Encyclopedia of GEOLOGY

Edited by
Richard C Selley
L Robin M Cocks
Ian R Plimer

ENCYCLOPEDIA OF
GEOLOGY

ENCYCLOPEDIA OF GEOLOGY

EDITED BY

RICHARD C. SELLEY
L. ROBIN M. COCKS
IAN R. PLIMER

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Elsevier Ltd., The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

First edition 2005

Library of Congress Control Number: 2004104445

A catalogue record for this book is available from the British Library

ISBN 0-12-636380-3 (set)

This book is printed on acid-free paper
Printed and bound in Spain

Editors

EDITORS

Richard C. Selley
Imperial College
London, UK

L. Robin M. Cocks
Natural History Museum
London, UK

Ian R. Plimer
University of Melbourne
Melbourne, VA
Australia

CONSULTANT EDITOR

Joe McCall
Cirencester
Gloucestershire, UK

Editorial Advisory Board

Foreword

Few areas of science can have changed as fast as geology has in the past forty years. In the first half of the last century geologists were divided, often bitterly, between the drifters and those who believed that the Earth and its continents were static. Neither side of this debate foresaw that the application of methods from physics, chemistry and mathematics to these speculations would revolutionize the study of all aspects of the Earth Sciences, and would lead to accurate and detailed reconstructions of world geography at former times, as well as to an understanding of the origin of the forces that maintain the continental movements. This change in world-view is no longer controversial, and is now embedded in every aspect of the Earth Sciences. It is a real pleasure to see this change, which has revitalized so many classic areas of research, reflected in the articles of this encyclopedia. Particularly affected are the articles on large-scale Earth processes, which discuss many of the new geological ideas that have come from geophysics and geochemistry. Forty years ago we had no understanding of these topics, which are fundamental to so many aspects of the Earth Sciences. The editors have decided, and in my view quite rightly, not to include detailed discussion of the present technology that is used to make geophysical and geochemical measurements. Such instrumental aspects are changing rapidly and become dated very quickly. They can easily be found in more technical publications. Instead the editors have concentrated on the influence such studies have had on our understanding of the Earth and its evolution, and in so doing have produced an excellent and accessible account of what is now known.

Any encyclopedia has to satisfy a wide variety of users, and in particular those who know that some subject like sedimentation or mineral exploration is part of geology, and go to an encyclopedia of geology to find out more. The editors have made a very thorough attempt to satisfy such users, and have included sections on such unexpected geological topics as the evolution of the Earth's atmosphere, the geology of Jupiter, Saturn, and their moons, aggregates, and creationism. I congratulate the editors and authors for producing such a fine summary of our present knowledge, and am particularly pleased that they intend to produce an online version of the encyclopedia. Though I have become addicted to using the Internet as my general encyclopedia, I will be delighted to be able to access something concerned with my own field that is as organized and scholarly as are these volumes.

Dan McKenzie
Royal Society Professor of Earth Sciences
Cambridge University, UK

Introduction

Civilization occurs by geological consent subject to change without notice....
Will Durant (1885 1981)

Richard de Bury, Bishop of Durham from 1333 to 1345, divided all knowledge into '*Geologia*', earthly knowledge, and '*Theologia*', heavenly knowledge. By the beginning of the last century, however, Geology was generally understood to be restricted to the study of rocks: according to the old dictum of the Geological Survey of Great Britain 'If you can hit it with a hammer, then it's geology.' Subsequently geology has been subsumed into Earth Science. This includes not only the study of rocks (the lithosphere), but also the atmosphere and hydrosphere and their relationship with the biosphere. Presently these relationships now form a nexus in Earth System Science.

The 'Encyclopedia of Geology' is what it says on the cover. What appealed to us when first approached to edit this work by Academic Press was a request that the encyclopedia should be rock-based. Readers are referred to the companion volumes, Encyclopedia of Atmospheric Sciences, Encyclopedia of the Solar System, Encyclopedia of Soils in the Environment and Encyclopedia of Ocean Sciences for knowledge on the other branches of Earth Science. Nonetheless we have extended our brief to include articles on the other planets and rocky detritus of our solar system, leaving others to argue, as no doubt Bishop Richard would have done, where the boundaries of earthly and heavenly knowledge might be. (His Grace would probably have charged the editors of the Encyclopedia of the Solar System with heresy.)

One of the first, and most difficult, tasks of editing this encyclopedia was to decide, not only which topics merited articles, but also how these articles should be grouped to facilitate the reader. This is easy for some branches of geology, but difficult for others. It is relatively easy to logically arrange articles on mineralogy and palaeontology, since they are defined by their chemistry and evolutionary biology. Articles that describe Earth history may be conveniently arranged in a chronological order, and articles on regional geology may be presented geographically. Other topics present problems, particularly in the area of sedimentology. There is, for example, a range of inter-related topics associated with deserts. This area could be described geomorphologically, and in terms of the aeolian and aqueous processes of deserts, aeolian sedimentary structures, and aeolian deposits. All of these aspects of deserts deserve mention, but there is no obvious logical way of arranging the discrete topics into articles. To help us in this task we relied heavily on our editorial board, whose individual members had more specialized knowledge of their field than we. To the Editorial Board Members, authors and anonymous referees of each article we give heartfelt thanks. We were also, of course, constrained by the willingness of expert authorities to contribute articles. To some degree therefore, the shape of the encylopedia owes as much to the enthusiasm of experts to write for us, as for our 'wish list' of articles. To facilitate readers finding their way around the Encyclopedia of Geology great care has been taken in cross-referencing within and between articles, in providing 'See Also' lists at the end of articles, and in the index. No doubt it will be easier for readers to navigate around the online version of the work, than to manipulate the several hard copy volumes.

As geological knowledge expands there is always more to learn and understand. While preparing the 'Encyclopedia of Geology' we have ourselves learned a great deal about geology, both within and beyond our own specialties. We invite you to read this encyclopedia and join us in the field trip of a lifetime.

Richard C. Selley
L. Robin M. Cocks
Ian R. Plimer
1 August 2004

References to related encyclopedia published by Elsevier, Academic Press:

Encyclopedia of the Solar System, 1998
Encyclopedia of Ocean Sciences, 2001
Encyclopedia of Atmospheric Sciences, 2002
Encyclopedia of Soils in the Environment, 2005

Guide to Use of the Encyclopedia

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of entries in alphabetical order. Most entries consist of several articles that deal with various aspects of a topic and are arranged in a logical sequence within an entry. Some entries comprise a single article.

To help you realize the full potential of the material in the Encyclopedia we have provided three features to help you find the topic of your choice: a Contents List, Cross-References and an Index.

1. Contents List

Your first point of reference will probably be the contents list. The complete contents lists, which appears at the front of each volume will provide you with both the volume number and the page number of the entry. On the opening page of an entry a contents list is provided so that the full details of the articles within the entry are immediately available.

Alternatively you may choose to browse through a volume using the alphabetical order of the entries as your guide. To assist you in identifying your location within the Encyclopedia a running headline indicates the current entry and the current article within that entry.

You will find 'dummy entries' where obvious synonyms exist for entries or where we have grouped together related topics. Dummy entries appear in both the contents lists and the body of the text.

Example

If you were attempting to locate material on erosional sedimentary structures via the contents list:

EROSION *see* **SEDIMENTARY PROCESSES: Fluxes and Budgets; Aeolian Processes; Erosional Sedimentary Structures.**

The dummy entry directs you to the **Erosional Sedimentary Structures** article, in the SEDIMENTARY PROCESSES entry. At the appropriate location in the contents list, the page numbers for articles under Sedimentary Processes are given.

If you were trying to locate the material by browsing through the text and you looked up **Erosion** then the following information would be provided in the dummy entry:

EROSION

See **SEDIMENTARY PROCESSES: Erosional Sedimentary Structures; Aeolian Processes; Fluxes and Budgets**

Alternatively, if you were looking up **Sedimentary Processes** the following information would be provided:

SEDIMENTARY PROCESSES

Contents

2. Cross-References

All of the articles in the Encyclopedia have been extensively cross-referenced.

The cross-references, which appear at the end of an article, serve three different functions. For example, at the end of the **PRECAMBRIAN: Overview** article, cross-references are used:

i. To indicate if a topic is discussed in greater detail elsewhere.

Africa: Pan-African Orogeny. **Antarctic**. **Asia:** Central. **Australia:** Proterozoic. **Biosediments and Biofilms**. **Earth Structure and Origins**. **Earth System Science**. **Europe:** East European Craton; Timanides of Northern Russia. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Indian Subcontinent**. **North America:** Precambrian Continental Nucleus; Continental Interior. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran. **Russia**. **Sedimentary Rocks:** Banded Iron Formations. **Shields**. **Terranes, Overview**.

ii. To draw the reader's attention to parallel discussions in other articles.

Africa: Pan-African Orogeny. **Antarctic**. **Asia:** Central. **Australia:** Proterozoic. **Biosediments and Biofilms**. **Earth Structure and Origins**. **Earth System Science**. **Europe:** East European Craton; Timanides of Northern Russia. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Indian Subcontinent**. **North America:** Precambrian Continental Nucleus; Continental Interior. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran. **Russia**. **Sedimentary Rocks:** Banded Iron Formations. **Shields**. **Terranes, Overview**.

iii. To indicate material that broadens the discussion.

Africa: Pan-African Orogeny. **Antarctic**. **Asia:** Central. **Australia:** Proterozoic. **Biosediments and Biofilms**. **Earth Structure and Origins**. **Earth System Science**. **Europe:** East European Craton; Timanides of Northern Russia. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Indian Subcontinent**. **North America:** Precambrian Continental Nucleus; Continental Interior. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran. **Russia**. **Sedimentary Rocks:** Banded Iron Formations. **Shields**. **Terranes, Overview**.

3. Index

The index will provide you with the page number where the material is located, and the index entries differentiate between material that is a whole article, is part of an article or is data presented in a figure or table. Detailed notes are provided on the opening page of the index.

4. Contributors

A full list of contributors appears at the beginning of each volume.

Contributors

Abart, R
University of Basel, Basel, Switzerland

Aldridge, R J
University of Leicester, Leicester, UK

Al-Jallal, I A
Sandroses Est. for Geological, Geophysical Petroleum Engineering Consultancy and Petroleum Services, Khobar, Saudi Arabia

Alkmim, F F
Universidade Federal de Ouro Preto, Ouro Preto, Brazil

Allen, P M
Bingham, Nottingham, UK

Allwood, A C
Macquarie University, Sydney, NSW, Australia

Al-Sharhan, A S
United Arab Emirates University, Al-Ain, United Arab Emirates

Anderson, L I
National Museums of Scotland, Edinburgh, UK

Arndt, N T
LCEA, Grenoble, France

Arnott, R
Oxford Institute for Energy Studies, Oxford, UK

Asimow, P D
California Institute of Technology, Pasadena, CA, USA

Atkinson, J
City University, London, UK

Bacon, M
Petro-Canada, London, UK

Bailey, J
Anglo-Australian Observatory and Australian Centre for Astrobiology, Sydney, Australia

Bani, P
Institut de la Recherche pour le Développement, Nouméa, New Caledonia

Bell, F G
British Geological Survey, Keyworth, UK

Bell, K
Carleton University, Ottawa, ON, Canada

Best, J
University of Leeds, Leeds, UK

Birch, W D
Museum Victoria, Melbourne, VIC, Australia

Bird, J F
Imperial College London, London, UK

Black, P
Auckland University, Auckland, New Zealand

Bleeker, W
Geological Survey of Canada, Ottawa, ON, Canada

Bogdanova, S V
Lund University, Lund, Sweden

Bommer, J J
Imperial College London, London, UK

Boore, D M
United States Geological Survey, Menlo Park, CA, USA

Bosence, D W J
Royal Holloway, University of London, Egham, UK

Boulanger, R W
University of California, Davis, CA, USA

Braga, J C
University of Granada, Granada, Spain

Branagan, D F
University of Sydney, Sydney, NSW, Australia

Brasier, M D
University of Oxford, Oxford, UK

Brewer, P A
University of Wales, Aberystwyth, UK

Bridge, M
University College London, London, UK

Brown, D
Instituto de Ciencias de la Tierra 'Jaume Almera' CSIC, Barcelona, Spain

Brown, A J
Macquarie University, Sydney, NSW, Australia

Brown, R J
University of Bristol, Bristol, UK

Bucher, K
University of Freiburg, Freiburg, Germany

Burns, S F
Portland State University, Portland, OR, USA

Byford, E
Broken Hill, NSW, Australia

Calder, E S
Open University, Milton Keynes, UK

Cameron, E M
Eion Cameron Geochemical Inc., Ottawa, ON, Canada

Carbotte, S M
Columbia University, New York, NY, USA

Carminati, E
Università La Sapienza, Rome, Italy

Chamberlain, S A
Macquarie University, Sydney, NSW, Australia

Charles, J A
Formerly Building Research Establishment
Hertfordshire, UK

Chiappe, L M
Natural History Museum of Los Angeles County
Los Angeles, CA, USA

Clack, J A
University of Cambridge, Cambridge, UK

Clayton, C
Eardiston, Tenbury Wells, UK

Clayton, G
Trinity College, Dublin, Ireland

Cocks, L R M
The Natural History Museum, London, UK

Coffin, M F
University of Tokyo, Tokyo, Japan

Collinson, J
John Collinson Consulting, Beech, UK

Comerford, G
The Natural History Museum, London, UK

Condie, K C
New Mexico Tech, Socorro, NM, USA

Cornford, C
Integrated Geochemical Interpretation Ltd, Bideford, UK

Cornish, L
The Natural History Museum, London, UK

Cosgrove, J W
Imperial College London, London, UK

Coxon, P
Trinity College, Dublin, Ireland

Cressey, G
The Natural History Museum, London, UK

Cribb, S J
Carraig Associates, Inverness, UK

Cronan, D S
Imperial College London, London, UK

Currant, A
The Natural History Museum, London, UK

Davies, H
University of Papua New Guinea, Port Moresby
Papua New Guinea

Davis, G R
Imperial College London, London, UK

DeCarli, P S
SRI International, Menlo Park, CA, USA

Dewey, J F
University of California Davis
Davis, CA, USA, and University of Oxford, Oxford, UK

Doglioni, C
Università La Sapienza, Rome, Italy

Dorning, K J
University of Sheffield, Sheffield, UK

Dott, Jr R H
University of Wisconsin, Madison, WI, USA

Doyle, P
University College London, London, UK

Dubbin, W E
The Natural History Museum, London, UK

Dyke, G J
University College Dublin, Dublin, Ireland

Echtler, H
GeoForschungsZentrum Potsdam, Potsdam, Germany

Eden, M A
Geomaterials Research Services Ltd, Basildon, UK

Eide, E A
Geological Survey of Norway, Trondheim, Norway

Eldholm, O
University of Bergen, Bergen, Norway

Elliott, D K
Northern Arizona University, Flagstaff, AZ, USA

Elliott, T
University of Liverpool, Liverpool, UK

Eriksen, A S
Zetica, Witney, UK

Fayers, S R
University of Aberdeen, Aberdeen, UK

Feenstra, A
GeoForschungsZentrum Potsdam, Potsdam, Germany

Felix, M
University of Leeds, Leeds, UK

Figueras, D
BFI, Houston, TX, USA

Fookes, P G
Winchester, UK

Forey, P L
The Natural History Museum, London, UK

Fortey, R A
The Natural History Museum, London, UK

Foster, D A
University of Florida, Gainesville, FL, USA

Frýda, J
Czech Geological Survey, Prague, Czech Republic

Franke, W
Johann Wolfgang Goethe-Universität
Frankfurt am Main, Germany

Franz, G
Technische Universität Berlin, Berlin, Germany

French, W J
Geomaterials Research Services Ltd, Basildon, UK

Fritscher, B
Munich University, Munich, Germany

Frostick, L
University of Hull, Hull, UK

Fuchs, Y
Université Marne la Vallée, Marne la Vallée, France

Gabbott, S E
University of Leicester, Leicester, UK

Garaebiti, E
Department of Geology and Mines, Port Vila, Vanuatu

Garetsky, R G
Institute of Geological Sciences, Minsk, Belarus

Garrard, P
Imperial College London, London, UK

Gascoyne, J K
Zetica, Witney, UK

Gee, D G
University of Uppsala, Uppsala, Sweden

Geshi, N
Geological Survey of Japan, Ibaraki, Japan

Giese, P
Freie Universität Berlin, Berlin, Germany

Giles, D P
University of Portsmouth, Portsmouth, UK

Glasser, N F
University of Wales, Aberystwyth, UK

Gluyas, J
Acorn Oil and Gas Ltd., Staines, UK

Gorbatschev, R
Lund University, Lund, Sweden

Gordon, J E
Scottish Natural Heritage, Edinburgh, UK

Gradstein, F M
University of Oslo, Oslo, Norway

Gray, D R
University of Melbourne, Melbourne, VIC, Australia

Greenwood, J R
Nottingham Trent University, Nottingham, UK

Grieve, R A F
Natural Resources Canada, Ottawa, ON, Canada

Griffiths, J S
University of Plymouth, Plymouth, UK

Hambrey, M J
University of Wales, Aberystwyth, UK

Hancock, J M[†]
Formerly Imperial College London, London, UK

Hansen, J M
Danish Research Agency, Copenhagen, Denmark

Harff, J
Baltic Sea Research Institute Warnemünde, Rostock, Germany

[†]Deceased

Harper, D A T
Geologisk Museum, Copenhagen, Denmark

Harper, E M
University of Cambridge, Cambridge, UK

Harrison, JP
Imperial College London, London, UK

Hatcher, Jr R D
University of Tennessee, Knoxville, TN, USA

Hatheway, A W
Rolla, MO and Big Arm, MT, USA

Hauzenberger, C A
University of Graz, Graz, Austria

Hawkins, A B
Charlotte House, Bristol, UK

Haymon, R M
University of California–Santa Barbara
Santa Barbara, CA, USA

He Guoqi
Peking University, Beijing, China

Head, J W
Brown University, Providence, RI, USA

Heim, N A
University of Georgia, Athens, GA, USA

Helvaci, C
Dokuz Eylül Üniversitesi, İzmir, Turkey

Hendriks, B W H
Geological Survey of Norway, Trondheim, Norway

Henk, A
Universität Freiburg, Freiburg, Germany

Herries Davies, G L
University of Dublin, Dublin, Ireland

Hey, R N
University of Hawaii at Manoa, Honolulu, HI, USA

Hoinkes, G
University of Graz, Graz, Austria

Hooker, J J
The Natural History Museum, London, UK

Horne, D J
University of London, London, UK

Hovland, M
Statoil, Stavanger, Norway

Howell, J
University of Bergen, Bergen, Norway

Howie, R A
Royal Holloway, University of London, London, UK

Hudson-Edwards, K
University of London, London, UK

Huggett, J M
Petroclays, Ashtead, UK and The Natural History Museum, London, UK

Hughes, N C
University of California, Riverside, CA, USA

Hutchinson, D R
US Geological Survey, Woods Hole, MA, USA

Idriss, I M
University of California, Davis, CA, USA

Ineson, J R
Geological Survey of Denmark and Greenland
Geocenter Copenhagen, Copenhagen, Denmark

Ivanov, M A
Russian Academy of Sciences, Moscow, Russia

Jäger, K D
Martin Luther University, Halle, Germany

Jarzembowski, E A
University of Reading, Reading, UK and Maidstone Museum and Bentlif Art Gallery, Maidstone, UK

Jones, B
University of Alberta, Edmonton, AB, Canada

Jones, G L
Conodate Geology, Dublin, Ireland

Joyner, L
Cardiff University, Cardiff, UK

Kaminski, M A
University College London, London, UK

Kay, S M
Cornell University, Ithaca, NY, USA

Kemp, A I S
University of Bristol, Bristol, UK

Kendall, A C
University of East Anglia, Norwich, UK

Kenrick, P
The Natural History Museum, London, UK

Kogiso, T
Japan Marine Science and Technology Center, Yokosuka, Japan

Krings, M
Bayerische Staatssammlung für Paläontologie und Geologie, Geo-Bio Center, Munich, Germany

Lancaster, N
Desert Research Institute, Reno, NV, and United States Geological Survey, Reston, VA, USA

Lang, K R
Tufts University, Medford, MA, USA

Laurent, G
Brest, France

Lee, E M
York, UK

Lemke, W
Baltic Sea Research Institute Warnemünde, Rostock Germany

Lesher, C M
Laurentian University, ON, Canada

Lewin, J
University of Wales, Aberystwyth, UK

Liu, J G
Imperial College London, London, UK

Long, J A
The Western Australian Museum, Perth WA, Australia

Loock, J C
University of the Free State Bloemfontein, South Africa

Lowell, R P
Georgia Institute of Technology, Atlanta, GA, USA

Lucas, S G
New Mexico Museum of Natural History Albuquerque, NM, USA

Lüning, S
University of Bremen, Bremen, Germany

Luo, Z-X
Carnegie Museum of Natural History Pittsburgh, PA, USA

Macdonald, K C
University of California–Santa Barbara Santa Barbara, CA, USA

Machel, H G
University of Alberta, Edmonton, Alberta, Canada

MacLeod, N
The Natural History Museum, London, UK

Maltman, A
University of Wales, Aberystwyth, UK

Martill, D M
University of Portsmouth, Portsmouth, UK

Martins-Neto, M A
Universidade Federal de Ouro Preto, Ouro Preto, Brazil

Marvin, U B
Harvard-Smithsonian Center for Astrophysics Cambridge, MA, USA

Mason, P J
HME Partnership, Romford, UK

Massonne, H-J
Universität Stuttgart, Stuttgart, Germany

Matte, P
University of Montpellier II, Montpellier, France

Mayor, A
Princeton, USA

McCaffrey, W
University of Leeds, Leeds, UK

McCall, G J H
Cirencester, Gloucester, UK

McCave, I N
University of Cambridge, Cambridge, UK

McGhee, G R
Rutgers University, New Brunswick, NJ, USA

McKibben, M A
University of California, CA, USA

McLaughlin, Jr P P
Delaware Geological Society, Newark, DE, USA

McManus, J
University of St. Andrews, St. Andrews, UK

McMenamin, M A S
Mount Holyoke College, South Hadley, MA, USA

Merriam, D F
University of Kansas, Lawrence, KS, USA

Metcalfe, I
University of New England, Armidale, NSW, Australia

Milke, R
University of Basel, Basel, Switzerland

Milner, A R
Birkbeck College, London, UK

Mojzsis, S J
University of Colorado, Boulder, CO, USA

Monger, J W H
Geological Survey of Canada, Vancouver, BC, Canada
and Simon Fraser University Burnaby, BC, Canada

Moore, P
Selsey, UK

Morris, N J
The Natural History Museum, London, UK

Mortimer, N
Institute of Geological and Nuclear Sciences, Dunedin
New Zealand

Mountney, N P
Keele University, Keele, UK

Mpodozis, C
SIPETROL SA, Santiago, Chile

Mungall, J E
University of Toronto, Toronto, ON, Canada

Myrow, P
Colorado College, Colorado Springs, CO, USA

Naish, D
University of Portsmouth, Portsmouth, UK

Nickel, E H
CSIRO Exploration and Mining, Wembley, WA, Australia

Nielsen, K C
The University of Texas at Dallas, Richardson, TX, USA

Nikishin, A M
Lomonosov Moscow State University, Moscow, Russia

Nokleberg, W J
United States Geological Survey, Menlo Park, CA, USA

Norbury, D
CL Associates, Wokingham, UK

O'Brien, P J
Universität Potsdam, Potsdam, Germany

Ogg, J G
Purdue University, West Lafayette, IN, USA

Oldershaw, C
St. Albans, UK

Oldroyd, D R
University of New South Wales, Sydney, Australia

Oneacre, J W
BFI, Houston, TX, USA

Orchard, M J
Geological Survey of Canada
Vancouver, BC, Canada

Orr, P J
University College Dublin, Dublin, Ireland

Owen, A W
University of Glasgow, Glasgow, UK

Pälike, H
Stockholm University, Stockholm, Sweden

Page, K N
University of Plymouth, Plymouth, UK

Paris, F
University of Rennes 1, Rennes, France

Parker, J R
Formerly Shell EP International, London, UK

Pfiffner, O A
University of Bern, Bern, Switzerland

Piper, D J W
Geological Survey of Canada, Dartmouth, NS, Canada

Price, R A
Queens University Kingston, ON, Canada

Prothero, D R
Occidental College, Los Angeles, CA, USA

Puche-Riart, O
Polytechnic University of Madrid, Madrid, Spain

Pye, K
Royal Holloway, University of London, Egham, UK

Rahn, P H
South Dakota School of Mines and Technology
Rapid City, SD, USA

Ramos, V A
Universidad de Buenos Aires, Buenos Aires, Argentina

Rankin, A H
Kingston University, Kingston-upon-Thames, UK

Rebesco, M
Istituto Nazionale di Oceanografia e di Geofisica
Sperimentale (OGS), Italy

Reedman, A J
Mapperley, UK

Reisz, R R
University of Toronto at Mississauga
Mississauga, ON, Canada

Retallack, G J
University of Oregon, Eugene, OR, USA

Rickards, R B
University of Cambridge, Cambridge, UK

Riding, R
Cardiff University, Cardiff, UK

Rigby, J K
Brigham Young University, Provo, UT, USA

Rigby, S
University of Edinburgh, Edinburgh, UK

Rodda, P
Mineral Resources Department, Suva, Fiji

Rona, P A
Rutgers University, New Brunswick, NJ, USA

Rose, E P F
Royal Holloway, University of London, Egham, UK

Rosenbaum, M S
Twickenham, UK

Rothwell, R G
Southampton Oceanography Centre, Southampton, UK

Roy, A B
Presidency College, Kolkata, India

Rushton, A W A
The Natural History Museum, London, UK

Russell, A J
University of Newcastle upon Tyne, Newcastle upon Tyne, UK

Schmid, R
ETH-centre, Zurich, Switzerland

Scott, E
National Center for Science Education
Berkeley, CA, USA

Scott, A C
Royal Holloway, University of London, Egham, UK

Scrutton, C T
Formerly University of Durham, Durham, UK

Searle, M
University of Oxford, Oxford, UK

Searle, R C
University of Durham, Durham, UK

Seibold, I
University Library, Freiburg, Germany

Selley, R C
Imperial College London, London, UK

Sellwood, B W
University of Reading, Reading, UK

Shields, G A
James Cook University, Townsville, QLD, Australia

Simms, M J
Ulster Museum, Belfast, UK

Slipper, I J
University of Greenwich, Chatham Maritime, UK

Smallwood, J R
Amerada Hess plc, London, UK

Smith, A B
The Natural History Museum, London, UK

Smith, I
Auckland University, Auckland, New Zealand

Snoke, A W
University of Wyoming, Laramie, WY, USA

Soligo, C
The Natural History Museum, London, UK

Stein, S
Northwestern University, Evanston, IL, USA

Steinberger, B
Japan Marine Science and Technology Center
Yokosuka, Japan

Stemmerik, L
Geological Survey of Denmark and Greenland,
Geocenter Copenhagen, Copenhagen, Denmark

Stern, R J
The University of Texas at Dallas, Richardson, TX, USA

Stewart, I
University of Plymouth, Plymouth, UK

Storey, B C
University of Canterbury, Christchurch, New Zealand

Storrs, G W
Cincinnati Museum Center, Museum of Natural History and Science, Cincinnati, OH, USA

Strachan, R A
University of Portsmouth, Portsmouth, UK

Suetsugu, D
Japan Marine Science and Technology Center, Yokosuka Japan

Surlyk, F
University of Copenhagen, Geocenter Copenhagen, Copenhagen, Denmark

Tait, J
Ludwig-Maximilians-Universität, München, Germany

Talbot, M R
University of Bergen, Bergen, Norway

Taylor, P D
The Natural History Museum, London, UK

Taylor, T N
University of Kansas, Lawrence, KS, USA

Taylor, W E G
University of Lancaster, Lancaster, UK

Tazawa, J
Niigata University, Niigata, Japan

Theodor, J M
Illinois State Museum, Springfield, IL, USA

Timmerman, M J
Universität Potsdam, Potsdam, Germany

Tollo, R P
George Washington University, Washington, DC, USA

Torsvik, T H
Geological Survey of Norway, Trondheim, Norway

Trendall, A
Curtin University of Technology, Perth, Australia

Trewin, N H
University of Aberdeen, Aberdeen, UK

Turner, A K
Colorado School of Mines, Colorado, USA

Twitchett, R J
University of Plymouth, Plymouth, UK

Tyler, I M
Geological Survey of Western Australia
East Perth, WA, Australia

Valdes, P J
University of Bristol, Bristol, UK

van Geuns, L C
Clingendael International Energy Programme
The Hague, The Netherlands

van Staal, C R
Geological Survey of Canada, Ottawa, ON, Canada

Vaněček, M
Charles University Prague, Prague, Czech Republic

Vaughan, D J
University of Manchester, Manchester, UK

Veevers, J J
Macquarie University, Sydney, NSW, Australia

Verniers, J
University of Ghent, Ghent, Belgium

Wadge, G
University of Reading, Reading, UK

Walter, M R
Macquarie University, Sydney, NSW, Australia

Wang, H
China University of Geosciences, Beijing, China

Ware, N G
Australian National University, Canberra, ACT, Australia

Warke, P A
Queen's University Belfast, Belfast, UK

Weber, K J
Technical University, Delft, The Netherlands

Welch, M D
The Natural History Museum, London, UK

Westbrook, G K
University of Birmingham, Birmingham, UK

Westermann, G E G
McMaster University, Hamilton, ON, Canada

Whalley, W B
Queen's University Belfast, Belfast, UK

White, N C
Brisbane, QLD, Australia

White, S M
University of South Carolina, Columbia, SC, USA

Wignall, P B
University of Leeds, Leeds, UK

Williams, P A
University of Western Sydney, Parramata, Australia

Wise, W S
University of California–Santa Barbara
Santa Barbara, CA, USA

Worden, R H
University of Liverpool, Liverpool, UK

Wyatt, A R
Sidmouth, UK

Xiao, S
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA

Yakubchuk, A S
The Natural History Museum, London, UK

Yates, A M
University of the Witwatersrand, Johannesburg
South Africa

Zhang Shihong
China University of Geosciences, Beijing, China

Ziegler, P A
University of Basel, Basel, Switzerland

Contents

Volume 1

A

B

C

D

E

Volume 2

F

Volume 3

G

H

I

†Deceased

J

L

M

Volume 4

N

O

P

Q

R

S

Volume 5

T

U

V

W

AFRICA

Contents

Pan-African Orogeny

A Kröner, Universität Mainz, Mainz, Germany
R J Stern, University of Texas-Dallas, Richardson TX, USA

Introduction

The term 'Pan-African' was coined by WQ Kennedy in 1964 on the basis of an assessment of available Rb–Sr and K–Ar ages in Africa. The Pan-African was interpreted as a tectono-thermal event, some 500 Ma ago, during which a number of mobile belts formed, surrounding older cratons. The concept was then extended to the Gondwana continents (Figure 1) although regional names were proposed such as Brasiliano for South America, Adelaidean for Australia, and Beardmore for Antarctica. This thermal event was later recognized to constitute the final part of an orogenic cycle, leading to orogenic belts which are currently interpreted to have resulted from the amalgamation of continental domains during the period ~870 to ~550 Ma. The term Pan-African is now used to describe tectonic, magmatic, and metamorphic activity of Neoproterozoic to earliest Palaeozoic age, especially for crust that was once part of Gondwana. Because of its tremendous geographical and temporal extent, the Pan-African cannot be a single orogeny but must be a protracted orogenic cycle reflecting the opening and closing of large oceanic realms as well as accretion and collision of buoyant crustal blocks. Pan-African events culminated in the formation of the Late Neoproterozoic supercontinent Gondwana (Figure 1). The Pan-African orogenic cycle is time-equivalent with the Cadomian Orogeny in western and central Europe and the Baikalian in Asia; in fact, these parts of Europe and Asia were probably part of Gondwana in pre-Palaeozoic times as were small Neoproterozoic crustal fragments identified in Turkey, Iran and Pakistan (Figure 1).

Within the Pan-African domains, two broad types of orogenic or mobile belts can be distinguished. One type consists predominantly of Neoproterozoic supracrustal and magmatic assemblages, many of juvenile (mantle-derived) origin, with structural and metamorphic histories that are similar to those in Phanerozoic collision and accretion belts. These belts expose upper to middle crustal levels and contain diagnostic features such as ophiolites, subduction- or collision-related granitoids, island-arc or passive continental margin assemblages as well as exotic terranes that permit reconstruction of their evolution in Phanerozoic-style plate tectonic scenarios. Such belts include the Arabian-Nubian shield of Arabia and north-east Africa (Figure 2), the Damara–Kaoko–Gariep Belt and Lufilian Arc of south-central and south-western Africa, the West Congo Belt of Angola and Congo Republic, the Trans-Sahara Belt of West Africa, and the Rokelide and Mauretanian belts along the western part of the West African Craton (Figure 1).

The other type of mobile belt generally contains polydeformed high-grade metamorphic assemblages, exposing middle to lower crustal levels, whose origin, environment of formation and structural evolution are more difficult to reconstruct. The protoliths of these assemblages consist predominantly of much older Mesoproterozoic to Archaean continental crust that was strongly reworked during the Neoproterozoic. Well studied examples are the Mozambique Belt of East Africa, including Madagascar (Figure 2) with extensions into western Antarctica, the Zambezi Belt of northern Zimbabwe and Zambia and, possibly, the little known migmatitic terranes of Chad, the Central African Republic, the Tibesti Massif in Libya and the western parts of Sudan and Egypt (Figure 1). It has been proposed that the latter type of belt represents the deeply eroded part of a collisional orogen and that the two types of Pan-African belts are not fundamentally different but constitute different crustal levels of collisional and/or accretional systems. For this reason, the term East African Orogen has been proposed for the combined upper crustal Arabian-Nubian Shield and lower crustal Mozambique Belt (Figure 2).

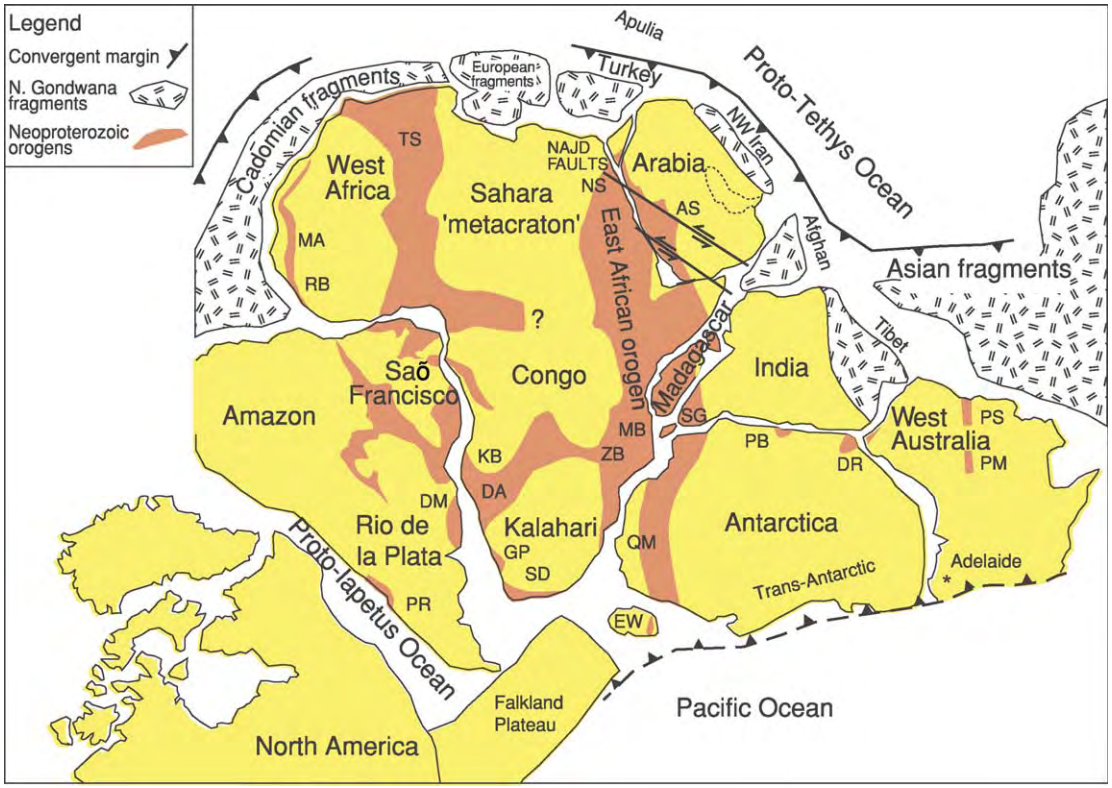


Figure 1 Map of Gondwana at the end of Neoproterozoic time (~540 Ma) showing the general arrangement of Pan African belts. AS, Arabian Shield; BR, Brasiliano; DA, Damara; DM, Dom Feliciano; DR, Denman Darling; EW, Ellsworth Whitmore Mountains; GP, Gariep; KB, Kaoko; MA, Mauretanides; MB, Mozambique Belt; NS, Nubian Shield; PM, Peterman Ranges; PB, Pryolz Bay; PR, Pampean Ranges; PS, Paterson; QM, Queen Maud Land; RB, Rokelides; SD, Saldania; SG, Southern Granulite Terrane; TS, Trans Sahara Belt; WB, West Congo; ZB, Zambezi. (Reproduced with permission from Kusky *et al.*, 2003.)

The Pan-African system of orogenic belts in Africa, Brazil and eastern Antarctica has been interpreted as a network surrounding older cratons (Figure 1) and essentially resulting from closure of several major Neoproterozoic oceans. These are the Mozambique Ocean between East Gondwana (Australia, Antarctica, southern India) and West Gondwana (Africa, South America), the Adamastor Ocean between Africa and South America, the Damara Ocean between the Kalahari and Congo cratons, and the Trans-Sahara Ocean between the West African Craton and a poorly known pre-Pan-African terrane in north-central Africa variously known as the Nile or Sahara Craton (Figure 1).

Arabian-Nubian Shield (ANS)

A broad region was uplifted in association with Cenozoic rifting to form the Red Sea, exposing a large tract of mostly juvenile Neoproterozoic crust. These exposures comprise the Arabian-Nubian Shield (ANS). The ANS makes up the northern half of the East African orogen and stretches from southern Israel and Jordan south as far as Ethiopia and Yemen, where the ANS transitions into the Mozambique Belt (Figure 2). The ANS is distinguished from the Mozambique Belt by its dominantly juvenile nature, relatively low grade of metamorphism, and abundance of island-arc rocks and ophiolites. The ANS, thus defined, extends about 3000 km north to south and >500 km on either side of the Red Sea (Figure 3). It is flanked to the west by a broad tract of older crust that was remobilized during Neoproterozoic time along with a significant amount of juvenile Neoproterozoic crust, known as the Nile Craton or 'Saharan Metacraton'. The extent of juvenile Neoproterozoic crust to the east in the subsurface of Arabia is not well defined, but it appears that Pan-African crust underlies most of this region. Scattered outcrops in Oman yielded mostly Neoproterozoic radiometric ages for igneous rocks, and there is no evidence that a significant body of pre-Pan-African crust underlies this region. The ANS is truncated to the north as a result of rifting at about the time of the Precambrian–Cambrian boundary, which generated crustal fragments now preserved in south-east Europe, Turkey and Iran.

The ANS is by far the largest tract of mostly juvenile Neoproterozoic crust among the regions of Africa that were affected by the Pan-African orogenic cycle. It

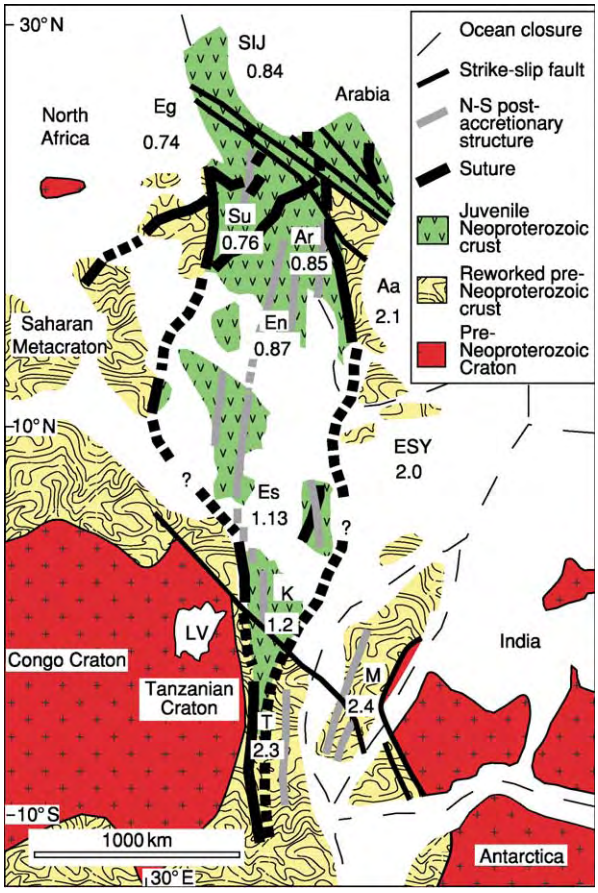


Figure 2 Pre Jurassic configuration of elements of the East African Orogen in Africa and surrounding regions. Regions include Egypt (Eg), Sudan (Su), Sinai Israel Jordan (SIJ), Afif terrane, Arabia (Aa), rest of Arabian Shield (Ar), Eritrea and northern Ethiopia (En), southern Ethiopia (Es), eastern Ethiopia, Somalia, and Yemen (ESY), Kenya (K), Tanzania (T), and Madagascar (M). Numbers in italics beneath each region label are mean Nd model ages in Gy.

formed as a result of a multistage process, whereby juvenile crust was produced above intra-oceanic convergent plate boundaries (juvenile arcs) and perhaps oceanic plateaux (*ca.* 870–630 Ma), and these juvenile terranes collided and coalesced to form larger composite terranes (Figure 4). There is also a significant amount of older continental crust (Mesoproterozoic age crust of the Afif terrane in Arabia; Palaeoproterozoic and Archaean crust in Yemen, Figure 2) that was overprinted by Pan-African tectonomagmatic events. ANS terrane boundaries (Figure 3) are frequently defined by suture zones that are marked by ophiolites, and the terranes are stitched together by abundant tonalitic to granodioritic plutons. Most ANS ophiolites have trace element chemical compositions suggesting formation above a convergent plate margin, either as part of a back-arc basin or in a fore-arc setting. Boninites have been identified in Sudan and Eritrea and suggest a fore-arc setting for at least some ANS sequences. Sediments are mostly immature sandstones and wackes derived from nearby arc volcanoes. Deposits that are diagnostic of Neoproterozoic 'snowball Earth' episodes have been recognized in parts of the ANS, and banded iron formations in the northern ANS may be deep-water expressions of snowball Earth events. Because it mostly lies in the Sahara and Arabian deserts, the ANS has almost no vegetation or soil and is excellently exposed. This makes it very amenable to study using imagery from remote sensing satellites.

Juvenile crust of the ANS was sandwiched between continental tracts of East and West Gondwana (Figure 4). The precise timing of the collision is still being resolved, but appears to have occurred after ~630 Ma when high-magnesium andesite 'schistose dykes' were emplaced in southern Israel but before the ~610 Ma post-tectonic 'Mereb' granites were emplaced in northern Ethiopia. By analogy with the continuing collision between India and Asia, the terminal collision between East and West Gondwana may have continued for a few tens of millions of years. Deformation in the ANS ended by the beginning of Cambrian time, although it has locally continued into Cambrian and Ordovician time farther south in Africa. The most intense collision (i.e. greatest shortening, highest relief, and greatest erosion) occurred south of the ANS, in the Mozambique belt. Compared to the strong deformation and metamorphism experienced during collision in the Mozambique belt, the ANS was considerably less affected by the collision. North-west trending left-lateral faults of the Najd fault system of Arabia and Egypt (Figures 1 and 2) formed as a result of escape tectonics associated with the collision and were active between about 630 and 560 Ma. Deformation associated with terminal collision is more intense in the southern ANS, with tight, upright folds, steep thrusts, and strike-slip shear zones controlling basement fabrics in Eritrea, Ethiopia, and southern Arabia. These north–south trending, collision-related structures obscure the earlier structures in the southern ANS that are related to arc accretion, and the intensity of this deformation has made it difficult to identify ophiolitic assemblages in southern Arabia, Ethiopia, and Eritrea. Thus, the transition between the ANS and the Mozambique Belt is marked by a change from less deformed and less metamorphosed, juvenile crust in the north to more deformed and more metamorphosed, remobilized older crust in the south, with the structural transition occurring farther north than the lithological transition.

The final stages in the evolution of the ANS witnessed the emplacement of post-tectonic 'A-type'

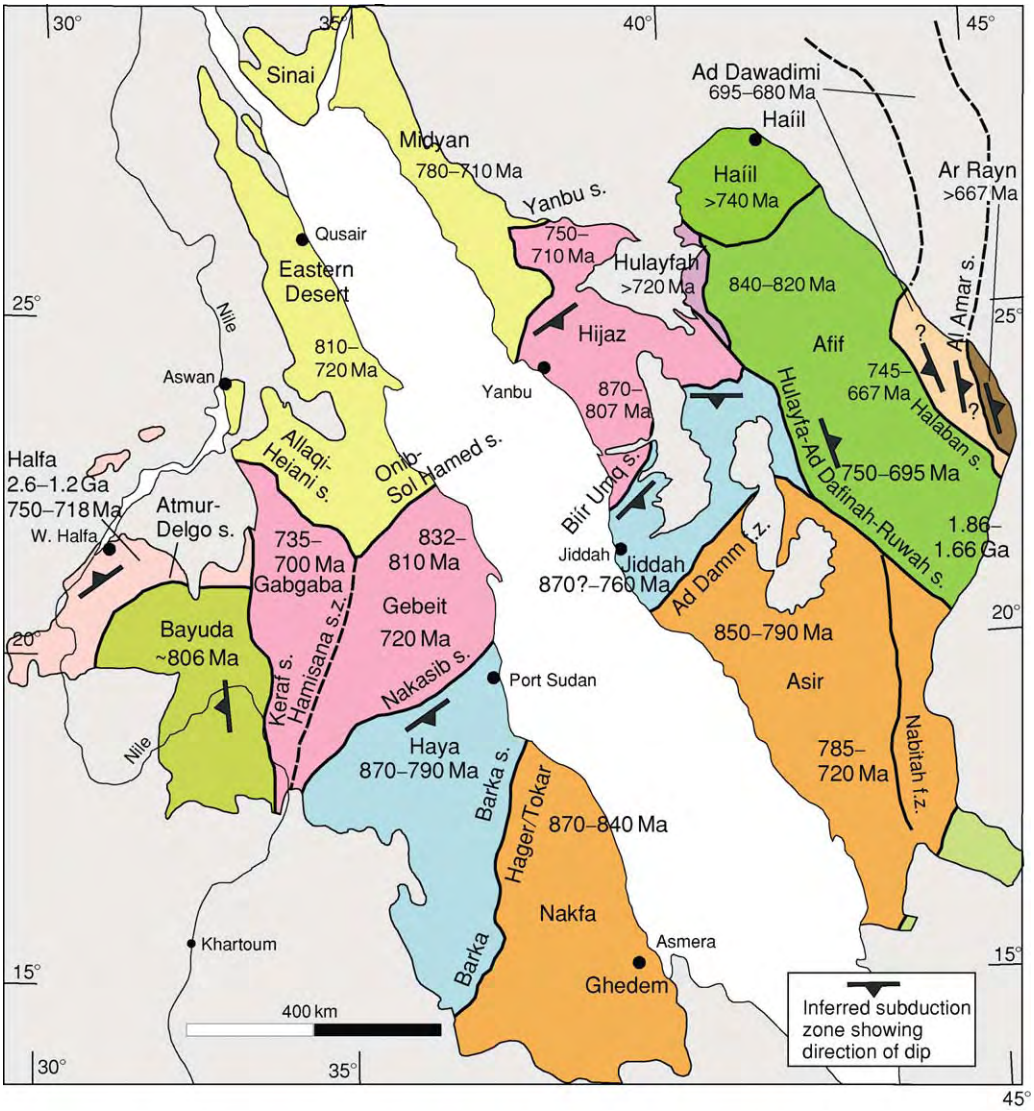


Figure 3 Terrane map of the Arabian Nubian Shield. (Reproduced with permission from Johnson PR and Woldehaimanot B (2003) Development of the Arabian Nubian Shield: perspectives on accretion and deformation in the northern East African Orogen and the assembly of Gondwana. In: Yoshida M, Windley BF and Dasgupta S (eds.) *Proterozoic East Gondwana: Supercontinent Assembly and Breakup*. Geological Society, London, Special Publications 206, pp. 289 325.)

granites, bimodal volcanics, and molassic sediments. These testify to strong extension caused by orogenic collapse at the end of the Neoproterozoic. Extension-related metamorphic and magmatic core complexes are recognized in the northern ANS but are even more likely to be found in the more deformed regions of the southern ANS and the Mozambique Belt. A well developed peneplain developed on top of the ANS crust before basal Cambrian sediments were deposited, possibly cut by a continental ice-sheet.

The ANS has been the source of gold since Pharaonic Egypt. There is now a resurgence of mining and exploration activity, especially in Sudan, Arabia, Eritrea, and Ethiopia.

Mozambique Belt (MB)

This broad belt defines the southern part of the East African Orogen and essentially consists of medium- to high-grade gneisses and voluminous granitoids. It extends south from the Arabian-Nubian Shield into southern Ethiopia, Kenya and Somalia via Tanzania to Malawi and Mozambique and also includes Madagascar (Figure 2). Southward continuation of the belt into Dronning Maud Land of East Antarctica (Figure 1) has been proposed on the basis of geophysical patterns, structural features and geochronology. Most parts of the belt are not covered by detailed mapping, making regional correlations difficult. There is no

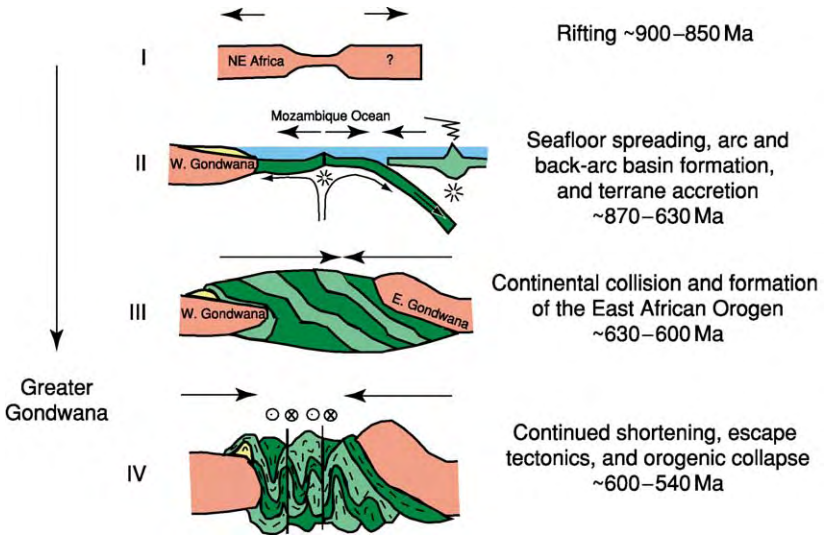


Figure 4 A diagram of the suggested evolution of the Arabian Nubian Shield.

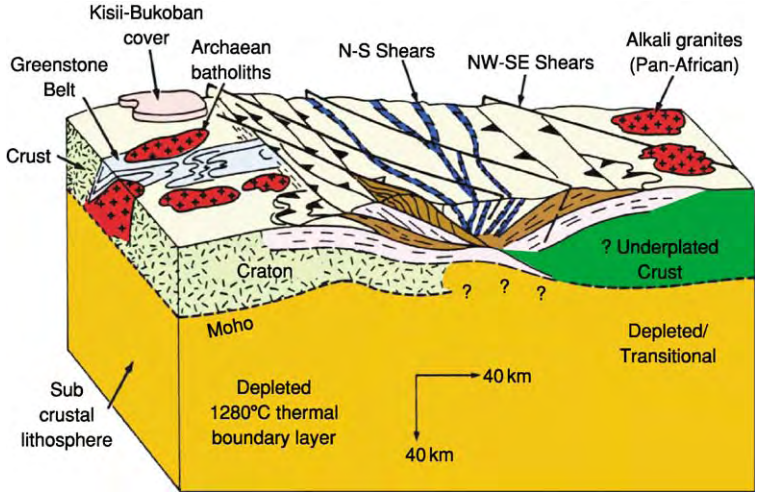


Figure 5 A schematic block diagram showing tectonic interdigitation of basement and cover rocks in the Mozambique Belt of Kenya. (Reproduced with permission from Mosley PN (1993) Geological evolution of the Late Proterozoic 'Mozambique Belt' of Kenya. *Tectonophysics* 221: 223 250.)

overall model for the evolution of the MB although most workers agree that it resulted from collision between East and West Gondwana. Significant differences in rock type, structural style, age and metamorphic evolution suggest that the belt as a whole constitutes a Pan-African Collage of terranes accreted to the eastern margin of the combined Congo and Tanzania cratons and that significant volumes of older crust of these cratons were reconstituted during this event.

Mapping and geochronology in Kenya have recognized undated Neoproterozoic supracrustal sequences that are structurally sandwiched between basement gneisses of Archaean and younger age (Figure 5). A ~700 Ma dismembered ophiolite complex at the Kenyan/Ethiopian border testifies to the consumption and obduction of marginal basin oceanic crust. Major deformation and high-grade metamorphism is ascribed to two major events at ~830 and ~620 Ma, based on Rb–Sr dating, but the older of these appears questionable.

A similar situation prevails in Tanzania where the metamorphic grade is generally high and many granulite-facies rocks of Neoproterozoic age show evidence of retrogression. Unquestionable Neoproterozoic supracrustal sequences are rare, whereas Late Archaean to Palaeoproterozoic granitoid gneisses volumetrically greatly dominate over juvenile Pan-African intrusives. These older rocks, strongly reworked during

the Pan-African orogenic cycle and locally migmatized and/or mylonitized, either represent eastward extensions of the Tanzania Craton that were structurally reworked during Pan-African events or are separate crustal entities (exotic blocks) of unknown origin. The significance of rare granitoid gneisses with protolith ages of ~1000–1100 Ma in southern Tanzania and Malawi is unknown. From these, some workers have postulated a major Kibaran (Grenvillian) event in the MB, but there is no geological evidence to relate these rocks to an orogeny. A layered gabbro-anorthosite complex was emplaced at ~695 Ma in Tanzania. The peak of granulite-facies metamorphism was dated at 620–640 Ma over wide areas of the MB in Tanzania, suggesting that this was the major collision and crustal-thickening event in this part of the belt.

In northern Mozambique the high-grade gneisses, granulites and migmatites of the MB were interpreted to have been deformed and metamorphosed during two distinct events, namely the Mozambican cycle at 1100–850 Ma, also known as Lurian Orogeny, and the Pan-African cycle at 800–550 Ma. Recent high-precision zircon geochronology has confirmed the older event to represent a major phase of granitoid plutonism, including emplacement of a large layered gabbro-anorthosite massif near Tete at ~1025 Ma, but there is as yet no conclusive evidence for deformation and granulite-facies metamorphism in these rocks during this time. The available evidence points to only one severe event of ductile deformation and high-grade metamorphism, with a peak some 615–540 Ma ago. A similar situation prevails in southern Malawi where high-grade granitoid gneisses with protolith ages of 1040–555 Ma were ductilely deformed together with supracrustal rocks and the peak of granulite-facies metamorphism was reached 550–570 Ma ago.

The Pan-African terrane of central and southern Madagascar primarily consists of high-grade ortho- and paragneisses as well as granitoids. Recent high-precision geochronology has shown that these rocks are either Archaean or Neoproterozoic in age and were probably structurally juxtaposed during Pan-African deformation. Several tectonic provinces have been recognized (Figure 6), including a domain consisting of low-grade Mesoproterozoic to Early Neoproterozoic metasediments known as the Itremo group which was thrust eastwards over high-grade gneisses. A Pan-African suture zone has been postulated in eastern Madagascar, the Betsimisaraka Belt (Figure 6), consisting of highly strained paragneisses decorated with lenses of mafic–ultramafic bodies containing podiform chromite and constituting a lithological and isotopic boundary with the Archaean gneisses and granites of the Antongil block east of this postulated suture which may correlate with similar rocks in southern India.

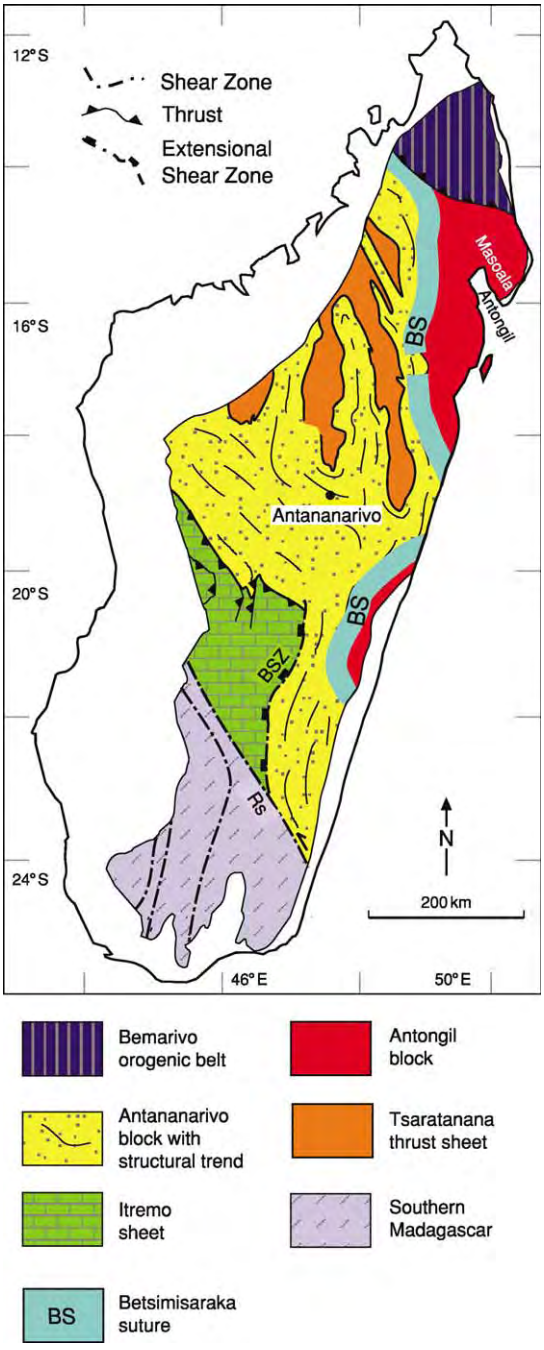


Figure 6 A simplified geological map showing the major tectonic units of the Precambrian basement in Madagascar. Rs, Ranotsara Shear Zone; BSZ, Betsileo Shear Zone. (Reproduced with permission from Collins and Windley 2002.)

Central and northern Madagascar are separated from southern Madagascar by the Ranotsara Shear Zone (Figure 6), showing sinistral displacement of >100 km and correlated with one of the major shear zones in southern India. Southern Madagascar consists of several north–south trending shear-bounded

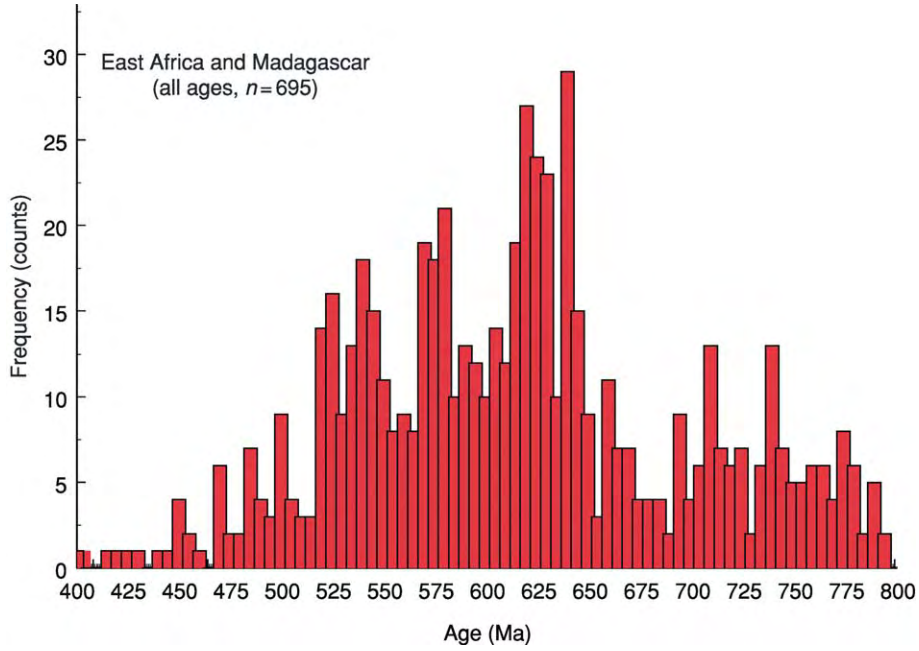


Figure 7 Histogram of radiometric ages for the Mozambique Belt of East Africa and Madagascar. Data from Meert JG (2003) A synopsis of events related to the assembly of eastern Gondwana. *Tectonophysics* 362: 1 40, with updates.

tectonic units consisting of upper amphibolite to granulite-facies para- and orthogneisses, partly of pre-Neoproterozoic age. The peak of granulite-facies metamorphism in central and southern Madagascar, including widespread formation of charnockites, was dated at 550–560 Ma.

The distribution of zircon radiometric ages in the MB suggests two distinct peaks at 610–660 and 530–570 Ma (Figure 7) from which two orogenic events have been postulated, the older East African Orogeny (~660–610 Ma) and the younger Kuunga Orogeny (~570–530 Ma). However, the are no reliable field criteria to distinguish between these postulated phases, and it is likely that the older age group characterizes syntectonic magmatism whereas the younger age group reflects post-tectonic granites and pegmatites which are widespread in the entire MB.

Zambezi Belt

The Zambezi Belt branches off to the west from the Mozambique Belt in northernmost Zimbabwe along what has been described as a triple junction and extends into Zambia (Figures 1 and 8). It consists predominantly of strongly deformed amphibolite- to granulite-facies, early Neoproterozoic ortho- and paragneisses which were locally intruded by ~860 Ma, layered gabbro-anorthosite bodies and generally displays south-verging thrusting and transpressional shearing. Lenses of eclogite record pressures up to 23 kbar. Although most of the above gneisses seem to be 850–870 Ma in age, there are tectonically interlayered granitoid gneisses with zircon ages around 1100 Ma. The peak of Pan-African metamorphism occurred at 540–535 Ma. The Zambezi Belt is in tectonic contact with lower-grade rocks of the Lufilian Arc in Zambia along the transcurrent Mwembeshi shear zone.

Lufilian Arc

The Lufilian Arc (Figure 8) has long been interpreted to be a continuation of the Damara Belt of Namibia, connected through isolated outcrops in northern Botswana (Figure 1). The outer part of this broad arc in the Congo Republic and Zambia is a north-east-verging thin-skinned, low-grade fold and thrust belt, whereas the higher-grade southern part is characterized by basement-involved thrusts. The main lithostratigraphic unit is the Neoproterozoic, copper-bearing Katanga succession which contains volcanic rocks dated between 765 and 735 Ma. Thrusting probably began shortly after deposition, and the main phase of thrusting and associated metamorphism occurred at 566–550 Ma.

Damara Belt

This broad belt exposed in central and northern Namibia branches north-west and south-east near

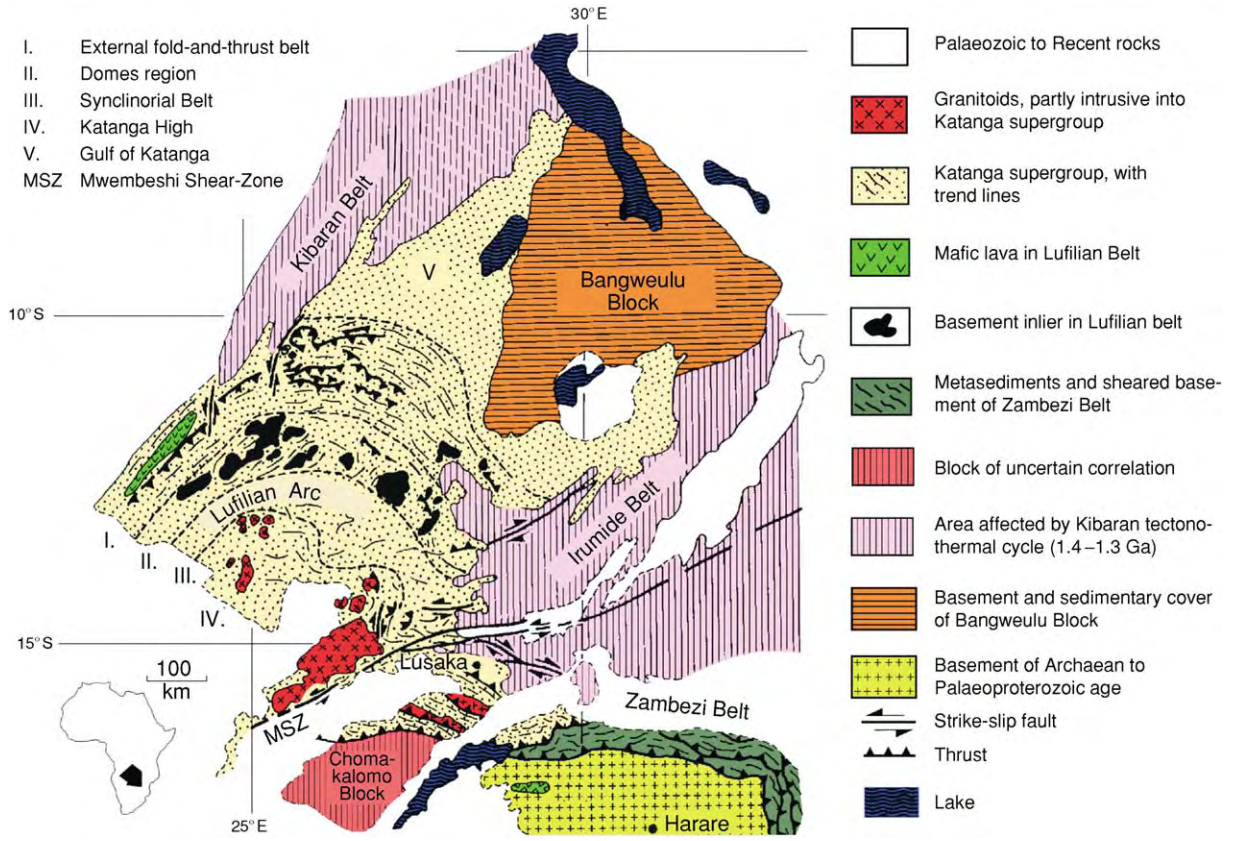


Figure 8 A simplified geological map of the Lufilian Arc and Zambezi Belt. (Reproduced with permission from Porada H and Berhorst V (2000) Towards a new understanding of the Neoproterozoic early Palaeozoic Lufilian and northern Zambezi belts in Zambia and the Democratic Republic of Congo.)

the Atlantic coast and continues southwards into the Gariep and Saldania belts and northwards into the Kaoko Belt (Figure 1). The triple junction so produced may have resulted from closure of the Adamastor Ocean, followed by closure of the Damara Ocean. The main lithostratigraphic unit is the Damara supergroup which records basin formation and rift-related magmatism at ~760 Ma, followed by the formation of a broad carbonate shelf in the north and a turbidite basin in the south. The turbidite sequence contains interlayered, locally pillowed, amphibolites and metagabbros which have been interpreted as remnants of a dismembered ophiolite. Of particular interest are two distinct horizons of glaciogenic rocks which can probably be correlated with similar strata in the Katanga sequence of south central Africa and reflect a severe glaciation currently explained by the snowball Earth hypothesis.

The Damara Belt underwent north- and south-verging thrusting along its respective margins, whereas the deeply eroded central zone exposes medium- to high-grade ductilely deformed rocks, widespread migmatization and anatexis in which both the Damara supracrustal sequence and a 1.0–2.0 Ga old basement are involved. Sinistral transpression is seen as the cause for this orogenic event which reached its peak at ~550–520 Ma. Voluminous pre-, syn- and post-tectonic granitoid plutons intruded the central part of the belt between ~650 and ~488 Ma, and highly differentiated granites, hosting one of the largest open-cast uranium mines in the world (Rössing), were dated at 460 Ma.

Uplift of the belt during the Damaran Orogeny led to erosion and deposition of two Late Neoproterozoic to Early Palaeozoic clastic molasse sequences, the Mulden group in the north and the Nama group in the south. The latter contains spectacular examples of the Late Neoproterozoic Ediacara fauna.

Gariep and Saldania Belts

These belts fringe the high-grade basement along the south-western and southern margin of the Kalahari craton (Figure 1) and are interpreted to result from oblique closure of the Adamastor Ocean. Deep marine fan and accretionary prism deposits, oceanic

seamounts and ophiolitic assemblages were thrust over Neoproterozoic shelf sequences on the craton margin containing a major Zn mineralization just north of the Orange River in Namibia. The main deformation and metamorphism occurred at 570–540 Ma, and post-tectonic granites were emplaced 536–507 Ma ago. The famous granite at Sea Point, Cape Town, which was described by Charles Darwin, belongs to this episode of Pan-African igneous activity.

Kaoko Belt

This little known Pan-African Belt branches off to the north-west from the Damara Belt and extends into south-western Angola. Here again a well developed Neoproterozoic continental margin sequence of the Congo Craton, including glacial deposits, was overthrust, eastwards, by a tectonic mixture of pre-Pan-African basement and Neoproterozoic rocks during an oblique transpressional event following closure of the Adamastor Ocean. A spectacular shear zone, the mylonite-decorated Puros lineament, exemplifies this event and can be followed into southern Angola. High-grade metamorphism and migmatization dated between 650 and 550 Ma affected both basement and cover rocks, and granitoids were emplaced between 733 and 550 Ma. Some of the strongly deformed basement rocks have ages between ~1450 and ~2030 Ma and may represent reworked material of the Congo Craton, whereas a small area of Late Archaean granitoid gneisses may constitute an exotic terrane. The western part of the belt consists of large volumes of *ca.* 550 Ma crustal melt granites and is poorly exposed below the Namib sand dunes. No island-arc, ophiolite or high-pressure assemblages have been described from the Kaoko Belt, and current tectonic models involving collision between the Congo and Rio de la Plata cratons are rather speculative.

West Congo Belt

This belt resulted from rifting between 999 and 912 Ma along the western margin of the Congo Craton (Figure 1), followed by subsidence and formation of a carbonate-rich foreland basin, in which the West Congolian group was deposited between *ca.* 900 and 570 Ma, including two glaciogenic horizons similar to those in the Katangan sequence of the Lufilian Arc. The structures are dominated by east-verging deformation and thrusting onto the Congo Craton, associated with dextral and sinistral transcurrent shearing, and metamorphism is low to medium grade. In the west, an allochthonous thrust-and-fold stack of Palaeo- to Mesoproterozoic basement rocks overrides the West Congolian foreland sequence. The West Congo Belt may only constitute the eastern part of an orogenic system with the western part, including an 800 Ma ophiolite, exposed in the Aracuaí Belt of Brazil.

Trans-Saharan Belt

This orogenic Belt is more than 3000 km long and occurs to the north and east of the >2 Ga West African Craton within the Anti-Atlas and bordering the Tuareg and Nigerian shields (Figure 1). It consists of pre-Neoproterozoic basement strongly reworked during the Pan-African event and of Neoproterozoic oceanic assemblages. The presence of ophiolites, accretionary prisms, island-arc magmatic suites and high-pressure metamorphic assemblages makes this one of the best documented Pan-African belts, revealing ocean opening, followed by a subduction- and collision-related evolution between 900 and 520 Ma (Figure 9). In southern Morocco, the ~740–720 Ma Sirwa-Bou Azzer ophiolitic mélange was thrust southwards, at ~660 Ma, over a Neoproterozoic continental margin sequence of the West African Craton, following northward subduction of oceanic lithosphere and preceding oblique collision with the Saghro Arc.

Farther south, in the Tuareg Shield of Algeria, Mali and Niger, several terranes with contrasting lithologies and origins have been recognized, and ocean closure during westward subduction produced a collision belt with Pan-African rocks, including oceanic terranes tectonically interlayered with older basement. The latter were thrust westwards over the West African Craton and to the east over the so-called LATEA (Laouni, Azrou-n-Fad, Tefedest, and Egéré-Aleksod, parts of a single passive margin in central Hoggar) Superterrane, a completely deformed composite crustal segment consisting of Archaean to Neoproterozoic assemblages (Figure 9). In Mali, the 730–710 Ma Tilemsi magmatic arc records ocean-floor and intra-oceanic island-arc formation, ending in collision at 620–600 Ma.

The southern part of the Trans-Saharan Belt is exposed in Benin, Togo and Ghana where it is known as the Dahomeyan Belt. The western part of this belt consists of a passive margin sedimentary sequence in the Volta basin which was overthrust, from the east, along a well delineated suture zone by an ophiolitic mélange and by a 613 My old high-pressure metamorphic assemblage (up to 14 kbar, ~700°C), including granulites and eclogites. The eastern part of the belt consists of a high-grade granitoid–gneiss terrane of the Nigerian province, partly consisting of Palaeoproterozoic rocks which were migmatized at ~600 Ma. This deformation and metamorphism is considered to have resulted from oblique collision of

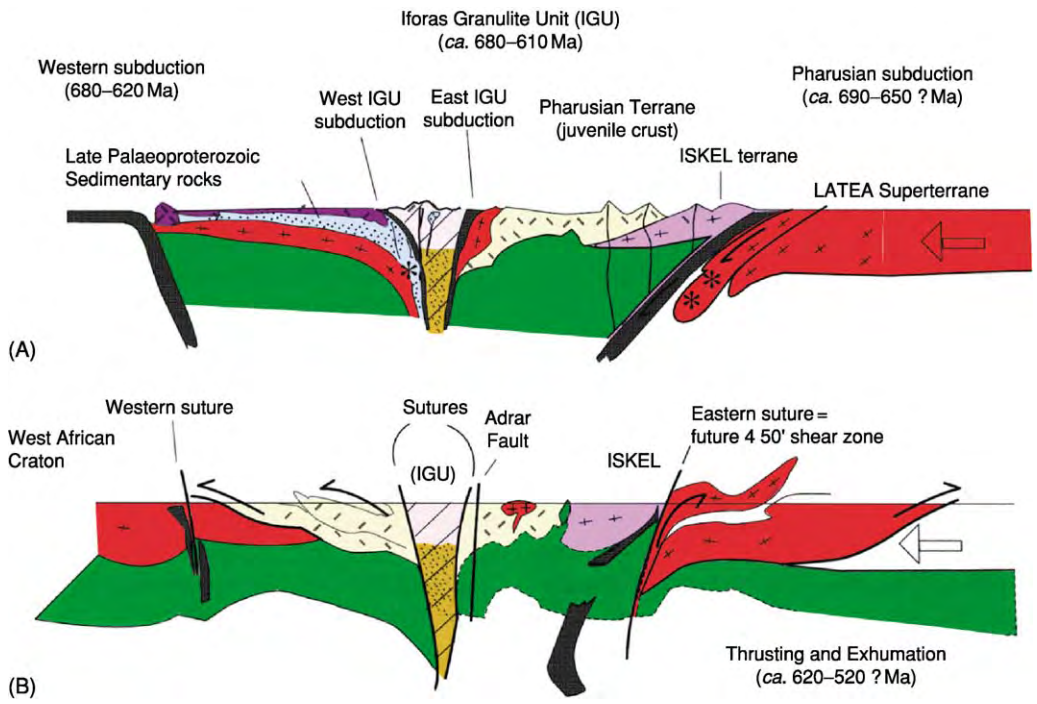


Figure 9 Diagrams showing the geodynamic evolution of western central Hoggar (Trans Sahara Belt) between ~900 and ~520 Ma. Stars denote high pressure rocks now exposed. (Reproduced with permission from Caby R (2003) Terrane assembly and geodynamic evolution of central western Hoggar: a synthesis.)

the Nigerian shield with the West African Craton, followed by anatectic doming and wrench faulting.

Pan-African Belt in Central Africa (Cameroon, Chad and Central African Republic)

The Pan-African Belt between the Congo Craton in the south and the Nigerian basement in the north-west consists of Neoproterozoic supracrustal assemblages and variously deformed granitoids with tectonically interlayered wedges of Palaeoproterozoic basement (Figure 10). The southern part displays medium- to high-grade Neoproterozoic rocks, including 620 Ma granulites, which are interpreted to have formed in a continental collision zone and were thrust over the Congo Craton, whereas the central and northern parts expose a giant shear belt characterized by thrust and shear zones which have been correlated with similar structures in north-eastern Brazil and which are late collisional features. The Pan-African Belt continues eastward into the little known Oubanguide Belt of the Central African Republic.

Pan-African Reworking of Older Crust in North-Eastern Africa

A large area between the western Hoggar and the river Nile largely consists of Archaean to Palaeoproterozoic basement, much of which was structurally and thermally overprinted during the Pan-African event and intruded by granitoids. The terrane is variously known in the literature as 'Nile Craton', 'East Sahara Craton' or 'Central Sahara Ghost Craton' and is geologically poorly known. Extensive reworking is ascribed by some to crustal instability following delamination of the subcrustal mantle lithosphere, and the term 'Sahara Metacraton' has been coined to characterize this region. A 'metacraton' refers to a craton that has been remobilized during an orogenic event but is still recognizable through its rheological, geochronological and isotopic characteristics.

Rokelide Belt

This belt occurs along the south-western margin of the Archaean Man Craton of West Africa (Figure 1) and is made up of high-grade gneisses, including granulites (Kasila group), lower-grade supracrustal sequences (Marampa group) and volcano-sedimentary rocks with calc-alkaline affinity (Rokel River group). Pan-African deformation was intense and culminated in extensive thrusting and sinistral strike-slip deformation. The peak of metamorphism reached 7 kb and 800°C and was dated at ~560 Ma. Late Pan-African emplacement ages for the protoliths of some of the granitoid gneisses contradict earlier hypotheses arguing for extensive overprinting of

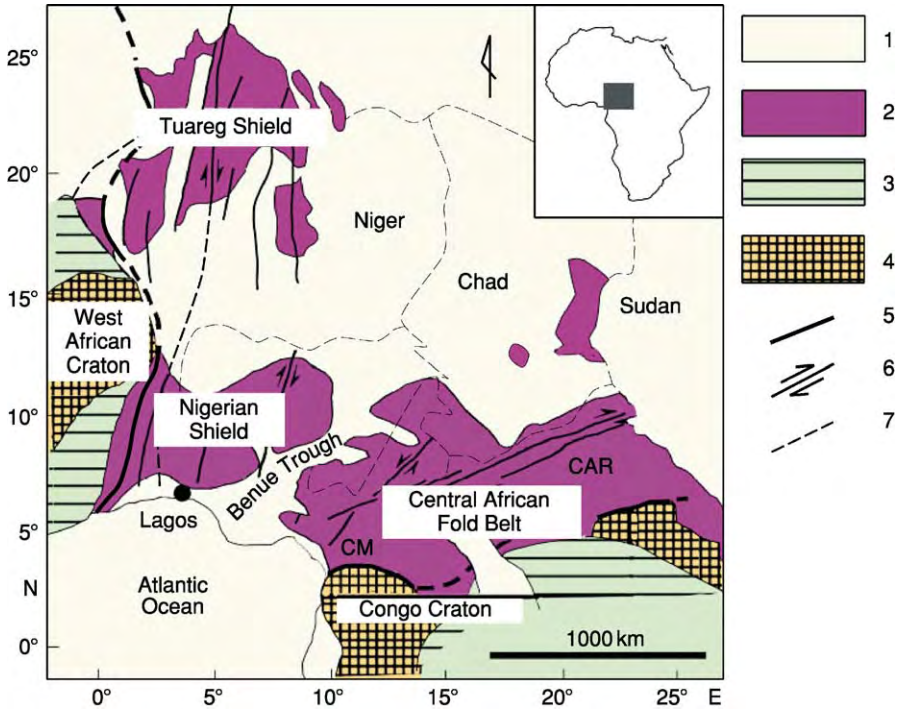


Figure 10 A sketch map showing Pan African domains in west central Africa. 1, Post Pan African cover; 2, Pan African domains; 3, pre Mesozoic platform deposits; 4, Archaean to Palaeoproterozoic cratons; 5, craton limits; 6, major strike slip faults; 7, state boundaries. CAR, Central African Republic; CM, Cameroun. (Reproduced with permission from Toteu SF, Penaye J and Djomani YP (2004).)

Archaean rocks. The Rokelides may be an accretionary belt, but there are no modern structural data and only speculative geodynamic interpretations.

Gondwana Correlations

The Pan-African orogenic cycle was the result of ocean closure, arc and microcontinent accretion and final suturing of continental fragments to form the supercontinent Gondwana. It has been suggested that the opening of large Neoproterozoic oceans between the Brazilian and African cratons (Adamastor Ocean), the West African and Sahara–Congo cratons (Pharusian Ocean) and the African cratons and India/Antarctica (Mozambique Ocean) (Figure 1) resulted from breakup of the Rodinia supercontinent some 800–850 Ma, but current data indicate that the African and South American cratons were never part of Rodinia. Although arc accretion and continent formation in the Arabian-Nubian shield are reasonably well understood, this process is still very speculative in the Mozambique Belt. It seems clear that Madagascar, Sri Lanka, southern India and parts of East Antarctica were part of this process (Figure 1), although the exact correlations between these fragments are not known. The Southern Granulite Terrane of India (Figure 1) consists predominantly of Late Archaean to Palaeoproterozoicc gneisses and granulites, deformed and metamorphosed during the Pan-African event and sutured against the Dharwar Craton. Areas in East Antarctica such as Lützow-Holm Bay, Central Dronning Maud Land and the Shackleton Range, previously considered to be Mesoproterozoic in age, are now interpreted to be part of the Pan-African Belt system (Figure 1). Correlations between the Pan-African belts in south-western Africa (Gariep–Damara–Kaoko) and the Brasiliano belts of south-eastern Brazil (Ribeira and Dom Feliciano) are equally uncertain, and typical hallmarks of continental collision such as ophiolite-decorated sutures or high-pressure metamorphic assemblages have not been found. The most convincing correlations exist between the southern end of the Trans-Saharan Belt in West Africa and Pan-African terranes in north-eastern Brazil (Figure 1). Following consolidation of the Gondwana supercontinent at the end of the Precambrian, rifting processes at the northern margin of Gondwana led to the formation of continental fragments (Figure 1) which drifted northwards and are now found as exotic terranes in Europe (Cadomian and Armorican terrane assemblages), in the Appalachian Belt of North

America (Avalonian Terrane assemblage) and in various parts of central and eastern Asia.

See Also

Arabia and The Gulf. **Australia:** Proterozoic. **Brazil**. **Gondwanaland and Gondwana**. **Palaeomagnetism**. **Tectonics:** Mountain Building and Orogeny. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Abdelsalam MG and Stern RJ (1997) Sutures and shear zones in the Arabian Nubian Shield. *Journal of African Earth Sciences* 23: 289 310.

Caby R (2003) Terrane assembly and geodynamic evolution of central western Hoggar: a synthesis. *Journal of African Earth Sciences* 37: 133 159.

Cahen L, Snelling NJ, Delhal J, and Vail JR (1984) *The Geochronology and Evolution of Africa*. Oxford: Clarendon Press.

Clifford TN (1968) Radiometric dating and the pre Silurian geology of Africa. In: Hamilton EI and Farquhar RM (eds.) *Radiometric Dating for Geologists*, pp. 299 416. London: Interscience.

Collins AS and Windley BF (2002) The tectonic evolution of central and northern Madagascar and its place in the final assembly of Gondwana. *Journal of Geology* 110: 325 339.

Fitzsimons ICW (2000) A review of tectonic events in the East Antarctic shield and their implications for Gon dwana and earlier supercontinents. *Journal of African Earth Sciences* 31: 3 23.

Hanson RE (2003) Proterozoic geochronology and tectonic evolution of southern Africa. In: Yoshida M, Windley BF, and Dasgupta S (eds.) *Proterozoic East Gondwana: Supercontinent Assembly and Breakup*. Geological Soci ety, London, Special Publications 206, pp. 427 463.

Hoffman PF and Schrag DP (2002) The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* 14: 129 155.

Johnson PR and Woldehaimanot B (2003) Development of the Arabian Nubian Shield: perspectives on accretion and deformation in the northern East African orogen and the assembly of Gondwana. In: Yoshida M, Windley BF, and Dasgupta S (eds.) *Proterozoic East Gondwana: Supercontinent Assembly and Breakup*. Geological Society, London, Special Publications 206, pp. 289 325.

Kröner A (2001) The Mozambique belt of East Africa and Madagascar; significance of zircon and Nd model ages for Rodinia and Gondwana supercontinent formation and dispersal. *South African Journal of Geology* 104: 151 166.

Kusky TM, Abdelsalam M, Stern RJ, and Tucker RD (eds.) (2003) Evolution of the East African and related oro gens, and the assembly of Gondwana. *Precambrian Res.* 123: 82 85.

Meert JG (2003) A synopsis of events related to the assembly of eastern Gondwana. *Tectonophysics* 362: 1 40.

Miller RMcG (ed.) (1983) *Evolution of the Damara Oro gen of South West Africa/Namiba*. Geological Society of South Africa, Special Publications, 11.

Mosley PN (1993) Geological evolution of the late Protero zoic 'Mozambique Belt' of Kenya. *Tectonophysics* 221: 223 250.

Porada H and Berhorst V (2000) Towards a new under standing of the Neoproterozoic early Palaeozoic Lufilian and northern Zambezi belts in Zambia and the Demo cratic Republic of Congo. *Journal of African Earth Sciences* 30: 727 771.

Stern RJ (1994) Arc assembly and continental collision in the Neoproterozoic East African Orogen: implications for the consolidation of Gondwanaland. *Annual Reviews Earth Planetary Sciences* 22: 319 351.

Toteu SF, Penaye J, and Djomani YP (2004) Geodynamic evolution of the Pan African belt in central Africa with special reference to Cameroon. *Canadian Journal of Earth Science* 41: 73 85.

Veevers JJ (2003) Pan African is Pan Gondwanaland: ob lique convergence drives rotation during 650 500 Ma assembly. *Geology* 31: 501 504.

North African Phanerozoic

S Lüning, University of Bremen, Bremen, Germany

Introduction

North Africa forms the northern margin of the African Plate and comprises the countries Morocco, Algeria, Tunisia, Libya, and Egypt (Figure 1). The region discussed here is bounded to the west by the Atlantic, to the north by the Mediterranean Sea, to the east by the Arabian Plate and to the south by political boundaries. Much of the geology across North Africa is remarkably uniform because many geological events affected the whole region (Figure 2). The geological study of North Africa benefits from large-scale desert exposures and an extensive subsurface database from hydrocarbon exploration. The region contains some 4% of the world's remaining oil (*see* **Petroleum Geology:** Overview) and gas reserves with fields mainly in Algeria, Libya, and

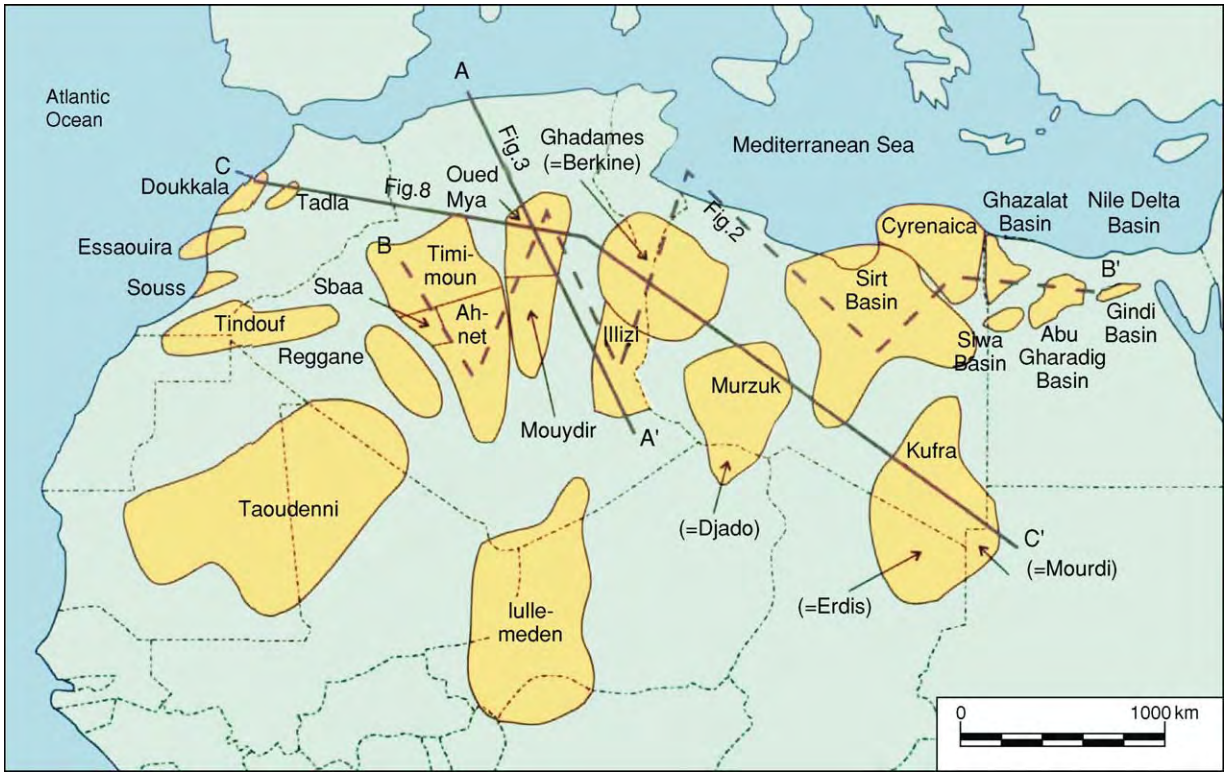


Figure 1 Location of major North African sedimentary basins. Lines indicate locations of cross sections in **Figures 2, 3** and **8**.

Egypt. Other natural resources that are exploited include Saharan fossil groundwater, phosphate, (*see* **Sedimentary Rocks:** Phosphates) and mineral ores.

Structural Evolution

Most of North Africa has formed part of a single plate throughout the Phanerozoic with the exception of the Atlas Mountains which became accreted during Late Carboniferous and Tertiary collisional events. North Africa can be structurally subdivided into a northern Mesozoic to Alpine deformed, mobile belt and the stable Saharan Platform (Figure 3). The latter became consolidated during the Proterozoic Pan-African Orogeny (*see* **Africa:** Pan-African Orogeny), a collisional amalgamation between the West African Craton and numerous island arcs, Andean-type magmatic arcs, and various microplates. The Late Neoproterozoic to Phanerozoic structural development of North Africa can be divided into six major tectonic (*see* **Plate Tectonics**) phases: (i) Infracambrian extension and wrenching; (ii) Cambrian to Carboniferous alternating extension and compression; (iii) mainly Late Carboniferous 'Hercynian' intraplate uplift; (iv) Late Triassic–Early Jurassic and Early Cretaceous rifting; (v) mid-Cretaceous 'Austrian' and Late Cretaceous–Tertiary 'Alpine' compression, and (vi) Oligo-Miocene rifting (*see* **Tectonics:** Rift Valleys).

Infracambrian Extension and Wrenching

The Late Neoproterozoic to Early Cambrian ('Infracambrian') in North Africa and Arabia was characterized by major extensional and strike-slip movements. Halfgrabens and pull-apart basins developed, for example, in the Taoudenni Basin (SW Algeria) and in the Kufra Basin (SE Libya). These features are considered to be a westward continuation of an Infracambrian system of salt basins extending across Gondwana from Australia, through Pakistan, Iran and Oman, to North Africa.

Post-Infracambrian – Pre-Hercynian

The structural evolution of North Africa between the Infracambrian extensional/wrenching phase and the Late Carboniferous 'Hercynian Orogeny' is complex. Local transpressional and transtensional reactivation processes dominated as a result of the interaction of intraplate stress fields with pre-existing fault systems of varying orientation and geometry. In some areas, such as the Murzuq Basin in SW Libya, these tectonic processes played an important role in the formation of hydrocarbon traps.

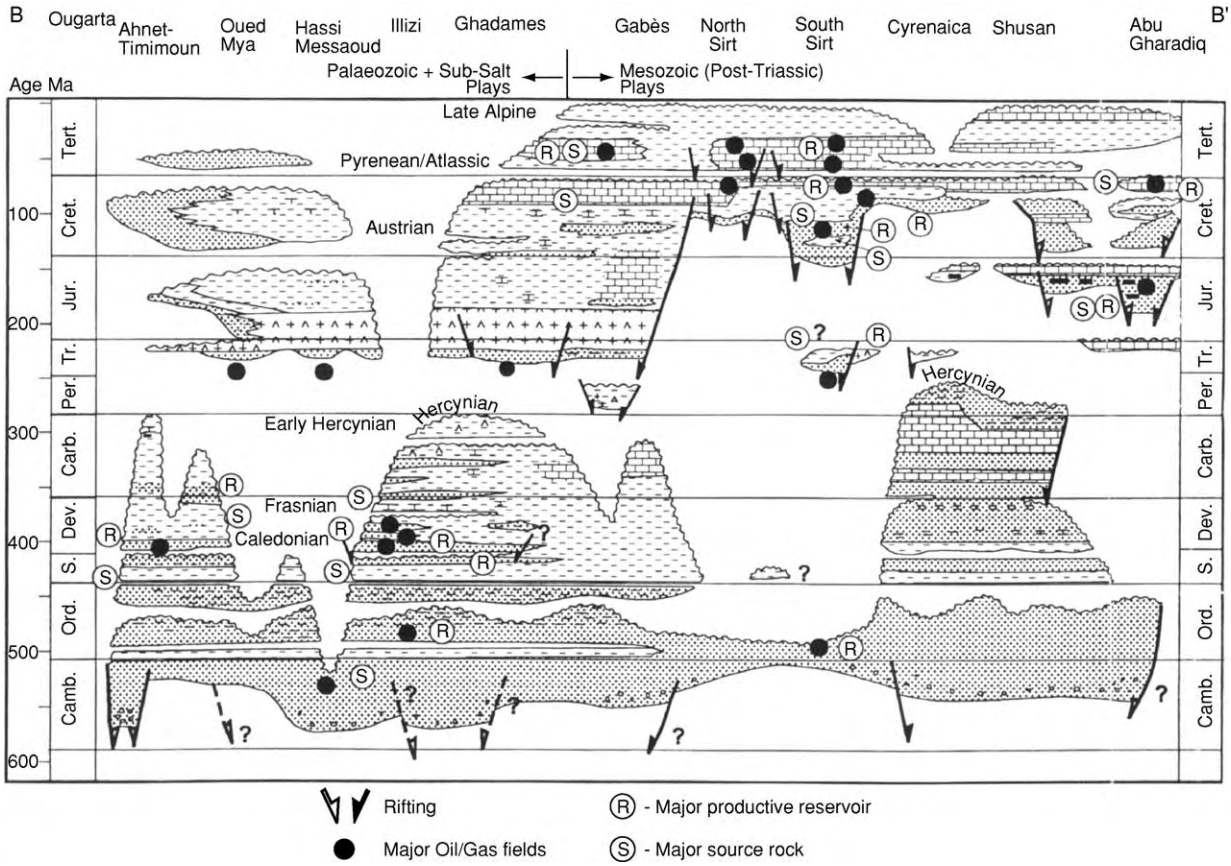


Figure 2 Phanerozoic chronostratigraphy of petroliferous provinces in North Africa. (From MacGregor (1998).)

During most of the Early Palaeozoic the Saharan Palaeozoic basins were part of a large, interconnected North African shelf system that was in a sagging phase. Some relief, however, was locally already created associated with local uplift and increased subsidence, including, for example, late Cambrian uplift in the Hoggar and increased sagging in the SE Libyan Kufra Basin, the latter leading to thinning of Cambro-Ordovician strata towards the present-day basin margins. The Saharan basins differentiated mainly from the Late Silurian/Early Devonian onwards when ridges were uplifted, associated with a basal unconformity, that in the regional literature has often been referred to as 'Caledonian unconformity'. This term, however, is inappropriate as tectonic events during the Silurian in North Africa were independent of those in the 'Caledonian' collisional zone, located many thousands of kilometres to the north, involving the continents of Laurentia, Baltica, Armorica, and Avalonia.

Hercynian Orogeny

Collision of Gondwana and Laurasia during the Late Carboniferous resulted in the compressional movements of the Hercynian Orogeny (Figure 4). In North Africa, the collisional zone was located in the north-west, leading to substantial thrusting and uplift in Morocco and western Algeria. Strong uplift associated with transpression on old faults occurred in the Algerian Hassi Massaoud region, leading to erosion into stratigraphic levels as deep as the Cambrian. The intensity of Hercynian deformation decreases eastwards across North Africa such that strong folding and erosion of anticlinal crests in the Algerian Sbaa and Ahnet basins is replaced towards the plate interior by low-angle unconformities and disconformities in the Murzuq Basin in south-west Libya. Notably, the present-day maturity levels of the main Palaeozoic hydrocarbon source rocks have a decreasing trend eastwards across North Africa (once present-day burial effects are removed) in parallel with the decrease in the intensity of the Hercynian deformation.

The gravitational collapse of the Hercynian Orogenic Belt in north-west Africa was accompanied by widespread Permo-Carboniferous volcanism in Morocco. The magmatism acted here as an 'exhaust valve' releasing the heat accumulated beneath the

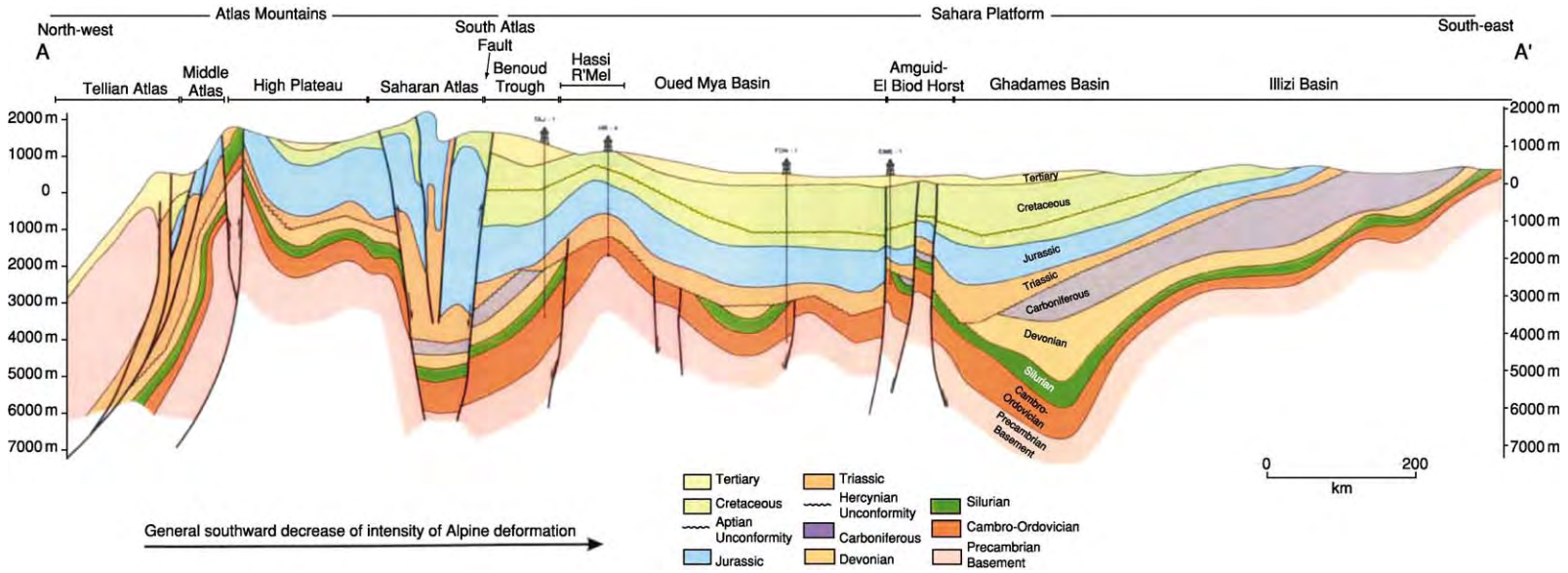


Figure 3 Cross section through Algeria illustrating typical the structural styles of North Africa. The Alpine deformed Atlas Mountains are separated from the Saharan Platform to the south by the South Atlas Fault. Location of section in **Figure 1**. (Courtesy E. Zanella.)

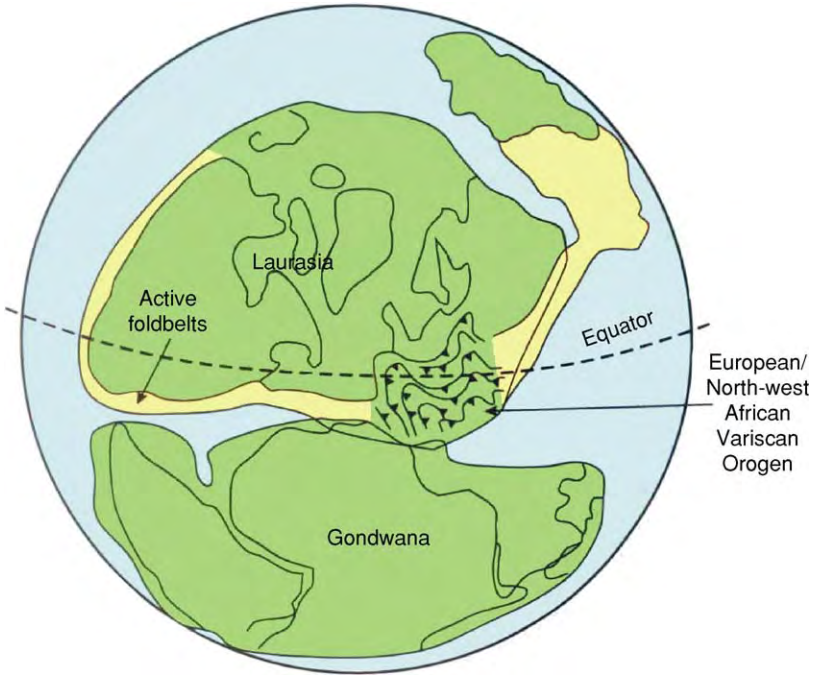


Figure 4 Hercynian compression as result of a Late Carboniferous plate collision between Laurasia and Gondwana. (After Doblas *et al.* (1998).)

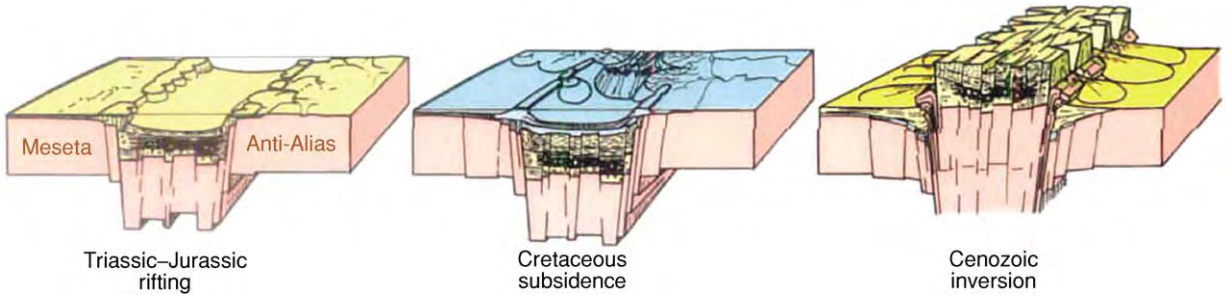


Figure 5 Block diagrams illustrating the geological evolution of the High Atlas, including Triassic Jurassic rifting, Cretaceous and Cenozoic inversion. (After Stets and Wurster (1981).)

Pangaean Supercontinent by insulation and blanketing processes which triggered large-scale mantle-wide upward convection and general instability of the supercontinent.

Mesozoic Extension

The opening of the Central Atlantic in the Triassic–Early Jurassic and contemporaneous separation of the Turkish–Apulian Terrane from north-east Africa initiated a significant extensional phase in North Africa which included graben formation in the Atlas region (Figure 5), rifting from Syria to Cyrenaica (NE Libya) and extension in offshore Libya and in the Oued Mya and Ghadames (=Berkine) basins in central and eastern Algeria. Rift-related Triassic volcanism occured in the northern Ghadames and Oued Maya basins.

A second important Mesozoic extensional phase in North Africa occurred during the Early Cretaceous, related to the opening of the South and Equatorial Atlantic Ocean. As a result, a complex of failed rift systems originated across North and Central Africa with the formation of half-grabens in, for example, the Egyptian Abu Gharadig Basin and in the Libyan Sirte Basin.

The Mesozoic extensional phase also triggered increased subsidence in several Saharan Palaeozoic basins, leading to deposition of thick, continental

deposits, for example, in the south-east Libyan Kufra Basin.

Alpine Orogeny

The onset of rifting in the northern North Atlantic during the Late Cretaceous led to an abrupt change in the motion of the European Plate which began to move eastwards with respect to Africa. The previous sinistral transtensional movements were quickly replaced by a prolonged phase of dextral transpression resulting in the collision of Africa and Europe. The 'Alpine Orogeny' led to an overall compressional regime in North Africa from the mid-Cretaceous through to the Recent. Changes in the collisional process, such as subduction of oceanic crust after accretion of a seamount in the Eastern Mediterranean, produced localized stress-neutral or even extensional pulses within the overall compressive regime.

An Aptian compressional event may be considered as a precursor to the 'Alpine Orogeny', in the narrow sense. It affected parts of North and Central Africa, inverting Early Cretaceous rift systems and reactivating older structures. Large Aptian-age anticlines occur in the Berkine Basin in Algeria and result from sinistral transpression along the N–S trending Transaharian fracture system.

The post-Cenomanian 'Alpine' compression in North Africa resulted in folding and thrusting within the north-west African collisional zones, as well as in intraplate inversion and uplift of Late Triassic-Early Jurassic grabens. Major orogens formed during this phase include the Atlas Mountains (Morocco, Algeria, Tunisia; Figure 5) and the 'Syrian Arc' Fold Belt in north-east Egypt and north-west Arabia. The Cyrenaica Platform (Jebel Akhdar) in north-east Libya also is an 'Alpine' deformed region.

The structural boundary between the Atlas Mountains and the Saharan Platform is the South Atlas Front (South Atlas Fault), a continuous structure from Agadir (Morocco) to Tunis (Tunisia). The fault separates a zone where the Mesozoic-Cenozoic cover is shortened and mostly detached from its basement from a zone where the cover remains horizontal and attached to its basement. Thrust-belt rocks north of the fault are structurally elevated by about 1.5 km above the Saharan Platform.

Apatite fission track data (*see* **Analytical Methods:** Fission Track Analysis) suggests that large parts of Libya and Algeria were uplifted by 1–2 km during the 'Alpine' deformational phase. As a consequence, Palaeozoic hydrocarbon source rocks were lifted out of the oil window in some parts of the Saharan Palaeozoic basins, resulting in termination of hydrocarbon generation.

Oligo-Miocene Rifting

Another major rifting phase in North Africa during the Oligo-Miocene was associated with the development of the Red Sea, Gulf of Suez, Gulf of Aqaba Rift system, which is the northern continuation of the Gulf of Aden, and East African rifts. Along the north-eastern margin of the Red Sea/Gulf of Suez axis, extension was associated with intrusion of a widespread network of dykes and other small intrusions. Rifting and separation of Arabia from Africa commenced in the southern Red Sea at about 30 Ma (Oligocene) and in the northern Red Sea and Gulf of Suez at about 20 Ma (Early Miocene). Subsequently, tectonic processes in the Arabian–Eurasian collisional zone changed the regional stress field in the northern Red Sea region, causing the rifting activity to switch from the Gulf of Suez to the Gulf of Aqaba. As a consequence the Gulf of Suez became a failed rift and was in part inverted.

Intense volcanic activity occurred in central and eastern North Africa during the Late Miocene to Late Quaternary. In places this had already commenced in the Late Eocene. Volcanic features include the plateau basalts in northern Libya, the volcanic field of Jebel Haruj in central Libya, the Tibesti volcanoes in south-east Libya and north-east Chad and the volcanism in the Hoggar (S Algeria, NE Mali, NW Niger). Some authors interpret this continental volcanism as related to a hot spot overlying a deep-seated mantle plume while others see the cause in intraplate stresses originating from the Africa–Europe collision that led to melting of rocks at the lithosphere/asthenosphere interface by adiabatic pressure release.

Depositional History

Infracambrian

The Infracambrian in North Africa is represented by carbonates, sandstones, siltstones, and shales, often infilling halfgrabens. In Morocco and Algeria, the unit includes stromatolitic carbonates as well as red and black shales, a facies similar to the Huqf Supergroup in Oman that represents an important hydrocarbon source rock there. Infracambrian siliciclastics are also known from several boreholes in the central Algerian Ahnet Basin and southern Cyrenaica (NE Libya). Infracambrian conglomeratic and shaly sandstones and siltstones occur at outcrop underneath Cambrian strata along the eastern margin of the Murzuq Basin and in some boreholes in the basin centre. In the Kufra Basin, the presence of some 1500 m of Infracambrian sedimentary rocks (of unknown lithology) is inferred for the southern basin centre, while

strata of similar age, including dolomites, have been reported from the eastern and western margins of this basin. Notably, salt deposits like those in Oman have not yet been confirmed from North Africa, although some features from seismic studies in the Kufra Basin may represent salt diapirs.

Cambro-Ordovician

The Cambro-Ordovician in North Africa is mostly represented by continental and shallow marine siliciclastics, dominated by sandstones with minor siltstone and shale intervals (Figure 6). Deposition occurred on the wide North African shelf in a generally low accommodation setting. The sediment source was the large Gondwanan hinterland to the south, with SE-NW directed palaeocurrents prevailing. The five reservoir horizons of the giant Hassi Messaoud oilfield are located in Upper Cambrian to Arenig quartzitic sandstones, including the Lower Ordovician Hamra Quartzite.

A major, shortlived ($\frac{1}{2}$–1 my) glaciation occurred in western Gondwana during the latest Ordovician, with the centre of the ice sheet located in central Africa. Features commonly attributed to pro- and sub-glacial processes reported from North Africa, Mauritania, Mali, the Arabian Peninsula, and Turkey include glacial striations, glacial pre-lithification tectonics, diamictites, microconglomeratic shales, and systems of km-scale channels. Several of these features, however, may also occur in deltaic systems unrelated to glaciation, complicating detailed reconstructions of the latest Ordovician glaciation in the region. The uppermost Ordovician in North Africa represents an important hydrocarbon reservoir horizon in Algeria (Unit IV) and Libya (Memouniyat Formation) (Figure 7).

Figure 6 Cambro Ordovician Skolithos ('Tigillites') in Jebel Dalma (Kufra Basin, SE Libya).

Silurian

Melting of the Late Ordovician icecap caused the Early Silurian sea-level to rise by more than 100 m, leading to a major transgression that flooded the North African Shelf to as far south as the northern parts of Mali, Niger, and Chad (Figure 8). Graptolitic, hemipelagic shales represent the dominant facies, while sandstone or non-deposition prevailed in palaeohigh areas, such as most of Egypt, which formed a peninsula at that time. In Libya, the total thickness of the shales (termed 'Tanezzuft Formation', Figure 7) increases north-westwards from 50 m in the proximal Kufra Basin, through 500 m in the Murzuq Basin to 700 m in the distal Ghadames Basin, reflecting the north-westward progradation of the overstepping sandy deltaic system ('Akakus Formation', Figure 7) during the mid-Llandovery to Ludlow/Přídolí (Figure 8).

The Silurian shales are generally organically lean, except for the Lower Llandovery (Rhuddanian) and Upper Llandovery/Lower Wenlock when anoxic phases occurred. During these phases, organically rich, black shales (often referred to as 'hot shale') with total organic carbon values of up to 16% were deposited. The older of the two black shale horizons is developed only in palaeodepressions that were already flooded in the Early Llandovery, while the upper black shale unit is restricted to areas that during the Late Llandovery/Early Wenlock had not yet been reached by the prograding sandy delta (Figure 8).

Silurian organic-rich shales are estimated to be the origin of 80–90% of all Palaeozoic-sourced hydrocarbons in North Africa. The same depositional system is also developed on the Arabian Peninsula, where age-equivalent black shales exist, for example, in Saudi Arabia, Syria, Jordan, and Iraq.

Characteristic limestone beds rich in '*Orthoceras*' are interbedded with the Ludlow-Přídolí shales in Morocco and western Algeria, the most distal parts of the North African shelf (Figure 8). In more

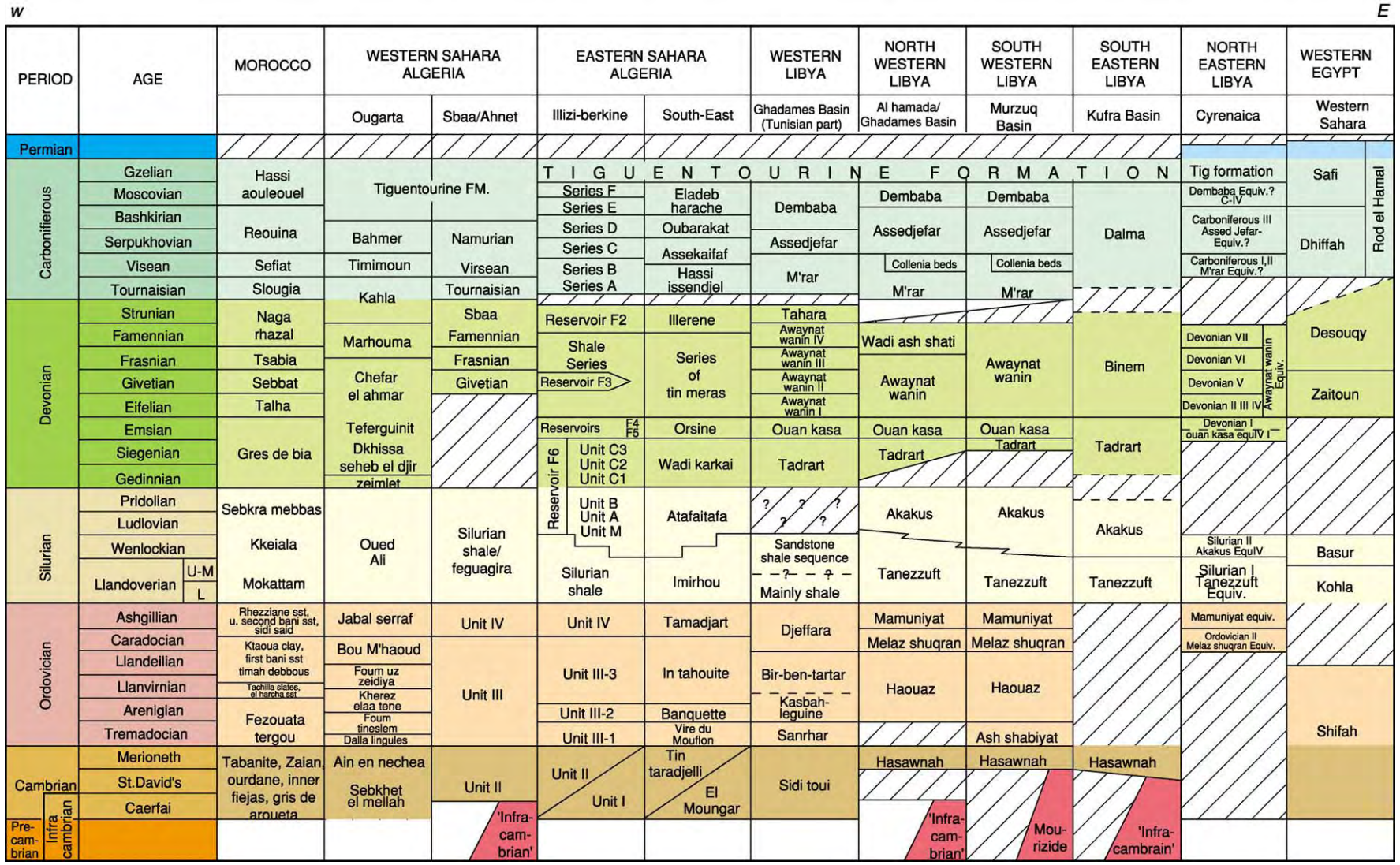


Figure 7 Correlation chart of Palaeozoic formations in North Africa.

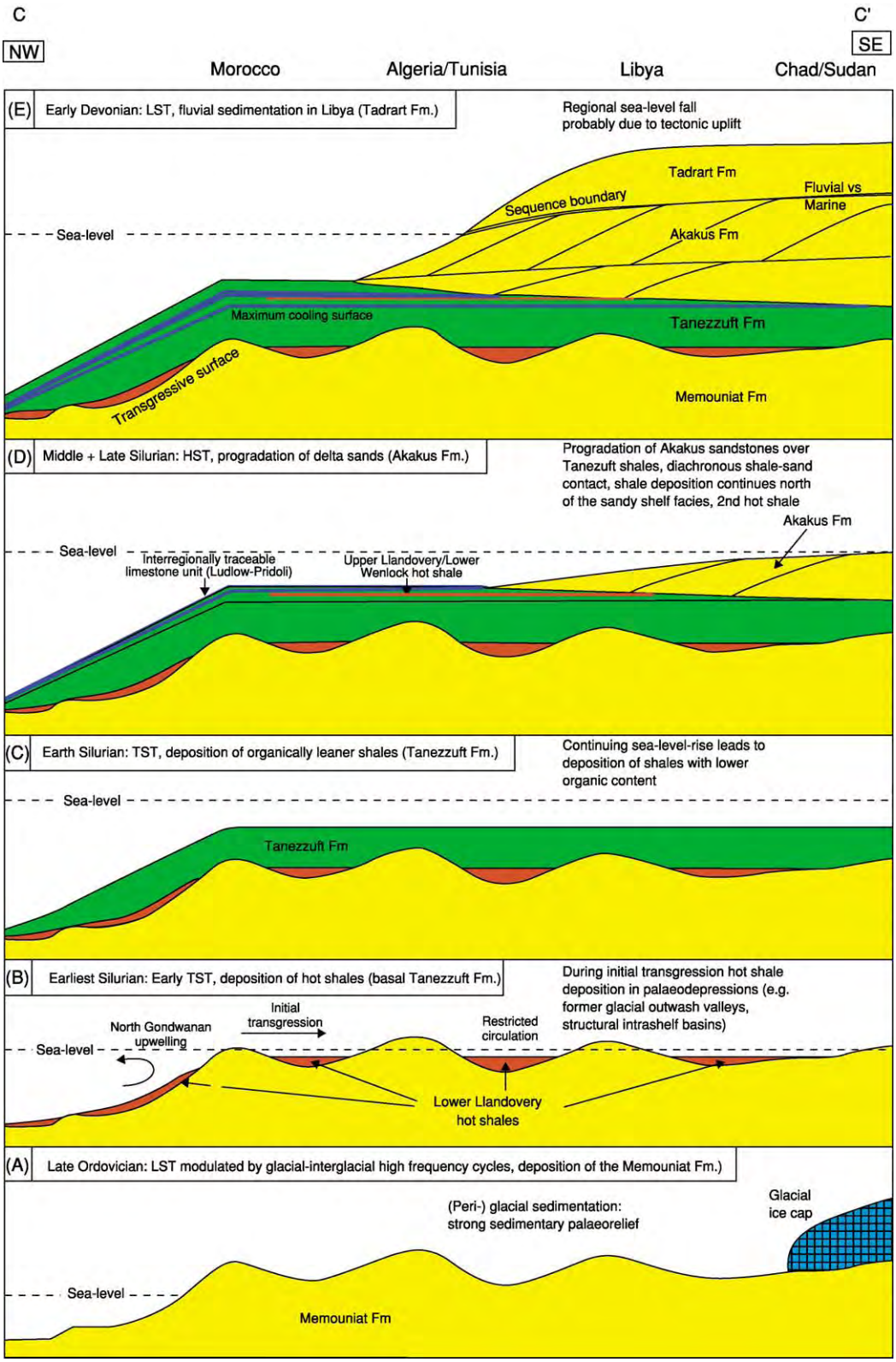


Figure 8 Depositional model for Late Ordovician to Early Devonian sediments in North Africa. (Modified after Luning *et al.* (2000).)

proximal shelfal locations, sand influx was already too great for limestones to develop. The '*Orthoceras* Limestone' in some areas is organic-rich. Similar age-equivalent limestones also occur in some peri-Gondwana terranes, such as in Saxo-Thuringia where the unit is termed 'Ockerkalk'.

Devonian

A major eustatic sea-level fall occurred during the latest Silurian/Early Devonian, resulting in a change to a shallow marine/continental facies in eastern and central North Africa. Coastal sand bar, tidal, and fluvial deposits form important hydrocarbon reservoir horizons, for example, in the Algerian Illizi Basin (unit F6, Figure 7) and the Ghadames (=Berkine) Basin ('Tadrart Formation') in north-west Libya (Figure 8). On the distal side of the North African shelf towards Morocco fully marine conditions still prevailed. The Lower Devonian of Morocco is well-known for its rich trilobite horizons. A sea-level rise during the later part of the Early Devonian led to deposition of shelfal shales and sandstones in central North Africa. In Algeria significant hydrocarbon reservoirs exist in sandstones of the Emsian (units F4, F5). In western Algeria the base of the Emsian lies under a limestone bed termed 'Muraille de Chine' ('Chinese Wall'), because at exposure it commonly forms a characteristic, long ridge.

Due to their distal position on the North African shelf and a minimum of siliciclastic dilution Morocco and western Algeria were dominated by carbonate sedimentation during the mid-Devonian. The facies here includes prominent mud mounds, for example, in the southern Moroccan area of Erfoud and in the central Algerian Azel Matti area. Further to the east, the facies becomes more siliciclastic. Eifelian-Givetian tidal bar sandstones form the main reservoir (unit F3) in the Alrar/Al Wafa gas-condensate fields in the eastern Illizi Basin.

The beginning of the Late Devonian was characterized by a major eustatic sea-level rise which resulted in deposition of hemipelagic shales, marls, and limestones over wide areas of North Africa. The Moroccan Middle to Upper Devonian typically contains rich cephalopods faunas (goniatites, clymeniids).

The 'Frasnian Event', an important goniatite extinction event and a phase of anoxia, occurred during the Early Frasnian and led to deposition of organic-rich shales and limestones in various places across North Africa. In the Algerian, Tunisian, and Libyan Berkine (=Ghadames) Basin, Frasnian black shales contain up to 16% organic carbon and represent an important hydrocarbon source rock (Figure 9). The organic-rich unit also occurs in South Morocco and north-west Eygpt. In parts of north-west Africa, a second organically enriched horizon exists around the Frasnian–Famennian boundary, associated with the worldwide Kellwasser biotic crisis. The deposits in southern Morocco include black limestones.

A major fall in sea-level occurred during the latest Devonian, triggering progradation of a Strunian (latest Devonian–earliest Carboniferous) delta in central North Africa. These clastics form an important hydrocarbon reservoir unit (F2) in Algeria.

Carboniferous

Sea-level rise during the Early Carboniferous resulted in the development of a widespread shallow marine to deltaic facies across large parts of North Africa. A carbonate platform was established in the Bechar Basin in western Algeria at this time. Early Carboniferous dolomites of the Um Bogma Formation in south-west Sinai host important Mn-Fe ores. Non-deposition and continental sandstone sedimentation occurred in southern and elevated areas, for example, in most of Egypt.

In the Late Carboniferous, deposition of marine siliciclastics was restricted to north-west Africa and the northernmost parts of north-east Africa, for example, Cyrenaica and the Gulf of Suez area. Paralic coal in the Westphalian of the Jerada Basin (NE Morocco) forms the only sizable Late Carboniferous coal deposit in North Africa. In the course of the latest Carboniferous Hercynian folding and thrusting, most of north-west Africa was uplifted, resulting in a change to a fully continental environment. Only Tunisa, north-west Libya and the Sinai Peninsula were still under marine influence at this time.

Permo–Triassic

Marine Permo-Triassic sedimentary rocks are restricted to the northernmost margin of central and eastern North Africa. For example, Permian marine carbonates and siliciclastics crop out in southern Tunisia representing the only exposed Palaeozoic unit in this country. Most of North Africa, however, remained subaerially exposed during the Permian to mid-Triassic. Continental red clastics (sandstones, shales, conglomerates) represent the most important lithologies. The Permian of Morocco is restricted to a series of intramontane basins located around the margin of the central Moroccan Hercynian massif. The main facies associations in the Triassic TAGI (Trias Argilo-Gréseux Inférieur) in the eastern Algerian Berkine (=Ghadames) Basin are fluvial channel sandstones, floodplain silts and palaeosols, crevasse splay deposits, lacustrine sediments, and shallow marine transgressive deposits. Fluvial sandstones of the TAGI are the main oil and gas reservoirs in the

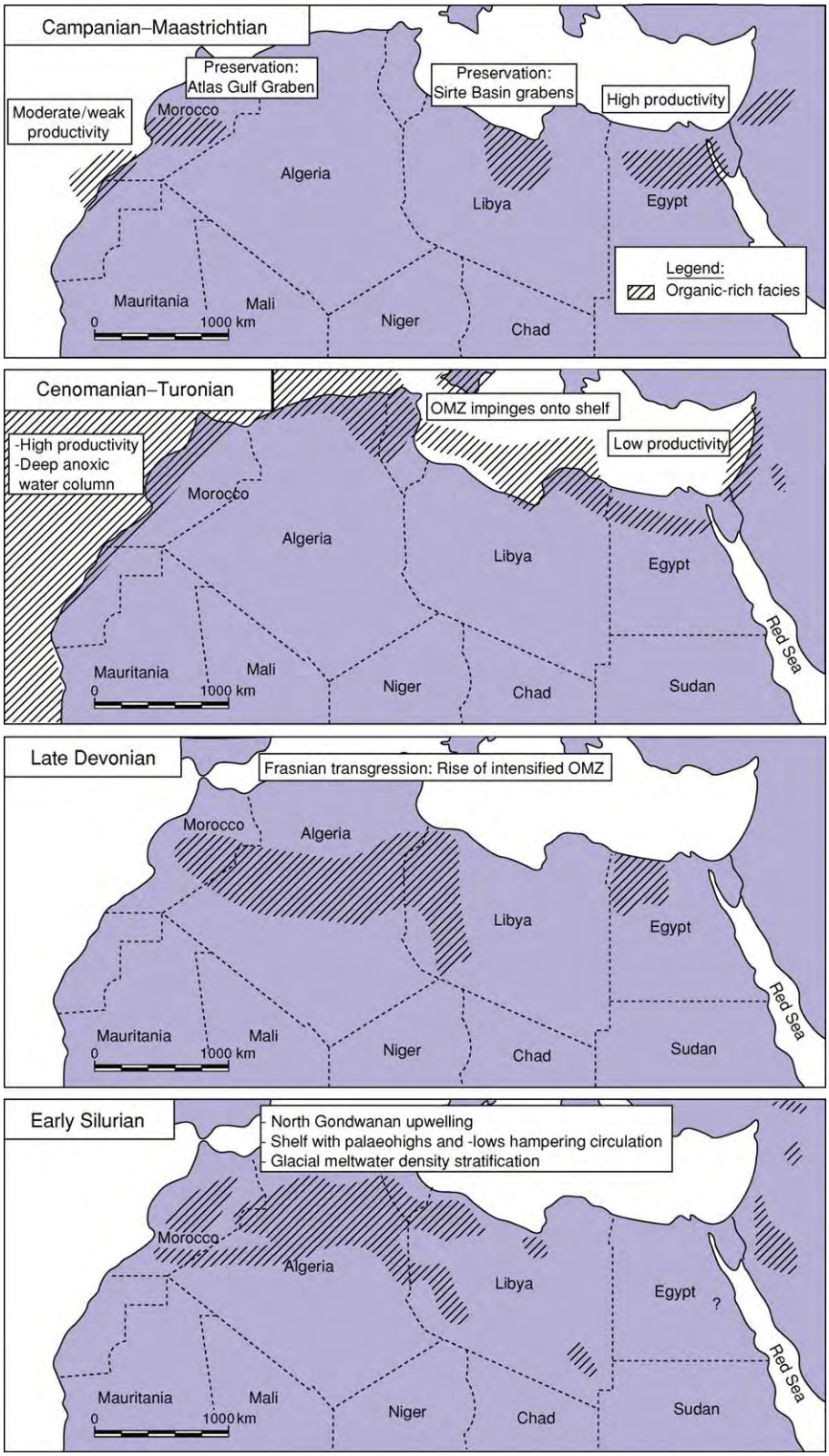


Figure 9 Known distribution of organic rich strata of Early Silurian, Late Devonian, Cenomanian Turonian, and Campanian Maastrichtian age in North Africa.

Algerian Berkine and Oued Mya basins, including the super-giant gas field in Hassi'R Mel. Similar Triassic sandstones also serve as a relatively minor hydrocarbon reservoir in the Sirt Basin, sourced from Cretaceous source rocks.

During the Late Triassic/Early Jurassic, evaporites were deposited in rift grabens associated with the opening of the Atlantic, and of the Atlas Gulf and with the separation of the Turkish-Apulian terrane from North Africa. Characteristic 'salt provinces' are located offshore along the Moroccan Atlantic coast, northern Algeria/Tunisia and offshore east-Tunisia/north-west Libya. In most areas the diapiric rise commenced in the Jurassic–Cretaceous.

The Late Triassic/Early Jurassic evaporites and shales in the north-east part of the Algerian Saharan Platform are up to 2 km thick and form a hydrocarbon caprock for the Triassic reservoir. In some cases, because of the Hercynian unconformity, they also form the caprock for Palaeozoic reservoirs such as at the super-giant Hassi Messaoud field in Algeria.

Jurassic

Marine sedimentation during the Jurassic was restricted to the northern and western rims of North Africa, including, for example, northernmost Egypt, the Atlas region, and the Tarfaya Platform in southern Morocco. Carbonate platforms and intraplatform basins were widespread, including development of reefal limestones and oolites. In the Gebel Maghara area in northern Sinai, paralic coal was deposited during the Middle Jurassic. Locally the Lower and Upper Jurassic of North Africa contain organically enriched horizons, corresponding in age to the prominent Jurassic black shales of central Europe (e.g., Posidonia Shale in Germany and Kimmeridge Clay in England). Such Jurassic bituminous pelites occur, for example, in the Atlantic Basin, Atlas Rift of Morocco, and the Egyptian Abu Gharadig Basin. South of the North African Jurassic marine facies belt, continental redbeds were deposited (Figure 10). In the Egyptian Western Desert the Jurassic–Cretaceous contains several prolific hydrocarbon reservoir horizons.

Cretaceous

Due to low eustatic sea level the Lower Cretaceous of North Africa is dominated by terrestrial clastics, termed the 'Nubian Sandstone' in Egypt and Libya ('Sarir Sandstone' in the Sirt Basin) (Figure 10). Once again, marine conditions existed only in a marine coastal belt in the north. During the Aptian to Maastrichtian, a series of transgressions gradually flooded the areas to the south. On the Sinai Peninsula, the transition phase is characterised by deltaic influenced,

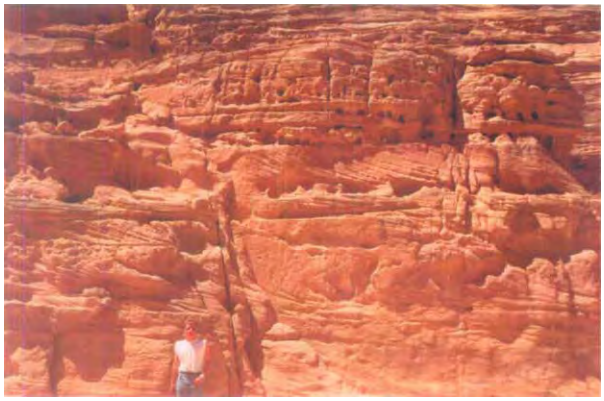


Figure 10 Cross bedded fluvial 'Nubian Sandstone', Jurassic Cretaceous, 'Coloured Canyon', central East Sinai (Egypt).

mixed siliciclastic-carbonate systems that during the Albian evolved into carbonate-dominated environments. During the latest Cenomanian, large parts of North Africa became submerged following a prominent eustatic sea-level rise that is thought to be one of the most intense Phanerozoic flooding event. As a consequence, the 'Transsaharan Seaway' was created, connecting the Tethys in central North Africa with the Atlantic in West Africa. Similar seaways and gulfs existed in north-west Africa into the Eocene. A seaway located within the Atlas rift system, the 'Atlas Gulf', was restricted temporally to the Cenomanian–Turonian.

The strong latest Cenomanian sea-level rise in combination with high productivity conditions in the southern North Atlantic are thought to form the basis for the Late Cenomanian–Early Turonian Oceanic Anoxic Event (OAE2) during which organic-rich strata were deposited in rift shelf basins and slopes across North Africa and in deep sea basins of the adjacent oceans. Characteristic sediments associated with this anoxia include oil shales in the Tarfaya Basin (southern Morocco), organic-rich limestones in north-west Algeria and northern Tunisia (Bahloul Formation), and black shales in offshore Cyrenaica, and the Egyptian Abu Gharadig Basin (Abu Roash Formation) (Figure 9). The unit represents a potential oil-prone hydrocarbon source rock in the region. A general decrease in peak organic richness and black shale thickness occurs in North Africa from west to east, which possibly is a result of upwelling along the Moroccan Atlantic coast and the absence of upwelling in the Eastern Mediterranean area.

The organic-rich Cenomanian-Turonian deposits also play an important role in the genesis of Zn/Pb ore deposits in northern Tunisia and eastern Algeria. The origin of these Zn/Pb ores is related to hypersaline basinal brines, made of ground water and dissolved Triassic evaporites, that leached metals

from the Triassic-Cretaceous sediments. Ore deposition occurred when these metal-bearing solutions mixed with microbially reduced sulphate solutions that were associated with the organic carbon of the Cenomanian-Turonian strata.

Due to the generally high sea-level, the marine Upper Cretaceous in North Africa is dominated by calcareous lithologies, namely dolomites/limestones, chalks, and marls (Figure 11). Lateral and vertical facies distributions are strongly related to sea-level changes of various orders as well as to the changing structural relief associated with Late Cretaceous syn-depositional compression. Great variations in thickness and facies as well as onlap features, for example, are developed around the domal anticlines of the Syrian Arc Foldbelt in Sinai and within rift grabens of the Sirt Basin (N. Libya).

The Campanian–Maastrichtian was characterised by very high sea-level, resulting in a widespread distribution of hemipelagic deposits, such as chalks and marls. These deposits often contain abundant foraminiferal faunas and calcareous nannofossil floras, which allow high-resolution biostratigraphic and palaeoecological studies in these horizons. As on the Arabian Peninsula, the Santonian–Maastrichtian interval in North Africa contains significant amounts of phosphorites, which are mined in, for example, Morocco/Western Sahara and Abu Tartour (Western Desert), making North Africa one of the world's largest producers of phosphate (*see* **Sedimentary Rocks:** Phosphates).

In places, the Campanian–Maastrichtian contains organic-rich intervals with total organic carbon contents of up to 16%, for example, in the Moroccan Tarfaya Basin and Atlas Gulf area, the Libyan Sirt Basin and the Egyptian southern Western Desert, Red Sea Coast and Gulf of Suez (Figure 9). Notably, Algeria, Tunisia, and West Libya are dominated by organically lean deposition during this time. Campanian–Maastrichtian black shales form important hydrocarbon source rocks in the Sirt Basin and the Gulf of Suez.

Figure 11 Contact between chalky limestones of the Early Eocene Bou Dabbous Formation (reddish) and the underlying Campanian Maastrichtian Abiod Formation (bluish) (Ain Rahma Quarry, Gulf of Hammamet area, Tunisia).

Palaeogene

Sea-level during most of the Paleocene–Eocene remained high resulting in deposition over wide areas (Egypt: Dakhla and Esna Shale) of hemipelagic marls and chalks that are rich in planktonic foraminifera. A sea-level fall occurred during the mid-Paleocene, resulting in the formation of a short-lived carbonate interbed ('Tarawan Chalk') in parts of Egypt. Within the Eocene, the facies typically changes here to hard dolomitic limestones with abundant chert nodules ('Thebes Limestone'). A similar Palaeogene facies development can also be found in parts of northern Libya and Tunisia.

The Eocene in Egypt, Libya, Tunisia, and Algeria includes nummulitic limestones up to several 100 metres thick, which were deposited in carbonate ramp settings. The unit forms major hydrocarbon reservoirs in offshore Libya and Tunisia. Well-exposed and continuous exposures occur in Jabal al Akhdar (Cyrenaica), where the nummulite body's geometry can best be studied (Figure 12). Notably, the Giza pyramids in Cairo are built from Eocene nummulite limestone.

The Eocene hydrocarbon play in the offshore of Tunisia is sourced by dark-brown marl and mudstone of the lower Eocene Bou Dabbous Formation. The unit contains type I and II kerogen and ranges in thickness from 50 to 300 m.

Neogene and Quaternary

Marine conditions during the Miocene were again restricted to the northernmost margin of North Africa

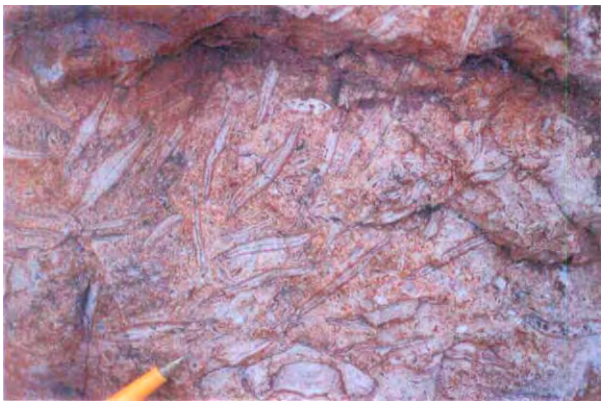


Figure 12 High energy nummulitic bank facies, Darnah Formation, Middle to Late Eocene, West Darnah Roadcut, Jebel Akhdar (Cyrenaica, Libya).

including the Atlas, Sirte Basin, Cyrenaica, and Red Sea. Carbonate platforms and ramps were developed in northern Morocco. The Miocene Gulf of Suez in Egypt is rich in hydrocarbons, containing more than 80 oilfields. Oils in the Gulf of Suez were mostly sourced from source rocks in the pre-rift succession, including the Campanian–Maastrichtian Brown Limestone. Hydrocarbon reservoir horizons include various Miocene syn-rift sandstones and carbonates as well as pre-rift reservoirs, including fractured Precambrian granites, Palaeozoic–Cretaceous sandstones, and fractured Eocene Thebes Limestone. The thickness distribution and facies of the syn-rift strata are strongly controlled by fault block tectonics. Shales and dense limestones of the pre-rift and the syn-rift units are the primary seals, while overlying Miocene evaporites form the ultimate hydrocarbon seals.

During the latest Miocene, more than 2 km thick evaporites were deposited in a deep and desiccated Mediterranean basin that had been repeatedly isolated from the Atlantic Ocean. In the near-offshore only a few tens to hundreds of metres of evaporites exist, whilst they are almost absent from the onshore area. As a consequence of the 'Messinian Salinity Crisis', a large fall in Mediterranean sea-level occurred, followed by erosion and deposition of non-marine sediments in a large 'Lago Mare' ('lake Sea') basin. Cyclic evaporite deposition is thought to be almost entirely related to circum-Mediterranean climate changes.

The Nile Delta system represents a major natural gas province. It was initiated during the Late Miocene with deep canyon incision into pre-existing Cenozoic/Mesozoic substrate, allowing transportation of huge amounts of sediments into the Mediterranean. The proximal infill of these canyons is thick, coarse alluvium becoming sandier with greater marine influence northwards. The far reaches of these canyon systems have proven to be a good Plio-Pleistocene hydrocarbon reservoir linked mainly to the lowstands, when sands were conveyed to the outer belts through incised canyons in the upper slopes which led to submarine fans farther northwards.

The Early Holocene (~9–7 kyr BP) was a relatively humid period in North Africa. During this phase, the African Humid Period, grasslands covered the Sahara/Sahel region, and many lakes and wetlands existed here. The humid conditions at this time were associated with a strengthening of the summer monsoon circulation due to an increase in the land–sea thermal contrast under the influence of relatively high summer insolation.

See Also

Africa: Pan-African Orogeny. **Analytical Methods:** Fission Track Analysis. **Petroleum Geology:** Overview. **Plate Tectonics**. **Sedimentary Rocks:** Phosphates. **Tectonics:** Rift Valleys.

Further Reading

Ben Ferjani A, Burollet PF, and Mejri F (1990) *Petroleum Geology of Tunisia*. Tunis: Entreprise Tunisienne d'Activités Pétrolières.

Beuf S, Biju Duval B, de Charpal O, Rognon P, Gariel O, and Bennacef F (1971) Les grès du Paléozoïque inférieur au Sahara, Sédimentation et discontinuités, évolution d'un craton. *Publications de l'Institut français du Pétrole* 18: 464.

Coward MP and Ries AC (2003) Tectonic development of North African basins. In: Arthur TJ, MacGregor DS, and Cameron NR (eds.) *Petroleum Geology of Africa: New Themes and Developing Technologies*. Geological Society London, Special Publication 207: 61 83.

Dercourt JM, Gaetani B Vrielynck E, *et al.* (eds.) (2000) *Atlas Peri Tethys, Palaeogeographical maps*. CCGM/CGMW, Paris.

Doblas M, Oyarzun R, Lopez Ruiz J, Cebria JM, Youbi N, Mahecha V, Lago M, Pocovi A, and Cabanis B (1998) Permo Carboniferous volcanism in Europe and northwest Africa: a superplume exhaust valve in the centre of Pangaea? *J. Afr. Earth Sciences* 26: 89 99.

Hallett D (2002) *Petroleum Geology of Libya*. Amsterdam: Elsevier.

Lüning S, Craig J, Loydell DK, Štorch P, and Fitches B (2000) Lower Silurian 'Hot Shales' in North Africa and Arabia: Regional Distribution and Depositional Model. *Earth Science Reviews* 49: 121 200.

Macgregor DS, Moody RTJ, and Clark Lowes DD (eds.) (1998) Petroleum Geology of North Africa. *Geological Society London Special Publication* 132: 7 68.

Maurin J C and Guiraud R (1993) Basement control in the development of the Early Cretaceous West and Central African Rift System. *Tectonophysics* 228: 81 95.

Piqué A (2002) *Geology of Northwest Africa*. Stuttgart: Gebr. Borntraeger.

Said R (1990) *The Geology of Egypt*. Rotterdam, Netherlands: Balkema Publishers.

Schandelmeier H and Reynolds PO (eds.) (1997) *Palaeogeographic Palaeotectonic Atlas of North eastern Africa and adjacent areas*. Rotterdam: Balkema.

Selley RC (1997) Sedimentary basins of the World: Africa. Amsterdam: Elsevier.

Stampfli GM, Borel G, Cavazza W, Mosar J, and Ziegler PA (2001) *The Paleotectonic Atlas of the Peritethyan Domain*. Strasburg European Geophysical Society.

Stets J and Wurster P (1981) Zur Strukturgeschichte des Hohen Atlas in Marokko. *Geologische Rundschau* vol. 70(3): 801 841.

Tawadros EE (2001) *Geology of Egypt and Libya*. Rotterdam: Balkema.

Rift Valley

L Frostick, University of Hull, Hull, UK

Introduction

The East African and Dead Sea rifts are famous examples of rifts that have played prominent parts in human evolution and history. They are both areas where the Earth's crust has been put under tension and ripped apart to give deep valleys that snake across the landscape. They are linked tectonically, via the Red Sea–Gulf of Aden, which is an incipient ocean separating the African and Arabian plates. The differences between the two rifts are caused by differences in the relative movement of the crust. In the East African rift the tension that formed the rift is close to 90° to the rift axis, whereas the movement of Jordan relative to Israel is northwards, almost parallel to the Dead Sea, which is a small section pulled apart as a result of splaying and bending of the faulted plate boundary.

The ancient crust of Africa has been subjected to rifting many times in its very long geological history. Recognizable rift basins can be identified in many locations around the continent, and they range in age from Palaeozoic to Quaternary, a time-span of over 500 Ma. In some areas there is evidence of repeated activity, and it appears that there have been at least seven phases of rifting over the past 300 Ma. The older rifts, for example the Benue trough in West Africa, have been inactive for many millions of years, but the most spectacular rift features are to be found in East Africa, where recent rifting has left a scar on the landscape that is visible from space (Figure 1). North of the zone where Africa touches Europe at the eastern end of the Mediterranean there is another famous rift, which is linked tectonically to East Africa. The Dead Sea Rift straddles the border between Israel and Jordan and is the lowest point on the surface of the Earth, reaching more than 800 m below sea-level.

Plate Tectonic Setting

Rifting occurs when the crust of the Earth is placed under tension, pulling it apart and causing faulting. The general term for the basins so produced is 'extensional' but they can occur in situations where the regional sense of movement is compressional or is tearing the crust, e.g. the Baikal and Dead Sea rifts, respectively. However, the main African rift basins were formed in a plate-tectonic setting that is dominated by extension, particularly during the Tertiary–Quaternary period. During this time the Great or East African Rift was formed as part of a larger plate-tectonic feature that stretches from south of Lake Malawi in Africa to the flanks of the Zagros mountains and the Persian Gulf in the north (Figure 2). It changes its nature along its length, resulting in a range of geological basins and geomorphological features. In Africa it is a volcanically active continental rift hundreds of kilometres wide that contains a range of river and lake sediments. As it quits Africa it passes into an incipient ocean with a newly formed seafloor spreading centre along the length of the Red Sea and the Gulf of Aden. The deposits in these basins include thick sequences of salt, which form effective traps for hydrocarbons generated from associated organic-rich shales. North of the Red Sea the type of plate margin alters as the boundary passes through the Gulf of Aqaba/Elat and

Figure 1 Satellite remote sensing image of the Horn of Africa and Arabia, showing the East African Rift system, the incipient ocean of the Red Sea Gulf of Aden, and the conservative plate boundary that runs through the Dead Sea. Images collected by the TERRA satellite using the MODIS instrument (moderate resolution imaging spectroradiometer) and enhanced with SRTM30 (Shuttle Radar Topography Mission 1km resolution) shaded relief.

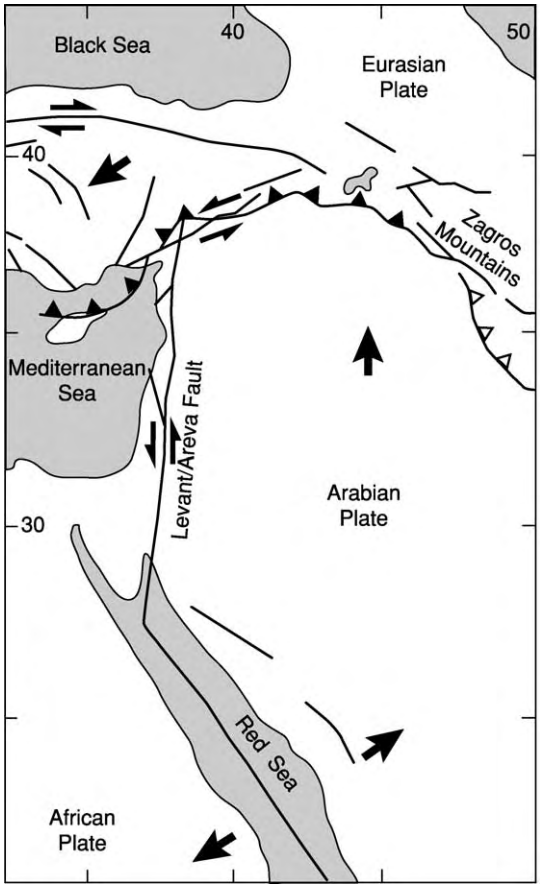


Figure 2 Diagrammatic representation of the plate tectonic setting of the area between the northern end of the East African Rift (Afar triangle) and the Zagros Mountains.

Figure 3 Satellite remote sensing image of the Sinai Arabian plate boundary, showing the Dead Sea and Sea of Galilee (Lake Kinneret). Image collected by the TERRA satellite using the MODIS instrument (moderate resolution imaging spectroradiometer).

into the Levant/Areva valley between Israel and Jordan. Here, the Arabian plate is moving past the European plate without significant extension or compression. However, localized tension associated with fault bends and splays has resulted in the formation of two very well-known biblical lake basins, the Dead Sea and Lake Kinneret (otherwise known as the Sea of Galilee), both of which can be seen in Figure 3. These are also termed 'rifts' although the setting and geological history are different from those of their larger East African contemporary. The system terminates in the Zagros mountains, where the crust created in the new Red Sea–Gulf of Aden ocean is compensated for by the crustal shortening inherent in mountain-building processes.

The East African Rift

Topography and Structure

Within Africa certain features of the topography and structure are common to all the basins.

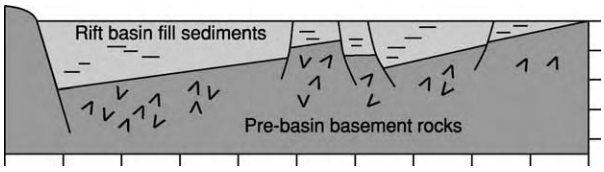


Figure 4 Stylized half graben structure typical of the basins in the East African Rift.

Topographically they comprise a central valley, often referred to as a 'graben', flanked by uplifted shoulders that are stepped down towards the rift axis by more or less parallel faults. Often, one flank is more faulted than the other, so that the rift valley is in fact asymmetrical and should be referred to as a 'half graben' (Figure 4). The width of the structure varies from 30 km to over 200 km, with the widest section at the northern extremity where the rift links to the Red Sea in the Afar region of Ethiopia. The main faulted margin alternates from one side of the rift to the other along its length, producing a series of

relatively separated basins, many of which contain lakes of varying depth and character (e.g. Lakes Tanganyika, Naivasha, and Malawi). These are separated into hydrologically distinct basins by topographical barriers crossing the rift axis where the border faults switch polarity. This surface separation reflects an underlying structure, the nature of which varies from basin to basin but often includes faulting with a tearing or scissor type of movement and flexing. Geologists are not agreed on the processes going on in these areas and have given these zones different names according to their assumptions about the mechanism of formation. These include transfer, relay, and accommodation zones, as well as ramps or just segment boundaries. The distance between adjacent boundaries varies from tens to hundreds of kilometres (Figure 5).

At each end of the individual border faults the displacement of the rift floor relative to rocks outside the valley reduces to zero. Displacement is greatest at the centre of the fault, and this leads to a subtle rise and fall of the rift floor along its length even without the intervention of major new cross-rift structures and processes.

The rift in Kenya is characterized by numerous caldera volcanoes and at least 3 to 4 phases of faulting, the most recent forming a narrow linear axial zone. the faulting ranges in age from Miocene to Recent.

In the southern half of the rift's 35 000 km length it divides into two distinct branches around Lake Victoria. The eastern branch contains only small, largely saline, lakes, while the western branch contains some of the largest and deepest lakes in the region, including Lake Tanganyika.

Doming and Volcanicity

The East African Rift contains two large domes centred on Robit in Ethiopia and Nakuru in Kenya. These domes are over 1000 km in diameter and extend far beyond the structural margins of the rift valley. Geophysical studies of these domes have shown that they are underlain by zones of hot low-density mantle rocks and that the surface crust is thinned significantly relative to adjacent areas. The domes are centres of volcanic activity that began more than 25 Ma ago and continues to the present day (Figure 6). Volcanic features are widespread in Ethiopia and extend southwards into Kenya along the eastern branch of the rift. It is estimated that there are more than 500 000 km^3 of volcanic rocks in this area, over a third of which occur in Kenya. In the branch to the west of Lake Victoria volcanism is spatially more

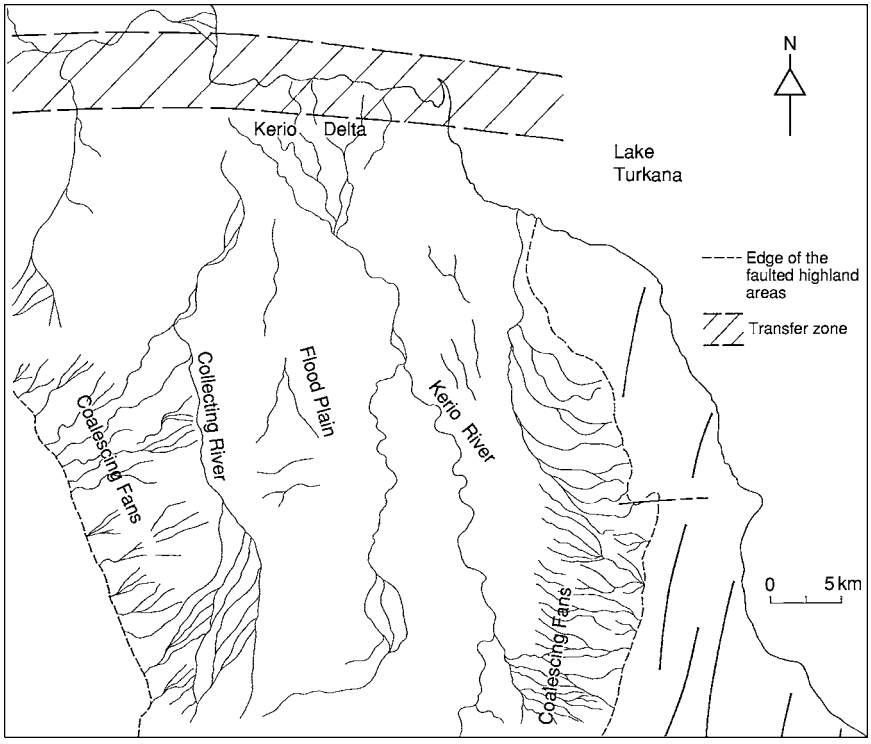


Figure 5 Diagrammatic representation of the river drainage close to the west shore of Lake Turkana, northern Kenya, showing the Kerio River flowing into the Lake at a transfer zone and the alluvial fans issuing from the fault scarps.

Figure 6 Geyser activity in the volcanically active area around Lake Bogoria, Kenya.

limited, occurring only to the north and south of Lake Tanganyika. This contributes to the different characters of the lakes in the two branches as not only can the volcanic rocks fill the basins, leaving less space for large lakes, but also many of the rock types are rich in salts, which contribute to the salinity of the lakes once they are released by weathering.

Large and active volcanoes that sit outside the rift structure are a striking feature of the landscape. Mounts Kilimanjaro and Kenya, for example, are favourite targets for climbers, and both sit on the flanks of the rift (Figure 7).

Hydrology and Climate

The East African Rift system sits astride the equator, extending from 12° N to 15° S, and this dictates the overall character of the climate. Superimposed on this are the effects of the rift topography, with its uplifted domes, faulted flanks, and depressed central valleys. Rainfall is lowest in the northern parts of Ethiopia and increases southwards into northern Kenya. The region is generally desert or semi-desert with vegetation limited to sparse grasses and scrub. South of where the rift branches the rainfall is higher, with the western branch being wetter than the eastern one. The uplifted mountains that make up the margins of the rift are wetter and cooler than the valley bottom; for example, an annual figure of over 2000 mm of rainfall has been recorded in the Ruwenzori Mountains near Lake Mobutu.

The doming that accompanied the rifting in East Africa has had a major impact on the present river systems. The development of the rift disrupted a pre-existing continental drainage system in which a few large rivers with vast integrated drainage basins dominated the landscape. As the area was domed and faulted and the new valley formed, the rivers adjusted to the new landscape: some lost their headwaters, others were created, some gained new areas to drain. The overall effect was to divert much of the drainage north into the Nile system and west into the Congo drainage, with only a few small rivers now reaching the Indian Ocean. Inside the valley, the rivers are generally short and small, ending in a lake not far from the river source, but a few rivers run along the rift, often caught between faulted hills, and discharge into lakes far from their original sources, e.g. the Kerio River in Kenya has its source near Lake Baringo but discharges into Lake Turkana more than 200 km to the north (Figure 8).

The segregation of the underlying structure into topographically distinct sections exerts an overriding control on the character and distribution of lakes throughout the rift. It provides the framework within which the balance between movement of water into the basin, from rainfall and rivers, and evaporation from the surface will work. The largest and deepest lake, Lake Tanganyika, is in the wetter western branch of the rift in a particularly deep section. It covers an area of over 40 000 km^2 and is more than 1400 m deep at its deepest point. Lakes in the eastern branch are smaller and shallower; for example Lake Bogoria is an average of less than 10 m deep, and if the climate changes and rainfall decreases they soon become ephemeral, drying out completely during periods of drought.

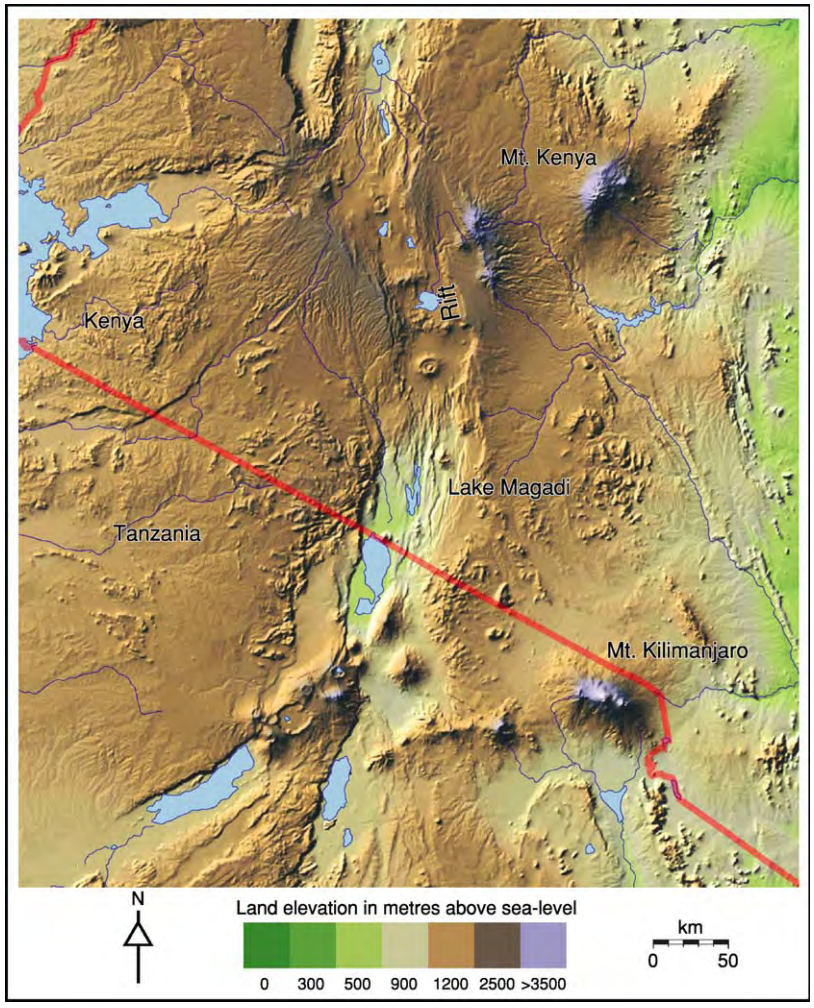


Figure 7 Satellite image of Mount Kilimanjaro and Mount Kenya, showing how they sit outside the main East African Rift structure. This is a shaded relief map produced from SRTM30 data with colour added to indicate land elevations.

Sedimentation and Basin Fills

As water flows into the rift basins it brings with it material dislodged and dissolved from the surrounding rocks, which is then deposited within the basin. How, where, and what is deposited depends on the shape of the basin and how surface processes work on and disperse the material. The overall shape of the basin fill is controlled by the pattern of faults and subsidence: deposits are thicker close to areas of the faults with greatest displacement (Figure 4). The geometry of the fill is therefore almost always asymmetric, thickening towards the main border fault and thinning in all other directions, giving a characteristic wedge shape.

There are no marine sediments in the rift: all the deposits are terrestrial and comprise river, delta, lake-coast, and lake sediments. Wind-blown sands and dunes are rare and of only local importance. The rivers vary in character from ephemeral, flowing only in response to seasonal rain storms, to perennial. The rivers carry and deposit sands and gravels in their beds, sweeping finer silts and clays into overbank lagoons and lake-shore deltas. The character of the lake deposits themselves depends on a variety of factors including the timing and character of river supplies, salinity, evaporation, water stratification, and animal and plant growth. In deep lakes such as Lake Tanganyika there are layered muds, which can be hundreds of metres thick and contain enough algal remains to generate oil. Shallower lakes can contain high numbers of diatoms, which leave deposits of a silica-rich rock called diatomite. Some lakes in volcanic areas of the rift have sufficiently high salt concentrations for precipitation and the development of exploitable salt deposits. One example is the trona, a complex carbonate of sodium, which is extracted seasonally from Lakes Magadi and Natron (Figure 9).

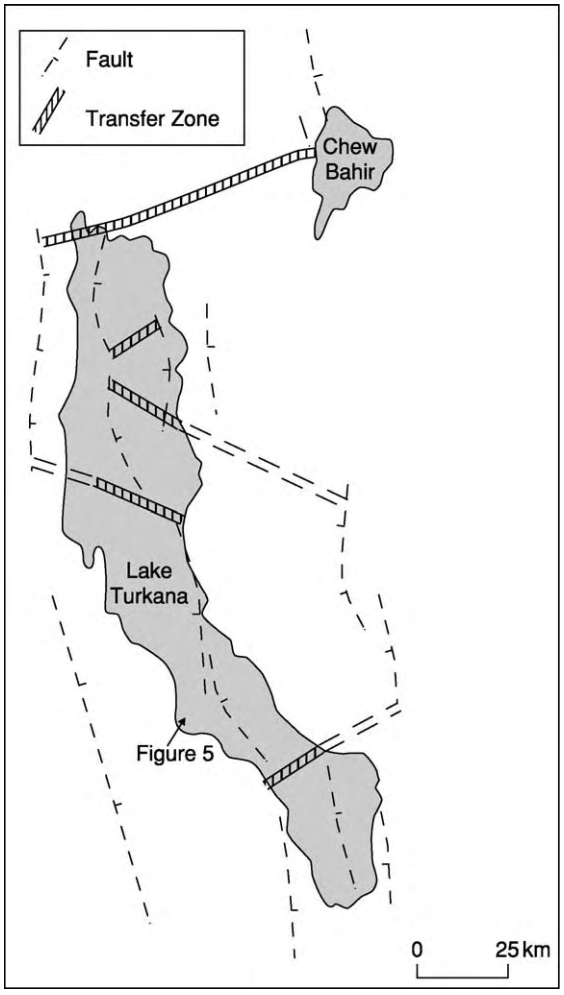


Figure 8 Stylized diagram of the Lake Turkana area at 3° N in the East African Rift, showing the main faults and transfer zones crossing the rift axis.

Hominid Finds and Evolution

The rift forms a striking geomorphological feature cutting across the African craton, segmenting the landscape, and controlling the local geology. Along most of its length it achieves a depth of in excess of 1 km and at its deepest, in Ethiopia, it is over 3 km deep. Its striking topography generates its own set of microclimatic and hydrological conditions, which have had a major impact on plant and animal distributions and evolution. It acts as a north–south corridor for the migration of animals and birds, but equally inhibits east–west movements. During periods of climatic stress at higher latitudes, when glaciers dominated much of the European and Asian continents, the lake basins of the rift were havens for animals, including early humans. Finds of early humans (hominids) in the rift are more numerous and more complete than in almost any other part of the world, and it has been postulated that all present-day humans are derived from ancestors that migrated out of the East African Rift (*see* **Fossil Vertebrates:** Hominids).

Dead Sea Rift

Topography and Structure

The Dead Sea Rift is superficially very similar to some of the individual lake basins in the East African Rift. It is a narrow depression in the surface of the Earth over 100 km long and only 25 km wide, reaching over 800 m deep at its lowest point (Figure 10). The Dead Sea is not, in reality, a sea at all but an enclosed salty lake, which occupies more than 80% of the surface area of the basin. It sits on the plate boundary that

Figure 9 Lake Magadi, Kenya, during the dry season, showing the surface of the lake completely encrusted with salt.

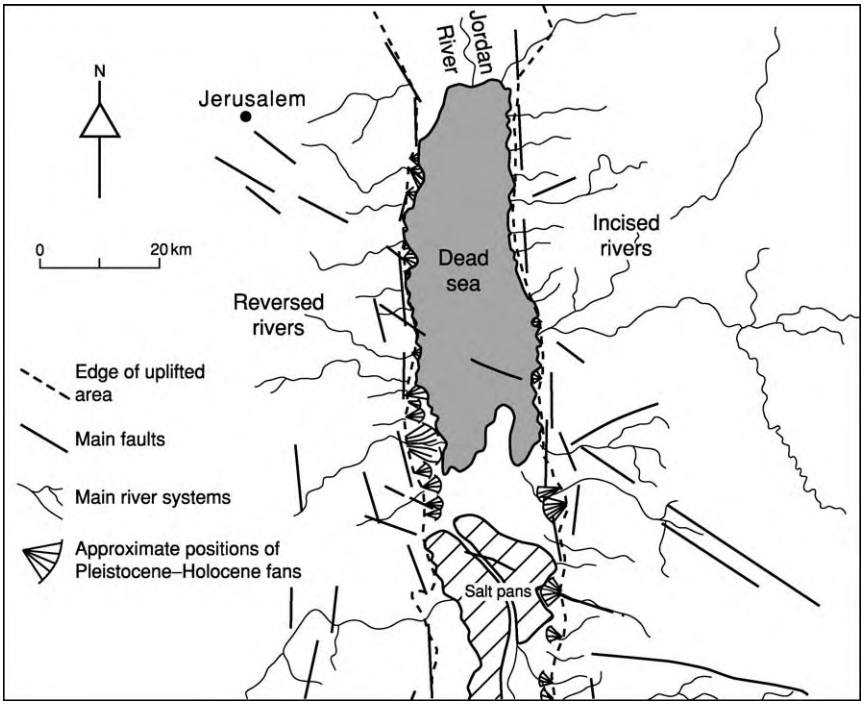


Figure 10 Patterns of faulting and their influence on the development of river systems around the Dead Sea pull apart basin.

spans the 1100 km between the Gulf of Elat/Aqaba and Turkey and separates the Arabian Plate to the east from the African Plate to the west (Figure 2). Since the Miocene, a period of about 20 Ma, Arabia is thought to have moved more than 105 km northwards, a type of movement that is termed strike-slip.

The Dead Sea Basin is a zone where the movement has resulted in local tension, producing faulting and leading to the sinking of a section of the crust. Such basins are termed 'pull-apart' basins and are characterized by very rapid subsidence and thick basin-fill sequences. The overall structure of the Dead Sea Rift is asymmetrical, not dissimilar to that of the East African Rift. The largest fault is in the eastern margin and forms the Jordanian shore of the lake. Here, the faulting exposes a spectacular rock sequence more than 1 km thick, which ranges in age from Precambrian (more than 544 Ma) to Pleistocene (less than 1 Ma). On the opposite side of the basin are a number of smaller subparallel faults, which cut the Cretaceous limestones of this margin into a series of structural steps (Figure 10). At either end of the basin are cross-rift structures that link movement along the Areva fault to the south with movement on the Jordan fault to the north.

A major feature of the southern part of the basin is the development of salt diapirs. These result from subsurface movements of thick deposits of rock salt, which can push up and punch through the overlying sediments and penetrate to the surface. One famous example of such a diapir is Mount Sedom, famous for its biblical links with the doomed and 'sinful' cities of Sodom and Gomorrah.

Rivers and Hydrology

The development of the Dead Sea Rift system disrupted a pre-existing drainage system that crossed from east to west across the Jordan plateau and drained into the Mediterranean. The headwaters of this system now run across the eastern scarp of the Dead Sea and have cut gorges over a kilometre deep to reach the lake shore (Figure 10). On the western shore a new set of rivers have evolved, which no longer drain into the Mediterranean Sea to the west but instead have been reversed and now drain from west to east. These rivers have also cut down into pre-rift rocks and run in gorges that are less deep than those of their eastern equivalents.

The present climate of the area is desert to semi-desert with rainfall of 50–200 mm year ^{1}. Because of this all rivers except the Jordan, which has headwaters in an area of higher rainfall to the north, are ephemeral and flow only in response to winter rain storms.

The lake water is renowned for its high salinity, which is 10 times that of normal seawater. Tourists

are attracted to 'swim' in the waters, which are so buoyant that individuals can sit unsupported and read a paper. The high salinity is a result of a combination of evaporation in a closed basin and the influence of brines coming from the solution of subsurface rock salt. The lake brines are particularly rich in chlorine and bromine, which are extracted in salt ponds and exported worldwide.

Climate Change and the Basin Fill

Lake levels in closed basins are very susceptible to the effects of climate change. Any increase in rainfall will upset the hydrological balance and cause lake levels to rise and salinity to fall. If rainfall decreases, lake levels will drop and evaporation will dominate, resulting in an increase in salinity. The surface of the Dead Sea shows evidence of having fluctuated between 180 m and 700 m below sea-level over the past 60 Ka in response to well-documented changes in climate. The rising and falling lake levels have a profound effect on the sedimentary deposits of the rift. High lake levels, such as those that prevailed during the deposition of the Pleistocene Lisan Formation, result in thick sequences of interlaminated chalk and silty clay (**Figure 11**). During periods of lower lake levels the river and fan deposits penetrate far into the basin and dominate the sequences. One surprising consequence of depressed lake levels is a change in the balance between saline and fresh groundwaters, with the latter penetrating further towards the axis of the basin. Since much of the basin axis is underlain by thick salt deposits, the fresh groundwater dissolves the preserved layers of salt, generating subsurface caverns and solution holes. This is currently happening in response to lake levels falling as a result of over abstraction of water from the Jordan River.

Earthquakes, Archaeology, and Sodom and Gomorrah

Earthquakes have been a feature of the Dead Sea Rift throughout its history. The earthquakes are generated by movement along the main fault zone and are often accompanied by the release of asphalt, gases, and tars, which are trapped in the layers of rock beneath the surface. The asphalt in particular is well documented and is found in layers within the older lake deposits. Fault movements tend to happen sporadically: long periods of quiescence are succeeded by times when earthquakes are regular events.

The Dead Sea Basin has been inhabited by local peoples for many thousands of years. The alluvial plains of the valley were rendered fertile by irrigation, and trading routes to the south, east, and west allowed early settlers to exploit the mineral wealth of the area, including gathering and trading materials from oil seeps and asphalt, which have been found as far away as Egypt in the tombs of the Pharaohs. The early Bronze Age was a time when the basin was well populated and was also a tectonically quiet period when few earthquakes occurred. Towards the end of this period there was a large earthquake, which may have resulted in the destruction of two major cities, Sodom and Gomorrah. There has been speculation about precisely how and why these cities were so

Figure 11 A section through the Lisan Formation of the Dead Sea, showing layers of chalk and silt (horizontal layers at the top and bottom of the section), some of which have been disturbed by earthquake activity (folded layers in the centre of the section).

comprehensively demolished that they were never rebuilt. One theory is that they were built on soft sediments that became liquid (liquefaction) as they were shaken, maximizing the instability of the ground (*see* **Engineering Geology:** Liquefaction). Interestingly, the occurrence of 'sulphurous' fires reported in the bible corresponds well with the release of the light fractions of oil from underground reservoirs as the ground moves and slides in response to shaking. It seems likely that the myths surrounding the destruction of Sodom and Gomorrah are based in fact and are a direct consequence of the unique geology of the area.

See Also

Biblical Geology. **Engineering Geology:** Liquefaction. **Fossil Vertebrates:** Hominids. **Geomorphology**. **Sedimentary Environments:** Lake Processes and Deposits. **Tectonics:** Earthquakes; Faults; Mid-Ocean Ridges; Rift Valleys.

Further Reading

Allen PA and Allen JR (1990) *Basin Analysis: Principles and Applications*. Oxford: Blackwells.

Enzel Y, Kadan G, and Eyal Y (2000) Holocene earthquakes inferred from a fan delta sequence in the Dead Sea graben. *Quaternary Research* 53: 34 48.

Frostick LE and Reid I (1989) Is structure the main control on river drainage and sedimentation in rifts? *Journal of African Earth Sciences* 8: 165 182.

Frostick LE and Steel RJ (eds.) (1993) *Tectonic Controls and Signatures in Sedimentary Successions*. International Association of Sedimentologists Special Publication 20. Oxford: Blackwells.

Frostick LE, Renaut RW, Reid I, and Tiercelin JJ (1986) *Sedimentation in the African Rifts*. Special Publication 25. London: Geological Society.

Girdler RW (1991) The Afro Arabian Rift System: an over view. *Tectonophysics* 197: 139 153.

Gupta S and Cowie P (2000) Processes and controls on the stratigraphic development of extensional basins. *Basin Research* 12: 185 194.

Neev D and Emery KO (1995) *The Destruction of Sodom, Gomorrah and Jericho*. Oxford: Oxford Univer sity Press.

Selley RC (ed.) (1997) *African Basins*. Sedimentary Basins of the World 3. Amsterdam: Elsevier.

Summerfield MA (1991) *Global Geomorphology: An Introduction to the Study of Landforms*. Harlow: Longman.

AGGREGATES

M A Eden and W J French, Geomaterials Research Services Ltd, Basildon, UK

Introduction

Aggregates are composed of particles of robust rock derived from natural sands and gravels or from the crushing of quarried rock. The strength and the elastic modulus of the rock should ideally match the anticipated properties of the final product.

Aggregates are used in concrete, mortar, road materials with a bituminous binder, and unbound construction (including railway-track ballast). They are also used as fill and as drainage filter media.

In England alone some 250 million tonnes of aggregate are consumed each year, representing the extraction of about 0.1 km^3 of rock, if necessary wastage is taken into account. Aggregates may be derived from rocks extracted from quarries and pits, or from less robust materials. For example, slate and clay can be turned, by heating, into useful expanded aggregates of low bulk density.

The principal sources of aggregate are sand and gravel pits, marine deposits extracted by dredging, and crushed rock from hard-rock quarries. As extracted, these materials would rarely make satisfactory aggregate. They need to be carefully prepared and cleaned to make them suitable for their intended purpose. The sources may also be rather variable in their composition and in the rock types present, so it is essential that potential sources are carefully evaluated. At the very least, the preparation of the aggregate involves washing to remove dust and riffling to separate specific size ranges.

The classification of aggregates varies greatly. An early classification involved the recognition of Trade Groups, which were aggregates consisting of rocks thought to have like properties and which could be used for a particular purpose. A fairly wide range of rock types was therefore included in a given Group. More recent classifications have been based on petrography. Again, these groups tend to be broad, and they focus on the macroscopic properties of the materials for use as aggregate rather than on detailed petrographic variation.

Because aggregates consist of particulate materials, whether crushed or obtained from naturally occurring sands and gravels, their properties are normally measured on the bulk prepared material. There are

therefore numerous standard tests that relate to the intended use of the material. Standard tests vary from country to country, and, in particular, collections of standard tests and expected test results are given in specific British and American Standards.

Many defective materials can occur within an aggregate. It is therefore essential that detailed petrographic evaluation is carried out, with particular reference to the intended use. An example of failure to do this was seen in the refurbishment of a small housing estate: white render was applied to face degraded brickwork. At first the result was splendid, but within 2–3 years brown rust spots appeared all over the white render because of the presence of very small amounts of iron sulphide (pyrite) in the sand used in the render.

Aggregate sources

Sands and gravels can be obtained from river or glacial deposits, many of which are relatively young unconsolidated superficial deposits of Quaternary age. They may also be derived from older geological deposits, such as Triassic and Devonian conglomerates (to take English examples). Flood plain and terrace gravels are particularly important sources of aggregate because nature has already sorted them and destroyed or removed much of the potentially deleterious material; however, they may still vary in composition and particle size. Glacial deposits tend to be less predictable than fluvial deposits and are most useful where they have been clearly sorted by fluvial processes.

Among the quarried rocks, limestones – particularly the Carboniferous limestones of the British Isles – have been widely used as aggregate. Similarly, many sandstones have suitable properties and are used as sources of aggregate, particularly where they have been thoroughly cemented. Compact greywackes have been widely used, notably the Palaeozoic greywackes of the South West and Wales.

Igneous rocks are also a very useful source of quarried stone when crushed to yield aggregates; their character depends on their mineralogy and texture. Coarsely crystalline rocks such as granite, syenite, diorite, and gabbro are widely used, as are their medium-grained equivalents. Some finer-grained igneous rocks are also used, but the very finest-grained rocks are liable to be unsatisfactory for a wide range of purposes. Reserves of rocks such as dolerite, microgranite, and basalt tend to be small in comparison with the coarse-grained intrusive plutons. Conversely, some of the high-quality granite sources lie within very large igneous bodies, which sustain large quarries and provide a considerable resource.

Regional metamorphic rock fabrics generally make poor aggregate sources. On crushing they develop an unsatisfactory flaky shape. Schists and gneisses can provide strong material, but of poor shape. On the other hand, metamorphism of some greywackes and sandstones can provide material of high quality, especially when it has involved contact metamorphism associated with the intrusion of igneous rocks, producing hornfels or marble. Such thermally metamorphosed rocks often have a good fabric and provide useful resources.

Investigation of Sources

There are three levels of investigation of the potential aggregate source. The first is the field investigation, in which the characteristics and distributions of the rocks present in the source can be established by mapping, geophysics, and borehole drilling. The second concerns the specific petrography of the materials. The third involves testing the physical and chemical properties of the materials. The material being extracted from the source must also be tested on a regular basis to ensure that there is no departure from the original test results and specification. Because sources are inevitably variable from place to place, there is always the risk that certain potentially deleterious components may appear in undesirable abundance.

A number of features may make the aggregate unsuitable for certain purposes; these include the presence of iron sulphide (pyrite, pyrrhotite, and marcasite). Iron sulphide minerals are unacceptable because they become oxidized on exposure to air in the presence of moisture, producing iron oxides (rust) and sulphate. This can result in spalling of material from the surface of concrete and rendering. The presence of gypsum in the aggregate is also highly undesirable from the point of view of concrete durability. Gypsum is commonly found in aggregates from arid regions. The presence of gypsum in concrete leads to medium- to long-term expansion and cracking. Other substances can create both durability and cosmetic problems.

Extraction of Aggregates

The development of aggregate quarries requires the removal of overburden and its disposal, the fragmentation of rock (usually by a scheme of blasting), and the collection and crushing of the blast product (*see* **Quarrying**). Critical to the success of the operation is the stability of the size of the feed material to the primary crusher. Screening is usually necessary to ensure that the particles are suitable for the crusher regime. At this stage it is also necessary to remove

degraded and waste material that is not required as part of the aggregate.

In sand and gravel workings, the source material is excavated in either dry or wet pit working. In marine environments, the process is based on suction and dredging using two techniques. In the first, the dredger is anchored and a pit is created in the seabed; production continues as consolidated materials fall into the excavation. In contrast, trail dredging is performed by a moving vessel, which excavates the deposit by cutting trenches in the seabed.

Extracted crushed rock, sand, and gravel are then prepared as aggregates through the use of jaw, gyratory, impact, and cone crushers. The type of crusher is selected according to the individual sizes of the feed material. Grading by screening is an adjunct to comminution and is also necessary in the production and preparation of the finished aggregate in cases where the particle-size distribution of the aggregate is important. The product is also washed and cleaned. The process of cleaning often uses density separation, with weak porous rock types of low density being removed from the more satisfactory gravel materials.

Classification

The classification of aggregates has changed significantly over the years but has always suffered from the need to satisfy many different interests. Most commonly aggregates are divided into natural and artificial and, if natural, into crushed rock, sand, and gravel. If the aggregate is a sand or gravel, it is further subdivided according to whether it is crushed, partly crushed, or uncrushed. It may then be important to state whether the material was derived from the land or from marine sources.

Once produced, the aggregate is identified by its particle size, particle shape, particle surface texture, colour, the presence of impurities (such as dust, silt, or clay), and the presence of surface coatings or encrustations on the individual particles.

Detailed petrographic examination is employed so that specific rock names can be included in the description. This also helps in the recognition of potentially deleterious substances. However, the diversity of rock names means that considerable simplification is required before this classification can be used to describe aggregates. Following recognition of the main category of rock from the field data, more specific names can be applied according to texture and mineral composition. Because aggregates are used for particular purposes, they are sometimes grouped according to their potential use. This means that they may be incorrectly named from a geological point of view. The most obvious example of this is where limestone is referred to as 'marble'. In 1913 a list of petrographically determined rock types was assembled, with the rocks being arranged in Trade Groups. This was thought to help the classification of road stone in particular. It was presumed that each Trade Group was composed of rocks with common properties. However, the range of properties in any one Group is so large as to make a nonsense of any expectation that the members of the Group will perform similarly, either in tests or in service. The Trade Groups were therefore replaced by a petrological group classification.

However, even rocks within a single petrographic group can vary substantially in their properties. For example, the basalt group includes rocks that are not basalt, such as andesite, epidiorite, lamprophyre, and spilite. Hence a wide range of properties are to be expected from among these diverse lithologies.

In the first place a classification describes the nature of the aggregate in a broad sense: quarried rock, sand, or gravel; crushed or otherwise. Second, the physical characteristics of the material are considered. Third, the petrography of the possibly diverse materials present must be established. This may require the examination of large and numerous samples. While it may be reasonable to describe as 'granite' the aggregate produced from a quarry in a mass of granite, that aggregate will inevitably contain a wide range of lithologies, including hydrothermally altered and weathered rocks. Whether a rock is geologically a granite, a granodiorite, or an adamellite may be less significant for the description of the aggregate than the recognition of the presence of strain within the quartz, alteration of the feldspar, or the presence of shear zones or veins.

Aggregate Grading

Aggregate grading is determined by sieve analyses. Material passing through the 5 mm sieve is termed fine aggregate, while coarse aggregate is wholly retained on this sieve (**Figure 1**). The fine aggregate is often divided into three (formerly four) subsets – coarse, medium, and fine – which fall within specified and partly overlapping particle-size envelopes. The size range is sometimes recorded as the ratio of the sieve sizes at which 60% passes and at which 10% passes. The shapes of the particles greatly affect the masses falling in given size ranges. For example, an aggregate with a high proportion of elongate grains of a given grain size would be coarser than an aggregate with flaky particles. This can affect the properties of materials made using the aggregate for, say, concrete, road materials, and filter design. Commonly materials needed for particular purposes have standard

Figure 1 Aggregate grades. (A) Fine sand suitable for mortars or render (width of image: 10 mm). (B) Coarse sharp sand or 'concreting' sand (width of image: 10 mm). (C) Coarse natural sand (width of image: 10 mm). (D) Flint gravel 5 10 mm (width of image: 100 mm). (E) Crushed granite 5 10 mm (width of image: 100 mm). (F) Crushed granite 10 20 mm (width of image: 100 mm).

aggregate gradings. These include, for example, mortars, concrete, and road-surface aggregates. It is sometimes useful to have rock particles that are much larger than the normal maximum, for example where large masses of concrete are to be placed. Commonly, however, the maximum particle size used in structural concrete is around 20 mm. An important parameter is the proportion of dust, which is often taken as the amount passing the 75 μm sieve. In blending aggregates for particular purposes, it is usually necessary to combine at least two and possibly more size ranges; for example, in a concrete the aggregate may be a mixture of suitable material in the size ranges 0–5 mm, 5–10 mm, and 10–20 mm.

The grading curve – a plot of the mass of material passing each sieve size – also determines the potential workability of mixtures and the space to be filled by binder and can be adjusted to suit particular purposes. The grading curve can be designed to reduce the volume of space to less than 10% of the total volume, but at this level the aggregate becomes almost completely unworkable.

Particle Shape

Particle shape is important in controlling the ability of the aggregate to compact, with or without a binder, and affects the adhesion of the binder to the aggregate surface. Shapes are described as rounded, irregular, angular, flaky, or elongate, and can be combinations of these (Figure 2). The first three are essentially equidimensional. The shape is assessed by measuring the longest, shortest, and intermediate axial diameters of the fragments. In the ideal equidimensional fragment, the three diameters are the same. Particles with ratios of the shortest to the intermediate and the intermediate to the longest diameters of above about 0.6 are normally regarded as equidimensional.

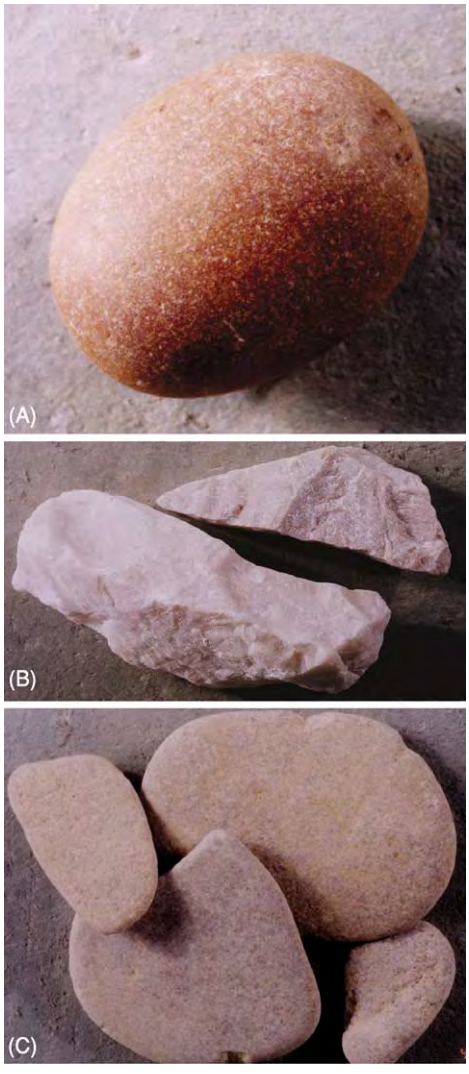


Figure 2 Examples of particular particle shapes. (A) Well rounded spherical metaquartzite. (B) Elongate angular quartzite. (C) Rounded flaky limestone.

For many purposes, it is important that the aggregate particles have equant shape: their maximum and minimum dimensions must be very similar. Spherical and equant particles of a given uniform size placed together have the lowest space between the particles. Highly angular particles and flaky particles with high aspect ratios of the same grading can have much more space between the particles. The shape of the particles can significantly affect the properties and composition of a mixture. The overall space is also determined by the grading curve. Sometimes highly flaky particles such as slate can be used in a mixture if they are accompanied by suitably graded and highly spherical particles.

Flakiness Index (British Standard 812)

The flakiness index is measured on particles larger than 6.5 mm and is the weight percentage of particles that have a least dimension of less than 0.6 times the mean dimension. The sample must be greater than 200 pieces. The test is carried out using a standard plate that has elongate holes of a given size; the proportion passing through the appropriate hole gives a measure of the flakiness index.

Elongation Index (BS 812)

The elongation index is the percentage of particles by mass having a long dimension that is more than 1.8 times the mean dimension. This measurement is made with a standard gauge in which pegs are placed an appropriate distance apart.

Petrography

The petrography of the aggregate is mainly assessed on the basis of hand picking particles from a bulk sample. Thin-section analysis either of selected pieces or of a crush or sand mounted in a resin is also employed. The petrographic analysis is essential to determine the rock types present and hence to identify potential difficulties in the use of the material. It allows recognition of potentially deleterious components and estimation of physical parameters. The experienced petrographer, for example, can estimate the parameters relevant to the use of a material for road surfacing.

Published standards provide procedures for petrographic description, including the standards published by the American Society for Testing and

Materials and the Rilem procedures. These standards list the minimum amounts of material to be examined in the petrographic examination. In BS 812, for example, it is specified that for an aggregate with a maximum particle size of 20 mm the laboratory sample should consist of 30 kg. The minimum mass of the test portion to be examined particle by particle is 6 kg. Normally the analysis would be carried out on duplicate portions. The samples are examined particle by particle, using a binocular stereoscopic microscope if necessary. Unfortunately, this procedure does not cover all eventualities, and some seriously deleterious constituents within the material may be missed. A rock particle passing a 20 mm sieve may have within it structures that give it potentially deleterious properties (Figure 3). It is therefore essential that the aggregate is examined in thin section as well as in the hand specimen. It is helpful if the aggregate sample is crushed and resampled to provide a representative portion for observation in thin section. A large thin section carrying several hundred particles is required. Some of the potentially deleterious ingredients may be present at relatively low abundance. For example, the presence of 1–2% of opaline vein silica would be likely to cause significant problems.

Where a sand or fine gravel is to be sorted by hand it is first divided into sieve fractions, typically using the size ranges <1.18 mm, 1.18–2.36 mm, 2.36–5 mm, and >5 mm. These size fractions are analysed quantitatively by hand sorting in the same way as for coarse aggregate. The stereoscopic microscope is used to help with identification. Thin sections are also prepared from the sample using either the fraction passing the 1.18 mm sieve or the whole fine aggregate. The sample is embedded in resin and a thin section is made of the briquette so produced.

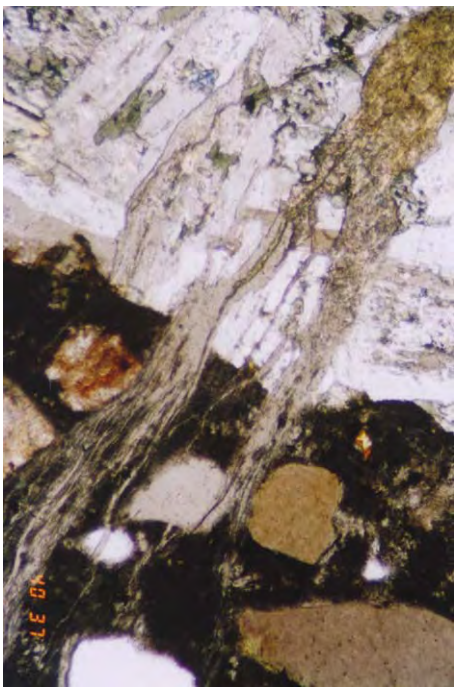


Figure 3 An alkali reactive granite coarse aggregate particle (top) with cracks filled with alkali silicate gel. The cracks run into the surrounding binder, which appears dark and contains quartz rich sand as a fine aggregate.

Specific Tests Measuring Strength, Elasticity, and Durability

For quarried rocks it is possible to take cores of the original source material and to measure the compressive and tensile strengths of that material directly. It may be necessary to take a large number of samples in order to obtain a reliable representative result. However, for sands and gravels the strength of the material can rarely be tested in this way, and so a series of tests has been developed that simulate the conditions in which the aggregate is to be used.

There is often a simple relationship between the flakiness index of the aggregate and its aggregate impact value (AIV) and aggregate crushing value (ACV). In general, the lower the flakiness index, the higher the AIV and ACV. Hence, comparing the AIV and ACV values with specifications requires knowledge of the flakiness index. Consideration also needs to be given to the shape of the aggregate following the test.

Density and Water Absorption

Some of the most important quantities measured for an aggregate are various density values. These include the bulk density, which is the total mass of material in a given volume, including the space between the aggregate particles. The saturated surface-dry density is the density of the actual rock material when fully saturated with water but having been dried at the surface. The dry density is the rock density after drying. In making these measurements, the water absorption is also recorded. These provide data that are essential for the design of composite mixes.

Aggregate Impact Value (BS 812)

The aggregate impact value provides an indirect measurement of strength and involves the impaction of a standard mass on a previously well-sorted sample. The result is obtained by measuring the amount of material of less than 2.36 mm produced from an aggregate of 10–14 mm. The lower the result, the greater the resistance of the rock to impaction. It is also useful to examine the material that does not pass the 2.36 mm sieve, and it is common to sieve the total

product at 9.5 mm to establish whether there is an overall general reduction in particle size.

Aggregate Crushing Value (BS 812)

The aggregate crushing value provides an indirect assessment of strength and elasticity in which a well-sorted sample is slowly compressed. The lower the degradation of the sample, the greater the resistance to crushing. The size ranges used are the same as for the AIV test.

10% Fines Value (BS 812)

The 10% fines value is the crushing load required to produce degradation such that 10% of the original mass of the material passes a 2.36 mm sieve, the original test sample being 10–14 mm. The samples are subjected to two different loads, and the amount passing the 2.36 mm sieve in each test is measured. Typically the two results should fall between 7.5% and 12.5% of the initial weight. The force required to produce 10% fines is then calculated.

Aggregate Abrasion Value (BS 812)

In determining the aggregate abrasion value, fixed aggregate particles are abraded with standard sand, and the mass of the aggregate is recorded before and after abrasion. The reduction in mass indicates the hardness, brittleness, and integrity of the rock.

The Los Angeles Abrasion Value (ASTM C131 and C535)

To determine the Los Angeles abrasion value, a sample charge is mixed with six to twelve steel balls, and together these are rotated in a steel cylinder for 500 or 1000 revolutions at 33 rpm. This causes attrition through tumbling and the mutual impact of the particles and the steel balls. The sample is screened after the rotations are completed using a 1.68 mm sieve. The coarser fraction is washed, oven dried, and weighed. The loss in mass as a percentage of the original mass is the Los Angeles abrasion value.

Micro Deval test

The Micro Deval test is widely used to determine the resistance of an aggregate to abrasion. Steel balls and the aggregate are placed in a rotating cylinder. The test may be carried out either wet or dry. The Micro Deval value is calculated from the mass of material that passes the 1.6 mm test sieve, as a percentage of the original aggregate mass.

Polished Stone Value (BS 812, Part 114)

To determine the polished stone value, the aggregate is mounted in resin and the exposed surface is polished using a wheel and standard abrasive. The result is measured using a standard pendulum, with the ability of the rock to reduce the motion of the pendulum giving an indication of the potential resistance of the aggregate to skidding. The sample is small and the result can vary according to the proportions of rock that are present. This test is difficult to perform reliably, and considerable practice is required to obtain a consistent result. In practice it is found that good skid resistance is derived from a varied texture in the rock with some variation in particle quality. Well-cemented sandstones and some dolerites tend to have high polished stone values, while rocks such as limestones and chert have very low polished stone values.

Franklin Point Load Strength

The Franklin point load strength can be directly assessed for large pieces of rough rock. A load is applied through conical platens. The specimen fails in tension at a fraction of the load required in the standard laboratory compressive-strength test. However, the values obtained in the test correlate reasonably well with those obtained from the laboratory-based uniaxial compressive test, so an estimated value for this can be obtained, if necessary, in the field.

Schmidt Rebound Hammer Value

The Schmidt Rebound Hammer test is a simple quantitative test in which a spring-loaded hammer travelling through a fixed distance strikes the rock in a given orientation. The rebound of the hammer from the rock is influenced by the elasticity of the rock and is recorded as a percentage of the initial forward travel. A sound rock will generally give a rebound value in excess of 50%, while weathered and altered rock will tend to give a much lower value.

Magnesium Sulphate Soundness Test (BS 812)

In the magnesium sulphate soundness test the degradation of the aggregate is measured following alternate wetting and drying in a solution of magnesium sulphate. The test provides a measure of the tendency of the rock to degrade through the crystallization of salts or ice formation. The result is influenced by the porosity and particularly by planes of weakness in the aggregate.

Freeze–Thaw Test

In the freeze–thaw test the aggregate is subjected to cycles of freezing and thawing in water. Each cycle lasts approximately 24 h. The temperature is reduced over a period of several hours and then

maintained at −15°C to −20°C for at least 4 h. The sample is then maintained in water at 20°C for 5 h. The cycle is repeated 10 times, and then the sample is dried and sieved, and the percentage loss in mass is determined.

Slake Durability Index

A number of small samples of known mass are placed in a wire-mesh drum. The drum is immersed in water and rotated for 10 min. The specimens are dried and weighed, and any loss in weight is expressed as a percentage of the initial weight. This is the slake durability index.

Methylene Blue Absorption Test

Methylene blue dye is dissolved in water to give a blue solution. It is absorbed from the solution by swelling clay minerals, such as montmorillonite. The quantity of potentially swelling clay minerals in a sample of rock is assessed by measuring the amount of methylene blue absorbed.

Chemical Tests

Aggregates are commonly tested by chemical analysis for a variety of constituents, including their organic, chloride, and sulphate contents. Organic material is readily separated from the aggregate by, for example, the alkalinity of cement paste. Its presence leads to severe staining of concrete and mortar surfaces. Sulphate causes long-term chemical changes in cement paste, leading to cracking and degradation. Chloride affects the durability of steel reinforcement in concrete, accelerating corrosion and the consequent reduction in strength.

Mortar Bar and Concrete Prism Tests

The durability of concrete made with a given aggregate is evaluated by measuring the dimensional change in bars made of mortar or larger prisms of concrete containing the specific aggregate. The mortar-bar test results can be obtained in a few weeks, but the prism test needs to run for many months or even years. The tests allow the recognition of components in the rocks or contaminants (e.g. artificial glass) that take part in expansive alkali–aggregate reactions.

Aggregates for Specific Purposes

Railway Track Ballasts

Railway track is normally placed on a bed of coarse aggregate. A lack of fines is required: the desirable particle size is generally 20–60 mm. The bed requires a free-draining base that is stable and able to maintain the track alignment with minimum maintenance. The aggregate is sometimes placed on a blanket of sand to prevent fines entering the coarse aggregate layer. The aggregate layer may be up to 400 mm thick.

The favoured rock types are medium-grained igneous rocks such as aplite and microgranite. Sometimes hornfels is used. Some of the more durable limestones and sandstones are also used. Weaker limestones and many sandstones are generally regarded as unsatisfactory because of their low durability and ready abrasion. The desirable qualities for an aggregate used for ballast are that it must be a strong rock, angular in shape, tending to be equidimensional, and free from dust and fines.

Aggregates for Use in Bituminous Construction Materials

Aggregates for use with a bitumen binder in building construction (as used in bridge decks and in the decks and ramps of multistorey car parks) require a high skid resistance. They must also be highly impermeable, protecting the underlying construction from water and frost attack and from the effects of deicing salts. The mix design is important: there should be a high bitumen content and a high content of fine aggregate and filler in the aggregate grading.

A wide range of rocks of diverse origin and a number of artificial materials are used in the bituminous mixes. The rocks must be durable, strong, and resistant to polishing. The aggregate must show good adhesion to the binder and have good shape. Skid resistance is also dependent on traffic density and, in some instances, a reduction in traffic has improved skid resistance. Visual aggregates have been developed where high skid resistance is required, and these include calcined bauxite, calcined flint, ballotini, and sinopal. Blast furnace slags yield moderately high polished stone values. The light-reflecting qualities are also important, and artificial aggregates such as sinopal, with their very high light reflectivity, are valued. Resistance to stripping, i.e., the breakdown of the bond between the aggregate and the bituminous binder, is also important. Stripping is likely to result in the failure of the wearing course and not necessarily in failure of the base course. The stripping tends to be most conspicuous in coarse-grained aggregates that contain quartz and feldspar. Basic rocks show little or no detachment. The aggregate has considerable strength, particularly in the wearing course. As an example, the aggregate crushing value for surface chasing and dense wearing courses will typically be 16 to 23, while for the base course it may be as high as 30. Similarly, the aggregate impact value might be 23 in the wearing course and 30 in the base course.

Aggregates in Unbound Pavement Construction

Aggregate is sometimes used in construction without cement or a bitumin binder. Examples are a working platform in advance of construction, structural layers beneath a road system, a drainage layer, and a replacement of unsuitable foundation material. Aggregates for these purposes must be resistant to crushing and impact effects during compaction and in use, and when in place they must resist breakdown by weathering or by chemical and physical processes and must be able to resist freeze–thaw processes.

It is likely that recycled aggregates will become increasingly important in these situations, although levels of potentially deleterious components, such as sulphate, may point to a need for caution in the use of such material. Aggregates for unbound construction often need to resist the ingress of moisture, since moisture rise and capillary transfer can cause progressive degradation.

Mortar

Mortar consists of a fine aggregate with a binding agent. It is used as a jointing or surface-rendering material. Sands for mortar production are excavated from sand and gravel pits in unconsolidated clastic deposits and are typically dominated by quartz. They are used in their natural form or processed by screening and washing. Rock fines of similar grade can also be used.

The most important feature of sand for mortar manufacture is that the space between the aggregate particles must generally be about 30% by volume. The volume of binder needs to be slightly greater than this volume, and hence a relatively high proportion of cement or lime may be required. Should the space be such that voids occur in the mix, the material will commonly show early signs of degradation and will be readily damaged by penetration of moisture. The space also appears to reduce the capacity of the mortar to bond with the substrate.

The workability and ease of use of the mixture also depends on the shape of the particles and the grading curve. Very uniform sand tends to have a high void space and therefore requires a high cementitious or water content and tends to develop a high voidage. On the other hand, the grading may be such that the space between the particles is too small and the mixture becomes stiff. The strength and elastic modulus of the rocks are also important because the resultant mixture of paste and aggregate must match the strength and elasticity of the material to which the mortar is applied. If it is not, then partings are liable to develop between the binder and the substrate. Similarly, the material must exhibit minimal shrinkage because again it might become detached from the substrate.

Concrete

This very widely used material has a very diverse structure and composition and serves many purposes. It is composed of aggregate graded for the specific purpose and a binder containing cement. In general, the properties of the aggregate must match the intended strength and elasticity of the product, and it must be highly durable. For many purposes a combination of coarse and fine aggregate with a maximum particle size of 20 mm is used. The grading curve is designed such that an appropriate amount of space occurs between the particles – typically around 25% by volume of the mixture. There are numerous components of aggregate that perform adversely in the medium and long term, so careful study of the material is required before use. The defective components are described in several standards, along with procedures for measuring their effects on the concrete. Some of these are described below.

In the 1940s it was recognized in the USA that certain siliceous aggregates could react with alkalis derived from Portland Cement. This led to spalling of concrete surfaces and cracking, sometimes in a spectacular manner. The phenomenon occurs throughout the world, and few rock sources are immune. An enormous amount of work has been carried out to evaluate the reaction, both in the laboratory and in structures. Major international conferences on the subject have been held. The alkalis for the reaction derive from the cement and are extracted into the pore fluid in the setting concrete. The concentration of alkali in the pore fluid can be affected by external factors as well as by the internal composition of the cement matrix. The rock reacting with the alkalis is typically extremely fine grained or has extremely small strain domains. Hence, fine-grained rocks, such as opaline silica within limestone, some cherts, volcanic glass, slate, and similar fine-grained metamorphic rocks, may exhibit a high degree of strain and so be able to take part in the reaction. More recently it has been found that certain dolomitic siliceous limestones are also to be avoided, again because they react with alkalis to cause significant expansion of the concrete and severe cracking.

See Also

Building Stone. **Geotechnical Engineering**. **Quarrying**. **Rock Mechanics**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Glaciers. **Sedimentary Rocks:** Limestones; Sandstones, Diagenesis and Porosity Evolution.

Further Reading

American Society for Testing and Materials (1994) *Annual Book of ASTM Standards (1994), Section 4, Construction, Volume 04.02, Concrete and Aggregates*. West Conshohocken: American Society for Testing and Materials.

Bérubé MA, Fournier B, and Durand B (eds.) (2000) *Alkali Aggregate Reaction in Concrete*. Proceedings of the 11th International Conference, Quebec, Canada.

British Standards Institution (1990) *BS812 Parts 1 to 3: Methods for Sampling and Testing of Mineral Aggregates, Sands and Fillers, Parts 100 Series Testing Aggregates*. British Standards Institution.

Dolor Mantuani L (1983) *Handbook of Concrete Aggregates: A Petrographic and Technological Evaluation*. New Jersey: Noyes Publications.

(1983) *FIP Manual of Leightweight Aggregate Concrete*, 2nd edn. Surrey University Press (Halsted Press).

Hobbs DW (1988) *Alkali Silica Reaction in Concrete*. Thomas Telford.

Latham J P (1998) *Advances in Aggregates and Armour stone Evaluation*. Engineering Geology Special Publication 13. London: Geological Society.

Popovics S (1979) *Concrete Making Materials*. Hemisphere Publishing Corporation, McGraw Hill Book Company.

Smith MR and Collis L (2001) *Aggregates, Sand, Gravel, and Crushed Rock for Construction Purposes*, 3rd edn. Engineering Geology Special Publication 17. London: Geological Society.

West G (1996) *Alkali Aggregate Reaction in Concrete Roads and Bridges*. Thomas Telford.

ALPS

See **EUROPE: The Alps**

ANALYTICAL METHODS

Contents

Fission Track Analysis

B W H Hendriks, Geological Survey of Norway, Trondheim, Norway

Introduction

Ages obtained from isotopic dating methods are based on the ratio of parent and daughter isotopes. Radioactive decay of parent isotopes causes daughter isotopes to accumulate over time, unless they decay further or are lost by diffusion or emission. In the case of the fission track method, the daughter product is not another isotope, but a trail of physical damage to the crystal lattice resulting from spontaneous fission of the parent nucleus. When the rate at which spontaneous fission occurs is known, the accumulation of such trails, known as fission tracks, can be used as a dating tool. Analogous to diffusional loss of daughter isotopes, the damage trails in the crystal lattice disappear above a threshold temperature by the fission track annealing process. Although the physics behind the annealing process are poorly understood, the outcome is empirically well known. Annealing initially causes the length of fission tracks to decrease and may eventually completely repair the damage to the crystal lattice. The latter is known as total annealing. The rate

at which annealing takes place is a function of both mineral properties and temperature history.

Fission tracks in geological samples have been well-studied in mica (*see* **Minerals:** Micas), volcanic glass, tektite glass; (*see* **Tektites**) titanite, and zircon (*see* **Minerals:** Zircons). However, most research has been done on fission tracks in apatite, a widely disseminated accessory mineral in all classes of rocks. Retention of fission tracks in natural minerals takes place only at temperatures well below that of their crystallization temperature. Fission track dating will, therefore, document the crystallization age of a crystal only when it has cooled rapidly to surface temperatures immediatley after crystallization (*see* **Analytical Methods:** Mineral Analysis). Fission track dating of volcanic rocks can provide an age of crystallization, while fission track dating of more slowly cooled rocks will always yield an age that is younger than the age of crystallization. The amount of fission tracks per volume and their length will be a sensitive function of the annealing process and of the cooling history of the sample being studied. A cooling history can be constrained by thermal history modelling of fission track data (fission track age and fission track length distribution). Fission track analysis and thermal history modelling of apatite fission track data provide powerful tools with which to assess regional cooling and denudation histories.

Following the rejuvenation of (U-Th)/He dating in the 1990s, the technique has become an important addition to the fission track method. (U-Th)/He dating can be applied to the same minerals as those commonly used in fission track analysis. (U-Th)/He dating is unique in its capability to constrain the very low temperature part of cooling histories of rock samples; the nominal closure temperature for apatite (U-Th)/He ages may be as low as ~50°C. Apatite (U-Th)/He dating today is a well-established technique in itself, but in most studies it is used in combination with apatite fission track analysis. Many fission track research groups now routinely apply (U-Th)/He dating in parallel with fission track analysis. An introduction to (U-Th)/He dating is, therefore, included here.

Fission Tracks

Fission tracks are linear damage trails in the crystal lattice. Natural fission tracks in geological samples are formed almost exclusively by the spontaneous fission of ^{238}U. Other naturally occurring isotopes, such as ^{235}U and ^{232}Th, also fission spontaneously, but the respective isotopes have such low fission decay rates that it is generally assumed that all spontaneous fission tracks in naturally occurring crystals are derived from ^{238}U. The frequency of fission events is low compared to α-particle decay events, about 1 fission event for every 2×10^6 α-particle decay events.

During spontaneous fission an unstable nucleus splits into two highly charged daughter nuclides (Figure 1). The two fission fragments are propelled in opposite directions, at random orientation with respect to the crystal lattice. The passage of the positively charged fission fragments through the host mineral damages the crystal lattice by ionization or electron stripping, causing electrostatic displacement. The end result is a cylindrical zone of atomic disorder with a diameter of a few nanometers – known as a fission track. Detailed information on the length of fission tracks is available for apatite only. Newly created apatite fission tracks have a length of ~16.3 ± 0.5 μm. Fission tracks can be observed directly through transmission electron microscopy, but with

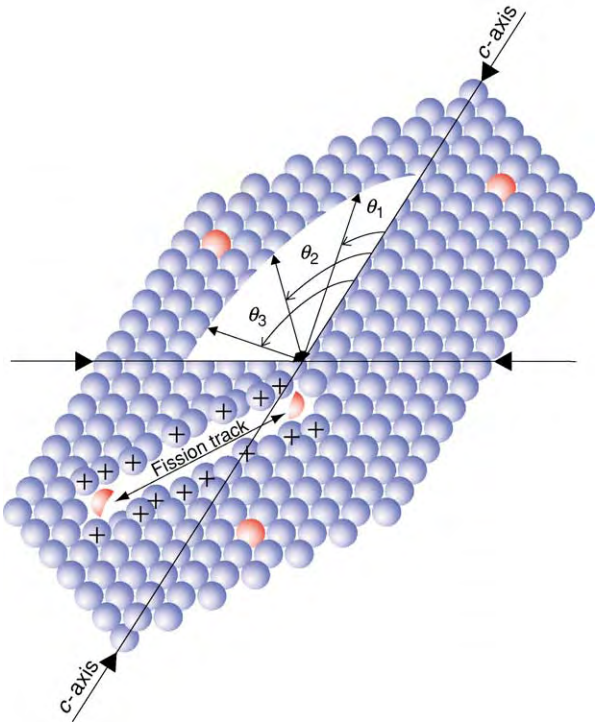


Figure 1 Spontaneous fission of ^{238}U (red spheres) produces two highly charged fission fragments (red half spheres) that recoil as a result of Coulomb repulsion. They interact with other atoms in the crystal lattice by electron stripping or ionization. This leads to further deformation of the crystal lattice as the ionized lattice atoms (blue spheres with plus sign) repel each other. After the fission fragments come to rest, a damage trail ('fission track') is left, which can be observed with an optical microscope after chemical etching. In apatite, the fission track annealing rate is higher for tracks at greater angle (θ) to the crystallographic *c* axis. Therefore, tracks perpendicular to the *c* axis are on average shorter than tracks that are parallel to the *c* axis.

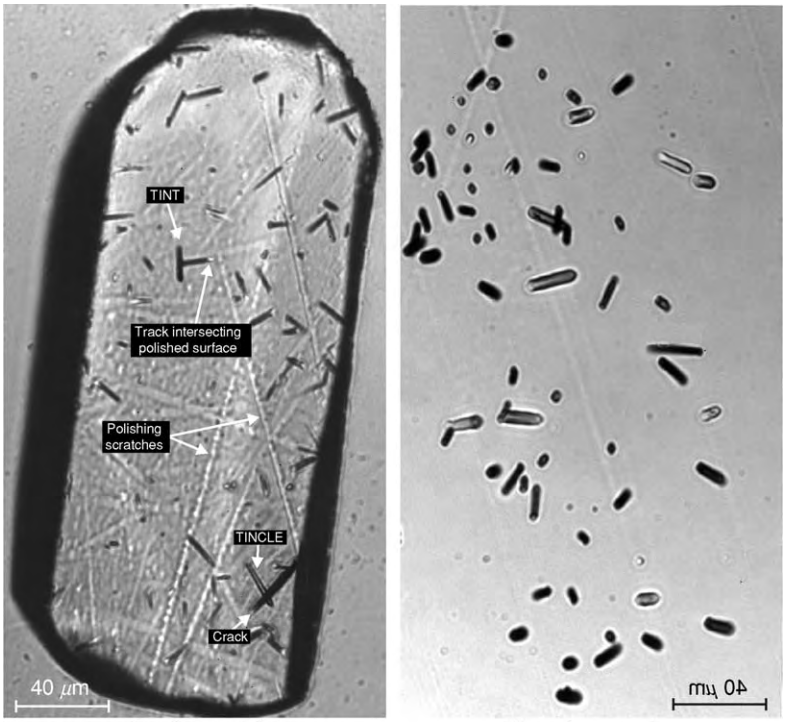


Figure 2 Fission tracks in apatite (left) resulting from the spontaneous fission of ^{238}U and induced fission tracks in mica (right) produced by irradiation in a nuclear reactor. Fission tracks in the mica outline a mirror image of the polished apatite crystal with which it was in close contact during irradiation. Fission tracks are revealed by chemical etching with HNO_3 (apatite) and HF (mica). Only fission tracks that intersect the polished surface, cracks (track in cleavage, TINCLE) or other tracks (track in track, TINT) can be reached and enlarged by the etchant.

an optical microscope they can only be observed after revelation by chemical etching. Seen through an optical microscope, chemically etched fission tracks appear as randomly oriented cigar-shaped features (Figure 2).

Fission Track Annealing

Laboratory experiments show that residence at elevated temperatures induces shortening of fission tracks. This process of track shortening by solid state diffusion is called fission track annealing. The rate of the annealing process is dependent on mineral properties and thermal history. Pressure and stress dependency have been suggested, but the evidence is ambiguous and highly controversial.

When a sample cools below the total annealing temperature, it enters the Partial Annealing Zone (PAZ; APAZ in the case of apatite, ZPAZ for zircon). As the sample cools within the PAZ, tracks shorten by lesser amounts until becoming relatively stable at low (<60°C) temperatures. Fission track ages are based on counts of tracks in a polished cross-section through a crystal. Shorter tracks have a smaller probability of intersecting the polished surface, and track length shortening in the PAZ consequently also leads to an apparent age reduction. This produces a characteristic pattern of fission track ages and mean track lengths within a PAZ (Figure 3). Such a pattern may be (partly) preserved in the case of very rapid cooling. This is referred to as a 'fossil PAZ' (Figure 4). The APAZ is sometimes referred to simply as the temperature interval between 60°C and 120°C, but this is an oversimplification because it neglects the impact of variations in chemical composition and cooling rate on the rate of the annealing process. The ZPAZ is more loosely constrained than the APAZ and probably lies somewhere between ~200°C and ~350°C.

A well-known problem with the interpretation of fission track data is that annealing can take place even below the temperatures normally associated with the PAZ. Apatite fission tracks that are formed in a nuclear reactor by irradiation with slow neutrons ('induced tracks') have an initial track length (l_0) immediately after irradiation of ~16.3 ± 0.5 μm. However, natural fission tracks ('spontaneous tracks') in apatite crystals separated from rapidly cooled volcanic rocks have mean track lengths in the order of 14.5–15 μm. Since this track length reduction can only have taken place at surface temperatures after eruption, it is

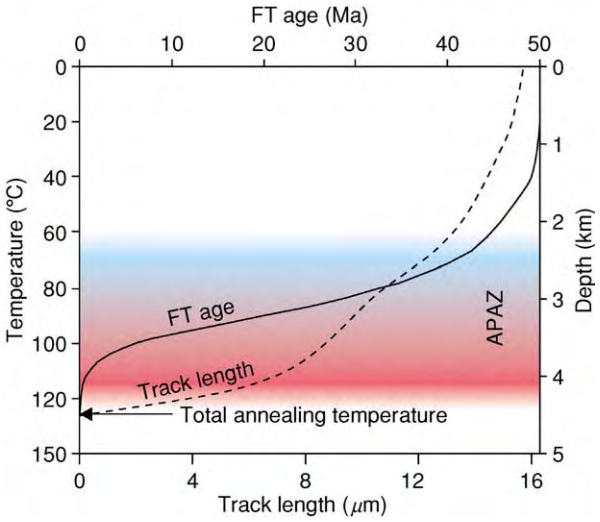


Figure 3 Pattern of apatite fission track ages and mean track lengths in a borehole. Example based on 5 km of rapid denudation at 50 Ma (30°C/km geothermal gradient) followed by 50 Ma of stable thermal conditions (no denudation). Samples close to the surface (<2 km depth) have fission track ages that approximate the timing of rapid denudation and have very long mean track lengths. Samples inside the apatite partial annealing zone (APAZ) have much younger fission track ages and shorter mean track lengths. Above the total annealing temperature fission tracks will be erased almost immediately after formation.

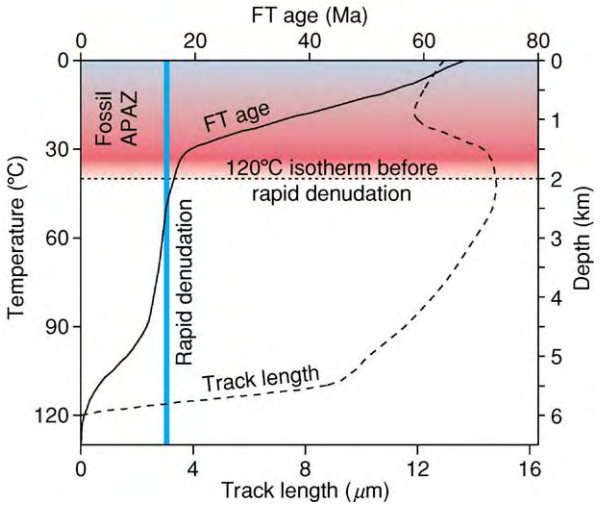


Figure 4 Pattern of apatite fission track ages and mean track lengths in a borehole after a late phase of rapid cooling. Example based on stable thermal conditions from 80 Ma to 15 Ma and a geothermal gradient of 20°C/km. At 15 Ma, a phase of rapid denudation removes 4 km of overburden. As a result of this rapid cooling, the pattern of fission track ages and mean track lengths in the APAZ as it existed before the onset of rapid cooling ('fossil APAZ') is largely preserved. The break in slope in the fission track age profile coincides with the onset of rapid denudation.

apparent that low-rate annealing in apatite takes place even at these low temperatures. However, it is very difficult to determine the rate at which fission track annealing takes place at such low temperatures. The annealing rate becomes very low and extrapolation of results from lab experiments to the geological timescale would introduce large uncertainties. Calibration with data from boreholes is also difficult, because the low temperature history (<60°C) of the borehole will almost never be accurately known in detail.

Fission track annealing behaviour in apatite is correlated with its unit-cell parameters and thus with crystallographic structure and chemical composition. The apatite group, with its simplified molecular formula $Ca_5(PO_4)_3F$, has a wide range of possible chemical compositions. Of all possible substitutions, that of chlorine (Cl) in place of fluorine (F) has the largest impact on the fission track annealing process. Apatite crystals with high chlorine contents are very resistant to annealing. Fission tracks in such crystals will survive higher temperatures than fission tracks in apatite crystals with lower chlorine contents. Even though chlorine content has a dominant control on annealing behaviour, significant variation in unit-cell parameters, and therefore in annealing behaviour, exists between chlorine-poor apatites as a result of a variety of possible substitutions, other than that of chlorine in place of fluorine. Because chlorine-poor apatites are the most common in nature, relying solely on chlorine content as a proxy for annealing behaviour does not account for the variation in annealing behaviour between most natural samples. The most obvious way to assess variation in chemical composition between apatite samples is to analyse every individual grain in which fission tracks are counted and measured, for example with a microprobe. Alternatively, etch pit size (Figure 5) the size of the intersection of a chemically etched fission track with the polished apatite surface, can be used as a measure for solubility and thereby as a proxy for bulk chemical composition. This method has been applied successfully and has the advantage that etch pit size is easily and inexpensively measured under the microscope at the same time that the actual fission track counting and measuring is done. No matter which approach is taken, it is essential to account for variations in chemical composition, because they exert such a strong influence on the annealing behaviour. Moreover, variation in chemical composition does not only occur between apatites from different rock samples, but also between apatite grains from the same rock sample. This is not only the case for (meta-)sedimentary rock samples, but also for igneous rocks.

Constraints on how the annealing process affects fission tracks can be combined into so-called annealing

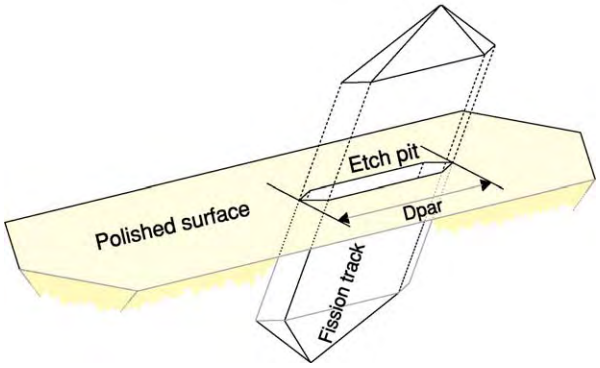


Figure 5 Chemically etched fission track intersecting a polished crystal surface. An etch pit is the intersection of the track with this surface. Dpar is the maximum diameter of the etch pit. Please note that the size of the etch pit in this example is exaggerated; real etch pits are orders of magnitude smaller relative to the grain sizes of natural crystals. Dpar can be used as a proxy for annealing behaviour.

models. Annealing models are mathematical approximations of how the annealing process relates to temperature history and mineral properties. Modern annealing models take into account variation in chemical composition and other parameters. They are calibrated with lab experiments and data from vertical arrays of samples in drill holes of which the thermal history is well known from independent data.

Fission Track Ages

A fission track age does not indicate the timing of cooling through a specific temperature boundary, but instead represents an integrated signal of the thermal history of a sample. Fission track ages, therefore, usually do not correspond directly to geological events. Additional fission track length data is needed to reconstruct a thermal history. Only in the case of very rapid cooling, may fission track ages correspond to a specific geological event.

The amount of spontaneous fission tracks in a crystal is dependent on its thermal history, its annealing characteristics, and the concentration of ^{238}U, the parent isotope. With the External Detector Method (EDM), which is the preferred analytical method in most fission track studies, it is possible to determine the ^{238}U concentration of every individual crystal in which tracks are counted. This is achieved by irradiating polished crystals that are mounted in epoxy (Figure 6A,B), together with a piece of low-uranium mica (the 'external detector') in direct contact with the polished surface. Spontaneous fission tracks in the crystals are revealed by chemical etching prior to irradiation (Figure 6C). Irradiation will induce the fission of ^{235}U and thereby produce new fission fragments in the crystals. A proportion of them will be emitted from the etched crystals into the covering mica detector (Figure 6D). After irradiation, the mica is etched (Figure 6E). Because ^{235}U/^{238}U has a constant ratio in normal rocks, the amount of fission tracks in the part of the mica that was in contact with a crystal ('induced tracks') can be used as a measure for the ^{238}U concentration in that particular crystal.

To monitor the neutron fluence during irradiation, samples are irradiated together with pieces of uranium-bearing glass that are also covered with mica detectors. Irradiation of the glass will produce fission tracks in the mica in the same way as natural crystals do and the amount of tracks per area (fission track density) will be proportional to the neutron fluence. After determining the density of spontaneous fission tracks in the crystals, of induced tracks in the matching areas in the mica and of the tracks in the mica covering the uranium-bearing glass, single crystal ages can be calculated. From the single crystal ages, a combined sample age can then be determined.

Statistical tests can be performed on the distribution of single crystal ages. Significant spread of single crystal ages can be a result of variation in annealing characteristics between crystals from the same sample, or, in the case of a (meta-)sedimentary sample, mixing of crystals from different source areas with different thermal histories. To test whether or not single crystal ages are from the same statistical population, a χ^2-test can be performed. Significant spread of single crystal fission track ages can also easily be detected in a radial plot (Figure 7). Zircon fission track dating is regularly used in provenance studies and the distribution of single crystal fission track ages is then used to tie age-populations to different source areas. Because of their greater resistance to annealing, zircon fission tracks will be preserved up to much higher temperatures than apatite fission tracks during burial in a sedimentary basin. Zircon crystals will, therefore, retain the source area signatures much better than apatite crystals and are also more resistant to abrasion during physical transport.

Fission track ages are normally calibrated against one or more age standards with the Zeta (ζ) method. Standards of known age are irradiated and analysed in the same manner as regular samples in order to establish a calibration factor ζ. Every analyst has to establish a personal ζ value in order to correct for observational bias. When a constant personal ζ value is obtained, every unknown fission track grain age can be calculated. The most commonly used apatite age standard is from Durango, Mexico, and has a ^{40}Ar/^{39}Ar age of 31.4 $\pm$ 0.5 Ma.

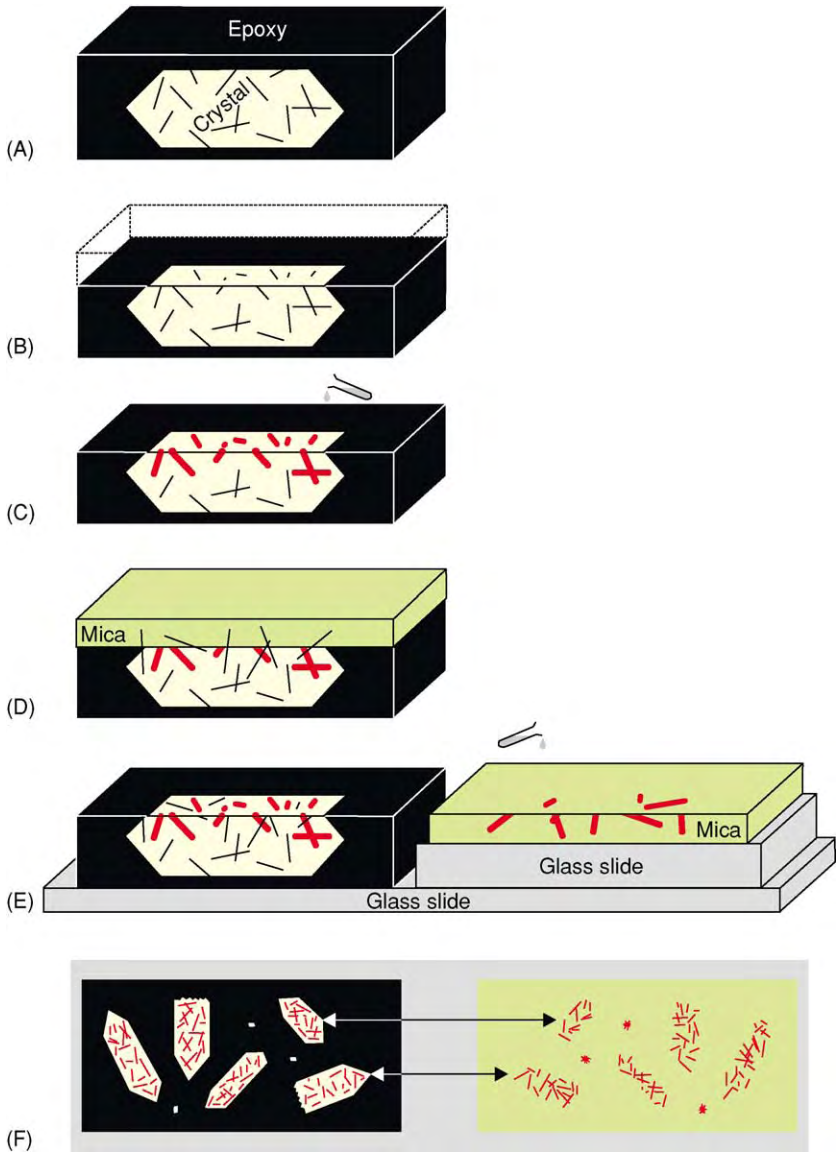


Figure 6 Schematic overview of the laboratory procedures involved in the External Detector Method. (A) Crystals with latent tracks are mounted in epoxy. These latent tracks cannot be observed with an optical microscope. (B) The mount is ground and polished until a sufficiently large internal surface in the crystal is exposed. (C) Chemical etching. Tracks that intersect the polished surface, or that can be reached by the etchant through other tracks or cracks, become enlarged (red) and can now be observed with an optical microscope. (D) External Detector Mica is mounted in close contact with the polished mount. A stack of mounts is placed in a cylinder and irradiated. A proportion of the fission fragments that are created inside the crystal by irradiation with neutrons, will be ejected into the mica. On top and bottom of the stack is a piece of uranium bearing glass, also covered with mica, to monitor the neutron fluence during irradiation. (E) After irradiation the mica is removed and chemically etched with HF. (E,F) Mounts and micas are glued onto a glass slide as mirror images. Track densities in crystals and the matching areas in the mica can now be determined under the microscope. (F) Plan view of several mounted crystals. Induced tracks in mica define outline of the crystals they were in contact with during irradiation.

Fission Track Length

Because detailed information on the length of fission tracks is available for apatite only, this section is solely applicable to apatite fission tracks.

Fission track length measurements are made on polished and chemically etched crystals (Figure 6A,B,C). To make sure that the full length of a track is measured, only tracks that are oriented parallel to the polished surface should be measured ('horizontal confined tracks'). This can easily be verified by focusing and defocusing the microscope; only tracks parallel to the polished surface will come into focus over their entire length at once.

The apatite crystal structure is anisotropic and so is the effect of the annealing process. The annealing rate

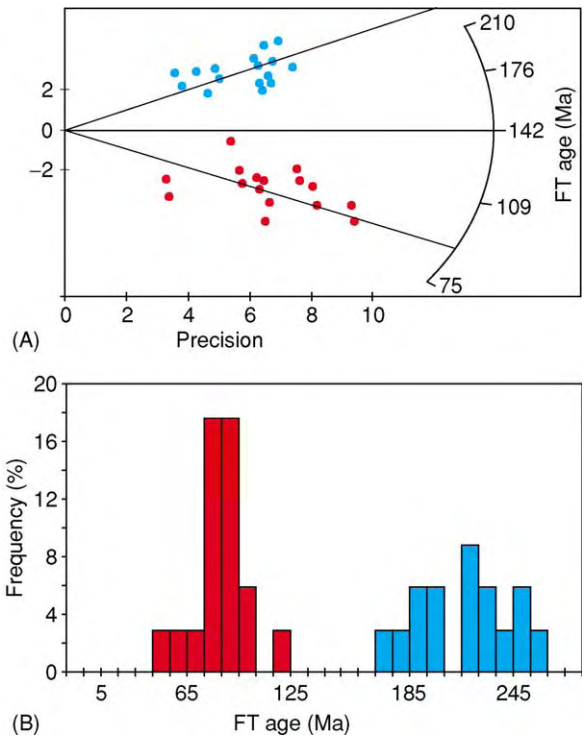


Figure 7 (A) Radial plot for a sample with two clearly defined age populations (red and blue). More precise fission track ages plot further from the origin of the *x* axis (precision). The *y* axis gives the fission track age and its error. The two populations have ages of ~90 Ma (red) and ~225 Ma (blue) and a combined age of 142 Ma. (B) Histogram of the individual crystal ages in the radial plot.

of fission tracks depends on the orientation of the fission tracks with respect to the crystallographic orientation of the apatite crystal (Figure 1). Tracks perpendicular to the *c*-axis anneal slightly faster than those parallel to it. For weakly annealed samples, this effect will be very small and usually it is ignored. For samples that experienced higher degrees of annealing, ignoring the orientation dependency of the track length will introduce some error. Anisotropy of annealing can, however, easily be taken into account by measuring θ, the angle of the fission track with the crystallographic c-axis, when the track length is measured.

Tracks that are revealed by etching because they intersect cracks (track-in-cleavage, TINCLE Figure 2), tend to be longer than tracks that are revealed because they intersect other tracks (track-in-track, TINT; Figure 2). Several explanations have been proposed, such as widening of cracks during polishing, increased etch rates along cracks, and the smaller likelihood for shorter tracks to intersect a crack. Because of their bias towards longer track lenghts, TINCLEs tend to conceal the anisotropy of annealing. It is recommended to measure only TINTs.

Thermal History Modelling

Fission tracks are produced continuously throughout the thermal history of a sample. Older fission tracks will, therefore, always have experienced a longer and older part of a sample's thermal history than younger fission tracks. This results in variation of fission track lengths within a sample and within single crystals. In the case of a purely cooling history, the older tracks will have experienced higher temperatures than younger tracks and so the older tracks will be annealed more, and thus will be shorter, than younger tracks. Track length reduction within the PAZ results in a mean length shorter than the initial length (l_0) and a skewed track length distribution. Different cooling histories and their resulting apatite fission track age and track length distributions are displayed in (Figure 8).

Thermal history modelling makes use of annealing models to calculate the apatite fission track age and track length distribution that would result from a particular thermal history. This synthetic modelling result can be compared to the fission track age and track length distribution obtained from a sample and the degree of fit can be assessed. By doing this for many different thermal histories, a range of thermal histories with a good degree of fit can be obtained. A geological interpretation can then be based on these thermal histories.

Independent geological observations and also additional age data such as those from (U-Th)/He dating can be used to constrain the range of thermal histories that a modelling program has to go through. For example, a time–temperature point with surface temperature at 100 Ma can be used as a constraint when modelling fission track data from a sedimentary sample of 100 Ma old (or data from a sample of the basement just below the sediment), because at the time of deposition the sediment (and the basement immediately below) must have been experiencing surface temperatures.

Thermal history modelling is a popular tool to retrieve a time–temperature path from apatite fission track data. However, in the absence of independent geological constraints, it cannot constrain anything other than cooling. The fission track age and fission track length distribution resulting from a thermal history that includes a reheating event will be virtually identical to that resulting from the same thermal history without the reheating event (Figure 9). Fission tracks essentially record cooling, and even in a thermal history that has included reheating, most of the record will come from the cooling segments.

A very common characteristic of thermal histories obtained from modelling of fission track data is that they include a late cooling event. The annealing

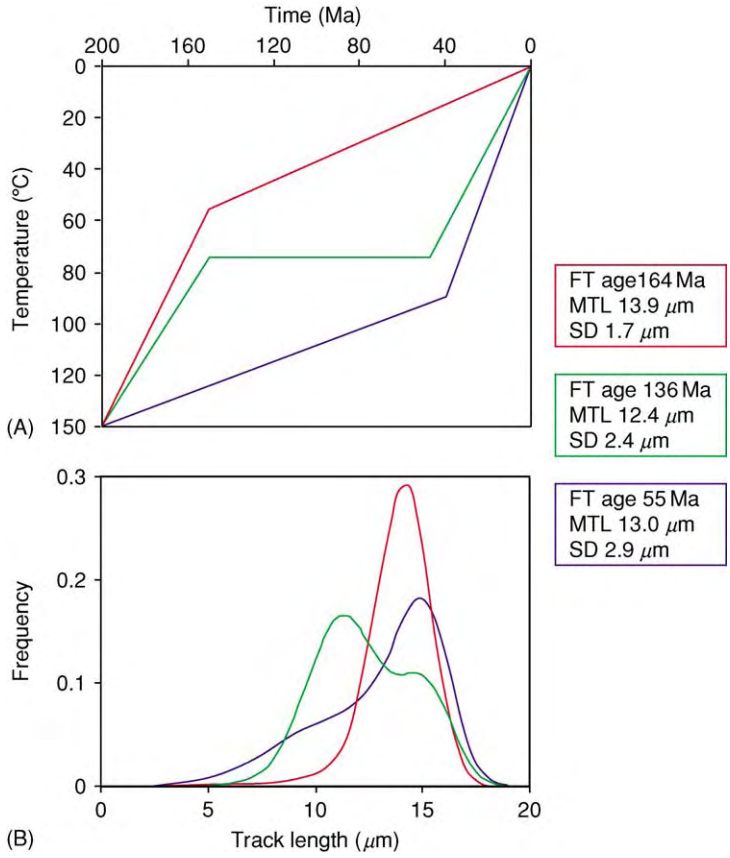


Figure 8 (A) Three cooling histories and (B) the resulting apatite fission track ages and track length distributions. Rapid early cooling (red) produces a narrow track length distribution with a long mean track length and an old age. Rapid initial cooling followed by a long period of time in the APAZ (green) results in a skewed track length distribution and a younger apatite fission track age. Slow cooling in the APAZ followed by rapid late cooling (blue) produces a negatively skewed track length distribution and a very young apatite fission track age. FT age, fission track age. MTL, mean track length. SD, standard deviation of the mean track length.

process is poorly understood for temperatures below ~60°C and, in general, annealing at these low temperatures is underestimated. Modelling of fission track data, therefore, often results in thermal histories that include a rapid late cooling event, which in many cases is merely a modelling artefact. Therefore, one should always be very critical about any geological interpretations based on this part of the modelled thermal history. The (U-Th)/He method gives much better time–temperature constraints for such low temperatures.

(U-Th)/He Dating

(U-Th)/He dating, like fission track analysis, provides information on a sample's low-temperature thermal history and not on its original, high-temperature igneous or metamorphic history. As is the case for fission track analysis, the (U-Th)/He technique is best established for apatite. (U-Th)/He dating of zircon and titanite has been explored, but at present this application is still in its infancy. At least for apatite, zircon, and titanite, the (U-Th)/He method is sensitive to lower temperatures than the fission track technique.

In nearly all minerals radiogenic helium is predominantly derived from the decay of uranium and thorium. Accumulation of α-particles (^{4}He nuclei) starts upon cooling into the Helium Partial Retention Zone (HePRZ). The HePRZ concept is similar to the Partial Annealing Zone concept of fission track analysis. At temperatures higher than that of the HePRZ, helium diffuses out of grains almost immediately after production. This means that there can be no accumulation of helium and the age will thus remain 0. At temperatures below that of the HePRZ, diffusion ceases and (U-Th)/He ages track calendar time. The nominal helium closure temperature (T_c) is dependent on mineral properties, cooling rate, and grain size. For apatite crystals in the size range of ~140–180 μm, T_c will be ~70°C at a cooling rate of 10°C/Ma. For smaller grain sizes and at lower cooling rates, it may be as low as ~50°C. Unlike fission track data,

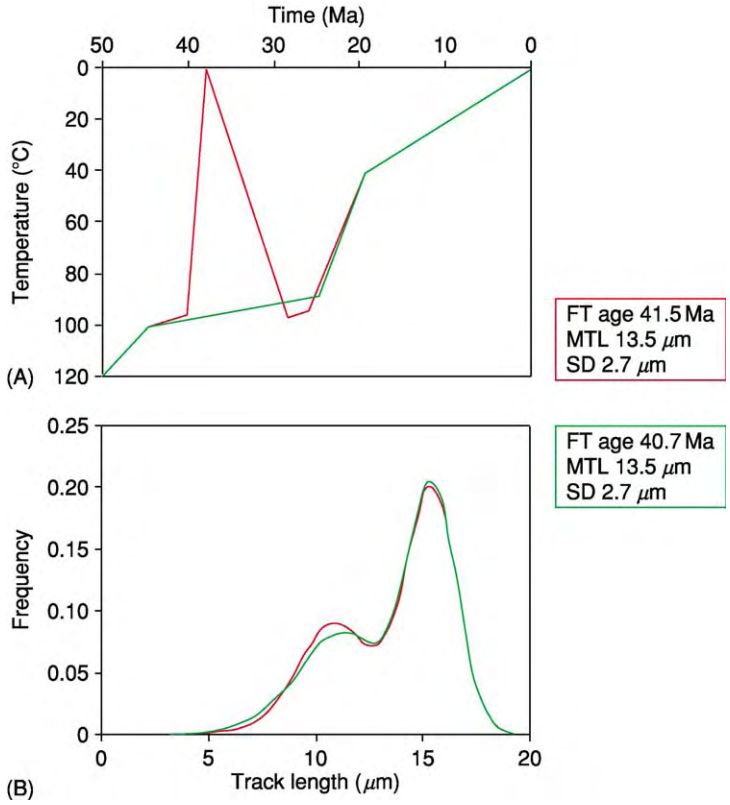


Figure 9 (A) Two thermal histories and (B) the resulting apatite fission track ages and track length distributions. Despite the difference in thermal history before the onset of final cooling, both scenarios produce essentially the same apatite fission track age and track length distribution. The difference in age and track length distribution is well within the uncertainty normally associated with fission track data.

apatite (U-Th)/He data do not seem to be influenced by chemical composition.

Because of the high kinetic energy with which α-particles are emitted from their parent nuclides, they may be lost from grains as a result of α-ejection. For apatite, the so-called α-stopping distance is ~19–23 μm, depending on the parent nuclide. Conditional on the distance of the parent nuclide from the physical grain boundary, α-particles may be either ejected from, or retained within grains (Figure 10A).

Because α-particles are emitted in random direction from their parent nuclei, they have a 50% chance to be ejected if their parent nuclei are located on the physical grain boundary itself. α-particles emitted from parent nuclei at more than the α-stopping distance from the grain boundary will never be ejected by α-emission. In principle, α-particles can also be implanted into grains when they are ejected from neighbouring grains. α-implantation can be ignored in most cases however, because the concentration of parent nuclei normally is much higher in minerals that are used for dating compared to the concentration in the host rock they were separated from. α-ejection must be corrected for by the so-called α-emission correction to obtain a geologically meaningful 'α-ejection corrected' (U-Th)/He age. The α-emission correction is dependent on the mineral and the appropriate α-stopping distances, grain size, and grain geometry. Smaller grains require bigger corrections because a larger percentage of parent nuclei will be within α-stopping distance from the grain boundary. A complicating factor for making the appropriate α-emission correction is that helium loss by ejection is interwoven with diffusional helium loss. The helium diffusion domain in some minerals (e.g., apatite) coincides with the physical grain. Within the HePRZ, helium will, therefore, be lost by diffusion at the grain boundary. Diffusion will tend to smooth the α-retention profile resulting from α-ejection, which is illustrated in a grain cross-section (Figure 10B). Because they are interlinked, the diffusion process and α-ejection cannot be accurately modelled separately and instead must be incorporated simultaneously in a numerical model. The diffusion process and its effect on the α-retention profile becomes more important with a longer residence time in the HePRZ, and application of such a model, therefore, is particularly important when working

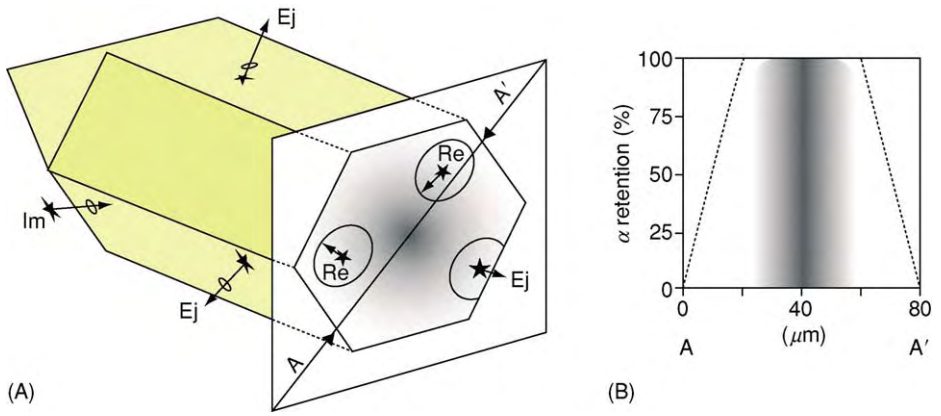


Figure 10 (A) Three possible outcomes of α emission in an apatite crystal: (Im) implantation, (Re) retention, and (Ej) ejection of α particles. Radius of circles is equivalent to the α stopping distance. (B) Diagram showing how α retention changes from rim to core to rim along a cross section perpendicular to the *c* axis of the crystal.

with samples that spend considerable time in the HePRZ.

Successful application of (U-Th)/He dating is very much dependent on the quality of the material being analysed. Several criteria must be satisfied before mineral grains are studied by (U-Th)/He dating. This somewhat restricts the range of rock types to which this technique can be applied. Crystals must be sufficiently large to keep the α-emission correction and its potential error within reasonable bounds. The minimum diameter of apatites that can be accurately dated is about ~75 μm. Helium diffusion and α-emission correction models are based on euhedral crystal morphology (or, for example, a finite cylinder approximation) and a homogeneous distribution of parent nuclei. Only intact, minimally abraded crystals lacking uranium and thorium zonation should therefore be used. Zonation of uranium and thorium is fairly common in apatite and ignoring this can introduce large errors. Crystals must be free of cracks and inclusions and should, therefore, be thoroughly inspected under the microscope before analysis. Small inclusions may be overlooked however, and uranium and thorium zonation is not easily detected in the grains that are being dated. Even very carefully selected grains may thus not fulfil all the selection criteria. (U-Th)/He analysis should, therefore, always be done in duplicate and preferably triplicate on separate batches of crystals from the same sample, so that reproducibility of the data can be verified. Although few rock types are explicitly excluded from (U-Th)/He dating as a result of the above-mentioned criteria, it is obvious that, for example, sedimentary rocks are less likely to yield unabraded crystals and that volcanic rocks may not contain sufficiently large crystals.

The analytical procedure for (U-Th)/He analysis consists of two stages. First, helium is extracted from crystals in vacuum by heating in a furnace or with a laser and then measured on a noble gas mass spectrometer. Subsequently, the sample is removed from the vacuum system, dissolved in acid and the uranium and thorium concentrations are determined, usually by Inductively Coupled Plasma Mass Spectrometry (ICP-MS). The entire procedure is then repeated for the sample duplicates. Depending on the expected helium, uranium, and thorium concentrations, a batch of apatites typically consists of ~1–20 crystals. When the helium, uranium, and thorium concentrations have been determined, a 'raw' (U-Th)/He age can be calculated. This must then be corrected for loss of α-particles to obtain a geologically meaningful age. A standard of known age should always be included in a series of age determinations to verify data quality. For apatite (U-Th)/He analysis, the Durango apatite is the most common monitor.

Applications of Fission Track Analysis and (U-Th)/He Dating

Fission track analysis can be applied to a large variety of geological problems and settings. In basins analysis, for example, apatite fission track analysis can help to estimate maximum palaeotemperatures and the postdepositional denudation history of the sedimentary basin, and zircon fission track data can be used to investigate sediment provenance. In orogenic belts (*see* **Tectonics:** Mountain Building and Orogeny), fission track analysis can provide estimates on erosion and denudation and, therefore, is important for mass balance studies. Non-orogenic settings, such as passive continental margins, are probably best suited to the capabilities of fission track analysis, because this type of setting in many cases will be dominated by cooling and denudation. As a conventional dating

method, fission track analysis can be applied to date, for example, volcanic glass or tuff. Because of its unique ability to constrain temperature histories in the uppermost part of the crust (upper ~3 km, depending on the geothermal gradient), apatite (U-Th)/He dating can be used to constrain the development of topography. It is, therefore, a valuable tool in geomorphological studies (*see* **Geomorphology**). (U-Th)/He dating also is an important technique to constrain the last part of thermal histories obtained from thermal history modelling of fission track data.

List of Units

μm	One millionth of a metre
$^{40}Ar/^{39}Ar$	Argon – argon dating method
Ca	Calcium
Cl	Chlorine
F	Fluorine
Fission	Splitting of an atomic nucleus
^{4}He	Helium – mass 4
HNO_3	Nitric acid
l_0	Length of a fission track immediately after formation
O	Oxygen
P	Phosphorus
T_c	Closure temperature
^{232}Th	Thorium – mass 232
^{235}U	Uranium – mass 235
^{238}U	Uranium – mass 238

Glossary

***α*-ejection** Loss of α-particles from a crystal as a result of α-emission.

***α*-emission** Emission of α-particles from parent nuclides.

***α*-implantation** Gain of α-particles by a crystal resulting from α-ejection from neighbouring crystals.

***α*-particles** ^{4}He nuclei.

***α*-retention** Retention of α-particles in a crystal after α-emission.

APAZ Apatite Partial Annealing Zone.

EDM External Detector Method.

HePRZ Helium Partial Retention Zone.

HF Hydrofluoric acid.

ICP-MS Inductively Coupled Plasma Mass Spectrometry.

Induced tracks Fission tracks produced by irradiation in a nuclear reactor.

PAZ Partial Annealing Zone.

Spontaneous tracks Natural fission tracks, formed by spontaneous fission.

TINCLE Track-in-cleavage, fission track intersecting a crack.

TINT Track-in-track, fission track intersected by another track.

ZPAZ Zircon Partial Annealing Zone.

See Also

Analytical Methods: Geochronological Techniques; Mineral Analysis. **Geomorphology**. **Minerals:** Micas; Zircons. **Petroleum Geology:** Exploration. **Tectonics:** Mountain Building and Orogeny. **Tektites**.

Further Reading

Barbarand J, Carter A, Wood I, and Hurford T (2003) Compositional and structural control of fission track annealing in apatite. *Chemical Geology* 198: 107 137.

Deer WA, Howie RA, and Zussman J (2004) *Rock Forming Minerals*, vol. 4B, *Silica Minerals, Feldspathoids and Zeolites*. London: The Geological Society.

Donelick RA, Ketcham RA, and Carlson WD (1999) Vari ability of apatite fission track annealing kinetics: II. Crys tallographic orientation effects. *American Mineralogist* 84: 1224 1234.

Farley KA (2002) (U Th)/He Dating: Techniques, Calibra tions and Applications. In: Porcelli DP, Ballentine CJ, and Wieler R (eds.) *Noble gases in geochemistry and cosmo chemistry,* pp. 819 843. Washington: Mineralogical Society of America and Geochemical Society.

Fleischer RL, Price PB, and Walker RM (1975) *Nuclear tracks in solids.* Berkeley: University of California Press.

Gallagher K, Brown RW, and Johnson C (1998) Fission track analysis and its applications to geological problems. *Annual Review of Earth and Planetary Sciences* 26: 519 572.

Gleadow AJW and Brown RW (2000) Fission track thermo chronology and the long term denudational response to tectonics. In: Summerfield MA (ed.) *Geomorphology and Global Tectonics,* pp. 57 75. Chichester: John Wiley and Sons Ltd.

Green PF, Duddy IR, and Hegarty KA (2002) Quantifying exhumation from apatite fission track analysis and vitri nite reflectance data: precision, accuracy and latest results from the Atlantic margin of NW Europe. In: Doré AG, Cartwright JA, Stoker MS, Turner JP, and White N (eds.) *Exhumation of the North Atlantic Margin: Timing, mech anisms and implications for petroleum exploration.* London: Geological Society of London.

Palmer DC (1994) Stuffed derivatives of the silica poly morphs. *Reviews in Mineralogy* 29: 83 118.

Van den Haute P and De Corte F (1998) *Advances in Fission Track Geochronology.* Dordrecht: Kluwer Academic Publishers.

Wagner G and Van den Haute P (1992) *Fission Track Dating.* Dordrecht: Kluwer Academic Publishers.

Zentilli M and Reynolds PH (1992) *Low temperature ther mochronology; short course handbook.* Mineralogical Association of Canada: Toronto.

Geochemical Analysis (Including X-ray)

R H Worden, University of Liverpool, Liverpool, UK

Introduction

Geochemistry is the study of the occurrence and distribution of elements, isotopes, minerals, and compounds in the natural environment. Geochemical analysis, the measurement of the quantities of elements, isotopes, minerals, and compounds in a rock, natural liquid, or naturally occurring gas, is a colossal topic.

The applications of geochemistry stretch from the core to the outer atmosphere and beyond. The objects of study include minerals, metals, and salts, organic biomolecules, coal, bitumen, kerogen, and petroleum, atmospheric-, river-, ground-, and formation-water, and gases in the crust, sediments, and atmosphere. Analysis is also a huge topic, with subjects ranging from rocks, minerals, compounds, and species to isotopes, elements, and atoms.

The Role of Analysis: Hypotheses, Questions, Problems, and Theories

In the context of Earth sciences, geochemical analysis commonly implies quantitative work with a numerical output. Geochemical analysis is thus the quantitative examination of the composition of natural Earth materials. Most geochemical work is undertaken to address a hypothesis, answer a question, or solve a problem. Even routine environmental geochemical monitoring is done in order to address the question of the continued safety of natural materials over time. It is typically important to formulate a hypothesis, question, or problem carefully before embarking on an analytical programme. Such an approach is best since the outcome should be cost-effective while being statistically rigorous and ultimately defensible.

Producing Geochemical Data

When any analytical geochemical device is used to quantify the composition of a material, the detector output is usually in the form of some sort of electrical signal. The electrical signal must be converted into a meaningful geochemical parameter, such as a unit of concentration. This conversion is usually achieved via a calibration curve. For any technique, a series of previously characterized standards must be analysed, in exactly the same way as an unknown sample, and the output signals measured. In essence, the calibration is a plot of known concentration versus the output signal over a range of concentrations (Figure 1). The detector outputs from the standards must be converted into some sort of equation that permits the eventual back-conversion of output signals into concentrations for unknown samples. In the example in Figure 1, the output must be determined as a function of concentration (equation 1) and then inverted to allow concentration to be determined from the output signal for unknown samples (using equation 2).

The quality of the analytical output from a device is fundamentally a function of the quality of the calibration curve. No calibration is perfect, and the degree of imperfection can be described in terms of accuracy and precision. These discrete characteristics are best described for a dataset composed of repeated analyses of the same standard sample (Figure 2). When the results are widely dispersed about the known answer, the output is said to be imprecise. When the results are tightly clustered about the known answer, the output is said to be precise. If the results cluster around the correct figure, they are said to be accurate, whereas if they cluster about a figure other than the correct output, they are said to be inaccurate (Figure 2). Accuracy and precision are the combined result of the quality of calibration, the sensitivity of the device,

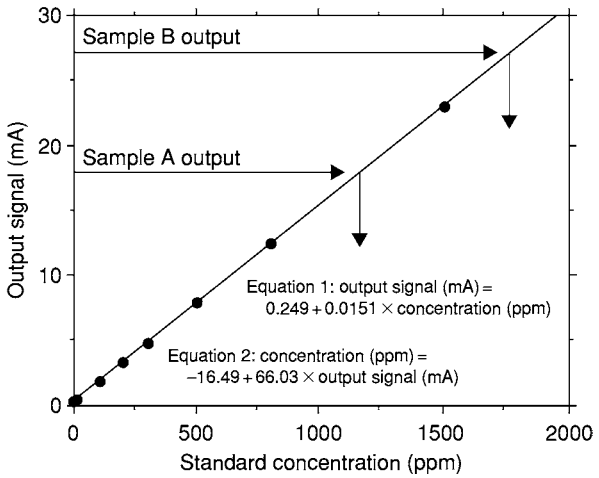


Figure 1 Example of a calibration curve using standards of known concentration and their output signal strengths. The small black dots represent the calibration. Equation 1 describes how concentration could be converted to signal strength. The more useful equation 2 describes how signal strength from an unknown sample can be converted into concentration. Unknown sample A yields a valid result since it falls within the calibration range. The analysis of sample B is not valid since the output signal falls outside the calibration range.

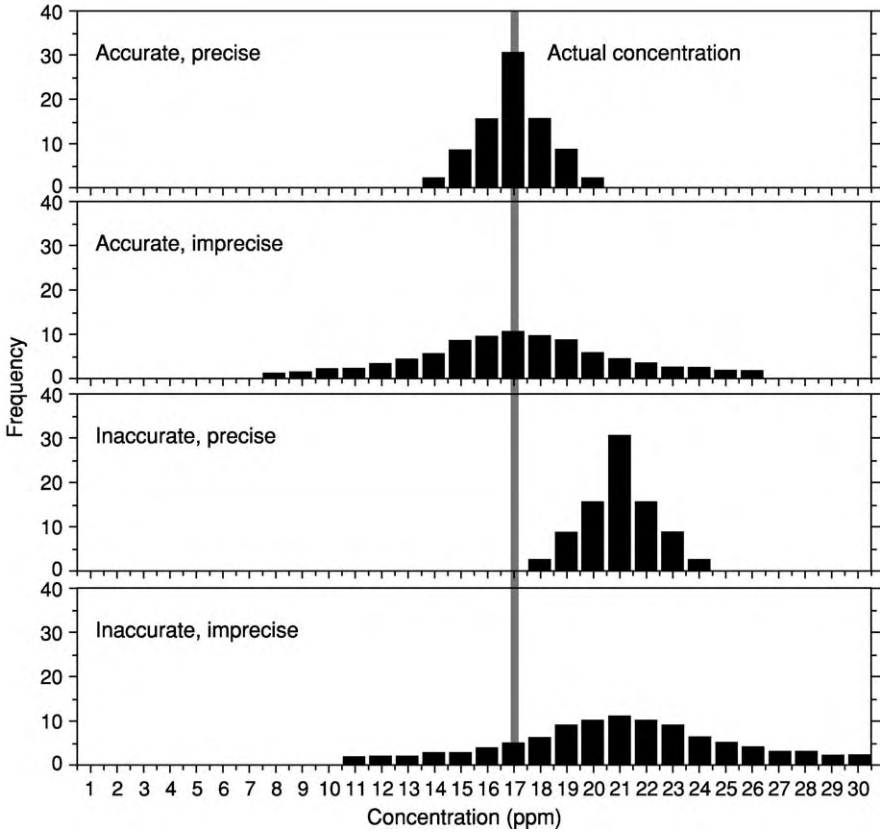


Figure 2 Representation of analytical precision and accuracy. Precision is a measure of how tightly the results cluster about the mean result. Accuracy measures the proximity of the mean result to the expected result. Accuracy can be improved by recalibrating the device.

the attainment of constant operating conditions, and the skill of the device operator.

Geochemical Analytical Protocol (Blanks, Standards, Repeats)

One of the most important aspects of any geochemical analytical protocol is to ascertain the credibility of the data. Every geochemical analysis has some degree of associated uncertainty (see Figure 2, where uncertainty is expressed in terms of precision). Every technique also has a minimum detection limit, below which the concentration cannot be determined. The theoretical lower detection limit is an intrinsic function of the technique, although the quality of the calibration and the cleanliness of the equipment used in sample and standard preparation also limit detection. It is seldom correct to report a figure of zero concentration. It is more correct to report the detection limit and state that the sample had a concentration below that value ('below detection' or 'none detected'). In any analytical run it is worthwhile running blanks to check the lower detection limit. It is also worth discretely running repeat samples to check the reproducibility of the analytical technique. Finally, it is worth running standards throughout an analytical run at regular intervals.

Developing and checking the calibration lines are the most crucial steps in geochemical analysis, and being aware of the upper and lower limits of the calibration is important. Many calibration lines do not vary linearly with concentration. Thus, if any geochemical analytical result falls outside the calibrated range of the detector and its output, then the result is theoretically invalid since it is not certain how to convert the instrumental signal (e.g. amps, volts etc) into geochemically significant units.

The Range of Geochemical Analytical Techniques

As stated earlier, geochemical analysis is a large topic, and many analytical techniques have been employed in geochemical research and studies over the last four or five decades. Techniques may be grouped according to the attribute being measured (Table 1). Many important techniques use X-rays in a variety of ways for analytical purposes. Other techniques use the optical effects of samples, typically taking advantage of electromagnetic radiation emitted or absorbed by the material at elevated temperatures.

Table 1 Summary of some of the main techniques of geochemical analysis and their sample types

Technique	*Output types*	*Abbreviation*	*Solid (rock, mineral, sediment, soil, filtrate)*	*Artificially dissolved solid*	*Water (natural solution)*	*Fluid organic*	*Solid organic*	*Gas*	*Fluids trapped in solids*
X ray techniques									
X ray analysis (electron microprobe)	Point mineral composition	MPA	Yes				Yes		Yes
X ray fluorescence	Rock composition	XRF	Yes			Yes	Yes		
X ray diffraction	Mineralogy	XRD	Yes				Yes		
Spectroscopic techniques									
Atomic absorption spectroscopy	Solution composition	AAS		Yes	Yes				Yes
Inductively coupled plasma optical emission spectroscopy	Solution composition	ICP OES		Yes	Yes				Yes
UV spectroscopy	Composition	UVS	Yes			Yes			Yes
Chromatography									
Gas chromatography	Gas composition	GC			Yes	Yes		Yes	Yes
Ion exchange chromatography	Aqueous liquid composition	IC		Yes	Yes				Yes
Mass spectroscopy									
Mass spectroscopy	Composition, isotope content	MS	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Isotope dilution mass spectroscopy	Isotope ratio	IRMS	Yes	Yes	Yes	Yes	Yes	Yes	
Inductively coupled plasma mass spectroscopy	Composition	ICP MS	Yes	Yes	Yes		Yes		Yes
Gas chromatography mass spectrocopy	Composition	GC MS			Yes	Yes		Yes	Yes
Thermal analysis techniques									
Pyrolysis	Elemental analysis (C,H,O,S,N), organic type					Yes	Yes		
Evolved water analysis, thermogravimetry	Clay content, clay mineralogy	EWA TG	Yes						
Fluid inclusion microthermometry	Temperature of trapping, composition	FI	Yes						Yes
Related techniques									
Wet chemistry techniques	Composition		Yes	Yes	Yes				
Microscopy (light and electron optics)	Mineralogy, fabric, crystallography	SEM,TEM	Yes				Yes		Yes

Other techniques use chromatography, the time taken for one substance to move through another or through a capillary under a given gradient. A wide family of geochemical analytical techniques use mass spectrometry to split propelled material, converted into charged particles, using electromagnets. Other techniques involve examining the products of heating Earth materials under controlled conditions and studying either the evolved fluids or the changes in the properties of the residual solids.

X-ray Techniques

Origin of X-rays

X-ray technologies have proved to be useful in geochemical analysis (Table 2). X-rays are part of the electromagnetic spectrum (Figure 3) and have wavelengths ranging between 0.01 nm and 10 nm (0.1–100 Å). They are waveforms that are part of a family that includes light, infrared, and radio waves. Since X-rays have no mass and no electrical charge, they are not influenced by electrical or magnetic fields and travel in straight lines. X-rays, like all parts of the electromagnetic spectrum, possess a dual character, being both particles and waves. The name that has been given to the small packets of energy with these characteristics is photon.

The simple model of the atom, proposed by Niels Bohr in 1915, is not completely correct, but it has many features that are approximately correct. The modern theory of the atom is called quantum mechanics; the Bohr model is an approximation to quantum mechanics that has the virtue of being much simpler than the full theory. In the Bohr model neutrons and protons occupy a dense central region (the nucleus), and electrons orbit the nucleus. The basic feature of quantum mechanics that is incorporated in the Bohr model is that the energies of the electrons in the Bohr atom are restricted to certain discrete values (the energy is quantized) – only certain electron orbits with certain radii are allowed.

X-rays are generated when free electrons from an electron gun give up some of their energy when they interact with the orbital electrons or the nucleus of an atom (Figure 4). The energy given up by the electron during this interaction reappears as emitted electromagnetic energy, known as X-radiation. Two different atomic processes can produce X-ray photons. One is called bremsstrahlung and the other is called electron-shell emission (Figure 5). Bremsstrahlung means 'braking rays'. When an electron approaches an atom, it is affected by the negative force from the electrons of the atom, and it may be slowed or completely stopped. The energy absorbed by the atom during the slowing of the electrons is excessive to the atom and will be radiated as X-radiation of equal energy to that absorbed. Bremsstrahlung X-rays tend to have a broad range of energies since the degree of slowing can be variable and materials composed of mixtures have atoms with different properties (Figure 6). Bremsstrahlung tend not to be used for geochemical analysis; that is the preserve of electron-shell emission.

Analysis of X-rays: Electron-Shell Emission

A common geochemical application of X-ray analysis is to direct a focussed electron beam at a polished rock or mineral surface and then collect and quantify the resulting secondary characteristic X-rays (Figure 7). The secondary X-rays help to reveal the elements present in that part of the sample that is directly under the electron beam. This technique is known as electron-beam microanalysis, or microprobe analysis, and gives spatially resolved major- and trace-element geochemical data from solid samples, including rocks, minerals, sediments, soils, and glass. Many ordinary electron microscopes are fitted with a secondary X-ray detector, making them suitable for geochemical analysis. All of these devices rely on electron optics, using electromagnetic lenses to focus and direct a stream of electrons, generated by an electron gun, onto a polished mineral or rock surface (Figure 7). The focused electron beam has a variable radius, but can typically be maintained at slightly greater than about 1 μm. The spatial resolution of a microprobe is actually somewhat greater than 1 μm. The impinging electron stream interacts with the polished surface and produces a wide range of signals, including secondary and backscattered electron and cathodoluminescence (light) as well as the secondary X-rays of concern here. There is an activation volume from which X-rays are generated, below the polished surface, which is several times larger than the primary beam. Samples must be highly polished (flat) to avoid scattering.

When a sample is bombarded by an electron beam, some electrons are knocked out of their quantum shells in a process called inner-shell ionization (Figure 5). Outer-shell electrons fall in to fill a vacancy in a process of self-neutralization. The shells are termed K, L, M, and N starting from the innermost most strongly bound shell.

Electrons moving from one shell to another produce characteristic X-rays. K-shell ionizations are commonly filled by electrons from the L shell (Kα radiation) or M shell (Kβ radiation). There are two Kα peaks (Kα1 and Kβ2) corresponding to two discrete states of the in-falling electron. When outer-shell electrons drop into inner shells, they emit a quantized

Table 2 X ray techniques commonly used in geochemical analysis

Technique	*Output 1*	*Output 2*	*Output 3*	*Sample type*	*Advantages*	*Disadvantages*
X ray analysis	Quantitative elemental composition of small volume (several cubic mm)	Major elements detected using energy dispersive spectrometer (SEM or microprobe)	Trace elements detected using wavelength dispersive spectrometer (microprobe)	Polished rock samples, or grains set in resin and polished	High spatial resolution; gives quantitative data with estimate of uncertainty; well established technology; wide range of elements	Sample preparation can be slow; errors can be large for some elements; problems of analysis statistics for heterogeneous minerals
X ray fluorescence spectroscopy	Quantitative major and trace elemental composition of a crushed homogenized sample			Crushed and homogenized sample then compressed or melted and quenched into small disks	Gives major and trace elements simultaneously; produces data for Al and Si in rocks as well as a range of metals; well established technology	Sample preparation can be slow; relies on sample type specific calibration curve; matrix effects can be large; instrumental set up can be arduous
X ray diffraction	Qualitative presence or absence of minerals in crushed sample	Relative proportion of minerals in a rock or sediment mixture for suitably prepared samples	Mineral composition		Fairly quick; well established technology; large database of minerals for comparison	Semi quantitative in most cases; difficult for minerals with solid solutions; difficult for poorly crystalline materials; difficult with very complex mixtures

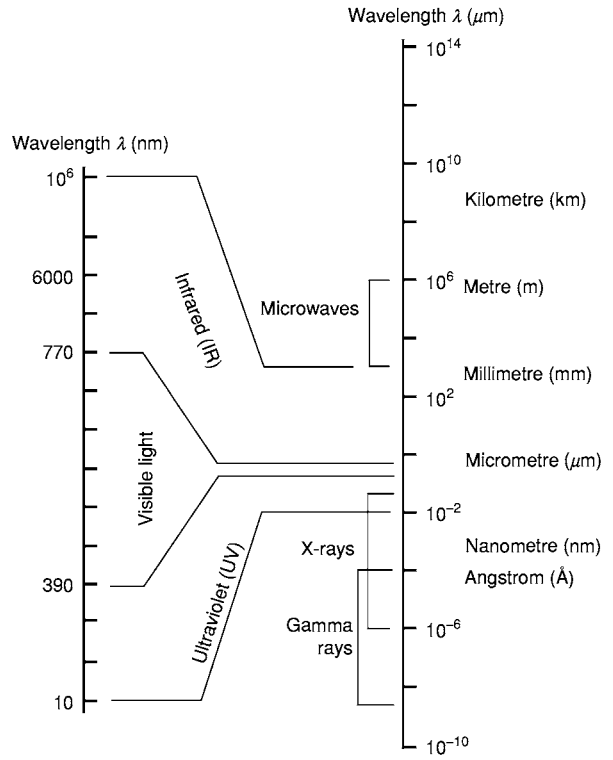


Figure 3 Representation of the electromagnetic spectrum with the various wavelengths employed in geochemical studies. Visible light is not discussed here but forms the basis of optical microscopy. Gamma rays are used to differentiate radioactive fission reactions and thus identify and quantify radioactive elements.

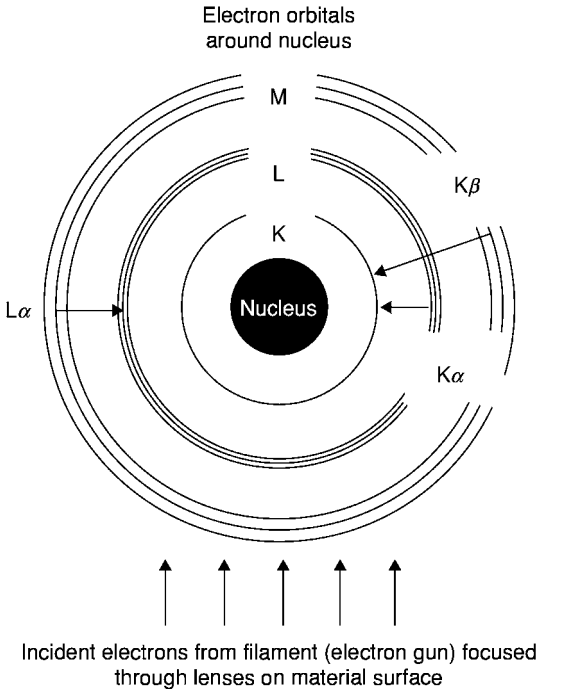


Figure 5 Incident electrons cause electrons in metal atoms to become excited and jump to outer orbitals. Kα and Kβ X rays (etc.) are emitted as electrons fall back to their original orbitals. The X ray energy is characteristic of the element and of the starting and final orbitals. Background 'white' radiation (bremsstrahlung) is due to collisions between incident and orbital electrons.

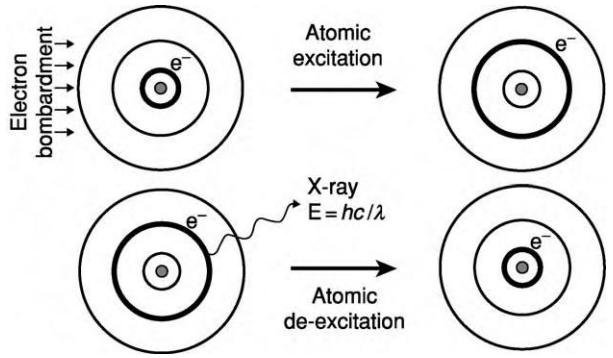


Figure 4 General set of processes involved in the generation of X rays by electron bombardment of atoms. The small grey filled circle represents the nucleus. The outer rings represent the quantized electron orbitals of the Bohr atomic model. The thicker black circle represents the location of a given electron (e). With electron bombardment, the highlighted electron jumps to a higher orbital. The energized electron quickly falls back to its original state, releasing an X ray.

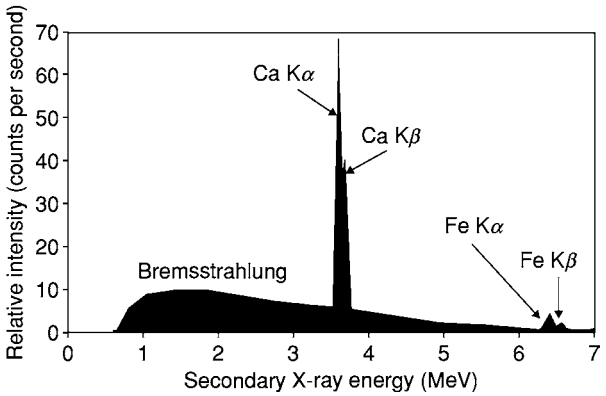


Figure 6 Example of a secondary X ray trace from a sample of calcite with a minor amount of iron substituting for the calcium (known as ferroan calcite). The X ray energy values are charac teristic of the elements. Note that the energy of the K lines in creases with increasing atomic number. Note also the increasing separation of the Kα and Kβ lines with increasing atomic number. The bremsstrahlung can cause problems with quantification, especially at low secondary X ray energies.

photon characteristic of the element. The resulting characteristic spectrum is superimposed on the brems-strahlung (Figure 6). An atom remains ionized for a very short time (about 10^{-14} s) and thus the incident electrons, which arrive about every 10^{-12} s, can repeatedly ionize an atom.

It is common practice to measure X-rays in units of thousand electron volts (KeV). One electron volt is

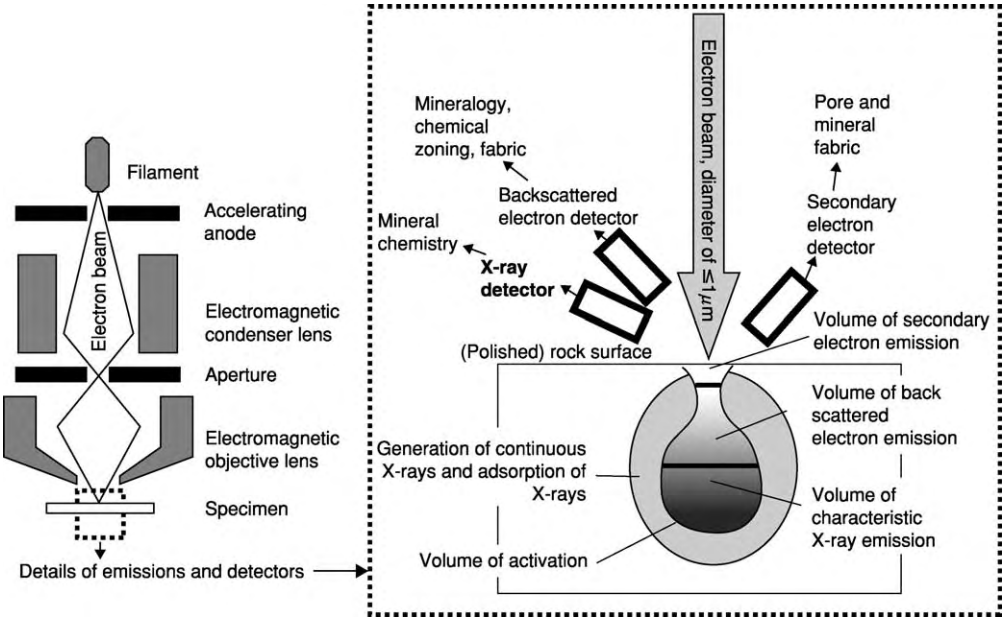


Figure 7 Basic equipment used for secondary X ray analysis following electron beam bombardment of Earth materials. The various types of signals generated in the samples and the various detectors are highlighted in the enlarged diagram. Characteristic X ray emission occurs several micrometres below the sample surface in a volume with a radius of a few times the electron beam radius (spot size). Bremsstrahlung are generated in the outlying volume where adsorption of X rays also occurs.

the same as 1.6021×10^{-19} J. An electron volt is the amount of energy gained by one electron when it is accelerated by one volt. The output characteristic X-rays have an energy and wavelength controlled by the element that is present and the specific nature of the electron orbital transition. Elements with low atomic numbers produce low-energy X-rays, and many X-ray detectors have difficulty quantifying the output from elements with atomic numbers of less than 11 (sodium). The energy and wavelength of the characteristic X-rays are related by the equation $E = hc/\lambda$, where E is the energy (in KeV), h is the Plank constant (6.626×10^{-34} J s), c is the speed of light (2.99782×10^{8} m s^{-1}), and λ is the element and electron orbital-specific wavelength of the X-rays. Conversion of X-ray energy into wavelength is thus achieved using E (eV) $= 123985/\lambda$ (in nm). X-ray spectra are typically plotted in terms of energy. The outputs of characteristic X-rays have intensities that are a complex function of the quantity of the element in the analysed volume. In general, the intensity is approximately proportional to the relative concentration. Several matrix (mineral) dependent processes alter the primary intensity of the secondary X-rays. The atomic mass of the element influences the efficiency of X-ray generation (Z-correction), some secondary X-rays are absorbed by the material (A-correction), and some of the absorbed X-rays result in localized X-ray fluorescence (F correction).

Secondary X-rays can be detected and quantified in two ways. One, which is best for major elements (>0.1 wt%), involves measuring the energy of the emitted X-rays using a scintillation counter (known as energy dispersive analysis). The other, best for trace elements (>1 ppm at best), measures the wavelength (and intensity) of the emitted X-rays (known as wavelength dispersive analysis). Energy dispersive analysis is typically faster than wavelength dispersive analysis but has much lower sensitivity.

Most modern devices are computerized and have inbuilt quantification and correction systems. The geochemical output from electron microprobes is in the form of oxide or element weight percentages.

X-ray Fluorescence

Although X-ray beams cannot be focussed or bent, they can be directed by a series of diffraction gratings, collimators, and slits. Such X-ray beams have been used for geochemical analysis by a technique known as X-ray fluorescence spectroscopy (XRF). The output from this technique gives the concentrations of most major elements (with atomic numbers of 11 (sodium) and above) and many useful trace elements.

It was stated earlier that some secondary X-rays can be absorbed by material and that some of this absorption can result in localized X-ray fluorescence (Figure 8). Just as with primary electron beams, when a primary X-ray excitation source (e.g. from an X-ray

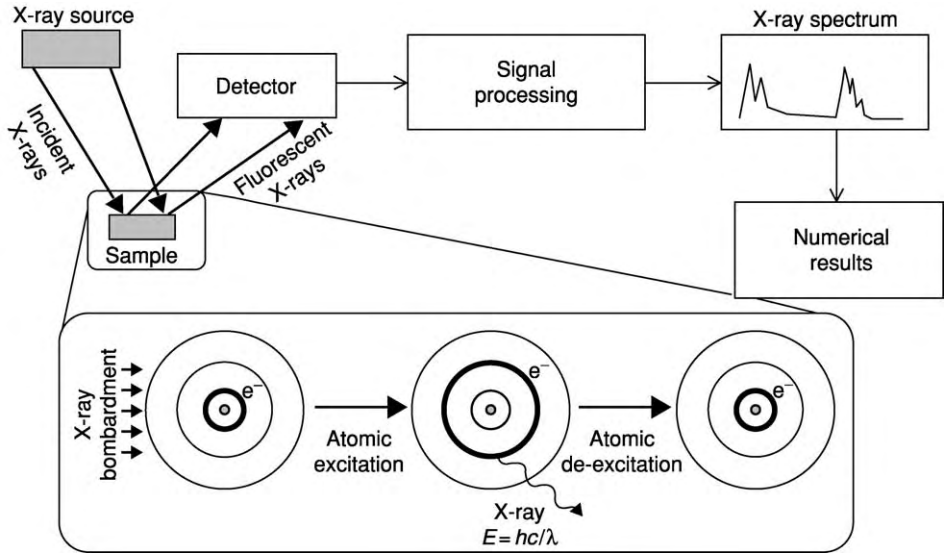


Figure 8 General set of processes involved in the generation of fluorescent X rays by X ray bombardment of atoms. The small grey filled circle represents the nucleus; the outer rings represent electron orbitals. The thicker black circle represents the location of a given electron (e). With X ray bombardment, the highlighted electron jumps to a higher orbital. The energized electron quickly falls back to its original state, releasing a secondary X ray. The range of elements in a sample leads to a range of characteristic fluorescent X rays with peak heights that are functions of the amounts of the elements in the sample.

tube or even a radioactive source) strikes a sample, the X-ray can be either absorbed by the atom or scattered through the material. Some incident X-rays are absorbed by the atom, and electrons are ejected from the inner shells to outer shells, creating vacancies. As the atom returns to its stable condition, electrons from the outer shells are transferred to the inner shells and in the process give off a characteristic X-ray with an energy equal to the difference between the two binding energies of the corresponding shells.

XRF is widely used to measure the elemental composition of natural materials, since it is fast and non-destructive to the sample. It can be used to measure many elements simultaneously. XRF can be used directly on rock surfaces, although there is a danger of natural heterogeneity resulting in variable results. Rock, soil, and sediment samples are typically crushed and made into pellets by compressing them or by melting the whole sample and then quenching to make a glass disk. XRF is useful for the geochemical analysis of a wide range of metals and refractory and amphoteric compounds (such as SiO_2 and Al_2O_3) and even some non-metals (chloride and bromide). XRF is also routinely used to measure the natural metal content of liquid petroleum samples. The quality of XRF data is a function of the selection of appropriate standards. It is considered to be best practice to use standards that are similar to the samples in question to minimize matrix effects. XRF can measure down to parts-per-million concentrations and lower, depending on the element and the material.

X-ray Diffraction

X-rays have a similar wavelength to the lattice spacing of common rock-forming minerals and have been used to characterize the crystal structure and mineralogy of Earth materials by using X-ray diffraction (XRD) analysis. This is most commonly used to define the presence of minerals, mineral proportions, mineral composition (in favourable circumstances), and other subtle mineralogical features of rocks, sediments, and soils.

X-rays, even from a pure elemental source bombarded with electrons, have a collection of peaks – X-rays characteristic of the quantized electron energy levels – and bremsstrahlung. X-ray beams of a tightly defined energy (and thus wavelength) have been used to investigate and characterize the minerals present in rocks, sediments, and soils by removing all but one X-ray peak from the spectrum of wavelengths generated by a source element. X-rays are useful in investigating mineral structure since they can be selected to have a wavelength that is only just smaller than the interlattice spacing (d-spacing) of common rock-forming minerals. A number of X-ray sources have historically been selected, but copper is the most commonly employed, and the copper Kα peak is the one that is directed onto samples. This has a characteristic wavelength of 1.5418 Å (0.15418 nm). This is ideal for many minerals since they have high-order (dominant, most obvious) lattice spacings of this size or up to 10–15 times greater than this wavelength.

The melee of X-radiation from copper can be reduced to the Kα peak and then directed by a series of diffraction gratings, collimators, and slits.

The main features of an XRD analyser include an X-ray source with collimators, slits, etc, a sample, and an X-ray detector (Figure 9). The source and detector are both rotated about the sample, an arbitrarily fixed point, and define the same angle (θ) relative to the sample. The angle between the source and the detector is thus 2θ relative to the sample.

Diffraction occurs when X-rays, light, or any other type of radiation passes into, but is then bounced back out of, a material with a regular series of layers. Diffraction occurs within the body of the material rather than from the surface (and so is quite different from reflection). Regular layers are a characteristic of all crystalline materials (minerals, metals, etc). Each rock-forming mineral has a well-defined set of these layers, which constitute the crystal lattice. No two minerals have exactly the same crystal structure, so fingerprinting a mineral by its characteristic set of lattice spacings helps to identify minerals. A radiation beam from a pure source has a defined wavelength, and the rays from such a pure source will be 'in phase'. Constructive interference occurs only when all the outgoing (diffracted) X-rays are also in phase. Destructive interference, the norm, occurs when the diffracted X-rays are no longer in phase. Constructive interference occurs when the extra distance that X-rays travel within the body of the material is an integer (whole number) multiple of the characteristic wavelength of the incident X-ray (Figure 10). The geometry of the XRD equipment, the wavelength of the incident radiation, and the lattice spacing are all important in defining whether constructive interference occurs. The key equation is known as the Bragg Law, which must be satisfied for constructive interference ('diffraction') to occur: $2d\sin\theta = n\lambda$, where d is the lattice spacing, λ is the wavelength of the incident X-ray source, and n is an integer (typically one in many cases). The value of θ, defined in Figure 9, is a function of the variable geometry of the XRD equipment.

X-ray diffraction is most commonly used on crushed (powered) rock samples to ensure homogeneity of the sample and randomness of the orientations of all the crystal lattices represented by different minerals. This is known as X-ray powder diffraction.

XRD works by rotating the X-ray source and the detector about the sample from small angles (e.g. 4°) through to angles of up to 70°. The low angles can detect large interlattice spacings (large values of d) while the high angles detect smaller interlattice spacings. For CuKα radiation these angles equate to d-spacings from about 30 Å down to about 1.5 Å, covering the dominant d-spacings of practically all rock-forming minerals.

For a pure mineral sample, the diffraction peaks from different lattice planes with discrete d-spacings have different relative intensities. This is a function of the details of the crystal structure of a particular mineral, but the maximum-intensity trace (peak) for many minerals has a low Miller Index value (a simple notation for describing the orientation of a crystal). For example, many clay minerals dominated by sheet-like crystal structures have (001) as the maximum-intensity peak. All other XRD traces have intensities that are fixed fractions of the intensity of the maximum-intensity trace. The result for each pure mineral is a fingerprint of XRD peaks on a chart of intensity on the y-axis and 2θ on the x-axis (Figure 11). This can be compared with collections of standards to identify the mineral.

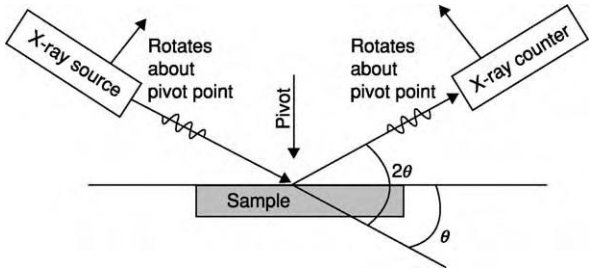


Figure 9 Basic elements of an X ray diffraction device. An X ray source is directed at a sample at a controlled and variable angle (θ). The X ray detector is at the equivalent angle on the opposite side of the pivot point. The source and detector are at an angle of 180° 2θ to one another. The source and detector are thus simultaneously rotated about the pivot point. When diffraction occurs the X ray detector records a signal above the background. The sample is usually a powder and preferably randomly orientated.

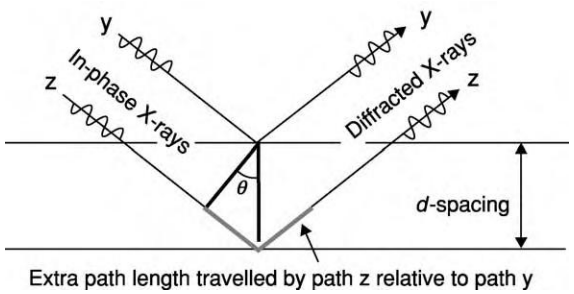


Figure 10 Diffraction from a crystal. The incident X rays are in phase as they hit the mineral surface. The grey line shows the path length difference between the two X rays. Constructive interference occurs when the extra path length ($2d\sin\theta$) is an integral multiple (typically one) of the wavelength of the X rays. Constructive interference leads to an X ray diffraction peak set against a low level of background noise.

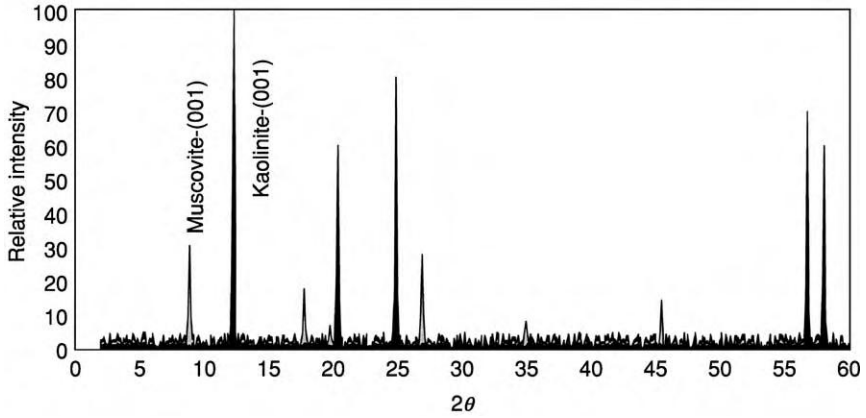


Figure 11 The X ray diffraction output from a rock composed of the clay minerals muscovite (pale grey peaks) and kaolinite (black peaks). The two minerals have discrete *d* spacings, which are reflected in their discrete peaks on the chart. The peaks at the low 2θ end have high *d* spacings, and vice versa. The dominant basal spacings of these two sheet silicates (the (001) planes) are labelled. The maximum intensity diffraction peaks from these minerals are produced by the (001) planes. From this diagram it would appear that there is much more kaolinite than muscovite in the mixture (approximately three times as much) since its maximum intensity peak is much more intense. The mixture therefore has about 25% muscovite and 75% kaolinite.

The intensity of the collection of diffraction peaks from a given mineral in a mixture (e.g. rock, soil, or sediment) is broadly a function of the proportion of the mineral in the mixture. This is a subject of ongoing research since the issue is complicated by different minerals having different efficiencies at diffracting X-rays. In simple terms, the intensities of the maximum-intensity peaks from a mixture of minerals give a guide to the relative proportions of the different minerals (Figure 11).

There is a range of problems for XRD when dealing with natural Earth materials. The most obvious is the simple identification of minerals when each one has its own collection of diffraction peaks. These must be carefully deconvoluted by a process of elimination. Computer-based programs can be of great help in this task. Many rock-forming minerals do not conform to a perfectly defined composition. The set of interlattice *d*-spacings in minerals is a function of the way atoms are packed together in the lattice, so that compositional variation affects the precise 2θ position of a given crystal orientation. In some cases, this variability can be put to good use since it can be used to identify the composition of a given mineral (Figure 12).

Another issue with XRD analysis, especially of sediments, soils, and sedimentary rocks is that many minerals have poorly defined crystal structure. This problem is typical of clay minerals, pedogenic oxides, etc. Poorly defined crystal structure translates into wider XRD peaks. This presents problems for quantification and results in a need to measure the area under an entire peak rather than the height of the maximum-intensity peak. However, this too can be put to good use by enabling the quantification of the transformation of a poorly defined crystal structure into a well-defined structure (typically as a function of time and temperature) during diagenesis or metamorphism (Figure 13).

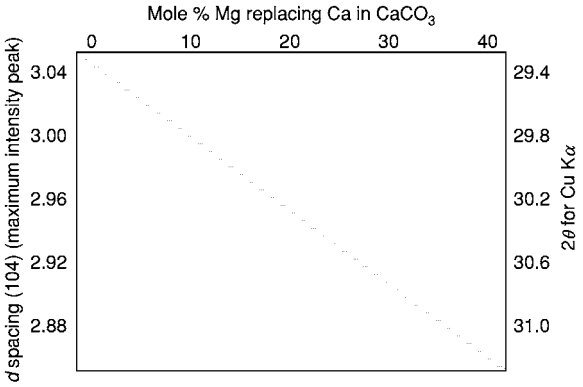


Figure 12 X ray diffraction for the same peak from the same mineral may occur at a different 2θ value if there is solid solution and the substituted element has a different ionic radius. For example, in calcite the dominant (104) diffraction peak moves systematically as magnesium replaces calcium. Magnesium ions are smaller than calcium ions so the structure collapses slightly. In favourable circumstances, this approach can be used to help determine mineral composition.

Optical Techniques

Spectroscopy is the use of the absorption, emission, or scattering of electromagnetic radiation by atoms or molecules (or atomic or molecular ions) to study the atoms or molecules qualitatively or quantitatively.

Isolated atoms can absorb and emit packets of electromagnetic radiation with discrete energies that

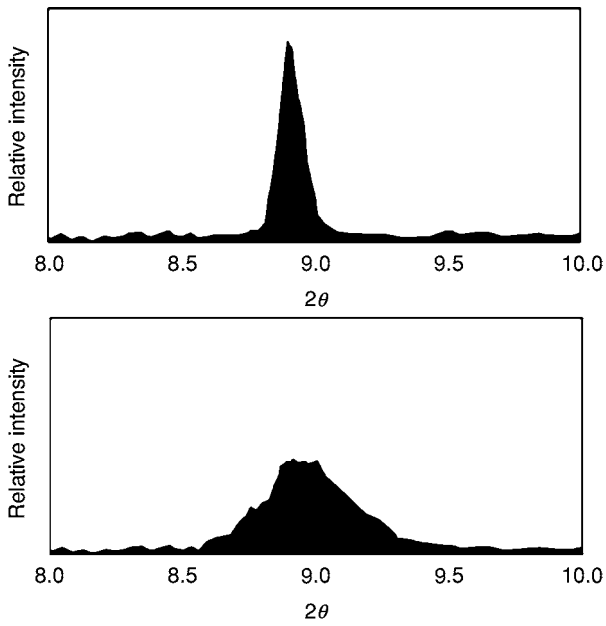


Figure 13 X ray diffraction outputs from a mineral (illite) with variable degrees of crystallinity. The top image represents a well crystalline sample, e.g. a slate. The lower image represents a poorly crystalline material, e.g. from a juvenile sedimentary rock, soil, or weathering horizon. The degree of crystallinity can be quantified by measuring the peak width at half the peak height.

are dictated by the detailed electronic structure of the atoms. When the resulting altered light is passed through a prism or spectrograph it is separated according to wavelength. Emission spectra are produced by high-temperature gases. The emission lines correspond to photons of discrete energies, which are emitted when excited atomic states in the gas make transitions back to lower-lying levels. Generally, solids, liquids, and dense gases emit light at all wavelengths when heated. An absorption spectrum occurs when light passes through a cold dilute gas and atoms in the gas absorb light at characteristic frequencies. Since the re-emitted light is unlikely to be emitted in the same direction as the absorbed photon was travelling, this gives rise to dark lines (absence of light) in the spectrum.

Emission and absorption spectroscopy have been used widely in geochemical analysis for many years (Table 3). They are commonly used to analyse waters but can also be employed to analyse rock and other solid samples that have been quantitatively dissolved (e.g. in acid solutions).

Atomic Absorption and Atomic Emission Spectroscopy

Atomic-absorption (AA) spectroscopy uses the absorption of light to measure the concentration of gas-phase atoms. Since samples are usually liquids or solids, the sample atoms or ions must be vaporized in a flame or graphite furnace (Figure 14). The atoms absorb ultraviolet or visible light. Concentrations are usually determined from a working curve after calibrating the instrument with standards of known concentration (Figure 15). The Lambert–Beer Law is the relationship between the change in light intensity for a given wavelength and the relative incident light energy: $\log I_o/I = aLc$, where I is the light intensity after the metal is added, I_o is the initial light intensity, a is a machine-dependent constant, L is the path length of light through the torch, and c is the concentration.

The light source is usually a hollow-cathode lamp of the element that is being measured. A major disadvantage of these narrow-band light sources is that only one element can be measured at a time. AA spectroscopy requires that the target atoms be in the gas phase. Ions or atoms in a sample must be desolvated and vaporized in a high-temperature source such as a flame or graphite furnace. Flame AA can analyse only solutions. Sample solutions are usually aspirated with the gas flow into a nebulizing and mixing chamber to form small droplets before entering the flame. AA spectrometers use monochromators and detectors for UV and visible light. The main purpose of the monochromator is to isolate the absorption line from background light due to interferences. Photomultiplier tubes are the most common detectors for AA spectroscopy.

This technique can be used to analyse aqueous samples with negligible sample preparation. It can also be used for rock and mineral samples if they are quantitatively dissolved (typically using acids of various strengths). Flame AA spectroscopy can typically detect concentrations as low as $mg\,l^{-1}$ (ppm), although graphite-furnace AA spectroscopy has been shown to be able to detect concentrations that are orders of magnitude below this.

Inductively Coupled Plasma Optical Emission Spectroscopy

ICP-OES is short for optical (or atomic) emission spectrometry with inductively coupled plasma. Plasma is a luminous volume of atoms and gas at extremely high temperature in an ionized state. The plasma is formed by argon flowing through a radio frequency field, where it is kept in a state of partial ionization, i.e. the gas consists partly of electrically charged particles. This allows it to reach very high temperatures of up to approximately 10 000°C. At these high temperatures, most elements emit light of characteristic wavelengths, which can be measured and used to determine the concentration of the elements in the solution.

Table 3 Optical spectroscopic techniques commonly used in geochemical analysis

Technique	*Output 1*	*Output 2*	*Output 3*	*Sample type*	*Advantages*	*Disadvantages*
Atomic absorption spectroscopy (AAS)	Element concentrations in water	Element concentration in quantitatively dissolved rock samples		Water sample from the natural or altered environment (or solid rocks and minerals dissolved in acid)	Well established technique; relatively low costs	Relatively high detection limit; relatively slow one element at a time; limited range of elements
Inductively coupled plasma optical emission spectroscopy (ICP OES)	Element concentrations in water	Element concentration in quantitatively dissolved rock samples		Water sample from the natural or altered environment (or solid rocks and minerals dissolved in acid)	Many elements analysed simultaneously; relatively fast; linear calibration over wide concentration range; wide concentration range; one calibration suitable for most material types; good for water samples; excellent (ppb) resolution for some trace elements	Expensive equipment; technique not for the novice analyst; carrier gases cause interference with some elements; minerals and rocks must be dissolved prior to analysis; relatively new technique still undergoing development
Infrared spectroscopy	Mineral proportions	Water content in quartz		Clay and other minerals	Ultra small samples; quantitative; relatively low cost; simple sample preparation	Complex to interpret absorption spectra; adsorbed water can interfere with diagnostic peaks
Ultraviolet spectroscopy	Presence of organic inclusions in minerals and rocks	Presence of liquid organics in porous materials (e.g. pollution or oil)	Approximate indication of the density and maturity of petroleum	Organic bearing rock, sediment, or soil	Simple to make qualitative observations; easily repeated; can be used on bulk and microscopic samples	Some minerals have masking fluorescence; difficult to quantity

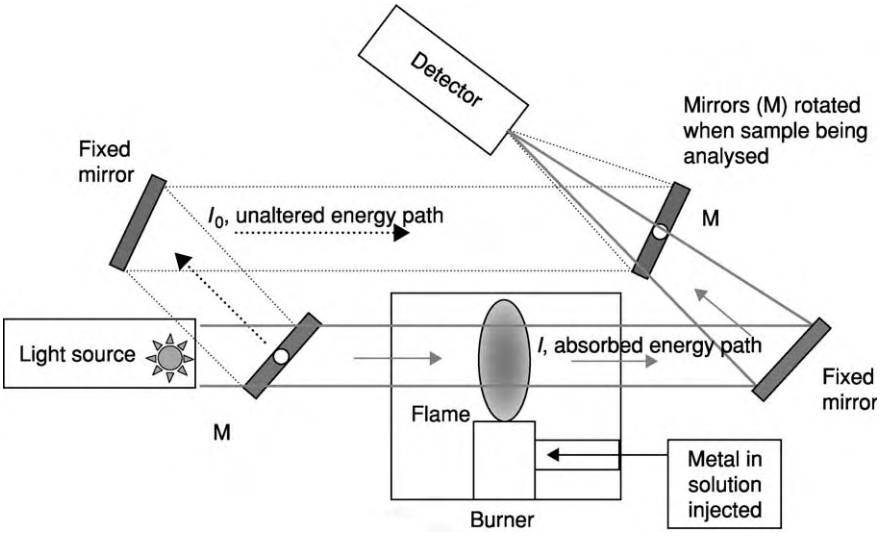


Figure 14 Atomic absorption spectroscopy equipment. The aqueous sample for analysis is injected into the burner and light is shone through the flame. The mirrors (labelled M) are rotated to measure original and the absorbed light characteristics. The path that is deflected away from the flame (labelled as unaltered energy path) measures the unaffected light intensity. The path that goes through the flame (labelled as absorbed energy path) measures the intensity of the light after it has passed through a burner containing the co injected dissolved sample.

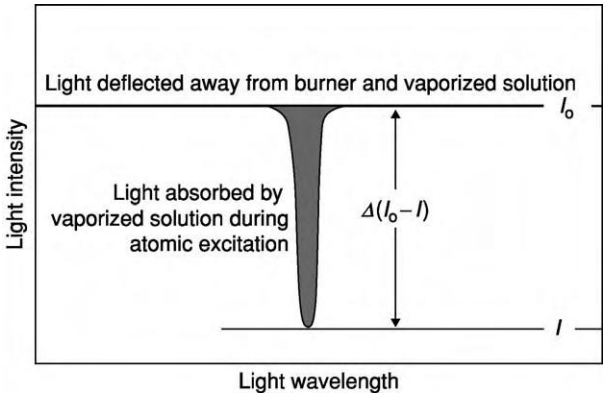


Figure 15 Light absorption due to atomic excitation with base line and post excitation light intensities indicated. The frequency of the absorbed light depends on the element. Concentration is proportional to I_o/I, all other things being equal.

The sample being analysed is introduced into the plasma as an aerosol of fine droplets (Figure 16). Light from the different elements is separated into different wavelengths by means of a grating and is captured simultaneously by light-sensitive detectors. This permits the simultaneous analysis of up to 40 elements, and ICP-OES is consequently a multi-element technique. In terms of sensitivity, ICP-OES is generally comparable with graphite furnace AA, i.e. detection limits are typically at the $\mu g\,l^{-1}$ level in aqueous solutions.

Ultraviolet Spectroscopy

Many organic-derived materials fluoresce under ultraviolet (UV) light – that is, they absorb light from the incident ultraviolet beam and release light, often in the visible range, that has a different wavelength from the primary UV light. UV spectroscopy can be used qualitatively to determine whether complex organic molecules are present in sediment or a rock. UV spectroscopy is commonly used to determine the presence of diffuse oil-shows in petroleum-reservoir cores.

There is a loose correspondence between the precise wavelength of the fluorescent light and the nature of the organic material. This relationship has been developed into an analytical technique to determine the thermal maturity of petroleum in rock samples. This technique has reached its apotheosis in its application to petroleum trapped in inclusions in mineral cements and healed fractures. These fluid inclusions have been used to track the petroleum generation of source rocks and the petroleum migration history into reservoirs (*see* **Fluid Inclusions**).

Infrared Spectroscopy

The electrical bonds between molecules continually vibrate as a function of interaction between neighbouring molecules. These bonds can be excited by infrared radiation, resulting in higher amplitudes of vibration. Only discrete (quantile) increases in vibration energy are possible, and this results in an

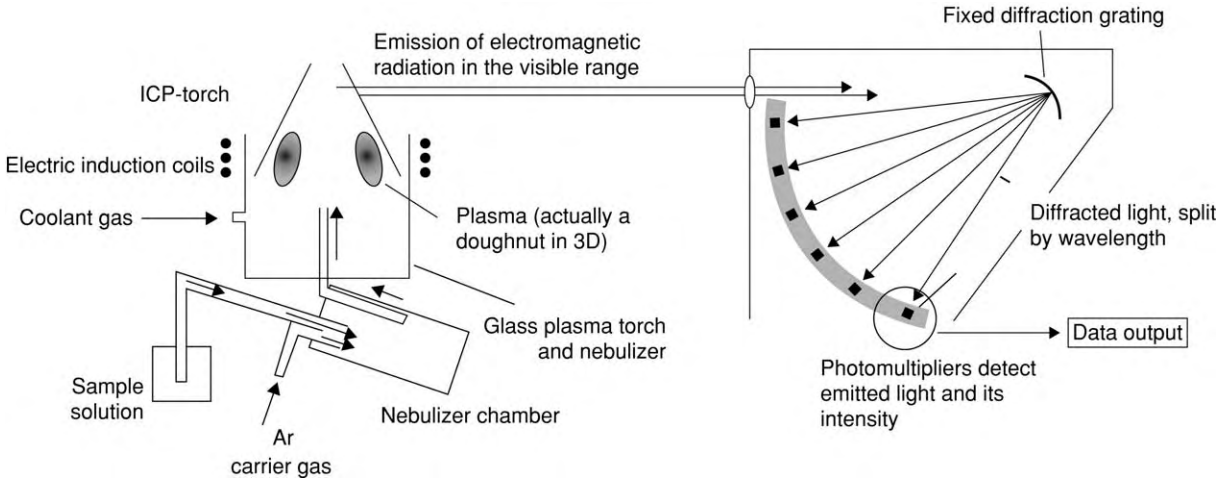


Figure 16 Essential components of an inductively coupled plasma optical emission spectrometer. The induction coils produce a ring of plasma at temperatures of *ca.* 10 000°C. The water sample, drawn in by the flowing carrier gas, is dispersed and drawn into the plasma. Elements emit light due to thermal excitation. A wide range of light frequencies can be analysed simultaneously using a diffraction grating to disperse the light.

absorption spectrum. The molecular environment is different in all minerals, even in minerals of similar structure (e.g. the clay family), so that infrared spectroscopy can be used to identify minerals in rock and sediment samples. The strength of the absorption spectrum depends on concentration, so the technique can be used for quantitative analysis of minerals under favourable circumstances. Infrared spectroscopy has been used to identify clay minerals in sandstones but is most useful for identifying organic inclusions and non-hydrocarbon gases trapped in fluid inclusions.

Chromatography

Chromatography has proved invaluable in geochemical analysis (Table 4). It is an analytical technique used to quantify and separate mixtures of fluid chemical compounds. There are many different kinds of chromatography, among them gas chromatography, organic liquid chromatography, and ion-exchange chromatography. All chromatographic methods share the same basic principles and mode of operation. In every case, a sample of the mixture to be analysed is applied to some stationary fixed material (the adsorbent or stationary phase) and then a second material (the eluent or mobile phase) is passed through or over the stationary phase. The compounds contained in the sample are then partitioned between the stationary phase and the mobile phase. The success of the approach depends on the fact that different materials adhere to the adsorbent with different forces. Those that adhere more strongly are moved through the adsorbent more slowly as the mobile phase flows over them. Other components of the sample that are less strongly adsorbed on the stationary phase and moved along more quickly by the moving phase. Thus as the mobile phase flows through the column, components in the sample move down the column at different rates and therefore separate from one another. At the end of the column, molecules or ions of the fastest-moving substance (least tightly bound to the stationary phase) emerge first, usually with each compound emerging over a well-defined time interval. A suitable detector analyses the output at the end of the column. Each time molecules or ions of the sample emerge from the chromatography column the detector generates a measurable signal, which is recorded as a peak on the chromatogram. The chromatogram is thus a record of detector output as a function of time and consists of a series of peaks corresponding to the different times at which components of the sample mixture emerge from the column.

By running standards and mixtures of known concentration, it is possible to relate the arrival time to species type and the size of the peak to concentration, making chromatography a valuable quantitative technique. There are three main types of chromatography employed in geochemical laboratories: liquid chromatography, gas chromatography, and ion chromatography. Gas chromatography is used to separate mixtures of gases or vaporized liquids. Ion chromatography is used to separate and analyse ions, typically but not exclusively anions, in aqueous solutions. Liquid chromatography is included in Table 4 but is mainly used as a sample-preparation procedure prior to gas chromatography to split petroleum, e.g. into

Table 4 Chromatographic techniques commonly used in geochemical analysis

Technique	*Output 1*	*Output 2*	*Output 3*	*Sample type*	*Advantages*	*Disadvantages*
Liquid chromatography	Physical separation of different components of complex liquid mixture (petroleum)			Whole petroleum or extracted bitumen samples	Excellent pre separation technique for GC and GC MS analyses; Gives quantities of groups of petroleum compounds	Limited separation capability
Gas chromatography	Physical separation of different components of complex gas phase mixture (petroleum gas or heated volatilized liquid)	Quantitative measure of proportions of different compounds in gas and liquid petroleum		Either whole petroleum or separate parts (achieved using liquid chromatography)	Splits gas and liquid range compounds; easily quantified; well established technology; good for samples with dominant alkanes	Co elution of different compounds; requires sample preparation; unknown GC peaks can give ambiguous interpretation
Ion chromatography	Physical separation of different charged (aqueous) components in complex natural solutions	Quantitative measure of proportions of different anions in water (common application)	Quantitative measure of proportions of different cations in water (less common application)	Water sample or solid sample quantitatively dissolved in water	Splits a range of anions in water; high resolution; relatively fast and simultaneously analyses all anions; no real alternative	Unsuitable for bicarbonate analysis

saturated, aromatic, and resin groups, and will not be covered further here.

Gas Chromatography

Gas chromatography (GC) – specifically gas–liquid chromatography – involves a sample being vaporized and injected onto the head of the chromatographic column (Table 4). The sample is transported through the column by the flow of an inert gaseous mobile phase. The column itself contains a liquid stationary phase that is adsorbed onto the surface of an inert solid (Figure 17).

The carrier gas is chemically inert (e.g. helium). A sample of gas or petroleum is injected into the column quickly as a slug to prevent peak broadening and loss of resolution. The temperature of the sample port is somewhat higher than the boiling point of the least-volatile component of the sample. Sample sizes typically range from tenths of a microlitre to 20 μl. The carrier gas enters a mixing chamber, where the sample vaporizes to form a mixture of carrier gas, vaporized solvent, and vaporized solutes. A proportion of this mixture passes onto the chromatography column. Chromatography columns have an internal diameter of a few tenths of a millimetre and walls coated with a liquid but stationary phase. The optimum column temperature depends on the boiling point of the sample; typically a temperature slightly above the average boiling point of the sample results in an elution time of 2–30 min. As the carrier gas containing the chromatographically separated sample passes out of the end of the column, it is passed into one of a number of detectors such as a flame ionization detector, which has high sensitivity, a large linear response range, and low noise. An example of a flame-ionization-detector signal from a whole-petroleum sample injected onto a GC column is given in Figure 18.

One of the problems of GC analysis of geochemical samples is that different compounds can have similar elution times, rendering identification and quantification difficult. However, the output stream from a gas chromatograph can be passed into other types of analytical instrument (e.g. a mass spectrometer) for further analysis of the separated compounds over and above simple quantification of compounds with a common elution time. GC is useful for analysing organic compounds and can be used to quantify mixtures if suitable standards have been employed.

Ion Chromatography

Ion chromatography can be used for both cations and anions. However, it is in the analysis of non-metal ions that the technique has proved most useful mainly because there are no real alternatives for the simultaneous quantitative analysis of these important species in waters or synthetic solutions.

Ion chromatography is used to analyse aqueous samples containing ppm quantities of common anions (such as fluoride, chloride, nitrite, nitrate, and sulphate). Ion chromatography is a form of liquid chromatography that uses ion-exchange resins to separate atomic or molecular ions based on their interaction with the resin (Figure 19). Its greatest utility is for the

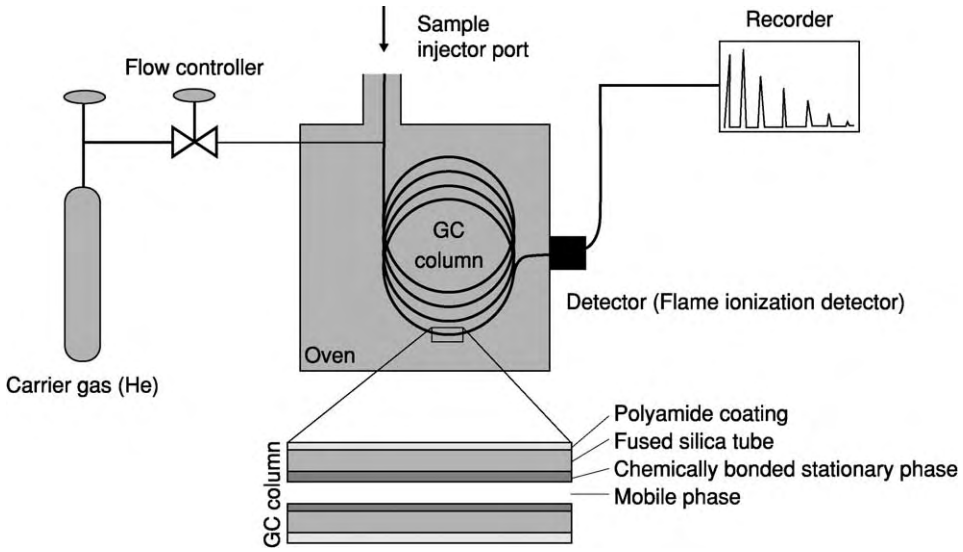


Figure 17 Essential components of a gas chromatograph. The sample injection port is heated to volatilize liquid phase organics. The GC column is held in an oven at a temperature above the boiling point of the compounds of interest. The inert carrier gas drives the sample through the capillary with its stationary phase. The stationary phase retards larger molecules more efficiently than small molecules. Smaller molecules thus pass out of the column more rapidly than large molecules.

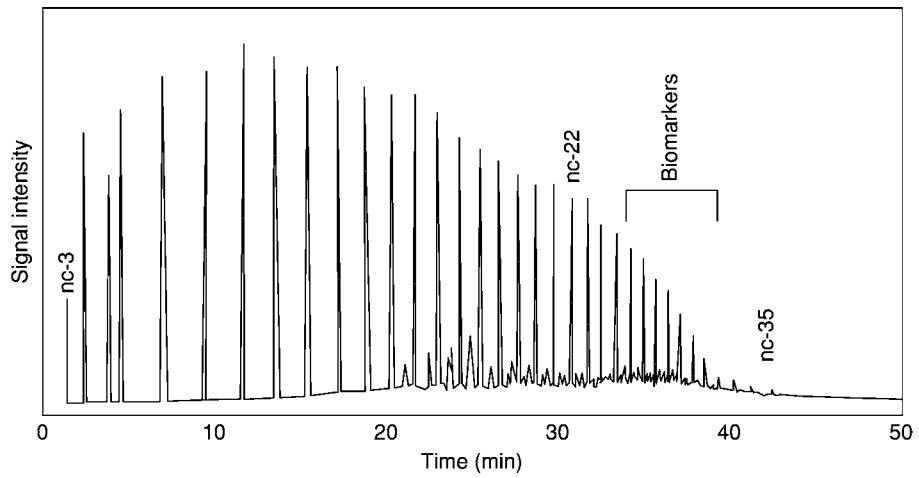


Figure 18 Typical GC output from a black oil sample. Most of the peaks correspond to normal alkanes. The longest chain alkane detected is $C_{35}H_{72}$. The biomarkers, molecules of clear biological origin, form an area of low level noise that is difficult to discern with this technique. On the figure nc 3, nc 22, nc 35 refer to normal alkanes with 3, 22 and 35 carbon atoms. The three labelled peaks are thus from C_3H_8, $C_{22}H_{46}$ and $C_{35}H_{72}$.

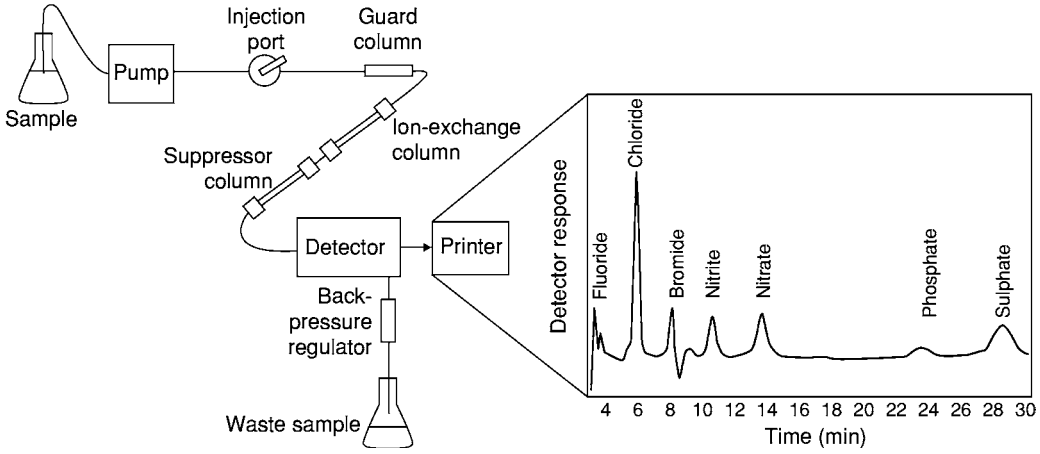


Figure 19 Essential components of an ion exchange chromatograph, and a typical output trace for a low concentration standard. The suppressor column removes the bicarbonate and carbonate that are released from the ion exchange chromatography column to avoid the real sample signal being swamped. The trace has 0.02 ppm fluoride and 0.1 ppm of all the other anions. Note that the transit time through the column increases with atomic number for the halogens and is greatest for the multivalent anions phosphate and sulphate.

analysis of anions for which there are no other rapid analytical methods. For anion chromatography the mobile phase is a dilute aqueous solution of sodium bicarbonate and sodium carbonate prepared with pure water.

The ion-exchange column is tightly packed with the stationary adsorbent. This adsorbent is usually composed of tiny polymer beads that have positively charged centres. These become coated with bicarbonate and carbonate anions if no sample is passing through the column. As anions in the sample enter the column, they are attracted to the positive centres on the polymer surface and may replace (exchange with) the bicarbonate and carbonate ions stuck to the surface. Usually, the greater the charge on the anion the more strongly it is attracted to the surface of the polymer bead. Also, larger anions generally move more slowly through the column than smaller anions. The result is that the sample separates into bands of different kinds of ion as it travels through the column.

The detector, usually a conductivity cell, measures the conductance of the solution passing through it. The conductance is proportional to the concentration of ions dissolved in the solution. It is essential to

pass the sample–mobile-phase mixture through a suppresser column – another ion-exchange column – to remove the bicarbonate and carbonate ions and avoid the sample signal being masked before entering the detector. Anions can be qualitatively identified by analysing standards and standard mixtures. The concentrations of anions are quantified by the usual geochemical technique of calibration (Figure 1).

Mass Spectroscopy

Principles of Mass Spectroscopy

If a moving charged molecule or atom is subjected to a sideways electromagnetic force, it will move in a curve. The amount of deflection in a given electromagnetic field depends on the velocity and mass of the ion and the number of charges on it. The latter two factors are combined in the mass–charge ratio (given the symbol m/z or sometimes m/e). Most of the ions passing through the mass spectrometer will have a charge of +1, so the mass–charge ratio will be the same as the mass of the ion. Atoms and molecules are ionized by removing one or more electron in an ionization chamber to leave a positive ion. The beam of ions passing through the machine is detected electrically. It is important that the ions produced in the ionization chamber have a free run through the device without hitting air molecules so that equipment is operated under a high vacuum. The vaporized sample passes into the ionization chamber. The electrically heated metal coil gives off electrons that are attracted to the electron trap, which is a positively charged plate. The particles in the sample (atoms or molecules) are therefore bombarded with a stream of electrons, and some of the collisions are energetic enough to knock one or more electrons out of the sample particles to make positive ions. Most of the positive ions formed will carry a charge of +1 because it is much more difficult to remove further electrons from an already positive ion. These positive ions are persuaded out into the rest of the machine by the ion repeller, which is another metal plate carrying a slight positive charge.

When an ion hits the detector, its charge is neutralized by an electron jumping from the metal to the ion. That leaves a space amongst the electrons in the metal, and the electrons in the wire shuffle along to fill it. A flow of electrons in the wire is detected as an electric current, which can be amplified and recorded.

If the magnetic field is varied, the ion stream can be deflected on to the detector to produce a current that is proportional to the number of ions arriving. The mass of the ion being detected is related to the size of the magnetic field used to bring it to the detector. The device can be calibrated to record current (which is a measure of the number of ions) directly against m/z. The output from the chart recorder is usually simplified into a 'stick diagram'. This shows the relative current produced by ions of varying mass–charge ratio (Figure 20). There are a wide range of mass-spectroscopy techniques; three have been summarized in Table 5.

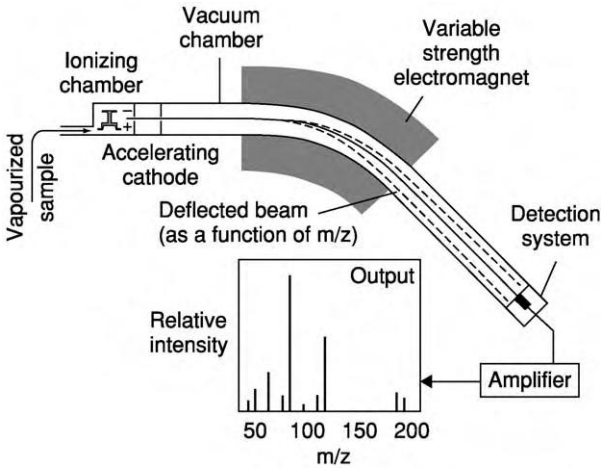


Figure 20 Essential components of a mass spectrometer. A source (from thermal ionization, inductively coupled plasma, a GC column, etc) feeds vaporized sample into an ionization chamber. The vapour is ionized to produce positively charged ions. These are drawn into the mass spectrometer via an accel erating electrode. The ions are passed into an electromagnet, where they are deflected as a function of the mass charge (m/z) ratio, their velocity, and the strength of the magnetic field. The electromagnets have their field strength varied to cause variable deflection. The signal strength is recorded as a function of mag netic field strength, and the signal output is given in terms of the relative intensity of the peak versus m/z.

Thermal-Ionization Isotope-Ratio Mass Spectroscopy

The technique of isotope dilution is being used increasingly to improve precision and accuracy by reducing the problems of calibration and sample-preparation effects. Some materials for analysis and some analytical methods result in a large degree of uncertainty. Variability caused by such problems is usually partly compensated for or monitored by using internal standards and surrogate samples. An isotope-dilution standard is the 'perfect' internal standard or surrogate.

An internal standard or surrogate is a compound similar to the sample of interest. An isotope-dilution standard is an isotope of an element or a molecular compound labelled with an isotope. A good example of this is ^{204}Pb, a minor isotope of lead. The natural stable lead isotopes are 204 (1.4%), 206 (24.1%), 207 (22.1%), and 208 (52.4%). By adding a known

Table 5 Mass spectrometry techniques commonly used in geochemical analysis

Technique	*Output 1*	*Sample type*	*Advantages*	*Disadvantages*
Thermal ionization mass spectroscopy	Concentrations of individual elements and isotopes, typically heavier elements	Solid salt of the element in question deposited on a filament	Relatively simple; gives isotope ratios; can be quantitative if sample diluted with known concentration of isotope; high temperature of filament causes ionization directly	Only for limited range of elements; only for solid samples
Inductively coupled plasma mass spectroscopy	Concentrations of trace and minor elements	Water and quantitatively dissolved solids	High analytical resolution; small sample sizes; wide range of elements; rapid analysis; no matrix effects due to sample dissolution	Expensive equipment; difficult technique requiring expert operator; solid samples require quantitative dissolution
Gas chromatography mass spectroscopy	Concentrations of trace organic compounds	Volatile or dissolved organic compounds	High analytical resolution; easy to quantify; large range of compounds can be determined from one sample	Problems with *ab initio* determination of unknown compounds; expensive and tricky technique

amount of ^{204}Pb to a sample before testing for total lead and by testing for each of the lead isotopes, it is possible to determine accurately total lead and individual lead-isotope ratios. This approach leads to much smaller errors than simply calibrating signal strength for individual m/z values.

Gas Chromatography Mass Spectroscopy

One of the main problems with analysis of GC column output is the co-elution (same rate of passage) of groups of compounds. This problem has been tackled by using a mass spectrometer as the detection system (rather than, for example, a flame ionization detector). Placed at the end of a chromatographic column in a similar manner to other GC detectors, a mass spectrometer detector is more complicated than other GC detector systems (Figure 21). A capillary column must be used in the chromatograph because the entire mass spectroscopy process must be carried out at very low pressures and in order to meet this requirement a vacuum is maintained via constant pumping using a vacuum pump.

The major components over and above the GC column are an ionization source, a mass separator, and an ion detector. There are two common mass analysers or separators commercially available for GC-MS: the quadrapole and the ion trap.

The power of this technique lies in its ability to produce mass spectra from each time-controlled GC peak (Figure 18). Thus, for each GC peak, coeluted molecules are ionized and separated into m/z fractions in the mass spectrometer. Complex organic molecules tend to fragment in predictable ways in the ion source, so a group of related molecules can be traced using the same m/z fraction for the different time-controlled GC fractions. A time plot of the same ionized molecular fragment can then be reconstructed from the individual intensities of the mass spectra. The data can be used to determine the identity and quantity of an unknown chromatographic component with an assuredness that is simply unavailable by other techniques. This approach can be quantified if the equipment is calibrated or if the sample is mixed with a known quantity of a standard. GC-MS analysis allows the quantification of trace organic components including biomarkers, thermally controlled optical isomers (e.g. steranes), and geological age-dependent molecules (e.g. olearane, derived from post-Cretaceous flowering plants).

Inductively Coupled Plasma Mass Spectroscopy

Inductively coupled plasma mass spectroscopy (ICP-MS) uses plasma of the same type as in ICP-OES (Figure 16), but here it is used to convert elements to ions that are then separated by mass in a mass spectrometer. This allows the different elements in a sample (and their natural isotopes) to be separated and their concentrations determined. The core of the ICP-MS system is the interface through which ions from the inductively coupled plasma source enter the high-vacuum chamber of the mass spectrometer.

ICP-MS combines the advantages of inductively coupled plasma (simple and rapid sample handling) and mass spectrometry (high sensitivity, isotope measurement) in a multielement technique. Detection limits can be much lower than in ICP-AES: certain

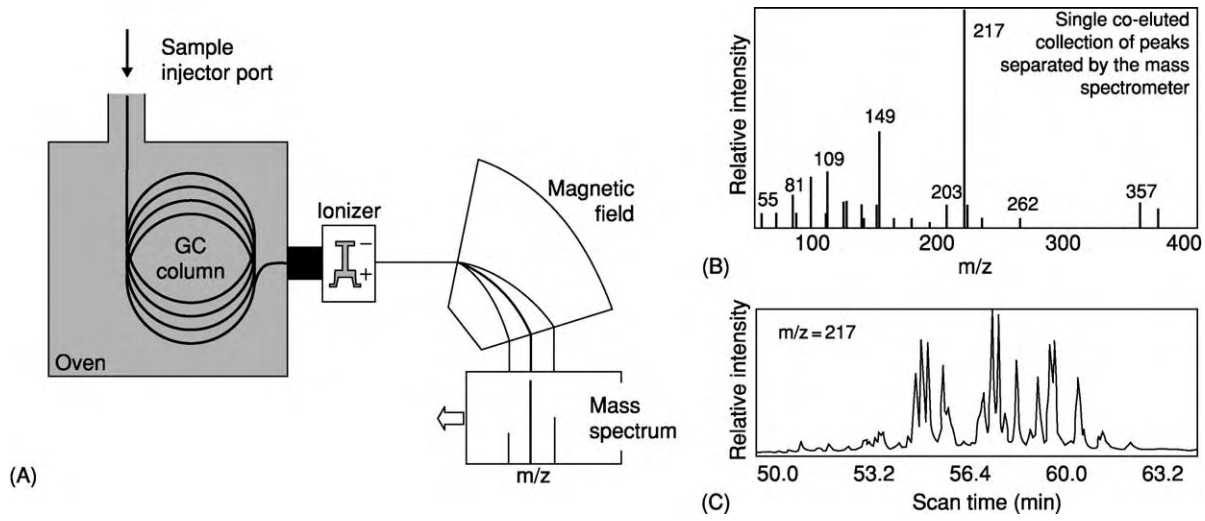


Figure 21 Essential components of a gas chromatography mass spectrometer and its output. The sample injection port is heated to volatilize liquid phase organics. The capillary GC column is held in an oven at a temperature above the boiling point of the compounds of interest. The GC column separates the sample into groups of molecules according to their elution rates. The GC separated sample is fed into the ionization chamber, where the organic molecules are fragmented, drawn into the mass spectrometer, and separated into different m/z fragments. (B) Each peak on a GC trace (**Figure 18**) has its own mass spectrum, permitting high resolution analysis of trace organic molecules. (C) The individual m/z fragments (e.g. 217) can be reconstructed as a plot of fragment concentration versus time. The high resolution of this technique permits the resolution of the biomarkers shown on the GC trace in **Figure 18**.

elements can be detected at the ng l^{-1} (parts per trillion) level in aqueous solutions.

Pyrolysis and Other Heating Techniques)

It is possible to characterize materials by heating them and either by studying the resulting change in optical and physical properties or by analysing the fluid evolved. This approach has been applied to organic materials, minerals, rocks, and fluid trapped as inclusions within mineral grains (Table 6).

Thermogravimetry and Evolved Water Analysis

If a small sample of rock is subjected to a controlled heating cycle and simultaneously weighed, the temperature at which volatiles are driven off from the sample can be accurately monitored. If the volatiles are carried to a detector using an inert gas (e.g. nitrogen), then it is also possible to analyse the evolved gases using an infrared water-vapour analyser to determine independently the exact quantity of water driven off at each stage in the heating cycle. The first technique is known as thermogravimetry; the second technique is known as evolved water analysis. Carbonate minerals also undergo volatile loss during heating, so it is important to pair the weight loss with the identification of the mineral undergoing volatile loss. The combined approach allows accurate assessment of volatile loss from clay minerals. These techniques can be used to determine the quantity of clay minerals in a rock under ideal conditions the exact types and quantities of clays can be determined since different clay minerals dehydrate at different temperatures.

Pyrolysis

Pyrolysis has been used to study organic material for a number of purposes. Heating organic matter is used to study the resulting total quantity of evolved CO_2, SO_2 etc. (when done in an oxygen atmosphere), which can be used to help determine the elemental composition of the organic matter. The elements in organic matter of all sorts, determined in this way, include carbon, hydrogen, sulphur, and nitrogen. The resulting evolved gas phases can be split using a GC column (see above) or analysed using various optical (e.g. infrared) techniques specific to the expected gas products. The output from this approach can be the total organic carbon content (e.g. of a rock or sediment) or the elemental analysis of pre-separated samples (Figure 22).

Another approach is to heat solid samples (e.g. of organic-rich sediment, coal, separated kerogen, or asphaltene exsolved from petroleum) in an oxygen-free environment and analyse the resulting fluid-phase products during a programmed heating cycle (known as rock eval pyrolysis). During heating the

Table 6 Pyrolysis and other thermal techniques commonly used in geochemical analysis

Technique	*Output 1*	*Output 2*	*Output 3*	*Sample type*	*Advantages*	*Disadvantages*
Thermogravimetry evolved water analysis	Estimate of total clay mineral content of rocks			Small samples of rocks, sediments, or minerals	Sub percentage level resolution; useful ally to XRD analysis; rapid	Difficult to resolve individual clay minerals in clastic rocks; difficult to repeat; destroys sample
Pyrolysis	Total organic carbon (on decarbonated samples)	Quantities of existing petroleum, petroleum generation potential, CO_2 generating potential (rock eval)	General character of petroleum that would be generated by organic matter with further heating (pyrolysis GC)	Organic bearing sediment or rock, solid asphaltene exolved from petroleum, bitumen	Rapid; approach allied to existing technology; quantitative; can reveal kinetic data about source rocks	Destroys sample; geologically unrealistic rates of heating (for rock eval); results sensitive to geological history of sample
Fluid inclusion microthermometry	Salinity of water trapped in minerals	Growth temperature of mineral		Polished wafers of rocks and minerals	High level of precision; easily repeated; high spatial resolution; reveals geological evolution of samples	Difficult sample preparation; requires assumption that salinity is due to NaCl; requires assumption that vapour saturated fluid was trapped

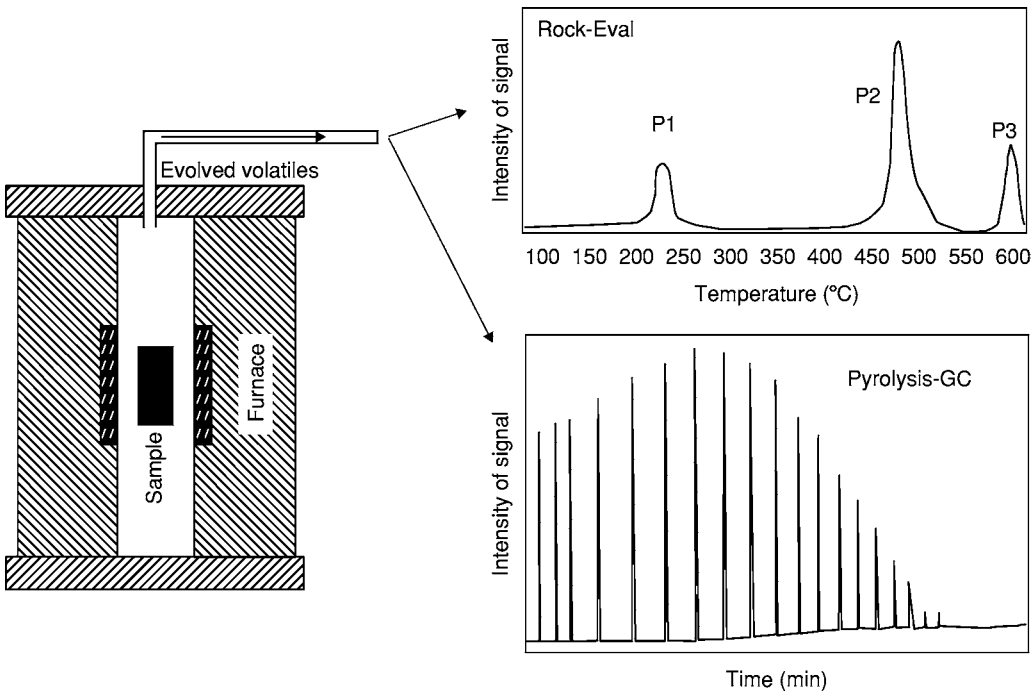


Figure 22 Essential components of pyrolysis equipment, with either the volatiles sequentially emitted during a programmed heating cycle or GC analysis of the pyrolysate. P1 represents evaporated pre existing petroleum. P2 represents the petroleum generated from kerogen. P3 represents the generation of oxygen bearing species (CO_2). T_{max} is the peak temperature of the P2 trace. The pyrolysis GC trace is typically used to determine the gas (short alkanes, small elution time) versus oil (longer alkanes, longer elution time) generation capability of immature kerogen.

free hydrocarbons contained in the sample are driven off first (the P1 peak) followed by experimentally generated petroleum (the P2 peak) and then oxygen-containing compounds (CO_2; the P3 peak). Detection in simple instruments is performed using a flame ionization detector (for free and laboratory-generated hydrocarbons) and a thermal conductivity detector for carbon dioxide. Three peaks are thus usually produced and quantified during heating and analysis. These peaks are expressed in terms of mg g ^{1}. P1 indicates how much petroleum exists in an organic-rich rock now. P2 indicates the petroleum-generating capability of the rock and is used to calculate the hydrogen index of the source rock. P3 is an indication of the amount of oxygen in the kerogen and is used to calculate the oxygen index. The temperature at which the maximum release of hydrocarbons from cracking of kerogen occurs during pyrolysis (P2 peak) is termed T_{max} and is an indication of the stage of maturation of the organic matter. The hydrogen index (HI = (100 × P2)/total organic carbon) and the oxygen index (OI = (100 × P3)/total organic carbon) correlate with the ratios of hydrogen to carbon and oxygen to carbon, respectively. These parameters have been usefully employed to study the type and thermal evolution of sedimentary organic matter.

The products of pyrolysis in the absence of oxygen can also be passed into a GC column to study their composition. This is known as pyrolysis-GC and can usefully simulate the type of petroleum that would be expected from the organic-rich source rock during thermal evolution.

Fluid Inclusion Microthermometry

Small samples of the fluid from which a mineral grew are commonly trapped in inclusions. When minerals are fractured, they sometimes re-heal, trapping the ambient fluid. Petroleum, which is immiscible with the aqueous mineral growth medium, can also get trapped in these inclusions. These fluid inclusions have proved to be very valuable to geochemists since they provide a snapshot of fluid evolution and rock geohistory. The fluid itself can be analysed if it is released by crushing. This approach is especially useful for petroleum inclusions. Analysis is by GC, GC-MS, or simply mass spectroscopy. UV spectroscopy can also be used to analyse petroleum trapped in inclusions to help reveal the broad characteristics of the petroleum.

Aqueous inclusions can be analysed either by freezing the sample and using electron micoprobe (secondary X-ray) analysis, or by crushing the sample and

collecting the fluid for conventional water analysis (inductively coupled plasma techniques, ion chromatography). Fluid inclusions are listed here since they are most commonly analysed by using a high powered microscope and heating–cooling stage. Most fluid inclusions are composed of discrete liquid and vapour phases even though they would have been trapped as a single-phase liquid, which is typically assumed to have been saturated with vapour at the time of trapping. During a heating cycle, the two phases homogenize; the precise temperature of homogenization reveals the temperature at which the mineral grew. The salinity of the water can be assessed by monitoring the temperature at which it starts freezing, since this temperature can be related to salt content (assuming that the water is dominated by dissolved NaCl). A combination of thorough petrography and thermometric studies of aqueous inclusions can help to reveal details of the thermal and mineral-growth history as well as the fluid evolution history.

Related Geochemical Techniques

The techniques listed and briefly discussed here are only some of the vast panoply of techniques that have been employed during geochemical studies over the last 50 years or so. Some have now fallen out of favour. For example, a technique called neutron activation analysis was used for a long time to measure trace elements in solids. It has fallen out of favour mainly owing to developments in ICP-OES and ICP-MS.

The vast range of light and electron optical techniques are routinely used in conjunction with a wide range of solid-state and even organic geochemical studies. Scanning electron microscopy has recently been extensively developed and can now give fabric, mineralogy, mineral chemistry, and high-resolution crystallographic information. Transmission electron microscopy can provide ultra-high spatial resolution (of the order of tens of nm) geochemical data (using secondary X-rays) as well as fabric and crystallographic data.

A wide range of wet geochemical techniques have been employed routinely in studies of natural waters from all near-surface and surface environments. Titration, electrochemical techniques, and colorimetry are essential techniques that are used routinely in many geochemical studies.

See Also

Clay Minerals. **Fluid Inclusions**. **Minerals:** Definition and Classification; Native Elements. **Petroleum Geology:** Chemical and Physical Properties; The Petroleum System. **Rocks and Their Classification**.

Further Reading

Emery D and Robinson AC (1993) *Inorganic Geochemistry: Applications to Petroleum Geology*. Oxford: Blackwells.

Farmer VC (1974) *The Infrared Spectra of Minerals*. London: Mineralogical Society.

Faure G (1986) *Principle of Isotope Geology*. New York: John Wiley and Sons.

Goldstein RH and Reynolds TJ (1994) *Systems of Fluid Inclusions*. Tulsa: Society of Sedimentary Geology.

Hagemann HW and Hollerbach A (1986) The fluorescence behaviour of crude oils with respect to their thermal maturation and degradation. *Organic Geochemistry* 10: 473 480.

Jarvis I and Jarvis KE (1992) Plasma spectrometry in the Earth sciences: techniques, applications and future trends. *Chemical Geology* 95: 1 33.

Jenkins R (1999) *X ray Fluorescence Spectrometry*. New York: Wiley Interscience.

Lico MS, Kharaka YK, Carothers WM, and Wright VA (1982) *Methods for the Collection and Analysis of Geo pressured Geothermal and Oil Field Waters*. Water Supply Paper 2194. United States Geological Survey.

Moore DM and Reynolds RC (1997) *X ray Diffraction and the Identification and Analysis of Clay Minerals*. Oxford: Oxford University Press.

Rollinson HR (1993) *Using Geochemical Data: Evalu ation, Presentation, Interpretation*. New York: Longman Scientific and Technical.

Tissot B and Welte D (1984) *Petroleum Formation and Occurrence*. Berlin: Springer Verlag.

Tucker ME (1988) *Techniques in Sedimentology*. Oxford: Blackwell Scientific.

Weiss J (2000) *Ion Chromatography*. New York: John Wiley and Sons.

Zussman J (1967) *Physical Methods in Determinative Mineralogy*. London: Academic Press.

Geochronological Techniques

E A Eide, Geological Survey of Norway, Trondheim, Norway

Introduction

Geochronology is the study of time as it relates to Earth history. As a distinct discipline within the natural sciences, geochronology emerged fully during the late nineteenth and early twentieth centuries with the discovery of radioactivity and the advent of radiometric dating methods. Importantly, the appearance of modern geochronology was the result of a strong interest in Earth history and the development of relative methods to estimate the age of Earth, both of which had been aspects of natural science research since at least the seventeenth century.

The human fascination with studying time and marking its passage can be traced to ancient cultures, exemplified through the precise astronomical calendars produced by numerous early civilizations (Figure 1). These calendars were based on calculations of the movements of celestial bodies relative to Earth and helped to raise speculations about the position and motion of Earth within this celestial system. These speculations led to efforts to understand Earth's origin and calculate its age, which today is generally agreed to be 4.5–4.6 billion years (By), and is the starting point for the geological time-scale (GTS) (Figure 2). The GTS is an iterative solution between 'absolute' and 'relative' ages determined by absolute and relative geochronological techniques. The formal distinction between absolute and relative ages has its roots in ancient calendars for which the passage of time was calculated from astronomical events linked to the solar year. Broadly, an absolute age is one that is based on processes affected only by the passage of time and which may thus be valid worldwide. In a strict sense, an absolute age should have direct correspondence to the absolute time-scale, determined on the basis of the solar year (Table 1). Relative ages are applicable to a restricted geographic area and usually pertain to a limited geological time period. Relative ages place the formation of different rock units or their physical features (faults, unconformities, etc.) in a relative chronological order. Though knowledge of the exact formation ages of different rock units is useful, numerical (absolute) ages are not prerequisite for establishing their relative chronology. Nonetheless, relative ages must eventually be calibrated against independently established (absolute) time-scales if they are to be extrapolated globally.

The framework for the GTS is based on relative ages, represented by the established, sequential subdivisions of geological time (Figure 2). The nomenclature of this framework was developed largely through the studies of natural scientists in the eighteenth and nineteenth centuries (*see* **Famous Geologists:** Sedgwick; Murchison; Darwin; Smith; Cuvier; Hutton). During the twentieth century, absolute age determinations for rocks around the globe allowed refinement of the GTS and adjustments were made to the initially imprecise or disputed boundaries between the geological systems. The absolute ages were derived using radiogenic isotope geochronological techniques. Calculating an age for a rock or mineral using these techniques combines precise measurement of naturally occurring, radioactive isotopes and their stable decay products with the physical principle that the radioactive decay of the isotopes occurred at a constant, known rate. Because radiogenic ages are 'absolute' in the sense that the decay of a radioactive isotope primarily depends only on the passage of time, radiogenic ages for rocks found in one area of the world should, in principle, be directly comparable 'in time' to other rocks dated with similar methods in other areas of the world. Regardless of the geochronological technique used, the combination of relative and absolute ages has yielded the opportunity not only to generate geological time-scales, but also to determine the ages of rocks and geological structures, the timing of geological 'events', and, importantly, the rates at which geological processes occur.

Today, the primary techniques for relative dating of geological materials include biostratigraphy, palaeomagnetism and magnetostratigraphy, and chemostratigraphy (*see* **Palaeomagnetism**, **Magnetostratigraphy**, **Analytical Methods:** Fission Track Analysis). Of the absolute dating methods, radiogenic isotope geochronology, astronomical time calibrations, and dendrochronology (*see* **Dendrochronology**) are the most widely used. However, it is the rock type that usually dictates the geochronological technique appropriate for obtaining the rock's age. Thus, basic knowledge of the relative and absolute geochronological techniques is useful not only to select the appropriate method to date the rock, but also to interpret the age(s) produced, and to give a higher degree of confidence to comparisons made between geological ages and the processes to which they are linked.

Figure 1 Shang oracle bones. Precise calendars developed by early civilizations were based on calculations combining the rotation of Earth on its axis (day), Earth's revolution about the Sun (year), and the Moon's revolution about Earth (month). These individual astronomical cycles are neither constant nor synchronous with one another, and ancient peoples had to determine the appropriate lengths for days, months, and years so as to allow the seasonal cycles of the sun to coincide with the monthly cycles of the moon. The two oracle bones from the Shang Dynasty, dating back to the fourteenth century BCE in China, show that the Chinese had established the solar year at $365\frac{1}{4}$ days and a lunar cycle at $29\frac{1}{2}$ days; in this way, they had recognized and accounted for the differences in astronomical cycles between the Sun, Moon, and Earth in a consistent manner. The causes for these shifts in the astronomical cycles are now known to be primarily the gravitational forces acting between the different celestial bodies (see **Figure 6** for modern use of astronomical calibrations). Figure used with permission from M Douma. (see http://webexhibits.org for additional information on calendars).

Historical Perspective

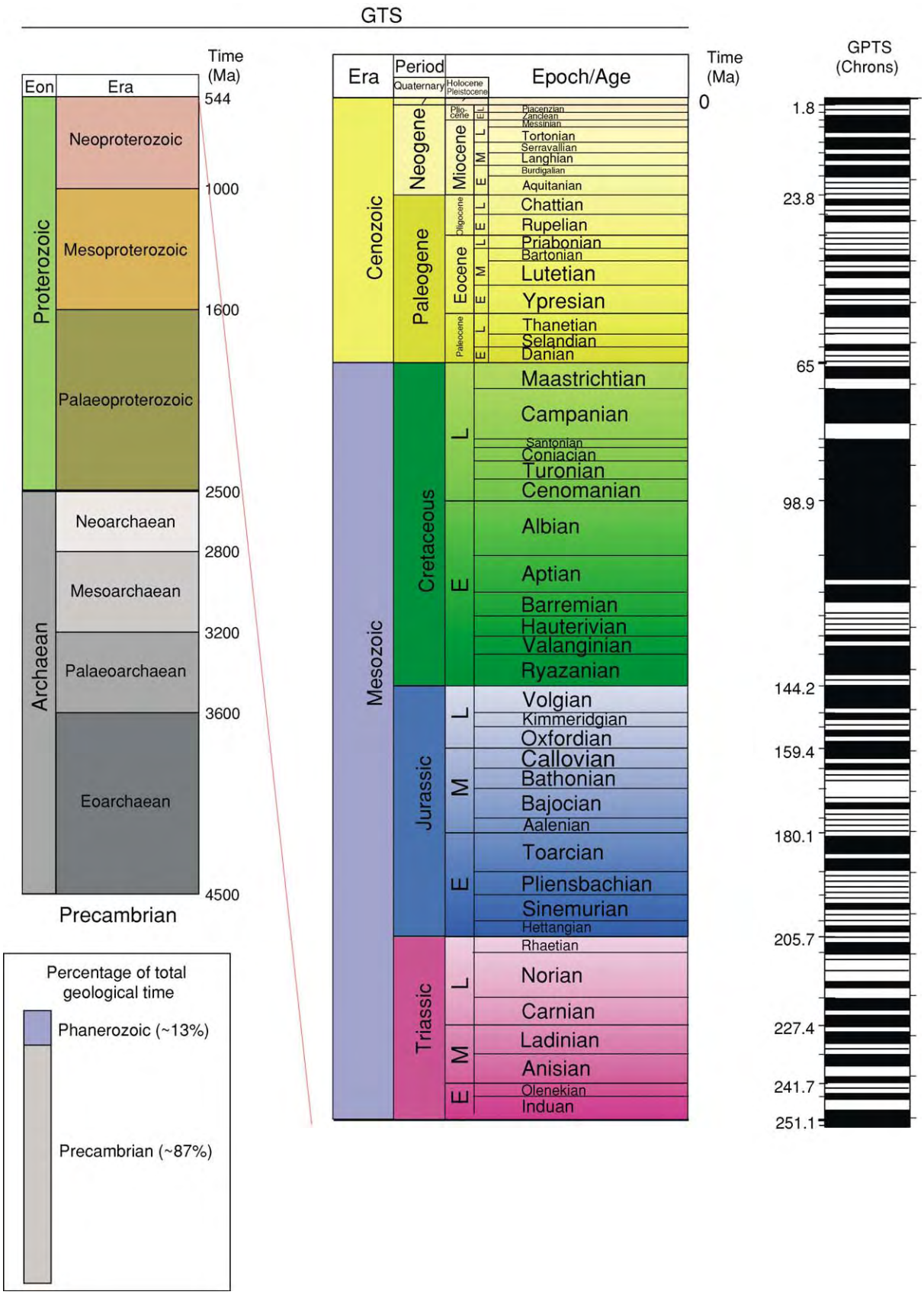
Relative Ages

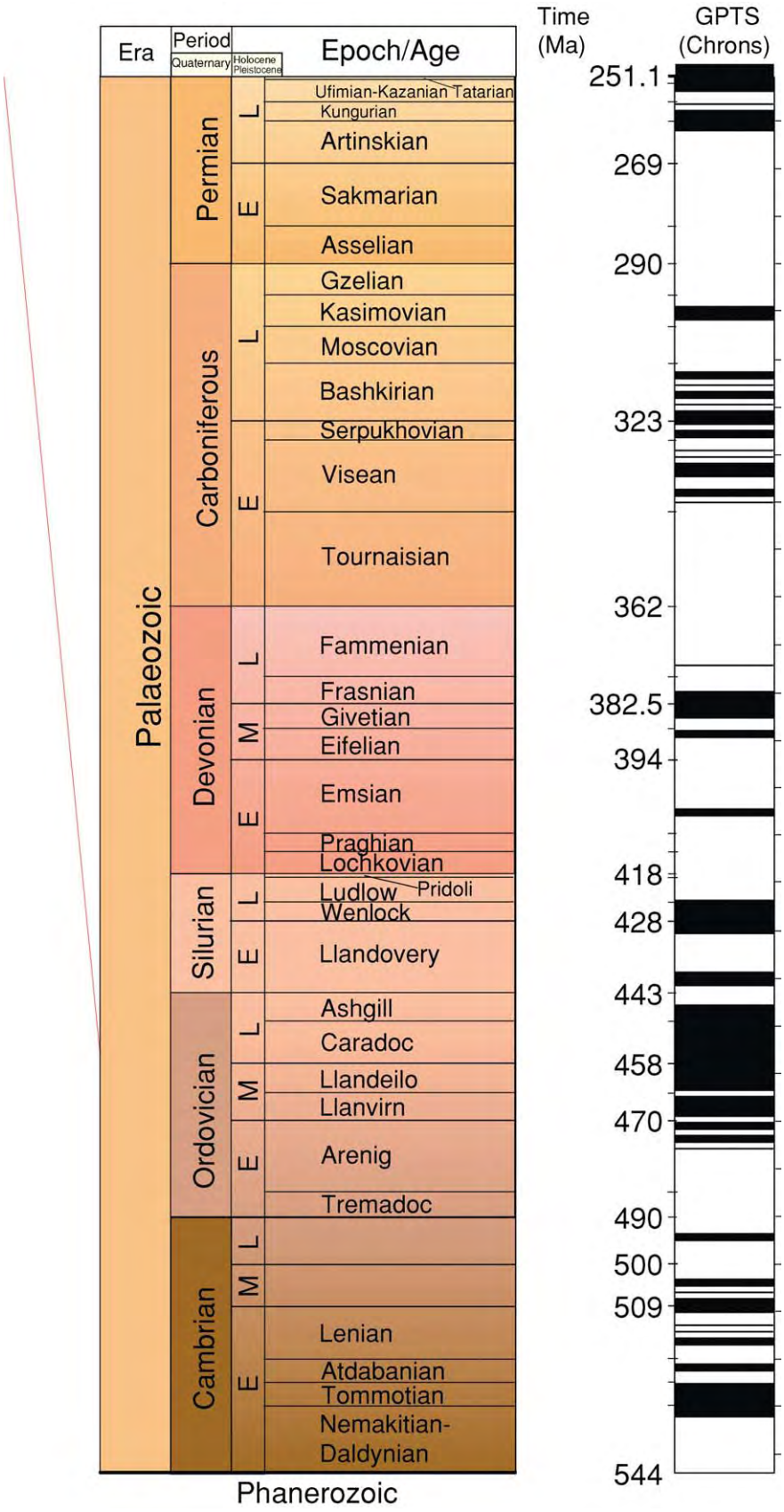
The conceptual background for modern geochronological techniques was established largely through the investigation of sedimentary rocks and the development of relative geochronology tools to study them. Studies of fossils and of the depositional order of sedimentary layers led to the principles of fossil succession and correlation and of superposition and relative chronology. These principles, accepted today as basic aspects in geoscience training, were crucial to the development of all relative geochronological techniques and can be attributed to the work of Steno in the seventeenth century, and of Smith, Cuvier, Brongniart, Lehmann, Füchsel, Pallas, and Arduino in the eighteenth century (*see* **Famous Geologists:** Smith; Cuvier).

Studies in the late eighteenth and early nineteenth centuries by Hutton, Lyell, Darwin, Murchison, and Sedgwick, among others, built on these early observations and the concept of relative geochronology. Murchison and Sedgwick named the Cambrian and Silurian systems in western Wales based on the systematic definition of sedimentary rock units using distinctive fossils; this procedure was fundamental for establishing all of the system subdivisions in the GTS. Hutton, Lyell, and Darwin each promoted the idea that Earth was very old. Darwin, in his first edition of the *Origin of Species*, estimated that about 300 million years (My) had passed since the end of the Mesozoic. Though today we know this estimate to be too high, Darwin's suggestion of this order of magnitude for the passage of geological time was important for carrying forward concepts such as evolution, and also for encouraging efforts to calculate absolute ages of rocks and of Earth (*see* **Famous Geologists:** Hutton; Lyell; Darwin; Murchison; Sedgwick).

Absolute Ages

In the latter half of the nineteenth century, calculations of Earth's age incorporated physical measurements in the field and laboratory and were, in this sense, quantitative; however, the methodologies initially employed were based on flawed assumptions that precluded their yielding accurate ages. Examples of these early attempts included calculating the rate of salinity increase over time for the world's oceans, and determining the age of the oldest sediment on Earth by estimating the total thickness and deposition rates for the sedimentary rock record. The salinity method assumed (incorrectly) that the world's oceans had initially been fresh and that no net exchange of sodium had occurred between seawater and rock.


GTS
Eon
Era
Time (Ma)
Proterozoic
Archaean
Neoproterozoic
Mesoproterozoic
Palaeoproterozoic
Neoarchaean
Mesoarchaean
Palaeoarchaean
Eoarchaean
544
1000
1600
2500
2800
3200
3600
4500
Precambrian
Era
Period
Epoch/Age
Time (Ma)
Cenozoic
Mesozoic
Neogene
Paleogene
Cretaceous
Jurassic
Triassic
Quaternary
Holocene
Pleistocene
Pliocene
Miocene
Oligocene
Eocene
Paleocene
Placenzian
Zanclean
Messinian
Tortonian
Serravallian
Langhian
Burdigalian
Aquitanian
Chattian
Rupelian
Priabonian
Bartonian
Lutetian
Ypresian
Thanetian
Selandian
Danian
Maastrichtian
Campanian
Santonian
Coniacian
Turonian
Cenomanian
Albian
Aptian
Barremian
Hauterivian
Valanginian
Ryazanian
Volgian
Kimmeridgian
Oxfordian
Callovian
Bathonian
Bajocian
Aalenian
Toarcian
Pliensbachian
Sinemurian
Hettangian
Rhaetian
Norian
Carnian
Ladinian
Anisian
Olenekian
Induan
0
GPTS (Chrons)
1.8
23.8
65
98.9
144.2
159.4
180.1
205.7
227.4
241.7
251.1
Percentage of total geological time
Phanerozoic (~13%)
Precambrian (~87%)


Era
Period
Epoch/Age
Time (Ma)
GPTS (Chrons)
Palaeozoic
Quaternary
Holocene
Pleistocene
Permian
Carboniferous
Devonian
Silurian
Ordovician
Cambrian
L
M
E
Ufimian-Kazanian
Tatarian
Kungurian
Artinskian
Sakmarian
Asselian
Gzelian
Kasimovian
Moscovian
Bashkirian
Serpukhovian
Visean
Tournaisian
Fammenian
Frasnian
Givetian
Eifelian
Emsian
Praghian
Lochkovian
Ludlow
Pridoli
Wenlock
Llandovery
Ashgill
Caradoc
Llandeilo
Llanvirn
Arenig
Tremadoc
Lenian
Atdabanian
Tommotian
Nemakitian-Daldynian
251.1
269
290
323
362
382.5
394
418
428
443
458
470
490
500
509
544
Phanerozoic

The calculations using the sedimentary rock record and deposition rates were inhibited by items such as missing sections and different rates of sedimentary deposition around the globe and throughout time. Both the salinity and sedimentation rate calculations yielded very low estimates for Earth's age (Table 2).

The first truly quantitative and influential effort to calculate an absolute age for Earth was made by the renowned physicist William Thomson (Lord Kelvin) during the middle to late nineteenth century. His concept was based on the idea that Earth had cooled from an originally molten state and was continuously losing heat from its surface through this cooling process. He made calculations for the length of time this process should have taken based on physical measurements of the rates of heat flow through a cooling body and of radiation of heat from the body's surface. Kelvin's calculations involved measurement of physical processes that were dependent only on the passage of time, so his conclusion that Earth was 20–40 My old fell technically within the realm of 'absolute' age determination and was widely accepted in the science community. This young age for Earth was at odds with the concepts put forward by Lyell and Darwin and produced an intense debate between Kelvin and promoters of evolution theory. However, a fundamental feature was missing from Kelvin's calculations – that of radioactive heat generation within Earth. At the time of Kelvin's initial calculations, radioactivity had not yet been discovered, so his equations greatly underestimated the amount of continuous heat generation within the crust and resulted in large underestimates of Earth's age (Table 2).

Henri Becquerel's discovery of radioactivity in 1896 launched the development of modern, radiogenic geochronological techniques. Radioactivity accounted for constant heat production from Earth's crust, as well as the production of heat from the sun, and eroded the premises of Kelvin's calculation. Soon after Becquerel had discovered that uranium (U) was radioactive, the radioactive properties of the elements radium, thorium, rubidium, and potassium (Ra, Th, Rb, and K) were also identified. The production of the isotopes helium (He), Th, and lead (Pb)

Table 1 Major time periods and definitions used for astronomical calendars[a]

Term	*Definition*	*Comment*
Solar or tropical year	Equal to 365.24219 days; the mean interval between two successive vernal equinoxes	The interval from one vernal equinox to the next may vary from this mean value by several minutes; this is because Earth's position in its orbit shifts slightly at the time of the equinoxes every year
Sidereal year	Equal to 365.25636 days; the time for Earth to make one revolution around the Sun, measured according to consecutive observations from Earth of the positions of stars	The precession of the equinoxes causes the sidereal year to be slightly variable and longer than the tropical year
Lunar or synodic month	Equal to 29.5305889 days; the mean period of time between new moons (or between exact conjunctions of the Sun and Moon); the lunar year contains 12 lunar months and is equal to 354.3671 days	The synchronization of calendar months with the lunar phases requires a combined sequence of months of 29 and 30 days in length; alternatively, as in Figure 1, the length of a month in days can be designated to be a non integer

[a]Time reference frames for astronomical calendars show the difficulties faced by early civilizations as they attempted to synchronize the movements of celestial bodies in a consistent calendar for measuring the passage of time. The cycles of the Moon and Sun relative to Earth change slowly with time, and a calendar year with an integral number of days cannot be perfectly synchronized to any of the astronomical reference frames. The astronomical formulas developed in the twentieth century to describe the changes in the orbital cycles of these celestial bodies yield the best approximations available for the length of any type of year (solar, sidereal, or synodic); however, the solutions to these formulas are descriptions of a constantly changing system and cannot be considered exact solutions. Thus the term 'absolute' age, in practice, when referring to astronomically calibrated time scales, is not strictly correct. A rather more general definition of absolute age is used herein.

Figure 2 The geological time scale (GTS; two coloured columns) and geomagnetic polarity time scale (GPTS; column with alternating black and white pattern) are often used together in geochronological studies. On the left side of the GPTS, the linear time scale hachures correspond to the epoch/age boundaries in the GTS; on the right side of the GPTS, the linear time scale hachures are placed at 10 My intervals. Note that different linear scales are used for denoting the Phanerozoic and Precambrian divisions. The true scale relationship between Precambrian and Phanerozoic times as a percentage of total geological time is shown on the lower left. Reproduced with permission from Eide EA (2002) Introduction plate reconstructions and integrated datasets. In: Eide EA (coord.) *BATLAS Mid Norway Plate Reconstruction Atlas with Global and Atlantic Perspectives*, pp. 8 17. Trondheim: Geological Survey of Norway.

Table 2 Selected historical review of estimates for the age of Earth[a]

Age of Earth (million years)	*Method*	*Year/author*
≃1973	Hindu chronology	*ca.* 120 150 BCE/priests
>300	Time for natural selection	1859/Darwin
100	Sediment thickness/deposition rate	1869/Huxley
<100	Cooling of Earth	1871/Kelvin
90	Sediment thickness/deposition rate	1890/de Lapparent
20 40	Cooling of Earth	1897/Kelvin
90	Salinity accumulation	1899/Joly
>1640	U Pb age of a Precambrian rock	1907/Boltwood
80	Sediment thickness/deposition rate	1908/Joly
>1300	Cooling of Earth	1917/Holmes
1600 3000	Decay of U to Pb in crust	1927/Holmes
3350	Terrestrial Pb isotope evolution	1947/Holmes
4000 5000	Radioactive isotope abundances	1949/Suess
4500 ± 300	Terrestrial Pb isotope evolution	1953/Houtermans
4540	Terrestrial Pb isotope evolution	1981/Tera

[a]In addition to these estimates, Jewish and Christian Biblical scholars from the second through seventeenth centuries suggested that the age of Earth ranged between 5000 and 7500 years, based on Julian, Gregorian, or Hebrew calendars. Some of the most well known sources for these age estimates include James Ussher, John Lightfoot, and St. Augustine. Regardless of the source, most ages of Earth published prior to the twentieth century were greatly underestimated. Research on the decay rates and processes for radioactive elements in Earth's crust finally led to more accurate calculations for Earth's age by the middle the 1900s. These calculations were based on the reconstruction of terrestrial Pb isotopic compositions from a primordial Pb reservoir, of composition similar to meteorites. The meteorite reference for these calculations has been the Canyon Diablo troilite.

from the radioactive decay of U was discovered at the start of the twentieth century by physicists Rutherford, Soddy, Strutt, Thomson, and Boltwood. Boltwood measured Pb–U ratios in unaltered minerals using a very rough estimate of the rate for the radioactive decay of U to Pb; he noted that the older the mineral, the greater the ratio (greater amount of the decay product, Pb). Rutherford applied the decay of U to He in a similar way to attempt to obtain ages for rock samples. At this important watershed for geochronological techniques, the realms of physics and geology became linked in a quantitative tool for measuring geological time. Through the first half of the twentieth century, great advances were made in understanding and applying radiogenic isotope geochronology to determine the ages of rocks and the age of Earth. Arthur Holmes was among those who made important contributions to the development of radiogenic geochronological techniques in this period (Table 2). Despite the progress through the middle of the twentieth century in producing absolute age constraints on Earth and its rocks, scientists lacked a cohesive Earth model in which to place the geological processes they were dating. In the 1950s and 1960s, the fundamental step was made in this regard through development of the plate tectonic paradigm and magnetic stratigraphy; plate tectonics and magnetostratigraphy also contributed significantly to development of high-fidelity time-scales and geochronological tools (*see* **History of Geology Since 1962**).

Oceanographic cruises in the 1950s identified the presence of alternating 'stripes' of high and low magnetic intensity on the ocean floor. This pattern was clarified in the 1960s marine geophysical work of Hess and Dietz, who proposed the theory of seafloor spreading, and Vine and Matthews, who suggested that new oceanic crust was generated at ocean ridges and became magnetized in the direction of Earth's magnetic field. The ocean-floor stripes revealed alternating periods in Earth's history during which the magnetic field had changed from normal to reversed polarity. When these theories were combined with new results from palaeomagnetic studies conducted on sedimentary and volcanic rocks onshore, a globally applicable pattern of periods of normal and reversed magnetic polarities was gradually defined (Figure 3). This magnetic 'stratigraphy' was a relative time-scale useful for global 'pattern matching' of magnetic anomalies and for relative geochronology. The potassium-argon (K–Ar) radiogenic isotope geochronological technique, employed since the 1950s, was used to determine ages for fine-grained basalts used in the palaeomagnetic studies and thus placed absolute age constraints on points in the magnetic anomaly stratigraphy. Through combination of palaeomagnetic and K–Ar dating methods, the magnetic stratigraphy became better defined and, eventually, globally correlatable in terms of geological time. From the 1970s to the present, ties between palaeomagnetism, radiogenic isotope geochronology,

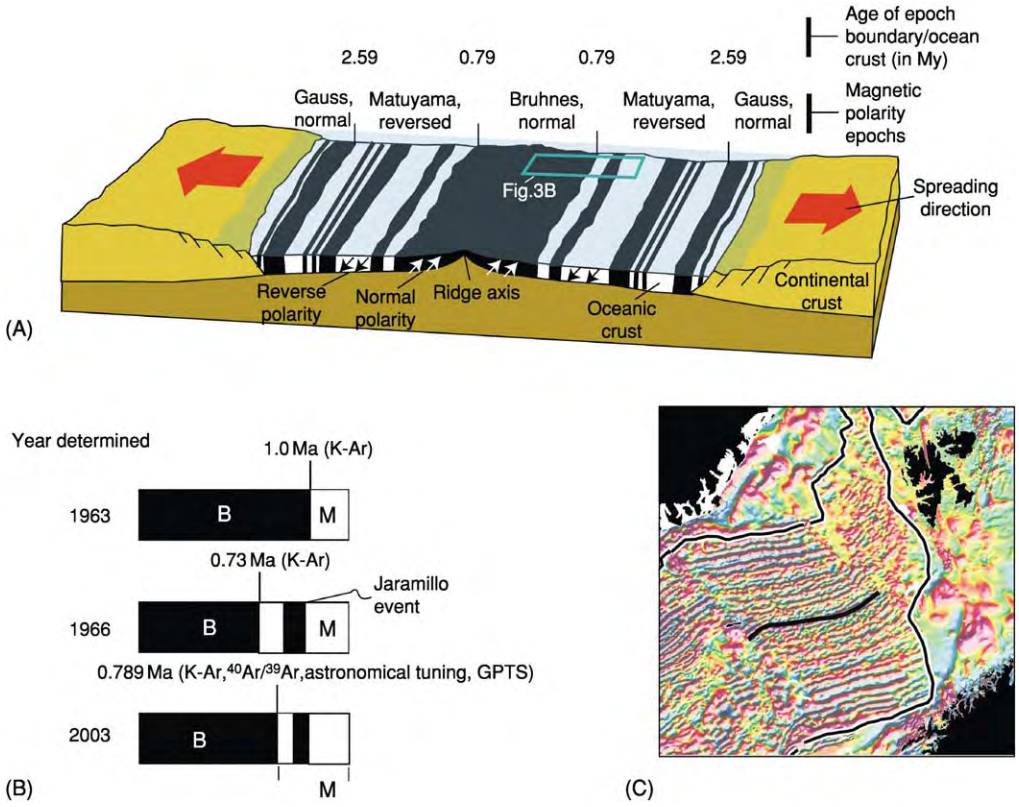


Figure 3 Seafloor spreading. (A) Genesis of mirror image, normal, and reversed magnetic polarity patterns in new oceanic crust, on either side of an oceanic ridge axis. The rifted continental margins yielded to new oceanic crust as seafloor spreading commenced. Alternating black (normal) and white (reversed) polarity patterns would normally be recorded by shipborne or satellite surveys. Historically, magnetic reversals were subdivided into major epochs (Bruhnes, normal; Matuyama, reversed; etc.); smaller normal and reversed 'events' were identified within these overall periods of normal or reversed polarity. Precise ages for these reversal epochs and, importantly, the boundaries between epochs were initially obtained with potassium argon (K Ar) geochronology. Refinements since the 1960s of the number and duration of magnetic reversals as well as their absolute ages have been accomplished by detailed comparison to biostratigraphy, the astronomically calibrated time scale, and ages from radiogenic isotope dating methods. (B)His torical refinement of the Bruhnes (B) Matuyama (M) boundary, where, in 1963, K Ar dating indicated the epoch boundary to be at ~1 Ma. The Jaramillo 'event' close to the Bruhnes Matuyama boundary had been discovered by 1966, and more precise K Ar dating placed the age of the epoch boundary at 0.73 Ma. By 2003, the combination of several dating methods, including K Ar and $^{40}Ar/^{39}Ar$ calibrations, astronomically calibrated time scales, and geomagnetic polarity time scales (GPTS), further refined the age of the boundary to a precise 0.789 Ma. (C) The magnetic anomaly map of the northern Atlantic Ocean between northern Norway, East Greenland, and Svalbard shows a real example of the alternating striped pattern of magnetic anomaly highs (red, normal polarity) and lows (blue, reversed polarity) on either side of the mid ocean ridge axis. The mid ocean ridge axis (trace identified with the single black line) separates a relatively symmetric, mirror image anomaly pattern in this part of the seafloor. Continent ocean boundaries are schematically indicated by thick black on white lines on the Norway and Greenland margins. (C) Reproduced with permission from Eide EA (coord.) *BATLAS Mid Norway Plate Reconstruction Atlas with Global and Atlantic Perspectives*, pp. 8 17. Trondheim: Geological Survey of Norway.

astronomically calibrated time-scales (ATSs), and biostratigraphy have facilitated definition of the geomagnetic polarity time-scale (GPTS) (Figure 2). Because of its tight calibration with these other methods, the GPTS provides the framework for most of the integrated time-scales presently in use for Jurassic and younger times (*see* **Plate Tectonics, Magnetostratigraphy**).

Today, the GTS, the GPTS, and the ATS have been intercalibrated for some geological time periods. Continued refinement and intercalibration of these time-scales will increase the possibility to make accurate age correlations for rocks and the geological events they represent. Important to recall is the fact that different geochronological techniques have been used to generate specific features of each time-scale, and that many techniques have particular geological time periods to which they are best suited; thus, complete intercalibration of these time-scales remains a challenging objective.

Relative Geochronological Techniques

Biostratigraphy

Methodology Biostratigraphy refers to correlation and age determination of rocks through use of fossils. Determining the environment in which the fossil species lived is inherent in this type of analysis. Theoretically, any fossil can be used to make physical correlations between stratigraphic horizons, but fossils that are best suited for making precise age correlations (time-stratigraphic correlations) represent organisms that (1) had wide geographic dispersal, (2) were short-lived, and/or (3) had distinct and rapidly developed evolutionary features by which they can now be identified. Fossils fulfilling these criteria are termed 'index' fossils. Both evolution and changes in local environment can cause the appearance or disappearance of a species, thus the time-significance of a particular index fossil must be demonstrated regionally through distinctions made between local environmental effects and time-significant events. Environmental effects may bring about the appearance/disappearance of a species because of local conditions, whereas time-significant effects may bring about the appearance/disappearance of a species because of evolution, extinction, or regional migration. Local environmental effects are not necessarily time significant and cannot be used in time correlations between different sedimentary units.

Application Fossils from the marine sedimentary record indicate existence of primitive life perhaps as early as 2.1 By ago, although the explosion of abundant life in the seas is usually tied to the start of the Palaeozoic era 544 million years ago (Ma). The continental sedimentary record indicates existence of plants and animals by Early Palaeozoic times, with recent indications of animals making forays from the seas onto land perhaps 530 Ma. Palaeozoic biostratigraphy, especially for the marine sedimentary record, is tied to precise, absolute ages for most period and stage boundaries, but gaps in the fossil record and/or the lack of isotopically datable rocks at key boundaries leave some discrepancies yet to be resolved. Biostratigraphy and fossil zone correlation are most precisely defined for the Mesozoic and Cenozoic eras; this is largely due to the ability to calibrate biostratigraphy not only with radiogenic isotope ages, but also with the GPTS and the ATS for these time periods.

Palaeomagnetism and Magnetostratigraphy

Methodology Earth's magnetic field, generated in the liquid outer core, undergoes periodic reversals, with magnetic reversal frequencies typically between 1 and 5 My. Some rock minerals (such as hematite or magnetite) may become magnetized in the same direction as Earth's magnetic field (normal or reversed), either when a magmatic rock cools or when sedimentary rocks are deposited. As geochronological tools, palaeomagnetism and magnetostratigraphy rely on determining the magnetic polarity, including magnetic declination and inclination, of the sample's remanent magnetic component. Palaeomagnetism uses these parameters to calculate a palaeomagnetic pole for the sampling site. An age for the pole is determined by matching the pole to a part of the apparent polar wander path (APWP) for that continent (Figure 4). Instead of using poles, magnetostratigraphy, as outlined previously, identifies a sequence of magnetic reversals in a sedimentary or volcanic section (Figure 2). The magnetostratigraphic profile is compared and matched to similar patterns in the GPTS and a chronology for the sampled interval is established. The absolute chronology of the GPTS is tied by radiogenic isotope methods, by calibration against the ATS, and/or by calibration with a well-defined biostratigraphic zone (*see* **Magnetostratigraphy, Palaeomagnetism**).

Application Palaeomagnetism and magnetostratigraphy are most successfully applied to fine-grained volcanic and sedimentary rocks; the latter include red beds, siltstones, mudstones, and limestones. Matching of palaeomagnetic poles to established APWPs yields imprecise ages for rocks, but is useful for reasonable, first-order age estimates, probably within about ± 10 My for Phanerozoic through Late Proterozoic rocks. The GPTS is most accurately refined through about 175 Ma because of the availability of marine magnetic anomaly profiles to which onshore data can be referenced; nonetheless, magnetic stratigraphy and the GPTS extend through the Palaeozoic to the earliest datable Cambrian sedimentary rocks (Figure 2). Well-constrained magnetostratigraphy yields very precise ages for the following reasons: (1) geomagnetic polarity reversals are rapid, globally synchronous events, and lend themselves well to global, time-significant correlations; (2) polarity reversals are not predictable and yield unique reversal patterns; (3) significant parts of the GPTS have been astronomically tuned, intercalibrated with detailed biostratigraphy, and/or constrained with absolute radiometric ages.

Chemostratigraphy

Methodology Non-radiogenic chemical geochronological tools for sedimentary rocks fall into one of

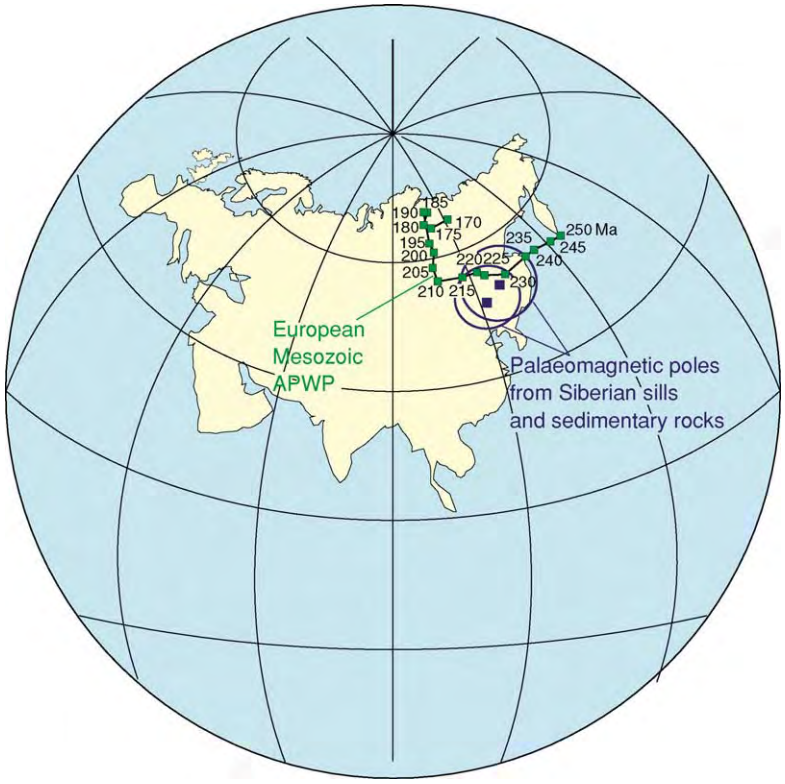


Figure 4 Palaeomagnetic poles from gabbroic sills and interleaved sedimentary rocks of initially unknown ages were obtained from a study in northern Siberia. The poles for these rocks were compared to the apparent polar wander path (APWP) for Europe in the Mesozoic. Well known ages are indicated in millions of years (Ma) for different segments of the APWP (designated with green squares). Within the uncertainty ellipses for the poles from the Siberian samples, the ages of the rocks were suggested to be between ~215 and 235 My. Subsequent radiogenic isotope age determinations on the sills confirmed this suggestion and refined the ages for the rocks to lie between 220 and 234 My.

three categories: pattern matching of time-stratigraphic shifts in stable isotope (O, C, or S) values and $^{87}Sr/^{86}Sr$ ratios, identification of siderophile element anomalies (Ir, Au, Pd, Pt, etc.), and chemical dating using amino acids. The principles for stable isotope methods are based on the fractionation of heavy and light isotopes of the stable elements O, C, and S. The heavy isotopes, ^{18}O, ^{13}C, and ^{34}S, are compared, respectively, to the lighter isotopes ^{16}O, ^{12}C, and ^{32}S. Stable isotopic compositions are reported as ratios (for example, $^{18}O/^{16}O$) relative to a standard for the same isotopic ratios. Processes causing fractionation of these isotopes depend primarily on temperature, isotope exchange reactions, and, in the case of S, change in oxidation state of sulphur compounds from action of anaerobic bacteria. The isotopic composition of Sr in sedimentary rocks is characterized by the $^{87}Sr/^{86}Sr$ ratio of the water from which the sediment precipitated; the water in the catchment area or in the ocean, in turn, will have an $^{87}Sr/^{86}Sr$ ratio that represents contributions from chemical weathering of rocks. Rocks of varying ages and different mineralogies have distinct $^{87}Sr/^{86}Sr$ ratios that will make different contributions of Sr to the water cycle. These contributions have been shown to vary over geological time in response to changes in the exposure and weathering of different landmasses.

For purposes of geochronology, the principle of 'pattern-matching' is also used with these isotopic methods (Figure 5). Measured isotopic ratios in a stratigraphic sample suite representing some interval of geological time yield a curve (or excursion pattern) that is compared to a global reference or supraregional curve for the same isotopes. The global reference curve must, in turn, be calibrated to an absolute timescale by some independent means, usually matching the stratigraphic section in question to another section that is tied either to the GPTS or to absolute ages.

Anomalously high concentrations of siderophile elements have been identified globally at three precisely determined time intervals: the Cretaceous–Tertiary boundary (65 Ma), the Eocene–Oligocene

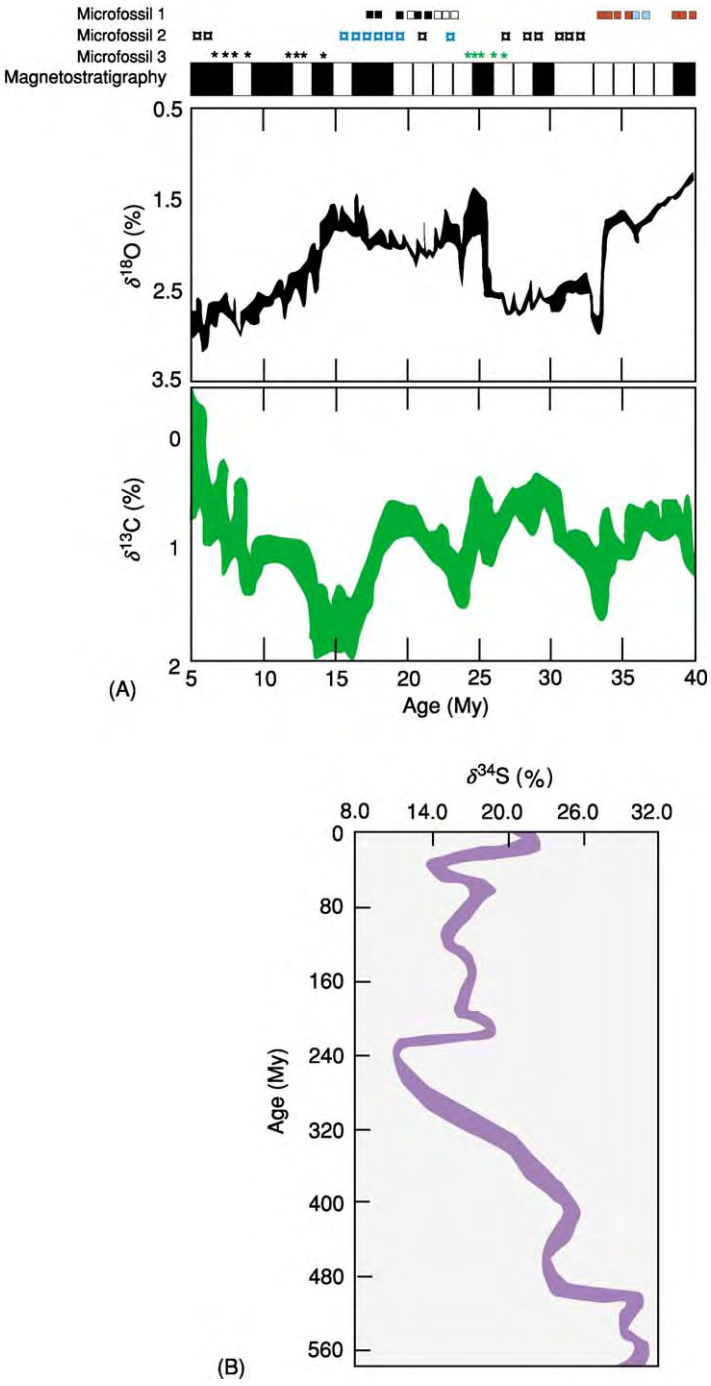


Figure 5 (A) Stable isotopes used in chemostratigraphy are commonly coupled with magnetostratigraphic and biostratigraphic information. In this fictive example, the stable isotope values for O and C were acquired for an entire sedimentary sequence of Cenozoic age. Magnetostratigraphy over the same zone may have revealed a pattern similar to that shown on the bar above the stable isotope curves, and this stratigraphy could then be correlated to the geomagnetic polarity time scale and used to calibrate the ages for the sedimentary column, which in this case spanned Pliocene through latest Eocene time. Biostratigraphy over the same stratigraphic column may have revealed a predominance of three types of microfossils, with different species within each microfossil group identified (designated here with different coloured symbols). Biostratigraphy might also be used to tie together and calibrate the stable isotope curves and make fine adjustments to ages determined with the magnetostratigraphic profile. Especially interesting would be to attempt to link any significant excursions in the isotope curves, either to changes observed in the microfossil distribution or to a specific time boundary. (B) Stable isotope stratigraphy can also be used over a larger time span for more regional or global correlations. This isotope curve for sulphur shows a marked change at about 240 Ma following a steady decrease through the Palaeozoic.

boundary (33.7 Ma), and 2.3 Ma. Other anomalies – specifically, spikes in iridium concentrations in sedimentary sequences – have been suggested at the Triassic–Jurassic boundary and at the Devonian–Carboniferous boundary. These anomalous concentrations have been associated with catastrophic events, usually meteor impacts or massive volcanic eruptions, and faunal crises or mass extinctions. Because of their global nature, limited duration, and precisely defined ages, anomalous siderophile concentrations can serve as indirect dating tools in sedimentary sequences (*see* **Impact Structures**).

The amino acid racemization (AAR) method uses the asymmetry of isomeric forms of several amino acids in fossil skeletal material to determine the time since the start of racemization. Racemization is the reversible conversion of one set of amino acid isomers to another set of isomers and begins with death of the organism. Sample materials are chemically treated and the amino acid types and isomer ratios are determined through chromatography methods. These ratios are used to calculate the time since the start of racemization through a formula containing a sample-site constant for the racemization rate. Because the racemization rate depends on external factors such as temperature, pH, and moisture, the rate varies between one sample site and another and must be calibrated for each site and each sample. This usually involves calibration against other samples (from the same sites) that have been dated by other methods.

Application Oxygen isotope stratigraphy may be applied to planktonic foraminiferal tests in pelagic sediments that are at least 1 My old. Sulphur isotopes are most commonly used to date marine evaporites with ages of deposition extending through ~650 Ma. Carbon isotopes may be used to date marine evaporites, marine carbonates, and (metamorphosed) marbles through Neoproterozoic age. Similarly, strontium, which substitutes readily for calcium, can also be used to date marine carbonates, apatite in marine sediments, and marbles through the Neoproterozoic. All of the isotope methods generally require samples that have been relatively unaltered by postdepositional events such as erosion, bioturbation, metamorphism, or recrystallization during diagenesis. Notably, work with metamorphosed marbles has indicated that C and Sr isotopes may maintain their original sedimentary deposition ratios despite having undergone extreme changes in pressure, temperature, and deformation subsequent to deposition.

Siderophile element anomalies are confined to the sedimentary rock record; the most well-documented anomaly is at the Cretaceous–Tertiary boundary (*see* **Mesozoic:** End Cretaceous Extinctions). The AAR method is restricted primarily to dating Holocene foraminifers extracted from pelagic sediments, although ages have also been determined for coprolites and mollusc shells.

Absolute Geochronological Techniques

Radiogenic Isotope Techniques

Methodology The natural decay of a radioactive isotope to a stable isotope occurs at a regular rate that is described by the decay constant (λ). The decay process is defined by an exponential function represented by the decay 'half-life' ($t_{1/2}$); the half-life is equivalent to the amount of time necessary for one-half of the radioactive nuclide to decay to a stable nuclide form. Radiogenic isotope techniques use this principle to calculate the age of a rock or mineral through measurement of the amount of radioactive 'parent' isotope and stable 'daughter' isotope in the sample material. The parent/daughter ratio and the decay constant for that isotope series are used to calculate how much time had to elapse for all of the stable daughter isotope to have been produced from an initial reservoir of radioactive parent isotope in the material (Table 3). This calculation presumes (1) no net transfer of radiogenic parent, stable daughter, and/or intermediate radioactive isotopes in or out of the sample material (mineral or rock) since time zero, (2) no unknown quantity of daughter isotope in the sample at time zero, and (3) that decay constants have not changed over the history of Earth. Many radiogenic isotope techniques are presently used to determine the ages of geological materials; the choice of appropriate isotopic system to determine an age of a sample depends primarily on the composition of the sample material, the geological 'event' or 'process' to be dated, and the sample's age. The latter is directly linked to the half-life of the isotope system: radionuclides with long half-lives can be used to date very old samples, whereas those with shorter half-lives are restricted to dating younger rocks. In addition to the naturally occurring radioactive isotopes, a number of nuclear reactions of cosmic rays with gas molecules will produce radionuclides, the so-called cosmogenic radionuclides. The most long-lived of these can be used for age determinations based on principles similar to those outlined for the other radioactive isotopes.

Applications The methods routinely used to date terrestrial metamorphic or igneous rocks and their minerals include techniques utilizing U/Th/Pb, Pb/Pb, Sm/Nd, Lu/Hf, Re/Os, Rb/Sr, K–Ar, and Ar/Ar (Table 3). All of these isotopes have half-lives >1 By,

Table 3 Common radiogenic isotope geochronological techniques

Method	*Radioactive parent*	*Stable daughter*	*Intermediate products*[a]	*Decay scheme*	*Half life (years)*	*Sample material*	*Typical geological 'events' dated*	*Comments*
U/Th/Pb, Pb/Pb	^{238}U	^{206}Pb	From ^{238}U: ^{234}Th, ^{234}Pa, ^{234}U, ^{230}Th, ^{226}Ra, ^{222}Rn, ^{218}Po, ^{218}At, ^{218}Rn, ^{214}Po, ^{210}Pb, ^{210}Bi, ^{210}Po	Chain: ^{238}U → ^{206}Pb, ^{235}U → ^{207}Pb, ^{232}Th → ^{208}Pb	^{238}U 4.468×10^9, ^{235}U 0.7038×10^9, ^{232}Th 14.01×10^9	Zircon, thorite, monazite, apatite, xenotime, titanite, uraninite, thorianite	Crystallization age (from melt or from medium to high metamorphic grade); age of Earth	U and Th are concentrated in the liquid phase and are typically incorporated in more silica rich fractions; half lives of the parent isotopes are much longer than those of intermediate products; Pb isotopes alone in rocks without U or Th can be used to calculate 'model ages' (with information on crustal growth)
	^{235}U	^{207}Pb	From ^{235}U: ^{231}Th, ^{231}Pa, ^{227}Ac, ^{227}Th, ^{223}Ra, ^{219}Rn, ^{215}Po, ^{214}At, ^{211}Bi, ^{211}Po					
	^{232}Th	^{208}Pb	From ^{232}Th: ^{228}Ra, ^{228}Ac, ^{228}Th, ^{224}Ra, ^{220}Rn, ^{216}Po, ^{212}Pb, ^{212}Bi, ^{212}Po, ^{208}Pb	Decay schemes produce alpha (^{4}He) particles; used for (U/Th)/He dating				
Sm/Nd	^{147}Sm	^{143}Nd	None	Simple: ^{147}Sm → ^{143}Nd (alpha decay)	1.06×10^{11}	Garnet, pyroxene, amphibole, plagioclase; mafic and ultramafic igneous and metamorphic whole rocks; lunar rocks	Crystallization age (from melt or from medium to high metamorphic grade)	Ages calculated from analysis of isotopes in separated minerals or cogenetic rocks; Sm and Nd are rare earth elements that tend to be less mobile during metamorphism and weathering
Lu/Hf	^{176}Lu	^{176}Hf	None	Branched: ^{176}Lu → ^{176}Hf (gamma ray emission); ^{176}Lu → ^{176}Yb (electron capture)	3.54×10^{10}	Apatite, garnet, monazite, zircon, xenotime, meteorites, lunar rocks	Meteorite formation; high grade metamorphism; igneous crystallization	Can also be used for information on differentiation of the mantle and crustal growth; ^{176}Yb branch of decay can be ignored for purpose of geochronology

Continued

Table 3 Continued

Method	*Radioactive parent*	*Stable daughter*	*Intermediate products*[a]	*Decay scheme*	*Half life (years)*	*Sample material*	*Typical geological 'events' dated*	*Comments*
Re/Os	^{187}Re	^{187}Os	None	Simple: ^{187}Re → ^{187}Os (beta particle emission)	4.56×10^{10}	molybdenite, osmiridium, laurite, columbite, tantalite, Cu sulphides; ores, meteorites	Ore deposit formation; iron meteorite formation	Enriched in metallic and sulphide phases; relatively depleted in silicates
Rb/Sr	^{87}Rb	^{86}Sr	None	Simple: ^{87}Rb → ^{86}Sr (beta particle emission)	4.88×10^{10}	Mica, feldspar, leucite, apatite, epidote, garnet, ilmenite, hornblende, pyroxene, clay minerals, some salts; felsic whole rocks, meteorites	Crystallization age (from melt or metamorphism); cooling (after high grade 'event'); diagenesis	Because Rb and Sr have close relationships to K and Ca, respectively, the method is especially useful for study of granitic rocks
K Ar, ^{40}Ar/^{39}Ar	^{40}K	^{40}Ar	None	Branched: ^{40}K → ^{40}Ca (beta emission); ^{40}K → ^{40}Ar (beta emission and electron capture)	1.25×10^{10}	Mica, feldspar, feldspathoids, amphibole, illite, volcanic rocks, lunar rocks, low grade metamorphic rocks, glass, salts, clay minerals, evaporites	Crystallization of quickly cooled igneous rocks; cooling of metamorphic and plutonic rocks	K Ar method involves splitting the sample to measure K and Ar; ^{40}Ar/^{39}Ar uses ^{39}Ar as a proxy for K and measures only Ar isotopes, with no sample splitting; the ^{40}Ar/^{39}Ar method is commonly used today
Carbon 14	^{14}C	^{14}N	^{14}C produced in atmosphere by collision of thermal neutrons (from cosmic rays) with ^{14}N; ^{14}C is oxidized rapidly and radioactive CO_2 enters the carbon cycle; radioactive ^{14}C decays	^{14}C → ^{14}N	5700	Organic matter: wood, charcoal, seeds, leaves, peat, bone, tissue, mollusc shells	Time since the organic material ceased to take up carbon	Dendrochronology and varve chronology are often used in carbon 14 dating to account for secular variation in the ^{14}C content in the atmosphere

[a]Note that the U Th Pb decay series involves numerous intermediate radioactive isotopes with short half lives ('chain' decay); only the direct intermediate products are listed here (products from branched decay have not been listed).

so the samples can be used to date Earth's oldest geological materials and events. Lunar and cosmogenic materials have also been dated with some of the same methods. The relatively shorter half-life of the K–Ar decay series, as well as the very short half-lives of the intermediate nuclides in the U and Th decay series, allow these isotope systems to be used for dating certain geological materials of Pleistocene (the U-series nuclides) and Holocene (the K–Ar and Ar/Ar methods) ages. Of the cosmogenic radionuclides, the most well known is probably carbon-14. The carbon-14 method is used to date organic materials; ^{14}C has a half-life of ~5700 years and is restricted to materials less than about 100 000 years old (Table 3). Aside from ^{14}C, other cosmogenic radionuclides include ^{10}Be, ^{26}Al, ^{36}Cl, ^{41}Ca, ^{53}Mn, ^{81}Kr, and ^{129}I; these can be used for dating relatively young materials (on the order of several 100 000 years for Ca and Kr and up to 1 My or more for Be, Al, Cl, Mn, and I). Though not treated in detail here, these isotopes can be applied to date a range of materials, including Quaternary sediments, ice, manganese nodules, groundwater, and soils, and to determine the age of exposure of terrestrial land surfaces and meteorites (*see* **Analytical Methods:** Fission Track Analysis).

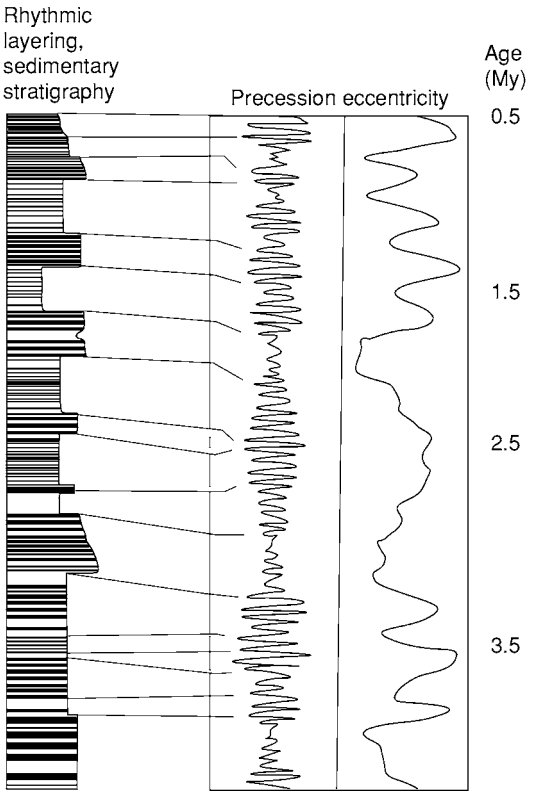


Figure 6 Astronomically calibrated time scales attempt to resolve the long term gravitational perturbations in Earth's orbit about the Sun. The mathematical solutions for the cyclicity of these perturbations are projected backward in time to determine the geological age of seasonal (solar) cycles preserved in the sedimentary rock record. Most astronomical calibrations define solutions for the precession and eccentricity of Earth's orbit. In this example, cyclical sedimentation patterns (alternating dark and light sedimentary layers) in a fictitious marine sequence were carefully logged, as on the left hand column. The log is matched to the calculated solutions for orbital precession and eccentricity that are tied to absolute time. Where possible, the stratigraphic column may also be tied to magnetostratigraphic, biostratigraphic, and/or radiogenic isotope geochronology data.

Astronomically Calibrated Time-Scales

Methodology Perturbations in the orbit of Earth about the sun are generated by gravitational interactions between Earth and the sun, moon, and other celestial bodies. These orbital perturbations cause cyclical climatic changes that are recorded in some sedimentary rocks. This principle was recognized by G K Gilbert in the nineteenth century, and he noted the potential to use this climatically driven, sedimentary cyclicity to place age constraints on certain parts of the rock record. Since Gilbert's time, astronomically calibrated time-scales have generated astronomical solutions for these perturbations in Earth's orbit that match sedimentary cycles recognized in nature, such as glacial varve sequences (Figure 6). These gravity-induced perturbations apply specifically to the obliquity of Earth's orbit, Earth's axial precession, and the eccentricity of Earth's orbit about the sun. Obliquity refers to the angle between Earth's axis of rotation and the orbital plane, whereas precession is the movement ('wobble') of the rotation axis about a circular path that describes a cone. Eccentricity is the elongation of Earth's orbit about the sun; this varies between a circular and an elliptical shape. The main periods of eccentricity of Earth's orbit are 100 000 and 413 000 years. The obliquity of Earth's axis has a main period of 41 000 years and precession of the axis has a main period of 21 000 years. Because the astronomically calibrated time-scales are based only on factors related to Earth's orbit about the sun, they are the only truly 'absolute' time-scales, following the strict definition of this word, and are mainstays for tying together or intercalibrating the other time-scales (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)).

Applications The geologically short periodicity of Earth's orbital perturbations has allowed calibration of precise astronomical time-scales for the past 15 My. Climate changes associated with ice ages have been the most easily recognized events in the rock record and the astronomical calibration of the Plio-Pleistocene time-scale remains one of the best. Although the Miocene-and-younger time-scales have been based primarily on the marine rock record,

continental sedimentary sections have increasingly been incorporated in these calibrations. Work with astronomically calibrated lacustrine sections of Triassic–Jurassic age has demonstrated that older rocks can also be anchored to the astronomical time-scale.

Dendrochronology

Methodology and Applications Dendrochronology applies the nonsystematic, climate-dependent variations in the thickness of annual tree rings of particular tree species to determine very exact dates for young events. Although restricted to use on Holocene samples, the high precision of the method (trees produce one ring per year, and uncertainties in ages determined with the method are usually ±1 year) has also been used to calibrate carbon-14 ages (see also Table 3).

Future Considerations

Geochronology furnishes the temporal framework for the study of geologic processes, giving data necessary to evaluate the rates, quantity, and significance of different rocks and geological 'events'. Both relative and absolute ages are important in this regard and should be viewed as complementary methods through which different rock types may be correlated in time. Today, a big challenge facing geochronologists is the intercalibration of the various time-scales. As part of this work, geologists working with radiogenic isotopes are attempting to refine the decay constants for a number of the commonly used radiogenic isotope dating methods. Inaccurate decay constants would clearly affect the accuracy of an age for a rock determined with a particular isotope system, and would have corresponding spin-off effects for ties made to magnetostratigraphic, biostratigraphic, chemostratigraphic, and astronomically calibrated datasets. Intercalibration of the various time-scales back through Mesozoic and Palaeozoic times will probably incorporate all of these methods, with extension of astronomical calibrations to the Palaeozoic probably involving 'floating' astronomical time-scales intercalibrated with the continually updated and refined GPTS and GTS.

Glossary

decay constant A number describing the probability that a radioactive atom will decay in a unit time.

half-life The time required for half of a quantity of radioactive atoms to decay.

isotopes Atoms with the same number of protons (= the same element), but a different number of neutrons (= different mass).

radioactive decay The spontaneous disintegration of certain atoms whereby energy is emitted in the form of radiation; a new, stable atom is the result.

siderophile An element preferring a metallic phase, with a weak affinity for oxygen or sulphur.

See Also

Analytical Methods: Fission Track Analysis. **Conservation of Geological Specimens**. **Creationism**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Dendrochronology**. **Famous Geologists:** Cuvier; Darwin; Hutton; Lyell; Murchison; Sedgwick; Smith; Steno. **Magnetostratigraphy**. **Mesozoic:** End Cretaceous Extinctions. **Palaeomagnetism**. **Palaeozoic:** Cambrian; End Permian Extinctions. **Plate Tectonics**. **Time Scale**.

Further Reading

Butler RF (1992) *Palaeomagnetism: Magnetic Domains and Geologic Terranes*. Cambridge, MA: Blackwell Scientific Publications.

Cox A (ed.) (1973) *Plate Tectonics and Geomagnetic Reversals*. San Francisco, CA: WH Freeman and Company.

Dalrymple BG (1991) *The Age of the Earth*. Palo Alto, CA: Stanford University Press.

Dickin AP (1995) *Radiogenic Isotope Geology*. Cambridge: Cambridge University Press.

Doyle P, Bennett MR, and Baxter AN (1994) *The Key to Earth History: An Introduction to Stratigraphy*. Chichester: John Wiley and Sons.

Eicher DL (1976) *Geologic Time*, 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

Eide EA (2002) Introduction plate reconstructions and integrated datasets. In: Eide EA (coord.) *BATLAS Mid Norway Plate Reconstruction Atlas with Global and Atlantic Perspectives*, pp. 8 17. Trondheim: Geological Survey of Norway.

Faure G (1986) *Principles of Isotope Geology*, 2nd edn. New York: John Wiley and Sons.

Geyh MA and Schleicher H (1990) *Absolute Age Determination: Physical and Chemical Dating Methods and Their Application*. Berlin: Springer Verlag.

Hilgen FJ, Krijgsman W, Langereis CG, and Lourens LJ (1997) Breakthrough made in dating of the geological record. *EOS* 78(28): 285, 288 289.

Lewis C (2000) *The Dating Game One Man's Search for the Age of the Earth*. Cambridge: Cambridge University Press.

Renne PR, Deino AL, Walter RC, *et al.* (1994) Intercalibration of astronomical and radioisotopic time. *Geology* 22: 783 786.

Gravity

J R Smallwood, Amerada Hess plc, London, UK

Introduction

The law of gravitational attraction between objects was deduced by Isaac Newton in the late seventeenth century. His 'inverse square' law stated that the force attracting two objects was proportional to the masses of the two objects and inversely proportional to the square of the distance between them (Table 1). Since the mass of the Earth is so great relative to the mass of objects on its surface, attraction of objects towards the Earth, i.e., their response to the Earth's gravity field, is often an important factor affecting geological processes. Measurement of the gravity field of the Earth is in itself a useful tool for investigating the sub-surface, as mass variations below the surface cause variations in the gravity field. The measurement of the shape of the Earth and its mass distribution have been important to defining the baseline gravity field from which deviations can be measured, as usually the anomaly rather than the overall field strength is useful for geological applications. There are now many ways of acquiring gravity data on land, sea, air, and from space, appropriate to the many scales on which gravity studies can be applied. Gravity variations over thousands of kilometres can be used for studies of mantle convection, variations over hundreds and tens of kilometres are relevant for studies such as lithospheric flexure, plate tectonics (*see* **Plate Tectonics**), crustal structure, and sedimentary basin development, hydrocarbon (*see* **Petroleum Geology: Exploration**) and mineral exploration (*see* **Mining Geology: Exploration**), while gravity variations over tens of metres can be used in civil engineering applications.

The Earth's Shape and its Gravity Field

The gravitational potential of a perfectly uniform sphere would be equal at all points on its surface. However, the Earth is not a perfect sphere; it is an oblate spheroid, and has a smaller radius at the poles than at the equator. Surveys in the early eighteenth century, under the direction of Ch-M de La Condamine and M de Maupertius found that a meridian degree measured at Quito, Equador, near the equator, was about 1500 m longer than a meridian degree near Tornio, Finland, near the Arctic circle. Subsequently, various standard reference spheroids or ellipsoids have been proposed as first-order approximations to the shape of the Earth, such as the World Geodetic System 1984 (Table 1). Given such an ellipsoid, a gravity field can be calculated analytically as a function of latitude. For example, a reference gravity formula was adopted by the International Association of Geodesy in 1967 (IGF67, Table 1), and another introduced in 1984 (WGS84, Table 1).

The mean density of the Earth, which is fundamental to the calculation of gravitational attraction, was first estimated following an experiment in 1775 by the Rev. Neville Maskelyne, using a technique suggested by Newton. If the Earth was perfectly spherical and of uniform density, then a plumbline would point down towards the centre of the Earth because of the force of gravity on the bob. However, any nearby mass would deflect the plumbline off this 'vertical'. Maskelyne and his co-workers measured plumb-bob deflections on the Scottish mountain, Schiehallion (Figure 1). They discovered that the mountain's gravitational pull deflected the plumb line by 11.7 seconds of arc. This allowed Charles Hutton to report in 1778 that the mean density of the Earth was approximately 4500 kg m ^{3}. This density value leads to an estimate of the mass of the Earth of about 5×10^{24} kg, not far from the currently accepted value of 5.97×10^{24} kg. The Schiehallion experiment had another distinction, in that in order to calculate the mass and centre of gravity of the mountain a detailed survey was carried out, and the contour map was invented by Hutton to present the data.

Since the mass of the Earth is not distributed uniformly, the real gravity field does not correspond to that calculated for an ellipsoid of uniform density. The 'geoid' is a surface which is defined by points of equal gravitational potential or equipotential (Table 1), which is chosen to coincide, on average, with mean sea-level. The geoid is not a perfect ellipsoid, because local and regional mass anomalies perturb the gravitational potential surface in their vicinity by several tens of metres. For example, a seamount on the ocean floor, which is denser than the surrounding seawater, will deflect the geoid downwards above it. 'Geoid anomalies' are defined as displacements of the geoid above or below a selected ellipsoid. The concept of the geoid as the global mean sea-level surface can be extended across areas occupied by land. This provides both a horizontal reference datum and a definition of the direction of the vertical, as a plumbline will hang perpendicular to the geoid.

Table 1 Gravity formulae

Quantity	*Formula*	*Constants and variables*	
Gravitational Force between two masses, F	$F \quad \frac{GMm}{r^2}$	G	Gravitational or Newtonian constant, $6.67 \times 10^{-11}\,m^3kg^{-1}s^{-1}$
		M	Mass of body (Mass of earth approx. 5.97×10^{24} kg)
		m	Mass of second body
		r	Distance
Gravitational Acceleration, a	$a \quad \frac{GM}{r^2}$		As above
Gravitational Potential, V	$V \quad \frac{GM}{r}$		As above
(Vertical) Gravity anomaly above a buried sphere, δg_z See **Figure 6**	$\delta g_z \quad \frac{4G\Delta\rho b^3 h}{3(x^2+h^2)^{3/2}}$	$\Delta\rho$	Density contrast
		b	Radius of sphere
		h	Depth of sphere
		x	Horizontal distance
International Gravity Formula 1967 Gravitational acceleration, g_t	$g_t \quad g_0(1+\alpha \sin^2\lambda+\beta \sin^4\lambda)$	g_0	Mean gravititational acceleration at equator, $9.7803185\,ms^{-2}$
		α	5.278895×10^{-3}
		β	2.3462×10^{-5}
		λ	Latitude
WGS84 Ellipsoidal Gravity Formula Gravitational acceleration, g_t	$g_t \quad \frac{g_0(1+d \sin^2\lambda)}{\sqrt{(1-e \sin^2\lambda)}}$	g_0	Mean gravititational acceleration at equator, $9.7803267714\,ms^{-2}$
		d	$1.93185138639 \times 10^{-3}$
		e	$6.6943999103 \times 10^{-3}$
		λ	Latitude
WGS Formula atmospheric correction, δg_t	$\delta g_t \quad 0.87 \times 10^{-5} \exp(-0.116h^{1.047})$	h	Elevation
Latitude correction for relative gravity measurements, δg_L	$\delta g_L \quad 8.12 \times 10^{-5} \sin 2\lambda \; \delta l$	δl	Distance in N–S direction between readings
		λ	Latitude
Bouguer plate correction, δg_B	$\delta g_B \quad 2\pi\rho Gh$	ρ	Bouguer correction density
		h	Elevation
Free air correction, δg_{FA}	$\delta g_{FA} \quad 308.6\,h$	h	Elevation
Free air anomaly, g_{FA}	$g_{FA} \quad g_{obs} - g_t + (\delta g_L + \delta g_{FA})$	g_{obs}	Observed gravity
Bouguer anomaly, g_B	$g_B \quad g_{obs} - g_t + (\delta g_L + \delta g_{FA} - \delta g_B + \delta g_T)$	δg_T	Terrain correction
Flattening factor for ellipsoid, f	$f \quad \frac{a-c}{c} \quad \frac{1}{298.26}$	a	Equatorial radius of Earth, 6378.14 km
		c	Polar radius of Earth, 6356.75 km

Measurement of Gravity

The first measurements of Earth's gravity, by timing the sliding of objects down inclined planes, were made by Galileo, after whom gravitational units were named. 1 Gal is $10^{-2}\,m\,s^{-2}$, and the gravitational acceleration at the Earth's surface is about 981 Gal. For convenience in geophysical studies of gravity anomalies, the mGal is usually used, or for local surveys 'gravity units' (g.u.) where 1 mGal = 10 g.u. Gravity may be measured as an absolute or relative quantity.

Classically, absolute gravity has been measured with a pendulum consisting of a heavy weight suspended by a thin fibre. The period of the oscillation is a function of gravitational acceleration and the length of the pendulum. H Kater designed a compound, or reverse, pendulum in 1815, that allowed some instrument-dependent factors to be cancelled out. The instrument was superceded by methods based on observations of falling objects. In a development of the free-fall method, a projectile is

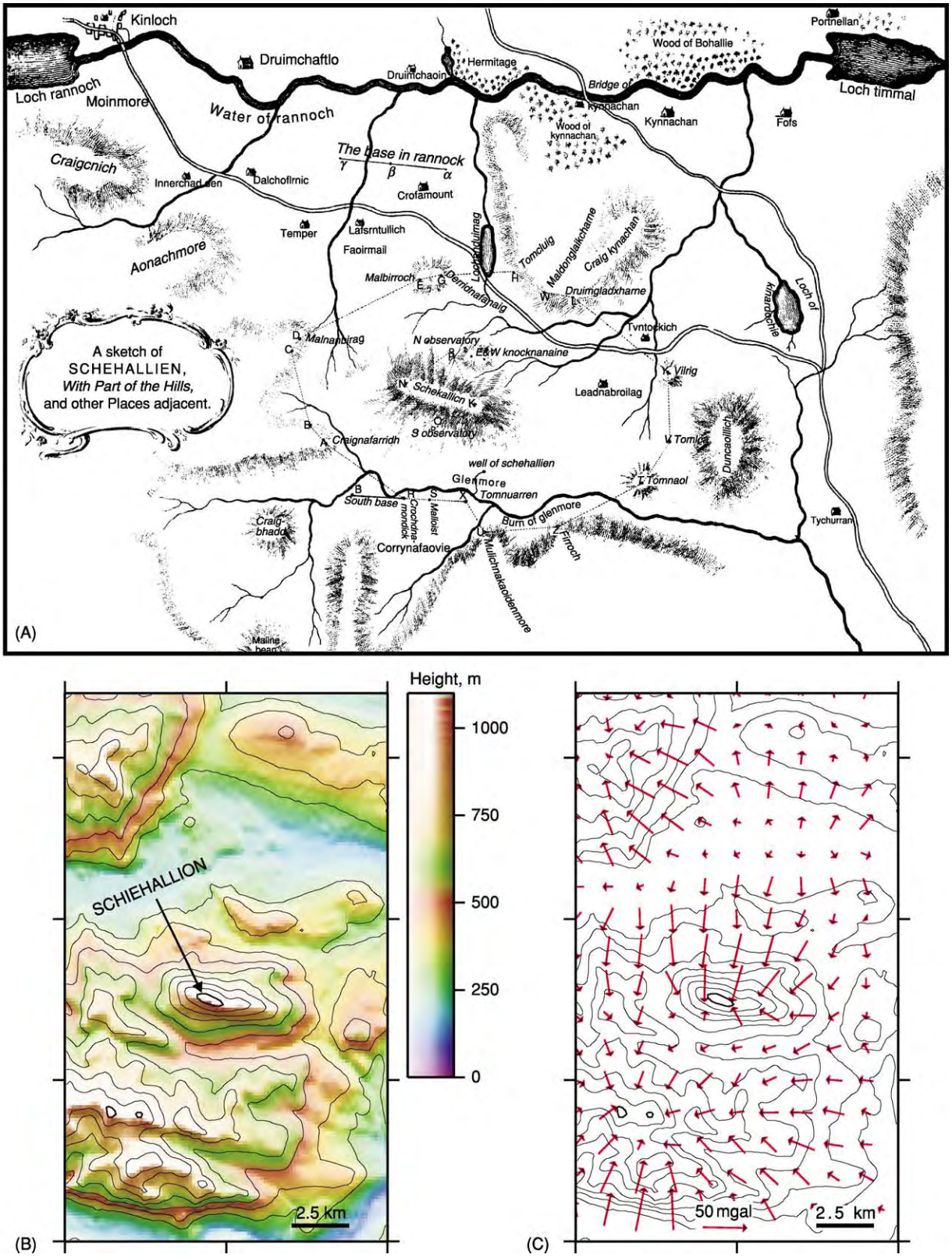


Figure 1 Sketch map of the area around Schiehallion, Scotland, by Charles Hutton. Plumb line deflections measured at stations north and south of the mountain allowed the first estimate of the density of the earth. (After R.M. Sillito with permission from Hutton (1778) © The Royal Society.)

fired vertically upwards and allowed to fall back along the same path. The gravity measurement depends on timing the upward and downward paths, which may be by light beam-controlled timers or interferometry.

Gravity differences can be measured on land with a stable gravity meter or gravimeter based on Hooke's law. A mass extends a spring under the influence of gravity and changes in extension are proportional to changes in the gravitational acceleration. More sensitive are 'astatic' gravity meters, which contain a mass supported by a 'zero-length' spring for which tension is proportional to extension. When the meter is in position, a measurement is made of an additional force needed to restore the mass to a standard position, supplied by an auxiliary spring or springs, an electrostatic system, or an adjustment of the zero-length spring itself. Gravity meters working on this principle measure differences in gravity between stations and surveys may be tied to one or more base stations at which repeated measurement can be made. Astatic gravity meters can have a sensitivity of about 0.01 mGal.

For applications where the gravity meter is subject to tilting and vibration, such as on board a ship or in an aircraft, isolation of the instrument is required such as providing a moving stabilised platform for the gravity meter and damping vibrations with appropriate shock absorption. When the gravity meter is moving, accurate data on the location and trajectory of the platform is required along with the gravity measurement. For airborne application, this requirement has been greatly assisted by the advent of the global positioning system (GPS) which allows rapid, precise, and accurate positioning (*see* **Remote Sensing:** GIS). Airborne gravity surveys, whether flown using fixed wing or helicopters, can provide economic, rapid, and non-invasive geophysical reconnaissance ideal for difficult terrain such as tundra, jungles, and wildlife reserves.

Deviations in artificial satellite orbits can be used to determine the long-wavelength components of the Earth's gravity field. Altimetry tools mounted on satellites have allowed much more detailed gravity mapping over the oceans, as sea surface height data can be processed to give the marine geoid. Geoid data can then be converted to gravity data with a series of numerical operations (Figure 2). Since the mean sea-level surface is the geoid, an equipotential surface, variations in sea surface height from the reference ellipsoid reflect density changes below the sea surface, largely from the density contrast at the seabed, but also from sub-seabed changes, such as crustal thickness changes.

Adjustments to Measured Gravity Signals

The first correction that can be applied to measured gravity values is the correction for latitude, to account for the centrifugal acceleration which is maximum at the equator and zero at the poles (Table 1). For gravity measurements made on land, several further corrections must be made (Table 1). The 'free-air correction' is made to adjust for difference in height between the measurement point and sea-level. This does not make any assumptions about the material between the sea-level datum level and the observation point and uses the inverse square law and the assumption of a spherical Earth. The 'Bouguer correction', named after the French mathematician and astronomer, is used to account for the gravitational effect of the mass of material between measurement point and sea-level. This requires assumptions to be made about the density of material, and the Bouguer plate or slab formula is applied (Table 1), which further assumes that this material is a uniform infinite plate. Historically a 'density correction' value of 2670 $kg\,m^{-3}$ has been used as a standard density for crustal material, and this corresponds to a Bouguer correction of 1.112 g.u./m, negative above sea-level. A 'terrain correction' may be applied to compensate for the effect of topography, again requiring assumptions about densities. Nearby mass above the gravity measurement station will decrease the reading and any nearby topographic lows will have been be artificially 'filled in' by the Bouguer correction so the correction is always positive. An additional correction to gravity measurements made on a moving vehicle such as an aeroplane or boat is the Eötvös correction, which depends on horizontal speed vector, latitude, and flight altitude.

Gravity Anomalies and Derivatives

Since for most geological applications the perturbations in the gravity field across an area or feature of interest are more important than the absolute gravity values, it is standard to compute gravity anomalies by subtracting the theoretical gravity value from the observed. The Bouguer gravity anomaly is the observed value of gravity minus the theoretical gravity value for a particular latitude and altitude, as outlined in Table 1. The Bouguer gravity is commonly used on land where maps of gravity anomalies can be used to view gravity data in plane view and it is convenient to have topographic effects (approximately) removed. Offshore, the free-air gravity anomaly is most useful, as the measurements are straightforward to correct to the sea-level datum.

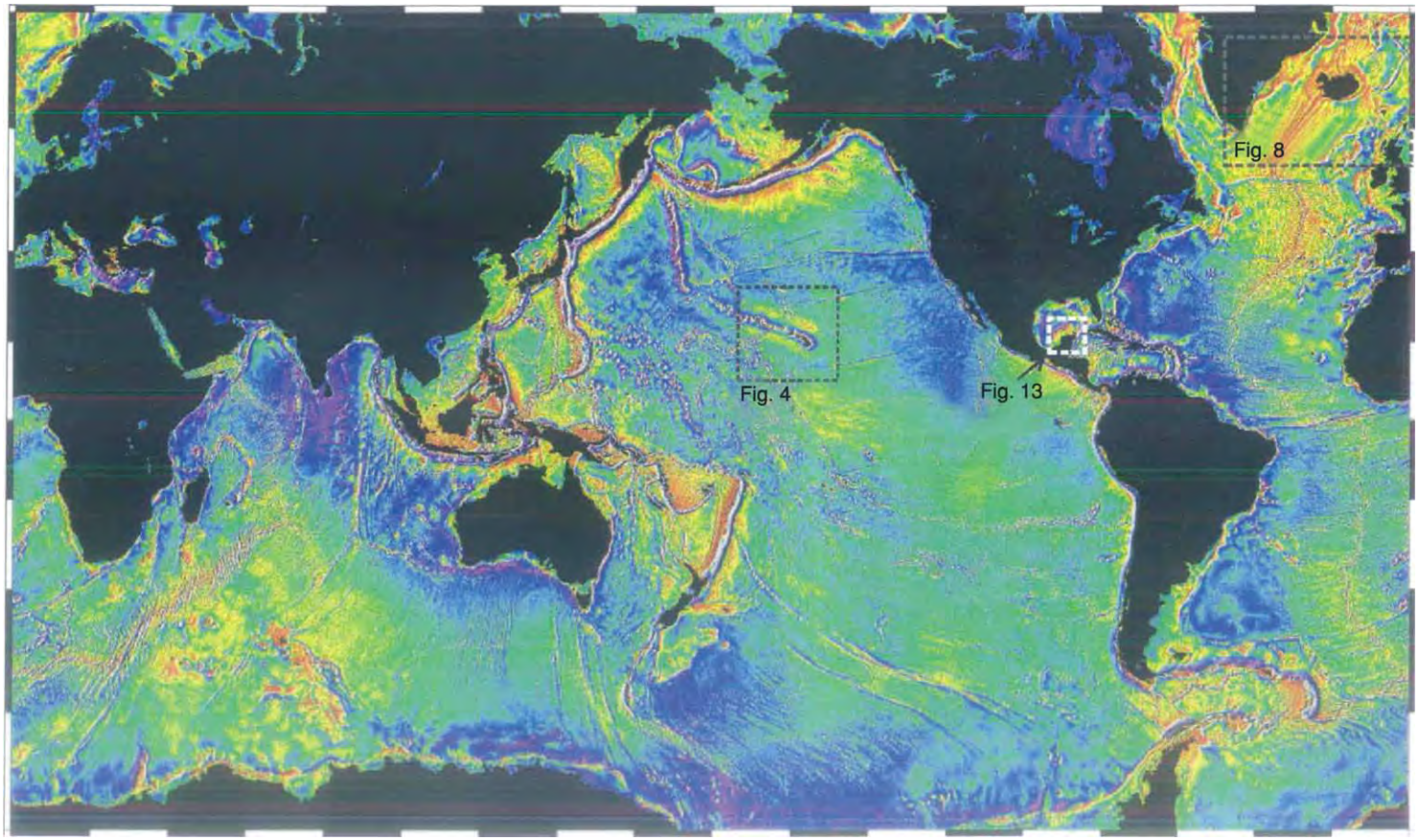


Figure 2 Marine free air gravity anomaly map derived from satellite altimetry (Sandwell and Smith (1997)). Warm colours indicate positive gravity anomalies. The gravity anomaly primarily indicates the shape of the seafloor, due to the strong density contrast from seawater to oceanic crust. Oceanic island chains, subduction zone trenches, and mid ocean ridges form features visible on this world map. Locations of **Figures 4, 8** and **13** indicated. (Image courtesy of NGDC.)

High-pass filtering or subtraction of a planar function from gravity anomaly data may be undertaken to remove a 'residual' or background trend if the feature of interest is known to be shallow or a subtle perturbation to a strong regional gradient. Other treatments of gravity data include upward and downward continuation, by which different observation levels can be simulated, and computation of vertical or horizontal derivatives, which may emphasise structural trends in the data.

Applications and Examples

Submarine Topography

The satellite-derived free-air gravity anomaly map over the oceans (Figure 2) strongly reflects the nearest significant density change, the seabed. There are positive gravity anomalies over seabed topographic highs such as submarine seamounts and mid-ocean ridges and negative anomalies over bathymetric deeps such as the trenches associated with subduction zones, although long-wavelength isostatically compensated structures have no gravity anomaly above them.

The coverage of the marine free-air gravity anomaly data can be exploited to produce sea-floor topography data (Figure 2). For this purpose, shipboard depth surveys, usually made with sonar equipment, are used to supply the long-wavelength part of the transfer function from gravity to topography. The shipboard data is usually considered accurate but limited in global coverage due to the spacing and orientation of survey ship tracks. Bathymetry interpolation using the satellite-derived gravity data highlights isostatically

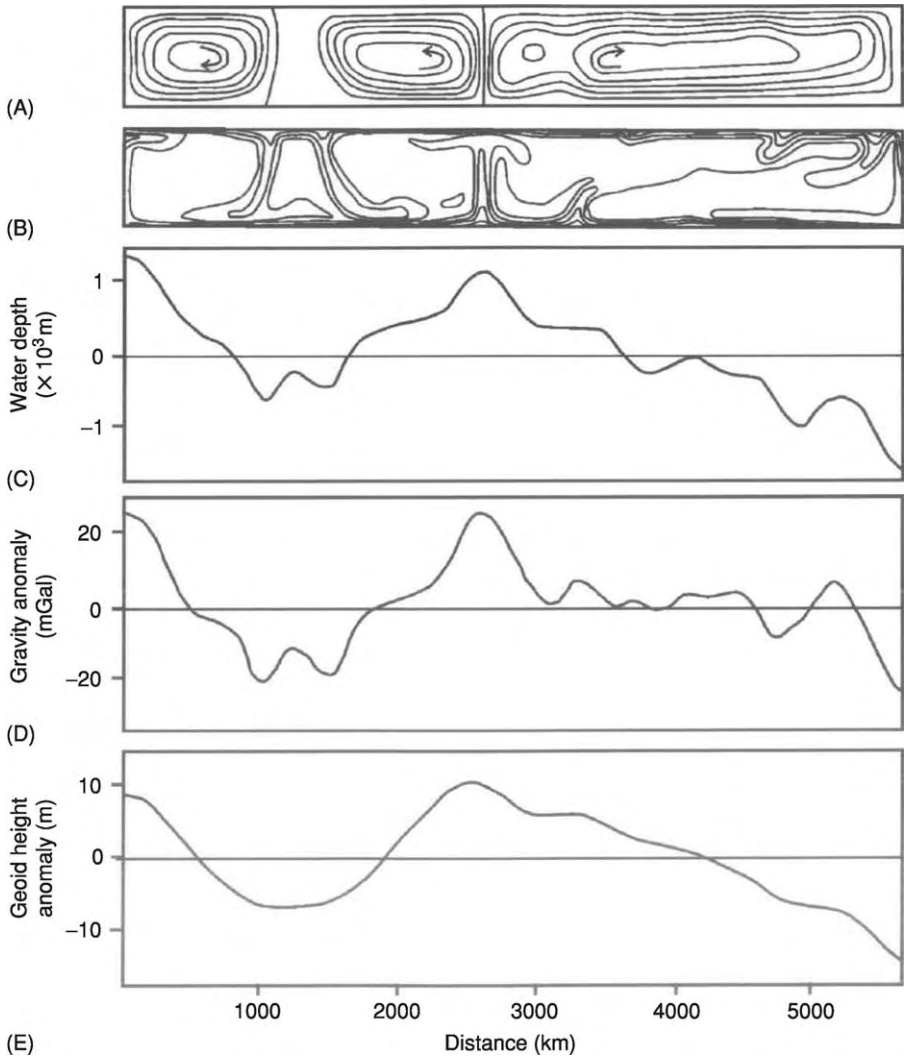


Figure 3 Model of the gravity effect of convection in the Earth's mantle. (A) Stream function of computer modelled mantle flow (B) 100°C temperature contours (C) Variation in seafloor depth given a 30 km thick elastic lithosphere above the convecting mantle (D) Modelled free air gravity anomaly (E) Modelled geoid (sea surface height) anomaly over the convecting mantle. (Reproduced with kind permission from McKenzie *et al.* (1980) © Nature Publishing Group. http:www.nature.com)

compensated topography which has no long-wavelength expression in the gravity data alone.

Mantle Convection

It is commonly accepted that the Earth's mantle convects, and the flow of mantle material gives rise to gravity anomalies. Where mantle material is anomalously hot, it has a lower density than surrounding cooler mantle, and it will give rise to a negative gravity anomaly at the surface. This effect is, however, overprinted by the positive gravity anomaly caused by the upward deflection of the lithosphere above the rising anomalously hot mantle column or sheet (*see* **Mantle Plumes and Hot Spots**). There will, therefore, be a positive gravity anomaly over rising mantle material and a negative gravity anomaly where mantle is cool and sinking (Figure 3).

Isostasy and Lithospheric Strength

Not all mountains would cause a gravitational plumb-line deflection such as that observed at Schiehallion. Bouguer had observed that a plumb-line was only deflected by 8 seconds of arc towards the mountains during Condamine's Quito survey, while his calculations suggested that it should have been deflected as much as 1′ 43″. This anomalous lack of deflection was attributed by R. Boscovich in 1755 to 'compensation' for the mass excess of the mountain by underlying mass deficiency at depth. This fed into the development of 'isostasy', which addresses the issue of support for topography on the Earth's surface. Two alternative early views of isostatic theory were put forward in the 1850s. John Henry Pratt suggested that the amount of matter in a vertical column from the surface to some reference level in the Earth was always equal, and that this was achieved by the material in the column having lower density material below mountains than below topographic lows. George Biddell Airy advanced the alternative view using the analogy of icebergs, that elevated surface topography was underlain by low-density crustal roots which effectively displaced denser underlying material. Subsequent studies have used gravity data to investigate these alternative models in different tectonic settings and included the additional factor of the strength of the lithosphere to support loads.

At wavelengths shorter than about 500 km, the relationship between the gravity anomaly and topography

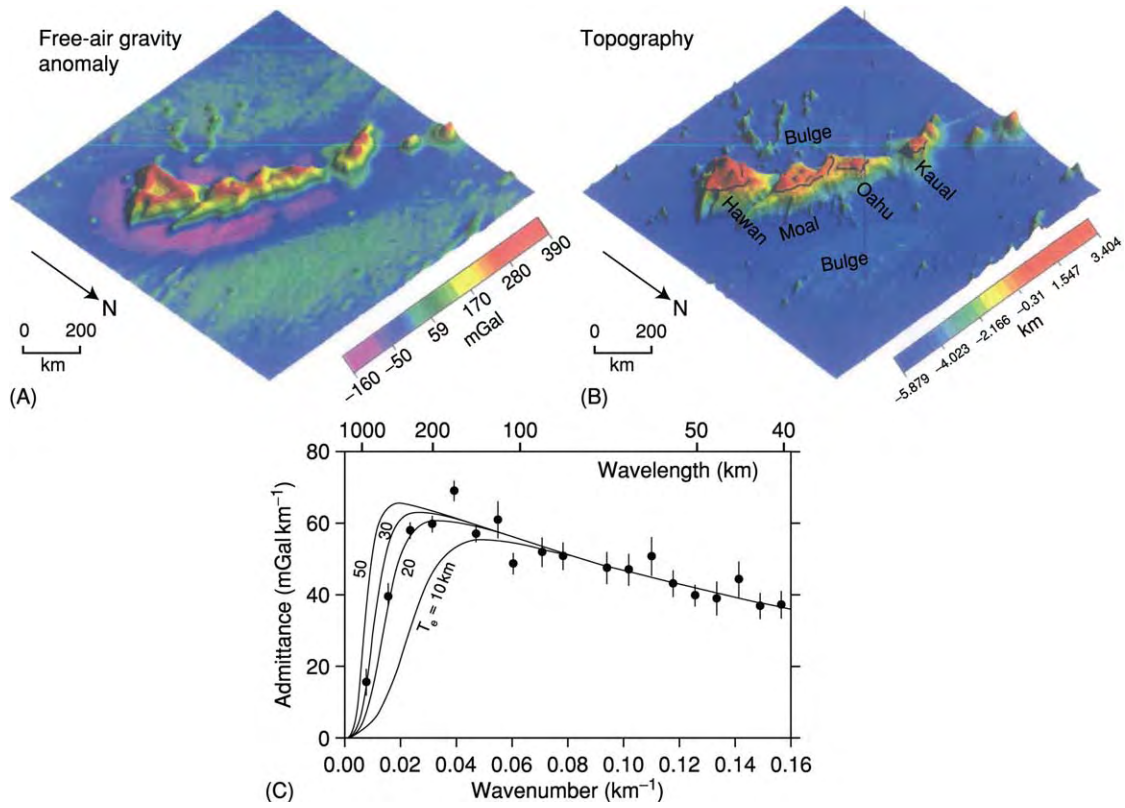


Figure 4 Free air gravity anomaly (A) and topography (B) in the region of the Hawaiian islands, Pacific Ocean (see Figure 2 for location). (C) A comparison of the observed admittance along the Hawaiian Emperor seamount chain (dots) with the predictions of a simple flexure model of isostasy, with varying elastic thickness, Te (lines). The observed admittance can be best explained with an elastic thickness for the lithosphere of 20 30 km. See Watts (2001) for more details. (Reproduced with kind permission from Watts (2001) © Cambridge University Press.)

is controlled by the mechanical properties of the lithosphere, which may be strong enough to support short wavelength loads, for example, isolated mountains. At longer wavelengths, the flexural strength of the lithosphere is commonly insufficient to support loads. The relationship between the gravity anomaly and topography can described by the wavelength-dependent 'admittance' function. The rate of change from flexurally-supported topography at short wavelength to topographic support by base-lithospheric pressure variations and regional density variation at long-wavelength depends on the effective 'elastic thickness' of the lithosphere.

Figure 4 shows the topography and gravity anomaly of some of the Hawaiian island chain and the calculated admittance. For these islands, a modelled elastic thickness of about 25 km matches the admittance data. Recently, methods have been developed to also include the effect of lithospheric loads both with and without topographic expression in estimation of the elastic thickness.

Density Contrasts, Analytical Models, and Non-Uniqueness

On a smaller scale, gravity anomaly maps provide the opportunity to identify and delineate sub-surface structures, as long as there are lateral density changes associated with the structure. Rocks at and near the surface of the Earth are much less dense than the Earth's average density of approximately 5155 kg m ^{3}, and crustal rocks are almost universally less dense than mantle rocks. An approximate density value of 2670 kg m ^{3} is often taken as an average value for upper crustal rocks while values of 2850 kg m ^{3} and 3300 kg m ^{3} have been used for overall crustal rocks and uppermost mantle, respectively, although these values vary with composition and temperature. Many sedimentary rocks are less dense than metamorphic and igneous rocks. Coal (1200–1500 kg m ^{3}) is one of the least dense rocks, while chalks and siliciclastic sedimentary rocks (1900–2100 kg m ^{3}) are generally less dense than massive carbonates (2600–2700 kg m ^{3}). With the exception of porous extrusive examples, crustal igneous rocks have densities approximately ranging from 2700 to 3000 kg m ^{3}. Density is not a diagnostic for lithology and variation in parameters such as porosity, temperature, and mineralogy can give significant density variability. Rocks with the lowest densities are those with very high porosities such as volcanic pumice, and in sub-aqueous environments recently deposited sediments. Density of sediments in a sedimentary basin tends to increase with depth as grains are compacted together (Figure 5). Igneous and metamorphic rocks tend to have higher densities than sediments as they frequently have negligible porosity and consist of relatively dense minerals.

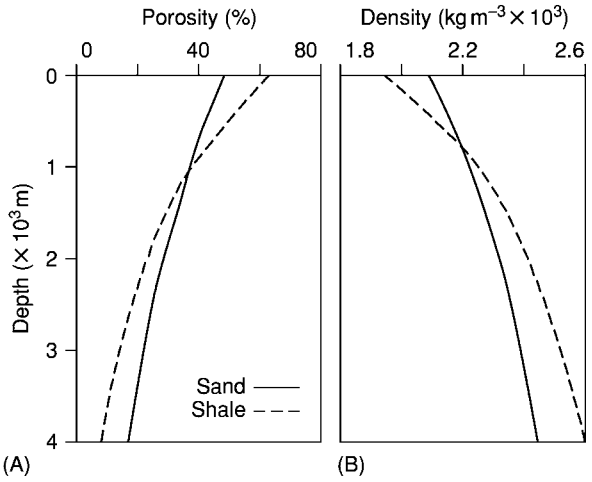


Figure 5 Typical variation of (A) porosity and (B) density with depth below seafloor for sands and shales in a sedimentary basin. Increasing vertical effective stress with depth causes compaction of the rock, reducing porosity and correspondingly increasing density. Deeply buried sedimentary rocks, therefore, have higher densities than shallower rocks of similar lithology.

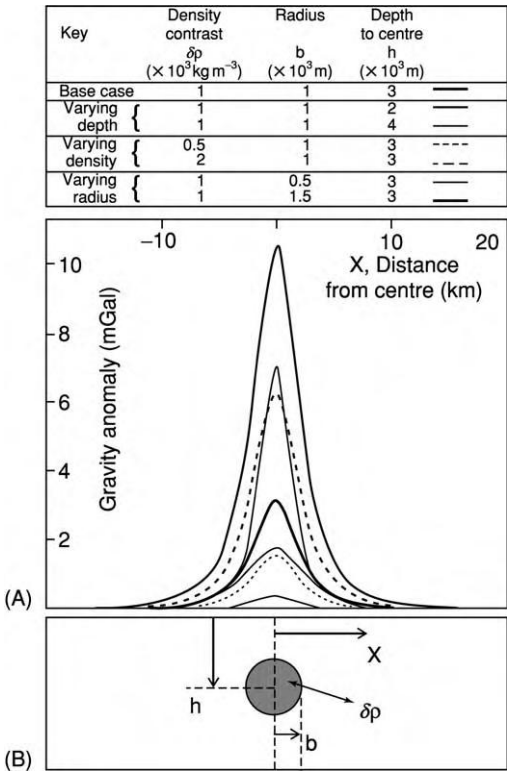


Figure 6 Modelled gravity anomaly (A) along a transect through the centre of a buried sphere (B) of varying radius b, density contrast $\delta\rho$, and depth of burial h. The similarity in shape between the various cases shown highlights the difficulties of interpretation of gravity anomalies, as there are no unique solutions to explain a particular gravity anomaly.

Any non-uniformity in mass distribution results in lateral variability of the gravity field. For some simple geometrical shapes, a gravity anomaly can be calculated analytically (Figure 6). The buried sphere example illustrates the observation that deep density anomalies give rise to an anomaly over a wider surface distance than otherwise similar shallow anomalies, while greater density contrasts give larger anomalies than small density contrasts. The similarity in the gravity anomaly curves for the example of a buried sphere (Figure 6) illustrates one of the problems that arises in interpreting gravity data: there is no unique density distribution that produces a particular gravity anomaly. Gravity models tend to be constructed using additional geological or geophysical data such as seismic refraction or reflection profiles, surface geology (Figure 7), borehole density measurements, magnetic, magneto-telluric, or

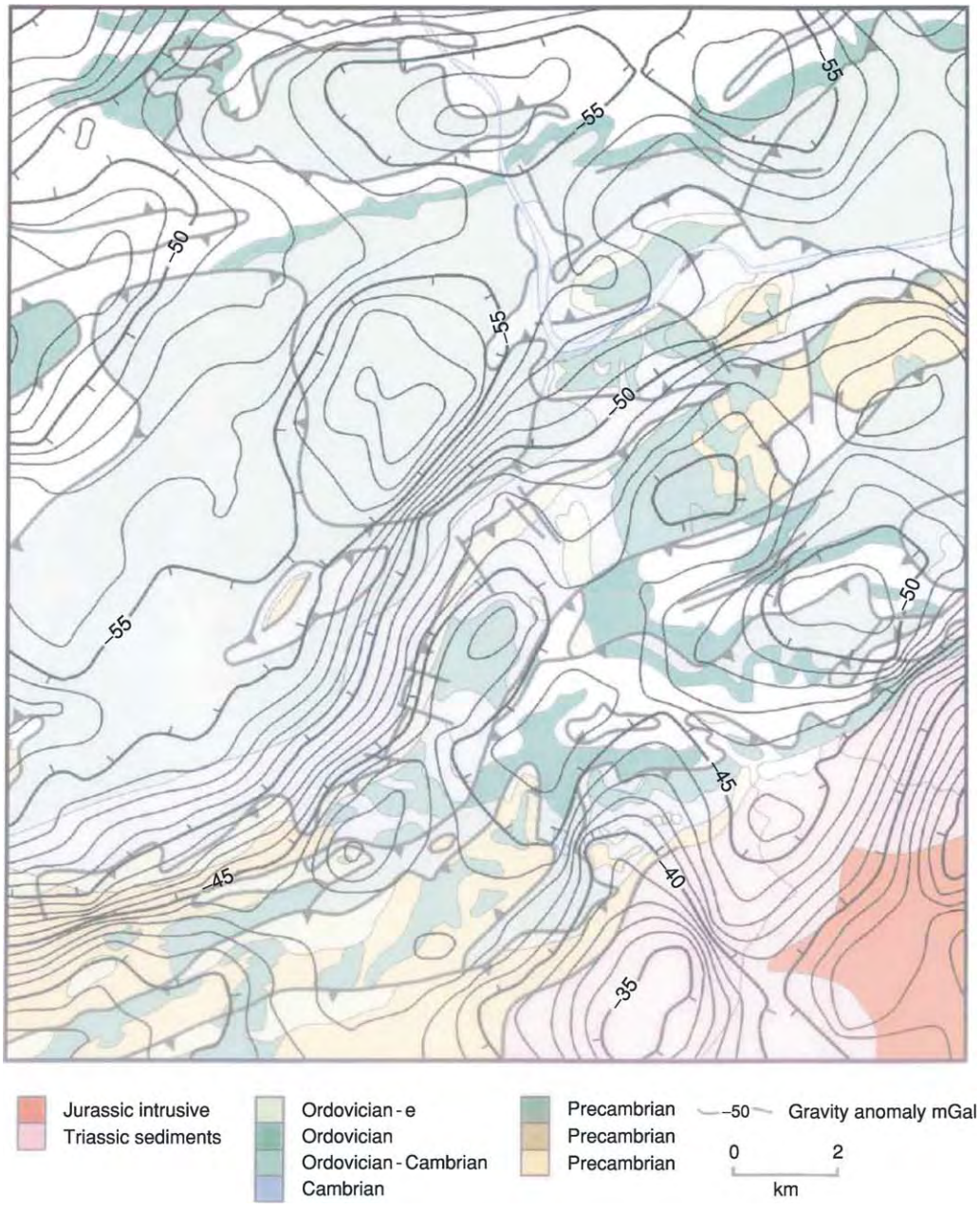


Figure 7 Bouguer gravity anomaly contours overlain on geological map of part of Eastern Pennsylvania, USA. Bouguer gravity anomaly highs occur over the horst blocks of dense Precambrian material and other lows and highs in the gravity field are associated with formations of varying densities. (Geological map courtesy of the Bureau of Topographic and Geologic Survey, Pennsyvania Department of Conservation and Natural Resources, gravity data courtesy of W. Gumert and Carson Services Inc. Aerogravity Division, PA, USA.)

electromagnetic surveys as appropriate to arrive at a plausible and consistent interpretation.

Crustal Observations from Satellite Gravity

Circular gravity anomalies similar in shape to that calculated for a buried sphere are observed over discrete igneous centres (Figure 8), where dense igneous rocks are inferred to have intruded less dense crustal rocks at a point of weakness in the lithosphere. Some of these 'bulls-eyes' in the gravity field have topographic expression but others may not have been identified without the satellite-derived gravity map. This map also allows identification of other large-scale crust-mantle interactions. One example is the set of south-ward pointing 'V-shaped' gravity anomalies flanking the mid-ocean ridge south of Iceland (Figure 8), which are caused by ridges and troughs in the top of the igneous crust. Although partially buried by sediment, these ridges have an expression in the gravity anomaly map because there is a significant density contrast between the igneous upper crustal rocks and the young pelagic sediments draping them.

Modelling in Conjunction with Other Data

A combination of gravity data and other data types is often productive. For example, oceanic fracture zones identified in the satellite-derived gravity anomaly map are useful in conjunction with 'sea-floor stripes' in magnetic anomaly data to determine the relative movement between tectonic plates (Figure 8).

Gravity data is commonly used to verify interpreted seismic models. Empirical relationships between seismic velocity and density can be used to convert a seismic (*see* **Seismic Surveys**) velocity model into a density model and the predicted gravity anomaly compared with observations. The example shown in Figure 9 shows a crustal velocity model along a 400 km line in the North Atlantic that has been

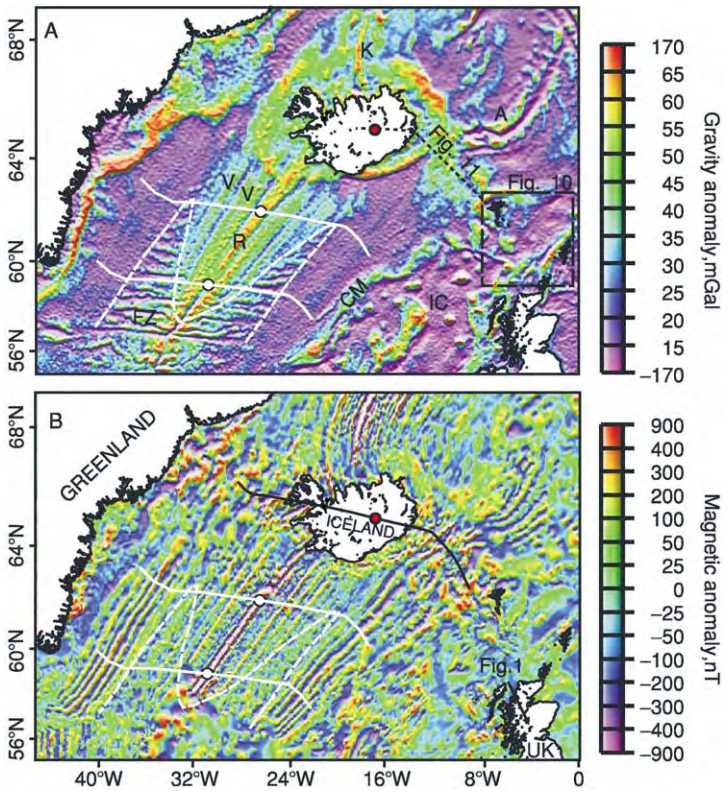


Figure 8 Free air gravity anomaly (A) and Magnetic anomaly (B) over the area surrounding Iceland (see **Figure 2** for location). The magnetic stripes form the record of magnetic field reversals during production of oceanic crust at the spreading centre. There are gravity anomaly highs over topographic highs such as the Reykjanes Ridge (R) and Kolbeinsey Ridge (K) spreading centres, and the extinct Aegir Ridge spreading centre (A). There are also circular gravity highs over igneous centres (IC) and linear anomalies along the continental margins (CM) and 'V shaped' ridges (V) which flank the Reykjanes Ridge and reflect propagating pulses of anomal ously hot mantle beneath the spreading centre. Red and white circles show the position of the present day spreading centre plate boundary. Solid white line shows flowlines from present spreading centre indicating direction of paleo seafloor spreading. These are determined from reconstruction of the magnetic stripes parallel to the fracture zones seen in the gravity data. Dashed white line shows area of oceanic crust disrupted by fracture zones (FZ); outside this area, oceanic crust was formed at a spreading centre without fracture zones on this scale. Dotted black line indicates approximate line of **Figure 12**. Location of **Figure 9** indicated. (After Smallwood and White (2002) with permission Geological Society of London.)

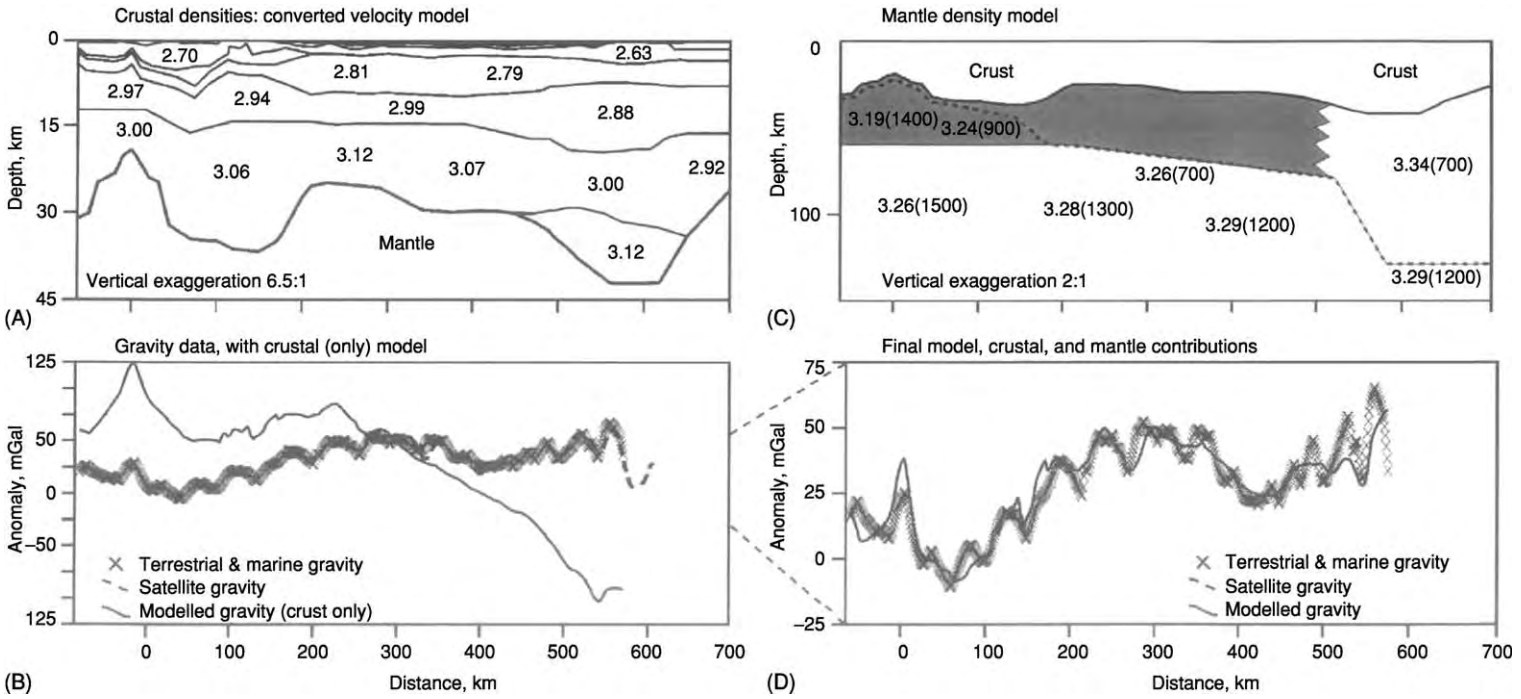


Figure 9 Gravity model along a profile from Iceland to the Faroe Islands (see **Figure 8** for location). The crustal density model (A) was constructed by converting a wide angle seismic velocity model to density using published empirical relationships. The gravity signature of the crustal model (B) did not match the satellite derived (crosses) or ship board/land based gravity anomaly measurements. When effects of varying lithosphere thickness and mantle density variations were included (C), a good match between model and data could be achieved (D). (After Smallwood *et al.* (1999) by permission © American Geophysical Union.)

converted to density. While the gravity anomaly signal expected from the crust alone does not match the observed gravity, when reasonable mantle temperature and compositional variations are included, a good match to the data can be obtained.

Gravity data is increasingly being incorporated into multivariable mathematical inversion projects in which multiple datasets are simultaneously modelled in order to increase confidence in a particular interpretation of the subsurface.

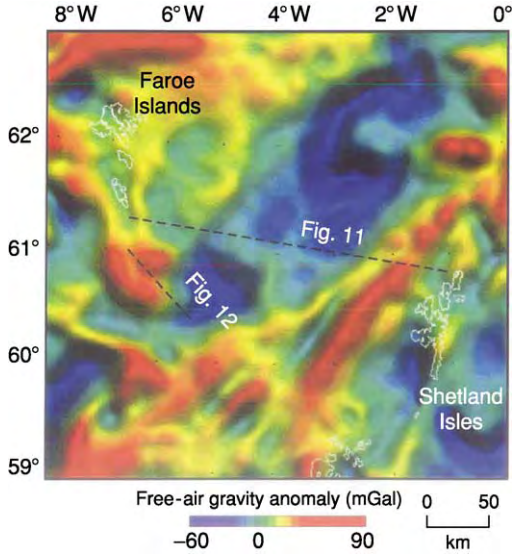


Figure 10 Free air gravity anomaly over Faroe Shetland area derived from satellite altimetry (Sandwell and Smith 1997) and shiptrack data (see **Figure 8** for location). The dominant signal is the NW SE gravity reflecting the area of deepest water between the Faroe Islands and the Shetland Isles. Shorter wavelength features arise from geological structures (see **Figures 11** and **12**).

Modelling Over Sedimentary Basins

Since there is often a significant density contrast between crustal and mantle rocks, gravity data may provide useful constraints on crustal thickness variations, which can occur in continental as well as oceanic settings. Lithospheric extension, for example, may thin the crust along with the rest of the lithosphere. As the relatively low density crust is thinned, it may isostatically subside and the resulting topographic low may form a sedimentary basin (**Figure 10**). If assumptions are made about rock densities, gravity anomaly data can be modelled to infer the extent of crustal thinning. Simplified models of the subsurface can be constructed and adjusted until a match or matches can be made to gravity observations. Mathematical inversion may assist by identifying a model which produces a gravity field that has a minimum misfit to observations.

In the example of this, shown in **Figures 11** and **12**, from the UK/Faroe-Shetland Basin, the gravity data is particularly valuable as flood basalts to the west of the basin make seismic imaging difficult. Although the top of the relatively dense mantle is elevated in the position where the crystalline crust is modelled to be most highly extended, there is a free-air gravity low caused by the dominance of the relatively low density water column and sedimentary fill which are constrained by seismic data, and the long wavelength effect of the thicker continental crust on the basin

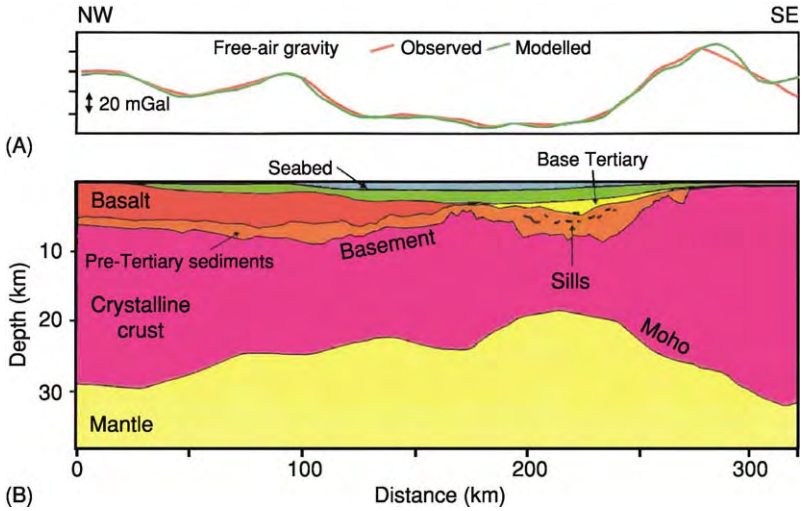


Figure 11 Modelled and observed free air gravity (A) along a profile between the Faroe Islands and the Shetland Isles (see **Figure 10** for location). The seafloor and other horizons (B) were partly constrained by seismic reflection data but beneath the basalt wedge reflections were not easy to interpret and the gravity modelling along this and other intersecting lines constrains a possible crustal model. (After Smallwood *et al.* (2001) with permission, Geological Society of London.)

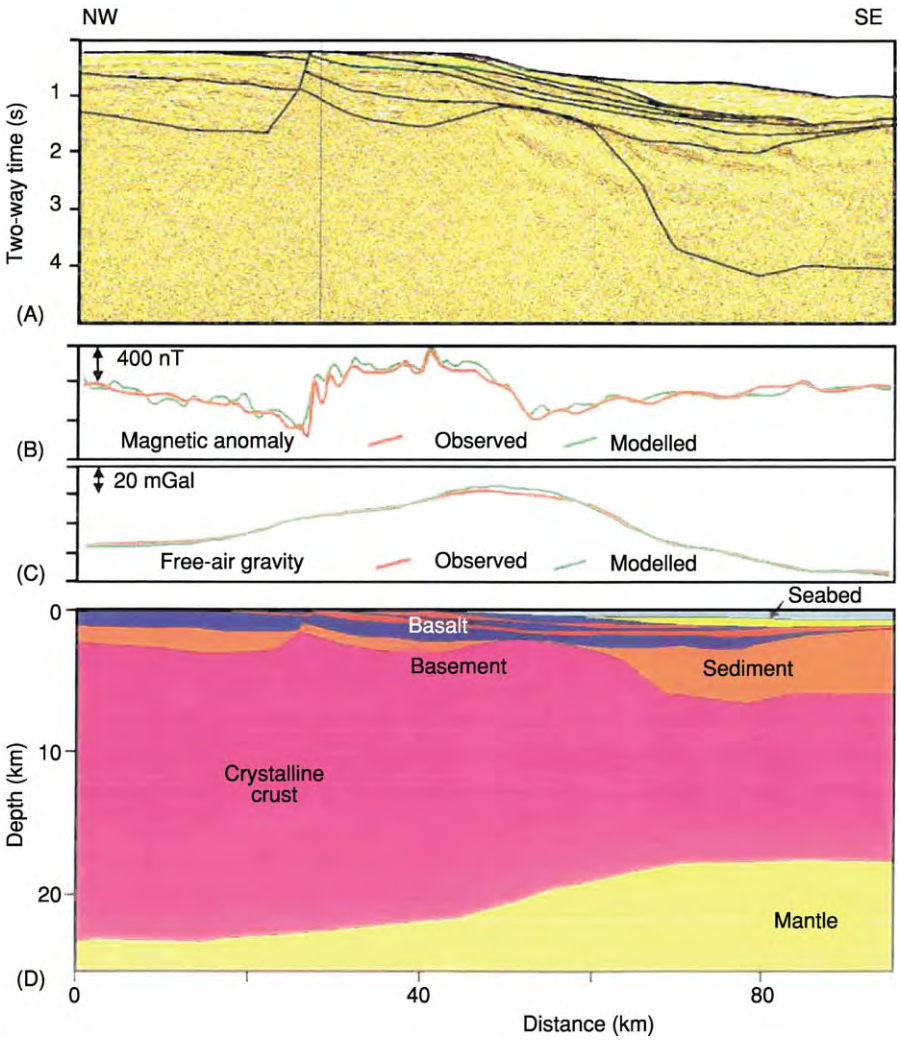


Figure 12 Modelled and observed free air gravity along a profile southeast of the Faroe Islands (see **Figure 10** for location). The seafloor and other horizons were partly constrained by seismic reflection data (A) but beneath the basalt reflections were not easy to interpret and magnetic anomaly (B) and gravity anomaly (C) modelling along this and other intersecting lines constrains a possible crustal model (D). (After Smallwood *et al.* (2001) with permission, Geological Society of London.)

margins. Another benefit added by gravity data to the understanding of this sedimentary basin was the requirement to add a unit with elevated density approximately 1 km thick in the centre of the basin to represent an interval intruded by igneous sills. The top of this unit was imaged well by seismic data but the thickness could not be estimated without the gravity model. Figure 12 shows the value of modelling magnetic anomaly data along with the gravity to constrain basalt thickness and internal structure.

Another geological structure for which gravity data provides a useful tool of investigation is the Chicxulub impact crater in the northern Yukatan peninsular of Mexico. There is no dramatic surface expression of the site, but there are concentric circular rings apparent in the gravity anomaly (Figure 13). The gravity anomaly arises as the crater has been infilled with relatively low-density breccias and Tertiary sediments. The double humped central gravity high is thought to correspond to a central uplift buried deep within the crater. The Chicxulub crater is one example where 3D gravity modelling has proved useful to constrain crustal structures in three dimensions.

Smaller Scale Surveys

Spatial deviation of gravity measurements is often used to infer lateral variations in density. If sufficiently accurate measurements can be made, then small-scale lateral variations in density can be inferred. Gravity surveying may be the best tool to identify mineral deposits if the target ores have densities contrasting with their host rocks. Massive

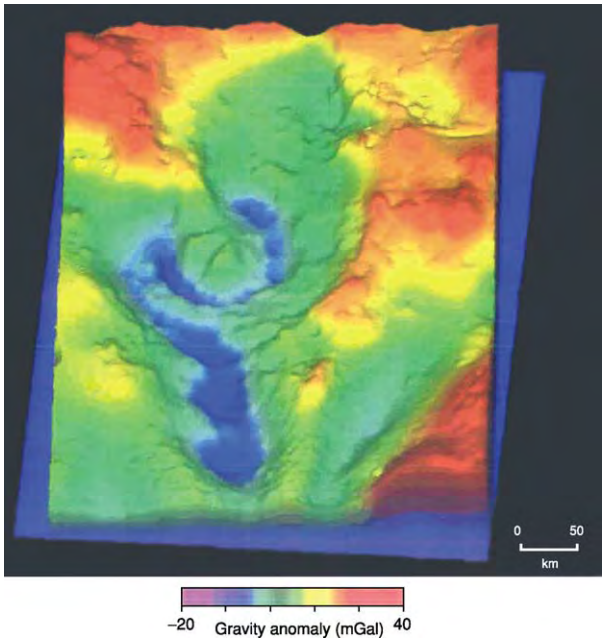


Figure 13 Merged free air gravity (offshore) and Bouguer (on shore) gravity anomalies across Chicxulub impact crater, Yucatan peninsular, Mexico (see **Figure 2** for location). Bouguer anomaly calculated with a reduction density of 2670 kg m ^{3}. Gravity anomaly over Chicxulub is a 30 mGal circular low with a 180 km diameter, with a central 20 mGal high. (Courtesy of Mark Pilkington, Natural Resources Canada.)

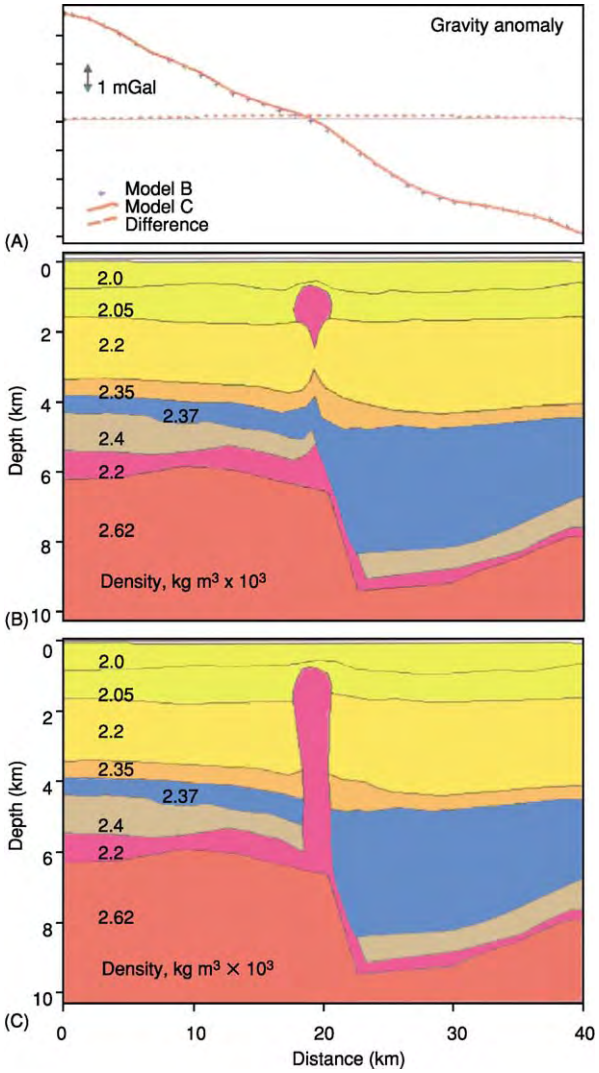


Figure 14 Cross sections across two gravity models. The modelled gravity response (A) for a cross section over a fault block and small detached salt pillow (B) is very similar to the response over a bigger salt diapir (C) offset by some other changes to the model layers. Since uncertainty and noise in marine gravity data may be at a 1 mGal level, gravity modelling of the total field may not be able to distinguish between these models. Seismic data is often poor below the top of the salt. Courtesy of A. Cunningham.

sulphides have densities ranging up to 4240 kg m ^{3}, and within host rocks of densities around 2750 kg m ^{3}, a sulphide body having a width of 50 m, a strike length of 500 m, and a depth extent of 300 m would give a gravity anomaly of about 3 mGal.

On a smaller scale, 'micro-gravity' surveys typically involve a large number of closely spaced gravity measurements aiming to detect gravity variations at levels below 1 mGal. These surveys may be designed for civil engineering projects where underground natural cavities in limestone or disused mine workings need to be detected, or depth to bedrock needs to be established. As with any gravity interpretation, any additional available information such as outcrop geological boundaries, density values of samples, or depths to important horizons may be incorporated in order to give a more realistic model.

Gravity Gradiometry

Sometimes knowledge of the magnitude of the gravity field is not sufficient to resolve between competing geological or structural models. In the example shown in **Figure 14**, the conventional gravity data is rather insensitive to the geometry of the salt diapir as a dominant long-wavelength gravity signal originates from an underlying fault block. In this case, the gradients of the gravity field may provide additional assistance. An instrument to measure the gradient of the gravity field was developed by Baron von Eötvös in 1886, and a unit of gravity gradient was named after him (1 Eötvös = 0.1 mGal km ^{1}). The concept of his torsion balance was that two weights were suspended from a beam at different heights from a single torsion fibre, and the different forces experienced by the two weights would deflect the beam. The torsion balance was accurate but somewhat cumbersome and slow, and it was superceded by the more

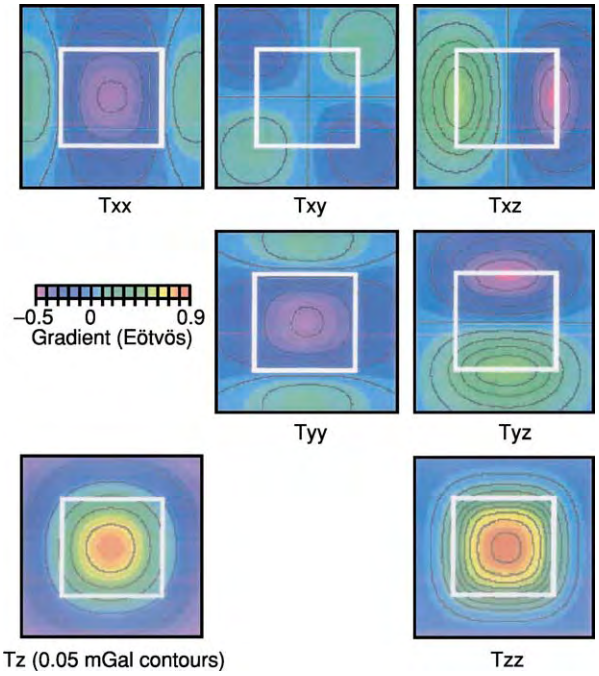


Figure 15 Modelled components of the gradients of the gravity field over a cube. The different tensors represent changes in the gravity field gradient in different directions, for example, Tzz is the vertical gradient of the field measured in a vertical direction. It highlights edges of subsurface density contrasts. Tz is the (directionless) full gravity anomaly field. (Courtesy of C. Murphy and Bell Geospace Inc.)

convenient astatic gravity meters. However, recently declassified military technology has seen a renaissance of gravity gradiometry, with the availability of a full tensor gradiometer consisting of 12 separate accelerometers arranged in orthogonal pairs on three separate rotating disks. This instrument has a quoted instrumental accuracy of 10^{-11} gal. Five independent tensor measurements are made and can be modelled, each sensitive to different aspects of subsurface density variation (Figure 15). In addition to the valuable insights gained from this multicomponent data, the components can be recombined to give a high resolution 'conventional' gravity map which benefits from the precision of the instrumentation. In cases similar to the salt diapir example, inversion of such precise gravity data has been used together with a correlation between seismic velocity and density to produce a modelled seismic velocity field which can be used in seismic depth processing.

Extra-terrestrial Gravity Fields

Gravity fields have been computed for the moon (*see* **Solar System:** Moon), and for Venus (*see* **Solar System:** Venus) and Mars (*see* **Solar System:** Mars), from observations of variation in artificial satellite orbits. The Doppler shift of spacecraft signals is observed, giving the spacecraft velocity in the 'line-of-sight' direction. The gravity field is calculated using a combination of many observations of line-of-sight acceleration.

In a similar method to that outline for terrestrial studies, the wavelength-dependent relationship between gravity anomalies and topography can be used to study the internal dynamics and the support of surface loads by the lithosphere. Gravity studies on Venus show that it has a similar lithospheric rigidity to the continents on Earth, despite its higher surface temperature, and that active mantle convection is responsible for the observed volcanic rises. In contrast, large gravity anomalies on Mars for example, a maximum anomaly of 344 mGal (from a spacecraft altitude of 275 km) over the crest of the Olympus Mons volcano, have led to the suggestion that the Martian lithosphere is extremely rigid. On the moon, circular positive gravity anomalies of up to 300 mGal have been identified, associated with basaltic lava flows infilling giant impact craters. These 'mascons' (mass concentrations) have provided a focus for debate on isostatic lunar history.

Conclusion

Gravity is a versatile tool for investigation and can provide constraints on sub-surface structure on a wide variety of scales from man-made structures to the size of an entire planet. To unlock the information contained within the gravity field, gravity observations are best used in conjunction with other types of data such as surface topography, geological mapping, borehole information, and seismic data.

See Also

Mantle Plumes and Hot Spots. **Mining Geology:** Exploration. **Petroleum Geology:** Exploration. **Plate Tectonics**. **Seismic Surveys**. **Solar System:** Venus; Moon; Mars.

Further Reading

Bott MHP (1982) *The Interior of the Earth*, 2nd edn. Amsterdam: Elsevier.

Fowler CMR (1990) *The Solid Earth: An Introduction to Global Geophysics*. Cambridge, UK: Cambridge University Press.

Gibson RI and Millegan PS (eds.) (1998) *Geologic Applications of Gravity and Magnetics: Case Histories*. SEG Geophysical Reference Series 8/AAPG Studies in Geology 43. Tulsa, OK: Society of Exploration Geophysicists and the American Association of Petroleum Geologists.

Gumert WR (1998) A historical review of airborne gravity. *The Leading Edge* 17: 113 117. http://www.aerogravity.com/carson2.htm.
Hansen R (1999) The gravity gradiometer: basic concepts and tradeoffs. *The Leading Edge* 18: 478, 480.
Hildebrand AR, Pilkington M, Connors M, Ortiz Aleman C, and Chavez RE (1995) Size and structure of the Chic xulub crater revealed by horizontal gravity gradients and cenotes. *Nature* 376: 415 417.
Hutton C (1778) An account of the calculations made from the survey and measures taken at Schiehallion, in order to ascertain the mean density of the Earth. *Phil. Trans. Royal Soc.* LXVIII: 689 788.
McKenzie DP, Watts AB, Parsons B, and Roufosse M (1980) Planform of mantle convection beneath the Pacific. *Nature* 288: 442 446.
McKenzie DP and Nimmo F (1997) Elastic thickness esti mates for Venus from line of sight accelerations. *Icarus* 130: 198 216.
Milsom J (2002) *Field Geophysics*, 3rd edn. Chichester, UK: John Wiley and Sons.
Sandwell DT and Smith WHF (1997) Marine gravity anom aly from Geosat and ERS 1 satellite altimetry. *Journal of Geophysical Research* 105: 10039 10054. (www.ngdc.noaa.gov)
Smallwood JR, Staples RK, Richardson KR, White RS, and the FIRE working group (1999) Crust formed above the Iceland mantle plume: from continental rift to oceanic spreading center. *Journal of Geophysics Research* 104(B10): 22885 22902.
Smallwood JR, Towns MJ, and White RS (2001) The struc ture of the Faeroe Shetland Trough from integrated deep seismic and potential field modelling. *Journal of the Geo logical Society of London* 158: 409 412.
Smallwood JR and White RS (2002) Ridge plume inter action in the North Atlantic and its influence on contin ental breakup and seafloor spreading. In: Jolley DW and Bell BR (eds.) *The North Atlantic Igneous Province: Stratigraphy, Tectonic, Volcanic and Magmatic Pro cesses*, pp. 15 37. London: Geological Society of London, Spec. Publ. 197.
Smith WH and Sandwell DT (1997) Global Sea Floor Top ography from Satellite Altimetry and Ship Depth Sound ings. *Science* 277: 1956 1962.
Telford WM, Geldart LP, and Sheriff RE (1990) *Applied Geophysics*, 2nd edn. Cambridge, UK: Cambridge Uni versity Press.
Watts AB (2001) *Isostasy and Flexure of the Lithosphere*. Cambridge, UK: Cambridge University Press.

Mineral Analysis

N G Ware, Australian National University, Canberra, ACT, Australia

Mineral Analysis

Mineral analysis involves determining the chemical relationships between and within mineral grains. Microanalytical techniques are essential, and methods include X-ray spectrometry and mass spectrometry. Electron probe and laser ablation procedures are commonly used techniques for major and trace element analysis, respectively (*see* **Analytical Methods:** Geochemical Analysis (Including X-Ray)).

A chemical analysis of a mineral is expressed as a table of weight percent (wt.%) of its component elements or oxides. Concentrations lower than about 0.5 wt.% are often expressed as parts per million (ppm) by weight of element. These mineral analyses are easily converted into atomic formulas and thence into percentages of the end-member 'molecules' within the mineral group (see **Table 1**). Mineral analyses are used in descriptive petrology, geothermometry, and geobarometry, and in the understanding of petrogenesis. Sometimes thousands of analyses are collected in the completion of a single research project. Large amounts of data are presented graphically, plotting concentrations of elements or ratios of elements against each other, thus illustrating chemical trends or chemical equilibrium (see **Figure 1**).

In addition to the chemical analysis, a complete description of a mineral requires a knowledge of its crystallography. Both chemical composition and crystallography are required to predict the behaviour of minerals, and hence rocks, in geological processes. The discovery of each new mineral involves the determination of its crystal structure as a matter of routine using X-ray and electron diffraction techniques. Thus, when a monomorphic mineral is identified from its composition, its crystallography follows. Polymorphs may be identified by optical microscopy. Whereas it is sometimes convenient to identify an unknown mineral from its diffraction pattern, and although cell parameters can be used as a rough measure of end-member composition, crystallography no longer plays a major role in quantitative mineral analysis.

It was once necessary to separate a mineral from its parent rock by crushing, followed by use of heavy liquids and magnetic/isodynamic separators. Up to a

Table 1 Analysis of garnet by EMPA/WDS for major elements and by LA ICP MS for trace elements[a]

Oxides (wt.%)		*Atoms (Oxygen = 12)*		*Trace (ppm)*		*End member (%)*	
SiO_2	37.12	Si	2.9985	Ti	172	Almandine	76.4
TiO_2	0.03	Ti	1.9798	Cr	37	Spessartine	7.9
Al_2O_3	20.80	Cr	0.0000	Mn	2716	Grossularite	2.7
Cr_2O_3	<0.01	Fe	2.3149	Ni	12	Pyrope	13.0
FeO	34.27	Mn	0.2386	Cu	32		
MnO	3.49	Zn	0.0024	Zn	331		
ZnO	0.04	Mg	0.3931	Pb	3.2		
MgO	3.27	Ca	0.0809	Y	12		
CaO	0.94	Na	0.0000	La	0.09		
Na_2O	<0.02	K	0.0000	Ce	0.11		
K_2O	<0.01			Yb	07.1		
Total	99.96	Total	8.0098			Total	100.0

[a]EMPA/WDS, Electron microprobe analysis/wavelength dispersive spectrometry; LA ICP MS, laser ablation inductively coupled mass spectrometry.

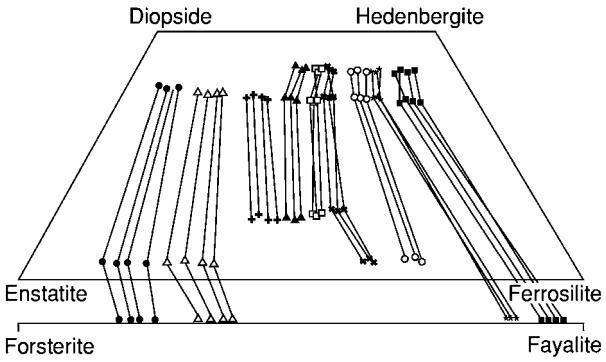


Figure 1 Depiction of plots of multiple mineral analyses: coexisting olivine, orthopyroxene, pigeonite, and two augites at various levels (different symbols) in a differentiated tholeiitic sill. Including duplicates, 246 analyses were used.

gram of a mineral was obtained in this laborious way and then analysed gravimetrically, destroying the aliquot in the process. With the advent of the electron microprobe X-ray analyser in the 1950s and reliable matrix correction methods in the 1960s, non-destructive microanalysis of minerals *in situ* in a polished thin section became possible. To the mineralogist, microanalysis refers not to microgram quantities, but to analysis on the micrometre scale. Today, a variety of microanalytical techniques are available and used almost to the exclusion of bulk techniques, which are reserved for petrological analysis and control of mining operations. Microanalytical techniques available to the mineralogist are listed with their acronyms in Table 2 and are discussed in the following sections. Nearly all major element mineral analysis is now carried out by electron microprobe analysis (EMPA), and laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) is fast becoming the technique of preference for trace element work.

Sample Preparation

The sample must be sectioned and have a polished surface and be of a form so that selection of any point on the surface is rapid and exact. Samples may be mounted in epoxy resin, often using a 25 mm diameter mould, then are well polished, typically finishing with 0.25 μm particle-size diamond paste. The sample is usually top-loaded in a specimen holder: this requires the sides of the mount to be orthogonal, and the polished surface should have minimal bevelling. Alternatively, rock sections may be prepared by gluing them to a glass slide, then grinding them to a 30 μm thickness and polishing well to create a section suitable for both optical microscopy and microanalysis. The thickness of the section must conform to the requirements of the analytical techniques used. In microanalysis, the term 'thin' is usually reserved for a section thin enough for the exciting beam to pass through it. This is essential in analytical electron microscopy (AEM), in which a <500 nm film prepared by ion beam milling is supported on a metallic grid. In proton-induced X-ray emission (PIXE) spectroscopy, it is sometimes advantageous to use sections 50–80 μm thick so that the proton beam can transit the section without causing damage. For micro-X-ray fluorescence (XRF) techniques, the sections should be no thicker than the grain size of the minerals.

When beams of charged particles are used, the sample must be an electrical conductor, and samples are coated with an evaporated conductive film. For

Table 2 Microanalytical techniques used in mineral analysis

Technique	Acronym	Probe	Product	Best practical concentration range	Analytical volume	
					Diameter	Depth
Electron microprobe analysis	EMPA	Electrons	X rays	5 ppm 100%	0.5 4 μm	0.3 8 μm
Scanning electron microscopy	SEM	Electrons	X rays	200 ppm 100%	0.5 4 μm	0.3 8 μm
Transmission electron microscopy (analytical electron microscopy)	TEM (AEM)	Electrons	X rays	100 ppm 100%	5 20 nm	50 100 nm
Proton induced X ray emission	PIXE	Protons	X rays	2 ppm 5%	1 5 μm	40 80 μm
Micro X ray fluorescence	XRF	X rays	X rays	10 ppm 1%	10 μm 1 mm	100 μm 2 mm
Synchrotron X ray fluorescence	SXRF	X rays	X rays	500 ppb 2%	2 μm 1 mm	100 μm 2 mm
Laser ablation inductively coupled plasma mass spectrometry	LA ICP MS	Laser	Ions	1 ppb 0.5%	10 200 μm	10 100 μm
Secondary ion mass spectrometry (sensitive high resolution ion microprobe analysis)	SIMS (SHRIMP)	Ions	Ions	0.1 ppb 1%	5 50 μm	100 nm 2 μm

electron excitation, the resistivity of the film should be less than 10 kΩ mm ^{1} and a carbon coat about 25 nm thick is usually sufficient. Passing a current of 50 to 150 A through sharpened electrodes in a 10 ^{5} torr vacuum or carbon thread at 10 ^{2} torr creates a homogeneous coating at a distance of 5 to 10 cm from the carbon. Secondary ion mass spectrometry and PIXE analysts mostly use 5 nm gold films, which are readily produced by heating a weighed quantity of gold in a tungsten wire basket.

Electron Microprobe Analysis

An electron probe is a beam of electrons accelerated through a high voltage and focused at the surface of the sample. The beam is commonly produced by passing an electric current through a hairpin tungsten filament that is maintained at up to 30 kV in the scanning electron microscope (SEM), 50 kV in the EMPA, and 400 kV in the AEM. The electrons are channeled through a hole in an anode maintained at zero potential and then through a series of magnetic lenses that shape the beam into a circular cross-section and focus it to a size less than the scattering expected in the sample. The energy of the electrons at the surface of the sample (depth 0) is denoted E_0 and is measured in thousands of electron volts (keV).

Electron Scattering

When an electron in the beam interacts with an atom in the sample, the most common result is for the electron to be scattered elastically, with a resulting change in direction and insignificant loss of energy. This is a simple Coulomb attraction between the approaching electron and the multiply charged nucleus. By comparison, the diffuse electron cloud in the atom, repelling in all directions, has a much smaller effect. The probability of elastic scattering increases with atomic number and decreases with higher E_0. It is possible for the electron to reverse direction, usually after several scattering events, and to leave the sample; this is known as back scattering and, at high atomic number and low E_0, more than half the electrons may be back scattered. In the AEM, the very high accelerating voltage ensures that the electrons pass through the thin sample with minimal scattering.

Electrons are also prone to inelastic scattering, which again involves a Coulomb reaction with the nucleus, but here the force of attraction by the opposite charges is actually translated into a reduction of momentum and the loss of kinetic energy is released in the form of an X-ray. These X-rays are known as *bremsstrahlung* ('braking radiation') and may have any energy up to the value of E_0. Inelastic scattering eventually causes the electron to come to rest and so defines the size of the interaction volume: the shape of the volume is determined by elastic scattering. Figure 2 illustrates such volumes.

Characteristic X-Ray Generation

Electrons may also cause ionization by ejecting an electron from its atomic orbital, provided E_0 is greater than the critical excitation energy (E_c). The probability of ionization is much lower than that of inelastic deceleration or inelastic scattering. If an inner orbital electron is ejected, an outer electron may fill the

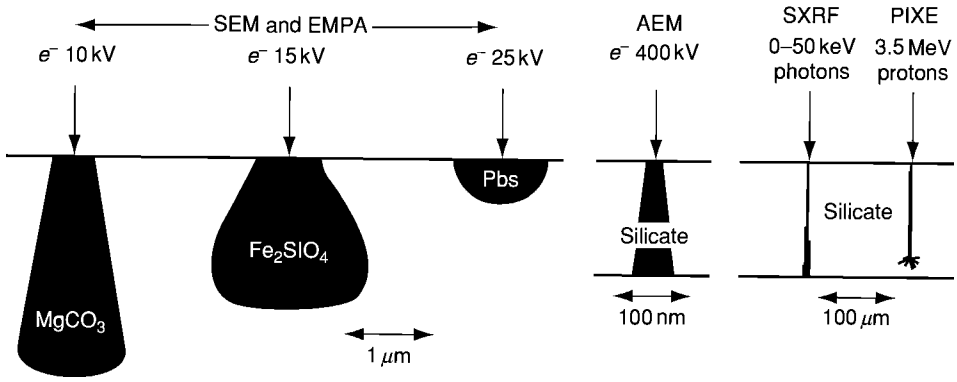


Figure 2 Reaction volumes; the shape of the volume is determined by elastic scattering. SEM, Scanning electron microscopy; EMPA, electron microprobe analysis; AEM, analytical electron microscopy; SXRF, synchrotron X ray fluorescence; PIXE, proton induced X ray emission.

vacancy and an X-ray photon having an energy equal to the difference in the energy levels of the two orbitals involved is emitted. The energy levels are a function of the number of protons in the nucleus, hence the energy of the X-ray is characteristic of the element. Thus, an electron-induced X-ray spectrum consists of characteristic lines superimposed on a continuum of *bremsstrahlung* of significant intensity.

Nearly all electron transitions occur by an electric dipole mechanism that is possible between only a limited number of orbitals; thus X-ray spectra are not complicated, containing only a few lines in any given energy range. The 1920s-era Siegbahn notation for X-ray lines is still in common use: for example, $L\beta_3$ denotes the third most intense line in the second most intense band in the L-shell spectrum; modern nomenclature would describe this line as L_1–M_3 which is the actual electron transition.

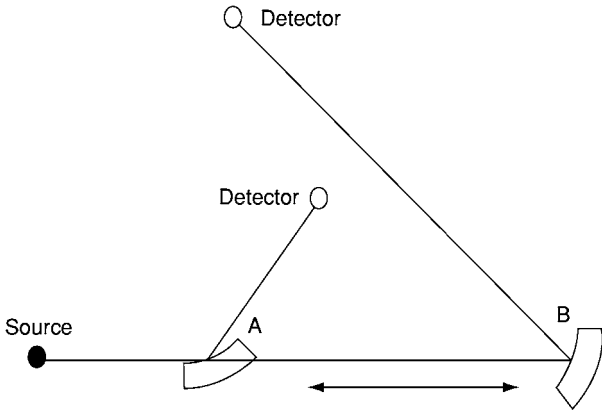


Figure 3 Linear focusing spectrometer, showing the low (A) and high (B) angles.

X-Ray Spectrometry

X-Rays may be measured using three types of spectrometer: (1) the Bragg-law crystal monochromator, usually (mis)named the wavelength-dispersive spectrometer (WDS), (2) the energy-dispersive spectrometer (EDS), which uses semiconductors such as lithium-drifted silicon or intrinsic germanium, and (3) the new X-ray bolometers or microcalorimeters, which operate at temperatures close to absolute zero.

Wavelength-dispersive spectrometer technology The linear focusing spectrometer has proved the most efficient WDS design for the point source of X-rays generated in the electron probe (Figure 3). The analysing crystal is curved in a barrel shape so that the diffracted X-rays fall on the entrance slit of the detector. The crystal is mounted on a worm gear and moves linearly with changing inclination. The detector moves in a parabola so that the angles subtended by the source and detector at the crystal are equal and so that the source-crystal and detector-crystal distances are equal. Most EMPA instruments are fitted with three to five spectrometers, each mounting a selection of two to four crystals that cover a full range of wavelengths. Lithium fluoride (LiF) is commonly used for short wavelengths, pentaerythritol (PET; $C(CH_2OH)_4$) is used for intermediate wavelengths, thallium acid phthalate (TAP; $C_8H_5O_4Tl$) is used for long wavelengths, and various vacuum-deposited multilayers (e.g., tungsten silicides, WSi_x) are used for ultralong wavelengths. Sealed xenon and gas-flow counters are used, the voltage on the wire being adjusted for operation in the proportional region with an internal gain of 10^3 to 10^5. A low-gain preamplifier is positioned as close as possible to the detector and a separate main amplifier shapes the counting pulses for digitization. The X-ray intensities are

measured by collecting counts sequentially at characteristic X-ray peak positions and at predetermined background positions, if possible, on either side of the peak.

The energy-dispersive spectrometer system The detector consisting of lithium-drifted silicon, Si(Li), is a wafer made from a single crystal of silicon having a surface area of 10–30 mm^2. The electrons in the outer atomic orbitals are shared by several neighbouring atoms, forming an essentially covalent bond. The energies of these electrons in the 'valence band' orbitals are about 1.1 eV lower than the semiconductor's 'conduction band levels'. However, it takes about 3.8 eV to promote an electron from the valence to the conduction band. When an X-ray enters the crystal, it may be absorbed in an interaction with an electron of one of the silicon atoms, producing a high-energy photoelectron. This photoelectron dissipates its energy in interactions that promote valence band electrons to the conduction band, leaving holes in the once-filled valence band. The number of electron-hole pairs formed is proportional to the energy of the X-ray: for example, Ca Kα (3.691 keV) would yield on average some 970 electron-hole pairs. A bias of 300–1500 V, depending on the thickness of the Si(Li) chip, is applied and the electrons are swept to the rear, where they enter the preamplifier as a weak pulse, the amplitude of which is proportional to the energy of the X-ray photon. Photo-optic feedback around a field-effect transistor is used in the preamplifier. Room temperature thermal excitation is sufficient to promote electrons from the valence to conduction band, so both the detector and the preamplifier are operated at liquid nitrogen temperatures. Sophisticated electronics are necessary to process the weak pulses in order that they may be stored accurately in a multichannel analyser. The whole X-ray spectrum is stored simultaneously, resulting in an X-ray histogram. Examples of EDS spectra are given in Figure 4; using these spectra as fingerprints, identification of minerals is possible with counting times as short as 100 ms. Within the EDS histogram, it is not always possible to measure background on either side of the peaks, and for quantitative analysis, spectrum deconvolution methods are necessary. Commercial software uses either background modelling or filter fitting.

Energy resolution The resolution of an EDS is quoted as the full-width at half-maximum (FWHM)

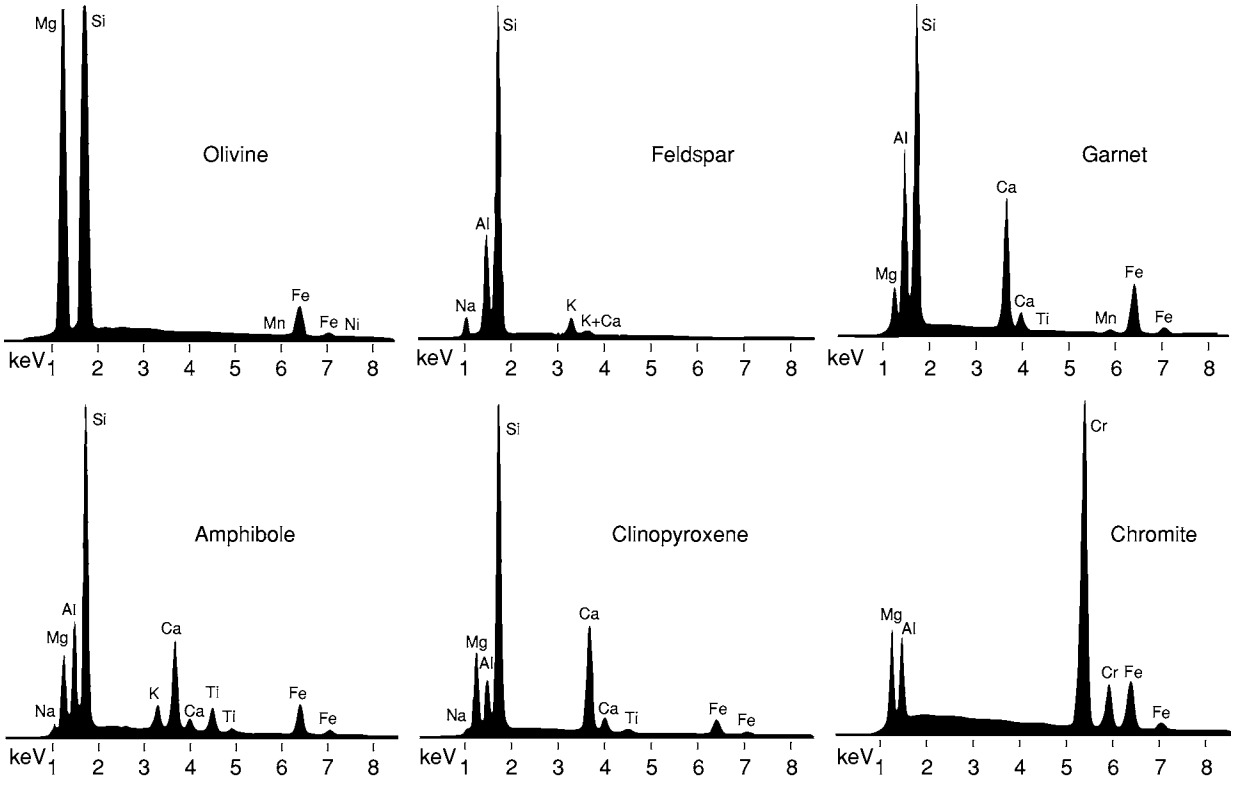


Figure 4 Scanning electron microscope/wavelength dispersive spectrometer spectra of minerals (15 kV, 1 nA, 20 s, 14 000 counts s^{-1}).

Table 3 X ray spectrometer devices

Crystal/device[a]	*Dispersion method*[b]	*Resolution FWHM (eV)*[c]	*Max practical c ount rate (kHZ)*	*Energy range (keV)*	*Collection area (mm^2)*
LiF	WDS	5 10	50	4 12	300 1200
PET	WDS	4 8	50	1.5 6	300 1200
TAP	WDS	3 5	50	0.8 2.2	300 1200
WSi_x	WDS	4 6	40	0.2 0.9	300 1200
Si(Li)	EDS	100 200	20	(0.3) 1 20	10 30
Ge	EDS	90 190	20	1 40	10 30
SDD	EDS	110 250	500	1 20	100 400
Bolometer	EDS	2 10	1	0.2 8	0.5 1

[a]LiF, Lithium fluoride; PET, pentaerythritol; TAP, thallium acid phthalate; WSi_x, tungsten silicides; Ge, germanium; SDD, silicon drift detector.
[b]WDS, Wavelength dispersive spectrometry; EDS, energy dispersive spectrometry.
[c]FWHM, Full width half maximum.

of the Mn Kα peak. This measure is chosen because readily available ^{55}Fe is a source of this X-ray line. For the Si(Li) detector, the FWHM at energy E is given by $\mathrm{FWHM}_E = (\mathrm{FWHM}_0^2 + 21.1FE)^{0.5}$, where FWHM_0 is the resolution at energy 0 and F is the Fano factor, which is a measure of the statistical fluctuations in the ionization and charge collection processes. Table 3 lists the performance of a range of X-ray spectrometers.

Other energy-dispersive spectrometer technology Germanium detectors have properties similar to those of Si(Li) detectors and are preferred for use at higher energies. They are found on AEM, PIXE, and synchrotron X-ray fluorescence (SXRF) instruments. The silicon drift detector (SDD) is based on charge-coupled semiconductor technology and can provide energy resolution similar to that of the monolithic Si-crystal EDS, but at a count rate of 500 kHz. Furthermore, an energy resolution of 140 eV can be achieved at only −13°C. The detector area can be made as large as 400 mm^2 so that low currents can be used for high count rates. It is possible to count at more than 1 MHz, but the resolution degrades as the count rate increases. This makes the detector unsuitable for quantitative analysis but ideal for mapping of mineral grains.

X-Ray bolometry has been developed using thin-film Ag microcalorimeters, transition edge sensors, and superconducting quantum interference devices. Such detectors have energy resolutions down to 2 eV and count rates of only 1 kHz. In theory, arrays of these millimetre-sized devices could be constructed giving a high overall count rate. The operating temperature is 70–100 mK and it is possible to achieve this using multistage Peltier cooling and an adiabatic demagnetization refrigerator.

Matrix Corrections

X-Ray intensities are measured in units of counts per second per nanoampere of beam current. The weight percent concentration of an element in a sample, C_{samp}, is related to the characteristic X-ray intensity, I_{samp}, by the equation $C_{\mathrm{samp}} = C_{\mathrm{std}}(I_{\mathrm{samp}}/I_{\mathrm{std}})([\mathrm{MATRIX}]_{\mathrm{samp}}/[\mathrm{MATRIX}]_{\mathrm{std}})$, where [MATRIX] denotes the effect of the chemical composition of the matrix on the X-ray intensity and 'std' refers to a standard of known composition.

There are four approaches to matrix corrections:

1. Empirical methods assume that each element linearly influences the X-ray intensity of each other element. A table of coefficients, analysed element against matrix element, is drawn up using extrapolations from measurements of binary alloys and solid solution series. These are known as alpha coefficients.
2. The ZAF corrections separately compute the effects of atomic number (Z), absorption (A), and secondary fluorescence (F): $ZAF = R/S\,f(\chi)(1+\gamma)$, where R is the back-scattering fraction and S is the X-ray generation factor due to stopping power; both of these are functions of atomic number. The function of the mass attenuation coefficient, $f(\chi)$, corrects for the absorption of the X-rays as they pass through the sample towards the detector. The additional contribution when a matrix X-ray fluoresces an analysed element ($E_{\mathrm{m}} > E_{\mathrm{c,a}}$) is represented by γ.
3. The $\phi(\rho z)$ methods: ϕ is defined as the ratio of the X-ray intensity from a thin layer, δz, of sample at a mass depth (ρz) to the X-ray intensity of a similar layer isolated in space. The $\phi(\rho z)$ procedures integrate this X-ray intensity ratio function, corrected for multicomponent systems, from the surface to a

depth where ϕ becomes zero. Work by several groups spanning 30 years has established $\phi(\rho z)$ methods that give reliable matrix corrections for nearly all mineral analysis.

4. The quantitative microanalysis procedure developed by J L Pouchou and F Pichoir, and so named the PAP procedure, has affinities with both the *ZAF* and the $\phi(\rho z)$ methods. The depth distribution of X-ray generation is modelled using parts of two adjoining parabolas, functions that are easy to integrate. Stopping power and absorption are carried out together. Fluorescence and back scattering are corrected separately.

The EMPA instrument is designed specifically for X-ray analysis and current designs provide up to five WDSs, one EDS, and an optical microscope built round the electron optical column. Scanning-beam imaging by back-scattered electrons, secondary electrons, and cathodoluminescence is also possible. The specimen stage and spectrometers are automated and software has been developed that has transformed the electron microprobe analyser into a turnkey instrument.

The Scanning Electron Microscope

Good quality mineral analyses may by obtained by scanning electron microscopy (SEM) and energy-dispersive spectrometry. SEM does not incorporate an optical microscope and the minerals must be located using back-scattered electron imaging. The energy-dispersive spectrometer must be robust and properly calibrated, requires stable electronics, and an appropriate spectrum deconvolution method must be employed. Many systems offer 'standardless analyses'; these methods work well over a restricted energy range but should be used with caution in mineral analysis, in which the energy range extends from Na (1.0 keV) to Fe (6.4 keV). Notwithstanding the high peak-to-background ratio, EDS analysis, properly executed, should have a better precision than WDS has for concentrations greater than 5 wt.%.

The Analytical Transmission Electron Microscope

In analytical electron microscopy, sufficient current may be focused on a region as small as 5 nm in diameter, so that an X-ray flux greater than 1000 counts s^{-1} measured by an EDS system may be generated in a 150-nm-thick film. Thus, extremely small domains of mineral grains may be analysed. The high value of E_0 helps to improve the peak-to-background ratios and enables the analysis of X-rays at the upper energy limits of the EDS detector. The intensity of the X-ray spectrum is a function of the indeterminate thickness of the sample. Quantification is attained by using ratios of peak intensities to that of a common element (Si for silicates, Fe for opaques) and assuming a value (usually 100%) for the sum of the analysed components: $C_x/C_{Si} = k_{x,Si}(I_x/I_{Si})$, $\sum C_x = 1.0$, where C is concentration and I is intensity. The calibration constants, $k_{x,Si}$, are evaluated empirically and correct for the difference in EDS efficiency and resolution at different energies. Matrix corrections are slight and are often ignored but can be applied using a rough estimation of the film thickness. Beryllium-window Si or Ge detectors that do not detect elements lower than sodium are used, but the technique of electron energy loss spectrometry (EELS) can be used in the analytical transmission electron microscope to determine elements down to beryllium.

Proton Induced X-Ray Emission

In the proton probe, the protons are typically accelerated through 2.5–3.5 MeV. This is below the threshold of the 8 MeV required for atom smashing yet high enough to generate a reasonable flux of X-rays. The mass of the proton is 1837 times that of the electron, and this large mass, combined with the higher energy, predicates very low scattering, either elastic or inelastic. The X-ray spectrum has a very low background and X-rays are generated to high E_c. The limiting factor for the use of high-energy characteristic X-rays is the capability of the X-ray detector. The protons may be focused down to submicrometre beams using collimation and a series of quadrupole magnetic lenses. The penetration of the proton beam is considerably longer than it is in the electron microprobe analyser: 3.5 MeV protons have a range of 100 μm in aluminium. Most of the X-rays are detected from depths of 20–50 μm, where the protons penetrate, causing very little damage. However, just before coming to rest, much of the kinetic energy is absorbed by the sample and considerable damage to the crystal lattice results (see Figure 2). The protons end up forming hydroxyl ions, free hydrogen, and hydrides.

The proton trajectory in the sample is essentially linear, with a smooth deceleration of about 100 eV per collision. The mechanisms of energy loss and ionization are well understood, and the algorithm of integration of PIXE X-ray yields along the path of the beam provides the foundation for a standardless microanalytical method. However, unavoidable uncertainties in

mass attenuation coefficients, when applied over the relatively long distances, can give rise to unacceptable errors in the analysis of high concentrations of elements such as Na, Mg, Al, and Si.

X-Ray Fluorescence

A beam of X-rays is not much scattered by solid matter and is absorbed only when ionization occurs. Ionization is caused when the energy of the X-ray waveform resonates with the energy of the electron orbital, increasing the amplitude of vibration so that the electron leaves the atom. The probability of ionization increases exponentially as the energy of the photon approaches the critical excitation energy. Thus, a primary X-ray beam is an efficient producer of characteristic X-rays in a sample. There is little background and the peak-to-background ratios are even better than is obtained in PIXE.

Commercial X-ray microprobes are available. A low-voltage X-ray tube and a waveguide focuses the X-ray beam down to a 2 μm spot. In common with PIXE, the secondary X-ray intensity is too low for WDS and the spectrometers used are conventional Be-window Si(Li) detectors.

The intense 'white' radiation in a synchrotron beam line has been used in mineral analysis since the mid-1980s; X-ray photon fluxes have since increased by factors of 10^5. Simple collimation with fine apertures can create a $<10\,\mu$m beam that may be used to detect X-rays up to 40 keV in energy. Focusing mirrors (the Kirkpatrick–Baez method) produce a $<2\,\mu$m spot, but the maximum useful energy is about 10 keV. Phase zone plates can obtain a nanometre focus and are used for X-ray mapping and microtomography.

Laser Ablation

Laser probing started in the 1970s with ultraviolet (UV; e.g., 266 nm) lasers focused with multiple-lens UV optics. The ablated material from a single laser pulse was ionized by a second laser beam that horizontally flooded the space above the sample, and the ions were then extracted through a tube in the optics and into a time-of-flight (TOF) mass spectrometer (MS). This technology has improved so that spatial resolution is $<0.5\,\mu$m and mass resolution is >5000 $M/\Delta M$. Resonant postionization techniques have made great improvements in sensitivity. These instruments are dedicated mass microprobes and are useful in sensitive qualitative applications, but there are problems in attempting quantitative analysis of refractory elements. Unfortunately, so many of the trace elements of interest to the mineralogist are refractory.

Since the mid-1980s, lasers have been used in conjunction with inductively coupled plasma mass spectrometers to form very successful laser microprobes. Today, the technique of laser ablation in conjunction with ICP-MS is used for the majority of published trace element mineral analyses. The polished sample is inserted in a cell and an inert gas is passed over the surface. The cell in Figure 5 is of a sophisticated design; most ablation cells use a cylindrical box and only one gas, which serves both as an ablating and a carrier medium. A UV laser, collimated to 20–200 μm and pulsing typically at 5 Hz, ablates the sample and the material is carried by the gas, usually argon, into a plasma torch, where most of the material is converted into monatomic cations. These ions are usually analysed by a quadrupole mass spectrometer.

The laser ablates the sample to a depth approximating the radius of the beam. Several hundred pulses are used and a stream of material enters the plasma torch over a period of up to several minutes. The ionized product of the ICP enters the high vacuum of the MS through a series of metal cones having small apertures in their tips. As little as 0.01% of the sample may enter the MS. The MS is cycled to detect the required isotopes in sequence; a cycle takes of the order of a second, and usually 10–30 isotopes are counted. Each cycle is deemed to analyse a 'slice' of the sample and the counts of each isotope in each slice are recorded by the on-board computer. Notwithstanding the inefficiency of transporting material into the MS, counts of 10^3 to 10^5 ppm ^{1} are obtained. As analysis proceeds, material can build up on the surfaces within the equipment and isotopes may be detected when the laser is switched off. The MS is cycled for a period before starting ablation so that the background levels may be determined. Some operators are concerned that the act of running the laser

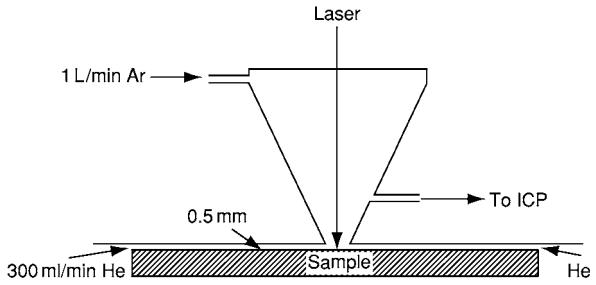


Figure 5 Laser ablation cell (the HelEx system). ICP, Inductively coupled plasma.

may of itself dislodge material in the system, thus adding to the background. As a precaution, an ultrapure silicon standard may be analysed to quantify this effect.

The National Institute of Standards and Technology (US Department of Commerce) prepares and issues a range of glass standards that contain 61 elements at trace concentration, at approximately 50 and 500 ppm. These standards are commonly used in LA-ICP-MS and are measured repeatedly throughout an analytical session, perhaps once to every 10 or 20 sample measurements. Corrections for drift in the performance of the equipment with each sample analysis are applied linearly between standard measurements.

Quantification of the isotope counts requires knowledge of the concentration of at least one element, used as an internal standard; concentrations of other elements may be determined from the ratios of isotopes to that of the internal standard. A convenient element to use is calcium, which is present as a major element in most silicates and has five stable isotopes of widely ranging abundance. It is usually possible to select a calcium isotope giving a signal similar to that of the trace elements. Other elements can be used for example, nickel in olivine, vanadium in oxides, and titanium in micas. With effort, the EMPA laboratory can provide such internal standard concentrations to ± 10 ppm at the 1000 ppm level, but isotopes of major elements such as Si and Mg can give good results. In sulphides, the sulphur concentration is usually known and, being an electronegative element, the cation signal is weak enough to be comparable with those of the trace elements. During ablation, much material condenses in and around the ablation pit, and the more refractory an element, the less likely it will be carried away by the gas. Thus fractionation processes occur even when the laser couples well with the mineral, and there is always a crater rim to the ablation pit (e.g., **Figure 6A**). In general the worse the coupling (**Figure 6D**), the greater the fractionation.

The first stage in quantification is to obtain isotope ratios corrected for background and fractionation. The background signal obtained with the laser off is averaged to give intensity values per slice, and these are subtracted (together with the values from the 'null' pure silicon standard if any) from the isotope signals measured with the laser on. Then ratios are calculated for each slice. These ratios, if plotted against slice number, will have a positive slope if the unknown undergoes less fractionation than the internal standard does; if not, then the slope will be negative. Either way, the plots are regressed to the point at which the laser is switched on and the value there is adopted for further calculation. Linear regression is often adequate; some operators prefer a polynomial. Anomalous slices, such as those containing inclusions in the mineral, may be excluded from the regression. Editing the data is facilitated by a good graphics computer program, but operations with a simple spreadsheet are adequate.

Quantification of the isotope ratios is continued by adjusting them with reference to the changes in ratios in the glass standard taken before and after the sample. Finally, the concentration of element x in the sample, $C_{x,samp}$, is given by the following equation:

$$C_{x,samp} = C_{int,samp}(I_{x,samp}/I_{int,samp})(I_{int,std}/I_{x,std})(C_{x,std}/C_{int,std})$$

where 'std' denotes the glass standard and 'int' is the internal element.

Differences in the coupling of the laser and hence the process of ablation between the glass standard and the sample are responsible for the major source of error in LA-ICP-MS. Another error is in the failure to predict overlaps on the analysed isotope. Overlap of isotopes of different elements and equal mass is either avoidable or readily quantified, but overlap from argon-sample dimers and from doubly charged ions may not be so obvious.

The Ion Microprobe

In secondary ion mass spectrometry (SIMS) and in sensitive high-resolution ion microprobe (SHRIMP) analysis (the 'big brother' of SIMS), beams of O^-, O^+, O_2^+, or Cs^+ at 10–20 keV sputter the surface of the sample, yielding a mixture of ions, molecules, and plasma. Three types of mass spectrometers are used: magnetic sector, quadrupole, and time-of-flight. Although few useful ions are produced, unlike the LA-ICP-MS, nearly all the ions can be analysed by the mass spectrometer, and SIMS is a more sensitive technique. Erosion of the sample is usually 1–10 nm s^{-1}, which is much slower than laser ablation and much less sample is required. In **Figure 6**, the volume of material excavated in the SIMS pit is 0.3% that of the laser pits. SIMS is primarily an isotope ratio technique, but quantitative elemental analysis at very low levels is possible.

In contrast to EMPA, a general theory for matrix corrections in SIMS may never eventuate. Several specialized methods have been applied; for example, in the infinite velocity method, emission velocities, calculated from experimentally measured energy distributions, are extrapolated to infinite velocity, where there are no matrix effects. This approach works well

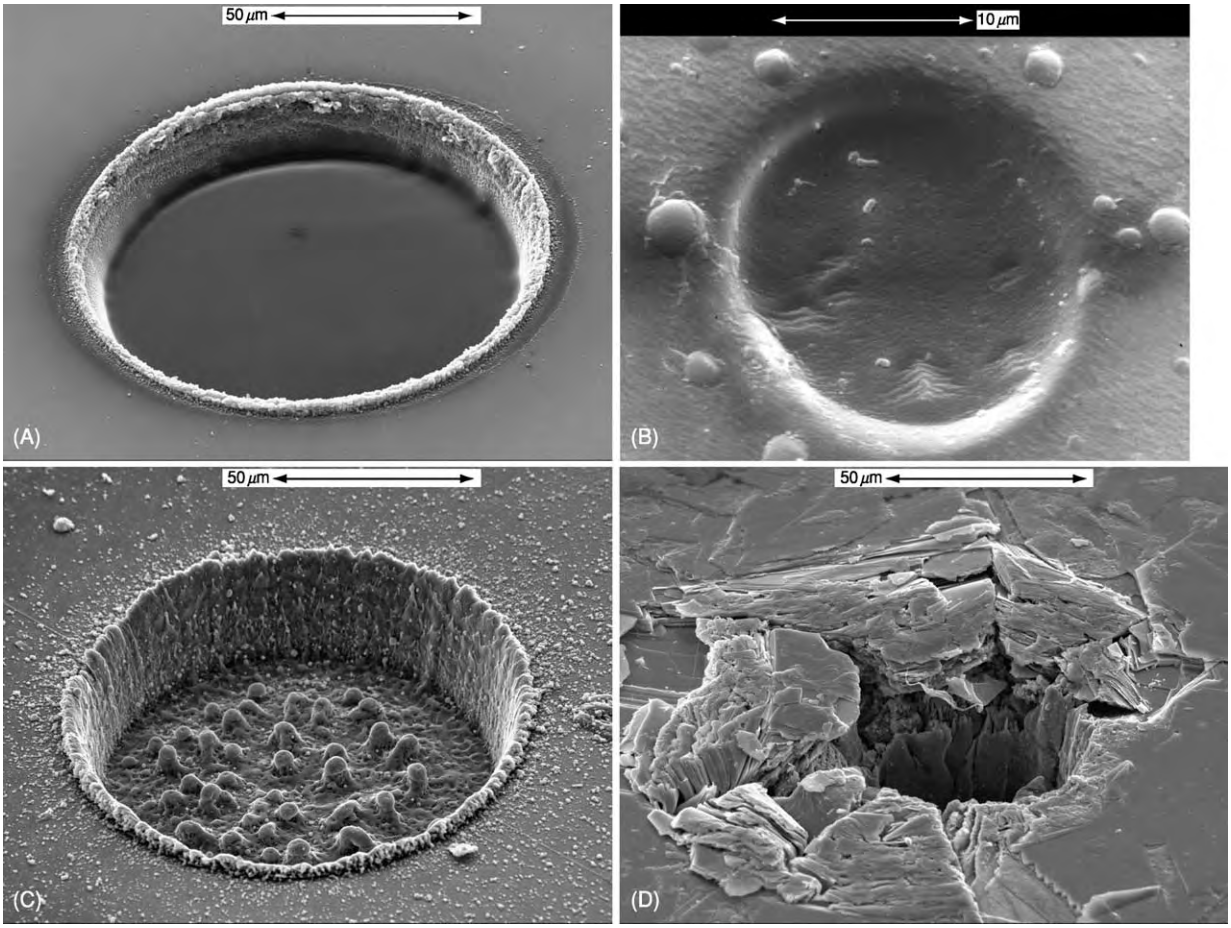


Figure 6 Scanning electron microscope images of ablation pits, using (A) a 193 nm laser on calcite, (B) secondary ion mass spectrometry, with O_2^- on zircon, and (C) 193 nm and (D) 226 nm lasers on molybdenite.

for trace elements implanted in simple matrices such as high-purity silicon. In mineral analysis, standards with the same crystal lattice as the unknown are required. Nevertheless, various laboratories have set up routine procedures for SIMS analysis in applications involving rare-earth elements, platinum group elements, and light elements, including hydrogen.

Compositional Mapping

By moving the automated stage under the beam in any microanalytical instrument, it is possible to build up an array of data that may be transformed into false-colour maps of the sort shown in Figure 7. This has become a routine overnight procedure in many EMPA laboratories: the X-ray peak intensities may be recorded for each position together with the backscattered electron signal, which furnishes both an image of the area scanned and a template for the *bremsstrahlung* background. For some applications, it is possible to obtain a full quantitative analysis at each point (i.e., pixel) with acceptable precision. Usually the colour scale is calibrated roughly (as in Figure 7) from the software's calibration file and matrix effects are ignored. In addition to EMPA, compositional maps have been obtained using PIXE, SXRF, and SIMS. LA-ICP-MS does not have high spatial resolution but line scans are attempted with useful results.

Other Mineral Analysis Methods

Analysis of OH , CO_3^{2} , B, Be, and Li and the allocation of iron between Fe^{2+} and Fe^{3+} pose problems in mineral analysis by the methods outlined in the preceding sections. Of the light elements, only F may be analysed by EMPA with an accuracy comparable with heavier elements. However, B, Be, and Li may

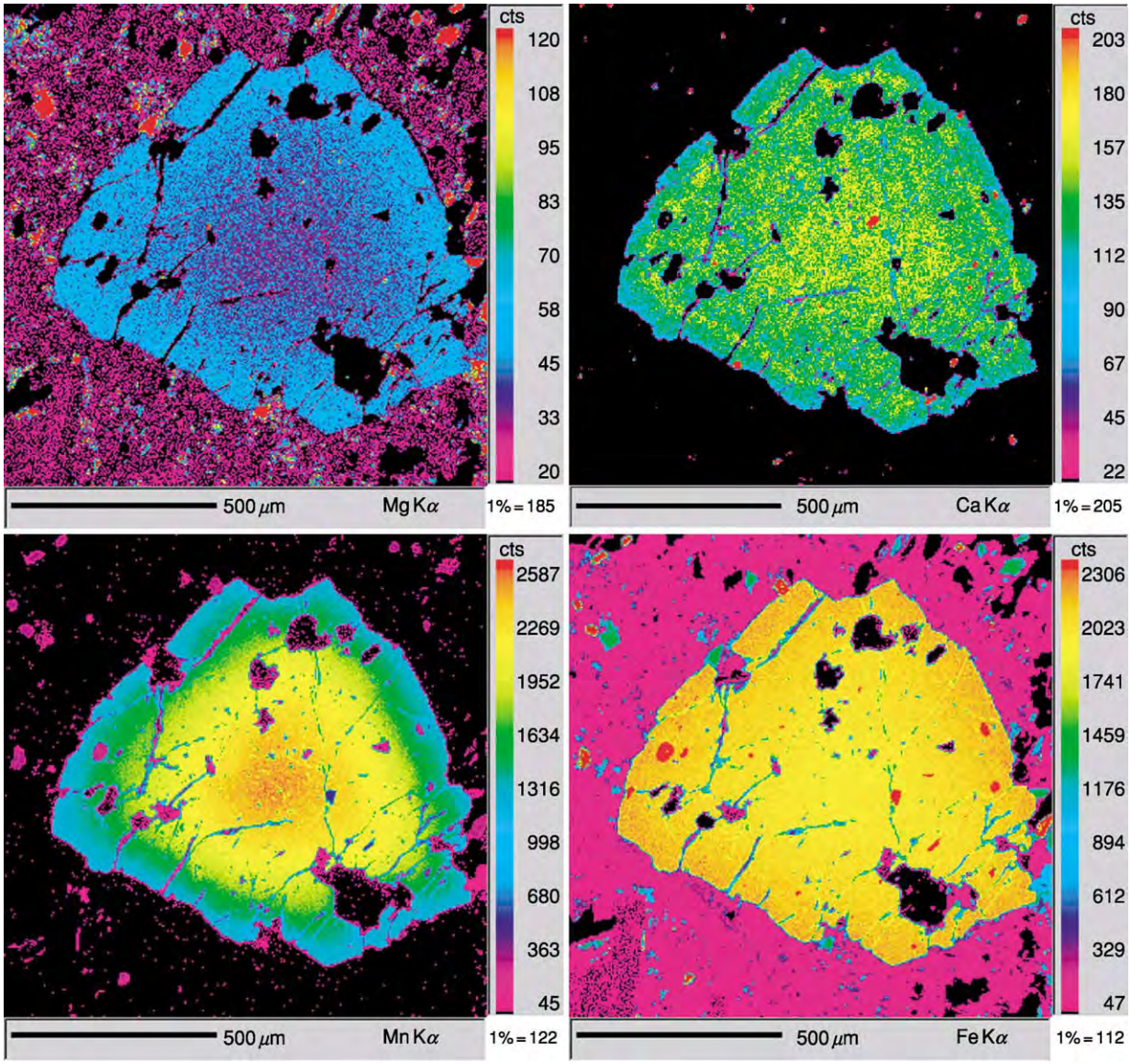


Figure 7 Electron microprobe analysis/wavelength dispersive spectrometry maps of garnet $(Mg,Ca,Mn,Fe)_3Al_2Si_3O_{12}$ (600×600 pixels, 50 ms pixel $^{-1}$, with a 25 kV, 100 nA, 1 μm beam).

be determined by LA-ICP-MS, though at a limit of detection >50 ppm and with indeterminate accuracy at concentrations corresponding to borates and beryllates. Light elements are readily detected by SIMS but quantification is beset with uncertainties.

The local atomic environment around the nucleus, including the electronic, chemical, and magnetic state, may be studied using Mössbauer spectroscopy. The Mössbauer effect is the recoilless absorption and emission of γ-rays by specific nuclei in a solid, and the spectroscopy of ^{57}Fe has been much studied with respect to mineral analysis.

Fourier transform infrared spectroscopy using an optical microscope can give quantitative information about anions such as OH and CO_3^{2-}, but the thickness of the slide must be measured accurately. Multiple valency may be determined by X-ray absorption near-edge spectroscopy (XANES), which is performed on the synchrotron, and by X-ray photoelectric spectroscopy (XPS), which requires an ultrahigh vacuum and analyses the outer 10 nm of the sample.

For most silicates, the FeO/Fe_2O_3 ratio may be estimated by allocation of Fe to FeO and Fe_2O_3 so

that the cation total equals the theoretical amount. A general equation can be used:

$$\text{wt.\% Fe(trivalent)} = (\text{total} - \text{theoretical})/\text{theoretical} \times \text{wt.\% O} \times 6.98125$$

In the garnet analysis in Table 1, the cation total is 8.0098 and the theoretical total is 8.0000; the wt.% oxygen is 39.54%, which is the sum of the oxides, 99.96%, less the sum of the elements. Application of this formula gives 0.48% Fe_2O_3 and 33.83% FeO, and recalculation of the mineral formula gives exactly 8.0000 cations. It is possible to analyse directly for oxygen with the EMPA, and if this is done with care, a similar result can be obtained but at the cost of extra instrument time.

See Also

Analytical Methods: Fission Track Analysis; Geochemical Analysis (Including X-Ray). **Minerals:** Definition and Classification; Micas; Olivines; Sulphides.

Further Reading

Cabri LJ and Vaughan DJ (eds.) (1998) *Modern Approaches to Ore and Environmental Mineralogy, Short Course Series*, vol. 27. Ottawa: Mineralogical Association of Canada.

Deer WA, Howie RA, and Zussman J (1997) *Rock Forming Minerals* (5 vols.). Bath, UK: Geological Society Publ. House.

Henderson G and Baker D (eds.) (2002) *Synchrotron Radiation: Earth, Environmental and Material Science Applications, Short Course Series*, vol. 30 Ottawa: Mineralogical Association of Canada.

Hurlbut CS and Sharp WE (1998) *Dana's Minerals and How to Study Them,* 4th edn. New York: Wiley.

Ireland TR (1995) Ion microprobe mass spectrometry: techniques and applications in cosmochemistry, geochemistry and goechronology. In: Hyman M and Rowe M (eds.) *Advances in Analytical Geochemistry,* vol. 2, pp. 1 118. Greenwich, CT: JAI Press.

Johansson SAE, Campbell JL, and Malmqvist KG (1995) *Particle Induced X ray Emission Spectrometry (PIXE).* New York: Wiley.

McCammon C (1995) Mossbauer spectroscopy of minerals. In: Ahrens TJ (ed.) *A Handbook of Physical Constants: Mineral Physics and Crystallography,* vol. 2, pp. 332 347. Washington, DC: American Geophysical Union.

Potts PJ (1992) *A Handbook of Silicate Rock Analysis.* London: Blackie Academic & Professional.

Reed SBJ (1993) *Electron Microprobe Analysis,* 2nd edn. Cambridge, UK: Cambridge University Press.

Ryan CG (1995) The nuclear microprobe as a probe of Earth structure and geological processes. *Nuclear Instruments and Methods* B104: 377 394.

Schulze D, Bertsch P, and Stucki J (eds.) (1999) *Synchrotron X ray Methods in Clay Science.* Aurora, CO: Clay Minerals Society of America.

Sylvester P (ed.) (2001) *Laser Ablation ICPMS in the Earth Sciences. Principles and Applications, Short Course Series*, vol. 29. Ottawa: Mineralogical Association of Canada.

ANDES

S M Kay, Cornell University, Ithaca, NY, USA
C Mpodozis, SIPETROL SA, Santiago, Chile
V A Ramos, Universidad de Buenos Aires, Buenos Aires, Argentina

Introduction

The Andean mountains are the type example of an 'Andean'-type subduction zone characterized by subduction of an oceanic plate beneath a continental margin and uplift of a mountain range without continental collision. They extend some 8000 km from Venezuela to Tierra del Fuego and are the major morphological feature of South America. On the Earth's continents, they include the highest active volcanoes (>6800 m), the highest peaks outside of the Himalayas (*ca.* 7000 m), the thickest crust (>70 km), the second greatest plateau in height and area (after Tibet), the most important volcanic plateau with the largest Tertiary ignimbrite calderas, and among the most shortened continental crust, deepest foreland sedimentary basins, and largest and richest precious metal (Cu, Au, Ag) and oil deposits. The central Andes are also the type example of the effects of shallowly subducting oceanic plates and of non-accreting margins where continental lithosphere has been removed by the subduction erosion process. The evolution of the Andes began in the Jurassic coincident with the arc system that developed above subducting oceanic plates along the western margin of South America during and after the breakup of the

Mesozoic Pangaean supercontinent. The history of the Andes is predominantly a story of magmatism, uplift related to contractional deformation, intervening episodes of oblique extension, collision of oceanic terranes in the north, formation of sedimentary basins, mineralization, loss of continental crust by fore-arc subduction erosion, and removal of the base of overthickened crust by delamination. The mechanisms of uplift and crustal thickening along with the amount, timing and fate of removed continental lithosphere are hotly debated topics.

Principal Geological Features of the Modern Andes

Subducting Oceanic Crust and Distribution and Character of Andean Magmatism

The morphology and geology of the modern Andes are strongly influenced by the age, geometry, and morphology of the subducting oceanic plates (Figures 1 and 2). A first-order feature related to these subducting plates is the division of the active volcanic arc into the Northern (NVZ), Central (CVZ), Southern (SVZ) and Austral (AVZ) Volcanic zones (Figures 3–6). The NVZ, CVZ and SVZ are underlain by segments of the Nazca Plate that are subducting at an angle of ~20–30°, and in which the magmas are principally generated by hydrous fluxing and melting of the mantle wedge. CVZ and northern SVZ magmas, erupted through the thick crust of the Central Andes, are primarily andesites and dacites, whereas central and southern SVZ and NVZ magmas, erupted through thinner crust, are primarily basalts and mafic andesites. Between these segments are the Peruvian and Chilean (or Pampean) amagmatic flat slab segments under which the subducting Nazca Plate forms a flat bench at ~100 km that can extend up to ~300 km east of the high Andes. The near absence of an asthenospheric wedge accounts for the lack of volcanism. The origin of the shallowly subducting segments of the Nazca Plate has been debated and variously attributed to subduction of the Juan Fernandez and Nazca ridges near their southern margins or complex interactions between the underriding and overriding plates. All of the models call for a 'collision' between a shallowly dipping Nazca Plate and the overriding South American Plate.

Other factors come into play at the northern and southern ends of the Andes where the geometry of the subducting plate is less well known. To the north, the NVZ is flanked by the amagmatic Bucaramanga segment under which the weakly defined subducting Caribbean Plate appears to dip at ~20°. In the south,

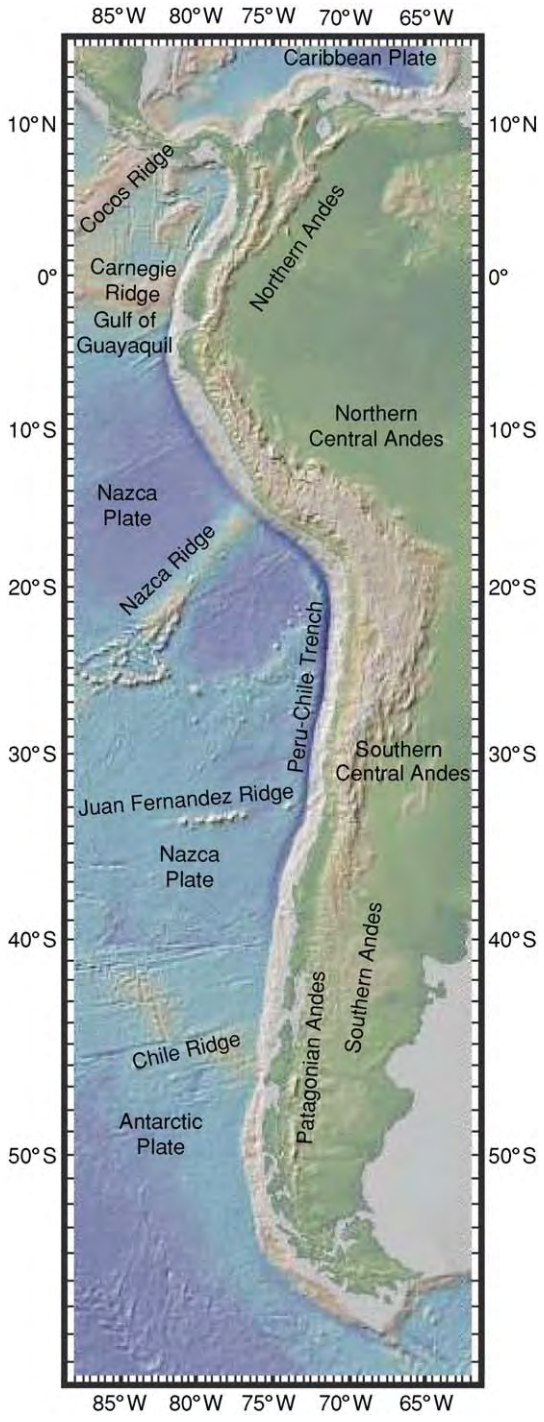


Figure 1 Digital elevation model (DEM) of western South America and surrounding oceans based on global bathymetry database at the Lamont Doherty Observatory of Columbia University. The figure shows major features on the subducting oceanic plate and the correspondence with the division of the Andes into the Northern Andes bounded to the south by the Gulf of Guayaquil, the Central Andes bounded to the south by the Juan Fernandez Ridge, and the Southern Andes.

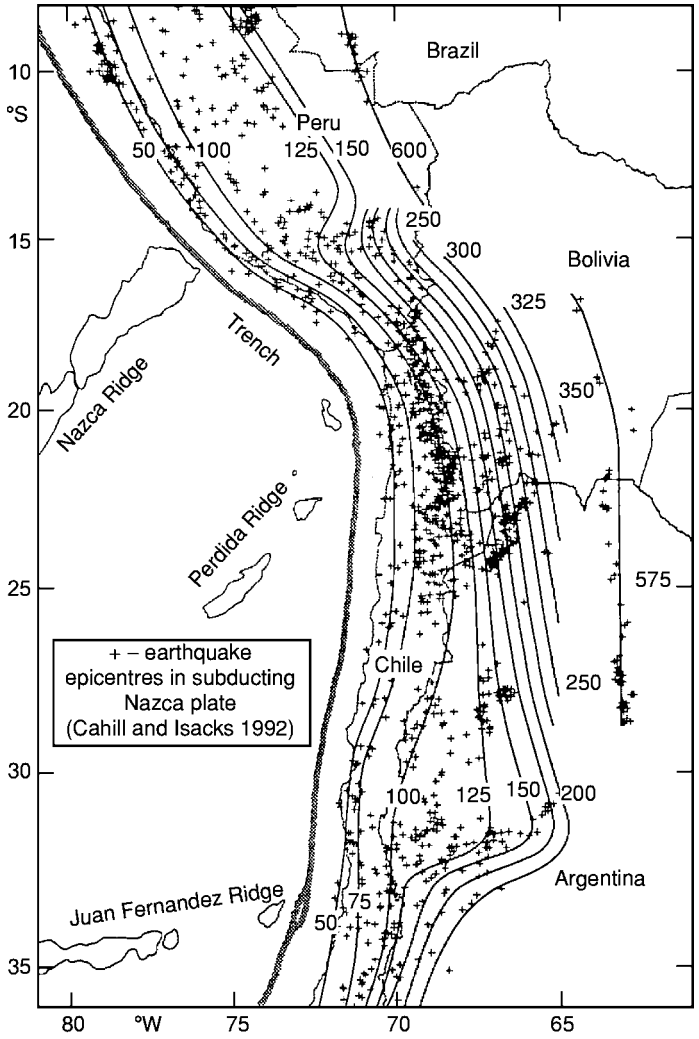


Figure 2 Map showing the Benioff zone geometry of the portion of the Nazca oceanic Plate subducting beneath the Central Andes. Major north to south changes in distribution and style of volcanism, basin development, and deformational styles can be correlated to a first order with the shape of the Nazca Plate. (Reproduced with permission from Cahill TA and Isacks BL (1992) Seismicity and shape of the subducted Nazca Plate. *Journal of Geophysical Research* 97: 17 503 17 529.)

the SVZ is separated from the AVZ by a volcanic gap that coincides with the Chile Triple Junction where the Chile Rise is colliding with the Chile Trench (Figure 7). The net convergence rate of the South American Plate with the Nazca Plate is ~9 cm year ^{1} whereas that with the Antarctic Plate is ~2 cm year ^{1}. Magmatism is absent in this region as the subducting slab is too young and hot to provide the volatiles to flux melting of the mantle wedge. The andesitic to dacitic 'adakitic' magmas of the AVZ are distinctive in that they are attributed to melting of the young hot subducting Antarctic Plate. The most convincing slab melt 'adakites' erupted anywhere in continental crust are the ~14–12 Ma Patagonian adakites (e.g. Cerro Pampa) that are attributed to melting of the trailing edge of the subducted Nazca Plate near the time of ridge collision.

Character of the Ranges, Basins and Faults of the Northern, Central and Southern Andes

The principal ranges and basins of the Andes reflect both the geometry of the subducting plate and the pre-existing continental crust and mantle lithosphere. The Andes are generally discussed in terms of a northern, a central, and a southern sector. Here the limit between the Northern and Central Andes is near the northern boundary of the Peruvian flat slab at ~4° S, and the limit between the Central and Southern Andes is at the southern margin of the Chilean Flat Slab near 33° S (Figure 1).

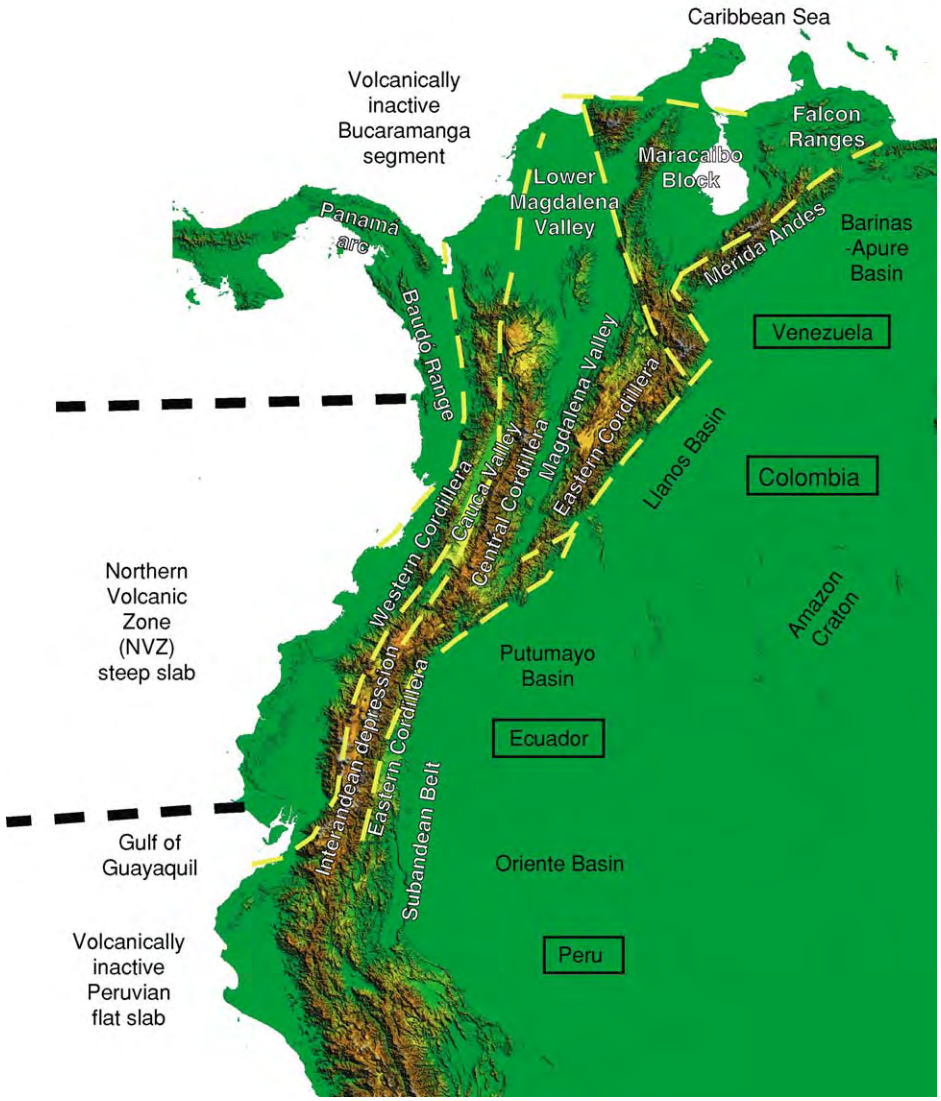


Figure 3 Digital elevation map (DEM) based on 1 km grid SRTM (Shuttle Radar Topographic Mission) satellite data (available through the United States Geological Survey) assembled by the Cornell University Andes group showing the major features of the northern Andes (Venezuelan, Colombian, and Ecuadorian Andes). Elevations are indicated by colours (green is lowest and white is highest). Dark regions are areas of rapid elevation change. Dashed yellow lines are principal faults. They include the fault zone east of the Baudó Block that corresponds with a Late Miocene suture, and the Romeral Fault that runs through the western part of the Central Cordillera and marks the eastern limit of oceanic terranes accreted in the Cretaceous. Faults surrounding the Maracaibo Block are the Bucaramanga Fault on the west, the Boconó Fault in the Mérida Andes on the east, and the Oca Fault on the north. Regions of active and inactive volcanism are bounded by dashed lines that intersect the coast at the boundaries.

Northern Andes The Northern Andes include the Andes of Venezuela, Colombia, and Ecuador, and are bounded to the south by the Gulf of Guayaquil at ~3° S (Figure 3). A distinctive feature of this region, well recognized by Gansser in the 1970s, is that the western part of the Northern Andes has been built on oceanic terranes accreted to South America as the Andes evolved, whereas, except for the southernmost part, the rest of the Andes has been built on continental basement that was already part of South America. Pervasive strike-slip deformation has played a major role in the evolution of the Northern Andes and is important in accounting for their extreme width. The Northern Andes are discussed relative to three segments that basically coincide with the political boundaries.

Venezuelan Andes The Venezuelan Andes face the Caribbean Plate and are geomorphologically the north-eastern extension of the Eastern Cordillera of

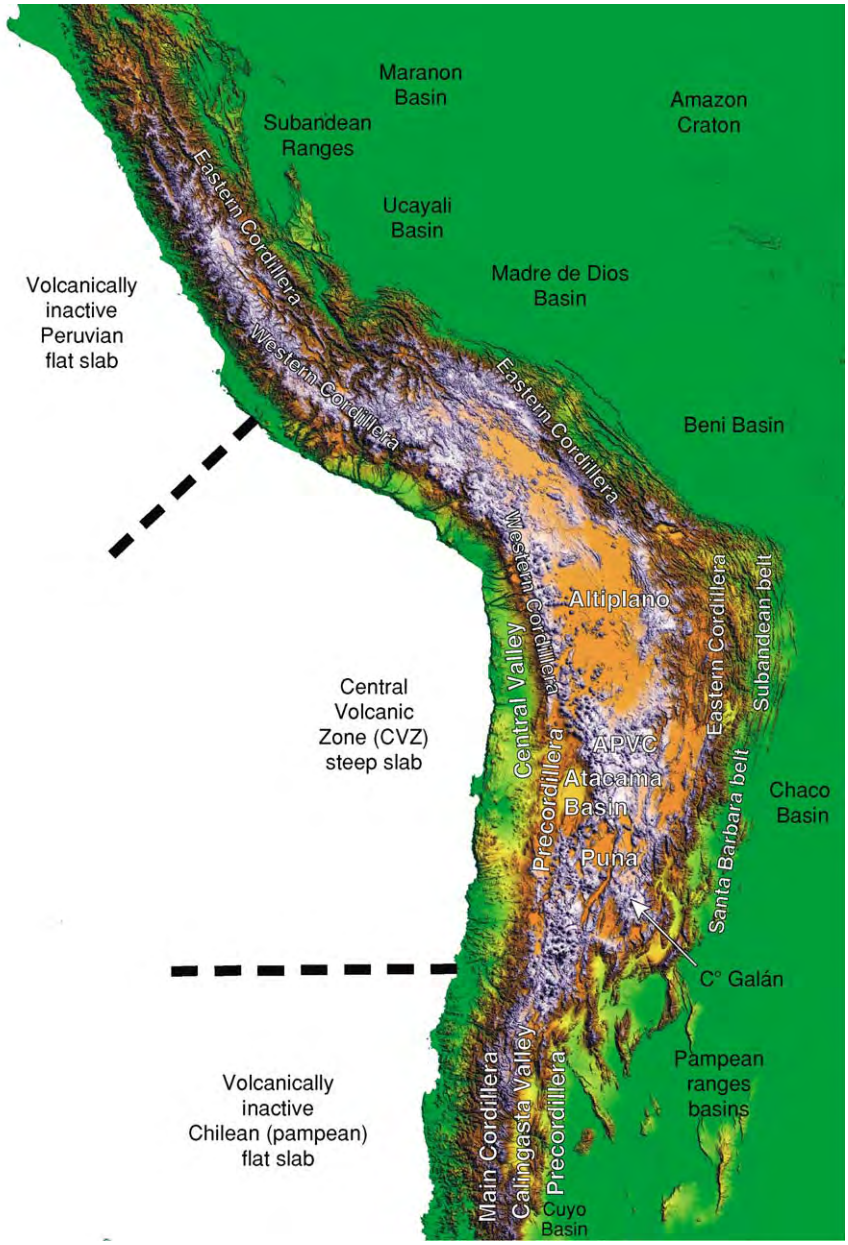


Figure 4 Continuation of digital elevation map (DEM) in **Figure 3** showing the principal features of the Central Andes of Peru, Bolivia, northern Chile and northern Argentina. The presence of white and blue colours attests to the higher elevation of the Central Andes. Individual faults are not indicated. Cerro Galán is the ~2 Ma ignimbrite caldera shown on the image on **Figure 5**. The highest average elevations and the greatest amounts of crustal shortening occur in this part of the Andes.

Colombia. They consist primarily of the north-east-trending Mérida Andes that merges northward into the Falcón fold–thrust belt and north-eastward into the Caribbean ranges. The Mérida Andes are largely composed of the Palaeozoic rocks of the Mérida Terrane that was accreted to South America in the Palaeozoic. The Caribbean ranges are a south-verging Cenozoic fold–thrust belt north of the oil-rich Venezuela Basin. The Mérida Andes are bounded to the north-west by the Maracaibo Block and to the south-east by the complex, multicycle Barinas–Apure foreland basin whose evolution initiated in the Late Palaeozoic. The Maracaibo Block is a triangular lithospheric scale wedge that is being squeezed northward between the right-lateral Boconó Fault to the east and the left-lateral Bucaramanga Fault to the west. The Oca-El Pilar right-lateral Fault lies to the north. The traces of the ~500 km long Boconó Fault system

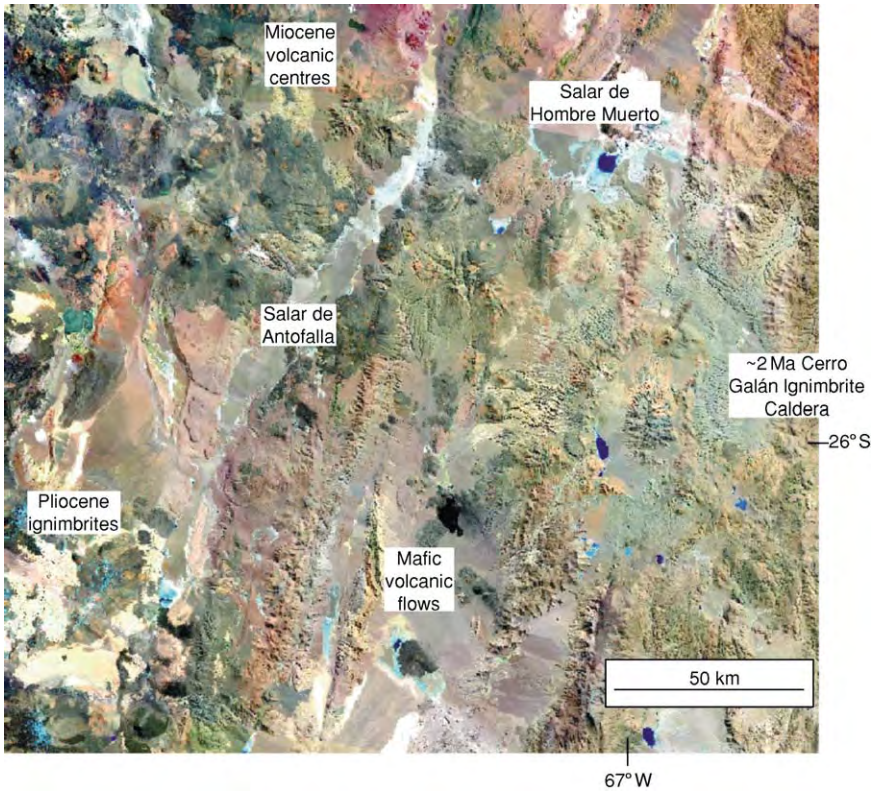


Figure 5 Thematic Mapper satellite image of the southern Puna plateau of the Central Andes showing the development of salars (closed evaporite basins), Neogene andesitic to dacitic volcanic centres, and ignimbrite flows and caldera complexes.

project north-east through the Mérida Andes into the Caribbean ranges. The Cretaceous to Cenozoic Maracaibo Basin is within the Maracaibo Block and is an important petroleum-producing basin.

Colombian Andes The Colombian Andes consist of the distinct Eastern, Central and Western Cordilleras. They are bounded on the east by the Borde Llanero thrust system that marks the boundary with the Putumayo Basin in the south and the Llanos Basin that connects with the Barinas–Apure foreland basin system to the north. The active volcanic arc over the Nazca Plate runs through the Central Cordillera. The Central and Eastern Cordilleras are largely composed of Precambrian continental crust covered by Palaeozoic to Tertiary sedimentary and volcanic sequences. The Central Cordillera was uplifted and subjected to pervasive Late Cretaceous to Early Tertiary deformation, whereas the Eastern Cordillera was most affected by Late Miocene to Recent compression. The Eastern and Central Cordilleras are separated by the upper and middle Magdalena Valley Neogene successor foreland basins. The asymmetrical shape of the basin results from post-Miocene west-verging thrusting on the east and the east-dipping Central Cordillera Block that partially underlies the basin. The Western Cordillera consists primarily of Cretaceous oceanic magmatic and deep-water sedimentary rocks. Along with the part of the Central Cordillera west of the Romeral Fault, the Western Cordillera is considered to be an amalgamation of terranes accreted to South America during the Cretaceous. The Central and Western Cordilleras are separated by the Cauca Valley whose western side is largely bounded by the Romeral Fault. This fault is considered to be a terrane suture that continues northward through the Lower Magdalena Valley Basin. On the west, the Western Cordillera is bounded by the San Juan Atrato fore-arc valley which is underlain by oceanic terranes. Part of this valley along with the Seranía de Baudó Block and the Panamá Arc were accreted in the Miocene (~12–13 Ma) along the Istmina deformed zone and the Uramita Fault.

Ecuadorian Andes The central Colombian Cordillera continues southward to form the Cordillera Real or Eastern Cordillera of Ecuador, which is composed

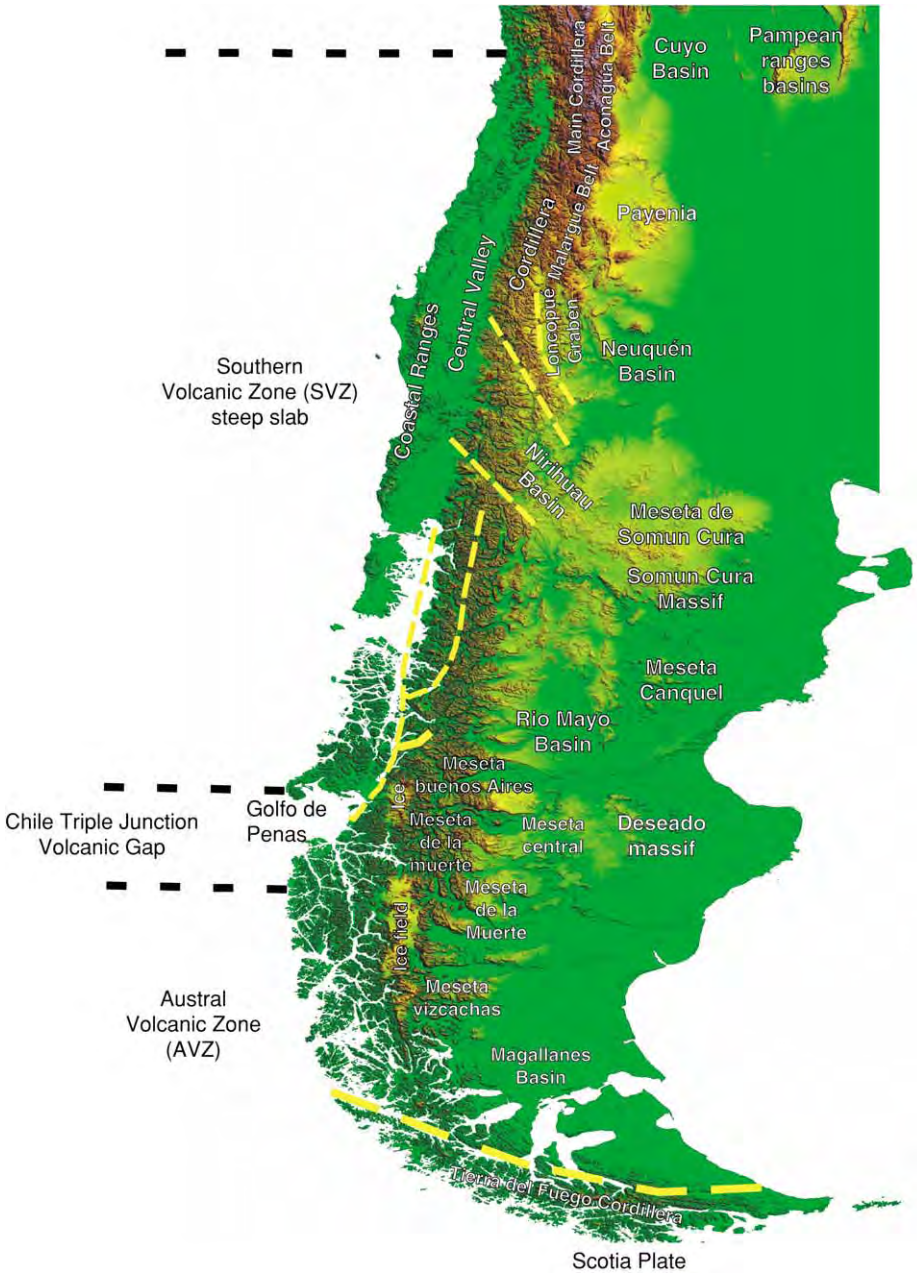


Figure 6 Continuation of digital elevation map (DEM) in **Figure 4** showing the principal features of the Southern Andes of Chile and Argentina. The evolution of the southern part of the region has been strongly influenced by successive collisions of segments of the Chile Rise with the trench as shown in more detail in **Figure 7**. Yellow lines are faults. Those along the Pacific coast belong to the Liquiñe Ofqui strike slip fault system.

of multiply deformed, variably metamorphosed Palaeozoic to Mesozoic rocks covered by Cenozoic volcanic and sedimentary units. The Putumayo and Oriente Basins to the east are part of a large composite Mesozoic to Cenozoic foreland basin that overlies older cratonic basement. The western 50–80 km of the basin, known as the Subandean Zone, has been uplifted to elevations of 500–1000 m by late Cenozoic thrusts. The Cordillera Real is separated from the Ecuadorian Western Cordillera by the Interandean Valley that is filled by Tertiary to Quaternary volcanic and volcaniclastic deposits. The valley is bounded to the west by the Pujili Fault and to the east by the Peltetec Fault, and is probably floored by Cordillera

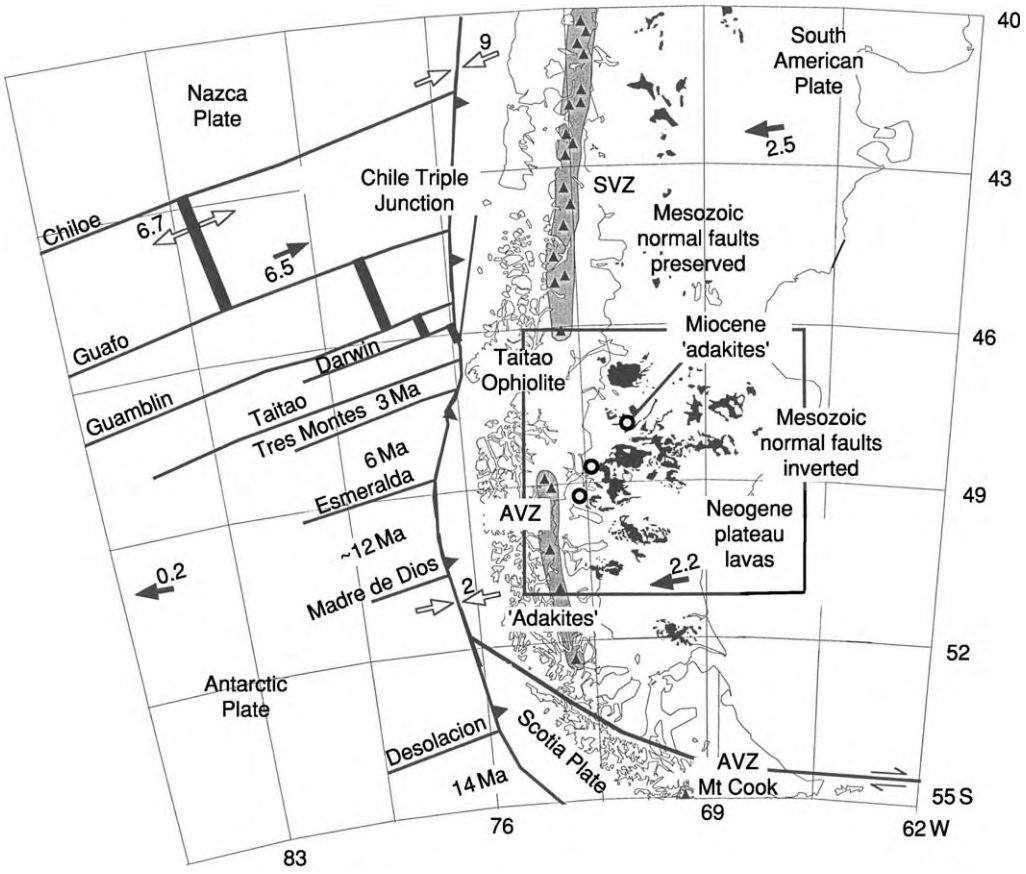


Figure 7 Map of southern Patagonia and surrounding ocean basins showing distribution of Southern Volcanic Zone (SVZ), Austral Volcanic Zone (AVZ adakites), and Miocene slab melt 'adakitic' volcanic centres, deformational styles, plate convergence velocities, and oceanic fracture zones and ridges. Active volcanism is absent east of where the Chile Ridge has collided in the last ~6 My. The boxed region is where features related to ridge collision are best developed. (Modified from Gorring ML, Kay SM, Zeitler PK, *et al.* (1997) Neogene Patagonian plateau lavas: continental magmas associated with ridge collision at the Chile Triple Junction. *Tectonics* 16: 1 17.)

Real-like basement. The axis of the Recent NVZ arc runs through the valley with volcanic centres occurring from the eastern part of the Western Cordillera into the Subandean Zone. The Western Cordillera consists largely of Cretaceous to Eocene oceanic crust, turbidites, and oceanic arc sequences covered by Late Eocene–Oligocene continental sediments and intruded by Eocene and younger magmatic rocks. The low-lying, Pacific coastal region is comprised of Cretaceous to Cenozoic basins underlain by oceanic basement.

Central Andes (~3° to 33° S) The best known part of the Andes is the Central Andes (Figure 4) which can be separated into the Peruvian flat slab segment, the Altiplano–Puna plateau segment that includes the Central Volcanic Zone Arc, and the Chilean flat slab segment. A surprising degree of bilateral symmetry exists in the shape of the subducting slab (Figure 2) and the topography of the land surface across the region (Figure 4).

Peruvian flat slab segment (~3° S to 15° S) The Peruvian flat slab segment largely coincides with the inactive part of the magmatic arc where active volcanism has been absent from 2.5° to 16° S for the last ~3–4 My. The Western and Eastern Cordilleras of Ecuador narrow into this region as the Interandean Valley virtually disappears and the Subandean belt broadens to a width of 120–250 km across into the composite Marañon–Ucayali foreland basin. A west to east profile shows a topographically low, narrow coastal region west of the Western Cordillera, a narrow Interandean depression, the Eastern Cordillera, the Subandean Zone, and the basins (e.g. Madre de Dios) of the eastern lowlands. A series of partly submerged Tertiary sediment-filled fore-arc basins (Trujillo, Lima and Pisco) along the narrow

continental shelf reflect the effects of crustal thinning due to fore-arc tectonic erosion. The basement of the southern coastal region consists of the high-grade metamorphic rocks of the Precambrian (~1 Ga) Arequipa block. To the east, the Western Cordillera is mainly composed of deformed metamorphic rocks and Mesozoic sediments that are cut and covered by Mesozoic to Tertiary magmatic rocks. The extensive Middle Cretaceous to Palaeogene Peruvian batholith that extends into the coastal region forms the western part. The highest peaks occur in the ~8 Ma Cordillera Blanca pluton whose western boundary, the Cordillera Blanca shear zone, has been interpreted as a low-angle detachment fault. The Eastern Cordillera is made up of pre-Mesozoic metamorphic rocks along with subordinate Palaeozoic and Tertiary intrusives. It hosts the Early Tertiary 'inner arc' and was the locus of intense Late Paleocene and Middle Eocene east-verging thrusting. The Subandean belt further east consists of Mesozoic and Tertiary sedimentary rocks and Palaeozoic basement cut by Late Miocene to Recent reverse faults that extend almost to the Brazilian border.

Central Volcanic Zone/Altiplano–Puna Plateau segment (~15° to 28° S) This segment includes the active Central Volcanic Zone arc and the widest part of the high Andean Cordillera that includes the Altiplano–Puna Plateau. The whole region is built on a variable age Precambrian and Palaeozoic basement. Uplift of the plateau is principally attributed to ductile thickening of the lower crust in response to compressional shortening, and subordinately to magmatic addition. Compressional shortening is reflected in the fold–thrust belts of the upper crust. Total amounts of shortening are widely debated with values in various transects ranging from 300 to 500 km in the north to <150 km in the south. The principal uplift is considered to be Eocene to Late Miocene in age in the north and Late Miocene in the south. A generalized west to east section across the central plateau crosses the coastal Cordillera, the fore-arc Central Valley, the Chilean Precordillera, the Western Cordillera, the Altiplano–Puna Plateau, the Eastern Cordillera, the Subandean Belt, and the Madre de Dios–Beni–Chaco composite foreland basin. These features all follow the bend in the trench (**Figures 1** and **2**) and coast (**Figure 4**) known as the Bolivian Orocline.

The Andes of northernmost Chile include the fore-arc region from the coast to the CVZ volcanic front. The Neogene sediment-filled Central Valley is largely flanked to the west by Jurassic to Middle Cretaceous arc sequences cut by the Cretaceous Atacama strike-slip Fault system, and to the east by the Cretaceous to Eocene volcanic and sedimentary rocks in the Chilean Precordillera that are cut by the Late Palaeozoic to Early Mesozoic Palaeogene Domeyko strike-slip Fault system. The Chilean Precordillera hosts the giant Chuquicamata and Escondita Cu deposits. The coastal Cordillera and Atacama Fault system run offshore at the bend in the coast in northernmost Chile where the western margin appears to have been removed by fore-arc subduction erosion. The Western Cordillera marks the western limit of the high Andes and contains the CVZ arc. Both diverge to the east around the pre-Andean Atacama Basin near 23° S to 24° S. The highest peaks reach ~5000–6550 m in the north and over 6800 m in the south. Underlying crustal thicknesses range from 70 km in the north to ~55 km in the south.

Flanking the Western Cordillera to the east is the most prominent feature of the segment: the ~700 km long and ~200 km wide Altiplano-Puna Plateau with an average elevation of 3700 m. This internally drained plateau hosts extensive Late Cenozoic sedimentary evaporate-filled basins (salars) and chains of Neogene volcanic centres (**Figure 5**) that extend to its eastern edge. The highest regions (> 6300 m) are Late Miocene to Pliocene andesitic to dacitic stratovolcanic and giant dacitic ignimbrite complexes. Only sequences older than ~10 My are significantly deformed. The Altiplano section in Bolivia north of ~22° S is mostly comprised of a relatively flat sediment-filled basin covering a largely Palaeozoic sedimentary sequence over a Brazilian shield-like Precambrian basement. Tertiary volcanic rocks are concentrated in the Western and Eastern Cordilleras where they can host important tin and silver (Potosi) deposits. One of the most famous is the silver deposit at Potosi in the Eastern Cordillera. Crustal thicknesses under the Altiplano range from ~66 to ~78 km. The Puna section in Argentina to the south is distinctive in being broken into ranges with high peaks and basins and east–west chains of Neogene volcanic rocks (**Figure 5**). This region also differs in being underlain by a basement that includes Palaeozoic mafic to silicic magmatic sequences formed in complex arc and back-arc settings. The greatest average elevation is in the southern Puna where crustal thicknesses appear to be <58 km.

The distribution of large <10–12 Ma ignimbrite calderas is uneven across the plateau. Late Miocene centres occur in the Altiplano–Puna Volcanic Complex (APVC) in the central plateau and along the eastern margin. Early Pliocene centres occur in the western APVC and near the southern Western Cordillera. The youngest major eruption, the ~2 Ma Cerro

Galán ignimbrite (see Figure 5), occurred in the southern Puna where latest Miocene to Recent monogenetic calc-alkaline and intraplate mafic flows not seen elsewhere on the plateau are also found. Small young northern Puna and Altiplano mafic flows have shoshonitic chemistry.

The plateau is bounded to the east by the Eastern Cordillera and adjoining Subandean Ranges. East of the Altiplano, the Eastern Cordillera of Bolivia consists mainly of deformed Palaeozoic sedimentary rocks overlain by Eocene to Miocene magmatic rocks that have been subjected to Middle Eocene to Late Miocene contractional deformation. Peaks can reach up to 6000 m. To the east, elevations drop off abruptly into the Subandean Belt where compressional deformation has been concentrated for the last ~10 My since the active thrust front moved eastward. The Subandean Belt of Bolivia and northernmost Argentina is largely comprised of deformed Palaeozoic sedimentary rocks cut by Late Miocene to Recent thin-skinned thrusts. The picture changes along the Puna where the principal Tertiary deformation of the Eastern Cordillera is Miocene in age and the Subandean Belt is characterized by the 400 km long Santa Barbara Fault system in which Neogene compressional deformation is characterized by inversion of normal faults related to the complex Cretaceous to Palaeogene Salta Group Rift system. A sharp contrast in structural style with the thin-skinned Subandean belt to the north and the Pampean basement uplifts that characterize the Chilean flat slab to the south corresponds with the boundaries of the Salta Group Rift basins.

Chilean flat slab segment (28° to 33° S) The Chilean (or Pampean) flat slab corresponds with the current gap in the active volcanic arc between 28° and 33° S over the shallow Benioff zone (Figure 2). In accord with the geometry of the subducting plate, the northern boundary is geologically transitional, whereas the southern boundary is relatively abrupt. A west to east traverse crosses the Coastal Cordillera, the high Andes consisting of the Main and Frontal Cordilleras, the Uspallata–Calingasta Valley, the Argentine Precordillera, and the Pampean Ranges that are flanked by a series of Neogene foreland basins. The fore-arc differs from the adjacent segments in that the Chilean Central Valley is missing. The Main Cordillera hosts Miocene volcanic arc rocks that along with Mesozoic and Palaeogene sedimentary and magmatic rocks cap major Late Palaeozoic/Triassic magmatic sequences similar to those that comprise the Frontal Cordillera. The 6962 m Aconcagua peak in the Main Cordillera near 32° S, the highest in the world outside of the Himalayas, is comprised of Miocene (~15–9 Ma) arc volcanic rocks thrust over Mesozoic and Tertiary sediments. The Main Cordillera hosts rich Miocene Au (El Indio–Veladero system) and Cu (e.g. Rio Blanco) deposits. The Uspallata–Calingasta Valley to the east is a piggy-back foreland basin filled with Miocene sedimentary and volcanic rocks. It coincides with a major Palaeozoic terrane suture and fills the volcanic gap between the presently active CVZ and SVZ arc segments. The Argentine Precordillera primarily consists of deformed Palaeozoic sedimentary sequences associated with Miocene sedimentary and minor dacitic volcanic rocks overlying ~1 Ga metamorphic rocks. These sequences are deformed by Middle to Late Miocene east-verging thrusts in the west and younger west-verging thrusts in the east. The west-verging thrusts are attributed to reactivation of basement structures. Further east, the Pampean ranges are thick-skinned blocks of Precambrian to early Palaeozoic crystalline rocks uplifted by Late Miocene to Pliocene high-angle reverse faults. Elevations in the north (Sierra de Aconquija) reach over 6000 m. The easternmost range, which is ~700 km from the trench, hosts small ~7–5 Ma volcanic rocks whose geochemical signatures indicate slab-derived components. The westernmost range (Pie de Palo) has been uplifted in the last 3 My. Total Neogene shortening estimates across the Chilean flat slab segment near 30° S are near 200 km with ~120 km of that occurring on thrust faults in the Precordillera.

Southern Andes (33° to 56° S) The Southern Andes begin south of the Chilean flat slab and the subducting Juan Fernandez Ridge on the Nazca Plate (Figures 1, 2 and 6). The northern part includes the Southern Volcanic Zone that overlies a moderately steep subduction zone in which the Nazca Plate decreases in age until reaching the Chile Triple Junction near 47° S. The southern segment is south of the Chile Triple Junction. *See* **Argentina** for more information on these regions.

Southern Volcanic Zone Segment (33° to 46° S) The Southern Volcanic Zone comprises the region of active volcanism from 33° to 46° S. From west to east, a section crosses the Late Palaeozoic metamorphic rocks of the Coastal Cordillera, the largely Tertiary sediment-filled Chilean Central Valley, Miocene arc rocks and uplifted and deformed Mesozoic sequences of the Main Cordillera upon which the active SVZ volcanic arc is built, and a southwardly narrowing retro-arc fold–thrust belt called the Aconcagua Belt in the northernmost part (~32° to 34° S), and the Malargue Belt to the south (~34° to 37° S).

The height of the Main Cordillera decreases from ~6600 to < 2000 m southward as the fore-arc Central Valley widens and back-arc shortening related to both thin- and thick-skinned thrusting decreases from ~100 to < 35 km at 37° S. Between 38° S and 40° S, the arc front has migrated westward since the Pliocene in conjunction with mild extension in the region of the Loncopué Graben. From ~38° S southward, the modern volcanic arc overlaps with the > 900 km long Liquiñe–Ofqui Fault system (Figure 6), which has had dextral displacement since at least the Early Miocene. The Mesozoic and Tertiary arc fronts are in the modern fore-arc north of 36° S and near the modern arc to the south. The giant Late Miocene El Teniente Cu deposit occurs in the westernmost Cordillera near 34° S.

The foreland has important basins with both rift and foreland histories that surround the pre-Mesozoic Somuncura and Deseado massifs. They include the Jurassic through Tertiary Neuquén Basin north of ~38° S, the Early Cretaceous Rio Mayo Embayment near 45° S and the Palaeogene to Early Neogene Nirihuau Basin near 41° S. A series of smaller segmented basins occur in the fore-arc.

Andean magmatism is well developed in the foreland in two regions. The first is in the Payenia area between ~35° S and 38° S where extensive volcanic sequences cover the northern half of the Neuquén Basin. These volcanic sequences, which can occur more than 500 km east of the modern trench, consist of Miocene andesitic/dacitic volcanic centres with arc geochemical signatures, and Early Miocene and Late Pliocene to Recent mafic alkaline flows. The second region is between ~41° S and 44° S where Palaeogene and Miocene andesitic to rhyolitic sequences, Eocene alkaline basalts, and the extensive Late Oligocene to Miocene mafic plateau flows associated with the Mesetas de Somun Cura and Canquel occur.

Chile Triple Junction and Austral Volcanic Zone Segment (47° to 56° S) Important changes occur south of the Chile Triple Junction near ~47° S where the Chile Ridge is colliding with the trench (Figure 7). The modern volcanic arc disappears between ~47° S and 49° S and resumes as the Austral Volcanic Zone in response to subduction of the Antarctic Plate. The high Andes at this latitude are dominantly composed of post-Triassic magmatic rocks with the principal part of the Jurassic to Miocene Patagonian Batholith forming the backbone of the Cordillera. Late Palaeozoic/Early Mesozoic fore-arc accretionary complexes occur in the fore-arc. The Patagonian Cordillera reaches a maximum elevation of ~4000 m near ~47° S east of where the Chile Ridge is currently colliding and decreases in elevation to the south. The 30 000 km^2 Patagonian ice-field, the world's third largest, occurs at the higher elevations.

The importance of back-arc crustal shortening abruptly increases at ~47° S as does the amount of Tertiary foreland sedimentary deposits. To the south, Mesozoic normal faults have largely been inverted by Tertiary compression whereas similar age normal faults are preserved to the north. The Patagonian region east of where the Chile Ridge has collided is notable for Neogene fore-arc volcanism (Taitao Ophiolite), abundant Late Neogene mafic plateau flows east of where ridge collision occurred at ~12 and 6 Ma, and widespread Pleistocene to Recent mafic flows. Extensive Eocene mafic retro-arc plateau lavas also occur in this area.

The southernmost part of Patagonia includes the Jurassic to Tertiary Magallanes (austral) Basin whose axis coincides with the Early Cretaceous Tortuga and Sarmiento ophiolite complexes. The Andes of Tierra de Fuego contain the Cretaceous metamorphic complexe of the Cordillera Darwin. They include an east to west trending fold–thrust belt cut by a major north-west to south-east trending left-lateral fault system that delimits the northern boundary of the Scotia Plate. The SVZ Mount Cook volcanic centre also occurs in this region.

Jurassic to Recent Evolution of the Andean Chain

The evolution of the Andes that began with the Mesozoic breakup of Pangaea can be divided into a Late Triassic to Early Cretaceous stage dominated by rifting processes and extensional arc systems, a Late Cretaceous to Early Oligocene stage in which the Andes evolved from an extensional dominated to a compressional regime, and a latest Oligocene to Recent stage in which most of the main Andean range was uplifted.

Stage 1: Rifting and Extensional Arc Systems

The Late Triassic to Early Cretaceous stage began with rifting associated with the breakup of Pangaea. The geometries of the rifts that began all across western South America in the Triassic and Jurassic reflect extensional directions, basement fabrics and old terrane boundaries. The north-east to south-west rifting *en echelon* pattern in the Northern Andes matches the counterclockwise rotation associated with rifting from the conjugate North America Yucatan Block. North-west trending rifts in the rest of the Northern, Central and Southern Andes are aligned with Gondwana Palaeozoic sutures. The north-west trending dextral shear pattern of rifts in Patagonia could have

been influenced by clockwise rotation of the Atlantic Peninsula initiating spreading of the Weddell Sea between 175 and 155 Ma.

This rifting was well underway when a subduction regime was diachronically established along the western margin of South America in the Early to Middle Jurassic. The Jurassic to Early Cretaceous history of this system is dominated by a complex series of fore-arc, intra-arc, and retro-arc basins in which extension was associated with subduction zone rollback and oblique convergence. Middle Jurassic to Early Cretaceous thermal subsidence in the back-arc began at different times in different places. The expansion of marine sedimentation in the Early Cretaceous reflects the thermal subsidence that followed this rifting and a long-term rise in global sea-level.

In the Southern Andes, extensional faulting in the Neuguen Basin was interrupted by inversion in the Jurassic (Araucanian Event). This inversion was followed by rifting along the east coast of northern Argentina, Uruguay and Africa at about the same time that the Late Jurassic magmatic arc in central Chile migrated ~30–40 km to the east and the Atacama intra-arc strike-slip fault system became active. Extension reached a height in the Southern Andes with the Late Jurassic formation of the oceanic Rocas Verdes Basin in southernmost Patagonia that incorporates the Sarmiento and Tortuga ophiolites. In the Northern Andes, Jurassic extension precipitated thermal sag that led to extensive shallow marine embayment by the end of the middle of the Early Cretaceous (Neocomian).

The Early Cretaceous (130–110 Ma) marks the opening of the South Atlantic Ocean off the shore of Brazil and rapid westward drift of South America relative to the underlying mantle. Active extension was most pronounced in Bolivia, northern Argentina and Chile as the central South Atlantic Ocean began opening at ~130 Ma at the time of the eruption of the Parana flood basalts in Brazil. Intracratonic extension occurred in the southern Altiplano and in the Salta Rift system and Pampean ranges where it was associated with basaltic volcanism. To the west, the Atacama intra-arc strike-slip fault system became transtensional at ~132–125 Ma and the Aptian period (~121–112 Ma) brought increased negative trench rollback velocity causing the intra-arc and back-arc extension that produced the marine sediment-filled aborted marginal basin in Peru associated with Casma volcanism (after ~112 Ma), the Sierra de Fraga low-angle detachment faults near 27° S, and the aborted marginal basin linked with the eruption of the extensive ~119–110 Ma Veta Negra and related volcanic groups in central and south central Chile (29 to 33° S).

In the Southern Andes, intracratonic rifting had essentially ceased by the Aptian as shown by the terminal stages of the Rocas Verdes Basin in southern Patagonia, a peak in the emplacement of the Patagonian batholith, and the end of marine sedimentation in the Neuquén Basin. In the Northern Andes, the presence of blueschist and high-pressure metamorphic rocks (~130 Ma) along the Romeral and Peltetec faults can be associated with the obduction of small island-arc systems to the Colombian Central Cordillera and western side of the Ecuadorian Cordillera Real. The Cretaceous (112–90 Ma) drowning of the Barinas–Apure Basin can be attributed to thermal subsidence.

Stage 2: Basin Inversion and Formation of the Early Andes

The main compressional stage that built the Andes began after 10 Ma as active spreading in the South Atlantic accelerated the separation of South America from Africa, and South America began to actively override the trench. A series of compressional events of variable intensity occurred all along the margin as foreland basins responded to flexural loading. Compressional events took place in the Late Cretaceous near ~105–95 Ma (Mochica Phase), near 85–75 Ma (Peruvian Phase), at the beginning of the Paleocene near 65–50 Ma, and during the Eocene (Incaic Phase). These are all times of changes in plate convergence rates and directions along the margin.

Late Cretaceous compression in Peru and Bolivia is marked by tectonic inversion in the Mochica Phase at ~105 Ma. In Chile, the final phase of pluton emplacement in the Coastal Cordillera took place at 106 Ma, after which the intra-arc Atacama Fault system was largely abandoned. In Patagonia, the Late Cretaceous (98–85 Ma) marks the closure of the Rocas Verde Basin a peak in production, of the Patagonian Batholith, and the formation of the thrust-loaded Magallanes foreland basin. An inversion at ~99 Ma marks the change from an extensional to a foreland setting for the Neuquén Basin.

The Peruvian phase corresponds with the final emplacement of the Late Cretaceous Coastal Batholith in Peru and the end of marine sedimentation in the large back-arc marine basin to the east in Peru and Bolivia. In Chile, a new arc and fault system was established at ~86 Ma in what is now the Central Valley, some 50 km east of the old Atacama system. The proto-Cordillera de Domeyko was uplifted at this time. In the Northern Andes, Late Cretaceous closure of an ocean basin led to accretion of the Cretaceous oceanic Piñon–Dagua Terrane to the Colombian and Ecuadorian Western Cordilleras between ~80 and 60 Ma (Calima Orogeny). The collision is generally

associated with an east-dipping subduction zone along the continental margin. The progressive southward drowning of the basin to the east that had been occurring through the Cretaceous ended in association with the deformation and uplift of the Central and Eastern Cordillera. Partial obduction of the Caribbean Plate at ~85–80 Ma caused the Latest Cretaceous to Early Eocene Caribbean orogeny and the southward migration of the Venezuelan fore-deep.

By the Palaeogene, marine deposition had ended everywhere but the northernmost Andes and the large basins in Patagonia; uplift had built a barrier to Pacific marine ingressions; and Andean detritus was being spread into foreland basins. In central Chile, the Paleocene was characterized by explosive silicic calderas. Progressive Palaeogene to Eocene southward collision of the Aluk–Farallon Ridge with the trench leading to a gap (slab window) between the subducting plates has been suggested to explain a pattern of arc volcanism at ~56–50 Ma near ~37° S, bimodal arc volcanism near 40° S, paucity of arc magmatism south of ~43° S and widespread basaltic volcanism further south.

Middle to Late Eocene deformation (47–32 Ma, Incaic Orogeny) occurred during a phase of fast oblique convergence. Compression produced the fold–thrust belts of the Western Cordillera of Peru, deformation associated with contractional and strike-slip faulting along the Cordillera de Domeyko in northern Chile, and the west-directed thrusting that resulted in uplift of the Bolivian Eastern Cordillera. This deformation, along with back-arc magmatism and deformation in the Zongo–San–Gabán Zone on the Altiplano/Eastern Cordillera, has been related to southward migration of a shallow subduction zone as well as to a more normal convergence direction here than to the north or south. These events were followed by a relative Oligocene magmatic lull along most of the Andean margin during a period of slow plate convergence.

Stage 3: Formation of the Modern Andes (~27–0 Ma)

The Neogene history of the Andes begins with the breakup of the Farallon Plate into the Nazca and Cocos Plates at ~28–26 Ma. This event is linked to the change in convergence direction and increase in relative convergence velocity that led to the major uplift of the Andes that culminated in the Late Miocene and Pliocene. Although the continuity of Miocene deformational phases in the broad Quechua Orogeny associated with this uplift has been widely debated, there is some agreement that general margin-wide changes occurred at ~19–16 Ma and 8–4 Ma. The formation of major Neogene Cu and Au deposits during this period correlates with periods of shallowing and steepening subduction zones, migrating arc fronts, and crustal thickening.

Events in the Central Andes can be related with shallowing and steepening of the subducting Nazca Plate. Shallowing beneath the Chilean flat slab can be correlated with Miocene uplift and cessation of arc volcanism in the Main Cordillera, Late Miocene thrusting in the Argentine Precordillera, Late Miocene to Pliocene uplift of the Sierras Pampeanas, and Pliocene termination of volcanism across the region. Deformational and magmatic events associated with shallowing of the subducting slab under the Peruvian flat slab region are superimposed on older Palaeogene features. Steepening of a formerly shallow subduction zone beneath the Central Altiplano–Puna Segment can be correlated with a Late Oligocene to Early Miocene magmatic gap associated with widespread compressional deformation that was followed by large volume Late Miocene to Pliocene ignimbrite eruptions. The large eruptions can be explained by hydrated mantle melts from above the slab intruding the crust and causing massive melting. The progressive westward narrowing of the magmatic arc fits with steepening of the subducting slab. Ductile deformation of the hot crust can explain Miocene plateau uplift as the thrust front migrated eastward into the Subandean belt. Magmatic and geophysical data support delamination of a thickened crustal root beneath the Altiplano–northern Puna during this time. The southern Puna and Altiplano regions were in transitional positions between the steepening and shallowing segments of the subducting slab. The presence of Late Pliocene ignimbrites and mafic magmas, a relatively thin crust and mantle lithosphere, a high average elevation, and a change in Late Miocene/Pliocene faulting patterns in the southern Puna have been related to Pliocene delamination of a dense, thickened crustal root.

The Neogene magmatic and deformational history of the Southern Volcanic Zone region can be tied to eastward migration of the Neogene arc front at ~20–16 Ma north of 36° S, and again at ~8–4 Ma north of ~34.5° S. The back-arc andesitic/dacitic centres and large retro-arc volcanic fields in the Puyenia region may be due to melting of hydrated mantle that formed as the result of transient Miocene shallowing of the subducting Nazca Plate. The Neogene evolution of the Andes south of 47° S can be tied to the northward propagation of the Chile Triple Point as segments of the Chile spreading centre collided progressively northward with the Chile Ridge. The effects of ridge collision include uplift and compressional deformation in the fore-arc and arc region before collision, emplacement of fore-arc lavas, eruption of

Late Miocene to Recent plateau lavas above the gap (slab window) that opened between the trailing edge of the Nazca and the leading edge of the Antarctic Plate, and eruption of arc magmas generated by melting of the hot young subducting plate. The presence of Paleocene to Recent Patagonian mafic back-arc magmatism at different times and different places (Payenia, Somun Cura region, and east of where the Chile Ridge has collided) attests to the readiness of the Patagonian mantle to melt given provocation.

Most of the present configuration of the Northern Andes is also due to Neogene tectonics. The Miocene opening of the Gulf of Guayaquil in Ecuador due to strike-slip motion on the Pujili–Cauca and Peltetec–Romeral faults led to the generation of pull-apart coastal and intermontane basins. The Late Miocene collision of the Carnegie Ridge against the coast of Ecuador resulted in compressional inversion of extensional structures and subsequent strike-slip displacement of the Northern Andes. The accretion of the Choco Terrane, which includes the Panama–Baudó Block, starting at ~13–12 Ma, caused deformation and uplift of the Colombian Eastern Cordillera and the Mérida Andes. The most intense phases of deformation in the Venezuelan Andes and the Maracaibo Block are linked to Caribbean oblique convergence and underthrusting in the last 10 My. The present convergence vectors of the Nazca and Caribbean plates and ongoing collision of the Panama–Baudó Arc have generated a state of intracontinental compressional stress partitioned into reverse and strike-slip faulting.

Concluding Remarks

This short synthesis presents the state of knowledge of the nature of the Andean mountain range. What is presented here should be viewed only as one stage in the understanding of the evolving picture of this magnificent and complex range which is a natural laboratory for investigating the origin and evolution of continental mountain belts and the formation and destruction of continental crust.

See Also

Antarctic. **Argentina**. **Brazil**. **Tectonics:** Mountain Building and Orogeny. **Volcanoes**.

Further Reading

Allmendinger RW, Jordan TE, Kay SM, and Isacks BL (1997) The evolution of the Altiplano Puna Plateau of the Central Andes. *Annual Reviews of Earth and Planet ary Sciences* 25: 139 174.

ANCORP Working Group (2003) Seismic imaging of a convergent continental margin and plateau in the Central Andes (Andean Continental Research Project 1996 ANCORP'96). *Journal of Geophysical Research* 108: 2328 doi:10.1029/2002JB001771.

Cahill TA and Isacks BL (1992) Seismicity and shape of the subducted Nazca Plate. *Journal of Geophysical Research* 97: 17 503 17 529.

Cordani UJ, Milani EJ, Thomaz F, Ilho A, and Campos DA (eds.) (2000) *Tectonic Evolution of South America.* 31st International Geological Congress, Río de Janeiro. Avail able through the Geological Society of America.

Gansser A (1973) Facts and theories on the Andes. *Journal of the Geological Society, London,* 129: 93 131.

Gorring ML, Kay SM, Zeitler PK, *et al.* (1997) Neogene Patagonian plateau lavas: continental magmas associated with ridge collision at the Chile Triple Junction. *Tectonics* 16: 1 17.

Gutscher MA (2002) Andean subduction styles and their effect on thermal structure and interplate coupling. *Jour nal of South American Earth Sciences* 15: 3 10.

Isacks BL (1988) Uplift of the central Andean plateau and bending of the Bolivian orocline. *Journal of Geophysical Research* 93: 3211 3231.

Jordan TE, Isacks BL, Allmendinger RW, *et al.* (1983) Andean tectonics related to geometry of subducted Nazca plate. *Bulletin of the Geological Society of America* 94: 341 361.

Kay SM, Godoy E, and Kurtz AJ (2004) Episodic arc mi gration, crustal thickening, subduction erosion, and Mio cene to Recent magmatism along the Andean Southern Volcanic Zone Margin. *Bulletin of the Geological Society of America* (in press).

Kley J, Monaldi CR, and Salfity JA (1999) Along strike segmentation of the Andean foreland: causes and conse quences. *Tectonophysics* 301: 75 96.

Skinner BJ (ed.) (1999) *Geology and Ore Deposits of the Central Andes.* Society of Economic Geology Special Publication No. 7.

Taboada A, Rivera LA, Fuenzalida A, *et al.* (2000) Geo dynamics of the northern Andes: subduction and intracontinental deformation (Colombia). *Tectonics* 19: 787 813.

Tankard AJ, Suarez Soruco R, and Welsink HJ (eds.) (1995) *Petroleum Basins of South America.* American Associ ation of Petroleum Geologists Memoir 62.

ANTARCTIC

B C Storey, University of Canterbury, Christchurch, New Zealand

Introduction

Antarctica has not always been the cold, isolated, polar continent that it is today, with all but a minor fraction covered by ice. Throughout geological time, Antarctica has amalgamated and combined with other continental fragments and has rifted to form new ocean seaways. For a period of nearly 300 million years (My), during much of the Phanerozoic, from 450 to 180 million years ago (Ma), Antarctica formed the keystone of the Gondwana supercontinent (Figure 1A). Prior to Gondwana, Antarctica may have formed part of a supercontinent called Rodinia (Figure 1B), whereby East Antarctica was joined to the western side of North America; this possibility has been developed within the framework of the Southwest US–East Antarctic (SWEAT) connection hypothesis. Today, Antarctica is isolated in a south polar position and, with the exception of one small segment, is completely surrounded by spreading ridges. Consequently, the continent has low seismicity and few active volcanoes.

The continent (Figure 2) is divided physiographically by the Transantarctic Mountains, one of Earth's major mountain ranges. These mountains, which extend for some 3500 km across the continent between the Ross and Weddell seas, are typically 100 to 200 km wide and reach elevations locally in excess of 4500 m. This spectacular topographic feature defines a fundamental lithospheric boundary that has profound crustal anisotrophy due to repeated cycles of tectonism, and marks a boundary between two regions, East and West Antarctica, with quite different geological histories. The main continental features are shown in Figure 3A and the main geological events are summarized in Figure 3B. East Antarctica is the old crystalline core of the continent (East Antarctic Shield) and is largely covered by the 4-km-thick East Antarctic ice-sheet. With the exception of the central Gamburtsev Mountains, a high sub glacial mountain range of unknown origin, East Antarctica has a subdued topography. West Antarctica, on the other hand, has a horst and graben topography, with the spectacular Ellsworth Mountains hosting the highest mountain in Antarctica, Mt Vinson (4820 m). A high Andean mountain chain occupies the spine of the Antarctic Peninsula.

The East Antarctic Shield

The East Antarctic Shield comprises a Precambrian to Ordovician basement of igneous and sedimentary rocks deformed and metamorphosed to varying degrees and intruded by syn- to posttectonic granites. This basement is locally overlain by undeformed Devonian to Jurassic sediments (Beacon Supergroup) and intruded by Jurassic tholeiitic plutonic and volcanic rocks. Although much of the East Antarctic Shield is covered by the thick East Antarctic ice-sheet, the application of traditional and modern U/Pb zircon and other dating techniques has shown that the shield has a three-stage tectonic history, the details of which are only just becoming evident. This history involves the following events:

1. The stabilization of various Archaean to Palaeoproterozoic cratons (dating from 3.0 to 1.6 Ga). These areas of ancient crust can be divided into an extensive central craton, the Mawson Continent, inferred to occupy much of the continental interior of the East Antarctic shield, and various marginal cratons exposed along the coast. The marginal Archaean cratons are correlated with similar cratons that rifted from Antarctica during the Mesozoic breakup of Gondwana (for example, the Kaapvaal–Zimbabwe Craton of southern Africa and the Dharwar Craton of southern India). One of these, the Napier Complex, contains the oldest rock, dated at 3930 ± 10 My, currently known from Antarctica.
2. The development of three high-grade Late Mesoproterozoic to Early Neoproterozoic mobile belts. These were previously assumed to form one single Grenville-age orogen around the coastline of East Antarctica, but now appear to represent distinct crustal fragments juxtaposed in the Cambrian. One of these belts, according to the SWEAT hypothesis, constituted a piercing point, linking North America and East Antarctica in the Rodinia Supercontinent between 1100 and 750 Ma.
3. Two Late Neoproterozoic to Cambrian 'Pan-African' mobile belts that rework, truncate, and offset the preceding mobile belts, indicating that the East Antarctic segment of Gondwana underwent significant reorganization during the assembly of Gondwana at the Precambrian–Cambrian boundary. One of these belts is a continuation of the East African Orogen and developed during closure of the Mozambique Ocean and the ultimate amalgamation of Gondwana.

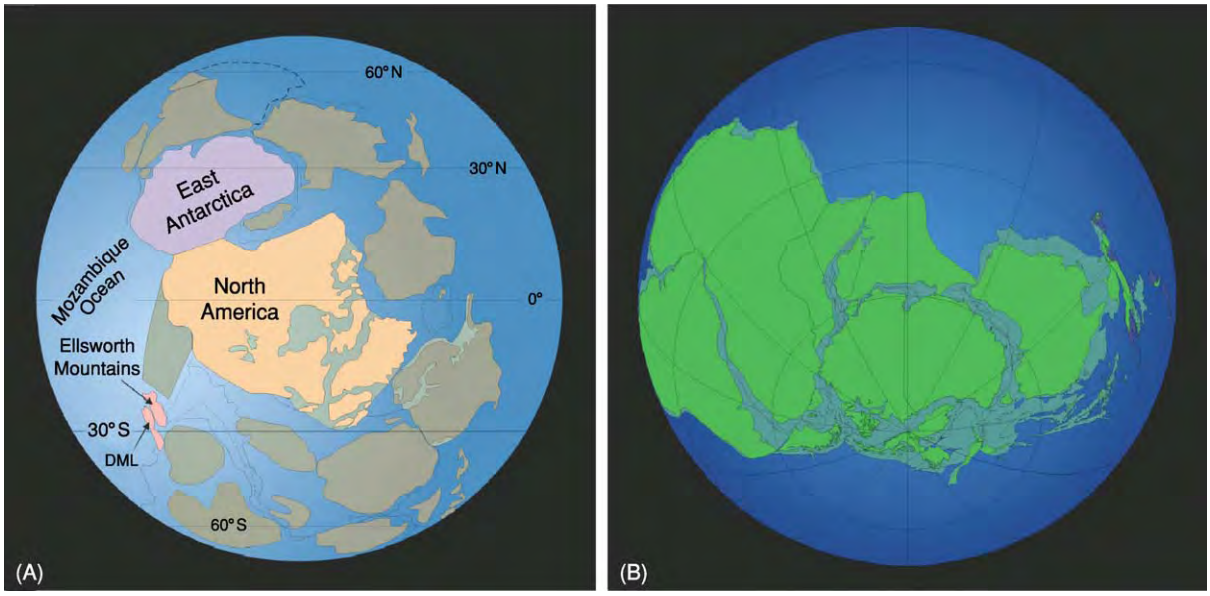


Figure 1 (A) A proposed Rodinia Supercontinent linking East Antarctica and North America, 1000 Ma, according to the 'SWEAT' hypothesis. (B) Antarctica within Gondwana 260 Ma (courtesy of R Livermore, British Antarctic Survey).

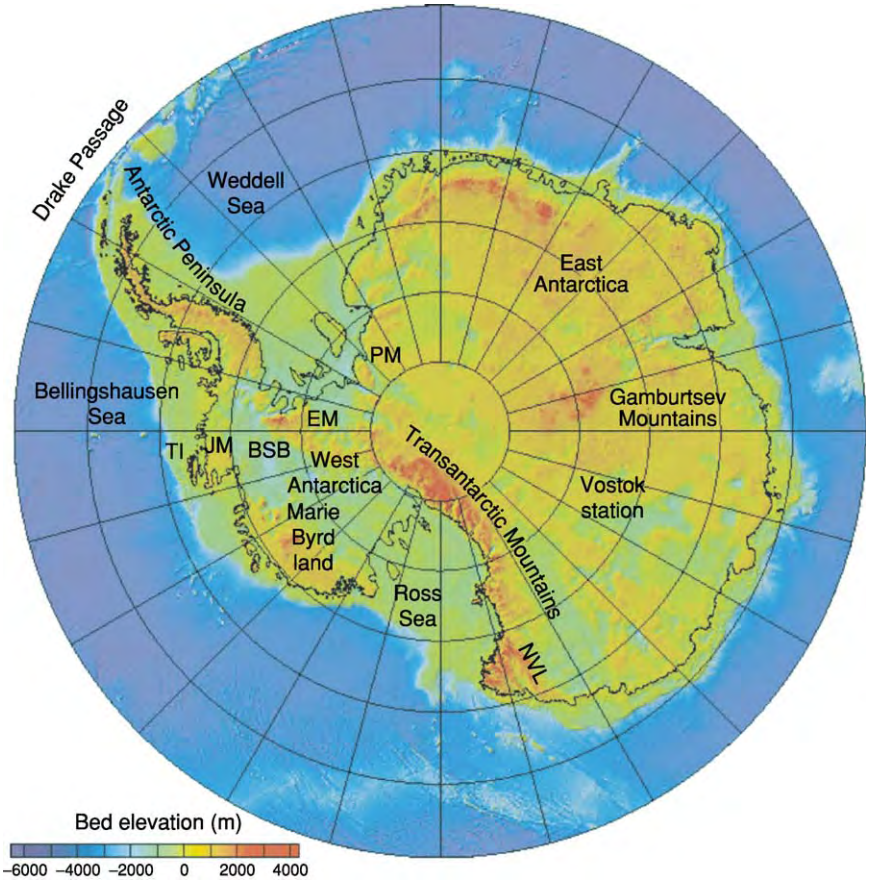


Figure 2 Subglacial topographic map of Antarctica. BSB, Byrd Subglacial Basin; EM, Ellsworth Mountains; JM, Jones Mountains; PM, Pensacola Mountains; NVL, North Victoria Land; TI, Thurston Island.

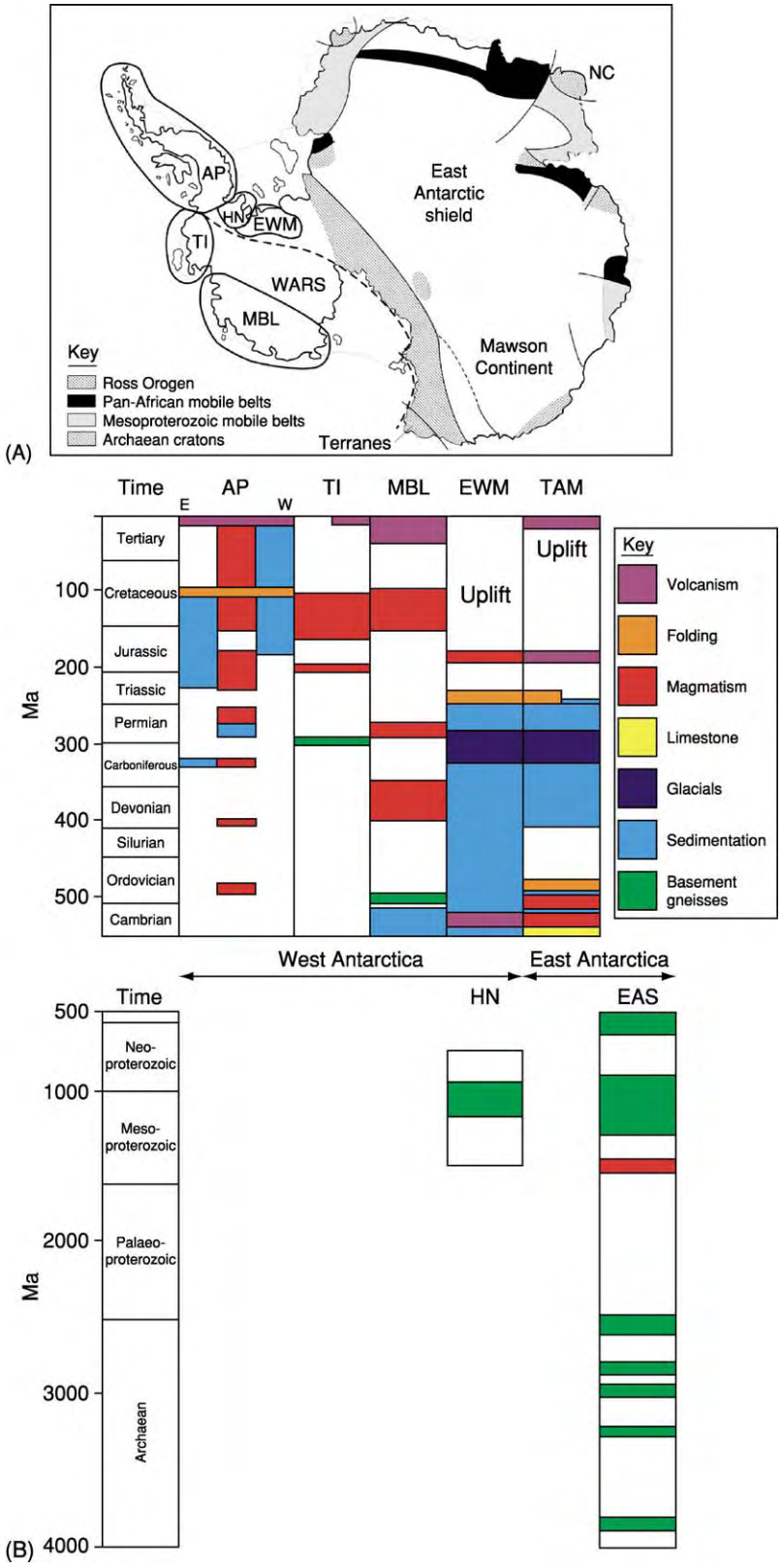


Figure 3 (A) Simplified geological map of Antarctica. AP, Antarctic Peninsula; EWM, Ellsworth Whitmore Mountains; HN, Haag Nunataks; MBL, Marie Byrd Land; TI, Thurston Island; WARS, West Antarctic Rift System. (B) Summary of the main geological events within the different crustal segments of Antarctica. E, East; W, West; EAS, East Antarctic Shield; TAM, Transantarctic Mountains; other abbreviations as in (A).

It is unclear as to what extent the other belt involved ocean closure or regional-scale transcurrent tectonics.

It is not known if the centre of the East Antarctic Shield is a single Palaeoproterozoic craton (Mawson Continent), as commonly assumed, or whether it is cut by Grenville-age or pan-African mobile belts, which would require modification of the boundaries of the Mawson Continent. This uncertainty means that the various terranes exposed on the coast can be joined in any number of ways and it follows that current models for Phanerozoic tectonics in East Antarctica and for global tectonics, given the crucial location of Antarctica in proposed supercontinents, will remain poorly constrained until some understanding of the continental structure beneath the ice-cap is achieved.

The Transantarctic Mountains

Marking the boundary between East and West Antarctica, the present-day intracratonic mountain chain has undergone episodic uplift since the Early Cretaceous and has been modelled as a major rift shoulder. The unifying geological feature of the mountains is a Middle Palaeozoic erosion surface (Kukri Peneplain) that separates gently tilted Devonian to Triassic sedimentary rocks (Gondwana cover sequence) and Jurassic continental tholeiites (Ferrar Supergroup) from a Proterozoic to Early Palaeozoic orogenic belt known as the Ross Orogen.

The Ross Orogen: The Palaeo-Pacific Margin of Gondwana

The recently formulated SWEAT hypothesis linking the North American Laurentian continent with East Antarctica has provided a powerful tectonic framework for interpreting the Late Proterozoic and Early Palaeozoic siliciclastic turbidites and volcanic rocks exposed along the Transantarctic Mountains as being deposited in a rift margin setting following the separation of Laurentia from Antarctica ~750 Ma. Following the Beardmore folding event, carbonates were deposited along the margin in Early Cambrian times. Outboard of the Early Cambrian limestone, Middle Cambrian carbonates, sedimentary rocks, and a bimodal volcanic sequence formed. The margin was subsequently transformed to an active Early Palaeozoic orogenic setting following the initiation of subduction of newly created proto-Pacific oceanic lithosphere beneath the rifted margin. Active deformation, volcanism, metamorphism, and emplacement of subduction-related igneous rocks (the Granite Harbour Intrusives) occurred along the mountain front from about 520 to 480 Ma. Large volumes of molasse sediment were shed into fore-arc marginal basins in the Middle Cambrian and Ordovician, primarily by erosion of volcanic rocks of the early Ross magmatic arc. The fore-arc deposits were intruded by late orogenic plutons as the locus of magmatism shifted offshore. Deposition of individual molasse sequences continued until 490–485 Ma. Three exotic terranes were emplaced along the North Victoria Land margin during the late stage of the Ross Orogeny.

Gondwana Cover Sequences: A Stable Continent

Unconformably overlying the Ross Orogen in the Transantarctic Mountains and in the once-neighbouring Gondwana continents is a flat-lying cover sequence up to 2.5 km thick. Known in Antarctica as the Beacon Supergroup, this represents a period of tectonic stability within the Gondwana continent that spans the Devonian to the Triassic. It is capped by basalt flows and is intruded by thick dolerite dykes and sills of Middle Jurassic age (~180 My old) in the Ferrar province, a large igneous province that heralded the breakup of Gondwana (Figure 4).

There are four phases of sedimentation within the Beacon Supergroup and its scattered equivalents around the periphery of the East Antarctic shield: the Devonian strata, Permo-Carboniferous glacial deposits, Permian sediments, and Triassic cross-bedded sandstones and mudstones. The Devonian strata, which were deposited on an extensive undulating surface known as the Kukri Peneplain, are largely mature quartzose sandstones with red and green siltstones deposited by a mix of shallow marine and alluvial sedimentation during warm and semiarid climatic conditions. Some of the strata contain fossil fish and characteristic trace fossils. The Permo-Carboniferous glacial deposits derive from Late Carboniferous times, when a thick ice-sheet covered a large part of Gondwana, including Antarctica, depositing a thick diamictite unit. The diamictite deposited from the continental ice-sheet ranges from 5 to 50 m thick and records at least four advance and retreat cycles. A thick glacial section in the northern part of the Ellsworth Mountains is considered to be glacial marine in origin. Glacial striae and associated features have been measured and provide a relatively simple picture of ice flow away from the central Transantarctic Mountains.

Permian sediments, the most widely distributed Beacon strata, are cross-bedded fine- to medium-grained sandstones, shale, and coal, the products of meandering braided rivers. Coal beds average about 1 m thick but range up to 10 m. The finer overbank deposits commonly contain the famous glossopterid

Figure 4 Middle Jurassic mafic sills related to the initial breakup of Gondwana, emplaced within a Triassic sedimentary sequence within the Theron Mountains, which are located in the Transantarctic Mountains.

leaves and fossilized stems and roots of the now-extinct glossopterid tree. The rapid and virtually complete disappearance of Late Palaeozoic ice is recorded in the earliest Permian strata. The *Glossopteris* leaves, dropped from the widespread woody deciduous tree or shrub, indicate a cool and wet rather than cold climate.

Following the end-Permian extinction event and the disappearance of the *Glossopteris* flora, the Triassic cross-bedded sandstones and mudstones were deposited by low-sinuosity rivers on a north-west-sloping plain. The reversal in palaeoslope is attributed to uplift associated with a Late Permian–Early Triassic Gondwanian folding that deformed the Cape Fold Belt and also the Permian and older strata of the Ellsworth and Pensacola mountains in Antarctica. The climate was mild and arid with a varied reptilian and amphibian fauna, and a flora characterized by the fossil fern known as *Dicroidium*.

Beacon sedimentation culminated in the Early Jurassic with explosive rhyolitic volcanism and volcanic debris flows, and ultimately by basaltic flows and associated intrusive rocks of the Ferrar Supergroup.

West Antarctica: A Collage of Crustal Blocks

The basin and range topography of West Antarctica can be used to delineate five physiographically defined crustal blocks (Figure 3) that have distinctive geological features and that may have existed as distinct microplates during the Mesozoic breakup of Gondwana.

Haag Nunataks: Part of the East Antarctic Shield

This small crustal block is situated between the southern tip of the Antarctic Peninsula and the Ellsworth Mountains. It is formed entirely of Proterozoic basement amphibolites and orthogneiss of Grenvillian ages (dating to between 1176 and 1003 Ma). Although the gneisses are exposed only on three small nunataks, aeromagnetic surveys show the full extent of the block and suggest that similar Proterozoic basement may underlie part of the Weddell Sea embayment region. Isotopic studies also suggest that this basement may be present beneath the neighbouring Antarctic Peninsula region. The gneisses correlate with Proterozoic basement gneisses within the East Antarctic Shield and may represent a fragment of the East Antarctic Shield displaced during the breakup of Gondwana.

Ellsworth Whitmore Mountains: A Displaced Fragment of the Gondwanian Fold Belt

The Ellsworth Mountains form a 415-km-long NNE–SSW-trending mountain range that contains the highest mountain in Antarctica. The mountains are situated along the northern periphery of the crustal block of the Ellsworth–Whitmore Mountains, which represents part of a displaced terrane once situated along the palaeo-Pacific margin of Gondwana, prior to supercontinent breakup, adjacent to South Africa and the Weddell Sea coast of East Antarctica. It was assembled in its present position by Late Cretaceous

times, following the Middle to Late Jurassic breakup of Gondwana. Some 13 km of sedimentary succession are exposed within the Ellsworth Mountains, representing a continuous Middle Cambrian to Permian succession that was deformed during the Late Permian to Early Triassic Gondwanian folding. Such a complete succession is unusual in Antarctica and is part of the Gondwana cover sequence that has similarities to the Transantarctic Mountains, including the Pensacola Mountains. The Early Palaeozoic succession of the Ellsworth Mountains comprises the Heritage Group (Early to Late Cambrian) and the Crashsite Group (Late Cambrian to Devonian). The Heritage Group is composed of clastic sedimentary and volcanic rocks that make up half the entire stratigraphic thickness within the Ellsworth Mountains. The volcanic rocks formed in a mid-Cambrian continental rift environment with mid-ocean ridge-type basalts erupted near the rift axis. The Heritage Group is overlain by 3000 m of quartzites of the Crashsite Group, Permo-Carboniferous glacial diamictites of the Whiteout conglomerate, and the Permian Polar Star Formation. Middle Jurassic granites related to the Gondwana breakup intrude the Ellsworth Mountain succession in scattered nunataks throughout the remainder of the Ellsworth-Whitmore Mountains crustal block.

Thurston Island: Pacific Margin Magmatic Arc

The Thurston Island block, which includes Thurston Island on the adjacent Eights Coast and Jones Mountains, records Pacific margin magmatism from Carboniferous to late Cretaceous times. The igneous rocks form a uniform calc-alkaline suite typical of subduction settings. On Thurston Island, the observable history began with Late Carboniferous (~300 Ma) emplacement of granitic protoliths of orthogneiss. Cumulate gabbros were emplaced soon after gneiss formation, followed by diorites that have Triassic ages. In the nearby Jones Mountains, the oldest exposed rock is a muscovite-bearing granite with an Early Jurassic age of 198 My. The subsequent evolution of the Thurston Island area was dominated by Late Jurassic and Early Cretaceous bimodal suites. Between 100 and 90 Ma, volcanism in the Jones Mountains became predominantly silicic prior to cessation of subduction along this part of the margin by collision or interaction of a spreading ridge with the trench. In the Jones Mountains, the basement rocks are unconformably overlain by postsubduction Miocene alkalic basalts.

Marie Byrd Land: Pacific Margin Magmatic Arc

Small scattered exposures throughout the coastal parts of Marie Byrd Land suggest that the block may contain two geological provinces or superterranes. In the western part, the oldest Palaeozoic rocks are a thick uniform sequence of folded sandstone-dominated quartzose turbidites of the Swanson Formation. In the eastern part, the older basement rocks include biotite paragneiss, calc-silicate gneiss, marble, amphibolite, and granitic orthogneiss with protolith ages of 504 My. In western Marie Byrd Land, the Swanson Formation was intruded by the Devonian–Carboniferous Ford Granodiorite and Cretaceous granitoids, whereas in eastern Marie Byrd Land, magmatic rocks are predominantly Permian and Cretaceous in age, indicating a long-lived magmatic history for Marie Byrd Land. The Cretaceous magmatic rocks include mafic dyke suites and anorogenic silicic rocks, including syenites and peralkaline granites that record an important change in the tectonic regime, from a subducting to an extensional margin, prior to separation of New Zealand from Marie Byrd Land and seafloor spreading from 84 Ma.

The Antarctic Peninsula: Long-Lived Andean-Type Margin

The Antarctic Peninsula is a long-lived magmatic arc built at least in part on continental crust with a record of magmatism and metamorphism that stretches back at least to Cambrian times. Pre-Mesozoic rocks are only sparsely exposed and include orthogneisses with protolith ages dating to ~450–550 Ma, paragneisses that form the basement to Triassic granitoids no older than Late Cambrian, and a few small granitic bodies ~400 My old. Locally, the basement underwent metamorphism, migmatization, and granite emplacement during Carboniferous (~325 Ma) and Permian (~260 Ma) times.

The Mesozoic geology of the Antarctic Peninsula has traditionally been interpreted in terms of a near-complete arc/trench system with Mesozoic accretion subduction complexes on the western Pacific margin of the Peninsula, a central magmatic arc active from 240 to 10 Ma (represented by the Antarctic Peninsula Batholith), and thick back-arc and retro-arc basin sequences on the eastern Weddell Sea side. The polarity of the system is consistent with east-directed subduction of proto-Pacific Ocean floor. However, the discovery of major ductile shear zones along the spine of the peninsula has led to the identification of separate domains and the possibility of the collision and accretion of separate terranes along the margin. Triassic and Early Jurassic plutons were emplaced along the palaeo-Pacific margin of Gondwana prior to the breakup of the supercontinent. The earliest plutons were peraluminous granites with S-type characteristics. By 205 Ma, metaluminous, type-I

granitoids were emplaced. Magmatism associated with the Jurassic breakup of Gondwana is represented by extensive silicic volcanism and associated subvolcanic plutonism that is part of the large Gondwana wide volcanic igneous province. Cretaceous and younger plutons were emplaced as a result of east-directed subduction of proto-Pacific ocean crust ranging in composition from gabbro to granodiorite, with a peak of activity between 125 and 100 Ma. The Tertiary part of the batholith is restricted to the west coast of the northern Antarctic Peninsula, signifying a major westward jump in the locus of the arc. With the exception of one small segment, subduction and its associated magmatism ceased in the Antarctic Peninsula between ~50 Ma and the present day, following a series of northward-younging (northward-facing) ridge trench collisions.

Gondwana Breakup: The Isolation of Antarctica

Four main episodes in the disintegration of the Gondwana continent led eventually to the isolation of Antarctica, to the development of the circumpolar current, and to an Antarctic continent covered in ice (Figure 5). The initial rifting stage started in Early Jurassic times (180 Ma) and led to formation of a seaway between West Gondwana (South America and

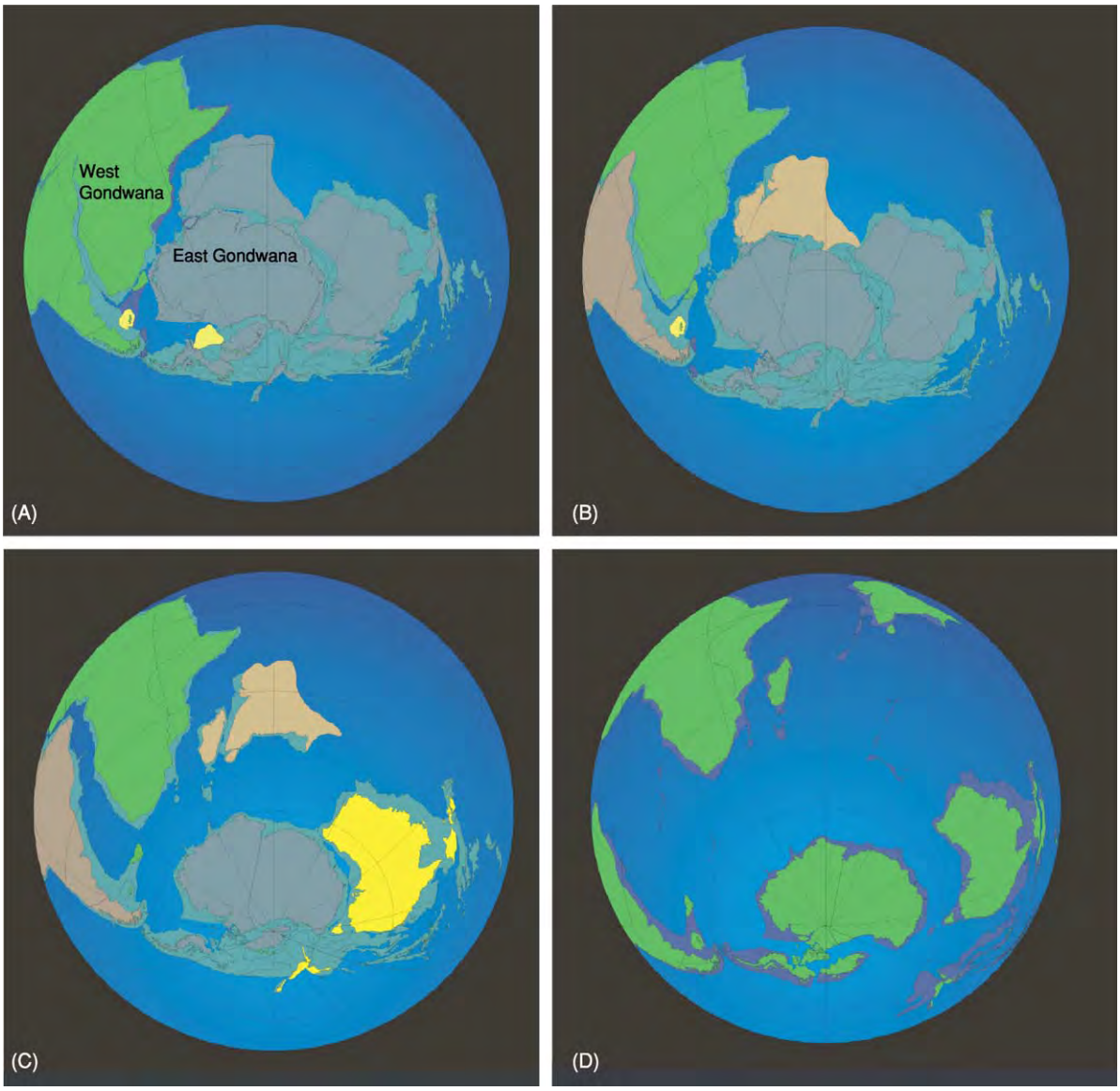


Figure 5 Snapshots of the distribution of the continents at (A) 150, (B)130, (C) 100, and (D) 35 Ma during the breakup of Gondwana, leading to the development of the circumpolar current and the polar isolation of Antarctica. Courtesy of Roy Livermore, British Antarctic Survey.

Africa) and East Gondwana (Antarctica, Australia, India, and New Zealand), and to seafloor spreading in the Somali, Mozambique, and possibly Weddell Sea basins. The second stage occurred in Early Cretaceous times (~130 Ma) when this two-plate system was replaced by three plates, with South America separating from an African–Indian plate, and the African–Indian plate separating from Antarctica. In Late Cretaceous times (90–100 Ma), New Zealand and South America started to separate from Antarctica until finally, approximately 32 Ma, the breakup of that once large continent was complete, when the tip of South America separated from the Antarctic Peninsula by opening of the Drake Passage, allowing the formation of the circumpolar current and thermal isolation of Antarctica.

The West Antarctic Rift System

Although the final isolation of Antarctica did not occur until opening of the Drake Passage ~32 Ma, a rift system formed within West Antarctica during Tertiary times. The rift system extends over a largely ice-covered area extending 3000 × 750 km, from the Ross Sea to the Bellinghausen Sea, comparable in area to the Basin and Range and the East African rift systems. A spectacular rift shoulder scarp, along which peaks reach 4 to 5 km maximum elevation, extends from northern Victoria Land to the Ellsworth Mountains. The rift shoulder has a maximum present-day physiographical relief of 7 km in the Ellsworth Mountains–Byrd Subglacial Basin area. The Transantarctic Mountains part of the rift shoulder has been rising episodically since Late Cretaceous times. The rift system is characterized by bimodal, mainly basaltic alkali volcanic rocks ranging in age from Oligocene or earlier to the present day. The large Cenozoic volcanic centres in Marie Byrd Land have been related to a mantle plume. Sedimentary basins within the rift system in the Ross Sea embayment contain several kilometres of Tertiary sediments that preserve a record of climate change from a greenhouse environment to an ice-covered world. There are 18 large central vent volcanoes in Marie Byrd Land that rise to elevations between 2000 and 4200 m above sea-level.

Antarctic Climate History: The Past 100 Million Years

The large-scale palaeogeography of the Antarctic region varied little from Late Cretaceous to Eocene times (~35 Ma), with the Antarctic continent still connected to South America and Australia and in a polar position. The West Antarctic rift system had begun to develop in the Late Mesozoic, with uplift in the Paleocene (45–50 Ma) as the Transantarctic Mountains began to rise and steadily erode. The earliest evidence of glaciers forming on the Antarctic continent comes as sand grains in fine-grained uppermost Lower Eocene and younger deep-sea sediments from the South Pacific, with isolated sand grains interpreted to record ice-rafting events centred on 51, 48, and 41 Ma. In the Ross Sea area, close to the Transantarctic Mountains, glaciers were calving at sea-level during the Eocene, becoming more extensive and spreading in earliest to Late Oligocene (~25 Ma), this being characterized by a number of thin till sheets separated by thin mudstone intervals. One of the mudstones contains a *Notofagus* leaf, which, along with contemporaneous beech palynomorphs, indicates a cool temperate climate on land during interglacial episodes, with many episodes of temperate ice-sheet growth and collapse. There was progressive disappearance of the *Notofagus*-dominated flora by the Late Oligocene (~24 Ma).

By the Early Miocene (~15 Ma), the Antarctic was completely isolated; there was development of the vigorous circumpolar currents and the present topography of the continent was in place. There was a large increase in ice cover beginning around 15 Ma. The majority view now is that since the Middle Miocene, Antarctic temperatures have persisted close to the present levels and that the East Antarctic ice-sheet has been a semipermanent feature during the past 15 My. This view is supported by work on well-preserved ice-desert landforms and deposits in mountains at the head of the Dry Valleys. These are dated from fresh volcanic ash deposits ranging between 4 and 15 My old, indicating that mean annual temperatures have been no more than 3°C above present at any time in the Pliocene. These and the geomorphic data suggest an enduring polar ice-sheet since Middle Miocene times. However, an alternative view of the post-Middle Miocene behaviour of the East Antarctic ice-sheet was presented as a consequence of finding a diverse biota of diatoms, sponge spicules, radiolarians, palynomorphs, and foraminifera in glacial diamictites or till deposits (the Sirius Group) at a number of locations high in the Transantarctic Mountains. These biota include Pliocene-age marine diatoms that may have been deposited in seas in the East Antarctic interior, subsequently to be glacially eroded and transported to their present sites by an enlarged East Antarctic ice-sheet. Although the transport processes and the depositional setting for the tills are well established, the origin of the age-diagnostic microfossils found in them has been in dispute. It is likely that some of the Pliocene-age diagnostic diatoms were deposited from the

atmosphere and thus cannot be used to date the associated deposit. The Sirius Group also includes terrestrial sequences that record many advances and retreats of inland ice through the Transantarctic Mountains. The uppermost strata include a shrubby vegetation that indicates a mean annual temperature 20° warmer than the present-day mean. Because the flora cover a long time range, the precise age of these deposits is not known.

The cyclical pattern of ice-volume change through the Quaternary is well established from the deep-sea isotopic records and is inferred from ice-core studies representing the past 400 000 years; the ice cores were taken from near the middle of the East Antarctic ice-sheet at Vostok Station. The data show a cyclical advance and retreat at 100 000-year intervals and that temperatures in the Antarctic interior, which were the same as present-day temperatures during the last interglacial, fell episodically to about 10° cooler during the Last Glacial Maximum, then rose rapidly to Holocene temperatures around 10 000 years ago.

See Also

Africa: Pan-African Orogeny; North African Phanerozoic; Rift Valley. **Andes**. **Argentina**. **Australia:** Proterozoic; Phanerozoic; Tasman Orogenic Belt. **Brazil**. **Gondwanaland and Gondwana**. **Indian Subcontinent**. **New Zealand**. **Oceania (Including Fiji, PNG and Solomons)**.

Further Reading

Barrett P (1999) Antarctic climate history over the last 100 million years. *Terra Antarctica Reports* 3: 53 72.

Collinson JW, Isbell JL, Elliot DH, Miller MF, and Miller JMG (1994) *Permian Triassic Transantarctic Basin. Geological Society of America Memoir 184,* pp. 173 222. Boulder, CO: Geological Society of America.

Dalziel IWD (1992) Antarctica; a tale of two supercontinents? *Annual Revue of Earth and Planetary Sciences* 20: 501 526.

Fitzsimons ICW (2000) A review of tectonic events in the East Antarctic Shield and their implications for Gondwana and earlier supercontinents. *Journal of African Earth Sciences* 31: 3 23.

Gamble JA, Skinner DNB, and Henrys S (eds.) (2002) *Antarctica at the Close of a Millennium*. Wellington: The Royal Society of New Zealand.

Lawver LA (1992) The development of paleoseaways around Antarctica. *Antarctic Research Series* 56: 7 30.

Leat PT, Scarrow JH, and Millar IL (1995) On the Antarctic Peninsula batholith. *Geological Magazine* 132: 399 412.

Macdonald DIM and Butterworth PJ (1990) The stratigraphy, setting and hydrocarbon potential of the Mesozoic sedimentary basins of the Antarctic Peninsula. *Antarctica as an Exploration Frontier* 139: 100 105.

Miller IL, Willan RCR, Wareham CD, and Boyce AJ (2001) The role of crustal and mantle sources in the genesis of granitoids of the Antarctic Peninsula and adjacent crustal blocks. *Journal of the Geological Society, London* 158: 855 867.

Moores EM (1991) The Southwest U.S East Antarctic (SWEAT) connection: A hypothesis. *Geology* 19: 425 428.

Storey BC (1996) Microplates and mantle plumes in Antarctica. *Terra Antarctica* 3: 91 102.

Stump E (1995) *The Ross Orogen of the Transantarctic Mountains*. Cambridge, UK: Cambridge University Press.

Tingey RJ (1991) *The Geology of Antarctica*. Oxford, UK: Oxford Science Publications.

Vaugham APM and Storey BC (2000) The eastern Palmer Land shear zone: a new terrane accretion model for the Mesozoic development of the Antarctic Peninsula. *Journal of the Geological Society, London* 157: 1243 1256.

ARABIA AND THE GULF

I A Al-Jallal, Sandroses Est. for Geological, Geophysical Petroleum Engineering Consultancy and Petroleum Services, Khobar, Saudi Arabia
A S Al-Sharhan, United Arab Emirates University, Al-Ain, United Arab Emirates

Introduction

Arabia consists of two main geological features: the Arabian shield and the Arabian shelf. Charles Doughty, who in 1888 produced the first geological map of Arabia, wrote "the Geology of the Peninsula of the Arabs consists of a stack of plutonic rock, whereupon lie sandstones, and upon the sand-rocks limestones. There are besides great land-breadths of lava and spent volcanoes". These two sentences encapsulate the geology of Arabia, and indeed of the whole of the southern shoreline of Palaeo-Tethys, from the modern Atlantic Ocean to the Arabian Gulf.

The Arabian shield is a Precambrian complex of igneous and metamorphic rocks that occupies roughly one-third of the western part of the Arabian Peninsula. The Arabian shield is a continuation of the adjacent African shield from which it is separated by the Red

Sea rift (*see* **Africa:** Rift Valley). The rocks in West Arabia, Yemen, Aden, and Oman are dated as Precambrian by radioisotopic dating and by the presence of Cambrian fossils in sediments above. Radiometric dates show that the Arabian shield was involved in the Pan-African Orogeny (*see* **Africa:** Pan-African Orogeny). The shield crops out along the east coast of the Red Sea rift, south to Yemen, and extend eastward into central Arabia with varied degrees of exposures.

The Arabian shield dips eastwards beneath some 6000 m of sedimentary rocks of Infracambrian to Recent age. These rocks crop out as belts surrounding the shield that dip gently east and north-east into the subsurface before they crop out again in Oman, United Arab Emirates, and Iran, mostly during the Mesozoic time that brought to surface the famous ophiolites of Oman. The geological formations are generally well exposed, due to a lack of vegetation, and can be traced along their outcrops for hundreds of kilometres. Geologists from Charles Doughty to the present day have noted that the uniform stratigraphy of Arabia can be traced westwards across North Africa to the Atlantic Ocean time (*see* **Africa:** North African Phanerozoic). This uniformity is most marked in the Lower Palaeozoic, and degrades thereafter through time.

Subsequent to the work of Doughty, geological mapping and fieldwork in Arabia in the 1930s was tied to oil exploration and covered more than 1 300 000 km^2. Figure 1 is the first published geological column of Arabia compiled by Powers and Ramirez in 1963.

During the early days of Aramco, geologists identified, measured, and mapped nearly 6000 m of sediments ranging in age from presumed Infracambrian to Recent. The main rock units have been identified and mapped since 1966 by Aramco and the US Geological Survey (USGS). Through the years much additional work has been performed by Saudi Aramco, USGS, and Bureau de Recherches Geologiques et Minieres (BRGM). Some formation names have changed and some new ones have been introduced in Arabia and the Gulf states, the result being a wealth of information both from outcrops and from the subsurface that has affected our knowledge of the formations, their contacts, ages, and nomenclatures; however, the original stratigraphic framework remains largely intact. This may be seen by comparing Figure 1 with Figure 2, which is the stratigraphic column for Saudi Arabia, recently produced by Aramco.

In central Arabia, the Palaeozoic and Mesozoic rocks crop out as curved belts bordering the Arabian shield, dominated by parallel west-facing escarpments capped by resistant limestone. In eastern Arabia the older sediments are mostly covered by a gently dipping belt of low relief Tertiary and younger deposits that include the Rub al Khali and north-eastern Arabia (Nafud) of Quaternary sands.

In north-western Arabia some 2000 m of largely lower Palaeozoic sandstone is exposed. The lower units of this Palaeozoic sandstone can be correlated into Jordan and across the Arabo-Nubian shield into North Africa. To the east lies a basin of Upper Cretaceous to Tertiary sediments. Volcanic rocks of Tertiary to Recent age cover substantial parts of the shield and adjacent cover. These result from deep crustal tension associated with the development of the Red Sea rift system.

The Stratigraphy of Arabia and the Gulf

The geological section above the Precambrian basement Arabia falls into eight major divisions separated by unconformities (Figures 1 and 2). These may be summarized as follows from base to top:

1. Infracambrian and Lower Palaeozoic clastic rocks (Cambrian through Lower Devonian);
2. Carboniferous, Permian, and Triassic carbonate/clastic rocks (Upper Permian through Upper Triassic);
3. Lower and Middle Jurassic clastic and carbonate rocks (Toarcian to Callovian?);
4. Upper Jurassic and early Lower Cretaceous carbonate rocks (Callovian through Valanginian);
5. Late Lower Cretaceous clastic rocks (Hauterivian through Aptian);
6. Middle Cretaceous clastic rocks (Cenomanian through Turonian?);
7. Upper Cretaceous to Eocene carbonate rocks (Campanian through Lutetian); and
8. Neogene clastic rocks (Miocene and Pliocene).

Infracambrian and Lower Palaeozoic Clastic Rocks (Cambrian through Lower Devonian)

Above the igneous and metamorphic basement of the Arabo-Nubian shield the Infracambrian cover consists of sandstone, carbonates, shale, and salts. The Infracambrian shows the oldest fossils. The Infracambrian Huqf Group of Oman contains potential petroleum source rocks. The Huqf Group includes the Mahara, Khufal, Shuram, Buah, and Ara Salt formations, ranging in age from Precambrian to Lower Cambrian. The evaporites are usually referred to as the Hormoz Salts, and have formed many diapiric structures in Oman and throughout the Gulf, many of which are petroliferous. In central Arabia the Precambrian shield is overlain by the Infracambrian

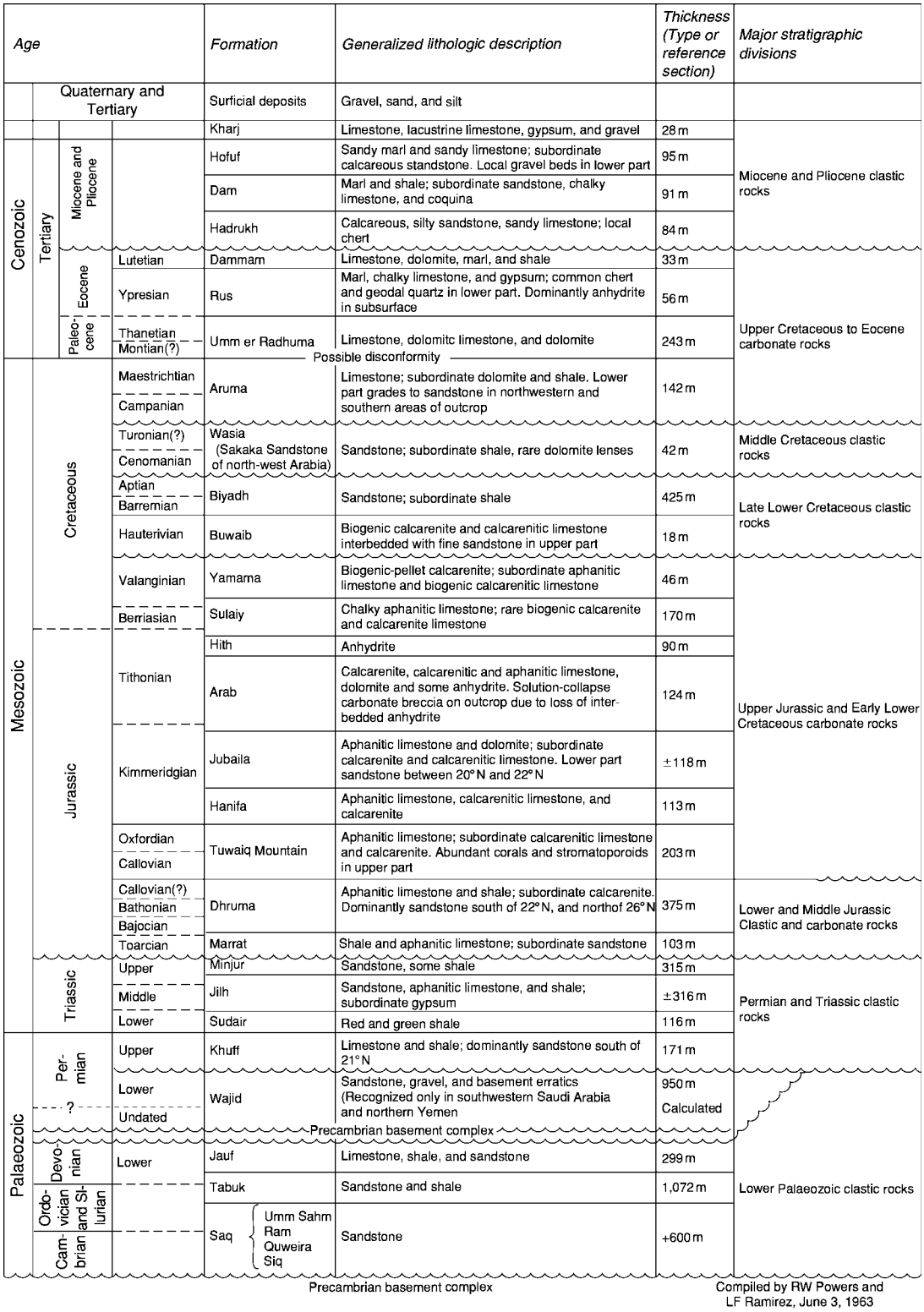


Figure 1 The first geological column of Saudi Arabia, compiled by RW Powers and LF Ramirez in 1963. After Powers *et al.* (1966).

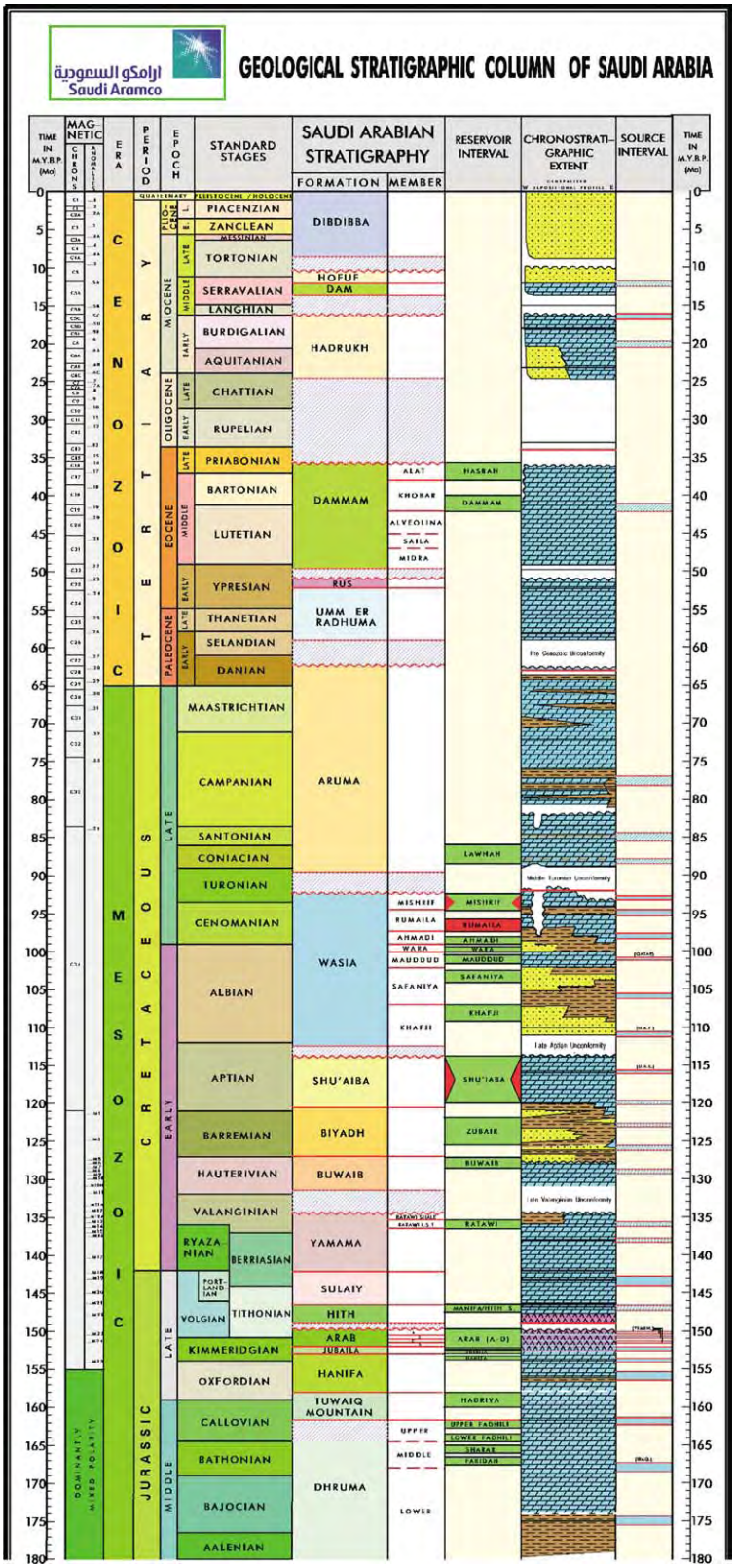


Figure 2 Modern geological stratigraphic column of Saudi Arabia. After Qahtani *et al.* (2004).

Continued

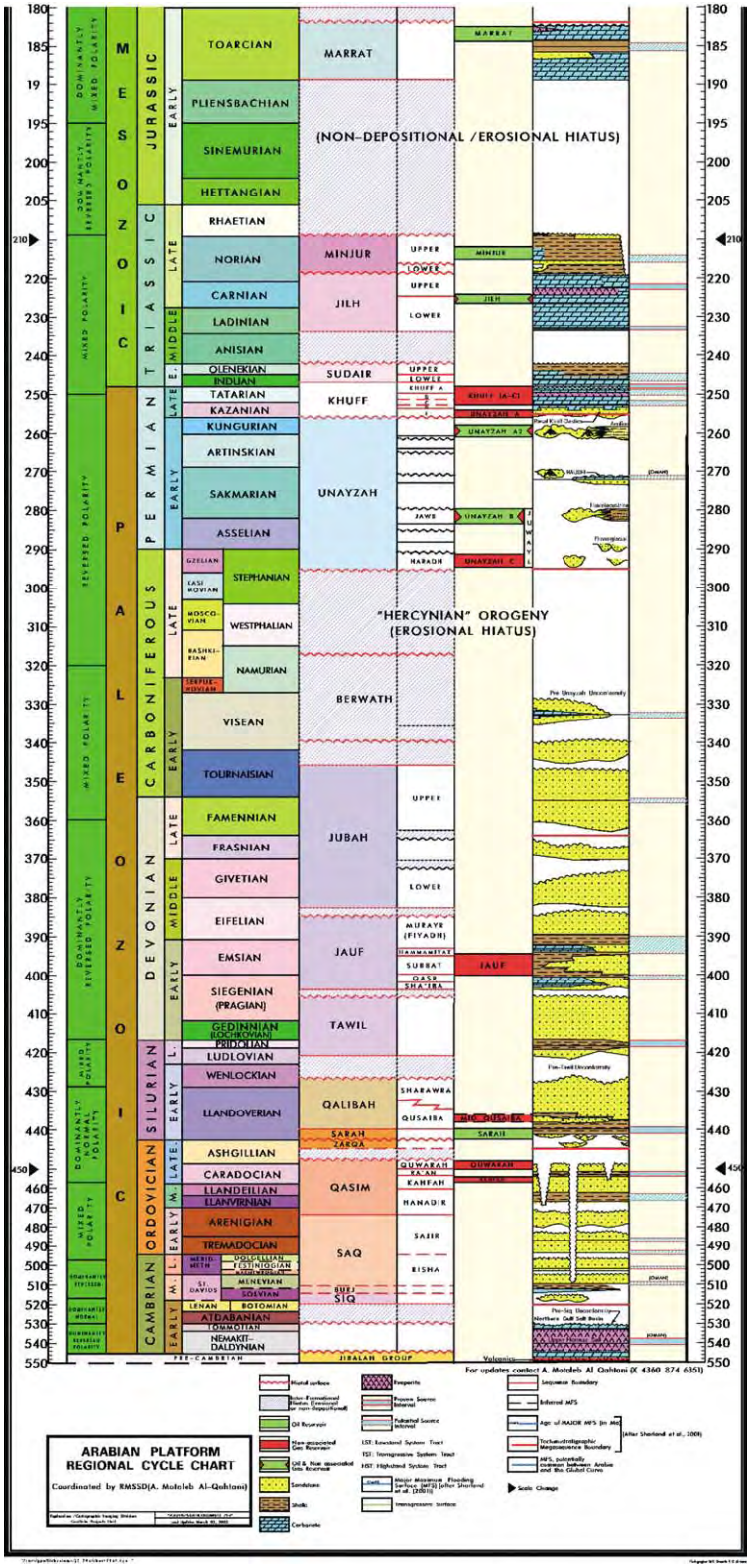


Figure 2 Continued

Jubailah Group, whose biostratigraphy is poorly known, but can be correlated with Huqf Group of Oman. The Lower Palaeozoic rocks are mainly quartz, sandstone, and shale, with some thin carbonate beds in the upper part. They range in age from Cambrian to Lower Devonian. These include the Saq, Tabauk, Tawil, and Jauf formations. Some of these old names have been replaced by newer names. For example, the Tabuk has been replaced by the Qasim. Additional formations have been added, such as the Kahfah and Juba. The 'old' Wajid Sandstone (Figure 1) in south-western Arabia has been divided into four new formations. The Cambro-Ordovician Saq and Qasim formations can be correlated with the Haima Group in Oman, which includes the Amin, Andam, Ghudun, and Safiq formations. In Jordan the Saq Sandstone can be divided into four units: the Siq, Quweira, Ram, and Um Sahm. The Siq is probably equivalent to the Yatib Formation of the BRGM in central Arabia. The Burj Formation of Jordan (Middle Cambrian carbonates and shale) was introduced to the Aramco chart in 1992, and Laboun suggested that it be renamed the Farwan Formation in al-Jauf near Jordan. The Burj carbonates overlie the Siq or Yatib in north-western Saudi Arabia. The Saq Sandstone is of mainly fluvial origin, though the presence of *Cruziana* tracks in abandoned shale channels towards the top suggest marine influence. The Saq Formation has two members, the Resha and the Sajir.

The Saq Formation is uncomfortably overlain by the Middle Ordovician Qasim Formation of sandstone with alternating Shale members. These include the Hanadir Shale, the Kahfah Sandstone, the Raan Shale, and the Qawarah Sandstone. The shales contain rare graptolites, of great biostratigraphic significance, and the sandstones locally contain abundant burrows variously termed *'Tigillites'*, *'Sabellarifex'*, and *'Scolithos'*, which are indicative of intertidal conditions.

The Qasim is unconformably overlain by the Zarga and Sarah formations of Late Ordovician and Early Silurian ages; these two formations are famous for their glacial features. The Mid- to Late Silurian age Qalibah Formation follows with its two members, Sharwrah and Qusaiba. Formerly the Qalibah was called the Tayyarat Formation developed by BRGM studies. The Qusaiba Shale is the main source rock for the Palaeozoic gas in the Jauf (Devonian), Unayzah (Permian), and Khuff (Late Permian–Early Triassic) reservoirs.

Sedimentological studies reveal that this Lower Palaeozoic sequence resulted from deposition on alluvial braid plains that passed down-slope into intertidal and shallow marine shelves, which in turn passed into basinal mud. Graptolites in the latter can be used to demonstrate the diachronous progradation of these facies across the Arabian shield. These events were coeval with similar progradational episodes northwards across the Saharan shelf.

The Tawil Formation (Late Silurian and Early Devonian) unconformably overlies the Sharwrah member of the Qalibah Formation. This sandstone is very different from other Palaeozoic sandstones because it contains an abundance of heavy minerals. These give a spiky character to the spectral gamma ray log, which aids subsurface correlation because the Palaeozoic sands commonly lack palynomorphs that can be used for age dating. The Tawil Formation is conformably overlain by the Jauf Formation. The Jauf Formation is remarkably different in character from where it crops out in the north to where it is found in the subsurface of eastern Arabia. Where it crops out near the town of Jauf it is mostly marine limestone, whereas in the subsurface in eastern Arabia it consists of deltaic sandstone. The Jauf Formation is one of the main Palaeozoic gas reservoirs in Arabia. In the Jauf Formation reservoir drilling problems are usually encountered due to considerable amounts of permeability-inhibiting illite. The palynology of the Jauf Formation is more useful for environmental than for stratigraphic analysis (Figures 1 and 2).

The Jauf Formation is overlain by the Jubah Formation. This consists of Middle Devonian to Lower Carboniferous cross-bedded sandstone that used to be considered part of the Jauf Formation, but is now separated from it. It is unconformably overlain by the Wasia Formation in the Sakaka area of northern Saudi Arabia. The Jubah Formation can be correlated with the Sakaka Sandstone, whose age was controversial, having once been considered Cretaceous, and equivalent to the Wasia Formation.

The Wajid Sandstone of south-western Arabia has been divided into five members: the Dibsiyah (Cambrian, equivalent to Saq), the Sanamah (Late Ordovician, equivalent to Sarah/Zarga), the Silurian Qusaiba, the Khusyyayn (Devonian–Carboniferous, equivalent to Tawil, Jauf, Jubah, and Berwath) and the Juwayl (Permo-Carboniferous) equivalent to Unayzah and its units. The lower parts of Unayzah has recently been interpreted by Aramco as the equivalent of the Juwayl or Wajid Sandstone and to the Al Khlata Formation of Oman.

Carboniferous, Permian, and Triassic Carbonate/Clastic Rocks (Upper Permian through Upper Triassic)

Lower Palaeozoic strata are succeeded by some 1000 m of interbedded carbonate and clastic sediment of Upper Permian to Triassic ages, with some

basal Permo-Carboniferous rocks. The old stratigraphic terminology included the Berwath (Carboniferous), Khuff, Sudair, Jilh, and Manjur Formations. Of these, only the Berwath has been renamed (Figures 1 and 2). The Unayzah Formation was introduced to include the Pre-Khuff clastics. The new formation includes the clastics below the Khuff together with the basal Khuff clastics. The Unayzah is one of main sweet gas reservoirs in the Palaeozoic rocks of Arabia. It is divided into the A, B, and C reservoirs. The age of this formation is still ill-defined. Currently it is believed to span the Carboniferous/Early Permian to Late Permian, usually referred as a Post-Hercynian orogeny event. Many intervals of the Unayzah reservoirs, however, differ from each other and from the lower unit (the Unayzah C). This is more cemented than the overlying B and A units, suggesting a very wide gap in age within the Hercynian period. This in turn led geologists to generate other nomenclatures to separate these sections. The Unayzah rocks are mostly alternating red beds with three sandstone reservoirs. Several depositional environments have been suggested for the Unayzah Formation. These include eolian dunes, meandering streams, incised valleys, deltas and parabolic and coastal plain deposits, and variations of the above. The Berwath Formation, however, is only known in the subsurface, and it has been suggested that the name be discarded. The Berwath rocks are similar to those of the Unayzah Formation. The Unayzah Formation in Saudi Arabia can be correlated with the Gharif and Al Khlata (Houshi group) in Oman and Faragan in Iran, while the Unayzah upper reservoirs (partly A and B) can be correlated with the Gharif Formation and the Unayzah C with the Al Khlata Formation. Debates on the nomenclature and correlation of the stratigraphic units of the Unayzah Formation continue. The Unayzah is one of the main Palaeozoic sweet gas reservoirs in Arabia.

The shallow marine sabkha carbonates and evaporites of the Khuff Formation unconformably overly the Unayzah Formation. The sequence starts with a basal shallow marine clastic unit. The Khuff reservoirs hold large volumes of gas in both Saudi and the Gulf states. The Khuff gas is usually sour due to its sulphur content, which increases northwards with increasing anhydrite. The Khuff reservoirs are normally dolomitized, with lenses of limestone. The reservoirs are heterogeneous, even though the main units are correlatable for long distances. The formation has undergone extensive diagenesis, including leaching, anhydrite cementation, and dissolution, that made it very hard to predict reservoir character. The Khuff Formation is equivalent to the Akhdar Formation of Oman. In the subsurface the Khuff has been divided into seven members: Khuff A, Khuff B, Khuff C, Khuff D, Khuff E, Northern Sandstone/Evaporate Member, and Southern Sandstone/Shale Member.

The Khuff Formation is overlain with local unconformities by the Sudair shale. This marks a change to a more restricted depositional environment from sabkha evaporites to terrestrial red beds. The Sudair Shale is followed by the Jilh Formation, which consists of interbedded sandstone, shale, limestone, anhydrite, and salt. It is often overpressured and hazardous to drill through. The formation has few oil shows in Arabia. The Jilh Formation is conformably overlain by the Minjur Formation (Upper Triassic), which consists mainly of sandstone and shale and is a very good aquifer for central Arabia. These rocks correlate with the lower part of the Sahtan unit in Oman.

Lower and Middle Jurassic Clastic/Carbonate Rocks (Toarcian to Callovian?)

Jurassic rocks are mainly marine shale interbedded with carbonates in central Arabia, grading to sandstone in northern and southern areas. These include the Marrat, Dhruma, and Tuwaiq Mountain Formations (Figures 1 and 2). The Marrat Formation unconformably overlies the Minjur Formation. The Jurassic system in central Arabia is dominated by the Tuwaiq Mountains escarpment in the outcrop with coral heads. Marrat and Dhruma are exposed in lower relief structures, marked by the red and green shales that alternate with resistant caps of yellowish limestone. These formations contain few oil reservoirs in Arabia.

The Lower and Upper Jurassic formation names have remained unchanged in Saudi Arabia, but other names have been introduced in other Gulf countries. For example, the Marrat Formation in Kuwait carries the same name, but the Tuwaiq Mountain Formation has been replaced by the Sargelo Formation.

Upper Jurassic and Early Lower Cretaceous Carbonate Rocks (Callovian through Valanginian)

The Upper Jurassic to Lower Cretaceous rocks are mostly cyclic carbonate sands and evaporites that close several stages of the Jurassic. These include the most important oil-bearing Arab formation in Arabia. The formation names are largely unchanged from the earliest days of research (Figures 1 and 2). The upper parts of the Tuwaiq Mountain Formation consist of mainly carbonate grainstone and packstone with corals and stromatoporoids, followed by the Hanifa Formation, which is composed of carbonate mudstone, wackestone, and grainstone. The Hanifa is the main source rock of the Jurassic oil of Arabia. This is overlain by the Jubaila Formation, which is composed of mainly mudstone and wackeo packstone; the famous Arab D reservoir can extend to include the upper parts of the Jubaila.

The Jubaila is succeeded by the Arab Formation (Figures 1 and 2), the most famous oil reservoir in Arabia, especially the Arab D unit. The Arab Formation includes four members A, B, C, and D, each of which consists of a carbonate reservoir with an anhydrite cap rock. The reservoirs include alternating ooidal, skeletal, and peloidal grainstone, wackestone, and packstone containing *Cladocropsis*, stromatoporoids, and foraminiferans. They also include some mudstone. The facies indicate shallowing upward sequences from high-energy shoal, through intertidal flat, to supratidal sabkha (salt marsh) and subaqeous anhydrites (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves). The Arab D Member of the Arab Formation contains most of the oil reserves of the Ghawar Field, the largest oil field in the world, more than 250 km long and 50 km wide.

The Arab Formation is well known in the subsurface by cores and logs, but the outcrop is poorly known from small hilly exposures that show signs of anhydrite karst terrain near Riyadh (*see* **Sedimentary Processes:** Karst and Palaeokarst). The Arab Formation extends into the Gulf states with minor modifications. For example, in Qatar, it is divided into the Fahahil and Qatar formations. The Arab Formation extends into Abu Dhabi where the amount of anhydrite increases offshore. Major facies changes of the Arab Formation occur in Kuwait where the Jubaila and Arab formations of the Gotnia Basin are largely evaporites.

The Jurassic rocks of Yemen are very different yet again, consisting of sandstone and shale with minor carbonates. The Naifa and Hajar formations (Kimmeridgian-Tithonian) are similar to the Arab Formation further north.

In Saudi the Arab Formation is overlain by the Hith Formation, which consists mainly of anhydrite with minor carbonate reservoirs and crops out at Dahl Al-Hith in the Al-Kharj area near Riyadh; elsewhere it has generated karst terrane. The Hith is overlain by the Early Cretaceous Sulaiy Formation, which is composed of fossiliferous chalky limestone with wackestone and packstone, and is overlain in turn by the Yamama Formation (Berriasian-Valanginian) composed of bioclastic and pelletal grainstone, wacke to grainstone and mudstone (Figures 1 and 2).

Late Lower Cretaceous Clastic Rocks (Hauterivian through Aptian)

The Late Cretaceous marks a change from predominantly carbonate to terrigenous sedimentation. The sequence commences with the Buwaib Formation, which is a thin basal carbonate composed of mainly bioclastic grainstone and packstone that pass up into sandstone, overlain by sandstone and shale of the Biyadh Formation. This is truncated by an unconformity between the Aptian and Cenomanian stages. In Kuwait, however, the Buwaib and Biyadh Formations are grouped together as the Zubair Formation, while in Abu Dhabi and Oman the Yamama, Buwaib, and parts of the Biyadh Formations are termed the Lekhwair Formation. The Biyadh Formation is overlain by the Shu'aiba Formation (Aptian). This does not crop out at the surface, and like most formations, thickens eastwards from the shield to the shelf. The Shuaiba Formation contains many rudist 'reefs' which are important oil fields, such as the Shaybah field in the Rub Al-Khali (empty quarter) of Arabia and the Bu Hasa field in the United Arab Emirates (*see* **Sedimentary Environments:** Reefs ('Build-Ups')).

Middle Cretaceous Clastic Rocks (Cenomanian through Turonian?)

The Shuiaba Formation is overlain by deltaic sandstone and shale of the Albian–Turonian Wasia Formation (Figures 1 and 2). This crops out occasionally as low-lying hills. The Wasia Formation includes the Ahmadi, Moudoud, Khafji, and Safaniyah members in the northern parts of Arabia. The Safaniyah and Khafji members contain huge oil reservoirs in northern Saudi Arabia and Kuwait. In some of the Gulf states the Wasia Formation and its constituent members have been raised to group and formation status, respectively. In Kuwait its equivalent is the Burgan Formation. In Oman its equivalent is the Nahr Umr Formation, a predominantly shaley unit that serves as a source rock and seal to the underlying Shuaiba reservoirs.

Upper Cretaceous to Eocene Carbonate Rocks (Campanian through Lutetian)

Wasia sedimentation culminated in an important regional unconformity of Turonian age, overlain by some 500 m of limestone of the Aruma Formation (Figures 1 and 2) that crops out in an escarpment extending northwards from Arabia into Iraq. The Aruma Formation ranges in age from Turonian to Danian, and is composed mainly of carbonate rocks with some shale towards the base and increasing amounts of sand towards the north. It is equivalent to the Fiqa Formation in Oman.

The Aruma Formation is overlain unconformably by the Um er Radhuma Formation (Selandian–Ypresian), which is a highly porous and permeable dolomitic carbonate and an important aquifer in eastern Arabia. This formation is overlain by the Ypresian Rus Formation, which consists of interbedded marls, chalky limestone, gypsum, and anhydrite with quartz geodes. The gypsum and anhydrite usually dissolve to form collapse breccias that create drilling problems and also increase the salinity of adjacent aquifers. The

Rus Formation is overlain by the Dammam Formation, whose type locality is the famous Dammam Dome, on which was drilled the first well to discover oil in Saudi Arabia. This was Dammam well 7, which found oil in the Jurassic Arab Formation in 1933. The Dammam Formation consists of a number of members that include shale and carbonates, are locally reefal, and serve as local aquifers. The formation is truncated by an Eocene–Miocene unconformity.

Neogene Clastic Rocks (Miocene through Pliocene)

Sediments of Neogene age (Miocene–Pliocene) include the Hadrukh, Dam, Hofuf, Kharj, and Dibdibah formations (Figures 1 and 2). These comprise alternating limestone, chalky limestone, marly sandstone, gravel, and gypsum. The upper part consists of about 200–600 m of nonmarine marly sand and sandy limestone that crop out across the Rub Al-Khali (Empty Quarter) and northeastern Arabia. The Tertiary in Oman is represented by the Hadhramout and Fars groups.

All of the sediment above is locally covered by unconsolidated Quaternary sand and gravel, which is the major contributor to the Rub Al-Khali sand in southern Arabia, an area of about 600 km^2. These include both eolian dunes in the sand seas and vast plains of fluvial sand and gravel.

The Structural Geology of Arabia and the Gulf

As mentioned in the Introduction, Arabia consists of two main structural elements, the Precambrian shield of igneous and metamorphic rocks in the west, and the shelf whose sediment thickens eastwards towards the great mobile belt of the Taurus, Zagros, and Oman Mountains. The Arabian Plate (Figure 3) extends from the eastern Mediterranean to the greater part of Arabia (Arabian shield, Arabian platform, and Arabian Gulf), and the western Zagros Thrust Zone—an area enclosed by latitudes 13° and 38° N and longitudes 35° and 60° E. The natural boundaries of the Arabian Plate are most easily defined to the north and north-east, where the Taurus Mountains pass eastwards into the Zagros Fold Belt, which passes in turn eastwards into the Makran ranges. The structures of the Zagros can be traced into the northern Oman Mountains. The region is bounded to the south by the Owen Fracture Zone in the Indian Ocean and the Gulf of Aden Rift, and to the west by the rift system of the Red Sea and the Gulf of Aqaba and Dead Sea.

The area of the Arabian Plate is more than 3 000 000 km^2. Geologically nearly one-third of the area is composed of Neoproterozoic igneous and metamorphic rocks of the Arabian shield, of which, by far, the greatest part lies within western Saudi Arabia with minor inliers in Yemen. This shield was formed by the accretion of a series of Precambrian volcanic island arcs that can be traced into north-east Africa. The Arabian shield represents a fragment of Gondwana that separated as a result of the Phanerozoic tectonic events that were involved with the demise of the Palaeo-Tethys Ocean. Prior to breakup, Gondwana was an important source of widespread clastic sedimentation in the Palaeozoic (from Cambrian to Mid-Permian) and its spread over the platform/interior homocline. The main structural elements of the Arabian Plate are shown in Figure 4.

Sedimentary rocks were deposited over the Arabian Plate from Late Precambrian to Late Cenozoic as a result of a series of major tectonic phases. These began with an intracratonic phase (Late Precambrian to Mid-Permian), followed by a passive margin phase (Mesozoic), and concluded with an active margin phase (Cenozoic). During the Palaeozoic era much of the Arabian Plate lay south of the tropics, and was affected by glaciation in the Late Proterozoic, in the Late Ordovician and in the Carboniferous–Permian. It was dominated by the deposition of clastics, but interrupted by episodes of warm-water carbonate deposition in the Middle Cambrian, the Devonian, and the Upper Permian.

In contrast, during the Late Permian to the Holocene the area lay in subtropical and equatorial latitudes and was dominated by carbonate deposition.

The Arabian Plate experienced a number of events, including the rifting and sea-floor spreading of the Red Sea and Gulf of Aden, collision along the Zagros and Bitlis sutures, subduction along the Makran zone, and transform movement along the Dead Sea and Owen-Sheba fault zones. The Makran and Zagros convergence zones separate the Arabian Plate from the microplates of interior Iran.

The main structural elements in the Arabian platform contain several inherited, mechanically weak trends. These include:

1. North-trending highs as exemplified by the En Nala (Ghawar) anticline and the Qatar Arch;
2. The north-west-trending Mesozoic grabens of Azraq (Wadi Sirhan and Jauf) and Ma'rib; and
3. North-east-trending systems like the south Syria Platform and the Khleissia and Mosul trends.

These trends suggest that rejuvenation of basement discontinuities played an important role in the evolution of Arabia.

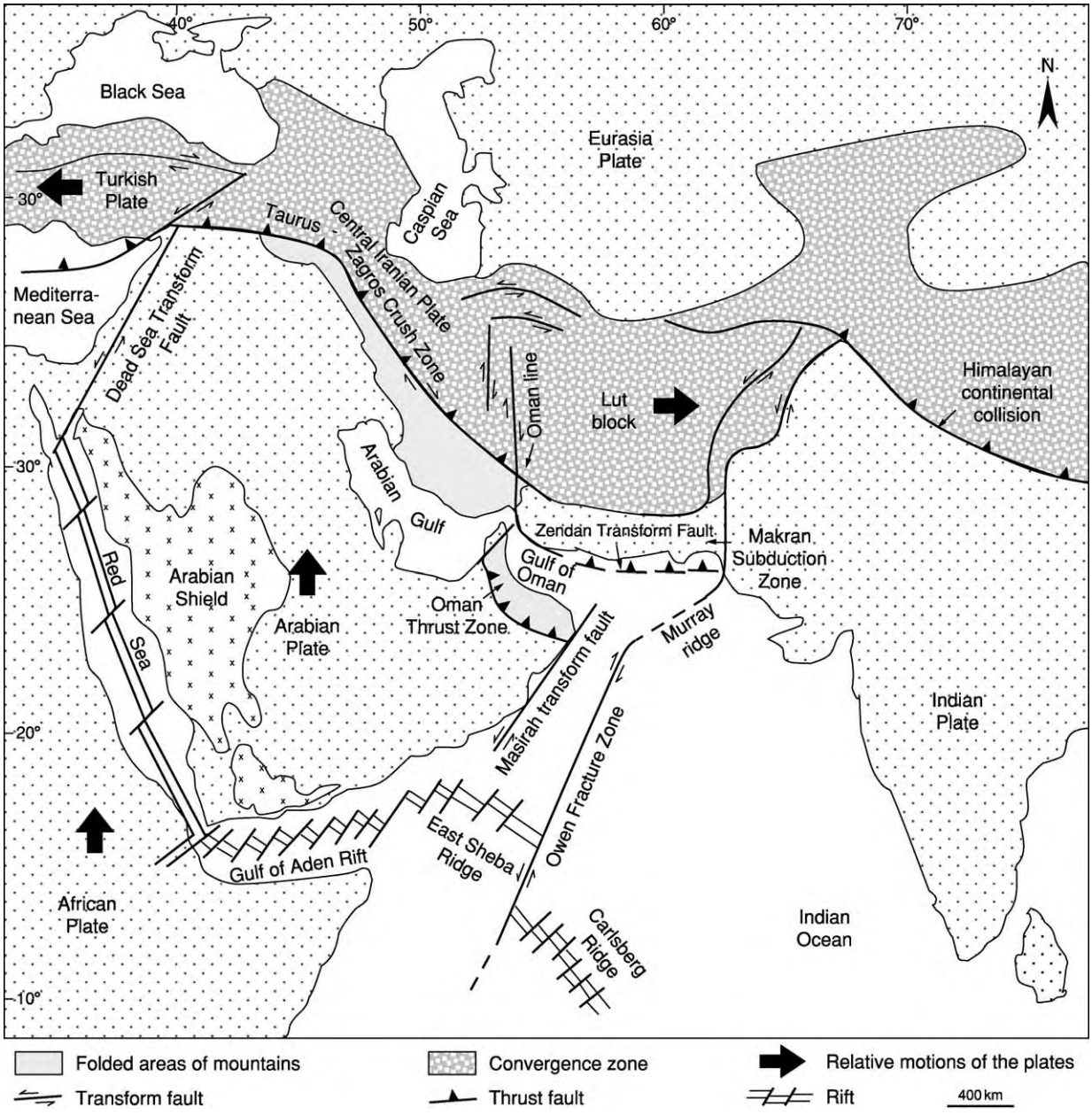


Figure 3 Tectonic setting of the Arabian Plate in relation to adjacent plates.

The Late Precambrian rocks of the Arabian Plate result from the accretion of a mosaic of terranes and ophiolitic sutures (dating to about 870–650 Ma) with later Neoproterozoic intrusions and depositional basins that together formed the basement of Arabia. The basement evolved and consolidated through the coalescence of several island-arc terranes over a long time span in the Proterozoic. Each closure and arc collision resulted in deformation and ophiolite obduction preserved as cryptic sutures.

Faulting in the Najd and the development of rift basins with thick salt sequences, including the salt basins of Oman and the Hormuz in the Arabian Gulf, occurred in the Late Precambrian to Early Cambrian (dated to about 610 to 520 Ma).

A Late Palaeozoic Orogeny is believed to have caused uplift and erosion over most of the Arabian Gulf and some parts of the Middle East Craton. Erosion related to this event tentatively dated as Early Carboniferous cut deeply into Cambrian and Precambrian strata. Following the earlier development of a carbonate shelf along the north and northwest margins in the Early to Middle Cambrian, the plate was covered by continental, deltaic, and

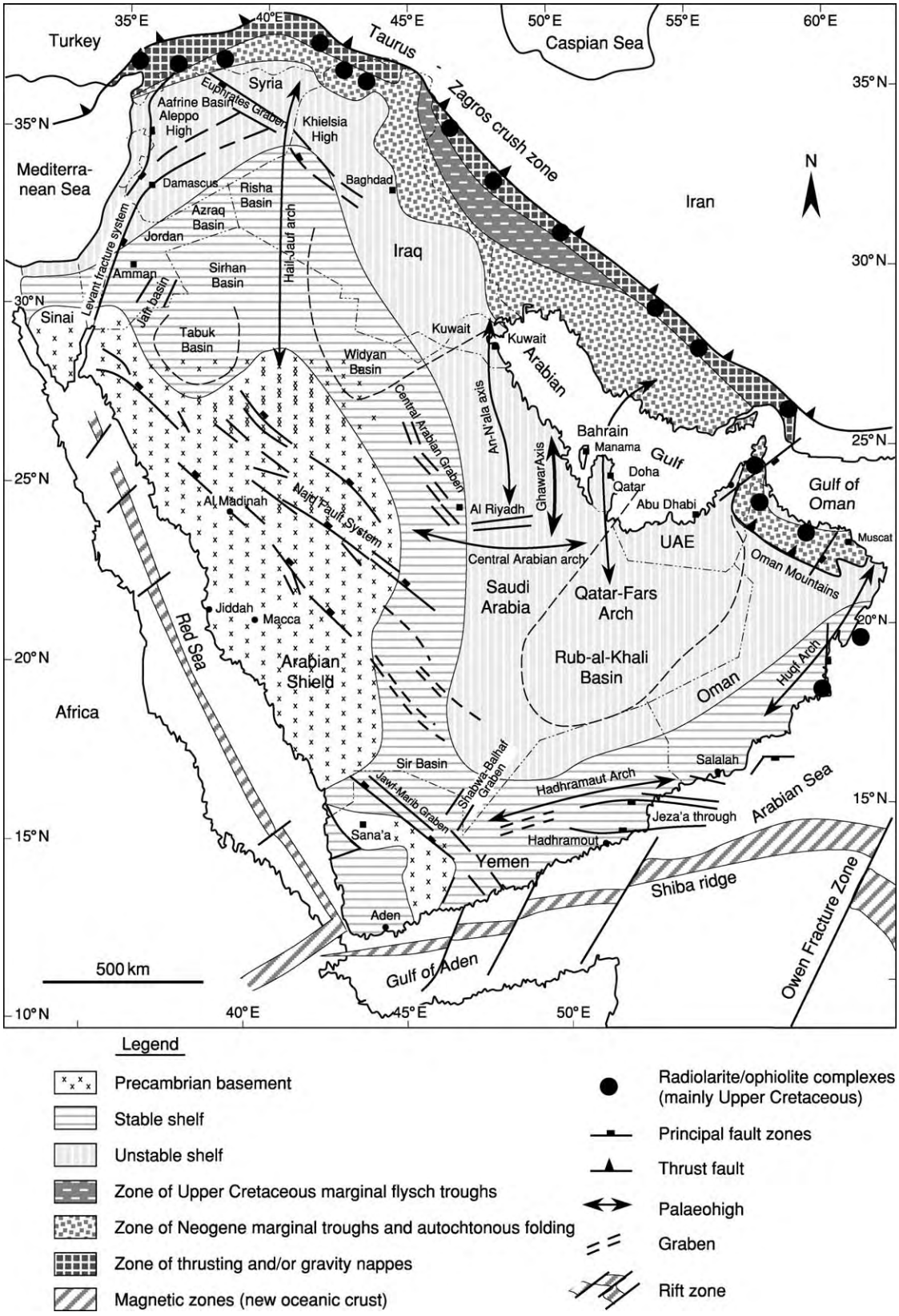


Figure 4 Main structural elements in the Arabian Plate. Reprinted with permission from Alsharhan and Nairn (1997) *Sedimentary Basins and Petroleum Geology of the Middle East*.

shallow-marine clastics. Carbonate deposition continued into the Early Ordovician on the Iranian terraces (Figure 4).

During the Late Ordovician and Early Silurian, the central and western parts of the Arabian Peninsula were covered by Saharan glaciers that advanced from the south pole, which was then located in African Gondwana. During this period, nondeposition, erosion, or marginal marine conditions prevailed in eastern and northern Arabia. Deglaciation in the Early Silurian led to a sharp sea-level rise, and the Palaeo-Tethys Ocean transgressed the Arabian and adjoining plates, depositing a thick, widespread, organic-rich shale directly over the glaciogenic and periglacial rocks of Arabia.

There is a general absence of Devonian deposits over the north-eastern Arabian shelf region with the exception of parts of north-eastern Iraq and locally in Oman, and the carbonate deposition reflects a return to lower latitudes. In the Late Devonian to Early Carboniferous the onset of south-west-directed subduction along the former passive margin initiated a phase of back-arc rifting and volcanism. Lower Carboniferous carbonates occur in northern Iraq, but elsewhere the Carboniferous is largely missing, reflecting regional emergence, nondeposition, or erosion.

Following the Hercynian orogeny in Late Carboniferous to Late Permian times the Central Arabian Arch developed as a nondepositional uplift, and the Rub Al Khali became a large nonmarine intracratonic basin. In south and south-east Arabia uplifted areas developed in the Hadhramaut–Huqf and in the vicinity of the present Oman Mountains.

Glaciation occurred in Oman, southern Saudi Arabia, and Yemen, and periglacial and fluviatile conditions existed in central Arabia. In Oman and Yemen tillites rest directly on a glacially striated Precambrian basement.

During the deposition of the Permo-Triassic sequence, back-arc rifting continued at the northern end of the Arabian Plate, and the new north-east passive margin was transgressed by a shallow Permian sea from which was deposited carbonates and evaporates, thickness variations of which indicate syndepositional tectonic activity. Over Arabia the thickness is almost uniform, ranging between 300 and 600 m, but thickens dramatically to more than 1200 m east- and southwards. The main depocentres for the Late Permian carbonates trend approximately north-west–south-east, parallel to the axes of the opening Neo-Tethyan and Hawasina oceans in Oman. There is a general thickening towards the north-east Arabian Gulf and Gulf of Oman and Iran.

A major period of Late Triassic uplift and erosion affected the southern part of the Arabian Gulf and led to the progradation of continental clastic sediment across the southern Arabian Gulf region.

In the Early Jurassic progressive back-arc rifting in the eastern Mediterranean led to the development of a new northern passive margin. During the Jurassic era rift basins in Syria and south-east Yemen were active, and intrashelf basins in the south-western Arabian Gulf, eastern Saudi Arabia, and southern Iraq–Kuwait were well developed. These intrashelf basins formed the main source and reservoir rocks for the large reserves of oil in Arabia.

The Neo-Tethys spreading ridge continued migrating north-eastwards and progressively subducted under Eurasia. This Early Cretaceous sedimentation is dominated by a carbonate sequence related to major flooding of the Arabian Peninsula.

The onset of Late Cretaceous thrusting in the Oman Mountains marks a distinctive change in the pattern of the basin subsidence, and represents the main phase of thrust tectonics in south-east Arabia. The Late Cretaceous thrusting during the closure of the Neo-Tethys is directly related to the change in plate translation (from a south-west to north-east direction) in response to the opening of the South Atlantic Ocean.

Significant and widespread breaks in sedimentation occurred across the Arabian Gulf region in Late Cenomanian and Turonian times. These stratigraphic breaks correspond to major tectonic events in eastern Arabia. In the Late Cretaceous the obduction of a series of ophiolites along the Neo-Tethys margin led to reactivation of some of the basement features in Arabia and localized basin inversions in Syria.

The end of the Cretaceous was marked by a regional unconformity which resulted in the Late Maastrichtian and Danian sediments being absent over most of the Arabian Plate. During the Early Cenozoic subduction of the Neo-Tethys beneath the Sanandaj–Sirjan Terrane along the northern margin of the Neo-Tethys caused the ocean basin to close, a process assisted by the initiation of rifting and opening of the Red Sea. The Late Paleocene–Eocene consists of predominantly shallow marine carbonates and evaporites. The onset of collision between the Arabian and Eurasian continents, which commenced in the Late Eocene, initiated the Zagros orogeny by suturing the Arabian and Eurasian plates. The collision created the Zagros foreland basin on the outer edge of the north-eastern Arabian shelf margin during the final closure of the Neo-Tethys.

Coeval with the Late Alpine Orogeny in Europe, the Neogene was a time of maximum compression between Arabia and Asia. During this period, the Arabian

Plate began to separate from Africa, the Gulf of Aden opened, and the Dead Sea transform fault acted as a complex sinistral strike-slip fault.

Economic Geology

The Precambrian basement rocks of Arabia contain a range of mineral deposits, notably gold, anciently mined in the Yemen. Important phosphate deposits occur in a belt that extends from Syria, through Jordan, and along the Palaeo-Tethyan shoreline into Egypt and across the Sahara to the Atlantic Ocean. These are found in Late Cretaceous and Eocene limestone (*see* **Sedimentary Rocks:** Phosphates). A range of evaporite minerals occurs in the Cambrian Hormuz salt, and in modern sabkhas of the Gulf coast, where magnesite occurs in significant quantities (*see* **Sedimentary Rocks:** Evaporites).

It is, of course, for its petroleum reserves that Arabia and the Gulf are best known. This basin is the largest petroleum province in the world. Various figures have been calculated for its reserves, but it is generally agreed (figures vary in time and with author) that it contains some 700×10^9 BOE (barrels of oil equivalent). This is about 40% of the world's petroleum reserves (*see* **Petroleum Geology:** Reserves). Petroleum is produced from reservoirs throughout the stratigraphic column. As a generalizations deep gas comes from Palaeozoic sandstones, and oil from Mesozoic carbonates. Petroleum source beds range from Infracambrian to Cretaceous. Petroleum entrapment is largely stratigraphic, occurring in truncated or onlapped sandstone, and in reefal or shoal carbonate, within which diagenesis in general, and dolomitization in particular, has often played an important part in controlling reservoir quality.

See Also

Africa: Pan-African Orogeny; North African Phanerozoic; Rift Valley. **Petroleum Geology:** Reserves. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Deserts; Reefs ('Build-Ups'). **Sedimentary Rocks:** Dolomites; Limestones; Phosphates. **Shields**.

Further Reading

Al Hajri S and Owens B (2000) Subsurface palynostratigraphy of the Palaeozoic of Saudi Arabia, joint study between Saudi Aramco and CIMP, in stratigraphic palynology of the paleozoic of Saudi Arabia, Special Publication 1. *GeoArabia*.

Al Jallal IA (1994) *The Khuff Formation, its regional reservoir potential in Saudi Arabia and other Gulf countries; depositional and stratigraphic approach,* GEO 1994, Middle East Petroleum Geosciences Conference, vol. 1, pp. 103 119. Bahrain: Gulf Petrolink.

Al Laboun A (1993) *Lexicon of the Paleozoic and Lower Mesozoic of Saudi Arabia*. Riyadh, Saudi Arabia: Al Hudhud Publishers.

Alshahran AS and Kendall CGStC (1986) Paleozoic to Early Mesozoic facies, depositional setting and hydrocarbon habitat in the Middle East: an overview and some play concepts. *American Association Petroleum Geologists Bulletin* 70: 977 1002.

Alsharhan AS and Nairn AEM (eds.) (1997) *Sedimentary Basins and Petroleum Geology of the Middle East*. Elsevier.

Alsharhan AS, Nairn AEM, and Mohammed AA (1993) Late Palaeozoic glacial sediments of the southern Arabian Peninsula: Their lithofacies and hydrocarbon potential. *Marine and Petroleum Geology* 10: 1 78.

Alsharhan AS and Scott RW (2000) Hydrocarbon potential of Mesozoic carbonate platform basin systems. In: Alsharhan AS and Scott RW (eds.) *Middle East Models of Jurassic/Cretaceous Carbonate Systems,* Special Publication 69, pp. 335 358. Society for Sedimentary Geology.

Beydoun ZR (1991) Arabian Plate hydrocarbon, geology and potential A plate tectonic approach. *American Association of Petroleum Geologists, Studies in Geology* 33: 77.

Doughty C (1888) *Travels in Arabia Deserta* (includes geological map and paper). London: Jonathan Cape.

Glennie KW, Boeuf MGA, Hughes Clarke MW, *et al.* (1974) Geology of the Oman Mountain. *Royal Geology and Mining Society (Netherlands) Transactions* 31: 423.

Konert G, Afifi AM, Al Hajri SA, and Droste HJ (2001) Paleozoic stratigraphy and hydrocarbon habitat of the Arabian Plate. *GeoArabia* 6(3): 407 441.

Murris RJ (1980) Middle East: Stratigraphic evolution and oil habitat. *American Association of Petroleum Geologists Bulletin* 64: 597 618.

Murris RJ (1981) Middle East: Stratigraphic evolution and oil habitat. *Geologie en Mijnbouw* 60: 467 480.

Sharland PR, Archer R, Casey DM, *et al.* (2001) Arabian Plate sequence stratigraphy, Special Publication 2. *GeoArabia*.

Stump TE and Van der Eem JG (1994) *Overview of the stratigraphy, depositional environments and periods of deformation of the Wajid outcrop belt, southwestern Saudi Arabia,* GEO 1994, Middle East Petroleum Geosciences Conference, pp. 867 876. Bahrain: Gulf Petrolink.

Vaslet D, Manivit J, Le Nindre YM, Brosse JM, Fourniquet J, and Delfour J. (1983). Explanatory notes to the geological maps of Wadi ar Rayn quadrangle.

Ziegler MA (2001) Late Permian to Holocene paleofacies evolution of the Arabian Plate and its hydrocarbon occurrences. *GeoArabia* 6(3): 445 503.

ARGENTINA

V A Ramos, Universidad de Buenos Aires, Buenos Aires, Argentina

Introduction

Argentina has a complete stratigraphic record of the Late Precambrian and Phanerozoic history of the southern sector of South America, and a large variety of present and past tectonic settings. As a whole, it can be divided into an Andean orogenic region and a stable basement area that is associated with one of the largest offshore continental platforms of the world. In order to understand its geology, a brief description of the present geological setting provides the key to the comprehensive record of its geological history. The Andes (see Figure 1) have controlled the distribution of volcanic and plutonic terrains, the sedimentary basins, and the structure of the different fold-and-thrust belts since the early Mesozoic. The opening of the Weddell Sea and the South Atlantic Ocean dominated the Mesozoic tectonics with a large period of extension. The Palaeozoic rocks were the result of an intricate interaction of sialic terranes or microcontinents that collided against the western Gondwana margin: ophiolitic belts, magmatic rocks, and different basins are testimony of these processes. Isolated patches of Late Precambrian rocks are widespread throughout most of Argentina, indicating the complex relationships of cratons, mobile belts, and shear zones.

Geological Setting

A large variety of geological provinces exist, from the active margin related to subduction of the oceanic Nazca and Antarctic plates beneath South America in the west, to the extra-Andean stable regions and large continental platform of the passive Atlantic margin in the east. The Andean chain in Argentina has the highest peaks of the western hemisphere; the Aconcagua, for example, a Miocene volcanic massif, rises 6965 m above sea-level (a.s.l.) (Figure 2). One of the main geological features of Argentina is the segmented nature of the Andes and the associated foreland regions, as depicted in Figures 3 and 4. This segmentation is the result of subhorizontal subduction in the central segment, as represented by the Pampean flat-slab region in the foreland. This contrasts with normal subduction in the northern and southern segments, which have Wadati–Benioff zones with a dip of about 30°.

Northern Segment

Four distinctive geological provinces characterize the northern segment. The first is the Cordillera Occidental, located along the international border with Chile; it is an active volcanic belt with large stratovolcanoes, several of them active, including the Lascar volcano, and has large ignimbritic fields of Late Cenozoic age. The second is the Puna, a high plain at 3750 m elevation; this southern extension of the Altiplano plateau is mainly composed of Miocene volcanic rocks and partially cannibalized foreland basins filled with Palaeogene to Miocene synorogenic deposits unconformably lying on a low-grade Palaeozoic basement. The third is the Cordillera Oriental; this thrust belt rises up to 6200 m high and has tectonic slices of Early Palaeozoic sedimentary and plutonic rocks. The fourth is the Sub-Andean system, consisting of middle- and low-elevation ranges ranging from 1500 to less than 1000 m high; the Sub-Andean province includes a series of large folds that expose a complete sequence of Silurian to Permian sedimentary rocks thrust on Cenozoic synorogenic deposits. The active orogenic front is located along the eastern foothills, 750 km east of the present oceanic trench. The Sub-Andean belt changes towards the south into the Santa Bárbara System, which is a range characterized by tectonic inversion of previous Cretaceous rifts (Figure 3).

Central Segment

The main feature of the central segment is the lack of active volcanoes along the Andean chain. The central segment includes the Cordillera Principal, the highest mountains in the Andes; the mountains are composed of Miocene uplifted volcanic rocks unconformably overlying marine Cretaceous and Jurassic carbonates, sandstones, and shales, interfingered with andesitic and dacitic lavas and tuffs. These rocks are in tectonic contact with the Cordillera Frontal, a series of north-trending mountains composed of Late Palaeozoic 5000- to >6000-m-high volcanics and granitoids that are overlain by continental Miocene to Pliocene synorogenic deposits. A wide tectonic valley, the Calingasta–Uspallata tectonic trench, separates the Cordillera Frontal from the Precordillera. The Precordillera fold-and-thrust belt, a 5000- to 3000-m-high range, is composed of Early Palaeozoic carbonate platform

Figure 1 Digital elevation model of Argentina. The extension and importance of the distinct cordilleras can be seen in the different segments of the Andean chain; note also the presence of plains, volcanic fields, basement uplifts, and large topographic massifs in the extra Andean regions. Image courtesy of United States Geological Survey.

deposits, clastic foreland basin Silurian and Devonian deposits, and Late Palaeozoic sedimentary rocks, including glacial deposits. The broken foreland is characterized by the crystalline basement uplifts of the Sierras Pampeanas, a block-mountain system formed during the shallowing of the subducted slab during the past 12 million to 9 million years. Isolated patches of Miocene to Pleistocene stratovolcanoes and volcanic

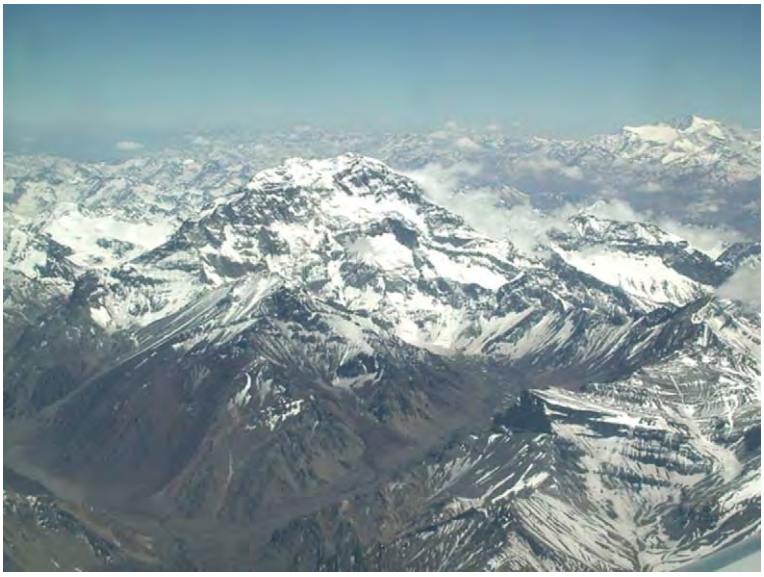


Figure 2 View of Mount Aconcagua, the highest peak in the western hemisphere (6965 m), characterized by Miocene volcanic rocks, unconformably overlying folded Mesozoic rocks.

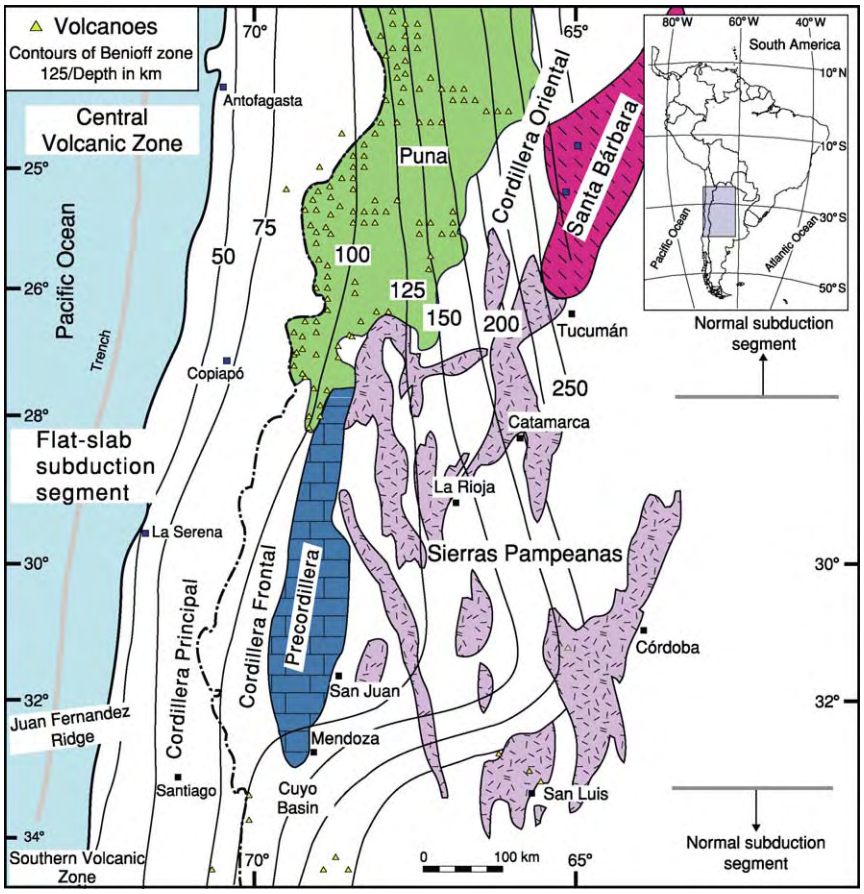


Figure 3 The different geological provinces of Argentina are controlled by the segmented nature of the Nazca oceanic slab. The northern and southern segments have a normal 30° dipping subduction zone that contrasts with the Pampean flat slab. Contours indicate depth (in kilometres) relative to the oceanic slab; triangles are Late Cenozoic volcanoes. The dash dot line represents the border between Argentina and Chile. From Ramos, V.A., Cristalliniy, E., Pérez, D.J. 2002. The Pampean flat slab of the Central Andes. Journal of South American Earth Sciences 15(1): 59 78.

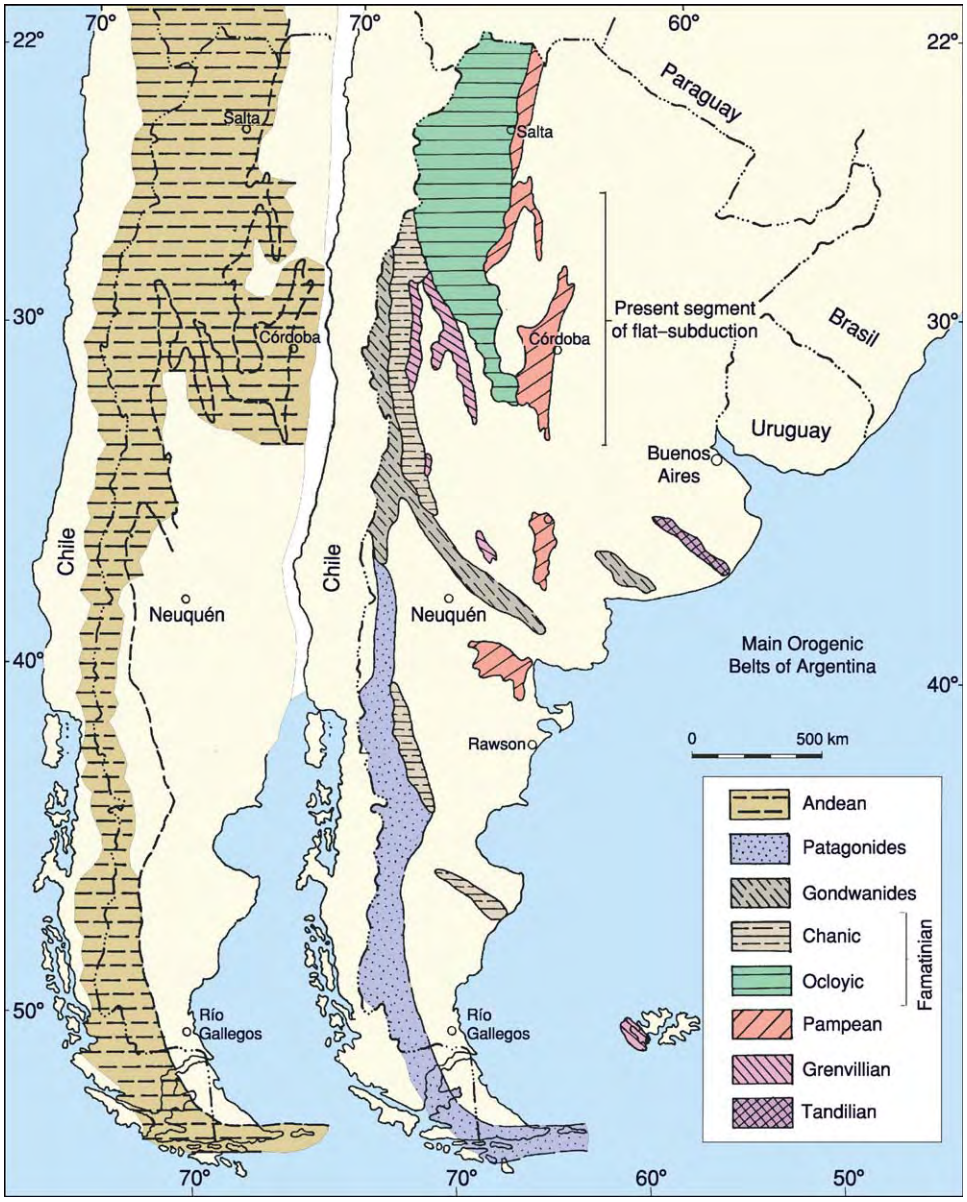


Figure 4 Main orogenic belts in the Andes of Argentina and Chile and adjacent regions. From the Tandilian to the Andean, the corresponding geological ages are Early Proterozoic, Middle Proterozoic, Late Proterozoic Early Cambrian, Early Palaeozoic (for both Chanic and Ocloyic), Late Palaeozoic, Late Mesozoic, and Cenozoic.

domes, with lavas and ignimbritic flows, spread over the Precordillera and the western part of the Sierras Pampeanas as a result of this flattening subduction process.

Southern Segment

South of the Tupungato stratovolcano, at about 33° 30′ S latitude, active volcanism resumes along the Cordillera Principal and continues down to the southernmost Andes. The main Cordillera is composed of a thin to thick-skinned fold-and-thrust belt formed by Jurassic and Cretaceous marine sediments derived from the Pacific Ocean. The altitudes, with the only exception being the active volcanic edifices, do not exceed 4000–3000 m and are even lower further to the south. The Cordillera Frontal, the Precordillera, and the Sierras Pampeanas disappear a few kilometres south of the flat subduction segment. The Palaeozoic rocks in the foothills are mildly uplifted in a series of basement blocks, including the San Rafael Block, yet still preserve the Late Palaeozoic peneplain. The foreland region has a Late Cenozoic large volcanic plateau that was formed by a few stratovolcanoes of andesitic to dacitic composition, and

large basaltic fields. This retroarc volcanism, developed between 35° and 38° S latitudes, is related to the late Cenozoic steepening of the subduction zone. Between 36° and 39° S, a large Mesozoic embayment exposes the Early Cretaceous and Jurassic marine clastic and carbonate sediments of the Neuquén retroarc basin, which are unconformably overlain by continental molasse of Late Cretaceous age.

Southernmost Segment

South of the Aysén Triple Junction between the Nazca, Antarctic, and South America plates, at about 46° 30′ S, the Patagonian Cordillera developed along the border between Argentina and Chile. The main characteristic of this segment, the continuous elongated batholith along the axis of the Patagonian Cordillera, is exposed for 2000 km and is thrust on Mesozoic volcanic and sedimentary rocks. The foothills are characterized by Late Cretaceous molasse deposits and synorogenic deposits of the Palaeogene and Neogene foreland basins. South of 52° S latitude, an oroclinal bend is characteristic of the Fueguian Cordillera in the southernmost tip of the Andes; it is composed of Jurassic volcanics and Cretaceous deposits that were heavily deformed during the Late Cretaceous closure of the Rocas Verdes marginal basin.

Andean Region

Volcanic Rocks

The Andes of Chile and Argentina are characterized by thick volcanic sequences developed in the past 26 million to 28 million years. These volcanic rocks are composed of andesite and dacite lavas interbedded with pyroclastic rocks in stratovolcanoes and volcanic domes, interfingered with large ignimbritic fields. The peaks of volcanic activity in the main arc are represented by thick (up to several kilometres) packages of highly differentiated rocks that constitute the Central Volcanic Zone of the Andes, as defined many years ago in southern Peru, Bolivia, and northern Chile and Argentina. The Ojos del Salado, the highest active volcano in the world (6800 m a.s.l.), is an example of a thick pile of andesites, dacites, and basalts that developed during Late Cenozoic times. These highly evolved rocks were the result of large eruptions through a thick crust (Figure 5); the eruptions contaminated the magmas by assimilation and differentiation through crystal fractionation. The eruptions alternated in Late Miocene and Pliocene times with rhyolitic tuffs and ignimbrites, which together make up large volcanic plateaux. These acidic plateaux are interpreted as the result of the lithospheric and crustal delamination that followed a period of steepening of the subduction zone after subhorizontal subduction north of 26° S.

The volcanic rocks south of this latitude record an opposite trend. Large stratovolcanoes developed during most of the Palaeogene along the border with Chile and they extended into Argentina in Early Neogene times. The large volcanic fields are younger towards the east and they gradually migrated to the foreland along corridors. The volcanic activity decreased in the Late Miocene and ceased along the main axis between 8 and 6 million years ago. Isolated high-K volcanoes, such as the Sierra del Morro in San Luis, are located 750 km away from the trench. South of 38° S

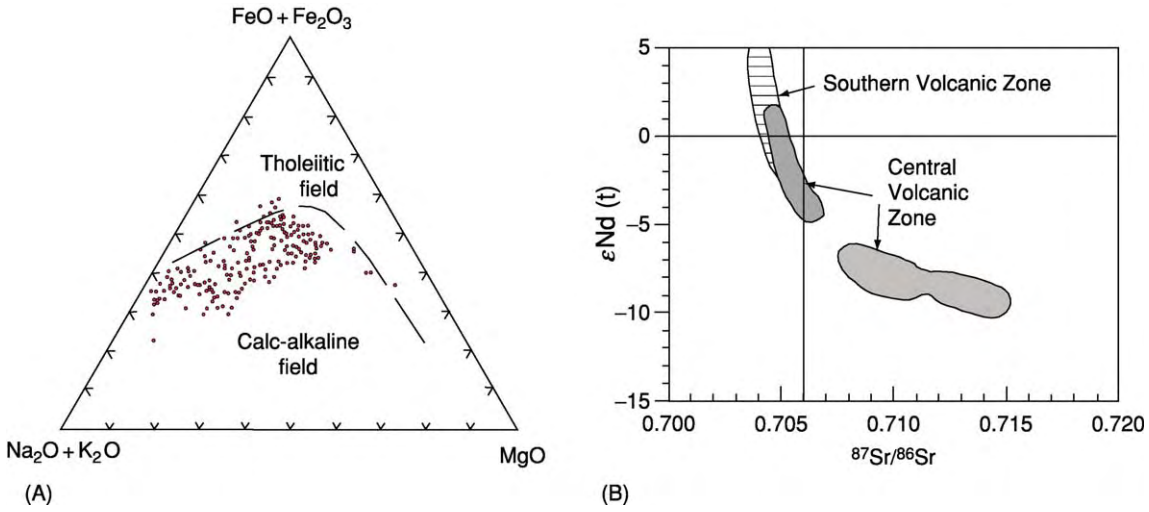


Figure 5 (A) Main geochemical characteristics of the subduction related volcanic rocks of Argentina, with a continuous calc alkaline trend. (B) Contrasting neodymium (Nd) and strontium (Sr) isotopic compositions of the central and southern volcanic rocks. Note the more primitive composition of the southern volcanic zone rocks, in comparison with rocks from the central volcanic zone.

and down to 46° 30′ S, large isolated patches of volcanic rocks represented the main arc along the Andes. These volcanic rocks, termed the Southern Volcanic Zone of the Andes, have two distinct segments. The northern one corresponds to andesites, dacites, and minor basalts, which erupted through large stratovolcanoes, such as the Tupungato, Marmolejo, and San José. The southern segment is represented by basalts, basandesites, and scarce dacites and rhyolites, which erupted through several Late Cenozoic volcanoes. The difference in composition is closely related to the thickness of the continental crust along the Andean chain; this varies from over 60 km north of 33° S down to 42 km at 46° S. South of 46° 30′ S there is a volcanic gap in arc volcanism; this is indicated by isolated adakite outcrops such as Cerro Pampa and Puesto Nuevo, which are related to exceptional oceanic slab melting of young oceanic crust prior to the subduction of an oceanic seismic ridge. Further south, from 48° S to 52° S, there are five small isolated volcanoes formed of basalts and basandesites; these (including the Lautaro, Aguilera, and Monte Cook volcanoes) constitute the Austral Volcanic Zone of the Andes. All of these rocks have an adakitic signature that indicates small partial melting of the oceanic slab superimposed onto the asthenospheric wedge magmas.

Fold-and-Thrust Belts and Their Synorogenic Deposits

The eastward migration of volcanic activity was coeval with the deformation, uplift, and development of synorogenic deposits. The eastern slope of the Andean chain records a series of discontinuous foreland basins, with thick sequences of continental deposits up to more than 10 000 m thick. The fold-and-thrust belts vary in style and kinematics from north to south and were controlled by the segmented nature of the subduction zones and the previous Palaeozoic and Mesozoic geological history.

Sub-Andean fold-and-thrust Belt The thin-skinned sub-Andean Belt, which moved on Late Ordovician, Silurian, and Devonian shales, has open folds and thrusts. The area was covered by marine deposits about 13.5 million years ago, showing that deformation in the sub-Andean Belt is later than Middle Miocene. The foreland basin deposits reach over 6000 metres. The syngrowth strata of Late Pliocene and Pleistocene age, as well as Global Positioning System (GPS) data and earthquake locations, indicate active tectonics in the thrust front. The Cenozoic orogenic shortening rate was of the order of 6.7–6.9 mm year ^{1}, but probably with periods of higher activity during Late Miocene and Plio-Quaternary times.

Santa Bárbara fold-and-thrust Belt The Santa Bárbara Belt is controlled by tectonic inversion of basement faults, and therefore shortening is less important than in the Sub-Andean Belt. Fault vergence is towards the west, controlled by the polarity of Mesozoic extension. Sag and synrift sequences of Cretaceous and Palaeogene age have marine deposits of Maastrichtian–Palaeogene age with a Pacific provenance. Most of the synorogenic deposits have been preserved in intermontaneous basins, up to 3-4 km thick.

Sierras Pampeanas Belt The Sierras Pampeanas Belt is a region of uplifted basement in which crystalline basement of Precambrian–Early Palaeozoic age is widely exposed. The reverse faults that bound the basement blocks have a dominant west vergence and are controlled by previous crustal sutures inherited from the Early Palaeozoic tectonics. This sector coincides with the subhorizontal subduction segment. The synorogenic deposits record two different stages. The older stage was an open-to-the-east large foreland basin associated with uplift and shortening in the main Andes. The second stage, which is characterized by a broken foreland, has the largest subsidence. Consequently, basins with more than 10 000 m of Late Miocene–Pliocene deposits have accumulated in a continental environment.

Cordillera Principal Belt The Cordillera Principal Belt comprises (1) a northern sector (30°–32° S latitude) characterized by the tectonic inversion of basement blocks where Late Palaeozoic–Triassic volcanic and plutonic rocks are exposed, (2) a central sector (32°–33° 30′ S latitude) where thin-skinned deformations of Jurassic and Early Cretaceous marine deposits are widely exposed, and (3) a southern sector (33° 30′–37° S latitude) where inversion tectonics occurred again. All of these belts were deformed in Late Cenozoic times, resulting in accumulation in the foothills of up to 3000-m-thick synorogenic deposits of continental sequences. The southern sector is linked to the Neuquén Embayment, a wide retroarc basin that developed during Early Mesozoic times and which covered most of the adjacent extra-Andean platform.

Patagonia fold-and-thrust Belt The Patagonia Belt also has two different segments. The northern segment (37°–46° S latitude) has a mild basement uplift with only local Cenozoic depocentres, because the main phase of uplift was produced in Cretaceous times. South of 46° 30′ S latitude, thick foreland basin sequences constitute the Austral or Magallanes retroarc Basin. The uplift and subsequent deposition

of the synorogenic deposits were controlled by collision of several segments of the Chile active spreading ridge from south to north in the past 14 million years, a collision that is still taking place at 46° 30′ S latitude.

Fueguian fold-and-thrust Belt The Fueguian Belt was controlled by the Late Cretaceous closure of the Rocas Verdes marginal basin, a back-arc basin of Late Jurassic–Early Cretaceous age. This segment is the only one that records the metamorphism and obduction of ophiolitic rocks during the Andean deformation. The basin development is linked to the formation of the Austral Basin; as a result of the northern vergence of the Fueguian fold-and-thrust Belt, the synorogenic deposits fill a series of east- and west-trending successor basins.

Stable Platform

The extra-Andean region consists of large Late Cenozoic plains, known as the Pampas, where thin Tertiary and Quaternary sediments cover the basement. These sequences preserve the first Atlantic transgressions, which occurred in Maastrichtian–Paleocene times, and several others, which were produced during the different episodes of high sea-level in the South Atlantic. Most of these sediments are capped by thick sequences of loessic deposits. The stable platform is characterized by several geological provinces that encompass uplifts and basinal areas.

Uplift Areas

Several old mountain systems emerge from the extra-Andean plains; from north to south, these regions consist of the Tandilia and Ventania systems and the Somuncurá and Deseado massifs.

The Tandilia System is a basement uplift produced during the Cretaceous at the time of opening of the South Atlantic Ocean, which exposed remnants of an Archaean basement (Figure 4), deformed by the Early Proterozoic Tandilian Orogen of Transamazonian age (~2000 Ma). The basement is covered by thin sequences of quartzites and carbonates of Late Proterozoic–Early Palaeozoic age. The Ventania System, a thin-skinned fold belt 600 km south of Buenos Aires, developed as a consequence of the Patagonia collision against Gondwana during Permian times. Thick sequences of clastic platform marine deposits, capped by the Sauce Grande Tillites in the Late Carboniferous, were thrust with a north vergence over the Gondwana margin. These rocks are a western extension of the Cape Fold Belt of South Africa, and the correlation between its glacial deposits with the Dwyka Tillite was one of the first geological arguments to support the existence of continental drift in the early twentieth century.

The Somuncurá basement massif is preserved in the northern sector of Patagonia (Figure 6). A thermally uplifted region during the development of a transient hotspot 27 Ma, the Somuncurá Massif is bounded by two east- and west-trending aulacogenic basins. The resulting thick alkaline basalts constitute the large Somuncurá shield volcano. These basalts are associated with other alkaline plugs and volcanic flows of Miocene age. This volcanic cover is unconformably overlying a Late Triassic–Early Jurassic rhyolitic plateau, Late Palaeozoic arc granitoids, and a metamorphic basement. The Deseado Massif is another uplifted basement; it is located in south central Patagonia (Figure 6), where metamorphic rocks of Early Palaeozoic age are covered by a thick pile of rhyolitic lavas and ignimbrites of Middle Jurassic age. The plateau basalts that characterize this sector of Patagonia, which was formed in the past 14 million years, were controlled by the thin sequences of alkaline basalts that are related to asthenospheric slab-windows generated during the Chile seismic ridge collision.

Basinal Areas

The platform area of the extra-Andean region has several basins, most of them with an orthogonal trend to the continental margin. The main basins are from north to south.

Chaco-Paraná Basin The 5-km-thick Chaco-Paraná Basin records a complex history. This region first consisted of Early Palaeozoic redbeds that were part of an early foreland basin associated with the Pampean deformation in Late Proterozoic–Early Cambrian times; this was followed by Silurian and Devonian marine and continental deposits related to a reactivation of the deformation in the western sector. Late Palaeozoic glacial and marine sequences subsequently formed during a period of generalized extension. Cretaceous redbeds and tholeiitic basalts then formed during the development of the Parana plume, about 130 Ma, and, finally, a thin sequence of distal foreland deposits of Cenozoic age was associated with dynamic subsidence of the stable platform during Andean deformation.

Salado and Colorado Basins The Salado Basin is an aulacogenic basin that formed during the opening of the South Atlantic Ocean. Synrift deposits of Late Jurassic–Early Cretaceous age are covered by marine Late Cretaceous and Cenozoic deposits. The sedimentary sequence is coeval with the development of the continental platform in the Atlantic margin. The

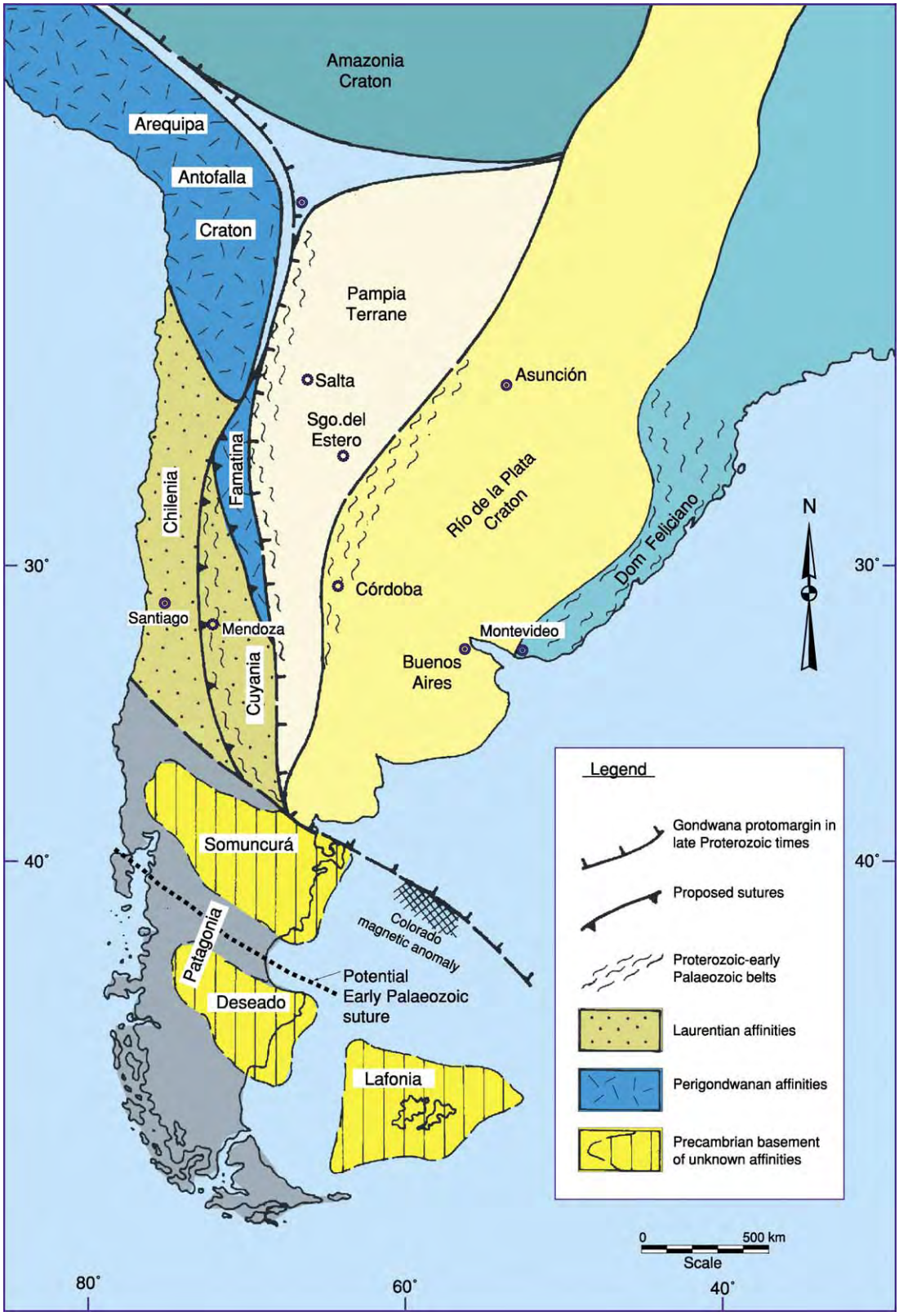


Figure 6 Generalized map of the Early Palaeozoic terranes and cratonic blocks of Argentina.

basin is over 5–6 km thick and opens towards the east. The Colorado Basin, another aulacogenic basin, developed along the suture between Patagonia and the south-western Gondwana continent. The timing and geometry of this basin are similar to those of the Salado Basin. Towards the west, the Colorado Basin is linked with the retroarc basin of the Neuquén embayment.

San Jorge Basin The intracratonic San Jorge Basin developed with an east–west trend during Middle Jurassic to Early Cretaceous times, during the opening of the Weddell Sea and the South Atlantic Ocean. During the Late Cretaceous, thick sequences of molasse deposits unconformably overlaid the previous rocks. Thick sequences of clastic deposits and tuff layers record the volcanic activity in the Andes during Cretaceous and Palaeogene times. Most of the basin was filled by continental deposits, with the first Atlantic marine transgression in Maastrichtian–Paleocene times.

Mesozoic Provinces

Important rift deposits that accumulated in north-western-trending depocentres are recorded in most of central and southern Argentina. The synrift deposits are of Late Triassic–Early Jurassic age, and consist of redbeds and alkaline basalts. A younger reactivation of these rift systems occurred in the Early Cretaceous, the time at which (~125 Ma) the first oceanic crust formed in the South Atlantic Ocean at these latitudes. The Cretaceous rifts are subparallel to the continental margin (Figure 7) and were controlled by the sutures of the Palaeozoic accreted terranes.

Two important rhyolitic geological provinces, the Choiyoi and the Chon Aike, were developed in the Andean and extra-Andean regions in Mesozoic times. Together these provinces constitute one of the largest rhyolitic provinces in the world, with an extension exceeding 1 000 000 km^2.

The Choiyoi Province

The Choiyoi province is characterized by an extensive rhyolitic plateau that developed between northern Chile, at 24° S latitude, and the northern part of Patagonia, at 40° S latitude. The region consists of thick piles of rhyolitic lavas, ignimbrites, and volcaniclastic deposits, occasionally interbedded with lacustrine limestones. These rocks were deposited in half-graben systems, with frequent thickness changes that indicate a synextensional deposition. The age of formation varies from 280 Ma in the north to 226 Ma in the southern sector. Volcanic activity had a climax 265 Ma, when ash-flow tuffs fell over most of the extra-Andean part of Argentina and southern Brazil, even reaching South Africa. This generalized extension followed terrane accretion and deformation in Late Palaeozoic times, which was associated with the cessation of subduction along the Pacific margin.

The Chon Aike Province

The rhyolitic Chon Aike Province covers most of Patagonia south of 40° S and also extends into the Antarctic Peninsula and further south. It is composed of abundant rhyolitic lavas, ignimbrites, and tuffs, with minor amounts of alkaline basalts interbedded in the lower part of the succession. Seismic lines through the Austral Basin show the half-graben geometry of these deposits. The volcanic sequences exposed in the Somuncurá Massif are between 185 and 170 My old, whereas the rhyolites exposed in the Deseado Massif are younger, between 165 and 155 My old. The equivalent rocks in the Tierra del Fuego Island are about 140 My old. All together, these rocks again show a decrease in age to south of the acidic volcanism, and are associated with the early rifting that led to the opening of the Weddell Sea and the South Atlantic Ocean.

Palaeozoic Terranes

Most of the basement of Argentina was the result of collision and amalgamation of continental fragments formed during Palaeozoic times. Ophiolitic assemblages, magmatic arcs, and older orogens cover most of the present stable extra-Andean platform. Among the main continental fragments are the Cuyania, Pampia, Chilenia, and Patagonia terranes (Figure 6).

Cuyania

Previously known as the Precordillera Terrane, Cuyania is one of the best known allochthonous pieces that docked against the western Gondwana margin during Middle Ordovician times. It consists of a middle Proterozoic metamorphic basement, with amphibolites, marbles, and granitoids with a typical Grenvillian signature. Cambrian limestones bearing *Olenellus* trilobites denote a Laurentian provenance; this has been confirmed by palaeomagnetic studies. The Early Ordovician carbonate rocks are followed by black shales that are unconformably overlaid by conglomerates and sandstones, interbedded with typical Gondwanian glacial deposits of latest Ordovician age. The arrival of this terrane is documented by an active calc-alkaline series of granitoids that developed along the continental margin of Gondwana from Late Cambrian to Early Ordovician times. These rocks,

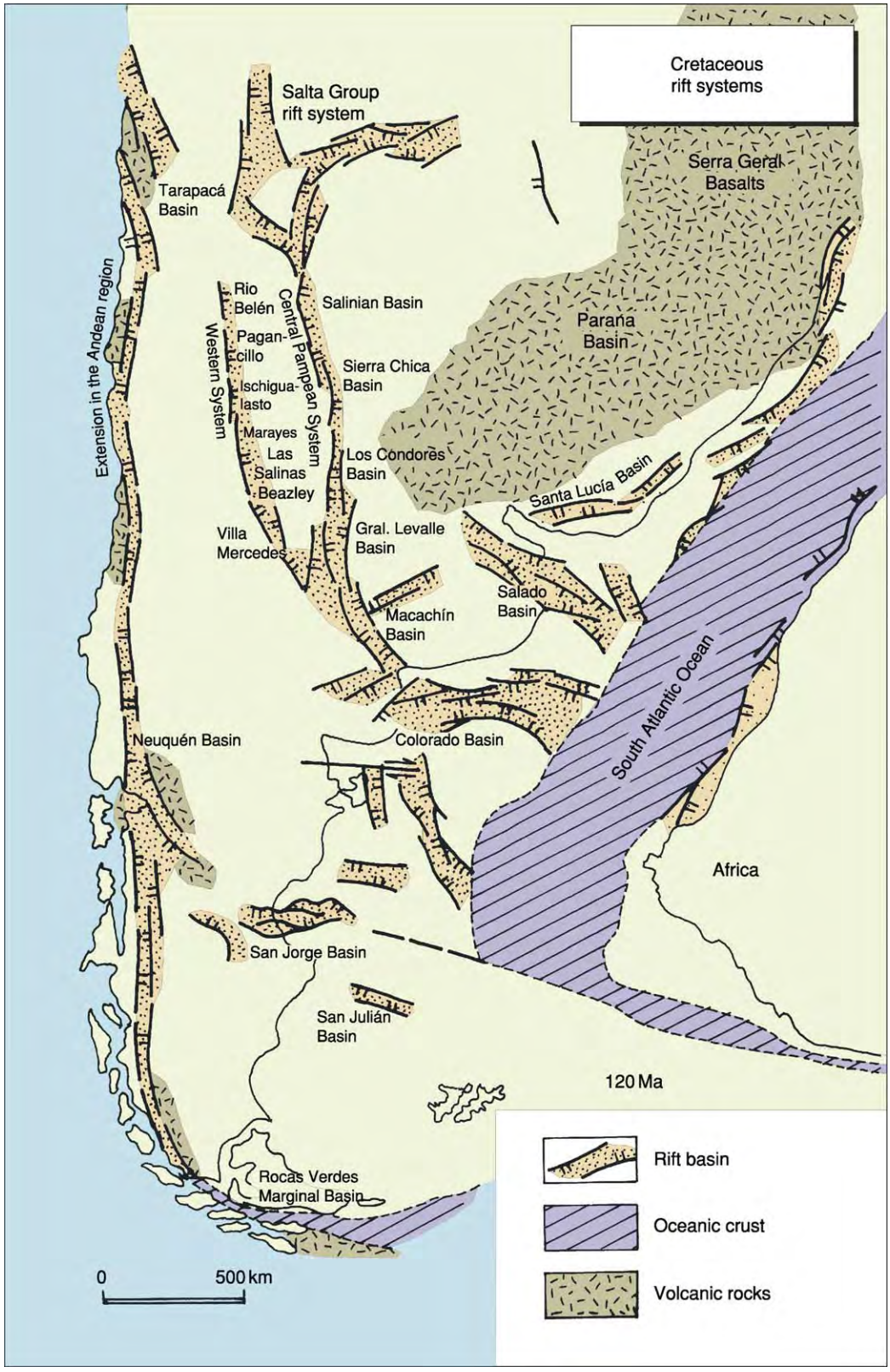


Figure 7 Mesozoic rift systems associated with the opening of the Atlantic margin. The intracratonic rifts were partially controlled by the sutures of the Palaeozoic accreted terranes. Along the main Andes, extensional basins were controlled by negative trench roll back velocity in the retroarc basins.

together with subsequent peraluminous granites, show a magmatic arc and postcollisional plutons along 1000 km of central Argentina. The Cuyania Terrane is bounded by two series of ophiolites, one in the western margin of the present Precordillera, and the other along the western margin of the Pampia Terrane.

Pampia

The Pampia cratonic block hosted an Early Palaeozoic magmatic arc, which consists of metamorphic and igneous rocks penetratively deformed during Late Proterozoic–Early Cambrian times, at the time of the Pampean Orogeny. A Late Proterozoic ophiolite bounds the eastern margin of the terrane. The protomargin of western Gondwana is exposed along the eastern Sierras Pampeanas, where relics of Late Proterozoic–Early Cambrian tonalites, granodiorites, and granites constituted the magmatic arc. The collision between the Pampia Terrane and the Gondwana protomargin is constrained by a series of rhyolites and other volcanic rocks of Early to Middle Cambrian age that unconformably overlie the deformed basement rocks. Low-grade metamorphic rocks, up to 5 km thick, constituted the synorogenic sequence preserved in the Cambrian foreland basin.

Chilenia

The amphibolites and schists with Grenvillian-age zircons, poorly exposed along the main Andes between Chile and Argentina, formed the basement of the Chilenia Terrane. Although the basement is well preserved along the Cordones del Plata and El Portillo, most of the metamorphic rocks are covered by several kilometres of sedimentary and volcanic rocks on the Principal and Frontal Cordilleras. An ophiolitic belt of Ordovician to Devonian age separates this basement from the Cuyania Terrane. The docking of the rocks is constrained by the Early Carboniferous marine deposits that finally amalgamated the allochthonous terranes of central Argentina.

Patagonia

The Patagonian basement records a Precambrian history that started in the Middle Proterozoic, as detected in detrital zircon fractions from the Deseado Massif and from the western part of the Somuncurá Massif. The northern sector of the Somuncurá Massif has evidence of a superimposed Brasiliano deformation.

Crustal growth of Patagonia began in the Early Palaeozoic, when a magmatic arc was developed in the northern margin of the Somuncurá Massif. Age constraints for this magmatic belt indicate Middle to Late Ordovician to Late Carboniferous and Early Permian ages, showing that an active margin lasted until Early Permian times. The Deseado Massif was also the locus of Early Palaeozoic magmatism. Granitoids of Ordovician to Silurian age show a north-western-trending arc, parallel to penetrative deformation preserved in slates and schists of Late Precambrian age. It has been proposed that both basement blocks were independent terranes that amalgamated during the Early Palaeozoic to form a microcontinent that collided with Gondwana in Permian times, to form the Ventania fold-and-thrust belt.

See Also

Andes. **Antarctic**. **Brazil**. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**.

Further Reading

Mpodozis C and Ramos VA (1990) The Andes of Chile and Argentina. In: Ericksen GE, Cañas, Pinochet MT, and Reinemud JA (eds.) *Geology of the Andes and Its Relation to Hydrocarbon and Mineral Resources, Earth Sciences Series*, vol. 11, pp. 59 90. Menlo Park, CA: Circum Pacific Council for Energy and Mineral Resources.

Ramos VA (1994) Geology of South America. *Encyclopae dia Britannica* 27: 580 583.

Ramos VA and Alemán A (2000) Tectonic evolution of the Andes. In: Cordani UJ, Milani EJ, Thomaz Filho A, and Campos DA (eds.) *Tectonic Evolution of South America*, pp. 635 685. Río de Janeiro: 31st International Geological Congress.

Ramos VA and Keppie D (eds.) (1999) *Laurentia Gondwana Connections before Pangea. Special Paper 336*. Boulder, CO: Geological Society of America.

Ramos VA and McNulty B (eds.) (2002) Flat slab sub duction in the Andes. *Journal of South American Earth Sciences* 15(1): Special Issue.

Tankard AJ, Suárez SR, and Welsink HJ (eds.) (1995) *Pet roleum Basins of South America. American Association of Petroleum Geologists Memoir* 62. Tulsa, OK: American Association of Petroleum Geologists.

ASIA

Contents

Central

S G Lucas, New Mexico Museum of Natural History, Albuquerque, NM, USA

Introduction

Central Asia encompasses a land area of 5.9 million km^2 and includes the countries of Kazakstan, Turkmenistan, Uzbekistan, Kyrgyzstan, and Tajikistan (Figure 1), which became independent of the Soviet Union in 1991. Some geographers also include portions of neighbouring Afghanistan, Pakistan, western China, and Mongolia in Central Asia, but this review focuses on the regional geology of the five former Soviet republics. Most of Central Asia is steppe (arid grassland) and desert, which contrasts sharply with its mountain ranges, which are some of the world's longest (the Tien Shan is 2500 km long) and tallest (more than 7000 m).

Kazakstan

The largest state in Central Asia, Kazakstan, is half the size of the USA and has a land area of 2 717 300 km^2. It thus occupies about half the total area of Central Asia. Kazakstan is mostly steppe and desert, but in the south and southeast the Tien Shan, Zailiski Alatau, and Dzhungarian mountain ranges have peaks that exceed 7000 m in altitude (Figure 2). In contrast, the Karagie Depression in western Kazakstan in the Caspian Sea basin is 132 m below sea level, the lowest point on land after the Dead Sea. Large Kazak lakes include one of the largest in the world, Lake Balkash, with a surface area in excess of 17 000 km^2.

Physiographically, Kazakstan is a vast country south of the western Siberian lowlands, north of the Tien Shan Mountains, east of the Caspian Sea and west of the Altai ranges that run along its eastern borders with China and Mongolia. Kazakstan includes part of the South Caspian Basin (Depression), with a sedimentary fill of Jurassic–Neogene age as much as 20 km thick, one of the most important oil-producing regions in the world. It also encompasses the southern tip of the Ural Mountains, most of which are in Russia. The extensive Kazak deserts and uplands cover most of the country. Internally-drained basins with saline lakes extend across the northern part of Kazakstan–the Tengiz, Balkash and Alakol basins. A chain of deserts (including the Betpak Dala) north of the Tien Shan Mountains and south of Lake Balkash grade northward into steppes and forested uplands.

Precambrian and Palaeozoic rocks are exposed in all of the major mountain ranges of Kazakstan, as well as in the Mangyshlak Peninsula (which juts into the Caspian Sea) and in the North Caspian basin. Mesozoic and Cenozoic rocks are the widespread fill of various sedimentary basins, including the South Caspian, North Caspian, Turan, Turgay, Balkash, Alakol, Ily (Chu), and Zaysan basins.

The Precambrian rocks exposed in the Kazak mountain ranges consist of mostly schists, gneisses, marbles, and igneous rocks. Palaeozoic rocks are mostly of sedimentary origin, being Cambrian–Ordovician marine strata and volcanic rocks, Silurian–Early Carboniferous clastic sediments, volcanic rocks, cherts and ophiolites and Late Pennsylvanian–Permian nonmarine and marine sedimentary rocks. Notably, in the North Caspian Basin, salt domes have formed in a thick (4–5 km) succession of Permian salt. From the Tien Shan to north of Lake Balkash, the Kazak mountain ranges also expose remnants of a huge magmatic belt of mid-Carboniferous to Permian age, composed of lavas, tuffs, and ignimbrites, with an aggregate thickness of 5 to 6 km.

In Kazakstan, Early Triassic rocks are mostly tuffs and basalts, but, younger Mesozoic rocks are mostly nonmarine sedimentary rocks, though some marine strata of Mesozoic age are present.

The large lakes of Kazakstan have long geological histories. A good example is Lake Zaysan in northeastern Kazakstan. Today, Lake Zaysan is located

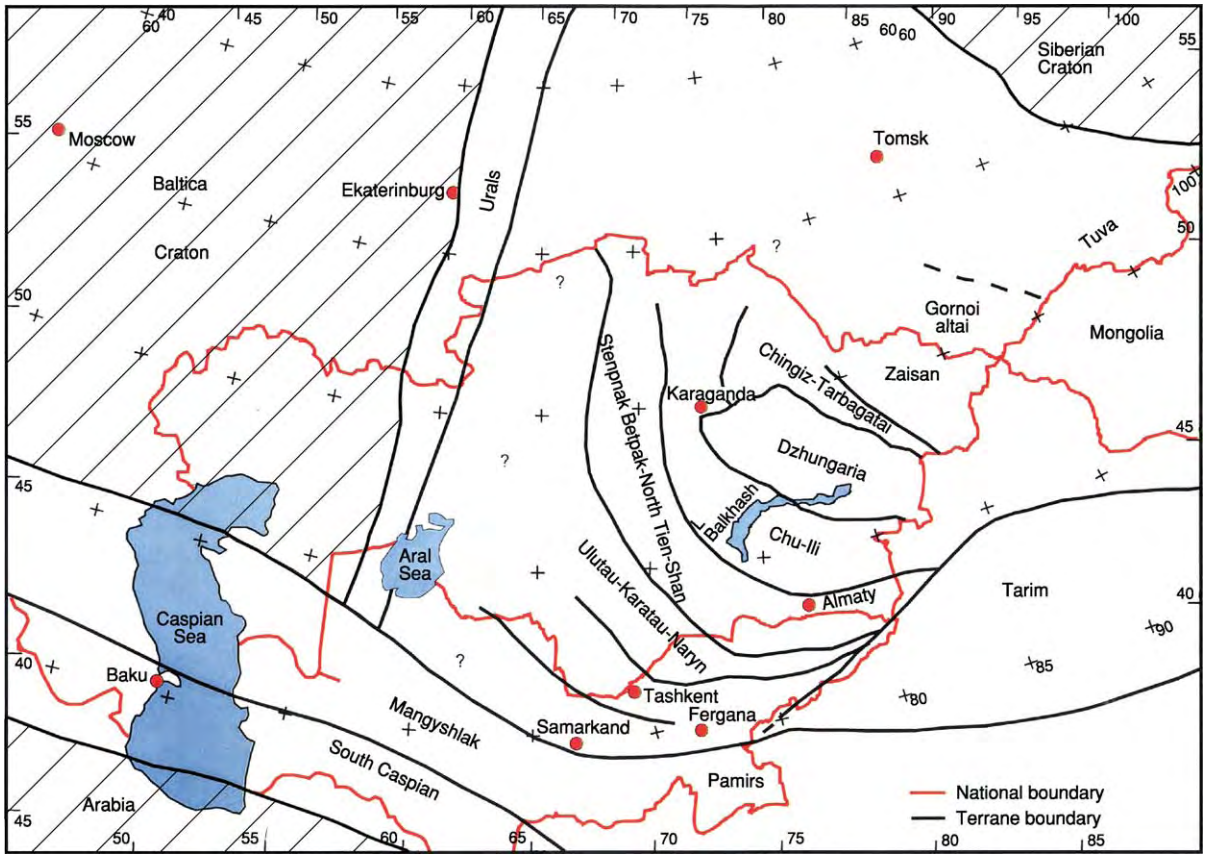


Figure 1 Map of Central Asia (modern topography) showing the boundaries of the Palaeozoic terranes. (With permission from Fortey and Cocks (2003).)

Figure 2 The Zaily Alatau in eastern Kazakstan is typical of the high mountain ranges of the Tien Shan Mountains that characterise much of Central Asia.

between the Altai and Targabatay ranges, is fed by the Cherny Itrysh River to the east and historically was about 100 km long, 32 km wide and 8 m deep. Beginning in the Late Cretaceous, a lake basin formed in the Zaysan basin and still exists today. The contracting and expanding lake basin thus has endured for more than 65 million years, and it has left nearly a kilometer of sedimentary basin fill, which is well exposed along the uplifted basin flanks (Figure 3A). These sedimentary rocks contain one of the most remarkable and continuous Cenozoic fossil records in Asia, including fossils of charophytes, land plants, ostracods, molluscs, fishes, turtles, crocodiles, birds, and mammals. Other Cenozoic lake basins in eastern Kazakstan also contain thick and fossiliferous sedimentary basin fills (Figure 3B).

Kazakstan is a country rich in oil, gas, and mineral resources. Oil and gas in Mesozoic and Palaeogene strata in the South Caspian basin are one of the world's great oil reserves. More than 300 coal deposits have been identified in Kazakstan, constituting proven reserves in excess of 170 billion tons. These coal deposits are mostly in Carboniferous rocks in the Kazak mountain ranges.

Kazakstan has a diverse and extensive mineral wealth, including important deposits of lead, zinc, iron, manganese, nickel, copper, tin, molybdenum, aluminum, and gold. Palaeozoic rocks are generally host to these metals.

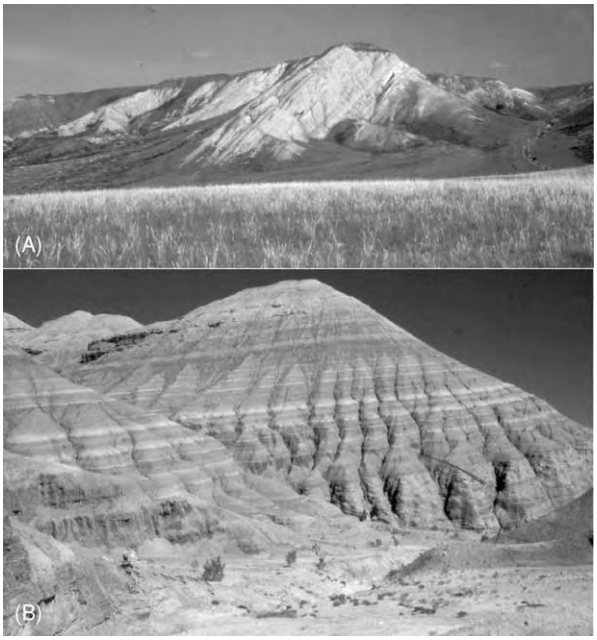


Figure 3 (A) Uplifted Palaeogene strata along the Kalmakpay River at the foot of the Targabatay Range in the Zaysan basin. (B) Miocene lake sediments near Aktau Mountain on the northern flank of the Ily basin in eastern Kazakstan.

Figure 4 This 10 metre long fishing boat was abandoned along the arid and receding shoreline of the Aral Sea.

Turkmenistan

Turkmenistan encompasses an area of 448 100 km^2, most of which (about 80%) is the vast Kara Kum (Black Sand) Desert. Elevations across this desert drop from 100–200 m in the east, to 28 m below sea level at the coastline of the Caspian Sea. Along the southwestern frontier of Turkmenistan, the Kopetdag and Balkhan Mountains reach elevations of up to 2912 m, and the western edge of the Tien Shan Mountains just reaches the eastern border of Turkmenistan. The vast Kara Kum is mostly blanketed by Quaternary deposits, and the mountain ranges in the southwestern part of Turkmenistan mostly expose Mesozoic (especially Triassic) rocks, as is true of the mountainous southeastern part of the country. The only significant outcrops of Palaeozoic strata, and of metamorphic and igneous rocks, are in the western Tien Shan Mountains.

Turkmenistan can thus be divided into two geologic regions, the Kara Kum platform and the southern mountainous (orogenic) belt. In the subsurface of the Kara Kum, extensive drilling has revealed Precambrian metamorphic rocks overlain by a thick, much deformed succession of Palaeozoic strata (mostly of Silurian, Devonian, and Carboniferous age), Palaeozoic granites and less deformed, mostly nonmarine, strata of Late Carboniferous, Permian, and Triassic age, with an aggregate thickness of 5 to 6 km. These strata are overlain by essentially flat-lying Jurassic, Cretaceous, and Cenozoic sedimentary rocks, that typically have an aggregate thickness of 1 to 2 km, though they reach thicknesses of 8 to 10 km in local sedimentary basins.

The mountainous belt of Turkmenistan is mostly made up of strongly deformed Mesozoic and Paleogene rocks that represent a wide array of environments, from nonmarine Triassic red beds, to Jurassic marine limestones, to Cretaceous shallow marine glauconitic claystones. The Tien Shan Mountains in Turkmenistan are a continuation of the same rocks better exposed in Uzbekistan.

Most of the topography in Turkmenistan is the result of Late Cenozoic tectonism. This tectonic activity continues today and is most evident in numerous earthquakes, some with Richter values greater than 7, especially in the southwestern part of the country. Pliocene–Pleistocene oscillations of the water level in the Caspian Sea drove water all the way out to the main part of the Turkmenistan plains. During the Soviet era, between the 1950s and 1980s, the Great Kara Kum Canal was built across southern Turkmenistan, from its eastern border to the Caspian Sea, a distance of some 1400 km. This canal, major irrigation projects along the Amu-Darya and Syr-Darya rivers, and the damming of other rivers helped aridify the Aral Sea region, turning it into a rapidly shrinking waterbody (**Figure 4**).

Turkmenistan is a major oil and gas producer, particularly in the area that adjoins the Caspian Sea. Deposits of potassium salt, halite, and sulphate in the Kara Kum platform have yielded important deposits of iodine and bromine. All economic production of these chemical elements is from rocks of Jurassic–Eocene age.

Uzbekistan

Uzbekistan encompasses an area of 447 400 km^2 and has a varied landscape from the western portion of the Tien Shan Mountains in the east and the Pamir-Altai ranges to the south-east, to the Kyzyl Kum ('Red Sand') Desert and shores of the Aral Sea to the northwest. The country is located between two of the great rivers of Central Asia, the Amu-Darya and the Syr-Darya, both of which have headwaters in the Tien Shan Mountains and flow to the Aral Sea.

The Tien Shan Mountains in Uzbekistan, as elsewhere, have a core of Palaeozoic (especially Silurian, Devonian, and Carboniferous age) sedimentary rocks and Precambrian–Early Palaeozoic metamorphic rocks that are intensely deformed. The metamorphic rocks include ophiolites, which are evidence of the ancient collapse of ocean basins as island arcs and other microcontinents were amalgamated to form what is now Central Asia. Granitic intrusions, that range in age from Cambrian to Triassic, are also present in the Uzbek portion of the Tien Shan. North of the Tien Shan, the bedrock of Uzbekistan is primarily Mesozoic and Cenozoic sedimentary rocks.

The Mesozoic rocks were mostly deposited in small, tectonically active basins in the southern and western parts of Uzbekistan. On the Ustyurt Plateau (the portion of Uzbekistan southwest of the Aral Sea) and in southern Uzbekistan, the Triassic rocks are mostly nonmarine sediments. Jurassic rocks, however, are much more widespread and are a mixture of marine and nonmarine sediments. The Cretaceous rocks are similarly widespread. The Lower Cretaceous strata are mostly of nonmarine origin, but some marine strata are present and are the host rocks for petroleum. The Upper Cretaceous rocks are mostly marine strata of limestone and gypsum; they contain important uranium deposits.

Palaeogene rocks in Uzbekistan are of both marine and nonmarine origin and contain reserves of oil and gas. The thickest cover of the low-lying regions of the country, however, is the Neogene strata, which are up to 6 km thick. Much of this thickness is alluvial sediments that were shed from the rising Tien Shan. Quaternary sediments are also products of the Tien Shan uplift and provide evidence of four pulses of uplift in the form of four prominent river terrace levels. To the north of the mountains, they are covered by a veneer of reddish orange sand dunes, 5 to 60 m thick, that make up the Kyzyl Kum Desert of northern Uzbekistan.

Oil, gas, and uranium deposits of Uzbekistan have just been mentioned. Mineral deposits are diverse and include copper, zinc, gold, and mercury, mostly in hydrothermal concentrations. Some Uzbek basaltic diatremes of Triassic age even yield diamonds.

Kyrgyzstan

A relatively small country, with an area of 198 500 km^2, Kyrgyztan is a mountainous land dominated by the Tien Shan Mountains, with peaks as high as 7439 m. Exposed rocks range in age from Archaean to Cenozoic. Much of the older bedrock of Kyrgyzstan (the bedrock core of the Tien Shan) is constructed primarily from a few Palaeozoic (primarily of Late Ordovician and Carboniferous age) island arcs that collided during the Late Palaeozoic–Early Mesozoic.

In Kyrgyzstan, Precambrian rocks have small outcrop areas in the mountains, which mostly expose Palaeozoic sedimentary rocks (especially of Ordovician, Silurian, Carboniferous, and Permian ages) as well as many Palaeozoic granitic intrusive rocks. Basin floors are covered with Cenozoic rocks, and Mesozoic strata are primarily exposed along uplifted basin margins.

The principal and largest basin in Kyrgyzstan is the Fergana Basin, which occupies much of the western part of the country and extends into eastern Uzbekistan. In the various Kyrgyz basins, Triassic and Jurassic strata are intensely folded, coal-bearing rocks that reach an impressive 5 km thick in the Fergana Basin. These rocks yield an extensive record of fossil plants (**Figure 5**). Overlying Cretaceous and Paleocene–Eocene rocks are a mixture of shallow marine, shoreline, and terrestrial sediments that are more than 2 km thick in the Fergana basin. Subsequent Cenozoic sediments are wholly nonmarine in origin, and are more than 4 km thick in the Fergana Basin. At various times, the Fergana Basin was filled with a lake surrounded by mountains several hundred meters high. Today, however, the largest lake basin in Kyrgyzstan is Issyk-Kul in the Tien Shan along the northern frontier with Kazakstan.

Tectonic activity in Kyrgyzstan during the Plio-Pleistocene produced the final uplift of the mountains and the deposition of thick continental sediments in the intermontane basins under cold and arid Pleistocene glacial conditions. This is when the current Tien Shan Mountains developed, and there was intense folding of the basinal Mesozoic and Cenozoic rocks. Kyrgyzstan is a highly seismic country where numerous large earthquakes occur, especially in the eastern Fergana Basin and in the Kyrgyz and Kungey mountain ranges along the northern frontier.

In Kyrgyzstan, oil and gas is produced in the Fergana Basin from Jurassic, Cretaceous, and Cenozoic

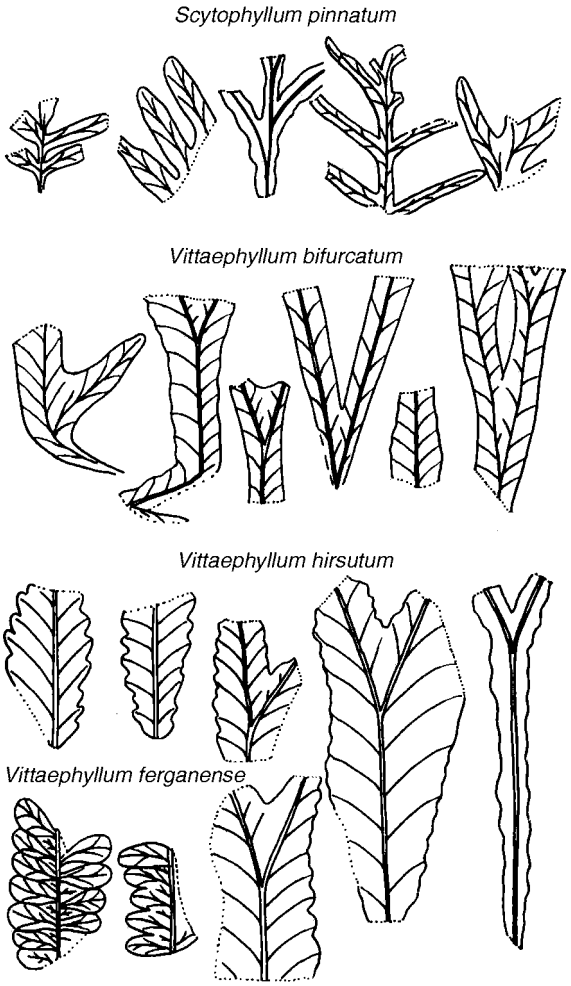


Figure 5 Triassic fern like plants from the Fergana Basin in Kyrgyzstan.

strata, and coal is also mined there. The Jurassic coal-bearing rocks in the Uzgen basin of western Kyrgyzstan are up to 5 km thick, and they are the most extensive exposures of Jurassic rocks in Central Asia. Minerals of economic importance are mercury, antimony, gold, and tin, mostly found in hydrothermal or skarn deposits.

Tajikistan

The smallest Central Asian country, with a land area of 143 100 km^2, more than 90% of Tajikistan is covered by mountains. The southern edge of the Tien Shan Mountains (also called the Trans-Altai Mountains) is along the northern edge of Tajikistan. Most notable are the Pamir Mountains, with glaciated peaks as high as 7495 m. The Pamirs (the name is from a Farsi word meaning 'high level valleys') are actually a high plateau with an average elevation of about 6000 m.

The eastern two-thirds of Tajikistan is the Trans-Altai and Pamir Mountains. The Pamirs are faulted belts of rocks that range from Late Palaeozoic in the north to Early Cenozoic in the south. In the western part of Tajikistan, where its capital city of Dushanbe is located, the Tajik depression is a basin with a sedimentary basin fill 6 km thick that consists of Jurassic salt deposits, Cretaceous–Oligocene marine strata, and younger Cenozoic nonmarine strata. These rocks are intensely folded, and the Jurassic salt includes a detachment surface that is being underthrust. Indeed, Tadjikistan (particularly under the Pamirs) is a place where the subduction of continental lithosphere is taking place. Because of this, it is a region of high seismicity, with earthquakes concentrated along the Pamirs and the southern margin of the Trans-Altai Mountains.

Geologic History

Central Asia is composed of a mosaic of ancient continental blocks which were microplates or independent terranes at various times in the Palaeozoic. Figure 1 shows a number of these terranes which are positioned between the old cratons of Baltica (most of northern and eastern Europe), Siberia (only part of modern Siberia) and the southern peri-Gondwanan terranes which include Mangyshlak, the South Caspian Terrane, and the various terranes which make up Arabia today. The many central Asian terranes shown together make up the tectonically complex 'Altaids': some have some Precambrian slivers in their cores, but most originated within island arc settings at various times in the Palaeozoic. The Lower Palaeozoic faunas contained within many of them are abundant and often distinctive, indicating liaisons and/or oceanic separations between their Lower Palaeozoic positions. These relatively small terranes progressively collided and amalgamated with each other, mostly during the Upper Palaeozoic during the formation of Pangaea. The results of the collisions are most evident in the bedrock cores.

In Central Asia, Late Palaeozoic compression, collision, and subduction was followed by Mesozoic extension, with widespread emplacement of granite batholiths and diabase dike intrusions, as well as the development of large sedimentary basins. During the Late Palaeozoic–Mesozoic, Central Asia was part of the Pangaean supercontinent, and as Pangaea split apart it remained part of southern Eurasia.

The Cenozoic geological history of Central Asia has been dominated by the collision of India with Eurasia during the Paleogene. This collision reactivated old structures along the Pangaean collisional zones. The

northward propagation of deformation from the collision saw India thrust under Tibet, subsidence of many of the Cenozoic sedimentary basins of Central Asia, rise of the southern and then the northern Tien Shan, and also the deformation of more northerly uplifts. This deformation thus produced the dramatic basins and ranges that characterize the Central Asian landscape today.

See Also

China and Mongolia. **Europe:** The Urals. **Pangaea**. **Plate Tectonics**. **Russia**.

Further Reading

Burtman VS (1980) Faults of Middle Asia. *American Journal of Science* 280: 725 744.

Burtman VS (1997) Kyrgyz Republic. In: Moores EM and Fairbridge RW (eds.) *Encyclopedia of European and Asian Regional Geology*, pp. 483 492. London: Chapman and Hall.

Dobruskina IA (1995) Keuper (Triassic) flora from Middle Asia (Madygen, southern Fergana). *New Mexico Museum of Natural History & Science*, Bulletin 5: 49p.

Fortey RA and Cocks LRM (2003) Palaeontological evidence bearing on global Ordovician Silurian reconstructions. *Earth Science Reviews* 61: 245 307.

Hendrix MS and Davis GA (eds.) (2001) Paleozoic and Mesozoic Tectonic Evolution of Central and Eastern Asia. *Geological Society of America Memoir* 194: 447p.

Leith RW (1982) Rock assemblages in Central Asia and the evolution of the southern Asian margin. *Tectonics* 1: 303 318.

Lucas SG, Emry RJ, Chkhikvadze V, *et al.* (2000) Upper Cretaceous Cenozoic lacustrine deposits of the Zaysan basin, eastern Kazakstan. In: Gierlowski Kordesch EH and Kelts KR (eds.) *Lake Basins Through Time and Space*, 46, pp. 335 340. AAPG Studies in Geology.

Moores EN (1997) Tajikistan. In: Moores EM and Fairbridge RW (eds.) *Encyclopedia of European and Asian Regional Geology*, pp. 71 80. London: Chapman and Hall.

Mukhin P (1997) Uzbekistan. In: Moores EM and Fairbridge RW (eds.) *Encyclopedia of European and Asian Regional Geology*, pp. 766 773. London: Chapman and Hall.

Rastsvetaev L (1997) Turkmenistan. In: Moores EM and Fairbridge RW (eds.) *Encyclopedia of European and Asian Regional Geology*, pp. 743 759. London: Chapman and Hall.

Yakubchuk A (1997) Kazakhstan. In: Moores EM and Fairbridge RW (eds.) *Encyclopedia of European and Asian Regional Geology*, pp. 450 464. London: Chapman and Hall.

Zonenshain LP, Kuzmin MI, and Natapov LM (1990) Geology of the USSR: A plate tectonic synthesis. Washington, DC, *AGU Geodynamic Series* 21: 242.

South-East

I Metcalfe, University of New England, Armidale, NSW, Australia

Introduction

East and South-east Asia is a giant 'jigsaw puzzle' of allochthonous continental lithospheric blocks and fragments (terranes) that are bounded by suture zones (representing the remnants of closed ocean basins) or by geological discontinuities such as major strike-slip faults (**Figure 1**).

The complex assemblage of South-east Asian continental terranes, accretionary complexes, ophiolites, volcanic arcs, and marginal ocean basins occurs in the zone of convergence between the Eurasian, Indo-Australian, and Pacific plates (**Figures 2** and **3**). In this region, two important biogeographical boundaries are recognized: the extant Wallace's Line and the Late Palaeozoic Gondwana–Cathaysia Divide (see **Figures 1**, **2**, and **3**). Wallace's Line marks the boundary between Eurasian faunas and floras to the north-west (including tigers, deer, and woodpeckers) and Australasian faunas and floras to the south-east (including wallabies, possums, and cockatoos) and was recognized by and named after Alfred Russel Wallace, now regarded as the Father of Biogeography. The Late Palaeozoic Gondwana–Cathaysia Divide forms the boundary between high-latitude cold-climate southern-hemisphere Gondwanan faunas and floras and equatorial or northern-hemisphere warm-climate sub-tropical or tropical Cathaysian faunas and floras. Both of these biogeographical boundaries are the result of convergent plate-tectonic processes bringing together allochthonous continental lithospheric terranes on which had developed contrasting faunas and floras owing to their prior geographical separation, different palaeoclimates, and biogeographical isolation. The regional geology of South-east Asia can be understood only in the context of the kinematic history and framework of the allochthonous continental terranes of the region and the evolution of the various ocean basins that once separated them.

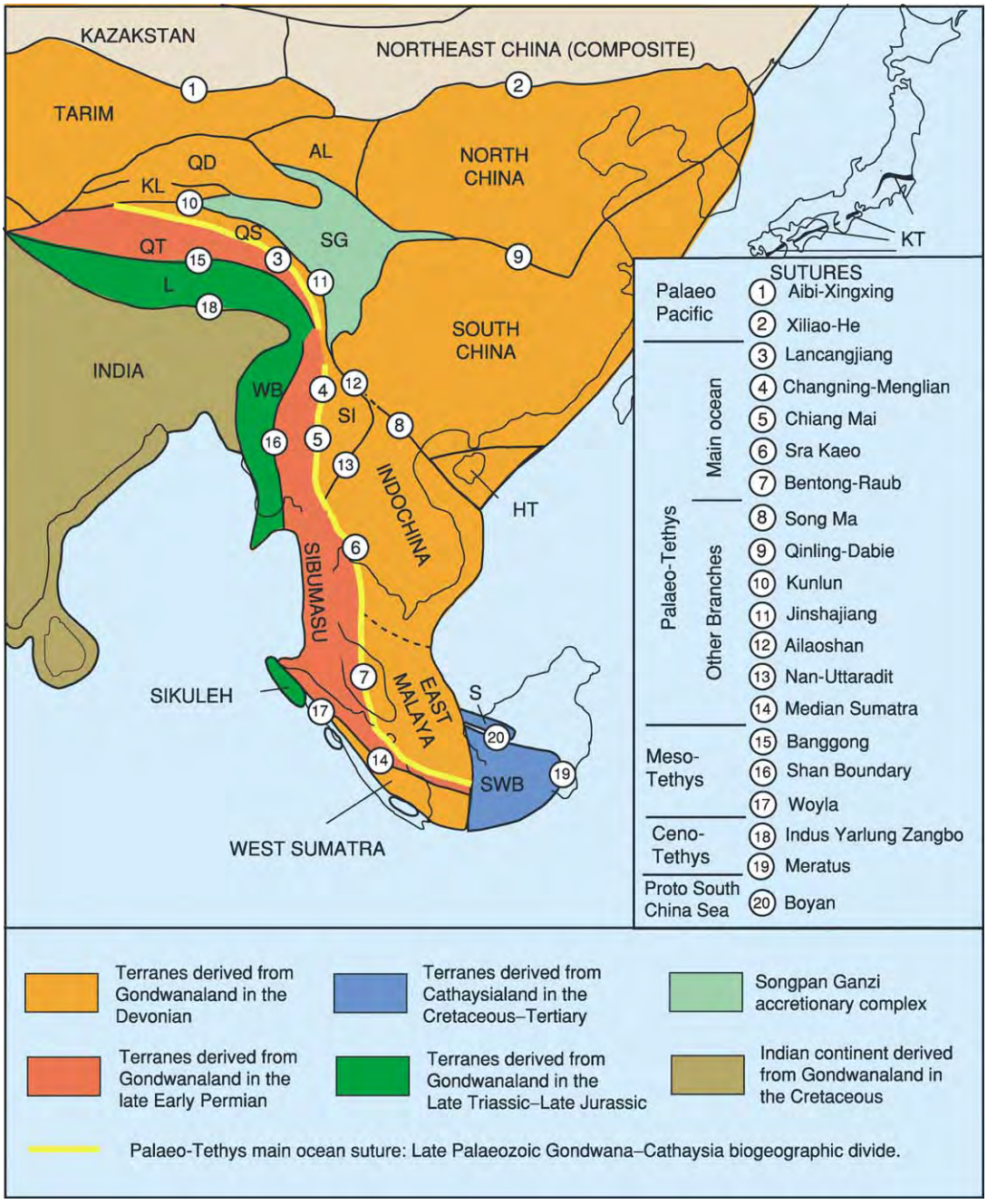


Figure 1 Distribution of the principal continental terranes and sutures of East and South east Asia. WB, West Burma; SWB, South West Borneo; S, Semitau Terrane; HT, Hainan Island terranes; L, Lhasa Terrane; QT, Qiangtang Terrane; QS, Qamdo Simao Terrane; SI, Simao Terrane; SG, Songpan Ganzi accretionary complex; KL, Kunlun Terrane; QD, Qaidam Terrane; AL; Ala Shan Terrane; KT, Kurosegawa Terrane.

Multidisciplinary data, including stratigraphic, biostratigraphic, biogeographical, palaeoclimatological, palaeomagnetic, and structural or tectonic information, indicate that in the Early Palaeozoic (545–410 Ma) all of the principal East and South-east Asian continental terranes were located on the margin of eastern Gondwana, where they formed an Indo-Australian 'Greater' Gondwana (*see* **Gondwanaland and Gondwana**) (Figure 4). They also probably formed part of the Indo-Australian element of the ancient supercontinent of Rodinia at 1000 Ma, prior to its breakup around 700 Ma and the subsequent formation of Gondwana at about 500 Ma (Figure 5).

Multidisciplinary data also suggest that the East and South-east Asian terranes rifted and separated from Gondwana as three continental slivers around

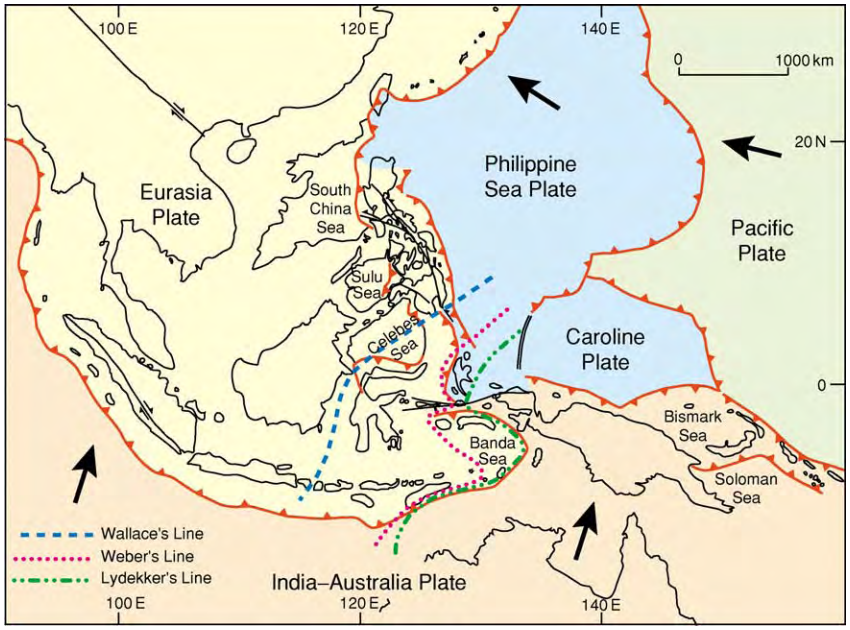


Figure 2 Principal tectonic plates of South east Asia. Arrows show relative plate motions.

350 Ma, 270 Ma, and 200–140 Ma, in the Devonian (North China, South China, Tarim, Indochina, East Malaya, West Sumatra), Lower Permian (Sibumasu, Qiangtang), and Late Triassic–Late Jurassic (Lhasa, West Burma, possibly Sikuleh, possibly West Sulawesi etc.), respectively. As these three continental slivers separated from Gondwana and drifted northwards, successive ocean basins opened between each sliver and Gondwana: the Palaeo-Tethys, Meso-Tethys and Ceno-Tethys, respectively (Figure 6). The Meso-Tethys and Ceno-Tethys are broadly equivalent to the Neo-Tethys of some workers. Destruction and closure of these ocean basins by subduction during Carboniferous to Cenozoic times led to the juxtaposition, by amalgamation and accretion (continental collisions), of once widely separated continental fragments, and the remnants of the ocean basins are now preserved in the suture zones of the region.

Smaller continental fragments, distributed in eastern South-east Asia, were derived from Indochina and South China during the opening of the South China Sea and southwards subduction and destruction of the Proto-South China Sea, or were transported westwards along major strike-slip faults from the northern Australian margin during its collision with the westwards-moving Philippine Sea, Caroline, and Pacific plates in a kind of 'bacon-slicer' tectonic mechanism.

The regional geology of South-east Asia is thus characterized by Gondwanan dispersion and Asian accretion of terranes and the subsequent collisions of India and Australia with these terranes following the breakup of Gondwana and their northwards drift.

The geological evolution of South-east Asia is thus essentially the combined and cumulative history of these terranes, the ocean basins that once separated them, and the plate-tectonic processes that have shaped the region. A variety of multidisciplinary data (Table 1) is used to constrain the origins of the terranes, their times of rifting and separation from the parent cratons, the timing, directions, and amount of drift, and the timing of suturing (collision and welding) of the terranes to each other. Some terranes sutured to each other (amalgamated) within a major ocean before, as an amalgamated composite terrane, they sutured (accreted) to proto-Asia.

Origins of the South-East Asian Terranes

Multidisciplinary data (Table 1) suggest that all the East and South-east Asian continental terranes originated on the Indian or northern or north-western Australian margin of Gondwana. Cambrian and Ordovician shallow-marine faunas of the North China, South China, and Sibumasu terranes have close affinities with those of eastern Gondwana, especially Australian Gondwana. This is observed in trilobites, brachiopods, corals and stromatoporoids, nautiloids, gastropods, and conodonts (Figure 4). More recently, the Gondwanan acritarch *Dicrodiacroium ancoriforme* Burmann has been reported from the Lower Ordovician of South China. The Cambrian–Ordovician faunas of Indochina are poorly known, but the Silurian brachiopods of Indochina

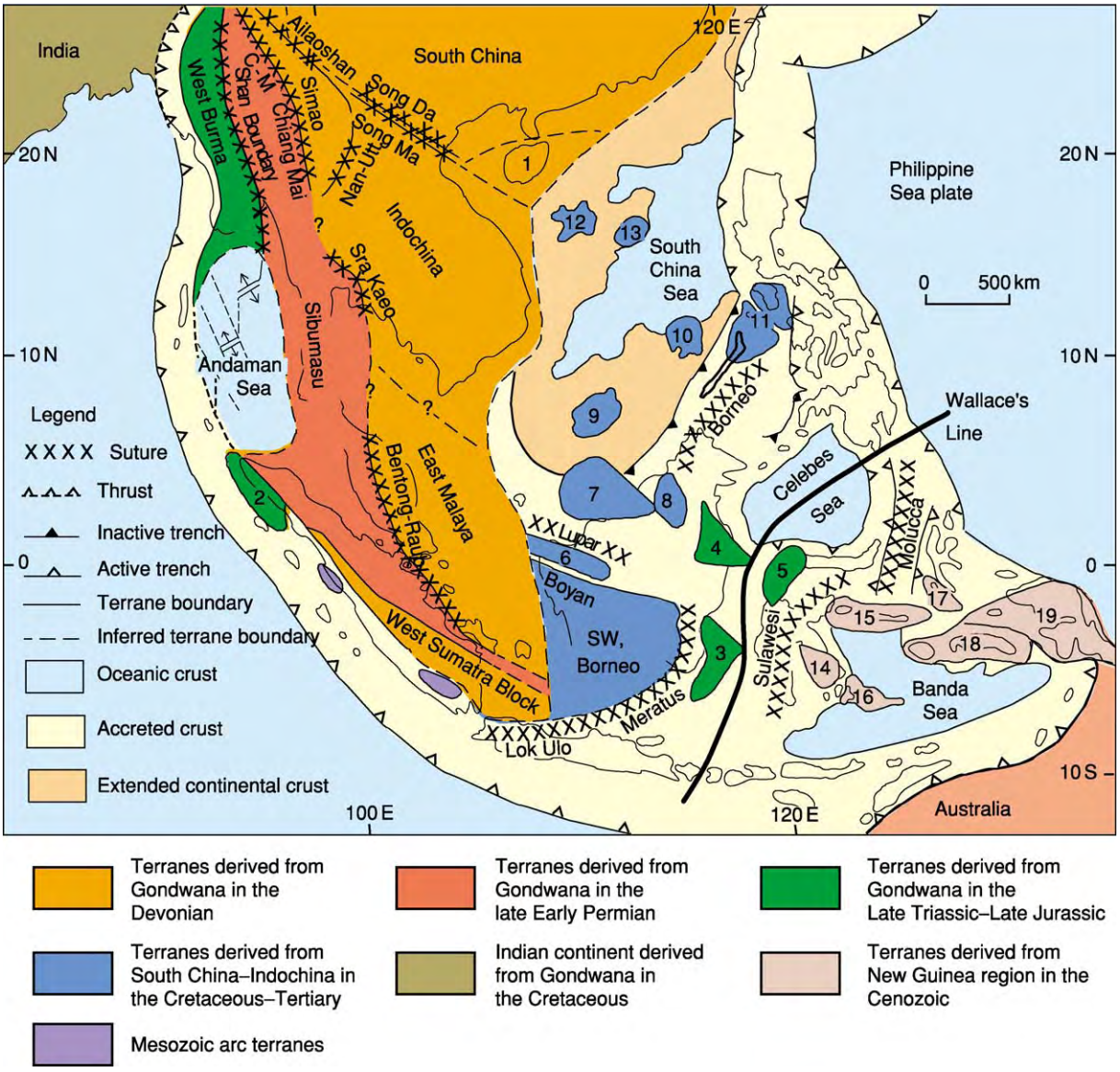


Figure 3 Distribution of continental blocks, fragments, and terranes, together with the principal sutures of South east Asia. Numbered microcontinental blocks: 1, Hainan Island terranes; 2, Sikuleh; 3, Paternoster; 4, Mangkalihat; 5, West Sulawesi; 6, Semitau; 7, Luconia; 8, Kelabit Longbowan; 9, Spratley Islands Dangerous Ground; 10, Reed Bank; 11, North Palawan; 12, Paracel Islands; 13, Macclesfield Bank; 14, East Sulawesi; 15, Bangai Sula; 16, Buton; 17, Obi Bacan; 18, Buru Seram; 19, West Irian Jaya. C M, Changning Menglian Suture.

along with those of South China, North China, Eastern Australia, and the Tarim terrane define a Sino-Australian province characterized by the *Retziella* fauna (Figures 4 and 7). Lower Palaeozoic sequences and faunas of the Qaidam, Kunlun, and Ala Shan blocks are similar to those of the Tarim block and South and North China, and these blocks are regarded as disrupted fragments of a larger Tarim terrane. These biogeographical data suggest that North China, South China, Tarim (here taken to include the Qaidam, Kunlun, and Ala Shan blocks), Sibumasu (with the contiguous Lhasa and Qiangtang blocks), Indochina, East Malaya, and West Sumatra formed the outer margin of northern Gondwana in the Early Palaeozoic. The close faunal affinities, at both lower and higher taxonomic levels, suggest continental contiguity of these blocks with each other and with Gondwana at this time rather than mere close proximity.

Early Palaeozoic palaeomagnetic data for the various East and South-east Asian terranes are often equivocal, varying in both quantity and quality. This makes reconstructions based purely on palaeomagnetic data difficult and suspect. However, in some cases, reasonable constraints on palaeolatitudes (but

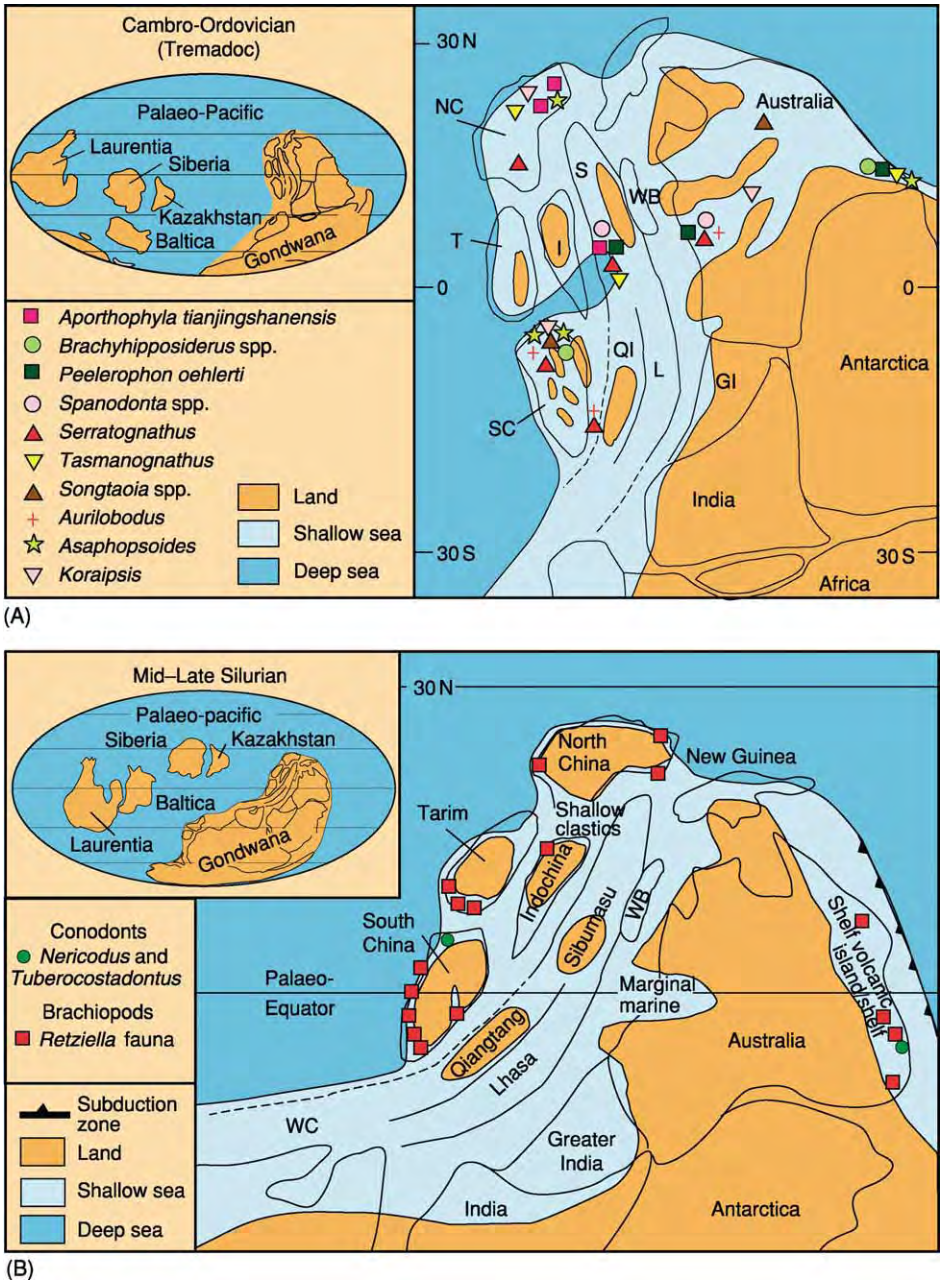


Figure 4 Reconstructions of eastern Gondwana for (A) the Cambro Ordovician (Tremadoc) and (B) the Mid Late Silurian, showing the postulated positions of the East and South east Asian terranes, the distribution of land and sea, and the distributions of shallow marine fossils that illustrate Asia Australia connections at these times. NC, North China; SC, South China; T, Tarim; I, Indochina East Malaya West Sumatra; QI, Qiangtang; L, Lhasa; S, Sibumasu; WB, West Burma; WC, Western Cimmerian Continent; GI, Greater India.

not always the hemisphere) and in some cases the actual position of attachment to Gondwana can be made. Data from North China provide a Cambrian–Late Devonian pole-path segment that, when rotated about an Euler pole to a position of fit with the Australian Cambrian–Late Devonian pole path, positions North China adjacent to North Australia. This position is consistent with reconstructions presented here. The gross stratigraphies of North China and the Arafura Basin show a remarkable similarity in the Early Palaeozoic, also supporting this position for North China. Positions for South China, Tarim, and Indochina are more equivocal, but latitudes of between 1° and 15° are indicated for South China in the Late Cambrian–Early Ordovician. The Tarim Block is placed between 6° S and 20° S for the same time period, consistent with a position on the Gondwanan margin between the North and South

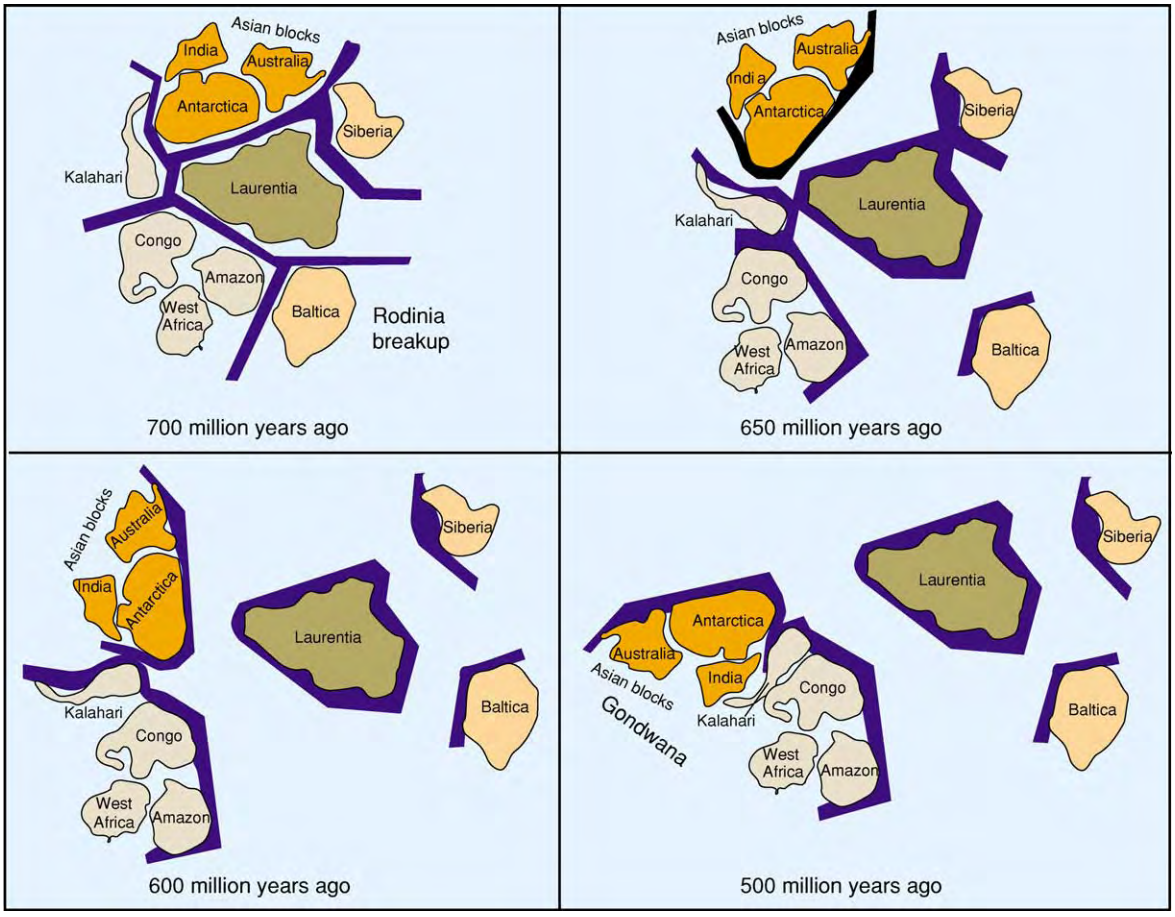


Figure 5 The supercontinent Rodinia at 700 Ma and its subsequent breakup, showing the formation of Gondwana at about 500 Ma.

China blocks. An origin for the Tarim Block outboard of the Kimberley region of Australia is suggested by comparisons of Precambrian sequences. The Cambrian–Early Permian faunas of the Sibumasu Terrane have unequivocal Gondwanan affinities and in particular show close relationships with western Australian faunas. Gondwanan plants and spores are also reported from this terrane. Glacial–marine diamictites, with associated cold-water faunas and sediments, of possibly Late Carboniferous and (mainly) Early Permian age are also distributed along the entire length of Sibumasu (Figure 8) and indicate attachment to the margin of Gondwana, where substantial ice reached the sea during glaciation. The most likely region for attachment of this terrane is north-western Australia, and a Late Carboniferous palaeolatitude of 42° S supports this placement. Striking similarities in the Cambrian–Early Permian gross stratigraphies of Sibumasu and the Canning Basin of north-western Australia also support a position outboard of the Canning Basin during this time. Both the Qiangtang and Lhasa blocks of Tibet exhibit Gondwanan faunas and floras up to the Early Permian and have glacial–marine diamictites, till, and associated cold-water faunas and sediments in the Late Carboniferous–Early Permian. Thus, all the East and South-east Asian continental terranes appear to have originated on the margin of Gondwana.

Rifting and Separation of South-East Asian Terranes from Gondwana

Devonian Rifting and Separation

South China, North China, Tarim, Indochina, East Malaya, and West Sumatra were attached to Gondwana in the Cambrian–Silurian but by Carboniferous times were separated from the parent craton, suggesting a Devonian rifting and separation of these blocks. This timing is supported by the presence of a conspicuous Devonian unconformity in South China and a subsequent Devonian–Triassic passive-margin sequence along its southern margin. Devonian basin formation in South China has also been shown to be related to rifting. The splitting of the Silurian Sino-Australian brachiopod province into two sub-provinces and the apparent loss of links

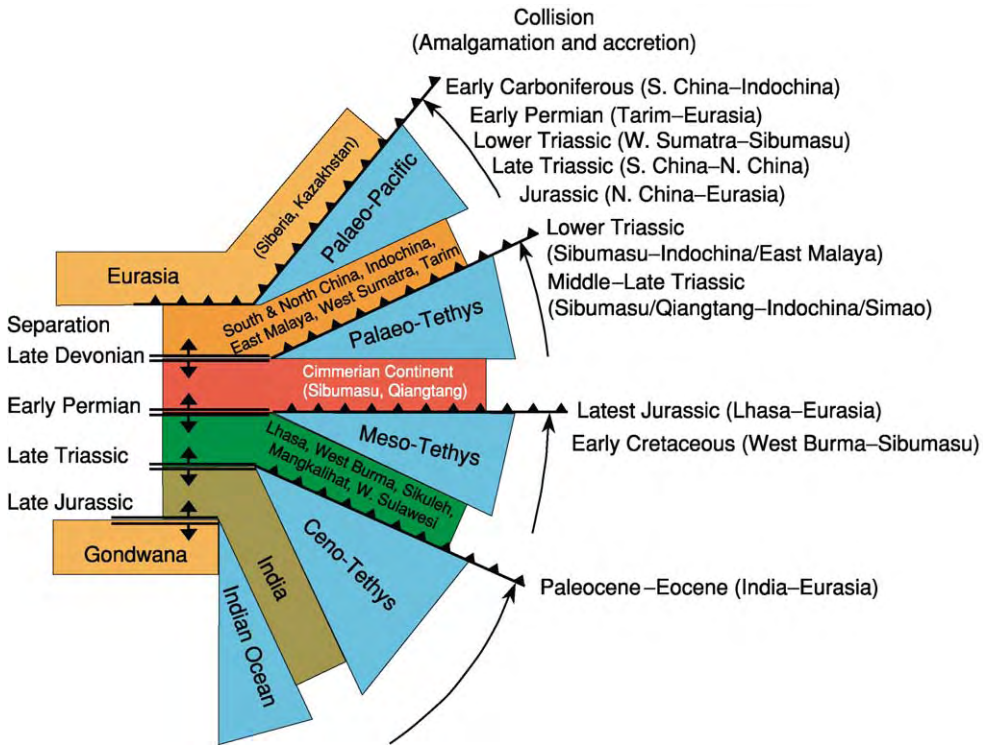


Figure 6 Times of separation and subsequent collision of the three continental slivers or collages of terranes that rifted from Gondwana and were translated northwards by the opening and closing of three successive oceans: the Palaeo Tethys, Meso Tethys, and Ceno Tethys.

between the Asian terranes and Australia in the Early Devonian may be consequences of the northwards movement and separation of these terranes from Gondwana. A counterclockwise rotation of Gondwana in the Devonian about an Euler pole in Australia is also consistent with clockwise rotation of the separating Asian terranes and spreading of the Palaeo-Tethys at this time.

Carboniferous–Permian Rifting and Early Permian Separation

There is now substantial evidence for rifting along the northern margin of Gondwana in the Carboniferous–Permian accompanied by rift-related magmatism. This rifting episode led to the late Early Permian (Late Sakmarian) separation of the Sibumasu and Qiangtang terranes, as part of the Cimmerian continent, from the Indo-Australian margin of Gondwana. The late Early Permian separation and Middle–Upper Permian rapid northwards drift of the Sibumasu Terrane are supported by palaeolatitude data, which indicate a change of latitude from 42° S in the Late Carboniferous to low northern latitudes by the Early Triassic. In addition, the faunas of the Sibumasu Terrane show a progressive change in marine provinciality from peri-Gondwanan Indoralian Province faunas in the Early Permian (Asselian–Early Sakmarian), to endemic Sibumasu province faunas in the Middle Permian, before being absorbed into the equatorial Cathaysian province in the Late Permian.

Late Triassic to Late Jurassic Rifting and Separation

The separation of the Lhasa Block from Gondwana has been proposed by different authors to have occurred in either the Permian or the Triassic. A Permian separation is advocated, either as part of the Cimmerian continent or as a 'Mega-Lhasa' Block that included Iran and Afghanistan. Permian rifting on the North Indian margin and in Tibet is here regarded as being related to separation of the Cimmerian continental strip, which included Iran, Afghanistan, and the Qiangtang Block of Tibet, but not the Lhasa Block. Sedimentological and stratigraphic studies in the Tibetan Himalayas and Nepal have documented the Triassic rifting and Late Triassic (Norian) separation of the Lhasa Block from northern Gondwana. This Late Triassic episode of rifting is also recognized along the north-western shelf of Australia, where it continued into the Late Jurassic, resulting in the separation of West Burma.

Table 1 Multidisciplinary constraining data for the origins and the rift drift suturing of terranes

Origin of terrane	*Age of rifting and separation*	*Drifting (palaeoposition of terrane)*	*Age of suturing (amalgamation/ accretion)*
Palaeobiogeographical constraints (fossil affinities with proposed parent craton)	Ocean floor ages and magnetic stripe data	Palaeomagnetism (palaeolatitude, orientation)	Ages of ophiolite and ophiolite obduction ages (pre suturing)
Tectonostratigraphic constraints (similarity of gross stratigraphy with that of parent craton, presence of distinctive lithologies characteristic of parent craton, e.g. glacials)	Divergence of apparent polar wander paths indicates separation	Palaeobiogeography (shifting from one biogeographical province to another with drift)	Melange ages (pre suturing)
Palaeolatitude and orientation from palaeomagnetism consistent with proposed origin	Divergence of palaeolatitudes (indicates separation)	Palaeoclimatology (indicates palaeolatitudinal zone)	Age of 'stitching' plutons (post suturing)
	Age of associated rift volcanism and intrusives		Age of collisional or post collisional plutons (syn to post suturing)
	Regional unconformities (formed during pre rift uplift and during block faulting)		Age of volcanic arc (pre suturing)
	Major block faulting episodes and slumping		Major changes in arc chemistry (syn collisional)
	Palaeobiogeography (development of separate biogeographical provinces after separation)		Convergence of apparent polar wander paths
	Stratigraphy rift sequences in grabens and half grabens		Loops or disruptions in apparent polar wander paths (indicates rapid rotations during collisions)
			Convergence of palaeolatitudes (may indicate suturing but no control on longitudinal separation)
			Age of blanketing strata (post suturing)
			Palaeobiogeography (migration of continental animals and plants from one terrane to another indicates terranes have sutured)
			Stratigraphy/sedimentology (e.g. provenence of sedimentary detritus from one terrane on to another)
			Structural geology (age of deformation associated with collision)

Reproduced with permission from Metcalfe I (1998) Palaeozoic and Mesozoic geological evolution of the SE Asian region: multidisciplinary constraints and implications for biogeography. In: Hall R and Holloway JD (eds.) (1998) *Biogeography and Geological Evolution of SE Asia*, pp. 25 41. Amsterdam: Backhuys Publishers.

Amalgamation and Accretion of Terranes

The continental terranes of East and South-east Asia progressively amalgamated during the Palaeozoic to Cenozoic. Most of the major terranes had coalesced by the end of the Cretaceous and proto South-east Asia had formed. The time of welding of one terrane to another can be determined using the various criteria given in Table 1. Table 2 lists the East and South-east Asian suture zones, colliding lithospheric blocks, interpreted times of suturing, and constraints on the times of suturing. The tectonostratigraphic record of each continental terrane in the region documents the geological history of that terrane, including variations in sedimentary environment, climate, faunal and

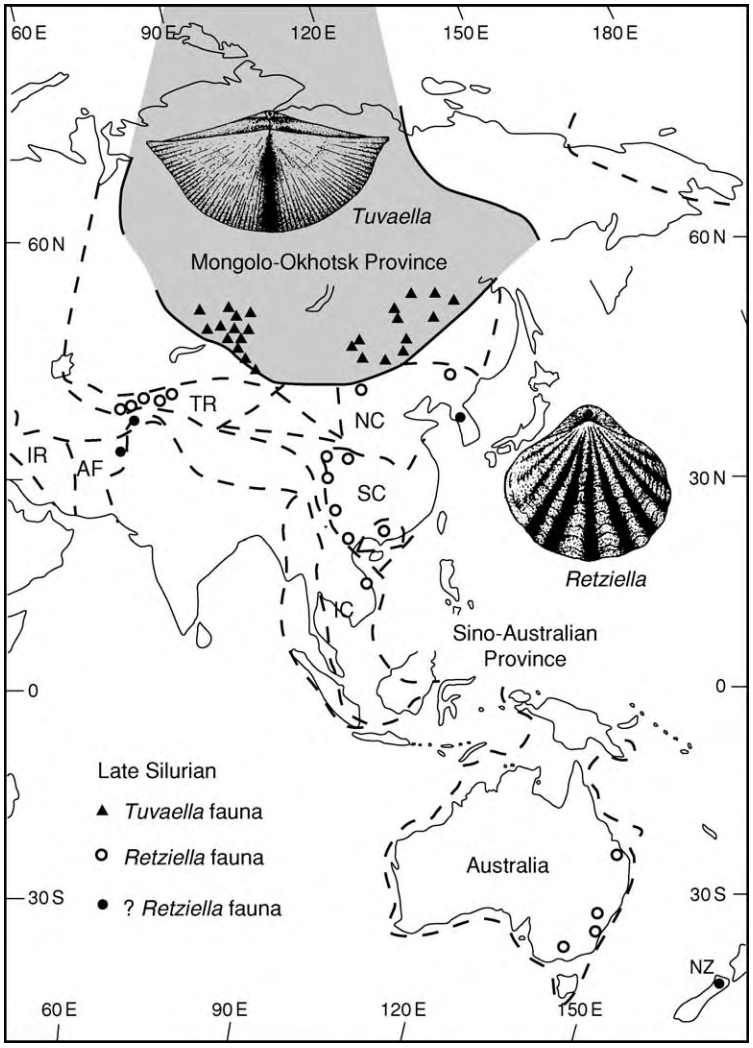


Figure 7 Late Silurian brachiopod provinces of Asia Australasia. IR, Iran; AF, Afghanistan; TR, Tarim; NC, North China; SC, South China; IC, Indochina; NZ, New Zealand.

floral affinities (changes in biogeographical regime) and latitudinal shifts, rifting events, episodes of deformation, and plutono-volcanic igneous activity. The regional geological history of South-east Asia is discussed below chronologically and in terms of the evolution of the various tectonic elements of the region.

Geological and Tectonic Evolution of South-East Asia

Proterozoic (2500–545 Ma)

The East and South-east Asian terranes, together with India and Australia, formed an integral part of the ancient supercontinent of Rodinia around 1000 Ma (*see* **Precambrian:** Overview). Fragmentation of this ancient supercontinent about 700 Ma ago led to Australia, India, Antarctica, and elements that now constitute South Africa and South America colliding and coalescing to form Gondwana about 500 Ma (Figure 5). Proterozoic basements of the mainland South-east Asian terranes are indicated by limited outcrops of schists and gneisses with Proterozoic radioisotopic ages in Thailand and Vietnam, by more substantial relatively unmetamorphosed Proterozoic sedimentary sequences in South China, and by Proterozoic radioisotopic ages of inherited zircons in granitoids in Peninsular Malaysia.

Phanerozoic (545–0 Ma)

Table 3 summarizes the principal geological events, and their timings, that have affected the South-east Asian region during the Palaeozoic, Mesozoic, and Cenozoic Eras. Continental collisions (amalgamation

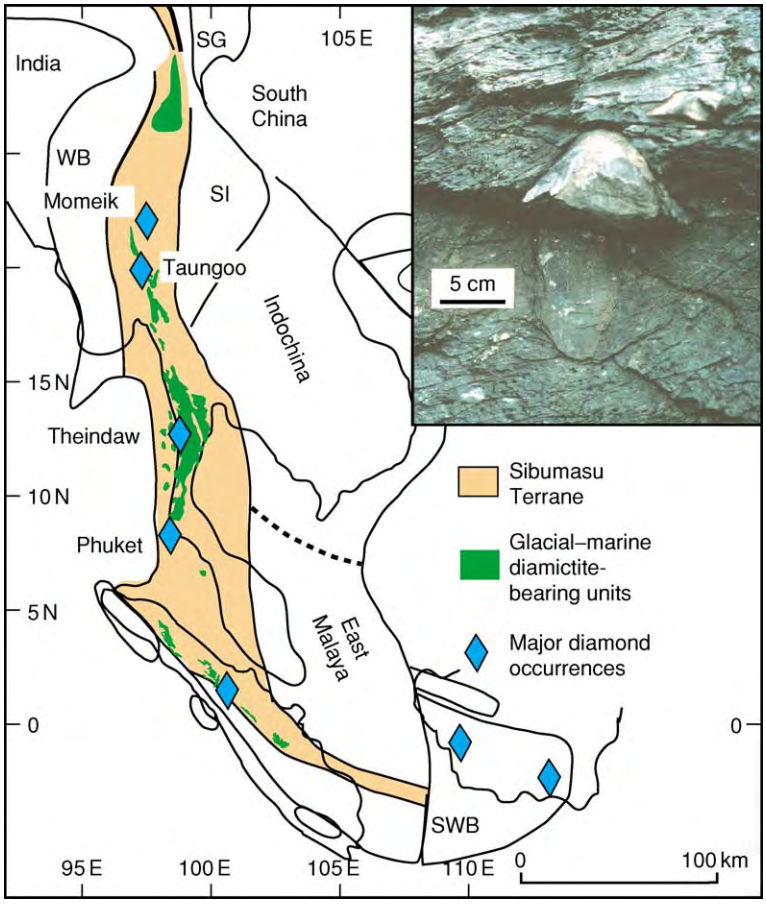


Figure 8 Map of mainland South east Asia, showing the distributions of Carboniferous Early Permian glacial marine sediments and major alluvial diamond deposits. Inset photo: dropstone in glacial marine diamictite orientated vertical to bedding; Singa Formation, Langkawi Islands, Peninsular Malaysia.

or accretion) are dated by the various suture zones (Table 2). The principal events that have affected the South-east Asian region during the Phanerozoic are outlined below chronologically.

Cambrian–Ordovician–Silurian (545–410 Ma) The East and South-east Asian terranes formed part of Indian–Australian 'Greater Gondwana'. Faunas of this age on the Asian blocks and in Australasia define Asian–Australian palaeo-equatorial warm-climate 'provinces', for example the Sino-Australian brachiopod province in the Silurian – see Figures 4 and 7.

Devonian (410–354 Ma) Australian eastern Gondwana continued to reside in low southern latitudes during the Devonian but rotated counterclockwise. This counterclockwise rotation mirrors a clockwise rotation of the North and South China, Tarim, Indochina, East Malaya, and West Sumatra terranes as they separated from Gondwana as an elongate continental sliver. Separation of this sliver from Gondwana opened the Palaeo-Tethys Ocean (Figure 9). Devonian faunas on the Chinese terranes still have some Australian connections. Early Devonian endemic faunas of South China, including some fish faunas and the distinctive *Chuiella* brachiopod fauna (Figure 9), are interpreted to be a result of the rifting process and isolation of South China on the rifting continental promontory, and do not necessarily imply continental separation of South China from the other Chinese blocks and Australia at this time.

Carboniferous (354–298 Ma) During the Carboniferous, Gondwana rotated clockwise and collided with Laurentia in the west to form Pangaea (*see* **Pangaea**). Australia, Sibumasu, Qiangtang, Lhasa, and West Burma were still attached to north-eastern Gondwana and drifted from low southern latitudes in Tournaisian–Visean times to high southern latitudes in Middle–Late Namurian times. The major Gondwanan glaciation commenced in the Namurian and extended through to the Early Permian. There were

Table 2 East and South east Asian sutures and their interpreted ages and age constraints. For location of sutures see **Figures 1** and **2**

No. on ***Fig. 1***	*Suture name*	*Colliding lithospheric blocks*	*Suture age*	*Age constraints*
1	Aibi Xingxing	Tarim, Kazakhstan	Permian	Lower Carboniferous ophiolites. Major arc magmatism ceased in the Late Carboniferous. Late Permian post orogenic subsidence and continental sedimentation in Junggar Basin. Palaeomagnetic data indicate convergence of Tarim and Kazakhstan by the Permian. Upper Permian continental clastics blanket the suture
2	Xiliao He	North China, Altaid terranes	Jurassic	Late Jurassic Early Cretaceous deformation and thrust faulting. Widespread Jurassic Cretaceous granites. Triassic Middle Jurassic deep marine cherts and clastics. Upper Jurassic Lower Cretaceous continental deposits blanket suture
3	Lancangjiang Suture	Qiangtang, Qamdo Simao	Early Triassic	Suture zone rocks include Devonian and Carboniferous turbiditic 'flysch'. Ocean floor basic extrusives of Permian age and Carboniferous Permian mélange. Carboniferous Permian island arc rocks are developed along the west side of the suture. Upper Triassic collisional granitoids are associated with the suture. Suture zone rocks are blanketed by Middle Triassic continental clastics
4	Changning Menglian Suture	Sibumasu, Simao	Late Permian Late Triassic	Oceanic ribbon bedded chert shale sequences have yielded graptolites, conodonts, and radiolarians indicating ages ranging from Lower Devonian to Middle Triassic. Limestone blocks and lenses dominantly found within the basalt sequence of the suture and interpreted as seamount caps, have yielded fusulinids indicative of Lower Carboniferous to Upper Permian ages
5	Chiang Mai Suture	Sibumasu, Simao	Middle Triassic	Basic volcanics (including pillow basalts) are dated as Carboniferous and Permian. Ages of oceanic deep marine bedded cherts within the suture zone range from Devonian to Middle Triassic. Seamount limestone caps are dated as Lower Carboniferous (Visean) to Upper Permian (Changxingian) in age
6	Sra Kaeo Suture	Sibumasu, Indochina	Late Triassic	Suture zone rocks include melange and chert clastic sequences, which include ultrabasics, serpentinites, Carboniferous pillow basalts, and Early Permian to Late Triassic oceanic sediments and associated pillow basalts. Limestone blocks in the mélange range from Upper Lower Permian to Middle Permian and a granitic lens has yielded a zircon U Pb age of 486 ± 5 Ma. Imbricate thrust slices dated as Middle Triassic by radiolarians. Jurassic continental sandstones blanket the suture zone
7	Bentong Raub Suture	Sibumasu, East Malaya	Early Triassic	Upper Devonian to Upper Permian oceanic ribbon bedded cherts. Mélange includes chert and limestone clasts with Lower Carboniferous to Upper Permian ages. The Main Range 'collisional' 'S' type granites of Peninsular Malaysia range from Late Triassic (230 ± 9 Ma) to earliest Jurassic (207 ± 14 Ma) in age, with a peak of around 210 Ma. Suture zone is covered by latest Triassic, Jurassic, and Cretaceous, mainly continental, overlap sequence
8	Song Ma	Indochina, South China	Late Devonian Early Carboniferous	Large scale folding and thrusting and nappe formation in the Early to Middle Carboniferous. Middle Carboniferous shallow marine carbonates are reported to blanket the Song Ma suture in North Vietnam. Pre middle Carboniferous faunas on each side of the Song Ma zone are different whilst the Middle Carboniferous faunas are essentially similar. Carboniferous floras on the Indochina block in north eastern Thailand indicate continental connection between Indochina and South China in the Carboniferous
9	Qinling Dabie	South China, North China	Triassic Jurassic	Late Triassic subduction related granite. Late Triassic U Pb dates of zircons from ultrahigh pressure eclogites. Late Triassic Early Jurassic convergence of apparent polar wander paths and palaeolatitudes of South China and North China. Initial contact between South and North China is indicated by isotopic data in Shandong and sedimentological records along the suture. Widespread Triassic to Early Jurassic deformation in the North China block north of the suture

Continued

Table 2 Continued

No. on ***Fig. 1***	*Suture name*	*Colliding lithospheric blocks*	*Suture age*	*Age constraints*
10	Kunlun	Qamdo Simao Kunlun	Triassic	Permo Triassic ophiolites and subduction zone mélange. Upper Permian calc alkaline volcanics, strongly deformed Triassic flysch and Late Triassic granitoids
11	Jinshajiang	Simao, South China	Late Permian Late Triassic	Ophiolites are regarded as Upper Permian to Lower Triassic in age. Mélange comprises Devonian, Carboniferous, and Permian exotics in a Triassic matrix. Upper Permian to Jurassic sediments unconformably overlie Lower Permian ophiolites in the Hoh Xil Range
12	Ailaoshan Suture	Simao, South China	Middle Triassic	Plagiogranite U Pb ages of 340 ± 3 Ma and 294 ± 3 Ma indicate that the oceanic lithosphere formed in the latest Devonian to earliest Carboniferous. Ophiolitic rocks are associated with deep marine sedimentary rocks including ribbon bedded cherts that have yielded some Lower Carboniferous and Lower Permian radiolarians. Upper Triassic sediments (Carnian conglomerates and sandstones, Norian limestones, and Rhaetian sandstones) blanket the suture
13	Nan Uttaradit Suture	Simao, Indochiona	Middle Triassic	Pre Permian ophiolitic mafic and ultramafic rocks with associated blueschists. Mafic and ultramafic blocks in the mélange comprise ocean island basalts, back arc basin basalts and andesites, island arc basalts and andesites and supra subduction cumulates generated in Carboniferous to Permo Triassic times. Permo Triassic dacites and rhyolites associated with relatively unmetamorphosed Lower Triassic sandstone shale turbidite sequence. Suture zone rocks are overlain unconformably by Jurassic redbeds and post Triassic intraplate continental basalts
14	Median Sumatra	West Burma, Sibumasu	Early Triassic	No remnants of the ocean basin that once separated West Sumatra and Sibumasu so far known. It is likely that West Sumatra was slid into juxtaposition with Sibumasu from east to west along a major strike slip fault associated with oblique subduction
15	Banggong Suture	Lhasa, Qiangtang	Late Jurassic Earliest Cretaceous	Suture is blanketed in Tibet by Cretaceous and Palaeogene rocks. Structural data indicate collision around the Jurassic Cretaceous boundary
16	Shan Boundary Suture	West Burma, Sibumasu	Early Cretaceous	Cretaceous thrusts in the back arc belt. Late Cretaceous age for the Western Belt tin bearing granites
17	Woyla Suture	Sikuleh, Sibumasu/W Sumatra	Late Cretaceous	Cretaceous ophiolites and accretionary complex material
18	Indus Yarlung Zangbo Suture	India, Eurasia	Late Cretaceous Eocene	Jurassic Cretaceous ophiolites and ophiolitic mélange with Jurassic Lower Cretaceous radiolarian cherts. Late Cretaceous blueschists. Eocene collision related plutons. Palaeomagnetic data indicates initial collision around 60 Ma. Palaeogene strata blanket the suture
19	Lok Ulo Meratus Suture	Paternoster, SW Borneo	Late Cretaceous	Subduction mélange of Middle Late Cretaceous age. Ophiolite with Jurassic ultramafic rocks. Pillow basalts of Jurassic and Early Cretaceous ages. Oceanic cherts of Late Jurassic to Late Cretaceous ages. Turbiditic flysche of Early Late Cretaceous age. Ophiolite obducted in Cenomanian. Suture overlain by Eocene strata
20	Boyan Suture	Semitau, SW Borneo	Late Cretaceous	Upper Cretaceous mélange
	Lupar Adio Borneo Suture	South China, Mangkalihat	Early to middle Miocene	Ophiolites of probable Cretaceous age. Subduction related melanges of Early Miocene and older ages. Melanges include tectonic melanges, olistostromes and mud diapirs. Blueschist metamorphism reported. Accretionary deformation in South Sabah until Early Miocene, continued until Middle Miocene in the north
	Sulawesi Suture	W. Sulawesi, Australian fragments	Late Oligocene Early Miocene	Complex accretion of ophiolites and arc and continental fragments (Australia derived). At least three phases of accretion. Cretaceous to Early Oligocene marine sediments associated with ophiolites. Oligo Miocene blueschist metamorphism. Large granite plutons in West Sulawesi dated at 12 4 Ma
	Molucca Suture	Arc arc collision	Pliocene and ongoing	Arc arc collision. Double subduction system with complete subduction of Molucca Sea Plate. Melange wedge/collision complex in Molucca Sea

Table 3 Palaeozoic, Mesozoic, and Cenozoic events and their ages in South east Asia

Process	*Age*
Palaeozoic evolution	
Rifting of South China, North China, Indochina, East Malaya, West Sumatra, Tarim, and Qaidam from Gondwana	Early Devonian
Initial spreading of the Palaeo Tethys ocean	Middle Late Devonian
Amalgamation of South China, Indochina, and East Malaya to form Cathaysialand	Late Devonian to Early Carboniferous
Development of the Ailaoshan Nan Uttaradit back arc basin and separation of the Simao Terrane by back arc spreading	Late Early Carboniferous
Rifting of Sibumasu and Qiangtang from Gondwana as part of the Cimmerian continent	Late Early Permian (Sakmarian)
Initial spreading of Meso Tethys ocean	Middle Permian
Collision and suturing of Sibumasu to Indochina and East Malaya	Latest Permian to Early Triassic
Initial collision of South and North China and development of Tanlu Fault	Late Permian to Triassic
Mesozoic evolution	
Suturing of South China with North China and final consolidation of Sundaland	Late Triassic to Early Jurassic
Rifting of Lhasa, West Burma and other small terranes	Late Triassic to Late Jurassic
Initial spreading of Ceno Tethys ocean	Late Triassic (Norian) in west (northern India) and Late Jurassic in east (NW Australia)
Northward drift of Lhasa, West Burma, and small terranes	Jurassic to Cretaceous
Collision of the Lhasa Block with Eurasia	Late Jurassic Earliest Cretaceous
Accretion of West Burma and Sikuleh? terranes to Sibumasu	Late Early Cretaceous
Suturing of Semitau to SW Borneo	Late Cretaceous
Cenozoic evolution	
Collision of India with Eurasia	Initial contact around 60 Ma, major indentation from around 45 Ma onwards
Final separation of Australia from Antarctica and establishment of the circum Antarctic ocean current	45 Ma, Late Eocene
Northwards drift of the isolated Australian continent over 27° of latitude. Gradual drying of the continent and evolution of the distinctive Australian fauna and flora from Gondwana ancestry	45 0 Ma, Late Eocene to present
Clockwise rotation and extrusion of Indochina and parts of northern Sibumasu	During the early Cenozoic but precise age not known
Major strike slip faulting	30 15 Ma, Middle Oligocene to Middle Miocene
Anti clockwise rotation of Borneo and the Malay Peninsula	Progressively during the Cenozoic but mostly during the Miocene
Initial spreading of the South China Sea basin	30 Ma, Middle Oligocene
Proto South China Sea destroyed by southwards subduction	40 15 Ma, Middle Eocene to Middle Miocene
Collision of the Australian continent with the Philippine Sea Plate (initiating clockwise rotation of Philippine Sea Plate and causing structural inversion in many Cenozoic basins of the region)	25 Ma, Late Oligocene
Clockwise rotation of the Philippine Sea Plate	From about 20 Ma onwards
Opening of the Sulu Sea	20 Ma, Early Miocene
Molucca Sea double subduction established	10 Ma, Late Miocene
Collision of the Philippine Arc and Eurasian continental margin in Taiwan	5 Ma, End Miocene

major global shifts in both plate configurations and climate during this time, and there was a change from warm to cold conditions in Australasia. This is reflected in the change from a high to a low diversity of faunas in Australasia, especially eastern Australia, where low-diversity endemic faunas developed in the Upper Carboniferous. The faunas and floras of North and South China, Indochina, East Malaya, West Sumatra, and Tarim are tropical or sub-tropical Cathaysian or Tethyan types during the Carboniferous and show no Gondwanan affinities (Figure 10). These terranes had already separated from Gondwana and were located in low-latitude or equatorial positions during the Carboniferous (Figure 11). Indochina and South China collided and amalgamated within Tethys during the Early Carboniferous along the Song Ma suture zone, which is now located in Laos and Vietnam. Ice-sheets and glaciers extended across much of eastern Gondwana during the Late Carboniferous: ice reached the marine environment of the

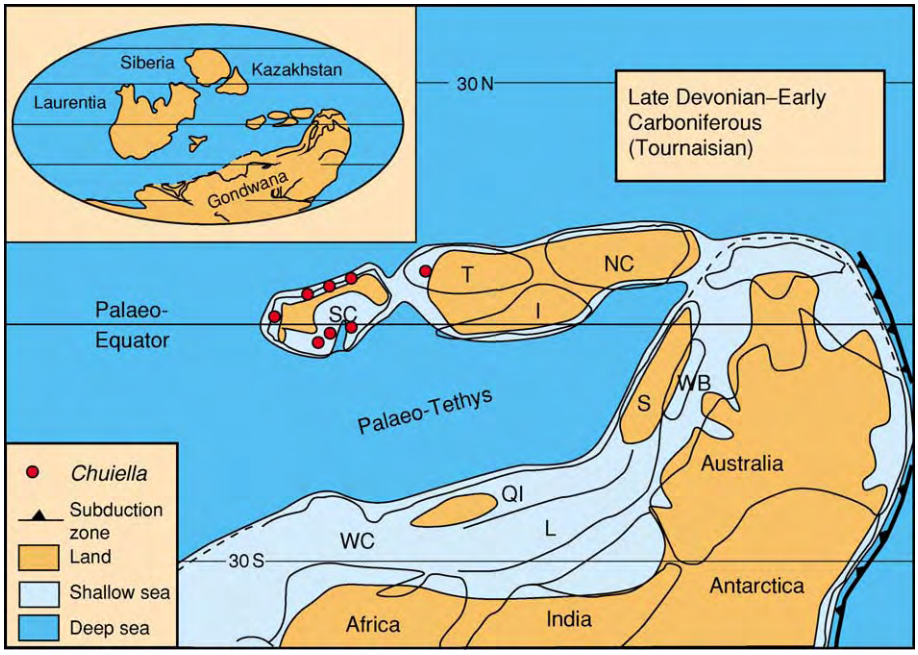


Figure 9 Reconstruction of eastern Gondwana for the Late Devonian to Lower Carboniferous (Tournaisian), showing the postulated positions of the East and South east Asian terranes, the distribution of land and sea, and the opening of the Palaeo Tethys ocean at this time. Also shown is the distribution of the endemic Tournaisian brachiopod genus *Chuiella*. NC, North China; SC, South China; T, Tarim; I, Indochina East Malaya West Sumatra; QI, Qiangtang; L, Lhasa; S, Sibumasu; WB, West Burma; WC, Western Cimmerian Continent.

India–Australian continental shelf of Gondwana, and glacial–marine sediments (diamictites; pebbly mudstones interbedded with normal marine shales and sands) were deposited on the continental shelf of eastern Gondwana. Subduction beneath South China–Indochina in the Carboniferous led to the development of the Ailaoshan-Nan-Uttaradit back-arc basin (now represented by the Ailaoshan and Nan-Uttaradit Suture Zones) and separation of the Simao Terrane by back-arc spreading.

Early Permian (298–270 Ma) During the Permian, Australia remained in high southern latitudes. Glacial ice continued to reach the marine environment of the north-eastern Gondwanan margin, and glacial–marine sediments continued to be deposited on the Sibumasu, Qiangtang, and Lhasa terranes (Figures 8, 11, and 12). Gondwanan cold-climate faunas and floras characterized the Sibumasu, Qiangtang, and Lhasa terranes at this time. In addition, the distinctive cool-water-tolerant conodont genus *Vjalovognathus* defines an eastern peri-Gondwanan cold-water province (Figure 11). Floral provinces are particularly marked at this time, and the distinctive Cathaysian (*Gigantopteris*) flora developed on North China, South China, Indochina, East Malaya, and West Sumatra, which were isolated within Tethys and located equatorially (Figure 13). During the late Early Permian (End-Sakmarian), the Cimmerian continental sliver separated from the north-eastern margin of Gondwana, and the Meso-Tethys Ocean opened by seafloor spreading between it and mainland Gondwana (Figure 11C).

Late Permian (270–252 Ma) By early Late Permian times the Sibumasu and Qiangtang terranes had separated from Gondwana, and the Meso-Tethys had opened between this continental sliver and Gondwana (Figure 11C). The Palaeo-Tethys continued to be destroyed by northwards subduction beneath Laurasia, North China, and the amalgamated South China–Indochina–East Malaya terranes. Following separation, and during their northwards drift, the Sibumasu and Qiangtang terranes initially developed a Cimmerian Province fauna and were then absorbed into the Cathaysian Province. North and South China begin to collide during the Late Permian, and a connection between mainland Pangaea and Indochina, via South and North China or via the western Cimmerian continent, is indicated by the occurrence of the genus *Dicynodon* in the Upper Permian of Indochina (Figure 11C). Sibumasu began to collide with Indochina and East Malaya in the Late Permian, and collision continued into the Early Triassic. Deformation associated with this event, and with the collision of North and south China, is known as the Indosinian Orogeny in South-east Asia and China.

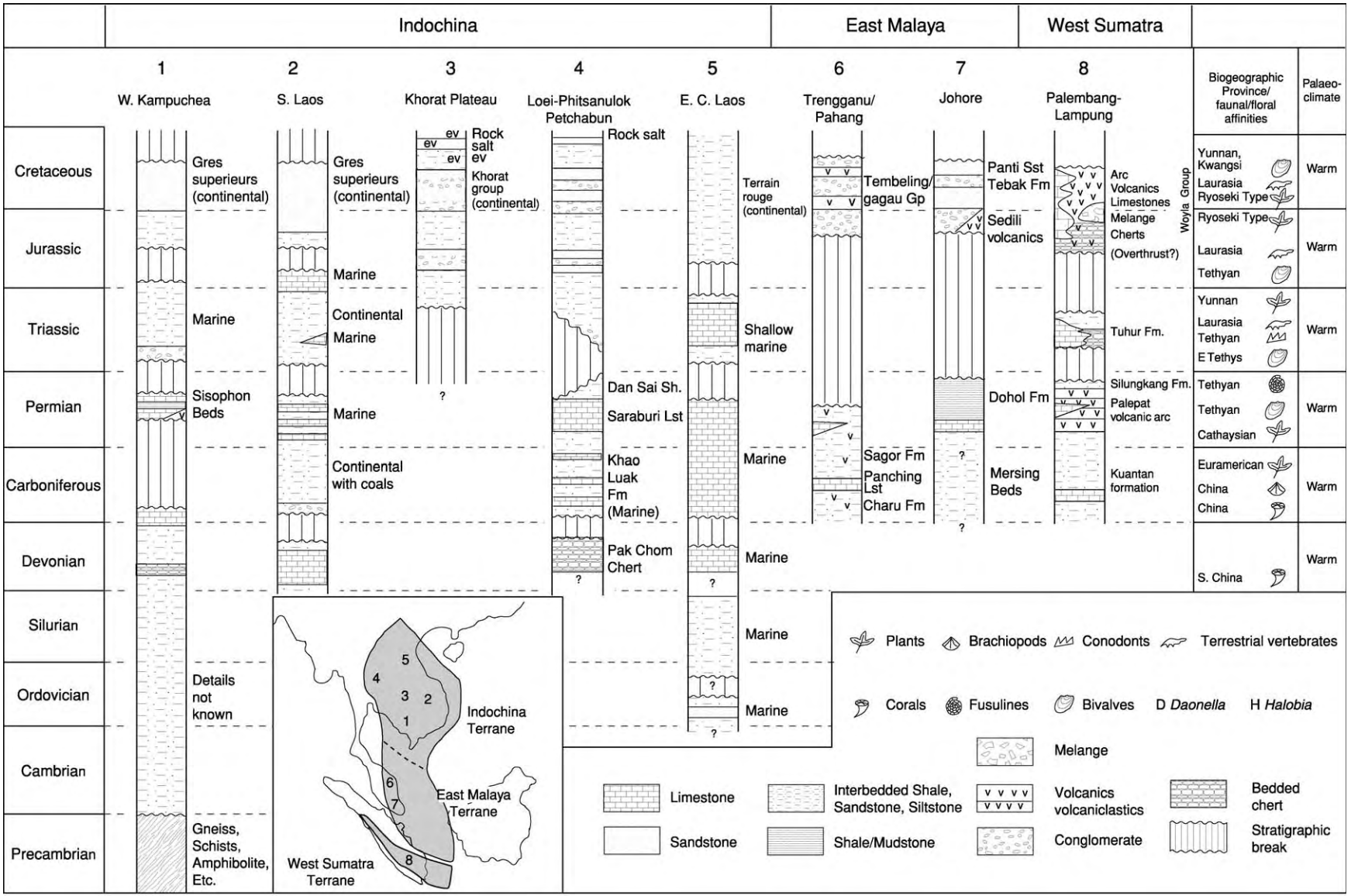


Figure 10 Representative stratigraphic columns for the Indochina, East Malaya, and West Sumatra terranes, and faunal provinces and affinities of faunas and floras. ev evaporites.

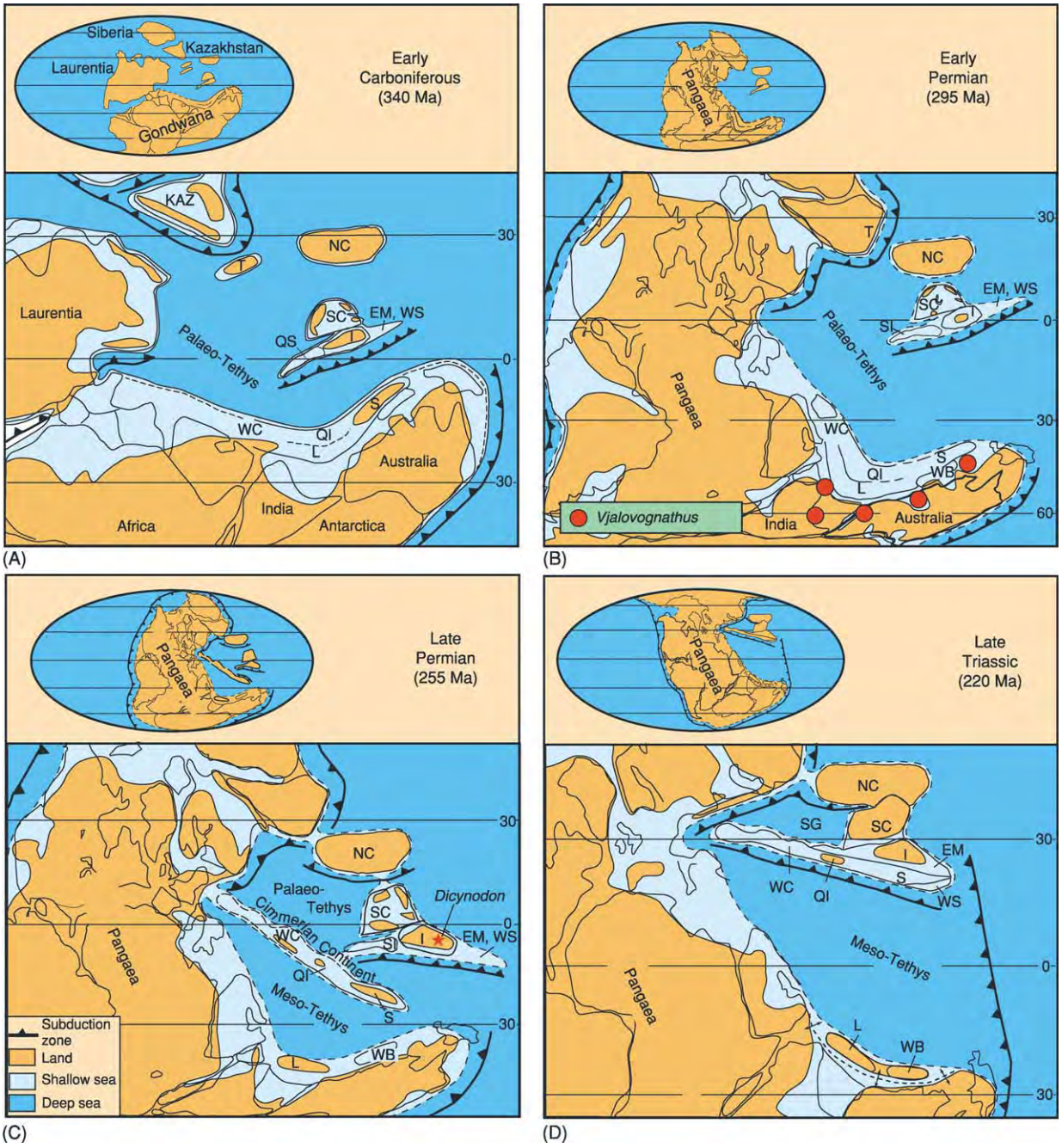


Figure 11 Palaeogeographical reconstructions of the Tethyan region for (A) the Early Carboniferous, (B) the Early Permian, (C) the Late Permian, and (D) the Late Triassic, showing the relative positions of the East and South east Asian terranes and the distribution of land and sea. Also shown are the distribution of the Early Permian cold water tolerant conodont, *Vjalovognathus*, and the Late Permian *Dicynodon* locality in Indochina. SC, South China; T, Tarim; I, Indochina; EM, East Malaya; WS, West Sumatra; NC, North China; SI, Simao; S, Sibumasu; WB, West Burma; QI, Qiangtang; L, Lhasa; WC, Western Cimmerian Continent; KAZ, Kazakhstan; QS, Qamdo Simao; SG, Songpan Ganzi.

Triassic (253–205 Ma) Australia was in low-to-moderate southern latitudes during the Triassic. The Sibumasu and Qiangtang terranes collided and sutured to the Indochina–South China amalgamated terrane. The West Sumatra Block was pushed westwards by the interaction of the westwards-subducting Palaeopacific Plate with the northwards-subducting Palaeo-Tethys during the Sibumasu–Indochina–East Malaya collisional process and was translated along major strike-slip faults to a position outboard of Sibumasu in the Early Triassic. The North–South China collision was nearly complete, and ultrahigh-pressure

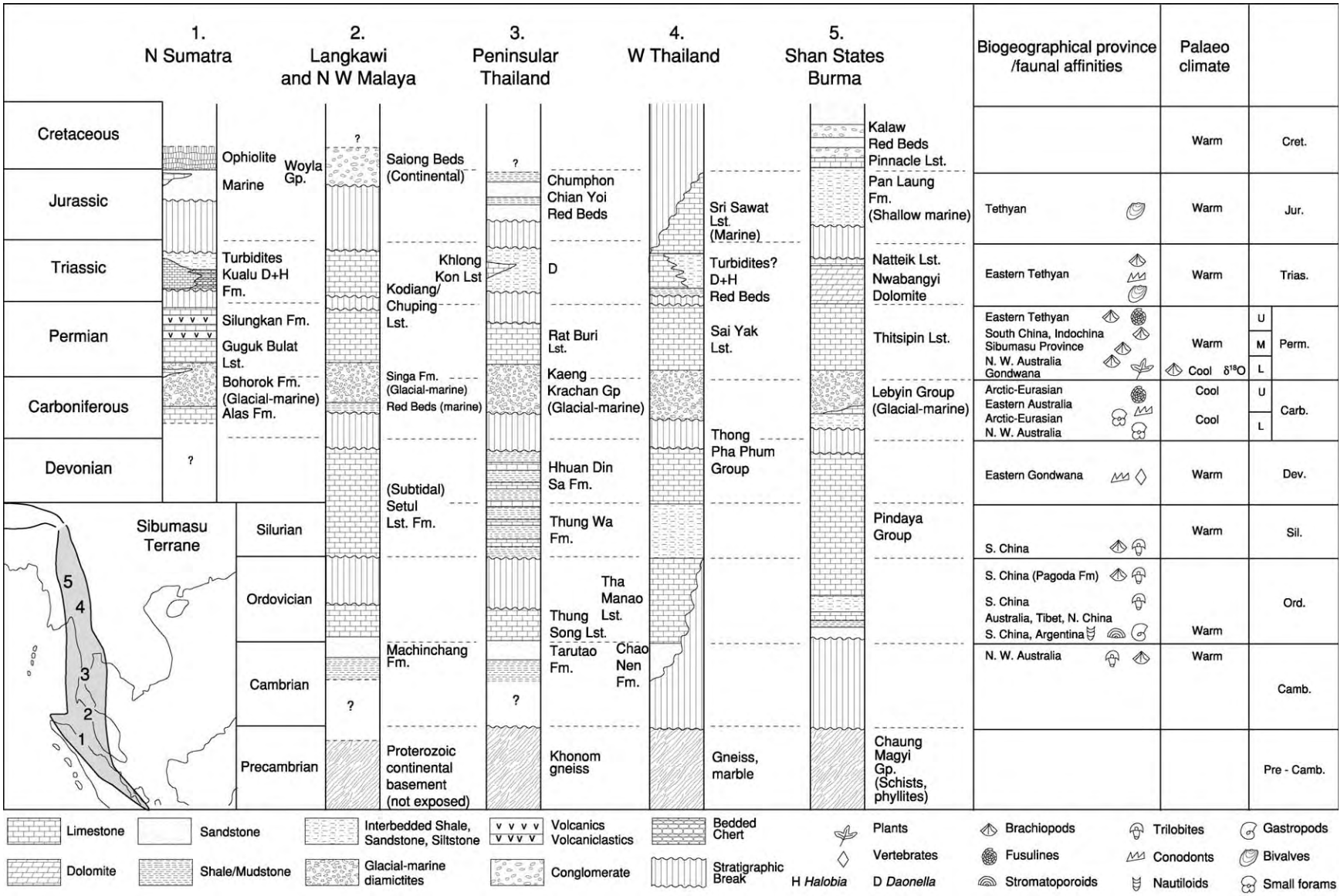


Figure 12 Representative stratigraphic columns for the Sibumasu Terrane, and faunal provinces and affinities of faunas and floras.

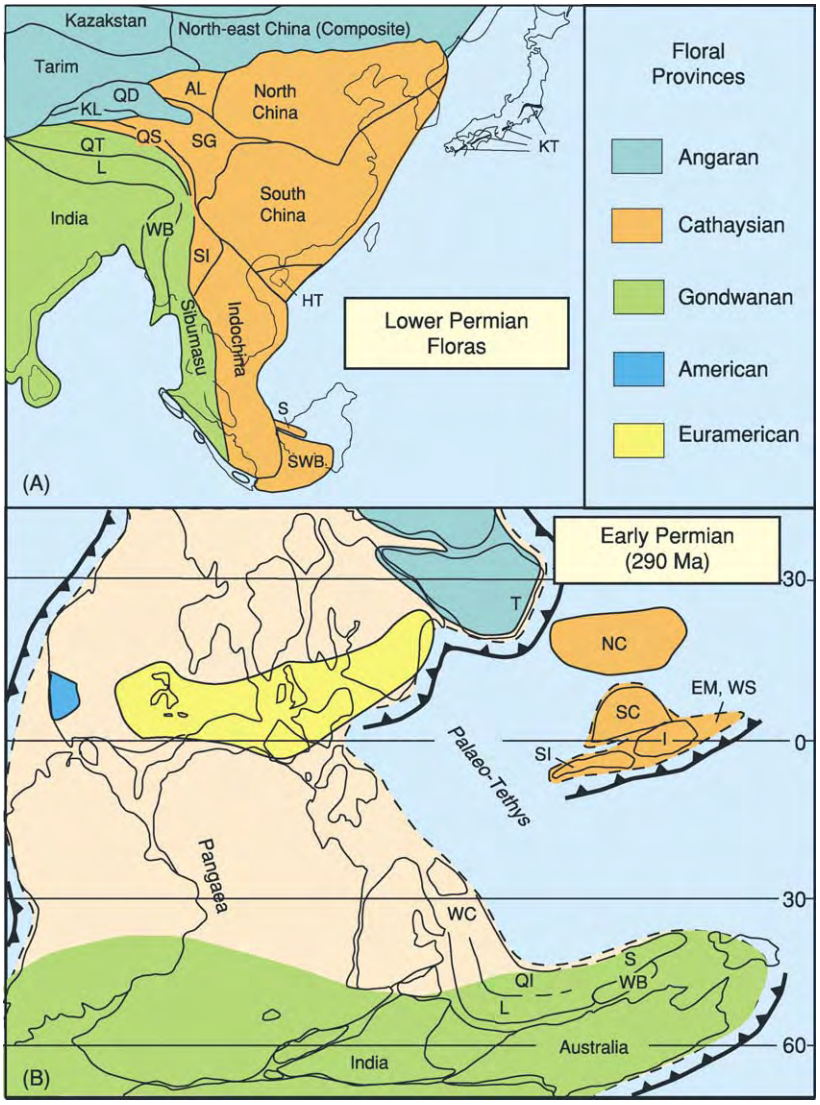


Figure 13 The distribution of Early Permian floral provinces plotted on (A) a present day geographical map, and (B) an Early Permian palaeogeographical map. (A) KL, Kunlun; QD, Qaidam; AL, Ala Shan; QT, Qiantang; L, Lhasa; QS, Qamdo Simao; SG, Songpan Ganzi; WB, West Burma; SI, Simao; HT, Hainan Island Terranes; S, Semitau; SWB, South West Borneo. (B) T, Tarim; NC, North China; SI, Simao; SC, South China; I, Indochina; EM, East Malaya; WS, West Sumatra; WC, Western Cimmerian Continent; QI, Qiangtang; S, Sibumasu; L, Lhasa; WB, West Burma.

metamorphics were exhumed along the Qinling-Dabie suture zone. Sediment derived from the North–South China collisional orogen poured into the Songpan Ganzi accretionary-complex basin producing huge thicknesses of flysch turbidites. The Ailaoshan-Nan-Uttaradit back-arc basin was closed when the Simao Terrane collided with South China–Indochina in the Middle to early Late Triassic. By Late Triassic (Norian) times, the North China, South China, Sibumasu, Indochina, East Malaya, West Sumatra, and Simao terranes had coalesced to form proto-East and South-east Asia (Figure 11D). During the final collisional consolidation of these terranes, the major economically important Late Triassic–Early Jurassic collisional tin-bearing Main Range granitoids were formed in South-east Asia (Figure 14).

Jurassic (205–141 Ma) Australia remained in low-to-moderate southern latitudes in the Jurassic. Rifting and separation of the Lhasa, West Burma, Sikuleh, Mangkalihat, and West Sulawesi terranes from north-western Australia occurred progressively from west to east during the Late Triassic to Late Jurassic. The Ceno-Tethys Ocean opened behind these terranes as they separated from Gondwana (Figure 15). Final welding of North China to Eurasia (the Yanshanian

deformational orogeny) took place in the Jurassic with the closure of the Mongol-Okhotsk Ocean. Initial pre-breakup rifting of the main Gondwanan supercontinent also began in the Jurassic.

Cretaceous (141–65 Ma) The Lhasa Block collided and amalgamated with Eurasia in latest Jurassic–earliest Cretaceous times. Gondwana broke up and India drifted north, making initial contact with Eurasia at the end of the Cretaceous (Figure 15). This early contact between India and Eurasia is indicated by palaeomagnetic data from the Ninetyeast Ridge and also by Late Cretaceous biogeographical links, including frogs and other vertebrates. The small West Burma and Sikuleh terranes also accreted to Sibumasu during the Cretaceous. Australia began to separate from Antarctica and drift northwards, but a connection with Antarctica via Tasmania remained.

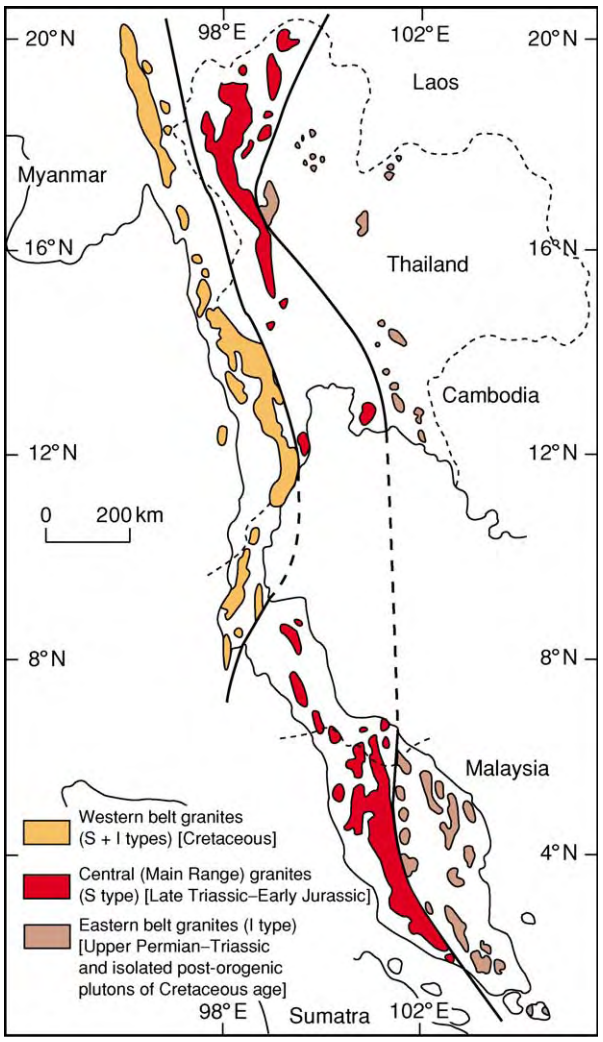


Figure 14 The three granitoid belts (provinces) of South east Asia.

Cenozoic (65–0 Ma) The Cenozoic evolution of East and South-east Asia involved substantial movements along and rotations of strike-slip faults, rotations of continental blocks and oceanic plates, the development and spreading of 'marginal' seas, and the formation of important hydrocarbon-bearing sedimentary basins. This evolution was essentially due to the combined effects of the interactions of the Eurasian, Pacific, and Indo-Australian plates and the collisions of India with Eurasia and of Australia with South-east Asia.

Various tectonic models have been proposed for the Cenozoic evolution of the region, which invoke different mechanisms for the India–Eurasia collision and different interpretations of rotations of continental blocks and oceanic plates. Three basic mechanisms have been proposed for the northwards motion of India into Eurasia. The first involves underthrusting of greater India beneath Eurasia; the second involves crustal shortening and thickening; and the third involves major eastwards lateral extrusion and progressive clockwise rotations of China, Indochina, and Sundaland.

Cenozoic clockwise rotations of crustal blocks are observed in Indochina and western Thailand, but major progressive counterclockwise rotations are seen in Borneo and the Malay Peninsula. The counterclockwise rotations observed in Borneo and Malaya are at variance with the extrusion model. Most models for the region neglect the clockwise rotation of the Philippine Sea Plate and/or the counterclockwise rotation of Borneo. The Cenozoic reconstruction model proposed by Robert Hall and illustrated here (Figures 16, 17, and 18), which takes into account both the clockwise rotation of the Philippine Sea Plate and the counterclockwise rotation of Borneo and Peninsular Malaysia, is preferred. Knowledge of the Cenozoic geodynamics of South-east Asia, the movements of microcontinents, and the shifting distribution of land and sea in the area (Figure 19) underpins and enhances our understanding of the biogeography of the region.

South-East Asian Geological Resources

Oil and Gas

Significant oil and gas accumulations occur widely in East and South-east Asia, generally in Cenozoic sedimentary basins (Figure 20), and these have contributed markedly to the economies of South-east Asian countries. The oil and gas accumulations are commonly associated with rocks of Middle and Upper Miocene age, with locally significant Oligocene and

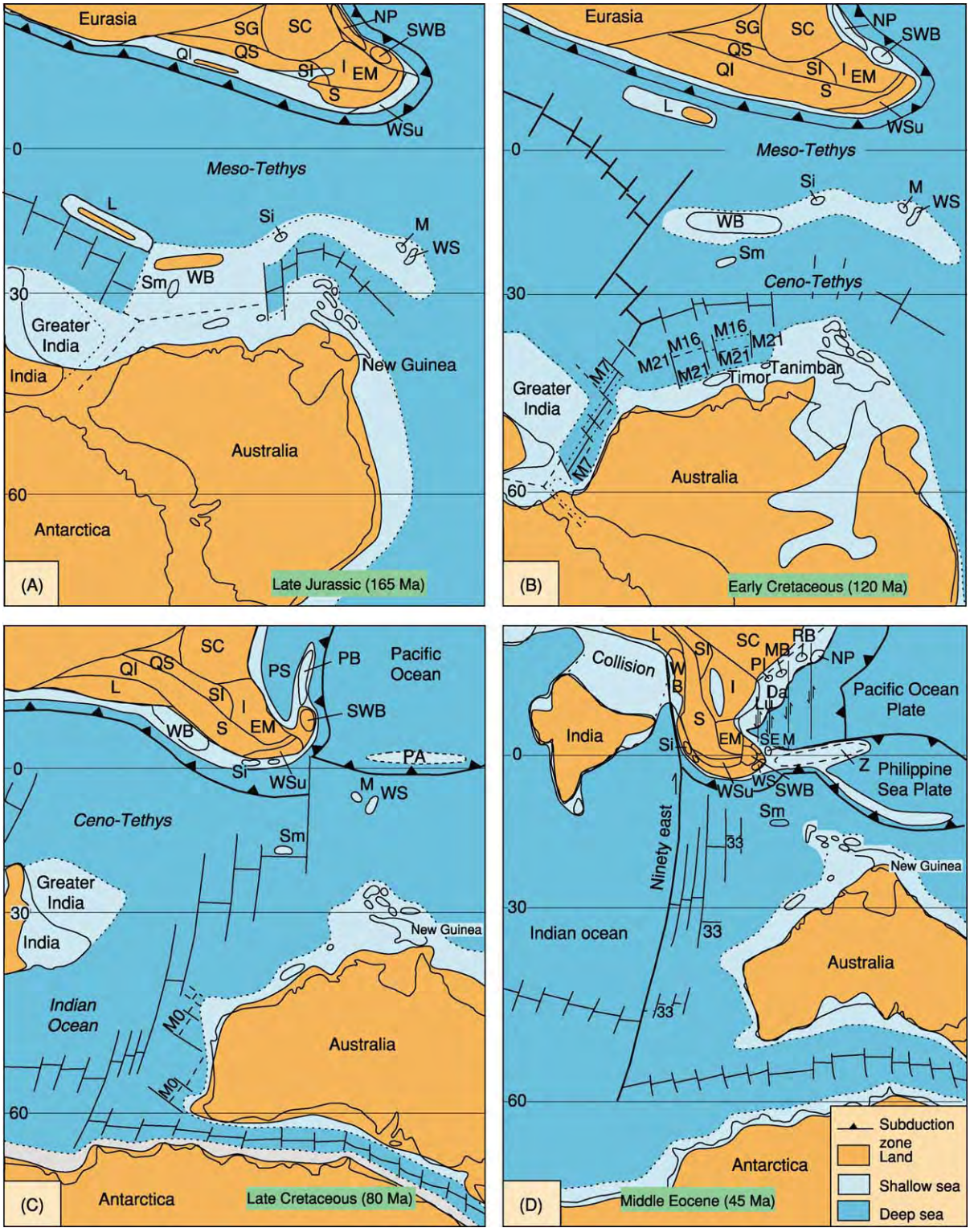
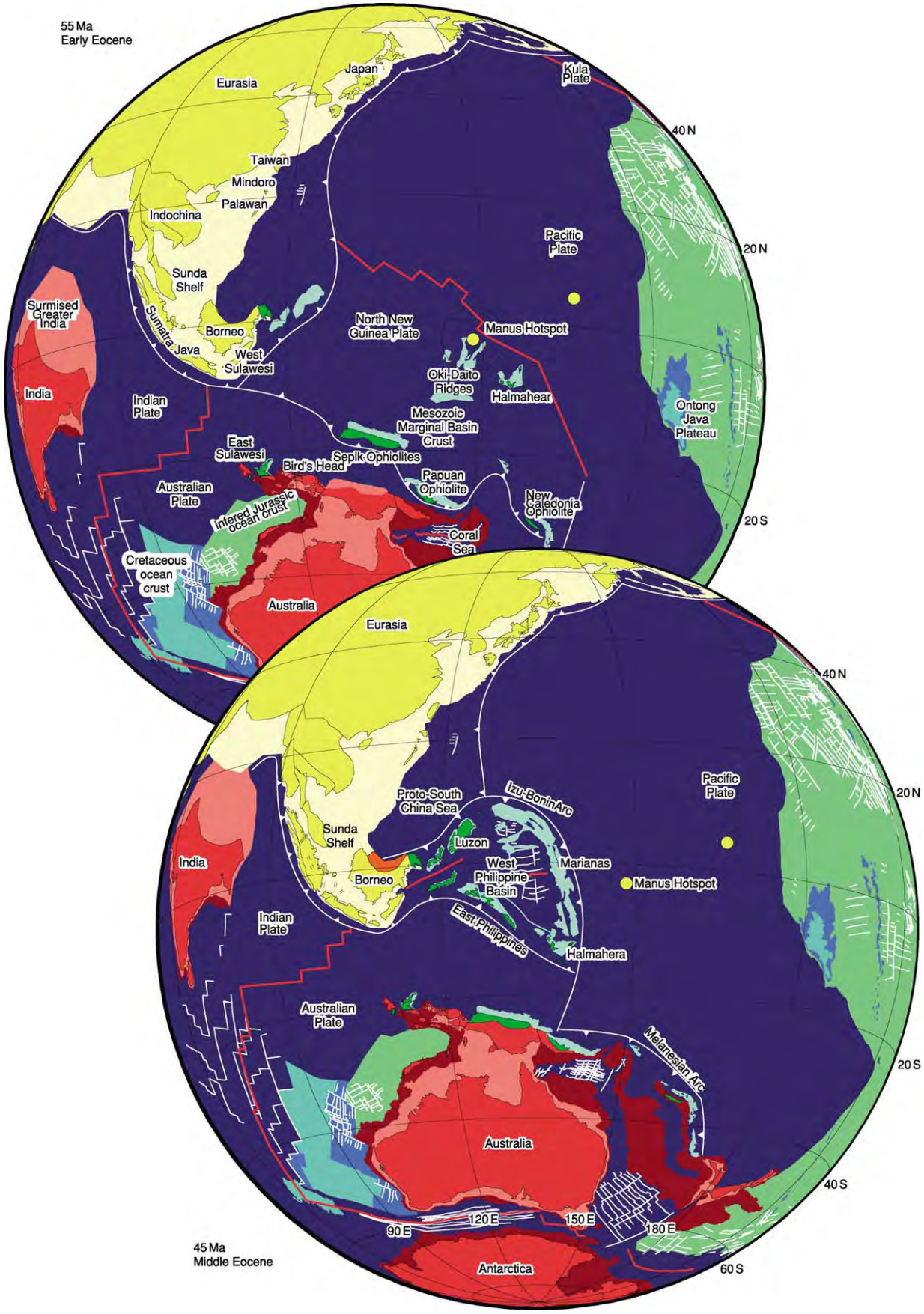


Figure 15 Palaeogeographical reconstructions for Eastern Tethys in (A) the Late Jurassic, (B) the Early Cretaceous, (C) the Late Cretaceous, and (D) the Middle Eocene, showing the distribution of continental blocks and fragments of South east Asia Australasia and the areas of land and sea. SG, Songpan Ganzi accretionary complex; SC, South China; QS, Qamdo Simao; SI, Simao; QI, Qiangtang; S, Sibumasu; I, Indochina; EM, East Malaya; WSu, West Sumatra; L, Lhasa; WB, West Burma; SWB, South West Borneo; SE, Semitau; NP, North Palawan and other small continental fragments now forming part of the Philippines basement; Si, Sikuleh; M, Mangkalihat; WS, West Sulawesi; PB, Philippine Basement; PA, Incipient East Philippine Arc; PS, Proto South China Sea; Z, Zambales Ophiolite; RB, Reed Bank; MB, Macclesfield Bank; PI, Paracel Islands; Da, Dangerous Ground; Lu, Luconia; Sm, Sumba. M numbers represent Indian Ocean magnetic anomalies.


55 Ma
Early Eocene
Eurasia
Japan
Kula Plate
40 N
Taiwan
Mindoro
Palawan
Indochina
Pacific Plate
20 N
Sunda Shelf
Surmised Greater India
Sumatra
Java
Borneo
West Sulawesi
North New Guinea Plate
Manus Hotspot
Oki-Daito Ridges
Halmahear
India
Indian Plate
Mesozoic Marginal Basin Crust
Ontong Java Plateau
East Sulawesi
Sepik Ophiolites
Bird's Head
Papuan Ophiolite
New Caledonia Ophiolite
Australian Plate
Infered Jurassic ocean crust
Coral Sea
20 S
Cretaceous ocean crust
Australia
Eurasia
40 N
Pacific Plate
20 N
Proto-South China Sea
Izu-Bonin Arc
Sunda Shelf
Luzon
West Philippine Basin
Marianas
India
Borneo
Manus Hotspot
Indian Plate
East Philippines
Halmahera
Australian Plate
Melanesian Arc
20 S
Australia
40 S
90 E
120 E
150 E
180 E
45 Ma
Middle Eocene
Antarctica
60 S

Pliocene occurrences. Some older Mesozoic accumulations do occur (e.g. in North Thailand), but, in general, the petroleum industry of the region is almost exclusively concerned with Tertiary sedimentary basins, and the pre-Tertiary is regarded as economic basement. However, in some cases, oil has migrated laterally and accumulated in fractured granitoids and other basement rocks, including Triassic limestones in the Gulf of Thailand.

Oil- and gas-bearing basins Space constraints preclude a full detailed description of the basins and hydrocarbon fields. A number of attempts have been made to classify South-east Asian hydrocarbon-bearing basins – particularly the Cenozoic basins of the region – genetically, but these attempts have failed owing to the complex and changing pattern of compressional and extensional tectonics in South-east Asia. Basins that were previously regarded as 'typical' back-arc basins are now being interpreted by some workers as the result of major strike-slip faults. Most basins were initiated in the Eocene or Oligocene, following a major Eocene break in sedimentation. Interpretation of the genesis of the South-east Asian basins is somewhat model-dependent, and there are competing models. Whichever model one favours, there has certainly been major strike-slip faulting in the region during the Cenozoic. This strike-slip faulting can clearly be related to basin formation, for example in North and Central Thailand and the Gulf of Thailand. However, the sense and displacement of many (if not most) of these strike-slip faults are poorly known. Other basins in the region are clearly related to plate convergence and subduction; for example, the North, Central, and South Sumatra basins and the Barito Basin of Borneo appear to be back-arc basins behind the arc developed on Sundaland as the India–Australian Plate was subducted northwards. The Palawan and Sabah basins could be classified as fore-arc basins. Other basins in the region, not related genetically to strike-slip faulting or to subduction processes, have been classified variously as continental failed rifts (aulacogens), cratonic basins, or basins that have formed on or between continental fragments.

Minerals

The distribution of the principal mineral deposits of South-east Asia is shown in Figure 21. Mineralization in South-east Asia is primarily associated with the ophiolites, volcanic arcs, and granitoid plutons of the region.

Mineralization associated with ophiolites Chromites are found in economic concentrations in ophiolites derived from the marginal-basin lithosphere of the Celebes, Sulu, and South China seas. Nickel sulphides and platinum also occur in dunite and serpentinite, and remobilized nickel also occurs as sulphides in veins in andesitic rocks. Deep tropical weathering of ophiolite ultramafics has resulted in nickel-bearing laterites that can be valuable ores (with a nickel oxide content of around 2.7%). Other important mineral deposits associated with ophiolites include manganese and Cyprus-type copper on pillow lavas and Besshi-type copper–iron massive sulphide deposits.

Mineralization associated with volcanic arcs Volcanic arcs are characterized by a wide range of mineralization types, ranging from epithermal vein deposits associated with near-surface fracture systems to higher-temperature vein and dissemination deposits associated with epizonal plutons. Ore deposits include porphyry copper, Kuroko-type copper–lead–zinc, and gold and gold–silver epithermal deposits.

Non-volcanic epithermal deposits The non-volcanic epithermal deposits of the region comprise stibnite, stibnite–gold, and stibnite–gold–scheelite mineralization, which is confined to 'Sundaland', i.e. the continental cratonic core of South-east Asia. The gold association is quite different from that of the Cenozoic volcanic arcs of the region and is confined to areas characterized by high-level felsic-to-intermediate plutons. There is commonly an important mercury association. This mineralization is younger than, and unrelated to, the tin-bearing granites of the region. Important deposits occur in western Borneo, Palawan, Sulawesi, the Malay Peninsula, Thailand, and Myanmar.

Tungsten deposits South-east Asia is a major tin–tungsten metallogenic province. The tungsten mineralization in South-east Asia is spatially and genetically related to the three main granite provinces of the region (Figure 14), and four types of deposit occur: hydrothermal quartz veins, scheelite skarn deposits, wolframite–scheelite–sulphide veins, and placer deposits (eluvial or marginally alluvial).

Figure 16 Plate reconstructions for South east Asia Australasia at 55 Ma (Early Eocene) and 45 Ma (Middle Eocene). The recon struction at 55 Ma precedes the collision of India with Eurasia. The reconstruction at 45 Ma coincides with a major period of plate reorganization. (Reproduced with permission from Hall R (2002) Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer based reconstructions and animations. *Journal of Asian Earth Sciences* 20: 353 434.)

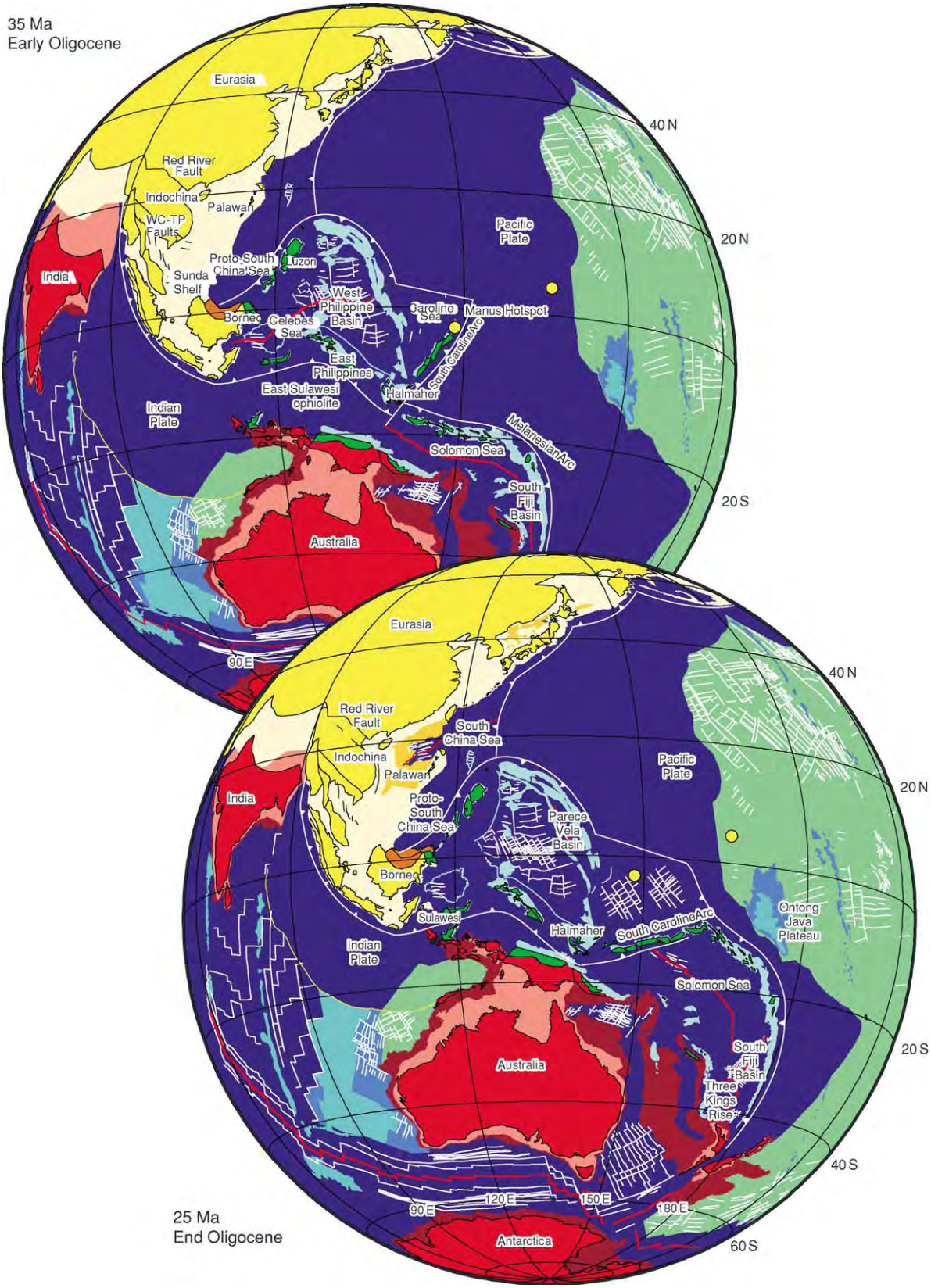


Figure 17 Plate reconstructions for South east Asia Australasia at 35 Ma (Early Oligocene) and 25 Ma (End Oligocene). (Repro duced with permission from Hall R (2002) Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer based reconstructions and animations. *Journal of Asian Earth Sciences* 20: 353 434.)

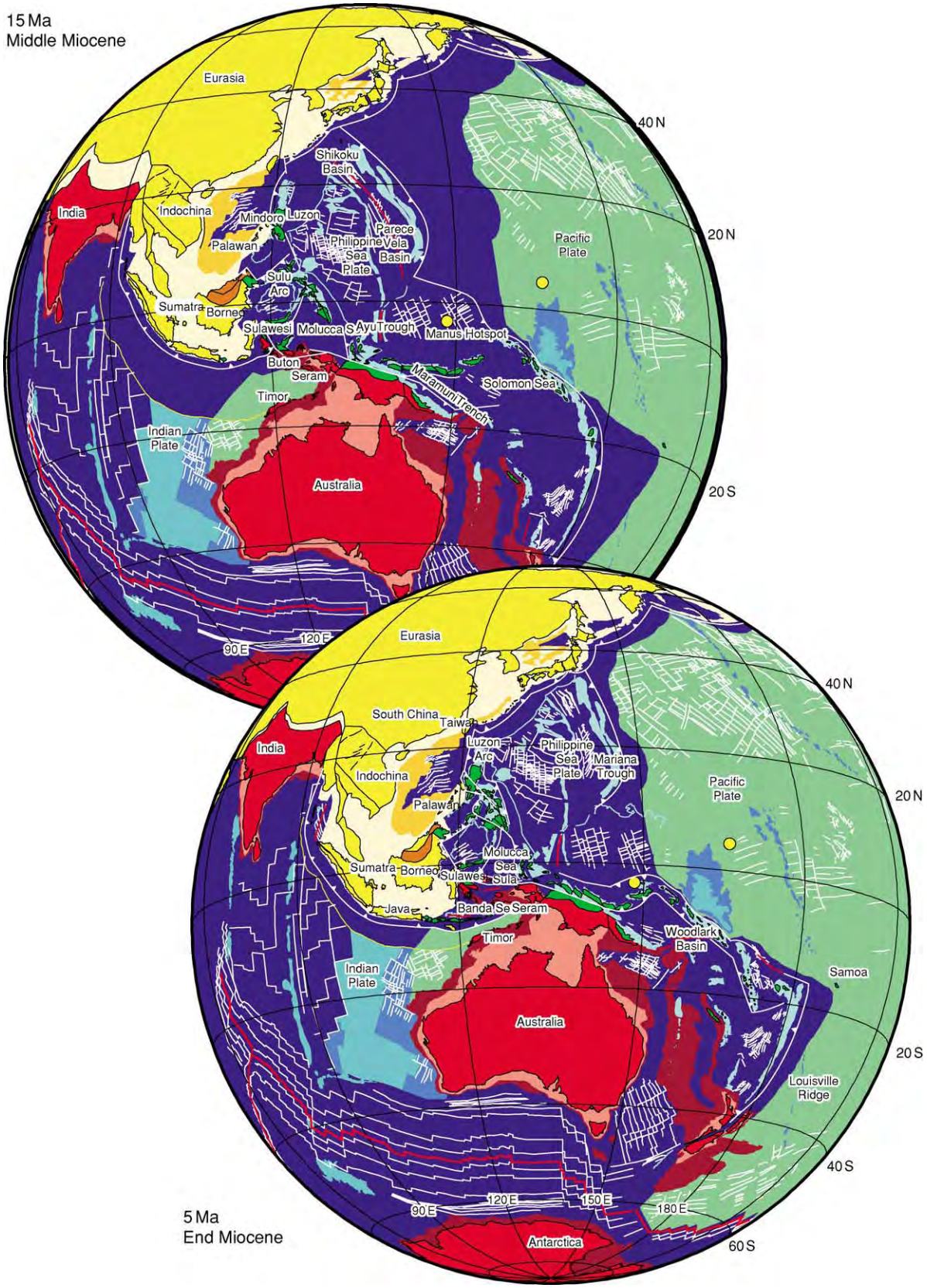


Figure 18 Plate reconstructions for South east Asia Australasia at 15 Ma (Middle Miocene) and 5 Ma (End Miocene). (Reproduced with permission from Hall R (2002) Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer based reconstructions and animations. *Journal of Asian Earth Sciences* 20: 353 434.)

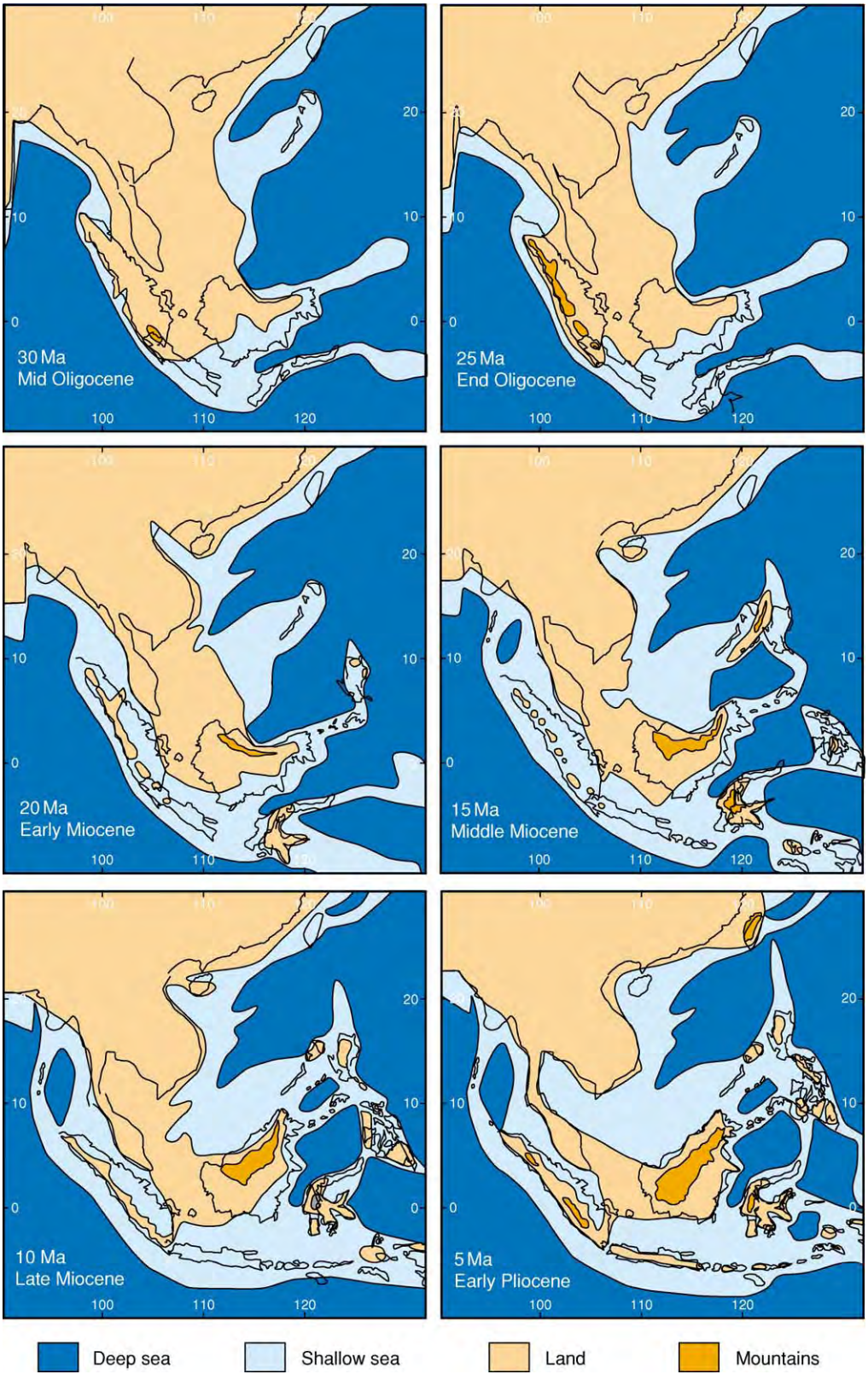


Figure 19 Reconstructions of East and South east Asia for the last 30 Ma, showing the variation in the distributions of deep sea, shallow sea, land, and mountains.

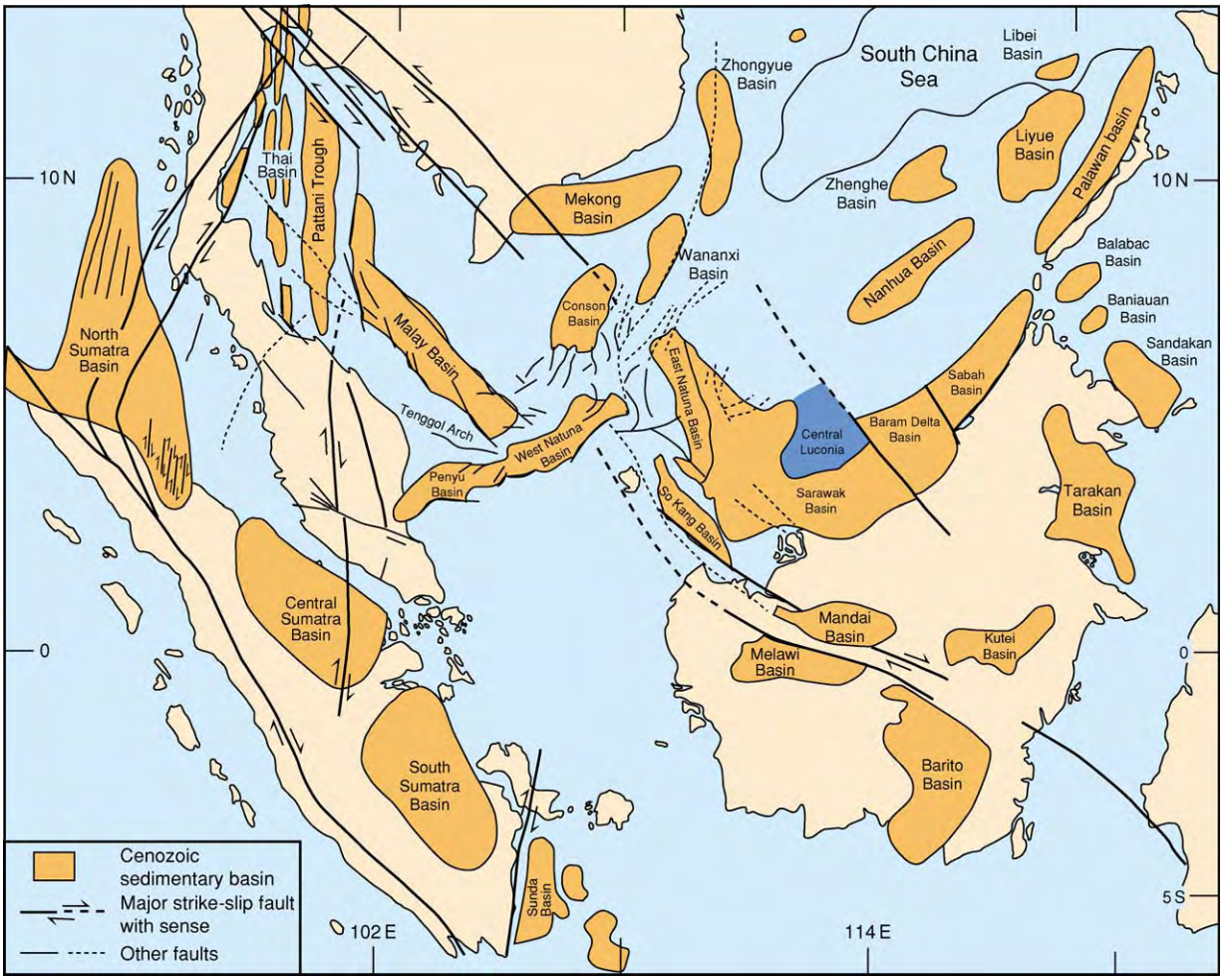


Figure 20 Oil and gas bearing Cenozoic sedimentary basins and major fault zones of South east Asia.

Tin deposits During the twentieth century, 70% of the tin mined in the world came from the South-east Asian tin belt. The tin mainly occurs as cassiterite ore in Late Cenozoic alluvial placer deposits that form part of the 400 km wide South-east Asian tin belt, which extends 2800 km from Burma and Thailand through Peninsular Malaysia to the Indonesian tin islands of Bangka and Belitung (Figure 21). Historically, these deposits have been extremely important to the economic development of Malaysia and South-east Asia, and, altogether, 9.6 million tonnes of tin (54% of the world's total tin production) has come from this tin belt. The tin deposits of the region are associated with the major granite/granitoid bodies of South-east Asia and formed during the emplacement of these intrusive igneous bodies, which were produced by major plate-tectonic processes as the continental blocks of the region converged and collided with each other. The South-east Asian granites form three major north–south belts or provinces: a Western Granitoid Province of Cretaceous–Cenozoic (149–22 Ma) age, a Central Granitoid Province of mainly Late Triassic and Early Jurassic (230–184 Ma) age, and an Eastern Granitoid Province of Permian, Triassic, and Early Jurassic (257–197 Ma) age (Figure 14). The Eastern Province was formed by subduction on the western margins of the Indochina and East Malaya terranes during convergence with the Sibumasu terrane. When these two continental terranes collided, the Central Province granites were generated. The younger Western Province granites were principally formed when the West Burma terrane collided with Sibumasu and by subsequent subduction processes. The Main Range and Eastern Belt granites of Peninsular Malaysia form parts of the Central and Eastern Granitoid Provinces, respectively. The tin mineralization occurred within the granites themselves and more especially in the country rocks that immediately surrounded them. Subsequent to mineralization, the granites and the country rocks into which they were intruded were uplifted and eroded, and the tin ore was deposited and concentrated in river alluvium.

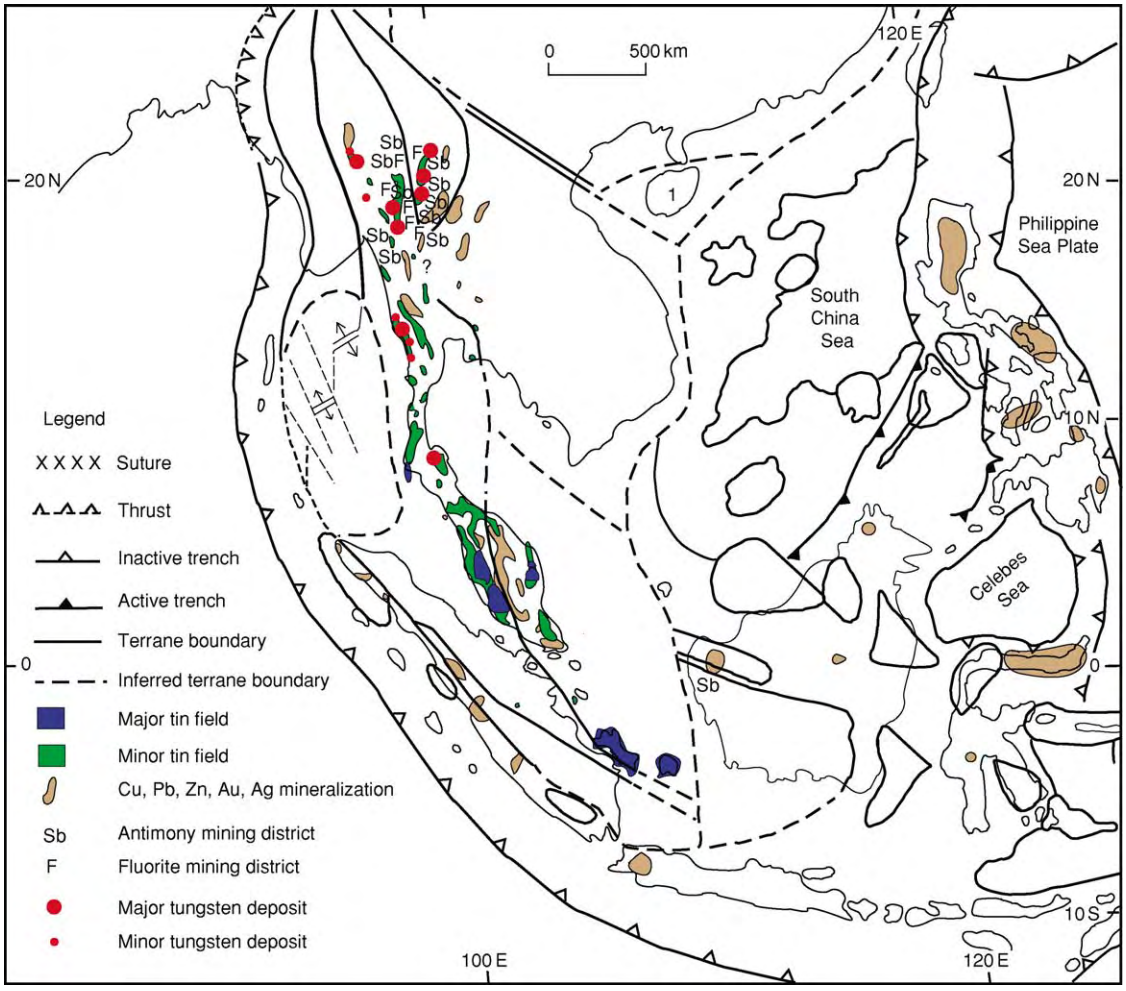


Figure 21 Distribution of the principal mineral deposits of South east Asia.

In addition to placer tin deposits there are also a wide variety of primary tin deposits. However, only two major primary tin fields have led to significant deep-mine production: the Sungei Lembing Mine near Kuantan, Pahang, Peninsular Malaysia, and Kelapa Kampit on Billiton Island. The lodes of these mines are developed in contact-metamorphosed Carboniferous–Permian argillaceous sedimentary and andesitic tuffaceous rocks within aureoles around high-level Permian–Triassic granitoids.

Mississippi Valley-type epigenetic deposits Stratiform deposits of predominantly galena and sphalerite are most commonly hosted in shallow-marine carbonate-dominated sequences on the rims of large sedimentary basins where changes in salinity caused deposition. The continental regions of South-east Asia, with their extensive Palaeozoic carbonate platforms, are suited to Mississippi Valley-type epigenetic deposits, and some have been identified. One of the best-known deposits in the region, now considered to be of this type, is at the Bawdwin lead–zinc mine north-north-east of Rangoon, Myanmar. Mining here can be traced back to 1412.

Iron ore deposits Iron ore deposits are widespread in the region, but significant production has been restricted to three areas: South-east China (Hainan, Kwangtung, Fukien), the Malay Peninsula, and the Philippines. The deposits are predominantly of contact pyrometasomatic type and are associated with plutons emplaced at a wide range of depths. Peninsular Malaysia was a major producer of iron ore, with production reaching 7 million tonnes in 1964. However, all major deposits have now been worked out.

Coal and lignite deposits South-east Asian coal deposits are almost entirely confined to deposits of subbituminous and lignitic coal in relatively small Tertiary terrestrial basins. There are minor occurrences of

bituminous coal and anthracite in Palaeozoic and Mesozoic formations of the older Sundaland continental core of the region, but these are generally uneconomic except for deposits of Triassic anthracite mined in northern Vietnam.

Diamonds, sapphires, and other gems Alluvial diamonds with no obvious sources ('headless placers') are found in several areas of South-east Asia, including Myanmar, southern Thailand (Phuket), Sumatra, and Kalimantan (Figure 8). These deposits occur in relatively young geological terranes, in contrast to the Archaean or Proterozoic terranes that host most primary diamond deposits and their associated alluvial workings.

Significant quantities of diamond have been recovered from two areas in Myanmar: Momeik in the northern part of the country and Theindaw in the southern part. The Momeik diamonds are recovered during mining of gemstone gravels; the Theindaw and Phuket diamonds are by-products of tin dredging. These sites, and those in Sumatra, lie within the Sibumasu terrane, which was detached from north-western Australia in Palaeozoic times and drifted northwards to become part of South-east Asia. The Myanmar–Thailand diamonds may be derived either from lamproitic intrusives within this terrane or secondarily from Permian glacial–marine sediments. The Kalimantan diamonds lie in another terrane, with a different origin, and, like eastern Australian diamonds, they may be directly or indirectly related to subduction processes.

Sapphire, ruby, zircon, and spinel gemstones and jade are worked at a number of localities in South-east Asia from alluvial and eluvial gravels derived from nearby outcrops of extensive Pliocene–Pleistocene high-alkali basalt fields. The basalts are usually basinite and nephelinite. The main centres for the mining of these gemstones are at Pailin, Bokeo, and Rovieng in Cambodia, Chanthaburi, Trat, and Kancharaburi in Thailand, south of Haikou on Hainan Island, in the Mogok and Lonkin areas of Myanmar, and in Taiwan. Most of the South-east Asian rubies come from the Trat area of Thailand, which supplies 70% of the world's production. Ruby, sapphire, and spinel are mined in the Mogok area of Myanmar from placers and eluvial deposits derived from marbles and calc-silicate migmatites associated with alaskite (Mogok Gneiss). Jadeite is mined in the Lonkin area of Myanmar, which is the world's foremost source of high-grade jadeite.

See Also

Asia: Central. **Australia:** Phanerozoic. **China and Mongolia**. **Gondwanaland and Gondwana**. **Palaeoclimates**. **Pangaea**. **Plate Tectonics**. **Precambrian:** Overview. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mountain Building and Orogeny.

Further Reading

Fraser AJ, Matthews SJ, and Murphy RW (eds.) (1997) *Petroleum Geology of Southeast Asia. Special Publication 126*. London: Geological Society.

Hall R and Blundell D (eds.) (1996) *Tectonic Evolution of Southeast Asia*. Special Publication 106. London: Geological Society.

Hall R and Holloway JD (eds.) (1998) *Biogeography and Geological Evolution of South East Asia*. Amsterdam: Backhuys Publishers.

Hamilton W (1979) *Tectonics of the Indonesian Region*. USGS Professional Paper 1078. Washington: US Geological Survey.

Hutchison CS (1989) *Geological Evolution of South East Asia*. Oxford: Clarendon Press.

Hutchison CS (1996) *South East Asian Oil, Gas, Coal and Mineral Deposits*. Oxford: Clarendon Press.

Meng LK (ed.) The Petroleum Geology and Resources of Malaysia. Kuala Lumpur. Petronal. P.

Metcalfe I (1988) Origin and assembly of South east Asian continental terranes. In: Audley Charles MG and Hallam A (eds.) *Gondwana and Tethys*, pp. 101 118. Special Publication 37. London and Oxford: Geological Society and Oxford University Press.

Metcalfe I (ed.) (1994) *Gondwana dispersion and Asian accretion Special Issue of Journal of Southeast Asian Earth Sciences* 9: 303 461.

Metcalfe I (ed.) (1999) *Gondwana Dispersion and Asian Accretion*. IGCP 321 Final Results Volume. Rotterdam: A A Balkema.

Metcalfe I and Allen MB (eds.) (2000) Suture zones of East and Southeast Asia Special Issue of *Journal of Asian Earth Sciences* 18: 635 808.

Metcalfe I, Smith JMB, Morwood M, and Davidson I (eds.) (2001) Faunal and floral migrations and evolution in SE Asia Australasia Lisse: A A Balkema.

Shi GR and Metcalfe I (eds.) (2002) Permian of Southeast Asia. Special Issue of *Journal of Asian Earth Sciences*, 20: 549 774.

ASTEROIDS

See **SOLAR SYSTEM: Asteroids, Comets and Space Dust**

ATMOSPHERE EVOLUTION

S J Mojzsis, University of Colorado, Boulder, CO, USA

Introduction

Earth is unique among the planets of the Solar System in having sustained conditions of temperature and pressure that permit stable liquid water at the surface for more than 4 Ga. The long-term stability of liquid water is fundamental to the origin and propagation of life, and Earth is exceptional among the known planets in having a biosphere. Abundant free oxygen accumulated in the atmosphere as a metabolic waste product of oxygenic photosynthesis; the high free-oxygen content contrasts markedly with the atmospheres of the neighbouring planets of the inner Solar System (Table 1). This globally oxidative condition was at first deleterious to microbial life, but led to the emergence of aerobic metabolisms, sexual reproduction, and multicellularity.

Appreciating how the current physical and chemical state of the atmosphere came to be requires knowledge of the initial conditions. The elemental ingredients that go into making a habitable world are ultimately derived from cosmologic nucleosynthesis – the formation of primarily hydrogen (^{1}H) and helium (^{4}He), with minor amounts of ^{2}D, ^{3}He, and ^{6}Li, at the Big Bang approximately 13 ± 1 Ga ago. All the other naturally occurring elements in the periodic table were created by stellar nucleosynthetic and supernova reactions. These events contributed gas and dust to the interstellar medium, and a cloud of such supernova ashes collapsed to form the solar nebula and, ultimately, the Sun and planets about 4.56 Ga ago (*see* **Solar System:** The Sun).

Table 1 Atmospheric constituents of the inner planets

	Earth	*Venus*	*Mars*
N_2^*	78.084%	3.5%	2.7%
O_2	20.946%	tr.	0.13%
H_2O	<4%	30 ppm	0.03%
Ar	9340 ppm	70 ppm	1.6%
CO_2	350 ppm	96.5%	95.32%
Ne	18.18 ppm	7 ppm	2.5 ppm
^{4}He	5.24 ppm	~12 ppm	nd
CH_4	1.7 ppm	nd	nd
Kr	1.14 ppm	~25 ppb	0.3 ppm
H_2	0.55 ppm	nd	nd
N_2O	320 ppb	nd	?
CO	125 ppb	17 45 ppm	0.08%
Xe	87 ppb	~1.9 ppb	0.08 ppm
O_3	10 100 ppb	nd	>0.04 ppm
HCl	1 ppb	~0.6 ppm	nd
SO_2	20 90 ppt	25 150 ppm	nd
H_2S	30 100 ppt	3 ppm	nd

*Abundance values are given for dry air. Abbreviations: tr., trace; ppm, 10^{-6} g g^{-1}; ppb, 10^{-9} g g^{-1}; ppt, 10^{-12} g g^{-1}; nd, not detected.

Origin of a Habitable World

Solar System and Planetary Formation: Relevance to Atmospheric Evolution

Following Big Bang nucleosynthesis, local concentrations of gas collapsed to form the first galaxies and the nebulae and stars that comprise them. Gravity binds these immense structures together, and it is the gravitational collapse when stars form that creates the pressures needed to initiate thermonuclear fusion reactions, which convert four ^{1}H nuclei to one ^{4}He nucleus and release energy. This energy provides an outward force that counters the inward gravitational collapse. Although seemingly violent and catastrophic in our Earth-bound (solid–liquid–gas) view of matter in the universe, such systems are remarkably stable and are sustained over many millions of years throughout the lifetime of a star on the main sequence. In smaller stars, such as the Sun, luminosity changes over the stellar lifetime, beginning with a short (less than 100 Ma) early phase of variable brightness, which lasts until the so-called main sequence of hydrogen fusion commences (Figure 1). Time in the main sequence is long for a star such as the Sun, of the order of 8 Ga, but it eventually draws to a close when much of the hydrogen fuel is exhausted. At that stage, the star expands and enters a

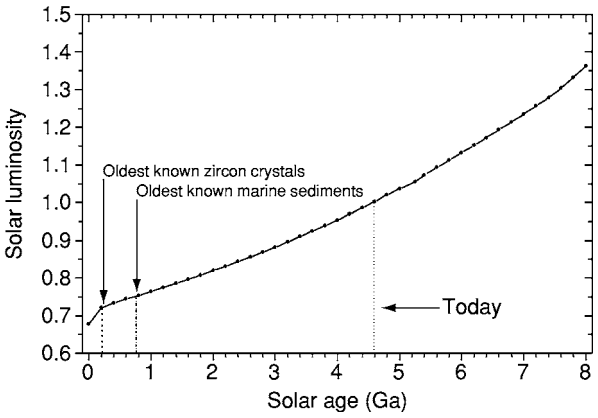


Figure 1 Secular changes in solar luminosity with time. Solar luminosity is expressed as fraction of current solar luminosity.

red giant phase, when helium is consumed to synthesize the isotopes ^{12}C and ^{16}O (via ^{8}Be). This is a rapid process in comparison to the length of time spent in the main sequence. When helium becomes depleted, the star collapses from a red giant into a white dwarf, which is dominantly composed of oxygen and carbon. Some of these stellar objects can continue to accrete mass from companion stars and become increasingly unstable, resulting in what are termed type-I supernovae. In the absence of a massive stellar companion, white dwarf stars slowly cool and darken over eons. The solar nebula of approximately 4.6 Ga ago was derived from the violent remains of perhaps several supernovae. There is evidence that more than one event occurred around the time of the collapse of the solar nebula, injecting it with short-lived nuclides such as ^{244}Pu, ^{60}Fe, and ^{26}Al, which were important heat-producing elements in the larger differentiated asteroids and proto-planets. During nebular collapse, the central region of the disk concentrated hydrogen and helium and eventually formed the Sun. The Sun entered its main sequence about 50 Ma after the initiation of cloud collapse. Solar radius and luminosity increased steadily as hydrogen was converted to helium; this is because, as the average density of the Sun increased, the rate of hydrogen fusion also increased.

For stars greater than ten times the solar mass, hydrogen fusion proceeds at higher temperatures and pressures. In such stars, time in the main sequence may be brief, as short as 10 Ma, before they enter the red giant phase. Since these stars burn hotter and at higher pressures, fusion of helium continues beyond ^{12}C to yield ^{20}Ne and ^{24}Mg, as well as beyond ^{16}O to create ^{32}S, ^{28}Si, and, finally, ^{56}Fe. Iron has the greatest binding energy per nucleon (Figure 2) and represents the limit of the energy-yielding thermonuclear-fusion products. When even this elemental fuel becomes exhausted, a super-massive star will rapidly collapse and can explode as a type-II supernova, one of the most exergonic events in the universe. Such an event was recently observed as SN 1987A, a supernova approximately 150 000 light-years distant, which proceeded with fusion beyond ^{56}Fe to create gold, lead, isotopes of thorium (^{232}Th) and uranium (^{235}U, ^{238}U), and transuranic elements up to ^{244}Pu and beyond. Supernova events randomly fertilize the galaxy with heavy nuclides, including the unstable radioactive heat-producing elements (potassium, uranium, and thorium) that are responsible for keeping planetary interiors warm and dynamic over geological timescales. It is estimated that about 50% of all interstellar dust is created in supernova explosions. Cosmic dust is the crucial raw material in the creation of a habitable world.

Formation of the Earth: Primary Atmosphere

A rapid accretion by asteroid-like bodies (termed 'planetesimals') formed the planets of the Solar System (*see* **Earth Structure and Origins**). A volatile envelope – captured from nebular gases and perhaps containing significant amounts of hydrogen, helium, methane, and ammonia – would have already started to accumulate when the growing proto-Earth reached a critical mass

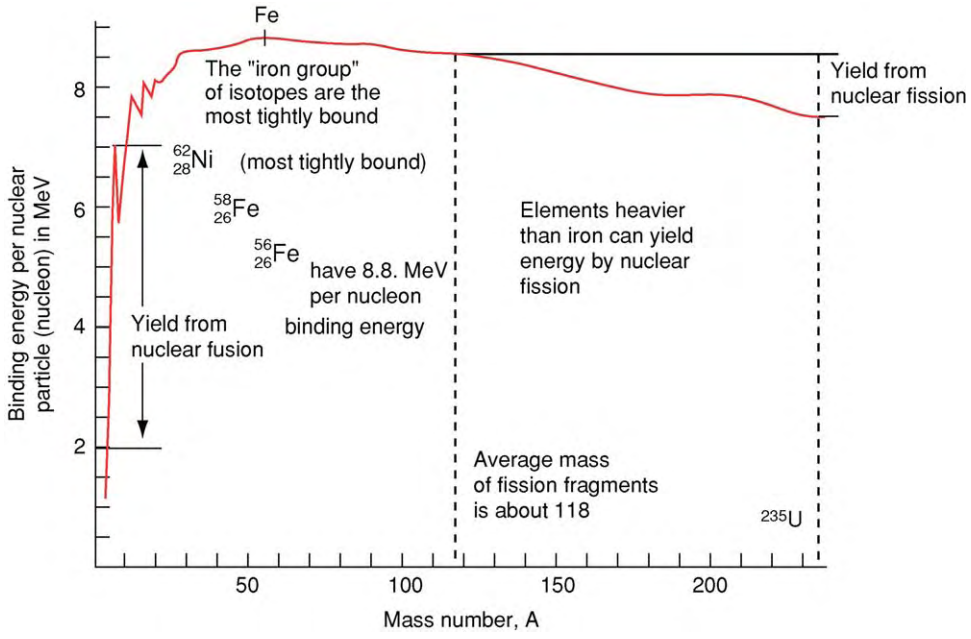


Figure 2 Binding energy curve of the nuclides.

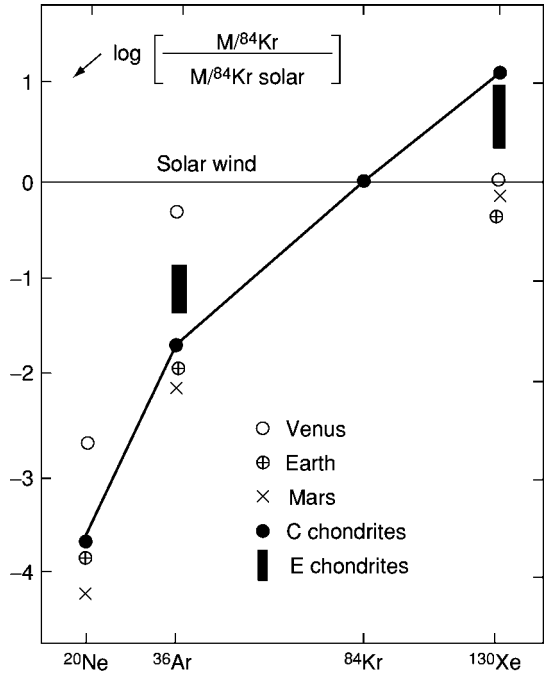


Figure 3 Noble gas abundances on Earth in comparison with cosmic abundances (adapted with permission from D. Hunten *et al.* in *Meteorites and the Early Solar System*, eds. John F. Kerridge and Mildred Shapley Matthews. © 1988 The Arizona Board of Regents).

of 10–15% of its present value. Whatever this primary atmosphere was, intense solar wind from the young Sun and later energetic impacts, culminating in the Moon-forming event, would have led to significant gas loss. Evidence for this can be seen in the concentrations of the noble gases (helium, neon, argon, krypton, xenon). Abundances of these chemically unreactive gases in Earth's atmosphere are far less than their solar or cosmic equivalents (Figure 3). This could be due to either solar wind or impact-induced erosion of planetary atmospheres (or both).

Formation of the Moon: Loss of the Primary Atmosphere

Evidence from a combination of lunar samples, meteoritics, interplanetary probes, isotopic studies, and physical models shows that the accretionary phase of the planets was intense and short-lived, lasting for less than 100 Ma after the initial collapse of the solar nebula. Under conditions of intense early bombardment, the conversion of the kinetic energy of millions of impacting planetesimals to thermal energy would have heated the proto-Earth to melting, further facilitating differentiation into core, mantle, crust, and atmosphere (*see* **Earth:** Mantle; Crust).

Sometime between 4.5 Ga and 4.45 Ga an object the mass of Mars entered an Earth-crossing orbit. The consequences of this event were profound. The giant impact hypothesis envisages total Earth melting, blow-out of the tattered remnants of primary atmosphere captured during accretion, and devolatilization of the material that coalesced to form the Moon (*see* **Solar System:** Moon).

Abatement and Cooling

As planetary accretion slowed, heat rapidly radiated from the surface into space, and the Earth's surface rapidly cooled, forming a solid chill crust. Outgassing due to the internal differentiation of the planet into core and mantle resulted in the formation of a dense atmosphere, probably composed of carbon dioxide, steam, and other gases. Condensation of oceans onto the primordial surface could then have occurred as a consequence of either intrinsic water from planetary outgassing or later contributions from comets and meteorites. Comets and meteorites are primitive undifferentiated objects rich in volatiles and reduced organic compounds – attractive procreative ingredients for the origin of life. Some postulate that all the water on the Earth's surface could have been delivered from comets without the need for outgassing of water vapour from the mantle. However, if the $^{2}D/^{1}H$ composition of the three comets measured so far is typical of these objects, the oceans could not have been produced solely by cometary water (Figure 4). To explain the data whilst still accepting the cometary origin of the hydrosphere, terrestrial water requires an additional component with a $^{2}D/^{1}H$ much less than that of average seawater. The weight of present evidence suggests that the likeliest source of this water is intrinsic planetary outgassing. However, it could be that prebiotic reactions on the early Earth leading to the origin of life required a significant contribution from comets, and several models predict that these contributions could have been large.

Liquid water, a source of energy, and organic raw materials are presumed to be necessary to initiate biological processes. Depending on the compositions of the earliest retained (secondary) atmosphere and hydrosphere, prebiotic molecules could have been synthesized in the atmosphere, formed at hydrothermal vents, or delivered by comets. The oceans were an early feature of the Earth. The oldest known rocks of sedimentary origin are >3.8 Ga ferruginous quartzites and banded iron formations from southern West Greenland and unequivocally support the presence of surface seas at that time. Rare zircons from the Jack Hills region of Western Australia are as old as 4.38 Ga and contain oxygen isotopic compositions that are consistent with the presence of substantial liquid water at or near the Earth's surface and interacting with the crust within 150 Ma of the origin of the Moon.

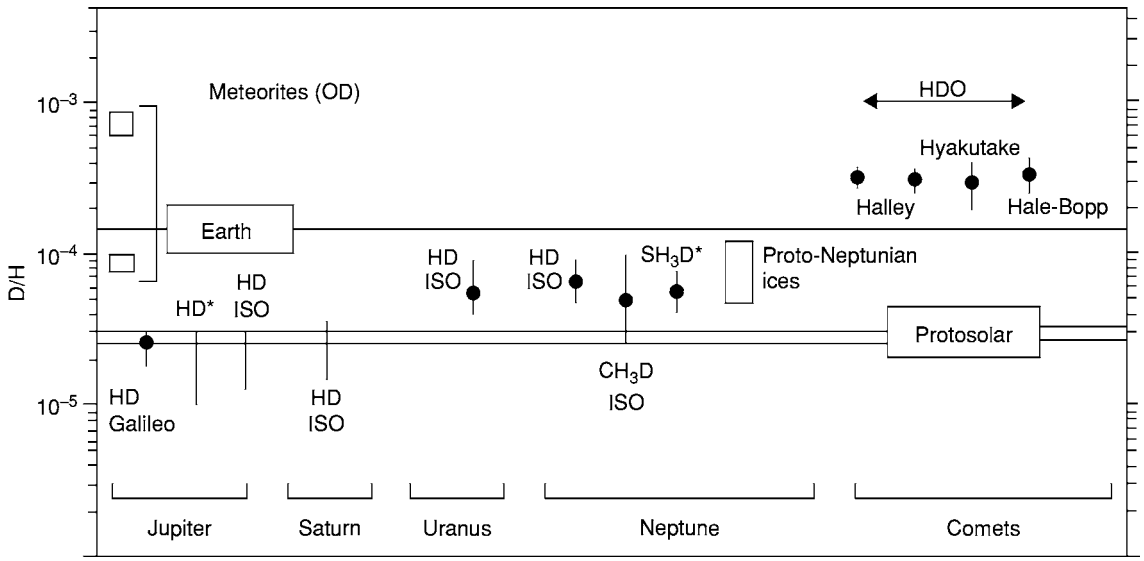


Figure 4 Hydrogen isotopic composition of various Solar System objects (adapted from A. Drouart *et al.* 1999).

Origin of the Secondary Atmosphere

Following the loss of the primary planetary atmosphere as a result of the energetic solar wind, bombardments, and the Moon-forming event, the secondary atmosphere began to accumulate. The large masses of the Earth (5.97×10^{24} g) and Venus (4.87×10^{24} g) and their planetary inventories of heat-producing elements allow the long-term persistence of high geothermal gradients and associated volcanism. In the case of the Earth, plate tectonics is possible in part because of a steep geothermal gradient that allows for the ultimate recycling of sediments and volatiles. On smaller bodies, such as the asteroids, the Moon, Mercury, and the icy moons of the outer planets, geothermal gradients are shallow; these bodies cooled quickly and early in Solar System history and are effectively cold dead worlds. Mars is an intermediate case between the warmer Earth and Venus and 'dead' worlds such as Mercury and the Moon. Comparisons of terrestrial volcanic-gas compositions with the atmospheres of the neighbouring planets show broad similarities (Tables 1 and 2). Water appears to have been lost on Venus due to thermal escape and was trapped in the crust of Mars by weathering and cold temperatures.

There is a growing consensus that the early (secondary) atmosphere of Earth was dominated by carbon dioxide, nitrogen, and water vapour, with minor components of noble gases, hydrogen, methane, and sulphurous compounds. To explain the early presence of liquid water when solar output was considerably less than today (Figure 1), it is necessary to assume that a dense carbon dioxide greenhouse atmosphere (possibly including methane) increased insolation on the early Earth. Such an atmosphere must have had a partial pressure of carbon dioxide (pCO_2) of at least 1 bar (and probably more) to keep the planet from freezing over (*see* **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Mercury; Venus; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus).

Table 2 Terrestrial volcanic gas compositions

	St Helens	*Etna*	*Kilauea*
H_2O*	91.58	53.69	78.7
H_2	0.269	0.57	1.065
CO_2	0.913	20.00	3.17
CO	0.0013	0.42	0.0584
SO_2	0.073	24.85	11.5
H_2S	0.137	0.22	3.21
OCS	2×10^{-5}	nd	0.0054
S_2	0.0003	0.21	1.89
SO	nd	0.03	nd
HCl	0.089	nd	0.167
HF	nd	nd	0.20

*Values are given in volume % (mol %); nd, not detected.

Weathering and hydrothermal reactions of crustal materials immediately commenced once the atmosphere and hydrosphere were established. Hydrothermal vents would have been ubiquitous on the early Earth, and it is estimated that the total volume of Earth's oceans could have cycled through the crust in less than 1 Ma. Carbon dioxide and water (as carbonic acid) react with silicate minerals to add carbonate and silica to the oceans, as represented in the Urey equation:

$$CO_2 + H_2O + CaSiO_2 = H_2CO_3 + SiO_2 + Ca^{2+}$$

This reaction was the major control on the amount of carbon dioxide in the atmosphere before the onset of photosynthesis, respiration, and organic-matter sequestration. During subduction, carbonate minerals carried in the descending oceanic crust are heated, releasing carbon dioxide and water as volatile components in arc magmas, completing the cycle back to the atmosphere. In hydrothermal vents within basaltic crust, water oxidizes Fe^{2+} in olivine, yielding hydrogen, magnetite, and serpentine. In gas-phase high-temperature reactions, hydrogen can reduce carbon dioxide to methane, but it is not likely that much hydrogen was maintained in the atmosphere, owing to its high rate of escape from the top of the atmosphere. The currently held view is that the early secondary atmosphere was, at most, weakly reducing, with the major components – carbon dioxide, nitrogen, water vapour, and carbon monoxide – supplemented by minor mixing fractions of hydrogen and methane. Estimates have been made of the rate of decline of carbon dioxide over geological time that provided the major greenhouse forcing to the atmosphere, while tracking secular changes in solar luminosity. Initially high partial pressures of carbon dioxide, possibly supplemented by biogenic methane, would have kept surface temperatures warm enough to avoid freezing of the planet until the crisis conditions of the snowball Earth in the Late Proterozoic. Since the era of severe Proterozoic glaciations, there has been a generally gentle decline in levels of carbon dioxide, maintaining the stability of the biosphere in the face of rising solar luminosity (of the order of about 6% per billion years).

Atmospheric Evolution

Early Anoxic Atmospheres

Free oxygen was rare or absent on the early Earth, so that the atmosphere was effectively anoxic. Constraints on ancient levels of atmospheric oxygen have been sought for many years, and most models of the early Earth consider oxygen levels to have been very low (less than 10^{-2} of present atmospheric level (PAL)) before about 2.3 Ga. Evidence to support this comes from detrital gold, uraninite, and pyrite grains in sediments older than 2.3 Ga, which it has been argued could not have been preserved in a high-oxygen atmosphere. Other geochemical indications, such as Fe^{2+}/Fe^{3+} in palaeosols and discussions of the possible origins of banded iron formations (*see* **Sedimentary Rocks:** Banded Iron Formations) and of carbon isotopic signatures in Archaean and Palaeoproterozoic organic matter and carbonates, have also been used to support the contention that the early atmosphere was essentially devoid of oxygen. In the absence of a ready source of free oxygen and, therefore, no effective ozone (O_3) screen from solar ultraviolet, atmospheric water vapour would have photodissociated into hydrogen and oxygen. Since the oxygen formed by this reaction produces just enough ozone to inhibit the reaction, accumulation of oxygen by this means is limited. Oxygen combines readily with hydrogen to form water vapour, so little oxygen is lost from the atmosphere to space. The bottom of the stratosphere acts as a cold trap and maintains a constant temperature of $-56°C$, keeping water from escaping from the top of the atmosphere (Figure 5).

Sulphur isotopes: a unique window onto early atmospheric chemistry Further evidence in support of an early anoxic atmosphere comes from the unusual chemistry of sulphur isotopes. Most physical processes fractionate the isotopes of an element because of the relative mass differences between the isotopic species. For instance, because the mass difference between ^{33}S and ^{32}S is half that between ^{34}S and ^{32}S, any mass-dependent (defined by $\Delta^{33}S = 0‰$; where $\Delta^{33}S = 0.515\delta^{34}S - \delta^{33}S$) physical effects on the ratio of ^{33}S to ^{32}S are expected to be about half those on the ratio of ^{34}S to ^{32}S. However, certain unusual gas-phase chemical reactions in the absence of free oxygen are capable of producing anomalous sulphur isotopic values that do not follow the mass-dependent relationship described above. The recent discovery of the signal of the mass-independent fractionation effect (where $\Delta^{33}S \neq 0‰$) in geological materials has important implications for understanding the chemical evolution of

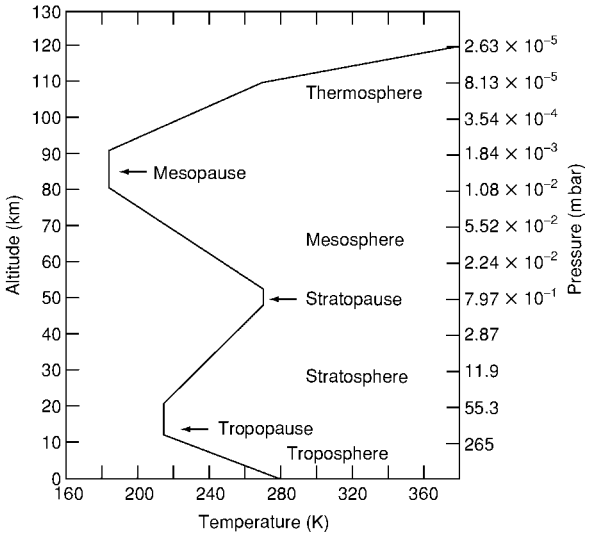


Figure 5 Temperature and pressure profile of the Earth's atmosphere.

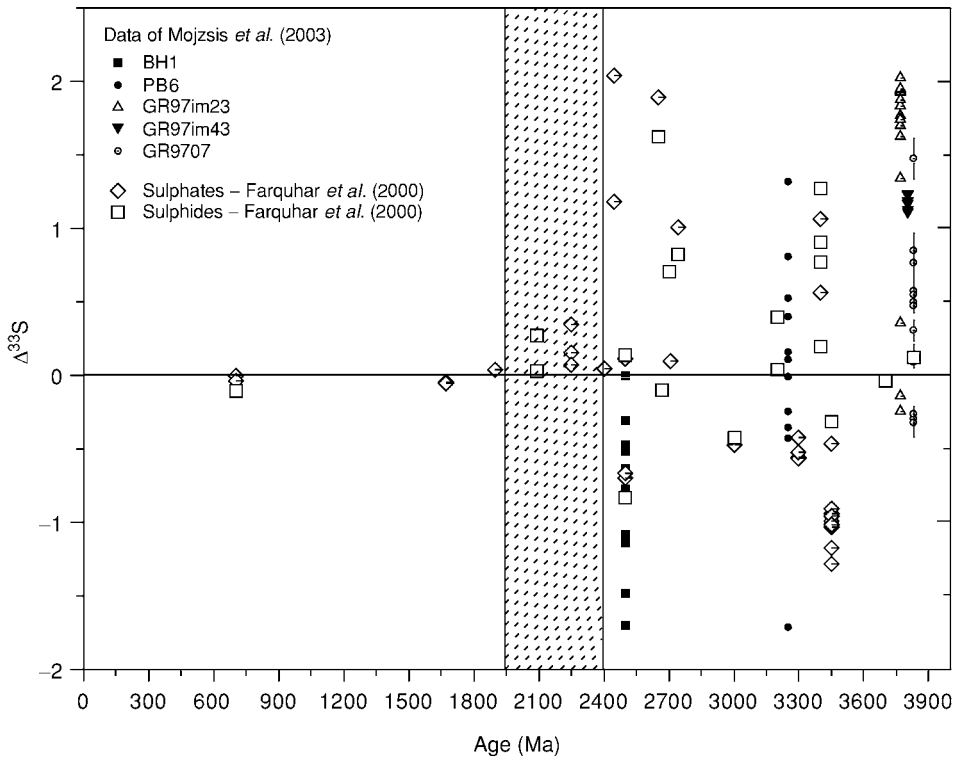


Figure 6 Mass independent sulphur isotopic compositions track changes in atmospheric composition with time. (Reproduced from Mojzsis *et al.* (2003).)

the atmosphere. It has been found that sedimentary sulphide and sulphate minerals in Archaean rocks contain sulphur where the levels of the minor isotope (^{33}S) deviate from the mass-fractionation line with other post-2.3 Ga terrestrial sulphur minerals. It appears that the signal of this chemistry can be transferred from the atmosphere to surface materials and thereafter to the geological record. The mass-independent fractionation effect of sulphur in ancient sulphides records chemical reactions in early anoxic terrestrial atmospheres, which were destroyed in the Early Proterozoic (2.4–2.3 Ga), and tracks the rise in levels of atmospheric oxygen (Figure 6).

Metabolic Energy and the Rise of Oxygen

Changes in atmospheric composition have had a profound effect on the evolution of life. Sequence analyses used to compare genomes have been used to infer the phylogenetic relatedness of living organisms and their evolutionary history in terms of molecular biology and metabolic style. It has been found that three great domains exist within the ribosomal RNA tree of life: Bacteria, Archaea, and Eukarya (Figure 7) (*see* **Precambrian:** Prokaryote Fossils). The organisms with the deepest branches in the tree, corresponding to the most ancient pedigrees, close to the root in Figure 7, are hyperthermophilic (heat-loving) chemoautotrophs (organisms that obtain their metabolic energy from chemical disequilibria). It appears that at least some of the earliest organisms lived off chemical energy, in much the same way as contemporary hot-spring microbial communities. Such environments are widespread in volcanic centres on land and on the seafloor, and they would have been even more widespread on a geologically restive early Earth. However, hyperthermophilic organisms are restricted to zones around hydrothermal vents where optimal chemical and temperature conditions persist. Photosynthesizers, on the other hand, are less restricted and have evolved to inhabit environments away from those of their chemosynthetic ancestors into the open ocean.

Light from the Sun provides a readily accessible, stable, and inexhaustible energy source to drive metabolic reactions. Photosynthesis uses light energy to capture electrons and move low-energy-state ions to higher energy states. The energy released can then be used to power metabolic reactions such as growth and reproduction. Photosystem I was the earliest photosynthetic pathway and used light-activated chlorophylls and electron donors such as H_2S and Fe^{2+} to yield S^0 and Fe^{3+} as oxidative products. At a later stage, photosystem II developed; this uses water as the electron

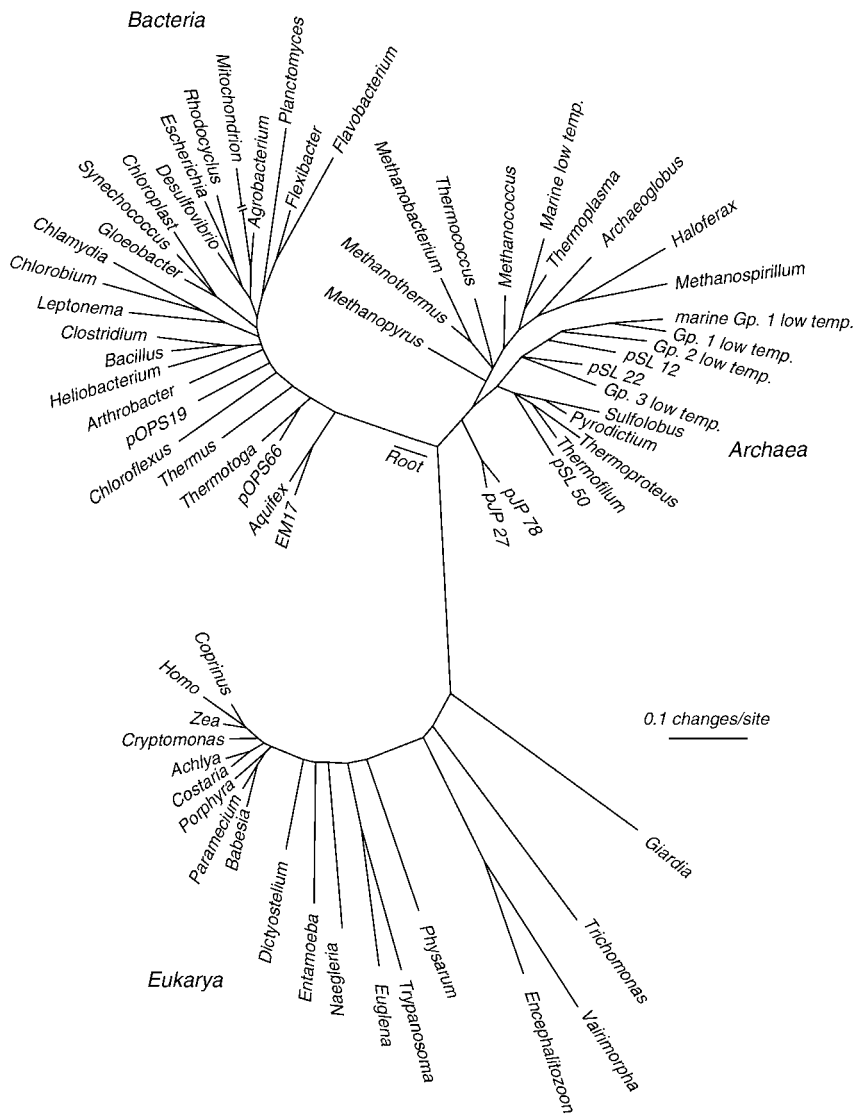


Figure 7 The phylogenetic tree of life based on comparative 16s rRNA sequences. (Reproduced from Pace NR (2001) The universal nature of biochemistry. *Proceedings of the National Academy of Sciences USA* 98: 805 808.)

donor to reduce carbon dioxide to biologically useful organic compounds such as carbohydrates. Liquid water and carbon dioxide are available to life in practically unlimited quantities; sunlight drives the reaction, and cyanobacteria that use it, although relative late-comers in phylogenetic terms, are arguably the most successful organisms in the history of life on Earth. It is photosystem II that releases free oxygen as a by-product in the reversible reaction:

$$6CO_2 + 6H_2O + \text{light energy} \leftrightarrow C_6H_{12}O_6 + 6O_2$$

The gradual build-up of oxygen in the atmosphere was accomplished by the slow toilsome efforts of cyanobacteria over many hundreds of millions of years. The immense importance of this metabolic reaction to the history of all life beyond the microbial level cannot be overstated (*see* **Earth System Science**).

Rise in atmospheric oxygen Initially, the vast amounts of reduced iron (Fe^{2+}) and other chemical species dissolved in the oceans, and perhaps some reduced gases in the atmosphere (such as methane), provided a sink for the oxygen produced by photosynthesis. As these sinks were exhausted, local concentrations of oxygen would have appeared, creating a dilemma for the early anaerobic microbial biosphere. High levels of oxygen produce toxic radicals in the environment, which cause extensive damage to cell components. This would have been a strong driving force for natural selection leading to adaptation: microbes that did not immediately

go extinct survived by retreating, for example, to anaerobic environments, such as deep in crustal rocks, organic-rich muds, and animal digestive tracts, where they continue to thrive today. New forms of life emerged with superoxide dysmutase and catalase enzymes, which catalyze the reduction of oxic free radicals to water. Finally, another group (the Eukarya; Figure 7) developed aerobic metabolism, providing energy yields approximately twenty times greater than those of anaerobic metabolism (*see* **Precambrian:** Eukaryote Fossils). The mitochondria – cell organelles contained in most Eukarya, which facilitate oxidative metabolism – arose from the endosymbiosis of proteobacteria. These organelles actually contain their own subunits of DNA that further implicate a eubacterial heritage.

The increased energy yield of aerobic metabolism set the stage for the evolution of all life above the unicellular level. The energy source that supports the aerobic biosphere is the Sun – thus established oxygenic photosynthesis and aerobic respiration governed the flow of carbon and oxygen through the atmosphere–hydrosphere system. During the Archaean the atmosphere and hydrosphere initially contained small amounts of free oxygen from photolysis of water vapour and then increasing amounts from oxygenic photosynthesis. By the end of the Archaean–Early Proterozoic, oxygen levels had begun to creep upwards, balanced by oxygen-consuming reactions such as weathering, hydrothermal activity, respiration, oxidation of organic matter, and differential rates of organic-matter burial. By about 2 Ga, levels of oxygen were about 1% of PAL. A transition era ensued, with oxygen levels fluctuating around the 1% PAL value and the oceans remaining reducing and sulphidic. After about 1.8 Ga, oxygen-consuming reactions were generally exhausted and free-oxygen concentrations reached about 10% of PAL. Over time, these levels stabilized at near 'normal' Phanerozoic atmospheric concentrations. It remains unquantified how rapidly oxygen levels increased during the Proterozoic.

The rise in atmospheric oxygen had an acute effect on the surface environment, not only because of its toxicity to many microbial organisms (providing the motive force to drive the evolution of more efficient aerobic metabolisms) but also by establishing an effective ultraviolet screen. Ozone is far more effective than diatomic oxygen at absorbing ultraviolet. The ozone screen formed in the stratosphere from accumulating O_2 that photodissociated to produce free oxygen radicals (O°), which then recombined with O_2 to make ozone. This ozone screen effectively made dry land habitable for plants and animals by the Palaeozoic. Industrial pollutants such as chlorofluorocarbons are now severely damaging the ozone layer.

A Neoproterozoic Snowball Earth

The increased oxidation of the surface zone in the Middle Proterozoic was probably a consequence of the sequestration of large quantities of organic carbon in sediments. From the oxygenic photosynthesis reaction, it can be seen how this scenario leads to a net loss of carbon dioxide and a net increase of oxygen in the atmospheric reservoir. The boosted greenhouse effect of heightened levels of carbon dioxide before about 2.5 Ga was lost to the Earth at a most inopportune time. Levels of solar luminosity were about 10%–15% lower at 2.2 Ga than at present (Figure 1). Loss of insolation from the carbon dioxide greenhouse spelled disaster for the Proterozoic Earth. The planet froze over, locking the oceans in ice and creating a high-albedo feedback loop that kept the planet frozen for extended periods, until levels of carbon dioxide increased due to passive volcanism and outgassing. Heightened levels of carbon dioxide warmed the atmosphere and dark dusty ice reduced the albedo, causing catastrophic melting and massive planetary warming. Subsequent weathering in a carbon dioxide-rich atmosphere, combined with massive algal blooms, led to enhanced burial of organic carbon, drawdown of carbon dioxide levels, and a renewed snowball Earth. The cycle is thought to have been broken by the secular increase in solar luminosity and a steady redistribution of the continents via plate motions. The survivors of these repeated 'ice-house' and 'hothouse' Earths were multicellular organisms that inherited a more stable environment high in oxygen (*see* **Palaeoclimates**).

The Phanerozoic Atmosphere

Phanerozoic Atmospheric Changes

Geochemical evidence from the study of palaeosols, coupled with data from carbonate and organic carbon in sediments as well as sedimentary pyrite and the chemistry of sedimentary silicate minerals, has been used to improve models of the carbon cycle of palaeoenvironments. These models have been used to document and explain fluctuations in levels of oxygen and carbon dioxide over Phanerozoic time. Long-term changes in the carbon dioxide and oxygen concentrations in the Phanerozoic atmosphere are summarized in Figure 8.

In the first part of the Phanerozoic, the Early Palaeozoic (Cambrian–Ordovician), evidence indicates that levels of carbon dioxide were about fifteen times PAL. These declined to within a few percent of

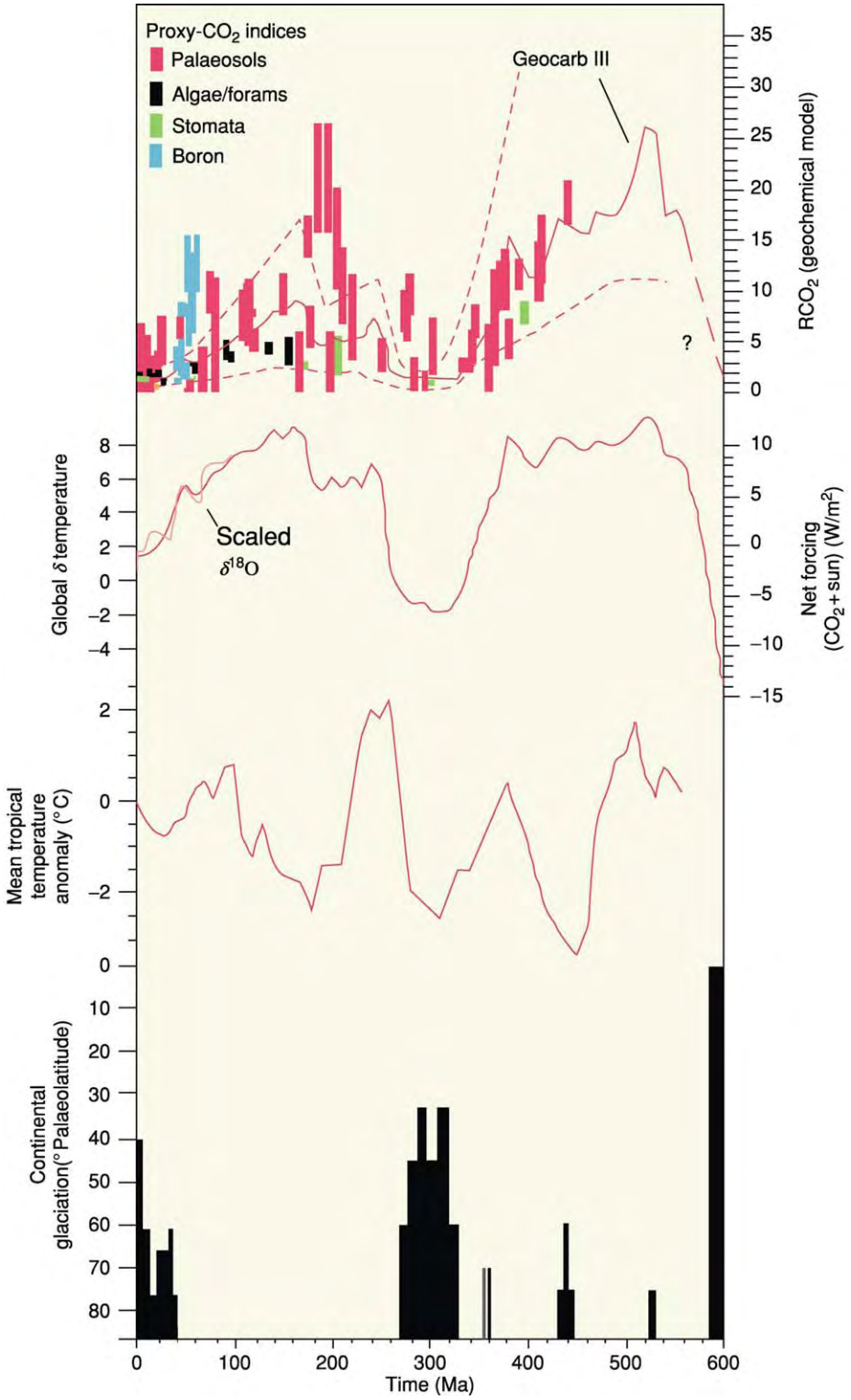


Figure 8 Phanerozoic temperature history. (Reproduced from with permission from Berner RA. Geocarb II: A revised model of atmospheric CO_2 over phanerozoic time, *Am. Jour. Sci.*, 294: 56 91)

PAL at the end of the Carboniferous. Massive carbon burial and increased solar radiation by the Carboniferous meant that atmospheric oxygen concentrations were greater than at any other time in Earth history (*see* **Palaeozoic:** Carboniferous). This may have allowed the existence of the huge Carboniferous insects observed in the fossil record (*see* **Fossil Invertebrates:** Insects). Insects rely on diffusion-limited flow

of oxygen through their exoskeletons: to be big they need to live in conditions of high atmospheric oxygen, so that sufficient oxygen passively diffuses into their blood to power muscles for flight.

Reduced greenhouse forcing and glaciation at the end of the Palaeozoic with the assembly of Pangaea reduced organic-matter burial, and there was a slow rise in carbon dioxide levels to around six times PAL in the Permian and Triassic. Levels of carbon dioxide gradually decreased, and stabilized near present-day levels in the Mesozoic. Oxygen tends to track carbon dioxide inversely; geological evidence and palaeoclimate models suggest a maximum of near 35% oxygen in the atmosphere at the beginning of the Permian. During the Permian, the oceans were highly stratified, with carbonate-rich water at depth that was depleted in oxygen. This system was unstable: ocean hypoxia could occur if ocean circulation intensified enough to mix deep-water carbon dioxide and hydrogen sulphide into surface waters. A protracted (20 Ma) whole-ocean hypoxia event is considered to be a major mechanism in the Permian–Triassic extinction event, which wiped out 90%–95% of all marine species (*see* **Palaeozoic:** End Permian Extinctions).

Carbon Dioxide and Climate Changes

High-resolution information about changes in atmospheric chemistry over the past 160 000 years can be obtained by studying the record of trapped gases in ice cores from the Greenland and Antarctic ice-caps. Furthermore, oxygen isotopic values from marine sediments, marine planktonic and benthonic fossils, cave deposits, and other sources can be used to estimate marine palaeotemperatures. Direct measurements of carbon dioxide and methane concentrations in ice cores permit assessment of past atmospheric levels of these gases, providing factors to incorporate into models of past air temperatures and sea-levels.

Data from deep ice cores taken in polar regions, coupled with complex palaeoclimate models (Figure 9), show large fluctuations in atmospheric carbon dioxide, oxygen and methane levels, leading to long-term temperature changes of the order of ±6°C or more. There is a strong correlation between levels of atmospheric greenhouse gases and palaeotemperature. The periodicities in these data provide clear evidence of the role of Milankovitch forcing by changes in the Earth's orbital parameters. The two strongest Milankovitch cycles observed correspond to the 26 000 year precession of the equinoxes and the 100 000 year period of rotation of the Earth's orbital axis (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). Changes in greenhouse-gas concentrations appear to follow rather than guide long-term climate, suggesting that Milankovitch

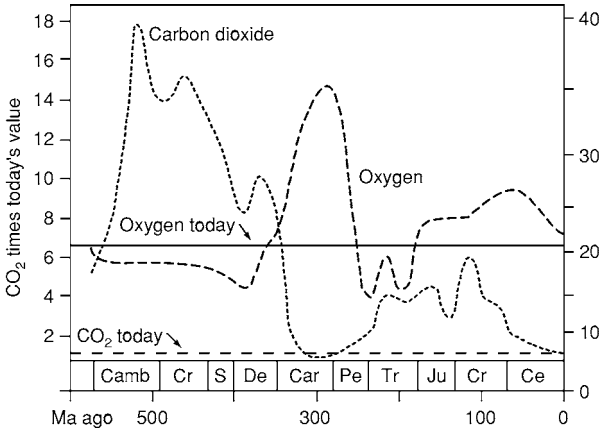


Figure 9 Phanerozoic carbon dioxide and oxygen concentrations.

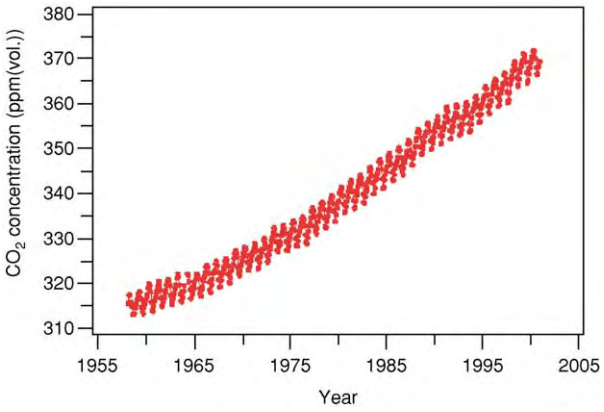


Figure 10 Changes in the concentration of atmospheric carbon dioxide over the last 60 years as measured at Mauna Loa, Hawaii. Data provided by D. Keeling and T. Whorf.

cycles are the prime mechanism bringing the Earth into a greenhouse or icehouse condition. Greenhouse gases do provide positive feedback at the beginning of temperature changes by boosting insolation. Anthropogenic emissions of greenhouse gases are not governed by Milankovitch cycles and represent a separate and increasingly important climate-forcing mechanism. Figure 9 shows that carbon dioxide concentrations changed by almost 100 ppm(vol.) towards the end of the last glaciation. Modern levels of carbon dioxide are near 370 ppm(vol.) and rising. Almost all of this change has occurred since the Industrial Revolution, and high-resolution monitoring at the Mauna Loa observatory shows clear diurnal and annual cycles in carbon dioxide levels (Figure 10), with a mean annual increase of 1.16 ppm(vol.) $year^{-1}$. This is over one hundred times the rate of increase of carbon dioxide levels inferred from all available

geological data, with no clear stabilization of these levels in sight.

Other greenhouse gases Neglecting the small contribution from internal heating driven by radioactive decay, planetary surface temperatures are governed by the amount of solar radiation received and its interaction with the atmosphere. Re-radiation to the surface of infrared radiation by greenhouse gases, chiefly carbon dioxide but also water vapour, methane, nitrous oxide species (NO_x), and chlorofluorocarbons (CFCs), keeps the present average surface temperature some 33 K above the black-body temperature of 255 K. Although they are present in small amounts relative to carbon dioxide and water vapour, methane, nitrous oxide species, and CFCs (which can form only from industrial processes) are very effective greenhouse gases. Methane in the atmosphere (concentration of 1.7 ppm(vol.)) is dominantly formed by biological processes and absorbs infrared radiation approximately 21 times more effectively than carbon dioxide; its levels have increased rapidly in the last several centuries. Gaseous nitrous oxide compounds are present at low concentrations (*ca*. 0.3 ppm(vol.)), are produced by biological nitrification, and have a long residence time in the atmosphere. The concentration of CFCs is very low (less than 0.003 ppm(vol.)), but they have a greenhouse effect ten thousand times greater than that of carbon dioxide and likewise have long residence times in the atmosphere.

Conclusions

The splendour of contemporary life on Earth, as revealed by the geochemical, palaeontological, and molecular phylogenetic records, took billions of years to achieve, and its evolution occurred for the most part within the envelope of the atmosphere and hydrosphere. In the context of astrophysical changes to the Sun and geophysical changes to the solid Earth, the atmosphere has evolved through a complex set of interrelated cycles, within which biology has been of central importance. Life has affected the planetary atmosphere, and biological evolution was in turn affected by it. The gaseous envelope of the planet evolved in response to changing geophysical conditions, such as mantle heat flow and solar luminosity. Early in Solar System history, Venus and Mars apparently had geochemical paths that paralleled that of the Earth, probably including liquid water at their surfaces. However, long ago they diverged from a physicochemical regime that promoted habitability. The nascent biosphere on these planets, if it ever existed, was either burned to a crisp (on Venus) or freeze-dried (on Mars). In the far future, as the luminosity of the Sun continues to increase, Earth will go the way of Venus.

See Also

Earth: Mantle; Crust; Orbital Variation (Including Milankovitch Cycles). **Earth Structure and Origins**. **Earth System Science**. **Fossil Invertebrates:** Insects. **Palaeoclimates**. **Palaeozoic:** Carboniferous; End Permian Extinctions. **Precambrian:** Overview; Eukaryote Fossils; Prokaryote Fossils. **Sedimentary Rocks:** Banded Iron Formations. **Solar System:** The Sun; Asteroids, Comets and Space Dust; Meteorites; Mercury; Venus; Moon; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus.

Further Reading

Hoffman PF and Schrag DP (2000) Snowball Earth. *Scientific American* 282: 68 75.

Holland HD (1984) *The Chemical Evolution of the Atmos phere and Oceans*. Princeton: Princeton University Press.

Margulis L (1984) *Early Life*. Boston: Jones and Bartlett.

Mojzsis SJ and Harrison TM (2000) Vestiges of a beginning: clues to the emergent biosphere recorded in the oldest known sedimentary rocks. *GSA Today* 10: 1 6.

Pace NR (2001) The universal nature of biochemistry. *Proceedings of the National Academy of Sciences USA* 98: 805 808.

Royer DL, Berner RA, Montañez IP, Tabor NJ, and Beerling DJ (2004) CO_2 as a primary driver of Phanerozoic climate. *GSA Today* 14: 4 10.

Sagan C and Mullen G (1972) Earth and Mars: evolution of atmospheres and surface temperatures. *Science* 177: 52 56.

AUSTRALIA

Contents

Proterozoic

I M Tyler, Geological Survey of Western Australia, East Perth, WA, Australia

Introduction

The Proterozoic is the period of geological time extending from the end of the Archaean, 2500 million years ago (Ma), to the start of the Phanerozoic (the base of the Cambrian System), 545 Ma. The Proterozoic is divided into the Palaeoproterozoic (2500–1600 Ma), the Mesoproterozoic (1600–1000 Ma), and the Neoproterozoic (1000–545 Ma). In Australia, Proterozoic rocks are present to the west of the 'Tasman Line' that separates 'Proterozoic Australia', where geophysical datasets show that Precambrian basement is continuous beneath Phanerozoic sedimentary basins, from the Tasmanides, where predominantly Palaeozoic basement is overlain by Mesozoic and younger sedimentary basins. Extensive exposures of Proterozoic rocks are present in western Australia, in northern, central, and north-eastern Australia, and in southern Australia and Tasmania. Proterozoic Australia is made up of three distinct cratons, the West Australian Craton, the North Australian Craton, and the South Australian Craton (Figure 1). These probably formed originally as parts of larger cratons or continental blocks (the South Australian Craton together with the previously adjacent part of Antarctica formed the Mawson Craton) and are dominated by Archaean and Palaeoproterozoic to Mesoproterozoic crust. The three cratons are separated by two predominantly Mesoproterozoic to Neoproterozoic orogenic belts, the Paterson Orogen and the Albany–Fraser Orogen. The Palaeoproterozoic to Neoproterozoic Pinjarra Orogen is present along the western margin of Australia.

Plate tectonic models can be applied to Proterozoic Australia. The increasing availability of high-quality geochronological data has highlighted the presence of distinct tectonostratigraphic terranes with differing geological histories within orogenic belts, and geophysical datasets reveal the heterogeneous nature of the crust throughout Proterozoic Australia. However, Proterozoic plate-tectonic processes may differ from modern processes, and the real lack of accretionary complexes and ophiolites, and significant differences in the geochemical and isotopic compositions of igneous rocks, may reflect changes in the nature and composition of the oceanic lithosphere through time. Palaeomagnetic evidence is placing greater controls on the movement and relative positions of the constituent crustal components. Diverse cratons and continental blocks aggregated during the Palaeoproterozoic and Early Mesoproterozoic to form Proterozoic Australia, which then played an integral part in the formation and breakup of the Meso- to Neoproterozoic supercontinent, Rodinia (Figures 2–7; summarized in Table 1). Proterozoic Australia is host to a wide variety of minerals , including world-class deposits of iron, uranium, gold, copper–gold, lead–zinc–silver, and diamond orebodies (Figure 8).

Neoarchaean to Palaeoproterozoic Assembling Proterozoic Australia: (2770–1600 Ma)

West Australian Craton

Within the West Australian Craton, large areas of Archaean rocks are exposed in the geologically distinct Pilbara and Yilgarn cratons (Figures 1 and 2). The Hamersley Basin was initiated on the southern part of the Pilbara Craton at the start of the Neoarchaean. West- to south-westerly directed rifting within cratonized granite–greenstone basement represents a distinct change to a Proterozoic and Phanerozoic tectonic style. The flood basalts (2770–2690 Ma) of the Fortescue Group dominate the rift-related lower part of the basin. These were buried beneath a breakup unconformity overlain by a passive margin sequence characterized by cherts and banded iron formations

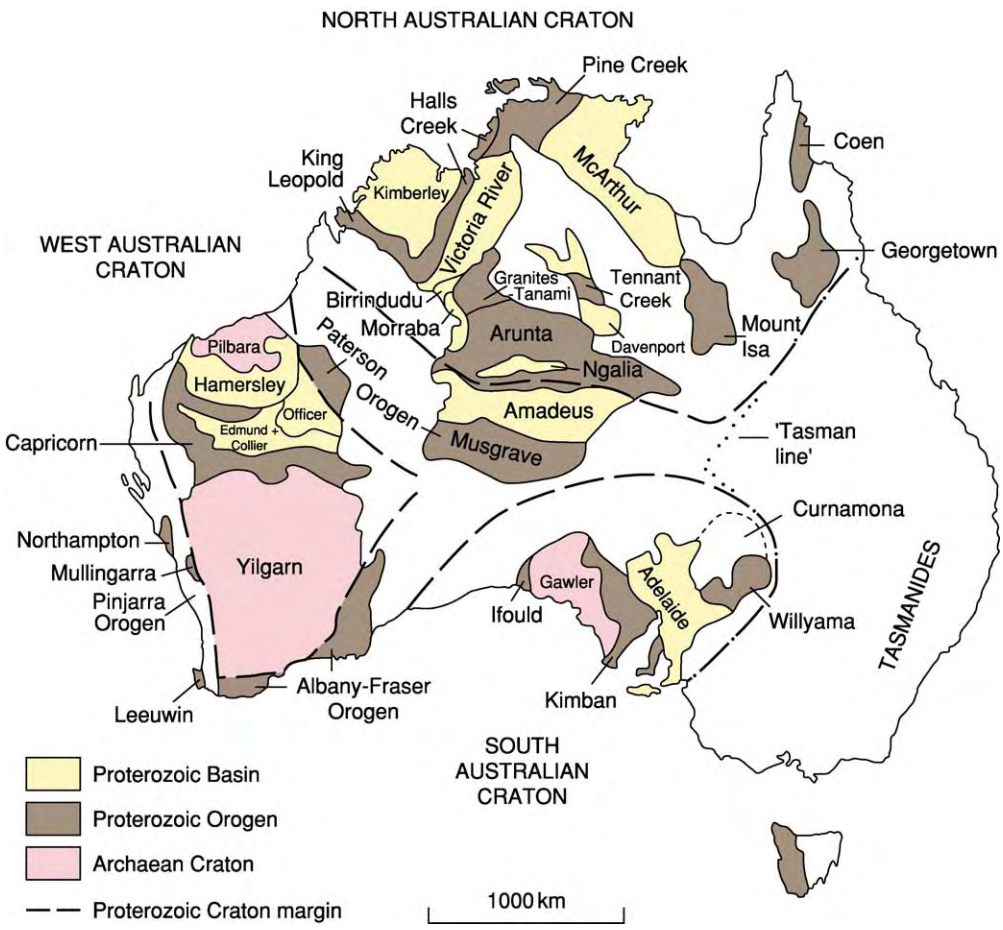


Figure 1 North, South, and West Australian cratons, showing the main outcrops of Precambrian rocks in Australia.

(BIFs) of the uppermost Fortescue Group and the Hamersley Group (2600–2440 Ma), which were deposited on a shelf or platform open to an ocean (*see* **Sedimentary Rocks:** Banded Iron Formations). A collisional, intracontinental back-arc setting has been proposed for the BIFs and mafic and felsic magmatic rocks of the upper part (2470–2440 Ma) of the Hamersley Group. The overlying Turee Creek Group, which includes probable glacial deposits, and the lower Wyloo Group were deposited in the McGrath Trough, a foreland basin developed in front of a northward-verging fold-and-thrust belt during the Ophthalmian Orogeny (~2200 Ma) (Figure 2). The Glenburgh Terrane in the southern part of the Gascoyne Complex within the Capricorn Orogen to the south contains basement from 2550–2450 Ma and may represent a remnant of the colliding continent that drove this orogeny.

In the Yilgarn Craton, tectonic processes typical of the Archaean developed the Eastern Goldfields granite–greenstone terrane 2700–2600 Ma, at the same time as the upper Fortescue Group on the Pilbara Craton. In the southern Capricorn Orogen, rifting and continental breakup took place in the gneiss terranes and granite–greenstones along the northern margin of the Yilgarn Craton. The ~2150-My-old rocks of the basal Yerrida Basin (Figure 2) formed initially in a sag basin followed by an abrupt change to a rift setting. The development of the Bryah and Padbury basins 2020–1900 Ma at the north-western margin of the Yilgarn Craton may reflect the development of a back-arc basin, which was then overlain by a foreland basin during the Glenburgh Orogeny 2005–1960 Ma (Figure 2). Geochemical and isotopic data indicate that suprasubduction zone magmatism formed the Dalgaringa Supersuite at an Andean-type margin during this event; this may represent the coming together of the combined Pilbara Craton and Glenburgh Terrane with the Yilgarn Craton.

The Capricorn Orogeny from 1830–1780 Ma (Figure 3) has previously been regarded as marking the collision between the Pilbara and Yilgarn Cratons. Neodymium isotopes show that melting of pre-existing, Early Palaeoproterozoic crust without the introduction of mantle-derived material formed

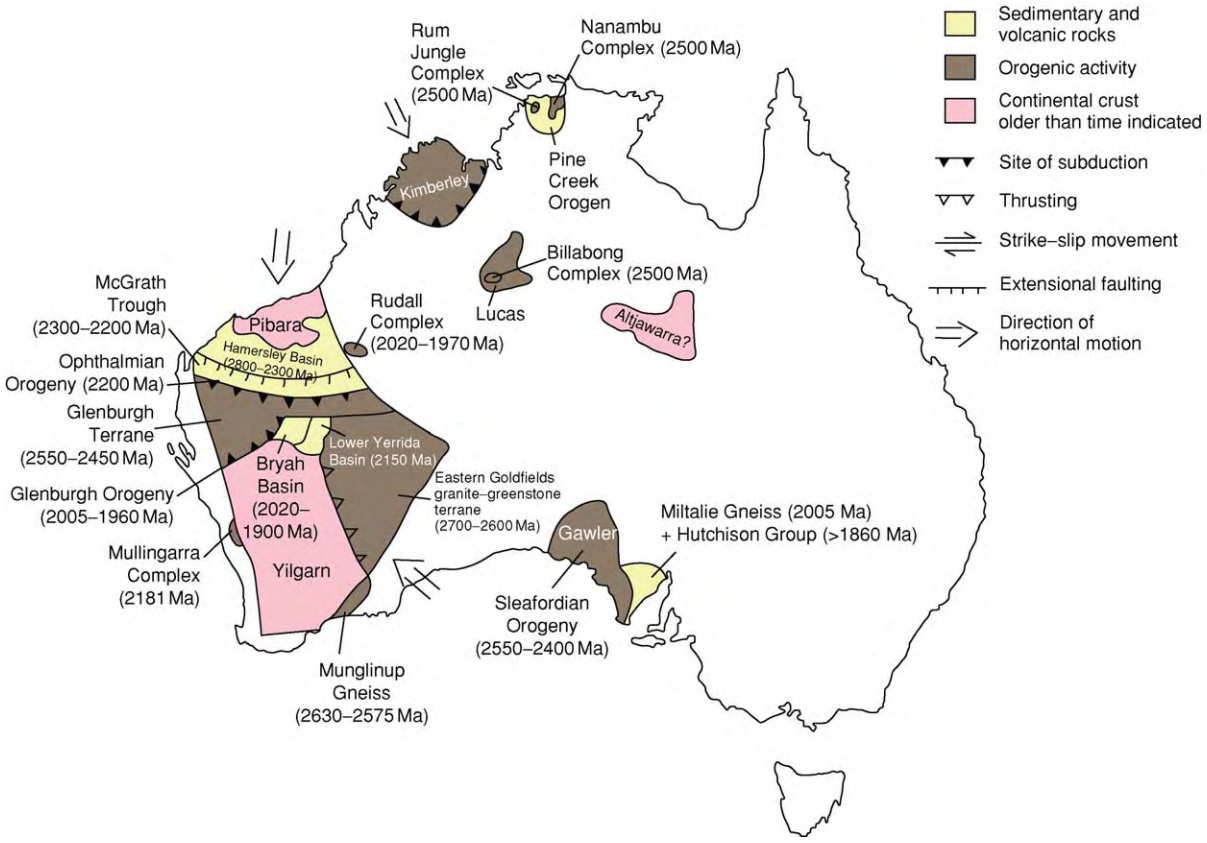


Figure 2 Assembly of the West Australian Craton, 2800 1900 Ma.

voluminous granitic magmatism, which, together with associated deformational and metamorphic events, extended across the entire orogen. The Capricorn Orogeny may represent an intracratonic event resulting from reactivation, possibly during amalgamation of the West Australian Craton with the North Australian Craton. The 1840- to 1805-My-old upper part of the Ashburton Basin and the ~1805-My-old Blair Basin developed as a foreland basin along the northern margin of the orogen at this time. Uplift of the Gascoyne Complex and the Yilgarn Craton, together with a presently unexposed Early Palaeoproterozoic terrane, provided the sediment deposited in the upper Ashburton Basin and the Blair Basin, the ~1840-My-old upper part of the Yerrida Basin and the 1840- to 1800-My-old Earaheedy Basin (Figure 3).

In the Rudall Complex along the eastern margin of the Pilbara Craton (Figure 3), the Talbot Terrane contains quartzite, amphibolite, and serpentinite as inclusions within complex orthogneisses that contain components that crystallized, respectively, ~2015, ~1970, and ~1800 Ma. A younger sequence (from <1790 Ma) of clastic metasedimentary rocks was possibly deposited in a foreland basin. In the adjacent Connaughton Terrane, a sequence of deformed and metamorphosed mafic volcanic rocks, and chemical and clastic sedimentary rocks, may have been deposited in a rift prior to ~1780 Ma. Deformation, metamorphism, and further granite intrusion took place in the Talbot and Connaughton terranes during the Yapungku Orogeny (1790–1760 Ma). West-verging thrusting and high-P metamorphism may have accompanied collision of the West Australian Craton with the North Australian Craton, which palaeomagnetic evidence indicates were unified by ~1700 Ma.

The Mount Barren Group and equivalent sedimentary rocks were deposited on the south-eastern part of the Yilgarn Craton ~1700 Ma (Figure 4). Reactivation of the Capricorn Orogen took place between 1670 and 1620 Ma with the intrusion of granitic rocks and the occurrence of medium- to high-grade metamorphism, which was synchronous with localized shear zones and hydrothermal alteration. In the Albany–Fraser Orogen, the Biranup Complex consists of granitic gneisses that include both Archaean (2630–2575 Ma) (Figure 2) and Palaeoproterozoic (1700–1600 Ma) protoliths (Figure 3). In the Pinjarra Orogen, a monzogranite in fault contact with

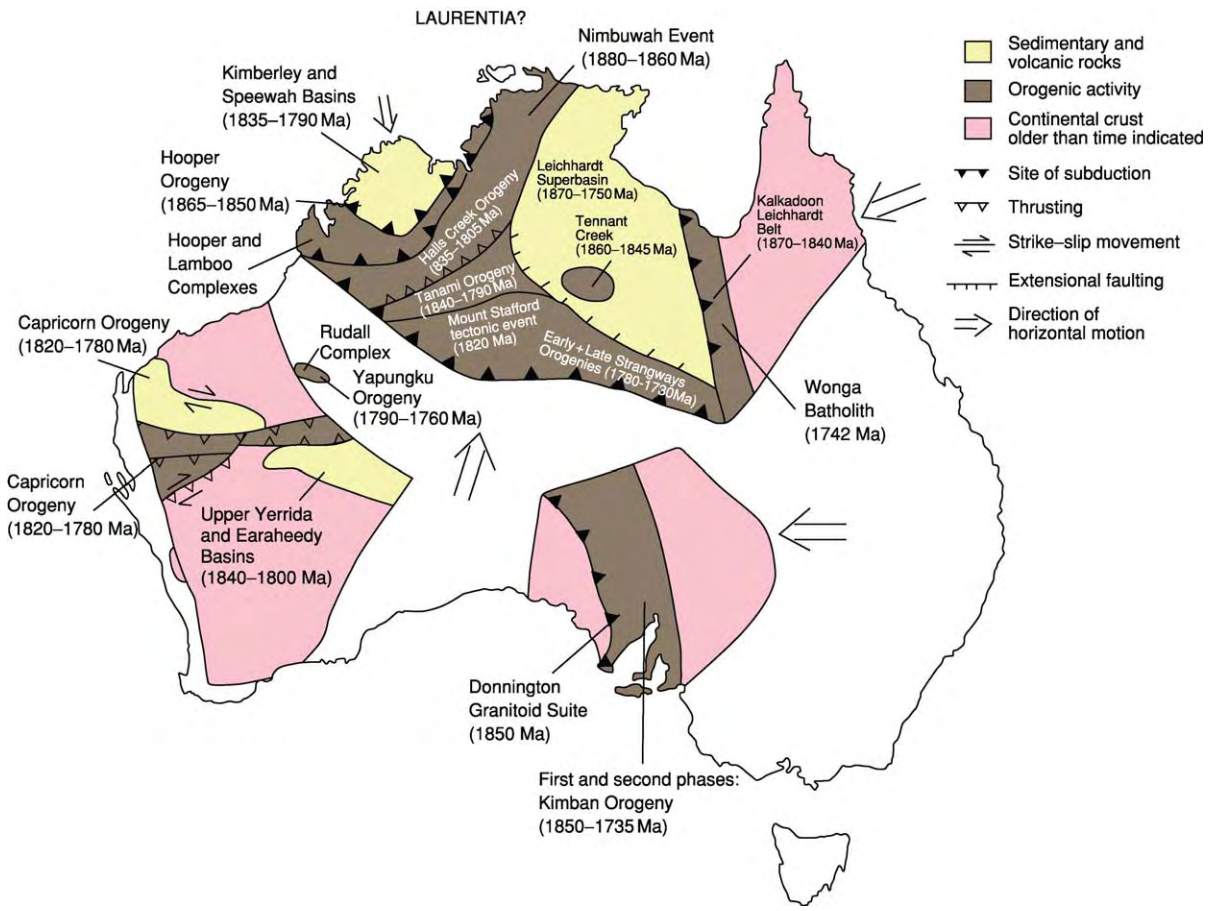


Figure 3 Assembly of the North Australian and South Australian cratons, 1900 1730 Ma.

metasedimentary gneiss in the Mullingarra Complex has been dated to ~2181 Ma (Figure 2). This is consistent with Sm–Nd model ages of 2280–2030 My from the Northampton and Mullingarra complexes, and boreholes in the Perth Basin, indicating the involvement of Early Palaeoproterozoic crust in the evolution of the orogen.

North Australian Craton

In the North Australian Craton, the presence of probable Archaean to Early Palaeoproterozoic cratons and continental blocks within the basement is inferred from geophysical datasets. Latest Neoarchaean basement is exposed as granite and granite gneiss within the Rum Jungle and Nanambu complexes in the Pine Creek Orogen (Figure 2). These basement inliers are overlain by arkosic rocks of the Namoona Group and the shallow-marine carbonates and mudstones and siltstones of the Mount Partridge Group deposited before ~1885 Ma. Archaean crust of the Lucas Craton is also present as inliers partly covered by arkosic and conglomeratic rocks within the Browns Range Dome and the Billabong Complex in the Granites–Tanami Complex (Figure 2). The presence of further concealed Archaean crust that may be as old as ~3600 My is indicated by Sm–Nd and Pb isotopic data, together with a U–Pb detrital zircon studies.

A widespread orogenic event took place throughout the North Australian Craton between 1910 and 1790 Ma. This event, originally defined with a shorter duration (between 1870 and 1840 Ma), has been termed the 'Barramundi Orogeny'. It was interpreted as an intracratonic event involving a linked polygonal system of rifts and sag basins, which evolved rapidly to became the sites of crustal shortening, voluminous magmatism, and high-temperature/low-pressure metamorphism. More recent interpretations have suggested that the 'Barramundi Orogeny' represents a series of linked collisional events that assembled the North Australian Craton. Based on similarities that can be recognized with the evolution of the Trans-Hudson Orogen, the North Australian Craton may have been originally contiguous with Laurentia, which would have lain to the north.

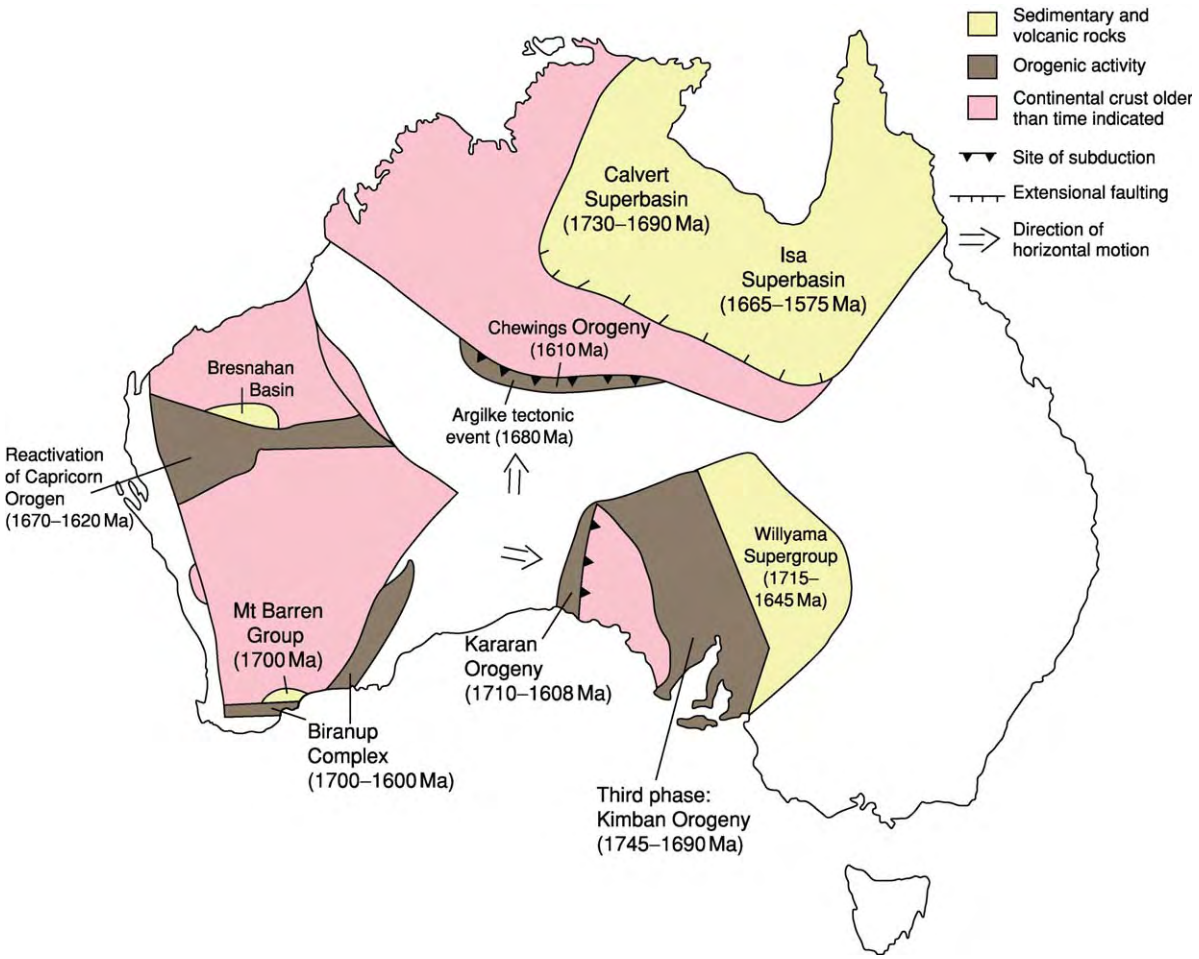


Figure 4 Final assembly of the cratonic elements of Proterozoic Australia, 1730 1600 Ma.

In the north-western part of the craton, the King Leopold and Halls Creek orogens were initiated by rifting along the western margin of the Lucas Craton ~1910 Ma. Renewed rifting is represented by deposition of the lower Halls Creek Group ~1880 Ma. Accretion of Neoarchaean to early Palaeoproterozoic continental fragments to the opposing eastern edge of the Kimberley Craton occurred before ~1910 Ma. Turbidites derived by erosion from these fragments were deposited ~1870 Ma. Deformation, metamorphism, and extensive felsic and mafic magmatism (Whitewater Volcanics, Paperbark Supersuite) occurred during the Hooper Orogeny (1865–1850 Ma) (Figure 3). The central part of the orogen (Tickalara Metamorphics) formed ~1865 Ma as an oceanic island arc above a north-west-dipping subduction zone. Layered mafic–ultramafic intrusions were emplaced ~1855 Ma. Deformation and metamorphism to high grade ~1845 Ma followed intrusion of numerous felsic and basic to intermediate sheetlike bodies during convergence and collision of the arc with the Kimberley Craton. Alkaline volcanism between 1870 and 1850 Ma marked further rifting along the Lucas Craton margin. A submarine fan deposited turbiditic rocks of the upper Halls Creek Group parallel to the craton margin.

Eruption of felsic and mafic volcanic rocks during rifting of the accreted arc ~1840 Ma was accompanied by the emplacement of further layered mafic–ultramafic intrusions. Continued subduction of oceanic crust to the north-west led to collision and suturing of the Kimberley Craton with the rest of the North Australian Craton by ~1820 Ma, during the Halls Creek Orogeny (Figure 3). Folding and thrusting accompanied metamorphism. During and immediately following collision, plutons of granite and gabbro were intruded to form the Sally Downs Supersuite (1835–1805 Ma) at the same time as the intrusion of additional large, layered mafic–ultramafic bodies. As the Sally Downs Supersuite was being intruded, the Speewah Group was deposited ~1835 Ma in a foreland basin on the Kimberley

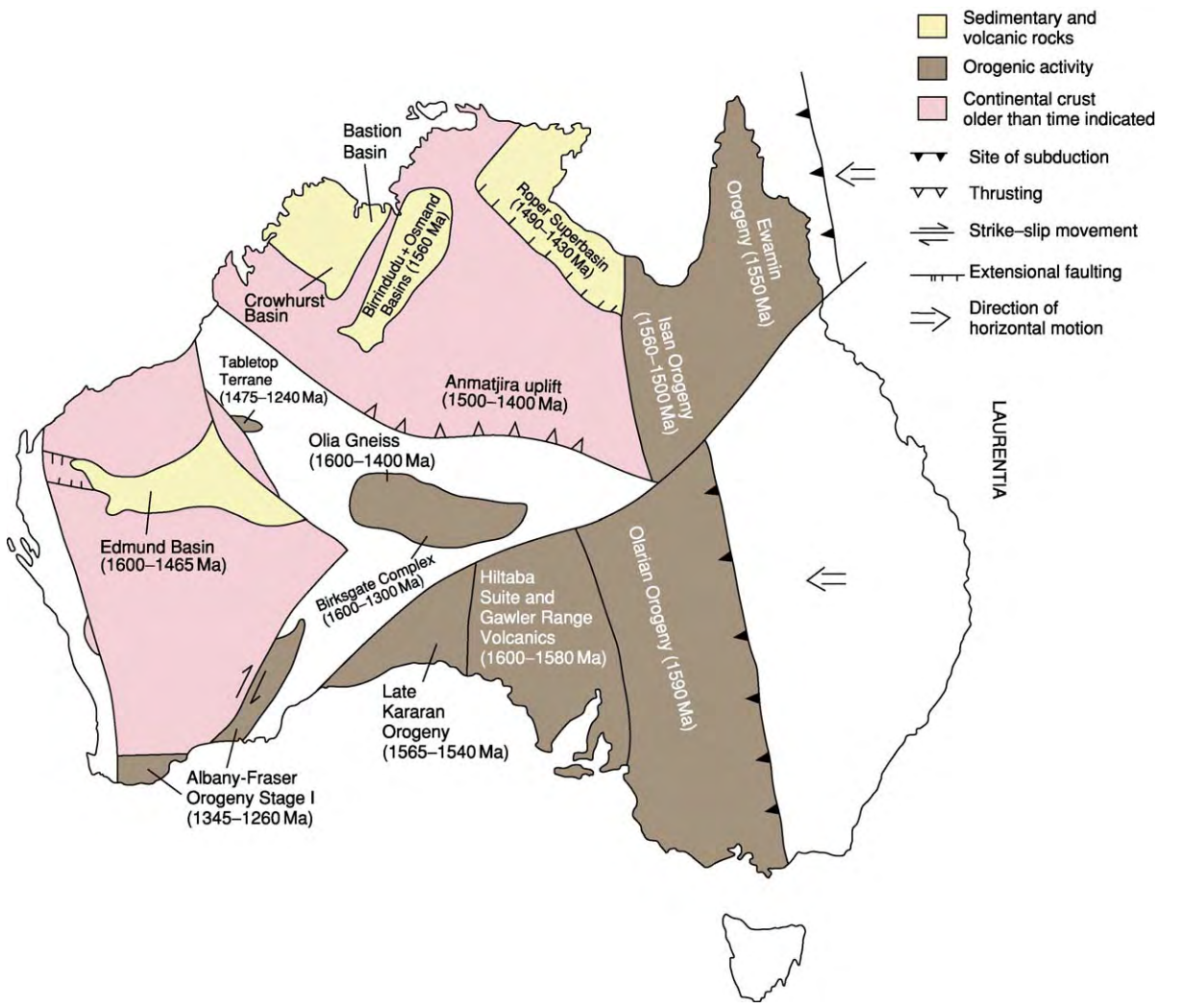


Figure 5 Docking with Laurentia, 1600 1200 Ma.

Craton. The overlying ~1800-My-old Kimberley Group was derived from the north. The intrusion of the voluminous Hart Dolerite ~1790 Ma may be related to a continental breakup, possibly from Laurentia, centred to the north (Figure 3).

To the south-east of the Halls Creek Orogen, the Granites–Tanami Complex and the northern part of the Arunta Inlier developed on the thinned Archaean to Early Palaeoproterozoic basement of the Lucas Craton (Figure 3). The <1877-My-old MacFarlane Peak Group and Dead Bullock Formation represent a rift, followed by a sag basin, with the overlying <1840-My-old Killi Killi Formation and Lander Rock beds turbidites being deposited during the Tanami Orogeny, which probably represents a within-plate response to the collisional Halls Creek Orogeny to the north-west. The orogenic event was followed by voluminous bimodal and granitic magmatism between 1825 and 1790 Ma in both the Granites–Tanami Complex and the northern Arunta Inlier, including the high-temperature/low-pressure metamorphism, deformation, and granite magmatism of the Mount Stafford tectonic event ~1820 Ma (Figure 3). Over the same time period, granite was intruded into the southern part of the Halls Creek Orogen. The Lichfield Complex is the north-eastward continuation of the Halls Creek Orogen, and the Pine Creek Orogen may represent a within-plate tectonic setting similar to that of the Granites–Tanami Complex (Figure 3). The ~1885-My-old South Alligator Group and the overlying Finniss River Group represent a transition from shallow-marine sediments to deep-marine turbidites, possibly related to the onset of the 'Barramundi Orogeny' in the form of the Nimbuwah event 1880–1860 Ma. Further deformation, equivalent to the Halls Creek and Tanami orogenies, accompanied extensive intrusion of granites between ~1840 and ~1820 Ma.

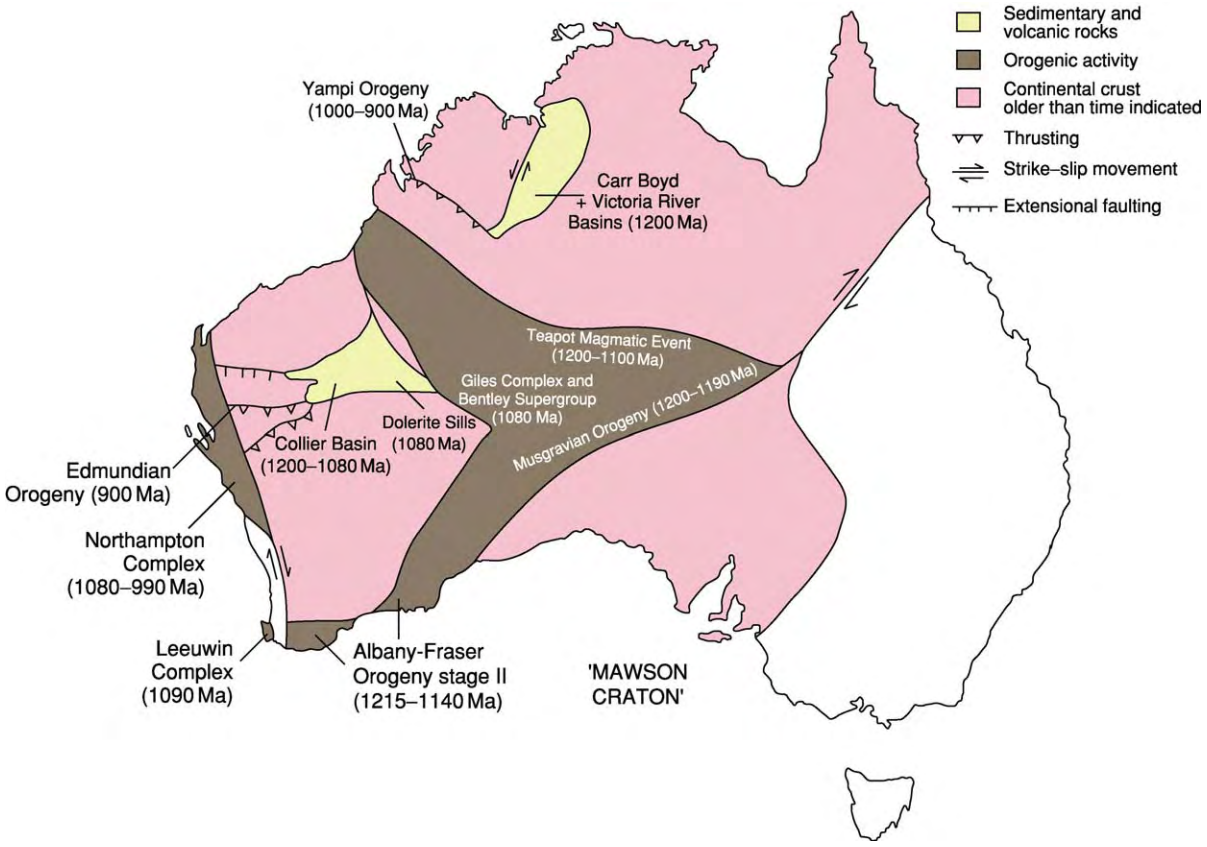


Figure 6 Intracratonic reactivation, 1200 900 Ma.

In the Mount Isa Inlier, the Kalkadoon–Leichhardt Belt from 1870–1840 Ma, now regarded as the remnants of a magmatic arc, represents the 'Barramundi Orogeny' (Figure 3). Along the southern edge of the northern Arunta Inlier, the Atnarpa igneous complex from 1880–1860 Ma has been interpreted as being generated by subduction at a convergent plate margin. Rocks of 'Barramundi Orogeny' age are also present in the Tennant Creek Inlier, where turbidites of the <1860-My-old Warramunga Formation are intruded by ~1850-My-old granites and are overlain by ~1845-My-old felsic volcanic rocks (Figure 3).

Between 1780 and 1730 Ma, a magmatic arc developed along the southern margin of the northern Arunta Inlier, above a northward-dipping subduction zone. The high-temperature/low-pressure metamorphism associated with the Early Strangways Orogeny 1780–1770 Ma overlapped the development of the magmatic arc. This event is contemporaneous with the Yapungku Orogeny (1790–1760 Ma) in the opposing margin of the West Australian Craton (Figure 3), and together they may record the suturing of those cratons, although probably not in their current configuration. The Late Strangways Orogeny (1740–1730 Ma) was accompanied by moderate-pressure metamorphism. Subduction of oceanic crust associated with the Early and Late Strangways orogenies was at a low angle to the north. A back-arc setting resulted, causing southward tilting within the crust to the north of the Arunta Inlier, together with high heat flows and extension followed by thermal subsidence. Between 1800 and 1670 Ma, two cycles of extensive volcano-sedimentary basin formation took place across the North Australian Craton, separated by a period of uplift and erosion. The first cycle (1800–1750 Ma) is represented by the development of the Leichhardt Superbasin (Figure 3). Bimodal volcanism, dominated by flood basalts, and fluvial to shallow-marine sedimentation occurred within the lower parts of the McArthur Basin successions (lower Katherine River Group, lower Donydji Group, Groote Eylandt Group, and lower Tawallah Group), the upper Hatches Creek Group in the Tennant Creek Inlier, and the lower parts of the successions within the Mount Isa Inlier (Haslingden Group, Quilalar Formation, Magna Lyn Metabasalt, Argylla Volcanics, Malbon Group, and Corella Formation). In the

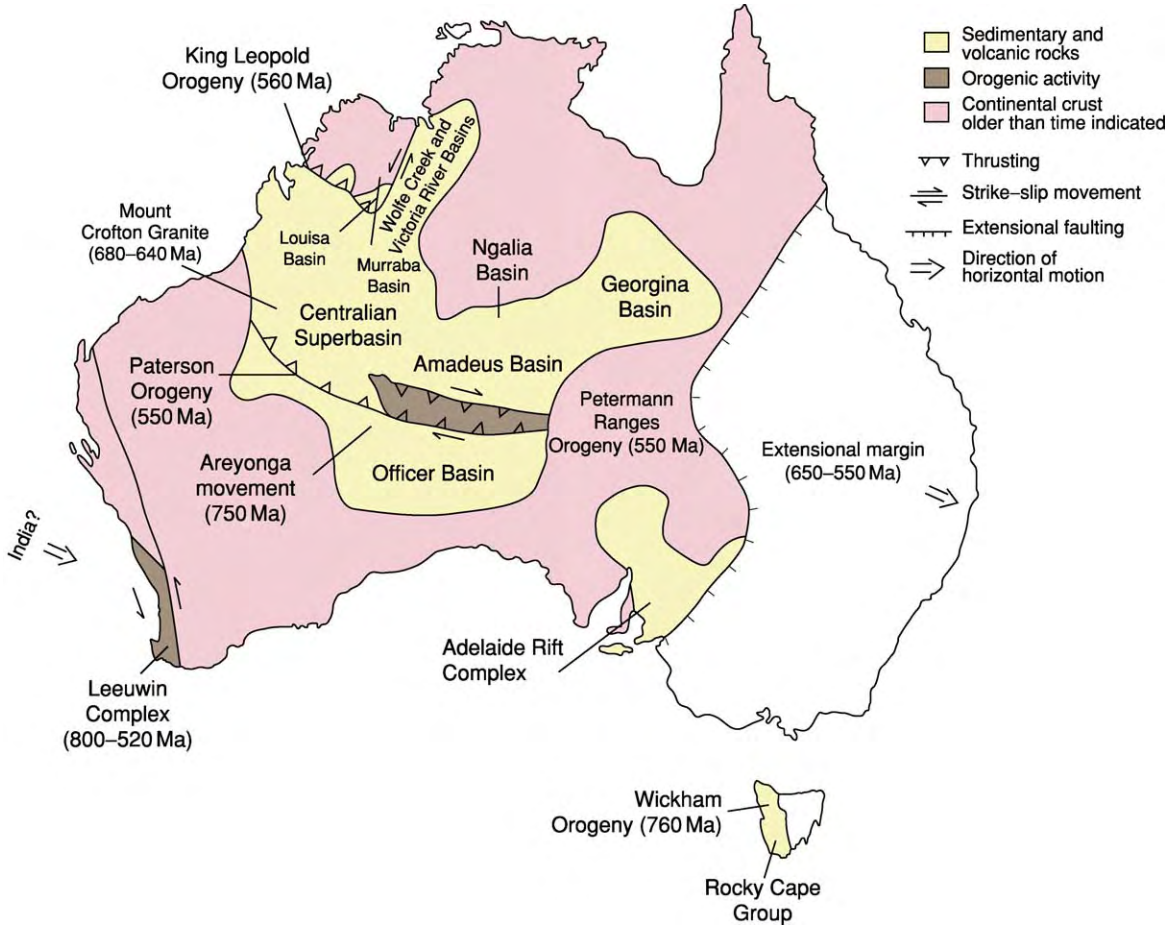


Figure 7 Rifting from Laurentia, collision with India, 900 520 Ma.

Mount Isa Inlier, rifting in the upper crust was accompanied by the development of a regional-scale extensional shear zone, high-temperature/low-pressure metamorphism, and igneous intrusion (the ~1742-My-old Wonga batholith) in the mid-crust.

A period of basin inversion ~1730 Ma produced a regionally extensive unconformity before the onset of the second depositional cycle (1730–1690 Ma), which formed the Calvert Superbasin, and the third depositional cycle (1665–1575 Ma), which formed the Isa Superbasin (Figure 4). The Sybella batholith was intruded in the Mount Isa Inlier ~1665 Ma. Again the successions are characterized by bimodal volcanics and rift-related sediments in the McArthur Basin (upper Katherine River Group, upper Donydji Group, Parsons Range Group, Spencer Creek Group, and upper Tawallah Group), the Mount Isa Inlier (Fiery Creek Volcanics, lower McNamara Group, lower Mount Isa Group, Kuridala Formation, and lower Soldiers Cap Group), and the Georgetown Inlier (lower Etheridge Group). The Redbank Thrust separates the southern Arunta Inlier from the northern Arunta Inlier. The Argilke tectonic event represents a significant magmatic and metamorphic event ~1680 Ma. A major north-directed thrust system, associated with granite intrusion, took place ~1610 Ma during the Chewings Orogeny (Figure 4).

South Australian Craton

In the Gawler Craton, Archaean gneissic rocks of the Sleaford and Mulgathing complexes were deformed, metamorphosed to granulite facies, and intruded by granites during the Sleafordian Orogeny 2550–2400 Ma (Figure 2). At the eastern margin of the Gawler Craton, orthogneiss of the Miltalie Gneiss (~2003 Ma) is unconformably overlain by shallow-water continental shelf deposits. These include metamorphosed amphibolite facies mafic volcanics, carbonates, and iron formation of the Hutchison Group >1860 Ma (Figure 2). The Kimban Orogeny has affected much of the Gawler Craton. It has been regarded as consisting of three tectonic events and dates from ~1850 to ~1690 Ma. It is contemporaneous with the Capricorn Orogeny and

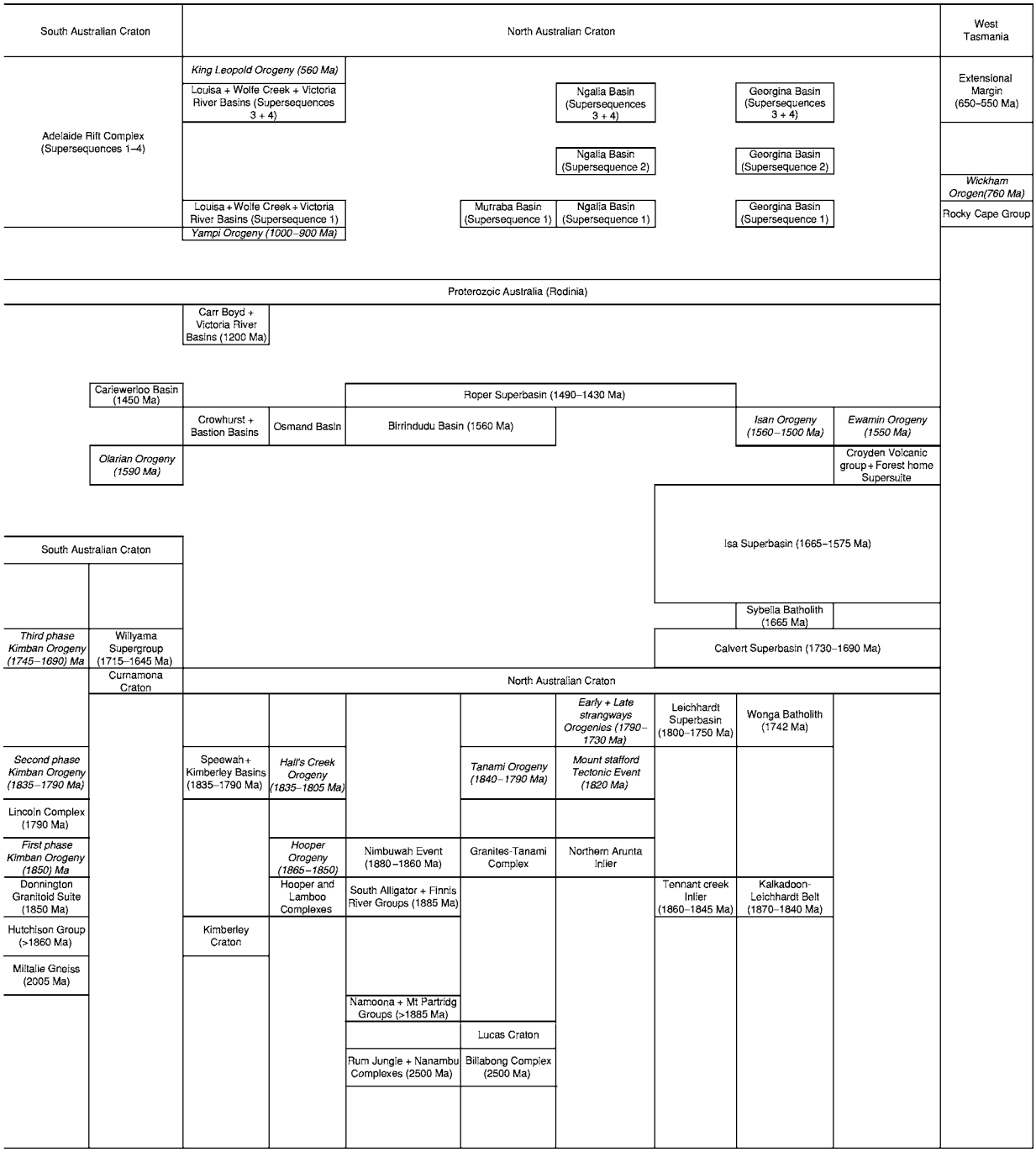
Table 1 Summary of Proterozoic tectonic events in Australia

Pinjarra Orogen	West Australian Craton			Albany-Fraser Orogen	Paterson Orogen			South Australian Craton	
Leeuwin Complex (800–520 Ma)					*Paterson Orogeny (550 Ma)*	*Petermann Ranges Orogeny (550 Ma)*			
	Western Officer Basin (Supersequences 3 + 4)						Amadeus Basin (supersequences 3 + 4)	Eastern Officer Basin (Supersequences 3 + 4)	
					Mount Crofton granite (680–640 Ma)		*Souths Range movement*		
							Amadeus Basin (Supersequence 2)	Eastern Officer Basin (Supersequence 2)	
				Aeryonga movement (750 Ma)					
	Officer + Amadeus Basins (Supersequence 1)								
Northampton + Leeuwin complexes (1090–990 Ma)	*Edmundian Orogeny (900 Ma)*								
	Dolerite sills (1080 Ma)					Giles Complex + Bentley Supergroup (1080 Ma)			
	Proterozoic Australia (Rodinia)								
	Collier Basin (1200–1080 Ma)			*Stage II Albany-Fraser Orogeny (1215–1140) Ma*		*Musgravian Orogeny (1200–1190 Ma)*	Teapot Magmatic Event (1200–1100 Ma)		
				Stage I Albany-Fraser Orogeny (1345–1260) Ma	Tabletop Terrane (1475–1290 Ma)				
	Edmund Basin (1600–1465 Ma)						*Anmatjira Uplift (1500–1400 Ma)*		
						Musgrave Complex: Olia gneiss + Birksgate Complex (1600–1300 Ma)		*Late Kararan Orogeny (1565–1540 Ma)*	
									Hiltaba Suite + Gawler Range volcanics (1600–1580 Ma)
	Bresnahan Basin							South Australian Craton	
	Reactivation of Capricorn Orogen (1670–1620 Ma)			Biranup Complex (1700–1600 Ma)			*Chewings Orogeny (1610 Ma)*	*Second Episode Kararan Orogeny (1630–1610 Ma)*	
			Mount Barren Group (1700 Ma)					St Peter's Suite (1630–1608 Ma)	
							Argilke Tectonic Event (1680 Ma)	*First episode Kararan Orogeny (1710–1670 Ma)*	
							Southern Arunta Inlier	Tunkillia Suite	
					Yapungku Orogeny (1790–1760 Ma)				
	Capricorn Orogeny (1830–1780 Ma)								
	Ashburton and Blair Basins (1840–1805 Ma)	Gascoyne Complex	Upper Yerrida and Earaheedy basins (1840–1800 Ma)		Rudall Complex (1800 Ma)				
	West Australian Craton								
		Glenburgh Orogeny 2005–1960 Ma	Bryah and Padbury Basins (2020–1900 Ma)		Rudall Complex (2020 1970 Ma)				
Mullingarra Complex (2181 Ma)	*Ophthalmian Orogeny (2200 Ma)*		Lower Yerrida Basin (2150 Ma)						
	McGrath Trough (2300–2200 Ma)								Gawler Craton
	Hamersley Basin (2800–2440 Ma)	Glenburgh Terrane (2550–2450 Ma)		Munglinup Gneiss (2630–2575 Ma)					Sleafordian Orogeny (2550–2400) Ma
	Pilbara Craton		Yilgarn Craton						Sleaford and Mulgathing complexes
	Granite-greenstones (3560–2800 Ma)		Granite-greenstones (3730–2600 Ma)						

the Yapungku Orogeny in the West Australian Craton, and the Early and Late Strangways orogenies in the North Australian Craton, and may represent a series of intraplate and plate margin events reflecting convergence and accretion of the Gawler Craton to what was then the south-eastern North Australian Craton.

Intrusive rocks of the Donnington Granitoid Suite (~1850 Ma) were emplaced synchronously with a deformation event that has been regarded as an early tectonic phase of the Kimban Orogeny (Figure 3). The second phase of the orogeny produced medium- to high-pressure/high-temperature metamorphism and associated deformation. Felsic and

Table 1 Continued



bimodal volcanics and interlayered metasedimentary rocks of the Myola Volcanics and the Broadview Schist, and the Peake Metamorphics were deposited on the Gawler Craton ~1790 Ma, contemporaneous with syntectonic granitic rocks of the Lincoln Complex. Further volcanic rocks and metasedimentary rocks were deposited between ~1765 and ~1735 Ma (McGregor Volcanics, Moonabie Formation, and Wallaroo Group). The third phase of the orogeny took place between 1745 and 1690 Ma and involved deformation, medium-grade metamorphism, and further granite intrusion (Figure 4).

Following accretion to the North Australian Craton, the site of subduction may have moved to

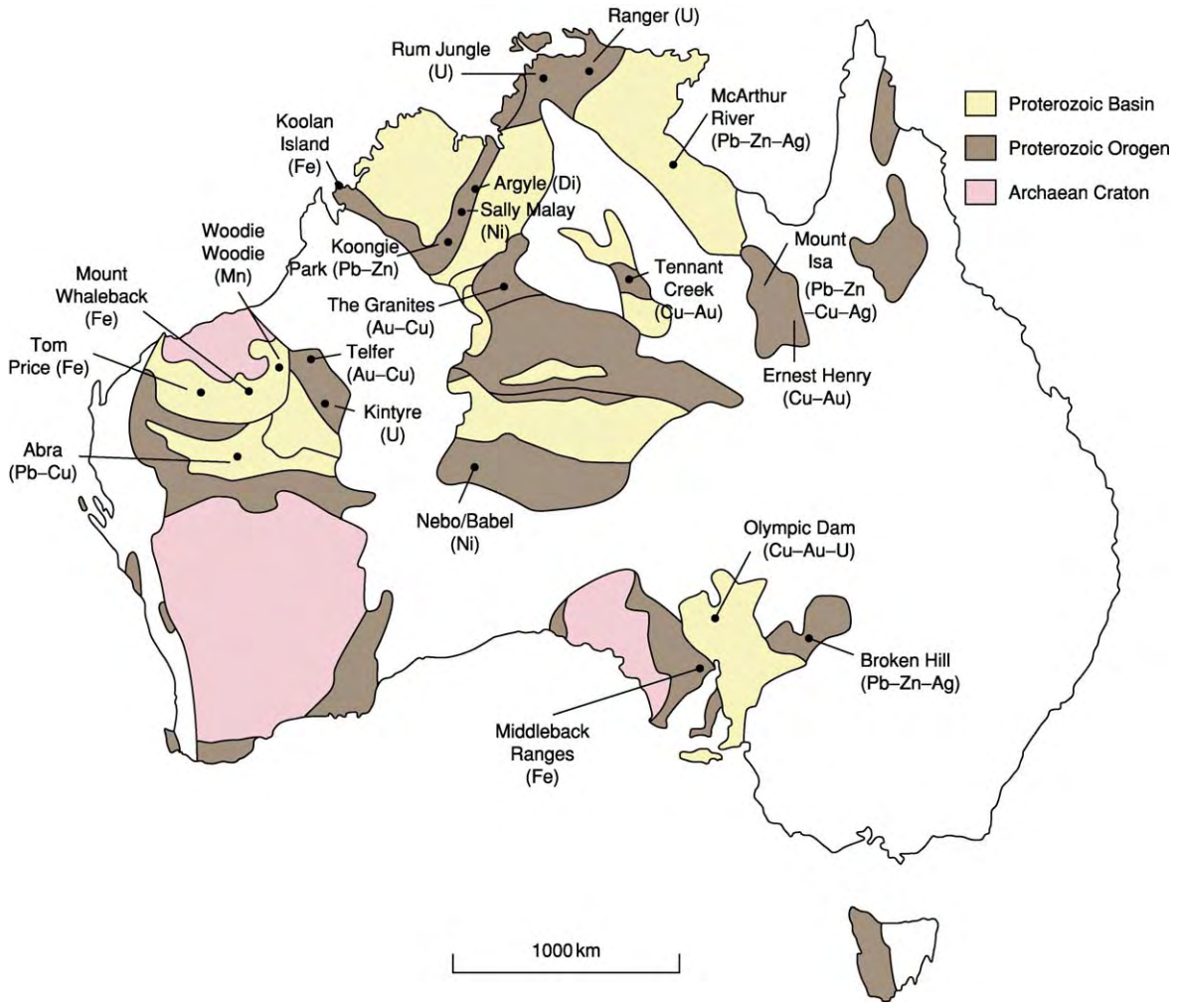


Figure 8 Significant mineral and diamond (Di) deposits in the Proterozoic of Australia.

the south-western Gawler Craton (Figure 4) with the development of the first tectonic episode of the Kararan Orogeny (1710–1670 Ma) and the intrusion of voluminous granitic rocks of the Tunkillia Suite in a possible back-arc setting. The second tectonic phase involved the intrusion of the St Peter Suite (1630–1608 Ma) as a continental magmatic arc, and probably reflects accretion of continental blocks from the west, now buried beneath the Phanerozoic Eucla Basin. The Kararan Orogeny is coincident with the Argilke and Chewings events in the southern Arunta Inlier, with the deposition of the Mount Barren Group and the reactivation of the Capricorn Orogen in the West Australian Craton, and with the development of the Calvert and Isa superbasins on the North Australian Craton (Figure 4). In the Curnamona Craton, the shallow-marine, immature clastic sedimentary rocks and bimodal volcanics of the Willyama Supergroup were deposited between ~1715 and ~1645 Ma. They may represent a continuation of the Calvert and Isa superbasins that has subsequently been displaced dextrally during the Mesoproterozoic (Figure 3).

Mesoproterozoic – The Assembly of Rodinia (1600–1000 Ma)

Along the eastern edge of the North Australian Craton and in the north-eastern part of the South Australian Craton, a major orogenic event took place between 1600 and 1500 Ma (Figure 5). This event included the Isan Orogeny (1560–1500 Ma) in the Mount Isa Inlier, the Ewamin Orogeny (1550 Ma) in the Georgetown Inlier, the Olarian Orogeny (1590 Ma) in the Curnamona Craton, and the late event (1565-1540 Ma) of the Kararan Orogeny. Subduction was to the west, culminating in collision with a continent to the east that may have been

Laurentia. In the Georgetown Inlier, the felsic rocks of the Croydon Volcanic Group and the Forest Home Supersuite represent a magmatic arc. Further to the west, the eruption of the voluminous Gawler Range Volcanics, and the intrusion of the Hiltaba Suite in the Gawler Craton, occurred between ~1600 and ~1580 Ma, contemporaneously with the early stages of this orogenic event.

Elsewhere in central and western Australia, Mesoproterozoic rocks of poorly understood tectonic setting include medium- to high-grade ~1600-My-old quartzofeldspathic gneisses (Olia Gneiss) to the north of the Woodroffe Thrust in the Musgrave Complex. These were intruded by granitic rocks ~1500 Ma, before being metamorphosed ~1400 Ma. South of the Woodroffe Thrust, metasedimentary gneisses and felsic and mafic orthogneisses (~1550 and 1330 Ma) represent a separate terrane. Granulites in the Birksgate Complex at the northern margin of the South Australian Craton were derived from metamorphism of felsic volcanics and sedimentary rocks (1600–1300 Ma). Uplift along the Redbank Thrust Zone at the northern margin of the southern Arunta Inlier took place 1500–1400 Ma (Anmatjira uplift phase). The eastern Tabletop Terrane of the Rudall Complex is formed by metasedimentary and metavolcanic rocks intruded by granitoids that date to ~1475 to ~1290 Ma.

On the North Australian Craton, clastic and carbonate sedimentary rocks of the Birrindudu Group were deposited in the Birrindudu Basin ~1560 Ma, unconformably overlying the Granites–Tanami Complex. In the Kimberley region, the Mount Parker Sandstone and the Bungle Bungle Dolomite in the Osmand Basin, and the Crowhurst Group and the Bastion Group, may be equivalent to the rocks of the Birrindudu Basin. The Roper Group and the South Nicholson Group were deposited in the intracratonic Roper Superbasin between ~1490 and ~1430 Ma, and subsidence may be related to late stages of the Isan Orogeny, and to the Anmatjira uplift phase in the Arunta Inlier (Figure 5). The intracratonic Cariewerloo Basin (~1450 Ma) developed on the South Australian Craton. On the West Australian Craton, the Edmund Group, the lower part of the Bangemall Supergroup, was deposited in the intracratonic Edmund Basin after ~1600 Ma. Initial deposition was restricted to narrow rift basins before the development of a broad, marine basin. Extensive intrusion of dolerite sills into the Edmund Group took place ~1465 Ma.

The Albany–Fraser Orogen developed between the South Australian Craton and the West Australian Craton (Figures 5 and 6). It has been interpreted as part of a Grenvillian collisional orogeny (1345–1140 Ma) that originally extended through the Musgrave Complex and possibly the now displaced eastern Rudall Complex, separating the South Australian Craton from the combined West and North Australian cratons. This event has been interpreted as marking the final assembly of Proterozoic Australia as part of the Rodinia Supercontinent. Two distinct tectonic stages took place in the Albany–Fraser Orogen, with the first producing deformation, high-grade metamorphism, and granitic magmatism between 1345 and 1260 Ma. The granulite facies layered mafic intrusions of the Fraser Complex crystallized ~1300 Ma. The second phase produced further granulite facies metamorphism and deformation ~1214 Ma and subsequent granite intrusion between 1190 and 1140 Ma. In the Tabletop Terrane of the Rudall Complex, granitic magmatism occurred ~1300 Ma, whereas in the southern part of the Musgrave Complex, granulite facies metamorphism occurred ~1200 Ma, with the intrusion of granitic rocks ~1190 Ma. Granitic rocks were also intruded into the southern Arunta Inlier during the Teapot magmatic event between 1200 and 1100 Ma. In northern Australia, the ~1200-My-old Auvergne Group and the Carr Boyd Group were deposited in the Victoria River and Carr Boyd Basins, respectively (Figure 6).

The Collier Group, the upper part of the Bangemall Supergroup, was deposited in the Collier Basin on the West Australian Craton and, together with the underlying Edmund Group, was intruded by a suite of syn-depositional dolerite sills ~1080 Ma. Large, layered mafic–ultramafic intrusions forming the Giles Complex (~1080 Ma) were intruded into the Musgrave Complex, followed by the eruption of the mafic and felsic volcanic rocks of the Bentley Supergroup 1080–1060 Ma. Together with the intrusion of the contemporaneous Stuart and Kulgera dyke swarms, the extensive mafic magmatism of 1080–1060 Ma has been described as a Large Igneous Province, and may represent the influence of a mantle plume centred beneath the Musgrave Complex (*see* **Large Igneous Provinces**).

In the Pinjarra Orogen, metasedimentary rocks in the Northampton Complex and Mullingarra Complex that may have had their source in the Albany–Fraser Orogen were metamorphosed, deformed, and intruded by granites between ~1080 and ~990 Ma. They are probably allochthonous and may have been transported and accreted in their present position by dextral movement along a proto-Darling Fault prior to 755 Ma (Figure 6). The protolith to ~1090-My-old granitic orthogneiss in the Leeuwin Complex may have a syn-collisional origin. Intracratonic reactivation of the King Leopold, Halls

Creek, and Capricorn orogenic belts took place between 1000 and 900 Ma during the Yampi and Edmundian orogenies (Figure 6).

Neoproterozoic-Proterozoic Australia in Rodinia (1000–545 Ma)

The Centralian Superbasin developed throughout much of Proterozoic Australia from ~830 Ma, either as an extensive intracratonic sag basin or as a series of interconnected basins that lapped onto intervening, emergent basement highs (Figure 7). The Adelaide Rift Complex developed to the south-east at the same time, possibly centred over a mantle plume. The lower part of the depositional succession in the superbasin, Supersequence 1, is contemporaneous with the intrusion of the Amata and Gairdner mafic dykes in the Gawler Craton and eastern Musgrave Complex, and the volcanics of the Callana Group in the Adelaide Rift Complex. Supersequence 1 is characterized by a basal sand sheet, overlain by stromatolitic carbonates and evaporates that are correlated throughout the component basins of the superbasin (Officer Basin, Amadeus Basin, Ngalia Basin, Georgina Basin, Wolfe Creek Basin, Louisa Basin, and Murraba Basin). Siliciclastic rocks of the Rocky Cape Group and the Burnie Formation in west Tasmania may be equivalent to Supersequence 1, deposited prior to the Wickham Orogeny ~760 Ma.

Deposition of Supersequence 1 was followed by a period of tectonic activity, the Areyonga movement, which may represent the breakup of Rodinia as Laurentia rifted away from the eastern margin of Proterozoic Australia to form the proto-Pacific Ocean. The northern part of the Adelaide Rift Complex between the Gawler and Curnamona Cratons has been interpreted as a failed arm, but there is no evidence of the development of an adjacent passive margin at this time. The main locus of Rodinia breakup may have developed much further to the east, now buried within the Tasmanides. Tilting and flexuring of basement fault blocks to the west within the Centralian Superbasin was controlled by the reactivation of major structures such as the Redbank Thrust, and a proto-Woodroffe Thrust. Widespread folding, uplift, and erosion took place within the basins. The 755-My-old Mundine Well dyke swarm was intruded into the West Australian Craton and the adjacent Northampton Complex.

Supersequece 2 is restricted to northern and central Australia and is marked by glacigene deposits correlated with the Sturtian glaciation ~700 Ma in the Adelaide Rift Complex. As sea-level rose, the glacial deposits were overlain by silt and mud deposited in a shallow epeiric sea. This was followed by a further period of uplift and erosion, the South Range Movement, and by the emplacement of the ~680- to 640-My-old Mount Crofton Granite and associated intrusions into the Lamil Group of the Yeneena Basin in the north-western part of the Paterson Orogen. Supersequence 3 is also marked by glacigene deposits, which extend into the Kimberley region and the north-west Officer Basin. These are correlated with the ~600-My-old Marinoan glaciation in the Adelaide Rift Complex. Again, rising sea-levels and the establishment of a shallow epeiric sea followed glaciation.

Supersequence 4 is dominated by sandstone and conglomerate deposited in submarine fan, deltaic, fluvial, and alluvial fan settings in local foreland basins. These are associated with the development of intracratonic orogenies, which resulted from the reactivation of the ancient sutures between the North Australian Craton, the West Australian Craton, and the South Australian Craton during the Paterson and Petermann Ranges orogenies ~550 Ma (Figure 7). In the Kimberley region, the King Leopold Orogeny reactivated the suture between the Kimberley Craton and the rest of the North Australian Craton at about the same time. Large dextral strike–slip movements produced folding and thrusting in the Musgrave Complex, with the exhumation of eclogite facies rocks as part of a crustal-scale flower-type structure. The Musgrave Complex now truncates the Albany–Fraser Orogen, and may have been displaced to the west at this time, extending as far as the Mesoproterozoic Tabletop Terrane of the Rudall Complex, with which the Warri–Anketell Gravity Ridge connects it beneath the Centralian Superbasin.

In the Pinjarra Orogen at the western margin of Proterozoic Australia, Leeuwin Complex orthogneisses from ~1090 and 800–650 Ma were metamorphosed at upper amphibolite to granulite facies conditions ~540 Ma. Further granitic rocks were intruded 540–520 Ma. These events were coincident with sinistral tectonic transport of the Leeuwin Complex along the Darling Fault during an oblique collision of Australia with India (Figure 7). This collision may be responsible for the contemporaneous intracratonic reactivations that produced major unconformities at the base of the Phanerozoic basins overlying Proterozoic Australia to the west of the Adelaide Rift Complex. An extensional margin developed to the east of the Adelaide Rift Complex between ~650 and ~550 Ma, and is now exposed in Tasmania and western Victoria (Figure 7). It is marked by the intrusion of a dolerite dyke swarm, the eruption of tholeiitic basalts, and the deposition of associated volcanogenic sediments, carbonates, and shallow-water siliciclastics. This margin was involved in an arc–continent collision ~505 Ma during the Delamarian Orogeny, initiating the development of

the Palaeozoic Lachlan Orogen. Extensive eruption of continental flood basalts (Kalkarinji large igneous province) of the ~510-My-old Antrim Plateau Volcanics and the Table Hill Volcanics in northern and western Australia may reflect further mantle plume activity coincident with this collision.

Mineral Deposits

The Proterozoic rocks of Australia are endowed with substantial mineral resources, containing significant deposits of iron ore, manganese, uranium, gold, copper, lead, zinc, silver, nickel, and diamond (Figure 8) (*see* **Mineral Deposits and Their Genesis**). World-class hematite deposits (Mount Whaleback, Tom Price) have been produced by enrichment of the banded iron formations of the Hamersley Basin during the Palaeoproterozoic. The highest grade low-P deposits (62–69% Fe), have been interpreted as the product of burial metamorphism of original supergene enrichment prior to ~1840 Ma. More recently, they have been reinterpreted as the product of hypogene processes related to the expulsion of fluids into the foreland of the Ophthalmian Orogeny (~2200 Ma). Proterozoic iron ore deposits have also been mined in the Palaeoproterozoic iron formations of the Hutchison Group in the Gawler Craton, and from enrichment of ferruginous placer deposits in the upper Kimberley Group in the Kimberley Basin (Koolan Island). Manganese has been mined from supergene deposits (Woodie Woodie) associated with Palaeoproterozoic karst developed on dolomites in the eastern Hamersley Basin.

'Unconformity-related' uranium (± Au ± platinum group elements) deposits (Rum Jungle, Alligator River) are found in Palaeoproterozoic sedimentary successions overlying predominantly Archaean basement rocks in northern Australia, and may be related to interaction of highly oxidized, acidic and Ca-rich meteoric brine with reduced basement fluids. Similar deposits have been found elsewhere in the Australian Proterozoic, including Kintyre, where Neoproterozoic metasedimentary rocks overlie Palaeoproterozoic basement. Metasomatic U deposits and the metamorphic-related Mary Kathleen deposit occur in Palaeoproterozoic rocks of the Mount Isa Inlier.

Gold is found in a number of settings throughout the Proterozoic, and historic production is related to generally small lode-gold occurrences and associated alluvial deposits. Major gold (Cu) mines have been developed at Telfer in the Yeneena Basin, and at The Granites in the Granites–Tanami Complex. Mineralization is centred on metasedimentary rocks in structural domes and is hypogene, related to the emplacement of granitic pluton(s) near a periodically reactivated, regional-scale, fluid-focusing structure, often a strike–slip fault. 'Proterozoic Cu–Au deposits' are found associated with iron oxide in the Tennant Creek Inlier, at Olympic Dam in the northern Gawler Craton, and in the Mount Isa Inlier (Ernest Henry). There is usually a spatial and temporal relationship between this style of deposit and granite intrusion (e.g., Olympic Dam and the ~1590-My-old Hiltaba Suite). Cu mineralization at Mount Isa is regarded as syn-deformational, late metamorphic, and is a separate event from Pb–Zn–Ag mineralization. The Mount Isa orebody is a world-class example of stratiform sediment-hosted Pb–Zn–Ag mineralization, produced by oxidized fluids moving through the sediment pile and being deposited either by seafloor exhalative processes or within the sediments. A similar orebody is present at McArthur River. The Broken Hill orebody is generally regarded as a metamorphosed example of stratiform sediment-hosted mineralization, although syn-metamorphic, skarn-type processes may have modified it. Volcanic-hosted massive sulphide (VHMS) Pb–Zn mineralization has been found at Koongie Park in the Halls Creek Orogen.

Nickel mineralization in the Proterozoic, together with platinum group elements, Cu, and V, is generally associated with layered mafic–ultramafic intrusions such as those in the Halls Creek Orogen (Sally Malay). Large, layered intrusions are also present in the Musgrave Complex (Giles Complex–Nebo and Babel deposits), the Albany–Fraser Orogen (Fraser Complex), and the Arunta Inlier.

The Argyle diamond mine, which is the world's largest, is developed on the ~1200-My-old AK1 lamproite pipe intruded into the Halls Creek Orogen. Proterozoic diamondiferous kimberlites from ~815 Ma are present in the Kimberley Basin (*see* **Igneous Rocks:** Kimberlite).

See Also

Igneous Processes. **Igneous Rocks:** Kimberlite. **Large Igneous Provinces**. **Mantle Plumes and Hot Spots**. **Metamorphic Rocks:** Classification, Nomenclature and Formation. **Mineral Deposits and Their Genesis**. **Mining Geology:** Exploration; Hydrothermal Ores. **Plate Tectonics**. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Mineralogy and Classification; Banded Iron Formations. **Tectonics:** Mountain Building and Orogeny. **Time Scale**.

Further Reading

Australian Geological Survey Organisation (1998) Geology and mineral potential of major Australian mineral provinces. *AGSO Journal of Australian Geology and Geophysics* 17(3): 1–260.

Australian Geological Survey Organisation (1998) Exploration models for major Australian mineral deposit types. *AGSO Journal of Australian Geology and Geophysics* 17(4): 1–313.

Cawood PA and Tyler IM (2004) Assembling and reactivating the Proterozoic Capricorn Orogen: lithotectonic elements, orogenies, and significance. *Precambrian Research* 128: 201–218.

Collins WJ and Shaw RD (eds.) (1995) *Time limits on tectonic events and crustal evolution using geochronology: some Australian examples Precambrian Research* 71: 1–346.

Direen NG and Crawford AJ (2003) The Tasman Line: where is it, what is it, and is it Australia's Rodinian breakup boundary? *Australian Journal of Earth Sciences* 50: 491–502.

Drexel JF and Parker AJ (1993) *The Geology of South Australia: Volume 1. The Precambrian. Geological Survey of South Australia, Bulletin 54*. Adelaide: Geological Survey of South Australia.

Fitzsimons ICW (2003) Proterozoic basement provinces of southern and south western Australia, and their correlation with Antarctica. In: Yoshida M, Windley BF, and Dasgupta S (eds.) *Proterozoic East Gondwana: Supercontinent Assembly and Breakup. Geological Society of London, Special Publication 206*, pp. 93–130. London: Geological Society of London.

Geological Survey of Western Australia (1990) *Geology and Mineral Resources of Western Australia. Western Australia Geological Survey, Memoir 3*. East Perth: Geological Survey of Western Australia.

Glikson AY, Stewart AJ, Ballhaus CG, Clarke GL, Feeken EHJ, Leven JH, *et al.* (1996) *Geology of the Western Musgrave Block, Central Australia, with Particular Reference to the Mafic–Ultramafic Giles Complex. Australian Geological Survey Organisation, Bulletin 239*. Canberra: Australian Geological Survey Organisation.

Hoatson DM and Blake DH (2000) *Geology and Economic Potential of the Palaeoproterozoic Layered Mafic–Ultramafic Intrusions in the East Kimberley, Western Australia. Australian Geological Survey Organisation, Bulletin 246*. Canberra: Australian Geological Survey Organisation.

Hunter DR (1981) *Precambrian of the Southern Hemisphere. Developments in Precambrian Geology 2*. Amsterdam: Elsevier.

Korsch RJ (ed.) (2002) Thematic issue: geodynamics of Australia and its mineral systems: technologies, syntheses and regional studies. *Australian Journal of Earth Sciences* 49(4): 593–771.

Myers JS, Shaw RD, and Tyler IM (1996) Tectonic evolution of Proterozoic Australia. *Tectonics* 15: 1431–1446.

Powell C McA and Meert JG (eds.) (2001) Assembly and breakup of Rodinia. *Precambrian Research* 110(1–4): 1–386.

Southgate PN (ed.) (2000) Thematic issue: Carpentaria–Mt Isa zinc belt: basement framework, chronostratigraphy and geodynamic evolution of Proterozoic successions. *Australian Journal of Earth Sciences* 47(3): 337–657.

Walter MR (ed.) (2000) Neoproterozoic of Australia. *Precambrian Research* 100(1–3): 1–433.

Phanerozoic

J J Veevers, Macquarie University, Sydney, NSW, Australia

Introduction

This account of Phanerozoic (544–0 Ma) Australia is told through a timetable of events (**Figure 1**), a set of fold-belt cross-sections (**Figure 2**), and a set of maps, which stretch from the initial Pan-Gondwanaland phase to the tectonics of today (**Figures 3–27**).

Phanerozoic Australia developed through three stratitectonic regimes (uniform plate tectonics and climate at similar or slowly changing latitude), which generated depositional successions of distinct facies that are bounded by unconformities at the margins and by stratigraphical lacunas in the interior (**Figure 1**). These stratitectonic regimes are summarized in **Table 1**.

Terranes successively broke off the west and northwest along divergent margins with rim basins. The eastern convergent margin grew by accretions of arc-subduction complexes during successive rollback of the trench. Changes of regime were driven by global events (*see* **Gondwanaland and Gondwana; Australia:** Tasman Orogenic Belt): the 544 Ma aftermath of the Pan-Gondwanaland collisional rotations and the onset of an active Pacific margin; the 320 Ma collisional merger of Gondwanaland and Laurussia to form Pangaea and the subsequent far-field contractional stress in central Australia, with ensuing tectonics driven by Pangaean-induced heat; and the 99 Ma change in the azimuth of subduction of the Pacific Plate from head-on to side-swipe.

The 650–580 Ma collision of East Gondwanaland and West Gondwanaland resulting from the closure of the Mozambique Ocean, together with continued Pan-Gondwanaland contraction until 500 Ma, generated the Mozambique Orogenic Belt and an

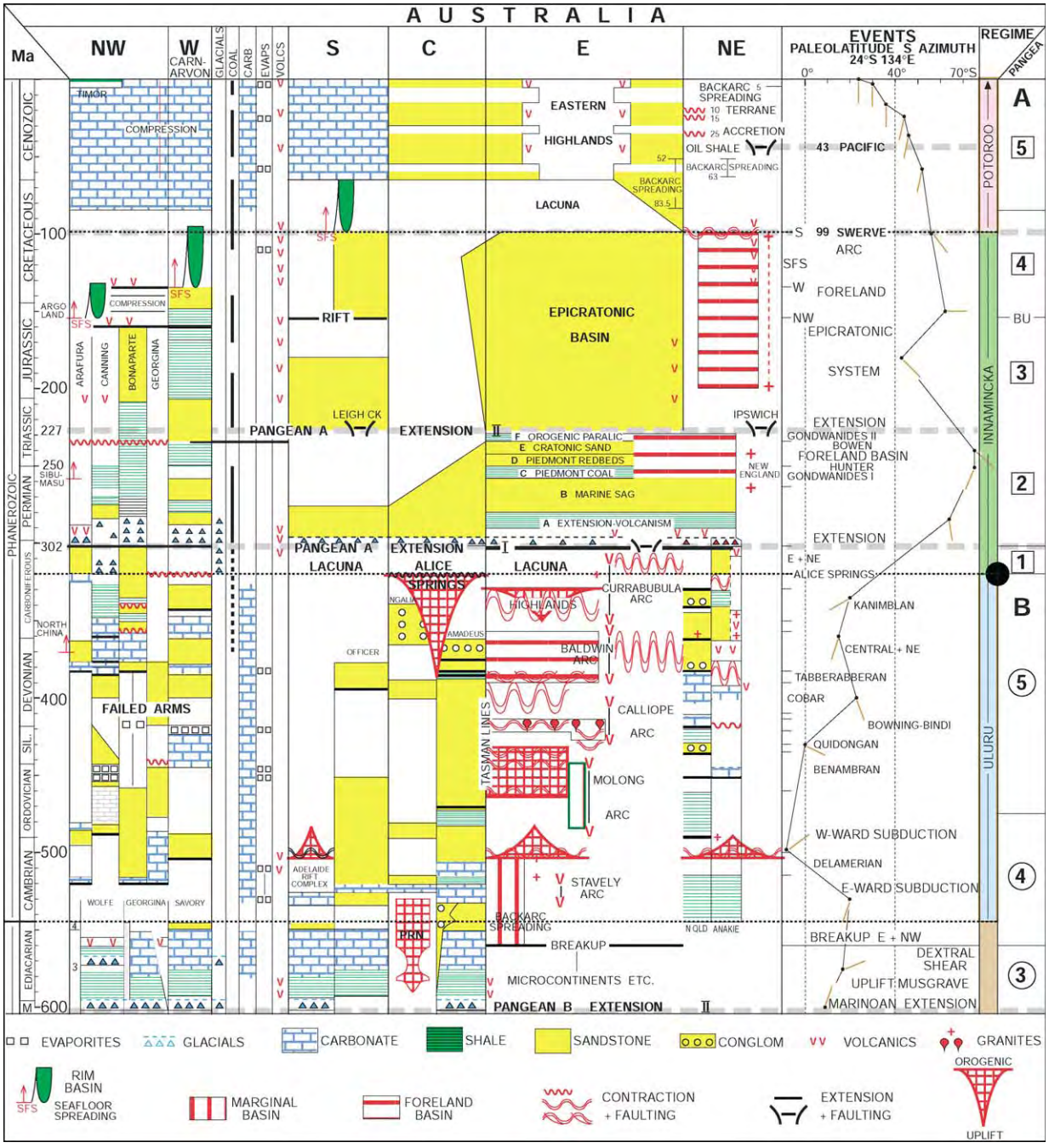


Figure 1 Timetable of events during the past 600 Ma. Columns from left to right are divergent margins (northwest (NW) with break off of terranes and west (W) represented by the Carnarvon Basin), depositional facies (glacial, coal, carbonate, evaporites, volcanics), and southern region (S), the centre (C), across the Tasman Lines to the Tasman Fold Belt System in the east (E) and north east (NE). Under 'events', the palaeolatitude of Alice Springs is plotted and the chief deformations are listed. Under 'regime', the last Neoproterozoic supersequences and the three Phanerozoic regimes are indicated. Under 'Pangaea', the stages of Pangaea B and A are indicated and the 320 Ma merger is marked by a filled circle. BU, breakup; carb, carbonate; evaps, evaporites; PRN, Petermann Ranges Nappe; QLD, Queensland; SFS, seafloor spreading; volcs, volcanics. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

orthogonal belt through Prydz Bay to Cape Leeuwin, with distributed thrusting and associated metamorphism in the rest of Australia. These events, together with the inception of subduction of the Palaeo-Pacific Plate beneath Antarctica–Australia at about 550 Ma, inaugurated Phanerozoic Australia.

The events along the Pacific margin during the Palaeozoic are shown in the set of cross-sections in

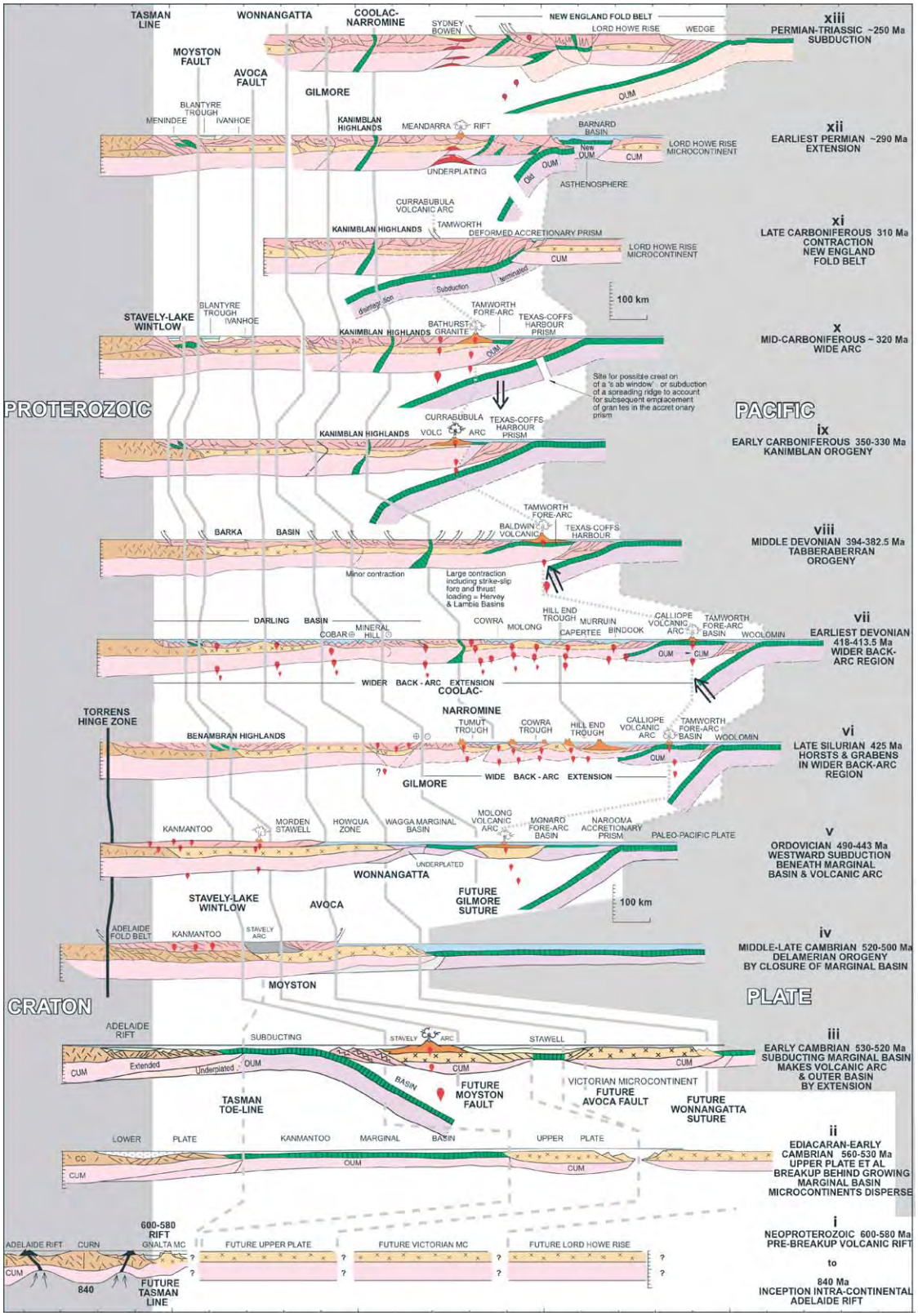


Figure 2 Cross sections of the lithosphere of the Tasman Fold Belt System developing between the fixed Proterozoic Craton and the active Pacific Plate. Ratio of vertical to horizontal scales is 1 : 1; the vertical scale has ticks every 10 km; the horizontal scale has ticks every 100 km. CC, continental crust; CUM, continental upper mantle; OUM, oceanic upper mantle; CURN, Curnamona Craton; MC, microcontinent. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

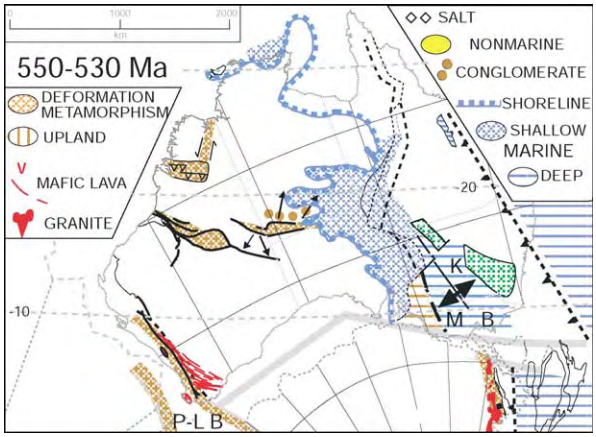


Figure 3 Latest Neoproterozoic earliest Cambrian (550 530 Ma) palaeogeography. The parallel lines crossing central eastern Australia are the Tasman Line (heavy broken line) and the Tasman Toe Line (light broken line), which mark the eastern outcrops of the Proterozoic craton and its wedge out, respect ively. KMB, Kanmantoo Marginal Basin; P LB, Prydz Leeuwin Belt. Reproduced with permission from Veevers JJ (ed.) (2000) *Bil lion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

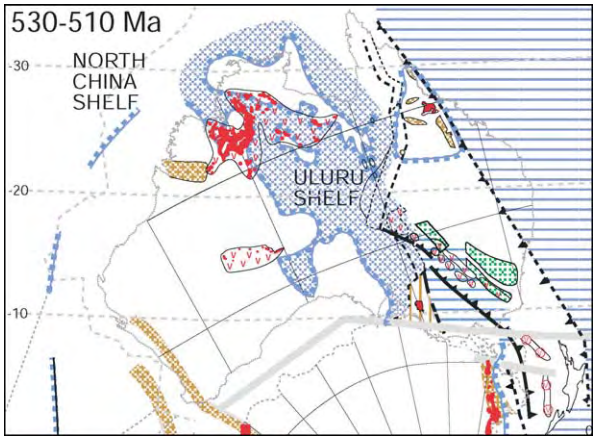


Figure 4 Early and Middle Cambrian (530 510 Ma) palaeo geography. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

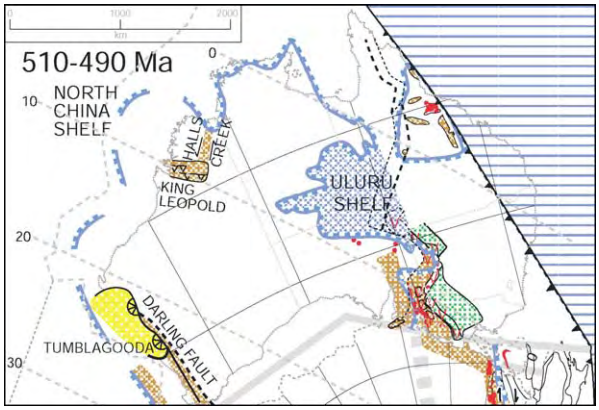


Figure 5 Late Cambrian (510 490 Ma) palaeogeography. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

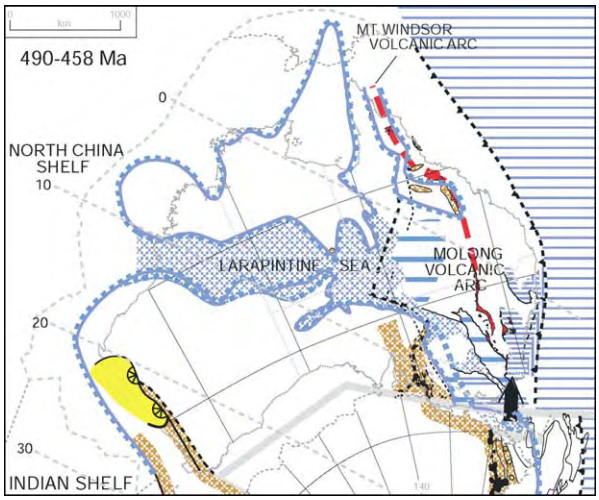


Figure 6 Early Ordovician (490 458 Ma) palaeogeography. The cooling Ross and Delamerian granites are shown in black. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Figure 2. The initial dispersal of microcontinents by growth of a marginal basin and their return by closure of the basin were followed by subduction, generating a magmatic arc behind which granites were emplaced in an extended region. When the subducted slab shallowed, the crust was shortened; when it steepened, the trench rolled back in front of a growing marginal basin.

The events across Australia and neighbouring Antarctica are shown in plan view in the set of palaeogeographical figures.

Latest Neoproterozoic–Earliest Cambrian (550–530 Ma)

The Uluru regime was inaugurated by collisions in West Gondwanaland and subduction in Antarctica, which deformed central and north-western Australia to produce mountain ranges and intervening basins where coarse arkose was deposited at the foots of the mountains, as at Uluru (Figure 3). In the south-west, the Prydz-Leeuwin Belt was terminally metamorphosed, and its periphery was intruded by a swarm of dykes.

In the south-east, the Kanmantoo Marginal Basin grew by back-arc spreading behind the subducting

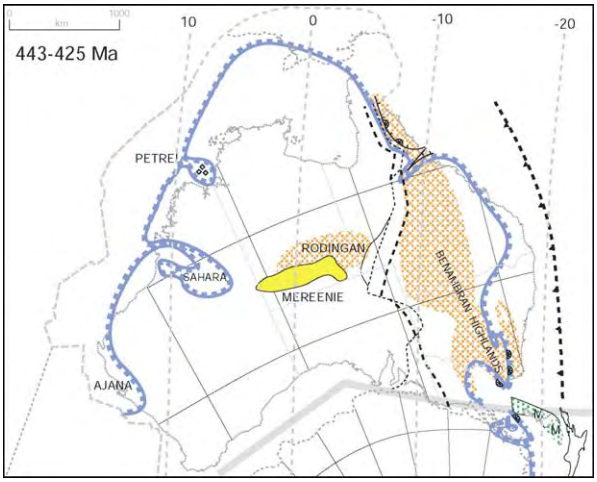


Figure 7 Early Silurian (443 425 Ma) palaeogeography. MM, Melbourne Mathinna Terrane. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

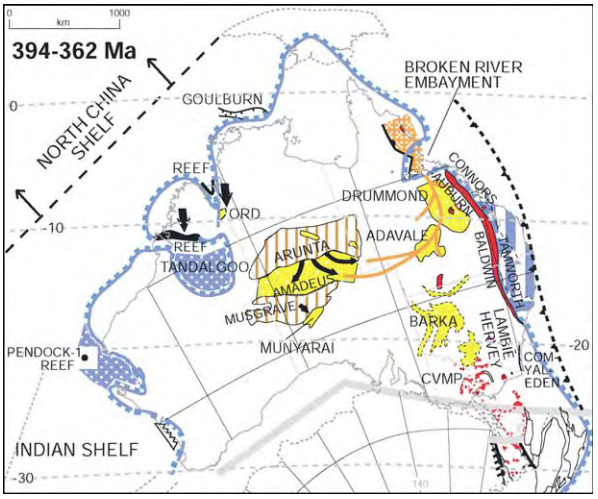


Figure 9 Middle and Late Devonian (394 362 Ma) palaeogeography. COM YAL EDEN, Comerong Yalwal Eden Rift Zone; ORD, Ord Basin; CVMP, Central Victorian Magmatic Province. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

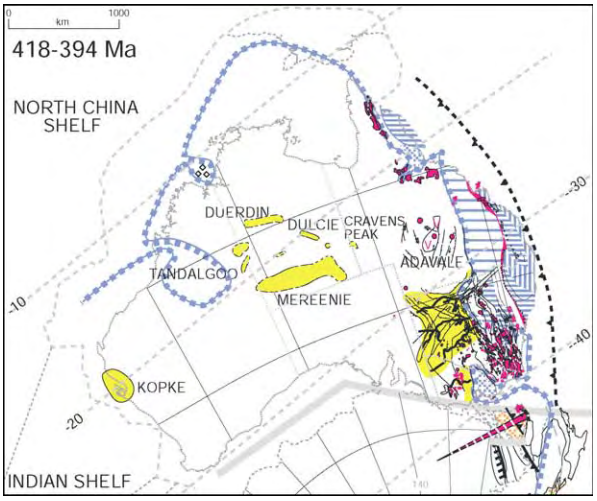


Figure 8 Early Devonian (418 394 Ma) palaeogeography. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

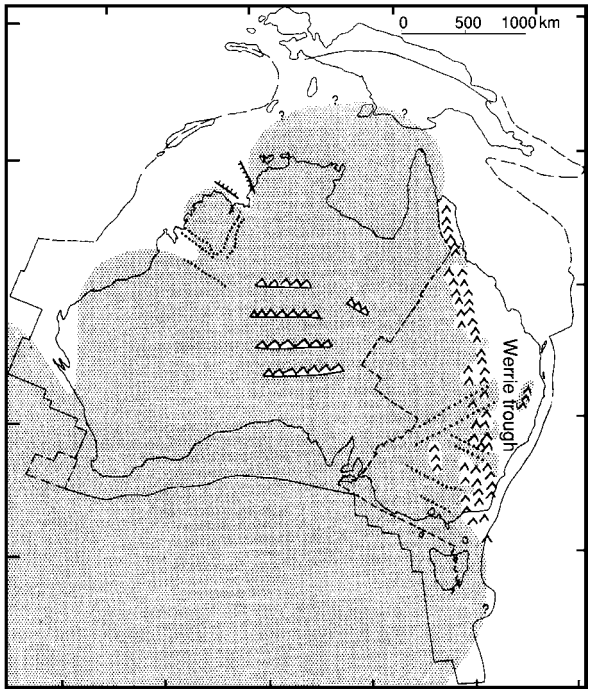


Figure 10 Postulated extent of the mid Carboniferous ice sheet. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Palaeo-Pacific Plate and wedged microcontinents off the mainland in the first episode of the Tasman Fold Belt System (**Figure 2ii**). In the north-east, an accretionary wedge grew in front of the trench. Central Australia was flooded by a shallow sea that lapped over the Uluru arkose, and another arm of the sea crossed the north. On the Antarctic side of a transform fault, another accretionary wedge grew in front of the trench and the first Ross granites were emplaced along the Transantarctic Mountains.

Early–Middle Cambrian 530–510 Ma

Basalt flows over much of the north and centre (**Figure 4**) were crossed by the Uluru Shelf, which connected with the North China Shelf. An island in

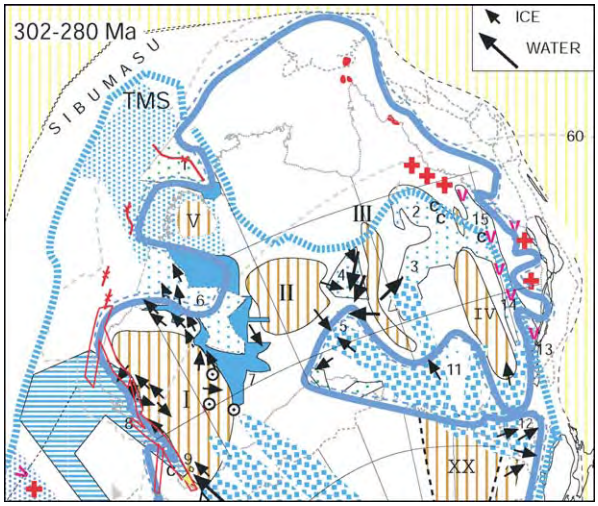


Figure 11 Latest Carboniferous earliest Permian (302 280 Ma) palaeogeography. Uplands are: I, ancestral Great Western Plateau; II, Central; III, Central Eastern; IV, ancestral Eastern Australian foreswell; V, ancestral Kimberley Plateau; XX, Ross. Basins are 1, Bonaparte; 2, Galilee; 3, Cooper; 4, Pedirka; 5, Arckaringa; 6, Canning; 7, Officer; 8, Carnarvon Perth; 9, Collie; 11, Oaklands; 12, Tasmania; 13, Sydney; 14, Gunnedah; 15, Bowen. Bullseyes indicate glacigenic sediment in mines. Coarse pattern, ice flow; solid blue, outcrops of glacigenic sediment; coarse blue stipple, subsurface glacigenic sediment; fine stipple, inferred glacigenic sediment. Red lines (west) indicate structure formed during Extension I. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

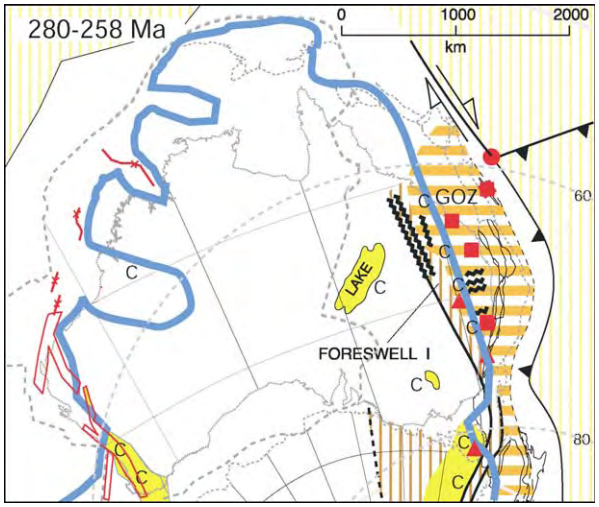


Figure 12 Permian (280 258 Ma) palaeogeography. C, coal; GOZ, Gogango Overfolded Zone; red squares, granite; red triangles, tuff. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

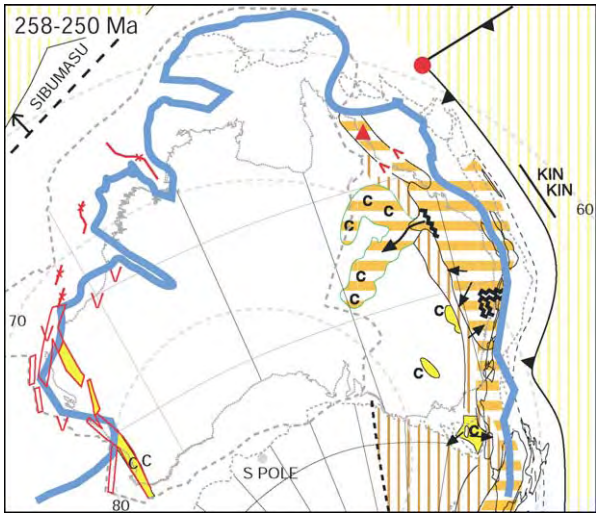


Figure 13 Late Permian (258 250 Ma) palaeogeography. Red triangle, tuff. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

the north-east was intruded by granite and deformed behind the trench. The Kanmantoo Marginal Basin started to close by north-eastward-directed subduction beneath an overlying volcanic arc (Figure 2iii). In Antarctica, an accretionary wedge continued to grow above the westward-subducting limb, and related Ross granites and marine volcanics were erupted.

Late Cambrian (510–490 Ma)

At the tectonothermal climax of Gondwanaland, granites stretched from the Antarctic Peninsula through the Transantarctic Mountains to Tasmania (Figure 5), and the Delamerian granites were generated during closure of the marginal basin to form the Kanmantoo Fold Belt (Figure 2iv). The trench continued past north Queensland along a margin with a wide accretionary prism that was metamorphosed and intruded by plutons at 500 Ma.

In the west, uplift along the Darling Fault and within the Prydz-Leeuwin Belt shed fans of Tumblagooda sand into the nascent Perth–Carnarvon Basin. In the north-east, the King Leopold and Halls Creek Fold Belts were terminally metamorphosed and thrusted, and the enclosed block was uplifted and stripped.

Early Ordovician (490–458 Ma)

The Delamerian–Ross Fold Belt, the zone of deformation with cooling granite (black dots), was intensely denuded in a tropical setting in which labile minerals and rock fragments were weathered away to provide a copious amount of quartzose sediment downslope to

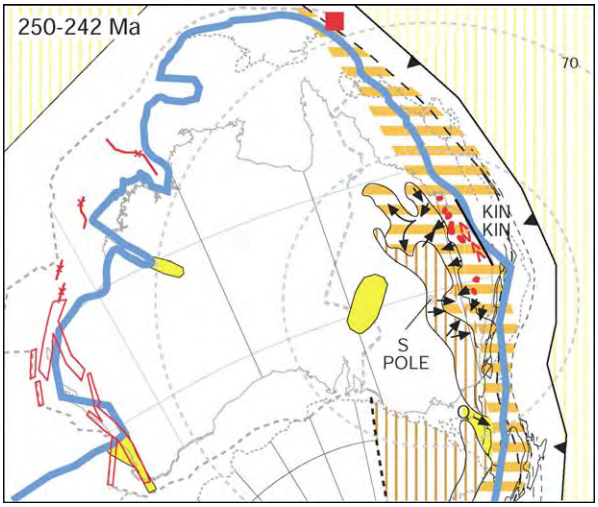


Figure 14 Early Triassic (250 242 Ma) palaeogeography. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

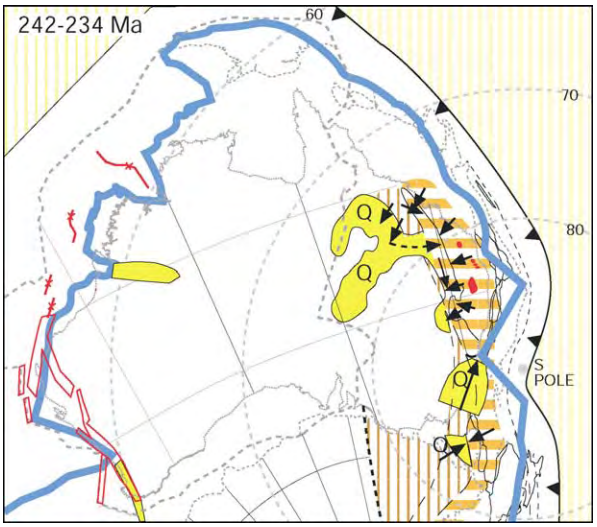


Figure 15 Early Middle Triassic (242 234 Ma) palaeogeography. Red, rejuvenated and initiated faults and depositional axes; Q, quartzose sandstone. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

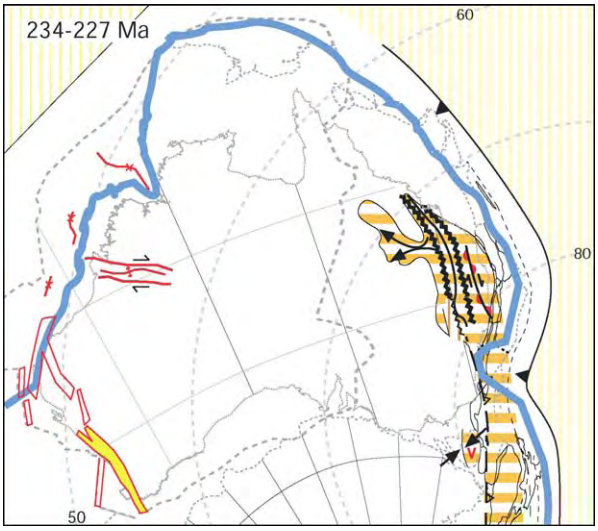


Figure 16 Mid Triassic (234 227 Ma) palaeogeography. Legend as in **Figure 3**. In the northern Canning Basin, an anticlinal axis (red line crossed by double headed arrow) formed between transcurrent faults (red lines) with black arrows. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

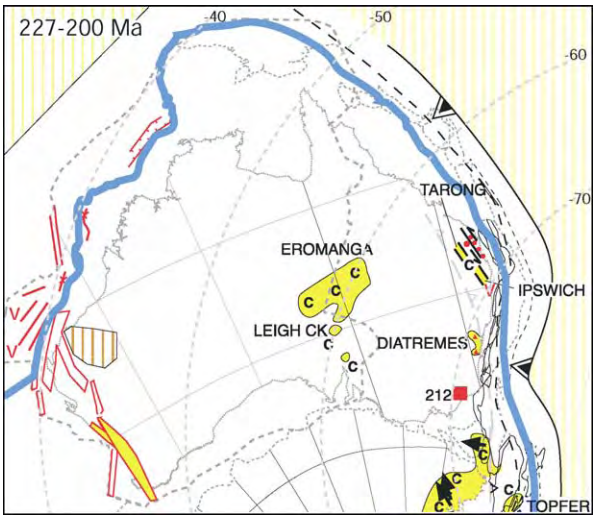


Figure 17 Late Triassic (227 200 Ma) palaeogeography. The double headed barbs on the trench indicate the change from shallowly dipping to steeply dipping subduction. Legend as in **Figure 3**. The red lines in the west represent faults. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Australia (Figure 6). Westward-directed subduction of the Palaeo-Pacific Plate beneath Eastern Australia generated a volcanic arc that stretched 3000 km northwards in front of a marginal basin (Figure 2v). Quartzose turbidites were shed eastwards (transversely) from the Delamerian upland, and volcanogenic sediment was shed from the volcanic arc. Turbidites were shed northwards (longitudinally) from Antarctica, as shown by 700–500 Ma zircons, along the fore-arc basin and abyssal plain in vast submarine fans, comparable to the modern Bengal fan. Behind the shelf, the Larapintine Sea crossed the interior towards the western sea, which intermittently advanced across the Tumblagooda sand.

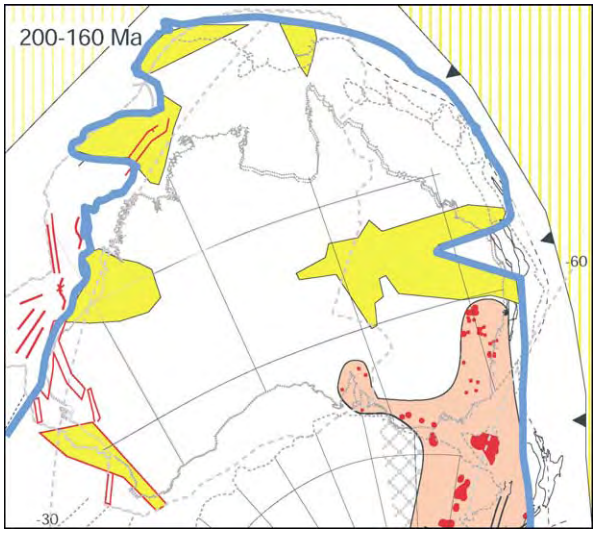


Figure 18 Early and Middle Jurassic (200 160 Ma) palaeo geography. The deep red shading represents basalt or dolerite within the large igneous field (pink). Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

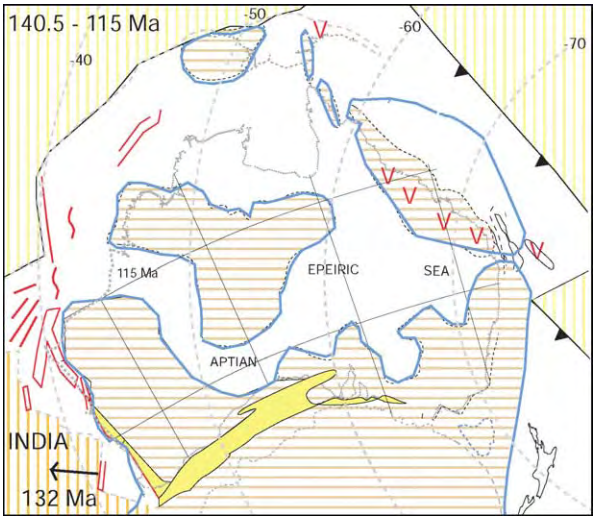


Figure 20 Neocomian to Aptian (140.5 115 Ma) palaeogeog raphy. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

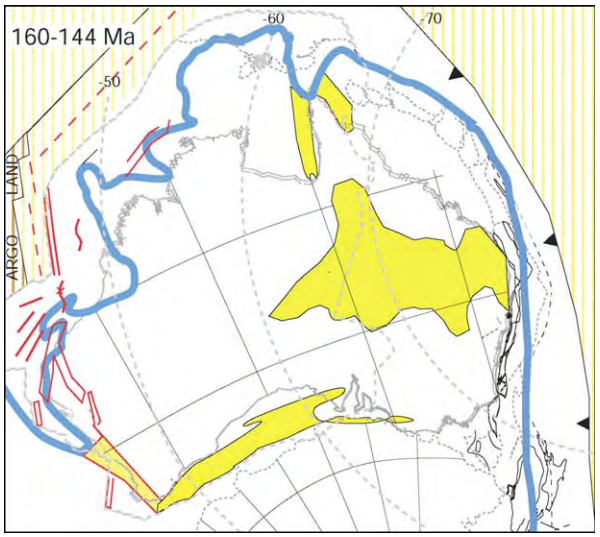


Figure 19 Late Jurassic (160 144 Ma) palaeogeography. Legend as in **Figure 3**. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

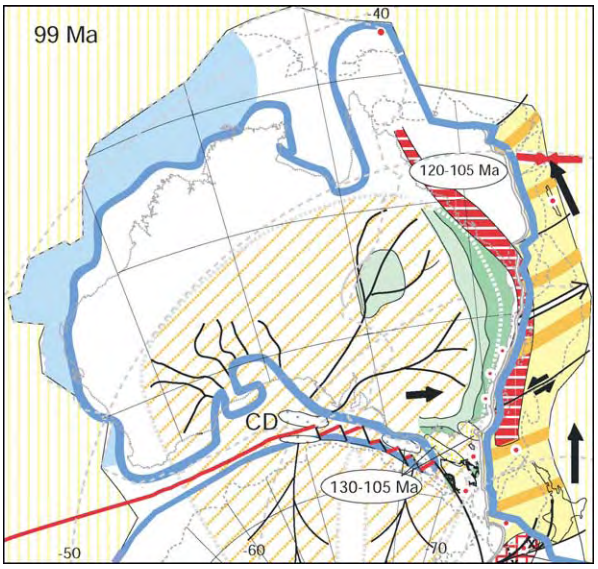


Figure 21 Cenomanian (99 93 Ma) palaeogeography. Red arrows, previous velocities; black arrows, prospective velocities (which are unchanged in Australia but swerve through 70° in the Pacific Plate); double line, detachment fault; yellow and tan shading, lower plate margin (Lord Howe Rise), white dotted line, division of drainage; CD, Ceduna depocentre; green, eroding uplands; oblique orange lines, south western drainage into Australian Antarctic depression; red horizontal shading, Whit sunday Volcanic Province; red dots, 99 Ma alkaline volcanics; light blue, carbonate deposition. Legend as in **Figure 3**. Repro duced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Early Silurian (443–425 Ma)

Closure of the marginal basins in eastern Australia generated the Benambran and related highlands (Figure 7 and Figure 2vi). The Rodingan uplift in the centre shed the Mereenie Sandstone. In the Canning Basin, the Sahara Formation was deposited in a

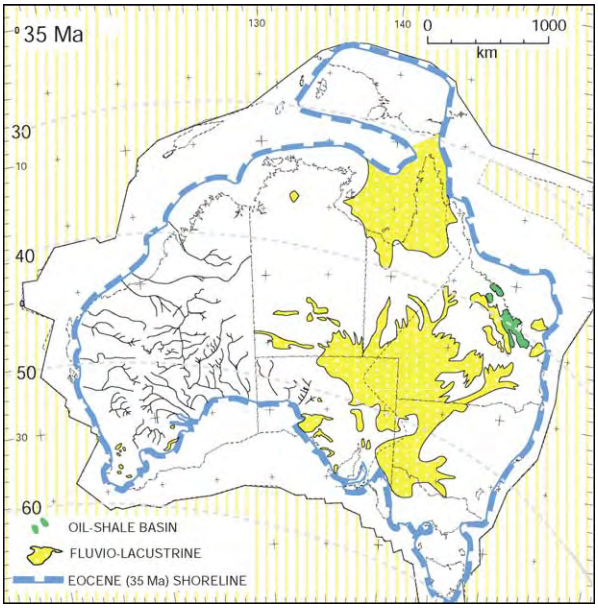


Figure 22 Eocene (35 Ma) palaeogeography. Legend as in Figure 3. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gond wanaland*. Sydney: GEMOC Press.

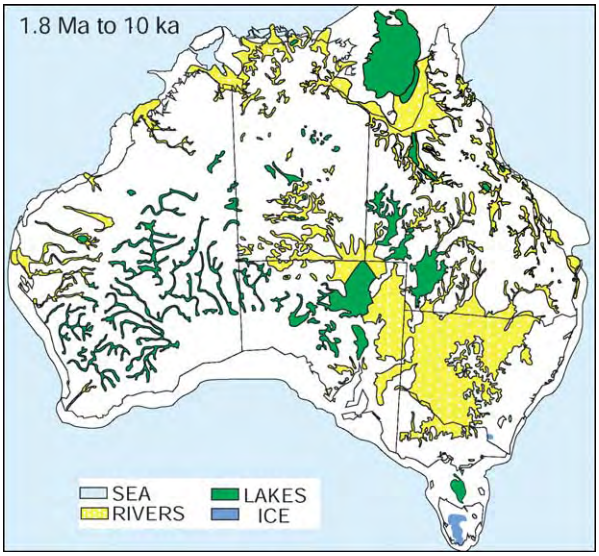


Figure 24 Pleistocene (1.8 Ma to 10 ka) palaeogeography. Re produced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

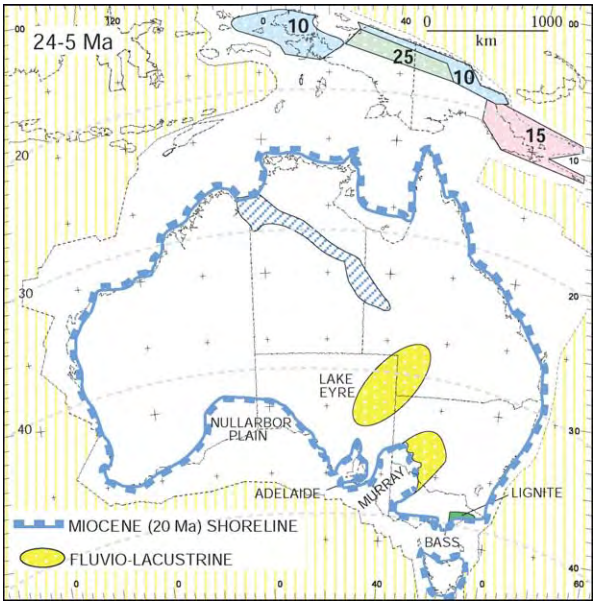


Figure 23 Miocene (24 5 Ma) palaeogeography. Numerals in dicate the times (Ma) at which the terranes docked. The blue oblique shading represents marine to nonmarine sediment. Legend as in Figure 3. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

restricted sea. In the Carnarvon Basin, the Ajana Formation succeeded the Tumblagooda Sandstone. In the Petrel Sub-Basin, salt was probably deposited during the Silurian and Devonian. Western Tasmania was covered by a shallow sea, while the Melbourne–Mathinna terrane slid past on its way to docking at 400 Ma.

Early Devonian (418–394 Ma)

The New Zealand terranes amalgamated at 415 Ma, and north-east and West Tasmania amalgamated at 400 Ma (Figure 8). In south-east Australia, relocation of the trench was accompanied by wide back-arc extension and emplacement of granite (Figure 2vii). The Adavale Basin was initiated with flows of andesite. Farther north, inner shallow-marine and outer deep-marine deposits were flanked by granitic plutons. A group of nonmarine deposits (Cravens Peak, Dulcie, Duerdin, and Mereenie) was deposited in the centre, salt in the Petrel Sub-Basin, aeolian and playa deposits of the Tandalgoo Formation in the Canning Basin, and the Kopke redbeds of the Carnarvon Basin.

Middle and Late Devonian (394–362 Ma)

Granodiorite was erupted in North Victoria Land during continuing westward-directed subduction (Figure 9). Following the contractional Tabberabberan Orogeny, plutonic rocks were emplaced and volcanic rocks were erupted in a broad zone in Tasmania, Victoria, and New South Wales, which included the Comerong–Yalwal–Eden volcanic rift. Fluvial sediment was deposited in the Barka, Lambie,

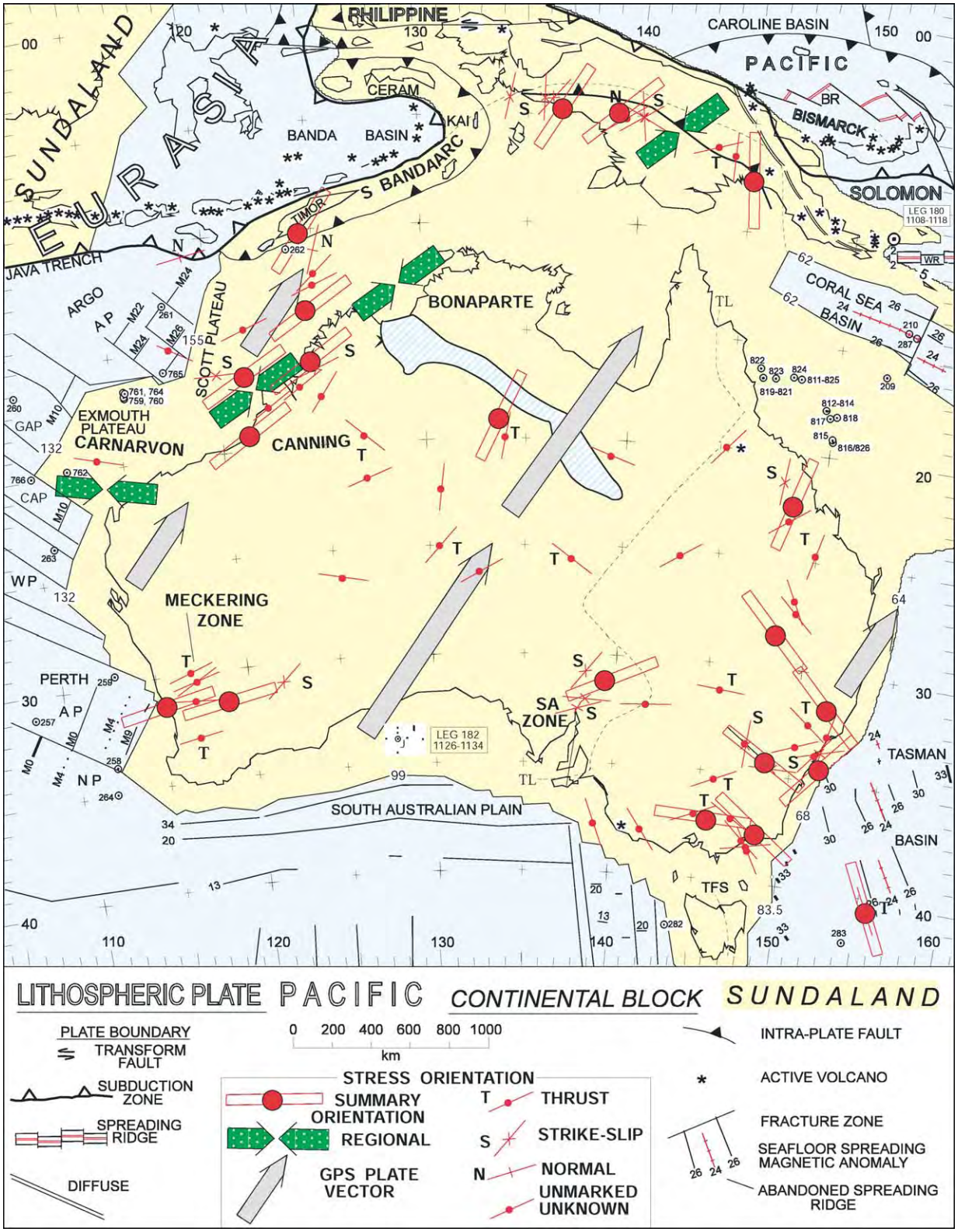


Figure 25 Tectonic map of the continental lithosphere of Australia New Guinea (yellow) and Sundaland (part of the Eurasian Plate; yellow) and the surrounding oceanic lithosphere (blue), including the Pacific Plate and the Solomon sub Plate to the north east. The ocean is marked by Deep Sea Drilling Project and Ocean Drilling Program sites (bullseyes) and selected seafloor spreading magnetic anomalies. The age of the continent ocean boundary is given in Ma. The oblique blue band across northern Australia signifies Miocene or younger sediment. Also shown are the orientation of maximum horizontal stress and the plate vector. AP, Abyssal Plain;

Continued

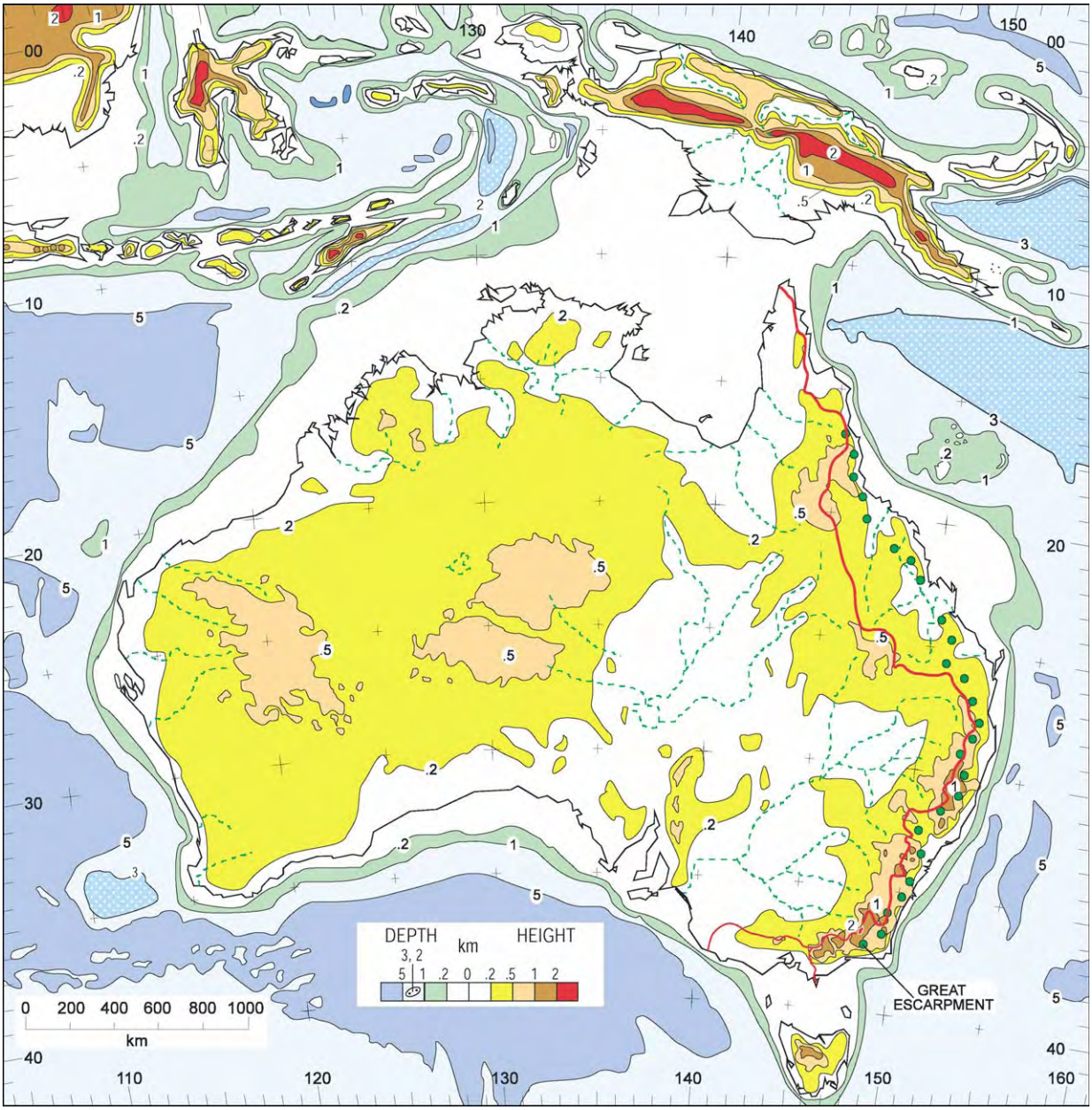


Figure 26 Morphology of Australia and the neighbouring land, sea, and ocean. Shown in eastern Australia are the Great Escarpment (green dots) and the drainage divide (red line). Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

and Hervey Basins behind the Baldwin arc. The fore-arc basin and accretionary prism (Figure 2viii) extended north as the Connors–Auburn arc.

In central Australia, the uplifted Arunta Block shed alluvial fans to the south, which were themselves folded and eroded. Fans in the Munyarai Trough were likewise shed from the overthrusting Musgrave Block. The eastward flow from central Australia through the Adavale Basin swung northwards (parallel to the Connors–Auburn arc) in the Drummond Basin and debouched in the Broken River Embayment, which was deformed in the latest Devonian.

BR, Bismarck Rift; C, Cuvier; G, Gascoyne; J, Jerboa 1; NP, Naturaliste Plateau; SA, South Australia; TL, Tasman Line; TFS, Tamar Fracture System; WP, Wallaby Plateau; WR, Woodlark Rift. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

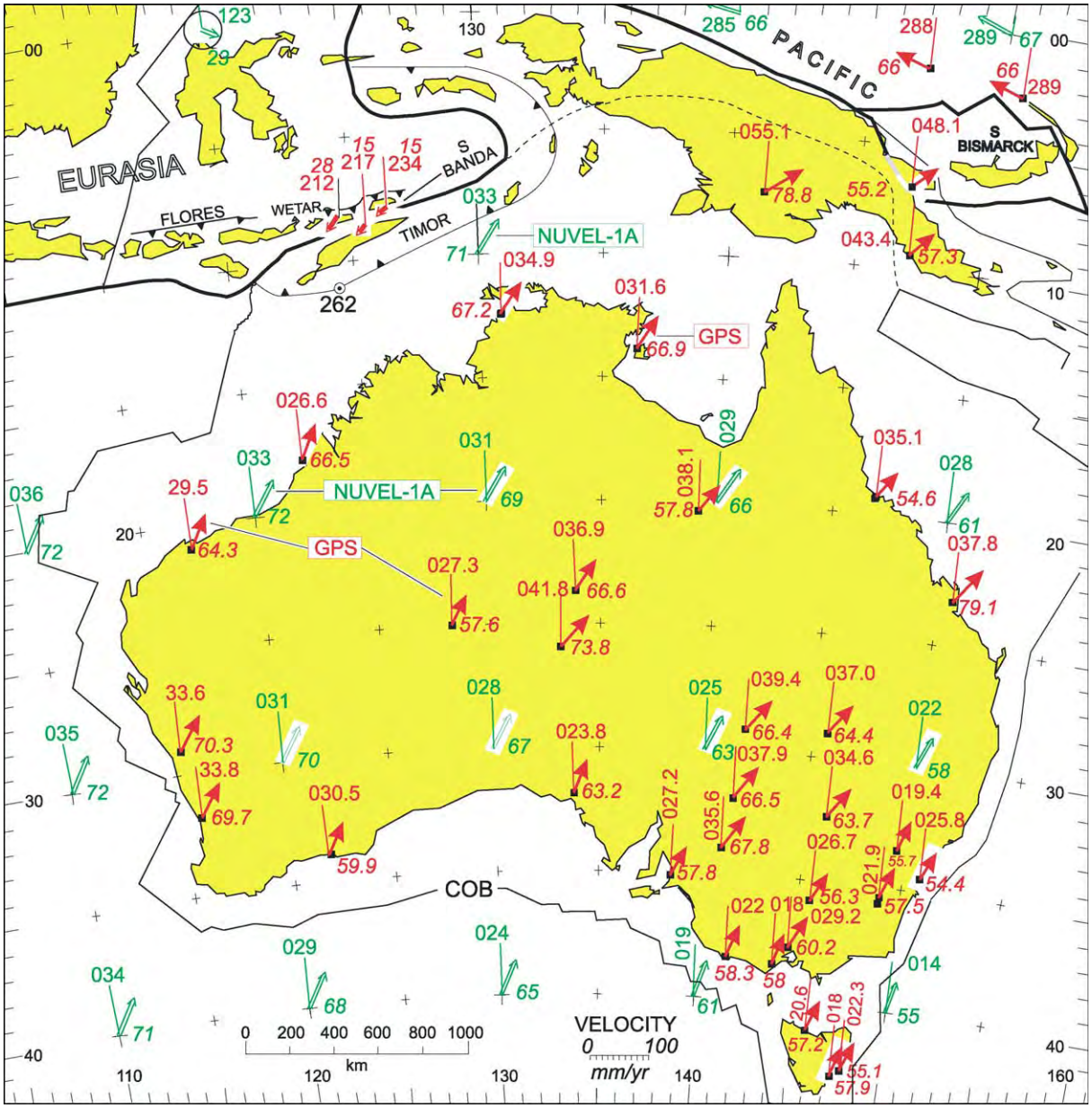


Figure 27 Velocities of sites on the Australian Plate and adjacent Pacific and Eurasian plates. Red arrows indicate velocities measured by GPS; double shafted green arrows indicate velocities calculated from NUVEL 1A at selected 10° crosses. For each arrow, the fine line is the north azimuth, the three digit numeral the bearing, and the italicized numeral the velocity in mm year ^{1}. The thrust line of Timor passes through DSDP (Deep Sea Drilling Project) site 262. Reproduced with permission from Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Table 1 The three stratitectonic regimes in the development of Phanerozoic Australia

Regime	*Age (Ma)*	*Platform stratigraphy*	*Tectonics*	*Latitude*
Potoroo	99 0	Quartz, carbonate, and evaporite	Post Pangaean	High to tropical
Innamincka	320 99	Glacigenic and coal measures	Pangaean	Polar
Uluru	544 320	Quartz and carbonate	Gondwanaland	Tropical

The North China terrane broke off the Palaeo-Tethyan margin. In the north, the Goulburn Graben was initiated; the Bonaparte Basin was rejuvenated and Frasnian alluvial fans were deposited at the foot of a fault and succeeded by a Famennian reef complex. In the Ord Basin, sand was deposited southwards from the same fault block. In the northern Canning Basin, gravel fans from the Kimberley Block mingled with a reef complex on a platform in front of a deep axis; to the south, the paralic Tandalgoo Formation was deposited in an arm of the sea. In the Carnarvon Basin, shelf limestone and inshore quartz sand give way in Pendock-1 to reef carbonate. Reworked Late Devonian conodonts in the Perth Basin suggest that the shoreline extended southwards.

Carboniferous (350–300 Ma)

South-eastern Australia was subjected to an Early Carboniferous (350–330 Ma) east–west contraction, followed by Mid-Carboniferous emplacement of granites (Figure 2x) and finally north–south contraction in megakinks. Similar events took place in North Queensland. In between, nonmarine deposition in the Drummond Basin ended at 330 Ma when gentle folding began. Central Australia was terminally folded and thrusted in the Mid-Carboniferous (320 Ma) Alice Springs Orogeny, which was echoed by transpression in the Canning Basin, by folding and faulting in the Carnarvon Basin, and by overthrusting in the Officer Basin. These movements, represented by the Mid-Carboniferous lacuna, mark the termination of the Uluru regime (when the Uluru Arkose was upended) and the inception of the Innamincka regime.

This rapid uplift and concomitant downwearing must have produced copious amounts of sediment, yet the depositional record over the Australian platform, except the eastern and north-western margins, is blank. Where could the sediment have gone? The paradox is resolved by postulating that uplift combined with rapid polar movement triggered not only an alpine glaciation in the east but also a continent-wide dry-based ice-sheet (Figure 10), which carried away sediment shed from the nunataks of the central uplifts and from the eastern cordillera and deposited nonmarine glacial sediment in the Werrie Trough. Only during its retreat in the earliest Permian did the ice release its load of sediment.

The Mid-Carboniferous marked a turning point in Australia, from the Uluru regime, with warm-water low-latitude carbonate facies, to the Innamincka regime, with cold-water high-latitude detrital facies.

Latest Carboniferous–Earliest Permian (302–280 Ma)

The Mid-Carboniferous lacuna ended with the latest Carboniferous (306 Ma) onset of Pangaean Extension I, which was accompanied by the eruption of S-type granites and felsic volcanics in north-eastern Australia and followed at 302 Ma by deposition of glacigenic sediment in the space provided by the extension (Figure 11). Eastern Australia was subjected to continuing dextral transtension, which produced an orocline, related pull-apart basins, and widespread magmatism (Figure 2xii). At polar latitudes ice flowed into depressions between uplands, and on the periphery into marine basins.

The west was extended into grabens and synclines. Ice flowed northwards over and around the ancestral Great Western Plateau and across the Canning Basin. Farther north, the ice possibly flowed through the Bonaparte Basin and through the restored position of the Tangchong–Malay–Sumatra Block in Sibumasu to the shore of Palaeo-Tethys. In central and eastern Australia, ice flowed from the central and central-eastern uplands into the Pedirka and Arckaringa basins, and from the central-eastern upland eastwards into the Cooper Basin, which also received ice from the ancestral eastern Australian foreswell to the east. From the Ross upland in Antarctica, ice flowed northwards into areas that were later, at 288 Ma, covered by an ephemeral shallow sea that followed the melting of the ice.

Early Permian (280–258 Ma)

By 265 Ma, the magmatic arc, marked by granite and tuff, had migrated northwards opposite central Queensland (Figure 12), and the following tracts, from craton to margin, can be identified: first, an epicontinental basin with nonmarine sediment; second, foreswell I (vertical bands), which shed quartzose sediment into, third, the initial foreland basin of the Bowen–Gunnedah–Sydney Basin, which is marked by coal alternating with shallow-marine deposits; and, fourth, an orogen and magmatic arc that were intensely deformed in the Gogango Overfolded Zone, which shed sediment back into the foreland basin. In adjoining Antarctica, foreswell I separated coal measures on the craton from the foreland basin.

In the west, marine sediment alternated with nonmarine sediment including coal.

Late Permian (258–250 Ma)

The magmatic arc propagated north-east to a position opposite 250 Ma tuff (Figure 13). Coal measures

reached a climax. The New England Fold Belt and adjacent basins were deformed (Figure 2xiii), and copious sediment flowed into the foreland basin and the distal Galilee Basin. The last ice melted. The Kin Kin terrane approached the trench.

In the west, Sibumasu broke off, the shoreline regressed, and rhyolitic to undersaturated volcanics were erupted.

Early Triassic (250–242 Ma)

The arc advanced to New Guinea, and plutons saturated the suture zone of the amalgamated Kin Kin terrane (Figure 14). In the foreland basin, cratonic and orogenic drainage joined in axial flow from a saddle just north of 30° S (present coordinates), and volcanolithic sediment poured westwards into the Galilee Basin. Tectonics was as before, but the surface environment was radically different. Coal measures and tuff were replaced by measures with redbeds devoid of both coal and tuff. The coal gap reflects the global extinction of peat-forming plants at the Permian–Triassic boundary, and tuff was eliminated by weathering in the warmer climate, even at polar latitudes.

In the west, shale was deposited behind a transgressive shoreline.

Early–Middle Triassic (242–234 Ma)

Volcanolithic sandstone was deposited in the foreland basin (Figure 15). Sediment flow in the Bowen Basin was now wholly southerly, and this was joined later in the stage by cratonic (quartzose) sediment flowing in an easterly direction from the Cooper–Galilee Basin. In the Sydney Basin the Hawkesbury Sandstone was derived from the Ross craton to the south-west, which also fed a quartz sandstone in Tasmania. The pole was nearby but ice is not indicated.

Mid-Triassic (234–227 Ma)

The complex of events called Gondwanides II included thrusting of the fold belt over the final deposits of the foreland basin and the spread of volcanolithic sediment to the west (Figure 16). The fold belt itself was intruded by granite along a transcurrent fault. At the same time, the northern Canning Basin was broadly folded during wrenching.

Late Triassic (227–200 Ma)

The Late Triassic saw the onset of Pangaean Extension II. In the east the change from contraction to extension reflected a change from shallowly to steeply dipping subduction (Figure 17). Maar diatremes were erupted in the Sydney Basin, and a 212 Ma syenite complex was emplaced in Victoria.

Deposition of coal resumed after a gap of 23 Ma. New intramontane basins included the Ipswich and Tarong coal basins; the Eromanga Basin with carbonaceous sediment subsided in the sump between the fold belt and the craton; the Leigh Creek and associated coal basins subsided on Proterozoic crust; and deposition of coal resumed in Tasmania and Antarctica.

In the west, the margins were rejuvenated by wrenching, downfaulting, and folding. Uplands in the Pilbara Block and Yilgarn Craton shed coarse detritus into the rifted Perth and Carnarvon Basins. Volcanics accompanied rifting in the Exmouth Plateau.

Early and Middle Jurassic (200–160 Ma)

The Permian–Triassic sedimentary successions were capped and intruded by a flood of 184–179 Ma tholeiitic basalts in southern Africa, Antarctica, and Tasmania, and by scattered volcanics in south-eastern Australia (Figure 18). The Eromanga Basin expanded across central-eastern Australia, with quartzose sediment derived from the craton and volcanolithic sediment from a presumed arc. In Tasmania, the dolerite rose into the Permo-Triassic sedimentary rocks of the Tasmania Basin as sills up to 400 m thick and dykes up to 1 km wide. Extrusive tholeiite is known near the southern tip only. In south-eastern Australia, the magma forms flows and shallow intrusions.

Thick sands accumulated in the western rifts, and sand sheets lay behind transgressive shorelines in the Canning and Bonaparte Basins.

Late Jurassic (160–144 Ma)

Argo Land split off north-western Australia at 156 Ma as a result of seafloor spreading in the Argo Abyssal Plain (Figure 19). The rifts along the western and southern margins continued to fill with sand and silt. The Eromanga Basin expanded northwards to New Guinea.

Neocomian to Aptian (140.5–115 Ma)

In a radical change, the shoreline advanced into the interior of Australia (Figure 20). The epeiric sea isolated large islands in the north-west and north and was confined by the rising rifted arch to the south and by the magmatic arc of the Pacific margin to the north-east. Volcanolithic sediment from the arc overran sediment from the craton.

The Southeast Indian Ocean opened at 132 Ma between India and Australia–Antarctica.

Cenomanian (99–93 Ma)

The change from the Innamincka regime to the Potoroo regime was marked by breakup along the southern margin, cessation of subduction in the east, and intense uplift of the eastern Highlands and epeirogenic uplift elsewhere, all occasioned by the 99 Ma change in the azimuth of subduction of the Pacific Plate from head-on (red arrows in Figure 21) to sideswipe (black arrows). A detachment fault along the continent–ocean boundary separated passive-margin mountains on the upper-plate margin, buoyed by thick underplating of mantle-derived melts, from the conjugate extremely attenuated lower-plate margin of the Lord Howe Rise. Drainage was divided near the present coast into a long south-western slope feeding the Ceduna depocentre, which accumulated 8 km of Late Cretaceous sediment, and a short eastern slope feeding the low-lying eastern borderland.

The emergence of the land, in particular the rising (and eroding) of the eastern highlands, forced a major Cenomanian (99 Ma) regression at a time of marine flooding elsewhere. The inception of south-western drainage into the Australian–Antarctic depression coincided with the inception of the juvenile ocean that grew out of the rifted arch and was preceded by 130–105 Ma deposition of volcanogenic sediment in the Otway, Bass, and Gippsland basins. In the east, the 120–105 Ma Whitsunday Volcanic Province was succeeded by *ca*. 99 Ma alkaline volcanics, and the previous volcanogenic sediment from the arc and in the south-eastern rift basins was succeeded by quartzose sediment. In the north and west, the shoreline regressed to and beyond its present position, and carbonate replaced detrital sediment on the margin of the widening Indian Ocean.

Eocene (35 Ma)

Following the opening of the Tasman Sea and Coral Sea, the Potoroo regime can be shown on a modern geographical base except for the higher palaeolatitudes (Figure 22). The sea transgressed the western half of the southern margin, and an ancestral Arafura Sea lay between northern Australia and New Guinea. Relaxation of the platform saw sediment accumulate in the central-eastern lowlands and in grabens in central Australia and south-eastern Queensland, including deposits of oil shale and, in isolated pockets, fluviolacustrine gravel, sand, silt, clay, and lignite, which were overlain by basalt in the eastern Highlands, too small to show in Figure 22. In the west, the uplift was etched by rivers except in an area of lakes in the south-west.

Miocene (24–5 Ma)

At 20 Ma, another transgression in the south crossed the Nullarbor Plain, the Adelaide region, the Murray Basin, and Bass Strait, and thick lignite accumulated in the onshore Gippsland Basin (Figure 23). Sheets of sand accumulated in the Lake Eyre drainage system and in the Murray Basin.

In New Guinea, the Sepik terrane had docked by 25 Ma and thrusted south-westwards to form a foreland basin; it was joined at 15 Ma by a composite terrane in the south-east, and at 10 Ma by terranes in the north and north-west. Miocene and younger marine to nonmarine sediment was deposited behind a foreswell across northern Australia. By the end of the Miocene the geography of the region was as it is today except that the palaeolatitude (at 16 Ma) was 10° farther south than today.

Pleistocene (1.8 Ma to 10 ka)

By the end of the Pleistocene, at 10 ka, sea-level was halfway between the level at the Last Glacial Maximum and its present level (Figure 24). Lakes occupied the Gulf of Carpentaria and the Bass Strait, and, as today, dry lakes covered much of interior Australia. The central-eastern lowlands in New South Wales were covered by sheets of fluvial sediment. Glacial deposits were confined to Tasmania and the Mount Kosciusko area. Aeolian dunes covered much of the arid interior, including the relict drainage of the south-west.

Present Tectonics and Morphology

The continental lithosphere of Australia–New Guinea is surrounded by the oceanic lithosphere of divergent oceans on its western, southern, and eastern margins and of convergent oceans on its northern margin from Timor to the Papuan Peninsula (Figure 25). The convergent boundary is marked by high relief, 3 km above sea-level in Timor and 5 km in New Guinea (Figure 26), intense Earth movement and seismicity, and scattered volcanicity. South of this boundary, as befits its intraplate position, Australia is low-lying, with a relief of 2.2 km in the south-east and little more than 1 km elsewhere, relatively quiescent, except for diffuse seismicity in the north-west sector, in zones about Meckering in the south-west, in South Australia, and in the south-east, and has only two areas of volcanicity, one in north-east Queensland, the other on the southern margin at 141° E.

The submarine features of the divergent margins have a certain bilateral symmetry about the meridian halfway across Australia: complexes of marginal plateaus and abyssal plains are backed by a broad shelf in the north, and narrow margins expand into long appendages (Naturaliste Plateau and Tasmania–South Tasman Rise) in the south. This pattern extends to the arrangement of seafloor-spreading magnetic anomalies in that the azimuth of the anomalies off the west coast is reflected in that off the east coast, while the set to the south is crossed by the line of symmetry. The age of onset of spreading follows another pattern, with anticlockwise propagation from 156 Ma in the north-west through 132 Ma in the west, and 99 Ma in the south, and as back-arc spreading at 83.5 Ma, 68 Ma, and 64 Ma in the south-east, 62 Ma in the north-east, and 5 Ma off eastern New Guinea.

Present Motion

Figure 27 shows the velocities (azimuth and magnitude) from a few years of Global Positioning System (GPS) measurements and from the NUVEL-1A global-plate-motion model averaged over several million years. For example, the GPS estimate in western Victoria of 58.3 mm year $^{-1}$ at 022° approximates the NUVEL-1A estimate of 61 mm year $^{-1}$ at 019°, and that near Darwin of 67.2 mm year $^{-1}$ at 034.9° matches the NUVEL-1A estimate of 71 mm year $^{-1}$ at 033°. Most of this motion comes from seafloor spreading in the Southeast Indian Ocean at a full rate of 60 mm year $^{-1}$ from an almost stationary Antarctica.

The stations on mainland New Guinea, including one near the plate boundary, record an azimuth of *ca.* 050°, which locates them on the south-western (Australian) side of the plate boundary. On the other side, the Pacific Plate is moving at a velocity of 66 mm year $^{-1}$ at 289°, almost at a right angle to Australia's motion, confirming the idea expressed by Alfred Wegener that the Pacific's westerly motion has planed off the obtruding anvil of New Guinea.

According to NUVEL-1A, the Eurasian Plate is moving slowly south-eastwards. GPS measurements on either side of the Flores and Wetar thrusts show that the southern Banda arc is virtually accreted to the Australian plate margin and that most of the convergence seems to be accommodated by left-lateral slip at the plate boundary, suggesting that the Timor Trough is now inactive as a thrust.

See Also

Antarctic. **Australia:** Proterozoic; Tasman Orogenic Belt. **Gondwanaland and Gondwana**. **New Zealand**. **Oceania (Including Fiji, PNG and Solomons)**. **Palaeoclimates**. **Pangaea**. **Plate Tectonics**. **Tectonics:** Mountain Building and Orogeny.

Further Reading

BMR Palaeogeographic Group (1990) *Evolution of a Continent*. Canberra: Bureau of Mineral Resources, Geology and Geophysics.

Veevers JJ (ed.) (1984) *Phanerozoic Earth History of Australia*. Oxford: Clarendon Press.

Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Veevers JJ (2001) *Atlas of Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Tasman Orogenic Belt

D R Gray, University of Melbourne, Melbourne, VIC, Australia
D A Foster, University of Florida, Gainesville, FL, USA

Introduction

The eastern part of Australia formed along the margin of Gondwana during the Palaeozoic due to the accretion of oceanic platform and basinal sequences (Figure 1). A region of eastern Australia over 1000 km wide now consists of three distinct orogenic belts (the Delamerian, Lachlan–Thomson, and New England orogens), which are collectively referred to as the Tasman Orogenic Belt (Figures 1 and 2). This deformed and metamorphosed tract of rocks was once part of a major orogenic system that extended some 20 000 km along the Gondwanan margin, incorporating parts of the Andes in South America (*see* **Andes**), the Cape Fold Belt in southern Africa, and the Ross Orogen in Antarctica (*see* **Antarctic**) (Figure 1). In eastern Australia accretion occurred in a stepwise fashion, with an eastward younging from the Cambrian to the Triassic, reflected

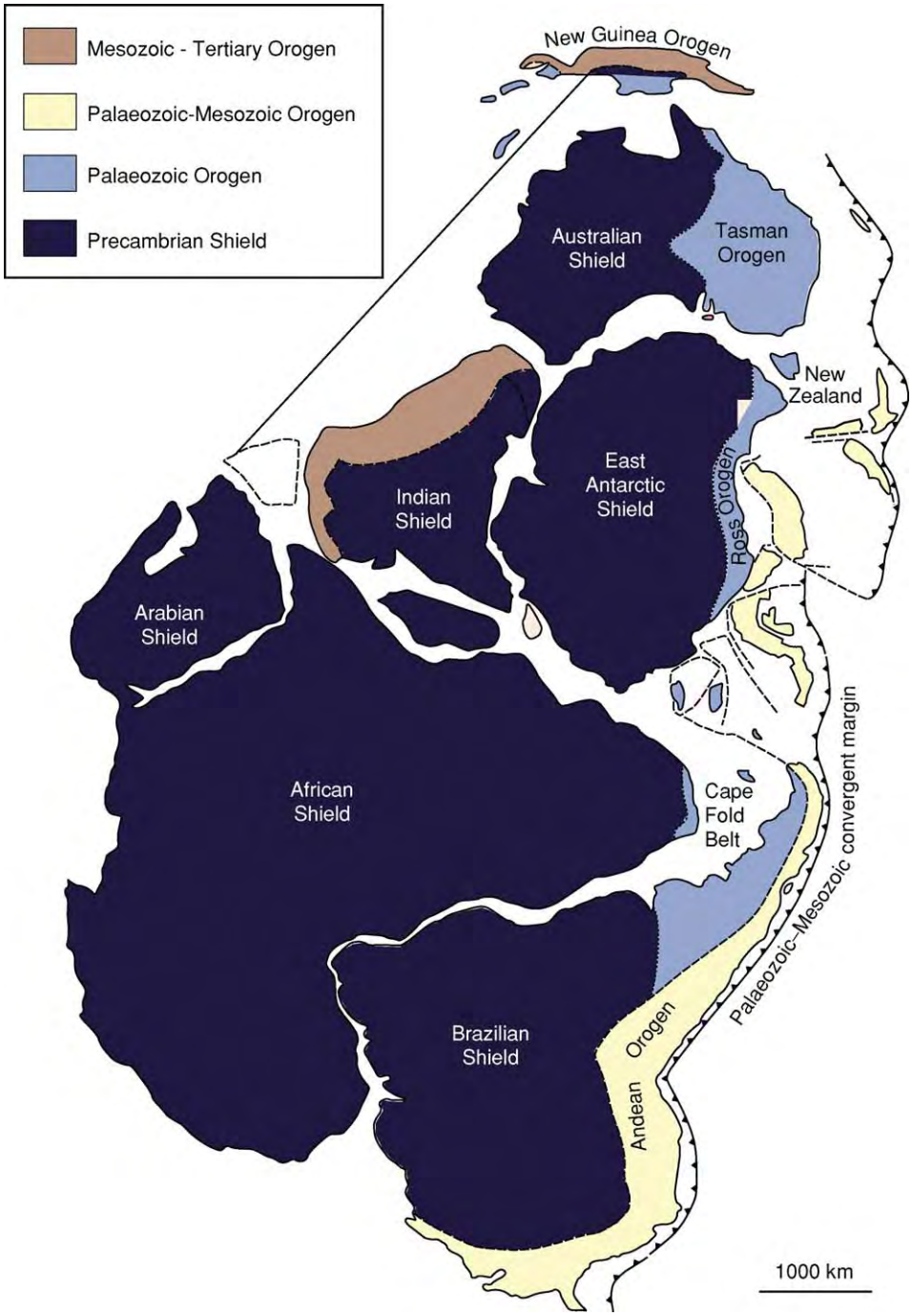


Figure 1 Simplified geological map of Gondwana with modern day continental outlines, showing regions of Precambrian shield and the younger orogenic belts. IGCP, Gondwana map base.

by peaks of deformation in the Early to Middle Cambrian, Late Ordovician–Silurian, and Permian–Triassic in the respective orogenic belts (Table 1).

Crustal growth occurred largely by the addition of turbidites (ocean-floor submarine-fan deposits), cherts (ocean-basin deposits), mid-ocean ridge basalts (oceanic crust), andesites (island arc), and granites. These younger, largely Palaeozoic, rocks occur to the east of the exposed Precambrian cratonic crystalline basement; the boundary between the two is known as the Tasman Line (Figure 2). They represent a significant addition (approximately 30% by surface area) to the Australian continent whilst it was part of Gondwana. Younger basinal sedimentary sequences

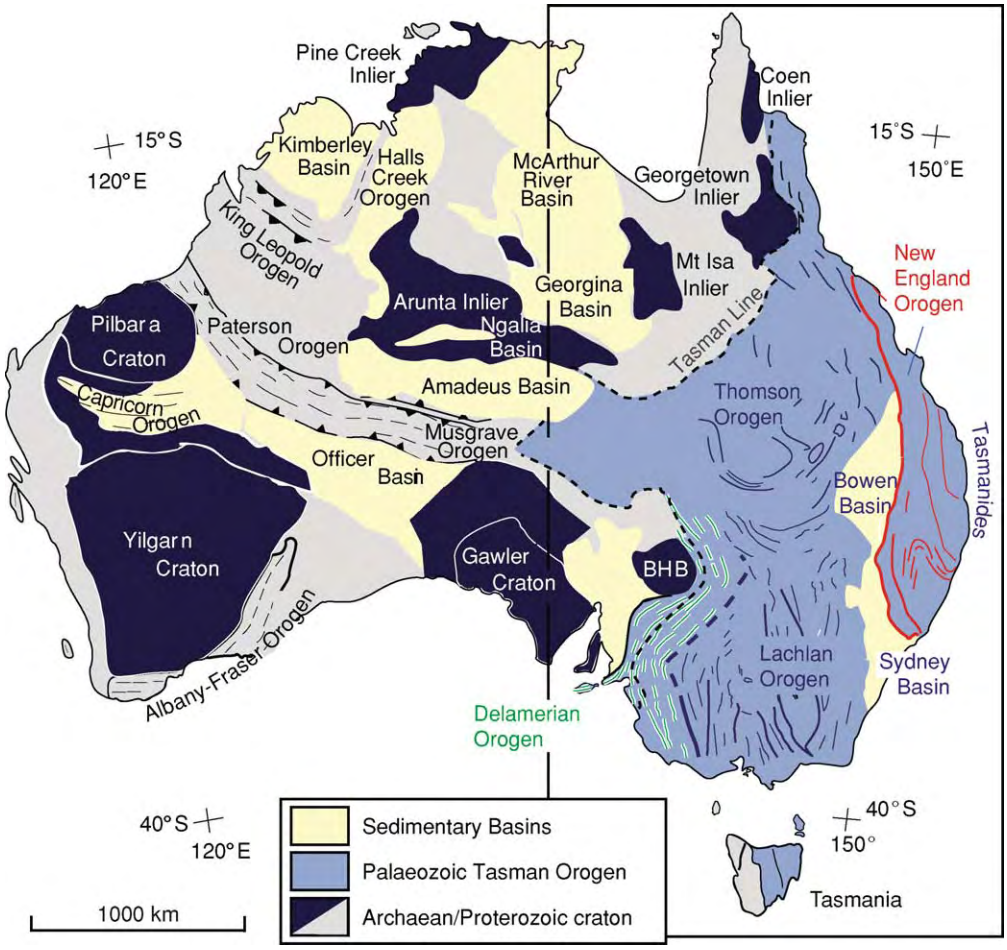


Figure 2 The major geological elements of Australia, showing the Tasman Orogen along the eastern margin. Location of the more detailed map area of **Figure 3** is shown.

(e.g. Great Artesian, Murray, and Sydney–Bowen Basins) now cover much of belt (**Figure 3**).

Accretion was accompanied by marked structural thickening and shortening (by about 75%) of Neoproterozoic platform and rift basinal sequences (Delamerian part), a Cambrian–Orodovician composite turbidite submarine fan overlying Cambrian oceanic crust (Lachlan part), and Carboniferous–Permian arc, fore-arc, and trench (subduction complex) and Permian foreland basin sequences (New England part) (**Figure 3**). For a large part of the history (Silurian–Early Carboniferous), deformation occurred by massive telescoping, with thrusting largely directed away from the cratonic core, together with components of strike-slip translation within a progressively deforming and prograding sediment fan along the Gondwanan margin. The margin was intruded by voluminous granitic plutons (accounting for up to 35% of the area), derived from subduction-generated mantle magmas and lower- to middle-crustal melting, between the Late Ordovician and the Triassic (**Figure 3**). The spatial and temporal variations in deformation, metamorphism, and magmatism across the Tasman Orogenic Zone show how subduction–accretion complexes are monotonous turbidite sequences evolve through time and eventually form crust of continental thickness and character.

Tasman Orogen Make-up

The Delamerian, Lachlan–Thomson, and New England Orogens, which form the composite Palaeozoic and Mesozoic Tasman Orogen (**Figure 2**), are distinguished by their lithofacies (**Figure 3**), differing tectonic settings, and timing of orogenesis and eventual consolidation with the Australian craton (**Table 1**). The Delamerian Orogen involved an inverted intracratonic rift, the Lachlan Orogen involved closure of a south-west-Pacific style marginal basin incorporating Bengal Fan-sized turbidite fans, and the New England Orogen involved a deformed arc–subduction complex belt.

Table 1 Orogens and subprovinces of the Tasman Orogenic Belt

Orogen	*Delamerian (West)*	*Western*	*Lachlan Central*	*Eastern*	*New England (East)*
Main lithofacies	Platform to deep water clastic margin sequence	Quartz rich turbidite sequence on oceanic crust	Quartz rich turbidite sequence on oceanic crust	Platform carbonates and clastics with rhyolites and dacitic tuffs	Volcanogenic clastic
Initial setting	Late Proterozoic rifted margin	Cambro Ordovician submarine fan in back arc basin	Ordovician submarine fan in back arc basin	Ordovician rifted arc	Late Palaeozoic fore arc arc trench
Oldest lithology	Basic volcanics	Cambrian basic volcanics (*ca.* 505 Ma)	Tremadocian chert	Ordovician andesitic volcanics (*ca.* 485 Ma)	Ordovician basic volcanics
Main deformation	Early to Middle Cambrian (*ca.* 520 500 Ma)	Late Ordovician Silurian (*ca.* 450 420 Ma)	Late Ordovician Early Devonian (*ca.* 450 410 Ma)	Silurian Early Carboniferous (*ca.* 430 360 Ma)	Late Permian Triassic (*ca.* 265 230 Ma)
Tectonic vergence	Westwards directed thrusting	Eastwards directed thrusting	Overall strike slip (*ca.* 410 400 Ma) with south westwards directed thrusting (*ca.* 440 420 Ma)	Eastwards directed thrusting	Westwards directed thrusting
Main plutonism	Cambrian (520 490 Ma)	Late Devonian (380 360 Ma)	Late Silurian (*ca.* 430 400 Ma)	Late Silurian Devonian (*ca.* 430 380 Ma)	Late Permian Triassic (*ca.* 270 210 Ma)

Delamerian Orogen

The Delamerian Orogen is the oldest and innermost part, immediately adjacent to the Precambrian cratonic core. The Delamerian Orogen was tectonically active from the Late Precambrian (*ca.* 650–600 Ma) to the Late Cambrian (*ca.* 500 Ma). It is now an arcuate craton-verging thrust belt with foreland-style folds and detachment-style thrusts. Duplexes in deformed Cambrian deep-water sandstones and mudstones (Kanmantoo Group) overlie less deformed and metamorphosed Neoproterozoic shelf sequences (Adelaidean). Metamorphic grade increases to the east, where polyphase deformation and amphibolite-grade metamorphism with local development of kyanite–sillimanite assemblages are spatially and temporally confined to the aureoles of synkinematic granites that are conformably aligned with the structural grain.

Lachlan Orogen

The Lachlan Orogen is the largest and the centrally located orogenic belt (Figure 2) with a marked similarity of sedimentary facies and structural style. It was active from the Ordovician and has a complex amalgamational and deformational history. The western and central parts of the Lachlan Orogen are dominated by a turbidite succession consisting of quartz-rich sandstones and black shales. These are laterally extensive over a present width of 800 km and have a current thickness exceeding 10 km. The eastern Lachlan Orogen consists of shoshonitic volcanics, volcaniclastic rocks, and limestone, as well as quartz-rich turbidites and extensive black shales in the easternmost part (Figure 3).

The western Lachlan Orogen consists of an eastwards-vergent thrust system with alternating zones of north-westwards- and northwards-trending structures. The central Lachlan Orogen is dominated by north-westwards-trending structures and consists of a south-westwards-vergent thrust belt linked to a fault-bounded metamorphic complex (Wagga-Omeo Metamorphic Belt). The eastern Lachlan Orogen is dominated by a northwards-trending structural grain and eastwards-directed thrusting, which caused inversion of extensional basins in the west. Tight-to-open chevron folds (accommodating between 50% and 70% shortening) cut by predominantly westwards-dipping high-angle reverse faults are part of different thrust systems within the Lachlan Orogen. The chevron folds are upright and gently plunging, but become

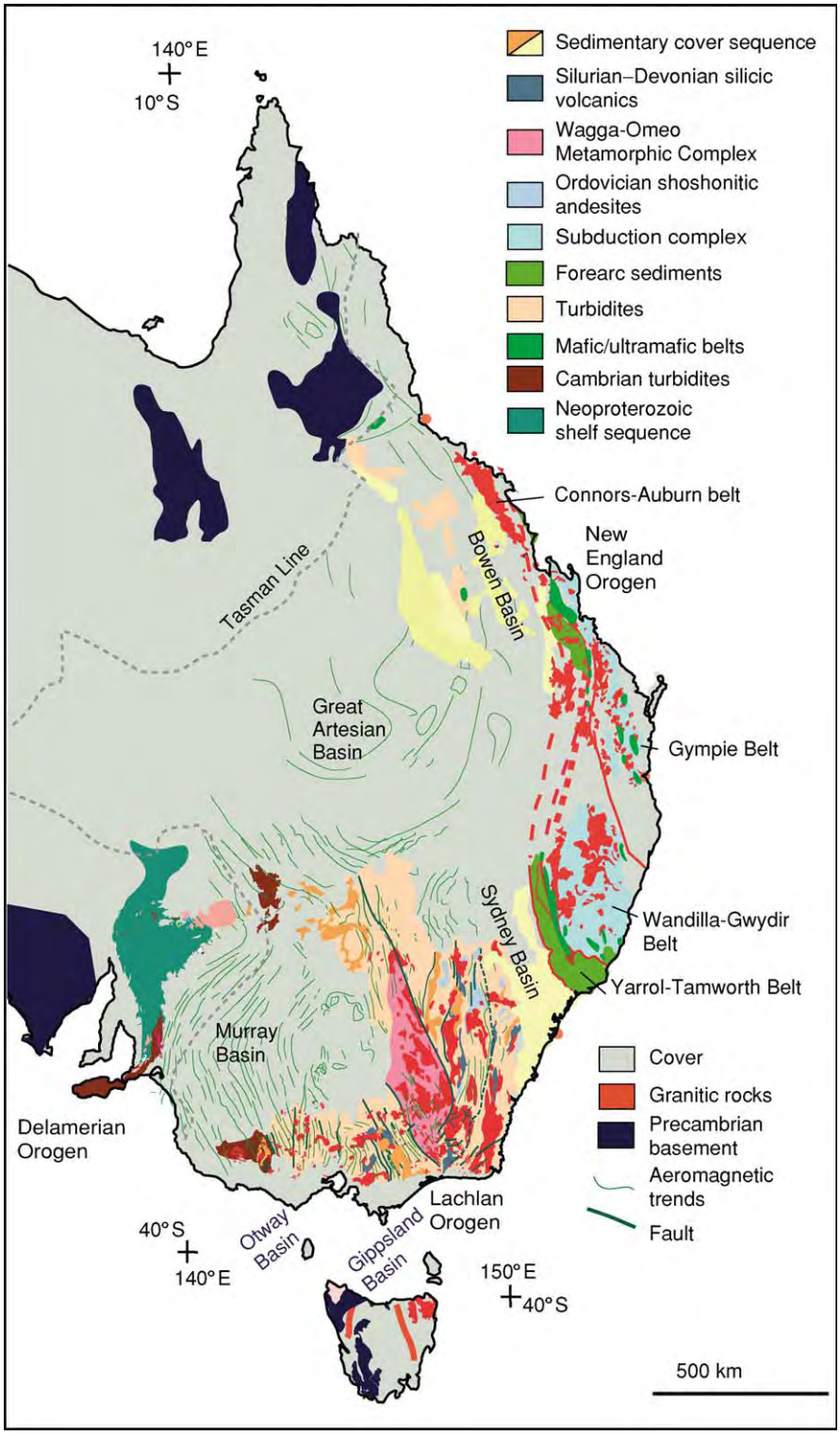


Figure 3 Geological map of the Tasman Orogenic Belt, showing the dominant lithologies, granite intrusives, major fault traces, and trend lines from regional aeromagnetic datasets.

inclined and polydeformed as they approach major faults. Metamorphism is greenschist facies or lower across the Lachlan Orogen, except in the fault-bounded Wagga-Omeo and several smaller (Cooma, Cambalong, Jerangle, and Kuark) metamorphic complexes, where high-temperature–low-pressure metamorphism is characterized by andalusite–sillimanite assemblages.

Thomson Orogen

The Thomson Orogen is meant to be coeval with, as well as the northern continuation of, the Lachlan Orogen (Figure 2). The orogen is hidden under a cover of younger sedimentary basins that make up the Great Artesian Basin, but geophysical aeromagnetic imagery shows that it has eastwards-trending structures that truncate the trends of the Lachlan Orogen. Drill samples and limited surface exposures have shown that it consists of low-metamorphic-grade turbiditic sandstone and mudstone lithologies similar to those of the Lachlan Orogen, and like the Lachlan Orogen it is intruded primarily by Silurian and Devonian granites.

New England Orogen

The New England Orogen is the youngest and most easterly part of the Tasman Orogen (Figure 2) and incorporates arc, fore-arc, and accretionary complexes. The New England Fold Belt was tectonically active from the Early Carboniferous to the mid-Triassic (a period of *ca.* 130 Ma). Westwards-directed Permian–Triassic thrusting caused interleaving and imbrication of the arc magmatic belt (Connors–Auburn Belt) and fore-arc (Yarrol–Tamworth Belt) and oceanic assemblages, including subduction complexes (Wandilla–Gwydir Belt) and ophiolite (Gympie Belt). Subduction-complex assemblages show a strong thrust-related fabric, polyphase deformation, and greenschist to amphibolite facies metamorphism. The Hunter–Bowen Orogeny (*ca.* 265–230 Ma) consolidated the terranes with Australia and resulted in the development of a Permo-Triassic foreland basin (the Sydney–Bowen Basin) (Figure 3).

Lithofacies

The Tasman Orogen is dominated by turbidites that were once part of large submarine fans overlying oceanic crust (Figure 3). Dismembered ophiolitic rocks of Neoproterozoic to Cambrian age and mafic to ultramafic affinities are preserved as slivers along major faults (Figure 4). The mafic rocks dominantly include tholeiitic pillow basalts and dolerites with some andesitic and boninitic volcanics, and are generally associated with hemipelagic black shales and cherts. These are interpreted as remnants of the oceanic crust that regionally underlies the Cambro-Ordovician turbidite fans. The ophiolitic slivers were incorporated into the turbidites as offscraped slices, as imbricated fault–duplex slices, and as blocks in mélange. In the eastern Lachlan Orogen Ordovician shoshonitic andesites are structurally interleaved with the Ordovician turbidites (Figure 3).

Inboard of the turbidites are Upper Proterozoic (Adelaidean) intracratonic rift sequences of marine to deltaic sandstones and shales, lagoonal evaporites, dolomites, and limestone, overridden by the Delamerian Orogen (Figure 3). These sediments were transgressed by Lower Cambrian shelf sediments that are transitional into deep-water sandstones and mudstones of the Kanmantoo Group. Outboard of the turbidites the New England Orogen consists of a collage of deformed and imbricated terranes of largely Middle to Upper Palaeozoic and Lower Mesozoic marine to terrestrial sedimentary and volcanic rocks, as well as strongly deformed flysch, argillite, chert, pillow basalts, ultramafics, and serpentinites.

Deformation

Based on unconformities, sedimentary facies changes, and, most recently, geochronology (Figure 5), deformation across the Tasman Orogenic Belt occurred in the Late Cambrian (Delamerian Orogen), Late Ordovician–Silurian and Early Devonian (Lachlan Orogen), and Late Permian–Middle Triassic (New England Orogen). The deformation was partitioned into regional-scale migrating thrust systems. Tectonic vergence, recorded largely by the dips of the major faults, is craton-directed for the Delamerian and New England Orogens, whereas the Lachlan Orogen shows mixed vergence, but with thrusting largely directed away from the craton (Figure 6).

The deformation style is chevron folding cut by high-angle reverse faults. Fault zones are characterized by higher than average strain and intense mica fabrics, transposition foliation, isoclinal folds, and polydeformation with overprinting crenulation cleavages.

Metamorphism

Metamorphic grade is generally low across the Tasman Orogen (Figure 7), with greenschist facies (epizonal) and subgreenschist facies (anchizonal) metamorphism of the turbidite sequences. Most of the turbidites are within the chlorite zone, with localized development of biotite in contact aureoles around granites. High-temperature–low-pressure metamorphism is indicated by localized migmatites and alkali feldspar–cordierite–andalusite–sillimanite gneisses in the Mount Lofty

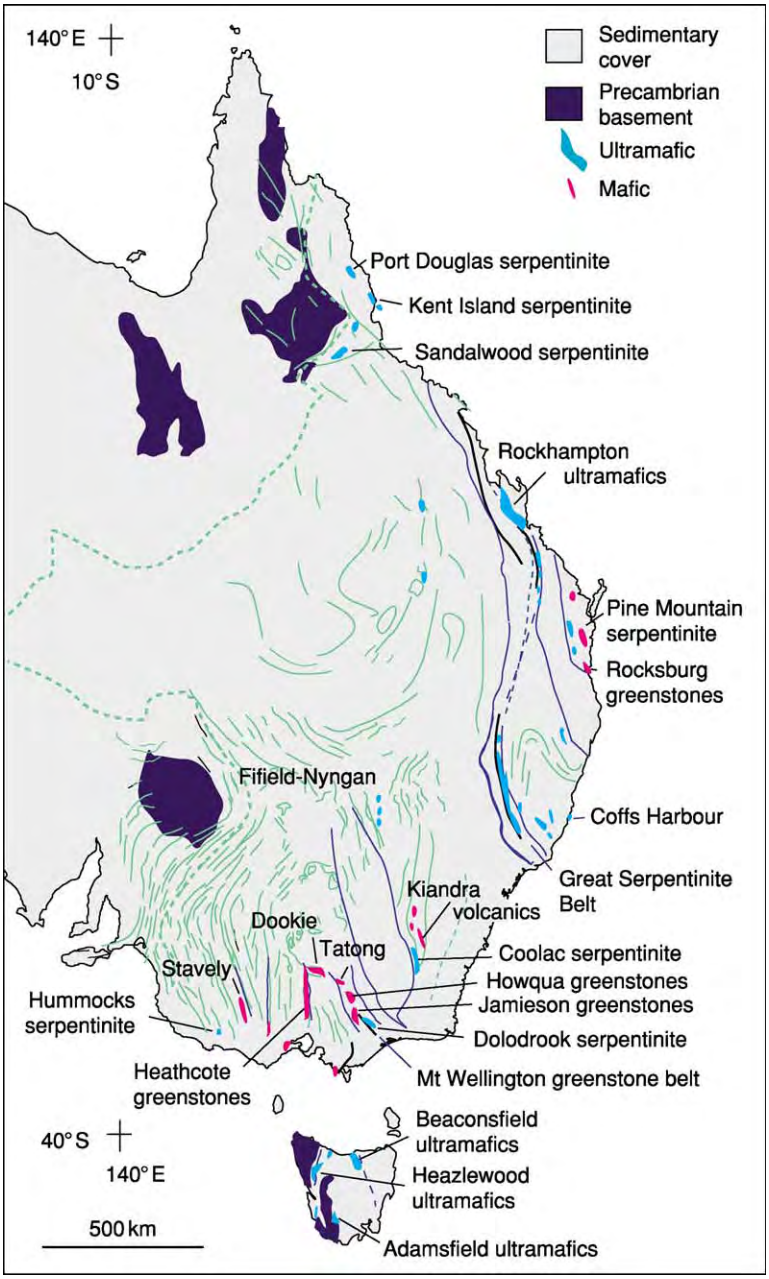


Figure 4 Geological map of the Tasman Orogenic Belt, showing the distribution of rocks with mafic and ultramafic affinities (dismembered ophiolites), major fault traces, and trend lines from regional aeromagnetic datasets.

Ranges, Glenelg Zone, and Moornambool Complex (Delamerian Orogen), Wagga-Omeo Complex and Cooma, Cambalong, and Kuark belts (Lachlan Orogen), and Wongawibinda Complex (New England Orogen) (Figure 7). Delamerian metamorphism is characterized by peak conditions of 650–700°C and 4–5 kbar in the Glenelg Zone, and 540–590°C and approximately 7–8 kbar in the Moornambool Complex. Peak metamorphic conditions in the Lachlan Orogen are 700°C and 3–4 kbar in the Wagga-Omeo Metamorphic Belt and Eastern Metamorphic complexes (e.g. Cooma Complex), but these belts are intimately associated with S-type granitic bodies.

Turbidites from the low-grade parts of the Lachlan Orogen show intermediate pressure series metamorphism, based on b_0 or x-ray diffraction lattice spacing measurements of phengitic micas, and are inconsistent with previous interpretations of widespread low-pressure 'regional-aureole' metamorphism. Moderately high-pressure metamorphism has been inferred from co-existing chlorite and actinolite in metabasites from one of the major fault zones. Furthermore,

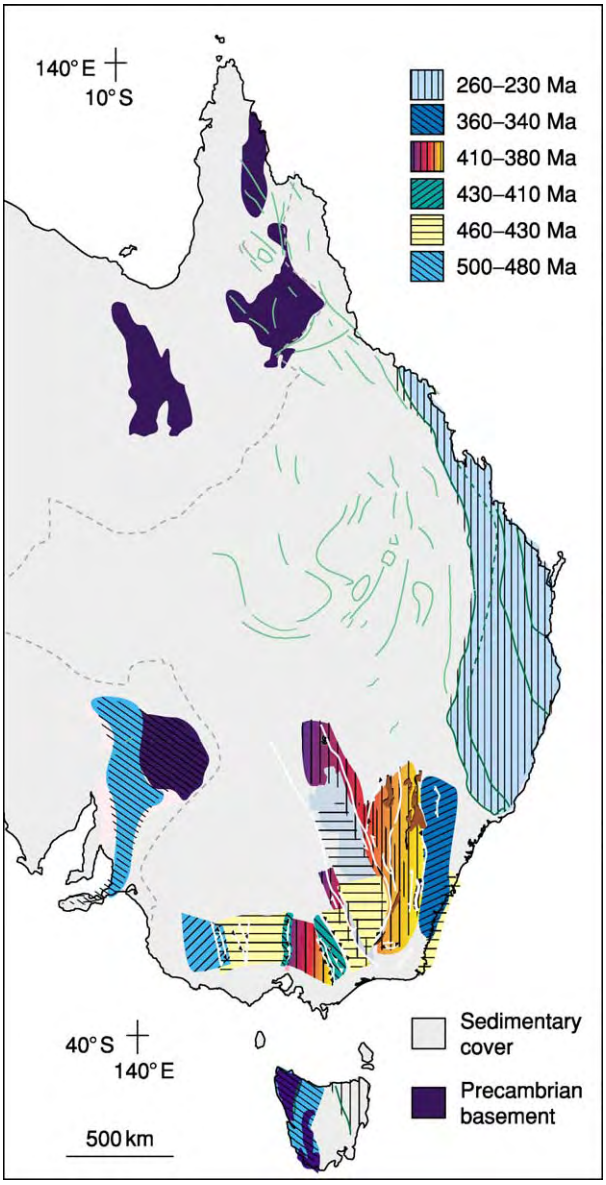


Figure 5 Geological map of the Tasman Orogenic Belt showing geochronological data.

low-temperature–intermediate-pressure blueschist facies metamorphism is recorded by blueschist blocks (knockers) in serpentinite–talc-matrix mélanges within other major fault zones of the Lachlan Orogen. Winchite (blue sodium–calcium amphibole) and glaucophane (blue sodium amphibole)–actinolite assemblages give estimated pressures of 6–7 kbar and 5–6.5 kbar, respectively, with temperatures of less than 450°C. These conditions match those inferred from metabasite blueschists from the Franciscan Complex, California and the Sanbagawa Belt of Japan. This low-temperature–intermediate-pressure metamorphism is coincident with the regional deformation at around 450–445 Ma (Lachlan Orogen).

Magmatism

Distinct pulses of magmatism are recorded by the granitic rocks that occupy up to 35% of the exposed area of the Tasman Orogenic Belt (Figure 8). Granitic bodies tend to be large (up to 10 000 km^2) and commonly elongated subparallel to the structural grain. Regional aureole, contact aureole, and subvolcanic field associations, as well as S and I types based on geochemistry and mineralogy, have been recognized. They are largely post-tectonic with undeformed and narrow (1–2 km wide) contact aureoles, although some (e.g. the Wagga-Omeo Metamorphic Complex and the Eastern metamorphic complexes of the Lachlan Orogen; Figure 7) are coeval with regional deformation. In the eastern Lachlan Orogen some granites coincident with the Eastern Metamorphic Belt were emplaced along major westwards-dipping shear zones. Some (e.g. the Central Victorian Magmatic Province; Figure 8) are subvolcanic granites associated with rhyolites and ash flows of similar composition, and reflect intrusion to shallow crustal levels (*ca.* 1–4 km). The regional-aureole types are less common and are associated with the high-temperature–low-pressure metamorphism, migmatites, and alkali feldspar–cordierite–andalusite–sillimanite gneisses (e.g. Mount Lofty Ranges, Glenelg Zone, Moornambool Complex, Cooma, Cambalong, and Kuark belts, Wongawibinda Complex; Figure 7).

Granite belts with distinctive shapes (Figure 8) reflect distinct tectonic regimes. The north-north-westwards-trending granites of the Central Lachlan Orogen (1 in Figure 8), the more northwards-trending granites of the eastern Lachlan Orogen (2 in Figure 8), and the northwards- to north-north-westwards-trending mid-Permian to Triassic granites of the New England Orogen (3 in Figure 8) are representative of arc and continental-margin–arc associations comparable to the Cordilleran batholiths of the western USA. Mid- to lower-crustal melting is due to crustal thickening and/or subduction, with some granitic activity – in particular the Silurian–Early Devonian granites of the eastern Lachlan Orogen – being linked to subduction rollback. Major magmatic activity occurred in the Cambrian–Ordovician (Delamerian part), Silurian–Early Devonian, Late Devonian, and Carboniferous (Lachlan part), and Carboniferous and mid-Permian–Triassic (New England part) (Figure 8).

Eastern Australian Plate Tectonic Evolution in the Gondwanan Context

During the Palaeozoic the Tasman Orogen was part of a greater Gondwanan oceanic accretionary system (Figures 9 and 10). The cycle of extension, sedimentation, and orogeny that formed eastern

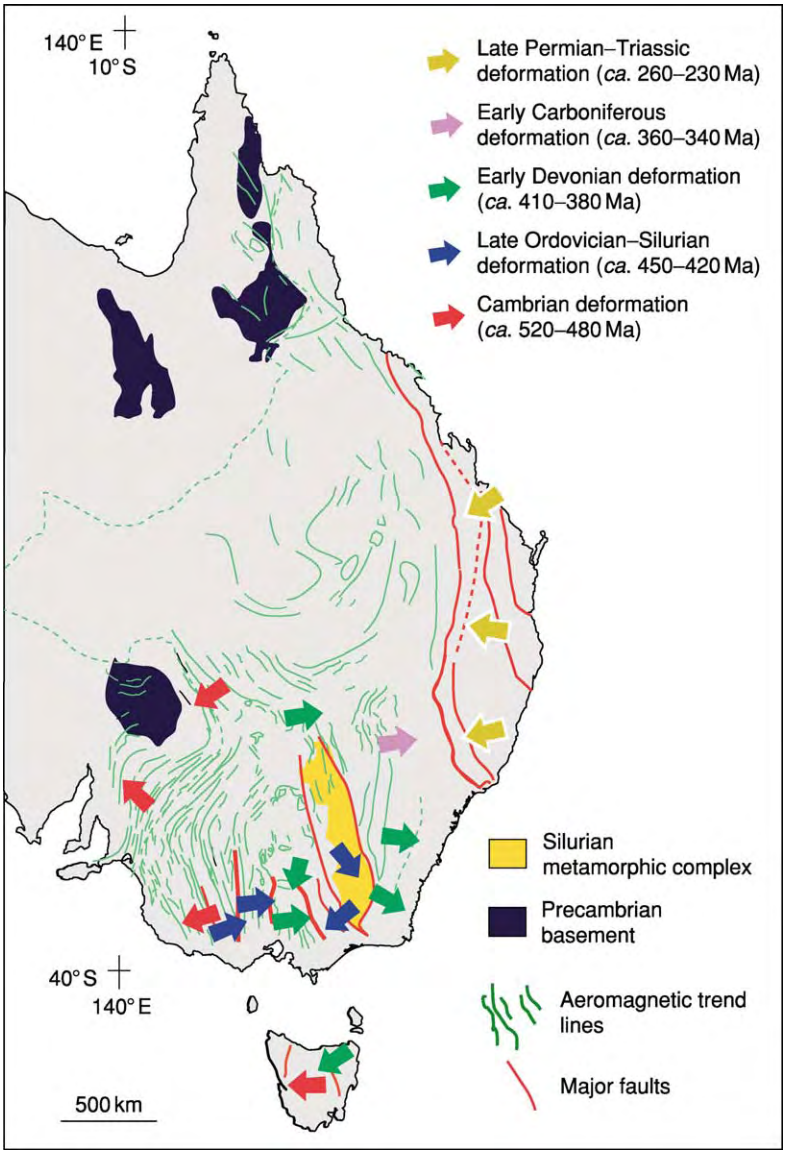


Figure 6 Geological map of the Tasman Orogenic Belt, showing tectonic vergence based on major fault dip. The major fault traces and trend lines from regional aeromagnetic datasets are also shown.

Australia was preceded during the Late Neoproterozoic by initial rifting between cratonic Australia and another large continental plate, probably Laurentia (North America) within the supercontinent of Rodinia (Figure 9A).

Tasman Orogen tectonic evolution has been broken down into a number of significant phases described below and is also presented as a series of Gondwanan-based maps (*see* Figures 9 and 10).

Rodinia Breakup (750–650 Ma): Rifting of Eastern Australia in the Late Neoproterozoic

Rifting began at about 800 Ma but did not lead to separation of the continents until about 700 Ma (Figure 9A). The early phases of rifting were epicontinental and are expressed as rift sequences in the Adelaide Rift and other graben structures. After separation, this sequence takes on a character that is more like that of a passive margin. In the Early Cambrian back-arc spreading in the south-east formed the deep-water Kanmantoo Basin, which rapidly subsided and was filled with clastic sediments. Very shortly after deposition of the Kanmantoo sediments the basin was inverted in the first stage of the Delamerian Orogeny.

Basin Inversion along the Gondwanan Margin (520–500 Ma): Ross–Delamerian Orogeny

The Delamerian orogenic event involved closure of the Kanmantoo Basin by folding and thrusting, metamorphism, partial melting, and intrusion of granitic

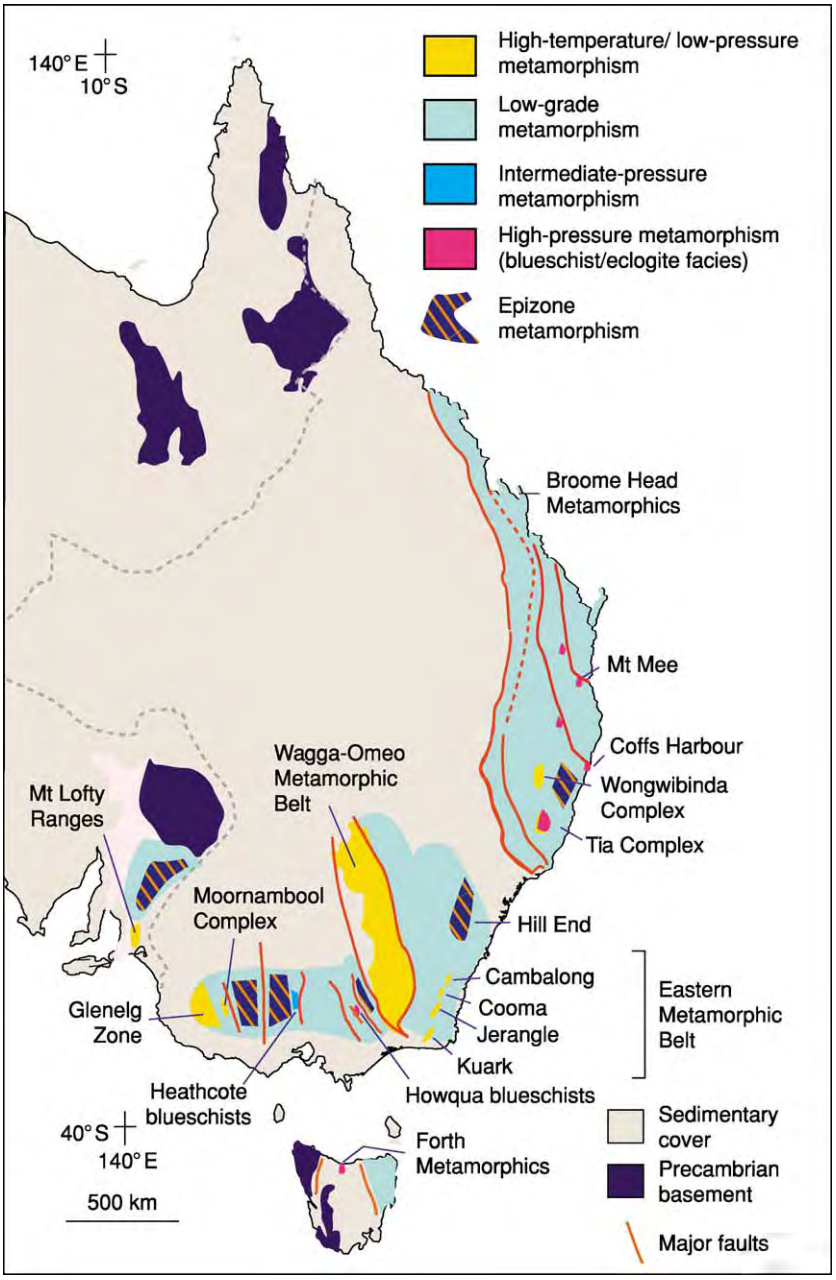


Figure 7 Simplified metamorphic map of the Tasman Orogenic Belt, showing the distribution of the main regional metamorphic facies and the locations of the high grade metamorphic complexes.

plutons (Figure 9B). The driving force for this orogenic event was presumably related to subduction that resulted in basin inversion. Subduction on the Pacific margin of the orogen is suggested by the presence of arc volcanics and plutons. The ultimate closure of the Kanmantoo Basin may have resulted from the collision of the margin with formerly rifted continental fragments that form the basement of the Glenelg Terrane or Tasmania or from ophiolite obduction. The Delamerian event began at about 520 Ma, continued until about 490 Ma, and involved westwards-directed thrusting. It is uncertain whether any oceanic crust was consumed between cratonic Australia and rifted continental ribbons in western Victoria, or whether the Kanmantoo Basin was developed only on thinned continental crust. The early phases of deformation were dominated by folding, metamorphism, and thrusting, followed shortly by syntectonic plutonism. The 'orogeny' culminated in local high-temperature–low-pressure metamorphism involving partial melting in

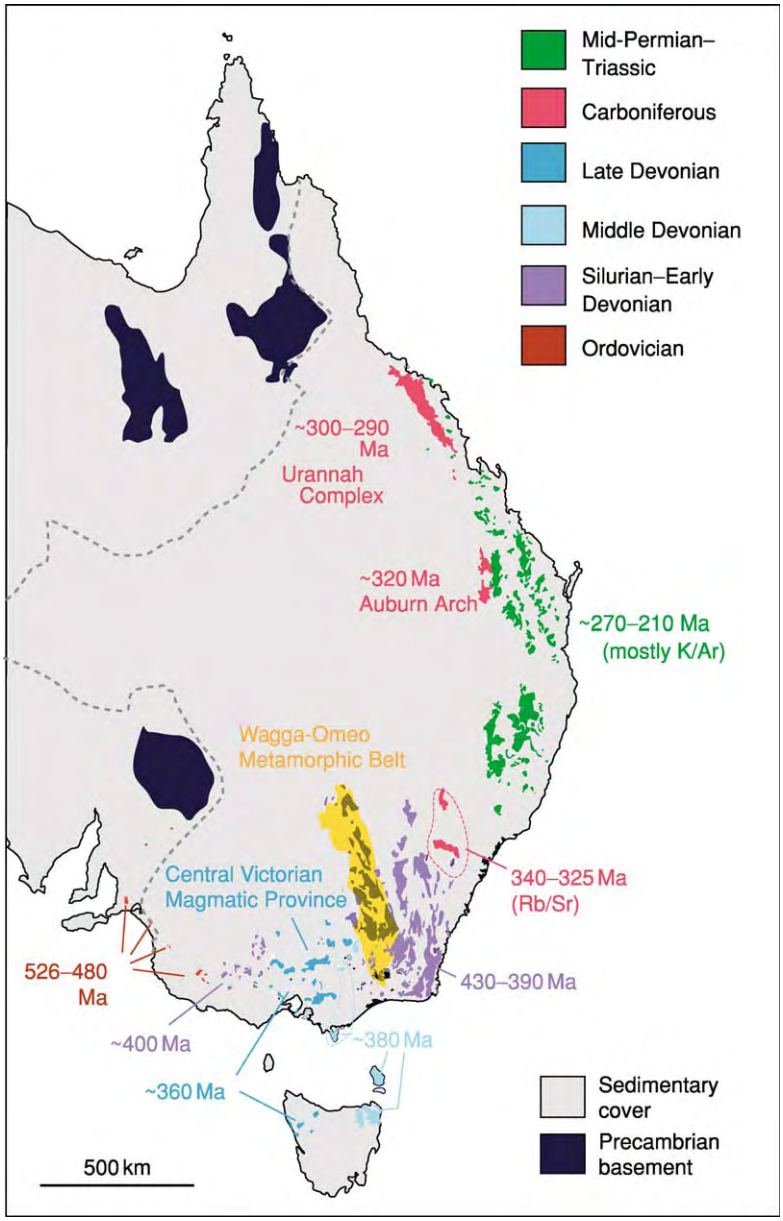


Figure 8 Ages and distribution of granite intrusives associated with the Tasman Orogen.

the highest-grade areas (at about 506 Ma) and was followed by post-tectonic plutonism (*ca*. 490–480 Ma).

Back-arc Basin Formation (520–500 Ma): Evolution of the Lachlan Orogen

A magmatic arc grew on the eastern side of the Delamerian Orogen, starting at about 510–500 Ma, and included plutons and volcanic rocks of the Stavely Belt in Victoria (Figure 4). Shortly after this major crustal thickening the Delamerian belt began to extend rapidly. Extension is suggested by, first, the dominantly alkaline and extension-related nature of the post-tectonic plutons (intrusion ages younger than about 490 Ma), second, the inferred rift setting of the Mount Stavely volcanic complex, and, third, thermochronological data indicating relatively rapid cooling of the higher-grade areas. Post-orogenic extension may have been caused by subduction roll-back after the Delamerian event, and this probably formed most of the oceanic (back-arc) basement for the Lachlan turbidites that were deposited during the Late Cambrian and Ordovician. Most of the mafic metavolcanic rocks and plutons of the central Victorian basement give ages of about 500 Ma. Extension induced by slab roll-back may also have rifted small continental fragments away from the cratonized Delamerian Orogen and distributed them within the developing back-arc basin.

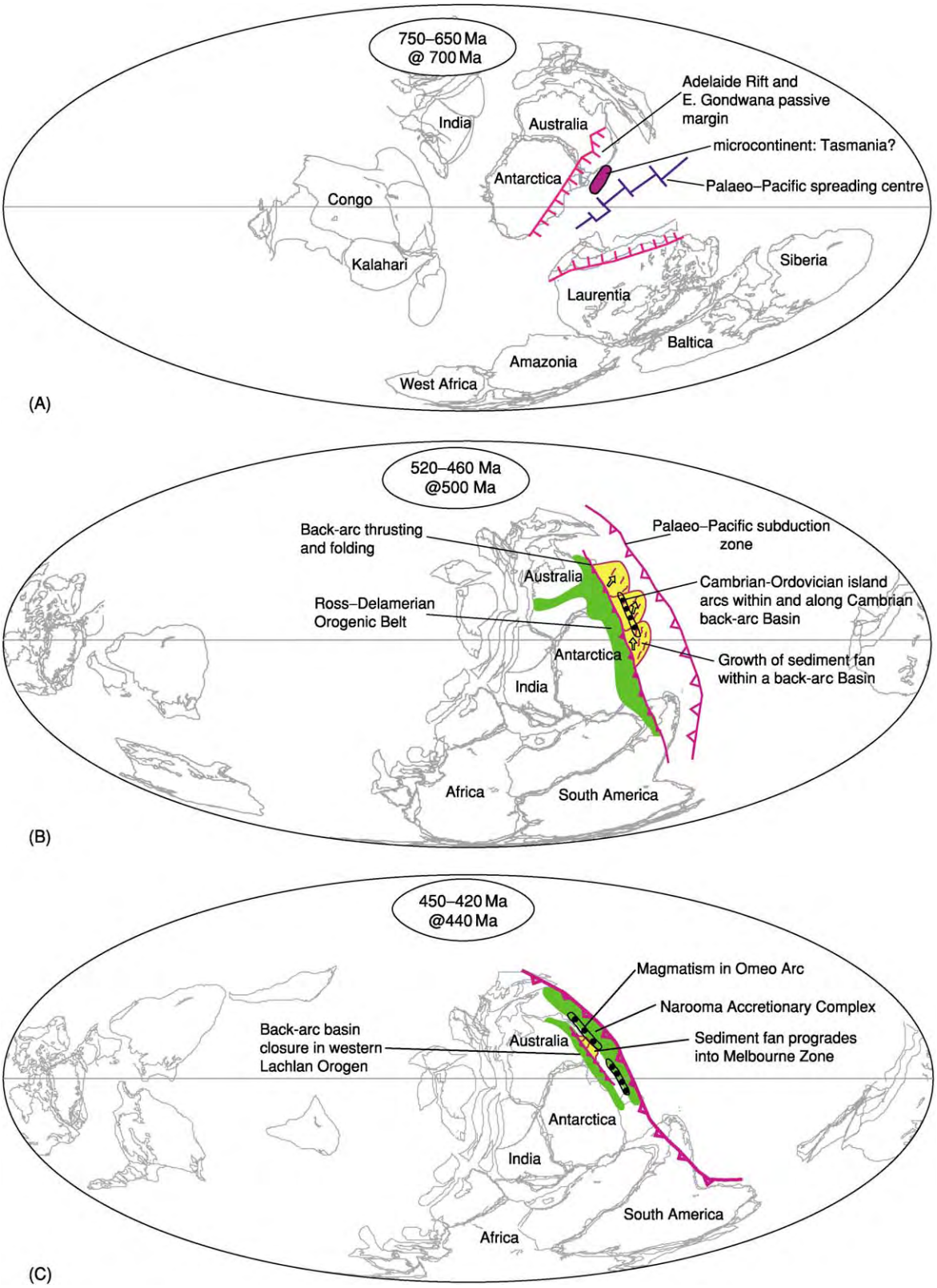


Figure 9 Plate tectonic reconstructions at (A) 750 650 Ma, (B) 520 460 Ma, and (C) 450 420 Ma. Adapted and reprinted, with permission, from the Annual Review of Earth and Planetary Sciences, Volume 28, © 2000 by Annual Reviews, www.annualreviews.org. The plate positions and continental outlines were calculated and drawn using data in Paleogeographic Information System/MacTM verison June 7, 1997 by MI Ross & CR Scotese (PALEOMAP Project, Arlington Texas).

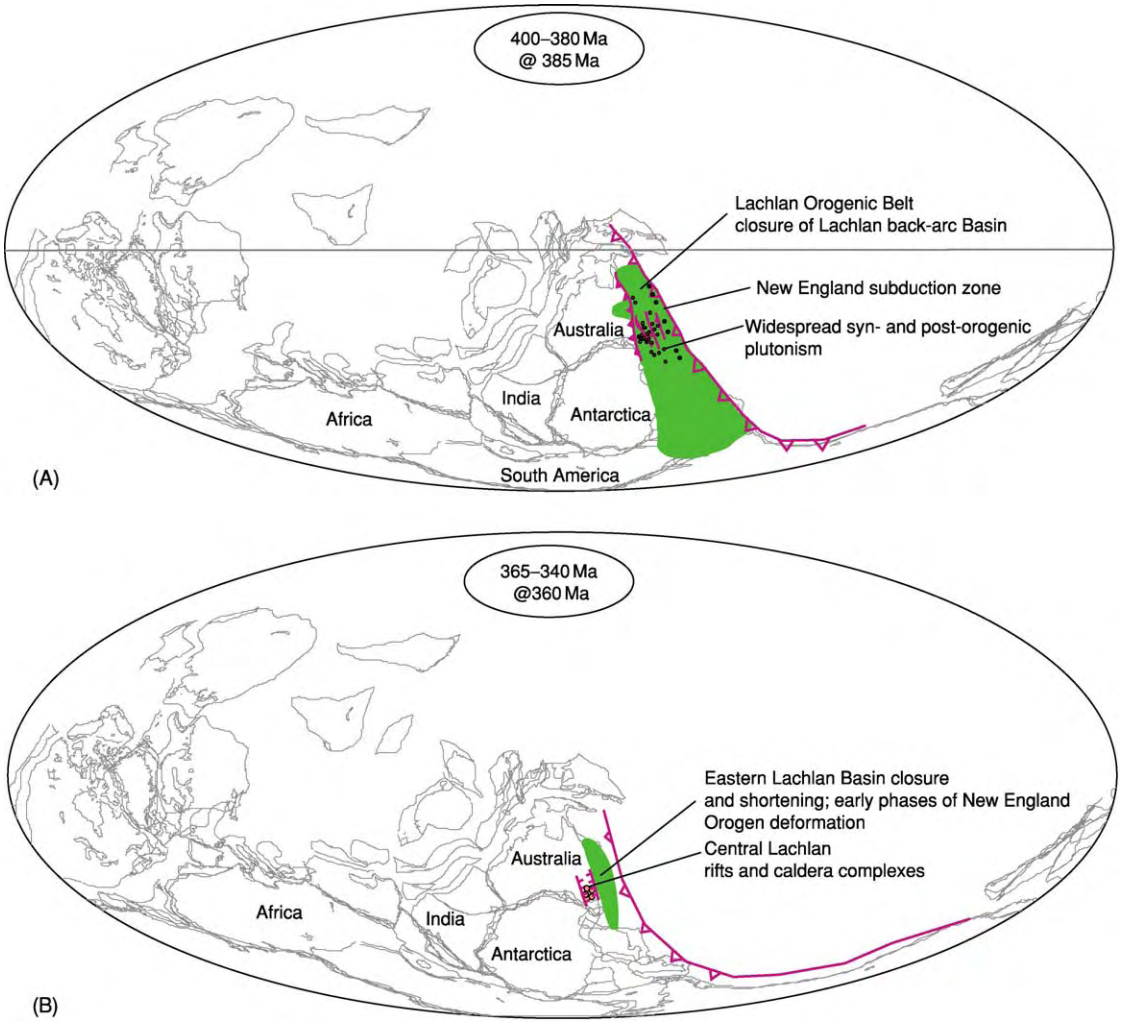


Figure 10 Plate tectonic reconstructions at (A) 400 380 Ma and (B) 365 340 Ma. Adapted and reprinted, with permission, from the Annual Review of Earth and Planetary Sciences, Volume 28, © 2000 by Annual Reviews, www.annualreviews.org. The plate positions and continental outlines were calculated and drawn using data in Paleogeographic Information System/MacTM verison June 7, 1997 by MI Ross & CR Scotese (PALEOMAP Project, Arlington Texas).

During this time large turbidite fans grew off the shore of the Ross–Delamerian mountain chain and spread onto the mafic crust of the developing marginal ocean (Figure 9B). Some thousands of kilometres off shore subduction had initiated within the oceanic plate, leading to the development of a volcanic arc complex at about 485 Ma, now represented by the Ordovican shoshonites in the eastern Lachlan Orogen, and an associated accretionary complex at about 445 Ma (Narooma Accretionary Complex), exposed in south-eastern New South Wales.

Back-arc Basin Closure (450–420 Ma): Evolution of the Lachlan Orogen

The Lachlan Orogen formed by amalgamation of a series of thick turbidite-dominated thrust wedges and volcanic-arc terranes in a tectonic setting similar to that of the Philippine, Mulucca, and New Guinea sectors of the western Pacific today (Figure 9C). The Lachlan Orogen resulted from the closure of a small ocean-basin–arc system situated along the Pacific margin of Gondwana, inboard of a larger major long-lived Palaeozoic subduction system that is now exposed in the New England Orogen (Figure 10A). The total amount of subducted oceanic lithosphere was relatively small (less than 1000 km).

Basin closure involved Woodlark Basin-style double divergent subduction (Figure 9C), inferred from the multiple subduction-related thrust systems (Figure 6) and the presence of the blueschist blocks in the serpentinite-matrix mélange along major faults (Figure 7). The thrust systems developed by duplexing

in the oceanic crust and imbrication in the overlying turbidite wedge. Shallow-angle subduction had probably initiated in the west, under the backstop of the Delamerian Orogen, by about 460 Ma, when major faults began forming in the western Lachlan Orogen. Sedimentation was continuing outboard in the basin as the western Lachlan sediment wedge thickened and the basal décollement propagated eastwards. A magmatic arc associated with this westwards-dipping subduction zone did not develop until Late Silurian times, probably because the subduction was of very shallow dip and the ocean basin that closed was relatively small, so the amount of subduction was limited. Shallow-dipping subduction is consistent with the moderately high metamorphic pressure–temperature ratios. The eastwards-dipping subduction system of the central Lachlan Orogen resulted in the high-temperature–low-pressure metamorphism and elongated north-westwards-trending Late Silurian to Early Devonian granitoids of the Wagga-Omeo Metamorphic Belt (Figures 7 and 8) as well as the emplacement of ophiolite slices (e.g. Howqua, Dookie, and Tatong greenstones; Figure 4) and the Howqua blueschists (Figure 7).

Andean-type Margin (400–380 Ma): Evolution of the Lachlan Orogen

During the Early Devonian, at around 400 Ma, the last undeformed part of the marginal basin in the west collided with the arc–metamorphic complexes of the central Lachlan Orogen (Figure 10A). This caused some crustal thickening and deformation of the younger sediments (former Melbourne 'trough'). Combined structural thickening and the removal of the oceanic lithosphere of the marginal basin formed crust of continental thickness and character, and led to a switch to continental sedimentation and the development of a major regional fold-belt-wide angular unconformity. These effects are related to the Tabberabberan orogenic event.

Collision followed by attainment of freeboard over most of the Lachlan Orogen led to the development of an Andean-type margin (Figure 10B). Continued westwards-dipping or craton-directed subduction of proto-Pacific lithosphere beneath this part of the Gondwanan margin along the major outboard long-lived subduction zone caused the magmatism that was responsible for the extensive northwards-trending elongated granitoids in the eastern Lachlan Orogen (Figure 8). An easterly younging in these granitoids suggests associated slab rollback during the mid-Late Silurian to Early Devonian, with rollback rates of the order of 6 mm year ^{1}. Thrusting was also taking place in the eastern Lachlan Orogen at this time.

Roll-back and Gondwanan Margin Post-Orogenic Extension (365–340 Ma)

Marked crustal-scale extension at around 380 Ma caused rifts, basin-and-range faulting, voluminous explosive volcanism and caldera development, and high-level plutonism in the western and central Lachlan Orogen (e.g. the Central Victorian magmatic province; Figure 8). By 340 Ma subduction in the eastern Lachlan Orogen had probably stopped, and reefs were growing around the volcanic edifices. The major subduction zone at this time probably stepped east more than 1000 km to form the New England Orogen (Figure 10B).

Andean Margin (350–280 Ma) and Arc–Continent Collision (260–230 Ma): Evolution of the New England Orogen

During the Carboniferous, westwards-dipping subduction beneath the Gondwanan margin led to silicic and intermediate volcanism from 350 Ma to 310 Ma, coincident with an inboard belt of Late Carboniferous rocks (330–325 Ma). At around 300 Ma magmatic activity migrated outboard into the developing associated Devonian–Carboniferous accretionary prism. Migration of the volcanic arc was associated with widespread extension in the Late Carboniferous to Early Permian due to slab retreat along the long-lived subduction zone, with the development of an outboard intraoceanic arc (Gympie Belt). Collision of this arc with the Gondwanan margin in the Late Permian led to the Hunter–Bowen Orogeny (260–230 Ma), which juxtaposed and accreted the Carboniferous flank–forearc sequence (Yarrol–Tamworth Belt), the Devonian arc complex, and the Devonian–Carboniferous subduction complex (Wandilla–Gwydir Belt) by craton-directed overthrusting. Margin overthrusting led to molasse sedimentation in a developing foreland basin (Sydney–Bowen Basin). Cordilleran-style granites (e.g. New England batholith) reflect renewed subduction magmatism at this time, with the development of a Late Permian–Early Triassic magmatic arc. This was followed by Late Triassic extension, which is recorded by silicic caldera-related volcanism and granitic plutonism as well as the development of small extensional basins with bimodal volcanism.

See Also

Andes. **Antarctic**. **Australia:** Proterozoic; Phanerozoic. **Gondwanaland and Gondwana**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mountain Building and Orogeny.

Further Reading

Ashley PM and Flood PG (eds.) (1997) *Tectonics and Metallogenesis of the New England Orogen*. Special Publication 19. Sydney: Geological Society of Australia.

Birch WD (ed.) (2003) *Geology of Victoria*. Special Publication 23. Sydney: Geological Society of Australia.

Burrett CF and Martin EL (eds.) (1989) *Geology and Mineral Resources of Tasmania*. Special Publication 15. Sydney: Geological Society of Australia.

Coney PJ, Edwards A, Hine R, Morrison F, and Windrum D (1990) The regional tectonics of the Tasman orogenic system, eastern Australia. *Journal of Structural Geology* 125: 19 43.

Flöttmann T, Klinschmidt D, and Funk T (1993) Thrust patterns of the Ross/Delamerian orogens in northern Victoria Land (Antarctica) and southeastern Australia and their implications for Gondwana reconstructions. In: Findlay RH, Unrug R, Banks HR, and Veevers JJ (eds.) *Gondwana Eight, Assembly, Evolution and Dispersal*, pp. 131 139. Rotterdam: Balkema.

Foster DA and Gray DR (2000) The structure and evo lution of the Lachlan Fold Belt (Orogen) of eastern Australia. *Annual Reviews of Earth and Planetary Sciences* 28: 47 80.

Foster DA, Gray DR, and Bucher M (1999) Chronology of deformation within the turbidite dominated Lach lan orogen: implications for the tectonic evolution of eastern Australia and Gondwana. *Tectonics* 18: 452 485.

Gray DR and Foster DA (1997) Orogenic concepts application and definition: Lachlan fold belt, eastern Australia. *American Journal of Science* 297: 859 891.

Scheibner E and Basden H (eds.) (1998) *Geology of New South Wales Synthesis. Volume 2: Geological Evolution*. Memoir Geology 13(2). p. 666. Sydney: Geo logical Survey of New South Wales.

Veevers JJ (ed.) (1984) *Phanerozoic Earth History of Australia*. Oxford Monographs on Geology and Geo physics 2. Oxford: Oxford University Press.

Veevers JJ (ed.) (2000) *Billion year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

BIBLICAL GEOLOGY

E Byford, Broken Hill, NSW, Australia

World-View of the Hebrew Scriptures

The world-view of those who wrote or edited the Hebrew Scriptures reflects the commonly held contemporary understandings of the structure of the universe. The editors of the Hebrew Scriptures (including those who wrote the first chapter of the Book of Genesis) did their work in the period immediately following the end of Babylonian exile. The return from exile followed the conquest of Babylon by Cyrus, King of Persia, in 538 BCE. The world-view in the scriptures was that of the desert people who lived in the Fertile Crescent of modern Iran, Iraq, Syria, Israel, and Jordan. The clearest, detailed articulation of the structure of the world and the universe is to be found in the Book of Job, Chapters 38–40.

What is described in Genesis 1:1 to 2:3 was the commonly accepted structure of the universe from at least late in the second millennium BCE to the fourth or third century BCE. It represents a coherent model for the experiences of the people of Mesopotamia through that period. It reflects a world-view that made sense of water coming from the sky and the ground as well as the regular apparent movements of the stars, sun, moon, and planets. There is a clear understanding of the restrictions on breeding between different species of animals and of the way in which human beings had gained control over what were, by then, domestic animals. There is also recognition of the ability of humans to change the environment in which they lived. This same understanding occurred also in the great creation stories of Mesopotamia; these stories formed the basis for the Jewish theological reflections of the Hebrew Scriptures concerning the creation of the world. The Jewish priests and theologians who constructed the narrative took accepted ideas about the structure of the world and reflected theologically on them in the light of their experience and faith. There was never any clash between Jewish and Babylonian people about the structure of the world, but only about who was responsible for it and its ultimate theological meaning.

The envisaged structure is simple: Earth was seen as being situated in the middle of a great volume of water, with water both above and below Earth. A great dome was thought to be set above Earth (like an inverted glass bowl), maintaining the water above Earth in its place. Earth was pictured as resting on foundations that go down into the deep. These foundations secured the stability of the land as something that is not floating on the water and so could not be tossed about by wind and wave. The waters surrounding Earth were thought to have been gathered together in their place. The stars, sun, moon, and planets moved in their allotted paths across the great dome above Earth, with their movements defining the months, seasons, and year.

In the world of the Jewish scribes and priests who assembled the Hebrew Scriptures, the security of the world (the universe) was 'guaranteed' by God. It was God who made the world secure. It was God who made Earth move or not move, as the case might be. When this is understood, the ancient sources can be understood, and many of the geological events that produced the theological reflections can be deciphered.

Geological Events with References in the Hebrew Scriptures

The Angel with the Flaming Sword (Genesis 3:24)

The location of the Garden of Eden was described as being "in the east". Four rivers were said to have flowed out of Eden. The naming of the Tigris, the Euphrates, and two other streams that have names that translate literally from the Hebrew to 'gusher' and 'bubbler' suggests that the garden was located at the top of the Persian Gulf. This area is well watered and would have been lush compared to the desert area that was the habitual location of the Jewish people from the beginning of the first millennium BCE.

The creatures with flaming sword are described as cherubim. In the ancient world, these were supposed awesome (hybrid) creatures and were the steeds that carried the high god of the Canaanites through the air. What is interesting is that the reference to the sword was made using the definite article, as if 'it' (whatever it was) were well known. This is a region well known for its petroleum deposits and oil and gas fires. A gas release or petroleum seep that caught fire could last for many years and extend hundreds of metres above the ground. It would have released much light, heat, and sound and would have been of the very nature described in the Biblical text as keeping people away from the locality.

The Flood: Genesis 6–9

Since, at latest, the middle of the nineteenth century, scholars, both scientific and theological, have rejected as impossible the idea of a flood that inundated all of Earth's landmasses. The flood story has been treated as ancient myth or as a theological metaphor for the return to supposed primeval chaos as a consequence of human sin. It can also be regarded as an expansion of local flood stories from the flood plains of great rivers and coastal areas to some sort of universal status. But since the work of William Ryan and Walter Pitman, there has been renewed interest in the occurrence of a catastrophic flood.

At the time of maximum glaciation at the height of the last Ice Age (about 16 000 BCE), sea-levels were about 120 m lower than present levels are. There was no longer a connection between the Black Sea Basin and the Mediterranean Sea. With the first great melting of the Eurasian Ice Sheet (12 500 BCE), the Black Sea filled with fresh water, which flowed through the Sakarya Outlet into the Sea of Marmara and down the Dardanelles Outlet into the Aegean. Later, with only seasonal flows down the rivers feeding the Black Sea, the level of this body of water fell far below its great melt maximum. But the level of the Mediterranean, connected to the waters of the world's oceans, continued to rise over the next 6000 years while at the same time the former channels from the Black Sea to the Sea of Marmara filled with rubble, thus creating a dam between the Mediterranean and the Black Sea Basin. By 6000 BCE, the sea-level in the Mediterranean was at least 100 m higher than the surface of the great Black Sea freshwater lake.

Ryan and Pitman have proposed a single geological event, namely, the breaching of that dam through what is now the Bosphorus Straight, from which followed the sudden and catastrophic flooding of the Black Sea Basin with salt water. Ryan and Pitman date the disappearance of freshwater molluscs and the introduction of saltwater molluscs to 5600 BCE. This happened suddenly, not gradually. Their submarine survey work has indicated that the Bosphorus was eroded by immense flows from the Mediterranean to the Black Sea. The inclination of the channel towards the Black Sea also accounts for the surface current of relatively fresh water towards the Mediterranean Sea and the lower more saline current directed towards the Black Sea. The dating of this geological event to early agricultural times links it to the Noah/Gilgamesh story.

There is some controversy in the early twenty-first-century geology literature concerning Ryan and Pitman's hypothesis of a single catastrophic breaching of the Bosphorus. At the time of publication, that controversy continues. However, Ryan and Pitman's hypothesis is not based simply on the dating of marine molluscs, but significantly on the erosion profiles of the Bosphorus Channel. These profiles reveal the continuation of the erosion channel beyond the escarpment, well below present sea-levels.

Most flood stories came from areas of flood plains and coastal districts and can be related to the sea-level rises associated with the end of the last Ice Age (10 000 to 12 000 years ago). The Noah legend is significantly different from these in that it is associated with the Anatolian highlands, well away from the coast and well above sea-level. The Black Sea had been a saltwater sea, but had become a large freshwater lake. Following the last Ice Age, it then shrank to become a lake with a surface about 100 m lower than the present level, but suddenly filled and became salty at about 5600 BCE. More recent discoveries by Robert Ballard indicate that there were quite sophisticated settlements on the shore of the freshwater lake before the reintroduction of salt water. These settlements seem to have been built on well-established agricultural and animal husbandry practices. Ryan and Pitman estimate that, after the breach of the Bosphorus, water would have flowed into the Black Sea Basin at about $40\,km^3\,day^{-1}$ and that the water level would have risen by about $15\,cm\,day^{-1}$. They estimate that the sound of the great inrush could have been heard up to nearly 500 km away. The water would have surged down the escarpment accompanied by large volumes of spray and deafening noise.

According to the Genesis account of the flood (Genesis 6–8), Noah was warned of the impending disaster and was told to make preparations for the preservation of animal and plant stock, which were to be preserved and transported to safety by means of a vessel that Noah was to build. The flood is described as coming from the sky and from the deep (Genesis 7:11). Noah had time to make the necessary preparations to save his domestic breeding and seed stock, for it would have taken nearly two years to fill the Black Sea Basin. In the flat country on the northern shores of the old freshwater lake, the advance of shoreline would have been about $1\,km\,day^{-1}$. The legend could well refer to an actual historical event, because people could have known of the rising water level on the Mediterranean side of the Bosphorus barrier, prior to its rupture.

It is suggested that those who fled the Black Sea Basin went in all directions, into northern and western Europe, and, significantly, across the Anatolian Plateau (adjacent to the location of the mountains of Ararat), into the Fertile Crescent, where the newcomers introduced their more sophisticated

agricultural practices and preserved the story of the destruction of their world in the great flood.

The Destruction of Sodom and Gomorrah (Genesis 19:24–26)

The description of the destruction of the Sodom and Gomorrah, two cities on the Plain, contains significant elements that correspond to massive earthquakes and the opening of sulphurous springs. Traditionally, Sodom and Gomorrah have been thought to have been located in the valley of the Dead Sea, perhaps even where the Dead Sea is now located. In spite of clear suggestions of a catastrophic geological event, no particular event has been satisfactorily identified with this story, nor have any identifiable ruins of early second-millennium-BCE towns or cities been found. Thus the description stands, but a particular event or location remains a matter of conjecture.

The Exodus

The Crossing of the Red Sea—Two Stories in Exodus 14 There are two accounts of the crossing of the Red Sea (the Hebrew name of which can also be translated as 'Sea of Reeds' or 'Distant Sea') in Chapter 14 of the Book of Exodus. The accounts are intricately intertwined but refer to two readily identifiable and distinguishable phenomena. It is generally agreed that the route of the Exodus from Egypt into the Sinai was close to the northern end of the Red Sea.

The first of these phenomena concerned the action of steady, strong winds on shallow expanses of water. Water can be removed a long way from the normal shoreline of one side of a lake or inlet, for the duration of the wind, and can then return to normal levels when the wind subsides. (This is a commonly observed phenomenon at Lake George, Australia, near Canberra.) At the beginning of the first account of the crossing (Exodus 14:21), we are told "the Lord drove the sea back by a strong east wind all night, and made the sea dry land". This account then tells of the attempt of the Egyptians to follow through after the Hebrews and that their chariot wheels became clogged and they fled. When it is recognized that the Hebrews had herds and were mostly on foot and would have further softened ground that was usually under water, then the clogging of the wheels of chariots would be expected. The dropping of the wind and the return of the water could have been a perfectly natural phenomenon.

The second account is much more dramatic, with references to great walls of water (Exodus 14:22). The destruction of the Egyptians was said to have been accomplished by the catastrophic return of the water (Exodus 14:28). This second account is of a far more destructive phenomenon, compared to the first, and probably more familiar one. In the second account, the sea pulls back and then surges over those who have ventured onto the ground from which the sea receded. This is a classic description of the destruction wrought by a tsunami.

The dating of the events of the Exodus has never been exactly agreed by scholars. The generally agreed range of dates is from the fifteenth century BCE to the middle of the thirteenth century BCE. Near the beginning of this period, there was, in fact, an event that could have produced a catastrophic tsunami of the type described in Exodus. In 1470 BCE, the volcano Santorini erupted with about the same force as that of Krakatau in 1883. These eruptions are the largest in historical times. The tsunami resulting from the Santorini eruption wrought havoc in northern Crete and on Milos and in the Peloponnese. This same tsunami would have swept south, causing immense damage along the Egyptian coast and especially in the low-lying agricultural areas of the Nile Delta. The sea would have withdrawn, possibly for 15 or 20 min, and then would have come the first return wave. The wave that hit Milos had a height of about 100 m and travelled at about 300 km h^{-1}. Even at half this speed and height, the destruction in Egypt would have been immense. Such an event would have been associated with the gods, and, for the Hebrews, with their own God. Thus, in Exodus 14, the accounts are of a real catastrophic event and of another event, of more common experience, combined into a single story of the power of the Hebrew God to save the Hebrew people.

The Plagues (Exodus 7–11) Evidence from the Greenland ice sheets implies that the Santorini eruption generated high levels of sulphuric acid. There would have been sustained acid rain in the eastern Mediterranean and cooling for many years as result of the dust blasted into the atmosphere. There would have been initial crop failures as a result of the acid rain and continued low yields as a result of lower temperatures. The dust blasted into the high atmosphere would have darkened the sky, making the sun and moon appear red, perhaps for several years.

Acid rain would also have contaminated water, causing destruction of aquatic plants and fishes. If this were combined with significant ash-falls, the waterways could have been rendered anoxic for a considerable length of time. Moreover, contaminated water and acid rain, with sudden loss of pasture, would be associated with a rise in stock disease and invasion of settled areas by insects and other vermin in search of food sources.

In the plague stories, descriptions of agricultural phenomena are consistent with the effects that would have followed the Santorini eruption. The sequence of plagues in Exodus were as follows: (1) blood (the Nile turns to 'blood' and everything in it died), (2) frogs, (3) gnats, (4) flies, (5) livestock disease, (6) boils, (7) hail and thunderstorms, (8) locusts, (9) darkness, and (10) death of the first-born child. Except for the last (which is clearly related to Jewish religious and cultic practice), each of the plagues could have been caused by the eruption of Santorini. Whatever else is concluded, it is clear that phenomena that would have been consequences of the eruption of Santorini were associated directly, in the stories and literature of the Hebrews, with the supposed divine intervention that led to the Exodus.

The Crossing of the River Jordan (Joshua 3–4) The crossing of the Jordan on dry land, by Joshua and descendants of those who had fled Egypt, is said to have been accomplished as a result of the waters "rising up in a single heap far off at Adam, the city that is beside Zarethan" (Joshua 3:16). Adam is at the junction of the Jordan and Jabbok rivers, about 25 km north (upstream) of the closest crossing point of the Jordan to Jericho. The Dead Sea/Jordan Valley is part of geologically active rift system where the Eurasian and African plates interact. Movements along the fault system could have produced a rubble dam that blocked the flow of the Jordan, as described in Joshua, thus allowing a crossing on dry ground or through very shallow water.

The Tablets of Stone with the Law Inscribed on Them The Ark of the Covenant, containing the tablets of stone that Moses brought down from the mountain, was the significant agent in the story of the Jordan crossing. The tablets were understood to have been given to Moses after God inscribed the law on them. The story of the initial giving of the tablets is in a series of events described in Exodus 24:12 and 31:18. The breaking of the stones is described in Exodus 32:19 and the giving of new stones is told in Exodus 34.

There is Precambrian graphic granite in the southern Sinai, the rock being formed by an intergrowth of quartz in feldspar derived from a quartz–feldspar melt, which gives the appearance of writing in stone (see Figure 1). Tablets of this material can be up to 20 by 20 cm. They are brittle, and when dropped, they cleave along the plane of crystallographic weakness.

Moses Strikes the Stone to Produce Water: Exodus 17:1–7 The regions associated with the wanderings of the Exodus are in the Sinai Peninsula, south of the Dead Sea and in the approach to the Holy Land on the eastern side of the Dead Sea and the Jordan. The river was crossed from the east and the siege of Jericho began. To the south and east of the Dead Sea, there are permeable sandstones overlaying impermeable granites. Water seeps down through the sandstones until it reaches the impermeable granite. Where there are vertical fractures, springs occur where water emerges from the rock face to produce a rock pool. A sequence of such springs occurs along the edge of the formations. One such pool, Wadi Musa (in modern Jordan), is identified by the local Bedouin with the story of Musa (Moses) striking the rock to get water.

Figure 1 Example of graphic granite (from Thackaringa Hills, west of Broken Hill, NSW, Australia). Photograph by Edwin Byford.

References to Earth Movements

The Hebrew and Christian scriptures originated at the eastern end of the Mediterranean, a region of geological activity throughout historical times. God was described as both securing the stability of the earth (Psalms 93 and 96) or as shaking the earth (Isaiah 29). God thus both protected people from earthquakes and caused them.

Earthquakes were used to indicate dates (e.g., Amos 1:1; Zechariah 14:5). Zechariah 14:5 referred back to the Amos earthquake (middle of the eighth century BCE), but the whole passage (Zechariah 14:4, 5) described what may now be construed as the lateral shift caused by an earthquake near Jerusalem. The Mount of Olives tear/wrench fault is part of the complex of tear faults that form the Dead Sea/Jordan Valley. In these regions, the rocks on the western side of the valley are displaced southwards compared to those on the east.

The Scientific Revolution Beginning in the Sixteenth Century, and Christian Responses Thereto

For a convenient point of departure, the beginning of the scientific revolution can be taken to be with Nicolas Copernicus (1473–1543), who, while a canon of Frauenburg (1497–1543), formulated a model of the solar system with the sun, not Earth, at its centre. It is commonly accepted that Copernicus was responsible for the seminal ideas that laid the foundations of the work of Kepler, Galileo, and Newton. In spite of great reservations about the new learning (note the condemnations of Galileo and the fact that the work of Copernicus remained on the Index from 1616 to 1757), Pope Gregory XIII revised and corrected the calendar in 1582, following advice from the Vatican Observatory. Although Protestant countries did not adopt the revisions for some centuries (in England and the American colonies not until 1752, and in the eastern Orthodox countries not until 1924), the Gregorian calendar is now the accepted norm.

Seventeenth-century England saw the theoretical foundations for modern science established in the work of Isaac Newton (1642–1727). But it was seventeenth-century Britain that also produced what has become the centrepiece of a perceived clash of science, and especially geology, and Christianity. In the politically and religiously charged atmosphere of the Protectorate, James Us(s)her (1581–1656), who was a graduate of Trinity College, Dublin, and Archbishop of Armagh, calculated the age of Earth, using the ages of the Patriarchs as they were recorded in the Old Testament, and interlocking these dates with those available from the great civilisations of the ancient world. His calculations were given in his *Annales Veteris et Novi Testamenti*, written between 1650 and 1654. Ussher dated the creation of Earth at 23 October 4004 BC. From 1679, Ussher's dates for various events were included as marginal notes in the 'Oxford Bibles' and in modified form in subsequent editions of the Authorised Version of the Bible (the 'Lloyd Bible', 1701), thereby establishing Ussher's dates in the minds of those who read the Bible in English.

By the eighteenth century, a mechanistic understanding of the universe was accepted among theologians and church leaders, especially in Britain and the Protestant maritime nations. Studies in botany, zoology, and anatomy, as well as the developments in mechanics and astronomy, provided data for arguments for the existence of God based on ideas of divine design. William Paley (1743–1805) published his *Natural Theology* (1802), which remained a basic text in apologetics until well into the twentieth century. The newly emerging sciences were understood to provide further proof concerning the existence and benevolent nature of God. The convergence of science and Christian faith was further encouraged by endowments such as that by Robert Boyle (1627–91), of Boyle's Law fame, for annual lectures on Christian apologetics. As yet, there was little to bring Ussher's date of the creation of Earth into question.

All this began to change from the middle of the eighteenth and early nineteenth centuries, with the beginnings of the classifications of fossils and sedimentary strata. For some decades, accommodation of Christian faith and the emerging pool of geological data and interpretation was achieved (more or less) through what has been called 'flood geology', based chiefly on the 'catastrophist' ideas of the Frenchman Georges Cuvier (*see* **Famous Geologists:** Cuvier). But by the middle of the nineteenth century, this was no longer accepted in learned scientific or theological circles. Following on the work of James Hutton (*see* **Famous Geologists:** Hutton), the influential geologist Charles Lyell (1830) (*see* **Famous Geologists:** Lyell) argued for noncatastrophist geology and, using evidence from his work at Mount Etna and elsewhere, for the idea that Earth was immensely old.

Many clergy, especially in England and Scotland, had been in the forefront of the new biological and geological sciences. But this did not prepare them for the publication of Charles Darwin's *Origin of Species* (1859) (*see* **Famous Geologists:** Darwin). The creation of the world was pushed back thousands of millions of years, and human beings could no longer be understood as a unique creation, qualitatively different from all other life forms. Though many scientists, including William Thomson (Lord Kelvin), questioned the extreme age of Earth advocated by Lyell and Darwin, it was certain clergy, most notably Samuel Wilberforce, Bishop of Oxford, who attacked Darwin's theory. (Thomson's objections were based on his understanding of Earth's cooling. He calculated that it would have 'died' were it as old as geological theory indicated. Radioactivity had not then been discovered, and so he had no mechanism by which Earth could be both very old and warm.)

Wilberforce's objections (which successfully drew attention to weaknesses in Darwin's original theory) were not long shared by the theological faculties of the universities of England and Scotland. On the contrary, by the early 1880s, Darwin's theory was being embraced as evidence of the immense providence of God. Frederick Temple's Bampton Lectures of 1884 and the *Lux Mundi* collection of essays edited by Charles Gore in 1889 embraced the new scientific developments.

Through the twentieth century, there has been a perceived opposition between science and Christianity, in particular, driven by those who adhere to Ussher's dating of the creation of the heavens and Earth. Generally known as Creationists, opponents of the 'new' geology and biology have portrayed Darwinian evolutionary theories as anti-Christian. But, interestingly, they have tried to conscript science to the Creationist cause. So-called Creation science (*see* **Creationism**) has tried to adapt scientific learning to prove that Earth is very young (only several thousand years old). In no way does Creation science satisfy the methodological rubrics of the scientific community, and Creationist interpretations of the data are universally rejected. It is interesting that Creation scientists want to appeal to science to establish their case. Even for religious literalists, it would appear that the best way to describe the physical universe is what is generally understood as scientifically. But this is understandable in that they operate chiefly in the 'scientistic' culture of the United States.

In the United States in the late twentieth century, objections to Creation science being taught in school science curricula have been led by mainstream Christians and Jews. (The plaintiffs in the 1981 action to set aside as unconstitutional Act 590 of the Arkansas Legislature, "Balanced Treatment for Creation-Science and Evolution-Science Act", included the resident Arkansas bishops of the United Methodist, Episcopal, Roman Catholic, and African Methodist Episcopal Churches, the Principal Officer of the Presbyterian churches in Arkansas, and other United Methodist, Southern Baptist, and Presbyterian clergy. Organisational plaintiffs included the American Jewish Congress, the Union of American Hebrew Congregations, and the American Jewish Committee.) For most Christians and Jews there is no perceived clash between their religious faith and the scientific enterprise.

The Bible is a collection of religious texts gathered and edited over a period in excess of a 1000 years, concluding in the early decades of the second century CE (anno domini). The texts are products of the times and places in which people have struggled to understand the ultimate significance and meaning of life. They reflect religious conviction and searching and have never been scientific texts, in the way that scientific texts have been understood since the late Middle Ages when modern science began to develop. However, events that may have appeared miraculous to the authors of the Old Testament can be understood satisfactorily in terms of modern geological knowledge. Religious knowledge and scientific knowledge deal with the same world and universe, but they operate with fundamentally different presuppositions of causality. This means that they can overlap and interact, but the causalities for science are proximate and immediate, and, for religion, are ultimate and perhaps eternal.

See Also

Creationism. **Famous Geologists:** Cuvier; Darwin; Hutton; Lyell. **Geomythology**. **Tectonics:** Earthquakes. **Volcanoes**.

Further Reading

Alter R (1996) *Genesis*. New York: WW Norton.

Görür N, Cagatay MN, Emre O, *et al.* (2001) Is the abrupt drowning of the Black Sea shelf at 7150 yr BP a myth? *Marine Geology* 176: 65 73.

Lewis CLE and Knell SJ (eds.) (2001) *The Age of the Earth: From 4004 BC to AD 2002*. London: The Geological Society.

Numbers RL (1992) *The Creationists*. New York and Toronto: Alfred A. Knopf .

Plimer I (2001) *A Short History of Planet Earth*. Sydney: ABC Books.

Ruse M (1999) *Mystery of Mysteries*. Cambridge: Harvard University Press.

Ryan W and Pitman W (2000) *Noah's Flood*. New York: Touchstone.

Wilson I (2001) *Before the Flood*. London: Orion Books.

BIODIVERSITY

A W Owen, University of Glasgow, Glasgow, UK

Introduction

In its broadest sense, biodiversity encompasses all hereditarily based variation at all levels of organization from genes through populations and species to communities and ecosystems. The term is a modern one, stemming from the early 1980s, and is a contraction of 'biological diversity', which has a long pedigree as an area of scientific investigation. Concerns over the rate of extinction of modern species have brought the study of biodiversity (and indeed the word itself) to the forefront of scientific, political, and popular concern. The fossil record both establishes the ecological and palaeobiogeographical origins of present-day biodiversity and provides an understanding of the longer-term influences on biodiversity of changes in climate, eustasy, and many other processes operating at the Earth's surface.

Biodiversity is measured in several ways, and there are considerable uncertainties in determining the number of living species, let alone the assessment of diversity through geological time. Databases compiled at a variety of geographical, taxonomic, and temporal scales have become powerful tools for assessing Phanerozoic biodiversity change, but there are significant concerns about the biases that might be inherent in the fossil data. The patterns emerging for the marine and terrestrial realms are very different, and there is a lively debate about how the global curves should be interpreted. Nonetheless, they have an important role to play in understanding the patterns and processes of biodiversity change in the modern world and therefore in the development of appropriate conservation strategies.

The Measurement of Biodiversity

Types of Biodiversity

Biodiversity is variously defined in terms of genes, species (or higher taxa), or ecosystems. The genetic and organism components have been combined by many authors and, in many palaeontological studies, a clear distinction is now commonly made between taxonomic diversity and disparity – the degree of morphological variability exhibited by a group of taxa. Disparity can be measured either by phylogenetic distance or by various phenetic indices. Within the history of a major clade, disparity may reach a peak before the peak in species richness is attained.

Setting aside disparity, two basic categories of biodiversity measurement have become widely used: inventory diversity, which records the number of taxa per unit area, and differentiation diversity, which provides a measure of the difference (or similarity) between levels of inventory diversity. Alpha (or within-habitat) diversity is the most common form of inventory diversity and records the number of taxa per area of homogenous habitat, thus reflecting species packing within a community. At its simplest, alpha diversity is species richness (i.e. the number of species in one area), but, especially in the study of modern diversity, its measurement also includes some assessment of abundance. Beta diversity, a differentiation metric, measures the variation in taxonomic composition between areas of alpha diversity. For larger areas, in many studies of terrestrial environments, the term gamma diversity has been used to reflect the number of taxa on an island or in a distinctive landscape, and epsilon diversity has been used for the inventory diversity of a large biogeographical region; the term delta diversity is occasionally used for the variation between areas of gamma diversity within such a region. However, in palaeontological analyses of marine faunas, many workers have used the term gamma diversity for differentiation diversity measuring taxonomic differentiation between geographical regions and therefore reflecting provinciality or the degree of endemicity.

Assessment of the numbers of higher taxa has been an important facet of the analysis of biodiversity – one point sample or community may contain more species than another while having fewer higher taxa and possibly, therefore, a lower genetic diversity and morphological disparity. Moreover, higher taxa are widely used as surrogates for species in biodiversity analyses, especially but not exclusively in the fossil record, where this approach reduces the effects of preservational biases, uncertainties in species identification, and the worst excesses of over-divisive species-level taxonomy. It also facilitates the analysis of biodiversity in biotas within which species-level identifications have not been undertaken because of lack of time or expertise. However, there is a need for caution: the correlation between number of species and number of higher taxa is at best a crude one and diminishes with increasing taxonomic rank. As a result, there is an increasing discordance between the diversity curves calculated from

progressively higher taxa and those calculated from species.

Modern Biodiversity

About 1.8 million modern species have been formally described, but there are enormous gaps in our knowledge of the true number of living species, and the data are patchy in terms of both taxonomic group and geographical coverage. Estimates of modern biodiversity vary considerably depending on the methods used and range from 3.5 to 111.5 million species, with about 13.5 million being considered a reasonable working figure. Less than 15% of the described living species are marine, but over 90% of all classes and virtually all phyla are represented in that environment, with two-thirds of all phyla known only from there. However, conservative estimates of the actual numbers of species of plants and animals (excluding micro-organisms) suggest that marine species may comprise as little as 4.1% of the total.

There is a crude inverse relationship between body size and species diversity, and the problems of quantifying prokaryotic diversity are the most acute. The development of molecular methods over the last few decades has revealed levels of species diversity that are orders of magnitude greater than those recorded previously using culture techniques and has more than trebled the number of major prokaryote divisions now recognized. Though still fraught with uncertainties, not least the definition of what constitutes a species, the estimation of prokaryote diversities from species-abundance curves holds considerable potential; recent calculations suggest that the oceans may have a total bacterial diversity of less than 2×10^6 species, whereas a ton of soil could contain twice this figure.

Ancient Biodiversity

The quantification of ancient biodiversity has a long history, with the 1860 compilation by John Phillips of the changing number of marine species through geological time being widely regarded as including the first published biodiversity curve. This was based on fossil faunas from the British Isles and was calibrated to take account of the total thickness of each stratigraphical interval. Its overall shape is remarkably similar to many of the biodiversity curves published over the last four decades based on global datasets. Since the 1960s, there has been a lively debate surrounding the compilation and interpretation of data at a range of taxonomic levels to show the changing patterns of biodiversity through geological time. This scientific endeavour has been greatly facilitated by the development of increasingly sophisticated large computer databases. These vary greatly in both the scope and refinement of their geographical, temporal, and taxonomic coverage and in the origins of the data included in them. The data range from first-hand sample data through assessments of the primary literature to secondary compilations. Similarly there is considerable variation in the metrics used to plot and analyse the data.

Global marine diversity curves for the Phanerozoic compiled at the family level from different databases have proved to be remarkably similar. Moreover, the most widely cited family- and genus-level curves, developed by JJ Sepkoski (e.g., Figure 1), have proved to be robust when recalculated to take account of substantial revisions of the underlying database in terms of the taxonomy and stratigraphy of many of the entries. This reflects the random distribution of such errors and inaccuracies and is a common phenomenon when analysing large palaeontological databases. There has been some debate over the likelihood of an artificial steepening of the Cenozoic

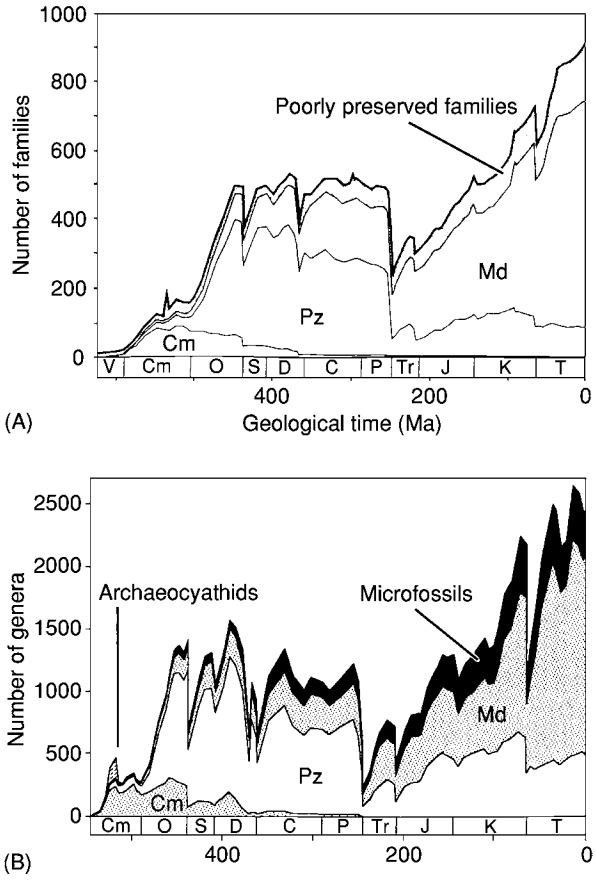


Figure 1 (A) Family level and (B) genus level marine diversity curves, showing the changing proportions of the Cambrian (Cm), Palaeozoic (Pz), and Modern (Md) evolutionary faunas through the Phanerozoic. (Adapted from Sepkoski JJ (1997) Biodiversity: past, present, and future. *Journal of Paleontology* 71: 533 539.)

parts of the diversity curves by the 'pull of the recent' – the possibility that the relatively well-sampled living fauna may extend the stratigraphical ranges of those taxa that also occur in the fossil record, whereas more ancient fossil taxa without modern representatives are less likely to have such complete range records. However, detailed analysis of those modern bivalve genera that have a fossil record has revealed that 95% occur in the Pliocene or Pleistocene, indicating that, for this major group at least, the modern occurrences have only a very small influence on the shape of the Cenozoic diversity curve.

Setting aside the 'pull of the recent', there have been increasing concerns that the global diversity curves may be strongly influenced by taxonomic practice, the temporal sampling pattern used, major geographical gaps in sampling, the absence of information on abundance, and, perhaps most significantly, major biases in the rock record from which the fossil data are derived.

Biases in the marine rock record principally reflect changes in sea-level and can be understood in terms of sequence stratigraphy (*see* **Sequence Stratigraphy**). They result both from variations in the total outcrop area of rock available for study and from the periodic restriction of the available record to particular environments. Thus, for example, in the post-Palaeozoic succession in western Europe, patterns of standing diversity and (apparent) origination and extinction on time-scales of tens of millions of years show a strong correlation with surface outcrop area. Perhaps most tellingly, the 'mass extinction' peaks are strongly correlated with second-order depositional cycles, coinciding with either sequence bases or maximum flooding surfaces. In the former case, the amount of fossiliferous marine rock is very limited (and biased towards shallow-water facies), whereas, in the latter case, there is a strong bias towards deep-water environments and hence a taphonomic bias against shallower-water faunas being available for sampling. In contrast, evidence from fossil tetrapods indicates that, despite earlier assumptions to the contrary, there is no correlation between sea-level change and the quality of the continental fossil record.

Biodiversity Change

Precambrian Biodiversity

From its appearance in the Archaean, life had a profound effect on the chemistry of the oceans and atmosphere during much of Precambrian time, but the overall pattern of diversity change during most of this interval is still weakly constrained. Given the difficulties of assessing diversity in living bacteria and the importance of molecular techniques therein, any understanding of their ancient diversity as represented by fossil material is severely limited. In addition, there is still debate about links between organism diversity and disparity of stromatolites and even about the biotic origin of some such structures (*see* **Biosediments and Biofilms**). The development of eukaryotic cells in the Early Proterozoic marked a major change in organism complexity, and the widely distributed fossil record of protists suggests that their diversity (probably, more strictly, disparity) remained low until about 1000 Ma, the age of the Mesoproterozoic–Neoproterozoic boundary, after which there were major increases in diversity and turnover rates. A trough in diversity during the interval of the Late Neoproterozoic that was characterized by major glacial events was immediately followed by another peak. This was a short-lived event, which overlapped only slightly with the flourishing of the enigmatic soft-bodied Ediacara biota. The dip in protist diversity during the latest Precambrian was followed by a major diversification that was coincident with that of marine invertebrates.

The Ediacara biota flourished for about 30 Ma and represents the first unequivocal record of metazoans in the body-fossil record. The oldest fossil representatives of this soft-bodied biota are recorded from a horizon in Newfoundland that is about 10–15 Ma younger than diamictites of the last of the Neoproterozoic glaciations, which have been dated at about 595 Ma, but the size (up to about 2 m) and complexity of these organisms indicate an earlier history. Whether this indicates an origin prior to the latest glacial event or very rapid evolution thereafter remains uncertain. Either way, the Late Neoproterozoic glaciations (the so-called 'Snowball Earth') probably had a profound effect on the early evolution of metazoans.

A few weakly mineralized metazoans are known from the latest Neoproterozoic, and a large fully mineralized metazoan, *Namapoikia*, of probable cnidarian or poriferan affinities has been described from rocks dated at about 549 Ma. Molecular data suggest that most benthic metazoan groups had a Proterozoic history prior to their acquisition of mineralized skeletons and their appearance in the fossil record during the Cambrian 'explosion'. Some of this earlier history may be indicated by trace fossils, but many trace fossils below the uppermost Neoproterozoic are controversial. Recent estimates using molecular data suggest that the metazoans originated at some time between 1000 Ma and 700 Ma, but the origins and early history of metazoan groups and hence their diversity during the Proterozoic remain unknown.

Phanerozoic Diversity Change

Marine biodiversity change The patterns of marine biodiversity change at the family level during the Phanerozoic emerging from the current databases consistently show a rise from a handful of taxa at the beginning of the Cambrian to some 1900 families at the present day (e.g. Figure 1A). This rise included steep increases in the Early to mid-Cambrian, the Early and mid-Ordovician, and from the Triassic to the Holocene. The diversity level attained in the Ordovician established a plateau, which was essentially maintained until the end Permian mass extinction event (*see* **Palaeozoic:** End Permian Extinctions). Other mass extinctions in the latest Ordovician and Late Devonian and at the ends of the Triassic and Cretaceous periods (*see* **Mesozoic:** End Cretaceous Extinctions) punctuated the overall pattern. The genus-level curve (Figure 1B) shows a similar overall pattern but, unsurprisingly, is much more saw-toothed, and major extinction events (not just the 'big five' listed above) had a much more profound effect at lower taxonomic levels.

The overall diversity curves have been resolved into three so-called evolutionary faunas (Figure 1), each characterized by the dominance of a set of major clades and together representing an increase in ecological complexity through time (Table 1). The Cambrian Fauna dominated its eponymous period but declined thereafter. The Palaeozoic Fauna had its origins in the Cambrian but rose during the Ordovician biodiversification event to a dominance that was largely maintained for the rest of the Palaeozoic. The Palaeozoic Fauna on the low-latitude Laurentian plate showed an onshore–offshore pattern of innovation and expansion of communities during the Ordovician. The Modern Fauna has been an important component of the marine biota, initially in nearshore habitats, since the Ordovician but did not become dominant until after the end Permian extinction event.

Mass extinctions produced the major troughs in the diversity curves but represent one extreme of a continuum of rates of species extinction per million years. The end-Palaeozoic, end-Mesozoic, and, to a lesser extent, Late Devonian mass extinctions did not simply reduce overall biodiversity; they disrupted the ecological patterns so severely that major changes in overall community and even ecosystem structure could occur. However, whilst some significant changes in the proportions of marine taxa grouped by fundamental differences in autecology and physiology took place after mass extinctions, many of the episodes of change or of relative stasis of such features did not mirror the concurrent trajectory in global diversity. Thus, for example, while there were stepped increases in the proportion of predator taxa following the end-Permian and end-Cretaceous extinction events, this proportion then fluctuated only within a fairly narrow band during the subsequent rises in overall diversity.

Non-marine biodiversity change The shape of the non-marine biodiversity curve is markedly different from that of the marine biodiversity curve (Figure 2), and its principal components are the plants, insects, and tetrapod vertebrates.

Whilst the land probably had a microbial flora extending back into the Precambrian, spores and phytodebris indicate that land plants were present

Table 1 The major components and ecological structures of the three marine evolutionary faunas

Evolutionary fauna	*Community diversity*	*Food webs*	*Tiering*	*Suspension feeders*	*Detritivores and carnivores*	*Planktonic food*	*Animals in water column*
Modern evolutionary fauna (demosponges, gastropods, bivalves, gymnolaemate bryozoans, malacostracans, echinoids, vertebrates)	High	Highly complex	Complex	Epifaunal and infaunal	Common	Abundant	Many
Palaeozoic evolutionary fauna (anthozoans, cephalopods, stenolaemate bryozoans, 'articulated' brachiopods, ostracods, stelleroids, crinoids, graptolites)	Intermediate	Intermediate	Develops	Epifaunal common	Common	Common	Many
Cambrian evolutionary fauna (hyolithids, monoplacophorans, 'inarticulate' brachiopods, trilobites, eocrinoids)	Low	Simple	Limited	Few	Dominant	Limited	Few

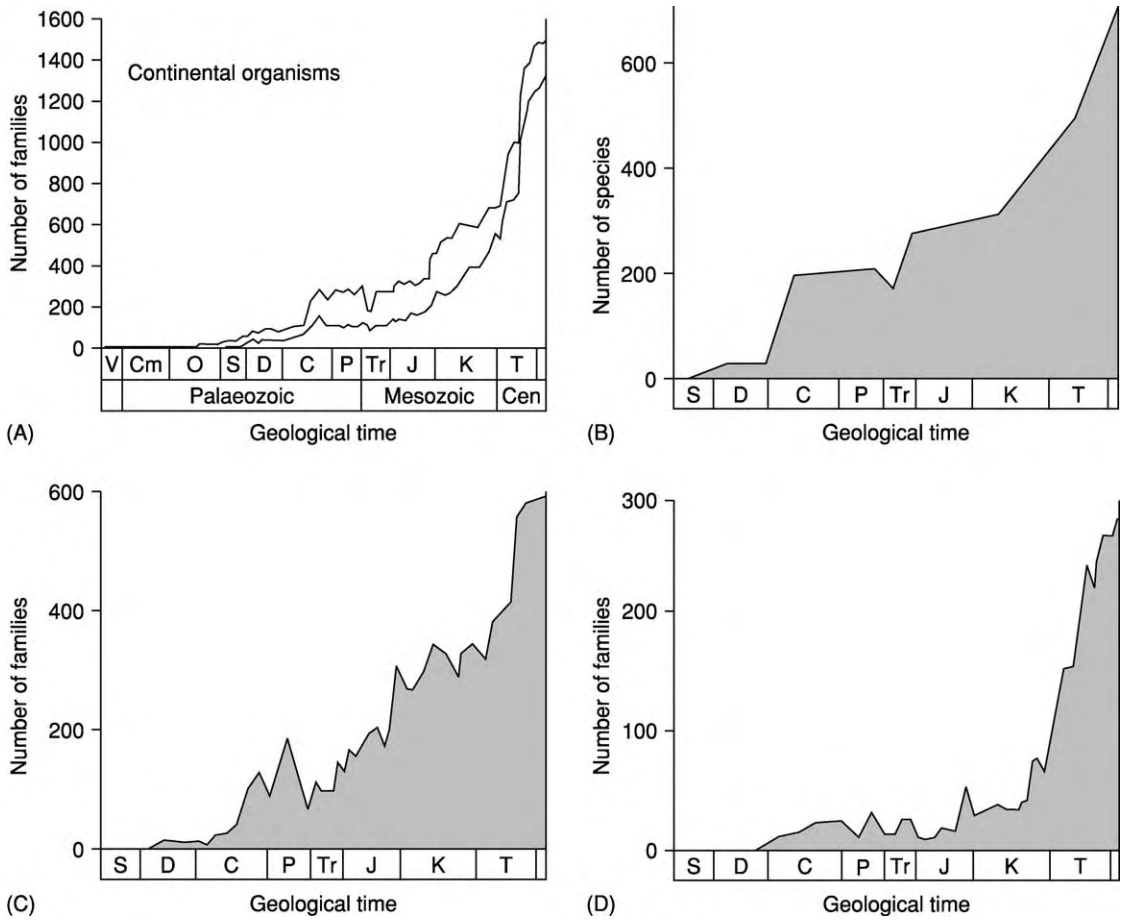


Figure 2 Terrestrial biodiversity through the Phanerozoic. (A) All multicellular families (lower curve represents families undoubtedly present within each stage and restricted to the non marine environment; upper curve includes less stratigraphically well attributed families and those that include environmentally equivocal that include marine or environmentally equivocal taxa); (B) vascular land plants; (C) insects; and (D) non marine tetrapods. (Adapted from Benton MJ (2001) Biodiversity through time. In: Briggs DEG and Crowther P (eds.) (2001) *Palaeobiology II*, pp. 211 220. Oxford: Blackwell Publishing.)

from the mid-Ordovician onwards. These earliest plants were probably bryophytes; spore evidence suggests that vascular plants or their immediate ancestors appeared in the Early Silurian. The first unequivocal land-plant megafossils are known from the mid-Silurian (Wenlock). Following the Silurian and Devonian development and radiation of early vascular-plant groups, the plant biodiversity curve (Figure 2B) reflects the successive acmes of the pteridophytes and early gymnosperms in the Carboniferous–Permian and the gymnosperms (*see* **Fossil Plants:** Gymnosperms) in the Triassic–Jurassic and the rise of the angiosperms from the Cretaceous onwards.

The diversity curve for the insects (Figure 2C) shows a rather saw-toothed exponential rise from the first appearance of the group in the Devonian. The tetrapods (Figures 2D and 3A) also show an essentially exponential pattern, with a long (Devonian to mid-Cretaceous) period of weak diversity increase followed by a steep rise. The curve primarily reflects the radiations and acmes of three major sets of groups: the basal tetrapods and synapsids in the Palaeozoic, archosaurs in the Mesozoic, and lissamphibians, lepidosaurs, birds, and mammals in the Cenozoic.

The colonization of freshwater and the increase in diversity therein was achieved both by air-breathing groups and by taxa from several major marine clades that independently became adapted to reduced salinity by movement upstream through estuarine environments. However, the occupation by invertebrates of freshwater habitats, especially those within the substrate, took a remarkably long time, and it was not until the Mesozoic that a significant range of originally marine groups developed the necessary osmoregulatory, reproductive, and dispersal capacities to occupy freshwater environments.

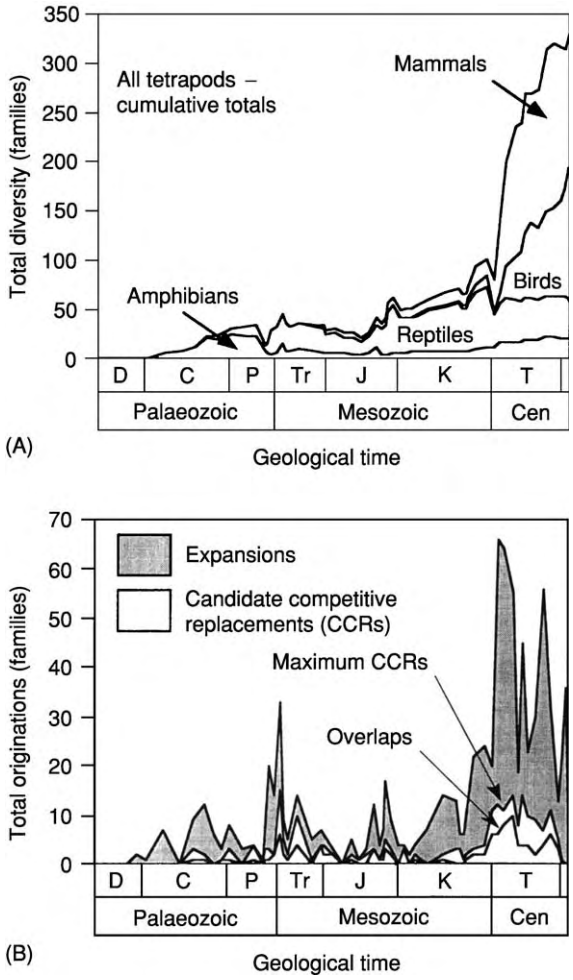


Figure 3 Biodiversity change in 840 non singleton tetrapod families, broken down into (A) the component groups and (B) the styles of origination of new families. Overlap candidate competi tive replacements are those in which the range of an originating family overlapped with that of another family rather than re placing it in the fossil record. (Adapted from Benton MJ (1999) The history of life: large databases in palaeontology. In: Harper DAT (ed.) *Numerical Palaeobiology*, pp. 249 283. Chichester: Wiley.)

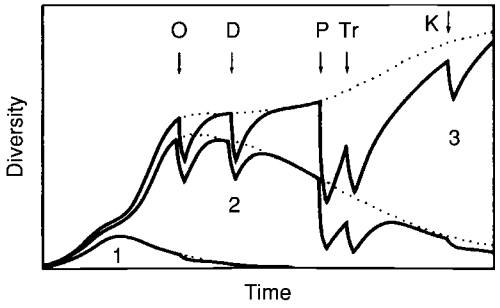


Figure 4 Three phase coupled logistic model of changes in family level marine biodiversity during the Phanerozoic, includ ing perturbations simulating the mass extinctions at the end of the Ordovician (O), in the late Devonian (D), and at the ends of the Permian (P), Triassic (Tr), and Cretaceous (K). The numbers 1, 2 and 3 refer to the Cambrian, Palaeozoic and Modern evolu tionary faunas shown on **Figure 1**. Dotted lines represent the trajectories of the unperturbed three phase system. (Adapted from Sepkoski JJ (1984) A kinetic model of Phanerozoic taxo nomic diversity. III. Post Paleozoic families and mass extinctions. *Paleobiology* 10: 246 267.)

Understanding Biodiversity Curves

If the existing curves are even an approximation of the true pattern, the great challenge is to interpret them in terms of the processes leading to biodiversity change through geological time. At the simplest level, the curves reflect the complex interplay of species originations, extinctions, and stasis, with the major changes in trajectory representing episodes of radiation punctuated by mass extinctions. Although the traditional assumption has been that radiations were 'adaptive', representing the acquisition of new characters by the radiating group that gave them superiority over their competitors, there is increasing evidence for the expansion of many clades into previously unoccupied ecospace. This is perhaps most strikingly illustrated by the tetrapods, for which it has been calculated that a maximum of only 13% of families could have originated by competitive replacement of earlier taxa (**Figure 3**). It has also been argued that clade radiations could arise randomly, without being driven by any deterministic cause such as competition or expansion associated with a new adaptation or opportunity.

The shapes of biodiversity curves Several attempts have been made to model mathematically the shapes of global Phanerozoic diversity curves and therefore to determine whether they reflect overarching global processes and patterns of evolution.

Continental biodiversity, either as a whole or of individual large clades (**Figure 2**), seems to show an exponential rise without a sustained levelling off. In contrast, the family-level marine curve has been modelled fairly closely in terms of logistic curves representing the Cambrian, Palaeozoic, and Modern Faunas perturbed by the major mass extinction events (**Figure 4**). In this coupled logistic model, each evolutionary fauna showed a slow initial increase in diversity, followed by a steep increase to a plateau and then a slow decline corresponding to the rise of the following evolutionary fauna (which had a lower initial rate of diversification and a higher maximum diversity level). Within such a model, the trajectory of the curve for the Modern Evolutionary Fauna is currently still in its exponential phase, but its roughly convex shape suggests that an asymptotic stage could be projected in the future. The effects of the mass extinction

events enhance the fit of the theoretical curves to the measured global patterns.

By extending the theory of island biogeography developed in the 1960s to the global scale, the logistic curves applied to marine familial diversity have been interpreted as reflecting evolution into new habitats and empty niches until a level of dynamic equilibrium is reached at the carrying capacity of the environment. This theory has received widespread, but not universal, acceptance. It has been argued, for example, that the shapes of the curves may represent not global-scale biotic interactions but rather the total effect of physical perturbations operating at a wide range of geographical and temporal scales. Moreover, the evidence for equilibrium, even on an ecological rather than a geological time-scale, and the upper limit on diversity that it would impose have also been widely contested. The Palaeozoic diversity plateau, taken to be strong evidence for equilibrium, may have been maintained by factors other than biotic interactions and may even be an artefact of the taxonomic level of the data. Importantly, it is clear that different clades or parts thereof may show very different patterns of biodiversity change.

The causes of biodiversity change Despite considerable research effort, there has been only limited success in identifying 'rules' governing species and community diversity and the relative abundances of species within communities. It is highly unlikely, even for mass extinction events, that a single factor can be invoked to explain global changes in biodiversity. In addition to factors intrinsic to individual clades, a host of interrelated extrinsic factors undoubtedly influence the evolution, distribution, and diversity of organisms across the spectrum of spatial and temporal scales. Crucially, the factors operating at one scale may be very different from those operating at another. Intriguingly though, there is some evidence to suggest that the family-level global diversity curve has properties of self-organized criticality, and, thus, Phanerozoic diversity change may be driven by the internal dynamics of life itself, as well as responding to external factors.

The present-day concerns over biodiversity change stem from the recognition of the deleterious consequences of anthropogenic activities, including habitat destruction, pollution, and influences on global climate. The fossil record provides a time perspective not only on the patterns of biodiversity change but also on the natural physical factors that drive it, from local fluctuations in environmental conditions to plate tectonics, eustasy, ocean circulation patterns, and climate changes. Heightened awareness of the quality and detail of the data involved in the generation and analysis of diversity curves at all spatial and temporal scales should result in greater confidence in the conclusions drawn from them.

See Also

Biological Radiations and Speciation. **Biosediments and Biofilms**. **Evolution**. **Fossil Plants:** Gymnosperms. **Mesozoic:** End Cretaceous Extinctions. **Palaeoecology**. **Palaeozoic:** End Permian Extinctions. **Sequence Stratigraphy**.

Further Reading

Bambach RK, Knoll AH, and Sepkoski JJ (2002) Anatom ical and ecological constraints on Phanerozoic animal diversity in the marine realm. *Proceedings of the Na tional Academy of Sciences USA* 99: 6854 6859.

Benton MJ (1999) The history of life: large databases in palaeontology. In: Harper DAT (ed.) *Numerical Palaeo biology,* pp. 249 283. Chichester: Wiley.

Benton MJ (2001) Biodiversity on land and in the sea. *Geological Journal* 36: 211 230.

Briggs DEG and Crowther P (eds.) (2001) *Palaeobiology II*. Oxford: Blackwell Publishing.

Crame JA and Owen AW (eds.) (2002) *Palaeobiogeography and Biodiversity Change: The Ordovician and Mesozoic Cenozoic Radiations*. Special Publication 194. London: Geological Society.

Gaston KJ and Spicer JI (1998) *Biodiversity: An Introduc tion*. Oxford: Blackwell Science.

Hewzulla D, Boulter MC, Benton MJ, and Halley JM (1999) Evolutionary patterns from mass originations and mass extinctions. *Philosophical Transactions of the Royal Society of London Series B* 354: 463 469.

Knoll AH (1994) Proterozoic and early Cambrian protists: evidence for accelerating evolutionary tempo. *Pro ceedings of the National Academy of Sciences USA* 91: 6743 6750.

Levin SA (ed.) (2001) *Encyclopedia of Biodiversity*, 5 vols. San Diego: Academic Press.

May RM (1992) How many species inhabit the Earth? *Scientific American* October 1992: 18 24.

Miller AI (2000) Conversations about Phanerozoic global diversity. *Paleobiology* 26(4 Suppl.): 53 73.

Sepkoski JJ (1984) A kinetic model of Phanerozoic taxo nomic diversity. III. Post Paleozoic families and mass extinctions. *Paleobiology* 10: 246 267.

Sepkoski JJ (1997) Biodiversity: past, present, and future. *Journal of Paleontology* 71: 533 539.

Smith AB (2001) Large scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. *Philosophical Transactions of the Royal Society of London, Series B* 356: 351 367.

Ward BB (2002) How many species of prokaryotes are there? *Proceedings of the National Academy of Sciences USA* 99: 10234 10236.

BIOLOGICAL RADIATIONS AND SPECIATION

P L Forey, The Natural History Museum, London, UK

Introduction

This entry discusses the evidence for speciation and the patterns of species multiplication and associated morphological change as revealed in the fossil record. Speciation is regarded as the key stage in the generation of biological diversity as it is the point at which reproductive isolation is achieved – the hallmark of the biological species (*see* **Evolution**). Many studies of modern species concern themselves with the short-term genetic fluctuations between populations of a single species, how those fluctuations may come about and how they are maintained. Such studies are not possible in the fossil record. On the other hand, the fossil record documents extended timescales and may, therefore, identify patterns of long-term changes relevant to different theories of speciation that are unavailable in the Recent world.

Irrespective of the mode of speciation, the fossil record demonstrates repeated instances, geologically speaking, of sudden increases in the numbers of species and these are usually associated with rapid morphological divergence. Such events are commonly referred to as radiations, more usually adaptive radiations, and are inferred to have come about by causes intrinsic to the organisms and/or extrinsic causes. Although studies of radiations involving populations and closely related species are studied in the Recent world, broad-scale patterns can only be studied through the time dimension supplied by the fossil record.

Species, Species Recognition and Speciation in the Fossil Record

There is a plethora of species definitions and criteria for recognizing species in the Recent world, but few are applicable to the fossil record where only a limited amount of morphological data, usually the skeleton, is available for study. Three such definitions that rely on some underlying process of speciation are shown in Figure 1. The biological species concept of Ernst Mayr (Figure 1A) is probably that most favoured by people studying the Recent world. Species status is achieved when populations diverge genetically to such an extent that reproductive cohesion is broken. There may be several reasons for fragmentation of reproductive continuity (geography, ecology, and behaviour are most commonly cited), although Mayr stressed geographical isolation. The parental species may continue to live alongside the daughter species. In practice, most modern species are not recognized on reproductive criteria: instead some measure of morphological, genetic, behavioural or ecological difference is used as a surrogate for reproductive incompatibility. As such, palaeontological species, recognized almost exclusively on morphological differences, are equally as valid as modern species. In fact, there has been empirical justification for equating genetic differences with morphological differences between modern species of some groups such as bryozoans and Darwin's finches. It is, however, recognized that morphological difference may not always indicate species differences (e.g., sibling species or polymorphic species including mimetic species).

The evolutionary concept of George Gaylord Simpson (Figure 1B) defines a species as an ancestor-descendent sequence of populations changing through time with their own trends and tendencies.

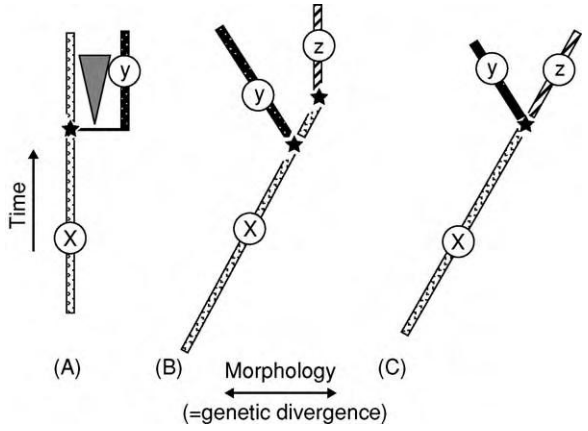


Figure 1 Three concepts of species. (A) Mayr's biological species. Populations reach species distinction when reproduct ive cohesion is broken, usually by the imposition of some isolat ing mechanism such as geographic separation of populations (triangle). The original species (X) may or may not continue to live contemporaneously with the daughter species (Y). (B) Simp son's evolutionary species concept. Species are ancestor descendent sequences of populations changing through time. A species may change gradually into another (X → Z) through a process of anagenesis or the lineage may split (X → Y) through cladogenesis. (C) The Hennigian species concept recognizes species as lineage segments. As soon as a cladogenetic event occurs there are automatically two new species (Y and Z), regardless of any perceived morphological change. Stars represent speciation events.

This definition embraces asexual species. Such a definition allows for one species to change gradually into another (anagenesis) as well as instances where a species is budded off or where one species splits into two or more (cladogenesis).

The Hennigian species concept (Figure 1C) (named after the German entomologist Willi Hennig), like the biological species concept, is based on the criterion of reproductive isolation. However, the species limits are recognized only at the point where one species splits to two or more and the interbreeding pattern of gene flow is disrupted, at which point the ancestral species cannot by definition live alongside the descendants. This species definition is strictly dependent upon the shape of the phylogenetic tree.

Recognition of Species

Species in the fossil record are nearly always recognized on morphological criteria. There are some exceptions, such as instances where species are distinguished from one another purely on the basis of stratigraphic occurrence or geographic location, but these are usually accepted as stopgap measures; to be revised when more information about morphology becomes available.

Monophyletic species are recognized on the basis of possessing one or more unique morphological characters and this corresponds most closely with the Hennigian species concept. Species viewed in this light have the same ontological status as higher taxa such as genera, families, orders, etc. In reality, very few species can be recognized in this way and logically it denies the existence of ancestors since ancestors have no unique features.

Typological species were once commonly recognized, whereby a single specimen (usually the first discovered or the most complete) was chosen as the Linnean type and this formed the nucleus of morphological variation considered worthy of species status. How much variation was to be allowed before a new species was recognized was never stated. Such species are usually frowned upon now as being steeped in the philosophy of pre-Darwinian essentialism. Yet, there remain many instances where such species are still used.

Phenetic species arose out of numerical taxonomy. Here species are measured for as many morphologically continuous variables as possible. The variables are then subjected to multivariate analysis, and clusters – (the species) – are then recognized. Although this may appear to be objective and theory free, it is heavily dependent upon the original samples analyzed and the multivariate analysis algorithms used. It is not always easy to identify which of the individual variables is contributing to the differences and, therefore, it is difficult to diagnose the species.

Phylogenetic species are the smallest diagnosable assemblages of specimens showing a unique combination of characters. Combination is the operative word here since there is no requirement for identifying unique characters (cf. monophyletic species recognition). Of all the recognition criteria this most faithfully agrees with all of the process-based definitions of species, as well as agreeing with the day-to-day practice of palaeontologists.

Estimates of species longevity may depend on which of the above ways the species have been recognized. Surveys across various fossil groups give average figures that range from 1 million years (for small mammals where generation time is very short) to 13 million years for dinoflagellates (which are largely asexual).

Speciation in the Fossil Record

It is generally accepted that most speciation events, with the possible exception of those caused by hybridization and polyploidy (chromosomal duplication), are too protracted to be detected in the Recent world and too brief to be seen in the fossil record. The lower limits of time that can be successively sampled in the fossil record appear to be in the order of 5000–10 000 years and such deposits are very rare, restricted to lake deposits of relatively short duration and small geographic range, and some deep-sea deposits. Nevertheless, there have been many studies seeking to establish the precise pattern of speciation.

There are two polarized theories on the tempo and pattern of speciation (Figure 2), punctuated equilibria and phyletic gradualism. Both have support from theoretical and empirical studies in the modern world (see legend to Figure 2) and it is likely that both modes of speciation can happen.

Phyletic gradualism Phyletic gradualism holds that morphological evolution is gradual and occurs independently of speciation. Most of the studies supporting this mode of evolution involve organisms that can be minutely sampled through continuous rock sequences.

Figure 3 shows an example of one gastropod species (*see* **Fossil Invertebrates:** Gastropods) changing or replacing another in an ancestor-descendent sequence (gradual speciation by anagenesis) by a shift in the modal shell form as measured by morphometric discriminant analysis. In this example the gradual change has been correlated with increasing depth of water. The populations intermediate between the two species were presumed to have lasted for 73 000–250 000 years. In this pattern there is no

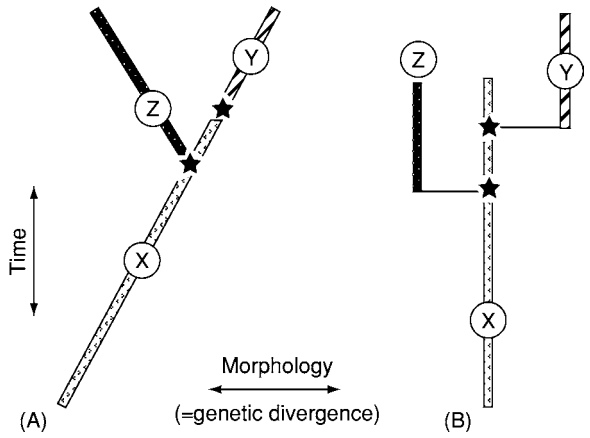


Figure 2 Two theories of speciation modes in the fossil record. (A) Phyletic gradualism suggests that speciation is a by product of changing ratios of gradual morphological variation (implied gene frequency) within populations through time and that speciation may happen through anagenesis (species X → Y) as well as cladogenesis (X → Z). Phyletic gradualism agrees with Simpson's evolutionary species concept (**Figure 1B**), coupled with Fisher's notion of 'directed' selection of gene frequencies within populations. (B) Punctuated equilibrium claims that individual species remain morphologically constant throughout time and that all or most of the morphological (and implied genetic) change is associated with cladogenesis: the speciation event. This theory agrees with Mayr's biological species concept (**Figure 1A**) coupled with Sewell Wright's idea of stabilizing selection. In these diagrams the species are denoted by letters and different shadings and the speciation events by stars.

rapid shift in morphology and, in studies of this kind, the species limits are usually recognized by comparing the morphological variation over a period of time with morphological variation of modern species and dividing up the continuum accordingly.

In another study (Figure 4) involving radiolarians recovered from Pacific Ocean deep-sea drilling cores, an instance of speciation was detected occurring during a period of about 500 000 years. In this case both the ancestral and daughter species showed marked deviation in many aspects of skeleton form. Further, it is interesting to note that morphological deviation, although slightly accelerated during the speciation event, continued for some time after the separation of morphotypes, suggesting strongly that morphological evolution is decoupled from speciation. In this instance both parental and daughter species showed morphological evolution and it may have been possible to suggest that a single ancestral species gave rise to two daughter species. The author of this study refrained from this because the ancestral species, which still lives today, shows a high degree of polymorphism and geographic variation, the scale of which is comparable to the variation seen in the lineage in this study.

It needs to be pointed out that many (but not all) of the studies that support phyletic gradualism, report information on continuously varying morphologies such as lengths and shapes. These parameters are inherently gradualist in their measurement, unlike many characters used in cladistic analysis such as presence/absence or four toes versus five toes in which there can be no intermediates.

Punctuated equilibrium Speciation following the punctuated equilibrium model (Figure 2B) has been claimed for many groups of organisms. This theory suggests that species themselves exhibit morphological stasis and that speciation is, geologically speaking, instantaneous and is accompanied by rapid morphological shifts. In order to demonstrate this at least three objections raised by advocates of phyletic gradualism must be overcome. The first objection is that what appear to be sudden changes only seem so because either the sampling intervals are too wide or the fossil record is lacking in the key years where intermediate populations might have lived. Therefore, it is necessary to demonstrate that the sampling is dense. A second objection is that it is possible that an isolated population, destined to become a new species, was accumulating genetic and morphological distinctiveness in a gradualist manner, but that the population was either so small that fossilization potential was virtually zero, or that it was living elsewhere. The sudden appearance is, therefore, explained not as any intrinsic property of speciation, but by the new species population growing large enough to leave a fossil record, or by immigration into the area of study. The third objection is a taxonomic argument. Many of the claims of punctuated equilibrium involve species recognized by qualitative characters where there are no intermediate states (see above). If this is true then the sudden appearance of morphologic changes must coincide with speciation and must also remain static until the next 'speciation'.

Despite the objections there are cases where punctuated equilibrium can be demonstrated. Figure 5 illustrates the case of speciation in the bryozoan genus *Metrarhabdotus* during the Neogene and Quaternary of tropical America (*see* **Fossil Invertebrates:** Bryozoans). The species were distinguished on both continuously varying characters, as well as qualitative differences. A stratophenetic tree was constructed. There was particularly dense sampling of the fossil record in the Dominican Republic between 8 and 4 million years ago, within which time there was the successive origin of 12 species. Each of the species appeared suddenly, accompanied by substantial morphological differences. The species themselves exhibited little or no change throughout their individual

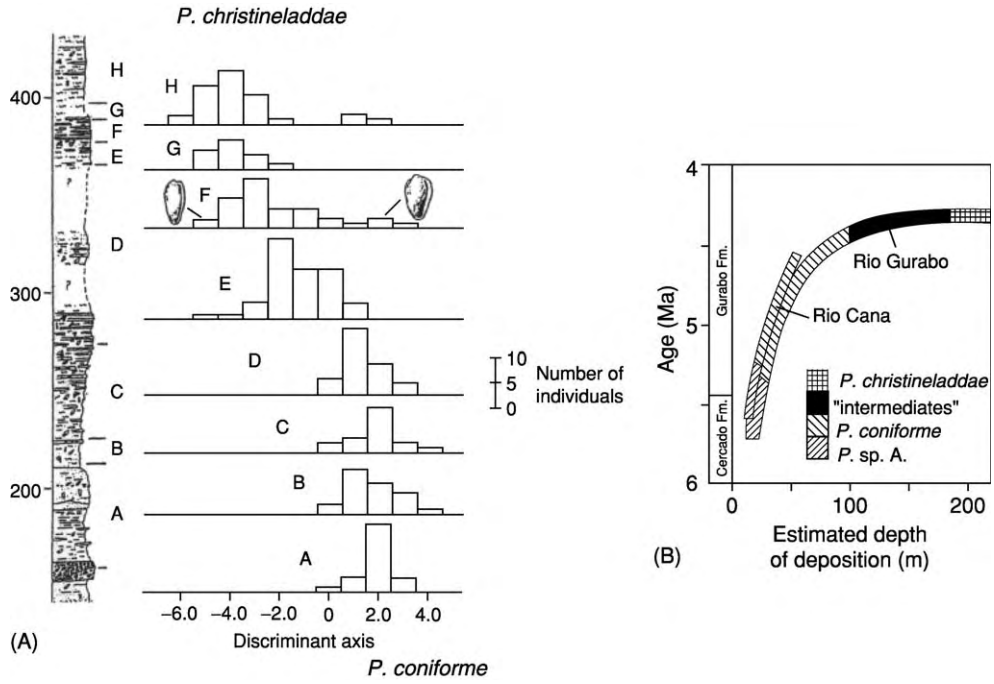


Figure 3 Speciation through anagenesis. In this study a species of marginellid gastropod (*Prunum coniforme*) living in the Mio Pliocene of the Dominican Republic was seen to show strong directional selection in shell morphology towards a new species (*Prunum christineladdae*). (A) Histogram of individuals on the discriminant axis that best distinguishes the two species (endpoints of sampling) collected from successively higher stratigraphic levels within the section at Rio Gurabo, Dominican Republic (section thickness marked in meters at left). Notice that at level F there appears to be a mixed population which appears to have occupied 0.6 2.5% of the entire range of the ancestral species (*P. coniforme*). (B) Over a longer time scale the species are distributed in sediments inferred to have been deposited in increasingly deeper water. Part of the history at Rio Gurabo is repeated at neighbouring Rio Cana. Reproduced from Nehm RH and Geary DH (1994) A gradual morphologic transition during a rapid speciation event in marginellid gastropods (Neogene; Dominican Republic). *Journal of Paleontology* 68: 787 795.

existence and this is a key feature of the punctuated equilibrium model. Although the possibility of immigration cannot be completely dismissed this is unlikely because the basal species of this lineage are morphologically very similar to species occurring earlier in the same area.

Therefore, it appears as though there is evidence for both patterns of speciation in the fossil record. Indeed, there may be instances where both kinds can be recognized within the same sequence, affecting closely related organisms. More commonly, species within the same lineage may show gradualism at some times and punctuated patterns at others (but here there is always the objection of differential sampling). Rather than trying to prove that all speciation complies with one or the other model it may be more productive to try to determine when and why one or the other is predominant.

Radiations

Radiations are episodes of increased rates of speciation relative to extinctions resulting in rapid net increases in diversity. They may be accompanied by evolution of more diverse body forms than in normal periods when extinction and origination rates are approximately equal. Radiations are nearly always explained in terms of adaptation, hence 'adaptive radiations', the latter being defined by Dolph Schluter as the evolution of ecological and morphological diversity within a rapidly speciating lineage. There are difficulties with defining both ecological and phenotypic diversity and it is not always easy to describe precisely the adaptation and even more difficult to isolate the cause, but the following factors have been suggested as triggers for radiations (adaptive or otherwise):

1. Abiotic causes – such as changing levels in O_2/CO_2 concentrations or significant changes in ambient temperature, changes in ocean current circulation, fragmentation of landmasses and consequent changes in both the length and ecological differentiation of coastlines, and changes in latitudinal distributions.
2. Biotic factors – such as the evolution of a particular character or characters that may change aspects of life history leading to rapid speciation and/or ecological differentiation.

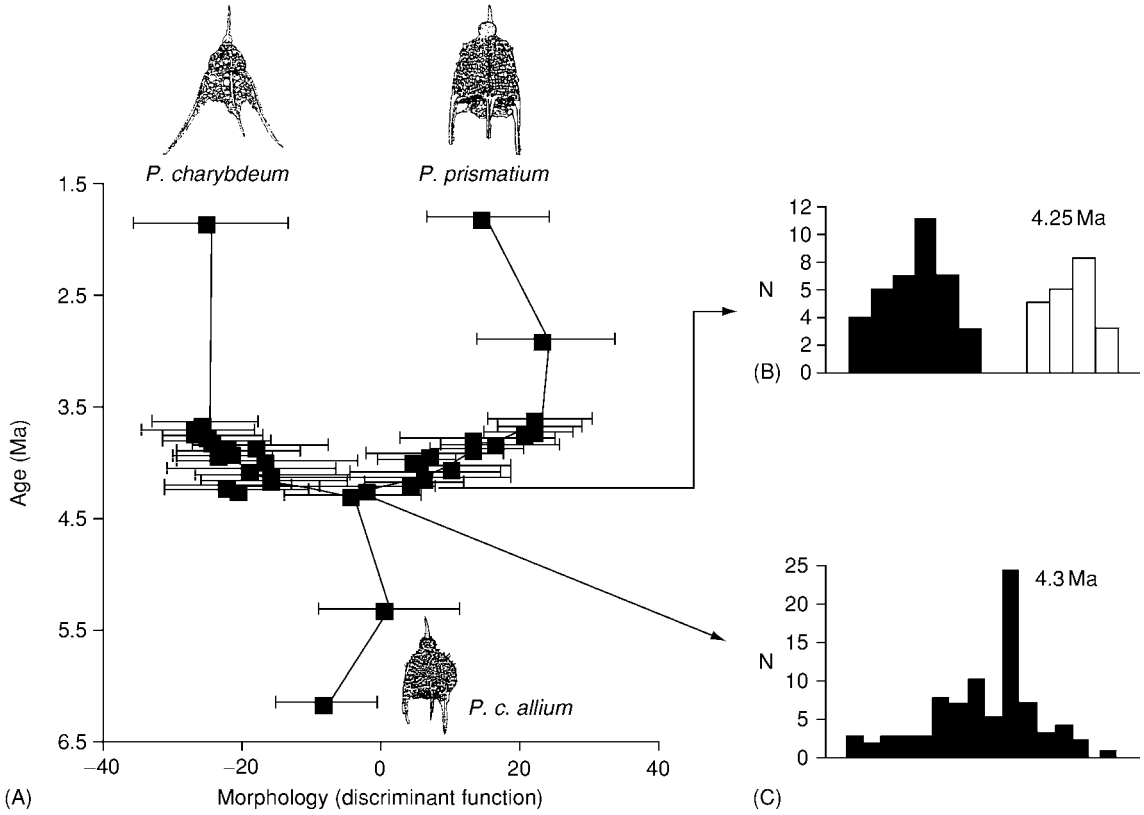


Figure 4 Speciation and the gradualist model. The present day radiolarian species *Pterocanium charybdeum* extends well back into the Miocene where a subspecies (*P. c. allium* is recognized). At about 4 million years ago a new species, *P. prismatium* arose and subsequently went extinct about 1.8 million years ago. To compile this diagram samples representing populations at 50 000 year intervals were taken from piston cores during the key 4.5 3.5 million year interval (much more widely spaced outside of this time band). The radiolarian tests were measured for 32 morphological characters such as length, breadth, angles and outlines and subjected to a multivariate discriminant analysis. The x axis of the diagram shows the value of the discriminant function that best separates the species. The boxes are the population means ± 1 standard error. The histograms on the right show discriminate scores at the point of the intermediate population (lower diagram) and the sampling level 50 000 years later (upper diagram), showing clear separation at this time; these two documenting the speciation event. Notice that both lineages continue to diverge morphologically in a gradualist manner long after the speciation event. Reproduced from Lazarus DB (2001) Speciation and morphological evolution. In: Briggs DE and Crowther PR (eds.) *Palaeobiology II*, pp. 133 137. Oxford: Blackwell Scientific.

3. Interactions between organisms – such as the extinction of one kind allowing ecological and phenotypic diversification in another. Alternatively it may be true that the morphological and diversification of one group allows similar phenomena in another (e.g., insects and plants). For example, in fishes, it has long been noted that the evolution of many durophagous (mollusc crushing) lineages coincides with the Mesozoic diversification of bivalve and gastropod molluscs. Many of these 'associations' are anecdotal and, while they may well be true, they need firm experimental evidence for their justification.

It is rarely possible to isolate a single cause and in all probability complex interactions between two or more factors are ultimately responsible for radiations. Therefore, the study of these evolutionary phenomena is both frustrating and challenging, demanding lengthy and detailed data collection.

Factors which may distort our view of radiations include the following:

1. Imperfections of the fossil record. Large hiatuses or differential preservation will inevitably distort our perception of radiations. For example, the sudden appearance and apparent radiation of many catfishes (most of which are freshwater) in the Early Tertiary must be judged against the knowledge that there are very few freshwater fish-bearing deposits in the underlying Upper Cretaceous.
2. Appearance of many diverse animals and plants associated with Lagerstätten deposits. Such deposits dramatically increase the numbers and morphological diversity of species that may, in reality, have had a long unrecorded history.

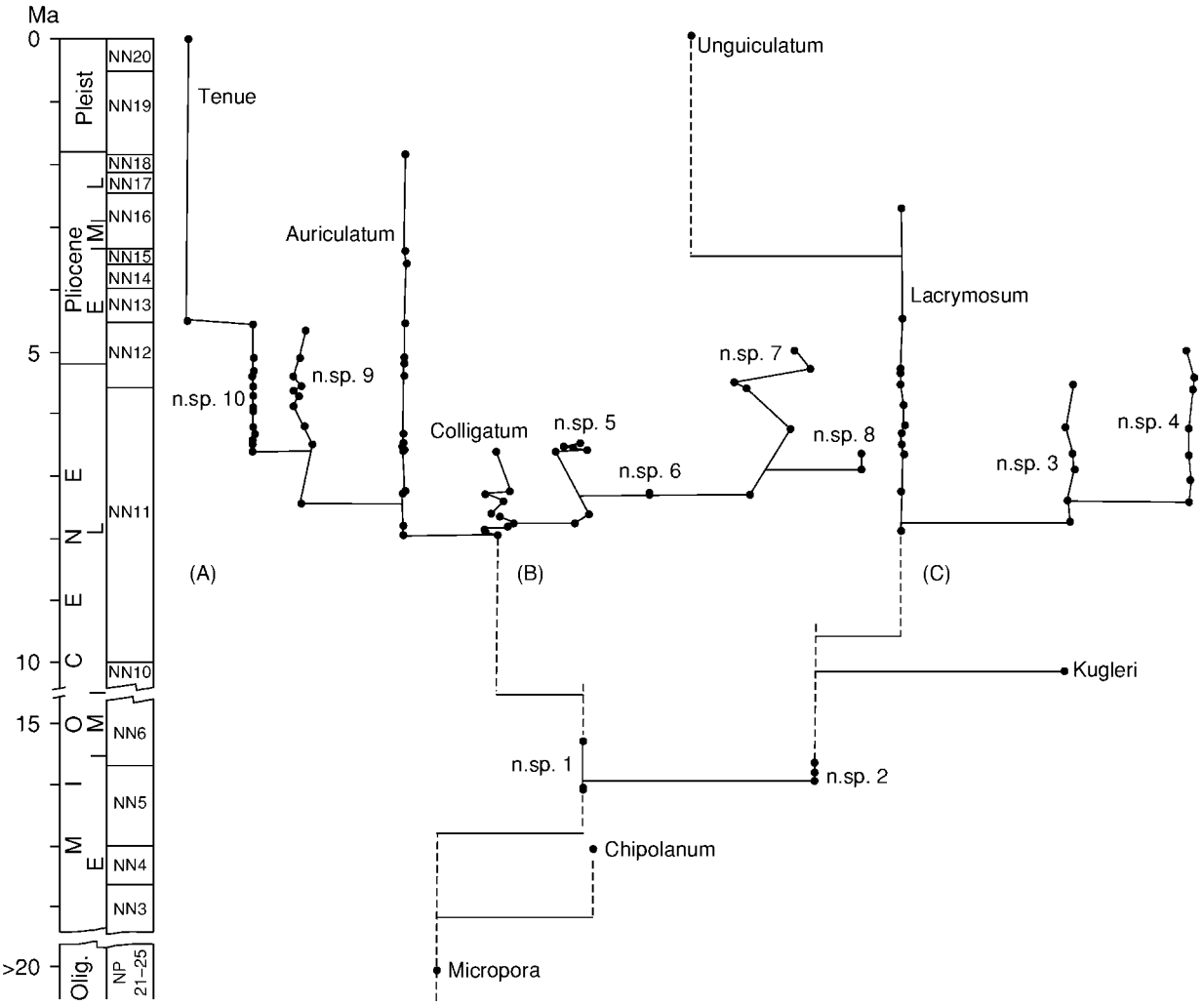


Figure 5 Punctuated equilibrium in Neogene bryozoans. The cheilostome genus *Metrarhabdotos* is widely distributed in Miocene and Pliocene deposits of the Caribbean. Sections in the Dominican Republic are particularly rich and uninterrupted, such that is possible to sample extensively. This study used 46 measurements, counts or codings of colony morphology of the species (some were left un named). These were subjected to multivariate analysis and the stratophenetic tree shown here was produced. The dots are sampling points. The x axis is a measure of the morphologic distance. Notice the sudden shift in morphological variation at points of cladogenesis. The within species variability is very small compared with the between species differences as predicted by the punctuated equilibrium model, as is also the prediction that the ancestral species continues to live alongside the descendant. The age, epoch and nannoplankton zones are shown on the y axis. Reproduced from Cheetham AH (1986) Tempo of evolution in a Neogene bryozoan: rates of morphological change within and across species boundaries. *Paleobiology* 12: 190 202.

3. Taxonomic artifact. Many diversity studies are based on estimates of numbers of genera or families. Dependent on how those taxa are recognized in successive stratigraphic levels, the expansion (or decrease) in diversity may be exaggerated or distorted. For example, it has been pointed out that the apparent sudden increase in new trilobite families in the earliest Ordovician may be an artifact of taxonomy of Cambrian trilobites, since many phylogenetic lines can now be drawn across the boundary linking families that were previously thought to be quite distinct.

Despite these problems there are good examples of association of cause with radiations.

Environmental Shift

One general pattern that has been identified after extensive collection of both taxic and geological data is that many invertebrate clades originated in nearshore marine environments and subsequently expanded in species numbers to occupy deeper water, at the same time becoming morphologically more diverse. Figure 6 shows one example of the

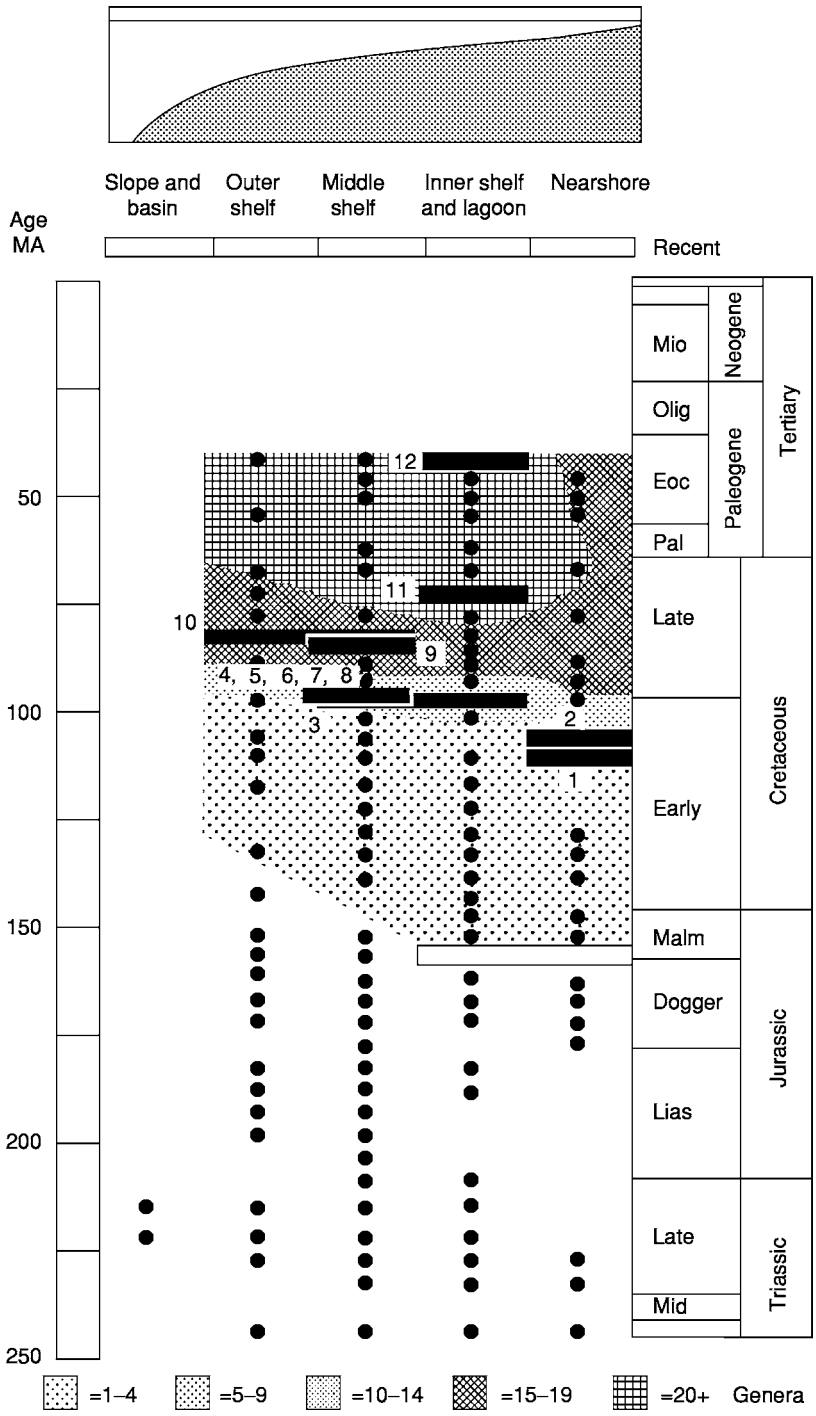


Figure 6 Study exemplifying that the early phases of a history of marine invertebrate clades began in nearshore environments and subsequently expanded to the outer shelf. In this study the environmental history of the cheilostome bryozoan radiation suggests that the group originated in nearshore and inner shelf environments (white box). As the group diversified in both numbers of genera (contours denoted by the different shadings) with the acquisition of evolutionary novelties (numbers associated with black boxes) they spread to middle and outer shelf habitats. The black dots represent localities yielding bryozoans at different horizons and depositional environments. Those dots below the shaded areas represent localities at which only non cheilostomes are found (i.e., cyclostomes and trepostomes) and these impose a taphonomic control documenting that the cheilostome pattern is not an artifact of preservation. The dots within the shaded area represent localities at which all kinds of bryozoan are found. Reproduced from Jablonski D, Lidgard S and Taylor PD (1997) Comparative ecology of bryozoan radiations: origin of novelties in cyclostomes and cheilostomes. *Palaios* 12: 505 523. SEPM (Society for Sedimentary Geology).

evolution of cheilostome bryozoans in the Cretaceous (*see* **Fossil Invertebrates:** Bryozoans). It needs to be emphasized that this is a general pattern; there is no implication that the ecological shift was the cause of speciation.

Abiotic Causes – Fragmentation of Areas

Radiations coincident with fragmentation of continental masses can be documented for a variety of organisms, but usually only over long timescales. One example, is the increased diversity of families of terrestrial amniotes between the Lower Jurassic and Paleocene following the breakup of Pangaea (Figure 7). It is very difficult to demonstrate a direct cause and effect, but there is a considerable shift in the latitudinal distribution of the continental masses such that substantial diversification in ecological conditions occurred leading to a tempting association with cladogenesis. The breakup of Pangaea also resulted in a considerable increase in available coastline,

	First	*1 (%)*	*2 (%)*	*3 (%)*
Early Jurassic (Hettangian Toarcian) Total first appearances: three (sample too small)				
Middle Jurassic (Aalenian-Callovian)	12	1 (8)	3 (25)	8 (67)
Late Jurassic (Oxfordian-Portlandian)	36	11 (30.5)	14 (39)	11 (30.5)
Early Cretaceous (Berriasian-Albian)	17	4 (23.5)	9 (53)	4 (23.5)
Early Late Cretaceous (Cenomanian-Santonian)	18	8 (44)	9 (50)	1 (6)
Latest Cretaceous (Campanian-Maastrichtian)	46	26 (56)	10 (22)	10 (22)
Paleocene (Danian-Thanetian)	74	49 (66)	21 (28)	4 (6)

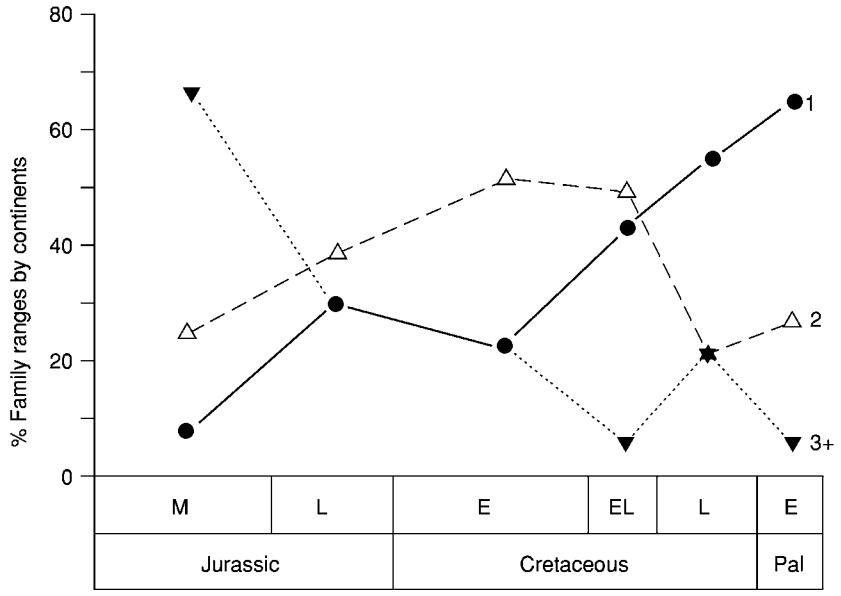


Figure 7 Radiations associated with fragmentation of continents. The breakup of Pangaea was associated with increasing provin ciality of terrestrial tetrapod families, plotted here from the Middle Jurassic to the Paleocene. This plot shows the percentage of families of terrestrial tetrapods that occupied one, two and three of the modern continental areas. Data given in table above. The plot shows that the percentage of families occupying only one of the modern continental areas rose from 8% of the families originating in the Middle Jurassic to 66% of those originating in the Paleocene. Conversely, 67% of families originating in the Middle Jurassic occupied three of the modern continental areas and this decreased to just 3% in the Paleocene. Reproduced from Benton MJ (1985) Patterns of diversification of Mesozoic non marine tetrapods and problems of historical diversity analysis. *Special Papers in Palaeon tology* 33: 185 202. Published by permission of the Palaeontological Association.

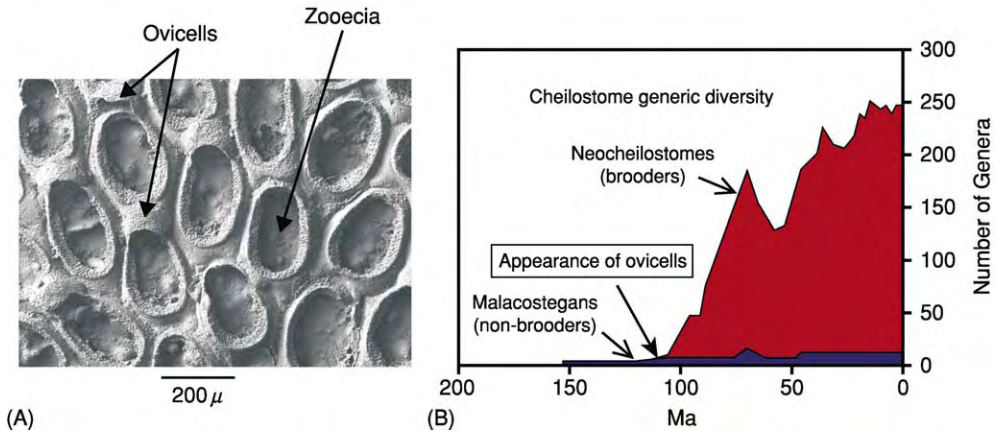


Figure 8 The cause of a clade radiation has often been identified as the evolution of a key innovation. One such example is the appearance of ovicells (brood chambers) in some cheilostome bryozoans. (A) *Welbertopora mutabilis*, upper Albian, Fort Worth Formation, Denton County, Texas. Part of the colony of one of the earliest neocheilostomes with ovicells in which larva develop. (B) Generic plots of non brooding and brooding cheilostomes showing the dramatic rise in the latter. The presence of Jurassic and Lower Cretaceous cheilostomes demonstrates that appearance of brooders is not a taphonomic artifact. Images supplied by Dr Paul Taylor, The Natural History Museum, from unpublished data.

along with environmental differentiation, and this almost certainly influenced the radiation of many marine organisms.

Origination of an Evolutionary Novelty Leading to Taxic Diversity

Often the radiation of a particular group has been linked with the appearance of a particularly significant character (e.g., development of the cleidoic egg, enabling the organisms to become reproductively independent of water, or wings to fly). In many of the usual examples there is very little taphonomic control, so that it is difficult to separate the appearance of an evolutionary novelty and increase in species diversity from the preservation bias that may be present in the fossil record. But there are some examples that avoid this problem. Cheilostomes are the dominant bryozoans (*see* **Fossil Invertebrates:** Bryozoans) living today and they are distinguished from other Bryozoa by the possession of box-shaped zooecia. The fossil record of cheilostomes begins in the Upper Jurassic, with a rapid rise in numbers of species and genera in late Lower Cretaceous. There are two kinds of cheilostomes: malacostegans (about 150 Recent species) and their species descendents, the neocheilostomes (about 5000 Recent species). Modern malacostegans release planktotrophic larvae that spend up to about 1 month free living in the water column before settling and beginning colony growth. During this time there is potential for considerable dispersal as well as associated potential for continued gene flow. Neocheilostomes brood their larvae within a specialized brood chamber: – the ovicell. The larvae have a very short, free life span before settling to the substrate (often adjacent to the parent) and beginning colony formation. Thus, gene flow between populations may be very restricted and as a consequence there is the greater potential for genetic divergence

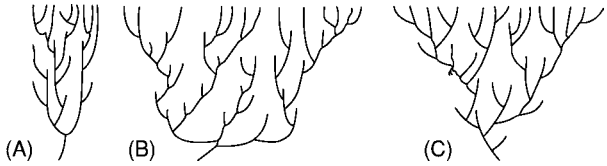
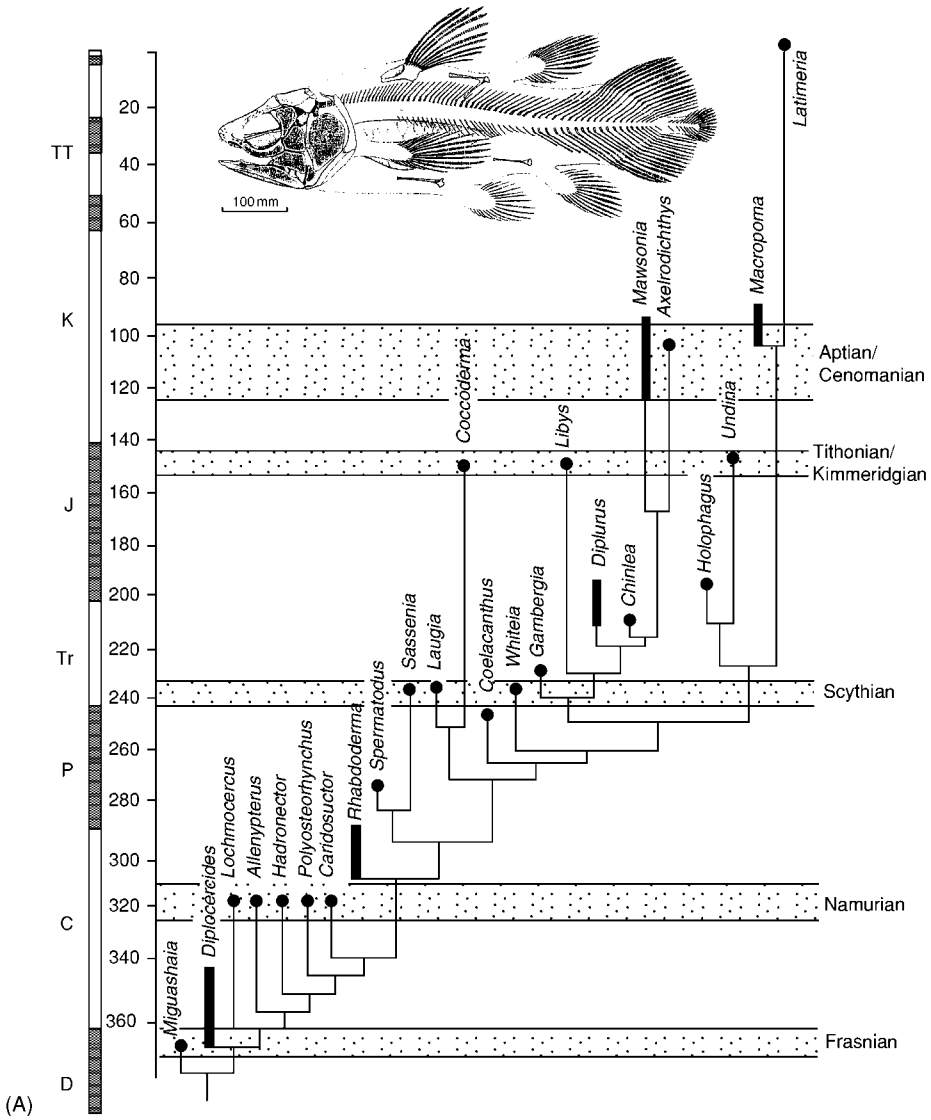
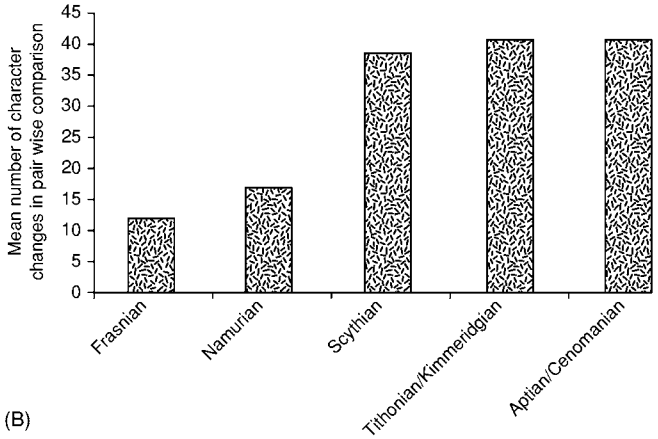
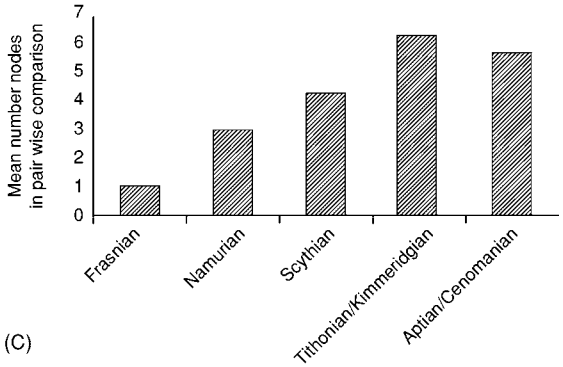


Figure 9 The history of a clade during phylogenetic diversifi cation may follow one of several patterns, idealized here. The horizontal axis represents morphological divergence, the verti cal axis time. (A). Morphological evolution constrained, species origination high. (B) Morphological evolution rapid in early periods relative to speciation. (C) Morphological evolution and speciation approximately in step with one another. From Foote M (1993).

Figure 10 Comparison of morphological and taxic evolution in coelacanth fishes using phylogenetic measures. (A) Phylogenetic tree showing 5 time bands selected as sample stages. Within each time band two types of comparison were made on a pair wise basis and then averaged. (B) The mean number of character changes occurring between taxa in each of the five time bands. This is a measure of morphological disparity. (C) The mean number of nodes (cladogenetic events) occurring between taxa in each of the five time bands. This is a measure of taxic diversity. In this case morphological diversity tracks taxic diversity except for a slight insignificant increase in taxic diversity relative to morphological diversity in the early stage (Namurian). Data from Forey PL (1998) *History of the coelacanth fishes*. London: Chapman & Hall.


100 mm
TT
K
J
Tr
P
C
D
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
Latimeria
Mawsonia
Axelrodichthys
Macropoma
Coccoderma
Libys
Undina
Holophagus
Diplurus
Chinlea
Sassenia
Laugia
Coelacanthus
Whiteia
Gambergia
Rhabdoderma
Spermatodus
Polyosteorhynchus
Caridosuctor
Hadronector
Allenypterus
Lochmocercus
Diplocercides
Miguashaia
Aptian/
Cenomanian
Tithonian/
Kimmeridgian
Scythian
Namurian
Frasnian
(A)


Mean number of character changes in pair wise comparison
0
5
10
15
20
25
30
35
40
45
Frasnian
Namurian
Scythian
Tithonian/Kimmeridgian
Aptian/Cenomanian
(B)


Mean number nodes in pair wise comparison
0
1
2
3
4
5
6
7
Frasnian
Namurian
Scythian
Tithonian/Kimmeridgian
Aptian/Cenomanian
(C)

and, eventually, speciation. The ovicells and hence brooding of larvae can be identified in the skeleton of the neocheilostomes (Figure 8A). Plots at the generic-level of malacostegans and neocheilostomes show little change in the former, but a dramatic rise in the latter (Figure 8B), suggesting strongly that the acquisition of brooding is responsible for the radiation. It is pertinent to note that very similar patterns involving similar contrasting reproductive strategies have been observed in turritellid gastropod evolution.

Morphological and Taxic Evolution During Radiations

Radiations involve particular patterns of morphological as well as taxic evolution. Figure 9 shows three types of relationship between morphological and taxic evolution during clade history. Studies examining the relationship between these two aspects may be phylogeny dependent or independent, and several different methods have been used to analyze both. Much literature suggests that the pattern in Figure 9B is common.

One study of the radiation of coelacanth fishes (*see* **Fossil Vertebrates:** Fish) used a strictly phylogenetic approach for both measures (Figure 10). The phylogeny of 24 genera was sampled at five time intervals, which collectively spanned the Upper Devonian to Upper Cretaceous. Therefore, this study dealt with a radiation that was long term and involved few taxa and, while demonstrating a method,

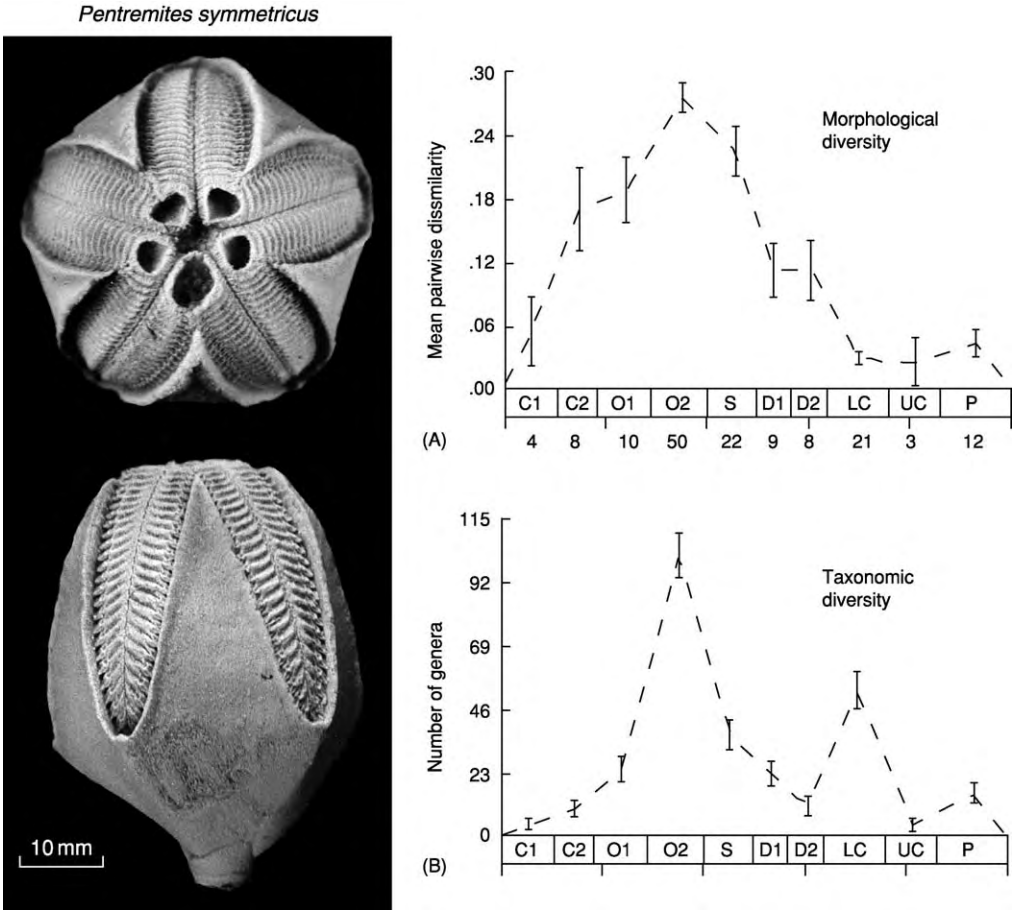


Figure 11 Comparison of morphological and taxic evolution in Blastozoa echinoderms using non phylogenetic measures taken at approximately equal time slices from Lower Cambrian to Permian. (A) Mean pair wise dissimilarity, calculated by comparing the number of dissimilar codes for 65 characters between each pair and averaging over the number of comparisons made. The number of genera (one species of each was examined) is given against each time slice examined. This gave an average measure of morphological disparity. (B) Number of genera at each time slice as a measure of taxic diversity. This study showed a rapid burst of morphological evolution that was not accompanied by taxic evolution in the early stages of clade history (C2 and O1). Note that quite the reverse trend was evident later in clade history (LC). Time slice abbreviations: C1 Lower Cambrian, C2 Middle and Upper Cambrian, O1 Lower Ordovician, O2 Middle and Upper Ordovician, S Silurian, D1 Lower Devonian, D2 Middle and Upper Devonian, LC Lower Carboniferous, UC Upper Carboniferous, P Permian. (After Foote 1992, with permission of author). Photograph supplied by Dr AB Smith.

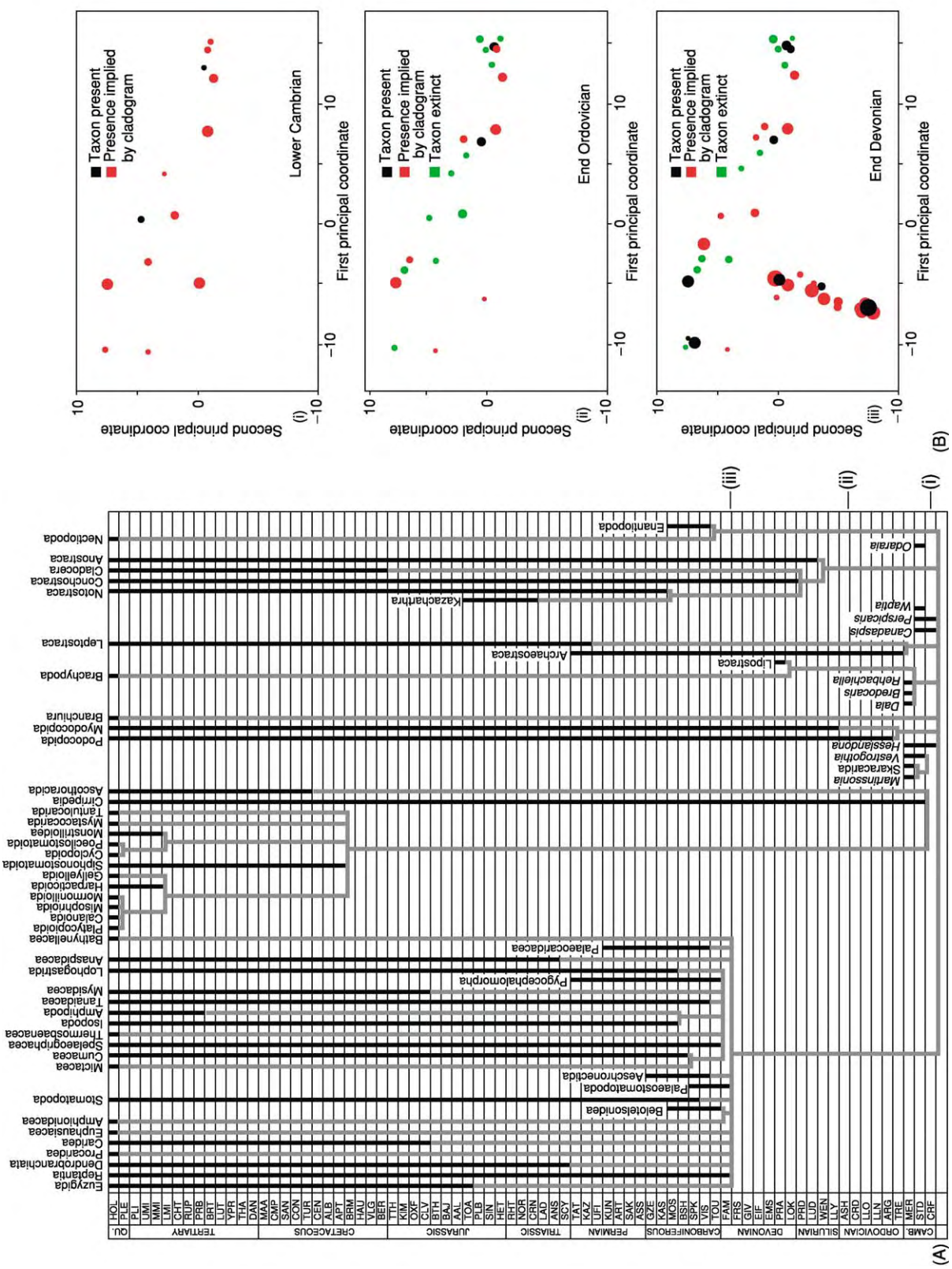


Figure 12 (Continued)

may not be typical. There was very little taphonomic control in this study. At each interval the distance, as measured by number of character changes along the inferred lineage between terminal taxa, was averaged over all pair wise comparisons. This gave an average morphological distance for successive time periods in the clade's history. The values for successive time periods could then be compared to each other and, in turn, compared to the number of cladogenetic events within each time period (which provided a measurement of taxic diversity). In this case the taxic diversity tracked morphological diversity except for the early stages where taxic diversity slightly exceeded morphological diversity.

Some other studies have not considered the phylogenetic relationships within clades (although monophyly of the entire clade has been accepted). Figure 11 shows the results of such a study of the radiation of members of the echinoderm subphylum Blastozoa (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids)). In this study, one species each from 147 genera, ranging in time from the Lower Cambrian to Permian, were examined for 65 morphological characters. Ten stratigraphic intervals were chosen and within each interval pair wise comparisons between characters coded for each of the genera were made and averaged. The phylogeny was not considered here. One common belief is that morphological diversity is disproportionately high compared to taxic diversity at the beginning of a radiation and that such a relationship is reversed later on. The blastozoan study appears to confirm this since the morphological diversity was far greater than the taxic diversity in the mid-Cambrian to Lower Ordovician; that is, in the early stages of clade history. The reverse relationship was true in the Lower Carboniferous.

The Cambrian explosion must be the most spectacular of all radiations that the planet has witnessed. Yet it remains one of the most enigmatic. Within 15 million years of the beginning of the Cambrian period, where tiny shelly faunas are first detected, a wide variety of skeletonized animals appeared, already well diversified and exquisitely adapted. Today they are classified amongst the major invertebrate phyla. Molecular evidence, alongside phylogenetic studies on the first appearing fossils all point to there being a long Precambrian history to many lineages. But exactly what form this history took is very uncertain. Steven Jay Gould, in his book *Wonderful Life*, – suggested that when skeletonized animals first appeared in abundance in the Cambrian there was a remarkably diverse array of body forms and that many of these were weeded out by extinctions leading to differential survival of a few body forms that eventually gave rise to our modern fauna. In other words, morphological disparity was greatest in the early phase of the animal lineage history and gradually became canalized. On a broad scale this theory echoed the radiation pattern shown by the blastozoan echinoderms mentioned above. Gould's thesis was grounded in a large part on the arthropods, and especially the crustacean-like animals, conspicuous in the famous Burgess Shale Lagerstätten, but was largely anecdotal, with little phylogenetic evidence or any precise measures of morphological disparity applied. More detailed studies comparing levels of disparity in Cambrian and Recent arthropods (*see* **Fossil Invertebrates:** Arthropods) showed that there was very little difference between then and now. More specifically detailed studies of the early evolution of the Crustacea have shown that morphological disparity grew in concert with taxic diversity (Figure 12).

The study of the pattern of radiations in the fossil record has implications for our understanding of post-extinction recovery and potentially for our efforts to put conservation efforts in place. If there are general patterns that emerge that can be associated with

Figure 12 Morphological disparity of crustaceans tracked through time against a phylogeny. In this study one exemplar taxon from each of the crustacean orders or suborders was coded for 135 characters. (A) A cladistic analysis plotted against stratigraphy resulted in the consensus tree shown here. The actual stratigraphic ranges are shown in black and those which must be assumed because of the shape of the tree in grey (ghost ranges). Notice that most of the species of modern crustacea did not appear until late Palaeozoic. (B) Plots showing disparity of the clades present through three time slices. The disparity is measured by making pair wise comparisons between the coded characters for each of the taxa present in any one time interval (the ghost range taxa were coded according to reconstructed ancestral states). Some differential weights were applied according to the particular type of character and in this it differed from the Blastozoan study (Figure 11). A principal coordinate analysis was then carried out to illustrate the difference in morphospace between the taxa. In the diagrams the first three principal coordinates are shown, the third coordinate being at right angles to the paper (filled symbols positive and open symbols negative). The black symbols represent taxa with a range extending through the stage illustrated, red symbols represent taxa implied by ghost ranges and the green symbols are taxa extinct at stage illustrated. This diagram shows that both morphological disparity and taxic diversity increased from the Cambrian to Upper Devonian at which time a new area of morphospace (bottom left) was occupied. Reproduced from Wills MA (1998) Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. *Biological Journal of the Linnean Society of London* 65: 455 500.

specific ecological conditions, then some predictions may be made as to where and what to protect.

See Also

Evolution. **Fossil Invertebrates:** Arthropods; Bryozoans; Echinoderms (Other Than Echinoids); Gastropods. **Fossil Vertebrates:** Fish.

Further Reading

Benton MJ (1985) Patterns of diversification of Mesozoic non marine tetrapods and problems of historical diversity analysis. *Special Papers in Palaeontology* 33: 185 202.

Cheetham AH (1986) Tempo of evolution in a Neogene bryozoan: rates of morphological change within and across species boundaries. *Paleobiology* 12: 190 202.

Erwin DH and Anstey RL (1995) Introduction. In: Erwin DH and Anstey RL (eds.) *New approaches to speciation in the fossil record*, pp. 1 10. New York: Columbia University Press.

Foote M (1992) Paleozoic record of morphological diversity in blastozoan echinoderms. *Proceedings of the Academy of Natural Sciences* 89: 7325 7329.

Foote M (1993) Discordance and correspondence between morphological and taxonomic diversity. *Paleobiology* 19: 185 204.

Forey PL (1998) *History of the coelacanth fishes*. London: Chapman & Hall.

Jablonski D, Lidgard S, and Taylor PD (1997) Comparative ecology of bryozoan radiations: origin of novelties in cyclostomes and cheilostomes. *Palaios* 12: 505 523.

Lazarus DB (2001) Speciation and morphological evolution. In: Briggs DE and Crowther PR (eds.) *Palaeobiology II*, pp. 133 137. Oxford: Blackwell Scientific.

Levington JS (2001) *Genetics, Palaeontology and Macroevolution*, 2nd edn. p. 617. Cambridge: Cambridge University Press.

Nehm RH and Geary DH (1994) A gradual morphologic transition during a rapid speciation event in marginellid gastropods (Neogene; Dominican Republic). *Journal of Paleontology* 68: 787 795.

Schluter D (2000) *The ecology of adaptive radiation*. Oxford: Oxford University Press.

Stanley SM (1979) *Macroevolution: pattern and process*. San Francisco: WH. Freeman.

Taylor PD and Larwood GP (eds.) (1990) *Major evolutionary radiations*. Oxford: Clarendon Press.

Wheeler QD and Platnick NI (eds.) (2000) *Species concepts and phylogenetic theory*. New York: Columbia University Press.

Wills MA (1998) Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. *Biological Journal of the Linnean Society of London* 65: 455 500.

BIOSEDIMENTS AND BIOFILMS

M R Walter and A C Allwood, Macquarie University, Sydney, NSW, Australia

Introduction

Throughout most of the geological record we find traces of the past inhabitants of our planet. Bones, woody stems, and hard outer carapaces provide evidence of the passing of species, entombed by the sediments that were deposited around them. However, the organisms that yield these relatively familiar kinds of fossils represent only the evolutionary branch tips of the 'Tree of Life' (**Figure 1**), and have risen only recently in the history of life on Earth. The bulk of the 'Tree of Life', and most of the history of life on Earth (**Figure 2**), is represented by microscopic organisms. Unfortunately, these tiny, soft-bodied creatures are not readily fossilized. Moreover, where fossils are preserved, their tiny size, simple morphology, and imperfect preservation make them extremely difficult to recognize. This challenge increases in the search for the earliest fossils on Earth within rocks that have been subjected to over 3 billion years of geological processes: not only are the fossils likely to be heavily degraded, but the very nature of the host sedimentary deposits can be quite different to younger deposits, so that the misinterpretation of unfamiliar structures can be a risk.

The challenging search for the oldest traces of life is further limited by the availability of suitable rocks. There are just three occurrences of Early Archaean sedimentary rocks, from remnant Archaean cratons in Greenland (*ca*. 3.8 Ga), South Africa, and Western Australia (3.4–3.5 Ga; **Figure 2**). The rocks of Greenland, however, have been subjected to pressures and temperatures that have obliterated original depositional fabrics. Thus, there are just two potentially fossiliferous sedimentary sequences remaining from the era during which life arose on Earth. Over recent

decades, these areas have become the focus of studies of the Earth's earliest biosphere and, despite the imagined impossibility of finding microbial remains in rocks of such antiquity, there have indeed been several 'fossil' discoveries. Evidence for life at *ca*. 3.5 Ga comprises three different types of 'fossil': (1) microscopic bacteria-like structures, interpreted as microfossils; (2) chemical signatures, interpreted as resulting from

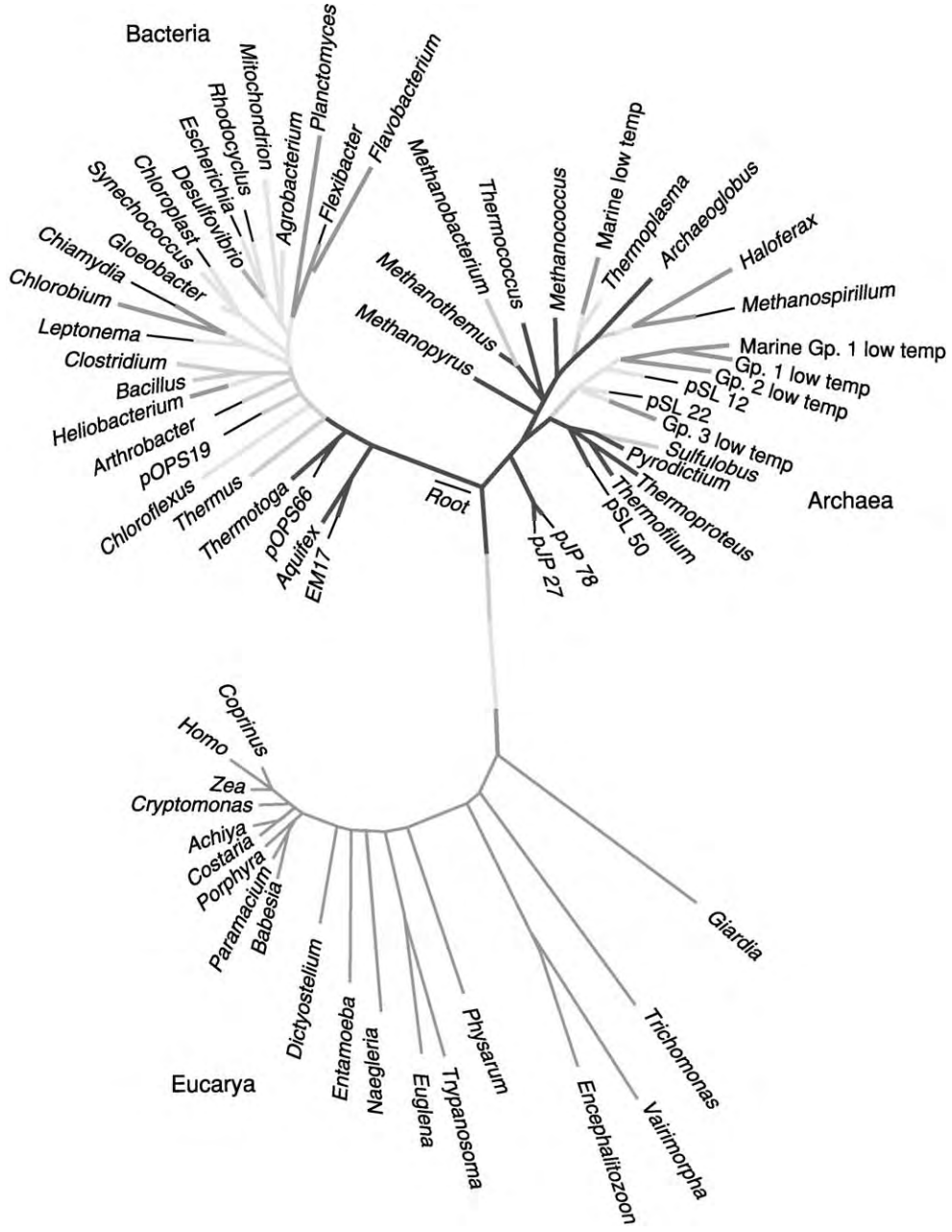
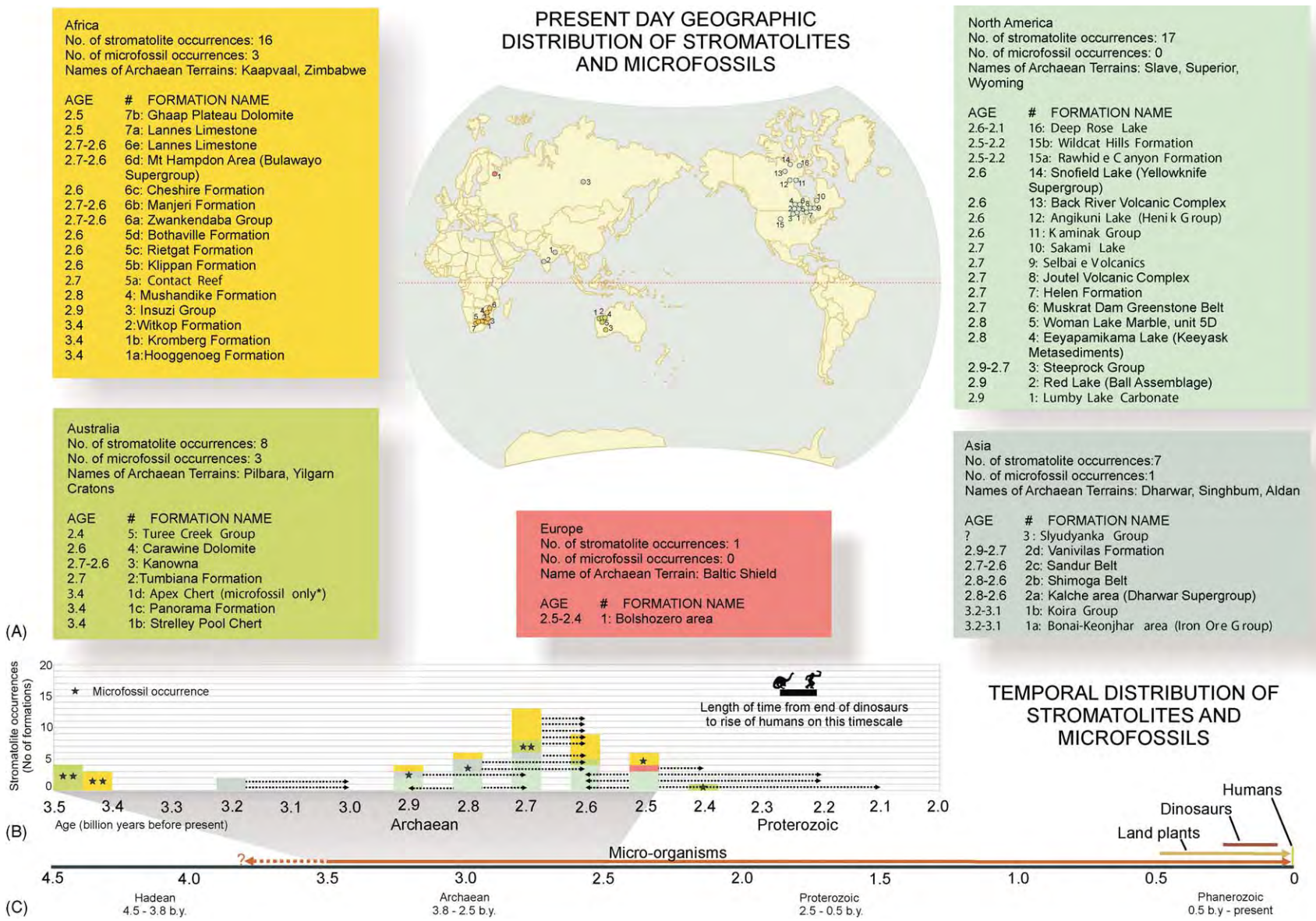


Figure 1 The 'Tree of Life' showing the genetic relationships between all living organisms. All branches are microbial, except for *Coprinus*, *Homo*, and *Zea*, at the extreme tip of the lower branch.

Figure 2 (A) Geographical distribution of Archaean stromatolite occurrences on the surface of the Earth today. All occurrences are clustered on remnants of Archaean crust in Africa, Australia, North America, Europe, and Asia. The oldest known stromatolites and microfossils occur in the Pilbara craton of Australia and the Kaapvaal craton in South Africa. (B) Temporal distribution of Archaean stromatolite and microfossil occurrences. Data are coloured by continent. The age of some occurrences is approximate, and some may extend into the Proterozoic, as indicated by dotted lines. (C) Geological time scale highlighting the dominance of microbes throughout most of the Earth's history. Higher organisms represent only the most recent fraction of the history of life on Earth.


PRESENT DAY GEOGRAPHIC DISTRIBUTION OF STROMATOLITES AND MICROFOSSILS
Africa
No. of stromatolite occurrences: 16
No. of microfossil occurrences: 3
Names of Archaean Terrains: Kaapvaal, Zimbabwe
AGE # FORMATION NAME
2.5 7b: Ghaap Plateau Dolomite
2.5 7a: Lannes Limestone
2.7-2.6 6e: Lannes Limestone
2.7-2.6 6d: Mt Hampdon Area (Bulawayo Supergroup)
2.6 6c: Cheshire Formation
2.7-2.6 6b: Manjeri Formation
2.7-2.6 6a: Zwankendaba Group
2.6 5d: Bothaville Formation
2.6 5c: Rietgat Formation
2.6 5b: Klippan Formation
2.7 5a: Contact Reef
2.8 4: Mushandike Formation
2.9 3: Insuzi Group
3.4 2:Witkop Formation
3.4 1b: Kromberg Formation
3.4 1a:Hooggenoeg Formation
North America
No. of stromatolite occurrences: 17
No. of microfossil occurrences: 0
Names of Archaean Terrains: Slave, Superior, Wyoming
AGE # FORMATION NAME
2.6-2.1 16: Deep Rose Lake
2.5-2.2 15b: Wildcat Hills Formation
2.5-2.2 15a: Rawhide Canyon Formation
2.6 14: Snofield Lake (Yellowknife Supergroup)
2.6 13: Back River Volcanic Complex
2.6 12: Angikuni Lake (Henik Group)
2.6 11: Kaminak Group
2.7 10: Sakami Lake
2.7 9: Selbaie Volcanics
2.7 8: Joutel Volcanic Complex
2.7 7: Helen Formation
2.7 6: Muskrat Dam Greenstone Belt
2.8 5: Woman Lake Marble, unit 5D
2.8 4: Eeyapamikama Lake (Keeyask Metasediments)
2.9-2.7 3: Steeprock Group
2.9 2: Red Lake (Ball Assemblage)
2.9 1: Lumby Lake Carbonate
Australia
No. of stromatolite occurrences: 8
No. of microfossil occurrences: 3
Names of Archaean Terrains: Pilbara, Yilgarn Cratons
AGE # FORMATION NAME
2.4 5: Turee Creek Group
2.6 4: Carawine Dolomite
2.7-2.6 3: Kanowna
2.7 2:Tumbiana Formation
3.4 1d: Apex Chert (microfossil only*)
3.4 1c: Panorama Formation
3.4 1b: Strelley Pool Chert
Europe
No. of stromatolite occurrences: 1
No. of microfossil occurrences: 0
Name of Archaean Terrain: Baltic Shield
AGE # FORMATION NAME
2.5-2.4 1: Bolshozero area
Asia
No. of stromatolite occurrences:7
No. of microfossil occurrences:1
Names of Archaean Terrains: Dharwar, Singhbum, Aldan
AGE # FORMATION NAME
? 3: Slyudyanka Group
2.9-2.7 2d: Vanivilas Formation
2.7-2.6 2c: Sandur Belt
2.8-2.6 2b: Shimoga Belt
2.8-2.6 2a: Kalche area (Dharwar Supergroup)
3.2-3.1 1b: Koira Group
3.2-3.1 1a: Bonai-Keonjhar area (Iron Ore Group)
(A)
Stromatolite occurrences (No of formations)
Microfossil occurrence
Length of time from end of dinosaurs to rise of humans on this timescale
Age (billion years before present)
Archaean
Proterozoic
(B)
TEMPORAL DISTRIBUTION OF STROMATOLITES AND MICROFOSSILS
Humans
Dinosaurs
Land plants
Micro-organisms
Hadean 4.5 - 3.8 b.y.
Archaean 3.8 - 2.5 b.y.
Proterozoic 2.5 - 0.5 b.y.
Phanerozoic 0.5 b.y - present
(C)

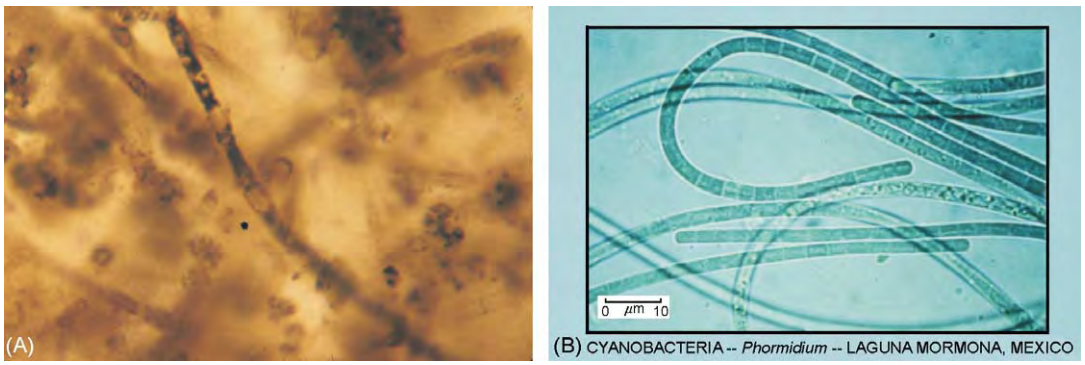


Figure 3 (A) Filamentous microfossils from the 1.9 Ga Gunflint Formation, Ontario, Canada, arguably the oldest widely accepted microfossils, although older ones have been reported. Photographs supplied by J. W. Schopf. (B) Modern filamentous microbes. Part A and B based on data from: Hoffman H (2000) Archean Stromatolites as Microbial Archives. In: Riding R and Awramik S (eds.) *Microbial Sediments*, pp. 315 327. Berlin: Springer Verlag.

microbial activity during formation of the rock; and (3) unique sedimentary structures, interpreted as stromatolites (laminated sedimentary structures formed by the activities of microbes) (Figure 3).

Arguably, these are the oldest traces of life on Earth. However, owing to the significance of determining when, where, and how life on our planet arose, a great deal of controversy surrounds the interpretation of the claimed fossils. As Carl Sagan once said, "extraordinary claims require extraordinary proof". Consequently, rigorous criteria have been developed for authenticating Early Archaean fossils. Moreover, early life on Earth is now recognized as an analogue for ancient microbial life on Mars, stimulating considerable research into the recognition of ancient microbial life using techniques that can be carried out by a rover or, eventually, a geologist on the surface of Mars.

Life is thought to have arisen on Earth sometime early during the 1.3 billion year period from 3.8 to 2.5 Ga, although the exact timing is still debated (*see* **Origin of Life**). In order to understand how we might interpret Archaean fossils, we will examine the way in which microbes interact with sediments in modern settings. Not only is this critical for interpreting the very oldest traces of life, but it is also important to hone our techniques in the familiar geological environments of Earth to aid us in the search for signs of life elsewhere.

Nature of Modern and Ancient Biosedimentary Systems

Microbial Sediments: Significance and Distribution

Today, almost no sedimentary environments are devoid of life. Through sheer numbers, microbes have played a fundamental role in shaping surface environments through time by their metabolic contributions to the global cycling of important elements, such as carbon, oxygen, sulphur, nitrogen, phosphorus, and iron. Microbes have also shaped the geological record of sedimentary environments, particularly through the activities of microbial communities living at the sediment–water interface. Although microbes live both as free-floating or swimming individuals (plankton), as well as structured, surface-adhered communities (biofilms), here we are particularly interested in the latter. This is because some of the activities undertaken by these film or mat-like communities can influence the way in which sediment is deposited and impart unique physical and chemical signatures that may be preserved in the geological record.

Although microbes are ubiquitous in all watery surface environments today, formation of microbial sedimentary structures, such as stromatolites, that could be recognized in the geological record is less common. Today, communities that form such structures occur mostly in environments that are inhospitable to higher, grazing organisms that feed upon the microbes. Such environments include those with extremes of temperature, pH, oxygen depletion (eg., hot springs), salinity (stranded lakes) or where the organisms are repeatedly exposed to drying conditions (intertidal zones). Sediment influx must also be minimal to prevent rapid burial of the mats. A famous modern example occurs in a hypersaline restricted coastal embayment at Hamelin Pool, Shark Bay, Western Australia (Figure 4). Salinity levels in Shark Bay are up to twice those of normal seawater, creating an inhospitable environment for most marine animals but hospitable for microorganisms associated with the stromatolites.

Although Shark Bay may not provide a perfect analogue for the first ecosystems, its lack of predators and

consequent free reign of microbial activity certainly mimic one important aspect of the early biosphere: the dominance of micro-organisms. Microbes were the only form of life on Earth during the Archaean (3.8–2.5 Ga) and much of the Proterozoic (2.5–0.5 Ga) aeons (Figure 2C). Indeed, it was the rise of the first grazing organisms at the end of the Proterozoic (0.5 Ga) that is thought to have triggered the near disappearance of stromatolite-forming microbial activity from that point onwards.

Nature of Mats and Biofilms

Biofilms are structured communities of micro-organisms adhering to a submersed surface. The species within a biofilm have planktonic equivalents (individual cells suspended in the water column), but only organisms in a biofilm are structured in a way that can significantly affect sedimentation. Several microbial species may be present within a biofilm, arranged in tiny mushroom- and pillar-type structures within an extracellular matrix of organic material (Figure 5). The

Figure 4 (A) Modern stromatolites from the subtidal realm of Hamelin Pool, Shark Bay, Western Australia. A range of stromatolitic microbial communities also occur in the intertidal zone, forming different structures, including tabular stromatolites and microbially bound ripples. (B) Shark Bay stromatolite in cross section, showing a clotted, slightly laminated texture.

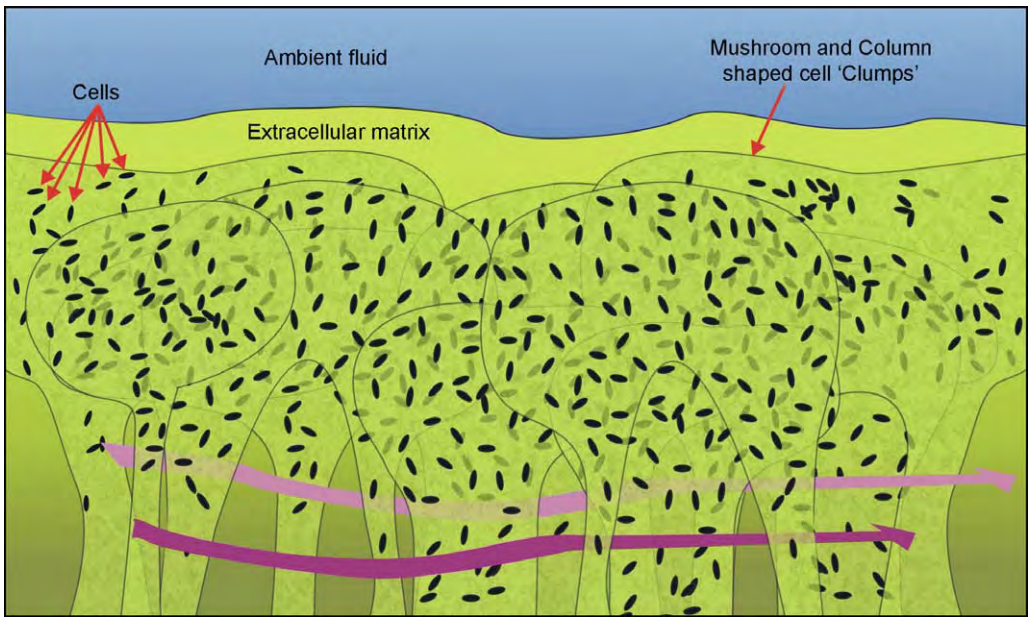


Figure 5 Schematic cross section of a biofilm. Cells are grouped together in mushroom and pillar type structures surrounded by a matrix of extracellular material. The intervening matrix provides a pathway for the diffusion of nutrients and wastes. Adapted from an illustration by Dirckx P (1999) Not Just Slime. In: Ben Ari E (ed.) *Bioscience* V. 49, no. 9. p. 691.

matrix provides a stable micro-environment that offers several advantages for cells, including:

1. Protection from harmful environmental fluxes and agents (e.g., temperature change, radiation damage, toxic materials).
2. Adhesive properties to anchor the community to the surface.
3. Efficient nutrient uptake from the surrounding environment.
4. Pretreatment of assimilated molecules, such as the breakdown of large organic molecules to smaller, more manageable components, which are then able to diffuse through the matrix and be taken up by the cells.
5. Conduits for the diffusion of metabolic by-products away from the cells.

Thus, cells in the biofilm are able to efficiently assimilate a range of organic and inorganic molecules from the environment through the matrix for use in reactions that provide energy and building blocks for cells. Moreover, the by-products from one species may provide the nutrients for another species within the biofilm. These processes are a significant component of the interaction between microbes and their environment.

A 'microbial mat' is a thicker and more complex type of biofilm, comprising many different species divided into specific layers. Different layers are favoured by different species because of their preference for certain light conditions, oxygen levels, sulphide concentrations, and so forth (Figure 6).

Figure 6 A piece of modern microbial mat from the shallow subaqueous sediments at Shark Bay (cross sectional view). The tufted, dark green layers at the top (photosynthetic cyanobac teria) are underlain by a light brown layer, a thin red brown layer, and then a grey black layer. The different colours in each layer reflect pigments, mineral matter, and organic material associated with different microbial species.

Effects of Microbial Activity upon Sedimentation

Mineral precipitation Mineral precipitation is commonly associated with microbial mat growth. However, precipitation may or may not be actively caused by the metabolic activities of the organisms in the mat. The difference between active and passive precipitation processes is described in Table 1.

Biotic (active) precipitation occurs when the by-products of microbial reactions alter the chemical balance within the micro-environment and cause the precipitation of compounds such as carbonate, iron oxide, iron sulphide, or silica. Microbial photosynthesis, for example, is one of the most important mechanisms for biological calcium carbonate (limestone) precipitation on a global scale. In microbial mats, this process involves the photosynthetic removal of CO_2 from a bicarbonate-bearing solution, which increases the carbonate (CO_3) concentration in the residual fluid. The carbonate may then bond with a cation, such as calcium, and precipitate out of the fluid as calcium carbonate. Photosynthesis is not the only mechanism for CO_2 removal and carbonate precipitation; the same can be achieved through certain biological oxidation, reduction, or hydrolysis reactions. These reactions could have provided a mechanism for biological carbonate precipitation prior to the rise of photosynthetic organisms. Mineral

Table 1 Comparison of non biological and biological precipitation processes

Passive (non biological) precipitation	*Active (biological) precipitation*
Mechanism	
Fluid saturation levels are sufficiently high such that mineral precipitation would occur without biological influence	Fluid saturation levels are lower than those necessary for spontaneous precipitation, yet precipitation continues due to microbial mediation of chemical reactions in the environment, i.e., microbes are responsible for maintaining physicochemical parameters, such as minimum levels of carbonate concentration (via photosynthesis) necessary for precipitation
Microbes simply act as nucleation sites, much as a twig or fallen piece of rock	
Microbially influenced crystal fabrics may form and be preserved within such deposits, but precipitation is not caused by microbial activity	
Example environment	
Common in hot spring environments where supersaturated fluids emanate from springs or vents	May occur in marine environments where minerals will not spontaneously precipitate from seawater

deposition can also occur when elements, such as iron, become concentrated by complexing (bonding) with microbially produced organic molecules.

Abiotic precipitation may occur in certain environments, such as hot springs, where fluids may be sufficiently saturated to spontaneously precipitate minerals even without microbial influence. These environments are commonly inhabited by microbes, perhaps simply because they are the only organisms that are able to survive under the prevailing conditions. Where fluid saturation approaches a level high enough for spontaneous precipitation, the microbes may simply act as templates for precipitation, a concept that is supported by the observation that fallen branches, insects, and other detritus in the same hot springs provide equally efficient substrates for mineral precipitation.

Trapping and binding In addition to causing the *in situ* precipitation of new minerals, thereby forming new, localized sediment deposits, microbial communities can also trap and bind detritus washed in from elsewhere. As material is trapped, the microbial community moves up through the sediment layer to form a new organic layer at the surface. Upward movement may be driven by, for example, a need to reach sunlight. The rate of sediment influx must be such so as to avoid either deeply burying the microbial mat or starving the mat of substrate material for growth.

Ancient Microbial Sediments

Microbes of the Early Archaean left traces of their passing in much the same way as microbes do today, forming distinctive fabrics, structures, and chemical signatures in the sediments that fell and precipitated around them. There are four basic kinds of fossil evidence, or biosignatures, that we can study in the geological record to unravel the nature of the oldest biosedimentary systems. These are listed in Table 2.

The interpretation of ancient biosedimentary systems is aided by the study of younger analogues, but this is limited by the fact that surface processes and environments on the early Earth were very different to modern systems. Thus, direct comparisons cannot always be made between the Archaean Earth and the Phanerozoic or modern Earth in order to understand the types of processes that were responsible for the formation of certain fossil-like structures. Nonetheless, careful and selective study of modern and younger analogues provides an extremely useful starting point for the interpretation of ancient systems. A discussion of the formation, occurrence, and interpretation of different microbial sedimentary products and their fossils follows.

Stromatolites

Stromatolites constitute by far the most abundant and important component of the early fossil record and thus provide the main body of evidence regarding the origins of life. The term 'stromatolite' has been widely used to refer to a diverse range of organic and/or laminated sedimentary structures. However, most researchers have used the definition by Awramik and Margulis, which defines a stromatolite as: "an organosedimentary structure produced by sediment trapping, binding, and/or precipitation resulting from the growth and metabolic activity of microorganisms".

Stromatolites accrete through the trapping and binding of sediment and/or by mineral precipitation that is influenced or caused by microbes at the sediment–water interface. Growth occurs because, as sediment accumulates on a stromatolite, either by deposition or precipitation, the microbes must move upwards to remain at the sediment–water interface. Thus, in living stromatolites, the constructing microbes generally live only in the top 1–10 mm of the structure. The physical shape of stromatolites may be columnar, domed, branching, tabular, or

Table 2 Summary of the main types of fossil evidence, or biosignatures, that may be sought in the geological record

Fossil type	*Type of evidence provided*	*Description*
Microfossils	Direct, morphological	Occurs where mineral precipitation has entombed and preserved actual microbes. Should be made of biological organic matter, but fossil casts occur where organic matter has been replaced by other minerals
Stromatolites	Indirect, morphological	Sedimentary structure formed by the combination of sedimentation and the activities of microbes. Classed as trace fossils (similar to a dinosaur footprint where the remains of the organism itself may be absent)
Chemical fossils	Indirect, chemical	Remnant environmental effects and by products of biochemical reactions that take place during microbial metabolism. Perhaps the subtlest of all fossils are the chemical fossils
Biomarkers	Indirect, biochemical	Remnant biological marker molecules that survive alteration processes that degrade morphological fossils and turn the original organic material to kerogen

mound-shaped (Figure 7). The combined attributes of a stromatolite reflect the interaction of physical, chemical, and biological processes (Figure 8) in the environment. For example, currents may cause elongation of the stromatolite along the current direction, a feature observed in modern stromatolites in Shark Bay and in Archaean stromatolites of Western Australia. When interpreting the morphology of stromatolites, it is difficult to determine what proportion of the physical attributes of a structure can be attributed to environmental influences and how much to purely biological factors.

The hard, laminated interior of the stromatolite records the depositional history of the stromatolite-building process. We can look at the nature of the laminations in order to understand the behaviour and character of a stromatolite-building community. Lamination within any sedimentary deposit reflects some past cyclicity in the sedimentation processes. In the case of stromatolites, lamination reflects cyclic variations in non-biological sedimentation processes, coupled with the variation in microbial activity. Lamination may accrete rapidly or slowly, depending on the causes of accretion and cyclicity. Table 3 describes four generalized mechanisms observed in modern stromatolites.

Interpretation of stromatolite-like structures Stromatolite-like structures can form by a variety of abiological mechanisms. In order to verify a biologic origin for purported stromatolites, all possible non-biological origins ought to be considered, including:

1. Mechanical sedimentary processes such as those caused by waves or currents.
2. Post-burial deformation of laminar deposits by dewatering processes.
3. Tectonic deformation, i.e., folding.
4. Soft sediment deformation or slumping, i.e., cohesive downhill movement of a portion of sedimentary layers.
5. Chemical precipitation, e.g., from saturated fluids, such as those associated with hot spring environments.

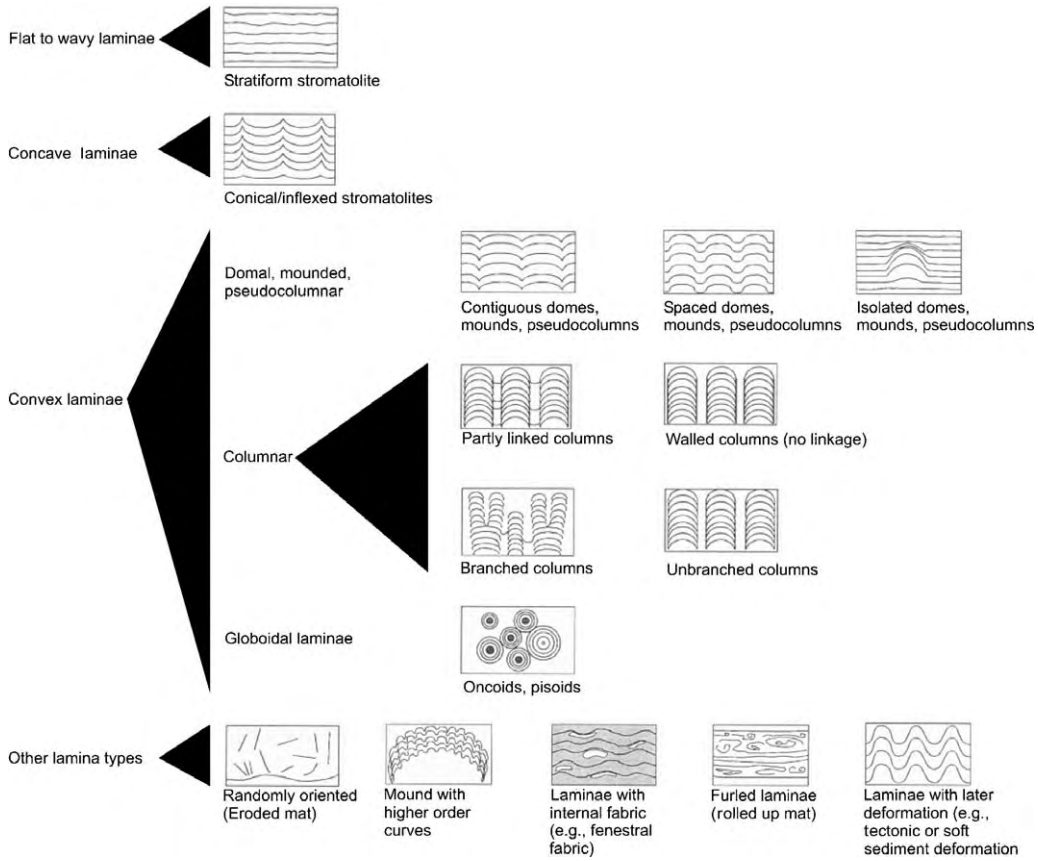


Figure 7 Physical attributes of Archaean stromatolites. Different attributes can be combined; for example, one 'community' may comprise spaced domes with higher order curvature and later deformation. These basic characteristics can be used to describe and classify stromatolites observed in the field. Adapted from Hoffman HJ (2000) Archaean stromatolites as microbial archives. In: Riding R and Awramik S (eds.) *Microbial Sediments*, pp. 315 327. Berlin, Heidelberg: Springer Verlag.

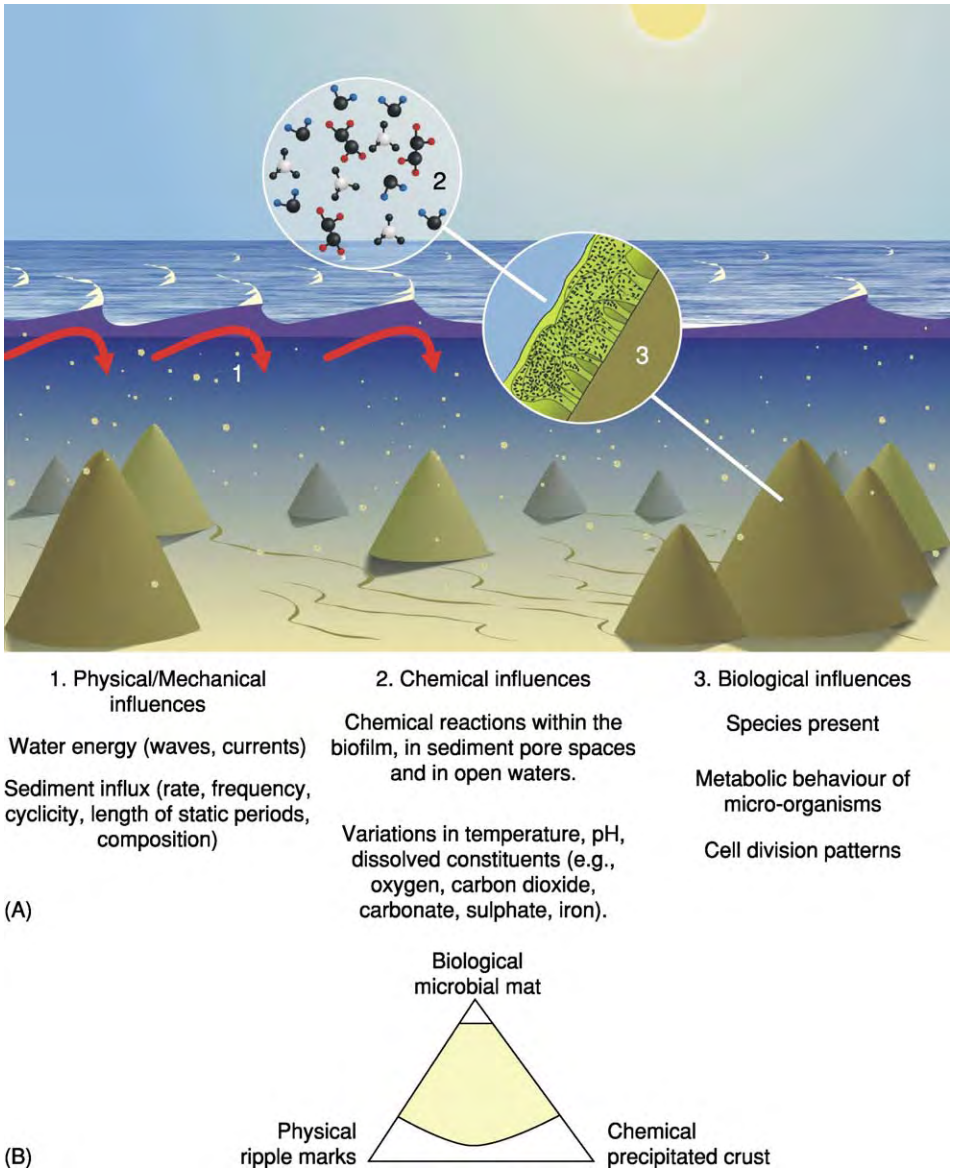


Figure 8 (A) Schematic representation of the processes affecting stromatolite formation. Although stromatolites form as a result of interactions between microbes and sediments, the final characteristics of the stromatolite are also affected by environmental parameters, such as waves and currents, sediment influx, water chemistry, temperature, and pH. All of these parameters can affect the structure and development of the microbial communities and their interaction with the materials around them. (B) Stromatolites are influenced by physical, chemical, and microbial processes, which can be represented on a triangular diagram as shown. Stromatolites in the geological record are clustered in the shaded area. The apices represent end members; for example, in a system in which only physical and/or chemical processes dominate, microbial communities, and therefore stromatolites, will not form.

Ideally, microfossils, chemical fossils, and microscale fabrics should be used in the interpretation of stromatolite-like structures, as the macroscale physical shapes created by biological and non-biological processes may be the same for a given set of environmental parameters. Unfortunately, microfossils and microfabrics are commonly destroyed by diagenesis (pressure- and temperature-related alteration during burial) (*see* **Diagenesis, Overview**). This is the case in all known Early Archaean stromatolites.

Archaean stromatolites Except for the latest Archaean Transvaal (Africa), Carawine and Fortescue (Australia) sequences, Archaean stromatolites occur mainly in thin, discontinuous, localized carbonate units formed at times of transient stability between tectonism and volcanism. These stromatolitic deposits are now preserved in the volcano-sedimentary sequences (greenstone belts) of Archaean cratons (Figure 2). A summary of the attributes of Archaean stromatolites is presented in Table 4 .

Table 3 Comparison of four generalized mechanisms for the formation of stromatolite lamination

Cause of accretion cycles	*Microbial response*	*Example*
Daily solar cycle. Independent of sedimentation	Phototactic (light responsive). Filamentous cells lie flat at night, but stand upright during the day in order to reach sunlight	*Phormidium hendersonii, Conophyton* (Yellowstone National Park)
Episodic sediment influx. Independent of light conditions	Normally flat lying filamentous organisms; glide vertically after new sediment influx in order to escape burial. Lamination is dependent on sediment supply	*Schizothrix* (subtidal stromatolites, Bahamas), *Microcoleus*
Seasonal changes in water chemistry	Low rainfall and warmer temperatures during the summer cause a change in organism dominance and behaviour	Lacustrine stromatolites, south western Australia
Extended pauses in sedimentation	Profound changes in species composition in the mat. Laminae construction styles vary due to different behaviours of the new community	Multispecies mats at Great Sippewissett Salt Marsh, Massachusetts

Based on data from Seong Joo *et al.* (2000) On Stromatolite Lamination. In: Riding R and Awramik S (eds.) *Microbial Sediments*, pp. 16 24. Berlin: Springer Verlag.

The geographical distribution of Early Archaean stromatolites today reflects the occurrence of preserved Early Archaean sedimentary rocks, which are restricted to narrow zones of sedimentary rocks, known as greenstone belts, of the Barberton (South Africa) and East Pilbara (Western Australia) regions. The oldest reputed stromatolites, from the 3.48 Ga Dresser Formation (Warrawoona Group, Pilbara), are domical laminated structures made of chert, barite, and iron oxide-rich laminae with minor carbonate (Figure 9; Dresser Formation stromatolites). The biological origin of these structures has been contested, with claims made that they could be artefacts of structural deformation (folding) and other post-burial processes. Moreover, the fact that these stromatolites occur in isolated clusters within structurally deformed rocks makes it difficult to acquire supporting contextual information.

Also in the Warrawoona Group lies the 3.46 Ga Strelley Pool Chert, a formation that contains abundant, well-preserved stromatolites (Figure 10; Strelley Pool Chert stromatolites). The stromatolites in the Strelley Pool Chert display a range of conical morphologies that grade into small ripple structures, suggesting that the stromatolites formed in a shallow environment intermittently washed by gentle waves. However, several abiological hypotheses have been proposed for the formation of these stromatolites, including non-biological chemical precipitation and/or wave/current deposition. Nonetheless, the Strelley Pool Chert stromatolites offer good prospects in the search for early evidence of life. Even better prospects are offered by the younger, but more complex, stromatolites in the 2.7 Ga Fortescue Group. These examples may provide a useful benchmark for the study of other Archaean stromatolites (Figure 11; Fortescue stromatolites).

Microfossils

Fossilization processes Microbial fossilization occurs only under special circumstances, typically where mineral precipitation entombs the microbes as or soon after they die. This may occur in highly precipitative environments, such as hot springs, in which fluids laden with dissolved chemicals emanate at the Earth's surface. The mineral deposits associated with these environments are called sinters (silica precipitates) or travertine (carbonate precipitates). Modern examples occur at Yellowstone Park (USA), Taupo Volcanic Zone (New Zealand), and Lake Bogoria (Kenya). However, even in these environments, microbial fossilization is never complete and perfect. In multispecies systems, some species may be more readily fossilized, whereas others are destroyed. Furthermore, studies of modern sinters have shown that all cellular level information of colonies is lost on geologically short time-scales as opaline silica recrystallizes to a more stable quartz phase. In many sedimentary environments, microbial remains degrade or are completely obliterated almost immediately (in geological terms) through physical and chemical processes, including oxidation and thermal degradation (in which organic material breaks down to kerogen as sediments are heated through burial, igneous intrusion, or circulation of hydrothermal waters). Under most circumstances, the remains of microbes will not survive intact.

Interpreting microfossil-like structures Some researchers have suggested that, for the important question of the oldest evidence of life, the 'null hypothesis' should be adopted, which stipulates that a fossil cannot be authenticated until all abiological explanations (e.g., abiotic sedimentation, post-burial alteration, sample contamination) can be refuted. The

Table 4 Summary of selected data on Archaean stromatolites. Entries in italics are questionably stromalites or questionably Archaean. Colours correspond to date in Figure 2

Continent	Age (Ga)	Formation name	Map ref. (Figure 2A)[a]	Depositional association[b]	No. Different stromatolite types	Size range[c]	Morphological attributes[d]
North America	*2.6–2.1*	*Deep Rose Lake*	*16*	*Volcanic*	*1*	*M L*	*p, d*
	2.5–2.2	*Wildcat Hills Formation*	*15b*	*Volcanic*	*3*	*S M L*	*s, p, b*
	2.5–2.2	*Rawhide Canyon Formation*	*15a*	*Volcanic*	*1*	*S M L XL*	*p, d, m*
	2.6	Snofield Lake (Yellowknife Super group)	14	Volcanic	7	S M L	s, p, d, co, b, o, m
	2.6	Back River Volcanic Complex	13	Volcanic	3	M L	s, p, d
	2.6	*Angikuni Lake (Henik Group)*	*12*	*Volcanic*	*2*	*L*	*s, d*
	2.6	*Kaminak Group*	*11*	*Volcanic*	*1*		*s*
	2.7	*Sakami Lake*	*10*	*Volcanic*	*1*	*M*	*s?*
	2.7	*Selbaie Volcanics*	*9*	*Volcanic*	*1*		*s?*
	2.7	Joutel Volcanic Complex	8	Volcanic	3	ML	s, p, c
	2.7	Helen Formation	7	Volcanic	1	M	s? c
	2.7	Muskrat Dam Greenstone Belt	6	Volcanic	1	M	p
	2.8	Woman Lake Marble, unit 5D	5	Volcanic	2	M	s, d
	2.8	Eeyapamikama Lake (Keeyask Metasediments)	4	Volcanic	3	M	s, d, c
	2.9–2.7	Steeprock Group	3	Volcanic	8	S M L XL XXL	s, p, d, c, co, b
	2.9	Red Lake (Ball Assemblage)	2	Volcanic	3	M L XL	w, o, m
	2.9	*Lumby Lake Carbonate*	*1*	*Volcanic*	*1*		*s, p, d*
Europe	2.5–2.4	Bolshozero area	1			M	s?
Asia	*?*	*Slyudyanka Group*	*3*		*1*		*o?*
	2.9–2.7	Vanivilas Formation	2d	Volcanic	4	M L	s, p, d, b, o, c
	2.7–2.6	Sandur Belt	2c	Volcanic	2	M L	d, c?
	2.8–2.6	Shimoga Belt	2b	Volcanic	4	M L	s, p, d, c, co, w
	2.8–2.6	Kalche area (Dharwar Supergroup)	2a	Volcanic	4	M L	s, p, d, c, co, w
	3.2–3.1	Koira Group	1b	Volcanic	4	M L	s, p, d, o
	3.2–3.1	Bonai–Keonjhar area (Iron Ore Group)	1a	Volcanic	4	M	s, p, d, o
Australia	*2.4*	*Turee Creek Group*	*5*	Volcanic	*1*	*M*	*p, c*
	2.6	Carawine Dolomite	4	Volcanic	7	M L XL	s, p, d, c, co, o
	2.7–2.6	Kanowna	3	Volcanic	1	S M	c
	2.7	Tumbiana Formation	2	Volcanic lacustrine	9	S M L	s, p, d, c, co, b
	3.5	Strelley Pool Chert	1c	Volcanic	3	M L	s, c, o
	3.5	Panorama Formation	1b	Volcanic	1	M	s, c
	3.5	Dresser Formation	1a	Volcanic	4	M L	s, p, d, o
Africa	2.5	Ghaap Plateau Dolomite	7b	Volcanic	14	S M L XL XXL	s, p, d, c, co, b, w, r, m
	2.5	Malmani Dolomite	7a	Volcanic	10	XS S M L XL XXL XXXL	s, p, d, c, co, b, o
	2.7–2.6	Lannes Limestone	6e	Volcanic	1		s?
	2.7–2.6	Mt Hampdon Area (Bulawayo Supergroup)	6d	Volcanic	1	M	p, co
	2.6	Cheshire Formation	6c	Volcanic	4	S M L	s, p, d, c, b, m

Continued

Table 4 Continued

Continent	Age (Ga)	Formation name	Map ref. (Figure 2A)[a]	Depositional association[b]	No. Different stromatolite types	Size range[c]	Morphological attributes[d]
Africa	2.7–2.6	Manjeri Formation	6b	Volcanic	2	M	s, p
	2.7–2.6	Zwankendaba Group	6a	Volcanic	4	M L XL	p, d, c, o, m
	2.6	Bothaville Formation	5d	Volcanic, fluvial	2	S M	s, c
	2.6	Rietgat Formation	5c	Volcanic, lacustrine	3	S M L	p, d, b, o, m
	2.6	Klippan Formation	5b	Volcanic, lacustrine	1	M	s, p
	2.7	Contact Reef	5a	Volcanic	1	M	o?
	2.8	Mushandike Formation	4	Volcanic	2	S M	s, p, d, b
	2.9	Insuzi Group	3	Volcanic	6	S M L XL	s, p, d, c, co, b, m
	3.4	Witkop Formation	2	Volcanic	2	M L	p, d
	3.4	Kromberg Formation	1b	Volcanic	4	M L	s, p, c, o
	3.4	*Hooggenoeg Formation*	*1a*	*Volcanic*	*1*		*s?*

[a]The references in the fourth column correlate with the map in **Figure 2**.
[b]Depositional association refers to the type of overall geological setting in that the stromatolites occur in. Note that most stromatolites are found in volcanic settings; the actual deposits are typically hot spring type sediments such as exist today at Yellowstone National Park (USA), Lake Bogoria (Kenya), and Taupo Volcanic Zone (New Zealand).
[c]Size ranges: XS micrometres, S millimetres, M centimetres, L tens of centimetres, XL metres, XXL tens of metres, XXXL hundreds of metres.
[d]Morphology codes refer to: b branching, c conical, co columnar, d domed, m ministromatolite, o oncoids, p Pseudocolumnar, r rolled up, s stratiform, w walled.
Based on data from Hoffman H (2000) Archean Stromatolites as Microbial Archives. In: Riding R and Awramik S (eds.) *Microbial Sediments*, pp. 315 327. Berlin: Springer Verlag.

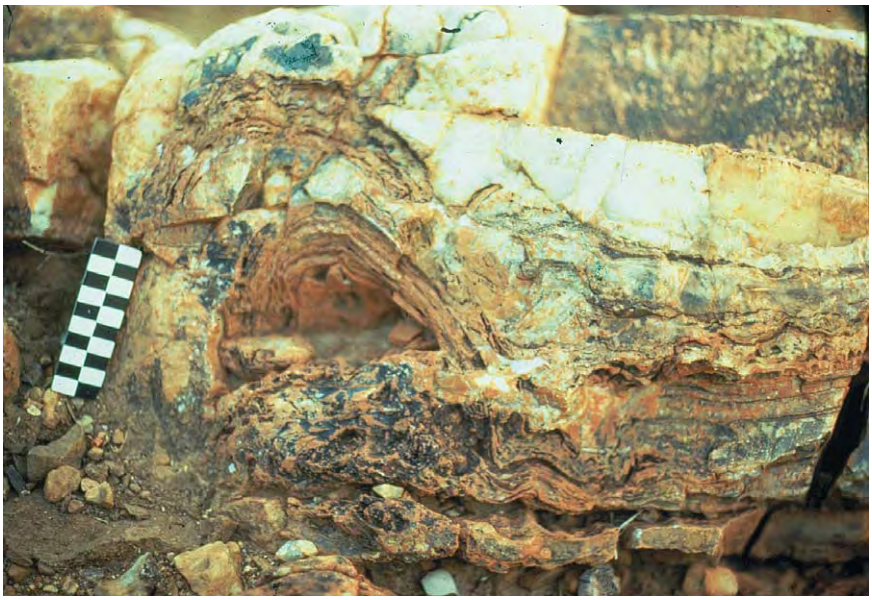


Figure 9 Domical stromatolite from the Dresser Formation, Pilbara region, Western Australia.

identification of microfossils is a topic of particular scrutiny in view of recent work demonstrating that microfossil-like objects can be reproduced abiotically in the laboratory. 'Biomorphs' made of silica and witherite (minerals that could occur in hydrothermal systems) were produced in the laboratory (**Figure 12**), displaying many features previously thought to be indicative of a biological origin, including spherical, segmented (septate), filamentous structures reminiscent of bacterial cells, and carbonaceous composition. Whereas kerogen is biological in origin, it should not be confused with

Figure 10 Conical stromatolite from the 3.42 Ga Strelley Pool Chert, Pilbara region, Western Australia. The layering comprises alternating fine chert and carbonate laminations. These are relatively flat, except in the stromatolite itself, where the laminae form a distinct cone. This kind of structure may be attributed to the activity of microbes.

non-biological organic carbonaceous material, which can form by a number of different processes, and has been shown experimentally to adhere to abiological fossil-like structures, lending them additional biological character.

The oldest microfossils The microfossils that were, for a long time, cited as the earliest evidence for life come from the Apex Chert in the Warrawoona Group (Pilbara region, Western Australia). This discovery comprised septate, filamentous structures (Figure 13) in what was thought to be a sedimentary chert layer. However, this chert layer was later shown to be a black chert dyke, which led to the contention that the putative microfossils could be hydrothermal artefacts or biomorphs (see above). Other carbonaceous structures interpreted as microfossils have been found in hydrothermal black chert deposits in the 3.49 Ga Dresser Formation (Warrawoona Group). Spheroidal structures of similar age, but more poorly preserved, have been reported from the Onverwacht Group in the Swaziland Sequence of South Africa. Although a biological origin for these discoveries is plausible, every report of microfossils from rocks of Early Archaean age has been contested and alternative abiotic explanations have been offered for their formation. Less contentious filamentous structures, interpreted as fossilized bacteria, have been found in the Late Archaean (2.7 Ga) Tumbiana Formation (Fortescue Group, Pilbara), but the oldest microfossils of widely accepted microbial origin come

Figure 11 Columnar stromatolites, exposed in cross section, from the 2.7 Ga Tumbiana Formation (Fortescue Group) in the Pilbara region, Western Australia. Hammer is approximately 40 cm in length.

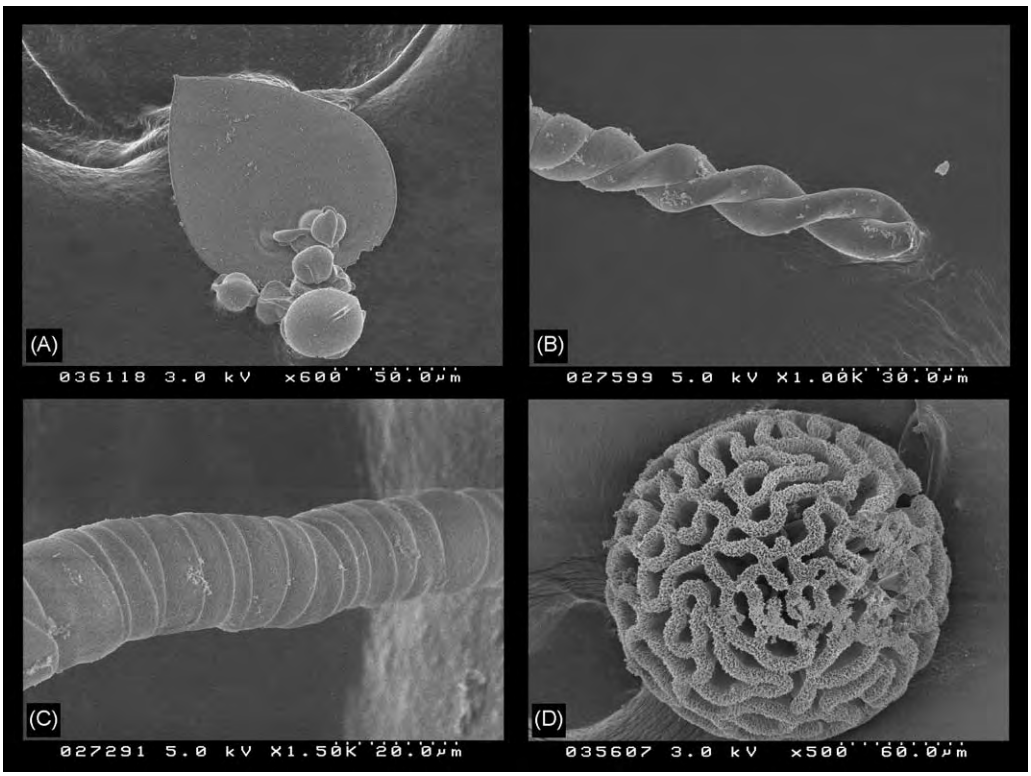


Figure 12 Scanning electron microphotographs of microfossil like structures that were produced abiologically in the laboratory. These 'biomorphs' display many complex characteristics that mimic structures of biological origin. Photographs courtesy of Stephen Hyde, Australian National University.

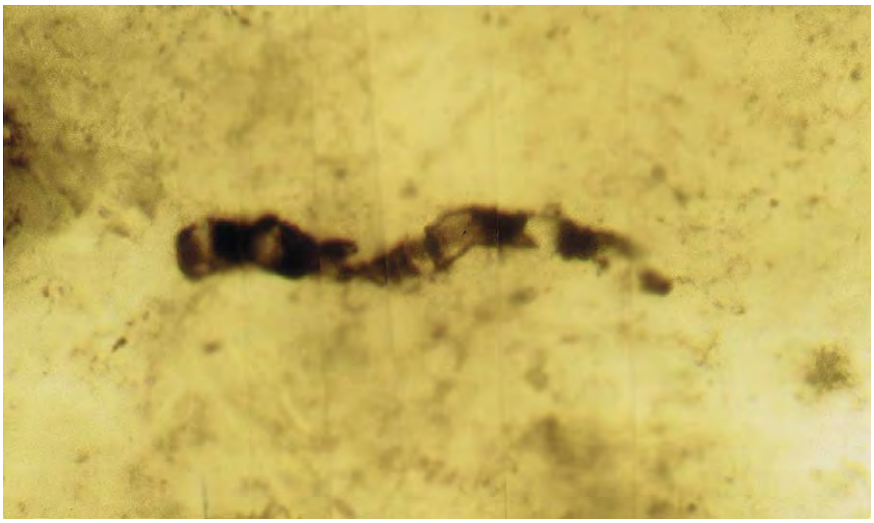


Figure 13 Putative microfossil from the Apex Chert, Pilbara region, Western Australia. Such structures have been interpreted as the fossilized remains of bacteria that flourished around 3.5 billion years ago. Photograph courtesy of J. W. Schopf.

from the 2.5 Ga Ghaap Plateau Dolomite in South Africa and the 2.1–1.8 Ga Gunflint Formation in western Ontario, Canada. These microfossils display coccoid (roughly spherical), septate filamentous, unbranched tubular, and budding bacteria-like structures (Figure 3).

Biomarkers

Biomarkers are biologically formed (biosynthesized) organic molecules that can be attributed to a specific biological origin. They possess molecular structures and isotopic characteristics that enable them to be distinguished from abiogenic organic compounds that

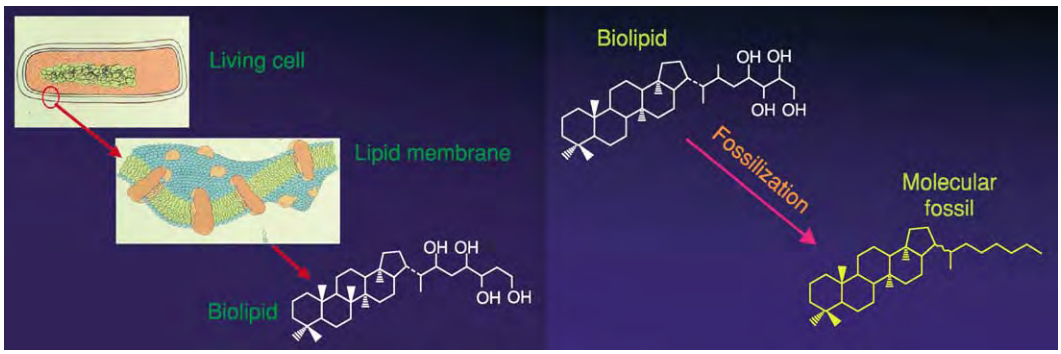


Figure 14 Schematic diagram of biomarker formation from the degradation of biological organic material. As the organic material is heated over millions of years, functional groups are lost from the molecular structure. In some cases, the remaining structure may still provide clues to the organism from which the organic material was derived. Adapted from original unpublished illustration by J. Brocks, with permission.

are distributed throughout the cosmos. In other words, they are a kind of molecular fossil that can: (1) provide evidence of life; and (2) provide information about the nature of those life forms, even in the absence of recognizable fossils. Figure 14 describes the formation of biomarker molecules as organisms degrade during burial and heating.

The oldest known biomarkers are from organic material in the 2.7 Ga Fortescue Group of the Pilbara region, Western Australia. The molecules were identified as originating from the group of organisms known as eukaryotes (*see* **Precambrian:** Eukaryote Fossils) (Figure 1).

Chemical Fossils

Microbial activity can be traced, in some instances, by the chemical signatures left behind in rocks. However, at present, our ability to detect these signatures may exceed our ability to interpret them. We know that microbes act as agents of dispersion, concentration, or fractionation of different chemical components and, in doing so, may impart specific chemical signatures upon their environment. However, we also know that similar patterns may be produced by abiological mechanisms. In the following paragraphs, we investigate some of the chemical signatures that can be used to trace biological activity, and some of the proposed problems with their interpretation.

Biological isotopic fractionation signatures form when organisms discriminate between light and heavy isotopes of elements, such as carbon, hydrogen, nitrogen, and sulphur. Both the light and heavy isotopes are involved in abiotic reactions, whereas organisms that utilize carbon compounds in their metabolic reactions almost always prefer the lighter ^{12}C isotope. The carbonaceous remains of the organisms are enriched in ^{12}C, whereas the source material from which they derived their carbon is enriched in ^{13}C. Carbon-bearing minerals (e.g., calcium carbonate or limestone) that precipitate from the residual ^{13}C-enriched material will acquire the heavy carbon signature. Other known biological fractionation processes involve the preferential assimilation of ^{14}N over ^{15}N, ^{32}S over ^{34}S, and hydrogen over deuterium. Thus, isotopic analyses of organic matter and their host mineral deposits can yield evidence of biological activity through fractionated C, N, S, and H signatures.

Carbon isotope fractionation patterns have been found in Early Archaean carbon-bearing rocks, such as the Strelley Pool Chert and Dresser Formation (Pilbara, Western Australia) and the Buck Reef Chert (Barberton, South Africa). Carbonaceous material (biogenic material?) with low values for $^{13}C/^{12}C$ from these formations may have derived from the biological fractionation of carbon isotopes. However, Fischer–Tropsch synthesis has been proposed as an alternative mechanism for carbon isotope fractionation. Fischer–Tropsch synthesis is a chemical reaction process involving carbon that is thought to occur in the mantle in the presence of certain catalytic compounds, such as Fe and Mn. Because Fischer–Tropsch synthesis is a high temperature process, and requires the presence of certain compounds, the geological setting of a fractionated C isotope signature may provide clues to differentiate Fischer–Tropsch-type occurrences from biological occurrences. However, the distinction is not always easily made and controversy persists with regard to the biological interpretation of light carbon signatures in Early Archaean rocks.

Beyond the morphological fossil record, which extends back to around 3.5 Ga, evidence for life may only be found via chemical fossils in *ca.* 3.8 Ga rocks from Greenland. These rocks are too metamorphosed to contain any morphological remains. However, carbon within the rocks is enriched in the light isotope, ^{12}C. This enrichment may have resulted from the biological fractionation of carbon, and has

thus been interpreted as the oldest evidence for life on Earth. However, in the absence of original sedimentary features and context, the interpretation of these signatures is contentious.

Clearly, the fossil that sets the irrefutable starting point for life on Earth has not been found. However, if the interpreted complexity of microbial communities seen in the fossil record at around 3.5 Ga is anything to go by, it seems likely that life on Earth began well before then. The remains of the very first microbes, or the rocks that they were part of, have been destroyed or metamorphosed beyond recognition.

Glossary

biogenic Of biological origin.
Ga Billion years before present/billion years of age.
lacustrine Associated with lake environments.
metamorphism Pressure- and temperature-related alteration of rocks during burial (involving processes that result in new textural and mineralogical characteristics).

See Also

Diagenesis, Overview. **Geysers and Hot Springs**. **Origin of Life**. **Precambrian:** Eukaryote Fossils; Vendian and Ediacaran. **Pseudofossils**. **Sedimentary Rocks:** Limestones.

Further Reading

Banfield JF and Nealson KH (1997) *Geomicrobiology: Interactions Between Microbes and Minerals*. Washington: Mineralogical Society of America.

Brocks JJ and Summons RE (2003) Sedimentary hydrocarbons, biomarkers for early life. In: Holland HD and Turekian K (eds.) *Treatise in Geochemistry,* pp. 63 115. Amsterdam: Elsevier.

Buick R (1990) Microfossil recognition in Archaean rocks: an appraisal of spheroids and filaments from a 3500 M.Y. old chert barite unit at North Pole, Western Australia *Palaios* 5: 441 459.

Ehrlich HL (2000) *Geomicrobiology*. New York: Marcel Dekker.

Grotzinger JP and Knoll AH (1999) Stromatolites in Precambrian carbonates: evolutionary mileposts or environmental dipsticks? *Annual Reviews, Earth and Planetary Science Letters* 27: 313 358.

Hoffman HJ (2000) Archaean stromatolites as microbial archives. In: Riding R and Awramik S (eds.) *Microbial Sediments,* pp. 315 327. Berlin, Heidelberg: Springer Verlag.

Lowe DR (1994) Abiological origin of described stromatolites older than 3.2 Ga. *Geology* 22: 387 390.

Schopf JW (1983) *Earth's Earliest Biosphere*. Princeton, NJ: Princeton University Press.

Schopf JW and Packer BM (1987) Early Archaean (3.3 billion to 3.5 billion year old) microfossils from Warrawoona Group, Australia. *Science* 237: 70 73.

Walter MR (1976) *Stromatolites*. Amsterdam: Elsevier.

BIOZONES

N MacLeod, The Natural History Museum, London, UK

Introduction

The contemporary concept of biozone is that of a stratigraphical interval defined on the basis of its biotic content. This deceptively simple definition serves as the door to vast realms of complexity, access to which is important not only because those realms contain much of the early history of geology, but also because they encompasses the single most important set of tools yet devised for reconstructing earth history. The biozone concept is built on the work of William Smith, who first demonstrated the importance of fossils for establishing sequences of geological events. Smith's ideas were extended by many nineteenth-century geologists and were instrumental in creating the geological time-scale (*see* **Time Scale**): one of the greatest scientific achievements of that century. For the last 100 years, various biozone types – range biozones of diverse types, assemblage biozones, interval biozones, acme biozones – have been employed throughout the Earth sciences where they have proven their worth in fields as diverse as palaeoceanography, palaeogeography, palaeoecology, and palaeoclimate studies; wherever there is a need for relative time correlation. In the twenty-first century, use of biozones remains at the forefront of stratigraphical analysis, with methodological improvements being made through use of the concept in quantitative stratigraphy (*see* **Stratigraphical Principles**).

In order to understand the concept of a biozone, one must understand its development. As noted above, foundations for an understanding of biozones were laid by William Smith who, in 1796, realised that the rock layers cropping out around the southern English town of Bath always occurred in the same

superposed order and that individual strata contained the same types of fossils. This insight provided Smith, who was engaged in surveying the routes of canals, with a practical tool whereby he could predict what rock types would be encountered during a canal's construction. Smith's observations of fossils were particularly important to his method because strata that superficially appeared similar (e.g., limestones, sandstones, shales), but that contained different sets of fossils, could be distinguished from one another and assigned to their correct positions within the overall sequence (*see* **Sequence Stratigraphy**). Using this approach, Smith was, by 1799, able to reconstruct the basic sequence of strata from the British Coal Measures (Carboniferous (*see* **Palaeozoic:** Carboniferous)) to the Chalk (Cretaceous (*see* **Mesozoic:** Cretaceous)). Extension of this method to other regions resulted in Smith producing the first geological map in 1815, the fossil evidence for which was publishing in his 1817 book *Stratigraphical System of Organized Fossils*.

Smith's work contained two further insights that proved crucial to development of the biozone concept. The first of these was that even a single stratum could be subdivided on the basis of its fossil content. This observation established variations in lithological type and fossil content as independent of one another, with fossil content often providing the more refined basis for subdivision. The second was that fossils from the same lithological types in different parts of the sequence often resembled one another. This observation established that there was a relation between the general types of fossils that occurred in different depositional environments.

In 1808 Smith showed his collection of fossils arranged stratigraphically to members of the Geological Society of London. By the 1820s Smith's methods had been accepted by most geologists and extensions had begun to appear, most notably by the work of Georges Cuvier and Alexander Brongniart (in 1822) in establishing the stratigraphical sequence of the Paris Basin. Later, Gerard Deshayes (in 1830), Heinrich Georg Bronn (in 1831), and especially Charles Lyell (in 1833) extended Smith's concepts by formulating subdivisions of Tertiary strata based on the sequence of fossils alone. This represented an important step toward the conceptualisation of the biozone in that it demonstrated the independence of palaeontological and lithological observations. Since changes in fossil morphologies were arranged in a recognisable and predictable sequence over both time and space, careful comparison of fossils with established sequences allowed stratigraphers to infer relative time correctly. Sequences of lithological changes failed to exhibit a similarly predictable pattern of variation over time.

The term 'zone' had been used informally by a number of geologists in the early and mid-1800s to denote a vertical interval of strata or assemblage of co-occurring fossil species. For example, Alcide d'Orbigny (1842–1851) used the term 'stage' and 'zone' interchangeably as subdivisions of Jurassic strata based on their ammonite content (*see* **Mesozoic:** Jurassic). The modern concept of the biostratigraphic zone, however, can be traced to the writings of Albert Oppel (e.g., *Die Juraforation Englands, Frankreichs und des südwestlichen Deutchlands*, 1856–1858) who used the term to denote 'the constant and exclusive occurrence of certain species [that] mark themselves off from their neighbours as distinct horizons' (translated in Arkell, *The Jurassic System of Great Britain*). Critical to Oppel's biozone concept was its abstract nature. Oppel-type zones existed independently of variations in the local lithological or palaeontological succession.

The success of Oppel's formulation for establishing long-range correlations within Europe and even between Europe and North America was beyond question by the late nineteenth century. Controversy remained though, regarding just what Oppel-type zones represented. In particular, Thomas Henry Huxley pointed out that, whereas Oppel's biozones could be construed to document the identity of, or homotaxic, arrangement in remote stratigraphical successions, the extent to which such zones represented equivalent intervals of time within such successions was unclear. In terms of practicality, such distinctions rarely mattered. Biostratigraphical analysis using Oppel-type zones had been established as an accurate method for reconstructing the local, regional, and (at least in principal) global sequence of strata, all of which were understood to represent some measure of geological time. Nevertheless, interest in the relation between biozones and measures of absolute time remained unabated, especially insofar as estimates of absolute time were critical to support for the concept of uniformitarianism that was held to underpin so much of late nineteenth-century geological theory.

The tools for radiometric – later radioisotopic – dating (*see* **Analytical Methods:** Geochronological Techniques), this volume) that were developed in the early 1900s offered an empirical way to resolve this controversy. Development of these methods was later augmented with other physio-chemical techniques, including magnetostratigraphy, chemostratigraphy, isotopic stratigraphy, and most recently, orbital stratigraphy. It should be noted though, that all of these latter, supplementary methods are, to a greater or lesser extent, dependent on accurate biostratigraphical analysis before their unique properties can be exploited with confidence. What has been made

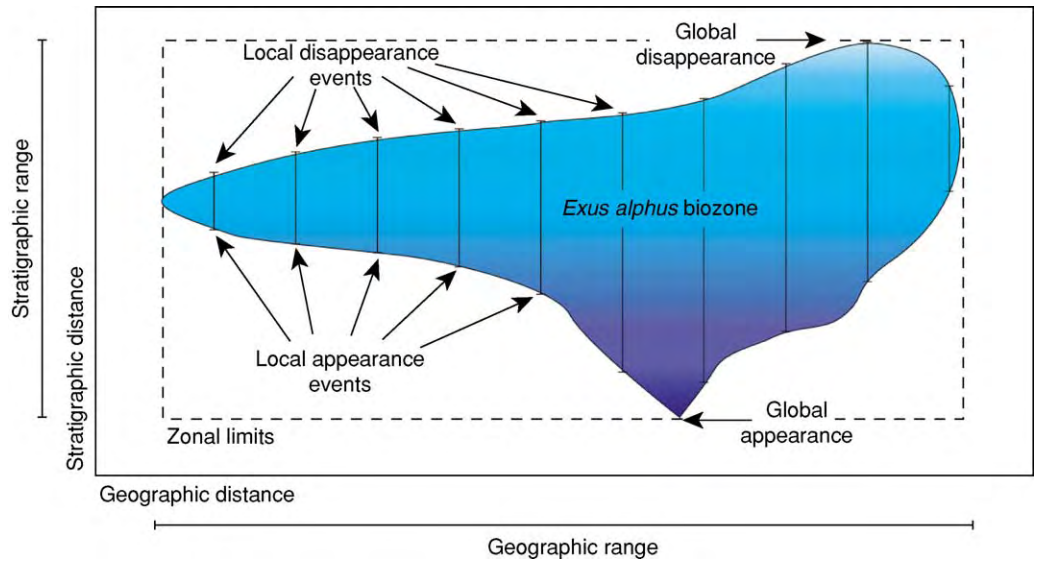


Figure 1 Relation between the spatio temporal concept of a biozone and various local and global events. A biozone is a three dimensional concept (two dimensions represented here) that encompasses all strata delimited by the zone's boundary definitions. All local identifications of the zone span the interval from the first appearance (equivalent to the appearance or stratigraphically lowest juxtaposition of defining criteria) to the last appearance (equivalent to disappearance or stratigraphically highest juxtaposition of defining criteria). The global maxima and minima of these juxtapositions also define a combined chronostratigraphical geographical interval within which the zone exists. Note that local first and last appearances of zone defining criteria are all diachronous. Note also that, even though the stratigraphic level or chronostratographic time represented in any local succession can, in principle, be correlated outside the geographic limits of the zone, the zone itself is unrecognizable outside these limits.

abundantly clear by the various methods of absolute dating, however, is that the boundaries between most biozones are often measurably diachronous on regional and global scales. This diachrony is often modest and, in many cases, can be ignored without fatally compromising the results of a biostratigraphical analyses. Nevertheless, biozones (Figure 1) are now conceptually recognised to have no necessary chronological implication in the sense that identification of the same biozone or biozone boundary cannot be taken as evidence of contemporaneous sediment deposition. This, in turn, has led to a proliferation of methods that can be used to combine various types of stratigraphical data into synthetic and mutually reinforcing comparison systems that can be used to make accurate geochronological assessments of stratigraphical successions (see below).

Types of Biozones

A wide variety of biozone types have been used to subdivide stratigraphical successions. Biozones are defined by single or combinations of biostratigraphical ranges for fossil species, genera, families, and so forth. Naturally, the ability to define and to recognise biozones is also strictly dependent on the existence of accurate systematic descriptions of fossil taxa such that individuals from locations remote from the primary study area can be quickly, easily, and correctly identified. A biostratigraphical range is the body of strata delineated by a fossil taxon's first appearance horizon or datum (FAD) and its last appearance horizon or datum (LAD, Figure 2). This 'first appearance' refers to the taxon's initial, or stratigraphically lowest appearance in a local succession while its 'last appearance' refers to the taxon's localised extinction at the stratigraphically highest level. Unfortunately, these first-lowest and last-highest conventions are usually reversed when working with wells that are being drilled, in which case the LAD is referred to as the 'first appearance' because it will be encountered before the FAD, and the FAD is referred to as the 'last appearance' because it will be encountered after the LAD.

Range Zones

A biostratigraphical range zone is a body of strata delineated by the total range of occurrences of any selected element(s) of the total fossil assemblage recovered from the body of strata (Figure 3). Although the conventional representation of a range zone is that of a one-dimensional interval (Figure 3A), a better mental image of a range zone is that of a two-dimensional body of rock delineated by the correlation of fossil distributions between stratigraphic successions (Figure 3B) or a three-dimensional

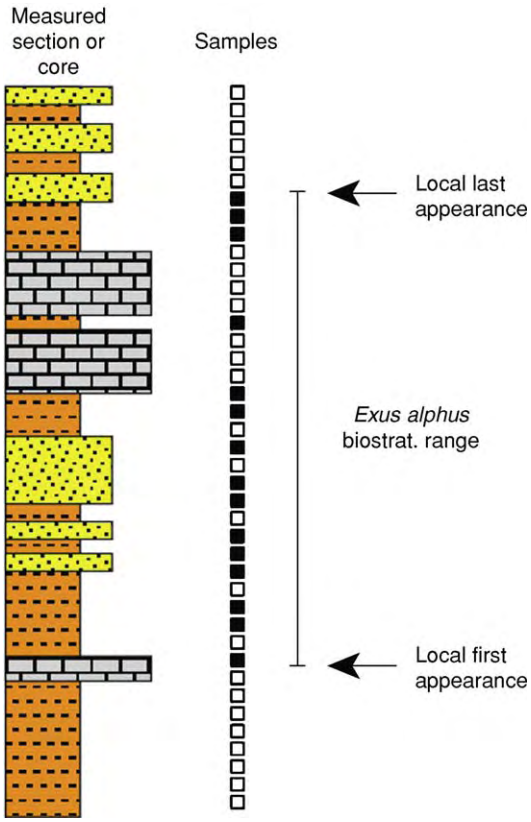


Figure 2 The manner in which the limits of a local biozone are established. Samples are taken up through a stratigraphic succession (or down through a core) and analyzed for their fo ssil content. The stratigraphically lowest sample containing a fossil species marks that fossil's local first appearance and the stratigraphically highest sample containing the species marks the local last appearance. The interval between these two datums represents the species' local biostratigraphic range. Note that gaps in the appearance pattern can occur within the species' biostratigraphical range. These gaps can be the result of many factors (e.g., low absolute abundance, small sample size, variable preservation, environmental exclusion, structural complications). The existence of such gaps means that the observed end points of each species' range represent minimum approximations of the true endpoints.

conceptualization that takes variation in occurrence and co-occurrence patterns across two geographical axes into account. This distinction between the conventional (one-dimensional) and actual (three-dimensional) conceptualization of biozone geometries holds true for all types of biozones. As better tools become available to reconstruct the three-dimensional geometries of stratigraphical bodies, the need for accurate three-dimensional concepts of biostratigraphical units will increase. There are four basic types of range zones: taxon range zones, concurrent range zones, Oppel zones, and lineage zones.

Taxon range zone A taxon range zone (also referred to as a teilzone, local zone, local-range zone, or topozone) is a body of strata delineated by the total occurrence range of any specimens belonging to a selected taxon (e.g., species, genus, family, Figure 3). Another way of thinking of this type of zone is as the region defined by all local appearances and extinctions of a selected fossil taxon (see Figure 1). While the geographic scope of such zones is set by the spatial occurrence pattern of their defining taxon, their boundaries are inherently diachronous insofar as the taxon's speciation and final extinction events will almost always vary from one locality to another (*see* **Mesozoic:** End Cretaceous Extinctions; **Palaeozoic:** End Permian Extinctions). Moreover, the boundaries of such zones are imprecise because taxon abundances usually diminish as both local and global speciation/appearance and extinction events are approached from within the zone. As a result, estimates of the local appearance and disappearance horizons of a taxon can be effected by sampling frequency and sample size near both ends of a taxon's stratigraphical range. Indeed, any abrupt appearance or disappearance of a taxon from a local succession is often taken as evidence of artificial truncation due to facies shift or the presence of a depositional hiatus.

Biostratigraphers estimate the boundaries of taxon range zones by assessing the boundaries of local range zones and correlating these boundaries between successions. By convention the names of taxon range zones refer to the defining taxon (e.g., *Dictyoha aculeata* Total Range Zone, *Parvularugoglobigerina* Total Range Zone).

Concurrent range zone A concurrent range zone (also referred to as an overlap zone or range-overlap zone) is a body of strata delineated by the those parts of the biostratigraphic ranges of two or more taxa that coincide in space and time (Figure 4). The term concurrent is a somewhat unfortunate choice for the name of this range-zone type since the colloquial definition of 'concurrent' is to 'happen at the same time'. This would, in principle, allow a concurrent range zone to be defined on the basis of two taxa that existed at the same time, but that never shared the same environment (e.g., a hypothetical *Tyrannosaurus rex–Abathomphalus mayaroensis* Concurrent Range Zone could be constructed as that interval of time represented by the overlap between these dinosaurian and planktonic foraminiferal species regardless of the fact that these two species have rarely, if ever, been found in the same stratigraphical succession). Such a zone would be of limited practical utility however, and the concept is understood to be restricted to those species whose stratigraphical ranges overlap in time and that occur together in the same samples.

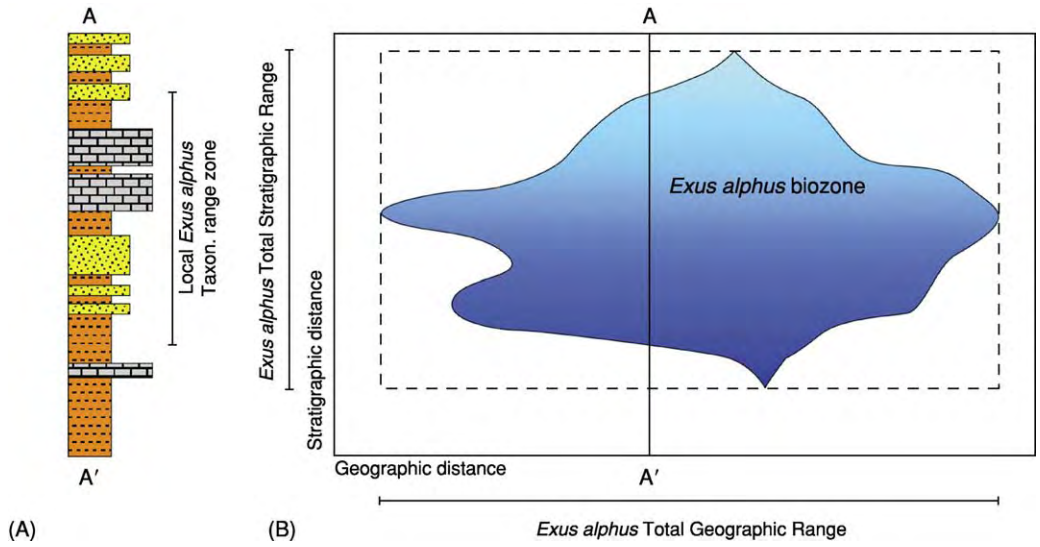


Figure 3 One dimensional (A) and two dimensional (B) representations of a taxon range biozone. The zone begins at the speciation event that gave rise to the taxon, encompasses all strata containing fossils assignable to the taxon, and ends at the taxon's global extinction horizon. Globally, these speciation and extinction events, in additional to the taxon's geographic limits of distribution, define a spatio temporal region within which the taxon's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizon tal) are defined (dashed box). In local sections however, the biozone will be represented by a one dimensional interval whose boundaries are included within the taxon's global biozone. Note the strong diachrony of the biozone boundaries in this example. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the taxon's range biochronozone (time interval represented by the Total Stratigraphic Range).

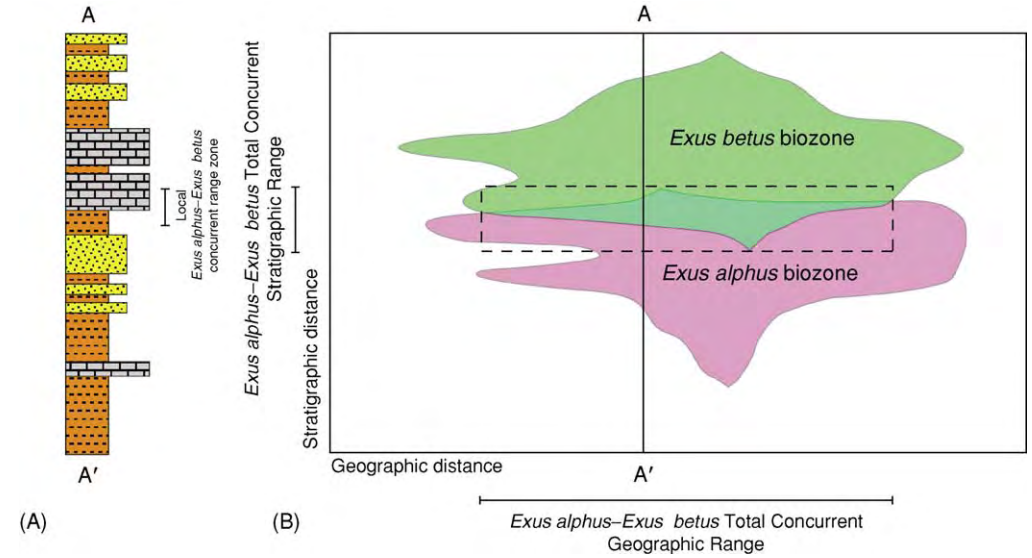


Figure 4 One dimensional (A) and two dimensional (B) representations of a concurrent range biozone. In this example the zone begins at the speciation event of the hypothetical species *Exus betus* and encompasses all strata containing fossils assignable to the taxon range zones of *Exus alphus* and *Exus betus*. Globally, the speciation event of *Exus betus*, the extinction event of *Exus alphus*, and the geographic limits of the concurrent distribution of these two species' range biozones define a spatio temporal region within which the taxon's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections however, the biozone will be represented by a one dimensional interval whose boundaries are included within this concurrent range biozone. Note the strong diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the taxons' concurrent range biochrono zone (= time interval represented by the Total Concurrent Stratigraphic Range). Other definitional geometries are also possible (e.g., two taxa, one of whose biozone is completely enclosed by the stratigraphic and geographic ranges of another) so long as the concurrent range criterion is respected.

The concurrent range zone concept is distinguished from the taxon range zone concept by being based on more than a single taxon, and from the assemblage zone concept by not including all elements of a naturally occurring assemblage of fossil organisms. Ideal taxa to define a concurrent range zone are those whose range overlap represents a unit with advantageous recognizability, similar ecological tolerance, wide geographic scope, and limited temporal duration. In order to distinguish the concurrent range zone from an Oppel zone, the presence of all zone-defining taxa should be used in recognising the zone, though this can serve to limit the concurrent range zone's scope. In practice, many biostratigraphers adopt a flexible approach to the use of defining taxa when recognising concurrent range zones, in which case the concept intrudes inevitably into the realm of the Oppel zone.

Properly defined concurrent range zones can have boundaries that are less diachronous than taxon range zones and less facies-controlled that many assemblage range zones. They also typically delineate finer stratigraphic and temporal intervals than taxon range zones. By convention, current range zones are named for the taxa that define them (e.g., *Globigerina selli–Pseudohastigerina barbadoensis* Concurrent Range Zone, *Neodenticula koizumii–Neodenticula kamtschatica* Concurrent Range Zone).

Oppel zone An Oppel zone is a body of strata delineated by the ranges and range-limits of a group of fossil taxa selected in such a way as to minimize zonal boundary diachrony and maximize the geographic scope of the interval so defined (Figure 5). This type of zone was made popular by Albert Oppel in studies of Jurassic stratigraphy in the mid-1900s and served as a 'type' example of a biozone in the days when distinctions between biostratigraphy and chronostratigraphy were less clearly drawn. Oppel even made use of certain lithological beds (e.g., 'spongy limestones') that he felt had chronostratigraphic significance in defining his zones.

Many stratigraphic reference works still list the Oppel zone as the 'preferred' or 'most useful' type of biozone; particularly those written by biostratigraphers specializing in invertebrate macrofossil taxa. This author's experience, however, suggests that in modern, high-resolution stratigraphical applications, and especially in the context of microfossil zonations which represent the contemporary stratigraphic standard for Jurassic through Recent marine sediments, explicit Oppel zones are rarely established in

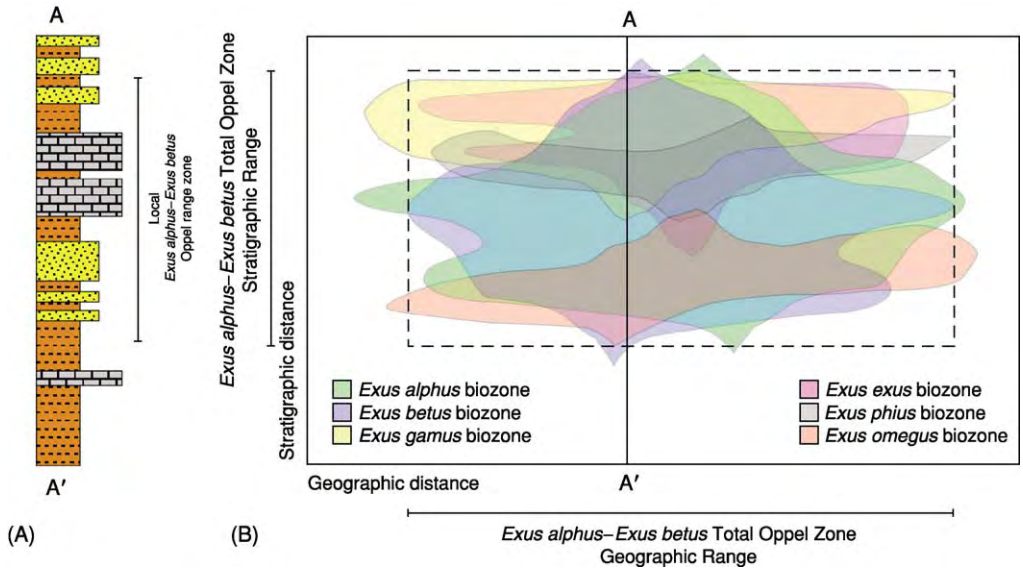


Figure 5 One dimensional (A) and two dimensional (B) representations of an Oppel range biozone. In this example the zone begins at the speciation event of the hypothetical species *Exus omegus* (representing the lowest interval of a concurrent range based on *Exus alphus, Exus betus*, and *Exus omegus*), and encompasses all strata containing fossils assignable to the upper limit of the *Exus alphus Exus beta Exus gamus* concurrent range zone. As is always the case with Oppel type biozones, these boundary definitions are chosen so as to enhance the ability of this zone to approximate a biochronozone. Globally, these two concurrent range biozones define a spatio temporal region within which the Oppel biozone's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections however, the biozone will be represented by a one dimensional interval whose boundaries are included within this total Oppel biozone. Note the relatively modest diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the corresponding biochronozone (= time interval represented by the Total Stratigraphic Range). Many other definitional geometries are also possible.

principle. Practice can be different though and, with the rise of neocatastrophism in the latter decades of the twentieth century, the common usage of certain standard microfossil zones has acquired a distinctly Oppelian character (e.g., the lowermost Paleocene planktonic foraminiferal Zone P0, also called the *Guembelitria cretacea* Zone) (*see* **Tertiary To Present:** Paleocene).

Oppel zones are more like abstract or Gestalt constructs than geometric juxtapositions of biostratigraphic ranges. They are largely constructed from sets of coincident range segments among taxa selected for the expressed purpose – or under the unexpressed assumption – of representing chronostratigraphic intervals. Unlike concurrent range zones (*sensu stricto*), Oppel zones do not rely on the necessary presence of all defining taxa to be recognised. Uniquely, this zonal concept embraces the idea that the biostratigrapher's judgement of the chronostratigraphical value of a particular taxon's presence (or absence) can be used in recognizing the zone. The rationale underlying this concept is that chronostratigraphical correlations are desirable and that, through long years of patient study, biostratigraphers can become sufficiently familiar with subtle morphological signals within their fossil faunas and floras that a higher level of time-based correlation than that afforded by any other zone concept can be achieved. The danger, of course, is that recognition of such a zone becomes an exercise in aesthetics with no obvious way to adjudicate judgemental disagreements between equally experienced stratigraphers. Interestingly, many – if not most – biostratigraphers approach the identification of fossil species in a similar manner. Employment of this Gestalt aesthetic in systematic palaeontology has led to the realisation that the reproducibility of many species lists is quite low. One cannot help but suspect that Oppel-zone recognisability is likely beset by similar problems.

By convention, Oppel zones are named for the dominant taxon used to recognise the zone's presence (e.g., *Subcolumbites-Prohungarites* Oppel Zone, *Globigerina selli–Pseudohastigerina barbadoensis* Oppel Zone).

Lineage zone A lineage zone (also referred to as a phylozone, evolutionary zone, lineage zone, morphogenic zone, phylogenetic zone, or morphoseries zone) is defined as a body of strata delineated by the combined biostratigraphic range of a series of ancestor–descendant species couplets (Figure 6). As such, a lineage zone is a type of taxon range zone in which

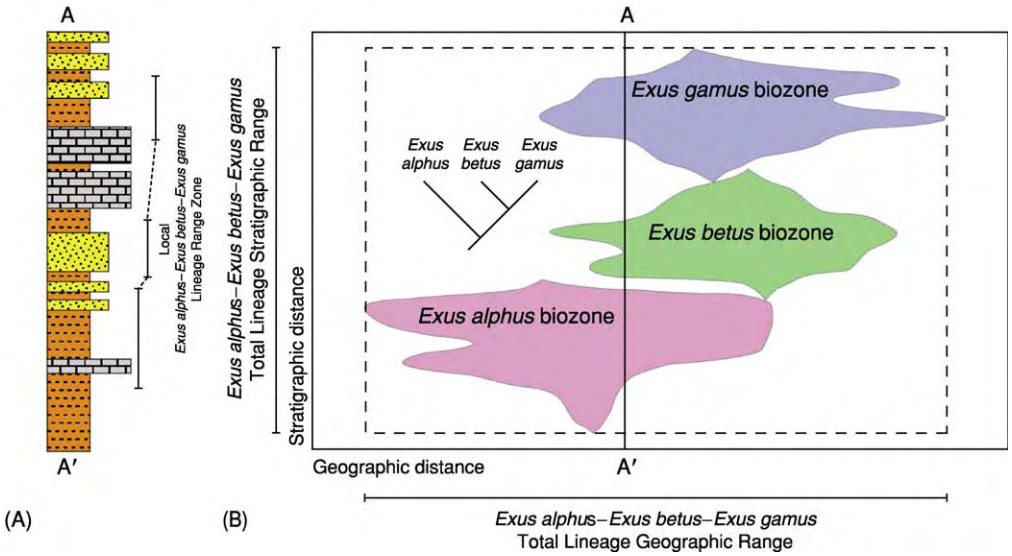


Figure 6 One dimensional (A) and two dimensional (B) representations of a lineage range biozone. In this example the zone begins at the speciation event of the hypothetical species *Exus alphus* (representing the oldest species of the lineage) and encompasses all strata containing fossils assignable to the lineage. In this example, the stacked taxon range biozones for a three species lineage define the overall lineage biozone. Globally, these taxon range biozones define a spatio temporal region within which the lineage's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections, however, the biozone will be represented by a one dimensional interval whose boundaries are included within the range of the global biozone. Note the strong diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the corresponding biochronozone (time interval represented by the lineage's Total Stratigraphic Range). Lineage range zones also have the added uncertainty that, since lineages are inferred not observed, their composition may change with new evidence and/or analyses.

the 'taxon' concept has been replaced by the evolutionary concept of a lineage or lineage segment. Given the controversial nature of attempts to identify ancestors in the fossil record using modern methods of phylogenetic systematics, one suspects that, were these systematic methods applied to some lineage zone-defining taxa, the basic rationale for recognizing many lineage zones might be called into question. Nevertheless, there are a number of instances in which morphological transitions between stratigraphically successive taxa are so well structured as to pass even the most stringent morphological tests.

A number of previous writers have commented that lineage zones should be particularly useful for chronostratigraphical correlations. I do not believe this to be the case. Lineage zones are prone to all the problems of taxon range zones (see above), and to the problem of being uniquely susceptible to the discovery of new fossils, or new characters that, by altering our understanding of phylogenetic relations between fossil species, would serve to change the zone's concept or rationale. A strictly comparative and geometric approach to the determination of FAD/LAD sequences in local sections, along with efforts to estimate reliable chronostratigraphic ages for those events based on global composite standard sections (see below), are likely to be of more practical use in achieving a chronostratigraphically-based biostratigraphy than efforts to mix phylogenetic inferences with biostratigraphic observations. Notwithstanding the issues highlighted above, lineage-based zonations remain popular, particularly among students of Tertiary planktonic microfossils where the tradition of qualitative phylogenetic analysis remains strong.

By convention, lineage zones are named for the lineages on which they are defined (e.g., *Globorotalia foshi* Lineage Zone, *Allevium praegallowayi–galloway-superbum* Lineage Zone).

Assemblage Zone

An assemblage zone (also referred to as a cenozone, ecozone, ecological zone, faunizone, biofacies zone, and association zone) is defined as that body of strata delineated by a natural assemblage of fossils (Figure 7). This assemblage can be defined in a variety of ways. For example, it may represent an ensemble of many groups (e.g., a coral–algal assemblage zone) or of a single group (e.g., a mollusc assemblage zone). It may also take a specific habitat into consideration (e.g., a planktonic foraminiferal assemblage zone), or cross habitat boundaries. The distinguishing character of this zone is that the assemblage be composed

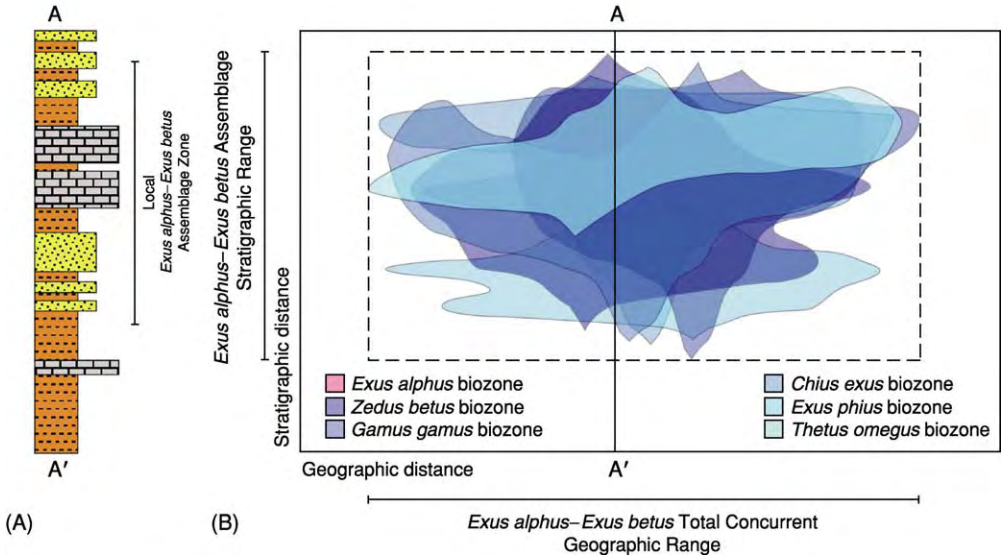


Figure 7 One dimensional (A) and two dimensional (B) representations of a assemblage range biozone. In this example the zone begins at the speciation event of the hypothetical species *Exus alphus* (representing the oldest species in an assemblage of six fossil species usually found together in the same environment) and encompasses all strata containing fossils assignable to this assemblage. As is always the case with assemblage range biozones, the boundary definitions are chosen so as to base the zone on a series of ecologically related taxa. Globally, this assemblage of taxon range biozones defines a spatio temporal region within which the assemblage biozone's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections, however, the biozone will be represented by a one dimensional interval whose boundaries are included within this global biozone. Note the strong diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the corresponding biochronozone (time interval represented by the Total Stratigraphic Range).

of fossils that actually lived together and (ideally) interacted with one another to the extent that the grouping itself achieved a degree of stability – and therefore recognizability – as a result of these interactions.

Because properly defined assemblage zones reflect ecological relations, this type of zone tends to be restricted to particular facies. The ability of such zones to track spatial shifts in environments through time provides them with a distinctive utility in terms of palaeoenvironmental analyses. However, this utility comes at a price, and that price is a relatively reduced ability to achieve long-distance chronostratigraphical correlations. For this reason, assemblage zones of benthic organisms tend to be most useful on local and regional scales. Assemblage zones of planktonic organisms do perform well in terms of chronological correlations, but the degree to which such organismal groupings are maintained by close inter-specific interactions is debatable.

Like Oppel zones, the specific criteria used to recognize assemblage zone boundaries are flexible. Not all of the groups present in the zone's 'type area' need be present to recognise the zone in remote locations. Unlike Oppel zones though, there is an objective and independent rationale underlying this definitional latitude. In the case of assemblage zones, one seeks to recognise a set of dependent ecological relations among species and between organisms and their environment that transcend mere faunal and/or floral lists. The objective reality of such patterns in nature is well established by numerous studies of modern faunas and floras and is reasonably well understood from a theoretical point of view. Oppel zones, on the other hand, are unified only in the vague sense that the species used to recognize the zone are thought to be useful in chronostratigraphical analysis. Although there is certainly ample justification for suspecting that, in many cases, the biostratigraphic ranges of the species in different regions and habitats will coincide, there is much less justification for regarding these organisms as part of transcendent causal association than is the case with assemblage zones.

As a result of the flexible manner in which assemblage zones are defined, the same groups can be used to define different assemblage zones (e.g., a coral–bryozoan assemblage zone and a coral–foraminiferal assemblage zone can have zone-defining taxa in common) and different members of the same ecological association can be used to define different assemblage zones. Assemblage zones have been used frequently in areas where suitably short-ranging taxa are not present or have not been studied.

By convention, the name of an assemblage zone should be based on two or more taxa that figure prominently in the zone's definition (e.g., *Eponides–Planorbulinella* Assemblage Zone, *Eodicynodon* Assemblage Zone).

Interval Zone

An interval zone (also referred to as an interbiohorizon zone, gap zone, or a partial-range zone) is defined as a body of strata delineated by the region between two distinctive biostratigraphic horizons, but that has no distinctive biostratigraphic identity of its own (**Figure 8**). The boundaries of interval zones can be marked by a wide variety of criteria. These zones typically represent the undefined regions between other types of zones; especially taxon range zones.

An interval zone's existence assumes a complementary relation with the underlying and overlying biostratigraphically defined horizons that serve as their inferior and superior boundaries. As with all other types of biozones, the traditional one-dimensional concept of biozone geometry (**Figure 8A**) can mask the more complex geometries evident in two and three-dimensional conceptualizations (**Figure 8B**). In particular, interval zones are confined geographically to only those regions in which the defining biozones overlap. So long as one's region of interest is confined to the geographical area encompassed jointly by the zone's defining taxa, recognition of the zone can be made with confidence. Outside this geographic envelope, though, recognition of an interval zone becomes problematic, if not impossible.

By convention, interval zones are either named for the taxa used to define their boundaries (e.g., *Globigerinoides sicanus–Orbulina suturalis* Interval Zone) or for a taxon that occurs in the interval, but is not itself used in the zone definition (e.g., *Globigerina ciperoensis* Zone).

Acme Zone

An acme zone (also referred to as a peak zone, flood zone, or epibole) is defined as a body of strata delineated by the region of 'maximal development' of a taxon (e.g., species, genus, family), but not its total range (**Figure 9**). In this context, the term 'maximal development' is meant to be used flexibly. In some cases it might refer to an initial increase and subsequent decrease in the relative abundance of a taxon that takes place within the confines of its biostratigraphic-geographic range. In others, it might refer to an increase/decrease in body size, an increase or decrease in diversity, etc. Since these aspects of a taxon's evolutionary/ecological history tend to be strongly associated with local and regional conditions, it is on these spatial scales that acme zones have their

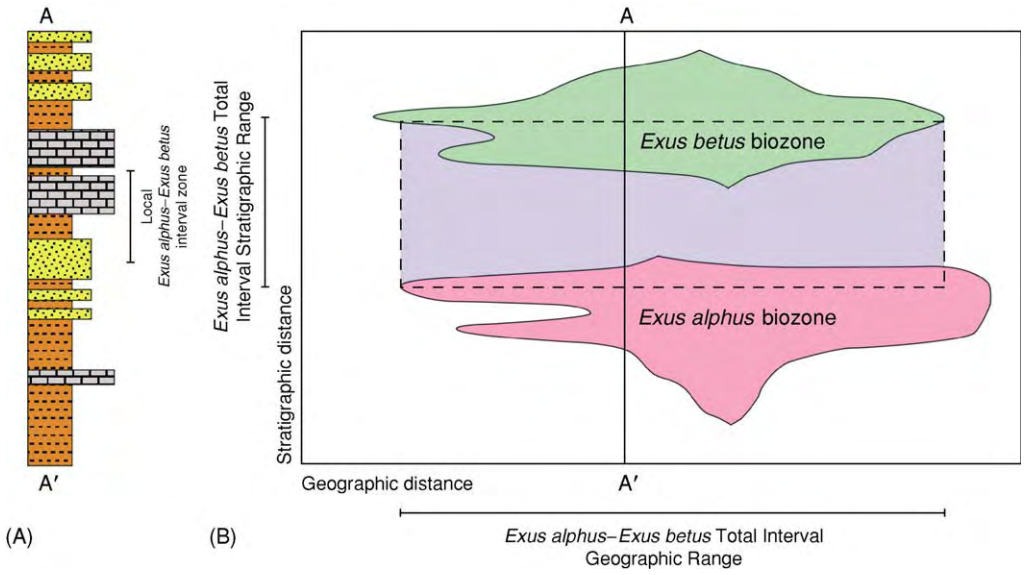


Figure 8 One dimensional (A) and two dimensional (B) representations of an interval biozone. In this example the zone begins at the geographic acme of the hypothetical species *Exus alphus* and encompasses all strata between this extinction datum and the overlying *Exus betus* geographic expansion datum. Globally, this interval between taxon range biozones defines a spatio temporal region within which the interval biozone's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections however, the biozone will be represented by a one dimensional interval whose boundaries are included within this global biozone. Note the strong diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the corresponding biochrono zone (time interval represented by the Total Stratigraphic Range). Many other definitional geometries are also possible (e.g., interval between two species' last appearances, interval between two species' first appearances).

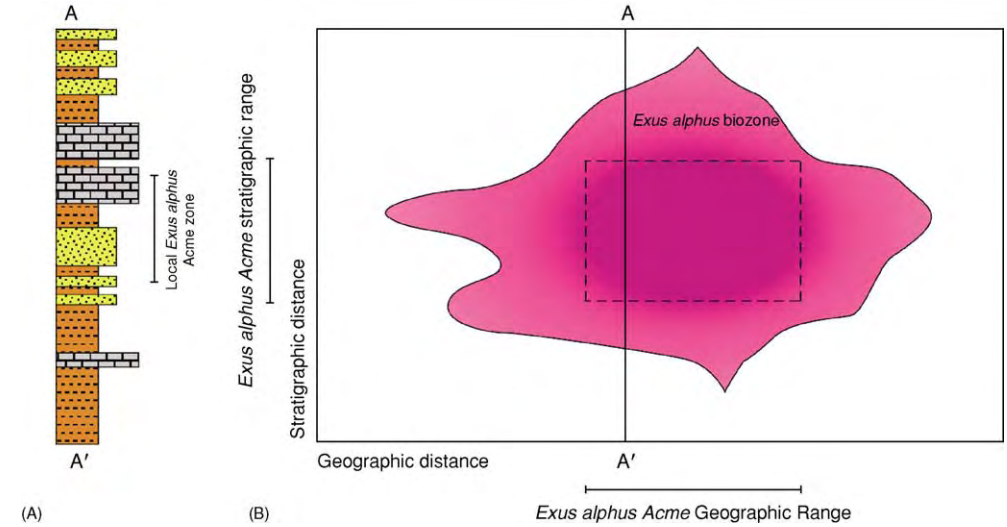


Figure 9 One dimensional (A) and two dimensional (B) representations of an acme biozone. In this example the zone begins at the first substantial increase in the relative abundance of the species (symbolized by dark shading) of the hypothetical species *Exus alphus* and encompasses all strata in which the species is regarded as abundant. Globally, these abundance based data define a spatio temporal region within which the interval biozone's inferred Total Stratigraphic Range (vertical) and Total Geographic Range (horizontal) are defined (dashed box). In local sections, however, the biozone will be represented by a one dimensional interval whose boundaries are included within this global biozone. Note the strong diachrony of the biozone boundaries. Intervals of strata that lie outside the biozones' geographical boundaries cannot be referred to the biozone *per se*, but can be placed within the corresponding biochronozone (= time interval represented by the Total Stratigraphic Range). Many other definitions of acme are also possible (e.g., increased size, increased ornamentation).

greatest utility. Care must be taken when defining such zones, that the recognition criteria employed are as explicit and objective as possible. Nevertheless, the popularity of such zones in regional basin analyses (particularly in commercial biozonations) stands as testimony to their practical utility. By convention, interval zones are named for the taxon used in its definition (e.g., *Emiliana huxleyi* Acme Zone, *Tylosaurus* Acme Zone).

Other Types of Biozones

Although the foregoing descriptions represent the most frequently used biozone types, other types do exist. These include such exotica as barren zones, coiling-direction zones, negative association zones, species pre-lap zone, species post-lap zone, etc. In addition, biozone types exist that are not defined on fossils *per se*, but rather on the trace fossils left in the sediment as a result of animal and plant activity (e.g., track zones). All such zones are valid only to the extent that they are useful to the stratigrapher and geologist, and only to the extent that their definition is based on unambiguous observational evidence. As was noted by Hedberg in 1971, 'Time and usage will then be the surest means of determining whether it is desirable for such [zones] to persist'.

Biozones and Biochronozones

Ever since the days of Smith, Lyell, and Oppel, the purpose of establishing biozones has been to achieve chronostratigraphical correlations. Other types of stratigraphic intervals (e.g., key beds, magnetochrons, isotope zones) may be based on boundaries whose emplacement is effectively synchronous over geological time-scale. None is uniquely identifiable in the same manner as biozones, however. Indeed, the use of biostratigraphic methods to place, in an approximate temporal framework, the observations upon which these other types of chronostratigraphic inference depend, is so common it is rarely even mentioned. This failure to give biostratigraphical data due credit for the basic contribution it makes would not be so unfortunate, were it not for the fact that biostratigraphy has come to be viewed by many as a lacklustre and unimaginative field of study. Nothing could be further from the truth. Biostratigraphic data underpin not only most of what we know about geological time over the last 600 million years, but are critical to studies of palaeogeography, palaeoecology, extinctions, diversifications, and patterns of morphological evolution. In order to remind oneself of the importance of biostratigraphical data, it is worth recalling that the lack of a detailed time-scale in which to place Proterozoic and Archaean events is not due to a lack of palaeomagnetic reversals, catastrophic depositional events, isotopic excursions, etc., but rather to the lack of fossil biotas on which to base a coherent, finely resolved, and widely accessible stratigraphy.

This having been said, the distinction between biozones and chronozones must always be borne in mind. All biozone boundaries are diachronous by definition and, without recourse to radioisotopic

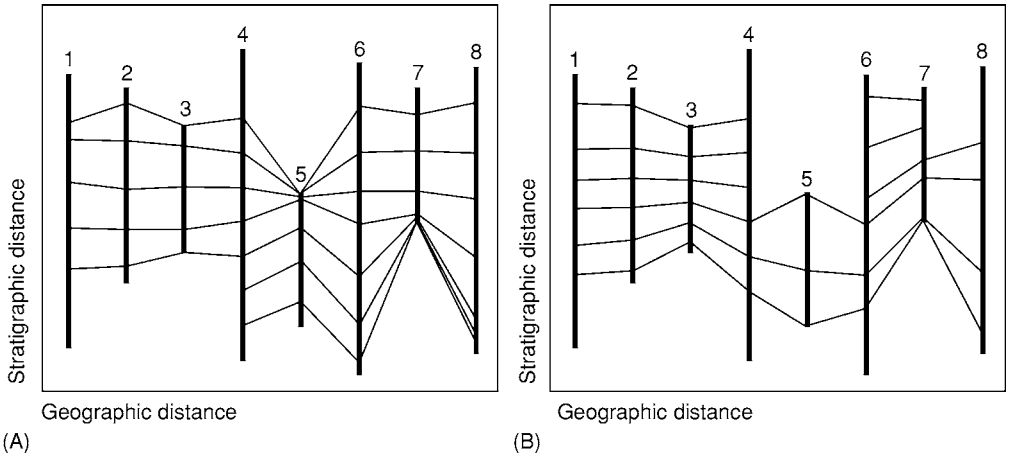


Figure 10 Results of multivariate biostratigraphic analyses on simulated biostratigraphic data from various geological sections numbered 1 to 8 using the graphic correlation (A), as implemented by constrained optimization) and ranking and scaling (B) approaches. These methods focus on the estimation of global, composite, taxon range biozones from datasets containing a series of biostratigraphic first and last appearance datums for two or more stratigraphic successions or cores. Such approaches offer more consistency, objectivity, and higher reproducibility of results than either qualitative or pairwise quantitative approaches to biostratigraphic analysis. In these diagrams, connecting lines between the eight sections represent best estimate lines of biochronostratigraphic correlation.

data, it is never correct to regard local biozones as being equivalent to chronozones. Even when the physical data necessary for absolute dating are available, it is no simple matter to accurately interpolate the ages of biozone boundaries from radioisotopically dated horizons.

Fortunately, several methods exist whereby stratigraphers can compare sets of biostratigraphical observations in different successions or cores and, through use of a few simple rules, create a summary of the chronostratigraphically relevant data contained in all such sections/cores. Graphic correlation (Figure 10A) and ranking and scaling (Figure 10B) are two of the more common quantitative methods used for bridging the gap between observational biostratigraphies and the ideal of a 'biochronostratigraphy'. The ongoing incorporation of raw biostratigraphical data into internally consistent and synthetic composite successions based on the application of these methods, along with improvements in consistent identifying fossil morphologies (e.g., via automated object recognition) promise to take the science of biostratigraphy, and its foundation concept – the biozone – into the twenty-first century. There it will, once again, serve as the single most useful contrivance in the practical stratigrapher's toolkit.

Glossary

Constrained optimization A mathematical technique involving the search for the optimal condition or structure of a system of observations given one or more consistently applied external rules.

Depositional hiatus A horizon within a body of sedimentary rock that represents a gap in time due to the nondeposition of sediment, active erosion, or structural complications.

Diachrony The condition of taking place at different times.

Facies A stratigraphic body distinguished from other such bodies by a difference in appearance or composition.

Homotaxis The condition of occupying the same position in a sequence.

Isochrony The condition of being created at the same time.

Microfossil Fossil shells or other hard parts of tiny organisms or parts of organisms studied with the aid of a microscope.

Phylogenetic systematics A method of determining evolutionary relations between taxa based on the nested patterns of shared characteristics derived from the same pre-existing condition.

Stratum (strata) A tabular section of a rock body that consists of the same type of rock material.

Taxon (taxa) Generalized term for any level within the Linnean hierarchy of biological classification (e.g., order, family, genus).

See Also

Analytical Methods: Geochronological Techniques. **Mesozoic:** Triassic; Jurassic; Cretaceous; End Cretaceous Extinctions. **Palaeozoic:** Cambrian; Ordovician; Silurian; Devonian; Carboniferous; Permian; End Permian Extinctions. **Precambrian:** Vendian and Ediacaran. **Sequence Stratigraphy**. **Stratigraphical Principles**. **Tertiary To Present:** Paleocene; Eocene; Oligocene; Miocene; Pliocene; Pleistocene and The Ice Age. **Time Scale**.

Further Reading

Cubitt JM and Reyment RA (1982) *Quantitative strati graphic correlation*. Chichester: Wiley.

Gradstein FM, Agterberg FP, Brower JC, *et al.* (1985) *Quantitative stratigraphy*. Dordrecht: D. Reidel.

Hedberg HD (1971) *Preliminary report on biostratigraphic units*. Montreal, Canada: International Subcommission on Stratigraphic Classification.

Hedberg HD (1976) *International stratigraphic guide: a guide to stratigraphic classification, terminology, and procedure*. New York: John Wiley & Sons.

Kauffman EG and Hazel JE (1977) *Concepts and methods in biostratigraphy*. Stroudsburg, Pennsylvania: Dowden, Hutchinson & Ross, Inc.

Mann K, Lane HR, and Stein J (1995) *Graphic correlation*. Tulsa: Society of Economic Paleontologists and Mineral ogists, Special Publication 53.

Miall AD (1984) *Principles of sedimentary basin analysis*. New York: Springer Verlag.

Rawson PF, Allen PM, Brenchley PJ, *et al.* (2002) *Strati graphical procedure*. London: The Geological Society.

Reyment RA (1980) *Morphometric methods in biostrati graphy*. London: Academic Press.

Shaw A (1964) *Time in stratigraphy*. New York: McGraw Hill.

BRAZIL

F F Alkmim and M A Martins-Neto, Universidade Federal de Ouro Preto, Ouro Preto, Brazil

Brazil in the Geological Scenario of South America

The South American continent comprises five major tectonic units: the Pacific active margin, the Atlantic passive margin, the Andean Orogen, and the Patagonian and South American platforms (Figure 1). The continental margins and the Andean Orogen are the younger portions of the continent; the platforms, on the other hand, correspond to the mature and ancient parts of South America.

Brazil is located entirely on the South American platform, which is defined as the Precambrian core of the continent, not affected by the Andean orogenies (Figure 1). The exposures of the South American platform are collectively referred to as the Brazilian Shield, which in reality encompasses three distinct morphotectonic domains: the Guyanas, the Central Brazil (also called Guaporé) and the Atlantic shields (Figure 1). Sedimentary basins, including large Palaeozoic sags, Cretaceous passive and transform margins, and Tertiary rifts make up the Phanerozoic cover of the South American platform (Figure 1).

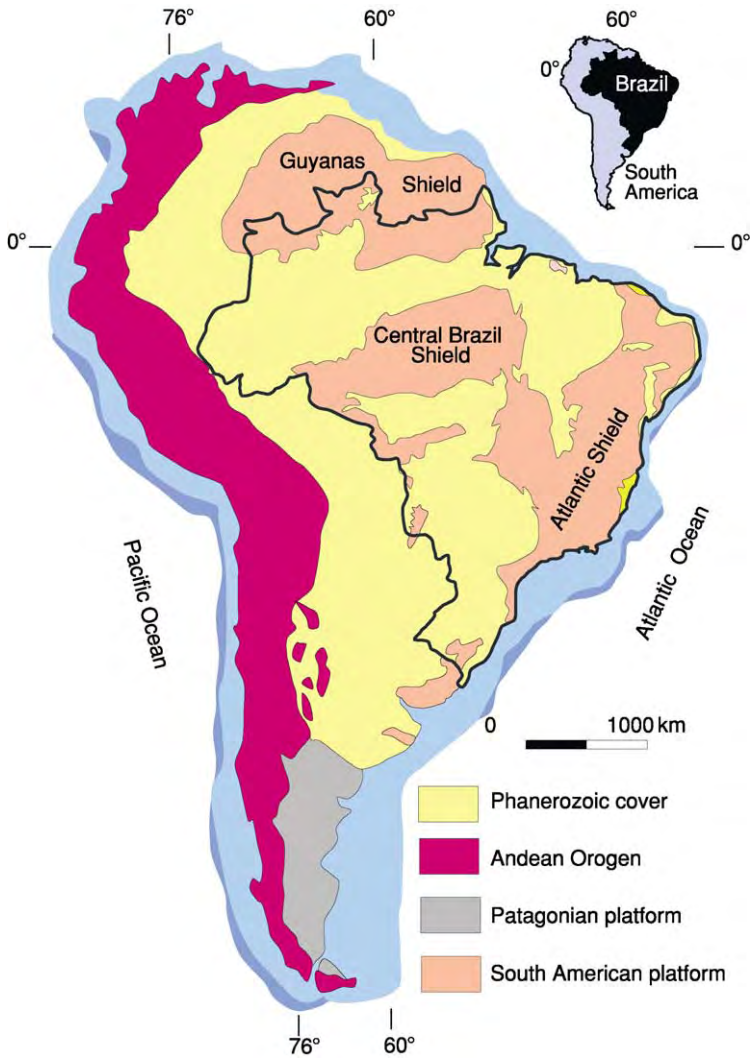


Figure 1 Tectonic map of South America, showing the main subdivisions. Modified from Almeida FFM, Hasui Y, Brito Neves BB, and Fuck RA (1981) Brazilian structural provinces: an introduction. *Earth Science Reviews* 17: 1 29, with permission from Elsevier.

Two fundamentally distinct lithospheric components, namely cratons and Neoproterozoic orogens, form the Precambrian core of South America. The cratons correspond to the relatively stable pieces of the continent that escaped the Neoproterozoic orogenies recorded in all the remaining shield areas of Brazil. Four cratons have been delimited in the South American platform: the São Francisco, Amazon, São Luis, and Rio de la Plata cratons (Figure 2). The noncratonic segments of the platform are the Neoproterozoic Mantiqueira, Tocantins, and Borborema orogenic domains (Figure 2). (In the Brazilian geological literature each component of the South American platform and its cover (i.e. cratons, Brasiliano orogens, and Phanerozoic basins) is referred to as a tectonic province.)

Together with Africa, the South American platform once lay in the western portion of Gondwana, the supercontinent assembled by the end of the Neoproterozoic and split apart in the Cretaceous, which also encompassed Antarctica, India, and Australia (Figure 3). The assembly of western Gondwana resulted from a series of diachronic collisions, predominantly between 640 Ma and 520 Ma, the

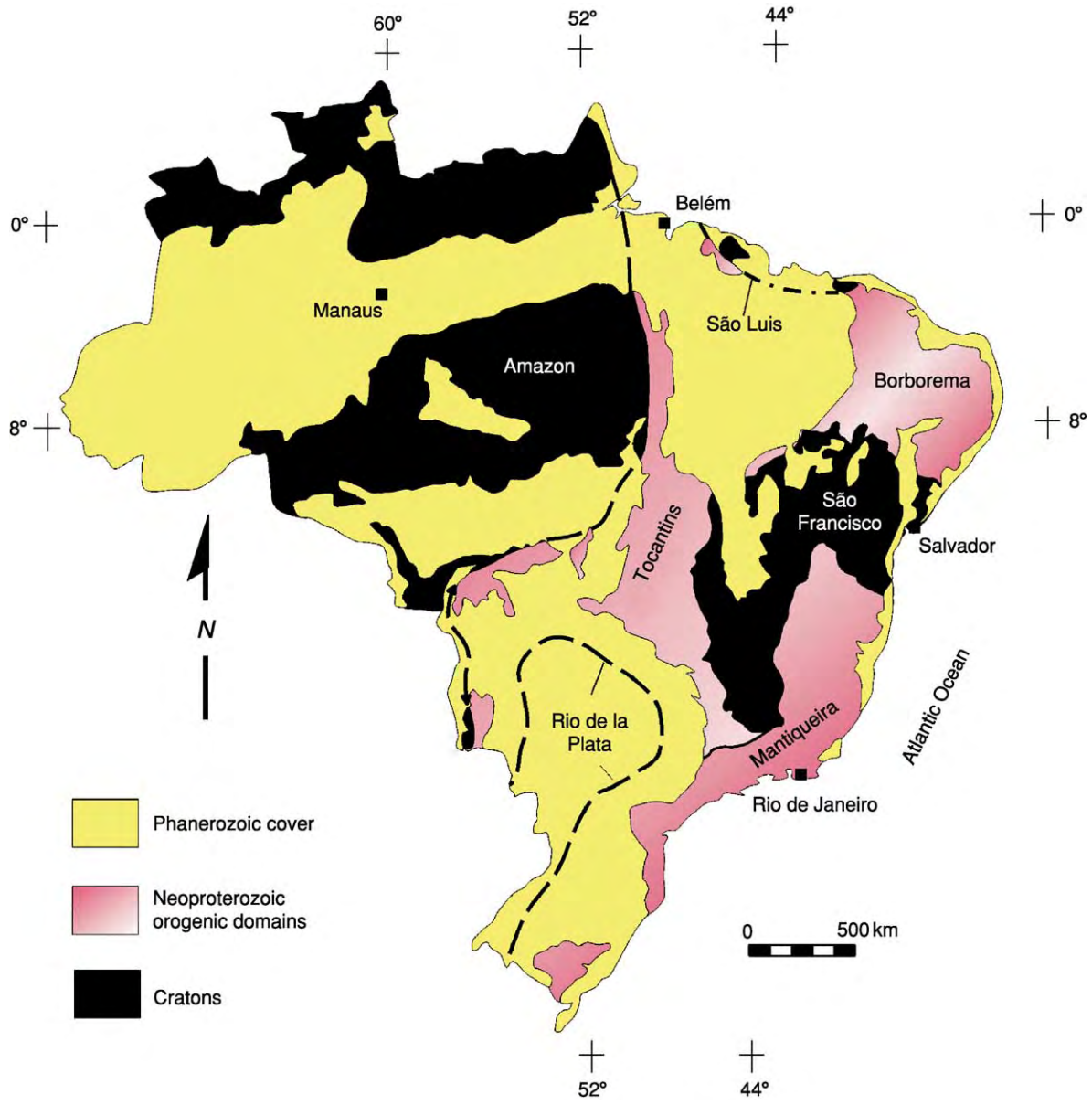


Figure 2 Simplified tectonic map of Brazil, showing the subdivisions of the South American platform and its Phanerozoic cover. Modified from Almeida FFM, Hasui Y, Brito Neves BB, and Fuck RA (1981) Brazilian structural provinces: an introduction. *Earth Science Reviews* 17: 1 29, with permission from Elsevier.

so-called Brasiliano or Pan-African orogenies. In this context, the cratons of South America and Africa are the preserved and more internal portions of the plates that collided to build up western Gondwana; the Neoproterozoic orogenic domains encompass the margins of these plates and other lithospheric pieces also involved in the tectonic collage that made up western Gondwana. The dispersal of western Gondwana and the opening of the South Atlantic in the Lower Cretaceous broke apart Neoproterozoic orogens and cratons. Consequently, the Neoproterozoic orogens and cratons of eastern Brazil have African counterparts (Figure 3).

At first glance the geological panorama of Brazil reflects only the Neoproterozoic assembly, Early Palaeozoic to Jurassic permanence, and Cretaceous

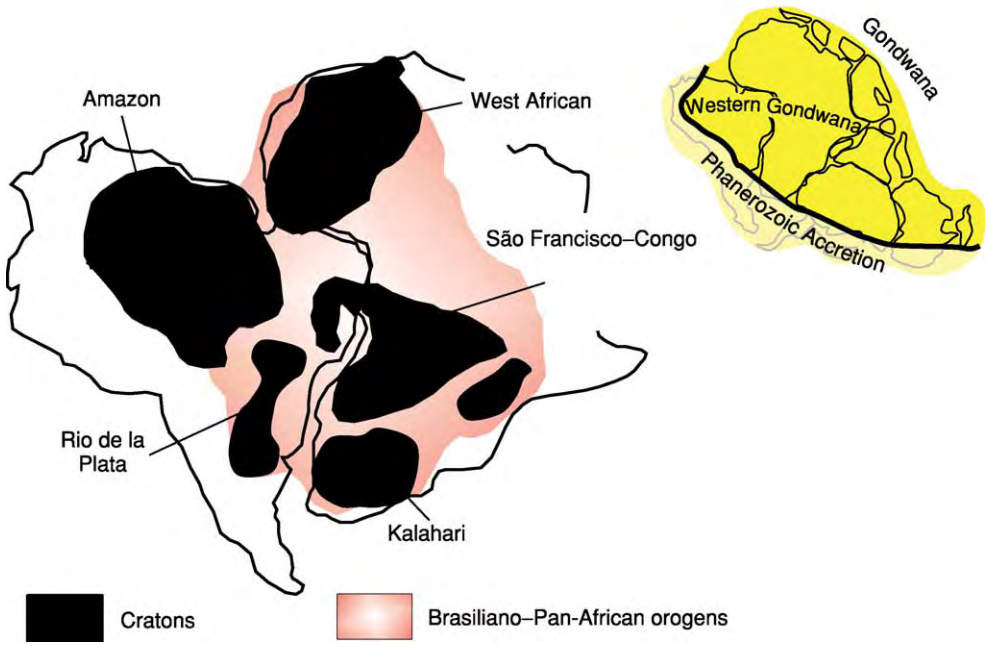


Figure 3 Schematic map of western Gondwana, showing the cratons and Neoproterozoic Brasiliano Pan African belts.

Table 1 The main thermotectonic events recorded in the South American platform

Event	*Age*	*Occurrence*	*Manifestation*	*Significance*
Tertiary reactivation	Eocene Oligocene	Eastern and Northern Brazil	Rifting, alkaline magmatism	
South Atlantic	130 78 Ma	Continental margin and interior	Flood basalts, followed by rifting and a late phase of alkaline magmatism	Gondwana breakup, opening of the South Atlantic
Brasiliano	640 520 Ma	Non cratonic areas	Deformation, metamorphism, magmatism, minor accretion	Western Gondwana assembly
Macaúbas rifting	900 800 Ma	São Francisco Craton and adjacent belts	Rifting, bimodal magmatism, glaciation	Rodinia breakup
Rondonian/ San Ignácio/ Sunsás	1500 1000 Ma	South western Amazon Craton	Deformation, metamorphism, magmatism	Rodinia assembly
Staterian rifting	~1750 Ma	São Francisco Craton and adjacent belts	Rifting, bimodal magmatism	Atlantica breakup
Uatumã	~1800 Ma	Amazon Craton	Anorogenic magmatism	Atlantica breakup, plume
Transamazonian	2200 1950 Ma	Amazon and São Francisco cratons, all Brasiliano orogenic domains	Deformation, metamorphism, magmatism, crustal accretion	Assembly of Atlantica supercontinent
Jequié, Rio Das Velhas, Aroense	2900 2780 Ma	São Francisco and Amazon cratons, some Brasiliano orogenic domains	Deformation, metamorphism, magmatism, crustal accretion	

dispersal of western Gondwana. A closer examination of the cratons and Neoproterozoic orogenic belts reveals, however, a long and diverse pre-Gondwana history, as well as a whole series of post-Gondwana features. Thus, in addition to the Brasiliano and South Atlantic events, other events of the same significance and extent are recorded in different ways in the South American platform and its cover. The most important of these events, together with their ages, areas of occurrence, and geotectonic significance, are shown in Table 1.

Regional Structures and Topography of Brazil

The topography of Brazil to a large extent reflects the constitution of the South American platform discussed in the previous section. The cratons underlie the low areas, hosting the main river basins; the highlands, on the other hand, have the Neoproterozoic orogenic domains as their substrata (Figure 4).

Each of the large-scale topographical highs and lows of the Brazilian territory is the expression of a particular regional tectonic structure. Some of these structures had already nucleated by the Palaeozoic. The majority, however, were initiated in the Mesozoic and underwent significant reactivation during the Cenozoic. The most prominent structural and topographical lows correspond to the Paraná, São Francisco, Parnaíba, and Amazonas basins. The Borborema Plateau, the Serra do Mar Uplift, and the Alto Paranaíba Arch are the largest structural and topographical highs (Figure 4).

Cratons

The cratons of the South American platform, consisting of Archaean crust with substantial

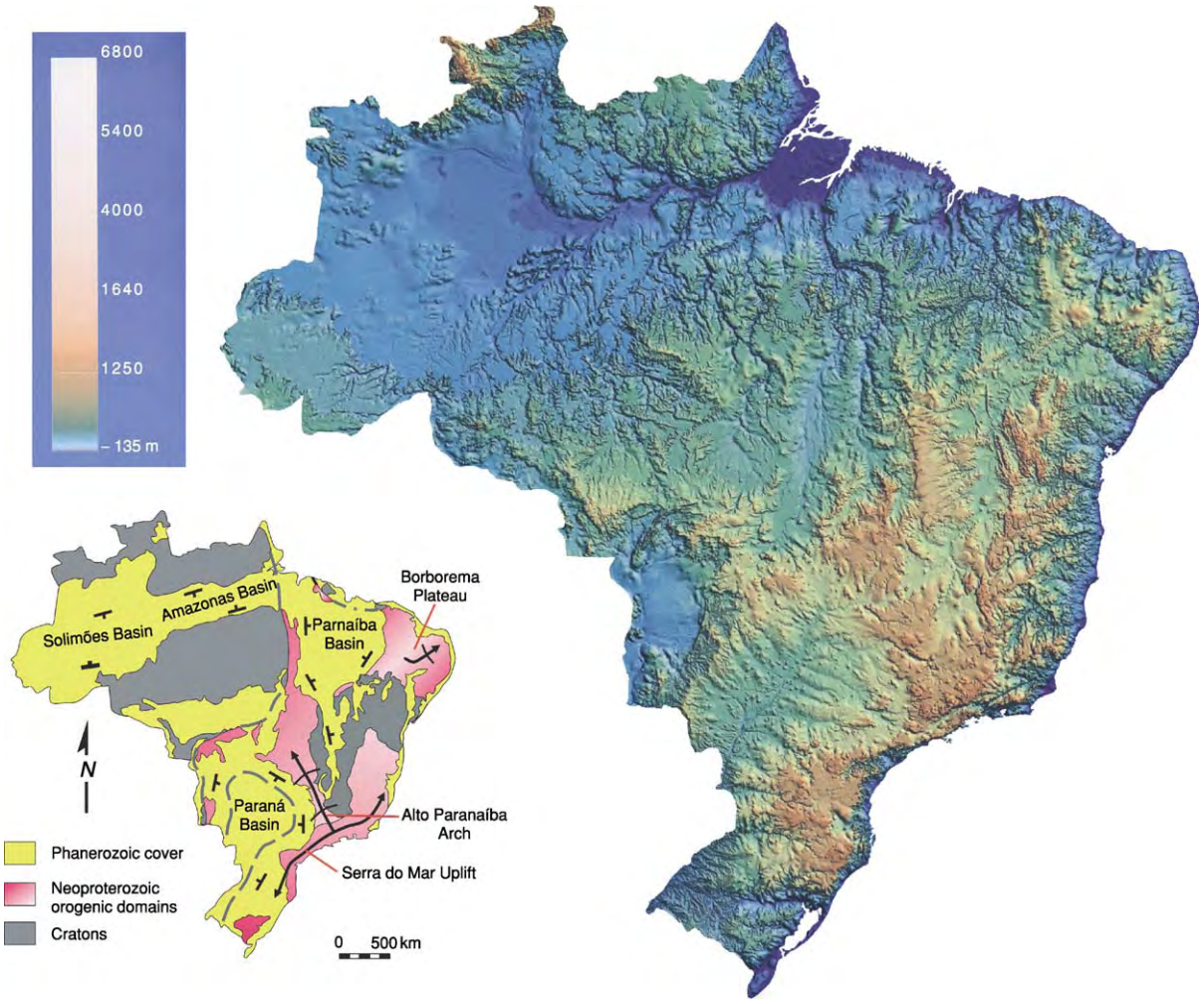


Figure 4 Correlation between the large scale tectonic structures and the topographical relief of Brazil. Relief map compiled by JBL Françolim based on EROS GTOPO 30 dataset; reproduced with the permission of the author.

Palaeoproterozoic accretions, are the portions of the Precambrian basement that were unaffected by the Brasiliano orogenies. Their boundaries are defined by the deformation styles of Neoproterozoic cover strata, patterns of geophysical anomalies, and geochronological data. Accordingly, the cratons of Brazil are bounded on all sides by Brasiliano basement-involved fold–thrust belts, but contain Brasiliano thin-skinned foreland fold–thrust belts. In this regard they differ from similar features delimited on other continents. The cratons of Brazil do, however, exhibit the typical attributes of their equivalents worldwide, such as mantle roots, low heat-flow values, and high lithospheric strength.

São Francisco Craton

A map view of the São Francisco Craton shows a shape that resembles a horse's head, with the northern segment merging with the east coast of Brazil (Figure 5). Except for the Atlantic coast, the São

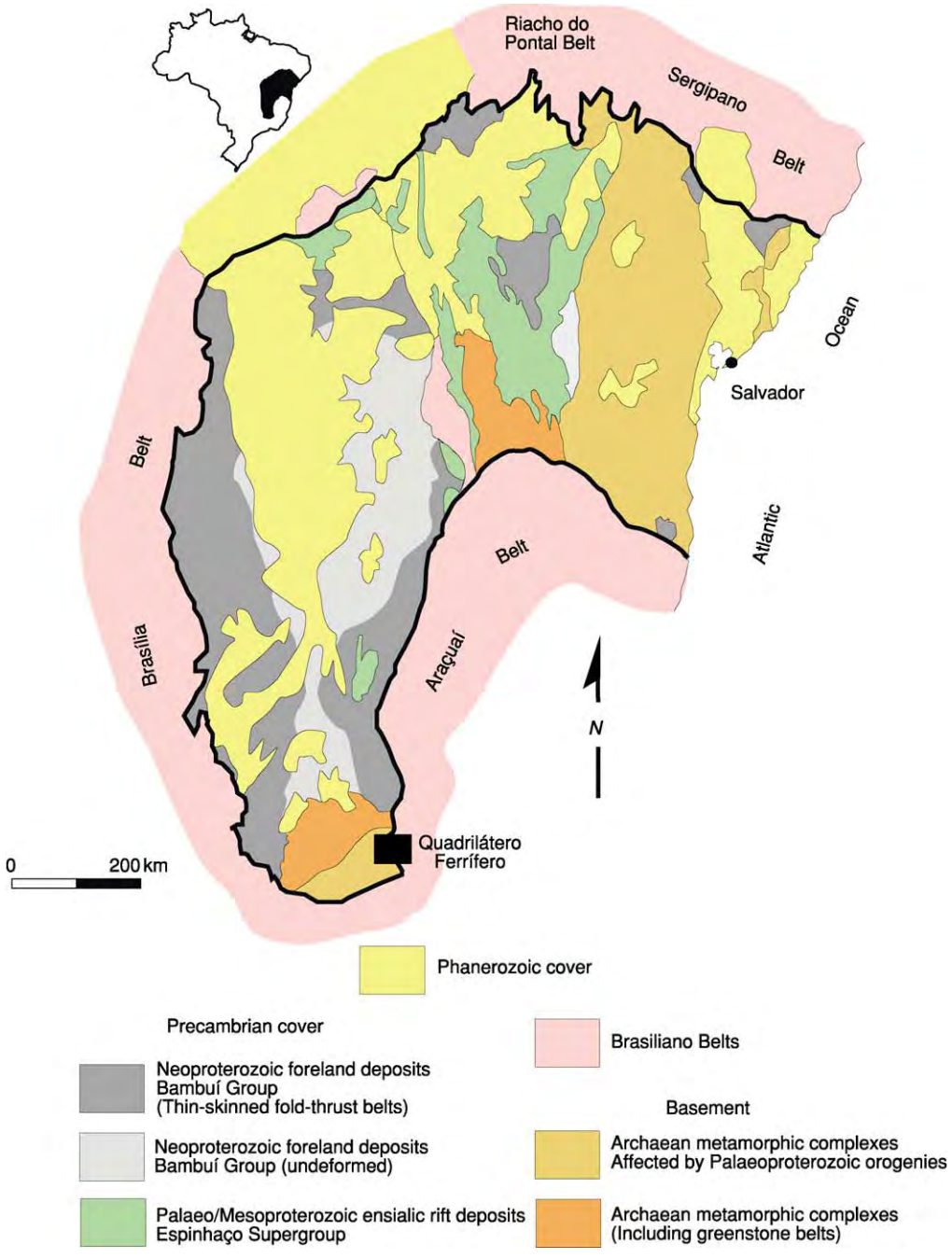


Figure 5 Simplified geological map of the São Francisco Craton showing the distribution of the major lithostratigraphical units.

Francisco Craton is bounded on all sides by the external fold–thrust belts of the Mantiqueira, Borborema, and Tocantins orogenic domains (Figures 2 and 5). As indicated by reconstructions of western Gondwana, the São Francisco Craton and the Congo Craton of Africa formed a single lithospheric unit from the Palaeoproterozoic until the Cretaceous opening of the South Atlantic.

The basement, made up of Archaean metamorphic complexes (tonalite–trondhjemite–granodiorite association and voluminous calc-alkaline plutons), Archaean greenstone belts, and Palaeoproterozoic metasedimentary successions, is exposed in the northern lobe of the craton and in a smaller area close to the southern boundary (Figure 5). In both exposures, Archaean nuclei are bounded on the east and south-east by Palaeoproterozoic Transamazonian fold–thrust belts, which involve the Archaean basement and Palaeoproterozoic metasedimentary units. The second-largest and best-studied mineral province of Brazil, the Quadrilátero Ferrífero (Iron Quadrangle), lies partly in the southern portion of the craton and partly in the adjacent Brasiliano Araçuaí belt. In this province, gold is found in an Archaean greenstone-belt sequence, and high-grade iron-ore deposits occur in a Palaeoproterozoic banded iron formation.

Alluvial to marine sediments with 1750 Ma acid-volcanic intercalations at the base (representing the fill of an ensialic rift system), glacial-influenced Tonian (*ca.* 850 Ma) rift sediments, and Cryogenian (720–600 Ma) foreland strata form the Precambrian cover of the craton. The foreland-basin strata are deformed in the areas adjacent to the craton boundaries, thereby defining two thin-skinned foreland fold–thrust belts with opposite vergences. These belts represent the orogenic fronts of the Neoproterozoic Brasília and Araçuaí belts, which fringe the craton to the west and to the east, respectively.

Amazon Craton

The Amazon Craton encompasses an area of approximately 430 000 km^2 in north-western South America, extending far beyond the Brazilian borders into Colombia, Venezuela, Guyana, Suriname, and French Guiana. The eastern and south-eastern limits of the craton are marked by the external portion of the Brasiliano Tocantins Orogen, represented by the Araguaia and Paraguay belts, respectively. The western and north-western boundaries are covered by the sub-Andean basins, and the northern limit is marked by the South American equatorial margin (Figures 2 and 6).

More than two-thirds of the Brazilian portion of the craton is covered by Phanerozoic sediments and the Amazon forest, so that the geological knowledge

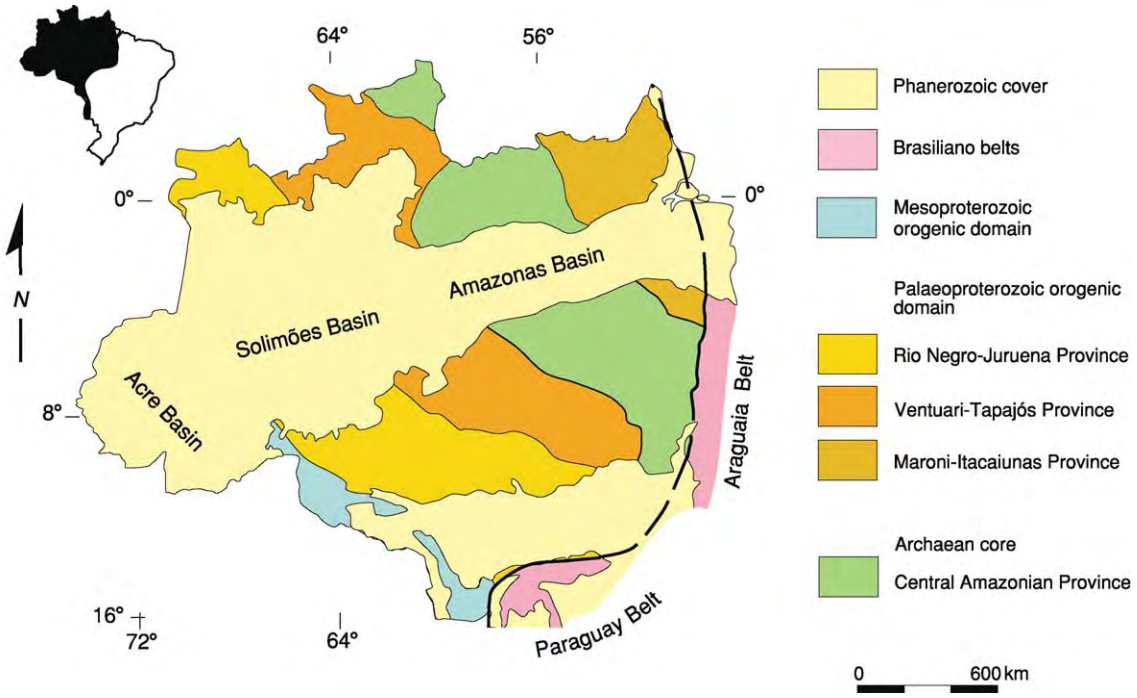


Figure 6 Schematic map of the Amazon Craton, illustrating the main basement components and cover units. Modified from Tassinari CG, Bettencourt S, Geraldes M, Macambira JB, and Lafon JM (2000) The Amazonian Craton. In: Cordani UG, Milani EJ, Thomaz FA, and Campos DA (eds.) *Tectonic Evolution of South America*, pp. 41 95. 31st International Geological Congress, Rio de Janeiro.

of the region is limited. The cratonic basement is composed of an Archaean core, a large Palaeoproterozoic orogenic domain, and a Mesoproterozoic orogenic zone to the south-west.

The Archaean core, or Central Amazonian geochronological province, is the portion of the craton not affected by the Palaeoproterozoic orogenies. It consists of granite–greenstone and high-grade terranes, overlain by Archaean volcanosedimentary sequences and very thick packages of undeformed Palaeoproterozoic to Mesoproterozoic sedimentary and volcanic units. The Carajás mining district, located in the north-western portion of the province, contains the largest high-grade iron-ore reserves in Brazil as well as important copper, manganese, gold, and nickel deposits.

According to recent estimates, approximately 70% of the Amazon Craton consists of Palaeoproterozoic juvenile crust, which forms the basement of three geochronological provinces: Maroni-Itacaiunas, Ventuari-Tapajós, and Rio Negro-Juruena (Figure 6). The Maroni-Itacaiunas Province encompasses a segment of a Palaeoproterozoic accretionary Orogen, which is the main manifestation of the 2200–1950 Ma Transamazonian event. The Ventuari-Tapajós and Rio Negro-Juruena provinces are also collages of juvenile terranes generated in the intervals 1950–1800 Ma and 1800–1550 Ma, respectively.

The south-western part of the Amazon Craton, along the border with Bolivia, comprises parts of a Mesoproterozoic Orogen developed between 1500 Ma and 1300 Ma. It involves a Palaeoproterozoic basement, a juvenile terrane, and late to post-tectonic granites.

São Luis and Rio de la Plata Cratons

The existence of the São Luis and Rio de la Plata cratons has been inferred from structural features of the various Brasiliano belts together with geochronological and geophysical data.

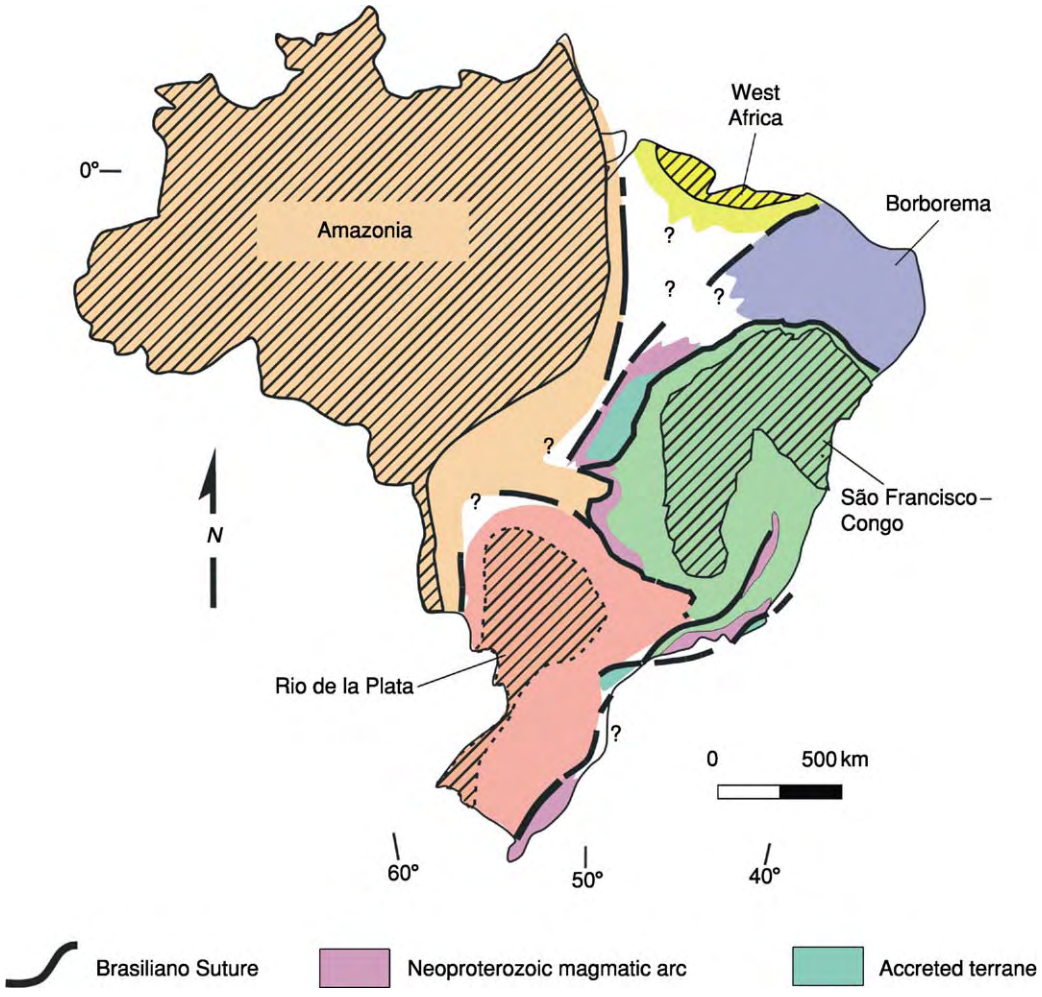


Figure 7 Neoproterozoic suture zones of Brazil, and the plates that collided to form the South American portion of western Gondwana. Note that the internal parts of the plates correspond to the cratons.

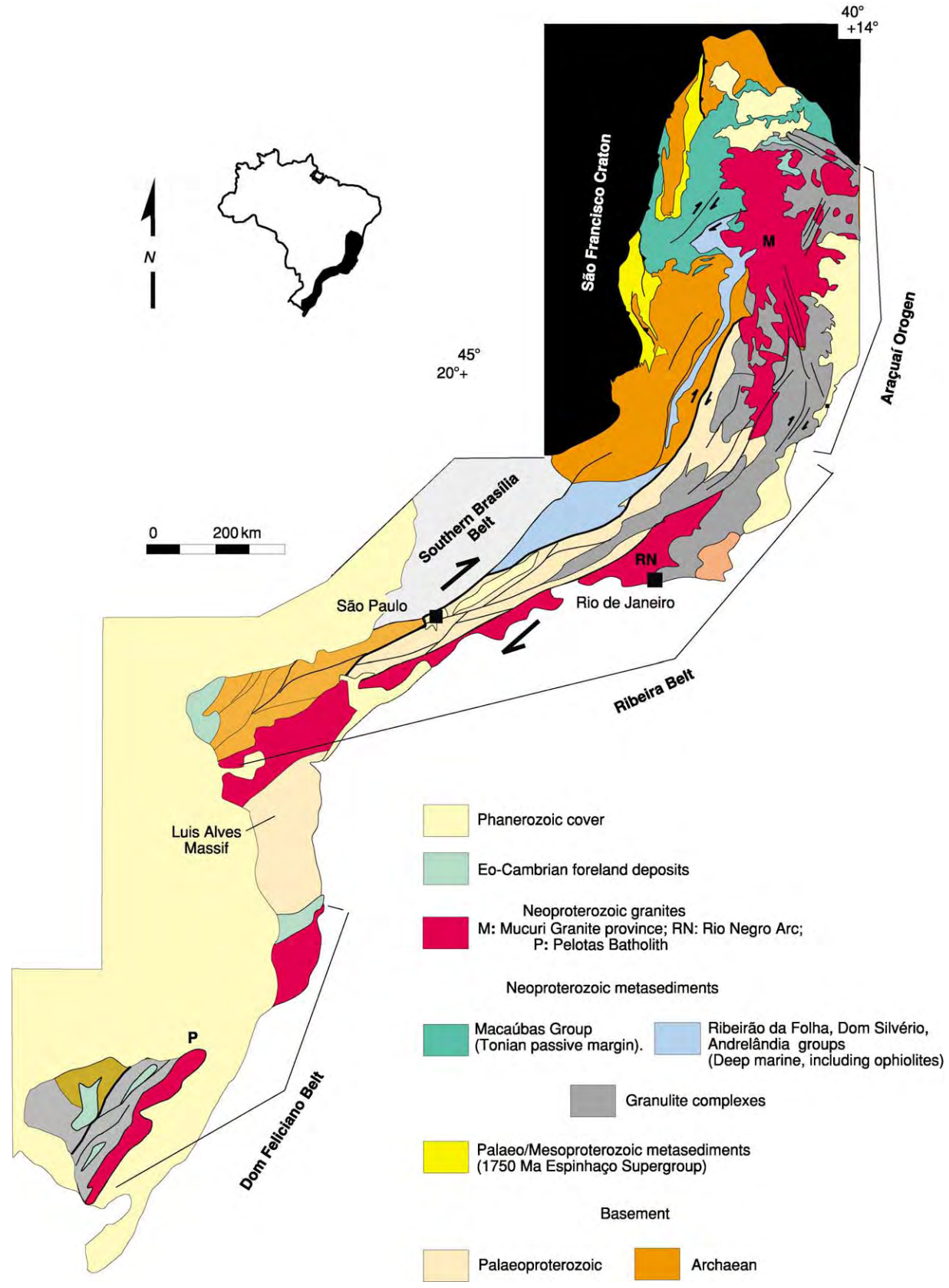


Figure 8 Geological map of the Mantiqueira orogenic system, showing the Araçuaí Orogen and the Ribeira and Dom Feliciano belts.

The São Luis Craton on the northern coast has only small exposures of a granite–greenstone-belt association, which yield potassium–argon cooling ages of around 2100 Ma. Potassium–argon dates obtained from the exposures of adjacent areas fall in the interval between 660 Ma and 500 Ma, indicating that they were formed during the Brasiliano event. This led to the interpretation that a small piece of the West African Craton, represented by the São Luis Block, was left in South America during the opening of the Atlantic (Figures 2 and 3).

The Rio de la Plata Craton underlies the Palaeozoic Paraná Basin in southern Brazil (Figure 2). The presence of a cratonic block that underlies the Paraná Basin, originally inferred from the tectonic polarity of the Neoproterozoic orogenic belts exposed in the region, was confirmed by gravity and teleseismic travel-time studies. Furthermore, exposures of Palaeoproterozoic crust unaffected by the Brasiliano event in Uruguay and Argentina are now viewed as extensions of the cratonic block that underlies the Paraná Basin.

Neoproterozoic Orogenic Domains

The Neoproterozoic orogenic domains form a complex network of interacting orogens and were generated by the collisions that led to the assembly of western Gondwana between 640 Ma and 520 Ma.

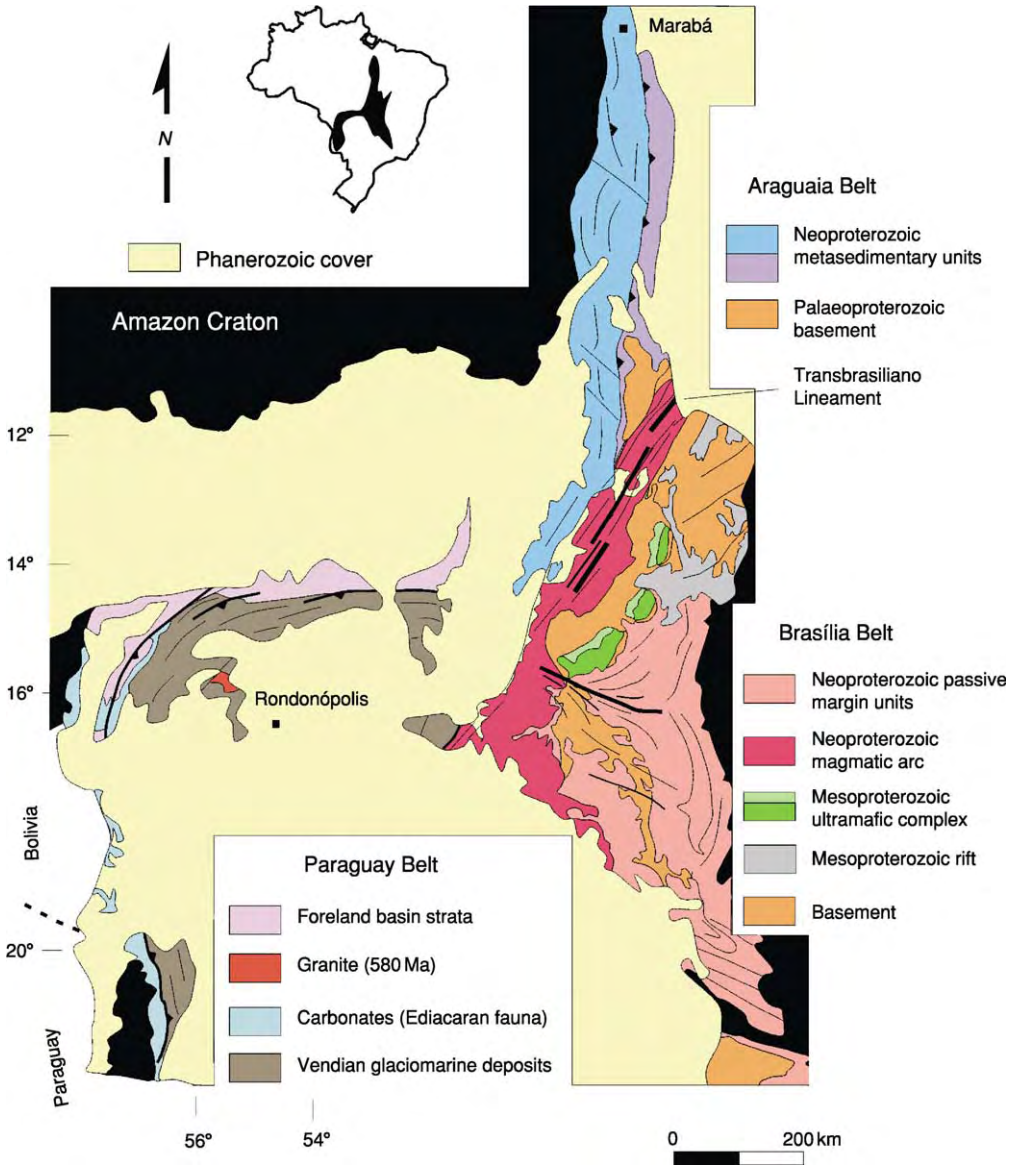


Figure 9 Geological map of the Tocantins orogenic system, showing the Araguaia, Paraguay, and Brasília belts.

Accretion of juvenile crust played a minor role in the Brasiliano orogenies; recycling of the pre-existing continental lithosphere was the dominant process. Besides wide zones of deformation and metamorphism, the Brasiliano orogenic domains contain a relatively large volume of syn- and post-collisional granites. Furthermore, a substantial part of the orogenic zone is either transpressional in origin or has at least been modified by strike-slip movement in the late stages of development.

Neoproterozoic suture zones have been directly mapped or inferred with the help of geophysics in almost all the orogenic domains, so that the plates involved in the creation of the Brazilian portion of western Gondwana can be delineated as shown in Figure 7.

Mantiqueira Orogenic System

The Mantiqueira orogenic system extends over a zone 2500 km long and 200 km wide along the south-east coast of Brazil, and comprises the Araçuaí Orogen and the Ribeira and Dom Feliciano belts (Figure 8); the West Congolian, Kaoko, and Gariep belts are their African counterparts. The Mantiqueira system overprints a branch of the Tocantins system in the area to the south of the São Francisco Craton and continues further south, reaching the Uruguayan shield beyond the Brazilian border.

Araçuaí Orogen The Araçuaí–West Congo Orogen (Figure 8) encompasses the terrane between the São Francisco Craton and the east coast of Brazil, as well as the West Congolian Belt of Africa. This orogen is the northern termination of the Mantiqueira system and its African counterpart, and it displays the unusual shape of a tongue, being surrounded on all sides, except the south, by cratonic areas. The Brazilian or Araçuaí portion of the orogen is made up of the Araçuaí fold–thrust belt to the west and a wide zone of high-grade and granitic rocks to the east, which represents its crystalline core. This internal zone is also known as the northern Ribeira, Atlantic, or Coastal Mobile Belt.

The Araçuaí fold–thrust belt verges towards the adjacent craton and involves a basement older than

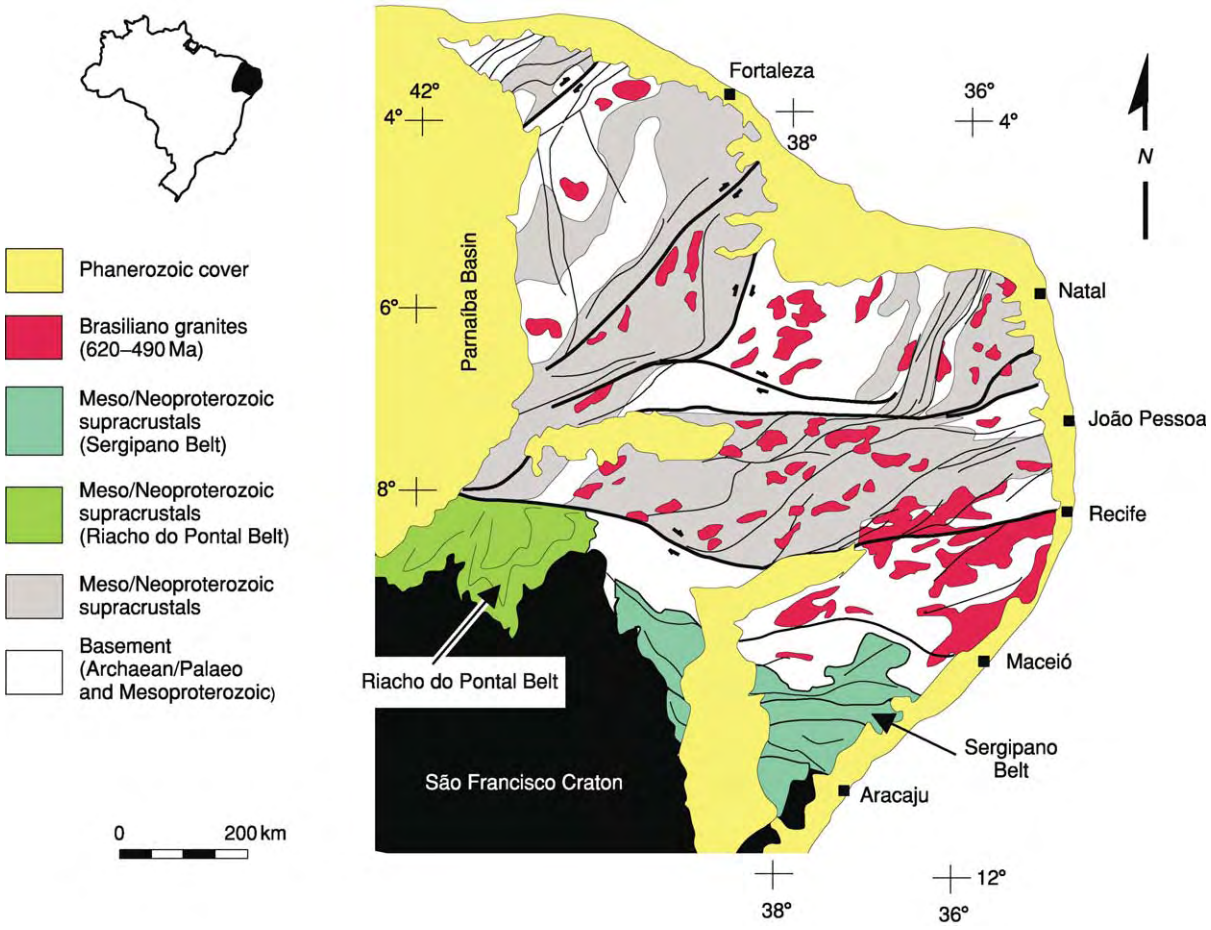


Figure 10 Geological map of the Borborema Strike slip System.

1800 Ma, a 1750 Ma ensialic rift assemblage, Neoproterozoic passive margin to foreland strata, and other correlative units. The crystalline core, dominated by an array of north-north-east–south-south-west-striking dextral transpressional zones, comprises a Palaeoproterozoic basement, amphibolite or granulite facies Neoproterozoic sedimentary units, and a batholithic body of Brasiliano granites (625–530 Ma).

The Araçuaí–West Congo Orogen is currently viewed as the product of the closure of a Red Sea type ocean, whose development started with rift formation in the Tonian between 900 Ma and 800 Ma.

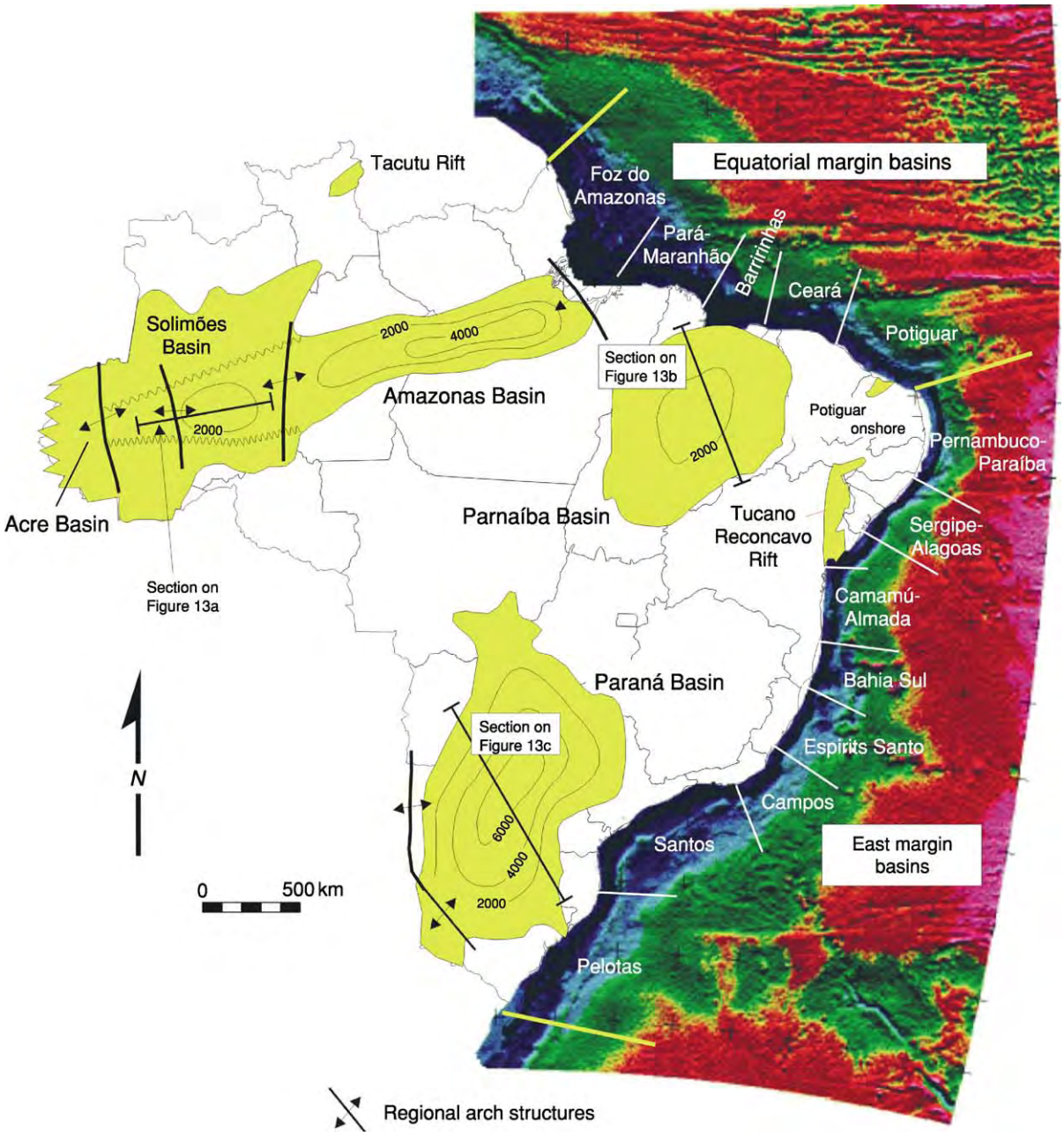


Figure 11 Locations of the most important Brazilian Phanerozoic sedimentary basins. Maps and isopachytes (in metres) of Palaeozoic sags, modified from Milani E and Thomaz Filho A (2000) Sedimentary basins of South America. In: Cordani UG, Milani EJ, Thomaz FA, Campos DA (eds.) *Tectonic Evolution of South America*, pp. 389 449. 31st International Geological Congress, Rio de Janeiro. Offshore Bouguer anomaly map from Mohriak (2003) Rift architecture and salt tectonics in South Atlantic sedimentary basins. In: Mohriak WU *Rifted Sedimentary Basins of South Atlantic: Turbidite Reservoirs, Sedimentary and Tectonic Processes*. 41st Brazilian Geological Congress and 1st International Conference of the Geophysical Society of Angola, Short Course Notes, CD ROM, Socıedade Brasiliera de Geologia, João Pessoa, Brazil.

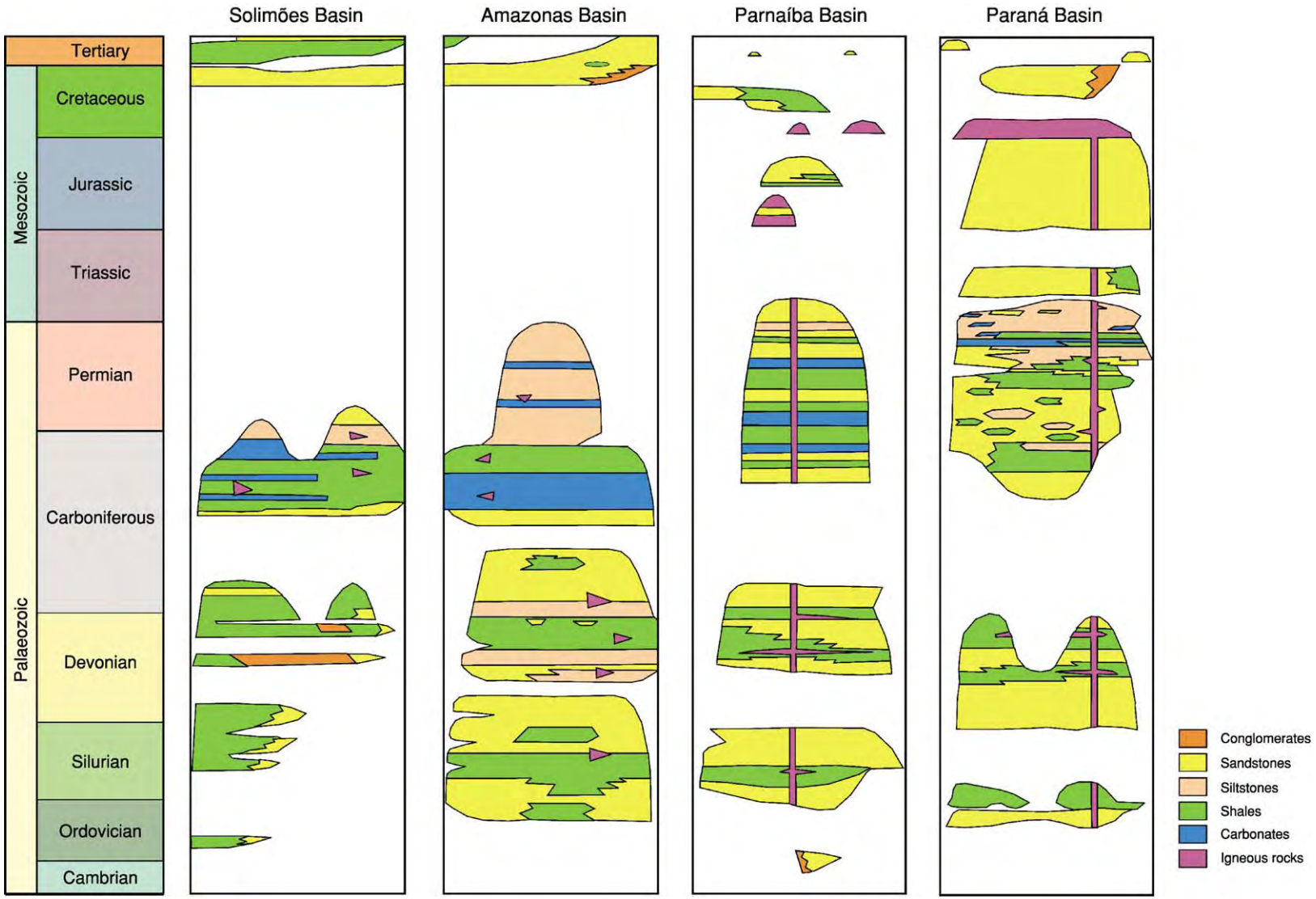


Figure 12 Stratigraphical chart of Brazilian Palaeozoic sag basins. Modified from the Brazilian Petroleum Agency (Agência Nacional do Petrōleo, ANP) home page at www.anp.gov.br

Ribeira and Dom Feliciano belts Together, the north-east–south-west orientated Ribeira and Dom Feliciano belts occupy a strip approximately 200 km wide and 1600 km long along the south-east coast of Brazil. The northern boundary with the Araçuaí Orogen is artificially taken to be where the fabric elements bend slightly to become north-north-east–south-south-west. The Ribeira Belt truncates the easternmost portion of the Tocantins orogenic system, represented by the southern Brasília Belt, and is overlain by the Paraná-Basin strata to the south, where the Luis Alves basement block separates it from the Dom Feliciano Belt.

A Palaeoproterozoic basement with some Archaean inliers, Mesoproterozoic to Neoproterozoic metasedimentary units, Eocambrian foreland deposits, and

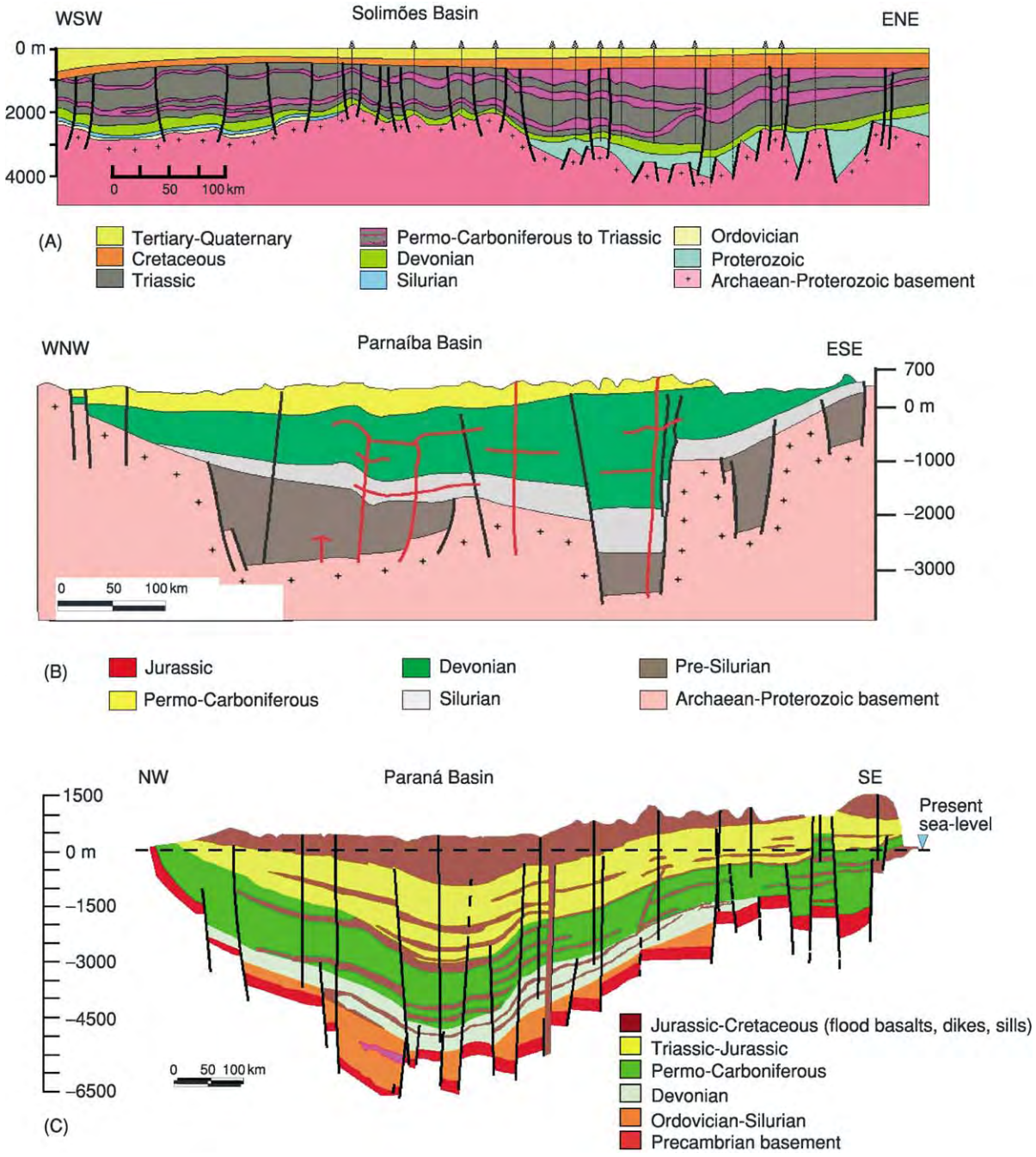


Figure 13 Schematic geological sections of (A) Solimões, (B) Parnaíba, and (C) Paraná sag basins. Modified from the Brazilian Petroleum Agency (Agência Nacional do Petrõleo, ANP) home page at www.anp.gov.br

voluminous granites emplaced in the pre-collisional to post-collisional stages are the main components of the Ribeira and Dom Feliciano belts. Various terranes and three magmatic arcs have been characterized along these belts.

The overall architecture of both the Ribeira and Dom Feliciano belts is a system of north-west-verging thrusts and nappes, which pass to the south-east and south into a transpressional zone. In the Ribeira Belt the internal transpressional zone is right lateral, whereas left lateral displacements dominate the Dom Feliciano Belt. This tectonic picture can be interpreted in two basic conflicting ways. It is portrayed either as a result of a late transpressional modification of a pre-existing frontal collisional orogen or as a product of an oblique plate convergence that led to the development of a transpressional orogen. Regardless of which hypothesis is correct, the evolution of the Ribeira and Dom Feliciano belts is associated with the interaction of three continental masses (the São Francisco–Congo, Rio de la Plata, and Kalahari cratons), which were separated by an ocean that contained magmatic arcs and microcontinents. The whole convergence stage lasted from 600 Ma to 520 Ma.

Tocantins Orogenic System

The Tocantins system comprises the orogens developed between the Amazon, Rio de la Plata, and São Francisco cratons, and occupies a large area in northern and central Brazil. The Tocantins System, together with the Borborema strike-slip province of north-eastern Brazil (see next section), resembles the present-day Himalayan system in architecture. The Araguaia, Paraguay, and Brasília belts are the exposed components of the Tocantins System (Figure 9).

Araguaia Belt The north–south trending Araguaia Belt on the eastern margin of the Amazon Craton comprises a west-verging basement-involved fold–thrust belt, which is 1200 km long and 150 km

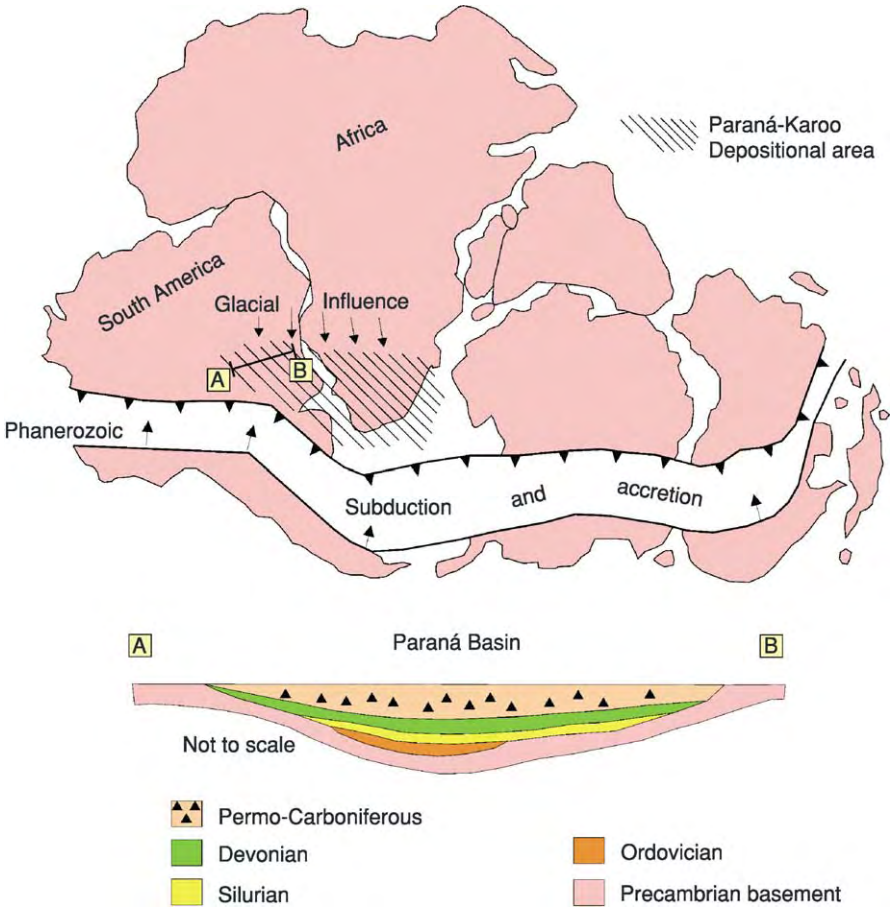


Figure 14 Palaeogeographical reconstruction of western Gondwana in the Late Palaeozoic, showing the location of the Paraná Basin in the foreland domain of accretionary processes on the Gondwanan margin. Modified from Milani EJ (1992) Intraplate tectonics and the evolution of the Paraná Basin, SE Brazil. In: De Wit MJ, Ransome IGD (eds.) *Inversion tectonics of the Cape Fold Belt, Karoo and Cretaceous Basins of Southern Africa*. Rotterdam, Balkema, pp. 101 108.

wide. To the south it is separated from the Brasília Belt by a basement block, which is affected by the north-east–south-west-trending Transbrasiliano Fault Zone; to the east and to the north it is covered by the sediments of the Palaeozoic Parnaíba Basin (Figure 9). The Archaean and Palaeoproterozoic basement is overlain by Neoproterozoic metasediments, which are the best exposed in the belt. A 720 Ma ophiolite occurs within thrust sheets in the central portion of the belt.

Paraguay Belt The Paraguay Belt, the youngest member of the Tocantins orogenic system, is a pronounced salient on the south-eastern border of the Amazon Craton (Figure 9). Vendian glaciomarine deposits overlain by carbonates containing an Ediacaran fauna, together with continental foreland deposits, are the main units exposed in the belt. Its overall architecture is characterized by a system of open to tight upright folds in the culmination zone of the large salient, which grade into two cratonward-verging systems of folds and thrusts at its north-eastern and southern ends. The deformation and metamorphism in the Paraguay Belt is estimated to have occurred between 550 Ma and 500 Ma and probably represents the last steps towards the final assembly of western Gondwana.

Brasilia Belt The Brasília Belt is an east-verging fold–thrust belt (1200 km long and up to 300 km wide) that fringes the São Francisco Craton to the west and south-west (Figure 9). The basement is an extension of the Archaean–Palaeoproterozoic substratum of the adjacent São Francisco Craton. In the internal zone, the Palaeoproterozoic basement received two Neoproterozoic accretions: an Archaean granite–greenstone terrane and a large magmatic arc (Figure 9). A suture zone juxtaposes the Rio de la Plata and São Francisco plates in the southern portion of the belt.

A 1770 Ma ensialic rift sequence, a thick Neoproterozoic passive-margin package, and deep-marine and probably ocean-floor assemblages are the main cover units. Mesoproterozoic layered mafic–ultramafic complexes and foreland-basin sediments are also involved.

From a structural standpoint the Brasília Belt consists of two distinct compartments that join along the pronounced east–west-trending Pirineus syntaxis: an east–south-east-verging basement-involved fold–thrust belt to the north, and a system of spoon-shaped east–south-east-verging nappes to the south. The Transbrasiliano Lineament, a north-east–south-west-orientated dextral strike-slip fault zone, overprints the internal portions of both the Brasília and

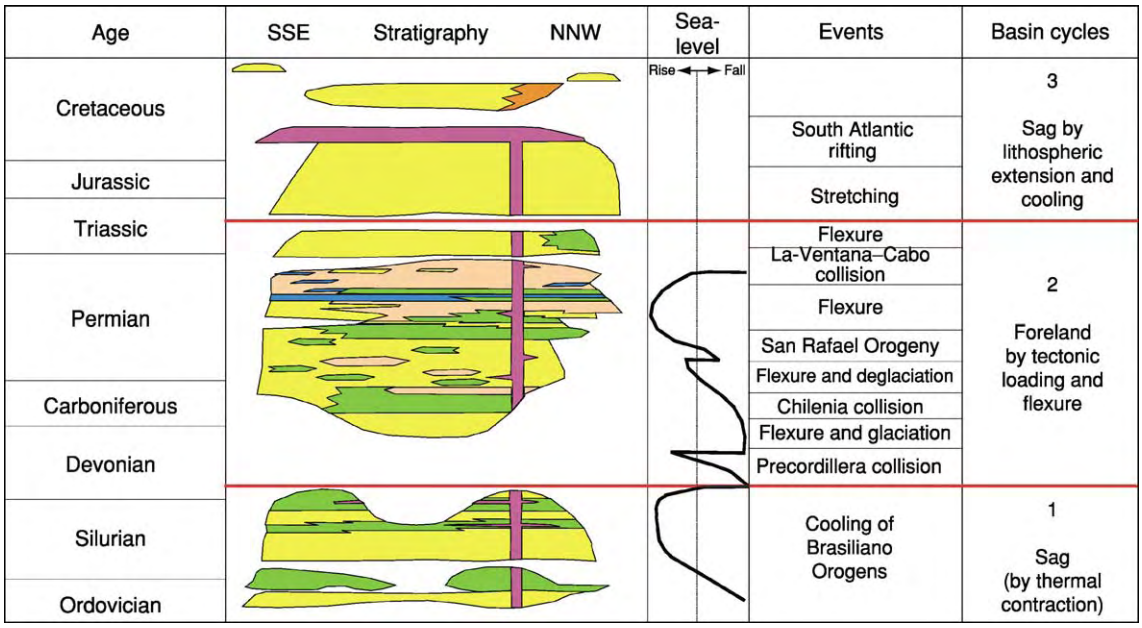


Figure 15 Stratigraphical chart of the Paraná Basin showing the evolutionary stages and their genetic controls. Note the corres pondence between depositional stages/unconformities and orogenic events in the Andean chain during the foreland stage. Based on: (1) Zalan PV, Wolff S, Astolfi MA, *et al.* (1990) The Paraná Basin, Brazil. In: Leighton MW, Kolata DR, Oltz DF, and Eidel JJ (eds.) *Interior Cratonic Basins*, pp. 681 708. AAPG Memoir 51. Tulsa: American Association of Petroleum Geologists; (2) Milani EJ (1992) Intraplate tectonics and the evolution of the Paraná Basin, SE Brazil. In: De Wit MJ, Ransome IGD (eds.) *Inversion tectonics of the Cape Fold Belt, Karoo and Cretaceous Basins of Southern Africa*. Rotterdam, Balkema, pp. 101 108.

Araguaia belts and seems to continue further north beneath the Phanerozoic cover, merging with the Borborema Strike-slip System (see next section).

The evolution of the Tocantins System probably started with an interaction between São Francisco–Congo and the Rio de la Plata at around 790 Ma. Arc magmatism took place between 900 Ma and 630 Ma, reflecting subduction and progressive closure of a large ocean separating São Francisco–Congo from Amazonia. After accretion of the arcs

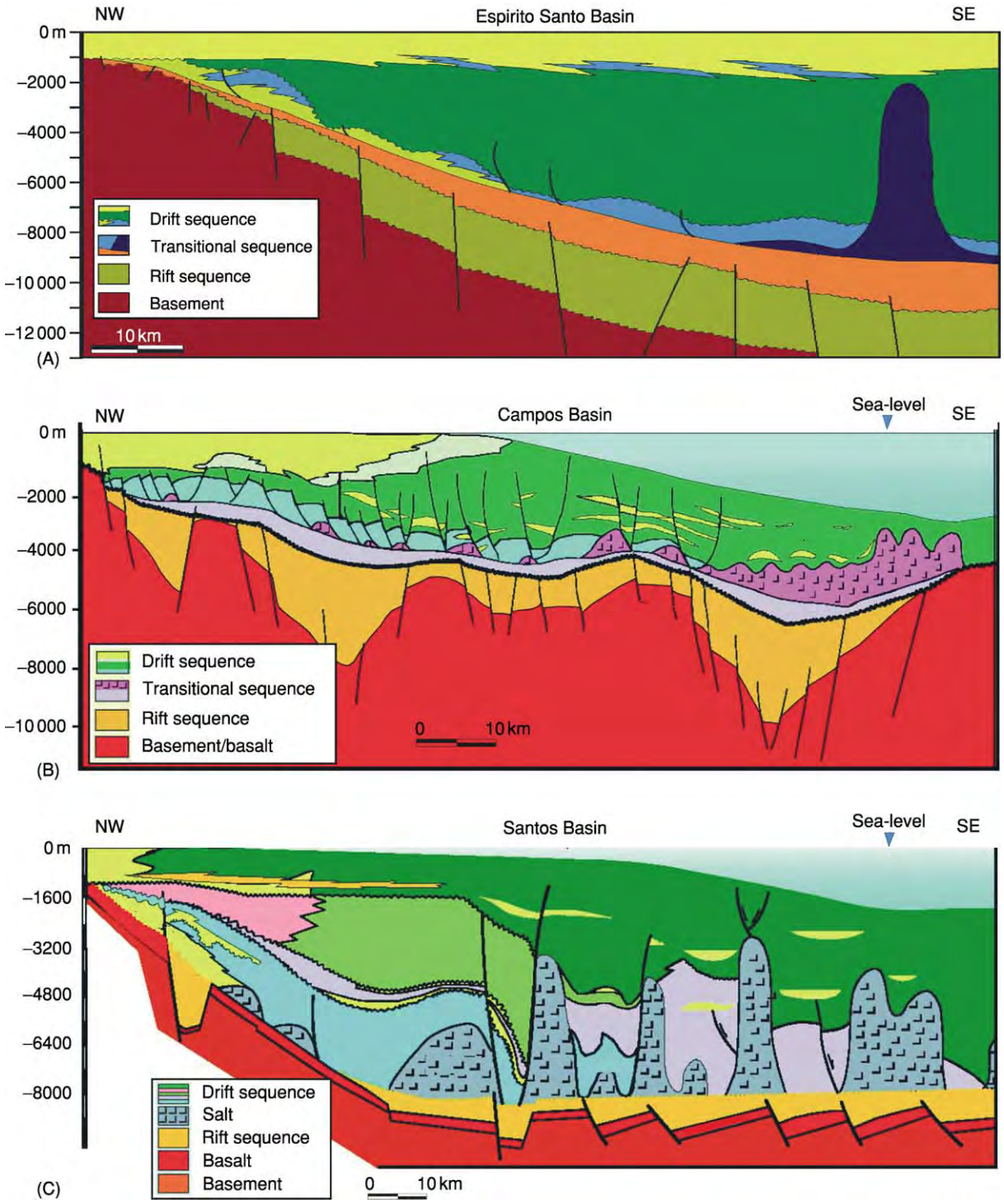


Figure 16 Schematic geological sections of Brazilian eastern margin (A) Espirito Santo, (B) Campos, and (C) Santos basins. Modified from the Brazilian Petroleum Agency (Agência National do Petrõleo, ANP) home page at www.anp.gov.br

Figure 17 Stratigraphical charts of the Sergipe Alagoas, Espirito Santo, Campos, and Santos Basins of the Brazilian eastern margin.

and the microcontinent, the convergence of São Francisco–Congo, Amazonia, and West Africa between 630 Ma and 590 Ma led to the development of the central and northern segments of the Tocantins System. The final collision between Rio de la Plata and Amazonia led to the development of the Paraguay Belt in the Early Cambrian.

Borborema Strike-Slip System and Associated Features

The Borborema Strike-slip System in north-eastern Brazil (Figure 10) comprises a gigantic Brasiliano strike-slip deformation zone associated with two south-verging fold–thrust belts, which were once continuous with the Nigerian, Hoggar, and Central African provinces. A fan-like array of dextral shear zones outlines elongated basement blocks, which correspond to the various massifs delimited in the region. These blocks are composed of Archaean, Palaeoproterozoic, and Mesoproterozoic rocks, covered by Mesoproterozoic to Neoproterozoic supracrustals, and intruded by voluminous Neoproterozoic granites. The metasedimentary units form schist belts, which wrap around the basement massifs.

The west-north-west–east-south-east-trending Sergipano Belt in the south-eastern portion of the Borborema System is made up of a system of south-south-west-verging folds and thrusts, involving an Archaean basement and a package of Neoproterozoic passive-margin to foreland strata. The thrusts of the Sergipano Belt are rooted in a dextral transpressional shear zone, which marks a Brasiliano suture, juxtaposing a continental mass represented by the remaining parts of the system to the north with the São Francisco–Congo Plate to the south. In Africa, the Oubanguides or Central African Belt on the northern

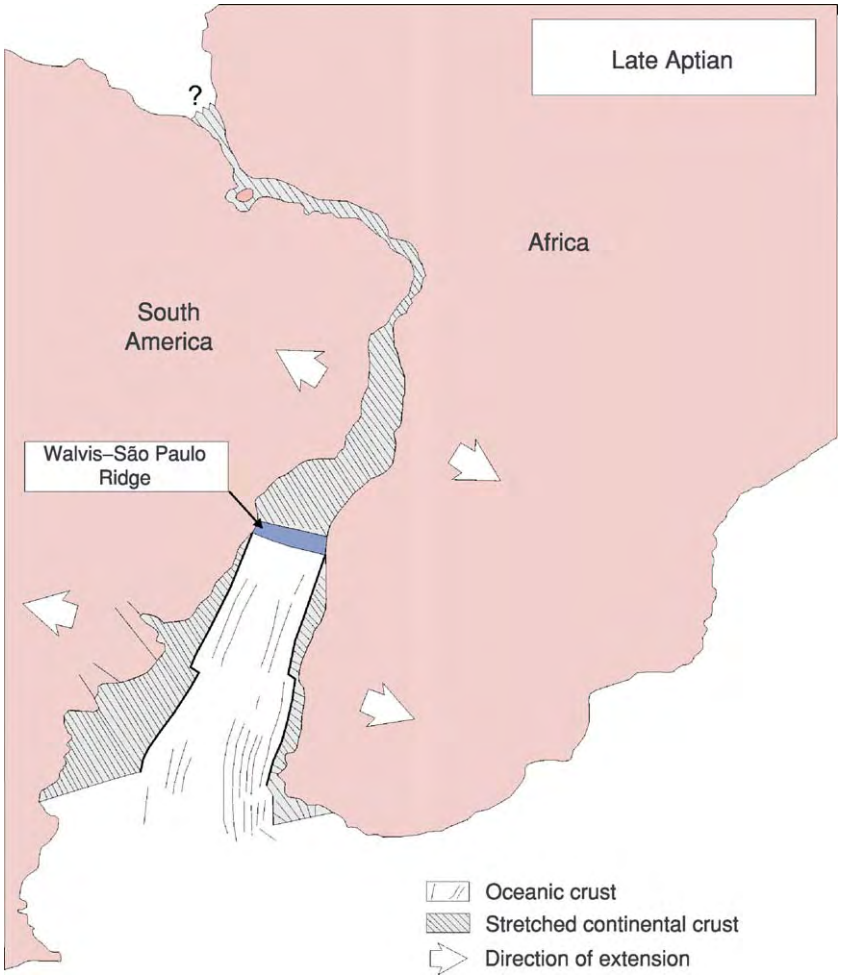


Figure 18 Palaeogeographical reconstruction of the South Atlantic in the Late Aptian, showing the location of the Walvis São Paulo Ridge, which acted as a barrier to free oceanic circulation, thereby creating conditions favourable to evaporitic (salt) deposition in the central segment of the eastern margin. Modified from Chang HK, Kowsmann RO, Figueiredo AMF, and Bender AA (1992) Tectonics and stratigraphy of the East Brazil rift system: an overview. *Tectonophysics* 213: 97 138, with permission from Elsevier.

border of the Congo Craton is an extension of the Sergipano Belt.

Separated from the Sergipano Belt by a basement high, the Riacho do Pontal Belt, comprises a pile of south-verging nappes involving Archaean basement and schists and quartzites of uncertain ages.

Peak metamorphism and the emplacement of syn-tectonic granites are dated by U-Pb and other methods at between 630 Ma and 580 Ma, whereas ^{39}Ar-^{40}Ar dates obtained in some of the main shear zones indicate cooling ages of between 580 Ma and 500 Ma.

Phanerozoic Sedimentary Basins

The Phanerozoic sedimentary record of Brazil (Figure 11) reflects three distinct plate-tectonic settings: the continental interior, where large Palaeozoic sag basins and Early Cretaceous intracontinental rifts are recognized; the passive margin, represented by the eastern border of South America; and the transform margin, represented by the Brazilian northern or equatorial margin. The continental margins developed from Mesozoic times onwards in response to the breakup of Gondwana and the spreading of the Atlantic Ocean.

Palaeozoic Sag Basins

During the Ordovician, following the cooling of Brasiliano–Pan-African orogenic systems, sedimentation started in the newly assembled western Gondwana in large interior sag basins, namely the Solimões (600 000 km^2), Amazonas (500 000 km^2), Parnaíba (600 000 km^2), and Paraná (1 400 000 km^2) basins (Figure 11). The stratigraphical records of these long-lived basins are characterized by a layer-cake arrangement, reflecting mostly shallow-marine sandy–shaly deposition (Figures 12 and 13). The major stratigraphical sequences are bounded by inter-regional unconformities, which are coeval along the basins (Figure 12), suggesting global control mechanisms. The existence of precursor rifts in these

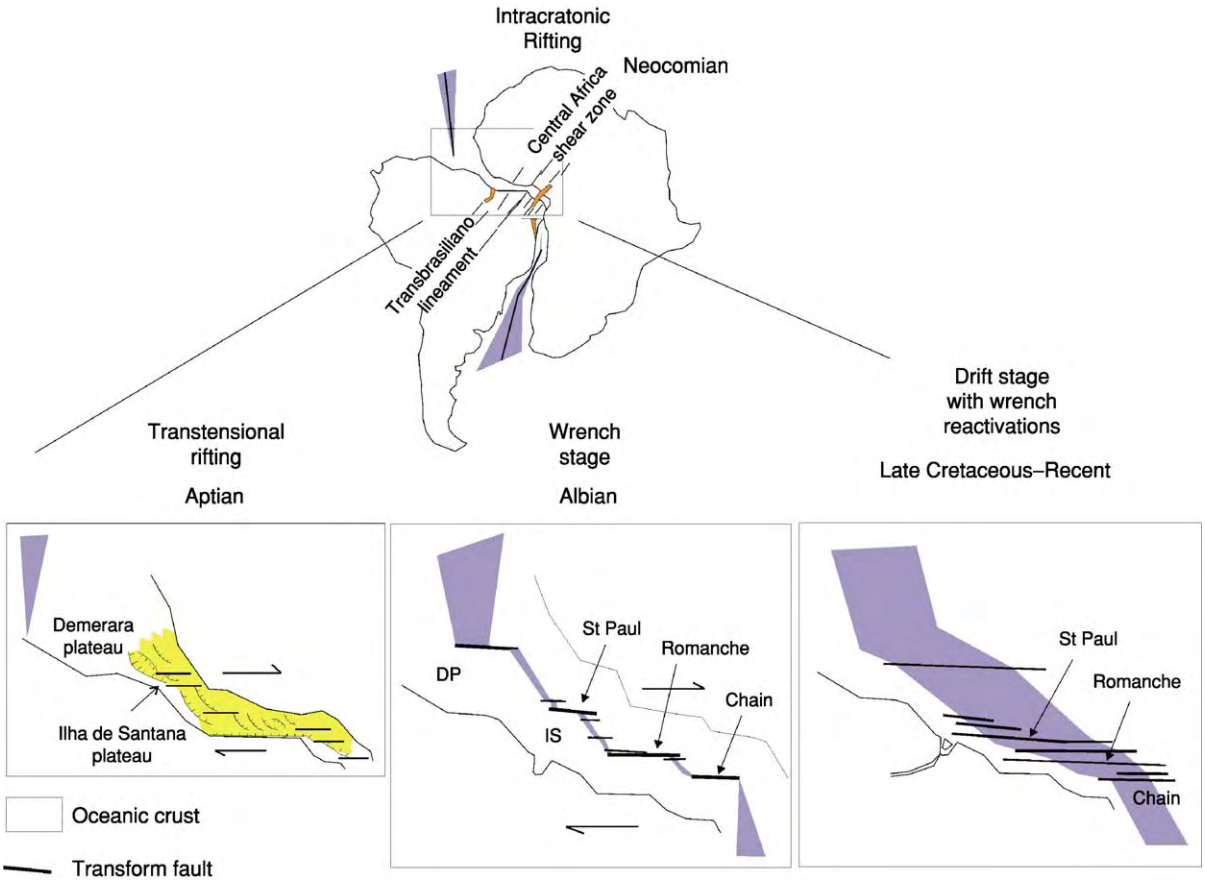


Figure 19 Tectonic evolution of the Brazilian equatorial transform margin. Modified from Falkenhein K and Martins Neto MA (eds.) (2003) *Equatorial Margin Project, Part 2: Sequence Stratigraphy, Turbidites, Tertiary Wrench, Petroleum Systems*. Multiclient project with Petrobras, Agip, Shell, Chevron Texaco, JNOC, ConocoPhillips, Encana, ElPaso, KerrMcGee, Statoil, Devon, Ouro Preto, Fundação Gorceix NUPETRO. Final report (unpublished).

basins is still controversial. Some deeper-processed seismic lines and gravity–magnetometry maps suggest a steer-head geometry, with a rift basin beneath the main sag (Figure 13).

Some authors have considered the Paraná and Acre basins to be foreland basins, because of their possible relationships to accretionary processes along the margin of western Gondwana (Figure 14). The Acre Basin is the name for the eastern pinch-out of the Peruvian sub-Andean Marañon-Ucayali foreland wedge. The Paraná Basin comprises three basin cycles: an Ordovician–Silurian thermal sag; a Devonian–Triassic foreland basin; and a Jurassic–Cretaceous sag associated with the opening of the Atlantic Ocean (Figure 15). Permo-Carboniferous glacial deposits and an enormous volume of Early Cretaceous (137–128 Ma) flood basalts also characterize the Paraná Basin fill.

Continental-Margin Basins and Associated Interior Rifts

The breakup of western Gondwana in the Early Cretaceous created the continental margins of South America and Africa. During the Neocomian, Barremian, and Aptian, rifting propagated from south to north, generating the eastern Brazilian margin basins. The north-west–south-east-orientated Brazilian equatorial margin opened later, between the Aptian and the Albian, as a transform margin. In the continental interior, Precambrian lineaments were reactivated to form interior rift basins.

Eastern Brazilian margin basins The eastern margin basins are classical passive-margin basins, comprising three sequences, reflecting precursor rift, transitional proto-oceanic, and thermal drift (Figure 16).

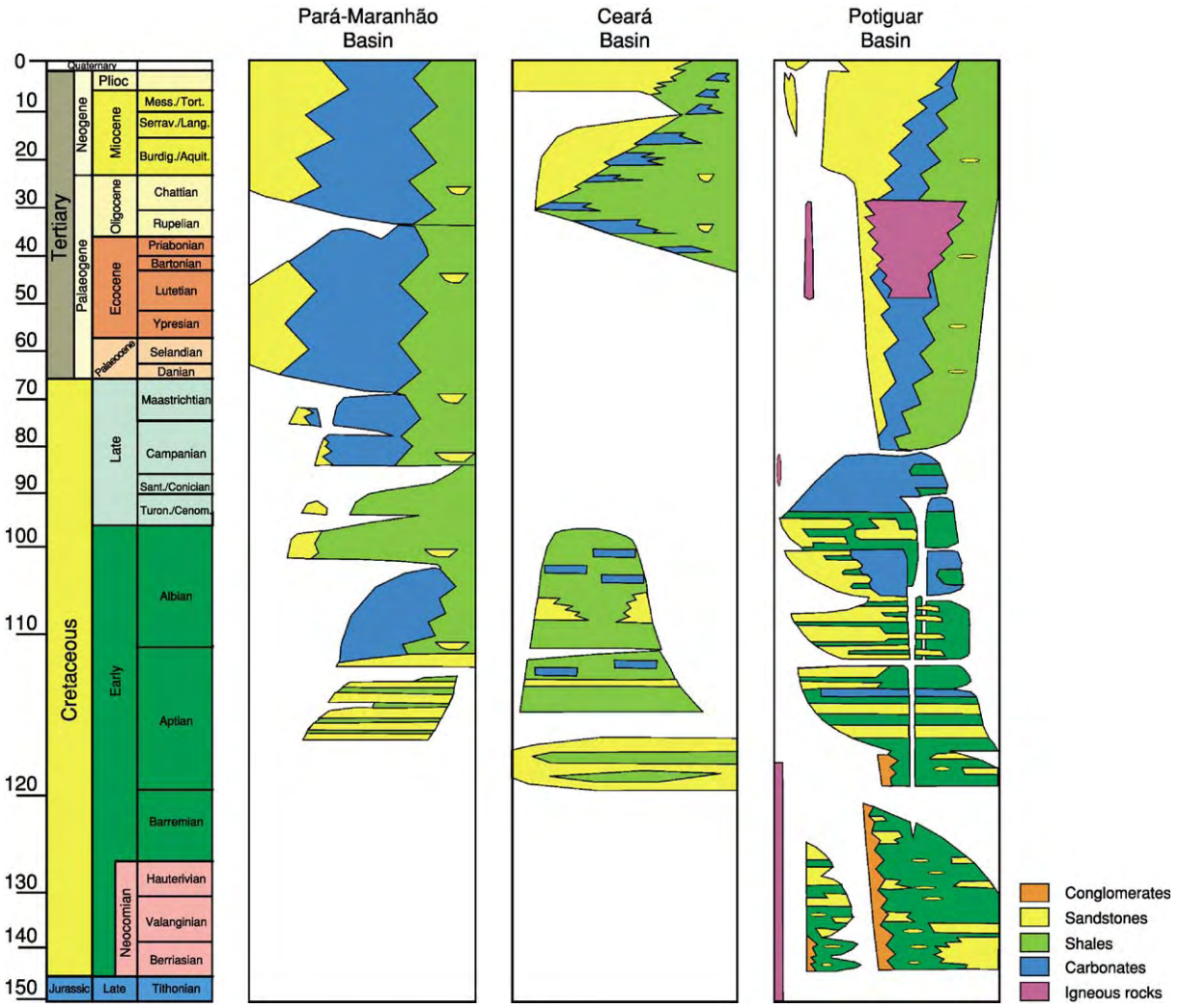


Figure 20 Stratigraphical charts of the Pará Maranhão, Ceará, and Potiguar Basins of the Brazilian equatorial margin. Modified from the Brazilian Petroleum Agency (Agência National do Petrõleo, ANP) home page at www.anp.gov.br

Neocomian to Early Aptian rift-stage deposits consist of basalts and continental siliciclastics, deposited in depocentres controlled by normal faulting in alluvial, fluvial, deltaic, and lacustrine environments.

The Aptian transitional proto-oceanic supersequence, characterized by low subsidence rates, marks the incursion of marine waters into the basin system, and comprises coastal to shallow-marine carbonates and mudstones and thick evaporites (Figures 16 and 17). Salt was deposited in the central segment of the eastern margin (from the Santos to the Sergipe-Alagoas basins; Figure 11) in an evaporative basin created by a volcanic barrier, the Walvis–São Paulo Ridge (Figure 18).

The first-order drift package comprises an Albian–Cenomanian retrogradational transgressive supersequence, a Cenomanian–Turonian maximum flooding interval, and a Late Cretaceous–Holocene progradational regressive supersequence. Whereas the transgressive supersequence is mostly carbonate, the regressive supersequence comprises unconformity-bounded third-order depositional sequences of turbidite-prone lowstand systems tracts and siliciclastic or, subordinately, mixed-carbonate transgressive highstand systems tracts (Figure 17). Salt tectonics (deformational structures in response to gravity driving motion of salt layers) controlled the sediment bypass and the main depocentres, as well as the architecture of the drift package (Figure 16).

The Campos Basin in the southern segment of the eastern margin (Figure 11) contains approximately 85% of the recoverable hydrocarbon reserves in Brazil, forming together with the adjacent Santos and Espirito Santo basins the most prolific oil province of Brazil.

Equatorial margin basins The Brazilian equatorial margin evolved in four tectonosedimentary stages (Figures 19, 20 and 21): Neocomian interior rift; Aptian transtensional rift; Albian wrench; and Cenomanian–Holocene drift with wrench reactivations.

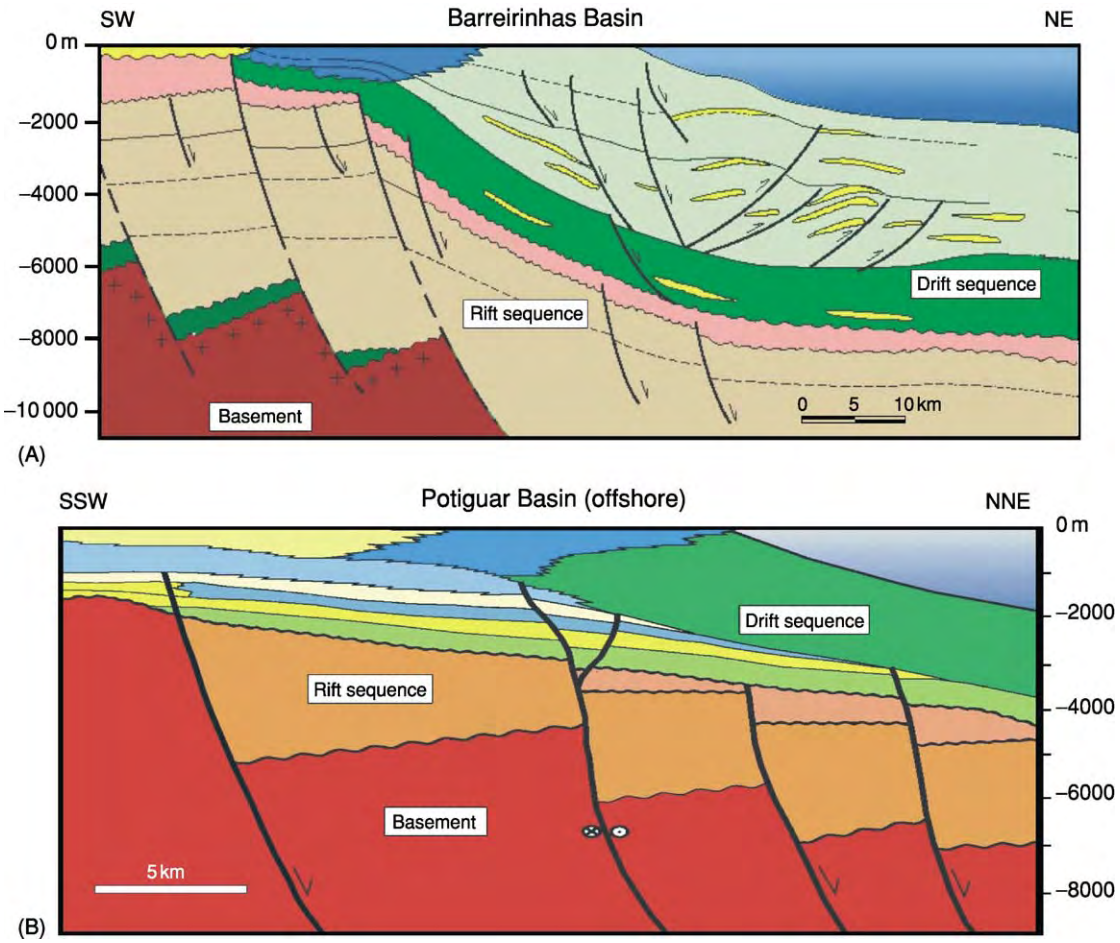


Figure 21 Schematic geological sections through the Brazilian equatorial margin (A) Barreirinhas and (B) Poriguar offshore basins. Modified from the Brazilian Petroleum Agency (Agência National do Petrõleo, ANP) home page at www.anp.gov.br

During the Aptian, after the development of the Neocomian interior rifts, the connection between the Central and South Atlantic oceans was initiated by the nucleation of an east–west-trending mega-shear corridor, which controlled the evolution of the Brazilian and West African equatorial margins until the full breakup in the Early Cenomanian. This dextral transtensional corridor was made up of north-west–south-east-orientated releasing bends, alternating with east–west-orientated transfer segments (Figure 19). The releasing bends hosted a series of frontal rifts, whereas the transfer segments evolved as sets of en echelon oblique rifts and east–west-orientated transfer faults. During the Albian, with the establishment of oceanic spreading centres in the north-west–south-east-trending releasing bends, a series of east–west-trending transform faults nucleated, among them the St Paul, Romanche, and Chain transforms. The previously formed rifts were partially deformed in the vicinity of the fault zones. Carbonate sedimentation prevailed in the releasing bends (Pará-Maranhão, Barreirinhas, and Potiguar basins), whereas syn-wrench siliciclastic sediments locally accumulated along the east–west-trending transfer segments. During the Late Cretaceous and Tertiary, strike-slip displacements along the oceanic fracture zones induced wrench deformation in the distal portions of the basins. Tertiary wrenching also triggered gravitational gliding in slope areas (Figure 21A). Siliciclastic sedimentation prevailed until the Early Eocene. Carbonate and mixed systems dominated later, except in the Amazon Cone area (see Figure 11), where a sand–shale package several kilometres thick accumulated from the mid-Miocene.

Early Cretaceous interior rifts Interior rifts formed during the Early Cretaceous in response to intraplate stresses related to the opening of the Central and South Atlantic oceans. The rift basins, the Potiguar Basin onshore, and the Tucano-Recôncavo-Jatobá and Tacutu rifts (Figure 11), display half-graben geometry and are filled with continental siliciclastic rocks (Figure 22).

Tertiary Rifts and Related Features

An episode of rifting, whose driving mechanism is still unknown, affected the whole eastern margin of Brazil at the end of the Eocene and beginning of the Oligocene. Reactivation of the Early Cretaceous rift structures, basic to alkaline magmatism, and nucleation of a series of small rifts along the Serra do Mar on the south-eastern coast are the main manifestations of this event.

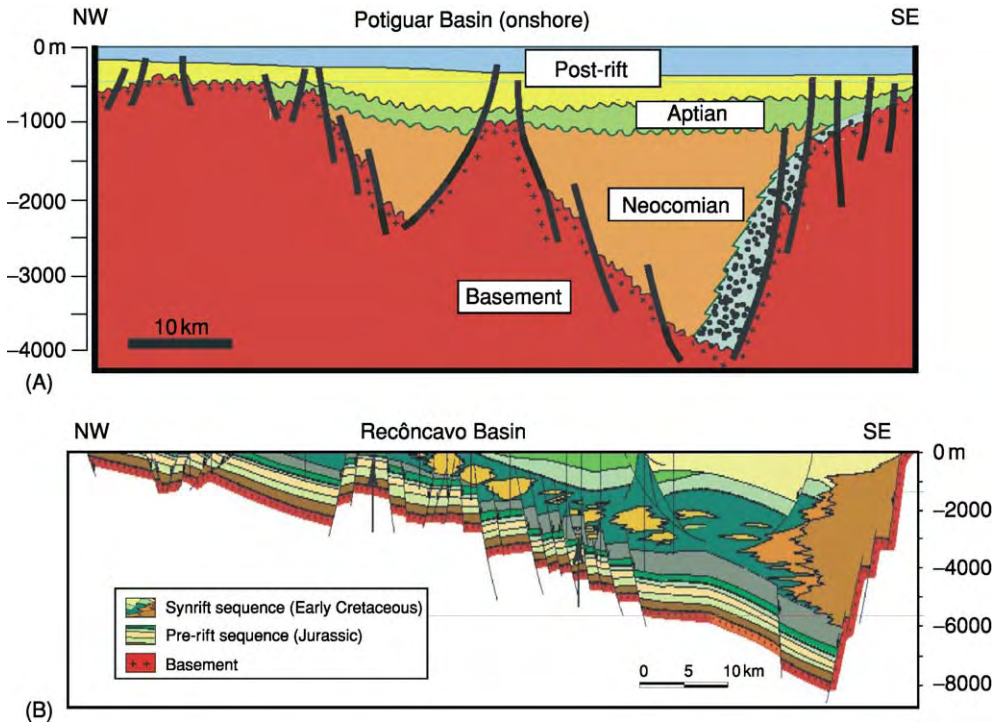


Figure 22 Schematic geological sections through the (A) Potiguar onshore and (B) Recôncavo interior rift basins. Modified from the Brazilian Petroleum Agency (Agência National do Petrõleo, ANP) home page at www.anp.gov.br

Glossary

Craton An ancient platform, i.e. a stable portion of a continent, not affected by Phanerozoic orogenies. Cratons also are differentiated pieces of continental lithosphere, characterized by a relatively thin crust coupled to a mantle root, whose thickness can reach up to 400 km.

Geochronological province Region in which a characteristic geochronological pattern predominates and dates obtained by distinct methods in different rock units are coherent.

Platform Relatively stable portion of a continent, not affected by a given orogenic event.

Shields Areas of the continental interior where Precambrian basement is exposed.

See Also

Africa: Pan-African Orogeny. **Gondwanaland and Gondwana**. **Mesozoic:** Cretaceous. **Plate Tectonics**. **Precambrian:** Vendian and Ediacaran. **Shields**. **Tectonics:** Mountain Building and Orogeny; Rift Valleys.

Further Reading

Almeida FFM, Hasui Y, Brito Neves BB, and Fuck RA (1981) Brazilian structural provinces: an introduction. *Earth Science Reviews* 17: 1 29.

Almeida FFM, Brito Neves BB, and Dal Ré Carneiro C (2000) The origin and evolution of the South American platform. *Earth Science Reviews* 50: 77 111.

Bizzi LA, Schobbenhaus C, Gonsalves JH, *et al.* (2001) *Geology, Tectonics and Mineral Resources of Brazil: Geographic Information System and Maps at the 1 : 2 500 000 scale.* Brasília: CPRM Serviço Geológico do Brazil. [CD ROM. Online at http://www.cprm.gov.br].

Brito Neves BB, Campos Neto MC, and Fuck RA (1999) From Rodinia to Western Gondwana: an approach to the Brasiliano Pan African cycle and orogenic collage. *Episodes* 22: 155 199.

Chang HK, Kowsmann RO, Figueiredo AMF, and Bender AA (1992) Tectonics and stratigraphy of the East Brazil rift system: an overview. *Tectonophysics* 213: 97 138.

Cordani UG and Sato K (1999) Crustal evolution of the South American Platform, based on Sm Nd isotopic systematics on granitoid rocks. *Episodes* 22: 167 173.

Cordani UG, Milani EJ, Thomaz FA, and Campos DA (eds.) (2000) *Tectonic Evolution of South America.* 31st International Geological Congress, Rio de Janeiro.

Heilbron M, Mohriak WU, Valeriano CM, *et al.* (2000) From collision to extension: the roots of the southeastern continental margin of Brazil. In: Mohriak WU and Talvani M (eds.) *Atlantic Rifts and Continental Margins,* pp. 1 32. Geophysical Monograph 115. Washington DC: American Geophysical Union.

Matos RMD (2000) Tectonic evolution of equatorial South Atlantic. In: Mohriak WU and Talvani M (eds.) *Atlantic Rifts and Continental Margins,* pp. 331 354. Geophysical Monograph 115. Washington DC: American Geophysical Union.

Mello MR and Katz BJ (2001) *Petroleum Systems of South Atlantic Margins.* AAPG Memoir 73. Tulsa: American Association of Petroleum Geologists.

Schobbenhaus C, Campos DA, Derze GR, and Asmus HE (1984) *Geologic Map of Brazil and Adjoining Ocean Floor Including Mineral Deposits Scale 1 : 2 500 000.* Brasilia: Ministério das Minas e Energia.

Tankard AJ, Soruco RS, and Welsink HJ (1998) *Petroleum Basins of South America.* AAPG Memoir 62. Tulsa: American Association of Petroleum Geologists.

Trompette R (1994) *Geology of Western Gondwana (2000 500 Ma). Pan African Brasiliano Aggregation of South America and Africa.* Rotterdam: Balkema.

BUILDING STONE

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

Unlike most other materials available to early man, stone was the most durable and could be stacked in courses to make walls to support roofs and thereby afford protection from the elements. Building stone continues to be used, valued for its strength and for aesthetic beauty purposes. Hence there is now and will remain a market for the technical talents of geologists capable of locating and selecting deposits of durable and attractive building stone, and in scoping and devising the environmental protection measures now necessary in extraction of building stone.

Historic Use of Building Stone

Historically, building stone has been employed for a wide variety of structural and load-bearing purposes (**Table 1**) Facade use of building stone was popular in the world's cities from the very advent of the high-rise building (about 1890) and this situation stimulated

Table 1 Historic and modern uses for building stone

Use	*Historic*	*Modern*
Housing	Generally structurally practical only to one or two storeys	Replaced since mid nineteenth century by brick and timber frames with occasional stone facades
Fortifications	Durable material; employed in massive, protective wall	Replaced by reinforced concrete since 1900
Road/street paving	Sole material for all weather roads and streets	Infrequently used and then only for aesthetic purposes; replaced since 1940 by use of crushed stone in asphaltic concrete
Towers	Watch towers and lighthouses; thick walls required for stability and resistance to attack	Replaced by reinforced concrete in order to meet cost of construction labour
Water supply	Traditional usage for aqueducts and for canal linings	Largely replaced by iron and steel conduits by 1900; lately by synthetic organics (plastics)
Sewers	Stone replaced by brick in mid nineteenth century	Obsolete today due to leakage and high cost of emplacement labour
Commercial buildings	Mainly as facades	Used as facades on steel framed buildings up to ten or more stories
Civil buildings	Limited to two storeys or requiring massive wall thickness	Restricted to facades, for aesthetic purposes
Mansions and institutions	Preferred construction material for wealthy patrons	Employed only for special esthetic reasons due to high costs of placement
Houses of worship	Traditional; thick walled or with flying buttresses	Now replaced with reinforced concrete, often with stone as cladding
Monuments	The preferred material	Remains the preferred material
Fireplaces and furnaces	Replaced by firebrick by ca 1840 for furnaces	Remains popular for fireplace facing; now used in conjunction with steel or ceramic flues
Wall dressing, interior and exterior	Seldom used as was trivial and unnecessary	Now the prime use of building stone
Bounding walls	Commonplace where rock outcrops are exposed for quarry removal	Ongoing demand in conservation areas and national parks

Table 2 Attractions and qualities of natural building stone

Quality	*Nature*	*Geological consideration*
Durability	When other than 'weak' rock, stone has been conditioned by the attack of geological agents and properties in preparation for its selection and use today	Most potential physical/chemical degradation has already occurred and the stone likely will be durable in its intended construction or architectural use
Resistance to abrasion	Necessary quality for stairs, walkways, and road and drive paving	Choice generally relegates to crystalline igneous or metamorphic rock type low in presence of micas
Variety of intact dimensions	Generally a result of geological depositional conditions of origin or of diagenetic or metamorphic changes that promote jointing and fissility as a result in natural separation into articulated blocks, preferably in a rectilinear pattern	Generally the result of either or both: 1. Joint spacing in igneous or crystalline metamorphic work 2. Foliation jointing in high grade metamorphic rock 3. Fluvial rounding in high energy stream regimes, often associated with paleglaciation
Planarity or flatness; when so selected	Considered 'flagstone' when length or breadth to thickness ratio is equal to or greater than about 10:1	Influenced by fissility imparted by sedimentologic (bedding) character or by metamorphically induced open foliation
Colour	Generally a wide variety, to please the client Colour variability generally greatest in volcanic terrain, such as Italy	Directly related to lithology and geological origin Locating geologist should examine for absence of reactive minerals such as iron pyrites
Texture	To please the client; but selected to avoid unsightly stains	Visible as a form of surface roughness Directly related to lithology and geological origin

Table 3 Basic geological controls over nature of building stone

Control	*Defined*	*Geological considerations*
Sedimentary rock	Most varieties are unsuitable for use as building stone when uniaxial compressive strength exceeds about 15 MPa	Rock younger than Cretaceous (as Tertiary era) aged tends to be 'weak' and unsuitable because of immature induration (density) High clay mineral content generally promotes unsuitability by way of spalling and deterioration in exposure to the elements
Glacial ice contact deposits	Often represents ground rich in durable and colour lithologic diverse specimens	Only the most durable rock survives the glacial plucking and transport experience Often excellent fireplace and pillar stone Requires considerable experience and labour in selection of useful individual stones
Rock lying in tectonic shear zones	Tends to be mylonitinised or otherwise shear damaged and therefore excessively fissile	Generally damaged beyond potential resistance to long term exposure to the elements in exterior applications
Pluton bounding zones of foliation in granitic rock	Lends itself well as a fall facing	Often has a pleasing and useful foliation character leading to slightly planar dimension ratio
Pluton capping rhyolites in post Jurassic terrains	Somewhat rare due to removal by erosion	Generally durable in climatically temperate regions while exhibiting good planar dimensions for wall facing
Streams and rivers draining tertiary volcanic centres of moderately high rainfall or subject to cloudburst type periodic events	Usually present as boulders of a variety of colours and textures	Fluvial action destroys weak varieties and delivers only the most durable specimens for collection
Foliated metamorphic rock in general	Foliated rock splits or rips well with minimum to no application of explosives	Tends to scar when removed and handled by machinery; requires care in recovery at quarries
High grade metamorphic rock in general (Gneisses)	Gneisses and migmatites generally suitable only as having experienced glacial action	As grain size diminishes to that of schist, desirable building stone qualities decrease; especially in the presence of micas
Rock within the contact aureole of a few hundred meters of intrusive igneous bodies	Sometimes visually undetectable cation exchange effects generated by hydrothermal effects of intrusion	Best to place samples in the weather for observation; best to use application of wet dry interludes of testing by laboratory cyclic wetting, and drying
Aggregate pits servicing great urban centers on the fringes of igneous metamorphic complexes	Pits will be located along modern rivers and streams colocated with Plio Pleistocene to recent fluvial deposits	Pits operating by suction dredging tend to have a 'bone pile' of vertebrate fossil bones

great interest in the national geological surveys. Most geological surveys produced serial sets of geological technical monographs on building stone and these books today are the most valuable published references for those geological professionals who are interested in building stone.

Geological Character of Building Stone

Stone possesses several strong qualities, valued by clients and architects and was so recognized in the early textbook on Engineering Geology in which Penning (1880) cautioned that:

> "In the selection of a stone, it is not merely in the testing it by hardness, composition, and appearance, that judg ment should be displayed, but also in ascertaining the conditions under which it lay before removal from its position in the quarry."

These qualities (Table 2) have geological origins, most effectively identified by professional geologists.

Geological Controls on Nature of Building Stone

Geological conditions exert first-order control over the kinds of building stone available within a given economic radius for recovery. These controls are largely

Table 4 Geologic tasks in prospecting for building stone

Task	*Defined*	*Considerations*
Determine general nature of stone desired	Nature of stone determines the likely source geological formation(s) to be prospected	Reduces the exploration task to identification of suitable terrain in lands available for recovery
Determine the economic radius for stone recovery	This defines the area of exploration	Acquire medium scale geological maps of the area of economic consideration
Assess suitability of deposit to produce stone of desirable end use qualities	1. Dimensions on initial recovery, without treatment 2. Durability as exposed in place, on construction	Most stone undergoes physical/chemical adjustment to stress removal on excavation and in the new environment of placement
Assess ultimate quantity of rock available	Must recognise costs involved in recovery site development	Quantity as affected by property bounds and stability of cuts made to recover the stone
Evaluate means of removal of stone	Cost of production of each cubic meter of stone prepared for shipment to market	Suitable rate of production must be balanced by lack of unacceptable damage to the produced stone
Evaluate cost of production of units of the stone (cubic meter)	Placement of accumulated stone produce between haul loads	Most stone today is palletised for unit transport packages of one cubic meter; avoids labour costs in distribution to building sites
Evaluate cost of transport of units of stone (cubic meter) to market areas	Resting place for spoil accumulated from stone removal and selection for transport	Ideally removed periodically to the final disposal site, in accordance with quarry closure plans or for sale as off site embankment or fill material
Assess unit cost impact of providing environmental protection	Planning and permitting, and meeting environmental requirements of the permit	1. Control of sedimentation and dust 2. Post closure land use 3. Implementing approved plan of closure of the quarry

Table 5 Types of stone masonry

Type	*Description*	*Considerations*
Ashlar	Individual blocks of rock are cut or trimmed, Broken: As placed, the joints are not continuous Small: When individual stones are less than 300 mm thick Rough: Squared stone masonry laid as 'range' work	Masons attempt to make use of a high percentage of rock as delivered from the quarry, sized in accordance with provisional specifications of supplier's contract, not necessarily of comparable size or ratio
Squared stone	Masonry in which individual stones are roughly squared in layers and roughly dressed on beds and joints	Distinction between Ashlar and Squared stone is that the latter generally is laid dry or with mortar spacing being 150 mm or greater between surfaces of stones
Quarry faced	Stone facing is left untreated as it comes from the quarry	Commonly jointed or bedded stone comes with iron or manganese stained natural planar surfaces
Cut and polished stone	A relatively new market with durable stone being sold mainly for interior use as flooring and counter surfaces	Durable stone is recovered from 'the ends of the earth' and sold at high prices controlled by vendor tradesmen
Rubble	Consists of stone unshaped by hand or machine and placed in its unaltered condition from the quarry	Applied mainly to river run fluvial boulders previously shaped by the stone to stone contact in transport from outcrop to mined deposit Also applies to informal recovery of naturally discontinuous rock recovered from small quarries by fairly intensive hand labour
Dressed rubble	Some attention to rough sizing by manual hammering	May be dressed by tensile splitting of field broken or split surfaces
Range work	Exposed fact of individual stone is roughly dressed	Laid in rough courses
Broken range work	Courses are not continuous throughout the wall or face	Facilitated with joint bounded rock blocks
Random work, or 'uncoursed rubble'	Unsquared stone without attempted regular courses	Laid without attempt to achieve horizontal or sub horizontal levelling between individual courses of stone
Coursed rubble	Sometimes roughly shaped by hammer so as to fit approximately	Unsquared stone levelled off at specific heights to an approximately horizontal surface of each course

Table 6 Sequential environmental permitting planning for recovery of building stone

Step	*Considerations*	*Geologic detail*
Topographic layout	Defines where the recovery is to take place	Scale of 1:1000 1:2000 Contour interval of one meter
	Enlarged ordnance map or plane tabled topographic map of deposit bounds	Property bounds to 100 m lateral buffer space Present hydrological features Present cultural features
Exposed geological detail	Contact of soil with bedrock, if exposed or as predicted	Include notion of geological structure
Proposed product stone	Dimensions and character How stored	1. Note possible variations in the character of stone across the breadth and depth of the planned quarry
	Frequency and method of removal from the site	2. Subdivide the quarry into separate ''domains'' if differences exceed the flexibility of marketing
Mining concept	Explains how the deposit will be opened up and how the stone will be accessed	1. Cutting bounds, crown of cut and toe of slopes 2. Sequential advance of the cutting with final proposed lateral bounds 3. Nature of machines to be employed 4. System of parting the stone from the outcrop or rock mass 5. Use of explosives
Management of runoff	Describe how precipitation and snowmelt will be managed	Indicate presence of drainage gradient and channelization to be placed to remove runoff to suitable discharge
Management of spoil	Describe dust, fine particles and sediment likely to be generated Rejection policy/procedure for discrimination of stone vs. spoil	Show locations for interim spoil placement
Determine general nature of stone desired	Nature of stone determines the likely source geological formation(s) to be prospected	Reduces the exploration task to Identification of suitable terrain in lands available for recovery
Determine the economic radius for stone recovery	This defines the area of exploration	Acquire medium scale geologic maps of the area of economic consideration
Access and haul roads	Within about 8 km of centre of area of interest	Examine road cuts and outcrops to ascertain the general presence of suitable stone
Product storage area	Placement of accumulated stone produce between haul loads	Most stone today is palletised for unit transport packages of one cubic meter; avoids labour costs in distribution to building sites
Spoil storage area	Resting place for spoil accumulated from stone removal and selection for transport	Ideally removed periodically to the final disposal site, in accordance with quarry closure plans or for sale as off site embankment or fill material
Sequential quarry reclamation	Ongoing plan of converting incremental phase areas of the overall stone removal site to suit the environmental permitting permit	Normally comes as a consequence of working within site ownership bounds and in consideration of stability of the highwall resulting from removal of stone from the bedrock outcrop
Management of incidental waters	1. Runoff from rainfall and snowmelt 2. Accommodation of seepage from the highwall	Requires temporary ditching and/or area sumps and pump withdrawal to authorised discharge point
Management of sediment	Dust and small particles subject to wind and runoff transport from the quarry to the environment	Generally managed by placement of sedimentation sumps and bales of cut vegetation such as 'hay' Requires ultimate on site manipulation as part of the closure plan
Closure plan	Presumes that the quarry will be provided with an acceptable after use	Should be worked into the long term operational plan and be in conformance with the required end use of the operating permit

related to lithology, conditions of origin, and geological history (Table 3) and are reflected in mineralogy and petrography, and include fabric and texture. Large expenditure on the acquisition of building stone should include a petrographical analysis for the latter features.

Locating Sources of Building Stone

Building stone has a relatively low economic value in the outcrop, its true value accruing from its placement work by stone mason tradesmen. Its recovery is expensive in terms of specialized skills and by the judicious use of explosives and of construction machinery for its recovery.

A number of factors related to prospecting for building stone deposits are well suited to engineering geological training (Table 4).

Environmental Planning for Recovery of Building Stone

As with most earth materials, there are two levels of activity that apply to recovery of building stone, informal and formal. For all but the weekend mason, all other recovery activities quickly will be recognized by environmental regulatory authorities as a form of 'quarrying', for which even a small stone recovery enterprise will require submittal of a Permit Application and some scheme of site development and integrated measures for protection of the environment (Tables 4 and 6).

Stone Masonry

Modern stone masonry nearly always serves the aesthetic purpose of the architect or landscape designer. No longer are we concerned with use of stone for its original structural role, but the manner in which the masonry is placed provides ample opportunity for meaningful artistic statement (Table 5).

Petrography

Petrographical examination of thin-sectioned hand specimens recovered from prospecting will be helpful in some instances in which the shape, colour, and/or texture meets the client's approval, but the geologist remains wary of the ability of the stone to meet expectations when placed in exterior locations exposed to the elements. As with basic identification of the stone, such manuals as listed in the **Further Reading** Section below provide an excellent desk reference.

Summary

Building stone provides a durable, aesthetically pleasing construction material where wear-resistant and climate-resistant natural stone can be used to enhance the qualities of modern construction. Significant costs must be paid for accessing and recovery machinery, palleting, and transportation to the vendor's yard. Application of geological knowledge is essential to providing appropriate attractive and durable stone at an economic production cost.

See Also

Aggregates. **Engineering Geology:** Rock Properties and Their Assessment; Site and Ground Investigation.

Further Reading

Bates RL (1987) *Stone, Clay, Glass; How Building Materials are Found and Used*: Enslow.

Bell FG (1993) Engineering Geology: Oxford: Blackwell Scientific Publications.

Blyth FGH and de Freitas MH (1984) *A Geology for Engineers*, 7th edn. New York: Elsevier.

Geological Society of London (1999) Stone; Building Stone, Rock Fill and Amourstone in Construction. *Engineering Geology Special Publications* 16: 480.

Parsons and David (eds.) (1990) *Stone; Quarrying and Building in England*. London: Phillmore, in Association with The Royal Architectural Institute.

Shadman A (1996) *Stone; An Introduction*. London: Intermediate Technical Publications.

Smith MR and Collis L (eds.) (1993) *Aggregates; Sand, Gravel and Crushed Rock Aggregates for Construction Purposes*, 2nd edn. London: Geological Society Special Publication no. 9, p. 339.

CALEDONIDE OROGENY

See **EUROPE: Caledonides of Britain and Ireland; Scandinavian Caledonides (with Greenland)**

CARBON CYCLE

G A Shields, James Cook University, Townsville, QLD, Australia

Introduction

The recycling of elements between Earth's interior, its sedimentary sheath, the oceans, and the atmosphere is vital to the continued functioning of Earth as a living planet. Of all the biogeochemical cycles operating on Earth, the carbon cycle (Figure 1) is probably the most fundamental process, without which life could not exist. Carbon, in the form of carbon dioxide (CO_2), represents the starting component in almost all food chains, while at the same time performing a complementary role as Earth's thermostat, providing an equable climate, suitable for the retention of liquid water on the surface of the planet. Carbon in the form of bicarbonate and carbonate ions helps to regulate ocean acidity and, in combination with calcium, provides a hard skeleton for many marine organisms. Without the biological sequestering of carbon and its eventual storage in the crust through geological processes over millions of years, there would be little or no free oxygen on Earth's surface. Similarly, without the deposition of organic carbon and calcium carbonate shells on the seafloor, Earth would inevitably become too hot for even the most radical of known extremophile metabolisms. All these important roles reflect different aspects of the global carbon cycle, which is in reality a hierarchy of subcycles, all operating on different time-scales.

From the moment a molecule of CO_2 first enters the surface environment, whether from a volcano or from the chimney stack of a power station, it enters a dynamically changing realm of continual recycling. Within a decade, the CO_2 molecule will very likely be consumed by photosynthesis and converted into organic carbohydrate, either as part of a land plant or in one of the trillions of single-celled algae in the world's oceans. In either case, death and decay will eventually release the CO_2 back into the surface environment, thus closing the cycle. This 'short-term' carbon cycle may replay hundreds of times before burial finally results in the addition of this carbon to the sedimentary pile, where some of it will remain for millions of years as part of the 'long-term' carbon cycle of sedimentation, tectonics, and weathering. On such huge time-scales, the sequestering of atmospheric CO_2 during chemical weathering and the eventual redeposition as carbonate rock are generally balanced by the release of CO_2 from volcanoes and through metamorphism. Although negative feedbacks tend to regulate CO_2 levels over the long term, shorter term perturbations to the carbon cycle may occur due to pulses of massive volcanism, carbon burial, or methane clathrate dissociation, all of which may have caused considerable climatic perturbation at different times in Earth history. The carbon cycle not only acts as a trigger for climate change but can react to such change, as happened during glacial–interglacial cycles, when regional warming trends appear to have preceded CO_2 increases. Interest in the short-term carbon cycle has increased greatly of late because of the observed elevation in atmospheric CO_2 concentrations due to fossil fuel combustion, cement manufacture, and land use changes. Fears that elevated CO_2 concentrations will change the future world climate have inspired research into the effects of an enhanced greenhouse effect on climate, sea-level, and biodiversity. This global research effort necessitates a good understanding of how Earth regulates climate through the carbon cycle, including the development of increasingly sophisticated models to describe the 'Earth system'.

Short-Term Carbon Cycle

The short-term carbon cycle refers to the natural cyclical processes of organic matter production and organic decay (Figure 2). The key step in this cycle is photosynthesis, whereby CO_2 is consumed by reaction with water to form organic carbohydrates, releasing oxygen in the process (eqn [1]):

$$CO_2 + H_2O = CH_2O + O_2 \qquad [1]$$

Photosynthesis $\rightarrow\leftarrow$ Respiration

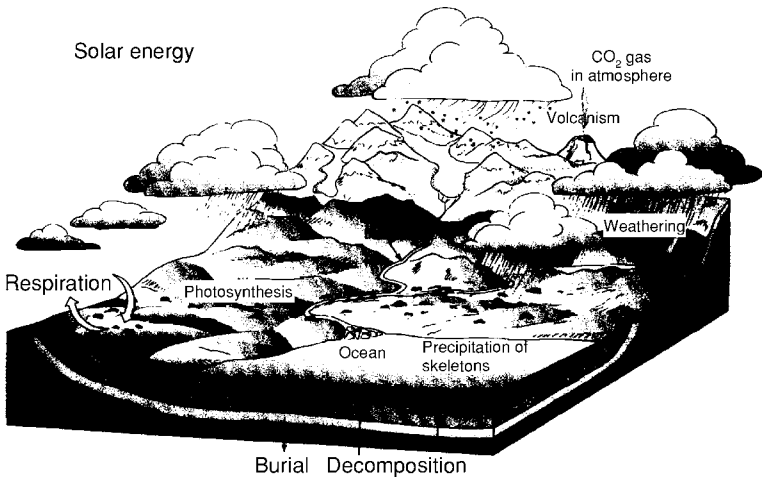


Figure 1 The natural global carbon cycle. Reprinted from Davidson JP, Reed WE, and Davis PM (1997) *Exploring Earth: An Introduction to Physical Geography.* Upper Saddle River, NJ: Prentice Hall.

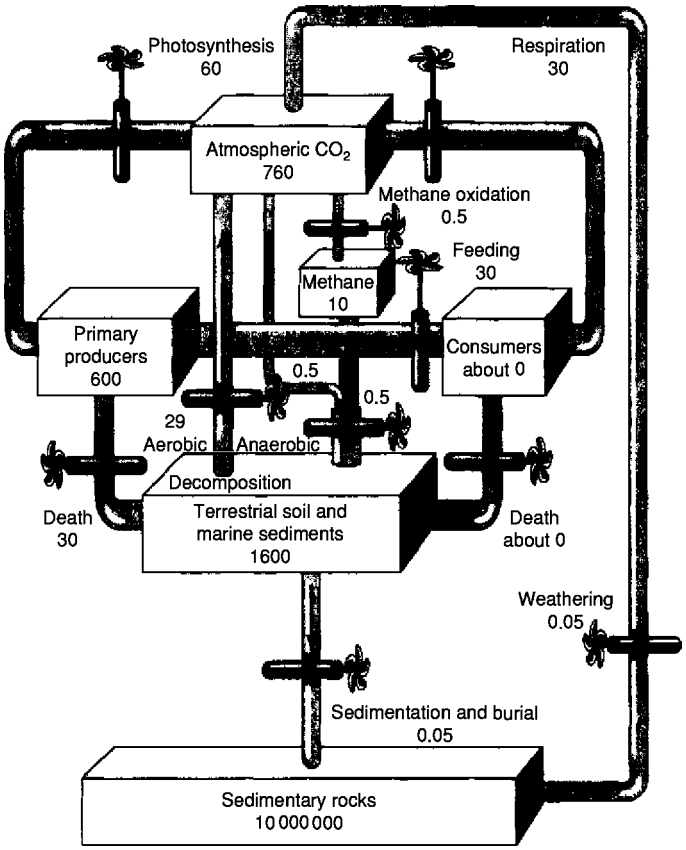


Figure 2 The short term and long term organic carbon cycles of productivity, sedimentation, burial, and weathering. Numbers refer to carbon reservoirs (in gigatons) and carbon fluxes (in gigatons year ^{1}). Reprinted from Kump LR, Kasting JF, and Crane RG (1999) *The Earth System.* Upper Saddle River, NJ: Prentice Hall.

Photosynthesis does not occur spontaneously but requires the input of solar energy. Consequently, photosynthesis can take place only where sunlight can freely penetrate, on the land or within the uppermost 100 m or so of oceans and lakes. On land, photosynthesis is carried out mostly by plants, whereas in the sea it is almost exclusively the domain of algae in the form of minute, free-floating silica-shelled diatoms

and calcite-shelled coccoliths. Just as photosynthesis always consumes CO_2, the reverse process, organic respiration or decay, returns that CO_2 after death of a plant or animal; this process is catalysed by the enzymatic action of chiefly anaerobic microbes that can be found in the digestive tracts of animals, in the soil, or beneath the seafloor. Because organic decay is essentially an oxidation process, oxygen is consumed instead of being released, which is why oceanic regions of high productivity are commonly underlain by 'oxygen minimum zones', i.e., zones of low oxygen concentration caused by the decomposition of organic matter in the water column. Organic raindown not only transfers CO_2 to the deep ocean but also acts as a biological pump for nutrients that, once stored in the deep ocean 'conveyor belt', generally reemerge only in regions of upwelling and high productivity. Reoxidation of organic matter that has reached the sediment may take place by a variety of mechanisms (oxic respiration, denitrification, or sulphate reduction). However, in addition, in the deepest, most reducing environments, methane may be produced by the actions of methanogenic bacteria using two main pathways (eqns [2a] and [2b]):

$$CH_3COOH \rightarrow CO_2 + CH_4 \qquad [2a]$$

or

$$4H_2 + CO_2 \rightarrow CH_4 + 2H_2O \qquad [2b]$$

Methane produced in this fashion may seep back into seawater to be reoxidized to CO_2 or may be stored temporarily for thousands to millions of years as the volatile methane clathrate.

Under normal circumstances, any particular ecosystem will be experiencing both photosynthesis and respiration, but in different proportions. If photosynthesis outweighs respiration, then the ecosystem will act as a sink for CO_2, whereas in the reverse case it will act as a source. This 'breathing' of the biosphere can best be seen in the annual fluctuations of atmospheric CO_2 in the northern hemisphere (**Figure 3**). During the northern hemisphere summer, when terrestrial photosynthesis and leaf growth surpass respiration and decomposition, CO_2 levels decrease by as much as 15 ppm in the boreal forest zone (55–65° N), a decrease almost completely balanced by the winter increase in CO_2 caused by the inevitable decomposition of fallen leaves. In the southern hemisphere, the cycle is reversed but the effect never attains more than 1 ppm. This hemisphere inequity confirms that such seasonal CO_2 fluctuations are driven by the terrestrial carbon cycle, rather than by the oceanic carbon cycle, there being far more terrestrial biomass in the northern than in the southern hemisphere. Because the fluxes involved in the short-term carbon cycle are so large, persistent imbalances would lead to intolerable fluctuations in atmospheric CO_2. Therefore, on timescales longer than about a century, this cycle of productivity and decay must be perfectly in balance. Nevertheless, not all CO_2 is eventually returned to the atmosphere. A small proportion of the carbon locked up in soils as organic matter may be washed away to be buried indefinitely in coastal and marine sedimentary basins. Similarly, a small proportion of marine organic matter will also escape the processes of decay by being swiftly buried in areas of high sedimentation rate. Organic burial constitutes a leak in the short-term carbon cycle, whereby a relatively small proportion of CO_2 is continually removed from the surface environment to be stored within Earth's crust. Some of this leaked CO_2 may eventually be

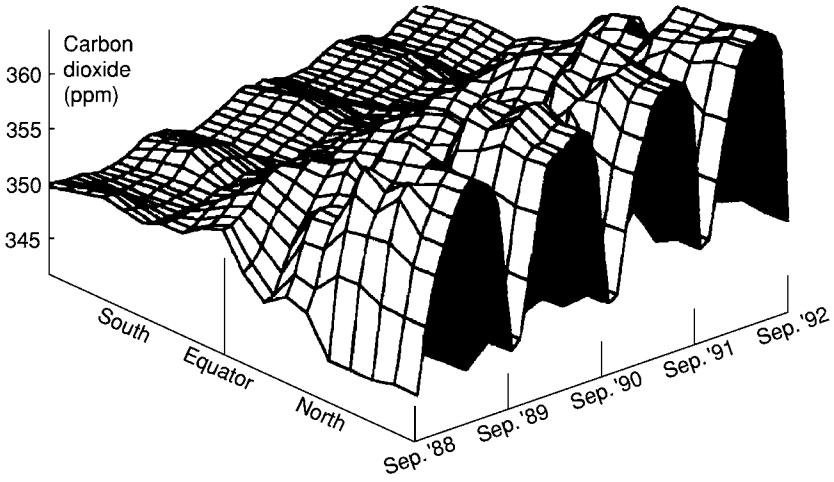


Figure 3 A three dimensional perspective of terrestrial biosphere breathing. Reprinted from Volk T (1998) *Gaia's Body: Toward a Physiology of the Earth*. New York: Springer Verlag.

returned to the surface environment by tectonic processes at convergent margins (deep-sea trenches), but not for many millions of years, as part of the 'long-term' carbon cycle of carbon burial, subduction, uplift, weathering, and carbonate deposition.

The Long-Term Carbon Cycle

Although some carbon reaches Earth from space in the form of impacting comets, most carbon first enters the surface environment as volcanic CO_2, originating from deep within Earth's interior at mid-ocean ridges, convergent margins, and the Great Rift Valley and other terrestrial volcanic provinces (e.g., Kamchatka). Some of this carbon is entirely new to the surface environment, being derived directly from the mantle, but some will be 'old' carbon, recycled from the sediment pile by tectonic processes such as metamorphism. At convergent plate margins, carbonate sediments can be carried to great depths, up to hundreds of kilometres, riding on subducting ocean lithosphere. The high pressures and temperatures at great depths encourage biogenic calcite (calcium carbonate) to react with biogenic and detrital silica (silicon dioxide) to form metamorphic calcium silicate minerals. This process releases CO_2 (eqn [3]) that may eventually find its way into the atmosphere via hot springs and seeps (**Figure 4**).

$$CaCO_3 + SiO_2 \rightarrow CaSiO_3 + CO_2 \qquad [3]$$

Much study is devoted to estimating the proportion of genuinely new CO_2 from the mantle relative to recycled CO_2 at convergent margins, but there is currently no consensus. If volcanic outgassing were allowed to proceed unchecked, then atmospheric partial pressure of CO_2 (pCO_2) would rise until an equilibrium concentration was reached, balanced only by escape of CO_2 into space. Exactly how hostile such an Earth would be can be appreciated by comparison with the furnace-like Venus, Earth's planetary neighbour and twin. Although Venus and Earth are of a similar size and are a similar distance from the sun, Venus has a much thicker atmosphere, made up almost entirely of CO_2, over 200 000 times more concentrated than in Earth's atmosphere. There is no reason to suspect that Venus overall is any richer in carbon than Earth is, so Earth must possess powerful mechanisms of CO_2 sequestration, unique in the solar system.

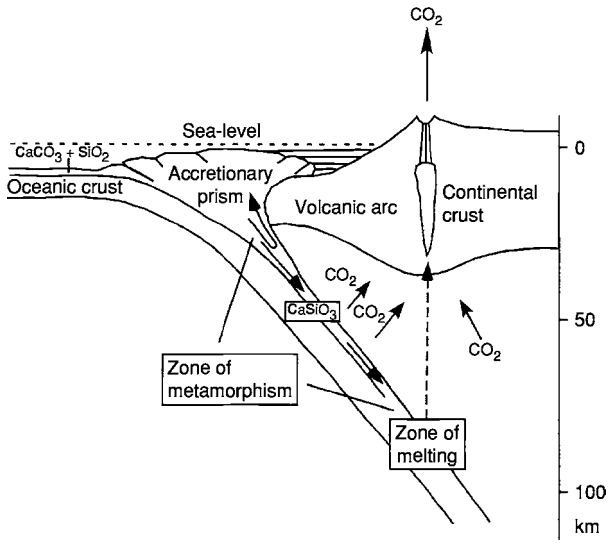


Figure 4 The long term carbon cycle of sedimentation, subduction, and outgassing.

The first step in the permanent removal of CO_2 from the atmosphere is chemical weathering (**Figure 5**), whereby rainwater, made acidic by the dissolution of aqueous CO_2, corrodes rock, forming minerals such as silicates and carbonates. In the presence of soil, this process is accelerated by plants, which help to concentrate CO_2 in the soil to levels generally 10 to 100 times higher than in the atmosphere. CO_2 used up in this way may then be transported via rivers and groundwater to the oceans, chiefly as bicarbonate anions (HCO_3), along with the weathered-out major cations such as Si, Al, Ca, Mg, K, Na, and Fe. Once in the ocean, some CO_2 may be removed permanently from the exogenic system by chemical precipitation as calcium carbonate, mostly in the form of shells and skeletal elements that rain down to the seafloor. Because $CaCO_3$ precipitation also releases CO_2 back into the water column (eqn [4]), CO_2 taken up by the weathering of carbonate rocks does not result in any long-lasting net change to atmospheric CO_2 levels. In other words, the CO_2 taken up during carbonate dissolution is simply returned during carbonate precipitation in the ocean:

$$CaCO_3 + CO_2 + H_2O = 2HCO_3 + Ca^{2+} \qquad [4]$$

Carbonate weathering →← Carbonate precipitation

Similarly, CO_2 consumed during the weathering of Na and K silicates does not result in any net change in atmospheric CO_2 because Na and K carbonate minerals are very soluble and do not readily precipitate from seawater. Equation [5] shows the net result of Ca–Mg silicate weathering, which is the most important mechanism of permanently removing CO_2 from the atmosphere:

$$(CaMg)SiO_3 + CO_2 \rightarrow (CaMg)CO_3 + SiO_2 \qquad [5]$$

Much research effort has been expended to better understand the controls on chemical weathering rates

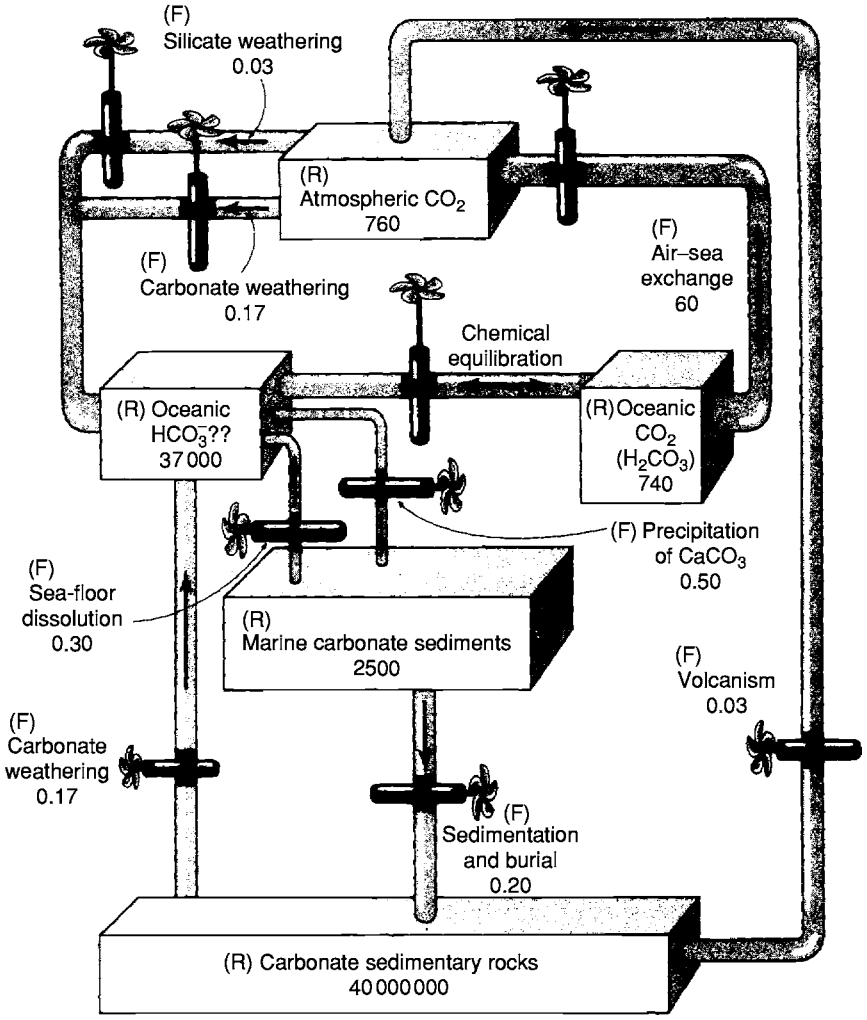


Figure 5 The global inorganic carbon cycle of weathering, runoff, carbonate precipitation, and burial. Numbers refer to carbon reservoirs (R; in gigatons) and carbon fluxes (F; in gigatons year ^{1}). Reprinted from Kump LR, Kasting JF, and Crane RG (1999) *The Earth System*. Upper Saddle River, NJ: Prentice Hall.

and therefore the effects of these controls on the long-term global carbon cycle. There is no consensus on these issues because several unrelated factors are involved, such as climate (both rainfall and temperature), denudation rates, weathering history, and rock type. Through a combination of these factors, certain tropical, humid areas of the world with freshly exposed volcanic rock can exert an outsized influence over the global carbon cycle. This is because fresh basalts are particularly susceptible to chemical weathering and are made up predominantly of Ca–Mg silicates. The weathering of basalts today may contribute as much as a third of all the silicate-weathering CO_2 flux. Thus volcanism acts not only as a source for atmospheric CO_2, but also as one of the major long-term sinks.

The products of chemical weathering, including Ca, Mg, and bicarbonate ions, arrive in the oceans in solution; concentrations build up in the oceans until an equilibrium state is reached between input and carbonate precipitation. Bicarbonate (HCO_3) and carbonate (CO_3^{2-}) ions, which together make up 99% of all dissolved carbon in seawater, help to regulate seawater pH by transferring protons (or hydrogen ions) between carbonate species (eqn [6]) according to the rules of carbonate equilibria. In effect, the carbonate system in seawater is capable of neutralizing a large portion of any extra CO_2 added from, say, fossil fuel burning or volcanism.

$$CO_2 + H_2O = H_2CO_3 = H^+ + HCO_3 = 2H^+ + CO_3^{2-} \quad [6]$$

Carbon dioxide $\leftrightarrow$ Carbonic acid $\leftrightarrow$ Bicarbonate $\leftrightarrow$ Carbonate

Many organisms catalyze the precipitation of $CaCO_3$ from seawater through active biomineralization and shell formation; for example, foraminifera, coccoliths, corals, and shellfish contribute to the permanent removal of atmospheric CO_2. Marine organisms generally form their shells at relatively shallow depths and sink to the shelf floor or abyssal depths after death. In regions where the water depth is no greater than about 1 km, the shells reach the seafloor more or less intact because these waters are saturated with respect to calcite and aragonite. However, at greater depths, where pressures are greater, temperatures are lower, and CO_2 concentrations are elevated, seawater is far more corrosive to $CaCO_3$. The level at which the rate of carbonate dissolution balances the downward flux of carbonate settling to the seafloor is called the carbonate compensation depth (CCD). This depth varies from ocean to ocean and has varied over time in response to carbonate productivity, temperature, atmospheric CO_2, and ocean circulation changes.

Geological Evolution of the Global Carbon Cycle

Many scientists suspect that the greenhouse effect and, in particular, the level of CO_2 in the atmosphere were different in the geological past. One major reason for this suspicion derives from the 'Faint Young Sun Paradox'. Because the sun is probably heating up over time – owing to the increasingly exothermic nuclear fusion of lighter elements, forming heavier elements in the sun's central core – it is probable that the input of solar energy to Earth was significantly lower in the past, by as much as 1% for every 100 million years, according to some estimates. However, Earth's sedimentary record provides evidence for the existence of liquid water at least 4 billion years ago, indicating that surface temperatures were not much lower then than they are now. Therefore, it appears that, over time, Earth's atmosphere has become increasingly less efficient in retaining solar energy. Model calculations indicate that, were CO_2 the only relevant greenhouse gas, CO_2 levels must have been as much as 10 000 times higher during the Early Precambrian. However, because methane is likely to have been a major greenhouse gas during early Earth history, before the increase of atmospheric oxygen levels, such estimates are likely to represent maximum CO_2 concentrations only.

On geological time-scales, atmospheric CO_2 levels are thought to be regulated by negative feedbacks between climate and silicate weathering rates. First, increased temperatures due to higher CO_2 levels would accelerate chemical weathering rates, thus effectively slowing any CO_2 increase. Second, weathering requires rainfall, which would most likely increase as temperatures rise, due to the acceleration of the hydrological cycle. Third, enhanced CO_2 fertilization of plant growth would also help to increase weathering rates by stabilizing soils and encouraging chemical weathering over physical weathering. This does not mean, however, that CO_2 levels have remained constant over geological time. Chemical weathering rates are imperfectly tied to CO_2 levels, being related also to additional parameters, including tectonics, vegetation cover, and palaeogeography, whereas atmospheric CO_2 may be influenced by independent changes in the input and output of CO_2. In the geological past, higher CO_2 levels could be sustained because chemical weathering rates on the continents were lower in the absence of deep soils, before the introduction of vascular plants in the Devonian period. A higher PCO_2 was also made possible by higher volcanic outgassing rates of mantle CO_2 and the abiotic nature of much carbonate deposition. Before the introduction of pelagic carbonate producers, such as planktic foraminifera by the Jurassic, carbonate deposition would have been restricted to shallow shelf environments, thus increasing calcite saturation levels and rendering deposition rates vulnerable to changing sea-level (palaeogeography). Although the introduction of land plants in the Devonian period does not appear to have had an immediate effect on climate (by sequestering more atmospheric CO_2), the deposition of massive peat deposits in shallow coastal environments during the Carboniferous and Permian periods almost certainly did. Long-term (10^7 years) increases in organic carbon burial during this interval are thought to have reduced CO_2 levels by as much as 70% (**Figure 6**), causing a series of glaciations, while allowing oxygen to build up in the atmosphere. One consequence of higher O_2 is to increase air pressure; it seems plausible that the well-recorded insect gigantism during this interval was related to the added buoyancy provided by greater air pressures. Such long-term, but ultimately unsustainable, imbalances in the proportion of carbon buried as organic carbon relative to carbonate carbon may have caused changes to both the CO_2 and the O_2 atmospheric budgets at other times, too, but the carbon isotopic record suggests that this proportion has been more or less constant at 1:4, respectively, on $>10^7$ year time-scales since 3 Ga. Geochemical models show that CO_2 levels during the Phanerozoic were also higher than at present (**Figure 6**), and reached peaks during the Early Phanerozoic and more recently during the Cretaceous period when high PCO_2 is likely to have contributed to the equable and balmy 'Greenhouse' climates typical of that interval in Earth history.

Despite the existence of regulatory feedbacks on atmospheric CO_2, the Earth system is still susceptible

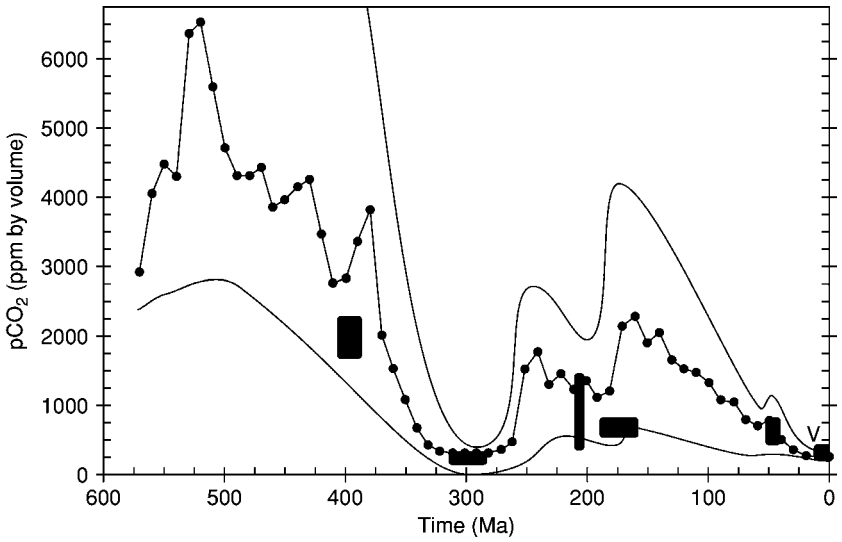


Figure 6 Comparison of model predictions of atmospheric CO_2 over the Phanerozoic, and plant stomata based data (black boxes). Reprinted from Royer DL, Berner RA, and Beerling DJ (2001) Phanerozoic atmospheric CO_2 change: evaluating geochemical and palaeobiological approaches. *Earth Science Reviews* 54: 349 392.

to shorter term ($<10^6$ years) perturbations caused by abrupt changes to the input or output fluxes of CO_2. Because volcanism is insensitive to surface negative feedbacks, short-term fluctuations in volcanic activity may have considerable, but short-lived, consequences for atmospheric composition and global climate due to the release of massive amounts of greenhouse CO_2. Such volcanic episodes, termed 'flood basalt events', are rare, but they have occurred several times over the course of the Phanerozoic; they resulted in Large Igneous Provinces, producing more than 1 million km^3 of volcanic rock, such as the Deccan and Siberian traps, in less than 1 million years (*see* **Large Igneous Provinces**). Correlations between the ages of rare igneous events and equally rare mass extinctions suggest that the resultant global warming may have caused mass extinctions, such as at the Permian–Triassic boundary (Siberian traps) (*see* **Palaeozoic:** End Permian Extinctions). Rapid episodes of global warming may also cause the sudden decomposition of volatile methane clathrate, which constitutes a massive reservoir of carbon, semi-permanently stored within sediments at high pressures and low temperatures, both on land and at sea. The abrupt release of methane (a greenhouse gas that quickly reverts to CO_2, another greenhouse gas) into the atmosphere would accelerate global warming. Precisely this may have happened at the Paleocene–Eocene boundary (Figure 7) as well as at other times in Earth history.

Glacial–Interglacial Cycles

CO_2 and CH_4 concentrations can be reconstructed back more than 200 000 years using air bubbles locked up in the Antarctic ice-cap; past concentrations clearly parallel changes in temperature during the more recent glacial–interglacial cycles (Figure 8), although they appear to lag consistently behind temperature by as much as 1000 years. This time lag suggests that whatever mechanisms are involved, a finite amount of warming is required before CO_2 and CH_4 outgassing becomes significant. Because temperature changes in Antarctica generally preceded temperature changes in the northern hemisphere, it is also possible that CO_2 changes may have been caused initially by changes in climate, but that increased CO_2 and other trace gases acted to amplify those climatic changes. The 'coral reef hypothesis' was originally proposed to explain the observed increase in atmospheric CO_2 recorded in glacial ice during times of deglaciation. This hypothesis considers that shallow shelf (or neritic) carbonate deposition is significantly lower when shelf space is decreased during glacial sea-level lowstands, whereas carbonate deposition would increase during sea-level transgressions. Because carbonate precipitation returns CO_2 back into the atmosphere, an increase in shelf space during deglaciation would serve to increase atmospheric pCO_2 and decrease the alkalinity of surface seawater, at least in the short term (Figure 9). Calculations suggest that sea-level can account for at least half of the observed increase in CO_2 levels. On longer time-scales, the effects of sea-level on carbonate deposition would be offset by changes to the deep-sea preservation of carbonate shells, but the efficacy of this feedback depends on the link between the shallow environment and the deep, which is limited by the sluggish pace of ocean

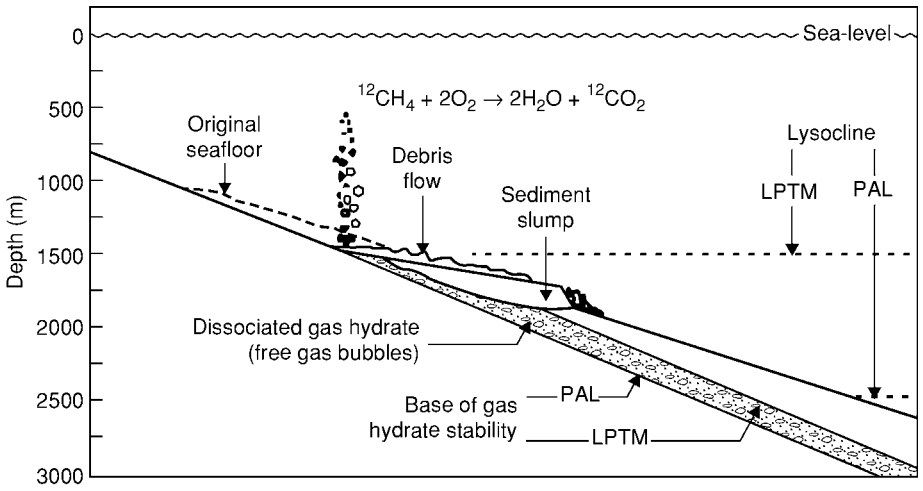


Figure 7 Schematic diagram showing the Latest Paleocene Thermal Maximum (LPTM) dissociation hypothesis. Massive quantities of ^{12}C rich CH_4 are released into the water column via sediment failure, causing short lived negative excursions in the marine carbonate $\delta^{13}C$ record and increases in atmospheric CO_2. PAL, Paleocene. Reprinted from Katz ME, Pak DK, Dickens GR, and Miller KG (1999) The source and fate of massive carbon input during the Latest Paleocene Thermal Maximum. *Science* 286: 1531 1533.

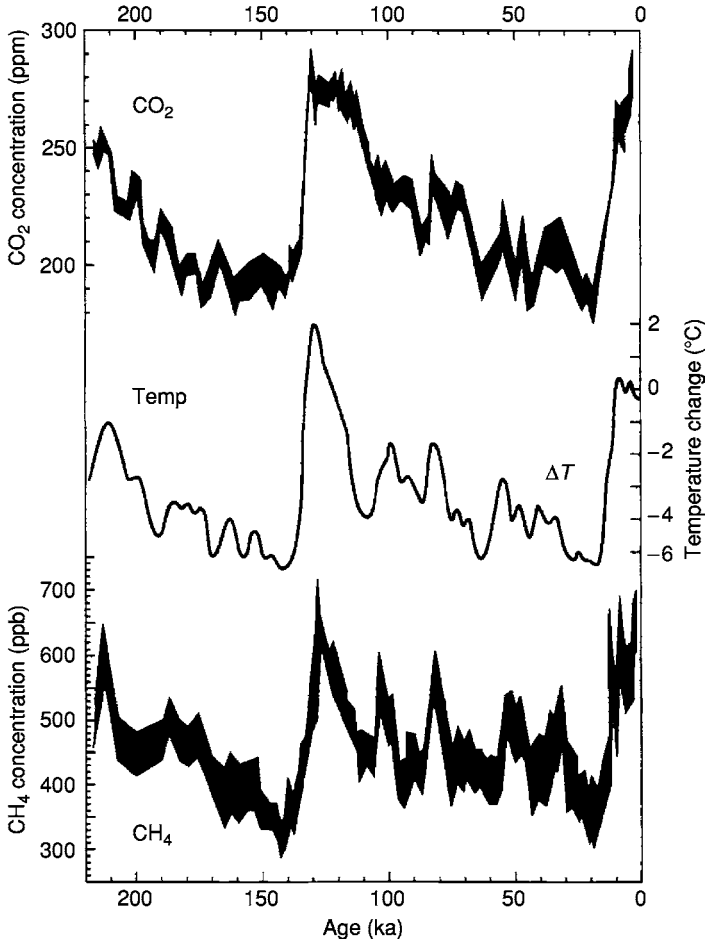


Figure 8 Atmospheric CO_2, temperature, and CH_4 records determined from the Vostok ice cores in Antarctica. Reprinted from Wigley TML and Schimel DS (eds.) (2000) *The Carbon Cycle*. Cambridge: Cambridge University Press.

circulation. Although the coral reef hypothesis seems unlikely to explain the entire CO_2 shift between glacials and interglacials, the same mechanism would have been more effective before the introduction of shelly plankton and before the establishment of a CCD.

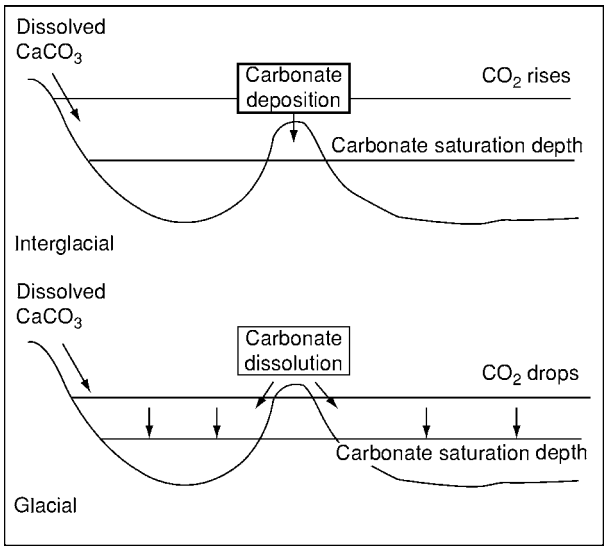


Figure 9 A qualitative comparison of $CaCO_3$ deposition and erosion to and from carbonate reef platforms from interglacial to glacial episodes. Reprinted from Milliman JD (1974) Marine carbonates. In: Milliman JD, Mueller G, and Foerstner U (eds.) *Recent Sedimentary Carbonates*, vol. 1. Berlin: Springer Verlag.

The causes of CH_4 changes are unlikely to be the same as for CO_2 changes. The largest sources of CH_4 are the methane clathrate reservoirs that form beneath cold ocean floors at depths of several hundreds of metres (Figure 7); on land, the CH_4 reservoirs are in deep lakes or deep beneath the tundra floor in Siberia, Greenland, and North America (*see* **Petroleum Geology:** Gas Hydrates). Rising CH_4 levels could represent a response to global warming in the form of decomposition of these volatile methane reservoirs, but whether on land, beneath the seafloor, or beneath retreating glaciers is unclear. Because CH_4 is unstable in the presence of oxygen and has a very short half-life in the atmosphere (about 10 years), any increase in CH_4 levels would translate eventually into a more sustained increase in the CO_2 content of the atmosphere.

Anthropogenically Induced CO_2 Increase and Future Predictions

Atmospheric CO_2 levels have been measured directly since 1957 at Mauna Loa, Hawaii; these and other data reveal that PCO_2 has risen by nearly 40% since 1800, most of half of this rise occurring during only the past half century (Figure 10). Ice-core

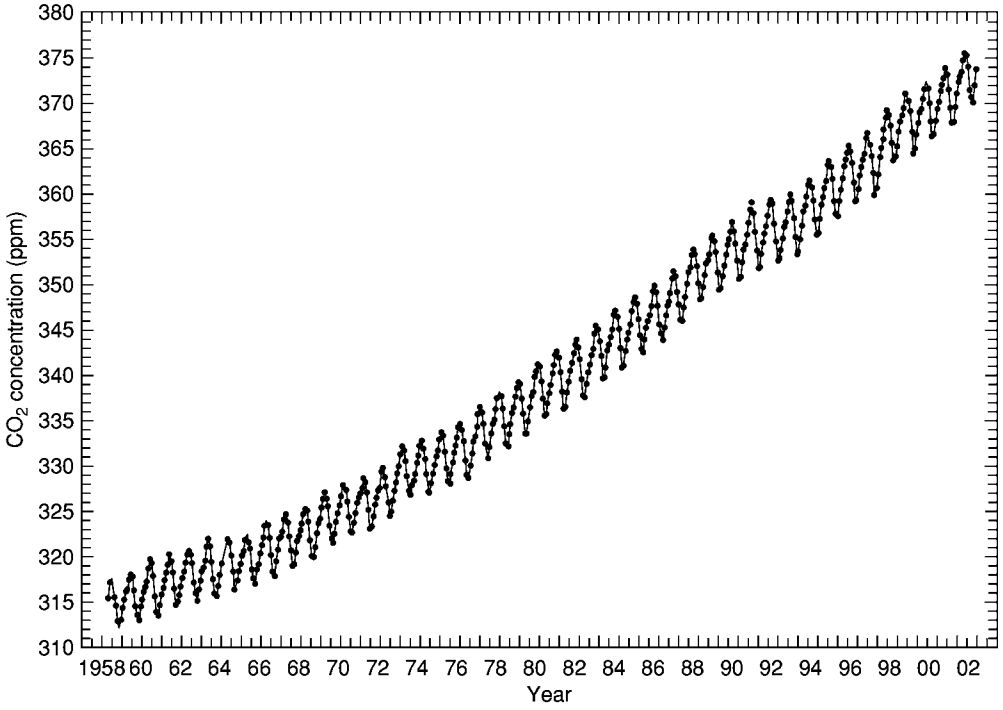


Figure 10 The 'Keeling' curve. Measurements of monthly average atmospheric CO_2 concentrations at the top of Mauna Loa, Mauna Loa Observatory, Hawaii. Source: CD Keeling and TP Wharf, Scripps Institute of Oceanography, La Jolla, California, USA (available on the Internet at http://cdiac.esd.ornl.gov/trends/co2/graphics/mlo144e.pdf).

measurements of trapped air demonstrate that PCO_2 for the 18 000 years before 1800 was much lower than it is today, and consistently averaged 280 ppm. There is a widespread agreement among scientists that the almost 100-ppm rise since the beginning of industrialization is unprecedented, being more than the increase during the last interglacial warming, which took place over thousands of years rather than over a mere 200 years. Isotopic analyses of atmospheric CO_2 confirm beyond doubt that this extra CO_2 derives in large part from the incineration of fossil fuels such as coal, oil, and gas.

The major components of the anthropogenic perturbation to the atmospheric carbon budget are anthropogenic emissions, ocean and terrestrial exchanges, and their effects on the atmospheric CO_2 increase (Figure 11). Models suggest that each year more than half of all the anthropogenic emissions from fossil fuel combustion, cement production, and tropical land clearing are offset by uptake in the oceans, by forest regrowth largely in the northern hemisphere, and by enhanced growth caused by increased levels of nutrients, mostly CO_2 and nitrate from fertilizers. Although several potential feedbacks exist to further counteract CO_2 increase through the effects of climate on ocean and terrestrial biosphere uptake of CO_2, it is not thought that these are significant in the short term. However, the possible effects of climate change on ocean circulation could be very significant indeed.

Much current research is devoted to predicting how CO_2 levels are likely to rise in the future and what impact this will have on global climate, sea-level, and biodiversity. The greenhouse effect of CO_2 is roughly logarithmic, which means that each factor-of-two increase in CO_2 produces roughly the same amount of warming. Global circulation models predict a doubling of CO_2 levels by the end of the twenty-first century (Figure 12) associated with global warming of between 1.5 and 4.5°C. Although the effects of global warming will be felt worldwide, changes in temperature will be distributed unevenly, with by far the greatest impact felt in the high latitudes of the northern hemisphere. Because the cycling of carbon in the terrestrial and ocean biospheres occurs slowly, on time-scales of decades to millennia, the effect of additional CO_2 through industrialization inevitably represents a long-lasting disturbance to the carbon cycle. Model predictions of future atmospheric CO_2 levels indicate that they will continue to

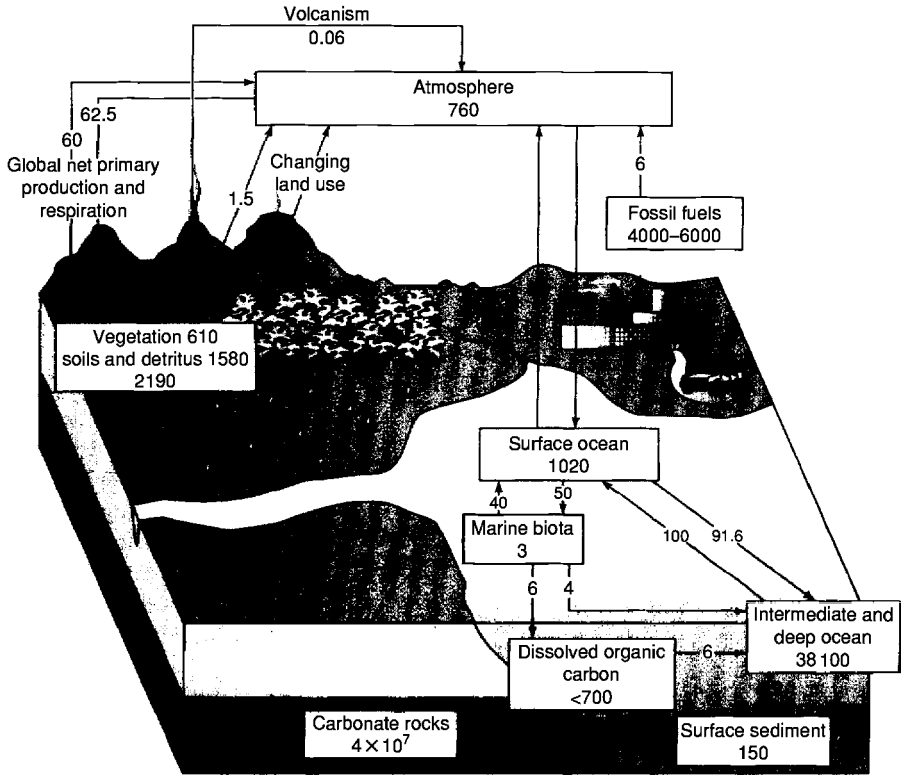


Figure 11 Major carbon reservoirs and fluxes (in gigatons and gigatons year $^{-1}$, respectively) in the global anthropogenically influenced carbon cycle. Reprinted from Kump LR, Kasting JF, and Crane RG (1999) *The Earth System*. Upper Saddle River, NJ: Prentice Hall.

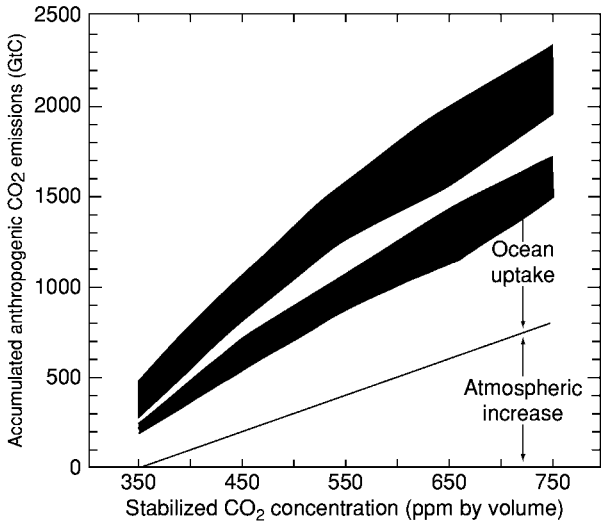


Figure 12 Accumulated anthropogenic CO_2 emissions over a period beginning in 1990 and projected out to 2200 (in gigatons of carbon), plotted against the final stabilized concentration level. Reprinted from Wigley TML and Schimel DS (eds.) (2000) *The Carbon Cycle*. Cambridge: Cambridge University Press.

rise throughout the present century, even if emissions are held constant. If the level at which CO_2 stabilizes is kept below 750 ppm, almost three times preindustrial levels, emissions will still have to be cut relative to today's emission rates.

There is continual improvement in our understanding of the modern-day carbon cycle, and increasingly sophisticated models are being developed to describe the Earth system (*see* **Earth System Science**). However, it is already clear that we have begun an unprecedented experiment, the course of which we cannot control over the short-term, but which will shortly become clear.

See Also

Atmosphere Evolution. **Earth System Science**. **Gaia**. **Palaeoclimates**. **Petroleum Geology:** Gas Hydrates; The Petroleum System. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Processes:** Fluxes and Budgets. **Sedimentary Rocks:** Limestones. **Solar System:** Venus. **Weathering**.

Further Reading

Berger WH (1982) Increase of carbon dioxide in the atmosphere during deglaciation: the coral reef hypothesis. *Naturwissenschaften* 69: 87 88.

Berner RA (1992) Weathering, plants and the long term carbon cycle. *Geochimica et Cosmochimica Acta* 56: 3225 3231.

Gaillardet J, Dupré B, Louvat P, and Allègre CJ (1999) Global silicate weathering and CO_2 consumption rates deduced from the chemistry of the large rivers. *Chemical Geology* 159: 3 30.

Katz ME, Pak DK, Dickens GR, and Miller KG (1999) The source and fate of massive carbon input during the Latest Paleocene Thermal Maximum. *Science* 286: 1531 1533.

Kump LR, Kasting JF, and Crane RG (1999) *The Earth System*. Upper Saddle River, NJ: Prentice Hall.

Pavlov AA, Kasting JF, Brown LL, Rages KA, and Freedman R (2000) Greenhouse warming by CH_4 in the atmosphere of early Earth. *Journal of Geophysical Research* 105: 11981 11990.

Royer DL, Berner RA, and Beerling DJ (2001) Phanerozoic atmospheric CO_2 change: evaluating geochemical and paleobiological approaches. *Earth Science Reviews* 54: 349 392.

Sunquist ET and Broecker WS (1985) *The Carbon Cycle and Atmospheric CO_2. American Geophysical Monograph 32*. Washington, DC: American Geophysical Union.

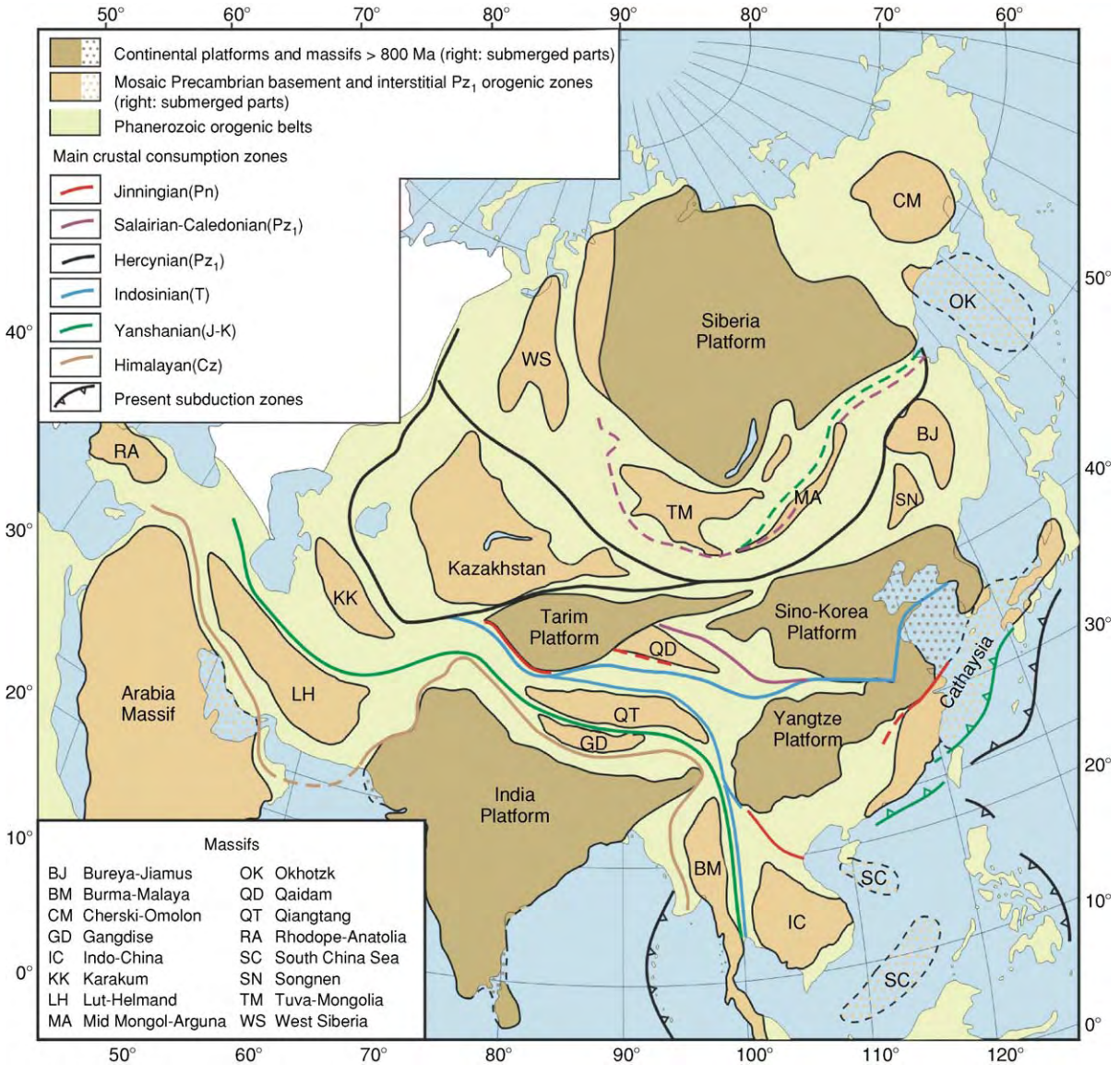
Wigley TML and Schimel DS (eds.) (2000) *The Carbon Cycle*. Cambridge: Cambridge University Press.

CHINA AND MONGOLIA

H Wang and Shihong Zhang, China University of Geosciences, Beijing, China
Guoqi He, Peking University, Beijing, China

Introduction

China and Mongolia are among the most composite in geological structures in the Asian countries. The tectonic units of different ranks comprise tectonic domains, platforms, and massifs with Precambrian basement, and Phanerozoic orogenic belts and zones. A tectonic domain usually encloses a platform (craton) and its surrounding orogenic belts, which are composed of orogenic zones of various ages and interstitial Precambrian median massifs. In the outline tectonic map of Asia (**Figure 1**), are shown the main cratons, the major massifs, the orogenic belts, and the major sutures of various ages. The history of crustal evolution may be subdivided into tectonic megastages based on fundamental



Color Image

Figure 1 Simplified tectonic map of Asia.

changes of tectonic frame of continental and even global scale, and tectonic stages marked by pronounced changes of widespread tectonic regimes, during which revolutionary orogenic epochs of short duration and rapid changes may be recognised (Figure 2). The global stratigraphic chart used follows the International Stratigraphic Chart (2000) with some changes in the Precambrian part. The tectonic development of China and Mongolia are discussed in terms of the megastages and the main geologic events that occurred in the various orogenic epochs. Emphasis are put on the Neoproterozoic Jinningian and the Upper Triassic Indosinian orogenies in China and on the Palaeozoic Salairian and Hercynian in Mongolia. The terms 'Caledonian' and 'Hercynian' for orogenies in China have different usages from their original locations in Europe. A world palaeocontinental reconstruction map for the Pangaea in mid-Permian time (Figure 4) is presented to show the possible position of the component parts of China and Mongolia and the world floral provinces at that time.

The Geology of China

The Main Tectonic Units and Crustal Evolution of China

The main tectonic units of China comprise three principal continental platforms (cratons), Sino-Korea, Yangtze, and Tarim, a deformed palaeocontinent (Cathaysia), and orogenic belts of various ages. The orogenic belts are situated between the platforms and

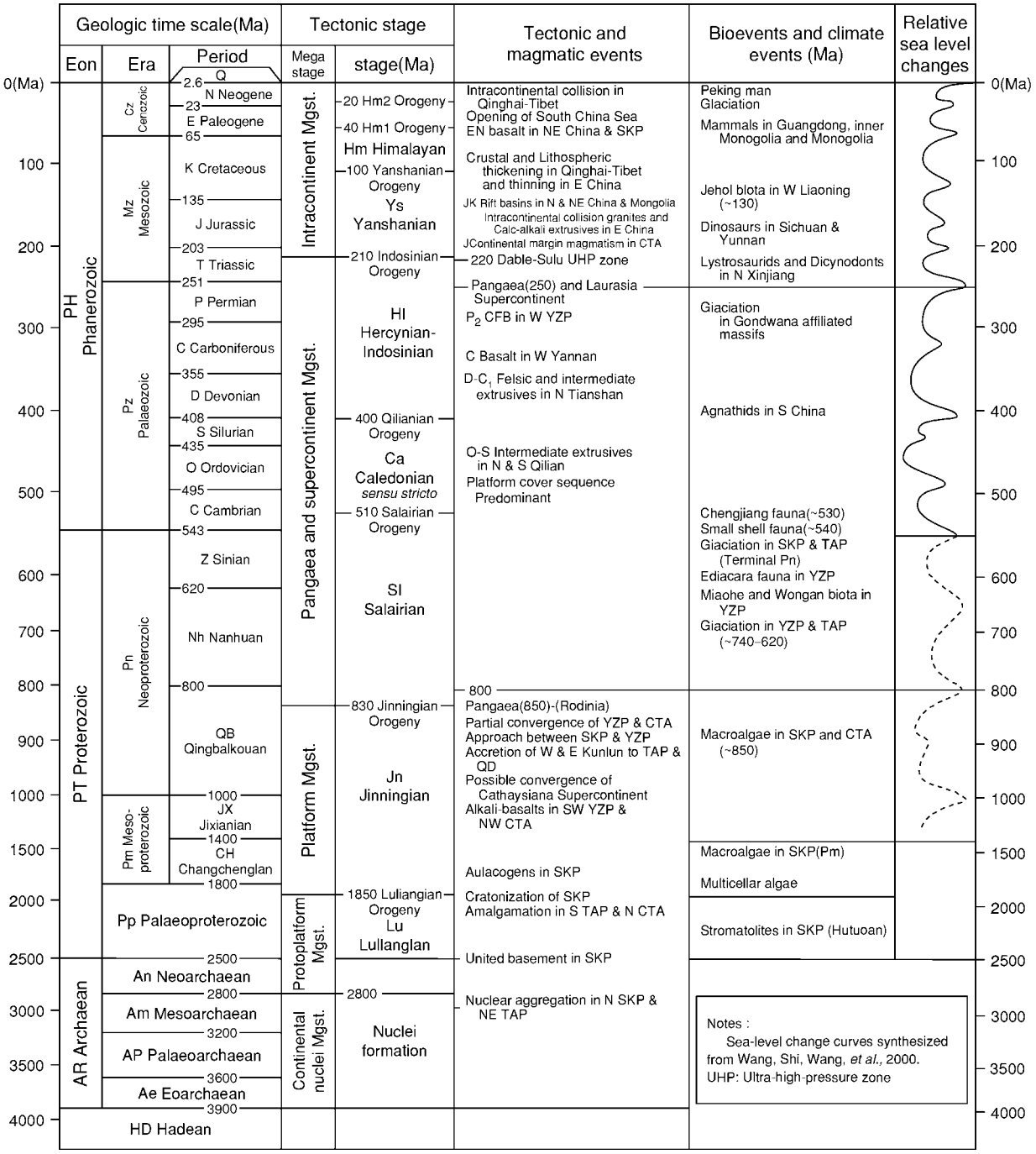


Figure 2 Tectonic megastages and geoevents in China and Mongolia.

major massifs, and contain small massifs that were probably split from neighbouring palaeocontinents. The plate boundary does not lie at the platform border but in a demarcation zone between the once distantly opposite marginal tracts of the palaeocontinents. We have called the major crustal sutures, or the demarcation zones that represent the consumed open seas, convergent crustal consumption zones (CZ), and the sutures that represent accreted island arcs onto the palaeocontinents, accretional crustal consumption zones (AZ).

Geologically, three main regions may be recognised in China. The northern region comprises the narrow Altai-Arguna Belt representing the southernmost part of the wide southern continent marginal tract of Siberia Platform, and the Tianshan-Xingan Belt representing the northern continent marginal tract of Tarim (TAP) and Sino-Korea (SKP) platforms.

The middle region consists of TAP and SKP, the Qilian Caledonides between them, and the Kunlun-Qinling Belt, the central orogenic belt of China, which is mainly composed of the Indosinides. The southern boundary of this region is the Indosinian Muztagh-Maqen convergent zone (MMCZ) in the west and the superimposed Indosinian Fengxian-Shucheng convergent zone (FSCZ) in the east. The southern region includes two parts: the eastern part, South China, consists of the Yangtze Platform (YZP), the Cathaysia palaeocontinent (CTA), and the Caledonides and Indosinides between them, while the western part covers the Qinhai-Tibet Plateau and the broad Indosinides to the south of MMCZ.

The history of crustal evolution of China may be subdivided into five megastages: (i) continental nucleus formation (*ca*. 2.7–2.8 Ga); (ii) protoplatform formation (*ca*. 1.9–1.8 Ga); (iii) platform formation (*ca*. 0.85–0.8 Ga); (iv) Laurasia Supercontinent or Pangaea formation (*ca*. 230–210 Ma); and (v) intracontinental development (210 Ma-Present) (Figure 2). Two orogenic epochs of revolutionary changes occurred in the Jinningian (*ca*. 1000–830 Ma) and the Indosinian (*ca*. 230–210 Ma). The close of the Jinningian Orogeny probably witnessed a convergence of the Chinese platforms and massifs to form the supercontinent of Cathaysiana, which was a part of the Neoproterozoic Rodinia (850 Ma). After the close of the Indosinian Orogeny, the tectonic frame of China underwent a basic change from a contrast between the north and the south to that between the east and the west. Thus, the crustal evolution of China underwent three stages; (i) the Jinningian and pre-Jinningian (Archaean to Early Neoproterozoic >800 Ma); (ii) the post-Jinningian to the Indosinian (Nanhuan to Triassic); and (iii) the post-Indosinian (Jurassic to Quaternary) (Figure 2).

China in the Pre-Jinningian and Jinningian (Archaean to Qingbaikouan)

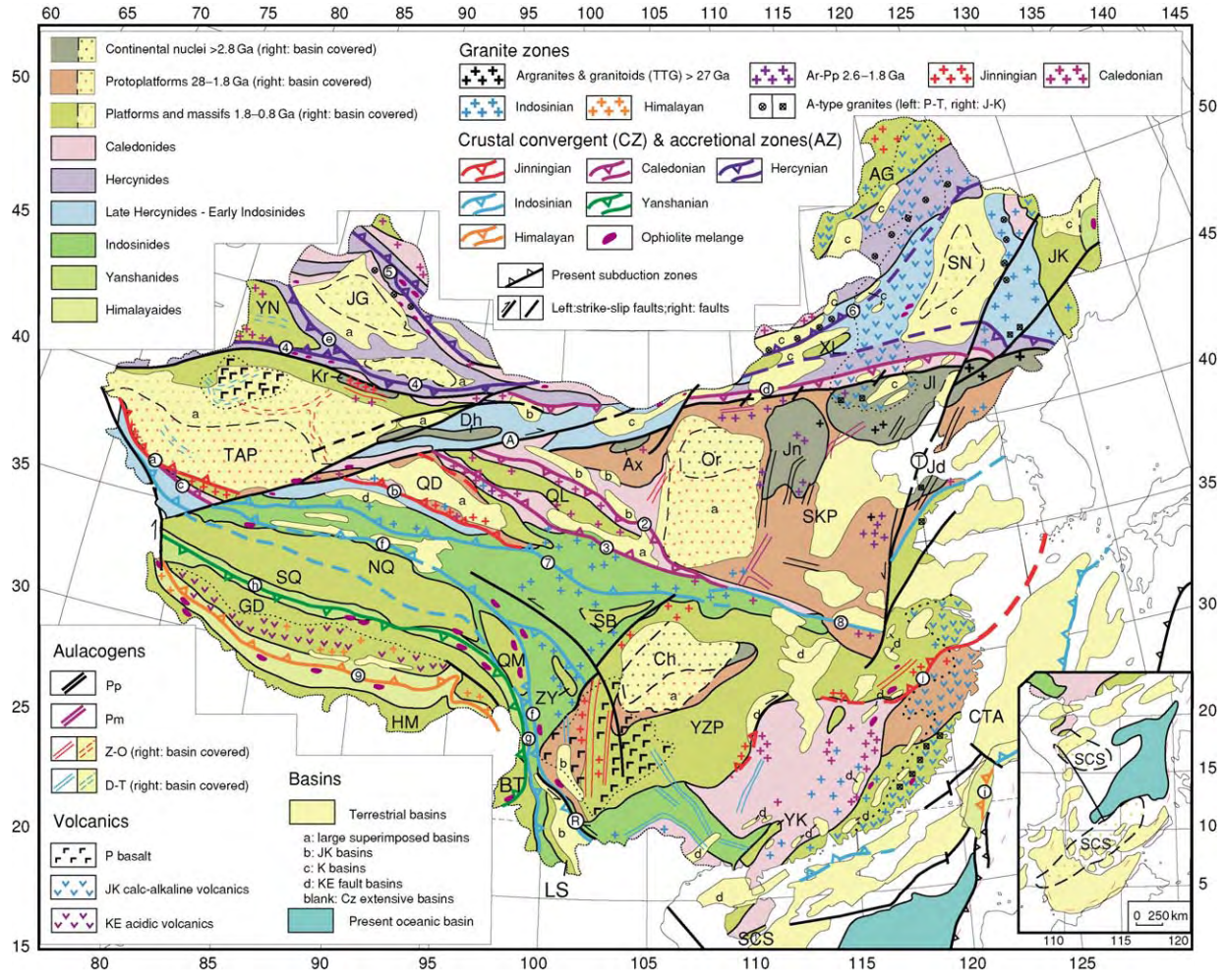
The time span from Archaean to Early Neoproterozoic (*ca*. 800 Ma) may be subdivided into three megastages which resulted respectively in the formation of continental nuclei at the end of the Mesoarchaean (*ca*. 2.8 Ga), the formation of protoplatforms at the end of the Palaeoproterozoic (*ca*. 1.8 Ga) through the Luliangian Orogeny, and the formation of platforms at the end of the Early Neoproterozoic (*ca*. 800 Ma) through the Jinningian Orogeny.

The continental nuclei Several continental nuclei may be recognized in SKP, by the presence of Mesoarchaean and older metamorphic supracrustal and TTG rocks. The Jiliao Nucleus (Jl) encloses northern Shanxi, northern and eastern Hebei, northern Liaoning and Jilin provinces (Figure 3). U-Pb and Pb-Pb ages of 3850–3550 Ma, representing the Eoarchaean primordial sialic crust and supercrustals, were identified from Caozhuang, eastern Hebei, and Anshan, Liaoning. Mesoarchaean rocks older than 2.8 Ga are known from the Jiaodong Nucleus (Jd) in eastern Shandong, the Sanggan gneiss, the Jining Nucleus (Jn) in the southern Inner Mongolia, and northern Shanxi. The Ordos Nucleus (Or) is inferred to exist under the northern part of the Ordos basin based on geophysical as well as geological data. Further in the west is the small Alxa Nucleus (Ax) near the western border of SKP.

In the Yangtze Platform (YZP), Mesoarchaean TTG rocks older than 2.8 Ga are known from the Kongling Group in the Yangtze Gorges region, probably representing the uplifted eastern margin of the Chuanzhong Nucleus (Cz) beneath the Sichuan basin. In the eastern part of the Tarim Platform (TAP), Mesoarchaean rocks are reported from the Quruktagh Nucleus (Qr) and from the Dunhuang Nucleus (Dh), where a single grain zircon U-Pb age of 3.6 Ga of Eoarchaean age was recently obtained. Generally, these ancient continental nuclei are distributed in the northern part of SKP and the north-eastern part of TAP. It is noteworthy that isotopic model age studies for various Archaean rocks in the SKP show a cluster around 2.8–2.7 Ga, which denotes an epoch of rapid continental growth, as in many other cratons in the world.

The Protoplatforms and the Luliangian Orogeny The Neoarchaean is widely distributed in the platform regions of China, especially in SKP. The juvenile crust generated in the Neoarchaean may be mostly assigned to the granite-greenstone terrains with a high proportion of granitoid intrusives, as in the Taishan Complex in Shandong. The Neoarchaean usually occurs as granite and TTG belts around and within the reworked nuclear regions in SKP. On the other hand, Neoarchaean greenstone belts, represented by the Wutai, Dengfeng and Taihua groups in Shanxi and Henan, contain more khondalites and mafic volcanics, and are probably of extensional origin. In general, the Neoarchaean witnessed a profuse intrusive and extrusive magmatism of 2.6–2.4 Ga age throughout SKP, which is responsible for the formation of united Archaean basement.

The Palaeoproterozoic (Pp) in SKP, generally called the Hutuoan 'System' (2500–1800 Ma), often occurs in long narrow belts and represents ancient aulacogens formed on the Archaean basement. The aulacogen sequence is mainly composed of a lower fluvial, a middle immature volcanisedimentary, and an upper molasse deposits. This sequence occurs in the Hutuo



Color Image

Figure 3 Outline map showing the tectonic units and crustal evolution of China. Hercynian and Yanshanian granites are wide spread and planar in distribution, and are not shown on the map.

Tectonic units: AG, Arguna Massif; Ax, Alxa Nucleus; BT, Baoshan Tengchong Massif; Ch, Chuanzhong Nucleus; CTA, Cathaysia palaeocontinent; Dh, Dunhuang Nucleus; GD, Gangdise Massif; HM, Himalaya Massif; Jd, Jiaodong Nucleus; JG, Junggar Massif; JK, Jamus Xingkai Massif; JI, Jiliao Nucleus; Jn, Jining Nucleus; Kr, Kuruktagh Nucleus; LS, Lanping Simao Massif; NQ, North Qiangtang Massif; Or, Ordos Nucleus; QL, Qilian Massif; QM, Qamdo Massif; SB, Songpan Bikou Massif; SCS, South China Sea Massif; SKP, Sino Korea Platform; SN, Songnen Massif; SQ, South Qiangtang Massif; TAP, Tarim Platform; XL, Xilinhot Massif; YN, Yining Massif; YZP, Yangtze Platform; ZY, Zhongza Yidun Massif.

Crustal consumption convergent zones (CZ): ① JSCZ, Jiangshan Shaoxing (Jinningian); ② QQCZ, Qilian Qinling (Caledonian); ③ SQCZ, South Qilian (Caledonian); ④ STCZ, South Tianshan (Hercynian); ⑤ EACZ, Ertix Almantai (Hercynian); ⑥ HGCZ, Hegen shan (Hercynian); ⑦ MMCZ, Muztagh Maqen (Indosinian); ⑧ FSCZ, Fengxian Shuchang (Indosinian); ⑨ YZCZ, Yarlung Zangbo (Himalayan).

Accretional zones (AZ): ⓐ WKAZ, West Kunlun (Jinningian); ⓑ EKAZ, East Kunlun (Jinningian); ⓒ KDAZ, Kudi (Indosinian); ⓓ OSAZ, Ondur Sum (Caledonian); ⓔ NTAZ, North Tianshan (Hercynian); ⓕ JSAZ, Jinshajiang (Indosinian); ⓖ CMAZ, Changning Menglian (Indosinian); ⓗ BNAZ, Bangong Nujiang (Yanshanian); ⓘ LJCZ, Liji (Himalayan).

Strike slip faults: Ⓐ Altun, Ⓣ Tanlu, Ⓡ Red River.

and Gantaohe aulacogens in the Wutai-Taihang region, in the Qinglong Aulacogen in north-eastern Hebei, in the Liaohe and Fenzishan aulacogens in eastern Shandong and southern Liaoning (Figure 3). Through the Luliangian Orogeny of 1.9–1.7 Ga age, these aulacogen rocks were intensely folded and regionally metamorphosed, sometimes to a high chlorite schist facies.

The Luliangian Orogeny brought about the formation of a protoplatform in North China, upon which was deposited the Meso- and Neoproterozoic paracover. This orogenic event is also seen in the TAP and Qaidam Massif (QD), where the Mesoproterozoic and Neoproterozoic sequences are always separated from the basement by a fragmented unconformity. In YZP, the Palaeoproterozoic basement is probably

present to the south of the Chuanzhong Nucleus (Ch), and the Kangdian belt in western YZP is mainly composed of Palaeoproterozoic metamorphic rocks. In CTA, Neoarchaean and Palaeoproterozoic rocks, with Luliangian metamorphism of amphibolite facies, occur in western Zhejiang and in the Wuyi Mountains in north-western Fujian.

The platforms and the Jinningian Orogeny In SKP, the Mesoproterozoic and Early Neoproterozoic strata are widespread and are divisible into the Changchengian (Ch), the Jixianian (Jx), and the Qinbaikouan (Qb) 'systems'. They are generally correlatable through acritarch and stromatolite assemblages between the main platforms of China. In the Yanshan-Taihang region, aulacogen (Figure 3) deposits consisting of the Lower Changchengian fluvial clastics, carbonates, and high potassium volcanics, are followed by the widely transgressive Gaoyuzhuang carbonates bearing the macro-alga *Grepania* in the upper part (1.5–1.4 Ga). The Jixianian is also widespread and contains stromatolite carbonates of great thickness. The Qingbaikouan represents a platform cover and is limited in distribution. The macroalgal assemblages of the Jixianian *Grepania* and the Qingbaikouan *Tawuia-Longfengshania* Assemblages (900–800 Ma) are almost identical to those found in the Greyson Shales in Montana and in the Little Dal Formation in the McKenzie Mountains in the Belt Supergroup of western North America. The *Tawuia* beds are also found in Hainan Island, which was probably a part of the Cathaysia palaeocontinent. These indicate that Laurentia and the Chinese palaeocontinents may have been near each other during Meso- and Neoproterozoic times.

Most platforms and massifs in China were dominated by an extensional tectonic regime in the Meso- and Neoproterozoic, as shown by the Xionger Aulacogen, with bimodal eruptives in Henan and southeastern Shanxi, and the Bayan Obo Aulacogen near the northern margin of SKP. Aulacogens of similar age are also developed in the Quruktagh region of northern TAP and in the Kunming region of southwestern YZP. All these aulacogens ended without any diastrophism. On the southern margin of SKP, there may have developed an island-arc system with the Qinling Group as the arc and the Kuanping Group as the back-arc basin. Both groups yield Mesoprotrozoic isotopic ages.

The Jinningian Orogeny of *ca*. 1000–830 Ma age has left clear records in many parts of China. The Qinling region, mainly South Qinling, was characterized by a complicated rift system probably formed in Late Mesoproterozoic to Early Neoproterozoic, which consisted of discrete massifs (Douling, Fuping) and trough deposits with bimodal volcanicism (e.g., the Xixiang group). The Jinningian Orogeny is represented by island arcs and marginal seas in the southern margin of North Qinling by the Songshugou ophiolite (983 Ma) and by an Early Neoproterozoic granite zone (e.g., the Dehe granite of *ca*. 950 Ma), denoting a northward subduction and accretion through an arc-continent collision. A similar southward accretion, with subduction-type granites is also found in the Hanzhong area of YZP. The common cover of the Sinian over the rifted elements implies the presence of a united South Qinling Belt. Therefore, the Jinningian Orogeny brought about the consolidation of the rifted region and the mutual approach of SKP and YZP.

Along the south-western margin of YZP in central Yunnan, to the east of the Early Mesoproterozoic Dahongshan metamorphic volcanisedimentary Belt, was developed a broad aulacogen trending north–south for hundreds of kilometres, composed mainly of the Mesoproterozoic Kunyang Group and equivalent strata of huge thickness. These rocks were intensely folded and intruded by the Jinningian granites (850–750 Ma), and may be interpreted as a wide back-arc basin lying to the east of the Dahongshan arc belt in central Yunnan.

In the Jiangnan uplift along the south-eastern margin of YZP, deformed arc-type turbidites with ophiolite mélange and Adakite zones dated at *ca*. 970 Ma represent an arc-continent collision of Early Jinningian age. In fact, the Jinningian orogenic zone was charaterized by volcanisedimentaries and granites extending from northern Zhejiang right to the border area between Hunan and Guangxi. At the close of the Jinningian Orogeny, a partial collision between YZP and CTA occurred along the Jiangshan-Shaoxing convergent zone (JSCZ), which closed at around 830 Ma. The ocean basin between YZP and CTA was probably separated by an archipelagic belt composed of small massifs near the Hunan-Jiangxi border. The north-western margin of CTA in the Neoproterozoic was also active and extends from Wuyi in North-western Fujian to Hainan Island.

Jinningian granites are distributed along the southern margin of TAP and the Qaidam Massif in the northern Kunlun Mountains and are also found in the northern margin of Qaidam. These granite zones represent active continental margins or continent-arc collision zones (Figure 3). The Jinningian granites are seldom found to the east of East Qinling, except in eastern Shandong along the Sulu Belt, although a Jinningian metamorphic event is suspected to be represented by the Dabie UHP belt. In summary, the main platforms and massifs in China seemed to have converged to form a loosely united Cathaysiana

Supercontinent at around 850 Ma, concurrently with the formation of the Neoproterozoic Rodinia.

China in the Post-Jinningian to the Indosinian (Nanhuan to Triassic)

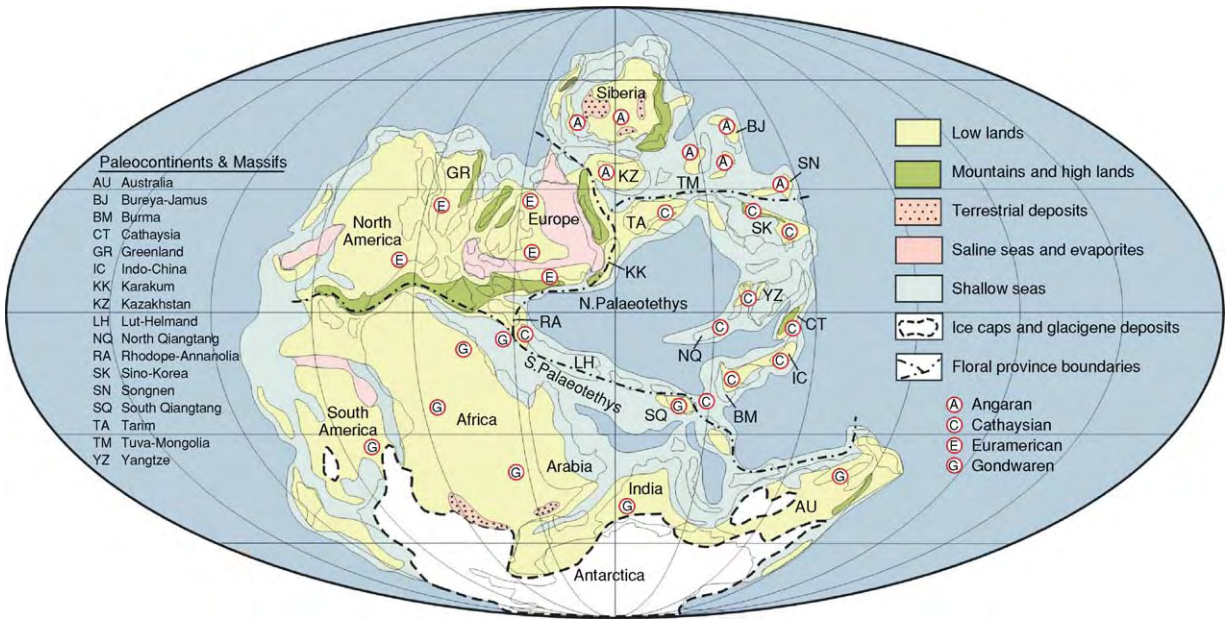
The time span from the Nanhuan to the Triassic (800–208 Ma) is traditionally subdivided into Caledonian, Hercynian, and Indosinian stages. The traditional Sinian (800–543 Ma) is subdivided into the Nanhuan (800–620 Ma) and the revised Sinian (620–543 Ma) according to the new Regional Chronostratigraphical Chart of China published by the Third All-China Stratigraphical Commission in 1999. The Hercynian and Indosinian are here integrated and called the Hercynian-Indosinian Stage (Figure 2).

The Caledonian stage The Caledonian in China is usually subdivided into a lower division of Xingkaian (Salairian) covering the Nanhuan and the Sinian, and an upper division covering the Middle Cambrian to the Silurian. The Caledonides are mainly distributed in the Qilian Mountains region between SKP and the Qaidam Massif and in South China between YZP and CTA. They are also found to the north of SKP and TAP (Figure 3).

The Nanhuan in YZP is characterized by clastic, glaciogenic, and cold-water deposits roughly corresponding to the Cryogenian in the International Stratigraphic Chart. Continental ice-sheets of *ca.* 740–620 Ma age seem to have been confined to western YZP and the Quruktagh region of TAP, but mountain and maritime glaciations are more widespread. In the type region of the Yangtze Gorges in the YZP, the Sinian includes the Doushantuo cap carbonates, phosphate and black shales bearing the Miaohe or Wengan biota (Pb-Pb isotopic age *ca.* 600 Ma), and the Dengying Formation bearing the Ediacara biota (*ca.* 590 Ma). In SKP, only the Sinian is developed in the peripheral parts and bears a rich metazoan fauna comparable with the Ediacara in northern Anhui. A higher Luoquan glacial horizon was reported all along the south-western border of SKP, the northern border of Qaidam Massif, the northern TAP and the Tianshan regions: it is probably younger than 600 Ma in age.

In most parts of eastern China, the Lower Palaeozoic began with an Early Cambrian transgression from southern YZP to SKP, with well-established trilobite zones for correlation. This indicates that the two platforms were not far from each other, although rifting seems to have begun in the Sinian, as is shown by entirely different Nanhua-Sinian sequences between the two platforms. The Early Cambrian Chengjiang Lagerstatte (*ca.* 525 Ma), of great significance in life evolution, contains arthropods and chordates, especially *Haikouichthys*, which might be the 'first fish' on Earth. No obvious break is known between the Sinian and the Cambrian, especially in South China, where bathyal carbonaceous silicolites are continuous in central Hunan and western Zhejiang. The continuous Lower Palaeozoic passive margin bathyal deposits have provided seldom seen complete sequences of Cambrian agnostid trilobite zones and Ordovician to Silurian graptolite zones, which are ideal for the designation of chronostratigraphical boundaries. In the residual sea between YZP and CTA, rifted uplifts and troughs were developed in which varied kinds of sediments were laid down. The marine realm began to shrink in the Late Ordovician, and the Caledonian Front, which started in the Wuyi Mountains, seems to have shifted westward, until the thick foreland basin deposits of Early Silurian age were formed within the YZP in westernmost Hunan. Caledonian granites and metamorphism are scattered, and no clear collision zone is found in the broad Caledonides in South China, which were 'filled up', rather than folded orogenic zones. In southern Hainan Island, Cambrian trilobites of Australian affinity are found, which may represent the northern margin of SCS (Figure 2), probably still far away to the south at that time.

The Caledonides between SKP and Qaidam on both sides of the Central Qilian Massif (QL) are represented by distinct collision belts and associated granite zones. In North Qilian a complete Caledonian orogenic sequence developed composed of Cambrian and Ordovician island arc volcanics, Middle and Upper Ordovician and Lower Silurian ophiolite suites, and Devonian molasse deposits. The QQCZ (Figure 4) is prominent and continues southeastwards into North Qinling. Late Cambrian ophiolite mélanges are also found in Lajishan in South Qilian, which may connected to the SQCZ. The SKP generally lacks the Upper Ordovician and Silurian except in the western margins, and is bordered to the north by a narrow Caledonide strip (south of OSAZ), composed of Cambro-Ordovician metavolcanics unconformably overlain by fossiliferous Silurian sediments. Caledonides are also known around the Junggar and Yining massifs in northern Xinjiang. In TAP, the Cambrian and Ordovician are of platform cover type, but display a slope and deep-sea facies in the Manjiar depression in the eastern part. The Silurian is incomplete and limited in distribution. Stable Lower Palaeozoic deposits reported from Himalaya, northern Gangdise, and north-western Qiangtang indicate the existence of Precambrian basement in these massifs. However, the age of the basement may be as young as *ca.* 600 Ma, comparable with the Pan-African of Gondwanaland.



Color Image

Figure 4 A mid Permian world palaeocontinental reconstruction.

Laurasia Supercontinent and the Hercynian-Indosinian stage In the Late Palaeozoic, the major part of SKP remained exposed after its upheaval in mid-Ordovician, until the Middle Carboniferous epicontinental seas began to inundate the platform. Marine facies prevailed up to the mid-Permian and are followed by Permo-Triassic paralic and terrestrial deposits. A Devonian oceanic basin existed along the China-Mongolia border, which was consumed and formed the main Hercynian HGCZ. The orogenic collision is manifested by Devonian to Early Carboniferous ophiolites and Permian A-type granites which mark the boundary between the Siberian continent marginal tract in the north and the Sino-Korean marginal tract in the south. In northern Xinjiang, the Hercynian EACZ is composed of *en echelon* sutures and represents the main boundary between the Altai-Hingan and the Tianshan-Inner Mongolia belts, and is continuous across southern Mongolia with the HGCZ. Two Hercynian collision zones documented by ophiolites occur to the south of the Junggar-Turpan and Yining massifs.

The Late Hercynian-Early Indosinian belts south of Beishan in the west, and extending via Linxi to Changchun in the east, are composed mostly of shallow marine and paralic sediments with occasional volcanics. The eastern segment is a clear boundary between the Angaran and the Cathaysian floras, but no ophiolite mélanges have been found. These marine troughs were probably filled at the end of the Palaeozoic, and are not genuine orogenic zones. Late Permian to Early Triassic terrestrial basins, formed after the Hercynian Orogeny, are widespread in northern Xinjiang, Gansu, and Inner Mongolia, and bear the tetrapod remains of *Dicynodon*, *Lystrosaurus*, and *Sinokanemeyeria*.

In the Late Palaeozoic, YZP was characterized in its northern marginal tract, the South Qinling, by thick Devonian continental slope flysch and marine to paralic Carboniferous to Triassic clastics and carbonates. To the south, both YZP and CTA witnessed Devonian transgressions from the residual Youjiang marine basin, and Late Carboniferous and Permian carbonate platforms were well developed in both regions. Extensional conditions prevailed in YZP, as is shown by the Late Permian Emei continental flood basalt. Devonian to Triassic aulacogens with different trends are known in south-western YZP, and isolated carbonate platforms were formed in the Youjiang oceanic basin, where Upper Palaeozoic deep-sea deposits with pelagic radiolarians are formed. The Youjiang marine basin was not closed until the Late Triassic, when the Indosinian Orogeny was predominant in most parts of southern China. The eastern part of YZP, now partly submerged under the sea, collided with SKP along the Dabie-Sulu HP and UHP zone through the Indosinian Orogeny, and was probably wedged in eastern SKP, with its northeastern promontary along the Imqingang Belt in Korea. Orogenic and post-orogenic Indosinian granite zones are widely distributed in southern China and in the Kunlun-Qinling Belt. Post-orogenic muscovite/two mica granites are known to the north of SKP as a result of intracontinental northward compression of SKP toward Inner Mongolia (Figure 3).

In western China, the main Muztagh-Maqen convergent zone (MMCZ) in the central orogenic belt closed in Late Hercynian to Indosinian times. The Indosinian Jinshajiang accretion zone, extending from West Kunlun southward to the Changning-Menglian zone (CMAZ) in western Yunnan, marks the boundary between the Yangtze-affiliated massifs in the east and the Gondwana-affiliated massifs in the west. In northern Tibet, an Indosinian suture is suspected to exist between the North and South Qiangtang massifs, mainly based on the occurrence of the Late Triassic Qiangtang ophiolite complex, which marks the southern margin of the Late Triassic flysch complex underthrust beneath South Qiangtang, and on the boundary between the Cathaysian and the Gondwanan floras (Figure 3). The wide Indosinides and their southern marginal massifs, North Qiangtang and Qamdo, may therefore represent the southern boundary of the newly amalgamated Laurasia Supercontinent which formed the northern half of Pangaea (Figure 4). The boundaries of the floral provinces are arbitrary, since mixed flora of different provinces are known; for example, the mixed Cathaysian and Angaran flora in northern Tarim. The most important world mass extinction at the end of the Permian (*ca*. 250 Ma), especially that of the marine organisms, is well represented and studied in China.

China in Post-Indosinian Times

The Indosinian Orogeny had caused a radical change of tectonic pattern of China from a north–south to an east–west demarcation, and the post-Indosinian marks a megastage of mainly intracontinental development. Jurassic and later seas retreated from China except in the Qinghai-Tibet region, and only sporadic marine ingressions occurred in eastern Heilongjiang, and in the border parts of TAP and CTA. The new tectonic framework and dynamics of China were chiefly controlled by interactions between the Siberia Plate in the north, the Pacific Plate in the east, and the India Plate in the south-west. The Indosinian Orogeny brought about the closure of the Late Palaeozoic Palaeotethys and the formation of the extensive Indosinides, including Kunlun-Qinling, Garze-Hon Xil, and down to Indochina and Malaysia, which formed the southern margin of the Eurasian palaeocontinent. Thus, an extensional system prevailed in East China and a successive northward accretion of the Gondwana-affiliated massifs onto Eurasia in the Tethys domain has dominated West China since the Late Mesozoic.

Post-Indosinian tectono-magmatism and basin development in eastern China In eastern China, the Indosinian Orogeny was followed by intracontinental collision and further welding of platforms and massifs, as is shown by the widespread Jurassic A-type granites and by the southerly-imbricated thrust zones in the Yanshan region and the westerly thrust zones in north-eastern Anhui and western Hunan. The large Tanlu fault, with a lateral shift of hundreds of kilometres in length, was probably mainly formed in the pre-Cretaceous. The newly-formed Circum-Pacific domain comprised a western belt of continental and maritime East China and an eastern belt of island arc-basin systems in the inner western Pacific. The Yanshanian Orogeny in eastern China is characterized by inner continent marginal type magmatism in the coastal belt (Figure 3), which originated by subduction of the Izanaqi Plate under East Asia. Consequently, inland eastern China was characterized by a combination of subduction and intra-continental collision types of magmatism, manifested by muscovite/two mica granites and high potassium calc-alkalic and shonshonitic volcanism. From a comparison between the crust and lithosphere thickness data of pre-Jurassic and Jurassic-Cretaceous in North China, based on petrogenic studies, it is found that a crustal thickening of *ca*. 15 km (40 against 50–60 km) and a lithospheric thinning of *ca*. 120 km (200–250 against 50–60 km) occurred in the Late Mesozoic. This proves that eastern China changed from a compressional to a tensional regime.

Mesozoic and Cenozoic basins in eastern China are distributed in three belts. The Ordos Basin and the Sichuan Basin in the western belt became terrestrial inland basins in the Late Permian and Late Triassic, respectively. Both were influenced by the Indosinian Orogeny that formed Late Triassic foreland basin deposits, in the southern border of Ordos north of the Qinling Orogen, and in the Longmenshan thrust belt of Sichuan Basin, respectively. The central belt comprises the Cretaceous rift basins, including the Songliao and the Liaoxi basins (bearing the Lower-Cretaceous Jehol biota (130–120 Ma), with the well preserved feathered dinosaur *Sinosauropterix* and the earliest flowering plant, *Sinocarpus*), the Cenozoic rift basins in SKP and Inner Mongolia, and the KE rifted and volcanic basins in southern China. The eastern belt consists of the offshore Cenozoic rift basins developed on the submerged palaeocontinents, which were mainly terrestrial in the Palaeogene and marine in the Neogene. Rifting and opening of the South China Sea occurred in the Oligocene and Miocene, when the SCS Massif was split into two fragments (Figure 3, inset map).

The northward accretion of the Gondwana-affiliated massifs to Eurasia and the formation of Qinghai-Tibet Plateau The Gondwana-affiliated massifs with

Precambrian basement include the Himalaya, the Gangise, the South Qiangtang, and the North Qiangtang. The Tianshuihai area of Karakorum, stable in the Lower Palaeozoic, may be a part of the North Qiangtang Massif. The Cathaysian *Gigantopteris*-flora found in the Shuanghu area of North Qiangtang, in contrast to the Gondwanan flora found in northern Gangdise, has led to a suspected biogeographic boundary between the North and South Qiangtang massifs (Figure 3).

The main Tethys oceanic basin to the north of the North Himalaya is represented by the main suture YZCZ. The recent discovery of Cambrian to Ordovician metamorphics in the Lhagoi Kangri zone may indicate the splitting of Gangdise from Himalaya-India in the Caledonian Stage. Continental marginal thick deposits appeared in the Permian to Late Triassic, and the marine basin may have been at its widest in the Jurassic. The northward subduction of the Himalaya oceanic plate beneath Gangdise probably began in the Late Cretaceous (*ca*. 70 Ma), and the collision of the two shown occurred in the mid-Eocene (*ca*. 45 Ma), as shown by the lower part of Linzizong volcanics in Gangdise, and the final disappearance of the residual seas. On the north side of Gangdise, a southerly subducted island-arc system of short-duration (Early Jurassic to Late Cretaceous) formed the Bangong-Nujiang suture (BNAZ). In the Cenozoic, intracontinental collisions resulting in vertical crustal thickening and shortening were prevalent in northern Tibet. Crustal thickening by subduction of the Himalaya beneath the Gangdise and subsequent collision produced muscovite/two mica granite of Early Miocene (*ca*. 20–17 Ma) age in the North Himalaya. In contrast, collision between adjacent massifs without subduction were more frequent and formed shonshonitic volcanism, as in the northern volcanic belt of North Qiangtang in the Oligocene. The Himalayan Orogeny and uplift of the Plateau occurred in two stages; Oligocene and Pliocene to Pleistocene (Figure 2).

It has been estimated that no less than 1000 km of north-south crustal shortening has occurred since the Early Palaeogene collision between Himalaya-India and the northern massifs, which has led to the eventual uplift of the Qinghai-Tibet Plateau. Except for the northward subduction of the Himalaya beneath Gangdise, the shortening was mostly accommodated by distributed vertical thickening of the crust, and the main strike-slip faults and thrust belts. The eastward extrusion of the Plateau in the Sanjiang belt of eastern Tibet is significant, but may not be vital in the Plateau construction. The northward indentation from Himalaya-India has been prevalent, but the southward indentation from the rigid Tarim craton and the Beishan massif is equally important in the overall dynamics of western China.

Geology of Mongolia

Tectonic Units and Tectonic Stages of Mongolia

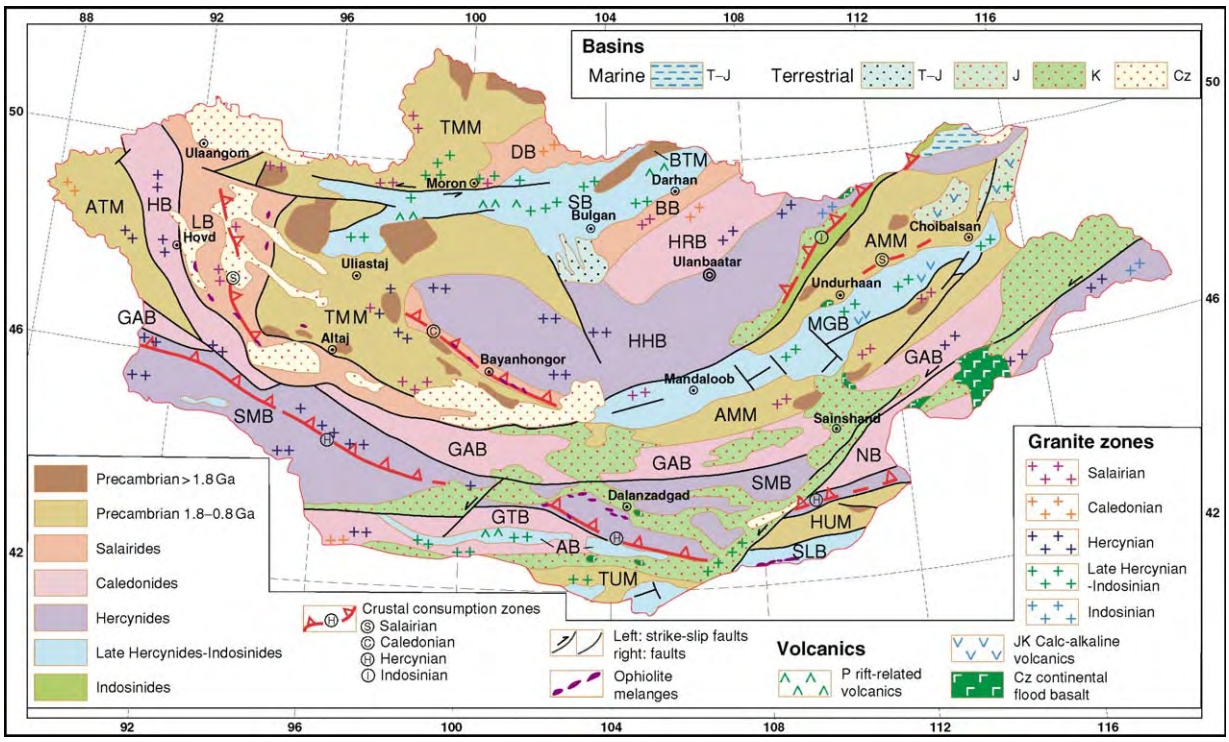
Mongolia is situated in the central part of the wide and complicated orogenic belts between the Siberian Platform in the north and the North Chinese platforms SKP and TAP in the south. Mongolia is subdivided into a northern domain and a southern domain, the boundary between which runs roughly along the southern margin of the Gobi-Altai-Mandaloovo (GAB) Belt of Caledonian age (Figure 5).

Three megastages may be recognised in the crustal evolution of Mongolia, approximately corresponding to the megastages of China. They are: (i) Neoarchaean (An) to Early Neoproterozoic (*ca*. 850 Ma); (ii) Late Neoproterozoic to Triassic, including the Salairian, Caledonian and Hercynian Orogenies; and (iii) Indosinian stages, the Mesozoic to Cenozoic, characterized by intracontinental development (Figure 2).

Mongolia in the Neoarchaean to the Early Neoproterozoic

The oldest rocks of Mongolia are found in the Neoarchaen Baidrag Complex in the southern part of the Tuva-Mongolia Massif (TMM), where a tonalite-gneiss has yielded a U-Pb zircon isotopic age of $2\,646 \pm 45$ Ma. In the same region, the Bumbuger Complex of granulite facies has a metamorphic zircon age of 1839.8 ± 0.6 Ma, which is coeval with the Luliangian Orogeny of China. Palaeoproterozoic massifs with reliable isotopic age dating are distributed in the TMM (Figure 5). A metamorphic age of *ca*. 500 Ma was obtained from the northern part of TMM (Songelin block), which may be attributed to the influence of the Salairian Orogeny in the region. In southern Mongolia, Palaeoproterozoic rocks may exist in the east-west trending Hutag Uul Massif (HUM) and Tzagan Uul Massif (TUM) near the southern border part of Mongolia.

Mesoproterozoic to Early Neoproterozoic rocks are widespread and form the main Precambrian basement in Mongolia. In northern TMM, the Hugiyngol Group, composed of metabasalts and metasediments including blue schists, have a metamorphic age of 829 ± 23 Ma. In the Mongol Altai Massif (ATM) and adjacent part of China, the thick Mongol-Altai Group, composed of highly mature terrigenous deposits, is unconformably overlain by fossil-bearing Ordovician beds. Sm-Nd model ages of *ca*. 1400–1000 Ma, obtained in westernmost Mongolia and adjacent parts of China, indicate the Precambrian age



Color Image

Figure 5 Outline tectonic map of Mongolia. Compiled and integrated from the Geological Map of Mongolia, 1:5 000 000, published by Institute of Geology and Mineral Resources, 1998, Ulaanbaatar, Mongolia.
Tectonic units. Northern Mongolia: TMM Tuva Mongolia Massif, AMM Arguna Mongolia Massif, ATM Altai Mongolia Massif, BTM Buteel Massif, LB Lake Salairian Belt, BB Bayangol Salairian Belt DB Dzhida Salairian Belt, HB Hovd Caledonian Belt, HRB Haraa Caledonian Belt, GAB Gobi Altai Mandalovoo Caledonian Belt. HHB Hangay Hentey Hercynian Belt, SB Selenge Late Hercynian Indosinian Belt, MGB Middle Gobi Late Hercynian Indosinian Belt. Southern Mongolia: TUM Tsagan Uul Massif, HUM Hutag Uul Massif, NB Nuketdavaa Caledonian Belt, GTB Gobi Tienshan Caledonian Belt, SMB South Mongolian Hercynian Belt. AB Atasbogd Late Hercynian Indosinian Belt, SLB Sulinheer Late Hercynian Indosinian Belt.

of the Mongol-Altai Group Meso- and Neoproterozoic sequences, including stromatolitic carbonates that occur in the southern belts of TMM and in AMM (Figure 5) in north-eastern Mongolia, which may be partly regarded as paracover strata on the ancient basement, as is the case in TAP and SKP of China.

Mongolia in the Late Neoproterozoic to the Triassic

The Salairian Stage (Late Neoproterozoic to Early Cambrian) The Salairian Stage, ranging from Late Neoproterozoic to Early Cambrian in age, is important in Mongolia. Well-developed Salairian orogenic belts are distributed in western Mongolia and are composed of Vendian to Cambrian siliciclastics, carbonates, and volcanics that were probably partly formed in an archipelagic ocean basin with island arcs and sea-mounts. Recent studies of the accompanying ophiolites have yielded U-Pb zircon ages of 568–573 Ma and Sm-Nd ages of *ca*. 520 Ma in many places, which are in conformity with the ages formerly determined by fossils. Synorogenic granites in the Lake area bear isotopic ages of Middle to Upper Cambrian. In northern Mongolia, there are two northeast-trending Salairian orogenic belts (DB and BB) extending eastwards into Russia, in which post-orogenic granites of Ordovician age are found. Research on the well-known Bayanhongor ophiolite zone, which is situated in a Caledonian Belt, has revealed that the ophiolites was emplaced by obduction in the time interval 540–450 Ma of Salairian age. A discontinuous belt, indicating the Salairian Orogeny, is known near Choibalsan in north-eastern Mongolia, which may have some connection with the Salairian Belt in the Okhotsk region (Figure 1). In view of the discovery of granites of Salairian age on the western border of the Jamus-Xingkai Massif of China (Figure 3), it seems that the Salairian Orogeny was active in both eastern Mongolia and north-eastern China.

The Caledonian stage (Middle Cambrian to Silurian) In the Hovd Caledonides of western Mongolia, the Lower Palaeozoic sequence, consisting of thick flysch of Tremadocian and older ages, and unconformably overlying Ordovician and Silurian carbonates and

clastics, with mafic and intermediate volcanics, are well developed. In central and eastern Mongolia, the Ordovician–Silurian sequence in the Gobi-Altai-Mandalovoo (GAB) Caledonian Belt includes a lower and an upper part separated by an unconformity, and both parts were much reworked in the Hercynian Orogeny. This extensive east-west trending belt (GAB) may mark the southern margin of the Early Palaeozoic northern Mongolian palaeocontinent. In southern Mongolia, in the Caledonian Gobi-Tianshan Belt (GTB), Ordovician to Silurian metasediments, with intercalated metavolcanics, are unconformably overlain by Devonian clastic desosits. Similar sequences are observed in the adjacent Junggar region of China. Thus the Caledonian Orogeny played an important role both in southern Mongolia and north-western China.

The Hercynian and Indosinian stage (Devonian to Triassic) The Early Hercynides of mainly Devonian to Early Carboniferous age are predominant in southern Mongolia. The South Mongolian Hercynian Belt (SMB) represents the main belt of Late Palaeozoic crustal increase in southern Mongolia and extends on both ends into China. To the west, it is continuous with the Ertix-Almantai zone in China and further to the west with the Zaysan fold zone in Russia (Figures 1 and 3). The Upper Palaeozoic in the central part of SMB is composed mainly of Devonian and Lower Carboniferous arc-related volcanics and clastics, including some Upper Silurian beds in the basal part. The ophiolite mélange zones within SMB are not continuous, and were evidently later dismembered. The lower part of the sequence includes the Berch Uul Formation consisting of a thick sequence of uppermost Silurian to Devonian tholeiitic pillow lavas, andesites, and tuffs, and Upper Devonian intermediate to felsic volcanics and olistostromes with coral limestone clastics. Frasnian conodonts and intrusive rocks dated at *ca*. 370 Ma are found in the arc-related volcanics. The upper part of the sequence comprises Lower Carboniferous fine-grained sandstones and mudstones with rich shallow marine fossils, which may have been formed after collision, although pyroclastic beds denoting volcanic activities are known to occur. In the Precambrian ATM, the Caledonian Hovd Belt (HB) and other parts of northern Mongolia, Devonian carbonates and clastics with felsic to intermediate volcanic beds are widespread as cover strata on the basement.

Thick Devonian flysch-like deposits are also developed in the broad central Mongolian Hangay-Hentey (HHB) Belt. The main part of HHB were folded after the Early Carboniferous, but a residual sea trough seems to have lingered on the eastern side of its north-eastern segment, which was essentially closed in Indosinian time. The resultant narrow Indosinide (Figure 5) sea extended to Russia and was probably continuous with the Mongol-Okhotsk seaway that finally closed in the Jurassic.

The Late Hercynian and Indosinian are usually inseparable in Mongolia. They are the Selenge Belt (SB) in northern Mongolia, the Middle Gobi Belt (MGB) in eastern Mongolia, the Sulinheer Belt (SLB), and the Atasbogd Belt (AB) in southern Mongolia. The latter belt extends both eastwards and westwards into China. All the belts are characterized by thick sequences of bimodal aulacogen volcanics, volcaniclastics, and pyroclastics of Late Carboniferous to Early Permian age dated by plant remains. To a certain extent, they are comparable with the sequence in the Bogda Mountains, which is a Carboniferous aulacogen that separated the Junggar and Tuha massifs in northern Xinjiang, China. There is, however, an alternative suggestion that a Carboniferous to Permian arc-basin system may have developed in the SLB on the southern border of Mongolia, which is continuous with the Late Hercynides–Early Indosinides belt in Inner Mongolia (Figure 3). The superimposed Triassic–Jurassic terrestrial basins developed on SB and HHB, in northern Mongolia, consist of Triassic molasse-like and coal-bearing deposits and trachyandesites and Jurassic clastic sediments, which are well dated by plant fossils. In the Noyon Basin in southern Mongolia, the Early Triassic *Lystrosaurus hedini* is found, which is almost identical to that known from the Ordos Basin.

Mongolia in the Post-Indosinian

After the Indosinian Orogeny, Mongolia, like the main parts of China, entered a new stage of intracontinental development. No marine deposits are known in Mongolia after the Triassic. Jurassic volcani-sedimentary basins with calc-alkaline volcanics occur in north-eastern Mongolia, which are the same as in the adjacent Xingan Mountains region of China. Cretaceous basins are widely distributed in southern Mongolia, for example the Zuunbayan Basin near Sainshand, and contain the renowned Jehol biota of Early Cretaceous age as in north-eastern China. Cenozoic basins are widespread in western Mongolia. In the Transaltai Basin situated to the south-east of Bayanhongor, abundant Palaeogene mammal remains have been discovered, which may be compared with those found in Inner Mongolia within China.

Conclusions

The history of China and Mongolia is discussed in terms of tectonic units and tectonic stages. The crustal

evolution of China included three megastages in the Precambrian, marked respectively by the aggregation of continental nuclei (2.8 Ga), the lateral growth and consolidation of proto-platforms through the Luliangian Orogeny (1.8 Ga), and the cratonization and coalescence of platforms into the Cathaysiana Supercontient through the Jinningian Orogeny (830 Ma). Until the Jinningian, the crustal evolution of China seems to have been dominated by continental growth, consolidation, and convergence to form a part of the Neoproterozoic Rodinia. In Mongolia, only the last megastage, ending at 830 Ma, marked by the formation of the main massifs, is recognized.

After the Jinningian, China and Mongolia entered a megastage characterized by a tectonic pattern consisting of discrete continents and ocean basins, until their reassembly at the close of the Indosinian Orogeny (210 Ma). The Cathaysiana Supercontinent began to dissociate in the Cambrian, and ocean basins were formed between Sino-Korea and Qaidam, which was entirely closed through the Caledonian Orogeny, with marked collision zones. The wide Caledonide between Yangtze and Cathaysia was, however, folded and uplifted without clear collision. To the north of Tarim and Sino-Korea, the narrow Caledonides represent continent-arc accretion. In Mongolia, the northern Mongolian massifs were successively accreted to the Siberia Platform, and the Mongolian massifs, the Salairides and Caledonides, together formed the northern Mongolian palaeocontinent, with the Gobi-Altai Caledonian Belt as its southern margin. Two main branches of Late Palaeozoic oceans, the Zaysan-South Mongolia-Hingan in the north, and the Ural-Tianshan in the south, were consumed mainly after the Early Carboniferous, and are represented respectively by the main Hercynian sutures (Figure 1). The Late Carboniferous to Early Triassic marine basins in southern Mongolia and Inner Mongolia of China probably formed an ocean with scattered islands that were filled up without appreciable collision. Furthermore, the Late Hercynides-Indosinides within northern Mongolia were actually intracontinental residual seas.

To the south of the Kunlun-Qinling central orogenic belt of China, an open sea had persisted since Early Palaeozoic, and the wide Indosinides are marked by the main Indosinian (Muztagh-Maqen) convergent zone in the north and the Jinshajiang zone in the south. The main collision zones usually coincide with older collision zones; in other words, they are polyphased or superimposed collision zones. It was at the close of the Indosinian Stage that the Laurasia Supercontinent took its final shape as the northern half of the Permian-Triassic Pangaea.

The post-Indosinian megastage of China and Mongolia witnessed an entirely new tectonic regime in East Asia, due to the appearance of the Circum-Pacific domain as a result of Pangaea disintegration and the opening of the Atlantic. The subduction of the western Pacific beneath East Asia in the Jurassic caused a continent marginal magmatism along eastern China, including the Hingan belt and eastern Mongolia. This new pattern brought about an apparent change of contrast between northern and southern China to that between eastern and western China. In eastern China, and to a certain extent in eastern Mongolia, there occurred a combination of continental margin type and intracontinental type of volcanism, which was followed by the Late Cretaceous to Cenozoic tensional regime of rifted basins and consequent crustal and lithospheric thinning. In western China, the tectonic process in the Qinghai-Tibet Plateau consisted of the northward accretion of the Gondwanan massifs to Eurasia, characterized by the northward subduction of the Himalaya beneath Gangise in the south, the distributed crustal thickening and shortening in the middle, and the southward indentation from Tarim and Mongol-Siberia in the northern part. The contrast between the compressional versus extensional, and between the crustal and lithospheric thickening versus thinning regimes between western China and eastern China are evident. These features may have reflected and induced the deeper process of an eastward flow of the asthenosphere from under western China, which might have, in turn, caused mantle upwelling and crustal and lithospheric thinning in eastern China.

See Also

Asia: Central; South-East. **Gondwanaland and Gondwana**. **Indian Subcontinent**. **Japan**. **Pangaea**. **Russia**.

Further Reading

Badarch G, Cunningham WD, and Windley BF (2002) A new terrane subdivision for Monglia: implications for the Phanerozoic crustal growth of Central Asia. *Journal of Asian Earth Sciences* 21: 87 110.

Deng JF, Zhao Hailing, Mo Xuanxue, Wu Zongxu, and Luo Zhaohua (1996) *Continental roots plume tectonics of China: key to the continental dynamics*. Beijing: Geological Publishing House. (In Chinese with English abstract.)

Dewey JF, Shackelton RM, Chang C, and Sun W (1994) The tectonic evolution of the Tibetan Plateau. *Philosoph ical Transactions of the Royal Society of London, Ser. A* 327: 379 413.

He Guoqi, Li Maosong, Liu Dequan, Tang Yanling, and Zhou Ruhong (1988) *Palaeozoic Crustal Evolution and*

Mineralization in Xinjiang of China. Urumqi: Xinjiang People's Publishing House and Educational and Cultural Press Ltd. (In Chinese with English abstract.)

Huang TK (1978) An outline of the tectonic characteristics of China. *Eclogae Geologicae Helvetiae* 71(3): 811 635.

Li Chunyu, Wang Quan, Liu Xueya, and Tang Yaoqijng (1982) *Explanatory Notes to the Tectonic Map of Asia*. Beijing: Cartographic Publishing House.

Liu Baojun and Xu Xiaosong (eds.) (1994) *Atlas of the Lithofacies and Palaeogeography of South China (Sinian Triassic)*. Beijing, New York: Science Press.

Ma Lifang, Qiao Xiufu, Min Longrui, Fan Benxian, and Ding Xiaozhong (2002) *Geological Atlas of China*. Beijing: Geological Publishing House. (61 maps and ex planations.)

Ren Jishun, Wang Zuoxun, Chen Bingwei, Jiang Chunfa, Niu Baogui, Li Jingyi, Xie Guanglian, He Zhengchun, and Liu Zhigang (1999) *The Tectonics of China from a Global View A Guide to the Tectonic Map of China and Adjacent Regions*. Beijing: Geological Publishing House.

Shi Xiaoying, Yin Jiaren, and Jia Caiping (1996) Mesozoic and Cenozoic sequence stratigraphy and sea level change cycles in the Northern Himalayas, South Tibet, China. *Newsletters on Stratigraphy* 33(1): 15 61.

Tomurtogoo O (ed.) (1998 1999) *Geological Map of Mon golia (1:1 000 000) and Attached Summary*. Mongolia: Ulanbaatar.

Wang Hongzhen (Chief Compiler) (1985) *Atlas of the Palaeogeography of China*. Beijing: Cartographic Publishing House. (143 maps, explanations in Chinese and English.)

Wang Hongzhen and Mo Xuanxue (1995) An outline of the tectonic evolution of China. *Episodes* 18(1 2): 6 16.

Wang Hongzhen and Zhang Shihong (2002) Tectonic pattern of the world Precambriam basement and problem of paleocontinent reconstruction. *Earth Science Journal of China University of Geosciences* 27(5): 467 481. (In Chinese with English abstract.)

Xiao Xuchang, Li Tingdong, Li Guangcen, Chang Chengfa, and Yuan Xuecheng (1988) *Tectonic Evolution of the Lithosphere of the Himalayas*. Beijing: Geological pub lishing House. (In Chinese with English abstract.)

Yin A and Harrison TM (2000) Geologic evolution of the Himalaya Tibetan Orogen. *Annual Review of Earth and Planetary Sciences* 28: 211 280.

Zhang Guowei, Meng Qingren, and Lai Shaocong (1995) Tectonics and structure of Qinling orogenic belt. *Science in China, Ser. B* 38(11): 1379 1394.

Zhong Dalai, *et al.* (2000) *Paleotethysides in West Yunnan and Sichuan, China*. Beijing, New York: Science Press.

Zonenshain LP, Kuzmin ML, and Natapov LM (1990) *Geology of USSR: A Plate Tectonic Synthesis*. Geody namic Series 21. Washington, DC: American Geophysical Union.

CLAY MINERALS

J M Huggett, Petroclays, Ashtead, UK and The Natural History Museum, London, UK

Introduction

Clay minerals are a diverse group of hydrous layer aluminosilicates that constitute the greater part of the phyllosilicate family of minerals. They are commonly defined by geologists as hydrous layer aluminosilicates with a particle size $<2\,\mu$m, while engineers and soil scientists define clay as any mineral particle $<4\,\mu$m (*see* **Soils:** Modern). However, clay minerals are commonly $>2\,\mu$m, or even $4\,\mu$m in at least one dimension. Their small size and large ratio of surface area to volume gives clay minerals a set of unique properties, including high cation exchange capacities, catalytic properties, and plastic behaviour when moist (*see* **Analytical Methods:** Mineral Analysis).

Clay minerals are the major constituent of fine-grained sediments and rocks (mudrocks, shales, claystones, clayey siltstones, clayey oozes, and argillites). They are an important constituent of soils, lake, estuarine, delta, and the ocean sediments that cover most of the Earth's surface. They are also present in almost all sedimentary rocks, the outcrops of which cover approximately 75% of the Earth's land surface. Clays which form in soils or through weathering principally reflect climate, drainage, and rock type (*see* **Weathering**; **Palaeoclimates**). It is now recognized that re-deposition as mudrock only infrequently preserves these assemblages, and clay assemblages in ocean sediments should not be interpreted in terms of climate alone, as has been done in the past. Most clay in sediments and sedimentary rocks is, in fact, reworked from older clay-bearing sediments, and many are metastable at the Earth's surface. This does not preclude the use of clays in stratigraphic correlation, indeed it can be a used in provenance studies. A few clays, notable the iron-rich clays form at the Earth's surface either by transformation of pre-existing clays or from solution. These clays are useful environmental indicators, so long as they are not reworked.

Clay Structure and Chemistry

Clays can be envisaged as comprising sheets of tetrahedra and sheets of octahedra (Figure 1). The general formula for the tetrahedra is T_2O_5, where T is mainly Si^4, but Al^{3+} frequently (and Fe^{3+} less frequently) substitutes for it. The tetrahedra have a hexagonal arrangement, if not distorted by substituting cations (Figure 2). The octahedral sheet comprises two planes of close-packed oxygen ions with cations occupying the resulting octahedral sites between the two planes (Figure 1B). The cations are most commonly Al^{3+}, Fe^{3+}, and Mg^{2+}, but the cations of other transition elements can occur. The composite layer formed by linking one tetrahedral and one octahedral sheet is known as a 1:1 layer. In such layers, the upper-most unshared plane of anions in the octahedral sheet consists entirely of OH groups. A composite layer of one octahedral layer sandwiched between two tetrahedral layers (both with the tetrahedra pointing towards the octahedral layer) is known as a 2:1 layer. 2:1 clays of the mica and chlorite families have multiple polytypes defined by differences in stacking parallel to the c axis. If the 1:1 or 2:1 layers are not electrostatically neutral (due to substitution of trivalent cations for Si^{4c} or of divalent for trivalent cations) the layer charge is neutralized by interlayer materials. These can be cations (most commonly K^+, Na^+, or NH_4^+), hydrated cations (most commonly Mg^{2+}, Ca^{2+}, or Na^+) or single sheets of hydroxide octahedral groups ($Al(OH)_3$ or $Mg(OH)_2$ (Figure 3). These categories approximately coincide with the illite, smectite, and chlorite-vermiculite families of clays. It is evident that the different types of interlayer cation will have a direct affect upon the thickness of the clay unit cell in the 001 dimension. This property, together with the ease with which the interlayer cations are hydrated or will interact with organic compounds, is much used in X-ray diffraction to identify clay mineral.

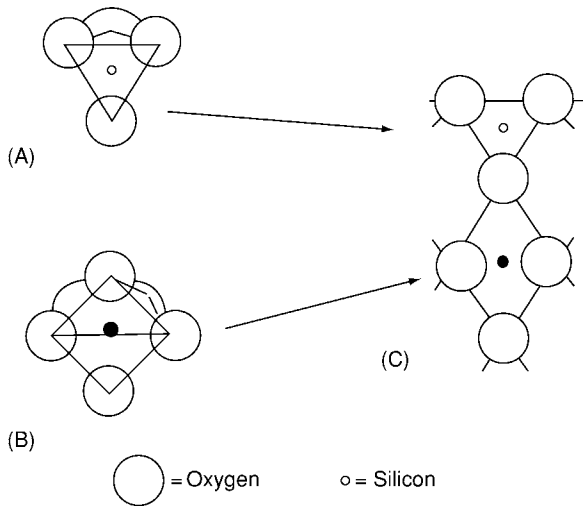


Figure 1 (A) Tetrahedrally co ordinated cation polyhedrons; (B) octahedrally co ordinated cation polyhedrons; (C) linked octahedral and tetrahedral polyhedrons.

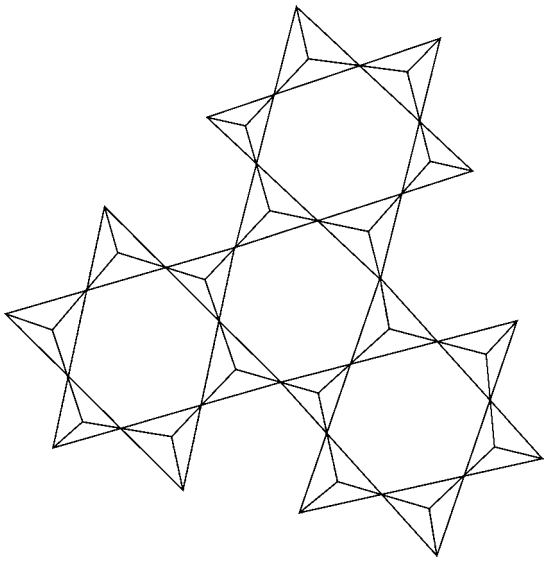


Figure 2 Hexagonal arrangement of edge linked tetrahedra.

The principal clay physical properties of interest to the geologist are cation exchange capacity and interaction with water. Clays have charges on (001) layer surfaces and at unsatisfied bond edge sites. An important consequence of these charges is that ions and molecules, most commonly water, are attracted to and weakly bonded to clay mineral particles. In most cases cations are attracted to layer surfaces and anions to edge sites. If the interlayer charge is low, cations between 1:1 layers can be exchanged for other cations and these cations can be hydrated by up to two water layers. Water in the interlayer site is controlled by several factors including the cation size and charge. In 'normal' pore fluid, one layer of water is associated with monovalent cations, two layers with divalent cations. Water can also weakly bond to the outer surface of clay particles. The relative ease with which one cation will replace another is usually:

$$Na^+ < K^+ < Ca^{2+} < Mg^{2+} < NH_4^+$$

i.e., NH_4^+ is usually more strongly fixed in the interlayer sites than is Na^+. The exchangeability of cations is measured as the cation exchange capacity (CEC). This technique is used to characterise clay reactivity, and each clay mineral has a characteristic range of CEC values (Table 1).

Classification

Clays are normally classified according to their layer type, with layer charge used to define subdivisions

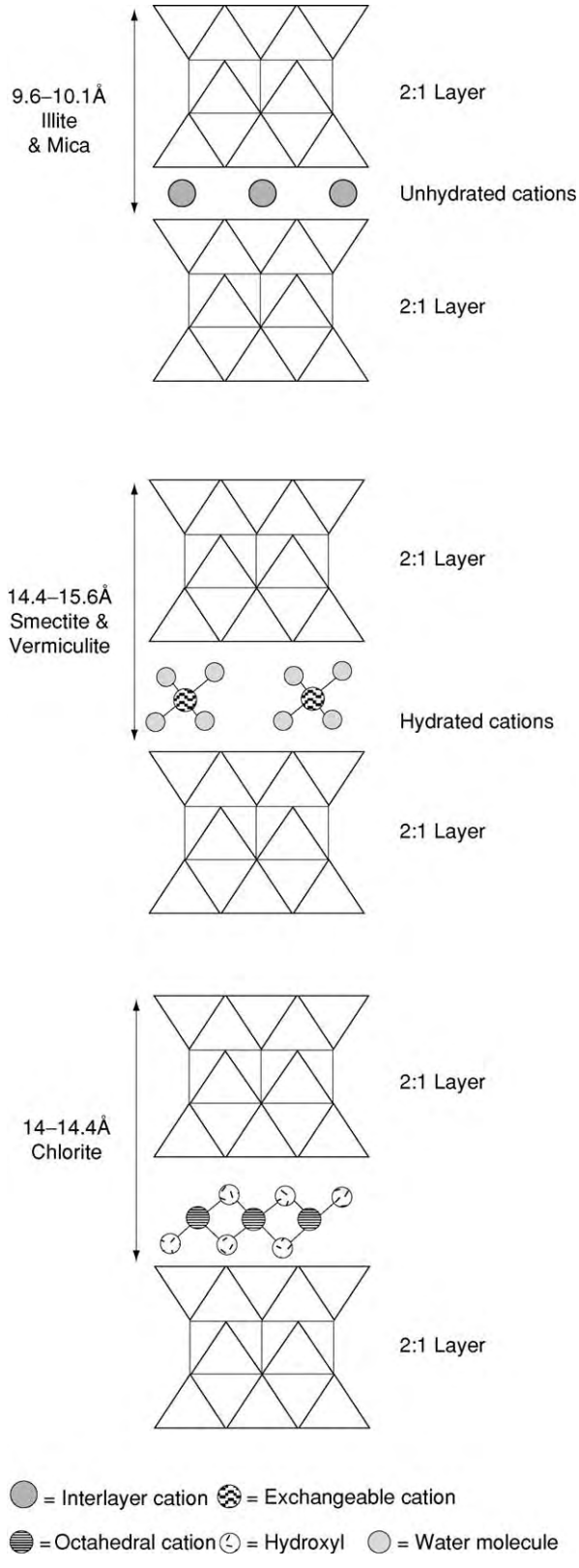


Figure 3 2:1 Clay structures.

Table 1 Typical values for cation exchange capacities. Cation exchange capacities (CEC) in milliequivalents/gram. (Data from multiple sources.)

	meq/100 g
Kaolinite	3 to 18
Halloysite	5 to 40
Chlorite	10 to 40
Illite	10 to 40
Montmorillonite	60 to 150
Vermiculite	100 to 215

(Table 2). Because of their fine particle size, clay minerals are not easily identified by optical methods, though the distinctive chemistry (and sometimes habit or morphology) of most allows identification by X-ray analysis in electron microscope studies. The most reliable method of clay identification, particularly in very fine grained rocks is X-ray diffraction, either of the bulk sample, or more reliably, of the fine fraction (usually $<2\ \mu$m). The response of the clays to glycol or glycerol solvation, cation saturation, and heat treatment is used to determine which clays are present and the extent of any interlayering (*see* **Analytical Methods:** Mineral Analysis).

Clays (Serpentine and Kaolin) (1:1)

Berthierine, $(FeMg)_3$-x(Fe_3Al)x$(Si_{2-x}Al_x)O_5(OH)_4$, is the FeII-rich member of the serpentine subgroup most commonly encountered in unmetamoprhosed sedimentary rocks and odinite is its FeIII-rich counterpart found (so far) in Eocene and younger sediments. Chemically, the kandite minerals are alumina octahedra and silica tetrahedra with occasional substitution of Fe^{3+} for Al^{3+}. Of the kandite group, kaolinite is by far the most abundant clay. Kaolinite and both halloysites are single-layer structures whereas dickite and nacrite are double layer polytypes, i.e., the repeat distance along the direction perpendicular to (001) is 14 Å, not 7 Å. Dickite and nacrite were originally distinguished from kaolinite on the basis of XRD data, however this is seldom easy, especially in the presence of feldspar and quartz. Infra-red and differential thermal analysis can be successfully used to make the distinction between these two clays.

Clays (Talc and Pyrophyllite) (2:1)

Ideal talc $(Mg_3Si_4O_{10}(OH)_2)$ and pyrophyllite $(Al_2Si_4O_{10}(OH)_2)$ are 2:1 clays with no substitution in either sheet and hence no layer charge or interlayer cations. However, minor substitution is common.

Table 2 Clay classification by layer type Tr trioctahedral, Di dioctahedral, x layer charge, Note the list of species is not exhaustive, and interstratified mixed layer clays abound (see below). (Adapted from Brindley and Brown, 1980.)

Layer type	*Group*	*Sub group*	*Species (clays only)*
1:1	Serpentine kaolin (x ~ 0)	Serpentines (Tr)	Berthierine, odinite
		Kandites (di)	Kaolinite, dickite, 7 Å & 10 Å halloysite, nacrite
2:1	Talc pyrophyllite (x ~ 0)	Talc (Tr) pyrophyllite (Di)	
2:1	Smectite (x ~ 0.2 0.6)	Tr smectites	Montmorillonite, hectorite, saponite beidellite, nontronite
		Di smectites	
2:1	Vermiculite (x ~ 0.6 0.9)	Tr vermiculites	
		Di vermiculites	
2:1	Mica illite (x ~ 0.8)	Tr illite?	
		Di illite	Illite, glauconite
2:1	Chlorite (x variable)	Tr Tr chlorites	Chamosite, clinochlore, ripidolite etc donbassite
		Di Di chlorites	
		Tr Di chlorites	
		Di Tr chlorites	Sudoite, cookeite
2:1	Sepiolite palygorskite (x variable)	Sepiolites	Sepiolite
		palygorskites	palygorskite (syn. attapulgite)

Clays (Smectite) (2:1)

The 2:1 clays with the lowest interlayer charge are the smectites. This group have the capacity to expand and contract with the addition or loss (through heating) of water and some organic molecules. It is this property that is used to identify smectites by glycol or glycerol solvation and heat treatments in XRD studies. This swelling is believed to be due to the greater attraction of the interlayer cations to water than to the weakly charged layer. Montmorillonite is a predominantly dioctahedral smectite with the charge primarily in the octahedral sheet ($R^{+}(Al_3Mg_{.33})Si_4O_{10}(OH)_2$), while beidellite ($R^{+}Al_2(Si_{3.67}Al_{.33})O_{10}(OH)_2$) and nontronite ($R^{+}$ $Fe(III)_2(Si_{3.67}Al_{.33})O_{10}(OH)_2$) are dioctahedral with the charge mainly in the tetrahedral sheet (where R = mono or divalent interlayer cations). Hectorite is a rare clay, similar to montmorillonite but it is predominantly trioctahedral with Mg^{2+} and Li^{2+} in the octahedral layer. Saponite is another uncommon smectite with a positive charge on the octahedral sheet and a negative one on the tetrahedral sheet. In smectite the interlayer cations are typically Ca^{2+}, Mg^{2+}, or Na^{+}; in high charge smectites K^{+} may be present. When smectite expands, the interlayer cation can be replaced by some other cation. Hence the cation exchange capacities of smectite is high compared with nonexpanding clay minerals. CEC can therefore be used to identify smectite.

Clays (Vermiculite) (2:1)

The second group of clays with exchangeable cations is vermiculite. Vermiculite has a talc-like structure in which some Fe^{3+} has been substituted for Mg^{2+} and some Al^{3+} for Si^{4+}, with the resulting charge balanced by hydrated interlayer cations, most commonly Mg^{2+}. The layer charge typically ranges from 0.6 to 0.9. Vermiculite is distinguished from smectite by XRD after saturating with $MgCl_2$ and solvation with glycerol. This results in expansion of the interlayer to 14.5 Å, rather than the 18 Å characteristic of smectite (though there may be exceptions to this rule). Vermiculite is much less often encountered in sedimentary rocks than is smectite, probably because it is most commonly a soil-formed clay, while coarsely crystalline vermiculite deposits are formed from alteration of igneous rocks.

Clays (Mica and Illite) (2:1)

Substitution of one Al^{3+} for one Si^{4+} results in a layer charge of 1, which in true mica is balanced by one monovalent interlayer cation (denoted R). In mica the cation is usually K^{+}, less often Na^{+} or Ca^{2+} and rarely NH^{+}. The term clay grade mica is sometimes used to describe mica which has been weathered resulting in loss of interlayer cations or formation of expandable smectite layers. 2:1 clays with layer charge ~0.8 are illite ($R^{+}(Al_{2-x}Mg,\ Fe\,II,\ Fe\,III)_xSi_{4-y}Al_yO_{10}(OH)_2$) and glauconite (the ferric iron-rich equivalent of illite). Like mica, illite and glauconite are characterised by a basal lattice spacing of d(001) = 10 Å which is unaffected by glycol or glycerol solvation, nor by heating. Illite is used as both a specific mineral term and as a term for a group of similar clays, including some with a small degree of mixed layering. Illite has been described as detrital clay-grade muscovite of the 2 M

polytype, plus true 1 Md and 1 M illite polytypes (with or without some smectite or chlorite mixed layers). Illite is, however, chemically distinct from muscovite in having less octahedral Al, less interlayer K and more Si, Mg, and H_2O.

Clays (Chlorite) (2:1)

Chlorite consists of a 2:1 layer with a negative charge $[(R^{2+},R^{2+})_3(_xSi_{4x}R^{2+}_y)O_{10}OH_2]$ that is balanced by a positively charged interlayer octahedral sheet $[(R^{2+},R^{3+})_3(OH)_6]^+$. R is most commonly Mg, Fe II, and Fe III, with Mg-rich chlorite (clinochlore) generally being metamorphic (high temperature) or associated with aeolian and sabkha sediments, while Fe-rich chlorite (chamosite) is typically diagenetic (low temperature). Ni-rich chlorite (nimite) and Mn-rich chlorite (pennantite) are the other two less common principal varieties of chlorite. Most chlorites are trioctahedral in both sheets, i.e., the ferric iron content is low. Chlorites with dioctahedral 2:1 layers and trioctahedral interlayer sheets are called ditrioctahedral chlorite (the reverse, tri-dioctahedral chlorite is unknown). Chlorite classification is further complicated by the existence of different stacking polytypes.

Mixed layer clays are those which consist of discrete crystals of interlayered clays, usually just two clay species are present, though three is known, mainly from soils. The stacking can be random, partially regular, or regular. Different types of ordering are described using the Reichweite terminology (Reichweite means 'the reach back'), denoted as R0 for random mixed layering, R1 for regular alternating layers of two clay types (also called rectorite), and R3 for ABAA for regularly stacked sequences. R2 (ABA) has not been positively identified. The most frequently encountered mixed layer clay is illite-smectite because smectite is progressively replaced by illite during deep burial diagenesis, mostly in mudrock. This clay is probably more abundant than either illite or smectite. Interstratified chlorite-smectite is associated with alteration of basic igneous rock or rock fragments in sandstone. Regularly interstratified chlorite-smectite and chlorite-vermiculite with 50% of each component are both known as corrensite. 14Å chlorite interstratified with 7Å chlorite is becoming a more widely recognised clay, particularly in sandstones which have undergone some diagenesis.

Fibrous Clays (Sepiolite and Palygorskite) (2:1)

Sepiolite and palygorskite (also known as attapulgite) are structurally different from other clays in two ways. Firstly, the tetrahedral sheets are divided into ribbons by inversion though they are still bonded in sheet form, and secondly the octahedral sheets are continuous in two dimensions only. They are consequently fibrous, though their macroscopic form may be flakes or fibres. An ideal sepiolite formula is approximately $Mg_8Si_{12}\ O_{30}OH_4(OH_2)_4 \cdot X[R^{2+},(H_2O)_8]$, and palygorskite is $MgAlS_8\ O_{20}OH_3(OH_2)_4 \cdot X[R^{2+},(H_2O)_4]$. Where X = octahedral sites in sepiolite which may contain Al, Fe, Mn, or Ni, while in palygorskite may be present Na, Fe, and Mn. Cation exchange capacities are intermediate between nonexpanding and expanding 2:1 clays, and the fibrous clay structure does not swell with addition of either water or organic solvents.

Clay Formation Through Weathering and Neoformation in Soils

Most of the clay in sedimentary rocks is probably formed by weathering or in soils. However, it is important to realize that firstly the amount of neoformed clay in soils is, at any one time, small relative to the total clay in sediments and sedimentary rocks and secondly that most clay in sediments is derived through reworking of older clay-rich sediment. Clay assemblages resulting from weathering reflect the pedoclimatic conditions (temperature, precipitation, drainage, vegetation), the composition of the rock being weathered and the length of time during which weathering occurs. This applies to palaeosols as well as recent soils (*see* **Soils:** Palaeosols), if they can be shown to be unaffected by clay diagenesis. principal process involved is hydrolysis, hence the degree of weathering increases with temperature and extent of exposure to water (precipitation and drainage), though plants and micro-organisms may also be important weathering agents (*see* **Weathering**). In tectonically unstable areas, rapid weathering and erosion may prevent formation of a stable soil clay assemblage. In general, very cold or hot and dry climate results in illite and chlorite formation. Temperate climates are characterised by illite, irregular mixed layers, vermiculite, and smectite (generally beidellite in soils, montmorillonite in altered tuff beds). Fe smectite and fibrous clays (usually palygorskite) prevail in subarid climates and hot wet climates are characterized by kaolinite (and the iron oxyhydroxide goethite). If soil-formed clays are preserved without further modification, either in the soil, during transport, or diagenesis, they may be preserved as palaeoclimatic indicators. Such preservation is, however, less common and less easily distinguished from diagenetic clays or clays reworked from older sedimentary rocks, than was formerly thought. These three modes of origin may be identified by careful provenance and petrographic studies.

Micaceous clay minerals in soils are mostly rock-derived and comprise both clay-grade micas and true illite. However, illite-rich illite-smectite can form through K fixation in smectite (especially high charge smectite) in soils subjected to repeated wetting and drying. Most soil illites, whatever their origin, are dioctahedral aluminium-rich illite. Smectite formation (Figure 4) is favoured by high pH, high silica activity, and an abundance of basic cations, consequently smectite is associated with poorly drained soils on base-rich parent material. Smectite formed in this way may be mostly ferruginous beidellites. However, smectites can form through transformation and neoformation in a wide range of soil types. The situation is further complicated by the frequent alteration of volcanic ash to form smectite (usually montmorillonite), which can occur dispersed through sediment or as discrete bentonite layers (*see* **Sedimentary Rocks:** Clays and Their Diagenesis).

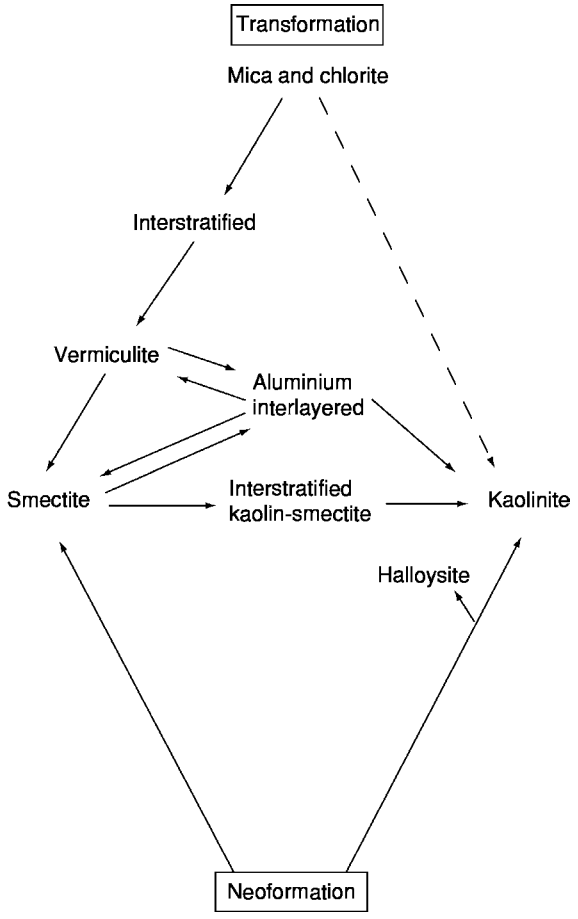


Figure 4 Pathways for the formation of clay minerals in soils. Most smectite and kaolinite is formed through neoformation while formation of kaolinite from mica and chlorite is not a true transformation due to the structural differences. (Adapted from Wilson (1999).)

Most soil vermiculite is trioctahedral and forms through weathering of biotite, hence many soil vermiculites are interstratified illite-vermiculite (called hydrobiotite when regularly interstratified). Vermiculite can also form from chlorite through breakdown of the interlayer hydroxide sheet, though again, the mixed layer clay is most frequently encountered in soils (vermiculite-chlorite is also known as swelling chlorite). Dioctahedral vermiculite is also present in soils but is less well known.

Most trioctahedral chlorite in soils is inherited. Primary chlorites are easily weathered to chlorite-vermiculite, or if weathering is intense, dissolved altogether. This is why chlorite is characteristic of high latitude soils where surface weathering is less intense than anywhere else on the Earth. So-called pedogenic chlorite and chlorite-rich interlayered clays apparently form entirely through transformation reactions involving the introduction of nonexchangeable interlayer hydroxy-Al polymers into the interlamellar sites of pre-existing smectite, and vermiculite or interstratified expandable clays. This transformation is favoured by pH 4–5.8, and it appears that in podzols at least, the reaction may neutralise acid rain.

Neoformed kaolinite is associated with tropical soils, particularly on stable cratonic areas. The reason for this association is that the intensity of weathering to form kaolinite from rock is greater than for other clays. Kaolinite formed through transformation reactions is less common but the existence of kaolinite interstratified with other clays suggests that these reactions do occur. Soil-formed kaolinite differs from diagenetic kaolinite in having some Fe substitution for Al, being less well ordered, and having a smaller particle size. Neoformed kaolinite (Figure 4) is also associated with limestone karst formation (*see* **Sedimentary Processes:** Karst and Palaeokarst). Commercial deposits of kaolinite (china clay) have formed through metasomatic alteration of granite (*see* **Clays, Economic Uses**). These are almost pure kaolinite and contain only minor quartz impurities. Kaolinite-rich beds known as underclays, tonsteins, or seat-earths occur beneath coal deposits, where they form as a consequence of intense plant-assisted weathering. Dickite is typically a high temperature diagenetic clay, while both dickite and nacrite are associated with hydrothermal deposits. Halloysite is mostly associated with young, volcanic-derived soils, in which it has a characteristic tubular or spherical morphology.

Interstratified clays are very commonly neoformed in soils. This is because the represent intermediate transformation products in reactions which proceed less fast at the Earth's surface than in during buried diagenesis. Various mica-vermiculites and illite-vermiculite are common in podzols, while

chlorite-vermiculite is found in a wide range of soil types on chlorite-bearing rock. Kaolinite-smectite (or more probably the kaolin mineral is halloysite, though this has only recently been recognized) is particularly abundant in soils formed on basic igneous rock.

Globally sepiolite and palygorskite are fairly rare though locally abundant. Recent and Quaternary deposits occur in North Africa and the Middle East, while there are well-known older Tertiary deposits in southern Spain and France. Neoformed sepiolite and palygorskite are commonly found in desert soils, ephemeral lakes, and playas, and occasionally in hydrothermal veins in basic and ultrabasic rocks. Formation of palygorskite requires confined conditions, either in shallow, enclosed water bodies at times of low lake level or in palaeosols and slow deposition. It occurs in mature, slow-formed calcrete, and in association with smectite and dolomite. Palygorskite does occur in deep water sediments. For example, the origin of palygorskite in deepwater sediments of the North Atlantic has been much debated. However, using SEM and TEM analysis, this clay has been demonstrated to be wind blown, reworked material from shallow water saline basins in Morocco.

Marine Clays

The clay mineral suite of marine environments is controlled by provenance (closely linked to climate), salinity, particle size, and neoformation. Detrital clay minerals distributions in estuaries and deltas reflect the different settling rates of the clays present and their tendency to flocculate. Where flocculation does not occur, the finer clay particles (which roughly equates with smectite) remain in suspension longest and are transported further offshore. However, no such depth profile will occur if mixing of freshwater suspended clay with saline water results in flocculation, or where marine currents are strong. The effective grain-size of clay particles is also dramatically increased through the formation of mucous-bound faecal pellets by marine invertebrates. Due to the incorporation of reactive organic matter into faecal pellets, they are prone to allogenic and early diagenetic mineralisation. An example is odinite, a ferric iron-rich clay that forms exclusively within the tropics, offshore from major river systems (which introduce large amounts of iron into the marine environment). Odinite replaces faecal pellets and also forms ooidal coatings on grains. In ancient sediments, odinite is believed to transform into ferrous iron-rich 7Å (berthierine) and 14Å (chamosite) clays. Further offshore in marine shelf sediments, the better known iron-rich clay, glauconite (and intermediate clays which range from glauconitic smectite to smectitic glauconite) forms just below the sediment water interface, again, largely through replacement of faecal pellets. Glauconite forms principally on the outer shelf, in a water depth of 100–300 m, at a temperature of 7–15°C and between 50° N and 50° S. It is also favoured by a low sedimentation rate because it is slow to form and has an He close to the redox boundary as it contains both ferrous and ferric iron. Hence glauconite is associated with transgressive sediments, and can be a useful environmental indicator, though there are many instances of reworked glauconite in turbidites, shallow marine, and nonmarine sediments.

Deep sea clays cover by far the greatest area, but are deposited very slowly. Much of the clay in these sediments is a combination of detrital and neoformed Fe-rich smectite, often associated with palygorskite (which may or may not be authigenic) and authigenic clinoptilolite. In the central Pacific, where there is perhaps the least detrital input, smectite makes up >70% of the clay fraction. Nontronite has been reported from deepsea areas of volcanic activity. Apart from neoformation of smectite, the deep ocean clays largely reflect the pedogenic processes occurring on adjacent landmasses.

Nonmarine Clays

Most continental, freshwater, and saline lacustrine clays are detrital. A few recent saline lakes contain unequivocal neoformed Fe and Mg-rich smectite. However, studies of lake clays do not always include sufficient investigation of the drainage basin to be certain that just because a lake clay is dominated by one clay type that it is neoformed. Some Tertiary lake sediments are almost monomineralic ferric rich illite, but the same provision regarding the drainage basin applies to these. There appear to be two distinct processes whereby this unusual clay forms, firstly in volcanic lakes with a high dissolved cation content, and secondly in ephemeral lakes through repeated wetting and drying of smectite. Laboratory experiments have demonstrated that such a process is possible.

Clay Stratigraphy

As a consequence of diagenetic replacement by illite, smectite is virtually unknown from pre-Mesozoic sediments, while throughout the Mesozoic and Early Tertiary, smectite is the principal clay mineral found in sediments of most types. Glauconite is most abundant at the end of the Cambrian and the end of the Cretaceous, while berthierine/chamosite are mostly found from the Ordovician to the end of the Devonian and from the earliest Jurassic until the end

of the Cretaceous. The reasons for these particular distributions through time may be related to the availability of suitable formation sites, which may in turn be linked to global climate and the formation/break up of the supercontinents. Global climatic shifts, especially in Tertiary and more recent sediments are clearly reflected in the clay minerals formed at the Earth's surface and consequently deposited in the oceans. The global cooling that began at the Eocene–Oligocene boundary and continued intermittently is reflected in the shift from a predominance of smectite to illite, plus minor chlorite and interstratified clays. At the same time, falling sea-level associated with global cooling and the build-up in polar ice resulted in the spread of nonmarine basins, hence the widespread freshwater clay deposits of the Oligocene. Recent research into clay stratigraphy is beginning to show that where post-formation modification is absent, clay mineral suites can accurately reflect Milankovitch cyclicity in sediments, and hence provide a geologically fine-scale dating tool.

Changes across stratigraphic boundaries, particularly where they are nonconformable, are not uncommon, and can be useful for dating purposes in nonmarine sediments, while within a single sedimentary unit, clear vertical changes in clay mineral suites in marine shales can be demonstrated to be linked to changes in ocean circulation and faunal changes. Note that lateral facies variations can result in lateral clay mineral suite variation, even within shales, and care should be taken when attempting to use clay stratigraphy in this way, especially if the intensity of diagenetic alteration varies between locations (*see* **Sequence Stratigraphy**).

See Also

Analytical Methods: Mineral Analysis. **Clays, Economic Uses**. **Palaeoclimates**. **Sedimentary Rocks:** Clays and Their Diagenesis. **Sequence Stratigraphy**. **Soils:** Modern; Palaeosols. **Weathering**.

Further Reading

Aplin AC, Fleet AJ, and MacQuaker JHS (1999) Muds and mudstones: physical and fluid flow properties. In: Aplin AC, Fleet AJ, and MacQuaker JHS (eds.) *Muds and Mudstones: Physical and Fluid Flow Properties*, Special Publications: 158, pp. 1 8. London: Geological Society.

Bailey SW (1980) Structures of Layer Silicates. In: Brindley GW and Brown G (eds.) *Crystal Structures of Clay Minerals and their X ray Identification*, Mineralogical Society Monograph No. 5, pp. 1 124. London: Mineral ogical Society.

Chamley H (1989) *Clay Sedimentology*. Berlin: Springer Verlag.

Colson J, Cojan I, and Thiry M (1998) A hydrogeological model for palygorskite formation in the Danian contin ental facies of the Provence Basin (France). *Clay Minerals* 33: 333 347.

Huggett JM (1992) Petrography, mineralogy and dia genesis of overpressured Tertiary and Late Cretaceous mudrocks from the East Shetland Basin. *Clay Minerals* 27: 487 506.

Huggett JM, Gale AS, and Clauer N (2001) The nature and origin of non marine 10Å green clays from the late Eocene and Oligocene of the Isle of Wight (Hampshire basin), UK. *Clay Minerals* 36: 447 464.

Moore DM and Reynolds RC (1997) *X ray Diffraction and the Identification and Analysis of Clay Minerals*. Oxford: Oxford University Press.

Merriman RJ (2002) Contrasting clay mineral assemblages in British Lower Palaeozoic slate belts: the influence of geotectonic setting. *Clay Minerals* 37: 207 219.

Robert C and Chamley H (1992) Late Eocene Early Oligocene evolution of climate and marine circulation: deep sea clay mineral evidence. *The Antarctic Paleoen vironment: A Perspective of Global Change Antarctic Research Series* 56: 97 117.

Rupert JP, Granquist WT, and Pinnavaia TJ (1987) Cata lytic properties of Clay Minerals. In: Newman ACD (ed.) *Chemistry of Clays and Clay Minerals*, Mineralogical Society Monograph No. 6: pp. 275 318. London: Mineralogical Society.

Srodon J (2002) Quantitative mineralogy of sedimentary rocks with emphasis on clays and with applications to K Ar dating. *Mineralogical Magazine* 66: 677 688.

Velde B (1992) *Introduction to Clay Minerals*. London: Chapman & Hall.

Vogt C, Lauterjung J, and Fischer RX (2002) Investigat ion of the clay fraction ($<2\,\mu$m) of the Clay Minerals Society reference clays. *Clays and Clay Minerals* 50: 388 400.

Wilson MJ (1999) The origin and formation of clay min erals in soils: past, present and future perspectives. *Clay Minerals* 34: 7 26.

CLAYS, ECONOMIC USES

Y Fuchs, Université Marne la Vallée, Marne la Vallée, France

Introduction

In common parlance, 'clays' are aggregates of minerals, soils, or rocks that commonly show plasticity. These materials contain not only minerals belonging to the clay family (*see* **Clay Minerals**) but also very fine particles with various proportions of quartz, feldspar, mica, and organic material. Clays are estimated to represent about 15% of the volume of Earth's crust as soils and as sedimentary rocks (mudstones, shales). In soils, clays play a major role in moisture retention, permeability, and adsorption of inorganic (metals) and organic solutes, as well as in the propensity for soils to shrink and swell. Thus the presence of clay minerals in soils is important for agriculture and construction. Clay has many other uses, however, and it is a particularly important commodity for industry. According to the United States Geological Survey, in 1998 there were approximately 770 active open-pit clay mines in 44 states of the USA. These mines, operated by ~240 companies that employed ~13 700 people in clay mining and milling, sold 43 million metric tons of clay products, with sales averaging $2.14 billion year^{-1}.

History

Clays have long been and are still used in different processes and products, from traditional house building (adobe in Latin America and toub in Africa, when mixed with straw) to personal-use products (shampoo, or 'rassoul', in North Africa); in AD 77, Pliny the Elder mentioned clays used as 'soap'. Advanced techniques now allow use of clays in the production of resistant, high-temperature ceramics and in remediation material and water-circulation barriers for waste repositories. Clay has been used as the building material of choice for millennia. Sun-dried bricks (adobe) and kiln-fired bricks were employed by the Sumerian and Babylonian civilizations of the Euphrates region for constructing ziggurats, city walls, palaces, and temples.

Clays have played an important role in pottery and vessel manufacture and in the trade in such goods. The earliest forms of clay pottery are about 10 000 years old. Stoneware items incorporating clays dating from 3400 years ago have been discovered in China. The oldest porcelain known was created in China during the Tang dynasty (AD 618–907). The import and trade of Chinese porcelain in Europe were important for centuries, and porcelain remained a very expensive product until it became more widely available in Florence at the end of sixteenth century.

Definitions

The term 'clay' has somewhat different meanings in different contexts (Figure 1). To the mineralogist, 'clay' is the term describing a mineral family characterized by its crystal structure. To the geologist, 'clay' designates a type of rock in which clay minerals are particularly abundant but in which other minerals, such as calcite and quartz, also contribute to the rock composition. For industrial purposes, the definition of 'clay' is based on grain size (typically $<2\,\mu$m in diameter). The importance of clays and clay minerals in economic geology derives from their fine grain size, their chemical composition, their layered structure, and the characteristics related to theses physical attributes, i.e., their role in ion exchange and their affinity for water. The importance of clays in economic geology is large and varied.

Industrial names of clay types do not correspond to the names as they are defined by the International Mineralogical Association (IMA) and as they are defined by researchers. Industrial names are generally based on the six categories established by the US Bureau of Mines: kaolin, ball clay, fire clay, bentonite, fuller's earth, and so-called common clay and shales. Kaolin, also called china clay, is composed mainly of kaolinite and other kaolinite group minerals (nacrite, halloysite, and dickite) and must be white for industrial uses (e.g., paper coating and filling, refractories, rubber, and paint). Ball clay is plastic, white-firing clay containing mainly kaolinite. It is used for manufacturing tiles, dinnerware, and bath and kitchen sinks and tubs. Fire clays withstand temperatures up to 1500°C and are used for refractories or to raise the vitrification limit. They are also composed mainly of kaolinite but may also include diaspore, an aluminium hydroxide (α-AlO(OH)), burley, ball clay, and common clay. Bentonite is an industrial clay that is composed mainly of clay minerals from the smectite group and principally of montmorillonite; it is used in foundry sands and as drilling muds. Fuller's earth is a non-plastic clay material that is characterized by its particular purifying and decolourizing properties. Fuller's earth is not very different from bentonite

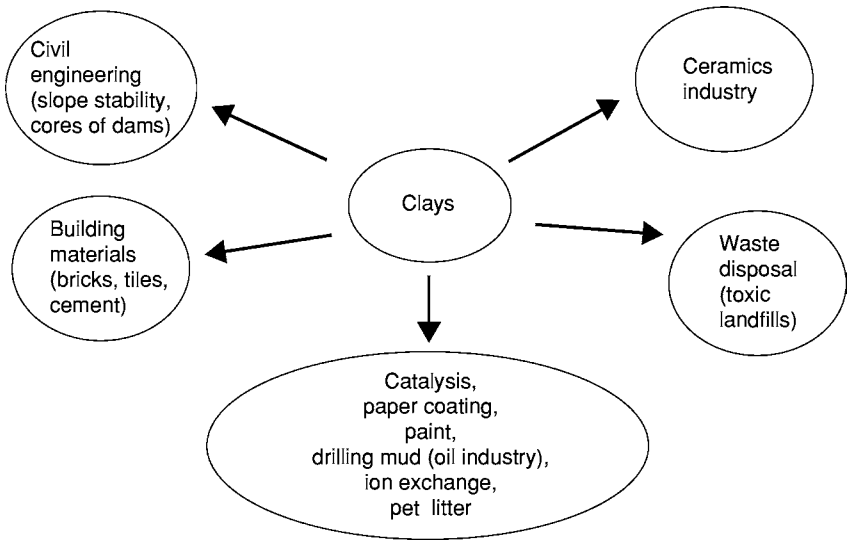


Figure 1 Uses of clays.

and is used mainly as pet litter and as an absorbent for oil and grease. Common clay is defined industrially as material that has a vitrification point below 1100°C. It consists of mixtures of different clay minerals, mainly illite and chlorite, but may also contain kaolinite and montmorillonite.

Civil Engineering

Volume changes in clay cause swelling and/or shrinking of soils, a major geological hazard for buildings and road pavement. Shrinkable clays are usually deposited as sediments in marine or brackish environments. They show little tendency to expand unless they have previously undergone a strong drying event. The shrinkage problems associated with this type of clay in soils are related to water uptake by deep tree roots, particularly in areas where the clay content exceeds 25%, the plasticity index is higher than 30%, and the water table is relatively deep seated. Expandable clays are characterized by their relatively high content of the clay mineral montmorillonite (and some other structurally related clay minerals of less importance). From an initial dry state, because of their montmorillonite content, these clays react to increasing moisture conditions by swelling. The same clays can show a reverse shrink–swell capacity with changed environment.

The presence of clays in soils is a problem for foundations of even small buildings. In all countries, structural foundations have to meet minimum requirements of building regulations. Under a building, the mass of soil that is appreciably affected by moisture content change is termed the 'active zone'. This zone represents the volume of soil that is influenced by the presence of an engineered structure. The impact can be related directly to the building (heat flow) or to other factors, such as changes caused by vegetation (effects of tree roots). In some countries, the depth of frost action is an important concern, particularly for structures that do not lose appreciable heat to the subsoil.

Building Material

Clay is a vital component in the manufacture of cements and bricks. To obtain a valuable clinker (a name for the material produced in the first stage of the cement manufacturing (after furnace)), the chemical composition of the raw material must be within defined limits. CaO is the most important component (from 60 to 69% in weight content), but the crude mix must also contain SiO_2 (18–24%), Al_2O_3 (4–8%), Fe_2O_3 (1–8%), and minor components such as MgO, K_2O_3, and Na_2O. The four principal oxides (CaO, SiO_2, Al_2O_3, and Fe_2O_3) are absolutely necessary. Clays extracted from a quarry have variable concentrations of these oxides and therefore it may be necessary to add material to obtain the required composition. Limestone is the main source of CaO. Clay or shale will provide most of the SiO_2, Al_2O_3, and Fe_2O_3.

A typical cement raw mix comprises around 80–90% limestone and ~10–15% clay or shale. According to the United States Geological Survey, the production of cement in the United States in the year 2000 reached 90.6 million metric tons. These data indicate that the cement industry was the most important consumer of clay in the United States, requiring about 25% of the domestic clay production (total, 40.8 million metric tons). Transport of huge quantities of

low-price raw material is not economical, thus the cement industry is best located near regions where geological conditions support mining of limestone and clay deposits.

Waste Disposal

Clay is used in containment systems for landfill waste deposits to control the water fluxes in and out of the system. For this purpose, industrial designers have developed a composite material that imposes a hydraulic constraint on the clay liner. The liner is composed of a layer of montmorillonitic clay (normally sodium 'bentonite') sandwiched between two nonwoven geotextiles or bonded to a geomembrane. The thickness of the clay liner is typically about 6 mm. It has the ability to self-repair if damaged; it is equivalent to many metres of compacted clay and is not associated with problems of moisture, drying, or tree roots that normally affect compacted clay barriers.

Ceramics Industry

The ceramics industry uses clay in the manufacture of bricks, ceramics (generic), glass, heavy clay products, aggregates, tiles, and refractory products; these accounted for two-thirds of the total clay production in the United States in 1985 (total, 45 million short tons). The most important clay mineral for ceramics is kaolinite, $Al_2Si_2O_5(OH)_4$. The basic chemical principle utilized by the ceramics industry is based on the structural and chemical transformation of clay minerals into minerals of the spinel type. In these reactions, kaolinite, heated at $\sim$550°C, loses its (OH) groups (dehydroxylation process) and is transformed into a product called metakaolinite, $Al_2Si_2O_7$, with formation of two molecules of H_2O. At approximately 980°C, metakaolinite transforms into a silicon spinel, $2Al_2O_3 \cdot 3SiO_2 + SiO_2$. This phase is rather complex and probably depends on the nature and ordering of the original clay mineral. At 1100°C, mullite, $2(Al_2O_3 \cdot SiO_2)$, begins to form in association with quartz (SiO_2). The reaction continues, and at about 1400°C all the material has been converted to mullite and cristobalite (a high-temperature form of SiO_2). Pure mullite would be stable up to 1850°C, but the presence of SiO_2 generates a liquid phase at about 1590°C that can induce silica exsolution in the material, and therefore the process is stopped at about 1650°C. At 1850°C, mullite would transform into corundum and a liquid-phase SiO_2. Ceramic mullite composition varies from $3Al_2O_3 \cdot 2SiO_2$ to $2Al_2O_3 \cdot SiO_2$, with possible partial substitution of Al by Fe: $3(Al_{0.9}Fe_{0.1})_2O_3 \cdot 2SiO_2$. The general process of ceramics formation at increasing temperature consists of going from an exsolution of silica from a primary raw material containing approximately 40% Al_2O_3, to mullite with an Al_2O_3 content up to 73%.

Refractory ceramics must have a particularly high alumina content. The phase line diagram of SiO_2–Al_2O_3 has its highest point theoretically at 100% Al_2O_3 and 2020°C, thus minerals such as bauxite, with higher Al_2O_3 content than kaolinite has, are added to natural kaolin to obtain the most refractory products. Earthenware, wall tile, sanitary ware, stoneware porcelain, and electrical porcelain are manufactured using kaolinite or ball clay, quartz, and fluxes. One of the most important requirements is that the product must be white or ivory in colour. Temperatures between 1150 and 1250°C are used for vitrification except for electrical porcelain, for which it is necessary to reach 1400°C. Because of its colours and its workability, kaolin is the main clay used, and the composition of various products ranges from 100% kaolinite for porcelain to 20–30% kaolinite for sanitary ware.

New applications for ceramics are under development, particularly as monolithic catalyst carriers for standard emission control equipment on vehicles. The ceramic for this purpose has a cordierite composition ($2MgO \cdot 2Al_2O_3 \cdot 5SiO_2$), obtained by firing a mixture of 70% talc and 30% kaolinite at 1400°C. Ceramic membranes are another new technology. These membranes are high-temperature (up to 1000°C) and corrosion-resistant materials that withstand the effects of corrosive products. They are used in many different processes, such as beer filtration and hot metal separation. The market for these products is steadily growing. Other applications for ceramics in modern industrial processes include permanent foundry moulds and heat exchangers.

Physicochemical Properties of Clay Minerals

The physicochemical properties of clay minerals are strongly related to their structures and compositions, thus it is possible to organize them into three basic categories: the kaolinite group, the smectites group, and palygorskite (attapulgite) and sepiolite.

Kaolinite is a two-layer clay. Substitution of Al for Si in the SiO_4 tetrahedron (Tschermak substitution) is limited, i.e., the charge on the kaolinite layer is limited. Because of this limited charge, kaolinite has a low exchange capacity, particularly regarding other clay minerals (smectites, bentonite, etc.). Related to the low surface charge, kaolinite has low absorption and adsorption capacities. Kaolinite is a hydrated aluminous silicate with very limited substitutions

(particularly by Fe) and is therefore white in colour. Particle size is very fine. Kaolinite is normally well crystallized, with broad dimensional distribution and low surface charge. These characteristics induce good flow properties and low viscosity even at high solid contents in water. Kaolinite is hydrophilic and will disperse easily in water with addition of a small quantity of dispersant. This quality makes kaolin an essential component in paper coating and in water-based interior paints. The paper industry is by far the largest user of kaolin. Coating paper with kaolin improves the printability of paper by making the paper surface smoother, more homogeneous, and more finely porous. To coat paper, a high-concentration solid/liquid suspension of pigment and binder is applied to the paper surface with a blade. Differently processed kaolinitic products must be further prepared for rotogravure and offset printing processes. Kaolin is also used as a filler during paper manufacture to enhance paper opacity and strength.

Smectites are a group of clay minerals characterized by their three-layered structure and their chemistry, consisting of sodium, calcium, magnesium, or lithium aluminium silicates. Chemical variations of smectite minerals include sodium montmorillonite, calcium montmorillonite, saponite, nontronite, and hectorite. Bentonites are types of sodium and calcium montmorillonites derived from alterations of volcanic glasses (usually volcanic ashes) and are of great industrial and economic importance. The three-layer structure of the smectites is very important; a single octahedral sheet is sandwiched between two tetrahedral sheets. In the octahedral sheet, Fe^{2+} (ferrous iron) and/or Mg^{2+} substitutes for Al^{3+}, creating charge deficiencies in the layers. The charge balance is also affected by Al substituting for Si in the tetrahedral sites. These deficiencies are charge balanced by exchangeable cations (Na^{+}, Ca^{2+}) located in interlayer positions. Two or four OH groups are associated with the exchangeable cations in the interlayer sites. The cations and OH groups located in the interlayer can be exchanged with other cations and also with polar organic molecules such as ethylene glycol, quaternary amines, and (poly)alcohols. Smectites are generally present in nature as very small particles and therefore they have a high specific surface. In association with a high charge, this induces a high degree of viscosity when smectites are mixed with water. High cationic exchange capacity and a high degree of viscosity when associated with water are the properties that determine the industrial uses of smectites.

The most important industrial user of smectites is the oil industry. Bentonite is used as drilling mud. Its excellent swelling capacity and high viscosity enhance the stabilizing effect of freshwater drilling mud. Bentonite mixed with sand of high silica content is also used in foundries. Adding a small quantity of water yields a plastic material that can be moulded. Bentonites are also used to pelletize iron ore and as sealants for water ponds, ditches, and household basement walls. Sodium bentonite can be produced from calcium bentonite by the addition of soda ash. In the presence of a soluble sodium salt, Na substitutes for Ca in the structure. These artificial bentonites, called sodium-activated bentonites, have better swelling and gelling properties than the natural ones have. Industrial products are frequently a mixture of natural and activated Na-bentonites.

Palygorskite (attapulgite) and sepiolite are hydrated magnesium aluminium silicates. Structurally, double silica tetrahedron chains are linked together through chain-like structures formed of oxygen and OH groups, forming octahedra in which the cationic site is occupied by aluminium or magnesium ions. Palygorskite and sepiolite are seen as fibres in scanning electron microscopy. This morphology is the result of the predominantly chain-like structure. Substitutions in the octahedral layer induce a moderate charge in the structure, and the exchange capacity is intermediate between that of kaolinite and smectite ($\sim$40 meq per 100 g for palygorskite–sepiolite and 60–100 meq per 100 g for smectites). The combination in palygorskite–sepiolite of particle charge, channel structure, and large surface area results in a high absorption and adsorption capacity that has many industrial applications. The high and very stable viscosity makes these clay minerals useful in saltwater drilling mud or high-electrolyte-content drilling mud. They are currently used in drilling in oilfields where brines are normally associated with oil.

In agriculture, palygorskite and sepiolite are used as retardants in fertilizers, pesticides, and herbicides. When mixed with these products, the high adsorption and absorption properties of the silicates allow slow and progressive product release; this is both economical and environmentally beneficial. Palygorskite and sepiolite, due to their great stability in suspension, are also used as suspending agents in paints, fertilizers, pharmaceuticals, and cosmetics. They are also used for industrial cleaning of factory floors and service station bays because of their capacity to absorb grease and oil. These clay minerals also find a widespread use as pet litter.

See Also

Clay Minerals. **Sedimentary Rocks:** Clays and Their Diagenesis.

Further Reading

Boormans P (2004) *Ceramics Are More Than Clay Alone*. Cambridge: Cambridge International Science Publishers.
Bundy WM (1991) Kaolin in paper filling and coating. *Applied Clay Science* 5: 397 420.
Burst JF (1991) The application of clay minerals in ceram ics. *Applied Clay Science* 5: 421 423.
Murray HH (1991) Overview clay mineral application. *Applied Clay Science* 5: 379 395.
Murray HH, Bundy WM, and Harvey CC (eds.) (1993) *Kaolin Genesis and Utilization*. Aurora, CO: Clay Minerals Society.
Vaccari A (1999) Clay and catalysis: a promising future. *Applied Clay Science* 14: 161 198.

COCCOLITHS

See **FOSSIL PLANTS: Calcareous Algae**

COLONIAL SURVEYS

A J Reedman, Mapperley, UK

Introduction

The organization of the systematic surveying of the geology of the vast territories that formerly comprised the British Empire, mainly through the establishment of country specific geological surveys (*see* **Geological Surveys**), has been a long process and one that is not yet entirely complete. This article outlines some of the major features of that process with special attention being given to the role of the organisation known as the Directorate of Colonial Geological Surveys. Though the latter organization was only in existence for less than twenty years, in the middle of the last century, its influence on geological surveying in many parts of the world is still felt to this day.

Early Days

In view of the fact that the Geological Survey of Great Britain was established in 1835, it may seem surprising that a systematic approach by the British Government to establishing and nurturing Geological Surveys in the many territories that constituted the British Empire took a further one hundred and twelve years to come to fruition, with the founding of the Directorate of Colonial Geological Surveys in1947. A considerable number of Geological Surveys had been established in various British overseas territories well before 1947 but largely through the lobbying efforts of enthusiastic individuals, and usually through the local colonial administration, rather than as a result of any overarching policy of the centralized colonial power vested in the British Crown. The first of these local initiatives was taken just seven years after the Geological Survey of Britain came into being. In 1840, the former provinces of Upper and Lower Canada were joined by an Act of Union into a single province and a country called Canada came into being. In 1842, the legislature of the new country decided to set up a geological survey and William Logan was appointed as its first director. Shortly afterwards, in 1853, a geological survey for India was established for India, based in Calcutta.

The Imperial Institute

The foundations for official British involvement in overseas geology, administered from London, the centre of the Empire, were laid through the opening of the Imperial Institute in 1893. This organization was established as the National Memorial of the Jubilee of Queen Victoria and was initially funded by grants from the Commissioners of the Great Exhibition held in London in 1851. It was reorganized on a number occasions during its relatively short life before its final demise in the 1950s, but during this time it contributed a great deal to geological knowledge of the vast territories that constituted the 'Empire'.

The prime purpose for the existence of the Imperial Institute was to promote trade in natural raw materials in the colonies and dependencies. Amongst these commodities, minerals (*see* **Mining Geology: Mineral Reserves**) were seen to be of great importance and a Mineral Resources Department, later to become a Division, was an important component of

the establishment. In its early years, it carried out a number of pioneering reconnaissance surveys in colonial territories, mainly involving just one or two geologists, often traversing whole countries on foot, without the aid of any but the most rudimentary of topographic maps. As interest in the Empire's resources grew, the Division increasingly concentrated on providing services, such as mineral identification, assaying, and economic assessment, to those fledgling geological surveys that were beginning to be established.

A report published in the Bulletin of The Imperial Institute in 1943 summarised the work of the Mineral Resources Department during the fifty years it had so far been in existence and also reviewed the likely requirements for the future. Noting that some half a million pounds had already been earmarked to extend scientific investigation into 'Colonial problems', excluding welfare and development projects, the report drew attention to "the wisdom of allocating a fair share of any available grants to the purpose of expanding the work of Geological Surveying in the Colonial Empire". It noted that by 1943, local (overseas) geological departments existed for territories with a total area of one and a quarter million square miles (excluding the Dominions) but that the total number of trained staff responsible for geologically surveying this vast area amounted to only forty six.

The early work of the Institute, mainly carrying out reconnaissance surveys, was seen as a necessary preliminary to the formation of a Geological Survey and in a number of colonies this was achieved. Such colonies included British Malaya (now Malaysia, 1903), Ceylon (now Sri Lanka, a Mineral Survey under various subsequent names was established in 1903), the Gold Coast (now Ghana, 1913), Uganda (1918), Nigeria (1919), Tanganyika (Tanzania, 1926), Sierra Leone (1927), Kenya (1933), and British Guiana (Guyana, 1933) all established Surveys, often attached to Mines Departments and with very few geological specialists.

Directorate of Colonial Geological Surveys

Though it was widely acknowledged that the Mineral Resources Division of the Imperial Institute had achieved a great deal in advancing knowledge of the mineral resources of the colonies since its inception, by 1943 it was also recognized that there was still a serious lack of geological knowledge of much of the extensive terrain that constituted the colonies and that a new organization was needed to help rectify this situation. A joint meeting of the Geological Society of London and the Institution of Mining and Metallurgy, after considering how geological surveys in the British colonies might be strengthened in the post-war era, made a number of recommendations. The Secretary of State for the Colonies accepted the advice of a committee set up to advise him on the need for an expansion of geological work and how this might be achieved, and on the first of January 1947, a central organization for overseas Geological Surveys was inaugurated in London with the appointment of Dr Frank Dixey as its Director and as the Geological Adviser to the Secretary of State. At the age of 54, Dr Dixey had considerable experience of geological surveying in the colonies, having successfully carried out a reconnaissance survey of Sierra Leone followed by almost eighteen years in Nyasaland, first as Government Geologist and then Director of the Geological Survey where he was required to spend much of his time and geological knowledge on the practical problems of establishing, and improving, rural water supplies. In 1939, the latter experience led him to be appointed to set up a Water Department in Northern Rhodesia (Zambia) and in 1944 he moved to Nigeria as Director of the Geological Survey.

The Directorate of Colonial Geological Surveys was initially housed in offices in the ornate Victorian building originally constructed for the Imperial Institute in South Kensington, London. Work commenced in a modest way; Dixey being joined by a geologist and then, in 1949, by a geophysicist. Two Deputy Director's were appointed and the task of establishing a photogeology section was commenced. In 1949, the Colonial Office took over responsibility for the Mineral Resources Division of the Imperial Institute which, as the Mineral Resources Division of the Directorate of Colonial Geological Surveys, continued its wide-ranging work on the mineral resources of the Colonial territories and compiling statistics on mineral production throughout the Empire.

It soon became apparent that an increasing workload demanded an increase in the establishment and investment on more modern equipment, with the cost being met from Colonial Development and Welfare funds. As well as continuing to service the specialist needs of the eleven Geological Surveys that had been previously set up in various the colonies with the assistance of the Imperial Institute, it became clear that some thirteen other territories could benefit from the establishment of their own Geological Surveys, and that the greatest need of both new and pre-existing Surveys would be for an increase in their staff complement of geologists, geophysicists, and geochemists together with buildings and equipment. While the costs were to be covered, during their first few years, from Colonial Development

and Welfare Funds, most territories eventually assumed responsibility for the cost of their own Geological Surveys.

By 1957, ten years after its foundation, the organization had grown very considerably and was able to offer a broad range of services to the colonies, including the recruitment of staff to serve in their Geological Surveys. Apart from Dr Dixey, the Director, the organization now comprised a Directorate including Geophysical and Photogeology Sections (total staff 9), a Mineral Resources Division (36 staff) and a Joint Services Division (total staff 14) which included a library and serviced both the Colonial Products Laboratory and Colonial Geological Surveys.

The Mineral Resources Division not only continued to give advice to Colonial Administrations on all aspects of mineral development but was also responsible for a new publication, 'Colonial Geology and Mineral Resources', a successor to the Mineral Resources section of the former 'Bulletin of the Imperial Institute', and also continued to publish the 'Statistical Summary of the Minerals Industry'. Perhaps most significant of all, at least as far as geological mapping of the Colonial territories was concerned, was the establishment of the Photogeological Section, shortly to be housed in the Headquarters of Colonial (Geodetic and Topographic) Surveys where it was able to access the considerable stock of air photographs being built up by the latter organization.

Before 1947, geologists working in the Geological Surveys of the colonies usually had to make their own topographic maps by conventional surveying methods and frequently in the absence of a reliable trigonometrical network. The advent of air photographs not only allowed the greatly increased production of topographic maps of many remote areas at a scale adequate for geological mapping but also made photographs available for geological interpretation. Dixey reviewed the first nine years of Colonial Geological Surveys, and was able to report that, together with the increase in geological staff, it was the topographic mapping programme of Colonial (Geodetic and Topographic) Surveys that was more and more responsible for the increased rate of geological mapping in the colonial territories. In the East African territories of Kenya and Uganda for example, only a total area of about 25 000 sq miles had been geologically mapped at various scales prior to 1947, but by 1993 that figure had increased by almost 75 000 square miles, all mapped at scales of either 1:50 000 or 1:125 000.

With respect to its establishment of a Geophysics Section and, more particularly, a Photogeological Section, Colonial Geological Surveys was ahead of its much older sister organization, the Geological Survey of Great Britain, with its nearby headquarters in Exhibition Road in South Kensington. More innovation was to follow under Dr Dixey's leadership. With the help of grants obtained from the Department of Scientific and Industrial Research and the Colonial Welfare and Development Fund, a Geochemical Prospecting Research Centre was set up at the Imperial College of Science and Technology in London and this laid the foundations for many subsequent geochemical prospecting programmes both in the then Colonial territories and, later, throughout many other parts of the world. In 1957, grants were also made available to set up an Isotope Research Group in the Department of Geology and Mineralogy of Oxford University and this group, expanding in future years, was devoted primarily to problems in overseas geology.

The task of promoting and assisting the formation of Geological Surveys in the Colonial Territories and Dependencies, started by the Imperial Institute, gained momentum after the inauguration of Colonial Geological Surveys. In the first nine years of its existence, Geological Surveys had been established in a further nine countries, as shown below:

The increase in the number of overseas 'colonial' Geological Surveys saw a marked increase in the number of geologists, geochemists, and geophysicists amongst their staff. In the first ten years of the Directorate, the total rose almost fourfold from 58 to 212, and to 258 by 1961. The Directorate acted as advisers to the Colonial Office on matters of staff recruitment and also ran a variety of training courses for both individual and groups of newly recruited or established staff of the Surveys. Secondment of specialists to posts in colonial administrations or overseas Surveys was also arranged. The demand for, and supply of qualified professionals in the geological field varied from year to year but the bulk of the new recruits were young men just completing either their first or second degrees and it is estimated that each year, between 1947 and 1955, Colonial Geological Surveys absorbed between one quarter and one half of the UK universities output of suitably qualified geologists.

Bechuanaland Protectorate (now Botswana)	1948
British Territories in Borneo (now mainly part of Malaysia)	1949
Jamaica	1949
Swaziland	1949
British Solomon Islands Protectorate	1950
Cyprus	1950
Northern Rhodesia (now Zambia)	1950
Fiji	1951
Cyprus	1952
New Hebrides	1959

When the grand old Victorian building that had once housed the Imperial Institute was demolished to make way for the new buildings of Imperial College, the Directorate of Colonial Geological Surveys moved, in 1960, to new purpose-built laboratories and offices in Greys Inn Road, London. By this time, inclusion of the term 'Colonial' in the organizations name was becoming politically unacceptable, and it was changed to the Directorate of Overseas Geological Surveys. Furthermore, an increasing number of Britain's overseas possessions were gaining independence and this led the British Government to set up a committee to consider, amongst other matters, the type of technical assistance which the United Kingdom should be in a position to provide in the geological and mining fields and, in this context, the future organization, structure, and functions of the Directorate of Overseas Geological Surveys.

An End and a Beginning

The report of the 'Brundrett' Committee was presented to the British Parliament in May 1964. Most of it recommendations were accepted, including the recommendation that "the functions of the Geological Survey of Great Britain should be expanded to cover overseas work in the geological and mineral (*see* **Mining Geology:** Exploration) assessment fields" and that "the Overseas Geological Surveys and the Atomic Energy Division of the Geological Survey of Great Britain should be amalgamated within the expanded Geological Survey of Great Britain". In June 1965, the amalgamation was completed with the establishment of the Institute of Geological Sciences. Overseas Geological Surveys, an integral part of the new organization, eventually evolved into the Overseas Division and then International Division of the British Geological Survey, later to be renamed the British Geological Survey where, building on the legacy of 'Colonial Geological Surveys', the work of geological surveying and assistance has continued in many countries of the former British Empire and in others worldwide.

See Also

Geological Field Mapping. **Geological Surveys**. **Mining Geology:** Exploration; Mineral Reserves.

Further Reading

Dixey F (1957) *Colonial Geological Surveys 1947 56*, Colonial Geology and Mineral Resources Supplement Series, Bulletin Supplement no 2, p. 129. London: HMSO.

Dunham KC (1983) Frank Dixey 1892 1982. *Bibliographical Memoirs of Fellows of the Royal Society* 159 176.

Intelligence Staff of the Mineral Resources Division, Imperial Institute (1943) *A Review of Geological Survey Work in the Colonies*. Bulletin of the Imperial Institute, vol. XLI, no 4. London: HMSO.

Pallister JW (1972) *British Overseas Aid in the Field of the Earth Sciences*. 24th IGC, Symposium 2.

Report of the Committee on Technical Assistance for Overseas Geology and Mining. (1964) London: HMSO.

Walshaw RD (1994) British Government Geologists Overseas A Brief History. *Geoscientist* 2: 10 12.

COMETS

See **SOLAR SYSTEM: Asteroids, Comets and Space Dust**

CONSERVATION OF GEOLOGICAL SPECIMENS

L Cornish and G Comerford, The Natural History Museum, London, UK

Introduction

This article covers the general ethics and methodologies of conservation, which can be applied to all geological material. It outlines best practice when carrying out both preventive and remedial conservation. Examples are given of unstable material and how to approach treatment. The further reading list will allow the reader to explore specific conservation issues.

There is no international standard for the care and conservation of fossils. Natural history specimens in

general do not automatically fall under the protection of the UNESCO treaty. A great deal depends on the enabling legislation passed by individual countries. A single collection may contain materials as diverse as unaltered organic material, bone, shell, amber, and a very wide range of rock, mineral, and sediment materials. Fossils by their very nature can be preserved in isolation or in a rock matrix, and as such the compositions of the specimen and the surrounding matrix have to be considered as one unit for conservation purposes. The term fossil has therefore been used to encompass a range of geological material.

Preventive Conservation

Preventive conservation is an important part of prolonging the life of an individual specimen or collection. There are two distinct aspects to preventive conservation: the organizational and the technical. This article concentrates on the technical aspect, whilst appreciating that in large organizations the management of human resources and the physical environment are required to produce good preventive conservation practice. Preventive conservation is everyone's responsibility and not just the preserve of an individual conservator. Preventive conservation involves indirect action to slow deterioration and prevent damage by creating conditions that are optimal for the preservation of the object.

Handling

All specimens should be handled with care; even those that appear robust can be damaged by inappropriate handling. The basic principles of handling include:

- cleanliness – use clean bare hands or disposable gloves;
- avoid unnecessary handling;
- assess the weight and condition of the object;
- check for any breakages, cracks, or old repairs;
- handle specimens one at a time;
- handle associated pieces separately;
- support the specimen fully when picking it up;
- use supports to carry specimens that are fragile or cannot carry their own weight;
- use a trolley or similar device to move objects that are heavy;
- do not drag or push a specimen across a surface; and
- provide protection against environmental changes for specimens that are being moved to an area with different environmental conditions.

The first step in preventive conservation is to ensure that specimens are handled correctly. Specimens should be fully supported during transportation from the storage area to the study area. Large specimens are in danger of failing in weak areas if they are held at only one point. Before picking up a specimen, a brief assessment should be made, noting any vulnerable areas. Strain should not be placed on cracks or joins, and surfaces that are flaking or friable should be avoided. Scratching of highly polished surfaces should be avoided. Absorption of dirt, oil, and salts from the hands can occur if the specimen has a porous surface (e.g. limestone). Large specimens require thick soft padding to support them (e.g. high-density polyethylene). Specimens that are being moved from one set of environmental conditions to another should be packed to ensure a slow acclimatization.

Storage

The correct storage of specimens will prevent unnecessary damage arising through abrasion and stacking. Ensuring that each specimen is housed correctly and fully supported in storage will avoid physical damage.

- Small objects stored in drawers may be placed in boxes or trays. These trays should be of adequate size and lined with a supporting material.
- Small specimens may be stored in drawers lined with foam and placed in cut-outs.
- Boxes or trays should be of such a size that they tessellate within a drawer, preventing trays moving when the drawer is opened or closed.
- Larger specimens may require bespoke mounts to ensure that they are fully supported.
- Mounts should be lined with inert foam to ensure that there is no adverse interaction with the specimen.
- Very heavy specimens should be stored on foam-lined boards that incorporate supports for vulnerable parts.

Packaging Materials

All materials that come into direct contact with a specimen should be inert, in that they should not damage the specimen through abrasion or chemical interaction. The materials used to store and display specimens should be made from conservation-grade substances. Foams that are used to support specimens and absorb vibration should not be made from materials that degrade rapidly. A closed-cell inert-nitrogen-blown polyethylene foam that is highly chemical resistant and stable in the presence of ultraviolet radiation is a suitable foam for use in storage. This type of foam can be purchased in varying densities, and the density should be chosen according to the weight of the specimen. Larger specimens should be stored on foam-lined shelves or pallets. Other materials such as

tissue used to pack specimens for travel should be acid free, as should the trays and boxes in which specimens are stored within collections.

Environmental Conditions

Many rocks and minerals are environmentally sensitive. Environmental factors affecting specimens include light, heat, dust, relative humidity, and pollutants. Some mineral specimens may change colour owing to their sensitivity to light. Crystals may fracture if they are exposed to heat or to cycles of hot and cold temperatures. Dust can be potentially damaging to specimens as it disfigures surfaces and encourages corrosion by providing nucleation sites for the absorption of water and other pollutants. Where specimens have been remedially treated with adhesives or consolidants, dust can sink into the polymer surface and subsequently be very difficult to remove. Pyrite, clay, shale, and subfossil-bone specimens may split and crack if they are stored at an inappropriate or fluctuating relative humidity. Pollutants can occur in materials associated with the specimen, in other objects in the collection, and in the surrounding air. Materials that are known to release harmful vapours at room temperature include wood – particularly oak, birch, and chipboard. These materials should be avoided for use in storage as the vapours given off may react with some mineral specimens. The best environment in which to store rock and mineral specimens is dust free and clean, with low light and moderate temperature and humidity levels.

Light

Light-sensitive specimens or specimens whose sensitivity to light is unknown should be housed in lightproof storage. Where specimens are to be displayed, the effects of light should be minimized by one or more of the following methods:

- excluding daylight;
- using low-ultraviolet light sources;
- using ultraviolet filters;
- using display methods that ensure that the specimens are lit only whilst they are being viewed; and
- rotating the specimens between display and low-light storage.

Relative Humidity and Temperature

In purpose-built collection storage it is possible to specify particular values of relative humidity and temperature. A stable environment is paramount to providing a stable collection. Heat can damage specimens through simple excess or indirectly when it is introduced suddenly into an unheated space, destabilizing the relative humidity. Fluctuating relative humidity can lead to physical damage of some specimens. Increased moisture in the air can lead to the dissolution of specimens through the absorption of water from the environment. Conversely, loss of water to the environment can lead to changes in the chemical composition and properties of some minerals. Suitable environments for mineral, rock, and fossil collections are those with stable relative humidity and temperature. A temperature between 15°C and 21°C, fluctuating by no more than ± 1°C, and a relative humidity of $45\% \pm 5\%$ would be suitable for most geological collections.

Where the specimen storage environment is unsuitable, microclimates may be created using humidity buffers. The buffering material, commonly a silica-based product available from conservation suppliers, can be purchased in bead, sheet, or cassette form. The buffers work by absorbing and releasing moisture from the enclosed air. Individual specimens can be placed in a sealed polyethylene container with an appropriate quantity of this material, and a constant humidity will be maintained for a period. The more often the container is opened, the more quickly the buffer will become spent and require recharging. Where specimens are too large to fit into polyethylene containers, microclimates can be produced by encapsulating the specimen in a buffer in a moisture-barrier film sealed with a heat sealer. These methods work very well for individual specimens but alternatives are required for larger collections of humidity-sensitive material.

Environmental Monitoring

The environment of the collection or specimen requires monitoring. This can be done in a number of ways, depending on the size of the task and the funds available. Simple humidity-indicator cards can be placed with the specimen to give a general indication of the amount of moisture in the atmosphere. Dial hair hygrometers will give a more specific reading, and recording thermohygrographs will give a weekly or monthly trace. These methods all have their uses, albeit limited. Electronic data loggers are more useful; they are programmed and the data downloaded using a computer. The data are then more easily analysed. Radio telemetric systems and building management systems give real-time data. This type of monitoring allows the environment to be measured remotely and is very useful inside microclimates such as display cases where buffering agents have been used or in places that are remote or difficult to access (Figure 1). Sensors are commonly used to control the environments of buildings housing collections. In the past, data from building management systems have been of little value to the preventive conservator for

Color Image

Figure 1 Moa bird on display monitored by telemetry.

recording the environmental conditions. Generally, these sensors were part of a complex control system, and expensive specialist software was needed to view the values. They also recorded data in formats that were difficult to export. Recording of building management system data can now be easily configured via an internet browser, and, importantly, the data can be retrieved simply and analysed using standard graphing software. High-quality building management system sensors have now been developed that allow simple checking and re-certification by conservators using low-cost certified handheld monitors.

Collection Surveys

The collection survey is an essential component of an environmental strategy and is used to determine the physical state and future needs of the collection (Figure 2). For large collections, statistical sampling methods may be needed to reduce the time and money spent on the survey. A formal documented inspection or survey will highlight problems in a collection. Typically information documented would include damage, surface pollutants, decay, environmental sensitivity, previous conservation treatment, style of storage, and environment. Once an overall picture of the collection has been built, a procedure for bringing the collection up to an acceptable standard can be determined.

Integrated Pest Management

Insect pests can cause major problems for a natural history collection. Although pests rarely damage geological material, the accompanying documentation may be affected. The pesticides that were generally used to stop pest infestations are no longer used because of the risk they pose to health. It is advisable to monitor collections through trapping so that potential infestations can be prevented. In large organizations the approach to preventing pest damage is known as integrated pest management.

Color Image

Figure 2 Survey being carried out on sub fossil bone mammal collection.

Mould

Preventing contamination by fungal spores within any building is impossible. However, mould growth can be prevented by ensuring that the conditions for germination do not arise in the stores or display areas. Uncontrolled indoor environments may experience extreme seasonal changes, allowing humidity levels to rise to the point where germination of spores occurs. If relative humidity is maintained at sufficiently low levels, outbreaks will not occur. Levels of between 50% and 60% relative humidity are considered safe and will not

allow such growth. When an infestation is encountered the steps to take are:

- confine it;
- stop its growth;
- eradicate it; and
- prevent it from reoccurring.

Reduced Oxygen Environments

Reduced oxygen environments are cost-effective low-impact methods of controlling the deterioration of specimens that are sensitive to oxygen, water vapour, or pollutants. Any rock or mineral specimen that is sensitive to oxygen, water vapour, or pollutants (such as rocks containing pyrite that is likely to oxidize or has begun oxidizing) can be stored in reduced oxygen environments to prolong their life. There are three components to a reduced oxygen environment: an oxygen scavenger or oxygen-purging system; an oxygen monitor; and a barrier film. Oxygen scavengers are composed of either iron filings or molecular sieves such as zeolite or mordenite. The optimum composition of a barrier film for this purpose is currently under research. Commercially available films are currently composed of materials chosen to suit either the food or the electronics industries, and, whilst these films may be useful for preventive conservation, they may not be the optimum. Barrier films are layered polymers combined to produce a film with good tear strength, low water migration, and low oxygen migration. The enclosure is made by wrapping the specimen in the barrier film and using a heat sealer to close the film and finish the encapsulation. The specimen should be supported in conservation grade materials to ensure that no physical stresses are imposed on the specimen and no damage is created by the abrasion of the film on its surface (Figure 3).

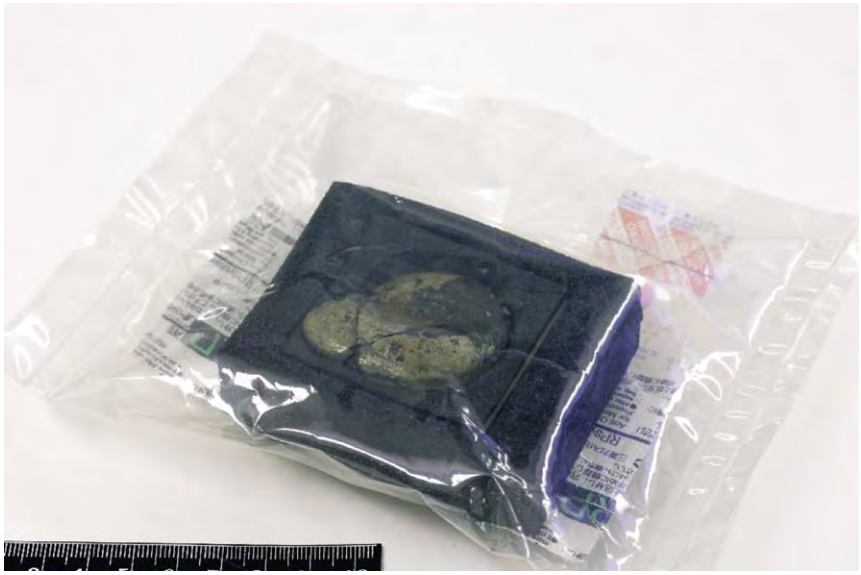
Remedial Conservation

Remedial conservation consists mainly of direct action carried out on an object with the aim of retarding further deterioration.

One of the most important and sometimes controversial stages of conservation treatment is surface cleaning. Irreversible damage can occur if inappropriate treatments are used. However, the removal from the surface of the object of contaminants or old consolidants that may otherwise cause harm is highly advantageous. For example dust can contain acidic particles that can cause surface damage. In some cases cleaning may also clarify an object's detail or reveal unseen damage that can be treated subsequently. Joins that require repair also have to be surface cleaned to improve the final bond. There are many types of surface-cleaning techniques available to the conservator, and they tend to be divided into mechanical cleaning and chemical cleaning.

Surface Cleaning: Mechanical

Abrasive Abrasive cleaning methods include all techniques that physically abrade the fossil surface to remove contaminants or coatings. Such techniques involve the use of materials that impact or abrade the surface under pressure, or abrasive tools and equipment. The use of water in combination with abrasive powder may also be classified as an abrasive cleaning method. Depending on the manner in which



Color Image

Figure 3 Pyritised ammonite encapsulated in barrier film in an oxygen free environment.

it is applied, water may soften the impact of the powder, but water that is too highly pressurized can be very abrasive.

Steam cleaning Steam, heat, and pressure provide the means for the immediate removal of particles from a given surface, cleaning it thoroughly. This technique is usually carried out using of a handheld steam pencil and is especially useful for removing old adhesives and labels.

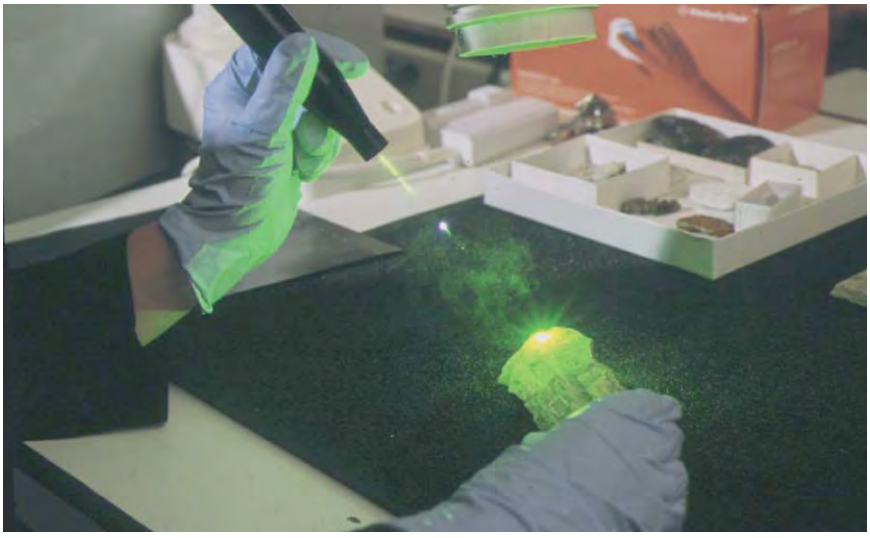
Ultrasonic Ultrasound is used in the cleaning of material because of its vibration rates. Large acoustic forces break off particles and contaminants from surfaces. This cleaning method is generally carried out either by immersing the specimen in a tank or by using a handheld pen. When the latter is used, its vibrating tip contacts the surface and fragmentation occurs. When fossils are immersed in a tank of liquid (usually water and detergent), sound waves from the transducer radiate through the tank, causing alternating high and low pressures in the solution. During the low-pressure stage, millions of microscopic bubbles form and grow. This process is called cavitation, meaning 'formation of cavities'. During the high-pressure stage, the bubbles collapse or implode, releasing enormous amounts of energy. These implosions work in all directions, on all surfaces.

Laser cleaning Surface cleaning by laser is a recent innovation in the treatment of geological material and presents several advantages when compared with standard cleaning methods. The laser offers the potential for selective and controlled removal of surface contaminants with a limited risk of damaging the underlying substrate. When the laser beam meets a boundary between two media, e.g. dirt and the sample surface, a proportion of the beam's energy is reflected, part is absorbed, and the rest is transmitted (Figure 4). The proportion of energy absorbed is determined by the wavelength of the radiation (primary source) and the chemical structure of the material. In order to remove dirt or another unwanted surface coating from an object, it is important that the dirt or coating absorbs energy much more strongly than the underlying object at the selected laser wavelength. If this is the case then, once the dirt layer has been removed, further pulses from the laser are reflected and the cleaned surface is left undamaged. The technique, therefore, is – to a certain degree – self limiting. Cleaning efficiency can also be enhanced by brushing a thin film of water onto the surface immediately prior to laser irradiation. The laser also effectively removes conducting coatings, for example gold or palladium, from microfossils; these have previously been almost impossible to remove.

Surface Cleaning: Chemical

Surface cleaning by using chemicals should be considered carefully so that the integrity of the object is not compromised. A disadvantage of using chemicals is the inability to control the chemical movement within an object, which can lead to over cleaning and the unwanted removal of fossil material. Usually a chemical-cleaning testing regime should be carried out, commencing with the least aggressive solvent. A solvent is a liquid capable of dissolving other substances.

Care should be taken to ensure that the physical and chemical properties of the fossil and surrounding matrix (where present) are taken into account so that



Color Image

Figure 4 Laser cleaning of fossil reptile (indet) and surrounding matrix. Visible green wavelength (533 nm) being used.

inappropriate use is avoided. For example, water is non-hazardous to the conservator and the strong polarity of its molecules means that it is effective in removing a wide range of pollutants. However, this apparently harmless material would have severe detrimental effects on humidity-sensitive pyritized fossils, causing them to deteriorate.

Poulticing is used to avoid deep penetration and to limit the action of chemicals on the object. Chemicals are mixed with absorbent powders to form a paste or poultice, which is applied to the surface. As the poultice dries, dirt is drawn out and has to be physically removed. A development of this technique is the use of solvent gels. These cleaning systems have an aqueous gel base composed of a polymer resin that thickens on the addition of water, and a surfactant – also a thickening agent – which improves the gel's contact with the surface to be cleaned. Any number of cleaning agents can be added to this gel base. Of particular concern to the conservator are the possible long-term effects on surfaces. The most pressing concern has been whether any residue of the gels is left on the treated surface that may cause future damage.

Consolidants, Adhesives, and Gap Fillers

The consolidants, adhesives, and gap fillers used at any given time reflect the existing knowledge of chemistry and technology; so, naturally, their diversity and quality change over time. From the viewpoint of the conservator, all should be reversible. Geological material, whether it is in scientific collections or in a private collection, must be considered valuable and irreplaceable. As such, conservationally sound and approved materials should be used wherever possible.

Consolidants When considering whether to use a consolidant on a specimen, it is important to remember that not all specimens require consolidation. The most important axiom of conservation is: minimal intervention is best.

A consolidant is a liquid solution of a resin (normally a synthetic polymer) that is used to impregnate a fragile object in order to strengthen its structure. Common solvents for the resins are water, acetone, alcohol, and toluene. Consolidants are generally available in two forms: pure resins and emulsions. Pure resins are mixed with their solvents to form a very thin watery solution, which is then applied to the specimen (or the specimen is immersed in the solution). The aim is to get the solution to penetrate the specimen's surface and carry resin down into the interior of the fossil bone: the consolidant must be thin otherwise it will be deposited on the surface of the bone only, like the shellac or varnish used in the past. Those treatments may have protected the surface, but did little to strengthen the whole bone.

The second class of consolidants, the emulsions, are mainly used to treat wet or moist specimens. Emulsions are suspensions, in water, of a resin and solvent solution, popularly polyvinyl acetate emulsion, and are generally white milky mixtures. Emulsions are not as desirable as pure resins. It is hard to reverse emulsions once they have dried and virtually impossible once they have crosslinked on exposure to ultraviolet light from the sun or from fluorescent bulbs. Emulsions also tend to turn yellow with age and increasing crosslinking.

Adhesives and gap fillers Historically, the most common adhesives have been animal glues made from bones, fishes, and hides, and these were used extensively for fossil repair and can still be purchased today. Owing to their inherent problems, such as yellowing, brittleness, and instability, they are no longer used on fossils. Consisting mainly of collagen–protein slurries, animal glues are also quite attractive to a variety of pests. This class of adhesives is mentioned here because of its long period of use. In museums and at fossil auctions, it is not uncommon to find specimens that have been repaired with these glues.

The twentieth century has given us many new classes of adhesives, all of which are organic polymers – large complex molecules formed from chains of simpler molecules called monomers. Manufacturers may extol the virtues of the newest adhesive from their laboratories, but only time can judge the effectiveness and longevity of an adhesive. Many turn yellow with age or are prone to brittle breakage, where even a slight jar or shock will cause the glued joint to break. Other polymers may crosslink with time or upon prolonged exposure to ultraviolet light, causing shrinkage that can seriously damage a fossil that has been repaired with them. Most of these polymers have unique properties and characteristics, which make some better for certain uses than others. Another class of recently developed adhesives used to repair fossils are the 'superglues' or cyanoacrylates. Their characteristics, which include rapid setting and strength, have made these adhesives increasingly popular for fossil repair; however, since they are so new (dating only from the 1980s), our knowledge of their long-term efficacy is limited. A major drawback is the difficulty of reversing bonds made with cyanoacrylates.

Adhesives are sometimes mixed with materials, e.g. glass micro balloons, to form a gap filler. Gap filling is

a process whereby a sympathetic replacement material is used to fill or bridge small or large gaps. The gap filler should not conceal original material and should be easy to remove (Figure 5).

Conservation of Sensitive Geological Material

Subfossil Bone

Subfossilized bone retains a fair amount of organic material (collagen) and original mineralized bone (hydroxyapatite). To maintain its integrity, subfossil bone needs to be kept in a stable environment. Once it is removed from the matrix in which it is found, it is likely to deteriorate rapidly, especially if it is wet and the bone is subjected to extreme fluctuations of temperature and humidity.

Color Image

Figure 5 Neanderthal skull (Tabun) showing gap filler which aids in its reconstruction.

Recently excavated material that may still be wet can be treated in two ways. The first method is the use of water-based consolidants, which can be applied by brushing, immersing, spraying, or vacuum impregnation. The second method is slow controlled drying to prevent cracking and delamination. The aim of controlling the drying procedure is to reduce the high relative humidity of the wet specimen slowly until it matches that of the storage area. Even if a water-based consolidant is used, it is advisable to apply controlled drying procedures until the specimen is stabilized.

Smaller specimens can be placed in plastic containers, and for larger specimens a chamber or tent can be constructed, for example by placing a piece of clear plastic sheeting over laboratory scaffolding. Barrier films can also be used to control drying stages. There needs to be a controlled exchange of the moist air inside with the lower-relative-humidity ambient air outside the box. To monitor changes in relative humidity, a humidity gauge can be placed in the containment area along with the specimen. Potential mould growth needs to be monitored, and placing a fungicide inside the containment area should control the problem. The optimum storage environment for subfossil bone has a relative humidity of 45–55% and a temperature between 15°C and 21°C. Low relative humidity can lead to cracking and shrinking as the specimen dries out, and high relative humidity (above 70%) encourages damaging mould and fungal growth.

Once the specimen has been dried and stabilized, an organic-solvent-based consolidant can be applied.

Color Image

Figure 6 *Dacosuarus maximums* showing severe cracking in response to fluctuating humidity conditions.

It is not advisable to apply water-based consolidants to specimens that have thoroughly dried because the high water content of the consolidant will cause the dry specimen to swell and crack multidimensionally.

If gap fillers are to be applied to repair the bone, particular attention should be given to the choice of filler. Materials that shrink or expand upon curing should be avoided as either action can damage the specimen.

Shale and Other Fine-Grained Sediments

Deterioration in the form of delamination, cracking, and shrinkage is common for shale and other fine-grained sediments, especially if they are stored in the wrong environment. The best approach is therefore to ensure that storage conditions are optimum, with a temperature of below 20°C and humidity level of around 50%.

Pyrite

The deterioration and even complete decomposition of pyritized fossils through oxidation is a common problem (Figure 6). Deterioration is best prevented by storing at less than 45% relative humidity. Once deterioration has occurred various conservation treatments are available, which have varying levels of success. For example, ethanolamine thioglycollate or ammonia gas treatment may be used. The latter is considered more conservationally sound.

Documentation

Existing documentation associated with individual geological specimens should be preserved as a matter of priority; without it the material is scientifically useless. Any treatments should be recorded accurately (with an image if possible) so that a history of treatment can be built up over time. This information will influence future conservation considerations.

See Also

Fake Fossils. **Minerals:** Definition and Classification; Sulphides; Zeolites. **Palaeontology**. **Rocks and Their Classification**.

Further Reading

Casaar M (1999) *Environmental Management Guidelines for Museums and Galleries*. London and New York: Routledge.

Cornish L and Doyle AM (1984) Use of ethanolamine thioglycollate in the conservation of pyritized fossils. *Palaeontology* 27: 421 424.

Cornish L and Jones CG (2003) Laser cleaning natural history specimens and subsequent SEM examination. Chapter 16 In: Townsend J, Eremin K, and Adriaens A (eds.) *Conservation Science 2002*, pp. 101 106. London: Archetype.

Cornish L, Doyle AM, and Swannell J (1995) The Gallery 30 Project: conservation of a collection of fossil marine reptiles. *The Conservator* 19: 20 28.

Croucher R and Woolley AR (1982) *Fossils, Minerals and Rocks Collection and Preservation*. London: Cambridge University Press.

Gilroy D and Godfrey I (1998) *A Practical Guide to the Conservation and Care of Collections*. Perth: Western Australian Museum.

Howie FM (ed.) (1992) *The Care and Conservation of Geological Material: Minerals, Rock, Meteorite and Lunar Finds*. Oxford: Butterworth Heinemann.

Institute of Paper Conservation. www.ipc.org.uk.

Larkin NR, Makridou E, and Comerford GM (1998) Plastic storage containers: a comparison. *The Conservator* 22: 81 87.

Museums, libraries and archives council. http://www.mla.gov.uk/index.asp.

Resource UK Council for Museums, Archives,Libraries (1993) *Standards in the Museum Care of Geological Collections*. London: Resource UK Council for Museums, Archives and Libraries.

Waller R (1987) *An Experimental Ammonia Gas Treatment Method for Oxidised Pyritic Mineral Specimens*, pp. 625 630. Triennial Report. Rome: ICOM Committee for Conservation.

CREATIONISM

E Scott, National Center for Science Education, Berkeley, CA, USA

Definitions

Although evolution is a component of many scientific fields, it generally has the same meaning across disciplines: cumulative change through time. The topic of this article, 'creationism', is a term with many definitions. In theology, creationism is the doctrine that God creates new souls for each person. In anthropology, creationism is the well-confirmed thesis that almost all human societies have origin stories about the acts of gods, or a God, or powerful spirits of some kind.

Viewed socially and politically, creationism refers to a number of twentieth-century religiously-based anti-evolution movements originating in the USA, but now spreading to many countries. The most familiar of these (and the movement to which the term is most frequently applied) is 'creation science', an attempt to demonstrate with scientific data and theory the theological view known as special creationism. According to special creationism, God created the universe – stars, galaxies, Earth, and living things – in essentially their present forms. Living things were created as 'kinds' that do not have a genealogical (evolutionary) relationship to one another. This biblical literalist theology views Genesis narratives, such as the creation of Adam and Eve, their sin and expulsion from Eden, and the Flood of Noah, as historical events. In creation science and its ancestor, 'Flood Geology', the Flood of Noah has shaped most of the Earth's geology. A recent creationist movement, Intelligent Design Theory, pays little attention to the Flood of Noah or to geology, or to fact claims of any sort, contenting itself with proclaiming God's intermittent creation of supposedly 'irreducibly complex' biochemical structures, such as the bacterial flagellum or the blood clotting cascade, rather than presenting a scientific alternative to evolution. A non-Christian creationism is promoted by the Krishna Consciousness movement, whose members agree with geologists about the age of the geological column and how it was shaped, but who argue that human artefacts are found from the Precambrian on, thus supporting a literal interpretation of the Vedas that humans have existed for billions of years.

Christians who opposed evolution during Darwin's time rarely referred to themselves as 'creationists'; they used the term 'creationism' generically to refer to the idea that God purposefully creates living things, in contrast with Darwin's naturalistic explanation for the appearance of humans and other creatures (*see* **Famous Geologists:** Darwin). Nineteenth-century clergy and scientists could choose from many models of creation beyond the Biblical literalist six 24-h days. Charles Lyell (*see* **Famous Geologists:** Lyell) proposed that God had created animals adapted to 'centres of creation' around the world; Cuvier (*see* **Famous Geologists:** Cuvier) proposed a series of geological catastrophes followed by a series of creations. Theologically, the 6 days of Genesis Creation could be interpreted as very long periods of time (the 'day-age' theory), or Genesis could be read as permitting a long period of time between the first and second verses (the 'gap theory'). Some doubtless clung to a literal Genesis of six 24-h days and a historical, universal Flood, but this view was not common amongst university-educated scientists or clergy.

The evolution of creationism and its relationship to geology are the subjects of this article, and thus, befitting an evolutionary approach, we begin with a historical perspective.

Static versus Dynamic Views of the Earth

Throughout much of the early European scientific period (1600–1700), two perspectives of the world competed: either it had remained unchanged since the special Creation described in Genesis, or it was changing now and had changed in the past. The shift from a static to a dynamic view of nature was stimulated by European exploration during the 1500–1700s. During these expeditions, vast amounts of natural history, including geology, were learned by travellers and settlers, and the new information proved to be difficult to fit into a biblical literalist framework.

The remains of molluscs and other sea creatures on mountaintops, found in the same groupings as living shellfish, encouraged da Vinci to question a literal Flood; he argued that the Flood would have mixed up the shells, not deposited them in life-like settings. Biological data also did not fit into the view of a static world: new species were discovered in the new lands that were not mentioned in the Bible, and geologists found remains of extinct species, troubling for a theology assuming a perfect Creation. Biogeography also made a literal Flood story problematic: how did marsupials in Australia and South America get there after the Ark landed on Ararat? Old views of a static Creation, unchanged since God rested on the seventh day, gradually gave way to an appreciation of an evolving world and, eventually, of the evolution of living things.

Geology is an evolutionary science, dealing as it does with cumulative change in the history of the planet. Geology came into its own as a scientific discipline during the 1700s, as more was learned about the geological characteristics of the planet, prompting speculations about the processes and mechanisms that produced them. The fruits of fieldwork and careful mapping illuminated such processes as sedimentation, erosion, volcanoes and earthquakes, mountain building, and the like, and it made sense that these processes had also operated in the past, changing the contours of the planet. The increased understanding of the Earth and the forces that produced its landforms led to the inevitable conclusion that the Earth was ancient.

An ancient Earth, however, conflicted with traditional scriptural interpretation that the Earth was

young; Archbishop Ussher's calculations of a 6000-year-old Earth were endorsed in the margins of commonly used Bibles. However, believing that the Book of Nature and the Book of God must be telling the same story, Christian geologists worked to find ways to interpret the geological data in a framework that allowed them to retain at least some of the Genesis story of Creation, whilst accepting the empirical evidence of their new science.

One accommodation was that of the English clergyman, Thomas Chalmers, who attempted to harmonize geology with the Bible through the 'gap theory'. In his 1815 book, *Evidence and Authority of the Christian Revelation*, Chalmers argued that there was a temporal gap between Chapters 1 and 2 of Genesis. This preserved a literal 6-day Creation, but placed it after a long pre-Adamite creation. The evidence of the rocks that the Earth was ancient was thus acknowledged without the wholesale abandonment of Genesis. The gap or 'ruin and restoration' compromise also had the advantage of allowing for a relatively recent Creation, which pleased religious conservatives. The gap theory remained popular through the nineteenth century and well into the twentieth.

Another compromise between the Bible and science was offered by Scottish stonemason Hugh Miller. In his popular 1847 book, *Footprints of the Creator*, he proposed the 'day-age theory', in which the 6 days of Genesis were not 24-h days, but long periods of time. This compromise was even more scientifically flexible than Chalmers' gap theory: it allowed for the acceptance of virtually all the geological data by requiring that Genesis be taken more figuratively.

Other religious views were more concerned with Darwinian evolution than geology. 'Progressive creationism' accepted the sequence of fossils in the geological column: God was believed to have created increasingly more advanced forms through time. The doctrine of 'theistic evolution', in which God was thought to use evolution and natural selection to bring about the current variety of living things, similarly had little impact on geology. During the early- and mid-nineteenth century, Christian geologists worked to harmonize the 'two books' and, working with professional clergy, eased the worries of Christians that modifications in Ussher's view of a 6000-year-old Earth would create irreparable rents in the fabric of Christianity.

Not all Christians agreed, however. Some felt that these compromises were unacceptable because they required the Bible to be 'interpreted' rather than being read at face value. What came to be called 'Scriptural Geology' took the position that, when the 'two books' were in apparent conflict, the book of God's word was to be given priority. The proponents of Scriptural Geology tended to be neither trained geologists nor university-educated clergy; rather, they were self-taught, educated laymen. Most of them promoted a young Earth and a historical Flood. At best Scriptural Geology was a rearguard movement that had little effect on the views of professional science. As will be discussed later, however, some of the same challenges to evolution promoted by the Scriptural Geologists reappeared in the twentieth century in the guise of 'creation science'.

By the end of the nineteenth century, practising scientists and professional clergy in both the USA and on the continent accepted an ancient age of the Earth and rejected the Flood of Noah as a universal historical event that shaped the planet's landforms. Virtually all scientists likewise accepted biological evolution, although not necessarily Darwin's mechanism of natural selection. In the USA, however, evolution was about to be attacked with the emergence of a religious position called 'Fundamentalism' and, even more importantly, by the efforts of a dedicated amateur geologist and, later, a hydraulic engineer.

Young Earth Creationism and Flood Geology

The Fundamentalist movement in American Protestantism began with a series of small booklets, collectively called *The Fundamentals*, published between 1910 and 1915. Fundamentalism was partly a reaction to a theological movement called Modernism that began in Germany in the 1880s. Modernism included viewing the Bible in its cultural, historical, and even literary contexts; biblical Creation and Flood stories, for example, were shown by comparison of ancient texts to have been influenced by similar stories from earlier non-Hebrew religions. With such interpretations, the Bible could be viewed as a product of human agency – with all that that suggests for the possibilities of error, misunderstanding, and contradiction – as well as a product of divine inspiration. Fundamentalists in response stressed divine inspiration and absolute accuracy of scripture, including Biblical miracles such as Noah's Flood.

Millions of copies of *The Fundamentals* booklets were printed and distributed. Most of the authors of *The Fundamentals* were day-age creationists, allowing for an old Earth, but insisting on a recent appearance of humans. Although not all *The Fundamentals* booklets were anti-evolutionary, the Fundamentalist position hardened against evolution fairly quickly. Fundamentalists were motivated by religious sentiments – if evolution were true, then what of the accuracy of the Bible? – and also a concern that evolution was the source of many corrosive

social practices, such as child labour, '*laissez-faire*' capitalism, and exploitation of workers and immigrants. The erroneous association of evolution with World War I German militarism further tainted evolution in the eyes of many.

Fundamentalism remained an almost exclusively American religious movement, not attracting much interest in Great Britain or Europe. However, the Bible-based religious view proved to be a fertile ground for anti-evolutionism, and science was pressed into service by a number of Fundamentalist evangelists to support the cause. Harry Rimmer, Arthur I. Brown, and others hammered evolution before large public audiences, and anti-evolution tracts and books sold well. The target audience of these evangelists was the general public, not the scientific community; the latter remained staunchly evolutionist, and professional scientists were scarce indeed among the anti-evolutionists associated with the growing Fundamentalist movement.

Into this environment came a Seventh-Day Adventist and self-trained geologist named George McCready Price, who was to develop the first version of the most influential creationist perspective of the twentieth century. Seventh-Day Adventist prophetess, Ellen G. White, had claimed visions of a 6-day Creation and a universal Flood; Price sought to scientifically validate her views. He recognized that geology was the key to disproving evolution because, if it could be demonstrated that the Earth was young, then biological evolution would be impossible. He therefore concentrated on attacking conventional interpretations of the geological record. For decades, he vigorously argued that fossils could not be used to determine the order of geological strata, because there were instances where fossils were 'in the wrong order'. He accused geologists of trying to 'shore up their theory' of evolution by proposing 'highly unlikely' theories, such as overthrusting. A favourite example was the Lewis Overthrust. Extending from central Alberta to the eastern edge of Glacier National Park in Montana, this massive structure is composed of Precambrian limestone overlying Cretaceous shales. Price argued that a limestone slab of such immense size could not have been positioned by natural forces. In general, Price ascribed all geological features to the 6 days of Creation, the period between Creation and the Flood, or the Flood itself, and called his view Flood Geology.

Price tried to distinguish his 'new catastrophism' from earlier catastrophic geological views, such as those proposed by Cuvier. He worked tirelessly to promote Flood Geology, and was given a boost when William Jennings Bryan referred to him as a noted scientist during the Scopes Trial. By the 1940s, Price's views had a wide influence on evangelical Christian views of creation and evolution. His major rivals within that community were the gap and day-age theories, but some evangelicals also believed in progressive creationism or theistic evolution. Price rejected all of these as being insufficiently biblical, and clung to the literalism of a young Earth and a global Flood.

Seventh-Day Adventists were considered to be theologically suspect by many conservative Christians. As a result, Price did not always receive the credit his pioneering work deserved. By the 1950s, Flood Geology had begun to slip out of favour with evangelicals, especially with the 1954 publication of Bernard Ramm's *The Christian View of Science and Scripture*, which promoted progressive creationism rather than the 6-day Creation and a universal Flood. The popularity of Ramm's book generated a backlash from more conservative evangelicals. Theologian John C. Whitcomb and hydraulic engineer Henry M. Morris presented the basics of Price's Flood Geology – without mentioning Price – in their 1961 book, *The Genesis Flood*, which proposed a universal Flood, Flood Geology, biblical literalist Genesis theology, and a young Earth as the only acceptable scientific and religious position for a true Christian.

In some respects, *The Genesis Flood* was even more conservative in its presentation of a purportedly scientific foundation for Genesis than were Price's publications. Whitcomb and Morris argued that the entire universe was created between 5000 and 7000 years before the present, whereas Price argued that only the solar system, the Earth, and living things were created during the Adam and Eve creation. Whitcomb and Morris revived the vapour canopy theory, an earlier view that, on the second day of creation, the Earth was shrouded in a canopy of water vapour. This provided the source of the 'waters above the firmament' that produced the forty days and nights of rain at the beginning of the Flood, and also produced a greenhouse effect enhancing the 'garden' of Eden. By shielding the surface of the planet from ionizing radiation, the vapour canopy also reduced cellular damage, allowing Methuselah and his contemporaries to have lifespans of hundreds of years. Consonant with creationists' objections to radiometric dating, the vapour canopy was also claimed to alter the ratios of ^{14}C in the atmosphere, making age estimations based on ^{14}C inaccurate. *The Genesis Flood* also presented the familiar young Earth model with the Flood as the source of sedimentary features around the planet, and reprised Price's idea of the fossil record being the result of 'hydrodynamic sorting'. Regular and streamlined animals would be naturally sorted by the Flood waters into lower strata; irregular (and more intelligent) fauna

would either temporarily escape to higher ground or would float there to be interred at higher elevations.

The book seized the attention of the evangelical community in a way that the Adventist Price's books did not. It sold tens of thousands of copies during its first decade, and remains in print today. The nineteenth century effort to support the literal 6-day Genesis Creation and a young Earth through scientific data and theory was finally institutionalized when Morris and his colleagues organized the Creation Research Society in 1963 as an association of conservative Christian scientists promoting young Earth creationism. In 1972, Morris organized the California-based Institute for Creation Research (ICR) to promote 'creation science'. Its staff and publications expanded; within a decade Morris could claim an ICR publication list of 55 books. The current book list numbers in the hundreds, and these and other creationist books have been translated into Afrikaans, Chinese, Czech, French, German, Hungarian, Italian, Japanese, Polish, Portuguese, Romanian, Russian, and Spanish. The ICR and other creation science ministries, such as Answers in Genesis, have been very effective in shaping the misunderstanding – and rejection – of evolution by millions of conservative Christians. As a result, the word 'creationism' itself is now usually understood to mean young Earth creation science.

Creation Science and Geology

Proponents of creation science argue that there are only two possible views: creationism and evolution; thus arguments against evolution are arguments in favour of creationism. Creation science literature thus centres on alleged examples of 'evidence against evolution', such as anomalies in scientific literature, that are then used to argue that the evidence for evolution is deficient. Echoing Price's contention that the geological column is not reliable, creation science proponents deny that missing strata at paraconformities ever existed: instead, they are evidence for rapid deposition by the Flood. Proponents also point to fossils that are out of order, such as an alleged human footprint occurring on top of a trilobite, or 'Moab man', a human skeleton found in the Morrison Formation (Late Jurassic) in Utah. Despite the report of the physical anthropologist who excavated the Moab skeleton that it was an intrusive burial and a ^{14}C date of 300 years, creationists continue to promote the burial as evidence of a distorted and misleading geological column.

The most famous 'out of place' fossils, however, are alleged human footprints occurring with dinosaur footprints, claimed to have been found in Pennsylvania, New Mexico, and Texas, the last being the best known. The Cretaceous limestones of the Paluxy River near Glen Rose, Texas, present many well-preserved tridactyl dinosaur footprints. Creationists have claimed since the 1930s that human footprints are found in the same strata, proving that dinosaurs and humans existed at the same time. Analysis of these footprints by palaeontologists and geologists has shown that none of the claimed 'mantracks' is valid: all can be explained as natural erosional features, eroded dinosaur tracks, or carvings. Although the ICR has retreated slightly from promoting the Paluxy River 'mantracks', other creationists continue to do so, as can be seen by even a cursory web search.

Like Price, modern creation science proponents know that the age of the Earth is a critical issue for evolution; without an old Earth, biological evolution could not have occurred. A great deal of creationist literature, therefore, focuses on efforts to demonstrate that the Earth is young rather than ancient. For example, when presented with evidence from ice cores, varves, or even tree rings, creationists argue that these annual indicators of age are flawed, because more than one layer, varve, or ring can be formed in a year, thus throwing off the counts.

Many of these young Earth arguments rely on extrapolations of various rates of change. When a measured rate is extrapolated backwards (or forwards), a result is produced that is incompatible with the accepted age of the Earth. Engineer Walter T. Brown's book, *In the Beginning*, includes numbered 'evidences' against evolution, including this one on the decay of the Earth's magnetic field:

> 84. Over the past 140 years, direct measurements of Earth's magnetic field show its steady and rapid decline in strength. This decay pattern is consistent with the theoretical view that a decaying electrical current inside Earth produces the magnetic field. If this is correct, then just 20 000 years ago the electrical current would have been so vast that Earth's structure could not have sur vived the heat produced. This implies Earth could not be older than 20 000 years.

Of course, the rate of decay of the Earth's magnetic field is not linear, but periodically reverses, and thus the fear of a too-powerful field 20 000 years ago is groundless. Other rate arguments similarly rely on taking a known rate and extrapolating it in a linear fashion, resulting in conclusions at variance with modern geology:

> 80. The continents are eroding at a rate that would level them in much less than 25 million years. However, evo lutionists believe fossils of animals and plants at high elevations have somehow avoided this erosion for more than 300 million years. Something is wrong.

83. Meteoritic dust is accumulating on Earth so fast that, after four billion years, the equivalent of more than 16 feet of this dust should have accumulated. Because this dust is high in nickel, Earth's crust should have an abun dance of nickel. No such concentration has been found on land or in the oceans. Therefore, Earth appears to be young.

Creation science proponents argue that, after the Flood, a strict uniformitarianism has held, but, during the Flood, God used processes that are different from those now in effect. This selective catastrophism is necessary to avoid the implications of standard uniformitarian interpretations, which are overwhelming evidence for a slowly changing planet and thus an ancient Earth. The young Earth model demands accelerated rates of change well beyond any observed today. Coral reefs would have needed to accumulate at 40 000 times their current observed rate to form as quickly as a Flood Geology model requires. The interior of the USA has sedimentary deposits of great thickness, averaging a kilometre in depth, some of which, such as calcareous muds, are the results of animal activity; these too would have needed to form at mind-boggling rates. In the Grand Canyon, the Kaibab and Redwall limestones together comprise about 300 m of sediment; to produce deposits of this thickness during the year of the Flood would have required deposition of 80 cm per day – clearly more than has ever been observed or even imagined.

As part of the effort to attack the age of the Earth, creation science proponents hold that radiometric dating methods are unreliable. The Grand Canyon Dating Project, a project of the ICR, recently attempted to demonstrate this by showing alleged discrepancies between dates of Grand Canyon lava flows collected by ICR staff and by others. In the ICR study, Pleistocene lava flows were dated earlier than the Proterozoic Cardenas basalts, but other geologists have pointed out that the technique used by the ICR (Rb/Sr) may give false isochrons by reflecting the age of the mantle source from which the lava flows were derived, rather than the eruption time of the lava itself.

A more recent ICR research project critiquing radiometric dating is the Radioisotopes and the Age of the Earth (RATE) project, begun in 1997. In 2000, Vardiman *et al*. published a 675-page book describing the project and its expectations for research. A final volume of conclusions from the project is expected in 2005.

The fact that ICR scientists have undertaken research distinguishes them from their nineteenth-century Scriptural Geology predecessors. The amount of research performed to date, however, is very small, and its quality has been questioned by professional geologists. Most of the literature consists of the selective search of scientific data for anomalies cited to disprove evolution, and the little actual research that has been performed, such as efforts to ^{14}C date dinosaur bones or the Grand Canyon Dating Project, has not inspired confidence in the quality of creationist research efforts. As a result, the professional geological community has largely ignored creation science, except as a social movement that has had remarkable success in popularizing poor science to the general public.

See Also

Analytical Methods: Geochronological Techniques. **Evolution**. **Famous Geologists:** Cuvier; Darwin; Lyell. **History of Geology From 1780 To 1835**. **History of Geology From 1835 To 1900**. **History of Geology From 1900 To 1962**. **Trace Fossils**.

Further Reading

Cole JR, Godfrey LR, Hastings RJ, and Schafersman SD (eds.) (1985) The Paluxy River Footprint Mystery Solved! *Creation/Evolution*, 5(1), pp 1 56.

Forrest B and Gross PR (2004) *Creationism's Trojan Horse, The Wedge of Intelligent Design*. Oxford: Oxford University Press.

Larson EJ (1997) *Summer for the Gods: The Scopes Trial and America's Continuing Debate Over Science and Religion*. New York: Basic Books.

Lindberg DL and Numbers RL (eds.) (1986) *God and Nature: Historical Essays on the Encounter Between Christianity and Science*. Berkeley, CA: University of California Press.

Livingstone DN, Hart DG, and Noll MA (eds.) (1999) *Evangelicals and Science in Historical Perspective*. New York, NY: Oxford University Press.

Lynch JM (ed.) (2002) *Creationism and Scriptural Geology, 1817 1857*. London: Bristol Thoemmes Press.

Numbers R (1992) *The Creationists*. New York: Knopf.

Scott EC (2004) *Creationism vs Evolution: An Introduc tion*. Westport, CT: Greenwood Publishing Group.

Stassen C (2003) *A Criticism of the ICR's Grand Canyon Dating Project*. Retrieved December 29, 2003, from http://www.talkorigins.org/faqs/icr science.html.

Strahler A (1999) *Science and Earth History, The Evolu tion/Creation Controversy* (revised edn.). Buffalo, NY: Prometheus Books.

Vardiman L, Snelling AA, and Chaffin EF (eds.) (2000) *Radioisotopes and the Age of the Earth: A Young Earth Creationist Research Initiative*. El Cajon, CA, and St. Joseph, MO: Institute for Creation Research and Creation Research Society.

Wise DU (1998) Creationism's geological time scale. *American Scientist* 86: 160 173.

DELTAS

See **SEDIMENTARY ENVIRONMENTS: Deltas**

DENDROCHRONOLOGY

M Bridge, University College London, London, UK

Introduction

In recent decades tree-ring studies have opened up a whole new area of climatic and other environmental reconstructions at various geographical and temporal scales. These generally go under the heading of dendrochronology – although some people reserve this term only for the use of tree rings in dating studies. Dendrochronological dating relies on the fact that whilst each tree-ring series reflects peculiarities of the life history of the individual tree, trees of the same species growing at the same time over wide areas respond similarly to the weather experienced both during an individual growth period, and perhaps also to conditioning throughout previous seasons.

Not all species have clearly distinct ring boundaries, and those without them are of little use to dendrochronology. Some species not only do have clear rings, but may have clear divisions within the rings, such that the ring can be divided into the 'earlywood' which forms during the spring and early summer, and the 'latewood' which forms through the summer and early autumn. Whilst many studies make use of the whole ring, in recent years many have focused on part of the ring, usually the latewood, which is usually more independent of the influences of previous growth in its variations. Most dendrochronological work has been done in temperate zones, although some tropical trees have been used.

With historical timber, samples of wood from naturally preserved timbers e.g. in river gravels, peat bogs and similar environments, usually take the form of cross-sections cut from trunk remains. Historical wood in situations such as standing buildings is generally cored down the radius of the trunk. Many of the studies outlined below, however, depend on samples from living, or recently felled, trees. Living trees can be sampled by extracting radial cores, usually of the order of 5 mm in diameter, taken using especially designed manually operated borers. The tree 'compartmentalizes' the wound, sealing it off from the rest of the living tissue, and any damage to the tree is generally minimal.

During the late 1960s and the 1970s, the development of radiocarbon dating and the realization that the formation of radiocarbon in the atmosphere had not been constant through time were the stimuli for several laboratories to attempt to build long (multimillennial) tree-ring chronologies. These provided wood samples of known calendar age with which to investigate changes in radiocarbon levels through time, and then to calibrate the radiocarbon time-scale. This has made possible 'wiggle matching', which in suitable circumstances can provide more accurate radiocarbon dates.

As the chronologies were being constructed it was noticed that some periods had scarce wood remains which was thought to reflect environmental changes, either in the vigour of tree growth, or in the conditions that favour the preservation of the wood. Once several long chronologies became available, it was noticed that certain periods showed common growth responses over wide geographical areas. For example, a downturn in oak growth lasting about a decade after 1628 BC was reported in Irish bog-oaks. This was subsequently found to occur at the same time in oaks growing in northern England and northern Germany.

This decade showed the lowest growth rates for many centuries in each locality, and must therefore reflect a large climatic forcing agent. It was suggested that the most likely candidate for this particular period of growth reduction was the Santorini (Thera) volcanic eruption, often linked to changes in the Minoan civilization. Accepting this as dating that particular event was controversial, but even if some were unable to accept this hypothesis, some major factor had to be responsible for the observed response.

Throughout the rapid growth of dendrochronology in the late twentieth century, many studies focused

on dating discrete events such as volcanic eruptions, earthquakes, landslides, avalanches, forest fires, floods, severe frosts etc. through history, giving important background information about the frequency and intensity of these phenomena, and in many instances giving the first calendar dates for them. These events may be represented in the tree-ring series by sudden growth rate changes, scars, or abnormal cells.

Longer-term variations and rates of change for such phenomena as river-flow changes, lake-level changes, saltwater ingression, changes in ocean currents through time (mapped by dating driftwood), glacial advance and retreat (reflected in subfossil trunks from valley sides), erosion rates, human forest clearance and the like have all been subjects for dendrochronological study. Periods of building of prehistoric lakeside settlements have been found to be coincident over areas of central Europe, perhaps reflecting migration patterns in response to climatic changes.

Dendroclimatology

It was not long before people realized that well replicated chronologies were themselves a proxy dataset of the major climatic influences on growth, and people looked for the best methods of extracting this information at annual, decadal, century and millennial time-scales. In some areas the relationships between climatic factors and growth are relatively simple, and the dominant limiting factor to growth exhibits a clear correlation with the width of the annual ring. In semi-arid areas for example, the ring width is generally a reflection of rainfall levels.

Some basic environmental information can be gained simply by looking at long-term trends in the ring width itself. For example, different centuries show quite different growth rates for similarly biologically aged oaks preserved in river gravels in Germany, used to construct a multimillennial chronology. Even without any complex calibration of the climate–growth relationship, it is possible to draw broad inferences about changing conditions at particular times, with narrow mean ring widths representing poorer growth conditions and wider mean ring widths more favourable growth conditions.

The sensitivity of trees to external changes in their environment changes with the site conditions. In more extreme environments it is possible that rings may be missing, or partially missing around the circumference, or if unfavourable conditions are experienced during the normal growth period, a slowing down and then resurgence of growth may produce apparent 'false' rings. These problems are generally detected when 'cross-matching' the samples during chronology construction. Unlike many scientific studies, sampling for dendroclimatological reconstruction relies on careful selection of the trees, not random selection. The reasons for this become clear when one considers the nature of the ring width itself. The ring width in any given growth season is the result of a number of factors: the biological age of the tree, environmental factors unique to that tree, stand-wide influences atypical of growth elsewhere, a large number of other factors, and of course the regional climatic signal that is of interest in these studies.

One can minimize the influence of non-climatic factors and maximize the climatic signal by careful site selection. Trees growing at the margins of the population, whether that be altitudinally or latitudinally are generally most sensitive to climatic factors. By choosing the dominant trees with no obvious signs of damage or disease, one enhances the relationship between ring width and regional climatic influence. By increasing the number of samples one reduces the individual tree responses and enhances the 'climatic signal'. The 'signal to noise ratio' has been studied in detail and will vary from location to location and species to species, but as a general guide a minimum of 15 to 20 trees seems to produce a representative sample.

The biological age of the tree generally has an effect on ring width, with a natural tendency to put on narrower rings as the tree increases in girth, even though overall productivity levels may be very similar. This so called 'age trend' can be readily removed by fitting curves to the overall series and then taking account of the difference of the observed ring width from the theoretical mean value at a given time. This process, known as 'standardization', results in the production of a series of indices of growth, and makes direct comparison between trees of different biological age more readily achievable.

Just how best to derive this series of indices for subsequent analysis has been the subject of a vast literature within dendroclimatology, and each new study really needs to justify the particular methods employed to remove age-related trends. The problem of course is that the more closely one fits curves to the original data, the more climatic information one may be subtracting, and whilst the great advantage of dendroclimatology is that it has the potential to produce annual resolution in the results, longer-term trends in growth may also reflect climatic information, and may be lost in the process of standardization. Even so, the resulting series is often highly autoregressive, that is to say that growth in any one year often reflects the overall vigour of the tree for the few prior years, and this autoregression needs to be removed to enhance

the year-to-year variation. Autoregressive Moving Average (ARMA) models have been perhaps the most frequently used standardizing models, though the literature is full of alternative methodologies, each of which may be more suitable in particular conditions.

Having satisfactorily arrived at a well replicated index series representing growth, the next stage is generally to model how the climatic information influences the observed growth. One of the great problems here is the availability of reliable meteorological records, and the length of those records for regions where the most environmentally sensitive trees are growing. The most useful tree-ring information is often from remote areas, and even where there are local records, these are generally only available over recent decades. Not surprisingly, an early study of oak growth in Britain showed that the variation in ring width attributable to mean monthly rainfall and temperature data varied according to distance from the meteorological station used.

Nevertheless, a process of stepwise principal component regressions generally allows a response function to be derived in which individual monthly rainfall and temperatures prior to, and during, the growth season can be recognized in order of their importance. For example, in a typical temperate northern-hemisphere model, April, May and June rainfall, and May, June and July temperature may be found to have the greatest influence on growth. A transfer function can then be derived so that these major elements are reconstructed throughout the length of well replicated tree-ring chronology. Typically, in the more sensitive trees growing at the margins of distribution, one factor can be derived as the major influence on growth, and the reconstructions are concerned with aspects of either rainfall or temperature.

Given that biological responses are often not linear, and that relatively crude data in the form of monthly means are the normal tools employed, it is surprising how reproducible the results have been. In 1982 a study of response functions for oaks growing at 16 sites in the British Isles was published. This showed many similarities in responses, including the surprising result that temperature in the December prior to growth was significantly negatively correlated with ring width in the following growth season. This was mathematically derived, without reference to the biology of the trees, and a tentative explanation was put forward. This suggested that a warm December resulted in the trees metabolizing the food reserves which were subsequently not available to the tree the following spring, resulting in less growth. In a cold December, the tree had 'shut down' allowing the reserves to be utilized for growth when conditions improved. This highlighted a need for better understanding of tree physiology which is still a requirement for better modelling of the climate–growth relationship.

Some early reconstructions were tempted to use very short periods of meteorological data to calibrate the model – often this was all that was available, although the ring series may go back several centuries. There is also a need to verify the model, so the short data series was divided into two parts, with one short series being used to calibrate the model and the other part to verify the reconstructed values over an equally short period. Hundreds of years were then reconstructed on this unsatisfactory model. Whilst this is tempting, and no doubt may highlight very different periods of climate within the tree-ring time-series, it is clearly open to inappropriate interpretation. Guidelines on methodology also suggest that in order to remove any effects of changes in response over time, the calibration and verification of the model should be done both ways around, i.e. using the outer years to calibrate and verify using the inner years, and then reversing the periods used for calibration and verification.

Critics of this response function approach to dendroclimatology point out that not only are responses seldom linear, and that dramatic changes can result from crossing threshold values, but that many influences on tree growth can occur at widely varying time-scales, perhaps in a matter of hours. For example, a severe frost on a single night in the late spring, when growth has not been going on for long and leaves are tender, may damage a significant proportion of the photosynthetic capacity of the tree and have a profound effect on growth for the remainder of that growth season. Such an event is unlikely to show up in a crude measure such as mean monthly temperature. Critics also point to the observable fact that whilst generally quite representative values are derived for much of the series, reconstructed values rarely recreate the more extreme years well, tending to underestimate the meteorological data (see **Figure 1**).

A number of studies, particularly based in Europe, have looked at growth in 'pointer' years – that is, years where a large proportion of the trees show a marked growth change in the same direction – and have managed to look in greater detail, often at daily weather records, to see what factors are responsible for these growth changes.

These two approaches both yield very valuable information, and are not mutually exclusive. It is by combining several approaches that the maximum information is likely to be gained in the future, although

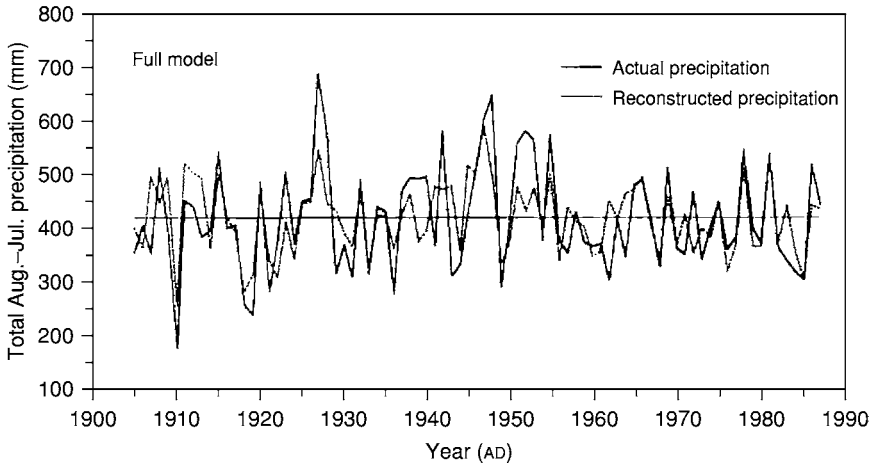


Figure 1 A comparison of the tree ring reconstructed and instrumentally recorded precipitation for the 1905 1987 calibration period. (Reproduced with permission from Case RA and MacDonald GM (1995) A dendroclimatic reconstruction of annual precipitation on the western Canadian prairies since AD 1505 from *Pinus flexilis* James. *Quaternary Research* 44: 267 275, copyright Elsevier 1995.)

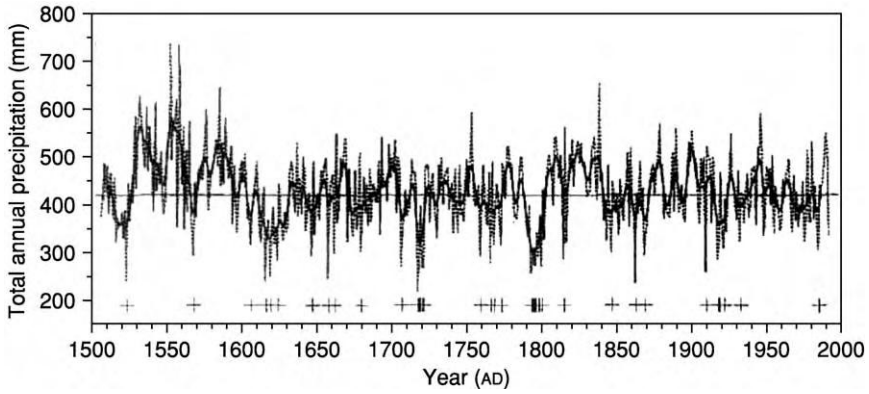


Figure 2 A 487 year (AD 1505 1992) reconstruction of August July annual precipitation. Drought years are indicated on the lower portion of the graph (+) and are defined as years when precipitation is more than one standard deviation less than the average of the complete series. The heavy line represents an 11 year low pass filter. (Reproduced with permission from Case RA and MacDonald GM (1995) A dendroclimatic reconstruction of annual precipitation on the Western Canadian prairies since AD 1505 from *Pinus flexilis* James. *Quaternary Research* 44: 267 275, copyright Elsevier 1995.)

of course the latter approach is only possible where the data are available.

Many dendroclimatic studies have concerned themselves with reconstructing one particular factor in a limited geographical area, for example summer temperature in a particular country or state. Another approach, which has been invaluable, is to produce a network of chronologies over national, continental or hemispherical ranges and to use these as the basis of more wide-scale reconstructions. This approach enables an assessment of the reconstruction values at the intermediate points to be made, and can identify areas within the network where further sampling is desirable. More ecologically based studies have shown changes in tree response that are elevation dependent (Figure 2).

Micro-Anatomical Variations

Whole ring, or earlywood/latewood width studies are the commonest form of dendroclimatological studies. In parallel with these studies however, there is a long history of studies in the relationship of intraring features with external growing conditions. Cell size, cell differentiation, cell distribution and other characteristics (e.g. wall thickness) have all been studied.

Stable Isotope Studies

This discussion has so far concentrated on the use of ring-width studies, but it is important to remember that dendrochronological studies (or tree-ring studies if one prefers) can and have used many other aspects

of the properties of the rings. Most notable amongst these is the use of X-ray densitometry. In this, uniform thickness cross-sections of the tree (usually from cores) are pre-treated to remove resins and other chemicals, and then subjected to an X-ray source, with film below. The resulting photomicrograph can be scanned to reveal not only ring-width information, but also the density of the cellulose across the ring. It has been found that the maximum latewood density in conifers is generally more highly correlated with meteorological factors than is simple ring width, and another advantage is that there is much less autocorrelation in density measurements. The disadvantages of this method are that is it far more costly and technologically demanding than pure ring-width studies.

Another area of tree-ring studies with a long history is the consideration of changing chemical isotopes in the wood itself. Both temporal and spatial variation can be found in the isotopic records of trees. Studies on variations in the deuterium, ^{13}C and ^{18}O isotopes have been used to look at long-term changes in temperature and precipitation.

Deuterium levels in tree rings contain information about the isotopic composition of the water taken up by the tree, and this is generally a reflection of the isotopic composition of the precipitation in the area where the tree grew. This isotope signal is closely associated with condensation temperatures, and deuterium levels in tree rings have therefore been used as records of palaeotemperature. Because of isotopic fractionation processes within the tree, deuterium levels have also been associated with humidity levels. It has also been shown that as a result of the covariation between temperature and relative humidity, that when this is taken into consideration, it is possible to produce a more sensitive temperature record from deuterium levels.

Carbon isotopes have been found to be linked with temperature, relative humidity, light intensity and water availability. Some of these associations, such as that with temperature, can vary considerably in different site conditions.

Stable carbon and hydrogen isotopes originate from different sources, the air and groundwater respectively, and hence a combination of isotope studies is likely to give a broader picture of environmental changes. Differences have been found between the isotopic composition of earlywood and latewood, and it is generally assumed that latewood values are more representative of the conditions at the time of wood formation, since there appears to be some utilization of stored food reserves during the formation of earlywood. Differences have also been found with the biological age of the tree, and therefore growth trends need to be removed as in many tree-ring studies.

In southern Germany, periods with high values in both isotope records occur around 5270 BC, 2990 BC and 2180 BC and are thought to represent times with high temperatures and low water availability. Similarly, low values around 6230 BC, 5600 BC and AD 390 probably represent unusually wet, cool periods. The 6230 BC corresponds with an event reported in North America in high-resolution palaeoclimate records, and it has been suggested that this may represent a time of rapid melting of the Laurentide ice-shield into the North Atlantic.

Other periods of coincidence in the hydrogen and carbon isotope records do not appear to be associated with information from other palaeoclimate records, and so some caution must be employed in their interpretation. It is hoped that the development of other proxy records – such as long-term varve chronologies – may help resolve these issues.

The ratio of ^{18}O to ^{16}O in the cell walls of plants results from changes in groundwater and evaporation. Levels vary with altitude and latitude, and oxygen isotope levels have long been regarded as records of palaeotemperature. It has been suggested from a study in India that relative humidity levels may be recorded in the tree rings, and it is likely that, as in other isotope studies, site conditions play a large role in determining the usefulness of particular studies.

Further research into tree physiology will undoubtedly unravel some of the more complex relationships between external environmental conditions and the incorporation of different stable isotopes into the wood. These will enhance the use of this specialist, and relatively expensive, area of tree-ring research.

Conclusions

Although very long chronologies have been constructed for many parts of the world, the replication of tree-ring data needed limits the time-frame for which these reconstructions are considered reliable, and the majority of those so far produced cover little more than five or six centuries, with a few notable longer contributions. The tree-ring-derived temperature reconstructions generally show greater century time-scale variability than is found from other proxy datasets, and suggest a cooler period in the late fifteenth and during the sixteenth centuries than is suggested by non-tree-ring data.

In recent years it has been suggested that the climate–growth relationship has fundamentally changed in character, perhaps as a result of increasing carbon dioxide levels or decreasing ozone levels. There seems to be a decoupling of the relationships

observed prior to the 1950s. Whilst this does not totally discredit the numerous reconstructions that have been carried out in recent decades, it does mean that their conclusions need to be used cautiously, and it poses serious problems for the calibration of future models used for climate reconstruction. Wide-scale ecologically based studies are highlighting variations in tree response at different site types in recent decades. It is perhaps more useful to extract climatic data from several different species growing in a variety of sites, but for which there is a limited historical dataset, rather than to attempt much longer reconstructions based on relatively few trees, in order to best understand the complexities of our changing climate.

See Also

Europe: Holocene. **Palaeoclimates**. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Baillie MGL (1982) *Tree Ring Dating and Archaeology*. London: Croom Helm.

Baillie MGL (1995) *A Slice through Time; Dendrochronology and Precision Dating*. London: Batsford Ltd.

Baillie M and Munro M (1988) Irish tree rings, Santorini and Volcanic dust veils. *Nature* 322: 344 346.

Becker B (1993) An 11,000 year German oak and pine dendrochronology for radiocarbon calibration. *Radiocarbon* 35: 201 213.

Becker B and Schirmer W (1977) Palaeoecological study on the Holocene valley development of the R. Main, southern Germany. *Boreas* 6: 303 321.

Briffa K, Jones PD, Wigley TML, Pilcher JR, and Baillie MGL (1983) Climate reconstruction from tree rings: part 1, Basic methodology and preliminary results for England. *Journal of Climatology* 3: 233 242.

Briffa K, Jones PD, Wigley TML, Pilcher JR, and Baillie MGL (1986) Climate reconstruction from tree rings: part 2, Spatial reconstructions of summer mean sea level pressure patterns over Great Britain. *Journal of Climatology* 6: 1 15.

Briffa K, Schweingruber FH, Jones PD, Osborn TJ, Harris IC, Shiyatov SG, Vaganov EA, and Grudd H (1998) Trees tell of past climates: but are they speaking less clearly today? *Philosophical Transactions of the Royal Society of London* 353B: 65 73.

Briffa KR, Osborn TJ, and Schweingruber FH (2004) Large scale temperature inferences from tree rings: a review. *Global and Planetary Change* 40(1+2): 11 26.

Case RA and MacDonald GM (1995) A dendroclimatic reconstruction of annual precipitation on the Western Canadian prairies since A.D. 1505 from *Pinus flexilis* James. *Quaternary Research* 44: 267 275.

Cook ER and Kariukstis LA (eds.) (1990) *Methods of Dendrochronology: Applications in the Environmental Sciences*. Dordrecht: Kluwer.

Cook ER, Bird T, Peterson M, Barbetti M, Buckley B, D'Arrigo R, and Francis R (1992) Climatic change over the last millennium in Tasmania reconstructed from tree rings. *The Holocene* 2(3): 205 217.

Dittmar C, Zech W, and Elling W (2003) Growth variations of common beech (*Fagus sylvatica* L.) under different climatic and environmental conditions in Europe a dendroecological study. *Forest Ecology and Management* 173: 63 78.

Fritts H (1976) *Tree Rings and Climate*. London: Academic Press.

Fritts HC (1991) *Reconstructing Large Scale Climatic Patterns from Tree Ring Data*. Tucson: University of Arizona Press.

Hughes MK, Kelly PM, Pilcher JR, and LaMarche VC (1982) *Climate from Tree Rings*. Cambridge: Cambridge University Press.

Jones PD, Bradley RS, and Jouzel J (eds.) (1996) *Climate Variations and Forcing Mechanisms of the Last 2000 Years*. Berlin: Springer Verlag.

Mayr C, Frenzel B, Friedrich M, Spurk M, Stichler W, and Trimborn P (2003) Stable carbon and hydrogen isotope ratios of subfossil oaks in southern Germany: methodology and application to composite record for the Holocene. *The Holocene* 13: 393 402.

McCarroll D, Jalkenen R, Hicks S, Tuovinen M, Pawellek F, Gagen M, Eckstein D, Schmitt U, Autio J, and Heikkenen O (2003) Multi proxy dendroclimatology: a pilot study in northern Finland. *The Holocene* 13: 829 838.

McCarroll D and Loader NJ (2004) Stable isotopes in tree rings. *Quaternary Science Reviews* 23: 771 801.

Pilcher J and Gray B (1982) The relationships between oak tree growth and climate in Britain. *Journal of Ecology* 70: 297 304.

Schweingruber FH (1988) *Tree Rings*. Dordrecht: Reidel.

Schweingruber FH (1996) *Tree Rings and Environment. Dendroecology*. Birmensdorf: Swiss Federal Institute for Forest, Snow and Landscape Research.

DESERTS

See **SEDIMENTARY ENVIRONMENTS: Deserts**

DIAGENESIS, OVERVIEW

R C Selley, Imperial College London, London, UK

Introduction

Lithification is the process by which unconsolidated sediment is turned into solid rock. In 1888, von Gumbel proposed the word 'diagenesis' to describe the processes of lithification. Subsequently, in 1957, Pettijohn wrote:

> Diagenesis refers primarily to the reactions which take place within a sediment between one mineral and an other, or between several minerals and the interstitial or supernatant fluids.

In 1968, Dunoyer de Segonzac gave a detailed account of the history of the concept of diagenesis, and pointed out that, in addition to the chemical processes mentioned by Pettijohn, some physical processes, such as compaction, are important in diagenesis. Thus, two types of diagenesis are now recognized: physical and chemical. These will be described in turn, but first it is appropriate to define the boundaries of diagenesis.

The Boundaries of Diagenesis

Diagenesis begins as soon as sediment is deposited. Fossilized beer bottles and other anthropogenic detritus found in modern 'beach rock' limestone picturesquely illustrate this. Ancient evidence of early diagenesis is further confirmed by the occurrence of intraformational conglomerates containing clasts, not only of contemporaneous limestone, but also of siderite-cemented sandstones and shales. As sediment is buried more deeply, temperature and pressure increase and, ultimately, diagenesis merges into metamorphism, with shale becoming slate, sandstone becoming quartzite, and limestone becoming marble. Field observation and laboratory experiments demonstrate that the boundary between these rock types, and hence diagenesis and metamorphism, is gradational. The sequence, deposition → diagenesis → metamorphism, is not a 'one-way street', however. At any time while sediment is on its way to metamorphism, it may be uplifted to the surface again. Rocks returned to the surface show a reversal of the trend of porosity decreasing with burial, and its enhancement by both physical and chemical processes. The term 'epidiagenesis' was applied by Fairbridge in 1967 to the diagenesis resulting from uplift and weathering. Epidiagenesis is of little significance in shales. It is, however, of great importance in sandstones and carbonates, because of the way in which it restores porosity and permeability to rocks that had previously lost these features. When buried beneath an unconformity, these epidiagenetically enhanced zones may provide excellent petroleum reservoirs. Epidiagenesis is also well known to mining geologists, being responsible for the 'gossan' sulphide ore bodies, such as those of Rio Tinto, Spain. Figure 1 delineates the boundaries of diagenesis.

Physical Diagenesis

The main processes of physical diagenesis are the expulsion of connate waters concomitant with the compaction of sediment. There is abundant evidence of compaction in sedimentary rocks, notably observations in thin sections of warped mica flakes and buckled fragmented shells. Arguably much of this compaction occurs during shallow burial, aided by seismic (including microseismic) shocks. Vibration causes the packing of sediment to tighten as the fabric adjusts and grains snuggle up close to each other.

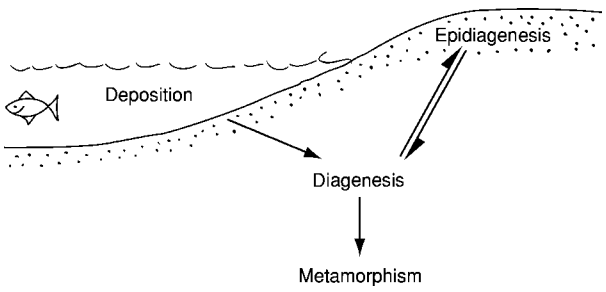


Figure 1 Diagram to show the interfaces and pathways of diagenesis.

Theoretical studies show that loosely packed sand may have a porosity of nearly 50%, but this is reduced to 27% for the tightest packing possible, a loss of 23%. Experiments reveal that poorly sorted muddy sediments compact more than clean well-sorted sands. This is because of the ease with which clay and finer particles may be squashed into the pores between the larger framework grains. Compaction gradually increases with burial depth, resulting in a concomitant loss of porosity. The exception to this general statement is the condition termed overpressure, in which the fluid pressure within pores diminishes the pressure at grain contacts and supports the sediment load. Once sediment is lithified, however, compaction ceases, although rocks have a degree of elasticity.

Graphs of porosity against burial depth for clay and sand are different. Clay burial graphs are curved, with the rate of porosity loss declining to become linear at depth. Sand burial curves, on the other hand, tend to be linear throughout. Intuitively, this suggests that clay loses porosity quickly by dewatering and compaction, whereas sands lose porosity more by cementation. Not all physical diagenesis results in porosity loss. Fracturing is an important process that takes place only in lithified rock: sedimentary, igneous, and metamorphic. Fracturing may occur in several ways, most commonly tectonic, but also, of course, anthropogenically around a borehole. Fracturing also occurs due to the release of stress when rocks are uplifted and the overburden pressure is diminished. Thus, rocks subjected to epidiagenesis become fractured, and fracturing is commonly well developed in truncated strata beneath an unconformity.

Fractures are extremely important, not so much because of the way in which they increase the porosity of a rock, which may be minimal, but because they cause a dramatic increase in the permeability of a previously impermeable rock. The resultant fractures may permit the flow of petroleum, water, and mineralizing fluids, which may, in turn, re-cement the very fractures that permitted their invasion.

Chemical Diagenesis

Chemical diagenesis includes mineral transformation, recrystallization, cementation, and dissolution. A wide range of minerals occur as cements in sedimentary rocks, precipitating in the pore spaces between the framework grains. The principal pore-filling cements are quartz, carbonate (calcite, siderite, and dolomite), and clay. Cement is obviously a major destroyer of porosity and permeability.

Chemical diagenesis, however, also includes the dissolution of grains, matrix, and cement by corrosive fluids. The grains most commonly dissolved are bioclasts, peloids, and ooids composed variously of aragonite (during shallow burial of limestone) and calcite. The matrix that is most commonly removed is lime mud, micrite. Similarly, the cement that is most commonly leached out is calcite, and other carbonates, rather than quartz. These observations demonstrate that dissolution results from the invasion of a sedimentary rock by acidic rather than alkaline pore fluids. Two sources of acidic fluid have been suggested. It is known that carbonic acid is expelled from shales prior to the emission of petroleum. It has been suggested that the carbonic acid serendipitously enhances the porosity and permeability of the potential reservoir, ahead of petroleum invasion. It is also noted that there is nothing new in acid rain. Sedimentary rocks exposed to epidiagenesis are subjected to flushing by acidic meteoric water. This leaches out carbonate grains, matrix, and cements. In limestones, it will enlarge existing fractures, and generate extensive zones of mouldic, vuggy, and cavernous porosity. The caves may synchronously fall in to form collapse breccias.

Dissolution, of whatever origin, will, of course, increase the porosity and permeability of sedimentary rock. Renewed burial will lead to renewed cementation, and the sedimentary rock will continue on its downward path to metamorphism.

Summary

Diagenesis is the term that describes the physical and chemical processes that take place in sediment after deposition and before it reaches the threshold of

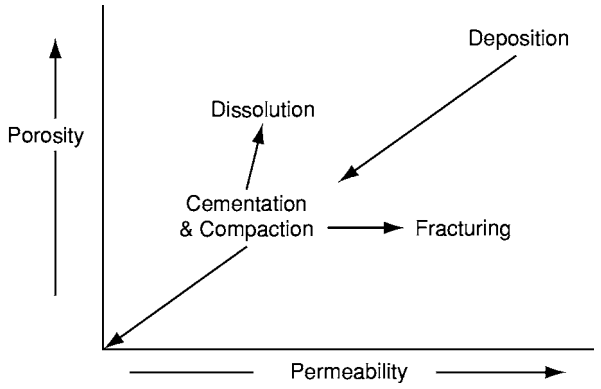


Figure 2 Diagram to illustrate diagenetic pathways and their petrophysical responses.

metamorphism. Physical processes include the loss of contained fluid and compaction, which destroys porosity, and fracturing, which enhances permeability. Chemical processes include porosity-destroying cementation and porosity-enhancing dissolution. The sequence, deposition → diagenesis → metamorphism, is not a 'one-way street', however. At any point along this route, the sediment may be uplifted to the surface again. Rocks returned to the surface show a reversal of the trend of porosity decreasing with burial, and its enhancement by the physical and chemical processes of fracturing and dissolution, respectively. Figure 2 summarizes the diagenetic pathways of sedimentary rocks.

Diagenesis is of great economic importance. Diagenetic studies form an integral part of the analysis of the evolution of porosity and permeability in aquifers and petroleum reservoirs. Diagenetic studies are also an important aid to understanding the formation of many mineral deposits.

See Also

Petroleum Geology: Production. **Sedimentary Rocks:** Chalk; Clays and Their Diagenesis; Dolomites; Limestones; Sandstones, Diagenesis and Porosity Evolution. **Weathering**.

Further Reading

Burley SD and Worden RH (2003) *Sandstone Diagenesis Recent and Ancient*. Oxford: Blackwell Science.

Dunoyer de Segonzac G (1968) The birth and development of the concept of diagenesis. *Earth Science Reviews* 4: 153 201.

Fairbridge RW (1967) Phases of diagenesis and authigen esis. In: Larsen G and Chilingar GV (eds.) *Developments in Sediments*, pp. 19 28. Amsterdam: Elsevier.

Giles MR (1997) *Diagenesis: A Quantitative Perspective*. Dordrecht: Kluwer Academic Publishers.

Leeder MR (1999) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Science.

Pettijohn FJ (1957) *Sedimentary Rocks*, 2nd edn. New York: Harper Geoscience.

Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

Tucker ME (1991) *Sedimentary Petrology*, 2nd edn. Oxford: Blackwell Scientific Publications.

DINOSAURS

See **FOSSIL VERTEBRATES: Dinosaurs**

EARTH

Contents

Mantle

G J H McCall, Cirencester, Gloucester, UK

Introduction

The mantle is the middle of the three primary concentric zones of the Earth, making up together with the core more than 99% of the planet's volume. It extends from the Mohorovicic Discontinuity (the base of the crust, which varies in depth from 25–35 km below the continents to 6–11 km beneath the ocean floor) (Figure 1A) to its boundary with the outer core at a depth of 2890 km.

The mantle itself is subdivided into an upper mantle, extending from the Mohorovicic Discontinuity to a depth of 670 km, and a much thicker lower mantle beneath. The upper mantle is further divided into a rigid upper zone, extending to a maximum depth of about 100 km, which together with the crust comprises the lithosphere, and a weaker ductile lower zone, which forms the asthenosphere. The ductile lower zone is commonly delineated at its base by a transition zone, extending from 400 to 670 km; the change to the lower mantle is transitional rather than abrupt. The lower mantle is delineated by a more rapid increase in density, amounting to 20%, as shown by an increase in the velocities of seismic waves. It is customary to consider the thickness of the lithosphere as varying from a maximum beneath the continents to zero beneath the mid-ocean ridges, the sites of eruptivity leading to sea-floor spreading (Figure 1A).

Direct Sampling of the Mantle is not Possible

The mantle, like the core beneath it, cannot be sampled directly. We shall never have revealed to us the actual rock material of which it is composed. Thus, all our knowledge of the mantle is derived from indirect methods, and this means that there remains a considerable degree of uncertainty as to its exact composition. The indirect methods that have been used to study the mantle are discussed below.

Seismology

Earthquake waves and similar seismic waves generated by artificial explosions (e.g. nuclear) provide a major line of evidence. Their velocities depend on the physical properties of the materials through which they pass. Experiments are carried out in the laboratory to establish the behaviour of such waves when passing through a known range of likely materials. Radial and lateral velocities can then be compared. Solutions, however, are seldom unique; there is also the difficulty of replicating the very high pressures of the deep mantle in the laboratory, so the method is less satisfactory when considering increasing depths in the mantle.

Chemical-equilibrium Studies

Experiments at high pressures and temperatures can examine the exchange reactions between different minerals and between minerals and magmas, and the stability fields of minerals at different pressures and temperatures, replicating the conditions at various levels in the mantle (Figures 2 and 3). From such experiments the possible or probable mineral and elemental make-up of the mantle can be deduced, but again there is no certainty in these results.

Peridotites and Oceanic Basalts

Materials erupting at the spreading centres in the oceans are considered to be virtually uncontaminated derivatives, ascending rapidly from layers in the mantle, and valuable conclusions can be drawn from the study of these. This method is believed to give reliable information about the mantle to a depth of about 300 km. Peridotite may well exist in the deeper mantle, and some authorities entertain a wholly peridotitic mantle.

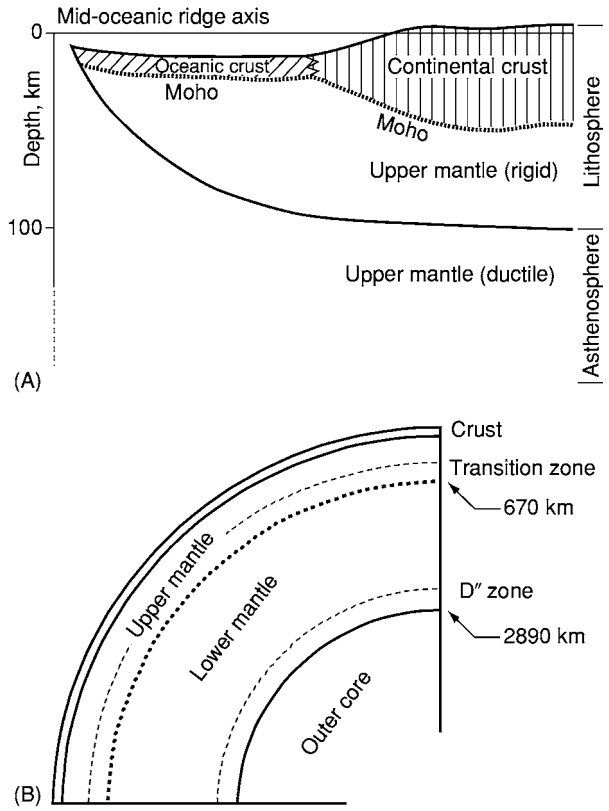


Figure 1 (A) Cross section from a mid ocean ridge to the inter ior of a continent, showing the relationship between the crust, lithosphere, Moho, and zones in the upper mantle. (B) The rela tionships between the upper mantle, lower mantle, and core (after Hancock PL and Skinner BJ (eds.) (2000) *The Oxford Companion to the Earth*. Oxford: Oxford University Press).

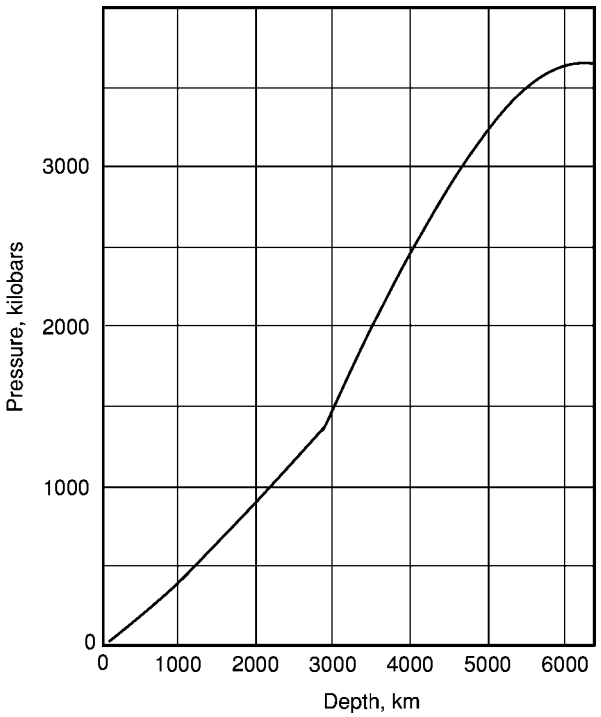


Figure 2 Pressure variation curve for the interior of the Earth (reproduced with permission from Mason B (1966) *Principles of Geochemistry*. New York, London, Sydney: John Wiley and Sons, Inc.).

Kimberlites

Kimberlites (*see* **Igneous Rocks:** Kimberlite) are mica peridotites, which, unlike oceanic peridotites, are volatile-rich. They carry mica and carbonate minerals. There are two ways in which these rocks, which are erupted very rapidly and explosively through continental crust in diatremes or pipes, sample the mantle. They contain diamonds with melt or mineral inclusions that come from great depths in the mantle, possibly as deep as the base at 2890 km. The diamonds in kimberlites are older than the host rocks, as shown by isotopic evidence, and have clearly been picked up during passage of the Kimberlite through the mantle to the surface. There may also be mineral or rock xenoliths, accidental fragments also picked up by the turbulent host magma on its ascent. These may range from comminuted debris to actual rock fragments. Kimberlites occur only within the continents, so they provide information only about the composition of the subcontinental mantle. Such accidental inclusions probably sample the mantle to a depth of at least 150 km.

Meteoritic Analogy

Geochemical modelling has been based on analogies with chondritic meteorites (*see* **Solar System:** Meteorites) because

- they are the most abundant extraterrestrial objects that strike the Earth (as measured by fall rates and excluding finds);
- they are essentially similar to terrestrial ultrabasic rocks, consisting of olivine-pyroxene assemblages with similar Fe/Mg ratios; and
- the stable-isotope data is in agreement with such modelling.

However, there are difficulties in accepting the 'chondrite Earth model', a major one being the difference in alkaline-element contents, and there is another problem involving the $^{13}C/^{12}C$ isotopic ratio, which is low in common chondrite meteorites and terrestrial basalts, but high in carbonaceous chondrites, terrestrial carbonates, and diamonds from kimberlite pipes (the latter must come from the mantle). Thus, a number of authorities believe that the mantle is derived from the accreted material of carbonaceous chondrites, rather than common chondrites. This conveniently provides a source for the water and other volatiles that are known to reside in the mantle (e.g. according to the Rubey model). The fact that

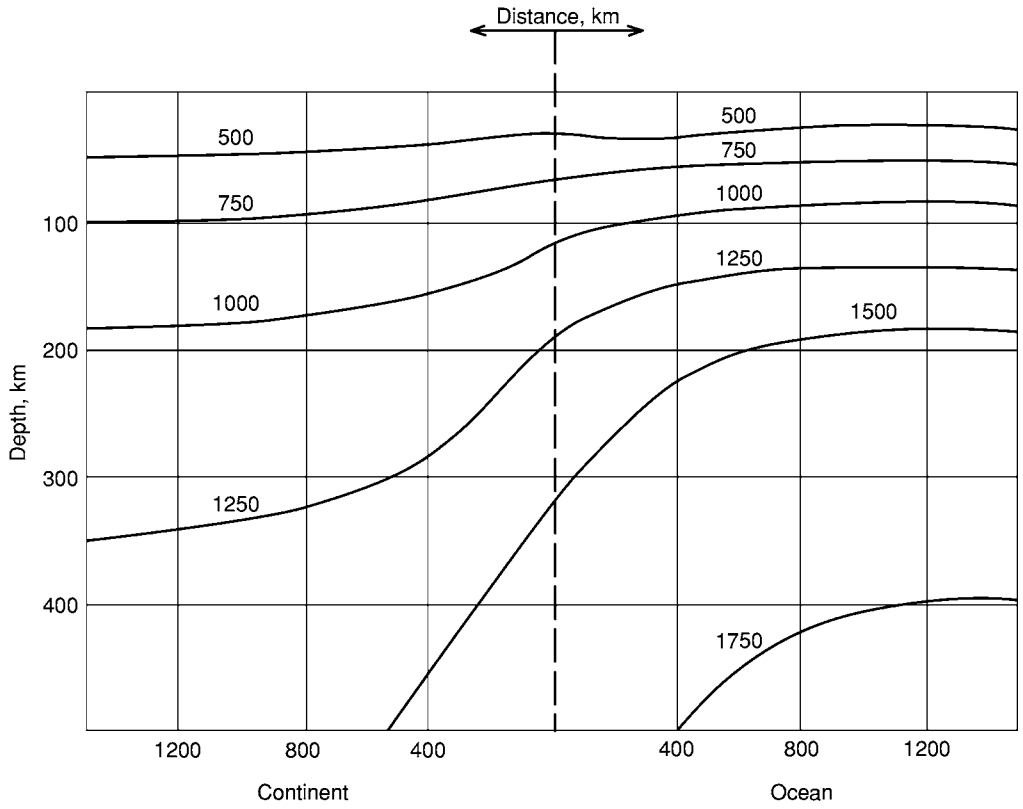


Figure 3 Temperature (degrees C) variation curve for the interior of the Earth. Lines are isotherms (reproduced with permission from Mason B (1966) *Principles of Geochemistry*. New York, London, Sydney: John Wiley and Sons, Inc.).

carbonaceous chondrites fall to Earth extremely rarely compared with common chondrites can be explained both by their fragility and by the fact that common chondrites, according to theory, are derived from them by metamorphic processes. However, doubts must remain about whether the material of chondrites, which accreted as asteroids in the interval between Mars and Jupiter at about 4500 Ma, is really representative of the material from which the Earth's mantle accreted at about the same time. There is an assumption of a homogeneous dust cloud from the Sun outwards.

The metal forming the core is supposed to have melted and gravitated to the centre of the planet (Figure 4). Whether accretion of the mantle from the dust cloud was homogeneous or heterogeneous (layered) remains undecided – both possibilities are shown in Figure 4.

Mantle Composition

Concepts concerning the composition of the mantle have inevitably changed since the acceptance of the plate-tectonic paradigm. It is relevant here to go back to the earlier statements. Those of Arthur Holmes and Brian Mason are concise summaries of what was then believed.

Holmes, in a 1965 reissue of his famous book *Physical Geology*, wrote

> It has long been thought that meteorites, stony and iron, might be direct clues to the nature of the Earth's mantle and core. Stony meteorites are like our terrestrial peridotites in many ways and, moreover a few varieties have compositions not unlike some of our basalts. For these and other reasons a similar range of composition for the mantle, with ultrabasic materials predominating, has seemed to be a plausible guess. The sudden change of seismic velocities at the Moho indicates either a change in composition (e.g. from basic rocks above to ultrabasic below) or a change in state (from, say, gabbro or amphibolite to a high pressure modification, which, for convenience, we may refer to as eclogite), without much change in composition. Olivine rich peridotites and garnet bearing varieties, at the appropriate temperatures and pressures, give seismic velocities that are about right for most parts of the upper mantle. Eclogite would do equally well, and various associations of the rocks would match seismic and density requirements. That eclogite does occur in the mantle is proved by the fact that it occurs in diamond pipes as inclusions. ... Moreover, basaltic magma ascends from the mantle and

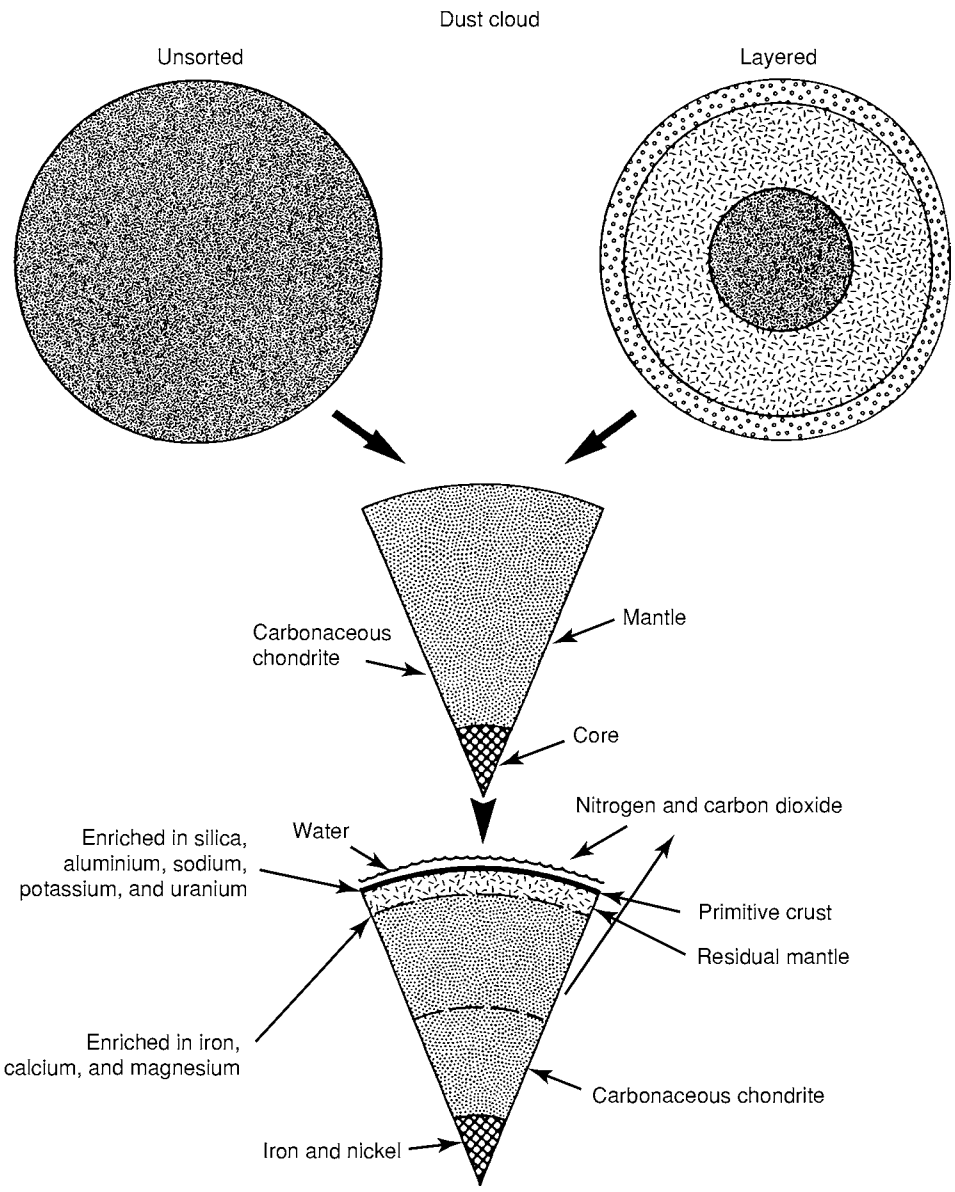


Figure 4 Two models for the accretion of the mantle and core (reproduced from Van Andel 1994).

one very probable source of basalt supply would be the melting of eclogite.

Although eclogites might represent primordial mantle or subducted oceanic crust, one problem with an eclogitic upper mantle is that eclogite xenoliths are scarce in kimberlites, whereas peridotite xenoliths are common. Mason, in a textbook reissued in 1966, stated that the alternative of a chemical explanation for the Moho was favoured by Clark and Ringwood in 1964; they preferred an ultrabasic model for the upper mantle, with the overall composition being one part basalt to three parts dunite (an olivine rock). This composition would equate to that of a pyroxene-olivine rock, which they called 'pyrolite'. This is a hypothetical composition, not a rock type. Fractional melting of this material would yield basaltic magma, injected into the crust over geological time, leaving residual dunite or peridotite. The pyrolite could crystallize in four ways, depending on the pressure and temperature in the mantle: olivine and amphibole;. olivine, aluminium-poor pyroxene, and plagioclase; olivine, aluminium-rich pyroxene, and spinel; or olivine, aluminium-poor pyroxene and garnet. Because of pressure and temperature differences, the compositions of the upper mantle beneath continents and oceans would differ. The proposed compositions under the continents and oceans are shown in Figure 5. Polymorphism of minerals was considered to be

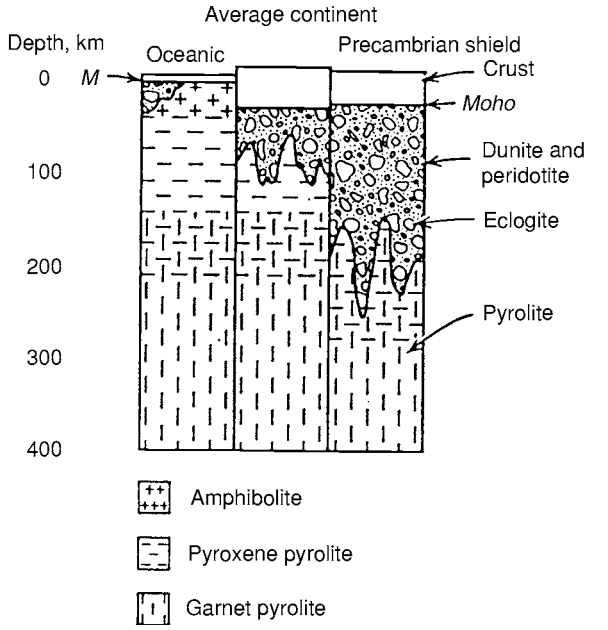


Figure 5 Suggested compositions of the mantle according to the original pyrolite hypothesis (reproduced with permission from Mason B (1966) *Principles of Geochemistry*. New York, London, Sydney: John Wiley and Sons, Inc.).

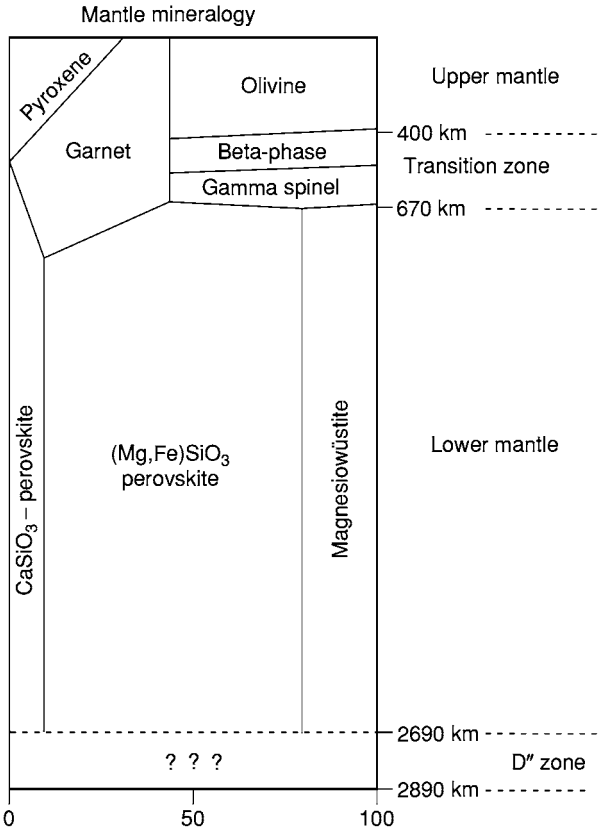


Figure 6 Suggested composition of the mantle at various depths. The verticle dimension is depth, the horizontal dimension is minerals (modified from Agee CE (2000) Mantle and core com position. In: Hancock PL and Skinner BJ (eds.) *The Oxford Compan ion to the Earth*, pp. 654 657. Oxford: Oxford University Press).

an important mantle process: olivine converting to an isometric spinel structure at higher density, pyroxene to an ilmenite structure, and quartz to the high-pressure polymorph that occurs naturally in impact structures – stishovite with rutile, stable at 130 kb and 1600 °C. Transformations into closer-packed forms were suggested to occur in the transition zone. The lower mantle was believed to consist of a mixture of $(Mg,Fe)SiO_3$ with an ilmenite structure and periclase (MgO), but the possibility of the ilmenite structure converting to that of perovskite (even denser) was suggested, as was the possibility of that then converting to a CaCl structure (though experimental evidence for these changes was lacking).

The pyrolite hypothesis, which was first stated by Clark and Ringwood in 1964, has survived the appearance of the plate-tectonic paradigm and was brought up to date in a summary by Agee published in 2000. He noted that a major point at issue is whether the mantle is homogenous or heterogeneous. The abrupt jumps or discontinuities in seismic properties within the transition zone are attributed to changes of minerals to denser, more closely-spaced phases; olivine is converted at a depth of 400 km to wadsleyite, an orthorhombic form found in meteorites, and at 520 km to ringwoodite (a gamma spinel form). At approximately 670 km olivine breaks down into perovskite ($[Mg,Fe]SiO_3$) and magnesiowustite ([Mg,Fe]O). Pyrolite has been given a hypothetical composition of 57% forsteritic olivine, 17% enstatitic pyroxene, 12% diopsidic pyroxene, and 14% pyrope garnet.

Pyrolite is a hypothetical analogue of the primordial composition of the mantle that existed after planetary accretion and core formation (the rigid part of the upper mantle, above the asthenosphere, is thus termed the 'residual mantle') (*see* **Earth Structure and Origins**). The shallowest part of the mantle is believed to have been continuously depleted in basaltic melt throughout geological time.

An up-to-date diagram of the likely composition of the mantle at various depths under the continents and oceans is given in Figure 6.

It has been noted above that there are two possibilities for the original accretion of the mantle from the dust cloud – homogeneous and layered. D L Anderson and others favour a compositionally layered mantle rather than the model of progressive downwards change under increasing conditions of pressure and temperature. The plate-tectonic paradigm requires the solid upper mantle to be rheid and slowly convecting as shown in Figure 7. Advocates of

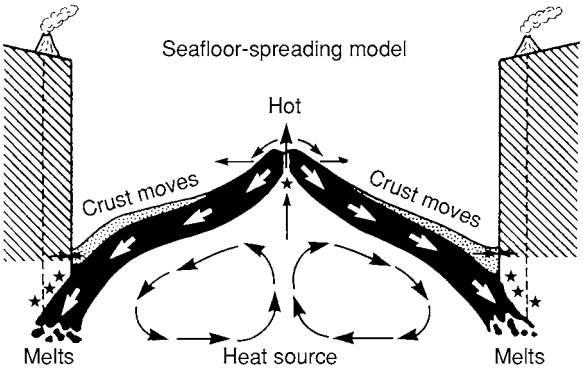


Figure 7 Mantle convection under a spreading ocean (reproduced from Van Andel 1994).

the compositionally layered mantle believe that the discontinuity at the base of the upper mantle, at approximately 670 km, is not only a phase-change boundary but also a compositional boundary. The layered model requires layered convection with little or no mass transport across this discontinuity. The pyrolite hypothesis is retained, but, whereas the upper mantle is taken to be composed of olivine-rich peridotite similar to pyrolite, the lower mantle is believed to be enriched in perovskite. This requires a lower mantle that is enriched in silica compared with the upper mantle, something that is in agreement with the chondritic Earth model. The compositional layering of the mantle could stem from the accretion phase, but could also be a product of an early melting stage, with the mantle crystallizing like a giant layered igneous intrusion.

The Problem of Subducted Slabs

It is a requirement of the plate-tectonic paradigm that lithospheric slabs descend into the mantle in subduction zones. One hypothesis is that they sink right through the mantle to a 'slab graveyard' at the core-mantle boundary. There is some seismological evidence for this in the basal 200 km of the mantle (the D″ zone). However, the seismological anomaly here may be explained in a number of ways, and advocates of the compositionally layered mantle believe that there is a high enough contrast of density and viscosity across the 670 km boundary to prevent any passage of material through it and that the slabs must remain in the upper mantle, the 'graveyard' being situated in the transition zone and affecting its composition. The sedimentary cover in these slabs contributes to the calc-alkaline eruptives in the crust above the descending plate (e.g. the Peruvian granitoid batholiths). There is no real certainty about what happens to the sediments descending into the mantle.

Tomography

Tomography delineates in three-dimensions regions of fast and slow seismic-wave in the Earth's interior, both radially and laterally. These may simply represent colder downwellings and hotter upwellings. They may alternatively represent compositional differences in a primordially layered mantle being complicated and eradicated by convection. Tomographical irregularities are invoked by advocates of slab descent to the base of the mantle and of plumes ascending through the 670 km discontinuity from that region (*see* **Earth Structure and Origins**), whereas advocates of 'closed upper mantle circulation of plate tectonics', such as Warren Hamilton, believe that plumes could not penetrate the 670 km discontinuity and that convection cells are restricted to the upper mantle.

What Drives Plates and Initiates the Pattern?

The plate-tectonic paradigm has been with us for four decades, but there is no certainty at all about what drives the plate motion. What actually starts the process? Is the rise of magma from the already convecting mantle at the site of the future mid-ocean ridge the splitting initiator? Or is initiation controlled by lithospheric irregularities above the mantle before convection begins? That is, is the primary drive exerted by the mantle or by the lithosphere? Though there has been much discussion of the driving forces of plate movement, none of the theories suggested seems adequate. However, we do have a clue in the Rift Valley of East Africa (*see* **Africa:** Rift Valley). This is manifestly an aborted ocean, a site of magmatism and rifting, which has, nevertheless, not developed into a spreading ocean – it has been 'halted in its tracks'. This surely is evidence that the process starts at the spreading centre, before any subduction takes place. It would seem to invalidate slab pull or the passive sinking of oceanic lithosphere, as favoured by Warren Hamilton, as the prime driving mechanism; rather, either already commenced mantle convection forces a split or the whole process is started by a lithospheric split along a favourable line of weakness in the crust– both mechanisms seem to be consistent with this evidence.

Conclusion

The mantle is of critical importance in understanding geological processes, yet remains elusive. Numerous models have been erected, and no doubt many more will be derived in the coming decades. An aspect that has not been touched on here, the variations in isotopic composition in basalts derived from the mantle,

illustrates the complexity of the problems. It is not possible to explain all the isotopic complexities by simple mixing of one enriched and one depleted reservoir, and the answer would seem to be that, even taking into account crustal contamination on ascent, there are a number of different reservoirs, a fact that argues for a heterogeneous mantle. Yet the exact petrological, spatial, and physical nature of the source regions remains as elusive as ever.

See Also

Africa: Rift Valley. **Earth:** Crust. **Earth Structure and Origins**. **Igneous Rocks:** Kimberlite. **Mantle Plumes and Hot Spots**. **Plate Tectonics**. **Solar System:** Meteorites. **Tectonics:** Earthquakes.

Further Reading

Agee CE (2000) Mantle and core composition. In: Hancock PL and Skinner BJ (eds.) *The Oxford Companion to the Earth,* pp. 654 657. Oxford: Oxford University Press.

Anderson DL (1989) *Theory of the Earth*. Oxford: Blackwell Scientific Publications.

Anderson DL (1999) A theory of the Earth. In: Craig GV and Hull JH (eds.) *James Hutton Present and Future,* pp. 13 35. Special Publications 150. London: Geological Society.

Clark SP and Ringwood AE (1964) Density distribution and constitution of the mantle. *Reviews of Geophysics* 2: 35 88.

Hamilton WB (2002) The closed upper mantle circulation of plate tectonics. In: Stein S and Freymueller JT (eds.) *Plate Boundary Zones,* pp. 359 410. Geodynamics Series 30. American Geophysical Union.

Holmes A (1965) *Physical Geology*. Edinburgh, London: Nelson.

Mason B (1966) *Principles of Geochemistry*. New York, London, and Sydney: John Wiley and Sons.

Price NJ (2001) *Major Impacts and Plate Tectonics*. London and New York: Routledge.

Ringwood AE (1975) *Composition of the Earth*. New York: McGraw Hill.

Rubey WW (1955) Development of the hydrosphere and atmosphere with special reference to the probable composition of the early atmosphere. In: Poldervaart A (ed.) *Crust of the Earth,* pp. 631 650. Colorado Geological Society of America Special paper 62 Part IV. Geological Society of America.

Van Andel TH (1994) *New Views on an Old Planet*. Cambridge: Cambridge University Press.

Wylie PJ (1967) *Ultramafic and Related Rocks*. New York, London, Sydney: John Wiley and Sons.

Crust

G J H McCall, Cirencester, Gloucester, UK

Introduction

Holmes, in 1965, defined the 'crust' or 'lithosphere' as the outer shell of the solid Earth, the two terms being synonymous at that time. He described it as being composed of a great variety of rocks, with a blanket of loose soil or superficial deposits (e.g., alluvium, desert sands). The term dates back to Descartes (1596–1650), who saw the crust as a shell of heavy rocks which was covered by lighter sands and clays and rested on a metallic interior. Leibnitz (1646–1716) believed that the Earth had cooled from an incandescent state and that the crust was the first consolidated rocky part covering a still molten interior. However, it became clear from the work of Kelvin (1862) that the Earth's tidal behaviour precluded a still molten interior beneath a thin solid crust.

The plate tectonics paradigm appeared in the late 1960s and required a rigid lithosphere extending to *ca.*100 km (*see* **Plate Tectonics**), including part of the upper mantle, above the convecting asthenosphere. This redefinition made the terms 'crust' and 'lithosphere' no longer synonymous. The crust is the rocky upper layer of the Earth with a loose superficial cover, and extends to the Mohorovičić discontinuity (Moho) at 5–35 km below the surface, whereas the lithosphere extends down to 100 km and includes the crust as its upper layer and also the uppermost part of the mantle beneath. The crust is thus the top compositional layer of the lithospheric plates, which move about the Earth's surface, the remainder of the plates being formed of the uppermost, rigid part of the mantle.

The oceanic crust and continental crust are of different character (**Figure 1**). The oceanic crust, at its thinnest, only extends down to 5 km minimum, but locally may be thicker, up to 15 km. It forms 59% of the total crust by area. The continental crust extends down to 30–80 km and forms 79% of the crust by volume. Under islands, continental margins, and island arcs, the crust is transitional and is 15–30 km thick.

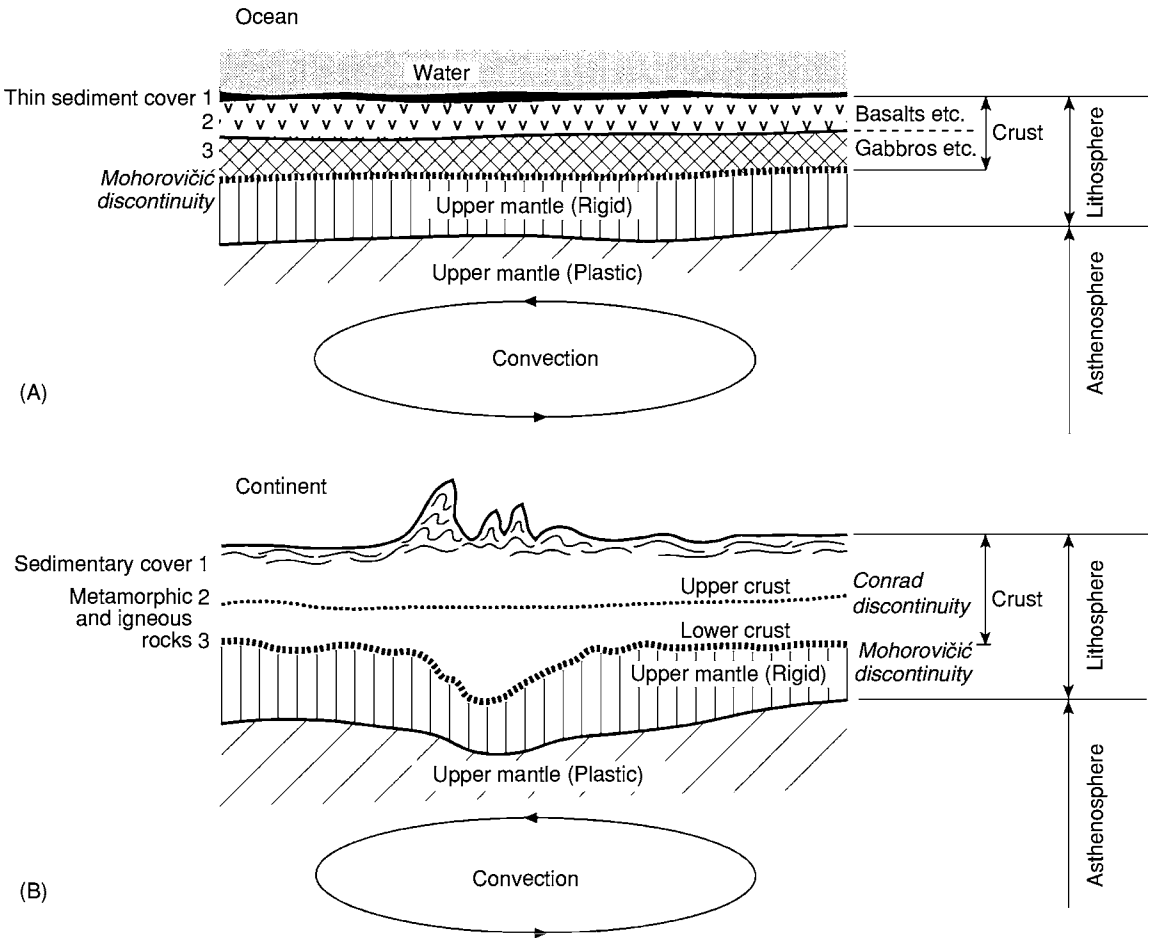


Figure 1 Diagrams showing the threefold layering of the oceanic (A) and continental (B) crust. The continental crust is much thicker than the oceanic crust. The two diagrams are not to scale.

Recycling of the Crust

Plate tectonics requires that crustal slabs are recycled in subduction zones and taken down to the mantle (Figure 2). However, this is by no means an even process, with oceanic crust being subducted very rapidly, but continental crust resisting recycling. The result of this is that the oceans are all products of the last plate dispersion cycle, which commenced about the beginning of the Mesozoic, and no oceanic crust is older than about 200 million years. In contrast, continental crust has been preserved almost indefinitely, and there is some continental crust in the Slave Province of Canada which is more than 4000 million years old. Minerals (zircons) in continental rocks in Western Australia are also up to 4300 million years old.

Physical Regions of the Crust

Continental crust is subdivided into a number of physical regions.

- Shields: deeply eroded expanses of low relief, which have been stable since Precambrian times.
- Platforms: similar to the above, but mantled by thick sedimentary cover, which may be entirely or in part Phanerozoic in age.
- Orogens: long, curved belts of folded rocks, usually forming mountain chains, mostly formed by continental collisions.
- Rifts: linear, fault-bounded depressions, traversing continents; these are the structures which originate crustal splitting and dispersion, and lead to mid-ocean ridge formation, but they may, as in the case of the East African Rifts, be aborted, i.e., never developed into oceans.

Oceanic crust is also subdivided into a number of physical regions.

- Volcanic islands: scatterings or chains of islands which mark mantle hotspots.
- Volcanic arcs: above subduction zones (e.g., Aleutian Islands), these may be represented by their

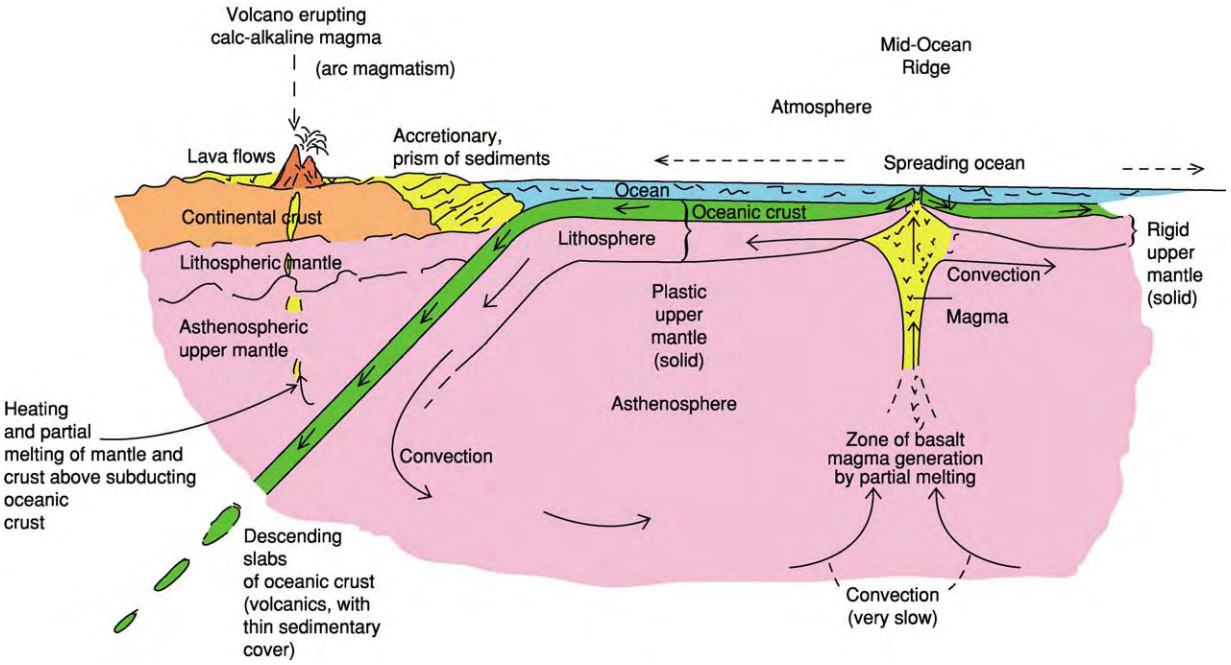


Figure 2 Diagram showing the recycling of the crust by subduction. After Edwards K and Rosen B (2000) *From the Beginning*. London: Natural History Museum.

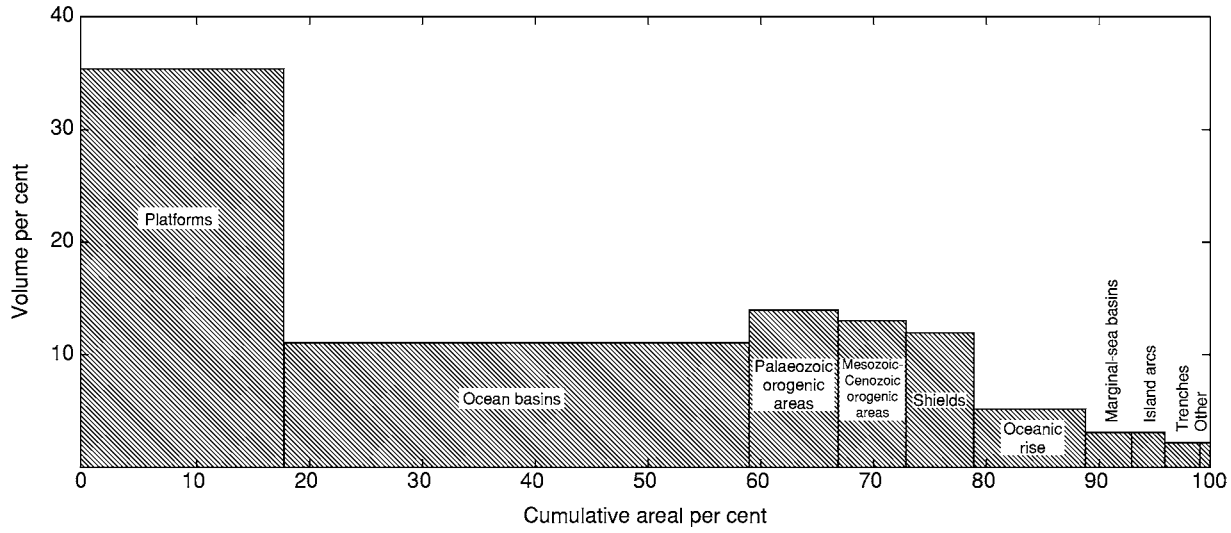


Figure 3 Area and volume proportions of the major crustal types. From *Plate Tectonics and Crustal Evolution* by K. C. Condie (1976). Reprinted by permission of Elsevier Ltd.

eroded underworks – chains of calc-alkaline batholiths (e.g., Peru).

- Trenches: the outer margin of subduction zones; the deepest parts of the oceans.
- Ocean basins/abyssal plains: the extensive flat, deep areas of oceans, beyond continental slopes.
- Marginal basins: small basins separating arcs, or landward from arcs (back-arc basins).
- Inland seas: seas within continents (e.g., Caspian).

The relative area and volume of the major crustal types are shown in Figure 3.

Crustal Structure

We can broadly determine the contrasting layers in the crustal structure by means of seismic wave velocities, determined from earthquake waves and artificial (e.g., nuclear) explosions. This shows us

that there are three contrasting layers in the oceanic crust and also three in the continental crust, but they are different. In the oceanic crust, we can use analogies with 'ophiolites' (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges), which have been recognized as old oceanic crust, subducted and later obducted onto the continental surface, where they form bodies of surface rocks. Ophiolites are present in surface rock assemblages of all ages, as far back as the Archaean upper boundary (2500 Ma). The three layers of the oceanic crust are believed to comprise a sediment layer (0–1 km thick), a basement layer (mainly basalt) (0.7–2.0 km thick), and an oceanic layer (sheeted basaltic dykes, gabbros) (3–7 km thick).

The continental crust can be divided into an upper layer (10–20 km thick) and a lower layer (15–25 km thick). Sedimentary cover, possibly up to several kilometres thick, may or may not be present above the upper layer. The boundary between the upper and lower layer may be well defined by a change in seismic velocities, although it may not be detectable. It is called the 'Conrad discontinuity'.

Chemical Composition of the Crust

Much of the crust is buried, and we have no direct access to the material from which it is composed; therefore, much reliance must be placed on indirect methods of determination.

Seismic Wave Velocities

An important method is based on the measurement of seismic wave velocities, which indicate what materials are possible for any particular layer according to density (*see* **Seismic Surveys**).

Chemical Analyses

The many different rock materials that make up the continental shields can be analysed directly and weighted according to observations of relative abundance to give an average for the upper continental crust. Fine sediments, such as clays, can be analysed, being assumed to match closely the average composition of the upper continental crust. This assumption is based on the fact that clays give consistent values for thorium and rare-earth elements, which are relatively insoluble in natural water; amounts of such elements vary widely in concentration between different rock types of the upper continental crust, and this suggests that mixing in such fine-grained sediments of the contributory source rocks is very thorough.

Studies of Sections of Deep Crustal Rocks Exposed by Tectonic Processes: Deep-Sourced Xenoliths in Volcanic Rocks

There are some regions in which tectonic processes have revealed sections of the crust, on end, so that a geologist can traverse across the surface terrain, encountering progressively deeper crustal levels, including deep metamorphic rocks, such as granulites, at the surface. One such case is the Kapuskasing Belt in Ontario, Canada. Such exposed sequences of the deeper crustal rocks indicate that, in the deeper continental crust, below the Conrad discontinuity, there is much variation in both the horizontal and vertical dimensions, and that models invoking uniform layering of, for example, a granulite layer are not valid. The evidence from both exposed sequences and deep-sourced xenoliths in volcanic rocks suggests, besides heterogeneity, that the lower continental crust is of more basic, less siliceous, composition than the upper continental crust – that is, more rocks of basaltic composition are represented and metamorphic granulites are characteristic.

Oceanic Crust

The composition of the oceanic crust can be derived from the study of ophiolitic sequences, which have a systematic layering from lavas at the top, through sheeted basaltic dykes, to gabbros or layered differentiates of gabbroic magma below. Ophiolites can commonly be studied in traverses on the ground, through sections exposed on end, and even, as in Oman and western Newfoundland, from exposures of the oceanic crust–mantle boundary. Deep ocean sediments are very thin and can be ignored in making a weighted assessment according to the observed abundance of these three components. Such an estimate yields a basaltic composition for the oceanic crust, but poor in K_2O.

The values derived from such methods and given in Table 1 seem to be reasonable.

Table 1 Chemical composition (wt.%) of the crust

Component	*Continental upper layer*	*Continental lower layer*	*Ocean crust*
SiO_2	65.5	49.2	49.6
Al_2O_3	15.0	15.0	16.8
TiO_2	0.5	1.5	1.5
FeO	4.3	13.0	8.8
MgO	2.2	7.8	7.2
CaO	4.2	10.4	11.8
Na_2O	3.6	2.2	2.7
K_2O	3.3	0.5	0.2
Total	98.6	99.6	98.6

Partial Melting within the Crust

At the mid-ocean ridges, where the oceanic crust is thinnest, partial melting of the upper mantle material leaves behind a depleted mantle. Some crustal material may also suffer partial melting, modifying the residual oceanic crust, but this is probably not a major process, because the ascent of magma is rapid here. However, in the island arc situations in the overriding plate above the subducting plate, the eruptives that form the volcanic arcs and batholiths beneath them, mainly of calc-alkaline igneous rocks, pass quite slowly through the quite thick crustal rocks of the overriding plate; therefore, here there is much mixture of partial melting products from both mantle and crust, and the crust is always suffering modification and displacement. This process makes for a very heterogeneous continental crust at all levels.

Crustal Growth and Loss

The fact that some isotopically derived ages of continental crustal rocks are more commonly obtained than others, e.g., 2700 Ma and 2000 Ma being particularly common, suggests that the rate of new continental crust growth compared with loss by subduction has not been constant through geological time. This may be explained in terms of the plate tectonics paradigm, but may also be explained, at least partly, if plate tectonics processes did not operate or did not operate significantly in Archaean times (something for which there is considerable evidence). This is, however, an area of considerable uncertainty. In any case, there are a number of processes by which continents grow: (1) by continental plate collision (which does not affect the overall total of continental crust, but can cause crustal loss by metamorphism of the deep roots of collision zones to denser *PT*-compatible mineral phases, gravitational instability, and sinking to the mantle); (2) by underplating by basaltic magma from mantle hotspots or plumes (assuming that plumes do exist), or surface eruption therefrom to form lava plateaux (e.g., the Deccan); (3) by the addition of ophiolite material of old oceans; (4) by obduction onto continental edges; or (5) by collisions with oceanic terranes (islands, island arcs, submarine plateaux). The return process may be the direct subduction of continental slabs, or a three-stage process of erosion of continental crustal material, deposition of it as sediments, and their subduction.

Oceanic crust is either rapidly subducted or lost to continents by obduction onto their edges or by collisions of terrains with continents as mentioned above. It does not have a protracted geological history as does continental crust. It can, however, be added to during its relatively ephemeral history by the addition of magma from mantle hotspots or plumes beneath (e.g., Ontong-Java Plateau).

The Primitive Crust

It is widely believed that the early Earth possessed a primitive globe-encircling basaltic crust. The present surfaces of Mercury, the Moon, or Mars may provide an analogue as they have apparently never experienced plate tectonics and there is evidence suggesting a basaltic surficial shell still covering the whole of Mars (*see* **Solar System:** Mars). We now know that there must have been some continental crust very early in the Earth's history, the evidence being in the form of zircons of continental crust provenance up to 4300 Ma at Mt Narryer in western Australia. The process of continental crust formation has been compared to a scum or slag of light material forming on the surface of a basaltic melt, repeated again and again, until extensive continents emerged after a repetition of cycles. Thick continental blocks appear to have been in place by 3500 Ma, being represented by the oldest cratons of South Africa and Australia, but the earlier protocontinents were probably small and thin. This is as far as we can go at the present state of knowledge.

The Crust and Isostasy

The concept that gravitational equilibrium controls the heights of continents and ocean floors was proposed by Dutton in America in 1889. The analogy with wooden blocks floating in water was elegantly illustrated by Holmes in 1965 (Figure 4). The concept is one of hydrostatic balance, the Earth's major relief being compensated by the underlying differences in density, whereas minor relief features are compensated by the strength of the underlying rock. The compensation depth was traditionally placed at the Moho, but the plate tectonics paradigm necessitates a rethink of this. Isostasy is, however, a real crustal process and can be seen in action in the central sag of Greenland under the weight of the enormous

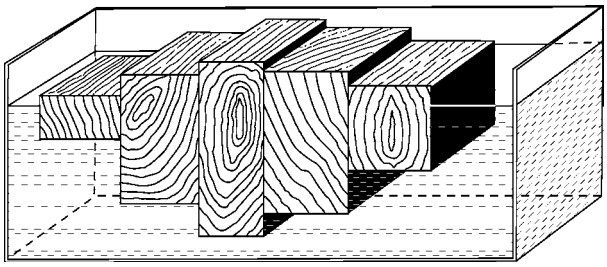


Figure 4 Representation of isostasy using floating wooden blocks of unequal dimensions. Reproduced with permission from Holmes A (1965) *Principles of Physical Geology*. London: Nelson.

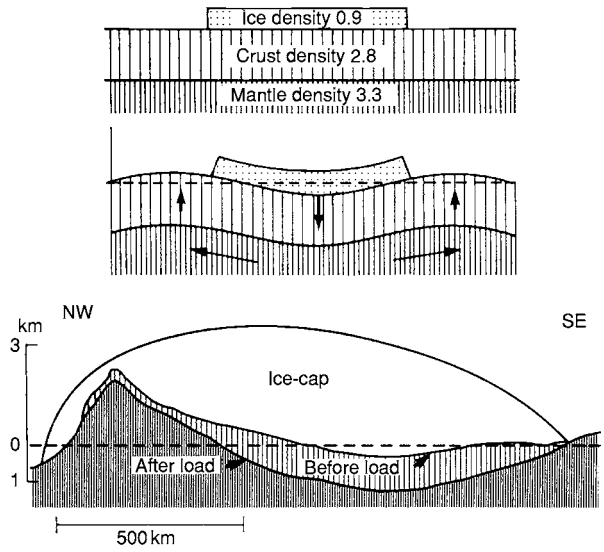


Figure 5 The Earth's crust floats on a dense mantle that behaves as a viscous liquid. If we place an ice cap on it, enough mantle material is displaced sideways to equal the additional crustal load, and a low broad welt is raised peripheral to the ice cap. The mantle adjustment is very slow, so that depression and rebound after removal of the ice cap are delayed. The situation at present in Greenland is shown in the lower part of the figure, and the delayed rebound is being experienced in the Baltic. From Van Andel TJ (1994) *New Views on an Old Planet*. Cambridge: Cambridge University Press.

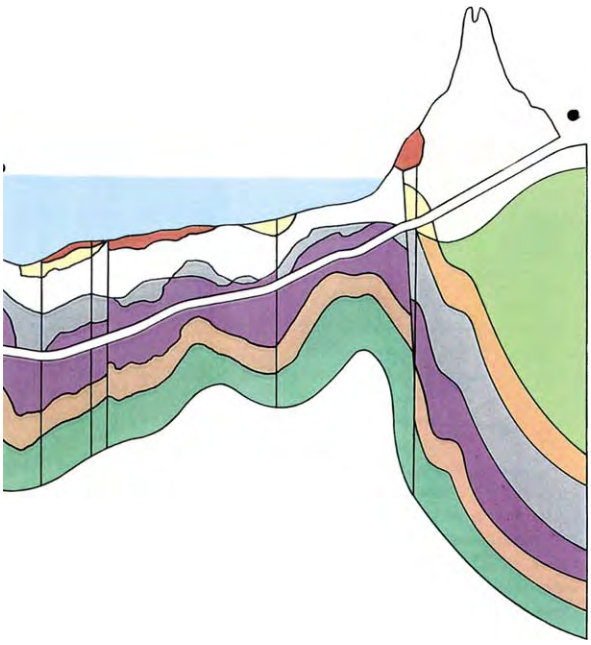


Figure 6 A sectional diagram of the French entry section to the Channel Tunnel showing tectonic deformation (folding, faulting) of the Cretaceous and Quaternary strata of the crust. Extract from BREAKTHROUGH by Derek Wilson published by Century Hutchinson. Reproduced from The Random House Group Limited.

ice-sheet which has sagged below sea-level in places (Figure 5), and also in the rebound occurring now in the Baltic after the ice-sheet has gone – the renowned Swedish geologist, Harry von Eckerman, used to go to the Alnö Island carbonatite complex yearly to see what new rocks had surfaced due to this rebound.

Heat Flow to the Crustal Surface

The continental surface heat flow is linearly related to the heat productivity of the near-surface granitic rocks. Variability from region to region is mainly related to the distribution of near-surface radioactivity, derived from certain minerals in the granitic rocks. Most models favour an exponential decrease in radioactive heat source with depth in the continental crust and a residual amount of heat rising from the upper mantle. The weighted average heat flow from both continental and oceanic crust is 1.5 heat flow units (HFU). This equivalence is explained by the heat flow to the oceanic crust surface coming mainly from the mantle, whereas it comes mainly from radioactive minerals in the crust at the continental surface. This requires increased radioactive sources in the mantle under the oceans, or that convective heat from the mantle under the oceans happens to equate to the radioactive-sourced heat from the continental crust, or that the mantle-derived convective heat is much the same under continental and oceanic lithosphere, but the thicker continental lithosphere is more depleted in radioactive heat sources than the oceanic lithosphere. The latter is preferred because it allows equal movements of both continental and oceanic crust-dominated plates.

Crustal Deformation

The crustal rocks of the Earth are subject to many and diverse deformation processes. The most significant are 'tectonic' processes (folding and faulting; Figure 6), which act very slowly through long periods of geological time and are mostly related to the movement of tectonic plates (especially collision and subduction). Stress on faults is subject to long build-up, but may be relieved by abrupt, almost instantaneous dislocation (which may form a linear earthquake scarp on the surface; Figure 7), or prolonged dissipation by creep movements without any earthquake. Earthquake foci are mostly situated within the crust, but some, especially those in margins of continental areas of plates, may have foci several hundred kilometres deep, within the mantle (*see* **Tectonics:** Earthquakes).

In contrast with tectonic deformation, 'superficial' deformation of the crust, such as landsliding, sink-hole formation, and submarine slumping, occurs abruptly, within a matter of seconds, minutes, hours, or days (Figure 8) (*see* **Sedimentary Processes:** Landslides).

Figure 7 A low fresh scarp of the Nojima Fault, Awaji Island, caused by the Kobe Earthquake, Japan, 1995. This was the almost instantaneous product of relief of one of many successive prolonged build ups of stress on an active fault. Reproduced with permission from Esper P and Tachibana E (1998) The Kobe Earthquake. In: JG Maurel and M Eddleston (eds.) *Geohazards in Engineering Geology, Geological Society of London Engineering Geology Special Publication 15*, pp. 105 116. London: Geological Society.

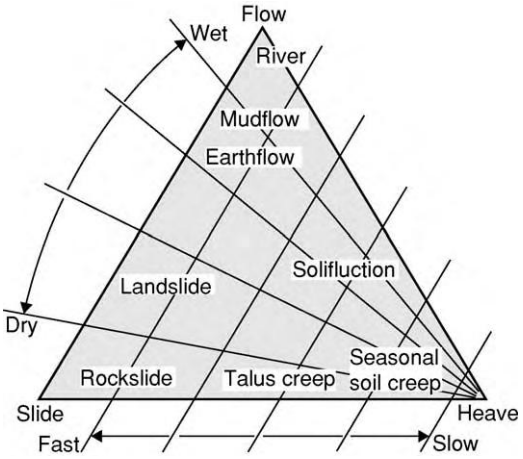


Figure 8 Classification of superficial mass movements on slopes and examples of some major types. Reproduced with permission from Goudie A (1989) *The Nature of the Environment*. Oxford: Blackwell.

Conclusion

The Earth's crust is the site of the great majority of processes studied by geologists, and thus is better known and understood than the mantle or core. However, there remain aspects of uncertainty, particularly the mechanisms of plate movement and earthquakes, and the exact nature of the lower parts of the crust.

See Also

Earth: Mantle. **Earth Structure and Origins**. **Mantle Plumes and Hot Spots**. **Plate Tectonics**. **Seismic Surveys**. **Sedimentary Processes:** Landslides. **Solar System:** Mars. **Shields**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Earthquakes; Mountain Building and Orogeny; Ocean Trenches.

Further Reading

Condie KC (1976) *Plate Tectonics and Crustal Evolution*. New York, Toronto, Oxford, Sydney, Braunschweig, Paris: Pergamon Press.

Edwards K and Rosen B (2000) *From the Beginning*. London: Natural History Museum.

Goudie A (1989) *The Nature of the Environment*. Oxford: Blackwell.

Hancock PL and Skinner BJ (2000) *Oxford Companion to the Earth*. Oxford: Oxford University Press.

Holmes A (1965) *Principles of Physical Geology*. London: Nelson.

Stamp LD (1951) *The Earth's Crust*. London, Toronto, Wellington, Sydney: GGHarrap.

Van Andel TJ (1994) *New Views on an Old Planet*. Cambridge: Cambridge University Press.

Orbital Variation (Including Milankovitch Cycles)

H Pälike, Stockholm University, Stockholm, Sweden

Introduction

Earth's orbital variations, caused by the mutual gravitational interaction between the sun, the planets, and their satellites, have been invoked as a mechanism for long-term variations (time-scales of 10^4–10^6years) of Earth's climate system. The orbital variations are known as 'Milankovitch cycles', after the Serbian mathematician Milutin Milankovitch. In this article, the relationships between the different parameters that affect orbital variations are explored, with a further view towards long-term patterns.

Celestial Mechanics

The time-varying motion of the planets and other satellites around the sun and around each other is controlled by mutual gravitational interactions. This relationship can be described by Newton's laws, corrected by Einstein's principles of general relativity. The overall behaviour of the entire solar system is even more complex; it is posed as an *N*-body problem (which is solved numerically), and is further complicated by physical parameters such as the tidal dissipation of energy and the detailed and changing distribution of mass within each body. Given this complexity, the planets do not follow stationary orbits around the Sun, but rather undergo perturbed quasiperiodic variations that, from a climatic point of view, this affects the amount, distribution, and timing of solar radiation received at the top of Earth's atmosphere. Milutin Milankovitch explained glaciations and de-glaciations (on time-scales of 10^4–10^6 years) by variations of solar radiation distribution at the top of Earth's atmosphere, coupled with the latitudinal distribution of land masses on Earth (*see* **Palaeoclimates**).

At any given time, the orbit of a body can be described by six parameters. These parameters define the position, shape, and orientation of an orbit and the location of a body in the orbit, with respect to a frame of reference. The trajectory of the orbiting body follows an ellipse; in the case of the Solar System, the sun is located at one focal point. Due to gravitational interactions between the different planets, the orientation and dimension of the ellipse change over time. Figure 1 illustrates how the relationships between the six 'Keplerian orbital elements' (a, e, i, λ, $\tilde{\omega}$, and Ω) apply to Earth in its orbit. A reference plane is fixed with respect to the stars; it is typically chosen as the orbital plane of Earth at a particular time (say, AD 1950 or 2000) and is called the 'ecliptic of date'. Alternatively one can choose a plane perpendicular to the total angular momentum of the Solar System (the 'invariant' plane). The 'invariant' plane almost coincides with the orbital plane of Jupiter due to its large mass. The reference plane is defined by two axes, both originating from the position of the Sun (S; Figure 1) on the reference plane. One of these, γ, is typically the position of the mean vernal (spring) equinox on the reference plane at a given time; the second axis, Z, is perpendicular to γ as well as to the reference plane. The parameter a, the semi-major axis of the orbital ellipse, corresponds to the average radius. The eccentricity e of the ellipse is defined as $e = (a^2 - b^2)^{1/2}/a$, where b is the semi-minor axis of the ellipse. The inclination of the orbit with respect to the reference plane is given by the angle i. The position of the ascending node N is specified by the angle Ω ('longitude of the node'), measured from the fixed direction in the reference plane (γ). The parameter $\tilde{\omega}$ specifies the position of the moving perihelion P (the closest approach to the Sun) and is defined as $\tilde{\omega} = \Omega + \omega$ ('longitude of the perihelion'). Finally, the sixth Keplerian element, λ, specifies the position of the orbiting body (J) on its elliptical orbit, and is defined as $\lambda = \omega + M$, where the mean anomaly M is an angle that is proportional to the area SPJ (Kepler's third law).

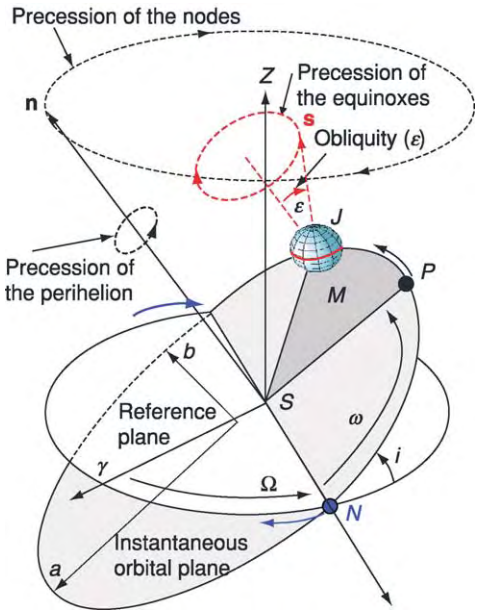


Figure 1 Orbital elements of the Earth's movement around the Sun.

Origin of Orbital Frequencies

If Earth were the only planet orbiting the sun, and in the absence of dissipative effects and other physical processes, the position and orientation of its orbit would remain fixed for all times. In this situation, the only Keplerian element that would change over time is λ, which describes the position of a body in its orbit. In this case, the position and velocity of the orbiting body would vary according to the Keplerian laws of motion around a fixed ellipse. However, gravitational interactions between all bodies of the Solar System cause changes in the shape and orientation of the elliptical orbit on various time-scales, which are typically of the order of 10^4–10^6 years. From a long-term climatic point of view, the relevant variations are those obtained after averaging the planetary orbits over their long-term orbital periods. These are termed 'secular variations', and they can be related to a set of fundamental frequencies. Variations in the orbital elements that characterize the secular variations can be separated into two categories that are related to different types of precession movements. The first category, variation within an orbital plane, is described by the variation of the eccentricity e and the rotation of the location of the perihelion (described by ω). The second category, variation of the orientation of an orbital plane, is described by the inclination angle i and the location of the ascending node N (described by Ω). These oscillations are coupled such that it is possible to investigate the behaviour of these parameters as pairs: (e, ω) and (i, Ω).

Fundamental Frequencies of the Solar System

Computing the orbital elements for the eight main planets (Pluto can be excluded due to its small mass) obtains eight characteristic modal frequencies for each of the paired elements (e, ω) and (i, Ω). Table 1 illustrates these fundamental frequencies, estimated over the past 20 My. Individual frequencies g_i are related to variations in the pair (e, ω), whereas frequencies s_i are related to variations in the pair (i, Ω). The individual g_i and s_i frequencies arise as eigenvalues from a matrix of linear differential Lagrange/Laplace equations that are used to expand the orbital elements for the eight planets. As eigenvalues of a matrix, they are not strictly associated with a particular planet. However, because the matrix from which they are obtained has a rear-diagonal structure, suppressing a planet removes one set of frequencies but does not change the other frequencies significantly. Thus, the indices g_i, s_i can be used to indicate which planet provides the strongest contribution to a particular frequency (indices g_1, s_1 correspond to Mercury; indices g_3, s_3 correspond to Earth; and so on). Note that all of the g_i terms are positive, indicating that the perihelia advance counterclockwise if viewed from the 'north' of the orbital axis shown in Figure 1. In contrast, seven out of the eight s_i terms are negative, indicating that the positions of the nodes, which mark the intersection of the orbital plane with the reference plane, regress (rotate clockwise). Due to considerations of angular momentum the eighth frequency is set to zero, and by conversion is chosen to be s_5 because the invariant plane is close to the orbital plane of Jupiter, due to Jupiter's large mass (*see* **Solar System:** The Sun; Asteroids, Comets and Space Dust; Meteorites; Mercury; Venus; Moon; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus).

General Precession of Earth

In addition to the fundamental orbital frequencies, which apply to the Solar System as a whole, two additional fundamental frequencies are necessary to

Table 1 Fundamental orbital frequencies[a]

Related to (e, ω)			*Related to* (*i*, Ω)			
Term	*Frequency (″ year* ^{1}*)*	*Period (ky)*	*Term*	*Frequency (″ year* ^{1}*)*	*Period (ky)*	*Associated planet*
g_1	5.596	231.0	s_1	5.618	230.0	Mercury
g_2	7.456	174.0	s_2	7.080	183.0	Venus
g_3	17.365	74.6	s_3	18.851	68.7	Earth
g_4	17.916	72.3	s_4	17.748	73.0	Mars
g_5	4.249	305.0	s_5	0.000		Jupiter
g_6	28.221	45.9	s_6	26.330	49.2	Saturn
g_7	3.089	419.0	s_7	3.005	431.0	Uranus
g_8	0.667	1940.0	s_8	0.692	1870.0	Neptune

[a]Fundamental orbital frequencies of the precession motions in the Solar System, computed as mean values over 20 million years. The g_i and s_i are eigenvalues that characterize the evolution of the orbital elements (e, ω) and (i, Ω), respectively, and are loosely associated with the eight planets considered; i.e., the indices g_1, s_1 correspond to Mercury, and the indices g_8, s_8 correspond to Neptune. The period a, in years, can be calculated from the frequency, in arcseconds (″) per year, as a $(360 \times 60 \times 60)/''$ year ^{1}. Data from Laskar J (1990) The chaotic motion of the Solar System a numerical estimate of the size of the chaotic zones. *Icarus* 88(2): 266 291.

describe the orbital motion of Earth. The formation of an equatorial bulge is caused by the rotation of Earth, and other processes (e.g., the formation of ice-caps at high latitudes and mantle convection) redistribute mass on Earth. These processes result in a torque that is applied to Earth by the Sun, the Moon, and the other planets. Approximately two thirds of the torque is caused by the Moon, and one third by the Sun. Similar to a spinning top, this applied torque results in the nutation and precession of Earth's spin axis. Nutation is the short-term period portion (periods $\leq$18.6 years) of these variations; precession is the long-period portion. The nutational component leads to a 'nodding' motion of Earth's spin axis, with main periods of ~13 days, 6 months, 1 year, 9.3 years, and ~18.6 years, whereas the precessional component makes Earth's spin axis trace out a cone shape. The short-term nutation component that is superimposed on top of the longer term precession component is illustrated in Figure 2, and the precession component is illustrated in Figures 1 and 3. From a climatic perspective on geological time-scales, only the precession component has a significant effect; nutational variations result in small, mainly atmospheric effects.

With respect to the fixed stars, the frequency p of the precessional cycle has a period of approximately 25.8 ky. The precession of Earth's spin axis has several effects on Earth's climate system, one of which is that the position of the seasons with respect to Earth's orbit, defined by the solstices and equinoxes with respect to the perihelion and aphelion of the orbit, changes over time. For this reason, the precession of Earth's spin axis is also called the 'precession of the equinoxes'. As shown in Figure 1, the precession of Earth's spin axis traces out a cone shape that forms an angle with Earth's orbital plane. This angle, denoted by ε, is the obliquity (tilt) of Earth. The angle ε changes due to the combined effect of the precession of Earth's spin axis and the changing orientation of Earth's orbital plane (this will be discussed in more detail in a following section).

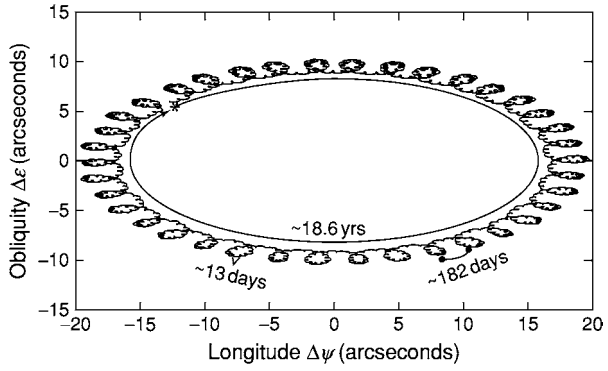


Figure 2 Short term nutation motion of Earth's spin axis, based on nutation and precession model IAU2000A of the International Astronomical Union. The nutation components in longitude and obliquity are with respect to the equinox and ecliptic of date. The figure is plotted for an ~18.5 year period, beginning on 1 January 2004.

As an approximation, the fundamental frequencies g_i and s_i can be used together with the precession constant p to explain the origin of almost all periodicities that affect the climate system, which arise from 'beats' between the fundamental frequencies. However, Jacques Laskar discovered that additional resonance terms are present in the Solar System, and these lead to the presence of chaos. The presence of chaos in the solar system has important consequences (see later). How the three orbital parameters, eccentricity, obliquity, and climatic precession, which are involved in the calculation of the solar radiation, are related to the fundamental frequencies of the Solar System is discussed in the following sections. The main parameters, known as 'Milankovitch cycles', are illustrated in Figure 3.

Eccentricity

Earth's orbital eccentricity e quantifies the deviation of Earth's orbital path from the shape of a circle. It is the only orbital parameter that controls the total amount of solar radiation received by Earth, averaged over the course of 1 year. The present eccentricity of Earth is $e \approx 0.01671$. In the past, it has varied between 0 and ~0.06. The eccentricity value can be used to compute the difference in the distance from Earth to the Sun between their closest and furthest approaches (perihelion and aphelion); presently, this amounts to $2e \approx 3.3\%$. At maximum eccentricity, the annual variation of solar insolation due to eccentricity is thus 24%. Although the exact values of orbital parameters should be computed by numerical integration, it is possible to approximate the calculation as a series of quasiperiodic terms, some of which are listed in Table 2. It is important to point out that the eccentricity frequencies are completely independent of the precession frequency p. Earth's eccentricity frequency component with the largest amplitude has a period of approximately 400 ky, which arises mainly from the interactions of the planets Venus and Jupiter, due to their close approach and large mass, respectively. This component is called the 'long' eccentricity cycle, and of all of Earth's orbital frequencies, it is considered to be the most stable. Additional terms can be found with periods clustered around ~96 and ~127 ky. These are called 'short' eccentricity cycles.

An important feature of all orbital components is the presence of 'beats'. These arise from the interaction between different frequency components and they produce a modulation in amplitude. This results, for example, in an amplitude modulation of the short eccentricity cycle, because the difference between the

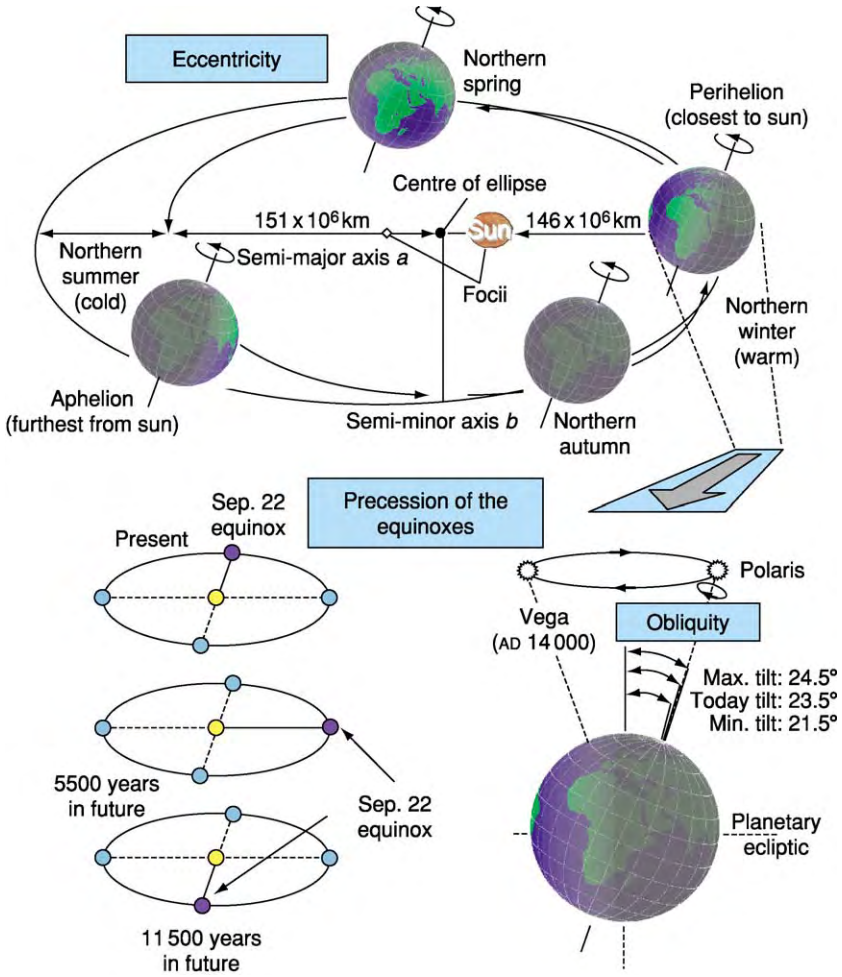


Figure 3 Orbital motions of Earth, showing the main Earth orbital parameters, eccentricity, obliquity, and precession of the equinoxes.

Table 2 Five leading terms for Earth's eccentricity[a]

Term	Frequency (″ year $^{-1}$)	Period (ky)	Amplitude
g_2 g_5	3.1996	406.182	0.0109
g_4 g_5	13.6665	94.830	0.0092
g_4 g_2	10.4615	123.882	0.0071
g_3 g_5	13.1430	98.607	0.0059
g_3 g_2	9.9677	130.019	0.0053

[a]Principal eccentricity frequency components in an astronomical solution analysed over the past 4 My. The frequency terms g_i refer to those given in Table 1.

Data from Laskar J (1999) The limits of Earth orbital calculations for geological time scale use. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical and Engineering Sciences* 357(1757): 1735 1759.

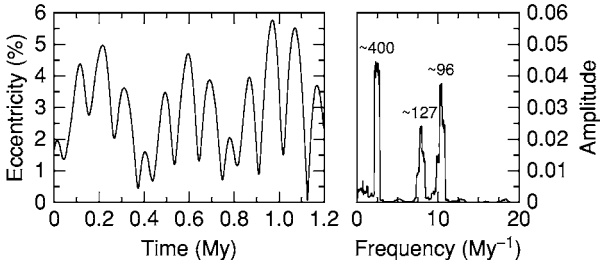


Figure 4 Earth's orbital eccentricity over time (1.2 million years) and frequency analysis for a 10 My time span. The peaks in the frequency analysis correspond to the frequencies given in Table 2; the numbers over the peaks represent the periods (in thousands of years).

second and third strongest eccentricity components is $(g_4 - g_5) - (g_4 - g_2) = (g_2 - g_5)$, which corresponds to the ~400-ky eccentricity cycle. The same modulation is observed for the fourth and fifth strongest terms. This type of amplitude modulation can be found in all orbital components of Earth. The nature of eccentricity variations is illustrated in Figure 4. The superposition of the long and short eccentricity cycles, and their variation in amplitude, are clearly visible. The right-hand side of the plot shows the

results of a frequency analysis, which was run over a 10-My-long interval to better resolve the position of individual peaks. The peaks correspond to the frequencies given in Table 2.

Obliquity

The obliquity (tilt) ε of Earth's axis with respect to the orbital plane (see Figure 1) is defined by the angle between Earth's spin vector **s** and that of the orbital plane **n**, and can be computed as $\cos\varepsilon = \mathbf{n} \diamond \mathbf{s}$, using unit vectors. As the inclination and orientation of the orbital plane vary, the obliquity is not constant, but oscillates due to the interference of the precession frequency p and the orbital elements s_i. As shown in Table 3, if the variation in obliquity is approximated by quasiperiodic terms, the result is a strong oscillation with a period of approximately 41 ky, with additional periods around 54 and 29 ky. The ~41-ky period arises from the simultaneous variation in Earth's orbital inclination, given by s_3, and the precession of Earth's spin direction, expressed by p. Table 3 also shows that the obliquity signal contains contributions from the g_i as well as the s_i fundamental frequencies, due to their combined effect on the change of the orbital plane normal. The present day obliquity of approximately 23.45° has varied between ~22.25° and ~24.5° during the past 1 million years. The main climatic effect of variations in Earth's obliquity is in control of the seasonal contrast. The total annual energy received on Earth is not affected, but the obliquity controls the distribution of heat as a function of latitude, and is strongest at high latitudes.

It is important to note that all of the obliquity frequency components contain the precession constant p. Due to tidal dissipation, the frequency of the precession constant p has been higher in the past, a fact that can be shown from geological observations, such as ancient growth rings in corals, and from tidal laminations. Figure 5 illustrates the variation in obliquity over 1.2 million years. The oscillation is dominated by a ~41-ky period cycle, and a variation in amplitude is also observed. This variation is due to beats arising from the presence of additional ~29- and ~54-ky periods, which are just visible in the frequency analysis shown on the right-hand side of Figure 5.

Table 3 Six leading terms for Earth's obliquity[a]

Term	Frequency (″ year $^{-1}$)	Period (ky)	Amplitude
$p + s_3$	31.613	40.996	0.0112
$p + s_4$	32.680	39.657	0.0044
$p + s_3 + g_4 \quad g_3$	32.183	40.270	0.0030
$p + s_6$	24.128	53.714	0.0029
$p + s_3 \quad g_4 + g_3$	31.098	41.674	0.0026
$p + s_1$	44.861	28.889	0.0015

[a]Principal obliquity frequency components analysed over the past 4 My. The frequency terms g_i and s_i refer to those given in Table 1.

Data from Laskar J (1999) The limits of Earth orbital calculations for geological time scale use. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical and Engineering Sciences* 357(1757): 1735 1759.

Climatic Precession

The precession of Earth's spin axis has a profound effect on Earth's climate, because it controls the timing of the approach of perihelion (the closest approach to the Sun) with respect to Earth's seasons. At present, perihelion occurs on the 4 January, close to the winter solstice. With respect to the stars, the precessional movement of Earth's spin axis traces out a cone shape with a period of ~25.8 ky. However, due to the precession of the perihelion within the orbital plane, the period of precession, measured with respect to the Sun and the seasons, is shorter. The motion of the perihelion is not steady but is caused by a superposition of the different g_i frequencies. For this reason, the precession of the equinoxes with respect to the orbital plane lurches with a superposition of several periods around ~19, 22, and 24 ky. The effect of the precession of the equinoxes on the amount of solar radiation received by Earth also depends on the eccentricity. If the eccentricity is zero (i.e., the orbit of Earth follows a circle), the effect of the precession of the equinoxes is also zero. From a climatic point of view, the eccentricity and longitude of the perihelion combine to create what is termed the 'climatic precession', defined as $e\sin(\tilde{\omega})$, where $e\sin(\tilde{\omega})$ is the longitude of perihelion from the moving equinox. This means that the climatic precession index is modulated in amplitude by variations in Earth's eccentricity.

A quasiperiodic approximation of the climatic precession time series reveals the contribution from different frequency components, as shown in Table 4.

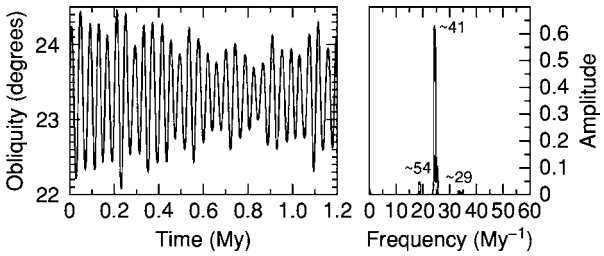


Figure 5 Earth's obliquity over time (1.2 million years) and frequency analysis for a 10 My time span. The peaks in the frequency analysis correspond to the frequencies given in Table 3; the numbers over the peaks represent the periods (in thousands of years).

Note that the components of climatic precession can be constructed from the precession constant p and the fundamental frequencies g_i. Figure 6 illustrates the variation in the climatic precession index, and its modulation in amplitude by eccentricity, over 1.2 million years. The frequency analysis, shown on the right-hand side of the plot, reveals three peaks corresponding to frequencies given in Table 4.

Insolation

Conceptually, the actual forcing of Earth's climate by orbital variations is applied through the radiative flux received at the top of the atmosphere at a particular latitude and time, which is then transferred through oceanic, atmospheric, and biological processes into the geological record. The integral of the radiative flux over a specified interval of time, termed 'insolation' (French from the Latin *insolare*), can be computed from the eccentricity e, the obliquity ε, and the climatic precession $e\sin(\tilde{\omega})$. The amount of solar radiation received at a particular location depends on the orientation towards the Sun of that location. Calculation of insolation becomes complex if it is to be calculated over a particular time interval. Averaged over 1 year and the whole Earth, the only factor that controls the total amount of insolation received, apart from the solar constant, is the changing distance from Earth to the Sun, which is determined by Earth's semi-major axis a and its eccentricity e.

If insolation variations are computed for a particular latitude, and over a particular length of time, the main contribution arises from the climatic precession signal, with additional contributions from the variation in obliquity. The exact nature of the insolation signal is complicated. Certain general statements can be made, though. The signal arising from the climatic precession is always present in insolation time series. Also, if the obliquity signal is present, it typically shows a larger amplitude towards high latitudes. In addition, the climatic precession signal in the insolation calculation depends on the latitude at which it is calculated, such that the signal in the southern-hemisphere summer shows opposite polarity to that in the northern-hemisphere summer (see Figure 7). If the mean monthly insolation is computed for a particular latitude, each month corresponds to a difference in phase (i.e., a difference in time of a particular insolation maximum or minimum) of approximately 2 ky (12 months approximately correspond to the (on average) ~24-ky-long climatic precession cycle).

It is unlikely that geological processes that encode the insolation signal are driven by variations at the same latitudes and times of the year throughout geological time. Depending on the latitude and the time interval over which insolation quantities are computed, the calculation can be very complex, and the question of time lags and forcing can be resolved only through the application of climate models. A very revealing study to this effect was reported by Short and colleagues in 1991. At the present level of understanding, it is probably appropriate to avoid a strict

Table 4 Five leading terms for Earth's climatic precession[a]

Term	Frequency (″ year $^{-1}$)	Period (ky)	Amplitude
$p+g_5$	54.7064	23.680	0.0188
$p+g_2$	57.8949	22.385	0.0170
$p+g_4$	68.3691	18.956	0.0148
$p+g_3$	67.8626	19.097	0.0101
$p+g_1$	56.0707	23.114	0.0042

[a]Principal climatic precession components analysed over the past 4 My. The frequency terms g_i refer to those given in Table 1. Data from Laskar J (1999) The limits of Earth orbital calculations for geological time scale use. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical and Engineering Sciences* 357(1757): 1735 1759.

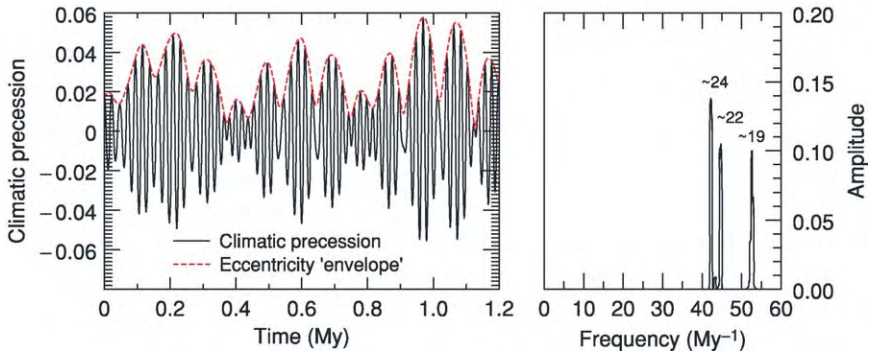


Figure 6 Climatic precession index and eccentricity envelope over time (1.2 million years) and frequency analysis for a 10 My time span. The peaks in the frequency analysis correspond to the frequencies given in Table 4; the numbers over the peaks represent the periods (in thousands of years).

interpretation of Milankovitch's theory, which would imply that the ice-ages are best explained by the summer insolation curve computed at 65° N. Instead, a better understanding of the complex mechanisms of the climate system will have to be achieved through the use of geological data providing boundary conditions for climate models.

Amplitude Modulation Patterns: The 'Fingerprint' of Orbital Cycles

A very important feature of Earth's eccentricity, obliquity, and climatic precession variations is that they display modulations in amplitude and frequency. These modulations provide a 'fingerprint' of a particular astronomical calculation at a given time. The modulation terms arise through the interference of individual cycles to produce 'beats', with periods ranging from hundreds of thousands to millions of years. The most prominent amplitude modulation cycles are listed in Table 5. An excellent visual representation of amplitude modulation cycles in astronomical calculations can be obtained by computing evolutionary or wavelet spectra, which show the variation in amplitude at different frequencies over time. This is shown in Figure 8 for the past 10 million years.

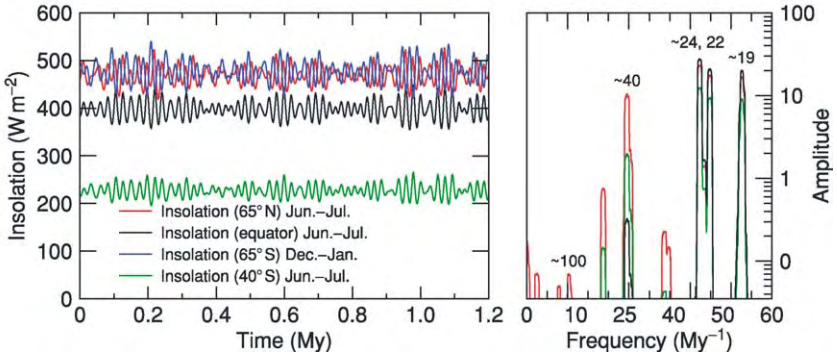


Figure 7 Earth's insolation over time (1.2 million years) at various latitudes and frequency analysis for a 10 My time span (on a logarithmic amplitude scale); the numbers over the peaks represent the periods (in thousands of years). The precession component is out of phase between the northern hemisphere summer and the southern hemisphere summer. Note that the eccentricity component in insolation computations is always very weak compared to the obliquity and precession components. Also note the latitudinal dependence of the relative strength of the obliquity cycle in the insolation curves.

Table 5 Modulation terms[a]

Type	*Interfering terms*	*'Beat' term*		*Period*
Short eccentricity amplitude modulation terms	$(g_4 \quad g_5) \quad (g_4 \quad g_2)$ $(g_3 \quad g_5) \quad (g_3 \quad g_2)$	$(g_2 \quad g_5)$	$\approx$	400 ky
	...	...		...
Short and long eccentricity amplitude modulation terms	$(g_4 \quad g_5) \quad (g_3 \quad g_5)$ $(g_4 \quad g_2) \quad (g_3 \quad g_2)$	$(g_4 \quad g_3)$	$\approx$	2.4 My
	...	...		...
Climatic precession amplitude modulation terms Identical to eccentricity frequencies and amplitude modulation terms				
Obliquity amplitude modulation terms	$(p+s_3) \quad (p+s_4)$	$(s_3 \quad s_4)$	$\approx$	1.2 My
	$(p+s_3+g_4 \quad g_3) \quad (p+s_3 \quad g_4+g_3)$	$(2g_4 \quad 2g_3)$	$\approx$	1.2 My
	$(p+s_3) \quad (p+s_3+g_4 \quad g_3)$ $(p+s_3) \quad (p+s_3+g_4 \quad g_3)$	$(g_4 \quad g_3)$	$\approx$	2.4 My
	$(p+s_3) \quad (p+s_6)$	$(s_3 \quad s_6)$	$\approx$	173 ky
	...	...		...

[a]Origin of amplitude modulation terms that affect Earth's eccentricity, obliquity, and climatic precession. Individual g_i and s_i terms refer to those given in Table 1. The ~100 ky (short) eccentricity cycles are modulated with a period of ~400 ky by the long eccentricity cycle. Both short and long eccentricity cycles are modulated with a period of 2.4 My, but with a phase difference of 180° (i.e., an amplitude maximum of the ~100 ky eccentricity coincides with an amplitude minimum of the ~400 ky eccentricity). Because eccentricity directly modulates the climatic precession, all eccentricity amplitude modulation terms are also present in the climatic precession signal. The obliquity signal is weakly amplitude modulated with a period of ~170 ky, and more strongly with a period of 1.2 My. This amplitude modulation cycle is dynamically linked with the 2.4 My cycle present in the eccentricity modulation. All period cycles listed here are shown in Figure 8.

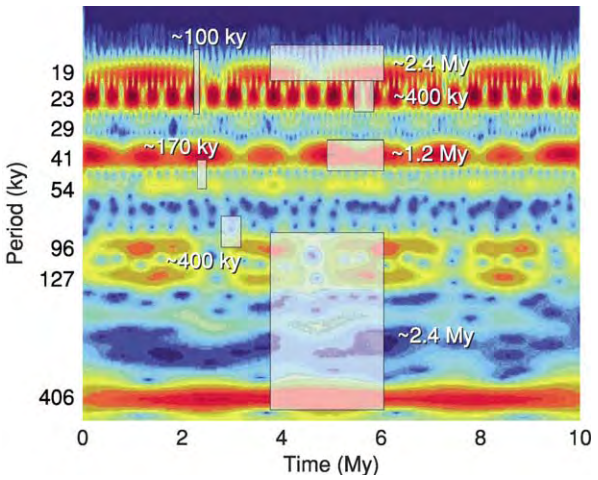


Figure 8 Joint time frequency analysis of an arbitrary mixture of the eccentricity, obliquity, and climatic precession signal to allow a better visual representation of the main climatically im portant orbital variations. The amplitude at a particular frequency and time is colour coded, red corresponding to a high relative amplitude. A selection of amplitude modulation terms is high lighted (boxes inside figure); these represent the 'fingerprint' of an astronomical model. The numbers near the fingerprints represent periods (see **Table 5**).

The significance of amplitude modulation cycles is twofold. First, if these cycles can be detected in the geological record, they allow the placement of geological data into a consistent framework within these amplitude modulation envelopes, even in the absence of individual cycles and in the presence of gaps. The extraction of long amplitude modulation cycles typically requires high-fidelity geological records that are millions of years long. Of particular value for the generation of geological time-scales beyond the Neogene is the ~400-ky-long eccentricity cycle, because it is considered to be very stable. In addition, if the eccentricity signal could be found in the geological data directly, as well as through its modulation of the climatic precession amplitude, it might be possible to evaluate phase lags between the astronomical forcing and the geological record. The second significant use of amplitude modulation cycles is that they are related to specific dynamical properties of a given astronomical model. These properties are related to the chaotic nature of the Solar System, and potentially allow the use of geological data to provide constraints on the dynamical evolution of the Solar System and astronomical models.

Tidal Dissipation and Dynamical Ellipticity

The mean fundamental orbital frequencies g_i and s_i, as well as the precession constant p, are likely to have changed throughout geological time. However, whereas the changes in g_i and s_i have probably been small, the precession constant p is likely to have changed significantly. Changes are caused by the effects of energy (tidal) dissipation as well as by redistribution of mass (due, e.g., to waxing and waning ice-caps and mantle convection). Changes in p are related to changes of the length of day, which is reflected in geological and palaeontological records. In particular, Earth's tidal response to the gravitational pull from the Sun and the Moon is not instantaneous. This means that the tidal bulge that develops on Earth (and on the Moon), is not aligned with the direction of the Moon's gravitational pull. This pull exerts a torque on Earth, which leads to a gradual decrease in its rotational velocity. In addition, conservation of angular momentum leads to an increase in the distance from Earth to the Moon over time, and a change in Earth's rotational velocity leads to a redistribution of mass on Earth ('dynamical ellipticity'). The dynamical ellipticity of Earth can also be affected by mantle convection and ice loading.

These processes modify Earth's precession constant p, the frequency of which has decreased over geological time. Because p is contained in the expressions for obliquity and climatic precession (see **Tables 3 and 4**), the periods of obliquity and climatic precession also change. In 1994, Berger and Loutre estimated possible values for changes of astronomical periods, based on astronomical and geological observations. Their results are illustrated in **Figure 9**. The effects of changing tidal dissipation and dynamical ellipticity values have a large impact on astronomical calculations, and have to be obtained from observation. Strictly speaking, astronomical calculations cannot be performed independently of the chosen Earth model, and particularly, there cannot be separate treatment of the Earth–Moon system. This is why numerical computations are invaluable.

Chaos in the Solar System

Probably the most significant development of astronomical theory in recent times has been the discovery of the chaotic behaviour of the Solar System by Jacques Laskar in 1990. Laskar established that the dynamics of the orbital elements in the Solar System are not fixed for all times, but rather are unpredictable over tens of millions of years. This is due to the non-linear gravitational interaction of the different bodies in the Solar System, which makes it theoretically impossible to calculate the exact movements of celestial bodies from their present-day masses, velocities, and positions over long periods of time. This feature poses limits on the use of astronomical theory for the purposes of creating astronomically calibrated

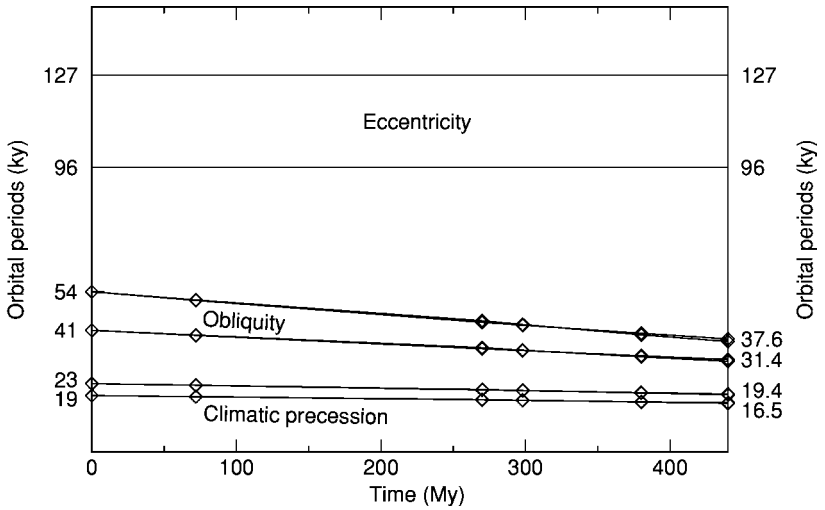


Figure 9 Estimated changes in the obliquity and climatic precession periods related to a changing distance between Earth and the Moon over the past 440 My, according to calculations by Berger and Loutre in 1994. Note that the periods of obliquity and climatic precession change due to a change of the precession constant *p*, whereas the eccentricity periods remain unaffected. Note that this diagram is for illustrative purposes only, and the exact variation of orbital frequencies over tens of millions of years is not yet known in detail.

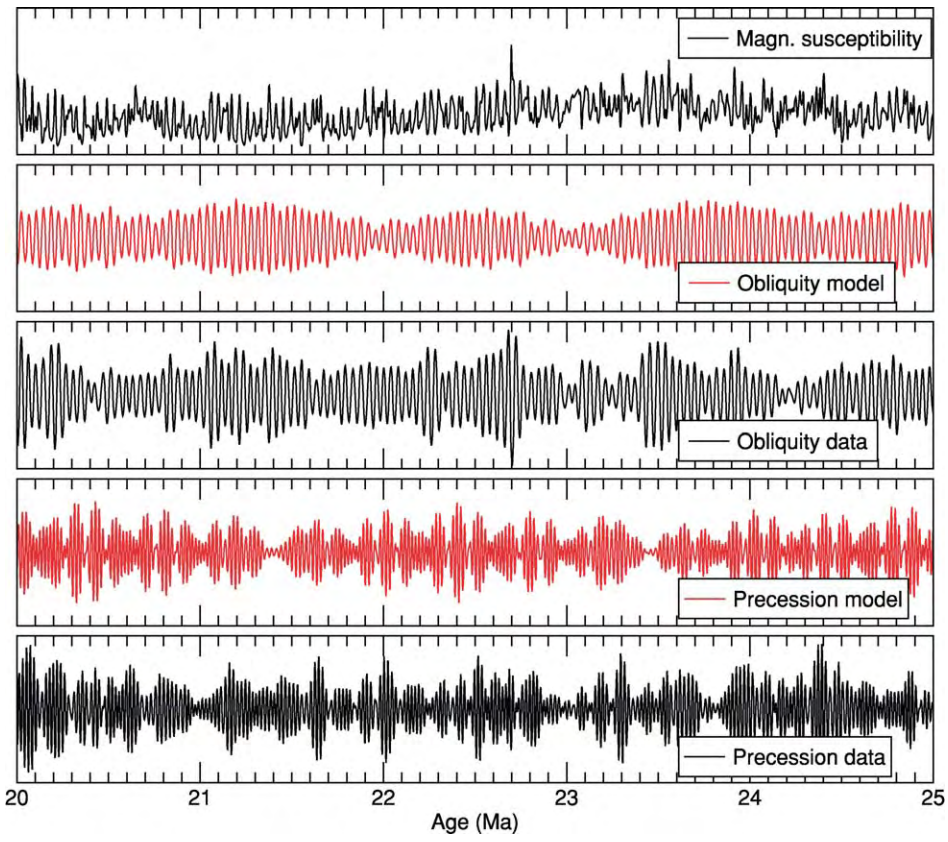


Figure 10 Lithological magnetic susceptibility data from Leg 154 of the Ocean Drilling Program. These data, analysed by Shackleton and colleagues in 1999, demonstrate the fidelity with which orbital variations can be encoded in the geological record. The magnetic susceptibility parameter, plotted together with filters that extract the visible cycles, chiefly reflects variations between terrestrial (clay rich) and marine (carbonate rich) sediments, being forced by a dominant obliquity signal.

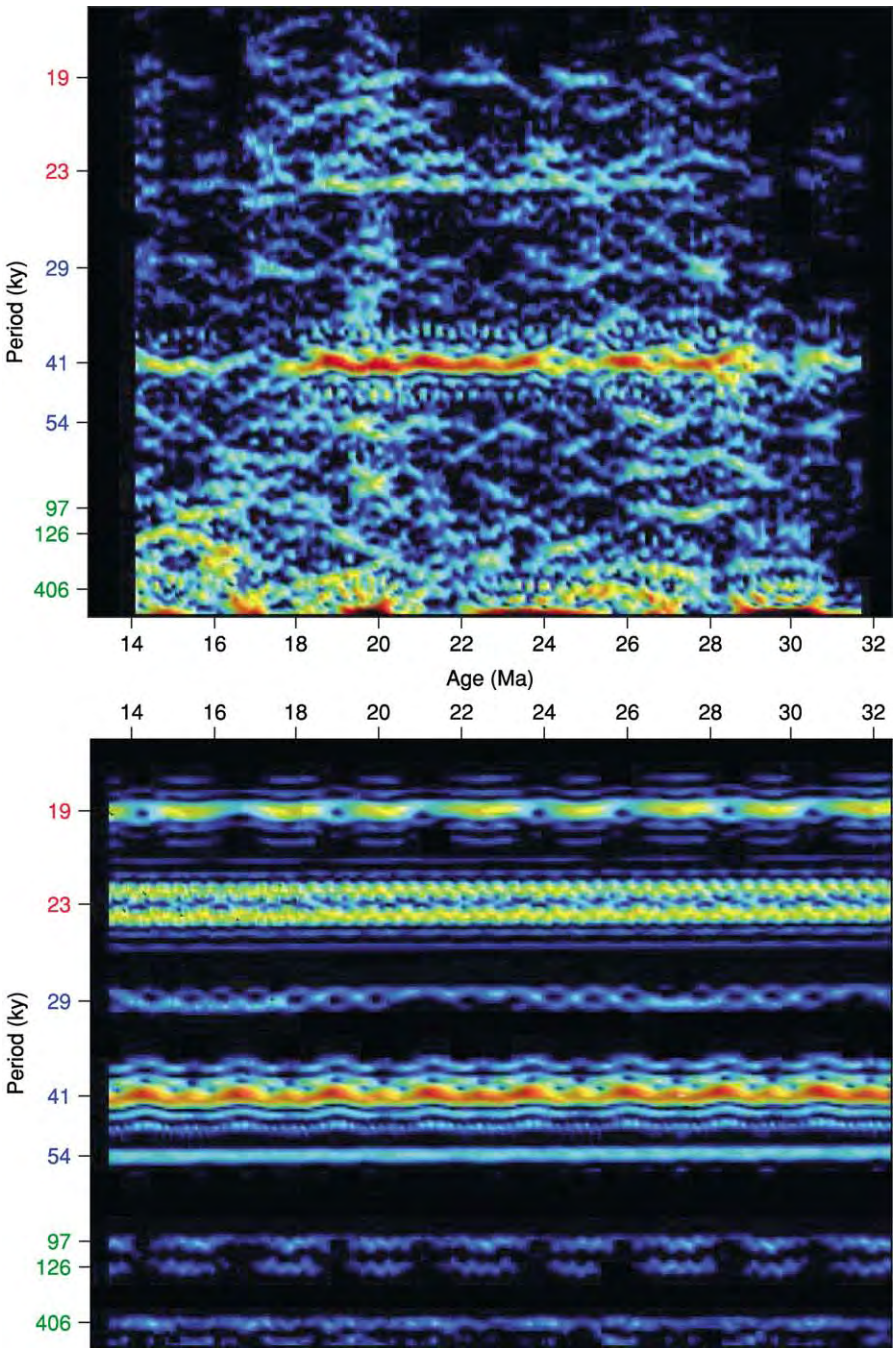


Figure 11 (Top) Evolutionary spectrum (frequency vs age, amplitude colour coded) of Ocean Drilling Program Leg 154 lithological data (combined magnetic susceptibility and colour reflectance) (cf. **Figure 10**). These data, analysed by Shackleton and colleagues in 1999, and contrasted to astronomical calculations of Earth's orbital variations, demonstrate that the geological data contain an imprint of long term amplitude modulations of astronomical forcing components. (Bottom) The astronomical solution provided by J Laskar.

geological time-scales. Laskar found that the calculations of orbits diverge exponentially with time for a given set of initial conditions. This implies, for example, that if the present position of a planet is known with a relative error of $\sim 10^{-10}$, this error increases to $\sim 10^{-9}$ after 10 My, and reaches the order of 1 after 100 My. Despite these limitations, the determination of astronomical parameters, such as the planetary masses, velocities, and positions, is improving due to additional satellite measurements.

Together with improved and more detailed astronomical models, the limits of accuracy of astronomical calculations are also likely to extend further back in time.

A clearer representation of chaos in the Solar System is provided by amplitude modulation terms. The ~2.4- and 1.2-My-long amplitude modulation terms that occur in the calculations of eccentricity and obliquity, respectively, are in resonance, and the expression $(s_4 - s_3) - 2(g_4 - g_3) = 0$ can evolve into a new state, where $(s_4 - s_3) - (g_4 - g_3) = 0$. This implies a change from a 1:2 resonance to a 1:1 resonance. Laskar found this behaviour to be the main source and representation of chaos in the Solar System. As shown in Table 5, these terms are present in several astronomical frequencies, and should be possible to detect in the geological record.

Earth's Orbital Variation Encoded in Geological Data

The imprint of Earth's orbital variation in geological records, first statistically demonstrated in 1976 for the recent geological past, in the seminal paper by Hays and colleagues, has now been found throughout parts of the Cenozoic and beyond, through variations of stable isotope measurements that act as a proxy for Earth's climate system, as well as in a large number of lithological parameters. A review of this body of data is beyond the scope of the present discussion, but it is illustrative to show at least one example of very high-quality data that demonstrate the imprint of Earth's orbital variations in the rock record. Figure 10 shows part of a record used to correlate geological data for the past ~30 My with astronomical calculations. The record shows exceptionally well-encoded obliquity and climatic precession cycles. Most importantly, this record also demonstrates the consistent variation in amplitude of the obliquity cycles. This can be illustrated with the help of evolutionary spectral analysis, whereby the relative amplitude at individual frequencies is evaluated at different times. This is shown in Figure 11.

Linking Earth's orbital variations (Milankovitch cycles) with geological records has caused controversies as to whether the theory that orbital variations driving Earth's climate conditions can be correct. The controversies stem from observations of palaeoclimatic proxies from the recent past that reveal an imprint that does not correspond to the expected strength of orbital variations in insolation calculations, and instead suggest a nonlinear response of the climate system at different orbital frequencies. In particular, during the past 800 ky, the imprint of eccentricity in stable isotope records has been much stronger than expected. It is now becoming clear that geological data show a wide variety of responses to individual orbital frequencies, depending on factors such as palaeolatitude, the prevailing oceanographic system at the study site, global ice volume, etc., with records showing much more variation than would be expected from a simple insolation calculation. A better understanding of the interaction between Earth's orbital variations and their imprint on the climate system and geological records is likely going to be gained from integrated Earth system modelling studies, making use of the growing body of observations that has been provided recently by ocean drilling. As a final note, orbital variations also affect the other planets of the Solar System, and recent attempts have been made to link these to climatic variations on Venus and Mars.

See Also

Analytical Methods: Geochronological Techniques. **Carbon Cycle**. **Earth Structure and Origins**. **Famous Geologists:** Agassiz. **Gaia**. **Magnetostratigraphy**. **Microfossils:** Foraminifera. **Palaeoclimates**. **Solar System:** The Sun; Asteroids, Comets and Space Dust; Meteorites; Mercury; Venus; Moon; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus. **Tektites**. **Tertiary To Present:** Pleistocene and The Ice Age. **Time Scale**.

Further Reading

Berger A, Imbrie J, Hays J, Kukla G, and Saltzman B. (eds.) (1984) *Milankovitch and Climate: Understanding the Response to Astronomical Forcing*. Dordrecht and Boston: D. Reidel Publishing Company.

Berger A and Loutre MF (1994) Astronomical forcing through geological time. In: de Boer PL and Smith DG (eds.) *Orbital Forcing and Cyclic Sequences (IAS Special Publication)*, vol. 19, pp. 15 24. Oxford: Blackwell Scientific.

Berger A, Loutre MF, and Tricot C (1993) Insolation and Earth's orbital periods. *Journal of Geophysical Research* 98(D6): 10341 10362.

de Boer PL and Smith DG (eds.) (1994) *Orbital Forcing and Cyclic Sequences (IAS Special Publication)*, vol. 19. Oxford: Blackwell Scientific Publications.

Croll J (1875) *Climate and Time in their Geological Relations: A Theory of Secular Changes of the Earth's Climate*. London: Daldy, Tsbister and Co.

Einsele G, Ricken W, and Seilacher A (eds.) (1991) *Cycles and Events in Stratigraphy*. Berlin: Springer Verlag.

Gilbert GK (1895) Sedimentary measurement of Cretaceous time. *Journal of Geology* III: 121 127.

Hays JD, Imbrie J, and Shackleton NJ (1976) Variations in the Earth's orbit: pacemaker of the Ice Ages. *Science* 194(4270): 1121 1131.

Hinnov LA (2000) New perspectives on orbitally forced stratigraphy. *Annual Review of Earth and Planetary Science* 28: 419 475.
House MR and Gale AS (eds.) (1995) *Orbital Forcing Timescales and Cyclostratigraphy, Geological Society Special Publication*, vol. 85. London: The Geological Society.
Imbrie J and Palmer Imbrie K (1979) *Ice Ages; Solving the Mystery.* Short Hills, NJ/Cambridge, MA: Enslow Publishers/Harvard University Press.
Lambeck K (1980) *The Earth's Variable Rotation: Geophysical Causes and Consequences.* Cambridge: Cambridge University Press.
Laskar J (1990) The chaotic motion of the solar system a numerical estimate of the size of the chaotic zones. *Icarus* 88(2): 266 291.
Laskar J (1999) The limits of Earth orbital calculations for geological time scale use. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical and Engineering Sciences* 357(1757): 1735 1759.
Milankovitch M (1941) *Kanon der Erdbestrahlung und seine Anwendung auf das Eiszeitenproblem, Special Publication 132*, vol. 33. Belgrade: Royal Serbian Sciences. ('Canon of Insolation and Ice Age Problem', English translation by Israël Program for Scientific Translation and published for the US Department of Commerce and the National Science Foundation, Washington, DC, 1969).
Schwarzacher W (1993) *Cyclostratigraphy and the Milankovitch Theory.* Amsterdam: Elsevier.
Shackleton NJ, Crowhurst SJ, Weedon GP, and Laskar J (1999) Astronomical calibration of Oligocene Miocene time. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical Physical and Engineering Sciences* 357(1757): 1907 1929.
Shackleton NJ, McCave IN, and Weedon GP (eds.) (1999) Astronomical (Milankovitch) calibration of the geological timescale (9 10 December 1998, London) *Philosophical Transactions of the Royal Society of London, Series A, Mathematical Physical and Engineering Sciences* 357(1757).
Short DA, Mengel JG, Crowley TJ, Hyde WT, and North GR (1991) Filtering of Milankovitch cycles by Earth's geography. *Quaternary Research* 35(2): 157 173.

EARTH STRUCTURE AND ORIGINS

G J H McCall, Cirencester, Gloucester, UK

Introduction

The Earth is a planet of the inner zone of the solar system, which is occupied by four 'terrestrial' planets – Mercury, Venus, Earth, and Mars from the Sun outwards. These planets are composed mainly of rock and metal, whereas the outer planets – the so called 'gas-giants' and 'ice-giants' – have immense gaseous envelopes. The four differ greatly in character. Mercury has a dead cratered surface closely resembling that of the Moon, though there is evidence that it is compositionally very different. Venus is shrouded by a dense greenhouse (carbon dioxide-rich) atmosphere that is not visually penetrable, but radar images reveal a largely volcanically moulded surface. Mars, like Venus and Earth, has a very complex surface, moulded by volcanism, impacts, and wind action, and has polar ice caps. The Earth is unique in possessing oceans that cover the majority of the surface and form part of the hydrosphere, which includes surface and underground water bodies. The Earth is also unique in possessing an oxygen-containing atmosphere dominated by nitrogen; Mars has a thin carbon dioxide-rich atmosphere with negligible oxygen, and Venus has a thick carbon dioxide-rich atmosphere. Mercury and the Moon have negligible atmospheres dominated by inert gases. The unique character of the Earth is well seen from Space (**Figure 1**): the oceans appear blue in reflected sunlight, and the single satellite is uniquely

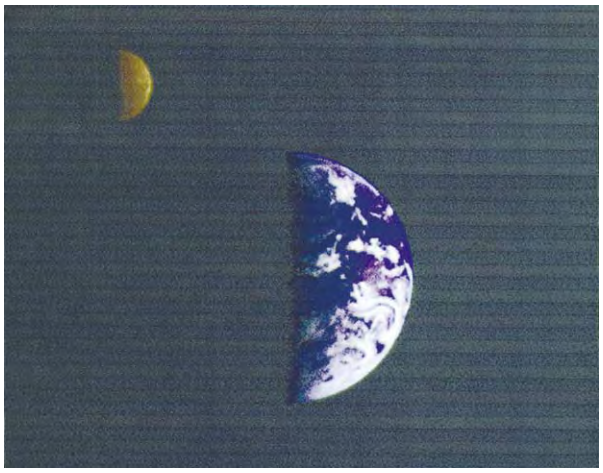


Figure 1 The Earth and Moon as seen from Space (NASA image kindly supplied by Sir Patrick Moore).

large compared with the mother planet, and uniquely orbits it at the same rate as it rotates so that the same lunar face is always Earthwards. Of the other inner planets, only Mars has moons (two of them), and they may be captured asteroids. The origin of the Earth's Moon remains highly controversial despite vigorous championing of various models.

The Unique Biosphere

As far as we know, the Earth is unique in having a biosphere on and below its surface, in its waters, and in the air above it – a quite thin zone compared with the planetary interior. Minute living organisms may be carried naturally high in the atmosphere, and the exact upper boundary of the biosphere is not closely defined. In both the oceanic and continental realms, bacteria can be found deep beneath the surface of the rocks: 92% of the Earth's bacteria occupy the deep biosphere, and they can live at depths of up to 4 km and at temperatures of up to 90°C.

Properties

The essential properties of the Earth are given in Table 1, where they are compared with those of the other inner planets and the Moon (*see* **Solar System:** Moon).

The Earth is not perfectly spherical but is an oblate spheroid; this was recognized by Newton to explain the precession of the equinoxes, the shift of the equinoctial point eastwards each year. The figure of 1/298.25 is widely accepted for the degree of flattening of the polar diameter compared with the equatorial diameter. The Earth is subject to various secular changes that affect the length of the day; these include secular orbital changes and rotational changes, amounting to several parts per 10^8 over time-scales ranging from days to millennia. Besides the rate of rotation, there are secular changes related to the celestial frame. Polar motion is a combination of the effect of the Chandler Wobble (departure from pure spin amounting to an axial deviation of 4″ of arc with a 428 day cycle) and a contributing effect of variation of the distribution of mass within the Earth and in the oceans and atmosphere. The precession of the equinoxes mentioned above is due to lunar and solar gravity acting on an oblate spheroid with an axis of rotation inclined at 23.5° to the perpendicular. There is also planetary precession, which reduces the obliquity of the axis to the ecliptic (the mean plane of the orbit around the sun) by 0.5″ per year. The combined precession has a period of 25870 years. The Milankovich model of climate change is generally accepted to be related to orbital eccentricity, obliquity of the ecliptic, and the precession of the equinoxes (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)), and this has a periodicities of 90 000–110 000, approximately 42 000 and approximately 26 000 years. This model is believed to control glaciations in the geological record. The world's temperate zone is advancing into the tropics, and this is the Milankovich cycle in operation before our eyes. All these secular changes complicate extremely accurate calculations of time and geodetic surveying, and there is the further complication of magnetism, in particular the Earth's magnetic field, which is at present subject to increasing variation (see below).

Table 1 The essential properties of the Earth, the other inner planets, and the Moon

	Radius (km)	*Mass (10^{20} kg)*	*Density (10^3 kg m $^{-3}$)*
Mercury	2439	3300	5.4
Venus	6051	48700	5.3
Earth	6371	59970	5.517
Moon	1738	734.9	3.34
Mars	3394	6420	3.9

Earth Tides

The gravitational attraction within the Earth–Moon system keeps them in orbit about the common centre of mass, a point about 4670 km from the centre of the Earth. The pull on the side of the Earth facing the Moon at any time is strongest, and the result is a tendency to pull the Earth into a prolate ellipsoid aligned with the Earth–Moon axis (Figure 2). The

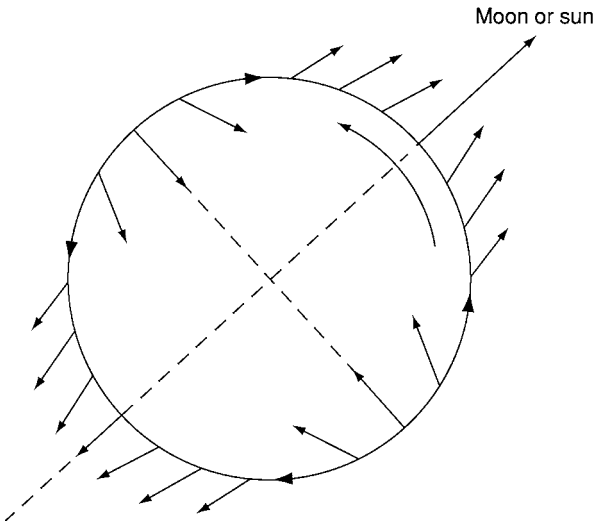


Figure 2 The pattern of tidal forces at the surface of the Earth resulting from the gravitational pull of the Moon and Sun (reproduced with permission from F D Stacey (2000) Earth Tides. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 280 281. Oxford: Oxford University Press.).

variation in gravitational pull imposed by the rotation of the Earth causes Earth tides. There is also a smaller contribution to Earth tides by the Sun.

Internal Configuration

What we know of the internal configuration of the Earth comes from the following sources:

- study of surface rocks and rocks brought up by eruptive processes (kimberlite diatremes are particularly important with respect to the mantle as they sample regions within it that would otherwise be unknown to us directly. (*see* **Volcanoes; Igneous Rocks:** Kimberlite);
- analogies with meteorites (in particular between nickel–iron meteorites and the core, and between carbonaceous chondrites and the mantle);
- instrumental measurements (geophysical), especially seismology (Figure 3); and
- geochemical research (relating to the Earth, other planetary bodies, and meteorites).

As a result of these approaches it is accepted that the Earth is made up of a series of concentric shells, each with its own physical and chemical properties (Figure 4). The outermost is the crust, beneath it is the mantle (divided into an upper mantle and lower mantle), and at the centre is the core (divided into the inner core and outer core). The plate-tectonics paradigm also accepts the existence of a lithosphere, which is not synonymous with the crust, but is a solid and rigid zone extending into the upper part of the mantle and bounded below by a transition zone. This separates the strong elastic layer from the weak and ductile asthenosphere, which is solid but very slowly convecting. The crust is bounded below by a discontinuity at which the velocity of seismic P waves increases abruptly; this is the Mohorovicic discontinuity and forms the lower boundary of the zone of isostatic adjustment (*see* **Earth:** Mantle; Crust).

The properties of the various concentric zones of the Earth are summarized in Table 2.

Magnetic Field

The Earth's magnetic field is environmentally critical because it forms a protective shield that blocks out solar radiation and harmful incoming particles. It has the character of a dipole, but has, throughout geological time, been subject to reversals, with the poles

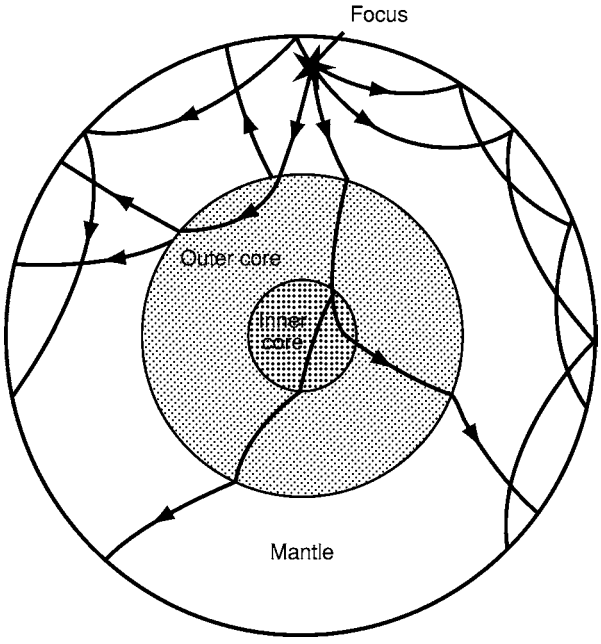


Figure 3 Earthquake energy travels down and out from the focus as shown, but at each internal boundary part of it is reflected or refracted and returns to the surface. Times of travel of the energy waves and their behaviour have given us much of the information that we have about the depths of the boundaries and the nature of the material forming the concentric shells of the Earth (reproduced from Van Amdel TJ (1994) (eds.) *New Views on an Old Planet*, 2nd edn. Cambridge: Cambridge University Press, with permission from the author and publisher).

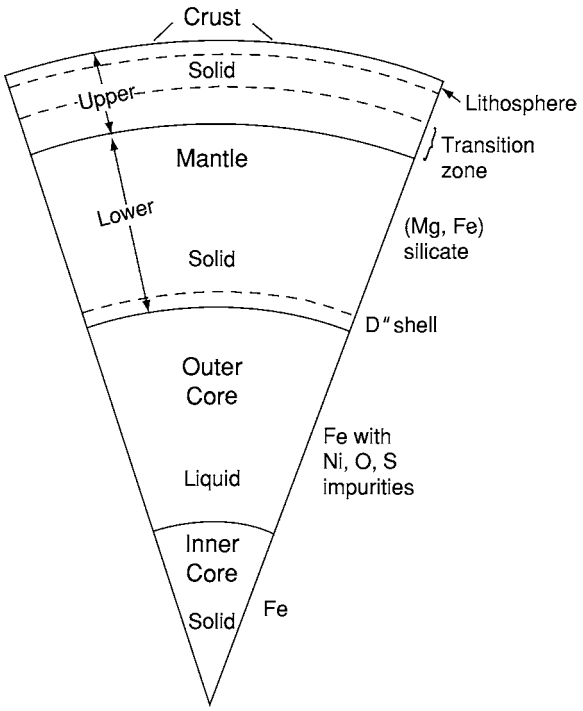


Figure 4 The configuration of the Earth's internal zones as deduced from the four indirect sources listed in the text (reproduced with permission from C M R Fowler (2000) Mantle convection, plumes, viscosity and dynamics. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 649 652. Oxford: Oxford University Press).

Table 2 Volumes, masses, and densities of the various concentric shells of the Earth (from C M R Fowler (2000) Earth Structure. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 276 280. Oxford: Oxford University Press)

	Depth (km)	*Volume (10^{18} m^3)*	*Volume (% of total)*	*Mass (10^{21} kg)*	*Mass (% of total)*	*Density (10^3 kg m^{-3})*
Crust	0 Moho*	10	0.9	28	0.5	2.60 2.90
Upper mantle	Moho 670	297	27.4	1064	17.8	3.35 3.99
Lower mantle	670 2891	600	55.4	2940	49.2	4.38 5.56
Outer core	2891 5150	169	15.6	1841	30.8	9.90 12.16
Inner core	5150 6371	8	0.7	102	1.7	12.76 13.8
Whole Earth		1083	100	5975	100	

*The Moho is 25 35 km deep under the continents and 6 7 km deep under the oceans.

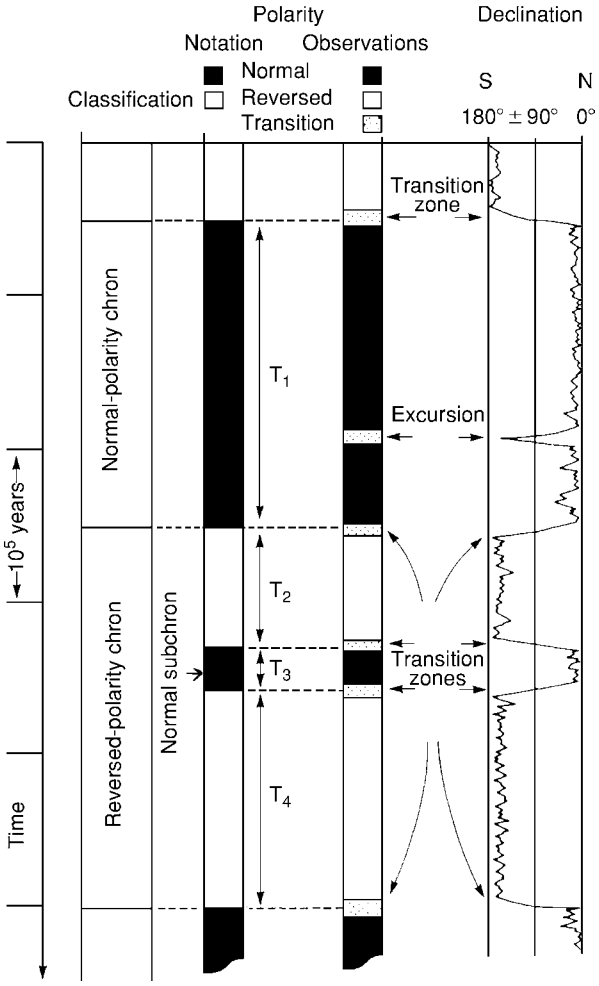


Figure 5 Magnetic polarity divisions of geological time (chrons, subchrons, transitional zones, and excursions) (reproduced from Harland WB, Armstrong RL, Cox AV, *et al.* (1990) *A Geologic Timescale 1989*. Cambridge: Cambridge University Press).

switching positions (Figure 5). Such reversals may occur over periods of a few hundred thousand years, but there have been longer periods without reversal of around 10 Ma. The intensity of the dipole field varies, being particularly strong during long stable periods and weakening just before reversal. The position of the north magnetic pole, first located by James Clark Ross in 1831 in the Boothia Peninsula, north-east Canada, shows secular displacement and is at present accelerating northwards towards Siberia (Figures 6A and 6B). The magnetic field is believed to be due to convection in the liquid outer core (Figure 7) (*see* **Magnetostratigraphy**).

Ozone Layer

There is a layer of ozone (O_3), a blue-green poisonous gas, in the stratosphere at an altitude of between 15 and 40 km (Figure 8). This is important because it absorbs the carcinogenic part of the solar spectrum (ultraviolet B). It is produced by photochemical reactions, and its concentration displays natural variations, being more abundant at the poles and less so in equatorial regions, though it is created at the equator and destroyed at the poles, where the concentration shows seasonal variation. Chlorofluorocarbon compounds, used as refrigerants and aerosol sprays, are believed to threaten the ozone layer; stratospheric jet planes would also do so, but relatively few are in use.

Plate Tectonic Movement and Mantle Convection

Space exploration has revealed no other body (planet, satellite, or asteroid) with evidence of lateral plate-tectonic displacement attributable to mantle convection, with accompanying ridge eruptivity, spreading away from ridges, subduction at plate margins, and continental collision (Figure 9). The upper mantle, though solid, is capable of very slow thermal convection over the vast periods of geological time. Mantle convection and plate tectonics probably did not operate in the Archaean and earliest Proterozoic (before 2000 Ma), though some minor shuffling of numerous small plates may have occurred. There is evidence that it operated through most of the Proterozoic and Phanerozoic, and the Earth's surface at present consists of relatively few

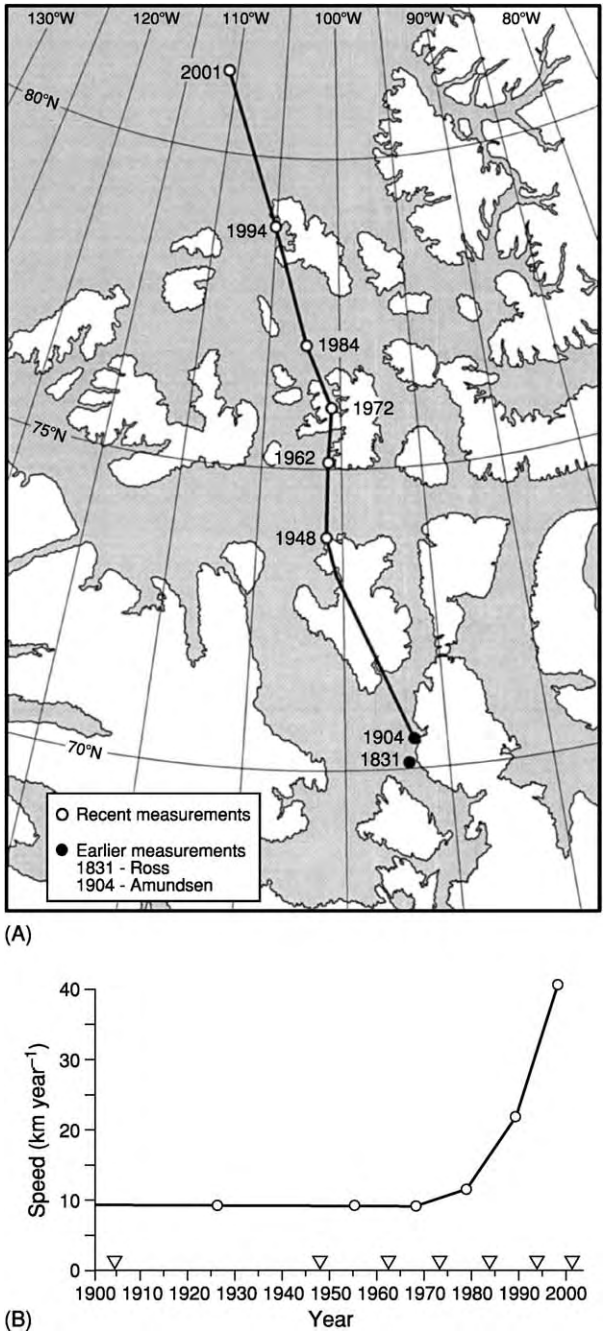


Figure 6 (A) The movement of the north magnetic pole since James Clark Ross located it in 1831. (B) The accelerating movement of the pole at the present time, which could herald a reversal (reproduced from McCall GJH (2003) Pole up the pole. *Geoscientist* 13: 9, with the permission of the Geological Society Publishing House).

plates, some of them immense (for example the one includes India, the Indian Ocean, Australia, and the Tasman Sea) (Figure 10). That plate movement has occurred and is occurring cannot be seriously doubted, though attempts at instrumental measurements have

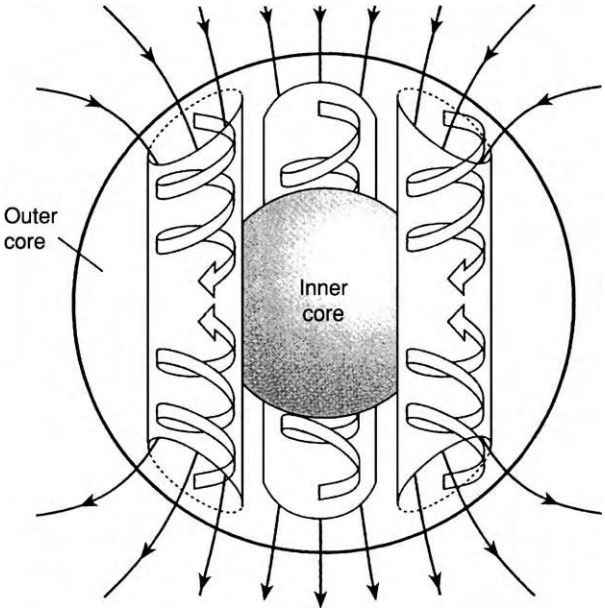


Figure 7 The magnetic field of the Earth is due to convection in the outer core, for which the configuration is not determined (reproduced from D Ravat (2000) Magnetic field, origin of the Earth's internal field. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 630 631. Oxford: Oxford University Press).

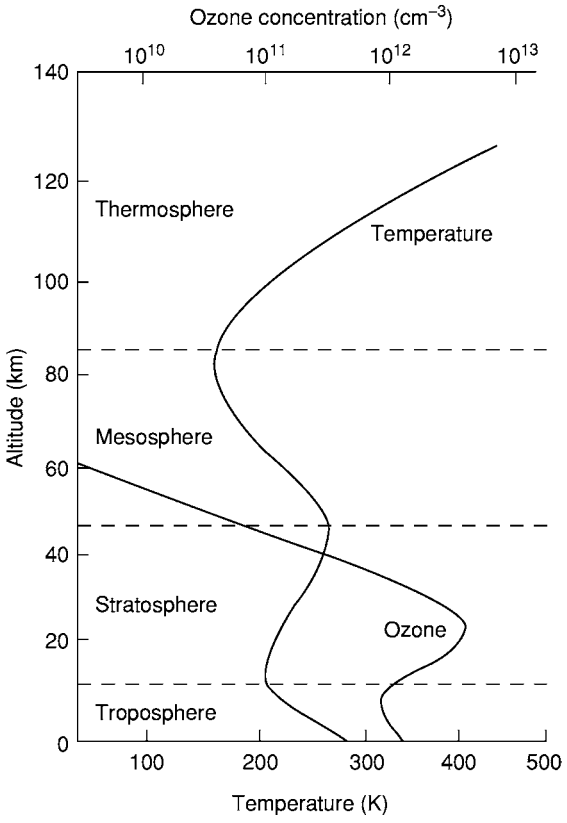


Figure 8 The configuration of the ozone layer (reproduced from F Drake (2000) Ozone layer chemistry. In: Hancock PL and Skinner BJ *Oxford Companion to the Earth*, pp. 772 773. Oxford: Oxford University Press).

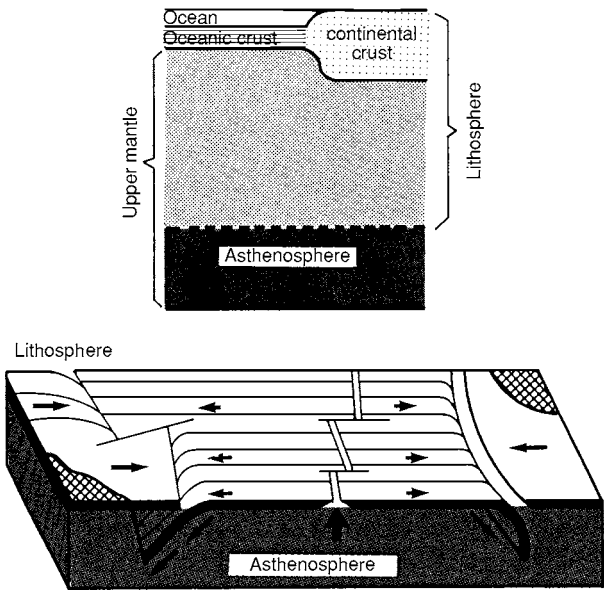


Figure 9 The widely accepted configuration of plate movement away from a mid ocean ridge and down subduction zones beneath continents (reproduced from Van Andel TJ (1994) *New Views on an Old Planet*, 2nd edn. Cambridge: Cambridge University Press, with permission from the author and publisher).

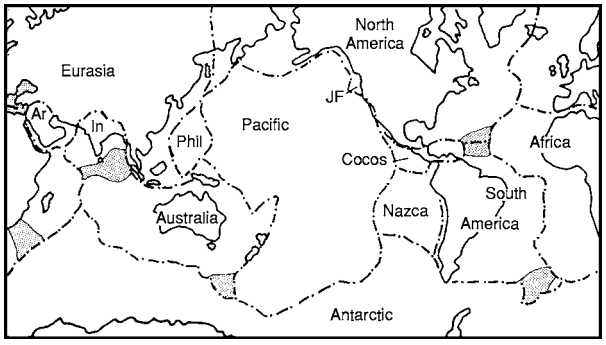


Figure 10 The tectonic plates of the Earth, Ar Arabia; Phil Philippines (reproduced from Van Andel TJ (1994) *New Views on an Old Planet*, 2nd edn. Cambridge: Cambridge University Press, with permission from the author and publisher).

yielded conflicting results and suggested differences in movements at margins and within plates. There are conflicting opinions as to whether convection operates down into the lower mantle (the more orthodox view) or only above the 670 km boundary, which forms a barrier. The latter view ('closed upper-mantle circulation') is held by D L Anderson and W B Hamilton and considers the reality or otherwise of deep plumes, rising from the base of the mantle. Plumes were originally invoked to explain chains of seamounts and/or islands such as the Emperor Seamount and Hawaiian Islands in the Pacific Ocean, where the ages of the individual islands become progressively younger towards the south-east. Oceanic lithosphere is hypothesized to move over a static hotspot. Sceptics believe that the 670 km barrier cannot be penetrated by a rising plume (*see* **Mantle Plumes and Hot Spots**). There is also no agreement on the mechanism of plate movement. Ridge push, subducting-slab pull, and a combination of both have all been favoured, as has a mechanism of gravity gliding. The plate movement must start with spreading away from the ridge before subduction begins to operate (no slab yet to pull), so slab pull can never provide the entire answer. It seems fair to say that, whereas the plate-tectonics paradigm must reasonably be accepted as fact, the mechanism remains unexplained, and mathematical calculations do not really explain how the forces needed to move the plates are derived (*see* **Plate Tectonics**).

Geochronological Comparisons – Earth and Other Solar System Bodies

It is early days yet in Space exploration, so, whereas we can derive a very detailed and accurate chronology for our own planet, based on the radioactive decay of isotopes, palaeontology, palaeomagnetism etc., we have very little chronological information about other bodies. There is no palaeontological data for other bodies (there may have been no life there). We have isotope-based data for the Moon, but there is really no geochronological data from there for the last 3000 Ma or so. Palaeomagnetism cannot be applied to bodies other than the Earth. There is a set of dates for the formation of the rocks comprising the SNC meteorites (shergottites, Nakhlites and chassignites), which are supposed to have come from Mars, and these have contributed to a Martian geochronology (which may yet prove to be wildly incorrect). Isotopic dates of around 4500 Ma for the formation of the asteroids have been derived from chondritic meteorites, and slightly younger dates, recording a slightly later melting process in the asteroidal parent bodies, have been derived for achondritic meteorites (*see* **Solar System:** Meteorites). For Venus and Mercury we have no yardsticks at all, although surmises have been made from the radar-derived images of Venus. Mercury resembles the Moon in its surface cratering and appears, like the Moon, to be a 'dead' multicratered body with no or negligible eruptivity since the deduced bombardment that formed the craters (*see* **Solar System:** Mercury; Venus). Sequential relationships without any benchmarks have been derived from the superposition orders of crater populations and volcanic structures. The comparison diagram (Figure 11) utilizes the very limited evidence from outside the Earth to make this comparison.

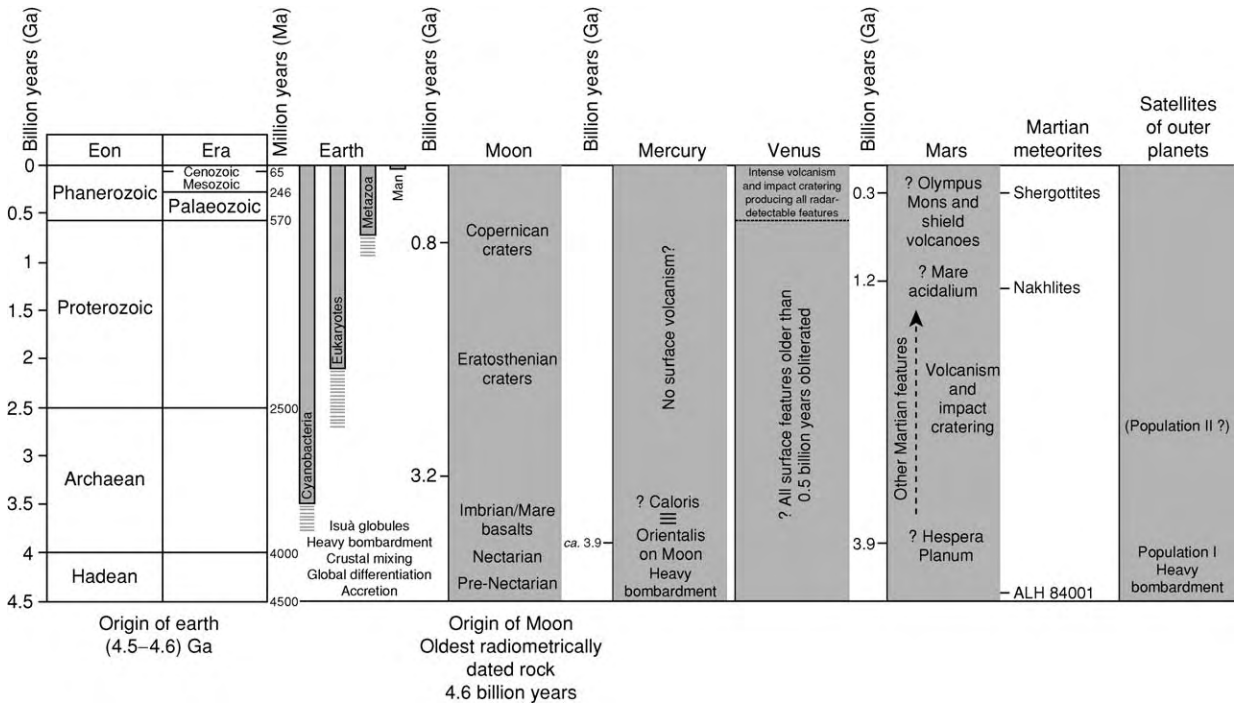


Figure 11 A comparison between the eons and eras of the geological column and sequences of events that have been deduced for the other inner planets, the Moon and satellites of the outer planets on the basis of the very limited evidence available from Space and meteoritical research (modified from G J H McCall (2000) Age and early evolution of the Earth and Solar System. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 8 11. Oxford: Oxford University Press).

The Origin of the Earth

The Earth was formed from the nebula that produced the Solar System. It is almost universally accepted that the Sun, the planets and their satellites, the asteroids, and the comets of the Oort 'cloud' grew from a cloud of gas and dust that contracted under its own gravity. The cloud of gas and dust had some degree of rotation (Figure 12) and, as the centre contracted, the angular momentum forced the remains into a flattened disk, in the plume perpendicular to the axis of rotation of the proto-Sun. This was apparently a very rapid process, perhaps taking place over 10 000 years. The dust particles accreted heterogeneously to produce lumps that formed planetesmals, the first small solid bodies, and cooling occurred.

We can estimate the composition of the original nebula from the solar abundances of elements, obtained spectrographically, and analysis of primitive meteorites (e.g. carbonaceous chondrites) (*see* **Solar System:** Meteorites); however, these are only estimates, and a second method is based on an assumption that may not be entirely accurate. Radioactivity-based dating methods (relying on the decay of radiogenic isotopes), as applied to terrestrial igneous rocks, allow us to date the condensation to 4500 Ma ago (the so-called 'age of the Earth'), and it is believed that condensation was rapid, taking about 100 000 years.

Accretion of the Earth and the other planets may have been slow and heterogeneous or rapid and homogeneous. Which of these two models is correct remains uncertain, but the first model, if correct, could have produced a layered mantle (*see* **Earth:** Mantle).

After condensation and accretion, the Earth evolved rapidly into a planet with most of its present properties. This could have taken no more than 500 Ma and was probably effected more rapidly: the oldest known minerals are zoned zircons from the Mount Narryer gneiss terrain, Western Australia, which give spot SHRIMP ages of 4400 Ma and are believed to have been formed by partial melting of pre-existing granitic crust (*see* **Minerals:** Zircons). This evidence indicates that there was solid rock at the Earth's surface astonishingly soon after condensation and accretion.

The oldest actual rock dated is a component of the Acasta Gneiss in the Slave Province of Canada, which has an age of 4000 Ma. Despite the existence of the Mount Narryer minerals, there is no other really significant evidence from geology of the Earth pertaining to the 500 Ma prior to the formation of the Acasta Gneiss, and this is called the 'Hadean Eon',

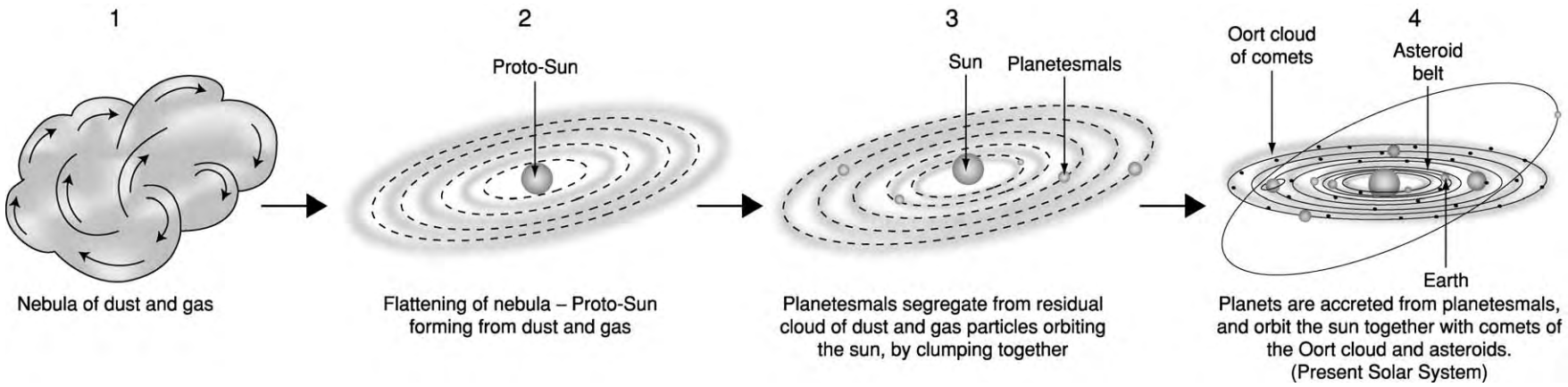


Figure 12 The development of the Earth as a planet in the Solar System from the nebula.

geology's dark ages. We do know something of this period from the Moon – the lunar rocks brought back by Apollo and Lunik date from this period – and it is reasonable to invoke a bombardment of the Earth, like the Moon, during this period. However, there is no trace of this in the geological record of the Earth. The lack of evidence of ejecta from the Moon in Space means that it is difficult to attribute all the heavy early cratering of the Moon to bombardment by impacting objects from Space (*see* **Solar System:** Meteorites). Nevertheless, the early Earth must have swept up asteroids, comets, and any other debris in its path for a few million years after it attained its present size.

The Earth, after accretion, must have heated up internally and melted in order to separate, gravitationally, the core (part liquid and part solid), mantle, and crust. This heating would have been caused by the radioactive decay of elements and by gravitational energy – the latter as the dust, or dust and planetesmals, condensed and accreted into a ball. It has also been suggested that heat could have been produced by impacts on the surface, though much of this heat would have dissipated into Space. The heavy elements, mainly iron, would have gravitated to the core, and the lighter silicate material would have ended up in the mantle and crust. A primitive and highly mobile surface layer, perhaps a magma ocean, would have initiated the crust, and from this small-scale proto-continents separated. More 'way-out' models for the formation of the crust invoke early impacts, the subsidence of large impact-generated basins in the basaltic proto-crust, and partial melting of the basalt to form the proto-continents. The early proto-continents were certainly small, and the extent of the continents increased until around 3200 Ma, when the Kaapvaal (South Africa) and Pilbara (Western Australia) cratons were formed.

The very lightest volatile elements would have separated at the same time to form the proto-ocean and proto-atmosphere. This volatile component would have had a bulk composition of 10–20% water and contents of other volatiles in line with those of carbonaceous chondrite meteorites, according to the chondritic Earth model. The first atmosphere was almost certainly reducing rather than oxygenating. Our present oxygen-containing atmosphere was derived later, though how much later is debatable, and relates, in its origin, to the onset of biological activity. How and when life originated is still a mystery – there are models that suggest that it originated within the early developing Earth, but there are also models that suggest that it arrived on the planet from comets or primitive meteorites (*see* **Origin of Life**).

See Also

Earth: Mantle; Crust; Orbital Variation (Including Milankovitch Cycles). **Igneous Rocks:** Kimberlite. **Magnetostratigraphy**. **Mantle Plumes and Hot Spots**. **Minerals:** Zircons. **Origin of Life**. **Plate Tectonics**. **Solar System:** Meteorites; Mercury; Moon; Venus. **Volcanoes**.

Further Reading

Anderson DL (1999) A theory of the Earth: Hutton, Humpty Dumpty and Holmes. In: Craig GY and Hull JH (eds.) *James Hutton: Present and Future,* pp. 13 35. Special Publication 150. London: Geological Society.

Edwards K and Rosen B (2000) *From the Beginning*. London: Natural History Museum.

Hamilton WB (2002) The closed upper mantle circulation in plate tectonics. In: *Plate Boundary Zones,* pp. 359 410. Geodynamics Series 30. American Geophysical Union.

Hancock PL and Skinner BJ (2000) *Oxford Companion to the Earth*. Oxford: Oxford University Press.

Harland WB, Armstrong RL, Cox AV, *et al.* (1990) *A Geologic Timescale 1989*. Cambridge: Cambridge University Press.

Holmes A (1965) *Principles of Geology,* revised edn. London and Edinburgh: Nelson.

McCall GJH (2000) Age and early evolution of the Earth and Solar System. In: Hancock PL and Skinner BJ (eds.) *Oxford Companion to the Earth*, pp. 8 11. Oxford: Oxford University Press.

McCall GJH (2003) Pole up the pole. *Geoscientist* 13: 9.

Price NJ (2001) *Major Impacts and Plate Tectonics*. London and New York: Routledge.

Rothery DA (1992) *Satellites of the Outer Planets*. Oxford: Clarendon.

Van Andel TJ (1994) *New Views on an Old Planet,* 2nd edn. Cambridge: Cambridge University Press.

Vita Finzi C (2002) *Monitoring the Earth*. Harpenden: Terra.

Wilde SA, Valley JW, Peck WH, and Graham CM (2001) Evidence for the early growth of continents and oceans from <4 Ga detrital zircons. *AGSO Geoscience Australia Record* 2002/3: 6 8.

EARTH SYSTEM SCIENCE

R C Selley, Imperial College London, London, UK

Introduction

What on Earth is Earth System Science?

Earth system science is founded on the precept that the Earth is a dynamic system that is essentially closed materially, but open with respect to energy. This statement needs to be qualified by noting that the Earth continues to accrete matter from space in the form of meteorites, asteroids, and comets. Incoming energy is principally derived from solar radiation at a rate that may fluctuate with time. Earth processes in the lithosphere, the biosphere, and the atmosphere are linked, with a change in one impacting on one or more of the others. Earth system science has grown out of an appreciation of the need to integrate geology with other scientific disciplines, not only to understand the planet on which we live, but also most particularly to predict its future in general and climate change in particular. Earth system science requires a holistic approach to education in which students learn geoscience, bioscience, climatic science, astroscience, and space science synchronously and seamlessly.

The Genesis of Earth System Science

In the fourteenth century Richard de Bury (Bishop of Durham 1333–45) divided all research into 'Geologia' (geology), the study of earthly things, and 'Theologia' (theology), the study of heavenly things. Masons, miners, and engineers were the first investigators of rocks. The Church taught them that the Earth was an inert mass of rock formed in 7 days. Subsequently natural philosophers pondered the meaning of fossiliferous strata and how it was that they were intruded by crystalline rocks and truncated by unconformities. Gradually it dawned on these natural philosophers that the rocks were not formed in an instant, but resulted from processes, many of which could be observed on the modern surface of the Earth. This was formulated in Hutton's principle of uniformitarianism, (*see* **Famous Geologists:** Hutton) and was epitomized in the dictum that 'the present is the key to the past.'

As Lyell (*see* **Famous Geologists:** Lyell) wrote in his *Principles of Geology* in 1834:

> The entire mass of stratified deposits in the Earth's crust is at once the monument and measure of the denudation which has taken place.

By the beginning of the twentieth century it was realized that the history of the Earth could be interpreted in terms of cycles. Davies (1850–1934) recognized the landscape cycle, commencing with uplift, followed by youthful, mature, and senile landforms, followed by rejuvenation. Stratigraphers recognized cycles of weathering (*see* **Weathering**), erosion, transportation, deposition, and diagenesis. Sedimentologists discovered that all sedimentation is cyclic, although some is more cyclic than others. The hydrologic cycle was revealed, in which water fell as rain on land, flowed into rivers, was discharged into the world's oceans, evaporated and reprecipitated. Geochemists identified the cycles of carbon (*see* **Carbon Cycle**) and other key elements, such as nitrogen, oxygen, and sulphur.

Agassiz (1807–73) (*see* **Famous Geologists:** Agassiz) was the first geologist to establish the existence of ancient glaciation. Subsequently evidence accumulated for past climatic cycles of alternating 'greenhouse' and 'icehouse' phases, as they became picturesquely termed.

Thus cyclicity was revealed in rocks, water, and air – or the lithosphere, the hydrosphere, and the atmosphere. The extent to which these cycles interrelated with one another was little understood. Initially palaeontologists took the view that the evolution of the biosphere responded to external changes, and had little inter-reaction with them. Subsequently it was realized that this is far from the case. The stromatolitic limestones (*see* **Fossil Plants:** Calcareous Algae) of the Late Precambrian, some 3400 My BP, are a dramatic example (Figure 1). These limestones are the relicts of primitive algae and cyanobacteria. They are found worldwide in Late Precambrian and Phanerozoic strata, and form today in tidal-flat environments.

Stromatolitic limestones provide the earliest case preserved in the stratigraphical record of the interaction of organic and inorganic processes to form rock. Their creators, the first abundant photosynthesizers, took carbon dioxide from the atmosphere, replaced it to some degree with oxygen, extracted calcium from sea water, and precipitated the vast limestone rock formations that are preserved all over the world to this day (*see* **Atmosphere Evolution**).

Analysis of ice cores from modern polar regions shows a strong positive correlation between carbon

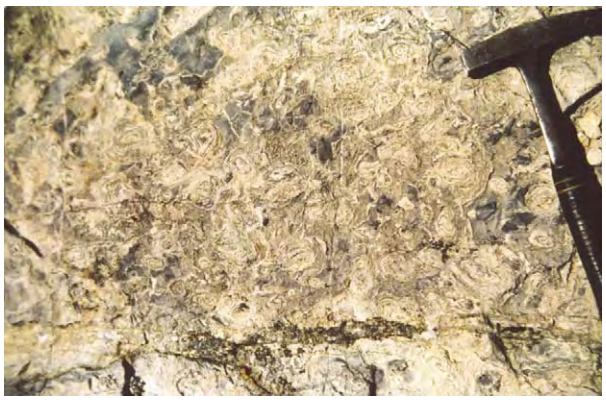


Figure 1 A bedding surface showing the characteristic collo form brassicamorphic structure of a stromatolitic limestone. Late Precambrian, Ella Island, East Greenland. Extensive Late Precam brian stromatolitic limestones provide early evidence of the inter relationship between the biosphere, the hydrosphere, and the atmosphere. Earth system science is all about these interrelation ships. For modern example *see* **Minerals:** Carbonates **Figure 2G**.

dioxide concentrations in the atmosphere and temperature. It is reasonable to assume such linkage in the past, and that the precipitation of stromatolitic and other limestones may be one of several causes of global cooling. In the history of the Earth extensive episodes of limestone formation have been followed by global cooling several times. The Late Precambrian stromatolitic limestones demonstrate the linkage between biosphere, lithosphere, and atmosphere, and provide a good starting point to examine what is now called Earth system science.

Biogeochemical Cycles

The appreciation of the interplay of atmosphere, biosphere, and hydrosphere, demonstrated by stromatolites, coupled with the realization that many Earth processes are cyclic, has lead to the concept of the biogeochemical cycle. This term is applied to the flux of material in and out of the lithosphere, hydrosphere, biosphere, pedosphere, and atmosphere (which collectively constitute the geosphere), and the chemical and physical changes that occur therein. The material Earth, or geosphere, is a closed dynamic system, within which there is a constant recycling of its components. It is not, however, a closed system with respect to energy, because this constantly reaches the Earth in the form of solar radiation (Figure 2).

A variation in solar energy will impact on the dynamics of biogeochemical cycles. As a generalization, an increase in solar radiation may result in global warming, and a decrease may result in global cooling. The situation is more complex than this, however, because of the interconnected nature of the cycles or systems.

Some Definitions Defined

Biogeochemical cycles are described and modelled in terms of reservoirs, material, fluxes, sources, sinks, and budgets. The reservoir is the amount of material in a given Earth system, such as oxygen in the atmosphere, or water in the ocean. The flux is the amount of material moved from one reservoir to another – for example, the amount of water lost from the ocean to the atmosphere by evaporation. The source is the flux of material into a reservoir, and the sink is the amount of material removed from it. The budget is the balance sheet of the amount of material lost or gained in a system. If source and flux are in equilibrium, such that the amount of material in the reservoir is constant, it is in a steady state. A system of two or more reservoirs, in which material is transferred cyclically without an external flux, is termed closed. Commonly Earth systems are interconnected or coupled: a variation in the flux of one affects the dynamics of the other (Figure 3).

Awareness that the geosphere consists of a multitude of coupled systems, where a variation in the flux of one will have a 'knock-on' effect on others, poses the question 'why is not the Earth in a permanent state of dynamic chaos?' Why, for example, does the Earth's climate alternate between glacial and interglacial episodes, without fluctuating wildly all the time? The answer lies in feedback, a concept familiar to physiologists and engineers. For example, the temperature of the human body remains constant, thanks to the feedback effect of the hypothalamus, and the temperature of a house remains constant owing to regulation of the central heating system by a thermostat. A feedback is when a process in one system causes changes in another one that in turn influences the first system. A positive feedback accelerates the original process; a negative feedback retards it. A positive feedback may be illustrated by considering the effect of increasing the CO_2 content of the atmosphere (by burning fossil fuels, by volcanic eruption, or whatever). This greenhouse gas will result in an increase in temperature, which will result in evaporation increasing from the oceans to the atmosphere and an increase in water vapour, another greenhouse gas, which will accelerate the increase in temperature. Negative feedbacks are rarer than positive feedbacks in the geosphere, but include heat transfer towards the poles and the cooling effect of volcanic ash clouds.

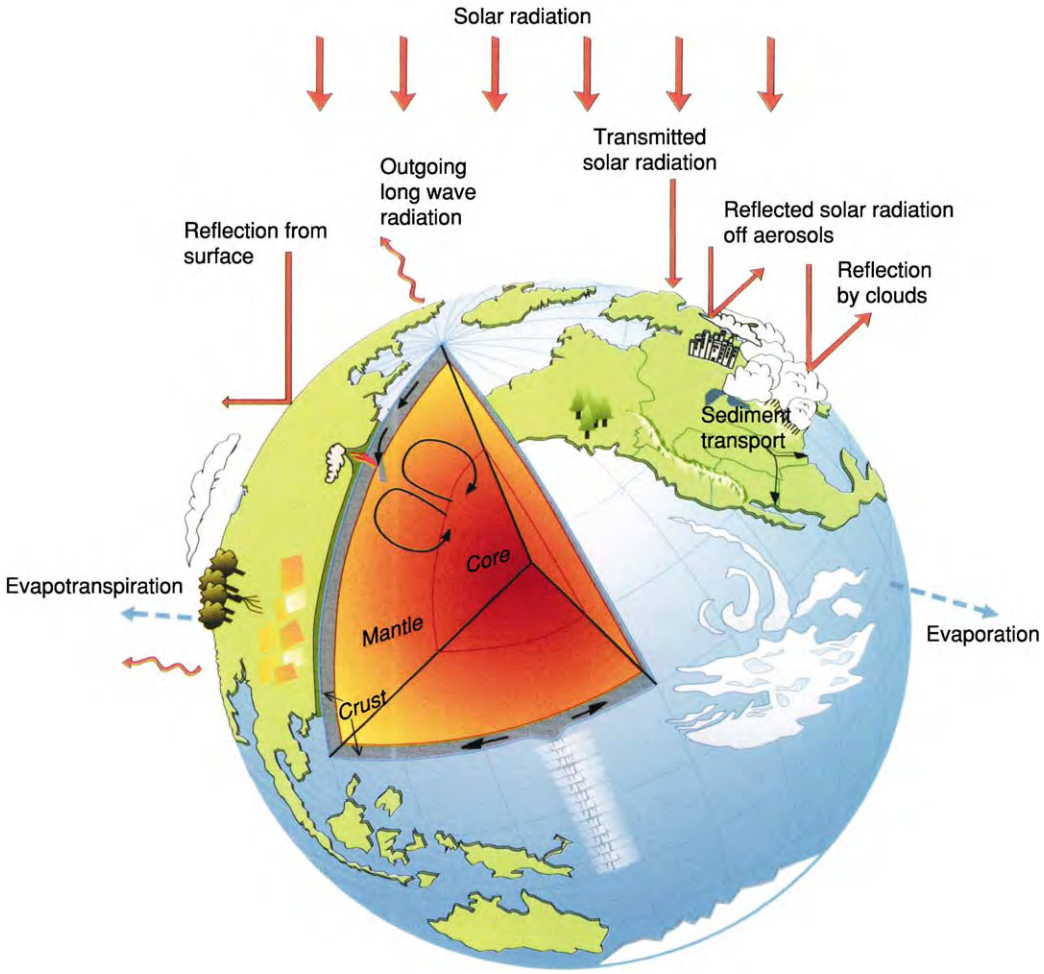


Figure 2 Overview of the cyclic processes on Earth. The placements of the geological and geographical features are not meant to represent their true positions on the Earth, but to provide an overview of some of the Earth systems. Reproduced with permission from Jacobson MC, Charlson RJ, and Rodhe H (2000) Introduction: Biogeochemical cycles as fundamental constructs for studying Earth system science and global change. In: Jacobson MC, Charlson RJ, Rodhe H, and Orians H. (eds.) *Earth System Science*, pp. 3 13. San Diego: Academic Press.

Earth System Science and the 'Gaia' Hypothesis

Since the 1970s James Lovelock developed the Gaia hypothesis, named after the ancient Greek goddess of the Earth (*See* **Gaia**). As originally conceived the 'Gaia' concept envisages the Earth as a super-organism that operates to regulate its own environment, principally temperature, to keep it habitable for the biosphere. Lovelock has never argued that the biosphere consciously anticipates environmental change, but only that it automatically responds to it. Nonetheless some sections of the public have construed it that way, and in the popular mind Gaia gained a quasi-mystical connotation, enhanced by its name. The great value of the Gaia hypothesis is that it presents the interdependence of the constituents of the geosphere in a media-friendly way. Earth system science also involves a holistic approach to the geosphere, but without the 'ghost in the machine'. Nonetheless Amazon, the internet book shop, still classifies books on Earth system science under 'Religion and Spirituality > New Age > Earth-Based Religions > Gaia'.

Impact of Earth System Science on Geology

When the dust of history has settled over the present period it may be argued that Earth system science has had as large an effect on geology as did plate tectonics some 40 years previously (*See* **History of Geology Since 1962**). Its import may be wider still however. This is because Earth system science is an important

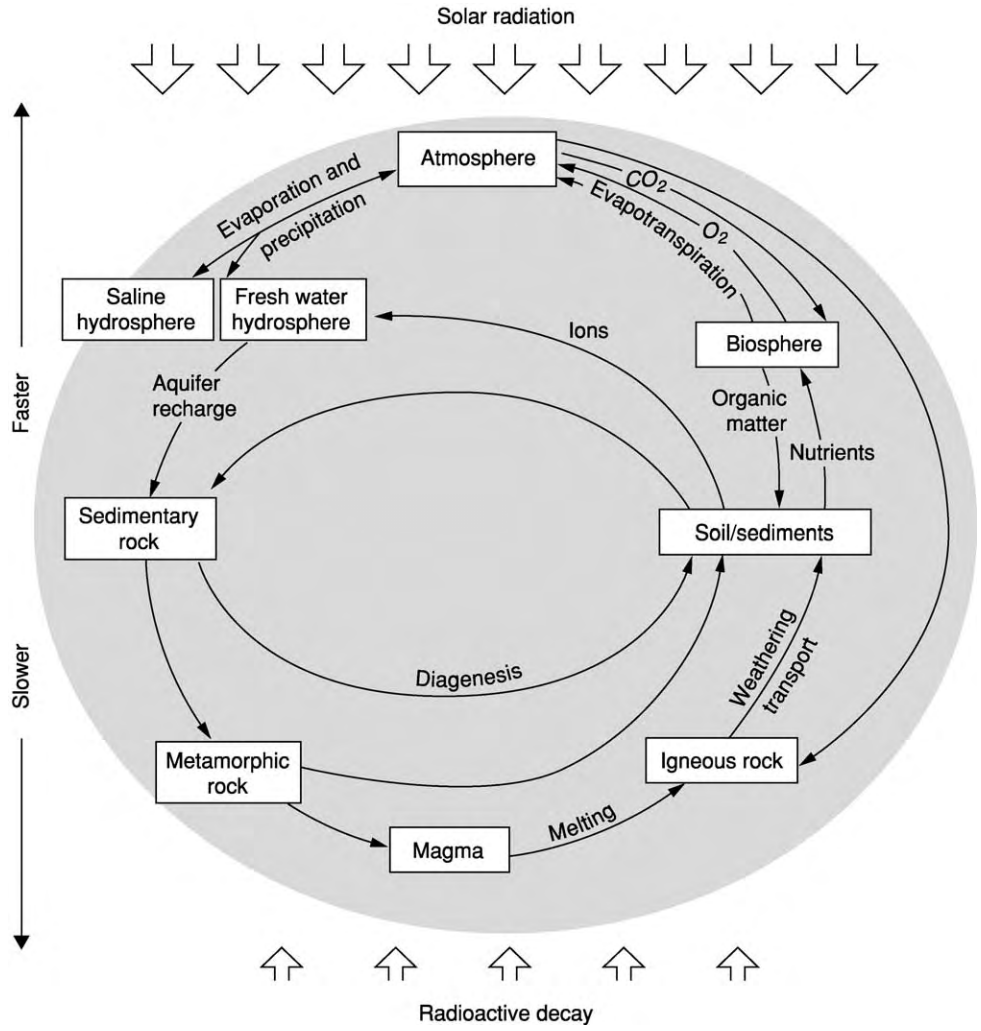


Figure 3 The cyclic processes and fluxes between major reservoirs on the Earth. Reproduced with permission from Jacobson MC, Charlson RJ, and Rodhe H (2000) Introduction: Biogeochemical cycles as fundamental constructs for studying Earth system science and global change. In: Jacobson MC, Charlson RJ, Rodhe H, and Orians H. (eds.) *Earth System Science*, pp. 3 13. San Diego: Academic Press.

tool in the study of climate change, a topic of great current concern. In particular it focuses attention on how anthropomorphic activities, such as the combustion of fossil fuels and deforestation, may have fed back into atmospheric systems, and hence affected global climate (*see* **Palaeoclimates**).

Moving from the practical to the theoretical, Earth system science has brought about a decline in the reductionist approach to science in general and to geology in particular. (Reductionism is the process of knowing more and more about less and less.) In its infancy geology was advanced by natural philosophers (the term 'scientist' was not popularized until 1858 by Huxley) who were polymaths. It was relatively easy to be a polymath in the nineteenth century, because the pool of scientific knowledge was limited. As the pool of knowledge expanded into a lake, then a sea, and finally an ocean, scientists have had to focus their attention on progressively smaller and smaller areas of knowledge, thus losing sight of the wood for the trees. Distinct disciplines of chemistry, physics, life science, and geology have evolved, all with their own specialized subsets. Earth system science, by taking a holistic view of the Earth, has had a beneficial effect on the development of interdisciplinary scientific research. Concomitant with this, however, has been a decline in the training of geologists who can identify rocks, minerals, and fossils, and map their distribution over the Earth. Geology has become disseminated into Earth science. University Geology Departments have metamorphosed into Departments of Earth Science, coupled with geography, environmental science, etc. These departments no longer produce graduates with focused geological knowledge, but Earth scientists who are Jacks-of-all-trades and masters of none.

Bishop Richard of Durham would have been saddened to see his word 'Geologia' fall into desuetude after 700 years.

See Also

Atmosphere Evolution. **Carbon Cycle**. **Famous Geologists:** Agassiz; Hutton; Lyell. **Fossil Plants:** Calcareous Algae. **Gaia**. **History of Geology Since 1962**. **Minerals:** Carbonates. **Palaeoclimates**. **Weathering**.

Further Reading

Ernst WGF (ed.) (2000) *Earth Systems: Processes and Issues*. Cambridge: Cambridge University Press.

Hamblin K (2002) *The Earth's Dynamic Systems: A Textbook in Physical Geology*. New York: Prentice Hall.

Jacobson MC, Charlson RJ, Rodhe H, and Orians H (2000) *Earth System Science*. San Diego: Academic Press.

Kump LR, Keating JF, Crane RG, and Kasting JF (2003) *The Earth System: An Introduction to Earth Systems Science*. New York: Prentice Hall.

Lovelock J (2000) *The Ages of Gaia: A Biography of Our Living Earth*. Oxford: Oxford University Press.

Nisbet E (2002) The influence of life on the face of the Earth. In: Fowler CMR, Ebinger CJ, and Hawkesworth CJ (eds.) *The Early Earth: Physical, Chemical and Biological Development*, pp. 275 307. Special Publication 199. London: Geological Society.

EARTHQUAKES

See **ENGINEERING GEOLOGY: Aspects of Earthquakes; TECTONICS: Earthquakes**

ECONOMIC GEOLOGY

G R Davis, Imperial College London, London, UK

Introduction

If you visit the strangely dimpled landscape known from Saxon times as Grimes Graves in Norfolk, England, you will see not an ancient burial ground, but hundreds of backfilled shafts in the white Cretaceous Chalk. Here Neolithic man discovered and exploited a thin subsurface layer of dark flint (**Figure 1**) so extensive and of such superior quality that, after due process of mining and treatment, the finished product could profitably be traded and fashioned into the finest flint tools and weapons at the cutting edge of technology. If you visit the flat forested landscape known as Weipa in the Yorke Peninsula, Australia, you will see where twentieth century man discovered and now exploits a surface layer of red bauxite (**Figure 2**) so extensive and of such superior quality that, after due process of mining and treatment, the finished product can profitably be traded and fashioned into the finest aluminium utensils and machines at the cutting edge of technology.

In his long march from hunter-gatherer to moonwalker, *Homo* has used his *sapiens* to exploit the Earth's bountiful mineral resources that, apart from food and clothing, have provided the materials for his advancement. At about 2000 BC flint from Grimes Graves and many other flint workings in the Cretaceous Chalk of Europe was an economic industrial mineral because insufficient copper and bronze was available to meet demand. The occurrence in nature of native copper and gold had already led to their use

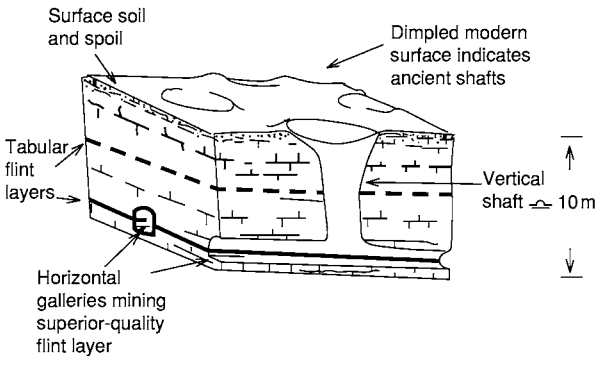


Figure 1 A sketch to illustrate how high quality flint was mined in the chalk of Britain about 4000 years ago.

Figure 2 The cliffs facing the Gulf of Carpentaria in Queensland, Australia, and a working face in the vast expanse of pisolitic bauxite. Scale shown by white penknife near the centre. Photos: GR Davis.

for over 2000 years as tools, ornaments and weapons. Supplies of metal increased when smelting was discovered to reduce the natural ores of Cu, Pb, Sn and later Fe into metal. The ownership of valuable metals and other minerals, such as salt, begat wealth and power. The rise and fall of ancient civilizations is intimately bound up with the riches derived from control of mineral resources, and the historical theme continues to this day. For example the shift of power and control of trade routes from the Persians to the Greeks was greatly influenced by wealth from the newly discovered Laurium Ag-Pb deposits near Athens, and then gold in Macedonia. Finance was available to build a powerful navy that enabled Themistocles to conquer Xerxes at the decisive battle of Salamis in 480 BC. The Greeks enjoyed a golden age of philosophy and art; the mining revenues assisted Philip of Macedonia to establish his dominance, and for his son Alexander the Great to finance the first campaign in his great Middle Eastern conquests about 330 BC. In that same region, the 20th century has witnessed the global economic and political impact of the discovery and development of major oilfields upon the countries of the Middle East.

Figure 3 (with logarithmic scale on the vertical axis) illustrates how world demand for mineral products over the last 150 years of the industrial age has been met by production growth rates that exceed even the 'explosive' growth rate of world population. Annual production rates for the basic industrial materials iron and petroleum have grown dramatically, and the production rate of copper, a primary base metal, has increased a hundredfold. During the twentieth century newer metals such as nickel became industrially important, and aluminium, once rare and precious, reached production rates exceeding that of copper, and at a lower cost per tonne. Many other 'new' elements continue to emerge in response to the growth of modern high-technology industry. For instance, demand is now growing rapidly for indium (mine production about 340 tonnes in 2002), which is needed for making flat-panel electronic displays as used in mobile telephones and plasma-screen televisions. World population is increasing, and the average per capita demand for mineral products, while heavily skewed by industrialized Western world usage, is also increasing. Economic geology plays an integral part in meeting this challenge to the mineral resource base of the world.

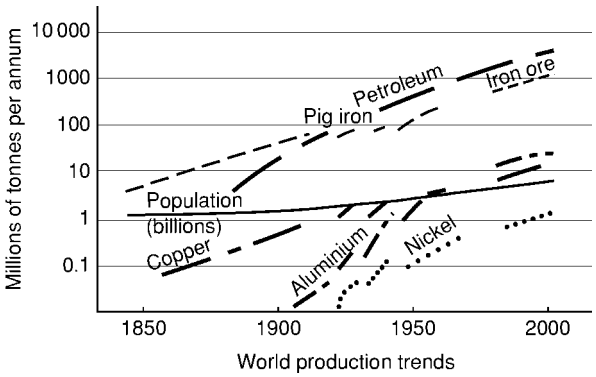


Figure 3 Over the past 150 years, world demand for mineral products has exceeded the 'explosive' growth in population, shown on a logarithmic scale.

Characteristics of Economic Mineral Deposits

The AGI Glossary of Geology defines economic geology as... "the application of geologic knowledge and theory to the search for and the understanding of mineral deposits; study and analysis of geologic bodies and materials that can be utilized profitably by man, including fuels, metals, non-metallic minerals, and water". In this work (Encyclopedia) water and hydrocarbon fuels are treated separately in view of their major industrial and economic importance. The solid mineral resources also fall into specialist groups, as indicated in the definition quoted above. Collectively, they comprise the non-renewable resource base of the extractive industries. To qualify for "profitable utilisation by man" they must all demonstrate certain characteristics that are, appropriately, both economic and geological.

Geological Attributes of Economic Deposits

- The size (tonnage) must be sufficient to sustain exploitation for a period long enough (mine life) to justify development.
- The valuable content (grade) must be high enough to repay all costs. Taken together, the tonnage and average grade express the comparative geological potential of the deposit.
- The shape, attitude, depth and physical properties of the deposit must be amenable to extraction by existing mining technology.
- The valuable constituent(s) must be amenable to separation from the unwanted portions (waste), and beneficiation to marketable product by existing technologies.

Economic Attributes of Economic Deposits

- The price received for the product(s) must cover all production costs and assure competitive advantage.
- The geographical situation must be amenable to mine development, including water and power supplies; and within economic transport distance to market, especially for bulky products.
- Socio-political conditions must be favourable. These include planning permission and mineral rights, various taxes and royalties, labour and safety laws and conditions, regulations for waste disposal, and rehabilitation when the mine is worked out.

A new mining development is typically capital intensive with a lead-time of years between discovery and profitable production. Brief consideration of the seven attributes listed reveals that the risk of financial failure is shared between measurable geological, engineering and logistical natural factors on the one hand and indefinite state-imposed constraints or incentives on the other. Economic geologists need a sound appreciation of mineral economics, mining engineering and mineral processing technology, but their core expertise rests upon their knowledge of mineral deposits.

Figure 4 illustrates the hierarchical progression from mineral occurrence to mineral prospect to mineral deposit, and then through mineral resources and proved mineral reserves to profitable mining. This progression rests for its success, above all else, on the quality of geological observation and interpretation, and its application in industry.

It is important also to appreciate that the mines of today are exploiting only those selected deposits that have met all the criteria for current economic development. They represent only the tip of the iceberg of the Earth's mineral resources. Large deposits of many mineral commodities, currently identified but subeconomic, may in future become payable because of improvements in mining or mineral extraction technology, commodity demands and prices, and similar conditions. Those deposits may one day join the stock of future discoveries that will provide the economic deposits and mines of tomorrow.

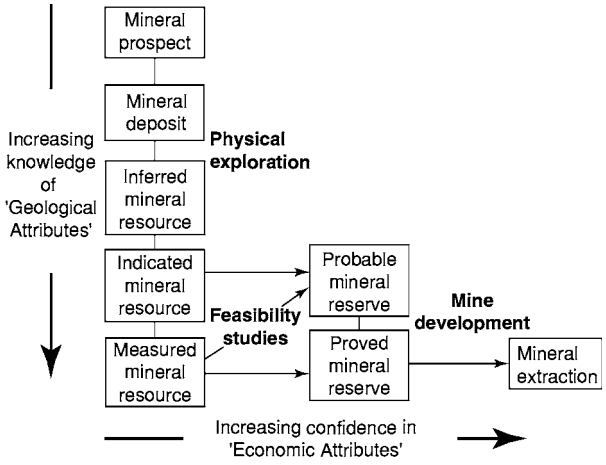


Figure 4 The progression from a mineral prospect to mineral extraction. Expenses increase as work becomes more detailed. Mine development requires capital expenditure before cash flow starts from mineral product sales. The categories of resources and reserves, and the technical terms proved, probable, meas ured, and indicated are strictly defined in codes of practice for mulated by professional mining institutions in some leading mining countries (Australia, Canada, South Africa, UK/Europe, and USA) and accepted by financial organizations such as stock exchanges.

Variety and Use of Mineral Deposits

The range of 'geologic bodies and materials' that is utilized by man has grown in scope and volume over the course of history, but falls into a few practical groups.

Construction Materials

These include building stones, sand, aggregates, clay, and cement raw materials. The high cost of transporting these bulk products requires source rocks to be as close as possible to the place of utilization, but typically the costs of extraction and treatment are low. A long history of use (the stone walls of Jericho were packed with a clay mortar circa 8000 BC) has built a body of experience concerning desirable source materials and their properties. In developed countries today, geological knowledge is applied to not only ensure that technical specifications are met, but also to discover and delineate those particular sources that best suit land development planning and minimize environmental problems. Some industrial rock products are valuable enough for export, such as clays for ceramics and ornamental stone for decorative use in buildings.

Fuel or Energy Mineral Deposits

These include hydrocarbon fluids, coal and uranium. Coal is a sedimentary rock derived from plant remains, the fossil fuel on which modern industrial development was built. Together with petroleum, this versatile material supplies most of the worlds' energy needs. It is also a major industrial raw material for the manufacture of chemicals, and coke for iron and steel production. Economic deposits of uranium-bearing minerals, the base on which the atomic age is built, occur at a scale of magnitude nowhere near that of coal and oilfields.

Industrial Minerals

Sometimes termed non-metallic minerals, these are valued for their chemical and/or physical properties and the fact that they are not of widespread occurrence. In general, prices are sensitive to market demand and product specifications (with premium prices for premium grade products), and quality is a major factor in the economic geological evaluation of mineral reserves and productive life of industrial mineral deposits. A vast range of industrial minerals is produced in an equally impressive range of tonnages. Minerals with valuable chemical properties, used mainly in the chemical and fertilizer industries, include rock phosphate, potash and mixed chloride salts, sulphur, nitrates and borates. Fluorspar and limestone are prominent as fluxes in the metallurgical industry and ceramics, and other process industries consume silica sand, feldspars, and kyanite. Physical attributes useful in filler and extender applications make talc, limestone and kaolin competitive in paints, paper and plastics. Other minerals with useful physical properties include asbestos, barytes, diatomaceous earth, and the lightweight aggregates perlite, pumice and vermiculite. Hardness is utilized in abrasives such as corundum, garnet and industrial diamond. The extensive list of industrially useful minerals makes it clear that a wide range of geological knowledge finds application in the search for industrial minerals and in ensuring products that conform to specifications set by the industrial user.

Metallic Mineral Deposits

The metallic ore minerals are commonly metal compounds such as sulphides and oxides in which the metal content is high compared with rock-forming minerals. Natural concentrations of ore minerals form discrete ore bodies that may typically contain only a few percent of the valuable metal. Unlike many industrial minerals that find direct use after mining and limited beneficiation, the ore minerals, in general, must be reduced to metal by complex processing. Modern mining and mineral extraction procedures are tending towards greater use of chemical and bacterial leaching methods for suitable ores, such as in-situ extraction of some uranium deposits, and heap leaching of some gold ores. The great bulk of metalliferous ore, however, is mined by surface or underground rock-breaking methods. Run of mine ore is first crushed and milled to a fine pulp, from which the desired ore minerals are separated from the gangue minerals by various methods to produce a concentrate. Metal is then extracted from the mineral concentrate by further treatment, which may include smelting or various chemical methods such as solvent extraction and electrowinning, followed by refining to market standards. The expensive multi-stage process of extraction results in complex engineering works at the site of large ore bodies, often in remote locations. The commonly used metals are sometimes grouped for convenience by their geochemical or industrial affinities. Base metals include Pb, Zn, Sn and Cu. Iron and ferroalloy metals include Cr, Co, Ni, Mn and V. Light metals include Al, Ti, Mg, Li and Be. The precious metals comprise Ag, Au and PGM (platinum group metals) (**Table 1**, **Figure 5**).

The attached statistical data illustrate some world production rates and the relative importance of the broad mineral groups in the world economy. It should be noted that most metal production tonnages are less by one or two orders of magnitude

Table 1 Annual world production levels of selected metals and industrial minerals

	1982 1983	*2000 2001*	*Main producing countries*	
	Tonnes per annum (millions)			
Coal	4000	4500	China	USA
Asbestos	4.1	2.0	Russia	Canada
Bauxite	78	139	Australia	Guinea
Chromite	8.2	13	South Africa	Kazakhstan
Copper	8.0	13.6	Chile	USA
Diatomite	1.9	1.7	USA	China
Fluorspar	4.4	4.2	China	Mexico
Iron ore	750	1150	China	Brazil
Kaolin	(21)	23	USA	UK
Lead	3.5	3.0	Australia	China
Manganese ore	24	22	China	South Africa
Phosphates	130	130	USA	China
Potash	25	27	Canada	Belarus
Talc	7.4	8.4	China	USA
Zinc	6.5	9.0	China	Australia
	Tonnes per annum (thousands)			
Antimony	?	123	China	Guatamala
Gold	1.4	2.5	South Africa	USA
Mercury	6.0	1.8	Kyrgyzstan	Spain
Platinum group metals	0.3	0.5	South Africa	Russia
Silver	12	19	Mexico	Peru
Tin	205	245	China	Indonesia
Uranium	50	34	Canada	Australia
Vermiculite	540	376	South Africa	USA

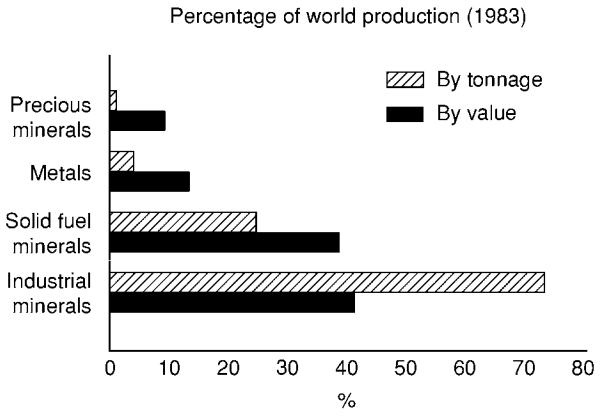


Figure 5 The importance of industrial minerals in the modern world is not always appreciated. The relative values of the four commodity groups is a better measure than tonnages, because the amount of rock mined to produce metals and precious metals from their ores is at least an order of magnitude higher. Compiled from data quoted in Evans (1995), after Noetstaller.

than the tonnages of rock mined to produce them. The relative value of metallic mineral and industrial mineral production varies considerably from one country to another. For instance, for the mean of the years 2001 and 2002, industrial minerals accounted for 78% of the total value of non-fuel minerals in the USA whereas in neighbouring Canada the metallic minerals accounted for 57% of the non-fuel group. (Compiled from various sources).

World Distribution of Economic Mineral Deposits

Like all other geological bodies, mineral resources are unevenly distributed in the Earth's crust. Many nations owe a major part of their wealth to discovery and development of economic mineral deposits, while others may be either very poorly endowed by nature, or ignorant of their undiscovered mineral resources. The people of Nauru are richly blessed because their small Pacific island homeland contains 40 million tonnes of high grade phosphate rock. In developed countries even with a large mining industry, the lack of domestic sources of certain vital commodities leads to the concept of 'strategic minerals' and defensive stockpiling against national emergencies. Before entry into World War II the USA included Cr, Mn and Sn in a list of 14 strategic commodities. Statistics of world mineral production illustrate the imbalances that cause geopolitical concern about national vulnerability to imported supplies (Figure 6), or to damaging price changes in key commodities. For example South Africa, a country about twice the size of France, dominates Pt production, supplies almost half the worlds Cr and major amounts of Au, Mn and V, but has inadequate domestic resources of Mo and Al, oil and gas, potash, sulphur and kaolin. China, a large country with a rapidly expanding economy, dominates world production of the rare earth elements (REE), produces over 80% of world Sb, over one-third of world V, Sn and W, and is a key nation in world minerals trade.

The observed major world patterns of distribution of economic mineral deposits become less arbitrary and more understandable when viewed in the light of metallogenic and geological maps. Relationships appear between mineral resources and the geological environments and rock systems in which they occur. For instance, oilfields and coalfields are associated with large sedimentary basins of upper Phanerozoic age, and one would not search hopefully for these

Figure 6 This map, showing the sites of tin production, is an example of the uneven global distribution of several mineral commodities. Modified from World Mineral Statistics (2003) by permission of the British Geological Survey. © NERC. All rights reserved. IPR/46 28CW.

great energy resources in sedimentary basins within ancient shield areas of Archaean rocks. The converse applies for the great banded iron formations (BIF) in Precambrian sedimentary basins that are host to most of the world's high-grade Fe ore deposits (*see* **Sedimentary Rocks:** Banded Iron Formations). Similarly, large and highly productive porphyry copper deposits are strongly grouped in the mountain chains made up of volcanic rocks and granitic intrusions along the western margins of the Americas. This pattern has been more readily understood over the past 40 years in the light of geological mapping and plate tectonic theory. Empirical associations between mineral deposits and their host rocks have long been noted and put to use by miners, and the list is still growing as geological knowledge of Earth's physico-chemical systems advances. Concepts have developed of mineral provinces and mineral epochs, and in recent decades regional patterns of mineralization have been related to the various kinds of tectonic plate boundaries (Figure 7) (*see* **Plate Tectonics**).

Economic Geology and the Extractive Minerals Industries

Every producing mineral deposit is a non-renewable resource with a finite life. In order to continue to meet world demand, the extractive industries are geared to a life cycle of activities as shown in Table 2.

These four activities are very briefly outlined below, and the reader is referred to specific topics for more comprehensive information.

The State of Relevant Geological Knowledge

Economic geologists are widely employed in the many specialized areas of the extractive industries. All, however, rely on the quality and appropriate application of their geological knowledge of mineral deposits (*see* **Mineral Deposits and Their Genesis**). Great advances have been made in the geological understanding of the 4500 million year history of the earth, and of the igneous, sedimentary and metamorphic processes in which the genesis of mineral deposits is an integral part. Concentrations of valuable minerals into economic deposits are no longer seen as special, isolated events, but as the result of processes, often sequential and superimposed, that operate in geological environments of every kind and in every age from the Archaean to the present. Advancing views on regional patterns of mineral deposition related to space and time (i.e., in mineral provinces and mineral epochs) were dramatically boosted through definition of crustal tectonic regimes by the theory of plate tectonics. The scientific overview of mineral deposits has moved away from worthy attempts at genetic classification, notably that of Lindgren, towards development of 'mineral

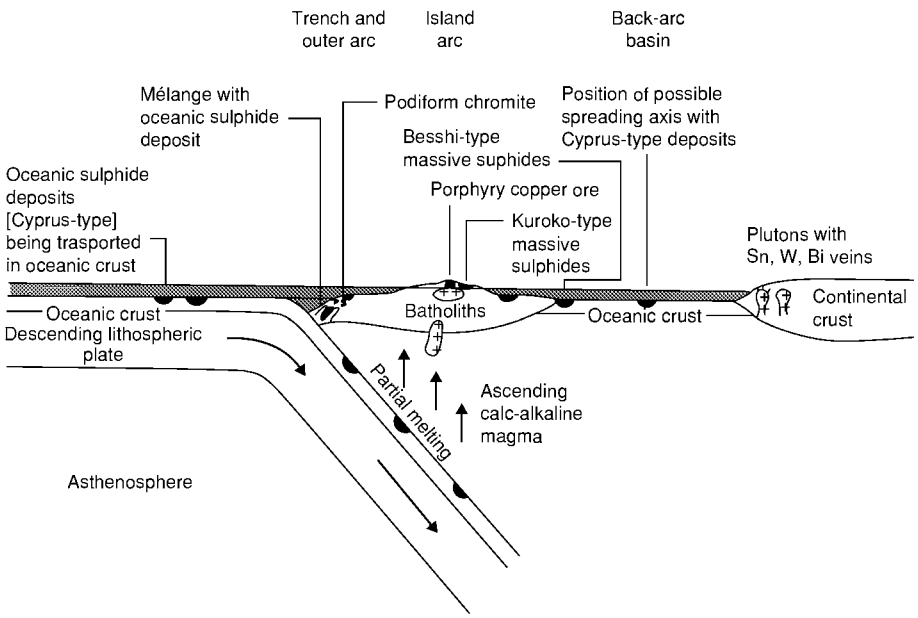


Figure 7 Development of the theory of plate tectonics in the 1960s led rapidly to a new view of the genesis of some ore deposits as shown on this early diagram in 1972. Reproduced with permission from Evans AM (1995) *Introduction to Mineral Exploration.* Oxford: Blackwell, p. 40 (modified from Sillitoe 1972).

Table 2 Life cycle of extractive industry activities

Activity	*Applied geology input*
Discovery of new deposits	Dominant role
Feasibility studies and mine development	Major role in defining reserves and critical technical parameters
Mineral extraction	Essential engineering role in operational efficiency
End of life of deposit	Long term environmental assessment

deposit models'. These models marshal all the pertinent facts concerning related deposits that are seen as 'types'. Modern advances in nearly all fields of geology are of relevance and application to some aspect of economic geology. For example structural geology from the scale of electron microscopy in rock deformation studies to megastructures on satellite imagery; evolution and structure of sedimentary basins; the geochemistry of sediments as source rocks for hydrocarbons and metalliferous brines; the volcanic realm and its diverse products; magma generation and crystallization; metamorphic effects on rock structures and mineral chemistry; and even the profound effects of major meteorite impacts on the Earth's crust.

Specialist fields of knowledge are required for each phase of the industrial cycle of activity (Table 2) in each branch of the extractive industries, and economic geologists have formed several specialist groups such as petroleum geologists, mining geologists and exploration geologists.

Discovery of New Deposits

To maintain a stable extractive industry long term, the huge tonnages of minerals extracted each year must be replaced, on average, by discovery and/or development of an equivalent amount of new materials. The difficulty imposed by this high demand is compounded by the fact that new mineral fields and individual deposits become progressively harder to discover as the easier finds are made. Scientifically based mineral exploration has developed into a sub-industry, in which the participants include government agencies, mineral resource companies of all sizes, and specialist groups offering contract and consulting services.

Exploration objectives are targeted on either certain commodities or certain regions of interest. The search area for bulk materials such as aggregates or low-grade coal is restricted by transportation costs and environmental factors; some mines concentrate their search close to existing infrastructure; and state organizations are of course interested in their own territories. Most mineral commodities are internationally traded, and exploration money and effort tend to favour countries that offer the most attractive combination of geology, political stability, mining taxation and operational infrastructure. Commodity-targeted exploration is based solidly on favourable

geological environments, the oil and gas sector being a prime example. Views on the factors that have controlled or generated mineral deposits (e.g., where in the world to look for kimberlite pipes?) are of paramount importance to cost-effective exploration. Many small specialized groups undertake exploration on this basis. Their successes may be sold to large companies, which have the policy options of either buying their future mineral resources or organizing their own exploration teams on a long-term basis.

Exploration programmes generally follow phases of target definition, reconnaissance, selective follow-up, and detailed exploration of prospects. A great and growing number of exploration technologies, including heavy mineral, geochemical and geophysical surveys, are available to supplement the essential art and science of sound geological mapping based on wide-awake observation. At the detailed stage of exploration, physical methods are used, such as pitting, trenching and especially the various types of drillholes now available.

Mineral exploration is a high-risk enterprise justified by potentially high reward, but a historically low success rate overall. Organizations with a relatively high success rate tend to be well managed, exploring in carefully researched and targeted areas, employing teams of well-qualified and motivated people to use the most cost-effective sequence of search technologies, with financial backing stable enough to ensure long-term effort. The constant application of sound geological judgement to ensure that exploration data at all stages make geological sense, and especially when deciding whether to abandon a prospect or press on at higher cost and effort, makes disciplined mineral exploration a professionally rewarding activity (Figure 8).

Feasibility Studies and Mine Development

At some point enough is known about a prospect through physical exploration to either abandon it or proceed to further detailed exploration work. A decision must then be made whether the prospect is worth the high cost risk of development into an operating mine. This is the period of feasibility study, when a multidisciplinary team will rigorously test the prospect against all the criteria for full economic status. The professional contribution from economic geology is the foundation stone upon which the entire edifice will be built, and confidence in the reliability of all geological data is essential. The main geological information will include the following:

- Knowledge of the general characteristics of the type of deposit involved.
- Complete mineralogical and chemical information, including grain size and distribution, and useful by-products or deleterious constituents, that affect amenability of the deposit to processing.
- The three-dimensional geometry and depth of the deposit, the physical condition of the ore and its wall rocks, and other factors affecting amenability to mining methods.
- A set of clear records including reports, geological and other survey maps, drillhole data and logs, and assay data.
- Mineral resource and reserve estimates, including estimates of the degree of error attached to such basic data as sampling methods, sample preparation and assays, bulk density of ore and gangue types, the range and distribution of values within the deposit, and the continuity of mineralization between drillholes or other exposures.

The process is illustrated by some geological aspects of the case history of the Lihir gold mine, situated in a volcanic caldera on the island of Lihir, Papua New Guinea (Figure 9). After 3 years of encouraging physical exploration, in 1985, the epithermal pyritic gold ore failed the test of amenability to conventional processing. Tests on an alternative advanced technology worked well, but meant that

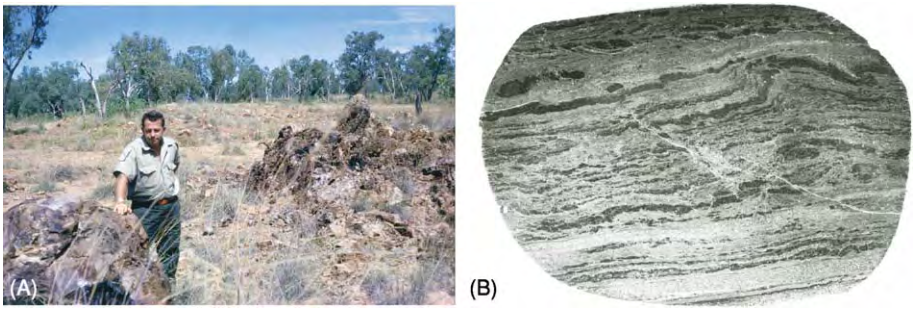


Figure 8 McArthur River, Queensland, Australia. A large stratiform pyritic Zn Pb sulphide deposit. (A) The gossanous discovery outcrop, broken open to reveal white secondary Pb and Zn minerals. (B) Polished surface (about 20 mm across) of very fine grained stratiform sulphide ore. The 2 5 micron grain size of the base metal sulphides in this large and easily mined deposit made it too difficult and expensive to process by conventional technology. Photographs: GR Davis.

Figure 9 The open pit gold mine at Lihir Island, Papua New Guinea, under development in 1998, twenty months after inception of mining operations. The two pits, then about 40 m deep, and centred on ore bodies 1.5 km apart, will merge into a pit about 2 km across, with a planned depth of 300 m. The background and foreground hills are the walls of the volcanic caldera in which the epithermal gold pyrite mineralization occurs. Photograph by courtesy of Rio Tinto PLC. Reproduced with permission from Evans AM (1995) Introduction to Mineral Exploration. Oxford: Blackwell p. 34.

only large-scale mining would provide an economic return. A further detailed drilling campaign was followed in 1992 by a feasibility study. This was based on a data bank that included 56 000 samples from 83 000 m of core from 343 drillholes, and 4479 specific gravity tests on the 12 ore types recognized. Four adjacent and partly overlapping ore zones amenable to open-pit mining were defined within an area of 2.0 km by 1.5 km. Using three computerized geostatistical methods to optimize confidence in ore continuity and distribution of gold values, an overall mineral resource of 478 million tonnes was estimated. Mining engineers then designed an open pit to mine 445 million tonnes in 15 years. Of this total, 341 million tonnes will be waste rock, 42 million tonnes will be high-grade ore reserves averaging 6.9 g/t Au to be processed through the plant, and the remaining 62 million tonnes will be stockpiled as low-grade ore averaging 2.9g/t Au to be processed during years 15 to 36. Gold production commenced in 1998, some 16 years after initial work on the discovery. As usual in a new mining area, exploration will continue, with a view to extending the life of the mining infrastructure beyond the presently planned project.

Mineral Extraction

Geological services are now widely employed in all forms of mineral extraction operations, the mainstay of the minerals industries. The proven usefulness of applied geology lies in its contribution towards optimizing the efficiency of the overall production engineering process, and the geologist is an integral member of the engineering team and its objectives. The nature of the work extends the quantitative aspects of exploration at the feasibility study stage into an even more detailed and practical realm where geological predictions are at once put to the test. Geological services fulfill the following main functions, differing only in detail across the spectrum from bulk material quarries through industrial minerals to metalliferous mines.

Maintaining mineral reserves A new mine is brought into production with only enough proved reserve to provide reasonable assurance that the invested capital will be returned. The greater part of the detailed exploration of the deposit must continue over the whole life of the operation. The planning

and execution of this vital work is the special responsibility of the geologist, as part of long-term management planning. It calls for skills in detailed geological mapping at scales up to 1/250, knowledge of sampling theory and practice, and competence in computerized data manipulation.

Services to the mining department Clean mining and low production costs depend on physically defining in the greatest detail the three kinds of material in a deposit – waste that must be dumped, payable ore or feed that goes to the processing plant, and marginal material that may be stockpiled for future processing. This daily task of the geologist at the working faces finalizes the preliminary outlines predicted from drillholes, sampling and assaying, and contributes also to quality control at the processing plant and towards planned production targets. This definitive work is as important to product quality and cost in a cement raw materials quarry as it is in an underground vein gold mine. Mining procedures rely also on information from rock mechanics, a specialized subject in which knowledge of the detailed structural geology of the deposit is linked with engineering parameters. This overlap with engineering geology occurs also in hydrogeology, as control of groundwater is important in nearly all mining operations.

Services to mineral processing plants Improvements in mineral recovery and efficiency over the life of a deposit may be brought about not only by processing technology, but by awareness of changes in the geological conditions to which the plant is tuned. The geologist is in the best position to forecast and make known changes in run-of-mine deliveries, such as grade and mineral composition of the ore, or clay content of the gangue, that are expected in long-life operations, and which upset the routine best performance of the processing plant.

Other engineering services Geological advice is often called for, especially in remote mining locations, in selecting sites for buildings, works, tailings disposal, stockpiles, etc., to select the best foundation conditions and especially to ensure that permanent works are not built in places that would sterilize future mineral resources. Geologists are also well placed to advise concerning natural hazards such as earthquakes, and landslides in mountainous terrain.

Even in so curtailed a summary, it must be mentioned that the overall usefulness of applied geology rests heavily on effective communication by geologists with the working industrial community. Good personal relationships at all levels are helpful, but clearly written reports that can be readily understood by non-geologists are essential.

End of Life of Deposit

In the environmentally conscious world of the twenty-first century, few mineral deposits are developed to the stage of extraction without a searching environmental impact investigation and statement. Similarly, when mineral workings reach the inevitable end of their productive lives, the impact of the closure may be planned so as to reduce deleterious effects and optimize beneficial effects upon the environment and the community that has grown around the deposit. For example, the Lihir gold mining project (mentioned in the feasibility study above) has from inception of operations provided finance and facilities for periodic review and action on environmental and community effects. Applied geology has entered the new professional field of environmental science in an age of positive and proactive approach.

Conclusion

From time to time ill-founded alarms have arisen that world resources of some minerals will be insufficient to meet mankind's growing demands. An influential, but ill-advised report of the Club of Rome in 1972, for instance, concluded that the limiting factor to world economic growth would be mineral resource deficiencies. Led by applied economic geology, exploration within the minerals sector has in practice demonstrated an abundant resource base open to ever-improving mining and processing technologies. There is no indication that future needs for traditional and new materials will not be met. In a crowded world, the concept of transition to sustainable development has become a serious global issue that geological science is well placed to appreciate. In the spirit of this new drive forward, the world's leading mining companies have proactively organized the Global Mining Initiative, a programme to ensure that the issues relevant to the extractive industries are positively identified and addressed. Both the theoretical and practical aspects of economic geology are involved in this broadening framework to its contribution in discovering and extracting the precious non-renewable mineral resources of our world.

See Also

Aggregates. **Geochemical Exploration**. **Mineral Deposits and Their Genesis**. **Mining Geology:** Exploration; Mineral Reserves. **Petroleum Geology:** Exploration; Reserves. **Plate Tectonics**. **Sedimentary Rocks:** Mineralogy and Classification; Banded Iron Formations; Ironstones; Phosphates.

Further Reading

British Geological Survey (2003) *World Mineral Statistics 1997 2001*. Minerals Programme Publication 13, Key worth.
Davis GR (1978) Geology in the minerals industry. In: Knill JL (ed.) *Industrial Geology*, pp. 78 110. Oxford: Oxford University Press.
Dixon CJ (1979) *Atlas of Economic Mineral Deposits*. London: Chapman & Hall.
Evans AM (1997) *An Introduction to Economic Geology and its Environmental Impact*. Oxford: Blackwell Scientific.
Evans AM (ed.) (1995) *Introduction to Mineral Explor ation*. Oxford: Blackwell Science.
Hartman L (1987) *Introductory Mining Engineering*. New York: Wiley.
Holland HD and Petersen U (1995) *Living Dangerously: The Earth, its Resources and the Environment*. New York: Princeton University Press.
Jones MP (1978) *Applied Mineralogy A Quantitative Approach*. London: Graham & Trotman.
Manning DAC (1995) *Introduction to Industrial Minerals*. London: Chapman & Hall.
Wills BA (1997) *Minerals Processing Technology*, 6th edn. Oxford: Pergamon Press.

ENGINEERING GEOLOGY

Contents

Overview

M S Rosenbaum, Twickenham, UK

Engineering geology embraces the whole of geoscience, gathering information pertinent to the (infra-)structure of society, preservation and enhancement of the environment, and sustainable development. The engineering geologist interprets this information with the relevant application of knowledge and judgement to support the engineering profession, and to protect the public and the environment.

An engineering geologist in essence reads the ground like a book, anticipating where adverse conditions might arise. The primary concern is with defining the likelihood of a geological hazard ('geohazard') occurring, whose existence may be predicted from an appropriate consideration of the

ground. This requires defining the state of the ground conditions: the geometry of the various soil and rock units, locating their boundaries; determining their structure, composition and (geotechnical) properties; identifying displacements or distortions (actual or potential), for instance those arising from neotectonic forces (stress), mining or excavation; and identifying the presence of fluids within and flow through the units comprising the ground profile.

Three disciplines are particularly important for the synthesis of engineering geology: hydrogeology, soil mechanics, and rock mechanics. Engineering geology facilitates their integration with the science of geology, particularly by bringing to bear the skills of observation, balancing idealization with reality, and injecting both experience and sound judgement. Hydrogeology concerns the character of fluids (water and gas) within the ground (as distinct from hydrology which concerns fluid flow, generally at the surface), whereas soil mechanics and rock mechanics are engineering disciplines, together forming the profession of geotechnical engineering. Each activity has its own distinct methodology and its own rigour. Each is interlinked and so there is a need to strive towards coherence and integration.

The term 'environmental geology' has evolved relatively recently, but as a discipline is really a component of engineering geology, concentrating on issues related to our industrial legacy, in particular the chemical character of the ground and groundwater, with focus on contaminated land.

The Profession

Another term requires consideration: 'geological engineering'. The distinction between this and 'engineering geology' lies in the requirements of the engineering institutions. Accreditation of education and training falls within the remit of these professional institutions. They set levels of training if direct registration of the individual is to be approved, such as the minimum time devoted to subjects like mathematics, mechanics, and design. In the United States, this is undertaken by the Accreditation Board for Engineering and Technology (ABET), leading to appointment as a Professional Engineer (PE); in the United Kingdom this is done by the Engineering Council (EC), leading to appointment as a Chartered Engineer (CEng). Such accredited courses are commonly entitled 'Geological Engineering', especially in the USA. The training of a 'Geological Engineer' in essence follows that of an engineer with additional geological knowledge, whereas the 'Engineering Geologist' remains a scientist; this difference has ramifications for professional registration and professional indemnity.

A relatively new route is now opening up for those working primarily within geoscience, accredited by the Geological Society of London, which leads to appointment as a Chartered Geologist (CGeol). This accreditation is also accepted by the European Federation of Geologists, qualifying the registrant as a Euro Geologist (Eur Geol), equivalent to a Euro Engineer (Eur Ing), but only the latter currently has legal standing.

Development

Engineering geology addresses the issues resulting from adverse ground behaviour. Records of early applications may be traced back five millennia, to the Ancient Egyptians mitigating the impact of flooding by the River Nile and constructing the Pyramids. There follow records of the Romans and Medieval Europeans utilising their understanding of engineering geology in a wide range of public works as well as for the extraction of mineral resources and support of military campaigns, notably sapping and mining. In the early nineteenth century, the creator of the first comprehensive geological map of England and Wales, the engineer William Smith, expressed the essence of an engineering geologist, applying geological principles (largely stratigraphic) to the design and construction of the British canal system as well as to groundwater control.

However, the development of engineering geology as a recognizable discipline in its own right arguably had to wait until the early twentieth century and realization of the important impact of geology on civil engineering. This was exemplified by failure of slopes along the Panama Canal in the first decade and the devastation brought about by the 1906 San Francisco earthquake, focusing attention to the wider potential adverse impact of the ground on the built environment. This encouraged the pioneering work by Karl Terzaghi, transcending the disciplinary boundaries between the science of geology and the art of engineering, seeding the growth of the geotechnical profession worldwide.

Publication lagged behind the increased understanding of ground behaviour that was being developed within the construction industry, from research underway in government-funded establishments, and from the lectures being delivered to engineering institutions and university students. Books containing 'engineering geology' in the title had been published by the 1880s, for instance the volume by Henry Penning, but were not widely distributed.

The compendium of case histories published in 1939 by Robert Legget has become an important milestone, helping establish the subject's profile, soon to be followed by 'A Geology for Engineers' by Blyth (in 1943) and the Berkey Volume (in 1950). The latter volume, entitled 'Application of Geology to Engineering Practice', was published by the Geological Society of America (GSA) in memory of the stimulating contributions by Charles Berkey, whose main professional work had been the provision of consultancy advice for dam construction but had also published his important paper, 'Responsibilities of the geologist in engineering projects' (AIME Tech. Pub. 215).

Collations of published works then followed in the early 1960s, notably the 'Reviews on Engineering Geology' series published by the Geological Society of America and the 'Engineering Geology Special Publications' produced by the Geological Society of London. The first Congress (in 1970) of the International Association of Engineering Geology (IAEG) established the discipline at an international level, since when it has provided a framework for promotion and development of the subject. The IAEG added 'and the Environment' to its title in 1997, thereby acknowledging the wider remit of the organisation.

Engineering Context

Ground behaviour of concern arises from the adverse combination of geological processes and ground conditions precipitated by human activity with the potential to cause harm. Examples may be drawn from a wide range of situations selected from many parts of the world. These include slope instability, subsidence, volume change, hazards relating to water, erosion, seismicity, volcanism, glacial and periglacial phenomena, and pollution. Knowledge concerning several processes, notably slope instability and subsidence, have reached a level of maturity but others, such as natural pollution, are still underrepresented and are in need of further research and development.

The causes of adverse ground behaviour are of scientific interest, but it is the consequences which are of greatest concern to the engineering profession. The properties of the geological profile and the processes operating within it determine where and when an adverse combination of circumstances might become linked together, quite possibly induced by human action, so bringing about an imbalance that triggers failure. Taking into account the likelihood of a geologically-related hazard occurring, together with an assessment of the vulnerability of people, property, and environment in the vicinity, provides support for decisions as to what should be done to mitigate the consequences.

Engineering geologists essentially interpret geological information for use by others, applying judgement as appropriate to provide the knowledge necessary for effective decision-making concerning resource abstraction and identification of constraints on development and regeneration. In essence, the aim is to identify the most probable geological conditions together with the most unfavourable conditions that might be plausible, and an estimate of their likelihood. This informs the broader hazard assessment and risk analysis procedures undertaken for most major engineering projects.

The general approach adopted is to build a three-dimensional model in advance of the site investigation, using existing knowledge of the distribution and properties of the geological materials and the geological processes acting upon them. Geological materials comprise three components: solids, liquids, and gases. Of equal relevance to the natural geological materials classically studied by geologists (bedrock, or 'Solid', geology) are the Quaternary deposits (superficial, or 'Drift' geology), artificial deposits ('Made Ground') and the weathered profile.

The conceptual ground model can then be used to plan the site investigation in an interactive way to reduce uncertainty without adopting unnecessarily expensive approaches (for example, close-spread fixed separations between boreholes). The potential exists for financially-based decision-making in terms of value, for example by linking reduction in geostatistical error of estimation with improvement in uncertainty (that is, risk reduction), and so develop an optimal site investigation design.

Social Context

The social and economic significance of ground failure is still largely underestimated, particularly in urban and suburban areas. Essentially the impact of each of a series of possible strategies which could be pursued needs to be measured, including the option of doing nothing. The aim should be to establish cost effective and socially acceptable management of the built environment. The scale of the effect brought about by the occurrence of adverse ground behaviour determines who might be affected and therefore who needs to know. This could range

from the owner of a property damaged by a local landslide up to the national government for a major catastrophe such as the devastation brought about by a high magnitude earthquake.

The overview and focus for advice is undertaken by the planning profession. The most important step for planners and the public alike is to become more aware of the importance of the nature of the ground and of the need to take this into account. The method of presenting relevant geological information is crucially important if it is to have effect. Users often have little or no knowledge of geology, the likely behaviour of the ground, or know what effect unexpected ground conditions and the causal processes might have on their use or enjoyment of that ground.

The most widely used tool for communicating ground conditions is the geological map, indicating the general (geological) character of the ground. However, this requires skilled interpretation to derive an indication of any potential ground-related problems. What the user requires is an indication of when and what advice is needed, and where it can be obtained, linked to a database of relevant information concerning each location.

The development of digital techniques for the manipulation and presentation of three-dimensional spatially defined data (notably Geographical Information Systems – GIS) is revolutionizing the way spatial data can be portrayed. It is now possible to interactively alter a geological model as new data becomes available from the site investigation, and soon it will be possible to overlay the proposed engineering works and iterate the design to reflect a variety of alternatives, observing the effects as the ground model changes. As important is the need to enhance public perception, increasing education, awareness, and information.

Further information

Further information on engineering geology as a profession may be obtained from the Engineering Group at the Geological Society of London (http://www.geolsoc.org.uk/template.cfm?name=geogroup10), from the Secretary of the Association of Engineering Geologists (http://www.aeg.com), and from the General Secretary of the International Association of Engineering and the Environment (http://cgi.ensmp.fr:88/iaeg/). General information on what is covered by engineering geology is summarized in Masters course introductions, for instance those at Leeds University (http://earth.leeds.ac.uk/msc/eng1.htm) and at Imperial College, London (http://www.cv.imperial.ac.uk/research/soils/engeo/enggeol%20home%20page1.html). General information on what is encompassed by the term 'geohazard' is summarized by the British Geological Survey (http://www.bgs.ac.uk/enquiries/hazards.html) and the Australian Geological Survey (http://www.ga.gov.au/urban/factsheets/geo index.jsp). The professional requirements to register as a Chartered Geologist are set out by the Geological Society of London (http://www.geolsoc.org.uk/template.cfm?name=chartered) and the requirements to register as a Euro Geologist are set out by the European Federation of Geologists (http://www.eurogeologists. de/).

See Also

Engineering Geology: Codes of Practice; Aspects of Earthquakes; Geological Maps; Geomorphology; Geophysics; Seismology; Natural and Anthropogenic Geohazards; Liquefaction; Made Ground; Problematic Rocks; Problematic Soils; Rock Properties and Their Assessment; Site and Ground Investigation; Site Classification; Subsidence; Ground Water Monitoring at Solid Waste Landfills. **Environmental Geology**. **Geological Engineering**. **Soil Mechanics**.

Further Reading

Anon (1993) *Without Site Investigation Ground is a Hazard*. Report of the Site Investigation Steering Group of the Insti tution of Civil Engineer. London: Thomas Telford.

Blyth FGH (1943) *A Geology for Engineers*. London: Edward Arnold.

Clayton CRI (2001) *Managing Geotechnical Risk*, p. 80. London: Thomas Telford.

Fookes PG (1997) Geology for Engineers: the Geological Model, Prediction and Performance. *Quarterly Journal of Engineering Geology* 30: 293 424.

Goodman RE (1999) *Karl Terzaghi. The Engineer as an Artist*. Reston; Virginia: American Society of Civil Engin eers.

Kiersch GA (ed.) (1991) *The Heritage of Engineering Geology; The First Hundred Years*. Boulder, Colorado: Geological Society of America, Centennial Special Volume 3.

Knill JL (2001) Environmental change and engineering geology: our global challenge. In: Marinos PG, Koukis GC, Tsiambaos GC, and Stournaras GC (eds.) *Engineer ing Geology and the Environment*, 4, pp. 3355 3361.

Knill JL (2003) Core values: the first Hans Cloos lecture. *Engineering Geology* 62: 1 34.

Legget RF (1939) *Geology and Engineering*. New York: McGraw Hill.

McCall GJH, de Mulder EFJ, and Marker BR (eds.) (1996) *Urban Geoscience*. Rotterdam: Balkema.

Müller Salzburg L (1976) Geology and engineering geo logy. Reflections on the occasion of the 25th anniversary

of the death of Hans Cloos. *Bulletin of the International Association of Engineering Geology* 13: 35 36.
Paige S (ed.) (1950) *Application of Geology to Engineering Practice*. Berkey Volume. New York: Geological Society of America.
Rosenbaum MS and Culshaw MG (2003) Communicating the risks arising from geohazards. *Journal of the Royal Statistical Society, Series A* 166(2): 261 270.
Terzaghi K (1925) *Erdbaumechanik auf Bodenphysikalischer. Grundlage*. Vienna: Franz Deuticke.

Codes of Practice

D Norbury, CL Associates, Wokingham, UK

Introduction

The practice of engineering geology requires effective communication of observations, test results, and a conceptual model of the ground. This communication has to be unambiguous and clearly understood if the works are to proceed smoothly. Engineering projects have become increasingly international, increasing the importance of clear communication. National codification of descriptive terminology and field and laboratory test procedures has been appearing over the last 30 years; the next step is for these national standards to be subsumed within international standards.

The History of Codification

Engineering geology as an established professional practice has been in existence for some 70 years, although some may argue that the practice has been around for as long as man has been carrying out engineering works in and on the ground. As the industry grew it became increasingly clear that the meanings of words, observations, and results were too often being misunderstood, making effective work increasingly difficult.

Initially there was no published guidance, but a range of publications aiming to standardize practice have emerged in two distinct areas. The procedures to be used in the field and in the laboratory have become standardized: guidelines have been prepared at a national level but with limited coordination between countries. However, the description of soils and rocks, which is arguably the basis of all engineering geological studies, has not seen the same rapid progress. Most of the guidance in this area has been advisory rather than compulsory, possibly because geologists tend to be independently minded practitioners.

As construction projects and engineering geology have become increasingly international, the need for common procedures and practices has increased. One of the primary aspirations of the International Organization for Standardization (ISO) is to provide such commonality, leading to better communication and fairer competitive tendering for work.

The development of standards essentially takes place in committee and is coordinated by the Comité Européen de Normalisation (CEN) and the ISO.

This article outlines the history of the development of codes in the practice of engineering geology, largely by reference to publications in the UK, which is the author's base of experience. Developments in other countries have been along similar lines at similar times, so the example of the UK situation is a useful illustration of the general picture. Examples of standards from other countries are included in the list of Further Reading. We are looking at a profession where the guidance is moving from national and advisory to international and normative.

What is the Role of Engineering Geology?

Engineering geology is a core component of the profession of ground engineering, which concerns engineering practice with, on, or in geological materials. Ground engineering is concerned with the well-being and advancement of society, including

- the safety of residential, commercial, and industrial structures,
- the essential supply of energy and mineral resources,
- the mitigation of geological hazards,
- the alleviation of human-induced hazards,
- the efficient functioning of the engineering infrastructure, and
- contributing towards a sustainable environment.

Ground engineering is based on the professional input of geologists and engineers, and specifically includes the scientific disciplines of engineering

geology, soil mechanics, rock mechanics, hydrogeology, and mining geomechanics. Examples of work activities are as follows.

- Geotechnics is concerned with the foundations of any type of building or structure, such as dams and bridges, and with excavations, slopes, embankments, tunnels, and other underground openings.
- The exploitation of natural resources involves surface and underground mining, the extraction and protection of groundwater, and the extraction of natural materials for construction, hydrocarbons, and geothermal energy.
- Geo-environmental considerations include protection and conservation of the geological environment, rehabilitation of contaminated land (soil and groundwater) and of mining areas, waste disposal (domestic and toxic), and the subsurface emplacement of chemical and radioactive waste.
- Geo-risk is the process of mitigating geological hazards (e.g. earthquakes, slope instabilities, collapsible ground, gas) in land-use planning.

Ground engineering is of considerable economic importance and benefit to society because it provides a means of building efficient structures and facilitating the sustainable use of resources and space. This is often not fully appreciated by the general public. In stark contrast to other engineered structures, most geoengineered solutions are hidden in the ground and so are not visible. Nevertheless, ground-engineered structures can present a major challenge to engineering design and construction and, if successfully completed, are testament to substantial technological and intellectual achievements.

The execution of such projects requires input from a range of scientific and engineering specialties, and the relevant specialists must be able to communicate with each other in order to agree on conceptual models and parameters to apply to the design and must leave an audit trail to ensure quality and safety. In addition, and perhaps even more importantly, there is a need to communicate with other interested parties, not least the owner of the project and the general public. Subjects on which efficient communication is required include observations of the condition of the ground in and around the works and the quality of that ground as revealed by physical records, the logging of cores or exposures, and parameters measured in field and laboratory tests. In order for such communication to be possible, an internationally agreed library of linguistic and scientific terminology, test procedures, and overall investigation processes needs to be available.

What are Codes in Engineering Geology?

Before considering the trends and requirements in the codification of the practice of engineering geology, it is important to remind ourselves of the role of the practitioner in this field. The fundamental role of an engineering geologist is to observe and record evidence of geological conditions at the site of proposed or current engineering works and to communicate these observations to other (non-geological) members of the team. The evidence for the ground conditions may be in the form of exposures, such as cliff or quarry faces, or may be in the form of cores or samples recovered from boreholes. It is almost universally the case that this geological information comes from the proximity, but not the actual location, of the proposed works. There may also be indirect information, such as the results of geological mapping or geophysical surveys or evidence from previous engineering works in the same area or geological setting. The engineering geologist therefore has to develop an understanding of the geology of the area and make predictions about the geology that will be encountered by, or will affect or be affected by, the engineering works. It is rare for the geologist to have sufficient information to understand the ground conditions fully, and there is always a point beyond which further investigation cannot be justified by a further reduction in uncertainty. It is therefore not uncommon for the geologist to have less information than might be obtained from a small number of boreholes. For instance, road and rail tunnels driven at low level through mountains cannot sensibly be investigated: borehole locations may not be available, and the cost of drilling hundreds of metres before reaching the zone of interest can be prohibitive.

Notwithstanding the source and detail of the information available, the engineering geologist has to collate and interpret the geological information in order to produce a realistic geological model that includes realistic assessments of the degree of uncertainty. The first stage is to create an essentially factual model, before moving on to the interpretation phase. The key aspect of the engineering geologist's role then comes into play: the communication of all aspects of this conceptual model to other members of the design team, the project owner or client, and, increasingly, the public.

To some extent this communication of information can be carried out using existing geological nomenclature in a qualitative sense. However, such an approach by geologists has often left listeners confused. Usually, standard geological nomenclature

is qualitatively, rather than quantitatively, defined, so even other geologists can be left uncertain as to the exact meaning intended. Over the years, a language of better-defined terms has developed, which should better enable geologists to communicate, not least because there is now a core of standard terminology with which the listeners will be familiar. It is the derivation and definition of this standard terminology that is one of the main reasons for recent advances in the drafting and implementation of codes in engineering geology.

The intention of the ISO and CEN in preparing international standards is to help raise levels of quality, safety, reliability, efficiency, compatibility, and interchangeability, and to provide these benefits economically. Standards contribute to making the development, manufacturing, and supply of products and services more efficient, safer, and cleaner. They make trade between countries easier and fairer. ISO and CEN standards also safeguard consumers, and users in general, of products and services and make their lives simpler.

The History of Codification in Description

Prior to 1970, there was no international standard terminology to allow the communication of descriptions of geological materials or their properties, although some of the larger contracting companies had developed internal guides. The first British Code of Practice (CP 2001) was published in the UK in 1957. This code laid down key underlying precepts for the description of soils, in particular that soils should be described in terms of their likely engineering behaviour. This basic need is often lost in today's welter of published guidance. However, CP 2001 did not cover the description of rocks. Geological sciences continued to develop an increasingly variable use of terminology, but the increasing size of the ground engineering industry made the terminology increasingly irrelevant and no longer tenable. The need for a defined and wide-ranging terminology had arrived.

The use of undefined and narrow terminologies caused confusion and ambiguity in communication and, as a result, frequent contractual arguments arose, often leading to claims based on 'unforeseen ground conditions'. This was hardly surprising as, if the terminology is variable and undefined, there will always be someone who could misread the ground conditions being predicted. For instance, terms such as 'highly fissured' and 'moderately jointed' were not defined and therefore meant different things to different readers. This was addressed when the Engineering Group of the Geological Society of London published, in the early 1970s, a series of Working Party Reports for guidance, for example on core logging (in 1970) and the preparation of maps and plans (in 1972). These reports formed the basis of UK practice and, as it turned out, international practice in many respects. Similar activities on the international scene resulted, by 1981, in publications from the International Association of Engineering Geologists and the International Society of Rock Mechanics on field investigation, geological mapping, and soil and rock description. In the UK, this decade of guidance culminated with BS 5930 (published in 1981), the seminal National Standard for engineering geological activities. It is important to note, however, that even at this stage BS 5930 was designated as a code of practice, meaning that the guidance was advisory rather than normative (i.e. mandatory). This designation was maintained in the updated version of BS 5930, published in 1999. The reasons for this relate to the preference of many geologists to not be given edicts on geological terminology. However, the designation of codes as advisory has little practical relevance. The codes are referenced in contract specification documents and thus become binding. In legal arguments about claims or failures, the courts will expect the national guidance to have been followed. Therefore, at least by default, the codes of practice have become *de facto* Standards.

The publication of international guidance has continued, albeit at a slower pace, with individual nations publishing national guidance documents. The designation of such guidance as mandatory or advisory varies, and it is this variance that is now being resolved in the name of international normalization.

Particular Problem Areas in Combining National Codes

Despite the codifications in various countries proceeding independently, there are many cases where the practices of one country have been adopted by another. For this reason, the preparation of international codes by the ISO has not been as difficult as might have been anticipated, at least as far as the description of soils and rocks is concerned. There have nevertheless been a few difficulties, as outlined below. However, the historical development of local codes has tended to reflect and emphasize local

geological conditions, and the classifications were rather more difficult to bring together into an all-embracing international standard. This proved particularly difficult for the classification of soils and resulted in the need for a simple and separate ISO Standard on this topic.

The Scandinavian countries have different types of soil (coarse glacial deposits and fine 'quick' clays) from Italy (volcanic sands) and Japan (silts and loose sands that are liable to liquefaction). National practice has, for sound technical reasons, tended to reflect the regional geological source materials, active geological processes, and impacts of climate and time of exposure, all of which vary across the world. The work carried out by the ISO Technical Committees has needed to incorporate these variations into practice.

Fine-Grained Materials

The definition of the grain-size boundaries used to classify soils and rocks for engineering purposes has, by and large, developed along similar lines throughout the world, as shown in Table 1. Exceptions to this are the USA, where a range of definitions are available, and some Pacific Rim countries. The agreement is wide ranging and complete except for the finer-grained materials, specifically rocks, where terms such as claystone, siltstone, mudstone, shale, and slate are all variously used. The simplification of the terminology to mirror the descriptive terms for soils is proposed in the latest codes.

Weathering Classifications

Weathering classifications are provided in a number of the existing national and international guidance documents, but there is little commonality between publications. The provision of such classifications within a descriptive framework has proved unhelpful, as weathering profiles are rarely amenable to global classification (because they depend strongly on climate and local topography) and fail to embrace the actual description of the weathering features that are present. For this reason, UK practice, since the 1999 publication of BS 5930, requires the description of weathering profiles. Further classification schemes are permitted only if relevant, available, useful, and unambiguous. However, this practice is not yet generally accepted at an international level.

Core Indices

Core indices such as Rock Quality Designation (RQD) have been proposed for the logging of rock not only from borehole cores but also from the mapping of exposures, in order to provide a ready indicator of rock quality. A number of indices and a range of definitions have been proposed, without much commonality. Recent International Standards have addressed this and provide unambiguous and clear definitions of which fractures should and should not be included in the quality assessment and of how the index should be measured. Although the RQD is only a crude indicator of rock quality, it is very widely used, on its own and correlated with other properties,

Table 1 Particle size definitions

Soil name	*Soil fractions*	*Dimensions*	*Sedimentary rock name*
Boulders		over 200 mm	
Cobbles		200 60 mm	
Gravel	Coarse	60 20 mm	
	Medium	20 6 mm	Conglomerate (boulders, cobbles and gravel)
	Fine	6 2 mm	
Sand	Coarse	2 0.6 mm	Coarse grained sandstone
	Medium	0.6 0.2 mm	Medium grained sandstone
	Fine	0.2 0.06 mm	Fine grained sandstone
Silt	Coarse	0.06 0.02 mm	
	Medium	0.02 0.006 mm	Siltstone (coarse, medium and fine)
	Fine	0.006 0.002 mm	
Clay		less than 0.002 mm	Mudstone or claystone

Few soils comprise merely a single principal size fraction. Soils usually include secondary constituents of a different size fraction that could significantly influence their engineering behaviour. There are also minor constituents that can help in geological identification but do not influence characteristic behaviour. Claystone is used in some countries to match the soil terms, but this can be confused with nodular concretions. In these cases the term Mudstone is preferred.

in rock-mass classification systems (e.g. compressibility, diggability, tunnel stand-up time). The lack of consistency in the measurement of the index is not widely appreciated, and so the international and normative standardization of the index is long overdue. This neatly illustrates the need for standardization wherever a property, however simple, is used to communicate a measure of ground quality. It will be valid only if the measurement process is consistently applied and understood.

The ongoing evolution of the codification of terminology and test procedures means that users need to be aware of the date of data collection when considering the descriptive and measurement standards that are likely to have been incorporated in borehole logs, maps, and records. In particular, archival records are likely to have been prepared according to different, outdated, Standards and definitions.

History of the Codification of Field and Laboratory Testing

The standardization of field and laboratory procedures is as important as the standardization of the description and includes forming the borehole or describing the exposure, executing field tests, recovering samples, and carrying out laboratory tests. If the results of any of these activities are to be applicable and relevant in the minds of others, the procedures used need to be clearly identifiable and standardized.

The codification of laboratory test procedures for soils commenced in the 1940s, when the first machines were developed. The procedures covered everything from how to drill a borehole, conduct the basic field tests, and take and describe samples, to their storage, transport, and laboratory testing. In fact, even the design of the testing machines had to be specified as test procedures evolved. Work by the ISO and CEN that is underway at the time of writing will describe the procedures and concentrate on the preparation of Technical Specifications (Table 2), building on earlier test procedures drafted by the International Society of Soil Mechanics.

The testing of rocks in commercial practice started somewhat later, by which time the need for international cooperation was better appreciated. It has therefore been possible for the rock-testing procedures to be better established, under the auspices of the International Society of Rock Mechanics, who have published a series of suggested methods. As there were no precedent procedures in place, these suggested methods have been rapidly adopted by the professional community and have become internationally recognized without the involvement of national standards bodies. It would therefore be comparatively straightforward, but not necessarily easy, to prepare normative international standards for most rock tests.

Professional Qualifications

A further aspect of codification is the identification of relevant qualifications and experience necessary for those who plan, execute, and interpret ground investigations. The guidance documents and codes prepared up to the end of the 1980s did not try to lay down rules about the qualifications and experience required by those working as specialists in engineering geology. In the very early days, this was felt to be unnecessary; the number of practitioners was small, and their capabilities and limitations were known through reputation. This has become increasingly less reliable, with even academic programmes of training changing significantly from previously accepted standards of coverage and achievement. It is now necessary to define the roles and who may be permitted to practice. It is interesting to consider who benefits most from such codification. Is it the client, who can feel better protected by the knowledge that they have properly trained and experienced professional advisors, is it the insurers, who feel they have lower exposure to risk, is it the individual practitioners, who feel this improves their status in society, is it the employers, who can recognize a qualified practitioner, or is it the companies, who can see a market with fair competition? The truth is probably a combination of all of these. The position taken by the Standards Institutions is based on the latter view, and ISO/CEN documents currently in preparation include definitions of specialist practitioners. These definitions will therefore become requirements in the practice of ground engineering.

The development of such defined roles is closely linked to the development of Directives in the European Union. In order to facilitate the mobility of workers, the availability of internationally recognized qualifications is essential. The Directive on this subject is likely to be enacted by about 2005. The requirement for practitioners to be able to practice, at least for limited periods, in any European Union country is the holding of a recognized qualification. This qualification is likely to be represented by the common platform of the European Federation titles of EurIng, awarded by the European Federation of Engineers, and EurGeol, awarded by the

Table 2 Codes in preparation by CEN/TC 341 and ISO/TC 182/SC 1 concerning geotechnical investigation and testing

ISO reference number	*Title*	*Committee draft*	*Draft International Standard*	*Final draft International Standard*	*Publication as EN/ISO standard*	*Remarks*
14688 1	Identification of soil			2002 March	2002 June	ISO
14688 2	Classification of soil		2001 June	2002 September	2002 December	ISO
14689	Identification of rock		2001 September	2002 September	2002 December	ISO
22475 1	Sampling methods	2003 September	2004 September	2006 March	2006 June	
22475 2	Sampling Qualification criteria	2003 September	2004 September	2006 March	2006 June	
22476 1	Cone penetration tests	2003 September	2004 September	2006 March	2006 June	
22476 2	Dynamic probing	2001 December	2002 September	2004 September	2004 December	
22476 3	Standard penetration test	2001 December	2002 September	2004 September	2004 December	
22476 4	Menard pressuremeter test	2003 September	2004 September	2006 March	2006 June	
22476 5	Flexible dilatometer test	2003 September	2004 September	2006 March	2006 June	
22476 6	Self boring pressuremeter test	2003 September	2004 September	2006 March	2006 June	
22476 7	Borehole jack test	2003 September	2004 September	2006 March	2006 June	
22476 8	Full displacement pressuremeter	2003 September	2004 September	2006 March	2006 June	
22476 9	Field vane test	2003 September	2004 September	2006 March	2006 June	
22476 10	Weight sounding test	2002 June			2002 December	TS
22476 11	Flat dilatometer test	2002 June			2002 December	TS
22477 1	Testing of piles	2003 September	2004 September	2006 March	2006 June	
22477 2	Testing of anchorages	2003 September	2004 September	2006 March	2006 June	
22477 3	Testing of shallow foundations	2003 September	2004 September	2006 March	2006 June	
22477 4	Testing of nailing	2003 September	2004 September	2006 March	2006 June	
22477 5	Testing of reinforced fill	2003 September	2004 September	2006 March	2006 June	
17892 1	Water content	2002 December			2003 June	TS
17892 2	Density of fine grained soils	2002 December			2003 June	TS
17892 3	Density of solid particles	2002 December			2003 June	TS
17892 4	Particle size distribution	2002 December			2003 June	TS
17892 5	Oedometer test	2002 December			2003 June	TS
17892 6	Fall cone test	2002 December			2003 June	TS
17892 7	Compression test	2002 December			2003 June	TS
17892 8	Unconsolidated triaxial test	2002 December			2003 June	TS
17892 9	Consolidated triaxial test	2002 December			2003 June	TS
17892 10	Direct shear test	2002 December			2003 June	TS
17892 11	Permeability test	2002 December			2003 June	TS
17892 12	Atterberg limits	2002 December			2003 June	TS

TS: technical specification, not initially normative.

All document drafting committees are led by CEN except where noted to the contrary.

European Federation of Geologists.These titles show that the bearer has undertaken an appropriate course of study, carried out appropriate training, and gained sufficient experience to be able to act as a professional engineer or geologist, and that this record has been submitted to the bearer's peers for validation. The holder of such a title agrees to work within the code of conduct operated by the awarding Federation and will be able to work in any European country, at least for limited periods, without needing to qualify separately in that country. These are major developments in providing commonality of professional standards and represent development exactly as hoped for by the ISO and CEN, but driven forward by the European Commission.

Codification into the Twenty-First Century

After 25 years in preparation, the suite of Eurocodes is, in the early 2000's, becoming a reality. These Eurocodes bring together codes of practice for building and civil engineering structures, and provide a world-class standard for all aspects of construction. Included in Eurocode 7 (geotechnical design) are elements of codes on the description and classification of soils and rocks, field investigation methods, field and laboratory testing, assessment of engineering parameters, and design procedures. For the first time, engineering geologists throughout Europe will be talking in a common language when reporting the findings of their field observations. This progress extends beyond Europe, as, in accordance with the Vienna agreement, the standards drawn up by CEN and ISO undergo parallel voting procedures for common adoption. Thus, for example, the proposals for the description of soils and rocks prepared by the ISO in 2003 will also be incorporated into Eurocode 7. Thus, engineering geologists around the world will be able to pass on their geological information without misunderstanding and ambiguity. The Japanese 'brown sandy clay' will be the same as the Swedish 'brown sandy clay'. Similarly, the results of field or laboratory testing will be transferable around the world. Major exceptions to this rule are China and the USA, who are not members of either of these standards bodies and who have had no input into the drafting of the codes.

A schedule of the codes and specifications being prepared by the ISO and CEN in this area is given in Table 2. The codes for the description and classification of soils and rocks were prepared by a Technical Committee of the ISO. The reason for this work item being proposed initially was, in accordance with the ISO mission, to encourage and allow international communication in applied science and therefore to allow more and fairer competition in international trade. This enshrines the concept of engineering geologists the world over sharing a single interpretation of a conceptual ground model. They will therefore be able to compete equally for contracts and cooperate more effectively on design briefs.

The primary benefit of such codes is that they will standardize ground investigation, reporting, and design approaches that are common over much of the world. The market for investigation and design, and thus the market for engineering geology, thus becomes both much larger and much more competitive, but with a level playing field.

However, the Eurocodes do not completely subsume the national practices, which have been built up over the years and which are pertinent to local situations. This is right and proper given that variations in engineering geological practice have their base in the different geological conditions in different countries. For instance the Scandinavian extensive crystalline basement crust and soft 'quick' clays varies significantly from the deep weathering profiles of the tropics. These different conditions require different approaches to investigation and testing. The description of the materials can nevertheless be based on a single standardized approach.

The national differences in the approach required can be incorporated into National Annexes, which allow key safety and technical issues to remain a national rather than a European responsibility and allow geological and climatic variations to be taken into account. However, these National Annexes are enhancements of, rather than local rewrites of, the overarching international codes.

Concluding Remarks

Over a remarkably short time-span, engineering geology has developed from a situation where the early professionals acted as individuals and managed without codes to today's world where teamwork predominates and there is an increasing need for accountability. The codes coming into place in the 2000s define how we should drill holes, take and test samples, and describe soils and rocks. Perhaps the biggest change will be that we will have to carry internationally recognized qualifications if we want to be able to practice all over the world.

See Also

Engineering Geology: Geological Maps; Site and Ground Investigation; Site Classification. **Geological Engineering**. **Geology, The Profession**. **Geotechnical Engineering**. **Sedimentary Rocks:** Mineralogy and Classification. **Tectonics:** Fractures (Including Joints). **Weathering**.

Further Reading

ASTM D 4879 89 (1989) *Standard Guide for Geotechnical Mapping of Large Underground Openings in Rock*. West Conshohocken: ASTM International.

ASTM D2487 93 (1993) *Classification of Soils for Engineering Purposes (Unified Soil Classification System)*. West Conshohocken: ASTM International.

ASTM D3282 93 (1993) *Standard Classifications of Soils and Soil Aggregate Mixtures for Highway Construction Purposes*. West Conshohocken: ASTM International.

ASTM D5878 95 (1995) *Standard Guide for using Rock Mass Classification Systems for Civil Engineering Purposes*. West Conshohocken: ASTM International.

ASTM D653 96 (1996) *Standard Terminology Relating to Soil, Rock, and Contained Fluids*. West Conshohocken: ASTM International.

Brown ET (ed.) (1978) *Rock Characterisation, Testing and Monitoring. ISRM Suggested Methods*. Oxford: Pergamon Press.

BS 5930 (1981) *Code of Practice for Site Investigations*. London: BSI.

BS 1377 (1990) *Methods of Test for Soils for Civil Engineering Purposes: Parts 1 to 9*. London: BSI.

BS 5930 (1999) *Code of Practice for Site Investigations*. London: BSI.

Deere DU and Deere DW (1988) The rock quality designation (RQD) index in practice. In: *Rock Classification Systems for Engineering Purposes*. ASTM STP 984. West Conshohocken: ASTM International.

DIN 4022 1 (1987) *Classification and Description of Soil and Rock. Borehole Logging of Rock and Soil not Involving Continuous Core Sample Recovery*. Berlin: Deutsche Norm.

DIN 18196 (1988) *Soil Classification for Civil Engineering Purposes*. Berlin: Deutsche Norm.

Engineering Group of the Geological Society of London (1970) The logging of rock cores for engineering purposes. Working Party Report of the Engineering Group of the Geological Society of London. *Quarterly Journal of Engineering Geology* 3: 1 25.

Engineering Group of the Geological Society of London (1972) The preparation of maps and plans in terms of engineering geology. Working Party Report of the Engineering Group of the Geological Society of London. *Quarterly Journal of Engineering Geology* 5: 297 367.

Engineering Group of the Geological Society of London (1977) The description of rock masses for engineering purposes. Working Party Report of the Engineering Group of the Geological Society of London. *Quarterly Journal of Engineering Geology* 10: 355 388.

Engineering Group of the Geological Society of London (1995) Description and classification of weathered rocks for engineering purposes. Working Party Report of the Engineering Group of the Geological Society of London. *Quarterly Journal of Engineering Geology* 28: 207 242.

IAEG (1981) Rock and soil description and classification for engineering geological mapping. Report by IAEG Commission on engineering geological mapping *Bulletin of the Engineering Geology and the Environmental* 24: 235 274.

ISO (2003) *Geotechnical Engineering Identification and Classification of Soil Part 1 Identification and Description*. EN 14688 1. ISO/TC 182/SC 1. Geneva: International Organization for Standardization.

ISO (2003) *Geotechnical Engineering Identification and Classification of Soil Part 2 Classification*. EN 14688 2. ISO/TC 182/SC 1. Geneva: International Organization for Standardization.

ISO (2003) *Geotechnical Engineering Identification and Classification of Rock*. EN 14689. ISO/TC 182/SC 1. Geneva: International Organization for Standardization.

Japanese Geotechnical Society (2000) *Method of Classification of Geomaterials for Engineering Purposes*. 0051 2000. Tokyo: Japanese Geotechnical Society.

Kulander BR, Dean SL, and Ward BJ (1990) *Fractured Core Analysis: Interpretation, Logging and Use of Natural and Induced Fractures in Core*. AAPG Methods in Exploration Series 8. Tulsa: American Association of Petroleum Geologists.

Swedish Geotechnical Society (1981) *Soil Classification and Identification D8:81*. SGF Laboratory Manual, Part 2. Byggforskringsrådet. Linköping: Laboratory Committee of the Swedish Geotechnical Society.

Aspects of Earthquakes

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

The main product arising from the work of engineering geologists is their site characterizations for engineered works. In terms of the impact of earthquakes and their mitigation, design engineers, architects, and planners need to both accommodate people and provide functionality. With these goals in mind, engineering geologists develop three-dimensional descriptions of the ground to the depth that will be affected by the static loads resulting from the engineered works and by the expected levels of earthquake-induced ground motion.

The damaging effects of earthquakes can be mitigated, but the earthquakes themselves cannot yet be controlled or stopped. Engineering geology can be applied in a variety of ways to assist in mitigating the effects of earthquakes; for instance by

- recommending the avoidance of sites that are likely to experience ground displacement as a result of fault movement;
- grading the site to a configuration that promotes stability of the ground mass;
- improving the engineering characteristics of the ground so as to resist damaging deformation;
- controlling groundwater that is likely to experience excess pore pressure leading to ground failure under the involved structural loads; and
- creating green space to isolate active faults and possible seismically induced slope failures.

Engineering geologists also record details of and analyse the damage done by earthquakes (**Figure 1**). In this context, the role of the geologist is expanded from considering the known structural-loading conditions imposed by the designers to considering the unknown, but anticipated, ground motion during a hypothetical earthquake felt to be appropriate to the role and function of the project. Worst-case-earthquake design predictions are appropriate for critical facilities or for projects housing dependent populations. Critical facilities are those projects for which the consequences of failure are intolerable.

Worst-case-earthquake design is based on the concept of the maximum credible earthquake (MCE), which is the greatest magnitude and duration of strong motion that can reasonably be expected to occur.

Ancillary to this concept is risk analysis to define magnitudes and durations of strong motion for which there are cost benefits and for which damage can be tolerated.

Figure 1 Engineering geological mapping records evidence of earthquake motion such as this railroad deformation indicating north to south shortening of the ground along the rail right of way. Most evidence of earthquake damage is fragile and likely to be destroyed by human activity within hours or days. (Photographed by Richard J. Proctor, then Chief Engineering Geologist, Metropol itan Water District of Southern California, at San Fernando Valley, California, February 1971.)

Why Ground Fails During an Earthquake

The behaviour of the ground during an earthquake depends not only on the character of the incidental ground motion but also on the properties of the site subsurface. Engineering geology is used in this context to detect and delimit bodies of ground that are most susceptible to earthquake damage and to identify, where possible, the interfaces between geological units where shaking displacements may be concentrated or otherwise focused.

Engineering geological investigation looks for horizons and pockets of damage-susceptible soil and potential surfaces of failure (displacement, created where differential stress is greatest and exceeds *in situ* shear strength). For rock masses, the failures representing displacements generally occur along pre-existing discontinuities.

Geometry plays a strong role in the development of earthquake-induced failure conditions. In the worst case, the gross motion vector generally enters a hillside mass and leaves at a hillside or some form of cut slope where there is no restraint to the passing, outward-bound ground motion, hence leading to failure. River banks and incised stream valleys are very susceptible to such conditions. Added to this basic geometry, if rainfall has been heavy for a week or more, the earth media have been infiltrated by rain or snowmelt, or, for reservoirs, drawdown has been recent and of some magnitude (a few metres or more) then a significantly unstable situation will result.

Where soil masses serve as the foundation and have little or no lateral ground support, new soil failure surfaces can be created leading to a condition of lateral spreading, where the foundation shifts in the direction of least lateral support, generally causing considerable damage to the affected component of the engineered structure.

Another poor geological situation is where major open rock joints strike at an oblique angle to the face of a hillside or cut, dipping towards the face at a dip angle greater than that of the internal friction of the failure surfaces, which themselves form a wedge of rock pointing out of the hillside.

Little warning of impending failure is given (perhaps just a couple of seconds in which audible grinding noises are heard) before the accelerated mass of debris, separated from the hill by the surfaces of the geometric wedge, begins to move.

Earthquake strong motion typically lasts between 10 and 30 seconds, which, given the presence of the instability factors listed above, is long enough to create a failure motion.

Linking Earthquakes to Ground Effects

As already noted, engineering geologists assess the likelihood of earthquakes occurring, based on data from observed and historic earthquakes in the same seismotectonic zone. Geological observation becomes the basis for predicting the potential occurrence of various ground effects.

Near-region geological evidence indicates the general seismogenic character of the site. Relevant factors include seismic history (magnitude and frequency of recurrence), palaeoseismic patterns, geomorphological evidence, fault sources, and the delimitation of seismogenic zones within which individual capable faults occur.

Displacement effects occur when bodies of earth materials are displaced along curvilinear failure surfaces (in soil) or along pre-existing discontinuities (in weak rock). Normally the displacement occurs in the direction of least lateral constraint.

Slope failures may affect unstable masses, defined by hillside geometry, valley walls, river banks, and cut slopes, when the arriving earthquake ground motion acts momentarily within the up-gradient mass to reactivate pre-existing rock discontinuities or newly formed curvilinear failure surfaces. Gravitational acceleration moves the disequilibrated earth mass down the slope.

Liquefaction is a dynamic process that mainly affects cohesionless unconsolidated soil in the presence of near-surface groundwater. Earthquake strong motion raises pore-water pressure in such strata, mobilizing the soil, often upward along linear or pipe-like channels, which may reach the surface to generate a fountain, leaving geomorphological evidence in the form of sand blows.

Geologically Based Mitigation

One of the most successful methods of mitigating potential earthquake damage is to employ engineering geological studies to devise methods by which human development can be governed or designed so as to be less susceptible to seismic damage.

Avoiding Damage-Prone Areas

Development projects should avoid damage-prone areas, and site characterization should be employed to discover and delimit ground at an unacceptably high risk. This is naturally difficult to accomplish where land development is governed by commercial interests and imperatives. There are two basic means by which development can be limited; both have been successfully employed in the State of California using local

laws (generally enacted as city or county ordinances) and State laws.

The first measure is the creation of a greenbelt, whereby ground felt to be prone to damage is assigned to park or promenade use, thereby remaining landscaped but available for use by residents rather than being given over to the construction of residential housing or commercial buildings.

The second measure is the zonation of capable (active) faults. This was devised by the California legislature in 1969. Named after its politician promoters, State Senator Alquist and State Assemblyman Priolo, the Alquist–Priolo Act requires the California Geological Survey to define exclusion zones (440 m wide) bounding the most likely central-axis location of the subject fault and also designed in band-width to meet the usual California alluvial fault-rupture situation in which repetitions of historic ground ruptures often occur with positional variance. The maps are updated regularly and are available for public inspection.

Improving Ground Conditions

For engineered works that must be built in seismically active regions, but not within or adjacent to active faults, there remains a variety of geotechnical methods that can improve the ability of the foundation or the surrounding ground to resist the expected MCE ground shaking. Such geotechnical methods generally fall into one of the following categories, all of which require engineering geological input in order to be successfully designed and emplaced.

- Dewatering avoids the pore-pressure build-up that may cause liquefaction or other forms of ground failure. Measures can be passive (e.g. French drains and vertical gravel-packed columns, wick drains and subhorizontal drain pipes) or active (e.g. dewatering points or wells).
- Retention generally uses iron or steel tiebacks and ground anchors to counter vectoral earthquake motion acting outwards from the foundation mass towards unconfined ground (such as hillsides and cuts) and upward acceleration towards the ground surface itself.
- Densification of the ground is usually achieved by dynamic compaction using falling weights (e.g. a demolition ball dropped from a crane).
- Less stable layers, pockets, or zones may be excavated and replaced. Selection of this option is generally controlled on the basis of (low) cost.

Predicting Collateral Damage

Earthquake-induced ground motion can dislodge or otherwise set in motion boulders, joint blocks, and slope masses. These situations generally can be identified from photogeological interpretation and confirmed by geological mapping. Each situation will have a slightly different nature, but is likely to fit into one of the following categories.

- Boulder dislodgment is the displacement of boulders by long-term geological processes (mainly slope erosion) or earthquake excitation. Much has been done by the Utah Geological and Mineralogical Survey to mitigate this hazard by educating land developers and home buyers. The situation is perhaps worst along the front of the Wasatch Mountains, at Salt Lake, and in Utah cities between Provo and Ogden.
- Rockfall is perhaps the most prevalent and dangerous of the slope-motion earthquake threats, recognizable mainly along major transportation routes through rugged mountain terrain, especially in areas and at elevations not supporting dense vegetation that would obstruct the passage of debris. The leader in the mitigation of individual-event rockfall is the Colorado Geological Survey.
- Slope failure has been noted primarily in the South American Andes and in the Coast Range immediately south of San Francisco, California. The South American slope failures have been particularly spectacular and perhaps the most devastating in terms of loss of life. Here, the dominant geological factors have been extreme topographical relief, narrow mountain-front valleys, concentrated human occupation in valley-sited towns, and, of course, the general position on the actively subducting Nazca tectonic plate. It is probable that each large earthquake in the greater Andean region will kill hundreds or thousands of people in these mountain-front towns, because huge masses of ground, often supported on naturally compressed air masses, will be dislodged so rapidly that the use of alarm sounds and evacuation principles will be impractical.
- Engineering geological evidence relies on the recognition of the topographical circumstances that will lead to the occurrence of tsunamis in coastal regions. Palaeoseismic evidence of tsunami inundation has been observed along the coast of Oregon. In continental interiors, vertical components of earthquake-induced ground rupture can create a similar inundation effect alongside lakes. All that is required is instantaneous displacement of the down-thrown side of a fault or an associated above-lake slope failure. Either case would displace masses of soil or rock into the body of the lake. This situation is very difficult to mitigate in terms of public rights and awareness.

Seismotectonic Zonation

Zonation is a technique whereby known and predicted earthquake damage can be assigned to areas for planning, zoning, and regulatory purposes. A number of concepts apply to zoning. Each determination should be made on site-specific needs.

Active Faults

Where ground rupture or the associated effects of surficial damage have been noted it is appropriate to assign a status of 'active' faulting. State-of-the-art determination is represented by the Alquist–Priolo Active Fault Zones, legislatively mandated in 1969 by the California Legislature. These zones are named, delineated, and maintained by the California Geological Survey and constitute 400 m wide strips centred along the best-identified trace or centre line of the subject fault. Zones are plotted on 1 : 24 000 US Geological Survey topographical quadrangle bases and are available for reference on the world wide web.

Credible Faults

The definition of credible faults was an innovation of the nuclear power plant safety programme of the US Nuclear Regulatory Commission (USNRC); credible faults are known faults or areas where there is strong geomorphological or structural geological evidence of fault displacement at least once during the Holocene and more than once in the past 500 000 years. The half-million-year bracket is an artefact of the fusilinid age determination resolution affinity of California marine seashore terraces, and was applied to the Diablo Canyon Nuclear Power Station (NPS) in the early 1960s. The concept, however, is not applicable elsewhere, though it is frequently specified. Credible faults are considered capable of experiencing ground rupture and propagation of ground motion along their entire known length. Bonilla was the first to associate ground-rupture length with the magnitude of the causative event.

Deterministic Design Input

Design-level input by engineering geologists forms the basis for carrying actual earth science evidence into 'seismic withstand' design by civil and other engineers. This geological input consists of the following process of analysis and evaluation. First, the area around the project is seismically zoned, and credible faults are identified or a most likely epicentre for a candidate earthquake is assigned for each seismogenic zone identified. Then, where active faults are defined, the design-input ground motion is considered to occur at the location closest (measured perpendicular to the fault trace) to the site. Where an active fault has not been identified, a most likely location is selected for each seismogenic zone; the expected ground motion is attenuated from the hypocentre to the site. This becomes the MCE for each of the candidate source faults or zones. Finally, the style of earthquake to be represented for each of the candidates is evaluated, in terms of vector displacement, duration of strong motion, and peak ground acceleration.

Attenuation

This assessment looks at the ground motion from the hypocentre to the project. Waves of ground motion move through the earth media, are concentrated in the bedrock, and there suffer the physical effects of trying to displace the surfaces of innumerable rock discontinuities encountered during the passage to the site. This infinite number of surface-to-surface grinding processes consumes much of the energy of the earthquake.

Attenuation curves have been developed empirically for a variety of geological terranes and are published in the literature, particularly in the reports of the former US Army Waterways Experiment Station (now Engineer Research and Development Command, Vicksburg, MS, USA).

Each geological terrane has its own directional effect on attenuation. The greatest effect is generally exerted perpendicular to strongly foliated older metamorphic terranes, such as the north-east-foliated Lower Palaeozoic sequence of New England, USA.

Comparison of Candidate Design Earthquakes

A selection of potentially impacting design earthquakes are defined in terms of the parameters listed above. Each candidate is moved through a defined scenario and attenuated from each stipulated source to the site. The worst cases are then declared the MCEs, and their characteristics are used as input in seismic-withstand design.

Engineering Geological Mapping

A primary method for addressing and meeting seismic threats uses engineering geological mapping, which records the evidence of past earthquake damage and uses this alongside geological conditions.

Earthquake damage was generally not mapped in detail until the Alaskan 'Good Friday' earthquake of 1964. The former (regrettably now defunct) Branch of Engineering Geology responded to this event under the constructive leadership of its Director, the late Edwin B. Eckel, also later the Executive Director of the Geological Society of America. Eckel's lobbying in Washington DC was immediate, and the Survey

mounted a field effort to observe and map the damage of the magnitude 8.3 event as a long series of Professional Papers, the content of which has served as the ongoing model for the application of engineering geology to earthquake mitigation learning.

Subsequent American earthquakes have been given similar, yet lesser overall, attention, for instance the San Fernando (California) earthquake of 9 February 1971 (magnitude 6.6) and the Loma Prieta event of 17 October 1989 (magnitude 7.1) located just south of San Francisco (California) on a San Andreas splinter fault.

Purpose

Strictly speaking, the purpose of engineering geological mapping in this context is to record the physical nature of ground-rupture earthquakes and to elucidate the related types of motion-induced damage (Table 1). Once defined, these can be applied on a worldwide basis, wherever similar conditions of tectonics and near-field geology and topography occur.

Engineering geological mapping for earthquake mitigation (Table 2) determines how engineered works and human safety can be protected by judicious design considerations, most of which are governed by geological conditions related to the site characterization.

Geological Profile (or Ground Profile)

This term geological profile has taken over in engineering geology from 'geological section' and considers only the depth of influence of engineering works (generally less than 15 m; Table 3).

Exploration Trenches and Trench Logging

During the peak of nuclear power-plant siting and construction in the 1970s considerable effort was made to confirm the absence of active faults within the footprint of the power block (i.e. the location of the reactor and the critical cooling linkage). The concept of avoiding active fault traces was promoted so that the risk of damage to the reactor and its containment as a result of rupture-type earthquake ground motion could be minimized. As a result, the technique of exploratory trenching advanced from the original method of observation from hand-dug pits (Table 4).

Site Characterization

As an integral part of site characterization in seismogenic regions, attention should be given to identifying geological conditions that may make the site susceptible to physical damage from strong motion (Table 5).

Post-Event Surveys

Certain elements of the earth media are particularly susceptible to being lifted, shifted, toppled, or cracked by earthquake strong motion. The patterns of damage reveal much about the frequency characteristics of the incident ground motion and the relative duration of the strong motion. Particularly affected are fine soils, boulders on slopes, blocks of rock defined by joints, overly steep stream and shore banks and cliffs, and hillside masses saturated with groundwater.

Geologists have but hours to locate, photograph, and map these features before they are destroyed, first by human visitors and soon after by rainfall and other natural erosive agents (Table 6).

Cultural features within the built environment (including engineered works) offer additional potential

Table 1 Engineering geological mapping of earthquake effects

Elements	*Purpose*	*Important considerations*
Stratigraphy	Identify and describe the geological formational units to be expected in design and construction	Individual engineering geological units
Groundwater regime	Define character of groundwater, as it is affected by ground motion and diminishes the shear strength of earth media to resist dynamic deformation	Perched water Vadose zone and fluctuations Peizometric surface Potentiometric surface
Rock mass characterization	Delimit observable or likely subsurface bounds of each detectable hard rock unit	Identify bodies of discontinuity bounded rock masses that may become unstable from shaking
Presence of weak rock	Basis of definition, including why the rock is determined to be weak	Recommendations as to how and why such weak rock may pose problems to design and/or construction, operation, and maintenance
Potential problems related to sedimentological, structural, or geomorphological conditions	Portions of surface or subsurface that appear to be related to mapped patterns of earthquake damage	Use special map symbols to portray these features; the Geological Society of London Engineering Geomorphological map symbols are ideal for this purpose

Table 2 Engineering geological mapping for earthquake mitigation

Elements	*Purpose*	*Important considerations*
Before earthquake		
Conceptual site geological model	Forms the basis for expectation and detection of geological features critical to the mission of defining geological controls over site specific earthquake threats	Define the presence and interrelationships between site geology and topographical configuration, to include hydrogeological conditions, as well as the character of existing or planned development
Existing evidence of presence of active or credible faults in near proximity to the project	Provide maximum means for avoidance of ground rupture potential to the project works	Regional geological maps Known active or credible faults Site explorations to preclude the presence of active or credible faults
Geomorphological evidence of past displacements or earthquake induced ground failure	Generally high potential for recurrence of failure or displacement in the future	Image interpretation of aerial photographs, especially those executed at low sun angle with shadows
After earthquake		
Ground deformational evidence	basis for reconstruction, lessons learned for basic knowledge Locate and plot evidence of ground rupture from activated faults	Map evidence immediately before it is destroyed Provides felt intensity information useful in meisoseismal mapping for quantification of attenuation
Groundwater perturbation	Establish fluctuation in groundwater surface from existing water wells and observation wells (if such are present)	Instructive as basis for pore pressure effects in deformation as well as liquefactions
Evidence of dislodgement	Helps to establish ground acceleration magnitude and vector of ground motion	Sometimes possible to locate fresh slope debris and to establish provenance of dislodgment
Displacements	Traces of offset pavement, curbs and traffic control lines Fracture pattern and offsets in sidewalks and retaining and structural walls	Establishes ground motion vectors and possible ground rupture
Distortion of infrastructure	Collapsed power lines, broken water and sewer pipes, deformed railroad rails	Facility may need significant improvement in replacement, if design or location is faulty in terms of future earthquake motion
Water displacement	Wave motion characteristics and indication of tsunami or seiche character	Evidence will persist for only a few days, even without human intervention
Survey triangulation	Establishes bounds of displaced ground	May indicate tilting of ground masses from ground shaking

Table 3 Ground profiles for earthquake mitigation

Elements	*Purpose*	*Important Considerations*
Primary profile	Placed perpendicular to known or most likely trace (strike) of active or credible fault	The profile integrates the Site Conceptual Model with actual site conditions and sets the stage for exploration trenching and logging
Display geological conditions related to expected worst case site response to and damage from the MCE	Promotes rational consideration of all geological site conditions that could conceivably affect performance of the planned works, in the context of earthquake threat	Generally means that one or more exploratory trenches will be placed and logged perpendicular to secondary geotechnical profiles
Secondary profile(s)	Examination of potential failure scenarios other than perpendicular to nearby active or credible fault	Prime example would be to construct a geotechnical profile perpendicular to the face of a slope or other face thought to be susceptible to earthquake damage

evidence to the discerning eye of the engineering geologist. Relatively tall and thin objects (such as gravestones, lamp standards, free-standing walls, plate window glass, roadway signboards, and utility poles and towers) tend to register a deformational response to ground motion. In addition to revealing much about the character of the underlying ground, such damage also gives an insight into the vector, acceleration, and duration of local ground motion.

Table 4 Exploration trenches and trench logging for earthquake mitigation

Elements	*Purpose*	*Important considerations*
Orientation of trench	Perpendicular to suspected fault trace (strike) or related geomorphological feature	Maximum exposure of displacement and age determination
Trench log	True scale representation as a hand drawn graphic image	Show running horizontal scale and repetitive vertical scale
Observation and recording	Capture dimensional relationships in true scale and position	Establish vertical and horizontal control Locate sample and photographic locations
Graphic scale	Faithful representation	Generally at 1 : 3 to 1 : 5, for detail
Sampling	Mainly for purposes of age determination of horizons and faulting displacements	Show location of samples on the graphical log
Labelling	Viewer identification of key geological features related to the nature and age determination of faulting	Mark also on the archival photograph and/or video image record of both faces of the trench
Faces	The surface portrays interrelationships of stratigraphy and relative ages of elements; normally only one of the two opposing faces is logged	Carbon 14 samples give minimum ages of offset strata; other techniques (e.g. fluid inclusion analysis) refine minimum age of displacement of trench intercepted faults

Table 5 Site characterization for earthquake mitigation

Elements	*Purpose*	*Important considerations*
Stratigraphy	Define horizons, zones, or pockets that may be improved by engineering to withstand earthquake induced failures or movement	Engineers to consider methods of improvement of the geotechnical characteristics of the critical slope failure mass
Displacement surfaces	Defines source and magnitude of ground rupture as input to design accommodation of future displacement on known faults	Generally applied only to earth fill dam embankments where it is deemed worth the effort, cost, and risk to attempt withstand design
Palaeoseismic evidence	Sand blows, as well as geomorphological evidence of ground displacements as seen in exploratory trenches, sand and gravel pits, and stream banks	Evidence mainly occurs as stratigraphical displacement or cross cutting relationships

Table 6 Post event geological mapping

Elements	*Purpose*	*Important considerations*
Ground rupture	Define nature of displacements from activation of a causative active fault	Displacement on active fault; displacement from ground heaving; tensile failures of lateral spreading
Liquefaction	Locates ground below which to excavate exploratory trenches	Need to define structural and stratigraphical locus of liquefaction
Lateral deformation	Define how cultural features have been affected	Transportation features, towers, pipelines, walls, gravestones, and many other thin and/or spindly objects that are moved in a bodily fashion
Ground failure	Delimit the failed ground then excavate, observe, sample, and define causes of damage	Forms more bases for lessons learned and then applied elsewhere

Summary

The engineering geologist's expertise in characterizing ground for all manner of engineered works can be extended to seismogenic zones. Engineering geologists can detect and describe geological, hydrogeological, and topographical conditions, defining the potential hazard at sites that are at risk of damage from earthquakes.

The description of fresh earthquake damage improves the ground model, helping to mitigate the adverse effects of future earthquakes. These analyses can be applied more widely to assist civil

engineers in the development of effective measures to withstand future events.

See Also

Engineering Geology: Seismology; Natural and Anthropogenic Geohazards; Liquefaction. **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes. **Seismic Surveys**. **Tectonics:** Earthquakes.

Further Reading

Bonilla MG (1967) Historic surface faulting in continental United States and adjacent parts of Mexico. Unnumbered open file report to the US Atomic Energy Commission. US Geological Survey. Menlo Park, California.

Bonilla MG (1973) Trench exposures across surface fault ruptures associated with San Fernando Earthquake. In: Murphy ML (ed.) *San Fernando, California Earthquake of February 9, 1971*, pp. 173 182. Washington DC: US Department of Commerce.

Bonilla MG, Mark RK, and Lienkaemper JJ (1984) Statis tical relations among earthquake magnitude, surface rupture length, and surface fault displacement. *Bulletin of the Seismological Society of America* 74: 2379 2412.

Brown RD Jr, Ward PL, and Plafker G (1973) *Geologic and Seismologic Aspects of the Managua, Nicaragua, Earthquakes of December 23, 1972*. USGS Professional Paper 838. Washington DC: US Geological Survey.

Eckel EB (1964) *The Alaska Earthquake, March 27, 1964: Lessons and Conclusions*. USGS Professional Paper 546. Washington DC: US Geological Survey.

Hays WW (1981) *Facing Geologic and Hydrologic Hazards: Earth Science Considerations*. USGS Profes sional Paper 1240 B. Washington DC: US Geological Survey.

Krinitzsky EL (2003) How to combine deterministic and probabilistic methods for assessing earthquake hazards. *Engineering Geology* 70: 157 163.

Obermeier SF, Jacobson RB, Smoot JP, *et al.* (1990) *Earth quake Induced Liquefaction Features in the Coastal Set ting of South Carolina and in the Fluvial Setting of the New Madrid Seismic Zones*. USGS Professional Paper 1504. Washington DC: US Geological Survey.

Scott GR (1970) *Quaternary Faulting and Potential Earth quakes in East Central Colorado*. USGS Professional Paper 700 C. Washington DC: US Geological Survey.

Youd TL and Hoose SN (1978) *Historic Ground Failures in Northern California Associated with Earthquakes*. USGS Professional Paper 993. Washington DC: US Geological Survey.

Wells DL and Coppersmith KJ (1994) New empirical rela tionships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America* 84: 974 1002.

Geological Maps

J S Griffiths, University of Plymouth, Plymouth, UK

Introduction

Engineering geological maps provide ground information of relevance to civil engineering planning, design, and construction. Given that geology and civil engineering deal with three-dimensional structures and how they behave through time, albeit usually on different time-scales (geology 10^1–10^9 years; civil engineering 10^1–10^2 years), it is to be expected that the preparation of maps and plans is a essential component of both disciplines. Hence engineering geology maps have been regarded as an effective means of conveying information between geologists and engineers from the earliest days of the emergence of engineering geology as an identifiable subject. Indeed, the first classic stratigraphic maps and sections of William Smith (*see* **Famous Geologists:** Smith) in the late eighteenth and early nineteenth centuries arose out of his work on canal construction and the need to anticipate ground conditions.

An engineering geological map differs from the standard geological map by not being limited to scientific categories of bedrock and/or superficial geology. The map may group materials according to similar engineering characteristics and also present relevant data on topography, hydrology, hydrogeology, geomorphology (*see* **Engineering Geology:** Geomorphology), and geotechnics, plus information on man-made structures such as landfill sites or earthworks. This additional ground information may be shown on the map itself, or by using tables, charts, and diagrams accompanying the map, as exemplified by Table 1 from the Hong Kong Geotechnical Area Studies Programme (GASP), and Figure 1 from the Vía Inter-Oceánica, Quito, Ecuador. Whilst most maps in the past were cartographically drafted and produced as hardcopy, there is increasing use of computer-based Geographical and Geoscience Information Systems (GIS and GSIS) for compiling the data in formats that can be readily updated and analyzed.

The data for an engineering geological map will be compiled from a wide range of data sources and will

Table 1 Example of an extended engineering geology legend based on bedrock data from the Hong Kong GASP programme

Material description				*Evaluation of material*		
Map unit	*Lithology*	*Topography*	*Weathering*	*Material properties*	*Engineering comment*	*Uses/excavation*
Lower Cretaceous Dolerite (dyke rock)	Black to very dark grey, fine to medium grained rock. Smooth joints normal to boundaries result of cooling	Generally occurs as linear structural features transecting the volcanic and granite units. May be of slightly depressed or elevated topographic form due to variable resistance of the country rocks. This geological structure often controls local surface runoff and may act as a loci for subsurface water concentrations	Weathers deeply to a dark red silty clay	Weathered mantle will contain a high proportion of clay and iron oxides leading to low Ø values. Intact rock strength will be very high, >100 Mpa when fresh	Restricted extent precludes detailed comment. Weathered mantle will have relative low permeability and will affect groundwater hydrology by forming barriers, and variable boundary conditions. Sub vertical dykes may dam groundwater leading to unexpectedly high groundwater levels.	Restricted extent precludes deliberate borrow or quarry activities. Weathered material would make poor fill but fresh rock would make suitable high density aggregate or railway ballast.
Upper Jurassic Hong Kong Granite	Pink to grey medium grained equigranular, non porphyritic rock Mineral include quartz, potassium feldspar, plagioclase, biotite and muscovite. Rough sheeting joints and widely spaced tectonic joints widespread	Forms extensive areas of moderate to steep convexo concave slopes. High level infilled valeys are common. Drainage pattern is often dendritic in nature and is commonly dislocated by major tectonic discontinuities. These units are characterised by moderate to severe gully and sheet erosion associated with hillcrest and upper sidelong terrain	Shallow to deep residual soils over weathered granite. Local development of less weathered outcrops in stream beds and occasional cliff faces. Residual core boulders common on surface of sidelong ground and gullies. Weathering depths >20 m	Material properties vary with depth within the weathering profile. For completely weathered granite typical values are $c' \approx 0$ 25 kPa, $Ø \approx 31$ 43°; permeablity $\approx 10^{-6}$ 10^{-8} m/s; dry density 1500 kg m^3. Moisture content 15% near surface, 30% at depth. Fresh rock UCS 80 150 MPa. Rock mass strength dependent on joint characteristics. Roughness angles for tectonic joints 5 10°; for sheet joints 10 15°; basic friction angle $\approx 39°$	Weathered mantle subject to sheet and gully erosion with landslides in steep slopes or if severely undercut. Perched water tables conform with highly permeable upper weathered zones. Rock is prone to discontinuity controlled failures in fresh to moderately weathered state. Stream and drainage lines align with geological weakness. Large structures may require deep foundations. Cut slope design may be governed by the large depths of weathered material	Extensively quarried and used as concrete aggregate. Weathered material widely used as fill as it is easily excavated by machinery. Core boulders can cause problems during excavation

Middle and Lower Jurassic Lok Ma Chau Formation	Metamorphosed sedimentary and volcanic rocks, including schist, phyllite, quarzite, metasediments and marble.	Forms hills of moderate to low relief due to its low resistance to erosion. Occurs extensively beneath collluvial and alluvial cover. Local areas of surface boulders and occasional rock outcrops on sidelong ground and in gullies	Metasediments generally weather to produce moderately deep (1 2 m), uniform or gradational red brown clay	Near surface completely weathered residual soil acts as a silt with a void ratio 0.25 0.33. Gradings show 5 15% clay, 40 60% silt, 20 30% fine sand. PL 25 35%; LL 34 40%. Typical shear strength $c' \approx 0$ 15 kPa, $\emptyset \approx 35^\circ$. Weathered materials dry density 1600 1800 kg m^3. Fresh rock UCS 40 90 MPa. Discontinuity strength parameters $c' \approx 0$ 5 kPa, $\emptyset \approx 25$ 30°	Considerable care is required during investigation, design, and construction. Bearing capacity reasonable for low and moderate loads. Stability is dependent on the very closely spaced discontinuities. Discontinuity surveys are essential for cut slope design. Material prone to failure along discontinuities when weathered and saturated	Material can be used as a source of bulk fill but may break down to silt if over compacted. Excavation by machine is relatively easy

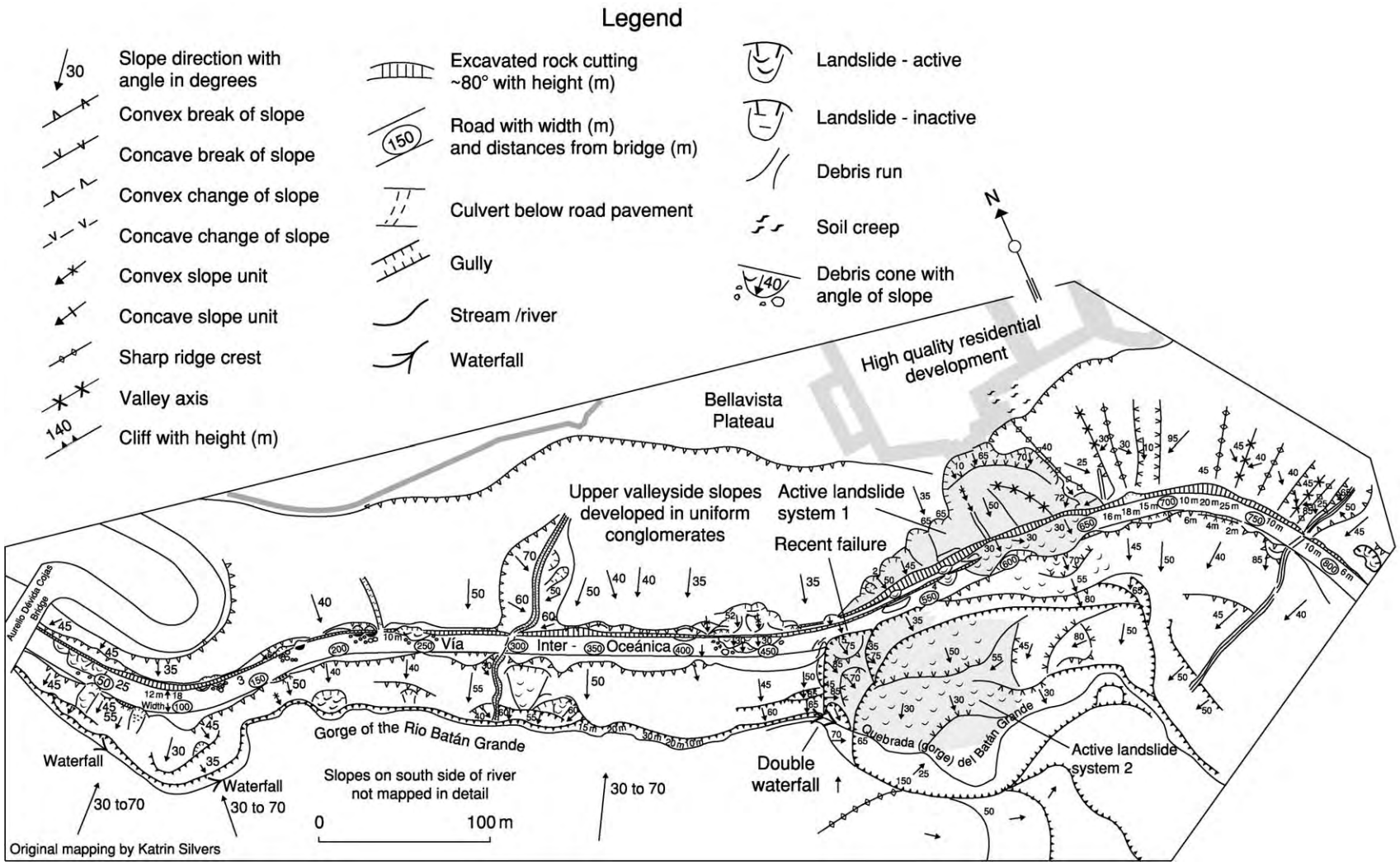


Figure 1 Large scale engineering geology map of a section of the Vía Inter Oceánica, Quito, Ecuador.

Table 2 Scale of Engineering Geology Maps suitable for different purposes

Large scale
- 1 : 10 000 or greater. Plans and sections up to 1 : 500 scale can be used in particular circumstances
- Based on detailed field mapping and ground investigation data, supplemented by interpretation of large scale aerial photographs or high resolution airborne and satellite imagery
- Uses: designing and interpreting ground investigations; background data for detailed design of foundations and structures; claims and litigation
- End users: civil engineers; engineering geologists; geotechnical engineers; quantity surveyors

Medium scale
- 1:10 000 to 1:100 000 scale
- Based on terrain evaluation techniques, remote sensing interpretation (small scale aerial photographs and satellite imagery) and ground truth mapping of selected areas
- Uses: geohazard evaluations; locating construction sites and route alignment planning; resource evaluation; hydrogeological studies; development planning at local and regional level; disaster relief planning
- End users: civil engineers; engineering geologists; local and regional planners; local and regional government; insurers; water authorities; emergency services; military

Small scale
- 1:100 000 or smaller
- Based on terrain evaluation techniques, remote sensing interpretation (primarily satellite imagery) and limited reconnaissance of field area
- Uses: generalized geohazard evaluation; general development planning at the regional or national level
- End users: local and regional planners; regional and central government; insurers; military

not be collected solely through field mapping. However, as with any map, the value of the engineering geological map is dependent on the accuracy of the information used in its compilation. Thus, engineering geologists responsible for producing engineering geology maps must have a broad range of skills and be able to recognize and correctly compile data from a wide spectrum of Earth science disciplines.

Engineering geological maps are not just academic interpretations of the ground; they are produced to meet the specific requirements of a project. Therefore, the material presented and the scale used on the map will vary to meet the requirements of the end-user. Consequently, there is no unique format or content for an engineering geology map and throughout the mapping programme, the engineering geologist will need to be aware of the end-use to ensure that collection and presentation of the data are appropriate to the project requirements. Table 2 provides an indication of the application of engineering geological maps at different scales to suit different purposes.

The Type of Data to be Recorded

Although the specific content of any one engineering geological map will depend on the application, the main aim of an engineering geological mapping programme is to produce a map on which the mapped units are defined by engineering properties or behaviour. The limits of the units are determined by changes in the physical and mechanical properties of the materials. The boundaries of the mapped units may not correlate or coincide with the underlying geological structure or the chrono-stratigraphic units as depicted on conventional geological maps. However, experience has shown that the lithology of engineering soils and rocks can often be effective in defining the engineering geological map units. Apart from these map units, Table 3 presents the type of additional data that are relevant and could be recorded on the engineering geological map through observations made in the field, supplemented by desk studies and ground investigations using exploratory pits and geophysics.

Whilst Table 3 provides an indication of the range of data that might be compiled, the requirements for an individual engineering geological map will be tailored to suit the specific issues to be investigated. For example, in an area of earthquake risk there is likely to be more emphasis on the location of active faults, extent of soils liable to liquefy under dynamic loading, unconsolidated deposits that can amplify ground shaking, and zones potentially liable to the affects of tsunami or seiche. Similar specific details will be appropriate for different types of geohazard evaluation (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards). For example, Figure 1 is an example of a large-scale engineering geological map/plan of a road (Vía Inter-Oceánica) east of Quito, Ecuador, which has been affected by landsliding. Because the concern of the engineering geologist in this study is slope instability, the map emphasises the landslides and man-made features but contains relatively little geological detail.

Some of the engineering geological data may be used to identify 'zones'. These areas on the map

Table 3 Data to be recorded on an Engineering Geology Map

Geological data
- Map units (chrono and/or lithostratigraphy)
- Geological boundaries (with accuracy indicated)
- Description of soils and rocks (using standard engineering codes of practice)
- Description of exposures (cross referenced to field notebooks)
- Description of state of weathering and alteration (note depth and degree of weathering)
- Description of discontinuties (as much detail as possible on the nature, frequency, inclination and orientation of all joints, bedding, cleavage, etc.)
- Structural geological data (folding, faulting, etc.)
- Tectonic activity (notably neotectonics, including rates of uplift)

Engineering geology data
- Engineering soil and rock units (based on their engineering geological properties)
- Subsurface conditions (provision of subsurface information if possible, e.g., rockhead isopachytes)
- Geotechnical data of the engineering soil and rock units
- Location of previous site investigations (i.e., the sites of boreholes, trial pits, and geophysical surveys)
- Location of mines and quarries, including whether active or abandoned, dates of working, materials extracted, and whether or not mine plans are available
- Contaminated ground (waste tips, landfill sites, old industrial sites)
- Man made features, such as earthworks (with measurements of design slope angles, drainage provision, etc.), bridges and culverts (including data on waterway areas), tunnels and dams

Hydrological and hydrogeological data
- Availability of Information (reference to existing maps, well logs, abstraction data)
- General hydrogeological conditions (notes on: groundwater flow lines; piezometric conditions; water quality; artesian conditions; potability)
- Hydrogeological properties of rocks and soils (aquifers, aquicludes, and aquitards; permeabilities; perched watertables);
- Springs and Seepages (flows to be quantified wherever possible)
- Streams, rivers, lakes, and estuaries (with data on flows, stage heights, and tidal limits)
- Man made features (canals, leats, drainage ditches, reservoirs)

Geomorphological data
- General geomorphological features (ground morphology; landforms; processes; Quaternary deposits)
- Ground Movement Features (e.g., landslides; subsidence; solifluction lobes; cambering)

Geohazards
- Mass movement (extent and nature of landslides, type and frequency of landsliding, possible estimates of runout hazard, snow avalanche tracks)
- Swelling and shrinking, or collapsible, soils (soil properties)
- Areas of natural and man made subsidence (karst, areas of mining, over extraction of groundwater)
- Flooding (areas at risk, flood magnitude and frequency, coastal or river flooding)
- Coastal erosion (cliff form, rate of coastal retreat, coastal processes, types of coastal protection)
- Seismicity (seismic hazard assessment)
- Vulcanicity (volcanic hazard assessment)

have approximately homogenous engineering geological conditions, usually defined by more than just the lithostratigraphy. The zoning system would be derived from the factual data contained on the base map. It would not, therefore, normally form part of the original mapping programme and represents a derivative or interpretative engineering geological map. Zoning maps can be particularly effective in geohazard studies where the magnitude of a particular hazard can be represented by an interpretative map containing data on probability of occurrence or frequency. However, interpretative zoning maps can be used in many ways relevant to engineering and examples can be found of maps showing: foundation conditions, excavatability of materials, sources of construction material, bearing capacity of soils and rocks, and general constraints to development.

Map Scale

For engineering work, mapping may be carried out during feasibility studies, before any site work is started, just prior to or during the site investigation phase (*see* **Engineering Geology:** Site and Ground Investigation), whilst construction is underway, or as a means of compiling data if there have been problems with a structure after it has been built. To meet this range of applications, three broad-scale categories are recognized (large, medium, and small). These categories are presented in Table 2, and whilst the boundaries to the categories should be regarded as flexible, there are differences in the way the maps are compiled, the techniques employed in their construction, and their intended end-use. A wide range of examples of all the categories of

engineering geology map can be found in the published literature, and Figure 1 is presented as an illustration of a typical large-scale engineering geological map. This map was produced specifically for the investigation of a single stretch of road and provides data of direct relevance to the design of remedial measures.

Medium-scale engineering geological maps are probably the most widely used and excellent examples can be found in the UK, where a national programme of applied geological mapping, predominantly by the British Geological Survey, has been underway for over 20 years. This programme has resulted in over 35 studies and produced a wide range of maps for use in engineering construction and planning development (Table 4). These maps have mainly been produced at a scale of 1 : 25 000 and represent compilations of engineering geological data in map form that can used as the basis for engineering feasibility studies. Similar applied geological mapping programmes were carried out in a number of other countries. In France, the ZERMOS programme has produced a number of 1 : 25 000 scale maps of selected area. Originally under the auspices of the Geotechnical Control Office, the former territory of Hong Kong was mapped at 1 : 20 000 during the GASP programme. The USA also has a long commitment to engineering geological mapping at a wide range of scales which is illustrated in the compilation of maps provided by the US Geological Survey Profession Report 950, entitled 'Nature to be Commanded.' However, despite their widespread production, the actual use of engineering geological maps in planning and development is variable.

Small-scale maps are best exemplified by the PUCE (Pattern-Unit-Component-Evaluation) system developed by the Commonwealth Scientific Research Organisation (CSIRO) in Australia, but also used in Papua New Guinea. This programme produced maps at a scale of 1 : 250 000 that define broad terrain patterns, within which particular assemblage of landforms, pedological soils, and vegetation sequences occur. These have been predominantly used in regional planning but there are examples of their use for aggregate resource surveys; route corridor alignment; water resources; military and off-road mobility; flood hazard; land capability assessment; and aesthetic landscape appreciation.

Table 4 Applied Geological Maps compiled in the United Kingdom 1983 96 utilizing Geological and Engineering Geological Data

Data points
- Location of exploratory holes and wells
- Distribution of geotechnical data test results
- Point rockhead information

Disturbed ground (human activity)
- Distribution (general)
- Distribution of mines and mine workings (all types, including surface and sub surface)
- Distribution of made ground

Superficial geology
- Soil types, extent, lithology, and thickness
- Drift thickness/rockhead contours
- Geotechnical properties of soils

Bedrock geology
- Rock types, extent, lithology, lithostratigraphy
- Structure contours
- Geological structure
- Geotechnical properties of rocks

Engineering geology
- Foundation conditions
- Hydrogeological conditions
- Ground conditions in relation to groundwater
- Nature and distribution of geohazards (subsidence, instability, flooding, earthquakes, etc.)
- Engineering geological zones (i.e., areas of homogenous engineering geological conditions)
- Aggregate and borrow material sources

Geomorphology
- Geomorphological landforms and process
- Drainage
- Areas of slope instability
- Flood frequency limits

Derived construction constraints maps
- Slope steepness
- Ground instability (e.g., subsidence, cambering, landslides, soft ground, etc.)
- Landslide hazard and risk maps
- Previous industrial usage (brownfield sites and contaminated land)

Derived resources maps
- Nature, extent, and properties of mineral resources (superficial and bedrock)
- Groundwater resources
- Distribution of aggregates
- Sites of Special Scientific Interest (SSSI)

Summary maps
- Development potential
- Summary of construction constraints
- Statutory protected land

Data Collection

Primary mapping for engineering geology follows the same basic rules and uses the same techniques established for conventional geological mapping. However, a number of additional decisions need to be made when undertaking engineering geological mapping. These are to identify the types of data that are to be collected to meet the survey requirements (Table 3); the scale of the mapping (Table 2); the methods to be used for data collection; and the intended final map products (Table 4).

Table 5 Example of an extended engineering geology legend based on the superficial geology map data in the Applied Geology Map of Stoke On Trent

			Planning & Engineering Considerations			
	Description	*Characteristics*	*Slope stability*	*Excavation*	*Foundations*	*Engineered fill*
Alluvium	Silts & clays 0.6 9 m thick. Occurs mainly in the Trent valley and tributaries	Very soft to firm, low to high plasticity, medium to high compressibility. May be desiccated near top	not applicable as occurs in flat areas	Diggable by excavator. Heave may occur at base of excavations. Trench support required	Low acceptable bearing capacity (<75 kPa)	Generally unsuitable
	Organic in places, with peat lenses	Very soft to soft, intermediate to extremely high plasticity. Highly compressible			Sulphate protection usually required for concrete	
	Sands and gravels often occur at base	Loose to very dense. Water bearing		Running ground conditions will require cut offs or dewatering		
Periglacial head	Variable soils derived from bedrock other superficial deposits. Composition varies according to the parent material. Generally consists of sandy, silty clays with gravel and cobbles. Forms a thin veneer on slopes and may thicken downslope. Perched water tables may occur within coarser horizons	Variable. Usually cohesive, soft to stiff, with low to high plasticity. Compressibility usually intermediate, but may be high. Pre existing shear surfaces may be present, with low residual friction angles	Natural slopes often marginally stable	Diggable by excavator	Consolidation settlement usually small. Differential settlement likely where soft compressible zones present	Generally unsuitable
			The variability of this deposit necessitates careful site investigation, preferably by trial pitting, to determine the site characteristics in relation to any proposed development. Clayey slopes may require drainage and other stabilisation works prior to construction			
Glacial sand and gravel	Coarse sand and subangular to subrounded gravel Occasional subrounded cobbles	Loose to dense granular deposit. Water bearing	not applicable as occurs in flat areas	Diggable by excavator Support and groundwater control required	Consolidation settlements small. Pile driving may be difficult in cobbles. Sulphate protection may be required for concrete	Suitable for use in embankments if the soft clay zones are removed
	Some horizons of laminated clay/ silt occur	Clay/silt horizons are usually soft to stiff, of low to intermediate plasticity				

Glacial till (boulder clay)	Variable deposit, generally sandy, silty clays with gravel, cobbles and occasional boulders Water bearing lenses of sand and gravel may occur	Generally firm to stiff, with low to intermediate plasticity and intermediate compressibility	Cut slope of 1V : 2.5H generally adequate for long term stability	May be difficult to dig and can require ripping. Excavations generally stable in the short term but deteriorate on exposure and wetting. Support required for deep excavations, and where sand lenses occur	Usually forms a good founding medium with acceptable bearing capacities typically 150 to 600+ kPa)	Suitable if placed in dry weather when moisture content is low
Landslide debris	In the field area all known occurrences occur in weathered mudstones of the Etruria Formation	Deposits contain shear surfaces with low residual strengths. Remoulded clay debris is generally poorly drained with possible perched water tables	Areas of landslide debris should be avoided if possible. Constructional activity is likely to reactivate slope movement unless appropriate remedial measures are taken Detailed site investigation is essential with extensive use of exploratory holes and geophysics. If construction is unavoidable groundwater and ground movement monitoring is essential			Unsuitable

In most engineering situations there will be four phases in the preparation of an engineering geological map: desk study; field mapping (*see* **Geological Field Mapping**); interpretation; and reporting. During the desk-study phase all existing data are compiled, remote sensing interpretation is carried out, a preliminary field reconnaissance may be undertaken, and the field programme is planned. Field mapping requires the collection of primary data in the field. Even if the available data is quite comprehensive and it is only intended to produce small-scale maps, some primary field mapping will be necessary. Interpretation of the data involves bringing together the field and desk study data and preparing the suite of maps that meet the project requirements. Finally, the maps will need to be supplemented by a written report for the end-user that expands on the details shown on the map and, in engineering situations, may provide some design guidance or recommendations.

Map Presentation

The presentation of engineering geological maps follows normal cartographic rules over scale, north arrow, and locational data, but the information displayed will be based on end-user requirements. Because the information on the map is variable, it is usually necessary to create a bespoke legend for the map, as exemplified by the key to Figure 1. However, general guidance on the typical symbols to use can be found in the standard literature on engineering geological maps listed below.

Often, with engineering geological maps, it is necessary to include quite comprehensive data in the map legend. These data will not have just been compiled from field observations but will include data from the desk studies and any detailed ground investigations carried out in the area. An example of a comprehensive, or extended, legend is provided in Table 5, based on the UK Applied Geological Map for Stoke-on-Trent. This uses the superficial geological map as the basis for identifying the engineering geological units. Table 1 provides an example of a similar compilation of data based on bedrock properties of the type used in the tropical weathering environment investigated during the Hong Kong GASP programme. In both examples, additional data on the geotechnical and engineering characteristics of the various materials are included in the tables as well as comments on

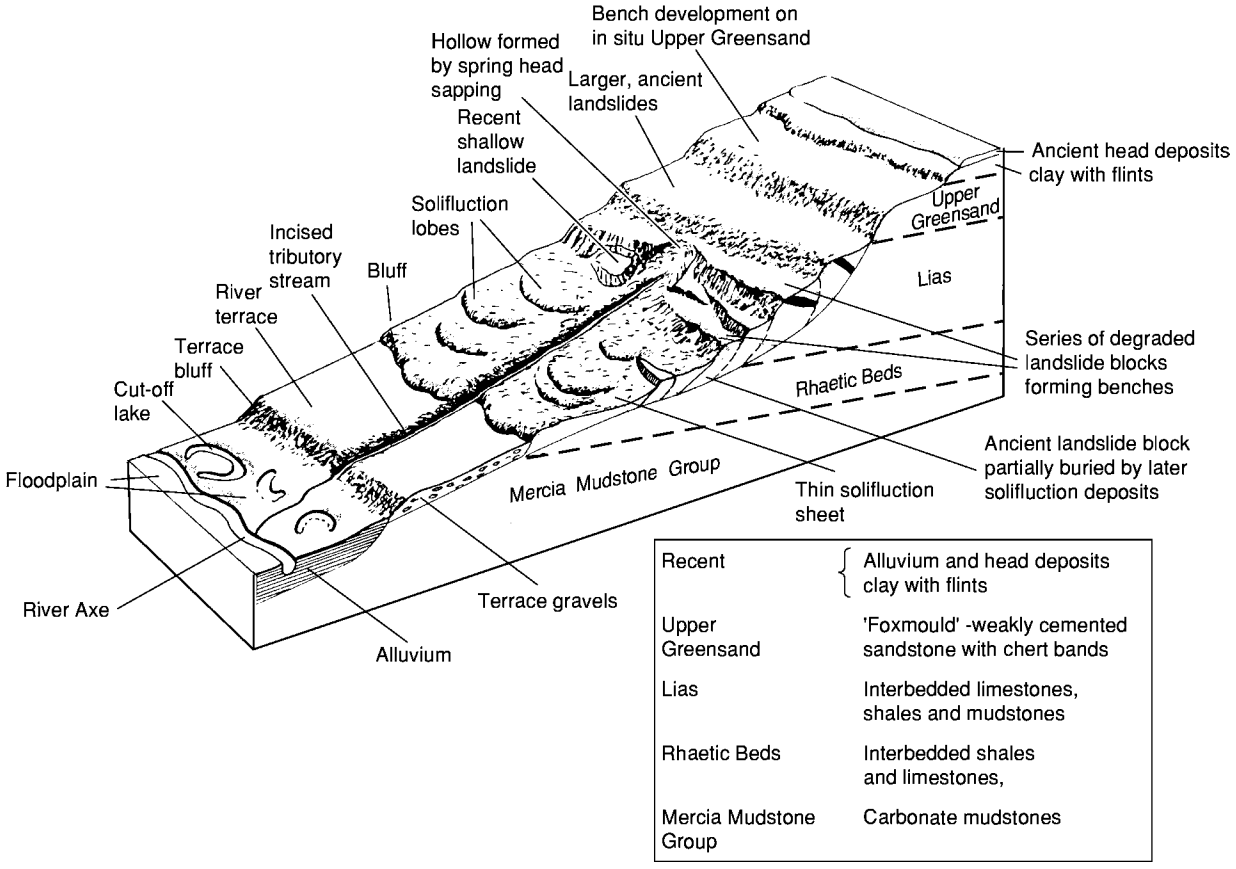


Figure 2 Three dimensional engineering geology ground model developed for the Axminster By pass, Devon, England. Reproduced with permission from Croot D and Griffiths JS (2001). Engineering geological significance of relict periglacial activity in South and East Devon. *Quarterly Journal of Engineering Geology and Hydrogeology* 34: 269 282.

engineering issues, such as slope stability, excavatability, and groundwater conditions. These data may be shown in summary on the actual map or in an accompanying report.

Another method of conveying engineering geological data that can be included on the map or in an accompanying report is shown in Figure 2. This three-dimensional block diagram, compiled during mapping for the Axminster by-pass in Devon, provides a synopsis of the ground conditions that illustrates how the bedrock geology, superficial geology, and geomorphology create a landscape that contains a number of technical problems for the engineering geologist. This type of figure is generally referred to as a 'ground model'.

These various techniques of data presentation are very effective in conveying detailed information in a way that can be readily understood by the non-specialist. However, there is always a concern that a report can become separated from its maps and, as a general recommendation, the map itself should be able to stand alone and be understood by all potential users, without having to refer to a separate report.

Integration with Site Investigation

Site investigation for engineering is the process by which data appropriate for the design and construction of structures is collected. Whilst this primarily involves the exploration of the ground using invasive techniques such as drilling and trial pitting, it is recommended that engineering geological mapping be integral to the process. Mapping has proved itself to be extremely cost effective and can be used to design a more efficient ground investigation by defining the engineering geological units that will represented by the exploratory holes. This is illustrated by the mapping carried out for the UK portal and terminal areas of the channel tunnel. The UK channel tunnel portal is located in a late-glacial multiple-rotational landslide that was subject to detailed mapping in order to plan the ground investigations. The mapping provided the basis for the development of a ground model against which additional data were checked as they were acquired. This demonstrates how mapping can ensure that any geohazards that might affect a project will be identified early on and thereby allowed for in the design.

See Also

Engineering Geology: Geomorphology; Natural and Anthropogenic Geohazards; Site and Ground Investigation. **Famous Geologists:** Smith. **Geological Field Mapping**. **Geological Maps and Their Interpretation**.

Further Reading

Anon (1976) *Engineering Geology Maps: A Guide to their Preparation*. Paris: The UNESCO Press.

Barnes J and Lisle RJ (2004) *Basic Geological Mapping*, 4th edn. Chichester: Wiley.

Brunsden D (2002) Geomorphological roulette for engin eers and planners: some insights into an old game. *Quar terly Journal of Engineering Geology and Hydrogeology* 35: 101 142.

Culshaw MG, Bell FG, Cripps JC, and O'Hara M (eds.) (1987) *Planning and Engineering Geology*. Geological Society Engineering Special Publication No. 4.

Dearman WR (1991) *Engineering Geological Mapping*. Oxford: Butterworth Heinemann.

Doornkamp JC, Brunsden D, Cooke RU, Jones DKC, and Griffiths JS (1987) Environmental geology mapping: an international review. In: Culshaw MG, Bell FG, Cripps JC, and O'Hara M (eds.) *Planning and Engineering Geol ogy*. Geological Society Engineering Special Publication, No. 4, 215 219.

Eddleston M, Walthall S, Cripps JC, and Culshaw MG (1995) *Engineering Geology of Construction*. Geological Society Engineering Special Publication No. 10.

Finlayson AA (1984) Land surface evaluation for engineer ing practice: applications of the Australian PUCE system for terrain analysis. *Quarterly Journal of Engineering Geology* 17: 149 158.

Fookes PG (1997) Geology for engineers: the geological model, prediction and performance. *Quarterly Journal of Engineering Geology* 30: 293 424.

Fookes PG, Baynes FJ, and Hutchinson JN (2000) Total geological history: a model approach to the anticipation, observation and understanding of site conditions. *GeoEng 2000, an International Conference on Geo technical & Geological Engineering,* Melbourne, **1**: 370 460.

Griffiths JS (ed.) (2001) *Land Surface Evaluation for En gineering Practice*. Geological Society Engineering Geol ogy Special Publication No. 18.

Griffiths JS (2002) *Mapping in Engineering Geology*. The Geological Society, Key Issues in Earth Scienes, 1.

Griffiths JS, Brunsden D, Lee EM, and Jones DKC (1995) Geomorphological investigations for the Channel Tunnel terminal and portal. *The Geographical Journal* 161(3): 275 284.

Hutchinson JN (2001) Reading the ground: morphology and geology in site appraisal. *Quarterly Journal of Engineering Geology and Hydrogeology* 34: 7 50.

Kiersch GA (ed.) (1991) *The Heritage of Engineering Geol ogy; the First Hundred Years*. Geological Society of America Centennial Special Volume 3.

Lawrence CJ, Byard RJ, and Beaven PJ (1993) *Terrain Evaluation*. Transportation Research Laboratory Report **SR 378,** TRRL, Crowthorne.

Maund JG and Eddleston M (1998) *Geohazards in Engin eering Geology*. Geological Society Engineering Geology Special Publication No. 15.

Porcher M and Guillope P (1979) Cartography des risques ZERMOS appliqués a des plans d'occupation des sols en

Normandie. *Bulletin Mason Laboratoire des Ponts et Chaussees* 99.

Robinson GD and Speiker AM (eds.) (1978) *Nature to be Commanded*. US Geological Survey, Professional Paper 950.

Rosenbaum MS and Turner AK (eds.) (2003) *Characterisation of Shallow Subsurface: Implications for Urban Infrastructure and Environmental Assessment*. Dusseldorf: Springer Verlag.

Smith A and Ellison RA (1999) Applied geological maps for planning and development: a review of examples from England and Wales, 1983 to 1996. *Quarterly Journal of Engineering Geology* 32: S1 S44.

Styles KA and Hansen A (1989) *Geotechnical Area Studies Area Programme: Territory of Hong Kong*. GASP Report XII, Geotechnical Control Office, Civil Engineering Services Department.

Geomorphology

E M Lee, York, UK
J S Griffiths, University of Plymouth, Plymouth, UK
P G Fookes, Winchester, UK

Introduction

Geomorphology is the study of landforms and landform change. Engineering geomorphology is concerned with evaluation of the implications of landform change for society. The focus of the engineering geomorphologist is primarily on the risks from current surface processes (i.e., the impact of geohazards), the characteristics of near-surface materials (i.e., the products of processes and changes, including landslide debris, river terrace gravels, duricrusts, and metastable soils), and the effects of development on the environment, notably on surface processes and any resulting changes to landforms or increased level of risk (i.e., environmental impacts).

Engineering geomorphology has developed in the past few decades to support a number of distinct areas of engineering, including river, coastal, geotechnical, and agricultural engineering. River engineering involves studies of the nature and causes of alluvial river channel change (e.g., channel migration, bank erosion, and bed scour); coastal engineering studies emphasize the understanding of the occurrence and significance of shoreline changes, especially in response to changes in sea-level and sediment supply. Geotechnical engineering has complemented engineering geology and has been proved to be valuable for rapid site reconnaissance. Geomorphology provides a spatial context for developing site models and for explaining the distribution and characteristics of particular ground-related problems (e.g. landslides, permafrost, or the presence of aggressive soils) and resources (e.g., sand and gravel deposits). In agricultural engineering geomorphology has contributed notably to the investigation and management of soil erosion problems. To a large degree, each of these applications of geomorphology has developed separately in response to specific engineering needs. In recent years, however, all have started to become integrated as a coherent discipline.

A Framework for Evaluating Change: Physical Systems

Earth's surface is dynamic, and landforms change through time in response to weathering and surface processes (e.g., erosion, mass movement, and deposition). Most of the changes occur in response to variations in the energy inputs into physical systems, including variations in rainfall intensity or total, in temperature, in river flows (discharge and sediment load), and in wave/tidal energy arriving at the coast, over a range of time-scales. Physical systems are a means of describing the interrelationships between different landforms. They form a useful spatial framework for evaluating how hazards and risks to a particular site can arise as a result of processes operating elsewhere. For example, changes in land use in the catchment headwaters can lead to changes in flood frequency and river channel change elsewhere. In addition, systems can be used to evaluate the potential impacts of a project on landforms at sites distant from the project; for example, reclamation of an intertidal wetland can have significant effects on the whole estuary, through the resulting changes to the tidal prism and mean water depth. Systems can be defined at a range of scales, from river drainage basins (watersheds or catchments) and coastal cells (sediment transport cells) to individual hillslopes, dunes, or cliffs. Irrespective of the scale, each system comprises an assemblage of individual components (i.e., the landforms) and transfers of energy and sediment (**Figure 1**).

Engineering geomorphology is directed towards understanding the way systems respond to relatively short- to medium-term changes in energy inputs

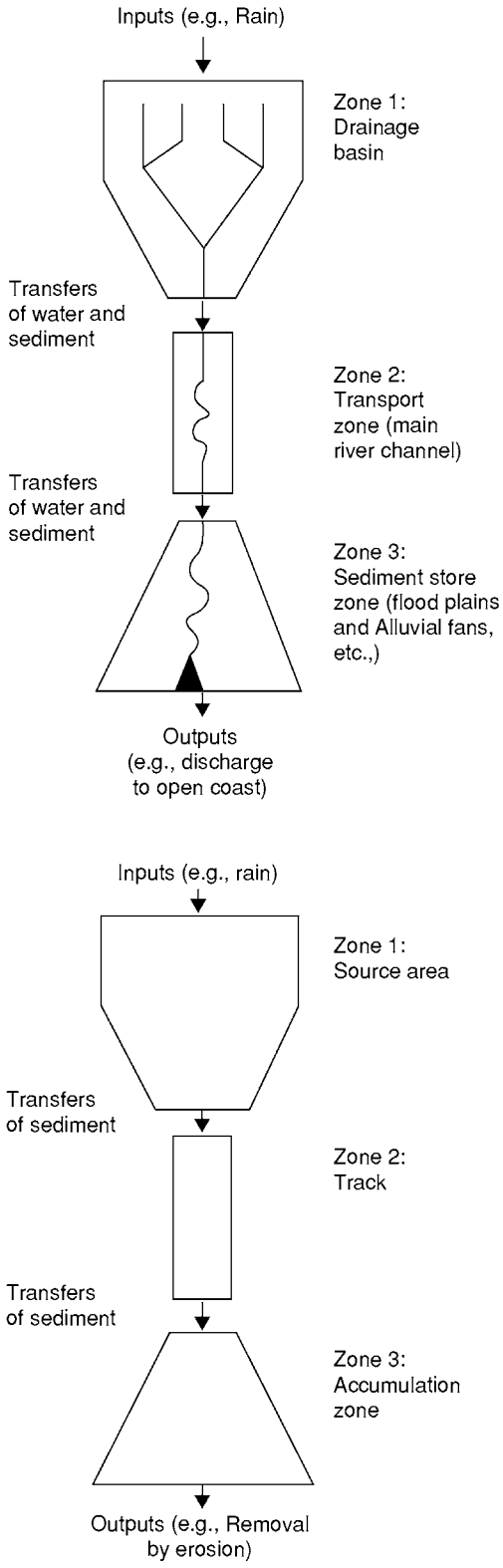


Figure 1 Examples of physical systems. (Top) A river catch ment; (bottom) a mudslide. Each system comprises an assem blage of landforms and transfers of energy and sediment. Reproduced from Fookes PG, Lee EM, and Griffiths JS (2004) *Foundations of Engineering Geomorphology*. Latheronwheel, Caithness: Whittles Publishing.

(e.g., resulting from climatic variability, changes in sediment supply, sea-level rise, or the effects of humans), rather than to long-term landscape denudation and evolution. However, there is also a need to be aware of the significance of longer term trends (e.g., the Holocene decline in sediment availability experienced on many temperate coastlines) and the presence of potential geohazards inherited from the past (e.g., ancient landslides, periglacial solifluction sheets, karst features). Engineering project cycles are generally in the order of 10–100 years (occasionally longer), and this relatively limited duration of 'engineering time' imposes a constraint on the types of landscape changes that are relevant to engineering geomorphology. Abrupt and dramatic landscape changes are likely to be significant over a time-scale of 10 to ~100 years. Relevant examples include establishment of gully systems, migration of sand dunes, river planform changes, coastal cliff recession, and the growth and breakdown of shingle barriers. High-probability to relatively high-probability events are important features and include wind-blown sand, soil erosion, shallow hillside failures, flooding, scour and river bank erosion, and coastal erosion and deposition. Low-probability events that could have a major impact on an engineering project or development are often a key issue and include flash floods, major first-time landslides, and tsunamis.

Investigation Methods

Engineering geomorphology provides practical support for engineering decision making (i.e., project planning, design, and construction). Engineering geomorphologists need to work as part of an integrated team and must provide information at many levels, ranging from crude prefeasibility-stage qualitative approximations to sophisticated quantitative analyses in support of detailed design and construction. The level of precision and sophistication required needs to be sufficient for a particular problem or context, so as to allow adequately informed decisions to be made.

An engineering geomorphologist is typically required to ask questions that address a number of situations:

- Will the project be at risk from instability, erosion, or deposition processes over its design lifetime (e.g., is the development set back sufficiently from a retreating cliff top)?
- How could problems arise (e.g., will removal of support during unregulated excavations at the base of a slope triggering a landslide)?
- What is the likelihood of the project being affected by instability, erosion, or deposition processes over

its design lifetime (e.g., is the chance of being affected by a channelized debris flow, or being engulfed by blown sand, very low)?
- What will be the effect of climate change or sea-level rise on the project risks (e.g., will there be accelerated cliff recession or a higher chance of landslide reactivation)?
- Why has a problem arisen (e.g., are there leaking water pipes or is there disruption of sediment transport along a shoreline)?
- What effects will the project have elsewhere (e.g., will there be reduction in floodplain storage or changes in surface run off and erosion potential)?
- What magnitude event should be designed for to provide a particular standard of defence (e.g., what is the expected volume/depth of the 1- in 50-year debris flow event)?

The answers to these questions can be both qualitative and quantitative; for example, recognition of pre-existing landslides with potential for reactivation can be related to rates of change and the magnitude/frequency of events.

Most projects will involve a combination of approaches, including application of historical data, taking measurements, and making maps and models. Historical records can be used to define the timing and frequency of past events, and historical maps, charts, and aerial photographs may aid in determining past rates of change. Measurement and monitoring of change are accomplished using a variety of field-based or remote-survey techniques (e.g., aerial photography and satellite imagery). Mapping and characterization of the landscape (terrain evaluation and geomorphological mapping) can be applied in developing geomorphological models to provide a framework for the prediction of hazards and future changes. Geographical information systems (GIS) are important tools for the storage, management, and analysis of the geomorphological information collected throughout a project.

Historical Records and Maps

In many countries, there exists a wide range of sources that can provide useful information on the past occurrence of events, including aerial photographs, topographic maps, satellite imagery, public records, local newspapers, consultants' reports, scientific papers, journals, diaries, and oral histories. For example, a

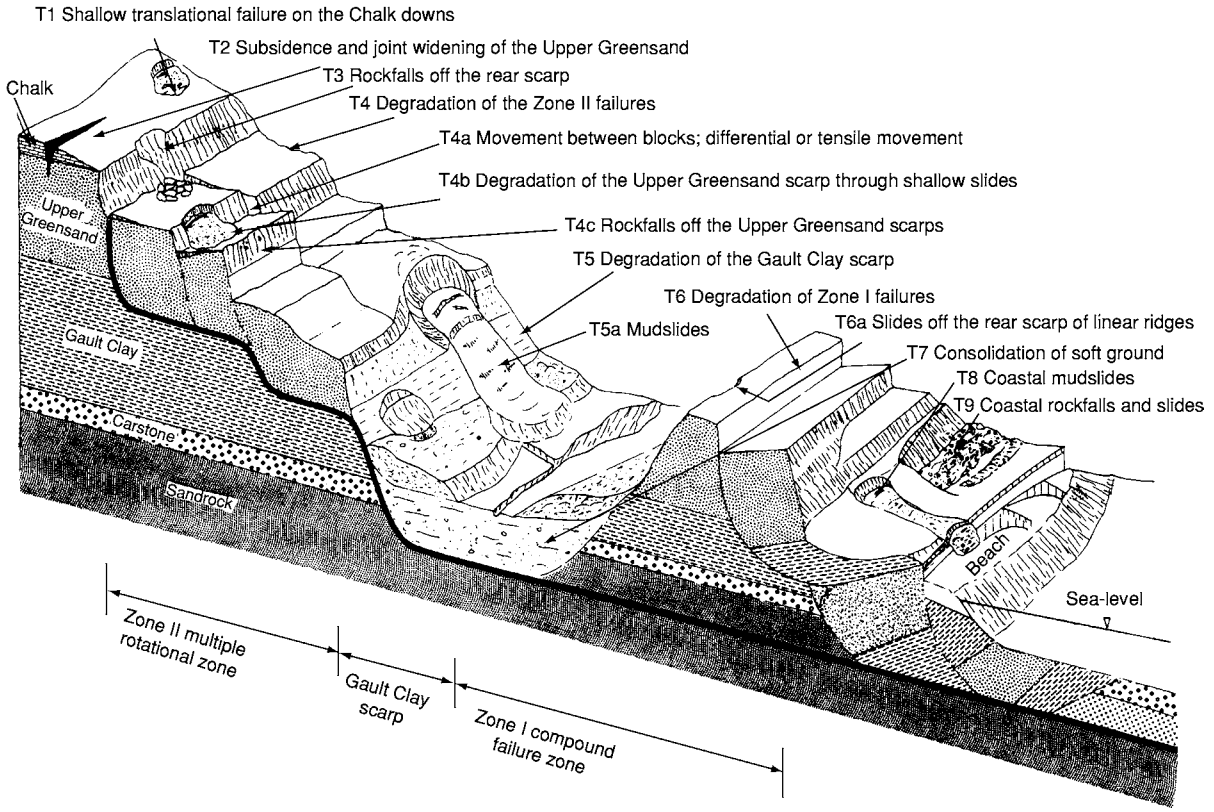


Figure 2 Landslide model of the Ventnor Undercliff, UK, showing different landslide processes (T1 T9). Systematic study of local records revealed that the contemporary movements are largely confined to the superficial degradation of the ancient, deep seated landslide complex. Reproduced from Lee EM and Moore R (1991) *Coastal Landslip Potential Assessment: Isle of Wight Undercliff, Ventnor*. London: Department of the Environment, UK.

systematic survey of local newspapers from 1855 to the present day was used to establish the pattern of contemporary ground movement in the Ventnor landslide complex, Isle of Wight, UK. The search identified over 200 individual incidents of ground movement and allowed a detailed understanding of the relationship between landslide activity and rainfall to be developed (Figures 2 and 3). Historical topographical maps, charts, and aerial photographs provide a record of the former positions of various

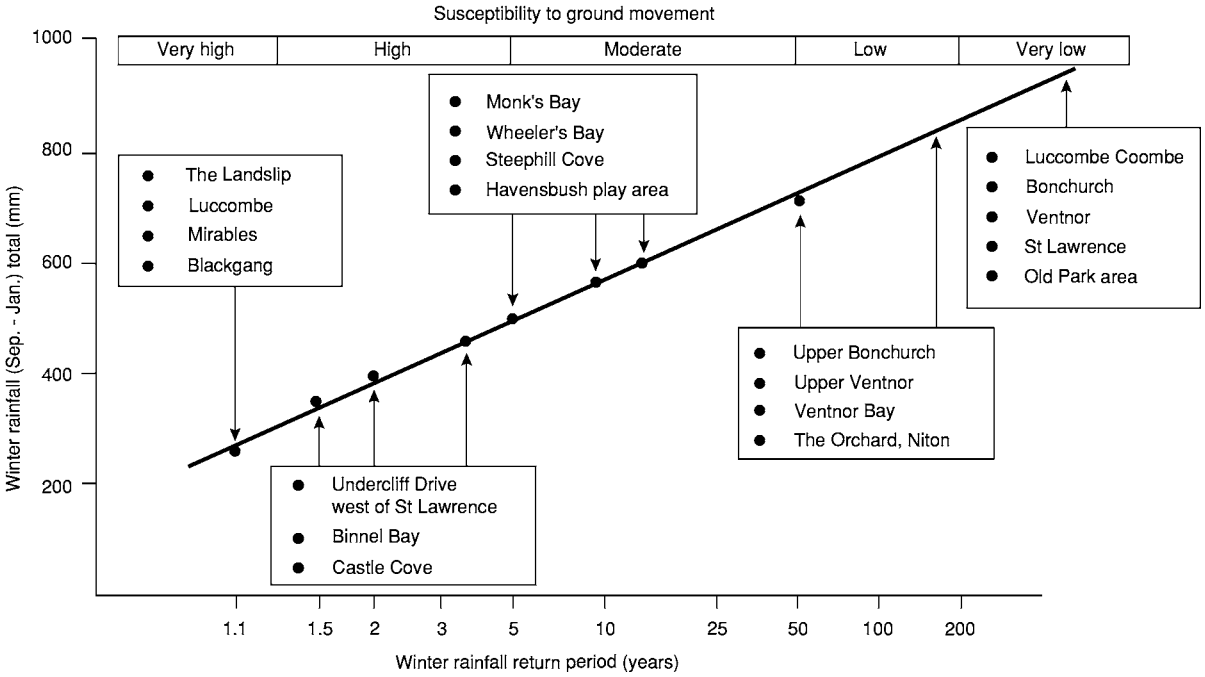


Figure 3 The relationship between rainfall and landslide activity in different parts of the Ventnor Undercliff, UK. This relationship was established by defining the minimum winter rainfall total associated with triggering ground movement over the past 150 years. Reproduced from Lee EM, Moore R, and McInnes RG (1998) Assessment of the probability of landslide reactivation: Isle of Wight Undercliff, UK. In: Moore D, and Hungr O (eds.) *Engineering Geology: The View from the Pacific Rim*, pp. 1315 1321. Rotterdam: Balkema.

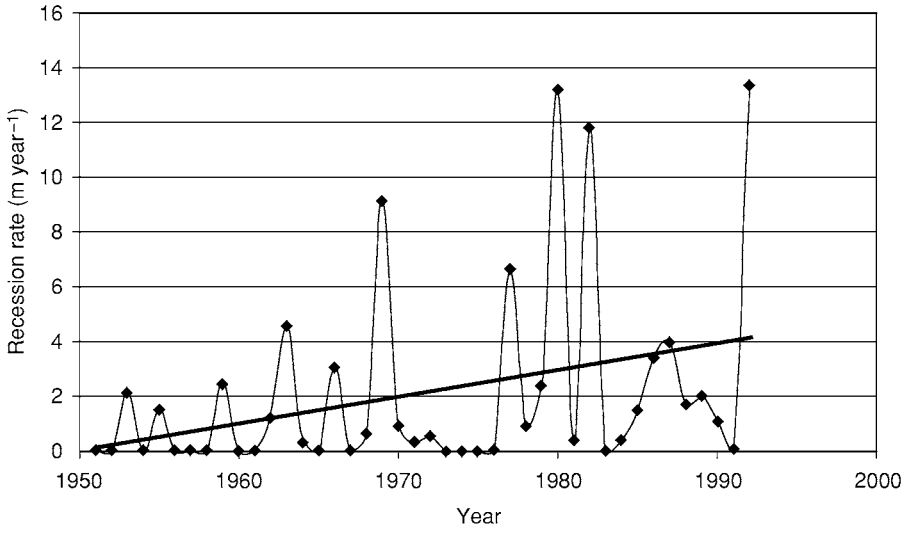


Figure 4 Annual measurements of cliff recession on the Holderness coast, UK. The measurements are for a single erosion marker post along the cliffline. Note that there has been a trend of increasing annual cliff recession over the 40 year period; this may reflect a combination of sea level rise and the impact of coastal defences on the adjacent coastline.

features, such as clifflines and river channel banks. In many cases, historical maps and charts may provide the only evidence of evolution over the past 100 years or more. When compared with recent surveys or photographs, these sources can provide the basis for estimating cumulative land loss and the average annual erosion rate between survey dates. However, great care is needed in using historical data because of the potential problems of accuracy and reliability.

Measurement and Monitoring

Rates of change can be determined through direct measurement of the positions of features at fixed points and at regular intervals. Channel cross-sections, beach profiles, or cliff tops, for example, can be assessed on an annual or biannual basis. In the early 1950s, on the Holderness coast, UK, the local authority initiated a programme of cliff recession measurement that has been continued on an annual basis ever since. A series of 71 marker posts were installed at 500-m intervals along the coastline, each post located at a distance of between 50 and 100 m normal to the coast. Annual measurements from each post to the cliff top commenced in 1953 (Figure 4).

Analytical photogrammetry can be used to quantify the nature and extent of landform changes over time, using aerial photographs taken on different dates, by comparing the three-dimensional coordinates of the same points. This approach was used to establish the rate of contemporary building movement in the Ventnor landslide complex, Isle of Wight. A total of 129 points distributed throughout the town were selected for measurement. The coordinates of each point were determined for each photograph date (1949, 1968, and 1988) and then compared to produce 'discrepancy' vectors. Where the coordinate discrepancy was greater than the standard deviation of coordinates, significant movement was assumed to have occurred.

Terrain Evaluation

Terrain evaluation has its origins in the need to organize and communicate specific earth science information or intelligence in a way that is of direct relevance

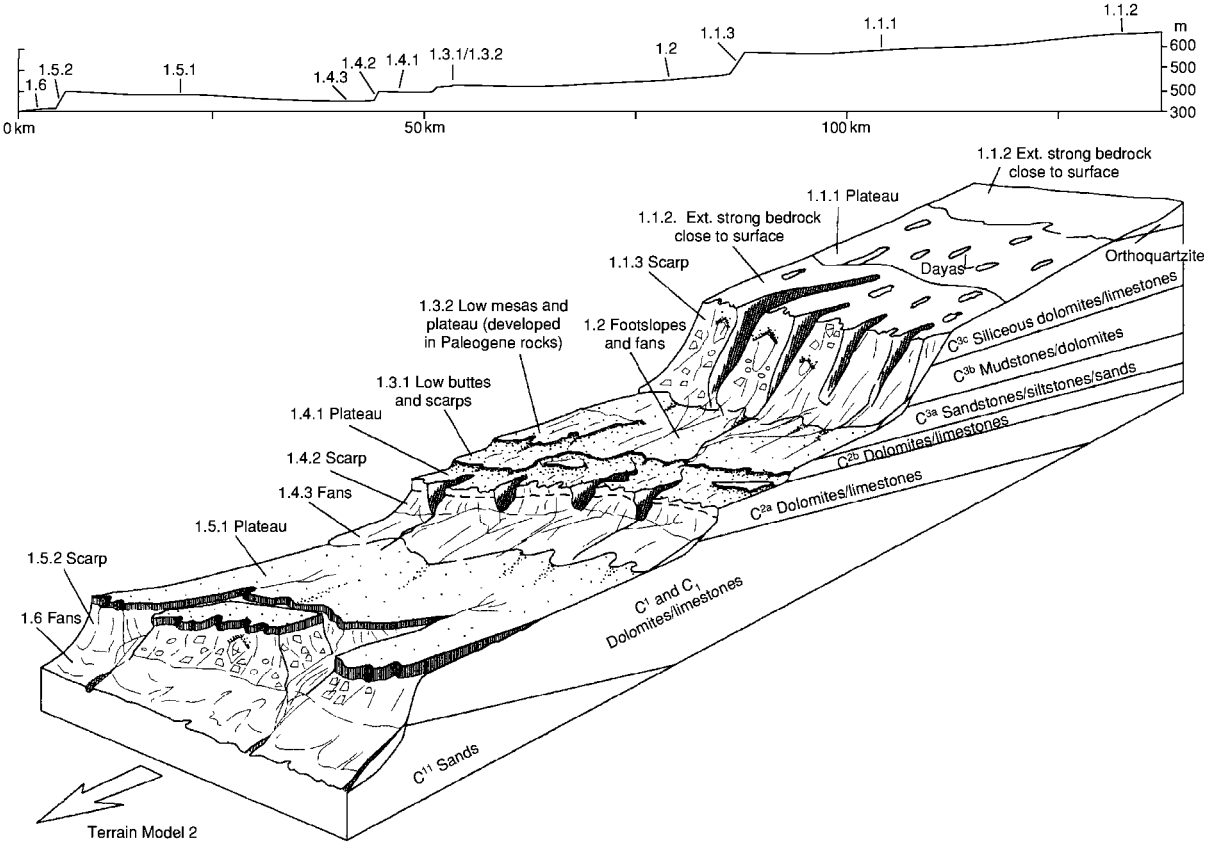


Figure 5 Terrain model 1 of the Tademait Plateau, site of the In Salah Gas pipeline project, Algeria. The numbers (e.g., 1.4.1) represent terrain units (see Table 1); geological codes (C^{3c} etc.) refer to notation on original geological maps of the region. Reproduced with permission from Fookes PG, Lee EM, and Sweeney M (2001) Terrain evaluation for pipeline route selection and characterisation, Algeria. In: Griffiths JS (ed.) *Land Surface Evaluation for Engineering Practice*, *Geological Society Special Publication 18*, pp. 115 121. Bath: Geological Society Publishing House.

Table 1 The In Salah gas pipeline, Algeria: characteristics for terrain unit 1.1.1[a]

Terrain type	*Plateau*
Surface form	Almost flat, very gently sloping (<1°) surface; extensive stone pavement surface, locally disrupted by vehicle traffic; frequent circular and linear enclosed silt/sand filled depressions (dayas) with no stone pavement
Geomorphological processes	Water: surface drainage is primarily into enclosed depressions (dayas); locally, channel flow occurs in straight, single thread ephemeral streams (wadis), less than 50 m wide, with sheet sand and nebkha floors
	Solution: dayas are likely to be formed by localized solution weathering of limestones
Hazards	Flooding: dayas will become flooded during and immediately after rain events; wadi flows are likely to be localized, with rapid transmission losses with limited potential for serious flooding and scouring; estimated Q_{max} <50 $m^3 s^{-1}$; V_{max} <1 $m s^{-1}$
	Subsidence: dayas may overlie infilled solution holes, with potential for subsidence
Superficial materials	Stone pavement over loose to dense calcareous silty fine sands (up to 3 m thick), with calcrete cobbles and gravels
Duricrust	Duricrust occurs as draped sequences on floors and flanks of broad wadi channels; strong to very strong siliceous calcrete within 0.5 m of surface
Bedrock	Massive crystalline limestones (very strong to strong), generally >3 m below surface; bedrock is partly karstified
Water table	Not observed
Rock mass structure	Not observed
Fracture spacing	Duricrust: 0.2 m (estimated)
Rock strength	Siliceous calcrete (strong very strong); limestone (strong very strong)
Excavatability	Easy to hard digging; hard to very hard ripping
Borrow	Random and select fill
Trafficability	No significant constraints
Principal constraints	Potential for subsidence within dayas needs to be considered; very strong materials within 3 m of the surface, especially along floors and flanks of wadis

[a]See **Figure 5**.

to the end user (civil and military engineers, land use planners, agriculturalists, and foresters). The principles of terrain evaluation involve defining areas of terrain that have similar physical characteristics, i.e., a typical range of topographic, geohazard, and engineering factors. The objectives of terrain evaluation are to identify clear associations between surface forms (the terrain units), near-surface materials, and processes; to simplify the complexity of ground conditions and surface processes within a particular area, highlighting those of significance to the planning of projects; and to provide a tool for predicting terrain characteristics within a particular area or region. This is based on the assumption that the terrain units are sufficiently homogeneous and mutually distinctive to allow valid prediction.

At the broadest scale, landscape types (terrain models or land systems) can be defined (e.g., mountains, desert plateaus, or coastal plains); this level of subdivision may be suitable for prefeasibility overviews of very large areas (see **Figure 5** and **Table 1**). Within a landscape type, it will be possible to identify a variety of landform assemblages (terrain units), such as river floodplains, escarpment faces, extensive areas of unstable hillslopes, and ridge crests; this level of detail may be sufficient for corridor assessment. Within a terrain unit, there will be numerous individual landforms (terrain sub-units) that will each present slightly different levels of challenge to a project; an escarpment face, for example, may contain a variety of sub-units, including bare rock faces and discrete landslide systems separated by stable ridges and spurs.

Geomorphological Mapping

The production of some form of map underpins many engineering geomorphological studies. The map might be the product of an intensive fieldwork programme, a sketch map of part of an area based on a walkover survey, or interpretation of remote imagery (e.g., aerial photographs or satellite images). Most map making follows three stages: recognition of landforms or landform elements (units) that provide a practical framework for addressing the problem in hand, characterization of these units in terms of the significant surface processes (these may be active or relict) and the near-surface materials, and interpretation of the significance of these forms, processes, and materials to the problem facing the project. This may involve producing some form of derivative map (e.g., a landslide hazard map or aggregate resource map). A good example is the extensive programme of geomorphological mapping that was carried out as part of the investigations at the Channel Tunnel portal and terminal areas near

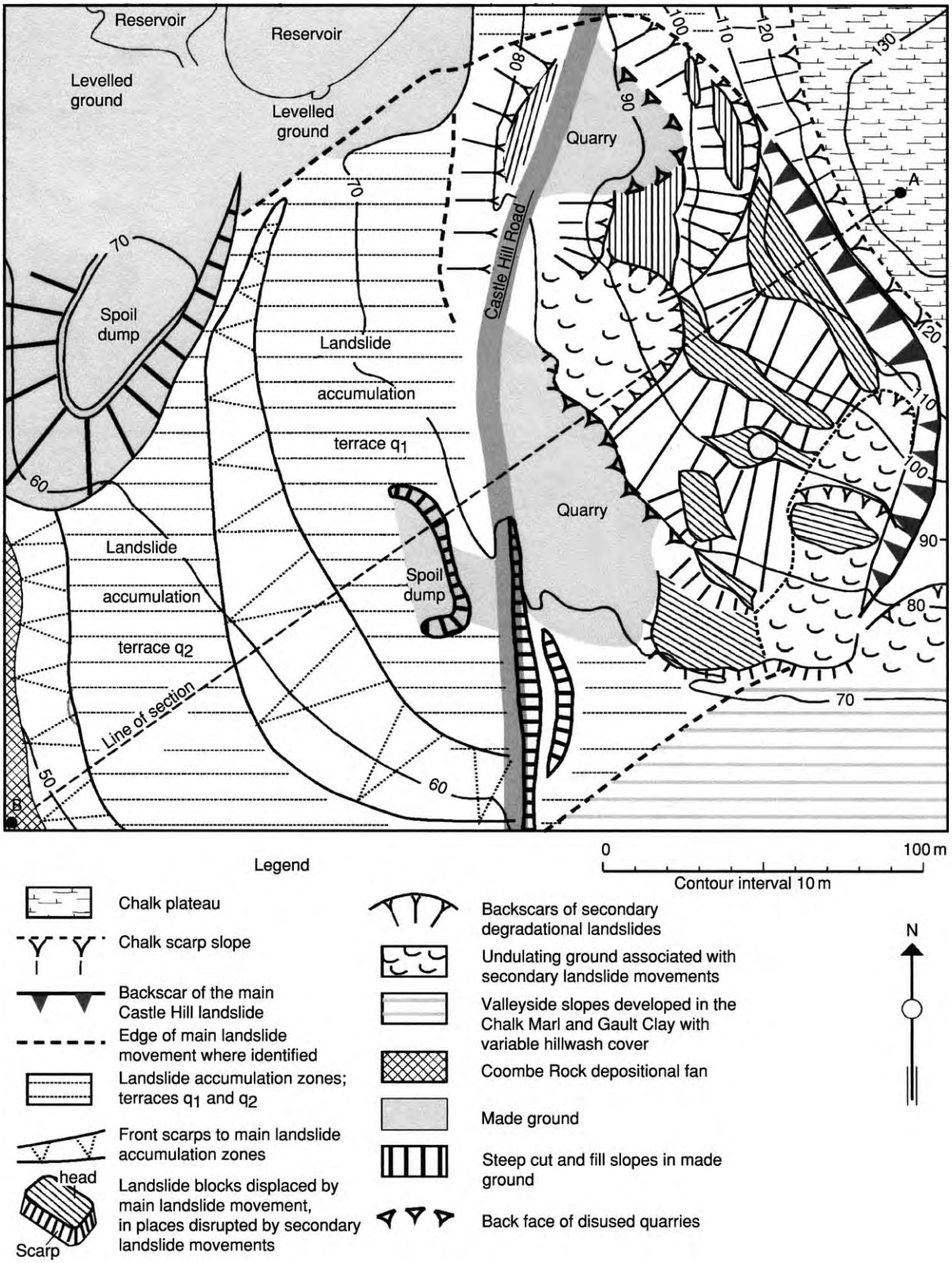


Figure 6 Engineering geomorphological map of part of the Channel Tunnel terminal area, Folkestone, UK. The map defines the extent of the Castle Hill landslide at the site of one of the portals for the tunnel. The line of section relates to another map in the source paper. Reproduced with permission from Griffiths JS, Brunsden D, Lee EM, and Jones DKC (1995) Geomorphological investigations for the Channel Tunnel Terminal and Portal. *The Geographical Journal* 161(3): 275 284.

Folkestone, UK (Figure 6). The Channel Tunnel terminal and portal on the UK side is located immediately below a Lower Chalk escarpment (the North Downs) that was known to be mantled by a series of large, ancient landslides. The objective of the mapping exercise was to delimit and define the nature of past and contemporary landslide activity within the general area of the construction site. Field mapping was undertaken at 1:500 scale. The large-scale geomorphological maps refined the boundaries of the landslides shown on earlier geological maps. The mapping was also of value in showing the form and complexity of the landslide units. This complexity could then be allowed for during the interpretation of borehole data, and for the engineering design.

Geomorphological Models

One of the main uncertainties in engineering is the risk of unexpected conditions (such as previously unrecognized hazards) or impacts and unanticipated changes. Failure to anticipate hazards and risks generally results from an inadequate understanding of a site and its context within a broader system. These problems can be minimized by the development of a model (a 'total geomorphological model') of the expected conditions at an early stage in the project.

Initial (conceptual) models can be based on desk study of available sources, concentrating on identifying the relevant systems and subsystems, along with the main environmental controls. The model should develop progressively to be site specific as the understanding of local conditions improves during the project, leading to a site model. Typical models include hazard models, which identify the nature and distribution of geohazards that might occur within an area, and ground models, which characterize the landforms and near-surface materials within an area. These models can be used to develop a checklist of questions to be answered during subsequent stages of investigations. Another commonly used model, the landform change model, identifies past changes within a system, including the interrelationships between system components. This type of model can provide the framework for developing future scenarios (e.g., the impact of climate change on the risks to a project).

See Also

Engineering Geology: Overview; Natural and Anthropogenic Geohazards; Made Ground; Site and Ground Investigation; Subsidence. **Remote Sensing:** GIS. **Sedimentary Processes:** Landslides; Fluxes and Budgets.

Further Reading

Fookes PG, Baynes FJ, and Hutchinson JN (2000) Total geological history: a model approach to the anticipation, observation and understanding of site conditions. In: *Proceedings, GeoEng 2000 International Conference on Geotechnical and Geological Engineering, 1, Mel bourne*, pp. 370 460. Basel: Technomic Publishing.

Fookes PG, Lee EM, and Griffiths JS (2005) *Foundations of Engineering Geomorphology*. Latheronwheel, Caithness: Whittles Publishing.

Fookes PG, Lee EM, and Milligan G (eds.) (2004) *Geo morphology for Engineers*. Latheronwheel, Caithness: Whittles Publishing.

Fookes PG, Lee EM, and Sweeney M (2001) Terrain evalu ation for pipeline route selection and characterisation, Algeria. In: Griffiths JS (ed.) *Land Surface Evaluation for Engineering Practice, Geological Society Special Pub lication 18*, pp. 115 121. Bath: Geological Society Publishing House.

Griffiths JS (ed.) (2001) *Land Surface Evaluation for Engineering Practice, Geological Society Special Publica tion 18*. Bath: Geological Society Publishing House.

Griffiths JS, Brunsden D, Lee EM, and Jones DKC (1995) Geomorphological investigations for the Channel Tunnel Terminal and Portal. *The Geographical Journal* 161(3): 275 284.

Lee EM and Clark AR (2002) *Investigation and Manage ment of Soft Rock Cliffs*. London: Thomas Telford.

Lee EM and Moore R (1991) *Coastal Landslip Potential Assessment: Isle of Wight Undercliff, Ventnor*. London: Department of the Environment, UK.

Lee EM, Moore R, and McInnes RG (1998) Assessment of the probability of landslide reactivation: Isle of Wight Undercliff, UK. In: Moore D and Hungr O (eds.) *Engin eering Geology: The View from the Pacific Rim*, pp. 1315 1321. Rotterdam: Balkema.

Morgan RPC (1986) *Soil Erosion and Conservation*. Harlow: Longman.

Thorne CR, Hey RD, and Newson MD (eds.) (1997) *Applied Fluvial Geomorphology for River Engineering and Management*. Chichester: John Wiley and Sons Ltd.

Geophysics

J K Gascoyne and A S Eriksen, Zetica, Witney, UK

Introduction

Geophysics can be defined as the study of the Earth through the measurement of its physical properties. Use of the discipline dates back to ancient times, but only since the advent of modern-day instrumentation has its application become widespread. The development of modern geophysical techniques and equipment was initially driven by oil and mineral exploration during the early to middle parts of the twentieth century, and many of the instruments used today in engineering geophysics owe their evolution to the field of exploration geophysics.

Engineering geophysics involves using geophysical techniques to investigate subsurface structures and materials that may be of significance to the design and safety of an engineered structure. Unlike the deeper investigations associated with exploration geophysics (up to 2–3 km), engineering surveys are usually concerned with investigation of the near-surface, at depths in the range of 1–100 m.

The key advantages of geophysics over intrusive site-investigation techniques, such as digging trial pits or drilling boreholes, are that geophysical methods are comprehensive and non-invasive. Large areas can be evaluated rapidly without direct access to the subsurface. One class of engineering geophysics, borehole geophysics, is an exception in that it makes use of boreholes already drilled to sample the local area around the borehole.

When combined with intrusive methods, geophysics provides a cost-effective means of analysing the undisturbed subsurface to aid selection of, and interpolation between, widely spaced sampling locations.

Engineering geophysics can be applied throughout the life cycle of an engineered structure, starting with the initial ground investigation to determine the suitability of a particular site and provide design-sensitive and critical safety information. This may be followed by materials testing during the various stages of construction, monitoring the impact of construction on surrounding structures, on-going monitoring of the integrity of structures after completion, and helping to determine when to schedule essential maintenance tasks, such as pavement or ballast renewal on a road or railway, respectively.

The success of all geophysical methods relies on there being a measurable contrast between the physical properties of the target and those of the surrounding medium. The properties used are typically density, elasticity, magnetic susceptibility, electrical conductivity, and radioactivity. Knowledge of the material properties likely to be associated with a target is thus essential to guide the selection of the correct method to be used and to interpret the results obtained. Often a combination of methods provides the best means of solving complex problems. It is sometimes the case that, if a target does not provide a measurable physical contrast, the association of the target with other measurable conditions may indirectly lead to detection.

Methods

Engineering geophysical methods can be split into two main categories – passive and active.

With passive methods, naturally occurring sources, such as the Earth's magnetic field, over which the observer has no control, are used to detect abnormal variations in background caused by the presence of the target. Interpretation of this data is non-unique and relies heavily on the knowledge of the interpreter.

Active methods involve generating signals in order to induce a measurable response associated with a target. The observer can control the level of energy input to the ground and measure variations in energy transmissibility over distance and time. Interpretation of this data can be more quantitative with improved depth control compared with passive methods, but ease of interpretation is not guaranteed.

Table 1 lists of some of the techniques most commonly used in engineering geophysics.

Measurements are commonly taken at the surface and from boreholes, underground mineworkings, over or under water, or from aircraft platforms. The advent of powerful computer-aided modelling has led to the development of a number of sophisticated imaging techniques, such as cross-hole seismic and resistivity tomography and reflective tomography, which are capable of imaging the properties of the ground in three dimensions between the surface and two or more boreholes or beyond the face of a tunnel.

Armed with a knowledge of the physical properties of a target (see Table 2), its burial setting, and the requirements of the survey, a feasibility assessment is carried out by a geophysicist to determine the likely deliverables of a geophysical survey. Based on the results of this assessment, an appropriate geophysical

Table 1 List of techniques commonly used in engineering geophysics

Technique	*Passsive/active*	*Physical property used*	*Source/signal*
Magnetics	Passive	Magnetic susceptibility	Earth's magnetic field
Microgravity	Passive	Density	Earth's gravitational field
Continuous wave and time domain electromagnetics	Active/passive	Electrical conductivity/resistivity	Radio frequency electromagnetics (Hz/kHz band)
Resistivity imaging/sounding	Active	Electrical resistivity	DC electrical (<1000 V)
Induced polarization	Active	Electrical resistivity/complex resistivity, and chargeability	DC electrical
Self potential	Passive	Electrokinetic	Streaming and diffusion potentials
Seismic refraction and reflection/sonic NDT	Active/passive	Density/elasticity	Explosives, weight drop, vibration, earthquakes
Radiometrics	Active/passive	Radioactivity	Natural or controlled radioactive source
Ground penetrating radar	Active	Dielectric (permittivity)	Pulsed or stepped frequency microwave electromagnetics (50 2000 MHz)
Wireline logging	Active/passive	Variety	Variety

Table 2 Typical range of physical property values for selected soils, geology, and man made materials

	Physical property and relevant example method(s)			
	Seismic P wave velocity range ($m\,s^{-1}$)	*Electrical resistivity range (Ωm)*	*Density range ($kg\,m^{-3}$)*	*Relative dielectric range*
Rock/soil/material type	*Seismic refraction and reflection, cross hole seismic tomography, borehole seismics*	*Two dimensional resistivity imaging, resistivity sounding, electromagnetics*	*Microgravity, seismic reflection, gamma gamma wireline logging*	*Ground penetrating radar*
Air	330		0	1
Water (fresh)	1450 1500	20 100	1000	81
Water (saline)	1450 1500	<10	1000	81
Ice	3100 4000	$>5 \times 10^4$	880 920	3 4
Granite	4000 6000	300 3×10^6	2500 2800	5 10
Basalt	5500 6500	10 1.3×10^7	2700 3300	8 12
Limestone	1700 7000	50 1×10^7	1900 2900	6 9
Chalk	1800 3500	40 200	1500 2600	8 15
Sandstone	1400 4500	1 7.4×10^8	1600 2800	4 6
Sand (dry)	200 1100	80 1000	1700 2300	3 6
Sand (saturated)	1500 2000	20 200		>20
Gravel	1000 2500	100 1400	1700 2500	5 20
Sand and gravel	400 2300	30 250	1700 2400	6 30
Glacial till	800 2300	15 50	1800 2100	8 30
Clay	1000 2500	4 150	1600 2600	3 15
Alluvium	1800 2000	10 100	1950 2000	>10
Concrete	3000 4000	10^1 10^6	1750 2400	5 30
Ballast				3 8
Made ground	160 600	20 2000	1400 2000	8 15
Tarmac	>2000			3.5 6

methodology can be designed to meet the survey objectives.

Once acquired, geophysical data needs to be processed and interpreted to provide meaningful information to the engineer. Data analysis plays a vital role in the successful application of engineering geophysics and is just as important as the selection of the survey method, survey design, and the practical skills needed to collect good data.

A key aspect of processing geophysical data ready for interpretation is the recognition and removal or separation of effects unrelated to the target being investigated. As an example, consider raw data collected during a microgravity survey. These data incorporate significant effects, such as the elevation of the site relative to sea-level, change in the Earth's radius with latitude, tides, local topography, and density of intervening materials, in addition to any

mass variation caused by the target itself. Successful use of microgravity data relies on measuring and removing the known influences related to position and elevation and contributions to the observed field caused by non-target sources. The importance of this is emphasized by the fact that these extraneous effects can be orders of magnitude greater than the signal related to the target itself.

Interpretation may be either qualitative or quantitative depending on the type of method employed, the end use of the data, and available budgets. Qualitative analysis simply involves delineating zones of property change (so-called anomalous areas) within the subsurface, with little or no attempt being made to determine the physical parameters of the target. This type of analysis is common in brownfield site characterization surveys where the objective of the survey is to locate buried objects such as utilities and underground storage tanks (Figure 1). Quantitative interpretation of the data, on the other hand, may include modelled estimates of physical properties, including target size and depth, and a model of the site ground conditions (Figure 2).

As evidenced above, the display of the data and final interpretation can take a variety of forms, depending on the method deployed and the objectives of the survey.

The detection of buried hazards, such as cavities or ordnance, generally involves the acquisition of data over a wide area, using techniques such as magnetic profiling or microgravity, and consequently results are normally presented as contoured or colour-coded plots.

In the case of linear surveys, such as bedrock profiling using seismic refraction or resistivity imaging, the results are generally presented as profiles or cross-sections. Three-dimensional visualization of processed geophysical data and geophysical models is also an effective tool to simplify communication to non-geophysicists. Figure 3 illustrates some typical data-presentation formats.

Survey Design

As we have seen, modern engineering geophysics encompasses the use of a wide suite of geophysical methods. The first job of the geophysicist when faced with a particular engineering problem is to design a suitable survey based on the survey objectives and the nature of the target and its surroundings (both above and below ground). The survey proposed must also be cost-effective and offer cost benefits over 'traditional' site-investigation methods (*see* **Engineering Geology:** Site and Ground Investigation).

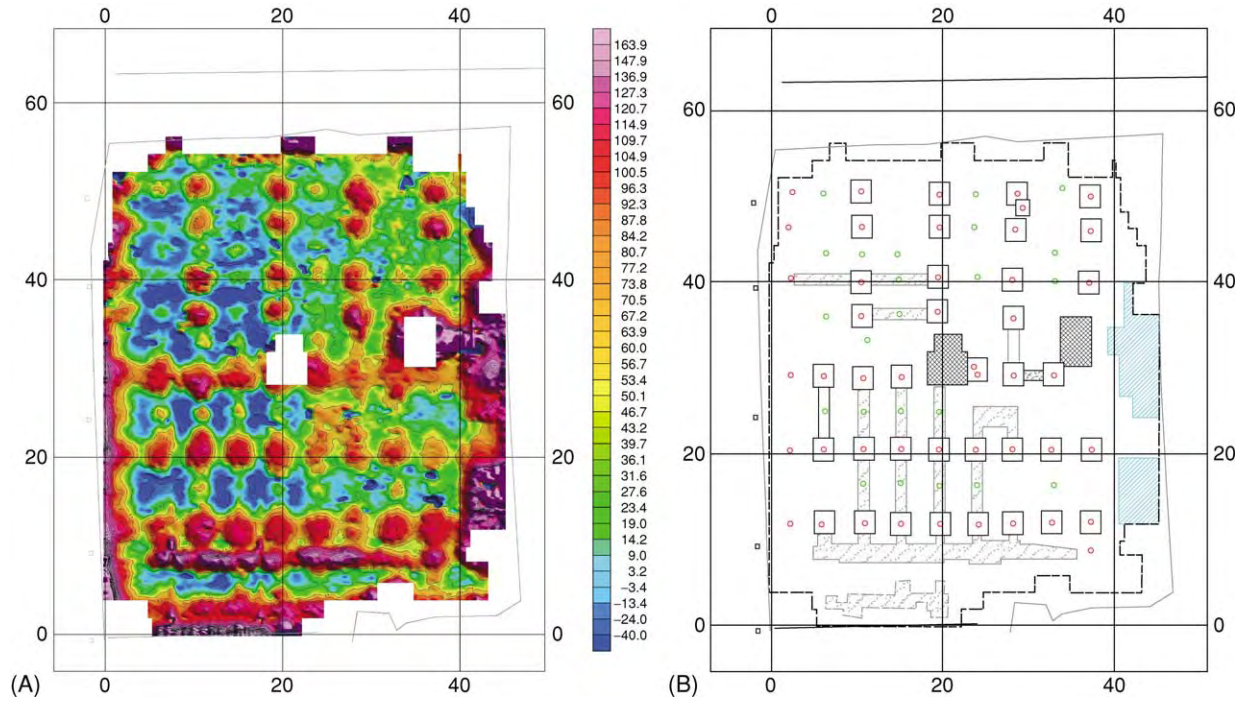


Figure 1 Example of qualitative interpretation of geophysical data to locate buried piles. (A) Time domain electromagnetic data collected on a 1 m × 0.5 m grid and interpolated onto a grid. (B) Interpretation showing the location of all in ground piles and other structures. The survey was 100% effective, facilitating the installation of new services without hitting any obstructions.

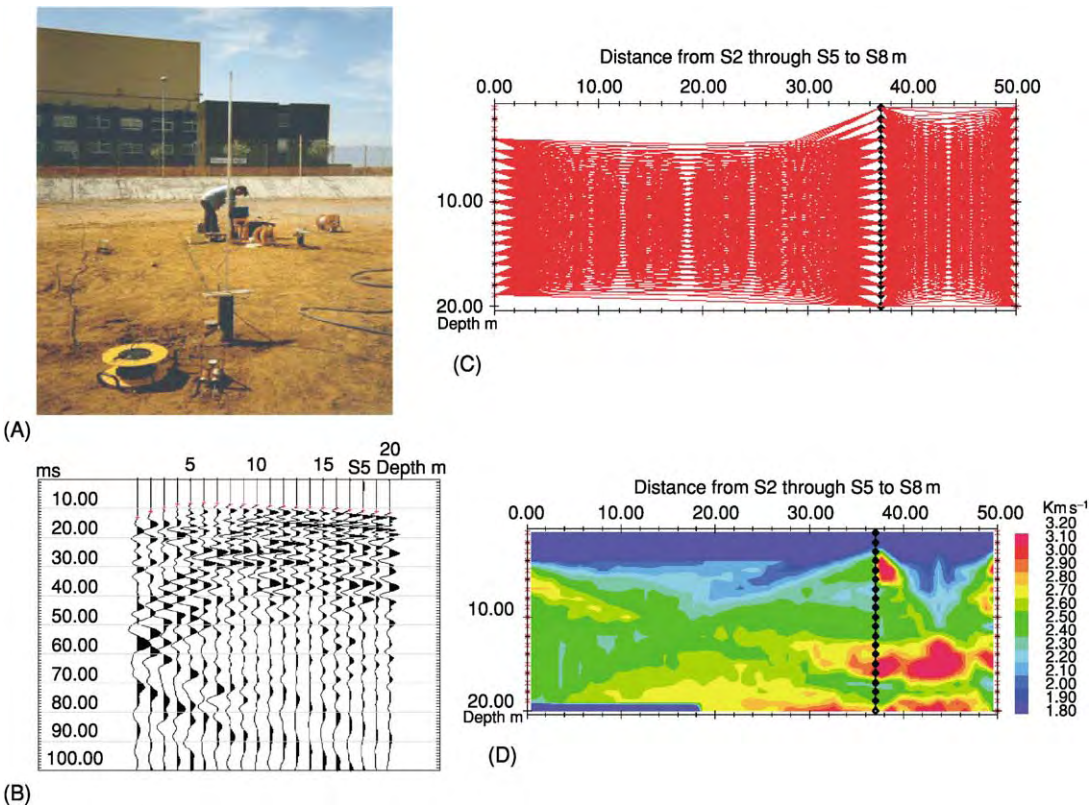


Figure 2 Example of quantitative interpretation of cross hole seismic data using two dimensional computer modelling. (A) Data collection on site using a pair of boreholes, one containing receivers and the other used for locating the seismic source, which is lowered at regular intervals. (B) Raw data showing first arrivals and later events. (C) Modelled ray trace based on matching travel times to measured first arrivals from the raw data. (D) Modelled velocities showing localized high velocity features (red).

It is worth re-emphasizing the fact that geophysical techniques cannot be applied indiscriminately. Knowledge of the material properties likely to be associated with a target is essential to choosing the correct method(s) and interpreting the results obtained.

The successful design and implementation of a survey requires careful consideration of the following main factors.

- Target discrimination. The nature and degree of contrast in physical properties between the target and its surroundings is a primary influence on the feasibility assessment of geophysics and on the choice of techniques. Information regarding the target and the expected surrounding materials may be limited or non-existent, and in these cases the geophysicist should recommend a trial survey or the application of multiple techniques on the site.
- Detection distance. All geophysical methods are sensitive to the relationship between target size and detection distance in addition to the material composition of the target and its surrounding materials. In general, the greater the detection distance, the larger a target's volumetric size and/or cross-sectional area must be for it to be detectable and the greater the required spacing between sample points (Figure 4).
- Survey resolution. The correct choice of sampling interval (frequency or spacing of sampling points) is critical to the success of the survey and its cost-effectiveness. For surveys conducted from the surface the sampling interval is dictated by the geophysical 'footprint' of the target, which may be tens of centimetres for small-diameter shallow pipes, a few metres for narrow fault zones, or tens of metres for large-diameter voids at depth. The optimum sampling interval will be the one that samples the anomaly adequately (commensurate with the perceived risk of non-detection) to meet the survey objectives with minimal loss of information. It is also important to note that undersampling can result in so-called 'aliasing' of the anomaly (Figure 5).
- Site conditions. The suitability of a site for collecting good-quality geophysical data is often overlooked in the design of works. The issues affecting data quality that could be of concern depend on the

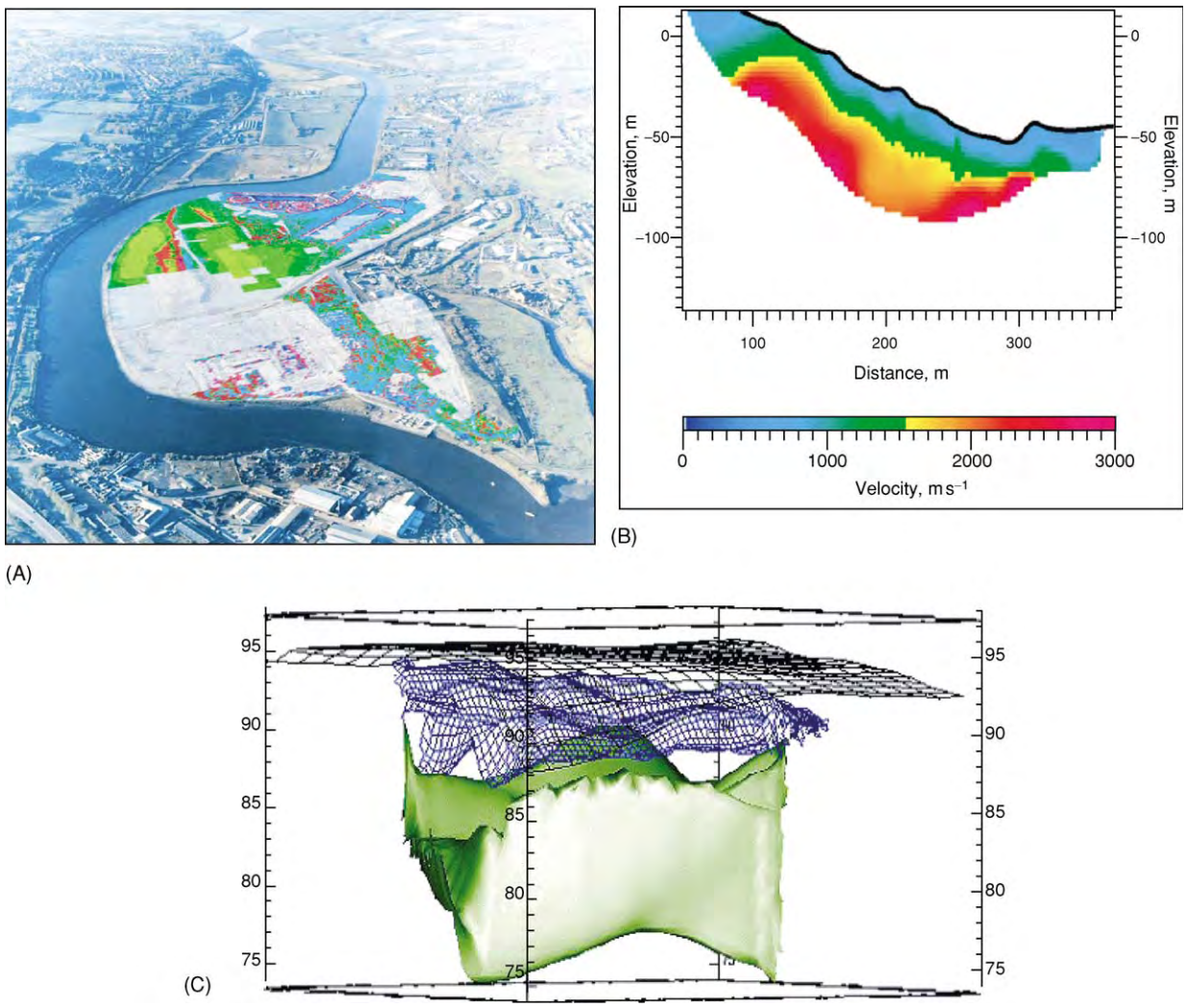


Figure 3 Examples of different display formats. (A) Plot of residual magnetics data showing ferrous targets, draped on an aerial photograph of the 80 ha brownfield site. (B) Modelled seismic velocity cross section for proposed pipeline route. (C) Three dimensional visualization of the geometry of the base of a landfill cell (green) and the top of leachate (blue mesh) derived from a combined two dimensional resistivity imaging and time domain electromagnetic survey.

method or methods being proposed. For example, in the case of electromagnetic and magnetic methods, signal degradation or geophysical 'noise' may be introduced by the presence of surface metallic structures and overhead power lines. For microgravity or seismic surveys, noise may result from traffic movements or wind and waves. Where the level of noise exceeds the amplitude of the anomaly due to the target and where this cannot be successfully removed, the target will not be detectable. The best way to assess the likely influence of site conditions is to visit the site at the design stage and/or carry out a trial survey.

The design of a survey can be aided by the use of powerful two-dimensional and three-dimensional forward-modelling geophysical software. With information on the expected target size, depth, and composition, a geophysicist can evaluate the feasibility of a particular method, including likely error bars on modelled size or depth, and can determine optimum survey design parameters such as the sampling interval and the optimum configuration for the method (e.g. transmission frequency for ground-penetrating radar and electromagnetic surveys or the geophone spacing in a seismic survey).

Figures 6 and 7 illustrate the results of a feasibility study for a tunnelling project in London, UK. Figures 6A and 7A are the client's conceptual model of a Victorian well thought to be associated with a risk of subsidence caused by tunnelling at 20 m depth. The feasibility analysis facilitated an early appreciation of

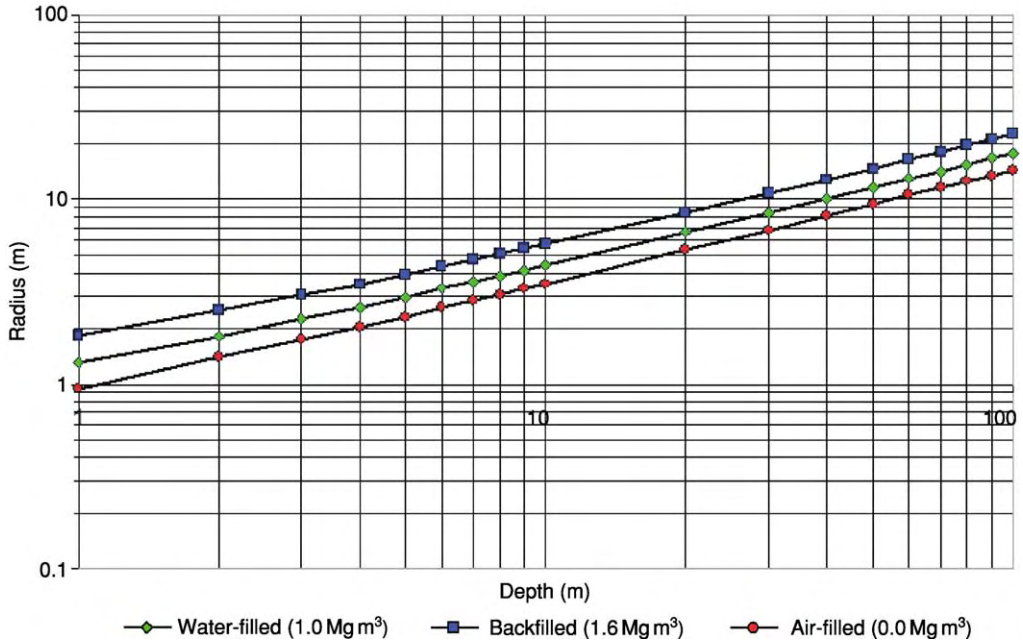


Figure 4 Approximate minimum radius of cavity that will produce a 15 μgal microgravity anomaly in host rock of density 2.3 Mg/m^3.

the limitations of the selected methodology and enabled a go/no-go decision to be made on carrying out a full-scale geophysical survey.

Working closely with the engineer, a geophysical advisor can clarify the likely value of using geophysics in a site investigation. Assuming that geophysics is viable for solving a particular problem, then the next step is to help the client decide what combination of geophysical screening and direct sampling would give the optimum information on the subsurface for the budget available (Figure 8).

Applications

Common applications of engineering geophysics can be broadly separated into seven main classes, each of which encompasses a range of different target types:

1. transport infrastructure – the application of geophysics in the design and monitoring of structures such as roads, railways, bridges, tunnels, and canals;
2. foundation design – the measurement of the engineering properties of soils and bedrock, for example in determining the depth to a pile-bearing layer;
3. pipeline route evaluation – the selection of appropriate route and design criteria, for example measuring soil corrosivity for cathodic protection;
4. hazards identification – the identification and monitoring of natural hazards such as cavities and landslides and man-made structures including mineshafts and unexploded ordnance;
5. non-destructive testing of structures;
6. containment structures – integrity testing and monitoring of structures such as dams and landfills; and
7. buried assets – mapping the location of utilities and underground storage tanks.

Targets tend to be either isolated structures – such as a mineshaft, tunnel, foundations, or underground tank – or laterally extensive – such as bedrock layers, overburden, or groundwater. Table 3 presents some examples of the use of geophysics to solve common engineering problems.

Transport Infrastructure

Geophysical methods have a valuable role to play in both the design and monitoring of linear engineering structures, such as roads, railways, and canals, and related structures, such as bridges and tunnels.

The near-continuous lateral coverage afforded by a technique such as continuous-wave electromagnetic profiling is particularly useful during route selection, as it provides a rapid and cost-effective means of assessing near-surface ground conditions. Reconnaissance surveys of this type may be followed up by more detailed studies using quantitative techniques such as seismic refraction profiling and two-dimensional resistivity imaging within areas identified as being of particular risk or that display anomalous ground conditions (Figure 9).

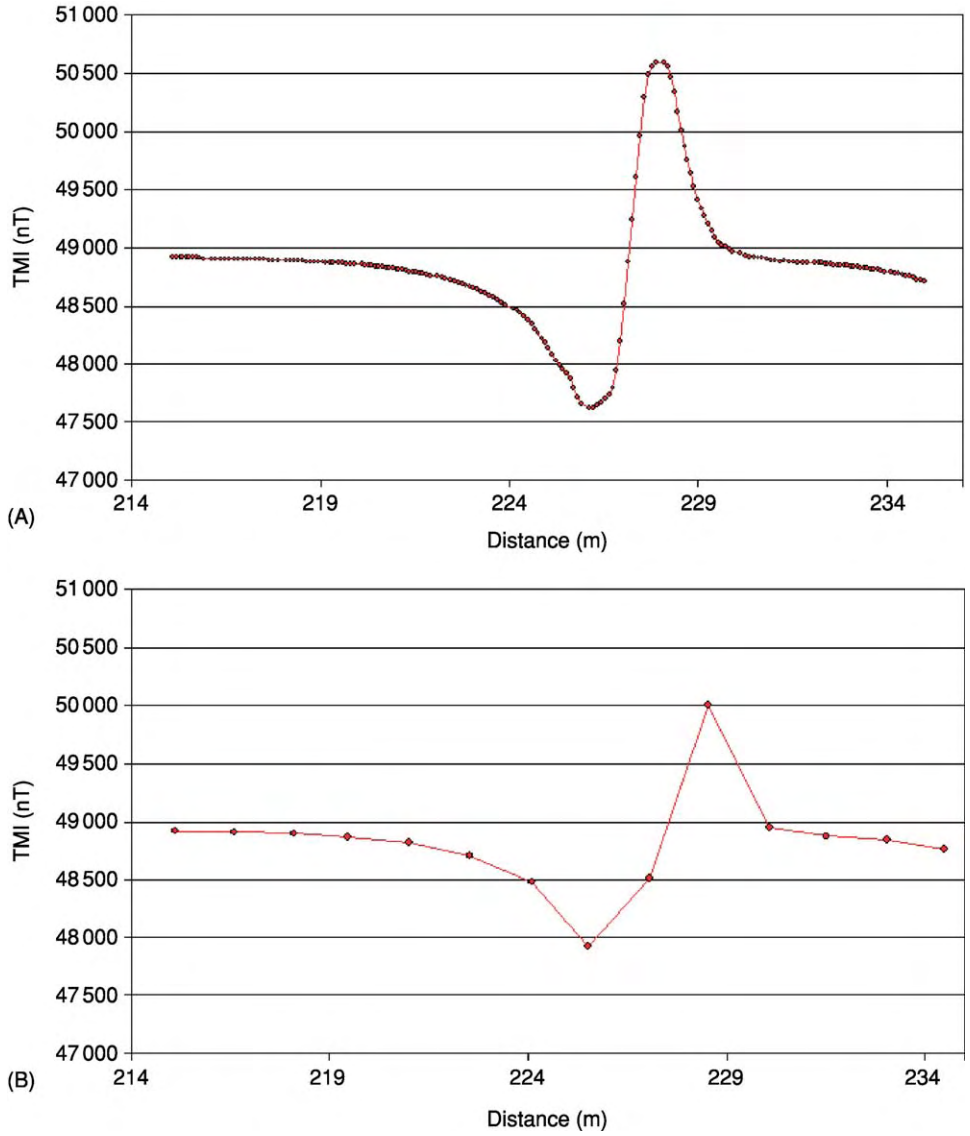


Figure 5 Example of the effects of aliasing on the detection of a pipe using magnetic profiling. The profile shown in **Figure 5A** illustrates correct sampling of the anomaly, whilst the profile shown in **Figure 5B** illustrates the smoothing effects of under sampling.

The main application of seismic refraction profiling is in the determination of bedrock depth and depth to groundwater (Figure 10). However, analysis of the P-wave velocity values obtained during such surveys can also be employed to map the variability in the strength of the bedrock along a route. The evaluation of so-called bedrock rippability is particularly useful in areas where excess material may need to be removed, such as in cuttings and the formation of embankments, in helping to determine the type of plant that will be required.

The heavy weight of traffic on roads and railways means that monitoring the condition of road pavements and track ballast is essential for proactive maintenance and planning of renewals. Ground penetrating radar (GPR) is particularly suited to this task. Modern digital GPR acquisition systems enable detailed mapping of changes in road-pavement construction, layer thicknesses, and ballast thickness and condition at speeds in excess of 100 km h^{-1} (Figure 11).

GPR can also play an important role in the evaluation of structures such as bridges and tunnels. Common applications include the identification of voiding behind brick or concrete tunnel linings and mapping the location and condition of bridge-deck reinforcing.

In order to determine layer thicknesses accurately the velocity of the GPR signal within the materials

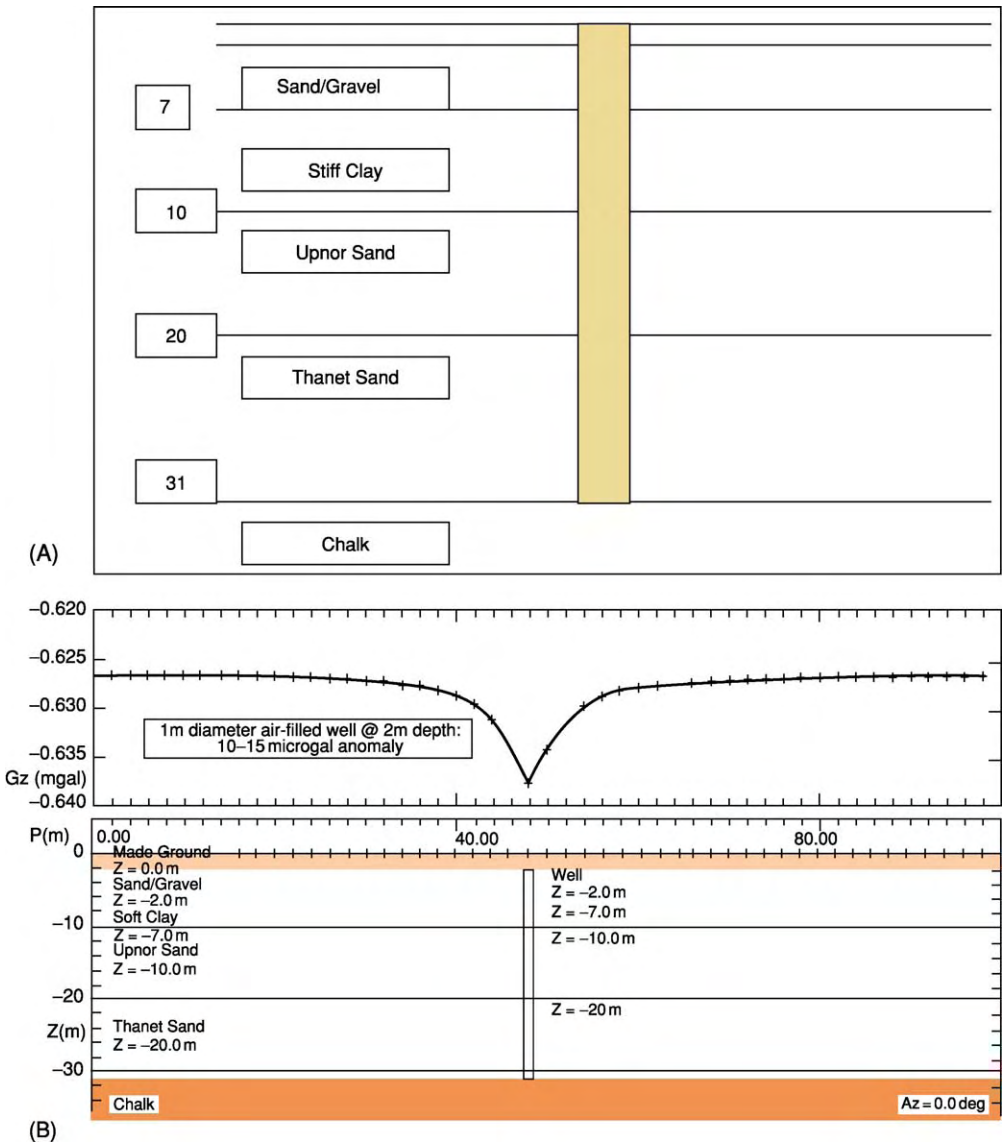


Figure 6 (B) The use of two dimensional forward modelling to determine the feasibility of using the microgravity method to detect (A) Victorian water wells. The size of the anomaly is 10 μGals, which, in the conditions prevalent on site (busy streets in London), would have been below the resolution of modern digital microgravity meters.

overlying an interface needs to be determined. This is normally done directly through the use of core samples but can also be derived indirectly using geometric techniques such as curve-fitting on hyperbolic anomalies and common mid-point analysis. The effective use of core samples requires that the data be accurately located.

Foundation Design

Geophysical methods are routinely used to measure the engineering properties of soils and bedrock as an input to the design of foundation structures including piles. These properties include but are not limited to soil resistivity, shear modulus, Poisson's ratio, and percentage moisture content. These properties are usually derived from measurements taken from the surface or from boreholes. Methods for deriving low-strain elastic moduli include surface-wave seismics, seismic refraction, cross-hole seismics, seismic tomography (Figure 2), and sonic logging.

Wireline logging has been used to correlate stratigraphy between boreholes and provide a clearer definition of stratal boundaries and type. In the

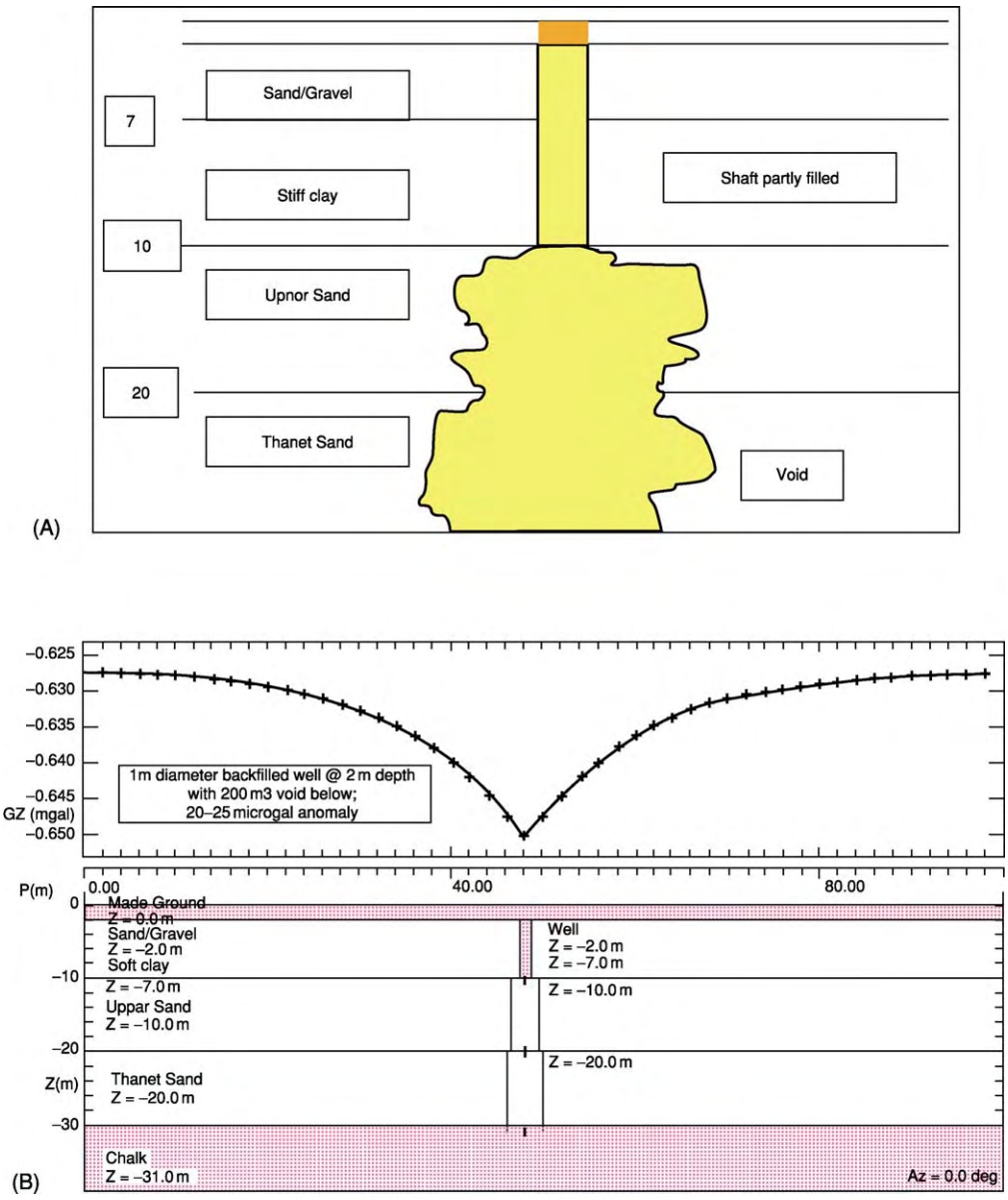


Figure 7 (B) The use of two dimensional forward modelling to determine the feasibility of using the microgravity method to detect (A) collapse structures associated with Victorian water wells. The size of the anomaly is 25 μGals, which is readily detectable using modern digital microgravity meters.

example shown in Figure 12, taken from a log from the Isle of Dogs in London, UK, the electrical log demonstrated a difference between the upper and lower parts of the Thanet Sand that was not obvious when looking at standard soils data such as gradings. The electrical log was matched to measurements of limit pressure in pressuremeter tests, and as a lower bound to SPT (Standard Penetration Test) data. The geophysical tool proved to be a powerful and cost-effective alternative for delineating the pile-bearing horizon in the Thanet Sand.

Pipeline Investigations

Determination of the most suitable route for a pipeline and the type of materials or protective measures required to prevent its corrosion in the ground can both be aided by the use of geophysical techniques.

As in the case of other linear structures such as roads, route evaluation may require the determination of bedrock depth, mapping of lateral variations in the nature of overburden materials, and the

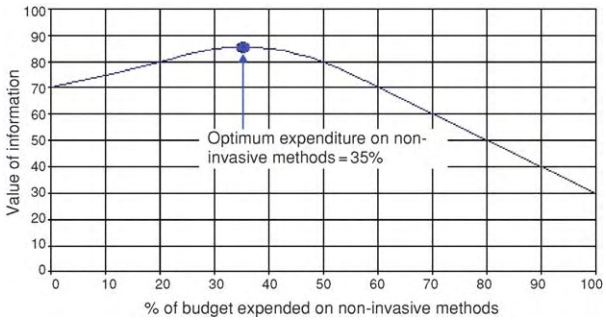


Figure 8 Example of a cost benefit analysis to determine the highest value spend (i.e. most information derived) for a fixed site investigation budget; in this brownfield case 35% of the budget was allocated for site wide geophysical screening, to target trial pits and boreholes that might otherwise miss important features.

location of subsurface hazards or anomalous ground conditions that may affect the construction or long-term integrity of the structure. Methods commonly employed include seismic refraction and electrical resistivity imaging for profiling bedrock depth, together with resistivity soundings and electromagnetic profiling for overburden characterization. The linear nature of pipelines means that they often traverse a diverse range of terrains, and, consequently, the type of hazard that may be encountered will be varied and not always obvious. Figure 13 illustrates data from a detailed microgravity survey undertaken along the proposed route of a pipeline in the UK with the objective of locating buried mineshafts. Forward and inverse modelling of the data allowed the likely size and depth of the shaft to be determined.

Table 3 Examples of the application of geophysical methods in engineering investigations

Application	*Example*	*Typical methods*
Transport infrastructure		
Route evaluation	Determination of thickness of overburden	Seismic refraction, 2D resistivity imaging depending on nature of expected lithologies and water table
	Location of buried channel	Seismic refraction or 2D resistivity imaging for large scale structures
	Identification of karstic structures (swallow holes, cavities)	Continuous wave electromagnetic profiling (reconnaissance), microgravity and seismic refraction for detailed modelling
Pavement/ballast assessment	Measurement of thickness of pavement or track layers and mapping location of construction or condition changes	High speed GPR (200 1200 MHz)
Structural assessment	Mapping location, depth, and condition of reinforcing within concrete	High frequency (900 1500 MHz) GPR, continuous wave electromagnetic profiling
	Identification of voiding, moisture retention behind tunnel wall lining	High frequency (900 1200 MHz) GPR, infrared thermography
Foundation design		
Stratigraphic	Determination of depth to pile bearing layer	Seismic refraction/reflection, cross hole seismics, seismic tomography, surface wave seismics, wireline geophysics
	Mapping boundaries of layers	Seismic refraction, seismic reflection, cross hole seismics, seismic tomography, surface wave seismics, wireline geophysics
	Measuring the stiffness of soils	Seismic refraction, seismic reflection, P and S wave cross hole, up hole, and down hole seismics, cross hole seismic tomography, surface wave seismics, wireline geophysics
	Rock rippability	Seismic refraction
	Groundwater table	Seismic refraction, 2D resistivity imaging, time domain electro magnetic sounding, 1D resistivity soundings, wireline logging
Structural	Mapping fault zones	Very low frequency continuous wave electromagnetic profiling, seismic reflection
Pipeline route evaluation		
Route evaluation geological	Determination of thickness of overburden	Seismic refraction, seismic reflection (P and S wave), 2D resistivity imaging, microgravity, continuous wave electromagnetic, time domain electromagnetic sounding
	Rock rippability	Seismic refraction
	Corrosivity of soils	Resistivity, continuous wave electromagnetic, redox potential

Continued

Table 3 Continued

Application	*Example*	*Typical methods*
Archaeological assessment	Map anthropogenic features such as buried pits, walls, foundations, crypts	Magnetics, resistivity profiling, GPR, 2D resistivity imaging, magnetic susceptibility
Hazards identification	See below	See below
Hazards identification		
Geological	Natural cavities and sinkholes	Microgravity, GPR, 2D resistivity imaging, continuous wave electromagnetic
Brownfield and current industrial	In ground obstructions	Continuous wave electromagnetic, magnetics, time domain electromagnetic, GPR, microgravity
	Mineshafts	Microgravity, continuous wave electromagnetic, magnetics, 2D resistivity imaging, GPR, infrared thermography
	Detection of live cables	Passive electromagnetic
	Detection of unexploded ordnance	Magnetics, time domain electromagnetic, GPR
	Mapping contaminant sources and pathways	Continuous wave electromagnetic, 2D resistivity imaging, spectral induced polarization, self potential, GPR, microgravity
	Mapping existence and number of underground storage tanks	Continuous wave electromagnetic, magnetics, time domain electromagnetic, GPR, microgravity
Containment structures		
Construction	Quality control of construction process	Electric leak location, GPR, resistivity imaging
	Integrity of engineered structure over time	Self potential, 2D/3D resistivity imaging, continuous wave electromagnetic, GPR, borehole geophysics
Buried assets		
Utilities	Verifying existing maps and producing 3D plans	Passive electromagnetic, continuous wave electromagnetic, magnetics, time domain electromagnetic, GPR
Underground storage tanks	Pre purchase audit	Magnetics and/or time domain electromagnetic, GPR, microgravity (for non metallic tanks)

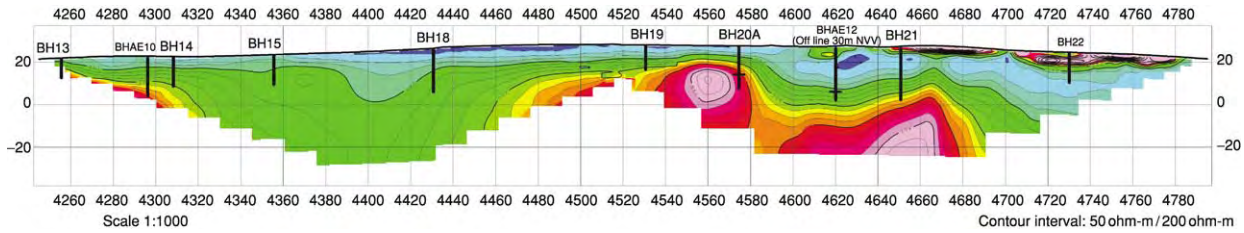


Figure 9 Cross section illustrating the results of a two dimensional resistivity imaging survey along a proposed road route in southern Ireland. The objective of the survey was to determine the thickness of the glacial overburden overlying the limestone bedrock. Areas of high resistivity, relating to the bedrock, appear orange and red, whilst areas of lower resistivity, due to the overlying glacial deposits, appear green and blue. The deep area of overburden on the left of the image relates to infilling of a solution feature.

The prevention of corrosion of buried pipelines requires knowledge of the aggressivity of the soils along the proposed route. A soil's aggressiveness depends on three main factors: its pH, redox potential, and electrical conductivity. Once these factors have been determined suitable preventative measures can be put in place. These can include cathodic protection, zinc coating, or the use of plastic rather than steel pipes.

The soil electrical conductivity (or resistivity) might be determined at selected positions along the route using techniques such as one-dimensional resistivity or time-domain electromagnetic sounding or in a more continuous manner using methods such as electromagnetic profiling and capacitively coupled towed resistivity imaging.

Resistivity soundings have the advantage of providing a profile of the variation in soil resistivity with depth, thereby enabling aggressive layers within the soil to be avoided by altering the depth of the pipe. However, as with other point-sampling investigation tools such as boreholes, areas of importance along a

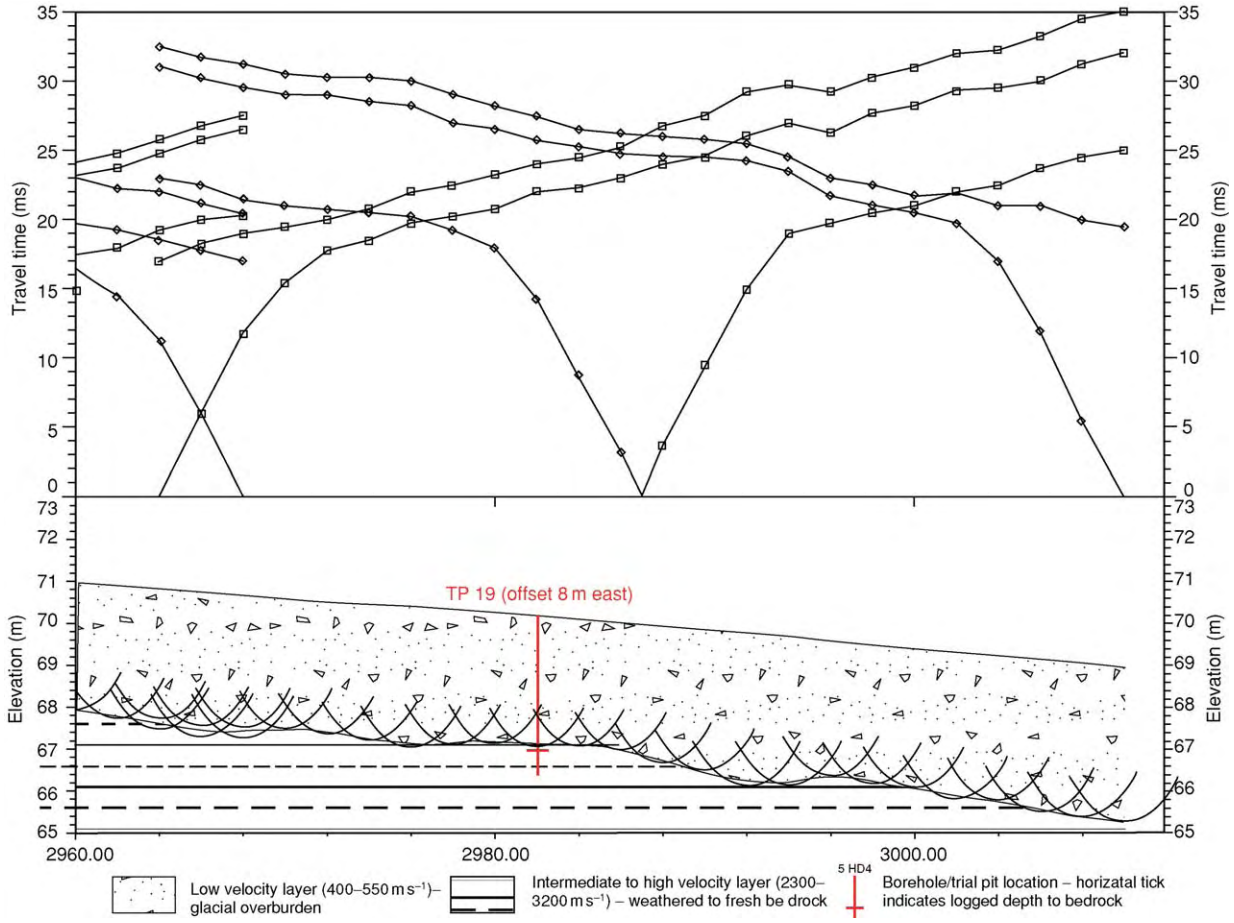


Figure 10 Seismic refraction survey to map depth to bedrock along a proposed road route.

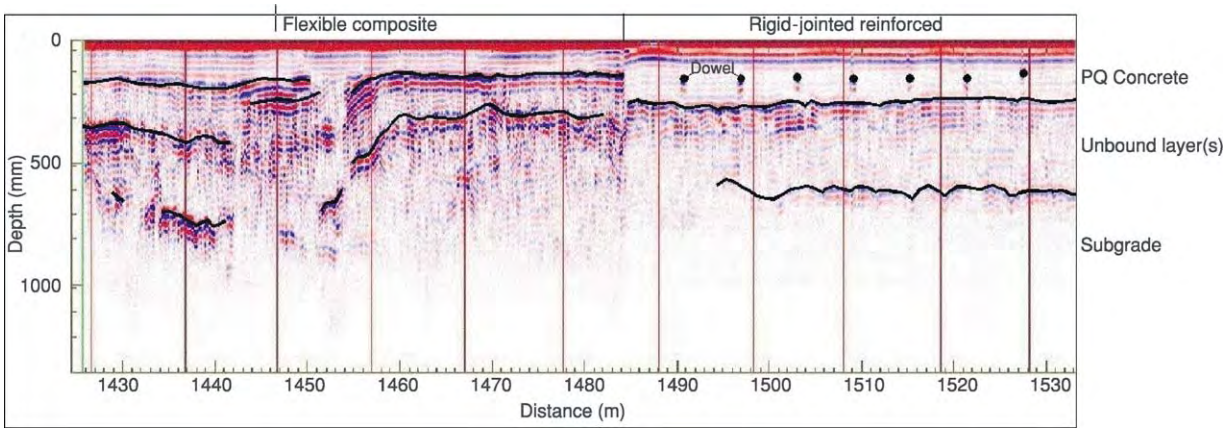


Figure 11 GPR pavement survey, illustrating two types of pavement construction.

route may be missed. Techniques such as continuous wave electromagnetic profiling enable more continuous coverage but require multiple passes with different instrument configurations to determine variations in resistivity with depth. As with other route investigations, it is often beneficial to undertake an initial reconnaissance survey using a profiling method and then target areas of interest with soundings.

Hazard Identification

Knowledge of potential subsurface hazards is critical to the safe development and maintenance of

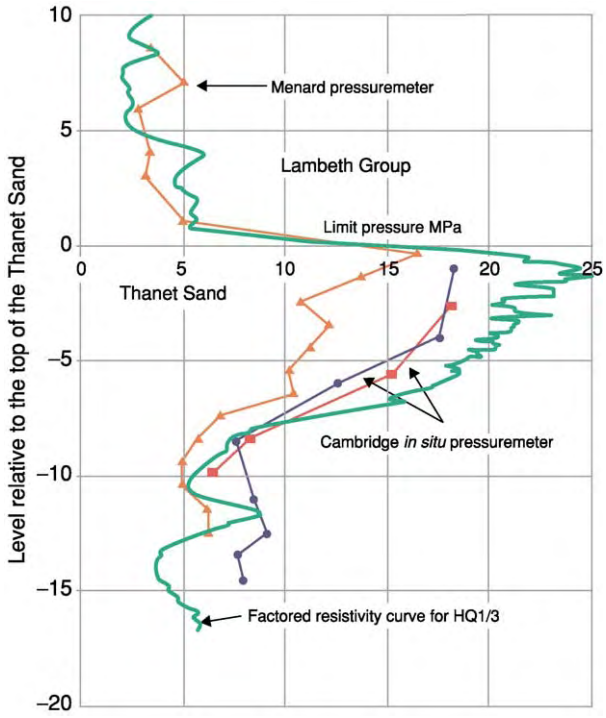


Figure 12 Comparison of geophysical (resistivity in green) and geotechnical (Menard pressuremeter and Cambridge *in situ* pressuremeter) results for a borehole in London, UK. (Published with permission of Arup 2003.)

an engineering structure. A lack of information on the location and nature of both natural and man-made hazards, such as caves, mineshafts, buried landfills, and unexploded ordnance, could potentially be catastrophic and could lead to costly delays.

An example of how geophysics can be used to mitigate the risk posed by natural sinkholes is illustrated in Figure 14. A previous site investigation, incorporating four boreholes located one in each corner of the site, had determined that the depth to chalk bedrock and the condition of the chalk were essentially similar in each borehole. The regulator was informed and permission was given to construct the landfill cell. Not long after commencement, the earthworks contractor encountered large sinkhole features in the chalk, and the works were halted. The regulator requested that a geophysical survey be carried out to clarify the extent of the problem. A ground-conductivity map of the 1 ha area highlighted numerous sinkhole features as ground-conductivity highs. These features were then further investigated using resistivity imaging and cone penetrometer testing to determine the depth of the sinkholes. A hypothetical exercise was carried out to determine how many boreholes would have been required to match the detail provided by the geophysical survey.

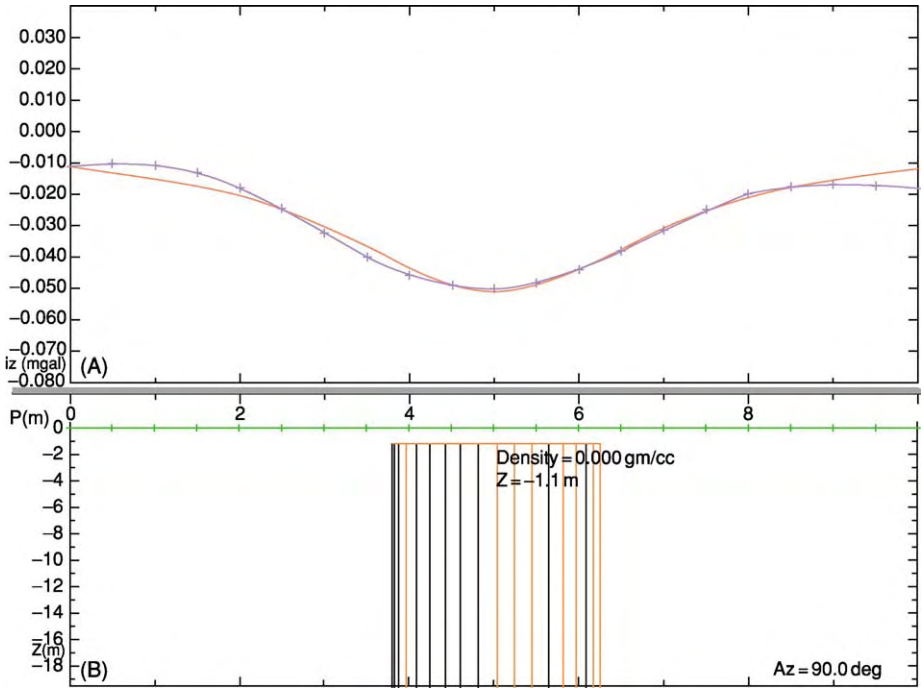


Figure 13 Modelling the microgravity response of a buried mineshaft. (A) The blue line represents the observed gravity data, whilst the red line represents the calculated response due to (B) a cylindrical source body.

Figure 14 shows that for a three-by-three (total of nine) or even a five-by-five (total of 25) grid of boreholes the information on the extent of sinkhole development would have been patchy at best. The most cost-effective sequence of events would have been to carry out the ground-conductivity survey first and then drill three or four boreholes in targeted areas that were revealed by geophysics to be representative of background and anomalies. Figure 15 presents an example of the use of ground penetrating radar to locate potentially hazardous voids within karstic limestone. The high resolution obtained from GPR surveys is ideal for the detection of small-scale near-surface cavities but under good conditions, such as in unweathered hard rock, the technique is also capable of providing information to significant depth. In these circumstances it is often used to complement data obtained from more traditional techniques such as microgravity profiling.

Projects on land associated with, or affected by, military activities may have an increased risk of encountering unexploded ordnance. Examples include

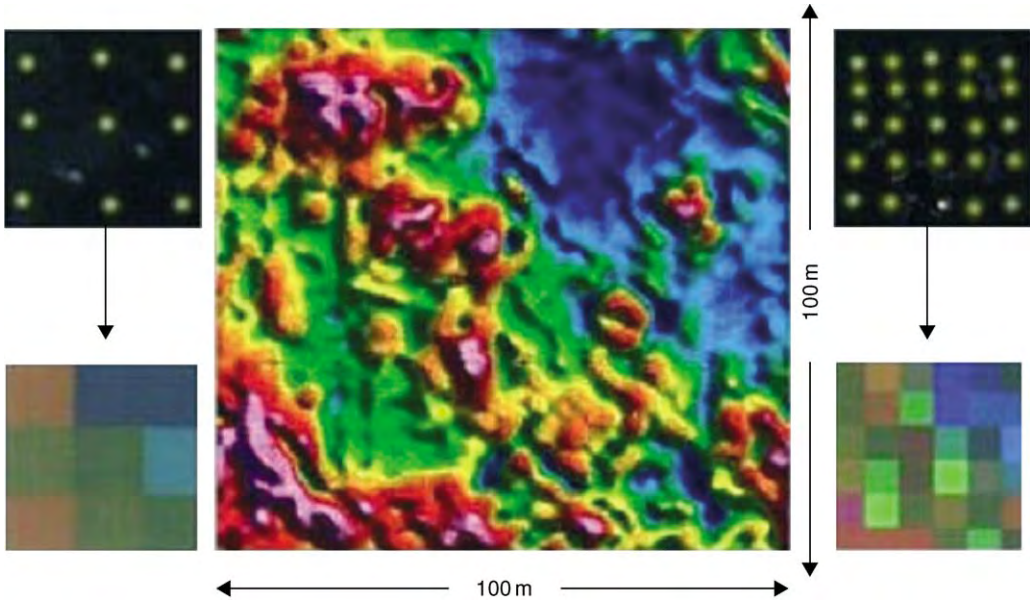


Figure 14 Comparison of the results of undertaking an electromagnetic (apparent conductivity) survey (centre) across a site to map solution features in chalk and a more traditional site investigation approach using just boreholes (left and right). Red and magenta areas on the apparent conductivity map represent sinkholes filled with moist sand. The blue and green areas correspond to competent chalk.

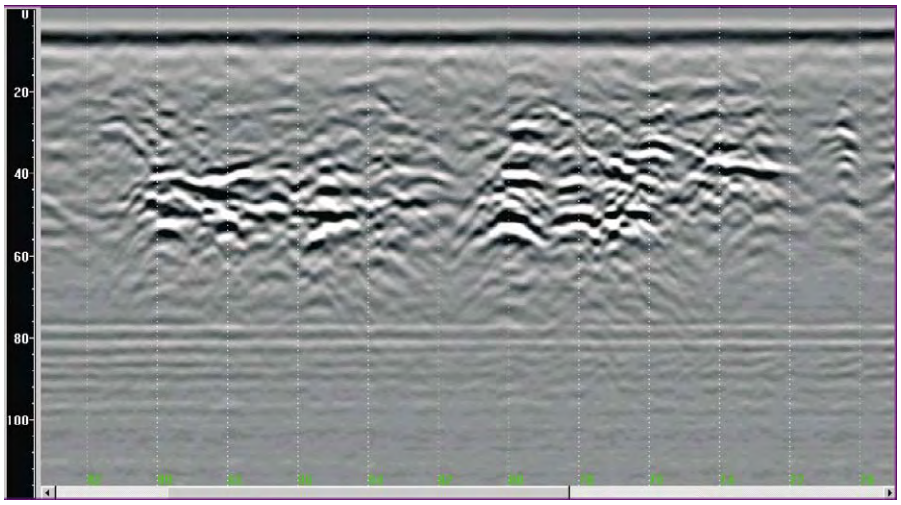


Figure 15 The use of GPR to locate voiding within karstic limestone for a port development in the Dominican Republic. The areas of white and black on this B scan plot (or 'radargram') are indicative of clustered small scale (less than 0.5 m in diameter) voids within otherwise competent limestone. The vertical axis on the plot represents the two way travel time of the GPR signal through the rock measured in nanoseconds.

munitions-factory sites, military bases, military training grounds, and wartime target areas. Apart from low-metal landmines, most ordnance is composed of metal casing, which can be targeted by geophysical methods (depending on the size and depth of burial). In London, UK, which was heavily bombed during the Second World War, there may be a risk of the presence of unexploded bombs to depths of 10 m or more, depending on the size of bomb dropped and soil conditions. Deep bombs such as these are undetectable using surface geophysical methods but can be detected using sensors lowered down boreholes or mounted in push rods. An example of how geophysics can be used to mitigate the risk posed by unexploded ordnance is shown in Figure 16. A site-wide gradient magnetometer survey was carried out on a close grid spacing to detect small ferrous ordnance items. These were then flagged in the field for identification and disposal. Following removal of identified targets a second survey was carried out to confirm that all located ordnance had been removed.

Non-Destructive Testing

Engineering geophysics forms a significant component of the field of non-destructive testing, which encompasses all methods used to detect and evaluate the integrity of materials. Non-destructive testing is used for in-service inspection and condition monitoring of structures and plant including the measurement of physical properties such as hardness and internal stress.

Examples of geophysical methods used in non-destructive testing are tabulated in Table 4.

Containment Structures

As well as playing a role in the assessment of ground conditions prior to construction of a containment structure such as a lagoon or dam, geophysics can also be used to effect in the assessment and monitoring of the integrity of a structure prior to and during use and in the investigation of breaches in containment during its lifetime. In the latter case, where the storage of hazardous materials such as chemical and nuclear wastes is involved, a geophysical survey is often a cost-effective means of investigating the problem as it avoids the need to remove the stored material.

The most common requirement for geophysics at the pre-commissioning stage is in the assessment of the integrity of lined structures such as domestic landfills and leachate lagoons using the electrical leak location method. This technique originated in the USA during the mid-1980s and has since been adopted widely. It now forms a recommended part of the Construction Quality Assurance testing of new landfill cells in the UK. The technique comprises a novel form of resistivity profiling, which takes advantage of the insulating properties of geomembrane materials, such as HDPE (high density polyethylene) and PVC (polyvinyl chloride), in order to pinpoint holes or tears as small as 0.5 mm^2. Variants of the basic method have been developed for testing soil-covered liners, cell side slopes (with no cover material), and water-filled lagoons and ponds.

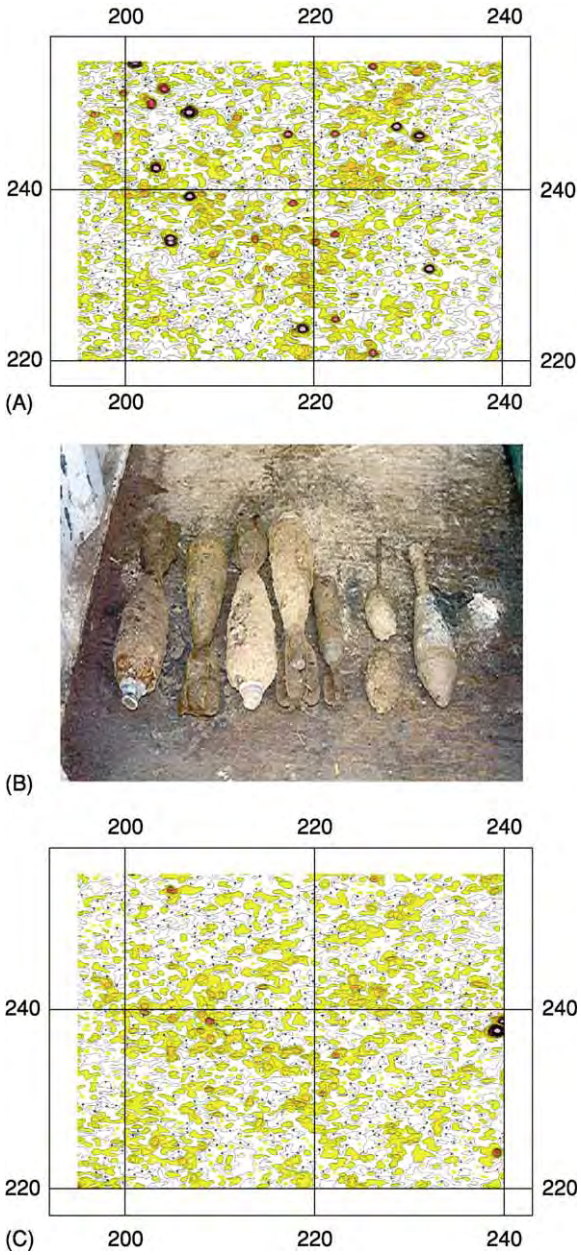


Figure 16 (A) The results of scanning an area using gradient magnetic profiling to detect ferrous ordnance items. (B) The located items were removed by trained EOC (Explosive Ordnance Clearance) personnel, and (C) the area was rescanned to prove that all items had been removed.

Table 4 Geophysical methods used in non destructive testing (after McDowell *et al.*, 2002)

Material	*Method*	*Application*
Concrete	Ultrasonic pulse velocity, GPR	Measure strength, voidage, reinforcing, cracking, cover, and fire damage
Metal	Ultrasonic pulse velocity	Measure thickness and cracking
Walls and roofs	Thermography, GPR, electromagnetic scanners, resistance/capacitance	Cavity insulation, wall ties, cladding fixings, moisture penetration, roof leaks
Buried objects	GPR, Cable Avoidance tool (CAT), magnetometer	Location of services and foundations, archaeological remains, buried objects
Machinery	Vibration metres, thermography, sound metres	Worn bearings, overheating (electrical)

Permanent geomembrane electrical leak location systems, comprising a grid of electrodes below or on either side of the geomembrane, allow continuous monitoring of the integrity of the liner throughout its lifetime. These systems can also be employed to create two- or three-dimensional resistivity models for the strata beneath the containment in order to monitor the spread of pollutants in the event of a breach.

Where access to a structure is unavailable or impossible owing to the hazardous nature of the contents, the structure can be monitored remotely using techniques such as electrical resistance tomography and excitation of the mass. Unlike electrical leak location, these methods also enable the detection of leaks from conductive containment structures such as steel underground storage tanks. Both methods involve the use of an array of electrodes placed around the structure, with excitation of the mass additionally employing an electrode placed within the containment.

Investigations into breaches in containment structures such as dams, locks, and canals can involve a wide variety of geophysical techniques, depending on the nature of the problem. Methods such as GPR and sonic testing can be used to identify voiding within and behind concrete and brick structures, whilst resistivity imaging might be used to map changes in electrical resistivity resulting from the outflow of water from a dam or canal. The self-potential method involves measuring low-amplitude voltages that result from the movement of fluids through soil and rock and can be used to locate seepage pathways within embankments and retaining walls.

Buried Assets

The inaccuracy and incompleteness of site plans showing services is infamous. Failing to confirm the location of underground services and other assets, such as underground storage tanks, can have serious cost and safety implications for a project. In the UK, the requirement to locate services prior to the commencement of any intrusive work has now been included in Health and Safety legislation. Confirmation of the position and depth of existing utility lines during the site-investigation phase allows for the planned diversion of essential services and ensures that exploratory boreholes and trial pits can be safely positioned.

The detection and tracing of services presents a particular problem for geophysics owing to the range of materials and target sizes that can be encountered. For this reason it is common to approach services detection with a broad suite of both passive and active techniques. Those most commonly used techniques include continuous-wave electromagnetic profiling (Figure 17), and passive electromagnetic methods such as the cable avoidance tool, magnetic profiling, and GPR (Figure 18). The latter is the only method capable of detecting non-metallic utility lines such as high-pressure water and gas mains.

In order to ensure optimum detection of linear targets such as services it may be necessary to undertake the survey in two orthogonal directions unless the majority of the targets are known to trend in a particular direction. This is particularly important for electromagnetic profiling and GPR, where the target response is highly dependent on the relative orientation of the instrument and target.

Where access to a particular pipe is available tracing may be simplified by using a pipe sonde. This comprises an electromagnetic transmitter mounted in a cylindrical housing that can be fed into the pipeline through an access cover or outfall. The signal from the transmitter is detected on the surface using a suitable hand-held receiver.

It is important to note that the detectability of services is dependant on burial setting and on the material type of the services being mapped. For example, an inaccessible concrete waste pipe 30 cm in diameter buried more than 2 m deep in clayey soils could be undetectable with any geophysical method. Services running beneath reinforced concrete slabs may be detectable only within 0.5 m of the surface with GPR as the presence of reinforcing will preclude

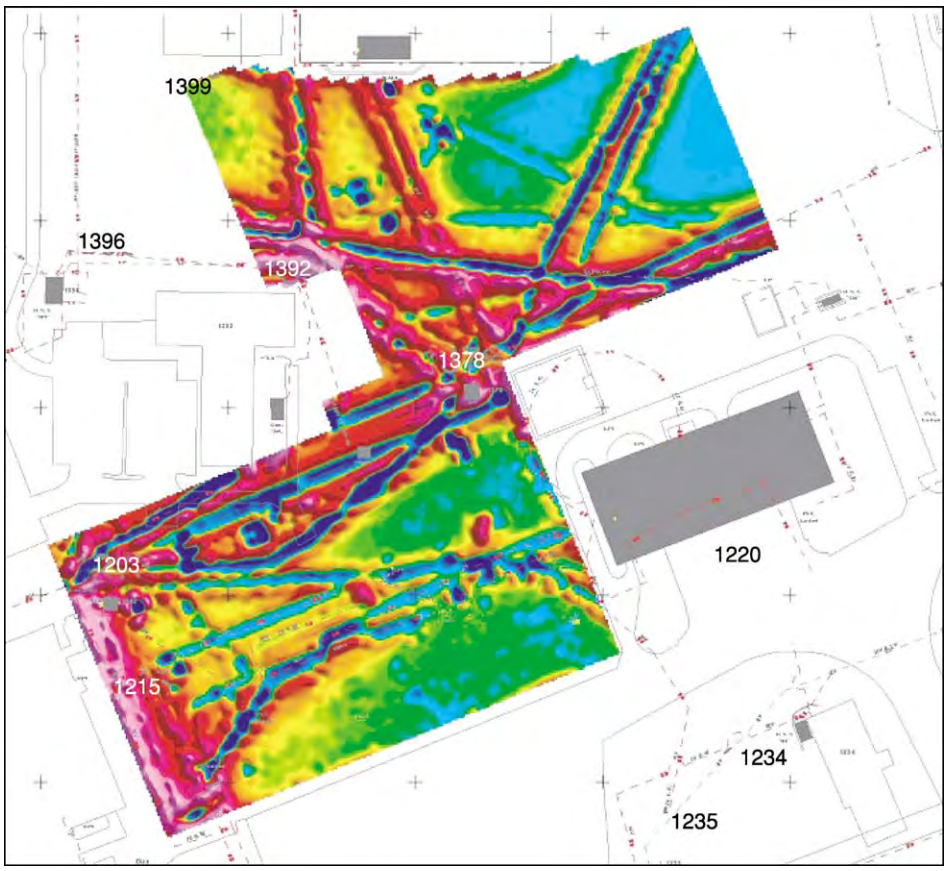


Figure 17 Electromagnetic profiling survey to map the location of services at an air base in the UK.

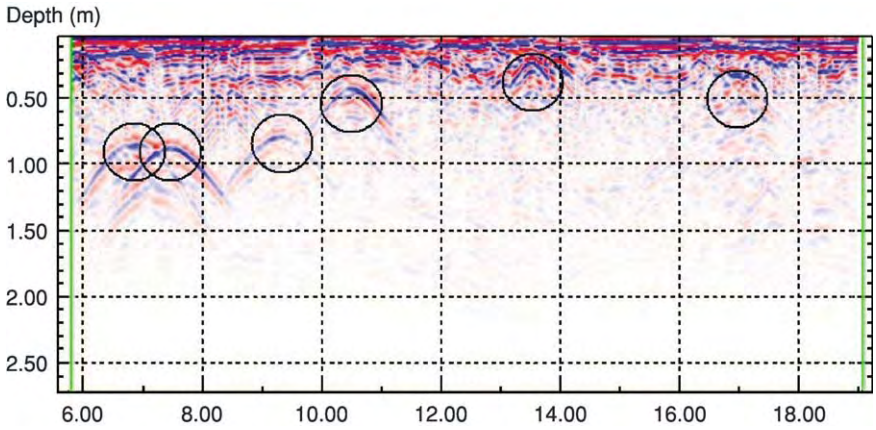


Figure 18 The location of buried utilities using GPR.

the use of electromagnetic methods. Exaggerated claims that '95% of all buried services will be mapped' should be weighed against the limitations of geophysical methods in general.

Estimating the depth to the top of a target is possible, although the degree of accuracy varies widely. Cable-avoidance tools provide a quick method of determining approximate depths on site, although these are generally accurate only to within ±20%. Advanced processing of magnetic data using techniques such as Euler Deconvolution or 2.5D inverse and forward modelling enables estimates of depths to within ±15–20%. In the case of GPR, calculation of target depths is dependent on determining the average velocity of the signal through the overlying materials.

See Also

Analytical Methods: Gravity. **Engineering Geology:** Seismology; Natural and Anthropogenic Geohazards; Site and Ground Investigation; Site Classification. **Military Geology**. **Rock Mechanics**. **Seismic Surveys**. **Soil Mechanics**.

Further Reading

Environmental and Engineering Geophysical Society. http://www.eegs.org/whatis/

Health and Safety Executive (UK) (2000) *Avoiding Danger from Underground Services*. Health and Safety Executive, London.

McCann DM, Eddleston M, Fenning PJ, and Reeves GM (eds.) (1997) *Modern Geophysics in Engineering Geology*. Engineering Geology Special Publication 12. London: Geological Society.

McDowell PW, Barker RD, and Butcher AP (eds.) (2002) *Geophysics in Engineering Investigations*. Engineering Geology Special Publication 19. London: CIRIA.

Milsom J (2003) *Field Geophysics*, 3rd edn., The Geological Field Guide Series. Chichester: John Wiley & Sons.

Reynolds JM (1997) *An Introduction to Applied and Environmental Geophysics*. Chichester: John Wiley & Sons.

Telford WM, Geldart LP, and Sheriff RE (1990) *Applied Geophysics*, 2nd edn. Cambridge: Cambridge University Press.

Zeltica. Geophysical Method Descriptons. http://www.geophysics.co.uk/methods.html

Seismology

J J Bommer, Imperial College London, London, UK
D M Boore, United States Geological Survey, Menlo Park, CA, USA

Introduction

Engineering seismology is an integral part of earthquake engineering, a specialized branch of civil engineering concerned with the protection of the built environment against the potentially destructive effects of earthquakes. The objective of earthquake engineering can be stated as the reduction or mitigation of seismic risk, which is understood as the possibility of losses – human, social or economic – being caused by earthquakes. Seismic risk exists because of the convolution of three factors: seismic hazard, exposure, and vulnerability. Seismic hazard refers to the effects of earthquakes that can cause damage in the built environment, such as the primary effects of ground shaking or ground rupture or secondary effects such as soil liquefaction or landslides. Exposure refers to the population, buildings, installations, and infrastructure encountered at the location where earthquake effects could occur. Vulnerability represents the likelihood of damage being sustained by a structure when it is exposed to a particular earthquake effect.

Earthquake Hazards and Seismic Risk

Risk can be reduced in two main ways, the first being to avoid exposure where seismic hazard is high. However, human settlement, the construction of industrial facilities, and the routing of lifelines such as roads, bridges, pipelines, telecommunications, and energy distribution systems are often governed by other factors that are of greater importance. Indeed, for lifelines, which may have total lengths of hundreds of kilometres, it will often be impossible to avoid seismic hazards by relocation. Millions of people are living in areas of the world where there is appreciable seismic hazard. The key to mitigating seismic risk therefore lies in control of vulnerability in the built environment, by designing and building structures and facilities with sufficient resistance to withstand the effects of earthquakes; this is the essence of earthquake engineering.

This is not to say, however, that the aim is to construct an earthquake-proof built environment that will suffer no damage in the case of a strong earthquake. The cost of such levels of protection would be extremely high, and, moreover, it might mean protecting the built environment against events that may not occur within the useful life of a particular building. It is generally not possible to justify such investments, especially when there are many competing demands on resources. Objectives of earthquake-resistant design are often stated in terms that relate different performance objectives to different levels of earthquake motion, such as no damage being sustained due to mild levels of shaking that may occur frequently, damage being limited to non-structural elements or to easily repairable levels in structural elements in moderate shaking that occurs occasionally, and collapse being avoided under severe

ground shaking that is only expected to occur rarely. These objectives are adjusted according to the consequences of damage to the structure; for rare occurrences of intense shaking, the performance target for a single-family dwelling will be to avoid collapse of structural elements, so that the occupants may escape from the building without injury. For a hospital or fire station, the performance target will be to remain fully operational under the same level of shaking, because the services provided will be particularly important in the aftermath of an earthquake. For a nuclear power plant or radioactive waste repository, the performance objective will be to maintain structural integrity even under extreme levels of shaking that may be expected to occur very rarely.

Once planners and developers have taken decisions regarding the location of civil engineering projects, or once people have begun to settle in an area, the level of exposure is determined. In general, seismic hazard cannot be altered, hence the key to mitigating seismic risk levels lies in the reduction of vulnerability or, stated another way, depends on the provision of earthquake resistance. In order to provide effective earthquake protection, the civil engineer requires quantitative information on the nature and likelihood of the expected earthquake hazards. As already indicated, this may mean defining the hazard, not in terms of the effects that a single earthquake event may produce at the site, but as a synthesis of the potential effects of many possible earthquake scenarios and quantitative definitions of the particular effects that may be expected to occur with different specified frequencies. This is the essence of engineering seismology: to provide quantitative assessments of earthquake hazards.

Earthquake generation creates a number of effects that are potentially threatening to the built environment (Figure 1). These effects are earthquake hazards, and engineering seismology, in the broadest sense, is concerned with assessing the likelihood and characteristics of each of these hazards and their possible impact in a given region or at a given site of interest. Earthquakes are caused by sudden rupture on geological faults, the slip on the fault rupture ranging from a few to tens of centimetres for moderate earthquakes (magnitudes from 5 to about 6.5) to many metres for large events (*see* **Tectonics:** Earthquakes). In those cases where the fault extends to the ground surface (which is often not the case), the relative displacement of the two sides of the fault presents an obvious hazard to any structure crossing the fault trace. In the 17 August 1999 Kocaeli earthquake in Turkey, several hundred houses and at least one industrial facility were severely damaged by the surface deformations associated with the slip on the

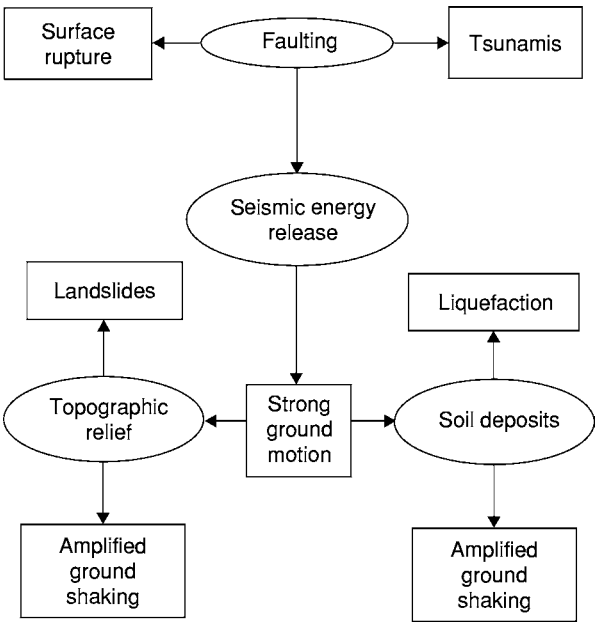


Figure 1 Potentially destructive effects of earthquakes, showing the elements of the earthquake generation process and the natural environment (ovals) and the resulting seismic hazards (rectangles).

North Anatolian fault. However, buildings must be situated within a few metres of the fault trace for surface rupture to be a hazard, whereas ground shaking effects can present a serious hazard even at tens of kilometres from the fault trace; hence, even if the fault rupture is tens or even hundreds of kilometres in length, the relative importance of surface rupture hazard is low. An exception to this is where lifelines, particularly pipelines, cross fault traces, although identification and quantification of the hazard allows the design to take into account the expected slip in the event of an earthquake: in the vicinity of the Denali Fault, the trans-Alaskan oil pipeline is mounted on sleepers that allow lateral movement, which permitted the pipeline to remain functional despite more than 4 m of lateral displacement on the fault in the magnitude 7.9 earthquake of 3 November 2002.

Surface fault ruptures in the ocean floor can give rise to tsunamis, seismic sea waves that are generated by the sudden displacement of the surface of the sea and travel with very high speeds. As the waves approach the shore and the water depth decreases, the amplitude of the waves increases to maintain the momentum, reaching heights of up to 30 m. When the waves impact on low-lying coastal areas, the destruction can be almost total.

All of the other hazards generated by earthquakes are directly related to the shaking of the ground caused by the passage of seismic waves. The rapid movement of building foundations during an earthquake

generates inertial loads that can lead to damage and collapse, which is the cause of the vast majority of fatalities due to earthquakes. For this reason, the main focus of engineering seismology, and also of this article, is the assessment of the hazard of ground shaking. Earthquake ground motion can be amplified by features of the natural environment, increasing the hazard to the built environment. Topographic features such as ridges can cause amplification of the shaking, and soft soil deposits also tend to increase the amplitude of the shaking with respect to rock sites. At the same time, the shaking can induce secondary geotechnical hazards by causing failure of the ground. In mountainous or hilly areas, earthquakes frequently trigger landslides, which can significantly compound the losses: the 6 March 1987 earthquake in Ecuador triggered landslides that interrupted a 40-km segment of the pipeline carrying oil from the production fields in the Amazon basin to the coast, thereby cutting one of the major exports of the country; the earthquake that struck El Salvador on 13 January 2001 killed about 850 people, and nearly all of them were buried by landslides. In areas where saturated sandy soils are encountered, the ground shaking can induce liquefaction (*see* **Engineering Geology:** Liquefaction) through the generation of high pore-water pressures, leading to reduced effective stress and a significant loss of shear strength, which in turns leads to the sinking of buildings into the ground and lateral spreading on river banks and along coasts. Extensive damage in the 17 January 1994 Kobe earthquake was caused by liquefaction of reclaimed land, leaving Japan's second port out of operation for 3 years.

The assessment of landslide and liquefaction hazard involves evaluating the susceptibility of slopes and soil deposits, and determining the expected level of earthquake ground motion. The basis for earthquake-resistant design of buildings and bridges also requires quantitative assessment of the ground motion that may be expected at the location of the project during its design life. Seismic hazard assessment in terms of strong ground motion is the activity that defines engineering seismology.

Measuring Earthquake Ground Motion

The measurement of seismic waves is fundamental to seismology. Earthquake locations and magnitudes are determined from recordings on sensitive instruments (called seismographs) installed throughout the world, detecting imperceptible motions of waves generated by events occurring hundreds or even thousands of kilometres away. Engineering seismology deals with ground motions sufficiently close to the causative rupture to be strong enough to present a threat to engineering structures. There are cases in which destructive motions have occurred at significant distances from the earthquake source, generally as the result of amplification of the motions by very soft soil deposits, such as in the San Francisco Marina District during the 18 October 1989 Loma Prieta earthquake, and even more spectacularly in Mexico City during the 19 September 1985 Michoacan earthquake, almost 400 km from the earthquake source. In general, however, the realm of interest of engineering seismology is limited to a few tens of kilometres from the earthquake source, perhaps extending to 100 km or a little more for the largest magnitude events.

Seismographs specifically designed for measuring the strong ground motion near the source of an earthquake are called accelerographs, and the records that they produce are accelerograms. The first accelerographs were installed in California in 1932, almost four decades after the first seismographs, the delay being caused by the challenge of constructing instruments that were simultaneously sensitive enough to produce accurate records of the ground acceleration while being of sufficient robustness to withstand the shaking without damage.

Prior to the development of the first accelerographs, the only way to quantify earthquake shaking was through the use of intensity scales, which provide an index reflecting the strength of ground shaking at a particular location during an earthquake. The index is evaluated on the basis of observations of how people, objects, and buildings respond to the shaking (Table 1). Some intensity scales also include the response of the ground with indicators such as slumping, ground cracking, and landslides, but these phenomena are generally considered to be dependent on too many variables to be reliable indicators of the strength of ground shaking. At the lower intensity degrees, the most important indicators are related to human perception of the shaking, whereas at the higher levels, the assessment is based primarily on the damage sustained by different classes of buildings. A common misconception is that intensity is a measure of damage, whereas it is in fact a measure of the strength of ground motion inferred from building damage, whence a single intensity degree can correspond to severe damage in vulnerable rural dwellings and minor damage in engineered constructions. The most widely used intensity scales, both of which have 12 degrees and which are broadly equivalent, are the Modified Mercalli (MM), used in the Americas, and the 1998 European Macroseismic Scale (EMS-98), which has replaced the Medvedev–Sponheuer–Karnik (MSK) scale.

Table 1 Summary of the 1998 European Macroseismic Scale

Intensity	*Definition*	*Effects on people and buildings*[a]
I	Not felt	Not felt; detected only by sensitive instruments
II	Scarcely felt	Felt by very few people, mainly those who are in particularly favourable conditions at rest and indoors
III	Weak	Felt by a few people, mostly those at rest
IV	Largely observed	Felt by many people indoors and very few outdoors; a few people are awakened by the shaking
V	Strong	Felt by most people indoors and a few outdoors; a few people are frightened and many who are sleeping are awakened by the shaking Fine cracks in a few of the most vulnerable types of buildings, such as adobe and unreinforced masonry
VI	Slightly damaging	Shaking felt by nearly everyone indoors and by many outdoors; a few people lose their balance Minor cracks in many vulnerable buildings and in a few poor quality RC structures; a few of the most vulnerable buildings have cracks in walls and spalling of fairly large pieces of plaster
VII	Damaging	Most people are frightened by the shaking, and many have difficulty standing, especially those on upper storeys of buildings Many adobe and unreinforced masonry buildings sustain large cracks and damage to roofs and chimneys; some will have serious failures in walls and partial collapse of roofs and floors; minor cracks in RC structures with some earthquake resistant design, more significant cracks in poor quality RC structures
VIII	Heavily damaged	Many find it difficult to stand Many vulnerable buildings experience partial collapse and some collapse completely; large and extensive cracks in poor quality RC buildings and many cracks in RC buildings with some degree of earthquake resistant design
IX	Destructive	General panic (this is the highest degree of intensity that can be assessed from human response) General and extensive damage in vulnerable buildings; cracks in columns, beams, and partition walls of RC buildings, with some of those of poor quality suffering heavy structural damage and partial collapse; non structural damage in RC structures with high level of earthquake resistant design
X	Very destructive	Partial or total collapse in nearly all vulnerable buildings; extensive structural damage in RC and steel structures
XI	Devastating	Extensive damage and widespread collapse in nearly all building types; very rarely observed, if ever
XII	Completely devastating	Cataclysmic damage; has never been observed and is probably not physically realizable

[a]RC, Reinforced concrete.

A very useful picture of the strength and distribution of ground shaking in an earthquake can be obtained by mapping intensity observations. The modal value is assigned to each given location, such as a village, from all the individual point observations gathered, and then lines called isoseismals are drawn to enclose areas where the intensity reached the same degree (**Figure 2**). Such isoseismal maps have many applications, one of the most important of which is to establish correlations between earthquake size (magnitude) and the area enclosed by isoseismals. These empirical relations can then be used to estimate the magnitude of historical earthquakes that occurred before the advent of the global seismograph network at the end of nineteenth century, but for which it is possible to compile isoseismal maps from written accounts of the earthquake effects. For engineering design and analysis, however, intensity values are of limited use because they cannot be reliably translated into numerical values related to the acceleration or displacement of the ground. Indeed, intensities are usually expressed in Roman numerals precisely to reinforce the idea that they are broad indices rather than numerical measurements. This is why accelerograms are invaluable to earthquake engineering, providing detailed measurements of the actual movement of the ground during strong earthquake-induced shaking.

Accelerograms generally consist of three mutually perpendicular components of motion, two horizontal and one vertical, registering the ground acceleration against time (**Figure 3**). The first generation of accelerographs produced analogue records on paper or film, which had to be digitized in order to be able to perform numerical analyses. The analogue recording and the digitization process both introduce noise into the signal; this noise then requires processing of the time-series to improve the signal-to-noise ratio, particularly at long periods. This processing generally involves the application of digital filters. The current models of accelerographs record digitally, thus bypassing the time-consuming and troublesome process of digitization, and producing records with much lower noise contamination. Another advantage of

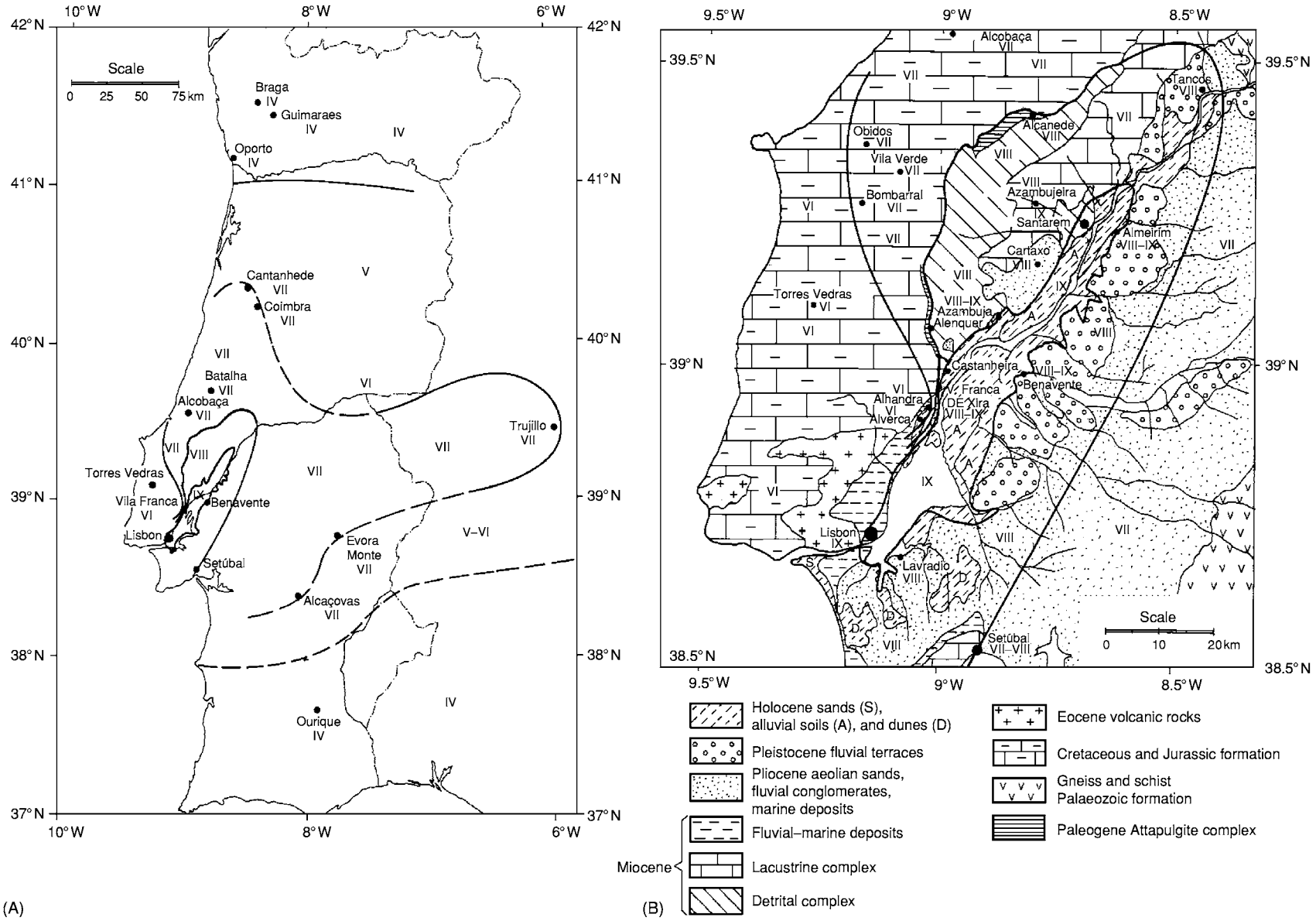


Figure 2 (A) Regional isoseismal map of the Lisbon earthquake of January 1531. (B) Isoseismal map for the epicentral area of the Lisbon earthquake of January 1531, superimposed on a map of the local geology. Reprinted with permission from Justo JL and Salwa C (1998) The 1531 Lisbon earthquake. *Bulletin of the Seismological Society of America* 88(2): 319 328. © Seismological Society of America.

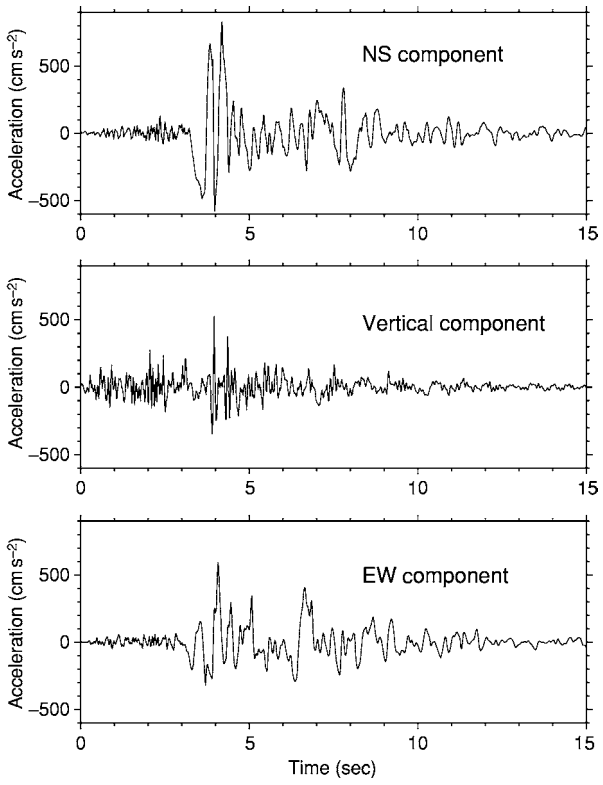


Figure 3 Three component accelerogram (NS, north south; EW, east west) digitized from an analogue recording at the Sylmar Hospital free field site during the Northridge, California, earthquake of January 1994.

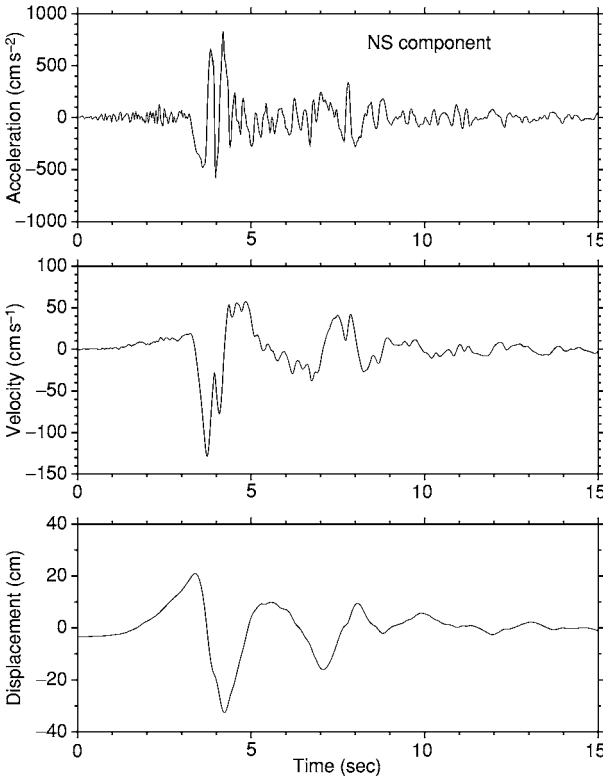


Figure 4 Filtered acceleration, velocity, and displacement time series from the north south (NS) horizontal component of the Sylmar Hospital free field site record (1994 earthquake, Northridge, California).

digital accelerographs is that they record continuously on reusable media, whereas optical–mechanical instruments remain on standby until triggered by a minimum-threshold acceleration, thus missing the very first-wave arrivals. Although the motions missed by an analogue instrument are generally very small, the advantage with a digital record is that the boundary conditions of initial velocity and displacement are known with greater confidence; the time-series of velocity and displacement are obtained by simple integration of the acceleration time-series, modified as needed by filtering and/or baseline correction to account for long-period noise (Figure 4).

Characterizing Strong Ground Motion

A number of parameters are used to characterize the nature of the earthquake ground motion captured on an accelerogram, although in isolation, no single parameter is able to represent fully all of the important features. The simplest and most widely used parameter is the peak ground acceleration (PGA), which is simply the largest absolute value of acceleration in the time-series; the horizontal PGA is generally treated separately from the peak vertical motion. Similarly the values of the peak ground velocity (PGV) and peak ground displacement (PGD) are also used to characterize the motion, although the latter is difficult to determine reliably from an accelerogram because of the influence of the unknown baseline on the records and the double integration of the long-period noise in the record. These parameters are particularly poor for characterizing the overall nature of the motion because they reflect only the amplitude of a single isolated peak (Figure 5). In engineering seismology, unlike geophysics, accelerations are generally expressed in units of g, the acceleration due to gravity (9.81 m s ^{2}).

Another important characteristic of the ground motion is the duration of shaking, particularly of the portion of the accelerogram where the motion is intense. There are many different ways in which the duration of the motion can be measured from an accelerogram, one of the more commonly used definitions being the total interval between the first and last excursions of a specified threshold, such as 0.05 g. Duration of shaking is particularly important in

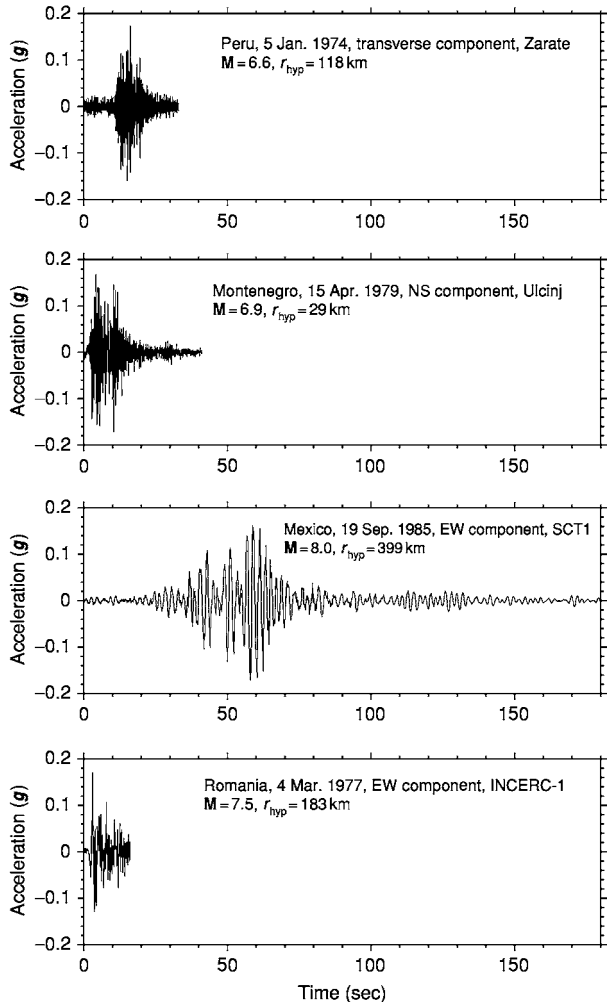


Figure 5 Four horizontal accelerogram components with nearly the same peak ground acceleration. **M**, Moment magnitude; r_{hyp}, distance from the hypocentre.

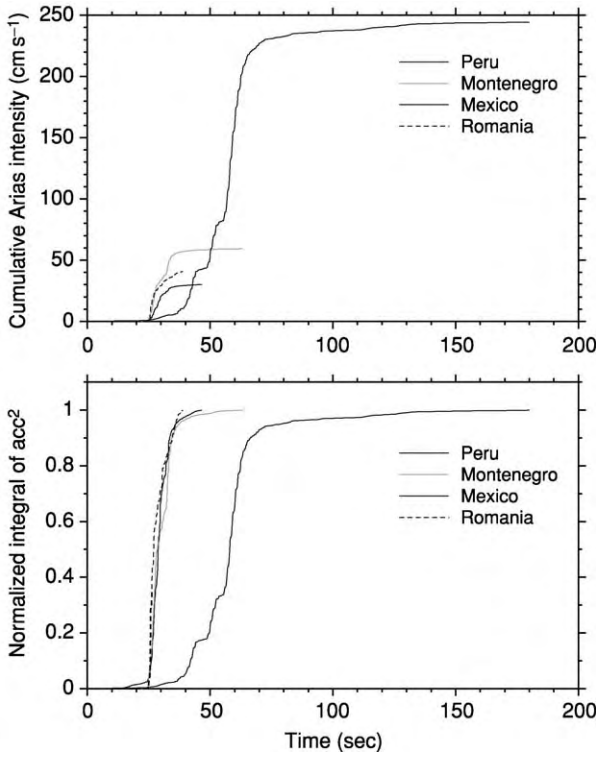


Figure 6 Husid plots for the four acceleration time series shown in Figure 5, showing the build up of Arias intensity with time. The upper plot shows the absolute values of Arias intensity; in the lower plot, the curves are normalized to the maximum value attained.

the assessment of liquefaction hazard because the build-up of pore water pressures is controlled by the number of cycles of motion as well as by the amplitude of the motion.

A parameter that measures the energy content of the ground shaking is the Arias intensity, which is the integral over time of the square of the acceleration. A plot of the build-up of Arias intensity with time is known as a Husid plot (Figure 6) and it serves to identify the interval over which the majority of the energy is imparted. The root-mean-square acceleration (a_{rms}) is the equivalent constant level of acceleration over any specified interval of the accelerogram; the a_{rms} is also the square root of the gradient of the Husid plot over the same interval. Arias intensity has been found to be a useful parameter to define thresholds of shaking that trigger landslides.

For engineering purposes, the most important representation of earthquake ground motion is the response spectrum, which is a graph of the maximum response experienced by a series of single-degree-of-freedom (SDOF) oscillators when subjected to the acceleration time-series at their bases (Figure 7). The dynamic characteristics of a SDOF system, such as an idealized inverted pendulum, are fully described by its natural period of free vibration and its damping, usually modelled as an equivalent viscous damping and expressed as a proportion of the critical damping that returns a displaced SDOF system to rest without vibrations. In earthquake engineering, the default value generally used is 5% of critical damping, the nominal level of damping in a reinforced concrete structure. The response can be measured in terms of the absolute acceleration of the mass of the system, or in terms of its velocity or displacement relative to the base. The most widely used spectrum currently is the acceleration response, because the acceleration at the natural period of the structure can be multiplied by the mass of the building to estimate the lateral force exerted on the structure by the earthquake shaking. The natural period of vibration of a building can be very approximately estimated, in seconds, as the number of storeys

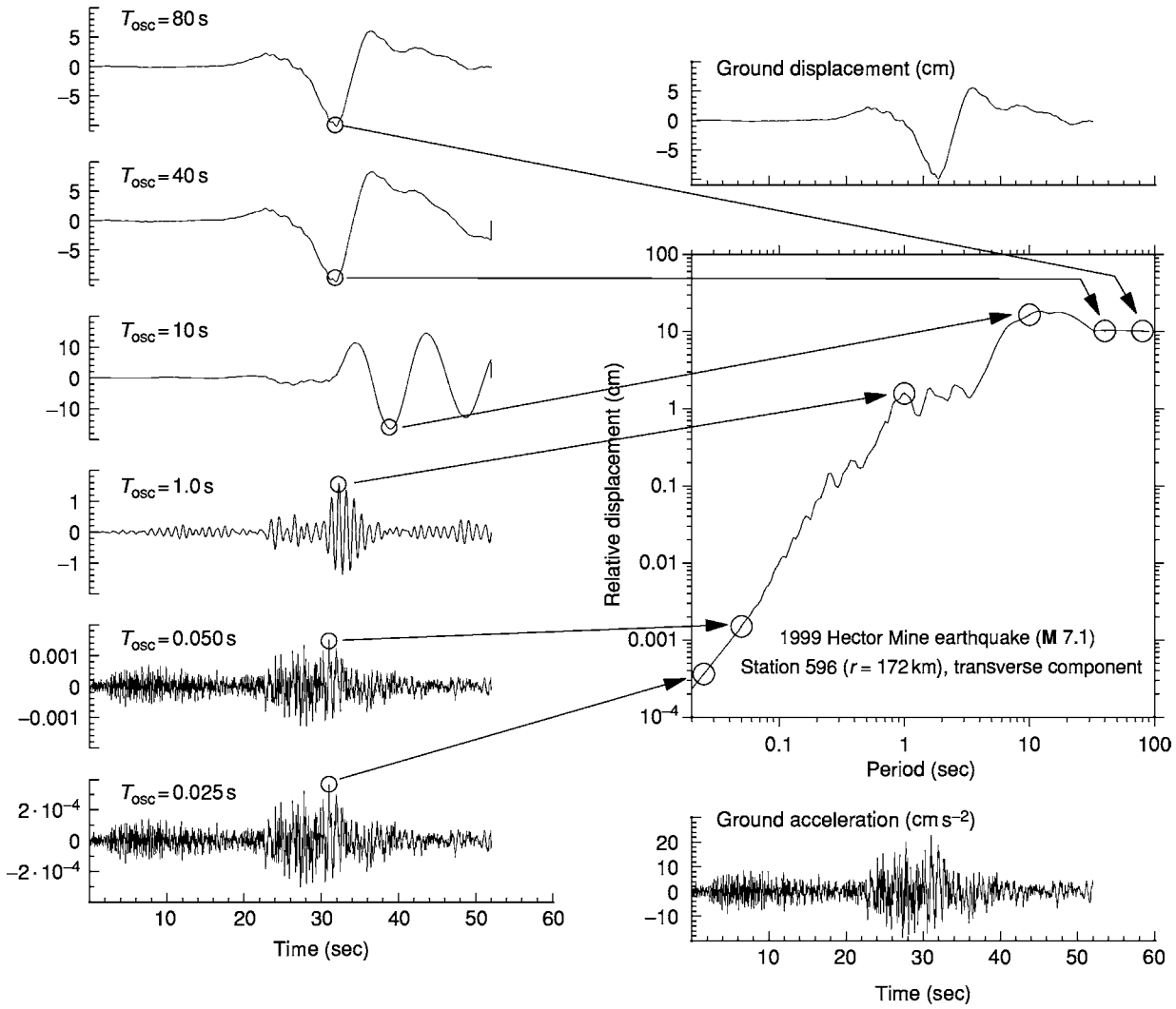


Figure 7 Illustration of the concept of the response spectrum. The trace in the lower right hand corner is the ground acceleration, and the traces on the left hand side are the displacement response time series for simple oscillators with different natural periods of vibration (T_{osc}). The maximum displacement of each oscillator is plotted against its natural period to construct the response spectrum of relative displacement. The lowest two traces in the figure illustrate that for short period oscillators, the response mimics the ground acceleration, whereas for long period oscillators, the response imitates the ground displacement (shown in the upper right hand corner). **M**, Moment magnitude; r, distance to fault rupture.

divided by 10. The acceleration response spectrum intersects the vertical axis at PGA (Figure 8). PGA as a ground-motion parameter has many shortcomings, including a very poor correlation with structural damage, but one of the main reasons that it is persistently used is the fact that it essentially defines the anchor point for the acceleration response spectrum. The relationship between the natural period of vibration of the building and the frequency content of the ground motion is a critical factor in determining the impact of earthquake shaking on buildings.

In recent years, there has been a gradual tendency to move away from the use of acceleration response spectra in force-based seismic design towards displacement-based approaches, because structural damage is much more closely related to displacements than to the forces that are imposed on buildings very briefly during ground shaking. Techniques for assessing the seismic capacity of existing buildings have already adopted displacement-based approaches, giving rise to greater interest in displacement response spectra (Figure 9).

Prediction of Earthquake Ground Motion

Recordings of ground motions in previous earthquakes are used to derive empirical equations that may be used to estimate values of particular ground-motion parameters for future earthquake

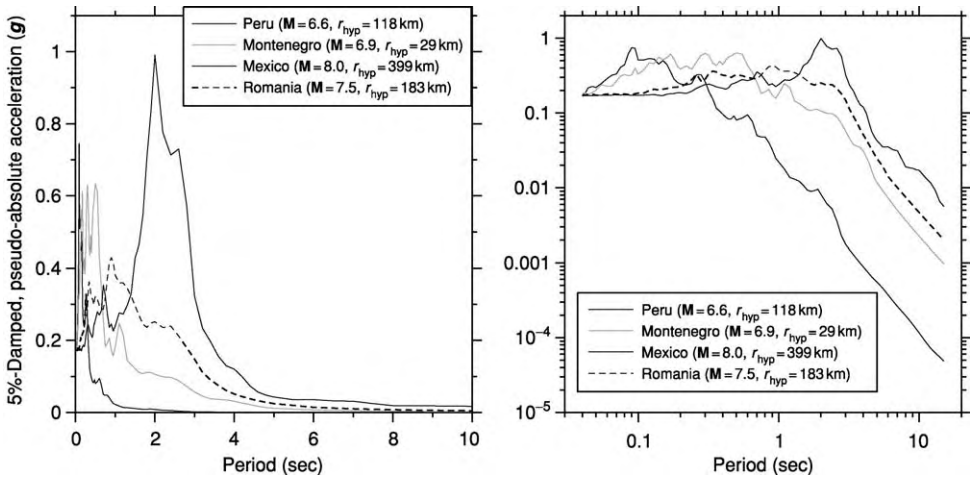


Figure 8 Acceleration response spectra (5% damped) of the four accelerograms shown in **Figure 5**; the plots are identical except that the one on the left uses linear axes and the one on the right uses logarithmic axes. Each type of presentation is useful for viewing particular aspects of the motion, depending on whether the interest is primarily at short or long periods of response. **M**, Moment magnitude; r_{hyp}, distance from the hypocentre.

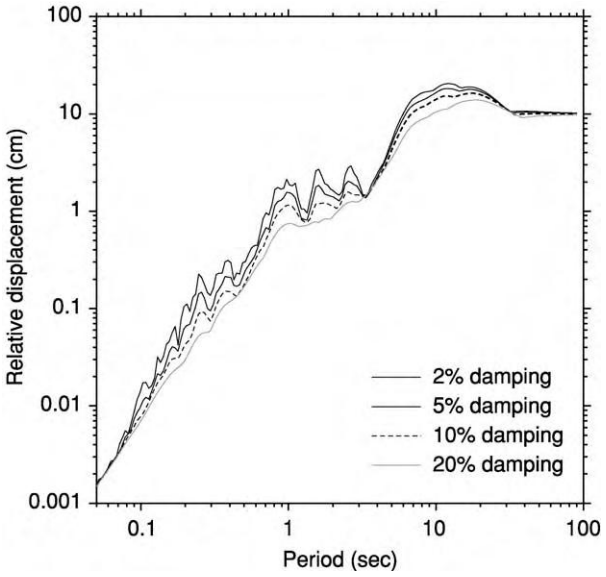


Figure 9 Displacement response spectra of a single accelero gram from the 1999 California Hector Mine earthquake (moment magnitude 7.1; station 596; r = 172 km, r is the distance from the fault rupture in km, transverse component), plotted for different levels of damping . The spectra all converge to the value of peak ground displacement at very long periods.

scenarios. Such empirical equations have been derived for a variety of parameters, but the most abundant are those for predicting PGA and ordinates of response spectra. The equations are often referred to as attenuation relationships, but this is a misnomer because the equations describe both the attenuation (decay) of the amplitudes with distance from the earthquake source and the scaling (increase) of the amplitudes with earthquake magnitude. These two parameters, earthquake magnitude and source-to-site distance, are always included in ground-motion prediction equations (**Figure 10**). The equations are simple models for a very complex phenomenon and as such there is generally a large amount of scatter about the fitted curve (**Figure 11**). The residuals of the logarithmic values of the observed data points are generally found to follow a normal or Gaussian distribution about the mean, and hence the scatter can be measured by the standard deviation. Predictions of PGA at the 84-percentile level (i.e., one standard deviation above the mean value) will generally be as much as 80% higher than the median predictions (**Figure 12**).

The nature of the surface geology can also exert a pronounced effect on the recorded ground motion. The presence of soft soil deposits of more than a few metres thickness will tend to amplify the ground motion as the waves propagate from the stiffer materials below to the surface (**Figures 2B and 13**). To account for this effect, site classification is generally included as a third explanatory variable in prediction equations for response spectrum ordinates, allowing spectra to be predicted for different sites (**Figure 14**). The modelling of site amplification effects in ground-motion prediction equations is generally crude, using simple site classification schemes based on the average shear-wave velocity of the upper 30 m at the site and assuming that the degree of amplification is independent of the amplitude of the input motion, whereas it is generally observed that weak motion is amplified more than strong motion; this non-linear response of soils is reflected only in a few predictive equations. For this

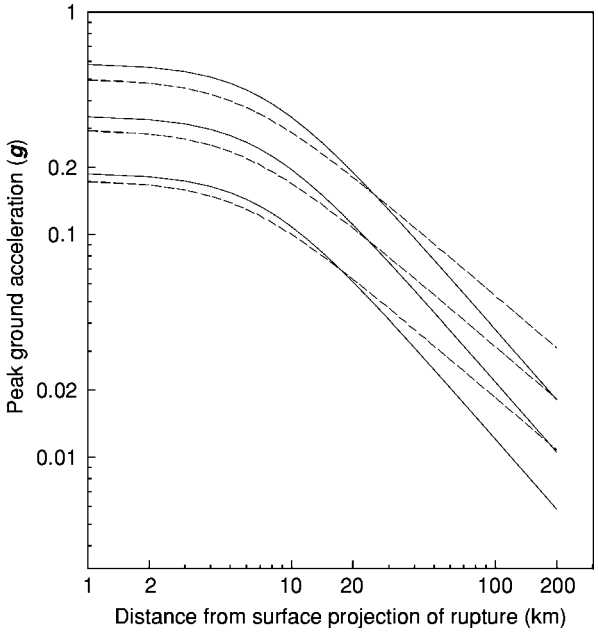


Figure 10 Predicted median values of peak ground acceleration as a function of distance for earthquakes of moment magnitude 5.5, 6.5, and 7.5, using equations derived from western North American (solid lines) and European (dashed lines) data sets. The European equation is based on surface wave magnitude, so an appropriate empirical conversion to moment magnitude has been used. The equations are based on different definitions of the horizontal component, the European equation using the larger of the two horizontal accelerations, the North American equation using their geometric mean; the former definition, on average, yields values about 1.17 times larger than the latter. Different criteria were used in the two studies for selecting records, especially at greater distances, for regression. In light of these various observations, it is not possible to draw conclusions about differences or similarities in strong ground motions between the two regions. The North American equation is from Boore DM, Joyner WB, and Fumal TE (1997) *Seismological Research Letters* 68(1): 128 153; the European equation is from Tromans IJ and Bommer JJ (2002) *Proceedings of the 12th European Conference on Earthquake Engineering*, Paper no. 394, London.

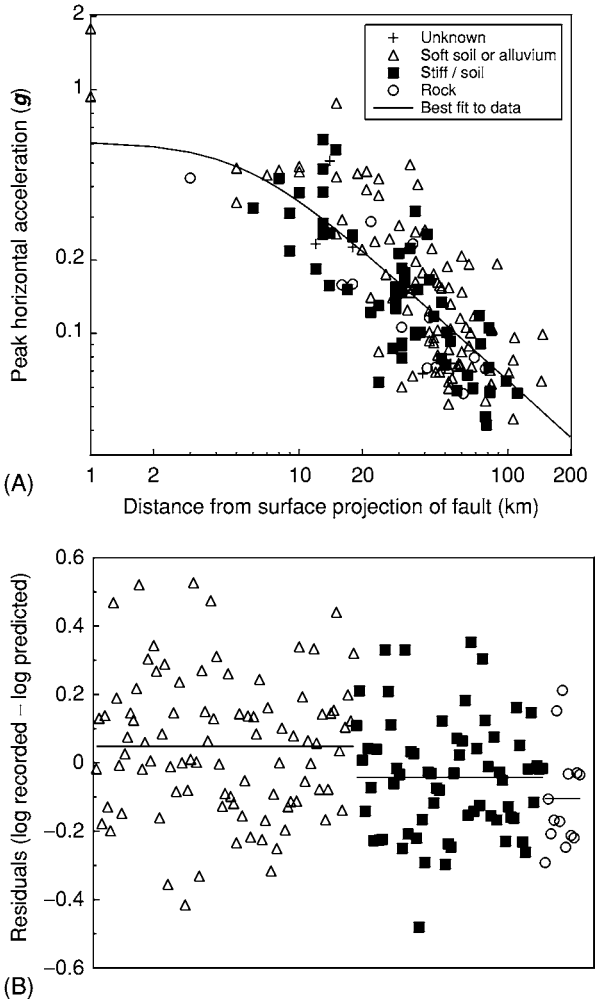


Figure 11 (A) Values of peak horizontal ground acceleration recorded in the 1994 Northridge earthquake in California, plotted as a function of distance from the earthquake source, with an indication of the surface geology at the recording site. The solid line is the result of a regression on the data using a typical model employed in ground motion prediction equations. (B) Residuals group by site class; the horizontal lines show the mean of the residuals in each class, indicating that there is a tendency for stronger motions on softer ground. However, the differences between the mean lines are small compared with the overall dispersion, thus inclusion of site classification in the predictive equation would result in only a modest reduction of the aleatory variability.

reason, for site-specific predictions, it is often preferred to predict the bedrock motions first and then model the dynamic response of the site separately (Figure 15).

The nature of ground shaking at a particular site during an earthquake is influenced by many factors, including the distribution and velocity of the slip on the fault rupture, the depth at which the fault rupture is located, the orientation of the fault rupture with respect to the travel path to the site, and the geological structure along the travel path and for several kilometres below the site. A situation that produces particularly destructive motions is the propagation of the fault rupture towards the site, which produces large-amplitude, high-energy pulses of velocity (Figure 16). The variable most often included in predictive equations after magnitude, distance, and site classification is the style of faulting; equations that include the rupture mechanism as an explanatory variable all predict higher amplitudes of motion from reverse-faulting earthquakes than from strike–slip events. The addition of this fourth explanatory variable, however, has an almost negligible impact on the scatter in the equations, with no appreciable reduction of the standard deviation.

Dense networks of accelerographs now exist in many countries around the world, but for many decades these were limited to a few regions such as California, Japan, Italy, Greece, and Yugoslavia. In

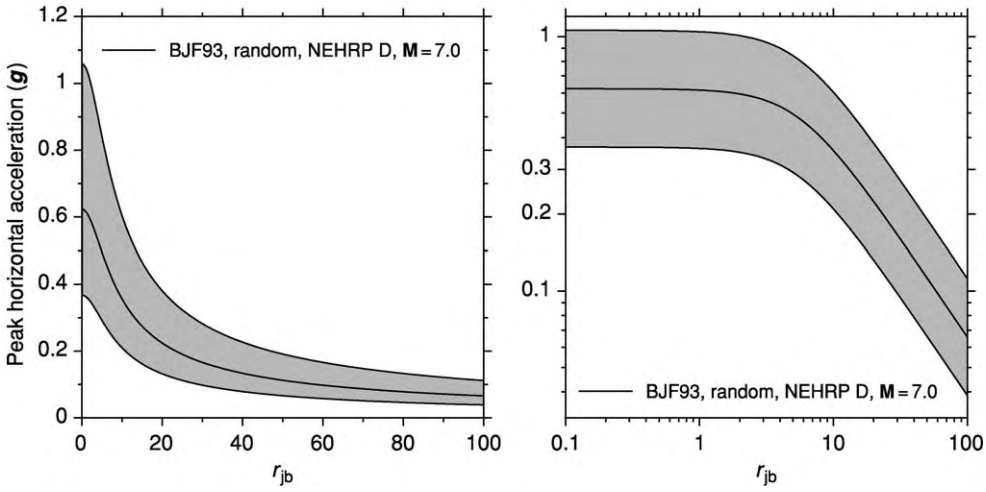


Figure 12 Predicted values of horizontal peak ground acceleration on soft soil (National Earthquake Hazards Reduction Program (NEHRP) class D) sites for an earthquake of magnitude 7 as a function of distance from the surface projection of the fault rupture (r_{jb}). The predictions are made using an equation derived from earthquakes recorded in western North America, by DM Boore, WB Joyner, and TE Fumal. The thick black line of Boore, Joyner, and Fumal (BJF93) shows the median predicted values and the grey bands indicate the range of the median multiplied and divided by 10^{σ}, where σ is the standard deviation of the logarithmic residual; on average, 68% of observations would be expected to fall within the grey area.

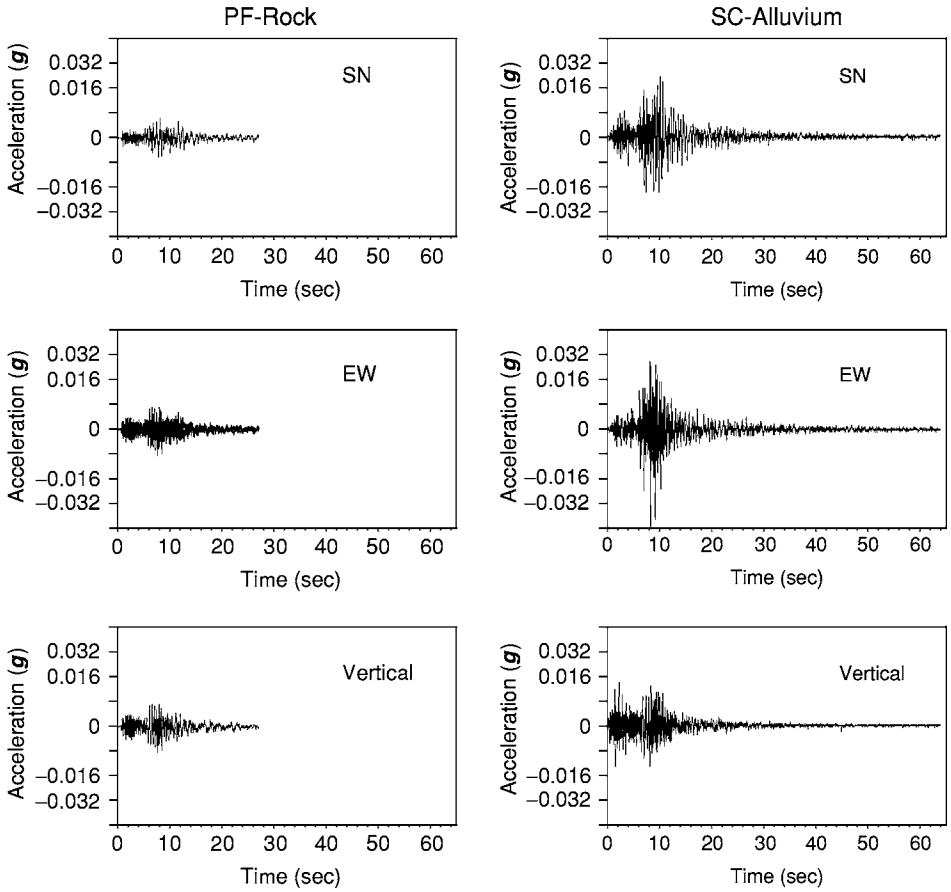


Figure 13 Accelerograms recorded on two instruments in the Gemona (Italy) array during the **M** 5.6 Bovak (Slovenia) earthquake of 12 April 1998 (**M** moment magnitude). Both instruments are 38 km from the earthquake epicentre, and are separated by about 700 m. The Piazza del Ferro station (PF) is on rock with a shear wave velocity reported as 2500 m s ^{1}, whereas the Scugelars station (SC) is on alluvium with a shear wave velocity of 500 m s ^{1}. Because the distance and azimuth of the two stations with respect to the earthquake source are almost identical, the difference between the amplitude of the two recordings is primarily due to the different surface geology at the two recording locations (SN, south north; EW, east west).

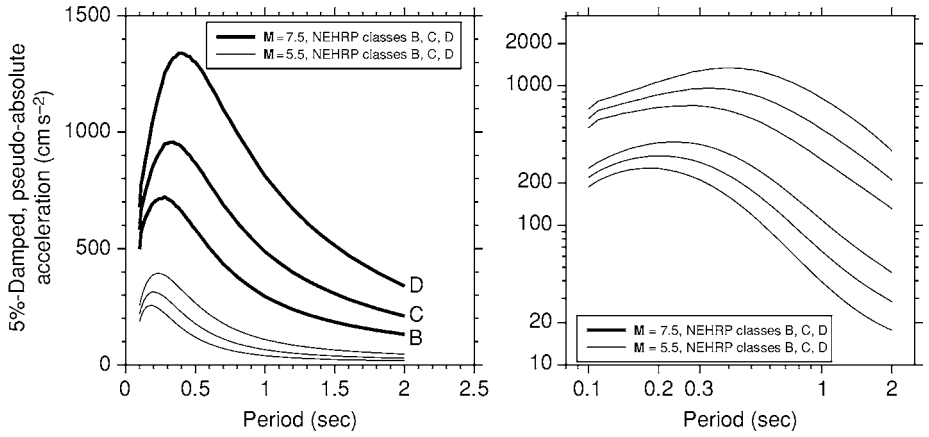


Figure 14 Predicted median ordinates of spectral acceleration (5% damped) for earthquakes of magnitudes 5.5 and 7.5, recorded at 10 km on sites classified as B, C, and D in the National Earthquake Hazards Reduction Program (NEHRP) scheme, corresponding to soft rock, stiff soil, and soft soil, respectively. The two graphs are identical except for the left hand plot being on linear axes and the right hand plot being on logarithmic axes.

other regions, the number of available accelerograms is further reduced by the relatively low frequency of felt earthquakes. Seismic hazard assessment in such regions is hampered by the absence of indigenous recordings, and a solution that is often employed is to use stochastic simulations to generate motions based on representing the ground shaking as random motion with frequency content and duration that are controlled by theoretical and empirical descriptions of seismic radiation and propagation (Figure 17).

Complete acceleration time-histories can also be generated using more sophisticated seismological models that represent the source as a finite fault rupture. These simulations usually take the form of a kinematic model of the fault rupture, with rupture velocities and slip amplitudes specified across the fault plane (although these can be specified using statistical distribution in order to be able to generate a suite of motions for a given rupture). Dynamic fault modelling, in which the stress conditions are specified and subsequent rupture is controlled by friction laws on the fault surface as well as the heterogeneous three-dimensional stress distribution, is less commonly used for the simulation of ground motions for engineering purposes because the necessary physical quantities are not known with sufficient precision and because the computational effort is much greater than with the kinematic models (which can be quite demanding if lateral variations in geology are taken into account). In the future, as computers become more powerful and the details of the geologic structure and velocity distributions are better determined, it is likely that the kinematic models will continue to be used, but the simulations will better represent the effects of wave propagation.

Seismic Hazard Assessment

Engineering seismology is often described as the link between Earth sciences and engineering, and this is most evident in seismic hazard assessment, the aim of which is to estimate the earthquake ground motions that can be expected at a particular location. Seismic hazard assessments may be performed for a number of reasons, the most common being to determine seismic loads to be considered in earthquake-resistant design of buildings. Another common application is the evaluation of the ground motions required as input to the assessment of landslide or liquefaction hazard. A rapidly expanding market for seismic hazard assessments is being created by the demand for earthquake loss models by local governments, emergency services, seismic code developers, and particularly the insurance and reinsurance industries.

There are many different ways to approach seismic hazard assessment, but all methods and approaches include two essential elements: a model for earthquake occurrence and a model for predicting ground motions from each earthquake. The starting point for building a model for earthquake occurrence, known as a seismicity model, is to first identify the locations where earthquakes will be expected to occur in the future. Data used to define seismic source zones include previous earthquake activity, the existence of geological faults, and crustal deformations. Ideally, all seismic sources should be identified as active geological faults, but this is very often not possible, hence zones are defined as areas within which future earthquakes are expected to occur (Figure 18). Some recent approaches to seismic hazard assessment dispense with source zones altogether, basing the sources of

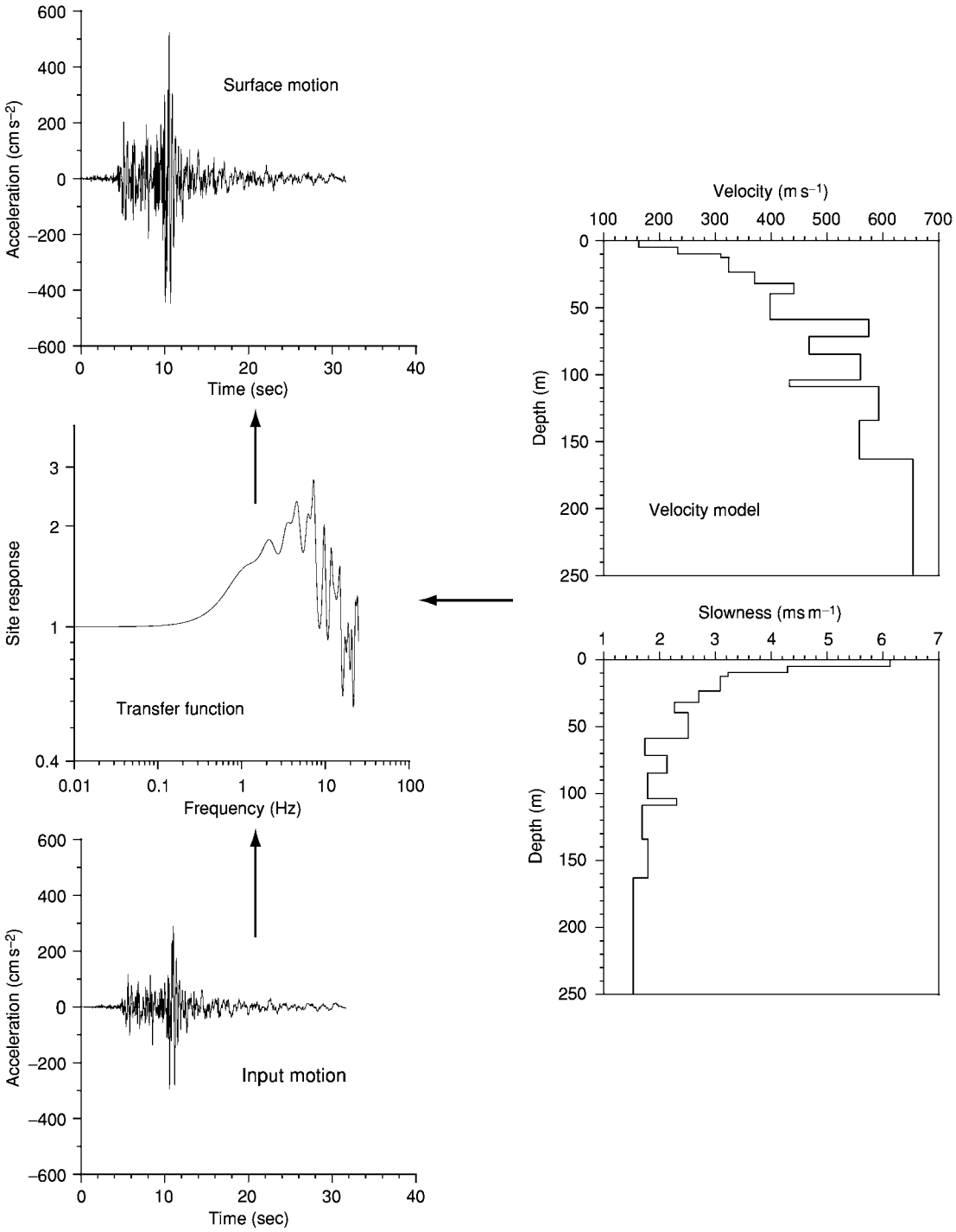


Figure 15 Illustration of site response analysis: The input motion in the underlying rock (bottom left) passes through a model of soil profile to produce the surface motions (top left). The soil characteristics are represented either by the shear wave velocity or the shear wave slowness (right); the slowness is simply the reciprocal of the shear wave velocity, but there are many reasons for using it in place of the velocity. The theoretical response of layered systems involves travel times across the layers, which are directly proportional to the slowness; the proportionality to velocity is inverse. Furthermore, to obtain a representative profile for a soil class using several boreholes, individual measures of slowness can be directly averaged, whereas it is not correct to do this for velocities. Finally, the largest influence on the surface motion is due to the differences in velocities near the surface, and these differences become more clearly apparent when the slowness profile is shown; larger velocity differences in the stiffer layers at greater depths, which can dominate a velocity profile, are less important.

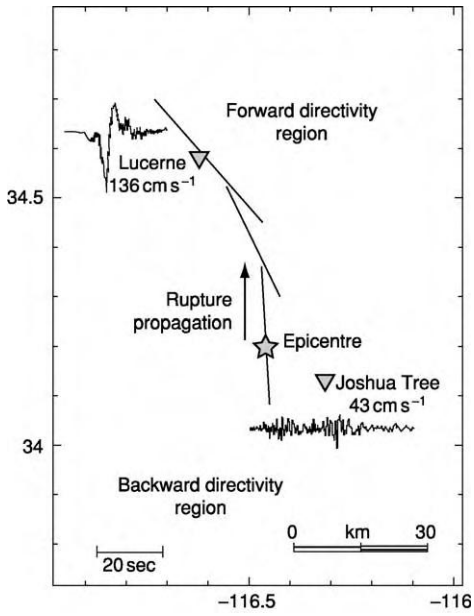


Figure 16 Velocity time series obtained from integration of horizontal accelerograms recorded at Lucerne and Joshua Tree stations during the 1992 Landers earthquake in California. The fault rupture propagated towards Lucerne and away from Joshua Tree, creating forward directivity effects (short duration shaking consisting of a concentrated high energy pulse of motion) at the former, and backward directivity effects (long duration shaking of several small pulses of motion) at the latter. Reprinted with permission from: Somerville PG, Smith NF, Graves RW, and Abrahamson NA (1997) Modification of empirical strong ground motion attenuation relations to include the amplitude and dur ation effects of rupture directivity. *Seismological Research Letters* 68(1): 199 222. © Seismological Society of America.

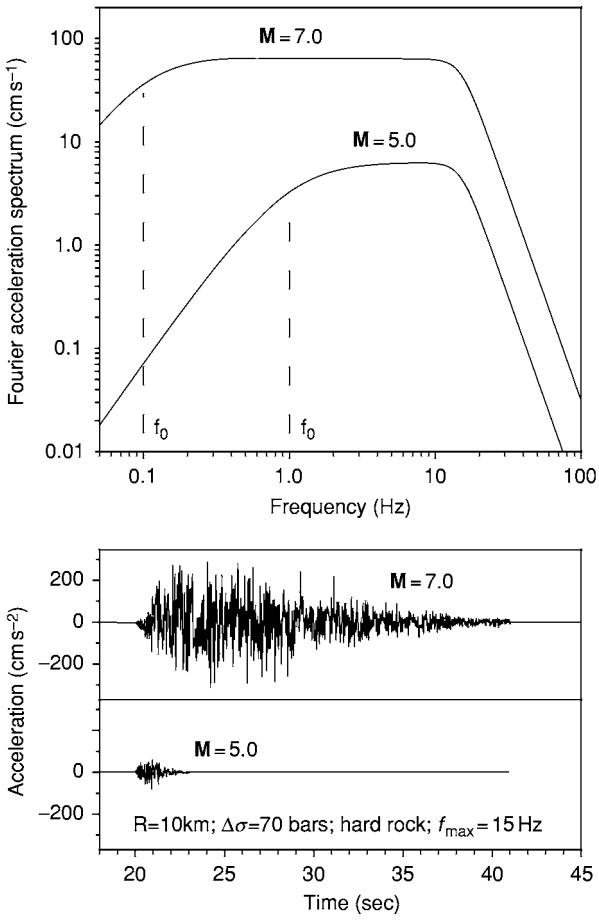


Figure 17 Theoretical Fourier spectra for earthquakes of mag nitudes 5 and 7 at 10 km from a rock site, assuming a stress parameter of 70 bars; f_0 is the corner frequency, which is inversely proportional to the duration of rupture. The lower part of the figure shows stochastic acceleration time series generated from these spectra, from which the influence of magnitude on amplitude and duration (or number of cycles) can be clearly observed. From the spectra in the upper part of the figure, it can be appreciated that scaling with magnitude is frequency dependent, with larger magnitude events generating proportionally more long period radiation.

postulated future events on past activity. The catalogue of instrumentally recorded earthquakes extends back at the very most to 1898, which is a very short period of observation for events for which recurrence intervals can extend to hundreds or even thousands of years. For this reason, great value is to be obtained from extending the catalogue through the careful interpretation of historical records, making use of empirical relationships between magnitude and intensity referred to previously. The seismic record can also be extended through palaeoseismology, by quantifying and dating coseismic displacements on geological faults.

Although the distinction masks many equally marked differences among the various methods that fall within each camp, a basic division exists between deterministic and probabilistic approaches to seismic hazard assessment. In the deterministic approach, only a few earthquake scenarios are considered, and sometimes just one is selected to represent an approximation to the worst case. The controlling earthquake will generally correspond to an event with the nominal maximum credible magnitude, located at the location closest to the site within the seismogenic source; the ground motion is generally calculated as the 50- or 84-percentile value from the prediction equation. In the probabilistic approach, all possible earthquake scenarios are considered, including events of every magnitude, from the minimum considered to be of engineering significance (~4) up to the maximum credible, occurring at every possible location within the source zones, and for each magnitude–distance combination various percentiles of the motion are considered to reflect the scatter in the ground-motion prediction equations. Alternative options for the input parameters may be considered by using a logic-tree formulation, in which weights are assigned to different options that reflect the

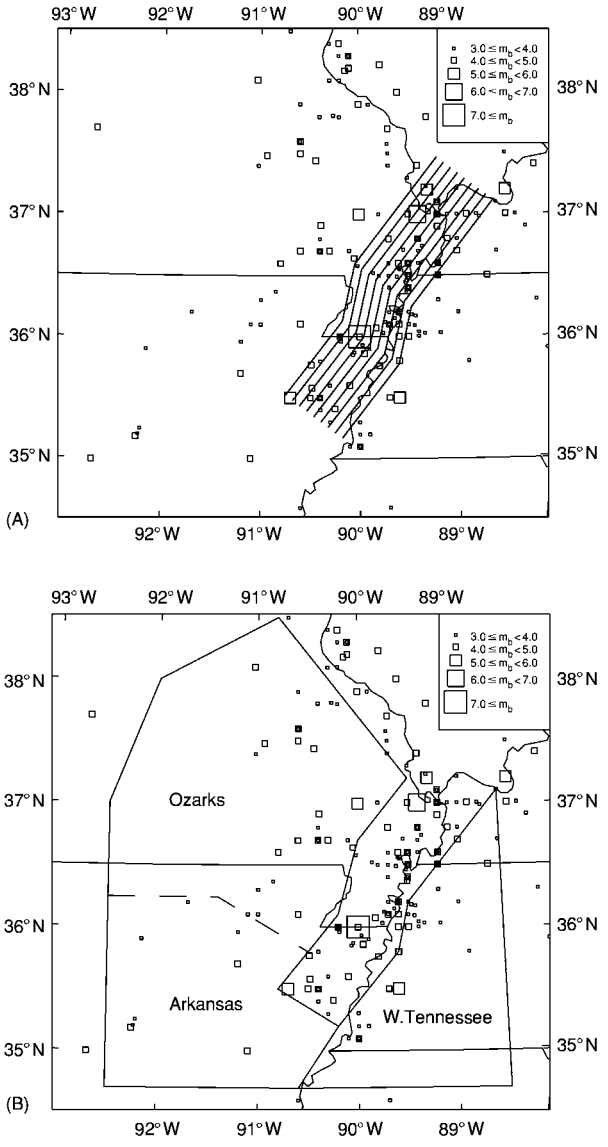


Figure 18 Seismic sources used for probabilistic hazard study of the Mississippi embayment. Small squares show the epicentres of earthquakes with the size corresponding to the body wave magnitude $\mathbf{M}_b$. (A) The New Madrid seismic source zone represented as a number of parallel hypothetical faults; (B) the area sources used to represent seismicity not associated with the New Madrid zone. Reprinted with permission from Toro GR, Silva WJ, McGuire RK, and Herrmann RB (1992) Probabilistic seismic hazard assessment of the Mississippi Embayment. *Seismological Research Letters* 63(3): 449 475. © Seismological Society of America.

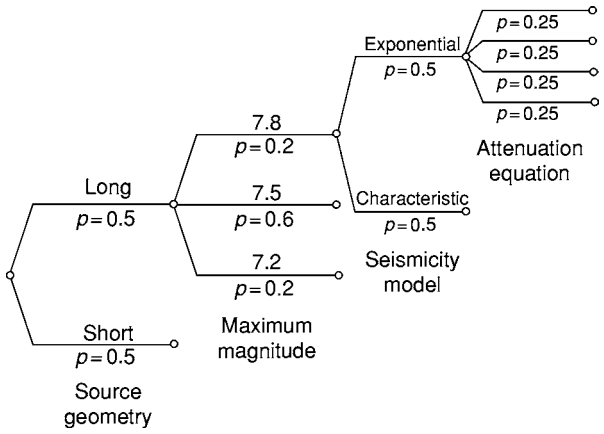


Figure 19 Logic tree formulation for probabilistic seismic hazard study of the Mississippi embayment. For different input parameters, different options are considered at each node, which are then assigned weights (p) to reflect the relative confidence in each option; the values of p at each node always sum to unity. The source geometry options are the faults shown in Figure 18, and an alternative model in which the faults are longer. Reprinted with permission from Toro GR, Silva WJ, McGuire RK, and Herrmann RB (1992) Probabilistic seismic hazard assessment of the Mississippi Embayment. *Seismological Research Letters* 63(3): 449 475. © Seismological Society of America.

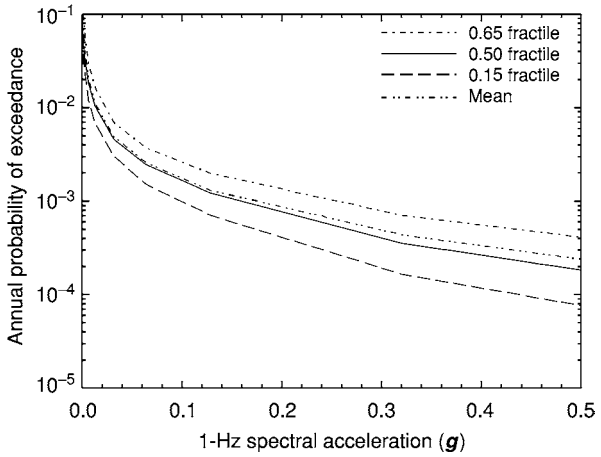


Figure 20 Seismic summary hazard curves for spectral acceleration at 1 s response period for Memphis (including site response), obtained using the logic tree formulation in Figure 19. Reprinted with permission from Toro GR, Silva WJ, McGuire RK, and Herrmann RB (1992) Probabilistic seismic hazard assessment of the Mississippi Embayment. *Seismological Research Letters* 63(3): 449 475. © Seismological Society of America.

relative confidence in their being the best representation (Figure 19). The output is a ranking of these earthquake scenarios in terms of the ground-motion amplitudes that they generate at the site and their frequency of occurrence (Figure 20).

Seismic hazard assessment can be carried out for an individual site, the output including a response spectrum calculated ordinate by ordinate, or even a suite of accelerograms. Hazard assessment is often performed for regions or countries, deriving hazard curves for a large number of locations and then drawing contours of a ground-motion parameter (most often PGA) for a given annual frequency of occurrence, which is often expressed in terms of its inverse, known as a return period (Figure 21). Such hazard maps in terms of PGA are the basis of

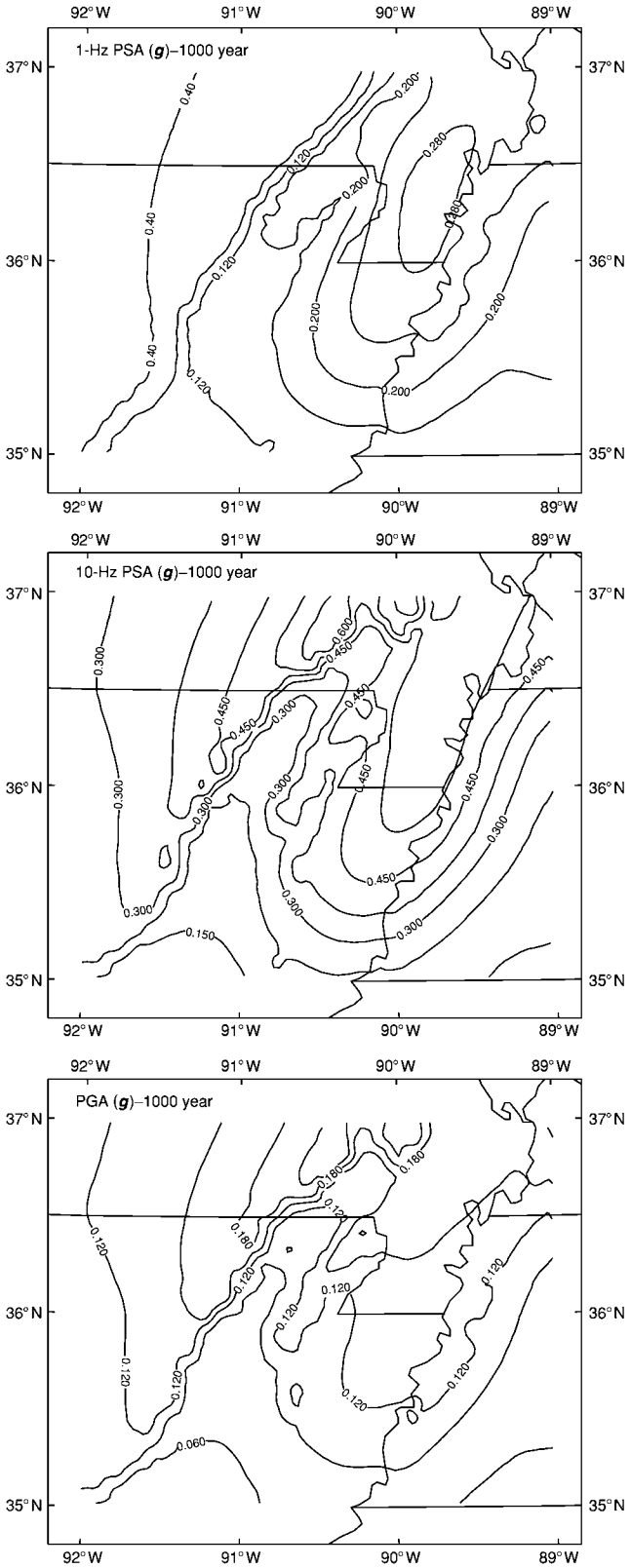


Figure 21 Contours of spectral accelerations in the Mississippi embayment with a 1000 year return period, for different response frequencies corresponding to the response periods of 1.0s (top) and 0.1s (middle) and to PGA (botom). The maps show the spectral accelerations read from the mean hazard curve. Reprinted with permission from Toro GR, Silva WJ, McGuire RK, and Herrmann RB (1992) Probabilistic seismic hazard assessment of the Mississippi Embayment. *Seismological Research Letters* 63(3): 449 475. © Seismological Society of America.

zonation maps included in seismic design codes, from which design response spectra are constructed by anchoring a spectral shape selected for the appropriate site class to the PGA value read from the zonation map.

See Also

Engineering Geology: Aspects of Earthquakes; Natural and Anthropogenic Geohazards; Liquefaction. **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes. **Tectonics:** Earthquakes; Faults; Neotectonics.

Further Reading

Abrahamson NA (2000) *State of the Practice of Seismic Hazard Evaluation. GeoEng 2000*, 19 24 November, Melbourne, Australia.

Ambraseys NN (1988) Engineering seismology. *Earthquake Engineering & Structural Dynamics* 17: 1 105.

Beskos DE and Anagnostopoulos SA (1997) *Computer Analysis and Design of Earthquake Resistant Structures: A Handbook*, chs 3 5. Southampton: Computational Mechanics Publications.

Bommer JJ (2003) Uncertainty about the uncertainty in seismic hazard analysis. *Engineering Geology* 70: 165 168.

Boore DM (1977) The motion of the ground in earthquakes. *Scientific American* 237: 68 78.

Boore DM (2003) Simulation of ground motion using the stochastic method. *Pure and Applied Geophysics* 160: 635 676.

Chen W F and Scawthorn C (2003) *Earthquake Engineering Handbook*, chs 1 10. Boca Raton, FL: CRC Press.

Douglas J (2003) Earthquake ground motion estimation using strong motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth Science Reviews* 61: 43 104.

Dowrick D (2003) *Earthquake Risk Reduction*. Chichester, England: John Wiley.

Giardini D (1999) The global seismic hazard assessment program (GSHAP) 1992 1999: summary volume. *Annali di Geofisica* 42(6): 957 1230.

Giardini D and Basham P (1993) Technical planning volume of the ILP's global seismic hazard assessment program for the UN/IDNDR. *Annali di Geofisica* 36(3/4): 3 257.

Hudson DL (1979) *Reading and Interpreting Strong Motion Accelerograms*. EERI Monograph, Earthquake Engineering Research Institute, Oakland, California.

Jackson JA (2001) Living with earthquakes: know your faults. *Journal of Earthquake Engineering* 5(special issue 1): 5 123.

Kramer SL (1996) *Geotechnical Earthquake Engineering*. Upper Saddle River, NJ: Prentice Hall.

Lee WHK, Kanamori H, Jennings PC, and Kisslinger C (2003) *International Handbook of Earthquake and Engineering Seismology, Part B*. chs. 57 64. Amsterdam: Academic Press.

McGuire RK (2004) *Seismic Hazard and Risk Analysis*. EERI Monograph MNO 10. Earthquake Engineering Research Institute, Oakland, California.

Naeim F (2001) *The Seismic Design Handbook*, 2nd edn. chs. 1 3. Boston: Kluwer Academic Publ.

Reiter L (1990) *Earthquake Hazard Analysis: Issues and Insights*. New York: Columbia University Press.

Natural and Anthropogenic Geohazards

G J H McCall, Cirencester, Gloucester, UK

Introduction

The topic of geohazards became popular with scientists and the media in the early 1990s at the time of the International Decade for Natural Disaster Reduction aimed, by the United Nations, specifically at developing nations in the Third World. It became popular with the media in the belief that we were moving into an age of disaster. In particular, the appreciation of the reality of global climate change and humankind's contribution in the Industrial Age to global warming has led to an awareness of the vulnerability of the world in which we live. Geohazards operate at local scales, e.g., in villages and towns, at a regional scale, and at the largest scale of all, global. Urban geohazards represent a specialized and increasingly important type of hazard.

An increasing incidence and scale of risk and disaster have recently occurred due to a number of factors:

- increased population concentrations;
- increased technological development;
- over-intensive agriculture and increased industrialization;
- excessive use of the internal combustion engine and other noxious fume emitters, and wasteful transport systems;
- poor technological practices in construction, water management, and waste disposal;
- excessive emphasis on commercial development;
- increased scientific tinkering with Nature without due concern for long-term effects.

Some hazards are natural, others are purely anthropogenic, and many have both natural and anthropogenic causal elements.

There is a need to consider geohazards from a sociological as well as a scientific viewpoint, as the perception and response of the local populace and all tiers of government are as important as technological understanding. Geohazards can also only be successfully evaluated with a full interdisciplinary approach: an international, possibly even a global, approach is also called for.

Definitions

There are two published definitions of geohazards.

> A geohazard is a hazard of geological, hydrological nature which poses a threat to Man and his activities. (McCall & Marker 1989)

and

> A geohazard is one that involves the interaction of man and any natural process on the planet. (McCall and others 1992)

The second is wider, but the first brings in the essential element of human vulnerability, which is well expressed by MR Degg in his diagram for the disaster equation (Figure 1) – the relationship between the hazard, disaster, and vulnerability. The larger and more severe the event and the more vulnerable the population, the greater the disaster; there is a threshold below which there is no disaster. If there are no human's present, strictly speaking there is no hazard.

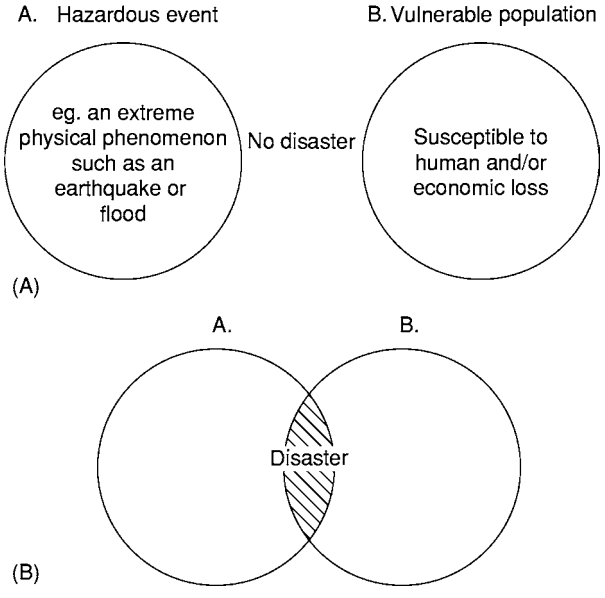


Figure 1 The 'disaster equation'. After Degg MR, in McCall GJH and Marker BR (1989) *Earth Science Mapping for Planning, Development and Conservation, Engineering Geology Special Publication No. 4.* London: Kluwer.

The words 'hazard' and 'risk' are often used interchangeably, but the risk is a quantification. RJ Blong defined 'risk' as the product of 'hazard' and 'vulnerability'. The word 'disaster' refers rather loosely to the consequence of hazards.

Types of Natural Geohazard

RJ Blong lists the following as the principal natural geohazards:

- earthquakes (r);
- volcanic eruptions (r);
- landslides (r);
- tsunamis (r);
- subsidence (s);
- coastal erosion (s);
- coastal progradation (s);
- soil erosion (s);
- expansive soils (s).

Already, we see two distinct classes of geohazard: the rapid-onset, intensive hazards (r) and the slow-onset, pervasive hazards (s). Attention has been traditionally focused by scientists, the media, and the populace on the former, but the latter can be equally damaging.

SA Thompson separated 'natural geohazards' from 'meteorological hazards':

- cyclones;
- tornadoes;
- floods;
- heatwaves;
- thunderstorms and gales.

All of these have some geological consequences, especially the first three.

GJH McCall later added to the list and separated those that are potentially catastrophic and cannot at present be reliably predicted, unless in general terms or not at all. These can be treated mainly by emergency planning, warning systems, education and communication, and evacuation. Mitigation and prevention procedures are of limited application. The list is:

- earthquake hazards (e.g., Kobe 1995);
- volcanic hazards (lava, pyroclastic flows, nuées ardentes, ash falls, lahars, blasts, and gas releases) (e.g., Etna, Martinique, Montserrat, Nevado del Ruiz 1985, Lake Nyos 1986);
- tsunamis (related to either of the above) (e.g., Hilo, Hawaii, 1960);
- extraterrestrial impacts (e.g., Tunguska 1908).

The remainder can be reasonably counteracted by engineering procedures at a cost. The list is:

- mass movements (landslides, avalanches, debris flows, ice-related movements) (e.g., Ventnor, Swiss Alps, Cotopaxi, Cordillera Blanca);
- subsidence (swallow holes, karstic processes, gypsum-related sink holes, sinking cities) (e.g., Pennines, Wuzhan, Ripon, Venice);
- flooding after abnormal rainfall and cyclones (Bangladesh);
- coastal erosion (sea-level rise) and coastal progradation (China, Burma);
- riverbank failure and silting up of rivers (Mississippi River);
- expansive and collapsing soils, thixotropic sands (e.g., Anchorage in Alaska);
- permafrost (e.g., Canada, Siberia);
- hazardous gas emission (radon) (e.g. Cornwall).

A third category was listed of those that fall between the two:

- combustion and wildfire (e.g., Australia);
- neotectonic deformation and fissuring (e.g., Xian);
- desertification (e.g., Sahel).

K Hewitt and I Burton, in 1975, published a list of parameters for 'selecting' hazard events:

- property damage extending to more than 20 families, or economic loss (including loss of income, a halt to production, costs of emergency action) in excess of US$50 000;
- major disruption of social services, including communications failure and closure of essential facilities of establishments of economic importance;
- a sudden, unexpected or unscheduled event, or series of events, which puts excessive strain on essential services (police, fire service, hospitals, public utilities) and/or requires the calling in of men, equipment, or funds from other jurisdictions;
- an event in which 10 or more persons are killed or 50 or more injured.

Such quantification of thresholds is applicable to rapid-onset, intensive hazards, but of little use for the slow-onset, pervasive hazards, such as karstic processes or coastal erosion. Also, it appears to be applicable to developed countries, but of little relevance to less developed countries, such as islands in the Pacific Ocean. Monetary loss and body count alone are thus poor indicators of the magnitude of an effect of a hazard on an afflicted community. Vulnerability should take into account not merely the risk, but also the endemic conditions inherent in the society. Deaths due to starvation consequent on hazards are not considered in this scheme.

Table 1 Frequency of hazards and deaths per event. From Thompson SA (1982)

Hazard	*Hazard frequency*	*% frequency*	*Deaths per event*
Geological hazards			
Landslides	29	2.7	190
Tsunamis	10	1.0	856
Volcanoes	18	1.7	525
Earthquakes	161	15.2	2652
Meteorological hazards			
Cyclones	211	19.9	2373
Tornadoes	127	12.0	66
Floods	343	32.3	571
Heatwaves	22	2.1	315
Thunderstorms and gales	36	3.4	587

SA Thompson, in 1982, produced a table combining frequency with deaths per event (Table 1). This table shows that earthquakes and cyclones are the most lethal per event on a global scale and both are of high frequency; however, flooding, because of its very high frequency, is nearly as lethal. The low frequency and number of deaths per event for landslides can be misleading – for example, in Basilicata, Italy, virtually all of the dense cluster of hilltowns are threatened by landslides, and it is unquestionably a major hazard, even if not a great killer. Likewise, in Nepal and the Pamirs (Figure 2), the steep topography means that landslides repeatedly wipe out villages and communications and present a major and intractable hazard problem.

SA Thompson also produced a table of fatalities covering Asia and Australasia between 1947 and 1981, showing how regional statistics can reveal extraordinary contrasts (Table 2). Asia had 85.8% of the global count of 1 208 044 global deaths from the hazards listed in Table 2 during this period. However, this figure is a combination of both geological and meteorological disasters, and, for the geological disasters, the figure is slightly below the global average. Australasia accounted for only 0.4% of the global deaths. The death counts and magnitude of the event cannot be correlated, because of factors such as variations in populations at risk, ground conditions, and building construction quality and type. In the case of volcanic eruptions, there is a wide variation in the expected magnitude and type of eruption according to the classification of the volcano and its petrological products.

Volcanoes and earthquakes are covered by separate entries in this encyclopedia (*see* **Tectonics: Earthquakes**, **Volcanoes**).

Figure 2 Beneath this peaceful scene in the Pamirs, Tadzhikistan, is buried Xait, one of 33 villages buried by landslides triggered by earthquakes in 1948, with 50 000 lives lost which geological knowledge could have saved. de Mulder EFJ, Holland, International Union of Geological Sciences 'COGEOENVIRONMENT' photograph.

Table 2 Fatalities per event in Asia and Australasia. From Thompson SA (1982)

Event	*Asia*	*Australasia*
Landslides	3576	0
Tsunamis	7864	44
Volcanoes	2806	4000
Earthquakes	33 3623	133

Procedures for the Mitigation of Natural Hazards

These include:

- the study of existing conditions and evidence of past activity, including mapping;
- prediction in time and space (when?, where?, and on what scale? risk assessment);
- prevention or mitigation by engineering or other methods;
- planning of land use;
- monitoring and installation of warning systems;
- emergency planning; proactive, including planning for evacuation; post-event relief.

Each of the above will have variable relevance to a particular hazard and subject site/area/region, and may have none at all.

Types of Anthropogenic Geohazard

JD Mather and others included:

- surface movements – subsidence (Figure 3) related to mineral or fluid (hydrocarbon, water) extraction and surface collapse into voids left by mineral workings;
- contaminated land – left by mineral workings, chemical plants, gas works, etc. ('brownfield sites');
- rising groundwater levels;
- modification of groundwater quality (unsewered sanitation, organic solvent pollution);
- waste disposal.

GJH McCall and others also included:

- loss of soil and agricultural land (including the effects of urbanization on the soil resource);
- reduction in biodiversity.

The anthropogenic hazards are mostly of long build up – even surface collapses due to old mine workings (Figure 4), which appear abruptly without warning, as the roof of the mined-out cavity gradually moves towards the surface due to the fall-out of blocks.

Three of these quiet, pervasive, slow build-up hazards are mentioned below to illustrate this type of geohazard.

Figure 3 Leaning church tower due to differential settlement in soft Holocene sediments, a foundation problem near Lake Ijssel, The Netherlands. de Mulder EFJ, Holland, International Union of Geological Sciences 'COGEOENVIRONMENT' photograph.

Rising Groundwater

Rising groundwater was described in 1992 by DJ George – a problem that has affected many cities in the Middle East (Kuwait, Doha, Cairo, Riyadh, Jeddah, Madinah), and is a major hazard in terms of cost, and loss of and injury to human life. A typical cause is the installation of new water supply systems piped in from outside a city area, which previously relied on numerous small wells within it. The water table under the city is no longer utilized and the water table rises. This produces:

- damage to buildings and structures;
- damage to services and roads;
- overloading of sewer systems and treatment plants;
- salting and waterlogging of soils;
- public health hazards.

This is essentially anthropogenic but, like all such hazards, is due to interference with the natural system.

Sea-Level Change on the China Coast

S Wang and X Zhao covered this hazard, which is likely to become more significant with global climate change – warming related to the 'greenhouse effect' – which has now been shown to have been unquestionably influenced since the beginning of the Industrial Age (see summary report by McCall GJH (2003); **Solar System:** The Sun). On the long, sinuous, and island-dotted coast of China, there are many large, industrialized cities and regions – e.g., Shanghai, Tianjin, Guangzhou – located on coastal plains and deltas. The evidence reveals a complex picture of overall oscillations of the order of a thousand years and different effects of Quaternary crustal movements, up and down in different regions; however, subtracting the latter, a general rise of 0.65 mm per year has been recognized. A future rise of several centimetres, which is predicted (also allowing for the effect of global warming), will have disastrous effects on the coastal and coastal hinterland environment and the industry and agriculture there. The wetlands will be inundated. Engineering constructions proposed in the Yangtze delta will be under the following main headings:

- control of land subsidence due to human activity;
- prevention of seawater seepage and windstorm tides;
- drainage of low-lying land and land improvement;
- taking into account sea-level rise in planning new urban areas.

Global Soil Loss: Biodiversity Loss

WS Fyfe has emphasized the fact that the explosive growth of the human population (a taboo subject as WI Stanton has emphasized) (Figure 5) is at the root of most of the quiet, pervasive hazards. He indicates that topsoil loss is occurring at a global rate of 0.7% per year. The global climate bioproductivity is vital to the functioning of the global thermostat, and species continue to be wiped out wholesale due to the activities of humankind.

As described by GJH McCall *et al.*, in an original definitive account, urban geohazards are becoming more and more significant as urbanization mushrooms in the early twenty-first century (Figure 6). These include natural and anthropogenic hazards.

Hazard and Risk Mapping

JG Doornkamp, in 1989, published a concise account of hazard mapping. He covered:

- landslips and avalanches;
- natural ground subsidence;

Figure 4 The result of a void migrating to the surface in old mine workings beneath a suburb of Glasgow. British Geological Survey photograph, in McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam: Balkema.

- hazards of quarrying and mining;
- erosion and deposition;
- flooding;
- saline soils;
- permafrost;
- seismological hazards;
- volcanic hazards.

The third in this list is an anthropogenic hazard – the general principles apply to both natural and anthropogenic hazards. JG Doornkamp noted that the distinction is ambiguous.

Every hazard has three key properties:

- magnitude;
- frequency;
- location.

The last of these is most amenable to mapping. Mapping may be taken for different reasons:

- to define the nature and extent of a historical event;
- to define the nature and extent of a recent event;
- to define present conditions in order to assess the likelihood of a recurrent event.

The social and economic content of the affected subject must be included in any such programme. A hazard cannot in itself be mapped, as it is a process. Rather, the results of a hazard are mapped. The resultant landforms, the effects on soil and rocks, and the effects on constructions are mapped.

JG Doornkamp devised a useful diagram showing the context of hazard mapping (Figure 7).

Aids to mapping include:

- satellite imagery;
- airborne multispectral scanning;
- photogrammetry;
- shallow geophysics;
- global positional systems.

DKC Jones has illustrated well the procedures in hazard mapping in an account of the application to landslides. The graphical methods are only part of the exercise, which includes:

- graphical methods, mainly involving mapping;
- the analysis of the empirical relationships between landslides and individual causal factors, such as slope steepness, material type, vegetation cover, rainfall intensity, and human constructions;
- multivariate analysis of the causes of landslides.

True hazard maps display the extent of past hazardous events (zoned), together with an internal division reflecting the magnitude frequency distribution

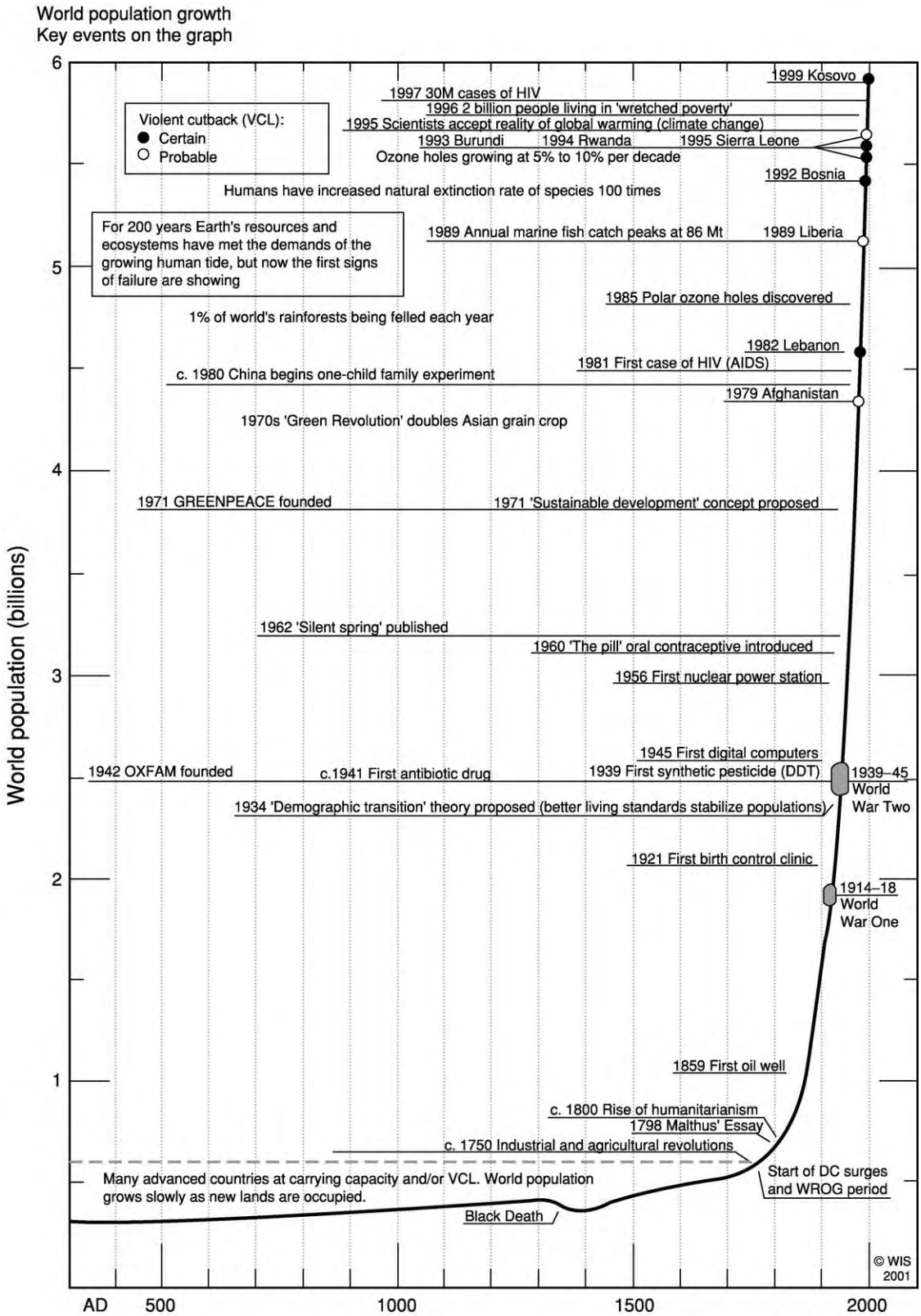


Figure 5 The escalation of the world population. After Stanton WI (2003) *The Rapid Growth of Human Populations: 1750 2000*. Brent wood: Multi science Publishers.

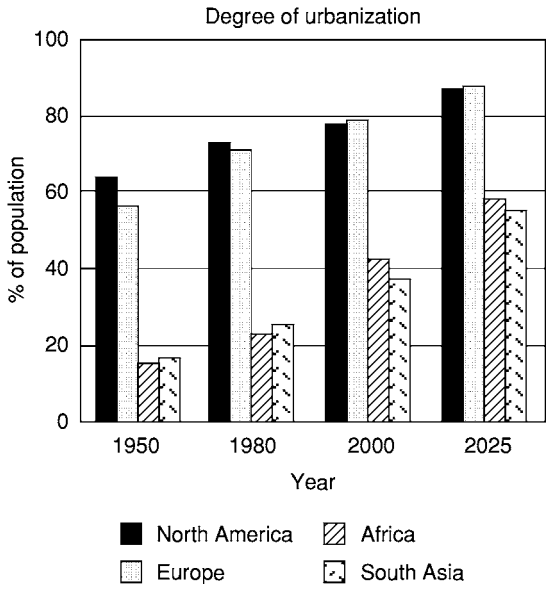


Figure 6 The escalation of urbanization. After de Mulder EFJ, in McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam: Balkema.

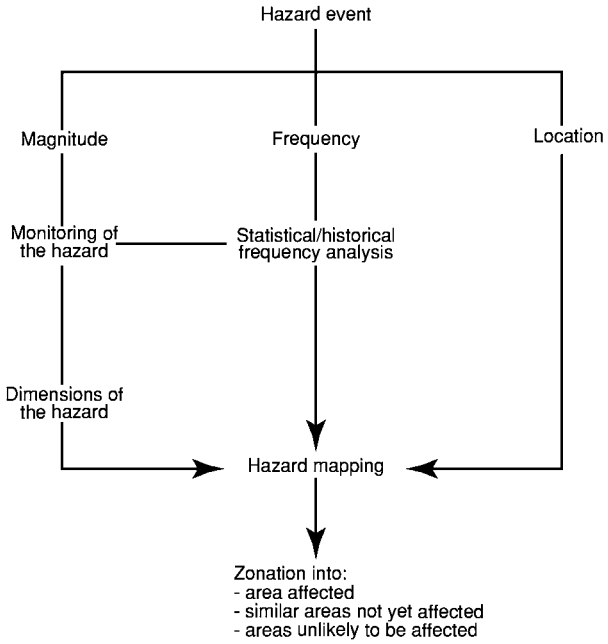


Figure 7 The context of hazard mapping. After Doornkamp JG, in McCall GJH and Marker BR (1989) *Earth Science Mapping for Planning, Development and Conservation, Engineering Geology Special Publication No. 4*. London: Graham and Trotman.

of past events (hazard rating). Evidence of past failures must be used to predict the likelihood of future events, their scale, and frequency. Risk maps are quite different in that they assess the potential losses that may be incurred by society. They may involve numerical quantification (for example, the table for the Central Pacific by Blong RJ (1988)) (Table 3). They may be followed by maps showing geographical divisions of recommended action (for example, the map of Basilicata; Figure 8). Hazard mapping may be very complex and involve numerous maps of different type. For example, DKC Jones, in 1992, listed the following as all used in landslide mapping:

- geological structure;
- chronostratigraphy of rocks;
- lithostratigraphy of rocks;
- lithostratigraphy of soils;
- rockhead contours;
- geotechnical properties of soils;
- geotechnical properties of rocks;
- hydrological conditions;
- geomorphological processes;
- geomorphological history, palaeodeposits, surfaces, or residual conditions;
- seismic activity;
- climate (including precipitation);
- ground morphology (especially slope height, length, and angle);
- land use;
- vegetation type, cover, root density, and strength;
- pedological soils, type and thickness of regolith;
- past landslide deposits;
- past landslide morphometry.

Wider Responses

JG Doornkamp listed four wider responses beyond mapping considerations that can be adopted to mitigate hazards:

- land use planning: place development in less hazardous places;
- economic planning: investment in the right places and maintenance of financial reserves to cover disasters; adequate insurance schemes;
- development control: designing built structures to cope with known hazards (e.g., earthquake-resistant designs);
- emergency planning: evacuation procedures to be planned and rehearsed in all areas at risk from major hazards; rescue teams trained and tested with modern equipment, able to move quickly to the site of a disaster.

Engineering and Geohazards

Experience has shown that geologists, on the one hand, and engineers and developers, on the other,

Table 3 Geological risk rating for selected towns in Papua New Guinea. After Blong RJ (1988)

Town	*Earthquake rating*	*Volcano rating*	*Tsunami rating*	*Risk rating*
Arawa	6	2	0.1	8
Daru	2	0.1	0.03	2
Goroka	5	0.1		5
Kavieng	3	0.1	0.1	3
Kiet	6	2	0.1	8
Kimbe	10	4	0.1	14
Kokopo	10	4	0.2	14
Lae	5	0.1	0.2	5
Madang	5	1	0.2	6
Mt Hagen	3	0.1		3
Port Moresby	3	1	0.01	3
Rabaul	10	5	0.2	15
Wewak	6	2	0.2	8

tend to have rather different viewpoints concerning geohazards. The differences concern site investigation, which is the keystone of any engineering project. There is a tendency amongst engineers and developers to omit consideration of the hazards pertaining to a particular site from previous site investigations, with the view that the problems will only be addressed if and when they arise. For the past three decades, environmental geologists have stressed that these factors must be evaluated as part of the previous site investigation, and this is becoming, certainly in the UK, a regulatory condition for construction and development.

Engineers are greatly involved in investigation, mitigation, and repair programmes related to both natural and anthropogenic hazards. For example,

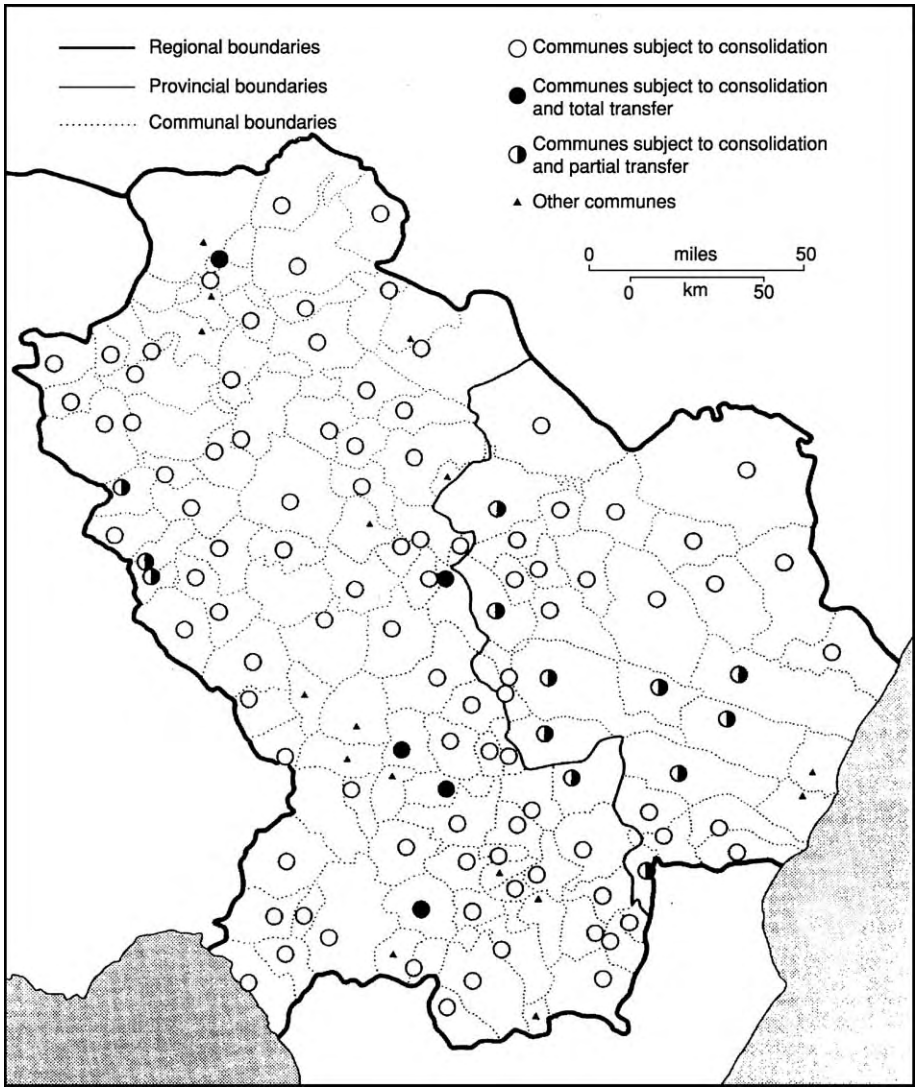


Figure 8 Landslide hazard assessment map, Basilicata, Italy. After Jones DKC, in McCall GJH, Laming DJC, and Scott SR (1992) G*eohazards Natural and Man Made*. London: Chapman and Hall.

they design earthquake-proof buildings in Japan and deviation works to prevent lava flows or lahars from reaching villages and towns (as in Sicily). To illustrate the variety of engineering applications to geohazards, a list is given below of the types of work in which engineers have been involved, and which are represented by reports to Special Publications of the Engineering Group of the Geological Society:

- liquefaction of sediments in the Fraser River Delta (Canada);
- flooding at Ladysmith (South Africa);
- fluvial hazards (Bihar, India);
- ice- and snow-related high-altitude problems in glacial lakes (Himalayas, Peru);
- cliff erosion (Isle of Wight, UK);
- lessons from the Kobe earthquake (Japan);
- risks from low-seismicity earthquakes (Holland);
- gypsum-related subsidence at Ripon, Yorkshire (UK);
- subsidence related to solution in chalk (south-east England);
- groundwater recharge under rapid urbanization (Mexico, Thailand);
- organic solvent pollution of groundwater (Coventry, UK);
- landfill disposal of urban wastes (Tanzania, Gambia, Mauritius);
- acid mine drainage (Transvaal);
- heavy metal contamination (Cornwall, UK);
- foundation conditions and site investigations (UK);
- land restoration, brownfield sites;
- detection of karst features by remote sensing (England).

Conclusion

There have been great advances in hazard-related earth science and engineering in recent decades, but these hazards are far from understood and there is an ongoing need for more data and statistics. Countries such as China, with its immense population, many of whom live in sites of high risk, and Colombia or Nepal, where the climate and physiography militate to give a situation in which natural geohazards are quite unavoidable, but can be mitigated, if only by emergency planning, are extreme examples; however, no country is free from these problems. With the growth of megacities, urban geohazards are assuming increasing importance (Figure 6).

See Also

Engineering Geology: Liquefaction; Made Ground; Problematic Rocks; Problematic Soils; Site and Ground Investigation; Subsidence. **Environmental Geology**. **Soil Mechanics**. **Solar System:** The Sun. **Tectonics:** Earthquakes. **Volcanoes**.

Further Reading

Appleton JD, Fuge R, and McCall GJH (1996) *Environmental Geochemistry and Health, Special Publication No. 113*. London: Geological Society.

Blong RJ (1988) Assessment of eruption consequences. *Kagoshima International Conference on Volcanoes Proceedings*, pp. 569 572.

Bullock P and Gregory PJ (1991) *Soils in the Urban Environment*. Oxford: Blackwell Scientific Publications.

Culshaw MG, Bell FG, Cripps JC, and O'Hara M (1987) *Planning and Engineering Geology, Engineering Geology Special Publication No. 4*. London: Geological Society.

de Mulder EFJ COGEOENVIRONMENT, International Union of Geological Sciences (IUGS) (undated leaflet) *Planning and Management, the Human Environment the Essential Role of the Geosciences*.

Maund JG and Eddleston M (1998) *Geohazards in Engineering Geology, Engineering Geology Special Publication No. 15*. London: Geological Society.

McCall GJH (2003) Global climate change a view from the floor. *Geoscientist* 13(6): 18 20.

McCall GJH, Laming DJC, and Scott SR (1992) *Geohazards Natural and Man Made*. London: Chapman and Hall.

McCall GJH and Marker BR (1989) *Earth Science Mapping for Planning, Development and Conservation, Engineering Geology Special Publication No. 4*. London: Graham and Trotman.

McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam: Balkema (see, especially, papers by McCall, Simpson, and Mather).

Stanton WI (2003) *The Rapid Growth of Human Populations: 1750 2000*. Brentwood: Multi science Publishers.

Thompson SA (1982) *Trends and Developments in Global Natural Disasters 1947 1981*. University of Colorado Institute of Behavioural Science Natural Hazards. Research Working Paper No. 45.

Wang S (1997) *Engineering Geology, Proceedings of the 30th International Geological Congress, Beijing, 1997*. Vol. 2 3. Utrecht: VSP.

Zhang Z, de Mulder EFJ, Liu T, and Zhou L (1997) *Geosciences and Human Survival, Environment, Natural Hazards, Global Change. Proceedings of the 30th International Geological Congress, Beijing, 1997*. Vol. 2 3. Utrecht: VSP.

Liquefaction

J F Bird, Imperial College London, London, UK
R W Boulanger and I M Idriss, University of California, Davis, CA, USA

Introduction

Liquefaction-induced ground failure has caused widespread damage and devastation; a recent example is the closure of the Port of Kobe, Japan's busiest port, following the 17 January 1995 earthquake, largely due to the liquefaction-related failure of reclaimed land (Figure 1). The potential consequences of liquefaction are far reaching, ranging from settlement or tilt of individual building foundations to the spread of fire when the water supply system is damaged by permanent ground deformations.

Liquefaction is the loss of shear strength of a saturated cohesionless soil due to increased pore water pressures and the corresponding reduction in effective stress during cyclic loading. Liquefaction and its associated ground deformation are very complex phenomena, and the term 'liquefaction' has been used to describe a wide range of soil behaviour. This article focuses on the susceptibility, triggering, and consequences of earthquake-induced liquefaction; liquefaction caused by non-earthquake-related loading is not included. Although the emphasis here is mainly the liquefaction of saturated cohesionless deposits such as sands or silty sands, it is important to consider that strength loss in fine-grained soils under earthquake loading can also pose a significant hazard. The assessment of liquefaction hazard is an essential part of any engineering project in seismic regions and is needed to make informed decisions with respect to mitigation options, foundation design, and emergency response and recovery plans based on what is considered to be an acceptable level of risk.

The Principles of Liquefaction

As a saturated cohesionless soil is cyclically loaded, its particle structure can tend to collapse to a denser arrangement. If the soil's permeability and the site stratigraphy are such that drainage cannot occur immediately, then, as the collapse occurs, stresses will be transferred from the soil grain contacts to the pore water, leading to an increase in pore water pressure. In simple terms, when the pore water pressure increases, the effective stress (total stress minus pore water pressure) on the particle structure will reduce, and as it approaches zero, the shear resistance of the soil will also approach zero. This loss of effective stress and shear resistance is known as liquefaction.

The stress–strain behaviour of liquefying sand depends strongly on its relative density. When loose sand liquefies, the gravitational static shear stresses may exceed the shear resistance of the soil and rapid deformation with very large shear strains can commence; this is referred to as flow deformation. The soil behaviour is termed 'contractive', and the shear resistance exhibited by the liquefied soil during flow deformation is termed the 'residual strength'. When moderately dense granular soils are cyclically sheared, pore pressures may similarly rise and liquefaction can be triggered. However, rather than undergoing flow deformation, the soil particle structure may try to expand as it reaches a certain level of shear strain, resulting in what is termed 'dilative' behaviour. For undrained conditions, this leads to a reduction in the pore water pressures and a corresponding increase in effective stress and shear resistance. A shear stress reversal, however, such as will occur many times during earthquake shaking, may cause the soil particle structure to be incrementally contractive and the state of zero effective stress may be temporarily reached once more. This continued cycle of zero effective stress and strength regain is termed 'cyclic mobility'. The cumulative deformations can be significant, particularly if the duration of shaking is long, but dilative soils do not exhibit very large flow deformations in the way that contractive soils do (Figure 2).

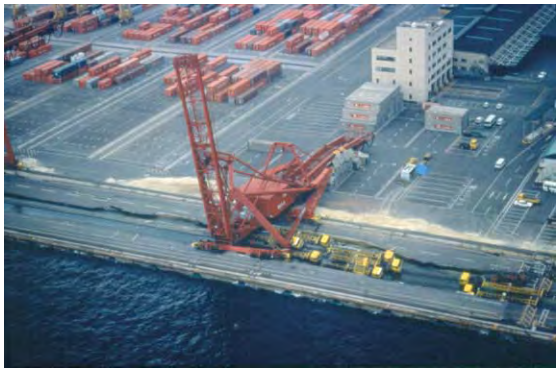


Figure 1 Liquefaction of reclaimed fills at the Port of Kobe in 1995 caused complete suspension of operations. The lateral displacement of the quay walls in this picture pulled apart the crane legs, causing collapse. Photograph by Leslie F. Harder, Jr.

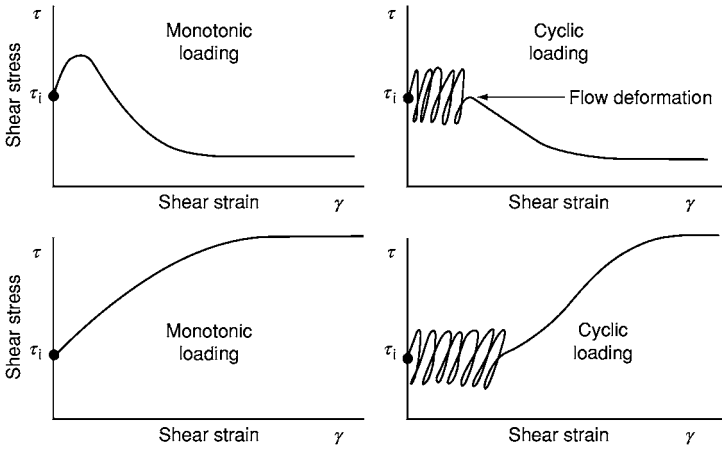


Figure 2 Stress strain behaviour of contractive (top) and dilative (bottom) soils under monotonic and cyclic loading. τ_i = initial (static) shear stress. In all cases, the steady state undrained shear strength is ultimately reached. Reproduced from Castro (1976).

Figure 3 Sand boil in liquefied soil following an earthquake; Japan, 1983.

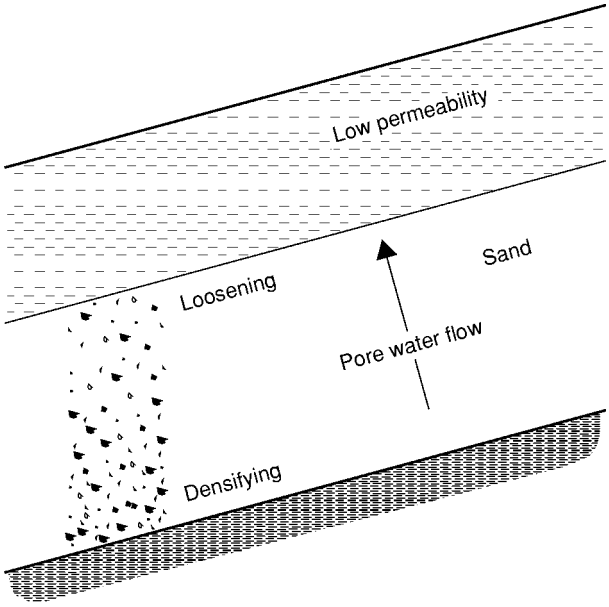


Figure 4 Schematic illustration of void redistribution resulting from spreading of pore pressure and global volume changes. Reproduced from the National Research Council (1985) *Liquefaction of Soils during Earthquakes. Report by the Participants in the Workshop on Liquefaction, Committee on Earthquake Engineering, National Research Council, Report No. CETS EE 001*. Washington, DC: National Academy Press.

Sand boils are caused by the tendency of the excess pore water pressures to dissipate upwards, towards the free surface, carrying soil particles up from the liquefied layer through cracks or channels in the overlying material and ejecting them at the ground surface (Figure 3). The upward seepage of pore water, driven by the earthquake-induced excess pore pressures, can be impeded by less permeable overlying soil layers, which can result in water accumulating near the interface between the liquefied soil and the overlying lower permeability soil. The accumulation of water can loosen the soil and possibly even result in the formation of water films, either of which greatly reduces the available shear resistance along the interface (Figure 4). This phenomenon is referred to as 'void redistribution'. A key consequence of void redistribution is that the residual shear strength of liquefied soil does not just depend on the pre-earthquake properties of the soil, but also depends on those factors affecting the dissipation and movement of pore water following earthquake shaking.

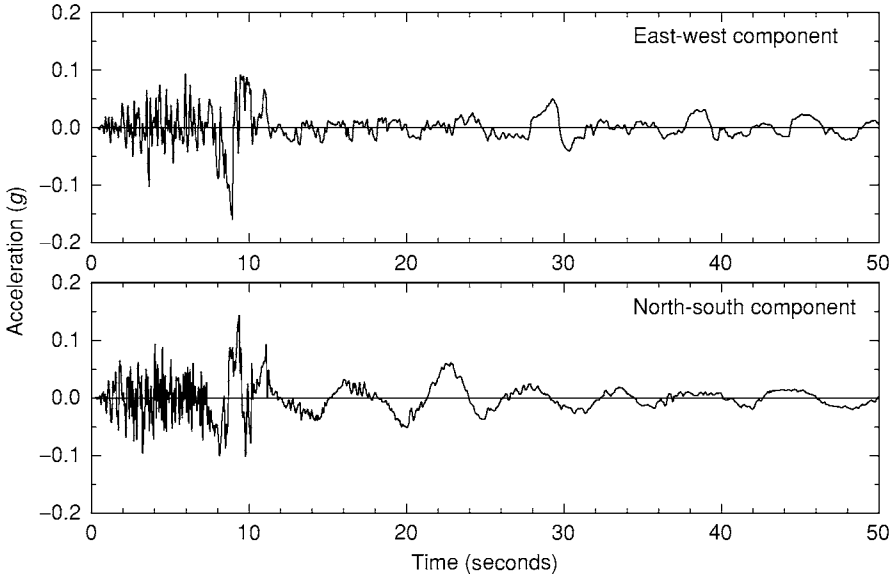


Figure 5 Horizontal accelerograms recorded in Niigata, Japan in the 1964 earthquake. The area surrounding the recording station was heavily damaged by liquefaction. The onset of liquefaction appears to have occurred at approximately 10 s after strong ground shaking commenced.

Another by-product of triggering of liquefaction is the potential effect it has on the characteristics of the strong ground motion. A number of accelerograms have clearly shown the triggering of liquefaction beneath the instrument (e.g., Figure 5).

Assessment of Liquefaction Hazard

The potential for liquefaction to cause damage is assessed in three stages: the susceptibility to liquefaction, the likelihood of liquefaction being triggered by the design earthquake scenario, and the consequences related to liquefaction (Figure 6). The variables that influence the onset of liquefaction include environmental factors such as the location of the water table, site stratigraphy, and depositional and seismic loading history of the soil; soil characteristics such as relative density, grain size distribution, mineralogy, and the presence of any cementing agents; and the characteristics of the earthquake under consideration, mainly the amplitude of ground shaking and its duration.

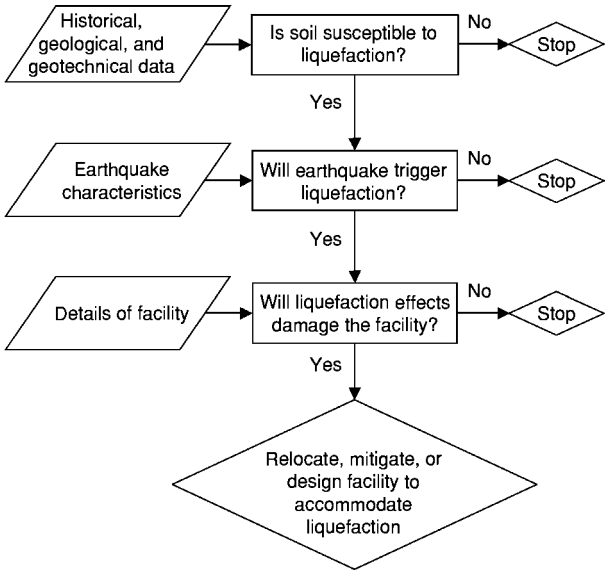


Figure 6 Flow chart for the three stages of assessing the potential of earthquake induced liquefaction to damage structures, lifelines, and other facilities.

Liquefaction Susceptibility

The most obvious evidence that a particular deposit is susceptible to liquefaction is historical precedence, either at the same site or in similar conditions elsewhere; it is not unusual for liquefaction to re-occur in the same location. In regions of potential hazard where there is no historical evidence, other information must be used to make an initial assessment of the liquefaction risk. The criteria in Table 1, classifying different levels of liquefaction susceptibility for various types and ages of soil deposits, are widely used by engineers and engineering geologists for the purposes of either producing regional maps of liquefaction hazard for planning purposes or for preliminary assessment of potential hazards for site selection purposes. Saturated loose cohesionless soils are the most

Table 1 Susceptibility of sedimentary deposits to liquefaction during strong seismic shaking[a]

Type of deposit	*Distribution of cohesionless sediments in deposit*	*Likelihood that cohesionless sediments, when saturated, would be susceptible to liquefaction*			
		<500 years	*Holocene*	*Pleistocene*	*Pre Pleistocene*
Continental					
River channel	Locally variable	Very high	High	Low	Very Low
Flood plain	Locally variable	High	Moderate	Low	Very Low
Alluvial fan and plains	Widespread	Moderate	Low	Low	Very low
Marine terraces and plains	Widespread		Low	Very low	Very low
Delta and fan delta	Widespread	High	Moderate	Low	Very low
Lacustrine and playa	Variable	High	Moderate	Low	Very low
Colluvium	Variable	High	Moderate	Low	Very low
Talus	Widespread	Low	Low	Very low	Very low
Dunes	Widespread	High	Moderate	Low	Very low
Loess	Variable	High	High	High	Unknown
Glacial till	Variable	Low	Low	Very low	Very low
Tuff	Rare	Low	Low	Very low	Very low
Tephra	Widespread	High	High	?	?
Residual soils	Rare	Low	Low	Very low	Very low
Sebka	Locally variable	High	Moderate	Low	Very low
Coastal zone					
Delta	Widespread	Very high	High	Low	Very low
Estuarine beach	Locally variable	High	Moderate	Low	Very low
High wave energy	Widespread	Moderate	Low	Very low	Very low
Low wave energy	Widespread	High	Moderate	Low	Very low
Lagoonal	Locally variable	High	Moderate	Low	Very low
Fore shore	Locally variable	High	Moderate	Low	Very low
Artificial fill					
Uncompacted fill	Variable	Very high			
Compacted fill	Variable	Low			

[a]Reproduced with permission from Youd TL and Perkins DM (1978) Mapping liquefaction induced ground failure potential. *Journal of the Geotechnical Engineering Division, ASCE* 104(GT4): 443 446.

vulnerable to liquefaction. Older deposits generally have higher densities and thus greater resistance to liquefaction.

The mode of deposition is also an important factor in liquefaction susceptibility, as can be seen in Table 1. Alluvial, fluvial, marine, deltaic, or wind-blown deposits generally have a higher susceptibility than do either residual soils or glacial tills; materials that are more highly consolidated. There are few case histories of liquefaction occurring at depths greater than about 10–15 m. This is partly due to the fact that soils at greater depths tend to be older, and therefore more resistant to liquefaction, and also that the earthquake-induced cyclic shear stress ratio (see eqn [1]) will be lower at greater depths. In addition, liquefaction at greater depths beneath level ground surfaces would not necessarily be evident at the ground surface, and thus may go undetected. The issue of a maximum depth for liquefaction is of particular importance for the construction and evaluation of large earth dams, where the presence of significant overburden above natural deposits should not be construed as a reduction in liquefaction susceptibility.

When some site-specific information is available, compositional criteria can be used to make preliminary assessments of liquefaction susceptibility. The fines content and clay content, the Atterberg limits, and the particle size distribution are all relevant. Though clays are generally resistant to liquefaction, it is also understood that plastic silts and clays may develop significant shear strains under sufficiently strong and sustained earthquake loading, leading to strength loss and rapid deformation or even instability. Compositional characteristics are used in various criteria, such as the often-cited 'Chinese criteria,' to evaluate a soil's liquefaction susceptibility, but a classification of 'non-liquefiable' should not be equated with the absence of a problem. Thus, liquefaction evaluation procedures may not be appropriate for plastic silts and clays, but the potential for such soils to fail and induce damage under earthquake loading should not be discounted.

Evaluation of Liquefaction Potential

The state-of-the-practice approach for evaluating liquefaction potential at a specific site, often referred to

as the 'simplified procedure', was first published by H B Seed and I M Idriss in 1971. This procedure compares the shear stresses induced by an earthquake with those required to cause liquefaction in the soil profile, where the liquefaction resistance is based on empirical data. The procedure is based on the following relationship:

$$\mathrm{CSR} = \frac{\tau_{av}}{\sigma'_v} = 0.65 \cdot \frac{a_{max}}{g} \cdot \frac{\sigma_v}{\sigma'_v} \cdot \frac{r_d}{\mathrm{MSF}} \qquad [1]$$

where CSR is the cyclic shear stress ratio, τ_{av}/σ'_v, at the depth of the soil layer under consideration, a_{max} is the peak ground acceleration at the surface, g is acceleration due to gravity, σ_v/σ'_v is the ratio of total to effective vertical stress in the soil layer, and r_d is a stress reduction coefficient (Figure 7). MSF is a magnitude scaling factor, equal to unity for $M = 7.5$ earthquakes, and greater than unity for $M < 7.5$, to correct in an implicit manner for the duration of shaking. Recommended procedures for the evaluation of r_d and MSF for use in conjunction with eqn [1] are described in a 2004 paper by Idriss and Boulanger.

The factor of safety (F) against liquefaction is then determined as follows:

$$F = \mathrm{CRR}/\mathrm{CSR} \qquad [2]$$

where CRR is the cyclic resistance ratio of the soil layer. The evaluation of CRR can be determined directly by laboratory tests on high-quality undisturbed samples and reconstituted samples (for new fills), or indirectly by *in situ* testing. An effective method (albeit an expensive one) of evaluating the cyclic resistance involves collection of high-quality undisturbed soil samples, using techniques such as ground freezing, and undertaking cyclic testing in the laboratory (e.g., cyclic simple shear or triaxial tests). The most common approach used by engineers to determine the CRR is based on *in situ* testing, and in particular the standard penetration test (SPT) and the cone penetration test (CPT). Shear wave velocity (V_s) measurements and the Becker penetration test (BPT) are also used in some specific applications.

Several publications provide guidance on the estimation of CRR using *in situ* measurements. Such correlations are obtained through back analysis of case history data, and as such are related to the variables used to define the CSR (eqn [1]) for the case histories, which can vary considerably among different authors. For this reason, it is important that practitioners are consistent in the methodology they adopt, because the combination of different recommendations could lead to incorrect results. An example of a recently completed correlation relating CRR to $(N_1)_{60}$ for clean sands (i.e., fines content $\leq 5\%$) is presented in Figure 8. The parameter $(N_1)_{60}$ represents the SPT blowcount corrected to a hammer energy level of 60% (to compensate for variations in testing equipment and procedures) and normalized to an effective vertical stress of 1 atmosphere (to compensate for variations in the depth of the *in situ* testing between the site and the case history data). Similar curves for cohesionless soils with higher fines content are also available, based on the same data used to construct the curve in Figure 8.

Permanent Ground Deformation

The extent of the permanent ground deformation as a result of liquefaction will dictate the engineering solutions in terms of mitigation or redesign. The interaction between liquefied and non-liquefied soils, foundations

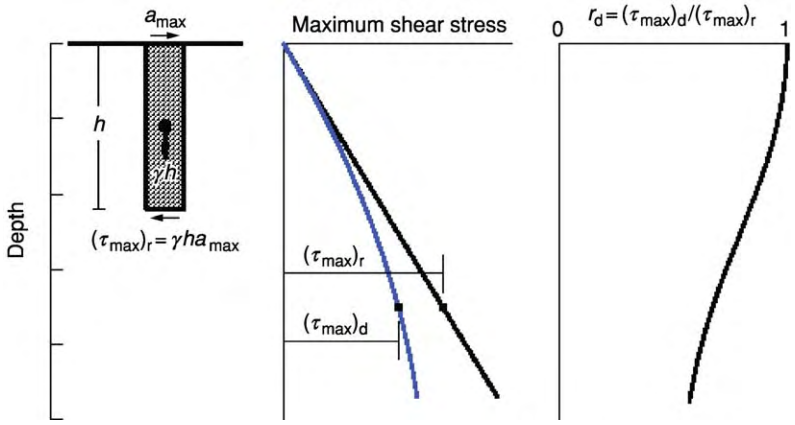


Figure 7 Schematic for determining maximum shear stress (τ_{max}) and the stress reduction coefficient (r_d). Reproduced with permission from Seed HB and Idriss IM (1971) Simplified procedure for evaluating soil liquefaction potential. *Journal of Soil Mechanics and Foundations Division, ASCE* 97: 1249 1273.

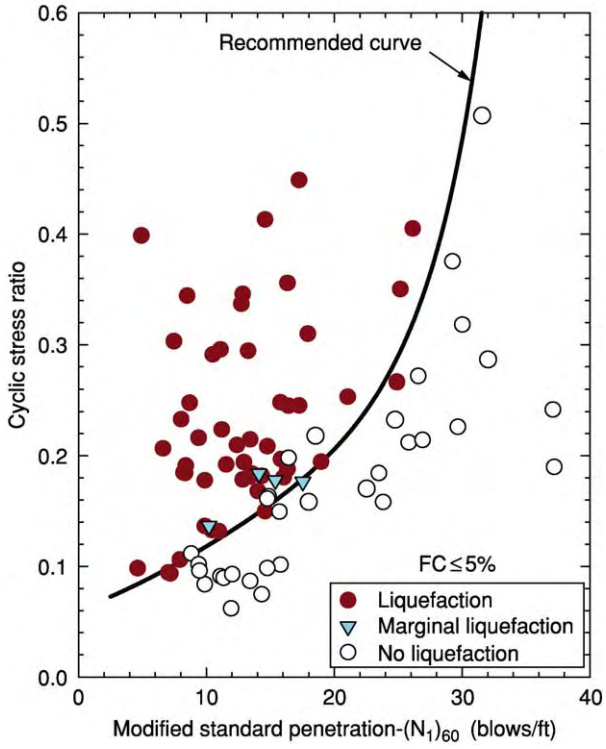


Figure 8 Correlation between cyclic shear stress ratio (for an earthquake of magnitude 7.5) causing liquefaction and standard penetration resistance for clean sands (FC, fines content ≤5%).

Figure 9 The failure of the upstream slope of the Lower San Fernando Dam shortly after the 1971 San Fernando earthquake. California narrowly avoided catastrophic flooding of 80 000 down stream homes. Liquefaction of the hydraulic fill caused loss of strength and slope instability, which left less than 1 m of free board after the earthquake. This picture shows the slide after emptying the reservoir; the paved road to the left indicates the former dam crest. Photograph by Leslie F. Harder, Jr.

and superstructures, and soils and embedded structures (e.g., pipelines, tunnels, tanks) will determine whether the potential for liquefaction-induced damage is unacceptable.

Flow failure Liquefaction flow failure occurs when the static shear stresses exceed the residual shear strength of a liquefied soil. Normally such failures occur on sloping ground that was stable in its static condition but became unstable due to the reduced shear resistance of the liquefied soil. Once movement is mobilized, displacements can be very large (up to tens of metres) and very rapid. The potential for flow failure can be assessed using standard slope stability analyses, substituting the residual undrained shear strength (discussed later) of the liquefied layer for its static properties, where a factor of safety below 1 indicates a flow failure hazard.

Lateral spreading Lateral spreading is also a downslope failure mechanism, related to cyclic mobility (Figure 9). The disturbing forces are a combination of the gravitational static downslope forces and the inertial loads generated by the earthquake. Lateral spreads can occur on gentle slopes or where there is a free face. The factor of safety against slope failure may remain above 1, and the ground deformation is a result of the progressive movement of surface layers as a result of the oscillation of the ground. Lateral spread-induced movements of as much as 10 m have been observed in past earthquakes.

The amount of horizontal movement can be estimated using relationships developed from empirical data, soil mechanics theory, or numerical modelling. Simplified relationships developed using empirical data are the simplest to employ, but users should consider the uncertainty associated with any simplification of this very complex phenomenon; simplifications necessarily neglect three-dimensional effects, local effects, or redistribution of pore water pressures, for example.

Ground oscillation Ground oscillation, another form of lateral ground deformation, can occur where the underlying soil has liquefied but there is no slope or free face for permanent lateral deformations to occur. Ground oscillation will manifest itself as large-amplitude transient ground waves with little or no resultant permanent deformation.

Settlement Liquefaction is a result of the tendency for saturated granular soils to densify under earthquake shaking. The eventual manifestation of this behaviour is settlement at the ground surface as excess pore pressures dissipate after the earthquake. Settlements are usually estimated using free-field, one-dimensional relationships for the volumetric strain induced in the soil as a function of both the relative density of the soil and the maximum shear

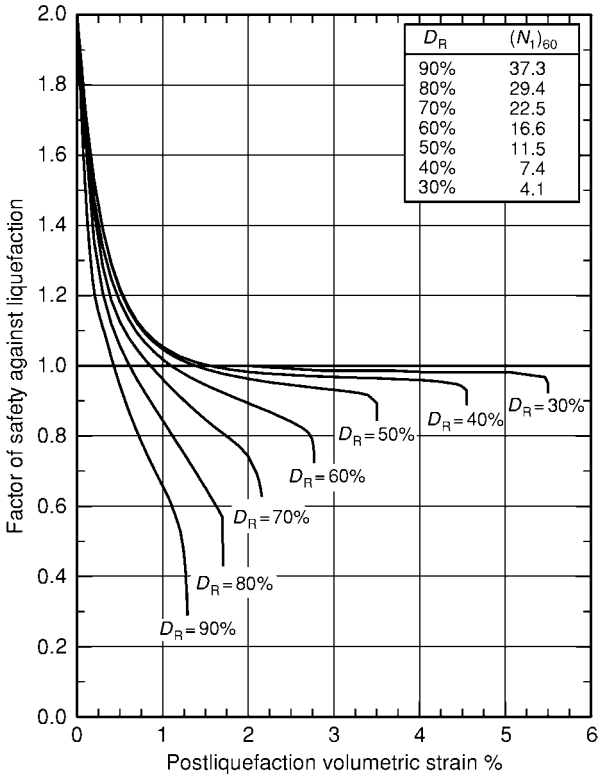


Figure 10 Chart for the estimation of volumetric strain as a function of the factor of safety against liquefaction (F; eqn [2]) and the relative density of the liquefied soil layer (D_R). $(N_1)_{60}$ is the standard penetration test blowcount normalized to an equivalent overburden pressure of 1 atmosphere and hammer energy of 60%.

Figure 11 Tilted buildings with shallow foundations in Niigata, Japan (1964) as a result of liquefaction induced bearing capacity failure.

Figure 12 An elevated transport link in Kobe following the 17 January 1995 earthquake. The bridge piers were piled, and the structure is generally undamaged despite the settlement of the surrounding soil by up to 1 m.

strain induced by the earthquake (Figure 10). In practice, the degree of settlement is complicated by the heterogeneity of the soil, the interaction between vertical and lateral movements, and the presence of the structure, and most simplified methods for predicting volumetric strains have quite high associated uncertainties.

Bearing failure If the residual shear strength of the liquefied soil beneath is sufficiently low, then bearing capacity failure will occur, causing the structure to settle or tilt uniformly (Figure 11). These types of failures are generally accompanied by heaving of the ground around the foundation. The potential for a bearing-capacity failure can be determined using simple static bearing-capacity formulas, substituting the residual undrained shear strength of the liquefied layers. Failure will occur if the factor of safety is less than unity. However, structures that are safe against bearing failure can still develop excessive settlements (total and differential), depending on the strains (shear and volumetric) that develop in the underlying soils.

Residual undrained shear strength The residual shear strength of a liquefied soil can be determined in the laboratory or can be estimated using correlations between *in situ* test data and liquefied shear strength obtained through back analysis of field case histories. Field data, compared to laboratory tests, tend to indicate much lower shear strengths. As well as the difficulty in obtaining true, undisturbed samples, some phenomena cannot be replicated in laboratory tests. The layered nature and contrasting permeability of *in situ* soil deposits can impede the flow of water as earthquake-induced excess pore water pressures dissipate; this can lead to localized weakening of the soil at permeability interfaces (see Figures 4 and 12). A further *in situ* phenomenon that cannot be replicated by testing small samples is the potential intermixing of soil layers with different characteristics, due to the shear deformation, which

can reduce the shear strength. Field data, however, are implicitly likely to include such phenomena. Consequently, the use of the empirical correlations to estimate postliquefaction shear strengths is recommended.

Consequences of Earthquake-Induced Liquefaction

The importance of liquefaction from an engineering aspect is damage to the built environment, as represented by the following examples:

- Settlements, which can be of the order of 5% of the thickness of the liquefied soil, causing uniform or differential settlement of foundations, pipelines, or transportation routes.
- Damage to piled foundations as result of loss of support and flow of soil around piles that extend into or through the liquefied soil layer(s).
- Loss of bearing support to shallow or piled foundations causing settlement or tilt.
- Lateral spreads damaging building foundations, bridge piers, highways and railways, river banks, or embankment dams or pipelines.
- Lateral flow failures.
- Embankment failures due to liquefaction of underlying material or embankment fill.
- Service interruption to buildings due to connection damage caused by foundation displacements.
- Increased lateral pressures against retaining structures such as quay walls.
- Induced hazards, such as fire caused by gas pipeline rupture, spread of fire due to interrupted water supply, or floods caused by dam failure.

The severity of the damage is dependent on the strength and duration of the earthquake ground shaking, the thickness of the liquefied layer, the material properties of the liquefied soil, and the proximity of free faces such as slopes or retaining structures, as well as other variables related to the structures. The response of a building to liquefaction-induced vertical or lateral permanent ground deformation depends to a large extent on the building's foundations (Figures 13–15). Bridge embankments are vulnerable to lateral spreading due to the combination of susceptible river channel deposits and sloping or free faces. Bridge piers in liquefied ground can rotate, twist, or displace laterally. In extreme cases, when liquefaction is combined with poor design in the form of insufficient bearing width of the bridge deck, excessive movement can cause loss of support to the bridge deck and hence collapse (Figure 15). Damage to roadways or railways due to liquefaction most commonly comprises settlement or lateral spreading of embankments, highway fills, or natural soil, which causes cracking and uneven surfaces. Such damage is relatively easy and inexpensive to repair but, nonetheless, the disruption and the indirect losses must be considered.

The performance of a region's lifelines networks, including water, electricity, and gas distribution systems, is an essential factor in the immediate emergency response and the subsequent recovery from the earthquake impact. Pipeline damage can be induced both by transient ground oscillations and by vertical

Figure 13 Tilted apartment block in Adapazari, Turkey, due to extensive liquefaction in the 1999 magnitude 7.4 Kocaeli earth quake. This building had thick, continuous, shallow foundations, which caused it to rotate as a rigid body. There is very little damage to load bearing elements, walls, or ceilings; nonethe less, the building would have been demolished. Photograph by Beyza Taskin.

Figure 14 Damage to the Marine Laboratory at Moss Landing, Monterey Bay, California, during the 1989 magnitude 6.9 Loma Prieta earthquake. Lateral spreading caused the building pad foundations to spread apart. The laboratory was subsequently demolished. Photograph by Leslie F. Harder, Jr.

or lateral permanent ground deformations, which can shear, compress, or pull apart pipelines at their joints. Foundation displacements are likely to damage pipeline connections to buildings, which can lead to building shutdown even in the absence of structural damage. Ruptured gas and water pipelines can trigger fires and impede the control of fires, respectively, creating a significant induced hazard related to liquefaction damage to pipelines. Buried pipelines, as well as other buried vessels such as tanks, can become buoyant in liquefied soils. Ports and harbours tend to be particularly vulnerable to liquefaction damage due to the combination of high water tables, reclaimed soils, and retained vertical faces. Dam failures due to liquefaction at the toe, or flow failure of the embankment, can cause catastrophic damage and massive loss of life as a result of the flooding induced by such a failure (see Figure 9). Mitigation of liquefaction hazard in the design of dams is therefore of the utmost importance.

Figure 15 This span of the Nishinomaya Bridge collapsed following the 1999 magnitude 6.9 Kobe earthquake in Japan as a result of liquefaction related foundation deformations. Ground cracks behind the quay walls and parallel to the water edge are indicative of the lateral ground movements that occurred. Sand boils are visible on the ground surface. Photograph by Leslie F. Harder, Jr.

Mitigation

When the potential liquefaction risk is considered to be unacceptable for the performance requirements of an engineering project, the three principal options are relocation, prevention, and effective design. Relocating a structure or facility to avoid susceptible zones may be the most straightforward option in terms of mitigating the effect of liquefaction. For regionally distributed facilities such as lifeline or transportation networks, this option is unlikely to be practicable. The use of ground improvement to limit or prevent the occurrence of liquefaction has been shown to be effective in many past earthquakes. Ground improvement may involve increasing the density of the liquefiable soil through compaction, vibration, or replacement; reinforcing and densifying the soil through jet grouting or deep soil mixing; or providing additional drainage to allow excess pore water pressures to dissipate more rapidly (Figure 16). Often a combination of several techniques is adopted. Again, this is unlikely to be an appropriate solution for regionally distributed facilities.

Accommodation through design is the third option to manage liquefaction risk. Foundations can be designed to withstand expected ground deformations

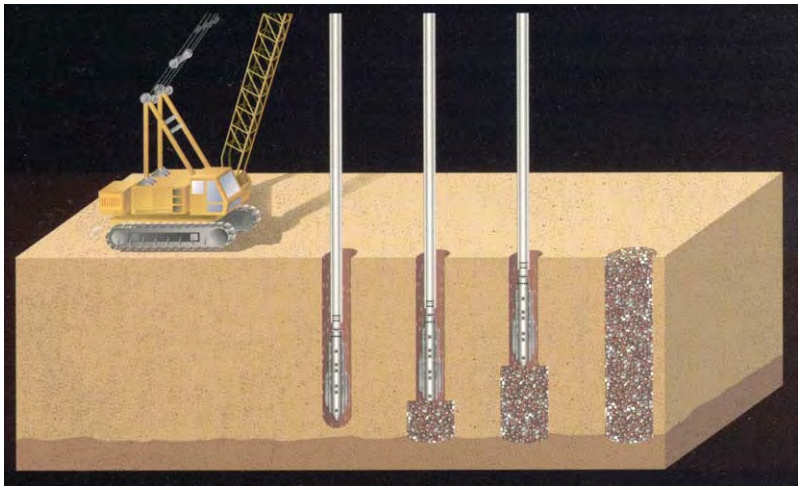


Figure 16 Installation of stone columns to mitigate liquefaction through densification and strengthening of liquefiable layers (courtesy of Hayward Baker, Inc.).

and to ensure that movements are not translated to the superstructure. Piled foundations can be designed to accommodate additional lateral loads imparted by soil movement, or to have sufficient vertical capacity even in the case of negative skin friction due to settlement. Shallow foundations can be designed to be sufficiently strong so as to behave as a rigid body when subjected to ground deformations. In the case of lifeline networks, the accommodation of the expected movements, and the implementation of appropriate response and repair measures, are frequently the only available solution.

See Also

Engineering Geology: Aspects of Earthquakes; Seismology; Site and Ground Investigation; Site Classification. **Geotechnical Engineering**. **Soil Mechanics**. **Tectonics:** Earthquakes.

Further Reading

Berrill J and Yasuda S (2002) Liquefaction and piled foundations: some issues. *Journal of Earthquake Engineering* 6(Special Issue 1): 1 41.

Boulanger RW and Idriss IM (2004) *State Normalization of Penetration Resistance and the Effect of Overburden on Liquefaction Resistance,* pp. 484 491. 11th International Conference on Soil Dynamics and Earthquake Engineering and 3rd International Conference on Earthquake Geotechnical Engineering, University of California, Berkeley.

Harder LF, Jr (1997) *Application of the Becker Penetration Test for Evaluating the Liquefaction Potential of Gravelly Soils*. Proceedings of the NCEER Workshop on Evaluation of Liquefaction Resistance of Soils, NCEER 97 0022.

Idriss IM and Boulanger RW (2004) *Semi empirical Procedures for Evaluating Liquefaction Potential during Earthquakes,* pp. 32 56. 11th International Conference on Soil Dynamics and Earthquake Engineering and 3rd International Conference on Earthquake Geotechnical Engineering, University of California, Berkeley.

Ishihara K (1993) Liquefaction and flow failures during earthquakes. (Rankine Lecture). *Geotechnique* 43: 351 415.

Ishihara K and Yoshimine M (1992) Evaluation of settlements in sand deposits following liquefaction during earthquakes. *Soils and Foundations* 32(No. 1): 173 188.

Kramer SL (1996) Geotechnical earthquake engineering. Upper Saddle River, NJ: Prentice Hall.

National Research Council (1985) *Liquefaction of Soils during Earthquakes. Report by the Participants in the Workshop on Liquefaction, Committee on Earthquake Engineering, National Research Council, Report No. CETS EE 001*. Washington, DC: National Academy Press.

Seed HB (1987) Design problems in soil liquefaction. *Journal of Geotechnical Engineering, ASCE* 113(8): 827 845.

Seed RB and Harder LF, Jr (1990) SPT based analysis of cyclic pore pressure generation and undrained residual strength. In: *H. Bolton Seed Volume 2. Memorial Symposium Proceedings,* pp. 351 376. Vancouver, BC: BiTech Publishers Ltd.

Seed HB and Idriss IM (1971) Simplified procedure for evaluating soil liquefaction potential. *Journal of Soil Mechanics and Foundations Division, ASCE* 97: 1249 1273.

Seed HB and Idriss IM (1982) *Ground Motions and Soil Liquefaction during Earthquakes. Monograph No. 5*. Oakland, CA: Earthquake Engineering Research Institute (Note: A new edition of this monograph is expected in 2005).

Tokimatsu K and Seed HB (1987) Evaluation of settlements in sand due to earthquake shaking. *Journal of Geotechnical Engineering, ASCE* 113(8): 861 878.

Youd TL, Hansen CM, and Bartlett SF (2002) Revised multilinear regression equations for prediction of lateral spread displacement. *Journal of Geotechnical and Geoenvironmental Engineering, ASCE* 128(12): 1007 1017.

Youd TL and Idriss IM (2001) Liquefaction resistance of soils: Summary report from the 1996 NCEER and 1998 NCEER and NSF workshops on evaluation of liquefaction resistance of soils. *Journal of Geotechnical and Geoenvironmental Engineering, ASCE* 127(4): 297 313.

Youd TL and Perkins DM (1978) Mapping liquefaction induced ground failure potential. *Journal of the Geotechnical Engineering Division, ASCE* 104(GT4): 433 446.

Made Ground

J A Charles, Formerly Building Research Establishment, Hertfordshire, UK

Introduction

The term 'made ground' is used to describe ground that has been formed by human activity rather than by natural geological processes. The material of which made ground is composed is described as 'fill', and in practice the terms 'made ground' and 'fill' are often used interchangeably. Fill materials include not only natural soils and rocks but also the waste products of mining and industrial processes, and commercial and domestic refuse.

Throughout history, mankind has deliberately adjusted the topography of the Earth by excavating soil and rock and placing the excavated material in more convenient locations. The casual disposal of waste materials has also changed landforms in urban areas. These processes greatly accelerated during the twentieth century, and in some localities, such as major conurbations and mining areas, a significant proportion of the surface area is now made ground. It is therefore important that the extent, depth, and nature of such deposits are reliably mapped and that the geotechnical behaviour of the fill materials is adequately understood.

A simple classification system has been developed for describing ground affected by human activity, and this classification is used in the geological mapping of the UK.

- Made ground – areas where it is known that fill material has been placed.
- Worked ground – areas where excavations have been made in natural ground.
- Disturbed ground – areas of surface and near-surface disturbance (including ground that has subsided), typically associated with mining.
- Landscaped ground – areas where the ground surface has been remodelled, but made ground and worked ground cannot be distinguished.

A broad distinction can be made between 'engineered fill' and 'non-engineered fill'. The distinction is essentially one of purpose: engineered fill is placed for a specific purpose, whereas non-engineered fill is a by-product of the disposal of waste material. Engineered fill is selected, placed, and compacted to an appropriate specification in order that it will exhibit the required engineering behaviour; engineering design focuses on the specification and control of filling. Non-engineered fill is placed with no subsequent engineering application in view. The distinction between engineered and non-engineered fill is clear in principle, but not always in practice.

History

Early earthmoving activities were undertaken to meet practical objectives, although in some cases that objective is not obvious to the modern mind. Large-scale earthmoving has provided many durable reminders of previous ages.

Silbury Hill, situated in Wiltshire in the south of England, is the largest prehistoric manmade mound in Europe and a remarkable early civil-engineering achievement. Archaeological evidence suggests that the 40 m high mound (Figure 1) is more than 4000 years old. It was carefully engineered in a series of six stepped horizontal layers created from concentric rings of chalk block walls. There are radial as well as circumferential walls, and the compartments between them are filled with chalk rubble. The steps in the outer slope were infilled to give a smooth slope. It has been suggested that the hill is a burial mound, but no graves have been found.

One practical reason for placing fill is to increase the area of land suitable for building on within the confines of a town or city. This has been achieved in hill country by cut-and-fill earthworks on hillsides and in low-lying areas by raising the level of marshy ground. Some 3000 years ago rock-fill platform terracing was formed on the eastern slopes of Jerusalem, thereby substantially increasing the building area. This is thought to be King Solomon's *Millo*, the

Figure 1 Silbury Hill, England.

construction of which is referred to in *I Kings* 9 : 15,24 and 11 : 27.

The construction of embankment dams to retain reservoirs of water has a long history. The 14 m high Sadd-el-Kafara was built with engineered fill in Egypt at the beginning of the age of the pyramids and impounded $0.5 \times 10^6\,m^3$ of water. The dam had a central core of silty sand supported by shoulders of rock fill. Inadequate provision was made for floods, and the dam was probably overtopped and breached during a flood.

Fills were placed to provide a suitable elevation for defence or for control of the local population. The mound for Clifford's Tower was built in York in 1069 by William I during his campaign to subdue the north of England. The 15 m high mound, which was built of horizontal layers of fill comprising stones, gravel, and clay, was founded on low-lying ground to provide a suitable elevation for the construction of a stronghold. The original tower was made of timber, and the present stone structure known as Clifford's Tower dates from the middle of the thirteenth century. Soon after the erection of the stone tower, severe floods in 1315–1316 softened the fill in the mound, and in 1358 the tower was described as being cracked from top to bottom in two places. These cracks, which were repaired at great expense before 1370, are still visible.

In many parts of the world, low-lying wet ground has been reclaimed by filling, and in the last few centuries this type of land reclamation has taken place on a large scale. Reclamation of the marshes bordering the Baltic Sea, on which St Petersburg is built, began in 1703. On the opposite side of the Atlantic Ocean, much of downtown Manhattan was built on made ground created before 1900. When present-day maps of cities are superimposed on old maps, the extent of such made ground is revealed.

In urban locations where the land has been continuously occupied for centuries, there are likely to be large areas of made ground. Fills have arisen inadvertently from the rubble of demolished buildings and the slow accumulation of refuse. Made ground of this type may contain soil, rubble, and refuse and may be very old. It can be quite extensive in area but is often relatively shallow. Some towns in the Middle East provide examples of the unplanned accumulation of fills. The most common building material was mud brick, and so the walls had to be thick. New construction took place on the ruins of older buildings, and in Syria and Iraq villages stand on mounds of their own making. The ruins of an ancient city may rise 30 m above the surrounding plain.

This gradual rise of debris has been much less common in Great Britain, although in some locations deep fills have accumulated. By the third century AD the Wallbrook in the City of London was already half buried, and mosaic pavements of Roman London lie 8–9 m below the streets of the modern city.

The Royal Scottish Academy in Edinburgh was completed in 1826 on the Mound, which was formed in the late 1700s using clay spoil from the construction of the New Town. The building was founded on square timber piles that, in the course of time, rotted because they were above the water-table, leaving large voids under the stone footings. Remedial works involving compensation grouting were carried out recently.

With the coming of the Industrial Revolution mankind's capacity to generate waste materials, and to cover significant portions of the Earth's surface with them, greatly increased. Where minerals were extracted from underground workings, it was impracticable to avoid extracting quantities of other materials with the desired mineral, and the resulting spoil was brought to the surface and placed in heaps.

The need to supply unpolluted water to the rapidly expanding industrial cities in the north of England led to the construction of large numbers of embankment dams in the nineteenth century. Dale Dyke was one of the dams that was built to supply water to Sheffield. The 29 m high embankment followed the traditional British form of dam construction, with a narrow central core of puddle clay forming the watertight element. The reservoir capacity was $3.2 \times 10^6\,m^3$. By 10 March 1864, during the first filling of the reservoir, the water level behind the newly built dam was 0.7 m below the crest of the overflow weir. In the late afternoon of 11 March 1864 a crack was observed along the downstream slope near the crest of the dam. At 23.30 the dam was breached, and the resulting flood destroyed property estimated to be worth half a million pounds sterling, and caused the loss of 244 lives. Developments in geotechnical engineering in the twentieth century have enabled safe embankment dams to be built with confidence.

Twentieth Century

The twentieth century saw a massive expansion of made ground. Large-scale earthmoving machinery made it possible to place fill rapidly and cheaply in quantities never before experienced. This applied both to engineered fills placed to construct embankment dams, road embankments, and sites for buildings, and to non-engineered fills placed as mining, industrial, chemical, building, dredging, commercial, and domestic wastes. Table 1 provides details of some fills placed over the last 4000 years and illustrates how the

Table 1 Some examples of made ground

Structure	*Location*	*Purpose*	*Date built*	*Height or depth (m)*	*Volume (10^6 m^3)*	*Surface area (ha)*
Silbury Hill	Wiltshire, England	Unknown	pre 2000 BC	40	0.25	2
Sadd el Kafara	Egypt	Retain water	pre 2000 BC	14	0.09	2
Clifford's Tower mound	York, England	Military	AD 1069	15	0.04	0.4
Dale Dyke dam	Sheffield, England	Retain water	AD 1864	29	0.4	4
Fort Peck dam	Montana, USA	Retain water	AD 1940	76	96	200
Scammonden dam	Huddersfield, England	Retain water + motorway embankment	AD 1969	76	4.3	15
Tarbela dam + upstream blanket	Pakistan	Retain water	AD 1976	148	118	700
Nurek dam	Tajikistan	Retain water	AD 1980	300	58	60
Gilow impoundment	South western Poland	Retain copper tailings	AD 1980	22	68	540
Lounge opencast site[a]	Ashby de la Zouch, England	Engineered fill for road embankment	AD 1990	40	3.7	15
Dixon opencast site	Chesterfield, England	Backfill	AD 1992	74	16	119
Kansai airport (Phase 1)	Osaka, Japan	Island for airport	AD 1994	33	180	510
Chek Lap Kok airport	Hong Kong	Platform for airport	AD 1996	25	194	1248

[a]Data refers only to that part of the backfill that was placed as an engineered fill.

scale of operations vastly increased in the twentieth century. Cross-sections of a number of the earth structures included in Table 1 are shown in Figure 2.

Although urban redevelopment has continued throughout history, modern programmes of urban regeneration are carried out at a rate and on a scale not seen before. Much of this redevelopment is carried out on made ground.

The reclamation of land from the sea can be achieved either by the construction of water-retaining embankments, which prevent the sea flooding land below sea-level, as in the Netherlands, or by the placement of fill to form made ground whose surface is above sea-level. Examples of the latter approach include the massive reclamation projects carried out in Hong Kong, Japan, and Singapore for the construction of new airports. Kansai airport is located 5 km off the Japanese mainland, and the placement of a 33 m thickness of fill to form the island has caused the underlying deposits to settle by 14 m. The construction of Phase 2 of Kansai airport has commenced, and this will increase the surface area of the made ground to 1100 ha.

The number and size of embankment dams increased greatly during the twentieth century. Fort Peck dam in the USA was built of hydraulically placed earthfill and consequently has flat slopes and contains a large volume of fill. When the embankment was nearly complete, there was a slide in the upstream slope involving 4×10^6 m^3 of fill. At the time of its completion in 1940 it was the largest dam in the world, and it remained so until Tarbela dam was completed in 1976 in Pakistan. Tarbela has an impervious upstream blanket, which continues for 2 km upstream of the upstream toe of the dam. The 300 m high Nurek dam in Tajikistan was completed in 1980 and is one of the world's highest dams.

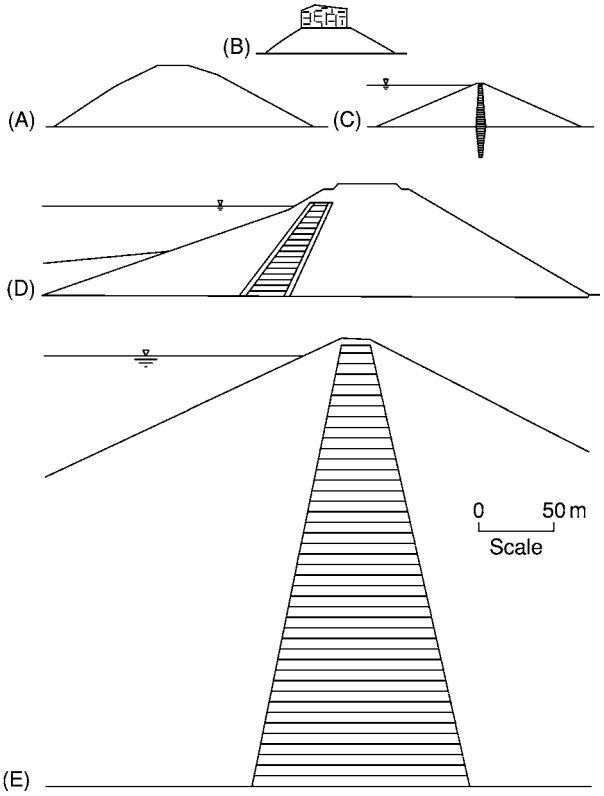


Figure 2 Fill structures: (A) Silbury Hill, England; (B) Clifford's Tower mound, England; (C) Dale Dyke dam, England; (D) Scam monden dam, England; and (E) Nurek dam, Tajikistan.

The 76 m high Scammonden dam, which was completed in 1969, is the highest embankment dam in England. The embankment was designed to serve the dual purpose of impounding a reservoir with a capacity of $7.9 \times 10^6\,m^3$ and carrying the M62 trans-Pennine motorway across the Scammonden valley. A cross-section of the embankment is shown in Figure 2D. The shoulders were built of compacted sandstone and mudstone rock fill, and it has an upstream-sloping rolled clay core in the upstream part of the embankment.

In metal mining, rock is crushed to extract the desired mineral, leaving large amounts of crushed rock as a waste material. The fine waste material, which is known as tailings, is discharged from the wet process as a saturated slurry and commonly pumped through a pipeline from the plant to an impoundment formed by an embankment dam. Large waste-impounding embankments have been built in the twentieth century, some more than 150 m high and some impounding more than $10^8\,m^3$ of waste materials. The Gilow impoundment in Poland has a surface area of 540 ha. The principal hazard posed by such dams and their retained waste impoundments is the risk of rapid discharge of the impounded waste material in an uncontrolled manner should there be a breach of the embankment. There have been a number of failures of large tailings dams in recent years, and some waste materials pose major hazards to the natural environment; for example, the breach on 24 April 1998 of the Aznalcollar tailings dam at Seville in Spain released $7 \times 10^6\,m^3$ of mining waste and water, threatening the Coto de Donana National Park.

Where iron ore, coal, and other minerals are found in thin seams not too far below the surface, they can be won by opencast or strip mining; that is, the overlying soil or rock is excavated to reach the mineral seams without the need for subsurface tunnelling. When a mineral has been extracted by opencast mining, the overburden soils and rocks are replaced in the excavation, and these mining operations have been major producers of deep non-engineered fills in many parts of the world. By 1986 more than 1000 residential buildings and farms had been established on deep uncompacted backfills in the Rhenish brown coal area of Germany. During the 1980s opencast coal production in Great Britain was about 14 million tonnes per annum, and the extraction of one tonne of coal typically involved the excavation of $15\,m^3$ of overburden. Sites were often restored to agricultural use without compaction during backfilling of the opencast excavation, but latterly systematic compaction has become more common. Two British opencast sites are included in Table 1.

Waste from the deep mining of coal, which is known as colliery spoil, is derived from the rocks adjacent to the coal seams, and during mining operations quantities of these rocks, unavoidably extracted with the coal or in driving the tunnels that give access to the coalface, are brought to the surface. Towards the end of the twentieth century, world coal production approached 5×10^9 tonnes per annum.

The coarse discard from coal mining used to be dumped in heaps, which could become very large. Following the Aberfan disaster in Wales in 1966, when the failure of an unstable colliery spoil heap caused great loss of life, it became the usual practice in Great Britain to place the spoil in thin layers with compaction. Geotechnical problems can be largely overcome by adequate compaction during placement. By 1974 there were 3×10^9 tonnes of colliery spoil in Great Britain, and in 1984 the coal mining industry produced 5×10^7 tonnes of coarse spoil annually, which was placed in tips adjacent to the collieries. In April 1988 there were 4700 ha of derelict land associated with colliery spoil heaps in England alone. The rapid decline of the coal industry in the 1990s meant that the annual production of these wastes reduced, but large stocks of colliery spoil remain in the coalfields. In 1996 it was planned to reclaim 900 ha of colliery land for residential, commercial, and retail uses.

A major proportion of the domestic waste generated in the UK is disposed of in landfill. In 1986, over 90% of controlled domestic, commercial, and industrial solid wastes (excluding mining and quarrying wastes) were disposed of by landfilling. In the year 1986/1987 nearly 2×10^7 tonnes of household or domestic waste were disposed of in England and Wales, and this figure did not change substantially over the subsequent 10 years. Despite environmental initiatives, landfilling has continued on a large scale. In the USA at the end of the twentieth century 1.2×10^8 tonnes of municipal solid waste were landfilled each year.

Functions of Made Ground

The function of non-engineered fills is to dispose of unwanted waste material. By contrast, the large quantities of fill material that are placed as engineered fills form part of carefully controlled civil-engineering works. Three major types of engineered fills are embankment dams, road embankments, and fills that support buildings.

As we have seen, embankment dams are constructed from fill and are built to retain reservoirs of water, which may be required for hydropower, water supply to towns, irrigation, or flood control. Similar

embankments can be used to retain canals. An embankment dam will usually be composed of several different types of fill: a low-permeability fill to form the watertight element, stronger fill to support the watertight element, and fills to act as filters, drains, and transition materials. Possible hazards affecting the dam include internal erosion, slope instability, and overtopping during floods.

Embankment dams are also used to retain lagoons of sedimented waste material from mining and industrial activities. If the waste is not toxic, these embankments may be designed so as to allow water to drain through. The embankment may be built in stages as waste disposal progresses.

Another use for a special type of embankment is to protect harbours and the shoreline from the sea. Such embankments are often built as rubble-mound breakwaters and require high-quality quarried rock for the fill. Placement to the required profile presents obvious difficulties.

Embankments to carry roads and railways are usually built to reduce gradients. The engineered fills use material excavated from adjacent cuttings during the cut-and-fill operations, so there is limited scope for material selection and whatever is excavated has to be placed in the adjacent embankment unless it is clearly unsuitable. In England, during the construction of the M6 motorway between Lancaster and Penrith in the 1960s, much of the soil excavated along the line of the road was very wet, and a geotechnical design was developed that involved the use of drainage layers built into the embankment during construction.

Buildings may be founded on made ground. Old excavations are infilled, and sometimes embankments are built above the level of the surrounding ground. The objective is to support buildings safely while minimizing the risk of damaging settlement, and, where structures sensitive to settlement are to be built on made ground, a high-quality fill that is not vulnerable to large post-construction movement is required.

Opencast-mining sites have often been restored for agricultural use, but where the sites are close to urban areas they may subsequently be used for housing and commercial developments. Loose backfill usually has considerable settlement potential, and some existing areas of loose fill have been improved by preloading with a temporary heavy surcharge of fill. Because of the free-draining nature of the loose fill, it soon consolidates, and this type of treatment has made some sites quite suitable for normal housing. Where building is foreseen prior to backfilling an opencast mining site, the fill material should be placed in thin layers and heavily compacted as an engineered fill.

Fill Placement

There are two basic elements in the quality management of engineered fills: placement of a fill with the required quality; and evidence that the fill has the required quality. Both an appropriate specification and rigorous quality-control procedures are required. It is not possible to prevent some variability in the made ground, as there will be a degree of heterogeneity in the source material and some segregation during placement. It is necessary to determine how the required properties can be achieved with an acceptable degree of uniformity.

Placement in thin layers with heavy compaction at an appropriate water content is the method usually adopted to obtain the required performance from an engineered fill used in dam, road, or foundation applications. Figure 3 shows the compaction of a clay fill, and Figure 4 shows a rock fill being watered during placement. There are three basic approaches to the specification of engineered fills.

Figure 3 Compacting clay fill.

Figure 4 Watering rock fill during placement.

- In a method specification, the procedure for placement and compaction is described. The type and mass of the compactor, the number of passes, and the layer thickness are specified. Reliance is placed on close inspection to ensure compliance with the specification.
- An end-product specification is based on required values for properties of the fill as it is placed. The basic measurements of the *in situ* state of compaction are density and water content, but these measurements on their own are not adequate indicators – the density needs to be interpreted in terms of the density at a specified water content under some standard type of compaction. For a clay fill, the specification could be in terms of percentage air voids or undrained shear strength. Compliance is tested as filling progresses.
- In a performance specification some facet of the post-construction behaviour of the fill is specified, such as a permissible post-construction settlement or a load-test result. With this approach the specification is directly related to one or more aspects of the performance requirements.

Non-engineered fills have usually been placed in the simplest and cheapest way feasible. The fill may have been tipped in high lifts with no systematic compaction. A non-engineered fill can be effectively converted to an engineered fill by *in situ* ground treatment subsequent to the completion of fill placement, although the limitations of such post-construction treatment should be recognized, since unsuitable fill material is still present after treatment and most forms of compaction applied to the ground surface are effective only to a limited depth.

Natural fine soils or waste materials can be mixed with enough water to enable them to be transported in suspension. Usually the suspension is pumped through a pipe and then discharged onto the surface being filled. The deposit is described as hydraulic fill. Figure 5 shows a lagoon of pulverized fuel ash; this waste product from coal-fired power stations has been mixed with sufficient water to enable it to be pumped to the lagoon.

Where a waste material contains particles of different sizes, segregation may occur during deposition. As the suspension flows away from the discharge point, the larger soil particles settle out almost immediately and the water and fines flow away. Eventually, the fine material also settles out but at a much greater distance from the discharge point. The placement of hydraulic sand fills beneath water involves settling from a slurry, and such subaqueous hydraulic fills can be placed by bottom dumping from a barge or by pipeline placement.

Figure 5 Lagoon filled with pulverized fuel ash.

Fill Properties and Behaviour

The required behaviour of made ground is closely linked to the purpose of the engineered fill and the processes and hazards to which it may be exposed. There may be performance requirements that can be expressed in terms of a wide range of geotechnical properties, such as shear strength, stiffness, compressibility, and permeability.

Made ground exhibits as wide a range of engineering properties as does natural ground; both the nature of the fill material and the mode of formation have a major influence on subsequent behaviour. Needless to say, there is a vast difference between the behaviour of an engineered heavily compacted sand-and-gravel fill and that of recently placed domestic refuse. Made ground may have been formed by fill placed in thin layers and heavily compacted or by fill end-tipped in high lifts under dry conditions or into standing water. The method of placement affects the density of the fill and the homogeneity of the made ground. Many of the problems with non-engineered fills are related to their heterogeneity. Depending on the method of placement and the degree of control exercised during placement, there may be variability in materials, density, and age.

The engineering properties of hydraulic fills can be expected to be very different from the properties of fills placed at lower water contents under dry conditions. Hydraulic transport and deposition of materials generally produces fills with a high water content in a relatively loose or soft condition. Segregation of different particle sizes is likely.

The amount of densification of a clayey soil that can be achieved for a given compactive effort is a function of water content. Laboratory compaction-test results are plotted as the variation of dry density with water content. Such plots are a simple and useful way of representing the condition of a partially

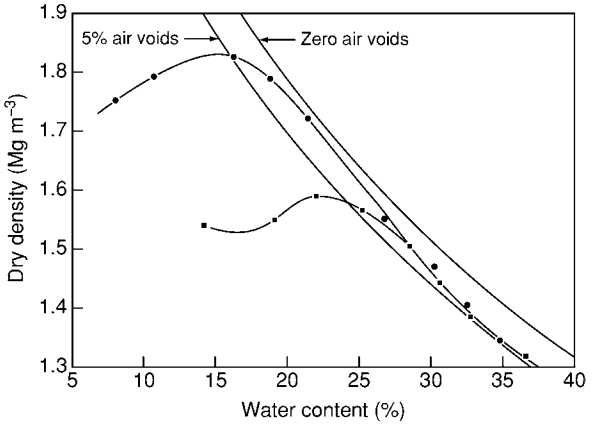


Figure 6 Laboratory compaction of clay fill using two different compactive efforts. —●—, Heavy compaction (4.5 kg rammer); —■—, Standard Proctor compaction (2.5 kg rammer).

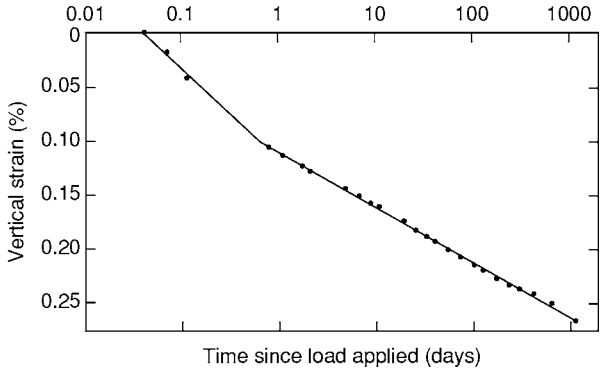


Figure 7 Creep compression of heavily compacted sample of sandstone rock fill.

saturated fill and can be used to identify two parameters of prime interest – the maximum dry density and the corresponding optimum water content. The terms maximum dry density and optimum water content refer to a specified compaction procedure and can be misleading if taken out of the context of that procedure. Figure 6 shows laboratory test results using two different compactive efforts. The heavier compaction produces a greater maximum dry density at a lower optimum water content. At a water content less than optimum, the specified compaction procedure may result in a fill with large air voids. At a water content significantly higher than optimum, the specified compaction procedure should produce a fill with a minimum of air voids, typically between 2% and 4%.

Where made ground is built upon, the magnitude and rate of long-term settlement is an important issue. Fill will undergo some creep compression, which occurs without changes in the load applied to the fill or the water content of the fill. For many fills there is a linear relationship between creep compression and the logarithm of the time that has elapsed since the load was applied. This type of relationship can be characterized by the parameter α, which is the vertical compression occurring during a log cycle of time (i.e. between, say, 1 year and 10 years). For loosely placed coarse fills, α is typically between 0.5% and 1%. For heavily compacted fills, α is much smaller and strongly dependent on stress. Figure 7 shows the rate of compression of a heavily compacted sandstone rock fill measured in a large laboratory oedometer test under a constant vertical applied stress of 0.7 MPa. From 1 day to 1000 days after the load was applied there was a linear relationship between creep compression and the logarithm of elapsed time, corresponding to $\alpha = 0.05\%$.

When first inundated, most types of partially saturated fill are susceptible to collapse compression if they have been placed in a relatively loose or dry condition. This reduction in volume can occur under a wide range of stresses and without a change in applied total stress. For buildings on fill, the risk of collapse compression is normally the main concern; where it occurs after construction has taken place, buildings may be seriously damaged. The causes of collapse compression on inundation fall into three categories: weakening of interparticle bonds; weakening of particles in coarse fills; and softening of lumps of clay in fine fills. Inundation can result from a rising groundwater table or from downward infiltration of surface water. An important objective of the specification and control procedures adopted for fills is to eliminate or at least minimize the potential for collapse compression.

Future Trends

It may be questioned whether the massive increase in the generation of fill materials that occurred in the twentieth century can or should continue. Large-scale earthmoving, which is visually intrusive and involves major modifications to the environment, is meeting increasing opposition, particularly where large dams are built. The principal objection to new dams has centred on the flooding of large tracts of land, but all engineering projects involving major earthmoving are likely to face similar opposition. While it is not unreasonable for earthmoving projects to be subjected to close scrutiny, it is important that the benefits conferred by many types of made ground are not overlooked.

Carefully controlled engineered fill is likely to form only a small fraction of new made ground, the bulk of which will be created by the landfilling of various types of waste material. Every year, billions of tonnes

of solid waste are generated by industrial and mining activities and deposited either as spoil heaps on the surface of the ground or as waste dumps within excavations. Mining has brought much prosperity to human societies, but it is now widely questioned whether these activities, which are being carried out on an unprecedented scale, are sustainable and whether such massive quantities of waste are environmentally acceptable. Despite the huge rate of consumption during the twentieth century, the known resources of many minerals have increased rather than diminished as new sources have been identified. It seems probable that, for the foreseeable future, the needs of successive generations will require the extraction of minerals from the surface of the Earth on a large scale and that there will be sufficient resources to meet this demand. These mining operations will mean that spoil heaps and waste dumps will cover ever greater portions of the Earth's surface.

While increased concern over environmental issues is not likely to halt the industrial and mining activities that generate waste, it will have a major effect on waste disposal. In most countries, plans for reclamation of the land must be approved before mining begins, and it is usually required that the land will be restored as closely as possible to its original state. Environmental factors will increasingly influence how, where, and in what form waste fills are deposited and the treatment that such deposits receive. There will be increasing pressure to reduce the amount of domestic waste that is placed as landfill.

See Also

Engineering Geology: Geological Maps; Natural and Anthropogenic Geohazards; Liquefaction; Site and Ground Investigation; Subsidence. **Environmental Geology**. **Geotechnical Engineering**. **Soil Mechanics**. **Urban Geology**.

Further Reading

Charles JA and Watts KS (2001) *Building on Fill: Geotechnical Aspects*, 2nd edn. Building Research Establishment Report BR 424. Garston: BRE Bookshop.

Charles JA and Watts KS (2002) *Treated Ground: Engineering Properties and Performance*. Report C572. London: CIRIA.

Clarke BG, Jones CJFP, and Moffat AIB (eds.) (1993) *Engineered Fills*. Proceedings of conference held in Newcastle upon Tyne, September 1993. London: Thomas Telford.

Grace H and Green PA (1978) The use of wet fill for the construction of embankments for motorways. In: *Clay Fills*. Proceedings of conference held in London, November 1978, pp. 113 118. London: Institution of Civil Engineers.

Johnston TA, Millmore JP, Charles JA, and Tedd P (1999) *An Engineering Guide to the Safety of Embankment Dams in the United Kingdom*, 2nd edn. Building Research Establishment Report BR 363. Garston: BRE Bookshop.

Lomborg B (2001) *The Skeptical Environmentalist: Measuring the Real State of the World*. Cambridge: Cambridge University Press.

Matsui T, Oda K, and Tabata T (2003) Structures on and within man made deposits Kansai airport. In: Vanicek I, Barvinek R, Bohac J, Jettmar J, Jirasko D, and Salak J (eds.) *Geotechnical Problems with Man made and Man Influenced Ground*. Proceedings of the 13th European Conference on Soil Mechanics and Geotechnical Engineering, Prague, vol. 3, pp. 315 328. Prague: Czech Geotechnical Society.

Parsons AW (1992) *Compaction of Soils and Granular Materials: A Review of Research Performed at the Transport Research Laboratory*. London: HMSO.

Penman ADM (2002) Tailings dam incidents and new methods. In: Tedd P (ed.) *Reservoirs in a Changing World*. Proceedings of the 12th Conference of the British Dam Society, Dublin, pp. 471 483. London: Thomas Telford.

Perry J, Pedley M, and Reid M (2001) *Infrastructure Embankments Condition Appraisal and Remedial Treatment*. CIRIA Report C550. London: CIRIA.

Proctor RR (1933) Fundamental principles of soil compaction. *Engineering News Record* 111: 245 248.

Rosenbaum MS, McMillan AA, Powell JH, *et al*. (2003) Classification of artificial (man made) ground. *Engineering Geology* 69: 399 409.

Schnitter NJ (1994) *A History of Dams: The Useful Pyramids*. Rotterdam: Balkema.

Trenter NA and Charles JA (1998) A model specification for engineered fills for building purposes. *Proceedings of the Institution of Civil Engineers*, *Geotechnical Engineering* 119: 219 230.

Problematic Rocks

F G Bell, British Geological Survey, Keyworth, UK

Introduction

As far as engineering behaviour is concerned, a distinction has to be made between rock as a material and the rock mass. 'Rock' usually refers to the intact rock, which may usually be considered as a continuum, that is, as a polycrystalline solid consisting of an aggregate of minerals or grains with void or pore space. The properties of intact rock are governed by the physical properties of the materials of which it is composed and by the manner in which they are bonded together. The properties that influence the engineering behaviour of rock material therefore include its mineralogical composition, texture, fabric, minor lithological characteristics, degree of weathering or alteration, density, porosity, strength, hardness, intrinsic (or primary) permeability, seismic velocity, and modulus of elasticity. Swelling and slaking are taken into account where appropriate, for example in argillaceous rocks.

On the other hand, a 'rock mass' includes the fissures and flaws as well as the rock material and may be regarded as a discontinuum of rock material transected by discontinuities. A discontinuity is a plane of weakness within the rock mass, across which the rock material is structurally discontinuous. Although discontinuities are not necessarily planes of separation, most of them are, and they possess little or no tensile strength. Discontinuities vary in size from small fissures to huge faults. The most common discontinuities in all rocks are joints and bedding planes. Other important discontinuities are planes of cleavage and schistosity, which occur in some metamorphic rock masses, and lamination, in some sedimentary rock masses. Obviously, discontinuities will have a significant influence on the engineering behaviour of rock in the ground. Indeed, the behaviour of a rock mass is, to a large extent, determined by the type, spacing, orientation, and characteristics of the discontinuities present. As a consequence, the parameters that should be used when describing a rock mass include the nature and geometry of the discontinuities, as well as overall strength, deformation modulus, secondary permeability, and seismic velocity of the rock mass.

The Influence of Weathering on Engineering Behaviour

The process of weathering represents an adjustment of the constituent minerals of a rock to the conditions prevailing at the surface of the Earth (*see* **Weathering**). Importantly, in terms of engineering behaviour, weathering weakens the rock fabric and exaggerates any structural discontinuities, thereby further aiding the breakdown processes. Rock may become more friable as a result of the development of fractures both between and within grains. Also, some weathered material may be removed, leaving a porous framework of individual grains.

Weathering is controlled by the presence of discontinuities because these provide access for the agents of weathering. Some of the earliest effects of weathering are seen along discontinuity surfaces. Weathering then proceeds inwards, so that the rock mass may develop a marked heterogeneity, with cores of relatively unweathered material within a highly weathered matrix (Figure 1). Ultimately, the whole of the rock mass can be reduced to a residual soil.

Weathering generally leads to a decrease in density and strength and to an increase in the deformability of the rock mass. An increase in the mass permeability frequently occurs during the initial stages of weathering owing to the development of fractures, but, if clay material is produced as minerals break down, the permeability may be reduced. Widening of discontinuities and the development of karstic features in carbonate rock masses lead to a progressive increase

Figure 1 Weathered basalt showing corestones with spheroidal weathering; basalt plateau, northern Lesotho.

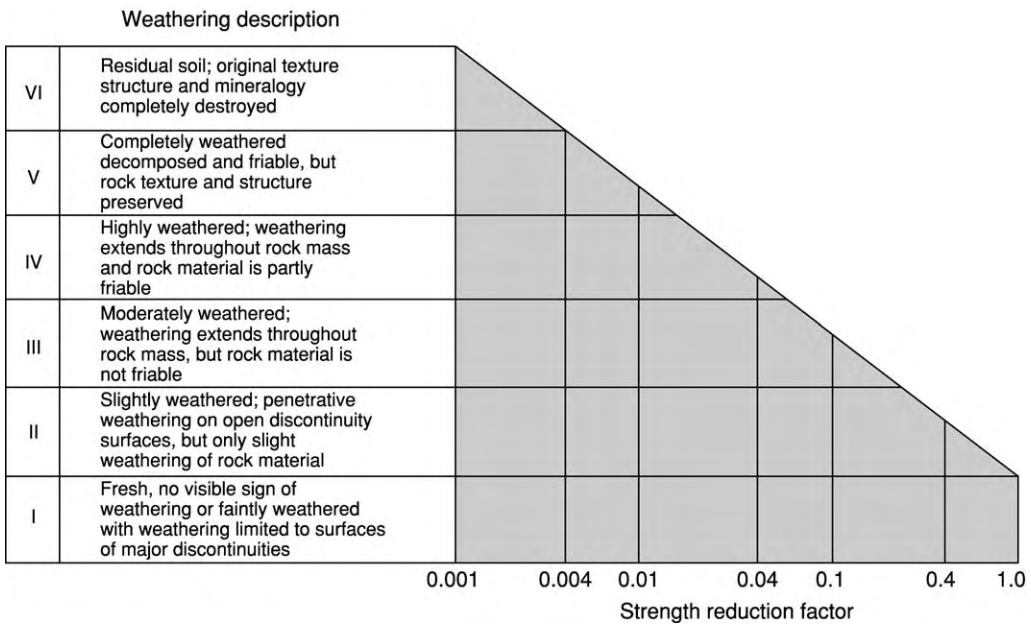


Figure 2 Generalized strength reduction due to increasing grade of weathering.

in secondary permeability. Weathering of carbonate and evaporite rock masses takes place primarily by dissolution. The dissolution is concentrated along discontinuities, which are thereby enlarged. Continuing dissolution ultimately leads to the development of sinkholes and cavities within the rock mass.

The grade of weathering refers to the stage that the weathering has reached. This varies between fresh rock on the one hand and residual soil on the other. Generally, about six grades of weathering are recognized, which, in turn, can be related to engineering performance (Figure 2). One of the first classifications of the engineering grade of weathering was devised for granite found in the vicinity of the Snowy Mountains hydroelectric power scheme in Australia. Subsequently, similar classifications have been developed for different rock types. Usually, rocks of the various grades lie one above the other in a weathered profile developed from a single rock type, the highest grade being found at the surface (Figure 3). However, this is not necessarily the case in complex geological conditions. Such a classification can be used to produce maps, sections, or models showing the distribution of the grade of weathering at a particular site. The dramatic effect of weathering on the strength of a rock is illustrated, according to the grade of weathering, in Figure 2.

Igneous and Metamorphic Rocks

Intrusive igneous rocks in their unaltered (unweathered) state are generally sound and durable, with adequate strength for any engineering requirement (Table 1). In some instances, however, they may be highly altered by weathering or hydrothermal processes (Figure 4). Furthermore, fissure zones are by no means uncommon in granites. Such granite rock masses may be highly fragmented along these zones; indeed, the granite may be reduced to sand-sized material and/or have undergone varying degrees of kaolinization (development of clay minerals). Generally, the weathered products of plutonic rocks have a large clay content. Granitic rocks can sometimes be porous and may have a permeability comparable with that of medium-grained sand. Some saprolites derived from granites that have been weathered in semi-arid climates may develop a metastable fabric and therefore be potentially collapsible. They may also be dispersive. Joints in plutonic rocks are often quite regular steeply dipping structures. Sheet joints tend to be approximately parallel to the topographical surface and develop as a result of stress relief following erosion (Figure 5). Consequently, they may introduce a dangerous element of weakness into valley slopes. The engineering properties and behaviour of gneisses are similar to those of granites.

Turning to extrusive igneous rocks, generally speaking, older volcanic rocks are not problematical; for instance, ancient lavas normally have strengths in excess of 200 MPa (Table 1). However, younger volcanic deposits have, at times, proved treacherous. This is because they often consist of markedly anisotropic sequences, in which lavas (generally strong), pyroclastics (generally weak or loose), and mudflows

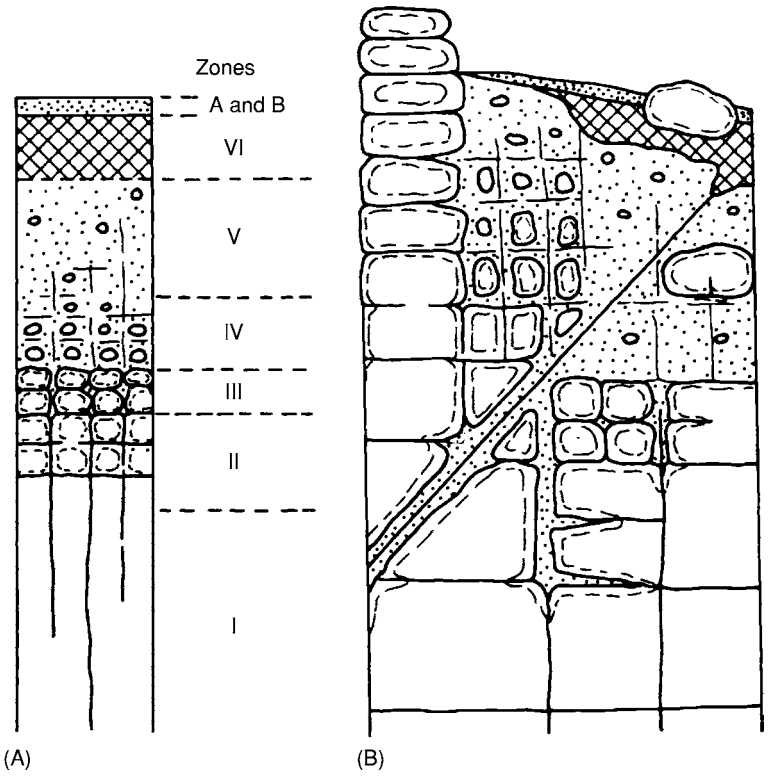


Figure 3 Weathering profile: (A) idealized profile; and (B) more complex profile.

Table 1 Geotechnical properties of some British igneous and metamorphic rocks

	Specific gravity	*Unconfined compressive strength (MPa)*[a]	*Young's modulus (GPa)*[b]
Mount Sorrel Granite (Leicestershire)	2.68	176.4 (VS)	60.6 (VL)
Eskdale Granite (Cumbria)	2.65	198.3 (VS)	56.6 (L)
Dalbeattie Granite (Kirkcudbrightshire)	2.67	147.8 (VS)	41.1 (L)
Markfieldite (Leicestershire)	2.68	185.2 (VS)	56.2 (L)
Granophyre (Cumbria)	2.65	204.7 (ES)	84.3 (VL)
Andesite (Somerset)	2.79	204.3 (ES)	77.0 (VL)
Basalt (Derbyshire)	2.91	321.0 (ES)	93.6 (VL)
Slate[c] (North Wales)	2.67	96.4 (S)	31.2 (L)
Slate[d] (North Wales)		72.3 (S)	
Schist[c] (Aberdeenshire)	2.66	82.7 (S)	35.5 (L)
Schist[d] (Aberdeenshire)		71.9 (S)	
Gneiss (Aberdeenshire)	2.66	162.0 (VS)	46.0 (L)
Hornfels (Cumbria)	2.68	303.1 (ES)	109.3 (VL)

[a]Classification of strength: ES, extremely strong, over 200 MPa; VS, very strong, 100 200 MPa; S, strong, 50 100 MPa.
[b]Classification of deformability: VL, very low, over 60 GPa; L, low, 30 60 GPa.
[c]Tested normal to cleavage or schistosity.
[d]Tested parallel to cleavage or schistosity.

are interbedded. Hence, foundation problems arise in volcanic sequences because weak beds of ash, tuff, and mudstone (formed from lahars) occur within lava piles, giving rise to problems of differential settlement and sliding. In addition, some volcanic materials weather relatively rapidly, so that weathering during periods of volcanic inactivity may readily lead to weakening and the development of soils, which are of much lower strength.

Clay minerals may be formed within newly erupted basaltic rocks by the alteration of primary minerals. Their presence can mean that the parent rhyolite, andesite, or basalt breaks down rapidly once exposed. The disintegration is exacerbated by the swelling of

Figure 4 Weathered granite grading into soil and showing a large corestone; near Hillcrest, Natal, South Africa.

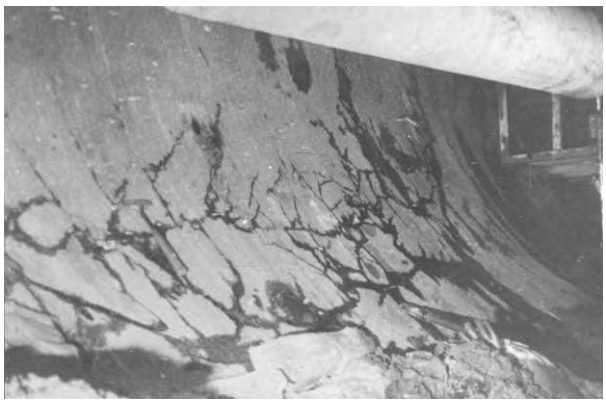


Figure 6 Crazing developed in basalt on exposure in the Transfer Tunnel; Lesotho Highlands Water Scheme.

Figure 5 An inselberg north of Mbabane, Swaziland, showing sheet joint up.

Figure 7 Amygdales and pipes in basalt, exposed in the Transfer Tunnel; Lesotho Highlands Water Scheme.

expansive clay minerals as they absorb water. This breakdown process has been referred to as slaking. These expansive clay minerals form when basic volcanic glass, olivine, pyroxene, or plagioclase is subjected to deuteric alteration brought about by hot gases and fluids from a magmatic source migrating through the rock. Disintegration can also be brought about by the absorption of water by zeolites, minerals that commonly develop in groundwater and are present in gas bubbles (amygdales). However, the breakdown of basalt also depends on its texture, since water must have access to the minerals that swell as water is absorbed. The disintegration of some basalts may take the form of crazing, an extensive microfracturing that develops on exposure to the atmosphere or to moisture (Figure 6). Such microfractures expand with time, causing the basalt to disintegrate into gravel-sized fragments. Some dolerites – intrusive rocks with a similar mineral composition to basalt – also undergo slaking.

Individual lava flows may be thin and transected by a polygonal pattern of cooling joints. They may also be vesicular or contain pipes, cavities, or even tunnels (Figure 7).

Pyroclastics (*see* **Pyroclastics**) usually give rise to extremely variable ground conditions owing to wide variations in their strength, durability, and permeability. Ashes tend to be weak and are often highly permeable. Those that are metastable are likely to undergo hydrocompaction on saturation. Moreover, ashes deposited on slopes are frequently prone to sliding, yet the irregular shapes of their constituent

particles, which can therefore interlock, may enable very steep slopes to be excavated that can stand up, at least in the short term. The strength and behaviour of tuffs depend on their degree of induration. However, the durability of some basaltic tuffs is poor or very poor, and they may be susceptible to frost. An ignimbrite is a pyroclastic rock consisting predominantly of pumiceous material that shows evidence of having been formed from a hot and concentrated pyroclastic flow. Once deposited, induration may be brought about by welding of viscous glassy fragments, by devitrification of glassy material, by deposition of material from escaping gases, and by compaction. Accordingly, ignimbrites have a wide variety of geotechnical characteristics, which are attributable to their modes of eruption, transportation, and deposition. At one extreme they are weak materials that behave as soils in the engineering sense; at the other extreme they are strong hard rocks in which extensive sets of essentially vertical cooling joints are developed. In fact, non-durable, intermediate, and highly durable ignimbrites have been recognized. The non-durable ignimbrites are characterized by low densities, high porosities, and low unconfined compressive strengths (5 MPa or less). A hyaloclastite consists of a mixture of rock and glass fragments. Initially, hyaloclastites are loose deposits, but as a result of weathering and diagenesis they usually become harder, particularly due to palagonitization, which is a solution–precipitation mechanism whereby the glass is hydrated and ions are leached out of it. The process slows down as the precipitation of authigenic minerals reduces the porosity and permeability of the rock mass. An increase in strength due to palagonitization can occur quickly; for example, loose pyroclastic material can be transformed into quite hard compact rocks within 20 years. Weaker varieties are probably more frequent than stronger types, often being hidden below the palagonized hyaloclastites that form the uppermost altered beds.

Slates, phyllites, and schists are characterized by textures that have a marked preferred orientation. Anisotropic structures such as cleavage and schistosity, attributable to such textures, not only adversely affect the strengths of metamorphic rocks but also make them much more susceptible to decay. Generally speaking, slates, phyllites, and schists weather relatively slowly, but the areas of regional metamorphism in which they occur have been subject to extensive folding so that, in places, such rocks may be extensively fractured and highly deformed. The quality of schists, slates, and phyllites is generally suspect. Care must be taken to detect weaker components; for instance, talc, chlorite, and sericite schists are weak rocks containing planes of schistosity only 1 mm or so apart. Some schists become slippery upon weathering and can therefore fail under a moderately light load.

Sandstones

Sandstones (*see* **Sedimentary Rocks:** Sandstones, Diagenesis and Porosity Evolution) exhibit a wide range of strengths (from less than 5.0 MPa to over 150 MPa), depending on their porosity, the amount and type of cement and/or matrix material, and the composition and texture of the individual grains. Higher cement or matrix content and lower porosity are characteristic of the stronger sandstones. In addition, their strength, like that of other rocks, is reduced by moisture content. The strength of saturated sandstone may be half of what it is when dry. For instance, the Kidderminster Sandstone (Triassic) has an average dry unconfined compressive strength of around 2.5 MPa, but when saturated this may be reduced to as low as 0.5 MPa. Nevertheless, sandstones generally do not give rise to notable engineering problems. Indeed, sandstones usually have sufficiently high coefficients of internal friction to give them high shearing strength when restrained under load. As a foundation rock, even poorly cemented sandstone is not normally susceptible to plastic deformation. Moreover, with the exceptions of shaly sandstone and rocks where the cement is readily soluble, clastic rocks are not subject to rapid deterioration on exposure. Nonetheless, salt action can give rise to honeycomb weathering in sandstone, which can cause relatively rapid disfigurement and deterioration when sandstone is used as a building stone. The process involves the progressive development of closely spaced cavities in the rock. Individual cavities range from a few millimetres to several centimetres in diameter, although larger cavernous weathering features, termed tafoni, can develop in sandstone exposures.

The presence of discontinuities can obviously adversely affect the behaviour of sandstone, reducing its mass strength. When inclined, discontinuities may cause rock to slide into unprotected excavations. Laminations impart a notable anisotropy to sandstone, reducing its strength to a significant degree along the planes of lamination. Furthermore, in certain engineering situations additional problems can develop. Friable sandstones, for example, can introduce problems of scour within dam foundations. Sandstones are also highly vulnerable to scouring and plucking actions in the overflow from dams and consequently have to be adequately protected by suitable hydraulic structures such as stilling basins. Quartzose sandstone in a tunnel being excavated by a tunnel-boring machine can prove highly abrasive to

the cutting head, and the cuttings produce an abrasive slurry when mixed with water.

Sandstones are frequently interbedded with shale. When such a formation is inclined, the layers of shale may represent potential sliding surfaces. Sometimes such interbedding accentuates the undesirable properties of the shale by concentrating water along the sandstone–shale contacts, thereby further weakening the shale.

Mudrocks

Mudrock is the commonest sedimentary rock, the two principal types being shale and mudstone. Shale is characterized by its lamination. Mudrock composed of grains of a similar size range and composition that is not laminated is usually referred to as mudstone. There is no sharp distinction between shale and mudstone; one grades into the other.

Shale is frequently regarded as an undesirable material to work with. Certainly, there are many examples of failures of structures founded on slopes in shales. Nonetheless, shales do vary in their engineering behaviour, largely according to their degree of compaction and cementation. Cemented shales are invariably stronger and more durable than compacted shales. The degree of packing – and hence the porosity, void ratio, and density of shale – depends on the mineral composition, grain-size distribution, mode of sedimentation, subsequent depth of burial, tectonic history, and effects of diagenesis. When the natural moisture content of shales exceeds 20%, they frequently become suspect and tend to develop potentially high pore-water pressures. Generally, shales with a cohesion of less than 20 MPa and an apparent angle of friction of less than 20° are likely to present engineering problems.

The higher the degree of fissility possessed by a shale, the greater the anisotropy with regard to strength, deformation, and permeability. For instance, the compressive strength and deformation modulus at right angles to the laminations may be more than twice those parallel to the laminations.

The greatest variation in the engineering properties of mudrocks can be attributed to the effects of weathering. Weathering reduces the amount of induration or removes it completely, leading to an increase in moisture content and a decrease in density. Indeed, weathering ultimately returns mudrock to a normally consolidated remoulded condition by destroying the bonds between the grains. Initially, mudrocks degrade rapidly to form a dominantly gravel-sized aggregate; this is facilitated by the presence of polygonal fracture patterns, joints, fissures, and bedding.

Depending on the relative humidity, some shales may slake almost immediately when exposed to the atmosphere. Alternate wetting and drying causes the rapid breakdown of compaction within the shale. Low-grade compaction shales, in particular, completely disintegrate after just a few cycles of drying and wetting. On the other hand, well-cemented shales are fairly resistant to slaking. If mudrocks undergo desiccation, air is drawn into the outer pores and capillaries as high suction pressures develop. On saturation, the entrapped air is pressurized as water is drawn into the rock by capillarity. Slaking therefore stresses the fabric of the rock. Disintegration consequently takes place as a result of air breakage after a sufficient number of wetting and drying cycles.

The swelling properties of certain mudrocks have proved to be extremely detrimental to the integrity of many civil engineering structures. Swelling, especially in clay shales, is attributable to the absorption of free water by expansive clay minerals, notably montmorillonite. Highly fissured overconsolidated shales have a greater tendency to swell than poorly fissured clayey shales, because the fissures provide access for water. The failure of poorly cemented mudrocks occurs during saturation, when the swelling pressure or internal saturation swelling stress developed by capillary suction pressures exceeds the tensile strength.

Uplift is common in excavations in shales, and can be attributed to swelling and heave. Rebound on unloading of the shale during excavation is attributed to heave due to the release of stored strain energy. Shale relaxes towards the newly excavated face, and sometimes this occurs as offsets along weaker seams in the shale. The greatest amount of rebound occurs in heavily overconsolidated compaction shales.

Settlement of shales can generally be managed by reducing the unit bearing load, for instance, by widening the base of a structure or by using spread footings. In some cases, appreciable differential settlement is provided for by designing an articulated structure that is capable of accommodating differential movement of individual sections without damage. Severe settlement may take place in low-grade compaction shales. However, compaction shales contain fewer open joints and fissures, which can be compressed beneath a heavy structure, than would the equivalent cemented shales.

When a load is applied to an essentially saturated shale foundation, the pore space in the shale decreases and the pore water attempts to migrate to regions of lesser load. Owing to the relative impermeability of shale, water becomes trapped in the voids and can migrate only slowly. As the load is increased, there comes a point at which it is partly transferred to the

pore water, resulting in a build-up of pore-water pressure. Depending on the permeability of the shale and the rate of loading, the pore-water pressure can increase, to the extent that it can equal the pressure imposed by the load. This greatly reduces the shear strength of the shale, and structures can fail under such conditions. By contrast, problems with pore-water pressure are generally less important in cemented shales.

Sulphur compounds are frequently present in argillaceous rocks. An expansion in volume large enough to cause structural damage can occur when sulphide minerals, such as pyrite and marcasite, oxidize, yielding products such as anhydrous and hydrous sulphates. Significant heave can occur when sulphur compounds resulting from the breakdown of pyrite combine with calcium to form gypsum and jarosite. Movements in excess of 100 mm have been recorded, with heave rates of about 2 mm per month. The reactions are exothermic, and the resulting increase in temperature can also have adverse effects, especially on the ventilation and insulation systems of basement and underground structures. The decomposition of sulphur compounds gives rise to aqueous solutions of sulphate and sulphuric acid, which react with the tricalcium aluminate in Portland cement to form calcium sulpho-aluminate or ettringite. This reaction is accompanied by significant expansion and leads to the deterioration of concrete.

Carbonate Rocks

Carbonate rocks are those that contain more than 50% carbonate minerals (such as calcite and dolomite). The term limestone (*see* **Sedimentary Rocks: Limestones**) is applied to those rocks in which the carbonate fraction exceeds 50%, over half of which is calcite or aragonite. If the carbonate material consists chiefly of dolomite, the rock is named dolostone.

Some representative geotechnical properties of carbonate rocks are listed in Table 2. It can be seen that, in general, the densities of these rocks increase with age, whilst the porosities are reduced. Furthermore, the porosity has a highly significant influence on the unconfined compressive strength: as the porosity increases, the strength declines. As both dolomitization and dedolomitization can give rise to increased porosity, both can be responsible for lower compressive strengths.

Both lithology and age frequently influence the strength and deformation characteristics of carbonate rocks. For instance, micritic (microcrystalline calcium carbonate) limestones may have a higher strength than sparitic (coarsely crystalline calcite) types. Most Silurian, Devonian, and Carboniferous limestones in Britain have compressive strengths of over 50 MPa, whereas Jurassic and Cretaceous limestones are often moderately weak, having unconfined compressive strengths of less than 12.5 MPa.

Table 2 Some geotechnical properties of British carbonate rocks

Property	*Limestone Carboniferous (Derbyshire)*	*Magnesian Limestone Permian (South Yorkshire)*	*Great Oolite Jurassic (Wiltshire)*	*Lower Chalk*	*Upper Chalk*
Dry density (Mg m ^{3})					
Range	2.55 2.61	2.46 2.58	1.91 2.21	1.85 2.13	1.35 1.61
Mean	2.58	2.51	1.98	2.08	1.44
Porosity (%)					
Range	2.4 3.6	8.5 12.0	13.8 23.7	17.2 30.2	29.6 45.7
Mean	2.9	10.4	17.7	20.6	41.7
Dry unconfined compressive strength (MPa)					
Range	65.2 170.9	34.6 69.6	8.9 20.1	19.1 32.7	4.8 6.2
Mean	106.2	54.6	15.6	26.4	5.5
Saturated unconfined compressive strength (MPa)					
Range	56.1 131.6	25.6 49.4	7.8 10.4	8.6 16.2	1.4 2.2
Mean	83.9	36.6	9.3	13.7	1.7
Young's modulus (GPa)					
Range	53.9 79.7	22.3 53.0	9.7 27.8	7.5 18.4	4.2 4.6
Mean	68.9	41.3	16.1	12.7	4.4

Dry density: very low, less than 1.8 Mg m ^{1}; low, 1.8 2.2 Mg m ^{1}; moderate, 2.2 2.55 Mg m ^{1}; high 2.55 2.75, Mg m ^{1}.
Porosity: low, 1 5%; medium, 5 15%; high, 15 30%; very high, over 30%.
Unconfined compressive strength: moderately weak, 5 12.5 MPa; moderately strong, 12.5 50 MPa; strong, 50 100 MPa; very strong, 100 200 MPa.
Young's modulus: very highly deformable, less than 5 GPa; highly deformable, 5 15 GPa; moderate, 15 30 GPa; low, 30 60 GPa; very low, over 60 GPa.

Similarly, the oldest limestones tend to possess the highest Young's moduli.

As can be seen from Table 2, the unconfined compressive strengths of the four limestones are reduced by saturation. The smallest reduction, 21%, is exhibited by the limestone of Carboniferous age, which is also the strongest. The reductions in strength of the Magnesian Limestone (Permian), Lincolnshire Limestone (Jurassic), and Great Oolite (Jurassic) are 35%, 40%, and 42% respectively. In other words, the strongest material undergoes the smallest reduction in strength on saturation, and there is a progressive increase in the average percentage reduction in strength after saturation as the dry strength of the limestone decreases; the weakest material shows the greatest reduction in strength.

Thickly bedded horizontally lying limestones that are relatively free from solution cavities afford excellent foundations. On the other hand, thinly bedded, highly folded, or cavernous limestones are likely to present serious foundation problems. The possibility of sliding may exist in highly bedded folded sequences. Similarly, when beds are separated by layers of clay or shale, especially when they are inclined, these may serve as sliding planes and result in failure.

Limestones are commonly transected by joints. These have generally been subjected to various degrees of dissolution, so much so that some may gape (Figure 8). Rain water is usually weakly acidic, and further acids may be taken into solution from carbon dioxide or from organic or mineral matter in the soil. The aggressiveness of the water to a limestone can be assessed on the basis of the relationship between the dissolved carbonate content, the pH, and the temperature of the water. At any given pH, the cooler the water, the more aggressive it is. If dissolution continues, its rate slackens, and it eventually ceases when the water becomes saturated. Hence, solution is greatest when the carbonate concentration is low. This occurs when water is circulating such that fresh supplies with low lime saturation are continually available. Freshwater can dissolve up to 400 mg l^{-1} of calcium carbonate. The solution of limestone, however, is a very slow process. For example, the mean rates of surface lowering of limestone areas in Britain tend to range from 0.04 mm $year^{-1}$ to 0.1 mm $year^{-1}$. Nevertheless, dissolution may be accelerated by manmade changes in the groundwater conditions or by a change in the character of the surface water that drains into the limestone.

Figure 8 Gaping joints due to dissolution, forming grykes in a limestone pavement above Malham Cove, North Yorkshire, England.

An important effect of dissolution within limestone is the enlargement of the pores, which enhances water circulation, thereby encouraging further dissolution. This increases the stress within the remaining rock fabric, which reduces the strength of the rock mass and leads to increasing stress corrosion. On loading, the volume of the voids is reduced by fracture of the weakened cement between the particles and by the reorientation of intact aggregates of rock that have been separated by the loss of bonding. Most of the resultant settlement takes place rapidly, within a few days of the application of the load.

The progressive opening of discontinuities by dissolution leads to an increase in mass permeability. Sometimes dissolution produces a highly irregular pinnacled surface, which may become exposed as a limestone pavement following soil erosion (Figure 8). The size, form, abundance, and downward extent of the aforementioned features depend on the geological structure and the presence of interbedded impervious layers. Solution cavities present numerous problems for the construction of large foundations.

Sinkholes may develop where opened joints intersect, and these may lead to an interlocking system of subterranean galleries and caverns (Figure 9A). Such features are associated with karstic landscapes (*see* **Sedimentary Processes:** Karst and Palaeokarst). The latter are characteristic of thick massive carbonate rocks. Sinkholes may be classified on the basis of origin into dissolution, collapse, caprock, dropout, suffusion, and buried sinkholes. Dissolution sinkholes form by slow dissolutional lowering of the outcrop or rockhead when surface water drains into carbonate rocks. Collapse sinkholes are formed by the collapse of cavern roofs that have become unstable owing to removal of support. Rapid collapse of a roof is very unusual; progressive collapse of a heavily fissured zone is more common. Caprock sinkholes are formed when an insoluble rock above a cavity in a carbonate rock collapses. Dropout sinkholes are formed by subsidence of a cohesive soil into cavities beneath. For example, clay soil

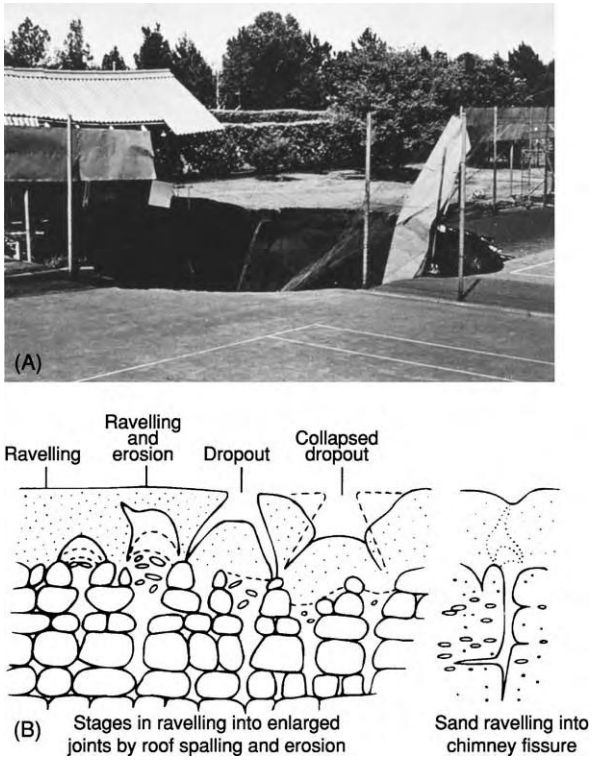


Figure 9 (A) A sinkhole developed as a result of groundwater lowering at a miner's recreation centre at Venterspoort, South Africa. (B) Mechanisms of ravelling.

may be capable of bridging a void developed in an underlying limestone. As the void is enlarged, the clay cover ravels until it eventually fails to form a dropout sinkhole (Figure 9B). Ravelling (breaking away) failures are the most widespread and probably the most dangerous of all the subsidence phenomena associated with karstic carbonate rocks. Rapid changes in moisture content lead to aggravated slabbing in clays. In particular, lowering of the water table increases the downward seepage gradient and accelerates downward erosion. It also reduces capillary attraction and increases the instability of flow through narrow openings, and gives rise to shrinkage cracks in highly plastic clays, which weaken the mass in dry weather and produce concentrated seepage during rains. Increased infiltration often initiates failure, especially when it follows a period when the water table has been lowered. Suffusion sinkholes develop in non-cohesive soils where percolating water washes the soil into a cavity in the underlying limestone. Buried sinkholes occur in limestone rockhead and are filled or buried with sediment. They may be developed by subsurface solution or as normal subaerial sinkholes that are later filled with sediment.

Unfortunately, the appearance of sinkholes at the surface is commonly influenced by human activity, especially groundwater withdrawal or an increased flow of water into the ground from a point source. As an example, more than 4000 sinkholes have been catalogued in Alabama as being caused by human activities, with the great majority developing since 1950. The largest is called the 'December Giant' because it developed suddenly in December 1972. It measures 102 m in diameter and is 26 m deep. Sinkholes are particularly dangerous when they form instantaneously by collapse, and they often occur in significant numbers within a short time-span. They have resulted in costly damage to a variety of structures and are a major local source of groundwater pollution.

Areas underlain by highly cavernous carbonate rocks possess the most sinkholes; hence, sinkhole density has proven to be a useful indicator of potential subsidence. Solution voids preferentially develop along zones of high secondary permeability because these concentrate groundwater flow. Data on fracture orientation and density, fracture intersection density, and the total length of fractures have been used to model the presence of solution cavities in limestone. Accordingly, the locations of areas at high risk of cavity collapse have been estimated from the intersections of lineaments formed by fracture traces and lineated depressions. Aerial photographs have proved particularly useful in this context. Subsidence-susceptibility maps can be developed using a geographical information system that incorporates the relevant data in a spatial context.

The chalk in southern England exhibits a wide range of dry density values, ranging from as low as 1.25 $Mg\,m^{-3}$ up to 2.5 $Mg\,m^{-3}$. Generally, chalk in northern England (i.e. Yorkshire) is denser than that in the south-east (Kent). The average porosity of chalk ranges between about 25% and 40%. The dry unconfined compressive strength of chalk varies from moderately weak to moderately strong; the chalk in Yorkshire tends to be appreciably stronger than that in Kent. When saturated, chalk undergoes a notable reduction in compressive strength, frequently by over 50%. The deformation properties of chalk *in situ* depend on its strength and on the spacing, tightness, and orientation of its discontinuities. In addition, the values of Young's modulus are influenced by the amount of weathering that the chalk has undergone. Six grades of weathering have been recognized, varying from completely unweathered material to a structureless mélange consisting of unweathered and partially weathered fragments of chalk in a matrix of deeply weathered chalk. Fresh chalk has a low compressibility and compresses elastically up to a critical pressure, which has been termed the apparent preconsolidation pressure (the yield stress). Marked breakdown and substantial consolidation occur at

high pressures. The yield stress of intact chalk may mark the point of pore collapse in high-porosity chalks (where the porosity exceeds 30%) as the cement bonds break down. Some interparticle slip and local grain crushing may also occur. With initial values of Young's modulus of 100 MPa and over, settlement in chalk is not normally a problem provided that the yield stress is not exceeded.

Chalk, being a relatively pure form of limestone, is subject to dissolution along discontinuities in the same way as the other carbonate rocks. However, subterranean solution features are less common in chalk, possibly because it is usually weaker than limestone and so collapses as solution occurs. Nevertheless, solution pipes and swallow holes are known to occur in chalk, commonly being found near the contact of Cretaceous chalk with the overlying Tertiary and Quaternary sediments. High concentrations of water, including run-off from roads, can lead to the reactivation of swallow holes and the formation of small pipes within a few years. Moreover, voids can gradually migrate upwards through chalk due to material collapse. Dissolution can lower the chalk surface beneath the overlying deposits, disturbing the latter and lowering their degree of packing. Hence, the chalk surface may be extremely irregular in places, causing problems of differential settlement to the foundations of buildings.

Evaporite Rocks

Evaporite deposits are formed by precipitation from saline waters in arid areas. They include anhydrite ($CaSO_4$), gypsum ($CaSO_4 \cdot nH_2O$), rock salt or halite (NaCl), and sylvite (KCl).

Dry densities and porosities for anhydrite and gypsum are given in Table 3, from which it can be seen that the ranges are relatively low. Anhydrite, according to its unconfined compressive strength, tends to be very strong, whereas gypsum tends to be only moderately strong. It appears that the impurity content of gypsum affects its strength and that material containing more impurities is stronger. This may be because impurities within calcium sulphate rocks tend to reduce the crystal size and consequently increase the strength. The Young's moduli are generally significantly higher for anhydrite than for gypsum. In other words, the deformability of anhydrite is either very low or low, whereas that of gypsum varies from low to high. In both rock types the amount of creep usually increases with increasing levels of constant loading.

The unconfined compressive strength of rock salt (halite) is generally moderately weak, whereas that of sylvite is moderately strong. These two rocks are either very highly or highly deformable (Table 3). In rock salt the yield strength may be as little as one-tenth of the ultimate compressive strength, whereas in sylvite plastic deformation is generally initiated at somewhere between 20% and 50% of the load at failure. Creep may account for anything between 20% and 60% of the strain at failure in these two evaporitic rock types.

Gypsum is more readily soluble than calcite. Sinkholes and caverns can therefore develop in thick beds of gypsum more rapidly than they would in

Table 3 Some geotechnical properties of British evaporite rocks

Property	*Anhydrite Cumbria*	*Gypsum North Yorkshire*	*Halite Cheshire*	*Sylvite Cleveland*
Dry density (Mg m ^{3})				
Range	2.77 2.82	2.16 2.32	1.92 2.09	1.86 1.99
Mean	2.79	2.21	2.03	1.94
Porosity (%)				
Range	3.1 3.7	3.4 9.1	2.7 7.9	3.2 8.7
Mean	3.3	5.1	4.8	5.4
Unconfined compressive strength (MPa)				
Range	77.9 126.8	19.0 40.8	9.4 14.9	18.5 31.8
Mean	102.9	27.5	11.7	25.8
Young's modulus (GPa) (E_{t50})				
Range	57.0 86.4	15.6 36.0	1.9 6.3	3.5 11.5
Mean	78.7	24.8	3.8	7.9

Dry density: very low, less than 1.8 Mg m ^{1}; low, 1.8 to 2.2 Mg m ^{1}; moderate, 2.2 to 2.55 Mg m ^{1}; high 2.55 to 2.75, Mg m ^{1}; very high, over 2.75 Mg m ^{1}.

Porosity: low, 1 to 5%; medium, 5 to 15%. Porosity derived by air porosimeter method.

Unconfined compressive strength: moderately weak, 5 to 12.5 MPa; moderately strong, 12.5 to 50 MPa; strong, 50 to 100 MPa; very strong, over 100 MPa.

Young's modulus: very highly deformable, less than 5 GPa; highly deformable, 5 to 15 GPa; moderately deformable, 15 to 30 GPa; low deformability, 30 to 60 GPa; very low deformability, over 60 GPa.

limestone. Indeed, in the USA, such features have been known to form within a few years, for instance where beds of gypsum are located beneath a dam. Extensive surface cracking and subsidence may be associated with the collapse of cavernous gypsum. The problem is accentuated by the fact that gypsum is weaker than limestone and therefore collapses more readily. Where beds of gypsum approach the surface, their presence is often indicated by collapse sinkholes. Such sinkholes can take only a matter of minutes to appear at the surface. However, where gypsum is effectively sealed from the ingress of water by overlying impermeable strata, such as mudstone, dissolution does not occur.

Massive deposits of gypsum are usually less dangerous than those of anhydrite, because gypsum tends to dissolve steadily, forming caverns or causing progressive settlement. In fact, the solution of massive gypsum is not likely to give rise to an accelerating deterioration beneath a foundation if precautions, such as grouting, are taken to keep seepage velocities low. On the other hand, massive anhydrite can be dissolved, leading to runaway situations in which seepage flow rates increase in a rapidly accelerating manner. Even small fissures in massive anhydrite can prove to be dangerous.

If gypsum or anhydrite occur in particulate form in the ground, their subsequent removal by dissolution can give rise to significant settlement. In such situations, the width of the solution zone and its rate of progress are obviously important as far as the location of hydraulic structures is concerned. Anhydrite is less likely than gypsum to undergo catastrophic solution in a fragmented or particulate form. Another point that should be borne in mind, and this particularly applies to conglomerates or breccias cemented with such soluble material, is that when this material is removed by solution the rock is greatly reduced in strength. In addition, when anhydrite comes into contact with water it may become hydrated to form gypsum. In so doing, there is a volume increase of between 30% and 58%, which exerts swelling pressures that are commonly between 1 MPa and 8 MPa and on rare occasions exceed 12 MPa. No great length of time is required to bring about such hydration.

Rock salt is even more soluble than gypsum, and the evidence of slumping, brecciation, and collapse structures in rocks that overlie saliferous strata bear witness to the fact that rock salt has gone into solution in past geological times. It is generally believed, however, that, in humid and semi-arid areas underlain by saliferous beds, measurable surface subsidence is unlikely to occur, except where salt is being extracted. Perhaps this is because an equilibrium has been attained between the supply of unsaturated groundwater and the salt available for solution. Exceptional cases of rapid subsidence have been recorded, such as the Meade salt sink in Kansas. Karstic features, particularly sinkholes, may develop in salt formations in arid areas. Salt extraction by some types of solution mining can give rise to serious or even catastrophic subsidence of the ground surface.

Organic Rocks: Coal

Coal is an organic deposit composed of different types of macerated plant tissue, which occurs in association with other sedimentary rocks such as shales, mudstones, and sandstones. Many coal seams have a composite character. At the bottom the coal is frequently softer, with bright coal in the centre and dull coal predominating in the upper part of the seam, reflecting changes in the type of plant material that accumulated and in the drainage conditions. Coal seams may be split, wholly or partially, by washouts. Coal usually breaks into small blocks that have three pairs of faces approximately parallel to each other. These surfaces are referred to as cleat. The cleat direction is fairly constant and is best developed in bright coal. Cleat may be coated with films of mineral matter, commonly calcite, ankerite, and iron pyrite. The breakdown of iron pyrite frequently gives rise to acid mine drainage, which is associated with coal mining activity and has caused problems in mining areas around the world. The heavy metals in, and the acidity and sulphate and iron contents of, acid mine drainage may pollute groundwater and surface water, and contaminate soils and sediments.

The unconfined compressive strength of coal varies, but generally it is less then 20 MPa. Exceptionally, the unconfined compressive strengths of some coals, such as the Barnsley Hard Coal, may exceed 50 MPa and their Young's moduli may be greater that 25 GPa. Consequently, the presence of coal seams in foundations does not usually present a problem. However, when coal has been mined at shallow depth, the presence of abandoned bell-pit or pillar-and-stall workings frequently cause foundation problems. Longwall mining of coal, which involves the total extraction of panels of coal, especially when it occurs at shallow depth, can give rise to notable subsidence at the ground surface, which, in turn, may damage buildings and structures or lead to flooding of agricultural land.

Another problem associated with coal is that it can spontaneously combust on exposure to air. This can be regarded as an atmospheric oxidation process in which self-heating occurs (i.e. an exothermic reaction emitting between 5 kCal and 10 kCal per

gram of coal). What is more, coal may be oxidized in the presence of air at temperatures below its ignition point. If heat is lost to the atmosphere, then the ignition temperature for coal is between 420°C and 480°C. However, where the heat of reaction is retained, the ignition point falls appreciably to between 35°C and 140°C. If the heat generated cannot be dissipated, the temperature rises, which increases the rate of oxidation so that the reaction becomes self-sustaining if there is a continuous supply of oxygen. Some coals ignite more easily than others; for example, high-rank coals are less prone to spontaneous combustion than are low-rank coals. It is not just coal seams that may spontaneously combust when exposed at or below the surface: colliery spoil heaps are also subject to spontaneous combustion. Obviously, the spontaneous combustion of coal, especially when it is being worked by opencast mining, and of spoil heaps when being reclaimed present appreciable problems.

See Also

Lava. **Pyroclastics**. **Rock Mechanics**. **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Chalk; Evaporites; Limestones; Sandstones, Diagenesis and Porosity Evolution. **Tectonics:** Fractures (Including Joints). **Weathering**.

Further Reading

Bell FG (ed.) (1992) *Engineering in Rock Masses*. Oxford: Butterworth Heinemann.

Bell FG (2000) *Engineering Properties of Soils and Rocks*, 4th edn. Oxford: Blackwell Scientific Publications.

Bell FG (2004) *Engineering Geology and Construction*. London: E & FN Spon.

Bell FG, Waltham AC, and Culshaw MG (2004) *Sink holes and Subsidence*. Chichester/New York: Praxis/Springer.

Goodman RE (1993) *Engineering Geology: Rock in Engineering Construction*. New York: Wiley.

Problematic Soils

F G Bell, British Geological Survey, Keyworth, UK

Introduction

Most civil engineering operations are founded in the uppermost layers of the ground and are therefore generally carried out in soil. Poor soil conditions in terms of engineering increase the cost of construction by necessitating special foundation structures and/or mean that some type of engineering soil treatment is required. Consequently, it important to understand the nature of the soil that is being dealt with. However, all soil types at times can be problematic, depending upon the conditions existing at a particular engineering site. For instance, if saturated gravel is to be excavated into, then it will have to be dewatered prior to the commencement of the operation in order to avoid flooding the site. Furthermore, soils can be contaminated and thereby present special engineering problems if the area where they occur is to be developed. Be that as it may, only the more troublesome soil types are dealt with here.

In the engineering sense, soil consists of an unconsolidated assemblage of particles between which are voids, which may contain air or water or both. As such, soil consists of three phases, that is, solids, water, and air. Under certain circumstances, soil can contain other gases such as methane or other liquids such as nonaqueous phase liquids. The interrelationships of the weights and volumes of the three phases are important since they help define the character of a soil. The solid phase of soil is derived from the breakdown of rock material by weathering and/or erosion, and it may have suffered a varying amount of transportation prior to deposition. It also may contain organic matter, the total organic content of soils varying from less than 1% in the case of some immature or desert soils to over 90% in the case of peats. The type of breakdown process(es) and the amount of transport undergone by sediments influence the nature of the macro- and microstructure of the soil that, in turn, influence its behaviour (Table 1). Furthermore, the same type of rock can give rise to different types of soils, depending on the climatic regime and the vegetative cover under which it develops. Indeed, the character of a soil frequently is influenced to a significant extent by the climatic regime in which it is formed and exists. This is especially the case with some soils formed in more extreme climates such as arid and semi-arid zones, notable examples being sabkha soils and dispersive soils, respectively, or quick clays in cold climates. Such soils possess their own peculiar characteristics that provide problems for the engineer. Time also is an important factor in the development of a mature soil.

Table 1 Effects of transportation on sediments

	Gravity	*Ice*	*Water*	*Air*
Size	Various	Varies from clay to boulders	Various sizes from boulder gravel to muds	Sand size and less
Sorting	Unsorted	Generally unsorted	Sorting takes place both laterally and vertically. Marine deposits often uniformly sorted. River deposits may be well sorted	Uniformly sorted
Shape	Angular	Angular	From angular to well rounded	Well rounded
Surface texture	Striated surfaces	Striated surfaces	Gravel: rugose surfaces. Sand: smooth, polished surfaces. Silt: little effect	Impact produces frosted surfaces

Any system of soil classification involves grouping soils into categories that possess similar properties, so providing a systematic method of soil description by which soils can be identified quickly. Although soils include materials of various origins, for purposes of engineering classification it is sufficient to consider their simple index properties, which can be assessed easily, such as their particle size distribution and consistency limits. For instance, coarse-grained soils are distinguished from fine on a basis of particle size, gravels and sands being the two principal types of coarse-grained soils. Plasticity also is used when classifying fine-grained soils, that is, silts and clays.

Quicksands

As water flows through sands or silts and slows down, its energy is transferred to the particles past which it is moving that, in turn, creates a drag effect on the particles. If the drag effect is in the same direction as the force of gravity, then the effective pressure is increased and the soil is stable. Conversely, if water flows towards the surface, then the drag effect is counter to gravity, thereby reducing the effective pressure between particles. If the upward flow velocity is sufficient, it can buoy up the particles so that the effective pressure is reduced to zero. This represents a critical condition where the weight of the submerged soil is balanced by the upward acting seepage force. If the upward velocity of flow increases beyond this critical hydraulic gradient, then a quick condition develops. As the velocity of the upward seepage force increases further from the critical gradient, the soil begins to boil more and more violently (**Figure 1A**). At such a point structures fail by sinking into the quicksand (**Figure 1B**).

Quicksands, if subjected to deformation or disturbance, can undergo a spontaneous loss of strength, causing them to flow like viscous liquids. A number of conditions must exist for quick conditions to develop. Firstly, the sand or silt concerned must be saturated and loosely packed. Secondly, on disturbance the constituent grains become more closely packed, which leads to an increase in pore-water pressure, reducing the forces acting between the grains. This brings about a reduction in strength. If the pore water can escape very rapidly the loss in strength is momentary. Hence, the third condition requires that pore water cannot escape readily. This is fulfilled if the sand or silt has a low permeability and/or the seepage path is long.

Quick conditions frequently are encountered in excavations made in fine sands that are below the watertable. Liquefaction of potential quicksands also may be brought about by sudden shocks caused by the action of heavy machinery (notably pile driving), blasting, and earthquakes (**Figure 1B**). Such shocks increase the stress carried by the pore water, and give rise to a decrease in the effective stress and shear strength of the soil. There is also a possibility of a quick condition developing in a layered soil sequence containing fine sands where the individual beds have different permeabilities.

Collapsible Soils

Soils, such as loess and brickearth, commonly are regarded as being of aeolian origin and some believe that the material of which they are composed was formed initially by glacial action. These soils often have a metastable fabric and so possess the potential to collapse. In addition, some wind-blown silts, such as those found in subtropical arid areas like Arizona and the Kalahari Desert in southern Africa, have metastable fabrics and are potentially collapsible. Certain residual soils also are prone to collapse. For example, when granites undergo notable chemical weathering in subtropical regions that involves appreciable leaching, then the resulting saprolite and

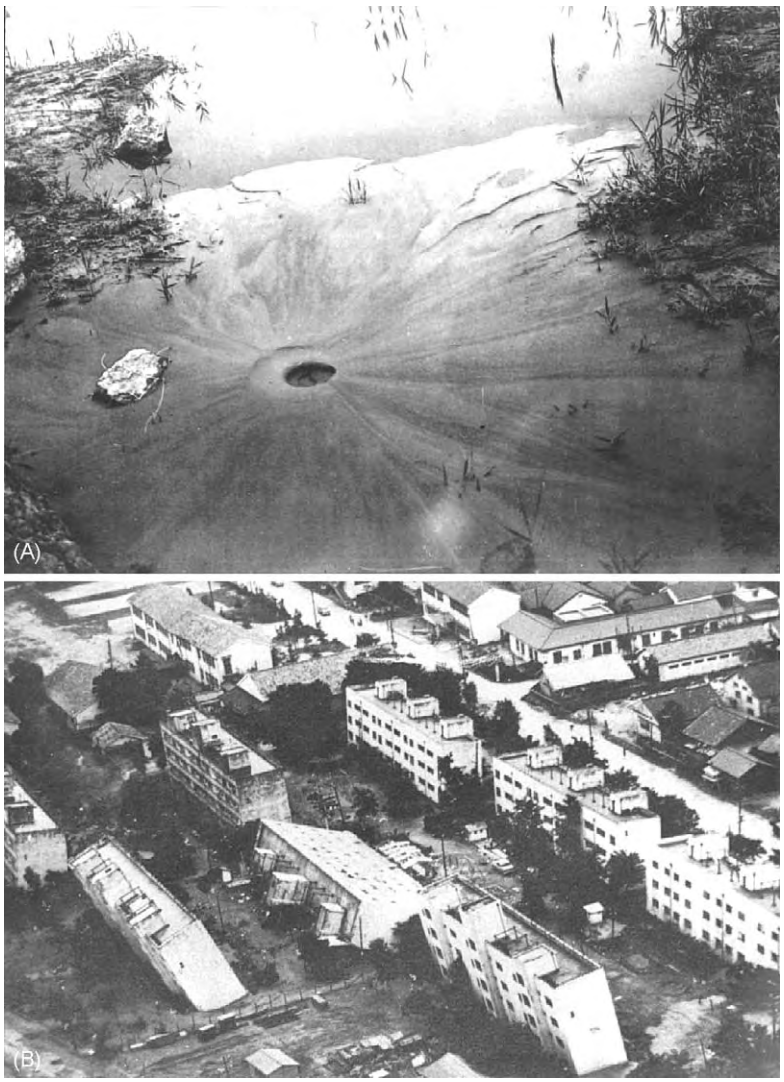


Figure 1 (A) Liquefaction of sand due to the Tangshan earthquake (28 July 1976) near Dougtian, China. (B) Collapse of buildings because of liquefaction of fine sands and silts, Niigata, Japan, 1964.

residual soil tend to develop a metastable fabric as fine particles are removed. As a consequence, the weathered products are potentially collapsible.

Truly collapsible soils are those in which collapse occurs on saturation since the soil fabric cannot support the weight of the overburden. When the saturation collapse pressure exceeds the overburden pressure, soils are capable of supporting a certain level of stress on saturation and can be regarded as conditionally collapsible soils. The maximum load that such soils can support is the difference between the saturation collapse and overburden pressures. These soils generally consist of 50 to 90% silt particles, and sandy, silty, and clayey types have been recognized, with most falling into the silty category (Figure 2). The fabric of collapsible soils generally takes the form of a loose skeleton of

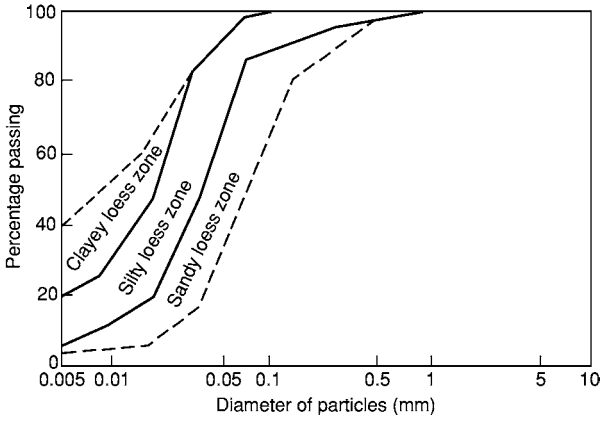


Figure 2 Particle size distribution of loess.

grains (generally quartz) and micro-aggregates (usually assemblages of clay particles). These tend to be separate from each other, being connected by bonds and bridges, with uniformly distributed pores (Figure 3). The bridges are formed of clay-sized minerals. As grains are not in contact, mechanical behaviour is governed by the structure and quality of the bonds and bridges.

The structural stability of collapsible soils also is related to the amount of weathering undergone. Younger weakly weathered loess generally has a high potential for collapse whereas older weathered loess is relatively stable. In addition, highly collapsible loess tends to occur in regions where the landscape and/or the climatic conditions are not conducive to development of long-term saturated conditions within the soil. The size of the pores is all important, collapse normally occurring as a result of pore-space reduction taking place in pores greater than 1 μm in size and more especially in those exceeding 10 μm in size. Hence, collapsible soils possess porous textures with high void ratios and relatively low densities. At their natural low moisture content, these soils possess high apparent strength but they are susceptible to large reductions in void ratio upon wetting. In other words, their metastable texture collapses as the bonds between the grains break down when the soil is wetted. Collapse on saturation normally only takes a short period of time, although the more clay particles such a soil contains, the longer the period tends to be.

From the above it may be concluded that significant settlements can take place beneath structures founded on collapsible soils after they have been wetted, in some cases in the order of metres. These have led to foundation failures. A number of techniques can be used to stabilize collapsible soils; these are summarized in Table 2.

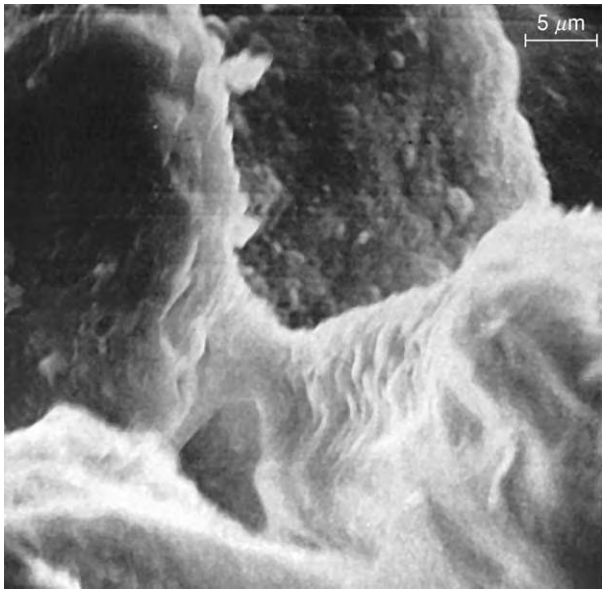


Figure 3 Scanning electron photomicrograph of brickearth from south east Essex, showing three silt particles, two of which are bridged by clay particles.

Expansive Clays

An important characteristic of some clay soils is their susceptibility to slow volume change that can occur independently of loading due to swelling or shrinkage. Some of the best examples of expansive clays are provided by vertisols or black clays. Such soils typically are developed in subtropical regions with well-defined wet and dry seasons. Unfortunately, expansive clays frequently are responsible for significant national costs due to damage caused to property. For example, in the United States, the cost of repair of damage to property built on expansive clays exceeds two billion dollars annually. This frequently is twice the cost of flood damage or landslide damage and 20 times the cost of earthquake damage. Many clay soils in Britain, especially south-east England, possess the potential for large volume change. However, the mild damp climate means that significant deficits in soil moisture developed during the summer are confined to the upper 1.0 to 1.5 m of the soil and that the field capacity is re-established during the winter. Even so, deeper permanent deficits can be brought about by

Table 2 Methods of treating collapsible foundations

Depth of subsoil treatment	
0 1.4 m	Moistening and compaction (conventional extra heavy impact or vibratory rollers)
1.5 10 m	Over excavation and recompaction (earth pads with or without stabilization by additives such as cement or lime). Vibro compaction (stone columns). Vibroreplacement. Dynamic compaction. Compaction piles. Injection of lime. Lime piles and columns. Jet grouting. Ponding or flooding (if no impervious layer exists). Heat treatment to solidify the soils in place.
Over 10 m	Any of the aforementioned or combinations of the aforementioned, where applicable. Ponding and infiltration wells, or ponding and infiltration wells with the use of explosive.

transpiration from large trees, and significant damage to property on potentially expansive clay can be caused during notably dry summers.

Generally, the clay fraction of expansive clay exceeds 50%, silty material varying between 20 and 40%, and sand forming the remainder. Montmorillonite normally is present in the clay fraction and is the principal factor determining the appreciable volume changes that take place in these soils on wetting and drying.

The depth of the active zone in expansive clays (i.e., the zone in which swelling and shrinkage occurs in wet and dry seasons, respectively) can be large. For instance, the maximum seasonal changes in the moisture content of expansive clays in Romania are around 20% at 0.4 m depth, 10% at 1.2 m depth, and less than 5% at 1.8 m depth. The corresponding cyclic movements of the ground surface are between 100 and 200 mm. Surface heaves approaching 500 mm have been recorded in some expansive clays in South Africa. During the dry season, profiles in some regions can dry out to depths of 15 to 20 m.

The potential for volume change in expansive clay soils depends on the initial moisture content, initial voids ratio, the microstructure, and the vertical stress, as well as the type and amount of clay minerals present. The type of clay minerals are responsible primarily for the intrinsic expansiveness, whilst the change in moisture content or suction controls the actual amount of volume change that a soil undergoes at a given applied pressure. Changes in soil suction are brought about by moisture movement through the soil, due to evaporation from its surface in dry weather, by transpiration from plants, or alternatively by recharge consequent upon precipitation. Alternating wet and dry seasons may produce significant vertical movements as soil suction changes. The rate of expansion depends upon the rate of accumulation of moisture in the soil. In semi-arid regions, it is limited by the availability of water and this, together with the available void volume, governs the rate of penetration of the heave front in the soil. When these expansive clays occur above the water table, they can undergo a high degree of shrinkage on drying. Seasonal changes in volume also produce shrinkage cracks so that expansive clays are often heavily fissured. Sometimes the soil is so desiccated that the fissures are wide open and the soil is shattered or micro-shattered.

Transpiration from vegetative cover is a major cause of water loss from soils in subtropical semi-arid regions. Indeed, the distribution of soil suction in soil is controlled primarily by transpiration from vegetation and this represents one of the most significant changes made in loading (i.e., to the state of stress in a soil). The suction induced by the withdrawal of moisture fluctuates with the seasons, reflecting the growth of vegetation. The maximum soil suction that can be developed is governed by the ability of vegetation to extract moisture from the soil. The level at which moisture is no longer available to plants is termed the permanent wilting point. In fact, the moisture content at the wilting point exceeds that of the shrinkage limit in soils with high clay contents and is less in those possessing low clay contents. This explains why settlement resulting from the desiccating effects of trees is more notable in low to moderately expansive soils than in highly expansive ones.

These volume changes can give rise to ground movements that may result in damage to buildings. Low-rise buildings are particularly vulnerable to such ground movements since they generally do not have sufficient weight or strength to resist. Be that as it may, three methods can be adopted when choosing a design solution for building on expansive soils. Firstly, a foundation and structure can be provided that can tolerate movements without unacceptable damage; secondly, the foundation and structure can be isolated from the effects of the soil; and thirdly, the ground conditions can be altered or controlled. These soils also represent a problem when they are encountered in road construction, and shrinkage settlement of embankments composed of such clay soils can lead to cracking and breakup of the roads they support (Figure 4).

Dispersive Soils

Dispersive soils occur in subtropical semi-arid regions, normally where the rainfall is less than 850 mm annually. Dispersion occurs in such soils when the repulsive forces between clay particles exceed the attractive forces, thus bringing about deflocculation so that in the presence of relatively pure water the particles repel each other to form colloidal suspensions. In non-dispersive soil there is a definite threshold velocity below which flowing water causes no erosion. By contrast, there is no threshold velocity for dispersive soil, the colloidal clay particles going into suspension even in quiet water. Therefore, these soils are highly susceptible to erosion and piping. Dispersive soils contain a moderate to high content of clay material but there are no significant differences in the clay fractions of dispersive and non-dispersive soils, except that soils with less than 10% clay particles may not have enough colloids to support dispersive piping. Dispersive soils contain a higher content of dissolved sodium (up to 12%) in their pore water than ordinary soils. The clay particles in soils with high salt contents exist as aggregates and coatings around silt and sand particles (Figure 5).

For a given eroding fluid the boundary between the flocculated and deflocculated states depends on the

Figure 4 Breakup along the CapeTown Johannesburg road, South Africa, due to expansive clay.

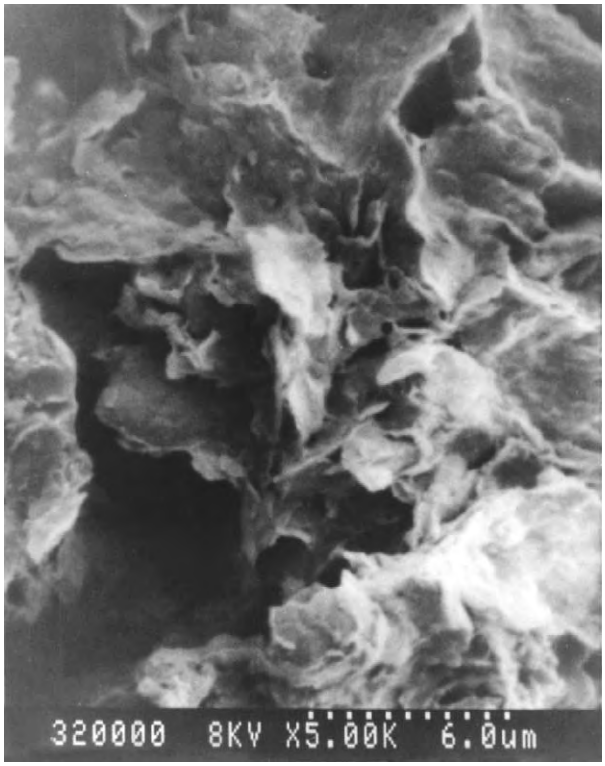


Figure 5 Scanning electron photomicrograph illustrating the fabric of dispersed soil from Natal, South Africa.

value of the sodium adsorption ratio, the salt concentration, the pH value, and the mineralogy. The sodium adsorption ratio (SAR) is used to quantify the role of sodium where free salts are present in the pore water and is defined as:

$$\mathrm{SAR} = \frac{\mathrm{Na}}{\sqrt{0.5(\mathrm{Ca} + \mathrm{Mg})}} \qquad [1]$$

with units expressed in meq/litre of the saturated extract. It has been suggested that an SAR value greater than 10 is indicative of dispersive soils, between 6 and 10 as intermediate, and less than 6 as non-dispersive. However, dispersion has occurred in soils with values lower than 6. The presence of exchangeable sodium is the main chemical factor contributing towards dispersive behaviour in soil. This is expressed in terms of the exchangeable sodium percentage (ESP):

$$\mathrm{ESP} = \frac{\text{exchangeable sodium}}{\text{cation exchange capacity}} \times 100 \qquad [2]$$

where the units are given in meq/100 g of dry clay. A threshold value of ESP of 10% has been recommended, above which soils that have their free salts leached by seepage of relatively pure water are prone to dispersion. Soils with ESP values above 15% are highly dispersive. Those with low cation exchange values (15 meq/100 g of clay) have been found to be completely non-dispersive at ESP values of 6% or below. Similarly, soils with high cation exchange capacity values and a plasticity index greater than 35% swell to such an extent that dispersion is not significant.

Severe erosion or worse, serious piping damage to embankments and piping failures of earth dams have occurred when dispersive soils have been used in their construction. Indications of piping take the form of small leakages of muddy coloured water from an earth dam after initial filling of the reservoir. The pipes become enlarged rapidly and this can lead to failure of the dam (Figure 6). Experience, however, indicates

Figure 6 Failed earth dam constructed of dispersive soil, showing piping outlets on the downstream side, near Ramsgate, South Africa.

that if an earth dam is built with careful construction control and incorporates filters, then it should be safe enough, even if it is constructed with dispersive soil. Alternatively, hydrated lime, pulverised fly ash, gypsum, or aluminium sulphate have been used to treat dispersive soils used in earth dams.

Humid Tropical Zone Soils

In humid tropical regions, weathering of rock is more intense and extends to greater depth than in other parts of the world. Residual soils develop in place as a consequence of weathering, primarily chemical weathering. The mineralogy of residual soils is partly inherited from the parent rock from which they were derived and partly produced by the processes of weathering. Hence, the mineralogy varies widely, as does grain size and unit weight. The particles and their arrangement evolve gradually as weathering proceeds. In addition, weathering of parent rock *in situ* may leave behind relict structures that may offer weak bonding even in extremely weathered material. Low strength along relict discontinuities may be attributable to particles being coated with low-friction iron/manganese organic compounds.

Reproducible results from some standard tests may be difficult to obtain from residual tropical soils. Different results can be obtained depending upon whether the soil is pre-dried prior to testing or kept close to its natural moisture content. Also, disaggregation of the soil structure, especially in relation to particle size analysis has proved problematic. Consequently, conventional index tests frequently have been modified in an attempt to make them more applicable for use with tropical residual soils. Of course, it would be wrong to assume that all tropical soils behave differently from those found in other climatic regions. For instance, alluvial clays and sands behave in the same manner and have similar geotechnical properties, regardless of the climatic conditions of the region of deposition.

Drying brings about changes in the properties of residual clay soils in that it initiates two important effects, namely, cementation by the sesquioxides and aggregate formation on the one hand, and loss of water from hydrated clay minerals on the other. In the case of halloysite, the latter causes an irreversible transformation to metahalloysite. Drying can cause almost total aggregation of clay size particles into silt and sand size ranges, and a reduction or loss of plasticity. Cycles of wetting and drying may increase the stiffness of the soil fabric, which increases its shear strength and decreases its compressibility.

Laterite can be regarded as a highly weathered material that forms as a result of the concentration of hydrated oxides of iron and aluminium in such a way that the character of the deposit in which they occur is affected. These oxides may be present in an unhardened soil, as a hardened layer, as concretionary nodules in a soil matrix or in a cemented matrix enclosing

other materials. When a hardened crust is present near the surface, then the strength of laterite beneath decreases with increasing depth.

Red clays and latosols are residual ferruginous soils formed primarily by chemical weathering of the parent rock. This results in the release of iron and aluminium sesquioxides, increasing loss of silica and increasing dominance of new clay minerals such as smectites, allophane, halloysite and, with increasing weathering, kaolinite. The microstructure also is developed by chemical weathering processes and consists of an open-bonded fabric of silt and sand size peds, which are formed mainly of clay minerals and fine disseminated iron oxides. Relatively weak bonds exist between the peds and are formed of iron oxides and/or amorphous aluminium silicate gels. Such soils differ from laterite in that they behave as clay and do not possess strong concretions. They do, however, grade into laterite.

Allophane-rich soils or andosols are developed from basic volcanic ashes in high temperature-rainfall regions. Allophane is an amorphous clay mineral. These soils have very high moisture contents, usually in the range 60% to 80% but values of up to 250% have been recorded; and corresponding high plasticity. The soils also are characterized by very low dry densities and high void ratios (sometimes as high as 6). Moisture content does affect the strength of andosols significantly as the degree of saturation can have an appreciable affect on cementation. Soils containing halloysite, or its partially dehydrated form metahalloysite, have high moisture contents (30% to 65%) and can possess high plasticity. Some of these soils are susceptible to collapse.

Soils of Hot Arid Regions

Most soils in arid regions consist of the products of physical weathering of rock material. This breakdown process gives rise to a variety of rock and mineral fragments that may be transported and deposited under the influence of gravity, wind, or water. Many arid soils are of aeolian origin and sands frequently are uniformly sorted. Uncemented silty soil may possess a metastable fabric and hence be potentially collapsible. The precipitation of salts in the upper horizons of an arid soil, due to evaporation of moisture from the surface, commonly means that some amount of cementation has occurred, which generally has been concentrated in layers, and that the pore water is likely to be saline. High rates of evaporation in hot arid areas may lead to ground heave due to the precipitation of minerals within the capillary fringe.

Where the watertable is at a shallow depth the soils may possess a salty crust and be chemically aggressive due to the precipitation of salts from saline groundwater. Occasional wetting and subsequent evaporation frequently are responsible for a patchy development of weak, mainly carbonate, and occasionally gypsum cement, often with clay material deposited between and around the coarser particles. These soils, therefore, may undergo collapse, especially where localized changes in the soil–water regime are brought about by construction activity. Collapse is attributed to a loss of strength in the binding agent and the amount of collapse undergone depends upon the initial void ratio. Loosely packed aeolian sandy soils, with a density of less than 1.6 Mg m^{-3}, commonly exhibit a tendency to collapse.

Silts may have been affected by periodic desiccation and be interbedded with evaporite deposits. The latter process leads to the development of a stiffened crust or, where this has occurred successively, to a series of hardened layers within the formation. Loosely packed aeolian silty soils formed under arid conditions often undergo considerable volume reduction or collapse when wetted. Such metastability arises from the loss in strength of interparticle bonds resulting from increases in water content. Thus, infiltration of surface water, including that applied during irrigation, leakage from pipes, and rise of water table, may cause large settlements to occur.

Low-lying coastal zones and inland plains in arid regions with shallow water tables, are areas in which sabkha conditions commonly develop. Sabkhas are extensive saline flats that are underlain by sand, silt, or clay and often are encrusted with salt. Groundwater is saline, containing calcium, sodium, chloride, and sulphate ions. Evaporative pumping, whereby brine moves upward from the water table under capillary action, appears to be the most effective mechanism for the concentration of salt in groundwater and the precipitation of minerals in sabkha. Salts are precipitated at the ground surface when the capillary fringe extends from the water table to the surface.

One of the main problems with sabkha is the decrease in density and strength, and increased permeability that occur, particularly in the uppermost layers, after rainfall, flash floods, or marine inundation, due to the dissolution of soluble salts that act as cementing materials. Changes in the hydration state of minerals, such as calcium sulphate, also cause significant volume change in soils. There is a possibility of differential settlement occurring on loading due to the different compressibility characteristics resulting from differential cementation of sediments. Excessive settlement also can occur, due to the removal of soluble salts by flowing groundwater. This can cause severe disruption to structures within months or a few years. Movement of groundwater also can lead to the dissolution of

minerals to an extent that small caverns, channels, and surface holes can be formed. On the other hand, heave resulting from the precipitation and growth of crystals can elevate the surface of a sabkha in places by as much as 1 m.

Because sabkha soils frequently are characterized by low strength, their bearing capacity and compressibility frequently do not meet routine design requirements. Various ground improvement techniques, therefore, have been used in relation to large construction projects such as vibro-replacement, dynamic compaction, compaction piles, and underdrainage. Soil replacement and preloading have been used when highway embankments have been constructed.

Cementation of sediments by precipitation of mineral matter from the groundwater may lead to the development of various crusts or cretes (e.g., precipitation of calcite gives rise to calcrete and of gypsum to gypcrete). These may form continuous sheet or isolated patch-like masses at the ground surface when the water table is at or near this level, or at some other position within the ground profile. Therefore, duricrusts or pedocretes are soils that have been, to a greater or lesser extent, cemented. These materials take three forms, namely, indurated (e.g., hardpans and nodules), non-indurated (soft or powder forms), and mixtures of the two (e.g., nodular pedocretes). Indurated pedocretes that occur at the surface may be underlain by loose or soft material that, at times, may be potentially expansive or collapsible. Nodular pedocrete with a smectitic clay matrix could have the potential to be expansive, whereas collapsibility may be associated with cemented soils, powder, or nodular calcretes. Small-scale karst-like features have been recorded in some weathered calcretes.

Soils Developed in Cold Regions

Tills are deposited by ice-sheets and consist of a variable assortment of rock debris ranging from fine rock flour to boulders; they are characteristically unsorted. On the one hand, tills may consist predominantly of sand and gravel with very little binder, whereas on the other they may have an excess of clayey material. The nature of a till deposit depends on the lithology of the material from which it was derived, on the position in which it was transported in the glacier, and on the mode of deposition. The underlying bedrock usually constitutes up to about 80% of basal tills, depending on its resistance to abrasion. Till sheets can comprise one or more layers of different material, not all of which are likely to be found at any one locality. Shrinking and reconstituting of an ice-sheet can complicate the sequence.

Distinction has been made between tills derived from rock debris carried along at the base of a glacier and those deposits that were transported within or at the terminus of the ice. The former is referred to as lodgement till, whilst the latter is termed ablation till. Lodgement till is commonly stiff, dense, and relatively incompressible. Because of the overlying weight of ice, such deposits are overconsolidated. Fissures frequently are present in lodgement till and influence its shear strength and therefore its stability. Ablation till accumulates on the surface of the ice when englacial debris melts out. It, therefore, is normally consolidated and non-fissile. The proportion of silt and clay size material is relatively high in lodgement till (e.g., the clay fraction typically varies from 15 to 40%). Ablation till is characterized by abundant large angular stones, the proportion of sand and gravel is high, and clay is present only in small amounts (usually less than 10%).

The most familiar pro-glacial deposits are varved clays. These sediments accumulated on the floors of glacial lakes and are characteristically composed of alternating laminae of finer and coarser grain size, each such couplet being termed a varve. The thickness of the individual varve is frequently less than 2 mm. Generally, the coarser layer is of silt size and the finer of clay size. Clay minerals tend to show a high degree of orientation parallel to the laminae.

Varved clays have a very wide range of plasticity, although most tend to be highly plastic and many possess the potential to swell. They tend to be normally consolidated or lightly overconsolidated. Furthermore, they often undergo a notable reduction in strength when remoulded, that is, they are medium sensitive to sensitive. However, the remoulded strength increases with time (Figure 7). The anisotropic behaviour of varved clay is explained by the influence of their lamination.

Quick clays were formed at the end or after the last ice age, when meltwater from glaciers carried large quantities of fine-grained sediments into the sea. The material of which quick clays are composed is predominantly smaller than 0.002 mm but some deposits seem to be poor in clay minerals, containing a high proportion of ground down fine quartz. The open fabric that is characteristic of quick clays has been attributed to their initial deposition, during which time colloidal particles interacted to form loose aggregations by gelation and flocculation. This fabric may have been retained to the present day.

Quick clays often exhibit little plasticity. However, the most extraordinary property possessed by quick clays is their very high sensitivity, that is, a large proportion of their undisturbed strength is permanently lost following shear (Figure 8). Consequently,

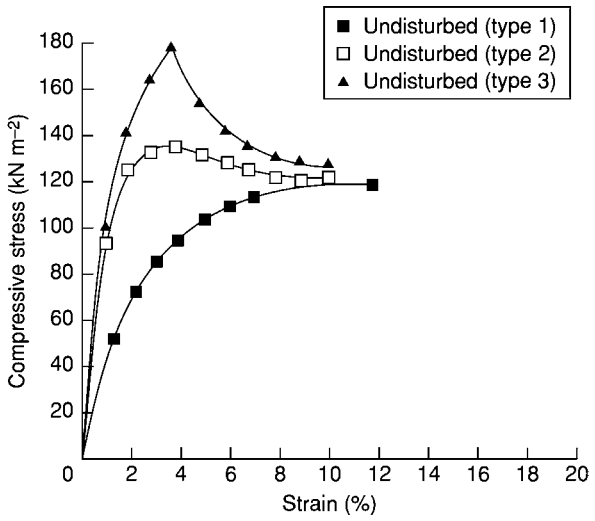


Figure 7 Examples of stress strain curves of undisturbed and remoulded Tees Laminated Clay tested in quick undrained tri axial conditions: (1) from Stockton, (2) from Cowpen Bewley, (3) from Middlesbrough.

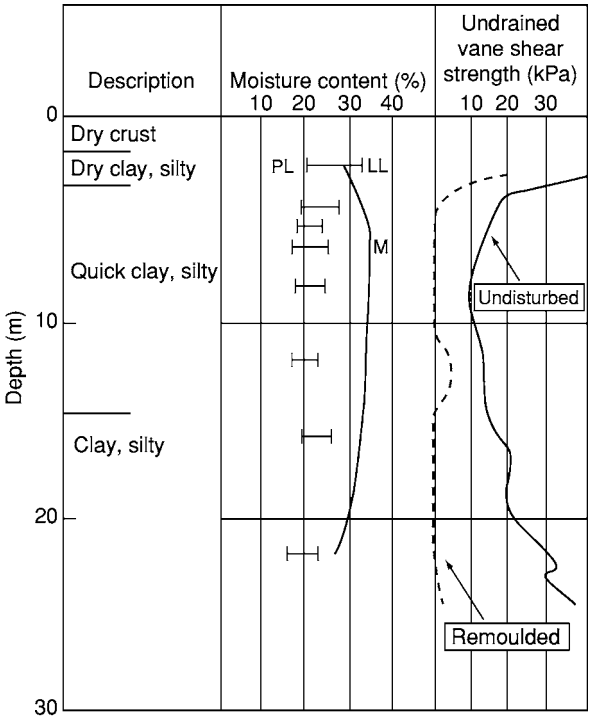


Figure 8 Moisture content, consistency indices, undrained shear strength, and sensitivity of quick clay from near Trondheim, Norway.

quick clays are associated with several serious engineering problems. Not only is their bearing capacity low, with large settlements occurring under load, but quick clays can liquefy on sudden shock. This can result in an almost instantaneous loss in strength. Slides, which have been referred to as mud-runs and bottleneck slides, therefore are a common feature of quick clay areas and sometimes have proved disastrous. Quick clays can be stabilized by adding salts or lime, which increase their remoulded shear strength.

Frozen ground phenomena are found in regions that experience a tundra climate, that is, in those regions where the winter temperatures rarely rise above freezing point and the summer temperatures are only warm enough to cause thawing in the upper metre or so of the soil. Beneath the upper or active zone, the subsoil is permanently frozen and so is known as permafrost. Because of the permafrost layer, summer meltwater cannot seep into the ground, the active zone then becomes waterlogged and soils on gentle slopes are liable to flow. Generally, the depth of thaw is less, the higher the latitude. It is at a minimum in peat or highly organic sediments and increases in clay, silt, and sand, to a maximum in gravel. Layers or lenses of unfrozen ground, termed taliks, may occur, often temporarily, in the permafrost.

The mechanical properties of frozen soil are influenced largely by the grain size distribution, mineral content, density, frozen and unfrozen water contents, and presence of ice lenses and layering. Ice has no long-term strength, that is, it flows under very small loads. When loaded, stresses at the point of contact between soil particles and ice bring about pressure melting of the ice. Because of differences in the surface tension of the meltwater, it tends to move into regions of lower stress, where it refreezes. The process of ice melting and the movement of unfrozen water are accompanied by breakdown of ice and bonding with grains of soil. This leads to plastic deformation of the ice in the voids and to a rearrangement of particle fabric. The net result is time-dependent deformation of the frozen soil, namely, creep. Frozen soil undergoes appreciable deformation under sustained loading, the magnitude and rate of creep being governed by the composition of the soil, especially the amount of ice present, the temperature, the stress, and the stress history.

As the soil thaws downwards the upper layers become saturated, and since water cannot drain through the frozen soil beneath, they may suffer a complete loss of strength. What is more, as ice melts, settlement occurs. Excess pore-water pressures develop when the rate of ice-melt is greater than the discharge capacity of the soil. This can lead to the failure of slopes and foundations.

Shrinkage, which gives rise to polygonal cracking of the ground, presents a problem when soil is subjected to freezing. Individual cracks may be over 1 m wide at their top and may penetrate to depths of 10 m. Water that accumulates in the cracks is frozen to form

ice wedges. When the ice disappears, an ice wedge pseudomorph is formed by sediment filling the crack. The ground also may undergo notable disturbance due to mutual interference of growing bodies of ice or excess pore-water pressures developed in confined water-bearing lenses. Involutions are plugs, pockets, or tongues of highly disturbed material, generally possessing inferior geotechnical properties, which have been intruded into overlying layers. Pseudomorphs of ice wedges and involutions usually mean that one material suddenly replaces another. This can cause problems in shallow excavations. Prolonged freezing gives rise to shattering in the frozen layer. Frost shattering may extend to significant depths and is responsible for an increase in deformability and permeability of the ground affected. It should be borne in mind that periglacial conditions extended over a much greater area of the land surface during the Pleistocene epoch and as a result the features mentioned are found in many present temperate climate regions such as Britain.

There are two types of methods of construction in permafrost, namely, passive and active methods. In the former, the frozen ground is not disturbed and heat from a structure is prevented from thawing the ground below, thereby reducing its stability. In other words, the aim is to maintain the ground beneath buildings in a frozen state. This is accomplished either by ventilation or by insulation. In the former case, a building is raised on piles above the ground so that air can circulate beneath it, thereby dissipating heat generated by the building. Alternatively, a building may be constructed on a layer of gravel to insulate the ground beneath. An insulating material such as polystyrene may be incorporated into the layer of gravel. The ground is thawed prior to construction in the active method. It is either kept thawed or removed and replaced by materials not affected by frost action. The active method is used where permafrost is thin, sporadic, or discontinuous, and where thawed ground has an acceptable bearing capacity. Foundations frequently are taken through the active layer into the permafrost beneath. Hence, piles often are used as foundation structures.

Frost action in a soil obviously is not restricted to periglacial regions. If frost penetrates down to the capillary fringe in fine-grained soils, especially silts then, under certain conditions, lenses of ice may be developed. The formation of such ice lenses may, in turn, cause frost heave and frost boil that may lead to the breakup of roads and the failure of slopes. Such problems are experienced in temperate climatic zones, as well as those of tundra regions.

Peat Soils

Peat represents an accumulation of partially decomposed and disintegrated plant remains that have been

Figure 9 Holme Post, a cast iron pillar erected in 1851 on the south west edge of Whittlesey Mere, near Peterborough, England, to indicate the amount of peat shrinkage and associated subsidence caused by drainage. The post was driven through 7 m of peat into clay until its top was flush with the ground. Within 10 years ground level had fallen 1.5 m through shrinkage. A second post was erected in 1957 with its top at the same level as that of the original post (right hand side). Between 1851 and 1971 the ground subsided by some 4 m.

preserved under conditions of incomplete aeration and high water content. It accumulates in areas where there is an excess of rainfall and the ground is poorly drained. Nonetheless, peat deposits tend to be most common in those regions with a comparatively cold wet climate. The high water-holding capacity of peat maintains a surplus of water, which ensures continued plant growth and consequent peat accumulation. Drying out, groundwater fluctuations, and snow loading bring about compression in the upper layers of a peat deposit. As the water table in peat normally is near the surface, the effective overburden pressure is negligible.

The void ratio of peat is very large, ranging from 9 up to 25. It usually tends to decrease with depth within a peat deposit. Such high void ratios give rise to phenomenally high water content, varying from a few hundreds per cent dry weight to over 3000% in some coarse fibrous varieties. Put another way, the water content may range from 75 to 98% by volume of peat. Peat, therefore, undergoes significant shrinkage on drying out. The magnitude of pore-water pressure is particularly significant in determining the stability of peat. With the exception of those peats with low water contents (less than 500%) and high mineral contents, the average bulk density of peat is slightly lower than that of water. Gas is formed in peat as plant material decomposes, the volume of gas varying from around 5 to 7.5%. Most of the gas is free and so has a significant influence on the rate of consolidation, pore pressure under load and permeability.

Differential and excessive settlement are the principal problems confronting the engineer working on a peat soil. Serious shearing stresses are induced, even by moderate loads. Worse still, should the loads exceed a given minimum, then settlement may be accompanied by creep, lateral spread or, in extreme cases, by rotational slip and upheaval of adjacent ground. At any given time, the total settlement in peat, due to loading, involves settlement with and without volume change. Settlement without volume change is the more serious for it can give rise to the types of failure mentioned above. What is more, it does not enhance the strength of peat.

Because of the potential problem of settlement arising from loading peat, especially in the construction of embankments carrying roads, some method of dealing with this problem has to be employed. Bulk excavation of peat frequently is undertaken if the deposit is less than 3 m in thickness. When a deposit exceeds 3 m or peat occurs as layers within soft sediments, precompression, involving surcharge loading, commonly is used. With few exceptions, improved drainage has no beneficial effect on the rate of consolidation.

When peatlands are drained artificially for reclamation purposes, the ground level can experience significant subsidence. The subsidence is not simply due to the consolidation that occurs as a result of the loss of the buoyant force of groundwater but also is attributable to desiccation and shrinkage associated with drying out in the zone of aeration and oxidation. For instance, in some parts of the Fenlands of eastern England, the thickness of peat has been almost halved as a result of drainage (e.g., Holme Post was installed in 1848 and by 1932 the thickness of the peat had been reduced from 6.7 m to 3.4 m; **Figure 9**).

See Also

Engineering Geology: Problematic Rocks; Rock Properties and Their Assessment; Site and Ground Investigation; Subsidence.

Further Reading

Andersland OB and Ladanyi B (1994) *An Introduction to Frozen Ground Engineering*. New York: Chapman and Hall.

Bell FG (1993) *Engineering Treatment of Soils*. London: E and FN Spon.

Bell FG (2000) *Engineering Properties of Soils and Rocks*, Fourth Edition. Oxford: Blackwell Scientific Publications.

Blight GE (ed.) (1997) *Mechanics of Residual Soils*. Rotterdam: AA Balkema.

Chen FH (1988) *Foundations on Expansive Soils*. Amsterdam: Elsevier.

Charman JM (1988) *Laterite in Road Pavements*. Special Publication 47, London: Construction Industry Research and Information Association (CIRIA).

Fookes PG (ed.) (1997) *Tropical Residual Soils*. Engineering Group Working Party Revised Report, London: Geological Society.

Fookes PG and Parry RHG (eds.) (1994) *Characteristics of Arid Soils*. Rotterdam: AA Balkema.

Jefferson I, Murray EJ, Faragher E, and Fleming PR (eds.) (2001) *Problematic Soils*. London: Thomas Telford Press.

Jefferson IF, Rosenbaum MS, and Smalley IJ (eds.) (2004) *Silt and Siltation, Problems and Engineering Solutions*. Berlin: Springer Verlag.

Rock Properties and Their Assessment

F G Bell, British Geological Survey, Keyworth, UK

Introduction

Those properties that can be used to describe rock materials in terms of engineering classification are referred to as index properties. Index properties frequently show a good correlation one with another. In order for the test of an index property to be useful it is desirable that it should be simple to obtain, rapidly performed, and inexpensive. The test results must be reproducible and the index properties must be relevant to the engineering requirement.

Density and Porosity

The density of a rock is one of its most fundamental properties. It is influenced principally by its mineral composition on the one hand and the amount of pore space on the other; as the proportion of pore space increases so the density decreases. Four different types of density are recognised. Firstly, grain density is the mass of the mineral aggregate per volume of solid material. The grain density is similar to the specific gravity (or relative density) of a rock except that the specific gravity is not expressed in units, it being the ratio of solid rock to that of an equal volume of water at a specified temperature. Secondly, dry density is the mass of the mineral aggregate per volume. Thirdly, bulk density is the mass of mineral aggregate and natural water content per volume. Fourthly, saturated density is the mass of mineral aggregate and saturated water content per volume. The unit weight can be used instead of density, and is expressed in terms of stress ($kN\,m^{-3}$). It can be derived from density simply by dividing by 98.8 (e.g. density of $2200\,kg\,m^{-3}$ $98.8 = 22.26\,kN\,m^{-3}$ unit weight; density is usually expressed in $Mg\,m^{-3}$).

The specific gravity can be determined by grinding a rock to powder and using a density bottle or alternatively by the immersion in water method. Determination of dry density, bulk density, and saturated density is dependent upon accurately weighing the rock specimen concerned and upon the accurate measurement of its volume. In the case of the dry density the rock specimen is dried in a ventilated oven at 105°C until a constant mass is reached and then allowed to cool in a desiccator. In order to obtain the saturated density, the rock specimen is first saturated by immersing in water under vacuum with periodic agitation to remove any trapped air. This method is not suitable for rocks that slake when immersed in water. The bulk density simply requires the mass of the specimen as obtained from the field, that is, with its natural moisture content. The volume of specimens can be determined by the caliper method or the buoyancy method. Regular shaped specimens are required for the caliper method so that accurate measurements of dimensions are obtained; the volume is obtained by immersion in water with the buoyancy method.

The porosity of a rock can be defined as the volume of the pore space divided by the total volume expressed as a percentage (Figure 1). Grades of dry density and porosity were suggested by the International Association of Engineering Geology (IAEG) and are provided in Table 1. Total or absolute porosity involves the total pore volume, that is, it includes the occluded pores (occluded pores are isolated and not interconnected with other pores). However, the effective or net porosity is a more practical measurement of porosity and it may be regarded as the pore space from which water or fluid can be removed. The

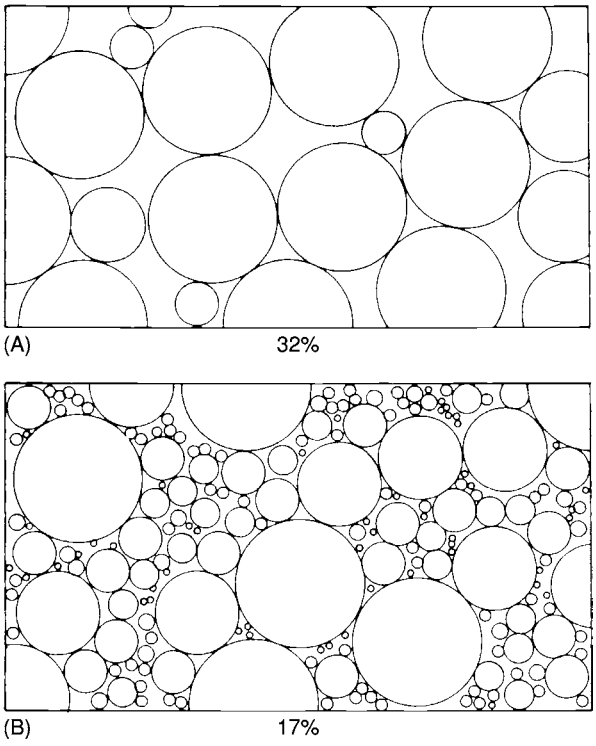


Figure 1 Different values of porosity attributable to different degrees of sorting. Numbers beneath diagrams refer to percentage porosity.

Table 1 Dry density and porosity

Class	Dry density (Mg m ^{3})	Description	Porosity (%)	Description
1	Less than 1.8	Very low	Over 30	Very high
2	1.8 2.2	Low	30 15	High
3	2.2 2.55	Moderate	15 5	Medium
4	2.55 2.75	High	5 1	Low
5	Over 2.75	Very high	Less than 1	Very low

Reproduced from Bulletin International Association Engineering Geology, No. 19, 364 371, 1979.

latter generally is determined by using the standard saturation method. Alternatively, an air porosimeter can be used. In both methods, the pore volume is obtained by saturating with water or with air, the total volume being found by the caliper or bouyancy method. The porosity as determined by these two tests does not provide an indication of the way in which the pore space is distributed within a rock, or whether it consists of many fine pores or a smaller number of coarse pores. Two tests have been used to investigate the distribution of pore sizes and the microporosity of a rock specimen, namely, the suction plate test and the mercury porosimeter test. Microporosity in the suction plate method is defined as the volume of water retained (expressed as a percentage of the total available pore space) when a suction equivalent to 6.4 m of head of water is applied to the specimen. In effect, this measures the percentage of pores with an effective diameter of less than 5 μm. Such pores are able to retain water against applied suction and influence the amount of damage that can be caused by frost or by the crystallization of soluble salts that a rock used for building stone may undergo. Mercury is forced to penetrate the pores of the specimen under an applied pressure in the mercury porosimeter test. Obviously, the finer the pores, the higher the pressure that must be used to bring about penetration. In this way it is possible to derive the dimensions and pore size distribution from a graph showing the distribution of pores sizes. A line is drawn on the curve at the position where 10% of the pore space has been filled with mercury. The pores below this size limit can be regarded as the microporosity.

Hardness

Hardness is one of the most investigated properties of materials, yet it is one of the most complex to understand. It does not lend itself to exact definition in terms of physical concepts. The numerical value of hardness is as much a function of the type of test used as a material property. The concept of hardness is usually associated with the surface of a material. For instance, the hardness of a rock can be considered as its resistance to a penetrating force, whether static or dynamic, or the resistance to displacement of surface particles by tangential abrasive force. As such, hardness is controlled by the efficiency of the bond between minerals or grains, as well as the strength of these two components.

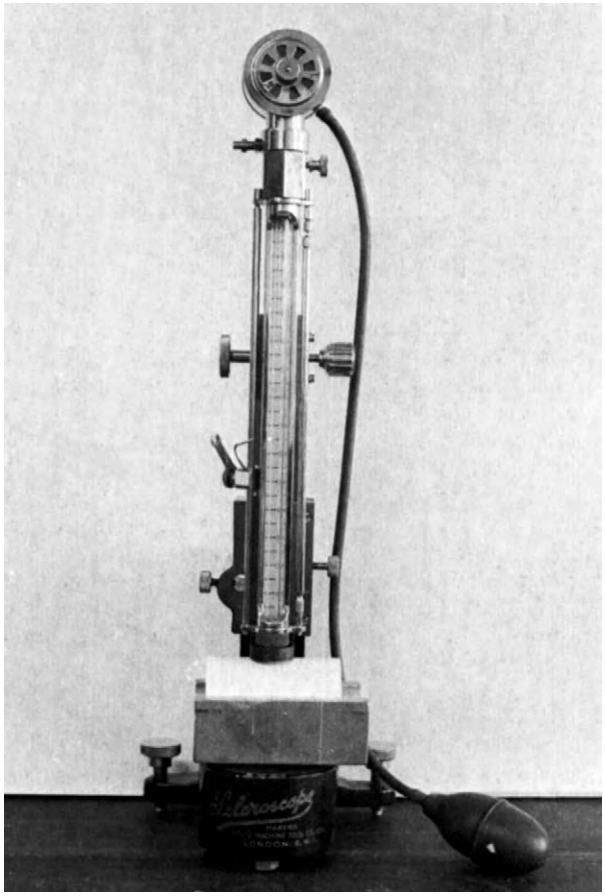


Figure 2 The Shore scleroscope.

A number of tests have been used to assess indentation hardness, of which the two most often used in rock mechanics testing are the Shore scleroscope and Schmidt hammer tests. The Shore scleroscope is a nondestructive hardness measuring device that indicates the relative values of hardness from the height of rebound of a small diamond-pointed hammer that is dropped vertically onto a securely clamped test surface from a height of 250 mm (Figure 2). Because a rock is not a homogeneous material, several hardness tests must be made over the surface of the specimen and the results averaged. Hence, at least 20 hardness determinations should be taken and each point of test should be at least 5 mm from any other. The Shore hardness value can be used to derive an approximate value of uniaxial compressive strength from Figure 3 for rocks with strengths in excess of 35 MPa.

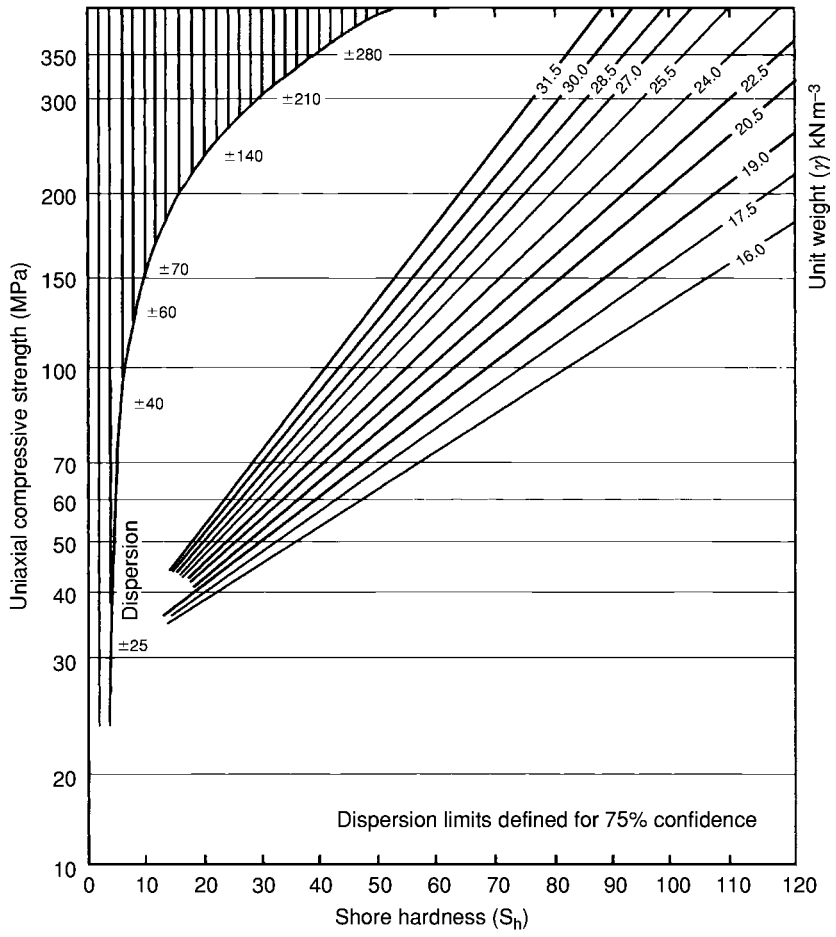


Figure 3 Correlation chart for Shore hardness, relating unit weight of rock, unconfined compressive strength and hardness value (After Deere DU and Miller RP (1966) *Engineering Classification of Engineering Properties for Intact Rock*. Technical Report AFWL TR 65 116, Air Force Weapons Laboratory, Kirkland Air Base, New Mexico).

The Schmidt hammer is a portable non-destructive device that expends a known amount of stored energy from a spring and indicates the degree of rebound of a hammer mass, following impact, within the instrument (Figure 4). Tests are made by placing the specimen in a rigid cradle and impacting the hammer at a series of points along its upper surface. The hammer is held vertically at right angles to the axis of the specimen. The specimens should have a flat smooth surface where tested and the rock beneath this area should be free from cracks. Test locations should be separated by at least the diameter of the plunger. At least 20 readings should be taken from each specimen. The lower 50% of the test values should be discarded and the average obtained from the upper 50%. This average is multiplied by the correction factor of the Schmidt hammer to obtain the hardness. However, the Schmidt hammer test is not a satisfactory method for the determination of the hardness of very soft or very hard rocks. Like the Shore scleroscope test, Schmidt hardness values can be used to derive approximate values of uniaxial compressive strength, as shown in Figure 5.

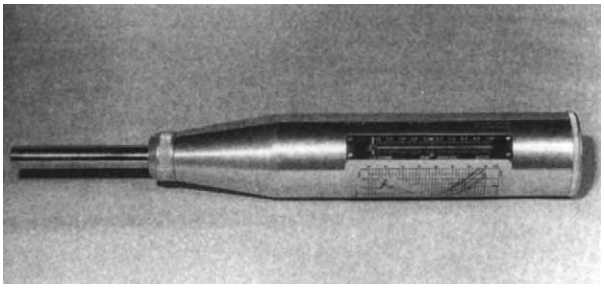


Figure 4 The Schmidt hammer. Reproduced from Engineering Geology, vol. 4, 1979, Elsevier.

Abrasion tests measure the resistance of rocks to wear. The two abrasion tests most frequently used are the Dorry and Los Angeles tests. As both these tests are used to assess the resistance to wear of aggregate for road making, they are not dealt with here.

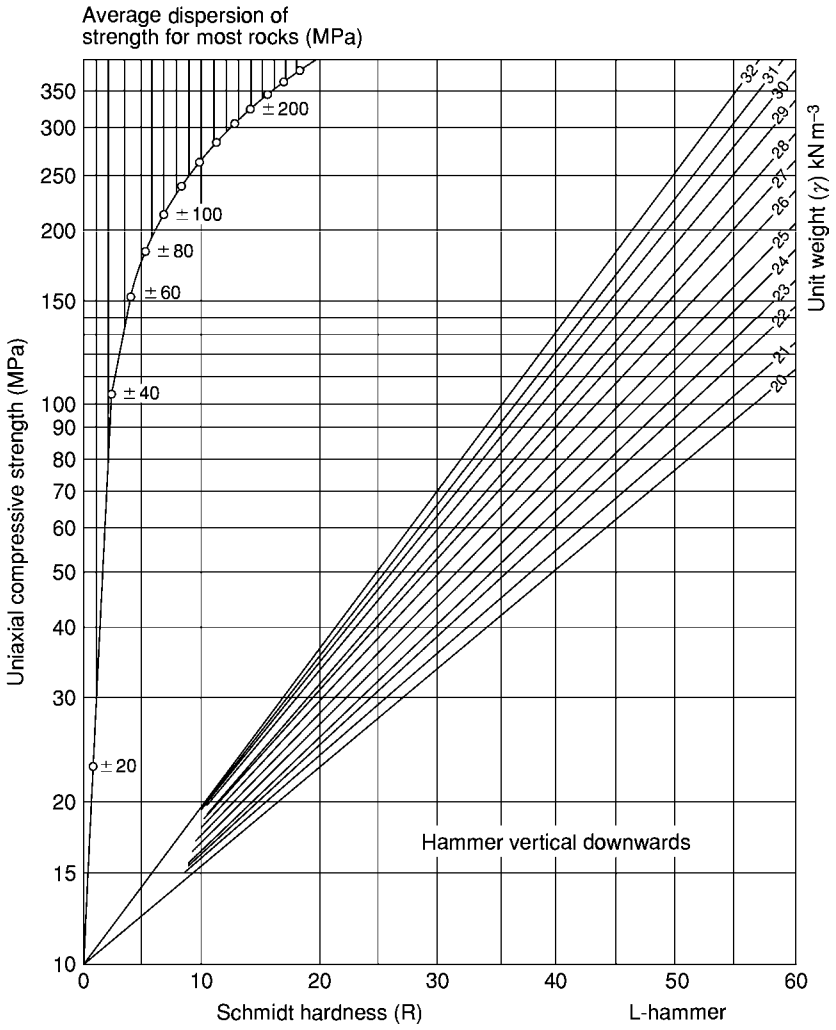


Figure 5 Correlation chart for Schmidt hardness, relating unit weight, unconfined compressive strength and hardness value (After Deere DU and Miller RP (1966) *Engineering Classification of Engineering Properties for Intact Rock*. Technical Report AFWL TR 65 116, Air Force Weapons Laboratory, Kirkland Air Base, New Mexico).

Deformation of Rocks

The factors that influence the deformation characteristics and the failure of rock can be divided into two categories: internal and external. The internal category includes the inherent properties of the rock itself, whilst the external category is that of its environment at a particular point in time. As far as the internal factors are concerned, the mineralogical composition and texture are obviously important, but fractures within a rock and the degree of mineral alteration are frequently more important. The temperature-pressure conditions under which a rock exists, significantly affect its mechanical behaviour, as does its pore water content. Importantly, the length of time that a rock experiences a change in stress, and the rate at which this is imposed, significantly affects its deformation characteristics.

The composition and texture of a rock are governed by its origin and geological history. Few rocks are composed of only one mineral species and even when they are the properties of that species vary slightly from mineral to mineral. Such variations within minerals may be due to cleavage, twinning, inclusions, cracking, and alteration, as well as to slight differences in composition. This, in turn, is reflected in the mechanical behaviour of the resultant rocks. As a consequence, few rocks can be regarded as homogeneous isotropic substances. The size and shape relationships of the component minerals/grains are also significant in this respect, generally the smaller the size, the stronger the rock. One of the most important features of texture as far as mechanical behaviour, particularly strength, is concerned, is the degree of

interlocking of the component minerals or grains. Breakage is more likely to take place along grain boundaries (intergranular fracture) than through grains (transgranular fracture) and therefore irregular boundaries make fracture more difficult. The bond between grains in many sedimentary rocks is provided by the cement and/or matrix, rather than by grains interlocking. The amount and, to a lesser extent, the type of cement/matrix is important, not only influencing strength and elasticity, but also density, porosity, and primary permeability. Rocks are not uniformly coherent materials, but contain defects that include microfractures, grain boundaries, mineral cleavage, twinning planes, inclusion trains, and elongated shell fragments. Obviously, such defects influence the ultimate strength of a rock and may act as surfaces of weakness that control the direction in which failure occurs. Crystal grain orientation in a particular direction facilitates breakage along that direction.

The presence of moisture in rocks adversely affects their engineering behaviour. For instance, moisture content increases the strain velocity and lowers the strength. More specifically the angle of internal friction is not affected significantly by changes in moisture content whereas the cohesion undergoes a notable reduction. It has therefore been suggested that the reduction in strength with increasing moisture content is due primarily to a lowering of the tensile strength, which is a function of the molecular cohesive strength of the material.

Although all rock types undergo a decrease in strength with increasing temperature and an increase in strength as the confining pressure is increased, the combined effect of these is notably different for different rock types. With increasing temperature there is a reduction in yield stress and strain hardening decreases. Heating enhances the ductility of rocks and their ability to deform permanently without loss of integrity. The transition from brittle to ductile deformation in porous rocks is characterized by an abrupt change from dilational behaviour at low stress to compaction during inelastic axial strain at high stress. This type of behaviour differs from that of rocks with low porosity. With the latter, dilatancy persists well into the ductile zone. The compaction that occurs during ductile deformation in porous rocks at high confining stress is due to collapse of the pore space and the rearrangement of grains to give more compact packing.

Four stages of deformation have been recognized, namely: elastic, elastico-viscous, plastic, and rupture. The stages are dependent upon the elasticity, viscosity, and rigidity of the rock, as well as on stress history, temperature, time, pore water, and anisotropy. An elastic deformation is defined as one that disappears when the stress responsible for it ceases. Ideal elasticity exists if the deformation on loading and its disappearance on unloading are both instantaneous. This is never the case with rocks since there is always some retardation, known as hysteresis, in the unloading process. With purely elastic deformation the strain is a linear function of stress, that is, the material obeys Hooke's law. Therefore, the relationship between stress and strain is constant, and is referred to as Young's modulus, E. Rock only approximates to an ideal Hookean solid. In fact, Young's modulus is not a simple constant but is related to the level of applied stress. Just how closely rock approximates to an ideal material depends on its homogeneity, isotropy, and continuity. Homogeneity refers to the physical continuity of a material, that is, the constituent particles are evenly distributed throughout its volume so that the elastic properties are the same at all points. Isotropy represents a measure of the directional properties of a rock. Hence, a rock is only isotropic if it is monomineralic and the crystals/grains have a random orientation. Since most rocks are composed of two or more essential minerals, which may possess preferred orientation, they are generally anisotropic. Continuity refers to the pore space and fractures within a rock. The degree of continuity affects the cohesion and so the transmission of stress throughout a rock.

The change in deformability at the elastic limit from elastic to plastic deformation is referred to as the yield point or yield strength. If the stress acting on a rock exceeds its elastic limit, then it becomes permanently strained, the latter being brought about by plastic flow. Within the zone of plastic flow there is a region where elastic stress is still important, referred to as the field of elastico-viscous flow. Plasticity may be regarded as time-independent, non-elastic, non-recoverable, stress-dependent deformation under uniform sustained load. Solids are classified as brittle or ductile according to the amount of plastic deformation they exhibit. In brittle materials the amount of plastic deformation is zero or very little, whereas it is large in ductile substances. Rupture occurs when the stress exceeds the strength of the material involved. It represents the maximum stress a rock is able to withstand prior to loss of cohesion by fracturing. The initiation of rupture is marked by an increase in strain velocity.

Most strong rocks exhibit little time-dependent strain or creep. However, creep in evaporitic rocks, notably salt, may greatly exceed the instantaneous elastic deformation. The time-strain pattern exhibited by such rocks, when subjected to a constant uniaxial stress, can be represented diagramatically as shown in Figure 6. The instantaneous elastic strain,

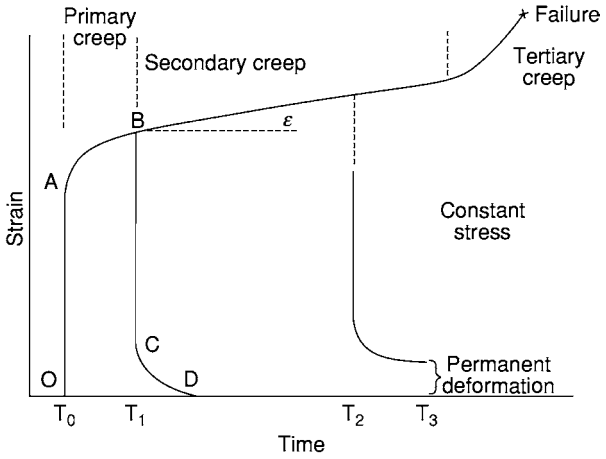


Figure 6 Theoretical strain curve at constant stress (creep curve).

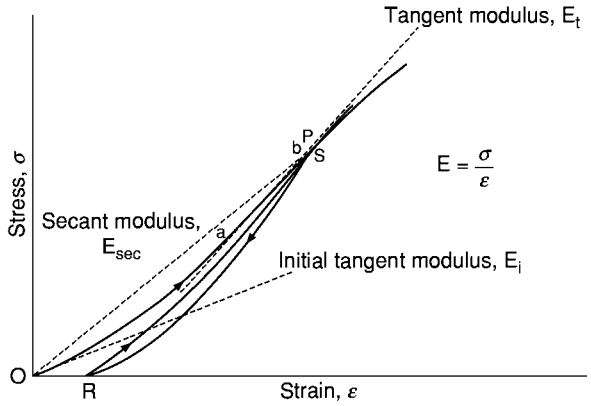


Figure 7 Representative stress strain curve for rock in unconfined compression, showing hysteresis.

which takes place when a load is applied, is represented by OA. There follows a period of primary or transient creep (AB) in which the rate of deformation decreases with time. If the stress is removed, the specimen recovers. At first this is instantaneous (BC), but this is followed by a time elastic recovery, illustrated by curve CD. On the other hand, if the loading continues, the specimen begins to exhibit secondary or pseudo-viscous creep. This type of creep represents a phase of deformation in which the rate of strain is constant. The deformation is permanent and is proportional to the length of time over which the stress is applied. If the loading is continued further, then the specimen suffers tertiary creep during which the strain rate accelerates with time and ultimately leads to failure.

Elastic Properties

Young's modulus, E, is the most important of the elastic constants and can be derived from the slope of the strain–stress curve obtained when a rock specimen is subjected to unconfined compression (i.e., static loading), it being the ratio of stress to strain. The strains are measured by attaching strain gauges to the test specimens, or by displacement transducers, and recording their outputs. Strain measurements on specimens less than 50 mm in diameter, however, are high and not representative of the material behaviour. Most crystalline rocks have S-shaped stress–strain curves (Figure 7). At low stresses the curve is non-linear and concave upwards, that is, Young's modulus increases as the stress increases. The initial tangent modulus is given by the slope of the stress-strain curve at the origin. Gradually a level of stress is reached where the slope of the curve becomes approximately linear. In this region Young's modulus is defined as the tangent modulus or secant modulus. At this stress level the secant modulus has a lower value than the tangent modulus because it includes the initial 'plastic' history of the curve. A classification of deformability has been suggested by the IAEG and is given in Table 2.

Table 2 Classification of deformability

Class	*Deformability (MPa × 10 ^{3})*	*Description*
1	Less than 5	Very high
2	5 15	High
3	15 30	Moderate
4	30 60	Low
5	Over 60	Very low

Reproduced from Bulletin International Association Engineering Geology, No. 19, 364 371, 1979.

In addition to their non-elastic behaviour, most rocks exhibit hysteresis. Under uniaxial stress the slope of the stress–strain curve during unloading initially is greater than during loading for all stress values (Figure 7). As stress is decreased to zero a residual strain, OR, is often exhibited. On reloading the curve RS is produced that, in turn, is somewhat steeper than OP. Further cycles of unloading and reloading to the same maximum stress give rise to hysteresis loops, which are shifted slightly to the right. The non-linear elastic behaviour and elastic hysteresis of brittle rocks under uniaxial compression is due to the presence of flaws or minute cracks in the rock. At low stresses these cracks are open but they close as the stress is increased and the rock becomes elastically stiffer, that is, E increases with stress. Once the cracks are closed the stress–strain curve becomes linear.

When a specimen undergoes compression it is shortened and this generally is accompanied by an increase in its cross-sectional area. The ratio of lateral

unit deformation to linear unit deformation, within the elastic range, is known as Poisson's ratio, v. This similarly can be obtained by monitoring strains during an unconfined compression test. The ideal geometrical value of Poisson's ratio is 0.333.

Another elastic constant is compressibility, K, which is the ratio of change in volume of an elastic solid to change in hydrostatic pressure. A further measure of elasticity is rigidity, G, which refers to the resistance of a body to shear. These four elastic constants (E, v, K, G) are not independent of each other and, if any two are known, it is possible to derive the other two from the following expressions:

$$G = E/2(1 + v) \quad [1]$$

and

$$K = E/3(1 - 2v) \quad [2]$$

Of the four constants, Young's modulus and Poisson's ratio are more readily determined experimentally.

Methods used to determine the dynamic values (as opposed to static values, see above) of Young's modulus and Poisson's ratio generally depend upon determining the velocities of propagation of elastic waves through a specimen of rock. These can be measured by using the high frequency ultrasonic pulse method, the low-frequency ultrasonic pulse technique, or the resonant method. For example, the high-frequency ultrasonic pulse method is used to determine the velocities of compressional, v_p, and shear, v_s, waves in rock specimens of effectively infinite extent compared to the wavelength of the pulse used. The condition of infinite extent is satisfied if the average grain size is less than the wavelength of the pulse that, in turn, is less than the minimum dimensions of the specimen. These two velocities can be substituted in the following expressions to derive the dynamic values of Young's modulus and Poisson's ratio:

$$E = \rho v_p^2 \frac{(1 + v)(1 - 2v)}{(1 - v)} \quad [3]$$

or

$$E = 2v_s^2 \rho (1 - v) \quad [4]$$

or

$$E = v_s^2 \rho \frac{[3(v_p/v_s)^2 - 4]}{[(v_p/v_s)^2 - 1]} \quad [5]$$

$$v = \frac{0.5(v_p/v_s)^2 - 1}{(v_p/v_s)^2 - 1} \quad [6]$$

where ρ is density.

Strength

Uniaxial Compression

The uniaxial strength, also known as the unconfined compressive strength, of a rock may be regarded as the highest stress that a rock specimen can carry when a unidirectional stress is applied, normally in an axial direction to the ends of a cylindrical specimen. It represents the maximum load supported by a specimen during the test divided by the cross-sectional area of the specimen. Grades of unconfined compressive strength are shown in Table 3. Although its application is limited, the uniaxial compressive strength allows comparisons to be made between rocks and affords some indication of rock behaviour under more complex stress systems.

The behaviour of rock in uniaxial compression is influenced to some extent by the test conditions. The most important of these is the length-diameter or slenderness ratio of the specimen, the most satisfactory slenderness ratio being 2.5 since it provides a reasonably good distribution of stress throughout

Table 3 Grades of unconfined compressive strength

Geological Society (1977)		*IAEG (1979)**		*ISRM (1981)***	
Term	*Strength (MPa)*	*Term*	*Strength (MPa)*	*Term*	*Strength (MPa)*
Very weak	Less than 1.25	Weak	Under 15	Very low	Under 6
Weak	1.25 5.00	Moderately strong	15 50	Low	6 10
Moderately weak	5.00 12.50	Strong	50 120	Moderate	20 60
Moderately strong	12.50 50	Very strong	120 230	High	60 200
Strong	50 100	Extremely strong	Over 230	Very high	Over 200
Very strong	100 200				
Extremely strong	Over 200				

*International Association of Engineering Geology.
**International Society for Rock Mechanics.

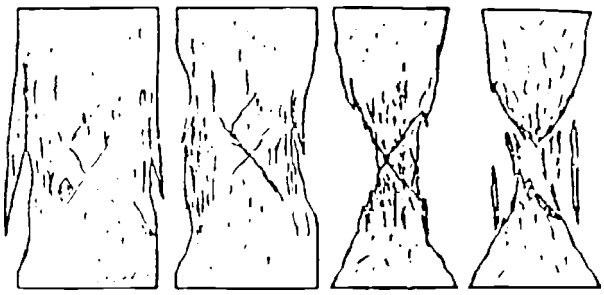


Figure 8 Stages in the development of fracturing with increasing unconfined compressive loading.

the specimen. Secondly, the rate at which loading occurs also influences the compressive strength, a loading rate of between 0.5 and 1.0 MPa normally is recommended. Thirdly, the ends of a specimen should be lapped so that they are exactly perpendicular to the long axis.

The onset of failure in a rock specimen subjected to compressive loading is marked initially by the formation of a large number of isolated fractures, which characterizes the relief of stress concentration produced by the mechanical inhomogeneities in the rock. Most cracks are orientated parallel to the applied stress (Figure 8). This is quickly followed by the development of two groups of macroscopic shear failures, at the boundary and in the interior of the specimen, which suggests that most of the major sources of induced lateral tensile stresses have been eliminated. The interior macroscopic shear failures are extended and become interconnected to form a conjugate set of open shear fractures. Two central cones are formed, which either abrade during the large shear displacement or produce major fractures in the remaining rock material. If one shear failure surface becomes dominant, then cones are not developed and the sample ultimately fails in two parts along a diagonal plane of shear.

Triaxial Compression Strength

A triaxial test is necessary if the complete nature of the failure of a rock is required. In this test a constant hydraulic pressure (the confining pressure) is applied to the cylindrical surface of the rock specimen, whilst applying an axial load to the ends of the sample. The axial load is increased up to the point where the specimen fails. Testing of the rock specimen is carried out within a special high pressure cell (Figure 9). A series of tests, each at higher confining pressure, are carried out on specimens from the same rock. These enable Mohr circles and their envelope to be drawn, from which the strength parameters, that is, the angle of friction (ϕ) and value of the cohesion (c) are obtained (Figure 10). The shear strength (τ) is then derived from the Mohr–Coulomb criterion:

$$\tau = c + \sigma \tan\phi \qquad [7]$$

where σ is the normal stress.

Direct Shear

A number of other tests can be used to deduce the shear strength of intact rock and of rock fracture surfaces, in addition to the triaxial test. These include the shear box test, the direct single and double shear tests, the punch shear test, and the torsion test. The most commonly used method is the shear box test, which is particularly useful for assessing the shear strength of weaker rocks and also for obtaining the shear strength along discontinuity surfaces. In this test a constant normal force is applied to the specimen, which then is sheared. A number of tests on the same material are undertaken, each at a higher normal stress, so that a shear strength versus normal stress graph can be drawn from which the value of cohesion and angle of friction are derived (Figure 11A). The other shear tests are used infrequently but are illustrated in Figure 11B.

Tensile Strength

Rocks have a much lower tensile strength than compressive strength. Brittle failure theory predicts a ratio of compressive strength to tensile strength of about 8:1 but in practice it is generally between 15:1 and 25:1. The direct tensile strength of a rock can be obtained by attaching metal end caps with epoxy resin to the specimen, which are then pulled into tension by wires. In direct tensile tests the slenderness ratio of cylindrical specimens should be 2.5 to 3.0 and the diameter preferably should not be less than 54 mm. The ratio of the diameter of the specimen to the largest crystal/grain in the rock should be at least 10:1. Unfortunately, the determination of the direct tensile strength often proves difficult since a satisfactory method has not yet been devised to grip the specimen without introducing bending stresses. Accordingly, most tensile tests have been carried out by indirect methods.

The Brazilian test is an indirect method of assessing the tensile strength of rocks, based on the observation that most rocks in biaxial stress fields fail in tension when one principal stress is compressive. In this test a cylindrical specimen of rock is loaded in a diametrical plane along its axis. The sample usually fails by splitting along the line of diametrical loading and the indirect tensile strength (T_b) can be obtained from:

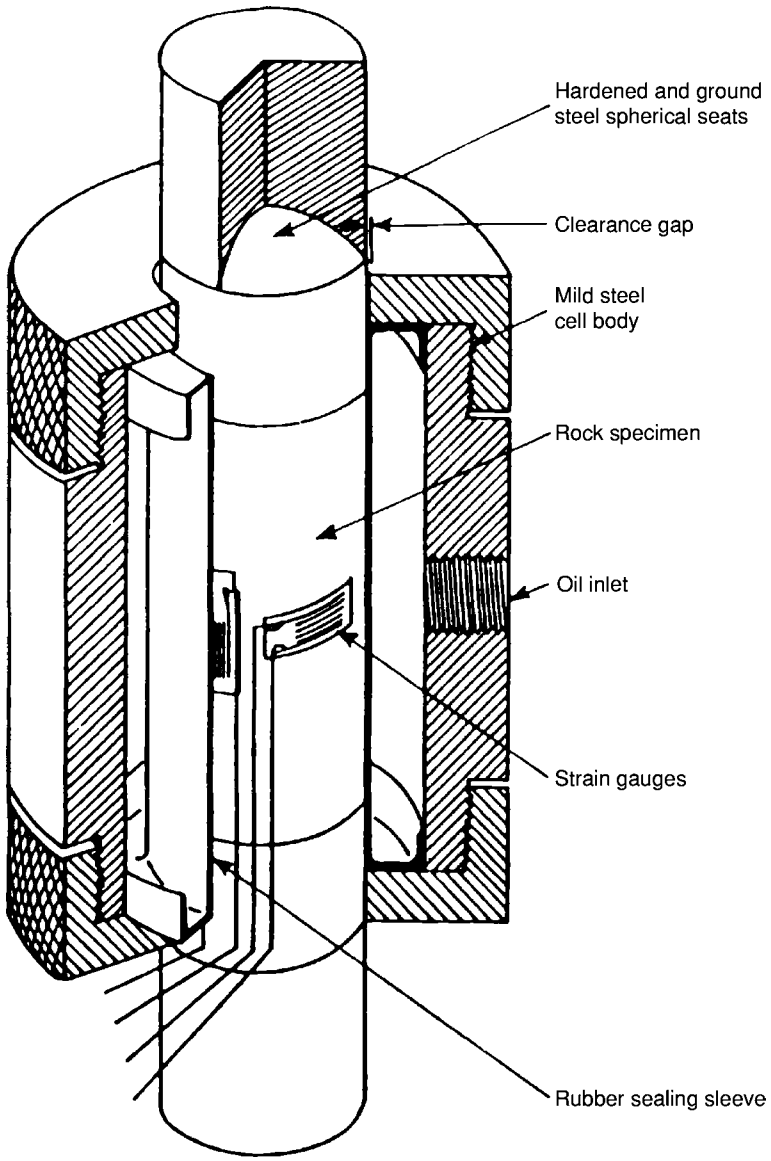


Figure 9 Cutaway view of a triaxial cell for testing rock.

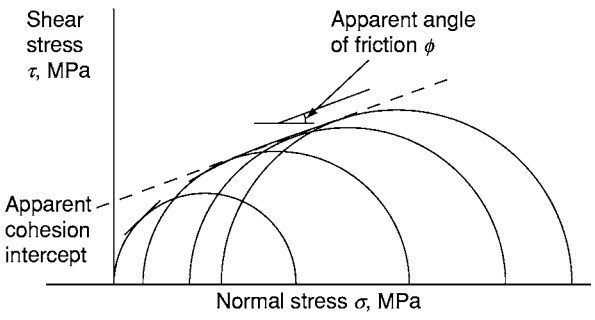


Figure 10 Mohr envelope and circles.

$$T_b = 2P/\pi LD \qquad [8]$$

where P is the load at failure, and L and D are the length and diameter of the specimen respectively. Disc-shaped specimens are used in the Brazilian disc test. In this case curved jaw loading platens are used to improve loading conditions. Uncertainties associated with the premature development of failure can be removed by drilling a hole in the centre of the disc-shaped specimen (sometimes this has been referred to as a ring test). A disc-shaped specimen should be wrapped around its periphery with a layer of masking tape and the specimen should not be less

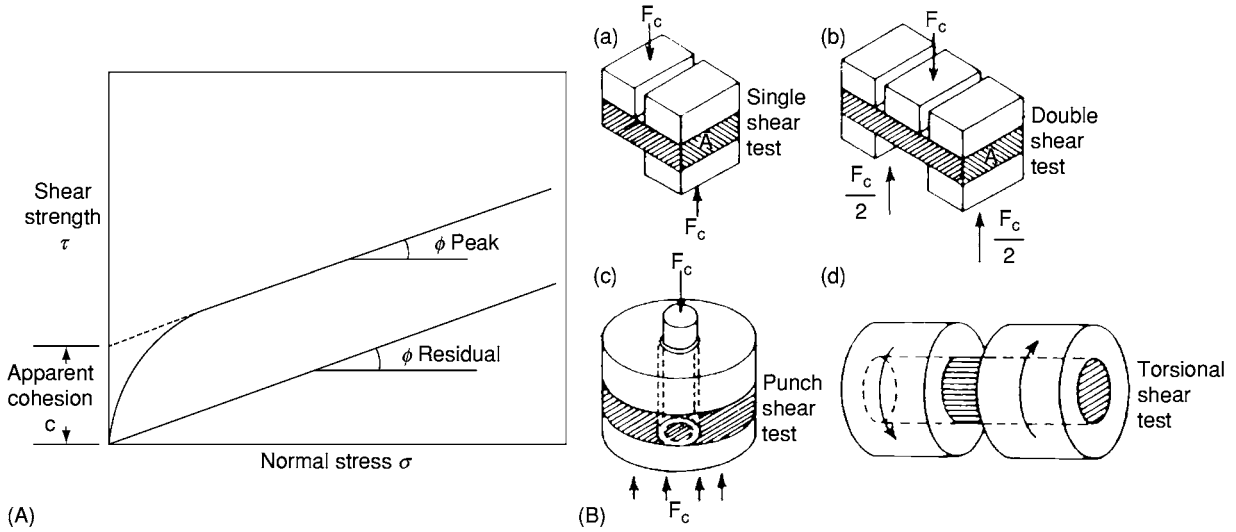


Figure 11 (A) Graph derived from shear test, showing curves derived for both peak and residual strength. (B) Other forms of shear strength tests.

54 mm in diameter, with a thickness that is approximately equal to the radius of the specimen. The tensile strength of the specimen is obtained as follows:

$$T_b = 0.636P/DH \quad [9]$$

where H is the thickness of the specimen. The Brazilian test is useful for brittle materials but for other materials it may give erroneous results.

In the point load test the specimen is placed between opposing cone-shaped platens and subjected to compression. This generates tensile stresses normal to the axis of loading and the indirect tensile strength (T_p) is then derived from:

$$T_p = P/D^2 \quad [10]$$

Loading can take place across the diameter of the specimen, as in Figure 12, or along the axis. Rocks that are anisotropic should be tested along and parallel to the lineation. Irregular-shaped specimens can also be tested, but at least 20 tests should be made on the same sample material and the results averaged to obtain a value. The point load test is limited to rocks with uniaxial compressive strengths exceeding 25 MPa (i.e., point load index above 1 MPa).

The effect of the size of specimens is greater in tensile than compression testing because in tension, cracks open and give rise to large strength reductions, whilst in compression the cracks close and so disturbances are appreciably reduced. This is especially the case in the axial and irregular lump point load tests. Accordingly, a standard distance between the two cones of 50 mm has been recommended, to which other sizes should be corrected by reference to a correction chart. Once determined, the point load index can be used to grade the indirect tensile strength of rocks, as shown in Table 4.

Finally, in the flexural test a cylindrical specimen of rock is loaded between one lower and two upper supports until the sample fails. The flexural strength gives a higher value of tensile strength than that determined in direct tension.

Durability of Rocks

Durability refers to the resistance that a rock offers to the various processes that lead to its breakdown and therefore durability tests can be used to provide a general impression of how a rock will behave in relation to weathering, especially mechanical weathering. Durability tests most frequently are used to assess the behaviour of suspect rocks, that is, those that tend to breakdown more readily such as mudrocks, some chalks, and certain basalts and dolerites. There are a large number of tests that have been used to assess the durability and many of them are used to determine the durability of rock as a material for building purposes, for aggregate, or for armourstone. The latter types of test are not dealt with here.

Some of the more simple tests include the water absorption test, the wet and dry test, the freeze-thaw test, and soak tests. The water absorption test involves oven drying a rock specimen at 105°C until it has attained a constant weight and then saturating it under vacuum. The percentage saturation is determined and reflects porosity. As rocks break down, their porosity increases and so the water absorption test has been used to indicate the degree of

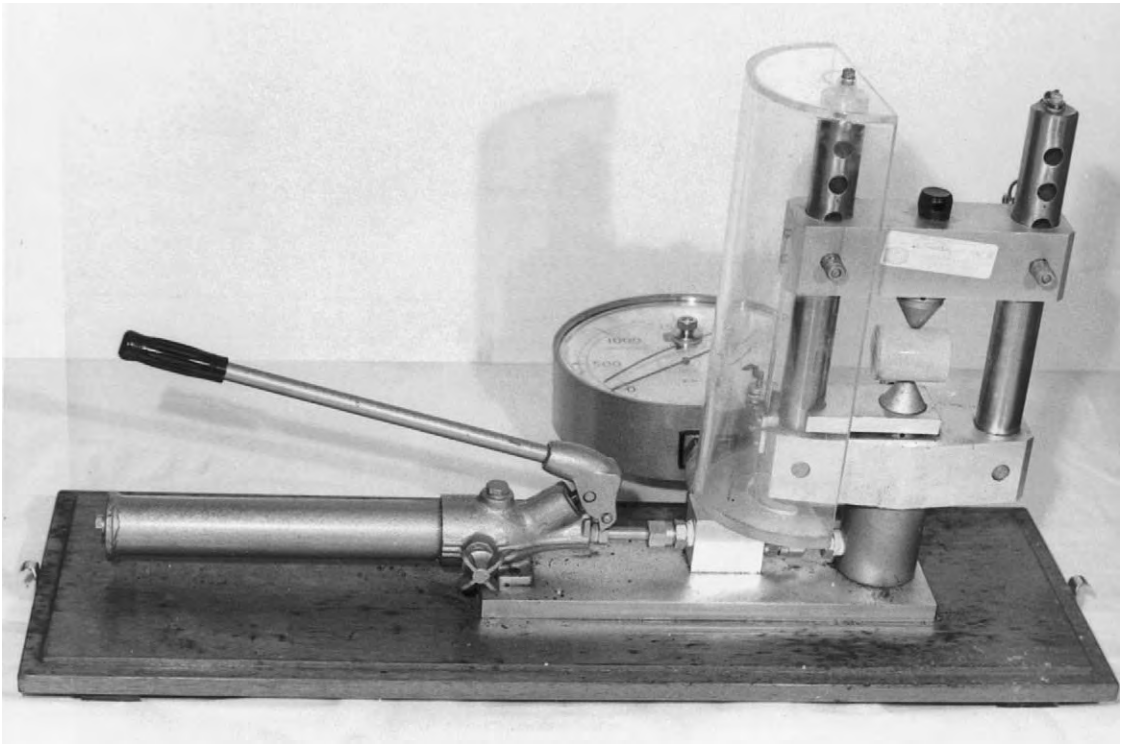


Figure 12 Point load test apparatus.

Table 4 Point load strength classification

Description	*Point load strength index (MPa)*	*Equivalent uniaxial compressive strength (MPa)*
Extremely high strength	Over 10	Over 160
Very high strength	3 10	50 160
High strength	1 3	15 60
Medium strength	0.3 1	5 16
Low strength	0.1 0.3	1.6 5
Very low strength	0.03 0.1	0.5 1.6
Extremely low strength	Less than 0.03	Less than 0.5

weathering a rock, especially a crystalline rock such as granite, has undergone. In other words, is it fresh, slightly weathered, moderately weathered, or highly weathered. However, some moderately and highly weathered rocks may break down before becoming saturated. Similarly, in the wet and dry test the rock specimen is first dried and then saturated, but this time for a given number of cycles. The effect that cyclic wetting and drying has on the specimen is recorded (i.e., no effect, softening, minor spalling, minor hairline cracking, severe hairline cracking, breakdown before last cycle noting cycle number). In the freeze-thaw test the specimen is saturated and then frozen for 24 h. The specimen is subjected to a given number of cycles and the effects are recorded in the same way as for the wet and dry test. The freeze-thaw test has been used to assess the frost resistance of building stone. However, it is no longer used for this purpose in Britain, it being regarded as unsatisfactory primarily because of the difficulty of interpreting the results in relation to a period of time over which a rock will perform as required. Soak tests are used to assess the breakdown of rocks as a result of swelling brought about by the absorption of water, especially of those rocks that contain swelling minerals. The rock specimen either may be soaked in water or ethylene glycol $(CH_2OH)_2$ for a given number of days. Ethylene glycol is much more effective than water as far as assessment of those rocks that contain swelling minerals are concerned. The soak test allows five classes of rock disintegration and the time when the worst condition occurs to be recognized:

Degree of disintegration:

Class 1 : No obvious effects, or only very minor spalling of sand-sized particles.
Class 2 : Flaking and/or swelling.
Class 3 : Fracturing without extensive spalling.
Class 4 : Fracturing with extensive spalling.
Class 5 : Complete disintegration.

Time required to reach worst condition:

Class 6 : 0–1 day
Class 5 : 2–3 days

Table 5 Ethylene glycol soak test index values

	Time class						
Degree of disintegration class	*6 0 1 day*	*5 2 3 days*	*4 4 10 days*	*3 11 15 days*	*2 16 20 days*	*1 21 30 days*	*0 Over 30 days*
1	7	6	5	4	3	2	1
2	8	7	6	5	4	3	2
3	9	8	7	6	5	4	3
4	10	9	8	7	6	5	4
5	11	10	9	8	7	6	5

Class 4 : 4–10 days
Class 3 : 11–15 days
Class 2 : 16–20 days
Class 1 : 21–30 days
Class 0 : Over 30 days

An index value can be derived from the integration of the degree of disintegration and time taken to reach the worst condition. In other words, a scale of relative durability values can be developed, ranging from 1 in which specimens are more or less unaffected to 11 when they are totally disintegrated (Table 5). The ethylene glycol soak test has been used to distinguish certain unsound basalts and dolerites, commonly described as slaking types, from sound types, particularly in relation to their use as road aggregate.

Slaking refers to the breakdown of rocks, especially mudrocks, by alternate wetting and drying. If a fragment of mudrock is allowed to dry out, air is drawn into the outer pores and high suction pressures develop. When the mudrock is next saturated the entrapped air is pressurised as water is drawn into the rock by capillary action. This slaking process causes the internal arrangement of grains to be stressed. Given enough cycles of wetting and drying, breakdown can occur as a result of air breakage, the process ultimately reducing the mudrock to gravel-sized fragments. The slake-durability test estimates the resistance to wetting and drying of a rock sample, and is particularly suitable for mudrocks and shales. The sample, which consists of 10 pieces of rock, each weighing about 40 g, is placed in a test drum, oven dried, and weighed. After this, the drum, with sample, is half immersed in a tank of water and attached to a rotor arm, which rotates the drum for a period of 10 min at 20 revolutions per minute (Figure 13). The cylindrical periphery of the drum is formed using a 2 mm sieve mesh, so that broken down material can be lost whilst the test is in progress. After slaking, the drum and the material retained are dried and weighed. The slake-durability index is then obtained by dividing the weight of the sample retained by its original weight, and expressing the answer as a percentage. The following scale of slake-durability is used:

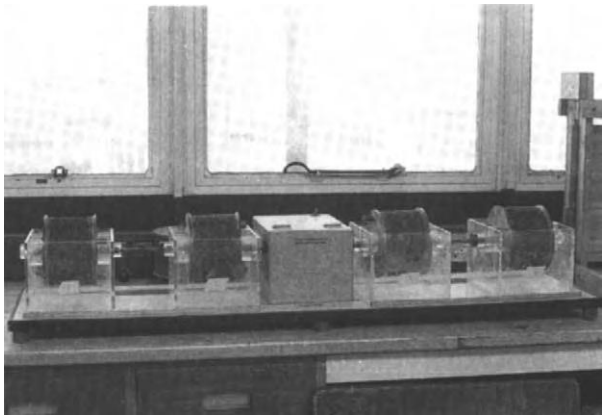


Figure 13 Slake durability test apparatus.

under 25%, very low
25–50%, low
50–75%, medium
75–90%, high
90–95%, very high
over 95%, extremely high.

However, it has been suggested that durable mudrocks may be better distinguished from non-durable types on the basis of compressive strength and three-cycle slake-durability index (i.e. those mudrocks with a compressive strength of over 3.6 MPa and a three cycle slake-durability index in excess of 60% are regarded as durable). In fact, the value of the slake durability test as a means of assessing mudrock durability has been questioned as the results obtained frequently do not compare well with those of other durability tests. This has led to a number of adaptations being made to the test, for example, ethylene glycol has been used instead of water, the time taken to carry out the test has been extended and the number of cycles has been increased.

Failure of weak rocks occurs during saturation when the swelling pressure (or internal saturation

swelling stress, σ_s) developed by capillary suction pressures, exceeds their tensile strength. An estimate of σ_s can be obtained from the modulus of deformation (E):

$$E = \sigma_s/\varepsilon_D \qquad [11]$$

where ε_D is the free swelling coefficient. The latter is determined by a sensitive dial gauge recording the amount of swelling of an oven-dried core specimen per unit height along the vertical axis during saturation in water for 12 h, ε_D being obtained as follows:

$$\varepsilon_D = \frac{\text{change in length after swelling}}{\text{initial length}} \qquad [12]$$

A durability classification has been developed based on the free-swelling coefficient and uniaxial compressive strength (Figure 14).

When the free expansion of rocks liable to swell is inhibited, stresses of sufficient magnitude to cause damage to engineering structures may develop. A number of tests may be conducted that provide an indication of the swelling behaviour of such rock. These include the swelling strain index test performed under unconfined conditions, and the swelling pressure index test, carried out under conditions of zero volume change. Swelling strain measurements can also be undertaken on radially confined specimens under various conditions of axial loading. The swelling strain represents maximum expansion of an unconfined rock specimen when it is submerged in water. Test specimens may either be oven-dried or retain their natural moisture content. As a test specimen is submerged in water it is advisable to prevent rock prone to slaking from collapsing before the ultimate swelling strain has developed by wrapping the specimen in muslin. During the test the specimen is supported by a frame between a fixed point and a measuring point. The latter may consist of a dial gauge or a deformation transducer. The specimen and frame are placed in a container that is filled

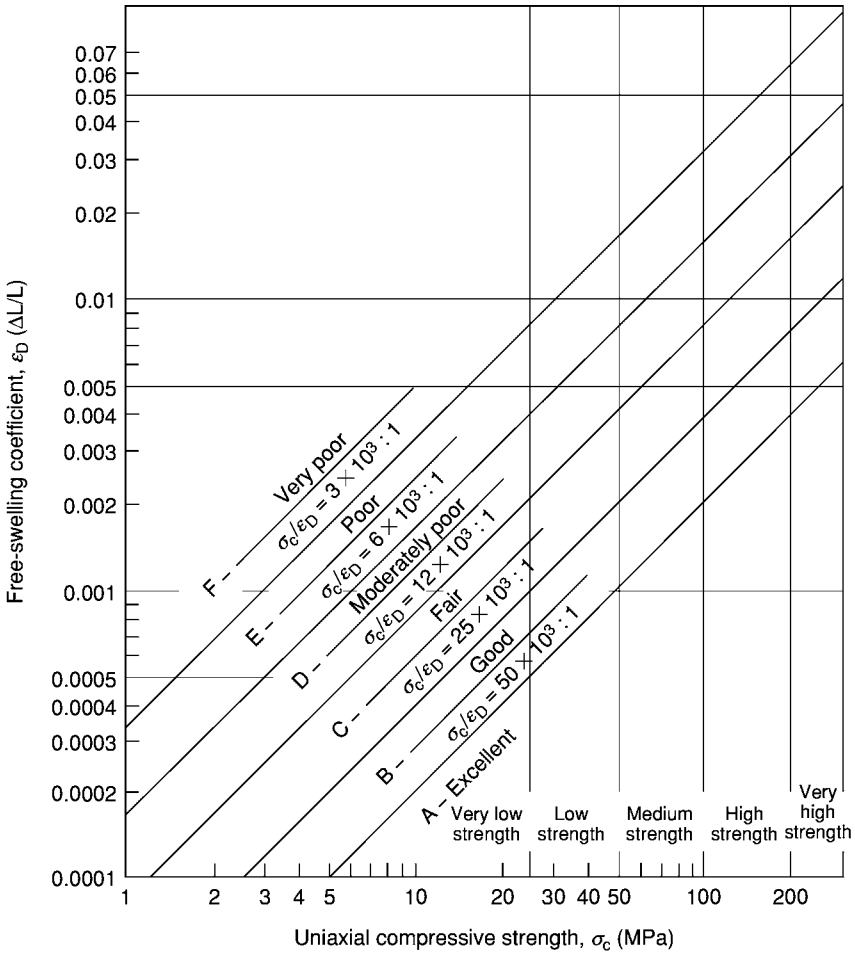


Figure 14 Geodurability classification chart (After Olivier HJ (1979) A new engineering geological rock durability classification. *Engineering Geology* 4: 255 279).

with water just above the level of the specimen. Strain is measured in all three perpendicular directions. No strain is allowed to develop in the test specimen during the swelling pressure test. This is accomplished by tightly mounting a cylindrical specimen inside a rigid ring that provides radial constraint. The specimen absorbs water through porous end plates. Any axial strain that develops is monitored and compensated for by increasing the axial load. The stress required to prevent expansion when equilibrium conditions are established is equal to the swelling pressure index.

Permeability

Permeability considers the ability of a rock to allow the passage of fluids into or through it without impairing its fabric. In ordinary hydraulic usage, a substance is called 'permeable' when it permits the passage of a measurable quantity of fluid in a finite period of time and 'impermeable' when the rate at which it transmits that fluid is slow enough to be negligible under existing temperature-pressure conditions. The permeability of a particular rock is defined by its coefficient of permeability or hydraulic conductivity. Grades of permeability are given in Table 6.

Determination of the permeability of many rock types in the laboratory is made by using a falling-head permeameter (Figure 15A). The sample is placed in the permeameter, which is then filled with water to a certain height in the standpipe. The stopcock is then opened and the water allowed to infiltrate the

Table 6 Grades of permeability (IAEG, 1979)*

	Permeability	
Class	*m s* ^{1}	*Description*
1	Greater than 10^{2}	Very highly
2	10^{2} 10^{4}	Highly
3	10^{4} 10^{5}	Moderately
4	10^{5} 10^{7}	Slightly
5	10^{7} 10^{9}	Very slightly
6	Less than 10^{9}	Practically impermeable

*International Association of Engineering Geology.

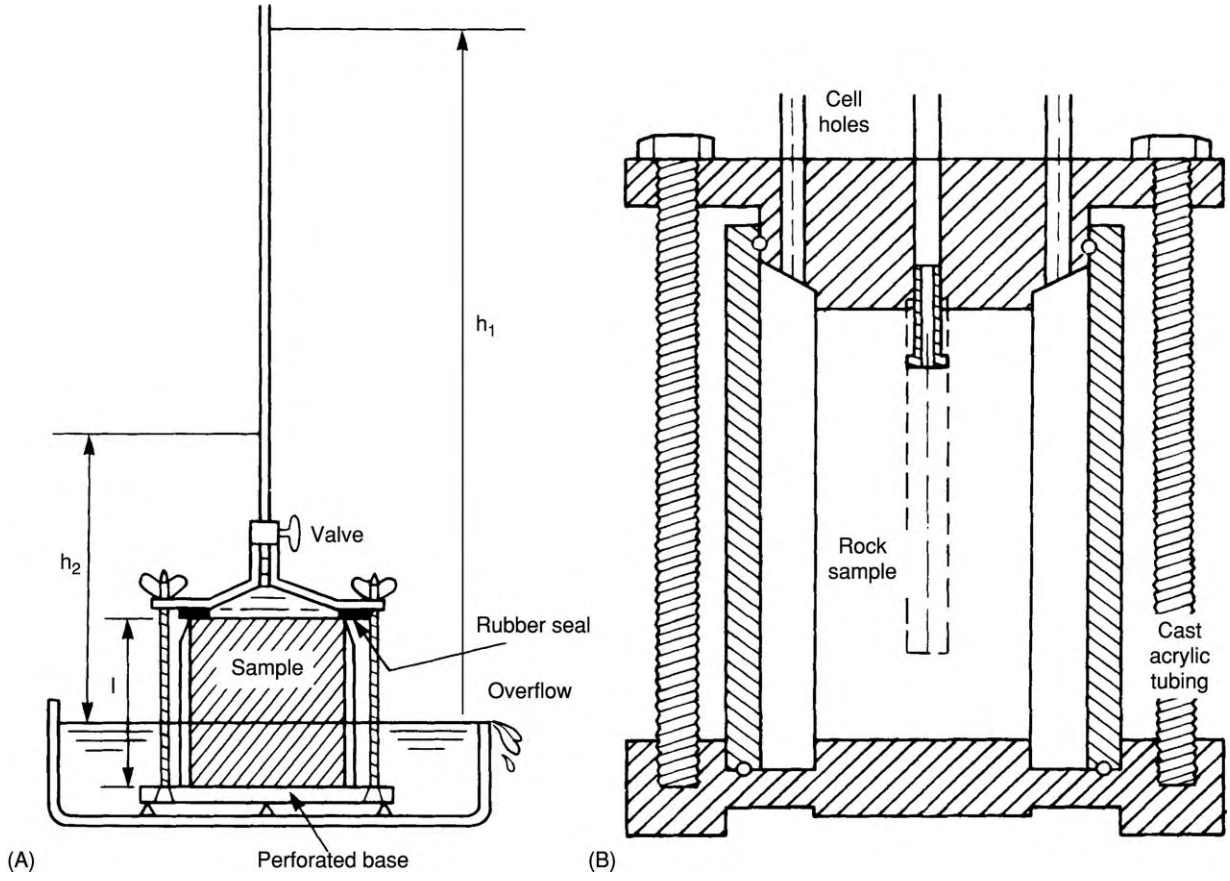


Figure 15 (A) Falling head permeameter. (B) Radial percolation test apparatus.

sample, the height of the water in the standpipe falling. The times at the beginning, t_1, and end, t_2, of the test are recorded and these, together with the two corresponding heights, h_1 and h_2, the cross-sectional area of the standpipe, a, and cross-sectional area, A, and length, l, of specimen are substituted in the following expression to derive the coefficient of permeability, k:

$$k = \frac{2.303al}{A(t_2 - t_1)} \times (\log_{10}h_1 - \log_{10}h_2) \qquad [13]$$

Variations of permeability in rocks under stress can be obtained by using a radial percolation test. A cylindrical specimen, in which an axial hole is drilled, is placed in the radial percolation cell. The latter can either contain water under pressure where the axial hole is in contact with atmospheric pressure; or water can be injected under pressure into the hole (Figure 15B). The flow is radial over almost the whole height of the sample and is convergent when the water pressure is applied to the outer face of the specimen, and divergent when the water is under pressure within the axial hole. Porous rocks remain more or less unaffected by pressure changes. On the other hand, fissured rocks exhibit far greater permeability in divergent flow than in convergent flow. Moreover, fissured rocks exhibit a continuous increase in permeability as the pressure attributable to divergent flow is increased.

See Also

Aggregates. **Engineering Geology:** Codes of Practice; Natural and Anthropogenic Geohazards; Problematic Rocks.

Further Reading

Anon (1975) *Methods of Sampling and Testing Mineral Aggregates, Sands and Fillers, BS 812*. London: British Standards Institution.

Anon (1982) *Standard Test Methods for Absorption and Bulk Specific Gravity of Natural Building Stone, C93 117*. Philadelphia: American Society for Testing Materials.

Bell FG (ed.) (1992) *Engineering in Rock Masses*. Oxford: Butterworth Heinemann.

Bell FG (2000) *Engineering Properties of Soils and Rocks*. Oxford: Blackwell Science.

Brown ET (ed.) (1981) *Rock Characterization, Testing and Monitoring*. Oxford: Pergamon Press.

Farmer IW (1983) *Engineering Behaviour of Rocks*, 2nd ed. London: Chapman and Hall.

Goodman RE (1989) *An Introduction to Rock Mechanics*, 2nd ed. New York: Wiley.

Hudson JA and Harrison JP (1997) *Engineering Rock Mechanics: An Introduction to the Principles*. Oxford: Pergamon.

Site and Ground Investigation

J R Greenwood, Nottingham Trent University, Nottingham, UK

Introduction and Terminology

The procedure of 'investigation' is fundamental to any project or activity involving the ground. The historical records need to be reviewed, current conditions need to be established, and the consequences of the proposed activity, works, or construction need to be carefully considered.

Investigation is an on-going process of establishing and reviewing the facts and processing the information to assist our future activities. With respect to construction works the following definitions are used:

- 'Site investigation' is a continuous process, as the construction project develops, involving both the site under consideration and the interaction with the surrounding areas. It is not confined to obtaining information on geotechnical aspects but may include hydrological, meteorological, geological, and environmental investigation.
- 'Ground investigation' is more site-specific and aims to investigate ground and groundwater conditions in and around the site of a proposed development or an identified post-construction problem.

The term 'site characterization' is now also used; it stems from the environmental specialist's study of contaminated sites but is equally applicable to any site. 'Characterization' perhaps implies the results of 'investigation'.

This article reviews the procedures necessary for quality site investigation to be carried out and describes some of the ground investigation techniques

commonly applied in advance of construction or remedial works.

Responsibilities

Investigation work must be entrusted to competent professionals with appropriate geotechnical engineering or engineering geology experience. Advice on the qualifications and experience required of such a professional is given in the publications of the Site Investigation Steering Group. The Geotechnical Advisor, who is an appropriately experienced Chartered Engineer or Chartered Geologist in the UK, will be a key figure in ensuring the correct geotechnical input to the project and should be appointed at an early stage and continue to work with the project team throughout the life of the project. Projects proceeding without the benefit of specialist geotechnical advice are more likely to encounter unforeseen problems, resulting in delay and budget overspend.

The Investigation Process

Investigation Stages

The investigation work for most projects can be divided into stages, as illustrated in Table 1. The Geotechnical Advisor will ensure appropriate geotechnical input at each stage (*see* **Engineering Geology:** Codes of Practice).

The Procedural Statement

The key to successful investigation lies in the planning process. If all aspects of the investigation work are considered in advance, together with necessary actions relating to the probable findings, then the outcome is likely to be satisfactory for all parties involved.

A convenient way to bring together and record the proposals for each stage of site and ground investigation is by preparing a 'Procedural Statement' or 'Statement of Intent'. This approach was formally introduced for the United Kingdom by the Department of Transport/Highways Agency in the 1980s and has now become widely accepted as good practice. An example of the topics covered in a Procedural Statement is given in Table 2. Headings and content will change slightly for each phase of the investigation process as more information is accumulated. The Procedural Statement is usually prepared by the geotechnical engineer or specialist responsible for the work and should be agreed by all interested parties, and in particular the client, before the investigation proceeds.

The Procedural Statement encourages the designer to consider relevant aspects of the proposed investigation and to seek authority to proceed. It forms a valuable document within a quality management system, and it becomes a base reference as the investigation proceeds in case changes are needed in the light of the findings.

The Desk Study

The desk study, sometimes referred to as the 'initial appraisal' or 'preliminary sources' study, is vital for gaining a preliminary understanding of the geology of the site and the likely ground behaviour. The term 'desk study' can be misleading because, in addition to collection and examination of existing information, it must include a walk-over survey. The study will determine what is already known about the site and how the ground should be investigated.

Before embarking on groundwork, much valuable information may be readily gleaned from existing

Table 1 Stages of an investigation (developed from Clayton CRI, Matthews MC, and Simons NE (1995) *Site Investigation: A Handbook for Engineers*, 2nd edn. Oxford: Blackwell Scientific)

Construction phase	*Investigation work*
Definition of project	Appointment of Geotechnical Advisor for advice on likely design issues
Site selection	Preliminary sources study (desk study) to provide information on the relative geotechnical merits of available sites
Conceptual design	Detailed preliminary sources study (desk study) and site inspections to provide expected ground conditions and recommendations for dealing with particular geotechnical design aspects and problems Plan ground investigation (Procedural Statement)
Detailed design	Full ground investigation and geotechnical design; (additional ground investigation if necessary for design changes or for problematic ground conditions)
Construction	Comparison of actual and anticipated ground conditions; assessment of new risks (additional ground investigation if necessary)
Performance/ maintenance	Monitoring, instrumentation, feedback reporting

Table 2 Example of a Procedural Statement's contents to be prepared before the ground investigation phase (HD 22/02)

The Procedural Statement (sometimes referred to as 'statement of intent' or the 'ground investigation brief') should be prepared by the responsible Geotechnical Advisor and agreed by the client and interested parties
1. Scheme
Details of scheme and any alternatives to be investigated; key location plan
2. Objectives
(For example) to provide information to confirm and amplify the geotechnical and geomorphological findings of the desk study as reported separately and to obtain detailed knowledge of the soils encountered and their likely behaviour and acceptability (for earthworks). To ascertain groundwater conditions and location of any underground workings (work limits to be defined)
3. Special problems to be investigated
Location of structures. Subsoil conditions below high embankments. Aquifers and likely water bearing strata affecting the proposed works. Rock stability problems. Manmade features to be encountered. Effects on adjacent properties
4. Existing information
List of all relevant reports and data
5. Proposed investigation work
Fieldwork Details of exploratory work proposed for specific areas with reasons for choice of investigation methods selected. Proposed sampling to match laboratory testing
Laboratory work Details of proposals with reasons for choice of tests and relevance to design
6. Site and working restrictions
Assessment of risk associated with proposals. Site safety, traffic management, difficult access, railway working
7. Specialist consultation
Details of specialist needed to support proposals
8. Programme, cost, and contract arrangements
Anticipated start date, work programme, contract arrangements, cost estimates, specification and conditions of contract. Arrangements for work supervision
9. Reporting
Responsibility for factual and interpretive reporting. Format of reports and topics to be covered

Table 3 Some sources of information for desk study work (preliminary sources study)

Topic	*Possible sources*
Site topography	Topographical maps, aerial photographs
Geology and soil conditions	Geological maps, regional guides and publications (British Geological Survey, sheet memoirs), learned society journals
Geotechnical problems and parameters	Published technical journal articles, civil engineering and geological journals, newspapers, previous ground investigation reports, local authorities
Groundwater conditions	Topographical maps, aerial photographs, well records, previous ground investigation reports, water authorities (flood records)
Meterological conditions	Meterorological Office
Existing site use and services	As built drawings, utility information, mining records, construction press, land use maps, commercial records, contamination records
Previous land use	Aerial photographs, old maps, archaeological records, agricultural records, mining records, local resident knowledge, local library records

sources, such as geological and Ordnance Survey maps, aerial photographs, and archival material. Such documents can yield significant information about site conditions and, following the walk-over survey, a geotechnical plan of the site may be prepared. A checklist of the type of information to be sought in a desk study is given in Table 3.

The desk study is often the most cost-effective element of the entire site-investigation process, revealing facts that cannot be discovered in any other way. The preliminary engineering concepts for the site are prepared and developed at the desk-study phase, based on the acquired information. The ground investigation in the field is then designed to confirm that the conditions are as predicted and to provide ground information for the detailed design and project construction.

Figure 1 illustrates the types of ground-related problem that might be investigated. It is important to identify the zone of influence relating to each example. Each project is unique and therefore requires a specially designed ground investigation.

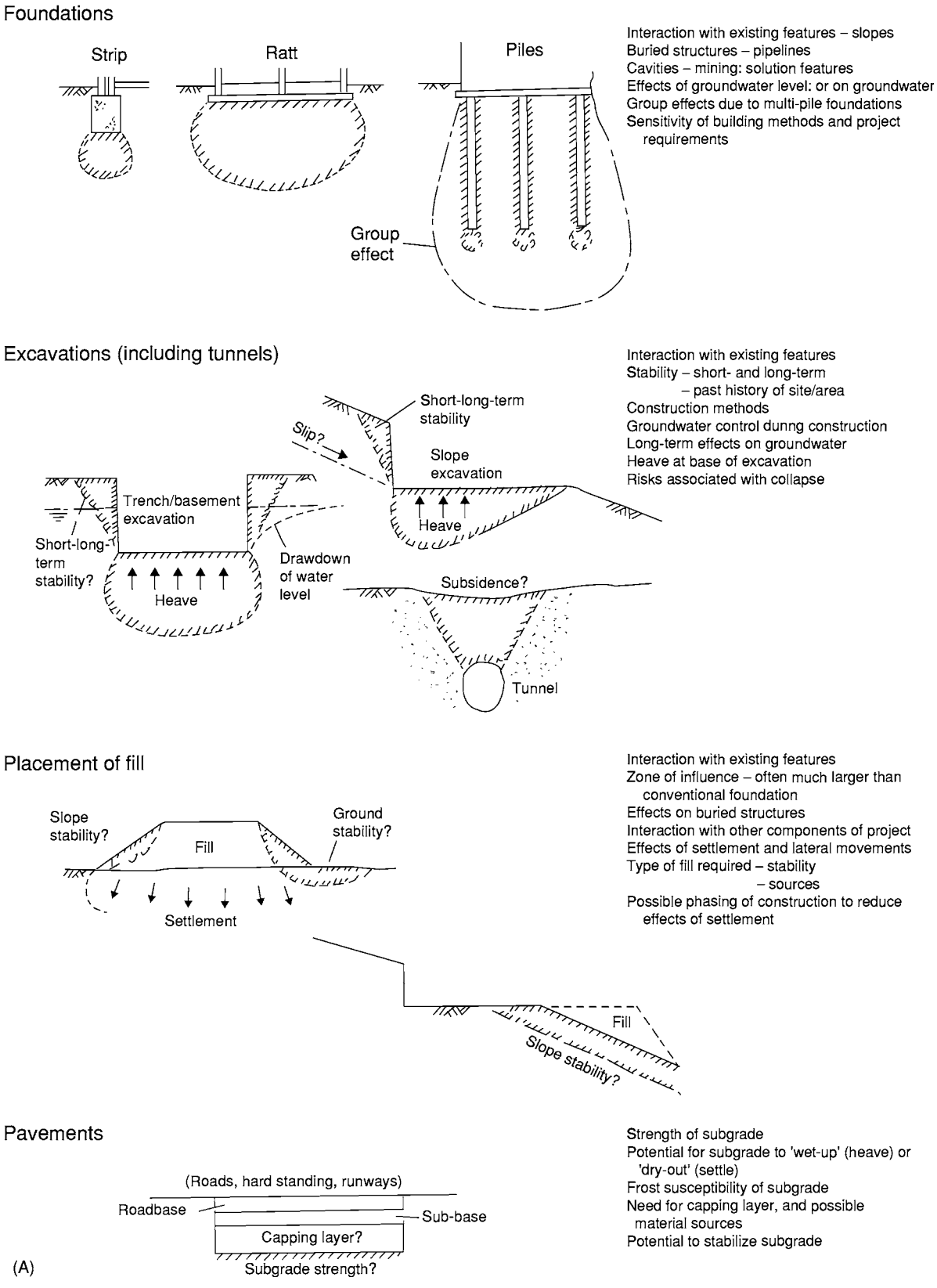
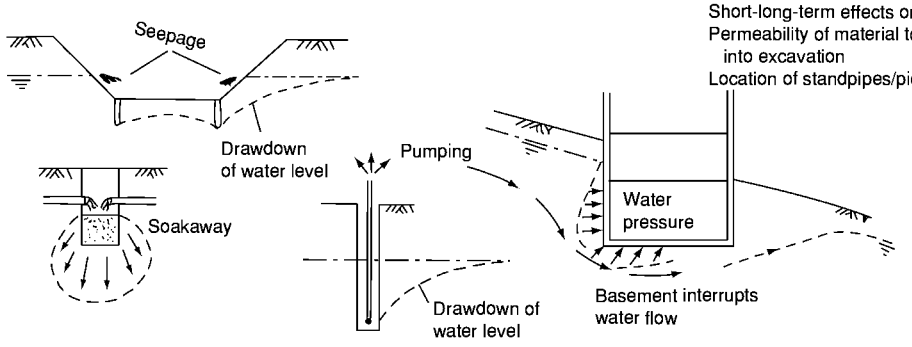


Figure 1 Construction zones to consider within site investigation. (Reproduced from Site Investigation Steering Group (1993) *Part 2 Planning, Procurement and Quality Management.* London: Thomas Telford.)

Continued

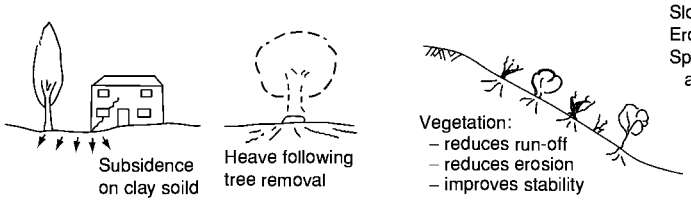
Drainage/groundwater effects

Type of drainage – soakaways, fin drains, slope drainage
Need for shor-term groundwater control
Short-long-term effects on groundwater levels
Permeability of material to be drained and/or rate of inflow into excavation
Location of standpipes/piezometers



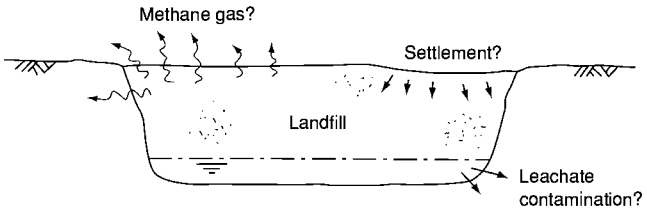
Vegetation: removal and planting

Effects of removel
Effects of planting
Slopes – depth/stability of topsoil
Erosion if topsoil omitted, e.g. chalk slopes
Special feature/requirements – liaise with landscape architect/horticulturalist



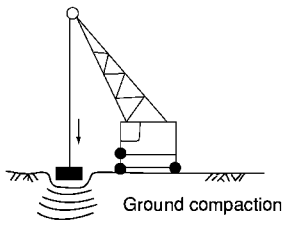
Environmental factors, e.g. landfills, contaminated land

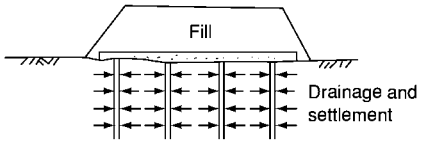
Nature and condition of existing material
Need to install monitoring equipment
Possible treatments prior to construction
Special factors e.g. isolation of vibration
Planning permission constraints
Designated areas – Sites of Special Scientific interest areas of outstanding natural beauty, protected flora/launa
Health and safety matters on contaminated sites



Ground improvement

Parameter to be improved. e.g. strength, compaction, construction speed
Methods available and information required to assess their suitability
Dewatering – effects on groundwater levels
Consult specialist contractors





(B)

Figure 1 *Continued*

The Walk-Over Survey

The walk-over survey is a detailed inspection of the site. It is often done in stages, with an initial visit for familiarization, photography, and checking of the current site conditions, and subsequent visits to check out features noted on historical maps and photographs. Features should be sketched at an appropriate scale on a base plan for inclusion in the desk-study report.

Derelict, Reused, and Contaminated Land

The desk study may reveal that a site has a history of previous use. Industrial use is likely to have left a legacy of pollution, which will require identification and possibly treatment depending on the intended reuse of the site. Guidance on the investigation of contaminated ground on the basis of 'source-pathway-receptor' is given in the key reference documents listed. The desk study will form an important element of the 'phase 1 risk assessment' that is required by contaminated-land legislation. An environmental specialist will need to be consulted whenever contamination is suspected, and appropriate safety precautions should be taken at all stages of the investigation work.

Reporting the Desk Study

On completion of the desk study, the results are formally presented in a report that brings together details of site topography, geology, geotechnical problems and parameters, groundwater conditions, existing construction and services, previous land use, expected construction risk, and proposed ground-investigation methods. A hazard plan is a good way of presenting a summary of the data accumulated during the desk study and walk-over survey.

Properties to be Determined

The prime purpose of a site investigation is generally to determine the ground profile and the continuity and properties of the various strata. This will include ranges of values and assessments of probability and risk, so that construction may proceed within a defined framework of knowledge.

Ground Investigation

Ground Investigation Design

The objective of a ground investigation is to ensure economical design and construction by reducing to an acceptable level the uncertainties and risks that the ground poses to the project.

The initial appraisal (desk study) will identify what is already known and will facilitate a preliminary understanding of the ground and its behaviour. This provides a basis for assessing the nature, location, extent, and duration of subsequent fieldwork and for preparing a programme of laboratory tests of samples obtained during the fieldwork. It is the geotechnical model against which every piece of acquired data can be checked. As the field programme progresses, so the model will be either confirmed or amended.

The design of the ground investigation can commence only once the following information has been obtained and client approval given:

1. a clearly defined purpose for the investigation;
2. an assessment of what information is required and when;
3. the areas and depths of ground to be investigated;
4. the time required for the investigation; and
5. an estimate of the cost.

These items can be conveniently covered by the Procedural Statement, as discussed previously.

Ground Investigation Methods

The ground investigation can include many different activities such as:

- trial pits with descriptions of the material exposed;
- boreholes with sampling for later laboratory testing;
- tests in boreholes using simple or sophisticated instruments;
- probing from the ground surface;
- loading tests at the surface or in excavations;
- geophysical testing; and
- geochemical sampling.

Ground-contamination investigations are often required, especially on derelict sites; appropriate expertise is essential as there are important health and safety considerations.

How Many? How Deep?

The scope and size of the ground investigation will depend both on what is known about the site and on the nature of the project. The ground investigation may, therefore, vary from a few trial pits dug by an excavator in one day, for a small housing project, to a major undertaking lasting many months, for a large earth dam. There are often alternative approaches that are equally acceptable technically, but sometimes one method is preferable to the other because of plant availability or access restraints. Boreholes may to some extent be replaced with trial pits, cone

penetrometer probes, or geophysical surveys to provide similar or complimentary information. Cost is always important, but the designer should prepare the 'right' investigation for the project rather than working to a fixed budget provided by the client.

Teamwork is important and involves pooling ideas and suggestions and taking account of restraints imposed by the client, structural designers, auditors, or environmental specialists, who may have valuable input that can be used to support the geotechnical engineer in preparing the investigation. Again, the Procedural Statement is a useful vehicle for conveying proposals to all interested parties.

Time should be allowed for innovative design and review of proposals. 'Sleeping' on ideas for a day or two can often lead to the development of better alternatives. Checklists are useful to remind the designer of aspects that should be covered.

Guidance on spacing and depth of and sampling in exploratory holes may be found in the publications of the Site Investigation Steering Group and other key texts.

There may be a temptation to believe that everything about a site should be discovered at the investigation stage. It may be more realistic to accept that certain local features (for example swallow holes in chalk, mine workings beneath a highway cutting, or precise founding levels in variable strata) are best picked up and reviewed during the construction or remediation phase. This could be more cost-effective than attempting to detect every void and strata variation during the investigation, but it does require an element of flexibility and the application of appropriate expertise during the construction work. Such flexibility is more readily available in the 'design and construct' type of construction contract and in the use of 'observation methods', which are gaining popularity with clients.

Ground Investigation Fieldwork

The various approaches that might be considered for gaining information on ground conditions are briefly reviewed in this section.

Trial Pitting

Trial pitting is a relatively cheap and efficient method of exploring the ground. Techniques vary from digging by hand – useful where services may be present or access by machine is difficult – and using mini excavators – where access is restricted – to using large-tracked back-hoe type excavators, which can reach depths of 6 m or more (Figure 2A).

Trial pits enable the stratification and the true nature of each soil horizon to be logged and representative samples selected. Block samples can be cut for subsequent laboratory testing, and localized samples can be taken from shear surfaces. However, access to the pit for close visual inspection is permitted only where temporary side support (shoring) is provided and a careful assessment of the risk of instability and the presence of gas has been made.

When inspecting the ground, a small wedge of soil is frequently sheared away from the side of the pit to reveal the true nature of the soil behind the zone of smear resulting from excavation. A moisture-content profile using close centres (typically 75 mm) can provide much information about the consistency of the soil. This is easily and accurately undertaken using the 'moisture in the bag' technique. The moisture

Figure 2 Two common methods of exploration. (A) The trial pit, with shoring and ladder for safe access and egress. (B) The percussion boring rig.

samples are sealed in a lightweight polythene bag (Figure 3). They are then weighed and dried in the bag, avoiding the need for separate moisture-content containers and reducing the error due to dehydration during sample storage.

Observation of the sides of the pit will provide an indication of the stability of the ground. The nature of the excavated material will provide information on possible reuse during earthworks. If a trial pit can be safely protected and left overnight, drying of the soil surface often reveals fissures, discontinuities, and possible existing slip surfaces within the strata.

Groundwater conditions and their effect on ground stability may be closely observed in the pit. Instability due to groundwater flow may prevent safe advancement of the pit in finer-grained granular soils. Permeability testing (by pumping or by water addition) may be carried out to give an indication of the ground permeability.

On completion of the excavation, strata should be replaced and compacted back to the depth excavated. The location (grid coordinates) of the pit must be carefully recorded as the ground is now disturbed and could present a hazard to future construction. For this reason, pits are generally excavated outside the planned area of the proposed foundations. The positions of exploratory holes are readily established by modern handheld global positioning systems.

Figure 3 'Moisture in the bag'; an efficient method of determining moisture contents.

Boring Techniques – Soft Ground

The light cable tool percussion ('shell and auger') technique of soft-ground boring at a diameter of 150 mm is perhaps the best-established, simplest, and most flexible method of boring vertical holes, particularly in the UK (Figure 2B). It generally allows data to be obtained for strata other than rock. A tubular cutter (for cohesive soil) or a shell with a flap valve (for granular soil) is repeatedly lifted and dropped using a winch and rope operated from a tripod frame. The soil that enters these tools is regularly removed and laid to one side for backfilling. Steel casing is used where necessary to prevent collapse of the sides of the borehole.

The technique can determine conditions to depths in excess of 30 m under suitable circumstances and usually causes less surface disturbance than trial pitting.

Small disturbed samples are taken on encountering each new stratum and at regular intervals within it. Larger samples of the materials are taken typically at 1 m intervals in the top 5 m of the borehole and subsequently at 1.5 m intervals. In cohesive or 'fine grained' soils, sampling is by nominal 100 mm diameter 0.45 m long open-tube drive sampling (U100), and the sample is often referred to as a (relatively) 'undisturbed' sample. These are sealed with wax to preserve moisture and capped at each end prior to transport to the laboratory for testing.

In granular soils a standard penetration test (SPT) is typically undertaken at 1 m to 1.5 m intervals, depending on depth, and a bulk sample is removed from the tested length for subsequent description and classification. The SPT involves driving a 50 mm diameter tube or cone into the ground using a 65 kg weight falling a distance of 760 mm and controlled using a trip mechanism. Blows are counted for successive penetrations of 75 mm for a distance of 450 mm. The SPT value is the number of blows required to penetrate the final 300 mm. This test is sometimes also used in firm clays. The results of the SPT are correlated empirically with the density and angle of shear resistance in granular soils and with shear strength in cohesive soils.

Alternative techniques that might be considered for advancing boreholes include the less reliable wash-boring method, where the soil is washed out of the cased hole, and flight auger boring, where disturbed samples are brought to the surface by the rotation of the auger and, where hole stability permits, the auger is removed for undisturbed sampling.

The hollow-stem flight auger is a more reliable system, which supports the hole whilst undisturbed samples are recovered through the central stem.

For shallow holes where access is difficult, hand augering may be employed. The use of a lightweight mechanical hammer associated with dynamic probing has led to the recent development of the 'window' sampling technique. This involves driving a tube with a section of the side wall removed to allow the soil sequence to be logged and samples to be selected from the side for classification and, increasingly, for contamination testing. Sample-tube lengths are typically 1 m or 1.5 m. Deeper sampling is possible by progressively reducing the tube diameter to reduce side friction from the strata above the sampling level.

Boring Techniques – Hard Ground

Rotary drilling produces rock cores by rotating an annular diamond-impregnated tube or barrel into the ground. This is the technique that is most appropriate for making investigation holes in rock or other hard strata. It may be used vertically or at any angle.

Core diameters of less than 100 mm are most common for site-investigation purposes. A flushing fluid such as air, water, mist, or foam is used to cool the bit and carry cuttings to the surface.

Examination of rock cores allows detailed description and generally enables angled discontinuity surfaces to be observed. However, it does not necessarily reveal the presence of near-vertical fissures or joint discontinuities. The core can be tested in the field or in the laboratory. Core recovery depends on rock type and the techniques employed.

Where open-hole rotary drilling is employed, descriptions of the strata depend on the examination at the surface of small particles ejected from the borehole in the flushing medium. Consequently, no indication of fissuring, bedding, consistency, or degree of weathering can be obtained. Depths in excess of 60 m can be investigated using rotary techniques with minimal surface disturbance.

Dynamic and Static Probing

Dynamic probing involves driving a rod with a fixed or detachable cone into the ground by allowing a falling weight to impact on an anvil attached at the top of the rod and counting the number of blows required to achieve successive penetrations, typically of 100 mm. The apparatus can be handheld (for example the Mackintosh probe), or a machine-mounted weight and guide can be used for larger-diameter rods (for example the Pennine probe) (Figure 4).

(A) Lightweight drilling, dynamic probe rig (Courtesy Exploration Associates)

(B) Lightweight dynamic (Mackintosh) hand probing for indication of soil density

Figure 4 Examples of dynamic probing (A) by lightweight drilling and (B) by hand. (A) A dynamic probe rig (photograph courtesy of Soil Mechanics). (B) Lightweight dynamic (Mackintosh) hand probing to obtain an indication of soil density.

Whilst the number of blows may be correlated empirically with the shear strength or density of granular soils, these are generally profiling tools best used for interpolating between boreholes or trial pits. Samples are not normally taken, but the technique has been extended by using the hammer apparatus to take 'window' samples, as described above.

The static cone technique (Figure 5) involves pushing an instrumented cone and sleeve into the ground at a defined rate and recording the point resistance and skin friction using electronic means. More sophisticated instruments can now monitor the pressure of pore water just behind the cone tip. By considering the ratio of end resistance to shaft friction resistance, an indication of the soil type and its density may be deduced.

The apparatus can measure changes in soil type over very short distances and avoids the problems of disturbance, particularly of granular soils, below the water table. It is a very rapid technique, often used as a profiling tool between boreholes, and is capable of electronic interpretation. It has been used extensively in North Sea oil and gas structure foundations and is now used widely for land-based investigations.

Figure 5 The static cone apparatus. (A) 'Dutch' cone penetrometer testing truck with (B) control console and (C) electric static cone (photographs courtesy of Fugro Ltd).

As with all probing techniques, the presence of large cobbles and boulders can prevent penetration in certain soils and in made ground. Although linked with boring techniques, the static and dynamic cones are regarded as a form of *in situ* test.

In Situ Testing

The main objectives of a ground investigation are generally to identify and classify the soil types into groups of materials that exhibit broadly similar engineering behaviour and to determine the parameters that are required for engineering design calculations.

Some soils, such as certain clays, may be readily sampled and transported to the laboratory for quality testing under controlled conditions. Other soils, such as very soft or sensitive clays, stony soils, sands, and gravels, and weak, fissile, or fractured rock are not easily sampled in an intact 'undisturbed' state, and therefore *in situ* testing is required to obtain the necessary engineering parameters. *In situ* tests may take the form of geophysical tests (as described in the following section), *in situ* soil-testing techniques, or links with field instrumentation.

In situ tests fall into three typical groups.

1. Empirical tests, where no fundamental analysis is possible and stress paths, drainage conditions, and rate of loading are either uncontrolled or inappropriate (example is SPT).
2. Semi-empirical tests, where a limited relationship between parameters and measurements may be developed (examples are cone penetrometer tests, California Bearing Ratio (CBR) and borehole permeability tests).
3. Analytical tests, where stress paths are controlled (but strain levels and drainage often are not) (examples are vane shear strength, pressuremeter tests, pumping tests and packer permeability).

Details of the tests and their application are given in British Standards 1377 and 5930 and other references listed in the further reading section. The main point to note with *in situ* testing is that the drainage conditions are almost impossible to control, and, therefore, there is uncertainty as to the drained/undrained nature of the test. However, the very act of testing *in situ* provides a good indication of the actual ground response provided that scale factors are taken into account. For example, a plate-load test having a plate width of B will affect the ground to a depth of approximately $1.5B$. The actual foundation may be many times the width of the test plate, and the depth of influence will therefore be much deeper.

The SPT is a simple cost-effective means of assessing granular soils. More analytically correct tools such as the pressuremeter are expensive and their results are not always straightforward to interpret. Judgement must be used to determine the level of sophistication required for a particular project.

Geophysical Investigation

Measurement of the geophysical properties of the ground may provide an indication of the location of strata boundaries and anomalous ground conditions. Such surveys may include measurements of ground conductivity, magnetic and gravity fields, electrical resistivity, and seismic responses (*see* **Analytical Methods:** Gravity, **Seismic Surveys**).

The use of geophysical techniques is generally non-intrusive, with little site disturbance, and is therefore sometimes carried out at the desk-study stage. However, the main benefit comes when some exploratory hole data are available to permit more accurate calibration, correlation, and processing of the geophysical data. The ability of the current generation of computers to process large amounts of data rapidly has led to a resurgence of interest in the potential value of geophysical investigation techniques.

Groundwater and Instrumentation

The presence of groundwater leads to many engineering problems, and it is vital that all water observations are carefully recorded. Whenever water is struck in a borehole or trial pit the point of entry must be noted. It is normal practice to cease the advancement of the hole for 20 min whilst any rise in the water level is observed. Longer-term observations of groundwater fluctuations may be made by installing a standpipe or standpipe piezometer. This typically comprises a porous filter at the base of a 19 mm diameter plastic tube inserted to an appropriate depth in a borehole (**Figure 6**). Water enters through the porous filter, and its level in the plastic tube can be monitored over a period of time using a dip meter that emits an audible signal when an electrical circuit is completed as it meets the water in the borehole.

Water-level records taken over an appropriate period of time will reveal the likely range of water conditions to be allowed for in the design and to be encountered during the project construction.

More sophisticated equipment such as hydraulic, electrical, or pneumatic piezometers with transducer systems and automatic recording can be used

where the information obtained is critical to the project.

Other instrumentation, such as settlement gauges, plates, and extensometers to measure vertical movement, and slip indicators and inclinometers to measure lateral movement at depth, are available for use in conjunction with the ground investigation or with advance trial constructions and main works constructed by 'observational' methods.

Laboratory Testing

It should be remembered that soil is an assemblage of solid particles, which may contain organic matter. The voids between the particles are filled with gas – usually air – and water. The geotechnical engineer will need to determine the composition of the ground *in situ* and also explore the changes that will occur to the soil as a result of the proposed construction.

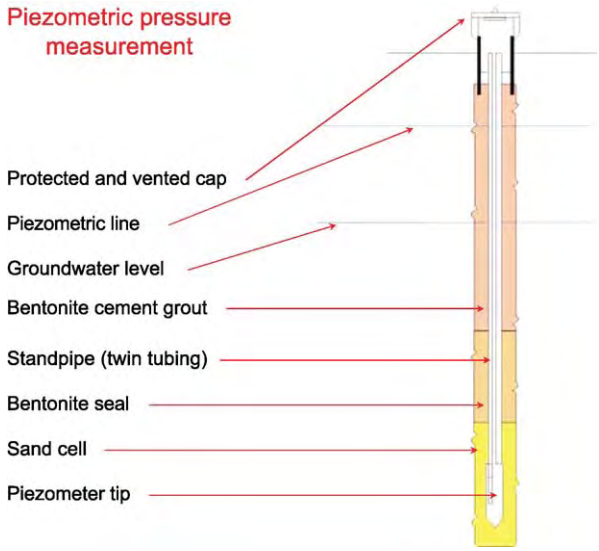


Figure 6 Schematic of standpipe piezometer (courtesy of Geotechnical Instruments UK Ltd).

Laboratory tests tend to be divided into two main classes.

Classification tests involve grouping soil types into categories of soils that possess similar properties. For coarse-grained soils the particle size distribution is often the key property, whereas for fine-grained soils the moisture content and mineralogy, as reflected in the Atterberg limits (index properties), most often hold the key to grouping soils of similar characteristics. For general classification purposes, the moisture content and index properties (liquid limit, plastic limit, and sometimes shrinkage limit) are used together with the particle size distribution, density of the soil, and, most importantly, an accurate description carried out in accordance with the standards in the code of practice.

Engineering tests assess the engineering properties of the soil, such as routine shear strength, compressibility (see Figure 7), compaction characteristics, and permeability. Where more specific soil properties are required for complex design functions, 'advanced' laboratory tests may be used, in which high-quality undisturbed samples are carefully prepared and tested under precise laboratory control and stress and small strains are measured.

Full details of laboratory tests and their relevance to engineering design are given in the appropriate British Standards and the key references in the further reading section.

It is important that the testing regime is planned in advance of the fieldwork in order that suitable samples are recovered from the ground. The Procedural Statement will help with this planning process, as the reasons for the sampling and subsequent

Figure 7 Commercial laboratory testing for shear strength and consolidation. (A) A 100 mm triaxial test cell. (B) A bank of consolidation (oedometer) test machines. (Photographs courtesy of CSL Ltd.)

laboratory testing can be logically considered together with the significance of the anticipated results.

Reporting

The reporting of site and ground investigation work is most important. The factual report is generally the only tangible output from the investigation and represents a considerable financial investment. The factual report should include:

- a statement as to the purpose and rationale of the investigation;
- a description of the work carried out, including references to the specification and standards adopted and any deviations from them;
- exploratory hole logs (including location and ground level);
- *in situ* test records;
- laboratory test results; and
- a site location plan.

The development by the Association of Geotechnical and Environmental Specialists of a standard format for geotechnical data files in digital form has led to the efficient collection, transfer, processing, presentation, and storage of the data for all parties involved in the investigation process. Data are ideally input manually to the computer only once or are collected directly in electronic form from monitoring and testing devices. Specialist software houses have developed programs to help the investigator collect, store, and process the data and to present it in the form most suited to the requirements of a particular project (Figure 8).

The exploratory hole log can perhaps be regarded as the heart of the factual report because it records all the information obtained at that particular location. An example log prepared from digital data is presented in Figure 9.

Interpretation of the geotechnical data to relate them to the proposed design and construction will be required for each project and is normally the responsibility of the Geotechnical Advisor. Formal interpretive reports (also known as assessment reports or appraisal reports) may be prepared as required by the design team. It is good practice to prepare a

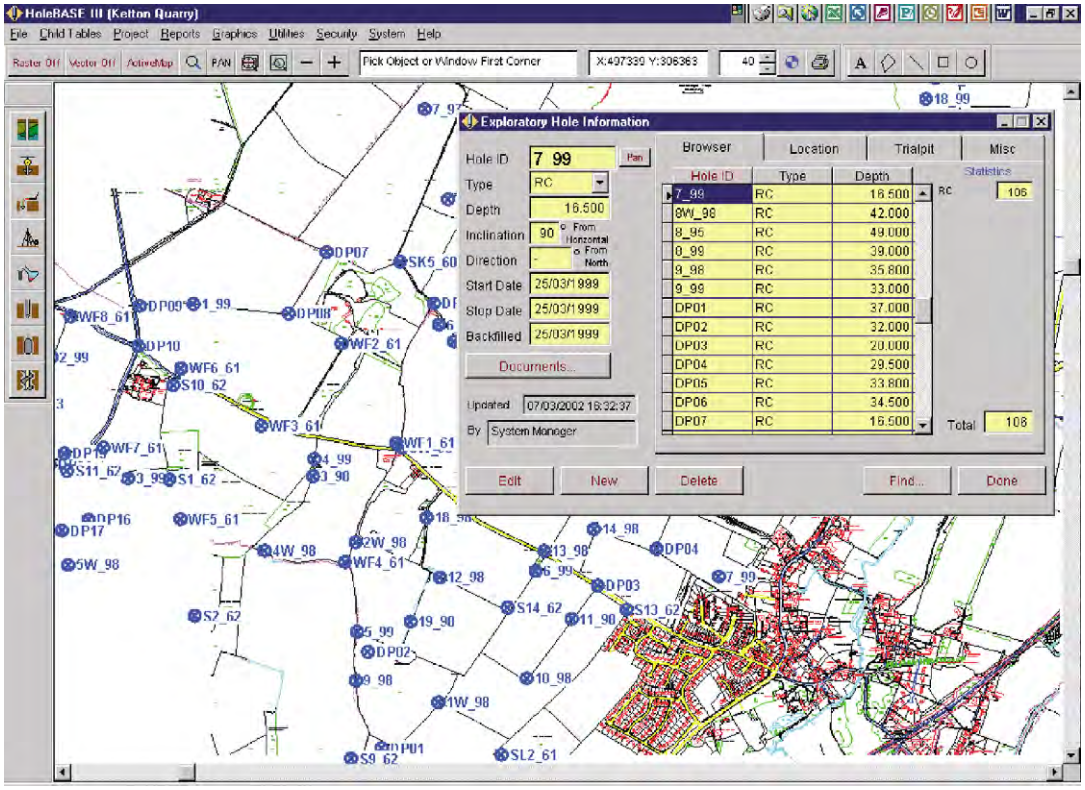


Figure 8 Use of digital techniques for efficient handling of site investigation data (image courtesy of Key Systems and Geotechnical Developments).

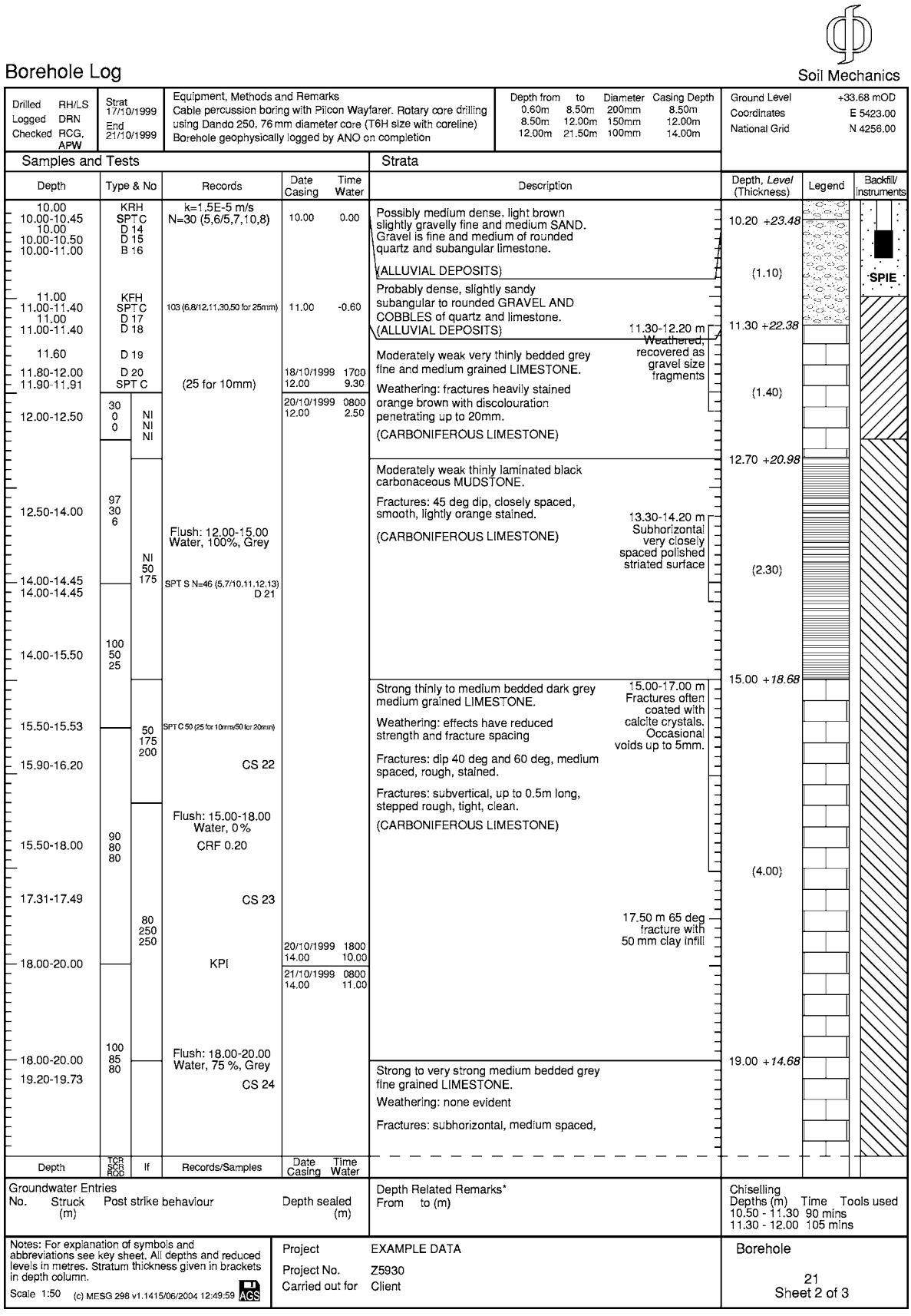


Figure 9 Example exploratory hole log (Courtesy Soil Mechanics).

geotechnical design brief to assist the construction team on site and, at the end of the project, to prepare a feedback (maintenance) report on any issues from which lessons can be learned and on which future scheme development can be based.

The various stages of quality control and reporting that help to ensure maintenance of the standards and accuracy of the output from a site and ground investigation are described in the publications of the Site Investigation Steering Group.

Concluding Remarks

Site investigation involves establishing facts concerning the ground; the excitement of discovery is tempered by the professional responsibility for much of the financial risk and safety associated with each engineering project.

The combination of appropriately experienced investigators working to agreed team procedures within a sensible permitted time frame with limited competition is likely to produce the best solutions to engineering challenges presented by particular ground conditions. This article has touched on some of the important issues relevant to site investigation but is in no way intended to be a complete guide. A good investigator will set out to define the investigation but to expect the unexpected.

See Also

Analytical Methods: Gravity. **Engineering Geology:** Codes of Practice; Site Classification; Ground Water Monitoring at Solid Waste Landfills. **Geotechnical Engineering**. **Seismic Surveys**. **Soil Mechanics**.

Further Reading

BS 5930 (1999) *Code of Practice for Site Investigation*. London: BSI.

BS 1377 (1990) *British Standard Methods of Test for Soils for Civil Engineering Parts 1–9*. London: BSI.

BS 10 (2000) *Contaminated Land*. London: BSI.

BS 10175 (2001) Investigation of potentially contaminated sites. London: BSI.

Clayton CRI, Matthews MC, and Simons NE (1995) *Site Investigation: A Handbook for Engineers*, 2nd edn. Oxford: Blackwell Scientific.

HD22/02 Managing geotechnical risk. DMRB. London: Highways Agency.

Institution of Civil Engineers (2003) *Conditions of Contract Ground Investigation*, 2nd edn. London: Thomas Telford.

McCann DM, Eddlestone M, Fenning PJ, and Reeves GM (eds.) (1997) *Modern Geophysics in Engineering*. Special Publication 12. London: Geological Society.

Perry J (1996) *Sources of Information for Site Investigations in Britain*. TRL Report 192. Crowthorne: Transport Research Laboratory.

Simons N, Menzies B, and Matthews M (2002) *A Short Course in Geotechnical Site Investigation*. London: Thomas Telford.

Site Investigation Steering Group (1993) *Part 1: Without Site Investigation Ground is Hazard*. London: Thomas Telford.

Site Investigation Steering Group (1993) *Part 2 Planning, Procurement and Quality Management*. London: Thomas Telford.

Site Investigation Steering Group (1993) *Part 3 Specification for Ground Investigation*. London: Thomas Telford.

Site Investigation Steering Group (1993) *Part 4 Guidelines for Safe Investigation by Drilling of Landfills and Contaminated Land*. London: Thomas Telford.

ENCYCLOPEDIA OF
GEOLOGY

ENCYCLOPEDIA OF GEOLOGY

EDITED BY

RICHARD C. SELLEY
L. ROBIN M. COCKS
IAN R. PLIMER

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Site Classification

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

Site characterization may be defined as the

> . . .three dimensional engineering geologic description of the surface and subsurface of the location for intended construction of engineered works, for habitation, com merce, resource development, mitigation of natural hazards or conduct of groundwater protection, waste management or environmental remediation.

Site characterization provides the design information necessary for project planning. Site characterization is frequently broken and discontinuous in nature, reflecting the naturally flawed ground that is being evaluated (Figure 1). The broad elements of site characterization are listed in Table 1. No such table can be wholly inclusive, but it serves to establish a format for discovering and predicting the elements that should be included within the scope of construction work.

The broad goals of site characterization include:

- identification of sites likely to experience ground displacement from fault movement;
- identification of the stability of a site's configuration;
- establishment of the need to improve the engineering characteristics of the ground so as to resist damaging deformation;
- establishment of the need to control groundwater to prevent excess pore pressure and subsequent failure under the involved structural loads; and
- identification of the potential for creating green space not only for social gain but also to incorporate active faults and existing geohazards.

Purpose and Scope of Site Characterization

Site characterization is a major component of site investigation, the purpose of which is to determine ground conditions at a proposed construction site. The current terminology signifies a broadening of the purpose and scope to encompass the determination of relevant design needs for a wide variety of construction, resource development, hazard mitigation, and environmental goals.

Site investigation matured from the end of World War II through the 1970s, and then expanded in the 1980s, providing broad scope for the work of engineering geologists.

There is some published guidance on site characterization, but it is the responsibility of the engineering geologist in charge of each project to define the scope of the characterization in order to meet the design needs of the planner and engineer. These needs fall into seven basic categories:

- a three-dimensional description of the project site to the depth of influence (i.e. where the maximum applied (live and dead) static load from the intended use of the site has dropped to 10%);
- a description of the length and breadth of the proposed site and of the area beyond the property boundaries to such a distance as will encompass any anticipated effect of the activities contemplated at the site (this distance will be sufficient to cover the origins of geological constraints that may impinge on the contemplated site, including slope instability, groundwater inundation, and loss of ground support);
- the identification of ground that may be affected by displacement, volumetric change (shrinkage or swelling), or other forms of loss of physical support or chemical integrity relating to the intended construction;
- the discovery or prediction of ground instability above, below, on, or adjacent to the site;
- the detection and measurement of groundwater conditions (including perched and transient conditions) that may affect the construction, operation, or maintenance of the proposed works (including the estimation of annual variations and chemical composition, and in particular the potential to harm human health or the environment);
- the discovery or prediction of conditions that may adversely affect the rate, progress, or method of construction of the project; and
- the detection or prediction of ground conditions that may be unsafe or endanger the health of persons engaged in construction, or in work, visitation, or occupancy of the developed site.

Typical Site Characterization

The owner has the primary and ultimate responsibility for ordering and funding a site characterization and for ensuring that any constraints identified by the

Figure 1 Engineering geologists generally work with already flawed earth materials. This weak rock of Pennsylvanian age at Little Rock, Arkansas, needs to be ripped on excavation, yet it is wholly discontinuous and platy.

site characterization are addressed in the design, construction, operation, and maintenance of the engineered works. Table 2 assesses the typical goals and components of the site characterization of a variety of engineered works.

Areas Subjected to Site Characterization

In reality, no two parties engaged in planning and conducting a site characterization will produce identical pieces of work. The quality and applicability of every site characterization reflects the experience and competence of its leadership. Strictly speaking, the basic (minimal) characterization must assess the site within the property boundaries. There are many circumstances, however, in which additional areas of ground should be considered in order fully to address the client's needs. Table 3 presents some of these considerations.

Pitfalls of Site Characterization

There are no hard and fast rules or standards that can be applied to all site characterizations. It is incumbent on the engineering geologist to inform the owner of those elements of the work that may best meet the needs of the design engineers and the owner's goals for the project.

If the owner chooses to underfund the site characterization or selects characterization activities on a price-sensitive basis, there are numerous ways in which the end product may fail to give adequate design-related information. Where price has not been a determining factor in the scope of the site

Table 1 General elements of site characterization

Element	*Purpose*	*Important considerations*
Stratigraphy	Identify and describe geological formational units expected in design and construction	Individual engineering geological units
Groundwater regime	Define character of groundwater	Perched water Vadose zone Peizometric surface Potentiometric surface (if present) 12 months of observation
Top of rock	Define elevation or surface below which excavation is difficult	Little's grades of weathering will be helpful here
Rock mass characterization	Delimit observable or likely subsurface bounds of each detectable hard rock unit	Contacts Lithological character and variations Geological structure represented by equal area stereographic projection; to include bedding or foliation attitude, discontinuities, faults, fault zones, shear planes, metamorphic foliation Zones of bad ground, by virtue of alteration, weathering, or shearing Application of one of the standard rock mass qualification concepts
Presence of weak rock	Basis of definition, including why the rock is determined to be weak	Recommendations as to how and why such weak rock may pose problems in design and/or construction, operation, and maintenance
Potential problems related to sedimentological, structural, or geomorphological conditions	Portions of surface or subsurface that may be affected by otherwise unanticipated features	Buried valleys Stratigraphical 'holidays' Facies changes Fault displaced blocks of ground Dissolution cavities (karst features) Glacial erratics (boulders) Irregular features of glacial and periglacial origin
Potential instability	Hillsides River banks Ground expanse of site	Unstable slope masses Subsurface mined or dissolution caverns
Likely excavation characteristics	Relative ease or difficulty of moving earth media	Outstanding troublesome departures include glacial, bedrock knobs at top of rock, and some types of hard igneous intrusions such as dykes and sills
Geological constraints	Observed or postulated geological conditions that may be known or felt to place the project concept in jeopardy; may require additional funds for resolution	Otherwise known as 'geological hazards' but more properly known as 'geological constraints' after the life work of Robert F Legget

characterization, there are still categories of characterization flaws that, if not heeded, could endanger the achievement of design goals. Some of these are presented in Table 4.

Conclusion

The majority of the project technical responsibilities accorded to the engineering geologist fall under the remit of site characterization relating to construction and waste disposal. The aim is to ensure that properly funded exploration and evaluation can establish the geological information required by the design engineers. This information is located by spatial coordinates and is usually managed and displayed by computer-assisted drawing and geographical information systems. The geologist thereby develops a database of usable, relevant information in the form of colour-coded perspective drawings that are easily assimilated by an array of (non-geoscience) professionals who can subsequently manipulate the data for parametric analysis or for their own better understanding, and can do so largely without fear of distorting the relevant facts.

Table 2 Elements of site characterization

Construction works	*Design citeria*	*Characterization*
Construction		
Foundations	Performance of foundation Stability of adjacent buildings	Optimal use of foundation geology to install minimum cost foundation
Stability of excavations	Basements and parking spaces Lateral cuts for building space Recognition of groundwater Stability of adjacent structures	Basements are an economic imperative in cities owing to the cost of land
Drainage and dewatering	Keep excavation free of water Floodworks integrity	Construction control of water versus long term maintenance of groundwater and earthworks pore water
Materials of construction	Use on the project Export to urban areas	Dealing with public opposition Preservation of water quality
Land takings	Provide value to owner	Possible mineral resource value
Transportation		
Highways and rapid transit	Stable roadway Borrow for above ground works Balanced cut and fill	Sufficient to move traffic from A to B with minimal impact on existing infrastructure
Airfields	Runway space Fuel storage Runoff and waste management	Huge land surfaces subject to incidental wastes from operation Anti terrorist considerations
Tunnels and underground caverns	Optimal excavation method Promote ground stability Ground control Management of groundwater	Consider spoil handling and disposal, and stability of overlying build environment
Ports and harbours	Wave protection Operational wastes	Associated dredging may disturb historic contaminated sediment
Navigable rivers	Floodworks and flood control Locks and dammed pools	Associated dredging may disturb historic contaminated sediment
Energy and resource		
Dams	Foundation integrity Seismic withstand behaviour Slope stability Emergency spillways	Locate an inherently stable dam axis position Provide nearby construction materials
Pipelines	Route versus geological costs Erosion protection Seismic stability Stability of adjacent ground	Objective of minimal excavation, siphonage, and above ground support Anti terrorist considerations
Power transmission	Mainly tower stability	Minimize grubbing impacts
Forest resources	Slope stability Haul road location Potential hydraulic damage	Harvest the resource without diminishing the condition of the remaining land
Mineral resources	Minimize exploration damage Opencast Underground workings	Minimize exploration damage Manage acid mine drainage Spoil management Protection of water and soil values
Mineral beneficiation	Concentrate ore value Wastewater effluent	Manage soil impacts
Reservoirs	Reduce bottom leakage Sedimentation protection Shoreline protection Water quality	Requires a considered evaluation of geomorphology and the primary and secondary porosities of geological units in contact with stored water
Groundwater recharge	Use of storm waters Use of treated wastewater	Introduce flood and grey waters to the groundwater regime before their loss to oceans
Environmental		
Waste disposal facility siting	Minimize adverse impacts Optimize geological conditions Locate sufficient daily cover	Control of leachate Protection of water quality
Water quality protection	Suspended solids Dissolved solids Discharge into governed water	Cation exchange leads to degraded water quality (pollution), which can lead to health effects (contamination)

Table 2 Continued

Construction works	*Design citeria*	*Characterization*
Military impacts	Manoeuvre and exercise grounds	Vehicular disturbance of ground leads to sedimentation
	Munitions impact zones	
	Battle damage	Spilled fuels and discharged munitions lead to environmental degradation
Derelict land	Waste dumps	Ground receives contamination and surface waters and groundwaters transport this to human and natural life receptors
	Industrial brownfields	
	Abandoned mined lands	
	Battle damage	
Environmental remediation	Uncontrolled hazardous waste	Characterization of site and waste is essential to formulation of remedial actions
		Without this actual risks become unacceptable
Saline intrusion	Control freshwater withdrawal	A mature art requiring only the application of sufficient funding
	Subsurface barriers	
Sedimentation	Soil particles	Results from disturbed ground
	Cation exchange	Diminishes reservoir capacity when inflow ceases
	River and stream transport	
		Damages wildlife habitat by turbidity
Hazards		
Flood hazards	Flood channels	Depends on geomorphological evidence of past flooding, integrated with runoff character of the related watershed
	No build zones in floodplains	
Rock falls	Gravitational dislodgment of rock and rock masses above inhabited or frequented areas	Delimit rock or masses of rock above endangered populations or transportation routes
Retain or devise protection from dislodgment or trajectory impact		
Snow avalanches	Recurrence intervals, and conditions and timing of recurrence	A sideline use of geological expertise in geological image interpretation
Seasonal advice on mitigation by explosive intervention or shed deflectors over railroad and highway routes		
Active faults	Ground trace	Dealing with vertical splaying of related fault splays in unconsolidated units
	Sense of displacement	
	Recurrence interval	
	Likely nature of strong motion	Blind thrust faults as unknowns
Unstable ground	Collapsing soils (water inflow)	Soil properties and character make them susceptible to damage
	Expansive soils (water inflow)	
	Liquefaction (earthquake strong motion in the presence of near surface groundwater)	
		Result damages foundations
Shore protection	Wave energy impact analysis	Evaluation requires geomorphological assessment in terms of the structural character of the shore rock
	Susceptibility of geological units to erosion and toppling	
		Wave mitigation usually only transfers wave energy to new targets
Volcanic	Lava flows	Termination or preclusion not yet possible
	Ash falls	Geological characterization can delimit sources, routes, and probable areas of impact
	Pyroclastic bombs	
	Glowing avalanches	
	Mudflows (lahars)	Mitigation involves avoidance or channelization

Continued

Table 2 Continued

Construction works	*Design citeria*	*Characterization*
Defence		
Military posts and installations	Characterization of ground for optimal use in training and staging, while striving for optimal environmental protection of military bases	World economic situation currently favours diminishing grounds allotted to military forces, while increasing environmental controls over lands used for military training
Use of terrain	Defence	Trafficability
	Offense	Obstacles and fields of fire
		River crossings
Potable water	Exploration and development	Alternative to use of water in the field of operations or to importation in competition with high priority food and munitions
	Replacement of damaged wells and reservoirs	
		Supplies for indigenous peoples and prisoners of war
Anti terrorism actions	Defence from attack	Use or exclusion of use of such features as caves, excavated tunnels, and unstable hillsides
	Actions against terrorists	

In all cases, the characterization should reflect the owner's objectives, available funding, the normal and standing considerations for public health, welfare and safety, and applicable and relevant regulations and codes, and should be conducted in compliance with or with reflection on relevant codes and standards, and be completed by or under the direct supervision of a registered (chartered) professional geologist. All specialization categories should consider the presence of geological constraints and natural hazards that may affect the design, construction, operation, and maintenance of the project in a manner that meets the client's objectives and applicable national, regional, and local laws and regulations.

Table 3 Key considerations of site characterization

Construction works	*Key considerations*	*Other considerations*
Construction		
Foundations	To property boundary	Adjacent ground that may be affected by the project or have conditions that may affect the project
Stability of excavations	Host ground to include likely active mass of wall or face subject to possible failure	Possible adversely orientated planes or wedges of geological weakness
Drainage and dewatering	Maximum radius of cones of depression	Search for possible hidden (geomorphological) pathways of high hydraulic transmissibility
Materials of construction	Secure projected quantities of acceptable material	Locate superior materials at somewhat greater distance
Land takings	Natural resource value	Impact on groundwater resources
Transportation		
Highways and rapid transit	Right of way plus additional width affecting existing infrastructure	Minimum of one boring or geophysical traverse for each suspected adverse geological possibility
		Minimize groundwater impact
Airfields	Footprints of load bearing areas	Toxic waste hotspots in areas of runway expansion
Tunnels and underground caverns	Route characterization for estimated ground conditions	Geormorphologic and structural hazards such as intercepted bad ground or groundwater flow pathways
	Excavation	
	Ground support	
	Groundwater management	
	Spoil management	
Ports and harbours	Toxic dredge spoils	Locations of toxic hotspots
Navigable rivers	Lock and dam foundations	Stability of caving river banks
Energy and resource		
Dams	Stability of embankment	Consider adjacent ground hazards
Reservoirs	Minimize bottom leakage	Minimize incoming sediment
Groundwater recharge	Maximize storativity	Use to protect groundwater regime
Pipelines	Assess geological excavation cost versus alternative routes	Minimize environmental impact on the traversed ground

Table 3 Continued

Construction works	*Key considerations*	*Other considerations*
Power transmission	Cost effective tower locations	Minimize impacts from unstable ground on or adjacent to route
Forest resources	Minimize hydraulic damage	Minimize sediment produced
Mineral resources	Minimize exploration and development impacts	Devise safe spoil disposal strategies and minimal cation exchange leaching
Mineral beneficiation	Proper management of spoil	Plans for plant closure
Environmental		
Waste disposal facility siting	Isolate from groundwater	Provide adequate cover sources
Water quality protection	Remove toxic hotspots	Control entry of freshwater
Military impacts	Remove explosive relics	Grade land for minimal impact
Derelict land	Remove toxic hotspots	Control entry of freshwater
Environmental remediation	Remove toxic hotspots	Control entry of freshwater
Saline intrusion	Control groundwater pumping	Inject flood or treated waters
Sedimentation	Reduce land disturbance	Define interception features
Hazards		
Flood hazards	Establish likely inundation	Evidence of catastrophic flooding
Rock falls	Unstable rock masses	Impact zones on adjacent ground
Snow avalanches	Photogeological evidence of pathways	Foundations and anchorage of structures to deflect snow
Active faults	Establish likely rupture location	Susceptibility to lateral spread or liquefaction
Unstable ground	Mainly related to regional physiography	Avoidance first; mitigation second
Shore protection	Delimit susceptible ground	Other shoreline influences, such as groundwater seepage
Volcanic	Pathway of impact	Methods to deflect impact
Defence		
Military posts and installations	Footprints of structures	Areas of intense use by vehicles
Use of terrain	Obstacles Trafficability	Interdiction of enemy movement Neutralization of underground space
Potable water	Exploration for wells	Water supply basin
Anti terrorism actions	Features for approaches	Unstable surrounding ground Wellfield susceptibility to sabotage

In all cases, the characterization should reflect the owner's objectives, the level of available funding for exploration and characterization, and the normal geotechnical, hydrological, and environmental practices employed in the region.

Table 4 Some pitfalls of site characterization

Potential pitfall	*Nature*
Failure to consider owner's goals and design conditions	Represent the stated truths of project impact on the geological environment. Owner should be informed of the consequences of these considerations, though it is the owner's responsibility to act on such information
Failure to separate factual from interpreted geological information	Discrimination should be made in order to represent the owner's rights and position, in which the construction contractor bears the normal risks associated with construction under the conditions reported in the design documents and construction specifications. Contractors are entitled to additional payments for geological conditions not adequately defined in the contract documents. Separation of factual from interpreted information more clearly defines the information available for consideration by the contractor in bid formulation
Failure to apply site conceptual geological model	Essentially a schematic drawing (cartoon) of known and possible geological conditions at the site, superimposed on one or two (perpendicular) sketch geotechnical profiles, from ground surface to maximum depth of expected impact on site design
'Geological possibilities'	It is proper to consider the possibility of certain sedimentological, stratigraphical, structural, and geomorphological anomalies at the site, resulting from its general regional physiographical affinity. These possibilities should be assessed in terms of the site development plan

Continued

Table 4 Continued

Potential pitfall	*Nature*
Regionally important geological material	Failure to recognize that the subject earth material in the physiographical province has special or unique characteristics that affect or its geotechnical behaviour
Geomorphological and structural geological features	Failure to consider the potential for anomalous subsurface zones, pockets, or buried valleys of materials that very significantly from those of the general host ground
Neglect to consider discontinuities	Failure to recognize that discontinuities represent preformed surfaces of weakness and therefore have the potential for displacement and sometimes groundwater or contaminant migration
Assessment of 'zones'	'Zones' should be invoked whenever evidence suggests the existence of pockets, bands, or other bounded bodies of earth material that could affect the design or performance of the project. 'Zones' are labelled and delimited by the characterizing engineering geologist
Deleterious minerals	Ground that has been geochemically altered should be expected to potentially produce *in situ* conditions or construction material that may not perform adequately in terms of project design. Often this occurs along the borders of intrusive bodies
Misassessment of karst	Karst terrain typically shows a concentration of dissolution along major joints and at their intersections. Much of the intervening ground may be devoid of dissolution features of sufficient magnitude to affect the project. Use of geophysical techniques is recommended.
Minimal time sensitive hydrogeological data	Only minimally accurate assessments of the character and seasonal behaviour of the groundwater can be made with a full 12 month record of hydrogeological conditions at and around the site
Inapplicable hydrological borehole testing	A prime example is the conduct of downhole packed hydraulic conductivity testing in a string of end to end positions, in which the conductivity is then averaged over the packed interval, and sometimes over the entire tested interval, whereas the true impact on project design and performance is more properly assigned to individual layers or zones, which themselves should be subjected to packed testing. On occasion this pitfall is used by litigation opposition in order to impart a condition more favourable to the client's interests
Belief in the 'pristine' brownfield (derelict) site	All brownfield sites should be considered contaminated until shown otherwise; brownfield sites cannot be adequately characterized for their toxic wastes without knowledge of site technical and operational history and an understanding of the waste generation nature of the industrial flow path most likely to have been present during operation
Notion of 'non aqueous' phase liquids	Most are toxic to people and the environment. 'Non aqueous' is a chemist's term; these compounds have the potential to dissolve in surface water and groundwater in concentrations inimical to health and the environment
Significance & VOCs and SVOCs	Volatile organic compounds (VOCs) migrate generally and in the hydrogeological system; semi volatile organic compounds (SVOCs) more commonly reside at their location of origin or dumping unless transported in sediment; SVOCs are characteristically recalcitrant to natural or induced degradation

See Also

Engineering Geology: Codes of Practice; Aspects of Earthquakes; Geological Maps; Natural and Anthropogenic Geohazards; Site and Ground Investigation; Subsidence. **Geological Field Mapping**.

Further Reading

Anon (1970) Logging of cores for engineering purposes. Working Party Report. *Quarterly Journal of Engineering Geology* 3: 1 24.

Bell FG (1975) *Site Investigations in Areas of Mining Subsidence*. London: Newnes Butterworths.

Bell FG (1992) Description and classification of rock masses. In: Bell FG (ed.) *Engineering in Rock Masses*, pp. 54 77. London: Butterworths Heinemann.

Bell FG, Culshaw MG, Cripps JC, and Coffey JR (eds.) (1990) *Field Testing in Engineering Geology*. Engineering Geology Special Publication No. 6. London: Geological Society.

British Standards Institution (1999) *Code of Practice on Site Investigations, BS 5930*. London: British Standards Institution.

Cabrera JG and Hatheway AW (1991) Investigation of preferred sites for selection and design. In: Kiersch GA (ed.) *Heritage of Engineering Geology The First Hundred Years*, pp. 395 428. Centennial Special Volume 3. Colorado: Geological Society of America.

Clayton CRI, Matthews MC, and Simons NE (1996) *Site Investigation*. 2nd edn. Oxford: Blackwell Science Limited.

Craig C (ed.) (1996) *Advances in Site Investigation Practice*. London: Thomas Telford Press.

Culshaw MG, Bell FG, Cripps JC, and O'Hara M (eds.) (1987) *Planning in Engineering Geology*, pp. 151 154.

Engineering Geology Special Publication No 4. London: The Geological Society.
Department of the Environment (1976) *Reclamation of Derelict Land: Procedure for Locating Abandoned Mine Shafts*. London: Department of the Environment.
Griffiths JS (ed.) (2001) *Land Surface Evaluation for Engineering Practice*. Engineering Geology Special Publication No. 18. London: Geological Society.
Hatheway AW (2002) Geoenvironmental protocol for site and waste characterization of former manufactured gas plants; worldwide remediation challenge in semi volatile organic wastes. *Engineering Geology* 64: 317 338.
Hawkins AB (ed.) (1986) *Site Investigation Practice: Assessing BS 5930*. Engineering Geology Special Publication No. 2. London: The Geological Society.
Hempen GL and Hatheway AW (1992) *Geophysical Methods for Hazardous Waste Site Characterization*. Special Publication number 3. Lakewood CO: Association of Engineering Geologists.
Hudson JA, Brown ET, Fairhurst C, and Hoek E (eds.) (1993) *Comprehensive Rock Engineering: Principles, Practice and Projects*. Oxford: Pergamon Press.
International Society for Testing and Materials. (1995 to present) *Various Site Characterization Standards Developed by Committee D 18; Soil and Rock and Committee E 50: Environmental Assessment*. West Conshohocken, PA: International Society for Testing and Materials
International Society for Testing and Materials (1995) *Guide to Site Characterization for Environmental Purposes with Emphasis on Soil, Rock, the Vadose Zone and Ground Water*. Standard D 5730. West Conshohocken PA: The International Society for Testing and Materials.
Lee LT, Davios WM, Goodson RA, Powell JF, and Register BA (1994) Site characterization and analysis penetrometer system (SCAPS) field investigation at the Sierra Army Depot, Herlong, California. In: Report WES/TR/GL 94 4, U.D. NTIS order number AD A277887. Vicksburg, MS: US Army Engineering and Waterways Experimental Station.
Little AL (1969) The engineering classification of residual tropical soils. In: Moh Z C (ed.) *Proceedings of the 7th International Conference on Soil Mechanics and Foundation Engineering: Specialty Session on Engineering Properties of Lateritic Soils 28 29 August, 1969, Mexico City, DF, Mexico*. pp. 1 10. Bangkok, Thailand: Asian Institute of Technology.
McCann DM, Eddleston M, Fenning PJ, and Reeves GM (eds.) (1998) *Modern Geophysics in Engineering Geology*. Engineering Geology Special Publication No. 12. London: Geological Society.
McDowell PW, Barker RD, Butcher AP, *et al.* (2002) *Geophysics in Engineering Investigations*. CIRIA C562. London: Construction Industry Research and Information Association.
Olson O (1992) *Site Characterization and Validation Final Report*. Report STRIPA TR 92 22, (US NTIS order number DE93 603603). Stockholm: Swedish Nuclear Fuel Supply Company (SKB; SvenskKaernbraenslefoersoerjning AB).
Sara MN (1993) *Standard Handbook for Solid and Hazardous Waste Facility Assessments*, 2nd edn. Boca Raton, FL: Lewis Publishers.
Sara MN (2003) *Site Assessment and Remediation Handbook*, 2nd edn. Boca Raton, FL: Lewis Publishers, CRC Press.
Simons NE, Menzies B, and Matthews M (2002) *A Short Course in Geotechnical Site Investigation*. London: Thomas Telford Press.
Weltman AJ and Head JM (1983) *Site Investigation Manual*. CIRIA Special Publication 25. London: Construction Industry Research and Information Association.

Subsidence

A B Hawkins, Charlotte House, Bristol, UK

Introduction

Ground subsidence that results in settlement or collapse of the ground surface is grouped into four main categories: (1) subsidence due to man-made voids and natural voids relatively close to the surface (e.g., due to coal mining, stone mining, or karstic features), (2) subsidence due to the removal of fluids from depth (e.g., water/oil extraction) and the consequential change in effective stress conditions, (3) subsidence due to the removal of soluble minerals (salt, gypsum) in groundwater, and (4) subsidence due to the removal of fines in suspension (piping). In addition, subsidence may occur related to alluvial deposits, shrink/swell, volcanism, and thermokarst.

Major Voids – Man-Made and Natural

Mining

The extraction of stone for building construction, and coal for energy, has taken place since time immemorial. Early stone mining invariably extended from adits excavated into a quarry face, either when moist stone was required for carving and/or the overburden was prohibitively thick. On the other hand, coal mining generally extends from shafts, except where coal seams outcrop on valley sides in areas such as

South Wales. There are three main methods of mining, long wall, pillar and room, and linear. In long-wall mining, ground subsidence is expected, whereas in pillar-and-room/stall mining, long-term subsidence is considered possible. Subsidence is not anticipated in linear mineral mining.

Long-wall mining is a relatively recent development. Up to 80% of the horizon may be removed, generally using machinery to win material from a long face. During the extraction process, the roof is supported; behind the advancing face, the roof is allowed to collapse. Although some bulking will occur as the material collapses, the depth of settlement at the surface is commonly one-third to two-thirds the thickness of the worked seam. Settlement is likely to extend over a larger area than that mined because the failure of the ground is controlled by the angle of draw, commonly 25–45° from the vertical, depending on the nature of the strata. This method of mining is applicable only when the material won can be fragmented, and hence in Britain has been used mainly for coal extraction and the working of the Fuller's Earth Bed at Bath.

Pillar-and-room/stall mining has been carried out for many hundreds of years. It is the main method of stone extraction and an early method of winning coal and salt. The size and shape of the support pillars depend very much on the nature of the material being won and the depth of the resource. In stone mines, the pillars may be narrower because the material is stronger, but because it is also more brittle, failure may occur by surface spalling. In weak materials such as coal, salt, and gypsum, creep within the material would be anticipated and hence the pillars are invariably bigger, the height/width ratio being an important consideration. As a consequence of the mining, the ground above the rooms/stalls is de-stressed and spalling of the roof material may occur, resulting in void migration. Important considerations are the dimensions of the rooms/stalls and the nature of the overlying strata. Mudrocks may experience stress-release fissuring and spall or may act as a competent material. Over shallow mines, despite bulking, the voids may extend through to the surface to form a crown hole collapse. Alternatively, the upwards migration of the void may be intercepted by a sandstone horizon with fewer discontinuities, with the result that the original height/width ratio of the pillar may be greatly changed. In this situation, the sandstone acts as a bridge, but if overstressing of the elongated pillars results in the collapse of one or more supports, the adjacent columns must take the additional weight, creating a scenario whereby larger areas may collapse as a result of the domino effect. Typical examples are in the Coal Measures at Bathgate, south-west of Edinburgh, and the Heidegroeve area in the Netherlands, where the roof of a calcarenite mine failed in June 1988. On both occasions, the sudden collapse of the ground produced 'earthquake' shocks.

In addition to the failure of the roof over pillar-and-room workings, collapses frequently occur close to adits and/or mine shafts where the ground experiences more intense climatic change than it does deep in the mine. Ground failure also commonly occurs when shafts or bell pits have been inadequately filled and/or capped.

Linear mineral mining is generally related to the extraction of lodes of metalliferous minerals. Such narrow workings rarely cause ground subsidence because the nature of the rock does not facilitate void migration and the rocks are sufficiently competent to arch over the void. However, subsidence may occur related to old shafts, unless they have been appropriately filled or capped.

Subsidence manifested through a particular rock horizon may be created by failure of mining undertaken in lower strata. Such a phenomenon is common in the West Midlands, where the Silurian limestones have been mined as a flux for the iron and steel industry but the subsidence is seen in the overlying Coal Measure rocks.

Mining frequently involves the necessity to drain the strata. This may result in consolidation of the overlying sediments and/or dissolution as groundwater passes down to much greater depths than would otherwise occur. The consolidation resulting from a modification of the groundwater regime in the area may result in structures above and adjacent to the workings experiencing settlement. If dissolution of cements occurs, the percolating groundwaters may carry a fine fraction in suspension, which can again result in settlement.

Increasingly, tunnels and other areas of underground space are now being developed. During the driving of a tunnel, the ground above the work first experiences extension, and then, as some settlement occurs, the stress regime changes to one of compression. Even when every attempt is made to reduce subsidence and its effect on transport routes and property above a tunnel, including the use of compensation grouting, a settlement trough is a common feature.

Karst

Calcareous rocks are prone to dissolution by slightly acidic water, whether the acidity is carbonic acid from the atmosphere, humic acids from the near-surface organic-rich horizons, or sulphuric acids associated with the weathering of iron sulphides. Although the older limestones (such as the Carboniferous limestones),

having experienced greater tectonic stress and have more pronounced fractures/discontinuities, younger limestones (such as the Chalk) also have cracks and hence significant secondary permeability. The high permeability results in limited near-surface erosion and hence the calcareous rocks invariably form uplands with low groundwater levels. Where the calcium carbonate goes into solution, it moves through and may go out of the limestone mass, leaving extensive openings/cave systems, particularly in the older limestones. These underground karstic features may develop at any level, depending mainly on the depth of groundwater in the geological past. Subvertical chimneys are commonly found connecting the semihorizontal caves. These low-pressure zones result in preferential percolation, above which swallow holes and other large surface features are common. In stronger limestones, the arching effect usually prevents collapse of the rock, although spectacular hollows have occurred where the roof of an underground passage has suddenly given way. The infill of wide solution hollows, such as grykes, swallets, or pipes, may be subject to sudden washout and may lead to collapse.

In both the older limestones (such as the Carboniferous Limestone) and the younger Chalk materials, man-made excavations occur. Although in the stronger older limestones it is easy to create stable mine passageways, where Chalk has been exploited for lime burning or to remove flints for building, it is much more difficult to assess the stability of the host rock. Quite spectacular subsidences have taken place related to old Chalk workings, such as occurred at Reading in the 1990s.

Fluid Extraction

Overburden pressure is supported by the volume and structure of the underlying material and its contained fluids. If the condition is confined, with no egress of fluids, the ground will be stable. However, if there is a reduction in the pore pressure due to a net egress of fluid (egress minus ingress), then the pressure experienced by the soil/rock supporting the overburden will be greater, the increase being [in] related to the [reduction in] stress previously taken by the liquid. [With the increase in] as a consequence of the increased amount of stress taken by the soil mass, some internal failure or distortion takes place. This causes non-elastic subsidence and, due to the rearrangement and/or decrease in volume of the material mass, settlement may be manifest at the ground surface.

Groundwater Extraction

Groundwater extraction has taken place under a number of large cities and has resulted in subsidence. Notable subsidence occurred in central London between 1865 and 1931 due to water abstraction from the Chalk. Beneath Mexico City, there is an aquifer between 50 and 500 m below ground level. Progressive pumping from this aquifer for over 100 years has resulted in much of the old city settling by 4 m, and in the north-east of the city, settlement of 7.5 m has been recorded.

Problems with settlement have been noted in the Shanghai area, where up to 300 m of Quaternary deposits are present in the Yangtze delta. These deposits contain five aquifers that were extensively pumped during the twentieth century. In the early years (1921–48), only some 9 million m^3 of water was pumped, but it was noted that a ground settlement of 23 mm per annum occurred and a 19-km^2 area experienced settlement of more than 0.5 m. With the increase in population of the city, between 1949 and 1956, the extraction rose to 140 million m^3 per year, resulting in an average settlement of 43 mm and an area of 93 km^2 experiencing in excess of 0.5 m subsidence. Between 1957 and 1961, the abstraction rose to 200 million m^3 per year and the ground settlement was 100 mm per year. Having recognized the problem that the water extraction was creating, the rate of pumping was reduced to 72 million m^3 per year between 1966 and 1989, and the settlement dropped to 2 mm per year. However, in the period 1990 to 2001, the rate of extraction increased to 113 million m^3 per year and the settlement increased to 16 mm per year. As a result of the groundwater extraction, the ground in the central part of Shanghai has settled by over 2.5 m. As the urban sprawl has extended, the settlement has expanded out such that now the whole of Shanghai has effectively suffered some ground settlement due to water extraction. This has caused major problems and damage to the infrastructure-sewers, roads, subway tunnels, and buildings.

Groundwater extraction near Pixley in California caused some 0.75 m of settlement between 1958 and 1963, the surface depression being mainly over the area pumped, with the effect decreasing with distance. In the Houston/Galveston region, the presence of faults has restricted the area affected by the dewatering such that a face up to a metre high has developed over a length of almost 17 km.

Oil and Gas Extraction

Oil extraction became important in the twentieth century and has also been associated with ground subsidence. In the Wilmington Oil Field of California, subsidence was first noted late in the 1930s, within 3 years of the commencement of oil production. By 1947, the rate of subsidence was some 0.3 m per year,

and by 1951, when 140 000 barrels a day was being extracted, the rate of subsidence had reached 0.7 m per year. In 1957, a decision was made to inject water into the sediments in an attempt to repressurize them, and by 1962, subsidence had effectively ceased over most of the field. Nevertheless, by 1966, up to 9 m of settlement had occurred over an elliptical area of more than 75 km^2.

The Ekofisk Oil Field in the Norwegian sector of the North Sea is a series of dome-like structures within the Cretaceous and Paleocene Chalk, sealed from other formations by a shale/low-permeability chalk cap and flanked by low-porosity chalks. The hydrocarbon is extracted from the Chalk reservoir some 3 km below the seabed. With hydrocarbon extraction, there was little natural replenishment of the pore fluids and hence the 3 km of overburden pressure caused breakdown of some of the Chalk within the reservoir rocks, resulting in a loss of support for the overlying strata and surface settlement. The initial subsidence was reportedly in the order of 0.5 m per year, but analyses carried out subsequently have indicated that a seafloor settlement in the order of 7 m has occurred. The exploration/production platform was raised by 6 m and recharge of the reservoir was undertaken by injecting fluids.

Settlement has been reported in the Groningen Gas Field, where there is exploitation of between 70 and 240 m of Permian sandstones at a depth of some 3 km. In this area it is estimated that the subsidence will exceed 0.25 m by 2025. Although this is not a major subsidence, in the low-lying Polder region of Holland, the effect of the ground settlement is significant.

Subsidence Related to Dissolution of Salts

The most important halite deposits in Britain occur between Birmingham and Manchester, where notices frequently draw attention to the effect of salt subsidence. Subsidence related to the halite of Triassic age beneath the plains of the north-west Midlands has developed as a consequence of natural events and human activity. In many areas, particularly around Northwich and Winsford, the salt subcrops beneath the Quaternary deposits, which are frequently quite granular in character. Percolating groundwaters dissolve the salt until the waters become saturated, producing a phenomenon referred to as 'wet rock head'. The dissolution of the halite removes support from the material above, leading to subsidence of the overlying materials and settlement at the ground surface.

Although mining of the salt has taken place in the north-west Midlands, most salt is now won by brine pumping. As the liquid being drawn to the low-pressure zones around the brine pumping well passes along the natural discontinuities within the rock mass, there is no predictable size or shape of the zone from which the salt has been removed. It may be roughly circular or almost linear, depending on the structure/permeability of the strata. Clearly the dissolution of the salt will be at its maximum where the incoming water is fresh, whereas around the extraction point, provided the rate of extraction is not too fast, the water will be saturated and the brine can be brought to the surface for the salt to be crystallized out. Large subsidence features have been formed as a consequence of brine pumping, such as in Bottom Flash, near the village of Winsford. On many occasions, the main collapse takes place after extraction in an area has ceased. Subsidence up to 1.6 m has been recorded near Hengelo in The Netherlands, with some of the surface settlements occurring several years after pumping ceased.

In other areas, salt has been mined by pillar-and-stall mining, which creates large underground cavities. Because the salt is prone to dissolution and creep, these mines become unstable in time; serious problems have occurred in areas such as Northwich, where the British Government has provided grant funding for the infilling of some of the mines.

Gypsum is another mineral that is prone to dissolution. A number of subsidence features have developed in the Ripon area of Yorkshire due to gypsum mining. General areas have sunk and/or hollows have developed, in part through collapse into voids created in these soluble mineral beds.

Flowing Water

The natural flow of water through granular/noncohesive material is capable of leaching the finer fraction. When the texture of the soil is not controlled by the interaction between the coarser particles, the loss of fines removes support to the overlying material such that the coarser fraction is repositioned and densified. When this takes place, the overlying material, if uncemented, loses [the] some support, resulting in settlement, sometimes referred to as hydrocompaction.

Rainfall penetrating vertically into the ground will be deflected at the groundwater level or at the depth where the soil becomes less permeable, commonly at the top of the bedrock. Flowing down-slope, the rainwater develops a series of very small channelettes that progressively merge to form more definite channels, and then pipes, at this water level. These soil pipes are generally 75 to 100 mm wide but may be up to 200 mm across. When the flow becomes interrupted, resulting in eddying, the water within the pipes may create underground cavities, which subsequently collapse. Such subsidence hollows may be up to a metre or so deep and in the order of 1 to 3 m across.

Alluvium

Most of the alluvial deposits in the world are estuarine in nature, although fluviatile sediments occur in valleys. Generally formed in the past 8000 years associated with the rising sea-level, alluvial deposits are normally consolidated, i.e., they have never had an overburden load greater than that existing today. For this reason the sediments often have a high moisture content and high porosity.

The main types of subsidence that occur in alluvial deposits are related to densification, when the deposits are either drained and/or are loaded in excess of the natural overburden pressure. In both situations, water is squeezed out through the pore throats of the sediment, particles are rearranged, and clay minerals are bent. Where the sediments have sufficient permeability that the water can be easily squeezed out, the settlement is relatively quick and is referred to as 'primary' consolidation, whereas finer sediments with smaller pore throats and lower permeabilities will experience a steady consolidation over a longer period, known as 'secondary' consolidation.

Where organic-rich sediments occur, the initial moisture content may be in the order of 1000%, with water accumulating both between and within the structure and cells of the various plants forming the marsh/peat material. When such deposits are loaded by a linear structure such as a road or railway, or by buildings, the overburden pressure squeezes out water from the organic-rich horizon, resulting in settlement. When the Holme Post was installed near Peterborough in Great Britain in 1851, to assess the shrinkage of the peat due to drainage, the top of the post was level with the ground surface. Within 10 years, the ground level had fallen by 1.5 m, and it had dropped by some 4 m by 1970.

Cuttings Resulting in Change in Groundwater Regime

Cuttings that create a modification to the groundwater regime will invariably result in subsidence as the zone of saturation is lowered and densification occurs.

Shrink/Swell

Groundwater levels vary in depth related to wet and dry periods. In areas where evaporation is comparatively strong, the evaporation front in the dry period may be relatively low such that the ground above loses most of its free water and some of the adsorbed water. Consequently, the ground shrinks and stiffens during the dry periods. Where hedges or trees exist, the nutrient-seeking roots remove considerable quantities of groundwater from the soils, producing a bowl-shaped zone of non-saturated soil. This frequently reaches 3 or 4 m in depth, and the lateral limit of dry soils may extend beyond the canopy of the tree. In the wet season, some of this effect is reversed and, if the trees are removed, rehydration may take place over the course of 5–10 years.

Subsidence Related to Volcanism

At Pozzuoli, on the coast adjacent to Mount Vesuvius in Italy, there is evidence that the ground has experienced both subsidence and uplift as a consequence of the change in ground stresses created by the build-up and release of pressures associated with volcanic activity. The uplift is clearly evidenced by the borings of marine molluscs found in columns some 2 m above sea-level whereas the subsidence is manifest by the presence of ancient floor slabs beneath the present sea-level.

Thermokarst

In Arctic areas, subsidence occurs when frozen ground is thawed. Near Edmonton, Canada, when a power station cooling lake was created, the perimeter dams subsided by more than 1.5 m when the ice lenses within the frozen ground thawed and the overburden pressure displaced the water.

Glossary

collapse The sudden falling in or falling down of the ground.

compaction Densification of the ground induced by human activity, such as the rolling of a road pavement, whereby the particles are forced together by compactive effort to stabilize the material and to reduce interparticle voids.

consolidation A natural densification generally accompanied by a loss of water and/or the precipitation of a cement; may or may not involve particle rearrangement and/or the flexing of particles such as clay minerals.

settlement The surface manifestation of sinking ground.

subsidence The sinking of the ground.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards; Problematic Rocks; Problematic Soils.

Further Reading

Bell FG and Stacey R (1992) Subsidence in rock masses. In: Bell FG (ed.) *Engineering in Rock Masses,* pp. 246 271. London: Butterworth & Heinemann.

Chai J C, Shen S L, Zhu H h, and Zhang X L (2004) Land susbidence due to groundwater drawdown in Shanghai. *Geotechnique, LIV* 2: 143 148.

Cooper AH (1989) Airborne multispectral scanning of sub sidence caused by Permian gypsum dissolution at Ripon, North Yorkshire. *QJEG* 23: 219 239.

Evans WB, Wilson AA, Taylor BJ, and Price D (1968) *Geology of the Country around Macclesfield, Congleton, Crewe and Middlewich*. London: Institute of Geological Sciences, HMSO.

Leddra MJ and Jones ME (1990) Influence of increased effective stress on the permeability of chalks under hydro carbon reervoir conditions. In: Proceedings International Chalk Symposium, Brighton, Thomas Telford, London, pp. 253 260.

Lofgren BN (1979) Changes in aquifer system properties with groundwater depletion. In: Saxene SK (ed.) *Evalu ation an Prediction of Subsidence*. Proc. Speciality Conf. Am. Soc. Civil Engineers, Gainsville, pp. 26 47.

Price DG (1989) The collapse of the Heidegroeve: a case history of subsidence over abandoned mine workings in Cretaceous calcarenites. In: Proceedings International Chalk Symposium, Brighton, Thomas Telford, London, pp. 503 509.

Wassmann TH (1980) Mining subsidence in Twente, east Netherlands. *Geologie en Mijnbouw* 59: 225 231.

Ground Water Monitoring at Solid Waste Landfills

J W Oneacre and D Figueras, BFI, Houston, TX, USA

Introduction

Subpart E of the Subtitle D regulations in the USA require substantial ground water monitoring at municipal solid waste landfills [MSWLF]. Owners and operators must adequately define the geological and hydrogeological conditions and must develop a sampling and analysis plan that includes statistical analysis of the geochemical data. These regulations and causes of ground water variability at MSWLFs are discussed in the Further Reading Section at the end of this article. Causes for variability include inadequate site characterization, improper well design, drilling, development, and sampling, laboratory artefacts, and misapplication of statistical methods.

Under 258.54(a)(1)(2) of the Federal regulations and Texas Administrative Code [TAC] 330.234. (a)(1)(2), the executive director may delete any constituent that the owner can document is not reasonably expected to be in or derived from the waste. Also, the director can establish an alternative list of inorganic indicator constituents in lieu of some or all of the heavy metals if the alternative constituents provide a reliable indication of inorganic releases from the MSWLF unit to the groundwater.

This paper presents and proposes several alternative and innovative methods for groundwater monitoring specific to MSWLFs, that the authors believe can meet the regulatory criteria and be cost-effective. These methods include free carbon dioxide determination for landfill gas impact, thermal surveys for gas and/or leachate migration, stable isotope analyses for fingerprinting discrete sources, and passive sampling using dedicated sondes.

Free Carbon Dioxide [CO_2] Determination

The various phases of gas formation in MSWLFs are presented in Figure 1. Four phases exist and are designated as: Phase I, Aerobic; Phase II, Anaerobic Non-Methanogenic; Phase III, Anaerobic Methanogenic Unsteady; and Phase IV, Anaerobic Methanogenic Steady. During Phase II, CO_2 blooms to 50% and 90% by volume. From this peak, CO_2 decreases to about 30% to 50% during Phase IV. Although CO_2 concentration in water is less than 1 mg l at one atmosphere, enrichment to groundwater occurs due to the concentration of CO_2 in landfill gas. CO_2 is much more soluble in water than methane [CH_4], 1700 mgl compared to about 50 mg l; therefore, as the CO_2 gas

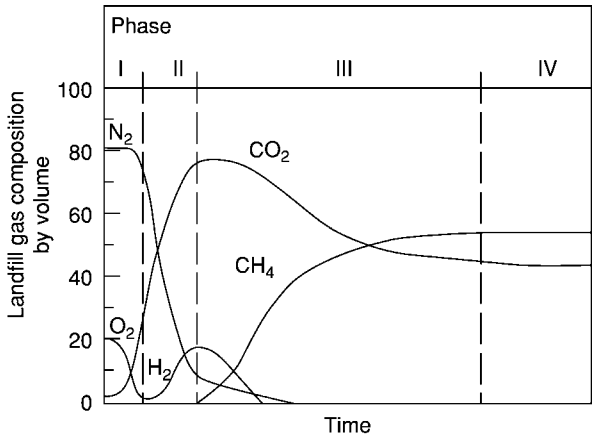


Figure 1 Phases of landfill gas formation (after Farquhar and Rovers (1973)).

pressure increases, CO_2 goes into solution, and formation of carbonic acid occurs which lowers the pH.

Landfill gas will travel through porous media via two processes; diffusion in response to a concentration gradient and convection from a pressure gradient. Gas will move more readily along paths of least resistance (zones of high hydraulic conductivity) and may move outward significant distances if upward movement is hindered by impervious layers such as frost, clay, concrete, etc. The depth to which landfill gas can penetrate is a function of site soil properties and soil moisture content. The groundwater surface represents a good lower limit of movement, particularly for methane; however, CO_2 will readily go into solution.

At MSWLFs with landfill gas migration, a full pH unit drop is typical in impacted wells as shown in Figure 2. Wells with no gas impact have pH values of about six whereas wells with most significant gas impact have a pH of five or lower. Since a relationship exists between pH, total alkalinity, and free CO_2; having values for two parameters allows the calculation of the third. For this site, the owner measures pH and total alkalinity and then determines the free CO_2 concentration. Background CO_2 values are below 50 mg l; however, gas impact has caused the CO_2 values to increase in excess of 400 mg l. Also, as CO_2 values increase, volatile organic compounds (VOCs) begin to appear and the percentages of sample with VOCs increases as CO_2 increases. In contrast, wells with background concentrations of CO_2, nearly 75% of all samples do not show VOCs (Figure 3).

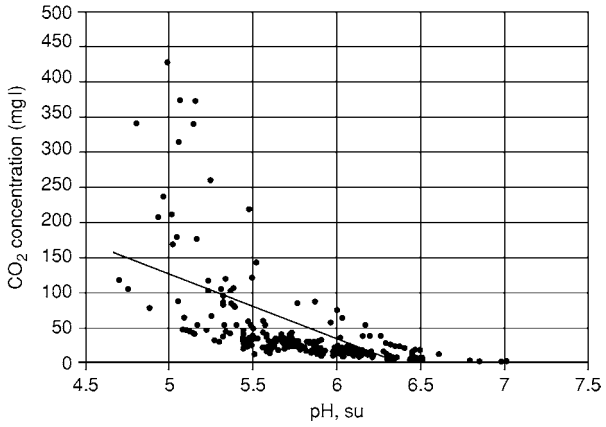


Figure 2 pH and CO_2 relationship in groundwater.

From the above discussion, there is a good correlation between elevated CO_2 values and the occurrence of VOCs. There are 149 samples with CO_2 concentrations less than 50 mg l and no detections of VOCs. In contrast, 47% of samples with CO_2 greater than 50 mg l contained VOCs and the likelihood of a VOC occurrence increases to 90% with increasing CO_2 concentration. Using free CO_2 concentrations as an indicator parameter could reduce the need for routine analyses of VOCs.

Thermal Surveys

The technical approach for using thermal surveying is based upon the fact that, in the absence of local influences, the Earth's temperature increases about 0.2 to 1.2°C with every 100 feet of depth below the ground surface. This temperature increase is due to conductive heat flow; heat moving through a medium such as

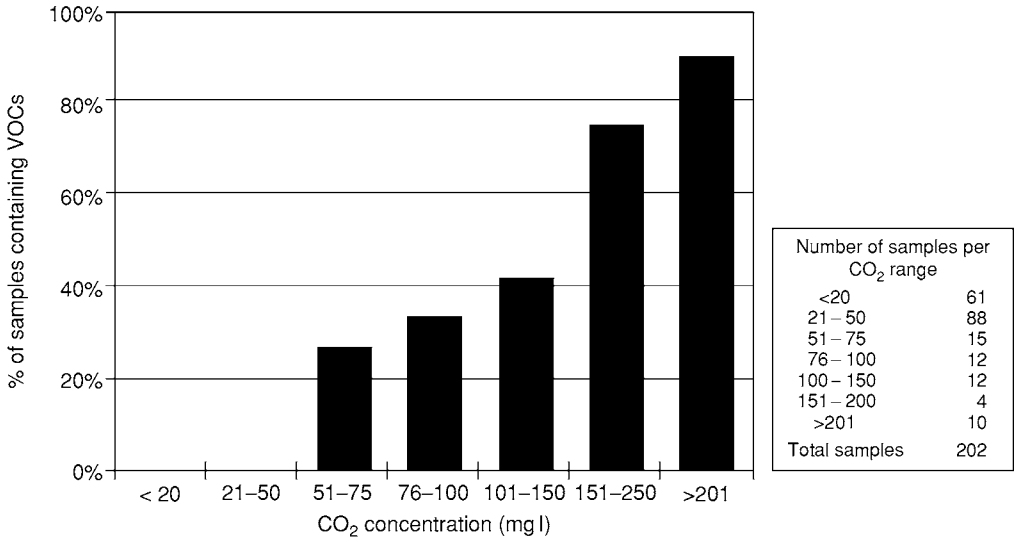


Figure 3 VOC and CO_2 relationship.

soil, rock, or groundwater as a result of temperature differences between the inner Earth and ground surface. A simple example of heat conduction is heat transferred through a knife held over an open flame; the handle will become hot to the touch as heat moves from the knife tip to the handle. At shallow depths, temperature is also affected by the diurnal, or day–night cycle, and seasonal, or summer–winter cycle. Diurnal influences extend only a few feet below the ground surface but seasonal influences can reach to depths of about 50 feet.

Heat in the Earth may also move due to advection. Gas or water can convey heat through porous media and thus create temperature differentials that may reflect subsurface media processes. A highly permeable gravel bed will carry away the interior Earth's heat more rapidly and therefore show a temperature anomaly that is lower than surrounding, less permeable material. Conversely, an anomaly high will exist where the groundwater is locally affected by heat above the normal geothermal gradient. MSWLFs generate heat as a byproduct of decomposition of wastes; this is demonstrated by the fact that the optimum temperature for methane generation is about 35°C (95°F). Once heated, water temperature will change slowly since it has a high specific heat capacity compared to most natural materials and is therefore a potentially useful tracer. In addition, increased temperature of water will alter the density and viscosity such that groundwater at 40°C will travel twice as fast as water at 5°C. Researchers have used temperature to trace thermal anomalies as much as five miles from the source. Traditional application of thermal surveys include location of groundwater supplies, identification of geothermal areas, and delineation of groundwater seepage around dams.

Precisely calibrated thermistors measure electrical resistance which varies proportionally with temperature. Thermistor probes are set at depths of about 10 feet to eliminate diurnal effects. Equilibrium requires about twenty-four hours; readings are taken with a resistance bridge and values converted to temperatures using calibration tables prepared for each sensor.

Analyses of δD and δ^{18}O help identify the probable source of groundwater. If the isotopic composition plots close to the meteoric water line in a position similar to that of present-day precipitation for a particular geographic region, the groundwater is almost certainly meteoric.

For the southern California area near a landfill test site with elevated chlorides, the expected δD and δ^{18}O values typically fall in the range of -50‰ and -7.5‰, respectively. Due to shifts in climatic patterns during the Tertiary, the time period associated with the bedrock at the test site, precipitation from the Tertiary differed isotopically from the isotopic composition of present-day precipitation. This compositional difference varies based upon the specific stratigraphic horizon within the Tertiary. The significance of the isotopic differences is discussed below.

D/H and ^{18}O in Landfill Leachates

It has been observed that significant deuterium enrichment in landfill leachates are due in part to the anaerobic decomposition of organic wastes within landfills. In addition, it was noted that increased δD and δ^{18}O in groundwater were contaminated by an adjacent landfill. Leachates from landfills in Illinois have δD values enriched about 30% to 60% relative to δD of local precipitation.

A significant portion of the hydrogen in landfill methane comes from the landfill leachate where microbes preferentially use the lighter hydrogen isotope, thereby enriching the leachate with deuterium. As the landfill ages and more methane is generated, more enrichment of deuterium occurs. Thus deuterium enrichment makes deuterium a very useful tracer for detecting groundwater impact from landfills.

^{13}C in Landfills

The major components of landfill gas, methane, and carbon dioxide (CH^4 and CO^2), will show unique isotopic signatures compared to the surrounding environment. Methanogenesis will cause the CO^2 portion of the landfill gas to become enriched relative to δ^{13}C; whereas the CH^4 portion will become depleted relative to δ^{13}C. δ^{13}C values of around $+20$‰ for CO^2 and -50‰ for CH^4 have been reported. Values for δ^{13}C will increase as the landfill ages since the microbes will prefer to use the isotopically light carbon to produce CH^4, thereby enriching the CO^2 at the expense of the CH^4.

Because CO^2 is much more soluble in water than CH^4 (around 1700 mg l compared to 50 mg l), landfill leachates will also become enriched with δ^{13}C due to increased alkalinity. Using the isotopes of hydrogen and carbon, researchers have been able to differentiate various sources for CH^4.

Environmental Isotopic Analyses

General Principles

Isotopes of specific elements, such as hydrogen, carbon, and oxygen, have the same number of protons but different number of neutrons in the

nucleus. This gives a different atomic weight for a specific element, even though the element maintains the same atomic number. As an example, hydrogen occurs as 1H with an atomic weight of 1 or as 2H (deuterium) or 3H (tritium) with atomic weights of 2 and 3, respectively. Stable isotopes do not undergo any natural radioactive decay and provide an understanding to the source of water or a process affecting water after release from the atmosphere. Any process that causes different isotopic ratios in different media is the result of isotopic fractionation. The most important process is vapour-liquid fractionation caused by evaporation and condensation. For instance, $^{18}O/^{16}O$ in rainwater is different from ocean water, which is different from oil brines.

For convenience, the δ (delta) notation is used to describe isotope ratios, which is defined as:

$$\delta \times [\text{sample}] = (\text{R}[\text{sample}] - \text{R}[\text{standard}])/\text{R}[\text{standard}] * 1000$$

where δX represents the relative difference in parts per thousand (called per mil (‰)) between the isotope ratio (R) in a sample and the ratio in some specified standard. For $\delta^{18}O$, the reference standard is SMOW, Standard Mean Ocean Water; $\delta^{13}C$ uses the PDB scale, a belemnite from the Pee Dee Formation of South Carolina; and $\delta^{34}S$ uses the Canyon Diablo troilite as the reference standard. Stable isotope values associated with landfill leachates and gases demonstrate pronounced signatures that are separate from the surrounding environment.

$^{18}O/^{16}O$ and D/H in Groundwater

The isotopic composition of seawater, is by definition, zero per mil for both $\delta^{18}O$ and δD. Upon evaporation from seawater, the water vapour has an isotopic concentration of about -80‰ for δD and -9‰ for $\delta^{18}O$. Fractionation processes affect the isotopes such that rain becomes progressively lighter in both δD and $\delta^{18}O$, as it moves farther from the ocean source. Data from modern rainfall plots on a straight line defined as:

$$\delta D = 8\delta^{18}O + 10 \quad [2]$$

It has been suggested that an upper limit of $+18$‰ $+/-2$‰ exists for $\delta^{13}C$ in landfills due to eventual steady state conditions that develop between CO^2 input and CH^4 production.

^{13}C in Oilfield Water

Data regarding $\delta^{13}C$ in oilfield water are not abundant; however, study of $\delta^{13}C$ values for Miocene petroleum source rocks in California showed high $\delta^{13}C$ values, greater than $+5$‰ PDB. $\delta^{13}C$ studies conducted on Miocene rocks in California showed $\delta^{13}C$ values typically above $+5$ and as high as $+27.8$‰ PDB. Thus, the very high $\delta^{13}C$ values were the result of degradation of organic acid anions by sulphate-reducing bacteria. Decarboxylation of short-chain aliphatic acid anions, principally acetate, is a probable major source for CO_2 in the oilfield waters.

Isotopic Data Interpretation

Isotopic data from the California landfill site have been plotted in various combinations in Figures 4 and 5. Figure 4 depicts $\delta^{18}O$ ratio versus δD ratio and shows the meteoric water line for rainfall in the southern California area. If data points plot off this line, this would indicate that the water (fluid) tested had been formed in an environment significantly different and unique with respect to normal precipitation. Site groundwater samples tend to plot on or below the meteoric water line, showing that they are strongly related to modern precipitation. The brine samples plot well above the line, indicating a different source or time period for this water. The leachate samples are far removed from either brine or groundwater, showing a totally different and distinct process of formation. This plot clearly shows that leachate, groundwater, and brine have distinct isotopic signatures. MW-1, the highest chloride well, and the 'probe', (completed in highly fractured bedrock) are strongly shifted toward the Pico #1 brine sample. There is a strong probability based on this plot, that there is some mixing between the oilfield brine being produced from the Pico #1 (1500 foot zone) and the water samples obtained from MW-1 and the probe. Absolutely no mixing of groundwater and leachate are indicated based on this plot.

Figure 5 is a simple plot that illustrates the wide variance in carbon isotopic analysis of samples from groundwater, brine, and leachate sources. The dissolved inorganic carbon 13/12 ratios have distinct and definitive isotopic signatures. This indicates that the carbon isotopes are extremely useful in delineating any mixing between brine, leachate, and groundwater. Based upon the simple comparison of carbon isotopes, there is no relationship between the groundwater samples and landfill leachate.

Passive *In Situ* Sampling

Much research and discussion have occurred in recent years regarding appropriate groundwater sampling practices.

It has been shown that variable speed centrifugal pumps reduce purge time, aid in slug and pump tests,

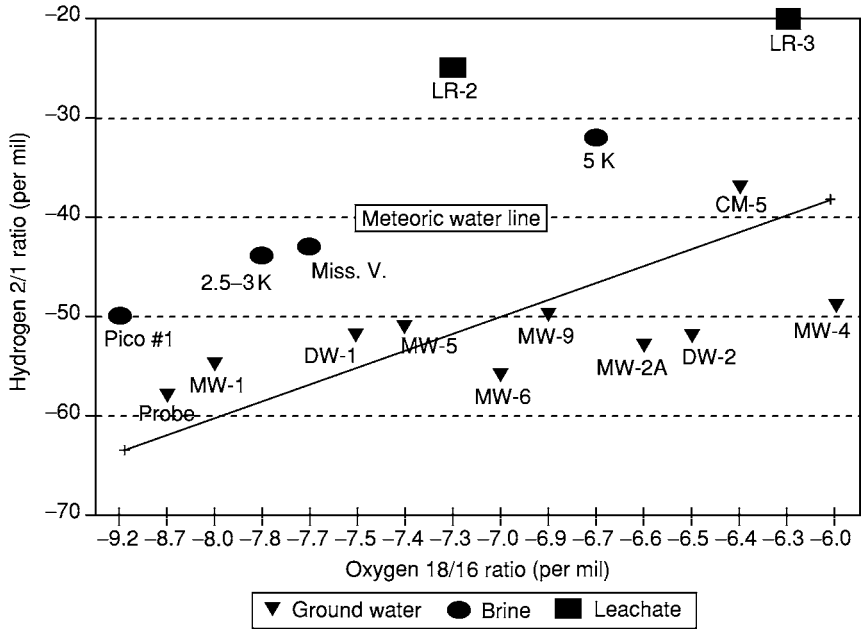


Figure 4 ^{18}O vs. D for brine, leachate, and groundwater for Californian landfill site.

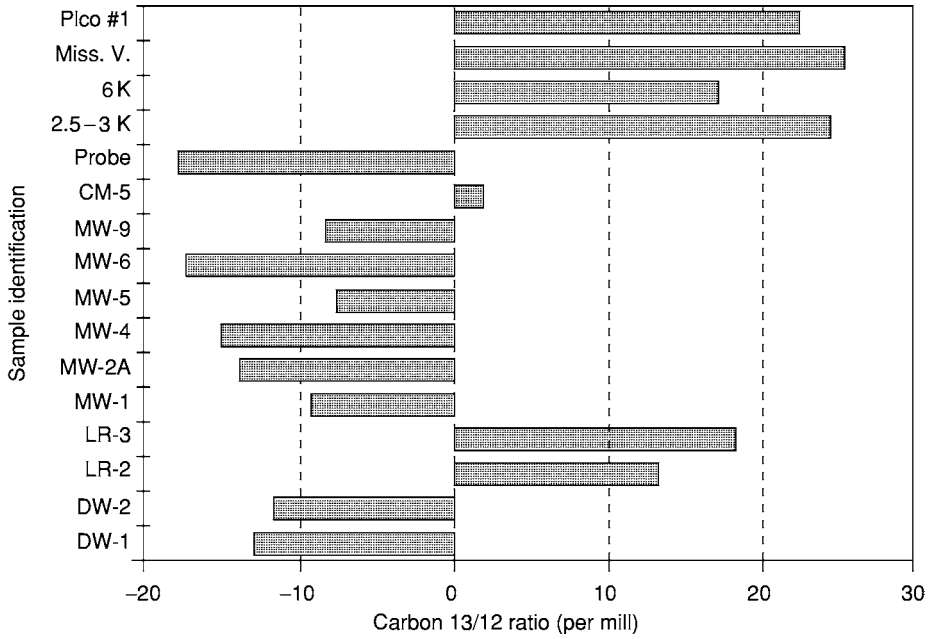


Figure 5 Carbon 13/12 ratios for brine, leachate, and groundwater for Californian landfill site.

and consistently provide low turbidity. Recent studies suggest that dissolved oxygen (DO) and Redox, key field parameters for groundwater sampling, are the last field parameters to stabilise. Unfortunately, these parameters are infrequently measured, even though DO and Redox play an important role in geochemical equilibrium of trace constituents such as heavy metals. Figure 6 shows upgradient and downgradient DO and Redox data from a MSWLF in New England with a defined leachate plume. Due to the anaerobic nature of leachate, impacted downgradient wells are depleted of DO and show a strongly reduced condition. Preliminary findings suggest that purging a nominal one well volume is necessary to establish equilibrium of DO and Redox and an average is two well volumes. In addition, atmospheric conditions

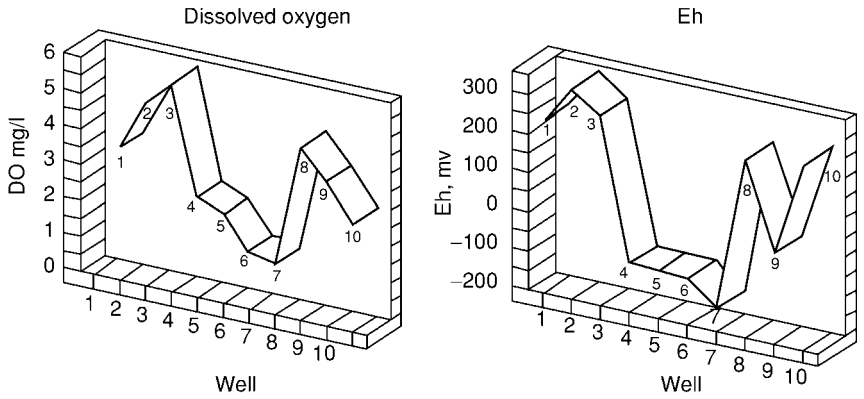


Figure 6 Reduction of DO and Eh in groundwater from leachate impact.

apparently influence DO and Redox in a standing column of well water to depths of 30 feet.

Low-flow purging that removes a minimal volume of water may not fully remove this atmospheric artifact. Figure 7 shows readings for DO and Redox, demonstrating an apparent 'stabilization' of the two parameters occurring at low-flow purging. Based upon USEPA guidelines for sampling, one would conclude that representative formation water is present in the well. However, increasing the flow rate shows a dramatic decrease in Redox of about 60%; a better representation of true formation water. This has significance for parameters sensitive to reducing conditions such as arsenic. Low-flow purging may have benefits for sites with known contamination, but for routine detection monitoring, may have limitations.

In situ passive sampling may have the benefit of eliminating purging while measuring key parameters specific to MSWLFs. *In situ* passive sampling can also provide continuous monitoring rather than semi-annual sampling. The owner can set the sampling intervals to weekly, daily, hourly, or by the minute. Maintenance is low, with only periodic recalibrations and battery replacement. Such *in situ* passive sampling devices can fit into 2 inch wells and monitor as many as eight parameters. For MSWLFs, key parameters would include: water level, temperature, pH, Specific Conductance, DO, Redox, chloride, ammonium, and CO_2. As stated previously, landfill gas impact is characterised by increased CO_2 and depressed pH. Conversely, leachate impact is characterized by increases in temperature, Specific Conductance, chloride, and ammonium, while DO and Redox will show dramatic decreases.

By obtaining data on a daily or weekly basis, short-term temporal or long-term seasonal changes can be noted and filtered from the data. For example, Figure 8 shows seasonal temperature data for the New England MSWLF.

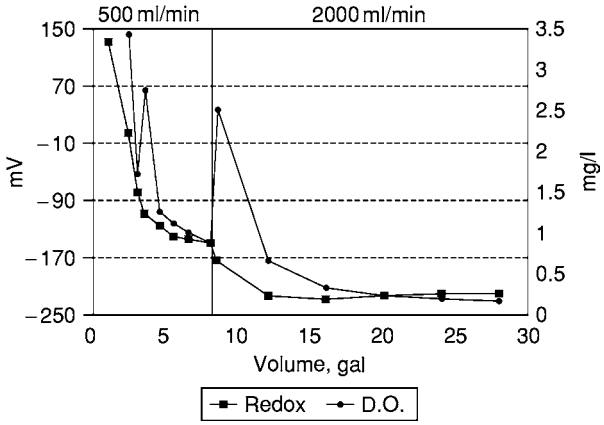


Figure 7 Changes in Eh and DO due to pumping rates in well MW 1 3D.

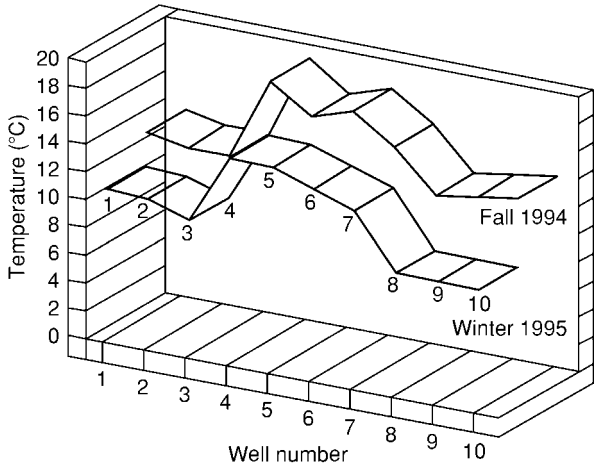


Figure 8 Seasonal temperature anomalies from leachate impact.

Note that seasonal fluctuations occur for temperature; however, the most impacted wells the GHR-11, maintains their high temperature anomaly from

early autumn to winter. By utilising *in situ* passive samplers, owners and regulatory agencies could realize several key benefits:

- *Parameters Unique to MSWLFs* – The Subtitle D parameter list has inherent problems. Firstly, most of the heavy metals are not critical to MSWLFs; secondly, some are soluble under oxidising conditions and/or low pH, a situation not pertinent due to the strongly reducing and buffering nature of MSWLF leachate; thirdly, the list implicitly assumes leachate impact only and does not address impact from landfill gas. Alternative parameters, such as those presented in this article, not only serve as better indications of groundwater impact but also will better delineate the means of impact, leachate, or landfill gas.
- *Continuous Monitoring of Ground Water* – Rather than quarterly or semi-annual monitoring, *in situ* passive sampling can provide much more data that depicts temporal, seasonal, or long-term trends in groundwater quality.
- *Reduced Frequency of Sampling* – Owners could reduce sampling frequency by using the data from *in situ* monitoring to determine if and when a sample should be taken. By establishing confidence intervals or other statistical criteria, owners would take samples only when exceedences of such criteria occurred.
- *Cost Control* – Reduction of groundwater monitoring costs, particularly post-closure accruals, would assist owners with small buckets, such as municipalities.
- *Error Reduction* – Data is simply downloaded from a datalogger to a laptop computer, thereby reducing a significant source of error, manual transcription.
- *Better Public Relationship* – Continuous monitoring could reduce public concern and increase confidence that a MSWLF is operated properly.

Conclusion

This article presents options to traditional groundwater monitoring at MSWLFs. By incorporating these options, owners can better delineate leachate versus gas impact; fingerprint definitive sources, and generate more and higher-quality groundwater data specific to MSWLFs. *In Situ* passive sampling may provide several benefits that include sampling frequency reduction, understanding of temporal trends, cost control, transcription error reduction, and better public confidence.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards; Liquefaction; Made Ground. **Quarrying**.

Further Reading

Baedecker MJ and Back W (1979) Hydrogeological Pro cesses and Chemical Reactions at a Landfill. *Ground Water* 17: 429 437.

Barcelona Michael, *et al*. (1990) *Contamination of Ground Water: Prevention, Assessment, Restoration*. Park Ridge, New Jersey: Noyes Data Corporation.

Beluche R (1968) *Degradation of Solid Substrate in a Sani tary landfill*. Ph.D. Thesis, University of Southern Cali fornia, Los Angeles, California.

Birman J (1990) *Handbook of Ground Water Development by Roscoe Moss Company*. New York: John Wiley and Sons.

Carothers WW and Kharaka YK (1980) Stable Carbon Isotopes Of HCO_3 in Oil field Waters Implications for the Origin of CO_2. *Geochimica et Cosmochimica Acta* 44: 323 332.

Coleman DD, *et al*. (1993) Identification of Landfill Methane using Carbon and Hydrogen Isotope Analysis. *Proceedings of 16th International Madison Waste Con ference*, pp. 303 314. University of Wisconsin Madison.

Drever JI (1982) *The Geochemistry of Natural Waters*. Englewood Cliffs: NJ Prentice Hall.

Farquhar GJ and Rovers FA (1973) *Gas Production During Refuse Decomposition, Water, Air and Soil Pollution*. Vol. 2.

Matthess FPG and Brown RM (1976) Deuterium and oxygen 18 as indicators of Leachwater Movement from a Sanitary Landfill. In: *Interpretation of Environ mental Isotope and Hydrochemical Data in Ground water Hydrology*. Vienna International Atomic Energy Agency.

Games LM and Hayes JM (1977) Carbon Isotopic Study of the Fate of Landfill Leachate in Groundwater. *Journal of Water Pollution Control Federation* 49: 668 677.

Hackley KC, Liu CL, and Coleman DD (1996) Environ mental Isotope Characteristics of Landfill Leachates and Gases. *Ground Water* 34(5): 827 834.

Liu, *et al*. (1992) *Application of Environmental Isotopes to Characterize Landfill Gases and Leachate*. Geological Society of America Abstracts with Programs, Cincinnati, Ohio, p A35.

Ludwig H (1967) *Final report: In Situ Investigation of Gases Produced from Decomposing Refuse*. California State Water Quality Control Board.

Murata KJ, Friedman I, and Madsen BM (1969) *Isotopic Composition of Diagenic Carbonates in marine Miocene Formations of California and Oregon*. US Geology Survey Professional Paper 614 B.

Oneacre JW (1992) *Solid waste Principles and Practice in the United States*. International Conference on Environ mental Protection and Control Technology, Enserach, Kuala Lumpur, Malaysia, pp. 661 674.

Oneacre JW (1993) Subtitle D Regulations Impact on Ground Water Monitoring. *Geotechnical News* 11(3): 49 52.

Oneacre JW and Figueras D (1996) *Ground Water Variabil ity at Sanitary Landfills: Causes and Solutions, Uncer tainty in the Geologic Environment*. Proceedings, ASCE, Madison, WI, pp. 965 987.

Ramawsamy JN (1970) *Effects of Acid and Gas Production as Sanitary Landfills*. Ph.D. Dissertation,West Virginia University.

Rank, *et al.* (1992) *Environmental Isotopes Study at the Breitenqu Experimental Landfill (Lower Austria). Tracer Hydrology*. In: Hotzl and Werner (eds.) Proceedings of the 6th International Symposium on Water Tracing, Karls ruhe, Germany, Sept 21 26, pp. 173 177. Rotterdam: Balkema.

Whiticar MJ and Faber D (1985) Methane Oxidation in Sediment and Water Column Environments Isotope Evidence. *Advances in Organic Geochemistry* 10: 759 768.

ENVIRONMENTAL GEOCHEMISTRY

W E Dubbin, The Natural History Museum, London, UK

Introduction

Soils and sediments occupying the Earth's surface lie at the interface of the lithosphere, atmosphere, biosphere, and hydrosphere. Within the weathered, complex, and porous milieu of these Earth-surface materials, myriad biogeochemical processes govern the movement of both nutrients and pollutants from the lithosphere to biota, where they are incorporated into plant and animal tissues, or to groundwaters, where they may be transported great distances to streams, rivers, and oceans. This chapter describes the most notable pollutants, both organic and inorganic, and the dominant processes governing their mobility and bioavailability in terrestrial and aquatic environments.

Trace Elements

Trace elements are those that occur in the lithosphere at concentrations typically less than $1\,g\,kg^{-1}$ (**Table 1**). Among the trace elements are the micronutrients (e.g., Cu, Ni, Zn), which are essential for the growth and development of micro-organisms, plants, and animals, and also the metalloids, which have characteristics of both metals and non-metals (e.g., As, B) (*see* **Minerals:** Arsenates). Heavy metals are defined as those trace elements with densities $>5.0\,g\,cm^{-3}$. Virtually all trace elements, even the micronutrients, exhibit toxicity to animals and plants when present at excessive concentrations. Radionuclides are a separate but important class of inorganic contaminant that may occur naturally (e.g., ^{222}Rn, ^{226}Ra, ^{238}U) or as a consequence of nuclear fission related to atomic weapons testing and nuclear power generation (e.g., ^{90}Sr, ^{137}Cs, ^{239}Pu).

Although trace elements occur naturally in all terrestrial environments, anthropogenic inputs may increase these concentrations considerably. The principal anthropogenic sources of trace elements are mining and smelting activities, fossil fuel combustion, chemical and electronics industries, as well as the addition of fertilisers and biosolids arising from agricultural operations. One notable example of severe trace metal pollution caused by smelting activities is found near Karabash, in the south Ural Mountains region of Russia, where for decades smelting operations have deposited metals on the surrounding landscape, destroying much of the vegetation and so contributing to widespread soil erosion (*see* **Environmental Geology**).

Toxic levels of trace elements (e.g., Cr, Ni) may also occur naturally, as in soils derived from serpentinitic rocks, which can lead to phytotoxicity and the consequent lack of vegetation over large areas of the landscape where the serpentine soils occur (*see* **Clay Minerals**). Alternatively, these elevated trace element concentrations may induce metal tolerance among certain plant species (e.g., *Thlaspi* spp.) as biological communities adapt to these metal rich environments.

Trace Element Bioavailability and Speciation

It is well established that total trace element content in soil or sediment is a poor indicator of toxicity. A more reliable measure of ecotoxicity is trace element bioavailability. A trace element is considered bioavailable if it can be utilized by biota. Bioavailability is therefore broadly equated with solubility, although some plants and micro-organisms are able to extract metals from solid phases normally considered insoluble. The solubility, and hence bioavailability, of a particular trace element is determined largely by its solid-phase speciation and mode of surface complexation. The main parameters

Oneacre JW (1993) Subtitle D Regulations Impact on Ground Water Monitoring. *Geotechnical News* 11(3): 49 52.

Oneacre JW and Figueras D (1996) *Ground Water Variabil ity at Sanitary Landfills: Causes and Solutions, Uncer tainty in the Geologic Environment*. Proceedings, ASCE, Madison, WI, pp. 965 987.

Ramawsamy JN (1970) *Effects of Acid and Gas Production as Sanitary Landfills*. Ph.D. Dissertation,West Virginia University.

Rank, *et al.* (1992) *Environmental Isotopes Study at the Breitenqu Experimental Landfill (Lower Austria). Tracer Hydrology*. In: Hotzl and Werner (eds.) Proceedings of the 6th International Symposium on Water Tracing, Karls ruhe, Germany, Sept 21 26, pp. 173 177. Rotterdam: Balkema.

Whiticar MJ and Faber D (1985) Methane Oxidation in Sediment and Water Column Environments Isotope Evidence. *Advances in Organic Geochemistry* 10: 759 768.

ENVIRONMENTAL GEOCHEMISTRY

W E Dubbin, The Natural History Museum, London, UK

Introduction

Soils and sediments occupying the Earth's surface lie at the interface of the lithosphere, atmosphere, biosphere, and hydrosphere. Within the weathered, complex, and porous milieu of these Earth-surface materials, myriad biogeochemical processes govern the movement of both nutrients and pollutants from the lithosphere to biota, where they are incorporated into plant and animal tissues, or to groundwaters, where they may be transported great distances to streams, rivers, and oceans. This chapter describes the most notable pollutants, both organic and inorganic, and the dominant processes governing their mobility and bioavailability in terrestrial and aquatic environments.

Trace Elements

Trace elements are those that occur in the lithosphere at concentrations typically less than $1\,g\,kg^{-1}$ (Table 1). Among the trace elements are the micronutrients (e.g., Cu, Ni, Zn), which are essential for the growth and development of micro-organisms, plants, and animals, and also the metalloids, which have characteristics of both metals and non-metals (e.g., As, B) (*see* **Minerals:** Arsenates). Heavy metals are defined as those trace elements with densities $>5.0\,g\,cm^{-3}$. Virtually all trace elements, even the micronutrients, exhibit toxicity to animals and plants when present at excessive concentrations. Radionuclides are a separate but important class of inorganic contaminant that may occur naturally (e.g., ^{222}Rn, ^{226}Ra, ^{238}U) or as a consequence of nuclear fission related to atomic weapons testing and nuclear power generation (e.g., ^{90}Sr, ^{137}Cs, ^{239}Pu).

Although trace elements occur naturally in all terrestrial environments, anthropogenic inputs may increase these concentrations considerably. The principal anthropogenic sources of trace elements are mining and smelting activities, fossil fuel combustion, chemical and electronics industries, as well as the addition of fertilisers and biosolids arising from agricultural operations. One notable example of severe trace metal pollution caused by smelting activities is found near Karabash, in the south Ural Mountains region of Russia, where for decades smelting operations have deposited metals on the surrounding landscape, destroying much of the vegetation and so contributing to widespread soil erosion (*see* **Environmental Geology**).

Toxic levels of trace elements (e.g., Cr, Ni) may also occur naturally, as in soils derived from serpentinitic rocks, which can lead to phytotoxicity and the consequent lack of vegetation over large areas of the landscape where the serpentine soils occur (*see* **Clay Minerals**). Alternatively, these elevated trace element concentrations may induce metal tolerance among certain plant species (e.g., *Thlaspi* spp.) as biological communities adapt to these metal rich environments.

Trace Element Bioavailability and Speciation

It is well established that total trace element content in soil or sediment is a poor indicator of toxicity. A more reliable measure of ecotoxicity is trace element bioavailability. A trace element is considered bioavailable if it can be utilized by biota. Bioavailability is therefore broadly equated with solubility, although some plants and micro-organisms are able to extract metals from solid phases normally considered insoluble. The solubility, and hence bioavailability, of a particular trace element is determined largely by its solid-phase speciation and mode of surface complexation. The main parameters

Table 1 Abundance, speciation, and toxicity of trace elements in soil and sediment

Element	*Median soil content*[a] *(mg kg ^{1})*	*Dominant solution species*[b]	*Function; Toxicity*
Ag	0.05	Ag^{+}, AgCl	None known; Plant and animal toxin
As	6	$As(OH)_3$, AsO_3^3 , H_2AsO_4 , $HAsO_4^2$	None known; Plant and animal toxin
B	20	H_3BO_3, H_2BO_3	Plant nutrient; Phytotoxin
Be	0.3	Be^{2+}, $Be(OH)_3$, $Be(OH)_4^2$	None known; Plant and animal toxin
Bi	0.2	Bi^{3+}	None known; Plant and animal toxin
Cd	0.35	Cd^{2+}, $CdSO_4$, CdCl , $CdHCO_3^{+}$	None known; Animal toxin
Cl	100	Cl	Plant nutrient; Phytotoxin
Co	8	Co^{2+}, $CoSO_4$, $Co(OH)_2$	Plant and animal nutrient; Plant and animal toxin
Cr	70	$Cr(OH)_2^{+}$, $Cr(OH)_4$, $HCrO_4$	Animal nutrient; Cr(VI) is a plant and animal toxin
Cs	4	Cs^{+}	None known; None known
Cu	30	Cu_2^{+}, CuCl , $CuCO_3$, $CuHCO_3^{+}$, $Cu(OH)^{+}$	Plant and animal nutrient; Plant and animal toxin
Hg	0.06	Hg^{2+}, $HgCl_2$, CH_3Hg^{+}, $Hg(OH)_2$	None known; Plant and animal toxin
Mo	1.2	H_2MoO_4, $HMoO_4$, MoO_4^2	Plant and animal nutrient; Plant and animal toxin
Ni	50	Ni^{2+}, $NiSO_4$, $NiHCO_3^{+}$, $NiCO_3$	Plant and animal nutrient; Plant and animal toxin
Pb	19	Pb^{2+}, $PbSO_4$, $PbHCO_3^{+}$, $PbCO_3$, $PbOH^{+}$	None known; Plant and animal toxin
Rb	67	Rb^{+}	None known; Phytotoxin
Sb	1	$Sb(OH)_2^{+}$, $Sb(OH)_3$, $Sb(OH)_4$, $Sb(OH)_6$	None known; Plant and animal toxin
Se	0.4	$HSeO_3$, SeO_3^2 , SeO_4^2	Animal nutrient; Plant and animal toxin
Sn	4	Sn^{4+}	None known; Plant and animal toxin
Tl	0.2	Tl^{+}	None known; Plant and animal toxin
U	2.7	$UO_{2(s)}$, UO_2^{2+}, UO_2CO_3, $UO_2(CO_3)_3^4$, $(UO_2)_3(OH)_7$	None known; Animal toxin
V	90	VO^{2+}, VO_2^{+}, $VO_2(OH)_2$, $VO_3(OH)^2$	Plant and animal nutrient; Plant and animal toxin
W	1.5	WO_4^2	None known; None known
Zn	70	Zn^{2+}, $ZnSO_4$, $ZnHCO_3^{+}$, $ZnCO_3$, $Zn(OH)^{+}$	Plant and animal nutrient; Plant and animal toxin

[a] Bowen (1979).
[b] Hayes and Traina (1998).

governing trace element speciation are pH and p*e*, the so-called 'master variables', as well as the total concentration of the ion in solution. These variables determine whether the ion in question will undergo hydrolysis, precipitation, redox reactions, or any number of complexation reactions at surfaces or in aqueous solution. Generally, metal cations are most soluble and available at low pH, principally pH <5, where hydrolysis is minimal and sorption to layer silicates and metal oxides is limited. Conversely, oxyanions such as chromate (*see* **Minerals:** Chromates) ($HCrO_4$) and selenite ($HSeO_3$) show greatest solubility at high pH, where electrostatic repulsion with negatively charged colloids minimizes surface complexation.

Redox-sensitive elements such as As, Cr, Fe, and Mn are subject to oxidation state changes under Earth surface conditions, with important implications for solubility and toxicity. Of the two As oxidation states predominating under Earth surface conditions (i.e., As(III) and As(V)), As(III) is more problematic because of its greater mobility and bioavailability. Given the ubiquity of manganese oxides (e.g., δ-MnO_2) in soils and sediments, these minerals have been proposed as key agents in the natural attenuation of As(III) contamination, as the oxidation of As(III) to As(V) by Mn(IV) is both rapid and thermodynamically favourable. The human health implications of As bioavailability is illustrated most clearly on the deltaic plains of Bangladesh and West Bengal, where groundwater As concentrations may exceed 400 $\mu g\,l^{-1}$ as a consequence of biologically mediated reductive dissolution of Fe(III) oxides, leading to the release to solution, in bioavailable form, of the previously sorbed As. Manganese(IV) oxides can also mediate the oxidation of Cr, from Cr(III) to Cr(VI). However, this redox reaction is undesirable because, unlike As, the oxidized form of Cr, Cr(VI), is the form of Cr most mobile and toxic.

Organic Contaminants

The environmental geochemistry of organic contaminants primarily concerns the sources, movement, and fate of petroleum hydrocarbons and their by-products, as well as the halogenated hydrocarbons, the group to which many pesticides belong. Petroleum hydrocarbons and associated compounds such as the oxygenates (e.g., methyl *t*-butyl ether (MTBE)) constitute a significant environmental risk by virtue of their widespread occurrence, mobility, and ecotoxicity (*see* **Geochemical Exploration**). Following release of hydrocarbons to the environment by multiple pathways, the low molecular weight volatile fraction ($<C_{15}$), often containing the carcinogenic benzene and polycyclic aromatic hydrocarbons, is largely lost to the atmosphere through volatilisation, leaving a relatively small but environmentally important portion of the light hydrocarbon pool to react with soil or sediment, or to enter groundwater. The less soluble, more chemically inert higher molecular weight hydrocarbons ($>C_{14}$) are potentially more disruptive to ecosystems, as illustrated by the spill in 1989 of crude oil from the *Exxon Valdez* oil tanker, which contaminated nearly 1750 km of Alaskan shoreline.

Halogenated hydrocarbons have both natural and anthropogenic origins, and many belong to the class of contaminant known as persistent organic pollutants (POPs), recalcitrant organic compounds that bioaccumulate and exhibit animal toxicity. POPs are dominated by the chlorinated hydrocarbons, which include many of the pesticides, such as aldrin, atrazine, chlordane, DDT, heptachlor, and the polychlorinated biphenyls. Despite the relative recalcitrance of POPs, their degradation can be mediated both by abiotic processes (e.g., oxidation by δ-MnO_2) as well as by the native soil microflora, particularly the fungi, which employ hydrolytic, reductive, or oxidative reactions to induce molecular dehalogenation. The hydroxylated compound so produced is thus rendered more susceptible to the degradation reactions of other soil microbiota.

The tendency of certain POPs to migrate from tropical and temperate climates to the colder polar regions has been the subject of study and debate for decades. A model of redistribution described as 'global distillation', involving POP evaporation followed by transport and condensation in colder regions, has received wide acceptance. Fractionation of POPs, during redistribution to higher latitudes, is driven by differential migration rates arising from variable POP vapour pressures and partition coefficients, with POP transport occurring in distinct jumps which are closely coupled to diurnal and seasonal temperature cycles.

Acidification of Terrestrial and Aquatic Environments

An important aspect of environmental geochemistry is acid deposition and the related acidification of Earth surface environments through both anthropogenic and natural processes. The burning of fossil fuels releases SO_2 and NO_x compounds which combine with atmospheric water to yield H_2SO_4 and HNO_3 that may be carried great distances before deposition as rain, mist, fog, or snow. Deposition of these acidic materials impacts negatively on soils, vegetation, and water bodies, particularly lakes which are poorly buffered and whose aquatic organisms are therefore at risk of increased soluble Al concentrations following a significant decrease in lake pH. Monuments and buildings constructed of limestone and marble, and which frequently represent much of our cultural heritage, are also at risk from acid deposition through dissolution of their constituent carbonate minerals (*see* **Minerals:** Carbonates). The effects of acid deposition are not entirely negative, however, as the additions of N and S to soils are beneficial, and these added nutrients frequently comprise a significant portion of the available soil N and S in highly industrialized regions.

An aspect of acidification of growing environmental importance concerns the oxidation of mining waste rich in reduced S, principally in the form of pyrite (FeS_2) (*see* **Environmental Geology, Minerals:** Sulphides). Oxidation of this S can yield vast amounts of H_2SO_4, giving rise to highly acidic waters, known as acid mine drainage (AMD) waters, containing toxic levels of soluble metals (**Figure 1**). The environmental significance of these acidic, metal-rich waters was emphasised in dramatic fashion with the collapse in 1998 of the Aznalcóllar mine tailings dam in southwest Spain. Failure of the dam led to the release of 1.3 million cubic metres of AMD waters, laden with Ag, As, Cd, Cu, Pb, and Zn, and the subsequent flooding of nearly 4,600 hectares of land with this toxic effluent.

Environmental Restoration

Decontamination of terrestrial environments is often costly and time-consuming, owing to the frequent occurrence of multiple pollutants, as well as the complexity of the contaminated matrices (e.g., soil or sediment). Strategies employed to remediate contaminated environments involve both *in situ* and *ex situ* techniques (**Table 2**), with the former generally receiving wider acceptance because of greater efficacy and lower implementation costs. *In situ* remediation may simply involve introducing a liming material

(e.g., CaO, $Ca(OH)_2$, $CaCO_3$) to raise pH and thus decrease metal bioavailability through sorption and precipitation. The addition of smectites or zeolites (*see* **Minerals:** Zeolites) serves to remove pollutants by means of ion exchange reactions, whereas apatite addition induces the formation of sparingly soluble metal phosphate precipitates.

Figure 1 Acid mine drainage waters near the Kristineberg Cu and Zn mine, Sweden (photo courtesy L. Lovgren, Umea University).

Phytoremediation encompasses a variety of *in situ* strategies involving the use of plants to remove or render environmental pollutants harmless. The most common of these strategies is phyto-extraction, which involves growing plants capable of hyperaccumulating the contaminant of interest, followed by plant harvest to remove both plant and contaminant from the site. Among the hundreds of hyperaccumulators recently identified are the brake fern (*Pteris vittata*), an effective As accumulator, and also the basket willow (*Salix viminalis*), which effectively sequesters both Zn and Cd. Biodegradation is a related *in situ* remediation strategy that utilises native microbial populations to degrade multiple organic contaminants. This microbial degradation is optimal within aerobic rhizosphere communities at circumneutral pH, particularly in the presence of abundant nutrients, which may be augmented with external sources to enhance microbial activity.

The more costly *ex situ* techniques are generally applied only to the most high-value contaminated sites. The simplest of the *ex situ* remediation strategies involves direct excavation of the contaminated material followed by landfill disposal. Soil washing (*see* **Soils:** Modern) is more complex, involving excavation followed by mechanical screening to obtain coarse ($>50\,\mu$m) and fine ($<50\,\mu$m) fractions which are subsequently treated with a series of surfactants and chelates to remove both organic and inorganic contaminants, including radionuclides. The aim of soil washing is to concentrate the pollutants within the fine fraction, which is then treated with a solidification/stabilizing agent (e.g., Portland cement, lime, fly ash) prior to landfill disposal.

Table 2 Summary of *in situ* and *ex situ* environmental restoration techniques

Method	*Contaminant*	*Processes*	*Limitations*
In situ			
Lime addition	Metals, radionuclides	pH increase causing sorption, precipitation	Ineffective for oxyanions
Smectite, zeolite, apatite addition	Metals, radionuclides	Ion exchange, sorption, precipitation	Selective, short term remediation
Phytoremediation	Metals, organics	Phytoaccumulation, phytodegradation	Unsuitable for highly contaminated sites
Biodegradation	Organics	Microbial degradation	Long term remediation
Volatilization	Volatile organics	Evaporative loss of volatile pollutants	Contaminant must be volatile
Ex situ			
Excavation and disposal	All	Removal and disposal	Costly, risk of pollutant dispersal
Soil washing	Metals, organics, radionuclides	Excavation, leaching of contaminants	Costly
Solidification/stabilization	Metals, radionuclides	Addition of stabilising agent	Volatile organics not immobilized

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Clay Minerals**. **Environmental Geology**. **Geochemical Exploration**. **Minerals:** Arsenates; Carbonates; Chromates; Sulphides; Zeolites. **Soils:** Modern. **Weathering**.

Further Reading

Adriaens P, Gruden C, and McCormick ML (2004) Biogeo chemistry of halogenated hydrocarbons. In: Lollar BS (ed.) *Environmental Geochemistry,* vol. 9, pp. 511 539. *Treatise on Geochemistry* Holland HD and Turekian KK (eds.). Oxford: Elsevier Pergamon.

Adriano DC (1986) *Trace Elements in the Terrestrial Envir onment.* New York: Springer Verlag.

Alexander M (1999) *Biodegradation and Bioremediation,* 2nd edn. London: Academic Press.

Alloway BJ (1995) *Heavy Metals in Soils,* 2nd edn. London: Blackie Academic and Professional.

Blowes RJ, Ptacek CJ, Jambor JL, and Weisener CG (2004) The geochemistry of acid mine drainage. In: Lollar BS (ed.) *Environmental Geochemistry,* vol. 9, pp. 149 204. *Treatise on Geochemistry* Holland HD and Turekian KK (eds.). Oxford: Elsevier Pergamon.

Bowen HJM (1979) *Environmental Chemistry of the Elem ents.* London: Academic Press.

Cozzarelli IM and Baehr AL (2004) Volatile fuel hydrocar bons and MTBE in the environment. In: Lollar BS (ed.) *Environmental Geochemistry,* vol. 9, pp. 435 474. *Trea tise on Geochemistry* Holland HD and Turekian KK (eds.). Oxford: Elsevier Pergamon.

Hayes KF and Traina SJ (1998) Metal ion speciation and its significance in ecosystem health. In: Huang PM (ed.) *Soil Chemistry and Ecosystem Health,* pp. 45 84. Madison, USA: Soil Science Society of America.

Schwarzenbach RP, Gschwend PM, and Imboden DM (1993) *Environmental Organic Chemistry.* New York: Wiley.

Sposito G (1996) *The Environmental Chemistry of Alumi num,* 2nd edn. London: CRC Press.

ENVIRONMENTAL GEOLOGY

P Doyle, University College London, London, UK

Introduction

Environmental geology has grown in stature as a discipline over the past 40 years, as considerations of economic geology have moved away from the simple exploitation of the late nineteenth and early twentieth centuries. Today, environmental geology is a broad area of geological endeavour and a major industry in its own right.

Defining Environmental Geology

Environmental geology may be defined as the interaction of humans with their – fundamentally geological – environment. The environment can be considered to consist of both the constituents of the Earth itself (rocks, sediments, and fluids) and its surface and the processes that operate to change it through time.

Environmental geology is a subset of environmental science, which is the study of the interaction of humans with all aspects of their environment – physical, atmospheric, and biological – and is linked directly with engineering geology (*see* **Engineering Geology:** Overview). This definition clearly indicates that it is the introduction of the human element to the equation that defines the concept of the environmental sciences – and, therefore, environmental geology – and it is a consideration of both the debits (impacts) and credits (benefits) of our existence. Environmental science is a way of managing our existence so as to maximize human success while minimizing the negative aspects. At the heart of environmental geology, as with all environmental sciences, is the concept of sustainable management – working with natural systems to sustain development but not at an unacceptable environmental cost.

Scope of Environmental Geology

It has been estimated that around 50% of the Earth's population live in urban centres, centres that cover just 1% of the Earth's surface. Given that environmental geology is the interaction of humans with their environment and given that the majority of humans live in cities, it follows that environmental geology can be considered as primarily an urban issue, with the most challenging problems occurring within the immediate hinterland of urban centres.

Environmental geology has been defined as an urban concept, a 'machine' that balances inputs, outputs and maintenance, most if not all of which have a geological component (**Figure 1**). The inputs include water (derived locally or from more distant locations, but feeding into the city), raw materials (in the form of

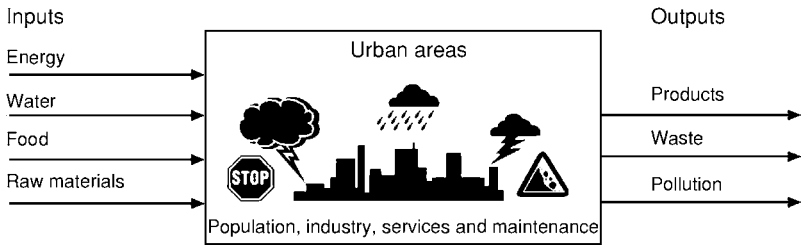


Figure 1 The urban machine. (Reproduced with permission from Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment.* Chichester: John Wiley.)

mineral resources for industry and construction), food resources (both locally produced and transported), and energy (the result in many cases of mineral resources such as coal, gas, and uranium). Outputs from the 'machine' include products from industry, wastes (in the form of worn-out materials, by-products of industry, and day-to-day wastes from domestic, industrial, and commercial sources), and pollution from poor waste-management strategies. The urban machine needs constant maintenance to replace its infrastructure and foundations, and to protect it from natural hazards.

The demands of the machine drive our demands on the environment, with the need for exploration and exploitation of mineral, water, and soil resources to provide the inputs; the need to combat environmental stress and natural hazards in protecting the workings of the machine, and the requirement to deal with the outputs, particularly wastes and their potential to pollute if not handled with appropriate sensitivity. All these aspects (and more) fall within the broad scope of environmental geology, and they can be distilled into the following themes.

1. Geology of resource management, including exploration and exploitation of resources (e.g. fuels, industrial minerals, and water) and the mitigation and limitation of associated adverse environmental impacts.
2. Geology of the built environment, particularly the constraints that ground conditions place on development.
3. Geology of waste management, specifically the disposal of wastes in the physical environment.
4. Geology of natural hazards.

Geology of Resource Management

Resources may be defined in the geological context as naturally occurring solids, liquids, or gases that are known or thought to exist in or on the Earth's crust in concentrations that make extraction economically feasible either at present or at some time in the future. This definition includes those geological materials that can be extracted with currently available technology (reserves) as well as those that are thought to exist but that will require further technological development to remove them.

In the short term, resource-management options focus on reserves, and environmental considerations have more power than at any other time in the past to reduce available reserves by altering the economics of their exploitation. Long-term planning must, of course, be based on the probability of new resources being discovered. The size of the available reserves relative to the total resources can be altered by several factors, such as commodity price, exploration, increasing the extent of the known resource, technological developments, and changes in regulation.

Economic Mineral Resources

Economic mineral resources are varied and can be defined in their broadest sense as any geological material that is of commercial value to society. This broad definition includes such diverse materials as fuels (e.g. coal, gas, uranium), construction materials, industrial minerals, metals, and precious minerals. Extraction of these materials varies from bulk extraction to specialist mining, and, although the basic concepts of exploration and exploitation may have many parallels, the environmental impacts posed by each one may well differ strongly.

Environmental Impacts of Mineral Extraction

The environmental impacts of mineral extraction vary according to the type of mineral and the extent of its deposit, with impacts varying throughout the working life of a mine or quarry, and with the issues often continuing long after the deposits worked are no longer economically viable. Typical issues are aspects of mine operation, mining subsidence, tackling mine wastes, and quarry or mine restoration.

There are a wide range of issues associated with mining. Quarries are troubled by blast noise and vibration, which can lead to increased rock-fall and

landslip activity. In deep mines, the emission of methane and other gases is a problem, while wastewaters may be charged with iron and may be acidified by the breakdown of pyrite and other sulphides. Acid mine drainage is a growing problem in many disused metal mines and has adverse affects on groundwater supplies and water bodies, such as estuaries, rivers, and lakes.

Mining subsidence is a result of the removal of geological materials underground, creating a void space, which subsequently collapses (Figure 2). Similar situations occur when voids are created for purposes other than mineral extraction. Longwall mining is associated with deep coal extraction and generally involves the planned subsidence of a relatively large area. Pillar-and-stall mining, which dates back centuries, can lead to the differential collapse of the pillars, giving rise to graben-like failure structures. In other cases, removal of deep salt deposits through the process known as brining – the pumping out of brine-charged waters – can lead to severe subsidence. Large-scale subsidence is often difficult to manage, and the possibility of constructing buildings over an unknown area of mining activity is a very live issue, even where the mining is known to date back to antiquity, as in Rome. Raft foundations, grouting, deep piling, and excavation and backfill are all possibilities, depending on the depth and extent of the problem.

Mine wastes are produced at two stages during the mineral-extraction process: during mining, when waste rock or spoil is produced; and during further processing of the materials extracted, which creates a further set of mineral wastes, generally much finer than simple spoil and referred to as tailings. Spoil is of variable grain size and generally accumulates in open tips close to the mine workings. Tailings produced by mineral processing tend to be much richer in minerals and more uniform and of a finer grain size owing to milling. In some cases these materials have a high toxicity, as hazardous chemicals are used in the separation of mineral particles. Tailings are also more likely to be distributed by weathering and erosion, raising the consequent possibility of hazardous air-borne particulate matter. The long-term stability of mine wastes is also of concern, as with the South Welsh Aberfan disaster in 1966, where 111 000 m^3 of debris moved down a 13° slope, enveloping a school with tragic loss of life.

Dealing with waste tips is a matter of concern for environmental geologists, and the solution depends very much on the nature of the material. In the UK, tax on aggregate extraction led to the reuse of some tax-exempt spoil, particularly slate and shale, as poor-quality aggregate. Other mineral spoil may be reused for rock fill or inert landfill or even in dealing with mine subsidence and quarry-restoration projects. Tailings, on the other hand, given the potentially high level of toxicity, have to be treated with caution. Wet tailings have to be ponded, treated, and removed, while dry tailings have to be carefully monitored so that water passing through the wastes does not lead to

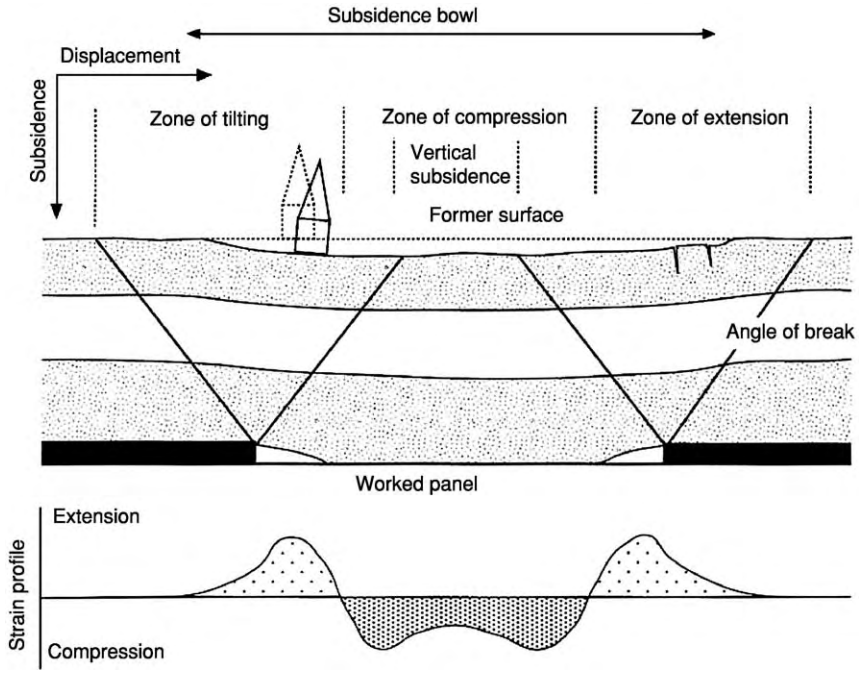


Figure 2 Subsidence associated with longwall coal mining.

groundwater pollution. Restoration and stabilization of wastes is essential, but this can be a tall order when vegetation that could be useful in binding surfaces is inhibited by the toxicity of the waste material itself.

Quarry and mine restoration is a major task. For mines, it is important first of all to ensure the safety of the underground workings themselves and to monitor such issues as the build-up of gases and acid mine drainage. In other cases, the mine workings may present a hazard in the form of future subsidence. Pit-head machinery and processing works, which may have used chemical processes requiring specialist attention, will also be an issue. Finally, the mine wastes themselves must be dealt with. For quarries, restoration issues reflect the nature of the materials quarried, with production-blasted hard-rock quarry faces representing an unsafe and unstable option. Restoration depends very much on the projected end use of the quarry; if left as a void, then faces will require engineering attention. This is particularly so given that many quarries are now being used to house industry or commercial concerns. In other cases, the void will be filled, most often as landfill.

Water Resources

Water is a vital geological resource that sustains life, yet most of the Earth's water is unavailable, with 98% being in the oceans and 1.6% in ice-sheets, leaving only 0.4% for drinking. Suitable drinking-water resources are distributed unevenly across the globe, and this, together with the predictions of environmental change, means that the search for drinking-water supplies is a major preoccupation of environmental geologists. Water-resource management is essentially a combination of three factors: acquisition, redistribution, and the treatment and disposal of wastewaters. The potential to pollute existing supplies through inadequate waste-disposal strategies is also a major issue, and this will be dealt with later when considering waste management.

Acquisition of water is primarily a process of extraction from groundwaters, which are controlled by rainfall and the porosity and permeability of the rocks present in a given area (**Figure 3**). These can be highly variable, and, even within a small area such as the UK, the percentage of water derived from this source is variable. Where groundwater is not an option, reliance on surface waters – rivers, lakes, and reservoirs – becomes much more significant, with desalination of seawater also possible.

The redistribution of supplies is an enormous undertaking and has lead to major international disputes when some countries have sought to dam rivers to the detriment of others downstream. Such international tensions are unlikely to recede in the future and could lead to 'water wars' if diplomacy breaks down. The creation of dams to contain waters and generate hydroelectric power has been an attractive but often highly controversial approach in some water-poor countries. However, inadequate site assessment has led to the failure of some dam sites, with leakage of waters or failure of the dam structure itself, so the high-cost deployment of such schemes requires a complete understanding and assessment of the issues involved.

Soil Resources

Together with water, soils are a fundamental geological resource, allowing us to grow sufficient crops to supply the food required to sustain the planet's growing population. The distribution and formation of soils across

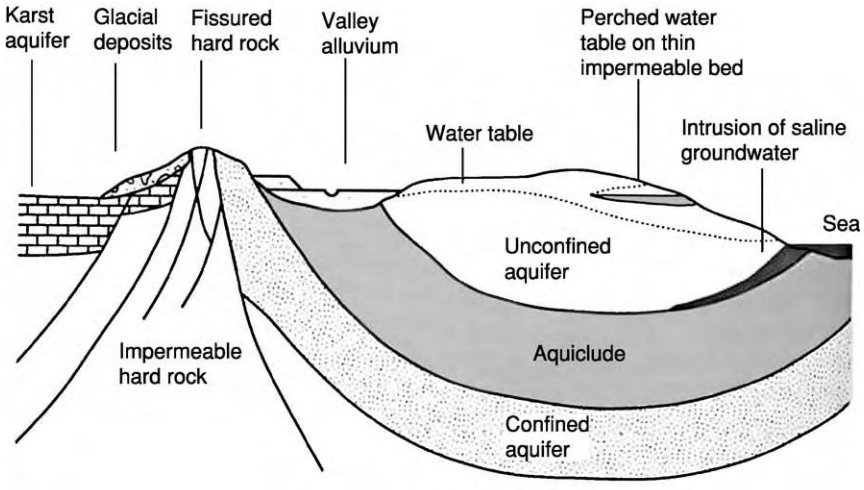


Figure 3 Schematic diagram of the main water bearing units. (Reproduced with permission from Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment.* Chichester: John Wiley.)

the world is contingent on climate, rainfall, biological activity, topography, and the underlying geology. Soils vary considerably, with an infinite variety – there are at least 19 000 different types in the USA alone – and widely differing classification schemes across the world (*see* **Soils:** Modern). The potential to erode and denude soils depends on the intensity of use, the ground cover, and so on. Soil conservation is therefore a major issue and can vary with land-use changes, new technologies, and colonization.

In the UK, raising the awareness of soils and the need to conserve them has led to the construction of five basic principles: that soils are an essential part of life support; that soils should enjoy the same level of protection as water and air; that integrated environmental management should include soils; that contaminated soils should be remediated; and that contamination should be avoided. These issues obviously have a global relevance.

Aesthetic and Scientific Resources

Conservation of aesthetic and scientific resources involves the conservation of areas of landscape and/or specific geological sites – known as geodiversity – for future generations (*see* **Geological Conservation**). The concept is based on four basic convictions: that they should be conserved for their own sake; that they form the basis for exploitation; that they form the basis for research and training; and that they have aesthetic and/or cultural value. This approach is growing in stature across the world, leading to a wider awareness of the value of geological features.

Geology of the Built Environment

The built environment – the towns and cities in which the vast majority of people live – is the central pillar of the urban machine, and, in many cases, the cause of many environmental issues. Built from geological resources – geomaterials – towns and cities are in constant need of renovation, renewal, and development, with the consequent need for greater exploration for aggregates, stone, and other materials (Figure 4). Engineering geology is an important factor in the construction of the built environment, particularly in the provision of sound foundations. Many of the older cities of the world are founded on ancient settlement sites, and this in turn creates its own issues of long-term stability.

Geomaterials

Construction materials, known as geomaterials, are quarried and mined from a wide variety of settings. The pressure to discover new resources is increasing, as the pressure to house the world's burgeoning population also increases. Stone and clay for adobe bricks have been the main building resources of the world since antiquity, and concrete has a long and venerable history. These materials are still in demand, and quarrying for them creates much the same environmental issues as any other type of extractive industry (*see* **Aggregates**).

Another important role for environmental geologists is in assessing the appropriateness of geomaterials for the job for which they are intended. This might include, for example: a durable, structurally strong, easily worked, and attractive stone for construction; large relatively uniform blocks of stone that can take a high degree of pounding without breaking down for use as armourstone blocks in coastal defences; or the provision of aggregates for roads that do not polish easily and can take a fair amount of crushing from large vehicles. Durability of stone and the effects of stone cleaning are major issues in some of the major cities of the world, needing constant monitoring (*see* **Building Stone**).

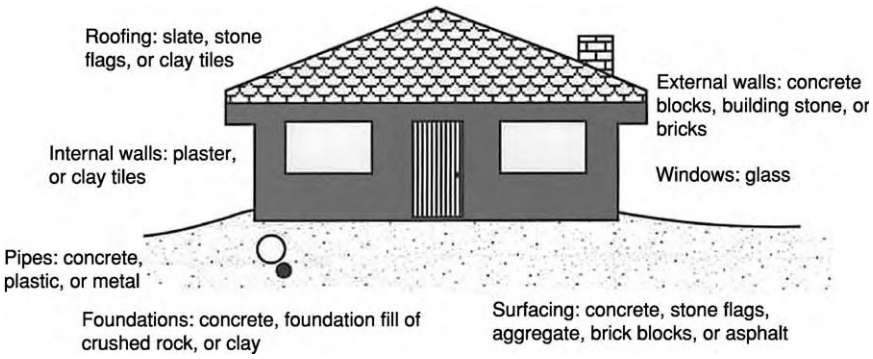


Figure 4 The main geomaterials and their uses. (Reproduced with permission from Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment.* Chichester: John Wiley.)

Engineering Geology

Major construction works depend upon sound site investigation, which documents the underlying geological setting and the strengths and behaviours of the rocks and engineering sols and recognizes hazardous ground conditions. In the UK alone, over 30% of all major civil engineering projects are delayed owing to poor ground conditions. It is essential to carry out thorough site investigations, using a range of geological techniques, including mapping and geophysical surveys, as well as committing significant time to material testing. Consideration of climate is a major issue, as in Hong Kong, where the failure of tower blocks has been attributed to the failure of thick tropical soils overlying crystalline basement, coupled with high rainfall (*see* **Engineering Geology:** Overview).

Geology of Waste Management

The urban machine creates an immense amount of waste, which may be defined as any substance that constitutes a scrap material, effluent, or unwanted surplus or any substance or article that requires disposal because it is worn-out or contaminated in some way. Although there is increasing emphasis on waste minimization and recycling, it is accepted that there will always be waste and that a strategy for its disposal will be required.

Wastes can conveniently be divided on the basis of whether they are managed (i.e. there is a management strategy in place to deal with their disposal that is in line with current practice and legislation) or unmanaged (i.e. there is no management strategy). Strategies for dealing with wastes have been based largely on two concepts: dilute and disperse, or concentrate and contain (Figure 5). The failure of these strategies ultimately leads to pollution, which is a by-product of inadequate waste-management strategies.

The dilute-and-disperse strategy is now largely discredited, because it relies on the potential of a given environmental setting to dilute the waste materials and then to disperse them. This has been, and continues to be, an issue for the dispersal of effluents in the sea and of gases in the atmosphere, as these natural systems are becoming choked with inadequately dispersed wastes. Concentrate and contain is now seen as the most viable option, with wastes being concentrated, compacted, or shredded where possible, before being placed in a safe totally enclosed (or 'sanitary') facility. At one level this entails the adequate preparation of a landfill site; at another, the containment of radioactive materials in an appropriate repository in perpetuity.

Contaminated Land

Contaminated land is any area that has been contaminated by its past industrial use or by the disposal of wastes. Contamination is therefore an inevitable effect of industrialization, and of waste disposal, as it involves the introduction of materials not naturally present at a given location. Contaminated land is a big issue in developed nations, with the decline in heavy industry and the targeting of former sites of industry as valuable real estate in inner-city locations. Increasingly, there is a need for adequate assessment of a site – assessing the type and level of contamination – before developing a clean-up strategy.

Sanitary Landfill

Employing the principle of concentrate and contain, sanitary landfills (Figure 5) ensure that appropriate voids are chosen for landfilling, appropriate containment strategies are employed in developing the landfill (using liners, separating the void into cells divided by bunds, and capping the site securely), the void is filled in an economical and organized manner, and

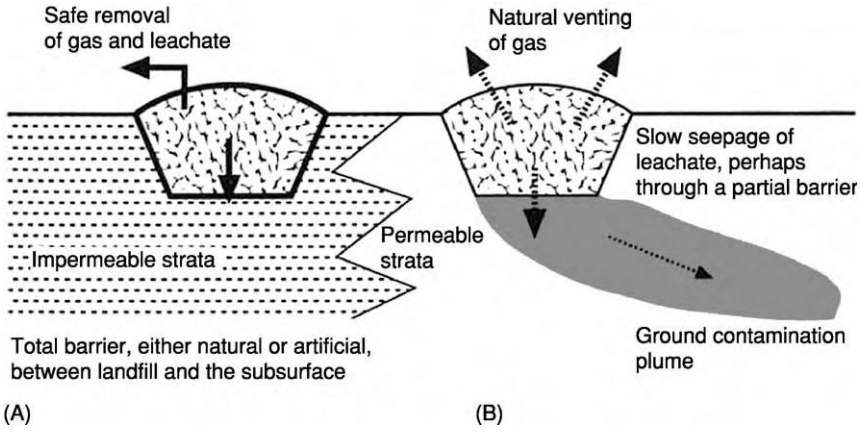


Figure 5 Landfill options: (A) total containment, and (B) slow dispersal. (Reproduced with permission from Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment.* Chichester: John Wiley.)

the by-products of the decay of putrescible wastes (i.e. landfill gas and leachates) are removed safely. The decomposition of wastes by bacteria takes place in three phases: first, aerobic bacteria break down the large organic molecules; second, anaerobic bacteria create simple compounds, such as hydrogen, ammonia, carbon dioxide, and organic acids; and, finally, the methane-rich landfill gas is created. In practice, this means that in the early years of a landfill the amount of gas produced is relatively small, rising to a maximum in around 20 years. Uncontrolled migration of this gas can be lethal, as in some cases the results have been explosive, with gases migrating through suitable geological pathways, as happened in 1986 at Loscoe in northern England. Leachates are dealt with by ponding and removal, in most cases, and then treated.

Deep Repositories for Nuclear Wastes

The disposal of nuclear radioactive wastes is a difficult problem, given the 'geological' time-scales of the half-lives of the isotopes of uranium used in the nuclear-power industry. One option is to concentrate wastes in a secure surface repository, which can be monitored to ensure that the multiple walls of the containment are secure. This obviously requires a long-term commitment. Over the last few decades, other options have been considered, including using a uniform densely crystalline deep-crustal rock body as a natural repository, which could be sealed in perpetuity (Figure 6). Lithologies considered have included fine-grained tuffs, basalts, and evaporates, and detailed investigations in the UK and the USA of the feasibility of sites in Cumbria and Texas, respectively, have been carried out. However, technical difficulties and political issues have stalled these geological solutions.

Geology of Natural Hazards

Natural hazards occur on a regular basis, cause disruption and destruction, and annually lead to large losses of life under tragic circumstances (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards). Geological hazards include high-magnitude low-frequency events (earthquakes, volcanoes, and tsunamis) and low-magnitude high-frequency events

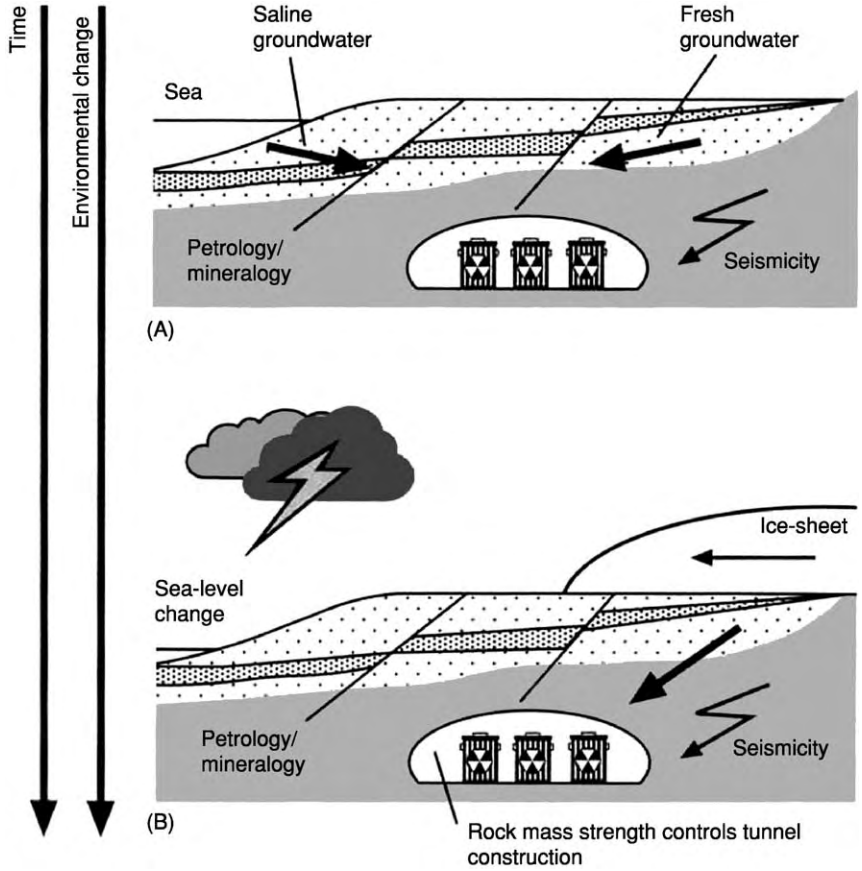


Figure 6 Deep repository option for nuclear wastes, illustrating (A) the current geology, and (B) the geology at some point in the future with the influx of fresh glacial meltwaters (arrow). (Reproduced with permission from Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment.* Chichester: John Wiley.)

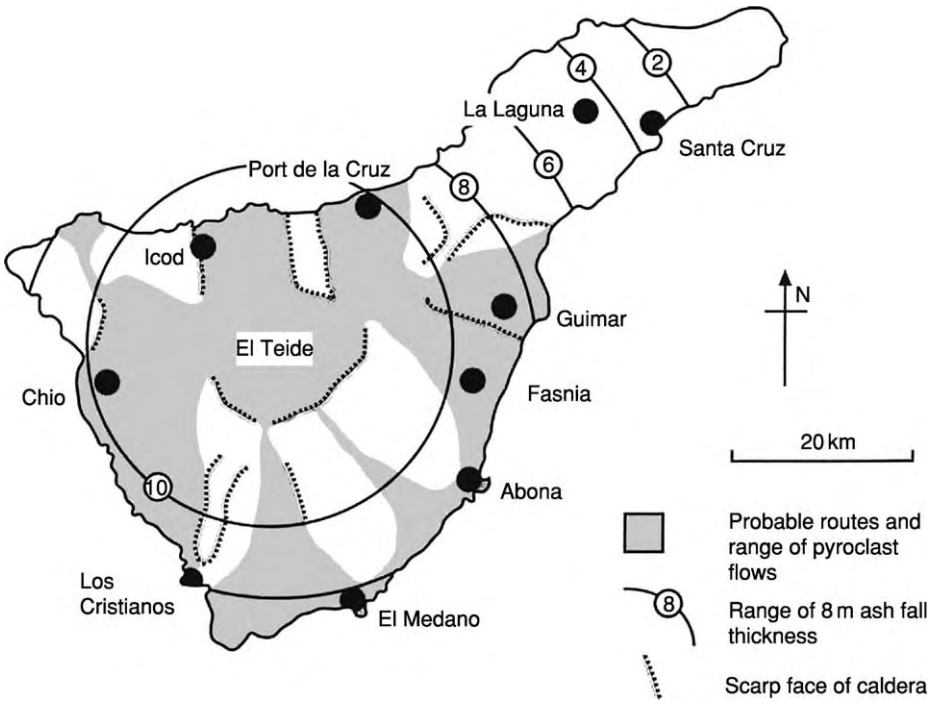


Figure 7 Volcanic hazard map of Tenerife, Canary Islands.

(such as soil, fluvial, and coastal erosion). Dealing with both types of event severely tests the local civil-defence infrastructures, and, with the frequency of disasters apparently on the increase – with attendant casualties and economic costs – planning for disaster management is a major preoccupation of local authorities.

There are three broad hazard types: atmospheric hazards, caused by tropical storms, hurricanes, and droughts; exogenic hazards, such as flooding, coastal erosion, and mass movement; and endogenic hazards, resulting from internal Earth processes, such as volcanoes and earthquakes. In many cases there can be considerable overlap between these events, as one can trigger another. In profiling natural hazards, it is important to gather data on the magnitude and frequency (including temporal spacing) of the events, the duration and speed of onset of a typical event, and the extent and dispersion of the effects. These issues have to be taken into account in civil-defence planning strategies, which will identify the risk to life, property, and the economy.

Risk assessment is an important task and will in many cases involve a predictive element, although prediction of earthquakes and volcanic eruptions is a tricky business, without a simple solution. Hazard mapping is a significant technique and had been used to good effect in identifying patterns of eruption in populated volcanic regions (Figure 7). This, together with a numerical assessment of the vulnerability of a given settlement, are important tools in the armoury of an environmental geologist.

See Also

Building Stone. **Clays, Economic Uses**. **Engineering Geology:** Overview; Natural and Anthropogenic Geohazards. **Geological Conservation**. **Geological Engineering**. **Quarrying**. **Urban Geology**.

Further Reading

Bailleul TA (1987) Disposal of high level nuclear waste in America. *Bulletin of the Association of Engineering Geologists* 24: 207 216.

Bell FG (1992) Salt mining and associated subsidence in mid Cheshire, England, and its influence on planning. *Bulletin of the Association of Engineering Geologists* 29: 371 386.

Bell FG (1993) *Engineering Geology*. Oxford: Blackwell.

Bell FG (1996) Dereliction: colliery spoil heaps and their rehabilitation. *Environmental and Engineering Geoscience* 2: 85 96.

Bell FG, Duane MJ, Bell AW, and Hytiris N (1996) Contaminated land: the British position and some case histories. *Environmental and Engineering Geoscience* 2: 355 368.

Bennett MR and Doyle P (1997) *Environmental Geology: Geology and the Human Environment*. Chichester: John Wiley.

Booth B (1979) Assessing volcanic risk. *Journal of the Geological Society of London* 136: 331 340.
Bryant EA (1991) *Natural Hazards*. Cambridge: Cambridge University Press.
Culshaw MG, Bell FG, Cripps JC, and O'Hara M (eds.) (1987) *Planning and Engineering Geology*. London: Geological Society.
Doyle P, Barton P, Rosenbaum MR, Vandewalle J, and Jacobs K (2002) Geoenvironmental implications of the underground war in Flanders, 1914 1918. *Environmental Geology* 43: 57 71.
Fookes PG and Poole AB (1981) Some preliminary considerations on the selection and durability of rock and concrete materials for breakwaters and coastal protection works. *Quarterly Journal of Engineering Geology* 14: 97 128.
Gray M (2004) *Geodiversity: valuing and conserving abiotic nature*. Chichester: John Wiley.
Gray RE and Bruhn RW (1984) Coal mine subsidence: eastern United States. *Geological Society of America Reviews in Engineering Geology* 6: 123 149.
Joseph JB (2004) Perception or reality waste, landfill and the environment. *Geology Today* 20: 107 112.
Lumsden GI (ed.) (1992) *Geology and the Environment in Western Europe*. Oxford: Oxford University Press.
McFeat Smith I, Workman DR, Burnett AD, and Chau EPY (1989) Geology of Hong Kong. *Bulletin of the Association of Engineering Geologists* 26: 23 107.
Prentice JE (1990) *Geology of Construction Materials*. London: Chapman & Hall.
Price M (1985) *Introducing Groundwater*. London: Allen & Unwin.
Smith K (1996) *Environmental Hazards: Assessing Risk and Reducing Disaster*, 2nd edn. London: Routledge.
Thomas RG (1989) Geology of Rome, Italy. *Bulletin of the Association of Engineering Geologists* 26: 415 476.
Waltham AC (1994) *Foundations of Engineering Geology*. Glasgow: Blackie.
Williams GM and Aitkenhead N (1991) Lessons from Loscoe: the uncontrolled migration of landfill gas. *Quarterly Journal of Engineering Geology* 24: 191 207.
Woodcock NH (1994) *Geology and Environment in Britain and Ireland*. London: UCL Press.

EROSION

See **SEDIMENTARY PROCESSES: Erosional Sedimentary Structures; Aeolian Processes; Fluxes and Budgets**

EUROPE

Contents

East European Craton

S V Bogdanova and R Gorbatschev, Lund University, Lund, Sweden
R G Garetsky, Institute of Geological Sciences, Minsk, Belarus

Introduction

The East European Craton is the coherent mass of Precambrian continental crust in the north-eastern half of Europe. Across a suture zone between the North Sea and the Black Sea, that crust meets the younger, thinner, warmer, and more mobile crust of western and southern Europe.

The overall area of the East European Craton is more than 6.7 million km^2, including the shelf seas.

Within the East European Craton, Precambrian crystalline crust is exposed in the Baltic (also Fennoscandian) and Ukrainian Shields as well as in minor areas in Belarus and the Voronezh Massif of south-western Russia. Elsewhere, the craton is covered by the Late Proterozoic and Phanerozoic sedimentary deposits of the Russian Platform (Figure 1).

The present state of the East European Craton is largely controlled by structures dating back to the time of its formation by the successive collision of three large, once independent, crustal segments – Fennoscandia, Sarmatia, and Volgo-Uralia – in the Palaeoproterozoic, *ca*. 2.1–1.7 Ga ago. This view was first proposed in 1993. Until then, the East European Craton was regarded as a rather uniform region of numerous minor 'blocks' of Archaean crust set in a matrix of Proterozoic folded belts.

Margins and Borders

Most of the margins of the East European Craton are characterized by the presence of younger mobile belts (Figure 1). In the north-west, the Scandinavian Caledonides (*see* **Europe:** Scandinavian Caledonides (with Greenland)) form a *ca*. 1800 km long strongly eroded, but still up to 15-km thick, pile of large thrust sheets built up of Early Palaeozoic and Precambrian rocks. These nappes derive from source areas far to the present west and were thrust atop the crystalline basement around 450–400 Ma ago. Autochthonous basement fabrics are still perceptible throughout the Caledonides, while geophysical and palaeogeographical data indicate that the Pre-Caledonian margin of the East European Craton may be as much as 400 km off the Atlantic coast of Scandinavia.

Along the north-eastern margin of the East European Craton, the Timanide Belt extends between the Urals and northernmost Norway (*see* **Europe:** Timanides of Northern Russia). It features initial sedimentation along a passive continental margin, followed by compression, metamorphism, and igneous activity due to subduction and collision 650–600 Ma ago.

The eastern limit of the East European Craton is the Late Palaeozoic Uralide Orogen, which represents the zone where ancient Europe collided with Asian terranes 350–300 Ma ago (*see* **Europe:** The Urals). Farther south-west, the margin of the East

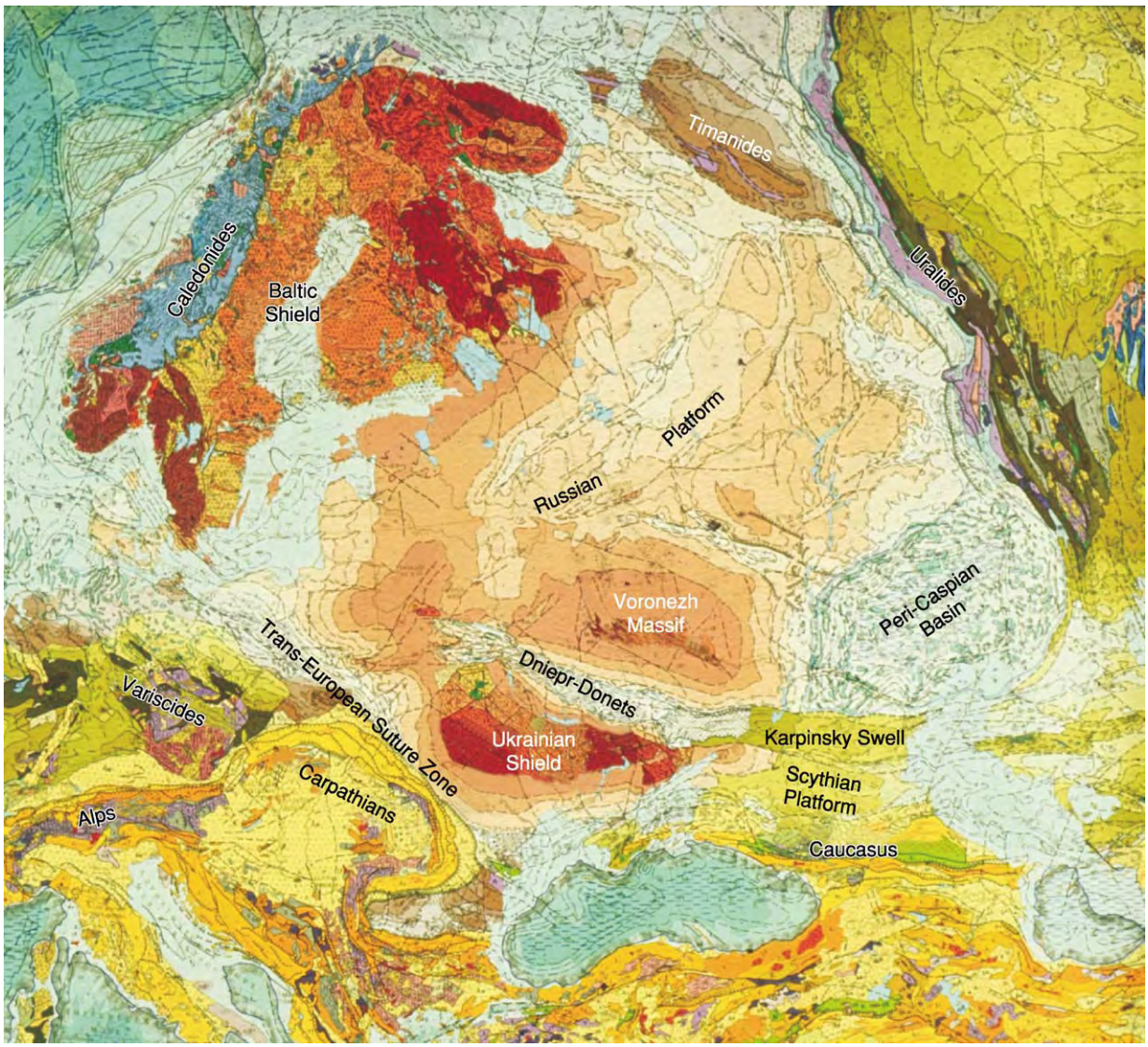


Figure 1 The East European Craton in its crustal setting between central Europe and western Siberia. The craton area is marked in reddish brown colours. The fainter the tint, the thicker the sedimentary cover. The areas of outcrop in the Baltic and Ukrainian Shields and the Voronezh Massif can be clearly seen. The cover is also thin in Belarus, but only in a minor uplift is the craton actually exposed. It can also be seen that the basement is more deeply buried in the eastern part of the Russian Platform than in the west. Between the eastern region and the basin beneath the southern Baltic Sea is the so called 'Scythian Rampart' of the early geologists, which connects the Baltic and Ukrainian Shields. The cover is thickest in the Peri Caspian Basin, where the mottled patterns denote salt dome tectonics. Among the large rifts and aulacogens characteristic of the East European Craton, the originally Proterozoic but subsequently rejuven ated central Russian and Pachelma systems can be clearly recognized. The latter is immediately north north east of the Voronezh Massif. The rift systems to the west of the Urals and in the south western foreland of the Timanides can also be seen. Prominent in the south is the Devonian Dniepr Donets Aulacogen and its inverted eastward continuation (the so called 'Karpinsky Swell') between the Scythian Platform and the Peri Caspian Basin and onwards into Asia. This figure is part of the 'Carte Tectonique de l'Europe et des Régions Avoicinantes,' scale 1 to 10 min. IUGS Commission for the Geological Map of the World, 1975 (by courtesy of the CGMW.)

European Craton was developed in the Neoproterozoic and Early Cambrian due to collision between the Craton and part of Gondwana. That margin was subsequently complicated by the uplifted southeastern continuation of the Devonian Dniepr–Donets Aulacogen (the Karpinsky Swell in Figure 1).

The southern edge of the East European Craton is outlined by the Alpine mountain belts of the Crimea and Caucasus. In that region, the Scythian Platform marks a slab of East European cratonic crust involved in the Alpine–Mediterranean orogenic process.

In the south-west, the Trans-European Suture Zone is the limit of the East European Craton. Across that boundary, the Craton abuts terranes successively formed during the Caledonian, Variscan, and Alpine–Mediterranean orogenies. Along the central course of the Trans-European Suture Zone, palaeontological data indicate that some terranes in Poland represent detached slices of the Craton. In northern Germany, a wedge of East European cratonic crust has been traced seismically far to the south-west of the Trans-European Suture Zone.

Crustal Thickness and Magnetic and Gravity Fields

The crust of the East European Craton is mostly around 35 km to somewhat more than 40 km thick. Moho depths of up to 50 km occur particularly in a wide central area near Moscow, in the south-west, and in the Ukrainian and Baltic Shields. Local maxima may even exceed 60 km. Under some of the Archaean parts of the East European Craton, cratonic roots in the lithospheric mantle reach down to 200–250 km.

The most pronounced Moho uplift, to a level of around 30 km, is associated with the northern part of the Peri-Caspian Basin, where a thick Proterozoic to Phanerozoic cover is also present. In consequence, the thickness of the Precambrian crystalline crust is only 10–15 km in that region.

At the borders of the East European Craton, Moho depths increase along the Uralides and the Caucasus collisional belts as well as in the central and southern parts of the Trans-European Suture Zone. Further north, continental crust thins markedly towards the North Atlantic Ocean.

In detail, steep gradients of Moho depth commonly follow one-time collisional and accretionary plate and terrane boundaries and, in general, the boundaries between Archaean and Proterozoic crust. However, in places, the original continental crust has been thickened by later mafic underplating or thinned by extension and magmatism.

The East European Craton differs from the neighbouring parts of Europe and western Siberia in featuring numerous belts of strong magnetic anomalies. These outline the boundaries of the different crustal units as well as later rifts and some major belts of granitic and high-grade metamorphic rocks (**Figure 2**). The patterns of the gravity field are similar, but strongly negative anomalies occur only along the Scandinavian Caledonides and just outside the limits of the craton along the Carpathian, Crimean, and Caucasian Alpine–Mediterranean mountains.

Morphology, Topography, and Sedimentary Cover

Morphologically, most of the East European Craton forms a vast low-lying plain. Some small-scale broken topography exists in the shield areas, particularly in the recently glaciated Baltic Shield. The largest flats, in contrast, extend towards the south-east, where the Peri-Caspian depression is largely below global sea-level (**Figure 1**). Pronouncedly mountainous areas occur only in western and northern Scandinavia, where the Caledonides and their Precambrian basement were uplifted and topographically rejuvenated during the opening of the North Atlantic.

The cover of the Russian Platform mostly ranges between some tens of metres and 2 km in thickness (**Figures 1 and 3A**). Several rifts and basins, however, contain sedimentary piles 3–5 km thick, while the Dniepr–Donets Aulacogen in the Ukraine and south-western Russia has a fill exceeding 15 km in thickness (**Figure 3B**). A general depression of the basement–cover boundary occurs towards the marginal mountain belts of the Timanides, Uralides, and Caucasus, the greatest cover thickness of approximately 20 km being reached in the area to the north of the Caspian Sea.

The covered parts of the East European Craton comprise several large basins of sedimentation, e.g. the Moscow, Baltic, and Peri-Caspian basins. These were largely formed in response to recurrent cycles of rifting, subsidence, and compression. Maxima of basin formation and filling occurred during the Riphean (Meso- to Neoproterozoic), the Early Vendian (terminal Neoproterozoic), the Late Cambian-Ordovician, the Middle-to-Late Devonian, the Carboniferous–Permian transition, and the Triassic (**Figures 3–5**). Generally, these maxima were related to orogenies, major plate-tectonic movements, rotations of the East European Craton, and its interaction with other proto-continents such as Laurentia, Greenland, and Avalonia.

Large rifts and aulacogens are characteristic elements of the East European Craton (**Figure 6**). Many of the oldest, i.e. of Meso- to Neoproterozoic age, are associated with the system of Palaeoproterozoic sutures that arose as a result of the formation of the craton from several colliding crustal segments. To this group belong the Pachelma, Volyn-Orsha, and Central Russian rifts and aulacogens (**Figure 6**). Other rifts of that age follow age-province boundaries, while several Neoproterozoic troughs in the Peri-Urals region were developed on a Mesoproterozoic passive continental margin. Unlike the Precambrian rifts, many Phanerozoic rifts, for instance the Dniepr-Donets Aulacogen, cut sharply across the grain of the crust.

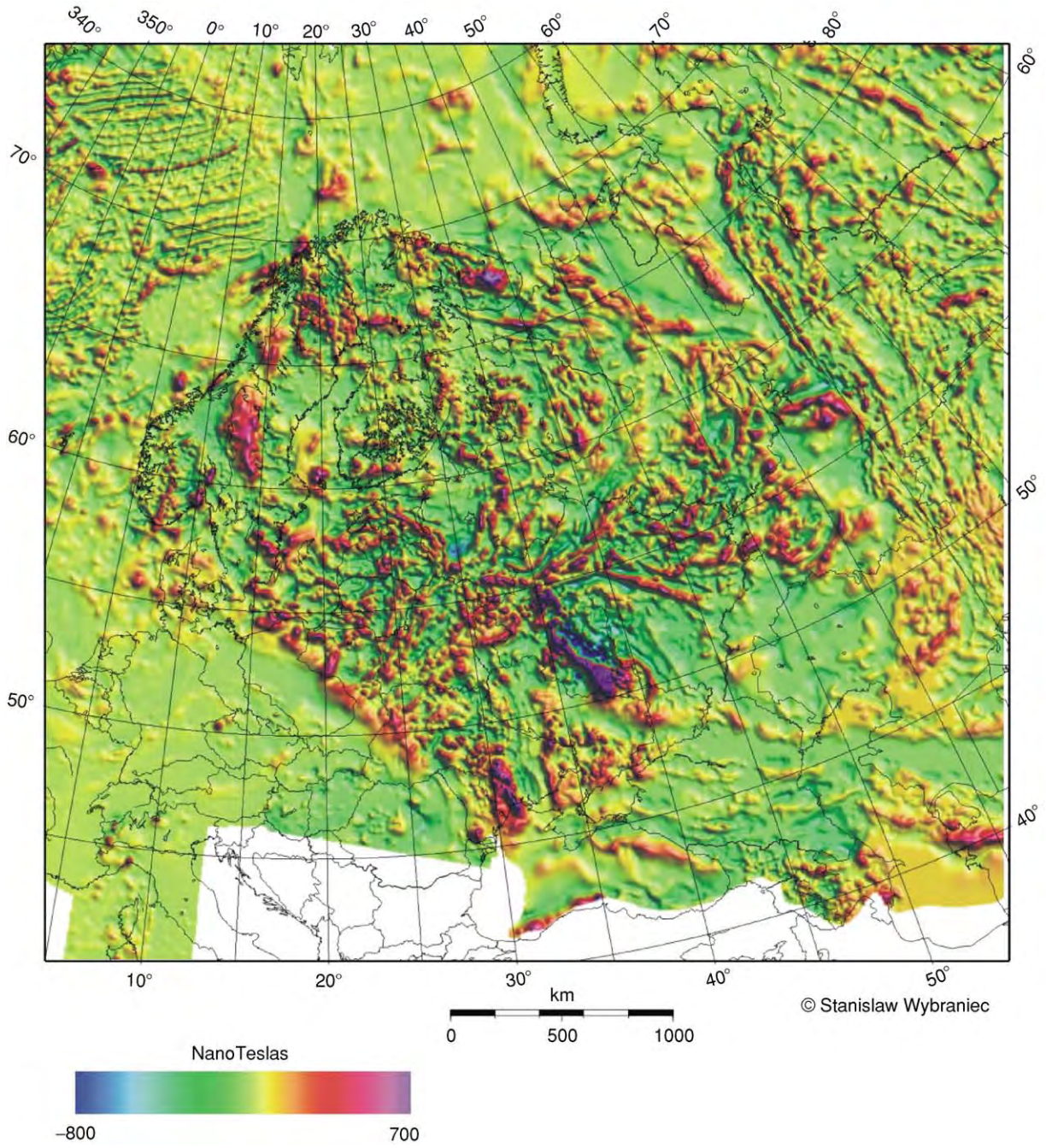


Figure 2 Magnetic anomaly patterns in the East European Craton. In general, these patterns describe very well the tripartite crustal segment structure of the East European Craton (cf. **Figure 8**). In addition, the strongest positive anomalies mark major occurrences of magnetic iron ores, for example the Kursk magnetic anomaly in the Voronezh Massif and the anomalies indicating the ores of northernmost Sweden (Kiruna, etc) and the Kriviy Rog ores in the Ukraine. Where sedimentary fillings are very thick in rifts and basins, ribbon shaped negative or alternating negative positive anomaly patterns prevail (e.g. in the Dniepr Donets and Pachelma Aulaco gens and the Peri Caspian Basin). Moderately positive anomalies are associated with amphibolite to granulite facies metamorphic belts and some granite provinces, such as the rapakivi massifs and parts of the Trans Scandinavian Igneous Belt. However, other rapakivi plutons (Viborg, for instance) have negative anomalies. Similar diversity is also found for the Archaean protocratons. Map by courtesy of S Wybaniec of the Polish Geological Institute.

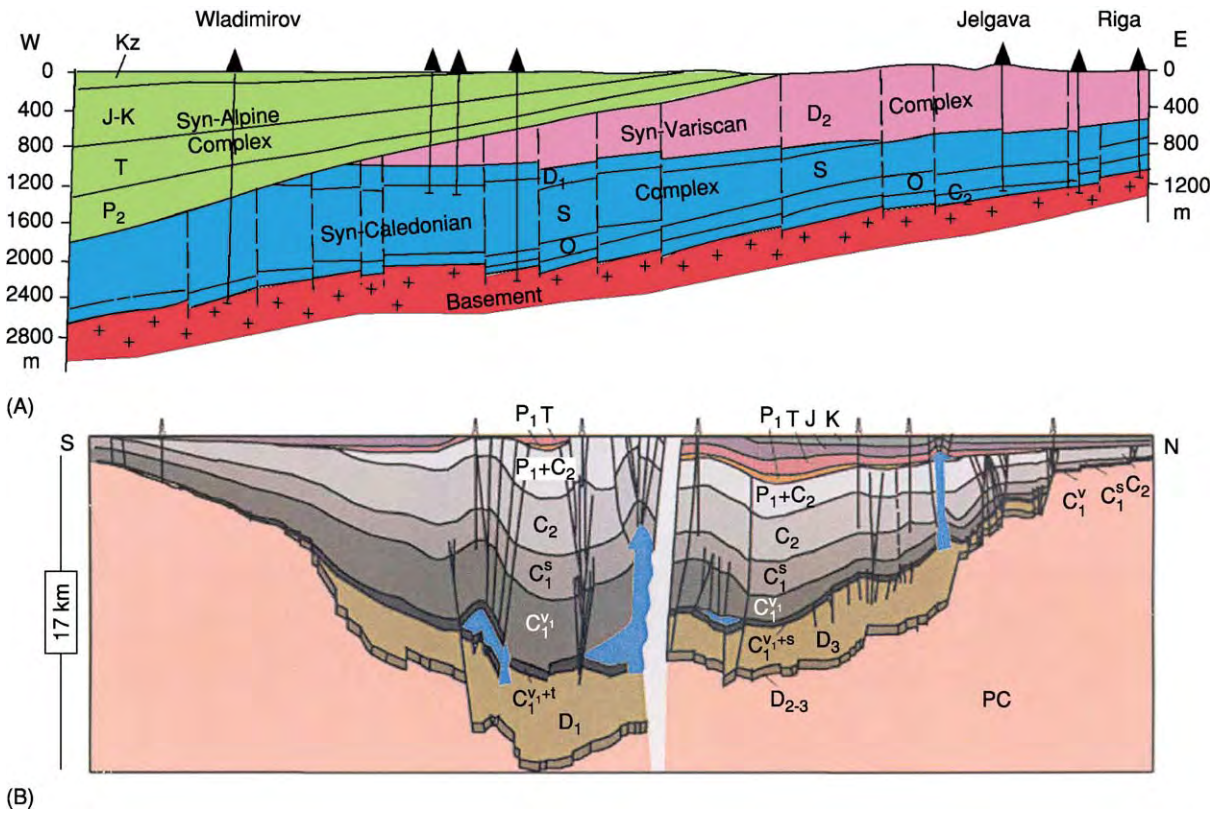


Figure 3 Sections across the Phanerozoic sedimentary cover of the Russian Platform and adjoining areas. (A) Profile of the eastern part of the Baltic Sea Depression, showing the disconformities between three different tectonostratigraphic complexes. These correspond roughly to the three Phanerozoic orogenies that shaped adjoining central Europe and changed the regimes of sedimenta tion and erosion within the East European Craton. C, Cambrian; O, Ordovician; S, Silurian; D, Devonian; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Kz, Cenozoic. (B) General seismic section across the Dniepr Donets Rift Basin. The ends of the section are approximately 170 km apart, but the middle portion of the profile has been omitted from the diagram. Salt diapirs are shown in blue. Pc, Precambrian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous. Seismic section by courtesy of the Secretariat of the EUROPROBE Programme.

Rifting in the East European Craton was commonly accompanied by within-plate mafic magmatism, which, during some periods, e.g. the Devonian, may have been associated with mantle-plume upwelling. Among the products are kimberlites (*see* **Igneous Rocks:** Kimberlite) found, for instance, in the Ukrainian Shield, Voronezh Massif, and Kola Peninsula region.

Crustal Segments of the East European Craton

The Meso- and Neoproterozoic Volyn-Orsha, central Russia, and Pachelma rifts and aulacogens divide the East European Craton into three different parts (Figures 6 and 8). Research during the last 15 years has demonstrated that these correspond roughly to three Precambrian crustal segments, each of which has its own Archaean–Palaeoproterozoic history. The three segments are Fennoscandia (including the Baltic-Fennoscandian Shield) in the north-west and north, Sarmatia (comprising the Ukrainian Shield and the Voronezh Massif) in the south, and Volgo-Uralia in the east.

Fennoscandia and Sarmatia both feature Archaean cores and accreted Palaeoproterozoic juvenile continental crust, whereas Volgo-Uralia is almost entirely Archaean but was strongly reworked in the Palaeoproterozoic.

Fennoscandia

The Fennoscandian crustal segment consists of two principal parts (Figure 9). In the north-east there is a large domain of Archaean crust, while the rest of the segment is made up of Palaeo- and Mesoproterozoic, mostly juvenile, mantle-derived rocks. These belong to several successively formed orogenic belts.

The demarcation between the Archaean and Proterozoic regions is fairly sharp, but the edge of the Archaean has been reworked tectonically and

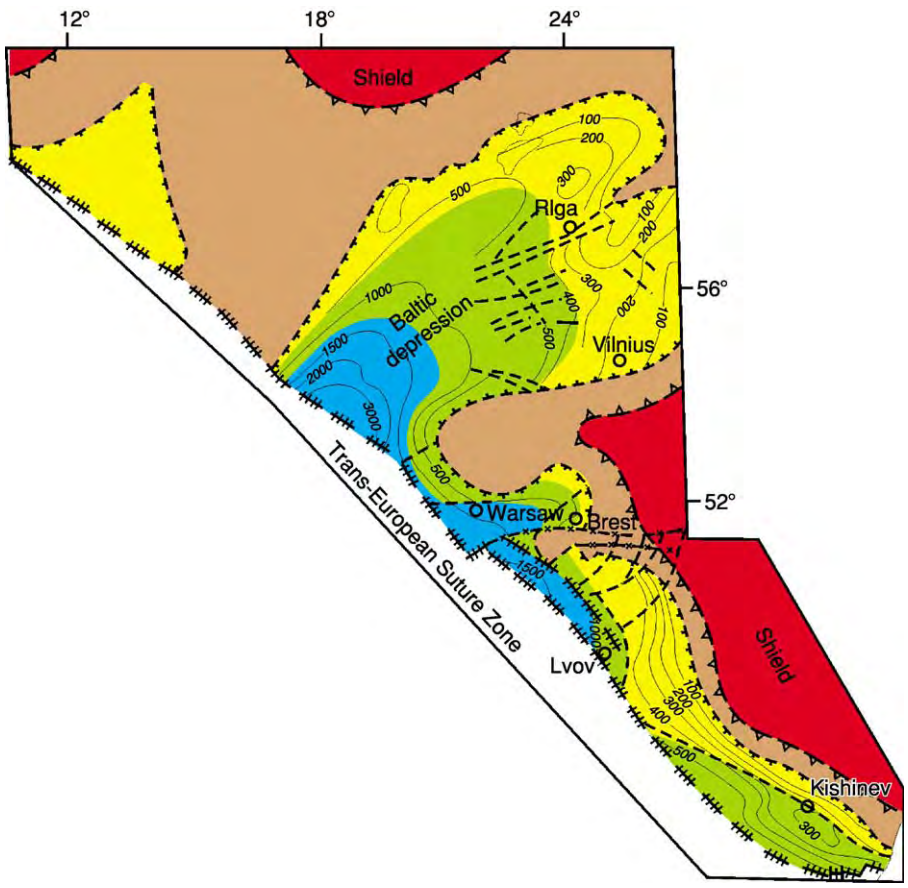


Figure 4 Lithologies of the Cambrian to Silurian rocks along the south western margin of the East European Craton. The lithofacies zones are: blue, deep water pelites; green, lower shelf calcareous and pelitic rocks; and yellow, shallow water calcareous rocks. The brown fields represent areas from which these and similar lithologies have been removed by erosion; red indicates other parts of the exposed Precambrian crystalline crust. © Svetlana Bogdanova.

magmatically, and is overlain or overthrusted by Palaeoproterozoic rock units. Inside the Archaean domain, however, is the north-west–south-east trending Late Palaeoproterozoic collisional Lapland–Kola Orogen, while isotopic data suggest that parts of the Palaeoproterozoic region in the south-west are underlain by blocks or wedges of Archaean crust.

The Archaean domain comprises the Karelian Protocraton, the Belomorian Belt along the White Sea, several different terranes in the Kola Peninsula, and a western province extending all the way to the Lofoten Islands of Norway but partly hidden beneath the Caledonide allochthon.

The Karelian Protocraton is the largest province of Archaean crust. It contains a sizable area and several minor occurrences of igneous rocks older than 3.1 Ga, the oldest zircon ages being about 3.5 Ga. These rock units are set in terranes dominated by 3.0–2.8 Ga greenstone belts rich in komatiites (*see* **Igneous Rocks:** Komatiite) and juvenile tonalite–trondhjemite–granodiorite (TTG) granitoids. The greenstone belts are of several generations and have been attributed to a number of different geodynamic settings such as continental rifts, oceanic arcs, and active continental margins. Remnants of approximately 2.8 Ga ophiolites are found in the north-eastern part of the Protocraton. In the literature, block mosaics as well as large gently dipping crustal wedge structures are indicated. This suggests a complex but still largely undeciphered story of Mesoarchaean continent assemblage. In the Neoarchaean, i.e. after 2.8 Ga, volcanism and granitoid magmatism continued, producing rocks ranging from TTG complexes to granites of collisional regimes and late post-kinematic intrusions.

Similar Neoarchaean developments characterize the adjacent Kola and Belomorian provinces, but these lack rocks older than 3.0 Ga. A notable feature of the Central Kola sub-province is the presence of the thick but subsequently highly metamorphosed 2.7 Ga Keivy cover sequence, which is known for its giant crystals of kyanite, staurolite, and other minerals.

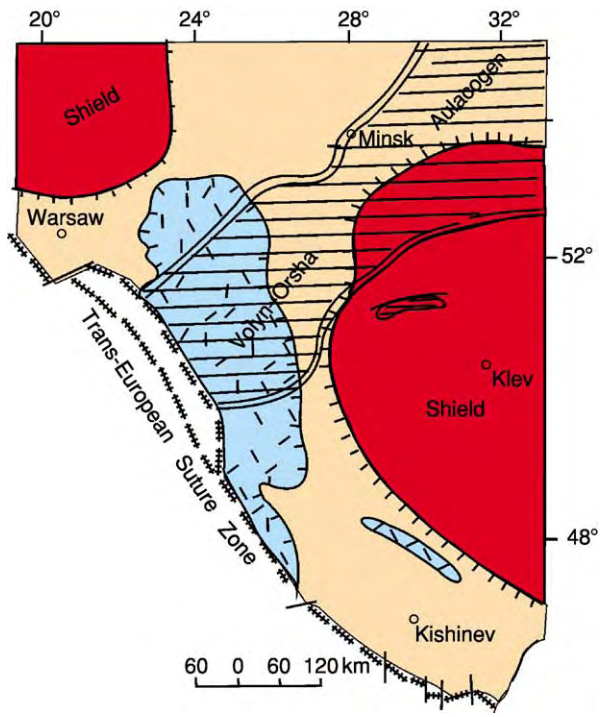


Figure 5 The south western margin of the East European Craton in the Neoproterozoic. The ruled area shows the Meso to Neoproterozoic Volyn Orsha Aulacogen, which is filled by mostly terrigenous sedimentary rocks. In the vicinity of the Trans European Suture Zone, however, the approximately 550 Ma Volynian flood basalts (blue markings) form a large outcrop set in Vendian and Early Cambrian sedimentary rocks (the sand coloured areas). The basalts fill a basin that trends across the Aulacogen and runs roughly parallel to the Trans European Suture Zone. This indicates that the south western margin of the craton was passive at the end of the Neoproterozoic. © Svetlana Bogdanova.

The Belomorian typically features extensive belts of mafic metavolcanics, continental-margin turbidites, and TTG-type granitoids, all arranged in large folded thrust sheets overriding the edge of the Karelian Protocraton. Thrusting here occurred at about 2.70–2.65 Ga, when the different terranes and crustal provinces – except possibly the western Archaean province – collided and merged to form a Neoarchaean continent. Concomitantly, parts of the Karelian province underwent high-grade metamorphism and migmatite formation. Along its boundary with the Belomorian Belt, 2.7 Ga eclogites are present. Exhumation and cooling then followed. Cratonization became complete at around 2.5 Ga.

Between 2.5 Ga and 2.0 Ga, Archaean Fennoscandia underwent extension, rifting, and even breakup, leading to the opening of small oceanic basins. Intermittently there was strong dominantly mafic but also komatiitic igneous activity. Most conspicuous are large greenstone rifts and belts, which tend to follow the boundaries between the different Archaean terranes. Ophiolites with ages of about 1.96 Ga occur in places.

The latter half of the Palaeoproterozoic was a period of intense orogenic activity across most of Fennoscandia. In the north, the Lapland–Kola collisional orogeny at about 1.96–1.85 Ga welded together the crustal units that constitute the present Archaean domain. Some of these appear to have been parts of a Neoarchaean continent that was broken up in the Early Palaeoproterozoic. In the case of other crustal units, an exotic provenance has been suggested. Altogether, the Lapland–Kola Orogen is a tectonic collage of reworked Late Archaean terranes and intervening belts of Palaeoproterozoic juvenile crust and turbidites (Figure 10). Prominent among the latter is the Lapland Granulite Belt, made famous by Eskola's work on the facies of metamorphism. Arc magmatism has been dated at 1.96 Ga; southwards thrusting and collision occurred at 1.91–1.90 Ga, while post-collisional magmatism, exhumation, and cooling lasted until 1.85 Ga.

To the south-west of the Archaean domain, the Proterozoic crust of central Fennoscandia was also created around and soon after 1.90 Ga in a series of orogenic events collectively named the Svecofennian Orogeny. Initial subduction between about 1.93 Ga and 1.91 Ga appears to have been towards the present south-west, creating a primitive volcanic arc, but soon its direction reversed to plunge beneath the Archaean Protocraton in the north-east.

In simplified outline, the Svecofennian Orogen consists of two volcanic belts and an intervening turbidite basin (the Bothnian Basin). In south central Finland there is a very large complex of granitoid plutons. The northern volcanic belt follows the edge of the Archaean Protocraton in Scandinavia but then swings south-eastwards to western and southern Finland, skirting the complex of granites. The southern belt extends between southernmost Finland and south central Sweden. Continuations of these terranes beyond Lake Ladoga into Russia are conceivable (Figure 9).

The ages of volcanism in the two belts overlap within the 1.90–1.88 Ga bracket; dominantly TTG-type plutonism lasted until about 1.87 Ga. The pressure–temperature ratios of Svecofennian metamorphism were mostly low to moderate, which attests to the absence of continent–continent collision. Between them, the two volcanic belts house the majority of the renowned Finnish and Swedish sulphide and iron ores. By 1.85 Ga, the Svecofennian Orogen had been consolidated; however, 1.86–1.84 Ga plutons of granitoids occur locally at its south-western margin and along some major intra-orogenic province boundaries.

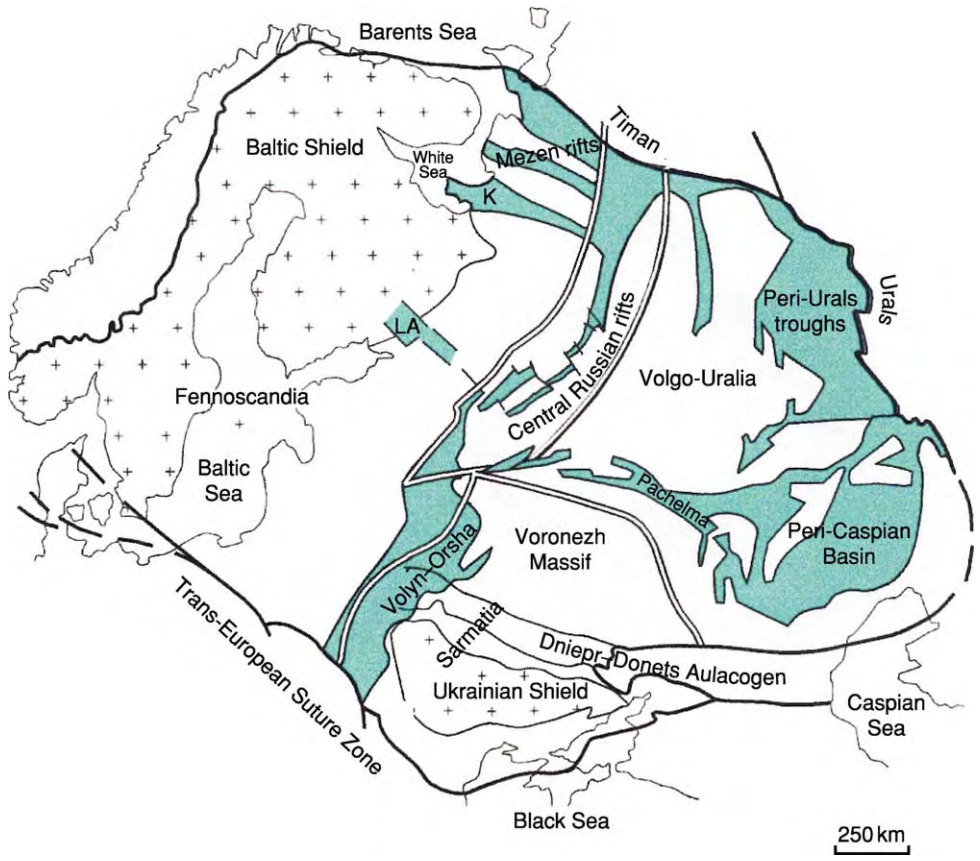


Figure 6 The patterns of the Meso to Neoproterozoic rifts and aulacogens beneath the sedimentary cover of the Russian Platform conform well with the three segment fundamental structure of the East European Craton. Thus, the Volyn Orsha, central Russian, and Pachelma rifts follow the Palaeoproterozoic sutures between Fennoscandia, Sarmatia, and Volgo Uralia. The Mezen rifts in the north align with the Lapland Kola collisional orogen and delimit the Timanides, while Volgo Uralia is completely surrounded by troughs, aulacogens, and the Peri Caspian Basin. © Svetlana Bogdanova.

Figure 7 The Mesoproterozoic (Riphean) cover sequence in the western foreland of the southern Urals was folded in the Palaeozoic. The sequence is approximately 1500 m thick. Its components are interbedded conglomerates, sandstones, silt stones, pelites, limestones, and dolostones, with numerous de posits of siderite ore. The picture shows rocks of the Bakal Formation in the Irkutskan siderite pit. © Svetlana Bogdanova.

Subsequently, granitic magmatism associated with metamorphism and migmatite formation but lacking connection with mafic rocks took place at various sites and times within the 1.84–1.78 Ga bracket. Current interpretation sees this as a response to compression and thickening of the Svecofennian crust, largely caused by orogeny outside the limits of the Svecofennian domain.

After 1.85 Ga, growth of new crust continued towards the south and the west, outwards from the cratonized Svecofennian territory. Several accretionary orogenic events at around 1.84–1.83 Ga and 1.81–1.78 Ga, and again during the Gothian Orogeny at 1.71–1.67 Ga, have been distinguished in a belt that stretches from Lithuania across southern and western Sweden to central Norway. Outside the limits of Svecofennian crust, the granitoid rocks are TTG-type calc-alkaline, whereas within a marginal western zone of the Svecofennian Orogen the 1.84–1.67 Ga

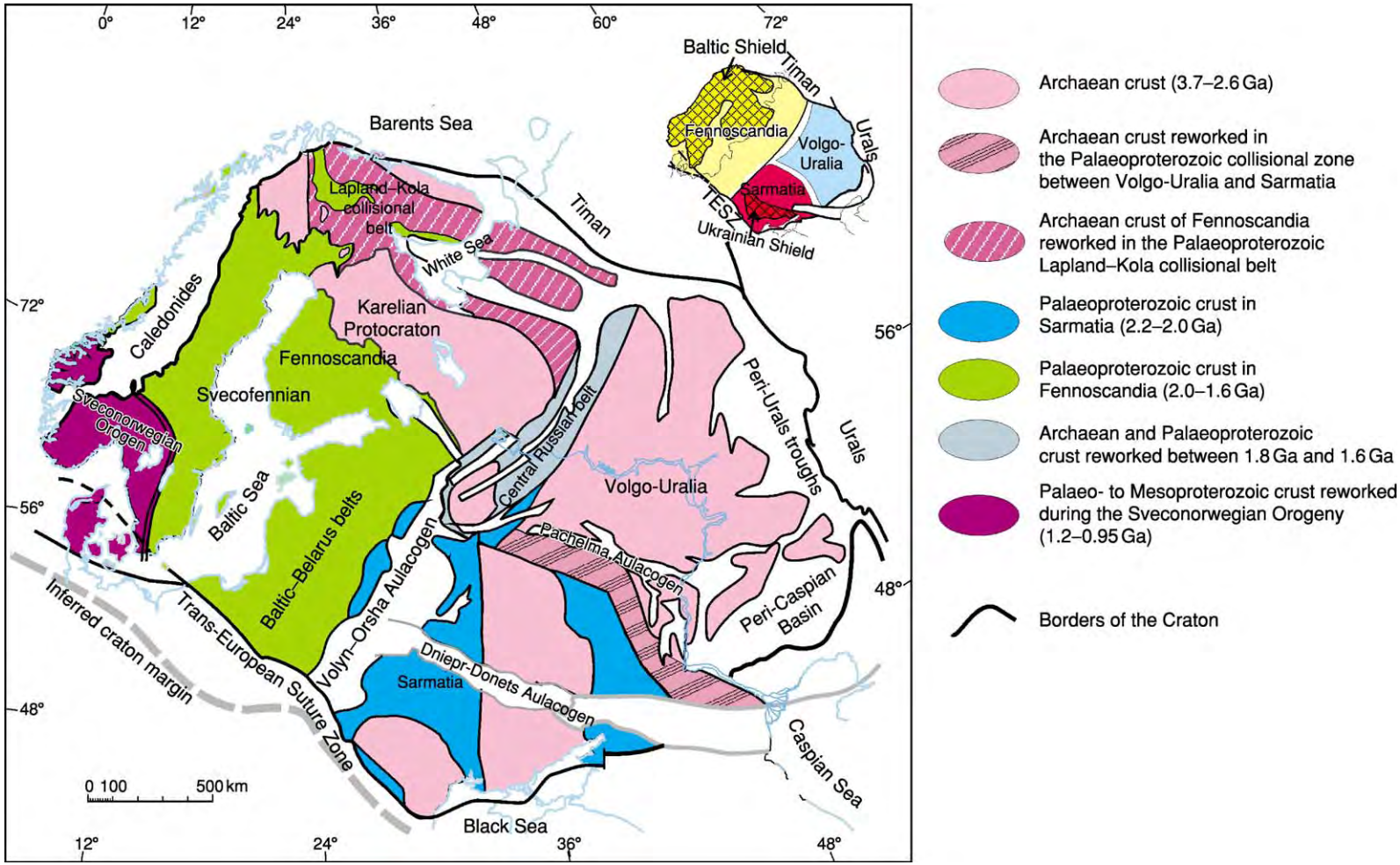


Figure 8 Major crustal provinces of the East European Craton. Apart from Palaeoproterozoic domal structures and some sedimentary and igneous rocks, Volgo Uralia is almost completely Archaean. Sarmatia, in contrast, comprises several large Archaean protocratonic units that were cemented together mostly during the Palaeoproterozoic. In Fennoscandia, Archaean crust is concentrated in the north eastern part, where it forms a separate domain. Sarmatia and Fennoscandia contain relics of approximately 3.7 3.5 Ga crust, while Volgo Uralia is wholly Neoarchaean. Another important difference is that the dominant 2.2 2.0 Ga Palaeoproterozoic crust in Sarmatia is substantially older than the 1.92 1.85 Ga crust of central Fennoscandia. Thus, the age patterns and structural arrangements in each of the three crustal segments are distinct. They are also largely discordant at the segment boundaries. The boundary between Sarmatia and Volgo Uralia is marked by belts of reworked Archaean crust, while the boundary between Fennoscandia and the other two crustal segments is different in its north eastern, largely Archaean, and south western, dominantly Proterozoic, parts. In the north east, the 200 km wide Central Russian Belt is built up of elongated wedged blocks derived from the adjacent crustal segments, whereas in the south west, beneath the Neoproterozoic Volyn Orsha Aulacogen, there are various Palaeoproterozoic terranes ranging from continental margin types to outboard volcanic arcs. These were stacked atop each other between about 1.8 Ga and 1.7 Ga. © Svetlana Bogdanova.

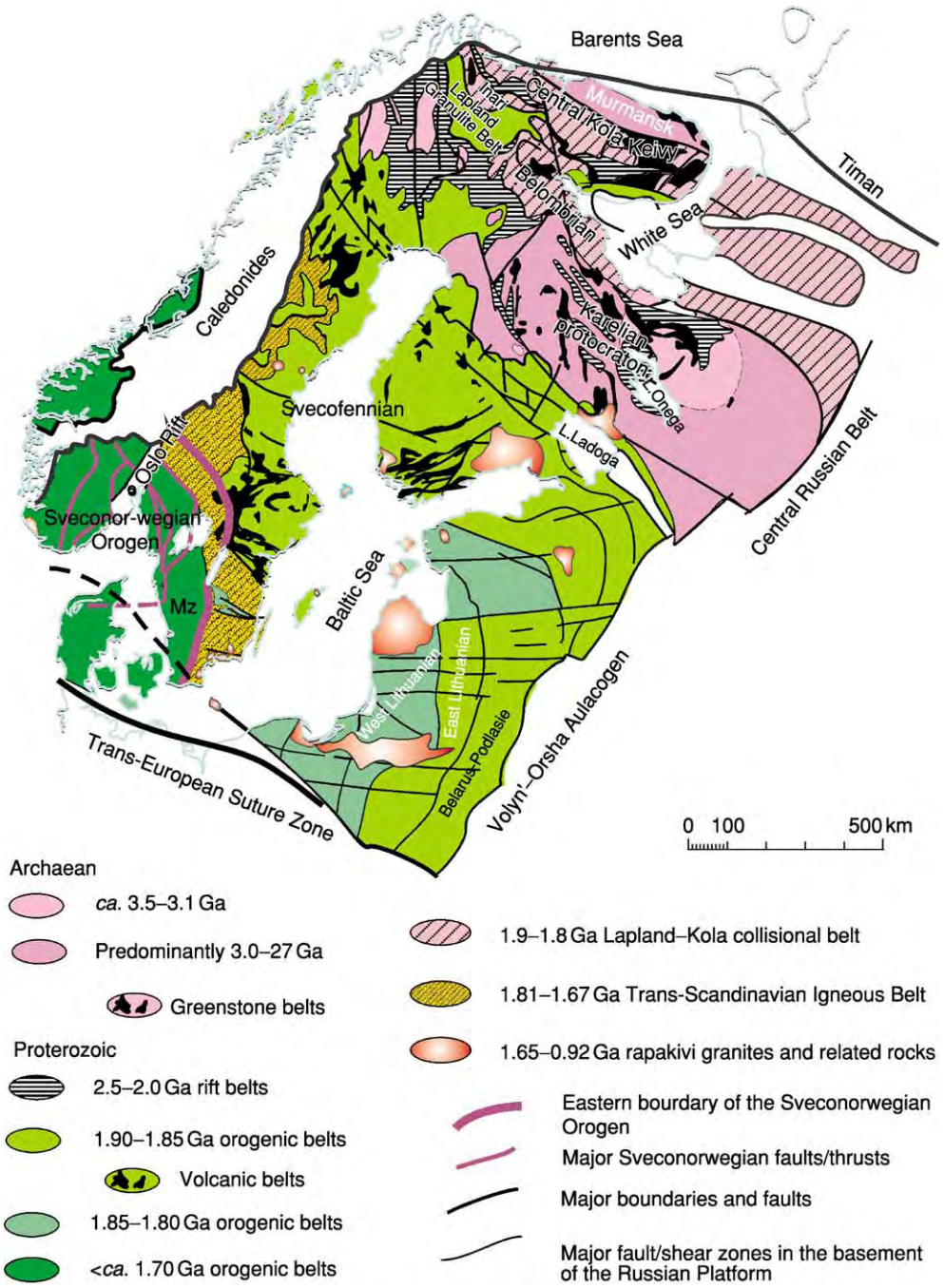


Figure 9 The major crustal provinces of Fennoscandia reflect the subdivision of that crustal segment into largely separate Archaean and Proterozoic parts. The master story of the Archaean part appears to have been the formation of a Mesoarchaean Protocraton in Karelia followed by a Neoarchaean continent. The latter was broken up in the Early Palaeoproterozoic but then brought together again during the Lapland Kola collisional Orogeny. In that process, Archaean terranes exotic to Fennoscandia may also have participated. Essentially coeval with the Palaeoproterozoic Lapland Kola collision in the present north was the first stage of the accretionary formation of Proterozoic crust in the rest of the Fennoscandian crustal segment. In the present west, this process continued episodically until about 1.55 1.50 Ga, apparently without involving major collisions of continental crustal units. In the south west, continental collision is indicated for the period between 1.5 Ga and 1.4 Ga, when interaction with a different craton possibly Proto Amazonia appears to have led to the Danopolonian Orogeny. The subsequent Sveconorwegian orogenic process, between about 1.1 Ga and 0.95 Ga, was pronouncedly collisional and approximately coeval with the Grenvillian Orogeny in Laurentia and the assembly of the Supercontinent Rodinia. Along the south eastern margin of Fennoscandia, the wide Baltic Belarus belts of stacked inboard and outboard terranes mark the Late Palaeoproterozoic (*ca.* 1.8 1.7 Ga) collision zone with Sarmatia. Here, the West Lithuanian Belt and its continuation into south eastern Poland differ sharply from the adjacent East Lithuanian and Belarus Podlasie belts in terms of both the direction of the tectonic grain and the age of crustal formation. In western Lithuania, the crust is nearly coeval with that in south western Sweden. Mz, Mylonite Zone. © Svetlana Bogdanova.

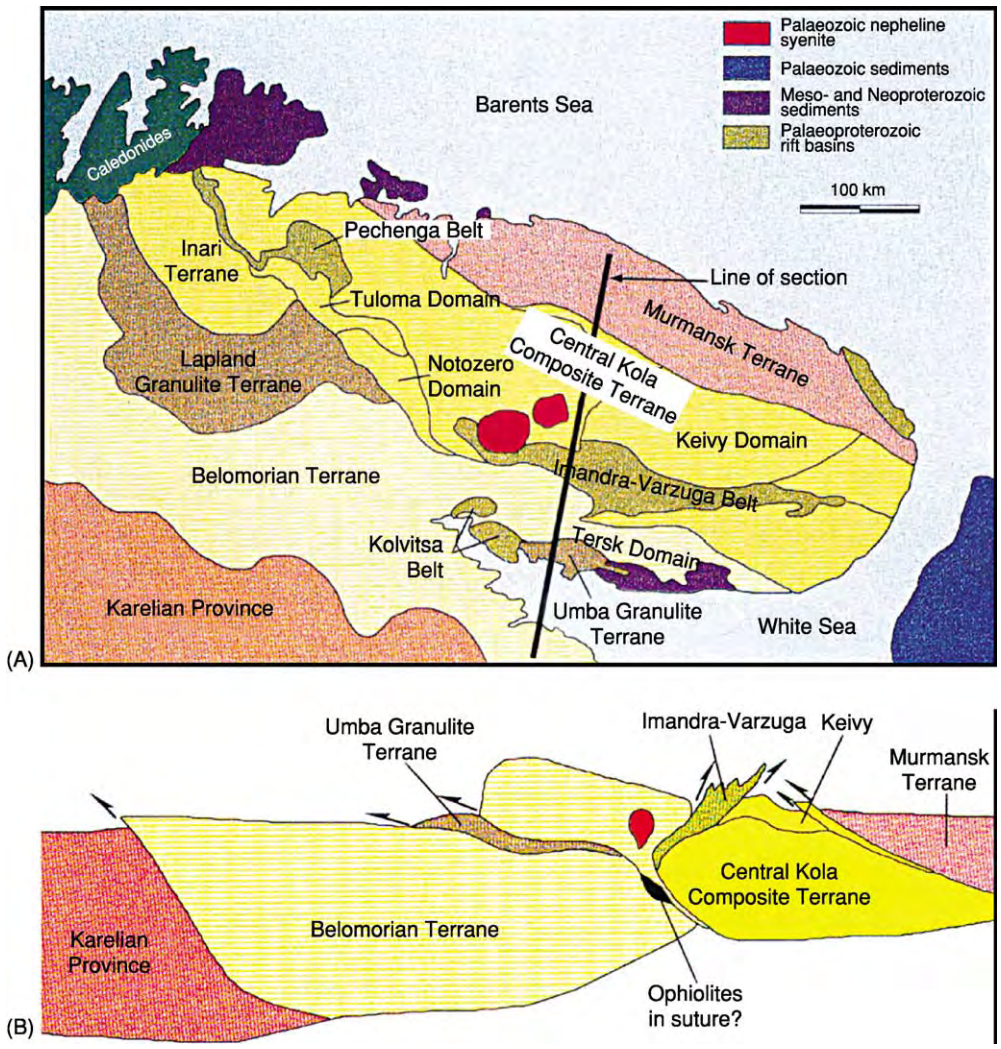


Figure 10 (A) Tectonics of the Lapland Kola collisional belt in the Kola Peninsula. (Map supplied by the Secretariat of the EUROPROBE Programme). (B) Cross section showing the juxtaposition of various Archaean terranes (the Karelian, Belomorian, Murmansk, and Central Kola Keivy) with the Palaeoproterozoic Imandra Varzuga volcanic belt and the Umba Granulite Terrane. Map and cross section by courtesy of the Secretariat of the EUROPROBE Programme.

magmatism was bimodal mafic to granitic and commonly alkali-calcic. As a cumulative result of the different igneous episodes, the large north–south trending Transscandinavian Igneous Belt was formed within the western margin (Figure 9). The Baltic–Belarus belts, some of them with a Palaeoproterozoic evolution similar to that of the Svecofennian in the Baltic Shield, characterize the south-western part of Fennoscandia; they are, however, overlain by phanerozoic sedimentary cover (Figure 3A).

During the Mesoproterozoic, renewed Orogeny between 1.6 Ga and 1.5 Ga created new crust in westernmost Sweden and southern Norway (the Kongsbergian event), etc. In contrast, orogenic events around 1.5–1.4 Ga (the Danopolonian Orogeny) and between 1.1 Ga and 0.9 Ga (the Sveconorwegian Orogeny) were largely collisional. The Sveconorwegian Orogeny involved only limited creation of new crust in south-western Norway but fairly voluminous granitic magmatism and granulitic to eclogitic metamorphism with migmatization in many places. Some of the metamorphic assemblages indicate pressures between 1 GPa and 1.5 GPa. High-grade granulitic rocks accompanied by charnockite, and numerous large granite bodies were also formed at around 1.4 Ga. During both these events, major more-or-less north–south-trending faults, thrusts, and suture zones were developed in most of southern Scandinavia. According to some hypotheses, all of southern Norway and western Sweden west of the so-called 'Mylonite Zone' (Figure 9) constitute terranes exotic to Fennoscandia, which were accreted to the East European Craton during the Sveconorwegian Orogeny.

At about 1.64–1.62 Ga, and again at 1.57 Ga, around 1.5 Ga, and 0.93 Ga, numerous large intrusions of rapakivi granites and associated charnockitic, mangeritic, and anorthositic rocks penetrated the Fennoscandian crustal segment. Extensive swarms of mafic dykes and sheets were formed throughout the Proterozoic, with maxima at 2.45–2.1 Ga, 1.6–1.5 Ga, and 1.3–1.2 Ga and 0.97–0.95 Ga. Some of these may be the results of superplume events in the Earth's mantle. In addition there are plugs and plutons of alkaline and carbonatitic rocks (*see* **Igneous Rocks:** Carbonatites), some of which are Precambrian and others Phanerozoic in age. Alkaline magmatism was particularly voluminous in the Kola Peninsula and along rifts such as the Permian Oslo Graben in Norway. Other intrusions formed isolated minor plugs.

Sarmatia

The Sarmatian segment is built up of several Archaean crustal provinces with ages ranging between 3.7 Ga and 2.7 Ga (Figure 11). Belts of Palaeoproterozoic crust intervene between some of these, but accretion and welding appear to have commenced in the Archaean.

The oldest crustal units are the Palaeo- to Mesoarchaean Podolian Block in the extreme southwest and the Oskol (Kursk-)–Azov Block in the east. Ages of 3.65–3.60 Ga have been obtained from ultramafic and tonalitic rocks, while major TTG intrusions are 3.4–3.3 Ga old. Both the Podolian Block and the Oskol–Azov Block represent one-time lower crust but have evolved through several events of high-grade granulitic metamorphism and multiphase magmatism in the Archaean and Palaeoproterozoic.

The Sumy–Dniepr gneiss–granite–greenstone terrane is generally somewhat younger. It was formed mainly between 3.2 Ga and 3.1 Ga and is characterized by numerous belts of principally mafic metavolcanic and sedimentary rocks. These contain basal sequences of komatiites, komatiitic basalts, and tholeiites, reaching thicknesses of 4–5 km. Atop the mafic volcanics rest calc-alkaline felsic volcanic rocks, tuffites, immature sandy metasediments, and banded iron formations with quartzites. These belts were metamorphosed and deformed concomitantly with the development of domal structures containing juvenile TTG-type granitoids and felsic gneisses with granulites. The alteration processes were related either to rifting and the opening of a Mesoarchaean ocean or to the evolution of a preexisting oceanic basin of Palaeo- to Mesoarchaean age.

In the Palaeoproterozoic, between about 2.6 Ga and 2.3 Ga, major belts of supracrustal rocks containing numerous banded iron formations were formed. Some of these outline the margins of Archaean terranes, while others appear to be controlled by rifts in the interior parts of the Oskol–Azov Block. The latter are the cause of the strong Kursk magnetic

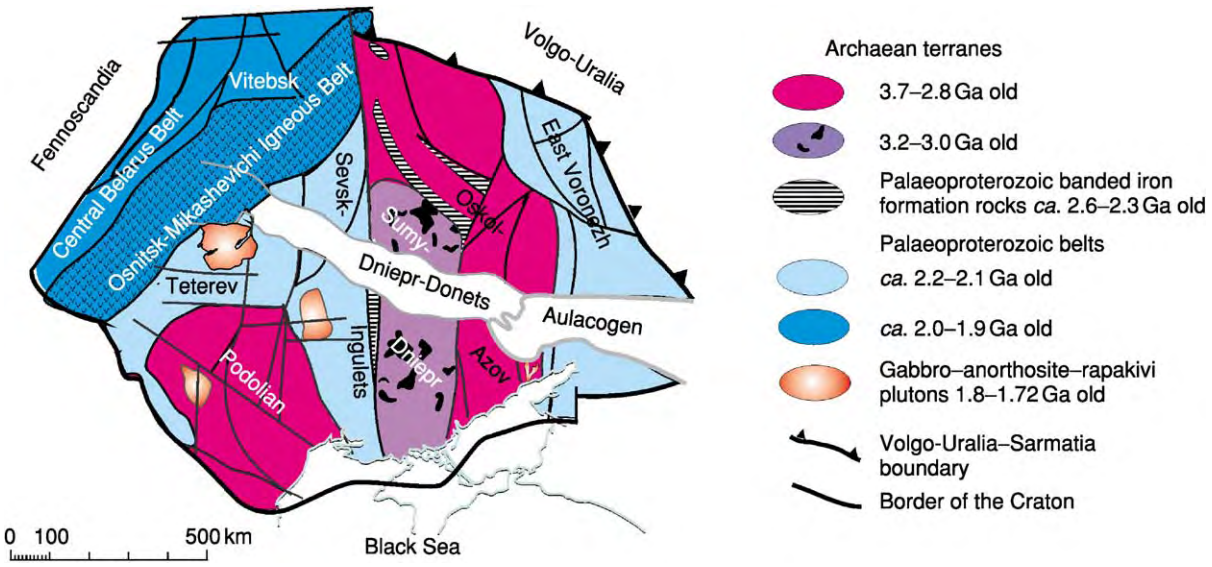


Figure 11 The processes that shaped the internal structure of Sarmatia were the docking of the older Archaean Oskol Azov and the younger Archaean Sumy Dniepr crustal blocks and, during the Proterozoic, the accretion of the Podolian Block to the evolving main mass of the Sarmatian crustal segment. The latter process was coeval with the generation of mantle derived juvenile Proterozoic crust elsewhere in Sarmatia. Directly related to the assembly of the East European Craton were the formation of the East Voronezh belts along the developing boundary between Sarmatia and Volgo Uralia about 2.1 2.0 Ga ago and the formation of the continental margin Osnitsk Mikashevichi Igneous Belt and the outboard Central Belarus Belt at around 2.0 Ga. These two belts faced the ocean basin that was consumed during the collision of united Volgo Uralia and Sarmatia (i.e. 'Volgo Sarmatia') with Fennoscandia. © Svetlana Bogdanova.

anomaly that marks an important iron-ore province. At about the time that these rock belts were formed, an ancient platform cover began to be deposited. The largest sedimentary basin was in the Sevsk-Ingulets (Kirovograd) domain.

In north-western Sarmatia, some Palaeoproterozoic volcanic rocks, e.g. those in the Teterev Belt, have mantle-related geochemical signatures and probably represent island arcs. The most conspicuous complexes of juvenile igneous rocks, however, are found in the East Voronezh belts in eastern Sarmatia, where an approximately 2.1 Ga volcanic arc and back-arc sedimentary basins with numerous ultramafic and mafic intrusions occur. Together with numerous granitoid intrusions of the same age along the edge of the Archaean Oskol–Azov Block, these indicate the presence of an ocean–continent interface. At this time Sarmatia was assembled and its dominant north–south-trending suture zones were created. These have crustal roots approximately 60 km deep.

Along the north-western margin of Sarmatia, the Early Palaeoproterozoic north–south fabric is truncated by several wide north-east-trending belts of Late Palaeoproterozoic rocks. These are the Osnitsk–Mikashevichi Igneous Belt, the Central Belarus Belt, and the Vitebsk granulite domain (**Figure 11**), all of which were formed between 2.0 Ga and 1.9 Ga, a top south-eastwards dipping subduction zones or zones of arc collision.

In particular, the presence of the extensive Osnitsk–Mikashevichi Igneous Belt suggests the existence at around 2.0 Ga of a unified Sarmatia, complete with a coherent north-western to northern margin. The Osnitsk–Mikashvichi Igneous Belt is 200–250 km wide and comprises volcanic, hypabyssal, and dominantly plutonic igneous rocks. The volcanic and dyke rocks are metabasalts and dolerites, meta-andesites, and meta-keratophyres, all metamorphosed in various ranges of the amphibolite facies. Granodiorites and granites with subordinate gabbros and diorites form large batholithic plutons, which are only weakly deformed and metamorphosed. The metavolcanics and minor interbedded metasedimentary rocks occupy small areas within the plutons or form septa between them.

A striking feature in Sarmatia is the occurrence of several large plutons of gabbro-anorthosites and rapakivi granites of 1.80–1.72 Ga age that are rich in titanium and other ore deposits. Traditionally, these intracratonic plutons have been considered to be anorogenic, but recent discussion tends to connect them with zones of preceding plate collision. New seismic results indicate that the anorthosite–rapakivi magmatism strongly influenced the composition of both the crust and the upper mantle.

In the Devonian, the formation of the Dniepr–Donets Aulacogen subdivided the Sarmatian crustal segment into two parts. The southern part contains the Ukrainian Shield, and the northern part contains the Voronezh Massif. However, no significant displacement along the aulacogen can have taken place, since the terrains on each side correlate excellently.

Volgo-Uralia

The crystalline crust of Volgo-Uralia is mostly Neoarchaean, the highest depleted-mantle Nd model ages being approximately 3.1 Ga. This crust is completely buried beneath a Phanerozoic, mostly Devonian to Triassic, sedimentary cover. The only exception is the Taratash complex (**Figure 12**), where a basement slice was brought up to the Earth's surface by thrusting connected with the Uralide Orogeny.

Under the Phanerozoic cover, however, there are numerous Meso- to Neoproterozoic aulacogens and troughs filled with 2–10 km thick sedimentary rocks intercalated with rare volcanics. Proterozoic rifting substantially complicated the margins of Volgo-Uralia and penetrated deeply into its interior. Because of the extensive cover, information on the crystalline crust is mostly derived from geophysics and from thousands of drill cores, which are particularly numerous because of the high oil and gas potential of the region.

Volgo-Uralia in general is a realm of high magnetization and dense upper crust. A characteristic feature of its magnetic field is the presence of numerous, in part very large (300 km and more), circular to oval concentric anomalies separated by wide belts of more or less linear anomaly patterns. Traditionally, such patterns have been explained by assuming that the non-linear nearly isometric anomalies represent stable Archaean massifs, while the linear zones relate to Palaeoproterozoic mobile belts. In Volgo-Uralia, however, this interpretation is at complete variance with the presently available geological information, which demonstrates that the circular anomalies represent Palaeoproterozoic domal, relatively uplifted, structures. In contrast, the linear belts consist of relatively well-preserved Archaean rocks. These are 3.0–2.7 Ga old and comprise metasedimentary and subordinate metaigneous granulites, these two types mostly forming separate belts. In addition, there are also greenstone sequences with komatiitic volcanics. Associated with the Archaean belts are zones of strong shearing and mylonitization defining fold–thrust structures and generally trending north-east–south-west to east–west.

Most of the domes in the Volga–Kama megablock (**Figure 12**) are made up of Archaean granulites reworked during the Palaeoproterozoic. In their

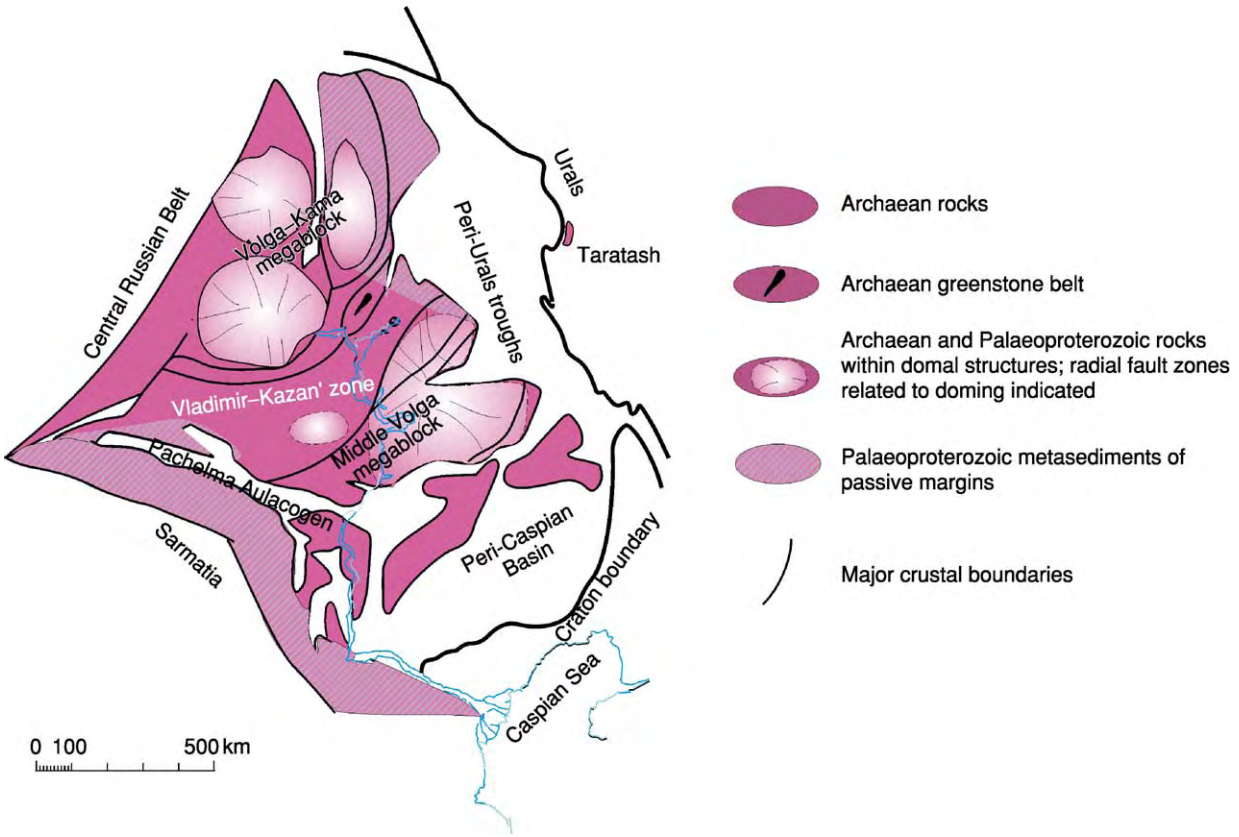


Figure 12 The specifics of Volgo Uralia were largely determined by the development of Palaeoproterozoic domal structures in the Neoarchaean continental crust. At about 2.1 2.0 Ga, the marginal belt facing Sarmatia was formed. This belt is composed of Palaeoproterozoic sedimentary commonly flysh like rocks, which suggest the one time existence of a passive margin. Subsequent Mesoproterozoic rifting and the formation of troughs, aulacogens, and basins influenced the structure of Volgo Uralia more than that of the other two East European crustal segments. The Peri Caspian Basin is unique in the East European Craton in having a relatively shallow Moho and a great thickness of sedimentary cover. © Svetlana Bogdanova.

central parts are Palaeoproterozoic granite and gneiss complexes. In contrast, the very large circular structure in the Middle Volga megablock has Palaeoproterozoic rocks and Palaeoproterozoic reworking on its flanks, while north-east–south-west striking Archaean granulites similar to those outside the dome occupy the centre.

All the domes feature radial and ring-shaped faults related to the deformation of the Archaean crust and associated with rare Palaeoproterozoic metasedimentary and metavolcanic rocks. The domes were formed between 2.3 and 2.0 Ga and may be related to the nearly coeval collision of Volgo-Uralia with Sarmatia. Along that collision zone and in the north-east of Volgo-Uralia there are large areas of Palaeoproterozoic turbiditic and metapelitic mica schists, silts, sandstones, and carbonaceous shales formed on the shelf margins of the crustal segment.

The Archaean and Palaeoproterozoic rocks of Volgo-Uralia are cut by numerous mafic dykes and sills, which were intruded during several stages of Meso- to Neoproterozoic rifting.

The Proterozoic fault systems of Volgo-Uralia strongly influenced the formation and structure of the sedimentary cover and the locations of the major Palaeozoic oil and gas resources.

Assembly of the East European Craton

The assembly of the East European Craton can be followed from palaeomagnetic data, isotopic ages, and analyses of rock formation and igneous activity at the margins of the three involved crustal segments (Figure 13).

First to collide were Volgo-Uralia and Sarmatia. The collision occurred at 2.1–2.05 Ga, when Volgo-Uralia was thrust onto Sarmatia. Geophysically, the suture can be followed for 30 km beneath Volgo-Uralia. At the Earth's surface, the junction zone contains extensive shelf deposits on the Volgo-Uralian

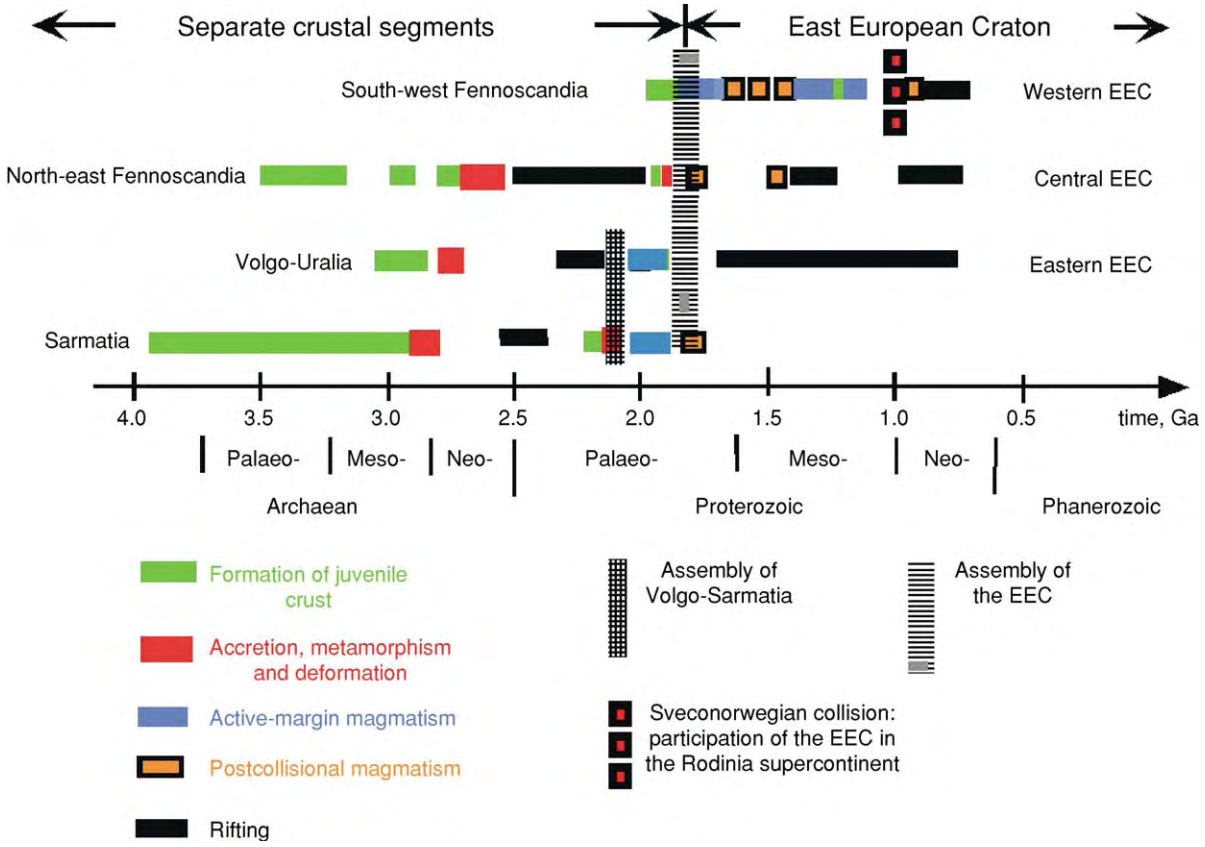


Figure 13 Fennoscandia, Sarmatia, and Volgo Uralia, the three crustal segments of the East European Craton (EEC), evolved differently during the Archaean and Palaeoproterozoic. Sarmatia and Volgo Uralia merged at around 2.1 2.0 Ga, creating a larger continental unit, Volgo Sarmatia, and thereafter shared a common history. Between about 1.8 Ga and 1.7 Ga, Fennoscandia and Volgo Sarmatia gradually amalgamated to form the EEC. The eastern and western margins of the EEC differed during the rest of the Proterozoic: mostly passive in the east and with episodes of activity in the west. In the west, the *ca.* 1.0 Ga Sveconorwegian collisional orogeny marks the beginning of the participation of the EEC in the Supercontinent Rodinia. © Svetlana Bogdanova.

side, while the so-called East Voronezh belts mark the edge of Sarmatia.

The south-westernmost belt represents a volcanic island arc, while the north-eastern belts preserve back-arc volcanic rocks and turbidites. Voluminous granitic magmatism occurred particularly in Volgo-Uralia. During the Mesoproterozoic, the Pachelma Aulacogen developed along the Sarmatia–Volgo-Uralia junction.

After that collision, it was several hundred million years before the combined Sarmatia and Volgo-Uralia ('Volgo-Sarmatia') approached Fennoscandia. Meanwhile, accretionary crustal growth occurred separately along the edges of both Fennoscandia and Sarmatia. At approximately 2.0 Ga, subduction of oceanic crust beneath the presently north-western edge of Volgo-Sarmatia commenced. This process eventually consumed all the oceanic crust between the protocraton and Fennoscandia. Within the continental margin of Sarmatia, the Osnitsk–Mikashevichi Belt began to develop, while the Central Belarus island arcs were formed outboard. Further west, the Middle Lithuanian Suture Zone, separating east–west trending rocks of the West Lithuanian Terrane from a belt of north-north-east-striking rocks in eastern Lithuania, was created. It may either coincide with the actual Fennoscandia–Sarmatia suture or be one of several components within a swarm of intersegment terrane boundaries. Palaeomagnetic data indicate that the colliding segments were still at least 2000 km apart between 1.80 Ga and 1.75 Ga.

The Palaeoproterozoic suture between Fennoscandia and Volgo-Sarmatia became the site where the Volyn–Orsha and central Russian systems of aulacogens, rifts and troughs, and sedimentary basins were developed in the Late Mesoproterozoic and Early Neoproterozoic.

See Also

Europe: Timanides of Northern Russia; Scandinavian Caledonides (with Greenland); The Urals. **Igneous Rocks:**

Carbonatites; Kimberlite; Komatiite. **Shields**. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Bogdanova SV (1986) *Zemnaya Kora Russkoy Plity v Rannem Dokembrii* [The Earth's Crust of the Russian Platform in the Early Precambrian]. Transactions of the Geological Institute of the USSR Academy of Sciences, vol. 408. Moscow: Nauka. [In Russian.]

Bogdanova S, Gorbatschev R, Stephenson RA, and Guterch A (eds.) (2001) EUROBRIDGE: Palaeoproterozoic accretion of Fennoscandia and Sarmatia. *Tectonophysics* 339: 1 237.

Gaál G and Gorbatschev R (1987) An outline of the Precambrian evolution of the Baltic Shield. *Precambrian Research* 35: 15 52.

Gee DG and Zeyen HJ (eds.) (1996) EUROPROBE 1996, Lithosphere Dynamics: Origin and Evolution of Continents. Published by the EUROPROBE Secretariate, Uppsala University, p. 138.

Glebovitsky VA (1997) *Early Precambrian of Russia*. London: Taylor and Francis Books Ltd.

Gorbatschev R and Bogdanova S (1993) Frontiers in the Baltic Shield. *Precambrian Research* 64: 3 21.

Gower CF, Rivers T, and Ryan B (eds.) (1990) *Mid Proterozoic Laurentia Baltica*. Special Paper 38. St Johns: Geological Association of Canada.

Khain VE (1985) *Geology of the USSR. First Part: Old Cratons and Paleozoic Fold Belts*. Berlin: Gebrüder Borntraeger.

Koistinen T, Stephens MB, Bogachev V, *et al.* (2001) *Geological Map of the Fennoscandian Shield, Scale 1 : 2 000 000*. Moscow: *Geological Surveys of Finland, Norway, and Sweden and North West Department of Natural Resources of Russia*.

Leonov YG and Khain VE (eds.) (1996) *International Tectonic Map of Europe and Adjacent Areas*, 3rd edn, *Scale 1 : 5 000 000, 5 sheets*. Paris: IUGS/UNESCO.

Nironen M (1997) The Svecofennian Orogen. *Precambrian Research* 86: 21 44.

Peive AV, Khain VE, Muratov MV, and Delany F (eds.) (1979) *International Tectonic Map of Europe and Adjacent Areas, Scale 1 : 2 500 000, 20 sheets*. Moscow: IUGS/UNESCO.

Stephenson RA, Wilson M, De Boorder H, and Starostenko VI (eds.) (1996) EUROPROBE: intraplate tectonics and basin geodynamics of the East European Platform. *Tectonophysics* 268: 1 309.

Thybo H, Pharaoh T, and Guterch A (eds.) (2002) Geophysical investigations on the Trans European Suture Zone II. *Tectonophysics* 360: 1 314.

Windley BF (1995) *The Evolving Continents*, 3rd edn. Chichester: John Wiley & Sons Ltd.

Ziegler P (1990) *Geological Atlas of Western and Central Europe*. Bath: Shell International Petroleum Mij BV and Geological Society Publishing House.

Timanides of Northern Russia

D G Gee, Uppsala University, Uppsala, Sweden

Introduction

Much of eastern and northern Europe, from the Ural Mountains to the high Arctic of Novaya Zemlya and westwards into the eastern Barents Sea and Pechora Basin (Figure 1), has a bedrock that is dominated by Palaeozoic and Mesozoic successions, which unconformably overlie deformed and metamorphosed Neoproterozoic rocks. The Late Neoproterozoic (Vendian) orogeny of the latter is referred to as Timanian, based on type areas in the Timan Range. Because it is largely covered by Phanerozoic successions and, towards the east in the Ural Mountains, much influenced by Late Palaeozoic to Early Mesozoic orogeny, the Timanide Orogen is less well known than many of the other old exhumed European mountain belts. Nevertheless, it dominates a large part of the Earth's crust in northeastern Europe, comprising the basement to thick hydrocarbon-bearing successions in the overlying Phanerozoic cover.

The concept of a Timanide mountain belt and related orogeny goes back at least a hundred years ago, based on type areas in the Timan Range. Other names that are widespread in the literature, for example, Baikalian, Cadomian and Assyntian, derived from contemporaneous orogenies in other parts of Eurasia, are less appropriate (and may even be misleading) for this East European Orogen.

The sedimentary rocks involved in the Timanide Orogen, exposed in the Timan Range and further south, in the western foreland of the Ural Mountains, were deposited along the eastern margin (present day coordinates) of the East European Craton (EEC) (*see* **Europe:** East European Craton) in the Neoproterozoic (1000–545 Ma); in some areas, sedimentation started in the Mesoproterozoic (1600–1000 Ma). The underlying EEC, dominating the European bedrock from Poland to the Uralian foreland and forming the ancient core of northern Europe, is composed of Archaean and Palaeoproterozoic crystalline complexes, mostly of

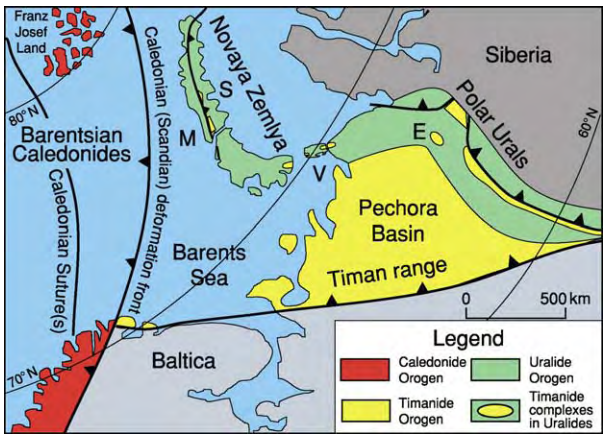


Figure 1 North western Russia: tectonic relationships. M: Mitushev Bay, S: Sulmenev Bay, V: Vaigach Island.

highly metamorphosed sedimentary and igneous rocks. From about 1650 Ma, the EEC was subject to repeated episodes of rifting, with the development of intracratonic basins (aulacogens). This rifting also controlled the structure of the eastern margin of the craton, with general subsidence leading to widespread deposition. The latter continued until the onset of Timanian Orogeny in the Vendian. Thereafter, uplift and deep erosion reduced most of the orogen to a peneplain which, early in the Palaeozoic, developed as a rifted margin to the craton-cored continent; the latter is generally referred to by palaeogeographers as Baltica. Deposition along the eastern edge of Baltica continued, mostly in passive margin shelf facies, until the beginning of the Uralian Orogeny in the Late Palaeozoic.

The name Baltica is applied to both a palaeocontinent (with a core of the EEC) and a plate (i.e. the continent with surrounding oceanic crust, reaching out to the plate boundaries) which existed in the Late Precambrian and Early Palaeozoic. Characteristic endemic faunas defined its independence from other major continents in the Early Cambrian. Baltica became a part of Laurussia during the mid-Palaeozoic Caledonian Orogeny. Baltica came into existence as an independent plate as the result of Neoproterozoic rifting and fragmentation of the supercontinent Rodinia; exactly when is not well defined. However, the eastern (Timanian) oceanic margin of Baltica was clearly established earlier in the Neoproterozoic than the north-western Baltoscandian margin.

The foreland folding and thrusting of the Timanide Orogen (Figure 2) extends for a length of at least 3000 km from the southern Urals (where it disappears beneath younger successions), northwards along the Uralian deformation front and then northwestwards via the Timan Range and northern

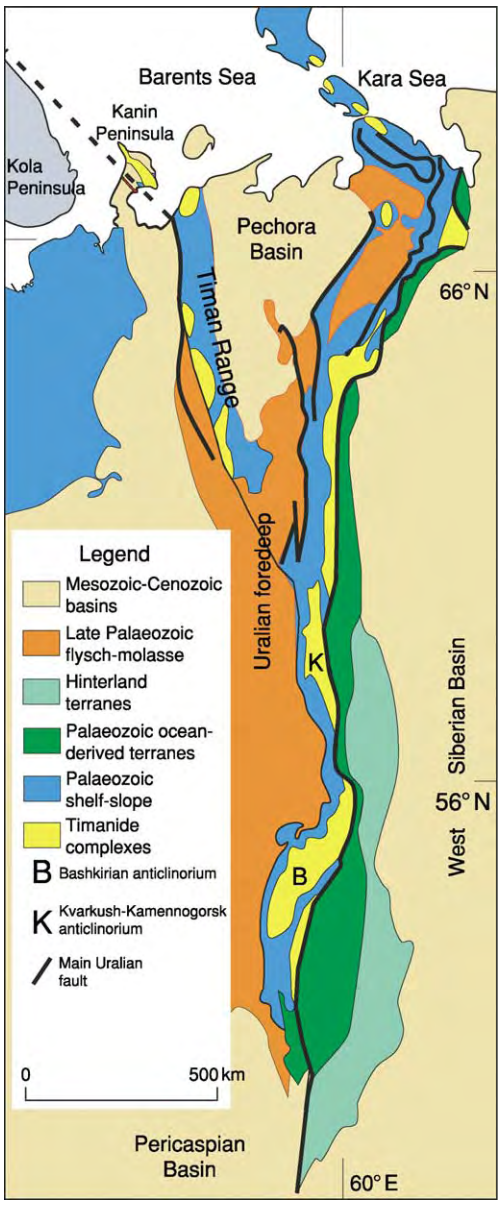


Figure 2 Timanide Complexes of the Urals and Timan Range.

edge of the Kola Peninsula to the Varanger Peninsula of northernmost Norway. Here, the Timanide Orogen is truncated by the mid-Palaeozoic, Scandian thrust front of the Scandinavian Caledonides and the latter is inferred to continue northwards through the Barents Sea to the northern edge of the Eurasian Shelf (Figure 1).

How far east the Timanides extended is not known. Timanian deformed and metamorphosed rocks occur throughout the western flank of Ural Mountains and into southern and central Novaya Zemlya. They dominate the Baltica margin complexes which underlie the main Palaeozoic allochthons of ocean-derived

ophiolites and island-arc volcano-sedimentary assemblages, and the latter, towards the east, are covered by Mesozoic successions of the West Siberian Basin.

The timing of the Timanian Orogeny is constrained by the ages of the youngest sedimentary successions involved in the Precambrian deformation and the oldest unconformably overlying sediments. The latter are generally Ordovician, but locally (beneath Kolguev Island and the Pechora Basin and on Novaya Zemlya) reach back into the Cambrian, the oldest strata (late Early Cambrian) occurring in central Novaya Zemlya. The youngest strata involved in the Timanian deformation include tillites in southern areas and contains Vendian acritarchs. Late Vendian successions in some Timanian foreland areas are developed in molasse facies. Thus, it can be inferred that the Timanian Orogeny occurred in the Vendian (Late Vendian in the front of the orogen; perhaps earlier in the hinterland) and lasted at least until the end of the Vendian, during uplift and erosion of the orogen, prior to Early Cambrian peneplanation and then the start of Palaeozoic platform deposition along the Baltica margin.

The different parts of the Timanide Orogen are summarized below, starting in the foreland fold-and-thrust belt and then continuing eastwards beneath the Pechora Basin to the metamorphic complexes of the Subarctic and Polar Urals and Novaya Zemlya.

Timanian Foreland Fold and Thrust Belt

Two very different Neoproterozoic sedimentary successions are preserved along the western front of the Timanide Orogen, the one in the Southern and Middle Urals and the other further to the north-west.

In the Timan Range and north-westwards to the Varanger Peninsula, thick turbidite-dominated successions occur in thrust sheets emplaced south-westwards onto platform facies, shallow marine carbonate (often stromatolite-bearing) and siliciclastic formations. These turbidites were apparently deposited along the margin of the East European Craton in continental slope and rise environments. These sedimentary rocks occur in upright to SW-vergent folds and are generally well preserved, with excellent characteristic sedimentary structures (graded bedding and Bouma sequences) and low grade of metamorphism (low greenschist facies). Only locally in the exposed thrust belt, on the Kanin Peninsula and northernmost Timan, are more deformed and metamorphosed sedimentary rocks present in the Timanide thrust sheets, providing evidence of the influence of regional high amphibolite facies metamorphism at depth within this part of the orogen. Blueschists have also been reported in the north-easternmost parts of the Kanin Peninsula. The turbidite-dominated successions are extensively intruded by pre-tectonic dolerite dykes and, in northern Timan, by an alkaline suite of gabbros, granites, and syenites, often nepheline-bearing, yielding zircon U/Pb ages of *ca*. 615 Ma.

To the south-west of the main Timanide deformation front, Late Vendian generally non-marine siliciclastic successions in the Varanger Peninsula, the Mezen Basin, and further south-east in the orogen were derived from the Timanide hinterland to the north-east, and are inferred to be molasse.

In the southern Urals (*see* **Europe:** The Urals), the Neoproterozoic turbidite facies is not exposed, and thick (up to 15 km) intracratonic successions were deposited from the beginning of the Mesoproterozoic to the Late Neoproterozoic, dominated by shallow-water siliciclastic and carbonate formations with some rift volcanics and sub-volcanic intrusions. These successions are called Riphean and are overlain by Early Vendian tillites and Late Vendian molasse, the latter a response to Timanian Orogeny further east.

The type area for Riphean stratigraphy is located in the core of a major Late Palaeozoic fold, the Bashkirian Anticlinorium, in the foreland fold-belt of the Urals (Figure 2). Beneath a major Ordovician unconformity, the folded and faulted Riphean and Vendian sedimentary rocks are overlain by a more deformed and, in part, more metamorphosed (amphibolite facies, locally with eclogites) allochthon, the Beloretsk Terrane. Neoproterozoic turbidites, similar to those in the Timan Range, may have been deposited in the areas at present occupied by the hinterland of the Southern Urals; if this was the case, they are now in the unexposed footwall beneath the Beloretsk Terrane.

Further north, in the middle Urals (Figure 2), another major Uralian fold, the Kvarkush-Kamennogorsk Anticlinorium, like the Bashkirian Anticlinorium (above), also contains Proterozoic successions below a post-Timanian, Ordovician unconformity. The siliciclastic and carbonate formations of this region are thought to be Neoproterozoic in age, reaching up into Early Vendian tillites and, probably, Late Vendian molasse. Of particular interest is the occurrence in the Kvarkush area of blue schists of Late Vendian (or possibly earliest Cambrian) age, thrust on to the other less metamorphosed, low greenschist facies sedimentary rocks. Ordovician quartzites overlie these Precambrian rocks and structures with major unconformity.

Basement of the Pechora Basin

From the Timan Range, the Timanide Orogen extends eastwards beneath the Palaeozoic and younger

strata of the Pechora Basin (Figure 2) to the foreland fold and thrust belt of the Polar Urals. The Phanerozoic cover thickens eastwards into the Uralian foredeep, where the unconformably overlying successions are 10–15 km thick and the character of the underlying basement is unknown. However, in some central parts of the Pechora Basin, along NW-trending axes of uplift, the basement reaches to within a few kilometres of the surface and deep drilling (about seventy holes) has succeeded in penetrating the entire cover succession to sample the basement. On the basis of this drilling, and also regional potential field (gravity and magnetic) anomaly maps, it has been possible to reconstruct a simple geological map of the pre-Palaeozoic rocks (Figure 3) beneath the Pechora Basin.

To the east of the Timan Range, three main tectonostratigraphic zones have been recognised. The western is referred to as the Izhma Zone and is a direct easterly extension of the Timan Range turbidite assemblages. Towards its eastern contact to the adjacent Pechora Zone, it is intruded by granites, mainly of calc-alkaline composition. The Pechora Zone is well seen on the potential field maps, it being well defined by a broad belt of NW-trending magnetic anomalies. Drillcores show that the Pechora Zone is dominated by volcanic and volcaniclastic rocks, extensively intruded by plutons, mainly of mafic and intermediate composition. Calc-alkaline granites also occur in this zone; together with those in the Izhma Zone, they yield ages of 550–560 Ma and apparently date the end of Timanian deformation and metamorphism in this part of the hinterland (Figure 4). Further east, the Bolshezemelskaya Zone also is preserved locally in structural highs, accessible to the drilling-rig. Acid volcanic rocks and granites have been sampled here from below the Ordovician unconformity, and somewhat older zircon ages of *ca* 570 and 620 Ma have been obtained from the latter, along with Grenville-age xenocrysts.

Previous literature, especially in Russia, contains several different names for the Bol'shezemel'skaya Terrane (e.g., Khoreyer and Novozemel'skaya) of continental blocks inferred (mostly on geophysical evidence) to exist at depth below the Pechora Basin and southern Barents Sea.

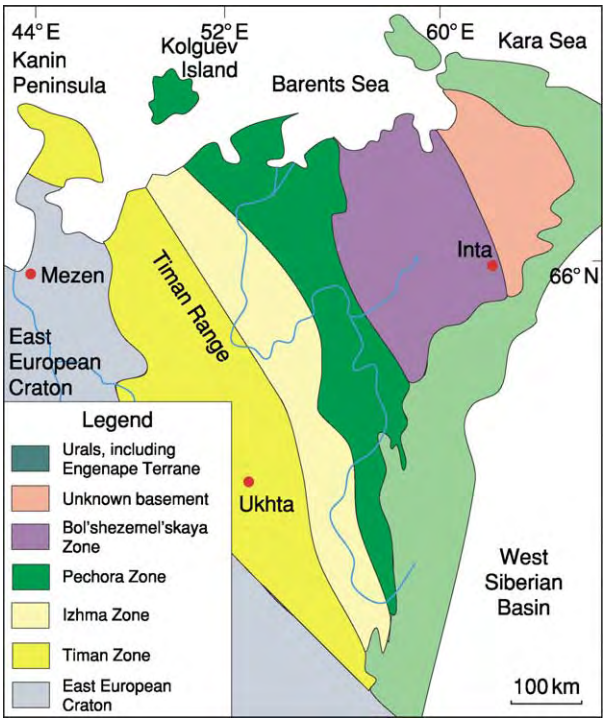


Figure 3 Pre Palaeozoic basement of the Timan Pechora region.

Pre-Ordovician Complexes of the Subarctic Urals

The Neoproterozoic complexes of the Pechora and Bol'shezemel'skaya zones, identified at depths of several kilometres beneath the Pechora Basin, strike south-eastwards into the mountain front of the Subarctic Urals. Within the Uralian frontal folds and lower thrust sheets, pre-Ordovician formations are exposed. Andesitic volcanic rocks dominate these Precambrian rocks in the northern Urals, along strike to the south from the Pechora Zone. Further to the north, in the subarctic Urals, a variety of granitic gneisses, granites, and metasediments occur, apparently as a southern continuation of the Bol'shezemel'skaya Zone. Isotopic ages are less reliable in these associations, but the presence of Mesoproterozoic metasediments and granites, and unconformably overlying Neoproterozoic siliciclastic (mainly quartzites) and carbonate formations, and acid volcanic rocks, are well documented. Early Ordovician conglomerates and quartzites overlie the Neoproterozoic successions with marked unconformity. Thus, it can be inferred that a Precambrian complex, differing in age and character from the Palaeoproterozoic and Archaean rocks of the EEC, occurs within the Timanide hinterland, outboard of the Pechora Zone calc-alkaline volcanic suites.

Polar Urals

North of the Arctic Circle in the Uralide Orogen, Neoproterozoic complexes comprise major components in the footwall to the Palaeozoic ocean-derived allochthons. In the core of foreland folds, beneath Ordovician quartzites, a fragmented ophiolite (Enganepe Terrane) has been dated by the U/Pb zircon method to 670 Ma, and is associated with arc

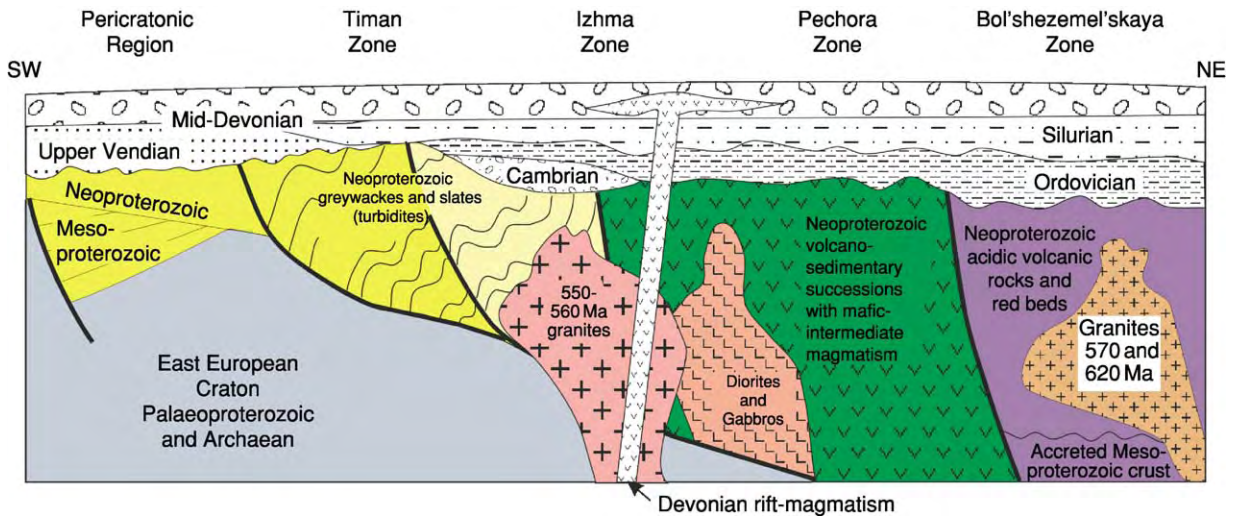


Figure 4 Diagramatic profile (SW NE) through the Timan Pechora region in the Devonian, prior to the Uralian Orogeny.

volcanites and a variety of volcaniclastic sedimentary rocks. Further east, in the Uralian thrust sheets, where the Palaeozoic metamorphic grade is higher, mafic island-arc related igneous complexes contain 580 Ma old gabbros, and, within the overlying Palaeozoic eclogite-bearing associations, similar ages characterise the igneous protoliths. Thus, in the most easterly exposed hinterland of the Timanide Orogen, the main Neoproterozoic component is derived from oceanic domains.

Novaya Zemlya

Northwards and north-eastwards from the Polar Urals, the Uralide Orogen continues via the Pai Khoi Peninsula, Vaigach Island, and the Kara Strait into Novaya Zemlya (Figure 1). Only the foreland folds and thrusts of the Uralide Orogen are preserved on land in Novaya Zemlya, the hinterland of the orogen being buried beneath the Mesozoic successions of the Kara Shelf.

As in the Polar Urals, major foreland anticlines and thrusts expose Neoproterozoic complexes beneath Lower Palaeozoic unconformities. On southern Novaya Zemlya and Vaigach Island, Lower and Middle Ordovician quartzites and limestones overlie greenschist facies metaturbidites. Further north, in the central parts of the archipelago (Figure 1), in the vicinity of the Mitushev Bay and Sulmenev Bay, Vendian (*ca* 600 Ma old) granites have been found, in the latter case intruded into schists, marbles, quartzites, and amphibolites. Whereas in the Mitushev Bay area, Silurian conglomerates rest unconformably on the granites, further north at Sulmenev Bay, both Ordovician and Cambrian strata overlie the unconformity, the oldest strata reaching back into the late Early Cambrian. Thus, in these northernmost hinterland areas, the unconformity may have been established in the Vendian. The amphibolite facies complex near Sulmenev Bay contains Mesoproterozoic detrital zircons; some authors infer that Precambrian orogeny there may have been older than Timanian (e.g., Grenville-age) and subsequently influenced by Timanian magmatism.

Barents Shelf

The Timanian rock units in the basement below the Pechora Basin strike north-westwards into the southern Barents Shelf and their associated geophysical (potential field) signatures can be followed towards the central parts of the Barents Sea, where they fade beneath the thickening Palaeozoic and Mesozoic successions of the Barents Shelf basins. Seismic studies, both near-vertical reflection and wide-angle refraction, have identified a thick wedge of Timanian metasediments along the southern margin of the Barents Sea towards the Kola Peninsula. Further out into the Barents Shelf, the identification of Timanian rocks is controversial and mainly based on the interpretation of seismic velocity data. Only on Franz Josef Land (Figure 1), is there clear evidence of a Timanide bedrock and even this is limited to a single deep drillhole located on the westernmost island, which penetrated a Lower Carboniferous unconformity to reach low greenschist facies, small folded Vendian meta-turbidites at 2–3 km depth.

Timanide Tectonic Evolution

Evidence of the Timanian Orogeny occurs over a wide area of north-eastern Europe. It provides a basis for reconstructing the Neoproterozoic to Early

Palaeozoic tectonic evolution of this part of Baltica; however, its fragmentary character and the extensive Phanerozoic cover of the Pechora and Barents Shelf basins, imply that these reconstructions are, at best, rudimentary. A major dilemma is the lack of evidence of the character of the Timanide hinterland to the east of the Urals – from the Palaeozoic ocean-derived allochthons and eastwards beneath the Mesozoic cover of the West Siberian Basin. In addition, it is probable that the Timanide Orogen extended both further south-eastwards, prior to the Uralian Orogeny, and north-westwards prior to the Caledonian Orogeny, but evidence is lacking.

There follows here a tectonic synthesis (Figure 5), based on the evidence presented above.

By the Early Neoproterozoic, after a long period (Mesoproterozoic) of intracontinental rifting, a passive margin had been established along the eastern edge of the EEC, at least from northern Norway to the Middle Urals. The change from an extensional to a compressional regime in the Vendian, with the onset of the Timanian Orogeny, is well defined in the foreland fold-and-thrust belt along this part of the EEC margin and extending southwards into the southern Urals. Deposition of a Late Vendian molasse facies in foreland basins along the length of the orogen provides evidence of the timing of hinterland uplift and foreland deformation.

The Timanian Ocean existing along the eastern margin (present-day coordinates) of the EEC in the

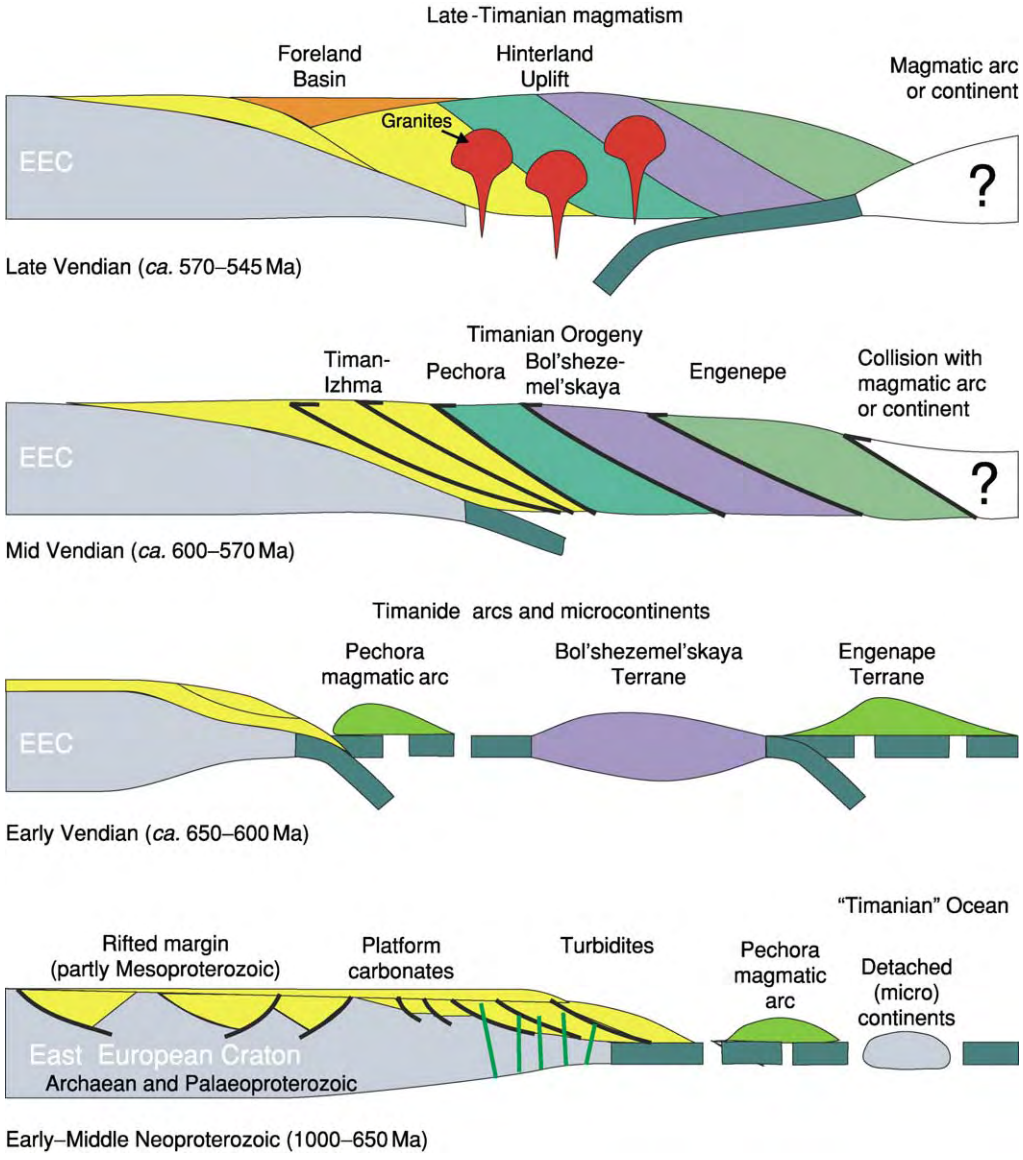


Figure 5 Timanian tectonic evolution, from the Early Neoproterozoic to the Late Vendian.

Neoproterozic, during deposition of the Timanian continental margin turbidites, development of the Pechora magmatic arc and Engane-Pe ophiolites, has been referred to in previous literature by several names, e.g., Palaeo-Asian, Turkistan, and Aegir, based on different interpretations of continent–ocean relationships. The latter are not well constrained and the informal term Timanian Ocean is preferred here.

Towards the hinterland, the dominance in the Pechora zone of mafic to intermediate calc-alkaline igneous suites and volcanogenic sedimentary rocks has been interpreted to imply thrusting of oceanic domains onto the thick passive margin turbidites of the EEC margin (Izhma Zone). The western contact of the Pechora Zone is not well defined, but the presence of Timanian blue schists along strike, both to the southeast, in the northern Urals (Kvarkush) and to the northwest, in the easternmost Kanin Peninsula, suggests that subduction-related arc magmatism culminated in thrust emplacement onto the EEC margin.

The existence of microcontinental terranes (blocks) in the Timanian Ocean has been widely inferred (indeed, some authors regard the Timanides as a number of NW-trending aulacogens with oceanic crust separating continental blocks). The microcontinental terranes differ from the EEC by having Grenville-age signatures with a Mesoproterozoic metasedimentary succession intruded by *ca*. 1000 Ma granites. The oceanic character of the adjacent Pechora Zone favours the interpretation that these outboard terranes (e.g., the Bol'shezemel'skaya Zone, and perhaps the Novaya Zemlya metamorphic rocks at Sulmenev Bay) were not accreted to the EEC until the Vendian.

To the east of the Bol'shezemel'skaya Zone, Late Neoproterozoic ophiolites in the front of the Polar Urals (Engane-Pe) and calc-alkaline intrusions in some of the Uralian allochthons further east, provide the only evidence of the oceanic character of the most internal parts of the Timanide hinterland. Taken together with the lack of evidence of the character of the Timanides east of the Urals, this has allowed the widely accepted hypothesis that Timanian Orogeny resulted from the subduction and accretion of Neoproterozoic ocean floor and island-arc assemblages, along with some fragments of continents (microcontinents), but without the involvement of a major Asian continent (e.g., Siberia). Continent–continent collision has not been favoured and the Neoproterozoic Timanian Ocean has been regarded as a forerunner of the Palaeozoic Uralian Ocean. This attractive hypothesis remains largely unconstrained.

Timanian orogenesis was contemporaneous with the passive margin evolution of north-western Baltica's Baltoscandian margin. It was also approximately contemporaneous with orogeny in many other parts of the world; for example, the Baikalian along the margin of the Siberian Craton, the Cadomian of western Europe and the Pan-African of many parts of Africa and the Middle East. Palaeomagnetic evidence has indicated that Baltica rotated about 120° counterclockwise between the Vendian and the Mid-Ordovician, prior to Caledonian collisional orogeny in the Silurian. This evidence has prompted a variety of palaeogeographic reconstructions relating the Timanides to these other orogens. It also implies that the rotation of Baltica may have accompanied the Timanian Orogeny, implying a significant component of dextral transpression during terrane accretion.

See Also

Europe: East European Craton; The Urals.

Further Reading

Belyakova LT and Stepanenko Vya (1991) Magmatism and geodynamics of the Baikalide Basement of the Pechora Syneclise. *Doklady Akademii nauk SSSR, (geologiya)* 106 117 (in Russian).

Bogatsky VI, Bogdanov NA, Kostyuchenko SL, Senin BV, Sobolev SF, and Khain VE (1996) *Tectonic map of the Barents Sea and the northern part of European Russia: explanatory notes*. Moscow: Institute of the Lithosphere, Russian Academy of Sciences Moscow.

Cocks LRM and Fortey RA (1998) The Lower Palaeozoic margins of Baltica. *Geologiska Foreningens Forhandlin ger* 120: 173 179.

Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomag netic review. *Journal of the Geological Society, London* 159: 631 644.

Dushin VA (1997) Magmatism and geodynamics of the palaeocontinental sector of the northern part of the Urals. *Moscow Nedra*. pp. 211 (in Russian).

Gee DG and Pease VL (2005) The Neoproterozoic Tima nide Orogen of Eastern Baltica. *Geological Society of London, Memoir*.

Gee DG, Belyakova LT, Pease V, Larionov AN, and Dovzhikova E (2000). New, single zircon (Pb evapo ration) ages from Vendian intrusions in the basement beneath the Pechora Basin, northeastern Baltica. *Polar forschung* 68: 161 170.

Korago EA, Kovaleva GN, Ilin VF, and Pavlov LG (1992) *Tectonics and metallogeny of the early Kimmeridgian of Novaya Zemlya,* pp. 1 196. St Petersburg: Nedra (in Russian).

Lopatin BG, Pavlov LG, Orgo VV, and Shkarubo SI (2001) Tectonic structure of Novaya Zemlya. *Polarforschung* 69: 131 135.

Olovyanishnikov VG, Roberts D, and Siedlecka A (2000) Tectonics and sedimentation of the Meso to Neoproter ozoic Timan Varanger Belt along the northeastern margin of Baltica. *Polarforschung* 68: 269 276.
Puchkov VN (1997) Structure and geodynamics of the Uralian orogen. *Geological Society of London Special Publications* 121: 201 236.
Roberts D and Siedlecka A (2002) Timanian orogenic deformation along the northeastern margin of Baltic, Northwest Russia and Northeast Norway, and Avalonian Cadomian connections. *Tectonophysics* 352: 169 184.
Siedlecka A (1975) Late Precambrian stratigraphy and structure of the northeastern margin of the Fennos candian Shield (East Finnmark Timan region). *Norges geologiske undersokelse* 316: 313 348.
Torsvik TH and Rehnström EF (2001) Cambrian palaeo magnetic data from Baltica: implications for true polar wander and Cambrian palaeogeography. *Journal of the Geological Society, London* 158: 321 329.
Willner AP, Ermolaeva T, and Stroink L (2001) Contrasting provenence signals in Riphean and Vendian sandstones in the SW Urals (Russia): constraints for a change from passive to active continental margin conditions in the Neoproterozoic. *Precambrian Research* 110: 215 239.

Caledonides of Britain and Ireland

J F Dewey, University of California Davis, Davis, CA, USA, and University of Oxford, Oxford, UK
R A Strachan, University of Portsmouth, Portsmouth, UK

Introduction

The term 'Caledonian' is derived from the Latin word for Scotland and has been used widely with reference to the Early to Mid-Palaeozoic orogenic belt that is exposed in the British Isles, eastern Greenland, and Scandinavia (**Figure 1**). When these areas are restored to their relative positions prior to the Cretaceous–Tertiary opening of the North Atlantic (**Figure 1**), it can be seen that they were formerly continuous, comprising the northern part of the Mesozoic supercontinent Pangaea. The Caledonian Orogeny is the oldest of three main Palaeozoic collisional events that formed Pangaea, the other two being the Variscan and Uralian orogenies.

The Caledonian orogenic belt has three main arms (**Figure 1**): the North Atlantic Caledonian Belt and its southern extension, the Appalachian Belt, which were both formed by the closure of the Iapetus Ocean, and the Tornquist Belt, which resulted from the closure of the Tornquist Sea. These belts separate areas of Archaean and Proterozoic crust, which formed the palaeocontinents of Laurentia (North America, Greenland, and the northern British Isles), Baltica (Scandinavia and the Baltic), and Eastern Avalonia (the southern British Isles, Belgium, and the adjacent crust). Eastern Avalonia is the northernmost of a group of continental fragments that were originally derived from Gondwana and include Western Avalonia, Iberia, and Armorica. The collision of these fragments through the Mid- to Late Palaeozoic resulted in the Variscan orogenic belt (*see* **Europe:** Variscan Orogeny), which just impinges on the south of the British Isles.

Palaeogeographical and Tectonic Framework

The Caledonian plate-tectonic cycle began in the Mid- to Late Neoproterozoic with the rifting and break-up of the supercontinent Rodinia and the subsequent dispersion of continental blocks to form the Iapetus Ocean. Palaeomagnetic studies indicate that, in the Late Cambrian to Early Ordovician, Laurentia occupied an equatorial position to the north of Baltica and Avalonia (the latter was still attached to Gondwana at this stage). The distance across the Iapetus Ocean between the locations of present-day Scotland and England was probably about 5000 km, and the width of the Tornquist Sea, which separated Gondwana from Baltica, was probably about 1300 km. Each continent was characterized by a distinctive trilobite fauna. The initial closure of Iapetus was associated with the development of a southwards-dipping intraoceanic subduction zone and volcanic arc off the eastern Laurentian margin. The collision during the Early to Mid-Ordovician between the Laurentian margin and the volcanic arc resulted in the Grampian orogenic event (**Figure 2**). Subduction polarity then reversed to dip northwards beneath the Laurentian margin.

Eastern Avalonia rifted from Gondwana in the Arenig and drifted rapidly northwards between 480 Ma and 460 Ma to reach a similar latitude to Baltica during the Caradoc. A southwards-dipping subduction zone was located along the northern Iapetan margin of Eastern Avalonia. As on the Laurentian margin, a suprasubduction zone ophiolite–arc

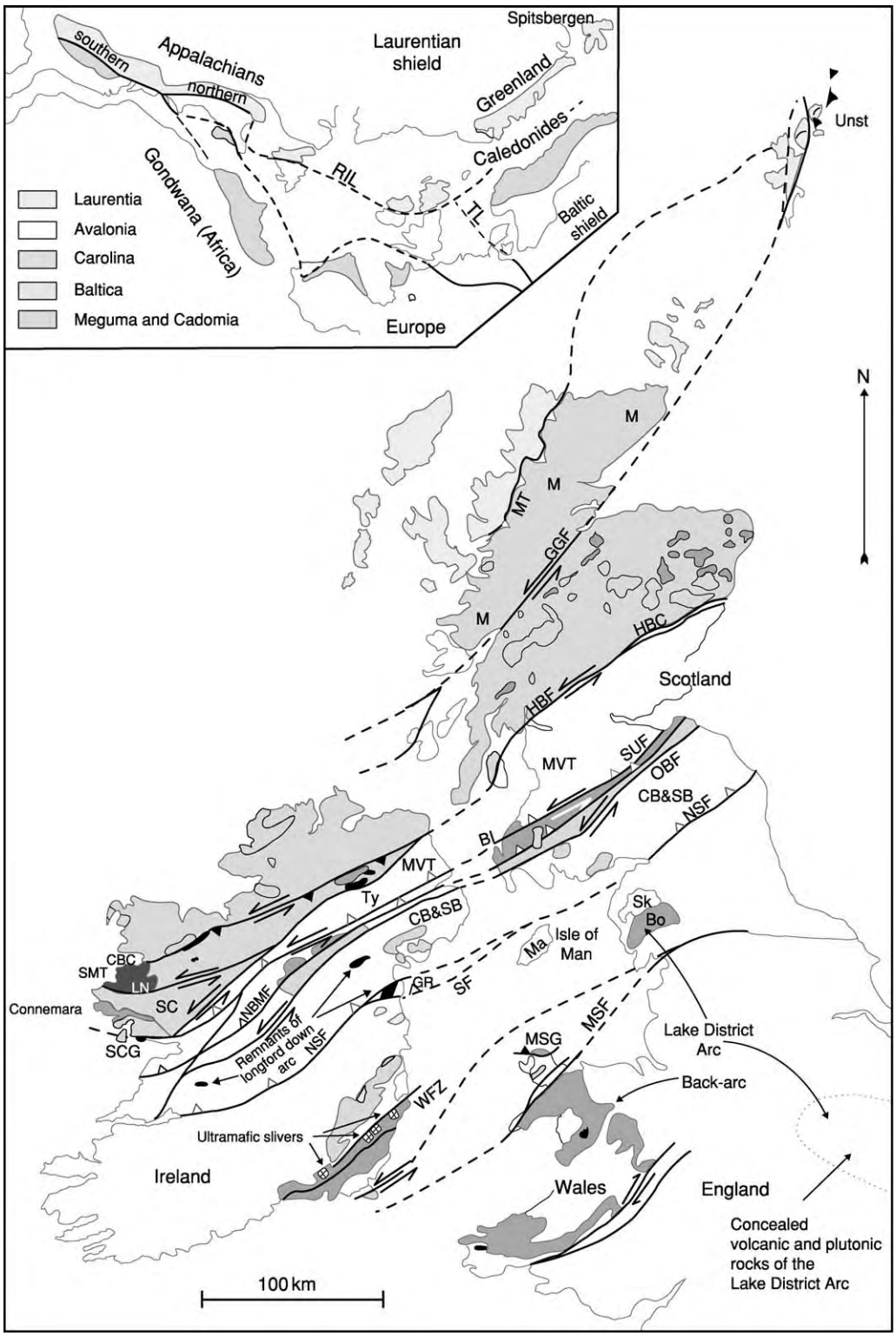


Figure 1 Simplified Caledonian tectonic map of the British Isles with their Appalachian continuation shown in the inset. BI, Ballantrae Ophiolite; BO, Borrowdale Volcanics; CB, Central Belts; CBC, Clew Bay Complex; GGF, Great Glen Fault; GR, Grangegeeth Terrane; HBC, Highland Border Complex; HBF, Highland Boundary Fault; LL, Leadhills Line; M, Moine; Ma, Manx Group; MSF, Menai Straits Fault; MSG, Monian Supergroup; MT, Moine Thrust; MVT, Midland Valley Terrane; OBF, Orlock Bridge Fault; RIL, Red Indian Line; SC, South Connemara Terrane; SCG, South Connemara Group; Sk, Skiddaw Group; SL, Slane Fault; SMT, South Mayo Trough; SUF, Southern Uplands Fault; TL, Tornquist Line; Ty, Tyrone Ophiolite; WFZ, Wicklow Fault Zone.

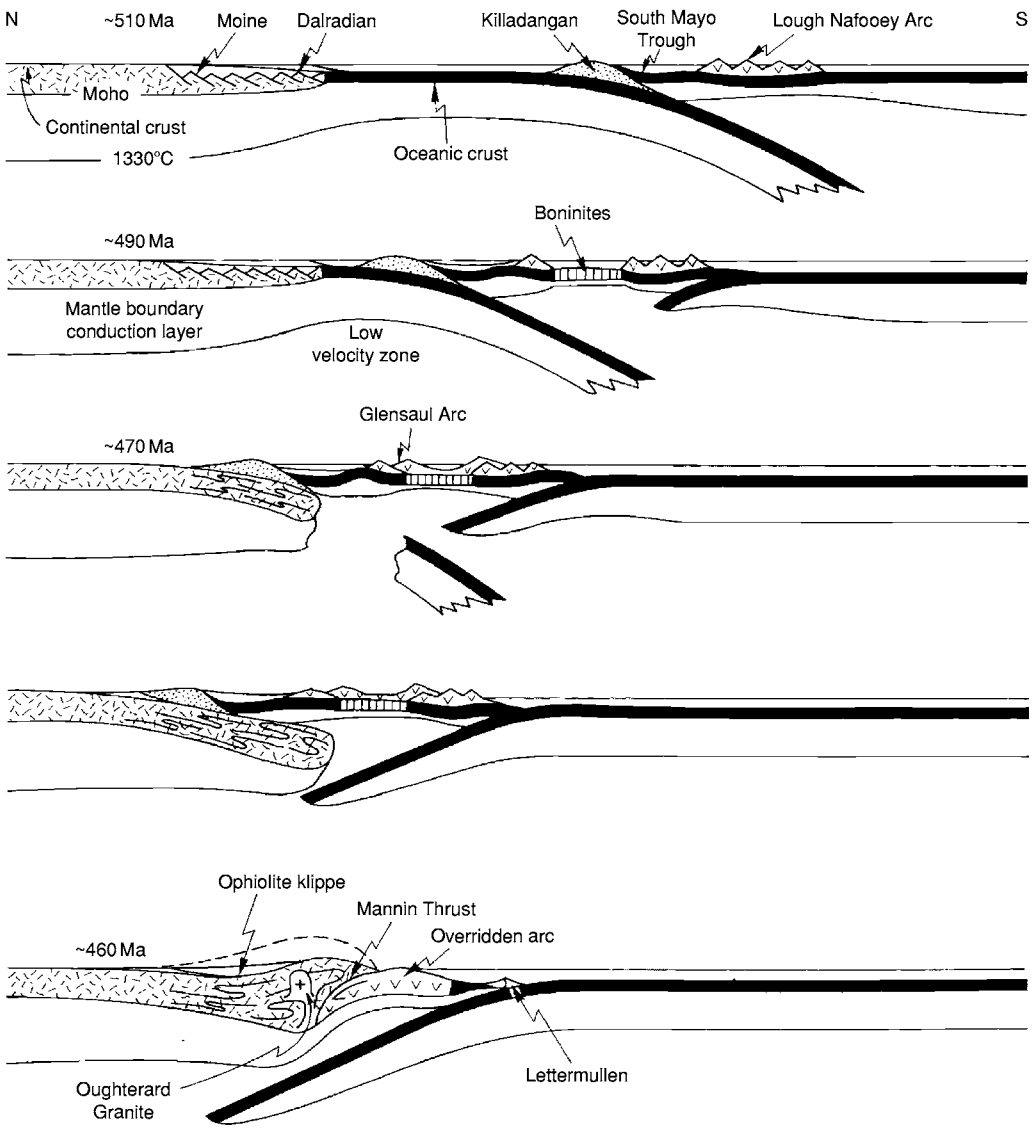


Figure 2 Schematic sections illustrating the Ordovician tectonic evolution of the Grampian Zone.

complex was obducted onto the Eastern Avalonian margin during that margin's collision with a northwards-dipping subduction zone, followed by polarity flip to southwards-directed subduction. The oceanic tract that opened between Eastern Avalonia and Gondwana is known as the Rheic Ocean. At 460 Ma, Eastern Avalonia was probably about 1000 km west of Baltica, and the width of the Iapetus Ocean was reduced to approximately 2000 km across the British sector. The intermediate position of Avalonia during its transit across the closing Iapetus Ocean is reflected in the changing affinities of its brachiopods, from Gondwanan to a mixture of genera with ancestors from Baltica and Laurentia, through the Ordovician. By Late Ordovician–Early Silurian times (440 Ma), the Iapetus Ocean and the Tornquist Sea had narrowed sufficiently to allow the interchange of the larvae of most benthic animals. Eastern Avalonia collided with Baltica during the Ashgill. By this time, the width of the Iapetus Ocean across the British sector had been reduced to around 1300 km.

Sinistrally oblique collision between the combined Avalonia–Baltica landmass and Laurentia occurred between 435 Ma and 425 Ma to form the continent Laurussia. The 'hard' collision of Baltica with the eastern Greenland–northern Scotland segment of Laurentia resulted in the Scandian orogenic event. In contrast, the 'soft' collision of Eastern Avalonia with Laurentia was not associated with regionally

significant shortening, suggesting that Eastern Avalonia and Baltica may have been only loosely coupled during the final collision. The Iapetus Suture is referred to as the Solway Line, because it is apparently located in the Solway Firth, albeit obscured entirely by unconformably overlying Carboniferous strata.

Between 425 Ma and 410 Ma, sinistral relative motion between Laurentia and Baltica–Avalonia became orogen-parallel and was associated mainly with displacement along the Great Glen, Highland Boundary, and Southern Uplands faults. From about 410 Ma, the Caledonides went into regional sinistral transtension, resulting in the formation of a series of transtensional Old Red Sandstone basins. These events were contemporaneous with the rapid contraction of the Rheic Ocean as Gondwana approached Laurussia. During the Early to Mid-Devonian, Eastern Avalonia was affected by widespread Acadian deformation. Although this has traditionally been assigned to the Caledonian orogenic cycle and the closure of Iapetus, it postdates the latter by *ca.* 20 Ma. Alternatively, the Acadian event could be the result of the collision of Armorica with the southern margin of Avalonia. The collision of Armorica with Avalonia and the collision of Iberia with the southern margin of Armorica during the Mid-Devonian could be viewed as the early stages of the protracted Variscan Orogeny, which ultimately resulted in the assembly of Pangaea.

The along-strike continuation of the Caledonides in Newfoundland provides, arguably, the best-exposed section across the orogen, which can be used as a template to aid the understanding of other sectors such as the British Isles. The Caledonides of the British Isles are subdivided by major faults into a series of terranes, each with a coherent internal history. The correlation of geological units and events across terrane-bounding faults is commonly problematic. These structures display evidence of long and complex reactivation histories, although the last major displacement for most faults north of the Iapetus Suture was during Silurian to Devonian time. Although the detail of the tectonic template is extremely complex, there are three main groups of terranes: those north of the Fair Head–Clew Bay Line and Highland Boundary Fault, which have Laurentian affinities; those south of the Solway Line, which have Gondwanan affinities; and the intermediate terranes, which represent fragments of continental margin, island arc, and oceanic lithosphere within a composite complex suture zone. Seismic imaging indicates that the Iapetus Suture dips moderately to the north, separating reflective Avalonian crust from overthrust non-reflective Laurentian crust.

Terranes and their Geotectonic Affinities

From north to south, the main terranes are characterized as follows.

Hebridean Terrane

In north-west Scotland, the Hebridean Terrane forms the foreland to the Caledonian belt and is limited to the east by the Moine Thrust Zone. The oldest component is the Archaean–Palaeoproterozoic Lewisian Gneiss Complex, which is correlated with Laurentian basement in south-east Greenland. The unconformably overlying continental Torridonian sedimentary rocks were deposited from 1200–1000 Ma and comprise two successions separated by an angular unconformity. The Lewisian and Torridonian rocks are overstepped by a highly distinctive succession of Cambrian–Ordovician shallow-marine quartzites and limestones, which accumulated on the Laurentian margin of Iapetus and correlate closely with similar sequences in north-western Newfoundland and eastern Greenland.

Northern Highland Terrane

The Northern Highland Terrane is bounded in north-western Scotland by the Moine Thrust Zone and the Great Glen Fault. The terrane is dominated by the Early Neoproterozoic Moine Supergroup, a thick succession of strongly deformed and metamorphosed marine sediments that were deposited in an intracratonic rift within the Laurentian sector of the Rodinia Supercontinent at 1000–870 Ma. Early deformation and metamorphism during a Knoydartian tectonothermal event at 800 Ma may have resulted from rift closure. Inliers of Archaean orthogneisses, variably reworked at 1000 Ma during the Grenville orogeny, represent the basement on which the Moine sediments were unconformably deposited.

Grampian Terrane

In Scotland and north-western Ireland, the Grampian Terrane occurs between the Great Glen Fault and the Fair Head–Clew Bay Line or Highland Boundary Fault. The terrane is dominated by the Dalradian Supergroup, a thick succession of variably deformed and metamorphosed metasandstones, metasiltstones, and metalimestones with minor contemporaneous mafic igneous rocks. The duration of sedimentation is uncertain, but it probably spanned at least the period 750–520 Ma. Tillites have been correlated with the widespread Sturtian and Varangerian glacial events. The Dalradian succession shows a transition from shallow-water to deep-water sedimentation,

reflecting the progressive thinning of the Laurentian crust during rifting of the Rodinia Supercontinent and the evolution of a passive continental margin adjacent to the developing Iapetus Ocean. Inliers of Palaeoproterozoic gneisses in the Inner Hebrides and north-western Ireland can be linked with similar units in south-eastern Greenland and probably represent the Laurentian basement on which the Dalradian sediments were unconformably deposited. Shear zone-bounded inliers of Early Neoproterozoic Moine-like rocks occur in the northern Grampian Highlands and may provide linkage across the Great Glen Fault.

Midland Valley Terrane

The Midland Valley Terrane is bounded to the north by the Fair Head–Clew Bay Line and the Highland Boundary Fault and to the south mainly by the Southern Upland Fault (in the west of Ireland it lies north of the Doon Rock Fault). Much of the pre-Late Palaeozoic geology, particularly in the Scottish Midland Valley, is obscured by Devonian and Carboniferous cover. However, the exposed Early Palaeozoic rocks have oceanic affinities. Early Ordovician ophiolites occur on Shetland, along the Highland Boundary Fault, and at Ballantrae, Tyrone, and Clew Bay. In western Ireland, remnants of an accretionary prism are present in Clew Bay and on Achill Island, locally associated with blueschists. Immediately to the south are the fore-arc Ordovician sediments of the South Mayo Trough and the mafic to intermediate calc-alkaline volcanics of the Lough Nafooey Group. Whether the unexposed continental basement that underlies the terrane has Laurentian affinities or is entirely exotic is unknown.

Connemara Terrane

The Connemara Terrane is located in western Ireland between the Doon Rock and Skirds Rock faults. It consists of metasedimentary rocks that have been correlated with the Dalradian Supergroup, and thus the terrane is interpreted as a tectonic slice of the Laurentian margin that became detached and interleaved with units of the Midland Valley Terrane during strike-slip displacements.

Southern Uplands Terrane

Located between the Southern Uplands Fault and the Solway Line, the Southern Uplands Terrane is characterized by thick sequences of Ordovician and Silurian sediments, which were deposited in deep-marine environments and subjected to complex deformation and low-grade metamorphism. The terrane itself comprises three major units bounded by strike-slip faults: the Northern, Central, and Southern belts. The Northern Belt comprises Ordovician sediments and rare volcanics; the Central Belt is formed of Ordovician and Silurian sediments, and the Southern Belt is entirely Silurian in age. Although opinions vary as to the tectonic setting of sedimentation, the balance of evidence suggests that the sediments represent an accretionary prism developed above the northwards-subducting Iapetus Ocean. In eastern Ireland, the Iapetus suture is drawn between the Ordovician inliers of the Grangegeeth and Bellewstown terranes, which show Laurentian and Avalonian affinities, respectively, in their brachiopod and trilobite faunas.

Lake District–Leinster Terrane

The Lake District Terrane is situated between the Solway Line and an unexposed boundary thought to lie to the north-west of Anglesey. It contains Cambrian–Arenig clastic rocks, mid- to Late Ordovician basic and acid volcanic and volcaniclastic rocks, such as the Borrowdale Volcanic Group of the English Lake District, and the unconformably overlying Silurian sediments of the Windermere Supergroup of northern England. The Lake District represents the northern edge of Eastern Avalonia, and the volcanic rocks are thought to have originated in a volcanic arc created by southwards-directed subduction on the Gondwanan side of the Iapetus Ocean.

Monian Terrane

The Monian Terrane is bounded to the south-east by the Menai Strait Fault System and comprises the Mona Complex of Anglesey and the Llyn Peninsula in north-western Wales and the Rosslare Complex of south-eastern Ireland. The terrane is highly segmented by shear zones and brittle faults. Proven Precambrian rocks include high-grade metasedimentary gneisses, a calc-alkaline granite pluton, and a belt of blueschist facies metamorphic rocks, which may represent an accretionary prism. Thick low-grade metasedimentary sequences, including a major melange unit, have been viewed as Precambrian but may be Cambrian, albeit very different from the Cambrian rocks of the adjacent Welsh Basin Terrane. The Monian Terrane may be an Allochthonous Gondwanan Terrane that was amalgamated with Eastern Avalonia by the Early Cambrian at the latest, after which it formed part of the southern foreland of the orogen.

Welsh Basin Terrane

Bounded to the south-east by the Welsh Borderland Fault System, the Welsh Basin Terrane comprises Cambrian–Silurian sedimentary successions, largely deposited in deep-water environments, and thick sequences of Ordovician tholeiitic to calc-alkaline

volcanics. Volcanism and sedimentation probably occurred in a back-arc basin floored by extended Avalonian basement, behind the Lake District–Leinster volcanic arc. A range of Ordovician calc-alkaline volcanic and plutonic rocks underlie the Mesozoic and younger cover of eastern England and may represent continuations of the Welsh Basin and/or Lake District volcanic belts.

Midlands Terrane

The Midlands Terrane is a triangular terrane that represents the southern foreland of Eastern Avalonia during the Caledonide Orogen and acted as a rigid indentor during the Acadian event. Late Neoproterozoic calc-alkaline igneous rocks and associated low-grade metasedimentary rocks, which developed along an active plate margin of Gondwana, are overlain unconformably by relatively thin sequences of Cambrian–Silurian shallow-marine platform sediments.

Tectonic Evolution of the Caledonides

The tectonic evolution of the Caledonides began in the Early to mid-Ordovician, with the commencement of closure of the Iapetus Ocean, and finished with the Early Devonian Acadian event. The main stages are as follows.

Early to mid-Ordovician Rifting of Eastern Avalonia and Arc–Continent Collision on the Laurentian Margin

Eastern Avalonia rifted away from Gondwana in the Early Ordovician (*ca*. 475 Ma), initiating closure of the Iapetus Ocean and opening, to the south, of the Rheic Ocean. A major southwards-dipping (in the present reference frame) subduction zone developed beneath the leading northern margin of Eastern Avalonia, and arc volcanism began in the Welsh Basin in the Early Ordovician. Sedimentary sequences deposited on Eastern Avalonia vary from shallow-marine on upstanding blocks such as the Midlands terrane and the Irish Sea landmass (effectively the Monian terrane) to deeper-marine sequences in the Welsh Basin.

By the Late Cambrian–Early Ordovician, oceanwards-dipping subduction zones had developed along the margins of Laurentia and Baltica, to form intraoceanic volcanic arcs. Localized orogenic events reflect the collision of these volcanic arcs with the continental margins. These include the Finnmarkian event (505 Ma) along the Baltica margin and the Grampian event (470–460 Ma), which affected the Northern Highlands and Grampian terranes. In western Ireland, the Grampian Arc is represented by the Lough Nafooey Volcanic Group, and the existence of a similar arc beneath the Late Palaeozoic cover of the Scottish Midland Valley is indicated by the presence within the Ordovician sediments of clasts of contemporaneous calc-alkaline plutonic rocks derived from a proximal source. By analogy with a similar-aged event in Newfoundland (the Taconic event), collision of the Laurentian margin with the arc at the subduction zone was associated with northward obduction of a major ophiolite nappe onto the Laurentian margin. Ophiolitic remnants of this nappe lie along the north-western margin of the Midland Valley Terrane, which effectively represents the Grampian Suture; the Ballantrae Ophiolite was probably obducted at about the same time. Within the Grampian Terrane, ophiolite obduction resulted in regional deformation and Barrovian metamorphism at greenschist–amphibolite facies of the Dalradian Supergroup. In north-eastern Scotland and Connemara, this was accompanied by the emplacement of major gabbros and granites. Late-stage backfolding and underthrusting of the Dalradian Supergroup formed south-eastwards-directed regional-scale fold nappes such as the Tay Nappe in Scotland and the Ballybofey Nappe in north-western Ireland. Within the Northern Highland Terrane, similar-aged folding and high-grade metamorphism of the eastern parts of the Moine Supergroup is assigned to the Grampian event. Following arc–continent collision, the polarity of subduction reversed to northward.

Mid-Ordovician–Silurian: Collision of Eastern Avalonia, Baltica, and Laurentia

On the Avalonian side of Iapetus, south-eastwards-directed subduction beneath Avalonia continued into the Late Ordovician. A major calc-alkaline volcanic arc lay through south-east Ireland and the Lake District, with the Welsh Basin being a volcanically active back-arc basin. Volcanism and subduction ceased during the Late Ordovician, probably because the Avalonian margin overran the ridge system in the Iapetus Ocean. Throughout the Late Ordovician and Silurian, the stable Midland Platform and Irish Sea landmass hosted shallow-marine or emergent conditions surrounded by deeper-marine basins, which were often filled with turbidites. Eastern Avalonia collided with Baltica during the Late Ordovician (440 Ma) as shown by palaeomagnetic data and evidence of progressive faunal integration. Southwestern Baltica records evidence of a Late Ordovician thermal event, contemporaneous with open folding and uplift in the Welsh Borderlands (the Shelveian event). This event is also synchronous with low-grade metamorphism in the North Sea, which was

probably also related to the collision between Baltica and Eastern Avalonia.

On the northern side of Iapetus, the uplands created by the Grampian Orogeny remained emergent through the Late Ordovician and Silurian. In the Midland Valley Terrane, the former volcanic arc was mostly buried under a blanket of marine and nonmarine sediments, most of which were sourced from the emergent Grampian Highlands to the north. Sedimentation occurred in a series of localized sinistrally transtensive basins. Outboard, to the south-east, the Southern Uplands Terrane is dominated by thick fault-bounded sequences of complexly deformed trench sediments, which define an accretionary prism. The dominant folds, cleavages, and associated detachments within these sediments formed progressively as an integral part of the development of the accretionary prism, rather than as the result of a distinct collisional event. A switch from early approximately orthogonal shortening to sinistral transpression within the accretionary prism occurred in the Early Silurian at 430 Ma.

The earliest linkage of Laurentian-derived and Gondwanan-derived crustal blocks occurred in Newfoundland in the mid-Ordovician, at approximately 450 Ma. However, the earliest closure events in the British Isles appear to be Silurian (*ca*. 425 Ma). With progressive collision of the two margins of Iapetus through the Silurian, deep-water basins tended to shallow whilst shallow-marine basins became nonmarine or emergent. In the Lake District, the Silurian Windermere Supergroup was deposited in a flexural foreland basin that resulted from overthrusting of the Southern Uplands Terrane onto the leading edge of Eastern Avalonia during the terminal stages of collision. The oblique collision between Eastern Avalonia and Laurentia along the Solway Line, which resulted in the final eradication of the Iapetus Ocean, was complete by the end of the Silurian. The collision here was relatively 'soft' and did not result in major folding or cleavage development in the immediately adjacent terranes. In western Ireland, the collision was 'harder' and was associated with sinistral transpression, which deformed the Ordovician and Silurian strata of the South Mayo Trough.

Silurian: Oblique Collision of Laurentia and Baltica, and Closure of Northern Iapetus

The Northern Highlands Terrane was probably located 700 km along strike and to the north-east of the Grampian Terrane when this part of Laurentia collided obliquely with Baltica at 430–425 Ma to result in the Scandian orogenic event. This was associated with regional nappe stacking in Scandinavia and eastern Greenland, with intense folding, ductile thrusting, and metamorphism of the Moine rocks in the Northern Highlands Terrane, and with amalgamation of the Northern Highlands and the Hebridean terranes along the Moine Thrust. A total north-west–south-east shortening displacement across the Moine Thrust of at least 150 km seems likely, and an equivalent total amount of shortening may also have occurred along internal ductile thrusts within the Moine. Analysis of Scandian structures in Scandinavia, eastern Greenland, and north-west Scotland is consistent with overall sinistral transpression, which became progressively partitioned into orogen-orthogonal components and orogen-parallel left-lateral strike slip.

Regional-scale nappe stacking was therefore followed by major sinistral strike-slip displacements from 425 Ma onwards along the Great Glen, Highland Boundary, and Southern Uplands faults. The magnitude of the displacements is uncertain, although movement of 700 km along the Great Glen Fault and minimum movements of 200 km along both the Highland Boundary Fault and the Southern Uplands Fault seem likely. Strike-slip faulting was largely synchronous with the emplacement of the Newer Granite Suite, a series of calc-alkaline mainly I-type plutons probably derived from the melting of lithospheric mantle and lower-crustal sources above the north-westwards-dipping subduction zone. Major plutons crop out extensively in the Grampian Highlands and in north-western Ireland, and eruption of coeval lavas and volcanics occurred in the area between Glencoe and Lorne in the south-west Grampian Highlands. Plutons were mainly emplaced in transtensional pull-aparts along relatively minor faults that were splays of, or Reidel shears related to, the major displacement faults.

Early to mid-Devonian: the Final Caledonian (Acadian) Collision – Closure of the Rheic Ocean?

Sinistral relative displacement of the terranes north of the Solway Line and subduction-related plutonism and volcanism continued into the Early Devonian. By this time, the marine basins on both margins had been largely uplifted following crustal shortening. In the Scottish Highlands, Early Devonian continental sediments were deposited unconformably on deeply eroded Moine and Dalradian rocks. The Early Devonian sediments of the Moray Firth–Orkney–Shetland area and eastern Greenland probably accumulated in transtensional basins, reflecting a change from orogen-parallel displacements to sinistrally oblique divergence between Laurentia and Baltica–Eastern Avalonia between 410 Ma and 395 Ma.

A similar transtensional setting may also apply to the Early Devonian basins of the Midland Valley and the Lake District. Fluvial and alluvial sedimentation also occurred on the Irish Sea landmass and in the Anglo-Welsh Basin, with major rivers sourced from uplifted segments of the Caledonides to the north and north-east.

In southern Britain, the Cornubian basins record south-deepening Devonian marine sedimentation on the southern rifted passive margin of Avalonia, which faced Armorica across an already very narrow arm of the Rheic Ocean. East–west trending Early Devonian growth faults have been identified in South Wales and North Devon. In southern Cornwall, Devonian sediments are characterized by thick sequences of distal turbidites, which were deposited either on highly thinned continental crust at the edge of the shelf or on oceanic crust. The ultramafic and associated rocks of the Lizard Complex in Cornwall have been interpreted as an ophiolite derived from the Rheic Ocean basin, which is inferred to have existed to the south.

During the Early to Mid-Devonian, widespread Acadian deformation and low-grade metamorphism occurred south of the Solway Line, forming the slate belts of the Lake District, south-eastern Ireland, and north-western Wales. In contrast to the area north of the Solway Line, where the Caledonian structural grain is mainly north-east–south-west, the trend of folds and cleavage in Avalonia is much more variable, showing a distinctly arcuate form. This structural pattern is apparently strongly controlled by the triangular shape of the old rigid continental block that underlies the Midlands terrane, which did not experience Caledonian deformation. The Welsh Borderlands Fault System broadly defines the south-eastern limit of Acadian deformation. There is a common tendency in the slate belts for the steeply dipping Acadian cleavage to cut gently across associated folds in a clockwise-deflected sense, indicating that the collision direction was sinistrally oblique. Although the Acadian phase of the Caledonian Orogeny has been ascribed to the closure of Iapetus, it is not clear why the shortening and uplift associated with this event should occur as late as the Early to Mid-Devonian. It is unlikely that forceful northward collision of Avalonia continued for a further 20 Ma after the last remnant of Iapetan crust was subducted, especially as there is growing evidence for sedimentation in sinistrally transtensive basins at this time, both south and north of the suture. Another possibility is that the Acadian deformation resulted from the collision of Armorica with Avalonia in the Early Devonian. This is supported by isotopic data that indicate that initial exhumation and northward-directed thrusting of the Lizard Complex in southern Cornwall occurred between 400 Ma and 380 Ma, overlapping in time with cleavage development in the slate belts to the north.

See Also

Europe: Scandinavian Caledonides (with Greenland); Variscan Orogeny; The Urals. **North America:** Northern Appalachians; Southern and Central Appalachians. **Palaeozoic:** Ordovician; Silurian; Devonian. **Pangaea**.

Further Reading

Dewey JF and Mange MA (1999) Petrology of Ordovician and Silurian sediments in the western Irish Caledonides: tracers of short lived Ordovician continent arc collision orogeny and the evolution of the Laurentian Appalachian Caledonian margin. In: MacNiocaill C and Ryan PD (eds.) *Continental Tectonics:* Geological Society, London, Special Publication 164, pp. 55 108.

Dewey JF and Shackleton RM (1984) A model for the evolution of the Grampian tract in the early Caledonides and Appalachians. *Nature* 312: 115 121.

Draut AE and Clift PD (2001) Geochemical evolution of arc magmatism during arc continent collision, South Mayo, Ireland. *Geology* 29: 543 546.

Soper NJ, Ryan PD, and Dewey JF (1999) Age of the Grampian Orogeny in Scotland and Ireland. *Journal of the Geological Society, London* 156: 1231 1236.

Van Staal CR, Dewey JF, MacNiocaill C, and McKerrow WS (1998) The Cambrian Silurian tectonic evolution of the northern Appalachians and British Caledonides: history of a complex, west and southwest Pacific type segment of lapetus. In: Blundell DJ and Scott AC (eds.) Lyell: the past is the key to the present: Geological Society [London] Special Publication, 143, pp. 199 242.

Scandinavian Caledonides (with Greenland)

D G Gee, University of Uppsala, Uppsala, Sweden

Introduction

The Caledonide Orogen of the North Atlantic region reaches northwards from type areas in the British Isles, along the eastern edge of Greenland and western Scandinavia, to the Barents Shelf and the Svalbard Archipelago. Prior to the opening of the Norwegian and Greenland seas and the Eurasian Basin in the Tertiary (Figure 1), this *ca*. 3000 km long segment of the orogen was about 1000 km wide. A substantial part of this width, perhaps as much as 30–40%, was the result of a long period of post-orogenic extension, lasting from the Late Palaeozoic into and through the Mesozoic, which was accompanied by the deposition of thick sedimentary successions; these now compose the continental shelves and host the main oil and gas resources of northern Europe. Thus, at the time of orogeny in the mid-Palaeozoic, this northern part of the Caledonides was a long and relatively narrow (*ca*. 600–700 km wide) mountain belt, similar in dimensions and majesty to today's Himalayas, separating the low-lying old cratons of eastern Europe and Greenland.

The Caledonian Orogeny, referred to as the Scandian Orogeny in these northern regions, resulted from the collision of two continents, Baltica and Laurentia. The former was much smaller than the latter and, during collision, played a similar role to that of India in the present-day Himalayan context. The Scandian Orogeny began in the Silurian and extended into and through the Early Devonian; the name is derived from the Scandes, the mountains of Norway and western Sweden. However, it is worth remembering that today's mountains, along the coasts of Scandinavia and eastern Greenland, are the result of Tertiary uplift during the opening of the Greenland and Norwegian seas; they are not the relics of Palaeozoic mountains, although they are dominated by Caledonian rocks.

Before the collision of Baltica and Laurentia and the Scandian Orogeny, these two continents were separated by the Iapetus Ocean. The closure of the Iapetus Ocean occurred over a period of about 80 Ma and involved the development of subduction systems along the margins of both Laurentia and Baltica and a wide range of tectonothermal activity. This complex situation, involving magmatism and sedimentation, deformation and metamorphism, was an essential part of the Early Caledonian evolution, prior to the final collision of the continents and a Devonian change in global stress regimes: compression and lateral shortening gave way to regional extension.

The tectonic evolution of the northern part of the Caledonide Orogen is discussed below, after a presentation of each of the three major regions of development – western Scandinavia, eastern Greenland, and the Barents Shelf.

Western Scandinavia

The Scandian mountains, with many peaks reaching a little over 2000 m, extend for nearly 2000 km along the length of Norway; they include substantial regions of western Sweden and the westernmost highest parts of Finland. The Caledonide Orogen, on land, is up to 300 km wide and extends off the Norwegian coast for a further 200–300 km beneath the shallow shelf areas of the Norwegian Sea. The orogen (Figure 2) is dominated by thrust sheets transported from west-north-west to east-south-east onto the Palaeozoic platform successions of the Baltoscandian margin of Baltica. The front of the orogen is generally marked by a prominent thrust scarp, clearly indicating that these allochthonous (i.e. transported) rocks originally, in the Devonian, extended much further eastwards, perhaps as far as 100 km, onto the platform.

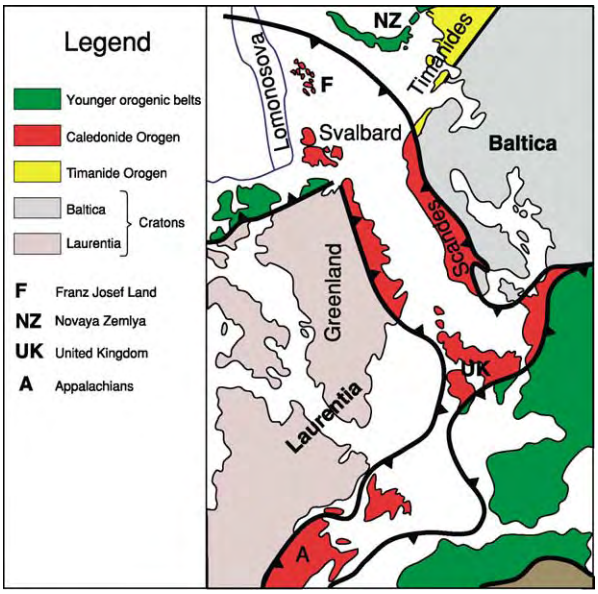


Figure 1 The North Atlantic Caledonides, from eastern Canada to the high Arctic Barents Shelf, in the Late Mesozoic.

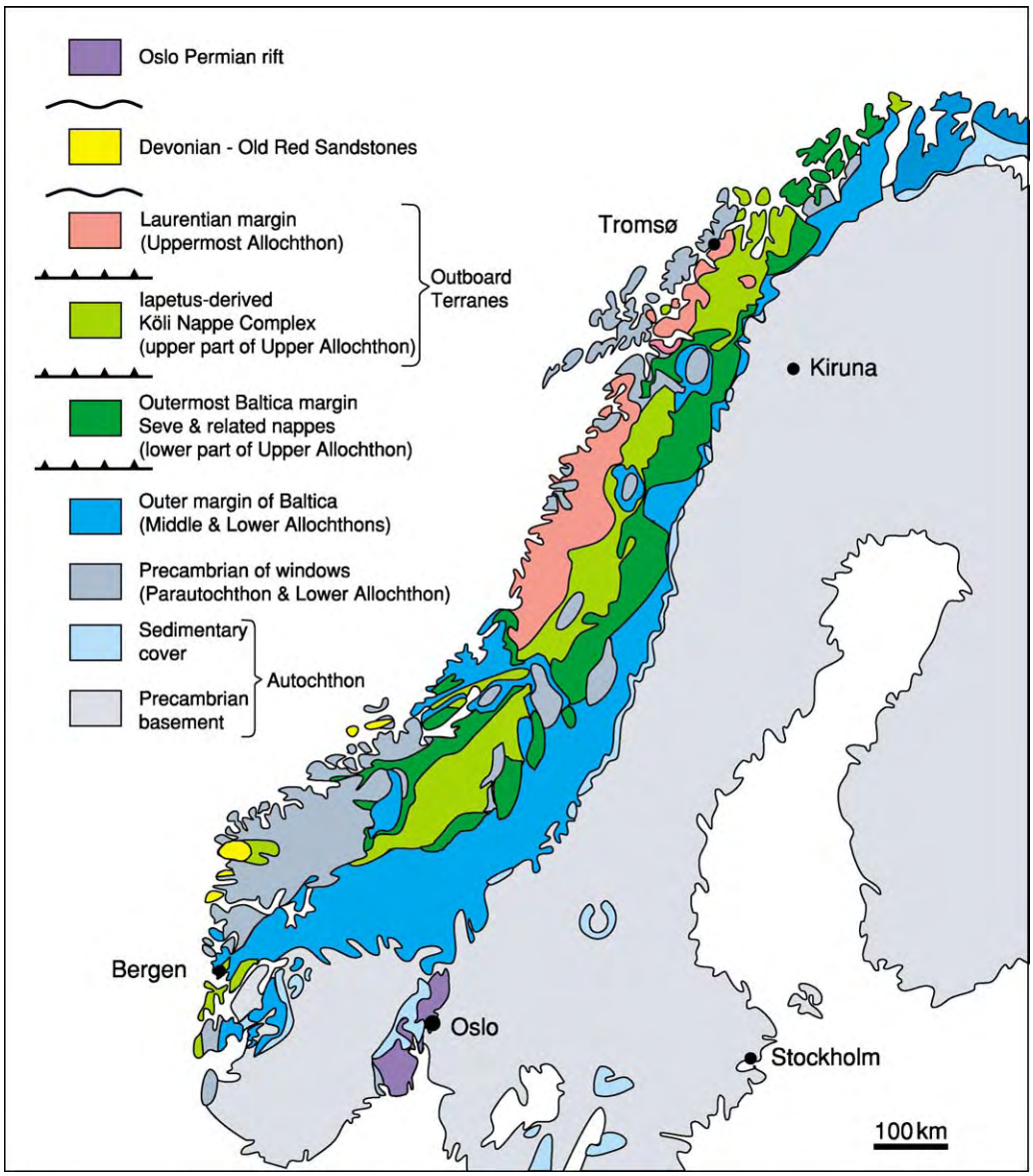


Figure 2 The Scandinavian Caledonides.

The Caledonide Orogen of Scandinavia is one of the world's classic areas for demonstrating the importance of very large scale near-horizontal thrusting of rocks during orogeny. Nearly 120 years ago, Alfred E Törnebohm demonstrated that the lateral transport of major allochthons from Norway into Sweden must have exceeded 100 km; subsequent work has shown that the total displacements amount to many hundreds of kilometres.

Along the eastern front of the orogen, the lowermost Scandian thrust sheets rest on thin (usually a few tens of metres) Cambrian sandstones and shales; locally, these autochthonous successions include underlying Vendian tillites and, in some central and southern areas, Ordovician and Silurian limestones and shales may also be present. These mainly Palaeozoic cover sediments were deposited on the deeply eroded and peneplained surface of the Fennoscandian Shield, which forms the basement along the entire Caledonian front. The shield is dominated by crystalline rocks, ranging in age from Late Mesoproterozoic (*ca.* 1000 Ma) in the south to Palaeoproterozoic and Archaean in the far north. The contrast in physical properties between these ancient granites and gneisses and the overlying sedimentary cover, especially the Cambrian alum shales, played an important role in controlling the character and geometry of the Scandian Orogeny. The alum shales are highly organic-rich (locally over 10%) and even today sometimes smell of bitumen. During

orogeny, these beds, up to a few tens of metres thick, provided a well-lubricated surface along the entire length of the mountain belt, over which the allochthons could easily be transported. These Cambrian shales are also characterized by unusually high levels of particular trace elements, such as uranium, vanadium, and molybdenum, allowing them to be identified further west, towards the hinterland of the orogen, where they are more metamorphosed and lack fossils.

The Scandian thrust sheets contain rocks that are derived both from the Baltoscandian margin (basement and cover) and from a variety of terranes that existed in the Early–mid-Palaeozoic outboard of Baltica. The latter are largely composed of ocean-floor and island-arc assemblages that formed in the Iapetus Ocean and were thrust onto the Baltoscandian margin during the Scandian Orogeny. Identification and correlation of the different allochthons across the orogen from the foreland, where they are relatively well preserved, to the hinterland, where they are generally highly attenuated and metamorphosed, has been possible only because all the thrust sheets are influenced by major late-orogenic folding with orogen-parallel axes. Within the cores of the anticlines, in tectonic windows, basement–cover complexes and overlying thrust sheets have been mapped, and allochthons have been correlated both along and across the orogen.

The many Scandian thrust sheets can be subdivided into four major groups of allochthons – Lower, Middle, Upper, and Uppermost. Where the rocks below the Lower Allochthon are disturbed but still comparable with the autochthon, the term parautochthon has been widely applied. The Lower and Middle Allochthons are composed of rocks derived from typical Baltoscandian-margin sedimentary-cover successions and their underlying crystalline basement. Mafic dykes intrude the uppermost units of the Middle Allochthon (Särv Nappes). The Upper Allochthon is more varied and is readily divisible into two parts: a lower group of thrust sheets (Seve Nappe Complex) that are compositionally similar to some of the underlying sedimentary cover but are much more highly metamorphosed, and an upper group (Köli Nappe Complex) of lower metamorphic grade that includes a wide range of igneous and sedimentary rocks derived from oceanic environments and including ophiolites and island-arc and back-arc assemblages. The Uppermost Allochthon has also proved to be complex, containing both continental-margin lithologies and ophiolites, intruded by major granite batholiths; these highest nappes were derived from the margin of Laurentia.

Baltoscandian Platform to Outer Margin (Lower and Middle Allochthons)

The Baltoscandian-platform successions of the Autochthon generally dip 1–2° westwards and reappear towards the hinterland in the antiformal windows. The overlying thrust sheets of the Lower Allochthon, riding on a décollement surface above the Cambrian alum shales, are dominated by a Cambro-Silurian stratigraphy that can be readily correlated with that in the Autochthon, but in which the formations are generally thicker. Prominent facies changes occur in the Ordovician, with the platform carbonates of the Autochthon giving way westwards to basinal shales and westerly derived turbidites. A return to shallow-marine environments characterizes the Late Ordovician and Early Silurian, before Late Llandovery deepening and the influx of Mid-Silurian turbidites (flysch). In southern Norway, in the Oslo graben and, locally, in the Swedish Caledonides, Early Silurian shallow-marine limestones and deeper-water turbidites shallow upwards in the Wenlock to a nonmarine siliciclastic Old Red Sandstone facies indicative of the development of a foreland molasse-filled basin. This basin is inferred to have existed along the entire Caledonian front, prior to Late Palaeozoic and Mesozoic erosion of the mountain belt.

In foreland parts of the orogen, the Lower Allochthon is dominated by low-grade sedimentary successions, with only subordinate slices of Precambrian crystalline rocks. Further west, towards the hinterland, the thrusts cut deeper into the basement. The overlying Middle Allochthon contains extensive sheets of highly mylonitized Precambrian granites and gneisses, along with metasedimentary successions that are generally of Neoproterozoic age. Prominent mylonites separate the Middle Allochthon from the Lower Allochthon, and the metamorphic grade (low greenschist facies) is generally somewhat higher in the former. Precambrian complexes (Jotun Nappe) of the Middle Allochthon comprise the highest mountains in Scandinavia.

Included in the upper part of the Middle Allochthon are thrust sheets that have a stratigraphy similar to that in the underlying nappes, including thick Neoproterozoic siliciclastic and carbonate successions, overlain by Vendian tillites and then sandstones. However, these allochthons (Särv Nappes) are remarkable because of the widespread occurrence of *ca.* 600 Ma old mafic dyke swarms, which are often composite and even occur as sheeted complexes. Igneous rocks of Neoproterozoic or Palaeozoic age are notably absent from all the underlying tectonic units in the Scandes, except where they are exposed in the deep hinterland along the Norwegian west coast.

Thus, it can be inferred that these allochthons were originally located off the coast of western Norway and, along with the overlying nappes, have been transported at least 300 km eastwards onto the Baltoscandian Platform.

Outermost Margin of Baltica (lower part of the Upper Allochthon)

The Särv Nappes are overthrust by a higher grade allochthon, which is itself dominated by psammites, pelites, and subordinate marbles, with numerous amphibolited dolerites and gabbros and occasional dunites and serpentinites. These rock units comprise the Seve Nappe Complex. Many of the lithologies are comparable with, but more heterogeneous than, the underlying Särv Nappes; they are always of higher metamorphic grade, reaching granulite and eclogite facies in parts of the mountain belt. The eclogite-facies metamorphism is of latest Cambrian to earliest Ordovician age (*ca*. 500–490 Ma), and at least some of the granulites are younger and related to the Scandian collisional orogeny.

The Seve Nappe Complex has been treated as a separate so-called suspect terrane, which was probably derived from the outermost edge of Baltica, where the extensive mafic magmatism and solitary ultramafites were concentrated in the transition zone between continental and oceanic crust.

Iapetus Ocean Terranes (the Köli Nappe Complex of the Upper Allochthon)

Well-preserved ophiolites, derived from the Early Palaeozoic ocean floor, along with volcanic island-arc igneous complexes and associated thick, mainly volcaniclastic, sedimentary successions characterize the outboard terranes of the Scandinavian Caledonides. The ophiolites are mostly of Early Ordovician (perhaps partly Late Cambrian) age, and the arc complexes date from throughout the Ordovician, providing evidence of the subduction systems controlling the closure of the Iapetus Ocean. The rock units are generally metamorphosed in greenschist facies and, within the sedimentary successions, fossils are locally preserved, providing evidence of both age and faunal affinity. Of particular interest has been the identification of Early Ordovician faunas (mainly trilobites and brachiopods, but also molluscs) with North American affinities, indicating that some of the ophiolite-bearing allochthons were derived from the Laurentian side of the Iapetus Ocean; others, apparently, were more centrally placed in the ocean.

These Scandian Iapetus-derived outboard terranes occur in the Köli Nappe Complex, which, in central parts of the mountain belt, can be readily divided into three parts – Lower, Middle, and Upper. Fragmented ophiolites occur locally in the base of the Lower Köli Nappes (e.g. at Otta and Handöl). Associated black-shale formations often host solitary ultramafites. The latter are sometimes mantled by detrital serpentinite and at one remarkable location (Otta) these sedimentary rocks are highly fossiliferous, giving good control of age (Llanvirn) and ocean-island affinity. Early Ordovician island-arc and rifted-arc volcanites dominate the lower parts of the Lower Köli Nappes, which are overlain by turbidites and sandstones; these shallow upwards in the Late Ordovician into continentally derived quartz sandstones and then limestones. The latter host a coral- and brachipod-bearing fauna with *Holorhynchus*, providing evidence of proximity to Baltica by Ashgill times. These shallow-marine environments deepen into basinal black shales, sometimes with pillow basalts, in the mid–Late Llandovery, after which the basins are filled with turbidites.

The Middle Köli Nappes are dominated by igneous complexes with calc-alkaline volcanic-arc affinities. Major plutons range in composition from gabbro and gabbro–diorite to tonalite, granodiorite, and trondhjemite, and the host rocks are mainly basalts and volcaniclastic formations, all generally metamorphosed in greenschist facies. Ordovician turbidites are conspicuous, and black shales have locally yielded Llandovery graptolites. The igneous complexes have provided mid-Ordovician uranium–lead zircon ages and evidence of Ordovician deformation; their chemistry indicates an origin in an oceanic setting, and most authors have considered them to be related to Laurentian-margin subduction systems.

The Upper Koli Nappes contain a major basal ophiolite (e.g. at Støren) and a range of mid-Ordovician calc-alkaline plutons. The ophiolite is directly overlain by Arenig limestones containing trilobites and brachiopods of unambiguous North American affinities; it has therefore been inferred that initial obduction of this ocean floor was onto the outer continental margin of Laurentia. Turbidites and fanglomerates dominate the overlying successions, which are also thought to be Ordovician in age; Silurian fossils have not been found in the Upper Koli Nappes.

Laurentian Continental Margin (Uppermost Allochthon)

Highest in the Scandian nappe pile, located along the west coast of central Norway, is a characteristic allochthon dominated by major granitoid batholiths, intruded into amphibolite facies sedimentary successions of continental affinity – marbles, schists, and psammites. Within this allochthon (e.g. in the

Helgeland Nappe Complex) there are fragmented ophiolites that may have affinities with those in the underlying Støren Nappe. The batholiths are of Ashgill to Llandovery age; thus they were intruded immediately prior to Scandian collision.

Scandian Collision

The Scandian collision, with thrusting of major allochthons many hundreds of kilometres onto the Baltoscandian platform, resulted in deep depression of the western edge of Baltica. Eclogite facies (high pressure and temperature) metamorphism, locally with mineralogical evidence of very high pressures (coesite and microdiamonds), was widespread in mafic rocks in the Precambrian basement of south-western Norway (Western Gneiss Region); it also occurs further north along the coast at a similar structural level. Exhumation of these deeply depressed rocks occurred during both Late Silurian thrusting and the extensional collapse of the orogen, with particularly rapid uplift in the Early Devonian. Crustal thickening during continental collision led to the build-up of the Caledonian mountain belt, and this was kept in balance during continued collisional compression by upper-crustal extension. This gravitational collapse resulted both in further migration of the nappe pile onto the Baltoscandian platform and in major westerly directed detachments in the hinterland, the latter accompanying the development of Old Red Sandstone basins.

Eastern Greenland

The Caledonian mountains of eastern Greenland, reaching from 70°–82° N, flank the eastern edge of the Greenland ice dome. As in Norway, the coastline is deeply penetrated by fjords, and the steep mountain sides, many rising vertically for 1000–2000 m out of the sea, provide spectacular outcrops for geological analysis. The continental shelf reaches only about 50 km offshore in the south, but widens northwards to about 300 km in north-east Greenland. Beneath these shelf areas of the Greenland Sea, the Caledonian bedrock is covered by Late Palaeozoic and younger successions. Tertiary mafic volcanics and intrusions, related to the opening of the Greenland Sea, are extensive south of Scoresby Sund (*ca*. 70° N).

The outcrop of Caledonian bedrock in eastern Greenland (Figure 3) is about 300 km wide in the south, at 70° N, and narrows northwards to about 100 km, striking obliquely offshore into the northern Greenland continental shelf. It forms the western part of the Caledonide Orogen and is dominated by thrust sheets, which were emplaced west-north-westwards onto the Laurentian Craton with its Early Palaeozoic cover. Characteristic of the autochthonous cover of the Laurentian margin of the entire Caledonides, from eastern Canada in the south, via north-western Scotland and eastern Greenland, to Svalbard, is a Cambro-Ordovician carbonate succession that contrasts markedly with the coeval succession of Baltica in terms of both lithology and fauna. In southern parts of the eastern Greenland autochthon, these formations, with underlying quartzites and, locally, tillites, rest on a deeply eroded and peneplained basement of Archaean and Palaeoproterozoic age. In northern areas, thick Mezoproterozoic sandstones and basalts separate this old basement from the Palaeozoic cover, and tillites are absent.

Most of the orogen front in eastern Greenland is covered by ice, and the relationships between the major Caledonian thrust sheets and the underlying cover are best seen in antiformal windows, similar to those in the Scandes. The allochthons have been divided into two groups and are referred to as 'thin-skinned', overlain by 'thick-skinned'. Whereas in the far north the former is a fold-and-thrust belt dominated by sedimentary successions, further south it incorporates extensive slices of Palaeoproterozoic and Archaean crystalline rocks. The upper 'thick-skinned' complex differs greatly from the underlying thrust sheets in the area south of 76° N, by incorporating latest Mesoproterozoic successions (Krummedal and Smallefjord groups), which are several kilometres thick and intruded by *ca*. 930 Ma granites; these are overlain, probably unconformably, by Neoproterozoic siliciclastic and carbonate successions (Eleonore Bay Supergroup), Vendian tillites, and platform Cambro-Ordovician carbonate formations. Thus the entire Caledonian mountain belt of eastern Greenland is composed of rocks derived from the Laurentian margin, including both Archaean and Palaeoproterozoic crystalline basement and younger generally shallow-marine sedimentary successions deposited on it. Outboard (e.g. Iapetus-related) terranes are notable by their absence. West-north-west-directed thrust transport of the allochthons has been estimated to be at least 200 km, along with some sinistral strike displacements; thus, prior to the Caledonian Orogeny, the platform margin of Laurentia was nearly twice as wide as it is today.

Only in north-easternmost Greenland do the Early Palaeozoic carbonate-dominated successions continue upwards through the Ordovician into the Silurian, to be replaced by easterly derived turbidites in the mid-Silurian. This evidence, together with the presence of mid–Late Devonian Old Red Sandstones in extensional basins in southern areas and many mid-Palaeozoic isotopic ages, indicates that, although

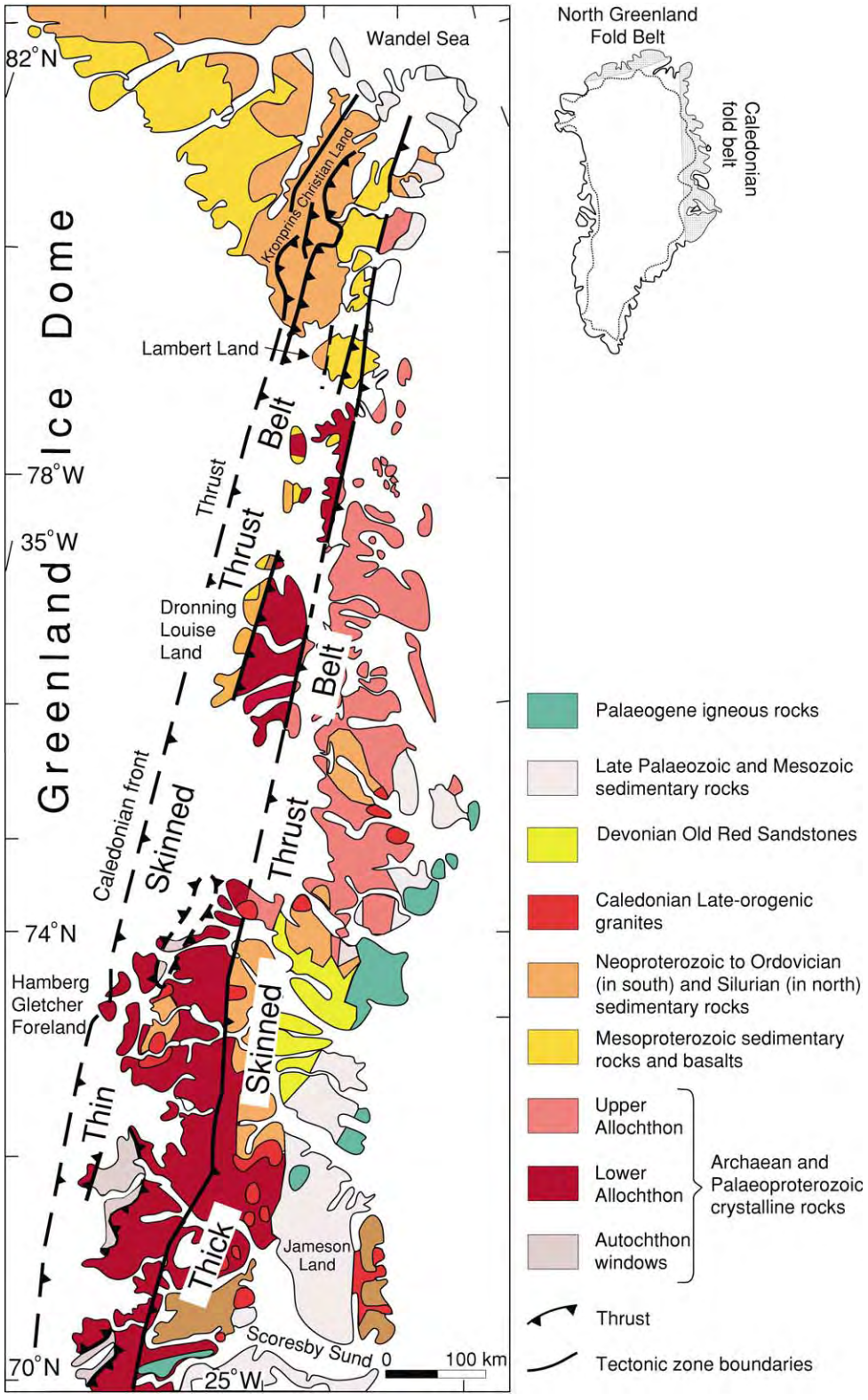


Figure 3 The Eastern Greenland Caledonides.

thrusting may have started in the Late Ordovician, the timing of the collisional orogeny in eastern Greenland coincided with that in the Scandes. In marked contrast to the Scandes, however, major post-kinematic granite plutons are present.

Svalbard

The Svalbard Archipelago (Figure 4), located on the north-western corner of the Barents Shelf, is dominated by a larger island, Spitsbergen, which is flanked to the east by Nordaustlandet, Barentsøya, and Edgeøya. Numerous other smaller islands are part of the archipelago, which reaches as far east as Kvitøya and south to Bjørnøya. The western and northern parts of the archipelago are dominated by Caledonian bedrock, including Old Red Sandstones; Carboniferous and younger successions unconformably overlie this 'basement', occurring in a major syncline that dominates the structure of the central and southern parts of the archipelago. These younger successions extend eastwards and southwards over much of the Barents Shelf. Along the western coast of Spitsbergen, a Tertiary east-verging fold-and-thrust belt is superimposed on the older structures.

Prior to the opening of the northernmost Atlantic and the Fram Strait, Svalbard's Caledonian bedrock was a direct northern extension of the East Greenland Caledonides (Figure 1). On Svalbard, the Caledonian rocks generally strike northwards and are split by major north-trending faults and Old Red Sandstone graben. Various provinces, or terranes, have been recognized, and nearly all can be related directly to the East Greenland Caledonides. However, the exception is important and occurs along the west coast of Spitsbergen.

Svalbard's eastern province (Figure 4) is readily divisible into two terranes, Nordaustlandet (including Kvitøya and eastern Ny Friesland) and West Ny Friesland. The Nordaustlandet Terrane is characterized by the typical Laurentian Cambro-Ordovician carbonate platform succession overlying tillites and thick Neoproterozoic dolomites, limestones, sandstones, and shales. Major unconformities separate

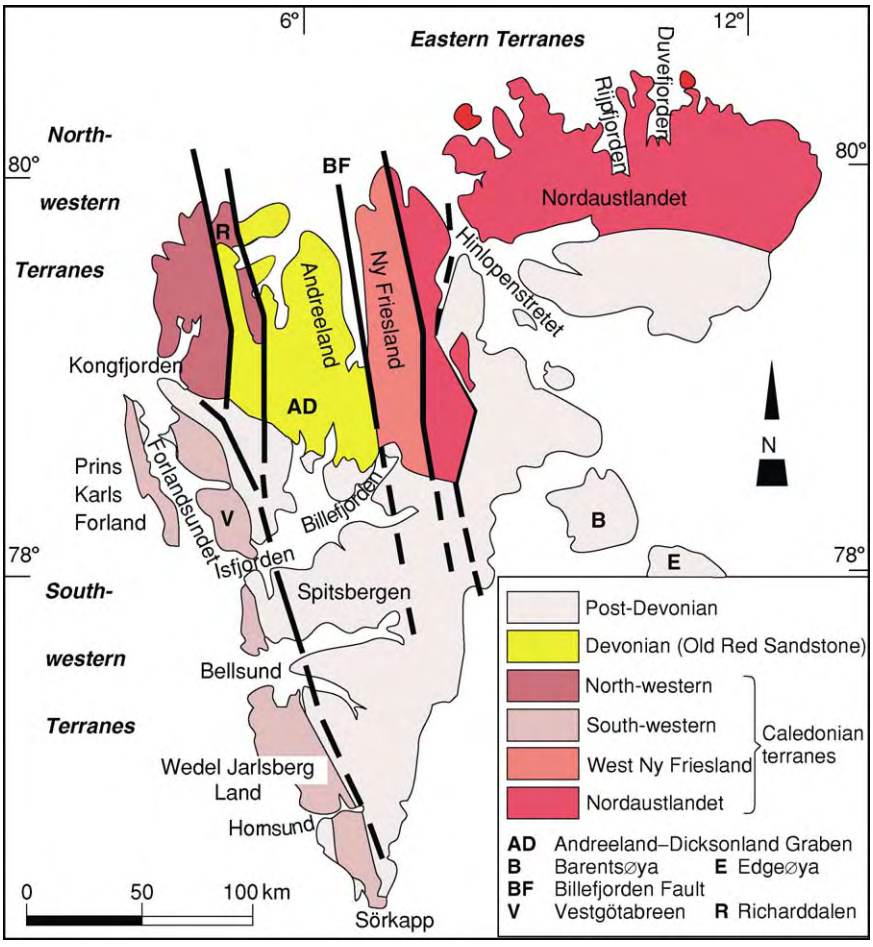


Figure 4 The Svalbard Caledonides.

this succession from a Late Grenvillian basement of andesites and rhyolites and underlying meta-turbidites of latest Mesoproterozoic or earliest Neoproterozoic age, intruded syntectonically by *ca*. 950 Ma granites. Upright to west-verging folding characterizes this easternmost Svalbard Terrane, and the grade of Caledonian metamorphism increases eastwards, with migmatization of the eastern parts of Nordaustlandet and the island of Kvitøya. Late-orogenic granites (*ca*. 420–410 Ma) also characterize this terrane, which, in all respects, compares closely with the highest allochthon of the central East Greenland Caledonides.

The Nordaustlandet Terrane is thrust westwards onto a succession of high amphibolite facies orthogneisses and isoclinally folded metasediments that is nearly 10 km thick – the Atomfjella Complex. The metasediments are dominated by quartzites, but also include marble and schist formations; the orthogneisses are mainly metagranites, dated at 1750 Ma. In both the metagranites and the quartzites, metadolerites are ubiquitous. Dating of detrital zircon has shown that the quartzites are Mesoproterozoic in age, but probably not younger than 1300 Ma (the age of the metadolerites), and that the marble and schist formations are latest Mesoproterozoic or younger. The Late Palaeoproterozoic 'basement' metagranites are repeated by thrusting at least three times in a major north-trending antiform that dominates the structure of western Ny Friesland. Only Caledonian (*ca*. 430–410 Ma) argon–argon ages of metamorphism have been obtained from the Atomfjella Complex; evidence of Grenvillian tectonothermal activity is notable by its absence.

The West Ny Friesland Terrane is closely comparable with the 'thick-skinned' allochthon of north-east Greenland, in terms of both the stratigraphy and the character of the Caledonian deformation and metamorphism. Taken together with the evidence (above) for the comparability of the Nordaustlandet Terrane and the central East Greenland allochthons, there can be little doubt that Svalbard's eastern Caledonian terranes are a direct northerly continuation of the East Greenland Caledonides (Figure 5). In much of

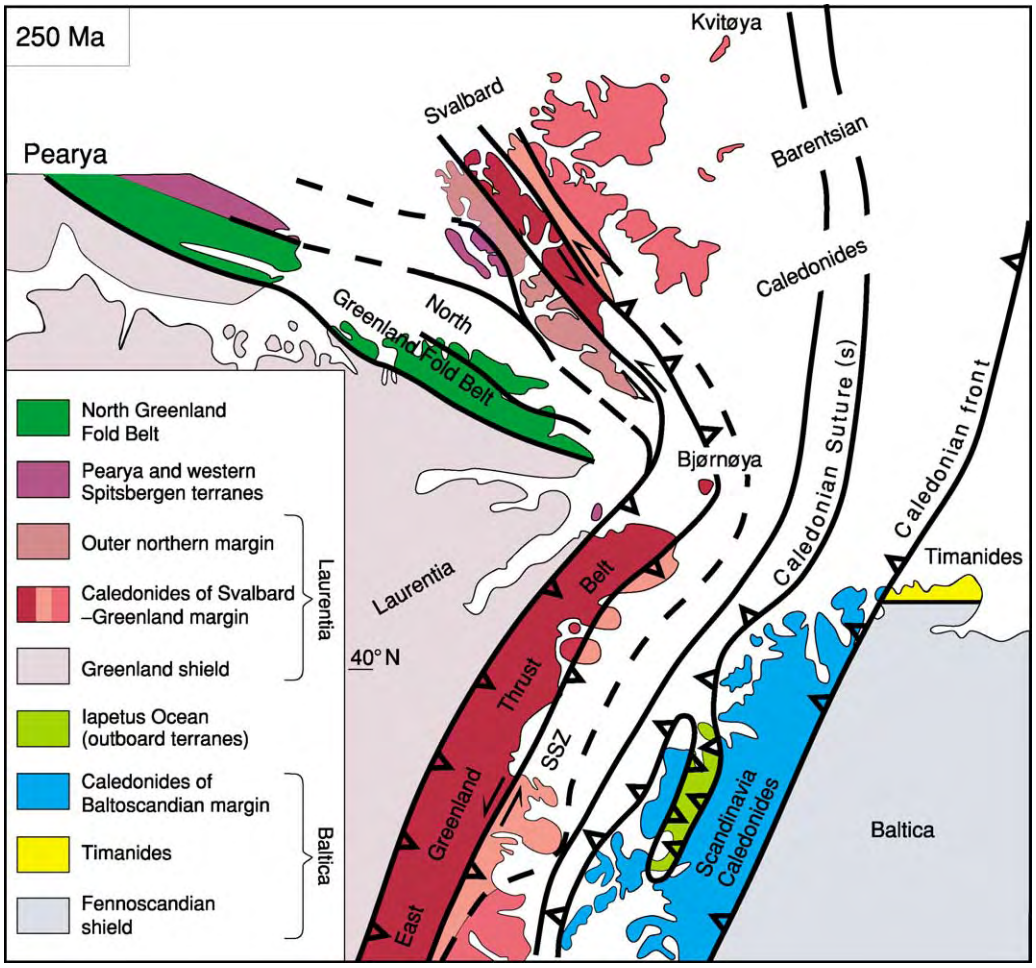


Figure 5 The Arctic Caledonides and Laurentia Baltica relationships in the Early Mesozoic.

the previous literature on the Svalbard Caledonides, a hypothesis has been favoured that these eastern terranes are derived from an off-shore area to the east of central East Greenland, implying sinistral transport over a distance of at least 1000 km; this hypothesis is not supported by recent investigations of Svalbard and Greenland.

Svalbard's eastern Caledonides are separated from the western terranes by the 50 km wide Andreéland–Dicksonland Old Red Sandstone graben. The Caledonian bedrock, from north-western Spitsbergen southwards via the west-coast Tertiary fold-and-thrust belt to southernmost areas and further south to Bjørnøya, provides evidence of both Laurentian-platform affinities and an outboard subduction-related terrane. Unambiguous Laurentian lithological and faunal signatures are found on Bjørnøya and southernmost Spitsbergen, where Cambro-Ordovician carbonate successions are closely related to strata of the same age in north-eastern Greenland. The underlying Neoproterozoic carbonate and siliciclastic successions are also comparable with those in this part of Greenland, but, interestingly, the metamorphic complexes unconformably underlying these sedimentary rocks are of Grenvillian age, a characteristic that is unknown in north-eastern Greenland.

In central western Spitsbergen, Neoproterozoic successions, including thick tillites, are overthrust by a blueschist and eclogite assemblage (the Vestgötabreen Complex). The high pressure–low temperature metamorphism occurred in the Early Ordovician and was followed by thrusting to high structural levels, erosion, and deposition of a mid–Late Ordovician conglomerate and limestone succession, passing up into Silurian turbidites. The subduction-related Vestgötabreen Complex is thought to be related to the Pearya Terrane of northernmost Canada (Ellesmere Island), both being foreign to Laurentia.

As shown in Figure 6, the Caledonian terranes of Svalbard comprise the western part of an orogen that extends northwards from the Scandes beneath the Barents Shelf to the edge of the Eurasian Basin. This part of the orogen is called the Barentsian Caledonides. Prior to the opening of the Eurasian Basin in the Tertiary, the Barentsian Caledonides are thought to have crossed Lomonosova (today's Lomonosov Ridge) and may have continued into the continental shelves of the Amerasian Basin.

The eastern side of the Barentsian Caledonides is flanked by the Timanides, a Late Neoproterozoic orogen that dominates the bedrock of north-easternmost Europe. The Timanides strike north-westwards

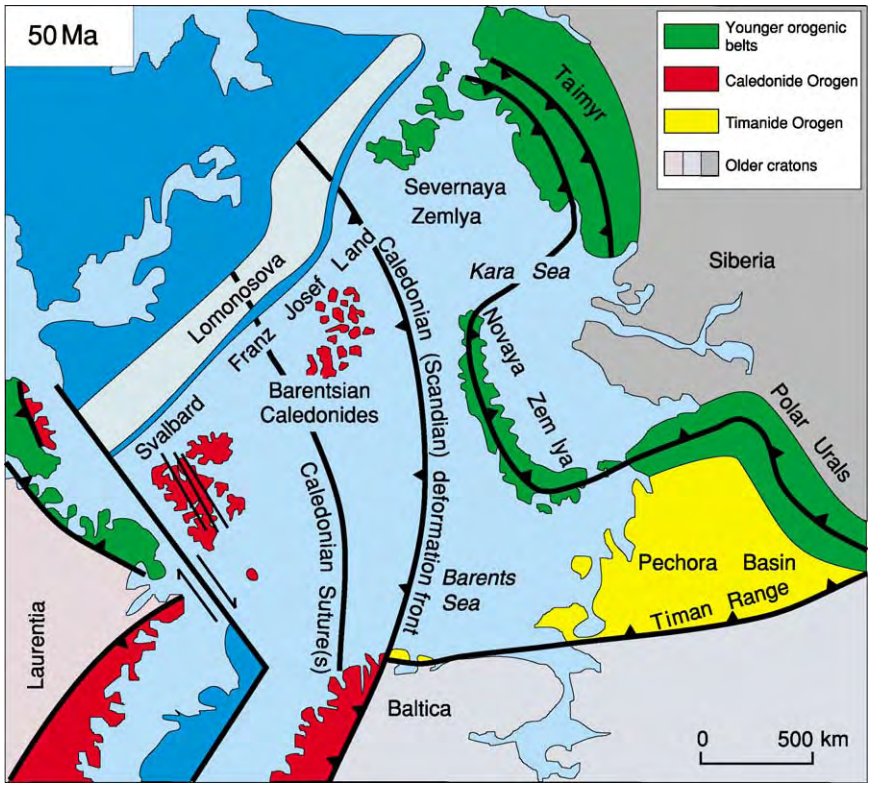


Figure 6 Tectonic elements of the western Eurasian Arctic in the Early Tertiary, during the initial opening of the Eurasian Basin and the Norwegian and Greenland seas (light blue, shelf and ridges; deep blue, ocean floor).

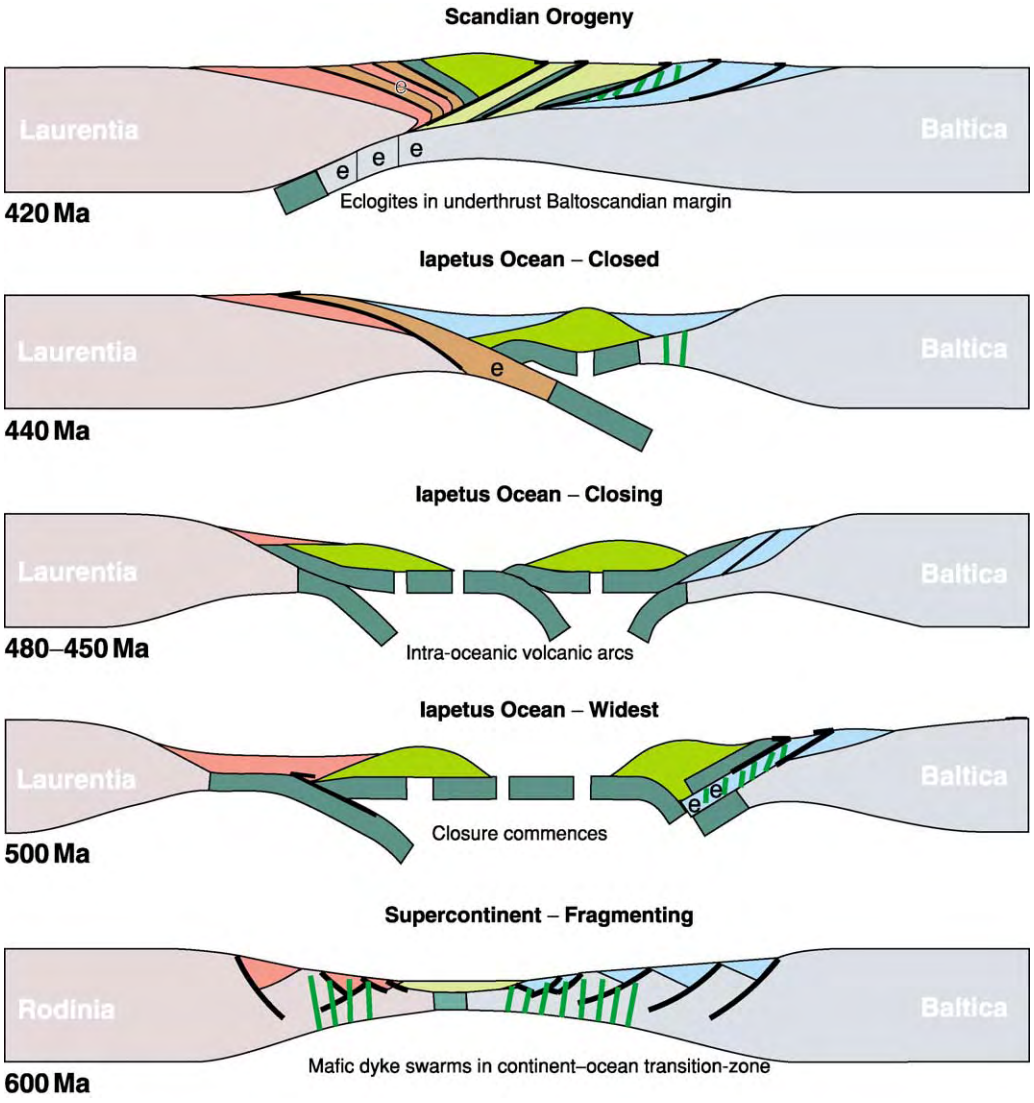


Figure 7 Caledonian tectonic evolution, from the Vendian to the Late Silurian. e; eclogite.

from a type area in the Timan Range, beneath the Pechora Basin, into the Barents Sea. A deep borehole on Franz Josef Land sampled folded Vendian turbidites at depth of 2 km; this evidence, along with geophysical data, suggests that the Barentsian Caledonides occur beneath Late Palaeozoic, Mesozoic, and Tertiary strata over much of the Barents Shelf.

Tectonic Evolution of the Northern Caledonides

The first interpretations of the North Atlantic Caledonides in terms of continental drift and plate tectonics inferred that Laurentia and Baltica were part of a single supercontinent (Rodina) in the Late Neoproterozoic. The Caledonian cycle was thought to involve rifting and separation of these continents, with the opening of a substantial ocean by seafloor spreading, followed by closure, resulting in a mid-Palaeozoic configuration that was similar to that which existed before ocean opening. Palaeomagnetic evidence has subsequently suggested that Baltica rotated about 120° between the Vendian and the Middle Ordovician; thus the model of simple opening and closing (the so-called 'Wilson cycle') is not generally accepted today. However, the structure of the Scandinavian Caledonides does provide evidence of a tectonic history that is readily divisible into three parts:

1. a period of Neoproterozoic rifting, leading to the separation of Baltica from a larger continental assemblage by the Early Cambrian;

2. opening of the Iapetus Ocean in the Cambrian and closing during the Ordovician, with subduction along the margins of both Laurentia and Baltica; and
3. Scandian collision of Baltica and Laurentia in the mid-Silurian and Early Devonian, with underthrusting of the latter by the former.

The Caledonides of eastern Greenland and Svalbard provide complementary evidence on the character of the Laurentian margin and allow the reconstruction of the tectonic evolution of the whole orogen, which is summarized schematically in Figure 7.

The vast Himalayan-type thrust systems of western Scandinavia and eastern Greenland, with many hundreds of kilometres of crustal shortening across the orogen, testify to more-or-less orthogonal collision of Laurentia and Baltica. Support for this conclusion is found in the kinematic evidence of west-north-west and east-south-east transport of the allochthons and the correlation of thrust complexes over distances of up to 1000 km along the orogen. This northern segment of the Caledonides differs markedly from the type areas of the Caledonide Orogen in the UK and further to the south-west, where sinistral transpression appears to have dominated the Caledonian collision of Avalonia and Laurentia.

See Also

Europe: Timanides of Northern Russia; Caledonides of Britain and Ireland; Permian to Recent Evolution. **Plate Tectonics**. **Tectonics:** Faults; Folding; Mountain Building and Orogeny.

Further Reading

Andreasson PG (1994) The Baltoscandian margin in Neoproterozoic early Palaeozoic times. Some constraints on terrane derivation and accretion in the Arctic Scandinavian Caledonides. *Tectonophysics* 231: 1 32.

Birkenmajer K (1981) The geology of Svalbard, the western part of the Barents Sea, and the continental margin of Scandinavia. In: Nairn AEM, Churkin M Jr, and Stehli FG (eds.) *The Arctic Ocean,* pp. 265 329. The Ocean Basins and Margins, Part 5. New York: Plenum.

Fortey RA and Bruton DL (1973) Cambrian Ordovician rocks adjacent Hinlopenstretet, north Ny Friesland, Spitsbergen. *Geological Society of America Bulletin* 84: 2227 2242.

Gee DG (1975) A tectonic model for the central part of the Scandinavian Caledonides. *American Journal of Science* 275A: 468 515.

Gee DG and Sturt BA (eds.) (1985) *The Caledonides Orogen Scandinavia and Related Areas*. Chichester: Wiley.

Gee DG and Tebenkov AM (2005) Svalbard: Fragments of the Laurentian Caledonian Margin. In: Gee DG and Pease VL (eds.) *The Neoproterozoic Timanide Orogen of Eastern Baltica*. London: Geological Society Memoir.

Grenne T, Ihlen PM, and Vokes FM (1999) Scandinavian Caledonide Metallogeny in a plate tectonic perspective. *Mineralium Deposita* 34: 422 471.

Harland WB (1997) *The Geology of Svalbard*. Memoir 17. London: Geoogical Society of London.

Henriksen N, Higgins AK, Kalsbeek F, and Pulvertaft TC (2000) Greenland from Archaean to Quaternary. Description text to the Geological map of Greenland 1 : 2 500 000. *Geology of Greenland Survey Bulletin* 185: 1 96.

Higgins AK and Leslie AG (2000) Restoring thrusting in the East Greenland Caledonides. *Geology* 28: 1019 1022.

Stephens M (1988) The Scandinavian Caledonides: a complexity of collisions. *Geology Today* 4: 20 26.

Stephens MB and Gee DG (1989) Terranes and polyphase accretionary history in the Scandinavian Caledonides. *Geological Society of America Special Papers* 230: 17 30.

Trettin HP (1989) The Arctic Islands. In: Bally AW and Palmer AR (eds.) *The Geology of North America An Overview,* pp. 349 370. Boulder, Colorado: Geological Society of America. V. A.

Variscan Orogeny

W Franke, Johann Wolfgang Goethe-Universität, Frankfurt am Main, Germany
P Matte, University of Montpellier II, Montpellier, France
J Tait, Ludwig-Maximilians-Universität, München, Germany

Introduction

Western and Central Europe were consolidated during an Upper Palaeozoic orogenic event, which has been named the 'Variscan' orogeny after a legendary Germanic tribe in north-eastern Bavaria, mentioned by the Roman author *Tacitus*. The term 'Variscan' should be given preference over 'Hercynian', which is often used for the same orogeny. The latter name is derived from the Harz Mountains in northern Germany, which do contain Variscan basement, but which represent a fault block thrust up during the Late Cretaceous.

Variscan rocks can be traced from Portugal to Poland and from the British Isles to the Mediterranean. The Variscan basement is exposed in 'massifs', which emerge from under younger Late Palaeozoic or Mesozoic cover. The most important massifs are the Iberian Massif in Spain, the Armorican Massif, the Massif Central, and the Maures Massif in France, parts of Corsica and Sardinia, the Vosges and the Black Forest, the Ardennes, the Rhenish Massif and the Harz Mountains in west-central Europe, and the Bohemian Massif further east. In the Alps and Carpathians, as well as in the Mediterranean realm, the Variscan basement has been much reworked by the Mesozoic–Cenozoic 'Alpine' orogenic processes (*see* **Europe:** The Alps).

The extraordinary width of the Variscan belt is the result of a complex palaeogeographical situation. The Variscan orogen is actually a collage of major and minor continental plates (Figures 1 and 2), which were once separated by at least three oceanic areas. These plates first spread apart (520–450 Ma), then converged, and eventually collided with each other (*ca*. 420–300 Ma). Sequential collision produced the huge landmass of Pangaea. By the Early Permian, the orogenic edifice was largely eroded, and was subsequently covered by Permian, Mesozoic, and Cenozoic deposits. The angular unconformity between Variscan basement and the cover is one of the key features of European geology (Figure 3).

There are two complementary ways to reconstruct such large-scale and long-term processes. Studies of palaeomagnetism (*see* **Palaeomagnetism**) can reveal fossilized magnetic fields in ancient rocks. The palaeodeclination reveals rotation of the sampling area with respect to the present-day magnetic meridian. The palaeoinclination records the palaeolatitude: inclination is vertical at the magnetic poles and horizontal at the magnetic equator. Since the Palaeozoic oceans in Europe were orientated more or less east–west, their closure implies changes in palaeolatitude, which are well documented in the palaeomagnetic record. Palaeomagnetic data can be compared with palaeoclimatic indicators, such as evaporites, coral reefs, tropical forests, or glacial deposits, and biogeography (areal distribution of fossil faunas and floras). With errors of $\pm$500 km, palaeomagnetism, palaeoclimatology and biogeography yield estimates of palaeolatitude and can therefore be used to deduce the larger-scale movements of the plates. It is important to note, however, that palaeomagnetism provides no information about palaeolongitude.

The second approach is geology. Even in ancient mountain belts, it is possible to identify characteristic elements of the plate-tectonic cycle, such as continental rifts, volcanic belts (magmatic arcs), remnants of oceanic lithosphere (ophiolites), and belts of high-grade metamorphic rocks, some of which have been subducted into the mantle before reascending to the surface. These features allow us to identify ancient active and passive plate margins, to establish a relative sequence of geological processes, and to date events. While the amount of ocean floor lost by subduction can be assessed only by palaeomagnetic methods, the deformation of the continental crust must be reconstructed by unravelling the polyphase tectonic deformation and metamorphism. In most orogenic belts, collisional deformation has reduced the width of the colliding blocks to 50% or even less.

In this article, we first address the large-scale migration of the continental blocks that are now welded together into the 'united plates of Europe', and then proceed to sketch out the 'ground truth' revealed by geological studies.

Palaeomagnetic and Biogeographical Record

The Palaeozoic geography of the crustal segments now contained in the continents bordering the Atlantic was dominated by the large plates of Laurentia, Gondwana, and Baltica. Between these plates there were a number of smaller microplates, which are now incorporated along the eastern margin of North America and

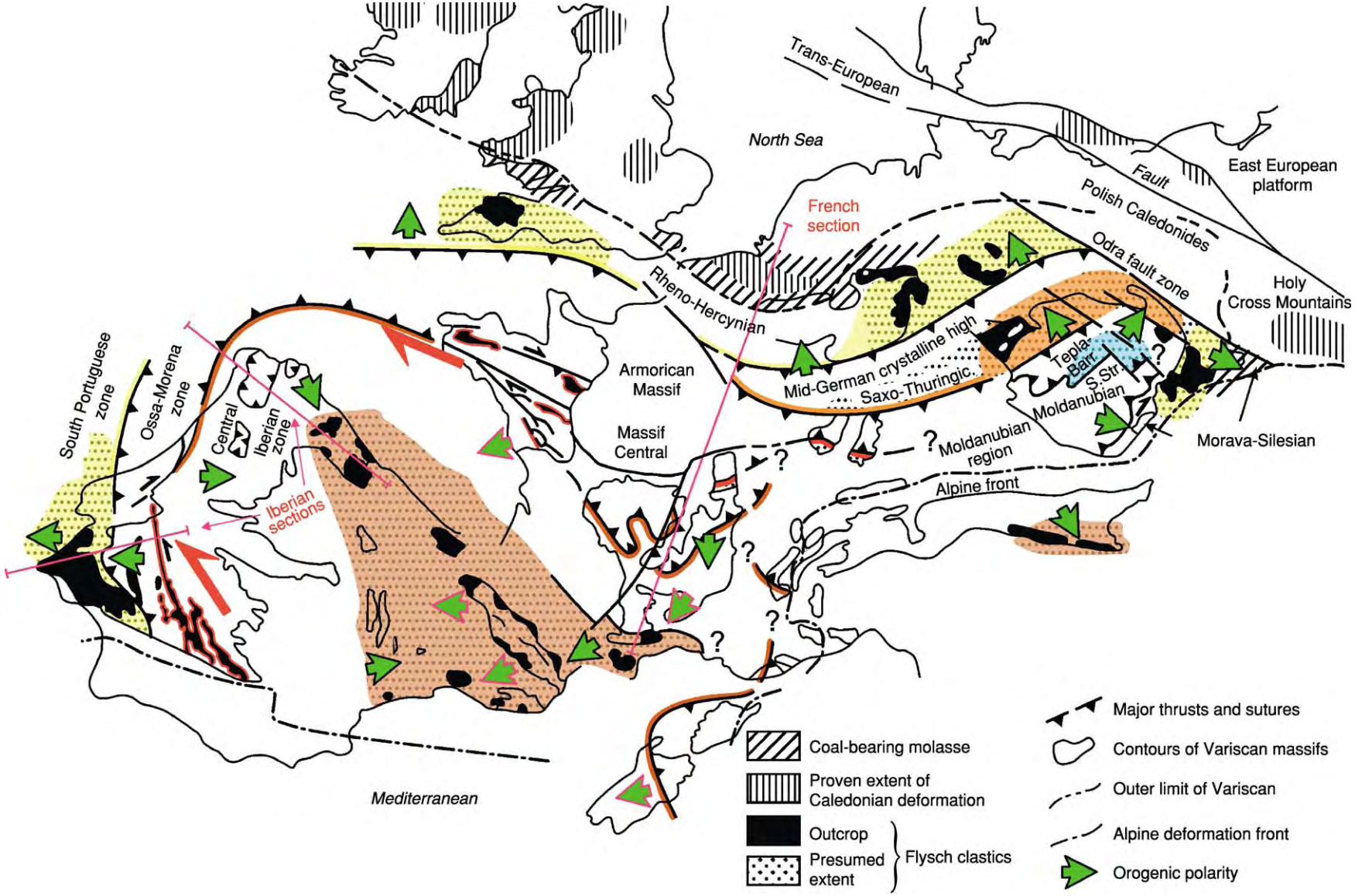


Figure 1 Tectonic and palaeogeographical subdivision of the Variscan basement in Europe, after Frank W (2000) The mid European segment of the Variscides: tectonostratigraphic units, terrane boundaries and plate tectonic evolution. In: Franke W, Haak V, Oncken, O, and Tanner D (eds.) *Orogenic Processes: Quantification and Modelling in the Variscan Belt*, pp. 35 62. Special Publication 179. London: Geological Society.

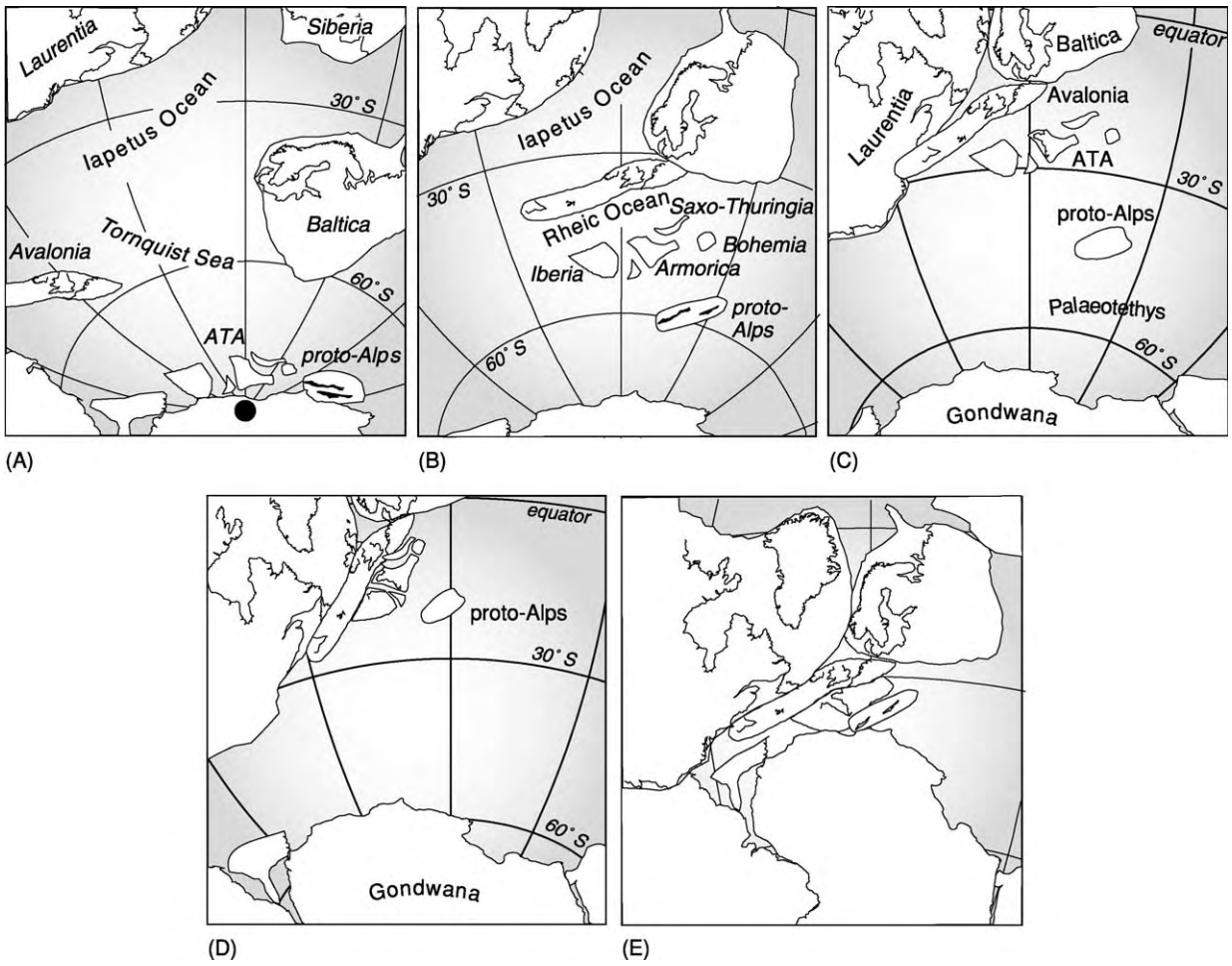


Figure 2 Positions of the continents and microplates bordering the Atlantic in (A) the Early Ordovician, (B) the Late Ordovician, (C) the Late Silurian, (D) the Middle Devonian, and (E) the Permian based on palaeomagnetic data.

Figure 3 Angular unconformity between folded Late Carboniferous deep water sandstones and Triassic fluvial deposits. Telheiro beach near Cabo Sao Vicente, Algarve, southern Portugal. Photograph taken by P. Matte.

the southern border of Eurasia. These terranes include Avalonia, the Armorican Terrane Assemblage, and the proto-Alps, which, in Cambrian to Early Ordovician times, were still attached to the north Gondwanan margin. Subduction of oceanic crust under this margin caused the Cadomian orogen (named after Caën, the Roman *Cadomus*, in Normandy). Cadomian basement is present in nearly all parts of the Variscan belt.

The positions and drift histories of Gondwana, Baltica, Laurentia, Avalonia, the Armorica Terrane Assemblage, and the proto-Alps are now fairly well constrained by palaeomagnetic data, and faunal and palaeoclimatic indicators are generally in good agreement. Palaeomagnetic data suggest the following palaeogeographical evolution (**Figure 2**).

Cambro-Ordovician

In the Early Ordovician, the northern margin of Gondwana was situated at high southerly palaeolatitudes, Baltica was situated between 30° S and 60° S and was inverted with respect to its present-day position, and Laurentia was in an equatorial position. The Cadomian basement rocks of Avalonia and the Armorican Terrane Assemblage clearly indicate that in Cambrian times these terranes were contiguous with the northern margin of Gondwana, and remained marginal to Gondwana until the Early Ordovician. By the late Early Ordovician (Tremadoc), data indicate that Avalonia had started to drift northwards, away from Gondwana, opening the Rheic Ocean in its wake (*see* **Palaeozoic: Ordovician**). It continued to move northwards throughout the Ordovician, gradually closing the Tornquist Sea and Iapetus Ocean, which separated it from Baltica and Laurentia, respectively. Palaeomagnetic data from different elements of the Armorican Terrane Assemblage indicate a similar, but independent, movement of these microplates, with separation from Gondwana being initiated slightly later in the Ordovician. No palaeomagnetic data are yet available from Ordovician rocks of the proto-Alps, but geological evidence suggests a continued Gondwanan affinity during this period.

Late Ordovician

By the Late Ordovician, Gondwana had moved some 30° northwards, and northern central Africa was situated over the south pole according to palaeomagnetic data from western Australia. Baltica was now in its present-day orientation, and its northern margin was at the equator. Laurentia, which did not move much throughout the Palaeozoic, remained straddling the equator and was separated from Baltica and Gondwana by the Iapetus Ocean. By Ashgillian times, palaeomagnetic and biogeographical data indicate that the Tornquist Sea, separating Baltica and Avalonia, had closed. Collision of Avalonia with Baltica created a narrow belt of deformation and metamorphism (the Polish Caledonides; **Figure 1**). The Rheic Ocean still separated Avalonia/Baltica from the Armorican Terrane Assemblage, which was situated at more southerly palaeolatitudes in the Ashgillian, based on palaeomagnetic evidence from the Bohemian Massif. The presence of glaciomarine sediments and cold-water faunas throughout the Armorican Terrane Assemblage reflects the Late Ordovician period of global cooling, which enabled the colonization of previously warmer-water realms by cold-water faunas. It has been shown that these glacial sediments of central Europe were deposited by seasonal or floating ice, in agreement with palaeomagnetic data from the Bohemian Massif, which clearly indicate intermediate to low palaeolatitudes. Strong faunal and lithological similarities in the Ordovician–Devonian successions of different massifs of the Armorican Terrane Assemblage indicate similar ecological conditions, demonstrating that they were all part of the same palaeogeographical domain. The proto-Alps were positioned at higher palaeolatitudes, between northern Gondwana and the southern margin of the Armorican Terrane Assemblage. This conclusion is based predominantly on faunal evidence, which indicates separation from northern Gondwana.

Siluro-Devonian

The palaeogeographical position of Gondwana from Silurian to Late Devonian times remains controversial on the basis of palaeomagnetic evidence, as two different models have been proposed in the literature. The more conservative model involves gradual northward movement of northern Gondwana throughout the Palaeozoic, with final closure of the ocean separating northern Africa from southern Europe in the Late Carboniferous. The alternative model is based primarily on palaeomagnetic data from south-east Australia and requires rapid northward movement of Gondwana in the Silurian, followed by rapid southerly movement in the Devonian. However, whether or not it is viable to use palaeomagnetic data from this region of Australia, whose autochthony with cratonic Australia is questioned, remains open to debate. In summary, taking all the palaeomagnetic, palaeoclimatic and biogeographical data into account, the more conservative model, involving gradual northward movement of Gondwana throughout the Palaeozoic, is considered more plausible.

Final closure of the Iapetus Ocean between Baltica/Avalonia and Laurentia occurred in the Siluro-Devonian, after which Baltica and Laurentia (Laurussia)

remained in equatorial palaeolatitudes until the end of the Palaeozoic era.

Late Silurian–Early Devonian palaeomagnetic data from a number of different crustal blocks of the Armorican Terrane Assemblage indicate palaeolatitudes of 20°–30° S. This implies gradual migration towards the southern margin of Baltica/Avalonia and closure of the intervening Rheic Ocean.

Late Devonian

By the Late Devonian, the Iapetus Ocean between Laurentia and Baltica/Avalonia had closed. Similarly, the Rheic Ocean between Avalonia and the Armorican Terrane Assemblage closed in the late Mid-Devonian. Closure of the ocean was essentially a longitudinal process and thus cannot be accurately constrained by palaeomagnetic data; however, invertebrate faunal differences between Bohemia and Avalonia persisted until the Emsian or Givetian. The now amalgamated Laurasian landmass moved southwards in Late Devonian times, but its northern border remained equatorial.

In Gondwana, however, the Late Devonian remains one of the more controversial periods. High-quality palaeomagnetic data from Australia clearly place central Africa over the south pole, requiring an ocean between the northern margin of Gondwana and the now amalgamated Laurasia. Faunal data from the southern Alps show poor similarity with the coeval fauna of northern Africa, and the sedimentary sequences of the southern Alps reflect a period of continuous sedimentation until the Late Carboniferous, with little evidence of any major deformation until Carboniferous times. However, the similarity of the fossil fish records of Gondwana (Australia) and Laurasia suggests that there was no oceanic separation of these two continents from late Early Devonian times onwards. These discrepancies in the faunal record, and between biogeographical and palaeomagnetic evidence, remain, as yet, unresolved. Nevertheless, there is general consensus that the collision of Gondwana with Laurasia to form the supercontinent Pangaea occurred in Late Carboniferous to Permian times.

A modern analogue of Palaeozoic plate dispersal might be the Indian Ocean: like Avalonia and the Armorican blocks, India, the Seychelles, and Madagascar have separated from Africa and, at least partly, made their way towards Asia (Laurussia).

Geological Record: Central Europe

Geological observations provide more detailed information about the plate-tectonic processes involved. Figure 4 is a diagrammatic representation of the plate-kinematic evolution of central Europe. Northward migration of the Armorican Terrane Assemblage was accommodated by northward-directed subduction of the Rheic Ocean beneath the southern margin of Avalonia, forming a magmatic arc, which is now preserved in the Mid-German Crystalline High (Figures 1, 4, and 5). By the late Early Devonian, the Rheic Ocean had more or less closed. At the same time, narrow oceans or seaways between the Armorican islands – Saxo-Thuringia and Bohemia – and between Bohemia and Moldanubia (possibly representing northern Gondwana) were being closed. The generally convergent movements were interrupted only by renewed extension along the Avalonia–Armorica boundary, creating a new narrow oceanic basin (the Rheno-Hercynian Ocean) whose remains can be traced from southern Portugal via south-west England and the Rhenish Massif into the Harz Mountains, with an easterly extension into Moravia (Figure 1).

The Rheno-Hercynian Ocean closed from the late Middle Devonian onwards, thus joining in the general convergence. The final collision of Avalonia with Franconia (the northernmost Armorican terrane) occurred in the earliest Carboniferous. Collision between the Armorican terranes further south (Saxo-Thuringia, Bohemia, and Moldanubia) took place in Late Devonian times.

Plate convergence was accommodated by a bilateral array of subduction or collision zones: the narrow Rheno-Hercynian and Saxo-Thuringian oceans were subducted towards the south, while the seaway between Bohemia and Moldanubia was subducted towards the north. This resulted in three collisional belts: the Rheno-Hercynian, Saxo-Thuringian, and Moldanubian (Figure 5), which were originally proposed by F Kossmat in 1927.

Collision resulted in the exhumation of previously subducted oceanic and continental rocks, which were emplaced as thrust sheets over the foreland areas. In each of the three collisional belts, the frontal thrust migrated towards the foreland, thus accreting foreland rocks to the orogenic wedge (Figure 5). Thrusting and folding reduced the widths of the original microplates considerably: collisional shortening in the three belts amounts to a minimum of *ca.* 800 km. This figure does not include the unknown width of oceanic crust subducted between the microplates.

Foreland basins filled with synorogenic clastic debris migrated ahead of the tectonic fronts. In the Rheno-Hercynian Belt, deposition kept pace with subsidence, so that sedimentation occurred in a shallow-marine to fluvial environment. The extensive coastal forests that developed in these tropical regions provided the raw material for economically important coal seams, which are exploited in Wales, the northern part of the Rhenish Massif, and Silesia (Figure 1).

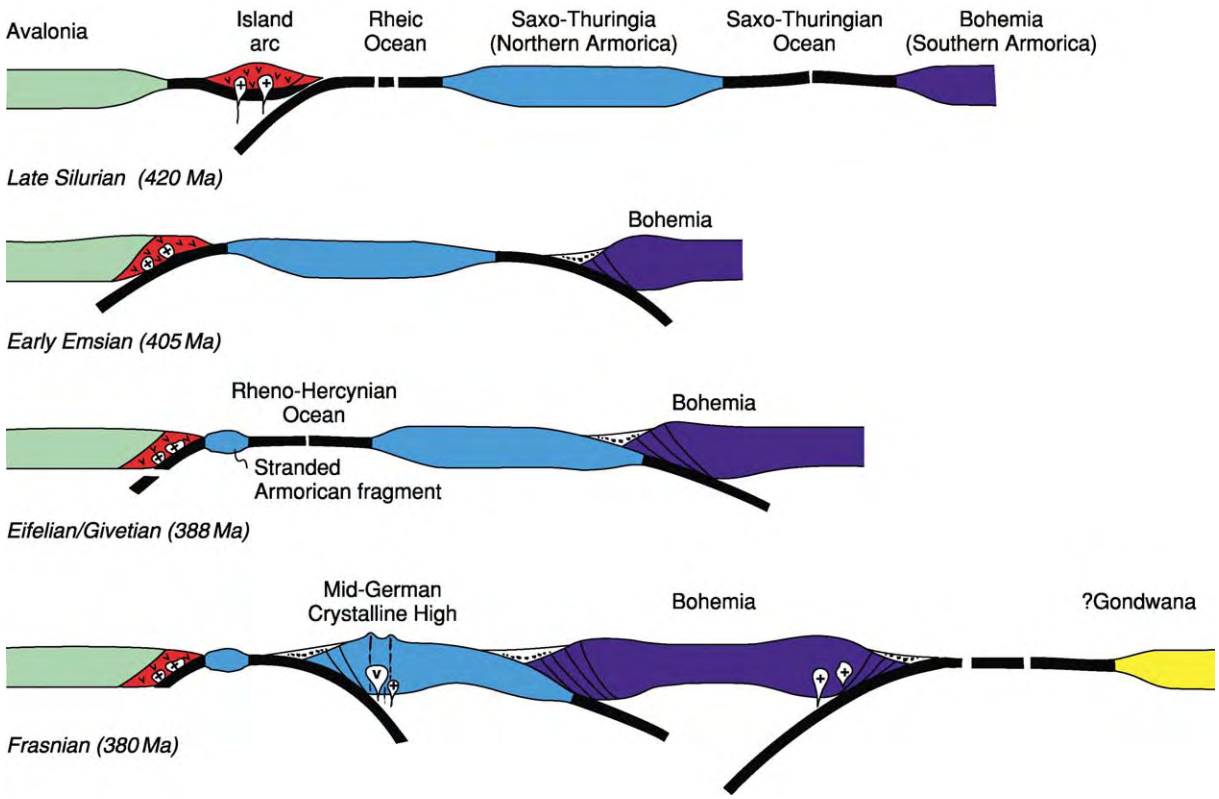


Figure 4 Plate kinematic evolution of central Europe, after Franke W (2000) The mid European segment of the Variscides: tectonostratigraphic units, terrane boundaries and plate tectonic evolution. In: Franke W, Haak V, Oncken, O, and Tanner D (eds.) *Orogenic Processes: Quantification and Modelling in the Variscan Belt*, pp. 35 62. Special Publication 179. London: Geological Society.

The present-day Variscan Belt of central Europe is not a linear feature. In Early Carboniferous times, the paralinear or linear subduction and collision zones were transported north-westwards, and drag along the south-western margin of Baltica caused it to rotate clockwise through approximately 90° (Figures 1 and 5). In about the latest Early Carboniferous, the south-eastern flank of this 'Bohemian Arc' was truncated by a huge transpressional fault zone (the Moldanubian 'Thrust'), which carried the south-eastern blocks south-westwards, possibly for a distance of about 1500 km.

Geological Record: Western Europe

An even more prominent arc structure is observed in the western part of the Variscan belt: if Iberia is rotated into its pre-Mesozoic position by closing the bay of Biscay, Variscan structures define a tight arc with a curvature of 180° in its internal part (the 'Ibero-Armorican' arc) (Figure 1).

The terranes of north-west Iberia are clearly linked with those of Brittany and the Massif Central in France and can be traced further eastwards into the Bohemian Massif, thus forming parts of the Armorican Terrane Assemblage. As in central Europe (see above), cross-sections through the two arms of the Ibero-Armorican virgation show a fan-like orogen with opposite vergences (verging towards the north-east and south-west in Iberia, and towards the south and north in France), i.e. convergent on its concave side and divergent on its convex side.

Two main sutures are visible in the internal metamorphic parts of the belt, and are the roots of nappes with opposing vergences (Figures 6 and 7). The Beja suture resulted from the closure of an oceanic realm between Avalonia (the south Portuguese foreland) and Armorica. There are no pre-Carboniferous rocks exposed in south Portugal, thus it remains uncertain whether this oceanic realm was the early Palaeozoic Rheic Ocean or a younger Devonian feature (as in south-west England and Germany). In any case, the Beja suture can be correlated with the Rheno-Hercynian ophiolites of the Lizard complex in south-west England and the Gießen-Harz Nappe in Germany. The second suture resulted from the closure of the Galice-Brittany ocean and can be traced from the French Massif Central, through the southern Vosges and Black Forest, into the Moldanubian belt of the Bohemian Massif (Figure 1). In contrast to

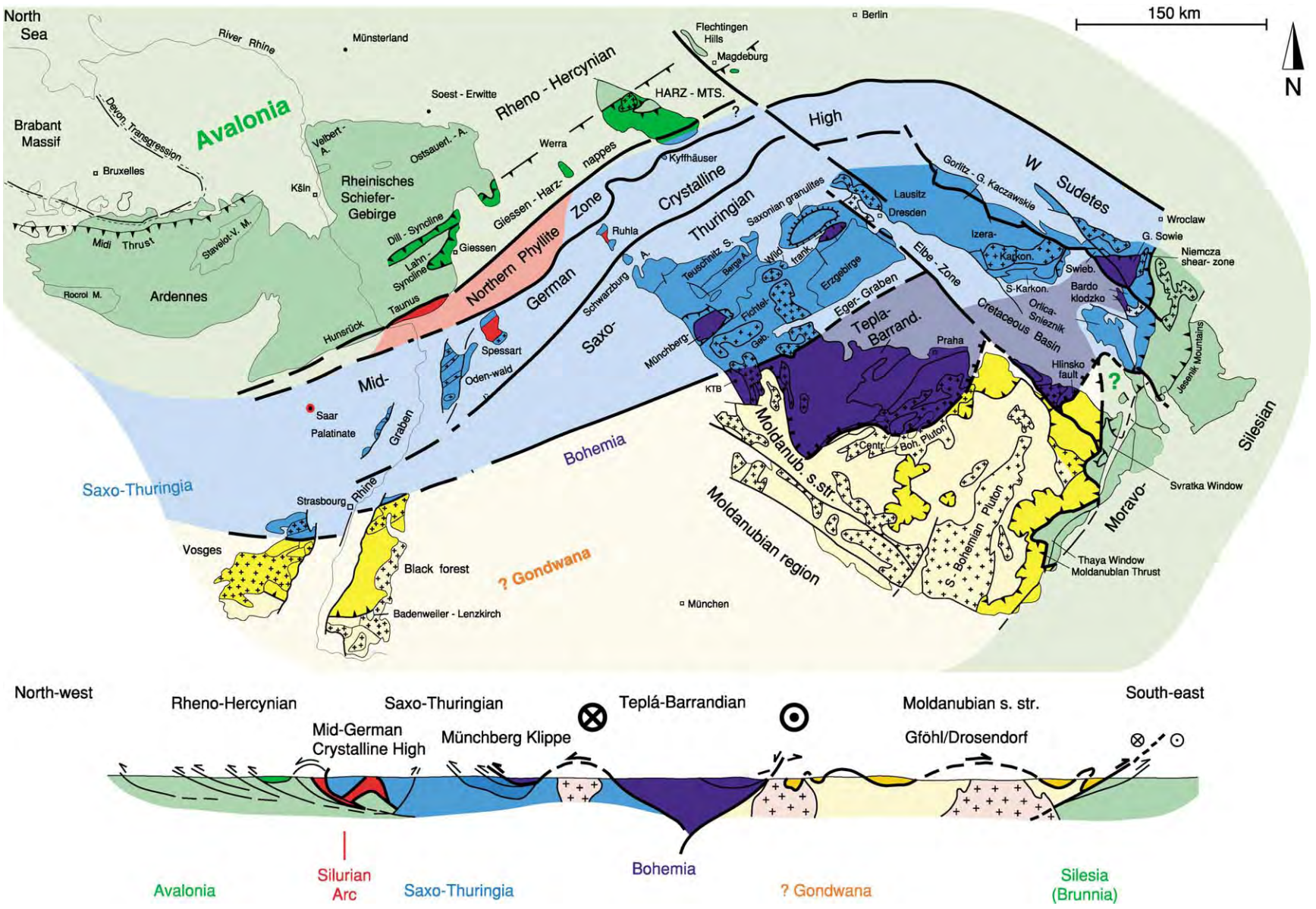


Figure 5 (A) Tectonic and palaeogeographical subdivision of the Variscan basement in central Europe. (B) Simplified tectonic cross section. A, Anticline; KTB, German Continental Deep Drilling site; M, Massif; S, South (as in S Bohemian Pluton); Teplá Barrand, Telpá Barrandian; Centr. Boh. Pluton, Central Bohemian Pluton; S karkon, South Karkonasze.

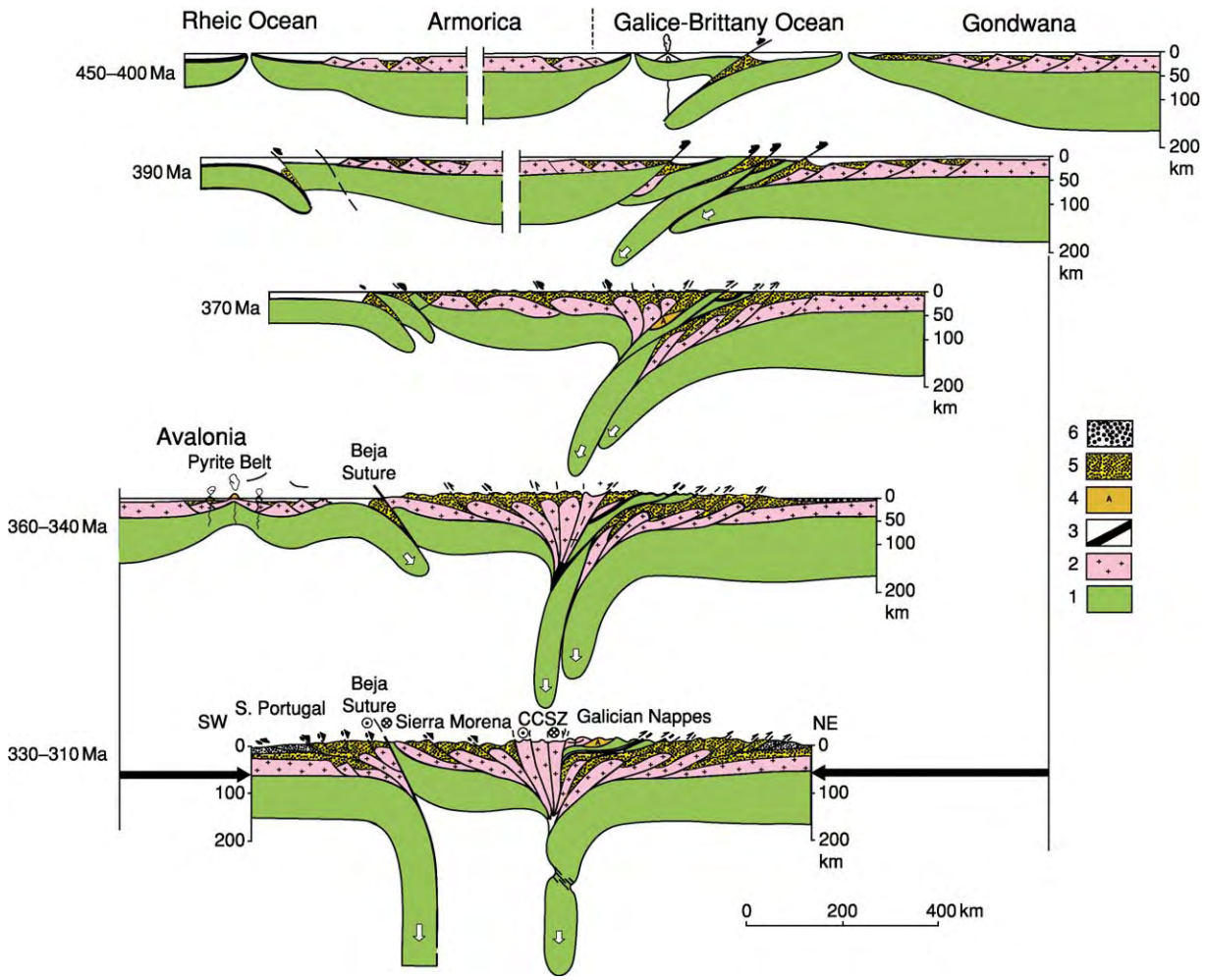


Figure 6 Plate kinematic evolution of the Variscan belt in Iberia: 1, mantle; 2, continental crust; 3, oceanic crust; 4, volcanic island arc; 5, Early Palaeozoic sediments; 6, Carboniferous foreland sediments. CCSZ, Coimbra Cordoba Shear Zone.

the situation in the central European section, there appears to be only one Armorican microcontinent (instead of Saxo-Thuringia and Bohemia). Nappes and sutures of the Ibero-Armorican domain are characterized by oceanic lithospheric rocks and early (420–370 Ma) ultrahigh-pressure, high-pressure and medium-pressure metamorphism. The most external parts of the belt comprise low-grade to non-metamorphic Carboniferous marine to paralic basins, which were deformed between 320 Ma and 290 Ma. The structures and tectonothermal histories of the Iberian and French sections of the Variscan Ibero-Armorican arc are described below (Figures 6 and 7). The Iberian section (Figures 6 and 7A) combines a southern segment from the South Portuguese Zone to the Central Iberian Zone and a northern segment based upon observations in north-western Spain (Figure 1). The northern segment originated from the closure of the Galicia-Brittany–Massif Central–Moldanubian ocean. Westward-directed subduction (in present-day coordinates) under an Armorican terrane affected first oceanic and then continental rocks between about 400 Ma and 300 Ma. Collision produced an eastward-facing accretionary wedge, in which deformation migrated from west to east. The lower autochthonous part consists of Early Palaeozoic shallow-water sediments deposited on the margin of Gondwana. The upper allochthonous part includes, from bottom to top, rocks of a passive margin thinned during the Ordovician, ophiolitic rocks (400–480 Ma) of arc or oceanic mantle, representing the remnants of the Galicia-Brittany–Massif Central ocean, and an uppermost ultra-nappe, which may represent the extended eastern margin of Armorica.

The southern segment of the Iberian section shows an opposite orogenic polarity, with south-west-facing folds and thrust sheets. Four main units are separated

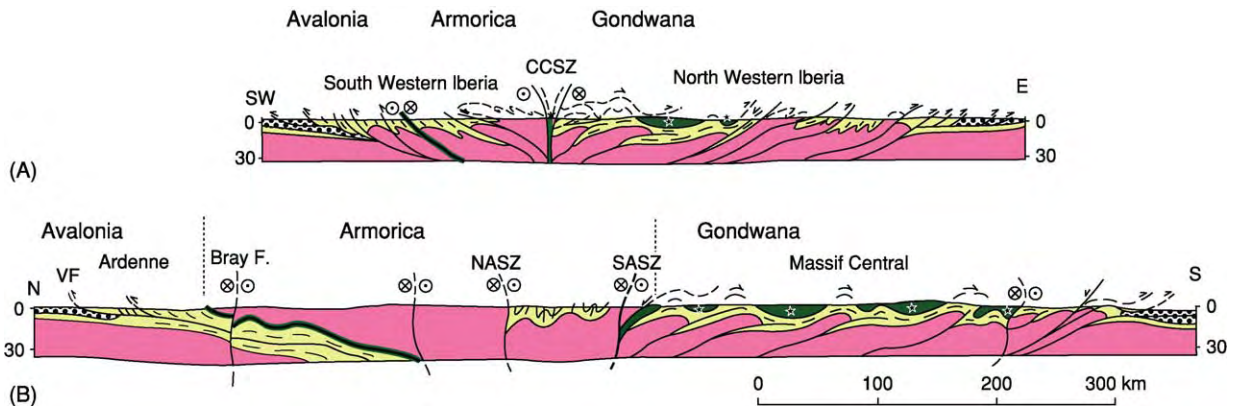


Figure 7 Simplified tectonic cross sections through the Variscan belt in (A) the Iberian peninsula and (B) France: pink, pre Palaeozoic basement and early Palaeozoic granitoids (undifferentiated); yellow, Early Palaeozoic sediments; green, thrust sheets with Early Palaeozoic ophiolites; dots, Carboniferous foreland basin deposits; VF, Variscan Front; Bray F, Bray Fault, NASZ, North Armoricn Shear Zone; SASZ, South Armorican Shear Zone; CCSZ, Coimbra Cordoba Shear Zone.

by major faults and/or sutures and are discussed below.

The Central Iberian Zone is the southern continuation of the autochthonous series of the northern Iberian section, with Early Palaeozoic rocks unconformably overlying low-grade Proterozoic sediments. A very important intracontinental sinistral shear zone separates the Central Iberian Zone from the Ossa Morena Zone and marks the boundary between tectonic domains with north-east and south-west vergences. This shear zone either represents or overprints the suture zone in which the north-west Iberian allochthon is rooted. There is evidence of very peculiar Ordovician magmatism with alkaline to peralkaline gneisses.

The Ossa Morena Zone shows a relatively complete fossiliferous Early Palaeozoic record from the lowermost Cambrian through to the Silurian and Devonian. Lower Cambrian sandstones and limestones unconformably overlie Upper Proterozoic rocks that are characterized by the presence of black cherts metamorphosed and intruded by Cadomian granitoids (550–500 Ma) as in northern Brittany. In some places (e.g. Cordoba) calc-alkaline andesites erupted during the latest Proterozoic/Early Cambrian. The Ossa Morena Zone shows large south-west-facing recumbent folds emplaced before the Early Carboniferous (which is, in part, coal-bearing). The Variscan metamorphism (epizonal to catazonal) is of a high-temperature–low-pressure type, and an important Variscan magmatism (diorites and granodiorites) is present. The present-day southern boundary of the Ossa Morena Zone is considered to be a suture, based on the existence of a belt of oceanic amphibolites (Beja-Acebuches) thrust over a possibly Devonian accretionary prism with slices of oceanic metabasalts. This north-west–south-east-trending boundary with the South Portuguese Zone has been reworked as a sinistral wrench fault.

The South Portuguese Zone exposes only Devonian and Carboniferous sediments and is characterized by important bimodal Tournaisian volcanic deposits, which contain the largest copper ore bodies in Western Europe (the South Portuguese Pyrite Belt). Time-equivalent volcanic rocks are known from south-west England and are widespread in the Rheno-Hercynian Belt of Germany. The Rheno-Hercynian and, hence, Avalonian affinities of the South Portuguese Zone are also suggested by a thick turbidite sequence deposited in a foreland basin ('Culm' facies), with a change towards paralic deposits in the south-west. It comprises the Viséan to Westphalian D. The South Portugese Zone is characterized by south-west-facing folds and thrusts. Deformation occurred during the Late Carboniferous.

The French section (see Figure 1 for location) can be subdivided into three segments. The Massif Central segment is the most complete section on the southern flank of the Variscan belt and is exposed over a distance of 400 km from the Montagne Noire to the southern margin of the Paris Basin. The Paris Basin segment is buried beneath Mesozoic sediments, but has been sampled in rare drillings and imaged at depth by a wide-angle reflection profile. Potentially equivalent rocks are exposed to the west in central Brittany. The Ardennes section is partly buried beneath the sediments of the Paris Basin, but is well known at depth from drilling, coal mining, and a

deep seismic-reflection profile. Approximately 100 km of section are well exposed along the Meuse River in the Ardennes.

In the Massif Central segment, three main tectonostratigraphical units may be distinguished. First, the Montagne Noire at the southern extremity of the Massif Central consists of fossiliferous Palaeozoic sediments including Lower Cambrian shallow-water sandstones and limestones, Lower Ordovician shales and sandstones, Devonian limestones, and a thick turbiditic syntectonic Visean–Namurian series. The whole of this lithostratigraphical sequence is involved in large (10 km) recumbent folds facing south to south-west. The lowermost para-autochthonous unit (Zone Axiale) consists of Upper Proterozoic sediments that were intruded by Ordovician granites at 450–460 Ma. Deformation and low-pressure metamorphism occurred between 330 Ma and 310 Ma.

Second, the 'Schistes des Cévennes-Albigeois' are a very thick (possibly around 4000 m) greenschist-grade series, probably derived from both Early Palaeozoic and Late Proterozoic protoliths. They were intruded by Cambrian–Ordovician diorites and granites at 540–460 Ma. The Cévennes series shows a northwards gently dipping slaty cleavage related to a pervasive southward or south-westward shearing. Barrovian-type metamorphism increases towards the top of the pile (inverted). Deformation and metamorphism occurred between 350 Ma and 340 Ma. Large granitic plutons were emplaced between 330 Ma and 305 Ma.

Third, the complex Leptyno-Amphibolitic Group consists of mafic and ultramafic rocks characterized by high-pressure to ultrahigh-pressure metamorphism (in places they consist of coesite-bearing eclogites). It is a very large (300 km) nappe with ophiolite fragments. The probable root zone in the southernmost part of the Paris Basin is marked by a significant positive gravity anomaly. The Leptyno-Amphibolitic Group is overlain by pelitic gneisses with slices of high-pressure granulites and peridotites. Metamorphism has been dated at 420 Ma (high-pressure) to 380 Ma (Barrovian). All these units are intruded by various types of Variscan granites, which were emplaced between 360 Ma and 300 Ma.

The Ardennes segment is well documented from drilling, coal-mining, and surface outcrops in the Ardennes. Like the more easterly Rhenish Massif, this segment represents the northern flank of the European Variscan Belt. Palaeozoic fossiliferous sediments with a thick Devonian clastic wedge are involved in a large north-west-facing fold and thrust belt, which has been well imaged by ECORS (France) and DEKORP (Germany) deep reflection profiles. The main frontal thrust (Faille du Midi), exposed at the France–Belgium border, carries Devonian rocks over the Late Carboniferous coal basin. Deformation occurred at around 300 Ma.

Features Characteristic of the Variscan Belt

The central parts of the Variscan belt are intruded by huge volumes of granite. In contrast to the situation in many other orogenic belts, most of these granites were not formed over oceanic subduction zones but originated from the melting of metasediments in the continental crust. Heating and melting of the crust probably occurred initially when the crust was thickened (two mica and cordierite leucogranites) and then later when subducted parts of the subcontinental lithospheric mantle became detached and sank back into the asthenospheric mantle. This permitted the upward ascent of hot asthenospheric mantle and advection of heat to the crust.

The Variscan crust is rich in unstable isotopes (mainly potassium, uranium, and thorium), whose decay may produce up to 30% of the heat flow observed in continental rocks. These heat-producing isotopes were extracted from the mantle by repeated magmatic episodes, shortly before and during Variscan plate convergence (Cadomian subduction magmatism, Cambro-Ordovician rift magmatism, and subduction- and collision-related magmatism in the Devonian and Carboniferous). These elements are also contained in mica and feldspar, two main constituents of the thick Early Palaeozoic shelf deposits, which piled up during the collision of the major and minor plates.

For these reasons, the Variscan orogen was 'hot' in comparison with 'cold' orogens such as the Alps (*see* **Europe:** The Alps), the Caledonides (*see* **Europe:** Scandinavian Caledonides (with Greenland); Caledonides of Britain and Ireland), and the Urals (*see* **Europe:** The Urals). The high temperatures prevailing during continental collision effected mechanical weakening of the crust. This is documented by the 'squeezing out' of melts or low-viscosity metamorphic rocks towards the forelands. The same effect is responsible for the rapid destruction of the orogenic 'root'. When heated, the deeper parts of the thickened crust spread laterally like oil on water. Therefore, the base of the continental crust (the Mohorovicic discontinuity) had already levelled out at a depth of about 30–35 km by Late Carboniferous or Permian times, i.e. shortly after the termination of crustal thickening (*see* **Moho Discontinuity**). This process was aided by the ascent of mantle-derived melts during the Permian, which initiated the break-up of Pangaea (*see* **Pangaea**).

See Also

Europe: Caledonides of Britain and Ireland; Scandinavian Caledonides (with Greenland); The Urals; The Alps. **Moho Discontinuity**. **Palaeomagnetism**. **Palaeozoic:** Ordovician. **Pangaea**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mountain Building and Orogeny.

Further Reading

Burg JP, Leyreloup A, Marchand J, and Matte P (1984) Inverted metamorphic zonation and large scale thrusting in the Variscan Belt: an example in the French Massif Central. In: Hutton DHM and Sanderson PJ (eds.) *Variscan Tectonics of the North Atlantic Region.* pp. 47 61. Special Publication 14. London: Geological Society.

Fortey RA and Cocks LRM (2003) Palaeontological evi dence bearing on global Ordovician Silurian continental reconstructions. *Earth Science Reviews* 61: 245 307.

Franke W (2000) The mid European segment of the Varis cides: tectonostratigraphic units, terrane boundaries and plate tectonic evolution. In: Franke W, Haak V, Oncken O, and Tanner D (eds.) *Orogenic Processes: Quantifica tion and Modelling in the Variscan Belt,* pp. 35 62. Special Publication 179. London: Geological Society.

Franke W and Stein E (2000) Exhumation of high grade rocks in the Saxo Thuringian Belt: geological con straints and geodynamic concepts. In: Franke W, Haak V, Oncken O, and Tanner D (eds.) *Orogenic Processes: Quantification and Modelling in the Variscan Belt,* pp. 337 354. Special Publication 179. London: Geological Society.

Franke W and Zelazniewicz A (2002) Structure and evolu tion of the Bohemian Arc. In: Winchester JA, Pharaoh TC, and Verniers J (eds.) *Palaeozoic Amalgamation of Central Europe,* pp. 279 293. Special Publication 201. London: Geological Society.

Kossmat F (1927) Gliederung des varistischen Gebirgs baues. *Abhandlungen des Sächsischen Geologischen Landesamtes, Neue Folge* 1: 1 39.

McKerrow WS, MacNiocaill C, Ahlberg PE, *et al.* (2000) The late Palaeozoic relations between Gondwana and Laurussia. In: Franke W, Haak V, Oncken O, and Tanner D (eds.) *Orogenic Processes: Quantification and Modelling in the Variscan Belt,* pp. 9 20. Special Publi cation 179. London: Geological Society.

Martinez Catalan JR, Arenas R, Diaz Garcia F, and Abati J (1997) Variscan accretionary complex of northwestern Iberia: terrane correlation and succession of tectonother mal events. *Geology* 25: 1103 1106.

Matte P (1998) Continental subduction and exhumation of HP rocks in Paleozoic belts: Uralides and Variscides. *Journal of the Geological Society of Sweden* 120: 209 222.

Matte P (2001) The Variscan collage and orogeny (480 290 Ma) and the tectonic definition of the Armorica micro plate: a review. *Terra Nova* 13: 122 128.

Owen AW, Harper DAT, and Rong Jia Yu (1991) Hirnan tian trilobites and brachiopods in space and time. In: Barnes CR and Williams SH (eds.) *Ordovician Geology,* pp. 179 190. Ontario, Canada: Geological Survey of Canada.

Robardet M, Verniers J, Feist R, and Paris F (1994) Le Paléozoique anté varisque de la France, contexte paléo géographique et géodynamique. *Géologie de la France* 3: 3 31.

Schönlaub HP (1992) Stratigraphy, biogeography and paleoclimatology of the Alpine Paleozoic and its implica tions for plate movements. *Jahrbuch der Geologischen Bundesanstalt* 135: 381 418.

Scotese CR, Boucot AJ, and McKerrow WS (1999) Gond wanan palaeogeography and palaeoclimatology. *Journal of African Earth Sciences* 28: 99 114.

Simancas JF, Carbonell R, Gonzalez Lodeiro F, *et al.* (2003) The crustal structure of the transpressional Variscan oro gen of SW Iberia: the IBERSEIS deep seismic reflection profile. *Tectonics* 22. DOI 10.1029/2002TC001479.

Stampfli GM (1996) The intra alpine terrain: a Paleoteth yan remnant in the Alpine Variscides. *Eclogae Geologi cae Helvetiae* 89: 13 42.

Suess E (1888) *Das Antlitz der Erde*, vol. IV. Prague, F. Tempsky.

Tait J, Schätz M, Bachtadse V, and Soffel H (2000) Palaeo magnetism and Palaeozoic palaeogeography of Gon dwana and European terranes. In: Franke W, Haak V, Oncken O, and Tanner D (eds.) *Orogenic Processes: Quantification and Modelling in the Variscan Belt,* pp. 21 34. Special Publication 179. London: Geological Society.

Torsvik TH, Smethurst MA, Meert JG, *et al.* (1996) Con tinental break up and collision in the Neoproterozoic and Palaeozoic: a tale of Baltica and Laurentia. *Earth Science Reviews* 40: 229 258.

Van der Voo R (1993) *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans.* Cambridge: Cambridge University Press.

Wegener A (1915) *Die Entstehung der Kontinente und Ozeane.* Braunschweig: Vieweg.

The Urals

D Brown, Instituto de Ciencias de la Tierra 'Jaume Almera', CSIC, Barcelona, Spain
H Echtler, GeoForschungsZentrum Potsdam, Potsdam, Germany

Introduction

The Uralide Orogen was one of the main mountain belts built during the Palaeozoic assembly of the supercontinent Pangaea. Since the breakup of Pangaea in the Mesozoic, the Uralides have remained intact and are today located in the interior of the Eurasia Plate. The current extent of the orogen is seen by its roughly north- and south-oriented magnetic signature, which abruptly interrupts that of Baltica, Kazakhstan, and Siberia, the tectonic plates that collided to form the Uralides (Figure 1A). Most of what is known about the Uralide orogen is confined to the present-day Ural Mountains, a narrow range of low to moderate topography extending for nearly 2500 km from near the Aral Sea in the south to the islands of Novaya Zemlya in the Arctic Ocean (Figure 1B). East and south of the Ural Mountains, much of the orogen is buried beneath Mesozoic and Cenozoic sediments of the West Siberian and Precaspian basins and is not well known. For descriptive purposes, the geology of the Uralide Orogen has been divided into a number of longitudinal zones that are largely based on the ages and palaeogeography of the dominant rocks within them. From west to east, these zones are the Pre-Uralian zone, the West Uralian zone, the Central Uralian zone, the Magnitogorsk–Tagil zone, the East Uralian zone, and the Trans-Uralian zone (Figure 1C). Additionally, the Uralides have been divided geographically into the South, Middle, North, Cis-Polar, and Polar Urals; the following discussions focus on the South and Middle Urals (Figure 2).

The Pre-Uralian, West Uralian, and Central Uralian zones, which together make up the western foreland thrust-and-fold belt, contain Late Carboniferous to Early Triassic sediments of the foreland basin, Palaeozoic platform and slope sediments of the Baltica margin, and Archaean and Proterozoic rocks of the East European Craton (that part of the cratonic nucleus of Baltica that took part in the Uralide orogeny) (*see* **Europe:** East European Craton). The Magnitogorsk–Tagil zone is made up of two volcanic arcs, the Magnitogorsk Arc (South Urals), which is composed of Lower Devonian to Middle Devonian basalts that are overlain by Upper Devonian volcanoclastic sediments, and the Tagil Arc (Middle Urals), which is composed of Silurian to Lower Devonian basalts and volcanoclastic sediments that are locally overlain by Lower and Middle Devonian sediments. The East Uralian zone is composed predominantly of deformed and metamorphosed volcanic arc fragments with minor amounts of Precambrian and Palaeozoic rocks thought to represent continental crust. The East Uralian zone was extensively intruded by Carboniferous and Permian granitoids, forming the 'main granite axis' of the Uralides. The Trans-Uralian zone is composed of Devonian and Carboniferous volcanic and plutonic complexes overlain by terrigenous redbeds and evaporites. Ophiolitic material (oceanic crust) and high-pressure rocks have also been reported.

Tectonic Evolution

The tectonic evolution of the Uralides (Figure 3) began during the Devonian as intra-oceanic subduction formed the Magnitogorsk and Tagil island arcs (island arcs form along convergent margins when one oceanic plate subducts beneath another oceanic plate). This was followed by the entry of the continental margin of Baltica into the subduction zone and the emplacement of an accretionary complex over the continental margin, and the exhumation of high-pressure rocks (the process of bringing the rocks in question from depth toward the surface) along the arc–continent collision boundary. At the same time, subcontinental subduction (in which oceanic crust subducts beneath continental crust) and volcanic arc formation appear to have been taking place along the margin of Kazakhstan, and to have continued throughout the Carboniferous. Throughout much of the Carboniferous, there was a deformation hiatus along the Baltica margin, which by then included the accreted volcanic arcs, and shallow-water platform margin sedimentation continued undisturbed. By the latest Carboniferous to Early Permian, the Uralian ocean basin had closed completely and the continent–continent collision between the Kazakhstan and Baltica plates had begun. As this collision progressed through the Early Permian to the Early Triassic, the western foreland thrust-and-fold belt and foreland basin of the Uralides developed, while widespread strike–slip faulting, with exhumation of lower crustal material accompanied by melt generation and granitoid emplacement, took place in the interior part of the orogen. An episode of Early Triassic extension and volcanism followed in the eastern

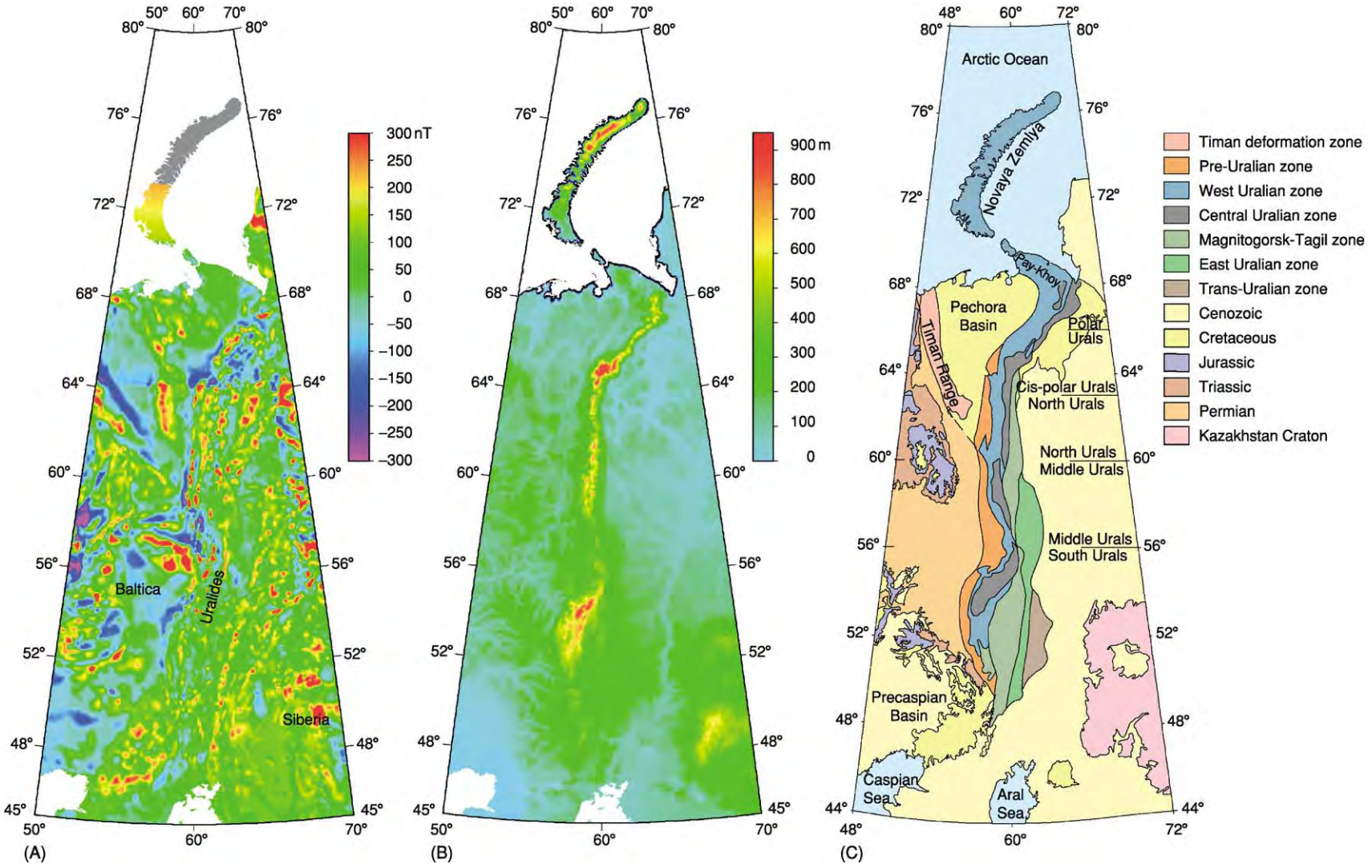


Figure 1 (A) Total field aeromagnetic map. The north east trending short wavelength anomalies represent the Uralides. Courtesy of National Geophysical Data Center. (B) Topography of the Ural Mountains. Note that the topography corresponds only to the western part of the orogen. Courtesy of National Geophysical Data Center. (C) Map of the Uralide Orogen outlining the extent of the different zones and geographical areas discussed in the text.

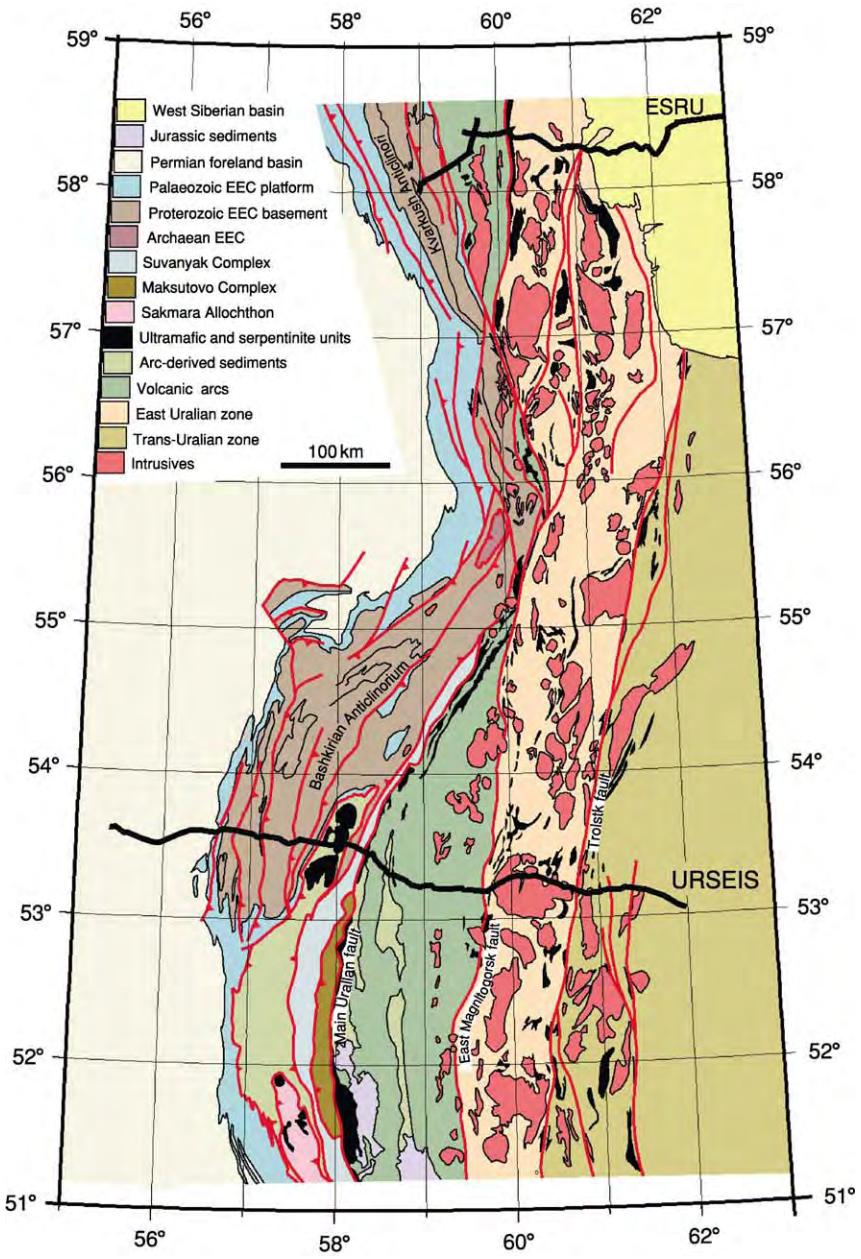


Figure 2 Geological map of the South and Middle Urals. The location of the Europrobe Seismic Reflection profiling in the Urals (ESRU) and Urals Seismic Experiment and Integrated Studies (URSEIS) transects are shown.

part of the Middle Urals and northward; by the mid-Jurassic, the intraplate Old Cimmerian deformation event led to the uplift of the Timan Range and formation of the Pay–Khoy–Novozemelian foldbelt in the northernmost Uralides (Figure 1C), and localized the deformation southward.

Arc–Continent Collision

Throughout geological time, intra-oceanic island arc development and its subsequent collision with a continental margin have been important processes in collisional orogenic belts, and among the most important means by which Earth's continental crust has grown. In the case of the Uralides, the Tagil and Magnitogorsk island arcs developed from the Silurian (Tagil) and Early Devonian (Magnitogorsk) and began to collide with the margin of Baltica in the Middle Devonian (Magnitogorsk) and the Early Carboniferous (Tagil) (Figure 3 shows this evolution schematically for the Magnitogorsk Arc). The Tagil Arc was pervasively deformed and metamorphosed to lower greenschist facies. Deformation in the

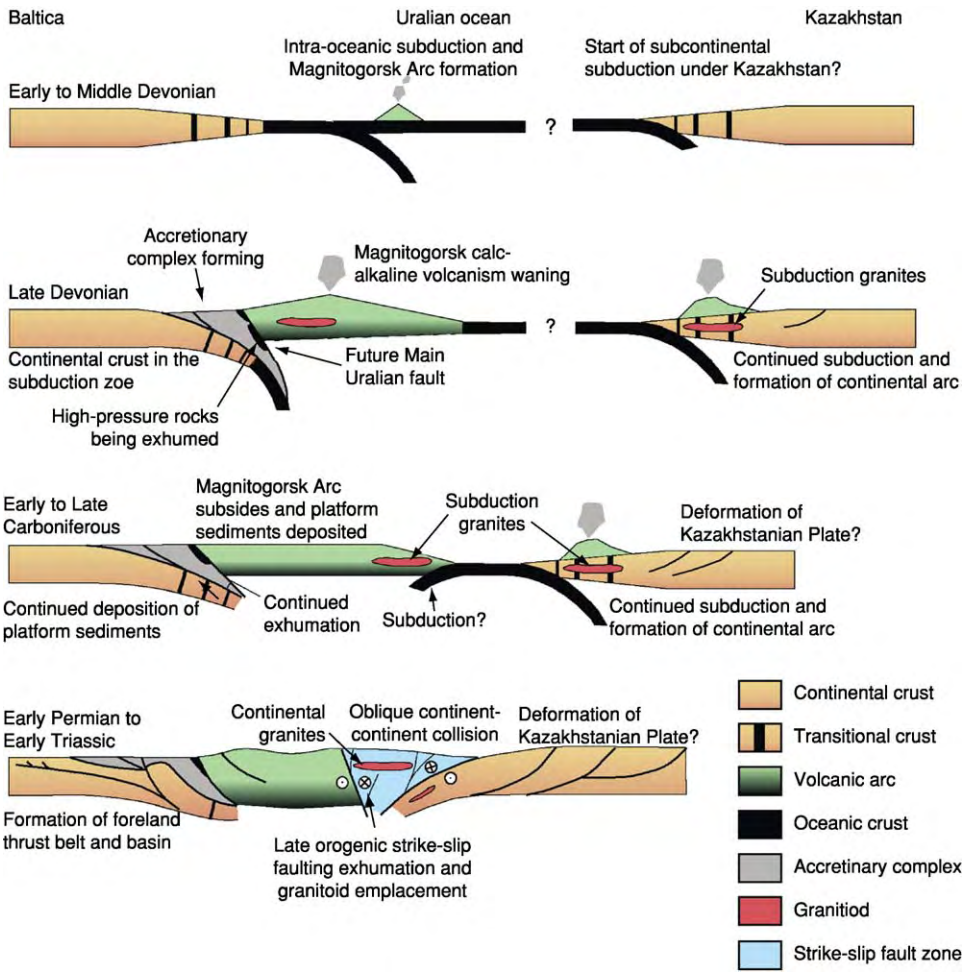


Figure 3 A simplified model for the tectonic evolution of the Uralides from the Early Devonian to the Early Triassic along the latitude of the Magnitogorsk Arc, highlighting the geodynamic processes that were active at each stage.

Magnitogorsk arc is low, with only minor, open folding and minor thrusting. The metamorphic grade is prenhite–pumpellyite facies. The Magnitogorsk arc–continent collision resulted in the development of an accretionary complex (Figure 3) that involved continental slope and platform sedimentary rocks (Suvanyak Complex) (Figure 2) that were detached from the margin of Baltica and thrust westward over the continental margin. The Suvanyak Complex is overthrust by syncollisional volcanoclastic sediments sourced from the accretionary complex and the Magnitogorsk arc (Zilair Formation). These units are flanked to the east by the high-pressure eclogite- and blueschist-bearing gneisses of the Maksutovo Complex (Figure 2), which records a peak metamorphic pressure of 20 ± 4 kbar (0.3 kbar is roughly equivalent to burial 1 km deep in Earth's crust) and temperature of 550 ± 50°C, and a peak metamorphic age of 380 to 370 Ma (during the Middle Devonian). The highest structural level of the accretionary complex is the Sakmara Allochthon in the south (Figure 2) and the Kraka lherzolite massif (a piece of oceanic mantle) in the north. The east-dipping Main Uralian fault, along which the arcs are sutured to the continental margin, is a fault melange that contains, among other things, several kilometre-scale fragments of oceanic crust and mantle.

Subcontinental Subduction

Little is known about what happened on the margin of the Kazakhstan Plate prior to or during its collision with Baltica, because no rocks that can be unequivocally assigned to the plate have been recognized in the Uralides. However, studies suggest subcontinental subduction and the development of a continental volcanic arc did occur. In part, the evidence for this comes from Silurian- to Devonian-age mafic to felsic gneisses and volcano-sedimentary rocks in the East Uralian zone that appear to represent a volcanic arc complex. The key piece of data for assigning these

rocks to a continental arc setting is the presence of subduction-related granitoids that intrude into them. These granitoids are thought to have formed in two subduction zone settings. The first subduction related magmatism occurred from about 370 to 350 Ma (during the Late Devonian to Early Carboniferous) and produced granitoids with a recognizable older continental component. These granitoids are interpreted to have been related to the development of a continental arc during subduction of oceanic crust under the continental margin of the Kazakhstan plate. A second phase of subduction magmatism occurred from about 335 to 315 Ma (during the Early to Late Carboniferous) and produced granitoids with little, if any, continental component. These granitoids are interpreted to have been related to melting of the earlier continental arc during a subsequent or continued subduction beneath the Kazakhstan Plate margin. Magmatic activity directly related to subduction ended after the Carboniferous.

The Foreland Thrust-and-Fold Belt

The western foreland thrust-and-fold belt and foreland basin of the Uralides developed from the Late Carboniferous to the Early Triassic. The foreland thrust-and-fold belt trends roughly north and south and measures ~50 to 150 km in width from the Main Uralian Fault to the deformation front. Rocks involved in the thrust belt include Proterozoic and Archaean basement of the East European Craton, Palaeozoic platform sediments of the Baltica margin, the arc–continent collision accretionary complex, and the Permian to Early Triassic foreland basin sediments. Along its eastern margin, rocks in the Kvarkush anticline (Middle Urals) and the Bashkirian anticline (South Urals) were deformed in the Neoproterozoic, and structures were reactivated and the rocks were deformed again during the Uralide Orogeny. The structural architecture of the foreland thrust-and-fold-belt is that of a west-verging thrust stack developed above a basal detachment that lies in the Proterozoic basement. Where it has been determined by balanced cross-section restoration (the procedure of restoring the layers depicted in a cross-section to the positions they had prior to deformation), the Palaeozoic shortening is approximately 20 km or less.

Late Orogenic Strike–Slip Faulting

The internal part of the Uralides was extensively affected by a late orogenic strike–slip fault system that extends north and south for more than 700 km before it disappears beneath Mesozoic and younger sedimentary cover. Throughout much of the Middle and South Urals, this strike–slip fault system coincides with the East Uralian zone (see **Figure 2**), although the currently defined Main Uralian fault appears to be its western limit in the Middle Urals, so it therefore includes the Tagil Arc. Estimates of displacement along some strands of this fault system range from a few tens of kilometres to more than 100 km. Isotopic dating on one segment of the fault system indicates an age of 247 to 240 Ma (during the Late Permian to Early Triassic) for the development of fault-related mylonites, and 305 to 291 Ma (during the latest Carboniferous) for associated metamorphic rocks. The late orogenic strike–slip fault system was extensively intruded by continental-type granitoids, first in the southern part, from 292 to 280 Ma (from the latest Carboniferous to earliest Permian), and then in the northern part, from 270 to 250 Ma (during the Late to Early Permian). These granitoids have an unusually primitive Sr and Nd isotopic composition that is thought to have resulted from remelting of the older continental arc.

Crustal Structure

The crustal structure of the Uralides has been determined from the integration of surface geology with a variety of geophysical data, including reflection and refraction seismic surveys, potential fields (gravity and magnetics), and the thermal regime. The Europrobe Seismic Reflection profiling in the Urals (ESRU) and the Urals Seismic Experiment and Integrated Studies (URSEIS) experiments, together with Russian reflection and refraction seismic surveys, and the Urals Wide-Angle Reflection Seismics (UWARS) experiment, provide a large dataset for interpreting the crustal structure of the Uralides. These data show that the Uralides still preserves its bivergent collisional structural architecture and confirms the existence of a crustal root along the central axis of the orogen (**Figure 4**). The crustal structure has been determined by physical properties along the ESRU and, in particular, the URSEIS transects (**Figure 2**).

Reflection and refraction seismic data show that the Uralide crust, in the East European Craton, thickens eastward, from ~40 to ~52 km to between 50 and 55 km, across the volcanic arcs, before thinning again to between 40 and 45 km in the easternmost part of the orogen. In the South (URSEIS) and Middle (ESRU) Urals, the East European Craton part of the Uralide crust is imaged by subhorizontal to east-dipping reflectivity that can be related to its Uralide and older orogenic events (**Figure 5**). In both datasets the East European Craton extends eastward beneath the Magnitogorsk–Tagil zone. The Magnitogorsk (URSEIS)–Tagil (ESRU) zone displays moderate to weak upper crustal reflectivity, but the middle and

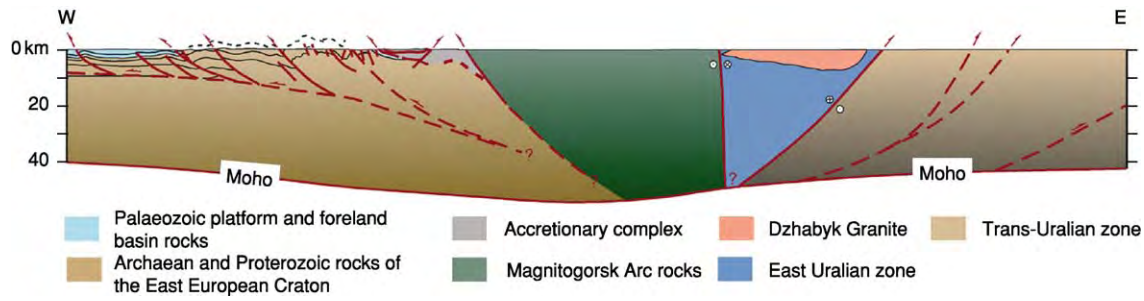


Figure 4 Schematic crustal cross section along the URSEIS profile (see Figure 2), showing the bivergence of the Uralides and the crustal root beneath the Magnitogorsk arc. The geometry of the western foreland thrust and fold belt is constrained by surface geology and reflection seismic and borehole data, and has been balanced and restored.

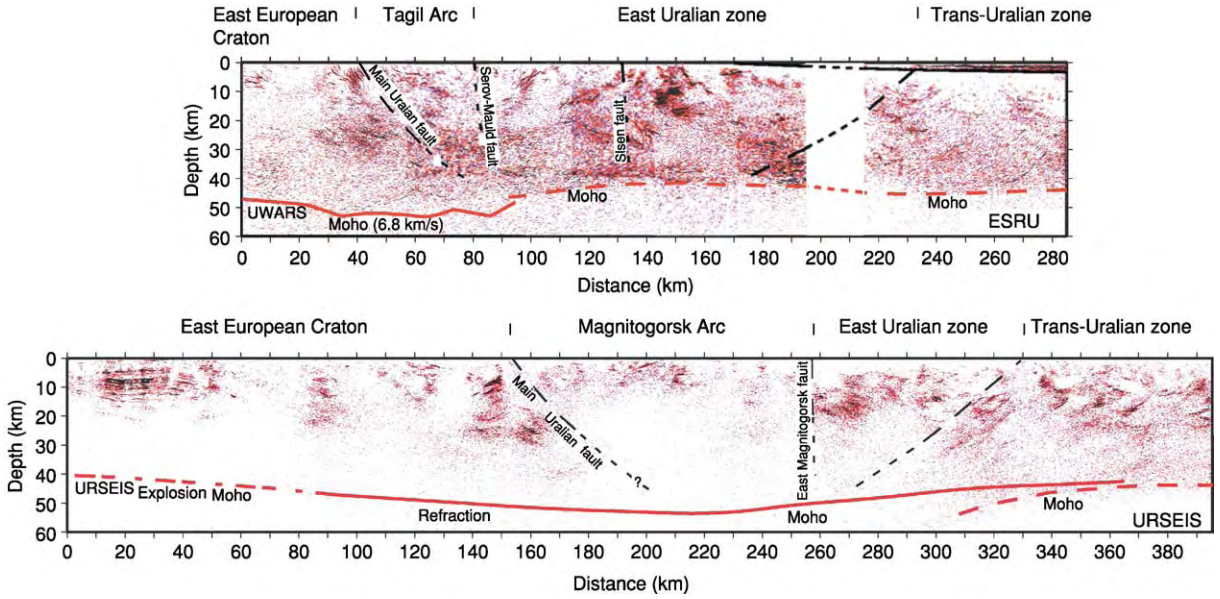


Figure 5 Automatic line drawings of the ESRU and URSEIS reflection seismic profiles, with the major crustal boundaries and the Moho indicated. These profiles clearly show the bivergent collision architecture of the Uralide orogen and the thickening of the crust towards its central axis.

lower crust reflectivity is diffuse in the case of the Magnitogorsk Arc, but quite strong in Tagil. The crust-to-mantle transition, or Moho, is not imaged beneath the Magnitogorsk zone, but is a fairly sharp transition beneath the Tagil zone. East of the arc complexes, the upper and middle crust region is imaged as clouds of diffuse reflectivity interspersed with, or cut by, sharp, predominantly west-dipping reflections that extend from the middle part of the crust into the lower crust, where it appears to merge with the Moho. In the Middle Urals (ESRU), it is characterized by abundant lower crustal reflectivity. In both the URSEIS and ESRU data, the Trans-Uralian zone dips westward beneath the East Uralian zone.

The velocity structure of the Uralide crust is best characterized along the URSEIS transect (Figure 6A and B). The upper crustal pressure wave velocities (V_p) reach up to 6.3 km s^{-1}, and the shear wave velocities (V_s) reach up to 3.9 km s^{-1}, with the higher values being in the Magnitogorsk Arc. In the middle and lower crust, V_p ranges from 6.5 to 6.8 km s^{-1}, reaching 7.1 km s^{-1} above the Moho in the central and eastern part of the transect; V_s ranges from 3.7 to 3.9 km s^{-1}, increasing to between 3.9 and 4.0 km s^{-1} at the Moho. The crust–mantle boundary is marked by an increase in V_p to >8.0 km s^{-1} and in V_s to >4.6 km s^{-1}. Using less resolved data, the Middle Urals appears to have a V_p structure similar to that of the URSEIS transect, although higher values (7.6–7.8 km s^{-1}) are reached near the Moho where the crust is thickest.

The Uralide heat flow density (HFD) is characterized by a strong minimum along the central part of the orogen (with values as low as 10 mW m^{-2}), reaching typical continental crustal values (up to

60 mW m $^{-2}$) on either side (Figure 7A). The short wavelength of the HFD anomaly is suggestive of a shallow origin for the minimum, although the heat production (k) data determined from surface samples are too high to allow the HFD minimum to be simulated along the URSEIS transect. This implies that more ultramafic material is present at depth beneath the Magnitogorsk zone, compared to below the East European Craton crust. Another option is that the HFD minimum is due to propagation of ground surface temperature changes to depth as a result of palaeoclimatic disturbances (such as recent climate change or glaciation). Modelling assuming the first alternative suggests that the Uralides thermal structure is characterized by relatively flat geotherms, with a Moho temperature of around 600°C (Figure 6C). This suggests that the root is not very cold and that the Magnitogorsk Arc rocks do not have a completely negligible heat production.

The Bouguer gravity anomaly in the South and Middle Urals (Figure 7B) is characterized by a low of between −60 and −45 mGal across the East European Craton, indicating upper and middle crustal densities of about 2.80 and 2.90 g cm $^{-3}$ (with small local variations) and lower crustal densities of 2.98 and 3.02 g cm $^{-3}$. There is an abrupt increase in the Bouguer anomaly to between about 0 to −40 mGal in the Magnitogorsk–Tagil zone, falling to between about −70 and −40 mGal in the East Uralian zone, and about −30 to −10 mGal across the East Uralian zone. This indicates an upper crustal density of between 2.71 and 2.80 g cm $^{-3}$ (somewhat higher in the Magnitogorsk zone), a middle crustal density of between 2.92 and 2.95 g cm $^{-3}$, and a lower crustal density of between 2.98 and 3.07 g cm $^{-3}$. The upper mantle has a density of 3.34 g cm $^{-3}$. The magnetic signature of the South and Middle Urals (Figure 7C) is characterized by short-wavelength features, with a long-wavelength low that reflects the magnetic character of the East European Craton. In the URSEIS transect, the magnetic crystalline basement is truncated about 50 km to the west of the Main Uralian fault, and magnetic susceptibilities of the rocks are 0 and 1.5 A m $^{-1}$ to the east and west of this truncation, respectively. Magnetic

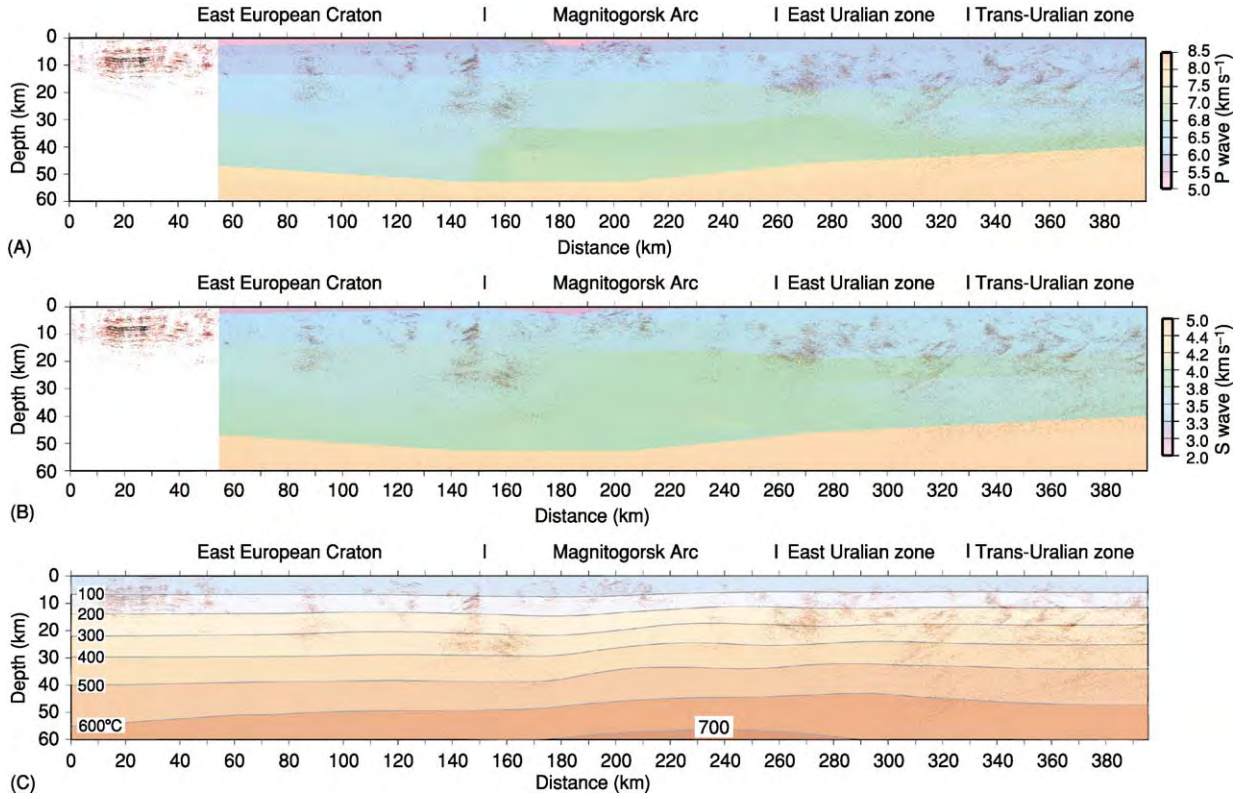


Figure 6 (A) The URSEIS reflection profile with the V_p data plotted over it; V_p is generally higher in the lower crust beneath the arc terranes. The location of the Moho is marked by the increase in V_p to greater than 8 km s $^{-1}$. (B) The URSEIS reflection profile with the V_s data plotted over it; V_s is generally higher in the middle and lower crust beneath the arc terranes. The location of the Moho is marked by the increase in V_p to greater than 4 km s $^{-1}$. (C) The URSEIS reflection profile with the present day geothermal gradient plotted over it. The temperature at the Moho is generally around 600°C, which suggests that the metamorphic grade of the Uralide crust currently does not reach granulite facies metamorphic conditions.

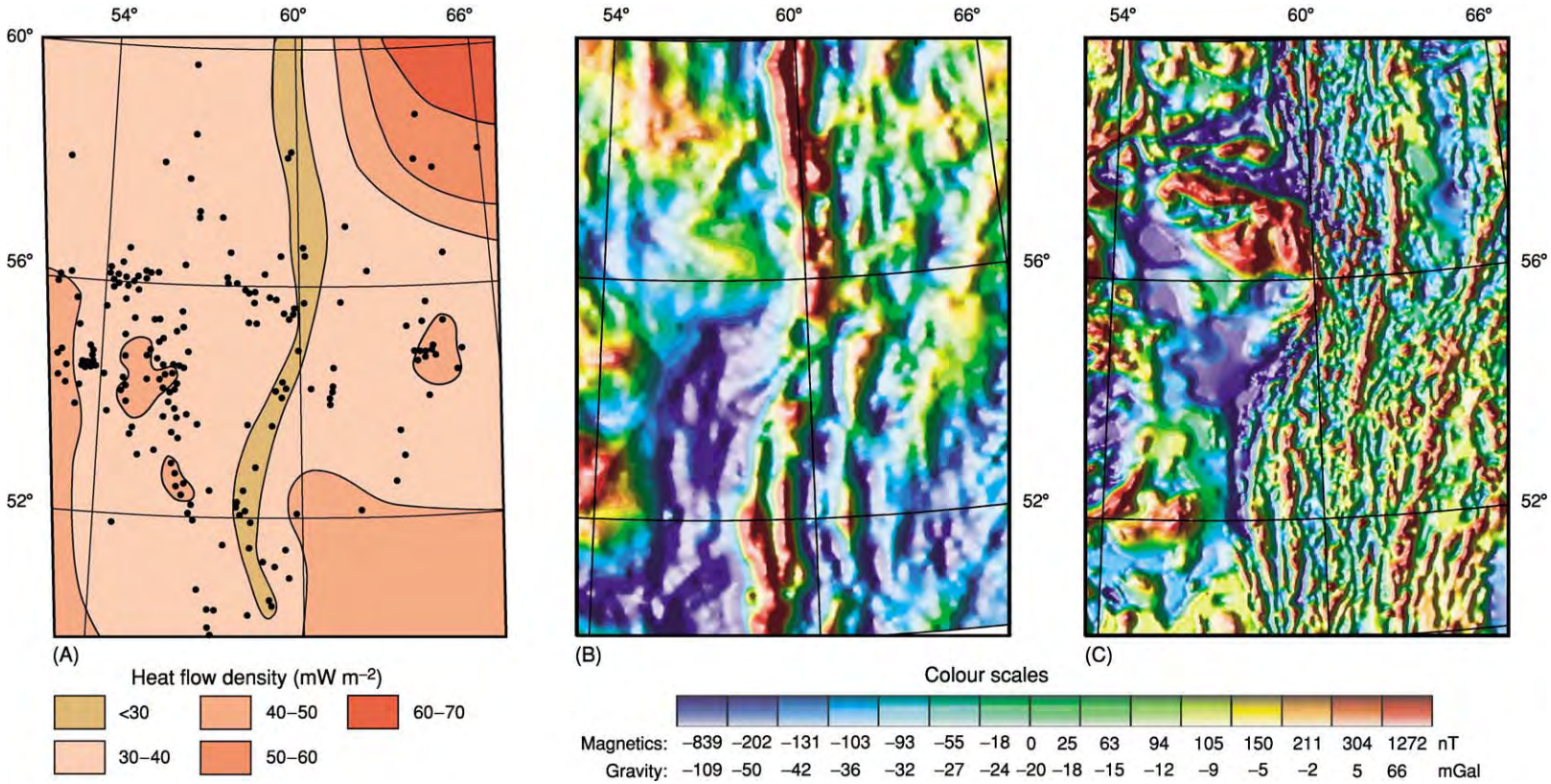


Figure 7 (A) Contoured heat flow density map of the Middle and South Urals. Dots represent data points. (B) Bouger gravity and (C) aeromagnetic maps of the Middle and South Urals. The maps are equal colour area and shaded relief, with illumination from the west. Maps courtesy of G Kimbell and C Ayala, BGS, British Geological Survey.

susceptibilities of the upper crust vary between 0.25 and 0.5 A m ^{1}, with areas reaching 2 and even 2.5 A m ^{1} locally. The middle and lower crust region has magnetic susceptibilities of 0.5 and 1.5 A m ^{1}.

Petrophysical modelling (assigning rock types based on the seismic velocity, density, and thermal data) of the Uralide crust along the URSEIS transect shows clear differences between the composition of the old continental crustal nucleus of the East European Craton and the newly added crust of the accreted arc terranes to the east. The crust of the East European Craton is more felsic than that of the Magnitogorsk and East Uralian zones, and the latter two zones have a lowermost crust with characteristics indicating a high garnet content (mafic garnet granulite) and/or the presence of hornblendite. The overall composition of the arc terranes is basaltic. The physical properties data suggest that eclogite is not present in the lower crust, or if present, it exists in such small amounts that it is below the resolution of the dataset.

Topography of the Ural Mountains

There is widespread evidence that much of the Uralides was eroded and peneplained by the Late Triassic to Jurassic. If the Uralides were peneplained by this time, when did the topography of the Ural Mountains, which locally reaches about 1900 m, form? The geomorphology of the South and Middle Ural Mountains is generally mature, being dominated by a system of smooth, north- and south-trending ridges, but with some younger features, such as deeply incised river valleys and elevated river terraces, that hint at recent uplift. The topography of the Ural Mountains is almost exclusively associated with the foreland thrust-and-fold belt and there is a strong correlation of topography with thrusts (thrusts lie in the valleys) (Figure 8). Low-temperature thermochronological studies using apatite fission track dating suggest that, with the exception of the Magnitogorsk Arc (which yields an age of 262 + 5 Ma, or Early Permian), much of the South and Middle Urals has been relatively stable since the Jurassic (ages range from 226 + 8 to 206 + 9 Ma, or Late Triassic) (Figure 8), and that since that time, very little cooling or erosion has taken place. However, there is a disturbance in the fission track age–altitude relationships across thrusts in the foreland thrust-and-fold belt, which suggests that some post-Jurassic reactivation has taken place along these faults and that the current relief of the Ural Mountains is therefore post-Uralide. (A fission track age taken from a higher altitude in the hanging wall of a thrust should be younger than one from the same stratigraphic unit at a lower level in the footwall, but this is not always the case in the Uralides.) Exactly why or when the recent relief formed is, however, not fully understood.

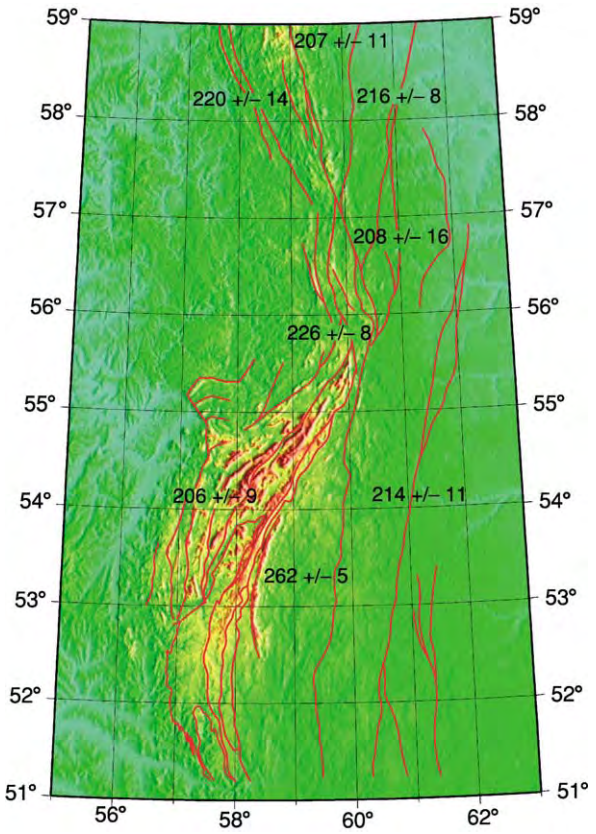


Figure 8 Topography of the South and Middle Urals, with average apatite fission track ages indicated (millions of years). Low temperature exhumation in much of the Uralides took place predominantly in the Late Triassic, although a notable exception is the Magnitogorsk Arc, where it took place in the Early Permian.

See Also

Europe: East European Craton; Timanides of Northern Russia.

Further Reading

Berzin R, Oncken O, Knapp JH, *et al.* (1996) Orogenic evolution of the Ural Mountains: results from an inte grated seismic experiment. *Science* 274: 220 221.

Brown D, Juhlin C, Alvarez Marron J, Perez Estaun A, and Oslianski A (1998) Crustal scale structure and evolution of an arc continent collision zone in the southern Urals, Russia. *Tectonics* 17: 158 170.

Brown D, Juhlin C, and Puchkov V (2002) *Mountain Building in the Uralides: Pangea to Present.* Geophysical Monograph 132. Washington, DC: American Geophysical Union.

Brown D, Carbonell R, Kukkonen I, Ayala C, and Golovonova I (2003) Composition of the Uralide crust from seismic velocities (V_p and V_s), heat flow, gravity, and magnetic data. *Earth and Planetary Science Letters* 210: 333 349.

Carbonell R, Perez Estaun A, Gallart J, *et al.* (1996) A crustal root beneath the Urals: wide angle seismic evi dence. *Science* 274: 222 224.

Echtler HP, Stiller M, Steinhoff F, *et al.* (1996) Preserved collisional crustal architecture of the Southern Urals Vibroseis CMP profiling. *Science* 274: 224 226.

Glodny J, Bingen B, Austrheim H, Molina JF, and Rusin A (2002) Precise eclogitization ages deduced from Rb/Sr mineral systematics: the Maksyutov complex, Southern Urals, Russia. *Geochimica et Cosmochimica Acta* 66: 1221 1235.

Juhlin C, Friberg M, Echtler H, *et al.* (1998) Crustal struc ture of the Middle Urals: results from the (ESRU) Europrobe Seismic Reflection Profiling in the Urals Experiments. *Tectonics* 17: 710 725.

Knapp JH, Steer DN, Brown LD, *et al.* (1996) A lithosphere scale image of the Southern Urals from explosion source seismic reflection profiling in URSEIS '95. *Science* 274: 226 228.

Meyer FM, Kisters AFM, and Stroink L (1999) Integrated geologic studies along the URSEIS '95 transect: contribu tions to the understanding of the orogenic evolution of the southern Urals. *Geologische Rundschau* 87.

Perez Estuan A, Brown D, and Gee D (1997) Europrobe's Uralides Project. *Tectonophysics* 276.

Puchkov VN (1997) Structure and geodynamics of the Ura lian orogen. In: Burg J P and Ford M (eds.) *Orogeny Through Time*, pp. 201 236. Special Publication 121. Oxford: Geological Society.

Savelieva GN and Nesbitt RW (1996) A synthesis of the stratigraphic and tectonic setting of the Uralian ophiolites. *Journal of the Geological Society* 153: 525 537.

Zonenshain LP, Kuzmin MI, and Natapov LM (1990) Ura lian foldbelt. In: Page BM (ed.) *Geology of the USSR: A Plate Tectonic Synthesis*, pp. 27 54. American Geo physical Union Geodynamics Series 21. Washington, DC: American Geophysical Union.

Permian Basins

A Henk, Universität Freiburg, Freiburg, Germany
M J Timmerman, Universität Potsdam, Potsdam, Germany

Introduction

The Late Carboniferous to Early Permian (Stephanian to Rotliegend) evolution of western and central Europe was characterized by the formation of numerous sedimentary basins as well as widespread and voluminous magmatism in the area of the Variscan orogen (*see* **Europe:** Variscan Orogeny) and its northern foreland. This phase of intense extension and thermal perturbation of the lithosphere led to an almost complete destruction of the just-formed Variscan orogen. Within 20 My, after convergence in the external zone of the orogen had ceased, extension and erosion had already removed most of the surface topography and crustal roots, respectively, and a uniform Moho depth was restored in the area of the former orogen. Towards the end of the Permian, the crust had locally subsided even below sea-level as is indicated by the rapid transgression of the Zechstein Sea over part of the former orogen.

In this article, some general characteristics of the Permo-Carboniferous evolution in western and central Europe are first summarized. Subsequently, the fill and structure of the sedimentary basins, as well as the composition and petrogenesis of the magmatic rocks formed during this epoch, are treated in some detail. Finally, an overview of the geodynamic setting and the potential driving forces for extension – topics that have been controversially debated in past years – is given.

General Characteristics

The end of convergence and the start of postconvergent evolution in the Variscan domain are documented by the youngest sediments that were still deformed by compressional movements. These occur in the most external parts of the Variscan fold belt (Ruhr basin) and are Westphalian D, about 305 Ma. However, the final stage of convergence was already accompanied by localized extension in the orogen's interior. This is documented by the rapid syn-collisional exhumation of metamorphic complexes, as well as the lateral extrusion of crustal blocks ('tectonic escape') and associated Namurian/Westphalian basin evolution next to major orogen-parallel strike–slip faults.

The subsequent Permo-Carboniferous (i.e., Stephanian to Early Permian) evolution of western and central Europe was largely controlled by a system of north-west- to south-east-trending strike–slip faults that intersected the Variscan fold belt and extended far into its foreland. Together with reactivation of older Variscan structures, movements along this fault system led to the formation of numerous transtensional half-grabens and pull-apart basins. Extension of the lithosphere and tectonic subsidence, as

indicated by rapid subsidence and sediment accumulation in these fault-controlled basins, can be dated to between about 305 and 285 Ma. This correlates with a phase of intense magmatism, peaking between 300 and 285 Ma, which occurred almost synchronously over large parts of Europe – both inside and outside the area affected by Variscan orogenic processes. Subsequent thermal equilibration of the intensely disturbed European lithosphere led to thermal subsidence. During this stage, sedimentation overstepped the initially fault-controlled basin margins. The individual basins were connected and a coherent depositional area formed in large parts of Europe. At about 250 Ma, the crust had locally subsided below sea-level, as is indicated by the rapid transgression of the Zechstein Sea over part of the former orogen. Even the subsequent Mesozoic basin evolution in large parts of Europe was still controlled by thermal equilibration of the anomalies formed by Permo-Carboniferous geodynamic processes.

Basin Formation

Late Stephanian to Early Permian basin evolution started after a regional hiatus of Early Stephanian age. This hiatus is documented within the Variscan domain as well as in its northern foreland. There, Stephanian and Early Permian sediments and volcanic rocks unconformably overlie the Westphalian foreland basin sequence. In the external fold-and-thrust belt, corresponding sediments were deposited on very-low-grade to low-grade metasediments, whereas in the internal part of the Variscides, they were often directly deposited on amphibolite-facies metamorphic rocks and granite intrusions. This demonstrates that by Late Carboniferous times, the Variscides had already been deeply eroded. Only locally Stephanian sediments unconformably overlie Namurian to Westphalian precursor basins. These basins, such as the Saar-Nahe, Central Armorican Basin (i.e., Laval), and the basins east of the Elbe Line, are usually located at major fault zones that exhibit strike–slip movements and may be related to tectonic escape during the final stage of the Variscan Orogeny (see Figure 1).

The fill of the basins consists mainly of continental siliciclastic sediments (shales, sandstones, and conglomerates) and interbedded volcanic and pyroclastic deposits. During the tectonic subsidence stage, between 305 and 285 Ma, the sedimentary sequences frequently show a typical facies evolution from fluvial to lacustrine and back to fluvio-deltaic depositional environments. Deposition in fluviatile and playa-lake environments was common during the subsequent thermal subsidence stage. In some basins, the earlier deposits contain economically important coal measures (e.g., the Saar-Nahe, Saale, Autun, and Sillon Houillier basins). The gradual decrease of coal-bearing strata from the Late Carboniferous to the Early Permian reflects the northward drift of Gondwana from an equatorial setting. Due to the

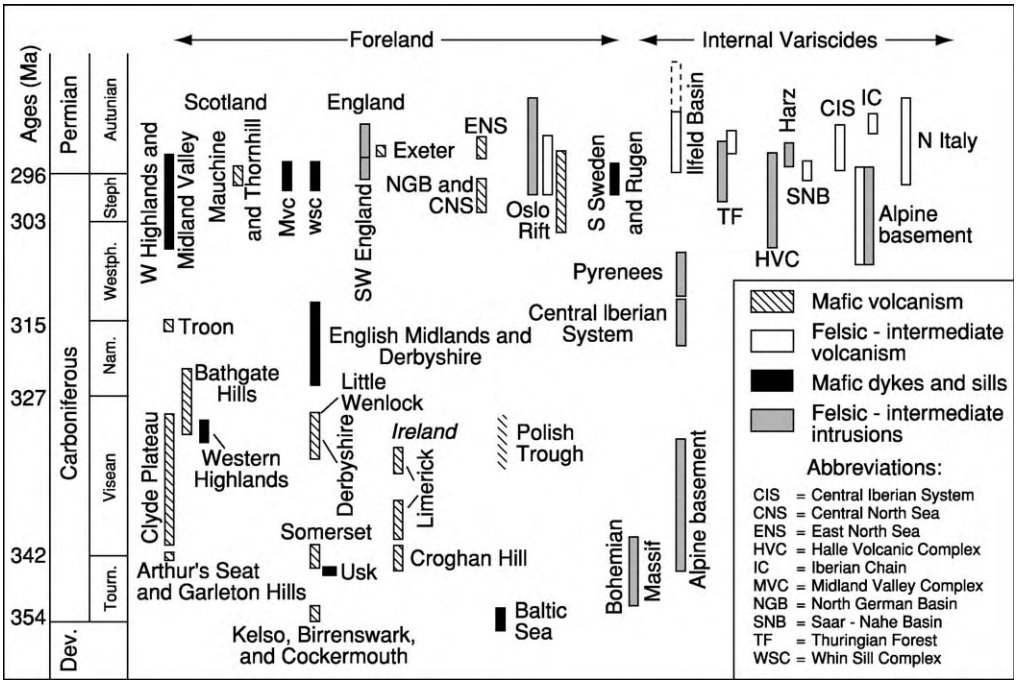


Figure 1 Age distribution of Permo Carboniferous magmatism in the Variscan domain.

resulting changes in climate, formation of substantial coal measures north of the Variscides had already ceased in the Westphalian D, whereas in the southern areas it continued until the end of the Stephanian.

The sediments and volcanic rocks accumulated during the syn-rift and subsequent thermal subsidence phases can locally reach total thicknesses of up to 8 km. However, it should be noted that the thicknesses presently observed are often only erosional remnants. At least in parts of the Variscan orogen it can be shown that a substantial part of the sediments deposited during the Stephanian to Early Permian was already eroded prior to the Zechstein transgression. In the Saar-Nahe Basin, for example, estimates based on shale compaction data and vitrinite reflectance modelling indicate removal of up to 3.7 km of sediments during this phase. Between these basins, the basement was exposed or was directly overlain by younger deposits.

The geometry and structural style of the individual basins were strongly controlled by the interplay of pre-existing structural elements and new faults that formed in conjunction with the dextral translation of Gondwana relative to Laurussia. A dextral strike–slip component is inferred from basin geometry and orientation, isopach maps of syn-rift deposits, migration direction of depocentres, the orientations and structural data from syn-sedimentary faults, simultaneous horst and basin formation, and the arrangement of volcanic centres and dykes. Thus, some of the basins trend obliquely to the structural grain of the orogen and intersect the Variscan deformation front, whereas others owe their formation to transtensional reactivation of older, compression-related faults. A prime example of the latter is the Saar-Nahe Basin in south-west Germany, which formed by oblique extensional reactivation of a crustal-scale fault structure, along which the internal zone of the Variscan orogen (Mid-German Crystalline Rise, part of the Saxothuringian Zone) was previously thrusted onto the external fold-and-thrust belt (Rhenohercynian Zone). Thus, the basins that were formed by transtensional reactivation of former thrusts frequently have half-graben structures. In contrast, pull-apart basins that opened during dextral strike–slip movements tend to be located next to the dominant north-west- to south-east-trending faults (continental-scale dextral shears parallel to the Tornquist–Teisseyre Line; e.g., the Elbe Line, Pays de Brays Fault, Bay of Biscay, Gibraltar-Minas, and Agadir fracture zones).

Several of the faults that controlled Permo-Carboniferous basin evolution were later reactivated, particularly in connection with the inversion of the Alpine foreland in the Late Cretaceous and Early Permian. These later events can sometimes obscure the Permo-Carboniferous basin geometry and fault displacement. Thus, identification of faults that were active in Permo-Carboniferous times requires careful analysis of indicators for syn-sedimentary tectonics, such as abrupt variations in thickness and facies architecture and syn-depositional deformation structures.

Magmatism

Late Carboniferous to Early Permian volcanic and plutonic activity of varying style and composition was coeval with extension and basin formation in both foreland and the internal Variscides, and occurred in crustal domains of various ages (**Figure 1**). Reliable U–Pb and $^{40}Ar/^{39}Ar$ mineral ages show that most of the activity occurred in the period 305 to 290 Ma. Magmatic rocks of mafic composition are common in the foreland, but are relatively rare in the internal Variscides. Layers of strongly altered air-fall tuffs, known as bentonites or tonsteins, occur in nearly every basin, but most of these originated from distal volcanic sources.

Extensional faulting took place before, during, and after volcanic activity. In the north-western part of the North German Basin, initial faulting caused the formation of horsts and grabens that strongly controlled the placement and thickness of the volcanic rocks. Faulting in the central North Sea mainly postdates the volcanic activity, whereas in the Oslo Rift, the main east and west extension was coeval with the main phase of trachyandesitic volcanism. In nearly all of the basins, lavas occur interbedded with sediments, showing that magmatic activity took place during subsidence and extension. The volcanic rocks are interbedded and alternate with Stephanian coal measures and Early to Late Permian, Rotliegend-facies alluvial and fluviatile clastic sediments, and lacustrine marls deposited in semiarid environments. Where not eroded, these are often followed by aeolian sandstones.

Foreland

In the foreland, a variety of volcanic and plutonic rocks are present in the Oslo Rift in south Norway, and large volumes of rhyolitic and andesitic volcanic rocks occur in northern Germany below younger cover and are known only from deep boreholes. More mafic magmatic activity resulted in the formation of dyke swarms and sills, such as the basaltic Whin Sill and the Midland Valley complexes in Great Britain, the basaltic dyke swarm in South Sweden, and the lamprophyre dyke swarms in the western Highlands of Scotland. Smaller volumes of mafic and felsic volcanic rock have been drilled in the

central and eastern North Sea, and the Midland Valley of Scotland is known for its large number of alkaline mafic sills, vents, and subvolcanic intrusions that often contain megacrysts and xenoliths of mantle material and high-grade metamorphic crust.

Several stages of magmatism have been recognized in the Oslo Rift. During the initial stages, felsic intruded Westphalian sediments, accompanied or shortly followed by the extrusion of up to 1500 m of primitive, alkaline basalt to phonotephrite lavas in the southern parts of the rift. The subsequent main rift stage resulted in fissure eruptions of thick sequences of porphyritic trachyandesite flows that constitute the largest volume of volcanic rocks. This was accompanied by the intrusion of large amounts of syenitic magmas. Following the main stage of magmatism, intrusion of dykes and batholiths of syenitic, monzonitic to granitic composition took place, often related to the collapse of large volcanoes. The Oslo Rift extends offshore into the Skagerrak to the south, where seismic data suggest the presence of a ~1-km sequence of Late Carboniferous–Early Permian lavas.

The Whin Sill and Midland Valley complexes in northern Britain consist of a series of dolerite sills and east–north-east- to east–west-trending dykes that extend eastwards into the North Sea. The dykes and sill have subalkaline to transitional basaltic compositions. In Scania in South Sweden, a north-west-trending swarm of subalkaline dolerite dykes may be related to a few isolated, north–north-west-trending dolerite dykes on the south-west coast and Bornholm Island, and to dolerite sills that intrude Cambrian shales in Västergötland. Volcanic rocks were encountered in the drill core in the Kattegat, and seismic data suggest the presence of Early Permian volcanic edifices offshore from eastern Denmark.

The Stephanian–Autunian volcanic rocks in the north-western part of the North German Basin are mainly limited to north–north-west- to north–north-east-trending grabens, and their thicknesses vary from a few metres to ~100 m. In the north-east, they form continuous layers up to 2 km thick and may have covered an area of ~180 000 km^2. Basalt lavas, dolerite, and gabbro intrusions are subordinate, and rhyolite lavas and ignimbrites, basaltic andesite and andesite lavas, and tuffs form the largest volumes. Five eruptive stages have been recognized, and the total volume has been estimated at ~48 000 km^3, approximately 70% of which are rhyolitic lavas and ignimbrites that predominate in the eastern part of the basin.

Elsewhere in the foreland, Stephanian to Autunian magmatic activity occurred in much smaller volumes. In the central and eastern North Sea, basalt, trachyandesite, and rhyolite flows and tuffs occur interbedded with Rotliegend-facies mudstones and sandstones. The volcanic rocks reach thicknesses of up to 160 m in the central North Sea, and ~680 m in the Horn Graben. In northern Britain, up to ~100- to 240-m-thick sequences of alkaline basalt lavas and tuffs occur in the Mauchline and Thornhill basins. These unconformably overlie Westphalian sediments and are interbedded with and overlain by Early Permian aeolian red sandstones. The volcanic rocks are probably related to the approximately 60 mafic vents that occur within a 20-km radius of the Mauchline Basin and predate the aeolian sandstones.

Variscan Internides

The magmatic rocks in the internal Variscides range from granite and diorite intrusions to rhyolitic–andesitic volcanic and subvolcanic rocks. Mafic rocks are limited to diatremes, small volumes of lavas, and mafic dykes. In addition, the volcanic rocks often comprise high proportions of ignimbrites and tuffs, typical for explosive volcanism. Examples are the granite and diorite intrusions in south-west England (Cornwall), the Harz, Erzgebirge, northern Italy, and the basement of the Alps, and the volcanic rocks in basins in the Thuringian Forest, Pyrenees, Iberia, northern Italy, the Provence, and the Saar-Nahe Basin. However, not all basins contain magmatic rocks; for example, in the Stephanian to Early Permian basins in the Massif Central in France, only small volumes of volcanic rocks are present.

In south-west England, postorogenic granite emplacement was probably accompanied by the extrusion of lavas, as indicated by clasts of rhyolite and tuff within Permian sediments. The main magmatic activity occurred within the relatively short period about 4.5 Ma. Granite intrusion and subsequent uplift and erosion partly overlapped with the formation of narrow, east–west-trending, fault-bounded grabens. These contain Rotliegend-facies clastic sediments and small volumes of olivine basalt and lamprophyric lavas and agglomerates. The lamprophyre lavas are genetically related to the minette dykes that occur throughout south-west England.

The Saar-Nahe Basin contains Early Permian andesite flows and pyroclastics that are associated with subvolcanic dacite and rhyolite domes and diatremes of subalkaline basalt. In the south-westwards extension in France, Early Permian andesites, rhyolites, and rhyolitic breccias occur in boreholes and reach a thickness of over 1 km south of Nancy. The small Ilfeld Basin in the south Harz contains a ~800-m-thick sequence of sediments and thick layers of tuffs, ignimbrites, and latitic, trachytic, and rhyolitic flows. Of similar age are the undeformed or only weakly

deformed granitoid and gabbro intrusions in the nearby Thuringian Forest and Harz area, and the high-level rhyolite sills near Halle. Differential uplift in the Thuringian Forest, associated with granite and diorite magmatism, led to block faulting and the formation of horsts and pull-apart basins, the latter filled by an up to 2-km-thick sequence of molasse sediments, lavas, and pyroclastic rocks of mainly trachyandesitic and rhyolitic composition.

Magmatic activity in Iberia and the Pyrenees took place in at least two stages, Late Carboniferous to Early Permian and mid-Permian to Triassic. The first phase is of predominantly calc-alkaline composition, whereas the younger magmatic rocks are more alkaline. Small basins and half-grabens contain sequences of Late Westphalian C to Autunian terrestrial sediments and volcanic rocks, the latter mainly of pyroclastic character, but also comprising volcanoclastic rocks and ash flows. Compositions range from andesite, to dacite, to rhyolite, but andesites predominate. Volcanism was accompanied by intrusion of hypabyssal sills, dykes, and domes, by high-level, often composite and hybrid, granitoid intrusions, and by several generations of granitic to dioritic dykes. The earliest volcanism in the Pyrenees may partly overlap with the intrusion of high-level calc-alkaline, often composite granitic to dioritic, plutons that have Westphalian (305–312 Ma) U–Pb zircon crystallization ages.

The pre-Mesozoic basement of the Alps, exposed in the external massifs and tectonic windows, contains Late Carboniferous to Early Permian granitoid intrusions and volcanic and sedimentary rocks. Narrow pull-apart basins contain andesite to rhyolite lavas and rhyolitic tuffs that were deposited on deformed and metamorphosed Visean rocks, suggesting several kilometres of uplift and erosion before 300 Ma. South of the Alps, near Bolzano in northern Italy, Early Permian latite, dacite, rhyodacite, and rhyolite lavas and tuffs cover an area of $\sim$4000 km^2 and locally reach a thickness of over 2 km. The Late Carboniferous to Early Permian granitoids in the basement of the Alps have a calc-alkaline, volcanic arc character, but in northern Italy, magmatic activity continued into mid-Permian times and changed to more alkaline compositions.

Petrogenesis

The varying compositions of the mafic rocks in the foreland show that they were derived from different mantle sources and evolved differently. The magmatic rocks in the Midland Valley of Scotland were derived by low-degree melting of deep mantle sources followed by rapid ascent, as indicated by their alkaline, magnesium-rich, and silica-poor composition, by the absence of low-pressure differentiates, by the presence of abundant megacrysts and of mantle and lower crustal xenoliths, and by the style of volcanism. The early alkaline mafic volcanic rocks in the Oslo Rift may have been derived from mantle sources that were metasomatically enriched by carbonatite fluids in earliest Palaeozoic times. Lithospheric extension causing decompression melting is a probable mechanism, but geochemical evidence suggests that a mantle plume component cannot be completely ruled out. In contrast, the subalkaline basalts of the Whin Sill and Midland Valley complexes indicate higher proportions of melting of shallow mantle sources. Despite being distributed over a relatively large area, their distribution need not reflect a mantle thermal anomaly of the same extent. The geometry and orientation of the dyke swarms suggest a magmatic focal region in the vicinity of the Denmark–Skagerrak region, and magma transport may have been horizontal, westwards into the North Sea and Britain. The position, trend, number, and size of the dykes may have been controlled by the regional dextral extensional stress field.

The magmatic rocks in the North German Basin and the internal Variscides have predominantly felsic to intermediate (rhyolitic to andesitic) compositions. Sr–Nd isotope data and the presence of garnet and crustal xenoliths indicate that their parent melts assimilated large amounts of crustal material, or that the rhyolitic end-members were derived by the melting of older crust. Furthermore, the scarcity of mafic magmatic rocks, in combination with the fractionated character and degree of alteration of the felsic to intermediate rocks, makes it difficult to establish the nature of the mantle sources. The granite intrusions and the felsic volcanic rocks often have calc-alkaline compositions suggesting subduction-related volcanic arc origins. This does not agree with the intracontinental extensional setting of the magmatic activity. Furthermore, most Variscan oceans had closed by Visean times. Instead, the volcanic arc signature may have been inherited from mantle sources that had been metasomatized by previous subduction events, caused by extensive assimilation of continental crust by the parent melts, and/or inherited through partial melting of older, calc-alkaline lower crust. The fact that Stephanian to Autunian mafic rocks are much rarer in the internal Variscides, compared to the foreland, suggests that the mantle-derived parent melts were unable to reach the surface directly. Instead, they stalled in lower to mid-crustal magma chambers, where they assimilated crustal material and fractionated to more felsic compositions before erupting. The large amount of crustal melts (rhyolites) in, for instance, the North German Basin may have been

generated by underplating of mantle-derived melts. However, there is little geophysical evidence for lower crustal mafic intrusions in this area. In the central and eastern North Sea, Denmark, and the Skagerrak Graben, geophysical evidence such as deep crustal reflectors and Bouguer anomalies suggest that, here, volcanism may have been accompanied by crustal underplating, although the age of these features remains uncertain.

The effects of Permo-Carboniferous magmatism at various crustal levels can be studied exemplarily in the Ivrea Zone of northern Italy. This tilted crustal fragment provides one of the rare opportunities in the world to study a coherent crustal succession from the Moho to upper crustal levels. The Ivrea Zone achieved its present structure largely at the end and shortly after the Variscan Orogeny. Intrusion of mantle-derived basic magmas at or near the base of the crust caused partial melting of lower crustal rocks. This in turn generated granitic magmas that ascended to middle and upper crustal levels. At the surface, magmatism is also documented by volcanic activity in the contemporaneous sedimentary basins.

Geodynamic Setting and Driving Forces for Extension

The geodynamic setting of the Variscan domain during the Permo-Carboniferous and the driving forces for crustal re-equilibration have been matters of intense debate. A prime feature is the strong thermal perturbation of the lithosphere, which has been attributed to diverse processes such as slab detachment, delamination and thermal erosion of the mantle lithosphere, crustal extension, or ascent of mantle plumes. Some explanations even manage without substantial crustal extension and assume eclogitization of the lower crust as the main process for the disappearance of the Variscan crustal roots. In this respect, it is interesting to note that basin formation and substantial magmatism occur also in the northern foreland of the Variscides (i.e., in an area not affected by Variscan crustal and lithospheric thickening). Thus, the processes often invoked for postconvergent settings may not hold in these areas. This leads to the fundamental question: were Permo-Carboniferous basin formation and magmatism genetically linked to the preceding Variscan Orogeny, or do they document a new geodynamic regime?

Because of the close temporal and spatial relationship between crustal thickening and postconvergent extension, several authors have suggested a gravitational collapse of the Variscan orogen as the dominant process controlling the geodynamic evolution of central Europe in latest Carboniferous to Permian times. The Tibetan Plateau of China and the Basin and Range Province of the United States have been proposed as modern analogues for theLate Palaeozoic destruction of the Variscan orogen. Other authors, however, have suggested that the postconvergent destruction of the Variscides was caused primarily not by body forces, but by a change in orientation of the far-field stress regime. The onset of the Stephanian to Early Permian basin formation indeed coincides with the change in the relative movement between Gondwana and Laurussia, from head-on collision to dextral translation.

To solve this dispute, it is crucial to establish whether the Variscan orogen still had excessive crustal and lithospheric thicknesses at the beginning of the Stephanian, which would have driven orogenic collapse. Unfortunately, no direct evidence is available, neither on the palaeotopography of the Variscan orogen nor on crustal thicknesses. Combining several lines of evidence, a maximum crustal thickness of about 50 km at the onset of the Permo-Carboniferous evolution seems likely. Estimates for the total lithospheric thickness can be constrained by the widespread Visean (~330–340 Ma) high-temperature/low-pressure metamorphism and granite magmatism that are so characteristic for the Variscan orogen. The magnitude of this thermal event, which reached lower crustal temperatures in excess of 900°C, indicates the absence of thick mantle lithosphere beneath the internal zone of the Variscan orogen at this time. Thus, if delamination or convective erosion of the mantle lithosphere occurred during the Variscan Orogeny, it must have occurred prior to 330 Ma. Consequently, none of these processes can be invoked to generate surplus potential energy and gravitational collapse for the Stephanian to Early Permian evolution.

Quantitative analysis using thermomechanical finite element models suggests that the gravitational instability of the Variscan orogen was insufficient to explain the observed amount and timing of crustal extension. In order to overcome the finite strength of the crust and restore a uniform crustal thickness in the area of the former orogen and its northern foreland, tensile plate boundary stresses are required to have operated until about 285 Ma. The cause of the far-field stresses can be attributed to the dextral translation of Laurussia relative to Gondwana during the Stephanian and Early Permian. Contemporaneous crustal shortening of 300–400 km in the linked Appalachian–Mauretanide orogen also provides an estimate for the total amount of regional extension that affected western and central Europe during this stage. Local extension estimates from, for instance, subsidence analysis and section balancing are available for only a few basins in the Variscan domain (see

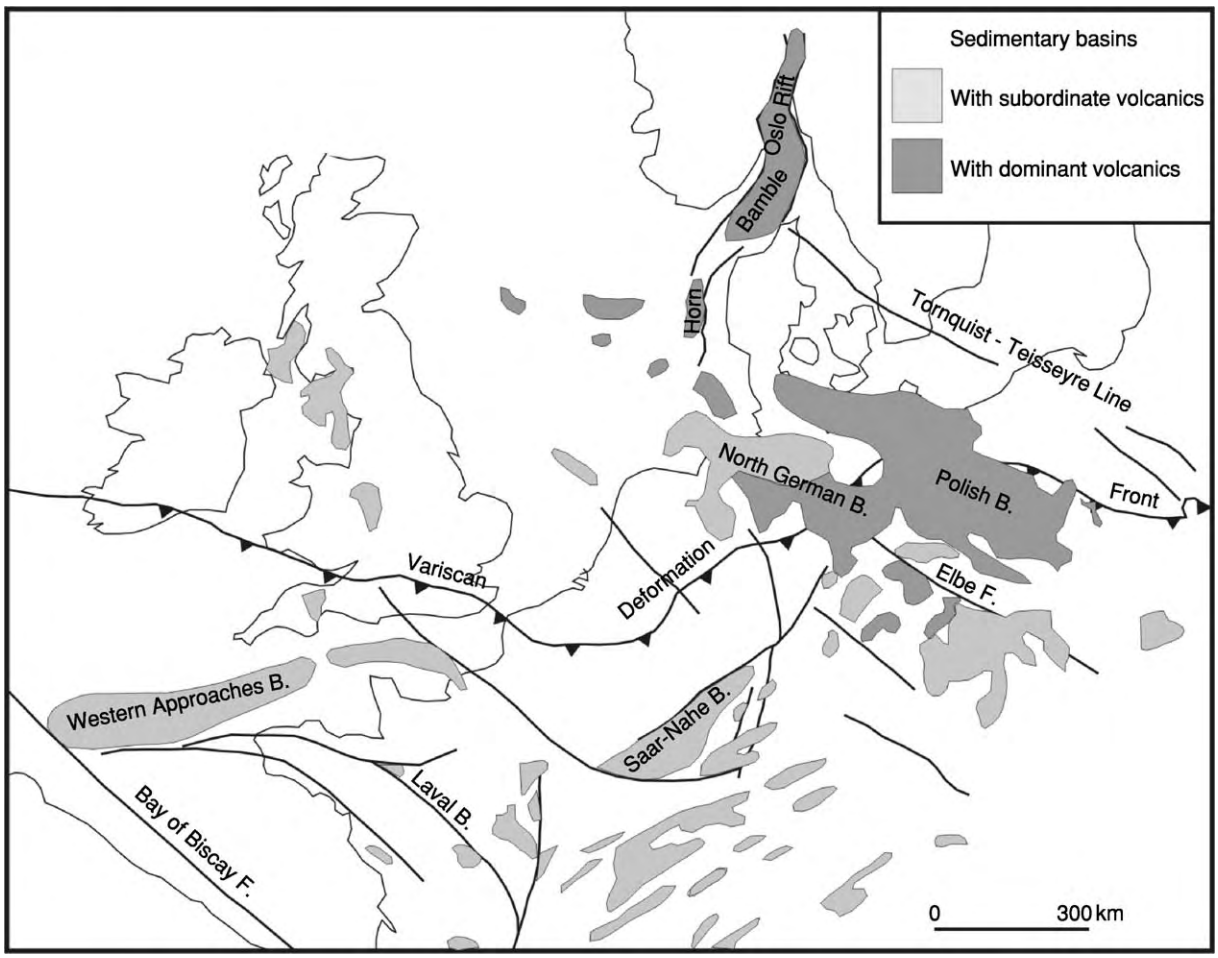


Figure 2 Permo Carboniferous sedimentary basins and volcanic centres in the Variscan realm. Modified with permission from Ziegler PA (1990) *Geological Atlas of Western and Central Europe*. The Hague: Shell Internationale Petroleum Maatschappij.

Figure 2). They indicate crustal stretching by a factor of at least 1.35. However, intrabasin highs, such as basement ridges and areas with postrift sedimentation only, do not necessarily represent unextended areas. This is because diffuse extension and ductile flow of lower crust towards the zones of maximum upper crustal extension may have contributed towards regional crustal thinning.

The decay of the thermal anomaly induced by Stephanian to Early Permian geodynamic processes, i.e., thermal contraction of the lithosphere and its re-equilibration with the asthenosphere, caused long-lasting subsidence and controlled the formation of Late Permian and Mesozoic thermal sag basins that were superimposed on the Permo-Carboniferous troughs.

See Also

Europe: Permian to Recent Evolution; Variscan Orogeny. **Lava**. **Palaeozoic:** Carboniferous; Permian. **Pyroclastics**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings; Deserts; Lake Processes and Deposits. **Tectonics:** Rift Valleys.

Further Reading

Arthaud F and Matte P (1977) Late Paleozoic strike slip faulting in southern Europe and northern Africa: results of a right lateral shear zone between the Appalachians and the Urals. *Geological Society of America, Bulletin* 88: 1305 1320.

Benek R, Kramer W, McCann T, *et al.* (1996) Permo Car boniferous magmatism of the Northeast German Basin. *Tectonophysics* 266: 379 404.

Burg J P, van den Driessche J, and Brun J P (1994) Syn to post thickening extension in the Variscan Belt of Western Europe: modes and structural consequences. *Géologie de la France* 3: 33 51.

Cortesogno L, Cassinism G, Dallagiovannam G, Gaggerom L, Oggiano G, Ronchi A, Seno S, and Vanossi M (1998) The Variscan post collisional volcanism in Late Car boniferous Permian sequences of Ligurian Alps,

Southern Alps and Sardinia (Italy): a synthesis. *Lithos* 45: 305 328.
Finger F, Roberts MP, Haunschmid B, Schermaier A, and Steyer HP (1997) Variscan granitoids of central Europe: their typology, potential sources and tectonothermal relations. *Mineralogy and Petrology* 61: 67 96.
Floyd PA, Exley CS, and Styles MT (1993) *Igneous Rocks of South West England. Geological Conservation Review.* vol. 5. London: Chapman & Hall.
Glennie KW (1999) Lower Permian Rotliegend. In: Glennie KW (ed.) *Petroleum Geology of the North Sea: Basic Concepts and Recent Advances,* pp. 137 173. Oxford: Blackwell Science.
Henk A (1999) Did the Variscides collapse or were they torn apart?: a quantitative evaluation of the driving forces for postconvergent extension in central Europe. *Tectonics* 18: 774 792.
Plein E (1995) Norddeutsches Rotliegend Becken. Rotlie gend monographie, teil II. *Courier Forschungsinstitut Senckenberg* 183: 1 193.
Schaltegger U (1997) Magma pulses in the Central Variscan Belt: episodic melt generation and emplacement during lithospheric thinning. *Terra Nova* 9: 242 245.
Sundvoll B, Neumann E R, Larsen BT, and Tuen E (1990) Age relations among Oslo Rift magmatic rocks: implica tions for tectonic and magmatic modelling. *Tectonophy sics* 178: 67 87.
Ziegler PA (1990) *Geological Atlas of Western and Central Europe.* The Hague: Shell Internationale Petroleum Maatschappij.

Permian to Recent Evolution

P A Ziegler, University of Basel, Basel, Switzerland

Introduction

Large parts of Western and Central Europe (WCE) are occupied by sedimentary basins that contain up to 8 km thick Permian to Cenozoic series (**Figure 1**). These basins are variably underlain by the Precambrian crust of the East-European-Fennoscandian Craton (*see* **Europe:** East European Craton) and its Late Precambrian to Early Palaeozoic sedimentary cover, by the Precambrian Hebridean Craton, by the Caledonian crust of the British Isles (*see* **Europe:** Caledonides of Britain and Ireland), the North Sea, Northern Germany and Poland and its Devonian and Carboniferous sedimentary cover, and by the Variscan fold belt in which orogenic activity had ceased at the end of the Westphalian. The present crustal configuration of WCE bears little relationship to the Caledonian and Variscan orogenic belts, but is closely related to the geometry of the Late Permian, Mesozoic and Cenozoic sedimentary basins and the Alpine orogen (**Figure 2**). This reflects that dynamic processes, which governed the evolution of the Late Permian and younger sedimentary basins, had a strong impact on the crustal configuration of WCE, and that the crustal roots of the Caledonian and Variscan orogens had been destroyed shortly after their consolidation.

During Permian to recent times, the megatectonic setting of WCE underwent repeated changes. Correspondingly, dynamic processes controlling the evolution and partial destruction of sedimentary basins also changed through time. Therefore, in some areas, basins of differing tectonic origin are stacked on top of one other.

The following main stage are recognized in the Late Permian to recent evolution of WCE, namely: (i) Late Permian–Early Cretaceous rifting during Pangaea breakup; (ii) Late Cretaceous–Paleocene rifting and early Alpine intraplate compression; and (iii) Eocene-recent opening of the Arctic–North Atlantic and collisional interaction of the Alpine Orogen with its foreland.

Background: Late Hercynian Wrench Tectonics and Magmatism

Following its Late Westphalian consolidation, the Variscan Orogen (*see* **Europe:** Variscan Orogeny) and its northern foreland were overprinted during the Stephanian to Early Permian by a system of continent-scale dextral shears, such as the Tornquist-Teisseyre, Bay of Biscay, Gibraltar-Minas and Agadir fracture zones which were linked by secondary sinistral and dextral shear systems. This deformation reflects a change in the Gondwana-Laurussia convergence from oblique collision to a dextral translation that was kinematically linked to continued crustal shortening in the Appalachian (Alleghanian Orogeny) and the Scythian orogens. Significantly, wrench tectonics and associated magmatic activity abated in the Variscan domain and its foreland at the transition to the Late Permian in tandem with the consolidation of the Appalachian Orogen.

Stephanian to Early Permian wrench-induced disruption of the Variscan Orogen and its foreland was accompanied by regional uplift, wide-spread extrusive and intrusive mantle-derived magmatic activity

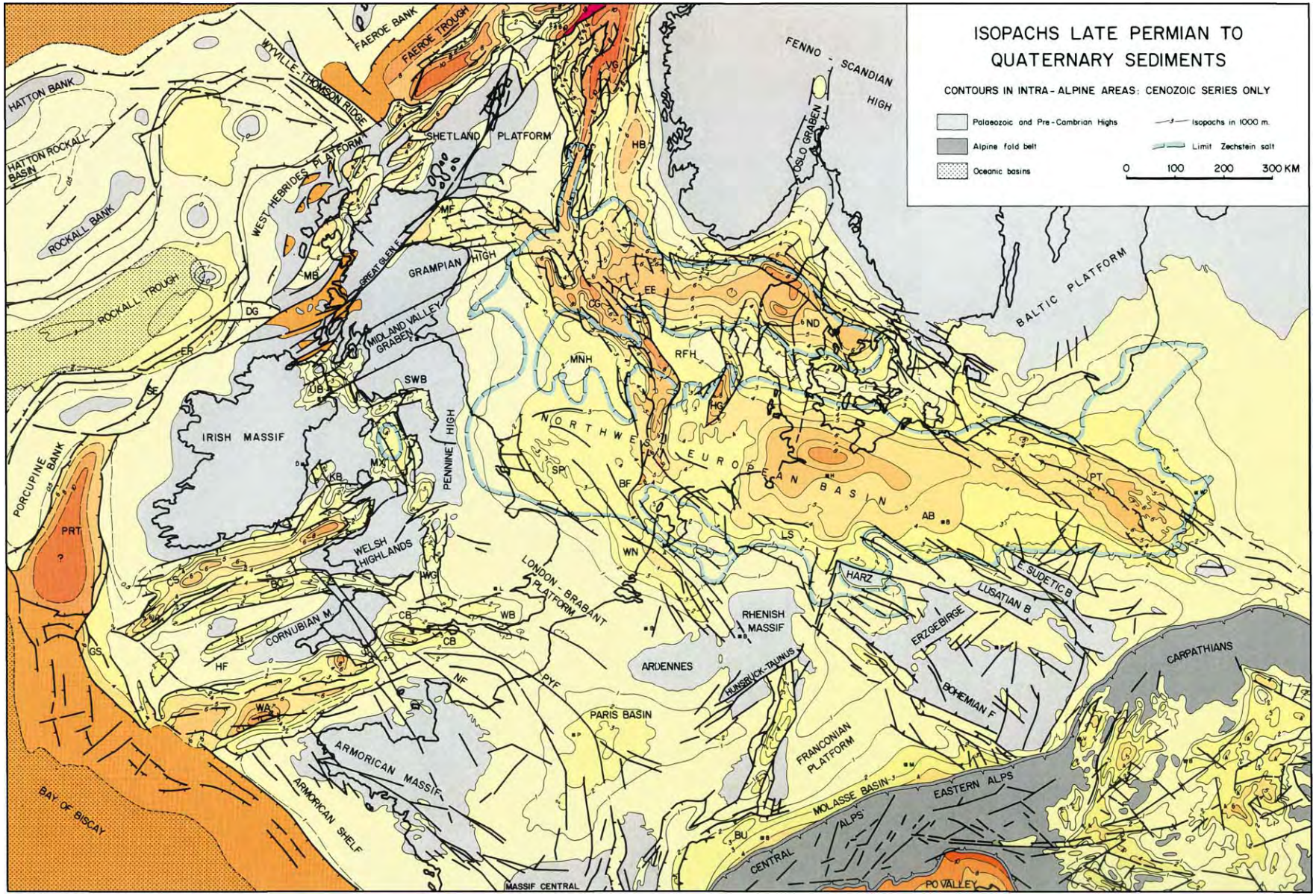


Figure 1 Total isopach of Late Permian to Cenozoic sediments.

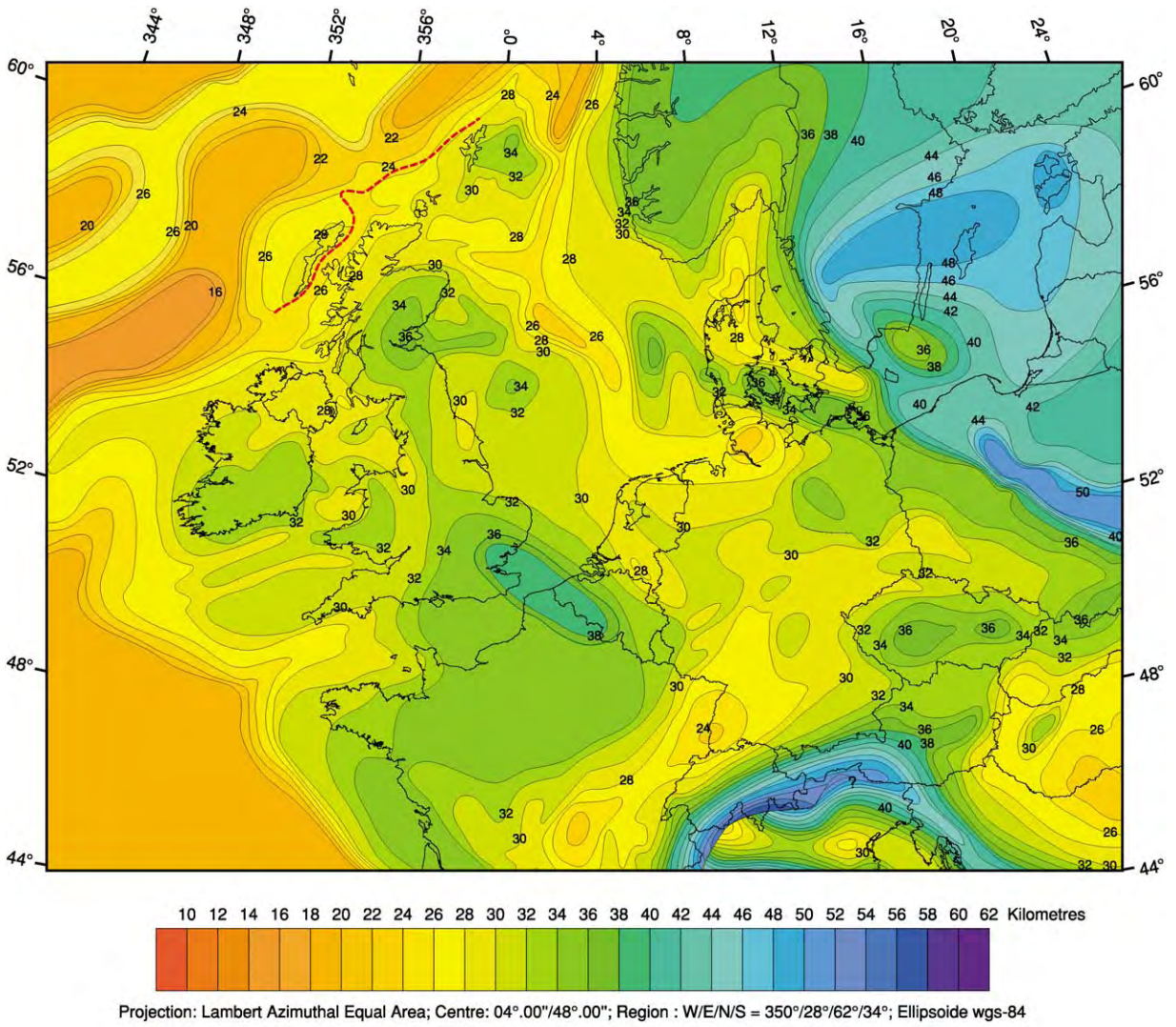


Figure 2 Moho Depth map of Western and Central Europe, contour interval 2 km (after Dèzes & Ziegler, 2004) (available in digital format).

that peaked during the Early Permian, and the subsidence of a multi-directional array of transtensional and pull-apart basins in which continental clastics accumulated. Basins that developed during this time span show a complex, polyphase structural evolution, including a late phase of transpressional deformation controlling their partial inversion. Although Stephanian to Early Permian wrench deformation locally gave rise to the uplift of extensional core complexes, crustal stretching factors were on a regional scale relatively low. Nevertheless, the model of the Cenozoic Basin-and-Range Province has been repeatedly invoked for the Stephanian-Early Permian collapse of the Variscan Orogen. Yet, kinematics controlling the development of these two provinces differed: whereas the collapse of the Variscides was wrench-dominated, extension dominated the collapse of the Cordillera.

The Stephanian to Early Permian magmatic activity can be related to wrench-induced reactivation of Variscan sutures which caused the detachment of subducted lithospheric slabs, upwelling of the asthenosphere, partial delamination and thermal thinning of the mantle-lithosphere, magmatic inflation of the remnant lithosphere and interaction of mantle-derived partial melts with the lower crust. In conjunction with slab detachment and a general reorganization of the asthenospheric flow patterns, a system of not very active mantle plumes apparently welled up to the base of the lithosphere in the area of the future Southern and Northern Permian basins, the British Isles and the Oslo Graben, causing thermal attenuation of the lithosphere and magmatic destabilisation of the crust-mantle boundary. In the domain of the Variscan Orogen, this accounted for the

destruction of its orogenic roots and regional uplift. By the end of the Early Permian, its crust was thinned down on a regional scale to 30–35 km, mainly by magmatic processes and erosional unroofing and only locally by its mechanical stretching. Moreover, the thickness of the mantle-lithosphere was reduced to as little as 50 to 10 km in areas which evolved during the Late Permian and Mesozoic into intracratonic thermal sag basins, such as the Southern Permian and the Paris basins.

Late Permian–Early Cretaceous Rifting During Pangaea Breakup

Following the Stephanian–Early Permian tectono-magmatic cycle, the potential temperature of the asthenosphere returned quickly to ambient levels (1300°C). This is indicated by the late Early Permian extinction of magmatic activity and the onset of regional thermal subsidence of the lithosphere, reflecting the decay of thermal anomalies that had been introduced during the Permo-Carboniferous. In combination with erosional degradation of the remnant topography of the Variscan Orogen and cyclically rising sea-levels, progressively larger areas subsided below the erosional base level and were incorporated into a new system of intracratonic basins, comprising the Northern and Southern Permian basins, the Hessian Depression, the Paris Basin, and the Franconian Platform (Figure 3). However, in large parts of the WCE, thermal subsidence of the lithosphere was overprinted and partly interrupted by the Late Permian–Early Triassic onset of a rifting cycle which preceded and accompanied the step-wise break-up of Pangaea. Major elements of this breakup system were the southward propagating Arctic-North Atlantic and the westward propagating Neotethys rift systems. Evolution of these mega-rift systems was paralleled by the development of multi-directional grabens in the WCE, major constituents of which are the North Sea rift, the North Danish-Polish Trough, the graben systems of the Atlantic shelves and the Bay of Biscay rift (Figure 3). Development of these grabens partly involved tensional reactivation of Permo-Carboniferous fracture systems.

Late Permian

Thermal subsidence of the Northern and Southern Permian Basin commenced during the late Early Permian and persisted into Early Jurassic times, as evidenced by quantitative subsidence analyses, facies patterns and isopach maps. In these basins sedimentation commenced during the late Early Permian with the accumulation of the continental Rotliegend red-bed series which attain a thickness of up to 2300 m in the axial parts of the-well defined Southern Permian Basin and of 600 m in the less well-defined Northern Permian Basin that was severely overprinted by the Mesozoic North Sea rift.

During the Late Permian, the Norwegian–Greenland Sea rift, which had come into evidence during the Late Carboniferous, propagated southward into the North-western Shelf of the British Isles, opening a seaway through which the Arctic seas transgressed via the Irish Sea and possibly the northernmost North Sea into the Northern and Southern Permian basins which, by this time, had subsided below the global sea-level (*see* **Europe:** Permian Basins). In these basins, the cyclical Zechstein carbonate, evaporite, and halite series were deposited under a tectonically quiescent regime. Following the initial transgression, the axial parts of the Northern and Southern Permian basins were characterized by deeper water conditions whereas along their margins basin-ward prograding carbonate and evaporitic shelves developed. Facies and thickness changes on these shelf series provide evidence for minor extensional faulting along the southern margin of the Southern Permian Basin. Oscillating sea-levels accounted for cyclical restriction of the Northern and Southern Permian basins in which up to 2000 m of halites, partly containing polyhalites, were deposited. During the end-Permian, global low-stand in sea-level, the Arctic seas withdrew from the WCE into the area between Norway and Greenland.

Triassic

During the Triassic, the Norwegian–Greenland Sea rift propagated southwards into the North and Central Atlantic domain, whilst the Neotethys rift systems propagated westwards through the Bay of Biscay and North-west Africa and linked up with the Atlantic rift system.

During the Early Triassic, the North Sea rift, consisting of the Horda half graben, and the Viking, Murray Firth, Central, and Horn grabens, was activated and transected the western parts of the Northern and Southern Permian basins, whereas their eastern parts were transected by the North Danish–Polish Trough (Figure 4). Simultaneously the rift systems of the Alpine domain, the Bay of Biscay, and the Western Shelves were activated. The latter included the Porcupine, Celtic Sea, and Western Approaches troughs. Crustal extension in the Celtic Sea and Western Approaches troughs involved at their eastern termination the reactivation of Permo-Carboniferous shear systems controlling the subsidence of the Channel Basin and intermittent destabilization of the Paris thermal sag basin. Significantly, Triassic–Early Jurassic rifting was

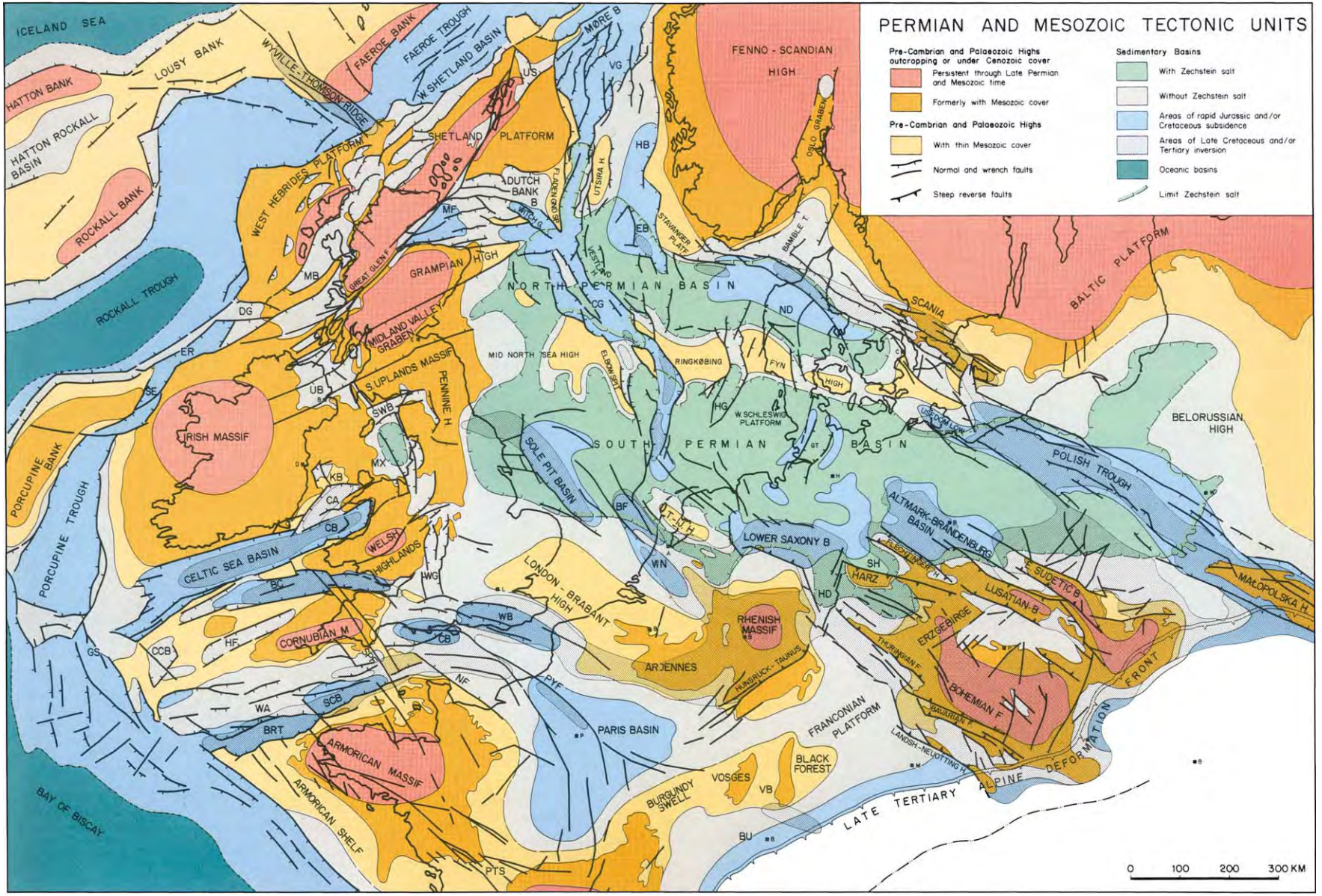


Figure 3 Permian and Mesozoic Tectonic units. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

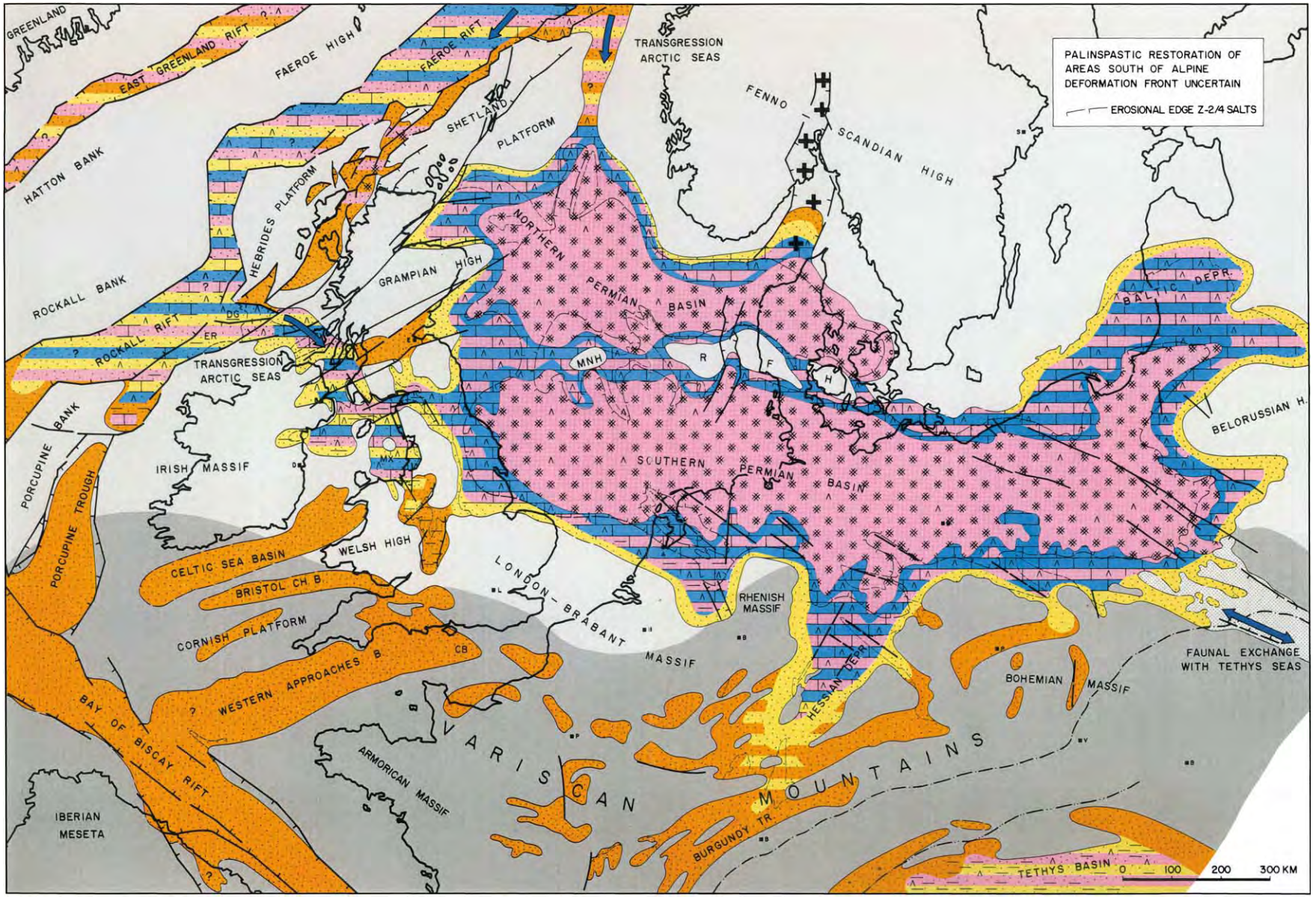


Figure 4 Late Permian, Zechstein palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

accompanied by a low level of volcanic activity with the exception of the Bay of Biscay, in the evolution of which wrench faulting compensating for crustal extension in the Atlantic domain played an important role. Areas not affected by rifting continued to subside during the Triassic and Early Jurassic in response to thermal relaxation of the lithosphere, accounting in conjunction with eustatically rising sea-levels for a progressive overstepping of basin margins.

During the Early Triassic, continental to lacustrine conditions prevailed in the evolving grabens and thermal sag basins of WCE, in which the 'Bunter' red-beds were deposited. Clastics were shed into these basins from adjacent Variscan and Caledonian highs, as well as from Fennoscandia. During the late Early Triassic, the Tethys Seas ingressed the continuously subsiding Southern Permian Basin via the Polish Trough, giving rise to the deposition of carbonates in Poland and distal halites in northern Germany and the Southern North Sea (Figure 5).

As during the Middle Triassic, the trend of highs that had separated the Northern and Southern Permian basins became gradually overstepped and with this these basins coalesced, thus forming the composite North-west European Basin. The Tethys Seas advanced further into this continuously subsiding basin complex via the Polish Trough as well as via the Burgundy Trough and the Trier and Hessian depressions, establishing a broad neritic basin in which the 'Muschelkalk' carbonates, evaporites, and halites were deposited (Figure 6). Furthermore, intermittent marine transgressions advanced from the Tethyan shelves via the Bay of Biscay rift into the grabens of the Western Shelves. By contrast, continental conditions continued to prevail in the grabens of the Central and Northern North Sea and the North-western Shelf.

With the beginning of the Late Triassic (*see* **Mesozoic:** Triassic), clastic influx from Fennoscandia and eastern sources increased, causing the replacement of the carbonate-dominated Muschelkalk depositional regime by the evaporitic 'Keuper' red-beds containing halites. Whilst the Polish seaway, which had linked the Tethys and the North-west European Basin, was closed, intermittent marine transgressions advanced through the Burgundy Trough into the evolving Paris Basin and the continuously subsiding North-west European Basin, as well as through the Bay of Biscay rift into the grabens of the Western Shelves. However, continental conditions persisted in the grabens of the Central and Northern North Sea and the North-western Shelf. Only during the Rhaetian did the Arctic Seas start to advance southwards into the rifted basins of the North-western Shelf, whilst neritic conditions were established in the broad North-west European Basin (Figure 7).

The Triassic series attains thicknesses of up to 3 km in the grabens of the Western Shelves, the North Sea, and in the Polish Trough, and up to 6 km in the grabens of the North-western Shelf. In the Northern and Southern Permian Basins, the diapirism of Permian salts commenced during the Triassic, and accounted for local subsidence anomalies.

Jurassic

In conjunction with continued rifting activity and cyclically rising sea-levels, the Arctic and Tethys Seas linked up during the Rhaetian–Hettangian, via the rift systems of the North-western and Western shelves and the continuously subsiding North-west European Basin (*see* **Mesozoic:** Jurassic). In the open marine, shale-dominated North-west European Basin, which occupied much of the Southern and Central North Sea, Denmark and Germany (Figure 7), the Belemnitidae (*see* **Fossil Invertebrates:** Cephalopods (Other Than Ammonites)) developed during the Hettangian and Sinemurian. Persisting clastic influx from the East-European Platform allowed only for temporary marine incursions via the Polish Trough. Similarly, fluvio-deltaic conditions prevailed in the grabens of the Northern North Sea until the end-Hettangian to Early Sinemurian when neritic conditions were also established in these basins. By Late Simemurian times, this facilitated a broad faunal exchange between the Boreal and Tethyan realms and the dispersal of the Belemnitidae.

In response to rising sea-levels and continued crustal extension, open marine conditions were established in the Central Atlantic during the Simemurian, permitting Tethyan faunas to reach the Pacific by Pliensbachian times. In basins which were dominated by the warmer Atlantic and Tethyan waters, carbonates and shales were deposited, whilst shales prevailed in the North-west European Basin, which was dominated by the cooler Arctic waters. During the Early Jurassic, repeated stagnant water stratification gave rise to the deposition of organic-rich shales, forming important oil source-rocks, for example, in the Paris Basin and the southern parts of the North-west European Basin.

During the Late Aalenian–Early Bajocian, the Arctic seas became separated from the Tethys and the Central Atlantic in conjunction with the uplift of a large arch in the Central North Sea from which clastics were shed into the adjacent continuously subsiding basins (Figure 8). Uplift of this arch was associated with major volcanism that may be related to the impingement of a short-lived mantle plume. Open marine communications between the Arctic and the Tethys–Atlantic seas were re-opened

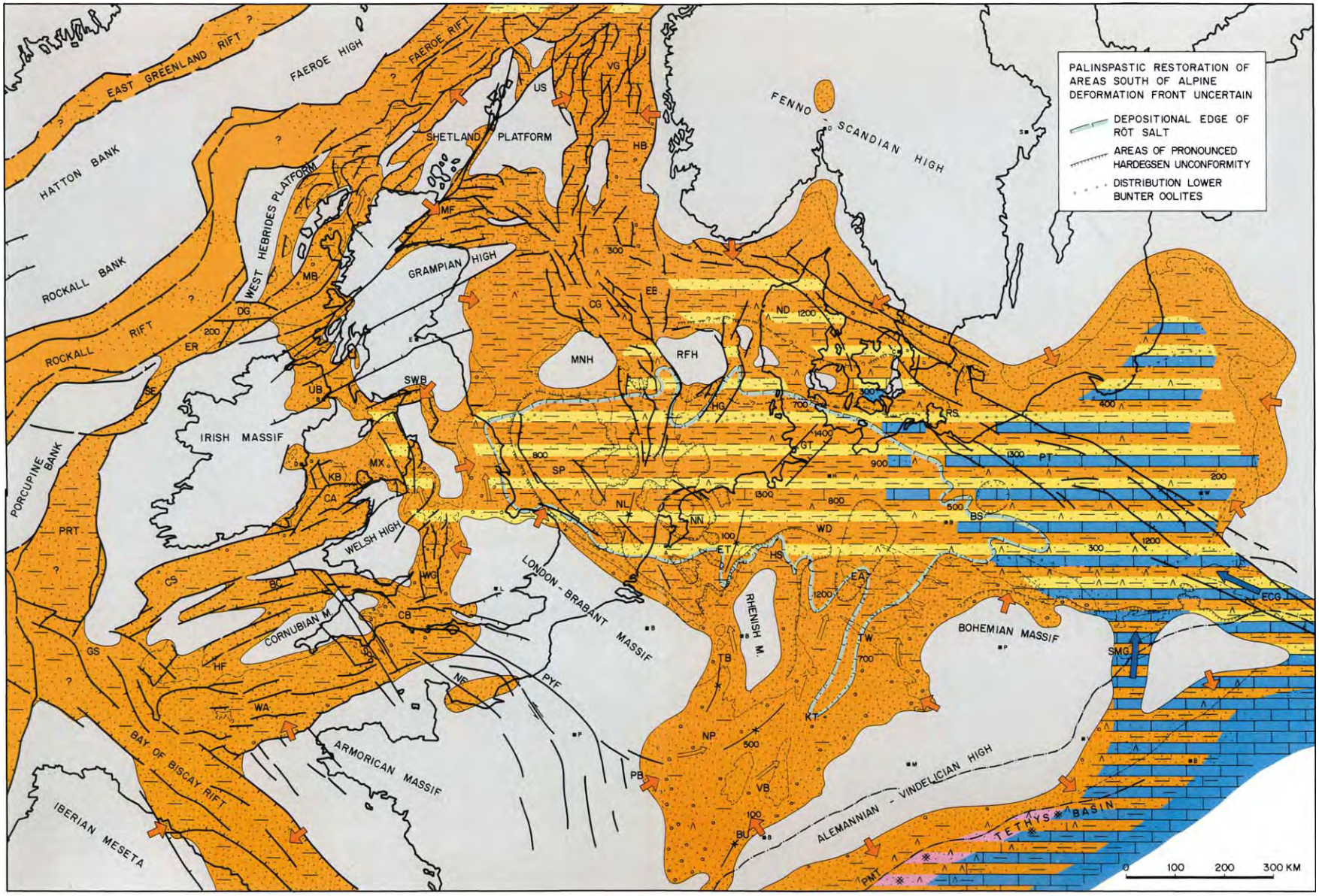


Figure 5 Scythian, Buntsandstein palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

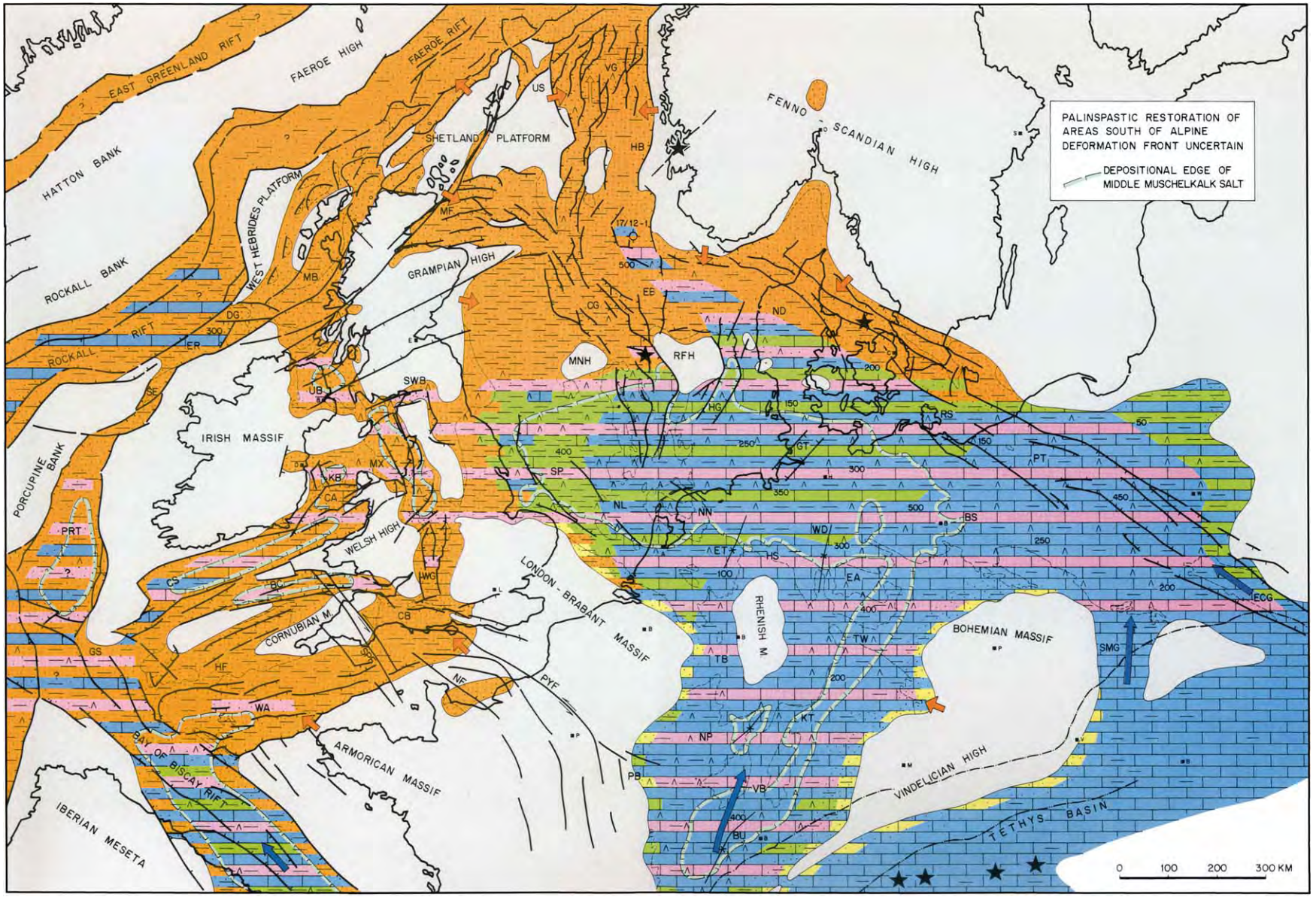


Figure 6 Anisian Ladnian, Muschelkalk palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

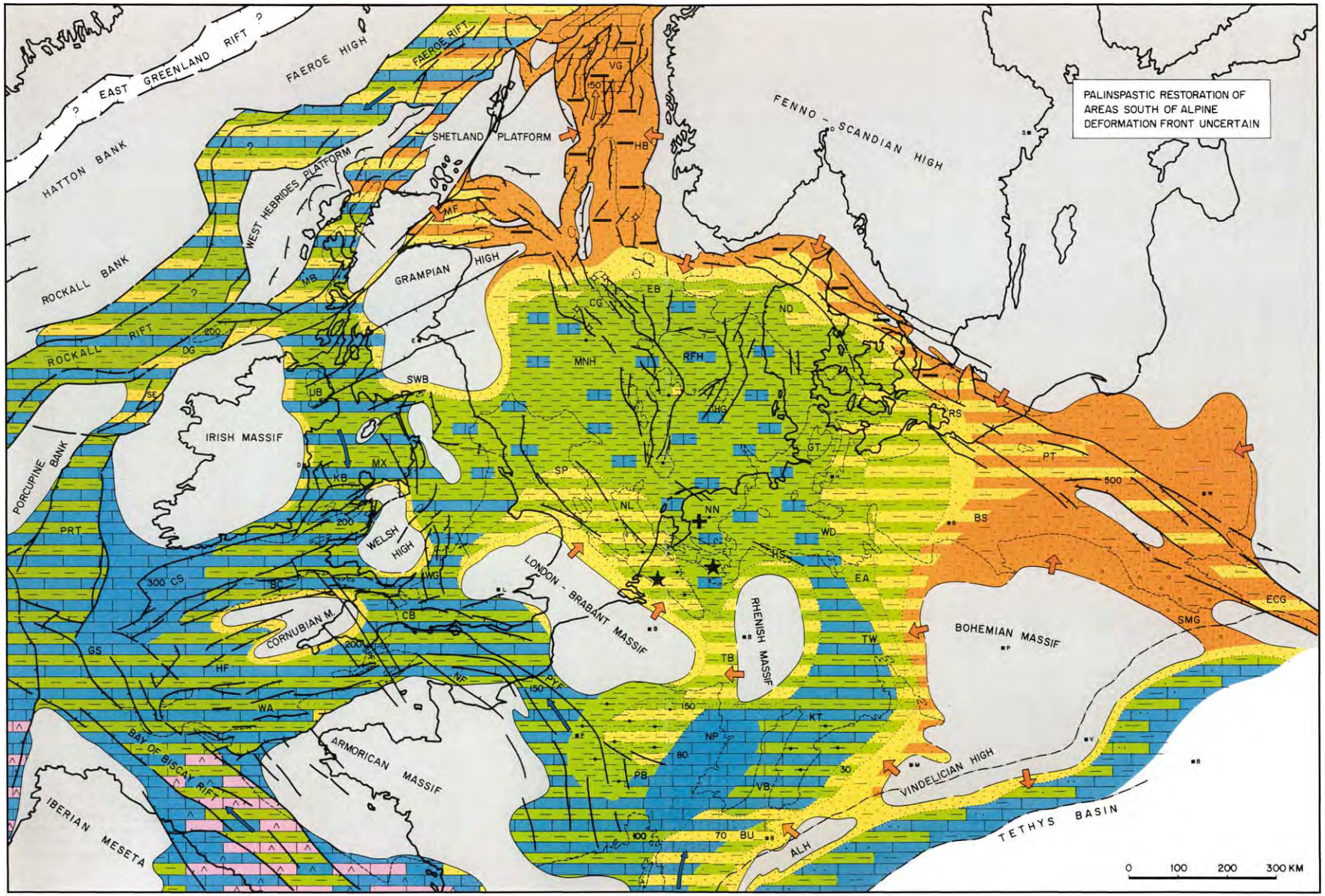


Figure 7 Rhaetian Hettangian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

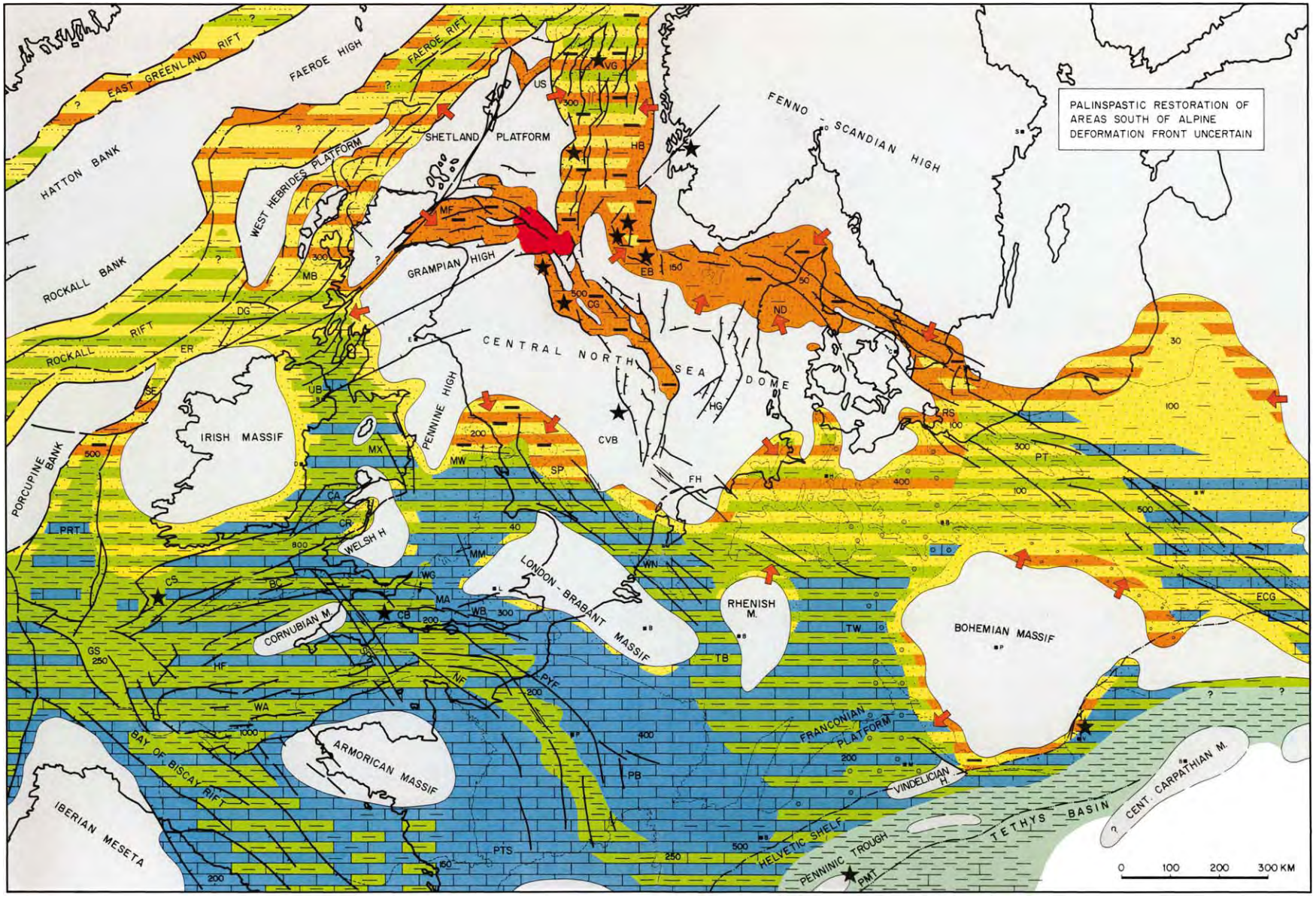


Figure 8 Bajocian Bathonian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

during the Callovian, due to the subsidence of the Central North Sea arch and continued crustal extension in the North Sea and the Faeroe-Rockall rifts. During the Middle Jurassic, carbonates prevailed in basins that were dominated by the Tethyan Seas, whereas clastics prevailed in basins dominated by Arctic waters.

Toarcian crustal separation in the Central Atlantic and Bajocian crustal separation in the Alpine Tethys caused a change in the stress systems that controlled the evolution of the WCE rifts. From then on their evolution was dominated by stresses controlling the northward propagation of the Atlantic rift system. During the Late Jurassic and earliest Cretaceous, accelerated crustal extension in the North Sea rift, now focused on its axial Viking and Central grabens, was accompanied by the development of sinistral shear systems at its southern termination which controlled the subsidence of the transtensional Sole Pit, Broad Fourteens, West Netherlands, Lower Saxony, and Sub-Hercynian basins, and the opening of a seaway cross-cutting the Bohemian Massif (**Figures 3 and 9**). This was accompanied by the gradual uplift of a high that extended from the London-Brabant Massif to the Bohemian Massif, partly separating the carbonate shelves of the Paris Basin and Franconian Platform from the shale and clastic dominated basins of the North Sea area. Contemporaneous accelerated rifting activity is also evident in the Western Approaches, Celtic Sea, and Porcupine troughs that opened into the Bay of Biscay and North Atlantic rifts. In the latter, crustal separation between Iberia and the Grand Banks of Newfoundland progressed gradually northwards during the Late Jurassic and Early Cretaceous. The Late Jurassic cycle of accelerated rifting and wrench tectonics was associated with only minor volcanic activity.

In the course of the Late Jurassic, deeper water conditions were established in the North Sea, Rockall-Faeroe, Porcupine, and Bay of Biscay rifts. Organic-rich Kimmeridgian–Tithonian shales deposited in these rifts and in flanking shallower water basins form the principal oil source rocks of the Central and Northern North Sea, Faeroe rift, and Channel Basin hydrocarbon provinces.

Early Cretaceous

During the earliest Cretaceous, large parts of the WCE were uplifted and subjected to erosion in response to wrench faulting and stress-induced buckling of the lithosphere, causing a relative low-stand in sea-level. Deeper marine conditions persisted; however, in the continuously subsiding Bay of Biscay, Porcupine, Rockall-Faeroe, and the North Sea rifts. On the other hand, deltaic to continental clastics were deposited in the rapidly subsiding transtensional Channel, Weald, Sole Pit, West Netherlands, Lower Saxony, and Altmark-Brandenburg basins, as well as in the Celtic Sea and Western Approaches grabens. However, the Tethys shelves, of which the Paris Basin formed an embayment, and the slowly subsiding North Danish-Polish trough were affected to a lesser extent by this tectonically-induced relative sea-level change (**Figure 10**).

From Neocomian to Albian times, basin margins were progressively overstepped in response to eustatically rising sea-levels and gradually abating tectonic activity in the rift- and wrench-induced basins of the WCE, whilst crustal extension progressively focused on the zone of future crustal separation between Europe and Greenland. With the onset of post-rift thermal subsidence, the North Sea Basin gradually expanded (**Figure 11**). Nevertheless, minor tectonic activity was still evident until Aptian times in the North Sea, the Sole Pit, Broad Fourteens, West Netherlands, Lower Saxony, and Sub-Hercynian basins, as well as along the Fennoscandian Border Zone and the border faults of the Bohemian Massif. Following Aptian crustal separation in the Bay of Biscay, the grabens on the Western Shelves became inactive and began to subside thermally. During the Early Cretaceous volcanic activity in the WCE was rather limited.

During the Neocomian, the North Sea Basin was partly separated from the Tethys by the wrench-induced Brabant-Rhenish-Bohemian high, and from the Atlantic by a high that extended from Ireland via Scotland to the Shetland Isles. However, in the course of the Aptian and Albian, shallow marine communications were reopened between the North Sea Basin and the Tethys and Western shelves via southern England and the Paris Basin. Throughout the Early Cretaceous, the WCE basin system was clastic-dominated except for the Armorican margin on which carbonates were deposited.

Late Cretaceous–Paleocene Rifting and Early Alpine Intraplate Compression

During the Late Cretaceous and Paleocene, rifting activity was centred on the Faeroe-Rockall Trough and the area between the Rockall-Hatton Bank and Greenland. Limited sea-floor spreading may have occurred in the southern parts of the Rockall Trough during the Cenomanian–Santonian. During the Campanian–Maastrichtian, the long-lived Iceland plume impinged on the Arctic-North Atlantic rift system,

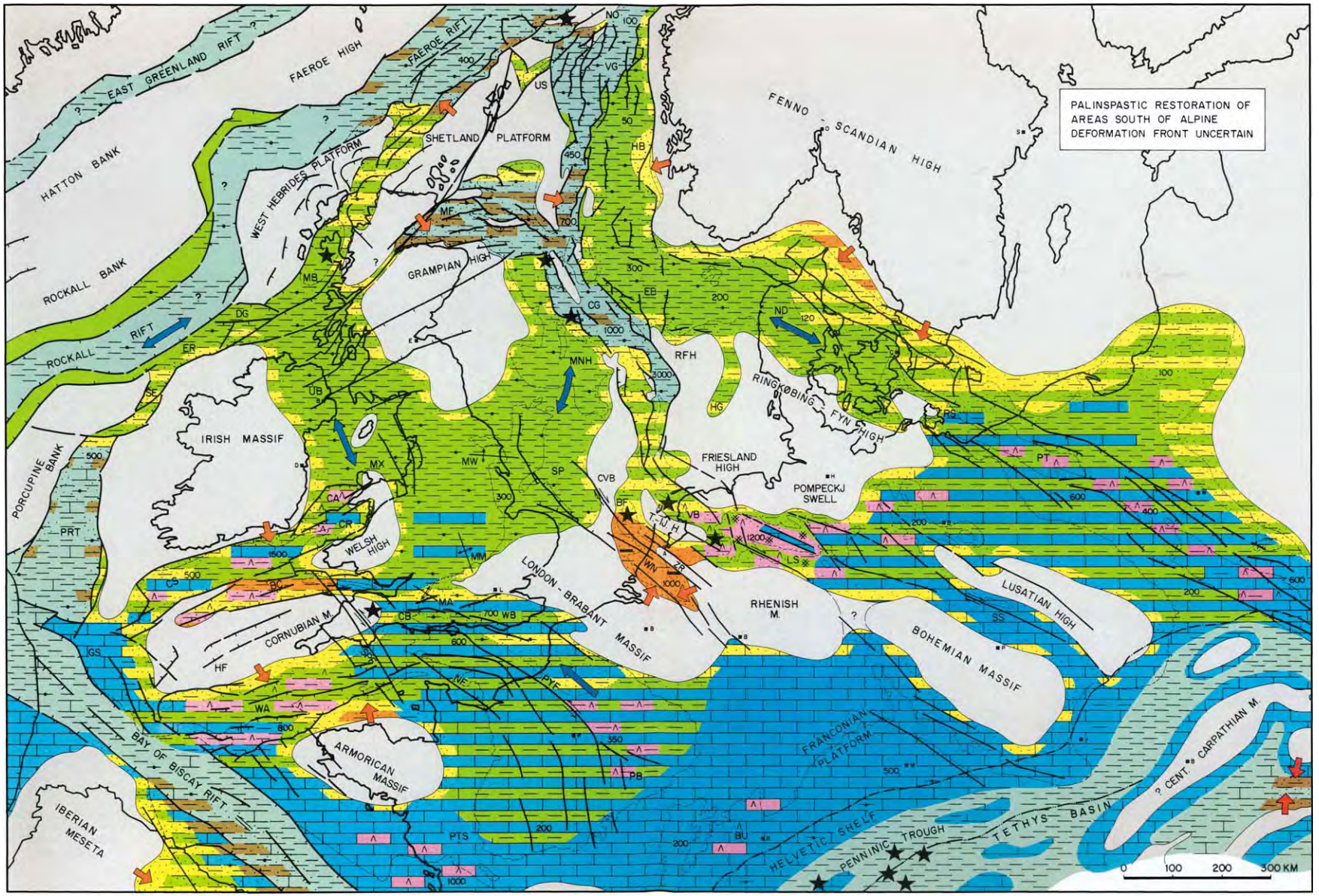


Figure 9 Kimmeridgian/Tithonian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

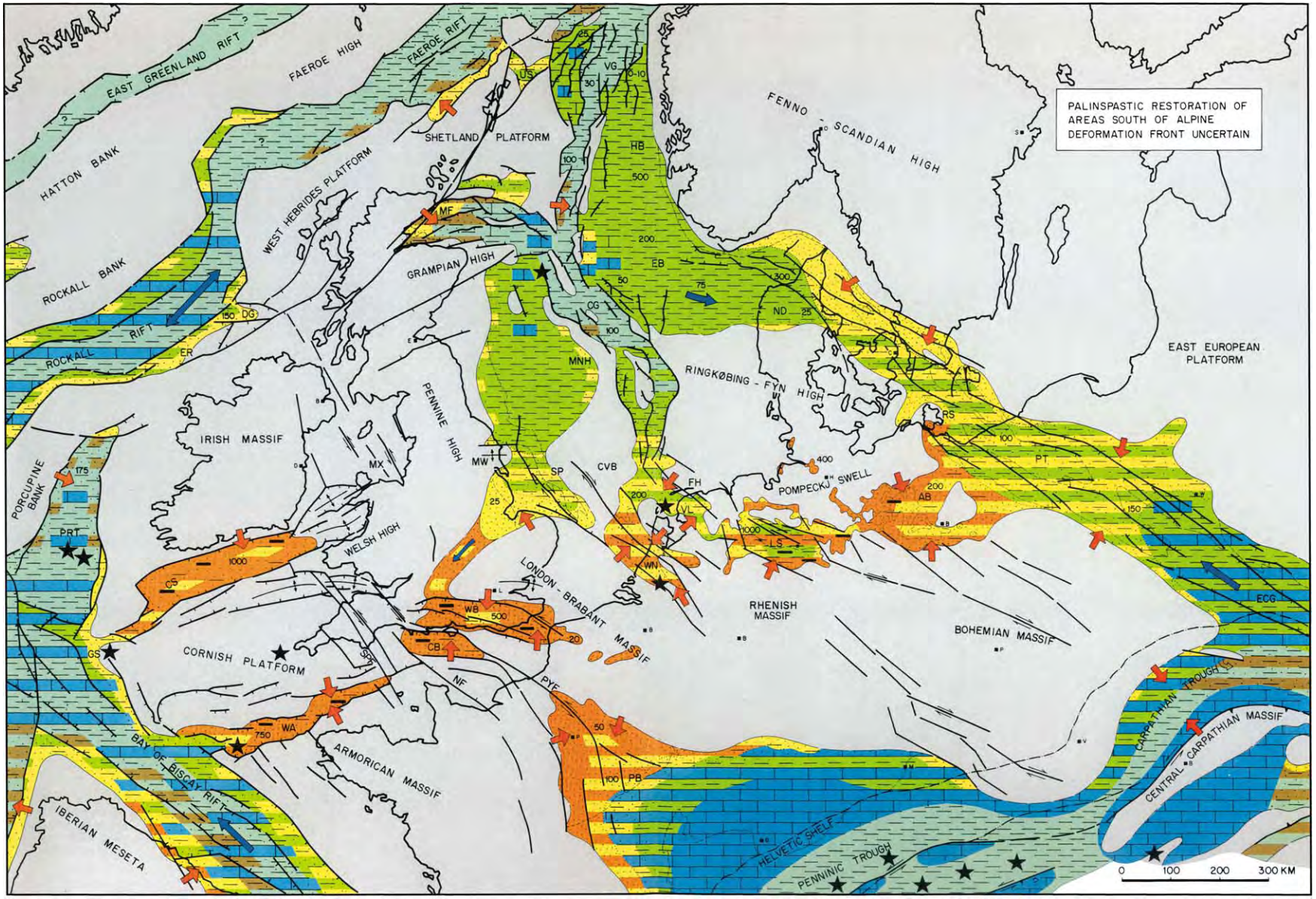


Figure 10 Berriasian Valanginian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

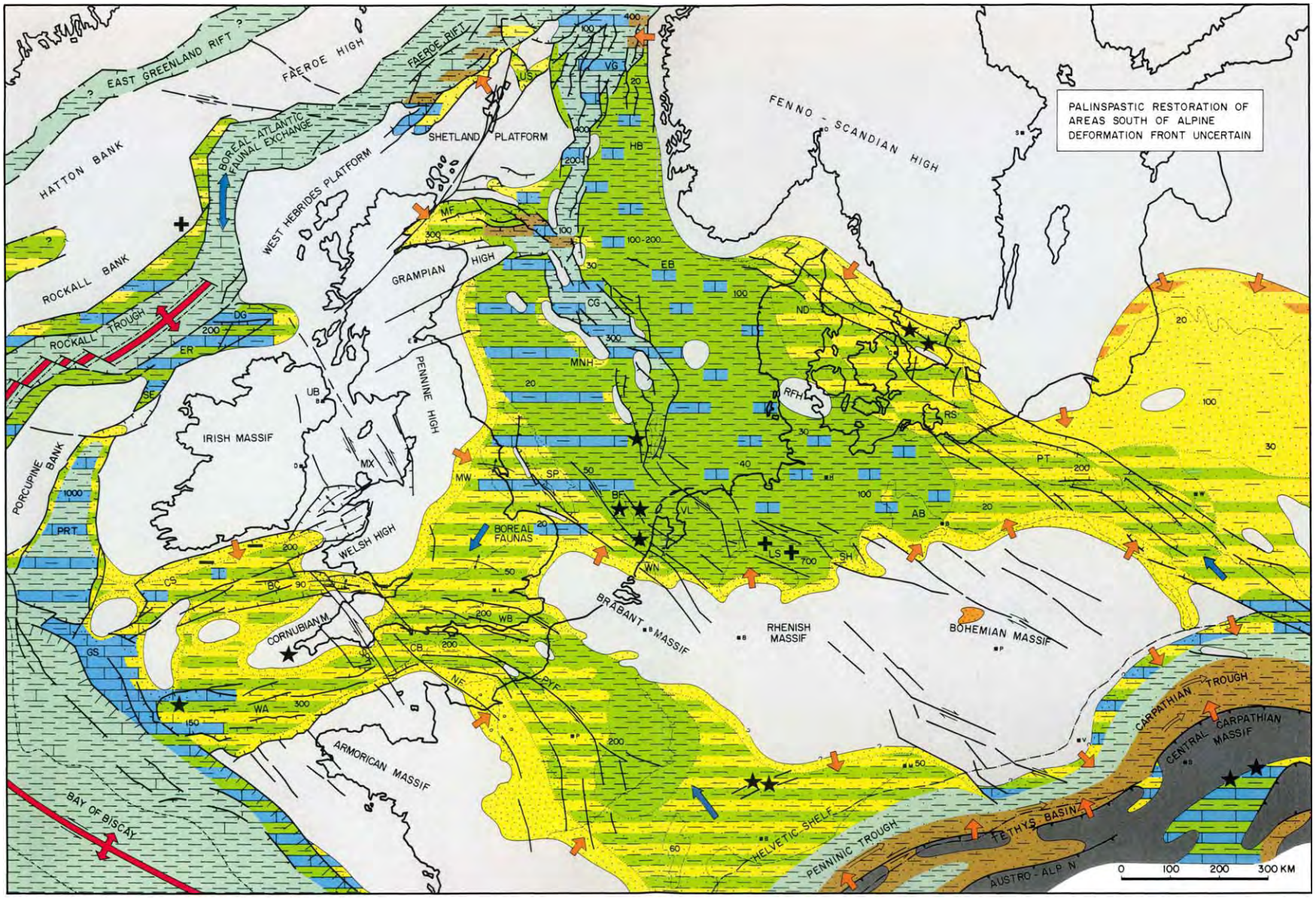


Figure 11 Aptian Albian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

giving rise to the development of the large Thulean flood basalt province that had, during the Paleocene, a radius of more than 1000 km. At the Paleocene–Eocene transition, crustal separation was achieved between Greenland and Europe to the west of the Rockall-Hatton Bank and in the Norwegian-Greenland Sea, paving the way for opening of the Arctic-North Atlantic Ocean.

During the Turonian–Santonian, Africa began to converge with Europe in a counter-clockwise rotational mode. Ensuing space constraints within the Tethyan belt were compensated by the activation of new subduction zones and the onset of closure of the Alpine Tethys and the Bay of Biscay.

Late Cretaceous

Late Cretaceous regional thermal subsidence of the North Sea Basin, the Polish Trough, the Paris Basin, and the Tethys shelves, as well as of the Western Shelves, combined with cyclically rising sea-levels, accounted for broad overstepping of the Mid-Cretaceous basin margins. This caused a drastic reduction of clastic supply to these basins in which sedimentation became carbonate-dominated from Cenomanian times onward. By end-Cretaceous times, much of the WCE was covered by the 'Chalk' sea, water depths of which ranged from neritic to bathyal, as evident in the axial parts of the North Sea Basin (Figure 12).

However, commencing in Late Turonian times, compressional stresses were exerted on the southern margin of the WCE, causing inversion of Mesozoic tensional basins and upthrusting of basement blocks by reactivation of pre-existing crustal discontinuities (Figure 3). The intra-Senonian pulse of intraplate compression, which affected the Polish Trough, the Fennoscandian Border Zone, the Bohemian Massive, the Sub-Hercynian, Lower Saxony, West Netherlands, and the Sole Pit Basins, as well as the southern parts of the North Sea rift, can be related to compressional stresses which were projected from the Alpine-Carpathian orogenic wedge through the oceanic lithosphere of the Alpine Tethys into the lithosphere of WCE. From the uplifted basement blocks of the Bohemian Massif and the Polish inversion axis, clastics began to be shed into the surrounding Chalk seas (Figure 12).

Paleocene

The second, more intense, Paleocene phase of intraplate compression, which affected about the same areas, and to a lesser degree also the Tethys shelves of the Western and Central Alps, the Paris Basin and the Channel area, probably marks the collision of the Alpine-Carpathian orogenic wedge with its East Alpine-Carpathian foreland and with the Briançonnais Terrane in the West and Central Alpine domain (*see* **Europe:** The Alps). The most distal intraplate compressional structures occurred at a distance of some 1500 km to the north-west of the contemporary collision front. The Paleocene phase of foreland compression, during which a Rocky Mountain-type array of basement blocks was upthrusted in the Bohemian Massif and the Polish Trough was deeply inverted, involved also broad lithospheric folding, causing a regional regression, the disruption of the Late Cretaceous carbonate platforms, and accelerated subsidence of the North Sea Basin. Combined with plume-related thermal doming of the British Isles during the development of the Thulean flood basalt province, sedimentation in the North Sea Basin changed at the end of the Danian from carbonates to clastics. At the same time, this basin became isolated from the Tethys and Atlantic seas but remained open to the Norwegian-Greenland Sea (Figure 13).

On the other hand, mixed carbonate-clastic environments persisted on the Western Shelves, that were not affected by this pulse of intraplate compression. However, Paleocene clastic influx into the Paris Basin from southern sources is probably related to the uplift of the Armorican-Massif Central arch in response to compressional stresses that were exerted on the shelves of Southern France during the early phases of the Pyrenean orogeny that had commenced during the Senonian. This is compatible with the Paleocene-Eocene development of the Languedoc-Provençal fold-and-thrust belt of Southern France, that involved inversion of Mesozoic extensional basins.

Paleocene intrusion of mafic dykes in the Massif Central, Vosges, Black Forest, and the Bohemian Massif, which must be related to partial melting of the lithospheric thermal boundary layer, probably reflected a mantle plume-related increase in the potential temperature of the asthenosphere, resulting in a renewed destabilization of the lithosphere. On the other hand, Paleocene compressional inversion of the Polish Trough and upthrusting of basement blocks in the Bohemian Massif apparently caused thickening of the crust (Figure 2).

Opening of the Arctic–North Atlantic and Collisional Interaction of the Alpine Orogen with its Foreland

With the Early Eocene onset of sea-floor spreading in the Arctic-North Atlantic, volcanic activity ceased on the shelves of the British Isles and the Rockall-Hatton-Faeroe Bank. With this, their post-rift thermal subsidence commenced, whilst thermal subsidence

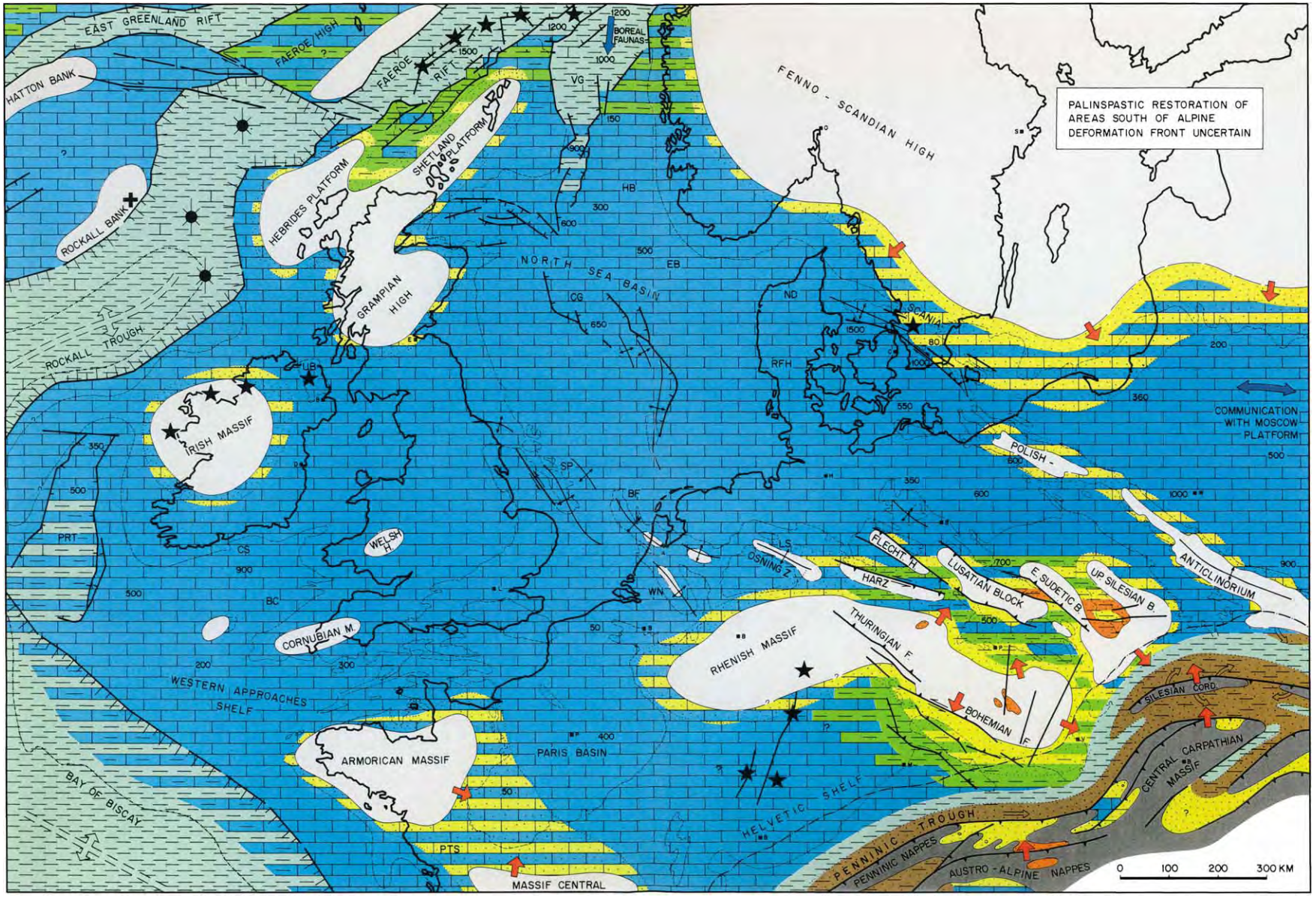


Figure 12 Senonian Danian palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

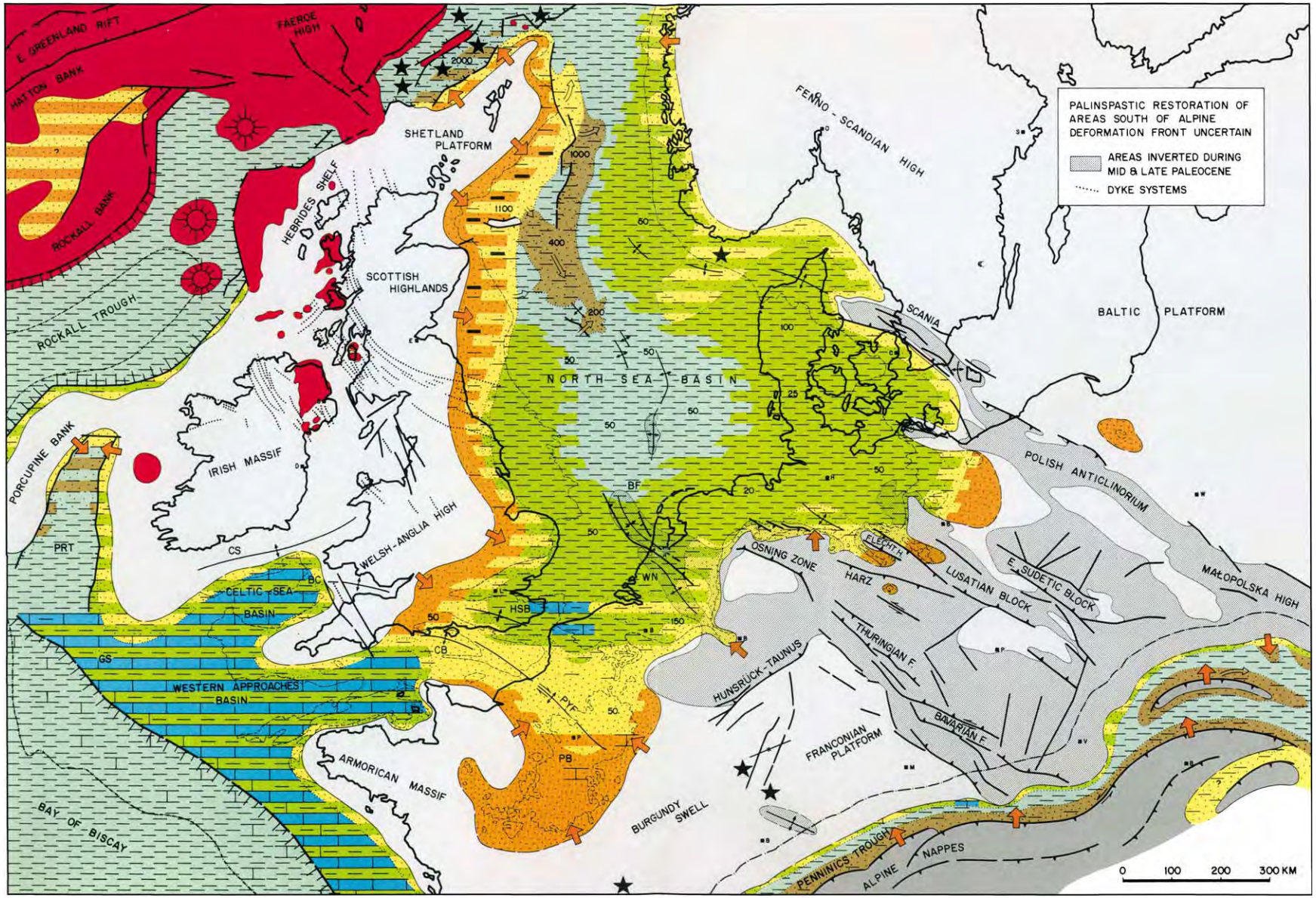


Figure 13 Late Paleocene palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

of the North Sea Basin and the Western Shelves continued. During the Late Eocene and Oligocene, reorganization of sea-floor spreading axes in the Norwegian-Greenland Sea, the shelves of the British Isles were destabilized by minor wrench faulting in the prolongation of the Iceland ridge and the Charlie Gibbs fracture zones, causing the subsidence of small transtensional basins in the Irish Sea area (**Figure 14**). Moreover, repeated pulses of basin inversion interfered with the thermal subsidence of the Celtic Sea, Western Approaches, and Channel basins (**Figure 3**).

During the Eocene, thrust-loaded flexural subsidence of the foreland of the Western, Central, and Eastern Alps, and also the Carpathian foreland, commenced. Oligocene to Miocene emplacement of the East-Alpine and Carpathian nappe systems was, however, not accompanied by further intraplate compressional deformation of their forelands, thus reflecting mechanical decoupling of these orogens from their forelands. By contrast, Late Eocene–Early Oligocene and Late Oligocene–Early Miocene inversion pulses evident in the Celtic Sea, Western Approaches, Channel, Weald, Sole Pit, Broad Fourteens, and West Netherlands basins testify to intermittent and increasing mechanical coupling of the evolving West and Central Alpine Orogen with its foreland (**Figures 3 and 14**). Crustal shortening in the Western and Central Alps persisted during the Late Miocene and Pliocene, as evident by folding of the Jura Mountains, and may indeed still be going on, as indicated by earthquake activity and geodetic data.

In the Alpine foreland, development of the tectonically still active European Cainozoic rift system (ECRIS) commenced during the Late Eocene. Today this rift system extends over a distance of more than 1000 km from the Dutch North Sea coast to the Mediterranean. Its southern elements are the northerly-striking Limagne and the Valence and Bresse grabens, which are superimposed on and flank the Massif Central, respectively. These grabens are linked via the Burgundy transfer zone to the northerly-striking Upper Rhine Graben which bifurcates northwards into the north-west-trending Roer Graben and the north-easterly trending Hessian grabens that transect the Rhenish Massif. The north-east-striking Eger Graben, which transects the Bohemian Massif, forms an integral part of the ECRIS (**Figure 14**). Localization of ECRIS involved the reactivation of Permo-Carboniferous shear systems. Although characterized by relatively low crustal stretching factors, the evolution of the ECRIS was accompanied by the development of major volcanic centres on the Massif Central, the Rhenish Massif and the Bohemian Massif, particularly during Miocene and Plio-Pleistocene times. Seismic tomography indicates that mantle plumes well up beneath the Massif Central and the Rhenish Massif but not beneath the Vosges-Black Forest arch; similar data are, however, not available for the Bohemian Massif. Despite this, the evolution of the ECRIS is considered to be a clear case of passive rifting.

During the Late Eocene, the Valence, Limagne, Bresse, Upper Rhine, and Hessian grabens began to subside in response to northerly-directed compressional stresses that can be related to the collisional interaction of the Pyrenees and the Alps with their forelands. These originally-separated rifted basins coalesced during their Oligocene main extensional phase, and the Roer and Eger Grabens started. During the Late Oligocene, rifting propagated southward across the Pyrenean Orogen into the Gulf of Lions and along coastal Spain in response to back-arc extension, that was controlled by eastward roll-back of the subducted Betic-Balearic slab. By Late Burdigalian times, crustal separation was achieved, the oceanic Provençal Basin began to open, and the grabens of southern France became inactive. By contrast, the intra-continental parts of the ECRIS remained tectonically active until the present, although their subsidence has been repeatedly interrupted, possibly in conjunction with stresses controlling far-field inversion tectonics. By end-Oligocene times, magmatic activity increased on the Rhenish Shield. At the same time, the area of the triple junction between the Upper Rhine, Roer, and Hessian grabens became uplifted, presumably in response to thermal thinning of the lithosphere, interrupting the Oligocene sea-way which had linked the North Sea Basin with the Alpine foreland basin. By Middle–Late Miocene times, the Massif Central, the Vosges-Black Forest arch, and slightly later, also the Bohemian Massif, were uplifted. This was accompanied by increased mantle-derived volcanic activity. At the level of the Moho, a broad anticlinal feature extends from the Massif Central via the Burgundy Transfer zone, the Vosges-Black Forest into the Bohemian Massif (**Figure 2**). Uplift of these arches probably involved folding of the lithosphere in response to increased collisional coupling of the Alpine Orogen with its foreland. Uplift of the Burgundy transfer zone entailed partial erosional isolation of the Paris Basin. Under the present north-west-directed stress regime, which had developed during the Miocene and intensified during the Pliocene and reflects a combination of Alpine collisional and Arctic-North Atlantic ridge-push forces, the Upper Rhine Graben is subjected to sinistral shear, the Roer Graben is under active extension, whilst thermal uplift of the Rhenish triple junction continues. Moreover, the late phase of accelerated subsidence of the North Sea Basin, commencing in the Pliocene, as well as the

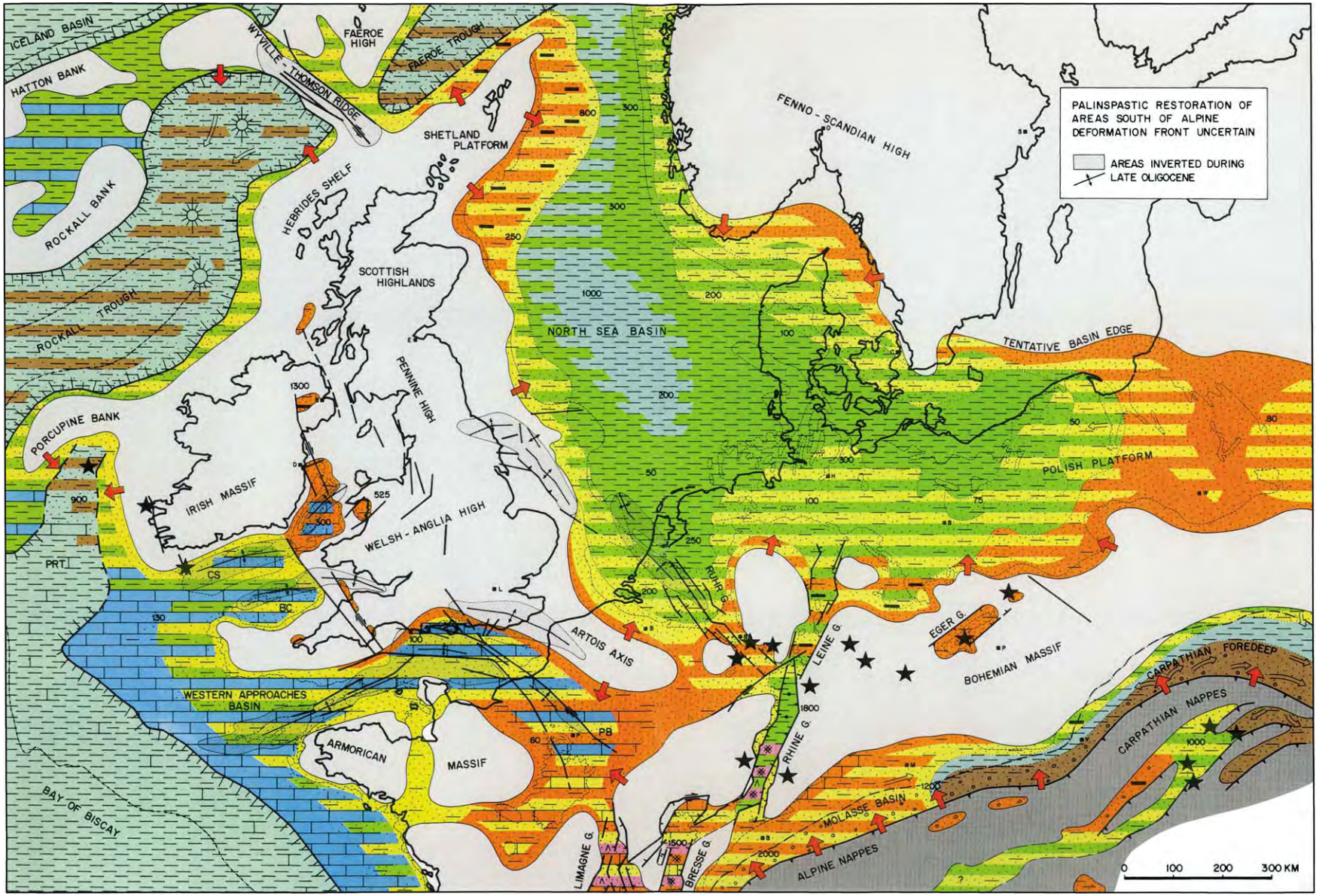


Figure 14 Oligocene palaeogeography. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

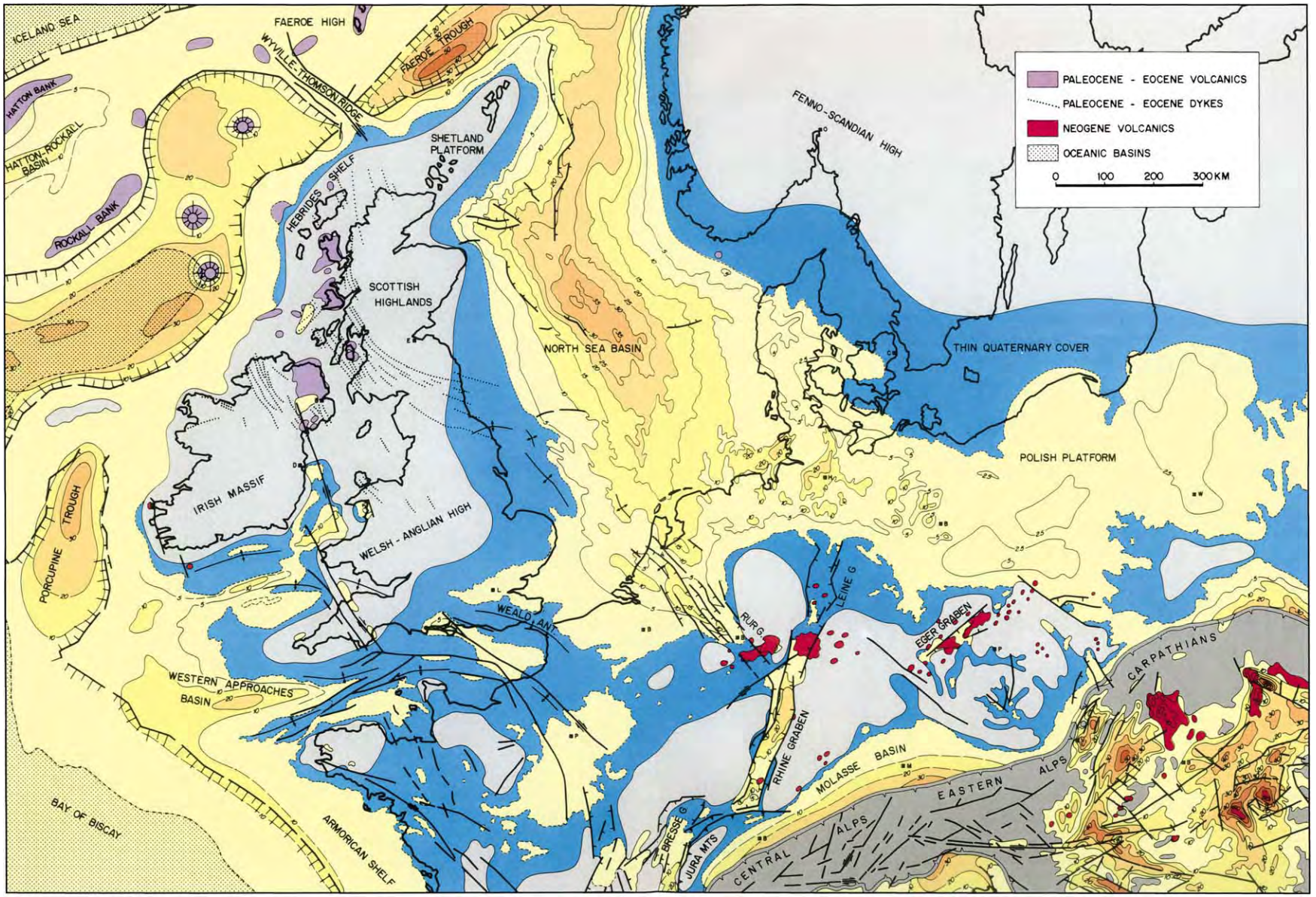


Figure 15 Isopach map of Cenozoic series. For legend see **Figure 16**. Details of Enclosure from *Geological Atlas of Western and Central Europe* 2nd Edition, Peter A. Zeigler, 1990, published by Shell International Petroleum Mij. B.V., distributed by Geological Society Publishing House, Bath.

Figure 16 Legend to palaeogeographic, isopach and tectonic maps and to stratigraphic correlation charts.

contemporaneous uplift of the British Isles and the Fennoscandian Shield, can be related to stress-induced deflection of the lithosphere.

Progressive uplift of the Rhenish and Bohemian massifs was coupled with the development of the modern drainage system of Central Europe and the shedding of massif clastics into the continuously subsiding North Sea Basin in which water depths gradually decreased. The Cenozoic isopach map (**Figure 15**) illustrates the broad saucer-shaped geometry of the North Sea thermal sag basin, the axis of which coincides with the trace of its underlying Mesozoic rift. This indicates that during Mesozoic rifting not only the crust but also the mantle-lithosphere of the North Sea Basin were significantly thinned (**Figure 2**). In contrast, Cenozoic sediments are generally thin on the Polish Platform, which is underlain by the deeply inverted Polish Trough that is marked by a crustal root. In northern Germany, sharp lateral Cenozoic thickness changes are related to the growth of Permian salt diapirs. A comparison of the present day erosional edges of Cenozoic basins and their depositional margins (**Figure 15**) illustrates the scope of Neogene to Recent uplift of the WCE in response to basin inversion and lithospheric folding.

Resources

The Permian and younger sedimentary basins of the WCE host a number of important hydrocarbon provinces, the most outstanding of which are tied to the Mesozoic North Sea rift system and the southern margin of the Southern Permian Basin. The hydrocarbon systems of the North Sea rift, as well as of the Faeroe trough, are largely tied to marine Kimmeridgian organic shales that charged reservoirs, ranging in age from Devonian to Palaeogene, with oil and gas. The gas-prone Southern Permian hydrocarbon province relies for hydrocarbon charge on Westphalian coal measures that were deposited in the Variscan foreland basin. Organic deeper water Zechstein shales and carbonates represent contributing source-rocks. Main reservoirs are Rotliegend sands, Zechstein carbonates, and Triassic sands. Early Jurassic marine organic shales control the petroleum systems of the Paris and Channel basins, as well as of the onshore parts of the North-west European basin in the Netherlands and Germany. The oil and gas province of the Aquitaine Basin relies for hydrocarbon charge mainly on marine Kimmeridgian and Berriasian shales. Hydrocarbons occurring in the Alpine foreland basin of Germany and Austria, as well as in the Upper Rhine Graben, were mainly derived from Oligocene marine shales.

Permian salts involved in diapric structures are widely exploited in the onshore parts of the WCE. Triassic salts are exploited in basins that are superimposed on Variscan crust. Polyhalites are associated with the Zechstein halites in the Southern Permian Basin and with Oligocene halites in the Upper Rhine Graben.

See Also

Africa: Rift Valley. **Europe:** Caledonides of Britain and Ireland; Permian Basins; The Alps. **Fossil Invertebrates:** Cephalopods (Other Than Ammonites). **Mesozoic:** Triassic; Jurassic. **Moho Discontinuity**.

Further Reading

BRGM, Societé Elf Aquitaine, Esso REP and SNPA (1973) *Géologie du Bassin d'Aquitaine*. Edition Bureau de Recherches Géologiques at Minières, Paris (Atlas).

Boldy SAR (ed.) (1995) Permian and Triassic Rifting in Northwest Europe. *Geological Society of London, Special Publication* 91: 263.

Cooper MA and Williams GD (eds.) (1989) Inversion Tectonics. *Geological Society of London, Special Publication* 44: 375.

Cope JCW, Ingham JK, and Rawson PF (eds.) (1992) Atlas of Palaeogeography and Lithofacies. *Geological Society of London, Mem.* 13: 152.

Dadlez R, Marek S, and Pokorski J (eds.) (1998) *Palaeogeographical Atlas of the Epicontinental Permian and Mesozoic in Poland*. 1:2,500,000. Panstwowy Instiytuy Geologiczny, Warsaw, 75 plates.

Debrand Passard S and Courbouleix S (1984) *Synthèse géologique du sud est de la France*. Vol. 2. Atlas: Stratigraphie et paléogeographie. Mém. B.R.G.M. 126.

Dèzes P, Schmid SM, and Ziegler PA (2004) Evolution of the European Cenozoic rift system: interaction of the Alpine and Pyrenean orogens with their foreland lithosphere. *Tectonophysics* (in press).

Glennie KW (ed.) (1998) *Petroleum Geology of the North Sea. Basic concepts and recent advances*, 4th Edn. Oxford: Blackwell Science.

Granet M, Wilson M, and Achauer U (1995) Imaging mantle plumes beneath the French Massif Central. *Earth Planetary Scientific Letters* 136: 199 203.

Kockel F (ed.) (1996) *Geotectonic Atlas of NW Germany, 1:300000*. Hannover, Germany: Federal Institute for Geosciences and Natural Resources.

Mégnien C (1980) Synthèse géologique du Bassin de Paris. *Mém. B.R.G.M.* 102: Vol. 2, Atlas.

Parker JR (ed.) (1993) *Petroleum Geology of Northwest Europe*. Proceedings of the 4th Conference. *Geological Society of London*, Vol. 1 & 2: 1542.

Parnell J (ed.) (1992) Basins of the Atlantic Seaboard: Petroleum geology, Sedimentology and Basin Evolution. *Geological Society of London, Special Publication* 62: 470.

Sissingh W (1998) Comparative stratigraphy of the Rhine Graben, Bresse graben and Molasse Basin: correlation of Alpine foreland events. *Tectonophysics* 300: 2249 284.
Stampfli G, Borel G, Cavazza W, Mosar J, and Ziegler PA (eds.) (2001) *The Paleotectonic Atlas of the PeriTethyan Domain*. CD ROM, European Geophysical Society.
Ziegler PA (1988) Evolution of the Arctic North Atlantic and the Western Tethys. *American Association Petroleum Geology, Mem*. 43; 198 p. and 30 plates.
Ziegler PA (1990) *Geological Atlas of Western and Central Europe,* 2nd Edn., Shell International Petroleum Minj. B.V., distrib. Geol. Soc., London, Publishing House, Bath, 238 p. and 56 encl.
Ziegler PA, Bertotti G, and Cloetingh S (2002) Dynamic processes controlling foreland development the role of mechanical (de)coupling of orogenic wedges and fore lands. *European Geophysical Society, Stephan Mueller Special Publication Series* 1: 29 91.

The Alps

O A Pfiffner, University of Bern, Bern, Switzerland

Introduction

The Alps as a Mountain Belt

The European Alps, a mountain chain with elevations reaching almost 5000 m, stretch from Nice to Vienna. The highest peak, Mont Blanc, reaches an elevation of 4807 m. Mont Blanc is part of a belt of granites that stretches from the Pelvoux massif in France to the High Tauern in Austria. The chain runs north–south from Nice, on northward, forming a 90° bend in Switzerland and then continuing eastward towards Vienna (Figure 1). It is narrowest in the transect of Switzerland. The mountain chain is dissected by numerous deeply incised valleys, some of which run parallel to the chain. To the north of the Alps, the Danube system drains into the Black Sea, the Rhine system drains into the North Sea, and the Rhone system drains into the Mediterranean. South of the Alps, the Po system drains into the Adriatic Sea.

The North-Alpine foreland basin, called the Molasse basin, stretches along the north side of the Alps. It was filled by sediments carried in by the rivers draining the Alps northward between 34 and 10 Ma. Similarly, the Po basin to the south of the Alps received the sediments from the Apennine chain and from the rivers draining the Alps southward. Both basins formed during the building of the mountain chain. The weight of the mountain chain flexed the tectonic plates on either side, creating depressions that readily filled and became shallow seas. Up to 35 km of rocks were eroded from the growing Alpine chain and accumulated in these depressions. Judging from the nature of the accumulated sediments, denudation of the growing chain kept pace with the vertical uplift. The rising mountain chain was probably never much higher than it is today.

Major Tectonic Units

The Alps formed as a result of the collision of the Eurasian and African plates, two continental tectonic plates that were initially separated by ocean basins. Starting around 100 Ma, these two plates moved closer to each other, closing the ocean basins between them and ultimately colliding. Consequently, on present-day tectonic maps of the Alps (Figure 2), it is possible to distinguish between rock suites pertaining to one or the other of these continents, or the ocean basins between them. The Helvetic zone, Jura Mountains, and the area north and west of the Alps pertain to the former European margin of the Eurasian Plate; the Austroalpine and Southalpine zones are parts of the former Adriatic margin of the African Plate. The Penninic zone is made up of sediments that accumulated in ocean basins that were located between the two continents, as well as the crustal rocks underlying these basins. During the closure of these basins and the ensuing collision of the two continental plates, the European margin was dragged down south-eastward beneath the Adriatic margin. The Eastern Alps are dominated by rocks of the upper plate, the Adriatic margin (Figure 2). Erosion has removed this upper plate almost completely in the Central and Western Alps. However, erosional remnants (termed 'klippen') of the upper plate in the Western and Central Alps, as well as pebbles carried out into the foreland by ancient rivers, prove that it once occupied much of the entire Alps.

The ocean basins between the Eurasian and Adriatic continental plates formed in response to the opening of the Atlantic Ocean. Figure 3 shows the palaeogeography at 170 and 130 Ma. Two basins, the Valais basin and the Piemont Ocean (sometimes called the Liguria–Piemont Ocean), formed between the European and Adriatic margins of the two continental plates. The two basins were separated by a microcontinent, the Briançonnais swell, which was connected to the Iberian Peninsula at the time. Plate

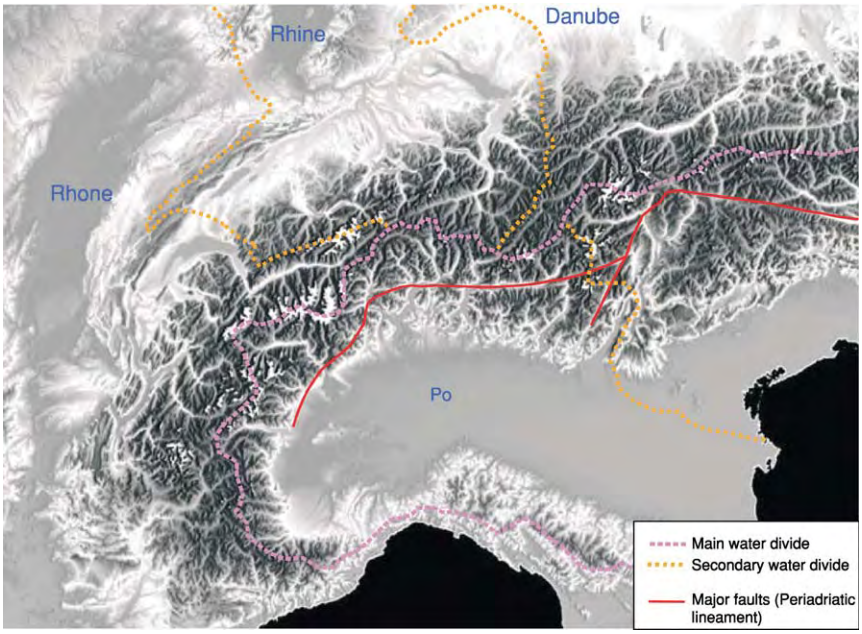


Figure 1 Digital elevation model, showing the large scale geomorphic features of the Alps and surrounding areas.

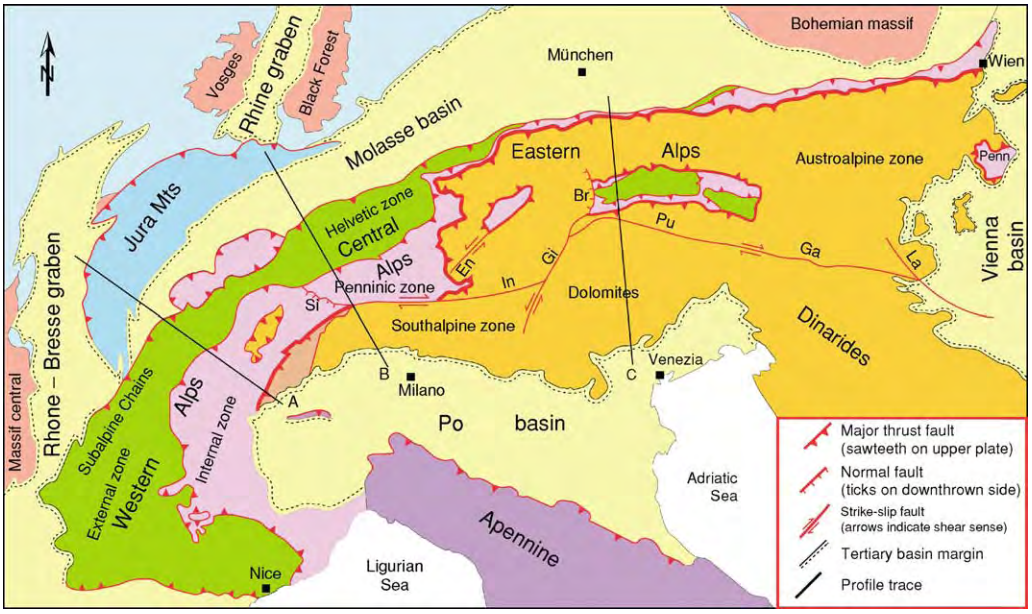


Figure 2 Tectonic map of the Alps, showing the major tectonic units. Sites A, B, and C correlate to orogen profiles A, B, and C in Figure 4. Faults: Br, Brenner; Si, Simplon; En, Engadine; In, Insubric; Gi, Giudicarie; Pu, Pustertal; Ga, Gailtal; La, Lavanttal.

movements during the opening of the Atlantic were such that the African plate moved eastward relative to the Eurasian plate, and both plates moved away from the North and South American plates. At an early stage (around 170 Ma), this sinistral plate movement occurred along a fracture zone that passed through Gibraltar. This opened the Piemont ocean in the area of the future Alps. At a later stage (around 130 Ma), when the opening of the Atlantic had proceeded further north, the sinistral movement occurred along a fault zone passing north of the Iberia–Briançonnais continental fragment along the Gulf of Biscay. This opened the Valais basin in the area of the future Alps.

The fate of the Piemont and Valais basins was controlled by convergence between the Eurasian and

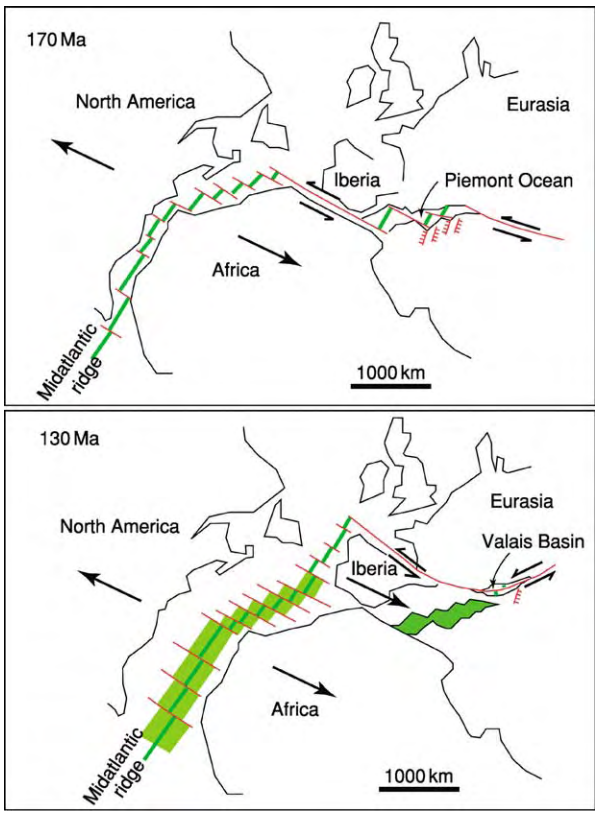


Figure 3 Palaeogeographical maps at 170 and 130 Ma, showing the future Alpine domain in the framework of the associated tectonic plates. Opening of the Piemont Ocean and the Valais Basin was closely linked to the opening of the Atlantic.

African plates. During this convergence, the basins were subducted beneath the Adriatic margin and were closed. Their lower part (the lithospheric mantle) was subducted and recycled into the mantle. Their upper parts (the crust and the sediments deposited in the two basins) were compressed and incorporated into the growing Alpine Orogen, where they now form the Penninic nappes. The European and Adriatic margins were also compressed during convergence and collision of the two plates. Consequently, the Helvetic nappes and the Jura Mountains, both of which consist essentially of Mesozoic shelf sediments, were formed on the European side. Similarly, the Austroalpine and Southalpine nappes represent the deformed Adriatic margin of the African plate.

Rock Types

When discussing the rock types that can be found in the Alps, it is useful to distinguish between rocks that formed prior to the opening of the Piemont and Valais basins and rocks that formed during and after the opening of the basins. In terms of Alpine geology, the older rocks are referred to as basement. This basement consists of two major units, crystalline rocks (granites and polymetamorphic gneisses and schists) and Palaeozoic sediments and volcanics, that pertain to mountain belts formed at 300 to 400 Ma. Granitic rocks are resistant to erosion, thus it is no surprise that they form many of the higher peaks in the Alps (including Mont Blanc).

The younger rocks are Mesozoic and Cenozoic sediments and volcanics ranging in age from 225 to 10 Ma. Large quantities of carbonates accumulated along the shelf seas of the continental margins, reaching thicknesses of more than 1 km. These carbonates now form the high cliffs that dominate the present-day morphology of the Alps. Sandstones and shales accumulated as basins on the continental slopes. In some instances, these basins were flanked by faults that formed in response to the breakup of the continents. Breccias accumulated at the foot of the steep fault scarps. The deepest part of the basins consisted of newly formed oceanic crust. Deep-sea sediments (radiolarian cherts) slowly covered the basaltic lava flows of the newly formed ocean floor.

Deep Structure of the Alps

A number of experiments have been designed to image the structure of the Alps to depths of over 50 km. Dynamite detonations and vibrator trucks located over a subsurface target generate seismic waves that travel downward and then are reflected back upward at various discontinuities in the Earth's crust. The upward-reflected waves (or echoes) are recorded by a surface array of geophones and processed into a coherent 2-dimensional image. The resulting seismic sections can then be interpreted in terms of subsurface geological structure. Figure 4 summarizes the findings for three transects through the Alps.

Western Alps

Within the framework of ECORS-CROP (CROP = CROsta Profonda = Deep Crust; ECORS = Etude de la Croûte Continentale et Océanique par Réflexion et Réfraction Sismiques = Study of the continental and oceanic crust by reflection and refraction sesimics), a joint project between France and Italy studying the deep continental crust by reflection and refraction sesimics, researchers have profiled a transect across the Western Alps (site A in Figure 2). The profile in Figure 4A summarizes the findings of the ECORS–CROP project. The Western Alps have an asymmetric structure. On the European margin, i.e., in the western part of the Alpine orogen, the crust–mantle

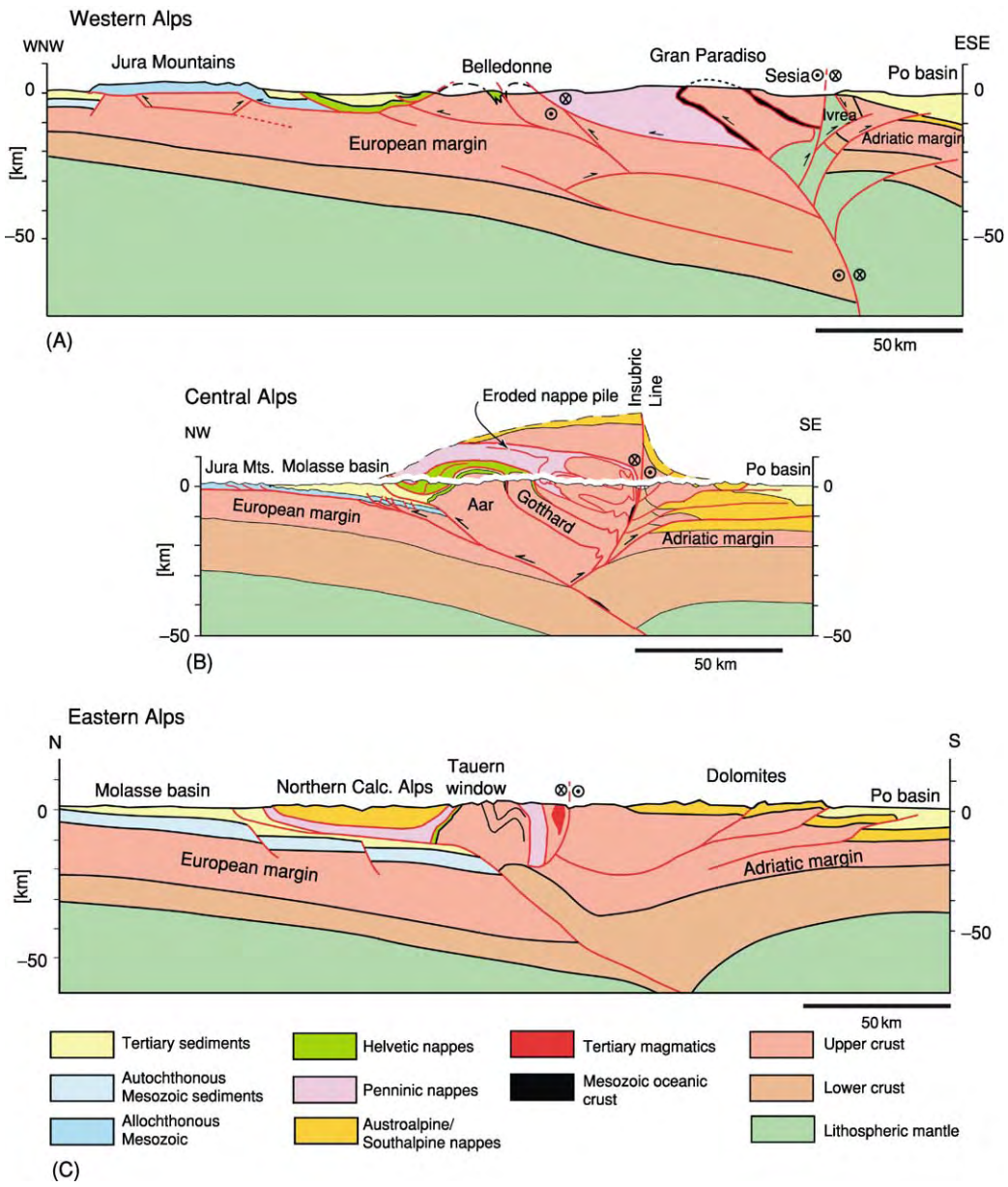


Figure 4 Three profiles through the Alps, showing the deep structure of the orogen. The profile sites are shown in **Figure 2**. (A) Transect through the Western Alps of France and Italy. (B) Transect through the Central Alps of Switzerland and Italy. (C) Transect through the Eastern Alps of Austria and Italy. See text for discussion.

boundary dips to the east, attaining depths exceeding 50 km beneath the orogen. Much of the crustal root is made up of lower crust, which appears to be tripled in this transect. This stacking was accomplished by thrust faults that moved entire blocks of crustal rocks upward towards the west. On the Adriatic margin, on the other hand, the crust–mantle boundary rises towards the surface, proceeding westward. This rise is accentuated by several thrust faults, which also affect the lower crust. In the case of the Ivrea zone, these lower crustal rocks and pieces of the mantle actually outcrop at the surface. The asymmetric structure of the Western Alps is a consequence of the collision of the Adriatic and European margins resulting from the convergence of the African and Eurasian plates. The ocean basins located between the two margins were highly deformed by these plate movements. The sediments, as well as their crustal substrate, were compressed, shortened, and stacked. These rocks now form the Penninic nappes shown in **Figure 4A**. The upper crust of the European margin was also shortened and thickened in the process of collision, forming a large-scale basement uplift (the Belledonne massif) within the Alps.

Further out towards the foreland, shelf sediments of the European margin were detached from their substrate, shortened by folding and thrusting, and transported towards the west. They now form the Jura Mountains. In addition to all of the thrusts and folds that represent WNW–ESE shortening, there were substantial movements in and out of the plane of the section shown in Figure 4A. These movements were related to strike–slip motions that displaced the Adriatic margin towards the north relative to the European margin.

Central Alps

Investigation of the Central Alps by the Swiss National Research Project (NRP) 20 has provided results complementary to those of the European GeoTraverse (EGT). The EGT study assessed the continental lithosphere that runs from the North Cape across Europe to Tunisia. Several transects were profiled across and within the Swiss Alps. Figure 4B, a profile along the central traverse of NRP 20, has been extended to include the Jura Mountains and the Po basin. As with the Western Alps, the asymmetric structure of the Central Alps evolved during convergence between Eurasia and Africa. The lower crust of the European margin extends at constant thickness beneath the Adriatic lower crustal wedge. The tip of the latter is exposed at the surface. The centre of the orogen consists of European margin upper crustal rocks that were stacked by thrust faults during plate convergence and collision. The associated heat and pressure transformed the rocks: granites became orthogneisses, sediments became paragneisses, and limestones became marbles. The mineral assemblages "frozen" in these metamorphic rocks allow determination of the temperature and pressure paths these rocks took during collision and the ensuing denudation.

Along one major fault, the Insubric Line, the nappe stack was moved upward and southward, but erosion kept pace with this uplift and removed a large section of the Austroalpine and Penninic nappes. Studies of the rocks outcropping at the surface indicate that they were once buried to depths exceeding 25 km. Strike–slip motion along the Insubric Line moved the Adriatic margin westward relative to the European margin in the later stages of the collision. In the north-western part of the orogen, the thrusting was chiefly towards the north-west. Conversely, thrusting was directed towards the south-east in the south-eastern part of the orogen. In both cases, the sequence of thrusting was from the centre of the orogen towards the forelands. This can be interpreted as the result of the collision of the European and Adriatic continental margins. Deformation of the crust occurred within the zone of contact and the deformed rocks were shoved on top of each other to form an orogenic wedge, similar to the wedge of snow forming in front of a moving snowplough. Shortening related to plate convergence and collision is particularly accentuated along this transect of the Alps. The associated uplift and erosion are indicated by the high degree of metamorphism of the exposed rocks outcropping at the surface. The focused horizontal shortening (and vertical stretching) explain why the Alpine chain is particularly narrow in this transect.

Eastern Alps

The Eastern Alps have been studied in a joint project (TRANSALP) between Germany, Austria, and Italy. A seismic transect through the Eastern Alps and the Dolomites has produced the seismic data and geological interpretations shown in Figure 4C. The crustal structure of the Eastern Alps is rather different from that of the western and central sections. The European margin shows a thin slab of lower crust dipping to the south. The Adriatic lower crust is thicker and has a piece protruding into the European upper crust. The upper crust of the European margin forms a large-scale basement uplift that is exposed in the Tauern window. The Helvetic nappes are virtually absent in this transect and the Penninic nappes make up only a proportionally small volume . The upper crust of the Adriatic margin, on the other hand, is much more voluminous. It is over 30 km thick and is shortened by thrust faults. Important strike–slip motions have displaced the Adriatic margin to the west, relative to the European margin along the Pustertal line.

Alpine Nappe Structures

Collision between the two margins of the Eurasian and African plates, like the crash of two cars, led to severe deformation (folding and fracturing) of the rocks within the zone of contact. Large-scale fractures in a zone of compression led to the formation of thrust faults that transported entire crustal blocks upward on a gently inclined fault surface. The displaced blocks were typically several hundreds of kilometres wide, 50–100 km long, and only a few kilometres thick. These thrust sheets are termed 'nappes', from the French for 'tablecloth', because of their shape. The (horizontal) shortening of the continental crust also led to the folding of layers of rock. The folds, which occur on every scale, from millimetres to kilometres, are an expression of the penetrative deformation of crustal rocks.

Thrust Faults

Geologists in the nineteenth century were puzzled when they saw older rocks lying on top of younger rocks. One of the classic locations where this is readily observed is in eastern Switzerland (Figure 5). The rugged peaks of the Tschingelhore are formed by Permian clastic sediments (roughly 260 Ma), but the rock immediately beneath the peaks consists of Jurassic limestones (formed about 150 Ma). The famous Glarus thrust is the sharp contact between the close proximity of the Permian and Jurassic units. The geographical extension of the Glarus thrust can be seen in a profile across eastern Switzerland (Figure 6). The Helvetic nappes were displaced northward along the Glarus thrust over a distance of 50 km. In order to displace a thrust sheet, a weak, basal lubricating layer must be present. In the case of the Glarus thrust, this lubricating layer is a highly sheared limestone, which was scraped off the footwall and dragged and drawn out along the thrust surface. It now forms a layer roughly 1 m thick and can be traced northward over a distance of some 30 km (in Figure 5 it is visible as a thin, light line along the horizontal contact).

As is evident from Figure 6, substantial internal deformation affected the rocks below and above the Glarus thrust. Highly sheared folds and thrust faults, repeating the various sedimentary layers many times, can be observed in the Mesozoic sediments beneath the Glarus thrust. Similarly, the rocks of the displaced block above the Glarus thrust show thrust faults and folds. One of these thrust faults, the Säntis thrust, displaced the younger Cretaceous strata much further north, compared to the older (Triassic to Jurassic) strata. In this case, the thrust fault was lubricated by a thick layer of shales, which have relatively low shear resistance.

Nappe Internal Deformation

As a thrust sheet is compressed, detached from its substrate, and transported, it undergoes internal deformation. The type of deformation depends largely on the nature of the rocks involved. In the case of a layered sequence of sedimentary rocks, folding prevails, whereas if the mechanical contrasts are low, such as sometimes occurs in a suite of granitic and gneissic rocks, the deformation may be more homogeneous. If mechanically weak layers of rocks are present, the deformation is focused and thrust faults may develop. All of these processes – folding, faulting, and homogeneous deformation – may take place jointly.

Figure 7 shows folded Cretaceous and Eocene limestones (100 to 50 Ma) that form an asymmetric anticline. The folded limestones are overlain by older rocks. Dolomites (~220 Ma) form the yellow cliff beneath the summit and crystalline basement rocks (>300 Ma) make up the dark summit. These older

Figure 5 The Glarus thrust in the Tschingelhore (between Flims and Elm in eastern Switzerland). The thrust fault is visible as a sharp horizontal contact between the older rocks that form the rugged peaks and the younger rocks that form the cliffs above the snowfields.

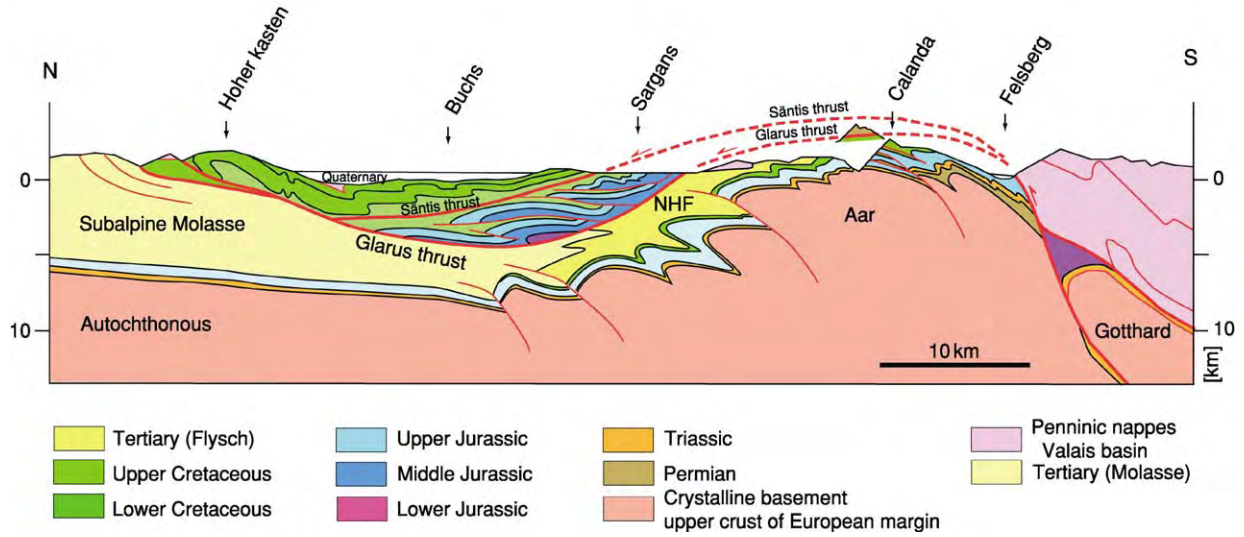


Figure 6 Profile through the Helvetic nappes of the eastern Swiss Alps. The Helvetic nappes were displaced along the Glarus thrust over a distance of around 50 km. But the rocks above and below the Glarus thrust were also intricately deformed, as is evident from the fold and thrust structures. The Santis thrust displaced the uppermost part, the Cretaceous strata, of the Helvetic nappes an additional 10 km to the north. Deeper down, the crystalline basement rocks of the Aar massif now form an anticlinal upwarp. NHF, North Helvetic Flysch.

Figure 7 Folded strata in the flank of Piz d'Artgas ('peak with arcs'), overthrust by older rocks forming the summit and the yellow cliff beneath.

rocks were emplaced along a thrust fault that is located near the base of the yellow cliff. Figure 8 is a profile across the Tauern window, where the upper crust of the European margin forms a large anticlinal fold. In the centre of the upwarp, erosion has removed the higher nappes, thus providing an insight into the formerly deeper parts of the orogen. The crystalline basement rocks in the core of the upwarp were compressed and internally shortened. From the deformed mineral grains of the rocks it is possible to determine how much horizontal shortening and vertical stretching actually occurred and to reconstruct the shape of the upwarp prior to this homogeneous deformation. The present-day shape of the Tauern upwarp (Figure 6), as well as its reconstructed geometry prior to homogeneous shortening, provide a

reminder of the ductile behavior that granitic rocks can exhibit in the course of plate collision.

The Klippen nappe, a Penninic nappe in the French–Swiss Alps, is a classic example of a style of internal deformation characterized by fold-and-thrust structures (Figure 9). The lubricating layer (evaporites) at the base of the nappe consists of a thick layer of anhydrite. This rock type, which has a particularly low shear strength, forms when very shallow areas of seawater evaporate. The great thickness of the weak evaporite layer in the northern part of the section shown in Figure 9 facilitated the formation of large-scale folds, and the anhydrite was able to flow into and fill the fold cores. In the southern part of the nappe, the anhydrite layer is thinner and the deformation style is characterized by imbricate thrusting. Each thrust fault is parallel to the strata and followed the weak anhydrite layer.

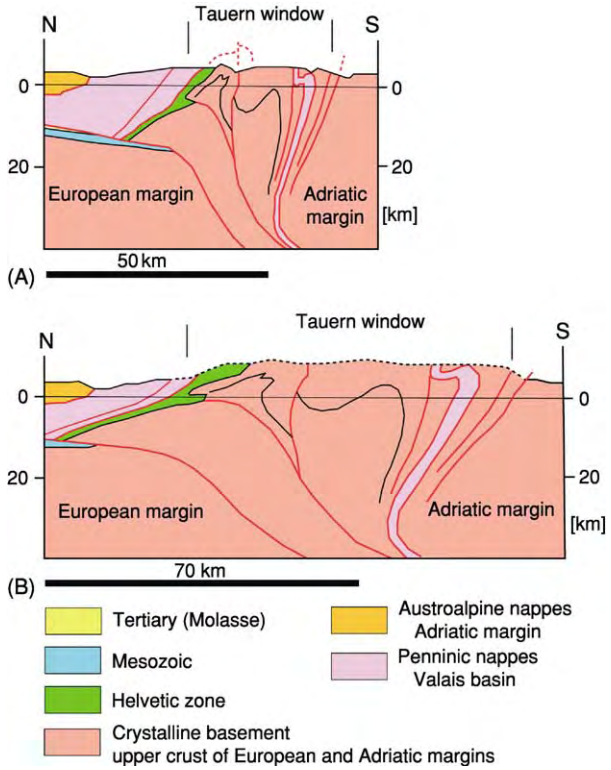


Figure 8 Profile across the Tauern window (Eastern Alps). (A) Present day geometry; (B) retrodeformed to the configuration that existed prior to homogeneous horizontal shortening and vertical stretching.

The Making of the Alps

Geologists working in the Alps had recognized early on that oceanic sediments occurred within the mountain range and were juxtaposed with rock units typical for continents. The pyramid of the Matterhorn (Figure 10), for example, is composed of crystalline basement rocks that were formed more than 300 Ma and which originated in the former (Adriatic) margin of the continental African Plate. In contrast, the base of the pyramid consists of volcanic and sedimentary rocks that formed in an ocean basin (the Piemont Ocean) 170 to 100 Ma ago. The Piemont Ocean formed in response to divergent motion between the Eurasian and African plates (see Figure 3). The Alpine Orogen evolved in a number of steps associated with relative movements between the Eurasian and African plates. The ocean basins between the two continental plates were closed in the process. The

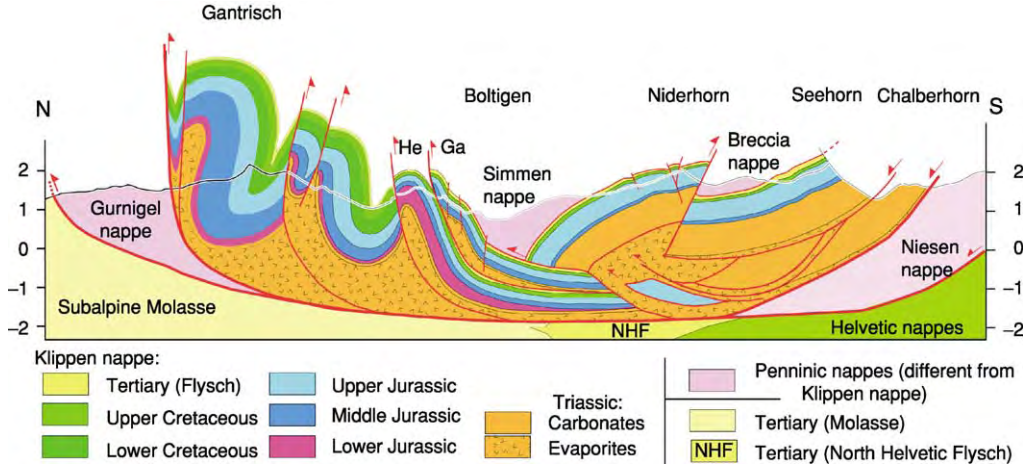


Figure 9 Profile across the Penninic Klippen nappe of the western Swiss Alps. The Klippen nappe consists of sediments of the former Briançonnais swell that have been overthrust onto sediments scraped off of the Valais basin and the Piemont ocean (the Niesen and Gurnigel nappes, respectively). The nappe internal structure of the frontal north west part of the Klippen nappe is dominated by folding, whereas in the internal south east part, imbrications stemming from thrust faulting prevail. Ga: Gastlosen thrust, He: Heiti thrust.

Figure 10 Crystalline basement rocks pertaining to the margin of the African continent build up the Matterhorn peak and overlie the younger volcanic and sedimentary rocks that formed in the Piemont Ocean.

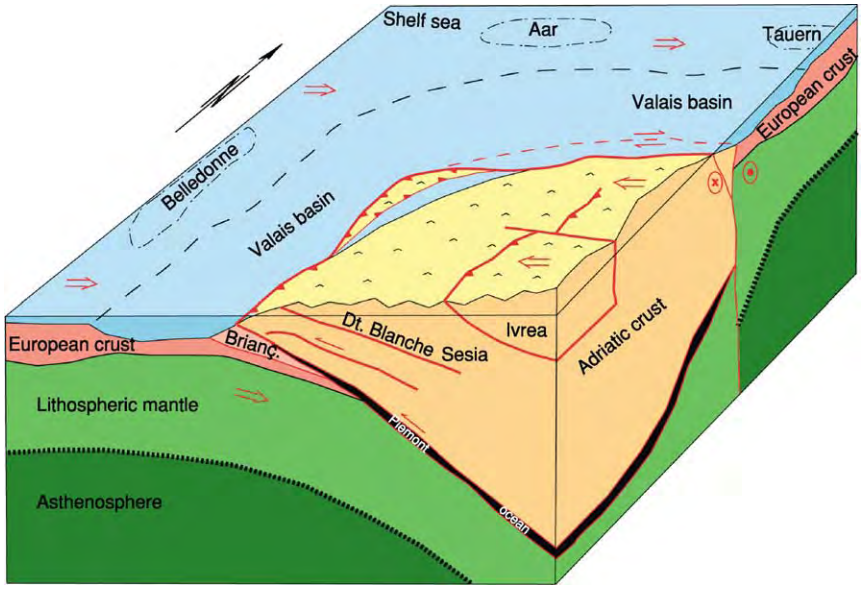


Figure 11 Block diagram showing the three dimensional geometry of the ancestral Alps at 90 Ma. An east dipping subduction zone in the Western Alps had consumed the Piemont Ocean. The Briançonnais continental fragment was entering this subduction zone. The Valais basin and the shelf seas of the European margin were the site of ongoing sedimentation.

first basin, the Piemont Ocean, closed in Cretaceous times (~100 Ma). The second basin, the Valais, closed in Tertiary times (~35 Ma). Closure of these basins resulted not only from head-on collision, but also involved strike–slip movements between the European and Adriatic margins.

During Cretaceous times, convergence between the Eurasian and African plates was directed east and west. The European margin (Figure 11) was approaching the Adriatic margin, which had already formed an ancestral mountain range. The Piemont Ocean had already been subducted along an east-dipping subduction zone. Small fragments of this ocean were scraped off of the descending plate and were attached to the upper plate, a process called 'underplating'. The Briançonnais microcontinent

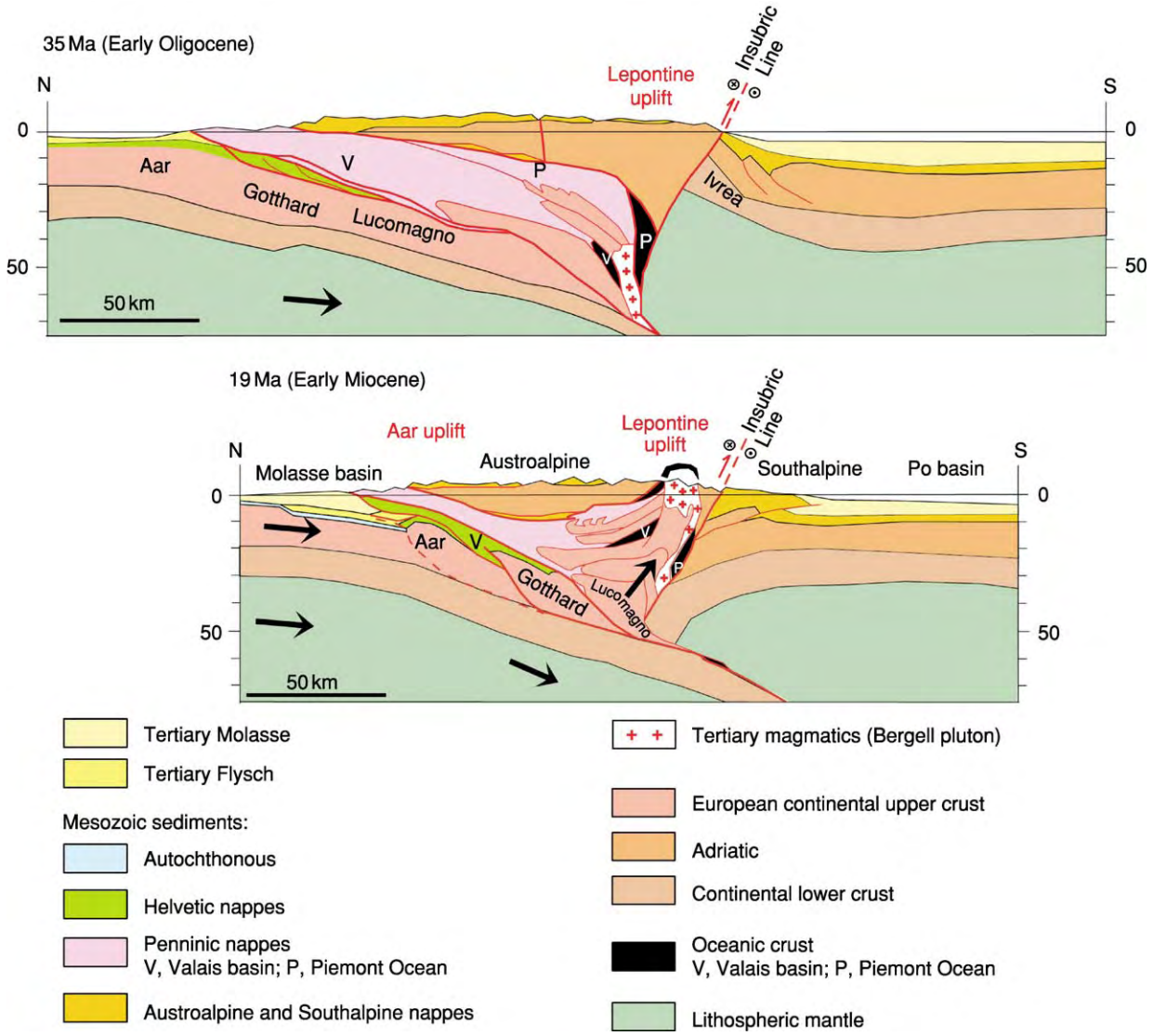


Figure 12 North south profile through the Central Alps of eastern Switzerland, reconstructed to the geometry at 32 and 19 Ma. Comparison of the two profiles reveals that the orogen grew outward with time on both sides, and that the units in the central part of the orogen were raised to a higher level by the combined action of folding and erosional denudation.

was just in the process of being subducted, but parts of it were also attached to the upper plate. The Valais basin was still the site of sedimentation, as was the shelf of the European margin. In the region of the future Central and Eastern Alps, the east–west convergence was expressed as east–west dextral strike–slip movements. At about 40 Ma, the convergence between the Eurasian and African plates changed to a north–south orientation. As a consequence, a south-dipping subduction zone evolved, into which the Valais basin gradually disappeared. Again, a number of fragments were scraped off of the descending plate and were accreted to the upper plate. About 35 Ma, the two margins started to collide. During this north–south convergence, strike–slip movements took place in the ancestral Western Alps. In the Central and Eastern Alps, the collision phase compressed the two margins and led to the stacking of crustal pieces, horizontal shortening, and vertical stretching. Figure 12 shows two stages of this collision phase in a cross-section trough the Central Alps, reconstructed for 35 and 19 Ma. The deformation of the two continental margins pushed crustal fragments up inclined thrust faults and uplifted parts of the orogen by large-scale folding and vertical stretching. As a consequence, the land surface of the ancestral Alps was uplifted. The ensuing high elevations caused precipitation and triggered enhanced erosion. Rivers built large fan deltas in the foreland of the Alps. As far as known, denudation kept pace with uplift during mountain building. Nevertheless, deep crustal fragments were exposed in the process, bringing to the surface samples of rock

that had been at depths of several tens of kilometres during the early stages of the formation of the Alps.

See Also

Europe: Mediterranean Tectonics; Variscan Orogeny; Permian to Recent Evolution. **Moho Discontinuity**.

Further Reading

Debelmas J (1974) *Géologie de la France*. Paris: Doin.

Debelmas J (1979) *Découverte Géologique des Alpes du Nord*. Orleans: BRGM.

Debelmas J (1982) *Découverte Géologique des Alpes du Sud*. Orleans: BRGM.

Krenmayr HG (ed.) (2000) *Rocky Austria: A Brief Earth History of Austria*. Wien: Geological Survey of Austria.

Labhart TP (2001) *Geologie der Schweiz,* 5th edn. Thun: Ott Verlag.

Labhart TP and Decrouez D (1997) *Géologie de la Suisse*. Lausanne: Delachaux et Niestlé.

Lemoine M, deGraciansky P C, and Tricart P (2000) *De l'Océan à la Chaîne de Montagnes: Tectonique des Plaques dans les Alpes*. New York: Gordon & Breach.

Marthaler M (2001) *Le Cervin est il Africain?* Lausanne: Loisir et Pédagogie.

Marthaler M (2002) *Das Matterhorn aus Afrika: Die Entstehung der Alpen in der Erdgeschichte*. Thun: Ott Verlag.

Neubauer F and Höck V (eds.) (2000) *Aspects of Geology in Austria. Reports of the Austrian Geological Society, Special Issue 92(1999)*. Wien: Austrian Geological Society.

Nicolas A, Polino R, Hirn A, Nicolich R, and ECORS CROP Working Group (1990) ECORS CROP tra verse and deep structure of the western Alps: a synthesis. In: Roure F, Heitzmann P, and Polino R (eds.) *Deep Structure of the Alps,* vol. 156, *Mémoires de la Societe Géologique de France,* pp. 15 28. Paris: Geological Society of France.

Pfiffner OA, Lehner P, Heitzmann P, Mueller St, and Steck A (eds.) (1997) *Deep Structure of the Swiss Alps. Results of NRP 20*. Basel: Birkhäuser.

Roure F, Bergerat F, Damotte B, Mugnier J L, and Polino R (1996) The ECORS CROP Alpine seismic traverse. *Mémoires de la Societe Géologique de France* 170.

TRANSALP Working Group (2002) First deep seismic reflec tion images of the Eastern Alps reveal giant crustal wedges. *Geophysical Research Letters* 29(10): 92 1 92 4.

Mediterranean Tectonics

E Carminati and C Doglioni, Università La Sapienza, Rome, Italy

Introduction

It is commonly accepted that Mediterranean geology has been shaped by the interplay between two plates, the African and European plates, and possibly also smaller intervening microplates. The Mediterranean was mainly affected by rifting after the Variscan Orogeny (*see* **Europe:** Variscan Orogeny): during the Mesozoic, oceanic Tethys areas and passive continental margins developed, where widespread carbonate platforms were formed. During the Late Mesozoic, the Mediterranean area was dominated by subduction zones (from east to west, the Cimmerian, Dinarides, and Alps–Betics), which inverted the extensional regime, consuming the previously formed Tethyan oceanic lithosphere and the adjacent continental margins. The composition (oceanic or continental), density, and thickness of the lithosphere inherited from the Mesozoic rift controlled the location, distribution, and evolution of the later subduction zones. The shorter wavelength of the Mediterranean orogens relative to other belts (for example, the Cordillera and the Himalayas) is due to the smaller wavelength of the lithospheric anisotropies inherited from the Tethyan rift.

The Mediterranean basin was, and still is, a collector of sediments derived from the erosion of the surrounding continents and orogens: the best examples are the Nile and Rhone deltas. In the past, other deltas deposited sediments in the bottom of the Mediterranean, and their rivers were later disconnected or abandoned: an example is the Upper Oligocene–Lower Miocene Numidian Sandstone, which was derived from Africa, deposited in the central Mediterranean basin, and partly uplifted by the Apennines accretionary prism. It is well known that, during the Messinian eustatic lowstand, the Mediterranean dried up several times, generating a salinity crisis during which thick sequences of evaporites were deposited in the basin. This generated a pulsating loading oscillation in the Mediterranean, because the repetitive removal of the water led to significant isostatic rebound across most of the basin, particularly where it was deeper, as in the Ionian, the Provençal, and the central Tyrrhenian seas.

The direction of the relative motion between Africa and Europe since the Neogene is still under debate.

Most reconstructions show directions of relative motion between north-west and north-east. Recent space geodesy data confirm this overall trend, in which Africa has a north–south component of convergence relative to Europe of about 5 mm year $^{-1}$, but they also show that the absolute plate-motion directions of both Europe and Africa are north-east and not north or north-west as is usually assumed (see the NASA database on present global plate motions, http://sideshow.jpl.nasa.gov:80/mbh/series.html).

The main Cenozoic subduction zones in the Mediterranean are the Alps–Betics, the Apennines–Maghrebides, and the Dinarides–Hellenides–Taurides. Closely related to the Mediterranean geodynamics are the Carpathian subduction and the Pyrenees (Figure 1). The Mediterranean orogens show two distinct signatures, which are similar to those occurring on opposite sides of the Pacific Ocean. High morphological and structural elevations, double vergence, thick crust, involvement of deep crustal rocks, and shallow foredeeps characterize eastwards- or north-eastwards-directed subduction zones (Alps–Betics and Dinarides–Hellenides–Taurides). Conversely, low morphological and structural elevations, single vergence, thin crust, involvement of shallow rocks, deep foredeeps, and a widely developed back-arc basin characterize the westwards-directed subduction zones of the Apennines and Carpathians. This asymmetry can be ascribed to the 'westward' drift of the lithosphere relative to the mantle, at rates of about 49 mm year $^{-1}$ as computed from the hotspots reference frame. All Mediterranean orogens show typical thrust-belt geometries with imbricate-fan and antiformal-stack associations of thrusts. The main factor that varies between orogens and within single belts is the depth of the basal décollement. The deeper it is, the higher is the structural and morphological elevation of the related orogen.

Extensional basins are superimposed on these orogenic belts: on the western side are the Valencia, Provençal, Alboran, Algerian, and Tyrrhenian basins, on the eastern side is the Aegean Basin, and to the north is the Pannonian Basin (Figures 2 and 3).

The Mediterranean can be divided into western, central, and eastern basins. The western Mediterranean is younger (mainly less than 30 Ma) than the central Mediterranean and eastern Mediterranean, which are mainly relics of the Mesozoic to possibly Cenozoic Tethys Ocean.

Positive gravity anomalies occur in the deep basins (the Provençal, Tyrrhenian, and Ionian seas), where the mantle has been uplifted by rifting processes. In contrast, negative gravity anomalies occur along the subduction zones.

Western Mediterranean

A characteristic feature of the western Mediterranean is the large variation in lithospheric and crustal thickness (Figure 5). The lithosphere has been thinned to less than 60 km in the basins (50–60 km in the Valencia trough, 40 km in the eastern Alboran Sea, and 20–25 km in the Tyrrhenian Sea), while it is 65–80 km thick below the continental swells (Corsica–Sardinia and the Balearic promontory). The crust mimics these differences, with a thickness of 8–15 km in the basins (Valencia trough, Alboran Sea, Ligurian Sea, and Tyrrhenian Sea) and 20–30 km underneath the swells (Balearic promontory and Corsica–Sardinia), as inferred by seismic and gravity data. These lateral variations in thickness and composition are related to the rifting process that affected the western Mediterranean, which is a coherent system of interrelated irregular troughs, mainly V-shaped, that began to develop in the Late Oligocene–Early Miocene in the westernmost parts (Alboran, Valencia, Provençal basins), becoming progressively younger eastwards (eastern Balearic and Algerian basins), culminating in the presently active east–west extension in the Tyrrhenian Sea (Figures 1, 2, 3, and 4). Heat flow data and thermal modelling show that the maximum heat flows are encountered in the basins: 120 mW m $^{-2}$ in the eastern Alboran Sea, 90–100 mW m $^{-2}$ in the Valencia trough, and more than 200 mW m $^{-2}$ in the Tyrrhenian Sea. All these sub-basins appear to be genetically linked to the back-arc opening related to the coeval 'eastwards' rollback of the westward-directed Apennines–Maghrebides subduction zone. Extreme stretching generated oceanic crust in the Provençal (20–15 Ma), Algerian (17–10 Ma), Vavilov and Marsili (7–0 Ma) basins. Between 25 Ma and 10 Ma, the Corsica–Sardinia block rotated 60° counterclockwise (Figures 1, 2, 3, and 5).

In the southern Apennines, the choking of the subduction zone with the thicker continental lithosphere of the Apulia Platform slowed the eastwards migration of the subduction hinge (Figure 6), whereas in the central and northern Apennines and in Calabria subduction is still active owing to the presence in the foreland of the thin continental lithosphere of the Adriatic Sea and the Mesozoic oceanic lithosphere of the Ionian Sea, allowing rollback of the subduction hinge.

The western Mediterranean basins tend to close both morphologically and structurally towards the south-west (Alboran Sea) and north-east (Ligurian Sea; Figures 1 and 6). The eastwards migration of the arc associated with the westwards-directed subduction generated right-lateral transpression along the entire east–west-trending northern African belt

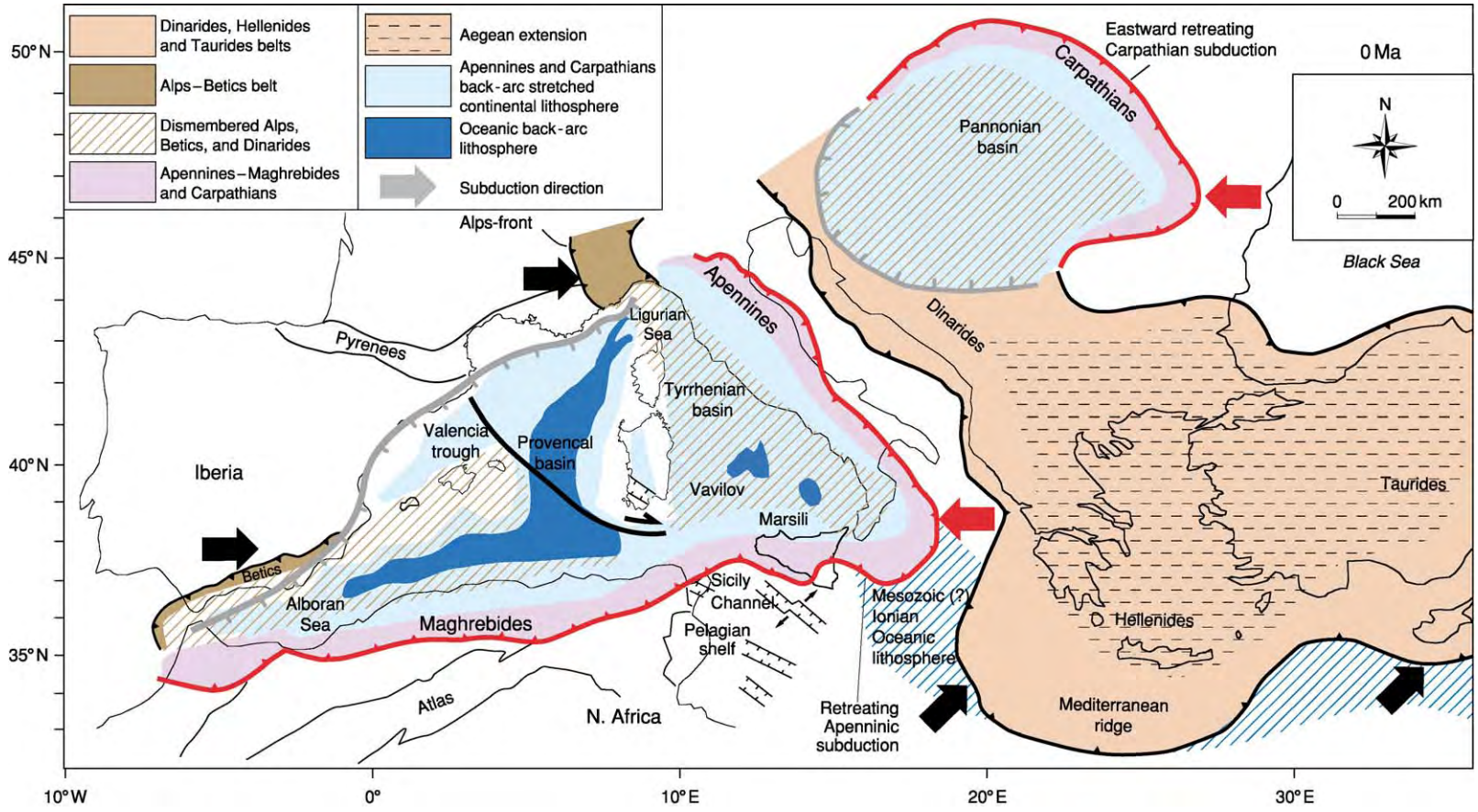


Figure 1 Present geodynamic framework. There are four subduction zones with variable active rates in the Mediterranean realm: the westwards directed Apennines Maghrebides; the westwards directed Carpathians; the north eastwards directed Dinarides Hellenides Taurides; and the south eastwards directed Alps. The Apennines Maghrebides subduction related back arc basin of the western Mediterranean stretched and scattered into segmented basins most of the products of the Alps Betics orogen.

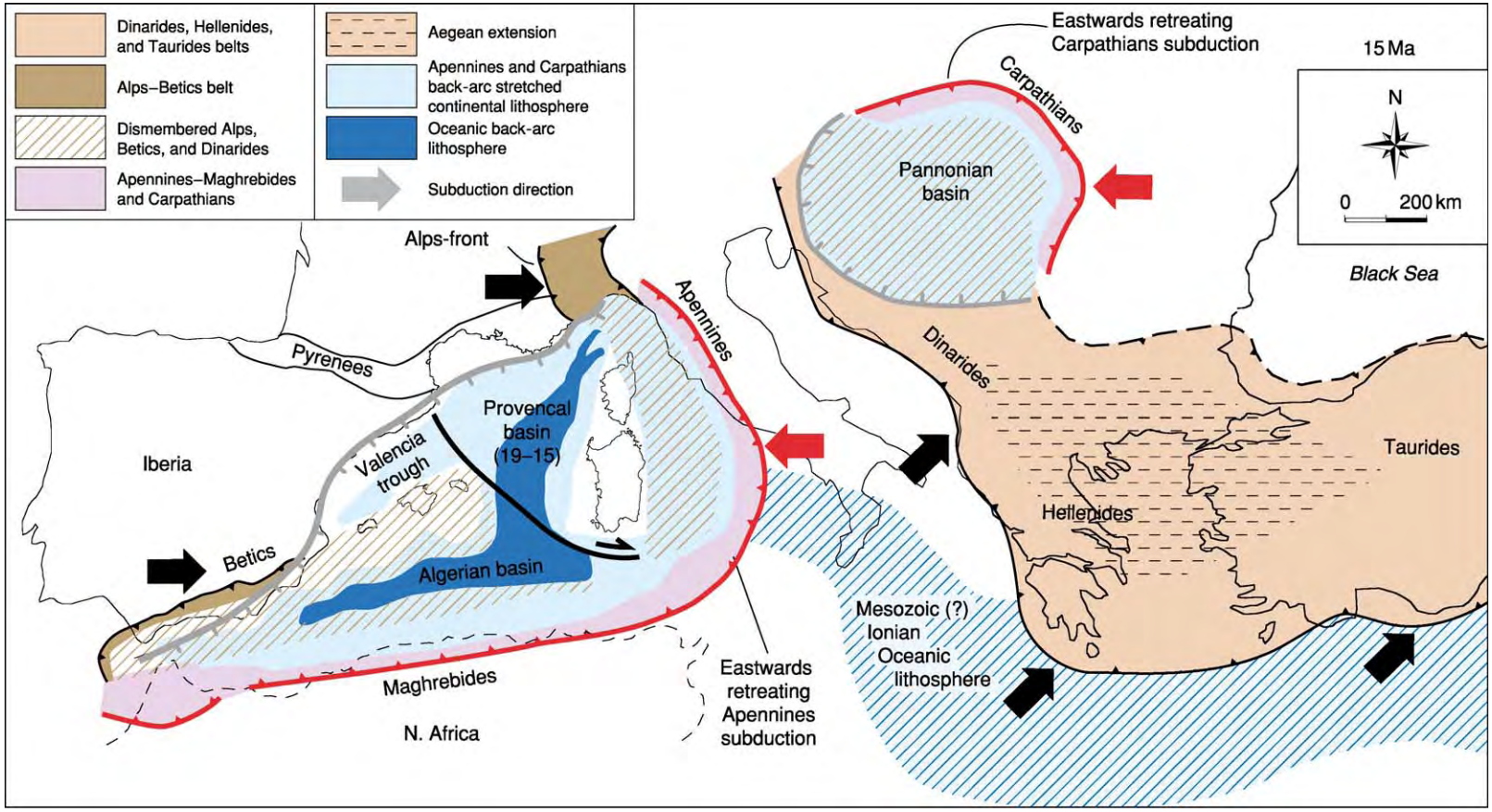


Figure 2 Palaeogeodynamics at about 15 Ma. Note the 'eastward' vergence of both the Apennines Maghrebides trench and the back arc extensional wave. The Liguro Provençal basin, the Valencia trough, and the North Algerian basin were almost completely opened at 10 Ma. The Dinarides subduction slowed down, owing to the presence of the thick Adriatic continental lithosphere to the west, whereas to the south the Hellenic subduction was very lively owing to the presence in the footwall plate of the Ionian oceanic lithosphere. The Carpathians migrated eastwards, generating the Pannonian back arc basin, with kinematics similar to those of the Apennines. Provencal basin (19 15) Age of the oceanic crust.

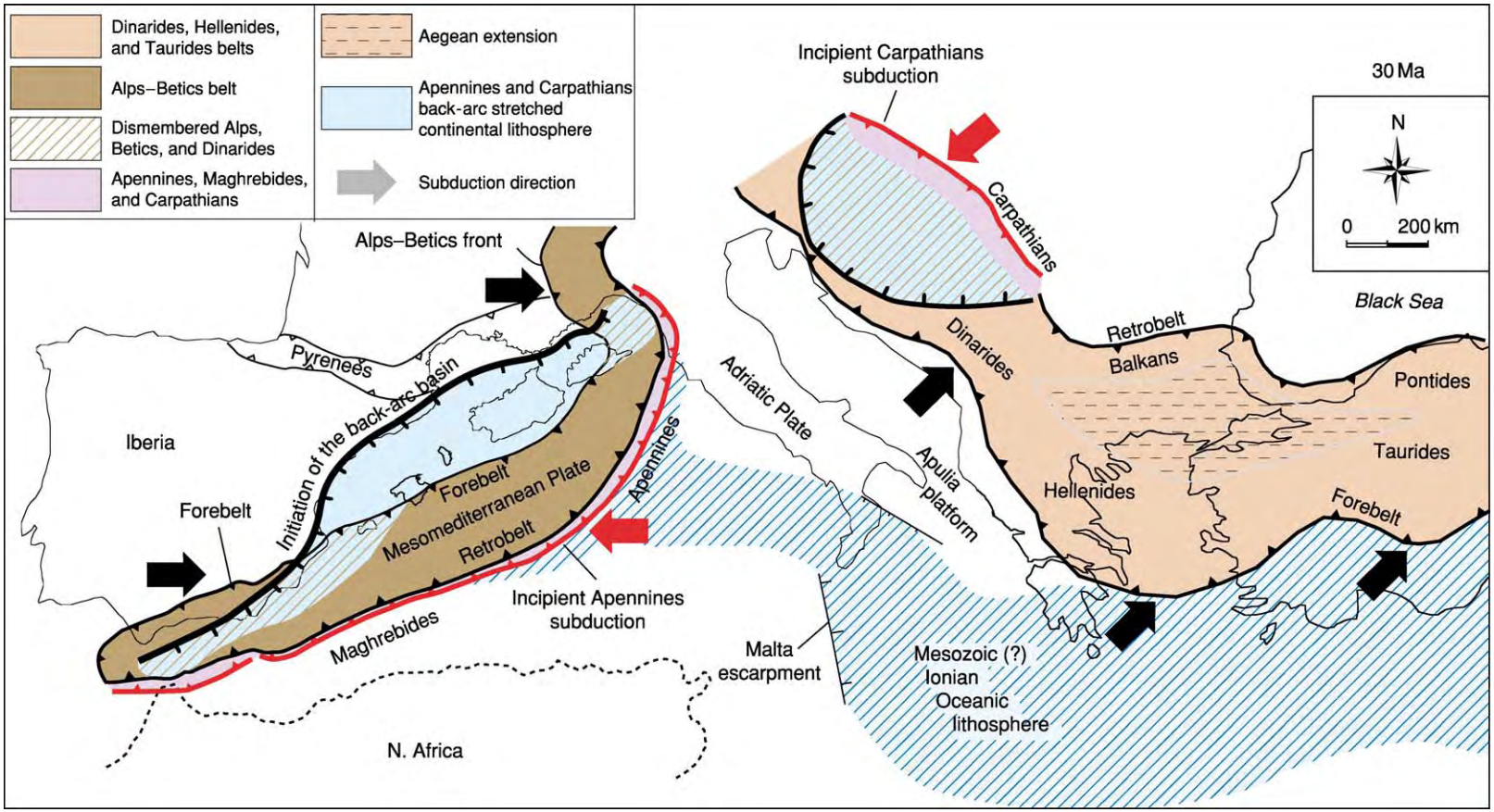


Figure 3 Palaeogeodynamics at about 30 Ma. The locations of the subduction zones were controlled by the Mesozoic palaeogeography. The Alps Betics formed along the south eastwards dipping subduction of Europe and Iberia underneath the Adriatic and Mesomediterranean plates. The Apennines developed along the Alps Betics retrobelt to the east, in which oceanic or thinned pre existing continental lithosphere was present. Similarly, the Carpathians started to develop along the Dinarides retrobelt (i.e. the Balkans). The fronts of the Alps Betics orogen were cross cut by the Apennines related subduction back arc extension.

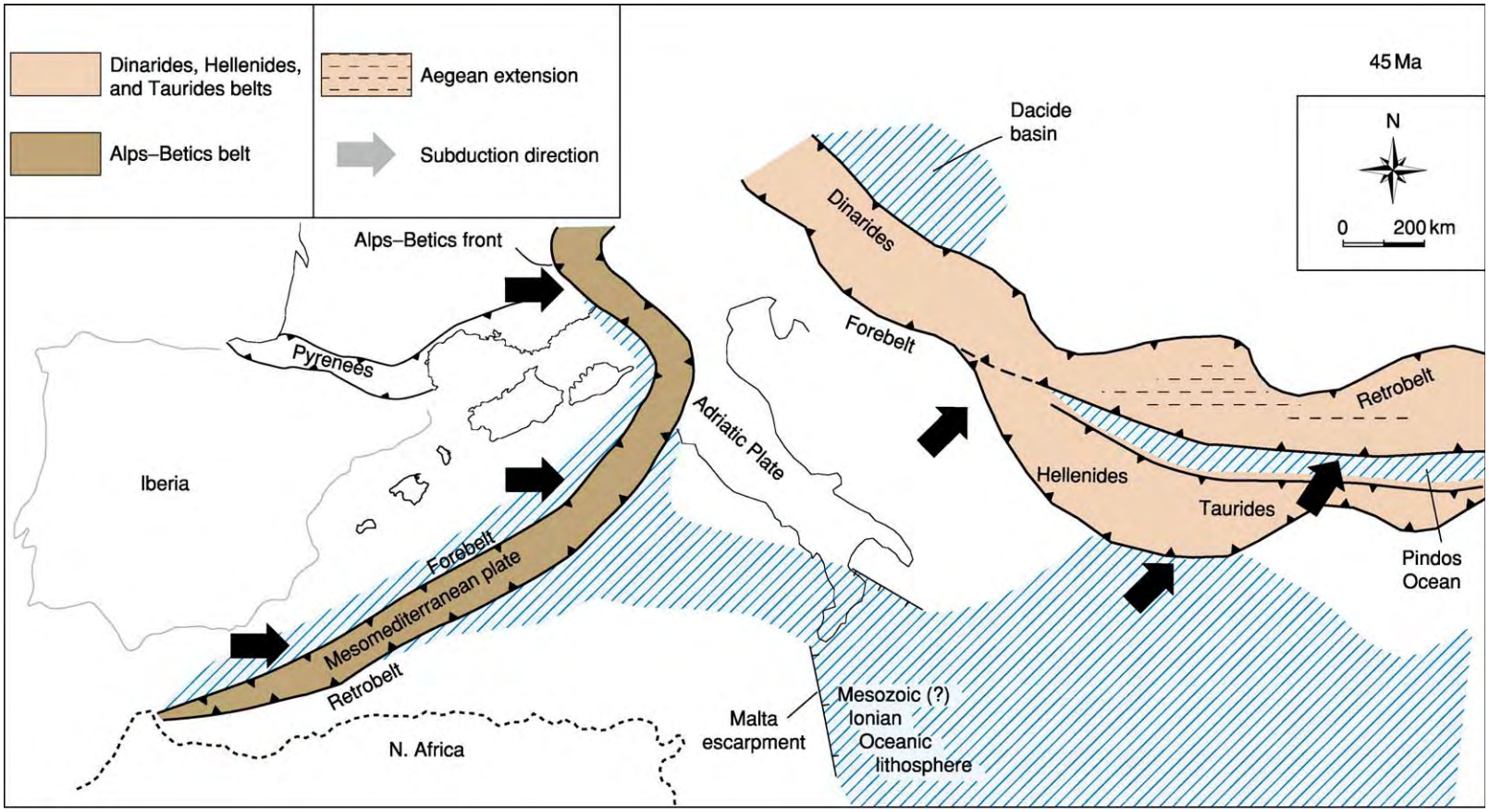


Figure 4 Palaeogeodynamics at about 45 Ma. The Alps were continuous with the Betics to Gibraltar, consuming an ocean located to the west.

(Maghrebides) and its Sicilian continuation, whereas left-lateral transtension occurs along the same trend in the back-arc setting just to the north of the African margin. The opposite tectonic setting is found in the northern margin of the arc.

Subduction retreat generated calc-alkaline and shoshonitic magmatic episodes – particularly in the western margins of the lithospheric boudins – which were followed by alkaline-tholeiitic magmatism in the back-arc to the west.

Extension partly originated in areas previously occupied by the Alps–Betics Orogen, which formed in the Cretaceous due to the 'eastwards'-directed subduction of Europe and Iberia underneath the Adriatic Plate and a hypothetical Mesomediterranean Plate (Figure 4). If Sardinia is restored to its position prior to rotation, it can be seen that during the Early Cenozoic the Alps were probably joined with the Betics in a double-vergent single belt. The western Alps, which are the forebelt of the Alps, were connected to the

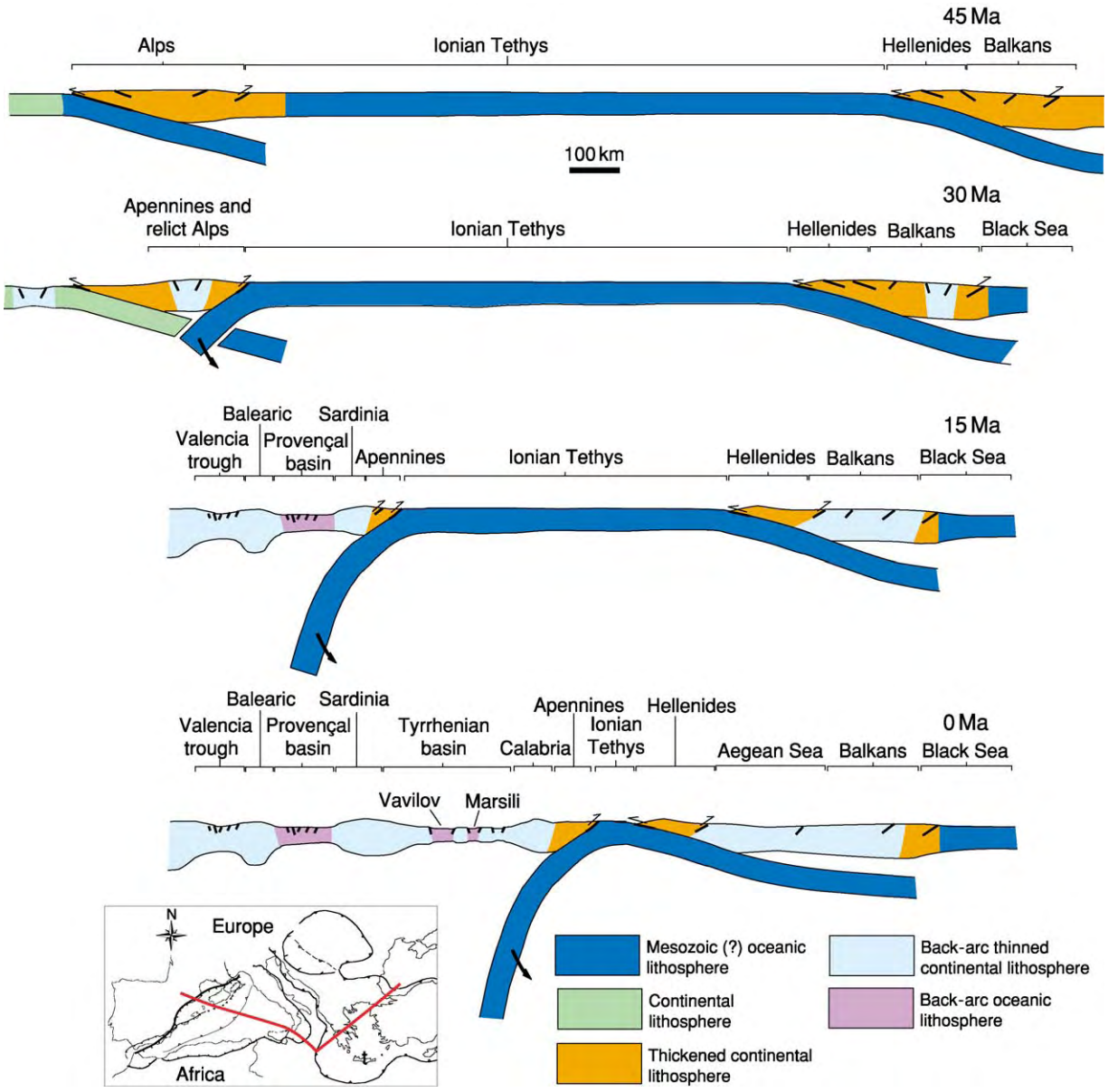


Figure 5 During the last 45 Ma, the evolution of the Mediterranean along the trace shown on the map (inset) is the result of three main subduction zones: the early eastwards directed Alpine subduction; the Apennines subduction switch along the Alps retrobelt; and the Dinarides Hellenides subduction. The last two slabs retreated at the expense of the inherited Tethyan Mesozoic oceanic or thinned continental lithosphere. In their hanging walls, a few rifts formed as back arc basins, which are progressively younger towards the subduction hinges. The slab is steeper underneath the Apennines, possibly owing to the westwards drift of the lithosphere relative to the mantle.

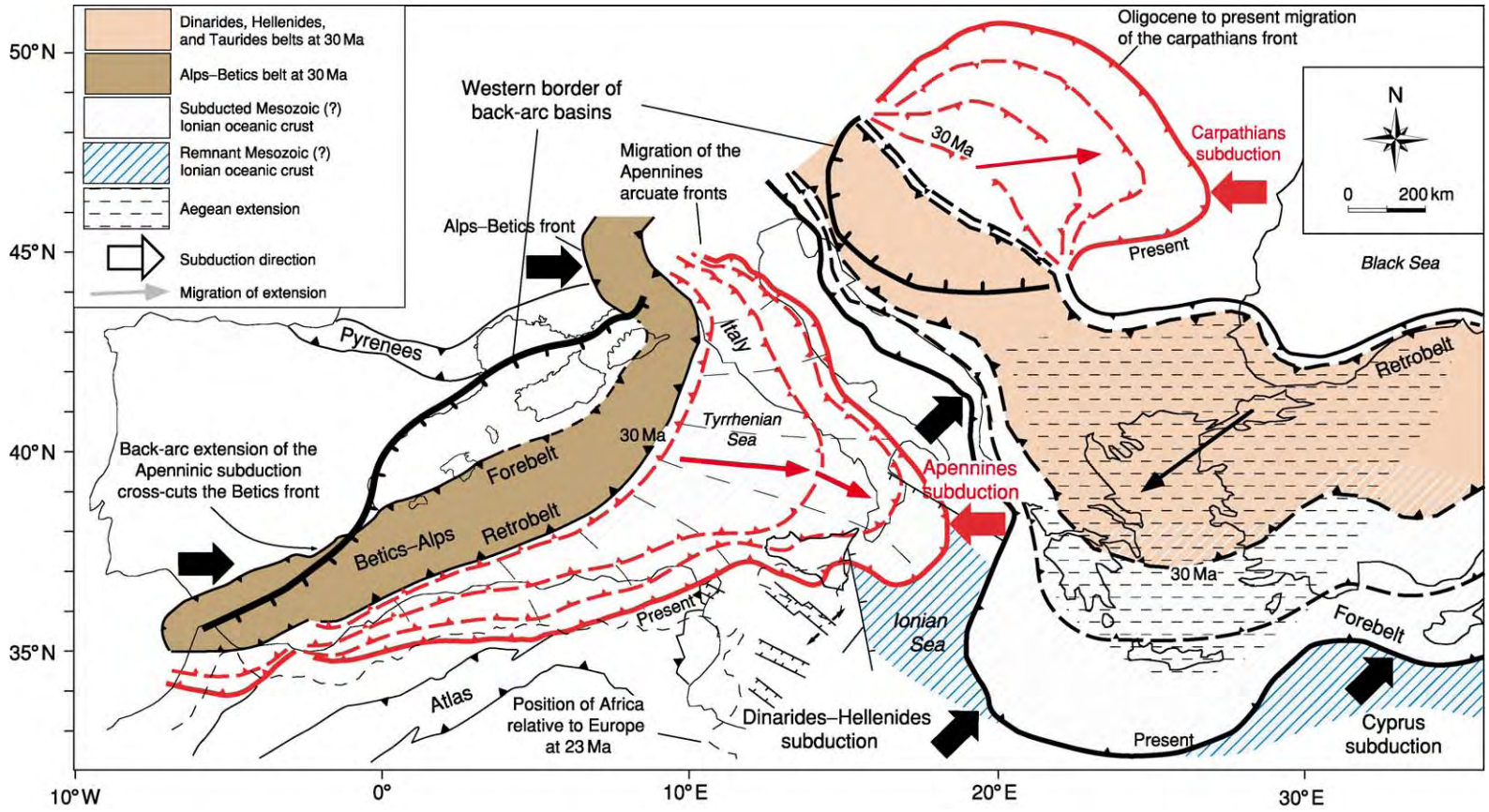


Figure 6 Main tectonic features of the Mediterranean realm, which has been shaped during the last 45 Ma by a number of subduction zones and related belts: the double vergent Alps Betics; the single eastwards vergent Apennines Maghrebides and the related western Mediterranean back arc basin; the double vergent Dinarides Hellenides Taurides and related Aegean extension; the single eastwards vergent Carpathians and the related Pannonian back arc basin; and the double vergent Pyrenees.

Alpine Corsica; the Alps continued south-westwards into the Balearic promontory and the Betics. The retrobelt of the Alps, the southern Alps, also continued from northern Italy towards the south-west. In a double-vergent orogen, the forebelt is the frontal part, which is synthetic to the subduction and verges towards the subducting plate; the retrobelt is the internal part, which is antithetic to the subduction and verges towards the interior of the overriding plate.

The westwards-directed Apennines–Maghrebides subduction started along the Alps–Betics retrobelt (**Figures 3 and 5**), where oceanic and thinned continental lithosphere occurred in the foreland to the east. Subduction underneath the Apennines–Maghrebides consumed inherited Tethyan domains (**Figures 5 and 6**). The subduction zone and the related arc migrated 'eastwards' at a speed of 25–30 mm year ^{1}.

The western Late Oligocene–Early Miocene basins of the Mediterranean nucleated both within the Betics orogen (e.g. the Alboran Sea) and in its foreland (e.g. the Valencia and Provençal troughs; **Figure 3**). At that time the direction of the grabens (40°–70°) was oblique to the trend of the coexisting Betics orogen (60°–80°), indicating its structural independence from the Betics Orogeny. Thus, as the extension cross-cut the orogen and also developed well outside the thrust-belt front, the westernmost basins of the Mediterranean developed independently of the Alps–Betics orogen, being related instead to the innermost early phases of back-arc extension in the hanging wall of the Apennines–Maghrebides subduction zone. In contrast to the 'eastwards'-migrating extensional basins and following the 'eastwards' retreat of the Apennines subduction zone, the Betics–Balearic thrust front was migrating 'westwards', producing interference or inversion structures.

The part of the Alps–Betics orogen that was located in the area of the Apennines–Maghrebides back-arc basin (**Figure 1**) has been disarticulated and spread out into the western Mediterranean (forming the metamorphic slices of Kabylie in northern Algeria and Calabria in southern Italy). Alpine type basement rocks have been dragged up in the Tyrrhenian Sea.

Similarly, boudinage of the pre-existing Alps and Dinarides orogens occurred in the Pannonian Basin, which is the Oligocene to Recent back-arc basin related to the eastwards-retreating westwards-directed Carpathian subduction zone (**Figures 1, 3, and 6**). In the Pannonian basin, the extension isolated boudins of continental lithosphere that had been thickened by the earlier Dinarides orogen, such as the Apuseni Mountains, which separate the Pannonian basin from the Transylvanian basin to the east. The western Mediterranean back-arc setting is comparable with Atlantic and western Pacific back-arc basins that show similar large-scale lithospheric boudinage, in which parts of earlier orogens have been scattered in the back-arc area, like the Central America Cordillera relicts that are dispersed in the Caribbean domain.

The Apennines accretionary prism formed in sequence at the front of the pre-existing Alpine retrobelt, and, therefore, the central western Apennines also contain the inherited Alpine orogen of Cretaceous to Miocene age. There was probably a temporary coexistence of opposite subductions during the Late Oligocene to Early Miocene (**Figure 5**). Structural and geophysical data support the presence of an eastwards-migrating asthenospheric wedge at the subduction hinge of the retreating Adriatic Plate. The subduction flip, from the Alpine eastwards-directed subduction to the Apennines westwards-directed subduction, could be reflected in the drastic increase in subsidence rates in the Apennines foredeep during the Late Oligocene to Early Miocene. Westwards-directed subduction zones, such as the Apennines, show foredeep subsidence rates that are up to 10 times higher (more than 1 mm year ^{1}) than those of the Alpine foredeeps. The subduction flip (**Figure 5**) could also be reflected in the larger involvement of the crust during the earlier Alpine stages than in the Apennines décollements, which mainly deformed the sedimentary cover and the phyllitic basement. It has been demonstrated that the load of the Apennine and Carpathian orogens is not sufficient to generate the 4–8 km deep Pliocene–Pleistocene foredeep basins, and a mantle origin has been proposed for the mechanism (slab pull and/or eastwards mantle flow).

Paradoxically, the extension that determined most of the western Mediterranean developed in the context of relative convergence between Africa and Europe. However, it appears that the north–south relative motion between Africa and Europe at the longitude of Tunisia has been about 135 km in the last 23 Ma, more than five times slower than the migration of the Apennines arc, which has moved more than 700 km eastwards during the last 23 Ma (**Figures 1 and 6**). Therefore, the eastwards migration of the Apennines–Maghrebides arc is not a consequence of the north–south relative convergence between Africa and Europe but is instead a consequence of the Apennines–Maghrebides subduction rollback, which was generated either by slab pull or by the 'eastwards' flow of the mantle relative to the lithosphere deduced from the hotspot reference frame.

The western Mediterranean developed mainly after the terminal convergence in the Pyrenees at about 20 Ma, which resulted from the Late Cretaceous to Early Tertiary counterclockwise rotation of Iberia, which was contemporaneous with the opening of the Biscay Basin.

In northern Africa, south of the Maghrebides (and the related Algerian Tell and Moroccan Riff), the Atlas Mountains represent an intraplate inversion structure, in which extensional (north-north-east-trending) and left-lateral (about east–west-trending) transtensional Mesozoic intercontinental rifts were later buckled and squeezed by Cenozoic compression and right-lateral transpression in the foreland of the Apennines–Maghrebides subduction zone. This is also indicated by the Mesozoic sequences in the Atlas ranges, which are thicker than the adjacent undeformed *mesetas*.

Central Mediterranean

The Malta escarpment (Figures 3 and 4), along the eastern coast of Sicily, is a physiographic feature that has been tectonically controlled since Triassic times. Rocks dredged from the Malta escarpment range from Mesozoic to Tertiary in age. The escarpment represents a Mesozoic continental margin that has been reactivated as a transtensional feature since the Pliocene. In spite of the Apennines and Hellenides Neogene subduction zones, two conjugate passive continental margins are preserved at the margins of the Ionian Sea, along the Malta escarpment to the south-west and the Apulian escarpment to the north-east. Based on the low heat flows (18–40 $mW\,m^{-2}$) and the 4–8 km of sedimentary cover, the Ionian Sea is probably a remnant of the Mesozoic Tethys Ocean, confined by the two conjugate passive continental margins. The transition from continental crust to oceanic crust appears to be sharper to the north-east than to the south-west. The basin between south-east Sicily and south-west Puglia was about 330 km wide. The inferred oceanic ridge could have been flattened by thermal cooling and buried by later sediments.

Stratigraphic and structural constraints to the north in the Apennines belt suggest that the Ionian Ocean continued to the north-west (Figure 5). This palaeogeography is supported by the seismicity of the Apennines slab underneath the southern Tyrrhenian Sea, which implies subducted oceanic lithosphere. The adjacent absence or paucity of deep seismicity does not imply the absence of subduction but can be interpreted as a reflection of the more ductile behaviour of the subducted continental lithosphere.

The Sicily Channel and the Pelagian shelf off the coast of eastern Tunisia have been undergoing extension since at least Pliocene times; in other words Africa is moving south-westwards in relation to Sicily (Figure 1). This process is responsible for the two grabens of Pantelleria and Malta deepening the seafloor and for the generation of active alkaline magmatism (e.g. the ephemeral Ferdinandea Island). The lithospheric extension was active whilst the Apennines–Maghrebides accretionary prism advanced, generating an interplay of two tectonic settings working together, with thrusts advancing over an orthogonal extending area, generating both thrusts cutting normal faults and normal faults offsetting thrusts. The rifting of the Sicily Channel seems to be physically connected north-westwards to the rift in south-western Sardinia (Campidano graben) and south-eastwards to the Sirte Basin, off the coast of Libya. One possibility is that this rift is linked through transfer zones in Egypt to the Red Sea and the East African Rift.

Eastern Mediterranean

The Dinarides, Hellenides, and Taurides are a polyphase orogen, representing the coalescence of at least two or three subduction zones since Mesozoic times (Figures 1, 4, 5, and 6). The orogen has a part synthetic to the north-eastwards-directed subduction, i.e. the forebelt verging south-westwards. The conjugate part of the orogen is the retrobelt, which verges north-eastwards and northwards (Balkans and Pontides). The existence of three subduction zones is supported by the occurrence of two distinct oceanic sutures, preserved as the ophiolitic suites of Vardar and the Sub-Pelagonian units, which represent two separated branches of the Mesozoic Tethyan Ocean and the present oceanic subduction of the Ionian Sea. It is commonly believed that the more internal (Vardar) suture zone is the older one.

The polyphase orogen exhibits a similar architecture to the Alps, but duplicated. The Rhodope–Serbo-Macedonian and Sakarya (northern Turkey) massifs mimic the internal massifs of the Alps, which represent the continental margin of the hanging-wall plate. On the other side, to the south-west of the Vardar oceanic suture, the Pelagonian (Macedonia–Greece) and Menderes (northern Turkey) massifs correspond to the external massifs of the Alps, representing the continental lithosphere of the footwall plate. The Pelagonian basement is at the same time the hanging-wall plate for the more external north-eastwards-directed subduction of the Sub-Pelagonian and Pindos Ocean, which was eventually closed by collision with the eastern margin of the Adriatic Plate.

However, unlike the Alps, widespread extension developed in the Dinarides–Hellenides–Taurides orogen (Figures 1 and 6). This extension resulted in the low topography of the orogen in comparison with belts such as the Alps and the Zagros or the Himalayas. In the Balkans, the Rhodope, and the Serbo-Macedonian massifs, structural and stratigraphic data indicate an interplay of compressional and

extensional tectonics. A Cretaceous to Eocene compressive deformation was followed by the generation of Eocene grabens. A later (possibly Miocene) compression inverted and uplifted these grabens, but it was followed by extensional tectonics that have affected the Balkan peninsula since Pliocene times, determining the north-west-trending normal faults and the related east–west right-lateral and north–south left-lateral transtensive transfer faults. North-eastwards-directed subduction is continuing along the eastern side of the Adriatic, in the Ionian Sea underneath the Mediterranean Ridge (the accretionary prism), and on the northern side of the Levantine Sea, i.e. in the eastern Mediterranean beneath Cyprus (Figure 1). The convergence rates are faster underneath the Mediterranean ridge (up to 40–50 mm year ^{1}), decrease eastwards along the Cyprus segment, and have minimum values along the Adriatic coast. The convergence rate appears to be controlled by the composition of the foreland lithosphere: where it is oceanic and dense, such as in the Ionian Sea, the subduction is faster than in the Adriatic and Cyprus segments, where the downgoing lithosphere is continental and transitional oceanic–continental, respectively. In the orogen, calc-alkaline and shoshonitic magmatism has accompanied most of the subduction since Cretaceous times. The later extensional process in the anomalously called 'back-arc' is possibly responsible for the transition to the alkaline magmatic signature.

One of the best-known ophiolitic sequences in the world crops out in Cyprus: a complete oceanic section is exposed (from harzburgites and peridotites of the upper mantle to gabbros, sheeted dykes, lavas, and pelagic sediments of the crust). The island is an anticline involving the whole crust, and its culmination coincides with the Erathostene seamount in the subducting foreland. The Erathostene seamount is a structural high inherited from the Mesozoic–Cenozoic rift.

Since at least Miocene times, there has been an independent and presently active subduction along the northern margin of the Black Sea, generating the Caucasus.

Geodynamic reconstructions of the eastern Mediterranean explain the extensional tectonics either by westwards Anatolian extrusion or by gravitational collapse of thickened lithosphere. However, these mechanisms can be ruled out because plate-velocity vectors increase from eastern Anatolia to the Aegean and Greece. This contradicts the basic rule that the velocity field decreases away from the source of the energy, i.e. the supposed squeezing of Anatolia by the Arabia indenter, or the collapse of the Anatolian orogen. Moreover, the topographic gradient between Anatolia and the Ionian deep basin is too small (less than 1°) to provide sufficient energy to explain the present deformation. Instead, the simplistic view of the westward Anatolian escape would close the Aegean Sea.

The plates involved in the geodynamic reconstructions of the eastern Mediterranean are Africa, Greece, Anatolia, Eurasia, and Arabia. Deformation is very active in all these areas. The most prominent geodynamic factor shaping the eastern Mediterranean is the north-east-directed subduction of Africa underneath Greece and the Anatolian Plate (Eurasia). Seismic lines across the Cyprus Arc at the southern margin of the Anatolian Plate show clear active compression and deformation of the seafloor.

The Aegean Sea is generally considered to be a back-arc basin resulting from the aforementioned subduction. However, the Aegean Sea is characterized by a relatively thick crust (20–25 km) in spite of long-standing subduction, which has probably been active since at least the Cretaceous. The subduction zone migrated south-westwards to the present position of the Cyprus-Hellenic subduction zone, and the associated orogen was later replaced by extension. In the Aegean Sea, Alpine-type crustal thickening with high pressures and low temperatures was followed by non-coaxial crustal-scale extension. This is consistent with the initial emplacement of thrust-sheets of basement slices, which were later cross-cut by extensional or transtensional faults. In addition, extension and associated magmatism were and are migrating south-south-westwards, and have developed particularly since the Oligocene, while subduction began much earlier. 'Normal' back-arc basins (e.g. the Tyrrhenian Sea) associated with westwards-directed subduction zones opened very fast (10–20 Ma) and are always contemporaneous with the subduction. Moreover, they are characterized by oceanization and eastwards migration of extension and related magmatism, features directly surrounded by a frontal accretionary wedge. In contrast, the accretionary wedge of the Hellenic subduction zone is the south-eastern prolongation of the Dinarides thrust belt, where no back-arc rift comparable to the Tyrrhenian Sea occurs.

The extension in western Turkey, the Aegean Sea, Greece, and Bulgaria appears to be the result of differential convergence rates in the north-eastwards-directed subduction of Africa relative to the hanging wall of disrupted Eurasian lithosphere. Relative to Africa, the faster south-eastwards motion of Greece than of Cyprus–Anatolia results in the Aegean extension. The differences in velocity can be ascribed to differential decoupling with the asthenosphere. In the back-arc basins of the western Pacific the asthenosphere replaces a subducted and retreated

slab; however, the Aegean rift represents a different type of extension associated with a subduction zone, in which the hanging-wall plate overrides the slab at different velocities, implying internal deformation.

According to this geodynamic scenario, during the compressive events associated with north-eastwards-directed subduction, basement rocks (both continental and ophiolitic slices) in western Anatolia and the Aegean Sea were uplifted and eroded. Later extension caused subsidence in the area, and the basement slices were partly covered by continental and marine sediments.

During its development, the Aegean extension migrated south-westwards (**Figures 5 and 6**). The Aegean rift affects the Aegean Sea and all of continental Greece, and it can be followed to the east, where it is widely expressed in Turkey, and to the north-west in Bulgaria, Albania, Macedonia, Serbia, and Bosnia. At the same time, from the Oligocene to the present, to the north, the Pannonian basin developed as the back-arc of the Carpathians subduction, but migrating eastwards, and affecting mainly eastern Austria, Slovenia, Croatia, Hungary, and Romania. Therefore, in the central part of the former Yugoslavia, the Pannonian and Aegean rifts meet with opposite directions of migration.

See Also

Europe: Variscan Orogeny; Permian to Recent Evolution; The Alps; Holocene. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mountain Building and Orogeny.

Further Reading

Berckhemer H and Hsü KJ (eds.) (1982) *Alpine Mediterranean Geodynamics*. Geodynamics Series 7. Washington: American Geophysical Union.

Calcagnile G and Panza GF (1980) The main characteristics of the lithosphere asthenosphere system in Italy and surrounding regions. *Pure and Applied Geophysics* 119: 865 879.

Carminati E, Wortel MJR, Spakman W, and Sabadini R (1998) The role of slab detachment processes in the opening of the western central Mediterranean basins: some geological and geophysical evidence. *Earth and Planetary Science Letters* 160: 651 665.

Catalano R, Doglioni C, and Merlini S (2001) On the Mesozoic Ionian basin. *Geophysical Journal International* 144: 49 64.

Cella F, Fedi M, Florio G, and Rapolla A (1998) Gravity modeling of the litho asthenosphere system in the Central Mediterranean. *Tectonophysics* 287: 117 138.

Christova C and Nikolova SB (1993) The Aegean region: deep structures and seismological properties. *Geophysical Journal International* 115: 635 653.

Dercourt J, Gaetani M, Vrielynck B, *et al.* (2000) *Atlas Peri Tethys, Paleogeographical Maps*. Geological Map of the World. Paris: CGMW.

de Voogd B, Truffert C, Chamot Rooke N, *et al.* (1992) Two ship deep seismic soundings in the basins of the Eastern Mediterranean Sea (Pasiphae cruise). *Geophysical Journal International* 109: 536 552.

Doglioni C, Gueguen E, Harabaglia P, and Mongelli F (1999) On the origin of W directed subduction zones and applications to the western Mediterranean. In: Durand B, Jolivet J, Horváth F, and Séranne M (eds.) *The Mediterranean Basins: Tertiary Extension Within The Alpine Orogen*, pp. 541 561. Special Publication 156. London: Geological Society.

Durand B, Jolivet J, Horváth F, and Séranne M (1999) *The Mediterranean Basins: Tertiary Extension Within the Alpine Orogen*. Special Publication 156. London: Geological Society.

Frizon de Lamotte D, Saint Bezar B, Bracene R, and Mercier E (2000) The two main steps of the Atlas building and geodynamics of the western Mediterranean. *Tectonics* 19: 740 761.

Gueguen E, Doglioni C, and Fernandez M (1998) On the post 25 Ma geodynamic evolution of the western Mediterranean. *Tectonophysics* 298: 259 269.

Guerrera F, Martin Algarra A, and Perrone V (1993) Late Oligocene Miocene syn / late orogenic successions in western and central Mediterranean chain from the Betic cordillera to the southern Apennines. *Terra Nova* 5: 525 544.

Huguen C, Mascle J, Chaumillon E, *et al.* (2001) Deformational styles of the eastern Mediterranean Ridge and surroundings from combined swath mapping and seismic reflection profiling. *Tectonophysics* 343: 21 47.

Kastens K, Mascle J, Auroux C, *et al.* (1988) ODP Leg 107 in the Tyrrhenian Sea: insights into passive margin and back arc basin evolution. *Geological Society of America Bulletin* 100: 1140 1156.

Réhault JP, Mascle J, and Boillot G (1984) Evolution géodynamique de la Méditerranée depuis l'Oligocène. *Memorie Società Geologica Italiana* 27: 85 96.

Robertson AHF and Grasso M (1995) Overview of the Late Tertiary Recent tectonic and palaeo environmental development of the Mediterranean region. *Terra Nova* 7: 114 127.

Stampfli G, Borel G, Cavazza W, Mosar J, and Ziegler PA (2001) *The Paleotectonic Atlas of the PeriTethyan Domain*. CD ROM. European Geophysical Society.

Stanley DJ and Wezel FC (eds.) (1985) *Geological Evolution of the Mediterranean Basin*. New York, USA: Springer Verlag.

Vai GB and Martini P (eds.) (2001) *Anatomy of an Orogen: the Apennines and Adjacent Mediterranean Basins*. Dordrecht: Kluwer Academic Publishers.

Wilson M and Bianchini G (1999) Tertiary Quaternary magmatism within the Mediterranean and surrounding regions. In: Durand B, Jolivet J, Horváth F, and Séranne M (eds.) *The Mediterranean Basins: Tertiary Extension Within The Alpine Orogen*, pp. 141 168. Special Publication 156. London: Geological Society.

Holocene

W Lemke and J Harff, Baltic Sea Research Institute Warnemünde, Rostock, Germany

Introduction

The Quaternary period comprises the shortest time interval of all geological systems. Compared to the preceding climatically stable and warm Tertiary, it is characterised by a multiple alternation of large-scale glaciations and short warm intervals in between. The latest interglacial period, which is still ongoing, is called Holocene after the Greek words 'holos' (entire) and 'ceno' (new). According to recent understanding, it began *ca*. 11 600 calendar years before present. In contrast to other epochs of the Earth's history, it is not defined and subdivided by certain floral or faunal assemblages but by climatic features. Another basic difference from former geological periods is the increasing human impact on the geosphere. In fact, some authors claim that the properties of the 'system Earth' have changed by human influence to an extent that it cannot be called natural anymore. Reconstructing the geological past, therefore, requires consideration of natural processes, as well as the results of human activity and to separate them from each other. Thus, Holocene geology is intensely interrelated not only with other natural sciences but also with human history, archaeology, and further social sciences. This adds a wealth of additional information to the data stored within geological archives. On the other hand, geological problems during the Holocene are not only a matter of actualism in the classical sense anymore. Due to the increasingly closer connection between geological processes and the development of the human society, forecasting of geological trends becomes more and more important. In this way, Charles Lyell's (*see* **Famous Geologists:** Lyell) statement about the principle of actualism could be extended to: "The knowledge about present and past is the key to the future."

Dating

When aiming for an accurate reconstruction of the geological past, dating becomes an essential issue. Looking back from recent times to the near past, dating of geological events is simply done on a high resolution by analysing the written historical archives. Further back in time, indirect methods (by using so-called proxy data) have to be used. Proxy data with a yearly resolution are related to processes which result in persistent and regularly successive yearly structures within sediments (e.g., varve sequences) or organic material like wood (dendrochronology). The latter is based on the study of tree ring patterns which are controlled mainly by climatic factors. In Europe it was used particularly for oaks in central and western Europe and for pines in northern Europe. Regionally generalised curves for these two tree species cover nearly all of the European Holocene.

Dendrochronological dates are highly valuable for calibrating dating results produced by other methods. This refers especially to isotopic dating by radiocarbon, which is widely used as a standard method for the dating of organic material within the Holocene. By comparing dendrochronological or varve counting dates with radiocarbon dating, inconsistencies within the later ones, particularly within the early Holocene, became obvious. Therefore, when looking at dates in the literature, it is crucial to consider if calibrated (calendar) years or radiocarbon years are referred to. Within this article calendar years before present (BP) are used (except for Figure 10). Other short-lived isotopes, such as ^{210}Pb are used to date processes and events in the more recent past on time-scales of decades and centuries.

Once the environmental history of a specific region is well known, assemblages of plants or animals might also help to assess the age of the deposits they are found in.

Climate

One of the most intensively studied subjects of Holocene development is climate. Ice and marine sediment cores have been used to assess climatic changes and they provide smoothed background data to more regional, or local and mostly more dramatic, climatic variations on the European continent. The onset of the Holocene is marked by a global drastic temperature increase of about 7°C at the end of the Younger Dryas, about 11 600 calendar years BP. This climate reorganisation happened during a period of not more than a few decades. Since then, the Holocene climate has been stable by comparison with the preceding glacial period. Nevertheless, minor climate fluctuations have been reconstructed. Several periods with cooler and warmer temperatures than the last century have left their traces in the geological and biological archives and also in human history (Figure 1).

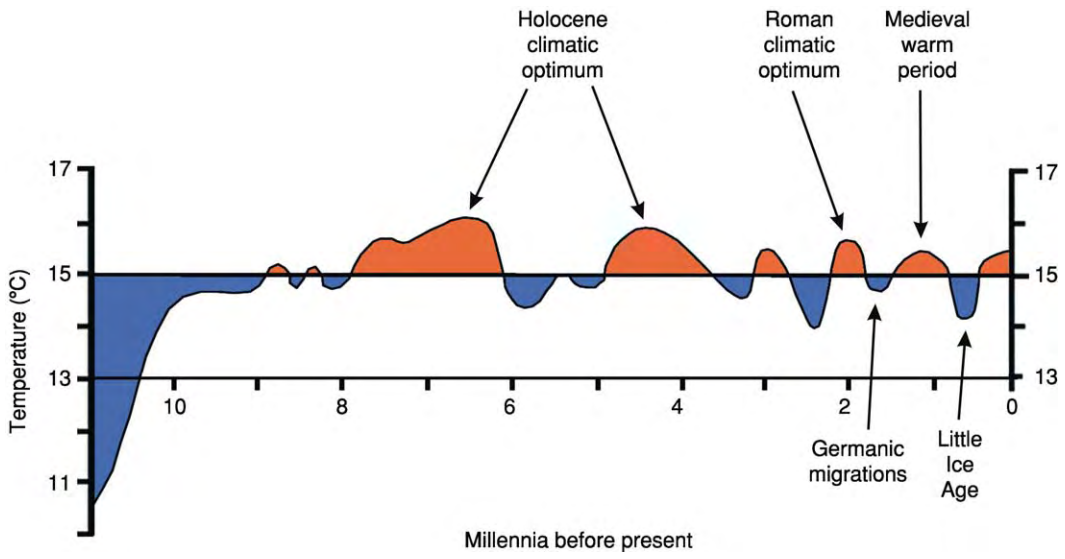


Figure 1 Mean Holocene annual temperatures. The horizontal axis is time before present (BP) in thousands of years (Ka). (Adapted from Schonwiese C (1995) *Klimaänderungen Daten, Analysen, Prognosen.* Berlin: Springer Verlag, with permission.)

During the first thousand years of the Holocene, the climate was possibly slightly cooler than today. Between 9000 and 8200 years BP, climatic conditions were slightly warmer and often moister than today. Cooler conditions throughout Europe have been interpreted from different proxy data for the period around 8200 years BP, when annual mean temperatures dropped by 2°C in central Europe and the Alpine timberline fell by about 200 m. This sudden cool phase lasted for about 200 years and wetter-than-present conditions in north-west Europe have been inferred.

Warmer temperatures prevailed in Europe during the so-called climatic optimum in the Early Holocene (8000-4500 years BP). The Early Holocene climatic optimum was characterised by warmer summers than today in Europe. For astronomical reasons, the northern hemisphere received nearly 8% more solar radiation during summers than in recent times. A northward shift of the inner tropical convergence zone (ITCZ) forced monsoonal rainfall as far north as the Mediterranean Basin. At around 5900 years BP, a short cold episode interrupted this warm phase.

Since 4500 years BP, the climatic conditions have fluctuated around a situation comparable to the recent one. Remarkable deviations occurred at about 3300 years BP, when intensified glaciation started in the Alps. About 2000 years ago BP, favourable climatic conditions promoted the development of the Roman Empire, while the Germanic migrations after its collapse went along with a cooler climate. In mediaeval times (*ca.* 1300–700 years BP), another warm period allowed the expansion of Scandinavian Vikings as far as Greenland and North America. Clear indications of warm temperatures during this time interval were also reported from northern Russia, central Europe, and the Mediterranean. The youngest climatic deterioration, known as the Little Ice Age, at about 700–150 years BP, destroyed the agricultural economic basis of the Norse settlers in Greenland, and by about 500 years BP, their population in Greenland had vanished. In the mid-seventeenth century, glaciers in the Swiss Alps advanced and rivers in England and the Netherlands often froze over during the winter. Severe cold winters have been deduced from borehole data in the Czech Republic for the time slice between 300 and 400 years BP, too. Since the mid-nineteenth century, the global temperatures have risen (Figure 1) again, a process which is still going on.

A general periodicity of 200 to 600 years for the whole Holocene climate can be inferred from various proxies. External processes including solar activity cycles and internal driving forces as volcanic eruptions are under debate as controlling factors of climate variability up until now. To what extent this climatic cyclicity is modified by human activity (e.g., extensive release of greenhouse gases) is a matter of current scientific discussion.

Naturally Changing Holocene Landscapes in Europe

The deglaciation processes, which had started within the Late Pleistocene, were accelerated in the Early Holocene. Within the first two thousand years of the

Holocene, the former prevailing tundra and steppe habitats (Figure 2) were nearly completely replaced by mixed deciduous forest (Figure 3). The first phase of tree invasion was characterised by birch and pine and later by hazel and elm. The local tree assemblages could differ considerably from each other and also from recent compositions. At around 10 000 years BP, in many parts of Europe the forest cover was still rather more open than at present, with more herbaceous glades. By 9000 years BP, the forest had become closed, but with conifers more abundant than at present in eastern Europe. Until that time, deciduous trees such as oaks or hornbeam were predominant or abundant, even in southern Europe (Figure 3). The typical recent Mediterranean vegetation with evergreen trees and shrubs started to develop after that time.

During the climatic optimum, thermophile plants and animals extended further to the north. Possibly due to elm disease, a drastic decline in the number of elm pollen is observed in the geological records throughout Europe at about 5800 years BP.

Since 4500 years BP, there has been an increasing human influence on the faunal and floral elements of the European ecosystem. Figure 4, showing the potential present vegetation without human influence, is obviously different from the coverage conditions today.

Another result of the changing Holocene climate was a rising global sea-level by meltwater supply and thermal expansion of sea-water (eustatic sea-level rise). During the maximum Weichselian glaciation, the global sea-level was about 125 metres deeper than today. In the Early Holocene, large deglaciated areas started to uplift because of the vanishing ice load (glacio-isostasy). At the centre of the last glaciation, around the Bothnian Bay, an uplift of more than 280 m is recorded within Holocene sediments. This isostatic uplift was compensated by subsidence in more distal regions (Figure 5) within the southern Baltic Basin. By some authors, this process is assumed to be in the context of the collapse of an asthenospheric bulge in front of the retreating Weichselian ice shield (Figure 6). The combination of eustatic

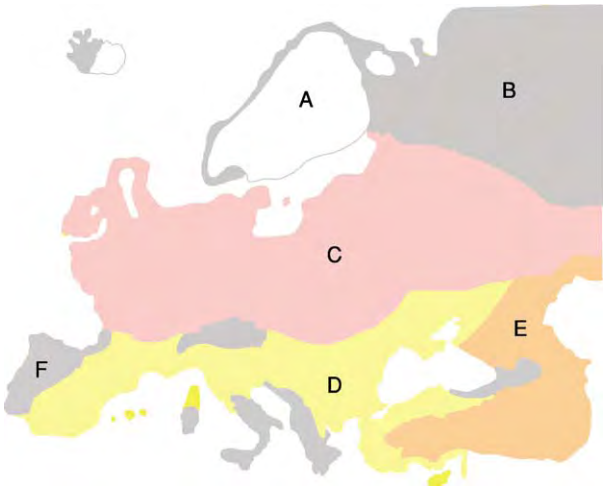


Figure 2 Vegetation zones in Europe during the Younger Dryas. (A Ice; B Polar Desert; C Steppe Tundra; D Dry Steppe; E Semi Desert; F Wooded Steppe. (Adapted from Adams 2002, with permission.)

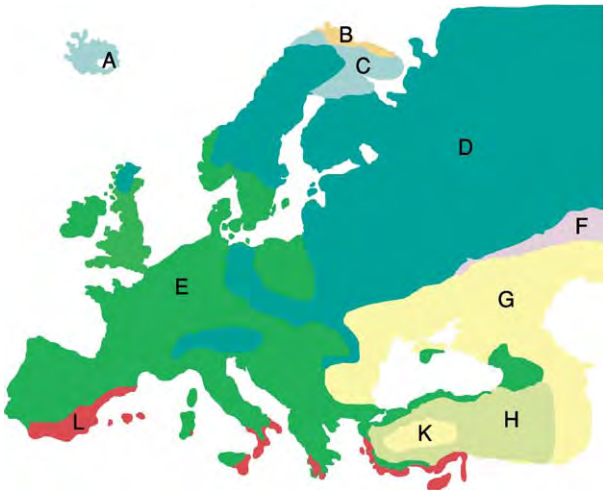


Figure 3 Vegetation zones in Europe about 9000 years BP (A Ice; B Tundra; C Open boreal woodland; D Boreal Forest; E Deciduous/Mixed Forest; F Forest Steppe; G Moist Steppe; H Woodland; K Dry Steppe; L Mediterranean Forest. (Adapted from Adams 2002, with permission.)

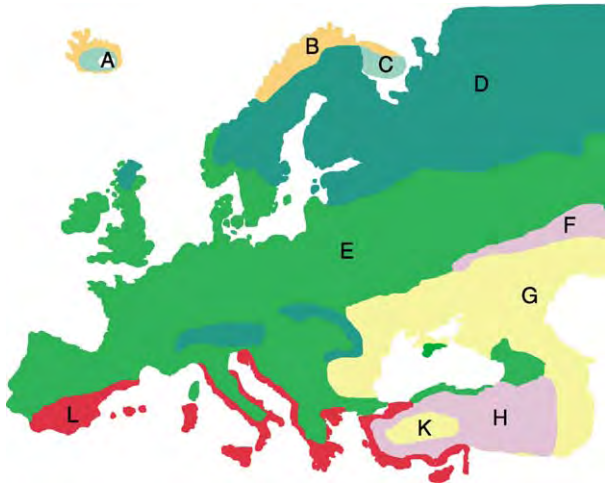


Figure 4 Potential present vegetation zones in Europe (A Ice; B Tundra; C Open boreal woodland; D Boreal Forest; E Deciduous/Mixed Forest; F Forest Steppe; G Moist Steppe; H Woodland/Wooded Steppe; K Dry Steppe; L Mediterranean Forest. (Adapted from Adams 2002, with permission.)

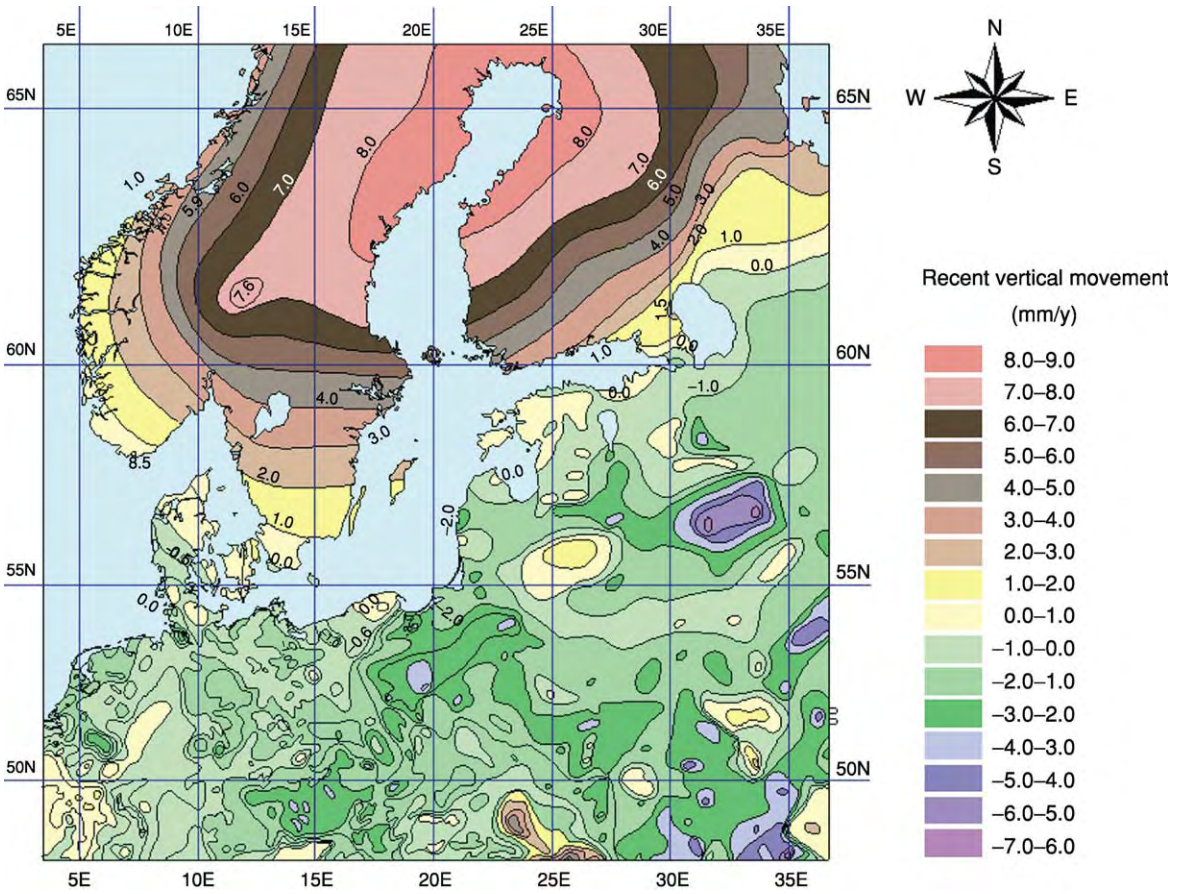


Figure 5 Recent vertical movement of the Earth's crust. In the northern part the map is dominated by the extensive, north east south west stretched uplift of Fennoscandia with maximum values of >8 mm year. The subsidence of a belt surrounding the Baltic Shield is less differentiated. (From Harff J, Frischbutter A, Lampe R, and Meyer M (2001) Sea level change in the Baltic Sea: interrelation of climatic and geological processes. In: Gerhard LC, Harrison WE, and Hanson BM (eds.) *Geological perspectives of global climate change*. Tulsa, Oklahoma, American Association of Petroleum Geologists in collaboration with the Kansas Geological Survey and the AAPG Division of Environmental Geosciences: 231 250. Reprinted by permission of the AAPG whose permission is required for further use.)

sea-level changes and isostatic uplift or subsidence, partly modified by tectonic movements, produced considerable changes in the geography of Europe during the Holocene. Depending on the geographic position of the affected area, large relative sea-level changes (positive or negative ones) have occurred (Figure 7). These changes are particularly obvious where large intracontinental basins like the recent Baltic Sea area were affected.

At the beginning of the Holocene, large parts of the Baltic Basin were filled with freshwater from the Baltic Ice Lake which was fed mainly by meltwater from a large glaciated area in North and north-eastern Europe. The only important outlet in the Öresound (between the recent Danish island Sealand and southern Sweden) was too narrow to serve as a sufficient spillway between the Baltic Ice Lake and the North Sea. The global sea-level was about 25 metres lower than in the Baltic Ice Lake. When the Scandinavian inland ice started to retreat from southern Sweden, a spillway through the central Swedish Depression was opened. As a dramatic process, half of the recent Baltic Sea's water volume drained into the Atlantic Ocean via the Kattegat and North Sea. This drainage took no longer than a few years and had an enormous impact along the former shores of the Baltic Ice Lake. Large areas previously covered by water became dry land, and southern Scandinavia became directly connected to central Europe. Saline waters of the Kattegat could enter the Baltic Basin for a time-span of a few hundred years, a stage of the Baltic Sea's development known as the *Yoldia Sea*, a phase which is dated from 11 570 to 10 700 years BP (Figure 8).

The connection between the Yoldia Sea and the Kattegat through central Sweden was located in a rapidly uplifting region. Therefore, the connection closed at about 10 700 years BP and a newly dammed-up freshwater lake was formed within the Baltic Basin. It

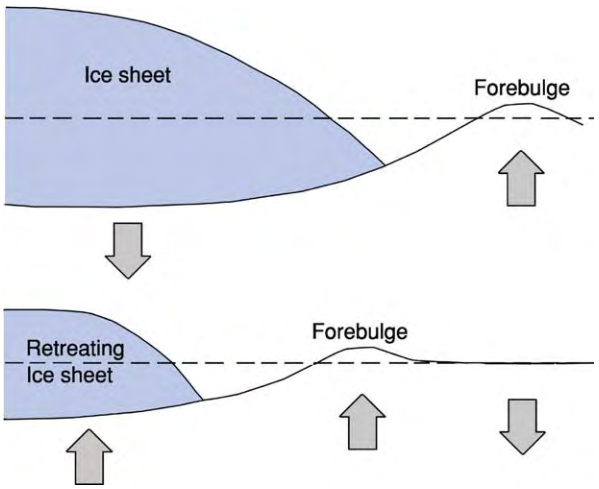


Figure 6 Principle of glacio isostatic vertical crustal move ments: (a) Ice load causes subsidence of the Earth's crust below and compensatory uplift beyond the ice margin as a fore bulge; (b) Uplift of the ice released Earth's crust and related displacement of the forebulge. (Modified after Daly 1934.)

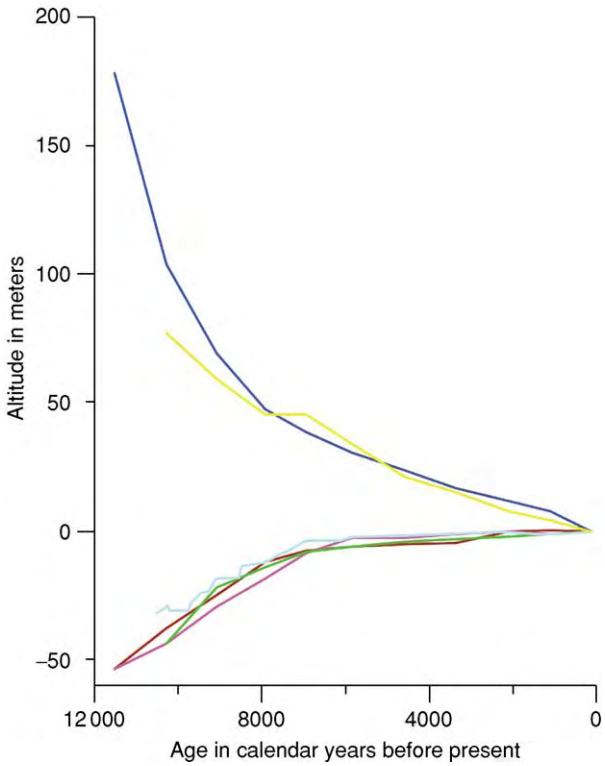


Figure 7 Schematic Holocene relative sea level curves for different European regions (blue = Oslo, Norway; yellow = south eastern Sweden; light blue = southern Baltic Sea; brown = French Atlantic coast; green = Netherlands; purple = central Mediterranean).

is called *Ancylus Lake*, with reference to a typical freshwater snail occurring in these waters (*Ancylus fluviatilis*) (Figure 9). This freshwater stage lasted more than two thousand years and was characterised by considerable short-term water level fluctuations. As differential isostatic uplift continued, the critical thresholds between the Baltic Basin and Kattegat moved from southern Sweden to the recent-formed Danish Straits. At about 8000 years BP, the eustatic sea-level rise led to the first ingressions of marine waters into the Baltic Basin. In pace with the rapidly rising global sea-level, the thresholds were flooded and a stable connection between the Kattegat and the Baltic Basin was formed. This crucial phase of the Baltic Sea's evolution is called the Littorina transgression, after a marine snail which is common in deposits of this period (*Littorina littorea*). At the onset of the Littorina transgression, the water level rose at a rate of 25 mm year within the south-western Baltic Basin (eustatic rise added to crustal subsidence, as described above), which slowed down later to about 3 mm year at about 4500 years BP (Figure 10). During a time-span of less than a thousand years, the sea-level rose by more than 20 m, implying enormous rates of coastal retreat within the southern Baltic Basin. Due to the rapidly rising sea-level at the beginning of the Littorina Stage, the glaciogenically-shaped land relief was drowned without any notable coastal erosion and longshore transport processes of sedimentary material. The resulting geographical situation is shown in Figure 11. Only after the sea-level rise slowed down at the end of the Littorina Stage (about 2000 years BP), and during the Post-Littorina Stage, was the recent spit and barrier coast formed by erosion and sediment transport. This process was mainly controlled by climatic factors, such as the wind-driven hydrographic regime which was superimposed on long-term eustatic and isostatic movements which have caused rising relative sea-levels and coastal retreat at the southern Baltic shores in recent times. Further north in Fennoscandia, isostatic uplift continuously exceeded the eustatic sea-level rise, resulting in a permanent general sea regression (Figure 7).

The changing Holocene sea level within the Mediterranean basins was and is mainly controlled by eustatic processes. In contrast to northern Europe, glacio-isostasy does not play a significant role. On the other hand, this region occupies the junction between the African-Arabian and the Eurasian plates which gives considerable tectonic activity in the different sedimentary basins of the Mediterranean Sea. Therefore, the general picture of the sea-level, development which reflects the eustatic curves, is superimposed on the regional and local tectonics.

A matter of ongoing discussion is the possibility of the reconnection of the Black Sea and the Mediterranean Sea during the period of Holocene sea-level rise. A catastrophic flood scenario at about 7500 years BP

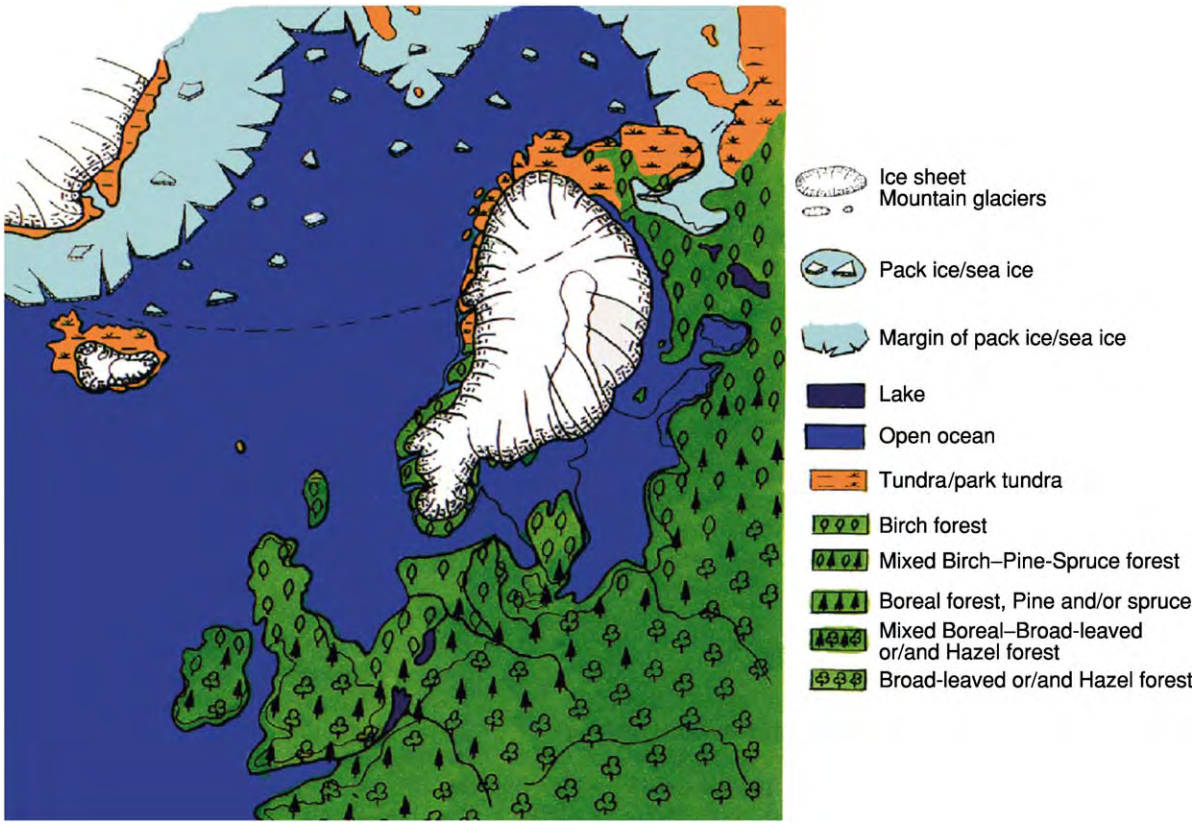


Figure 8 North western Europe during the Yoldia Sea stage of the Baltic Sea's history (about 11 000 years BP). (From Andersen BG and Borns HW (1994) *The ice age world: an introduction to Quaternary history and research with emphasis on North America and Europe during the last 2.5 million years*. Oslo: Scandinavian University Press, with permission.)

is questioned because of indications of a more complex and progressive transgression process over the past 12 000 years.

Interrelation of Human Activity and Natural Environment

In terms of human history, the onset of the Holocene is equivalent to the beginning of the Mesolithic period, which is characterised by prevailing hunter, fisher, and gatherer societies in Europe. The temperature rise at the end of the Younger Dryas caused an accelerated deglaciation. Alpine glaciers retreated to historical dimensions and the receding inland ice in Scandinavia was followed successively by tundra, steppe, and finally forests. This was accompanied by an enhanced northward migration of animals and their Mesolithic hunters. The rapidly changing natural environment, possibly in combination with human activity, resulted in the extinction of some characteristic genera of the Pleistocene megafauna such as the mammoth (*see* **Tertiary To Present:** Pleistocene and The Ice Age). Another consequence of the rapidly changing landscapes was a very variable migration pattern of the Mesolithic hunter, fisher, and gatherer groups in central and north-western Europe.

The following Neolithic period was closely connected with the introduction of farming in the various geographical regions. The first Neolithic settlements in southern Europe (Greece) are dated to about 9000 calendar years BP. In northern Germany, southern Scandinavia and the British Isles, they are more than 3000 years younger. Coming from the Near East, the new method of food-procurement spread to the Great Hungarian Plain in a first wave from 8200–7800 years BP. A second leap entered the North European Plain at around 7400 years BP. Neolithic settlers from Southeast Europe migrating along the rivers Danube and Rhine were probably responsible for the consequent social and cultural changes. In parts of Europe (e.g., western Mediterranean and northern Europe) the native Mesolithic population adopted agricultural methods to form a transitional economy.

While the hunter-fisher-gatherer societies of the Mesolithic used and manipulated the natural ecosystem without altering it considerably, the Neolithic farmers started to transform the environment according to their

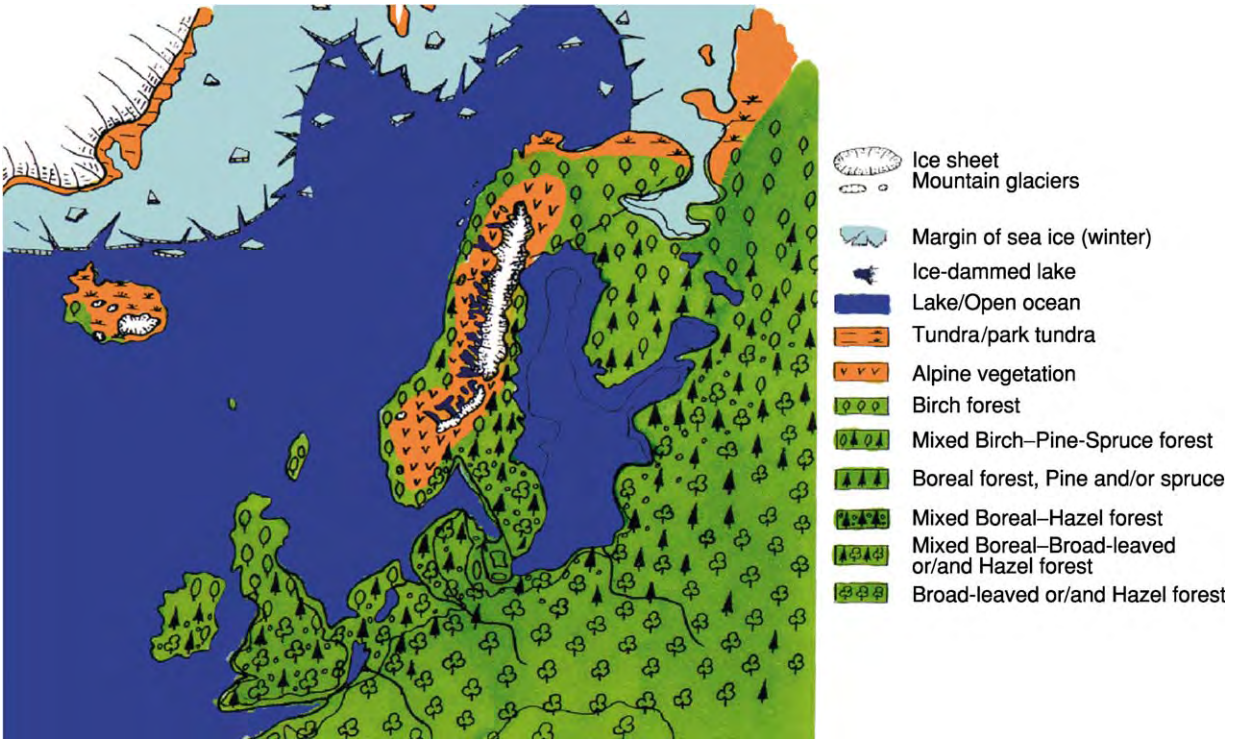


Figure 9 North western Europe during the Ancylus Lake stage of the Baltic Sea's history (about 9500 years BP). (From Andersen BG and Borns HW (1994) *The ice age world: an introduction to Quaternary history and research with emphasis on North America and Europe during the last 2.5 million years*. Oslo: Scandinavian University Press, with permission.)

needs. Cultivation started within rather small (perhaps pre-existing) open land patches in the Early Neolithic. In this phase clearances remained localised and the early Neolithic farmers were still highly dependent on intact natural habitats surrounding the cultivated land patches.

This situation changed in the later part of the European Holocene, when increasing demand for agricultural products induced a major transformation of the yet still mainly natural environment into an agricultural one. This process was certainly time-transgressive and affected some European regions more than others. But, far from being controlled solely by human action, natural feedback, combined with differing vulnerability of the existing ecosystems, amplified, shifted, or interfered with the initial direction of processes initiated by human activity. In order to adapt to the partly self-induced new conditions, human societies had to react by further cultural development which accelerated the general transformation process of the natural environment. This is exemplified by the domestication of animals which were advantaged by the anthropogenically influenced environment, while on the other hand natural competitors and predators became progressively extinct by hunting kill-off or simply by loss of habitat (extensive land use by farming). The same applies to floral assemblages, as initially, their composition was governed by natural conditions but it became increasingly influenced by agriculture. This development finally produced an increasing interdependence of cultivated plants and animals with mankind. Societies like the Irish people in nineteenth century, for example, were greatly dependent on potato growth. When the Late Potato Blight ruined all the potato crop in Ireland in the 1840s, the resulting famine lead to dramatic consequences. About 1 million people died while another 1.5 million people emigrated.

The stability of the cultural landscape which had evolved out of the interplay between natural and human influences, as well as that of human societies living there, depended on the state of their equilibrium. Changes of the natural component could cause considerable impact to the progressively complex human society and this is exemplified by climatic or geological influences. Some historical epochs with a prospering economy and politically stable conditions are connected with warmer periods.

Colder climate deviations often were characterised by political and economical instability. This refers, for example, to the time of the Germanic migrations about 450–700 AD and the Little Ice Age between 1500 and 1800 AD. Rather short-term, but possibly catastrophic, impacts derived from events like

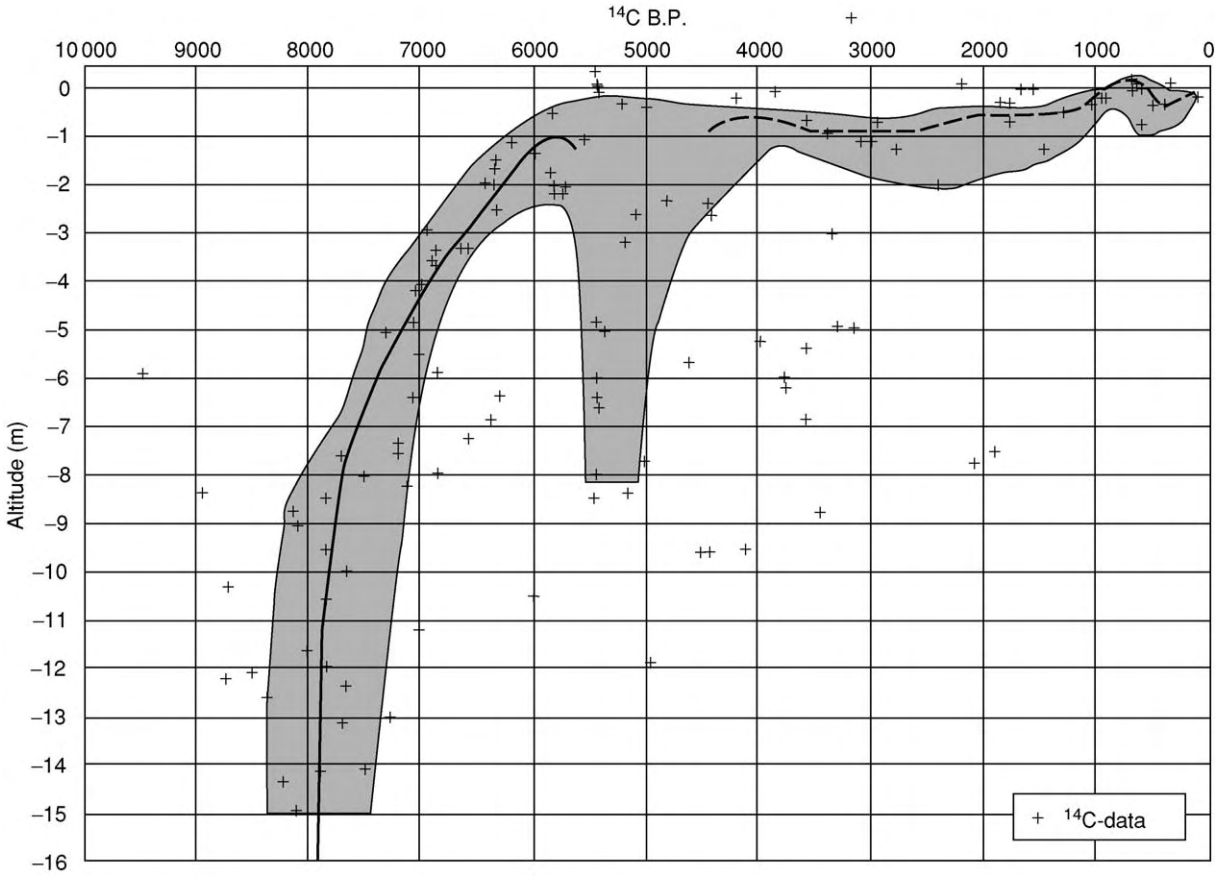


Figure 10 Preliminary curve of relative sea level change for western Pomerania, shaded area: interval of 'data confidence', solid line: local trend for western Pomerania, dashed line: estimated trend, for the time span between 6000 and 4500 years BP tectonically controlled local uplift is assumed. Note: Ages are given here in ^{14}C years BP. (From Harff J, Frischbutter A, Lampe R, and Meyer M (2001) Sea level change in the Baltic Sea: interrelation of climatic and geological processes. In: Gerhard LC, Harrison WE, and Hanson BM (eds.) *Geological perspectives of global climate change*. Tulsa, Oklahoma, American Association of Petroleum Geologists in collaboration with the Kansas Geological Survey and the AAPG Division of Environmental Geosciences: 231 250. Reprinted by permission of the AAPG whose permission is required for further use.)

volcanism (e.g., the eruption of Vesuvius in AD 79 described by Pliny the Younger), earthquakes (e.g., the disastrous one of Lisboa in AD 1755) or floods along rivers and seashores.

On the other hand, changing social conditions also induced dramatic changes of their natural environment. Once human activity stopped or declined, the surrounding environment developed depending on its natural stability. Robust ecosystems changed back to a state similar to the original one being controlled by the natural conditions. An example for such processes is the re-forestation after depopulation during the Thirty Years' War (1618–1648 AD) in central Europe.

In other cases, human activity led to irreversible effects on the ecosystem's stability when the social structures collapsed. During the Roman period there was a land use maximum partly on metastable soils. Here, terraces were maintained, preventing soil erosion. After the invasion of eastern nomads, the land was partly abandoned and soil erosion started at a greater extent and in the worst case, barren badlands were the final result. Therefore, maximum erosion is not connected with maximum land use, but with subsequent phases in different cultural environments. Generally, when judging human impact on the natural environment, it is often regarded only as negative. However, in contrast, the creation of new metastable ecosystems more diverse than before, particularly in the early phases of the Holocene, would not have been possible without human activity.

Human Activity and Environmental Conservation

During the last 500 years, human impact on the European environment have become much more important than the natural conditions. Particularly

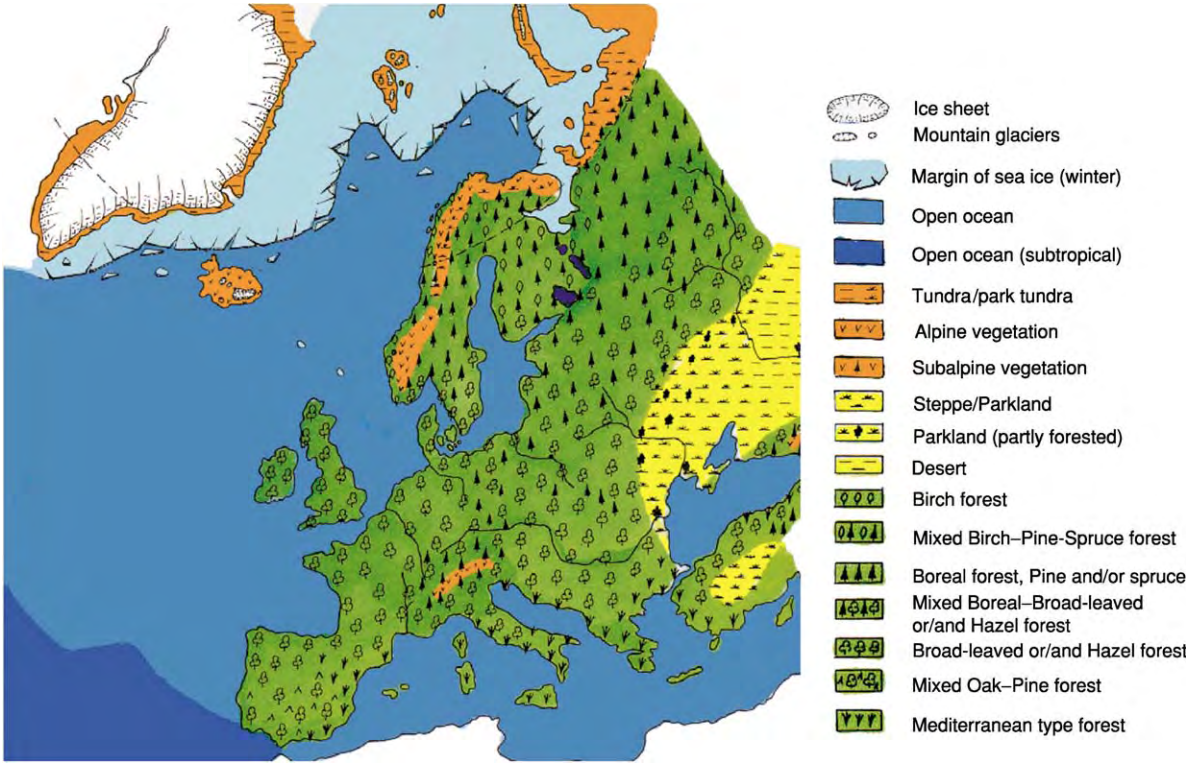


Figure 11 Europe about 4500 years BP. (From Andersen BG and Borns HW (1994) *The ice age world: an introduction to Quaternary history and research with emphasis on North America and Europe during the last 2.5 million years*. Oslo: Scandinavian University Press, with permission.)

within the last hundred years, industrial methods in agriculture have resulted in a widespread conversion of natural habitats into farming land. Accelerated human-induced soil loss has become a major problem. Measures initiated to solve the problems of modern agriculture, for example, to protect the soil against nutrient depletion, partly result in adverse effects such as eutrophication.

One of the most severe impacts on environment is connected with the beginning of industrialisation. Natural resources which had accumulated over millions of years have been exploited within decades or less. Large industrial facilities have been built on terrain of formerly less affected landscapes. Moreover, military needs have modified great parts of the terrain. Growing populations all over Europe have enhanced the conversion of natural habitats into settlement areas. At the same time, the management of industrial and municipal waste has become a major issue. The effects of industrialisation are reflected by many proxies such as the concentration of heavy metals in sub-recent deposits (Figure 12).

All these tendencies have developed ideas on nature conservation, sometimes with the idealistic approach of going back to a state where human influence is negligible. Nature conservation in this context may be regarded as a contradiction in itself because it does not recognise the vital role of human society for the natural environment. Large areas along the Netherlands's coast would have been flooded if the coast was not protected by coastal engineering. Furthermore, the dynamic character of the environmental status must be considered. For the last 8000 years of the Baltic Sea's history it can be shown, for example, that high nutrient levels already existed immediately after the Littorina Transgression. Without any remarkable human influence, organic substances and nutrients could accumulate in the sediments because of restricted vertical convection. This process has been intensified during the last centuries and decades by agriculture and the industrial release of nutrients (Figure 13). Thus, if the Helsinki Commission for the protection of the Baltic Sea aims to restore the eutrophication level of the 1950s, it must be stated that that was just one time slice of the Baltic Sea's Holocene development. One might be successful in re-establishing the concentrations of certain nutrients, but it would be in a completely different new context. As human society develops further, it cannot be expected that there will be no response by the natural environment. A certain equilibrium, including the benefit for as many species as possible including

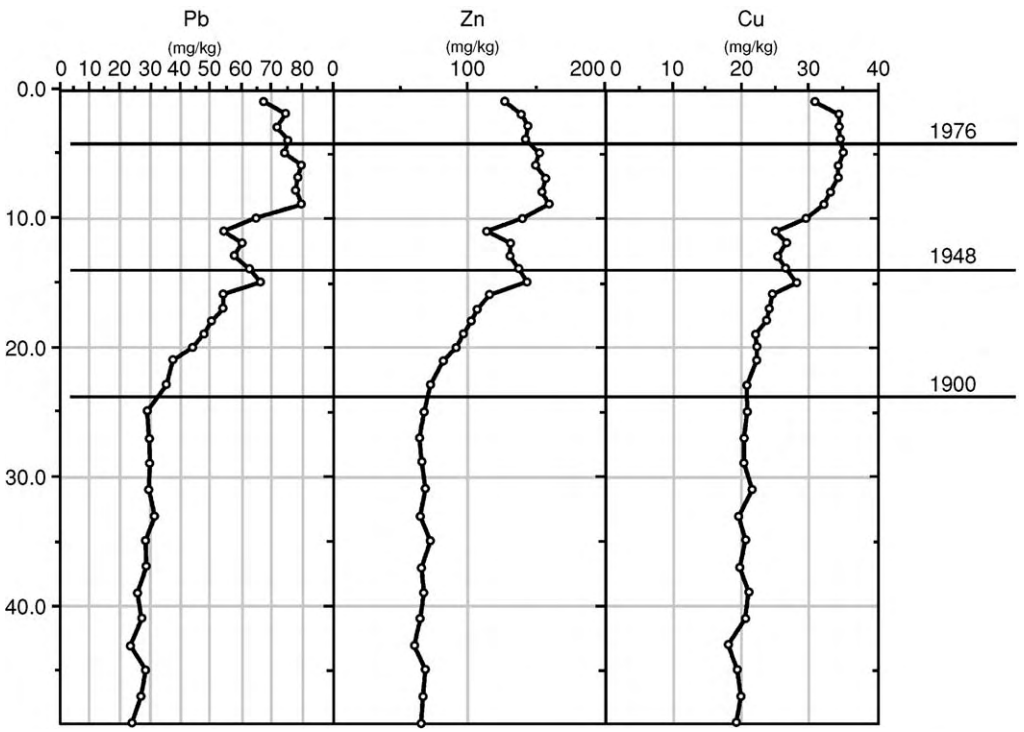


Figure 12 Enrichment of heavy metals within the last century, as recorded in a sediment core from the Arkona Basin (western Baltic Sea). (From Th. Leipe, Baltic Sea Research Institute Warnemunde, with permission.)

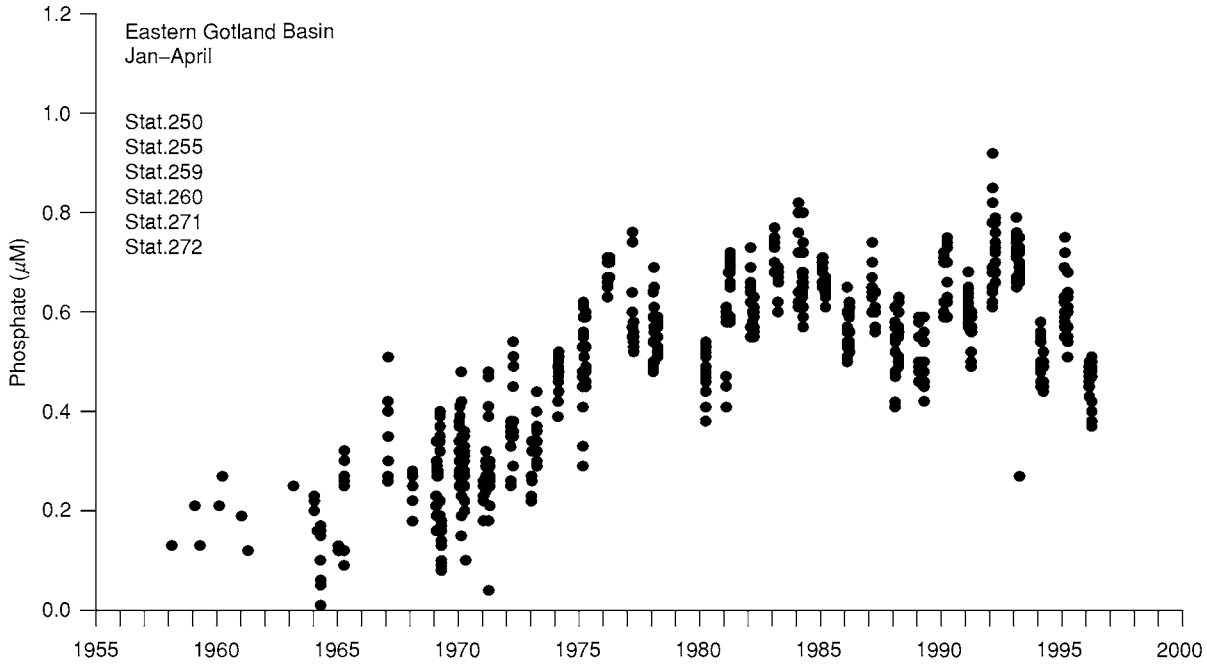


Figure 13 Trends of winter phosphate concentrations in the surface layer (0 10 m) of the eastern Gotland Basin (central Baltic Sea). (From G. Nausch, Baltic Sea Research Institute Warnemunde, with permission.)

Table 1 Summary of the climatic and historical development during the European Holocene

Calendar years BP	*Environmental Period*	*General climatic features*	*Selected historical events*	*Stages of the Baltic Sea's development*
		Modern climatic optimum, warm	World War I and II	Mya Sea
200	Subatlantic			
			French Revolution	
		Little Ice Age, mean annual temperature in Europe 1°C lower than today, cold winters, but pronounced fluctuations, glacier advances		Less brackish
400			Thirty Years' War	Mya Sea Lymnaea Sea
			Renaissance and reformation	Brackish
600				
			Bubonic plague kills about one third of the European population	
		Transition to colder conditions		
800				
		Medieval climatic optimum, mean annual temperature 1 1.5°C higher than today, winegrowing as far north as to the British Isles, first dry, later wet		Brackish
	Subatlantic			
1000				
			Expansion of Norman people as far as Iceland, Greenland, North America and southern Russia	
1200				
		Cold and wet period, many glacier advances		
1400				
	Subatlantic		End of the Roman Empire	Slightly brackish
1600				
			Invasion of the Huns forces the emigration of nations	Lymnaea Sea
1800				
2000		Roman climatic optimum, as warm as the Medieval optimum, mostly very wet, towards the end more dry	Foundation of the Roman Empire	
			First southward migrations of Gothic tribes	
2200				
		Pronounced cold period, mean annual temperature 1 1.5°C lower than today, very cool summers, very wet, large glacier advances	Celtic La Tène culture in big parts of Europe	
2400				
			Beginning of the Greek classic period Formation of the Roman Republic	
2600			Celtic Hallstatt culture in central and western	
	Subatlantic		Europe introduces the iron age	
	Subboreal		Greek Archaic period	
2800				
3000			First Celtic tribes in eastern and central Europe	
				Brackish
3200				

(Continued)

Table 1 Continued

Calendar years BP	*Environmental Period*	*General climatic features*	*Selected historical events*	*Stages of the Baltic Sea's development*
		Predominantly warm with Distinct fluctuations, less precipitation than during Subatlantic		
			Urn field culture in central and South eastern Europe	
3400				
			First bloom of the Mycene culture in Greece	
3600				
			Beginning of the Bronze Age in northern Europe	
3800	Subboreal			
		Cold period with glacier advances, initially dry, later more wet		
4000				
4200			In Europe exists an extensive trade route network	
4400				
			Early Minoic culture at Crete Island launches the European Bronze Age	
4600				
4800				
5000			Increasing influence of Indo European people all over Europe	
5200			Introduction of the wheel in Europe	
5400			Megalithic monuments in many regions of central and western Europe	Brackish
5600	Subboreal			
5800	Atlantic			
6000			Narva culture in North eastern Europe	
6200		Rapid increase of humidity in eastern central Europe	First stone buildings at the Orkney Islands	
6400			First Neolithic settlements at the British Isles	
		Warm period, mean annual temperatures 2 3°C warmer than today, especially warm winters, very moist, former predominant pines are replaced in the forests by oaks, lime and hazel		
6600				
			Late Mesolithic Ertebølle culture in northern central Europe	
6800				
7000			Neolithic Karanovo culture in South eastern Europe	
7200			Late Mesolithic Ertebolle culture in northern Europe	
7400			Neolithic 'Bandkeramik' culture in the Loess areas of central Europe	
7600				Strongly brackish
7800				
			Mesolithic Kongemose culture in northern Europe	Littorina Sea
8000				

(Continued)

Table 1 Continued

Calendar years BP	*Environmental Period*	*General climatic features*	*Selected historical events*	*Stages of the Baltic Sea's development*
		Short lived cold interval, drop of the mean annual temperature by 2°C		Ancylus Lake
8200			First Neolithic agricultural societies in South eastern Europe and Greece	Freshwater
8400			The English Channel separates the British	
8600		Slightly warmer and moister than today	Isles from the continent	
8800				
9000	Atlantic		Hunter and gatherer societies	
	Boreal	Cold phase in the Carpathian Basin		
9200				
		During summers generally warmer than today, mainly open winters	Maglemose culture in northern Europe	
9400				
			Azilian culture in western Europe	
9600				
			Microliths become common	
9800				
			Use of log boats is proven	
10000				
10200	Boreal			
	Preboreal		Nomadic hunters arrive in England	
10400				
			Ahrensburg culture in northern Germany	
		Freshwater		
10600				
		Summers as warm as today, but very cold winters		
10800				Ancylus Lake
				Yoldia Sea
11000				Freshwater
11200				Regionally brackish
11400		Quick warming	Bromme culture in Denmark	
				Freshwater
11600	Preboreal	Beginning of the Holocene	Mesolithic	Yoldia Sea
	Younger Dryas	End of the Pleistocene	Palaeolithic	Baltic Ice Lake
11800				
12000		Cold period, mean annual temperatures by 5 9°C lower than today		Freshwater

mankind is desirable. Closed production cycles might be one of the important targets to achieve this.

Actualism in a New Context

The special character of the Holocene as a period which is not only part of the geological past but also an interface with the future, gives the principle of actualism a new dimension. In addition to regarding the present as being the key to the past, past and present times might be regarded as a key to the future. Predictions of future developments become more and more important to cope with possible changes in the natural environment. For this purpose, detailed knowledge about similar processes in the past is indispensable. In order to calibrate proxies from the geological record, it is necessary to analyse recent proxies by comparison with older ones, and also the

written record in order to separate historical trends. Improvements in dating methods, and additional information from geological, archaeological, biological, historical and other sources will help to develop scenarios which might help the recognition and response to future challenges.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards. **Famous Geologists:** Lyell. **Fossil Vertebrates:** Hominids. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Adams, J *Europe during the last 150 000 years* [online at http://www.esd.ornl.gov/projects/qen/nercEurope.html]

Andersen BG and Borns HW (1994) *The ice age world: an introduction to Quaternary history and research with emphasis on North America and Europe during the last 2.5 million years*. Oslo: Scandinavian University Press.

Björck S (1995) A review of the history of the Baltic Sea, 13.0 8.0 ka BP. *Quaternary International* 27: 19 40.

Cunlifffe B (ed.) (1994) *The Oxford Illustrated Prehistory of Europe*. Oxford New York: Oxford University Press.

Donner J (1995) *The Quaternary history of Scandinavia*. Cambridge: Cambridge University Press.

Emeis K C and Dawson AG (2003) Holocene palaeoclimate records over Europe and the North Atlantic: modelling and field studies. *The Holocene* 13: 305 464.

Grove JM (1988) *The Little Ice Age*. London, New York: Routledge.

Harff J, Frischbutter A, Lampe R, and Meyer M (2001) Sea level change in the Baltic Sea: interrelation of climatic and geological processes. In: Gerhard LC, Harrison WE, and Hanson BM (eds.) *Geological perspectives of global climate change*. Tulsa, Oklahoma, American Association of Petroleum Geologists in collaboration with the Kansas Geological Survey and the AAPG Division of Environ mental Geosciences: 231 250.

Litt T, *et al.* (2003) Environmental response to climate and human impact in central Europe during the last 15 000 years a German contribution to PAGES PEPIII. *Quater nary Science Reviews* 22: 1 124.

Pirazzoli PA (1991) *World atlas of Holocene sea level changes*. Elsevier Oceanography Series 58, Amsterdam, London, New York, Tokyo: Elsevier Science Publishers B.V.

Roberts N (1998) *The Holocene An environmental his tory*. Oxford: Blackwell Publishers Ltd.

Schönwiese C (1995) *Klimaänderungen Daten, Analysen, Prognosen*. Berlin: Springer Verlag.

EVOLUTION

S Rigby, University of Edinburgh, Edinburgh, UK
E M Harper, University of Cambridge, Cambridge, UK

Introduction

The theory of evolution by natural selection, put forward by Darwin in 1859 (*see* **Famous Geologists:** Darwin), is the greatest unifying theory of biology and palaeontology. In this context, evolution is the change that occurs between successive populations of organisms, due to their modification in response to selection pressures. The potential to change is provided by genetic variability within populations and by genetic change through time (mutation). The pressure for change to occur exists outside an organism and is provided by interactions within the environment. These interactions may be predominantly physical or biological effects. Small-scale changes in populations, giving rise to new species, are defined as microevolution. Larger-scale changes, such as the origin of new higher taxa – which may have new body plans or new organs – are defined as macroevolution. The study of evolution also includes the study of patterns of diversification and extinction. Macroevolution may be the end result of microevolution working over a long time-scale or it may be a suite of emergent properties that require unique interpretations.

Historical Background

The presence of large numbers of species on the Earth and the means by which they appeared were discussed throughout the Enlightenment, though the use of the word evolution did not become common until the twentieth century. The possibility that one species might change into another was of interest to Charles Darwin's grandfather, Erasmus Darwin, for example. Studies that focused on the ways in which species transform were begun by Jean-Baptiste Lamarck and published in his *Philosophie Zoologique* in 1809.

Lamarck argued that an 'internal force' caused offspring to differ slightly from their parents and also that acquired characters could be passed on to the next generation. One of his examples was that of giraffes, whose long necks were assumed to be a

product of successive generations reaching for higher and higher leaves. He suggested that each giraffe lengthened its neck slightly by this activity and, in turn, passed on to its descendents the capacity to grow longer necks. He visualized species as forming a chain of being, from simplest to most complicated, with each species being capable of transforming into the next in line, and all existing indefinitely. In Britain this work was disseminated by both Richard Owen, who was generally supportive of the theory, and Charles Lyell (*see* **Famous Geologists:** Lyell), who was critical of it.

Charles Darwin encountered work by both of these scholars and also explored huge tracts of the natural world during his 5 years study on the *Beagle* (1831–1836). His work on a number of organisms, notably finches collected from the Galapagos Islands in the Pacific, persuaded him that organisms were adapted to their particular niche and that species were capable of change. The process by which this change could occur was a preoccupation of Darwin's in the succeeding years. As early as 1838, he had read the seminal work of Malthus on populations, but he was still working on the scope and implications of his theory when he was forced to publish by correspondence from Alfred Russel Wallace. A joint paper presented to the Linnaean Society in 1858 was followed the next year by his classic work *On the Origin of Species*.

Darwin's theory of species originating through natural selection can be set out in a small number of propositions. First, organisms produce more offspring than are able to survive and reproduce. Second, successful organisms – those that survive long enough to breed themselves – are usually those that are best adapted to the environment in which they live. Third, the characters of these parents appear in their offspring. Fourth, the repetition of this process over a long time-scale and many generations will produce new species from older ones.

The consequences of this theory are enormous. Not least, they caused scientists at the time to reconsider their assumption of a chain of life. Evolution by natural selection is a response to the local environment and is not predetermined on a grand scale. Organisms do not necessarily evolve into more complicated species over time. Amongst the general public, the theory was seen as being in conflict with a literal reading of the Bible, a view that persists amongst a religiously conservative minority.

In the years after publication, the most significant weakness of Darwin's theory was perceived to be its failure to supply a plausible mechanism for the inheritance of characters. However, this mechanism was supplied when Gregor Mendel's (1865) work on heredity was rediscovered in the early twentieth century. Mendel observed that characters were passed from parent to child in a predictable fashion depending on the relative dominance of the traits carried by each sexual partner. Characters did not 'blend' in the offspring, which is what Darwin had suggested and which astute critics had pointed out would actually have prevented evolution from occurring. These observations opened the door to the modern study of genetics. After some decades of debate, a modern consensus was reached in the 1940s, which is the basis for our current understanding of Darwin's ideas.

Evolution and Genetics: The Living Record

Evolution is possible because the genetic transmission of information from parent to offspring works as it does, in a Mendelian fashion. Subsequent work on genetics has elucidated the exact means by which this occurs and has shown how variation can be developed and sustained in a population.

The information that can be passed from one generation to the next in a population is contained on strands of DNA (deoxyribonucleic acid), or occasionally RNA (ribonucleic acid), within each cell. A DNA molecule forms from a series of nucleotides, which are joined up like beads on a string. Each nucleotide has, as one of its elements, a base. The four types of base DNA are adenine, thymine, guanine, and cytosine (usually abbreviated to A, T, G, and C). Two strings of nucleotides join via base pairs to make the double-helix shape of DNA. A always joins to T, and C always joins to G. Sequences of bases are the code that stores the information needed to produce an organism. This includes information about making the various parts of the cell or set of cells and also information about the rates at which different processes should occur and their relative timings. Each piece of information that the DNA holds is called a gene. Genes can be sequences of DNA or can be little pieces of DNA separated by other sets of bases. Most of the DNA appears to have no purpose and is called non-coding DNA. A human is produced from about 30 000 genes that use about 5% of the nucleotides of our DNA (**Figure 1**).

When sexual reproduction occurs, one copy of the DNA (carried on chromosomes) of each parent is passed to the children. The offspring therefore have two sets of instructions within their DNA. The pair of genes that share a common function are called alleles, and the combination of alleles controls the effect on the bearer. However, this effect will not be passed to

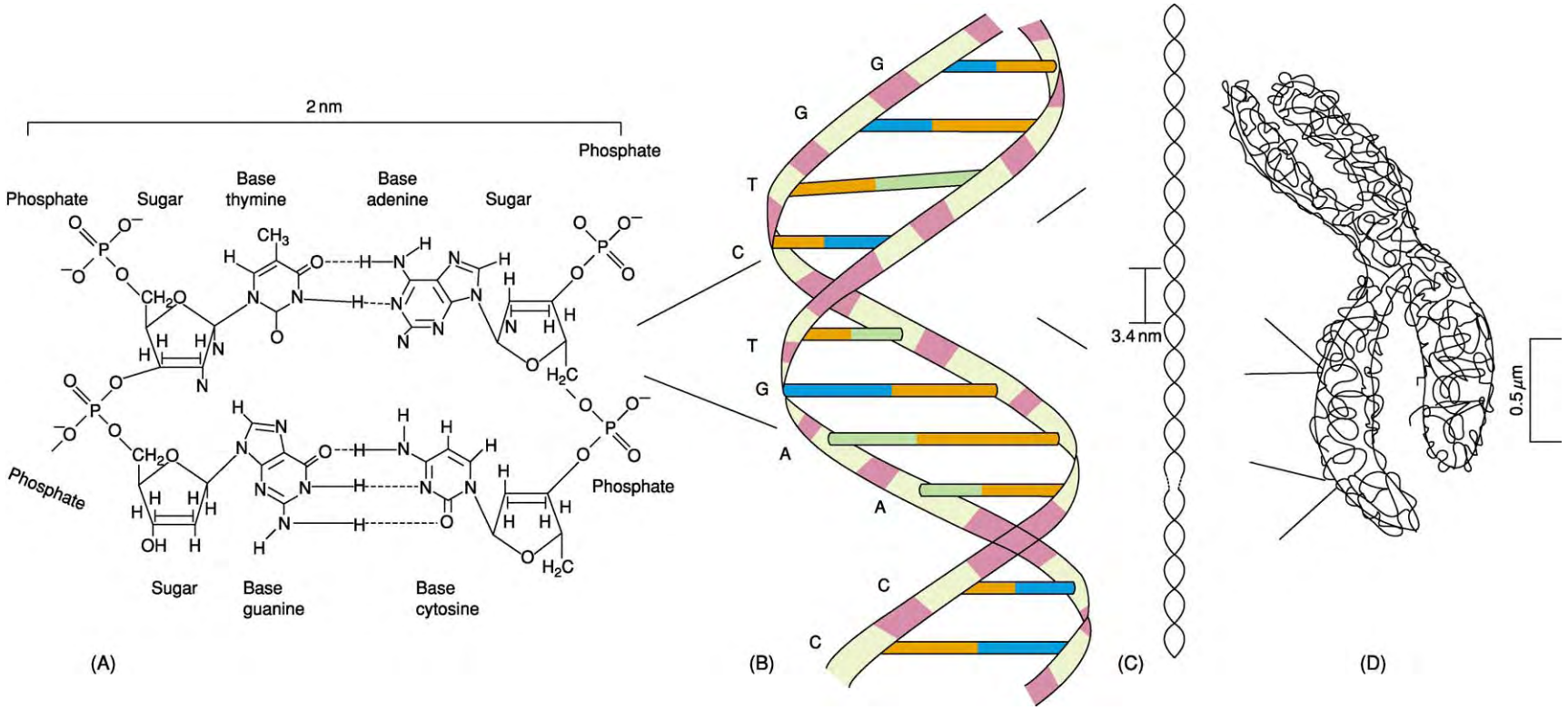


Figure 1 A diagrammatic representation of DNA, genes, and chromosomes. (A) The molecular structure of a double strand of DNA. Each strand is made up of a chain of sugars (yellow) and phosphates (purple), linked together by a set of four bases: thymine (orange), adenine (green), guanine (blue), and cytosine (red). The shapes of these bases cause adenine to bond to thymine and guanine to bond to cytosine. This makes each strand of DNA a mirror image of the other. (B) A piece of DNA carries information in the form of sets of bases, in this example GGTCTGAAC. (C) A gene is a set of useful bits of DNA, which code for a particular protein or carry out a particular instruction. Genes may be formed in pieces separated by long intervals. (D) A chromosome is a folded up cluster of DNA found in the nuclei of eukaryotic cells.

the next generation, but rather one of the alleles will. This will then combine with another allele to generate another product. Although the results of allele combination can have a complicated range of expressions in a cell or a body, the alleles don't mix, so variation is maintained.

A wide range of errors can occur when the DNA strand is replicated during reproduction. These can affect the non-coding DNA or the genes and can produce mutations of varying effect depending on whether it is genes that control the production of the body or the timing or duration of elements of this production process that are affected.

Time and Narrative: The Fossil Record

Biologists have explored theories of evolution in tandem with palaeontologists, who can retrieve narratives of evolutionary change from the fossil record. The ability to study change over millions of years is a great advantage of using fossils. Theoretically, it should be possible to study aspects of the morphologies of fossils collected bed-by-bed throughout a rock sequence in order to elucidate patterns of evolution. However, the preservation of individual fossils is often poor; most depositional events produce significant time-averaging, and the fidelity of long records of sedimentary sequences is often questionable. At some scale, all deposition is intermittent, and this means that there are gaps of some scale in all narratives retrieved from the fossil record. Fossils preserved in lakes or deep-sea cores may be less affected by this problem than fossils from more dynamic environments, and research has generally concentrated on these locations.

Microevolution

The set of potentially interbreeding members of a population forms a species, which contains a range of variation in its appearance (the phenotype) and in its genetic codes (the genotype). In practice, most living species are defined on the basis of phenotypic characteristics rather than genetic information or reproductive potential. In fossil studies of evolution, only the phenotypes are available, and the definition of a species must be based on clusters of phenotypic characters, which are taken as proxies for the potential to interbreed.

Natural selection acts on a set of individuals, so that the physical characteristics of the group and the underlying genotypes change over time. This process eventually gives rise to new species and is known as microevolution. Biologists class only gene shifts within populations as microevolution and define anything larger, including the appearance of new species, as macroevolution. To palaeontologists, the distinction is usually between speciation and anything higher, such as the emergence of new genera or of new organs.

Sometimes a species gradually changes through time until the point comes where the fossil representatives of successive populations are recognized as a different species. However, a parent species often splits into more than one offspring species or evolves into an offspring species that coexists with the parent species for some time. In this case, the original population must split into two or more subsets that cannot interbreed with one another. The two best-known methods of achieving this are called allopatric speciation and sympatric speciation.

In allopatric-speciation events a single original population is split into two geographically isolated elements. This is a common phenomenon over geological time as continents fragment, mountains rise, or sea-levels change. Each geographically separated fragment of the initial population contains only a fraction of the original genetic variation, so it may tend towards difference from the original population without any active selection, although this is now regarded as a minor component in the formation of new species. More importantly, different geographical regions will tend to produce different environmental stresses from those that were experienced before separation, leading to the selection of different successful characters in the separated populations. This eventually leads to significant changes of form in the isolated populations, which may finally produce new species.

An example of allopatric speciation has been recovered from the fossil record of Plio-Pleistocene (3–0.4 Ma) radiolarians, which are siliceous planktonic protists, collected in the North Pacific. A divergence in the forms of two sister species of the genus *Eucyrtidium* was found to have occurred at around 1.9 Ma, following a short period when the populations had been separated from one another.

During sympatric speciation the emerging species share a geographical range but may become separated over time by differences in behaviour or in resource exploitation. Adaptive pressures act differently on these populations, and different characteristics will be favoured, leading to a progressive change of form and eventually to reproductive isolation. At this point a new species will have appeared. There is some doubt about the mechanism by which species first begin to diverge without becoming geographically isolated, although the generation of new species in this way has been demonstrated for a number of types of animal. Studies on cichlid fishes in African lakes show that the most closely related species of fish

often live in the same lake, rather than in adjacent lakes, as might be expected if allopatric speciation had occurred (Figure 2). Although it seems intuitively obvious that populations that become physically dissimilar will eventually be unable to produce offspring, the genetic basis for this change can be demonstrated in the laboratory but not yet fully explained.

The fossil record can be used as a tool to help in the understanding of evolution and the formation of new species. It may be that evolution progresses gradually for most of the time, an idea known as phyletic gradualism. The classic fossil example of this slow continuous process of morphological change is the study by Peter Sheldon of Ordovician trilobites recovered from deep-water shales in central Wales. Eight different genera of trilobite, including well-known forms such as *Ogygiocarella*, were found to exhibit incremental changes in rib number through the duration of one graptolite zone, which probably represents significantly less than 1 Ma. Gradual change is generally difficult to observe in the imperfect fossil record. It could be argued that in a less continuous sedimentary record (or one sampled less finely) this sequence of events would appear as a series of abrupt changes. Commonly, what is preserved is a long period where little or no change is observed followed by the abrupt appearance of a new form.

The theory of punctuated equilibrium attempts to explain this phenomenon not as the product of an imperfect fossil record but as a common pattern of evolutionary change. This is done by applying the concept of allopatric speciation to the problem. Eldridge and Gould, who developed the idea of punctuated equilibrium, argue that most species probably arise in small, geographically isolated areas and that they arise rapidly as they encounter new selection pressures. At some later time the evolved offspring species may move back into areas where it encounters its parent species and may out-compete this form. In most areas where this happens, the geological record will show one species – the parent – abruptly replaced by another – the offspring – with no intermediate steps. The chance of the isolated population being represented in the fossil record during the short period of its evolution into a new species is very slim (Figure 3). It may be that Williamson, in a study of molluscs in Plio-Pleistocene sediments from Lake Turkana, found one such rare fossil example of punctuated equilibrium. Species of gastropod and bivalve both appeared to remain static in shape for long periods of time, punctuated by brief periods when their shape changed abruptly.

Although some studies seem to show a punctuated-equilibrium style of evolution, others appear to show that evolution has progressed via phyletic gradualism, and a consensus has yet to emerge regarding these theories. In practice, most evolution is probably the result of a mixture of punctuated and gradual periods of change, partly depending on the scale of

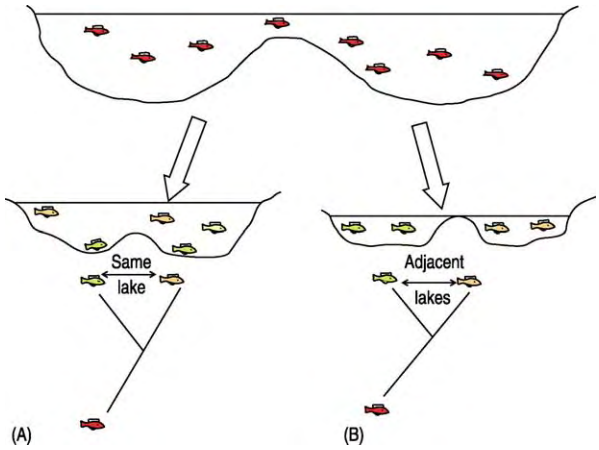


Figure 2 The difference between allopatric speciation and sympatric speciation, using the example of fishes living in lakes. (A) Sympatric speciation occurs due to changes in behav iour or mode of life, in this case by a partitioning of the original population into limnetic and benthic groups. Here, descendent species are most closely related to species living in the same lake. (B) Allopatric speciation occurs following geographical sep aration of the populations, in this case caused by a fall in lake level. Descendent species are most closely related to fishes living in adjacent lakes.

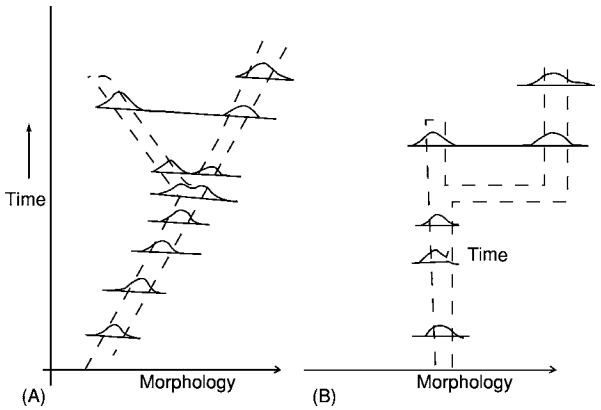


Figure 3 The differences between phyletic gradualism and punctuated equilibrium models of speciation. (A) In phyletic gradualism the shape change is gradual and populations are seen to move across morphospace continuously. Periods of spe ciation are relatively long and can be recorded in the fossil record. (B) In punctuated equilibrium the shape change is inter mittant, rapid, and related to geographical separation of a part of the population. For most of the time the form of the population is static. In most areas no speciation event is seen, and the fossil record shows abrupt changes of morphology with no intermediate stages.

observation. Work by Johnson on Jurassic oysters (*Gryphaea*) from across western Europe provides a good example of this aggregate pattern. Change over approximately 6 Ma was generally slow, but rapid periods of change in isolated populations were also observed. One unfortunate result has been the suggestion that punctuated equilibrium is antithetical to Darwinian evolution. In this usage it is not, as even the rapid bursts of evolution implied by the theory would take place via a series of gradual (i.e. small-scale) changes in the form of the organism concerned.

Macroevolution

Macroevolution is the study of all evolutionary events or effects larger than the appearance of a new species. This includes studies of long-term change in the geological record and of the emergence of new higher taxa, for example new phyla. Linked to both of these topics is the difficult issue of how significant new structures or organs can evolve. Palaeontology is central to this study, as it provides a measure of time and can identify the most likely dates of appearance of new characters or taxa.

The single biggest and most important argument about macroevolution is whether it is a scaled-up version of microevolution or something different. If it is different, then those differences may be a reflection of the emergent properties of this complicated system and hence still reliant on microevolutionary processes occurring. More controversially, it has been argued that macroevolution includes rapid and large-scale changes of form that necessitate steps that might initially produce organisms that are less successful than their ancestors. This is completely counter to Darwinian ideas of evolution. An example of these issues can be presented via a consideration of the evolution of major groups of tetrapods.

All living vertebrates with pentadactyl limbs (that is mammals, reptiles, amphibians, and birds) evolved from an ancestral fish, with the process beginning in freshwater lakes and rivers in the Devonian (Figure 4). Since then a wide variety of adaptations have appeared in these higher groups, such as feathers, fur, and wings. It can be convincingly demonstrated that some lineages, or evolving lines, acquired these characters gradually, by microevolutionary processes. The classic example of this is the origin of mammals from reptile ancestors through the Triassic and Early Jurassic. Character change occurred at a relatively constant rate throughout this 100 Ma period, and intermediate forms are well known in the fossil record. However, the process by which these new characters appeared may have controls that are not seen in microevolution and which are hinted at by

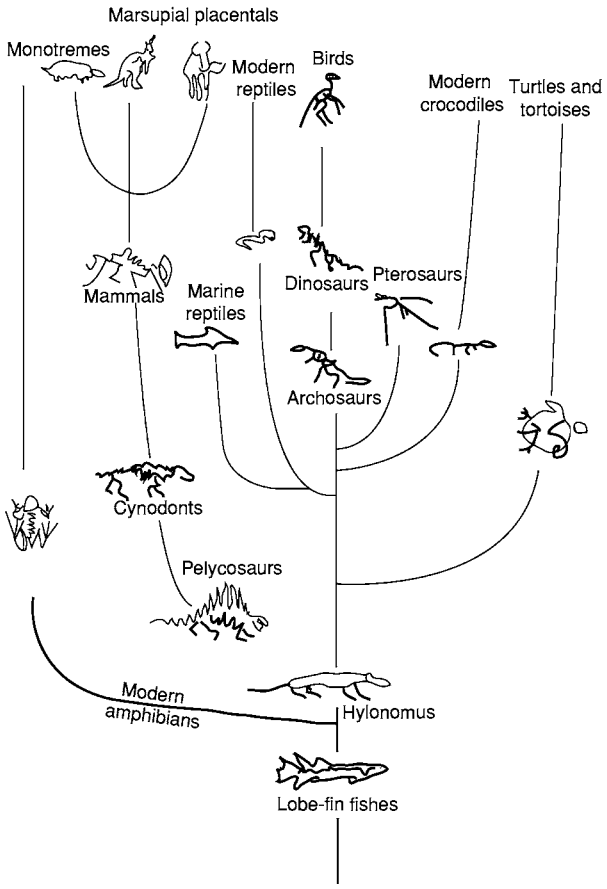


Figure 4 A simplified evolutionary tree for tetrapods, those vertebrates with pentadactyl limbs. This group of organisms evolved from lobe finned fishes during the Devonian. Whilst the evolution of mammals from cynodonts was a gradual process in which macroevolution appears to conform to microevolutionary expectations, the origin of wings in pterosaurs and birds is more difficult to explain without evoking some new process unique to evolution at this scale.

the suggestion that, during this event, significant evolution tended to occur in small-carnivore groups. More controversially it has been argued that some new characters, for example wings, could not have evolved by gradual steps as they would have been useless in their early stages of development. Complicated counter arguments that invoke the possible uses of wings without flight, for example, have not cleared up the controversy.

Looking at the history of life, it is clear that there have been periods of major increases in diversity and periods of major innovation. Significant increases in diversity tend to happen after mass extinctions and are called evolutionary radiations (*see* **Biological Radiations and Speciation**). Empty niches created by the extinction event are quickly filled as organisms radiate to form new species that are able to exploit the available resources. Evolutionary radiations may

also be facilitated by the appearance of major innovations, such as the evolution of hard parts by a variety of different taxa close to the Precambrian–Cambrian boundary. The two are not necessarily coupled. For example, eukaryotes, the complicated internally divided cells with which we are most familiar, evolved more than 2 Ga ago (possibly much more), but they did not become widespread or common until much later at around 1 Ga ago. They could not radiate until there was adequate oxygen present in the Earth's atmosphere and oceans, as they depended on this molecule for respiration.

It seems likely that evolutionary selection can work at the species level as well as at the level of an individual within a population. Species-level selection favours species that have lower extinction rates and higher origination rates than their 'competitors'. In the long-term, these species become more abundant at the expense of their less-successful competitors. Distinguishing between the levels of selection is extremely difficult in practice, but the theory helps to demonstrate the emergent properties of species.

Some types of extinction, or reductions in diversity, may also be explained by macroevolution and, in turn, throw light on the mechanisms of evolution. It is clear that the evolution of a new species will increase competition for resources and may force another species to become extinct if it is unable to compete successfully. The pattern of species extinction would be expected to be one of increasing chance of extinction with species age, but in some cases this does not seem to be so. Instead, species age does not appear to correlate with the likelihood of extinction. Van Valen has used this observation (which is itself somewhat contentious) to suggest a novel hypothesis for macroevolutionary patterns. He suggested that competition for resources produces a dynamic equilibrium between species, in which each will continue to evolve in order to survive. This is the core of the Red Queen hypothesis, which suggests that organisms evolve to keep their biological place or, to paraphrase the quotation from Alice in *Alice Through the Looking Glass*, they 'run to keep still'.

The characters that help organisms to survive at times of low extinction rate may be different from those that make survival of mass extinctions more likely. In other words, the criteria by which species are selected may vary with extinction rate. Specialist species tend to have greater survival potential at times when extinction rates are low and reduced survival potential when extinction rates are high. In addition, it has been suggested that small species have a greater chance of surviving mass extinctions than larger species, though the overall trend in evolution is clearly not towards smaller species.

The level of understanding of genetics is now so great that it is possible to explore macroevolution in this way. In traditional views of macroevolution, a set of ways in which different forms could be produced with small changes in the genome was known as heterochrony. The idea was that different parts of the body grew at different rates. In some examples, this might be a difference in the rate at which sexual maturity was reached relative to the rate at which the rest of the organism (the somatic portion) developed. If sexual reproduction became possible at an earlier stage in body development, this was known as pedomorphosis. The classic living example of this is the axolotl. This resembles a juvenile salamander, complete with external gills, but reproduces at this stage of development. If it is injected with extract from the thyroid gland, an axolotl will develop into an adult salamander. A genetic view of this kind of evolution is that there has been a change in the regulatory genes that switch on and off the protein-coding gene sequences within cells. If these genes start to operate at new rates, then the phenotype will change shape, in some cases dramatically.

It is now known that some genes, especially a group known as *Hox* genes, control development by instructing the different parts of the growing embryo on which part of the body should be built. It is known that these genes are more common in vertebrates than in other groups of animals and that there was a single period when these genes duplicated (or rather, duplicated twice), so that vertebrates carry four times as many of these genes as do invertebrates. This multiplication occurred between the evolution of the cephalochordates and proper vertebrates, probably during the Cambrian period. It is tempting to assume that this evolutionary event facilitated the increase in complexity needed to produce vertebrates and may have made them more 'evolvable' since. Whether or not cause and effect can be proved in this example, it points to a growing understanding of the relationship between genes and macroevolution.

See Also

Biodiversity. **Biological Radiations and Speciation**. **Famous Geologists:** Darwin; Lyell. **Fossil Invertebrates:** Trilobites. **Origin of Life**. **Palaeozoic:** Cambrian. **Precambrian:** Eukaryote Fossils.

Further Reading

Darwin C (1859) *On the Origin of Species*. Penguin Books (edited by J W Burrow).

Eldredge N and Gould SJ (1972) Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf TJ (ed.)

Models in Paleobiology, pp. 82 115. San Francisco: Freeman, Cooper.
Gingerich PD (1985) Species in the fossil record: concepts, trends and transitions. *Paleobiology* 11: 27 41.
Greenwood PH (1974) *Cichlid Fishes of Lake Victoria, East Africa: The Biology and Evolution of a Fish Flock*. London: The British Museum (Natural History).
Johnson ALA and Lennon CD (1990) Evolution of gryphae ate oysters in the Mid Jurassic of Western Europe. *Palae ontology* 33: 453 485.
Ridley M (1996) *Evolution.* Oxford: Blackwell.
Sheldon PR (1987) Parallel gradualistic evolution of Ordo vician trilobites. *Nature* 330: 561 563.
Skelton PW (ed.) (1993) *Evolution: A Biological and Palae ontological Approach*. Wokingham: Addison Wesley Publishing Company.
Van Valen L (1973) A new evolutionary law. *Evolutionary Theory* 1: 1 30.
Williamson PG (1981) Palaeontological documentation of speciation in Cenozoic molluscs from Turkana Basin. *Nature* 293: 437 443.

FAKE FOSSILS

D M Martill, University of Portsmouth, Portsmouth, UK

Introduction

This article considers fake fossils and the part that forgers have played in ruining scientific reputations and hindering the development of science. In some cases, financial reasons appear to have been a motive for forgery, but the impact has, nonetheless, been detrimental to science. This article is written from a certain amount of experience, the author having fallen foul of at least one fossil fabrication. Some tips are provided for those who might encounter fake fossils.

Forged fossils fall into a number of categories. Some are complete fabrications and should be considered as sculptures. They may be cast from materials that resemble rock, such as plaster or cement, or carved from real rocks. Some forgeries represent embellishments of genuine fossils, and include the addition of elements from another fossil simply to make an incomplete specimen appear more complete. Other forgeries are conversions whereby a common fossil is transformed to resemble something much rarer; others are chimeras whereby two or more fossils of different animals are united to produce quite fabulous creatures. Some composites are not manufactured deliberately to deceive; rather, many simply represent attempts to fill gaps for aesthetic purposes and to make museum displays more informative. In this latter case, no deceit is intended, but when past curators have failed to keep records of which fossils were amalgamated, taxonomic problems have arisen several years later.

In some unusual cases of forgery, remains of modern animals and plants are transmogrified into fossils by being embedded in resins or by being glued onto bedding planes (Figure 1). Deciding what constitutes a fossil forgery can be difficult. Purists might argue that any modification of a fossil represents an act of forgery, although a museum display specimen might be enhanced simply to demonstrate what a skeleton may have looked like when complete, or a damaged piece might be skilfully repaired to obscure an ugly scar or hole, perhaps caused by bad collecting practice. Certainly, Victorian museum curators thought it perfectly acceptable to construct a complete skeleton from the remains of a number of partial skeletons. One of the most famous examples includes the mounted skeleton of the giant sauropod dinosaur *Brachiosaurus brancai* that forms the centrepiece to the Humboldt Museum in Berlin. This magnificent skeleton is thought to contain the parts of at least five different individual fossils.

Cruel Hoaxes

Fake fossils represent deliberate attempts by the unscrupulous to hoodwink the unsuspecting into believing that an object is a genuine fossil. Such is human nature that as long as fossils have a financial value or can result in prestige for the discoverer or describer, then there are going to be disreputable people prepared to exploit this for their own ends, be they greed, spite, or self-betterment. This is not a new phenomenon, and has been a practice from the earliest days of palaeontology. Some faking of fossils is indeed a consequence of criminal intent to obtain money through deception, but in a number of cases, fossils have been faked in what appear to have been either jokes that have gone seriously wrong or deliberate attempts to ruin scientific reputations.

Such is the case of the now famous lying stones of Eibelstadt, near Wurzburg, Germany. This is one of the oldest, well-documented cases of fossil forgery,

Figure 1 In this crude attempt to forge a fossil, a recently dead dragonfly has been glued to the surface of a piece of limestone. Such forgeries at first can appear to be examples of excellent preservation. Be alert if a thin veneer of varnish prevents direct access to the surface of the fossil.

and concerns a certain Dr Johann Beringer. Apparently, Beringer was an extremely pompous fellow, and was despised by a number of local academics. The academics generated an enormous number of crude forgeries that they passed to Beringer via hired helpers. Despite the crude nature of the fossils, and even despite later attempts by the forgers to reveal their cruel act, Beringer published a scientific account of the fossils in 1726. The book, *Lithographiae Wirceburgensis*, figured hundreds of the faked specimens, which included sculptures of spiders in their webs, frogs, birds, and even shooting stars and moons. The flagrant act of forgery came to light just before publication, but the book includes a note dismissing the claims of forgery, such was Beringer's arrogance.

Perhaps the most famous example of deception is the case of the Piltdown Man forgeries. Many books have been written detailing this hoax and speculating on the identity of the perpetrator. Essentially, the lower jaw of an orang-utan was substantially altered and buried in a gravel pit along with some fragments of human cranium in Sussex. The remains were discovered by Charles Dawson in 1912 and then described by leading vertebrate palaeontologist Dr Arthur Smith Woodward, who was, at the time, the Keeper of the British Museum of Natural History, London. Although several observers had wondered if the fossil was a forgery, it was not definitively shown to be so until 1953, as a result of a fluorine analysis on the jawbone. This was a sad postscript to the scientific career of Sir Arthur, who had been a brilliant palaeontologist. The hoax was a scandal for British science, and it held back palaeoanthropology for several decades.

At least the Piltdown forgery concerned genuine organic remains and Arthur Smith Woodward could perhaps be forgiven for making a mistake; after all, someone had set out deliberately to deceive. In 1966, at the age of 91, noted German vertebrate palaeontologist Professor Freiderick von Huene, of Tübingen University, described a juvenile skeleton of the ichthyosaur *Leptonectes* (then called *Leptopterygius*) that had been made from cement, stained brown, and placed on a slab of rock from the Early Jurassic Posidonia Shale Formation. Not a single fossil bone was present on the specimen; it was nothing more than a sculpture, and not a very accurate one at that. Proof of this forgery came to light only when the sculpture was being cleaned by a preparator some 4 years later, and it was not revealed to the scientific world until 1976. Huene never had to face the embarrassment of this exposé because he died in 1969, and in this particular case, the published paper of Huene was not of great scientific consequence. This is in marked contrast to the paper of Arthur Smith Woodward, on the Piltdown 'fossils', which announced the presence of the oldest hominid fossils in Europe and purported to show that large human brains were an early evolutionary development. But there are some similarities in the two hoaxes. In both cases, the scientists concerned were extremely eminent and had enjoyed careers in which they had risen to the very top of their profession. It would be no surprise to learn that they had made enemies on the way up, and that some embittered rascal had sought cruel revenge. These, fortunately we hope, are rare cases.

Too Much Haste

A more recent (November 1999) case of fossil forgery resulted in considerable embarrassment for North American palaeontologist Philip J Currie, artist Stephen Czerkas, and especially for the senior assistant editor of *National Geographic Magazine*, Chris Sloan. This sorry story concerned a strange case whereby two spectacular, and quite genuine fossils, were merged together to construct a chimera comprising the back end of a small dinosaur, *Microraptor zhaoianus*, and the front end of a small fossil bird, *Yanornis martini*, both from the famous Early Cretaceous Yixian Formation of Liaoning Province, China. The two incomplete specimens were joined together to make a single, complete feathered dinosaur. Unfortunately, so much excitement was generated over the specimen that the *National Geographic Magazine* printed an article on its discovery and its perceived relevance to the 'birds are dinosaurs debate' just before the specimen was shown to be a forgery. An even more unfortunate aspect of this case occurred because, unusually for an article in the *National Geographic*, the fossil chimera was given a scientific name, *Archaeoraptor liaoningensis*, which, according to the rules of scientific nomenclature, was valid for at least part of the specimen. Paradoxically, it turned out that both halves of the chimera represented important scientific discoveries, and both were new to science. Suspicions surrounding the nature of the fossil came to light when the specimen was scanned using computed axial tomography (CAT), and it became clear that the pieces did not fit together well. A more careful examination then revealed the forgery, and although Phil Currie highlighted some problems with the fossil, these were not relayed to *National Geographic*. It was only when Chinese palaeontologist Xu Xing had met with a Chinese fossil dealer that the sorry story of the forging really emerged. But by then it was too late; the article had already appeared in the November 1999 issue and the proof of the forgery came one month later. It is to the

relief of many palaeontologists that this forgery was discovered before too much damage had been done to the scientific case being made for the bird–dinosaur hypothesis, but, sadly, the furore over the forgery has distracted from the undoubted importance of the two genuine fossils.

Not all fossil forging has serious consequences for science, and indeed, some forging is undertaken in an attempt to 'improve' fossils for the fossil-buying public. Such forgeries usually involve adding embellishments to genuine fossils, or converting fossils from one type to another (**Figure 2**). Such embellishments do not enhance the value of the fossil; indeed, they destroy the scientific value, but they might make a fossil look attractive to the unwary purchaser. This type of forgery is common among the fossil dealers of Brazil who raid the spectacular fossil fish beds of the Santana and Crato formations of north-eastern Brazil. Here it is common to find forged fishes that have heads and tails belonging to different species. Fins may be added, and some specimens might be artificially lengthened by the insertion of several bodies into one example. Conversions are common, and it is frequent to find heads of large fossil fishes converted into frogs, and small specimens of the gonorhynchiform fish *Dastilbe* converted into lizards. To the unwary, the presence of some genuine bones is enough to encourage belief that the entire fossil is genuine.

Until recently, most of the fossils available commercially from the fossil beds of Brazil were collected to supply flea markets in the tourist centres of Brazil. The fossils were often enhanced to make them visually more attractive to tourists who probably knew very little about fossils, but who wanted to have an unusual souvenir of Brazil (**Figure 3**). More recently, the genuine fossils have become highly sought after by museums, and many of the rarer fossils from Brazil, such as pterosaurs and dinosaurs, command very high prices. There has thus arisen a new financial incentive for the forging of fossils. Previously, forgeries were rapidly constructed, using a sharp chisel, by the addition of a crudely engraved outline of a fish. Now, elaborate constructions are made by glueing together numerous pieces of real fossils to produce such things as pterosaur bones and crocodile heads. Much time and effort goes into these constructions, but by and large they remain crude and are easy to recognise. However, in a skull of a dinosaur that had been obtained by the Museum für Naturkunde, Stuttgart, Germany, a sagittal crest at the back of the skull was revealed to be fake only after CAT scanning. The crest, in fact, was a part of the lower jaw repositioned to make the specimen look more spectacular. There was no need for the forger to have executed this

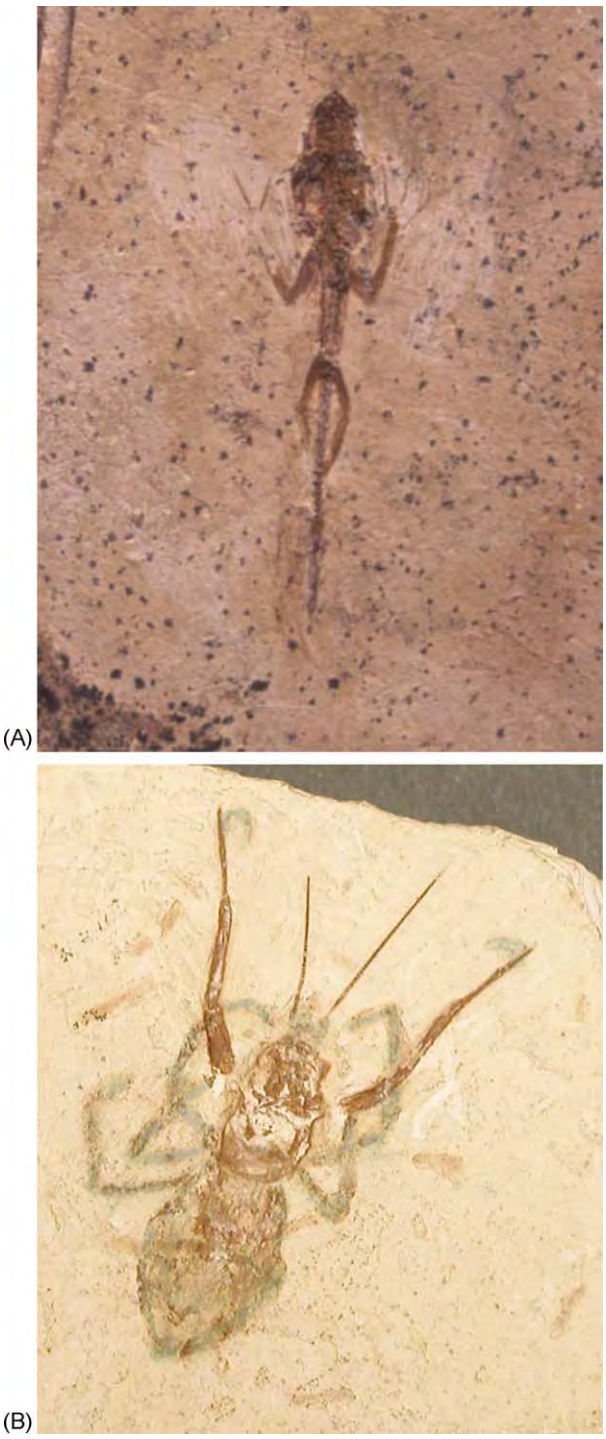


Figure 2 Examples of genuine fossils that have been altered to resemble something rare. Both are from the Nova Olinda Member of the Crato Formation, Ceará, Brazil. (A) The fossil gonorhynchiform fish *Dastilbe* has had limbs added to make it look like a lizard. (B) A fossil insect has had extra legs and claws added in ink to make it resemble something new, perhaps a spider. This fake was revealed by dropping industrial methyl ated spirit onto the fossil. The ink of the faked legs bled into the rock, whereas the real limbs remained intact. Both photographs by Robert Loveridge.

Figure 3 A forged composite group of fossil fishes. The fishes here are all exceptionally well preserved, but were not found in this association. They are from the famous Santana Formation, Ceará, Brazil. They have been glued together with a mix of car body filler and ground rock.

embellishment, because the skull represented a new genus and species of dinosaur and represented the most complete skull of a spinosaurid dinosaur ever found.

Amber

Amber has long been famous for its fossil inclusions, and has been used in the jewellery industry with or without fossil inclusions for thousands of years. New discoveries of amber in the Dominican Republic have resulted in a large number of forgeries. A majority of these are offered to unsuspecting tourists. Most are sold cheaply, but a number of higher priced specimens containing lizards and frogs have proved to be cleverly executed forgeries. It is not always easy to distinguish forged amber from the real thing. The hot needle test, whereby a red-hot needle is pressed into the specimen, will give off a resinous smell if the specimen is genuine amber, whereas the smell will be acrid if it is a synthetic resin; however, the test inevitably marks the specimen.

Religious Zealots

There have been several attempts by those creationists (*see* **Creationism**) who appear to feel threatened by palaeontological and geological evidence that runs counter to biblical interpretations that Earth is not terribly old and was created in a very short span of time. Rather than accept the findings of science, some supporters of a biblically based creation theory have challenged the data on which certain scientific claims are made by attempting to discredit palaeontology. Attempts to do this using logical argument have proved difficult, and so some unscrupulous individuals have attempted to undermine scientific findings by forging data. Perhaps the most notable attempt was the claim that human footprints occurred alongside those of dinosaurs at the Paluxy River site in Texas, USA. The Paluxy River site is famous for lengthy trackways of footprints of Cretaceous dinosaurs and has been made into a National Park. Reports that human footprints had been found in the same layers as the dinosaur footprints had always been treated with scepticism by the palaeontological community. Wrapped in pseudoscientific jargon, photographs of the human footprints side by side with dinosaur footprints were used as 'evidence' that humans and dinosaurs were around at the same time, and it would therefore have followed that dinosaurs could not have been millions of years old. Despite considerable protestations by scientists, it was only later that the perpetrators of the forgeries admitted that the human prints were handmade rather than footmade.

When a Fossil is Not a Forgery

One of the most important fossils, historically, is the London specimen of the small feathered bird/dinosaur *Archaeopteryx lithographica*. This fossil was

widely hailed as a missing link between reptiles and birds because its exceptional preservation in the fine-grained lithographic limestones of Bavaria showed it to have a dinosaurian skeleton that included a long tail with numerous vertebrae and a beak in which the jaws possessed teeth. And yet the animal was clothed in feathers, a feature known today only in birds. This was just what devotees of Darwin needed to support the theory of evolution, and indeed *Archaeopteryx* became the archetypal missing link; it is an animal that appears to be transitional between two groups of animals, a status that is still claimed for it today. During the 1980s, two eminent scientists, but not palaeontologists or geologists, Sir Fred Hoyle and Chandra Wickramasinghe, published a claim that the London specimen of *Archaeopteryx* was a forgery. If their claim had been correct, there is no doubt it would have had important implications, but such an upheld claim would have had even more dire consequences had it been made in Victorian London. However, several independent analyses of the evidence for forgery showed quite categorically that Hoyle and Wickramasinghe were out of their depth and did not understand the nature and diversity of fossilization processes. Nevertheless, the claim, coming as it did from such noted scientists, generated considerable excitement in the media, and a number of books and papers resulted from the claim. Sadly, many palaeontologists had to devote considerable time and effort to debunk these incorrect claims.

Detecting Forgeries

It is advisable always to be suspicious of fossils bought commercially and to be very sceptical of any exceptional fossil that is provided by a 'friendly' noncolleague. Fossils traded commercially are quite likely to have been enhanced in order to increase their aesthetic appeal, but such improvements are usually easily detected by experienced palaeontologists. Some traders of ammonites increase the size of the ammonite by carving extra whorls into the rock.

Fabricated parts of fossils are often constructed using plastic-based fillers. These plastics will melt when probed with a hot needle, giving off an acrid smoke. Rock does not do this. Many of the spectacular trilobites from the Ordovician and Devonian of Morocco have been enhanced with fillers, and indeed some are simply casts made from moulds of genuine fossils. The casts are glued to blocks of limestone and coloured black with boot polish.Where there has been a real intent to deceive, the workmanship of the forgery is often very good and almost impossible to detect by casual inspection. Examination under a microscope may help, but when suspicions are raised, proof may come only after expensive CAT scanning or chemical analysis.

See Also

Creationism.

Further Reading

Charig AJ, Greenaway F, Milner AC, Walker CA, and Whybrow PJ (1986) *Archaeopteryx* is not a forgery. *Science* 232: 622 626.

Hoyle F and Wickramasinghe C (1986) *Archaeopteryx: The Primordial Bird*. Swansea: Christopher Davies.

Martill D (1994) Fake fossils from Brazil. *Geology Today* 1994: 36 40.

Nield T (1986) The lying stones of Eibelstadt. *Geology Today* 1986: 78 82.

Ross A (1998) *Amber: The Natural Time Capsule*. London: The Natural History Museum.

Russell M (2003) *Piltdown Man: The Secret Life of Charles Dawson*. Stroud: Tempus Publishing.

Sloan CP (1999) Feathers for *T. rex*. *National Geographic* 196(5): 98 107.

Simons LM (2000) *Archaeoraptor* fossil trail. *National Geographic* 197: 128 132.

Suess H D, Frey E, Martill D, and Scott D (2002) *Irritator challengeri*, a spinosaurid (Dinosauria: Theropoda) from the Lower Cretaceous of Brazil. *Journal of Vertebrate Paleontology* 22: 535 547.

Wild R (1976) Eine Ichthyosaurier Fälschung. *Neues Jahrbuch für Geologie und Paläontologie, Monatshefte* 1979: 382 384.

FAMOUS GEOLOGISTS

Contents

Agassiz

D R Oldroyd, University of New South Wales, Sydney, Australia

Son of a clergyman, Jean Louis Rodolphe Agassiz (Figure 1) was born in the village of Môtier in Canton of Fribourg, Switzerland. After schooling at Bienne and Lausanne and early acquiring an interest in natural history (particularly of fishes), he attended the universities of Zurich, Heidelberg, and Munich, intending to take a medical degree. However, at Heidelberg he began studying palaeontology under Heinrich Bronn and embryology under Friedrich Tiedemann. He also became friends with Alexander Braun and his family (later marrying his sister Cécile) and Karl Schimper, and the three young men went on botanical excursions together. At Munich, he came under the influence of Friedrich Wilhelm Schelling, Lorenz Oken and German *Naturphilosophie*. Additionally, he studied botany under Carl Friedrich von Martius, and embryology under Ignatius Döllinger. A natural philosopher should, as Agassiz represented Oken's view in his autobiography: "[construct] the universe out of his own brain, deducing from *a priori* conceptions all the relations of . . . living things". In practice, Agassiz certainly did not discount empirical information. Indeed, his hunt for 'facts' was one of his chief priorities. But this search was propelled by the desire to apprehend the activity of the Absolute Being in nature. Indeed, it was strongly influenced by the notion that he was examining the works of a divine Creator.

Agassiz's PhD (Erlangen/Munich) (*Selecta Genera et Species Piscium quos in Itinere per Brasiliam Annis MDCCCXVIII–MDCCCXX* [1829]) was devoted to the Brazilian fishes collected from Amazonia by the recently deceased Johann Baptist von Spix. It was tactfully dedicated to Georges Cuvier (*see* **Famous Geologists:** Cuvier). Agassiz also obtained an MD at Munich in 1830, but by then he was determined to be a naturalist not a physician. The following year Agassiz went to Paris to study comparative anatomy under Cuvier, having already examined numerous collections of fossil fish in leading museums. Cuvier was greatly impressed by the young man's work and took him under his wing, introducing him to Alexander von Humboldt, and teaching him the principles of comparative anatomy and how to reconstruct fossil fish. So Agassiz gave up the German idea of the unity of the animal kingdom and followed Cuvier's notion of there being four fundamental types in the animal kingdom. Cuvier was so impressed by Agassiz's abilities that he passed on the notes, drawings, and specimens that he had collected on fossil fish for him to study. He also ensured that other institutions made their collections available to Agassiz.

Figure 1 Louis Agassiz (1807 1873).

Cuvier died of cholera in 1832, but his influence on Agassiz was strong and permanent, particularly respecting the idea of successive geological catastrophes and the creation of new species. Agassiz's studies of fossil fish eventually yielded his great treatise *Recherches sur les Poissons Fossiles* (5 vols, 1833–1843), with the figures mostly drawn by the artist Joseph Dinkel (whom he employed over a long period); and *Monographie des Poissons du Vieux Grès Rouge* [*Old Red Sandstone*] *ou Système Dévonien des Iles Britanniques et de Russie* (1844–1845). Agassiz received the Geological Society's Wollaston Medal for his ichthyological work in 1836.

At a youthful 25 years of age, Agassiz was appointed Professor of Natural History at the small new Lyceum or Academy at Neuchâtel, back in his home region of Switzerland, and soon began to establish that institution's reputation. His early *magnum opus* made use of specimens sent to him from all over Europe, and in particular from the Old Red Sandstone of Scotland, to which country he made two visits. The later association with the amateur stonemason Hugh Miller, who arranged for Agassiz to receive specimens of Devonian fossil fish, is particularly well known through Miller's popular book *The Old Red Sandstone* (1841), and his contributions were incorporated into Agassiz's work on Devonian ichthyology.

Unfortunately, Agassiz's first marriage to Cécile Braun failed, in part because he gave so much attention to his work and partly because he came under the sway of his assistant, the geologist Edouard Desor, who pushed his way into the Agassiz household despite Cécile's objections. Moreover, Agassiz's ambitious publishing projects led to financial problems and life became difficult for him in Neuchâtel. He therefore sought the assistance of von Humboldt and Charles Lyell (*see* **Famous Geologists:** Lyell) to travel to North America, and in 1846 he went to Boston at the invitation of James Avory Lowell to give a lecture series on natural history. These were outstandingly successful, and led to his appointment as Professor of Zoology and Geology at Harvard in 1848, where he soon became one of the country's leading scientists. In 1852, he was additionally professor at the Medical School at Charleston, South Carolina, and also at Cornell University in 1868.

Declining a chair in Paris, despite the offer of most favourable terms, Agassiz committed himself to American science, pushing, with the help of endowments from Francis Calley Gray and others, for the foundation the famous Museum of Comparative Zoology at Harvard in 1858–1859 (which opened in 1860). In 1863, he helped persuade Abraham Lincoln to establish the National Academy of Sciences; the same year Agassiz was appointed a regent of the Smithsonian Institution. Agassiz had reached the top of the tree.

Subsequently, Agassiz travelled widely on both land and sea and wrote numerous scientific papers in the USA, as well as popular essays, reviews, and educational works, his writings on classification being the most influential. However, while revelling in the hospitality and opportunities that America offered, he retained a belief in the superiority of European science and culture, which later alienated some colleagues.

Left behind in Europe, Agassiz's wife had died of tuberculosis in 1848. His son joined him in America, and subsequently his two daughters. In 1850 he married Elizabeth Cary, who later founded a girls' school that later developed into Radcliffe College at Harvard. A number of Agassiz's European *epigone* followed him to America, including Desor, with whom Agassiz eventually fell out, after an unpleasant episode involving accusations of plagiarism, financial malfeasance, and worse, for which Agassiz was found to be without fault. His first wife's intuitions were more than vindicated. Desor withdrew to Europe.

Apart from collecting, naming, and describing modern and fossil fish, Agassiz also proposed a scheme for fish classification, based on their scales. This was not 'biologically' ideal, but suited the study of fossil fish, for which in many cases the scales are the best preserved remains, the bones having been cartilaginous. Thus four main orders of fish were proposed, based on their scales, rather than their crania:

1. Having plate-like scales, often with tubercles or bony points, detached from one another and irregularly arranged on a tough skin (Placoid).
2. Having large, bony, usually shiny (enameled) plate-like scales, not normally overlapping, but often interlocking in some way (Ganoid).
3. Having thin, horny, overlapping plates, each having one side with a jagged edge or comb-like projections (Ctenoid).
4. Having thin, flexible, overlapping, horny scales, smooth in outline and circular or elliptical in form (Cycloid).

For the Ganoids, Agassiz was especially interested in the modern *Lepisosteus*, which was the sole surviving modern representative of the group. So, like Cuvier (who worked on elephants, mammoths and mastodons), he specifically sought to compare living and extinct types.

Agassiz's taxonomic system was later superseded by various others, based principally on bones rather than scales, though his introduction of the Ganoidei was a substantial contribution. His taxonomy was problematic, for, while comparing fossilized and modern forms, he saw no evolutionary connection between them. On the other hand, he offered something new by the use of fossil fish for stratigraphic purposes. Moreover, in a manner that would have appealed to Cuvier, he sought to find out something about the 'conditions of existence' of his specimens as to temperature, salinity, and mode of locomotion. He supposed that prior to the Cretaceous there seemed to be less distinction between fresh-water and marine forms than at present and it might be the case that these two environments were not so marked previously as they are at present.

However, Agassiz's most important contribution to geology was his advocacy of the concept of an Ice-Age (*Eiszeit*), fundamental to Pleistocene geomorphology and stratigraphy (*see* **Tertiary To Present:** Pleistocene and The Ice Age). Curiously, it was linked to the biological ideas that he imbibed from Cuvier.

In Switzerland, the idea that the country's glaciers were formerly of greater extent had been recognized by observers back in the eighteenth century, such as the minister Bernard Friedrich Kuhn (1787). There is a report of a manuscript by a mountaineer Jean-Pierre Perraudin (1818), which described the extent of moraines and erratic boulders, and regarded striated and polished rocks as evidence of glacial action. It was perhaps Perraudin who really initiated the glacial theory in Switzerland. The highway engineer Ignaz Venetz accepted Perraudin's ideas and read a paper on the topic at Neuchâtel in 1829. The mining engineer Jean de Charpentier, director of the salt mines at Bex, also obtained information from Perraudin and in 1834 read a paper at Lucerne about the former greater extent of glaciers. (Agassiz met Charpentier when he was still at school and was partly inspired by him to become a naturalist.) However, Charpentier's paper was regarded as mistaken and was mocked, Agassiz being one of the opponents. (Historians examining Agassiz's students' lecture notes from that period have shown that he was then critical of the theory.)

But in 1836 Agassiz was in the Bex area and was shown around by Charpentier, and after calling on Venetz and examining the evidences in other parts of Switzerland he became a convert to the theory. While in Bex, Agassiz met his old student friend, the botanist and palaeontologist Karl Schimper, and the two also discussed the glacial evidence. In February 1837, Schimper gave a botanical talk at Neuchâtel, at the conclusion of which he passed round a copy of a poem that introduced the new word *Eiszeit*. By then, Agassiz had picked up the evidences and ideas in their entirety and was running with them. He presented a first outline of his views in public at the meeting of the *Société Helvétique des Sciences Naturelles* at Neuchâtel in July 1837, in what became known as the *Discours de Neuchâtel*. By 1840 Agassiz published his major study on the topic, and his most important contribution to geology: *Études sur les Glaciers*. In publishing this, he got ahead of Charpentier's *Essai sur les Glaciers* (1841), and recriminations followed, stirred up, it has been suggested, by Desor. Schimper was also annoyed with Agassiz for failing to mention him in *Études* (though he was mentioned in the *Discours*). There followed a further work on glaciers co-authored with Arnold Guyot and Desor, describing the different types of glaciers, their component parts, their motions, and a detailed account of the Aar Glacier: *Système Glaciaire: Ou Recherches sur les Glaciers, leur Mécanisme, leur Ancienne Extension et le Rôle qu'ils ont Joué dans l'Histoire de la Terre* (1847).

The *Discours* was written in haste, but provided persuasive evidence for the former extension of glaciers, at least in Switzerland, and strong arguments against the floating iceberg theory favoured by Lyell in Britain, or the common idea of glacial erratics being emplaced by catastrophic floods. On the other hand, Agassiz thought that erratic boulders might have fallen into their present positions rather than being directly transported by ice. Agassiz's *Études* was a sumptuous volume, beautifully illustrated, providing all the documentation necessary to convince readers of the former extension of glaciers. The theory could also explain the existence of the vast extent of superficial deposits ('till') over northern Europe, then

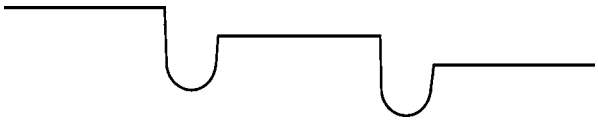


Figure 2 Sketch in Agassiz's *Discours de Neuchâtel* showing hypothetical fall of global temperature over time.

known as 'diluvium' (by association with the notion of catastrophic floods) or 'drift' (by association with the floating iceberg theory).

However, the explanatory theory advanced by Agassiz was much less persuasive than that which it was supposed to explain. It was widely held at the time (in accordance with the views of Élie de Beaumont, e.g.) that the earth was cooling. Agassiz arbitrarily assumed that it did so in a fashion indicated by his sketch (Figure 2), which appeared in both the *Discours* and the *Études*.

This graph was supposed to represent not just a cooling inorganic planet, but one inhabited by living organisms that were wiped out, however, from time to time by 'Cuvierian' catastrophes, and then replaced by different sets of organisms. The abrupt falls of temperature corresponded to the sudden disappearance of life forms, with the temperatures remaining approximately constant so long as the (supposedly heat sustaining) life forms continued in existence. Agassiz supposed that the formation of the Alps themselves was an event of recent occurrence, was preceded by a 'catastrophic' fall in temperature, and was then followed by the establishment of modern forms of life. On this view, then, the epoch preceding the present could have been of extreme cold, producing the former extended glaciation evidenced in the Alps. The onset of cold must have been sudden, from the appearance of mammoth remains in Russia. Agassiz suggested that the glaciation could have extended from the North Pole right down to the Mediterranean and Caspian seas. Thus the Great Ice-Age.

This theory was perhaps the most 'catastrophist' ever propounded by a 'respectable' geologist (other than bolide aficionados). Agassiz wrote (1838: 382):

> [T]he epoch of extreme cold which preceded the present creation . . . was attended by the disappearance of the animals of the diluvian epoch of geologists, as the mam moths of Siberia still attest, and preceded the uprising of the Alps, and the appearance of the animated nature of our day, as is proved by the moraines, and the presence of fish in our lakes. There was thus a complete separ ation between the existing creation and those which have preceded it; and, if the living species sometimes resemble in our apprehension those which are hid in the bowels of the earth, it nevertheless cannot be affirmed that they have regularly descended from them in primogeniture, or, what is the same thing, that they are identical species.

Thus Agassiz set his face against transformism or evolution and offered hyper-catastrophism and the doctrine of special creations (assuming but not then stating) that they occurred by some divine means.

Agassiz visited Britain in 1834 and 1835, chiefly in connection with his interests in fossil fish, but he also made the acquaintance of the 'diluvialist' William Buckland, who in turn visited Agassiz in Switzerland in 1838. Buckland had long been interested in the drift deposits, which he earlier has ascribed to the Noachian Flood, and introduced the distinction between 'diluvium' and 'alluvium'. He was, however, converted to Agassiz new theory during the course of his 1838 visit, realizing that features of British geology that had long puzzled him could be successfully explained in terms of the land-ice theory. In 1840, Agassiz attended meetings of the Geological Society in London and the British Association in Glasgow, and presented his glacial theory, prompting much discussion in British geological circles. However, the theory, as presented in Glasgow, tried to reconcile the new doctrine with the older idea of glacial submergence, for after the melting of the glaciers flood waters could have moved boulders and gravels (thus accounting for glaciofluvial materials).

Following the meeting, Agassiz and Buckland went on a tour of Scotland, and were satisfied that they could see most satisfactory evidence in favour of the land-ice theory, and successfully interpreted the 'Parallel Roads of Glen Roy', which Darwin had the previous year interpreted as marine shore-lines, as being due to the successive shore lines of an ice-dammed lake, an interpretation that was rapidly published in the newspaper *The Scotsman*. Following his Scottish tour, Agassiz proceeded to Ireland, where again he found ample evidence for glaciation. Returning to Scotland, he then journeyed back to London, seeing many more evidences of glaciation, and spoke at the Geological Society.

Debates about the land-ice theory rumbled on in Britain for the next quarter century. Lyell was initially converted to Agassiz's ideas, but most other influential geologists such as Roderick Murchison (*see* **Famous Geologists:** Murchison) were not. Not long after Agassiz returned to Switzerland, Lyell recanted: "he found the proposed departure from present temperature conditions too much to accept for his uniformitarianism, and he reverted to the glacial submergence theory and floating icebergs". It was not until the 1860s that more general acceptance of the land-ice theory began, with the suggestions of the surveyor Andrew Ramsay as to how glaciers might excavate the basins that are now

occupied by mountain lakes; and land ice could have moved uphill to deposit marine shells on the tops of hills in North Wales. Eventually, through the theoretical work of James Croll (1875), an astronomical theory of the origin of climate change was developed, and such theory has been under discussion through to the present.

In America, Agassiz successfully applied his ideas to the interpretation of observations in the Great Lakes area, which he explored in 1848, finding new fish for his examination and ample evidences of glaciation (see his *Lake Superior; Its Physical Character, Vegetation, and Animals, Compared with those of Other and Similar Regions* [1850]), and many other regions. But most of his work in the USA was zoological rather geological. In particular, and in keeping with his long-held Cuvierian views, he was active in his criticism of Darwin's evolutionary theory, which ran counter to Agassiz's long-held beliefs about the special creation of life forms. Agassiz even resisted the idea that different varieties of animals of the same species could be produced through time, from which it followed that the different human races were essentially different species! Thus, he gave 'scientific' comfort to racial bigots. Agassiz could not comprehend how similar but different creatures of the same species could have been produced worldwide. As a special creationist, that seemed to him to be the required alternative to his 'polygenism', and as such had to be rejected. Agassiz's views in fact succeeded in driving Lyell further into the evolutionist camp. Also, because he was opposed to the idea of variation over time, Agassiz was inclined to suppose that every variety of fish he encountered represented a different species. Hence his classification became inordinately unwieldy. It is interesting that a figure, published in 1844, depicting the genealogy of his four main groups of fish, looks quite like a modern evolutionary tree, yet none of the 'branches' are shown as linking at their bases, though they 'lean towards' one another, so to speak, in a way that a later evolutionist might regard as suggestive.

The source of Agassiz's anti-evolutionism can be traced to his contacts with Oken and German *Naturphilosophie*, and associated Platonism (fused with Christian beliefs), as well as Cuvier. Species, for Agassiz, could be regarded as 'types' representing the 'thoughts' of the Creator. Because there could be no substantial natural variation over time, events such as the Ice-Age represented catastrophes of divine origin that also offered the possibility of renewed creative activity. The 'plan' of Creation was, he supposed, better understood by the natural historian than the theologian.

Agassiz was not a great geologist, despite his outstanding capacity for grasping and ordering information, and his powers as a teacher. In 1865–1866, he visited South America, funded by a wealthy Bostonian, Henry Thayer, hoping to find evidences of glaciation in the tropics. Seriously perturbed by Darwin's theory, Agassiz sought new evidence to support of his long-held ideas about catastrophes and the great Ice-Age. He wished to show that the event was of worldwide extent: so it should be possible to find evidence for it in the southern hemisphere, even in Amazonia. In Brazil, he thought he had found the evidence he sought, but he mistook boulders produced by tropical weathering for glacial erratics, and soil produced by weathering was misidentified as glacial till. His co-workers did not all agree, but Agassiz thought he had the experience and expertise to recognize glacial evidence when he saw it (though he admitted he saw no glacial striations).

Agassiz's attempt to extend his Ice-Age to equatorial regions was a failure and provided a classic example of 'theory-laden' observations. On the other hand, his recognition and advocacy of the concept of a glacial epoch and the land-ice theory (even if not original to him) was of fundamental importance, marking the beginning of glaciology and all that followed in the study of Pleistocene geology.

See Also

Creationism. **Evolution**. **Famous Geologists:** Cuvier; Darwin; Lyell; Murchison. **Fossil Vertebrates:** Fish. **History of Geology From 1835 To 1900**. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Agassiz L (1887) *Geological Sketches*. New York: Houghton, Mifflin & Co.

Agassiz L (1967) *Studies on Glaciers Preceded by the Dis course of Neuchâtel Translated and Edited by Albert V. Carozzi*. New York and London: Hafner Publishing Company. (This volume contains an English translation of Agassiz's *Discours de Neuchâtel*.)

Andrews SM (1982) *The Discovery of Fossil Fishes in Scotland up to 1845 with Checklists of Agassiz's Figured Specimens*. Edinburgh: Royal Scottish Museums.

Brice WB and Figueirôa SFdeM (2001) Charles Hartt, Louis Agassiz, and the controversy over Pleistocene glaciation in Brazil. *History of Science* 39: 161 184.

Carozzi AV (1973) Agassiz's Influence on Geological Thinking in America. *Archives des Sciences Genève* 21: 5 38.

Davies GL (1969) *The Earth in Decay: A History of British Geomorphology 1758 1878*. London: Macdonald Technical and Scientific.

Gaudant J (1980) Louis Agassiz (1807 1873), fondateur de la paléoichthyologie. *Revue d'Histoire des Sciences* 33: 151 162.
Lurie E (1960) *Louis Agassiz: A Life in Science*. Chicago and London: Chicago University Press.
Marcou J (1896) *Life, Letters, and Works of Louis Agassiz,* 2 vols. New York: Macmillan (reprinted Gregg International, 1971).
North FJ (1943) Centenary of the glacial theory. *Proceedings of the Geologists' Association* 54: 1 28.

Cuvier

G Laurent, Brest, France

Introduction

Georges Cuvier was one of the grand masters of zoology in the first third of the nineteenth century. He laid the foundations of vertebrate palaeontology, and his work led to the development of the idea of stratigraphical stages through the work of Alcide d'Orbigny. Cuvier's rivalry with Lamarck, the founder of invertebrate palaeontology, manifested itself in their disputes over the subjects of catastrophism in the history of the Earth and transformism in the history of life.

Biography

Jean-Léopold-Nicolas-Frédéric (called Georges) Cuvier was born on 23 August 1769 in Montbéliard, which at that time belonged to the Duchy of Wurttemberg (Germany) but retained French as its language. At an early age he showed an interest in the study of the natural world. As he came from a Protestant family, his parents intended that he should become a pastor, but he failed his entrance examination to the seminary. Nevertheless, he obtained a scholarship to the Caroline Academy in Stuttgart, where, during the years 1784 to 1788, he received training appropriate for a future official in the service of the Duchy. In accordance with his personal interests, he also attended courses in natural history. He became friendly with Christian Heinrich Pfaff (1772–1852) and more particularly with Karl Friedrich Kielmeyer (1765–1844), who was similarly devoted to zoology and who became Professor of Zoology at the Caroline. It was he who taught Cuvier the art of dissection and gave him his 'first ideas about philosophical anatomy'.

After failing to obtain a post in the bureaucracy at the end of his period of training, Cuvier found employment as a tutor to an aristocratic Protestant family in Normandy, where he spent the years 1788 to 1795, the most disturbed period of the Revolution. He devoted his leisure time to studying botany and the anatomy of animals, particularly molluscs, which he encountered in the neighbouring coastal area. Thanks to his friends Pfaff and Kielmeyer, Cuvier maintained his links with German naturalists.

In April 1795, with the assistance of the physician and agronomist Abbé Alexandre Tessier (1742–1837), a refugee at Fécamp, Cuvier was able to establish himself in Paris. He was well received there, particularly by Etienne Geoffroy Saint-Hilaire, who was already a Professor at the Museum and with whom he became friendly. They collaborated with one another and coauthored some articles. Upon his arrival, Cuvier obtained a teaching position at the newly established college at the Panthéon. The same year he was chosen by Antoine Mertrud to fill a vacancy at the Muséum d'Histoire Naturelle. This marked the beginning of Cuvier's distinguished teaching career, both there and in the university. He was named a Member of the First Class of the Institut de France (subsequently the Académie des Sciences) when it was formed in 1795. In 1800, he was appointed to Jean Daubenton's former chair at the Collège de France. In 1802, when Mertrud died, he became titular Professor of Comparative Anatomy at the Muséum d'Histoire Naturelle. In 1803, he became Permanent Secretary of the First Class of the Institut de France.

Simultaneously, Cuvier pursued an administrative career. In 1802, he was appointed Inspector General of Public Education. In 1808, Napoleon named him Councillor of the University, which he was re-establishing, and in 1810–1811 Cuvier was one of the leading lights in the reform of higher education, first in France and subsequently in Italy, Germany, and the Netherlands. In 1813, he was a Councillor of State as 'Maître des Requêtes'. The Restoration brought him still more honours. Louis XVIII appointed him Chief Councillor of Public Education and made him a Baron in 1819. The same year he was named President of the Section of the Interior in the State Council, representing the interests of non-Catholics. In 1824, Charles X conferred on him the honour of Officer of the Légion d'Honneur, of which he had been a Chevalier since the time of the Empire. Louis-Philippe

named him Pair de France in 1831. Cuvier was a member of three sections of the Institut de France: the Académie Française, the Académie des Sciences, and the Académie des Inspections et Belles-Lettres, as well as numerous foreign academies.

In 1803, Cuvier married the widow of the former fermier général Duvaucel, who had been guillotined in 1793. None of their four children survived, and their deaths caused Cuvier great distress. Cuvier died on 13 May 1832, at the height of his fame, after a short illness, the precise nature of which is unknown (although it may have been cholera or myelitis).

Cuvier's Work and Achievements

The eighteenth and nineteenth centuries were dominated by a desire to emulate the astronomical achievements of Newton in other areas of science. Cuvier aspired to be the Newton of natural history. He wanted to introduce into this field the approach that henceforth would govern all physical sciences: analyse facts, isolate them, compare them, and then try to ascertain general causes to explain the facts thus ordered according to common laws or principles. His best-known law – the one that made possible his fossil reconstructions – was the law of the correlation of organs or parts: all the parts of an organism must be suitably correlated so as to make a viable whole, capable of coping with the conditions of existence. He adopted the 'comparative' approach in the late eighteenth century when endeavouring to restore the remains of mastodons that had been sent to France from America for examination. The task was accomplished using anatomical analogies with modern elephants (for which he regarded the African and Indian types as being distinct, as were the remains of the Siberian mammoth). Applying this principle, Cuvier succeeded in reconstructing a large number of extinct forms. A single tooth, so to speak, told him everything about an organism, he triumphantly proclaimed apropos his reconstruction of the *Mosasaurus*. The immutable laws of zoology, with their wonderful constancy, which are not contradicted in any class or family, served Cuvier admirably in his arduous task of 'resurrecting' (his word) the past.

The notion of species obviously underpinned all attitudes towards, and classifications of, animated nature. It was one of the most clearly defined concepts in Cuvier's work. The most important concepts in nature were those of the individual and the species, and they were connected through the process of generation. Organized beings had two bases for natural classification: the individual, resulting from the common action of all the organs; and the species, resulting from the bonds created by the generation of individuals. From his earliest publications, and particularly in his *Tableau Èlémentaire de l'Histoire Naturelle des Animaux* (published in 1797), Cuvier gave a definition to which he remained steadfast:

> The collection of all organized bodies born one from another, or having parents in common, and all those that resemble them in the same way as they resemble each other, is called a species. [Cuvier G (1797) *Tableau Èlémentaire de l'Histoire Naturelle des Animaux*. Paris: Baudouin. p. 11]

But, in practice, in many cases – and whenever considering the past – one cannot use the descent of forms to define species. So, they must be classified by their distinctive external, and more particularly their internal, parts. Form becomes the prime consideration in the study of living bodies, and gives anatomy a role that is almost as important as that of chemistry.

Although Cuvier seems at times to have supposed that there was really nothing in nature other than the species and the individual, nevertheless the study of living forms led him to ascribe a concrete reality to another type of organization, namely that of *embranchements*. An *embranchement* was an ensemble of animal forms that had a common structural plan, which served as the basis for all external modifications. Cuvier's four *embranchements*-vertebrates, molluscs, articulata (jointed or segmented animals), and zoophytes or radiata-are still well known. If there was a 'closed system' in Cuvier's mind it would seem to have been in systematics, at the level of the *embranchements*. Each of these formed a separate whole; there was no transition or gradation from one *embranchement* to another. Other organisms would not be viable because they would not meet the conditions of existence.

The 'construction plans' of the different *embranchements* were entirely different. There is, for example, no passage from vertebrates to molluscs.

> Whatever arrangement is given to animals with back bones and those without them, one can never place one of their large classes at the end of one group, and some what similar animals at the head of the other so that the two are linked together [Cuvier G (1800) *Leçons d'Anatomie Comparée: 1*. Paris: Baudouin. p. 60]

Similarly,

> There can be no intermediary between mollusca and articulata, nor between them and the radiata, for one cannot fail to recognise the profound interval or 'salta tion' there is when one passes from one construction plan to another.

It was in this spirit that Cuvier undertook the palaeontological investigations for which he became famous.

One of his first concerns was to evaluate the significance of fossils in the reconstruction of the Earth's past. The 'documents' furnished by the successive layers – the 'charters' or 'diplomas of the history of the globe' – revealed that all organisms were not created simultaneously. There was a 'definite succession in the forms of living organisms'. If there were only unfossiliferous strata, one might claim that the various terrains were created at one and the same time. But palaeontology showed that the various classes of vertebrates do not date from the same epoch. Cuvier was certain that the oviparous quadrupeds appeared much earlier than the viviparous types, for he thought that they began with the fishes, whilst the terrestrial quadrupeds appeared long after. Moreover, there is not only an order of succession between classes but also a pronounced order of the species within the stratigraphical column.

In establishing this chronological sequence, thanks to the collaboration of his friend Alexandre Brongniart (1770–1847), who was more of a geologist than he was, Cuvier clearly affirmed the connection between fossils and geological strata – between palaeontology and stratigraphy – to the benefit of the 'true' theory of the Earth. Indeed, in 1806, he proposed to the Académie des Sciences a programme of palaeontological research that would qualify as stratigraphical. Some of the main tasks were to ascertain whether there was any regularity in the succession of fossils, to determine which species appeared first and which came later, and to discover whether these two kinds of species are never found together, or whether there are alternations in their reappearance.

In his *Recherches sur les Ossemens fossiles de Quadrupèdes* (published in 1812), Cuvier applied himself to this programme. With Brongniart, he had proposed to resolve the following questions by means of his studies. Are there animals or plants that are proper to certain strata, and which do not occur in others? Which species appear first, and which come after? Do these two sorts of species sometimes occur together? Are there alternations in their recurrence; in other words, do the first forms recur and the second ones then disappear? Have these animals and plants perished in the places where their remains are found or have they been transported there?

Cuvier noted that the fossils of the Calcaire grossier of the Paris Basin are entirely different from those of the Chalk, and he clearly enunciated the concept of 'characteristic fossils' in stratigraphy. If a formation could be characterized by its fossils, it was because it contained organisms that were entirely distinct from those of other formations. Fossils were the fundamental tools of stratigraphical determination, and Cuvier could use them to recognize a particular formation in a large number of calcareous beds. A formation previously observed in some distant canton could be recognized by the nature of the fossils in each bed. Fossils were marker signals that never failed. Indeed, there was a constant relationship between the strata and the animal and plant remains found in them. Thus an immense field of observation and research was opened up, and Cuvier never doubted that reality would correspond progressively with this programme.

Precisely determining fossil species and delimiting the places where bones were discovered would make it possible to compare not only the superposed strata but also strata that were juxtaposed, in a parallel geological situation, neither above nor below but adjacent to one another, in the same basin or at a distance in two separate basins. Cuvier was not content simply to study the Paris Basin: he extended his observations to other regions. He thought it was important to study the calcareous strata of other basins and to compare them with those of the Paris Basin. Applied successively to other cantons, this method would soon yield important generalizations, and palaeontology, too long fed by illusory conjectures, would evolve a rigorous progress similar to that of other natural sciences.

From stratigraphy, the true history of the Earth began to emerge. When fossils were studied *in situ*, or in relation to strata, they ceased to be simple curiosities and became 'historical records'. Thanks to their study according to this perspective, one could show that there had been successive epochs in the formation of the globe and that a series of different operations or processes had operated at different times.

The historical key provided by palaeontology was thus well established by Cuvier. He was enthusiastic about the grand prospect of studying the past. He wanted to be able to arrange organisms in their chronological order, know about the development of life, determine precisely which forms appeared first, and recognize the simultaneous appearance of certain species and their gradual destruction. His vision provided a research programme for geologists once they had renounced their 'just-so stories' and begun, instead, to do the work of historians. The problem of the history of the Earth was correctly posed in palaeontological and stratigraphical terms. Even though, after Cuvier and Alexandre Brongniart, there was still much to discover, the method for discovery left nothing to be desired.

Cuvier's early works made him one of the masters of comparative anatomy and also opened up a new field to him – palaeontology – that was seemingly full of promise and in which his knowledge of German science, which was then more advanced than French

science, gave him a privileged position in France. His *Recherches sur les Ossemens Fossiles de Quadrupèdes* was published in 1812, but the work had been published earlier in the *Annales du Muséum*. In a communication made to the Institut de France, Cuvier had, in 1801, enunciated three hypotheses that were, even then, already known to his colleagues: former species had been entirely destroyed, or they had been modified in form, or they had been transported from one climatic zone to another. The second of these three alternative explanations had originally been suggested by Jean-Baptiste Lamarck (1744–1829), while the third was proposed by Barthélémy Faujas de St. Fond (1741–1819) and others. Cuvier preferred the first, which involved not only the disappearance but also the destruction of ancient forms. Thus, from the beginning of his scientific career, the young naturalist adopted the postulate of what came to be known as catastrophism as the basis for his palaeontological researches.

With this end in view, Cuvier applied himself to distinguishing carefully between fossil and modern forms. He was certain that none of the ancient forms had 'living analogues'. The 23 species that he had already been able to restore all appeared to have been 'destroyed' and to have become extinct. This assumption had direct consequences for geological theories: the lost species had "belonged to beings from a world anterior to our own and to beings that were destroyed by some revolutions of the globe" (Mémoire sur les espéces d'Eléphans vivantes et fossiles, Mémoire de l'institut national des Sciences et des Arts, Fructidor an VII (août-septembre, 1799), 2, 1–22: cit. p. 21). The disappearance, or, as Cuvier put it, the 'destruction', of former beings could only be explained by a 'general revolution of nature'.

The master palaeontologist laboured hard to establish the reality of the 'last catastrophe', which was related to the ideas favoured by believers of Holy Scripture. Although the picture of a total destruction of organisms did not correspond with the facts given in the Bible, Cuvier presented himself as a defender of the Noachian Deluge. He, along with Jean-André Deluc and Déodat Dolomieu (who were defenders of the idea of the Flood as a geological agent), thought that if anything was certain in geology it was that the surface of the globe had undergone a sudden revolution no more than five or six thousand years ago. But this catastrophe was, for Cuvier, only 'the last universal inundation'. That Cuvier was a supporter of general catastrophes does not, however, mean that he did not also uphold the idea of limited or local catastrophes. In the series of revolutions that he proposed, some were only partial. But just one universal catastrophe was enough to raise the problem that it posed for the continuity of life. Thus, although there had been numerous revolutions, there had not been so many creations, for migrations could play a role in some cases, as he suggested could potentially occur in New Holland (Australia). If there was an inundation of Australia that destroyed all its marsupials and the continent was subsequently colonized by animals from Asia, then the stratigraphical record in Australia would show a general catastrophic flood followed by the new creation of forms.

Cuvier never proposed a precise number of revolutions or creations. It was his follower Alcide d'Orbigny (1802–1857) who devoted himself to such calculations. He divided the stratigraphical column into 27 stages and therefore proposed a total of 28 creations, which came to serve as the basis for later stratigraphical work. Nowhere in Cuvier's *oeuvre* do we find the expression 'successive creations'. However, he originated the idea of repeated creations. From the moment when he envisaged general irruptions that "destroyed all the quadrupeds that they reached" and "caused the entire classe to perish" new creations were required to make the animals reappear. Cuvier's position on this was quite explicit, and so well known that from the beginning of the nineteenth century until his death he was considered to be the catastrophists' leader, as the eminent geologists who knew him would have agreed. Could Cuvier, who was able to speak so clearly and on occasion defend himself so well, have been misunderstood to such an extent by his contemporaries? This is inconceivable.

Cuvier had other good reasons for rejecting transformism, which were doubtless less significant for him than his catastrophism, but which had greater scientific validity. He raised a palaeontological objection that was valid even from a non-catastrophist perspective, namely the absence of intermediate or transitional forms between the former creatures and those that were more recent or extant. If the most ancient forms were the ancestors of those that followed, then one would expect to find the remains of the genealogical intermediaries. But, Cuvier objected, such transitional forms are never found. Between the *Paleotherium* and today's most similar species no such forms had at that time been found. Cephalopods do not lead on to fishes – a fact that even Lamarck did not dispute. However, Cuvier did not say how he thought new forms could have been created.

Influence: The 'Domination of Cuvier'

According to David Hull, in 1860, "on the continent, especially in France, catastrophism still reigned". But this view is mistaken.

Cuvier died in 1832 at the height of his fame. He was, however, already a controversial figure in politics, well known for his participation in the service of a succession of different regimes, and his intellectual and scientific worth were questioned by some of his contemporaries. The judgment of Goethe is well known: "No one described a fact better than he did. But he is almost devoid of philosophy. He will produce pupils well trained but with little depth" Eckernonn J.P. (1948). Gespr̈ache mir Goethe. Muncher, Deutsches Verläghaus Bong, pp. 329–330. Alexander von Humboldt, who admired Cuvier's "memorable studies of fossil bones", revealed his disagreements with him during a lecture Cuvier gave at the Collège de France, in which Cuvier criticized the ideas of Ètienne Geoffroy Saint-Hilaire.

Cuvier undoubtedly had faithful followers, especially in England; William Buckland is the best known but many others could be cited. Cuvier also had disciples in France, of whom Adolphe Brongniart is the best known and was the most devoted but there were others, for example Léonce Èlie de Beaumont and d'Orbigny (although they had some reservations). Èlie de Beaumont followed Cuvier in his rejection of fossil man, but did not do so when it came to the general destruction of life, as Darwin noted in 1859. Even d'Orbigny questioned some of Cuvier's conclusions about the Earth's past.

It is among the members of the thriving and internationally esteemed Société Géologique de France, founded in 1830 – of which, significantly, Cuvier was not a member, although Darwin (*see* **Famous Geologists:** Darwin) and Lyell (*see* **Famous Geologists:** Lyell) were – where one should look for the opponents of his catastrophist and anti-transformist opinions. Even before his death, many opposed him, including Constant Prévost (who was one of Lyell's teachers), Jean-Baptiste d'Omalius d'Halloy (the father of Belgian geology), Jules Desnoyers, André de Férussac, Marcel de Serres, and Ami Boué (one of the Society's founders). In 1830, Ami Boué, expounding Adolphe Brongniart's work on fossil plants, maintained that Cuvier's ideas were contested or rejected by 'the majority of geologists', whose names he gave. Shortly before Cuvier's death Boué declared himself to be radically opposed to "the idea put out by Mssrs Cuvier and Buckland on the universal Deluge, the universality of which was shown to be false by the most obvious facts" and also to "other opinions of M. Cuvier, namely... his hypothesis which, contrary to the natural order and to facts, admits universal cataclysms at several epochs prior to the Noachian Deluge". Gérard-Paul Deshayes, another of Lyell's tutors, recalled that in 1835 French zoologists were divided into two camps, one following Cuvier and the other following Lamarck. In 1836, Lyell himself placed Cuvier and Lamarck on the same rung. Several members of the Society, including Boué, openly declared themselves to be supporters of Lamarck.

Informed historians will not be surprised to read the evidence of Camille Dareste, who, in 1859, before Darwin became widely known, attested that the transformist theory of Lamarck had penetrated deep into the French scientific community. The domination of Cuvier after his death is a mistaken historical legend, which conscientious historians should not perpetuate.

See Also

Biblical Geology. **Creationism**. **Evolution**. **Famous Geologists:** Darwin; Lyell. **History of Geology From 1780 To 1835**. **Stratigraphical Principles**.

Further Reading

Boué A (1831) Compte rendu de la traduction allemande, par Nöggerath, des *Révolutions du Globe*, par Cuvier. *Bulletin des Sciences Naturelles et Géologie* 24: 129 130.

Coleman W (1964) *Georges Cuvier Zoologist: A Study in the History of Evolution.* Cambridge, MA: Harvard University Press.

'Collectif' (1932) *Centenaire de Cuvier, Archives du Muséum National d' Histoire Naturelle.* Paris: Masson et Cie.

Cuvier G (1797) *Tableau Èlémentaire de l' Histoire Natur elle des Animaux*. Paris: Baudouin.

Cuvier G (1800) *Leçons d' Anatomie Comparée: 1.* Paris: Baudouin.

Cuvier G (1801) Extrait d'un ouvrage sur les espèces de Quadrupèdes. *Journal de Physique* 52: 253 267.

Cuvier G (1812) *Recherches sur les Ossemens Fossiles de Quadrupèdes: 1.* Paris: Déterville.

Cuvier G (1812) *Essai sur la Géographie Minéralogique des Environs de Paris.* Paris: Baudouin.

Cuvier G (1817) *Mémoire pour servir à l'Histoire et à l' Anatomie des Mollusques, Mémoire sur la Scyllée.* Paris: Déterville.

Cuvier G (1825) *Discours sur les Révolutions de la Surface du Globe.* Paris: Dufour.

Cuvier G (1827) *Rapport Historique sur les Progrès des Sciences Naturelles depuis 1789*. Paris: Imprimerie Imperrales.

Dareste C (1859) Biographie de Lamarck. In: Hoefer JCF (ed.) *Nouvelle Biographie Générale*, 29, pp. 55 62.

Flourens P (1859) *Histoire des Travaux de Georges Cuvier,* 3rd edn. Paris: Garnier.

Hull D (1973) *Darwin and his Critics: The Reception of Darwin's Theory of Evolution by the Scientific Commu nity.* Chicago: Chicago University Press.

Laurent G (1987) *Paléontologie et Évolution en France, 1800 1860: De Cuvier Lamarck à Darwin*. Paris: Comité des Travaux Historiques et Scientifiques.
Laurent G (2000) Paléontologie(s) et évolution au début du XIXe siècle: Cuvier et Lamarck. *Asclepio* 52: 133 212.
Outram D (1980) *Georges Cuvier, Vocation, Science and Authority in Postrevolutionary France*. Manchester: Manchester University Press.
Rudwick MJS (1997) *Georges Cuvier, Fossil Bones, and Geological Catastrophes: New Translations and Interpretations of the Primary Texts*. Chicago and London: The University of Chicago Press.
Smith JC (1993) *Georges Cuvier: An Annotated Bibliography of his Published Works*. Washington DC: Smithsonian Institution Press.

Darwin

D R Oldroyd, University of New South Wales, Sydney, Australia

Introduction

Charles Darwin (Figure 1) is chiefly remembered for his celebrated theory of the evolution of life forms and speciation, by means of natural selection. But his considerable contributions to geology should not be forgotten. Darwin, born in 1809, was brought up at Maer, Staffordshire, UK, son of a prosperous doctor. He attended school at Shrewsbury, and at age 16 proceeded to Edinburgh University to study medicine, but withdrew from the course because of his distaste for dissections and operations conducted without anaesthetics. He then went to Cambridge to take the standard arts degree, with a view to becoming a clergyman.

Darwin's interest in natural history developed while he was still at school and was furthered in Edinburgh by studies of plankton in the waters of the Firth of Forth. He attended some of Robert Jameson's mineralogical/geological lectures, which were presented according to the 'geognostic' principles of the famous eighteenth-century Freiberg teacher of 'Neptunist' theory, Abraham Werner; Darwin found the ideas taught unacceptable and he discontinued his attendance. However, he may have learnt more geology there than he later acknowledged in his autobiography. Both Darwin and his father described his time at Cambridge as wasted, which it was, so far as theological training was concerned, but Darwin continued his private studies in natural history (especially beetle collecting) and became an epigone (and later, a personal friend) of the botany professor John Henslow. Henslow imparted some geological understanding to Darwin, having earlier done a fair amount of geological work, notably in Anglesey.

Beagle Voyage

Enthused by the writings of Alexander von Humboldt, Darwin wanted to travel. In his last year at Cambridge, he planned an informal journey with friends to Tenerife, which necessitated his brushing up on geology, having earlier largely ignored the subject at Cambridge, being "so sickened with the lectures at Edinburgh". Henslow taught him to use a clinometer and gave him geological advice on his project. Possibly Darwin also attended some lectures of the geology professor Adam Sedgwick (*see* **Famous Geologists:** Sedgwick), and certainly participated in the field excursions Sedgwick ran, around Cambridge. Then, in 1831, at Henslow's suggestion, Sedgwick took Darwin along on a short field excursion in North Wales as assistant and companion, also with the idea of teaching him the rudiments of field geology. Darwin was an apt pupil, but this was essentially

Figure 1 Charles Darwin (1809 82), from a photograph (1854?). Engraved for *Harper's Magazine*, October 1884. Reproduced from Darwin F (1887), 1, Frontispiece.

all the training he had in geology. Nevertheless, when later that year Darwin joined the *Beagle* on its epic journey and circumnavigation of the globe, he initially regarded geology as his principal scientific interest and even skill, having also done some solo field trips in south-west England.

Darwin studied Charles Lyell's *Principles of Geology* (*see* **Famous Geologists:** Lyell), the first volume of which (published in 1830) he had on board the *Beagle*; according to Darwin's later published accounts, he immediately sought to apply his understanding of Lyell's ideas to the observations made at St Iago, in the Cape Verde Islands, the *Beagle*'s first port of call. However, Darwin's field notes made at a village called St Domingo, situated in a wadi-like inland valley, reveal that he was initially thinking in catastrophist terms, such as he might have picked up from Sedgwick. Darwin thought that the valley had been created by some "great force" followed by "the agency of large bodies of water", which moved large boulders and seemed also to have deposited material similar to diluvium (which he had encountered aplenty in Wales) near the coast. Subsequently, when Darwin published his geological observations, his initial catastrophist ideas were set aside, his Cape Verde observations being presented as supportive of Lyellian geology. Thus, Darwin reported evidence for uplift of land, in the form of a fossiliferous limestone about 40 feet above sea-level, with similar material forming at sea-level today. A volcano there also seemed to have depressed strata, but the published comparison of past and present processes was unquestionably Lyellian.

Along the east coast of South America, Darwin made important collections of megafaunal remains. He made inland excursions towards the Andes and observed what appeared to be a 'stepped' structure for the Patagonian plains, indicating successive elevations. Near Tierra del Fuego, he saw snow-covered mountains, and glaciers descending close to sea-level. Travelling up the Chilean coast in 1835, he experienced a major earthquake at Concepción, with considerable uplift directly evident and in seeming accord with Lyellian geology. Darwin also travelled across Andean passes and formed the idea that the earth had been ripped apart along the line of the mountains and that igneous matter had entered the huge fissure. He even prepared a sketch map (in about 1834) of the southern part of the continent, distinguishing three main units: Tertiary strata; granites, gneiss, mica, slate, quartz rock, and clay slate; and lavas, tufas, and porphyries. A more detailed, also undated, map of the more southerly part of the continent from about the same period depicted seven units: 1. Granite, Mica slate; 2. Trappean rock and Porphyries; 3. Purple Porphyry and Tufa. Metamorphics; 4. Clay Slates; 5. Tertiary (Pliocene?); 6. d[itt]o—Recent; 7. Basaltic Lava.

Crossing the Pacific, Darwin noted that some coral islands seemed to have been elevated, as indicated by the dead coral above sea-level. Elsewhere, there seemed to have been subsidence, but the corals were growing upwards at about the same rate as the land was sinking, so that fringing reefs were formed. This evidence cohered with Lyell's idea that some parts of Earth's crust were rising while others were sinking. In Australia, Darwin crossed the Blue Mountains and observed its great cliffs and valleys, but having only a restricted view of the topography, he mistakenly supposed that the valleys were produced by marine rather than by fluvial erosion. In Tasmania, he quickly got the hang of the geology around Hobart, and possibly, on the basis of 'drop-stones', had the idea of there having been a glaciation in what he referred to as Carboniferous times (actually Permian).

On his return to England, Darwin soon became acquainted with Lyell and remained a lifelong friend; Darwin was elected a Fellow of the Geological Society and was soon on its Council, then became a Secretary and, later, Vice-President. In 1838, he published a general theoretical paper, *On the Connexion of Certain Volcanic Phenomena in South America, and on the Formation of Mountain Chains and Volcanos, as the Effects of the Same Power by which Continents are Elevated*. Published a year after he had come to accept 'transformism', this work was less 'steady state' than Lyell's geology envisaged. Darwin linked earthquakes, elevations, and volcanic eruptions. He thought that (for the Andes, at least) there were repeated uplifts and intrusions along the axis of the range, followed by cooling and consolidation. Rejecting global contraction, differential sedimentary loads, or interplanetary forces, he had no concrete suggestion as to the cause of the elevations. But he was convinced that the uplift proceeded in small stages and was ongoing, rather than occurring in one great catastrophic episode. He concluded that "the configuration of the fluid surface of the earth's nucleus is subject to some change, —its cause completely unknown, —its action slow, intermittent but irresistible". (Modern students of plume theory and the effects of actions occurring at the mantle/core boundary may find this remarkably prescient!)

Geological Publications and ideas on Glacial Phenomena

Following the *Beagle* voyage, Darwin published three major geological books: *The Structure and Distribution of Coral Reefs* (1842), *Geological Observations*

on the Volcanic Islands Visited during the Voyage of H.M.S. Beagle (1844), and *Geological Observations on South America* (1846). In addition to matters previously discussed, he distinguished in 1846 between stratification, cleavage, and foliation, but the distinction between cleavage and bedding was probably imparted to him by Sedgwick. Also, and importantly, in 1844, Darwin initiated for petrologists the idea of gravity settling, based particularly on his observations of igneous rocks in the Galapagos Islands; this was based on the idea that crystals that first form from a cooling magma may separate out and thereby alter the chemical composition of the remaining fluid, thus producing magmatic differentiation.

Another important piece of work undertaken by Darwin post-voyage was his attempt in 1839 to explain the strange set of markings, the so-called Parallel Roads of Glen Roy, on the sides of Glen Roy in central Scotland. These controversial parallel and horizontal markings evidently marked former shorelines of some kind. In Darwin's view, they represented different marine shorelines, being formed (by analogy with ideas developed in South America) by a succession of land elevations. Erratic granite boulders were also to be found, and Darwin ascribed their deposition to floating icebergs. Indeed, the whole situation was seen and interpreted in terms of what he had seen in the Tierra del Fuego area, with Glen Roy being in some ways comparable to the Beagle Channel. So the observations were thought to accord with the notion of subsidence of the land, associated with cold and extended glaciation in Scotland. On subsequent elevation and amelioration of the climate, the supposed shorelines and erratics would be exposed. Subsequently, in 1842, Darwin thought the glacial submergence and floating iceberg theories could also be applied in North Wales.

Darwin's ideas about Glen Roy were later superseded by the idea that the marks were due to glacial lakes: i.e., water ponded in the valley by glaciers blocking its mouth at different altitudes, water having escaped over different passes at different altitudes, according to the size of the barriers. Thus, there were several distinct former lake margins at different levels. Darwin later acknowledged the superiority of this theory, and called his Glen Roy paper a "great failure" in his autobiography.

Later Years, Evolution, and the Age of the Earth

By the 1840s, Darwin's health was deteriorating and he gave up substantial fieldwork: he was beginning to focus more attention on his grand theory of evolution by natural selection, first adumbrated in 1837. His last geological paper proper (on the geology of the Falkland Islands) appeared in 1846, and that year he turned to a taxonomic study of barnacles, both modern and fossil, continuing this work at his home for eight years. His study of modern forms led to his discovery of males living as 'parasites' within the female forms, and also a gradation from hermaphrodite types, through forms with females having an 'attached' male organ and one that was parasitic but physically detached, to types whereby there was sexual dimorphism, but in which the males were 'parasitic' on the females. Thus, Darwin saw, in barnacles, evidence for the evolutionary emergence of sexual dimorphism. He then turned to the study of fossil barnacles, publishing a two-volume monograph (in 1851 and 1854). In the light of hindsight, it can be seen that these works were ordered (or the organisms classified) from an evolutionary perspective.

When Darwin eventually published his *Origin of Species* in 1859, there were two issues of principal geological interest. First, he wanted to present to the public the idea of the history of living forms as being analogous in form to a branching tree; second, he had to deal with the problem of the age of Earth. There was also the problem of the origin of life and the apparent appearance of quite new forms from time to time, particularly the appearance in the Cambrian (or Silurian as he termed it then, following Roderick Murchison (*see* **Famous Geologists:** Murchison)) of quite well-developed forms apparently without ancestors. But the actual stratigraphic record showed anything but continuity or smooth transitions. Examples of trends, with a complete presentation of the various forms in an evolving continuum, or cases of branching and speciation, were conspicuous for their absence in the fossil record. Darwin sought to answer this difficulty by appealing to the incompleteness of the stratigraphic record: many pages of the evolutionary record were missing due to weathering and erosion, or had been destroyed by metamorphism. Moreover, intermediate forms might not all have been preserved at the same locality, so it would be unreasonable to expect smooth transitions in an ascending section. Similarly, the abrupt appearance of new types in strata was a problem for Darwin (and was used as an argument against him by his contemporaries, and by critics ever since). Again he appealed to the immensity of time, the imperfection of the geological record, and geologists' incomplete knowledge of that imperfect record. Also, he pointed out that much time might be required to evolve some particularly advantageous character (such as the ability to fly), but once

acquired, the increase of that character would be very rapid. So the fossil record might give the appearance of sudden changes; but the reality could actually have been one of continuous change. And sometimes anticipated ancestors might be found, as in the case of fossil whales then quite recently found in the Cretaceous. Darwin long sought the occurrence of fossil sessile barnacles and was delighted when some were eventually found in the Chalk, as he had expected would be the case one day. As to the absence of fossils older than the Cambrian (Silurian), Darwin thought that metamorphism might chiefly be responsible. He did not know that soft-bodied Precambrian fossils would one day be found; but it would have been in accordance with his expectations.

As to the age of Earth, comparison of the thickness of preserved sediments with the rate of deposition seemed to reveal the immensity of time. Assuming the erosion of the valley of the Weald in Kent as being chiefly due to the action of the sea, and thinking of erosion as proceeding at 1 inch per century, Darwin gave a figure of 306 662 400 years for the formation of the valley. This was obviously a crude estimate. Darwin halved this estimate in the second edition of the *Origin* and subsequently withdrew it all together. Modern opinion has it that Darwin's figure was much too large, but it is evident that he had a clear vision of the immensity of geological time, and thought that it could brush away many of the objections to his evolutionary theory.

In his last major work, *The Descent of Man* (published in 1871), Darwin boldly applied his theory to humans, but said little about fossil forms, and the book had little geological content. He did, however, speculate that humans had first evolved in Africa, because that was where our nearest animal relatives were found; this suggestion is still thought to be correct.

See Also

Biological Radiations and Speciation. **Evolution**. **Famous Geologists:** Lyell; Murchison; Sedgwick. **History of Geology From 1835 To 1900**.

Further Reading

Barrett PH, Gautrey PJ, Herbert S, Kohn D, and Smith S (eds.) (1987) *Charles Darwin's Notebooks, 1836 1844: Geology, Transmutation of Species, Metaphysical Enquiries*. London: British Museum (Natural History); Ithaca: Cornell University Press.

Darwin C (1842) *Geology of the Voyage of the Beagle: The Structure and Distribution of Coral Reefs*. London: Smith Elder & Co.

Darwin C (1844) *Geological Observations on the Volcanic Islands Visited during the Voyage of H.M.S. Beagle*. London: Smith Elder & Co.

Darwin C (1846) *Geological Observations on South Amer ica*. London: Smith Elder & Co. (Also numerous later editions and different publishers.)

Darwin F (ed.) (1887) *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter* 3 vols. London: John Murray.

Herbert S (1986) Darwin as a geologist. *Scientific American* 254(May): 116 123.

Herbert S (1991) Charles Darwin as a prospective geo logical author. *British Journal for the History of Science* 24: 159 192.

Herbert S (2005) *Charles Darwin, Geologist*. Ithaca and London: Cornell University Press. In press.

Pearson PN (1996) Charles Darwin on the origin and diver sity of igneous rocks. *Earth Sciences History* 15: 49 67.

Pearson PN and Nicholas CJ (2003) Charles Darwin's geo logical observations at Santiago (St Jago), Cape Verde Islands. *International Commission on the History of Geological Sciences 28th International Symposium. Trin ity College, Dublin, Ireland . . . Programme, Abstracts & Delegates*, 41.

Rhodes FHT (1991) Darwin's search for a theory of the Earth: symmetry, simplicity and speculation. *British Journal for the History of Science* 24: 193 229.

Roberts MB (2000) I coloured a map: Darwin's attempts at geological mapping in 1831. *Archives of Natural History* 27: 69 79.

Rudwick MJS (1974) Darwin and Glen Roy: a "great failure" in scientific method? *Studies in History and Philosophy of Science* 5: 97 185.

Secord JR (1991) The discovery of a vocation: Darwin's early geology. *British Journal for the History of Science* 24: 133 157.

Stoddart DR (1976) Darwin, Lyell, and the geological sig nificance of coral reefs. *British Journal for the History of Science* 9: 199 218.

Du Toit

J C Loock, University of the Free State, Bloemfontein, South Africa
D F Branagan, University of Sydney, Sydney, NSW, Australia

Introduction

Alexander Logie du Toit, South Africa's greatest and best-known geologist, started his career in 1903 and, over the next 20 years, mapped and investigated a large area of South Africa. He collected data simultaneously on the Dwyka Tillite, other Karoo rocks (formerly spelled Karroo), vertebrate fossils, the southern *Glossopteris* flora, Karoo dolerites and the Drakensberg basalts. During his career he expanded his interests to include other fragments of Gondwana. This led him to propagate the idea of the former existence of a southern supercontinent. He was very much aware of the pioneering work of his predecessors, especially Wegener (*see* **Famous Geologists:** Wegener), but he carried their theories further. After a short period of consolidation and publication in the early 1920s, he settled down to write his book on the 'wandering continents'. His *Our Wandering Continents* (published in 1937) cast the gauntlet at sceptics and anti-drifters. Du Toit was vindicated nearly three decades later, when geophysical studies, mainly of the ocean floors, led to the theory of plate tectonics.

Ancestry and Opportunities

Alexander du Toit (Figure 1) was fortunate in birth and ancestry. Early in the nineteenth century, Alexander Logie from Fochabers, Banffshire, Scotland, served as an officer in the 72nd Regiment in South Africa. Captain Logie married Henrietta Elizabeth Susanna du Toit, a descendant of Francois du Toit from Lille in France, who arrived at the Cape of Good Hope in 1686 as a Huguenot. As the marriage was childless, the couple adopted the infant son of Stephanus Hendrik du Toit, the brother of Mrs Logie, and his wife, Barbara Stadler. The boy was christened Alexander Logie, but he retained his du Toit surname, growing up on the family estate on the outskirts of Cape Town. Alexander married Anna Logie, daughter of Robert Clunie Logie, a brother of Captain Alexander Logie. They had four children, one of whom was Alexander Logie du Toit, born at Rondebosch near Cape Town on 14 March 1878. His relatives included some of the most prominent families then living in Cape Town. When he passed the matriculation examination in 1893 he was among the top scholars in Cape Colony.

Du Toit was sent to the University of the Cape of Good Hope, where he passed the intermediate examination with distinction in 1894 and went on to obtain the Bachelor of Arts degree in mathematics and natural science the following year, again passing with distinction.

Du Toit's descent dictated that he should study in Scotland. He qualified in mining engineering at the Royal Technical College of Glasgow in 1899 and then studied at the Royal College of Science in London. In 1901 he was appointed a lecturer at the Royal Technical College of Glasgow and also at the University of Glasgow. His studies and travels in Scotland and further afield gave him an insight into the geology and stratigraphy of the northern hemisphere and into the nature, origin, and structure of fold mountains. This newly acquired knowledge later proved crucial in his theories on continental drift. He kept his contacts with Scotland and submitted a thesis on the copper–nickeliferous layered intrusion of Insizwa in 1910, for which he was awarded a DSc degree.

In Scotland all was not study. Du Toit played the oboe, a hobby that he maintained for many years. He married Adelaide Walker in Glasgow and returned

Figure 1 Portrait of Alexander du Toit. (Photograph reproduced by courtesy of the Geological Society of South Africa.)

with her to South Africa late in 1902, where a new world and a long career stretched before him.

The Formative Years

The Geological Commission of the Cape of Good Hope, 1903–1912

The Geological Commission of the Cape of Good Hope, established in 1896 by the Cape Parliament, consisted of a veteran politician and senior civil servants. Du Toit assumed the post of geologist at the beginning of 1903, being listed on the scientific staff as Alex L du Toit BA FGS. When he left for his first field session, he was accompanied by the Commission Director, Arthur Rogers, who introduced him to the upper beds of the Cape Supergroup (Devonian) and the overlying basal units of the Karoo Supergroup (Late Carboniferous to Early Permian). This was du Toit's first contact with the Dwyka Tillite, which is exposed in the south-western corner of the Karoo outcrop area in a desert environment. We can picture him walking on the unweathered outcrops with a huge number of loose erratics lying around, which he identified according to rock type. Later, he was to trace these erratics in the Northern Cape to actual outcrops of Precambrian rocks. Little did he know at that stage that he was destined, three decades later, to be an international expert on the Dwyka Tillite and its equivalents in other fragments of Gondwana.

During his 9 years service with the Commission, du Toit mapped and studied the rocks and the strata in three areas. In the western corner of the huge area underlain by rocks of the Karoo Supergroup, he unravelled the stratigraphy of the two basal units, namely the Dwyka and Ecca Groups, paying special attention to the Dwyka glacials. Du Toit also had an interest in the Karoo Dolerite (Jurassic) and the diatremes that pierce the Karoo beds. In the north-eastern Cape, he covered an area containing rocks ranging from Early Precambrian to Quaternary. Diatremes and kimberlite intrusions again attracted his attention. Du Toit is, however, chiefly remembered as a field geologist for his detailed and accurate maps of the north-eastern Cape, where he concentrated on the Beaufort and Stormberg Groups and the plateau basalts of the Drakensberg.

An image of du Toit as a competent field geologist now emerged. He was a wiry and energetic man, who could cover long distances on foot, on a bicycle, or, on occasion, on horseback. In the more open areas to the west, his caravan-like wagon was pulled by a donkey team (Figure 2). Mapping, often in areas where large-scale maps showed farm boundaries only, was done using a small plane table and alidade. He was renowned in the geological community for being able to judge distances very accurately. When mapping dolerites, he could visualize an intrusion into Karoo beds as a three-dimensional body and hence accurately predict the locations of the dolerite outcrops. One of du Toit's ways of winning the confidence of the local inhabitants was to encourage the infirm and to dispense aspirins or coloured pills to those who feigned or claimed illness. As his fame as a 'doctor' spread, he found that he could rely on the locals for advice on the geography of a mountainous area

Figure 2 During his first field excursions, du Toit was provided with a wagon and a team of donkeys. (Photograph reproduced by courtesy of the Natural History Division of the South African Museum, Iziko Museums of Cape Town.)

or the route of a cattle track through bush and over mountains.

The results of du Toit's labours in the field are to be found in the Geological Commission's annual reports, and other observations were published in scientific journals. The 12 maps with which he was involved, either on his own or mainly in conjunction with Rogers, covered an area of 180 000 km^2 and were a source of wonder to his successors. One of his publications, on the evolution of the river systems in Griqualand West, is consulted to this day by geologists prospecting for alluvial diamonds. In the first edition (1905) of his book on the geology of the Cape Colony, Rogers referred to 'Gondwanaland', the term introduced by Suess in his *Das Antlitz der Erde* (1885) (*see* **Famous Geologists:** Suess). When Rogers prepared the second edition, he invited du Toit to be his co-author. The second edition contained expanded versions of the Gondwana hypothesis and of South African stratigraphy.

Between 1905 and 1910, du Toit worked mainly in the more arid northern parts of Cape Colony, on unfossiliferous sediments and various igneous rocks. Here, he became friends with, and was impressed by, the Dutch geologist Gustaaf Molengraaff. Du Toit's reports on the nature and petrology of the kimberlite pipes, which were often richly diamond bearing, were widely read and used by prospectors and mining interests.

The Cape Geological Commission was disbanded at the end of 1911, but du Toit had already collected in his memory and in notebooks a vast number of facts, observations, interpretations, and ideas. He was to prepare a great synthesis in the following decade.

Geological Survey, 1912–1920

Du Toit was transferred to the Geological Survey of the Union of South Africa when the four colonial surveys or commissions were amalgamated in 1912, and he suddenly found that he was free to study geological problems over the whole of South Africa. By this time, du Toit had transferred his interests to minerals and rocks. He spent more and more time on the investigation of specific sites and less time on mapping. He participated in the activities of the Geological Society of South Africa and served as President in 1918 and again in 1928. His first book, *A Physical Geography for South African Schools*, appeared in 1912. In 1914 he visited Australia to attend the meeting of the British Association, where he met the Sydney professor T W Edgeworth David and was able to examine the rock succession in eastern Australia and the evidence for Late Palaeozoic glaciation. The remarkable similarities between the records of events in two widely separated southern continents, with evidence of glaciation in Australia at the same time as that indicated by the Karoo rocks in South Africa, was striking.

As the First World War had broken out, du Toit was called up to serve as a geologist and was charged with finding water for the Union Defence Force. When the South Africans invaded German South-west Africa (now Namibia), du Toit had to find suitable sites at which to drill for water in the desert. His military service must be seen as a bonus because he found time to study the basement rocks, the Late Precambrian Nama beds, and, more importantly, the basal Karoo rocks.

Du Toit's last publication while a member of the Survey was his monumental compilation on the Karoo dolerites.

Department of Irrigation, 1920–1927

The Department of Irrigation requested du Toit's services because it needed his expertise. His relatively brief period of service saw both tragedy and triumph. Adelaide du Toit died in 1923, leaving her husband and a grown-up son. Two years later, du Toit married Evelyn Harvey.

Du Toit's many reports from this period, now mostly filed away and forgotten, dealt chiefly with dam sites and geomorphology. Nearly two decades had passed since du Toit had started his career, and his observations and the synthesis of the facts were ready for a wider audience. When the South African Association for the Advancement of Science met in Durban in 1921, du Toit was invited to deliver the popular evening lecture, for which he chose to speak on land connections between South Africa and other continents. He presented evidence in the form of vertebrate life, the migration of vertebrates, palaeoclimates, volcanism, and fold mountains to an audience that included sceptics. His main thrust was a resumé and analysis of the glacial deposits at the base of the Karoo. This was augmented by a map showing Gondwana at the close of the Carboniferous.

In the same year, the Geological Society of South Africa published du Toit's summary and analyses of Carboniferous glaciation in South Africa. The references to the pre-glacial topography, the direction of flow of continental ice-sheets, and the distribution of erratics placed southern Africa in the wider context of Gondwana. A later generation of stratigraphers provided evidence that the Dwyka glacials range in age from the Late Carboniferous to the early Permian.

Other adventures followed. In 1923, a grant from the Carnegie Institution in Washington enabled du

Toit to visit Brazil, Uruguay, Argentina, and Chile, which were previously poorly mapped and not well understood. He carried out a remarkable amount of field exploration in South America, meeting local geologists, including his old friend David Draper from South Africa, who was briefly managing the Boa Vista diamond mine in Minas Geraes. Du Toit mastered the difficult literature, which was mainly in Portuguese, Spanish, and German. His tour allowed him to study Devonian beds, fold ranges, and, as might be expected, the Karoo equivalents. Back in South Africa, he was able to show Edgeworth David, *en route* to England, the key local sites relevant to the displacement hypothesis. It was a fruitful meeting, and both geologists strengthened their support for the notion of continental drift.

The Carnegie Institution published du Toit's *Geological Comparison of South Africa with South America* in 1927. The book contained, *inter alia*, a chart showing the stratigraphic column from the Devonian to the Jurassic for selected South American countries. The boundaries of the Afro-American landmass were shown to have a bearing on the displacement hypothesis, which was becoming more widely known through the English translation of Wegener's *Origin of Continents and Oceans* (1924).

By this time, du Toit had made contact with the Dutch geologist Willem van Waterschoot van der Gracht, who was then working in the petroleum industry in the USA. Van der Gracht had become an apostle of continental drift and persuaded the American Association of Petroleum Geologists to organise a conference in New York in late 1926 to discuss the theory. Despite the contributions of several supportive American speakers, including Reginald Daly, a prejudiced group, led by Charles Schuchert, condemned the theory out of hand. Unable to attend, du Toit was dismayed by the intolerant attitudes of some people and their personal attacks on Wegener and others. He sent a paper to support Van der Gracht's publication of the proceedings, which appeared in 1928, and in the following years added further publications in journals such as the *American Journal of Science*; even as late as 1944 he made a rejoinder to G G Simpson regarding his ideas on Tertiary mammals and continental drift.

For some years, du Toit had been planning to write a textbook on the geology of South Africa. The first edition, published in 1926, was the first synthesis of its kind. A second revised edition followed in 1939. Du Toit died before he could complete his revision of the third edition, but his old friend and colleague Sidney Haughton carried on the task and the book was published in 1954.

Years of Work and Wandering

In 1927, du Toit was invited to join De Beers Consolidated Mines as a consulting geologist, specializing in diamondiferous kimberlite pipes and alluvial gravels. Once again, he could travel extensively in Africa, but now in an official capacity. These travels, to areas in which De Beers had an interest, afforded him many opportunities to study the local geology. He could also visit other countries: the USA and Canada in 1932; the USSR in 1937; and India in 1938. Little concerning the areas visited survives in published form. However, from a cache of photographs and other documents discovered recently, we know that he travelled widely in the western USSR, from the Urals down to the Ukraine and the coastal area of the Black Sea.

Du Toit's greatest contribution to geology, and also his swansong, was his book *Our Wandering Continents* (1937), in which he assembled the observations, deductions, comparisons, and syntheses of facts and theories of a lifetime. A brief description of the features of the book is in order. After a review in which he acknowledged the work of his predecessors, he referred to tectonism, volcanism, palaeoclimates, plant fossils, and geosynclines to explain and describe his grouping of the ancient continents and his theory of drifting. The distribution of glacials formed, as before, the core of his arguments. We should note that du Toit now used the term 'Gondwana' for the southern supercontinent. For the northern supercontinent he introduced the term 'Laurasia', derived from Laurentia (the eastern North American shield) and Asia. Additionally, the book contained a discussion of Arthur Holmes's suggestion of fracturing of continental blocks by subcrustal convection movements to explain spreading. Du Toit also included a figure showing the development of continental rifting that owed something to Holmes's famous figure of 1929, representing convection and continental fracture. But du Toit doubted whether subcrustal convection was 'wholly competent to account for continental drift'.

The evidence for the former linkage of the southern continents was illustrated in several convincing diagrams (**Figures 3, 4,** and **5**), and the structural correspondences between western Europe and North America were also depicted, along with a suggested pattern for the opening of the North Atlantic (see **Figures 6** and **7**).

Throughout the book, du Toit challenged the geological community to accept his theories, but he did not live long enough to witness the acceptance of continental drift nearly two decades later.

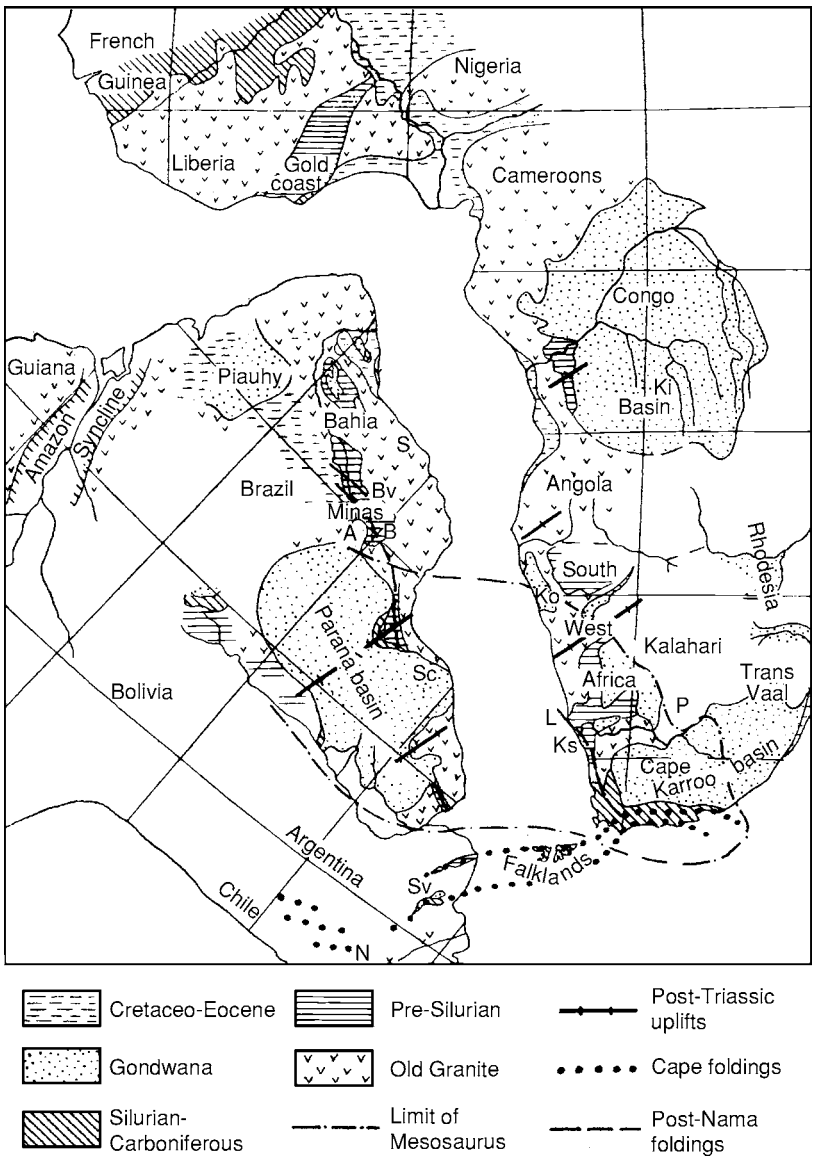


Figure 3 Analogies between South America and southern Africa, according to du Toit. (Reproduced from Du Toit AL (1937) *Our Wandering Continents: A Hypothesis of Continental Drifting*. Edinburgh: Oliver and Boyd.)

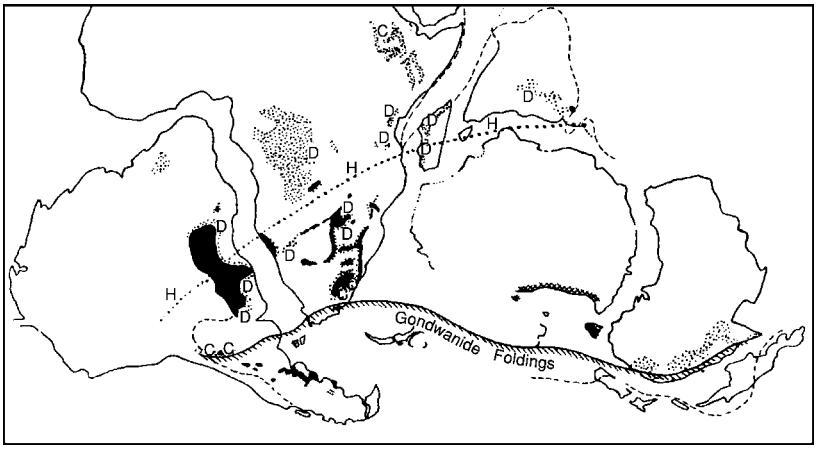


Figure 4 Structural features in 'restored' southern continents, Late Triassic and Rhaetic, according to du Toit (1937: 93).

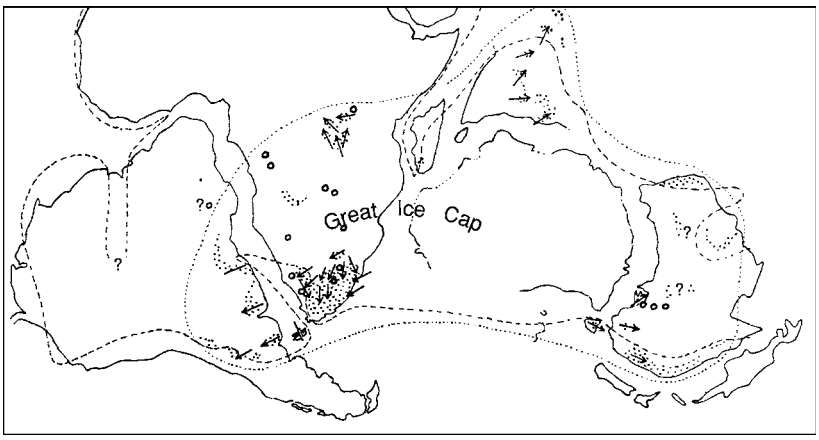


Figure 5 Arrangement of continents and areas of glaciation in the Late Carboniferous, according to du Toit (1937: 76).

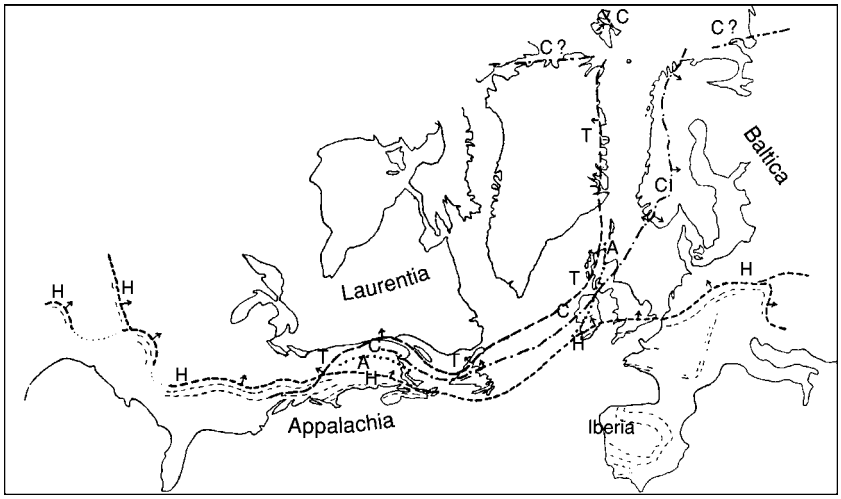


Figure 6 Structural analogies between North America and Europe, according to du Toit (1937: 145).

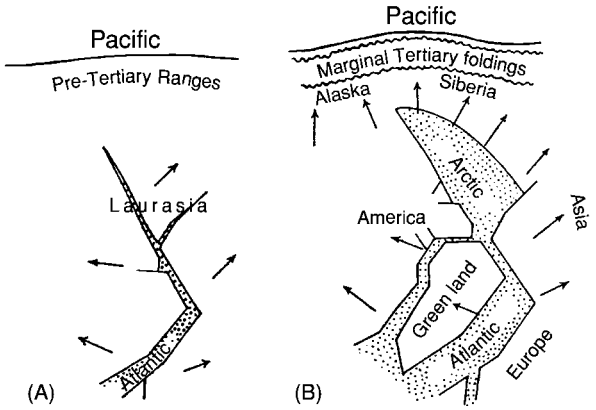


Figure 7 Evolution of the Atlantic Arctic rift, according to du Toit (1937: 222).

Retirement and Honours, 1941–1948

After retiring from De Beers, Alexander and Evelyn returned to Cape Town, where du Toit embarked on a frenzied round of activity, as his diaries for 1945 and 1946 reveal. He corresponded with friends and colleagues in South Africa and other countries, made social calls, and undertook short visits to the countryside. He continued writing scientific papers and assisted aspiring authors. His final diary entry (31 December 1946) contained a vague but ominous reference to a medical problem that was to carry him away on 25 February 1948.

During his lifetime, du Toit was awarded many medals and honorary doctorates from five South African Universities. His richly deserved Fellowship of the Royal Society of London was not awarded until 1943, possibly delayed by English prejudice against his continental drift ideas. The greatest honour of all came when the Geological Society of South Africa instituted the biennial Alex L du Toit Memorial Lecture Series in 1949. A few years later, his image appeared on a South African stamp.

A collection of his notebooks is owned by the University of Cape Town. The diaries for 1905, 1906, 1908, 1931, 1945, and 1946 are kept by the South African Museum in Cape Town. A cache of photographs and documents, together with his academic gown, was recently donated to the Museum. His awards, medals, and certificates are in the hands of a grandson.

See Also

Famous Geologists: Suess; Wegener. **Gondwanaland and Gondwana**. **History of Geology From 1835 To 1900**. **History of Geology From 1900 To 1962**. **Pangaea**. **Plate Tectonics**.

Further Reading

Branagan DF (2004) *The Knight in the Old Brown Hat: A Life of Sir T W Edgeworth David, Geologist*. Canberra: National Library of Australia.

Du Toit AL (1921) Land connections between the other continents and South Africa in the past. *South African Journal of Science* 18: 120 140.

Du Toit AL (1921) The Carboniferous glaciation of South Africa. *Transactions of the Geological Society of South Africa* 24: 188 227.

Du Toit AL (1926) *Geology of South Africa*. Edinburgh: Oliver & Boyd.

Du Toit AL (1927) *A Geological Comparison of South America with South Africa. With a Palaeontological Contribution by F R Cowper Read*. Washington: The Carnegie Institution.

Du Toit AL (1937) *Our Wandering Continents: A Hypoth esis of Continental Drifting*. Edinburgh: Oliver & Boyd.

Gevers TW (1949) *The Life and Work of Alex L du Toit*. Alex L du Toit Memorial Lecture 1. Johannesburg: Geological Society of South Africa.

Haughton SH (1949 [1950]) Memorial to AL du Toit (1878 1948). *Proceedings of the Geological Society of America*: 141 149.

Haughton SH (1949) Alexander du Toit 1878 1948. *Ob ituary Notices of Fellows of the Royal Society of London* 6: 385 395.

Holmes A (1929) Radioactivity and Earth movements. *Transactions of the Geological Society of Glasgow* 18: 559 606.

Rogers AW and du Toit AL (1909) *An Introduction to the Geology of Cape Colony With a Chapter on the Fossil Reptiles of the Karroo Formation by Prof R Broom,* 2nd edn. London: Longmans.

Waterschoot van der Gracht WAJM (ed.) (1928) *The Theory of Continental Drift: A Symposium*. Tulsa: American Association of Petroleum Geologists.

Hall

R H Dott Jr, University of Wisconsin, Madison, WI, USA

Introduction

James Hall (1811–1898) of New York was North America's pre-eminent invertebrate palaeontologist and geologist of the nineteenth century. That he was a giant among early American geologists is evidenced by the facts that he served as President of the American Association for the Advancement of Science (1856), was a charter member of the National Academy of Sciences (1863), and was chosen to be the first President of the Geological Society of America (1889). Hall was also the best-known American geologist on the international scene in his time. As early as 1837, he was elected to membership of the Imperial Mineralogical Society of St Petersburg. Later he was the Organizing President of the International Geological Congress meetings in Buffalo, New York (1876) and in Paris (1878); he was a Vice-President of the congresses in Bologna (1881) and Berlin (1885), and he was Honorary President of the Congress in St Petersburg (1897). Hall was elected Foreign Correspondent to the Academy of Sciences of France in 1884, being its first English-speaking member. It was primarily the *Paleontology of New York*, published in 13 volumes between 1847 and 1894, that initially brought Hall his fame. However, the broader community of geologists chiefly remembers him more for the curious theory of mountains presented in his Presidential Address to the American Association for the Advancement of Science in 1857.

Early Life and Education

Hall was born near Boston in Hingham, Massachusetts on 12 September 1811. His parents had emigrated from England two years earlier, and James was their first of five children. The family was of modest means, but the young Hall was fortunate to have a gifted teacher in his public school, who stimulated an interest in nature. Through his teacher, James met several leading members of the Boston Society of Natural History. Having developed a strong interest in science, Hall was attracted to a new college in Troy, New York, which emphasized science and employed revolutionary new approaches to learning, with an active role for the student coupled with hands-on laboratory and field-trip instruction. This Rensselaer Plan was developed by Amos Eaton, with financial backing from his patron, Stephen van Rensselaer. Unable to afford commercial transportation, Hall walked the two hundred miles to Troy. At Rensselaer, he was instructed by Eaton and Ebenezer Emmons, and had for classmates such geologists-to-be as Douglas Houghton, Abram Sager, Eben Horsford, and Ezra Carr. Hall graduated with honours in 1832 and undertook a tour on foot to the Helderberg Mountains in south-eastern New York to collect Silurian and Devonian fossils. A job as a librarian allowed him to continue at Rensselaer for another year and to earn a Master of Arts degree with honours (1833). He then held an assistantship in chemistry for several more years. In 1838, he married Sarah Aikin, the daughter of a Troy lawyer; they had two daughters and two sons. Sarah died in 1895.

The New York Survey

In 1836, the New York legislature authorized a 4-year geological and natural history survey; an extension of 2 years was later authorized. Four men – William W Mather, Ebenezor Emmons, Timothy A Conrad, and Lardner Vanuxem – were in charge of four respective districts, and Lewis C Beck was mineralogist for the survey. Botanist John Torrey and zoologist James De Kay conducted the biological survey. Hall was engaged to assist his former teacher, Emmons, in the Second District in north-eastern New York, where his first assignment was to study iron deposits in the Adirondack mountains. A year later the districts were revised; Conrad was appointed State Paleontologist, and Hall was put in charge of a new Fourth District in western New York, with former Rensselaer students Horsford, Carr, and George W Boyd as his assistants. When the survey ended in 1841, only Hall and Emmons remained in New York. Hall became State Paleontologist, and Emmons became State Agriculturalist.

Lardner Vanuxem, who had studied in France, had been instrumental in introducing to America the value of fossils for subdividing strata and correlating those of similar age from place to place based upon similar fossils. Meanwhile, Timothy Conrad had gained a reputation for studies of Cenozoic fossils of the coastal plain. Thus the survey had strength in palaeontology from the start, and its staff soon developed a New York stratigraphy that set the precedent for naming stratigraphical divisions after geographical localities. Young Hall's career blossomed quickly after the monograph on the fossils and stratigraphy of the Fourth District was published in 1843. This and the other survey reports soon aroused much interest in Europe, where Palaeozoic fossils were being used to define stratigraphical subdivisions during the mid-nineteenth century. For example, Roderick Murchison's *Silurian System* appeared in 1839 (*see* **Famous Geologists:** Murchison), John Phillips's Palaeozoic Series was proposed in 1840, and Joachim Barrande's monographs on the lower Palaeozoic fossils of Bohemia began to appear in 1852. These and other authors began corresponding with Hall, and European geologists began beating a path to Albany – most notably Charles Lyell (*see* **Famous Geologists:** Lyell) during several American visits in the 1840s. During a visit in 1846, Eduard de Verneuil, a close associate of Murchison, tried to persuade Hall not to introduce the name Cambrian to the New World, but rather to use only Silurian for the lowest Palaeozoic strata – a reflection of the famous Murchison–Sedgwick feud then raging in Britain. Hall, however, was not swayed, for he was a leading exponent of the widely held 'nationalistic' view that an American stratigraphical classification was best for America.

As geological investigations in America began to mature, stratigraphical nomenclature was becoming important, especially for comparisons among different regions. Hall and others proposed that an organization be created to deal with nomenclature and other mutual problems, and so in 1838 in Albany the American Association of Geologists was created; the first formal meeting was held in Philadelphia in 1840. From this organization evolved in 1857 the American Association for the Advancement of Science, which was modelled on the British Association. Still later, the Geological Society of America was spawned in 1888 from a division of the American Association for the Advancement of Science. Hall was promptly elected President (**Figure 1**).

Figure 1 James Hall in 1856 at the peak of his career and only one year before his famous Presidential Address to the American Association for the Advancement of Science, in which he first proposed his theory of mountain formation. (From Clarke JM (1921) *James Hall of Albany Geologist and Paleontologist, 1811 1898*. C. Ayer Company Publishers.)

The Albany Training Ground

In 1857, Hall constructed a substantial brick laboratory building, where he worked for the rest of his life. This Albany laboratory became a veritable training school for a host of budding geologists who would distinguish themselves in the history of American science. Although universities were beginning to offer formal instruction in geology during the mid-nineteenth century, there was practically no instruction in palaeontology. As a result, apprenticeship was the principal route into that field, and Hall's laboratory was the place to apprentice. Among the many who profited from association with Hall were Charles E Beecher, Ezra S Carr, John M Clarke, Nelson H Darton, Grove K Gilbert, Ferdinand V Hayden, Eban N Horsford, Joseph Leidy, W J McGee, Fielding B Meek, Charles S Prosser, Carl Rominger, Charles Schuchert, Charles D Walcott, Charles A White, Robert P Whitfield, Josiah D Whitney, Charles Whittlesey, and Amos H Worthen.

Hall's assistants learned more from him than just palaeontology, however, for they also experienced a strong, egotistical, and irascible personality. Although his sharpest attacks were reserved for his legislative enemies, most of his assistants were also treated to his notorious outbursts. In addition to throwing vituperative verbal daggers, he sometimes brandished menacingly a stout cane or even a shotgun, kept at the ready near his desk. Perhaps the most extreme self-righteous attack was upon James T Foster, a school teacher in Greenbush, New York. Foster had the audacity to publish a popularized geological chart in 1849. Hall was so outraged that he stole aboard a boat bound for New York City and threw the entire printing of the offensive chart into the Hudson River. He had quite a time fighting the subsequent libel suit, which entangled him, Louis Agassiz (*see* **Famous Geologists:** Agassiz), James D Dana, and several other notables for several years.

Another celebrated example of Hall's irascible temper involved the prominent British geologist Charles Lyell, during his first visit to America in 1841–42. At first Hall and others were greatly flattered by the attentions of their famous visitor, but Lyell's insatiable questioning, which earned him the nickname 'Pump', and his copying of the Americans' geological maps gradually provoked resentment and a fear of being pre-empted. In March 1842, an anonymous letter signed 'Hamlet' appeared in a Boston newspaper, charging Lyell with geological piracy. It was written by Hall after some of his compatriots criticized him for being too generous in sharing the results of his research with Lyell, especially by giving him a copy of his as yet unpublished *Geologic Map of the Western and Middle United States*. Needless to say, this letter cast a chill upon the Association of American Geologists' meeting a month later, but Lyell participated as if nothing had happened. Although the charge was largely true, Hall was afterwards mortified by his rash act. For once, however, he managed to mend the damage done by his intemperate action, and he remained thereafter on good terms with Lyell.

Almost as legendary as his paranoiac outbursts was Hall's acquisitiveness for fossils. He employed every conceivable means to acquire outstanding collections. An effective technique was to flatter and invite collectors to work with him in Albany and to bring their collections. Commonly, however, when the apprentice moved on his collection did not. Hall was a workaholic who drove himself as mercilessly as he did his assistants. He could rarely say 'no' to even the most ridiculous schemes, and he ignored the entreaties of friends to ease his pace for the sake of his health.

Beyond New York

As he completed his Fourth District studies, Hall decided to see how far the New York stratigraphical classification might apply beyond his state. In 1841, he made the first of several odysseys west. With geologist David Dale Owen he made a boat trip down the Ohio River to Owen's base at New Harmony, Indiana, and, from there, he proceeded across Illinois to Missouri, Iowa, and Wisconsin. Hall was amply rewarded with evidence to support the extension of the New York stratigraphy in a broad way across the entire region. There were some significant differences, however, which he, and perhaps only he, could recognize. For example, he found that the Palaeozoic strata were much thinner to the west of New York and that there were important differences in sedimentary facies, with more clastic sediments in the east and more carbonate strata to the west. In effect, Hall had discovered the contrast between what would, much later, be termed the stable craton and the Appalachian orogenic belt. This trip also provided information to allow him to complete the *Geologic Map of the Middle and Western States*, which was incorporated into Hall's Fourth District report of 1843 – the map that Lyell had used to help prepare his own geological map of the then United States, which was published in 1845 in *Travels in North America*.

Hall's finances were always tenuous. He was easily drawn into risky ventures and also had his salary cut, or even suspended, by a frequently hostile state legislature. At least once he had to sell some of his fossil collections in order to raise money. As his reputation grew, however, opportunities for temporary outside employment helped to tide him over his New York financial droughts. These ventures also allowed him to expand his knowledge widely. One of the first such ventures took him to the Lake Superior region in 1845 to examine copper deposits for a private company. In 1847, the Federal Government authorized a geological survey to evaluate the mineral resources of northern Michigan and Wisconsin. In 1850, Hall was engaged to provide his expertise on Palaeozoic stratigraphy and palaeontology for that survey. He made two brief trips to the area (1850 and 1851), from which he gained further insights into the stratigraphy of the Great Lakes region and added to his ever-growing fossil collections. Perhaps the most important result of his work for this survey, however, was the recognition of fossil reefs in the Silurian strata of south-eastern Wisconsin. This was the first recognition of ancient reefs in North America, and perhaps in the world.

When asked to study fossils from western regions, which others had collected during various expeditions, he willingly obliged. He recognized the first known Mesozoic fossils collected by John C Fremont in the 1840s. In 1853, he agreed to let his assistants Fielding B Meek and Ferdinand V Hayden go to the White River badlands of Nebraska Territory (now in South Dakota) to collect newly discovered Cenozoic non-marine invertebrate and mammalian fossils. Meek, whose artistic as well as collecting skills were vital to Hall's enterprise, was glad to escape from his mentor for a few months. Eventually he extricated himself from Hall's empire to join the new United States Geological Survey. Meek never forgave his perceived exploitation by Hall.

When Iowa decided to undertake a geological survey in 1855 and needed a director, the Governor looked to New York, which had eclipsed all other states as well as the Federal Government in the calibre of its geological survey. Hall accepted the position with alacrity as his New York salary had been suspended in 1850 by a more than usually hostile legislature. Moreover, he welcomed the opportunity to obtain and study fossils from the new state. He soon suggested Amos Dean of Albany as the first Chancellor of the University of Iowa. Hall himself was identified as the first Professor of Geology but apparently he never lectured there. In fact, Hall mostly directed the survey from Albany and spent little time in Iowa. Four assistants did most of the actual work: Josiah D Whitney concentrated upon mineral resources, while Amos H Worthen of Illinois dealt with palaeontology, assisted by F B Meek and R P Whitfield. Hall knew that Worthen had the finest collection of crinoids in the country, so a condition of his employment was that Hall be allowed to describe them, which he did in the Iowa Survey report. Hall came to Iowa for the winter meetings of the legislature to lobby on behalf of the Survey, but payment of salaries was so erratic that he had to borrow money in Albany to keep the effort going. Finally in 1859 the survey was suspended, but two volumes had appeared in 1858.

In 1857, Illinois undertook a geological survey, and Worthen was one of three applicants to direct it. Hall wrote a glowing endorsement of him, but he also supported the other two applicants. This lapse of judgement earned him the animosity of all three applicants, and, in the end, Hall was denied access to the fossils collected by the Survey, which was a great disappointment to him.

While still working in New York and Iowa and for the Canadian Geological Survey, in 1856 Hall accepted an affiliation with Wisconsin. He joined a former Renssalaer colleague, Ezra Carr, now a professor at the University of Wisconsin, and Edward Daniels for this new effort. Hall devoted little time

to the Wisconsin initiative, so Carr and Daniels were really in charge. Whitney was engaged to study the lead deposits of south-western Wisconsin, and Charles Whittlesey was engaged to study the mineral deposits of northern Wisconsin. A large volume was published in 1862, but a hostile Wisconsin legislature abruptly terminated the endeavour because it judged the results to be insufficient. It cared only about potentially economic results, so a frustrated Hall and his assistant, Robert P Whitfield, published Wisconsin's palaeontology within a New York report in 1867 (and again separately in 1871). This ingenious solution to a publication problem was typical of Hall. Much earlier he had circumvented a New York legislative edict to limit the number of expensive palaeontological monographs published simply by issuing several volumes as subdivisions of a single Part of the series, resulting ultimately in 13 separate monographs – at least twice the intended limit.

Hall became involved in several other state surveys in various capacities, ranging from advising about personnel to acting as a consultant for palaeontology or the titular head of a survey. Included were surveys of Missouri (1853 and 1871) and California (1853–1856), the transcontinental railroad survey (1853–1857), and surveys of New Jersey (1854–1857), Ohio (1854–1857), Texas (1858), Mississippi (1858), Michigan (1869–1870), and Pennsylvania (1870–1875). While this list is a testimony to his prominence, Hall's contributions to these many surveys were minor except for the identification of fossils.

Hall made his last trip to the Midwest in 1889, at the age of 77, while first President of the new Geological Society of America. His purpose was to obtain brachiopods by any and all means necessary for his latest project, namely to revise the description and classification of that great group of Palaeozoic fossils. In addition to successfully obtaining many specimens, he also met and lured to Albany a young Charles Schuchert of Cincinnati, who was destined to become his most famous protégé and a professor at Yale. The ambitious brachiopod study culminated in the last volume, Part 8, of the *Paleontology of New York*, which appeared in 1894.

During the completion of the brachiopod monograph, Hall had his last and sweetest wrangle with New York bureaucracy. The Executive Secretary of the Regents, which oversaw his programme, had become overly zealous in trying to impose strict accounting and efficiency procedures. Such a fuss developed that the legislature had to intervene. To resolve the fracas, it appointed crotchety old Hall as State Paleontologist and State Geologist for life, with complete managerial freedom. Doubtless the legislators realized that Hall's days were numbered, and, in fact, he died 3 years later. Hall must have recalled with satisfaction an earlier observation, when a particularly vicious political enemy died suddenly, that "Providence was usually on my side".

The Origin of Mountains

Hall is most widely known for his theory of mountains, which embodied the concept of the geosyncline, a term coined not by Hall but by James D Dana in 1873. In his 1857 Presidential Address to the American Association for the Advancement of Science, Hall startled his audience with a discourse on the origin of mountains rather than speaking about palaeontology and stratigraphy. In stating that "the greater the accumulation, the higher will be the mountain range", he pronounced that a great thickness of strata was a prerequisite for a mountain range composed of folded strata. Hall rejected the then popular theories of mountains of Frenchman Élie de Beaumont – that mountains formed as a result of global cooling and contraction – and the American brothers William B Rogers and Henry D Rogers, who postulated that catastrophic wrinkling of the crust resulted from wave-like movements in a fluid subcrustal zone. Instead, Hall was influenced by a suggestion by the British astronomer John F W Herschel in 1836, which anticipated the modern theory of isostasy. Herschel argued that vertical movements of the crust are caused by changes in pressure and heat at depth, which in turn are the result of erosion and deposition at the Earth's surface. The vertical adjustments towards gravitational equilibrium were accommodated by a pliable subcrust. The key element for Hall was the accumulation of thick sedimentary layers, which he imagined must depress the crust and, in the process, become wrinkled to form the structures seen in ranges such as the Appalachians. He envisioned compression of the upper layers and tension of the lower ones during subsidence – much as occurs when bending a ream of paper (**Figure 2**).

In 1859, Hall published the following, in the most commonly quoted source for his theory, Part 6 of the *Paleontology of New York* (*Volume 3: Descriptions and Figures of the Organic Remains of the Lower Helderberg Group and the Oriskany Sandstone. 1855–1859*, pp. 70–73. Albany: New York State Geological Survey):

> ...[t]he line of greatest depression would be along the line of greatest accumulation [that is] the course of the ori ginal transporting current. By this process of subsidence ... the diminished width of surface above caused by this curving below, will produce wrinkles and folding of the [upper] strata. That there may be rents or fractures of

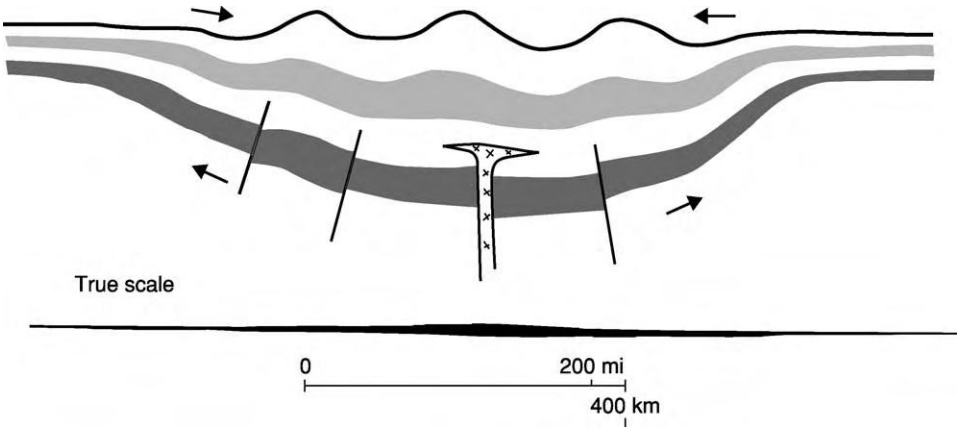


Figure 2 James Hall's theory of down warping resulting from sedimentation. The upper layers are crumpled as their circumference diminishes, whereas the lower layers are broken by tension, which allows dykes to be intruded from below. Hall never published diagrams of his theory, so this was constructed from his verbal discussion. Modified with permission from Dott RH Jr (1985) James Hall's discovery of the craton. In: Drake ET and Jordan WM (eds.) *Geologists and Ideas: A History of North American Geology*, pp. 157 167. Centennial Special, Volume 1. Boulder: Geological Society of America.

> the strata beneath is very probable, and into these may rush the fluid or semi fluid matter from below, produ cing trapdykes, but the folding of strata seems to be a very natural and inevitable consequence of the process of subsidence.

A year earlier, in the report of the Iowa Survey (1858), Hall had also emphasized the contrasts in thickness between the Appalachian region and the Midwest, with detailed remarks about contrasting facies as well as thicknesses in various portions of the Palaeozoic successions of the two regions. Here, too, he included a brief summary of his theory of mountains by stating that "The thickness of the entire series of sedimentary rocks, no matter how much disturbed or denuded, is not here great enough to produce mountain features" (Vol. 1, p. 42).

Hall was vague about the cause of mountain uplift. He simply ascribed it to continental-scale elevation, which he thought had no direct relation to the folding of strata within the mountains and which he did not attempt to explain. Contemporaries were quick to challenge him on this point, with Dana noting that Hall had presented "a nice theory of mountains with the mountains left out". Hall lamely denied that he ever intended to offer a complete theory of mountain building. His failure to publish the Presidential Address until 1883 may have been because of such criticisms, but his first priority was always palaeontology and he knew that the essence of his theory was to appear in both the Iowa and New York reports (as well as in an abstract in Canada) soon after his oral address.

Hall's contribution to mountain-building theory was marginal at best and was soon eclipsed by Dana's more profound and comprehensive contraction theory, which postulated that thick strata were a result of mountain-building processes rather than the cause. Nonetheless, Hall's emphasis on a cause-and-effect relationship between orogenic belts and very thick strata had a significant influence on three generations of geologists. By coining the term 'geosynclinal', which was later converted to the noun 'geosyncline', Dana formalized Hall's demonstration that Palaeozoic strata are ten times thicker in the Appalachian Mountains than in the lowlands to the west (the craton).

Even though Hall was wrong about the cause of mountain building, he was nevertheless the first person to underscore the profound stratigraphical contrasts between orogenic belts and what are now termed stable cratons. He drew attention at an early stage to large-scale stratigraphical patterns among some of the larger tectonic elements of the Earth's crust and had other shrewd stratigraphical insights that were ahead of the times. By virtue of his breadth of experience in both the cratonic and the orogenic regions of eastern North America, he was uniquely equipped to see this fundamental distinction. Coupled with his prodigious contributions to palaeontology, this assured James Hall of a prominent niche in the history of geology.

See Also

Analytical Methods: Geochronological Techniques. **Famous Geologists:** Agassiz; Lyell; Murchison; Sedgwick. **Geological Surveys**. **Stratigraphical Principles**. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Clarke JM (1921) *James Hall of Albany Geologist and Paleontologist, 1811 1898*. Albany: Privately printed.

Dana JD (1873) On some results of the Earth's contraction from cooling, including a discussion of the origin of mountains, and the nature of the Earth's interior. *American Journal of Science, 3rd series* 5: 423 495; 6: 6 14; 104 115; 161 172; 381 382.

Dott RH Jr (1979) The geosyncline first major geological concept 'Made in America':. In: Schneer CJ (ed.) *Two Hundred Years of American Geology*, pp. 239 264. University Press of New England: Durham, New Hampshire.

Dott RH Jr (1985) James Hall's discovery of the craton. In: Drake ET and Jordan WM (eds.) *Geologists and Ideas: A History of North American Geology*, pp. 157 167. Centennial Special, Volume 1. Boulder: Geological Society of America.

Fisher DW (1978) James Hall patriarch of American paleontology, geological organizations, and state geological surveys. *Journal of Geological Education* 26: 146 152.

Hall J (1842) Notes upon the geology of the western states. *American Journal of Science and Arts*, 1st series, 42: 51 62.

Hall J (1989) *The Natural History of New York. Part 6. Palaeontology of New York. Vol 3. Descriptions and Figures of the Organic Remains of the Lower Helderberg Group and the Oriskany Sandstone. 1855 1859*. Albany: New York State Geological Survey.

Hall J (1883) Contributions to the geological history of the North American continent. *Proceedings of the American Association for the Advancement of Science* 31: 24 69.

Hall J and Whitney JD (1858) *Report on the Geological Survey of the State of Iowa*. Des Moines: Legislature of Iowa.

Hutton

D R Oldroyd, University of New South Wales, Sydney, Australia

Introduction

The Scottish geologist, physician, farmer, philosopher, chemist, businessman, and industrialist James Hutton (1726–1797) is commonly regarded as the 'founder of modern geology', though a similar claim could be made for some others, and in some ways Hutton's thinking was not at all modern by today's standards.

Hutton's Early Career and the Beginning of His Interest in Geology

Hutton was born in Edinburgh, the son of a businessman who served for a time as City Treasurer. The young Hutton went to Edinburgh University at the age of fourteen, where he studied humanities, attended the mathematics lectures of Colin Maclaurin, and acquired a keen interest in chemistry. After his time as a student, he was briefly apprenticed to a solicitor, but eventually decided to study medicine. After taking the Edinburgh course, he went to Paris in 1747 and thence to Leiden, where he submitted a doctoral thesis in 1749. This dealt with the circulation of the blood and matters of human physiology and had a distinct chemical slant. The thesis title referred to the human body as the 'microcosm', which was traditionally regarded as having analogies with the Earth or with the whole cosmos (the 'macrocosm'). The thesis may have been the seed from which sprang Hutton's later cyclic theory of the Earth.

On returning to Britain, Hutton did not take up medicine. Instead, he went into partnership in an industrial process for extracting sal ammoniac (ammonium chloride) from soot. But Hutton sired a son, probably out of wedlock, and 'tactfully' left Edinburgh for several years to pursue a career in agriculture on two farms in Berwickshire, which he had inherited from his father. He wished to do his farming on a scientific basis, so he went to East Anglia to study the latest methods of agriculture, which he subsequently brought to Scotland. During his two years away Hutton travelled extensively and, as he became increasingly interested in the Earth, recognized the ubiquity and perpetuity of erosion and deposition and that sedimentary rocks were consolidated sediments.

Hutton worked his farms himself and experimented with agricultural techniques. Under Maclaurin, he had become acquainted with the principles of 'deism', and he had apparently lost his Christian faith at an early stage of his life. According to the deist view, God had created the Earth 'in wisdom' as a suitable place for human habitation. The existence of God was not known by courtesy of Jesus Christ, the Bible, the Church, or any other agent of revelation, but by human reason. For Hutton, divine design was manifest in Nature itself, both in the way organisms functioned and were structured and in the way the Earth was apparently well 'contrived' for human existence (with air, water, soil, animals, plants, etc. all suited to us).

But it was obvious that soil was constantly being washed into the sea, and, since it was essential for human well-being, it had somehow to be replenished. As a deist, rather than a biblical literalist, Hutton could take a grand view of time. The Earth could be millions of years old, but in that case the land would eventually be eroded to a plain and the good soil would end up as sediments in the seas. So Hutton asked himself how high ground could be regenerated to provide a source of new soil.

Hutton's Theory of Cyclic Earth Processes

In 1764, Hutton made a journey into the Highlands and began to collect geological information and specimens in a systematic manner. His farms were by then profitable and the sal ammoniac business was prospering. So he began to think of returning to Edinburgh, now as a gentleman–farmer. Probably in 1767, he rented out his farms and returned to Edinburgh (his old scandal had by then been forgotten or forgiven) to enjoy the pleasures of intellectual life in one of the great cities of the eighteenth-century Enlightenment.

Among Hutton's new friends were the economist Adam Smith, the chemist Joseph Black, who conducted experiments on heat, and the engineer and steam-engine inventor James Watt. It is likely that Watt's engines encouraged Hutton to think of heat as an agent of geological change. Perhaps the Earth had a central source of heat that might somehow drive the cycle essential for a theory of the Earth that provided for a renewal of soil? The Earth's internal heat could be analogous to the fire of Watt's engine, which drove the complicated mechanism of the engine and the machinery of a factory. But Hutton did not imagine that the Earth's internal heat was due to combustion.

Hutton's theory of the Earth was, then, developed as follows. The Earth, he thought, had a central reservoir of heat, the source or means of maintenance of which was unspecified. Rocks on the surface were broken down by weathering and erosion to form soils. Sediment was deposited in the seas by rivers, which also carved valleys. Sediments accumulated in layers on the ocean floors, and the lower layers were compressed and consolidated by the sediments deposited on top of them, assisted by the Earth's internal heat. The rock-salt deposits of Cheshire seemed to Hutton to have been melted at some time. Likewise, the grains of sand in quartzites seemed to show evidence of fusion at their edges in the process of consolidation by heat.

In time, the consolidated materials, under pressure, might become so hot that they would melt. Veins of crystalline rock, dykes or sills, could be emplaced. Moreover, Hutton supposed, great masses of molten material (which we would call magma) could be intruded into the Earth's crust, heaving it up. On cooling, this magma might crystallize to form subterranean masses of granite, which might subsequently be exposed by weathering and erosion. Thus the land would be renewed and Hutton's Earth, 'designed in wisdom', would continue indefinitely as a place suited to human habitation. The upheaval of strata was confirmed by the presence of marine fossils in strata well above sea-level. However, at the time of the first public presentation of his theory, Hutton appeared to have personal knowledge only of mineral or metallic ore veins, not granitic veins, and he did not then describe any personal examinations of large granitic bodies. He went looking for these systematically only after the preliminary presentation of his ideas.

Be that as it may, the whole process envisaged by Hutton was cyclic, for the upheaved strata would be eroded to form a new surface, on which other sediments might subsequently be deposited. So one might hope to find places where the lower layers were inclined to the horizontal and the overlying ones lie over them horizontally. Such a structure came to be known as an unconformity, and the subsequent discovery of unconformities was considered a triumph for Hutton's theory, as he apparently had the idea of such structures before he actually saw them. An unconformity could be taken to mark the end of one cycle and the commencement of the next.

Hutton's cyclic Earth processes were continuous and open ended. He did not say that the Earth was infinitely old, but as he put it in a famous sentence: 'we find no vestige of a beginning—no prospect of an end' Hutton (1788 p. 304). His cycle has been called the 'geostrophic cycle' (see **Figure 1**, which explicates Hutton's notion of unconformity).

Hutton's theory was formally read before the Royal Society of Edinburgh in 1785 and published in 1788. It appeared in expanded form in two volumes in his *Theory of the Earth* in 1795. Two further incomplete volumes remained unpublished in his lifetime, but the manuscripts were found in the nineteenth century and published in 1899 as Volume 3. This book described Hutton's fieldwork after the presentation of his 1785 paper.

Hutton's 1785/1788 paper did not explain the Earth's internal heat, but he tried to use field specimens to support its existence. Many of the materials that bind sediments together, such as calcareous spar, silica, etc., are not themselves water soluble.

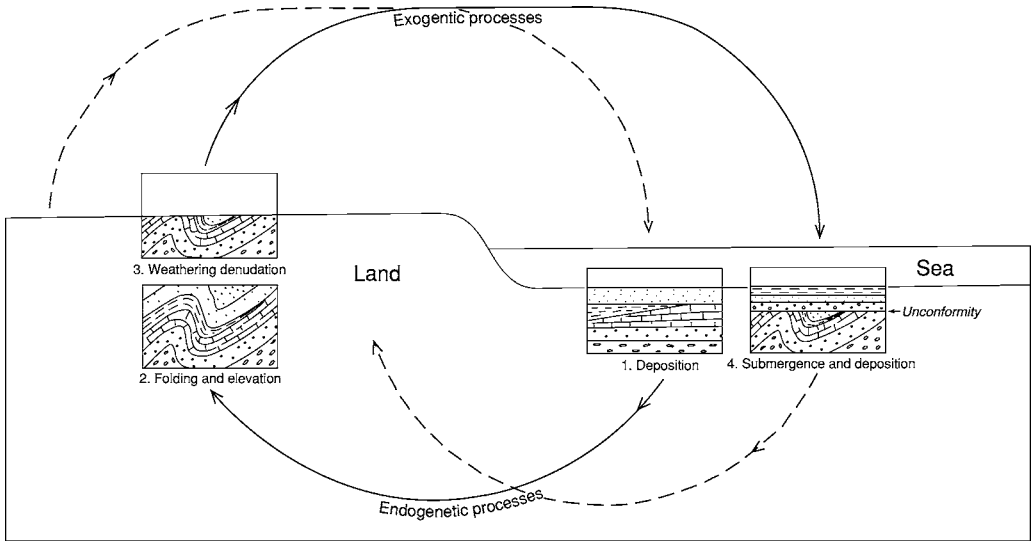


Figure 1 Representation of the 'geostrophic cycle'. Reproduced by permission of The Geologists' Association from *Proceedings of the Geologists' Association*, Tomkeieff SI, Unconformity an historical study 1962, **73**, pp. 383 417, fig. 6. © 1962 The Geologists' Association.

Therefore, Hutton argued, water could not have been the prime agent causing their consolidation. But heat could penetrate into bodies and, by fusion, could cause consolidation. So, he thought, heat must have been responsible for the penetration of sediments by flint nodules (which were injected while molten). Likewise, nodules containing crystalline spar that did not extend to their outer surfaces could not, Hutton supposed, have acquired the crystalline matter by transmission of solutions. Hutton also exhibited a specimen of 'graphic' granite, which contained crystals of quartz within feldspar within quartz. Such a texture could not, he maintained, have been produced by crystallization from aqueous solution.

There were evidently gaps in the evidence for the cyclic chain of Hutton's theory. Hutton argued that there had to be heat within the Earth and there had to be some means of elevation, even if he did not know precisely how that process worked. The hot interior was supported by the evidence of volcanoes, of course, and mines seemed to have higher temperatures at greater depths (but that was not proven by careful measurements until the nineteenth century).

Geological Evidence to Support Hutton's Theory

So Hutton's 1785/1788 paper was not in itself sufficient to persuade all his auditors or readers. At the time of its presentation, he had not recorded observations of veins of granite penetrating other rocks, nor, so far as we know, had he discovered any unconformities (although they had been reported by others without their theoretical significance being recognized). But following the public presentation of his ideas Hutton made excursions to various parts of Scotland to look for confirmatory field evidence.

In September 1785, Hutton went into the Grampians to hunt for contacts between granite and surrounding rocks into which it might have been injected while in a fused state. He was accompanied by a friend, John Clerk of Eldin, who made excellent drawings of what they saw. Hutton thought that he would find the evidence he wanted to the west of the mass of Aberdeen granite. But he may have received some hints of where to look (possibly from Clerk), for they headed directly for the valley of the River Tilt, which runs north-east from Blair Atholl. Complicated outcrops of limestones and schists were found in the valley floor, and not far up the glen they came across fine exposures of granitic veins, which sometimes cut across the country rock and elsewhere could be seen anastomosing between, or across, the laminae of the country rock. Hutton got so excited that his guides imagined that 'nothing less than . . . a vein of silver or gold . . . could call forth such strong marks of joy and exultation'! The granitic veins were also traced back to the large mass of granite on the north side of the glen. Hutton's joy was, of course, due to the fact that he had found what he had predicted on the basis of his theory, and the geometry of the veins was compatible only with the granite having worked its way into the country rock from below and from the granite mass.

In 1787, Hutton visited the Isle of Arran, which has a large mass of granite at its northern end, with

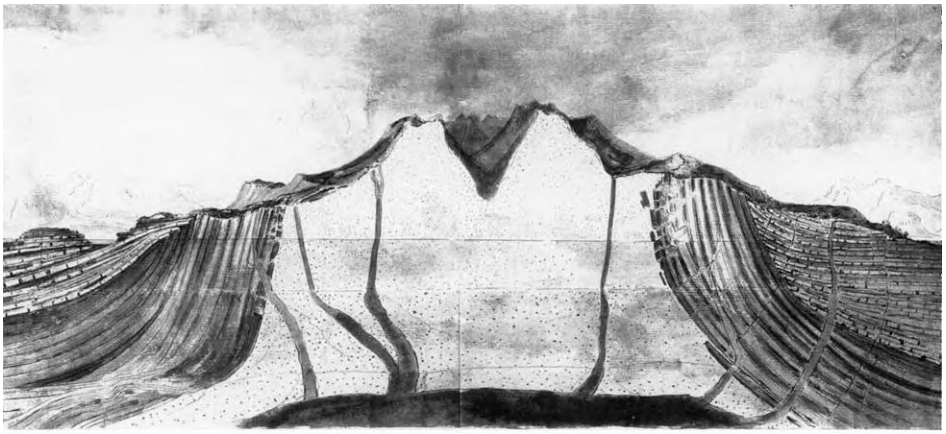


Figure 2 Cross section of the northern part of Arran, drawn by John Clerk Jr (1787). Reproduced by permission of Sir Robert Clerk of Penecuik.

surrounding tilted-up layers of schist and beds of sandstones and other sediments that are stratigraphically above the schists (but lower in terms of altitude). A large block of schist traversed by granite veins was brought back to Edinburgh to convince critics of the virtues of his theory. John Clerk's son (of the same name) (1757–1812) accompanied Hutton and produced a wonderful cross-section of the island, construed in terms of Hutton's theory (Figure 2). This section is congruent with a geological map of the area, based upon modern knowledge, which shows a domed structure of schists and sedimentary rocks, disposed around a central core of granite.

Also, at Loch Ranza on the northern tip of Arran, Hutton found his first unconformity, with the sedimentary strata (sandstones and limestones) lying over the inclined or almost vertical schists (Figure 3). Again he had found a state of affairs that he had predicted from his theory. A 'swarm' of basaltic dykes was also observed by the shore of the southern part of the island.

The Loch Ranza unconformity was not, however, wholly convincing, for the rocks were obscured by vegetation. But on their way home the travellers found a much clearer example near Jedburgh, where the road ran by the banks of the River Jed and a section revealed a splendid view of Old Red Sandstone lying horizontally over the up-ended grey gritty sandstone that is now known as Silurian greywacke (which Hutton called 'schistus' although it was not a schist).

The most famous discovery of an unconformity occurred in 1788. Hutton's upland farm was situated on 'schistus', while his main farm was on soil derived from Old Red Sandstone. The contact between the two rock types ran northwards to the coast. Hutton must have been aware of the two rock types, which

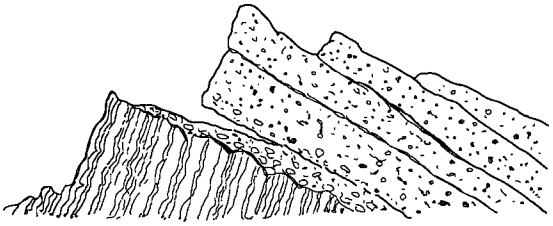


Figure 3 Unconformity at Loch Ranza, Arran, as described by Hutton and figured by Sir Archibald Geikie. Reproduced from Hutton (1899). *Theory of the Earth. . .Vol. III* Edited by Sir Archibald Geikie, p. 235. London: Geological Society.

were similar to those that he had seen at Jedburgh. It seemed a good plan to examine the coast, where an unconformity might be exposed. Accordingly, with his friends John Playfair (1748–1819), Professor of Mathematics at Edinburgh, and a local landowner, Sir James Hall (1761–1815), a keen amateur scientist who did some of the first experimental geology, Hutton sailed along the Berwickshire coast, past the schist terrain to that of sandstone. As anticipated, they encountered an unconformable contact, at a place called 'Siccar Point', with Old Red Sandstone overlying the schist, as at Jedburgh, but exposed in such a way that the three-dimensional structure of the contact could be examined.

The famous excursion was described in Playfair's biography of Hutton. The three men were aware that if the Siccar Point exposure was interpreted through the lens of Hutton's theory then it entailed the passage of a vast amount of time. The sediments of the greywacke were first deposited horizontally under the sea and consolidated by heat and the pressure of superincumbent material. Then the area was upheaved by forces acting from within the Earth, with hardening of the sediments by heat and pressure. The

forces were such that the 'schistus' now stood almost vertically. But the strata were then subjected to weathering and erosion so that the upheaved rocks were reduced to an approximately level surface. In time, the now vertical strata subsided below the sea once more (by an inadequately explained process) and were covered by layers of sediment derived from adjacent high ground. Again there was consolidation, following which uplift occurred, exposing the strata to the elements once more, but without the overlying sandstones being folded or inclined. Thus the disposition of the rocks observed at Siccar Point could be understood – provided that time was unlimited. Playfair wrote in his biographical memoir of Hutton:

> Revolutions still more remote appeared in the distance of this extraordinary perspective. The mind seemed to grow giddy by looking so far into the abyss of time; and while we listened with earnestness and admiration to the philosopher who was now unfolding to us the order and series of these wonderful events, we became sensible how much further reason may sometimes go than imagination can venture to follow.
>
> Playfair (1805, p. 73)

Thus at Siccar Point Hutton provided evidence for (but not formal proof of) the Earth's great age and the cyclic nature of geological processes. The locality has long been recognized as one of geology's most significant field sites. Hutton acquired Playfair as a convert to his theory, and it was Playfair who popularized Hutton's ideas – Hutton's prolix style and confusing theory of heat (see below) did not gain him many adherents.

Hutton's Later Work on the Theory of Heat

In his old age, Hutton tried to give some kind of physicochemical explanation of the forces causing elevation and subsidence, but he had little success. He knew that bodies expanded when heated, and the kind of heat that produced this effect he called 'sensible heat'. He also knew that when heat was applied to a solid it increased in temperature, but on reaching the melting point it would melt without changing temperature, even though it was still being supplied with heat. In the change of state, the heat supplied to produce melting was somehow hidden. Black had called this 'latent heat'.

But the nature of heat was uncertain. Hutton thought it was a kind of weightless 'substance'. He knew that everyday objects have mass and that massive bodies are attracted to one another by gravitation. But there also seemed to be repulsive forces at work, as for example when water is boiled: steam engines exert pressure in their cylinders.

Today, we distinguish between radiant heat and heat transmitted by conduction. Hutton had no adequate concept of radiation, but he knew that heat from the sun shines on us, across space. He called it 'solar substance', and, though weightless, it somehow seemed to be absorbed by plants, though Hutton did not know how. Adding to the complications, Hutton accepted the old 'phlogiston theory' of combustion (which was collapsing at the end of the eighteenth century), according to which an inflammable material contains a weightless 'substance' or 'principle' called 'phlogiston', which is dispersed into the atmosphere during combustion. Hutton was inclined to suppose that 'solar substance' and 'phlogiston' were one and the same. (Actually, if one regards 'phlogiston' as energy, then some of the problems that Hutton was trying to understand fall into place for us.)

Hutton grappled with such problems in two books: *Dissertations on Different Subjects in Natural Philosophy* (1792); and *A Dissertation upon the Philosophy of Light, Heat, and Fire* (1794). All his arguments cannot be followed here, but he tried out the idea that objects normally attracted one another according to the inverse-square law of gravitation. Thus he spoke of 'gravitating matter'. At very close quarters, however, objects supposedly began to repel one another, according to a force law in which the distance between particles was raised to a power greater than two. The repulsive force (or 'solar substance') could supposedly take various guises: 'sensible' heat, manifested by expansion; latent heat; light; electricity; and phlogiston. So, when sediments were under extreme pressure, they might move from a compressive phase to an expansive (expanding) phase. Hence, in the geostrophic cycle, there could be alternating periods of contraction (compression or consolidation) and expansion (producing land elevation).

Hutton's theory depended on a balance of attractive (gravitational, cohesive, and concretive) and repulsive (specific, or sensible, and latent heats) forces. There could be different resultant states, arising from the forces producing elevation (expansion) and subsidence (contraction) at different times and places. But when, lacking the concept of energy, Hutton started talking about 'solar substance' in reference to solar radiation (as we would say) and thought that this 'substance' was immaterial, confusion and misunderstanding followed amongst his contemporaries; it is scarcely possible for us to make sense of his theory of heat.

Hutton's Legacy

We can see, therefore, that, for all Hutton's success in looking into the 'abyss of time' and his successful predictions of granitic veins and unconformities, his theory had significant lacunae, and the physical explanation of expansion and uplift was not really integrated into his *Theory of the Earth* and attracted little or no following. Expansion was the Achilles Heel of his theory, and the problem remained unsettled for generations. In the end, expansion and elevation were simply assumed by Huttonian cyclists.

People eventually accepted Hutton's cyclic theory, even though they could make little sense of the physical basis he envisaged. But acceptance took time, and geological theory was racked with controversy until well into the 1820s. The Professor of Natural History, Robert Jameson (1774–1854), gained control of geology teaching at Edinburgh, and even Hutton's specimens, for many years. So Huttonian theory tended to be eclipsed in Scotland for decades, despite the best efforts of Playfair and Hall, who, after Hutton's death, conducted experiments that sought to simulate the consolidation of sediments, aided by heat, and to show that limestone heated in a sealed gun barrel could be converted into something like marble without loss of carbon dioxide.

In addition to having excellent ideas about weathering and erosion, the deposition and consolidation of sediments, rates of geological change, the immense age of the Earth, and arguments in favour of geological cycles supported by evidence of unconformities, Hutton clearly appreciated the fact that many valleys have been carved by the rivers that now flow in them. Through second-hand knowledge of the Alps, he suggested that glaciers might have been much larger in the past than at present and could have deposited large blocks considerable distances from where the rock types are found *in situ*. Thus it seems that he envisaged a former colder climate than at present (due to the mountains being higher and carrying more snow) and appreciated the enormous erosive powers of glaciers.

In 1802, Playfair published his *Illustrations of the Huttonian Theory of the Earth*, which set out Hutton's doctrines in improved literary form. Hutton had referred to the ideas of the Swiss geologist Horace Bénédict de Saussure (1740–1799) about the transport of glacial debris by glaciers extended from the Alps to the Jura Mountains, and these ideas were given greater prominence by Playfair, who also wrote about the patterns of river drainage systems. But neither Hutton nor Playfair had the idea of an Ice Age. That came later, principally through the advocacy of Louis Agassiz (1807–1873) (*see* **Famous Geologists:** Agassiz).

Another Scottish geologist, the influential Charles Lyell (1797–1875) (*see* **Famous Geologists:** Lyell), accepted many of Hutton's ideas and made them almost paradigmatic, handing them on to another Scot, Archibald Geikie (1835–1924), who coined the methodological maxim: 'the present is the key to the past'. But that principle was already well established by Hutton. He used his knowledge of what he could see going on around him – on his farms and during his travels – to develop a theory about how the Earth operated as a system and how it might have been in the remote past. But Hutton's cycles were not identical. There could be local variations from one phase to the next. The Earth had a history, while operating in a law-like manner, so as to be in a steady-state when viewed on a grand scale. We can also credit Hutton with advancing the concept known today as 'deep time' – and for doing so by geological reasoning.

See Also

Famous Geologists: Agassiz; Lyell. **History of Geology Up To 1780**. **History of Geology From 1780 To 1835**. **Igneous Rocks:** Granite. **Unconformities**. **Weathering**.

Further Reading

Baxter S (2003) *Revolutions in the Earth: James Hutton and the True Age of the World*. London: Weidenfeld & Nicolson.

Dean DR (1992) *James Hutton and the History of Geology*. Ithaca: Cornell University Press.

Donovan A (1978) James Hutton, Joseph Black and the chemical theory of heat. *Ambix* 25: 176 190.

Gerstner PA (1968) James Hutton's theory of the Earth and his theory of matter. *Isis* 59: 26 31.

Gerstner PA (1971) The reaction to James Hutton's use of heat as a geological agent. *British Journal for the History of Science* 5: 353 362.

Hutton J (1788) Theory of the Earth; or an investigation of the laws observable in the composition, dissolution, and restoration of land upon the globe. *Transactions of the Royal Society of Edinburgh* 1: 209 304.

Hutton J (1795) *Theory of the Earth, with Proofs and Illustrations*. London: Edinburgh: William Creech; London: Cadell, Junior, and Davies (republished in facsimile 1972).

Hutton J (1899) *Theory of the Earth. . .Vol. III* Edited by Sir Archibald Geikie, p. 235. London: Geological Society.

Hutton J (1997) *James Hutton in the Field and in the Study edited by Dennis R. Dean: Being an Augmented Reprinting of Vol. III of Hutton's Theory of the Earth (I, II, 1795), as First Published by Sir Archibald Geikie (1899)*. New York: Scholars' Facsimiles & Reprints, Delmar.

Jones J (1985) James Hutton's agricultural research and his life as a farmer. *Annals of Science* 42: 574 601.
McIntyre DB (1997) James Hutton's Edinburgh: the histor ical, social and political background. *Earth Sciences History* 16: 100 157.
McIntyre DB and McKirdy A (2001) *James Hutton: The Founder of Modern Geology*. Edinburgh: National Museums of Scotland (1st edn, 1997).
Oldroyd DR (2000) James Hutton's 'Theory of the Earth' (1788). *Episodes* 23: 196 202.
Playfair J (1805) Biographical account of the life of Dr James Hutton, F.R.S.Edin. *Transactions of the Royal Society of Edinburgh* 5: 39 99.
Şengör AMC (2001) *Is the Present the Key to the Past or the Past the Key to the Present? James Hutton and Adam Smith versus Abraham Gottlob Werner and Karl Marx in Interpreting History*. Special Paper 355. Boulder: Geological Society of America.
Tomkeieff SI (1962) Unconformity an historical study. *Proceedings of the Geologists' Association* 73: 383 417.

Lyell

D R Oldroyd, University of New South Wales, Sydney, Australia

Charles Lyell (Figure 1) was arguably the most important geologist of the nineteenth century, and his *Principles of Geology* (1st edn, 3 vols, 1830–1833; 11th edn, 1872) was a classic text that exerted much influence on the development of geology, as well as helping to shape the development of Charles Darwin's thought. Lyell's other major works were his *Elements of Geology* (titled *Manual of Elementary Geology* in some editions) (1st edn, 1838; 6th edn, 1865) and *Geological Evidences of the Antiquity of Man* (1st edn, 1863; 4th edn, 1873). He also published two books on his travels in North America.

Lyell was born into a well-to-do family at Kinnordy House, Forfarshire, Scotland, but much of his youth was spent at the family's second home in Hampshire (with a more agreeable climate). He attended a private school in Salisbury and then at Midhurst; thereafter he attended Exeter College, Oxford, where he studied mathematics and classics, but also became greatly interested in geology through the lectures of William Buckland, which students could attend as optional additions to their main curriculum. Lyell's family was considerably interested in natural history, and during his vacations they travelled extensively on the continent. Lyell also made observations on the Kinnordy estate. Even while a student, he was elected Fellow of both the Linnean and Geological Societies.

On leaving Oxford, Lyell started to train for the law at an office in London, but found the work uncongenial and complained of problems with his eyesight, and so, having (limited) independent means, he did not continue in this line of work. Rather, consorting with many of the leading geologists of the day, and travelling widely, he became virtually a full-time gentleman-geologist, being elected to the Royal Society as early as 1826. Two years later when travelling on the Continent and meeting important figures in Paris, etc., he decided to give up legal work altogether. Eventually, he acquired significant income from his geological writings. In 1832, he married Mary Horner, daughter of Leonard Horner, himself a geologist and educationist, who had learned Huttonian theory in Edinburgh. The couple, who had no children, settled in London, where Lyell became established as one of its leading scientists.

Buckland's Oxford lectures were informative and entertaining. He taught the essentials of stratigraphy, and particularly William Smith's idea that strata could be identified and correlated by their fossil contents (*see* **Famous Geologists:** Smith). But Buckland, in the religious atmosphere of Oxford, and trying to show that his science was compatible with the Bible, laid much emphasis on his studies of superficial deposits and cave remains (about which he was an authority) and sought to show that such materials

Figure 1 Charles Lyell (1797 1875).

could be explained as the result of the Noachian Flood, which in terms of biblical chronology, occurred only a few thousand years ago. Thus, Buckland's geology, which had sources in the work of Cuvier (*see* **Famous Geologists:** Cuvier), could (supposedly) lend material support to theological claims. But such a global flood would have been impossible according to the laws of nature as presently acting, and would be incompatible with geological processes seen at work today.

Lyell travelled and thought much during the 1820s; in Scotland he visited such sites as Glen Tilt and Siccar Point (*see* **Famous Geologists:** Hutton). Though greatly stimulated by Buckland, he came to reject his idea on the great role ascribed to catastrophic floods, and leaned towards the geology of Hutton, his father-in-law-to-be having attended John Playfair's 'Huttonian' lectures in Edinburgh. Abraham Werner's Neptunist theories were rejected as being incompatible with the limited solvent power of water and with Hutton's observations. In Forfarshire (which he mapped in 1824), Lyell saw marls being deposited, or already deposited, in fresh-water lakes fed by springs and associated with shells and plant remains. He knew that in the Paris area Cuvier and Brongniart had found similar fresh-water limestones, which, they thought, had no modern analogues. Thus Lyell's thinking was turned towards the idea of explaining geological phenomena in terms of presently occurring processes.

In Huttonian theory, an immense amount of time was required to make possible the great cycles of geological change that he envisaged. The question of time was one that Lyell, therefore, had to consider. Evidence for the Earth's great antiquity was produced during Lyell's journey to Sicily in 1828. He saw the huge still active volcano, Mount Etna, and it was evident that it was made up of successive lava flows. Historical information about recent flows gave an approximate idea of the rate of accumulation of the flows and the build-up of the mountain. The height of the mountain being known, one could thus form an approximate idea of its age. Further, Lyell examined shells in recent-looking strata lying below the lavas. Nearly all were still to be found today in the Mediterranean. So strata and shells of geologically recent appearance were in rocks that were very ancient in human terms, being older than flows from the volcano. (In a subsequent letter to his sister, Lyell offered that on a 'moderate computation' the shells might be 100 000 years old.) So if geologically recent rocks were ancient in human terms, rocks lower in the stratigraphic column must be exceedingly ancient. Evidently the Earth was of enormous age. In this argument, Lyell was assuming that the rate of flows at Etna occurred at approximately equal rates. He was applying the principle that nature was uniform in her operations: what was later dubbed the principle of uniformity. Lyell was a 'uniformitarian' with regard to Etna, as he had been with respect to the lake deposits in Forfarshire.

Also on his Italian journey of 1828, Lyell visited Pozzuoli on the coast near Naples. There he observed three standing columns of a Roman building, then thought to be an ancient temple. These had marks of the borings of marine organisms half way up, which suggested to Lyell that the land there had fallen below sea level since Roman times, and had subsequently risen; all this having happened without the columns toppling over. From this, Lyell inferred that the level of land was rising or falling in different places, as Hutton had previously proposed. Moreover, the processes were not sudden or catastrophic, but gradual.

Following his return to Britain, Lyell began to write his major book, which sought to establish the working methods and procedures of geology. It was to give geology its proper method and fundamental principles: hence its title, *Principles of Geology*. These may be summed up by the adage (as later stated by Archibald Geikie) that 'the present is the key to the past'. Also, for Lyell, geological processes were assumed to be 'gradual'.

Hutton's geology envisaged grand cycles of rock formation, erosion, transport, deposition, consolidation, and subsequent elevation. The rocks of each cycle were not necessarily identical in any given place, and the geologist needed to work out the history of what had happened at each locality. But overall, the earth did not have an historical direction: it did not 'progress'. Things were much the same in the past and present (humans excepted).

Lyell's views were much the same, but he placed more emphasis on fossils. He supposed that conditions were constantly changing at any given locality from one period to the next, because of the local changes of relative levels of land and sea. Climate could change too, according to whether more high land happened to be near the poles at a given time, or nearer the equator, the former state of affairs producing cooler conditions overall. So some forms would become extinct if they failed to meet the conditions of existence.

On this basis, new types of organisms also needed to come into existence from time to time. Lyell presumed that they did so, even though he did not know how this occurred. Further, he assumed that the basic animal types had always been found on the earth. On this view, there was a gradual turnover of species. His model can be represented as shown in Figure 2.

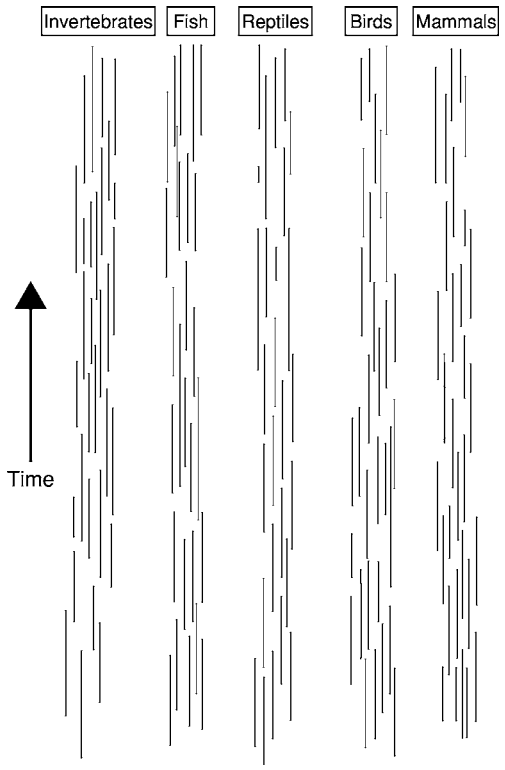


Figure 2 Diagram illustrative of Lyell's theory of species change, with 'random' creations and extinctions.

According to this model, Lyell assumed that the further back the geologist explored in time, the fewer extant species there would be. In fact, nearly all species before the beginning of the Tertiary in Europe would now be extinct. Further the Tertiary could be subdivided according to its proportions of extant fossils. The subdivisions that Lyell proposed were:

- Newer Pliocene 96% recent fossils
- Older Pliocene 52% recent fossils
- Miocene 19% recent fossils
- Eocene 3% recent fossils

Between the Secondary (later called Mesozoic) rocks and the Tertiary there was a period of non-deposition in Europe, so that Secondary fossils were now virtually all extinct. There had been a complete turnover of forms during the stratigraphic time-gap. Likewise, there was a large time-gap and turnover of forms between the Primary (Palaeozoic) forms and the Secondary types. Moreover, the further back one went through the Secondary epoch, the smaller was the percentage of forms resembling those at the top of the Secondary (i.e., in the Cretaceous); likewise, through the Palaeozoic. Lyell regarded marsupial remains found in Secondary rocks in the Stonesfield Slate near Oxford as vindication of his idea that the major animal types went back into the indefinite past. That they could not be found in the most ancient rocks was due to the fact that they had been lost by erosion or metamorphism (an important notion that Lyell first explicitly enunciated). All this was compatible with Lyell's uniformitarianism, but he had no modern empirical warrant for the supposition that new species could somehow come into being.

Lyell's ideas attracted much attention, though most geologists, looking at the stratigraphic record, found it hard to accept that there was no evidence of progress in the fossil record through time. On the other hand, his desire for geology to have its own procedures, with geological processes operating in accordance with the presently observable laws of nature, met with approval, as did his mastery of facts and grasp of palaeontological and stratigraphic detail. He wanted geology to be a science, *sui generis*, distinct from cosmology. Geologists, he thought, did not need a general 'theory of the earth' such as his 18th-century predecessors had sought to offer (though Lyell's cyclic geology was in fact fundamentally the Huttonian theory). But Lyell focused on stratigraphy and palaeontology, not 'hard rocks' such as granite.

In 1831, following the success of Volume I of his *Principles*, Lyell obtained a chair in geology at King's College, London, a newly founded Church of England establishment. By then, the putting aside of the Noachian Flood as a geological agent seemed not to concern the authorities, and Lyell's religious view were considered 'sound'. However, he only gave lectures there in 1832 and 1833. Ladies were allowed to be present for the first course, but were thought to present an undesirable distraction and their further attendance was terminated. In consequence, the attendance fell sharply, and Lyell decided that he was in part wasting his time there, so he resigned to return to his publishing and life as a gentleman geologist. In this he was eminently successful, continuing his extensive fieldwork, and involvement with the Geological Society and the British Association. Lyell served as President of the Geological Society in 1835–37, and again in 1849–51. Subsequently, he was knighted (1848), was awarded the Royal Society's Copley Medal in 1858, and served as President of the BA in 1864.

Lyell was seriously concerned with French geology. He acknowledged Cuvier's mastery of palaeontology, but rejected his 'catastrophist' theory. Lyell's *Principles* did much to counter this doctrine in contemporary Britain. In Paris in 1828, he met with the conchologist and palaeontologist Gérard Deshayes, who assisted him in the identification and stratigraphic placement of the shells he collected that year. Lyell reacted negatively to the tectonic theory of Léonce Élie de Beaumont (which envisaged mountain ranges as having been formed as a result of the

Earth's cooling and contraction) and significantly hindered its acceptance in Britain. Most importantly, Lyell gave close attention to the 'transformist' (evolutionary) theory of Jean Baptiste Lamarck in Volume II of *Principles*. Changing conditions cause new needs for organisms. To adjust to changing circumstances, organisms may alter their habits, and consequently their forms. These changes may be transmitted to subsequent generations, producing a gradual transformation of species. The first simple forms of life appeared naturally (without divine action) by spontaneous generation.

Such ideas were rejected by Lyell over many pages. His principal objection was that the stratigraphic record did not reveal smooth transitions such as Lamarck's theory would lead one to expect to find. But there were other objections, such as the inability to produce new species by breeding; and hybrids were sterile. Nevertheless, Lyell devoted much energy to thinking about what the concept of species meant, the 'laws' of distribution of species, and the extent to which they could or could not show modification due to different or changing circumstances. The problem of species and speciation was one of the main features of his book, and it set the scene for Darwin's work, and his seeing his fundamental problem to be 'the origin of species' (*see* **Famous Geologists:** Darwin).

A major problem for geologists in the first half of the nineteenth century was the large quantities of superficial deposits: gravel, tenacious clay containing unsorted rock fragments and fossil remains, and large boulders of rock distant from the nearest 'solid' outcrops of rock of that type. Such phenomena were eventually explained by the work of Louis Agassiz and his theory of an Ice Age (*see* **Famous Geologists:** Agassiz). In the early nineteenth century, these deposits were ascribed to the Noachian deluge or some like catastrophe, and William Buckland distinguished between 'diluvium' (Flood deposits) and 'alluvium' (materials deposited by rivers in the normal course of events). It was supposed that a great inundation(s) could have swept over the globe, even depositing the erratic boulders and marine shells loose at the tops of hills or mountains.

But according to Agassiz, the better explanation was that there had formerly been a colder climate with the whole of northern Europe once covered by ice, which had transported boulders, ground up the underlying rock, and deposited it, along with river gravels, over the land. The ice also could have scratched the underlying rock and transported shells to hill tops. Agassiz lectured on this to the BA in 1840, and some geologists were converted to his ideas, including Lyell. His general theory was presumed to be capable of accounting for a period of extreme cold, such as to cause widespread glaciation, if much of the high land at that time happened to be in the polar regions. But Lyell's conversion was short lived. Agassiz's theory seemed to take him too far from present analogies or present climatic conditions. So he adopted the theory that came to be called 'glacial submergence': there was a period of great cold, but not such as to produce an all-enveloping mass of land-ice. Rather, there was a general fall of land surface, causing marine submergence, accompanied by cooling, causing extension of ice-fields and the transport of boulders by drifting ice-bergs (hence the diluvial deposits are now generally called 'drift'). During his North American trip of 1845 Lyell saw floating ice in the St Lawrence River, which modern observation seemed to account for the occurrence of erratics satisfactorily in accordance with his methodology. While Lyell later accepted Agassiz's theory for the Alpine regions he never accepted the general land-ice theory, preferring the glacial submergence model.

After Darwin returned from his *Beagle* voyage in 1836, he and Lyell became close friends, but during the years before the publication of *The Origin of Species* Darwin mostly kept his emerging transformist ideas to himself. Lyell was opposed to transformism for reasons that he developed back in the 1820s, and like many he was concerned about evolution's implications for 'revealed religion' and social stability. In his Presidential Address to the Geological Society (1851), he spoke against evolutionary ideas. Man, he thought, was a very recent creation, subsequent to the mammoths. However, after Darwin revealed his ideas to Lyell about 1856, he was reluctantly converted and did his best to see the early publication of Darwin's ideas in 1858. In his *The Antiquity of Man* (1863), Lyell set forth ideas on transformism and stated his acceptance of the Darwinian theory of evolution by natural selection (though he represented it as a 'modification' of Lamarck's doctrine).

In the 1850s, Lyell had devoted a considerable amount of travel and fieldwork to the study of ancient humans, which was consistent with his general interest in the younger parts of the stratigraphic column. By that time, considerable numbers of cave deposits and flint implements had been discovered, as well as some human-like remains, notably the Neanderthal skull, found near Dusseldorf in 1857. This seemed, according to Thomas Henry Huxley's description, which Lyell quoted, to be intermediate between that of a modern human and a chimpanzee's. But Lyell cautiously (and rightly) stated that "it is at present to too exceptional, and its age too uncertain, to warrant us in relying on its abnormal and ape-like characters, as bearing on the question whether the farther back we trace Man into the past, the more

we shall find him approach in bodily conformation to those species of the anthropoid quadrumana which are most akin to him in structure" (*Antiquity*, p. 375). Lyell also reported on the recently discovered *Archaeopteryx*, which might seem to be a missing link, but he also deferred to the anatomist Richard Owen's opinion that it was actually a bird. Thus, Lyell supported Darwin's evolutionism in a way that was valuable to its acceptance. But at the same time he did not push all the evidence to what we might regard as its logical conclusion.

On reading Lyell's works, one is struck by his mastery of exposition and his command of the literature, especially in stratigraphy. His influence was very great, both in his own day and subsequently. There is, however, ambiguity in the concept of 'uniformitarianism' (gradualism, steady-statism, naturalism, and 'actualism' – or the idea that modern, actually observable, processes should be used to provide geological explanations). Lyell held to all these positions. Modern geologists commonly make obeisance to uniformitarianism, without making the foregoing distinctions. Modern geology does not necessarily adhere to any of them, except in its rhetoric; for Lyell convinced people that *his* approach was the right one to adopt for geology to be regarded as a science.

See Also

Famous Geologists: Agassiz; Cuvier; Darwin; Hutton; Smith. **History of Geology From 1780 To 1835**. **History of Geology From 1835 To 1900**.

Further Reading

Gould SJ (1987) Charles Lyell, historian of time's cycle. In: Gould SJ (ed.) *Time's Arrow Time's Cycle: Myth and Metaphor in the Discovery of Geological Time,* pp. 99 179. Cambridge (Mass) and London: Harvard University Press.

Hooykaas R (1963) *Natural Law and Divine Miracle: The Principle of Uniformity in Geology, Biology and Theology.* Leiden: EJ Brill.

Lyell C (1997) *Principles of Geology* edited with an intro duction by James A. Secord. London, New York, Ring wood, Toronto and Auckland: Penguin Books.

British Society for the History of Science (1976) *The British Journal for the History of Science: Lyell Centenary Issue* 9(2).

Rudwick MJS (1969) Lyell on Etna, and the antiquity of the Earth. In: Schneer CJ (ed.) *Toward a History of Geology,* pp. 288 304. Cambridge (Mass) and London: The M.I.T. Press.

Rudwick MJS (1971) Uniformity and progression: reflec tions on the structure of geological theory in the age of Lyell. In: Roller DHD (ed.) *Perspectives in the History of Science and Technology,* pp. 209 227. Norman: University of Oklahoma Press.

Rudwick MJS (1978) Charles Lyell's dream of a statistical palaeontology. *Palaeontology* 21: 225 244.

Rudwick MJS (1990) *"Introduction," Principles of Geology, First Edition* [in Facsimile] *Volume I Charles Lyell*, pp. vii lviii. Chicago: University of Chicago Press.

Wilson LG (1972) *Charles Lyell: The Years to 1841*. New Haven: Yale University Press.

Wilson LG (1998) *Lyell in America Transatlantic Geology, 1841 1853*. Baltimore and London: The Johns Hopkins University Press.

Murchison

D R Oldroyd, University of New South Wales, Sydney, Australia

Roderick Murchison (Figure 1) was the eldest son of a wealthy Scottish landowner at Tarradale estate, Ross-shire. Though born in Scotland, and always emphasizing his Scottish ancestry, he spent most of his career in England and spoke with an English accent. Following his father's death and mother's remarriage, Roderick was sent to school at Durham, aged 7 years, soon forming the ambition to be a soldier. At 13 years old, he attended the military college at Great Marlow where his training gave him a good 'eye for country'. He was soon involved in the 'Peninsula War' in Portugal, fighting at the Battle of Vimieira, aged only 16 years. From this victory, his unit moved into Spain where things went badly for the British army in winter conditions, with forced marches, defeat in the Battle of Coruna, and withdrawal in disarray. After a spell in Sicily, Murchison was posted to Ireland, where he led a dissolute and expensive life, and later likewise in London as a half-pay captain. With the end of the wars, he was fortunate to meet a general's daughter, Charlotte Hugonin, 3 years his senior, and they were married in 1816. They then took a leisurely tour through France, Switzerland, and Italy, and under his wife's influence his self-education began, learning French and Italian, visiting museums and galleries and some scientists and scientific institutions.

Figure 1 Roderick Murchison (1792 1871).

On their return, the Tarradale property was sold and the couple (who remained childless) resided for some years in Barnard Castle, County Durham, where Murchison devoted himself to the sporting country life. However, the two also participated in local gatherings of literary and scientific people, and met Sir Humphry Davy. Charlotte had interests in botany and mineralogy, but her husband's fox-hunting passion continued and they moved to Melton Mowbray in the hunting shires. However, urged by his wife and Davy, Murchison determined to take up geology, moving to London in 1824 and attending chemistry lectures. He joined the Geological Society, went on field trips with Buckland and others, and began serious study of the science.

Work in southern England yielded Murchison's first paper in 1825 (in which year he was elected to the Geological Society; and in 1826 to the Royal Society!), he journeyed to Scotland in 1826 and 1827, the former trip being made to Jurassic strata at Brora, following instruction on Secondary stratigraphy from William Smith (*see* **Famous Geologists:** Smith), whom he visited in Scarborough; the latter trip being with Adam Sedgwick (*see* **Famous Geologists:** Sedgwick). Murchison also visited the Continent with Lyell (*see* **Famous Geologists:** Lyell). Having gained some knowledge of how to study older rocks with Sedgwick, from 1831 Murchison and his friend and mentor determined to study the then rather little-known rocks of Wales; the so-called 'Transition Series' of Werner's Neptunist geology. He began to work from the known base of the Old Red Sandstone in the Welsh Border country, from the Wye Valley through to Cheshire.

Historians have rather detailed knowledge of Murchison's travels and scientific work as his field notebooks have been preserved, together with much correspondence. There is also a multi-volume 'journal', based on the notebooks, copied out by an amanuensis in Murchison's old age, which was intended for biographical purposes, sometimes being judiciously 'improved' by Murchison to give a favourable view of his accomplishments and ideas.

It is known, then, that in 1831 Murchison and Sedgwick planned to work out the geology of the Welsh (and Border region) Transition Series but, Sedgwick being otherwise occupied, Murchison set out alone that year, and Sedgwick arrived later, starting in North Wales, briefly with Darwin as an assistant (*see* **Famous Geologists:** Darwin). Near Ludlow, Murchison found richly fossiliferous rocks and the structure was made out successfully. He worked in his area for the next few seasons, subdividing the 'Upper Grauwacke Series' into the 'Ludlow Series', the 'Wenlock Limestone', the 'Horderley and May Hill Rocks' (later called the Caradoc Series), and the 'Builth and Llandeilo Flags', the first two being placed in the Upper Silurian and the latter two in the Lower Silurian (1835). The name Silurian was coined for a new geological system after the Silures tribe that formerly inhabited that part of Britain. The arrangement Murchison envisaged is shown in Figure 2, reproduced from his great treatise, *The Silurian System* (1839, p. 196). This *magnum opus* provided immense detail concerning the different units, figures of their characteristic fossils, and a valuable map of the geology of his Silurian 'domain'.

In 1835, Sedgwick and Murchison introduced the terms Cambrian and Silurian, though the word 'System' was not used at that time. Moreover, the boundary between the two was not then clearly defined, though following their fieldwork of 1834 Murchison stated that the upper and lower rocks had been 'dovetailed' in a manner that was satisfactory to both geologists. Unfortunately, this proved later not to have been the case and in the years that followed a serious controversy developed between the two former friends (*see* **Famous Geologists:** Sedgwick, **Palaeozoic:** Ordovician).

Murchison's approach to geology was considerably shaped by his military background. He came to regard 'his' Silurian System as personal territory, and the more parts of a map that could receive

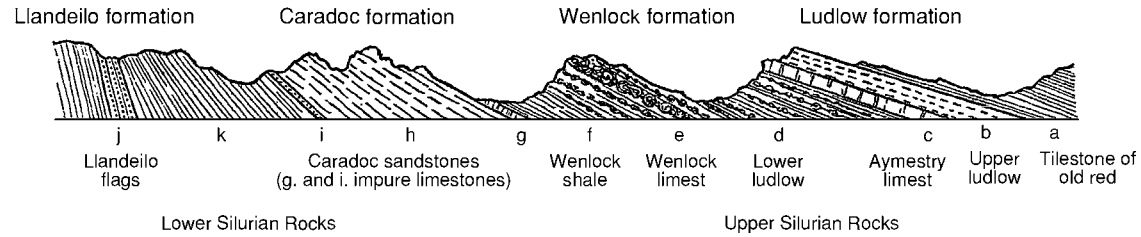


Figure 2 Arrangement of Silurian Strata, according to Murchison (1839).

Silurian colours the greater was his satisfaction! Indeed, he became known as the 'King of Siluria'. From the geological point of view it was evidently necessary to see whether the Silurian System, established in the Welsh Border area, was of general application, preferably worldwide. This grand task of spreading Silurian colours was undertaken personally by Murchison, in western and eastern Europe, the vast tracts of western Russia, and in Scotland in his old age. Other geologists were encouraged to find Silurian formations elsewhere, where Murchison's own feet had not trodden.

This expansion of Siluria also occurred in time as well as space, as Murchison sought to extend the Silurian down into the regions of Wales, where Sedgwick had established the Cambrian. So, when Murchison produced a condensed version of his stratigraphy in *Siluria* (1854), a name that suggested a kind of 'kingdom', it pushed down into Sedgwick's territory calling his rather unfossiliferous rocks of North Wales 'Lower Silurian'. The book was subtitled *The History of the Oldest Known Rocks Containing Organic Remains*. That is, Murchison claimed all the Palaeozoic rocks below the Devonian as belonging to his system.

This state of affairs arose in part because, after Sedgwick had put forward his Cambrian System, he did not there and then describe its characteristic fossils, and when this was eventually done some of his types had already been classified as Lower Silurian by Murchison. (Early on, Murchison divided his system into Lower and Upper divisions.) In *The Silurian System*, Murchison allowed the existence of fossil-containing Cambrian rocks below the Silurian, but it was supposed that they were linked by 'passage beds' to the Llandeilo. So the possibility of territorial and temporal expansion was already there. It occurred again in Russia (see below). In *Siluria* it was stated that the fossiliferous Cambrians were lateral extensions (undulations) of Silurian strata (an opinion said to have been reached in 1841), and all that was left on the map of Sedgwick's Cambrian was the apparently unfossiliferous rocks of the Harlech Dome area, some on the southern side of the Menai Strait between Anglesey and the rest of North Wales, and the unfossiliferous Longmynd rocks near Church Stretton in Shropshire. The remainder of Sedgwick's 'Cambria' was now depicted in Silurian colours. It was this encroachment, which had been going on through the 1840s, that so incensed Sedgwick (*see* **Famous Geologists:** Sedgwick). Murchison did well out of it all. He was awarded the Royal Society's Copley Medal in 1849, having been knighted in 1846.

But Murchison did not gain his honours lightly. After visiting German colleagues, he was in Russia in 1840 and 1841; Poland in 1843; Germany and Russia in 1844 and 1845; Scandinavia in 1844 and 1845; France and Germany in 1839 and 1843, and again in 1846 and 1847; and in Italy and Switzerland in 1847 and 1848. There were also journeys in Britain. In addition, Murchison served as General Secretary of the British Association, President of the Geological Society in 1841–1843 and of the Royal Geographical Society in 1843–1844. He also maintained an expensive but hospitable life style in Belgravia, London. Murchison's energy was truly remarkable.

Murchison's journeys in Russia, conducted with the French stratigrapher Philippe Edouard de Verneuil and the Russian zoologist Count Alexander Keyserling, were, for their time, of epic proportions. From St Petersburg in 1840, they travelled up to Archangel, then by indirect route to Moscow, and back to St Petersburg. Though much of the terrain is covered by drift and offers few good sections, information about Carboniferous, Devonian, and Silurian strata were obtained, partly from informants with knowledge of wells, etc. The area under St Petersburg is remarkable for having Cambrian *clays*, which by their fossils were construed by Murchison as Lower Silurian. In *The Geology of Russia* (see below) he stated that the clay contained fossils "belonging to the very oldest known Silurian or protozoic type . . . [and was] the true base of the Palaeozoic series, as indicated by a gradual dwindling out of animal life in the deposit of a region, where no eruptions ha[d] taken place, and where the strata are wholly unaltered". A further notable discovery, agreeable to the existence of the Devonian as a System, was the discovery of Old Red Sandstone fish in rocks that otherwise resembled the Magnesian

Limestone of England or the Zechstein of Thuringia, but were unlike the Old Red Sandstone rocks of Scotland. This confirmed the idea of the Devonian as a palaeontologically characterized system, which had different lithologies in different localities (*see* **Palaeozoic:** Devonian).

Early in his journey of 1841, Murchison met Czar Nicholas I in St Petersburg and established a good rapport with him. He was 'duchessed' by the Russian aristocracy, and formed a high opinion of it and of Russia. From Moscow, the geologists travelled east to Perm, to the west of the Urals, and thence further east to the point where they could see the plains of Siberia. Turning south on the western side of the mountains, they reached Orsk, and then headed west again, crossing the Volga and reaching the Sea of Azov, before going north again to Moscow and St Petersburg. It was on the basis of the rocks in the region of Perm that, later that year, Murchison proposed a new system, the Permian, after the ancient kingdom of Permia. Silurian, Devonian, and Carboniferous rocks were confirmed. The travellers saw considerable mining activity in the Urals. A central granitic nucleus appeared to be flanked by Silurian, Devonian, and Carboniferous strata. The Urals were eventually crossed and re-crossed on seven parallels between 60° and 54° N. As he had done for the Silurian and Devonian, Murchison named the Permian after a region with strata containing characteristic fossils.

On returning to St Petersburg, Murchison presented the Czar with a geological map, reports on the coal deposits of the Donetz Basin, and information about the alluvial gold of the Urals. It was a highly successful 'campaign' and Murchison's self-esteem rose to new heights. It was further fortified on his return to Britain by finding that the officers of the Geological Survey were obtaining results in Wales that seemed to support Murchison's views about the Silurian vis-à-vis the Cambrian.

In 1843, Murchison returned to eastern Europe, where he compared the Tatra Mountains of Poland with the Urals, and met von Humboldt in Berlin and the notable Palaeozoic palaeontologist Joachim Barrande in Prague. The same year Keyserling pursued the Russian researches in the northern Urals. In 1844, Murchison paid visits to Denmark, Norway, Sweden and Russia, and was able to find the basement rocks in Scandinavia underlying the Palaeozoics. In 1845, he was again in the north, receiving from Czar Nicholas the award of the Great Cross of St Stanislaus, for his notable contributions to Russian geology, and geologizing in the Baltic region. *The Geology of Russia in Europe and the Ural Mountains*, published in 1845 with Keyserling and Verneuil as co-authors and incorporating information from many others, was a monumental work, providing accounts of the geologists' journeys and descriptions of the geology of the half-continent. There were copious illustrations, including many sections and two large coloured geological maps; also descriptions and beautiful figures of fossils. Additionally, the book contained considerable theoretical discussion. It marked Murchison as the master stratigrapher and geological traveller of his age.

Besides becoming a leading geologist, Murchison was also active in encouraging geographical exploration, which he saw as essential to the expansion of the British Empire. As a man of influence, he interested himself in the geological appointments being made in the colonies, and his powers of patronage were considerable. For example, he assisted in the placement of Frederick McCoy in a chair at Melbourne University, which may have seemed advantageous to Murchison as it got one of Sedgwick's main allies out of the country! By 'placing' or maintaining contacts with people in different parts of the world, Murchison also gained advantage by having information channeled through his hands. He was gratified to have numerous topographical features named in his honour.

Murchison's influence became so great that he even felt qualified to offer opinions about the geology of countries that he had never visited. For example, he thought it likely that gold might be found in eastern Australia, adjacent to the hills of the 'Australian Cordillera' (mostly a rather inconspicuous topographic feature), on the basis of examination of some non-auriferous rock specimens shown him by the Polish explorer Paul Strzelecki, and on the grounds that the range contained rocks somewhat like those observed in the Urals (where alluvial gold had been found on the eastern flanks) and was similarly aligned, approximately N–S. In the event, this 'prediction' (1844) proved correct and was followed by gold rushes in the 1850s. Murchison regarded his lucky forecast as evidence that he was a "sort of authority" on Australian gold deposits. He recommended (1846) migration to Cornish tin miners, some of whom benefited from his fortuitously useful advice.

Murchison thought that the Russian gold was emplaced by quite recent tectonic activity in the Ural region, and he suggested that the range had undergone several distinct upheavals. Nevertheless, he gave credence to the theory of Élie de Beaumont that mountain ranges with different alignments were of different ages. Thus, the old Palaeozoics of Scandinavia were aligned SW–NE; the supposedly post-Carboniferous/Permian Urals ran N–S; while the post-'oolitic/chalk' of the Caucasus ran WNW–ESE.

The position of Director-General of the Geological Survey of Great Britain fell vacant in 1855, upon the decease of its founder Sir Henry De la Beche, and Murchison was appointed in his place, though already in his sixties. It proved to be an astute move, so far as the Survey's progress was concerned, for Murchison had innumerable contacts and used them to advantage to build up the organization considerably. He was indefatigably a man of organization, and competent, with his experience in the running of several societies and associations. The appointment was gratifying to Murchison as it ensured that the official maps should be constructed and coloured according to *his* interpretation and subdivisions of Palaeozoic geology, to the extent that the Cambrian was almost driven off the map for British geology.

Much of Murchison's geological work in his later years was focused on Scotland, where a separate branch of the Survey was established in 1867, though surveying had begun there back in 1854. Murchison's Scottish work involved him in the last of his three great controversies, and involved the attempted expansion of Silurian colours over the greater part of northern Scotland. There is today thought to be a great thrust-plane (the 'Moine Thrust') that runs from the north coast near Lochs Durness and Eriboll to the south-west, terminating in the southern part of Skye. To the west, one finds 'Fundamental [Lewisian] Gneiss' (so called by Murchison), overlain unconformably by the unfossiliferous Torridonian Sandstone. Lying unconformably on this there is a series of sediments, including a 'quartz rock' and the fossiliferous Durness Limestone. Over this lies the complex unit called the Moine Schists, above the thrust plane and extending eastwards until it is itself overlain unconformably by Old Red Sandstone on the eastern side of the country. But the structure near the thrust plane is complicated, with folding, inversions, and apparent duplication or repetitions of strata; in places the gneiss 'reappears', both near the thrust fault and again further east.

Murchison visited the north-west Highlands of Scotland in 1855 (with the Aberdeen University geologist James Nicol), in 1858 (with the local amateur naturalist Charles Peach), in 1859 (with the Survey officer Andrew Ramsay), and in 1860 (with the young surveyor Archibald Geikie). Fossils regarded by Murchison as Lower Silurian were found by Peach in the Durness Limestone (at a lower horizon than the thrust plane). All the strata appeared to dip gently to the southeast, with a strike approximately parallel with what is now thought to be the thrust-fault system. The outcome of all this work was that in the view of Murchison (and also Ramsay and Geikie) there was an essentially simple ascending sequence (with unconformities) from 'Fundamental Gneiss' on the west (regarded as lying at the bottom of the whole stratigraphic column for Britain) through to the Old Red Sandstone on the east, with a repetition of quartz rock into distinct upper and lower units, and also repetition of the gneiss. This meant that the Moine Schists, lying between the supposed Lower Silurian Durness Limestone and the Devonian Old Red Sandstone, though unfossiliferous, could be regarded as Silurian. So when a geological map of Scotland was published by Murchison and Geikie in 1861, large areas of northern Scotland were represented in Silurian colours. Murchison's empire was again expanding in a manner that he found most satisfactory. As to the Cambrian, Murchison allocated the unfossiliferous Torridonian Sandstone to that System, so Sedgwick saw some expansion his empire, but not by rocks with well-characterized fossils.

But Nicol's reading of the structure was very different from Murchison's. He came to the view that there was repetition of the western and eastern metamorphic rocks due to a large (high-angle) fault, and that the resultant fissure had been filled in part by some kind of igneous rock. (He was perhaps mistaking some gneiss for igneous rock.) If this interpretation were correct then placement of the Moine Schists in the Silurian would be suspect. So Nicol and Murchison fell out, and Nicol thereafter conducted his work separately from the Survey chief. The issues were debated at the British Association meeting in Aberdeen in 1859, where both geologists put forward their cases. From his stronger social position, Murchison was judged the winner by most geologists, and in fact Nicol's idea was by no means wholly correct.

The results of this encounter were most satisfactory to Murchison and Geikie, who became his mentor's advocate and eventually his sympathetic biographer. Murchison got more Silurian colour onto the geological map of Britain. In time, Geikie was appointed head of the Scottish branch of the Survey, and, when Murchison endowed a chair in geology at Edinburgh University (with Geikie's urging), it was Geikie who moved smoothly into the position, holding it concurrently with his post in the Survey. Later he became Director-General of the Survey, President of the Royal Society, and one of Britain's leading geologists.

However, the Murchison theory of the structure of the north-west Highlands was shown to be in error by Charles Callaway's and Charles Lapworth's mapwork in the early 1880s, and the reputation of the deceased Nicol was restored. Lapworth showed that the structure involved folding and thrust-faulting (a

term proposed by Geikie), and suggested that the Moine Schists were in fact formed by the earth movements that gave rise to the folding and faulting, while the repetitions of rock types could be attributed to the S-shaped folding. (This suggestion was eventually taken up by Geikie's own staff, surveying in the 1880s, well after Murchison's death.) The whole episode illustrates Murchison's dominating personality and commanding social role towards the end of his career.

The reasons underlying the Cambrian–Silurian debate have been analyzed by Rudwick (1976) in the following terms. At one level, it arose because Murchison's structural interpretations were not always correct and because he confused the May Hill Sandstone (Wenlock) with the lithologically similar Caradoc Sandstone (Caradoc). Both geologists gained ideas about how to do stratigraphy from William Smith (*see* **Famous Geologists:** Smith). Smith himself started from the observation of superposed sections of rocks of characteristic structure and lithologies. Subsequently, he remarked that each rock suite had its own characteristic fossils, but he saw no reason in principle why one set of fossils should not graduate into another. Thus there could, in principle, be overlap between Cambrian and Silurian fossils. Murchison started off on a similar basis, but gradually shifted towards thinking that it was fossils that *defined* a system. Once this had happened, and he began to find 'Silurian' fossils in Sedgwick's Cambrian, then annexation of territory 'naturally' followed (given that Sedgwick was so slow in getting his 'Cambrian' fossils published). It seemed to Murchison that he was dealing with a *bona fide* system, as it preceded land plants, had few vertebrates, and was apparently distributed widely round the world. By contrast, when Sedgwick got round to palaeontological analysis about a decade after his initial fieldwork in North Wales, he thought that the break should, if anywhere, lie between Murchison's Lower and Upper Silurian; so that for Sedgwick the Cambrian should incorporate Murchison's Lower Silurian. But by then the Lower Silurian was already well established, with its fossils described.

Rudwick further points out that both geologists were opposed to Lyell's 'steady-statism' (*see* **Famous Geologists:** Lyell). They both believed that life originated at some point in the past, and Murchison wished 'his' system to be the one that contained the first evidences of life with hard-bodied remains. Hence he sought to cannibalize Murchison's Cambrian. When Barrande in Bohemia found a 'Primordial' fauna below Murchison's Lower Silurian (palaeontologically defined), it could have served as palaeontological basis for a Cambrian System. But Murchison declined to follow this path, and did not practise what he preached in the matter of the Cambrian.

Like many geologists of his day, Murchison gave considerable attention to the problem of the superficial 'drift' deposits that blanket much of Europe, and which he saw in abundance in Scandinavia, Russia, Britain, and elsewhere. In the early nineteenth century, such materials were commonly ascribed to the Noachian Flood, or later to catastrophic floods but not necessarily universal or of divine origin. In the 1840s, there were two further contending theories: that of climatic change producing an Ice Age, with land ice as the agent for the emplacement of the 'drift', as advocated by Louis Agassiz (*see* **Famous Geologists:** Agassiz); and various versions of 'glacial submergence', with cooling and changes of sea-level relative to the land such that floating icebergs could carry detritus and deposit mud and 'erratic' boulders, as envisaged by Darwin and Lyell. The 'flood theory' received some theoretical support from the Cambridge mathematician and geologist, William Hopkins, who advocate the idea of 'waves of translation'. A sudden uplift of the sea-floor might, it was suggested, produce not only waves at the ocean surface, but also wholesale lateral movement of masses of water, capable of transporting ('translating') large boulders and finer debris.

It was Hopkins' theory that Murchison favoured, in part because it was seemingly in accord with the evidences familiar to him in the Alps and elsewhere of huge earth movements, foldings, faulting, and even inversions. (Murchison had seen evidence of stratigraphic inversion in the Glarus Canton, Switzerland, when he visited the area in 1848, but subsequently disregarded it in his thoughts about the north-west Highlands of Scotland.) He accepted that retreating glaciers left moraine material in the Alpine regions, and was happy with the idea of icebergs transporting drift material. But for long he could not accept land-ice as being responsible for the huge tracts of drift on land of low relief that he saw in Russia. Besides, the evidence of striations did not seem to accord with the land-ice theory. For example, in the area of the Gulf of Bothnia he saw scratch-marks directed southeastwards, from an area of Sweden of low altitude. He did not imagine that glaciers could have come from further north, from the mountains of Arctic Sweden. Nor could he imagine that land-ice could on occasions travel uphill, transporting marine shells to hill tops. It was only in 1862 that Murchison conceded to Agassiz's land-ice theory.

Murchison was one of the heroes of the heroic age of geology. His contributions to stratigraphy, and the broadening of geological knowledge generally, were immense. He was extraordinarily energetic, and generally amiable. Other than Lyell, he was far

the most influential British geologist of his day. But he exercised that influence through the hand of administrative power as much as by reasoned argument. He was a conservative in both politics and geological theory. Even his protégé Geikie, who owed him so much, concluded that Murchison "was not gifted with the philosophical spirit which evolves broad laws and principles in science. He had hardly any imaginative power. He wanted . . . the genius for dealing with questions of theory . . ." Possibly things would have been different if Murchison had received a university, rather than a military, education.

See Also

Famous Geologists: Agassiz; Darwin; Lyell; Sedgwick; Smith. **History of Geology From 1780 To 1835**. **History of Geology From 1835 To 1900**. **Palaeozoic:** Ordovician; Silurian; Devonian; Carboniferous.

Further Reading

Geikie A (1875) *Life of Sir Roderick I. Murchison. . . Based on his Journals and Letters With Notices of his Scientific Contemporaries and a Sketch of the Rise and Growth of Palaeozoic Geology.* London: John Murray (reprinted in facsimile by Gregg International Publishers Ltd, 1972).

Oldroyd DR (1990) *The Highlands Controversy: Constructing Geological Knowledge through Fieldwork in Nineteenth Century Britain.* Chicago and London: Chicago University Press.

Rudwick MJS (1972) Levels of Disagreement in the Sedg wick Murchison Controversy. *Journal of the Geological Society* 132: 373 375.

Rudwick MJS (1985) *The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists*. Chicago and London: Chicago University Press.

Secord JE (1986) *Controversy in Victorian Geology: The Cambrian Silurian Dispute*. Princeton: Princeton University Press.

Stafford RA (1989) *Scientist of Empire: Sir Roderick Murchison, Scientific Exploration and Victorian Imperialism.* Cambridge, New York, Port Chester, Melbourne and Sydney: Cambridge University Press.

Thackray JC (1976) The Murchison Sedgwick Controversy. *Journal of the Geological Society* 132: 367 372.

Sedgwick

D R Oldroyd, University of New South Wales, Sydney, Australia

Introduction

Adam Sedgwick (Figure 1) was born in 1785 in the village of Dent in the Yorkshire Dales, northern England, son of the local vicar and third of a family of seven. He died as a Fellow of Trinity College and Professor of Geology at Cambridge in 1873. He attended Sedbergh School near Dent, and with help from a notable local amateur mathematician, John Dawson of Garsdale, obtained a scholarship to Trinity, where he studied mathematics. He was named 5th Wrangler (fifth in the list of first-class honours students) in 1808. Following further examination, Sedgwick obtained a College Fellowship in 1810 and taught undergraduate mathematics. He was ordained in 1817. On becoming a permanent member of college, Sedgwick also committed himself to bachelorhood. During his life as a geologist, he proved to be extremely energetic in the field, covering large distances in a day. In Cambridge, he was quite often indisposed, but his chronic health problems apparently disappeared once he got into fieldwork.

Though a gifted mathematician, Sedgwick did not make a career in that discipline. From fragmentary autobiographical notes, he evidently had some geological interests from an early age, and he 'geologized' on the Continent in 1816. Also, he was 'introduced' to the Geological Society of London in 1818. Even so, it is surprising that his scientific accomplishments were thought sufficient to secure the Cambridge chair in geology that year. He was elected Fellow of the Royal Society in 1821, John Herschel heading the list of those who nominated him. Sedgwick was President of the Geological Society in 1829–31, and President of the British Association when it met in Cambridge in 1833.

Geological Work

On obtaining his chair, Sedgwick threw himself into geology. He started his annual fieldwork in southern England, then worked his way northwards to Northumberland, and in the years 1822–24 he made the first systematic survey of the Lake District. He obtained topographic maps of the region, 'recognized' certain rock units, and systematically covered

Figure 1 Portrait of Adam Sedgwick (1785 1873); the original is in the Sedgwick Museum, Cambridge. Copyright: The Sedgwick Museum of Earth Sciences.

the region over three seasons, colouring in his maps according to his selected lithological units. Sedgwick did not look much for fossils, but measured strata and cleavage dips and the alignments of beds, folds, faults, joints, and cleavage planes. Faults sometimes could be seen on the ground. At other times, they became apparent when the different rock types were coloured onto the maps. Sedgwick's labour and energy were immense. Armed with hammer, acid bottle, map, compass, clinometer, and notebooks, Sedgwick tried to determine the structure of that complicated region.

Trained in mathematics, the neophyte geologist was trying to ascertain whether the strata displayed any regular geometric pattern. They hardly did, but when Sedgwick published his work he sought to subsume it under the theory of Léonce Élie de Beaumont, according to which, as the Earth cooled and contracted, it supposedly formed a regular fold pattern in its crust, with mountain ranges of similar age having similar alignments. The theory never exerted much influence in Britain, and Sedgwick soon gave up the idea; but the fact that he sought to deploy the French theory suggests that he was interested in a geometrical (mathematical) theory of the earth. This was consistent with the Cambridge tradition, which found expression in the activities of the Cambridge Philosophical Society, which Sedgwick helped found in 1819. He wanted a quantitative geology, with mathematically formulated laws. However, his later Lakeland work (assisted by amateur collectors) used fossils, and by the end of his career, he had worked out a stratigraphic order for the sedimentary rocks compatible with that later developed on biostratigraphic principles. Sedgwick also recognized the rocks of central Lakeland for what they were: the relics of ancient volcanoes. He referred to water-deposited volcanic ash deposits as volcanic mud. From his Lakeland work, Sedgwick came to understand (and publish) the distinction between bedding and cleavage, but the distinction was acquired from the local amateur Jonathan Otley, who probably got it from quarrymen.

Sedgwick's Lakeland mapping revealed the existence of large tear-faults in some of the strata, and the eastern boundary of the region was marked by a huge normal fault. Hence, he suggested that the area had been affected by great earth movements. Using terminology proposed in the 1830s by his Trinity colleague, William Whewell, Sedgwick was a 'catastrophist'.

Sedgwick soon met Roderick Murchison (*see* **Famous Geologists:** Murchison), who wanted someone to show him how to make sense of ancient rocks in mountainous regions, and together they undertook a lengthy reconnaissance journey round the north coast of Scotland in 1827, unfortunately correlating the eastern and western sandstones of northern Scotland (now regarded as Devonian and Precambrian, respectively). In 1829, they made an extensive tour of the Continent, meeting European professors and travelling to Germany, Bohemia, Hungary, Austria, Switzerland, and Italy. This greatly extended Sedgwick's experience, especially in the "focus of Wernerian geology," southern Germany, which he found to be "the most decidedly volcanic secondary country I ever saw". He observed granite veins and inclined or even overturned Secondary rocks (a term used in the nineteenth century as a synonym for the German term *Floetz*; later for the strata ranging from Silurian to Cretaceous; and later restricted to the Mesozoic Era). This demolished his earlier adherence to Neptunism and he publicly repudiated the doctrine in 1831. Observations in Italy clearly suggested uplift, which was not part of the Wernerian repertoire. On the other hand, Sedgwick was inclined to ascribe the movement of (glacial) erratic boulders to the action of catastrophic floods, but some Swiss deposits could have been emplaced by the bursting of lake barriers. He rejected the idea that 'diluvium' was all deposited in the Noachian Flood, but was critical of Charles Lyell's belief that conditions on Earth were essentially similar through time (*see* **Famous Geologists:** Lyell).

In the 1830s, Sedgwick collaborated with Murchison in Wales, trying to bring order to the strata there. The strata in the mountains of Snowdonia seemed to have analogy with those of the lakes. Sedgwick tackled them, working on somewhat similar lines, and making a traverse north-west to south-east across Snowdonia in 1832. He unravelled the

structure to some extent, and although the rocks seemed to have few fossils, the Cambrian System was introduced, largely on the basis of Sedgwick's work. Murchison, starting in the Welsh border region in gentler country with fossiliferous rocks, had an easier time and revealed what appeared to be a new system with its characteristic fossils, dubbed the Silurian.

However, the line of boundary between the two systems was not established at the outset of the investigations in 1834, with the result that a bitter feud subsequently broke out between Sedgwick and Murchison, with the former trying to extend 'his' system upwards, the latter endeavouring to expand 'his' downwards, eventually to the very lowest fossiliferous rocks. (The issues were very complex; it was not until after the protagonists' deaths that the issues were resolved, by Charles Lapworth, who, in 1879, proposed the Ordovician System to occupy the disputed territory between the Cambrian and Silurian. (*see* **Palaeozoic:** Ordovician) Sedgwick had complained, with reason, that materials he had sent to the Geological Society were changed so as to accord with Murchison's views, without Sedgwick's knowledge or consent. The situation grew so bad between the two that after 1853, the Geological Society declined to accept further papers by the protagonists of Siluria or Cambria. Sedgwick felt grievously ill treated and snubbed by the Society. The battle became transferred to the forum of the British Association, but after 1854, Sedgwick withdrew from that body also, so far as the Cambrian and Silurian were concerned, and continued the battle from Cambridge and in the pages of the *Philosophical Magazine*. Some of the stratigraphic formations and their classifications are shown in Figure 2. It should be noted that the rocks that Murchison allowed to be Cambrian in 1859 were the unfossiliferous Longmynd rocks, later classified as Precambrian. (Murchison also allowed Sedgwick the unfossiliferous Torridonian Sandstone in Scotland, also now regarded as Precambrian.)

Initially, Murchison won the battle, partly because he had better fossil evidence, and placed full reliance on it. Also, he had strong influence in the Geological Society and became Director of the Geological Survey in 1855, and his classifications were used by the survey officers. Sedgwick had fewer allies, mostly at Cambridge. For the rocks he was dealing with, he had to rely on structural understanding and lithologically based mapping to a greater extent than did Murchison. (Graptolites were not regularly used for stratigraphic correlation in the mid-nineteenth century.) However, Sedgwick succeeded in showing, on palaeontological grounds, that the claimed unity of Murchison's Silurian System was flawed. In 1852, Sedgwick and his assistant Frederick McCoy found that one of Murchison's Silurian formations, the Caradoc, had rocks containing two distinct faunas, as shown by the palaeontological determinations of McCoy and John Salter. There had been erroneous correlations; the same term, 'Caradoc Sandstone', had been applied to different series of rocks; and there should be an unconformity within the Caradoc, as Murchison then understood it. Sedgwick proposed the division of Murchison's Caradoc into the Caradoc Sandstone, containing fossils such as *Trinucleus*, and an upper May Hill Sandstone, containing *Pentamerus* species. This eventually turned the tide against his Silurian being regarded as a coherent system. The Survey sought to retrieve the situation by adopting the terminology 'May Hill Sandstone Llandovery' rocks, regarding them as a kind of passage or 'Intermediate Series' between the Upper and Lower Silurian. There were repercussions, too, for the interpretation of Murchison's Llandeilo.

This battle was fought with extreme vehemence. Both geologists attached their names and reputations to 'their' system. The battle seemed to exemplify the height of the colonial era, with Sedgwick and Murchison trying to extend their empires. Murchison was popularly called the 'King of Siluria'. Both men tried to rewrite history in their historical accounts of the events. In Sedgwick's case, this may have been partly due to failing memory. Earlier, there had been a bitter controversy in Devonshire, where, while still friends, Sedgwick and Murchison began to unravel the structure and stratigraphy of the area, in the process becoming involved in controversy with Henry De La Beche, the first Director of the Geological Survey. From this acrimonious debate emerged the concept of the Devonian System.

The so-called Old Red Sandstone was the unit well known in eastern Scotland as lying unconformably over Silurian strata. It was mapped by William Smith as 'Red and Dunstone'. He placed it below the Coal Measures and below a limestone that cropped out in Derbyshire and elsewhere. In Devonshire, there occurred rocks with plant remains that appeared to De la Beche to belong to the old Transition/Greywacke series (Cambrian or Silurian rocks). In Murchison's opinion, however, the plant-bearing rocks could not be so old: they must be from the Coal Measures. But he had not then been to Devon to see the rocks for himself. Murchison then combined forces with Sedgwick to combat De la Beche's interpretation. It was an issue of more than academic significance because it bore on the question of the possible extent of coal-bearing rocks.

The plant-bearing rocks overlay contorted rocks of ancient appearance, but these contained corals

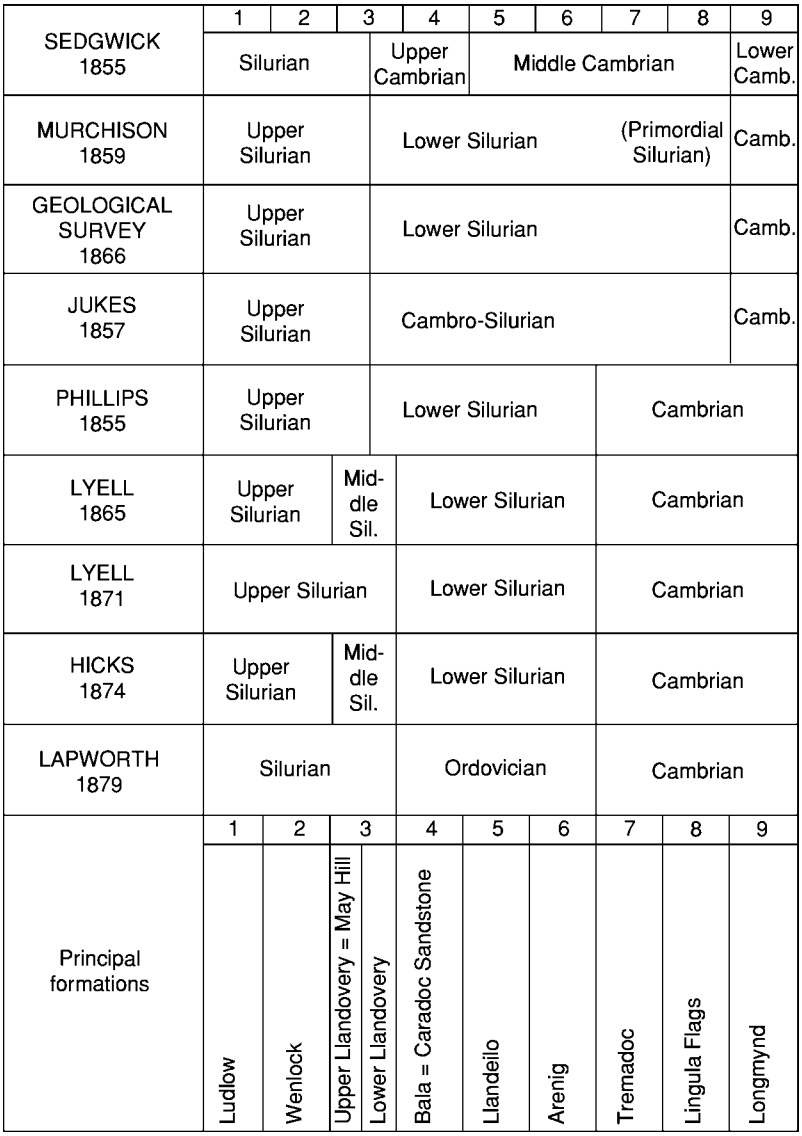


Figure 2 Classification of British Lower Palaeozoic Rocks. Reproduced from Secord JA (1986) *Controversy in Victorian Geology: The Cambrian Siurian Dispute*, p. 287. Copyright ©1986 by P.U. Press. Reprinted by permission of Princeton University Press.

different from those in the Silurian. In the opinion of the coral expert William Lonsdale, the corals were intermediate between Silurian and Coal Measure types. Thus, the Devon rocks might be situated between the Silurian and the Carboniferous. So, in 1839, Sedgwick and Murchison proposed the Devonian System, being one that had different facies in different localities. Here the palaeontological evidence of corals was taken to outweigh the uncertain stratigraphic reliability of plant fossils and the structural arguments advanced by De la Beche. The issues were debated with considerable rancour and again illustrate the character of stratigraphic controversies in the nineteenth century. But this time, Sedgwick and Murchison were on the same side and the debate never got quite so out of hand as did the Silurian/Cambrian contest.

Sedgwick as a Teacher; Other Activities, Beliefs, and Character

At Cambridge, Sedgwick gave an annual course of lectures and built up the university's geological collections, partly from his own collected specimens, but also by donations and purchases. His summer fieldwork was done at his own expense. Partly for this reason, he took a 'second job' in 1834, as a canon at Norwich Cathedral. This might have been a sinecure, but Sedgwick took his responsibilities seriously, and resided in Norwich for several months each year, also

encouraging the development of a museum in the city and giving geological lectures. Sedgwick was a strong supporter of amateur science and assisted the natural history society in Kendal, near Dent.

Sedgwick was a renowned orator – or preacher and lecturer. Science lectures were not a required part of the Cambridge curriculum when he gave his first course in 1819, but he attracted many students and dons to his lectures. His course was repeated until 1859, when he was 74. He spoke extempore about geological principles and his recent fieldwork, rather than about unnecessary minutiae. His lectures, utilizing specimens and maps and diagrams to explain his ideas, were popular, and he also led groups on horseback on field excursions round Cambridge. On a famous occasion at the British Association meeting at Newcastle in 1838, he spoke in the morning at Tynemouth beach to a group attending the meeting; and by the afternoon he had attracted a crowd of thousands, expounding the relationships (as he saw them) between geology, political economy, natural theology, and patriotism, reportedly drawing tears of emotion from some auditors. As John Herschel described it, Sedgwick

> led them on from the scene around them to the wonders of the coal country below them, thence to the economy of a coal field, then to their relations to the coal owners and capitalists, then to the great principles of morality and happiness, and last to their relation to God and their own future prospects.
> (Clarke and Hughes (1890).)

In this can be seen the relationship between Sedgwick's science, social, religious, and political philosophy. Implausibly, he supposed that Britain had been specially favoured by God for its place in the world, with its deposits of coal, limestone, and iron ore.

Sedgwick was, then, devoutly religious, and a preacher as well as a teacher. From a relatively obscure Anglican background, he rose to be a Cambridge Professor and Vice-Master of Trinity, and one of the leaders of the heroic age of geology. He even met with Prince Albert (whose election to the Chancellorship of the University he promoted) to discuss reforms at Cambridge, and many of Sedgwick's suggestions were implemented. Though generally amiable, greatly liked and admired, and able to communicate with the full range of society, from quarrymen, to famous writers such as William Wordsworth or Walter Scott, to Royalty, he was uncompromising and dogmatic. He favoured Catholic emancipation, but having become an establishment figure, he did not wish to see the regular order of things upset by scientific theories that seemed to him subversive, or at odds with orthodox Anglican theology. For such reasons, he was bitterly and publicly opposed to the transmutationist ideas expressed in 1844 by Robert Chambers in his *Vestiges of the Natural History of Creation*, and was privately grieved by Charles Darwin's ideas in *The Origin of Species* (*see* **Famous Geologists:** Darwin). He also rejected the land–ice theory of Louis Agassiz (*see* **Famous Geologists:** Agassiz).

See Also

Famous Geologists: Agassiz; Darwin; Lyell; Murchison; Smith. **History of Geology From 1780 To 1835**. **History of Geology From 1835 To 1900**. **Palaeozoic:** Cambrian; Ordovician; Silurian.

Further Reading

Clark JW and Hughes TMcK (1890) *The Life and Letters of the Reverend Adam Sedgwick, LL.D., D.C.L., F.R.S., Fellow of Trinity College, Cambridge, Prebendary of Norwich, Woodwardian Professor of Geology, 1818 1873*. Cambridge: Cambridge University Press.

Oldroyd DR (2002) Adam Sedgwick: a confident mind in turmoil. In: Harman P and Mitton S (eds.) *Cambridge Scientific Minds,* pp. 64 78. Cambridge: Cambridge University Press.

Oldroyd DR (2002) *Earth, Water, Ice and Fire: Two Hun dred Years of Geological Research in the English Lake District*. London: The Geological Society.

Rudwick MJS (1972) Levels of disagreement in the Sedgwick Murchison controversy. *Journal of the Geological Society* 132: 373 375.

Rudwick MJS (1985) *The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentle manly Specialists*. Chicago and London: Chicago University Press.

Rudwick MJS (1988) A year in the life of Adam Sedgwick and company, geologists. *Archives of Natural History* 15: 243 268.

Secord JA (1986) *Controversy in Victorian Geology: The Cambrian Silurian Dispute*. Princeton and Guildford: Princeton University Press.

Sedgwick A and Murchison RI (1835/36) On the Silurian and Cambrian Systems, exhibiting the order in which the older sedimentary strata succeed each other in England and Wales. *Report of the Fifth Meeting of the British Association for the Advancement of Science held at Dublin in 1835,* pp. 59 61. London: John Murray.

Smith C (1985) Geology and mathematicians: the rise of physical geology. In: Harman PM (ed.) *Wranglers and Physicists: Studies on Cambridge Physics in the Nineteenth Century,* pp. 49 83. Manchester: Manchester University Press.

Speakman C (1969) *Adam Sedgwick Geologist and Dalesman, 1785 1873: A Biography in Twelve Themes*. Broad Oak, London, and Cambridge: The Broad Oak Press Ltd.

Smith

D R Oldroyd, University of New South Wales, Sydney, Australia

Introduction

William Smith (Figure 1) is renowned in the history of geology for differentiating and listing in order the English strata, from the Chalk down to the Coal Measures, and hence enunciating the stratigraphic principle that strata have a generally regular order of superposition and may be characterized by their fossil contents. The great geological map of much of Britain that Smith published in 1815 has also brought him much posthumous fame. As Smith put it in the memoir accompanying his map:

> [T]here is a great deal of regularity in the position and thickness of... strata; and although considerable dis locations are found in collieries and mines, and some vacancies [gaps] in the superficial courses of them, yet... the general order is preserved; and... each stratum is... possessed of properties peculiar to itself, has the same exterior characters and chemical qualities, and the same extraneous or organized fossils throughout its course. (Smith W (1815, p.2))

Background

Smith was born in 1769 in the village of Churchill, Oxfordshire, near Chipping Norton, an attractive part of Britain where Jurassic rocks (Oolitic Limestone and Lias) crop out well. His father, a blacksmith, died when he was only 7 years old, and he was then brought up by a farmer uncle at Over Norton, a few miles to the north-east, in similar type of country. Even at an early age, Smith was a keen collector of fossils. Not wishing for a life as a farm labourer, he began studying mathematics, geometry, and surveying techniques, and at age 18 became an assistant to a local surveyor.

Figure 1 William Smith (1769 1839) aged 69, engraved by TA Dean.

Smith soon became autonomous, and his work as a land surveyor evolved into the business of surveying for canal construction (initially the Somerset Coal Canal). This type of construction was then forging ahead in the days of the Industrial Revolution, and in effect, Smith became what would today be called a civil engineer. He advised on tunnel constructions, borings for coal, and mining activity in the Somerset coalfield. In 1794, he travelled to northern England on a 'fact-finding tour', in connection with canal work. He later advised on drainage projects for wealthy landowners who sought to develop their lands agriculturally, but these landowners were also interested in the mineral wealth that their estates might hold.

Development of Smith's Stratigraphic Principle

The varied experience Smith acquired, and especially that resulting from the canal cutting through different stratigraphic horizons, led, before 1796, to Smith recognizing the general aspects of his stratigraphic principle. He realized that the regular stratifications found within the coal mines could also be observed outside the mines. The Somerset Canal cuttings, which cut through two similar sections at two separate but neighbouring localities, revealed the lateral extents of strata, and Smith began to get the 'feel' of the internal structure of the earth as regards the strata of his region. In 1795, he took up residence in Bath and pondered what his work was beginning to reveal, namely, that "Nature has disposed of these singular productions [fossils] and assigned to each class its peculiar Stratum". In 1797, he wrote down a list of strata for the Bath district, listing 28 units, from Chalk down to Carboniferous Limestone, but without mentioning fossils (this is his earliest dated list that has survived). A revised version of this document (with fossils given) was dictated to two local clergymen/naturalists, the Reverends Benjamin Richardson and Joseph Townsend in 1799, who were also shown a circular map of the Bath district and one of

Somerset that Smith had geologically coloured. Not long before, Smith's employment with the Somerset Coal Canal Company had been terminated and it is possible that he was beginning to think of finding some practical and remunerated application of his new ideas. During the next few years, he was largely involved with drainage schemes, but by 1804, he was chiefly employed in 'mineral surveying'. He travelled great distances in these lines of work, and thus accumulated information that he later synthesized in the form of his celebrated geological map of 1815.

The revised stratigraphic table of 1799 had 23 stratigraphic units, from Chalk down to Coal, with thicknesses indicated, along with localities where the rocks cropped out. 'Fossils and Petrifactions' were given for each unit, but the fossil categories, as stated by him, were imprecise (corals, cochleae, ostreae, impressions of ferns, etc.). Also, some strata were characterized lithologically, or the entry was stated "No fossils known". So Smith's earliest table of strata was not based wholly on fossils, and appears to have been primarily a list of the lithological sequence of distinctly identifiable strata. At that time, he had no 'scientific' knowledge of fossils, but he was certainly collecting fossils well before 1799; his collections survive today at the British Museum. In any case, his differentiation of four blue clays and two different oolitic limestones indicates that he was making more use of fossils than is evident from the table dictated to Richardson and Townsend. They were impressed by the practical man's revelations and encouraged Smith to continue his investigations, which he did, producing a simple geological map of England and Wales in 1801. It showed quite a clear representation of the distribution of several major stratigraphic subdivisions, notably those now known as the Carboniferous, Jurassic, and Cretaceous.

With the loss of forests for shipbuilding and development of agriculture, and increasing demands for iron, there was need for coal in the years of the Napoleonic Wars, and various prospecting attempts were made in southern England, often on the illusory promise of the discovery of lignites or the occurrence of blue clays, thought by their appearance to be associated with coal measures. But Smith knew the correct order of strata by their fossils and realized that these attempts were doomed to failure. The prospectors were looking too high in the stratigraphic column. On the other hand, though Smith also gave sometimes successful advice as to where coal might be found in known coal areas, he was not always successful, due to unforeseen structural complexities, so that the coal beds present were unfortunately missed.

In fact, Smith's own entrepreneurial ventures were by no means successful. His intended book, *Accurate Delineations and Descriptions of the Natural Order of the Various Strata that are Found in the Different Parts of England and Wales, with Practical Observations Thereon*, for which a prospectus was printed in 1801, was never published because of the double bankruptcy of the intended publisher in 1801 and 1804. (The text of this prospectus was published in a 1942 article by LR Cox.) Worse, in 1798, Smith purchased a small estate at Tucking Mill near Bath, where quarrying of Bath stone was later attempted. This proved a failure because of the unexpectedly poor quality of the stone, and the financial problems that flowed from this led to Smith's subsequent financial collapse. (It is a sad irony that the experienced engineer and emerging geologist should have misjudged the stone quality and the difficulty of getting it out of the quarry.) On the other, hand, through agricultural contacts at Woburn, seat of the Duke of Bedford, Smith was introduced to the influential Sir Joseph Banks, President of the Royal Society, from whom Smith received both moral and financial support. Also, in 1801, Smith attracted, as a kind of 'pupil', the polymathic John Farey (likewise a practical man, but one of many accomplishments, including mathematics and music). They did fieldwork together and Farey became a constant advocate of Smith and his work. Smith also became acquainted with the map publisher and engraver John Cary, whose map of England and Wales and various county maps became the basis of Smith's great geological map (see later).

In 1802, Smith opened an office in Bath to conduct his affairs, and in 1804, he moved to London, where he displayed his fossils, arranging them on sloping shelves so that the fossils of each stratum were displayed in the order that they appeared in the English rocks, producing a kind of three-dimensional effect. But his efforts could only be spread thinly, and not necessarily systematically, because of the exigencies of his work. Although it is clear that Smith long intended to attempt to publish a geological map of England, Wales, and southern Scotland, this project was not in fact fulfilled until 1815, partly because of difficulties in reaching an agreement with a publisher and partly because he was continuing with his observations and collections. Smith's business took him all over the country, and he developed his topographic and geological knowledge wherever he went, to the point where he was almost overwhelmed with information and specimens.

The Geological Society of London was founded in 1807, with the wealthy George Bellas Greenough as its president. It might have seemed natural or appropriate for Smith to have joined the Society, but this did not happen. Smith was from a lower social class than were the Society's founders, and his patron,

Banks, was at odds with the new group, which he saw as a rival to 'his' Royal Society. Additionally, the Society's early Fellows were chiefly interested in mineralogical matters, and they and others of influence, such as the Board of Agriculture or Professor John Kidd at Oxford, doubted the value of Smith's 'biostratigraphy'. So, to an extent, Smith was on his own, and he had to carry through his project using his own uncertain financial resources.

Eventually, however, in 1815, Smith issued his great map and its accompanying memoir – *A Delineation of the Strata of England and Wales, with Part of Scotland* and *A Memoir to the Map and Delineation of the Strata of England and Wales, with Part of Scotland*. The geological information was entered on a map specially engraved for the purpose by Cary. There followed *Strata Identified by Organized Fossils, Containing Prints on Coloured Paper of the Most Characteristic Specimens in each Stratum*, which was issued in four parts between 1816 and 1819, providing descriptions of Smith's chosen stratigraphic units and beautiful coloured illustrations of their typical fossils, and *Stratigraphical System of Organized Fossils*, which was issued in 1817. (Neither publication was fully completed.)

The main map (dedicated to Banks) was issued on 16 sheets, one being an index sheet. (The Banks copy at the British Library may be the 'master' copy, but this is not certain.) On a scale of 5 miles to the inch, the map was approximately 260 by 180 cm in size. Each copy was hand coloured (apparently using colourists employed by Cary, not always to Smith's satisfaction) and there were five issues between 1815 and 1817 (or perhaps later). Examination of surviving copies has shown that Smith added information to the new issues as it became available to him. From 1819 to 1824, he issued also various 'county maps', which depicted the geology of individual counties. These lacked 'geological rationale', in the sense that counties were not 'natural' geological regions, but they were useful commercial products. Together they formed *Smith's Geological Atlas*. A smaller country map on the scale of 15 miles to the inch, with revisions of the earlier map, was issued in 1820, and this map was also variously revised until at least 1828. Additionally, Smith produced several geological sections, including one from London to Snowdon in North Wales, in 1817, and one of the strata south of London, in 1817. A manuscript section from 1824, from Flamborough Head on the east Yorkshire coast to Whitehaven on the Cumberland coast on the west of England, is preserved at Oxford.

The 1815 map was a mighty contribution to geology, achieved largely single-handedly. Surviving pristine copies are objects of great beauty, ingeniously and impressively coloured, so as to convey almost a three-dimensional effect, by increasing the intensity of colouration towards the lower boundaries of the outcrops of the various units. The map depicted 23 stratigraphic subdivisions, and some of the colours that Smith chose (e.g., green for Chalk) survive into modern maps.

But Smith's position was financially precarious, as it had been ever since his unsuccessful quarrying venture near Bath, and eventually, in 1817, he was obliged to sell his fossil collections to the British Museum for £700, the catalogue for this being his *Stratigraphical System of Organized Fossils*. The sale only postponed Smith's financial crisis, however, and he found himself languishing in a debtors' prison for 10 weeks in 1819. He obtained release by sale of his property near Bath, but withdrew to the north of England, making a living by continuing his survey work and giving lectures in Yorkshire's major towns. He had previously been assisted by his gifted nephew John Phillips, and their association continued. Phillips was perhaps chiefly responsible for the aforementioned east–west section of the north of England, which revealed an understanding of the faulted structure of the Vale of Eden, to the east of the Lake District. Phillips also did some of the lecturing. (Phillips subsequently became one of Britain's leading geologists, a Fellow of the Royal Society and Professor at Oxford.) While in the north, Smith put both Adam Sedgwick (*see* **Famous Geologists:** Sedgwick) and Roderick Murchison (*see* **Famous Geologists:** Murchison) on the right track as to the use of fossils for stratigraphic purposes. Smith eventually settled in Scarborough and assisted in the founding of the Scarborough Philosophical Society and Museum (which was made largely to Smith's design).

Back in 1808, Smith had been visited by Greenough and other leaders of the Geological Society, but they seemed unimpressed by his work, and subsequently started compiling a collaborative Society map, based on lithological principles such as were typically used by German geologists. However, at some point during the year after Smith was imprisoned, under the influence of the publication of Smith's map, the Geological Society group changed their approach and issued their own map, using some of Smith's fossil-based data, though Greenough asserted that the utility of fossils had been "greatly over-rated". Smith, at that point in time, had been walking to the north with his nephew Phillips as companion. Smith and his supporters, such as Farey, claimed his priority rights on several occasions, but these were not fully acknowledged until there was a change of personnel in the Geological Society. Eventually, in 1831, acknowledgement was accorded Smith by making him the

first recipient of the Society's Wollaston Medal (though there were objections made as to whether it was appropriate to make an award for work first announced in 1799). From the chair, the President Adam Sedgwick acknowledged his personal indebtedness to Smith's advice and dubbed him the 'Father of English Geology'. Smith's reputation was thus securely sealed, and the following year he received an annuity of £100 per year from the government. Moreover, his stratigraphic subdivisions set the pattern for work in other countries: the world followed British stratigraphy. If geology had emerged in the United States, China, or New Zealand, say, the stratigraphic column would look substantially different from that which is now used.

Influence of Smith's Work

The question of Smith's theoretical ideas in geology is important. His sections showed the strata of southern England in their correct order, where they conveniently form a 'layer-cake' stratigraphy. His main expertise was in the stratigraphy of these Mesozoic sediments (as they are now called). It is evident that strata of different lithologies were recognized first, and then Smith realized that each stratum had its characteristic fossils. Soon, he could reverse the argument and use fossils to identify the strata. Sometimes, however, he encountered problems. He thought, for example, that the poorly fossiliferous Magnesian Limestone of north Yorkshire (subsequently designated as Permian) and the Lias (now Jurassic) belonged to the same stratum, because they contained rather similar fossil fishes. Also, what is now thought of as Carboniferous Limestone and Lias were regarded on occasion by Smith as different facies of the same unit, there being no locality where the Carboniferous Limestone, Magnesian Limestone, and Lias appear in what is today regarded as the correct order, and some Carboniferous Limestone does occur in places reworked into Lias. This is not to blame Smith. He was pioneering, and mistakes were to be expected in those early days of biostratigraphy.

Throughout his career (started as a surveyor), Smith was always primarily interested in the geometrical arrangement of rocks, because this was what counted for agricultural, mining, and engineering purposes. His livelihood depended on knowing that order. Nevertheless, though it was not his primary concern, he did ask himself why the order was the way it was, how the strata came to be formed, and how long it took for them to be deposited. Smith's religious beliefs appear to have been conventional, or characteristic of his time, and involved use of the 'argument from design'. So one answer (1817) was simply that "[t]he interior of the earth…is formed upon the wisest and best principles", and that the inclinations of the strata evidenced design by making the different rock types available for human use. Fossils must "strike the admirers of nature with a degree of reverential awe and grateful admiration of the Almighty Creator". Earlier, in 1802, in a preface to a book that was never published, Smith had supported an older eighteenth-century idea that the inclinations of the strata were the result of Earth's rotation when the materials were still "in soft state or of pulpy consistence". But Smith apparently dropped this idea, which would imply that stratigraphic order did not represent chronological order. Even earlier, according to an 1844 memoir of Smith by J Phillips, Smith thought (in about 1795) that "each stratum had been successively the bed of the sea, and contained in it the mineralized monuments of the races of organic beings then in existence".

Another shred of evidence on this matter is provided by Farey, in a review he published in 1810 of Georges Cuvier and Alexandre Brongniart's 1808 memoir on the geology of the Paris area. Farey claimed that soon after Smith began his investigations, he "discovered an important law regulating all the *known alluvia*, or that which consisted of or contained the fragments and reliquia of known strata, [namely that they] were moved *from the south-east* towards the south-west" (italics in original) for material from any particular stratum seemed to have been transported beyond its western edge. This appears to have been a reference to observations of boulder clay, or 'drift', which material was ascribed by Smith (or Farey?) to "vast tidal currents which have swept over all the surface from SE. to NW., since or at the time, that the deposition of regular strata ceased". This suggests some support on Smith's part for the catastrophist doctrines espoused by Cuvier (*see* **Famous Geologists:** Cuvier). This could have accorded with Smith's religious views, but would also have involved the notion of time for the emplacement of superficial materials.

Additionally, HS Torrens has drawn attention to a Smith manuscript from about 1806; the manuscript indicates that Smith was then thinking of a vast extent of geological time: "the time required for the Perfection and Decay, and subsequent formation, into Strata which have evidently been formed in deep and quiet water". This time "would stagger the faith of Many". But Smith's lectures in Leeds in 1825 referred specifically to geological proofs of the occurrence of the deluge. He seems to have been impressed with William Buckland's recently claimed evidence for the occurrence of the Noachian Flood from cave excavations in Yorkshire. For further variety,

there are 40 proof sheets at Oxford of a work to be titled *Abstract Views of Geology*, which was apparently in press at the time of Smith's death. He was speculating again about the formation of strata "from a chemical conversion of liquids and gases into the solid state, —the layering being the effect of an uncombinable excess of one of the ingredients in the layer then formed, and the vertical joints in that layer the effect of solidification. . .." But he was not advancing geology through such suggestions. Continuing through his late lectures and this last work, he kept reverting to his great principle of identifying and ordering strata by their fossil contents, and the utility of knowledge of this kind. His speculations about time and process had little influence on the development of geology.

Smith's strata, as given in his table of 1817, with approximate modern equivalents, following JCM Fuller (1995), were as follows:

1. London Clay – Tertiary, Lower Eocene.
2. Sand – Tertiary, Lower Eocene (Woolwich and Reading Beds).
3. Crag – Tertiary, Pliocene (Shelly Sand).
4. Sand – Tertiary, Paleocene (Thanet Sand).
5. Chalk – Upper Cretaceous, Cenomanian to Senonian).
6. Greensand – Upper Cretaceous, Albian (Upper Greensand).
7. Brickearth – Upper Cretaceous, Albian (Gault Clay).
8. Sand – Lower Cretaceous, Aptan (Lower Greensand).
9. Portland Rock – Upper Jurassic, Portlandian–Purbeck.
10. Sand – Lower Cretaceous, Wealden (Ashdown Sand).
11. Oaktree Clay – Upper Jurassic (Kimmeridge Clay) and Lower Cretaceous (Wealden).
12. Coral Rag and Pisolite – Upper Jurassic, Corallian.
13. Sand – Upper Jurassic, Corallian (Lower Calcareous Grit).
14. Clunch Clay and Shale – Upper Jurassic, Oxfordian (Oxford Clay).
15. Kelloways Stone – Upper Jurassic, Callovian.
16. Cornbrash – Middle Jurassic, Bathonian and Upper Jurassic, Callovian.
17. Sand and Sandstone – Middle Jurassic (Hinton Sand).
18. Forest Marble – Middle Jurassic, Bathonian (and Wychwood Sandstone).
19. Clay over Upper Oolite – Middle Jurassic (Bradford Clay).
20. Upper Oolite – Middle Jurassic, Bathonian (Great Oolite Limestone).
21. Fuller's Earth and Rock – Middle Jurassic, Bathonian.
22. Under Oolite – Middle Jurassic, Bajocian (Inferior Oolite).
23. Sand – Lower Jurassic, Upper Lias (Midford Sand).
24. Marlstone – Lower Jurassic, Middle Lias, Domerian.
25. Blue Marl – Lower Jurassic (Lower Lias Clay).
26. Blue Lias – Lower Jurassic (Lower Lias).
27. White Lias – Rhaetic (Lower Lias).
28. Red Marl – Triassic (Keuper Marl).
29. Redland Limestone – Permian (Magnesian Limestone).
30. Coal Measures – Pennsylvanian.
31. Mountain Limestone – Mississippian.
32. Red Rhab and Dunstone – Devonian (Old Red Sandstone).
33. Killas – Devonian and older (slates, grits).
34. Granite, Syenite, Gneiss.

It is clear, then, that in its essentials, Smith's stratigraphic order still stands to this day.

See Also

Economic Geology. **Famous Geologists:** Cuvier; Murchison; Sedgwick. **Geological Maps and Their Interpretation**. **History of Geology From 1780 To 1835**. **Palaeontology**. **Stratigraphical Principles**.

Further Reading

Cox LR (1942) New light on William Smith and his work. *Proceedings of the Yorkshire Geological Society* 25: 1 99.

Eyles JM (1969) William Smith (1769 1839): a chronology of significant dates in his life. *Proceedings of the Geological Society of London* 1657: 173 176.

Fuller JGCM (1995) *"Strata Smith" and his Stratigraphic Cross Sections, 1819: A Review of Facts Worth Knowing about the Origin of Stratigraphic Geology in the Mind of William Smith (1769 1839), an English Country Sur veyor and Civil Engineer.* Tulsa: American Association of Petroleum Geologists; Bath: Geological Society Publishing House.

Knell SJ (2000) *The Culture of English Geology, 1815 1851*. Aldershot, Burlington, Singapore, and Sydney: Ashgate.

Phillips J (1844) *Memoirs of William Smith, LL.D. Author of the "Map of the Strata of England and Wales," by his Nephew and Pupil*. London: John Murray. Reprinted (1978). New York: Arno Press; and (2003) Bath: The Bath Royal Literary and Scientific Institution (with add itional essays on Smith by HS Torrens).

Sheppard T (1917) William Smith: his maps and memoirs. *Proceedings of the Yorkshire Geological Society* 19: 75 253.

Smith W (1815) *A Delineation of the Strata of England and Wales, with Part of Scotland; Exhibiting the Collieries and Mines, the Marshes and Fen Lands Originally Overflowed by the Sea, and the Varieties of Soil According to the Vari ations in the Substrata, Illustrated by the Most Descriptive Names by W. Smith*. London: (see copy of original map at http://www.unh.edu/esci/greatmap.html).

Smith W (1815) *A Memoir to the Map and Delineation of the Strata of England and Wales, with Part of Scotland*. London: John Cary.

Smith W (1816) *Strata Identified by Organized Fossils, Containing Prints on Coloured Paper of the Most Char acteristic Specimens in Each Stratum*. London: Printed by W Arding; sold by W Smith; J Sowerby; Sherwood, Neely and Jones; and Longman, Hurst, Rees, Orme and Brown. (see copy of original at http://www.unh.edu/esci/wmsmith.html).

Rudwick MJS (1996) Cuvier and Brongniart, William Smith, and the reconstruction of geohistory. *Earth Sciences History* 15: 25 36.

Torrens HS (2001) Timeless order: William Smith (1769 1839) and the search for raw materials 1800 1820. In: Lewis CLE and Knell SJ (eds.) *The Age of the Earth: From 4004* BC *to* AD *2002, The Geological Society, Special Publication No. 190,* pp. 61 83. London: The Geological Society.

Torrens HS (2002) *The Practice of British Geology, 1750 1850*. Aldershot and Burlington: Ashgate Variorum.

Winchester S (2001) *The Map that Changed the World*. London: Viking.

Steno

J M Hansen, Danish Research Agency, Copenhagen, Denmark

Introduction

Nicolaus Steno (Niels Stensen; **Figure 1**) was an anatomist, geologist, and bishop, often considered to be the founder of geology as a science. He was the first to describe the most fundamental principles of stratigraphy and crystallography, and the first to claim by rigorous arguments that fossils are the remains of former life on Earth. Steno's *principle of superposition* is simple but fundamental for all geologists and belongs to the first steps of understanding that a geology student must acquire. The principle states that geological strata were originally deposited horizontally under the action of gravity, and that the upper strata are younger than the lower are. The principle also states that inclined or folded strata have been tilted or disturbed subsequent to their deposition. The principle of superposition, theoretically, builds on Steno's statement that a crystal, sediment layer, or other kind of geological structure that takes the shape of the surface of another solid body is younger than the body from which it takes its shape. In conjunction with the principle of superposition, Steno's *principle of intersection* says that a body of rock or other geological structure is younger than those rocks or structures it is found cutting through.

Steno held forth yet another important stratigraphic principle, the *principle of reconstruction*. This states that it is possible to 'backstrip' a series of geological changes in reverse order, from the present to the past, having regard to the principles of superposition and intersection. In this way it is possible to obtain knowledge about still older situations. Thereafter, with this knowledge about past situations and their order of occurrence, the geological history of a locality, from the past to the present, can be reconstructed.

In mineralogy, Steno was the first to describe the principle of crystal growth, which leads to constant and specific angles between the sides of crystals of specific minerals (Steno's Law). In the philosophy of science and natural history, Steno founded the *principle of recognitive induction*, which made it possible to separate palaeontology and historical geology from theology.

Career, Science, and Beliefs

Steno travelled through large parts of Europe, visiting renowned scientists, academies, and universities, as if receiving scientific inspiration from the landscapes he saw. Before he was 30 years old, his anatomical studies had made him famous; in 1667, he was attached to the Medici Court in Florence, where Grand Duke Ferdinand II, impressed by Steno's anatomical and preliminary studies on fossils, made him a member of Accademia del Cimento ('Academy of Experiments'). During the next 2 years, Steno established the most important and permanent principles of what were to become the geological core disciplines: palaeontology, stratigraphy, and mineralogy.

Steno was born in Copenhagen of a Danish mother, Anne, and a Scanian father, Sten Pedersen. Niels was a fragile child and was brought up in the orthodox Protestantism of the Copenhagen of those days. Due to illness, he was kept indoors from his third to his sixth year. Isolated from other children, he listened to his parents and their friends' religious conversations

Figure 1 Steno (Danish: *Niels Stensen* or occasionally *Steensen* i.e., *Niels, the son of Sten*; Latin: *Nicolai Stenonis* or abbreviated *Steno*; French: *Nicolas Sténon*; Italian: *Niccolò Stenone*), as he most likely appeared *ca.*1667 69. Contemporaneous portrait, by the Dutch court painter Justus Sustermanns (1597 1681). The original is in the Uffizzi Gallery, Florence; a copy hangs in the Institute of Medical Anatomy, Copenhagen University. This version is from a poster made for a Steno exhibition at Tromsø University, Norway, in 1999. In other portraits of Steno can be seen his *sigilum*, a stylized asymmetrical heart from which a cross rises, a suggestion that scientific knowledge provides the highest praise to God.

and became familiar with mechanical and chemical crafts in his father's respected goldsmith shop. But shortly after Niels' recovery from illness, his father suddenly died. Because the boy was gifted, he was sent to Vor Frue Skole, a Lutheran academy, where he learned his fluent Latin from the enthusiastic Ole Borch. Borch also took Niels and the other pupils on botanical excursions around Copenhagen. Subsequently, Borch became one of Denmark's most distinguished intellectuals, holding professorships at the university in poetry, philology, chemistry, and botany. More than anyone else, Borch turned the young Steno towards becoming a scientist. He visited Steno in Holland and their friendship continued until the end of Steno's life.

In his eighteenth year, Steno enrolled at Copenhagen University, where he came under the influence of Thomas Bartholin, who, as head of the Faculty of Medicine, was famous as discoverer of the lymphatic vessels. At the time, Denmark and Sweden were at war. The city was besieged, the university was closed, and Steno assisted the students' defence of the barricades. Following the cease-fire in 1659, he managed to get out of the city by taking a ship to Rostock. Shortly afterwards, Steno turned up in Amsterdam, encountering a new world of scientific opportunities.

Steno soon moved to Leiden, where he continued to study the glands, the muscles, and the heart, and where he developed friendships with Jan Swammerdam and Baruch Spinoza. Steno went to Paris in 1665, where he presented his theories on the human brain and on muscles. According to a contemporary reviewer, Steno "turned upside down what is basic in medicine". Thus, besides Steno's geological and philosophical contributions, he is also famous for some important anatomical discoveries. The discovery of the duct from the parotid gland to the mouth is named *ductus stenonianus*, in recognition of just one of Steno's many anatomical contributions. Also important, but hardly recognized, was his description and understanding of the threefold division of the body fluids. In 1665, Steno gave the first modern description of the human brain, contradicting the interpretations of Galen, Willis, and Descartes. Moreover, the modern understanding of the anatomy and function of muscles and muscle fibres should be attributed to Steno.

After a year in Paris, Steno travelled to Italy, where his geological interest was to flourish as a member of the Medici court in Florence from 1667. On his way, he passed through southern France, where, at the University of Montpellier, he met the Englishmen Martin Lister and John Ray. Steno's geological achievements from his years in Tuscany, and his contribution to the principles of modern science, as they developed from Bacon, Galileo, and Descartes, have hardly received the reputation they deserve. Being a Dane (writing for the most part in a beautiful Latin, which was then in decline, rather than in the up-and-coming French and Italian, and only privately in English and German), he did not contribute to the national pride and fame of any large country or court. Moreover, his contemporary reputation was hindered by his criticisms of some of his most influential scientific contemporaries and by a superficial understanding of his religious conversion to Catholicism in the year (1669) that he wrote his most important geological work, *De Solido Intra Solidum Naturaliter Contento Dissertationis Prodromus* (translation: *The Prodromus to a Dissertation Concerning Solids Naturally Enclosed in Other Solids*). Eighteenth-century writers were puzzled by Steno's conversion from Protestantism to Catholicism and by his shift from an academic to a clerical career. His geological methods were, however, promptly applied in England, Germany, and Italy, but his name was rarely mentioned before Lyell, Humboldt, and Élie de Beaumont drew attention to his work in the 1830s.

In 1671, Steno was recalled to Copenhagen by Christian V, who, due to Steno's Catholic faith, could

not employ him as a university professor. Instead, the king made him Royal Anatomist, a title invented for the purpose. After a long journey through large parts of south-eastern Europe, Steno finally arrived in Copenhagen. However, in 1675, he obtained the king's permission to leave, and shortly after he became a Catholic priest in Italy. Soon the Pope called him to Hannover and Steno was appointed 'Bishop of Titiopolis', a now-defunct city of the Eastern Roman empire in Asia Minor. In reality, Steno was called by the Pope to lead the Catholic mission in northern Germany, Denmark, and Norway.

In Hannover, Steno met the Duke's librarian, Leibniz, who, after having read Steno's geological work, *De Solido*, in 1669, was inspired to write his theory, *Protogea*, concerning the Earth's origin. Leibniz became an admirer of Steno's geology and used him as scientific mentor and 'censor'. However, Leibniz was unsuccessful in persuading Steno to recommence his geological work. Most likely, Steno told Leibniz that he had, while in Florence, expanded his geological 'prodromus' (*De Solido*) into a more comprehensive geological dissertation, and had given it to Holger Jacobaeus, who was later Professor of Geography at Copenhagen University. After Steno's death, Leibniz wrote to several scientists, attempting to find out what had happened to Steno's geological papers. However, the extended version of *De Solido* has never been found.

After the period in Hannover, Steno dealt with numerous theological matters. These writings are collected in his *Opera Theologica* and *Letters*, and Steno did not really write on science any more, except for a philosophical letter to Spinoza. In the letter, written a few days before he was made 'Bishop of Titiopolis', Steno criticized Spinoza for having adopted a materialistic 'religion'. Steno died in Schwerin when he was 48 years old, weakened through several years of religiously inspired poverty and self-inflicted fasting. He was beatified in 1988.

Philosophy of Fossils and Recognition

In 1667, at the request of Ferdinand II, Steno wrote a brief dissertation (*Canis carchariae dissectum caput*) on the similarity between the teeth of living sharks and so-called tongue-stones, or glossopetrae (*glossa*: tongue; *petrus*: stone), now interpreted as fossilized shark teeth. Through this work, Steno is considered to be the first scientist to have established a series of empirical and rigorous arguments in order to describe fossils and interpret them as the geologically preserved remains of former living organisms. In six 'conjectures' (*conjecturae*), Steno explained that solid 'bodies' resembling parts of marine animals are indeed the remains of the things they resemble, provided the resemblance is found on every scale and in every visible detail. Further, Steno stated that such things do not grow in the Earth but have been deposited there by natural processes in the past, and that fossils should not be understood as inexplicable imprints of God's finger, but as representatives of the things they resemble (**Figure 2**).

In *De Solido* (1669), Steno made a general statement of his six 'conjectures' on the origin on fossils, formulated 2 years earlier. Now Steno enunciated a general geological, as well as a basic philosophical, principle: the *principle of recognition*: "If a solid body resembles another solid body in all respects, not only in the state of its surface but also in the internal arrangement of the parts and particles, it will resemble it also in the method and place of production". Steno mentioned that the similarity may only be structural and textural, and not necessarily chemical. He noticed that some fossil shells have been petrified, or the material substituted by other 'smallest parts' (elements, minerals, and sedimentary particles), different from the materials of which they were originally formed. The shape and visible structures of the original body may still be preserved even though the original material has been substituted by sediment or some type of mineral other than that of the original crystal, bone, or shell. Giving examples and descriptions, Steno further declared that the 'principle of recognition' is valid not only for fossils but also for geological strata, crystals, and any other solid body embedded in the earth. Nature's laws are 'univocal', thus similar conditions produce similar products.

A scientist should believe in direct observations and in reasoning derived from the observations, even if such reasoning implies a historical development of Earth and dramatic changes in the distribution of land and sea through time. A scientist should not trust speculations when they are contradicted by observations. The human ability to recognize things is inherent and makes possible the basic method of empirical science. In 1673, Steno further explained his conception of the senses and human reasoning: "It is not the function of the senses to display things as they are or to judge them, but to transmit to the reason those conditions of the things to be examined, which are sufficient for acquiring a knowledge of things appropriate to man's purpose". In other words, Steno realized that recognition is the most fundamental cognitive capacity in humans. It is prior to recognitive induction, or generalization, in science. Recognition is a prerequisite for generalization. *A priori* skills are 'above' *a posteriori* skills. Thus, Steno's principle of recognition is not only a geological principle, but also a general theory about cause and effect and regularity in nature. Furthermore,

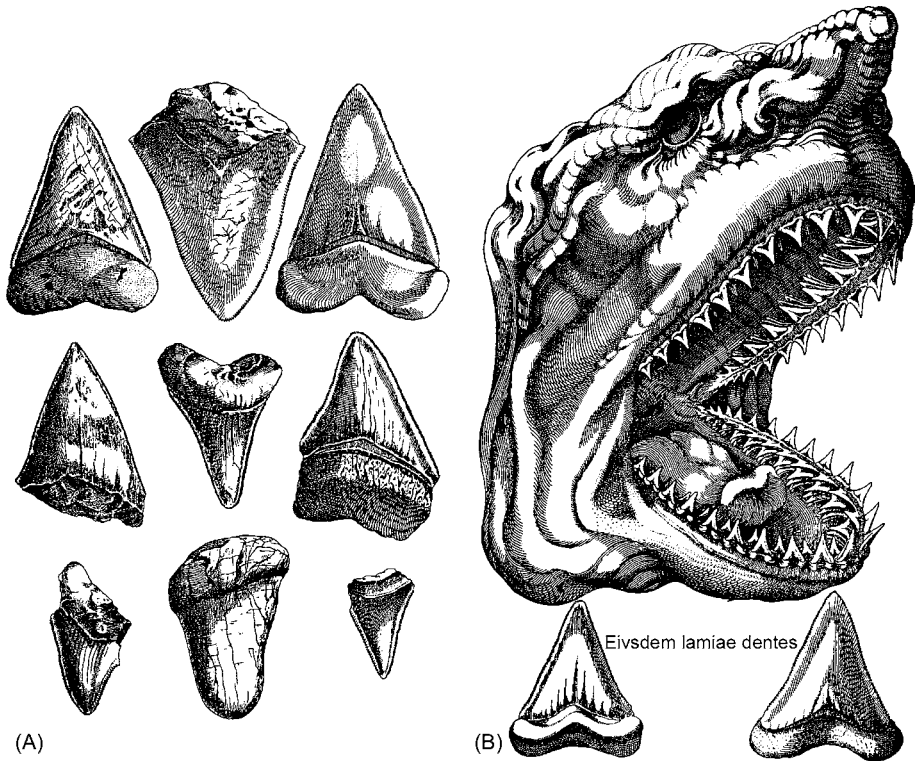


Figure 2 Steno's personal drawings of fossil shark teeth (A), compared with a contemporaneous artist's imaginative drawing of a monster shark (B) caught near Livorno and dissected by Steno. Note how the artist has been instructed to draw the anatomically correct arrangement of the teeth, which, when they are worn off, are substituted by new teeth rolling forward (another discovery made by Steno).

his principle was probably history's first theory of perception that was built on anatomical arguments relating to the capacity of the sensing organs and the brain to obtain relevant, albeit partial, knowledge from the evidence provided by nature.

Philosophy of Stratigraphy and Reconstruction

During his travels through the mountainous regions of southern Europe and during his numerous excursions in Tuscany, Steno noticed a large number of geological structures. In Steno's time, these structures had found no scientific explanation. At the time, most, if not all, scientists believed that landscapes and structures in Earth's crust derived from the time of Creation or had been formed during the turmoil of the Flood. However, through his studies of fossils, Steno became convinced that Earth had a history, accessible to scientific and human understanding over and above (though not generally conflicting with) the explanations in the Scripture. His interest was directed to all scales, not simply the small-scale structures of minerals and fossils. The structure of rocks, strata, and formations and their similarity on both sides of gorges and valleys became another philosophical and perceptual problem that Steno wanted to solve. Similarly, he wanted to find methods for exploring the history of the large-scale structures of mountain ranges.

Through his work on fossils from 1667, and 2 years later in *De Solido*, Steno dared to formulate and apply the core of his philosophy of science. This led to his definition of the fundamental stratigraphic principles of superposition and intersection, on how to find chronological and causal clues in geological bodies, in order to reconstruct their history. Moreover, it led to understanding of the general principles of crystal growth (Steno's Law). In consequence, *De Solido* is generally considered as the first scientific work on geology.

The basis of Steno's geological methods was a combined actualistic and 'forensic' procedure, proclaimed in the introduction of *De Solido*: "Given a substance endowed with a certain shape, and produced according to the laws of nature, to find in the body itself clues disclosing the place and manner of its production". Steno proclaimed that geological structures should be read according to the assumption that the present laws of nature were also in operation in the past.

Moreover, the natural structures of solid rocks and strata provided clues that could lead to 'demonstrative certain' understandings of how and in which environment (place) any given geological structure has been produced.

De Solido is structured in five parts: (1) the aforementioned introduction to the Grand Duke, (2) a philosophical part, in which the fundamental principles and methods are explained and formulated in general terms, (3) an empirical part, with numerous examples on how to apply the fundamental principles, (4) a description and interpretation of Tuscany's geological history, based on Steno's fundamental principles, and (5) acknowledgements from Steno's 'peer reviewers', the eminent scientists Vincenzio Vivianni and Francesco Redi. In the second (philosophical) part of *De Solido*, Steno summarized his geological understanding in three general 'propositions' about the way nature works and how it should be 'read'. The proposition on fossils and recognition was derived from Steno's previous work on fossils. A new proposition described the basis of stratigraphy and reconstruction: "If a solid body is enclosed on all sides by another body, the first of the two to harden [to attain a certain form] was that one which, when both touch, transferred its own surface characteristics to the surface of the other". From this general 'proposition' Steno developed a series of chronological principles, including the principles of superposition, intersection, and reconstruction. He gave a long series of examples from sediments, volcanic rocks, veins, crystals, fossils, etc., on how his principles works in practice. He realized that the principles would make it possible to reconstruct a scientifically plausible description of the historic development of Earth. He emphasized that the reconstructive method could show the succession and type of geological events, though the duration of the geological history was still unknown: "On this issue Nature is silent, only Scripture speaks".

By the help of his general 'propositions' and the associated superposition principle in Part 4 of *De Solido*, Steno showed how the geological history of Tuscany could be separated into six stages. There were two stages when the region was flooded by water and when its geological strata were deposited, two stages when it was flat and dry land formed by crustal uplifting of the strata previously deposited in water, and two stages when it was an uneven mountainous landscape eroded by rivers and deformed so that previously horizontal strata had been tilted, and again covered by younger horizontal strata.

In a cartoon-like series of didactic drawings (see **Figure 3A**), Steno showed how to reconstruct a region's geological history. The reconstruction must begin with the present state of affairs. Then, by the help of the superposition principle, it must be discovered what the situation was immediately prior to the present. When that is known, the situation immediately before this second-last situation must be discovered, and so on, with the third-last, fourth-last, etc., until it is impossible to identify any older situation. Then, when the different situations and their order of occurrence are known, the geological history can be reconstructed, beginning with the oldest known situation.

By good fortune, Steno's approval from the 'peer reviewers' had been easy to obtain, because he had worked in an area that could be interpreted relatively easily, and where there appeared to be no serious contradiction between Nature and Scripture. Tuscany had been flooded twice, first at the time of the Creation, before animals and plants lived on Earth (for which reason no fossils are found in the first sediments deposited by water), and again during the Flood and other marine transgressions (that is, after the creation of animals and plants, for which reason fossils are found in the sediments deposited during the Flood or later).

Philosophy of Crystals and Growth

Steno's third proposition in *De Solido* deals with the nature of growth. This included all kinds of natural growth, no matter whether it takes place in the inorganic or in the organic realm. Steno was inspired by Kepler's mathematical study of dense packing of 'atoms' and how snow crystals become a certain shape. From his own studies of sediments and crystals, however, Steno realized that crystal growth will give rise to regular external forms that cannot be produced by sedimentary processes, but also to structures more complex than those that can be produced by the packing of identical 'atoms' (see **Figure 3B**). Steno insisted that growth must be understood as a general problem, not only for crystals. He concluded that "if a solid body was produced according to the laws of nature, it was produced from a fluid." On reading *De Solido* and Steno's earlier anatomical papers, it is easy to understand how he reached this perceptive view. Superficially, the third proposition may seem odd. However, it reflects deep insight into change: all changes are results of motion. Motion is expressed in three basic forms: (1) as when we make a journey or an animal is running (i.e., change of location), (2) as when water runs in a river (i.e., flow), and (3) as "the first and hitherto unknown cause of motion", which Steno had already (in *De Thermis*) described as heat, and now also described as the motion of matter's smallest parts (i.e., diffusion). Thus, Steno envisaged three fundamental types of change: in modern terminology, this is change of location (or dislocation), flow, and diffusion.

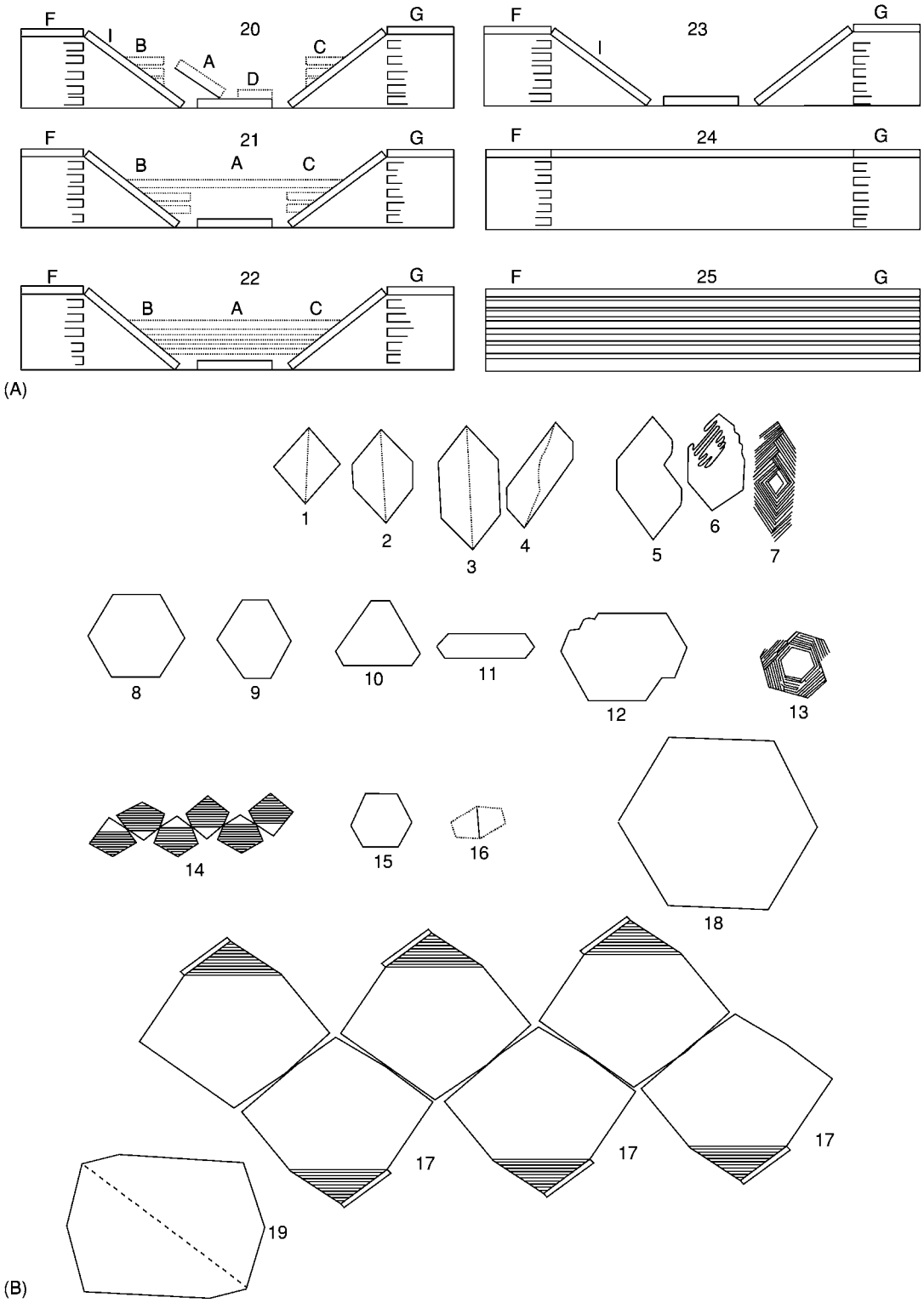


Figure 3 (A) Steno's model showing how to reconstruct the geological history of Tuscany. (B) Drawings of various crystal forms and indications of how crystals grow and dissolve.

All other kinds of change can be reduced to combinations of these three basic types!

Now, after this third proposition, Steno explained geological types of growth. Sediments increase with the help of gravity, which adds sand grains and other dispersed particles to the bottom from a slurry, or by traction of particles along the bottom until they hit obstacles. Crystals also grow by external addition, but for different reasons. The attractor is not gravity, but some unknown force, because crystals may grow even from the roof of a cavity or cavern. Steno explained geological growth as follows:

> Additions made directly to a solid from an external fluid sometimes fall to the bottom because of their own weight, as is the case with sediments; sometimes the additions are made from a penetrating fluid that directs material to the solid on all sides, as is the case of incrust ations, or only to certain parts of the solid, as is the case of those bodies that show thread like forms, branches, and angular bodies.

There is no doubt that Steno held that all kinds of solid growth, whether inorganic or organic, sooner or later could be explained as being due to an external addition of dissolved or dispersed material to pre-existing surfaces. The unsolved problem appeared not to be how growth takes place, but how crystals begin to crystallize and maintain a certain form as expressed by the constant angle between their crystal faces. The solution to these problems had to be found in the imperceptible smallest parts of the crystallizing matter, because neither the first 'germ' of a crystal nor its developing faces take shape from the substrate on which it grows. The growing crystal 'moulds' the substrate, but its faces are different from the latter. Kepler's paper on the dense packing of identical 'atoms' did not explain this, because it would imply that crystals could only be hexagonal, trigonal, or cubic, but not rhombic, and certainly not monoclinic or triclinic, as Steno knew some crystals to be.

Philosophy of Science and the Limits of Knowledge

The fundamentals of Steno's inductive/empirical philosophy of science were formulated in 1665 in Paris, when Steno gave history's first modern description of the human brain, including realistic drawings, completely different from those of his contemporaries Willis and Descartes. Steno opposed Descartes' conception of the brain and showed that Descartes' theory was built on pure speculation about God's impact on the human will, acting through the pineal gland. In Descartes' opinion, God controlled the soul and the human will by vibrating and rotating the pineal gland at the centre of the brain. Then, when the pineal gland, by such 'divine impacts', made the gland touch various parts of the brain, the contacts supposedly made the body react correspondingly, as if the brain were a mechanical machine and God its driver. On the contrary, said Steno, the brain is so fragile, and its structures so fine and complicated, that it "cannot even comprehend itself".

Prompted by his anatomical researches, matured through his founding of a scientific basis for the study of Earth, and made humble by his strong religious beliefs, Steno came to a clear and rigorous philosophy of science, close to that of modern scientists. During his geological studies of Tuscany and under the influence of what he had come to understand about Earth's history and changes, Steno developed his philosophy of science in order to separate geology and medicine from theology. When encountering something that is not understood, it is necessary to find something in it that is intelligible and compare it with something that is known and can be produced. The philosophical basis for *De Solido* can thus be summarized as follows: Specific recognitive induction (recognition) and specific empirical investigations (experiments, dissections, fieldwork, etc.) must regulate more general deductions and speculations (generalizations), but must not overrule deductive reasoning and perspectives, which for obvious reasons cannot be observed by the human senses or comprehended by the human mind. The past must be studied through knowledge about the present, but the past and present realities are much greater than scientific knowledge about it can ever be.

This should be understood to mean that it is not possible to observe the past *per se*, but only 'imprints' of past events. So, when seeking to interpret the past, primary emphasis must be placed on those clues that can actually be observed. This should be understood so that explanations about the inability to know anything directly are not neglected. However, such perspectives must, in contrast to Descartes' misuse of deduction, always respect what is known with the aid of the senses and by rigorous reasoning.

This way of thinking led to a general 'Kantian' (though pre-Kant) theory of human perception and interpretation of nature. In his Copenhagen lecture of 1673, Steno generalized his views on the difference between things 'as we see them' and things as they are 'in themselves'. There will always be a difference between nature as it is, and nature as humans interpret it. Modesty, caution, and scientific rigour should be key in attempts to understand things. Humans must believe in their immediate sensory capacities, and in what scientifically founded investigations reveal to the senses. Finally, it is important to believe in what the senses

transmit to the reasoning. However, because human senses are imperfect and reasoning capacity is incomplete, the truth can only be approached, and complete understanding cannot be obtained.

Steno concentrated his philosophy of science in the following sentence, which has been cited more than anything else from his hand:

> *Pulchra sunt quae videntur,*
> *Pulchriora quae sciuntur,*
> *Longe pulcherrima quae ignorantur.*

In English, this reads as follows:

> Beautiful is what we see,
> More beautiful is what we know,
> Most beautiful is that about which we are insensible.

This has erroneously been interpreted to mean that Steno ranked religious belief above scientific knowledge. But he explained the aphorism in this way:

> Yes indeed, after having rejected all the errors of the senses, who would not repeat: beautiful is what appears to the senses without dissection; more beautiful what dissection draws forth from the hidden interior; yet far the most beautiful is what, escaping the senses, is revealed by reasoning helped by what the senses have already perceived.

In Steno's philosophy, humility about scientific principles and scientific understanding offered the highest praise to God. Scientific knowledge must never be ruled by clerical beliefs and powers. On the contrary, science will guide us towards the truth, which, however, will never be fully understood because of our limited sense capacity and imperfect intellectual resources.

See Also

History of Geology Up To 1780.

Further Reading

Cutler A (2003) *The Seashell on the Mountaintop. A Story of Science, Sainthood, and the Humble Genius who Discovered a New History of the Earth*. New York: EP Dutton.

Garboe A (1954) Nicolaus Steno (Niels Stensen) and Eras mus Bartholinus: two 17th century Danish scientists and the foundation of exact geology and crystallography. *Bulletin of the Geological Survey of Denmark. 4th Series* 3: 1 12.

Garboe A (1960) Niels Stensen's (Steno's) lost geo logical manuscript. *Bulletin of the Geological Survey of Denmark* 14: 243 246.

Gould SJ (1981) The titular bishop of Titiopolis. *Natural History* 90: 20 24.

Kardel T (1994) Steno: life, science, philosophy (with Niels Stensen's *Prooemium* or preface to a demonstration in the Copenhagen Anatomical Theater in the year 1673, and Holger Jacobaeus: Niels Stensen's Anatomical dem onstration No. XVI, and other texts translated from Latin). *Acta Historica Scientiarum Naturalium et Medi cinalium* 42: 1 159.

Moe H (1988) *Nicolaus Steno: An Illustrated Biography*. Copenhagen: Rhodos.

Noe Nygaard A (1986) Nicolaus Steno, paleontologist, geologist, crystallographer. In: Poulsen JE and Snorrason E (eds.) *Nicolaus Steno 1638 1686. A Re consideration by Danish Scientists*, pp. 167 190. Copenhagen: Nordisk Insulinlaboratorium.

Rodolico F (1971) Niels Stensen, founder of the geology of Tuscany. *Acta Historica Scientiarum Naturalium et Medicinalium* 23: 237 243.

Rudwick MJS (1972) *The Meaning of Fossils*. New York: MacDonald, Elsevier.

Scherz G (ed.) (1969) *Steno: Geological Papers* (translated by AJ Pollock). Odense: Odense University Press.

Steno N (1669) *De Solido Intra Solidum Naturaliter Contento Dissertationis Prodromus*. Florence: Ex Typo graphia Sub Signo Stellae (English translation in Scherz, 1969).

Suess

B Fritscher, Munich University, Munich, Germany

Introduction

Eduard Suess (Figure 1) was the most influential European geologist of late nineteenth and early twentieth centuries. As a professor of geology, he taught at the University of Vienna for nearly 45 years. In his major works, *Die Entstehung der Alpen* (*The Origin of the Alps*; 1875), and in the comprehensive *Das Antlitz der Erde* (*The Face of the Earth*; 1883–1909), he elaborated a 'global tectonics', based on the contracting hypothesis. By his works he created a new image of the structure and the formation of fold mountains and introduced basic terms of twentieth-century structural geology, such as the Laurentian and Angara Shields, the huge southern continent Gondwana, and the Tethys (as a former central sea, the precursor of the Mediterranean). Moreover, Suess was one of the pioneers of the doctrine of nappe folding in the Alps and he

Figure 1 Eduard Suess (1831 1914), in a portrait published in his posthumous 1916 memoir, *Erinnerungen*.

founded the concept of 'eustatic' sea-level changes. Suess is also remembered as an engineer and politician. He was a long-serving member of the Austrian national parliament and promoted and planned a new water supply for his home town, thus becoming one of the creators of modern Vienna.

Scientist, Engineer, and Politician

The founder of 'global tectonics', Suess was born into a well-connected bourgeois family. His father, a wool merchant, born in Germany, had worked and travelled in various European countries before setting up a wool business in London in 1828. Here, Eduard Suess was born on 20 August 1831. Three years later, the family moved to Prague, where Suess, as he later recalled, arrived as a 'complete English child' who understood not a single German word. In 1845, his father took over a leather factory near Vienna.

The young Eduard got a polyglot education from English, Belgian, and German tutors. In 1847, he entered the Polytechnic 'High School' (now Technical University) at Vienna, but soon left the revolutionary town of 1848, where he had participated in demonstrations, for the University of Prague. His early subjects were higher mathematics, physics, and descriptive geometry. The impressive collection of Silurian fossils at the museum in Prague roused his interest in geology, and he started to make excursions to nearby fossil-rich areas. Back at the Vienna Polytechnic in 1849, he continued his palaeontological studies in the surroundings of Vienna. In 1850, he presented a scientific paper (published in 1851) on Bohemian graptolites to Wilhelm Haidinger (1795–1871), who was then director of the newly established Austrian Geological Survey. His very first publication, however, had been a chapter on geology for a tourist guide of the Carlsbad region; Suess had written the chapter in 1850 during a visit to the region for a 'cure'.

After returning to Vienna, Suess was imprisoned for his participation in the demonstrations in 1848. Although he was released just a few weeks later, he had to leave the Polytechnic School, thus never acquiring a doctorate or any other formal university qualification. Nevertheless, in 1852, he was appointed an assistant in the Imperial Mineralogical Collection in Vienna. Assigned to classify the brachiopods of the collection, he published some pioneering studies in this field. His efforts to become a Privatdozent (private lecturer) at the university failed for his lack of a doctorate. By his palaeontological work, however, and by early travels accompanying well-known geologists such as Franz von Hauer (1822–99), Arnold Escher von der Linth (1807–72), Paul Deshayes (1797–1875), and Ernst Beyrich (1815–96), Suess already had a name among earth scientists. Thus, in 1857, on the recommendation of leading Vienna geologists, he was appointed 'professor extraordinary' of palaeontology at the university, and 5 years later was appointed the same position in geology. In 1867, Suess was appointed to a full professorship in geology, which he retained until 1901. For about three decades, Suess also travelled extensively throughout Europe.

An early engagement in school and university education marked the beginnings of Suess' political career. In 1862, he published an essay on the soils and the water supply of Vienna, showing that the numerous epidemics of that time, particularly typhoid, resulted from the city's water supply, which was at that time mainly based on wells. The following year, Suess was elected a member of the town council and was named head of a commission to study the water supply. He suggested that water should be brought by an aqueduct from mountains springs, about 70 km away; 10 years later, in 1873, the new pipeline began to operate, and the number of deaths from typhoid fever was subsequently reduced to one-fourth. Suess' second famous engineering project was the regulation of the Danube, designed to prevent the frequent flooding of the lower lying areas of Vienna. A canal was opened in 1875, and after 1876, there were no more major floods. Suess was also a member of the Diet of Lower Austria from 1869 to 1874, and he held a seat in the Austrian

Parliament from 1873 to 1896, being mainly engaged in implementing liberal reforms in the school system.

Suess was also an ordinary member of the Austrian Academy of Sciences from 1867, and served as its President between 1898 and 1911. The advancement of scientific cooperation between different scientific disciplines and different national Academies was one of his main goals. He also promoted the foundation of the Institute of Radium Research in Vienna in 1910. Suess was elected a member of various European scientific academies and the Geological Society of London awarded to him the Wollaston Medal in 1896. Suess died in Vienna on 26 April 1914 and was buried at Marz (Burgenland, Austria), where his family owned a small estate.

Palaeontology, Stratigraphy, and Eustatic Sea-Level Changes

Following his early studies of Silurian graptolites in Bohemia, Suess' main fields of study at the Imperial Mineralogical Collection were the palaeontology and stratigraphy of the Tertiary strata of the Vienna Basin. He entered scientific virgin territory by his brachiopod studies and was the first Austrian palaeontologist to be engaged in the classification of Tertiary mammals, thus providing an overview of the mammalian fauna of the Vienna Basin. Suess set the comparative method against the prevailing emphasis on merely taxonomic classification. Focusing on the anatomy of fossil species and comparing their modes of life with those of existing species, he created an early form of palaeobiology. As early as 1859, he discussed the ecology of brachiopods, and in 1875 he first used the term 'biosphere' (Biosphäre) to denote the distinct terrestrial sphere of the living organisms, which could be regarded as the surface of the lithosphere. These approaches were continued by his long-standing assistant Othenio Abel (1875–1946), who became one of the founders of modern palaeobiology.

Suess' stratigraphic work was mainly concerned with the Alps. It brought a new structural view to alpine stratigraphy, distinguishing chronological and spatial units. Suess cleared up stratigraphic problems of the European Rhaetian and Miocene, the latter in particular in the region of Eggenburg (lower Austria), north of Vienna. Basic studies related to the correlation of Alpine Triassic, Jurassic, and Cretaceous formations with their equivalents outside the Alps, and, together with Albert Oppel (1831–65), Suess correlated the development of the Alpine and Swabian Triassic strata.

The region of Eggenburg was the area where Suess first developed his concept of eustatic sea-level changes. In surveying the Tertiary beds, he remarked

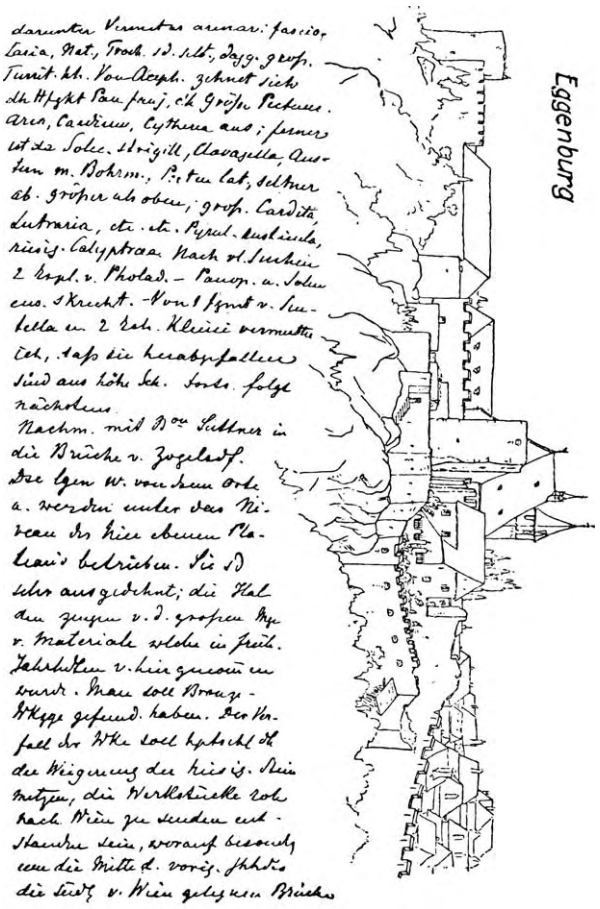
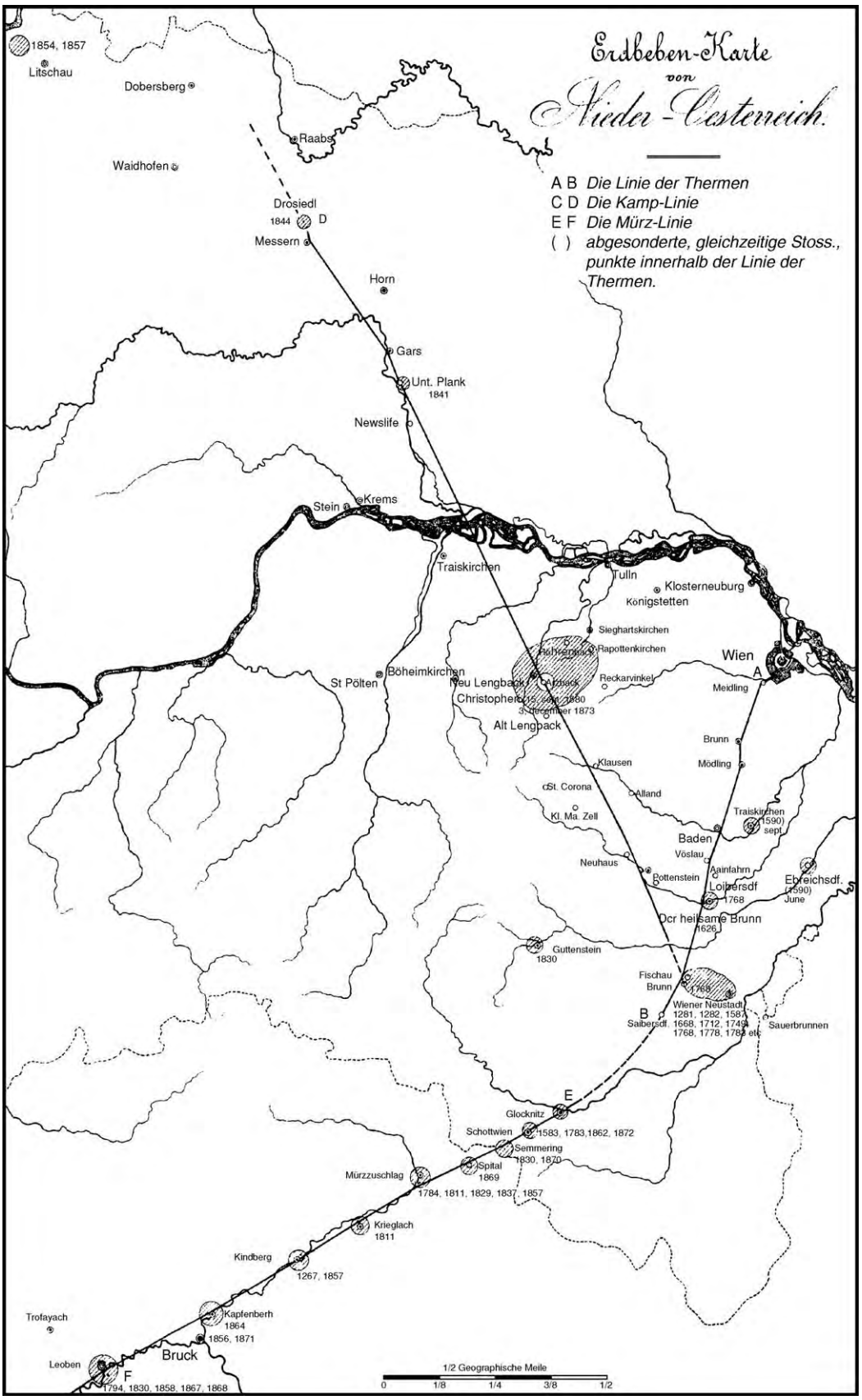


Figure 2 Sketch of Eggenburg (lower Austria) by Eduard Suess, from Suess' geological diaries, as published in *Erinnerungen* in 1916. The region of Eggenburg was crucial for Suess' ideas on large scale variations of sea level. In surveying the regularity of the ancient shorelines of this area, Suess first thought of what were later called 'eustatic' movements of the sea.

the regular height of the ancient shorelines of this area (Figure 2). These regularities seemed hardly explicable by an uplift of the land but, rather, by a fall of the sea-level. In 1885, after visiting Norway, he thought his ideas of the fall of the sea-level were confirmed by the stepped, horizontal terraces he had observed on the sides of the fjords and other valleys. Thus, 3 years later, he presented his theory of 'eustatic movements', i.e., of large-scale changes of the sea-level (separate from orogenic belts), which could be observed at approximately the same height over large parts of the earth.

The Origin of the Alps

In 1865, Suess was commissioned to produce a survey of the geology of the Austrian Empire, comprising at that time Hungary, Czechoslovakia, and parts of Romania and Poland. In the following


Erdbeben-Karte
von
Nieder-Oesterreich.
A B Die Linie der Thermen
C D Die Kamp-Linie
E F Die Mürz-Linie
( ) abgesonderte, gleichzeitige Stoss., punkte innerhalb der Linie der Thermen.
1854, 1857
Litschau
Dobersberg
Raabs
Waidhofen
Drosiedl
1844
D
Messern
Horn
Gars
Unt. Plank
1841
Newslife
Stein
Krems
Traiskirchen
Tulln
Klosterneuburg
Königstetten
Sieghartskirchen
Rapottenkirchen
Wien
A
Meidling
Reckarvinkel
St Pölten
Böheimkirchen
Neu Lengbach
Christophen
Alt Lengbach
Brunn
Mödling
Klausen
St. Corona
Alland
Kl. Ma. Zell
Traiskirchen
1590)
sept
Baden
Vöslau
Neuhaus
Pottenstein
Ainfahrn
Loipersdf
1768
Ebreichsdf.
1590)
June
Der heilsame Brunn
1626
Guttenstein
1830
Fischau
Brunn
1768
B
Saibersdf
Wiener Neustadt
1281, 1282, 1583
1668, 1712, 1749
1768, 1778, 1783 etc
Sauerbrunnen
E
Glocknitz
Schottwien
1583, 1783, 1862, 1872
Semmering
1830, 1870
Spital
1869
Mürzzuschlag
1784, 1811, 1829, 1837, 1857
Krieglach
1811
Kindberg
1267, 1857
Trofayach
Kapfenberh
1864
1856, 1871
Bruck
Leoben
F
1794, 1830, 1858, 1867, 1868
1/2 Geographische Meile
0
1/8
1/4
3/8
1/2

years, he extended his field studies to the Carpathians, the Sudetes, and the Apennines. In applying the comparative method, which he had already used in his palaeontological work, he found that all these mountain ranges (including the Alps) had certain structures in common. As a result, in 1875, he published his first major book, *The Origin of the Alps*, a comprehensive discussion of the origin and the structure of mountain chains, anticipating most of his later ideas on tectonics.

Suess's actual entry to these ideas involved studies on earthquakes, following a visit to Calabria in 1871, where Suess witnessed the devastation of the great earthquake of 1870. This gave rise to the idea of a systematic compilation of historical accounts of earth tremors in lower Austria, i.e., in a region far away from any volcanic activity. Suess found that these Austrian earthquakes were distributed along specific lines that cut across quite different rock formations. Thus, in 1873, he published two major articles on earthquakes in lower Austria and in southern Italy, concluding that earthquakes are restricted to specific structures within Earth's crust and are thus due to the same forces as those that gave rise to the formation of mountains (**Figure 3**).

The revolutionary concept that Suess now set forth in *The Origin of the Alps* abandoned the idea of similarity of the structure of mountain chains, which had dominated geology for nearly a century. Contrary to the theory of mountain formation by vertical upheavals due to eruptive rocks, favoured by Leopold von Buch (1774–1853) and Leonce Élie de Beaumont (1798–1874), Suess set forth his view of horizontal movements as the essential cause of the formation of folded mountain chains, entertaining the idea of unilateral horizontal overthrustings by tangential pressures, in the case of the Alps, directed from south to north. And Suess stated a fundamental difference between the mountain chains and their older, rigid 'forelands' (Vorländer), which act, so to speak, as 'earth dams' against the mobile chains. As a further characteristic feature of developing mountain chains, he put forward the idea of 'hinterlands', i.e., their usually curved and relatively depressed 'inner' sides (**Figure 4**). These inner sides were the location of volcanism and earthquakes, due to the tensions caused by the movement of the newly forming mountains towards the forelands. Already during his visit to the volcanic areas of southern Italy in 1871, Suess had thought of volcanoes and earthquakes as being associated with regions of rupture within Earth's crust. Thus, earthquakes and volcanism were closely related to the process of mountain formation. Suess explicitly stated that volcanoes and intrusions were just the side effects of mountain formation.

Figure 4 Sketch of the main lines of strike for the folds of the Carpathians and the Balkans, from the first volume (1885) of Suess' *Das Antlitz der Erde*. The Carpathian Mountains, in particular, initiated Suess' thoughts about groups of curved lines, and tangential and unilateral movements, as opposed to the prevailing assumption of a symmetrical structure of folded mountain chains. The actual structure of the mountain chains (i.e., the disposition of the folds) depended on the form of the foreland and the resistance of its materials.

A Contracting Earth

In *The Origin of the Alps*, Suess also first introduced his ideas on the forces that could have brought about mountain ranges and their accompanying phenomena: namely, the contraction or shrinking of the earth due to its continuous cooling. In the 1840s, the American geologist James Dwight Dana (1813–1895) had proposed the idea of unequal radial contraction of the earth due to cooling, linking this to the origin of earthquakes. In the early 1870s, Dana and other geologists extended the contraction hypothesis as the basic assumption of the formation of mountains. Thus, Suess was neither the creator nor the chief advocate of the contraction hypothesis. However, it was Suess who – referring to Dana, to Robert Mallet's (1810–81) ideas on 'volcanic energy', and to Charles Darwin's (1809–82) theory of the origin of earthquakes due to the formation of fissures – most clearly used the contraction hypothesis for more than

Figure 3 A map from Suess' 1873 book *Die Erdbeben Nieder Österreichs*, showing the distribution of earthquakes in lower Austria. According to the position of the strongest effect of the respective earthquakes (years given near the names of the villages), Suess constructed three 'earthquake lines' (Erdbebenlinien). The Thermen Linie (A B) was named according to the thermal and sulphur springs along this line (for instance, at Brunn, Voslau, and Meidling). The Kamp Linie (C D) and the Murz Linie (E F) were both named according to rivers of the area.

30 years as a basis for explaining all the tectonic features of Earth's crust.

Arguing against the theory of mountain formation by volcanic elevations, Suess pointed to his studies in the Alps, which had offered only a single example that might confirm this theory: in the Euganean Hills (Italy), where a mass of Jurassic and Cretaceous limestones seemed to have been moved by trachyte. In the final chapter of *The Origin of the Alps*, Suess related the structures of mountain ranges to irregular earth contraction, and first used his famous phrase of 'the face of the earth'. He emphasized that, notwithstanding the irregularity of the contraction, its direction seemed to have remained the same over large regions and extremely long periods of time. In this, Suess was anticipating his later distinction of Caledonian, Armorican, and Variscan folding.

Global View: The Face of the Earth

In 1885, Suess published the first volume of his best known work, *The Face of the Earth*. The second volume followed in 1888, and the two parts of Volume 3 appeared in 1901 and 1909, respectively. An English translation in five volumes was published between 1904 and 1924. The most important edition became the French translation under the direction of Emmanuel de Margerie (1862–1953), in which thousands of new footnotes and about 500 figures were added. The monumental work was not simply an extension and more detailed discussion of Suess' previously formulated ideas. Rather, *The Face of the Earth* provided a 'global tectonics'. Embodying the results of his own travels and research all over Europe, Suess compiled the then-known materials relating to the tectonic structures of Earth's crust and created the 'language' that made possible a global view of the planet's tectonic features (**Figure 5**).

At the commencement of the tectonic development of the present landmasses, Suess distinguished four ancient continents: Laurentia, Angaraland (in what is now northern/central Siberia), Gondwana, and Antarctica (**Figure 6**). A further ancient block was the Baltic Shield. The later term for this, Fennoscandia, however, was actually not used by Suess. Between the Eurasian and the Indo-African blocks extended a series of younger mountain chains. Suess thought of them as having originated in the Tethys, an ancient Mediterranean sea encompassing half the globe, from Central America to the Sunda Islands (the name was coined according to the Greek goddess of the sea). A second series of young mountain chains, forming festoons and garlands, encircled the Pacific Ocean.

In addition to the youngest epoch of folding, the Alpine Orogeny, Suess distinguished two more principal stages of mountain formation in Europe: the Caledonian and the Variscan (also known as the Armorican and the Hercynian) orogenies. Their ages increased in passing from south to north. Though the most recent (Alpine) chains around the Mediterranean were already subsiding, the older ones, now eroded and covered by younger deposits, are presently at rest, but could be reactivated. According to Suess' theory of mountain formation, the subsidence and the formation of large grabens were predominant characteristics of Earth's crust, as against horizontal dislocations, such as folds and overthrusts. With reference to Suess' hypothesized periods of mountain formation, in 1887 the French geologist Marcel Bertrand (1847–1907) developed a tectonic classification of Earth's history, also correlating the different periods of folding with those of increased igneous intrusions. Thus, Suess' periods of folding became a continuous process that had built up the European continent from south to north.

In addition to these basics of a 'global tectonics', Suess introduced a series of further large-scale features of Earth's surface, such as his distinction of different types of coasts. According to the direction of the axes of fold mountains, he distinguished a 'Pacific' and an 'Atlantic' type. Around the Pacific Ocean, the fold lines were running more or less parallel to the coast, whereas in the case of the Atlantic Ocean, they ran approximately at right angles to the coast.

In the first volume of *The Face of the Earth*, Suess also further elaborated his ideas on contraction. Meanwhile, the hypothesis had received strong support by the Swiss geologist Albert Heim (1849–1937), who, in 1878, had calculated the supposed reduction of the circumference of Earth due to its shrinking or contraction. For the Jura mountains and the Alps, Heim estimated a relative compression of about four-fifths and one-half, respectively (compared to the original width, which was estimated by mentally 'smoothing out' both chains). Relative to the full circumference of Earth, the shrinking of Earth due to the folding would be almost 1%. Such a process (for which Heim thought a cooling of 200°C would be sufficient) going on throughout the whole history of Earth was quite plausible according to the contraction hypothesis.

Relying on his ideas on earthquake lines, and on further detailed discussions of the various phenomena of Earth tremors, Suess constructed a system of dislocations in rock formations, due to the reduction of the volume of the globe. The tensions produced by the process of contraction would tend to differentiate into tangential and radial tensions, thus producing both horizontal (i.e., pushing and folding) and

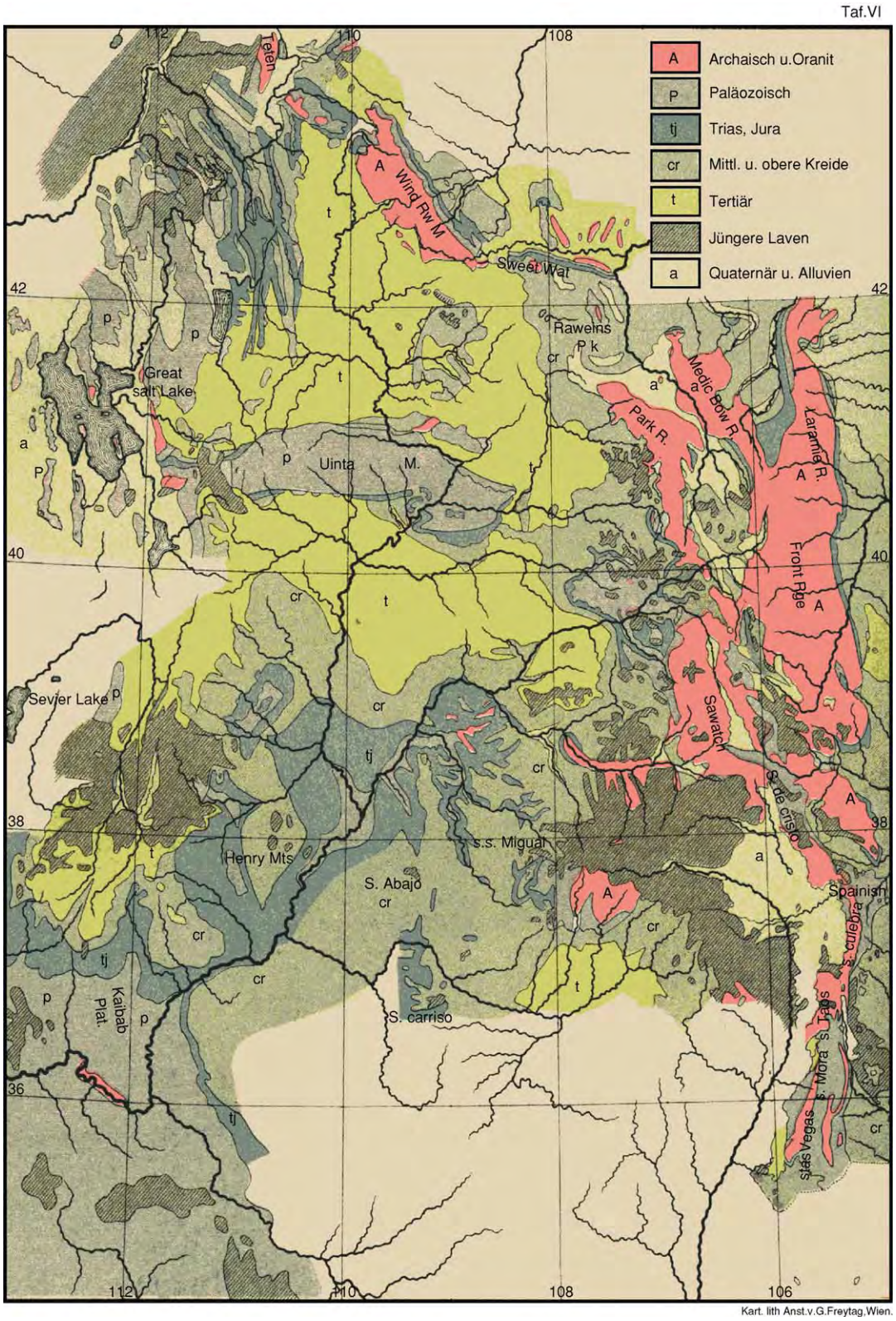


Figure 5 Map showing divergent branching (Virgation) of the Rocky Mountains, from the first volume (1885) of Suess' *Das Antlitz der Erde*. To convey the complexity of the structures of folded mountain chains, Suess often used the term 'Virgation', meaning the rodlike spreading out, or dispersion, of the individual branches towards their forelands. Such an order also meant that a region that was the 'hinterland' for one branch served as the 'foreland' for another. Translation of the key: A, Archaean rocks and granite; p, Palaeozoic; tj, Triassic and Jurassic; cr, Cretaceous; t, Tertiary; cross hatching, younger lava; a, Quaternary and alluvium.

Figure 6 Section of Suess' map (from the third volume, part 2, of *Das Antlitz der Erde*), showing the arrangement of the tectonic units of Earth. The rose coloured areas are Laurentia and Gondwana; brown, Caledonides and Saharides; green, Asian structures with Angaraland; pale violet, Oceaniden, Australia, and Antarctica; yellow, Cape Mountains; cross hatched, volcanic islands of Atlantic type.

vertical (or subsiding) movements. Thus, Suess distinguished two groups of dislocations, one produced by more or less horizontal movements of mountains, the other one by more or less vertical movements (i.e., by subsidence). At the end of the first volume of *The Face of the Earth*, Suess gave his famous statement of the history (and the future) of Earth: "What we are witnessing is the collapse of the terrestrial globe". He also linked Earth's contraction to the development of life. Subsidences had made possible the accumulation of water in the deep oceans, and, at the same time, the emergence of the continents, which became the home of organisms that breathe with lungs.

This idea of the linkage of the development of life on Earth to its tectonic history was elaborated at the end of the last volume of *The Face of the Earth*. Suess claimed that his ancient shields (Laurentia, Angaraland, Gondwana, and Antarctica) were the essential regions for the development of life. These areas supposedly did not participate in folding and transgressions for a long time. Consequently, the development of life in these areas should show fewer disturbances than elsewhere, and therefore Suess called them 'asylums' (Asyle). From the asylums, the distribution of the vegetation should have started again after the great tectonic changes.

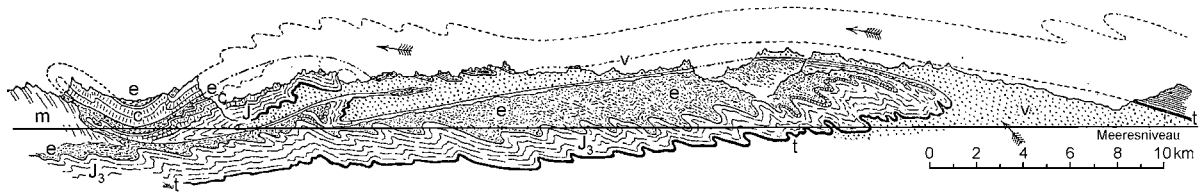


Figure 7 The Glarus double fold (according to a sketch by Albert Heim, as published in volume 3, part 2, of Suess' *Das Antlitz der Erde*), explained as a single overthrust towards the north. m, Tertiary Nagelfluh; e, Lower Tertiary flysch; c, Cretaceous; J, Jurassic; t, Helvetic Triassic; V, Verrucano. The sea level (Meeresniveau) is indicated.

Nappe Folding

The publication of *The Face of the Earth* covered a period of 24 years, and some of Suess' early concepts were changed during this time. The most striking example was the doctrine of nappe fold structures, as developed from the 1880s. Suess may have implicitly accepted large overthrusts as processes of mountain formation, at least since the 1870s. And he actually played a significant role in the development of the new doctrine, which is commonly ascribed to Marcel Bertrand. As early as 1883, Suess gave a new interpretation of Heim's 'Glarus double fold' (Glarner Doppelfalte) in the Glarus Canton of Switzerland (Figure 7). Whereas Heim had the idea of two folds from both north and south, Suess suggested a single overthrust toward the north. Nevertheless, it was not until publication of the third volume of *The Face of the Earth*, and in particular its second part, that Suess included nappe folding in his discussion of the processes of mountain formation. He discussed several of these structures, as found in the young European mountain ranges, accepting also large amounts of overthrusting, such as, for instance, in the case of the Dent Blanche massif in Switzerland, which had recently been proposed by the Swiss geologist Émile Argand (1879–1940).

Suess in the Twentieth Century

European geological thinking was deeply influenced by Suess' tectonics: a great number of his concepts and terms became standard in twentieth century earth sciences and are still current. On the other hand, Suess' basic theoretical assumption, i.e., the hypothesis of the contracting Earth, was subjected to critiques, even during the publication of *The Face of the Earth*. In 1912, only 3 years after the publication of the last volume and 2 years before Suess' death, Alfred Wegener (1880–1930) came up with quite new ideas on the origin of continents and oceans. Wegener was Suess' most serious critic. Moreover, the recognition of radiogenic heat, produced within Earth, threw grave doubt on the idea of a cooling and contracting planet. Nevertheless, it was the Viennese geologist who provided Wegener with essential information about the large-scale features of Earth, indicating former connections and movements. Thus, notwithstanding Suess' errors, his work remains among the most impressive and comprehensive scientific theories of Earth ever written.

See Also

Africa: Pan-African Orogeny; North African Phanerozoic; Rift Valley. **Andes**. **Antarctic**. **Argentina**. **Australia:** Proterozoic; Tasman Orogenic Belt. **Brazil**. **Famous Geologists:** Wegener. **History of Geology From 1835 To 1900**. **New Zealand**. **Oceania (Including Fiji, PNG and Solomons)**. **Plate Tectonics**. **Shields**. **Tectonics:** Earthquakes; Folding; Mountain Building and Orogeny. **Volcanoes**.

Further Reading

Cernajsek T, Csendes P, Mentschl C, and Seidl J (1999) "...hat durch bedeutende Leistungen...das Wohl der Gemeinde mächtig gefördert." *Eduard Suess und die Entwicklung Wiens zur modernen Großstadt*, Österreichisches Biographisches Lexikon Schriftenreihe 5. Vienna: Institut Österreichisches Biographisches Lexikon.

Greene MT (1982) *Geology in the Nineteenth Century. Changing Views of a Changing World*. Ithaca and London: Cornell University Press.

Hamann G (ed.) (1983) *Eduard Suess zum Gedenken (20.VIII.1831 26.IV.1914)*. Sitzungsberichte der Österreichischen Akademie der Wissenschaften, Philologisch historische Klasse 422. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Şengör AMC (1982) Eduard Suess' relations to the pre 1950 schools of thought in global tectonics. *Geologische Rundschau* 71: 381 420.

Şengör AMC (1998) Die Tethys: vor 100 Jahren und heute. *Mitteilungen der Österreichischen Geologischen Gesellschaft* 89: 5 177.

Şengör AMC (2003) *The Large Wave Deformations of the Lithosphere: Materials for a History of the Evolution of Thought from the Earliest Times to Plate Tectonics, Memoir 196*. Boulder: Geological Society of America.

Suess E (1862) *Der Boden der Stadt Wien nach seiner Bildungsweise, Beschaffenheit und seinen Beziehungen zum bürgerlichen Leben: Eine geologische Studie*. Vienna: W Braumüller.
Suess E (1873) *Die Erdbeben Nieder Österreichs*. Vienna: K Gerold's Sohn.
Suess E (1875) *Die Entstehung der Alpen*. Vienna: W Braumüller.
Suess E (1885 1909) *Das Antlitz der Erde* (three volumes; vol. 3 in two parts). Prague and Leipzig: F Tempsky and G Freytag.
Suess E (1897 1918) *La Face de la Terre* (three volumes; vol. 3 in four parts) (Traduit et annoté sous la direction de E de Margerie. Avec une préface par M Bertrand). Paris: Librairie Armand Colin.
Suess E (1904 1924) *The Face of the Earth* (five volumes). (Translated by HBC Sollas, under the direction of WJ Sollas.) Oxford: Clarendon Press.
Suess E (1916) *Erinnerungen*. Leipzig: S Hirzel.
Tollmann A (1981/1982) Die Bedeutung von Eduard Suess für die Deckenlehre. *Mitteilungen der Österreichischen Geologischen Gesellschaft,* (special volume to mark the 150th anniversary of E. Suess' birth) 74/75: 27 40.
Wegmann E (1976) Eduard Suess. In: Gillispie CC (ed.) *Dictionary of Scientific Biography* 13, pp. 143 149. New York: Charles Scribner.

Walther

I Seibold, University Library, Freiburg, Germany

Career

Johannes Walther (Figure 1) was one of the early pioneers in sedimentology, introducing a modern approach that combined both lithological and biological aspects.

Walther was born on 20 July 1860, the son of a vicar in Neustadt/Orla in the German province of Thuringia. In his boyhood he was handicapped by a nervous disease (probably caused by a serious fall) that prevented his regular attendance at school and as a result he did not receive a leaving certificate. But his intelligence and enthusiasm for science were so evident that he obtained special permission to study at the University of Jena and his health was restored during his period as a student. Walther was awarded his PhD in zoology in 1882, following which he went to Leipzig and Munich to study geology and palaeontology more intensively. At the same time, he started his first investigations of the sea floor in the Bay of Naples, where Anton Dohrn's marine biology station was based. In 1886, Walther became a lecturer in Jena; in the following years he undertook extended geological expeditions abroad. During his time as lecturer Walther was almost without income, for his position had no salary, apart from the small lecture fees. He therefore had to make a living by writing (e.g., newspaper articles) and giving public lectures. However, in 1894 he was appointed associate professor (Haeckel Professor) at Jena, a post endowed by a wealthy Swiss admirer of Walther's teacher, the famous zoologist Ernst Haeckel.

In 1899, Walther married Janna Hentschel. They had two children, a son and a daughter. From 1906 until his retirement, he was Director of the Geological Institute of Halle University. His work comprised more than 120 publications, including a dozen books, most of which appeared in several editions. Walther received honorary degrees from Perth and Melbourne (the latter 2 weeks after the outbreak of World War I!). In 1928, he was Visiting Professor at John Hopkins University, Baltimore. From 1924 to 1931, he served as President of the German Academy of Scientists, *Leopoldina*, in Halle. He died at Hofgastein, Austria, on 4 May 1937.

Figure 1 Johannes Walther.

Early in his career, Walther set himself the aim of 'modernizing' Lyell's uniformitarianism. Thus he undertook journeys to study as many geological conditions and environments as possible, but he focused on two fields in particular: shallow marine environments (including reefs) and deserts. At the time when he wrote his principal works, palaeontology, stratigraphy, and tectonics dominated geology in the universities, but Walther did not focus on these fields. Sedimentary rocks were chiefly studied because of their fossil content, useful for stratigraphy, and not because they illustrated former environments. Strongly influenced by Haeckel, the German apostle of Darwinism, Walther applied his ideas to geology by looking at the mode of formation of sediments, and the processes that formed rocks, not just their characteristics. His credo was: "Aus dem Sein erklären wir das Werden" (From the present state [of a rock] we explain its origin). But he was well aware that some past processes do not occur today. Though physical laws remain the same, geological conditions vary and may even be unique.

With this dynamic approach, Walther was able to integrate climatological, sedimentological, palaeontological, and other aspects into Amanz Gressly's 'facies' concept (1838), which was the basis for Walther's comprehensive consideration of facies. His view of the succession of palaeoenvironments was expressed in the 'Law of Facies', which was subsequently named after him (even though it had already been found by Gressly): "Es ist ein Grundsatz von weittragender Bedeutung, dass primaer sich nur solche Facies und Faciesbezirke geologisch überlagern können, die in der Gegenwart nebeneinander zu beobachten sind" ("It is a principle of far-reaching importance that only the facies or facies areas that are at present adjacent to one another can be geologically superimposed upon one another") (Walther [1894], p. 979; see also Middleton [1972]). In other words, the relative horizontal distribution of sediments with their organic content will be transformed into a vertical distribution, having a chronological order. Gressly found the rule during his extended fieldwork for his admirable monograph on the Swiss Jurassic near the town of Solothurn, where he carefully studied the facies changes. Fifty years later, Walther discovered Gressly's study of 1838 anew, when he was working for his volume on lithogenesis. He explained and discussed this idea in detail in the first of three chapters on facies in his most interesting third volume (*Lithogenesis of the Present*) of his fundamental work, *Einleitung in die Geologie als historische Wissenschaft* (1894). This discussion of the 'law' was only a minor part of the total corpus of his work, but it is on this that his present reputation chiefly rests.

Walther stressed the importance of organisms in geological processes and *vice versa*: biogeology. The dependence of biocenoses (groups of organisms living together, forming natural ecological units) on their substrates means that lithology should have priority over palaeontology. He mentioned, for example, that the empty shells of index fossils can be transported over long distances and may, therefore, give false stratigraphic results. With his zoological background of comparative anatomy Walther thus advocated 'comparative lithology'. This concept improved stratigraphy substantially and Amadeus W. Grabau (1870–1946) paid tribute to it by dedicating his classic *Principles of Stratigraphy* (1913) to Walther.

The first of the other two volumes of Walther's *Introduction to Geology as Historical Science* (see above), *Bionomie des Meeres*, was a treatise on marine biogeology that dealt with the interactions between the sea floor, fauna, and flora; beyond that it also had chapters on oceanography. (He had previously published a popular booklet on general oceanography in 1893.) The second volume was mostly a collection of faunal lists.

All Walther's pioneering work was undertaken in his years in Jena. One of his later publications was the voluminous *Allgemeine Paläontologie* (*General Palaeontology*) (1927). He wanted to give a summary of all the topics he had thought and taught about, and this was a lot. The book contained a wealth of varied ideas, though regrettably a number of them were already outdated by the time the book appeared. With its 809 pages, it presents difficult, yet still inspiring reading. For Walther, theories were much more important than details.

Special Contributions

Shallow Marine

Walther's final study on the sediments of Dove Bank (*Taubenbank*, 1910), a shoal in the Gulf of Naples, was a classic. He compared his first maps of 1884 with the latest findings of 1910 to evaluate the changes caused by volcanic activity (Vesuvius produced a great ash-fall in 1906) and the unusual storms of the intervening years (an early look at event stratigraphy!). His former studies of the rock-forming calcareous algae (1885) enabled him to compare his old and new results. He showed that coralline algae spread rapidly, consolidating sediments within 25 years.

Perhaps even more interesting was Walther's application of biological experiments to explain bioturbation, which he stated could occur down to 15 cm. He also measured the digging velocity of mussels. Using

a 'hyperactualist' method, he explained the formation of calcareous sands by keeping crayfish and mussels together in an aquarium and finding that four crayfish of 12–18 cm length worked a load of 580 g of mussels into detritus within 12 days. The final weight was 240 g.

A 'fossil example' was provided by his study of the fauna of the Solnhofen sediments (Upper Jurassic, Bavaria, South Germany) (1904), a contribution for a volume in honour of Haeckel's 70th birthday. Walther tried to show all features of a biotope by determining the frequency of marine and land fossils at various localities in the Solnhofen sediments. He noted the lack of freshwater and brackish species, and localized the presence of plant debris; by the decreasing frequency of occurrence of the planktonic crinoid *Saccocoma*, from a centre outwards in all directions, he concluded that the greatest depth of water had been in this central area. Besides his own observations, Walther used data from all available collections and fossil lists for his demonstration of the Solnhofen biotope. Unfortunately, this exemplary study was almost hidden among the zoological articles in the Haeckel volume and in consequence was overlooked by many geologists.

Reefs

During his early studies (1885) in the Gulf of Naples, Walther became interested in the growth of calcareous algae. He compared his marine findings with what could be found in Tertiary sediments in Sicily and with Triassic alpine rocks, and concluded that lithification takes place concomitantly with deposition, leading to the formation of structureless limestones.

A further step towards greater understanding of reefs was achieved as a result of his voyages to the south coast of Sinai (1886) and to the Palk Strait (India) (1888–1889). Walther also emphasized that reefs are traps for considerable amounts of sediments, which form up to 60% of the whole reef complex. Further, he noted the importance of tectonic movements, sea level changes, and topography for the reef growth (elevations are preferred places), laterally or vertically.

Deserts

The journey of 1886 along the Sinai coast, with a return along the Egyptian side of the Red Sea through the Galala Desert offered Walther splendid opportunities for the study of desert environments, which he was subsequently able to extend in the USA (1891), central Asia (1897), Egypt (1911), and Australia (1914). He soon recognized the importance of aeolian erosion, which previously had not been much taken into account. Consequently, his first publication on deserts had the provocative title *Die Denudation in der Wüste und ihre geologische Bedeutung* (1892) (*Desert Denudation and its Geological Importance*) and provoked many controversies. In fact, he tended to exaggerate the notion of desert erosion and neglected the influence of periods with strong fluviatile erosion.

Walther was the first to describe ventifacts and *Dreikanter*. The form of sand grains gave him indications of their aeolian origin, an approach that was systematically utilized by André Cailleux in the 1930s. Walther dealt with the phenomena of fossil deserts, which he classified as tropical, coastal, plantless volcanic, rain-shadow, and glacial deserts. The fact that the different types could be associated with younger or older sediments from other climatic zones was explained by changes in the Earth's axis of rotation. With this idea, he was well ahead of his time. In a special publication, Walther described the fauna of a lake in the *Buntsandstein* desert in Thuringia (1904). From his study of the Nubian Sandstone in Egypt he became convinced of the desert character of the *Buntsandstein*. His ideas about the formation of (minor) salt deposits are still accepted, along with Ochsenius's bar theory. Walther was the first to describe laterite profiles in the deserts of Western Australia, interpreting them as weathering products of a former climate (1915).

A revised and enlarged version of his first desert book was published in 1900 and went through four editions until 1924. The last edition has recently been translated into English with commentary (1997). This book was the main reason for his being invited to Australia by the British Association for the Advancement of Science in 1914. In Germany, he was called '*Wüsten* (desert) Walther'.

National Education

The nineteenth century was a golden age for national science education and Walther's teacher, Haeckel, was indefatigable (and extremely successful) in this field. Walther himself wrote a flowing, somewhat poetic, prose, which was very apt for such work. His books for a broader public enthused more than one generation and were popular with both school teachers and their students. The *Vorschule für Geologie* (1905) (*Elementary Course in Geology*), for example, sold 22 000 copies and was translated into Russian and Czech. Its last Russian edition appeared as late as 1940. Another successful book was the *Geologie von Deutschland* (1910). Walther saw an important task in the training of schoolteachers in geology and held many vacation courses for them.

However, his extensive work in this field affected his academic reputation in later years.

Impact

This may be one of the reasons why Walther's ideas were not fully appreciated by his contemporary fellow geologists. Another reason could be that he tended to go to extremes when proposing and defending a new idea. Walther was still young when he published his pioneering works and was then of low academic status. Many of his colleagues, especially in Germany, did not recognize or appreciate the views he introduced. They did not fit into the current academic trends. The recognition of his books in Austria and Switzerland was greater than at home. But some of the outstanding German geologists of the time (Wilhelm von Gümbel, Karl von Zittel, Hermann Credner, Edmund von Mojsisovics) appreciated his revolutionary steps in the direction of biogeology, as opposed to palaeontology or stratigraphy.

Walther had particular influence in Russia. His desert book was translated in 1911 and lithology became one of the main fields of Russian geology. In Britain, his reputation was mostly based on the desert book and was soon forgotten after his death. In the USA, only a few geologists, such as William Henry Twenhofel, mentioned his books. Walther's somewhat 'baroque' style of writing caused language difficulties. Also the World War I may have hindered the spread of his works and ideas. Interest in them was revived during the second half of the twentieth century by the facies research in the oil industry. After some delay, Walther eventually became better known in the USA than in Europe.

See Also

Fossil Plants: Calcareous Algae. **History of Geology From 1900 To 1962**. **Sedimentary Environments:** Depositional Systems and Facies; Carbonate Shorelines and Shelves; Deltas; Deserts; Reefs ('Build-Ups'). **Sedimentary Processes:** Aeolian Processes. **Stratigraphical Principles**.

Further Reading

Ginsburg RN, Gischler E, and Schlager W (eds.) (1994) *Johannes Walther on Reefs*. English translation with commentary. Miami: University of Miami, Rosenstiel School of Marine and Atmospheric Science (*Geological Milestones II*).

Gischler E and Glennie KW (eds.) (1997) *The Law of Desert Formation: Present and Past*. English translation, with preface and introduction, of Johannes Walther (1924). Miami: University of Miami, Rosenstiel School of Marine and Atmospheric Science (*Geological Milestones IV*).

Gressly A (1838) *Observations Géologique sur le Jura Solenrois. Nouvelles Mémoires de la Société Helvétiques des Sciences Naturelles*. Volume 2. Neuchâtel.

Middleton GV (1972) Johannes Walther's law of the cor relation of facies. *Bulletin of the Geological Society of America* 84: 979 988.

Seibold I (1992) *Der Weg zur Biogeologie: Johannes Walther 1860 1937*. Berlin, Heidelberg and New York: Springer.

Vissotzky WR (1965) *Johannes Walther and his Role in the Progress of Geology*. Moscow: Nauka (in Russian).

Walther J (1885) Die gesteinsbildenden Kalkalgen des Golfes von Neapel und die Entstehung structurloser Kalke. *Zeitschrift deutsch Geologische Gesellschaft* 37: 329 357.

Walther J (1888) Die Korallenriffe der Sinaihalbinsel: Geo logische und biologische Beobachtungen. *Abhandlungen der mathematisch physikalischen Classe der königlisch Sächsischen Gesellschaft der Wissenschaften zu Leipzig* 14: 435 506.

Walther J (1891) Die Adamsbrücke und die Korallenriffe der Palkstrasse: Sedimentstudien im tropischen Litoralge biet. *Petermanns Geographische Mitteilungen* 22: 40.

Walther J (1891) Die Denudation in der Wüste und ihre Geologische Bedeutung. *Abhandlungen der mathema tisch physikalischen Classe der königlisch Sächsischen Gesellschaft der Wissenschaften zu Leipzig* 16: 345 570.

Walther J (1893) *Allgemeine Meereskunde*. Leipzig: Weber.

Walther J (1893 1894) *Einleitung in die Geologie als his torische Wissenschaft*. 3 vols. Jena: Fischer. I. *Bionomie des Meeres: Beobachtungen über die marinen Lebensbe zirke und Existenzbedingungen*: 1 196; II. *Die Lebens weise der Meeresthiere: Beobachtungen über das Leben der geologisch wichtigen Thiere*: 200 531; III. *Lithogen esis der Gegenwart: Beobachtungen über die Bildung der Gesteine an der heutigen Erdoberfläche*: 535 1055.

Walther J (1900) *Das Gesetz der Wüstenbildung in Gegenwart und Vorzeit,* 1st edn. Berlin: Reimer.

Walther J (1904) *Die Fauna der Solnhofener Plattenkalke. Festschrift. 70. Geburtstag von Ernst Haeckel,* pp. 133 214. Jena: Fischer.

Walther J (1904) Über die Fauna eines Binnensees in der Buntsandsteinwüste. *Zentralblatt für Mineralogie, Geo logie und Paläontologie* un numbered volume (for 1904): 5 12.

Walther J (1905) *Vorschule der Geologie*. Jena: Fischer.

Walther J (1910) Die Sedimente der Taubenbank im Golfe von Neapel. *Abhhandlungen der königlich Preussischen Akademie der Wissenschaften, Physikalisch Mathema tische Classe* 3: 1 49.

Walther J (1910) *Lehrbuch der Geologie Deutschlands*. Leipzig: Quelle & Meyer.

Walther J (1915) Laterit in Westaustralien. *Zeitschrift der deutschen Geologischen Gesellschaft* 67: 113 132.

Walther J (1927) *Allgemeine Paläontologie. Geologische Fragen in biologischer Betrachtung*. Berlin: Borntraeger.

Wegener

B Fritscher, Munich University, Munich, Germany

Introduction

Alfred Wegener (Figure 1) was the 'father' of the theory of continental drift, one of the most influential theories in modern earth sciences. From a geophysical point of view, Wegener constructed a new picture of a dynamic Earth, postulating large-scale, ongoing horizontal movements of the continents, contrary to the theory of the permanence of continents and oceans that prevailed in the early twentieth century. First published in 1912, Wegener's theory had to wait for more than 50 years to become fully acknowledged, due in particular to the lack of a credible mechanism to explain, and direct empirical confirmation of, continental movement. The revival of Wegener's theory in the late 1950s came from studies of the ocean floors, an approach that Wegener had never considered. Though modern plate tectonics differs significantly from Wegener's original theory, there was nevertheless hardly any geological idea in the twentieth century that was subjected to greater scientific and public dispute than the idea of 'drifting' continents.

Meteorology and Polar Research

Born on 1 November 1880, Alfred Lothar Wegener studied astronomy, meteorology, and physics at Heidelberg, Innsbruck, and Berlin, earning a PhD in astronomy from the University of Berlin in 1905. Following his older brother Kurt Wegener (1878–1964), he was appointed an assistant at the Aeronautical Observatory at Lindenberg, near Berlin. There, he became acquainted with modern methods for the study of the higher atmosphere, including free balloon riding; together with his brother, he broke the world endurance record for balloon riding in 1906 by staying aloft for more than 52 h. That same year, Wegener joined a Danish expedition to Greenland under Ludvig Mylius-Erichsen (1872–1907); the goal was to map Greenland's north-east coast. Wegener became the first to use kites and tethered balloons to study the atmosphere in an Arctic climate. His Arctic research earned him a position at the University of Marburg, where he lectured on meteorology and practical astronomy from 1909. He published several papers on meteorological subjects, including a monograph on *The Thermodynamics of the Atmosphere* (1911), which became a standard textbook throughout Germany. By his studies on the chemical composition of the atmosphere, and its temperature distribution, Wegener pioneered the new science of aerology.

Wegener first presented his theory of continental drift to the public at the beginning of 1912. He found little time to reply to his numerous critics, for only a few month later he was in Greenland again. Together with the Dane, Captain Johan Peter Koch (1879–1928), Wegener became the first to winter on the icecap, and in the following spring, they undertook the longest crossing of the great ice sheet ever made up to that time. After his return from Greenland, Wegener married Else Köppen (1892–1992), daughter of Wladimir Köppen (1846–1940), a leading European meteorologist who became Wegener's lifelong mentor and collaborator.

In the summer of 1914, Wegener was drafted into the German army but was soon released from combat duty after being twice wounded. The fall of a meteorite in April 1916, near Marburg in Hesse,

Figure 1 Alfred Wegener in 1910. Reproduced from the Deutsches Museum, Munich.

turned his attention to these bodies and to the origin of the craters of the moon, which he thought were formed by impacts of bodies belonging to the solar system. In 1919, following his father-in-law, he became head of the department of theoretical meteorology at the German Marine Observatory at Grossborstel near Hamburg, and a 'professor extraordinary' (außerordentlicher professor) at the newly founded University of Hamburg. From 1924, Wegener held a professorship in meteorology and geophysics at the University of Graz. Soon he was preparing another expedition to Greenland, for a systematic study of the great icecap and its climate; the expedition departed in 1930. Wegener died at the beginning of November 1930, a day or two after his 50th birthday, while returning from a rescue expedition that brought food to a party of his colleagues camped in the middle of the Greenland icecap. His body was eventually recovered in May 1931.

A New Image of the Earth

Wegener's famous theory of continental drift was actually the work of just a few months. In the autumn of 1911, he became aware of a paper summarizing the evidence for the close relationship of the older fauna of South America and West Africa. He also remembered an earlier observation of the striking congruence of the coastlines on either side of the Atlantic Ocean; towards the end of the year, in 1911, he gave a preliminary account of his basic ideas in a letter to Wladimir Köppen. On 6 January 1912, Wegener presented his new theory at a meeting of the Geologische Vereinigung (Geological Society) in Frankfurt, and promptly published a preliminary paper on his ideas in *Petermann's Geographische Mitteilungen* (Figure 2).

For most earth scientists of Wegener's day, the hypotheses of wandering continents sounded rather fantastic, chiefly because of the lack of a satisfactory explanation of the moving forces. However, with the help of the German geologist Hans Cloos (1885–1951), Wegener extended his early paper and published his first book on his theory as *Die Entstehung der Kontinente und Ozeane* (*The Origin of Continents and Oceans*) in 1915, now also referring to some earlier (rather speculative) ideas on continental displacements by William Henry Pickering

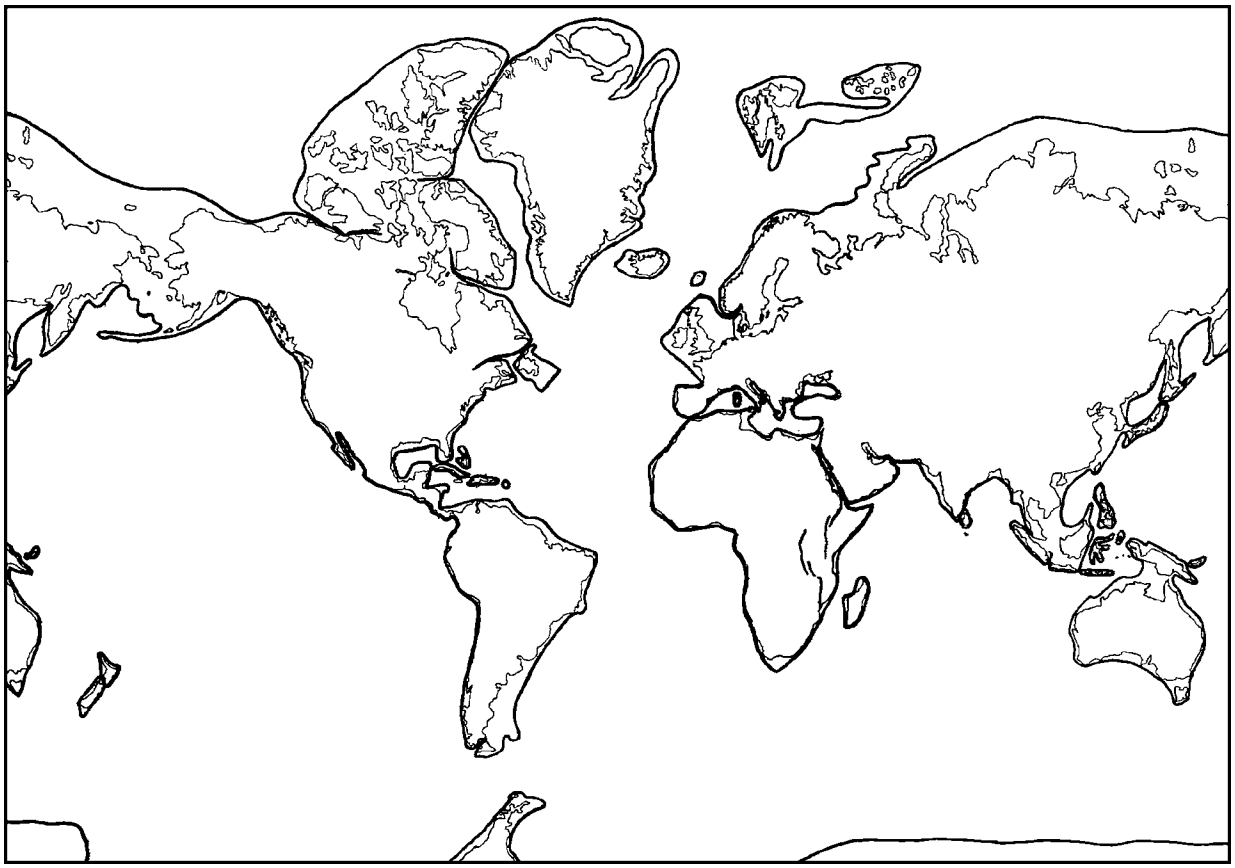


Figure 2 Map from Wegener's 1912 paper, *Die Entstehung der Kontinente*, showing the boundaries of the continental blocks (including the continental shelves).

(1858–1938) and Frank Bursley Taylor (1860–1938). After World War I, the second and third editions of his book, both revised and enlarged, were published (in 1920 and 1922, respectively). Palaeoclimatology was an essential empirical background of his theory, and, together with Köppen, Wegener published in 1924 a detailed discussion of the climates of the geological past (*Die Klimate der geologischen Vorzeit*). A fourth edition of *The Origin of Continents and Oceans* followed in 1929. From 1924 onwards, there were also translations (of the third edition) into English, French, and other languages, by which the theory became known internationally.

Wegener's original intention had been to give a genetic explanation of the large-scale features of Earth's surface (the continental blocks and the ocean basins) according to a single comprehensive principle. This principle of horizontal mobility, i.e., of the splitting off and drifting apart of continental blocks, did not emerge from new experimental research, but, rather, from a 'rearrangement' of known geophysical and geological information. This was set against two widely accepted and closely related assumptions in the earth sciences around 1900, namely, the hypotheses of former (now sunken) land bridges between continents (postulated due to striking geological and palaeontological conformities) and the theory of earth contraction due to its general cooling, advocated by leading geologists such as Eduard Suess (1831–1914) and Albert Heim (1849–1937).

Contraction would not have been sufficient to account for the large folds of Earth's crust. The great arching forces required to transmit the full shrinkage of a whole great-circle to one point of it have been proved to be physically impossible. The idea of a shrinking Earth, commonly illustrated by the simile of a drying apple, contradicted gravity measurements, i.e., the doctrine of isostasy (the rocks under the oceans are denser than are those under the continents, so altitudinal differences are compensated and equilibrium of pressure or 'isostasy' prevails; accordingly, the less dense continental blocks may be thought of as 'swimming' on the underlying mass, like an iceberg floating in the sea).

Wegener referred to Suess' distinction of the 'Sial' (silica/alumina-rich) and 'Sima' (silica/magnesia-rich) layers, emphasizing their different densities (2.5–2.7 for the Sial, or the continental blocks, and about 3.0 for the Sima of the ocean floors) and their different melting points (with the Sial's melting point being 200–300°C above that of the Sima). Concerning the temperature increase towards Earth's interior, the difference might not be sufficient to justify the assumption that solid Sialic blocks 'swim' in a fluid Sima (**Figure 3**). Rather, both layers had to be thought of as being viscous, i.e., plastic (with greater plasticity for the Sima). The characteristics of such viscous fluids are paradoxical in that the duration of the action of the forces determines whether the body behaves as a solid or a fluid. Consequently, within geological time, large horizontal displacements of the

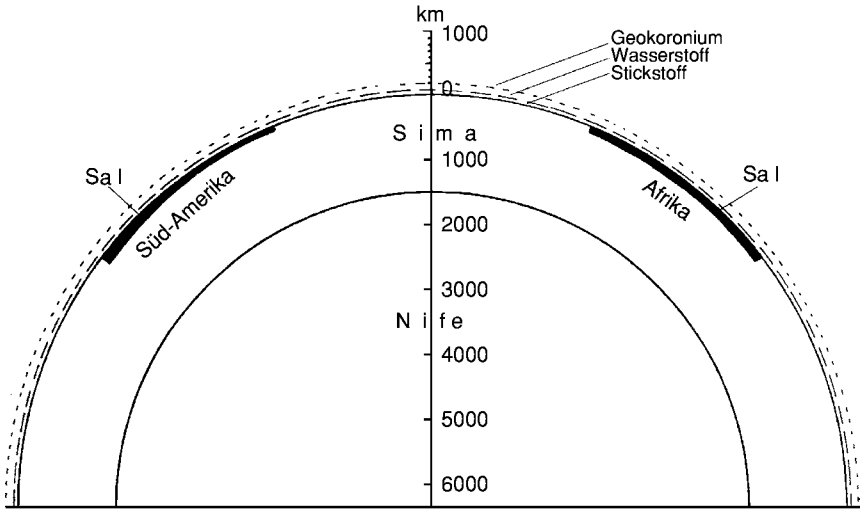


Figure 3 Cross section along a great circle through South America and Africa, from Wegener's 1912 paper, *Die Entstehung der Kontinente*, showing the 'swimming' continents within the Sima (silica/magnesia rich region). Also represented are the Nife (nickel/iron rich core of the Earth), and the atmospheric layers of nitrogen (Stickstoff), hydrogen (Wasserstoff), and 'geokoronium' (a hypothetical gas, which Wegener had introduced to account for the typical green spectrum line of the aurora borealis) (figure in correct scale).

continents would be possible, though the motion would be very slow.

Mountain Formation, Volcanism, and Rift Valleys

Wegener first discussed the origin of continents and oceans and the formation of folded mountain ranges. Relating to the nappe-fault structures of the Alps, where the original areas that are now thrust up as mountains were apparently up to four or five times wider than at present, Wegener thought that mountain formation was a unilateral, irreversible process: each pressure brings about an increase of the thickness and a shortening of the surface, while, on the other hand, tension leads to splitting of the continental blocks. The individual stages of perceived as mountain formation comprised continual processes of splitting and compression, whereby the original Sialic crust (for which Wegener assumed a thickness of about 30–35 km) gradually decreased in surface area, split into separate pieces, and increased in thickness. Along with the movement of the continental blocks, a hypothesized universal ocean ('Panthalassa') began to divide into a shallow sea and a deep sea.

Volcanism, for Wegener, was mainly related to the continental 'fronts'. Areas where tension prevailed, such as the Atlantic Ocean, and also opening faults, seemed to be relatively poor in volcanoes as compared with areas such as the Pacific Ocean, where pressure was increasing. The fronts of moving blocks made conditions more favorable to volcanism than did the 'backs'. Nevertheless, Wegener wondered whether the mid-Atlantic ridge might be considered as a zone where, with the continuing expansion of the Atlantic, the floor was continuously breaking up, making room for fresh, relatively fluid and high-temperature Sima from below! Moreover, increased volcanic activity in some periods of Earth history might be due to large displacements (as, for instance, during the Tertiary).

Trench faults (Grabenbrüche), i.e., rift valleys, acquired new meaning as representing the beginnings of new continental separations. Gravity measurements had shown that beneath such lines lay material of greater density, compared to that on either side. Therefore, these lines could be seen as incipient fissures within the continental blocks (into which the denser Sima was rising according to the principle of isostasy). The best examples of such separations were provided by the East African trenches and their continuation through the Red Sea. At the majority of the trenches, the measurable mass deficit was not compensated by greater density of the matter beneath it. Thus, the trenches must be youthful disruptions of a continental block.

Wegener's theory of mountain formation was further supported by the fact that the folding of the Andes seems to have been essentially simultaneous to the opening of the Atlantic Ocean. The American blocks, during their westward drifting, had encountered resistance at the presumably very old and relatively rigid floor of the Pacific Ocean. Thus, the extended shelf, with its mighty sediments, forming the western border of the continental block, was compressed to a range of fold mountains. For the Tertiary folds of the Himalayas, Wegener assumed that lower India had formed an extended peninsula prior to compression, the southern end of which lay next to that of South Africa. The folds had been produced by 'impact' of the Indian subcontinent and the main mass of Asia.

Geological and Palaeontological Evidence

The palaeontological evidence indicating a former connection between the organic components of different continents had already given rise to the doctrine of former land bridges. Among the most striking findings were the distributions of the *Glossopteris* flora on the southern continents and the occurrence of *Mesosaurus* at the turn of the Permian and the Carboniferous exclusively in south-eastern South America and the western parts of Africa; both of these discoveries suggested a former connection of the two continents. Using these relationships also allowed calculations of when the continents were separated (either by horizontal displacements or by sinking of the land bridges). South America and Africa had been connected during the Mesozoic, but were separated at the end of the Eocene or Early Oligocene. The connection between Europe and North America seemed to have been maintained during the older Tertiary period, but separation occurred in the Miocene, although it might have continued in the far north (over Scandinavia and Greenland) into the Pleistocene. The connection of Lower India with southern Africa, which Wegener had postulated based on his ideas on the formation of the Himalayan range, was also confirmed by palaeontological evidence. Zoogeographers had long assumed a former elongated Indian–Madagascan peninsula (called 'Lemuria'), separated from the African block by the Mozambique Channel.

The zoogeographic concept of Lemuria had given rise to Suess' notion of a great southern continent, Gondwana, comprising parts of South America, Africa, Lower India, Australia, and Antarctica. Assuming the unchanged positions of its present-day

relics, however, required ascribing a huge extent to this continent. Wegener, by contrast, proposed a much reduced primeval continent, Pangaea. In the Permian, i.e., until some 300 Ma ago, all the continents were supposedly joined in one land mass extending from pole to pole. During the Triassic, about 200 Ma ago, Pangaea began to break up and the newly emerging continents started moving into their current positions. In the Jurassic, there were few remaining connections except at the northern and southern ends. Just as northern Europe and North America remained connected until the older Tertiary period, a connection of the southern continents seems to have persisted, running from the southern coast of Australia over Antarctica to South America. Later, the Antarctic block, like the South American block in the Tertiary, moved over from South Africa towards the side of the Pacific Ocean. Only in the Quaternary period, then, did the Australian block become detached (**Figure 4**).

For geological and tectonic evidence, Wegener referred particularly to Suess' *magnum opus*, published in three volumes during 1885–1909, *Das Antlitz der Erde* (*The Face of the Earth*). Considering the tectonic relations, Europe/Africa and both Americas seemed to represent the edges of an immense expanded fissure. In the north, for instance, the Greenland massif was matched by Scandinavia, both consisting of gneiss, and the less mountainous North America corresponded to the likewise less mountainous Europe. The most striking example, however, was the Carboniferous mountain range, called the Armorican mountains (Suess' 'transatlantic Altaides'), which made the coalfields of North America appear to be the direct continuation of the European ones.

Wegener's theory of mountain formation was also confirmed by remarkable differences between the Atlantic and the Pacific hemispheres, such as the distinction between Pacific and Atlantic types of coasts (marginal chains and ocean trenches in front

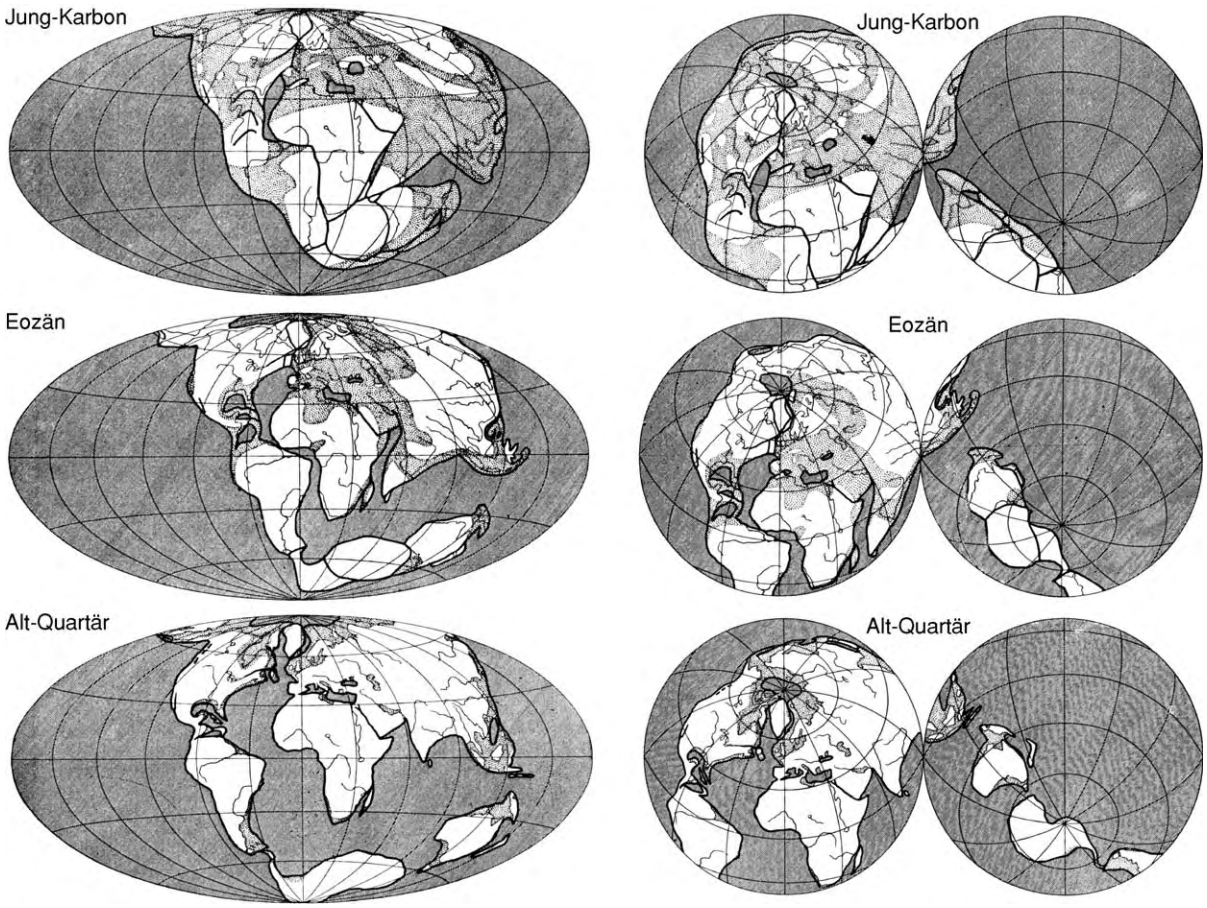


Figure 4 Wegener's reconstruction of the separation of the continents from the primeval Pangaea, from his 1926 paper *Paläogeographische Darstellung der Theorie der Kontinentalverschiebungen*, showing the relative positions of the continents during the Upper Carboniferous (Jung Karbon), Eocene (Eozan), and Lower Quaternary (Alt Quartar) (in two different projections). Cross hatching represents deep seas, dotted regions represent shallow seas; rivers, recent coastlines, and outlines are shown only for orientation.

of the Pacific coasts, as contrasted to the wild, irregular 'ria' Atlantic coastlines). There were also differences in the volcanic lavas of the two hemispheres, as emphasized by the Vienna petrographer Friedrich Becke (1855–1931) and others. The Atlantic lavas contained a greater proportion of sodium, whereas calcium and magnesium prevailed in the Pacific lavas. Such differences were intelligible according to the assumptions of continental movements. The opening of the Atlantic was matched by the general pressing of the continents against the region of the Pacific Ocean: pressure and compression prevailed at the coasts of the latter whereas tension and splitting occurred at the latter.

Palaeoclimatology

Traces of glaciation during the Permian (ground moraines lying on scratched bedrock) were to be found on the southern continents, e.g., in East India and Australia. If the present-day arrangement of the land masses had prevailed at that time, this 'Permian ice age' would have required an icecap of seemingly impossible size. And the north pole would have been in Mexico, where no trace of glaciation during that period was recorded. Following the idea of horizontal displacements, however, all regions subjected to glaciation came together concentric to the southern margin of Africa. And one had only to place the south pole in this much reduced glaciated area to give the Permian ice age a much more plausible form.

Wegener had discussed these palaeoclimatological features since 1912. In 1924, he gave a detailed description of the climatological changes from the Carboniferous through to recent times, following the traces of glaciations, swamps, and deserts, i.e., moraines, coal, salt, and gypsum, throughout Earth's history (Figure 5). In reconstructing the respective polar shifts, Wegener emphasized that they obviously took place along with the great displacements of the continental blocks. In particular, there was temporal coincidence of the best confirmed polar shift, in the Tertiary, and the opening of the Atlantic (Figure 6). Movement of the poles since the Pleistocene might also be related to the final separations of the continents in the north and the south.

Motive Forces

Wegener was very cautious about the forces that might have caused continental displacements. First, it was necessary to demonstrate the reality and the manner of the displacements before indulging in the hope of finding their cause. Nevertheless, he tentatively suggested two candidates: centrifugal forces caused by the rotation of Earth and tidal-type waves within Earth, generated by the gravitational pull of the sun and the moon. In the 1929 revision of Wegener's theory in

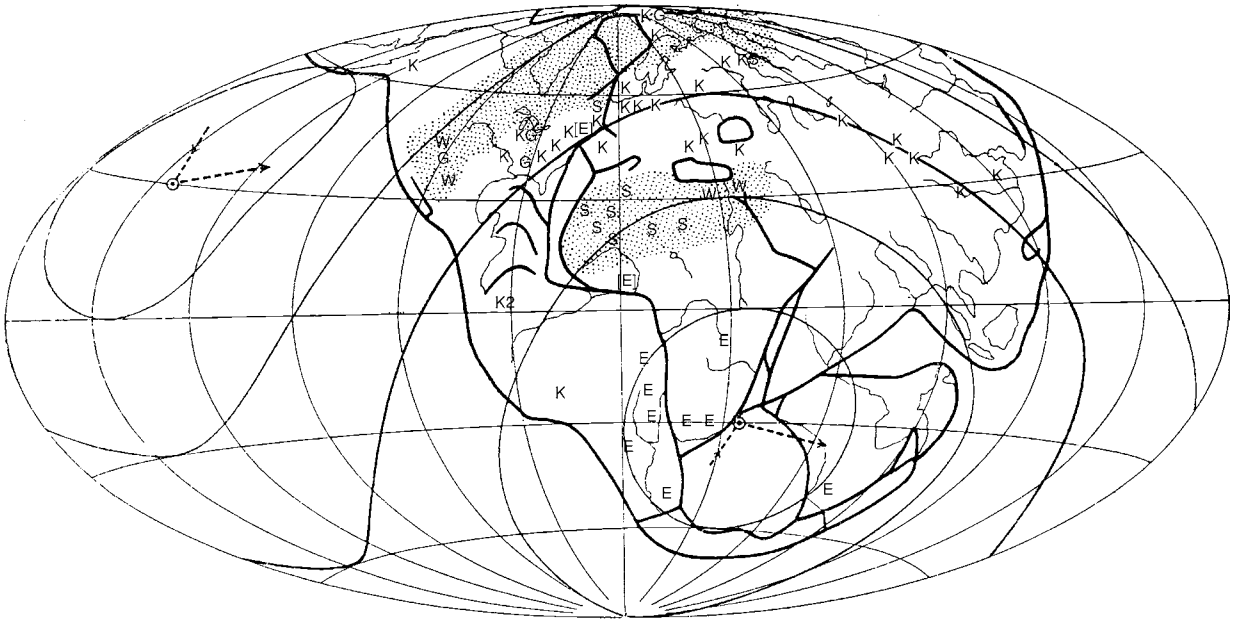


Figure 5 Wegener thought continental drift was the key to the climatic changes during Earth's history. This map, published in the 1924 book by Koppen and Wegener, *Die Klimate der geologischen Vorzeit*, shows traces of glaciation, swamps, and deserts for the Carboniferous. E, Traces of glaciation; K, coal; S, salt; G, gypsum; W, desert sandstone. Dotted regions indicate arid areas, dashed lines indicate the positions (i.e., the pathways) of the poles, and the bold curved line indicates the respective position of the equator.

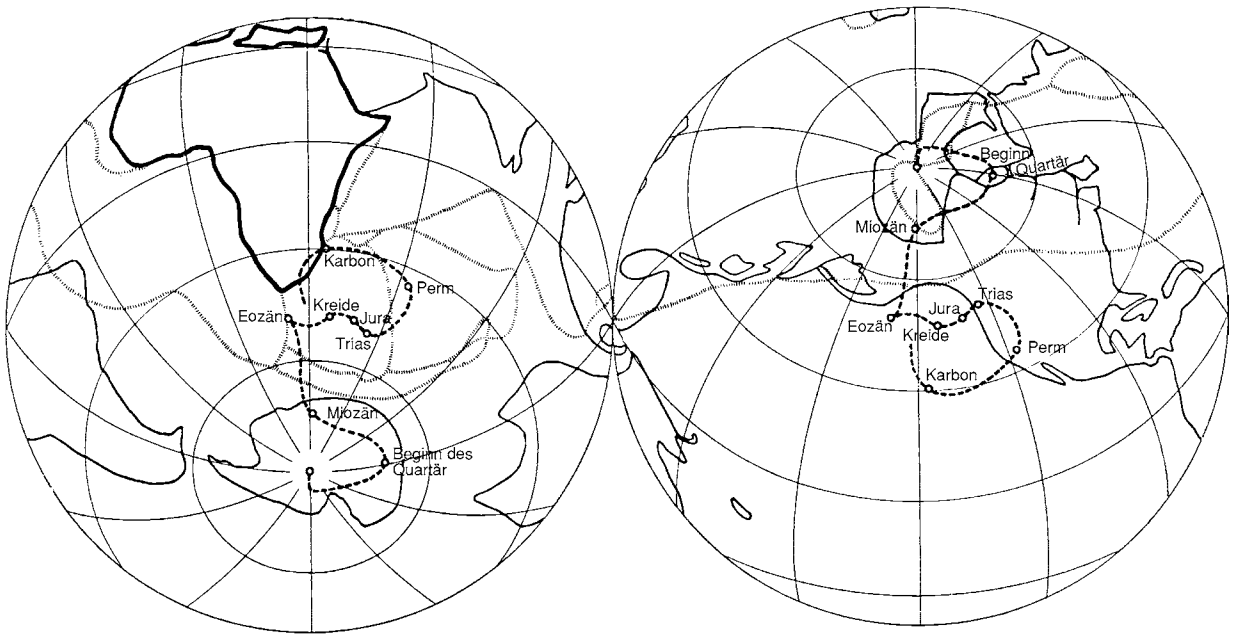


Figure 6 Map published in the 1924 book by Koppen and Wegener, *Die Klimate der geologischen Vorzeit*, showing polar shifts (dashed lines) from the Carboniferous to recent, related to the African table (left, south pole; right, north pole). Bold lines outline the continental blocks; hatched lines represent the Carboniferous (Karbon) period. Perm, Permian; Jura, Jurassic; Trias, Triassic; Kreide, Cretaceous; Eozan, Eocene; Miozan, Miocene; Beginn des Quartar, beginning of the Quaternary.

Die Klimate, he also mentioned convection currents within the Sima; these had been first discussed as a cause of mountain formation by the Vienna geologist Otto Ampferer (1875–1947) in 1906.

Wegener also endeavoured to calculate the recent velocity of the relative motion of the continents, though he was well aware that these values must be quite uncertain. In his 1912 paper, comparing various longitude determinations for Greenland, he had deduced an increase of the distance to Europe of 11 m year ^{1}. Referring to the lengths of transatlantic cables, he suggested that North America was drifting away from Europe at about 4 m year ^{1}.

From Continental Drift to Plate Tectonics

The theory of continental drift was long rejected by the majority of geologists. Among Wegener's few followers were the South African Alexander Du Toit (1878–1948), for whom continental drift provided the best explanation of the close similarities between the strata and fossils of Africa and South America, and the Swiss geologist Émile Argand (1879–1940), who saw continental collisions as the only means of producing the folded and buckled strata he had observed in the Alps (*see* **Famous Geologists:** Du Toit). Nevertheless, Wegener's explanation of the Permo-Carboniferous ice age impressed even his critics.

Wegener's reputation as a meteorologist and a polar explorer contributed to keeping his theory alive. His work was immediately remembered when, around 1960, surprising data were obtained from the ocean floor: palaeomagnetic patterns alongside the mid-ocean ridges clearly suggested the spreading of the seafloor. Within about two decades, Wegener's principle of horizontal displacements of parts of Earth's crust became almost universally accepted, although, ironically, the process still lacked a consensus as to its causes, though convection currents in the internal mantle are most commonly advocated.

It should be noted that Wegener's original concept differed from modern plate tectonics in essential points, particularly with regard to the Sial and the Sima. According to modern theory, the (Sialic) continents do not 'plough' through the (oceanic) Sima. Instead, both continents and ocean floor are regarded as forming solid plates, 'floating' on the asthenosphere, which, due to tremendous heat and pressure, behaves like an extremely viscous liquid (as Wegener had thought the Sima did). Therefore, the older term 'continental drift', still often used today, is not quite appropriate for the modern concept. Notwithstanding these differences, Wegener's basic ideas remain sound, and the lines of evidence that he used to support his theory are still valid. He first envisaged a dynamic Earth, connecting its major features and various geological processes – continental

movements, folded mountain ranges, rift systems, earthquakes, volcanism, ocean transgressions, palaeoclimatological changes, etc. – on a global scale. In this sense, Wegener's theory was a true forerunner of plate tectonics.

See Also

Africa: Rift Valley. **Famous Geologists:** Du Toit; Suess. **Gondwanaland and Gondwana**. **History of Geology From 1900 To 1962**. **History of Geology Since 1962**. **Palaeoclimates**. **Pangaea**. **Plate Tectonics**. **Tectonics:** Mid-Ocean Ridges; Mountain Building and Orogeny.

Further Reading

Carozzi AV (1985) The reaction of continental Europe to Wegener's theory of continental drift. *Earth Sciences History* 4: 122 137.

Fritscher B (2002) Alfred Wegener's 'The origin of continents, 1912'. *Episodes* 25: 100 106.

Jacoby WR (2001) Translation of 'Die Entstehung der Kontinente, Dr Alfred Wegener, *Petermann's Geographische Mitteilungen*, 58 (1912)'. *Journal of Geodynamics* 32: 29 63.

Köppen V and Wegener A (1924) *Die Klimate der geologischen Vorzeit*. Berlin: Bornträger.

Lüdecke C (1994) Stratigraphische Methode der Rekonstruktion von Expeditionsergebnissen am Beispiel des Todes von Alfred Wegener während der Grönlandexpedition (1930 31). In: Fritscher B and Brey G (eds.) *Cosmographica et Geographica: Festschrift für Heribert M. Nobis zum 70. Geburtstag, Algorismus,* vol. 13, pp. 347 367. Munich: Institut für Geschichte der Naturwissenschaften.

Oreskes N (1999) *The Rejection of Continental Drift: Theory and Method in American Earth Science*. New York and Oxford: Oxford University Press.

Runcorn SK (ed.) (1966) *Continental Drift*. New York and London: Academic Press.

Schwarzbach M (1986) *Alfred Wegener: The Father of Continental Drift*. Madison. WI: Science Tech Publications.

Sengör AMC (1991) Timing of orogenic events: a persistent geological controversy. In: Müller DW, McKenzie JA, and Weissert H (eds.) *Controversies in Modern Geology: Evolution of Geological Theories in Sedimentology, Earth History and Tectonics,* pp. 403 473. London: Academic Press.

Wegener A (1912) Die Entstehung der Kontinente. *Petermann's Mitteilungen aus Justus Perthes' Geographischer Anstalt* 58: 185 195, 253 256, 305 309.

Wegener A (1926) Paläogeographische Darstellung der Theorie der Kontinentalverschiebungen. In: Dacqué E (ed.) *Paläogeographie,* pp. 171 189. Leipzig and Wien: F Deuticke.

Wegener A (1971) *The Origin of Continents and Oceans*. (Translation from the 4th revised German edition by J Biram, with an introduction by BC King.) London: Methuen.

Wegener A (1980) *Die Entstehung der Kontinente und Ozeane*. (Reprint of the 1st and 4th editions, edited by A Vogel.) Braunschweig: Vieweg.

Wegener E (1960) *Alfred Wegener: Tagebücher, Briefe, Erinnerungen*. Wiesbaden: Brockhaus.

Wutzke U (1998) *Kommentiertes Verzeichnis der schriftlichen Dokumente seines Lebens und Wirkens,* Berichte zur Polarforschung 288. Bremerhaven: Alfred Wegener Institut für Polar und Meeresforschung.

FLUID INCLUSIONS

A H Rankin, Kingston University, Kingston-upon-Thames, UK

Introduction

Fluid inclusions are small droplets of fluid that have been trapped within crystals either during primary growth from solution or at some later stage, usually as a result of recrystallization along healed microfractures. They are ubiquitous in both naturally occurring minerals and in laboratory-grown crystals. To the chemist or materials scientist, these gross defects cause endless obstacles in their quest to grow near-perfect crystals. However, to the geologist, they provide a unique fossil record of the various fluids responsible for the formation and evolution of rocks and minerals throughout the history of the Earth.

Despite their small size (usually less than 20 μm), their chemical composition and physical properties can be readily determined, and the data may be used to estimate the temperatures, pressures, and physicochemical nature of the fluid at the time of trapping. This information has made an immense contribution to the development of modern theories of ore genesis, petrogenesis, diagenesis, and petroleum migration and accumulation, and to our understanding of the importance of the fluid phase in a wide range of geological processes.

Occurrence and General Characteristics

Formation and Genetic Classification of Fluid Inclusions

Small changes in the chemical or physical properties of fluids near to a growing crystal face can lead to perturbations in the stability of crystal growth and the development of gross defects, manifested as embayments, along crystal faces. These embayments will seal over during a period of greater stability, trapping a portion of fluids to form 'primary' (P) fluid inclusions. In many instances, the trapped fluid will be 'homogeneous' at the time of trapping. In others, where immiscible fluids are present or where mechanical entrapment of other coexisting crystalline phases has occurred, trapping will be 'heterogeneous'.

At some stage after primary growth, 'secondary' (S) fluid inclusions can form from later fluids, particularly as a result of recrystallization along microfractures. The chemical and physical properties of these inclusions may be very different from those of the earlier mineral-forming fluids. However, if fracturing and rehealing take place during primary growth, the fluids may be indistinguishable, and the terms 'pseudosecondary' or 'primary–secondary' (PS) appropriately describe such inclusions. A schematic representation of this genetic classification of inclusions is shown in Figure 1.

For most geological applications, it is necessary to establish whether the inclusions are primary, secondary, or pseudosecondary, and also whether heterogeneous trapping has occurred. Heterogeneous trapping may be recognized by the variable proportions of liquids and solids in a single group or generation of inclusions. Various criteria may be used to distinguish between P, PS, and S inclusions, but these may be difficult to apply and it may be difficult to identify primary inclusions in many samples.

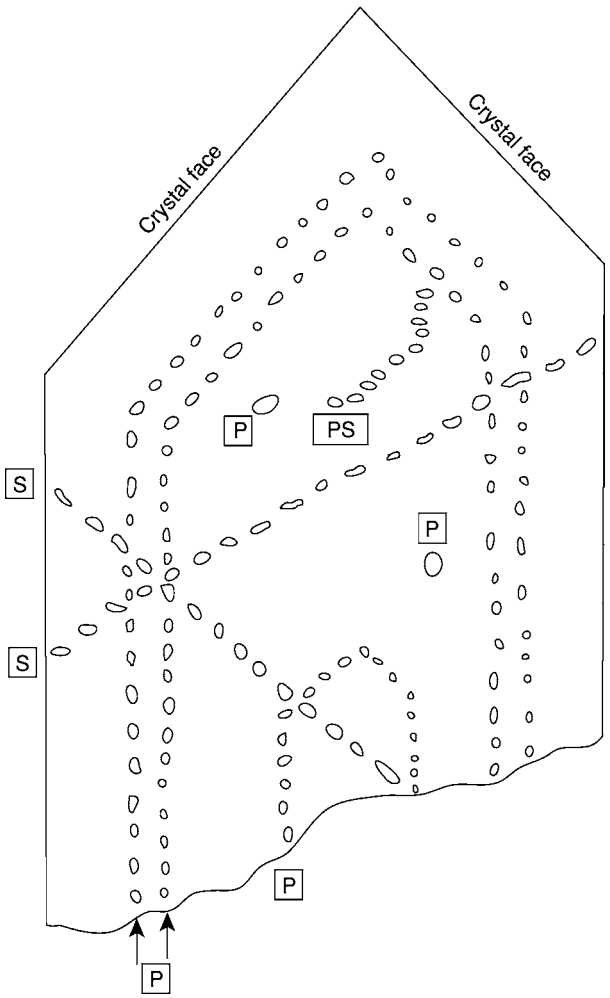


Figure 1 Schematic representation of the distribution of primary (P), secondary (S), and pseudosecondary (PS) fluid inclusions in a quartz crystal. Modified from Rankin AH (1989) Fluid inclusions. *Geology Today* 5: 21–24.

Choice of Material for Study

The successful application of fluid inclusion studies depends partly on serendipity and partly on the type and quality of material available for study. Due to their small size, observations on fluid inclusions are carried out under a microscope using polished wafers around 1–2 mm thick. In most cases, clear, transparent minerals are needed, but it is also possible to study inclusions in some deeply coloured, semi-transparent minerals in very thin ($<50\,\mu$m) polished sections. Care must be taken with soft, easily cleaved minerals, such as calcite and fluorite, because of the possibility of leakage during sample preparation or analysis. Fluid inclusions are particularly difficult to observe in turbid or translucent minerals, such as feldspar. Quartz is usually the preferred host.

Size and Shape of Inclusions

Fluid inclusions seldom exceed 1 mm in size; most are less than $20\,\mu$m, and those greater than 1 cm are exceptionally rare and usually regarded as museum specimens. Fluid inclusions display a variety of shapes. They may be flattened and irregular, rounded, or regular with three-dimensional 'negative crystal' shapes mimicking the crystal symmetry of the host crystal (Figure 2).

Phases Present at Room Temperature

Fluid inclusions contain varying proportions of liquid (L), solid (S) and gas (G) depending on the composition (X), temperature (T), pressure (P), and volume (V) of the enclosed fluid at the time of entrapment.

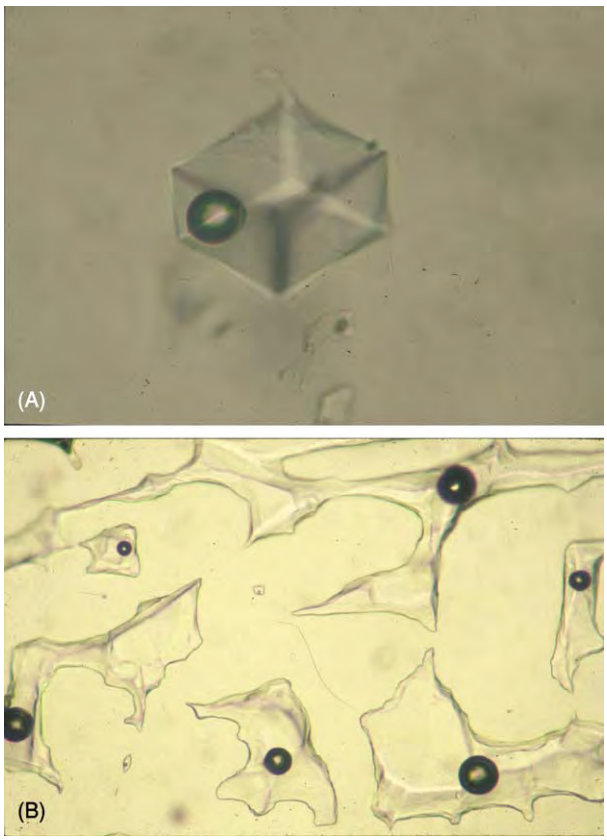


Figure 2 Contrasting shapes of two phase, liquid gas (L + G) aqueous inclusions in fluorite. (A) Negative crystal cavity mimicking the cubic symmetry of the host mineral. (B) Group of irregularly shaped inclusions showing evidence of necking down.

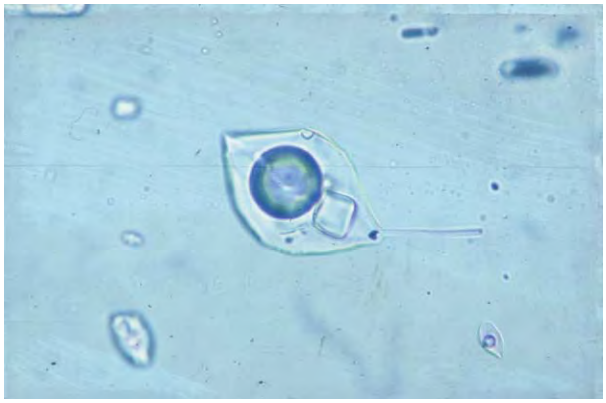


Figure 3 Three phase, liquid gas solid (L + G + S) aqueous inclusion in topaz containing a cubic halite daughter mineral and several smaller, unidentified daughter or captive minerals.

Figure 4 Three phase, liquid liquid gas (L + L + G) CO_2–H_2O inclusion with the characteristic 'double bubble' comprising an outer rim of liquid CO_2 enclosing a bubble of CO_2 gas.

The nature of these phases will vary according to the geological conditions during formation.

At room temperature, inclusions in magmatic or hydrothermal minerals usually contain gas bubbles in addition to entrapped liquids (e.g., aqueous brines and silicate melts). In aqueous inclusions, the bubbles are mobile and often move in response to small temperature changes, e.g., through heat accompanying an increase in the intensity of microscope illumination. In silicate melt inclusions, the bubbles are immobile, either locked within a quenched silicate glass or squashed in bundles of crystals and crystallites. H. C. Sorby, often regarded as the 'father of modern petrography', explained the origin of these bubbles in his classic paper of 1858, which laid the foundations for the widespread application of fluid inclusions as geothermometers and geobarometers. He demonstrated that the gas bubbles in two-phase, gas–liquid inclusions were 'shrinkage bubbles' resulting from the differential thermal contraction of the contained liquids and host crystal on cooling. A simple analogue is the mercury-in-glass thermometer in which the headspace represents the shrinkage bubble.

Any dissolved salts that reach their saturation point on cooling from the trapping temperature will crystallize *in situ* as daughter minerals in aqueous inclusions (Figure 3). These can sometimes be identified on the basis of their optical and crystallographic properties and behaviour on heating and cooling. The most common daughter minerals are halite (NaCl), sylvite (KCl), nahcolite ($NaHCO_3$), gypsum/anhydrite ($CaSO_4$), and calcite ($CaCO_3$). Dissolved gases, such as CO_2, CH_4, and N_2, will also partition preferentially on cooling into the shrinkage bubble. In gas-rich aqueous inclusions, a 'double bubble' containing CO_2-rich liquid and CO_2-rich gas is sometimes apparent at temperatures below the critical point of CO_2 at 31.1°C (Figure 4).

The phases present in silicate melts include bubbles and daughter crystals, but this will vary depending on the cooling history and whether the host rock is plutonic or volcanic/subvolcanic. They may be glassy or crystalline and vapour bubbles may or may not be present.

Figure 5 Aqueous inclusion containing two immiscible drop lets of brown coloured oil in fluorite. Note the presence of a gas bubble in the larger oil droplet.

Heterogeneous trapping of coexisting liquids, melts, and volatiles is also evident in some geological environments. For example, immiscible droplets of oil- or methane-rich vapour may be trapped in aqueous inclusions from low-temperature sedimentary/diagenetic environments (Figure 5). Recognition of these immiscible assemblages provides important information on magmatic differentiation processes, such as the exsolution of brines from silicate melts in granitic rocks, and silicate–carbonate immiscibility in alkaline and carbonatitic environments. In hydrothermal environments, the coexistence of low-density gas-rich and high-density aqueous inclusions in hydrothermal minerals can be used to infer boiling conditions in a hydrothermal system.

Heterogeneous trapping of coexisting solid and liquid can also lead to the development of inclusions containing varying proportions of captive rather than true daughter minerals.

Geological Applications – Information Available

Prerequisites and Assumptions

There are a number of potential pitfalls in the application of fluid inclusion studies to geological problems.

Is the material trapped within an inclusion a representative sample? For most applications, the assumption that the fluid was trapped homogeneously is necessary. Although this is not always the case, it is often possible to infer homogenous trapping from the consistency of the liquid, solid, and gas phase proportions in a given generation of inclusions. Varying proportions will indicate either heterogeneous trapping, multiple fluid generations, or necking down.

A more fundamental issue is whether the compositional differences known to occur within the submicroscopic boundary layer between the crystal and growth solution are significant. Under most geological growth conditions, this layer is generally assumed to be small (less than 1 μm) and the effect is only likely to be significant in very small inclusions, less than a few micrometres, which are too small to study anyway. A rule of thumb is that larger inclusions are much less likely to suffer from these boundary layer effects. This is backed up by evidence from studies of synthetic inclusions, where compositions usually match those in the parent growth solutions.

Have components been lost or added since entrapment? There has been much debate in the past as to whether fluid inclusions have retained their compositional integrity over geological time. The consensus, based on laboratory studies of natural and synthetic inclusions, is that in many situations bulk leakage is unlikely. In others, it can often be recognized or avoided. Bulk leakage and decrepitation (the explosive release of inclusion contents usually as a result of rapid heating) are relatively easy to recognize in natural samples, either because the inclusions appear 'empty' or are surrounded by characteristic decrepitation haloes of smaller inclusions. In soft, easily cleaved minerals, such as halite and calcite, the inclusions are more susceptible to leakage. It is partly for this reason that fluid inclusion researchers usually prefer to study inclusions in more resilient minerals, such as quartz. Bulk leakage and decrepitation are also likely when pre-existing, low-temperature inclusions are subjected to higher temperatures (and pressures), e.g., in low-temperature mineral veins or evaporite minerals in the contact zone of igneous intrusions. Experimental and theoretical studies have shown that, under high thermal pressure and chemical gradients, components such as H_2O and H^+ may move in or out of inclusions via diffusion rather than bulk leakage. Experts have argued that the high gradients required to maintain an effective rate of diffusion are insufficient over the time-scales involved in many geological processes, except under extreme P–T conditions, e.g., during high-grade metamorphism.

Has the volume of the inclusion remained constant since trapping; what are the effects of recrystallization? This question is particularly important if fluid inclusions are to be used as effective geothermometers and geobarometers. Irregularly shaped fluid and melt inclusions have a tendency to change towards energetically more stable shapes, such as negative crystals or spheres, via a process of internal recrystallization

of the inclusion walls. This usually occurs without any change in volume.

Crystallization of material on the walls of inclusions in hydrothermal and metamorphic minerals after trapping is usually considered to be insignificant because of the low solubilities of host minerals, such as quartz, in aqueous or CO_2-rich fluids. In highly water-soluble minerals, such as halite, or in melt inclusions formed in slow cooling plutonic environments, significant crystallization may occur on the walls of the cavity. Although there will be some apparent change in volume after trapping, the situation can be reversed by heating the sample at an appropriate rate in the laboratory.

The internal recrystallization processes responsible for changing the shape of large irregular inclusions can also cause the inclusion to split up or 'neck down' into a number of smaller inclusions. This is not usually a problem if necking down has occurred before the development of vapour bubbles or daughter phases in the inclusions, because each necked-down inclusion will still retain the original *P–V–T–X* properties at the time of trapping. Problems arise when necking down has taken place after heterogenization, in which case inclusions with variable phase proportions can develop (Figures 2B and 6).

Can a particular fluid inclusion assemblage be linked to a specific stage of crystal development or geological process? If primary (or pseudosecondary) inclusions are present, given the caveats above, they should represent the primary crystallization conditions and processes responsible for the formation of the host mineral. If only secondary inclusions are present, all that can be inferred are the *P–V–T–X* conditions during some recrystallization processes. Several generations of secondary inclusions, each representing a different stage in the evolutionary history of a rock or mineral sample, may be present, and careful petrographical examination is required to distinguish between them. Cathodoluminescence studies of quartz or carbonate hosts can sometimes be used to establish relative fluid inclusion chronologies. With care and patience, it is sometimes possible to link them to particular stages of crystallization or fluid activity (Figure 7).

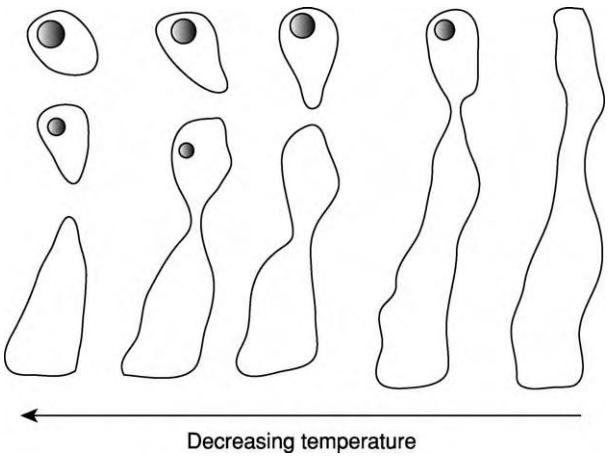


Figure 6 Schematic representation of necking down processes in aqueous fluid inclusions after trapping, showing the development of contraction bubbles at different stages in the necking down process. Modified from Roedder E (1984) Fluid inclusions. *Mineralogical Society of America, Reviews in Mineralogy* 12, and Shepherd TJ, Rankin AH, and Alderton DHM (1985) *A Practical Guide to Fluid Inclusion Studies*. London: Blackie and Sons.

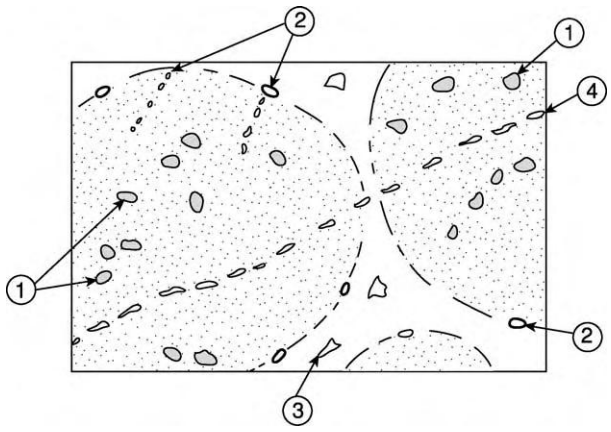


Figure 7 Schematic representation of four different generations of fluid inclusions in a quartz cemented sandstone. The earliest inclusions (1) are contained within the quartz grains and represent fluid processes prior to sedimentation and diagenesis. The latest inclusions (4) cross cut both the sand grains and the quartz cement and are also unrelated to the cementation process. Inclusions related to early cementation (2) may occur at the quartz grain boundary and along healed microfractures in the quartz grains. Inclusions related to later cementation processes (3) occur within the centre of the cement.

Fluid inclusion studies on hydrothermal ore deposits are usually carried out on transparent gangue minerals because the associated ore minerals are usually opaque. Even when primary inclusions are present, it is critically important to establish a clear mineral paragenesis to ensure that a particular generation of transparent gangue is coeval with a particular stage of ore mineralization.

Principles of Fluid Inclusion Geothermometry and Geobarometry

One of the most important applications of fluid inclusion studies is their use as palaeogeothermometers and palaeogeobarometers. This involves the careful measurement of the temperatures at which the heterogeneous contents of the inclusion became

homogeneous on heating, referred to as the homogenization temperature (T_h). This involves direct observations, under the microscope, using specially constructed heating (and cooling) stages covering the temperature range from −196°C (the boiling point of liquid nitrogen used as a coolant) up to about 1200°C. In the case of high-temperature melt inclusions, an alternative is to carry out observations on the inclusions after a series of successive heat/quench runs using conventional muffle furnaces.

The homogenization temperature may be used to place constraints on the *P–T* trapping conditions and, hence, the formation conditions if the inclusions in the host rocks are primary, provided that:

1. The inclusions have maintained their integrity over geological time.
2. Homogeneous versus heterogeneous trapping conditions can be recognized.
3. The bulk composition of the inclusions (aqueous brine, silicate melt, H_2O fluid, liquid petroleum, etc.) is known and the *P–V–T* data for the relevant fluid systems are available from the literature or can be calculated from thermodynamic data.

Aqueous inclusions and gaseous inclusions The principles are illustrated in Figure 8 with reference to two-phase aqueous inclusions modelled on the simple H_2O system. The homogenization temperature defines a unique point on the vapour–liquid or boiling point curve and represents the minimum *P–T* trapping conditions. As the composition (X) is known, this point also uniquely defines the density ($1/V$) of the fluid. The line of equal density emanating from this point is known as an isochore. The true *P–T* trapping conditions must have taken place at some point along this isochore. If depths or pressures are known, the trapping temperature (T_t) can be estimated. Reciprocally, if pressure-independent methods, such as oxygen isotope geothermometry, are available, the pressure or depth can be established.

The same principles apply to gas-rich inclusions in the system CO_2 N_2 CH_4, which characterize many metamorphic environments and associated mineral deposits. Homogenization will take place either to liquid or vapour at temperatures below 31°C, the highest critical point for an end-member gaseous component in this system (CO_2). By studying different generations of gas-rich inclusions, it may be possible to establish a reasonable idea of the *P–T* evolution and peak conditions during the metamorphic history of a particular sample.

Coexisting aqueous and gas-rich or petroleum-bearing inclusions In cases in which it can be demonstrated that aqueous and gaseous fluid inclusions represent the homogeneously trapped end members of a 'boiling fluid', trapping must have occurred on the G + L curve. In these cases, T_h will be the same as the trapping temperature (T_t). An extension of this principle is the 'intersecting isochore' method of geothermobarometry which can be applied to homogenization temperature data for coexisting aqueous and hydrocarbon inclusions, using the relevant bubble point or methane saturation curves, as illustrated in Figure 9.

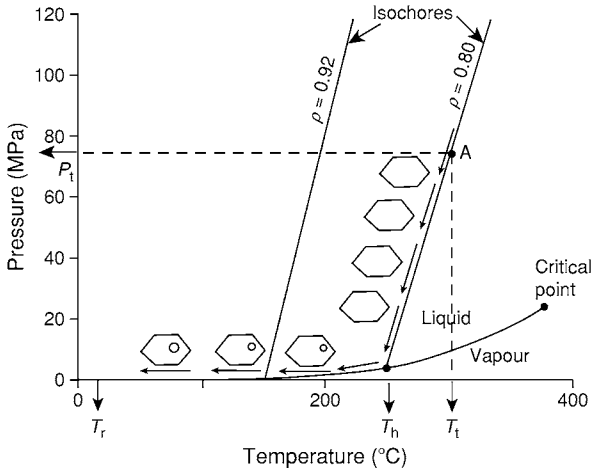


Figure 8 *P T* phase diagram for pure water illustrating the principles of fluid inclusion geothermometry and geobarometry. An inclusion trapped at P_t and T_t will follow the *P T* cooling path defined by a line of equal fluid density (an isochore). The inclusion remains homogeneous until the isochore meets the vapour liquid equilibrium curve, at which point a vapour bubble nucleates. Continued cooling takes place along this curve, causing further shrinkage of liquid and growth of the bubble down to room temperature. On heating, the liquid expands, completely filling the inclusion at a point on the vapour liquid curve referred to as the 'homogenization tempera ture' (T_h), which represents the minimum *P T* trapping condi tions. However, trapping could have occurred anywhere within the single phase region above this curve along the *P T* path defined by the isochore. An estimate of P_t (and salinity) is re quired to determine T_t. The difference between T_t and T_h is often referred to as the 'pressure correction'.

Melt inclusions It is generally accepted that the vapour–liquid homogenization temperatures of silicate melt or glass inclusions in minerals from intrusive or subvolcanic igneous rocks approximate to their trapping temperature without the need for pressure corrections. This is because silicate melts are far less compressible than other fluids, such as water or CO_2. In deeper seated intrusions, care is needed to ensure that equilibrium conditions have been maintained during the homogenization of multicomponent crystalline melt inclusions by using slow heating rates.

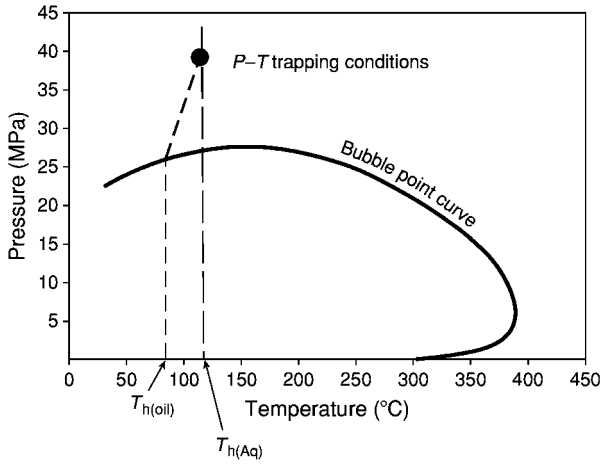


Figure 9 Principle of the isochore intersection method applied to coexisting oil and aqueous inclusions. The bubble point curve can be calculated from estimated oil compositions (e.g., from gas chromatography analysis of the inclusions). The homogenization temperature of the oil, $T_{h(oil)}$, can be used to calculate the *P T* trapping path of the oil inclusion in the single phase region above this curve. The point at which the isochore intersects the homogenization temperature line for coexisting aqueous inclu sions, $T_{h(Aq)}$, defines the *P T* trapping conditions. See Munz In: Andersen T, Frezotti M L, and Burke EAJ (eds.) (2001) Petroleum inclusions in sedimentary basins: systematics, analytical methods and applications. *Lithos* 55: 185 212.

The T_h values of melt inclusions can be used effectively as independent geobarometers in cases in which silicate melts and associated fluid inclusions have been trapped coevally. The isochores for these associated aqueous or gaseous inclusions can be used to fix the pressure or depth of formation.

Chemical Compositions of Geological Fluids from the Analysis of Inclusions

A wide variety of methods are available to determine the chemical compositions of fluid and melt inclusions.

Simple optical methods Simple optical methods are often very effective. For example, the Becke line test can be used to distinguish between aqueous fluid and silicate glass phases because of their contrasting refractive indices. Ultraviolet (UV) fluorescence microscopy can be used to identify petroleum inclusions, and the identification of daughter minerals can sometimes be based on their optical and crystallographic properties.

Microthermometry Microscope heating and freezing stage studies (microthermometry) are frequently used to identify CO_2 and associated volatiles and to estimate salinities of aqueous inclusions. Pure CO_2 should melt at $-56.6°C$ (the triple point). Other volatiles, such as CH_4 and N_2, will lower this melting temperature. Similarly, the melting point of pure water is lowered by the addition of salts, such as NaCl, KCl, and $CaCl_2$, in solution. By measuring this depression and referring to published freezing point depression curves, it is possible to estimate the salt content or salinity of aqueous inclusions. Mostly, this is performed with reference to the simple NaCl H_2O system, and the results are usually expressed as equivalent weight per cent NaCl to take into account the fact that components other than NaCl are also usually present. In practice, because of supercooling effects, the melting temperatures of frozen inclusions, rather than their freezing temperatures, are used to estimate salinities.

Bulk methods A range of instrumental methods with the required sensitivity (parts per million to parts per billion range) are available for the analysis of inclusions in bulk samples (100 mg to 1000 g). Gas chromatography and mass spectrometry are widely used to determine volatile and hydrocarbon contents and the stable isotopic compositions of inclusion fluids released by crushing or heating. The analysis of cations and anions in aqueous leachates from crushed samples has often been used to determine the chemical compositions of aqueous inclusions. However, in common with other bulk methods, there may be significant problems of contamination from the host or admixed minerals, and it is not always possible to obtain a large enough sample with a single or dominant generation of inclusions.

Point methods Various spectroscopic and microbeam methods have been applied to the *in situ* analysis of fluid and melt inclusions. These include electron microprobe analysis, which is particularly suited to the analysis of melt inclusions, and proton-induced X-ray emission (PIXE) and synchrotron X-ray fluorescence (SXRF) spectroscopy, which are particularly suited to the analysis of trace elements and ore metals in melt and brine inclusions. However, the methods which have had the largest impact in recent years are laser Raman microprobe analysis (LRM) and laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). LRM is non-destructive and is widely used to identify the solid, liquid, and volatile components in unopened inclusions down to a few micrometres in size. It is particularly useful for determining the molar volumes of CO_2, N_2, CH_4, and other Raman-active volatiles in aqueous inclusions and H_2O in silicate melt inclusions. It has also been used to identify a wide variety

Figure 10 Crystallized melt inclusion in flux grown synthetic ruby. Note the deformed vapour bubble (transparent) and crystalline flux material.

of daughter minerals in inclusions. The LA-ICP-MS method is destructive, but is capable of multielemental and isotopic analysis of individual inclusions down to about 20 μm in size. It is particularly suited to the analysis of trace elements (e.g., ore metals and rare earth elements) in inclusions.

Fluid Inclusion Studies as Tools in Exploration

The distribution and abundance of particular types of fluid inclusion have been used, with variable degrees of success, in the exploration for mineral, petroleum, and geothermal energy resources. Examples include:

1. The logging of the distribution of oil-bearing inclusions in exploration wells, which can be used to infer the migration of oil in particular lithologies and the former presence of oil in an otherwise dry well.
2. The mapping of the overall abundance and distribution of fluid inclusions in crystalline rocks, which often show regional and localized increases (on metre to kilometre scales) near to zones of intense hydrothermal alteration and mineralization.
3. The recognition of 'boiling' assemblages, which may be used to help locate high-grade bonanza gold areas in epithermal gold veins or the depths of boiling zones in modern geothermal systems.

Fluid inclusion studies are useful in assessing the suitability of proposed sites for nuclear waste disposal, because they serve as indicators of the fluid processes that have affected the proposed repository site throughout its geological history.

Finally, in gemmology, the nature and distribution of various solid and fluid inclusions are often used to help distinguish natural from artificial gems or simulates (Figure 10). They can be used to characterize the source or geographical location of a particular gem mineral, and also to assess whether colour has been enhanced through artificial heat treatment.

See Also

Diagenesis, Overview. **Gemstones**. **Gold**. **Igneous Processes**. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Petroleum Geology:** The Petroleum System. **Tectonics:** Hydrothermal Activity.

Futher Reading

Andersen T, Frezotti M L, and Burke EAJ (eds.) (2001) A series of review papers in this special issue of *Lithos* dealing with all aspects of fluid inclusion studies. *Lithos* 55: 1 321.

Goldstein RH and Reynolds TJ (1994) Systematics of fluid inclusions in diagenetic minerals. Short Course No. 31, Society of Economic Palaeontologists and Mineralogists, Tulsa, USA.

Gubelin EJ and Koivula JI (1986) *Photoatlas of Inclusions in Gemstones*. Zurich: ABC Edition.

Lowenstern JB (1995) Applications of silicate melt inclu sions to the study of magmatic volatiles. In: Thompson JFH (ed.) *Magmas, Fluids and Ore Deposits. Short Course*, vol. 23. pp. 71 99. Mineralogical Association of Canada (Vancouver, B.C.).

Rankin AH (1989) Fluid inclusions. *Geology Today* 5: 21 24.

Richards JP and Larson PB (eds.) (1998) Techniques in hydrothermal ore deposits: geology. *Reviews in Economic Geology* 10 (contains reviews on Fluid Inclusion Tech niques of Analysis (T. J. Shepherd & A. H. Rankin) and Fluid Inclusion Modeling for Hydrothermal Systems (P. E. Brown)).

Roedder E (1972) The composition of fluid inclusions. *US Geological Survey, Professional Paper* 440 JJ.

Roedder E (1984) Fluid inclusions. *Mineralogical Society of America, Reviews in Mineralogy* 12.

Roedder E and Bodnar RJ (1980) Geologic pressure deter minations from fluid inclusion studies. *Annual Review of Earth and Planetary Sciences* 8: 263 301.

Roedder E and Bodnar RJ (1997) Fluid inclusion studies of hydrothermal ore deposits. In: Barnes HL (ed.) *Geochemistry of Hydrothermal Ore Deposits*, 3rd edn. ch. 13, pp. 657 697. New York: Wiley and Sons.

Shepherd TJ, Rankin AH, and Alderton DHM (1985) *A Practical Guide to Fluid Inclusion Studies*. London: Blackie and Sons.

Sorby HC (1858) On the microscopical structure of crystals, indicating the origin of minerals and rocks. *Quarterly Journal of the Geological Society of London* 14: 453 500.

FORENSIC GEOLOGY

K Pye, Royal Holloway, University of London, Egham, UK

Introduction

Forensic geology is concerned with the application of geological data and techniques to issues that may come before a court of law. It is closely related to environmental forensics, forensic engineering, and forensic archaeology. Environmental forensics is somewhat broader in scope than forensic geology and involves a wider range of environmental data, knowledge, and expertise. It frequently involves investigations of environmental problems such as water and air pollution. Forensic engineering also overlaps with environmental forensics and is typically concerned with such issues as ground stability, the failure of buildings and other engineering structures, flooding, wind damage, fires, and explosions.

All subdisciplines of the geosciences have potential forensic applications, but sedimentology, mineralogy, petrology, geochemistry, palaeontology, and geophysics have so far made the greatest contributions. Shallow geophysical prospecting methods have been widely used by forensic archaeologists and others to locate and characterize clandestine graves and buried objects such as drugs and weapons (*see* **Engineering Geology:** Geophysics). However, probably the most widely recognized application of forensic geology is the use of geological materials as trace evidence, which can be of value in linking a suspect to a crime scene. In the wider forensic and legal literature, sediment, soil, dust, and rock fragments have often been grouped together under the loose term 'soil' evidence.

Some of the earliest users of geological and soil evidence were not, in fact, geologists. As early as 1893 Hans Gross, an Austrian professor of criminology, had pointed out the value of examining 'dirt' on a suspect's shoes as a possible indicator of their movements, and the German chemist Georg Popp is widely credited with being the first to undertake systematic 'soil'-comparison studies in the early 1900s. Around this time, the English writer Sir Arthur Conan Doyle, literary creator of the investigator Sherlock Holmes, also used soil comparison in one of three cases in which he became a real-life investigator. The work of these men had an important influence on Edmund Locard, initially a student of forensic medicine, who later went on to be Director of the Technical Police Laboratory in Lyons, France. Locard developed the first detailed scientific procedures for the analysis of dust traces and established the famous Locard exchange principle:

> Whenever two objects come into contact, there is always a transfer of material. The methods of detection may not be sensitive enough to demonstrate this, or the decay rate may be so rapid that all evidence of transfer has vanished after a given time. Nonetheless, the transfer has taken place. Locard (1930a Part1), cited by Murray and Tedrow 1992, p7)

In the USA, the Federal Bureau of Investigation Laboratory at Quantico, Virginia, was using soil and mineral analyses extensively in criminal cases as early as 1935 and has maintained a strong interest in this area. Other important forensic work relating to particulates was undertaken in the 1960s and 1970s at McCrone Associates in Chicago, where the first Particle Atlas was developed. Outside the USA, scientists at the UK Home Office Forensic Laboratory at Aldermaston, now closed, made extensive use of soil evidence in the 1960s and 1970s. In Japan, extensive use has been made of geological and soil evidence since the early 1980s by scientists at the National Research Institute of Police Science.

Since the early 1990s the true potential of geological and soil evidence, and that of other related subdisciplines such as forensic botany, forensic entomology, and forensic anthropology, has become much more widely recognized amongst police forces and forensic scientists worldwide. A much broader range of geological techniques and approaches is now being applied in the context of both criminal and civil-law investigations. Many of these techniques also have extensive application in other areas, including the war against terrorism, international drug smuggling, and broader environmental-quality and public-health campaigns. There is considerable current interest in applying forensic-geological techniques and principles to such issues as illegal trading in ivory and rhino horn, archaeological artefacts and works of art, and in traceability studies related to a wide range of foodstuffs and other commodities.

Despite this long-standing interest, there is still only a relatively small specialist published literature in forensic geology, and until recently Murray and Tedrow's book *Forensic Geology*, originally published in 1975 and reprinted in 1992, was the only one available in the field.

Types of Information Provided by Geological and Soil Evidence

There are many questions and issues to which forensic-geological techniques and information can be applied. These include:

whether an individual, motor vehicle, or other specified item was present at a particular location (e.g. crime scene),
the sequence and possible timing of a visit to that location and possibly others,
the location of buried objects (e.g. bodies, arms caches, drugs),
the source of imported or smuggled items,
the cause of death (especially in cases of possible drowning or suffocation),
the geographical origin of unidentified human remains, and
the length of time a body has been present at a location and the length of the post-mortem interval.

Traces of rock, sediment, soil, and dust can be present on a whole variety of items of interest, but amongst those most frequently submitted to the crime laboratory for examination are footwear, clothing, vehicles, flooring materials, digging implements, washing-machine filters, polythene bags in which items have been stored, firearms, and knives (Figures 1 and 2). Samples associated with the human body are also sometimes subject to examination. These include tapings of the skin, finger-nail scrapings, washings from the hair, nasal passages, trachea, and lungs, the contents of the gastrointestinal tract, and faeces.

With modern analytical techniques, very small traces of mud on an exhibit can provide ample sample material for a whole battery of tests to be undertaken (Figure 3). However, the greater the amount of material that is available for analysis, the wider the options in terms of the analytical techniques that can be employed. If only small amounts of mud or particulates are present, it is usually necessary to preserve as much of the evidence as possible for possible re-examination. Consequently, non-destructive tests, or those that are minimally destructive, are preferable to destructive tests that require a relatively large amount of sample material.

Even exhibits that on first examination might seem unpromising from the viewpoint of preservation of forensic evidence, such as a burnt-out car (Figure 4), can in fact be a source of useful geological evidence. Many geological materials are not affected by the temperatures that typically occur in standard vehicle fires, and in many instances particles of gravel and sand can be recovered from suitable parts of such vehicles, such as the footwells and suspension arms.

Figure 1 Boots and spade seized from an individual suspected of digging illegal treasure trove from a national heritage site.

Comparison of Suspect Samples with Crime-Scene Samples and Other Reference Samples from Known Locations

One of the simplest situations faced by the forensic geologist is to compare mud or soil on a suspect's footwear or clothing with reference samples taken from a known crime scene. In most cases, comparison is also made with reference samples from one or more other locations (e.g. the suspect's home address or place of work) for elimination purposes. Comparisons of the samples should be based on several criteria, chosen from the list shown in Table 1, the choice depending on a number of factors such as the amount and type of material present in the forensic samples, any available background information that might indicate which criteria are likely to provide the best discrimination, and time, cost, and equipment-availability limitations. Clearly, the greater the number of lines of comparison that can be used, the greater is the potential for discrimination between samples. In some circumstances a single method of comparison may be sufficient to screen samples and to identify those that can be eliminated from further consideration. However, great care needs to be taken in selecting the screening method, since some soil and sediment properties, such as colour and particle size, may vary considerably over short distances and with time. As a general rule, at least three independent methods should be used for sample comparison, and considerably more may be required where an apparent similarity between 'suspect' and crime-scene samples is identified.

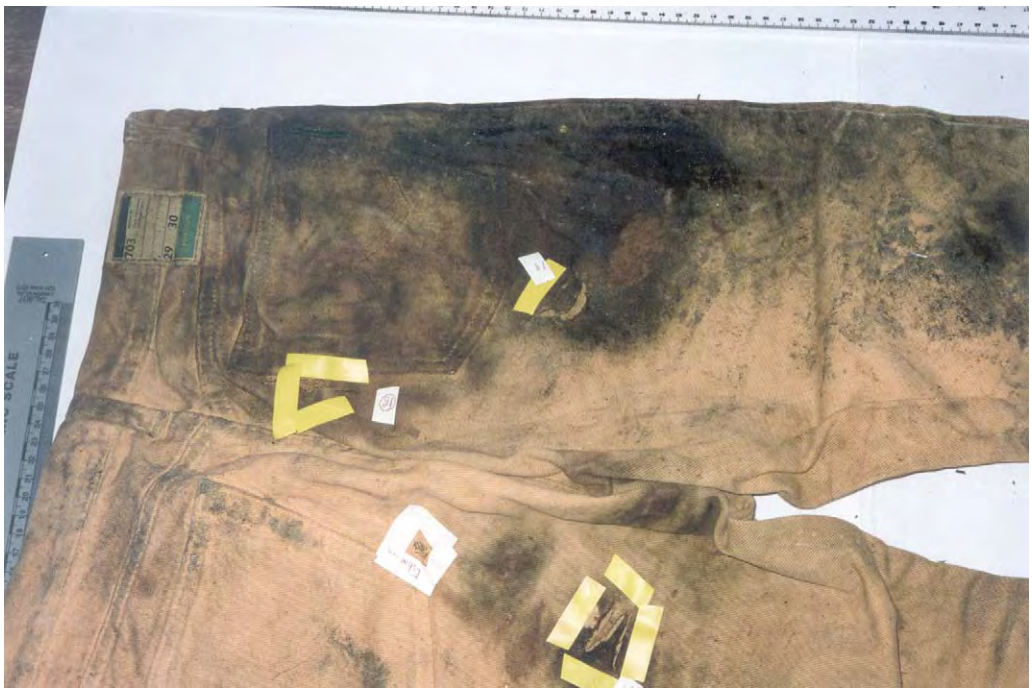


Figure 2 Extensive soil staining on a pair of jeans taken from the victim of a multiple stabbing dumped in woodland.

Figure 3 Mud spots on the jersey of an individual suspected of having buried the body of a murder victim.

Figure 5 shows a soil-covered house brick that was recovered from a hold-all containing dismembered body parts found in a London canal. The identity of the victim was unknown, and forensic examination of the hold-all showed it to be of a type widely available in the UK. Initial questions raised by the police therefore included whether anything could be said about the origin of the brick based on an analysis of the brick itself and also of the adhering soil. The approach in this case was to remove the soil and to analyse both the bulk material and different size fractions using a combination of techniques. These included quantitative colour analysis, particle-size analysis, mineralogical analysis, major and trace element analysis, and pollen analysis. The results indicated that the probable source was a garden adjacent to a domestic property in a part of north London, and subsequent comparison with reference samples taken from the garden of a suspect who came into the enquiry showed a very high degree of similarity.

Figure 6 shows a boot with extensive red soil staining taken from the body of an illegal immigrant who was found in the wheel-well of a Boeing 747 aircraft that landed at London's Heathrow airport. The man carried no formal identification or indication of his origin. Since the plane had made a number of flights to several different countries since the wheel-well had last been subject to detailed examination, there were three possible places where the unidentified individual could have managed to stow away. In order

Figure 4 A burnt out car used by a man who was later convicted of having stabbed his wife to death.

Table 1 Aspects of sediments and soils that have been used for the purposes of forensic comparison

(A) Bulk sample properties

Bulk sample properties	*Main techniques and equipment used*
Rock/sediment/soil texture	X radiography, microtomography, optical and scanning electron microscopy, image analysis
Particle size distribution	Dry and wet sieving, laser granulometry
Particle shape properties	Image analysis
Surface area	Nitrogen gas adsorption
Colour	Colour charts, spectrophotometry
pH	pH electrode, colorimetry
Water soluble cations and anions	Atomic absorption, ion chromatography
Enzymes	Enzymatic extraction
Bacteria	Culture experiments, microscopy
Lipid biomarkers	Gas chromatography mass spectrometry
Carbon, nitrogen, and sulphur contents	Wet chemistry, CHNOS elemental analyzer
Bulk organic matter content	Walkley Black colorimetric method, Fourier transform infrared spectroscopy, pyrolysis gas chromatography mass spectrometry
Polyaromatic hydrocarbons	Gas chromatography mass spectrometry, high pressure liquid chromatography
Calcium carbonate content	Collins calcimeter, Chittick apparatus
Thermoluminescence characteristics	Heat induced photon emission
Fluorescence characteristics	Fluorescence microscopy
Major and trace element composition	X ray fluorescence, inductively coupled plasma spectrometry, neutron activation
Bulk mineralogy	Optical microscopy, point counting, automated scanning electron microscopy, X ray chemical microanalysis, X ray diffraction
Clay mineralogy	X ray diffraction, infrared spectroscopy
Mineral magnetics	Magnetic susceptibility, frequency dependent susceptibility, isothermal remanent magnetization
Stable carbon, nitrogen, and sulphur isotopes	Continuous flow mass spectrometry, laser fluorination mass spectrometry

Table 1 Continued

(B) Individual particle type properties and assemblages

Individual particle type	*Main techniques and equipment used*
Radiogenic isotopes	Thermal ionization mass spectrometry, quadrupole mass spectrometry, laser ablation mass spectrometry
Radioactive isotopes	Alpha counting, beta counting, gamma counting
Opal phytoliths	Optical and scanning electron microscopy, supplemented by energy dispersive X ray chemical analysis
Foraminifera	As above
Coccoliths	
Coralline particles	
Molluscs	
Gastropods	
Ostracods	
Diatoms	
Insect remains	
Pollen and spores	
Plant seeds	
Leaf and stem fragments	
Coal fragments	
Charcoal fragments	
Wood fragments	
Quartz sand grain surface textures	
Gravel surface textures	
Coatings on mineral grains	
Light fraction mineral grains	As above, plus cathodoluminescence microscopy, laser Raman spectroscopy, electron probe analysis, ion probe analysis, laser ablation inductively coupled plasma spectrometry, microspectrophotometry, dating by Ar Ar and U Pb series methods
Heavy minerals	As above
Slag and ash	As above, with exception of dating
Spherules	As above
Brick	
Concrete	
Pottery	
Glass	
Alloys and pure metals	
Fibres	Optical microscopy, fluorescence microscopy, microspectrophotometry, scanning electron microscopy, X ray chemical microanalysis, Fourier transform infrared spectroscopy and microscopy, ultraviolet spectroscopy
Paint	As above
Paper	As above

to identify the location, red soil from the boots was examined using a combination of techniques including quantitative colour analysis, bulk sample and clay-fraction mineralogy by X-ray diffraction, chemical analysis by inductively coupled plasma spectrometry, and pollen analysis. The results clearly indicated a source in a wet tropical country. Comparison was made with control samples taken from adjacent to the airport in the country (Ghana) from which the plane had last departed prior to its arrival at Heathrow, and a very high degree of similarity was obtained in terms of all comparison criteria (Figure 7).

The first stage in any forensic comparison of soil or other geological samples is to determine whether or not a possible 'match' can be excluded. If it can, then no further attention need be given to that sample. If it cannot, then further investigation may be warranted. A conclusion of an exact match can sometimes be drawn with virtual certainty when the samples in question make a physical fit and have the same texture and chemical composition. This may occur, for example, with two halves or several broken pieces of rock or ornamental stone. In other circumstances a physical fit may be observed, for example, between a shoe impression in mud and a shoe seized from a suspect that is of the same size and has the same tread pattern as the shoe that made the impression. However, there may be several thousand such shoes in circulation, and a specific 'match' with an individual shoe often cannot be made. In this instance, analysis of mud adhering to the shoe, if shown to be indistinguishable from that in which

Figure 5 House brick with soil staining recovered from a hold all containing the dismembered remains of a prostitute dumped in a canal.

Figure 6 Soil stained boot from a deceased male found in the undercarriage stowage space of a Boeing 747.

the shoe impression was found, may provide strong supportive evidence that the particular shoe under consideration made the mark.

However, there are many cases where mud-stained footwear is recovered during an investigation but an exact spot at the crime scene where it may have been acquired has not been identified. In such cases, comparisons of the soil on the shoe with several different reference samples from the crime scene, and usually elsewhere, have to be made on the basis of statistical and graphical comparisons, and the results can be interpreted only in probabilistic terms. The degree of similarity between samples can be expressed in several semi-quantitative and quantitative ways, but meaningful statistics about the likelihood of such a degree of apparent 'match' being due to chance are often difficult to provide. This is because the full range of variation that exists in natural soils is impossible to determine and can be estimated only on the basis of sampling. The availability of database information relating to suitable comparison samples varies greatly from one region to another, and there may be a total absence of pre-existing information in some parts of the world. In such cases, it is necessary to undertake a suitable background investigation, involving collection and analysis of a sufficiently large number of reference samples, in order to provide adequate contextual information for interpretation.

Where no physical fit has been identified, the nearest thing to a definitive connection between two questioned samples is usually provided by the identification of one, or more commonly several, highly unusual (or 'exotic') particle types in both samples. These may be naturally occurring particles or they may be of human or animal origin. They need not be considered 'unique' in themselves, but should be sufficiently rare, either alone or in combination with other unusual particles in the same sample, that the chance of them occurring in any two samples under investigation is extremely low. Examples of two particles that fall into this category are shown in Figures 8 and 9. Waste-dumps, industrial premises, and roadside verges are examples of locations that often contain mixtures of particles that have a more restricted distribution than natural soils. The assemblages of particles present in such locations often show considerable local variation, and it may be possible to limit a potential source area to just a few square metres.

Persistence of Geological Evidence

Geological evidence may persist for a considerable period of time after it has been picked up from the source location. For example, gravel, sand, or mud that enters

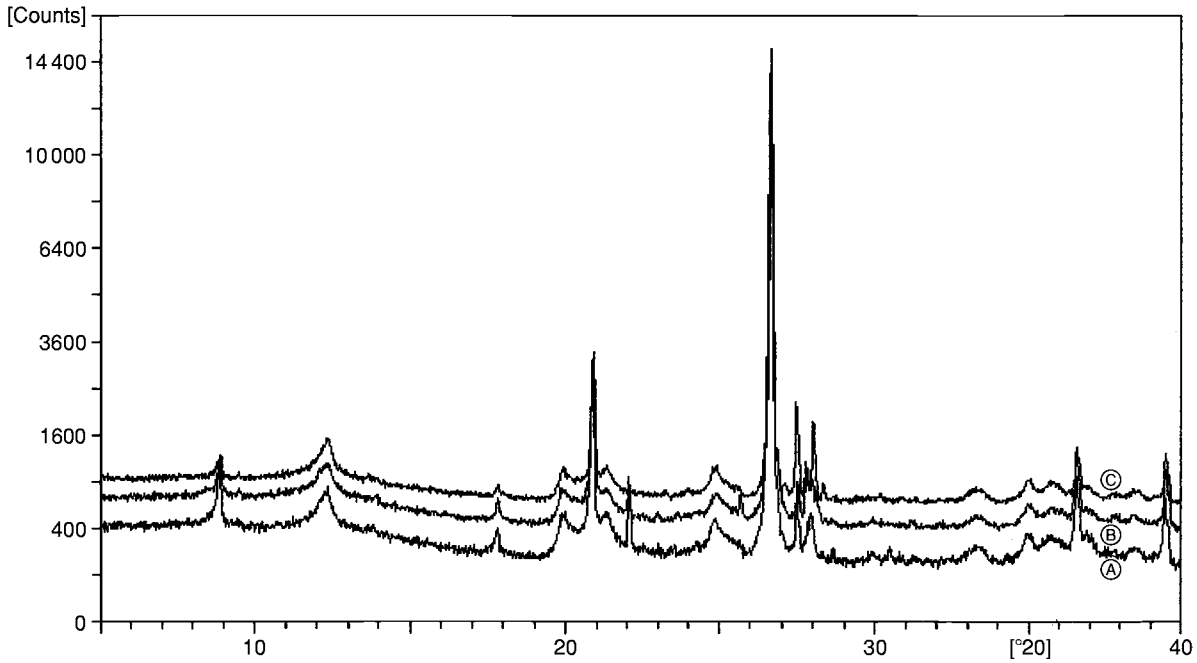


Figure 7 Comparison of X ray powder diffractograms for the $<150\,\mu$m fraction of soil samples from the right and left boots (A and B) of a deceased stowaway found in the undercarriage stowage space of a Boeing 747 jet after landing in London, compared with a control sample from Accra airport, Ghana (C).

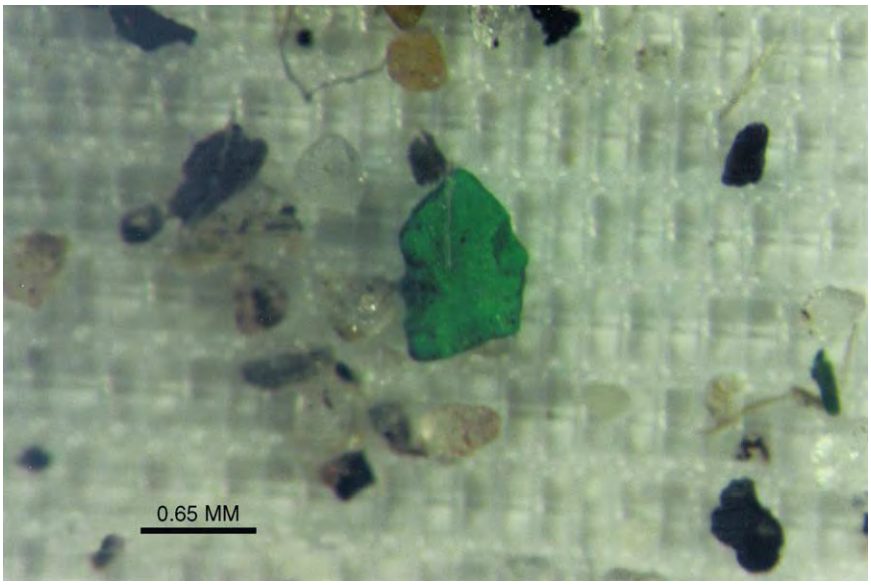


Figure 8 Example of an 'exotic' synthetic chromium rich particle, identified by optical microscopy, which was interpreted to be of 'industrial' origin.

the interior of a car on footwear, clothing, or other items (such as a spade) will stay there, subject to some loss due to outward transfer on the footwear or clothing of later occupants, until such time as the vehicle is thoroughly cleaned. Even after cleaning, traces may remain in certain hard-to-access locations. Whereas some forms of forensic botanical evidence degenerate as a result of oxidation or fungal decomposition, most inorganic sediment particles are very resistant and may be immune to changes over time. Consequently, they provide useful clues years or even decades after a crime was committed, providing that exhibits have been retained and suitably stored to prevent environmental and cross-exhibit contamination.

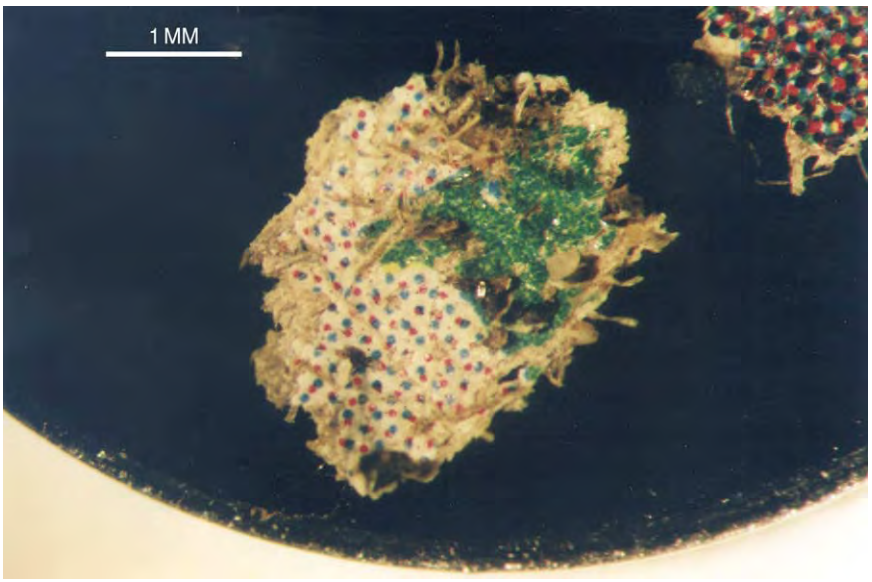


Figure 9 A further example of 'unusual' particles found in soil on the boots of a suspect who was later convicted of murder. The particles are from the cover of a particular issue of a glossy magazine, copies of which had been dumped at the murder scene.

Modification of Primary-Transfer Soil Evidence

It should always be borne in mind that material initially picked up from a location, a process referred to as primary transfer, may subsequently be modified as additional particles are picked up from other locations or some of the primary particles are lost during subsequent movement. During primary transfer, the material transferred may not exactly reflect the nature of the material at the source point, depending on the nature of the material and the nature of the contact involved. For example, if a person sits or lies on wet ground there is frequently a tendency for the finer particles to be selectively transferred and retained on the clothing. In other circumstances, only a certain size range of coarse particles may be retained, for example coarser particles trapped within the detail of footwear sole treads or gravel particles trapped in tyre treads. For this reason, it is important that sample comparisons are undertaken on narrowly defined size fractions as well as on bulk samples.

Secondary Transfer

The possibility of secondary transfer of soil and other geological evidence should always be borne in mind when exhibits are examined. For example, consider a case where Person A walks across a muddy car park and gets into the passenger seat of a vehicle, which is driven by another person (B) to another location, where A gets out. During this process, mud from the car park is transferred via the footwear of Person A to the front passenger footwell of the vehicle. The driver of the car (B) then drives to a third location and picks up another person, C, who also sits in the front passenger seat. The shoes of Person C come into contact with mud in the front passenger footwell deposited by Person A, and this is then transferred out of the car, via Person C's footwear, onto the hallway carpet at Person C's home address. If it subsequently emerges that a crime has been committed in the car park, and Person C becomes a suspect, simple comparison of mud on Person C's footwear and hall carpet with control samples from the car park might lead to the spurious suggestion that Person C had been present at the crime scene. For this reason, great care needs to be taken by the forensic geologist to document the amount, distribution, layering, and nature of any mud or similar evidence present on items, including footwear, submitted for analysis. In this connection police and scenes-of-crime examiners have an important responsibility to provide the forensic geologist with necessary background information and to ensure that comparison samples are taken from all locations and exhibits of possible relevance for comparative examination.

Location of Crime Scenes, Buried Bodies, and Weapons and Drugs Caches

A frequent problem that the forensic geologist is asked to address concerns the identification of the location where a crime has taken place, perhaps

involving the disposal of a body, or where weapons, money, and drugs have been stored or buried. In many instances one or more suspects are identified and their vehicles seized for possible identification of evidence that might indicate the location of the deposition or burial site. Both the inside and outside of the vehicle are then examined in detail, in parallel with searches for blood, other forms of DNA, fibres, hairs, and fingerprints. Numerous samples are normally taken from the footwells, wheel arches, mud flaps, and other parts of the bodywork and chassis to build up as detailed a picture as possible of the vehicle's recent movements. Similar examination and sampling is often undertaken on associated items such as petrol cans, car jacks, mats, spades, footwear, and clothing belonging to the suspect. The full spectrum of the samples' physical, chemical, and biological make-up is examined, the objective being to create an environmental profile of the samples, which may assist the direction of further police enquiries.

In this type of work, individual particle types, which may be either inorganic or biological, can be highly diagnostic. Particular pollen types may indicate specific ecological habitats, such as moorland, coniferous forest, deciduous broad-leaf woodland, or salt-marsh. Diatom assemblages may indicate saline, brackish, or freshwater environments. Highly diverse assemblages of particles of industrial or human origin may indicate waste dumps or industrial estates. Natural rock particles, such as chalk, coal, slate, and basalt, or mineral assemblages may indicate particular areas within specific geological outcrops. Even the relative abundance of different morphological types and surface textures of common minerals such as quartz may suggest specific localized areas with a particular surface geology and soil type (**Figure 10**). Viewed in polished section under an optical or scanning electron microscope, rock types that have very similar chemical compositions can be seen to have quite different depositional and diagenetic textures, which may be specific to particular lithostratigraphic units only a few tens of millimetres in thickness (**Figure 11**). Such precise identification is often aided by the analysis of microfossil assemblages, including foraminifera and shell debris.

Figure 12 shows a number of gravel- and coarse-sand-sized particles that were recovered at post mortem from the trachea and bronchi of a murder victim who had been shot in the head and whose burning body was found dumped on a farm track outside Edinburgh. The lithological assemblage of the gravel clasts, which were characterized by numerous *in situ* freshwater diatoms, indicated that they had originated from a river or river-marginal setting in the Airdrie area near Glasgow. The large size of the particles and the depth of penetration into the lungs meant that

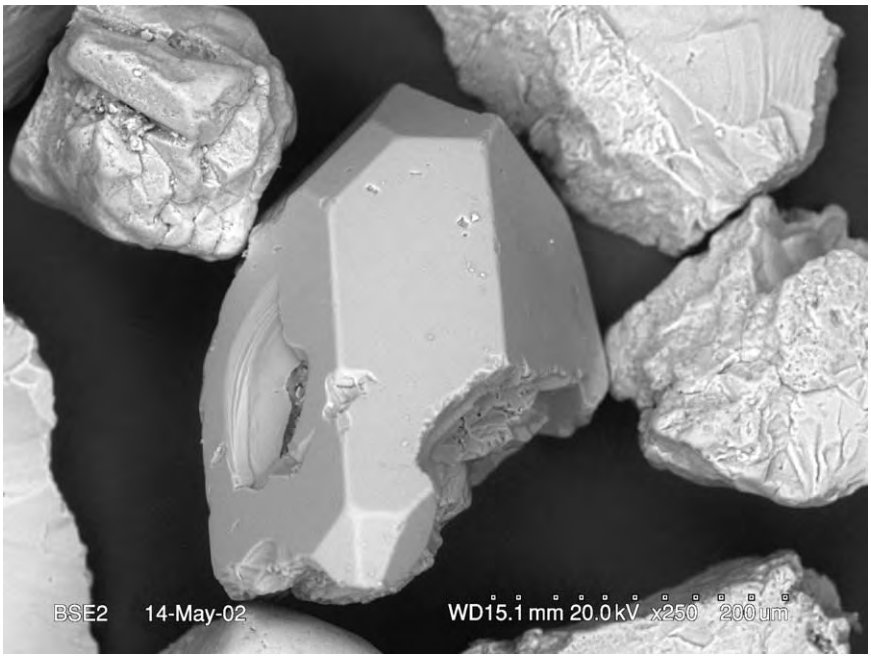


Figure 10 Scanning electron microscope image of a very fresh, unabraded, and uncorroded euhedral quartz grain. Such particles are normally found only very close to the rock source.

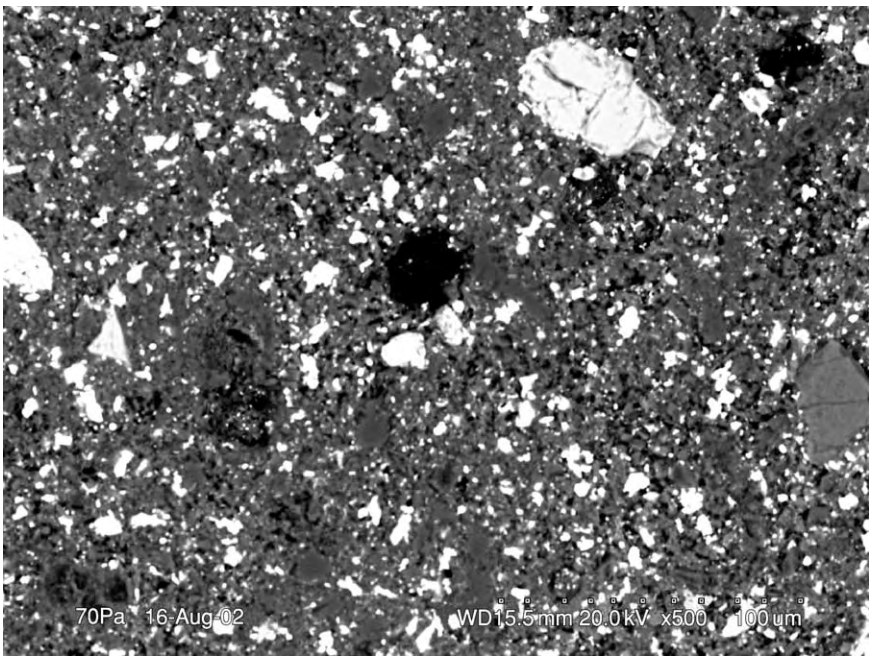


Figure 11 Scanning electron microscope image of a polished section of a piece of Chalk recovered from the suspension of a car owned by an individual who was later convicted of murdering two young girls and using the car to deposit their bodies.

they could have been sucked in only while the victim was alive, possibly being held face down with his head in water just prior to being shot.

Studies of Human Remains

Unidentified human remains are not infrequently found dumped by the roadside or washed up on the coast or in rivers and lakes. The remains may consist of whole bodies, with or without clothing, or of torsos or even isolated limbs. If identification proves impossible using dental records, fingerprints, or DNA, alternative means must be used to determine the geographical origin and identity of the victim. There are several ways in which the forensic geologist may contribute. First, studies of sand, mud, and dust particles present on clothing or on the outside of the body may indicate the area from which the body came (Figure 13). Particles that are exotic to the site where the body was discovered, such as coralline algae on a body found in the UK, would clearly indicate a recent tropical or subtropical association. Studies of pollen may provide further information about botanical exposure.

Studies of particles within the body may provide indications of environmental exposure in the hours or days immediately prior to death. Washings of the nasal passages, lungs, and hair may be useful in this regard, as may fingernail scrapings and particulates contained within the gastrointestinal tract and faeces. The absence of exotic pollen and inorganic particles within the gastrointestinal tract of a known recent immigrant may provide significant evidence that he or she had been present in the country where they were found for at least several days.

In cases where the cause of death cannot be determined with certainty by conventional post-mortem examination, examination of particulates in the lungs and other body tissues may assist in determining whether death was due to true drowning or some other cause prior to, or during, immersion. Analysis of diatoms present in the lungs, liver, spleen, blood, and bone marrow has for many years been undertaken as a confirmatory test in possible drowning cases (Figure 14). However, the diatom test has been controversial because numerous cases of false negative and false positive results have been documented. However, if the test is combined with studies of other particles that are known to be associated with a particular water body or type, its reliability is potentially much improved. As previously noted, quite large quantities of particles up to fine gravel in size can be aspirated into the trachea and lungs through the open mouth in the presence of water. Such particles may also be swallowed and carried into the stomach of a drowning person. Water is not always required, however: death may quite frequently be due

Figure 12 Group of 23 gravel sized particles recovered from the trachea and upper bronchii of a murder victim who had been shot in the head and his body set on fire, apparently after having his head immersed in a freshwater stream.

to suffocation when dry silt, sand, soil, or even fine gravel is forcibly or otherwise involuntarily inhaled. Careful analysis of diatoms and other environmentally sensitive biotic indicators associated with the inorganic particles is a useful means of determining whether death occurred while the face was submerged in water. Similar environmental discrimination can be achieved using sediment and soil sometimes found in other body orifices of deceased persons, including the anus and vagina.

Analysis of the trace-element and isotopic compositions of nail, hair, bones, and teeth can provide information about environmental exposure and diet over time periods ranging from a few days to an entire lifespan. Stable isotopes of carbon, oxygen, hydrogen, and nitrogen provide information about diet (including drinking water) and hence about climate; radiogenic isotopes of strontium, lead, and neodymium reflect aspects of diet, geological source terrain, and atmospheric exposure. In general, hair and nails provide information about time periods of a few days to a few months, bones provide information relating to the last few years of life, and tooth enamel provides information relating to the first few years of life, from *in vivo* until approximately age fifteen. By comparing data for teeth and parts of several different bones, information can be gained about human and animal migration during life.

Analysis of radioactive-isotope ratios in human remains may help to determine the post-mortem interval (i.e. the time since death). In the case of bodies that are between about a year and several decades old, lead and polonium isotopes provide the most useful information, while for longer time periods other isotopes, such as caesium and radiocarbon, can help to distinguish modern from 'archaeological' bone.

Increasingly, the work of forensic geologists is being used not only for intelligence purposes in criminal investigations but also as expert-witness evidence presented in court. Geological evidence has made a

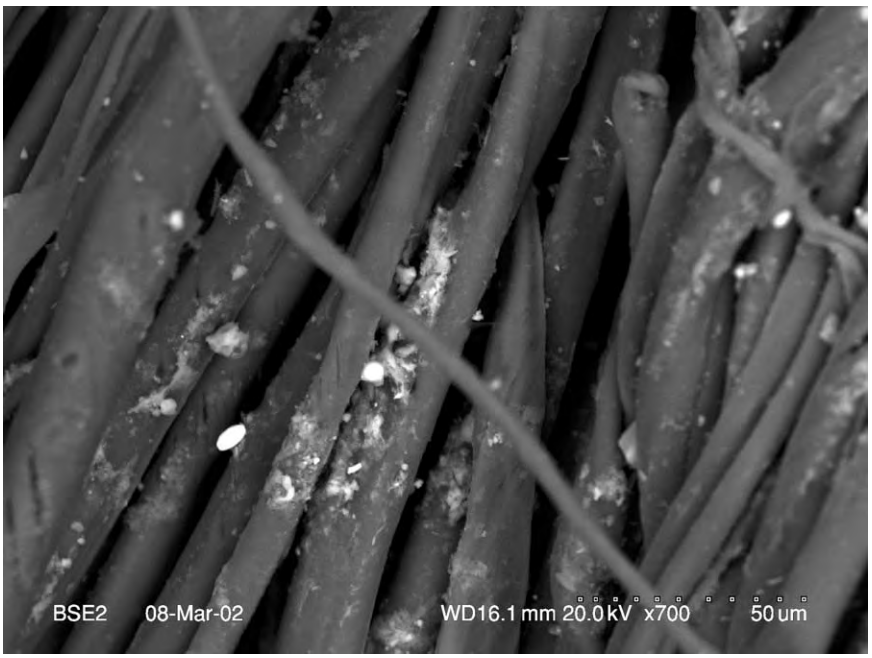


Figure 13 Scanning electron microscope image showing sediment particles adhering to fibres in a pair of orange shorts that had been placed on the torso of a murdered child found in the River Thames.

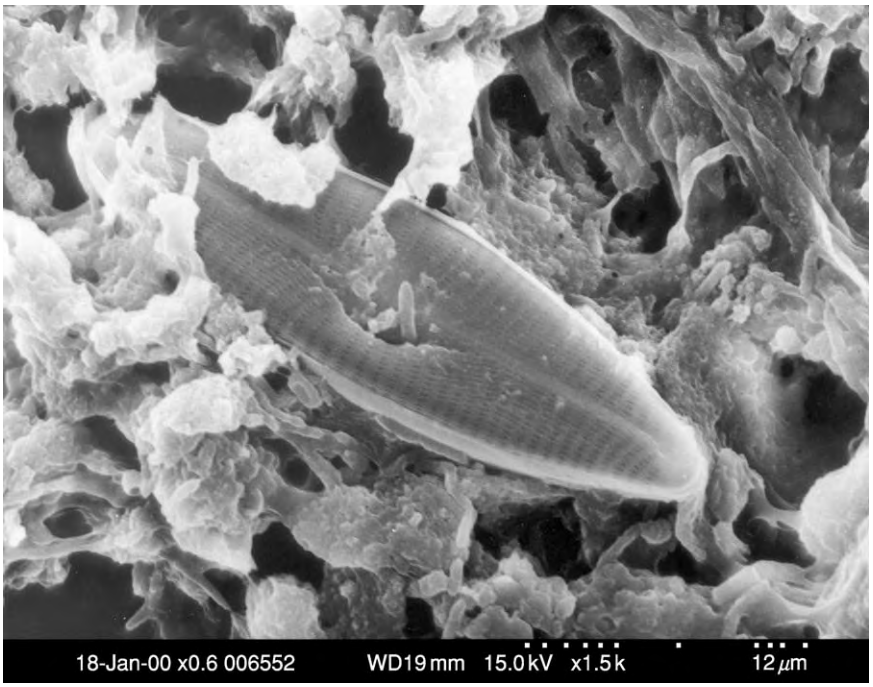


Figure 14 Scanning electron microscope image of a pennate diatom found in the lung tissue of a suspected drowning victim.

significant contribution in recent years to a number of high-profile trials involving murder, terrorism, international drug smuggling, and people trafficking. Other areas of frequent geological expert-witness testimony include various aspects of environmental forensics (e.g. contamination studies), engineering-geology failures, and traceability studies relating to food-stuffs and raw materials.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Engineering Geology:** Geophysics. **Geoarchaeology**. **Microfossils:** Palynology. **Soils:** Modern.

Further Reading

Brown A (2000) Going to ground. *Police Review* 4 February: 18 20.

Croft DJ and Pye K (2004) Multi technique comparison of source and primary transfer soil samples: an experi mental investigation. *Science and Justice* 44: 21 28.

Demmelmeyer H and Adam J (1995) Forensic investigation of soil and vegetable materials. *Forensic Science Review* 7: 119 142.

Foster IDL (ed.) (2000) *Tracers in Geomorphology.* Chichester: Wiley.

Hall DW (1997) Forensic botany. In: Haglund WD and Sorg MH (eds.) *Forensic Taphonomy,* pp. 353 363. Boca Raton: CRC Press.

Hunter J, Roberts C, and Martin A (1997) *Studies in Crime: An Introduction to Forensic Archaeology.* London: Routledge.

Kubic T and Petraco N (2002) Microanalysis and examin ation of trace evidence. In: James SH and Nordby JJ (eds.) *Forensic Science. An Introduction to Scientific and Inves tigative Techniques,* pp. 251 296. CRC Press: Boca Raton.

Locard E (1930) The analysis of dust traces. Part I. *American Journal of Police Science* 1930: 276 278.

Locard E (1930) The analysis of dust traces. Part II. *American Journal of Police Science* 1930: 401 418.

Locard E (1930) The analysis of dust traces. Part III. *American Journal of Police Science* 1930: 496 514.

Marumo Y, Sugita R, and Seta S (1999) Soil as evidence in crime investigation. *International Criminal Police Review* 474 475: 75 84.

McCrone WC, Delly JG, and Palenik S (1973) *The Particle Atlas. Volumes I to VI. An Enclopedia of Techniques for Small Particle Identification.* Chicago: Ann Arbor Science.

Morrison RD (2000) *Environmental Forensics. Principles and Applications.* Boca Raton: CRC Press.

Munroe R (1995) Forensic geology. *Royal Canadian Mounted Police Gazette* 57: 10 17.

Murphy BL and Morrison RD (eds.) (2002) *Introduction to Environmental Forensics.* San Diego: Academic Press.

Murray RC (2000) Devil in the details: the science of forensic geology. *Geotimes* 45: 14 17.

Murray R and Tedrow JFC (1992) *Forensic Geology,* 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

Pollanen MS (1998) *Forensic Diatomology and Drowning.* Amsterdam: Elsevier.

Pye K and Blott SJ (2004) Particle size analysis of sediments, soils and related particulate materials for forensic purposes using laser granulometry. *Forensic Science International* 144: 19 27.

Pye K and Croft DJ (eds.) (2004) *Forensic Geoscience Principle, Techniques and Applications.* Special Publica tion 232. London: Geological Society.

Sabine PA (1991) Geologists at war: a forensic investigation in the field of war time diplomacy. *Proceedings of the Geologists Association* 101: 139 143.

Shuirman G and Slosson JE (1992) *Forensic Engineering. Environmental Case Histories for Civil Engineers and Geologists.* San Diego: Academic Press.

FOSSIL INVERTEBRATES

Contents

Arthropods

L I Anderson, National Museums of Scotland, Edinburgh, UK

Introduction

Arthropods are the most diverse group of fossil and living invertebrate animals. They have colonized land, sea, and air due to a relatively plastic and adaptable body plan organization. Many different phyla of metazoans, including the arthropods, show repetition of structure along their long axis. This is generally termed metamerism.

Arthropod bodies are characteristically segmented, and the different major groups are defined by various patterns of tagmosis (segment fusion) to form heads, bodies, and tails. The phylum derives its name from the characteristic jointed appendages found in all members (Greek: αρθροσ '*arthros*', jointed; ποδοσ '*podos*', foot). These jointed appendages, which consist of individual elements called podomeres, take the form of walking legs, gill branches, or antennae. Broad divisions can be made within the arthropods depending on the exact nature of the leg appendages. Uniramians, or arthropods with single-branched appendages, retain only a walking or swimming leg. Biramians have an appendage which consists of two separate elements: a leg branch for movement and a gill branch used in respiration.

Arthropods all possess a rigid exoskeleton which provides support and protection, and acts as a buffer between the external environment and internal body processes. The exoskeleton of arthropods is composed of a carbohydrate polymer called chitin. In some arthropod groups, the chitin superstructure is reinforced with mineral salts, such as calcium carbonate or calcium phosphate. Arthropods with this biomineralization show a higher preservation potential than non-mineralized forms, and consequently have a more complete fossil record through time. For non-mineralized forms, we have to look to sites of exceptional preservation (Fossil Konservat-Lagerstätten) to trace their evolutionary history and discover aspects of their palaeobiology. Localities, such as the Middle Cambrian Burgess Shale, the Ordovician Soom Shale, the Devonian Rhynie chert, the Carboniferous Mazon Creek nodule hosted biota, the Jurassic Solnhofen Lithographic limestone, and Baltic, Dominican, and Mexican ambers, all provide windows onto the arthropod faunas at stratigraphical points

in geological time in which preservational failure would otherwise obscure our view of evolution of the various groups (*see* **Lagerstätten**). Furthermore, there is a gradation of fidelity of preservation even in these special examples. To this end, it is entirely possible to study the exquisite three-dimensional outer surfaces of Tertiary flies in amber (Figure 1) or the internalized respiratory systems of Devonian arachnids.

A consequence of the possession of a rigid exoskeleton composed of cuticle is that, in order for the arthropod to increase in body size over time, the rigid cuticle has to be shed or moulted and a new cuticle grown below the surface to replace it. Arthropods differ from many other animal groups in that they go through the process of ecdysis (literally shedding the exoskeleton as a moult) in order to attain greater body size. Ecdysis is facilitated by the presence of sutures in the exoskeleton, lines of weakness which can be broken through so that the arthropod can struggle free of the old skin. The ontogeny of ecdysis allows segment addition and tagmosis to be identified from the first instars right up to the adult forms. Unfortunately, for arthropods, moulting of the old skin and the time taken to harden the new exoskeleton leaves them vulnerable to predator attack or adverse environmental conditions. For this reason, many arthropods (both in aquatic and terrestrial environments) seek out a 'moulting refugia' in which the process of ecdysis and subsequent hardening of the new exoskeleton can take place in relative safety. Where mortality occurs during this post-moult stage, the cuticle of the arthropod fossil formed can display wrinkles, creases, or even tears in the structure.

Figure 1 A fungus gnat preserved in Baltic amber from the Upper Oligocene Lower Eocene of Palanga, Lithuania (×55). © The Trustees of the National Museums of Scotland, Edinburgh, UK (NMS G.2002.6.17).

The cuticle itself takes the form of hardened sclerites, one per body segment. Dorsal sclerites are termed tergites, ventral sclerites are termed sternites. Where the tergites and sternites are fused together, they form a simple ring structure enclosing the arthropod soft tissues. Lying between these discrete segments are areas of connective and unmineralized cuticle. The presence of this less robust construction material explains why, upon death, well-mineralized arthropods, such as trilobites, can disarticulate to a greater or lesser degree into their constituent parts (cephalon, thorax, thoracic segments, pygidium, etc.). The leg appendages in arthropods are similarly constructed with cylindrical reinforced cuticular podomeres with intervening, less well-sclerotized arthrodial membranes connecting them. Muscle attachments required to move the individual leg podomeres are attached to small boss-like projections from the inside of the exoskeleton, called apodemes.

The cuticle of arthropods explains in part their success in the colonization of such a wide variety of different environments with widely varying challenges. The cuticle provides a physical barrier between the internal organs and the external environment (which helps to reduce desiccation through water loss in terrestrial forms) and a chemical barrier in aquatic conditions (where hypo- or hypersalinity could upset the osmotic balance, leading to dehydration of the organism and ultimately death). As well as a physical barrier, waxes produced by the cuticle layer help to waterproof this material. Some other more unusual properties of arthropod cuticle have been identified recently, including the fluorescence under ultraviolet (UV) light of 'hyaline' cuticle within the exoskeleton of scorpions. It happens that this 'hyaline' layer is actually very resistant to bacterial degradation and is therefore often the main remaining constituent material of fossil scorpions. For example, the Early Carboniferous East Kirkton biota of West Lothian, Scotland, contained a number of different scorpion taxa which were recovered by hydrofluoric acid maceration of the enclosing limestone matrix.

The level of morphological detail made available through this technique was as good as studying extant species.

Classification Schemes

There still remains scope for significant disagreement in the higher level taxonomy of this group of animals. Historically, the main dichotomy relates to whether the phylum Arthropoda is a monophyletic or polyphyletic group of organisms. Some variation occurs in the higher level classification of this group, with some authors recognizing Arthropoda at phylum level and the constituent members variously as subphyla or orders. A monophyletic grouping for the arthropods would imply that all of the lower level groupings originated from a single common ancestor. This was the largely historic viewpoint based on the striking common appearance of the various groups. In order to counter what seems to be the logical conclusion based on many shared characters in common, the argument for polyphyly has to promote plausible explanations for the observed similarities between groups as well as providing proof of separate points of origin. The fossil record may provide just such evidence, allowing the evolution of the various arthropods to be considered with the benefit of deep time hindsight. However, since the work of Sidnie Manton, who suggested that three major groups, the Crustacea, the Chelicerata, and the Uniramia, would more comfortably be accommodated as phyla of their own, the consensus view has swung back to the likelihood of the monophyletic status of the group being the correct conclusion. This has been due in part to the advent of molecular sequencing techniques. Novel (non-morphological) hypotheses linking crustaceans and hexapods were primarily based on the analysis of single or a handful of genes. However, as the science advances and the techniques of extraction improve, the case for monophyly appears to be more solid. Indeed, one study not only supports the crustacean–hexapod clade, but also recognizes pycnogonids as the sister group to other euarthropods. The historical convention has been to subdivide the Arthropoda into the following equal-biased groupings:

- Trilobita (containing all trilobites, but excluding naroiid arthropods);
- Chelicerata (Arachnida, Scorpionida, Xiphosura, Eurypterida, Chasmataspida);
- Crustacea;
- Uniramia (including both insects and myriapods).

However, as more and more examples of basal arthropods and 'oddball' taxa are discovered, primarily from Fossil Konservat-Lagerstätten, other less conventional higher taxon groupings can be identified. Some examples include:

- Atelocerata (Diplopoda, Chilopoda, Pauropoda, Symphyla, Hexapoda);
- Arachnomorpha/Arachnata (Trilobites, Chelicerates, Trilobitomorpha).

Arachnomorpha is a clade which basically consists of all taxa more closely related to Chelicerates than crustaceans. Panarthropoda is again a clade, but this time containing all true arthropods and their soft-bodied relatives, such as onychophorans and tardigrades.

To this end, the term 'arthropod' should probably be used to represent a grade of organization in which a soft-bodied ancestor (presumably a segmented worm) developed a toughened cuticle. Gene sequencing and molecular markers may provide fascinating new insights into the interrelationships of the various high-level arthropod groups.

What the Fossil Record Says

Tracing arthropods back through geological time indicates that the divisions between the major groups extend almost as far back as the fossil record of Metazoa. Convergence of form and channelling by functional morphology of multiple arthropodization events through forming a rigid cuticle could have led to polyphyly; however, the more complete our record of the group becomes, the less likely this seems.

A Brief History

The first arthropods appear in the fossil record of life on Earth in the Lower Cambrian some 540 Ma. The arthropod faunas of the Chengjiang biota of China demonstrate that, even at this early stage, a diverse range of arthropod body plans had evolved. By Middle Cambrian times, rapid and explosive radiation and evolution of arthropods led to the development of many different body plans, including the majority of those still extant. This 'Cambrian Explosion' in the number and diversity of arthropodan groundplans was one of the first major radiations of the group and took place in the marine environment. By the Upper Cambrian, trilobites had diversified markedly and this carried through into Ordovician times.

Mid-Silurian and Lower Devonian rocks yield evidence of the next major step in the evolution of the arthropods: the colonization of the terrestrial environment, a process termed terrestrialization. Different terrestrial arthropod groups probably took different terrestrialization pathways. Although this

is difficult to determine from the fossil forms, extant members of the ancient lineages can lend clues as to what these pathways might have been. For example, osmotic concentrations of body fluids in myriapods suggest that this group terrestrialized straight from the sea without an intermediate step in freshwater.

One particular locality which has yielded a disproportionately large amount of data on early terrestrial ecosystems is the Devonian Rhynie Chert Lagerstätte in Aberdeenshire, Scotland. Here, the silica-mineralizing action of hot spring and geyser fluids captured, in three-dimensional detail, early terrestrial arthropods, such as arachnids, insects, and centipedes, as well as aquatic forms, such as freshwater crustaceans and euthycarcinoids. The remarkable morphological stasis in early terrestrial arthropods, in relation to extant forms, suggests that arthropod-dominated terrestrial ecosystems stabilized early on in the Earth's history and that soil and litter habitats and inhabitants have changed very little ever since.

One of the novel innovations of arthropods was the evolution of muscle-powered flight in the insects (*see* **Fossil Invertebrates:** Insects). The earliest insects or apterygotes (insects without wings) first appear in sediments of Devonian age in the form of collembolans and thysanurans. In earliest Carboniferous times, the major expansion and evolution of land-based flora provided new ecological niches available to this highly adaptable group. The advent of insects and the development of flight led to the colonization of an entirely new ecospace. Interestingly, the earliest non-scorpion arachnids appear to have relied solely on poisoned fangs to capture and subdue prey items. Orb web-weaving using spider silk could only have developed in response to the presence of aerial insects.

By Lower Carboniferous times, the first winged insects had developed and an explosion in the diversity of insects occurred due to coal swamp forest ecosystems. Preservation and discovery are aided by the presence of siderite concretions in the roof shales to many coal seams which, in turn, are commercially exploitable. During the Middle Carboniferous, insects became diverse and widespread for the first time in their fossil record. This trend continued through the Carboniferous and into the Permian period, whereupon an explosion in the total diversity of fossil insects took place. The depositional environments in which these fossils are found do not differ significantly, suggesting a real increase in diversity. During the Jurassic and Cretaceous, favourable shallow (epieric) sea conditions saw the emergence and dominance of decapod crustaceans as a major player in nearshore marine ecosystems. This development within the Crustacea continues to the present day, with arthropods still developing new and successive waves of colonization, particularly in terms of the terrestrialization of some decapod crustaceans, such as the land crabs.

Respiration

From the same basic body plans, different solutions to the challenges of respiration in aquatic and terrestrial environments have evolved. In the smallest of microarthropods, cutaneous respiration, by which the skin acts as a gas exchange surface, takes place. The ratio of surface area to internal volume limits the maximum size of arthropod which can utilise this method. In larger aquatic arthropods, respiratory gas exchange is through structures termed 'book gills'. These thin lamellate cuticular structures are bathed in oxygen-containing water, either wholly external to the body, as in the gill branches of trilobites, or in a specialized cavity, as in the horseshoe crabs (Figure 2). Terrestrial arthropods show adaptation to the particular challenges of living in an environment in which the constant danger of desiccation means that

Figure 2 The xiphosuran, *Rolfeia* sp., from the Lower Carboniferous of Mumbie Quarry, Glencartholm, Dumfries and Galloway, Scotland (×1.1). © The Trustees of the National Museums of Scotland, Edinburgh, UK (NMS G.1998.35.3).

book gills would cease to function. In terrestrial arthropods, respiration is via book lungs (analogous and probably derived in part from book gills) or by a tracheal system. The lamellate structure of book lungs requires additional supporting struts so that they do not collapse in air. Furthermore, the book lungs tend to be positioned internally within the body of the animal as a means of further reducing water loss through desiccation. The tracheal system is a passive method of diffusing oxygen into the body of the arthropod via a branching system of tubes opening on the surface of the exoskeleton at a tracheole. The insects show the greatest development of the tracheal system, related no doubt to the oxygen demands of muscle-powered flight. Other uniramians, such as millipedes, also respire in a similar manner through the presence of a series of paired spiracles along the length of the body. The earliest example of this form of respiration is recorded in the mid-Silurian millipede, *Pneumodesmus newmani*, from Stonehaven, Scotland.

The diffusive process involved in tracheal respiration places an upper limit on the maximum size of terrestrial arthropods which utilize this method, unless a means of secondary 'pumping' derived from muscle contraction can improve efficiency. Book gills, on the other hand, operate in water and are not subject to the same constraints on size which are faced by a land-based book lung system. In some respects, this may help to explain why the largest arthropod fossils found are those of aquatic organisms, such as the extinct eurypterids. Indeed, some very active arachnids, such as solfugids (camel spiders), possess both book lung and tracheal systems, allowing them to meet the oxygen demands of their highly active and voracious life style. The antiquity of the development of tracheal respiratory systems in terrestrial arthropods stretches back as far as the earliest opilionid (harvestman spider) found in the Early Devonian Rhynie Chert, Aberdeenshire, Scotland.

Reproduction

Aquatic and terrestrial environments present different challenges to reproductive strategies in arthropods. Terrestrial arachnids tackled the problem of the desiccation of reproductive material by producing a spermatophore, which is either passed from male to female using specially adapted leg appendages or pedipalps (spiders), or left on the ground attached to a stalk (scorpions).

Internal fertilization is practised by a few groups, including the opilionids (harvestman spiders), and the preservation of internalized organs in arachnids, such as the Rhynie Chert opilionids, in which penis and ovipositor structures are known in separate specimens, demonstrates morphological conservatism.

Direct morphological evidence of sexual dimorphism, or reproductive structures, has long been identified in the chelicerate arthropod group, the eurypterids. In these arthropods, Type A and Type B appendages have been identified on the ventral surfaces. A recent study went further and demonstrated that the longer Type A appendage was probably female, whilst the Type B appendage belonged to the male. An elaborate system of flexure along predefined sutures in the cuticle allowed the female to retrieve an unstalked spermatophore from the substrate prior to storage in the spermathecae.

Feeding

An effective demonstration of how the understanding of the form and feeding function of extant arthropods could be 'reverse engineered' to understand those of the fossil forms was provided by one particular example: the Lipostracan crustacean, *Lepidocaris*, known from three-dimensional fossils in the Early Devonian Rhynie Chert.

The most recent synthesis of feeding in trilobites (*see* **Fossil Invertebrates:** Trilobites) suggested that, as a group, the various trilobites employed different strategies. Indeed, it was argued that much of the variety of form seen in the exoskeletal structure of trilobites was probably driven in response to specific feeding modes (Figure 3). In the main, the structures most responsive to this driving factor were the cephalon (or head region) and the hypostome, a small plate lying anterior to the mouth cavity and variously free or fixed depending on the feeding strategy. The primitive mode of feeding in trilobites was identified as predatory or scavenging, with cephalon and accompanying hypostome morphology (coupled with overall body shape) used as a guide to indicate possible feeding strategies within different trilobite groups. However, this may even be a relict from the arthropods ancestral to the trilobites.

On the other hand, detritivorous trilobites possessed a hypostome detached from the doublure of the cephalon. The glabella (the axially aligned raised area atop the cephalon), where the stomach was thought to reside, would be relatively small in these forms, reflecting the small quantities of food processed at any one time. Filter-feeding trilobites exhibit a different overall morphology again, and often possess a highly vaulted cephalon which would have allowed the leg appendages to stir into suspension food-bearing sediment particles. The demonstration of an elegant scheme such as this in a reasonably

Figure 3 The trilobite, *Phacops* sp., from the Devonian of Morocco, North Africa (×1.7). © The Trustees of the National Museums of Scotland, Edinburgh, UK (NMS G.1987.20.4).

well-known fossil group should be applied to the less well-known Palaeozoic arthropods.

Relevance

Arthropod fossils are used by palaeontologists and geologists in a number of different ways for different reasons. In rocks of Palaeozoic age, the trilobites have long been used as biostratigraphical marker fossils. They find application in this particular field of study as they are readily fossilized (carbonate-impregnated exoskeleton), numerically abundant (each individual moults many times over a lifetime, 'amplifying' their number of potential fossils), and show sufficiently rapid speciation events to make them useful as index fossils. In Early Palaeozoic rocks, nektonic species are useful in palaeogeographical reconstructions of former oceans, seas, and, by default, landmasses; the nektonic life style has led to a widespread occurrence of the resultant fossil remains, ideal for broad stratigraphical correlations. Another use for the numerically abundant trilobites in a stratigraphical framework is the evaluation of evolutionary rates within well-defined lineages. Such frameworks allow trilobites to be utilized in studies of evolutionary changes within a species, up until the point at which a new species evolves, a field of study known as micro-evolution.

Finally, many micro-palaeontological samples, such as fossils of ostracodes (*see* **Microfossils:** Ostracoda), are similarly numerically abundant and readily preserved, and have found application in palaeontological and archaeological studies of past environments due to their environmental sensitivity. The calcium carbonate-impregnated shells of these small bivalve-carapaced crustaceans are a common constituent in rocks of all ages from the Cambrian onwards.

Visual Systems

Two main types of eyes are recognized in the arthropods: compound (multifaceted) and simple ocellar eyes (Figure 4). Compound eyes probably developed separately in a number of different arthropod groups, giving rise to analogous if not homologous structure. Many arthropods possess both compound and simple eyes, as demonstrated by the extant chelicerate horseshoe crab, *Limulus*. The ocellar eyes are positioned on the top of the dorsal surface of the carapace and are primarily used to detect light levels, indicating whether or not the top of the carapace is covered by sand; this information is of particular use to burrowing arthropods. The compound eyes of trilobites are further subdivided into holochroal and schizochroal forms, with the schizochroal structure developing as a result of paedomorphosis. Holochroal trilobite eyes consist of many round or polygonal lenses, the edges of which are all in contact and covered by a single corneal membrane. Schizochroal eyes are a unique system restricted to certain trilobite groups (Phacopina). The lenses are large and separated from one another. Each separate lens has its own separate corneal covering. Mounting the eye lens on upstanding and curved turrets may have allowed certain trilobite species a full 360° field of vision.

Trace Fossils

The combination of rigid appendages and the ability to be motile in a wide range of ecological habitats has provided a rich and diverse trace fossil record which can be assigned to the activities of arthropods. Trace fossils can indicate the presence of arthropods in depositional environments in which preservation would normally fail, and body fossils are non-existent. A good example of this is the Permian Coconino Sandstone Formation of Arizona. Various studies of the desert sandstones have revealed the presence of scorpion and tarantula trackways, the body fossils of which are absent from these environments.

Figure 4 The enigmatic arthropod, *Ainiktozoon loganense*, from the Silurian Fossil Konservat Lagerstatten of Lesmahagow, Lanark shire, Scotland (×1.5). © The Trustees of the National Museums of Scotland, Edinburgh, UK (NMS G.1996.40.1).

Distinctive trace fossils, such as *Kouphichnium*, can reveal the presence of xiphosuran arthropods (horseshoe crabs) and, by analogy with their extant relatives, such as *Limulus*, can be used to imply near-shore or freshwater palaeoenvironments in otherwise sediments devoid of body fossils. Trace fossils, when found in conjunction with body fossil assemblages, may allow behaviour to be interpreted, such as the speed of movement or different actions (nesting, foraging, etc.). Of particular importance in the fossil record of insects is the formation of leaf mines and bite marks in accompanying floral assemblages, giving a clear indication of an escalating arms race between plants and arthropods.

Final Remarks

In conclusion, the Arthropoda can be marked out as one of the most important groups of organisms on the Earth. Their fossil record spans from the earliest recognizable invertebrate faunas right up until the present day. Their numerical abundance, both in the present and in the past, positions them as the major players in ecological structure: predators and prey in all three major Earth environments (land, sea, and air). The group provides excellent opportunities to investigate the effects of extinction events, adaptive radiation, macro- and micro-evolution, and the consequences of palaeobiology in extinct groups. Understanding the relationships between arthropods and other plant and animal groups remains important. Arthropods were an important part of the food chain in the past, just as they are today. Further advances in DNA sequencing and other types of molecular data will no doubt increase our knowledge of the inter-relationships of this diverse and fascinating group.

See Also

Evolution. **Fossil Invertebrates:** Trilobites; Insects. **Lagerstätten**. **Microfossils:** Ostracoda. **Palaeoecology**. **Trace Fossils**.

Further Reading

Allen KC and Briggs DEG (1989) *Evolution and the Fossil Record*. London: Belhaven Press, Pinter Publishers.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman and Hall.

Braddy SJ and Dunlop JA (1997) The functional morphology of mating in the Silurian eurypterid, *Baltoeurypterus tetragonophthalmus* (Fischer, 1839). *The Zoological Journal of the Linnean Society* 121: 435 461.

Briggs DEG and Crowther PR (eds.) (1990) *Palaeobiology: A Synthesis*. Cambridge: Cambridge University Press.

Edgecombe DG, Wilson GDF, Colgan DJ, Gray MR, and Cassis G (2000) Arthropod Cladistics: combined analysis of histone H3 and u2 snRNA sequences and morphology. *Cladistics* 16: 155 203.

Fortey RA and Owens RM (1999) Feeding habits in trilobites. *Palaeontology* 43: 429 465.
Fryer G (1985) Structure and habits of living branchiopod crustaceans and their bearing on the interpretation of fossil forms. *Transactions of the Royal Society of Edinburgh: Earth Sciences* 76: 103 113.
Rasnitsyn AP and Quicke DLJ (2002) *History of Insects*. Dordrecht, Boston, London: Kluwer Academic Publishers.
Willmer P (1993) *Invertebrate Relationships: Patterns in Animal Evolution*. Cambridge: Cambridge University Press.

Trilobites

A W A Rushton, The Natural History Museum, London, UK

Introduction

The trilobites are a large group of extinct marine arthropods, the Class Trilobita, characterized by the longitudinal division of the exoskeleton into three lobes – a convex central (axial) lobe flanked on each side by a flatter pleural area (Figure 1) – hence their name. Trilobites were widely distributed throughout the Palaeozoic era, first appearing in the later part of the Early Cambrian period (at about 520 Ma) and disappearing just prior to the end-Permian mass extinction (*ca*. 250 Ma). During their 270 million year existence, trilobites spread worldwide, occupied a wide range of marine habitats, and evolved a great variety of morphologies.

Trilobites are a relatively well-studied group and a good deal is known, or inferred, about their life, growth, and activities. Their extensive fossil record has provided illuminating case histories in evolution and ecological adaptation, and, in addition, they have proved to be of great value in stratigraphy, especially in the Cambrian and Ordovician periods, and in interpreting Palaeozoic palaeobiogeography.

Thanks to their striking appearance, strange-looking yet evidently organic in origin, trilobites have always been attractive fossils. Specimens have been recorded in prehistoric burial sites and they have found a place in folk-lore; in more recent times, avid collectors have built up magnificent collections of fine specimens, some of which have commanded high prices. The scientific study of trilobites commenced at the end of the seventeenth century, and, during the eighteenth and nineteenth centuries, European workers, including Linnaeus, published descriptions and illustrations. The benchmark was Barrande's illustrated account of the Bohemian trilobites, published in 1852, that showed their astonishing variety and stimulated interest in the group amongst all scientific communities; this interest continues to this day, such that more than 5000 genera have now been described and a huge number of species named, perhaps approaching 20 000.

Form of the Exoskeleton

The general features of the exoskeleton are labelled in Figure 2. The exoskeleton typically has an oval outline in plan view, and consists of a cephalon (headshield), furnished with a pair of compound eyes, a jointed thorax of few to many segments, and a pygidium (tail-piece) made up of fused segments. Typically the animal was 2–10 cm in length, but a species of *Acanthopleurella* is just over 1 mm in length when fully grown, and the largest known *Isotelus* is 72 cm in length.

The trilobite exoskeleton, unlike that of many arthropods, was mineralized by calcite and phosphate, thus greatly increasing its potential for preservation

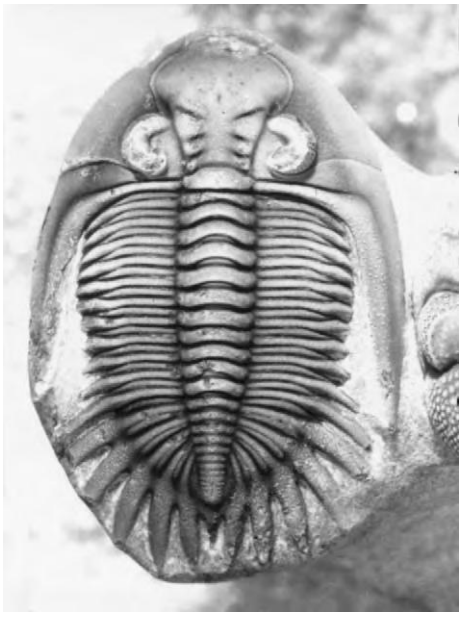


Figure 1 The dorsal exoskeleton of *Neometacryphaeus* (sub order Phacopina) from the Devonian of Morocco. About natural size. Courtesy of Professor R. A. Fortey.

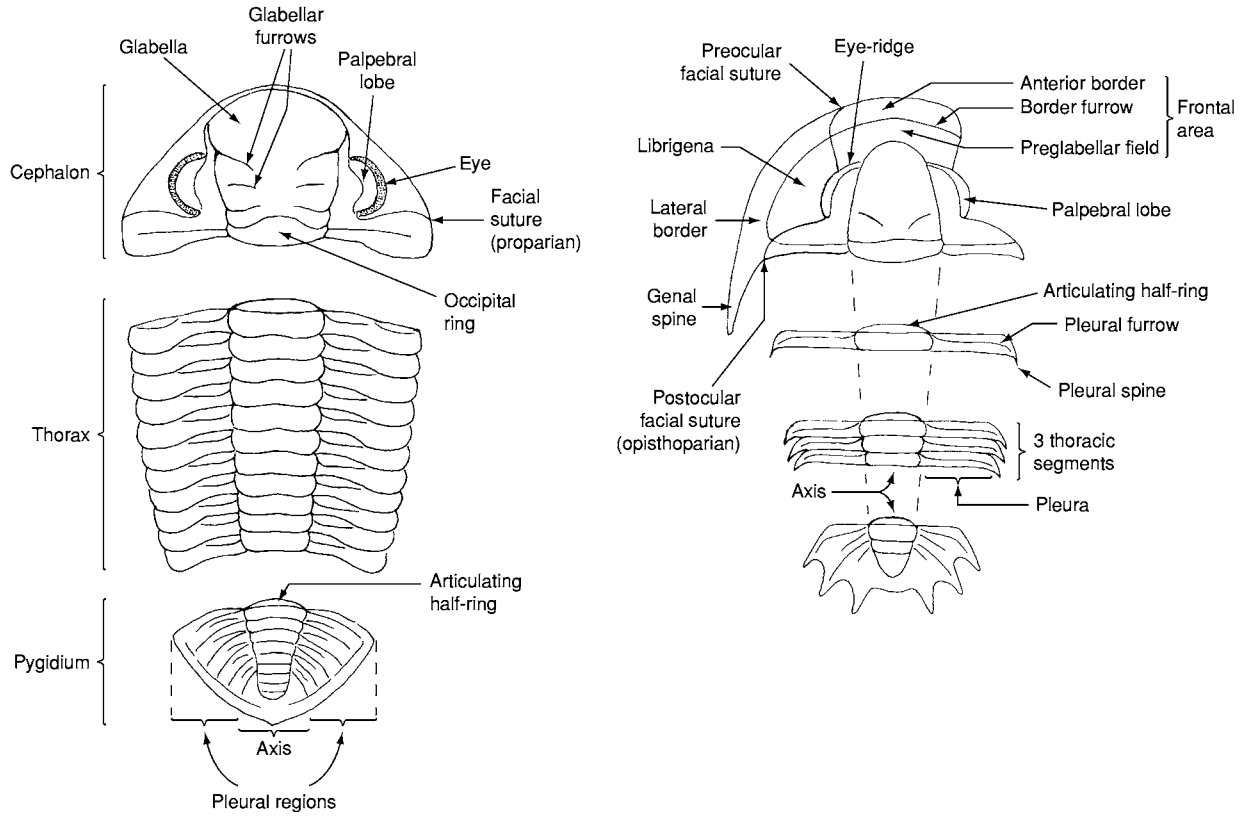


Figure 2 Terminology based on the phacopine *Acaste* and the ptychopariid *Marjumia*.

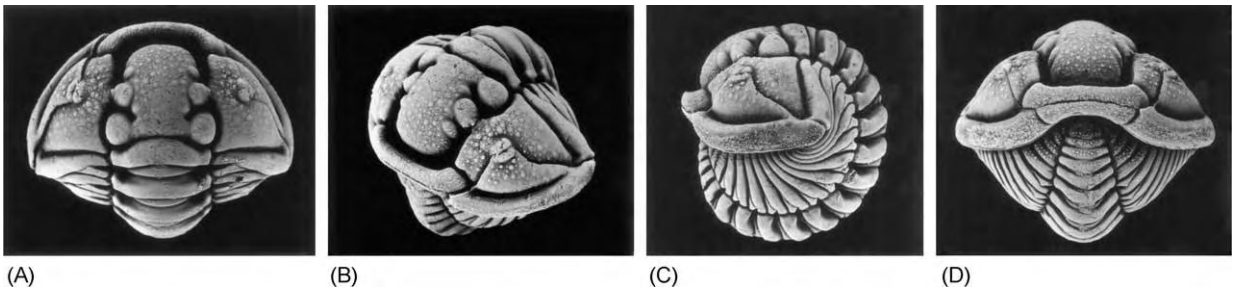


Figure 3 An enrolled *Calymene* from the Silurian. Width of cephalon about 3 cm. Courtesy of Dr Derek Siveter.

and leading to the good fossil record of the class. In contrast, the appendages (antennae and limbs) were not mineralized and are found only rarely, in conditions of special preservation, and the ventral membrane is scarcely known. The depressed form of many early trilobites suggests that they originated as benthic forms, with the unprotected ventral side on the substrate, and the dorsal shield protecting the animal from above. They subsequently evolved a great variety of forms that allowed a range of feeding strategies to be employed and enabled the occupation of many marine environments. So great is their variety that almost every generalization given here could be modified by some exception.

The dorsal exoskeleton is more or less vaulted, especially the axial lobe, which housed most of the musculature, the alimentary tract, and the main circulatory and nervous conduits. The pleural lobes are less convex and served to cover the limbs, namely the walking legs and the gill-filaments. The thoracic segments articulate with each other, so that many trilobites could roll up, thus protecting the ventral side (Figure 3). The outer edge of the exoskeleton is recurved under the body to form a rim, the doublure (Figure 4), on which the animal could rest on the seafloor.

In most trilobites, the cephalon is divided by the facial sutures into three main sclerites: the cranidium,

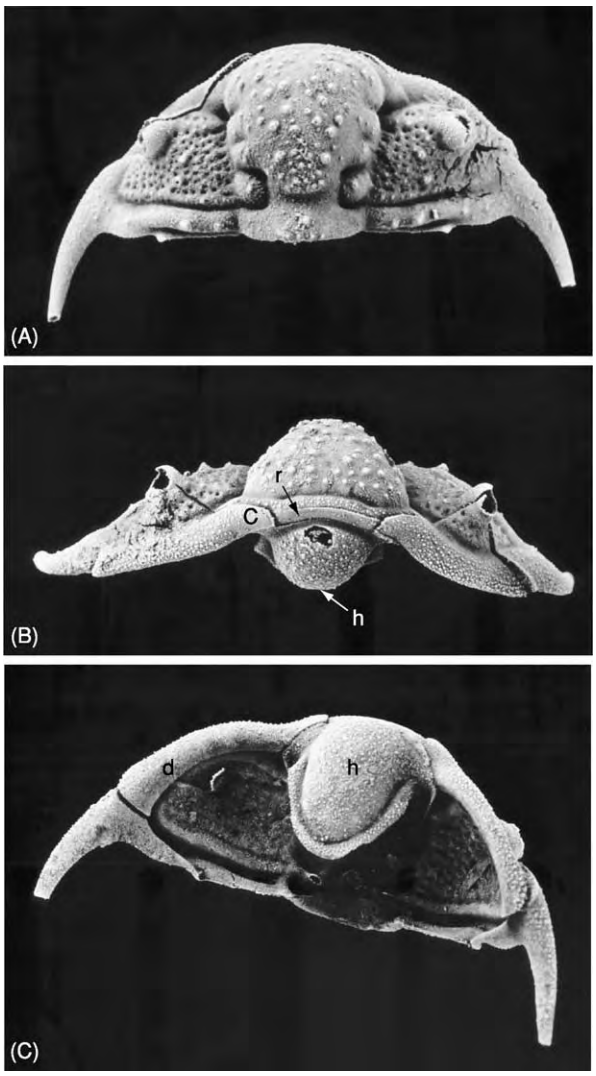


Figure 4 A silicified cephalon, about 8 mm in length, of *Cer aurus* (Cheirurina) from the mid Ordovician: views from top (A), front (B), and below (C). The hypostome (h) underlies the fore part of the glabella and its anterior edge is set against a narrow rostrum (r). Reproduced with permission from Whittington HB (1992) *Fossils Illustrated 2. Trilobites.* Woodbridge: Boydell Press.

flanked by a pair of librigenae. The glabella, whose shape and lateral furrows are important features for classification, housed the fore-gut and stomach. The posterior part of the glabella, the occipital ring, articulates with the thorax and commonly mimics features of the thoracic axis. The palpebral lobes overlay the eyes, but the ocular surface was commonly attached to the librigena, and the facial suture almost invariably passed around the upper edge of the eye to facilitate moulting of the exoskeleton, whilst avoiding damage to the ocular surface. The course of the preocular and postocular sutures is very varied. The preocular sutures curve inwards to meet in front of the cranidium and commonly bound a separate mineralized plate, the rostrum (not always present), that underlies the frontal margin (Figures 4 and 5). Underneath the glabella is the hypostome, a mineralized plate lying anterior of the mouth and under the fore-gut. The anterior end of the hypostome was either set rigidly against the rostrum or doublure (the conterminant condition, Figures 4 and 5B; sometimes modified to impendent, Figure 5C), or was separated from the rostrum and supported only by musculature and the ventral membrane (the natant condition; Figure 5A). In some instances, the hypostome was docked against the rostrum, but was not rigidly attached to it; it may have been joined by a sort of hinge of unmineralized integument.

The thorax is made up of several more or less similar segments, commonly 8–12 in number, but as few as two in some forms and more than 40 in others. Sometimes one or more segments show special adaptations, such as axial spines or macropleurae. Each axial ring has an articulating half-ring that fits under the segment in front and, during enrollment, they protect parts of the axial lobe from being exposed; at the same time, the pleural tips of each segment can slide under the pleurae in front (Figure 3).

The pygidium is composed of few to many fused segments and its size varies from tiny to larger than the cephalon (Figures 6A and 6B). In some pygidia, the segmentation remains obvious, but others are well fused so that the segmentation is not visible. During enrollment, the underside of the pygidium could, if it was small, be tucked into the cephalic chamber; alternatively, it could be pressed against the underside of the cephalon, the margin of which commonly shows pits or grooves – coaptive structures – that allow a perfect fit.

The trilobite exoskeleton may be smooth, granulose or tuberculate, pitted (Figure 4A), or covered with terrace lines (the latter being especially characteristic of the ventral doublure; Figure 7). Spines may be developed almost anywhere and some species are extravagantly spiny (Figure 17B). Many exoskeletal features, such as tubercles, were perforated with pores that might have housed setae and presumably had sensory functions. There is a trend in many major trilobite groups for some species to become 'effaced', whereby the axial, glabellar, and border furrows are smoothed out and the 'trilobed' appearance is lost (Figure 6).

Eyes

Most trilobites had well-developed eyes, those from the Early Cambrian being amongst the earliest eyes known in the fossil record. The eyes are generally placed high on the fixigenae and, in certain forms,

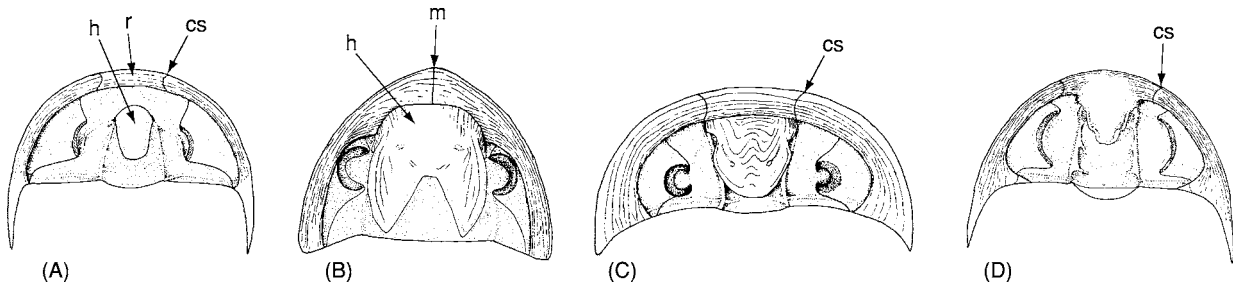


Figure 5 Ventral views of cephala. (A) *Elrathia* (Ptychopariina), with simple natant hypostome (h) and rostrum (r) separated from the librigenae by connective sutures (cs). (B) *Isotelus* (Asaphina), with forked conterminant hypostome (h) and a median suture (m) in place of the rostrum and connective sutures. (C) *Raymondaspis* (Styginina), with impendent hypostome, similar to the conterminant condition, but the anterior edge of the glabella does not correspond to that of the hypostome. (D) *Fieldaspis* (Corynexochida), in which the hypostome and rostrum are fused into a single plate; cs, connective sutures. Reproduced with permission from Fortey RA (1990) Ontogeny, hypostome attachment and trilobite classification. *Palaeontology* 33: 529 576.

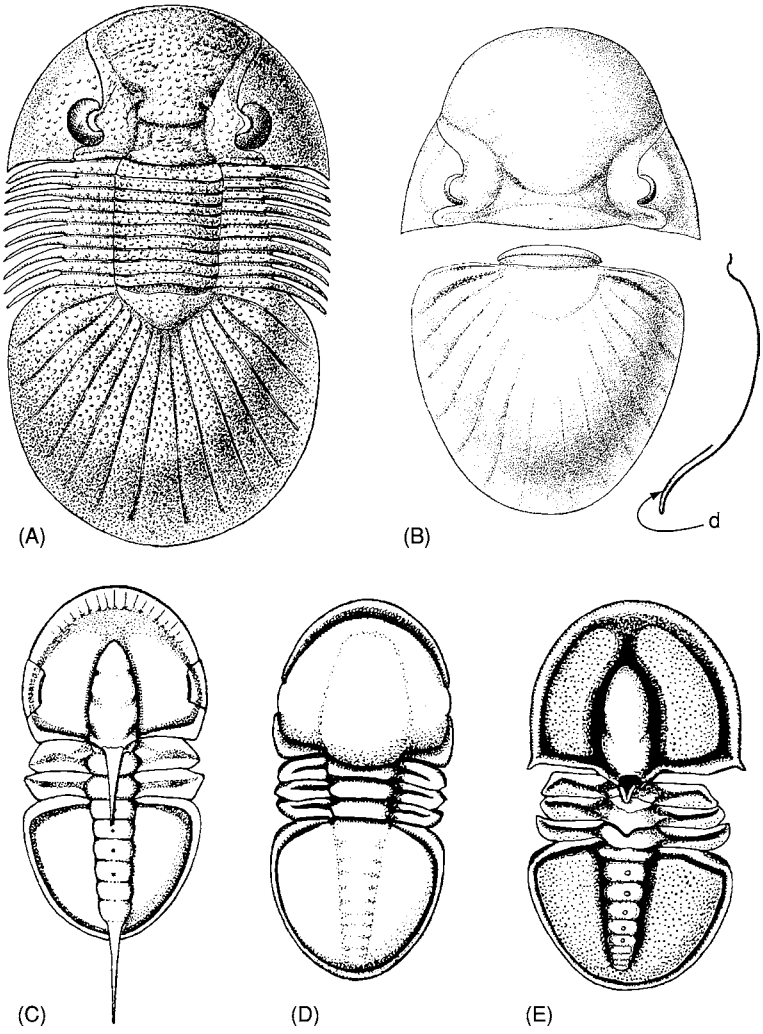


Figure 6 Effacement of external features in Styginina (A, B) and Eodiscina (C E). (A) *Scutellum*, Devonian, 10 cm in length. (B) *Paralejurus*, Devonian, cephalon (5 cm in length) and pygidium with longitudinal profile, showing doublure (d). (C) *Pagetia*, mid Cambrian. (D) *Pagetiellus*, Early Cambrian, cephalic and pygidial axis effaced, eyes marginal. (E) *Eodiscus*, mid Cambrian, not effaced, but eyes absent. (C E) each about 8 mm in length.

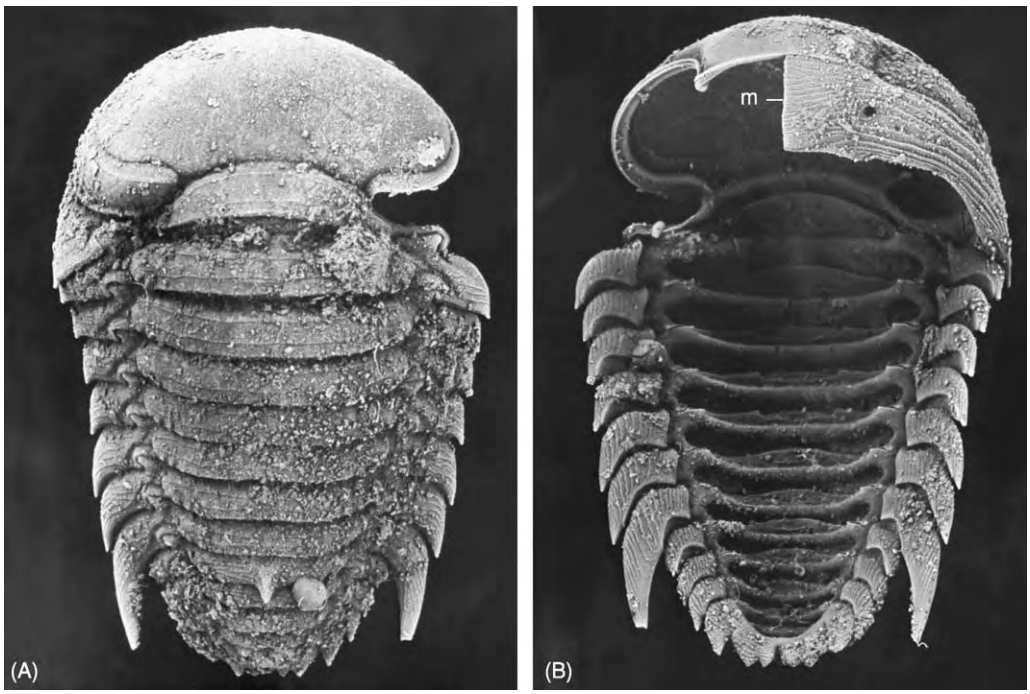


Figure 7 Dorsal and ventral views of a silicified dorsal exoskeleton of *Remopleurides* (Asaphina), with one librigena in place. Length about 20 mm. The ventral view shows the median suture and terrace lines striating the doublure. Reproduced with permission from Whittington HB (1992) *Fossils Illustrated 2. Trilobites.* Woodbridge: Boydell Press.

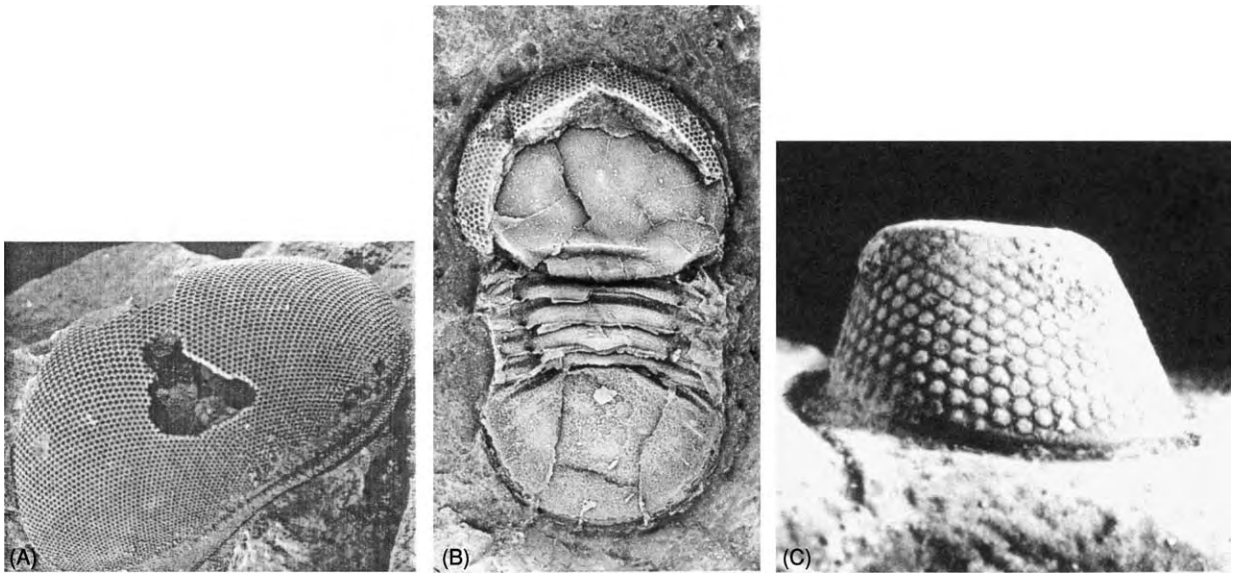


Figure 8 (A, B) Holochroal eye in Cyclopygidae (Asaphina). (A) *Pricyclopyge* left eye, about 10 mm in length. (B) *Microparia lusca*, a species in which the eyes have grown forward and joined up in front. (C) Schizochroal eye of *Dalmanites* (Phacopina), side view of right eye, about 5 mm in length; each lens is about 0.25 mm in diameter. Reproduced with permission from Treatise O, revised vol. 1, Owens RM (2002) Cyclopygid trilobites from the Ordovician Builth Llandrindod inlier, Central Wales. *Palaeontology* 45: 469 485, and Clarkson ENK (1966) Schizochroal eyes and vision in some phacopid trilobites. *Palaeontology* 9: 464 487.

lie at the summit of eye-stalks. Being arcuate in plan, they afforded a wide field of view. Most trilobites had eyes consisting of many small contiguous lenses covered by a cornea of calcite (holochroal eyes; Figure 8A), whereas the suborder Phacopina had highly evolved eyes with fewer but larger biconvex lenses, each with their own cornea, that were separated by interlensar sclera (schizochroal eyes;

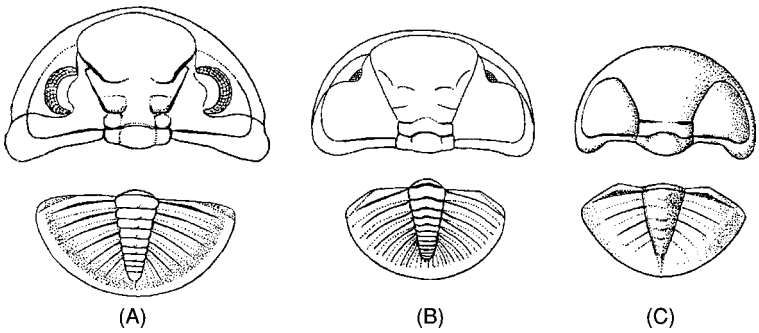


Figure 9 Reduction and loss of eyes in some Phacopidellinae (Phacopina) from the Silurian and Devonian. (A) *Phacopidella*, ×1. (B) *Denkmannites*, ×1, with reduced eyes. (C) *Ductina*, ×3, blind.

Figure 8C). These lenses were composed of calcite and a study of their optics has shown that they had features to overcome birefringence and chromatic aberration.

In pelagic or mesopelagic trilobites, the holochroal eyes became enlarged and bulbous, even merging in front of the glabella, giving all-round vision (Figure 8B). In contrast, secondary eye reduction or loss is present in various groups of trilobites, and is commonly associated with species from deep-water settings. Some lineages have been found that show progressive reduction of eyes, leading to their loss, and blindness is a feature of species in many major clades (Figures 6E and 9).

Limbs

The ventral limbs of trilobites are known only in about a dozen taxa, but these exemplify several major groups of various ages from the Cambrian to Devonian. The anterior appendages were a pair of uniramous antennae, all the succeeding appendages being deeply biramous and uniform in structure, apart from a pair of posterior antenniform cerci known only in *Olenoides*. There were generally three pairs of biramous limbs under the posterior part of the cephalon (but four pairs have been claimed in one or two instances), and one under each of the thoracic and pygidial segments (Figure 10). The coxa at the base of the limb is spinose and presumably acted as a gnathobase. Attached to the coxa was a lower branch, the endite, of seven podomeres (joints), including a small terminal claw; this branch was used for walking and digging. It is inferred that the hinge between the coxa and the ventral side of the exoskeleton allowed forward and backward movement of the endite, whilst the joints between the podomeres allowed the limb to be curled in under the body. The upper branch, the exite, was filamentous and is assumed to have been a gill. It occupied the small space above the endite and under the pleural region (Figure 10).

Other Features

Sites of the insertion of muscles are seen on parts of the exoskeleton, notably as paired furrows on the glabella, but also as apodemes on the ventral side of the axial furrows and, more sporadically, as pits on the pygidial axis; they may also appear as colour patterns, and may thus be visible in effaced trilobites in which the glabellar furrows are obscure. Many trilobites, especially in the Cambrian, show caeca, i.e., traces of fine anastomosing ridges that cross the preglabellar and genal fields (Figure 11). They have been regarded as an exterior expression of part of the respiratory system or, alternatively, diverticula of the digestive tract.

Two trinucleid trilobites have been found that, thanks to freakish weathering, show the alimentary tract extending back from the glabella, along the axial region to the pygidium. Wonderfully preserved juvenile examples from Sweden of the highly derived form *Agnostus pisiformis* have revealed not only the peculiar limbs of that species, but also features whose presence had been inferred in trilobites but not otherwise seen, e.g., the mouth, gut, anus, ventral membrane, and minute sensory pits and setae. Several trilobites have been found that show exoskeletal modifications caused by teratology, parasites, healed injury, or partial failure in moulting. Exoskeletal features attributable to sexual dimorphism have not been positively identified. However, bulbous structures present in front of the glabella in certain species have been interpreted as brood-pouches.

Growth

Trilobites, in common with other arthropods, grew by ecdysis (moulting), so that any animal that reached

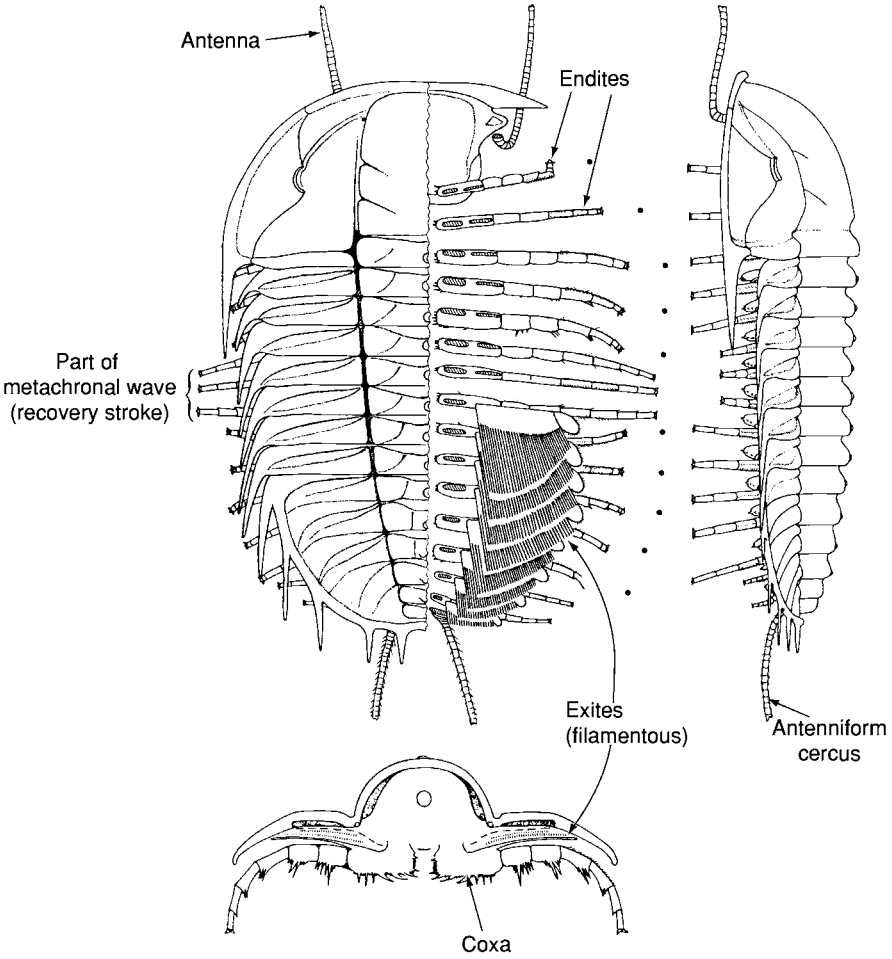


Figure 10 Reconstruction of *Olenoides* (Corynexochida), showing limbs and supposed gait when walking. Sagittal length about 7 cm.

Figure 11 Cranidium and librigena of *Parabolina* (Olenina), showing genal caeca. Length of cranidium, 10 mm. Reproduced from Rushton AWA (1982) The biostratigraphy and correlation of the Merioneth Tremadoc Series boundary in North Wales. In the Cambrian Ordovician Boundary: © National Museum of Wales, Geological Series No. 3, Cardiff.

maturity could have left remains of several exoskeletons; there are records of bedding planes with large accumulations of moulted remains. Most often the exuvia became disarranged, being preserved merely as scattered sclerites, but many examples have been found that show arrangements suggestive of the moulting strategies employed by various species. In some instances, the facial sutures opened and the librigenae were sloughed off separately from the cranidium; in others, such as the phacopids, the thoracopygon (thorax + pygidium) may be found only a little disarranged, whilst the cephalon or cranidium is found nearby, separated and overturned; in such a case, the animal crept out from the cast shell by way of the split between the cephalon and thorax (Figure 12).

The earliest growth stage, the protaspis, is a single shield commonly a fraction of a millimetre in length. Later moults show the developing cranidium and the addition of a protopygidium with fused thoracic segments (Figure 13C); successive segments were generated at the posterior end of the pygidium and, as they grew forward, were released into the thorax (meraspid stages) until the full complement of segments was achieved (holaspid stage), as shown in Figure 13. Full or partial series of growth stages

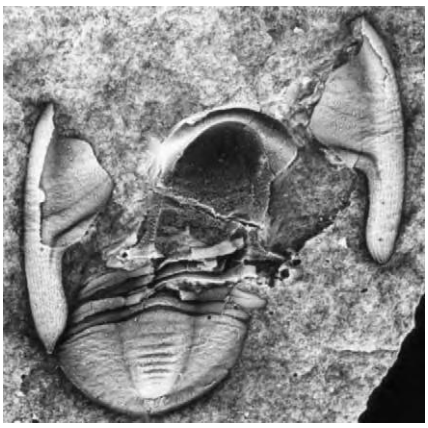


Figure 12 Moulted exoskeleton of *Spatulina* (Proetida) from the Early Carboniferous. The librigenae are cast off, the thorax is slightly disarranged, and the cranidium is inverted in relation to the other parts; the hypostome is missing. Sagittal length, as preserved, about 10 mm. Reproduced, with permission from Cambridge University Press, from Owens RM and Tilsley JW (1995) An atheloptic trilobite assemblage from the Carboniferous of North Devon. *Geological Magazine* 132: 713 728.

have been reconstructed for a considerable range of species, and some of the growth patterns are characteristic of major groups, e.g., the Asaphida, in which advanced members of the group have relatively large globular protaspides (Figures 13I and 13J). With all trilobites, there were great changes in the shape as well as the size of the exoskeleton throughout growth, and such changes provided opportunities for the derivation of new morphologies by the operation of heterochronic processes, whereby forms developed sexual maturity whilst still showing juvenile features (paedomorphosis).

Activities

Walking

Tracks made by arthropods, including trilobites, in unconsolidated sediments may be preserved as trace fossils, although it is generally impossible to be sure which animal was responsible for which traces (*see* **Trace Fossils**). It is thought that trilobites, like other arthropods with many walking limbs, crawled forwards using their limbs in a metachronal wave (whereby groups of limbs push forwards whilst intervening groups of limbs make a return stroke, ready to give their next push, as represented in Figure 10). Tracks that are of a size compatible with contemporaneous trilobites have been analysed and are indicative of such motion. There are also trace fossils which show that trilobites could progress by making a series of short hops or jumps.

Digging

Trilobites used their endites to excavate resting 'nests' or to probe the substrate for prey. The bilobed burrowing traces, *Rusophycus*, have been positively associated with trilobites because they sometimes show the outline of the cephalic shield or genal spines, and one example has been found with a complete *Flexicalymene* above it. *Rusophycus* shows transverse scratch marks that were made by the tips of the endites. Complete specimens of calymenids and homalonotids found with the thorax extended in a concave curve may represent animals that died whilst occupying such a resting burrow. The longer bilobed trackways, known as *Cruziana*, are thought by some to represent digging or ploughing by an advancing trilobite, although the means by which the locomotion and digging activities could have been combined has not been worked out.

Swimming

A number of trilobites have been identified on the basis of several lines of evidence as having been adapted to pelagic or mesopelagic habits, all the best documented being Ordovician in age (Figure 14). All have huge eyes with a wide field of vision. Some, like *Carolinites*, have narrow pleural areas and a vaulted axis that gives space for strong musculature; they are found in association with benthic faunas of all facies, being independent of any of them. These are pelagic forms. Other less vaulted and muscular forms, typified by the cyclopygids (Figure 14B), are likewise associated with various faunas, but in their case only in outer-shelf settings. These are regarded as mesopelagic. Certain highly streamlined genera are interpreted as rapidly swimming predators (Figure 14C), but as the limbs are not known in any pelagic form, the actual action of swimming is conjectural.

Feeding

Various morphotypes associated with particular feeding strategies were developed independently at different times in different major trilobite groups. Predators and scavengers tended to be relatively large forms with well-developed eyes and a big conterminant hypostome that is braced against the rostrum, and sometimes fused with it, as in Corynexochids (Figure 5D) and *Paradoxides*. The posterior end of the hypostome, near the mouth, commonly has special features, such as a notch or forked and serrated margin, presumably adaptations to manipulate food (Figure 5B). A large and forwardly expanded glabella and an impendent hypostome characterize advanced forms like *Phacops* and its relatives (Figure 17D).

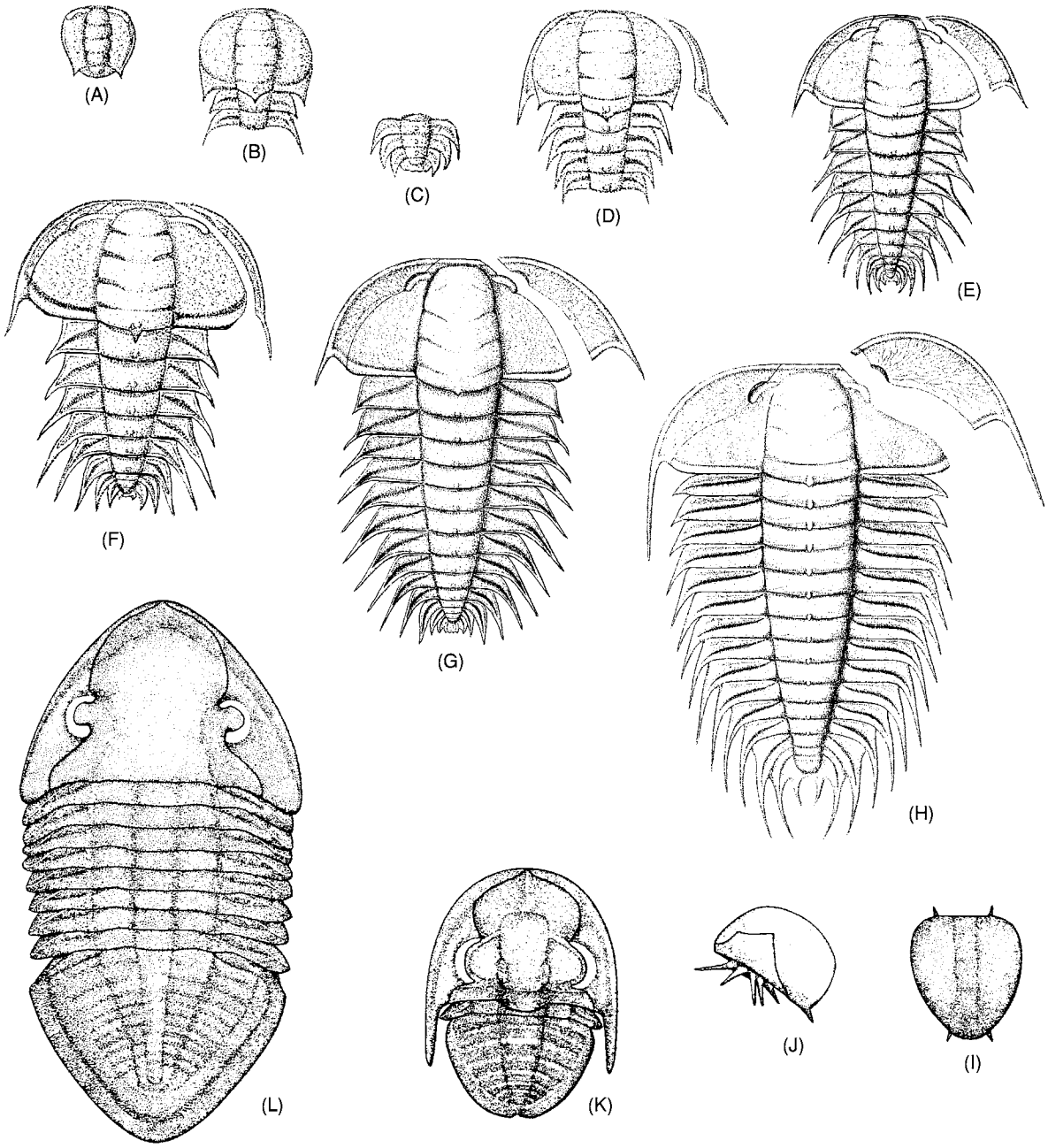


Figure 13 Growth of trilobites, exemplified by *Parabolina* (Olenina) (A H) and *Isotelus* (Asaphina) (I L). *Parabolina*: (A) protaspis, 0.5 mm in length; (B) meraspid cranidium, 0.6 mm in length with two free segments; (C) transitory pygidium (to same scale) with unreleased thoracic segments; (D, E) larger meraspides (to same scale) with four and six free segments; (F, G) meraspides with eight and 10 segments, respectively, 3.3 and 4.5 mm in length; (H) holaspis form with 12 segments, 16 mm in length. *Isotelus*: (I, J) globose protaspis, 0.6 mm in length, top and side views; (J) shows the spinous hypostome; (K) meraspides degree 1, 3 mm in length; (L) holaspis, 60 mm in length. A H Reproduced by permission of the Royal Society of Edinburgh and E. N. K. Clarkson, C. M. Taylor, and P. Ahlberg from 'Ontogeny of the trilobite *Parabolina spinulosa* (Wahlenberg, 1818) from the upper Cambrian Alum Shales of Sweden' by Euan N. K. Clarkson, Cecilia M. Taylor, and Per Ahlberg in *Transactions of the Royal Society of Edinburgh: Earth Sciences*, volume 88 (1997) pp. 69 89. I L Treatise O, revised, vol. 1.

In contrast, detritus feeders tended to be smaller, less active animals with a smaller glabella and relatively wide preglabellar and pleural areas. The hypostome was commonly natant and conservatively ovoid in shape (Figure 5A). An enormous number of Cambrian species of this type have been described and presumably they signify a subtle partitioning of resources; fewer are known from the Ordovician and later periods when such trilobites had to compete with newly evolved Mollusca. An offshoot of

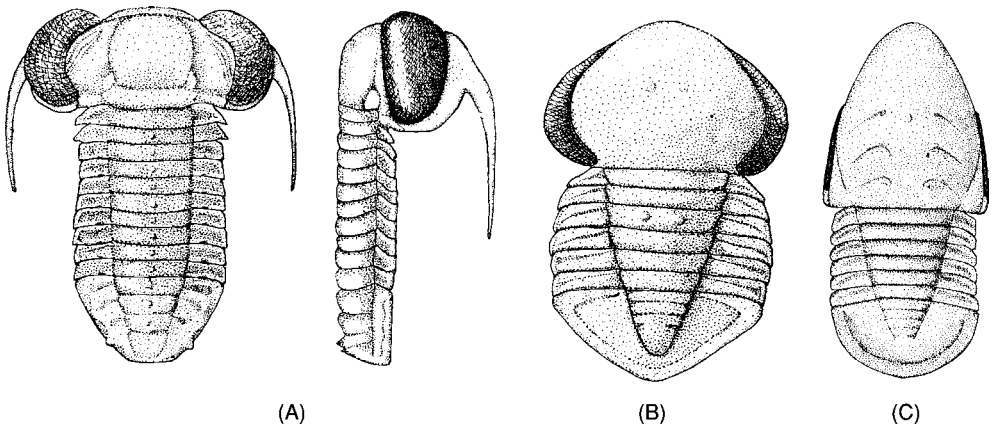


Figure 14 Large eyed trilobites. (A) *Carolinites* (Proetida), top and side views of a pelagic form (about 3 cm in length). (B) *Pricyclopyge*, a broad mesopelagic cyclopygid (Asaphina) about 3 cm in length. (C) *Degamella*, a streamlined cyclopygid, 5 cm in length. Reproduced by permission of the Royal Society of Edinburgh and R. A. Fortey from 'Pelagic trilobites as an example of deducing the life habits of extinct arthropods' by R. A. Fortey in *Transactions of the Royal Society of Edinburgh: Earth Sciences*, volume 76 (1985) pp. 219 230.

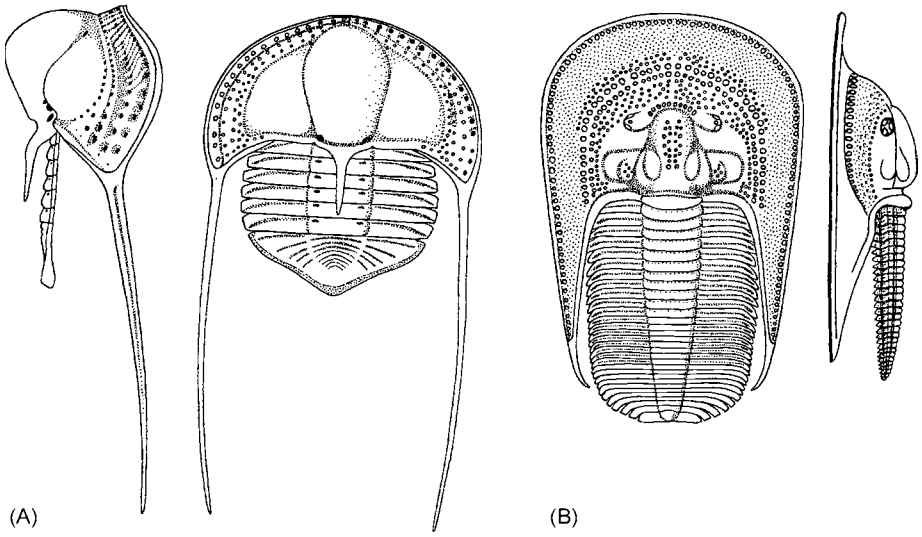


Figure 15 Top and side views: (A) *Cryptolithus* (axial length, 3 cm), an Ordovician trinucleoid in the Order Asaphida; (B) *Harpes* (axial length, 6 cm), a Devonian member of the Harpina, Order Ptychopariida; both show the large cephalon and relatively small thorax that are taken to characterize a chamber filter feeding mode of life. Although not closely related, these groups show a remarkable convergence, both having developed a broad bilamellar fringe on the cephalon. The upper lamella is part of the cranidium; the lower lamella includes the genal spines and is separated from the upper lamella by the facial suture, which extends around the margin of the cephalon.

the detritus feeder type is the possible development in the Olenidae (e.g., Figure 13H) of a chemoautotrophic metabolic strategy. The olenids had a multiplicity of thoracic segments and conservative hypostomes that were degenerate in some instances, and they survived in conditions of unusually low oxidation and high sulphur content. It has been suggested that they fed on symbiotic sulphur bacteria that they cultivated on their exites.

A group of trilobites with large head-shields is regarded as chamber filter feeders (Figure 15). In these, the cephalon is considerably deeper than the thoracopygon and the hypostome is elevated well above the substrate. It is thought that the animal stirred up the sediment in the chamber beneath the cephalon and sorted edible particles from suspension. The best known of these trilobites are the harpetids, whose massive genal prolongations extended the filter chamber backwards alongside the thorax, and the trinucleids. These groups independently developed broad cephalic fringes composed of an upper and a lower lamella that were separated by a facial suture

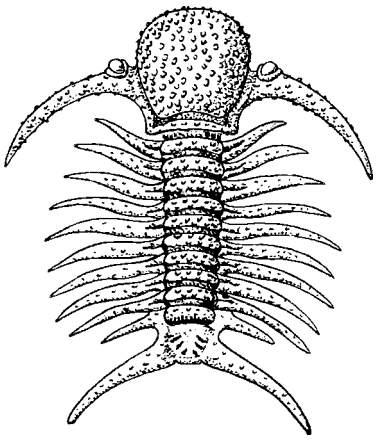


Figure 16 *Deiphon* (Cheirurina), an enigmatic Silurian trilobite, about 3 cm in length.

extending around the margin of the cephalon. Of the trinucleid group, *Cryptolithus* is significant because the limbs are known and *Rusophycus* excavations made by the animal have also been recognized.

Of the pelagic trilobites mentioned above, most were probably plankton feeders; the large streamlined forms (*Degamella*, *Parabarrandia*) were likely to have been predatory upon smaller plankton feeders. Besides the foregoing, there is a host of other forms whose feeding strategies are undetermined, including many species that occupied specialist niches within carbonate mounds, as well as bizarre forms such as *Deiphon* (Figure 16).

Classification

Cladistic analysis shows that Class Trilobita is not primitive in relation to other early arachnomorph arthropods, and the trilobites are taken to be a natural group characterized by the calcified exoskeleton and ocular surface, the facial sutures, and features of the hypostome and pygidium. However, a comprehensive natural classification of the trilobites has not yet been devised. Thousands of genera have been grouped into hundreds of subfamilies and families, but the content and limits of many of these are subjective, and it is debated how they should be grouped into higher categories. Several major natural groups are widely accepted, especially amongst the post-Cambrian trilobites, but the origins of some of them are cryptic and how the various groups are interrelated remains uncertain. Obstacles to classification have included the relatively poor knowledge of the ontogenies of many trilobites, their limbs (and ventral features generally), the tendency for iterative evolutionary trends to yield misleading homeomorphic forms, and the unresolved questions that surround the great mass of plesiomorphic taxa in the Cambrian. None the less, great advances in the description and interpretation of trilobites in the latter half of the twentieth century has led to an improved understanding of the problems, and has yielded such provisional classifications as that presented in outline in Table 1. Cladistic and morphological analyses have provided some unexpected, although fundamentally satisfying, results, e.g., the transfer of the trinucleid group to the Asaphida.

Table 1 Major orders and suborders of trilobites. Some examples are shown in Figure 17

Order	*Suborders and some other major groups*
Agnostida	Agnostina, Eodiscina
Redlichiida	Redlichiina, Olenellina
Corynexochida	Corynexochina, Styginina, Leiostegiina
Lichida	(including superfamily Odontopleuroidea)
Phacopida	Phacopina, Cheirurina, Calymenina
Asaphida	(including superfamilies Asaphoidea, Remopleuridioidea, Dikelokephaloidea, Anomocaroidea, Cyclopygoidea, Trinucleoidea, etc.)
Proetida	
Ptychopariida	Ptychopariina, Olenina, Harpina

Distribution in Time and Space

The distribution of trilobites was determined by the availability of environmental settings with appropriate water depth, temperature, turbulence, substrates, oxygenation, etc. Barriers to migration included oceans at unfavourable temperatures, or too deep and wide to cross. Hence, faunas from shallow continental shelf areas are more endemic than those from continental slopes. This kind of evidence, especially from the most endemic types of trilobites (and other fossils), is exploited to provide lines of evidence, independent of plate tectonics, for reconstructing continental palaeogeographies. Thus, the occurrence of the Gondwanan trilobite *Plaesiocomia* in subsurface Florida, USA, indicated that, during the Ordovician, that state was a fragment of Gondwana, and was only later accreted on to Laurentia.

Cambrian

The earliest history of the trilobites is unknown. The oldest known species appeared in shallow marine settings around several palaeocontinents and are referable to a number of separate clades, indicating that the trilobites had an earlier history (probably as unmineralized or weakly mineralized forms) that is not yet known. One of the earliest to appear is the olenelloid *Fallotaspis* (Figure 17M), a genus

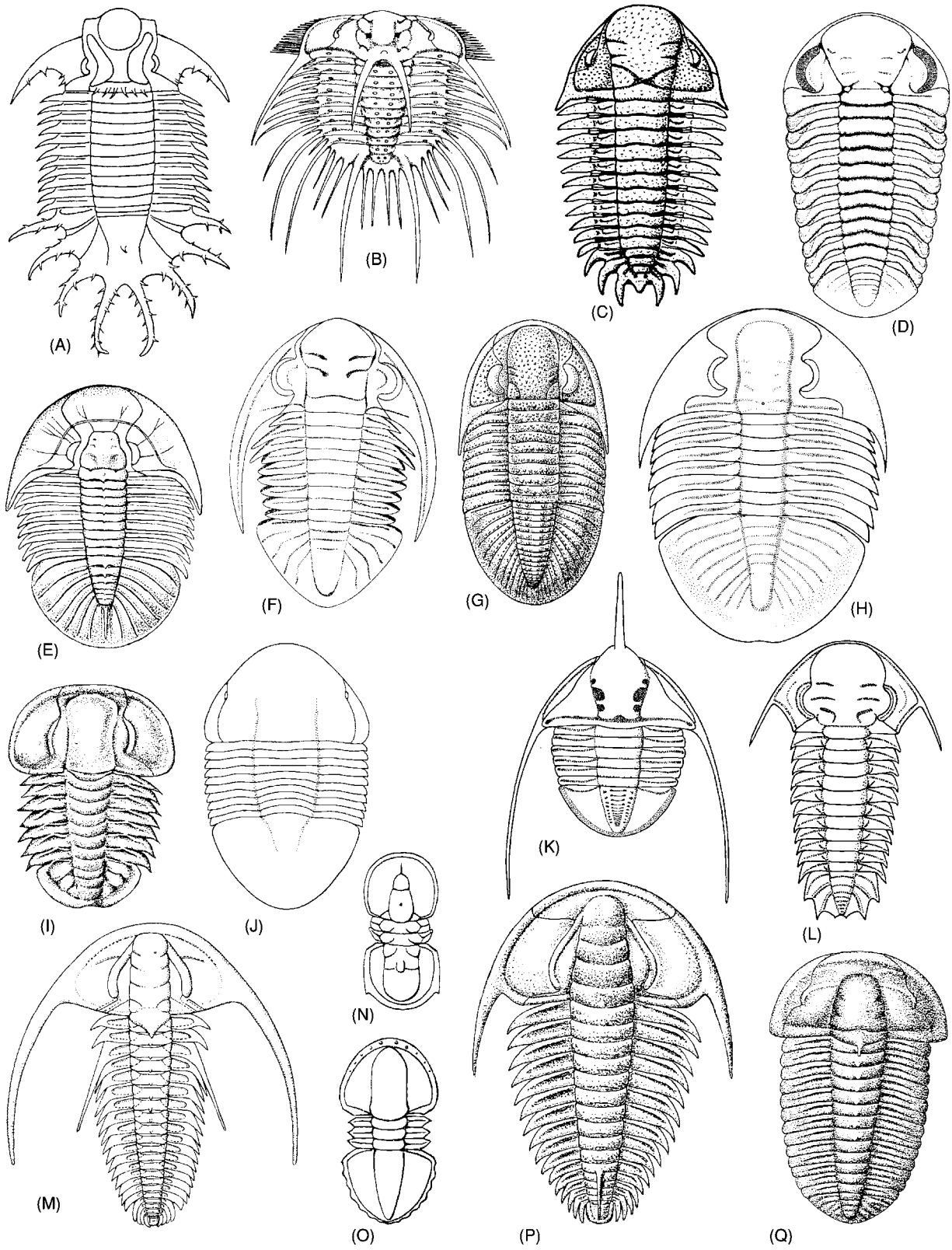


Figure 17 Representatives of some of the principal groups of trilobites. Typical sagittal lengths of each exoskeleton are given. (A, B) Order Lichida this includes the lichids and odontopleurids, which have distinctive cephalic and pygidial structures; although probably related to the Dameselloids, the derivation of the whole group is unknown. (A) *Terataspis* (Lichoidea), Devonian, 500 mm. (B) *Odontopleura* (Odontopleuroidea), Silurian, 40 mm. (C, D) Order Phacopida this order contains the suborders Phacopina,

that makes an early appearance in Siberia, Laurentia, and on the edge of Gondwana (Morocco). The Olenellina went on to radiate particularly in the Early Cambrian of Laurentia. They are not found in south China, where Redlichiida (Figure 17P) dominated, accompanied by oculate eodiscoids, and later by Ptychopariida (e.g., Figure 17Q). The redlichiids extended to Morocco, where they are associated with endemic olenelloids and opisthoparian forms, such as bigotinids and ellipsocephalids. The provinciality of the Early Cambrian shallow-water faunas continued into the Middle Cambrian, but as trilobites colonized deeper water environments, there arose taxa with the capacity to cross from one continental margin to another, e.g., *Arthricocephalus*, certain Agnostida, and *Centropleura*. The Late Cambrian saw an expansion of ptychopariid deposit feeders, with endemic faunas in Laurentia, north and south China, Australia, Kazakhstan, and north-west Siberia, whilst the dysaerobic environment of the 'olenid sea' around Baltica provided conditions that allowed the Olenidae to flourish. Some very widely distributed taxa, such as *Glyptagnostus* and *Irvingella* (Figure 18), which reached all these sites, may have been pelagic.

Ordovician

During the Ordovician, the trilobites reached their acme; the exoskeleton developed an even greater range of modifications than was seen in the Cambrian: well-documented pelagic forms evolved, schizochroal eyes appeared, and some benthic lineages evolved cephalic fringes as an aid to chamber filter feeding. Many Cambrian groups disappeared, but at least as many new groups appeared to replace them, and spread to occupy new niches. The Ordovician is characterized especially by trilobites of the Asaphida, including such groups as the remopleuridids, trinucleids, and cyclopygids (Figures 17H, 17K, and 17L). As in the Cambrian, Early Ordovician trilobite faunas on the shallow platformal areas were highly endemic; the bathyurids (Figure 17F) of Laurentia and Siberia, dikelokephalinids of eastern Gondwana (south China and Australia), and calymenoideans of west Gondwana show little or no intermixing, whereas the deeper shelf and planktonic forms are less exclusive. During the mid- to later Ordovician, there was some blurring of this provinciality as Laurentia, Baltica, and Avalonia approached one another, and, before the onset of the major Late Ordovician glaciation, latitudinal disparities in temperature seem to have been less marked. The extinction associated with that glaciation affected many groups of trilobites (e.g., agnostoids, bathyurids, the last ptychopariids, very nearly all the asaphoids), especially those that exploited the open ocean at some stage in their development, and greatly reduced the diversity of the class.

Silurian

The trilobites that survived the end-Ordovician extinction (Phacopida, Proetida, styginids, lichids, and odontopleurids; see Figure 17) mostly survived into the Late Devonian, giving these faunas a distinctive stamp. Many of the genera were cosmopolitan, spreading throughout the tropical and temperate parts of the world, thanks to the equable conditions and the closeness of the major continental masses during the period. However, in the cold waters of the southern hemisphere, a new province (the Malvinokaffric Province) was becoming established.

Cheirurina, and Calymenina; they are united partly by the form of their protaspides, which suggests that they may be related to an ancestral ptychopariid. (C) *Cheirurus*, Silurian, 50 mm. (D) *Ananaspis* (Phacopina), Silurian, 30 mm (for *Calymene*, see Figure 3). (E G) Order Proetida compact opisthoparian trilobites with a natant hypostome (conterminant in some derived taxa) and a distinctive form of protaspis. (E) *Tropidocoryphe* (Proetoidea), Devonian, 20 mm. (F) *Bathyurus* (Bathyuroidea), Ordovician, 50 mm. (G) *Phillipsia* (Phillipsiidae), Carboniferous, 20 mm. (H, K, L) Order Asaphida a very large group that includes several suborders, probably derived from a ptychopariid ancestor. A globular protaspis is characteristic, and a median suture is generally present and the rostrum absent in the asaphoid, remopleuridoid, and cyclopygoid groups; in the trinucleoid groups, the eyes are lost and the suture has become marginal. (H) *Ogygiocaris* (Asaphoidea), Ordovician, 70 mm; (K) *Ampyx* (Trinucleoidea (see also Figure 15A)), Ordovician, 40 mm; (L) *Robergia* (Remopleuroidea), Ordovician, 20 mm (for Cyclopygoidea, see Figure 14). (I, J) Corynexochida conterminant Cambrian forms typically with a forward expanding glabella and a large pygidium, which is thought to have given rise to post Cambrian Styginina, including the Illaenoidea. (I) *Amphoton* (Corynexochida), Cambrian, 25 mm; (J) *Ectillaenus* (Illaenoidea), Ordovician, 70 mm (for *Scutellum* (Styginina), see Figure 6). (M, P) Redlichiida large eyed, commonly conterminant forms with many thoracic segments. The Redlichiina are primitive opisthoparian forms, but the Olenellina primitively lacked facial sutures. (M) *Fallotaspis* (Olenellina), Early Cambrian, 60 mm. (P) *Redlichia* (Redlichiina), Cambrian, 40 mm. (N, O) Agnostida small highly derived forms, treated here as trilobites, although excluded from the class by some workers. The Eodiscina, indubitably trilobites, are here included in the Agnostida. (N) *Agnostus* (*Homagnostus*), Cambrian, 5 mm. (O) *Serrodiscus* (Eodiscina), Cambrian, 15 mm. (Q) Ptychopariida a huge and intractable paraphyletic group, primitively opisthoparian and with a natant hypostome, mainly Cambrian in age. It includes the suborders Olenina and Harpina, and may have given rise to other orders, e.g., Asaphida, Phacopida, Proetida. (Q) *Parasolenopleura* (Ptychopariina), Cambrian, 25 mm (for Olenina, see Figure 13; for Harpina, see Figure 15B).

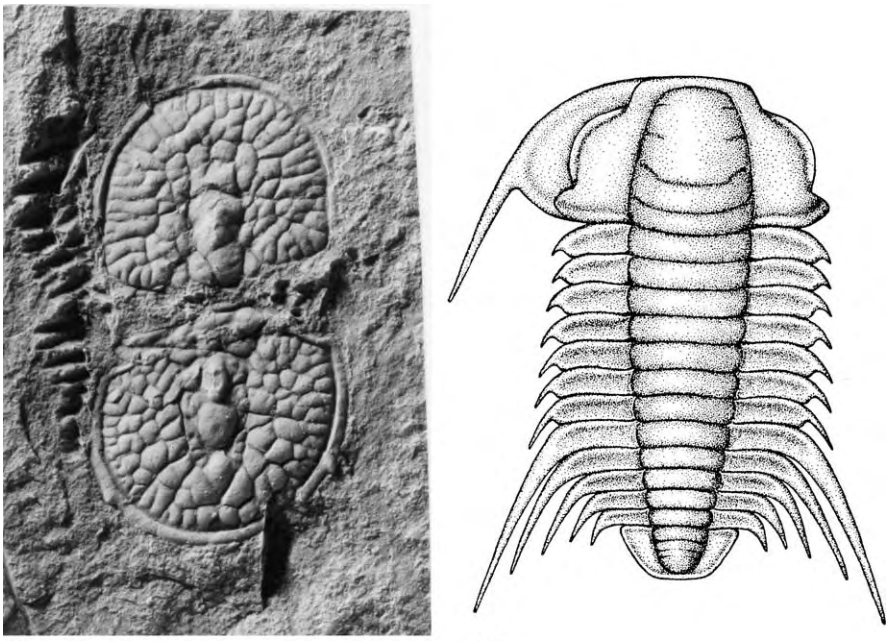


Figure 18 Two globally distributed trilobites that are useful for intercontinental correlation. (A) *Glyptagnostus reticulatus*, characteristic of an early Late Cambrian age; length, 6 mm. (B) *Irvingella*, Late Cambrian; length, 20 mm. Reproduced with permission from Rushton AWA (1983) Trilobites from the Upper Cambrian *Olenus* zone in central England. *Spec. Pal. Paleont.* 30: 107 139, and Rushton AWA (1967) The Upper Cambrian trilobite *Irvingella nuneatonensis*. *Palaeontology* 10: 339 348.

Devonian

Trilobite faunas of the Malvinokaffric Province were widespread in cool-water clastic rocks in the southern hemisphere and were dominated by radiation of the calmoniids (a clade of phacopids), accompanied by homalonotids, aulacopleurids, and odontopleurids. In temperate and tropical latitudes, other families that formerly characterized the Silurian now diversified in the warmer, more varied conditions obtaining during the earlier Devonian, only to vanish later during extinction events precipitated by eustatic changes of sea-level; the cheirurids, calymenids, and lichids became extinct in the Givetian Stage; odontopleurids, harpetids, and styginids in the Frasnian; and the last phacopids at the end of the Famennian. Only Proetida survived to the Carboniferous.

Carboniferous

Despite the extinction of so many clades, the remaining proetid trilobites, represented by four families, diversified to occupy a range of niches in inner-shelf, carbonate mound, and outer-shelf settings during the Early Carboniferous (Mississippian). The morphology of certain outer-shelf taxa is convergent on some of the unrelated, small-eyed (atheloptic) species that characterized comparable environments in the Ordovician and Devonian, and several phillipsiids (Figure 17G) developed features supposedly typical of predators or scavengers – forwardly expanding glabellas and conterminant hypostomes. After the mid-Carboniferous crisis, many forms disappeared and the Late Carboniferous (Pennsylvanian) and Permian trilobites mainly lived in shallow shelf settings.

Permian

Three families of Proetida survived in the Permian, and these, especially the phillipsiids, diversified somewhat in the mid-Permian. However, the major marine regression towards the end of the Permian so limited the habitats available to trilobites that the group finally became extinct just before the end of the Palaeozoic era.

Biostratigraphy

Despite the endemic character of many taxa, trilobites have proved to be of great importance in biostratigraphy, especially for the Cambrian period. The superposition of faunas, empirically determined, has provided one of the most widely applicable means for correlating sedimentary successions, both locally

and worldwide. As trilobites dominated the faunas of the Cambrian and are rich in morphological features, they are the obvious choice for characterizing local and regional successions. The Early Cambrian in Morocco, Siberia, and China is divided into about 10 trilobite biozones in each area; likewise, the Middle Cambrian of Australia and China is divided into some 8–10 biozones. In the Late Cambrian of Australia, 17 trilobite zones are recognized, and about 15 in Kazakhstan. In the olenid facies of Scandinavia, there are eight zones divided into about 30 subzones. Correlation between the successions around different continents is difficult, especially in the earlier Cambrian, but becomes more secure in the later half of the Cambrian. The occasional taxa that crossed the barriers between continents (e.g., *Irvingella* and some Agnostida, such as *Glyptagnostus*) are important because they provide key correlative ties (**Figure 18**). In the Ordovician and Silurian, graptolites are used widely for intercontinental correlation, but the pelagic trilobite, *Carolinites* (**Figure 14A**), is of equal value in the earlier Ordovician. Other trilobites retain their value in local successions, especially those with few graptolites, e.g., the Early Ordovician of the Laurentian platform and the Middle Ordovician of Baltica. Elucidation of the complex patterns of pits in the fringes of trinucleid trilobites has enabled detailed correlations locally, e.g., in the Middle and Late Ordovician of Britain. Trilobites also have biostratigraphical value in certain facies of the Devonian and Carboniferous.

See Also

Biozones. **Evolution**. **Fossil Invertebrates:** Arthropods. **Palaeoecology**. **Palaeozoic:** Cambrian; Ordovician; Silurian; Devonian; Carboniferous; Permian. **Trace Fossils**.

Further Reading

Clarkson ENK (1979) The visual systems of trilobites. *Palaeontology* 22: 1 22.

Fortey RA (2000) *Trilobite! Eyewitness to Evolution*. London: HarperCollins.

Fortey RA and Owens RM (1999) Feeding habits in trilo bites. *Palaeontology* 42: 429 465.

Harrington HJ, Henningsmoen G, Howell BF, *et al.* (1959) Trilobita. In: Moore RC (ed.) *Treatise on Invertebrate Paleontology*, Part O, *Arthropoda* 1. Lawrence, KS: Geo logical Society of America and University of Kansas Press.

Jell PA and Adrain JM (2003) Available generic names for frilobites. *Memoirs of the Queensland Museum* 48: 331 553.

Levi Setti R (1993) *Trilobites*, 2nd edn. Chicago and London: University of Chicago Press.

Šnajdr M (1990) *Bohemian Trilobites*. Prague: Czech Geo logical Survey.

Whittington HB (1992) *Fossils Illustrated 2. Trilobites*. Woodbridge: Boydell Press.

Whittington HB, Chatterton BDE, Speyer SE, *et al.* (1997) *Treatise on Invertebrate Paleontology*, Part O, *Arthro poda* 1, *Trilobita*, revised, vol. 1. Boulder, CO and Law rence, KS: Geological Society of America, Inc., and University of Kansas Press.

Insects

E A Jarzembowski, University of Reading, Reading, UK and Maidstone Museum and Bentlif Art Gallery, Maidstone, UK

Introduction

Insects, which belong to the taxonomic group Hexapoda ('six legs'; Table 1), are the most successful organisms on Earth, if biodiversity is measured as a count of the number of species. More than 1.4 million species of insects have been described in the past 250 years, comprising 65% of all known species of life on Earth (Figure 1A). Insects are also the most successful group in the fossil record, if palaeodiversity is measured as a count of the number of families documented, as many palaeontologists do (Figure 2).

How many insects are there? Nobody knows for sure, although all estimates suggest that millions of insect species remain to be described. The great majority of these live in exotic places. It seems unlikely that description will keep pace with global habitat loss and extinction brought about by human

Table 1 Insects systematic position

Taxonomic division	*Group/number*
Kingdom	Animalia
Phylum	Arthropoda
Superclass	Hexapoda (insects in the broad sense)
Order	About 43
Family	Over 1500
Genus	?
Species	3 20 million

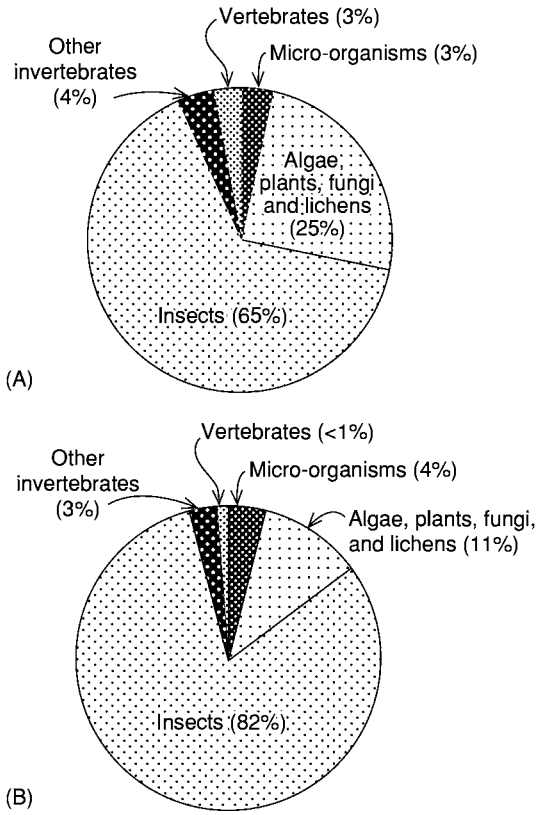


Figure 1 Global biodiversity, showing percentage breakdown for insects and other organisms. (A) Percentages of the approximate number of known species worldwide (1 454 000). (B) Percentages of the estimated total number of species worldwide, which includes those thought to be undiscovered (65 654 000).

development. Simple extrapolation from the fossil record suggests that the total number of insect species is probably less than 20 million, although some biologists prefer a higher figure. Everyone is agreed, nevertheless, that insects represent over 50% of all known species and that they belong to an exclusive group of hyperdiverse organisms.

Origins

Where do insects come from? Insects are undoubtedly a class of arthropods, or 'joint-legged' animals (Table 1). Arthropods (*see* **Fossil Invertebrates:** Arthropods) also include trilobites, crustaceans (prawns, etc.), chelicerates (spiders, etc.), and myriapods (millipedes and centipedes). Insects show closest relationships to crustaceans (e.g., in the structure of their compound eyes) and myriapods (e.g., in their tubular or tracheate respiratory system) (Figure 3). The exact relationships are currently a subject of debate.

Classification

Insects are divided into two main groups – winged and wingless hexapods (Figure 4). The wingless (apterygotan) insects are a mixed group and only some (silverfish) are thought to share a common ancestor with winged (pterygotan) insects. The pterygotes are divided into two main groups (Figure 5) – those that can fold their wings over the body (Neoptera) and those that cannot (Paleoptera). The neopterans, in turn, can be divided into two groups – those that undergo complete metamorphosis (Holometabola) and those that undergo incomplete metamorphosis (cockroach and grasshopper orders and bug orders, or Polyneoptera and Paraneoptera, respectively); the Polyneoptera and Paraneoptera are also known as exopterygotes, because the wings develop on the outside in the young stages (Figure 6). In contrast, the wings develop inside holometabolous insects, or endopterygotes. For holometabolans, the chrysalis, or pupa, is the 'resting' stage between the caterpillar, maggot, or grub stage and the flying adult stage. The holometabolans are the most diverse insects and apterygotes are the least diverse (Figure 7). Some pterygotes have, however, lost their wings (e.g., fleas). Very high diversities (100 000 or more species) are reached in only four (holometabolous) orders: Coleoptera (beetles), Lepidoptera (moths and butterflies), Hymenoptera (wasps, ants, and bees), and Diptera (true flies).

Geological History

The origin of insects is a mystery, the Cambrian Burgess Shale arthropods being too early to cast any light on the subject. The oldest definite hexapod is *Rhyniella praecursor* from the Early Devonian Rhynie Chert. *Rhyniella praecursor* is a springtail belonging to the living apterygote order Collembola. The earliest true insect is currently considered to be *Rhyniognatha hirsti*, also from the Rhynie Chert. In the latest Lower Carboniferous and Upper Carboniferous there is evidence of the radiation of the pterygotes, including paleopterans and polyneopterans. These insects were the world's first flying animals, long before vertebrates took to the air. In the succeeding Permian, the paraneopterans and holometabolans became established. Insects are essentially terrestrial organisms, but the first definite freshwater forms appeared in the Permian. After a setback in the Early Triassic extinction (Figure 8), insects regained their ordinal strength by the Tertiary, establishing some new innovations on the way, e.g., evolving parasitic and parasitoid forms as well as insect societies (Figure 9). The extinction at the start of the Mesozoic seems to have

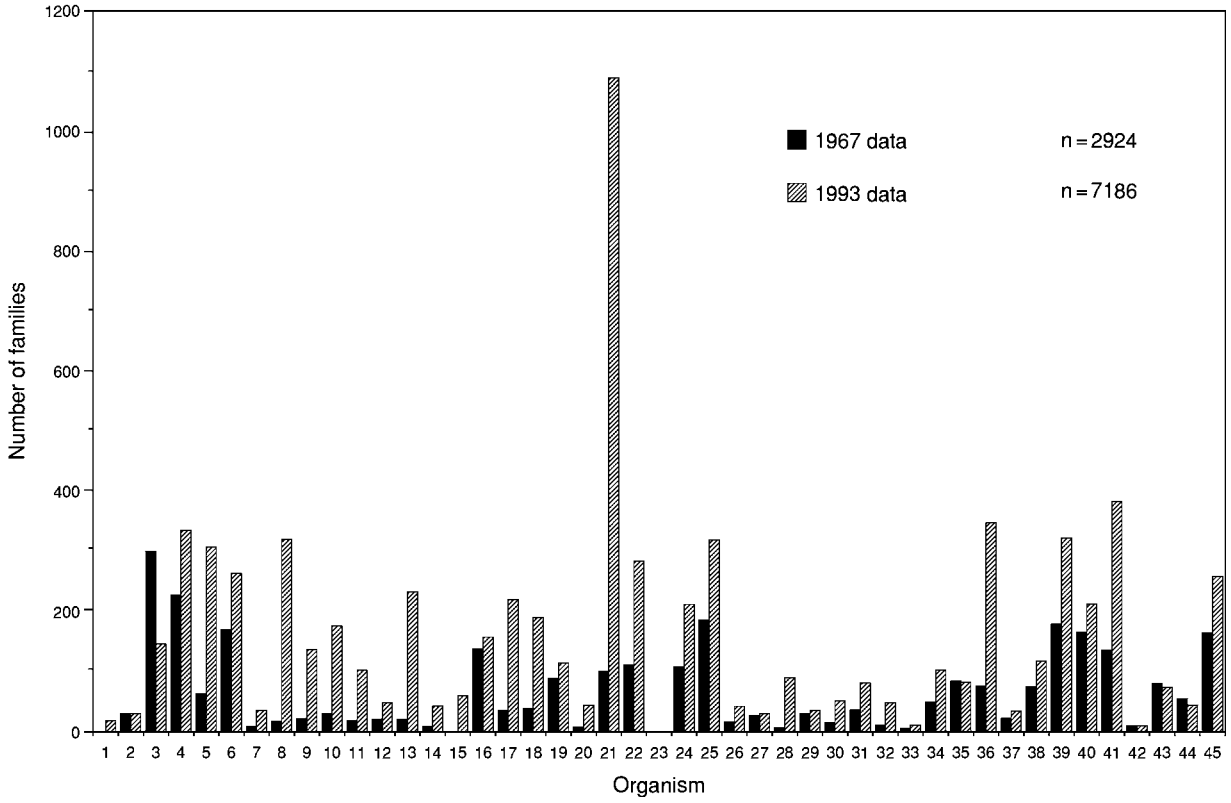


Figure 2 Palaeodiversity of organisms. Key to numbers: 1, bacteria and blue green algae; 2, fungi; 3, other algae; 4, single cell organisms; 5, sponges; 6, corals, etc.; 7, chitons, etc.; 8, snails; 9, nautiluses; 10, ceratites; 11, ammonites; 12, belemnites; 13, bivalves and tusk shells; 14, uncertain molluscs; 15, segmented worms; 16, trilobites; 17, spiders, etc.; 18, crustaceans (excluding seed shrimps); 19, seed shrimps; 20, millipedes, etc.; 21, insects; 22, lamp shells; 23, phoronids; 24, moss animals; 25, sea urchins, etc.; 26, primitive chordates; 27, graptolites; 28, problematica; 29, miscellaneous; 30, conodonts; 31, lampreys, etc.; 32, cyclostomes; 33, primitive fish; 34, sharks, etc.; 35, primitive bony fish; 36, advanced bony fish; 37, more bony fish; 38, amphibians; 39, reptiles; 40, birds; 41, mammals; 42, mosses, etc.; 43, ferns, etc.; 44, seed plants (excluding 45, flowering plants).

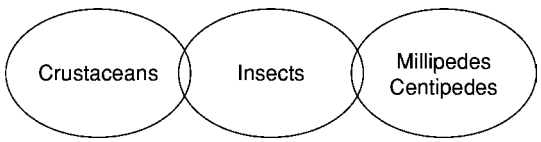


Figure 3 Relationship of insects with other arthropods.

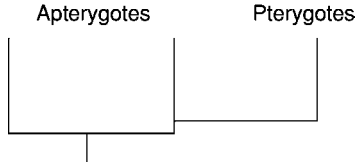


Figure 4 Major insect groups; apterygotes include springtails and silverfish.

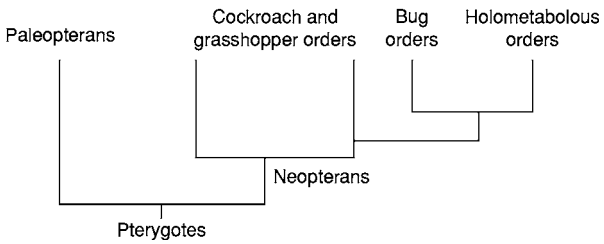


Figure 5 The two main groups of pterygotes are the paleopter ans and the neopterans, which are further divided into subgroups. Paleopterans include dragonflies and mayflies. Cockroaches and grasshoppers, along with stoneflies, stick insects, crickets, locusts, earwigs, termites, and praying mantises comprise the polyneopteran orders. Paraneopteran orders include bugs and lice. Holometabolous orders include beetles, lacewings, wasps, ants, bees, caddisflies, moths, butterflies, flies, and fleas.

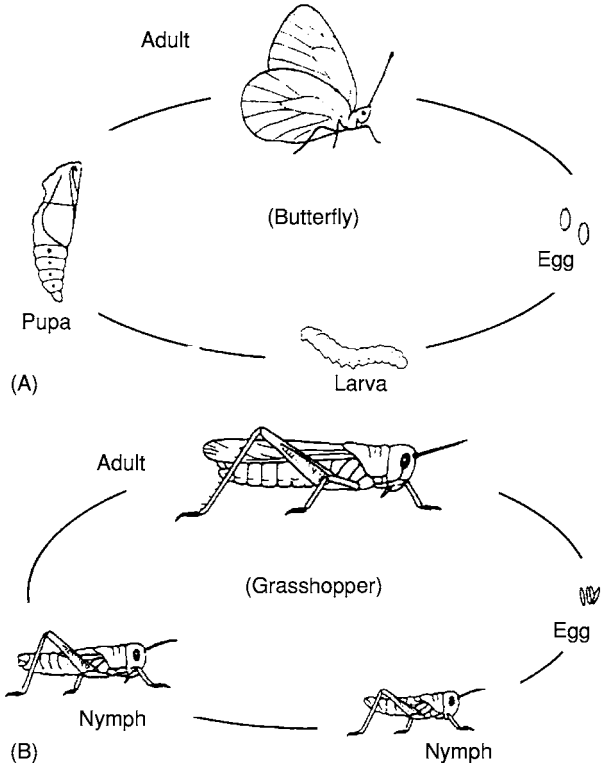


Figure 6 Life cycles of insects. (A) Endopterygotes undergo a complete metamorphosis. (B) Exopterygotes undergo an incom plete metamorphosis.

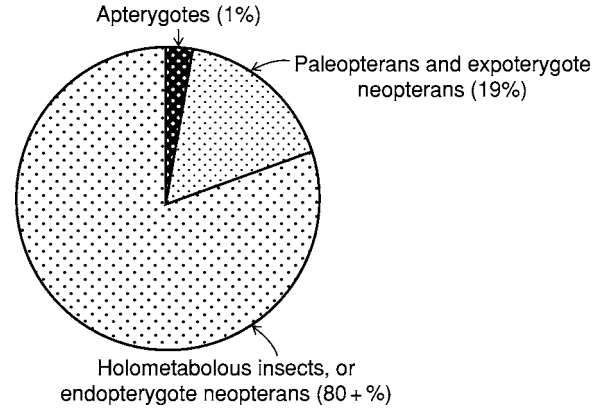


Figure 7 Approximate biodiversity of major insect groupings, showing percentages of primitively wingless insects (aptery gotes), primitively winged insects with incomplete metamorphosis (paleopterans and exopterygote neopterans), and primitively winged insects with complete metamorphosis (holometabolous insects, or endopterygote neopterans).

been the biggest in insect history, although losses were not really catastrophic. Indeed, the successful order Diptera (true flies), which were opportunists ('specials'), arose in the Triassic (Figure 8). The extinction showed, however, that even hyperdiverse

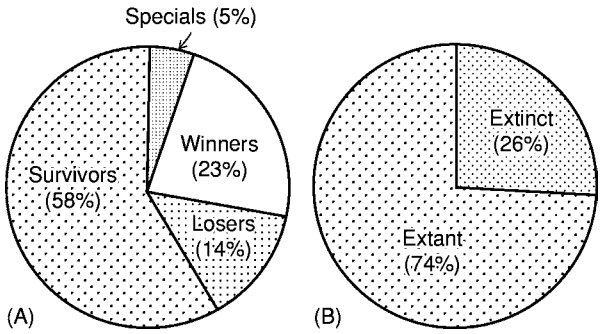


Figure 8 Insect orders during the (A) Triassic extinction; (B) Phanerozoic composition. Opportunists are categorized as 'specials'.

organisms are affected by global environmental change. The principal groups (orders) of insects in the fossil record are outlined in Table 2.

Collecting and Documentation

Fossil insects are more common than is generally supposed, especially their disarticulated remains. They occur in a variety of sedimentary environments, including marine and non-marine deposits, in both organic and fine-grained clastic rocks. Insects often occur in early diagenetic concretions, including ferruginous and phosphatic ones, and in calcareous mudstones. Some of the best preserved insects occur in amber from the Lower Cretaceous onwards. More unusual modes of preservation include pyritized or silicified insects and inclusions in gypsum crystals. Insects are found as trace fossils as well as body fossils. Fossil insects have been used in palaeoenvironmental reconstruction in addition to phylogenetic analysis, especially in Quaternary deposits.

Just as there are millions of insects to be described in today's hot countries, so there are thousands of fossil species to be described in places that were once warmer. The process of collection and documentation of fossil insects is of scientific as well as cultural value. It is possible to find more new fossil insect species on one field trip than in an entire lifetime of collecting of the more popular fossil groups (e.g., vertebrates). Fossil insects are thus ideal for satisfying the goal of finding something new (and they usually require less storage space on account of their small size). Knowledge of the pre-Quaternary insects has largely been forgotten since the era of the pioneer Victorian geologists and naturalists. There is now, however, a revival of interest, the wider search for early (Carboniferous pre-Namurian)

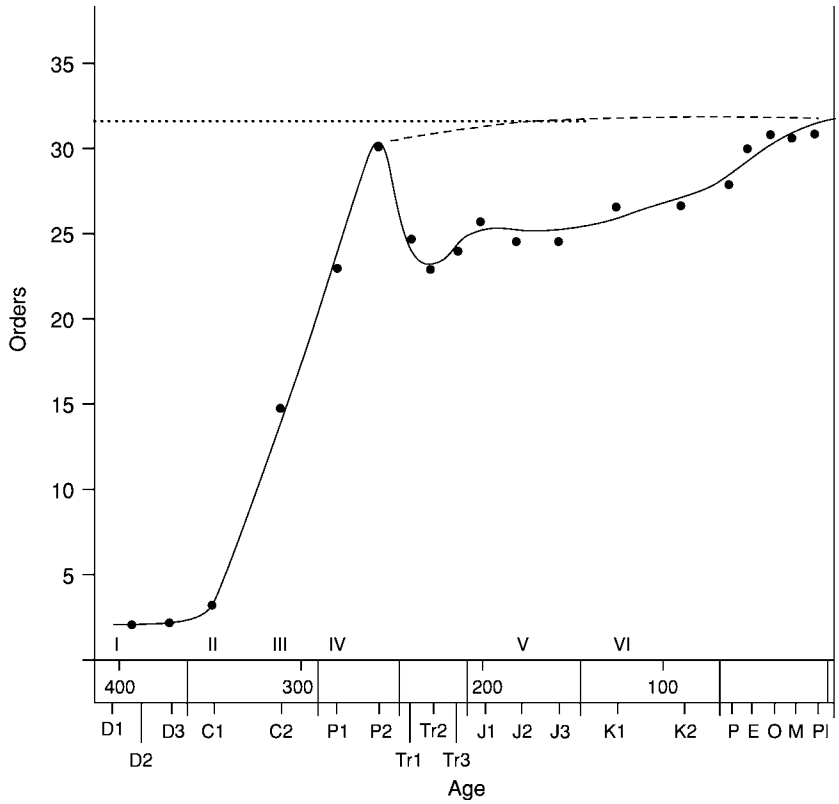


Figure 9 Orders through time and key events. Dashed line represents extrapolation, dots represent equilibrium (saturation) value. Key: Pl, Pliocene; M, Miocene; O, Oligocene; E, Eocene; P, Paleocene; K, Cretaceous; J, Jurassic; Tr, Triassic; P, Permian; C, Carboniferous; D, Devonian. Numerals 1, 2, and 3 represent Lower, Middle, and Upper subperiods and epochs; I, earliest hexapods; II, earliest pterygotes; III, near modern range of plant feeding strategies; IV, evolution of holometabolous insects; V, earliest insectan parasites and parasitoid radiation; VI, earliest amberized and definite social insects.

Table 2 Principal groups of insects in the fossil record

Principal group	*Order*	*Age/description*
Apterygota	Collembola (springtails)	Lower Devonian Recent
	Diplura (two pronged bristletails)	Upper Carboniferous Recent
	'Thysanura' (three tailed bristletails)	This order is split into two groups: Archaeognatha (Middle Devonian? Recent) and Zygentoma (Upper Carboniferous Recent; e.g., silverfish). The extinct order Monura (Upper Carboniferous Permian) with single 'tails' is probably related to the Archaeognatha, whereas Zygentoma is related to the Pterygota
Pterygota		
Paleoptera	Palaeodictyopteroid group	An Upper Carboniferous Upper Permian group of three or four extinct orders (Palaeodictyoptera, Permothemistida, Megasecoptera, Diaphanopterodea) with beak like mouthparts and an additional pair of 'winglets'
	Ephemeroptera (mayflies)	Upper Carboniferous Recent
	Protodonata	Upper Carboniferous Triassic, including the giant dragonflies with wingspans up to 70 cm, the largest insects of all time
	Odonata (dragonflies, damselflies)	Upper Carboniferous Recent
Neoptera	'Protorthoptera'	Lower Carboniferous Triassic. A taxonomic wastebasket of early neopterans

Continued

Table 2 Continued

Principal group	*Order*	*Age/description*
Polyneoptera	Plecoptera (stoneflies)	Lower Permian Recent
	Embioptera (web spinners)	Lower Permian Recent
	Phasmatodea (stick insects)	Upper Permian Recent
	Orthoptera (crickets, grasshoppers, katydids, locusts)	Upper Carboniferous Recent
	Titanoptera	Extinct Triassic order allied to Orthoptera
	Grylloblattodea (ice bugs)	Lower Permian Recent
	Mantophasmatodea	Eocene Recent
	Protelytroptera	Extinct Permian earwiglike insects
	Dermaptera (earwigs)	Lower Jurassic Recent
	Miomoptera	Extinct Upper Carboniferous Lower Jurassic insects
	Blattodea (cockroaches)	Upper Carboniferous Recent
	Isoptera (termites)	Lower Cretaceous Recent
	Mantodea (praying mantises)	Lower Cretaceous Recent
	Caloneurodea	Extinct Upper Carboniferous Permian insects
Paraneoptera	Zoraptera (angel insects)	Oligocene Recent
	Psocoptera (bark and book lice)	Lower Permian Recent
	Phthiraptera (lice)	Eocene Recent
	Thysanoptera (thrips)	Lower Permian Recent
	Hemiptera (true bugs)	Upper Carboniferous Recent
Holometabola (Oligoneoptera)	Glosselytrodea	Extinct Lower Permian Upper Jurassic insects
	Strepsiptera (stylopids)	Eocene Recent
	Coleoptera (beetles)	Lower Permian Recent
	Raphidioptera (snake flies)	Upper Permian Recent
	Megaloptera (alder flies)	Lower Permian Recent
	Neuroptera (lacewings)	Lower Permian Recent
	Hymenoptera (wasps, ants, bees)	Upper Triassic Recent
	Trichoptera (caddis flies)	Lower Permian Recent
	Lepidoptera (moths, butterflies)	Lower Jurassic Recent
	Diptera (true flies)	Lower Triassic Recent
	Siphonaptera (fleas)	Lower Cretaceous Recent
	Mecoptera (scorpionflies)	Lower Permian Recent

insects being of paramount importance. For logistic reasons, the study of fossil insects (palaeoentomology) relies on international co-operation; to facilitate this objective and to promote knowledge, the International Palaeoentomological Society was founded in 2001.

See Also

Fossil Invertebrates: Arthropods.

Further Reading

In addition to the print literature, several groups have web sites that are sources of information about insects: the Arthropod Laboratory of the Russian Academy of Sciences (http://www.palaeoentomolog.ru), the International Palaeoentomological Society (http://www.cwru.edu/affil/fossilinsects), and the University of Barcelona's Meganeura Palaeoentomological Newsletter (http://www.ub.es/dpep/meganeura/meganeura.htm).

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Carpenter FM (1992) *Superclass Hexapoda. Treatise on Invertebrate Paleontology, Part R, Arthropoda 4, 3 & 4*.

Earl of Cranbrook (1996) The scientific value of collec tions. *Sarawak Museum Journal* 50(71): 73 86.

Jarzembowski EA (2001) Insect "bioerosion". *Acta Geolo gica Leopoldensia* 26(52/53): 161 164.

Jarzembowski EA (2003) Palaeoentomology: towards the big picture. *Acta Zoologica Cracoviensia, Krakow* 46(suppl.): 25 36.

Jarzembowski EA and Ross A (1993) The geological record of insects. *Geology Today* 9(6): 218 223.

Rasnitsyn AP and Quicke DLJ (eds.) (2002) *History of Insects*. Dordrecht: Kluwer Academic Publishers.

Wilson EO (1992) *The Diversity of Life*. Cambridge, MA: Harvard University Press.

Brachiopods

D A T Harper, Geologisk Museum, Copenhagen, Denmark

Introduction

The brachiopods or lamp-shells form a distinctive and diverse group of marine, mainly sessile, benthic invertebrates with a long and varied geological history dating back to the Early Cambrian. Over 12 000 fossil species and approximately 350 living species have been reported, belonging to nearly 6000 genera. Brachiopods consist of two shells or valves, the dorsal and ventral, which grow by accretion, recording the ontogenetic history of the animal. The group is distinguished by the possession of a ciliated feeding organ (the lophophore), together with, in most taxa, a fleshy attachment stalk or pedicle. The phylum has a characteristic set of muscles that act in opposition to open and close the valves, together with a variety of skeletal structures that support both the lophophore and musculature. Despite their minimalist metabolism, requiring little food and oxygen, the phylum has adopted a huge range of morphologies and a wide range of ecological strategies during a history of nearly 600 million years.

Brachiopod Animal

The brachiopod animal is enclosed by two, morphologically different shells or valves, opened and closed by a variety of muscles (Figure 1). In contrast with the bivalves (*see* **Fossil Invertebrates:** Bivalves), where the right valve is a mirror image of the left, the plane of symmetry bisects both valves perpendicular to the commissure. The larger of the two valves is generally named the ventral or pedicle valve; in many brachiopods, the fleshy stalk or pedicle emerges from the apex of this valve and attaches the animal to the seabed. The pedicle can vary from a thick, fleshy stalk to a bunch of delicate, thread-like strands, which can anchor the brachiopod in fine mud. Some brachiopods lost their pedicles during ontogeny and adopted a free-living mode of life, lying recumbent on or partially in the sediments on the seafloor. The dorsal or brachial valve contains the extendable food-gathering organ or lophophore, together with its supports. A number of styles of lophophore have evolved. The earliest growth stage, the trocholophe, is an incomplete ring of filaments, still retained by the paedomorphic microbrachiopod *Gwynia*; by the schizolophe stage, a bilobed outline has developed, which probably characterized many of the smaller Palaeozoic taxa. The more complex plectolophe, ptycholophe, and spirolophe lophophore types are characteristic of the more advanced articulated brachiopods.

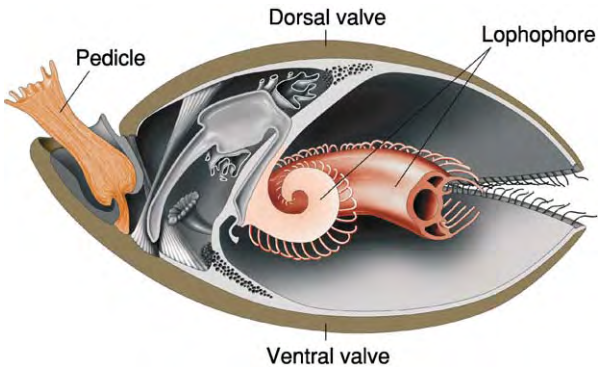


Figure 1 The brachiopod animal showing the key apomorphies of the phylum: the lophophore, pedicle, and dorsal and ventral valves. Redrawn and modified from Kaesler RL (1997).

The linguliformeans (Figure 2A) have organophosphatic shells with pedicles that either emerge between both valves or through a foramen. The shells develop from a planktotrophic larval stage without mantle reversal; the group is characterized by an alimentary tract ending in an anus. In the lingulates, the opening and closing of the valves is achieved by a complex system of muscles and the pedicle emerges between both valves. Some authors have suggested that the withdrawal of the soft parts posteriorly causes a space problem that can force the valves apart; relaxation allows the animal to expand again forwards, allowing the valves to close. The paterinates are the oldest group of brachiopods, appearing in the lowest Cambrian Tommotian Stage. Although linked to the other linguliformeans on the basis of an organophosphatic shell substance, the shell structure of the group is quite different, and the shells have true interareas, delthyria, notothyria, and apparently had functional diductor muscle systems.

The craniiformeans (Figure 2B) contain a diverse, yet probably monophyletic, group of morphologies centered on *Crania*, but including *Craniops* and the bizarre trimerellids. The shells consist of organocarbonate, and the animal developed separate dorsal and ventral mantle lobes after the settlement of a nektobenthonic larval stage.

The rhynchonelliformeans (Figure 3) have a pair of calcitic valves that contain a fibrous secondary layer,

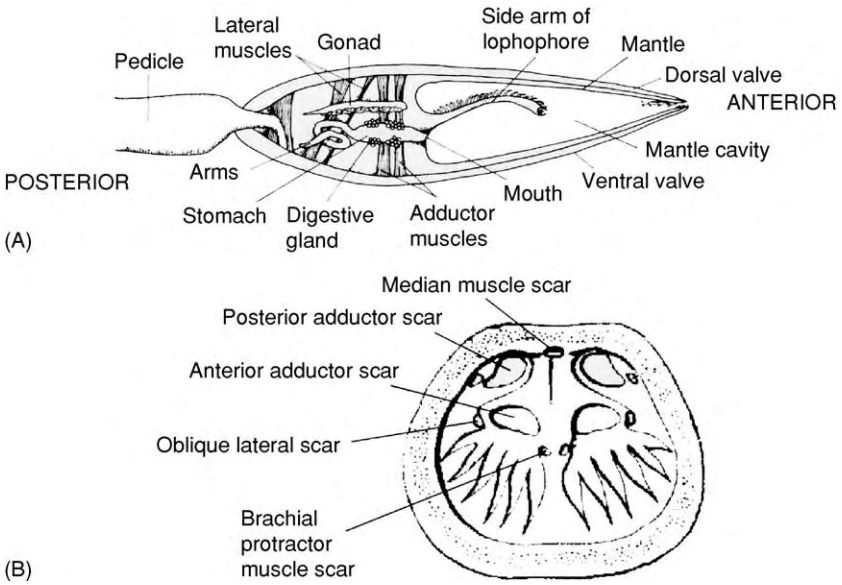


Figure 2 (A) Linguliformean morphology based on *Lingula*. (B) Craniiformean morphology based on *Neocrania*.

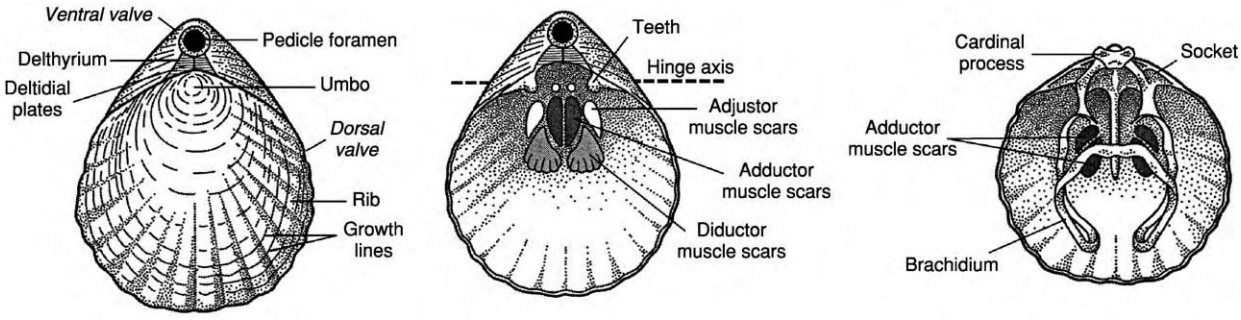


Figure 3 Rhynchonelliformean morphology showing, from left to right, features of the exterior, ventral, and dorsal valves; based on the living genus *Magellania*. Adapted from Clarkson ENK (1998) *Invertebrate Palaeontology and Evolution*, 4th edn. Cambridge: Blackwell Science.

with variable convexity, hinged posteriorly and opening anteriorly along the commissure. The mantle lobes are fused posteriorly, where the interareas are secreted; their margins form the hinge between the ventral and dorsal valves. Articulation was achieved by a pair of ventral teeth and dorsal sockets and the valves were opened and closed by opposing diductor and adductor muscle scars. In the majority of rhynchonelliformeans, the valves were attached to the substrate by a pedicle, developed from a larval rudiment, and emerging through a foramen in the delthyrial region. The subphylum contains five classes: the Chileata, the Obolellata, the Kutorginata, the Strophomenata, and the Rhynchonellata. Already by the Early Cambrian, representatives of four of the five classes were present. However, the last two classes, containing over 1500 and 2700 genera, respectively, dominated Phanerozoic brachiopod faunas.

Brachiopod Shell

The brachiopod shell is a multilayered complex of both organic and inorganic material that has proven to be of fundamental importance in the classification of the phylum. The shells of most rhynchonelliformean brachiopods consist of three layers (**Figure 4**). The outer layer (periostracum) is organic, whereas underneath are the mineralized primary and secondary layers. These layers are sequentially secreted by cells within the generative zone of the mantle, forming first a gelatinous sheath, followed by the organic periostracum, and then the granular calcite of the primary layer; the subsequent secondary layer is thicker, composed of calcite fibres and, in some brachiopods, a third prismatic layer is secreted. There are a number of variations of this basic template. The linguliformeans, for example,

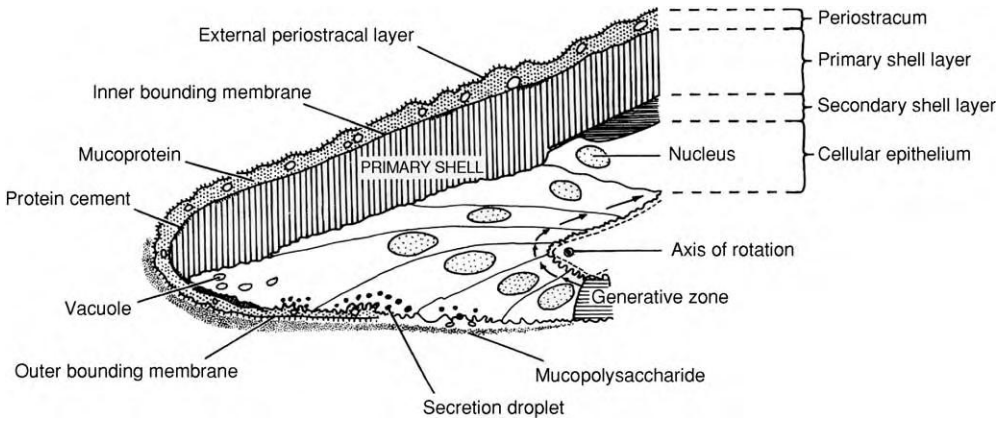


Figure 4 Brachiopod shell structure based on sections through living *Notosaria*. Adapted from Clarkson ENK (1998) *Invertebrate Palaeontology and Evolution*, 4th edn. Cambridge: Blackwell Science.

have phosphatic material combined into the shell fabric. The rhynchonelliformean brachiopods are composed of low-magnesian calcite; these organocarbonate shells may have fibrous, laminar, or cross-bladed laminar shell fabrics in their secondary layers.

Many shells are perforated by small holes or punctae, in life holding finger-like extensions of the mantle or caeca. Some strophomenides have pseudopunctae, with fine inclined rods or taleolae embedded in the shell fabric.

Classification and Phylogeny

The traditional split of the phylum into the Inarticulata and Articulata has been discarded in favour of three subphyla: the Linguliformea, Craniiformea, and Rhynchonelliformea (Table 1). All three have quite different body plans and shell fabrics (Figure 5). The linguliformeans contain five orders united by organophosphatic shells; the inclusion of the paterinides is the most problematic, as the group shares some morphological characters with the rhynchonelliforms. The craniiformeans include three rather disparate groups with quite different morphologies, but which together possess an organocarbonate shell. Most scientists accept eight articulated orders, mainly based on the nature of the cardinalia and the morphology of the other internal structures associated with the attachment of muscles and the support of the lophophore; recently, the more deviant chileides, obolellides, and kutorginides have been added to the subphylum. In addition, the articulated taxa have been split into those with deltidiodont (simple) and cyrtomatodont (complex) dentitions; the former group includes the orthides and strophomenides, and the latter includes the spire bearers.

Cladistic-based investigations have developed a phylogenetic framework for the phylum, supporting the three subphyla (Figure 6); their defining characters are based on shell structure and substance. The mutual relationships between these groups are still unclear, as are the relationships between the many primitive articulated and nonarticulated groups that appeared during the Cambrian explosion. During and following the Ordovician radiation, the relationships between taxa are more clearly defined.

Origin and Affinities

To date, no definite brachiopods have been described from Precambrian rocks. Molecular data and the inclusion of the brachiopods within the other lophophorates (the bryozoans (*see* **Fossil Invertebrates: Bryozoans**) and phoronids) suggest that the phylum belongs to the protostomes. This would support the evolution of the Brachiopoda from a slug-like ancestor, such as *Halkieria* (with anterior and dorsal shells), by the folding of the mantle along a line perpendicular to the length of the animal. On the other hand, morphological evidence suggests that brachiopods are in fact deuterostomes. The ancestral brachiopod may have possessed a planktotrophic larva with a straight gut with the adult form attached by a pedicle. The group diverged to form the organophosphatic linguliformeans with a displaced anus, and the craniiformeans and rhynchonelliformeans with calcitic shells; the rhynchonelliformeans lost the anus and acquired complex hinging mechanisms.

Ecology

Living and fossil brachiopods have developed a wide range of life styles (Figure 7). The majority were attached by a pedicle cemented to a hard substrate or rooted into soft sediment. A number of quite different inarticulated and articulated taxa were cemented to the substrate, whereas some groups

Table 1 The Linguliformea, Craniiformea, and Rhynchonelliformea

Subphylum	*Order*	*Key characteristics*	*Stratigraphical range*
Linguliformea	Lingulida	Spatulate valves with pedicle usually emerging between both shells	Cambrian Recent
	Acrotretida	Micromorphic forms with conical ventral valve; dorsal valve with platforms	Cambrian Devonian
	Discinida	Subcircular shells with conical ventral valve and distinctive pedicle foramen	Ordovician Recent
	Siphonotretida	Subcircular, biconvex valves with spines and elongate pedicle foramen	Cambrian Ordovician
	Paterinida	Strophic shells with variably developed interareas	Cambrian Ordovician
Craniiformea	Craniida	Usually attached by ventral valve; dorsal valve with quadripartite muscle scars	Ordovician Recent
	Craniopsida	Small oval valves with internal platforms and marked concentric growth lines	Ordovician Carboniferous
	Trimerellida	Commonly gigantic, aragonitic shells, with platforms and umbonal cavities	Ordovician Silurian
Rhynchonelliformea	Chileida	Strophic shells lacking articulatory structures but with umbonal perforation	Cambrian
	Dictyonellida	Biconvex valves with large umbonal opening commonly covered by a colleplax	Ordovician Permian
	Naukatida	Biconvex shells with articulatory structures and apical foramen	Cambrian
	Obolellida	Oval valves with primitive articulatory structures	Cambrian
	Kutorginida	Strophic valves with interareas but lacking articulatory structures	Cambrian
	Orthotetida	Biconvex shells, commonly cemented, with bilobed cardinal process	Ordovician Permian
	Billingsellida	Usually biconvex with transverse teeth and simple cardinal process	Cambrian Ordovician
	Strophomenida	Concavoconvex, usually bilobed cardinal process; recumbent life mode; cross laminar shell structure with pseudopunctae	Ordovician Permian
	Productida	Concavoconvex valves with complex cardinalia; recumbent or cemented life mode; often with external spines	Ordovician Triassic
	Protorthida	Well developed interareas, primitive articulation and ventral free spondylium	Cambrian Devonian
	Orthida	Biconvex, usually simple cardinal process; pedunculate; delthyria and notothyria open	Cambrian Permian
	Pentamerida	Biconvex, rostrate valves with cruralia and spondylia variably developed	Cambrian Devonian
	Rhynchonellida	Usually biconvex, rostrate valves with variably developed crurae	Ordovician Recent
	Atrypida	Biconvex valves with dorsally directed spiralia and variably developed jugum	Ordovician Devonian
	Athyridida	Usually biconvex valves with short hingeline and posterolaterally directed spiralia	Ordovician Jurassic
	Spiriferida	Wide strophic valves with laterally directed spiralia; both punctate and impunctate taxa	Ordovician Jurassic
	Thecideida	Small, strophic shells with complex spiralia including brachial ridges and median septum	Triassic Recent
	Terebratulida	Biconvex valves with variably developed long or short loops	Devonian Recent

evolved clasping spines and possibly extended mantle fibres to help stabilize their shells. In a number of groups, the pedicle atrophied during ontogeny. Many taxa thus developed strategies involving inverted, pseudoinfaunal, and recumbent life modes; a number lived in cosupportive clusters and others mimicked corals. Not all brachiopods were sessile; a few, such as *Lingula*, adopted an infaunal life style, whereas the articulated forms, *Camerisma* and *Magadina*, were semi-infaunal.

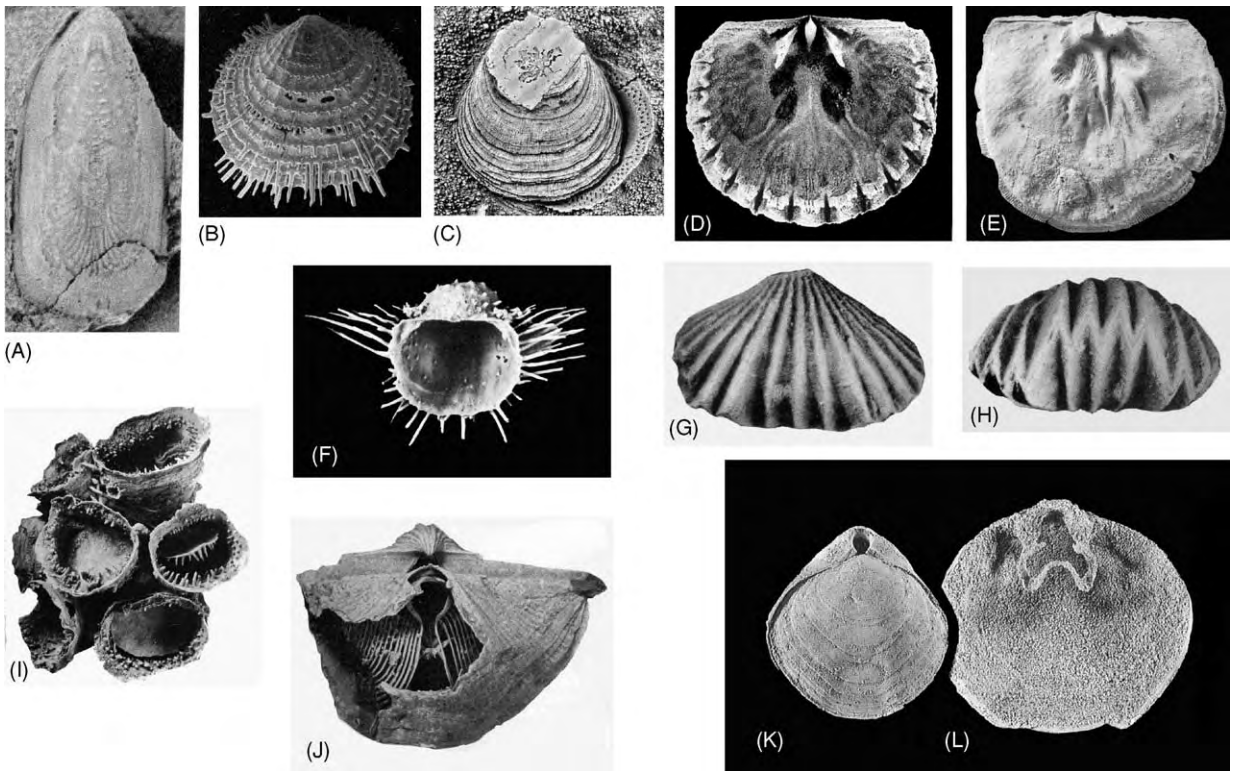


Figure 5 Range of brachiopod morphologies: (A) *Pseudolingula* (Ordovician lingulide: ×2); (B) *Nushbiella* (Ordovician siphonotretide: ×10); (C) *Crania* (Palaeogene craniide: ×2.5); (D) *Sulevorthis* (Ordovician orthide: ×3); (E) *Rafinesquina* (Ordovician strophomenide: ×1/3); (F) *Grandaurispina* (Permian productide: ×1); (G, H) *Rostricellula* (Ordovician rhynchonellide: ×3); (I) *Cyclacantharia* (Permian richthofeniid: ×0.5); (J) *Neospirifer* (Permian spiriferide: ×2/3); (K,L) *Seymourella* (Palaeogene terebratulide: ×1).

Throughout the Phanerozoic, the brachiopods have participated in a spectrum of level-bottom, benthic palaeocommunities. Pioneer studies on Silurian brachiopods suggested that their palaeocommunities were depth related (Figure 8A). The onshore–offshore assemblages of the *Lingula*, *Eocoelia*, *Pentamerus*, *Stricklandia*, and *Clorinda* palaeocommunities have been amplified and modified to form the basis of Benthic Assemblage (BA) zones 1–5, ranging from intertidal environments to the edge of the continental slope; more basinal environments are included in an extra BA 6. Parallel studies on Mesozoic brachiopods have, on the other hand, suggested that brachiopod-dominated palaeocommunities were controlled by substrate rather than depth (Figure 8B). Clearly, a combination of these and other factors in reality controlled the distributions of the Brachiopoda in a complex system of suspension-feeding guilds.

Brachiopods have also acted as substrates for a variety of small epifaunal animals. The progressive and sequential colonization of Devonian spiriferids by *Spirorbis*, *Hederella*, *Paleschara*, and *Aulopora* marks the development of sere and climax palaeocommunities on a brachiopod shell. It has been suggested that such animals congregated adjacent to the inhalant currents on the median parts of the anterior commissure. An alternative hypothesis reverses the direction of flow through the brachiopod mantle cavity, and thus these commensal organisms took advantage of waste being ejected from the brachiopod.

Geographical Distribution

The biogeographical patterns of the linguliformean brachiopods were quite different from those of the craniiformeans and rhynchonelliformeans. The former had planktotrophic larval phases with a facility for wide dispersal; in contrast, the lecithotrophic larvae of the latter were short-lived. Brachiopods dominated the benthos of the Palaeozoic evolutionary fauna. Nevertheless, Cambrian brachiopods were organized into tropical and natal realms, where linguliformeans developed widespread distributions in shelf and slope settings, and rhynchonelliformeans were more diverse in the tropics, preferring shallow-water carbonate and mixed carbonate–siliciclastic environments. Ordovician brachiopods generally showed a decreased provincialism during

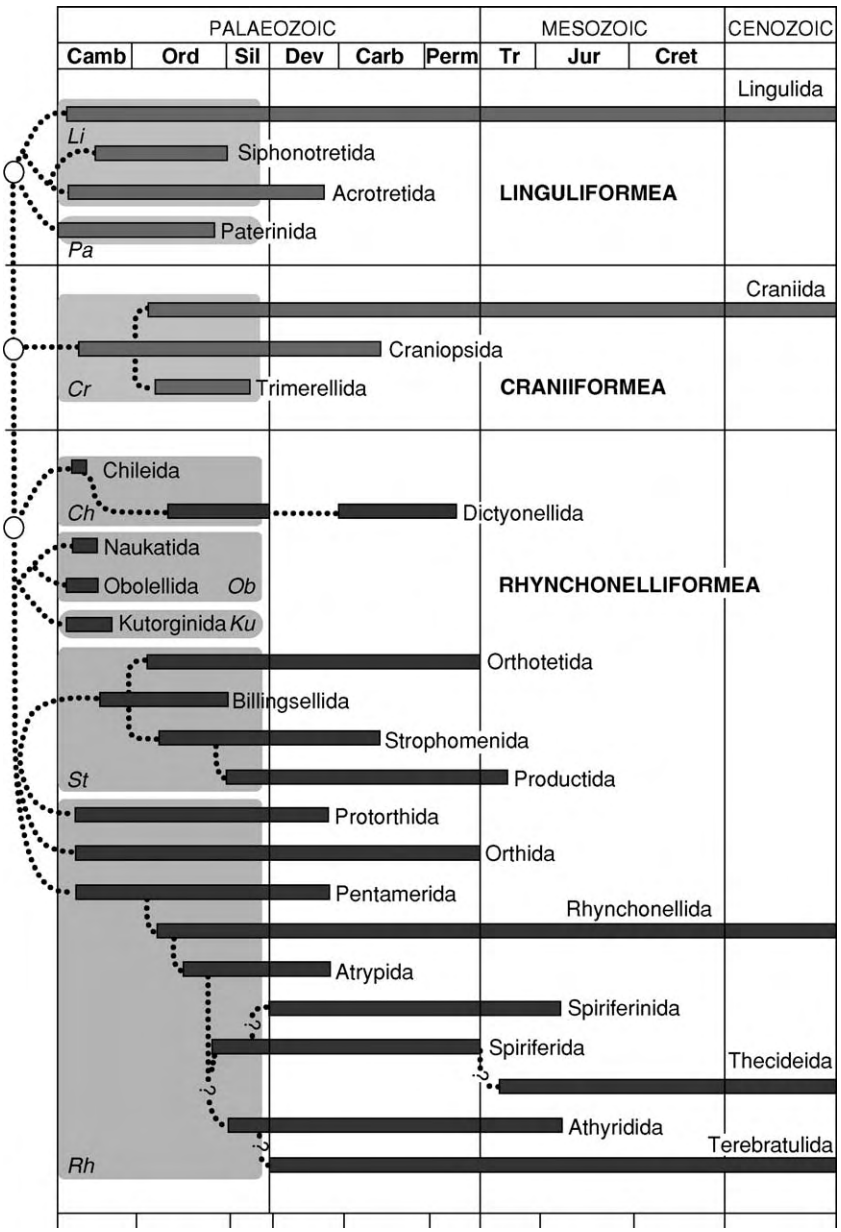


Figure 6 Phylogeny and stratigraphical distribution of the brachiopod orders.

the period; provinciality was most marked during the Early Ordovician, when a range of platform provinces associated with the continents of Baltica, Gondwana, Laurentia, and Siberia were supplemented by loci of endemism associated with a range of microcontinents and volcanic arcs and island complexes. Provincialism was reduced during the Silurian with the close proximity of many major continents; by the Wenlock, however, two broad provinces, the cool-water *Clarkeia* and the mid-latitudinal *Tuvaella* faunas, emphasize an increasing endemism, climaxing during the Ludlow and Prídolí epochs. Provinciality was particularly marked during the mid-Devonian, coincident with peak diversities in the phylum. Although clear biogeographical patterns continued into the Carboniferous, the Permian was characterized by high degrees of provinciality probably associated with steep climatic gradients.

During the Triassic, brachiopod faunas, following an interval of cosmopolitan disaster taxa, were organized into Boreal (high-latitude) and Tethyan (low-latitude) realms. This pattern continued throughout the Mesozoic with loci of endemics and occasional modifications due to ecological factors, such as the circulation of ocean currents and the

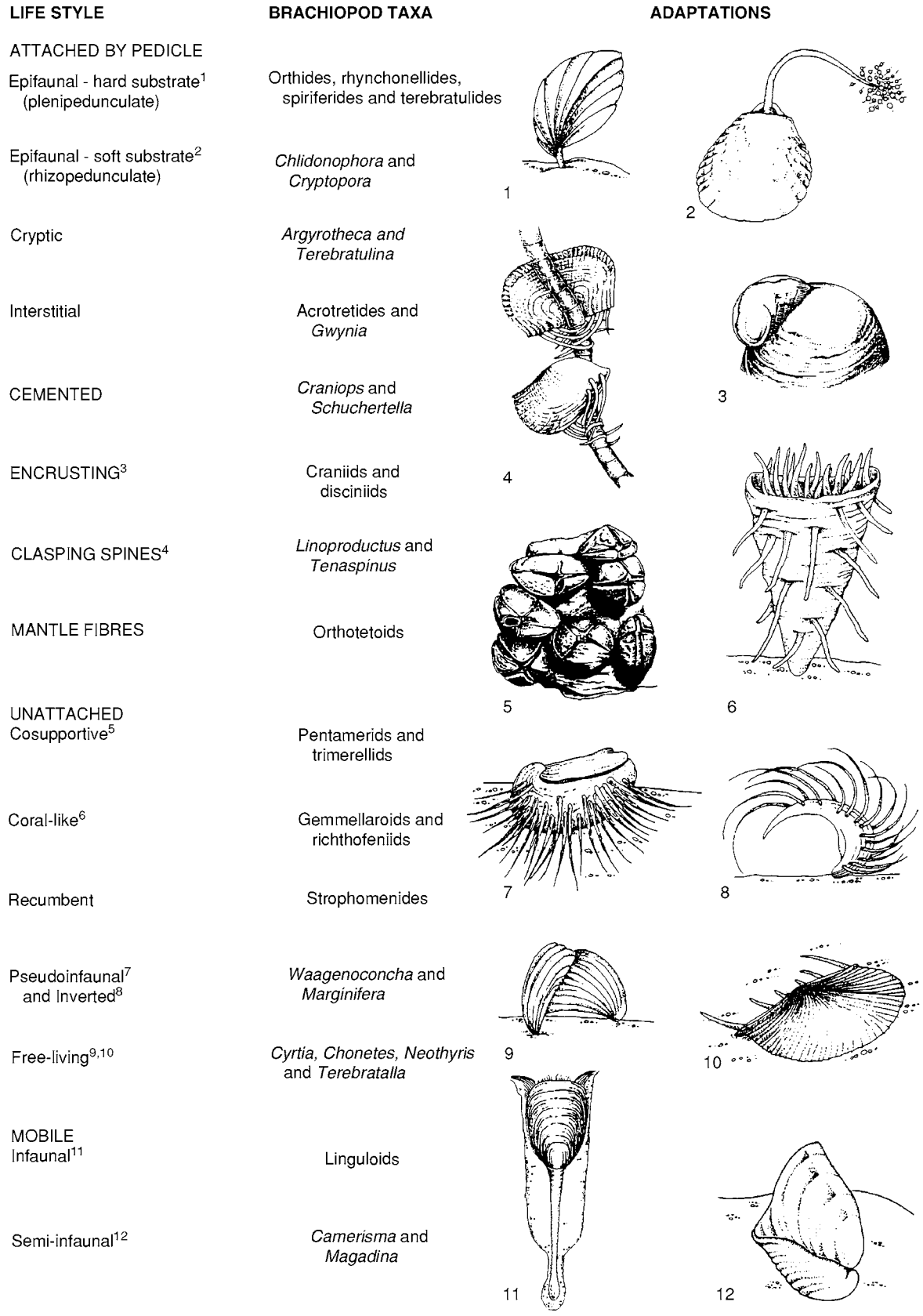


Figure 7 Brachiopod life styles. Reprinted from Harper DAT and Moran R (1997) Brachiopod life styles. *Geology Today* 13: 235 238.

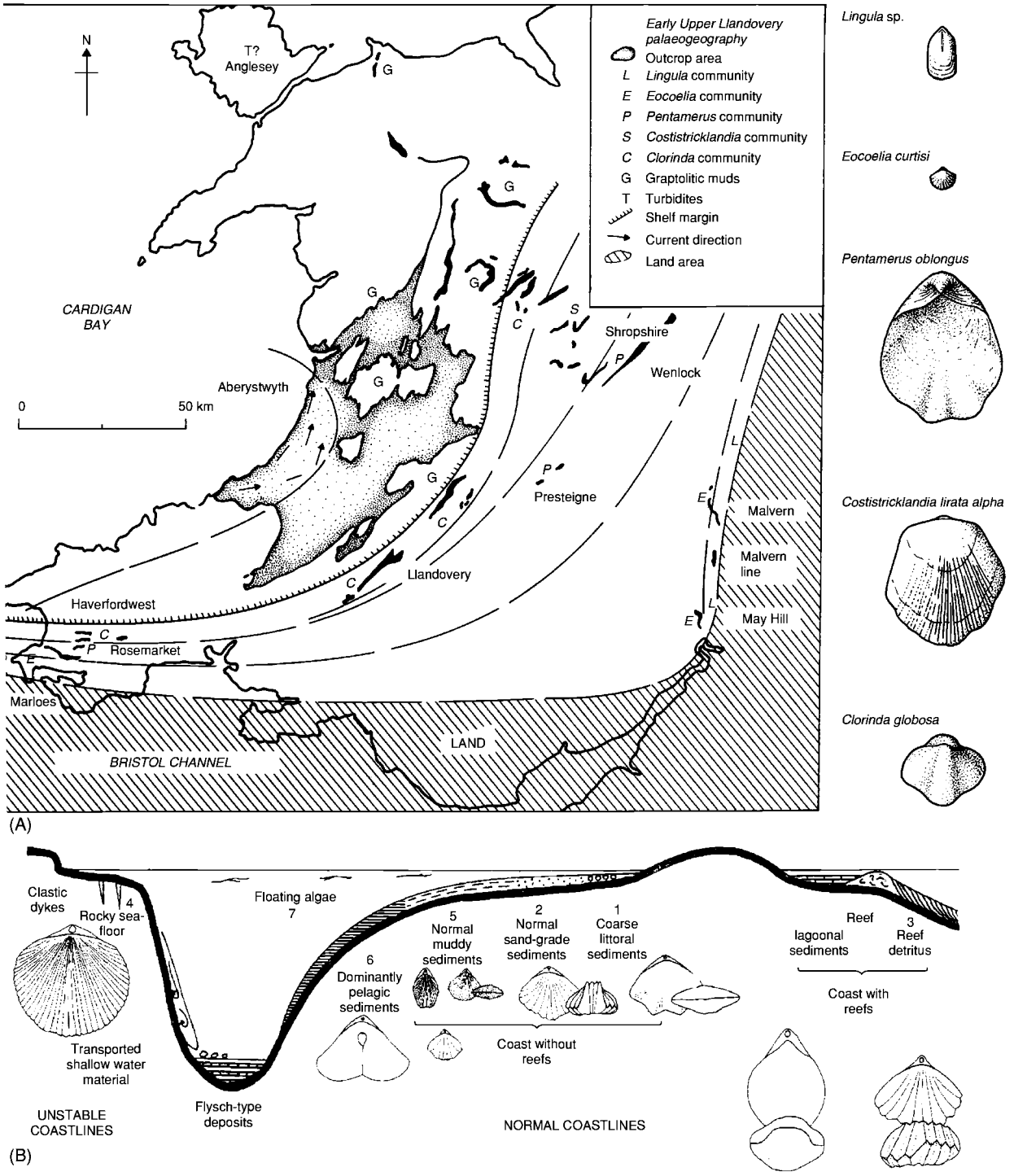


Figure 8 (A) Depth related brachiopod dominated palaeocommunities, first established in the Silurian palaeoenvironments of the Welsh Basin. Modified from Ziegler AM, Cocks LRM and McKerrow WS (1968) *Palaeontology*. Cambridge: Blackwell Science. (B) Substrate related brachiopod dominated palaeocommunities from the Mesozoic, established across the Alpine region. Adapted from Benton MJ and Harper DAT (1997) *Basic Palaeontology*. Harlow: Addison Wesley Longman.

local development of chemosynthetic environments. Biogeographical patterns amongst living forms reflect their Cenozoic roots; a Southern Area, Northern Pacific, and a Northern Area (Atlantic, Mediterranean, North Sea, and the circumpolar northern oceans) are based on a variety of articulated brachiopod associations; the linguliformeans have more widespread, near-cosmopolitan distributions.

Stratigraphical Distribution

In terms of the Phanerozoic evolutionary faunas, the Cambrian, Palaeozoic, and Modern brachiopod faunas are fundamentally different. Cambrian faunas were dominated by a range of nonarticulated groups, together with groups of disparate articulated taxa, such as the chileides, naukatides, obolellides, kutorginides, billingsellides, protorthides, orthides, and pentamerides. These groups participated in a variety of loosely structured, nearshore palaeocommunities. During the Ordovician radiation, the deltidiodont orthides and strophomenides dominated faunas (Figure 9), many first generated around Early Ordovician island complexes and later dominating the platforms, where they participated in an offshore movement of palaeocommunities. Following the end-Ordovician extinction event, spire-bearing brachiopods reached their dominance, particularly in the carbonate environments of the mid-Palaeozoic. The Carboniferous and, particularly, the Permian were intervals of spectacular experimentation: some brachiopods mimicked corals or developed extravagant clusters of spines, whereas a number of groups reduced their shells, thus presenting soft tissues to the outside environment. The end-Permian extinction removed this diversity. Mesozoic and Cenozoic brachiopods were and are overwhelmingly dominated by cyrtomatodont rhynchonelliformeans with either crurae (rhynchonellides) or loops (terebratulides).

Extinctions and Radiations

The brachiopods experienced five main extinction events followed by recoveries and radiations of varying magnitudes (Figure 9). The end-Ordovician event occurred in two phases against a background of glaciation and accounted for almost 80% of the existing brachiopod families. The recovery and subsequent radiation are marked by the decline of deltidiodont groups, such as the orthides and strophomenides, whereas the spire-bearing atrypides, athyridides, and spiriferides, together with the pentamerides, achieved a greater dominance, particularly in carbonate environments. Late Devonian events, at the Frasnian–Famennian stage boundary, also associated with climate change, removed the atrypides and pentamerides, and severely affected the orthides and strophomenides, whereas the spiriferides and rhynchonellides survived in deeper water environments and staged an impressive recovery. A particular feature of the post-Frasnian fauna is the diversity of recumbent brachiopod megaguilds, dominated by the productides.

The end-Permian event (*see* **Palaeozoic:** End Permian Extinctions) has been ascribed to a range of different causes and was associated with the disappearance of over 90% of all living species, including some of the most ecologically and taxonomically diverse brachiopods. The post-extinction fauna was first dominated by a variety of disaster taxa, including lingulids; nevertheless, the brachiopod fauna later diversified within a relatively few clades dominated by the rhynchonellides and terebratulides. The end-Triassic event removed the majority of the remaining spiriferides and the last strophomenides. The agenda set by the end-Permian event, involving the subsequent dominance of rhynchonellide and terebratulide groups, was continued after the end-Triassic event. The end-Cretaceous event (*see* **Mesozoic:** End Cretaceous Extinctions) may have been responsible for the loss of about 70% of Chalk brachiopod faunas in North-West Europe; nevertheless, the subsequent radiation in the Danian limestone facies involved many of the pre-extinction taxa at the generic level.

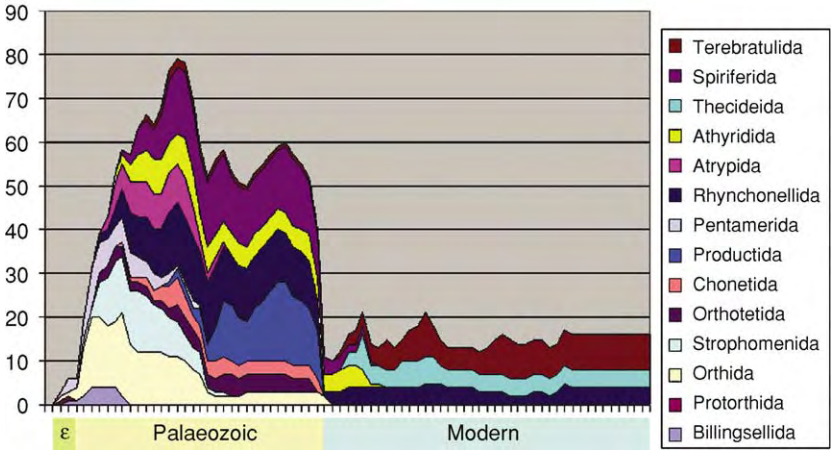


Figure 9 Stratigraphical abundance and distribution of the main orders of rhynchonelliformean brachiopod.

Modern Brachiopod Faunas

Despite their relative rarity today as compared with the Palaeozoic, living brachiopods are actually widespread, represented mainly by pedunculate forms attached to a variety of substrates across a spectrum of water depths. At high latitudes, brachiopods range from intertidal depths to basinal environments at depths of over 6000 m. They are most common in fjord settings in Canada, Norway, and Scotland and in the seas around Antarctica and New Zealand. The association between the horse mussel, *Modiolus modiolus*, and *Terebratulina retusa* is particularly widespread in the Northern Hemisphere. In the tropics, however, many species are micromorphic, exploiting cryptic habitats in reef crevices or in the shade of corals and sponges. Larger forms live in deeper water environments, evading groups of predators that might graze on meadows of newly attached larvae.

See Also

Biological Radiations and Speciation. **Fossil Invertebrates:** Bryozoans; Bivalves. **Mesozoic:** End Cretaceous Extinctions. **Palaeozoic:** Cambrian; Ordovician; End Permian Extinctions. **Sedimentary Environments:** Carbonate Shorelines and Shelves.

Further Reading

Benton MJ and Harper DAT (1997) *Basic Palaeontology.* Harlow: Addison Wesley Longman.

Brunton CHC, Cocks LRM, and Long SL (eds.) (2001) *Brachiopods Past and Present. Systematics Association Special Volume Series 63*. London and New York: Taylor and Francis.

Carlson SJ and Sandy MR (eds.) (2001) *Brachiopods Ancient and Modern. A Tribute to G. Arthur Cooper. Paleontological Society Papers 7*. New Haven: Yale University Reprographics.

Clarkson ENK (1998) *Invertebrate Palaeontology and Evolution,* 4th edn. Cambridge: Blackwell Science.

Harper DAT and Moran R (1997) Brachiopod life styles. *Geology Today* 13: 235 238.

Harper DAT and Rong J (2001) Palaeozoic brachiopod extinctions, survival and recovery: patterns within the rhynchonelliformeans. *Geological Journal* 36: 317 328.

Kaesler RL (ed.) (1997 2002) *Treatise on Invertebrate Paleontology,* Part H, *Brachiopoda* (revised), vol. 1 4. Boulder, CO and Lawrence, KS: The Geological Society of America and the University of Kansas (continuing).

Nielsen C (2002) *Animal Evolution: Interrelationships of the Living Phyla,* 2nd edn. Oxford: Oxford University Press.

Williams A, Carlson SJ, Brunton CHC, Holmer LE, and Popov LE (1996) A supra ordinal classification of the Brachiopoda. *Philosophical Transactions of the Royal Society, Biological Sciences* 351: 443 481.

Bryozoans

P D Taylor, The Natural History Museum, London, UK

Introduction

The Bryozoa are a phylum of colonial invertebrates occasionally referred to as Ectoprocta or by the obsolete name Polyzoa. Because the great majority of bryozoan species possess calcareous skeletons, they are among the commonest groups of macrofossils found in the post-Cambrian marine fossil record. Bryozoans are active suspension feeders. In common with most other suspension feeders, they do not need to move in search of food and, with a few exceptions, are sessile, living permanently anchored to a hard surface, such as a rock, shell, or seaweed. Although present-day bryozoans can be found living in freshwater lakes and slow-flowing rivers, the majority are marine, inhabiting depths from the intertidal to the abyssal. Most modern species are stenohaline, and fossil bryozoans are almost exclusively marine owing to the absence of mineralized skeletons in freshwater species.

All bryozoans are colonial. Each colony comprises a group of genetically identical modular units called zooids. Zooid size enables bryozoans to be distinguished from colonial corals (*see* **Fossil Invertebrates:** Corals and Other Cnidaria), with which they are occasionally confused: bryozoan zooids seldom exceed 2 mm in maximum surface dimension, whereas the equivalent coral structures (corallites) are typically centimetric in scale. Colony growth occurs mostly by the addition of new zooids, a process termed budding, supplemented in some groups by lengthening of existing zooids. Fully grown colonies of some species contain fewer than 10 zooids, whereas other species may develop colonies containing many thousands of zooids, which may on occasion reach 50 cm in diameter.

Anatomy and Feeding

The basic bryozoan zooid consists of a body wall enclosing a fluid-filled cavity (coelom or pseudocoel)

in which is suspended the digestive system and various other organs (Figure 1). The body walls of most species incorporate a calcareous layer, which forms the hard skeleton of the zooid and is sometimes termed the zooecium (the hard skeleton of the colony as a whole is termed the zoarium). Bryozoans have a U-shaped gut with separate mouth and anus. They feed actively on particles in suspension, mainly phytoplankton, using an inverted-V-shaped or bell-shaped lophophore comprising a ring of between 8 and 30 or more tentacles. Cilia on the tentacles beat in unison to create a current of water that enters the open end of the lophophore and exits at the sides between the tentacles. Some food particles are thrown directly towards the mouth at the bottom of the lophophore, while others are sieved by the cilia and/or batted back into the fast flow at the centre of the lophophore by tentacular flicking. Zooids in some species feed more or less independently, but in others the feeding currents of groups of zooids combine. Spaced a few millimetres apart over the colony surface, excurrent chimneys may develop to which zooids channel their filtered water. These chimneys are often marked by raised areas on colony surfaces termed monticules, which can be clearly distinguished in fossil bryozoans. In other species with mesh-like erect colonies, zooidal lophophores create a one-way flow through the holes or fenestrules.

When not feeding, the lophophore can be withdrawn into the safety of the body walls, where it is further

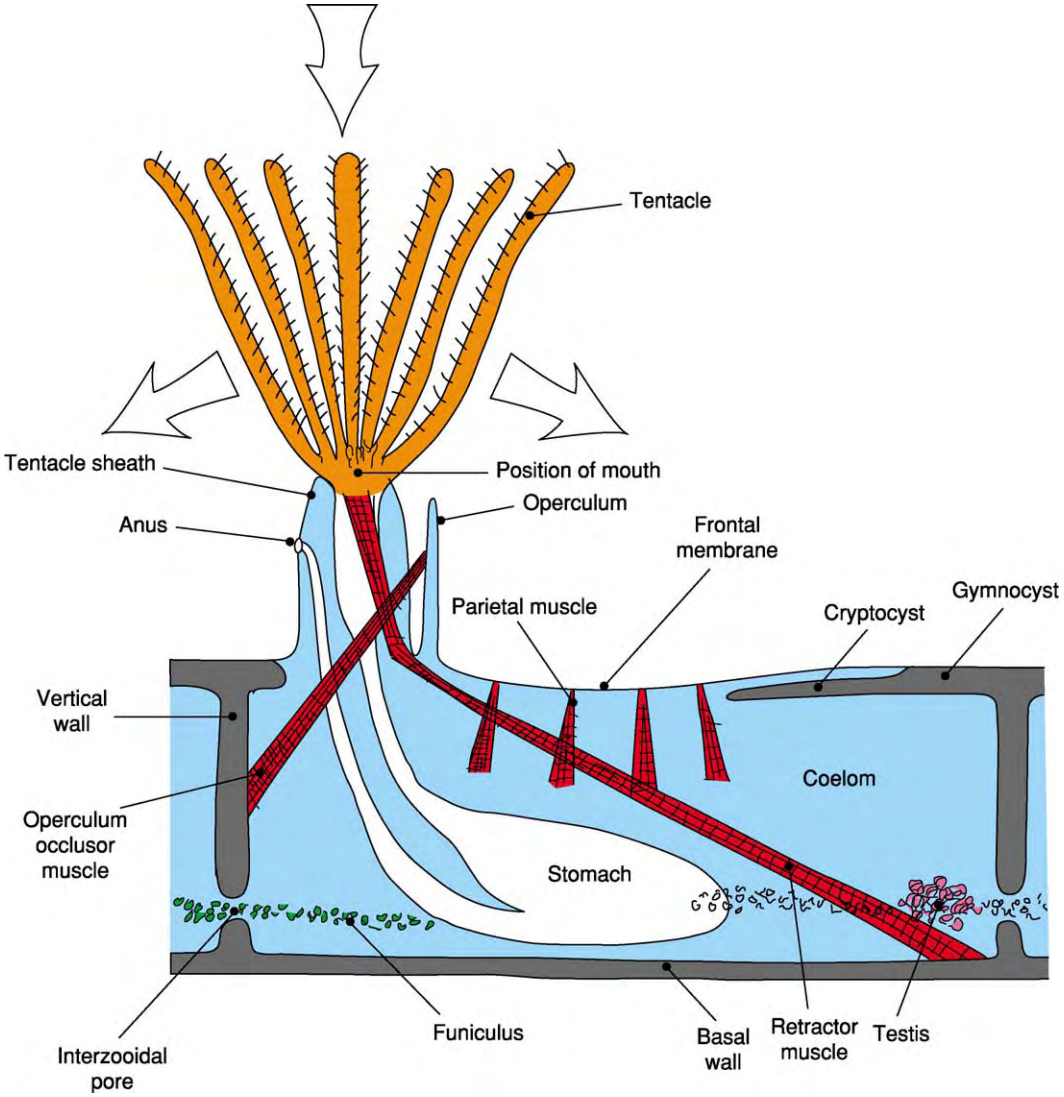


Figure 1 Anatomy of a basic bryozoan zooid illustrated by a diagrammatic vertical section through a cheilostome zooid with the lophophore extended. The gut and tentacle crown (lophophore) are shown in orange, the coelom in blue, the muscles in red, the funiculus in green, the testis in purple, and the calcified skeleton in dark grey. Arrows above and to the sides of the lophophore indicate feeding current flow. The parietal muscles are attached to the basal wall outside the plane of section.

protected by a hinged flap (operculum) in some species. Withdrawal is accomplished by the fast contraction of a retractor muscle running between the base of the lophophore and a point on the body wall. Different taxonomic groups employ different mechanisms for protruding the lophophore. However, all use a hydrostatic system entailing muscles acting on body walls to drive coelomic fluid into the tentacle sheath, forcing it to evert and push the lophophore out through the aperture of the zooid.

Bryozoans have a primitive nervous system. Nervous activity can be demonstrated by touching a lophophore, thereby inducing its immediate retraction into the zooid. The simultaneous retraction of the lophophores of adjacent zooids shows that the nervous systems of zooids are interconnected. Zooids can also be linked by a system called the funiculus. This is thought to be a homologue of the circulatory systems found in many other animals, but it does not carry blood or oxygen or have a heart. Instead, the funiculus, which is linked to the stomachs of the zooids, transports lipids and other metabolites both within and between zooids. Movement of food resources across the colony is undoubtedly important in the provisioning of nonfeeding zooids, developing larvae, and newly budded zooids at the growing edges of the colony. Pores in the skeletal walls allow both nervous and funicular connections between zooids. In some species soft-tissue linkages also occur over the outer ends of the skeletal walls between the zooids.

Reproduction and Growth

The life cycles of bryozoans are complex and varied. Colony growth occurs mostly by zooidal budding, an asexual process involving mitotic cell division. However, the formation of new colonies is usually a sexual process, with meiotic cell division occurring in the testes and ovaries, which are located alongside the funiculus. Bryozoan colonies are hermaphroditic, with each zooid usually producing sperm and eggs sequentially, although some species have separate male and female zooids. Released through tiny pores in the tips of the tentacles, sperm are carried away in the exhalent feeding currents and ambient flow, and a small proportion survive to fertilize an egg in another colony. Fertilized eggs develop into larvae. In a minority of bryozoan species the larvae are planktotrophic – they feed while in the plankton for weeks or months. However, most bryozoan species have non-planktotrophic larvae, which are brooded by the parent colony before being released into the plankton for a short period (hours or days). In both cases, the larvae eventually settle on a firm or hard surface (e.g. a shell or stone) and undergo metamorphosis to form the first zooid (ancestrula) of a new colony. Budding from the ancestrula produces the first generation of asexual zooids, which in turn bud further zooids, and so on.

New zooids are usually budded at specific locations, such as branch tips in tree-like colonies and the peripheral growing edge in sheet-like encrusting colonies (**Figure 2B**). The location and growth orientation of new buds largely determines the shape of the colony. Modular zooids of similar shape can be 'assembled' into disparate colony forms, many of which have evolved in parallel in unrelated groups. Various schemes have been devised to classify these growth forms, some using geometric terms (e.g. reticulate) and others based on genera having the growth form (e.g. reteporiform). A simple division is into encrusting, erect, and free-living colonies. Encrusting colonies tend to be either sheet-like (**Figure 2B**), with zooids arranged multiserially, or

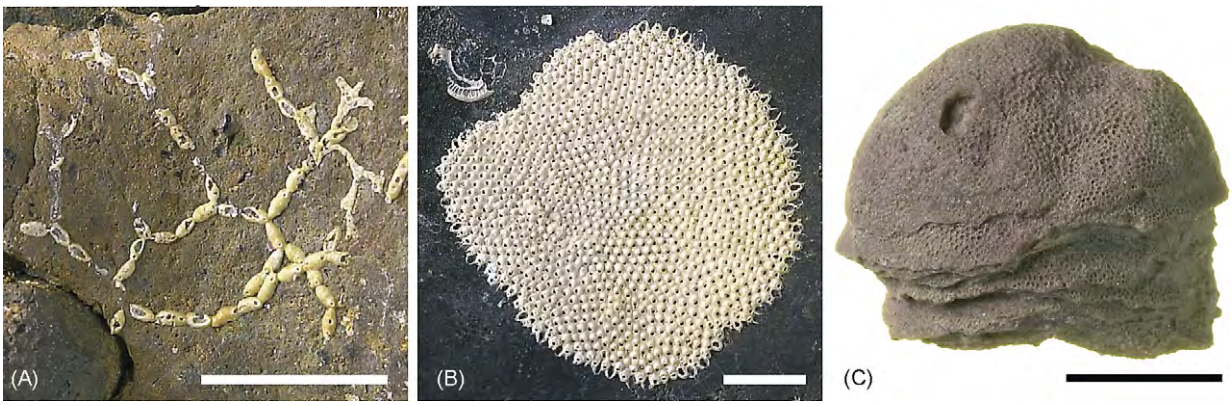


Figure 2 Bryozoan colony forms. (A) Encrusting runner (cyclostome *Voigtopora*, Cretaceous, England, scale bar 5 mm), (B) encrusting sheet (cheilostome *Porella*, Plio Pleistocene, New Zealand, scale bar 2 mm), and (C) dome shaped colony (trepostome '*Dianulites*', Ordovician, Russia, scale bar 10 mm).

runner-like, with uniserially or pauciserially arranged zooids forming a branching colony (Figure 2A). These two kinds of encrusting colonies represent alternative survival strategies. Sheet-like colonies are compact and place a premium on the defence of the living space that they occupy (phalanx strategy), whereas runner-like colonies disperse their zooids widely across the substrate, maximizing the chances that some will survive (fugitive or guerrilla strategy). New zooids are sometimes budded on top of older zooids to give a multilayered colony. Among erect bryozoans, common colony forms are tree-like (Figure 3A), frondose (Figure 3B), and net-like or reticulate (Figure 3C). Some erect colonies are rigid, others flexible, by virtue of having either elastic joints linking stiff internodes or a weakly mineralized skeleton. Flexible colonies are less likely to snap under the bending stresses imposed by ambient currents. Free-living colonies outgrow their substrates (e.g. sand grains) and rest directly on the sediment surface. They include massive dome-shaped colonies (Figure 2C), commonest in the Lower Palaeozoic, and smaller concavo-convex colonies (lunulites) found in the Cretaceous–Holocene. The latter support themselves above the seabed using structures called setae and in one genus (*Selenaria*) are even able to 'walk'.

Fragmentation provides an alternative mode of colony formation, especially in erect species with fragile branches that are easily broken off, transported away and able to resume colony growth elsewhere. Colonies formed in this way are clones. The ability to produce new colonies by fragmentation depends on colonies being able to survive after the death of some of their constituent zooids. The death of zooids within still-living colonies (partial mortality) is routine in bryozoans. The oldest zooids within large long-lived colonies are frequently dead (or at least dormant and no longer feeding), and predators can kill individual zooids in living colonies. If the loss of zooids to predators or other causes exceeds the budding of new zooids at the growing edges of the colony, then 'negative growth' occurs, and the colony actually decreases in size with time and age. Another complexity in bryozoan life cycles is introduced by the potential for different colonies of the same species to fuse with one another if they come into contact during growth. The fused colony will be a genetic chimera unless the original colonies were clonal, which may be the case in cyclostome bryozoans where embryos in brood chambers divide into many genetically identical larvae, a process called polyembryony.

Cyclical polypide degeneration and regeneration is characteristic of bryozoans. This process occurs when the polypide – essentially the gut, lophophore, and certain associated tissues – breaks down, giving a mass called a 'brown body' that can be expelled. A new polypide is subsequently formed within the body walls and feeding recommences. Individual zooids may pass through several polypide cycles during their lifespan. The 'brown deposits' found in some fossil bryozoans with zooidal chambers filled by diagenetic cement are probably remnants of non-expelled brown bodies.

Polymorphism and Within-colony Zooidal Variations

One of the most intriguing features of bryozoans is zooidal polymorphism. This is where the colony contains more than one type of zooid, each having a distinctive morphology and fulfilling a particular

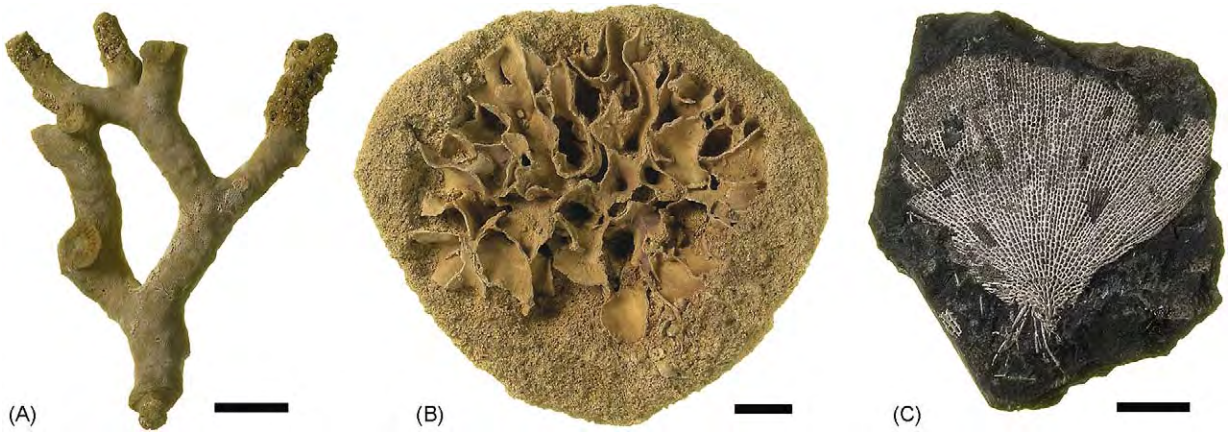


Figure 3 Erect bryozoan colony forms. (A) Tree like (cyclostome *Heteropora*, Pliocene, England, scale bar 10 mm), (B) frondose (cheilostome *Pentapora*, Pliocene, England, scale bar 20 mm), and (C) reticulate (fenestrate *Fenestella* s.l., Carboniferous, Wales, scale bar 10 mm).

functional role. Feeding zooids are known as autozooids, and non-feeding polymorphs are known as heterozooids. Heterozooids include the avicularia found in the majority of cheilostome bryozoans. Relative to their overall size, avicularia have large opercula called mandibles, which act like jaws and have been observed capturing and holding onto predators until they die. Some avicularia replace autozooids in the normal budding sequence of the colony and are of similar size, whereas others are much smaller and are budded onto the surfaces of the autozooids, often positioned so as to guard the vulnerable aperture. A subcategory of avicularia, vibracula, have elongated mandibles (setae) apparently used for cleaning detritus from the colony surface and also forming the appendages of lunulites. Polymorphs called gonozooids function as larval brood chambers in cyclostome bryozoans. Other heterozooids found in extant bryozoans include rhizozooids (roots) and kenozooids (space fillers). Palaeozoic stenolaemate bryozoans with long tubular zooids may contain polymorphs of smaller diameter between the autozooids. If these polymorphs possess diaphragms (cross partitions), they are called mesozooids, otherwise they are termed exilazooids. It is likely that both mesozooids and exilazooids functioned passively as space fillers to prevent overlap between the lophophores of the feeding zooids.

Aside from polymorphism, there are three other sources of variation in zooids within colonies: ontogeny, astogeny, and microenvironment. Ontogenetic variation occurs because zooids within a colony are of different ages and hence differ in their developmental stage: newly budded zooids may have thinner skeletons than older zooids and can look very different. Astogeny refers to the development of the colony: zooids budded early in astogeny are often smaller and simpler than those budded later during colony growth. Additional variations within colonies can be related to differences in the microenvironment experienced by the zooids; for example, obstacles on the substratum surface may cause localized distortion of zooids in encrusting colonies.

Skeleton

The mineralized skeleton of bryozoans consists of an extracellular secretion of tiny crystallites of calcium carbonate set in an organic matrix. There are clear similarities with the shells of brachiopods and molluscs in the calcareous composition of the skeleton, the morphology of the crystallites, and the mode of secretion. However, bryozoan skeletons are typically more intricate and comprise two fundamentally different types of wall. Exterior walls resemble brachiopod and mollusc shells in growing on an organic template (cuticle) and thickening through accretion on one side only. Interior walls develop initially as invaginations of the epithelia forming the exterior walls, lack a cuticular layer, and thicken through accretion on both sides. Exterior walls usually occur at the interface between the bryozoan and the external environment, whereas interior walls typically form the boundaries between zooids or intrazooidal structures. Some stenolaemate bryozoans have skeletal rods (styles) embedded in their interior walls, often with a cone-in-cone internal structure, and projecting from the colony surface as spines.

Whether exterior or interior, each skeletal wall typically contains several layers of contrasting fabric. Three basic categories of wall fabric are: finely granular; lamellar, with crystallite long axes subparallel to the wall surface; and perpendicular, with crystallite axes at a high angle to the wall surface. Granular fabrics have very small crystallites and in some fossils are apparently neomorphic in origin. Lamellar fabrics often comprise lath-like crystallites imbricated at low angles or stacks of platey crystallites. Perpendicular fabrics are usually fibrous.

Mineralogically, most bryozoan skeletons are made of calcite, but an increasing proportion of cheilostomes from the Upper Cretaceous onwards use aragonite, either alone or overlying a calcite basal skeleton. The amount of magnesium in bryozoan calcite varies from low (less than 4% wt) to high (more than 10% wt). High-magnesium calcite skeletons are found in some post-Palaeozoic cheilostomes and may also have been present in a few Palaeozoic trepostomes and cystoporates with granular neomorphized walls containing microdolomites. Slight deviations from equilibrium in carbon and oxygen isotopes have been detected in bryozoan skeletons, possibly due to kinetic effects associated with rapid growth rates.

Classification and Evolutionary History

Three classes of bryozoans are generally recognized: Phylactolaemata, Gymnolaemata and Stenolaemata (Table 1). Extant phylactolaemates live entirely in freshwater lakes and slow-moving rivers. They are of relatively low diversity (about 80 species) and lack mineralized skeletons. Body fossils of phylactolaemates are unknown, but the chitinous disc-shaped statoblasts manufactured by phylactolaemates for overwintering and dispersal, like plant seeds, are very occasionally recorded in the fossil record as far back as the Permian. Anatomical evidence indicates

Table 1 Major bryozoan groups with their main skeletal features and geological ranges

Class	*Order*	*Skeletal morphology*	*Geological range*
Phylactolaemata	[none]	(Soft bodied)	Permian Recent
Gymnolaemata[a]	Ctenostomata[a]	(Soft bodied)	Ordovician Recent
	Cheilostomata	Box shaped zooids often with complex frontal morphology and intricately shaped apertures; avicularia; ovicells; very diverse colony forms; occasionally with monticules	Jurassic Recent
Stenolaemata	Cyclostomata[a]	Cylindrical zooids often with exterior frontal walls; porous interior walls; gonozooids; diverse colony forms; occasionally with monticules	Ordovician Recent
	Trepostomata[a]	Cylindrical zooids without exterior frontal walls; nonporous interior walls; diaphragms and styles common; some with mesozooids or exilazooids; tree like, sheet like or mound like colony forms, some frondose colonies; monticules common	Ordovician Triassic
	Cystoporata	Cylindrical zooids without exterior frontal walls; interior walls nonporous in most; extrazooidal vesicles and/or lunaria present; some with exilazooids; sheet like or mound like colony forms, some tree like and frondose colonies; some with monticules	Ordovician Triassic
	Cryptostomata	Box shaped to cylindrical zooids without exterior frontal walls, regularly arranged; interior walls nonporous; tree like or frondose colony forms, some articulated	Ordovician Triassic
	Fenestrata	Box shaped zooids without exterior frontal walls, regularly arranged; interior walls nonporous; narrow branched erect colonies, often reticulate or pinnate with zooids opening on only one side of the branch	Ordovician Permian

[a]Probable paraphyletic taxa.

that phylactolaemates are the most primitive of bryozoans. Gymnolaemates are a paraphyletic grouping within which the third bryozoan class – stenolaemates – is nested. Two orders of gymnolaemates are recognized, the soft-bodied ctenostomes and the calcareous cheilostomes.

Ctenostomes

Ctenostomes are predominantly marine, but a small number of the approximately 300 extant species are found in freshwater. Their colonies tend to be inconspicuous and 'weedy', often forming delicate encrustations with prostrate zooids sometimes aligned along a colonial stolon, although more robust encrusting and erect colonies occur in some gelatinous species. Despite their lack of hard parts, ctenostomes can be preserved as fossils in two ways. Some species bore into calcareous substrates, especially mollusc shells. These leave a trace fossil comprising branching arrays of thread-like tunnels bearing zooids at regular intervals, which communicate with the surface of the substrate via a small circular aperture. Although most examples of ctenostome borings are of Mesozoic or Cenozoic age, the earliest examples come from the Arenig Stage of the Ordovician. The second mode of preservation of ctenostomes is bioimmuration. This is where organic overgrowth, by an oyster for example, leaves a natural mould of the bryozoan on the underside of the overgrowing organism. Bioimmured ctenostomes occur quite commonly in the post-Palaeozoic. Most have oval zooids arranged in branching uniserial chains. Many bioimmured ctenostomes from the Jurassic have very similar zooids and colony growth patterns to cheilostomes, and some also possess an operculum, normally regarded as diagnostic of cheilostomes, supporting the hypothesis that cheilostomes originated from a ctenostome ancestor that acquired a mineralized skeleton in the Late Jurassic.

Cheilostomes

Over 1000 genera of cheilostomes have been recognized, and, with an estimated 4300 extant species, cheilostomes are the dominant group of present-day bryozoans. As noted above, however, this gymnolaemate order made a relatively late appearance about 155 Ma ago and even then did not become abundant in the fossil record until the mid-Cretaceous. Cheilostomes underwent an explosive radiation during the Late Cretaceous. The onset of this spectacular diversification, which was accompanied by equally impressive increases in morphological disparity and abundance, coincided with the appearance of ovicells, hood-like chambers for brooding larvae (**Figure 4A**). This correlates with the evolution of short-lived brooded larvae from the long-lived nonbrooded larvae characteristic of more primitive cheilostomes. Such a switch in larval type is predicted to have had

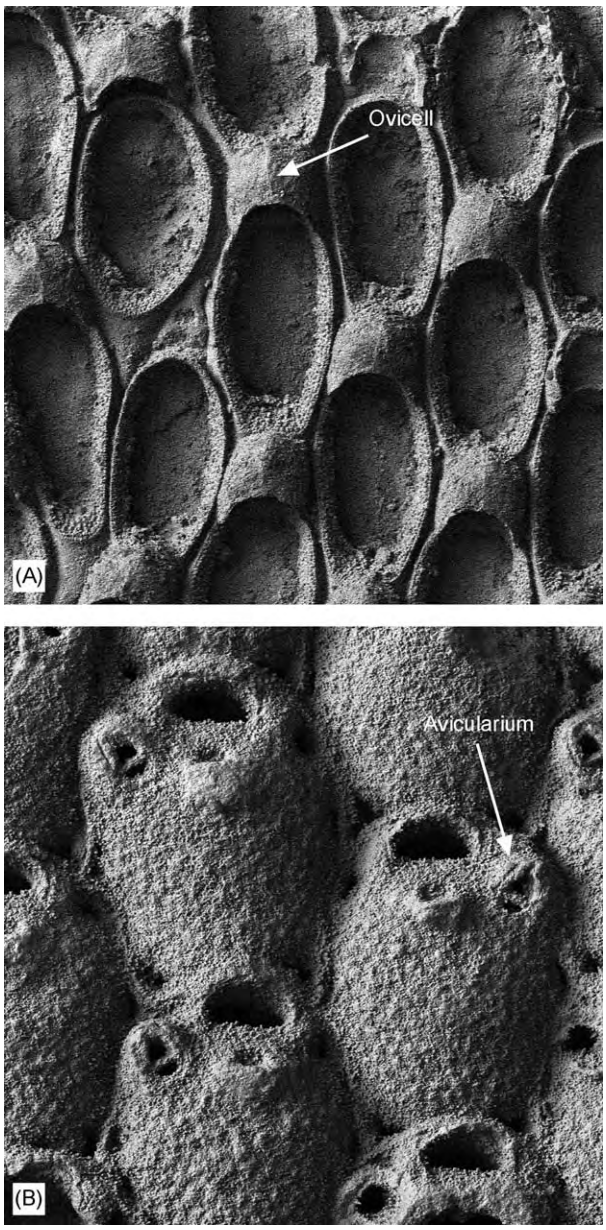


Figure 4 Fossil cheilostome bryozoan zooids (scanning electron micrographs). (A) Group of zooids from a colony of *Wilberto pora*, an anascan cheilostome with a mostly uncalcified frontal area; note the larval brood chambers (ovicells) (Cretaceous, Texas, USA; magnification × 55). (B) Zooids of *Microporella*, an ascophoran cheilostome with a cryptocystal frontal shield and a small semielliptical orifice; note the tiny adventitious avicularia (Pliocene, New Zealand; magnification ×65).

profound consequences for dispersal, population genetic structure, and speciation rate. The good larval dispersal and gene flow of nonbrooders is predicted to result in genetically homogeneous species that are unlikely to become fragmented into incipient new species, whereas poor larval dispersal and gene flow in brooders should give rise to genetically heterogeneous species, which are more prone to speciate.

At the colony level, cheilostomes exhibit a vast range of forms, some homeomorphic with colony forms seen in stenolaemate bryozoans but others unique to the cheilostomes. The most primitive cheilostomes have simple encrusting colonies with zooids arranged one after the other along ramifying branches or irregularly in lobate sheets. Somewhat more advanced cheilostomes show a more regular patterning of the zooids in coherent sheets (**Figure 2B**). While the majority of cheilostome species have encrusting colonies, erect species can be visually striking and are more likely to be noticed as fossils. Some have tree-like colonies, others are frondose (**Figure 3B**), and yet others are reticulate with lace-like colonies. Articulated colonies and flexible erect colonies with light calcification (and consequently poor fossilization potential) also occur among cheilostomes. Small spindle-shaped colonies anchored into particulate sediments by rootlets are found in some species, while the cap-shaped concavoconvex colonies of free-living lunulites are without close parallel among the other bryozoan orders.

The wide range of zooid-level skeletal features in cheilostomes provides useful characters for taxonomy. Indeed, species identification is often possible on the basis of just a few zooids, while colony level features are generally of less importance than in the stenolaemates. Cheilostome zooidal skeletons are typically box shaped and are rounded-rhombic or rectangular in frontal view (**Figure 4**). The vertical walls of the zooid are completely calcified, except where penetrated by pores that allow communication between adjacent zooids (**Figure 1**). Basal walls may be completely calcified or have a central uncalcified window. However, it is the frontal walls of the zooid that provide most of the variation useful in taxonomy. These may be largely uncalcified, as in most anascan cheilostomes (**Figure 4A**) where the frontal surface of the zooid is occupied by a nonmineralized frontal membrane with, at its distal end, the orifice. Through time an increasing proportion of cheilostomes, including the ascophorans, which predominate at the present day, have evolved protective frontal shields covering the frontal membrane (**Figure 4B**). Frontal shields are constructed of three main types of skeleton: gymnocyst is a planar exterior wall; spinocyst is another type of exterior wall comprising hollow spines that grow centripetally to overarch the frontal membrane; and cryptocyst is an interior wall, often with a pustulose surface texture. Many cheilostome zooids have erect spines, which are often hollow, articulated basally, and located near the orifice or around the inner edge (mural rim) of the gymnocyst.

Avicularian polymorphs are widespread among cheilostomes.

Stenolaemates

Stenolaemates were the dominant bryozoans throughout the Palaeozoic, where they are represented by five orders – Trepostomata, Cystoporata, Cryptostomata, Fenestrata, and Cyclostomata – of which only the cyclostomes are extant. The oldest stenolaemates are trepostomes reported from the Tremadoc of China. By the Arenig, representatives of the other four orders had appeared, initiating a phase of exponential diversification that continued into the Caradoc. The origin of the mineralized skeleton from a ctenostome-grade ancestor may have been the primary trigger for stenolaemate radiation, opening the way for the evolution of a much greater variety of colony forms. A plateau of stenolaemate family diversity characterizes the remainder of the Palaeozoic, before the catastrophic decline in diversity marking the end-Permian mass extinction (Figure 6). Although only one order (Fenestrata) is thought to have become extinct at this time, stenolaemates were rare in the Triassic, and the Palaeozoic 'holdover' orders (trepostomes, cystoporates and cryptostomes) did not survive beyond this period. Bryozoan faunas of Jurassic and Early Cretaceous age, before the rise of the cheilostomes, consist predominantly of cyclostomes.

Trepostomes Trepostomes (approximately 200 genera) are the 'stony bryozoans' that abound in Palaeozoic deposits such as the Upper Ordovician around Cincinnati in the USA. A few trepostome species have sheet-like encrusting or frondose erect colonies, but most have dome-shaped colonies (Figure 2C) or, more commonly, robust tree-like colonies, which are typically broken into individual branches during fossilization. Trepostome zooids are characteristically long and tubular, terminating in a polygonal aperture (Figure 5A). Many species have

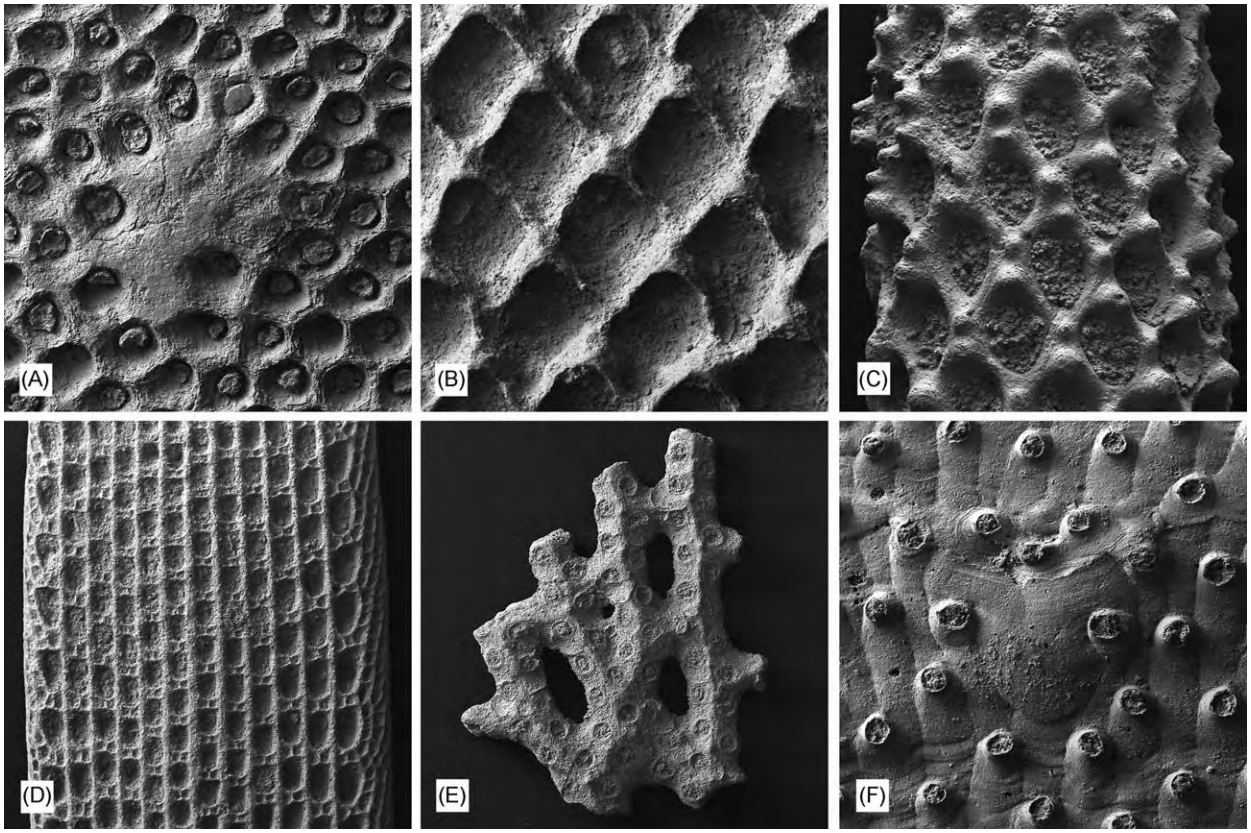


Figure 5 Fossil stenolaemate bryozoans (scanning electron micrographs). (A) Unidentified trepostome with polygonal zooidal apertures surrounding a monticule (Ordovician, Utah, USA, magnification ×30). (B) Diamond shaped zooidal apertures with lunaria in the cystoporate *Favositella* (Silurian, England, magnification ×30). (C) Branch of the rhabdomesine cryptostome *Rhabdomeson* with styles protruding as blunt spines (Carboniferous, Scotland, magnification ×50). (D) Flattened frond of the ptilodictyine cryptostome *Phaenopora* showing autozooids arranged in well defined rows and some smaller polymorphs (Silurian, Sweden, magnification ×15). (E) Fragment of a colony of the fenestrate bryozoan *Fenestella* s.l. with branches bearing two rows of apertures, short dissepiments and fenestrules (Carboniferous, Scotland, magnification ×15). (F) Autozooids and a gonozooid (centre) in the cyclostome *Hyporosopora* (Jurassic, England, magnification ×25).

mesozooids or exilazooids, and monticules are widespread. Styles are often present, as are diaphragms, which may be planar and complete, partial (hemiphragms), or curved (cystphragms). Like those of the other dominant Palaeozoic stenolaemate orders, trepostome skeletons lack frontal exterior walls. Branches of tree-like species usually have an inner thin-walled region (endozone) where zooidal budding is concentrated and zooidal axes are subparallel to the branch axis. This is surrounded by an outer thick-walled region (exozone) where zooid axes are almost perpendicular to the branch surface.

Cystoporates Cystoporates (approximately 90 genera) resemble trepostomes but can be distinguished from them by the presence of one or both of the following features: lunaria – hood-like projections over zooidal apertures (Figure 5B); and extrazooidal vesicles – 'bubbly' skeletal tissue between the zooids. They also often have monticules, sometimes with centres filled by vesicular tissue. Most cystoporates are sheet-like encrusters or dome shaped, but frondose and tree-like forms can also occur. A few have irregular pores in the walls between the zooids.

Cryptostomes Cryptostomes (approximately 80 genera) differ from trepostomes and cystoporates in having zooidal budding focused in restricted zones and zooids that tend to be shorter, giving colonies of a generally more delicate appearance with a regular pattern of apertures on the surface. All cryptostomes have erect colonies, and these are divided into two main sorts, the tree-like colonies of rhabdomesines (Figure 5C) and the frondose or flat-branched colonies of ptilodictyines (Figure 5D), which have a median lamina from which the zooids are budded. Articulations occur in some cryptostomes, either at the base of the colony in certain ptilodictyines or at regular intervals throughout the bushy colonies of one family of rhabdomesines. Styles are frequently present, and there may be more than one type of style per colony, some protruding from the colony surface as spines (Figure 5C). Short diaphragms (hemisepta) occur close to the aperture in many species.

Fenestrates Fenestrates (approximately 150 genera) are typified by short and often almost box-shaped zooids budded in a precise pattern. Colonies always have narrow erect branches bearing between two and about 10 series of zooids with subcircular apertures opening on only one side of the branch; the side without apertures may be reinforced by thick calcification. Often branches bifurcate and coalesce to form a planar mesh, but a similar mesh can be produced as a result of solid skeletal dissepiments linking branches at intervals (Figures 3C and 5E). In both cases, the holes in the mesh would have functioned as conduits for water flow during feeding. Pinnate colonies also occur. Probably the best-known and most diverse group of fenestrates is the Family Fenestellidae (fenestellids), which dominates many Carboniferous and Permian bryozoan assemblages. Fenestellid colonies have two rows of zooids along the branches, which are linked by dissepiments, and often develop nodes along the crests of the branches between apertures. In some genera, these nodes are greatly enlarged to form a superstructure affording protection to the zooids beneath. The peculiar screw-like colonies of *Archimedes* make it the most distinctive of all fenestellids. The screw is a massively thickened central axis that supports a helically wound mesh. Many populations of *Archimedes* relied on asexual propagation, with broken-off screws coming to rest prostrate on the seabed before producing daughter screws on their upper sides that restored vertical growth.

Cyclostomes Cyclostomes (approximately 370 genera) are first recorded in the Arenig but were uncommon until the Jurassic (Figure 6). Palaeozoic cyclostomes differ from contemperaneous orders in having exterior frontal walls. All Palaeozoic species have delicate 'weedy' colonies, which can be erect or encrusting and runner-like or sheet-like. The evolutionary zenith for cyclostomes was between the mid-Jurassic and the Late Cretaceous when a much greater spectrum of colony forms evolved. Species with exterior frontal walls (Figure 5F) were joined by others without frontal walls that are convergent with the extinct orders of stenolaemates of the Palaeozoic (Figure 3A). Cyclostomes possess regularly porous skeletal walls, permitting interzooidal communication, and nearly all species have bulbous gonozooids (Figure 5F) for brooding the polyembryonic larvae. Apertures of autozooids can be polygonal or subcircular and are often surrounded by a tubular prolongation (peristome).

Occurrence

Sedimentary environments most favourable to bryozoans are marine, shallow water, fully oxygenated with good circulation furnishing planktonic food resources, well provisioned with stable hard or firm substrates, and lacking a high influx of particulate sediment. Although bryozoans do occur intertidally, the majority of species are subtidal and inhabit the mid- to outer continental shelf. While

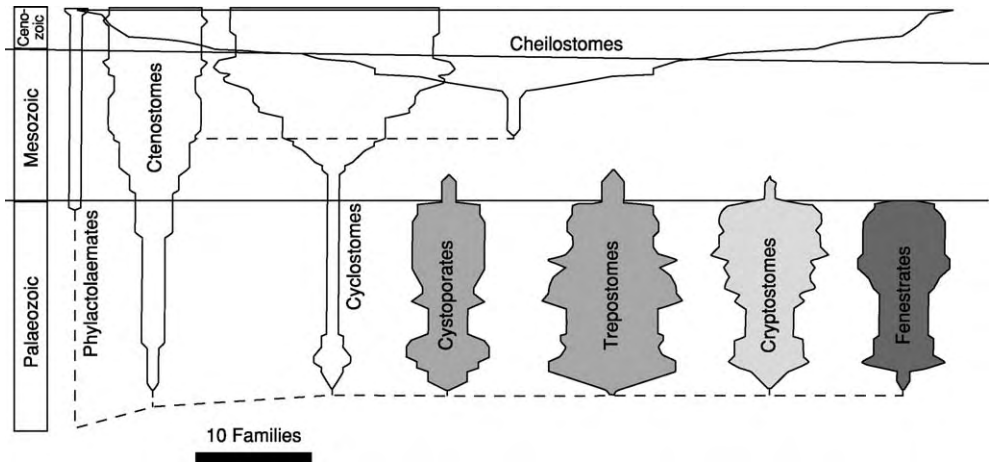


Figure 6 Diagram of changing family diversities in the bryozoan orders. Tentative phylogenetic relationships are indicated by dashed lines.

a few modern species have restricted bathymetric distributions, many have depth ranges exceeding 100 m. High diversities of bryozoans can be found in all major climatic zones, but tropical species tend to be inconspicuous, typically occurring as small encrusting colonies in cryptic habitats such as the undersides of stones and shells. Bryozoans are a major contributor to modern cool-water carbonates, such as those forming on the southern Australian continental shelf.

In the geological record bryozoans are often found in association with carbonate facies. The best collecting opportunities are in unconsolidated fine-grained clastics, where surface picking, bulk sampling, and gathering of shells that may be encrusted with bryozoans can yield diverse faunas. In some instances, bryozoans are present in rock-forming abundance, usually comprising branch fragments of parautochthonous erect colonies (Figure 7) and less often small bioherms of robust in-situ colonies. Bryozoan limestones in the Cenozoic are typically clean-washed cross-bedded calcarenites or calcirudites. Mud mounds, including the Waulsortian reefs of the Carboniferous, often contain bryozoans. It has been suggested that the bryozoans assisted in mud sedimentation by acting as baffles and through the production of faecal pellets.

Bryozoan-rich deposits show a striking change in latitudinal distribution through time, from a panlatitudinal pattern in the Palaeozoic to an extratropical distribution in the post-Palaeozoic. The apparent displacement from the tropics of bryozoans contributing significant amounts of carbonate sediment may be related to the end-Permian mass extinction event, which had a severe effect on stenolaemate orders, and/or to increasing levels of colony destruction by grazing predators and exclusion by fast-growing corals and calcareous algae in the post-Palaeozoic.

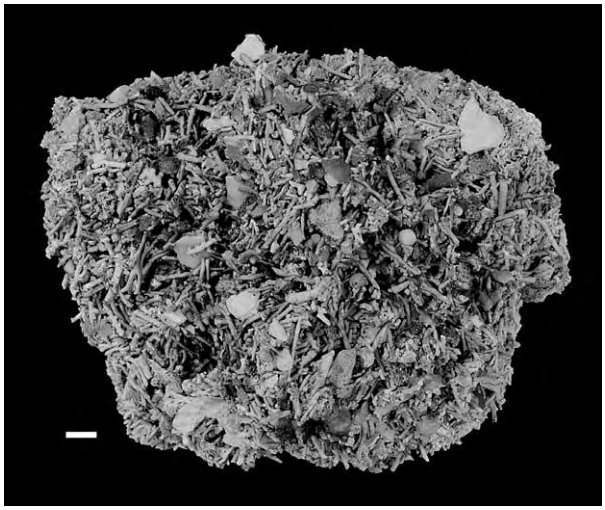


Figure 7 Bryozoan limestone from the Eocene of North Carolina, USA. A poorly cemented facies of the Castle Hayne Limestone consisting mainly of broken branches of tree like cyclostome and cheilostome colonies. Scale bar 10 mm.

The continuous presence of bryozoans through thick sequences deposited over long intervals is unusual. Instead, they tend to be patchily distributed, which is one of the factors that has hindered their use in biostratigraphy. Another limitation arises from typically long species durations; e.g. for cheilostomes currently living in the Mediterranean Sea and represented in the fossil record, average species longevity is about 15 Ma.

Palaeoecology

Much interest has focused on the use of bryozoan growth forms in palaeoenvironmental analysis. Pioneering work suggested that particular growth forms were adapted to particular hydrodynamic and sedimentational regimes. However, these findings are being re-evaluated in the light of the frequent occurrence of many different growth forms together. Nonetheless, the ratio of encrusting to erect species does decline with water depth in modern environments, while a dominance of either erect articulated colonies or free-living colonies usually points to relatively high sedimentation rates and/or a lack of sizeable substrates.

Shallow-water colonies of some erect species have been shown to possess thicker branches than deeper water colonies, giving the potential for detecting depth gradients in the geological record. The plasticity in colony form evident within some species is known to be ecophenotypic. For example, in the Mediterranean species *Schizoporella errata*, mound-like colonies develop in fast flow, and hollow branching colonies develop in slow flow. Variation in zooid size within species is inversely proportional to temperature. Two palaeoclimatic applications result from this relationship. First, comparisons of mean zooid size between colonies can yield information on the relative temperatures in which they grew, data that may be calibrated to an absolute value if the species concerned is extant. Second, zooid size within perennial colonies tracks seasonality – the higher the coefficient of variation in zooid size, the greater the annual range of temperature experienced by the colony.

With the exception of small circular borings possibly made by tiny gastropods, fossil bryozoans seldom show clear evidence of having suffered predation. Other biotic interactions are, however, frequently fossilized. Symbioses between bryozoans and hermit crabs are recorded back to the Jurassic. The crabs themselves are not fossilized, but the shelters provided by their bryozoan symbionts are diagnostic, forming thick encrustations on the surfaces of gastropod shells and extending the helicospiral coiling beyond the original shell aperture. Competitive interactions for substrate space between bryozoans and other encrusters with hard skeletons are evident on many fossil shells and clasts. Such competition between cyclostome and co-occurring cheilostome bryozoans has been documented by tallying overgrowths when sheet-like colonies of the two orders meet. The results have shown that cheilostomes have maintained a remarkably constant competitive advantage over cyclostomes for the past 100 Ma, on average winning about two-thirds of encounters.

See Also

Biological Radiations and Speciation. **Evolution**. **Fossil Invertebrates:** Brachiopods; Corals and Other Cnidaria. **Palaeoecology**. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups').

Further Reading

Boardman RS and Cheetham AH (1987) Phylum Bryozoa. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates,* pp. 497 549. Oxford: Blackwell.

Boardman RS, Cheetham AH, Blake DB, *et al.* (1983) Bryozoa (revised). Volume 1. In: Moore RC and Robison RA (eds.) *Treatise on Invertebrate Paleontology. Part G,* p. 1 625. Boulder, Colorado and Lawrence, Kansas: Geological Society of America and University of Kansas Press.

Hageman SJ, Bock PE, Bone Y, and McGowran B (1998) Bryozoan growth habits: classification and analysis. *Journal of Paleontology* 72: 418 436.

Jackson JBC and Cheetham AH (1994) Phylogeny reconstruction and the tempo of speciation in cheilostome Bryozoa. *Paleobiology* 20: 407 423.

McKinney FK (1986) Historical record of erect bryozoan growth forms. *Proceedings of the Royal Society of London* B 228: 133 148.

McKinney FK (1995) One hundred million years of competitive interactions between clades: asymmetrical but not escalating. *Biological Journal of the Linnean Society* 56: 465 481.

McKinney FK and Jackson JBC (1989) *Bryozoan Evolution.* London: Unwin Hyman.

O'Dea A and Jackson JBC (2002) Bryozoan growth mirrors contrasting seasonal regimes across the Isthmus of Panama. *Palaeogeography, Palaeoclimatology, Palaeoecology* 185: 77 94.

Ryland JS (1970) *Bryozoans.* London: Hutchinson.

Smith AM (1995) Palaeoenvironmental interpretation using bryozoans: a review. *Geological Society of London Special Publication* 83: 231 243.

Taylor PD (1999) Bryozoa. In: Savazzi E (ed.) *Functional Morphology of the Invertebrate Skeleton,* pp. 623 646. Chichester: Wiley.

Taylor PD and Allison PA (1998) Bryozoan carbonates in space and time. *Geology* 26: 459 462.

Woollacott R and Zimmer RL (eds.) (1977) *Biology of Bryozoans.* New York: Academic Press.

Corals and Other Cnidaria

C T Scrutton, Formerly University of Durham, Durham, UK

Introduction

The phylum Cnidaria encompasses a varied group of primitive, dominantly marine, diploblastic metazoans of which the corals are by far the most important in the fossil record. Other cnidarians generally have poor preservation potential but nevertheless exist as scattered fossils extending back probably to the Late Precambrian.

Cnidarian Organization and Classification

All cnidarians consist fundamentally of a sac-like body with a mouth surrounded by tentacles (Figure 1). There are two variants: when the sac rests on its aboral surface with the mouth uppermost, it is a polyp; when it floats in the water with the mouth facing downwards, it is a medusa. When both forms exist in the same species, medusae are formed on the polyp, from which they separate (bud off) and produce male and female gametes. These fuse to form a free-floating larval stage, a planula, which settles and metamorphoses into a new polyp. The phylum is also characterized by the presence of stinging cells for the capture of prey (nematocysts), which are particularly abundant in the tentacles and which give the phylum its name.

There are three main subdivisions (classes) of the Cnidaria, which differ principally in the interior structure of the sac, the enteron or gut, and in the prominence of the two body forms (Figure 2). The Hydrozoa have no internal partitions in the gut, and in most species both polyps and medusae are well developed. The Scyphozoa have four radial flanges extending towards the centre of the gut, and the medusoid, or jellyfish, stage is dominant. The polyp is reduced to a specialized factory for producing medusae (scyphistoma). In the Anthozoa, the gut is subdivided in a more complex and varied way by radial partitions called mesenteries, whilst the medusoid stage is entirely absent, and gametes are produced directly by the polyp. The polyp stage may become modular by budding in both hydrozoans and anthozoans. Anthozoan colonies are simple, usually with no differentiation of the modules. However, in the Hydrozoa, modules may become specialized for different functions such as feeding and reproduction. The ultimate expression of this is the formation of complex colonies of specialized modules, not only of polypoid but also of medusoid origin, which function as superindividuals. These are the chondrophores and siphonophores, represented respectively in modern oceans by *Vellela* (by-the-wind-sailor) and *Physalia* (Portuguese Man-o'-War).

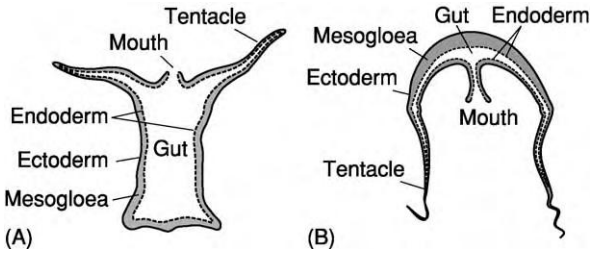


Figure 1 Basic cnidarian anatomy. (A) Polyp. The polyp is shown without partitions in the gut, as in the Hydrozoa. (B) Medusa. The jelly like appearance of jellyfish is due to the thickening of the mesogloea, gelatinous connective tissue separating the ectoderm and endoderm in the roof of the medusoid bell.

The corals are anthozoans that have secreted a calcium carbonate skeleton beneath the polyp or polyp colony. A few hydrozoan polyp colonies also secrete a carbonate skeleton. Some other anthozoans secrete spicules or rarely coral-like skeletons, but most are soft-bodied – for example, the sea anemones. Some hydrozoan polyps and scyphozoan scyphistomas may have an external organic periderm, but the medusae are essentially soft-bodied. The bias in the cnidarian fossil record is directly related to the relative skeletonization of the different groups.

Rare solitary coral-like fossils from the Lower Cambrian of Siberia have been assigned to a fourth, extinct, class of cnidarians, the Hydroconozoa.

Geological History

Precambrian Origins

All three of the principal classes have been claimed to be present in the Late Precambrian Ediacaran fauna (**Precambrian:** Vendian and Ediacaran). Here, internal and external moulds of nonmineralized organisms are preserved in siltstones and fine sandstones. They are controversial, and alternative interpretations have been advanced for these fossils. However, some are convincingly cnidarian. As well as unassigned jellyfish, there are others of likely scyphozoan affinity (*Kimberella*; Figure 3B) and structures reasonably interpreted as the floats of hydrozoan chondrophores (*Ovatoscutum* and *Chondroplon*). The most striking anthozoan is the

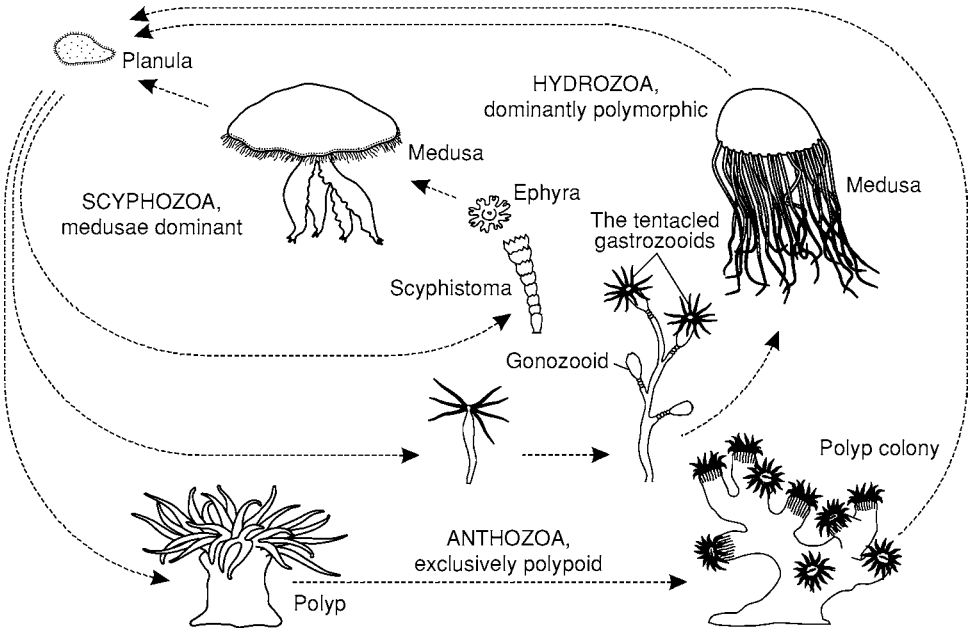


Figure 2 Life cycles of the three cnidarian classes. In the polymorphic Hydrozoa, the polypoid stage is usually colonial, often with specialization of individuals. The tentacled gastrozooids are specialized for feeding and the gonozooids for the production of medusae. The polyp stage in Scyphozoa is reduced to a benthonic structure that buds off medusae from its distal surface. Anthozoans are exclusively polypoid, and those that secrete a calcium carbonate skeleton constitute the corals. Adapted from Hill D and Wells JW (1956) Cnidaria general features, p. 7, **Figure 4**. In: *Treatise on Invertebrate Paleontology*, *part F*. Courtesy of and © 1956, The Geological Society of America and The University of Kansas Press.

sea pen *Charniodiscus* (**Figure 3A**). Other structures may represent the remains of cnidarians that are now extinct.

Unfortunately, even assuming their correct interpretation, these remains throw no light on the relationships within the Cnidaria, but suggest an earlier period of diversification not represented in the fossil record, at least to date. The origin of the classes has to be inferred by indirect means and has been the subject of much discussion. The internal structure of the gut and the character of their nematocysts suggest that the Hydrozoa are the most primitive cnidarians and the Anthozoa are the most advanced. The most favoured scenario envisages a hydrozoan medusa, derived from a planuloid ancestor, as the ancestral cnidarian, with the polyp evolving later as a prolonged juvenile stage.

Phanerozoic Diversity

The type of preservation represented by the Ediacaran fauna is not known after the Precambrian. This is attributed to the rise of active macrophagous predators and scavengers, disruption by burrowing organisms, and possibly the increasing influence of the meiofauna all contributing to the early breakdown of soft tissue in comparable environments. The fossil record of soft-bodied cnidarians becomes sparse and patchy.

Hydrozoa The Hydrozoa are represented by a very few, sometimes doubtfully assigned, medusae such as the questionable *Crucimedusina* (Late Carboniferous) and the more convincing *Hydrocraspedota* (Middle–Upper Jurassic) and *Kirklandia* (Lower Cretaceous). Hydroid polyp colonies are better represented. A rich fauna is known from the Ordovician and Silurian of Scandinavia, some so well preserved that the scleroprotein periderm can be recovered from the calcareous matrix by acid digestion. In younger rocks, hydroid colonies commensal with serpulid worms have been preserved as moulds and casts by overgrowth of the basal stoloniferous network by the calcareous worm tube (*Protulophila gestroi*, Mesozoic – Early Tertiary, Europe and the Middle East; **Figure 3D**). *Drevotella*, a freshwater hydroid (rare even among living hydroids), is recorded from the Upper Carboniferous Mazon Creek Lagerstätte of Illinois.

Structures interpreted as the floats of chondrophores are relatively common in the Early and Middle Palaeozoic. Examples are *Scenella* (Lower Palaeozoic; previously referred to the Mollusca) and *Plectodiscus* (Lower Devonian). Preservation is helped by the stiff scleroprotein cover surrounding the float. After the Carboniferous, however, for reasons unknown, only a single further example is recorded. The siphonophores appear to have left no fossil record at all.

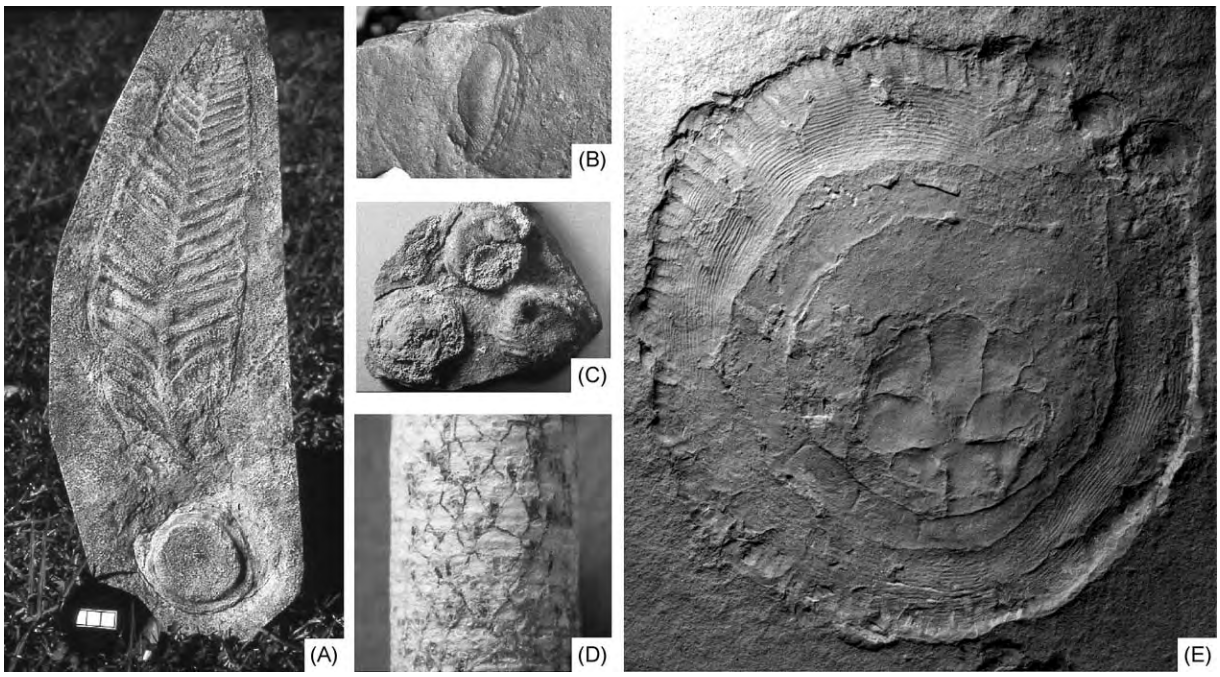


Figure 3 Representative non coralline cnidarians. (A) *Charniodiscus arboreus*; Precambrian, Vendian, Flinders Ranges, South Australia, ×0.2. (B) *Kimberella quadrata*, flank of bell with serial gonads; Precambrian, Vendian, Ediacara, South Australia, ×0.4. (C) *Bergaueria* sp., undersurface of fine sandstone bed with pedal impressions of anemones; Cambrian, Herrería Formation, Cantabria, Spain, ×0.5. (D) *Protulophila gestroi*, pyritized stolonal network and polyp bases exposed by the decortication of the worm tube *Protula* sp.; Cretaceous, Gault Clay, Ford Place, Kent, UK ×3. (E) *Rhizostomites admirandus*, a scyphomedusa; Upper Jurassic, Solenhofen Limestone, Solenhofen, South Germany, ×0.3.

Finally, there are a few calcified hydrozoan polyp colonies in the fossil record (excluding the Stromatoporata and Chaetetida, both formerly assigned to the Cnidaria but now referred to the Porifera). The Milleporina (*Millepora* is found on living reefs, particularly in the Caribbean) and Stylasterina can be traced back to the Late Cretaceous. The hydroid *Hydractinia*, which is now most commonly found encrusting gastropod shells, extends back to the Eocene and possibly the Mesozoic.

Scyphozoa Convincing post-Ediacaran fossil scyphomedusae are extremely rare and virtually restricted to the Solenhofen Lagerstätte (Upper Jurassic, southern Germany). Here beautiful impressions of *Rhizostomites admirandus* reach a diameter of 0.5 m (Figure 3E). Scyphopolyps (scyphistoma) are also rare, the most convincing being the conical scleroprotein tube of *Byronia* (Upper Cambrian–Middle Ordovician, Canada, Poland), which is very similar to the living scyphopolyp *Stephanocyphus*.

A group of relatively common chitinophosphatic tubular fossils from the Lower Ordovician to the Triassic, the Conulata, have been assigned more or less closely to the Scyphozoa by some, although others favour their separation as a discrete phylum. There is evidence of attachment possibly benthonic as juveniles and a pelagic adult stage, although interpretation of their lifestyle is controversial. Their cnidarian affinity is based on a commonly occurring quadrate cross-section and rare examples of internal ribs or septa in multiples of four, comparable to the tetrameral symmetry of scyphozoans.

Anthozoa The exclusively polypoid anthozoans have a much more complex partitioning of the gut than do other cnidarians. Beneath the mouth, a tube (stomodaeum) extends down into the gut, which is subdivided by radial partitions called mesenteries (Figure 4). Those that extend to meet the stomodaeum are termed complete. The mesenteries are the site of digestion, absorption, excretion, and gonad development, and at least some anthozoans can extrude long mesenterial filaments from the mouth to aid feeding.

The three subclasses are distinguished by the pattern of mesenteries in the gut. Relationships between them are not clear from the fossil record so are instead based on the morphology and development of living forms. The ancestral anthozoan is thought to have been a polyp with six unpaired complete mesenteries, which gave rise to the Ceriantipatharia on the one hand and, with the addition of two more complete mesenteries, to the Octocorallia on the other. The latter stock, with the addition of four incomplete

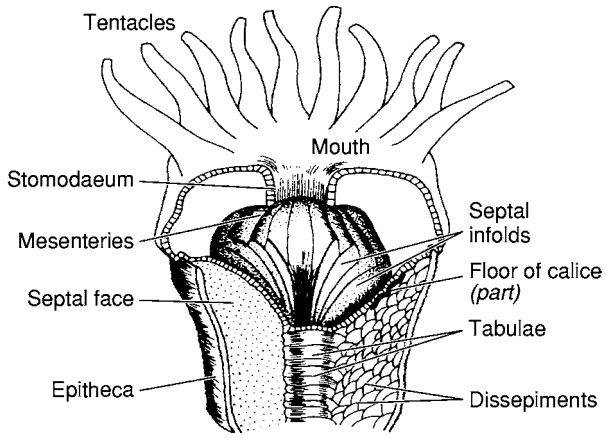


Figure 4 Hypothetical relationship between polyp and skeleton in a solitary rugose coral. Modified from Scrutton (1978, **Figure 2A**) with the permission of Springer Verlag. Scrutton, CT (1978) Periodic growth features in fossil organisms and the length of the day and month. In: Brosche P and Sundermann J. *Tidal Friction and the Earth's Rotation.* Berlin: Springer Verlag.

mesenteries and their arrangement into six pairs, gave rise to the Zoantharia, the subclass that includes the corals and true anemones.

The Ceriantipatharia has few living representatives and one doubtful fossil from the Miocene. However, it includes solitary, burrowing, anemone-like polyps, which may have left as yet unidentified trace fossils. The Octocorallia is much more important as a living group and is slightly better represented by fossils. The Pennatulacea (sea pens), convincingly present in the Precambrian Ediacaran fauna, have a few doubtful records in the Palaeozoic but more credible remains in the Mesozoic. Gorgonians (sea fans) are also sparsely and doubtfully represented in the Palaeozoic, apart from Ordovician material such as *Pragnellia*, but from the Jurassic onwards scattered fossils of the solid calcified horny axes of the polyp colonies are recorded. Alcyonarian spicules have been recorded from the Silurian of Sweden. One group of octocorals, the Coenothecalia, distinguished by a solid calcareous skeleton, includes the 'blue coral' *Heliopora* and has a record stretching back to the Upper Cretaceous. *Heliopora* and the living stoloniferous octocoral *Tubipora* are homoeomorphic with the Palaeozoic tabulate corals *Heliolites* and *Syringopora*, respectively, leading some formerly to classify the Tabulata as octocorals. The similarities are superficial, however, and, although they are anthozoans, the Tabulata are not considered to be closely related. Of the Zoantharians, the fossil record of anemones is extremely poor. The oldest appears to be a recent record from the Cambrian of China. *Mackenzia costalis* (Middle Cambrian Burgess Shale, British Columbia) is now considered to be an actinian anemone but is in need of revision, and Palaeactinia is another questionable actinian (Middle Ordovician, New York). Trace fossils of actinian burrows have also been claimed, including *Intrites* (Vendian) and the relatively common *Bergaueria* (Vendian–Middle Ordovician; **Figure 3C**). *Dolopichnus* (Early Triassic), if correctly interpreted, is the only record of a zoantharian from that interval and has a special significance in the evolution of the corals.

The corals The corals are essentially anemones that have secreted a calcium carbonate skeleton beneath the polyp (**Figure 4**), which then occupies a cup or platform at the top of the skeleton (the calice). The skeleton (corallum) is basically discoidal, conical, or cylindrical, with or without an external wall (epitheca) and radial disposed vertical plates or spines (septa), and usually with horizontal partitions that range from simple flat plates to different combinations of variously shaped plates and vesicles (tabulae and dissepiments). When the coral is a modular colony, each module secretes its own skeleton, the form of which is dependent on whether the modules separate completely, remain in contact, or are only partially separated.

Eight orders of corals are now recognized, of which three are by far the most important: the Rugosa, Tabulata, and Scleractinia (**Figure 5**). They are divided into two broad stocks on the basis of their known or inferred mesenterial and septal development and skeletal mineralogy. The aragonitic Scleractinia and the closely related actinian/corallimorpharian anemone group have a primary arrangement of six paired mesenteries, with subsequent sets of mesenteries inserted cyclically between the preexisting cycles (**Figure 6C and 6D**). This leads to a sequence of 6, 6, 12, 24, 48, etc., paired mesenteries in successive cycles, which is generally reflected in the insertion of septa in the coralla of scleractinian corals. This pattern of septal insertion can be recognized in fossil scleractinians and in two minor extinct orders of aragonitic corals: the Kilbuchophyllida (**Figure 7E**) and less certainly the Numidiaphyllida. The two remaining major groups of corals, the Rugosa and Tabulata, are both extinct. The calcitic Rugosa (**Figure 8**) mostly have well-developed septa, which show a pattern of insertion quite distinct from that in the Scleractinia (**Figure 6A**). They also have six protosepta, reflecting an initial set of six paired mesenteries, but subsequent insertion of major septa is serial in only four of the loci thus defined. In addition, minor septa are usually present between the major septa. There is some dispute concerning the significance and interpretation of this, but by far the

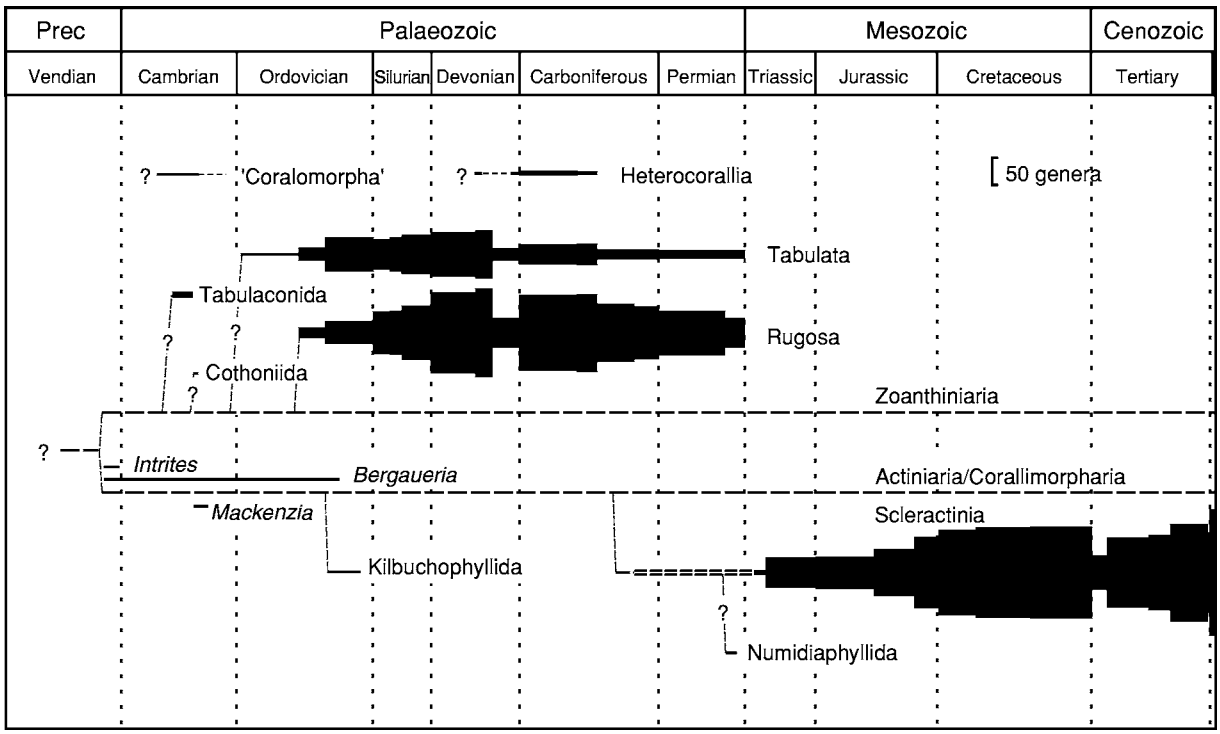


Figure 5 Geological ranges of the zoantharian corals. Their relationships to the two stocks of living anemones are indicated. The double dashed line leading to the Scleractinia notionally indicates its polyphyletic origins in the Middle Triassic. The Hetercorallia are sufficiently dissimilar from the other corals to suggest that they might have evolved from a quite separate anemone stock, which is now extinct. The Coralomorpha is an informal grouping of coral like fossils, the affinities of which are not yet determined. *Intrites* and *Bergaueria* are thought to be actinian trace fossils, and *Mackenzia* is a doubtful fossil actinian anemone. The generic diversities of the Tabulata, Rugosa and Scleractinia are indicated. Modified from Scrutton (1997, **Figure 2**) The Palaeozoic corals, I: origins and relationships. *Proceedings of the Yorkshire Geological Society*, 51: 177 208 with the permission of the Yorkshire Geological Society.

majority opinion is that these corals have an inferred pattern of mesenterial development similar but not identical to that in a small group of living anemones, the Zoanthiniaria (**Figure 6B**). Septa in the Tabulata (**Figure 9**) are also calcitic and are often poorly developed, but one small group has a pattern of insertion similar to that in the Rugosa. Assuming the integrity of the Tabulata as a monophyletic clade, which has been controversial, this suggests that they may, like the Rugosa, be descended from a zoanthiniarian ancestor. Not only have the tabulates been regarded as a mixed bag of unrelated corals, but some if not all of them have been claimed, among other things, to be calcareous sponges. Indeed, some fossils formerly assigned to the Tabulata, such as the Chaetetida, have been properly reclassified as sponges, but the group as currently understood is generally accepted as a valid coral taxon, and favositid tabulates have been found with convincing polyps preserved in the calices (Silurian, Anticosti Island; **Figure 9C**). Of the remaining orders, which are all apparently calcitic, the most important is the Heterocorallia, in which septal insertion is disputed but is quite unlike either pattern described above (**Figure 7F**). Its relationships to other corals are unclear, and it may have evolved from a separate polyp stock that is now extinct. The other minor orders, the Cothoniida and Tabulaconida, have not yielded identifiable patterns of septal insertion but are generally considered to be closest to the zoanthiniarian stock. It seems likely that the two polypoid stocks, zoanthiniarian and actinian/corallimorpharian, had already diverged by the beginning of the Cambrian (**Figure 5**).

Corals first appear in the Lower Cambrian. The Tabulaconida (*Moorowipora*, *Arrowipora*, *Tabulaconus*) have features similar to those of tabulate corals, to which some have referred them. However, the origination of new corallites (increase), at least in *Moorowipora* (**Figure 7A–C**), is more like that found in rugosans. Altogether, they possess features that do not suggest that they are ancestral to any of the Ordovician corals, in addition to which they disappear with the collapse of the archaeocyathan reef system, with which many are associated, at the end of the Lower Cambrian. These reefs and other carbonate environments also yield a range of other coralline organisms, which are doubtfully true corals and in some cases doubtfully cnidarians. *Flindersipora*, for example, has more in common with the enigmatic

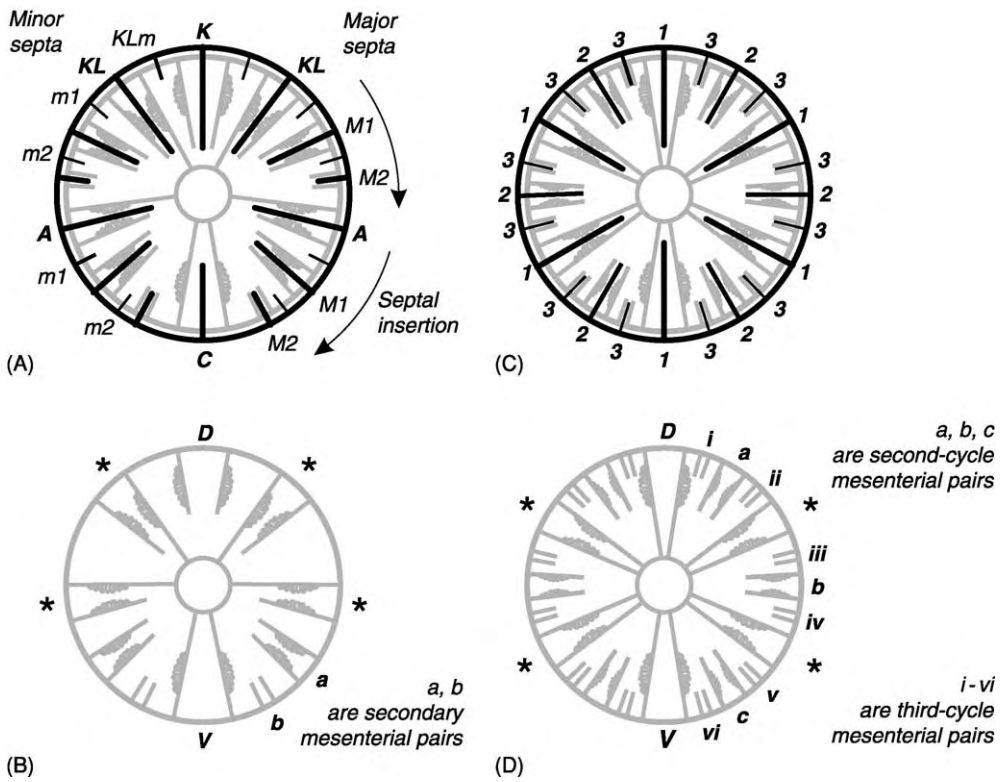


Figure 6 Relationship between septal insertion in (A) the Rugosa and (C) the Scleractinia and mesenterial insertion in (B, D) their presumed ancestral anemone stocks. (C, D) Scleractinian polyps and corallimorpharian anemones appear to be virtually identical except for the skeleton secreted by the former. (A, B) The relationship between the Rugosa and a zoanthiniarian like ancestor is more speculative. Modified from Scrutton (1997, **Figure 7**) The Palaeozoic corals, I: origins and relationships. *Proceedings of the Yorkshire Geological Society*, 51: 177 208 with the permission of the Yorkshire Geological Society.

Ordovician tetradiids, formerly grouped with the Tabulata but now considered of uncertain affinity. The Cothoniida from the Middle Cambrian (*Cothonion*; **Figure 7D**) are clustered cup-shaped corals with opercula, on the inside of which septa are well developed. However, no insertion pattern can be deduced. They are most similar to Rugosa, but with no features suggesting direct descent to the first true rugosans, which did not appear until the Mid-Ordovician. All the Cambrian corals are considered to be short-lived experiments in skeletonization, probably from the same anemone stock that later gave rise to the Rugosa and Tabulata.

The earliest accepted tabulate coral, *Lichenaria*, appeared in the Lower Ordovician. Diversity remained low until the Middle Ordovician, when the first Rugosa appeared and both orders, but particularly the Tabulata, underwent a modest diversification. By the time of the Late Ordovician extinction event, some 60 genera of Tabulata and 40 genera of Rugosa were established. The Kilbuchophyllida (*Kilbuchophyllia*; **Figure 7E**) also appeared in the Middle Ordovician but remained restricted to a small area of the UK and quickly became extinct.

The Tabulata were more seriously affected than the Rugosa by the Late Ordovician extinction event, after which the Rugosa were always generically the more diverse. Both peaked in diversity in the Middle Devonian, and both suffered severely in a series of extinction pulses beginning at the end of the Middle Devonian and continuing until the end of the period. Thereafter, the Tabulata remained a much-reduced presence, whilst the Rugosa recovered almost to peak diversity in the Lower Carboniferous. Both were affected by early Upper Carboniferous extinction before beginning a long, slow decline until final extinction in the Late Permian. The enigmatic Heterocorallia appeared in the Devonian, reached their acme in the Lower Carboniferous, and became extinct in the early Upper Carboniferous (**Figures 7F and 7G**). In addition, the Numidiaphyllida (*Numidiaphyllum*) made a brief and local appearance in the Late Permian.

The Scleractinian corals (**Figures 10 and 11**) first appear in the Middle Triassic. Formerly it was proposed that the scleractinians evolved from the rugosans of the Late Palaeozoic, but this is now generally discounted. Several lines of evidence support this

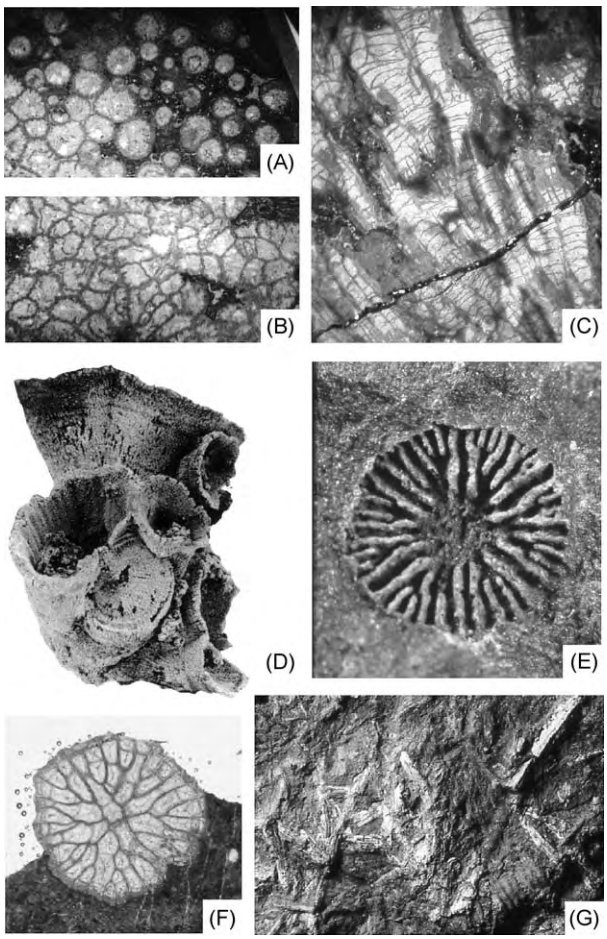


Figure 7 Representative Cambrian and minor group corals. (A C) *Moorowipora chamberensis*; Lower Cambrian, Moorowie For mation, Flinders Ranges, South Australia, ×5: (A, B) cross sections and (C) longitudinal section; photographs courtesy of Margaret Fuller. (D) *Cothonion sympomatum*; early Middle Cam brian, Coonigan Formation, New South Wales, Australia, ×4.5; photograph courtesy of John Jell and reproduced by permission of the Australasian Association of Palaeontologists; **figure 2** in Jell PA and Jell JS (1976) Early Middle Cambrian corals from western New South Wales. Alcheringa 1: 181 195. (E) *Kilbucho phyllia discoidea*, mould of a juvenile specimen showing clear cyclic septal insertion; Ordovician, Caradoc Series, Kirkcolm Formation, Southern Uplands, Scotland, UK, ×3.5. (F) *Heterophyl lia grandis*, cross section showing distinctive heterocoral septal pattern; Carboniferous, Dinantian, Northumberland, England, UK; ×5. (G) *Heterophyllia ornata*, scattered specimens on a bedding plane; note the spines on the edges of the corallite top right; Carboniferous, Dinantian, Northumberland, England, UK; ×1.

conclusion: the lack of Early Triassic corals; a change in skeletal mineralogy from calcite (Rugosa) to aragonite (Scleractinia); contrasting patterns of septal insertion between the two orders; and proof of the antiquity of the scleractinian septal-insertion pattern with the discovery of *Kilbuchophyllia* in the Ordovician. This implies that the ancestral anemone stock survived from the Palaeozoic, and the Early Triassic trace fossil *Dolopichnus* may be proof of that. Furthermore, the earliest scleractinians suggest several already divergent stocks, which became mineralized at about the same time, the origins of which recent DNA work on living descendants suggests can be traced back to the Carboniferous (**Figure 5**). Following severe extinctions at the end of the Triassic, scleractinians steadily diversified through the Mesozoic, with a peak of reef-building activity in the Late Jurassic and less significant losses at the end of that period. The end-Cretaceous extinction event severely affected coral stocks, but from the survivors evolved the highly diverse and ecologically successful scleractinian faunas of the present day. No other group of zoantharian corals is recognized after the end of the Palaeozoic.

Coral Structure and Taxonomy

Coral taxonomy and classification is based principally on the skeletal characteristics of the corallum, the skeletal microstructure, and growth structure and form. Coral skeletons are often highly variable in internal structure and growth form. In addition, similar structures often reoccur, and it can be difficult to separate genetic relationships from convergent homoeomorphy. Given that coral structure is relatively simple to start with, the result is that considerable problems exist with both taxonomy and classification within the major orders.

Many species were erected in the past with little or no regard to intraspecific variation. Only in recent years have more serious attempts been made to take variation into account, with the result that many species and genera are now being placed in synonymy. Rather more stability is apparent in the grouping of genera at the family level, at least in Palaeozoic corals. Here the *Treatise on Invertebrate Paleontology* and its revision have been very influential, although now much modified by more recent work. However, evolutionary relationships within the Rugosa particularly are rather poorly understood. The *Treatise* has also been a major influence in Scleractinian classification, but here major revisions are being made as a result of more recent studies of better-preserved material, particularly from faunas in which the original aragonitic mineralogy has been preserved, allowing septal microstructure to contribute to unravelling relationships. Microstructure has yet to play a very significant role in classifying Palaeozoic corals. Their calcitic skeletons contain up to 8 mol% $MgCO_3$, and there is great controversy over which textures are original and which are secondary as a result of recrystallization.

The Kilbuchophyllida are solitary corals, whilst the Heterocorallia, the Rugosa and the Scleractinia

contain both solitary and modular growth forms. All tabulate corals are modular colonies, as are all of the accepted Cambrian corals and the Numidiaphyllida. Overall, solitary corals show a wide variety in form, but this may be consistently developed within many species. Discoidal (Figures 8B and 10A–C) and similar broad, flat growth forms (tympanoid, cupolate) and those with a noncircular cross-section (calceoloid, flabellate, cuneiform, etc.) tend to be the least variable at the species level. Conicocylindrical solitary corals (Figure 8C) may show more irregular forms as the result of disturbance during growth, at least among the Rugosa where adult solitary corals are generally free-living. Colonial growth form in the skeleton varies from branching to massive, ramose, and foliose. In many branching (fasciculate) forms, the polyps separate completely, and the clonal group is held together only by the corallites they secrete beneath themselves (Figure 8H and 8I). The cluster of corallites may be loose and divergent (dendroid; Figure 10E) or close and parallel (phaceloid), and the arrangement appears to be under genetic constraint in many cases. Corallites may remain attached on opposite sides, like the palisades in a fence, the ranks joined together to enclose lacunae in the skeleton (cateniform; Figure 9G and 9H). An external wall may enclose the ranks but along them adjacent corallites may be in contact or separated by colonial tissue, suggesting neural connectivity between the polyps. A range of structural relationships between the modules exist in massive, ramose, and foliose growth forms. Corallites may be packed more or less tightly together, becoming polygonal in cross-section, each retaining their own wall (rugosan cerioid, suggesting complete separation of the polyps; Figure 8J–M). However, various levels of integration of the polyps may be reflected in the loss of individual corallite walls (for example scleractinian cerioid, Figure 10D)

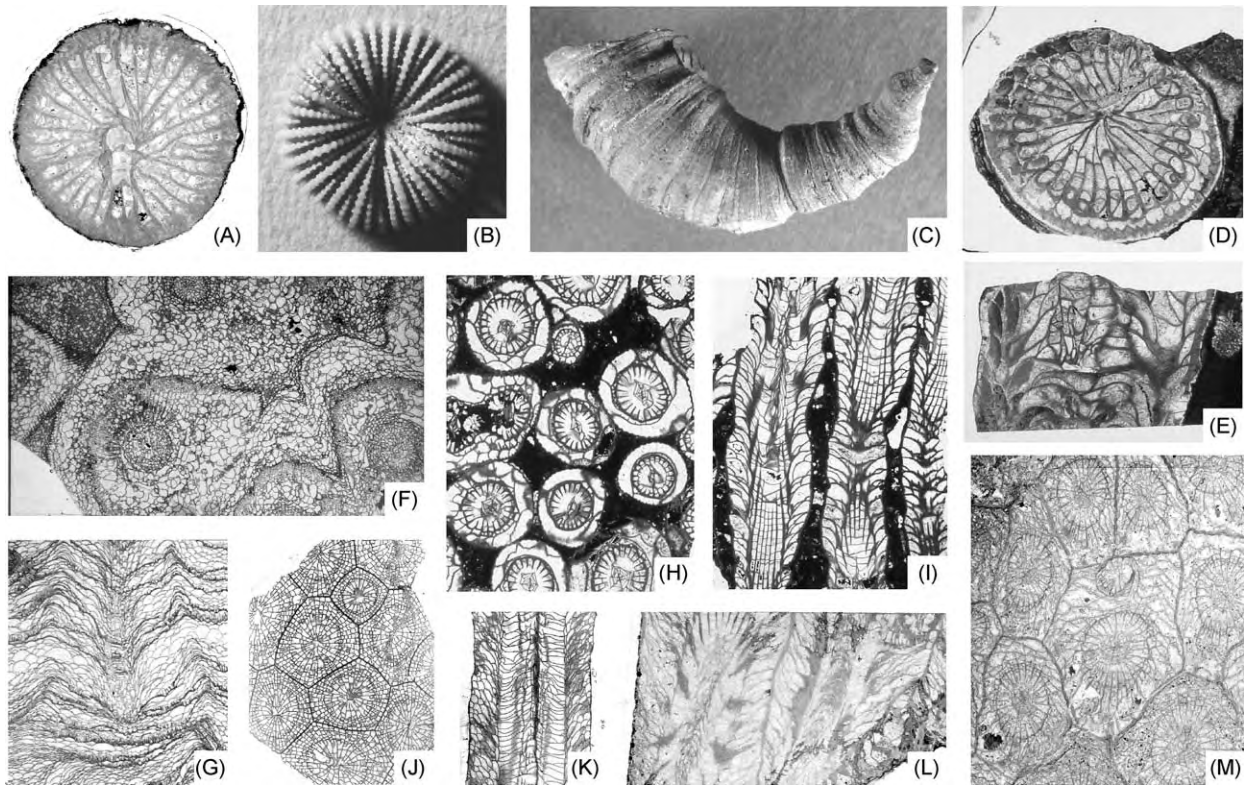


Figure 8 Representative Rugose corals. (A) *Amplexizaphrentis* sp. showing a pinnate arrangement of septa reflecting the pattern of insertion; Lower Carboniferous, Dinantian, Northern England, UK; ×1.5. (B) *Palaeocyclus porpita*, discoidal corallum showing a pinnate arrangement of septa particularly about the very short cardinal septum, bottom centre; lower Silurian, Llandovery Series, Gotland, Sweden ×2.5. (C) Typical solitary horn coral, shown in growth orientation; Carboniferous, locality unknown; ×2. (D, E) *Tabulophyllum* sp., showing well developed septa, dissepiments, and tabulae; Upper Devonian, Frasnian, East Ogwell Limestone, south Devon, UK; ×1.5: (D) cross section and (E) longitudinal section. (F, G) *Arachnophyllum murchisoni*, an amural colony with (G) pseudo walls formed of stacks of dissepiments; Silurian, Much Wenlock Limestone, Shropshire, UK; ×1: (F) cross section and (G) longitudinal section. (H, I) *Lonsdaleia duplicata*, phaceloid fasciculate colony; Carboniferous, Dinantian, Cumbria, UK; ×1: (H) cross section and (I) longitudinal section. (J, K) *Lithostrotion araneum*, massive cerioid colony; Carboniferous, Dinantian, northern England, UK; ×1.3: (J) cross section and (K) longitudinal section. (L, M) *Actinocyathus floriformis*, massive cerioid colony showing corallites with well developed axial structures; Carboniferous, Namurian, Great Limestone, Co. Durham, UK; ×1.3: (L) longitudinal section and (M) cross section.

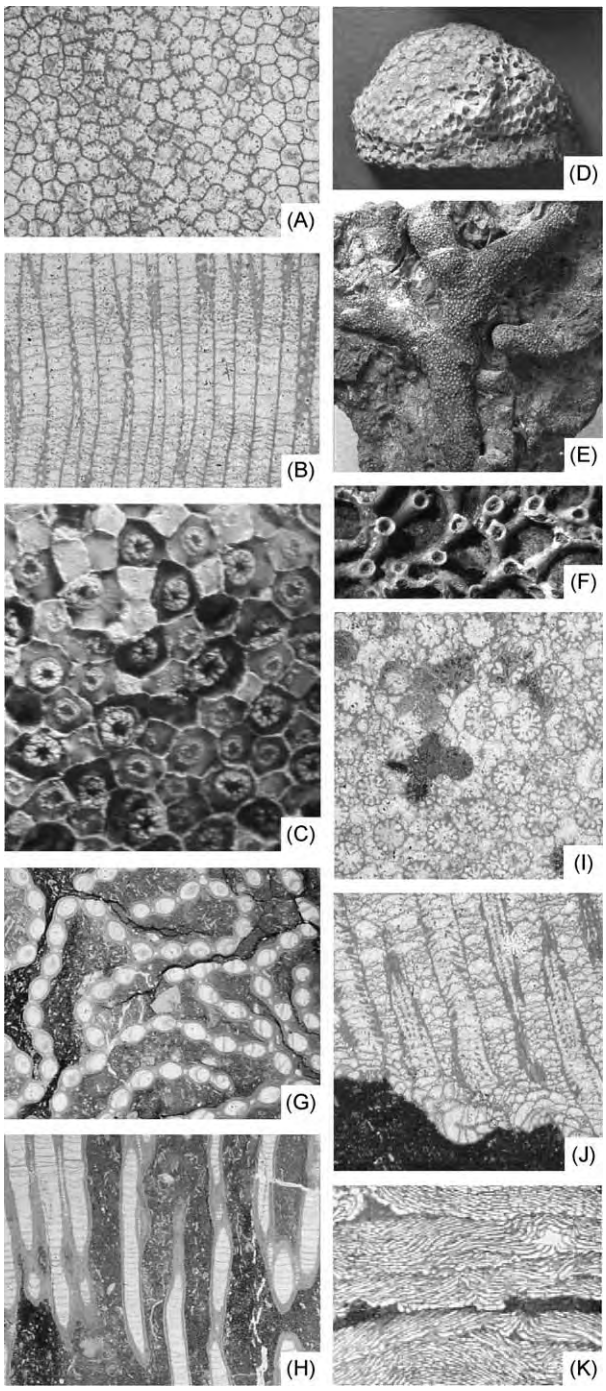


Figure 9 Representative Tabulate corals. (A, B) *Paleofavosites asper*, massive cerioid colony with well developed pores connecting adjacent corallites; septa are well developed in denser bands, with more crowded tabulae (reflecting an annual growth cycle), and weak to absent elsewhere; Silurian, Much Wenlock Limestone, Shropshire, UK; ×3: (A) Cross section and (B) longitudinal section. (C) *Favosites* sp., with polyps preserved in many calices; Silurian, Llandovery Series, Jupiter Formation, Anticosti Island, Canada; ×4. Photograph courtesy of Paul Copper. (D) *Paleofavosites rugosus*, a small domal colony with a flat base: Silurian, Much Wenlock Limestone, Shropshire, UK; ×0.75. (E) *Thamnopora* sp., ramose growth form, the Palaeozoic coral that is structurally closest to the highly successful *Acropora* of living reefs (see Figure 11); Devonian, Portilla Formation, Bernesga Valley, northern Spain; ×0.5. (F) *Aulopora* sp., encrusting colonial meshwork of horn shaped coralla; Silurian, Klinteberg Formation, Gotland, Sweden; ×2. (G, H) *Halysites catenularius*, cateniform colony with corallites separated by thin tubules of colonial tissue along the ranks; Silurian, Much Wenlock Limestone, Shropshire, UK; ×2: (G) cross section and (H) longitudinal section. (I, J) *Propora* sp., coenenchymal colony with vesicular colonial tissue, from which corallites arise at the base of the colony; Silurian, Llandovery Series, Petalocrinus Limestone, Woolhope, Herefordshire, UK; ×3: (I) cross section and (J) longitudinal section. (K) *Subalveolites panderi*, longitudinal section of a colony formed by encrusting layers of oblique corallites; Silurian, Llandovery Series, Hughley Shales, Devil's Dingle, Shropshire, UK; ×2.5.

and, ultimately, by calical centres set in continuous colonial tissue (amural (Figure 8F and 8G), coenenchymal (Figure 9I and 9J) and plocoid (Figures 10G, 11A and 11B in ramose growth forms)). In the Rugosa and Tabulata, individuals are still sufficiently well defined to suggest that polyp integration was restricted to neural connectivity in most cases and only rarely was there confluence of body cavities. However, some massive and branching tabulates have tubes or pores connecting adjacent corallites (Figure 9A and 9B), suggesting interconnected gastric cavities. In some Scleractinia, separation of polyps may be more or less incomplete, resulting in small clusters (polycentric phaceloid; Figure 11B) or linear series (flabellate, meandroid; Figure 10F) of mouths opening into a single gut and surrounded by a single series of tentacles.

Corallite structure in colonial corals is either relatively invariant or restricted to a well-defined and narrow range at species level and often also at generic level. Growth form is more variable. In the past, it has been considered to reflect environmental influences, but increasingly the importance of underlying genetic controls has become apparent, at least for some corals.

Coral Ecology and Palaeoecology

Scleractinia

Living scleractinian corals range from the tropics to high latitudes (<75°) and from the surface to abyssal depths (Figure 12). However, two ecological groups can be distinguished, one with symbiotic algae (zooxanthellae), the other without (Table 1). Zooxanthellate scleractinians dominate living coral reefs and are restricted in their distribution by their algal symbionts to water temperatures of 16–40°C and shallow water depths, generally less than 190 m. Azooxanthellate scleractinians can exist through a much broader range of temperature and depth. They are minor components of shallow-water tropical reefs but build structures of their own (coral thickets and banks) in deeper and cooler waters.

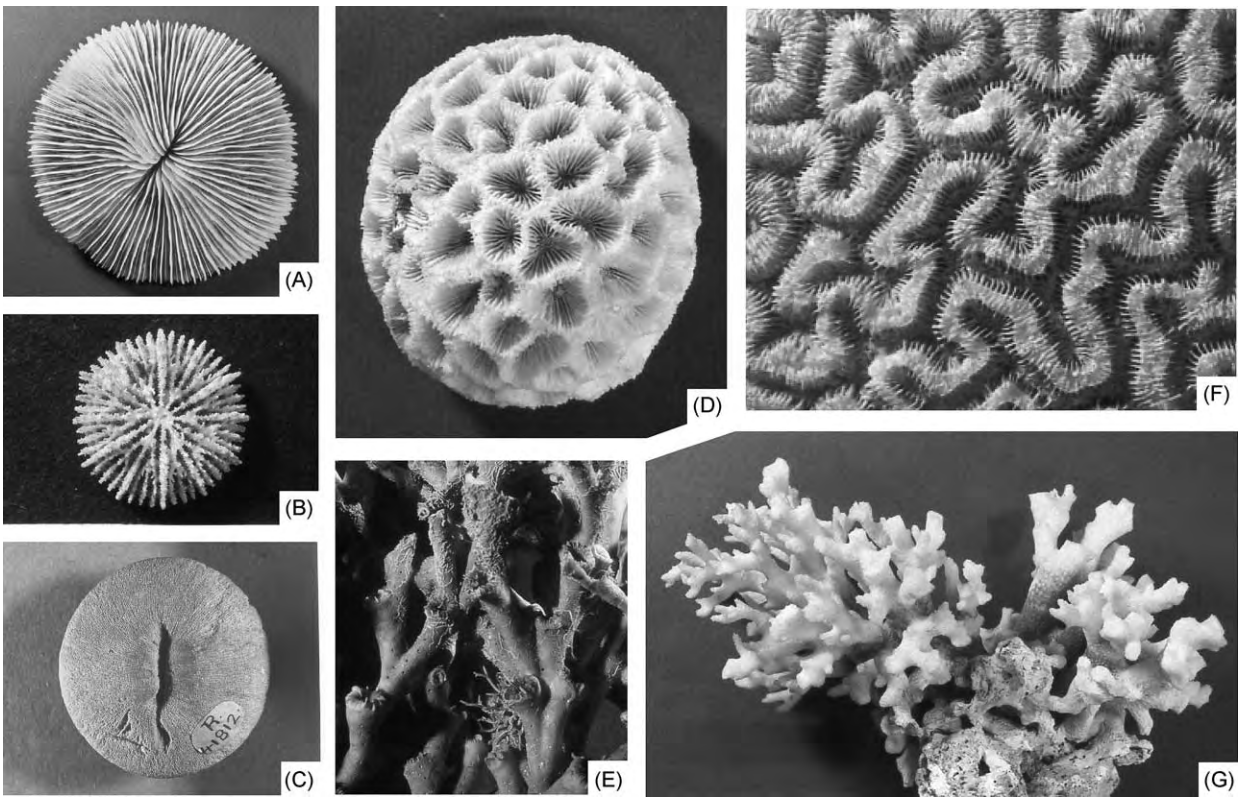


Figure 10 Representative Scleractinian corals. (A) *Fungia fungites*, calical view of discoidal zooxanthellate corallum; Recent, Great Barrier Reef, Queensland, Australia; ×0.25. (B) *Fungiacyathus symmetricus*, discoidal azooxanthellate corallum; Recent, locality unknown; ×3.5. (C) *Cyclolites* sp., calical view of zooxanthellate cupolate corallum with large numbers of closely spaced perforate septa; Cretaceous, Spain; ×0.8. (D) *Favites* sp., cerioid (scleractinian usage) colony in which intercorallite walls are formed by the interlocking of the peripheral ends of the septa (a septotheca); Recent, Great Barrier Reef, Queensland, Australia; ×0.7. (E) *Lophelia prolifera*, azooxanthellate dendroid colony with well spaced corallites; Recent, coral bank on continental shelf, North West Atlantic; ×2.5. (F) *Diploria* sp., typical zooxanthellate meandroid colony (brain coral); Recent, West Indies; ×1. (G) *Stylophora pistillata*, ramose zooxanthellate colony in which each branch is made up of many plocoid corallites; Recent, Great Barrier Reef, Queensland, Australia; ×0.5.

The role played by zooxanthellae is not entirely clear, but they make an important contribution to the energy budget of the coral, contributing up to 98% of its fixed carbon requirement. They appear to accelerate calcification, which in the extreme case of the Stag's horn coral *Acropora* can achieve a linear growth rate of 270 mm yr^{-1}. However, many massive Scleractinia appear not to have enhanced growth rates, and recent evidence suggests that, in this case, zooxanthellae may act to switch calcification on and off, rather than accelerate it, thus conserving energy. The algae also remove waste from the coral and may thereby allow very large colonies (in excess of 10 m) of tiny individuals (approximately 1 mm in diameter) to develop. There is a striking contrast in coloniality between zooxanthellate and azooxanthellate corals. The former are 95% modular by genus, whereas the latter are about 30%. The symbiosis with algae appears to have developed in the early Jurassic.

A major factor in the success of scleractinian corals in building rigid structures is their ability to cement their skeletons securely to other skeletal material and to encrust hard substrates (Figure 11). Shallow-water scleractinian-dominated reefs became important from the Jurassic onwards, although deep-water banks built by azooxanthellate scleractinians extend back into the Triassic.

Palaeozoic Corals

Rugose and tabulate corals enjoyed neither the advantages of algal symbiosis nor the ability to encrust securely to the extent seen in the scleractinians. The first point has been contentious. However, scleractinian zooxanthellate colonial corals show flattening of the skeleton and a significant reduction in growth rate with depth, as light levels decrease. A recent study of a widely distributed tabulate coral, among those considered most likely to have harboured algal symbionts, showed no signs of these effects. In addition, only 35% of rugosans (but all tabulates) are colonial, and massive growth forms of any Palaeozoic coral appear not to exceed about 2 m in diameter. Apart from some

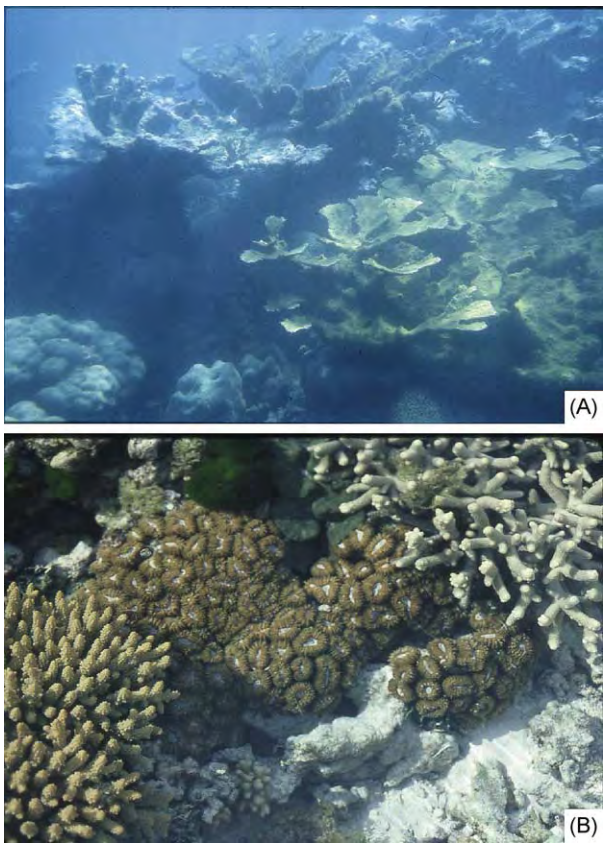


Figure 11 Reef forming scleractinian corals. (A) Reef edge dominated by *Acropora palmata*; open ocean to the right; East Rio Bueno, Jamaica. (B) Reef flat with phaceloid, partly polycentric colony of *Lobophyllia hemprichii* in the centre, flanked by ramose, plocoid colonies of *Acropora* spp; Heron Island, Great Barrier Reef, Queensland, Australia.

specialized forms, rugose and tabulate corals showing secure encrusting attachment or overgrowth as mature coralla are rare. This reflects the presence of an external wall (epitheca or holotheca) around the corallum and the lack of an extensive edge zone of soft tissue, the reverse of the situation in most scleractinians. Instead, the Palaeozoic corals were predominantly adapted to life on soft substrates. They were subsidiary contributors to reef cores, although common in peri-reefal communities, and reached their greatest importance in bioherms and biostromes, with little evidence of frame building and were often associated with mud-grade carbonate environments. Many show evidence of instability and smothering as soft substrates were disturbed by storm activity and sediment movement. They reached their maximum diversity in warm shallow waters, with diversity decreasing with depth and towards the contemporary poles. In addition, there was a distinctive low-diversity community of small, thick-walled, tabulate and solitary rugose corals characteristic of dysaerobic, dysphotic and other marginal environments that persisted from the Silurian to the

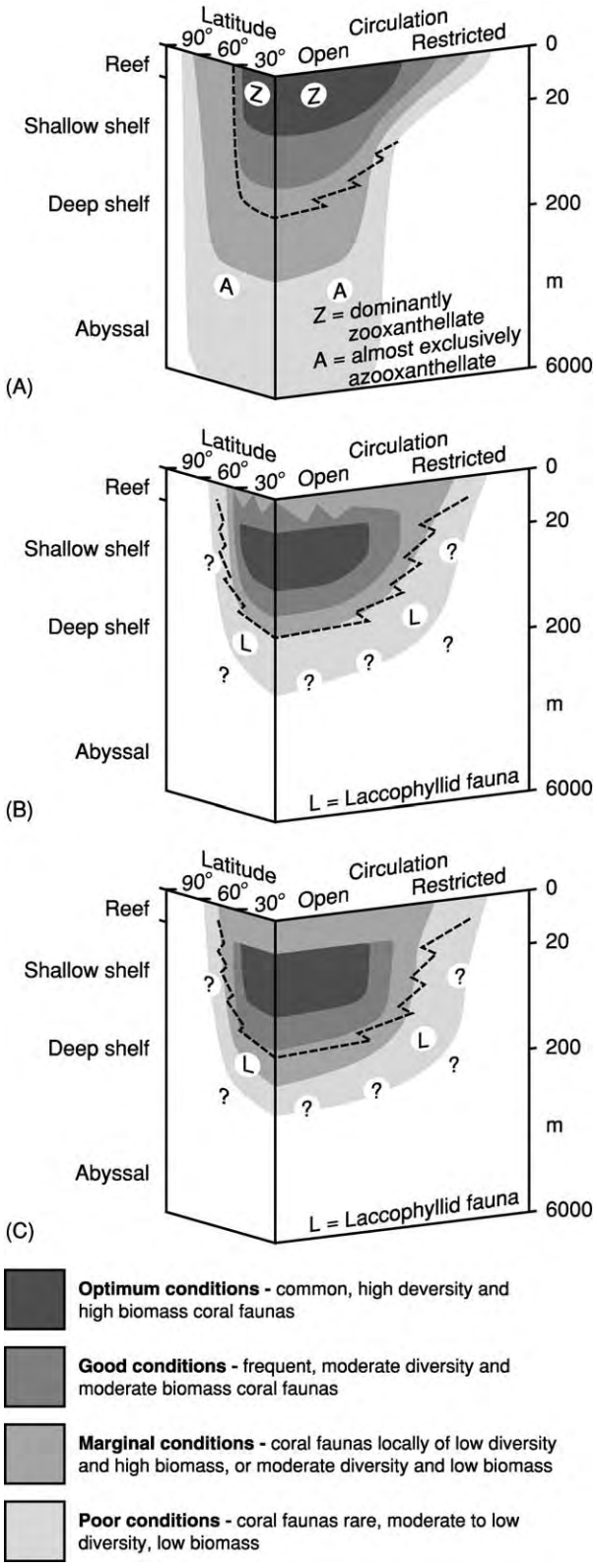


Figure 12 Generalized ecological and palaeoecological ranges of (A) the Scleractinia, (B) the Tabulata, and (C) the Rugosa. Depth not to scale, and divisions diagrammatic. For further discussion see text. Modified from Scrutton (1998, Figure 30) The Palaeozoic corals, II: structure, variation and palaeo ecology. *Proceedings of the Yorkshire Geological Society*, 52: 1 57 with the permission of the Yorkshire Geological Society.

Table 1 Comparative features of the two ecological groups of the Scleractinia, the Rugosa, and the Tabulata

	Scleractinia			
	Zooxanthellate	*Azooxanthellate*	*Rugosa*	*Tabulata*
Algal symbiosis	Present	Lacking	Lacking	Lacking
Epitheca/ holotheca	Mainly absent	Present in some	Almost universal	Present in most
Solitary forms	<1 m+ in diameter	Small	*c.* 4 140 mm in diameter	None
Colonial forms[a]	95% colonial; colonies may be large up to 10 m+ in diameter; corallites often very small, *c.* 1 50 mm in diameter	30% colonial; mainly dendroid; corallites well spaced, medium sized	36% colonial; colonies mainly small 30 400 mm, some up to 1 m or rarely 2 m in diameter; fasciculate rarely larger	All colonial; colonies mainly small 30 600 mm, but up to 2 m; corallites 1 20 mm in diameter
Growth rates	*Acropora* 100 150 mm yr ^{1}, up to 270 mm yr ^{1}; massive forms 5 15 mm yr ^{1}	5 15 mm yr ^{1}, up to 40 mm yr ^{1} in some shallow branching forms	3 10 mm yr ^{1} (massive), up to 27 mm yr ^{1} (solitary),? rarely up to 70 mm yr ^{1}	2 9 mm yr ^{1} in most but up to 20 mm yr ^{1}
Skeleton	Often light, porous	Usually solid, non porous	Usually solid, rarely porous	Usually solid, rarely porous
Depth	Most 0 190 m, abundant at less than 70 m, optimum 1 25 m	0 6200 m, most less than 800 m	Usually less than 60 m, recorded down to 120 m, probably extending much deeper	Common down to about 60 m, probably extending much deeper
Temperature	16 40°C, 25 29°C optimum	1°C 29°C	Most tropical to temperate	Most tropical to temperate
Ecology	Mostly hermatypic[b]; major component of shallow water reefs	Mostly ahermatypic[b] but fasciculate colonies form banks at depths of 100 1500 m	Common particularly in shelf bioherms and biostromes; minor component of reefs	Common particularly in shelf bioherms and biostromes; locally significant in reefs

[a]Percentage coloniality is based on numbers of genera.
[b]Hermatypic forms are those associated with shallow water reefs.
Modified from Scrutton (1998, Tables 3 and 4) The Palaeozoic corals, II: structure, variation and palaeoecology. *Proceedings of the Yorkshire Geological Society*, 52:1 57 with the permission of the Yorkshire Geological Society.

Late Permian (the Laccophyllid fauna; Figures 12B and 12C). It was characteristically cosmopolitan compared with the provincialism of other coral faunas.

Glossary

Amural Colonial rugose corals in which the walls separating individual corallites are lost

Ahermatypic Non-shallow-water reef-building scleractinian corals

Azooxanthellate Scleractinian corals lacking symbiotic algae (zooxanthellae)

Calceoloid Growth form of solitary rugose corals in which the corallite is semicircular in section and rests on its flattened face. Sometimes called 'slipper-shaped'

Calice Upper surface of the corallite on which the polyp rests. Calices may vary from convex to concave but most are shallow to deep and funnel or cup shaped

Cateniform Colonial corals in which corallites are in contact along their opposite sides like posts in a fence. This arrangement is called a rank, and ranks join at various points in the colony to enclose more or less irregularly shaped lacunae. Ranks in cross-section look like chains, hence 'chain corals'

Cerioid Structural arrangement of corallites in colonial corals such that each corallite is polygonal in section and in close contact with its neighbours. Usage differs in the Rugosa and Scleractinia in that in the former each corallite retains its external wall (epitheca), whereas in the latter the corallites are separated by the merging of peripheral septal ends from adjacent corallites to form a septotheca

Coenenchymal Structural arrangement of corallites in colonial corals such that corallite centres are set in common colonial tissue (coenenchyme). Coenenchyme may have various forms, but is usually tubular or vesicular

Corallite, corallum The skeleton secreted beneath a single polyp. Corallum is also used for the skeleton of a colonial coral consisting of many corallites

Cuneiform Growth form of solitary scleractinian corals in which the corallite is wedge shaped

Cupolate Growth form of solitary scleractinian corals in which the corallite has a flat base and a highly domed upper (calicular) surface

Dendroid Structural arrangement of corallites in a branching (fasciculate) coral in which each branch is a single corallite and the branches are irregularly divergent

Diploblastic Level of organization in primitive invertebrates in which the body wall consists of external and internal cellular layers (ectoderm and endoderm), which may be separated by a layer of noncellular connective tissue

Discoidal Growth form of solitary corals in which the corallite has a flat base and a flat to gently concave upper (calicular) surface

Dissepiment A vesicular skeletal plate, subhorizontal to inclined zones of which occupy the peripheral areas of many corallites (dissepimentarium). In scleractinian corals, vesicular and tabular dissepiments are distinguished

Dysaerobic Environment with 0.1–1 ml O_2 per litre of water

Dysphotic Environment with low light levels

Enteron Cavity in the body of a cnidarian polyp or medusa; the gut

Epitheca The external wall of a corallite

Flabellate Growth form of scleractinian corals, either a fan-shaped solitary coral with an elliptical cross-section or a meandroid corallum consisting of a single, linear, polycentric corallite

Foliose Growth form of colonial corals in which the branches are broad and more or less flat laminae

Hermatypic Scleractinian corals that build shallow-water reefs

Holotheca The external wall of a colonial coral in which the individual corallites have lost their own walls (epithecae)

Homoeomorphy Similarity of form resulting from convergence rather than genetic relationship

Increase The process of new corallite formation in a colonial coral

Lagerstätte Fossil deposit in which organisms not normally preserved, such as those lacking mineralized skeletons, are found

Meandroid Structural arrangement of corallites in scleractinian colonial corals in which a number of linear polycentric corallites are intertwined (a brain coral)

Medusa The free-floating stage of the cnidarian body plan in which the mouth faces downwards surrounded by tentacles

Mesentery A radially disposed fleshy infold of the endoderm of a cnidarian polyp, attached to the inner surfaces of the oral and basal discs. The inner end of the lamina may be attached to the stomodaeum, in which case the mesentery is said to be complete

Nematocysts Stinging or adhesive structures housed in specialized cells of the ectoderm, characteristic of cnidarians

Operculum Skeletal plate or plates covering the calice of a corallite in some corals

Periderm Outer stiff organic (scleroprotein) skeleton of some polypoid hydrozoans

Phaceloid Structural arrangement of corallites in a branching (fasciculate) coral in which each branch is a single corallite and the branches are relatively closely spaced and subparallel

Planula Larval stage of a cnidarian

Plocoid Structural arrangement of corallites in colonial scleractinian corals such that corallites are separated by septal costae, dissepiments, or other common colonial tissue

Polyp The fixed benthonic stage of the cnidarian body plan in which the mouth faces upwards surrounded by tentacles. In the corals and some other cnidarians the polyp secretes a calcium carbonate skeleton at its base

Ramose Branching growth form in which each branch consists of many corallites. The structural arrangement of the corallites may be cerioid, amural, coenenchymal, plocoid, or meandroid

Scyphistoma Benthonic polypoid stage in scyphozoans specialized for budding off medusae

Septum In corals, a radial skeletal element, which may be a solid or perforate vertical plate or be reduced to a series of spines. In Scyphozoa, one of four radial fleshy partitions of the gut

Stomodaeum Cylindrical collar extending down from the mouth into the gut of anthozoan polyps

Tabula A horizontally disposed skeletal plate, which may be flat, dished, arched, domed, or vesicular, stacks of which occupy the axial area of most rugosan corallites (tabularium) or the whole width of the corallite in some rugosans and most other Palaeozoic corals

Zooxanthellate Scleractinian corals with symbiotic algae (zooxanthellae) present in the endoderm

See Also

Fossil Invertebrates: Porifera. **Lagerstätten**. **Palaeoclimates**. **Palaeoecology**. **Precambrian:** Vendian and Ediacaran. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups').

Further Reading

Babcock LE and Feldmann RM (1986) The phylum Con ulariida. In: Hoffman A and Nitecki MH (eds.) *Problem atic Fossil Taxa,* pp. 135 147. New York: Oxford University Press.

Coates AG and Oliver WA Jr (1973) Coloniality in zoantharian corals. In: Boardman RS, Cheetham AH, and Oliver WA Jr (eds.) *Animal Colonies, Development and Function Through Time,* pp. 3 27. Stroudsburg: Dowden, Hutchinson and Ross.

Hill D (1981) Rugosa and Tabulata. In: Teichert C (ed.) *Treatise on Invertebrate Paleontology, Part F. Coelenterata (Supplement 1).* pp. xl + 762. Boulder, Colorado and Lawrence, Kansas: Geological Society of America and University of Kansas Press.

Jenkins RJF (1992) Functional and ecological aspects of Ediacaran assemblages. In: Lipps JH and Signor PW (eds.) *Origin and Early Evolution of the Metazoa,* pp. 131 176. New York: Plenum Press.

Moore RC (ed.) (1956) *Treatise on Invertebrate Paleontology, Part F. Coelenterata.* Boulder, Colorado and Laurence, Kansas: Geological Society of America and University of Kansas Press.

Nudds JR and Sepkoski JJ Jr (1993) Coelenterata. In: Benton MJ (ed.) *The Fossil Record, 2,* pp. 101 124. London: Chapman & Hall.

Oliver WA Jr and Coates AG (1987) Phylum Cnidaria. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates,* pp. 140 193. Oxford: Blackwell.

Scrutton CT (1979) Early fossil cnidarians. In: House MR (ed.) *The Origin of Major Invertebrate Groups,* pp. 161 207. London: Academic Press.

Scrutton CT (1997) The Palaeozoic corals, I: origins and relationships. *Proceedings of the Yorkshire Geological Society* 51: 177 208.

Scrutton CT (1998) The Palaeozoic corals, II: structure, variation and palaeoecology. *Proceedings of the Yorkshire Geological Society* 52: 1 57.

Scrutton CT (1999) Palaeozoic corals: their evolution and palaeoecology. *Geology Today* 15: 184 193.

Scrutton CT and Rosen BR (1985) Cnidaria. In: Murray JW (ed.) *Atlas of Invertebrate Macrofossils,* pp. 11 46. Harlow: Longman.

Seilacher A (1989) Vendozoa: organismic constructions in the Proterozoic biosphere. *Lethaia* 22: 229 239.

Stanley GD Jr (1986) Chondrophorine hydrozoans as problematic fossils. In: Hoffman A and Nitecki MH (eds.) *Problematic Fossil Taxa,* pp. 68 86. New York: Oxford University Press.

Stanley GD Jr (ed.) (1996) *Paleobiology and Biology of Corals.* Lawrence, Kansas: The Paleontological Society.

van Iten H (1992) Microstructure and growth of the conulariid test: implications for conulariid affinities. *Palaeontology* 35: 359 372.

Veron JEN (1995) *Corals in Space and Time: Biogeography and Evolution of the Scleractinia.* Ithaca: Cornell University Press.

Echinoderms (Other Than Echinoids)

A B Smith, The Natural History Museum, London, UK

Introduction

Echinoderms are an important group of marine invertebrates with multiplated calcite skeletons and a long and rich fossil record. Echinoderms first appeared in the Early Cambrian and today are represented by approximately 6000 species distributed throughout all latitudes and from the intertidal zone to the deep ocean trenches. They have one limitation, however: they can tolerate only slightly reduced salinities and are absent from brackish and estuarine habitats. Sessile forms such as crinoids and blastoids are filtration feeders, using their network of arms or tentacles to capture organic particles from the water. Vagile forms are more diverse in their feeding strategies, with scavengers, algivores, carnivores, deposit feeders, and filter feeders all represented. Almost all are benthonic, living on the seafloor or buried just beneath it, although there are rare pseudoplanktonic and true nektonic species.

Key Attributes of Echinoderms

Crown-group echinoderms share several characteristics that mark them apart from other invertebrates and show them to be a monophyletic group.

- They possess multiplated skeletons whose elements are composed of high-magnesium calcite and under magnification have a highly distinctive mesh-like structure termed stereom (**Figure 1**). Because of this, even isolated plates in petrographic thin section are instantly recognizable as being derived from echinoderms. Stereom can have a variety of three-dimensional arrangements, largely depending on the nature of the investing soft tissue.
- Their body plan is basically pentaradiate. This is clearly expressed in all living groups, although there are exceptions, especially amongst some of the more primitive Early Palaeozoic blastozoan groups.
- All crown-group echinoderms possess a water vascular system built to a common plan. The water vascular system is a hydraulically operated system of tube feet and linking canals that is derived from the larval left mesocoel. It is composed of a central

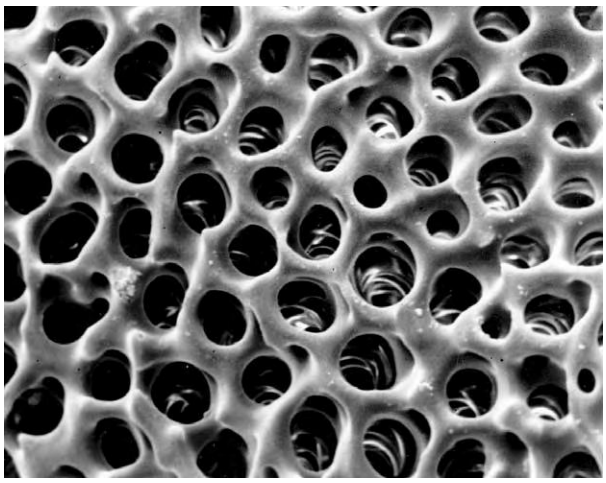


Figure 1 Scanning electron micrograph (magnification ×250) of an echinoid plate showing the three dimensional mesh (stereom) of which the echinoderm skeleton is composed.

ring canal, which surrounds the start of the digestive system just internal to the mouth. Five primary branches bud off from this ring forming long radial water vessels, which end blindly and which give rise on either side to a series of external tentacles termed tube feet. An additional short branch arising from the circumoral ring leads to the exterior via the hydropore and bears no tentacles. The skeletal elements associated with the radial canals and their tube feet comprise the ambulacral series. Ambulacral-plate series are thus always arranged radially around the mouth.

Of these characteristics, only possession of stereom is ubiquitous to all stem-group echinoderms. Pentameral symmetry is confined to crown-group echinoderms, stem-group members either being completely asymmetrical or showing bilateral or triradiate symmetry. Similarly, only some stem-group echinoderms show evidence of an ambulacral system and ambulacral plating.

Phylogenetic Relationships

Echinoderms are triploblastic Metazoa, which, despite their unique pentaradiate adult body plan, belong to the Bilateria. More specifically, echinoderms are members of the group Deuterostoma, i.e. bilaterians in which the embryonic blastopore develops into the anus rather than the mouth. This aligns echinoderms with chordates and hemichordates, the latter comprising acorn worms, pterobranchs, and the extinct graptolites (*see* **Fossil Invertebrates:** Graptolites). Recent molecular and morphological analyses unambiguously indicate that echinoderms and hemichordates are the more closely related, and together these form the taxon Ambulararia. Both hemichordates and echinoderms have a vascular system of tubes (the water vascular system of echinoderms and the tentacular system of pterobranchs) that originates from the homologous primary body coeloms. In hemichordates the tentacular system is symmetrically paired and forms from both the left and the right mesocoel. In echinoderms only the left mesocoel forms the water vascular system, and hence there is only a single hydropore. As both hemichordates and chordates are bilaterally symmetrical as adults and bear gill slits (pharyngeal openings), the latest common ancestor of echinoderms and hemichordates must also have been bilaterally symmetrical and have possessed gill slits.

Echinodermata are divided into five extant classes, which together form the crown group of the phylum. These are the Crinoidea (feather stars and sea lilies) (*see* **Fossil Invertebrates:** Crinoids), Asteroidea (starfishes), Ophiuroidea (brittle stars), Echinoidea (sea urchins) (*see* **Fossil Invertebrates:** Echinoids) and Holothuroidea (sea cucumbers). Major differences separate the body plan of crinoids from those characterizing the other four classes, and molecular and comparative data all point to crinoids as the most primitive of the extant echinoderm classes. The crinoids and their extinct stalked relatives the blastozoans together form the group Pelmatozoa. Asteroidea, Ophiuroidea, and the Echinozoa form a sister clade, the Eleutherozoa, and are free-living. Molecular evidence suggests that amongst eleutherozoans the Echinoidea and Holothuroidea are the most closely related, and together they form the group Echinozoa. Echinozoa have a globular to cylindrical body form with the aboral surface highly reduced. By contrast, asteroids and ophiuroids have a stellate body plan and their oral and aboral surfaces are equally developed. The extinct edrioasteroids are an important early eleutherozoan group.

There are some numerically small but interesting stem-group echinoderms; most important amongst these are the carpoids (see below). Carpoids all lack pentaradiate symmetry, and most lack an ambulacral system while retaining pharyngeal gill slits.

Geological History

The geological history of echinoderms is summarized in Figure 2. The earliest records we have are of isolated plates with the distinctive stereom structure from the Botomian (Early Lower Cambrian) of America. By the end of the Lower Cambrian we have evidence of carpoids, helicoplacoids, blastozoans, and edrioasteroids, suggesting that the crown-group split between pelmatozoans and eleutherozoans had

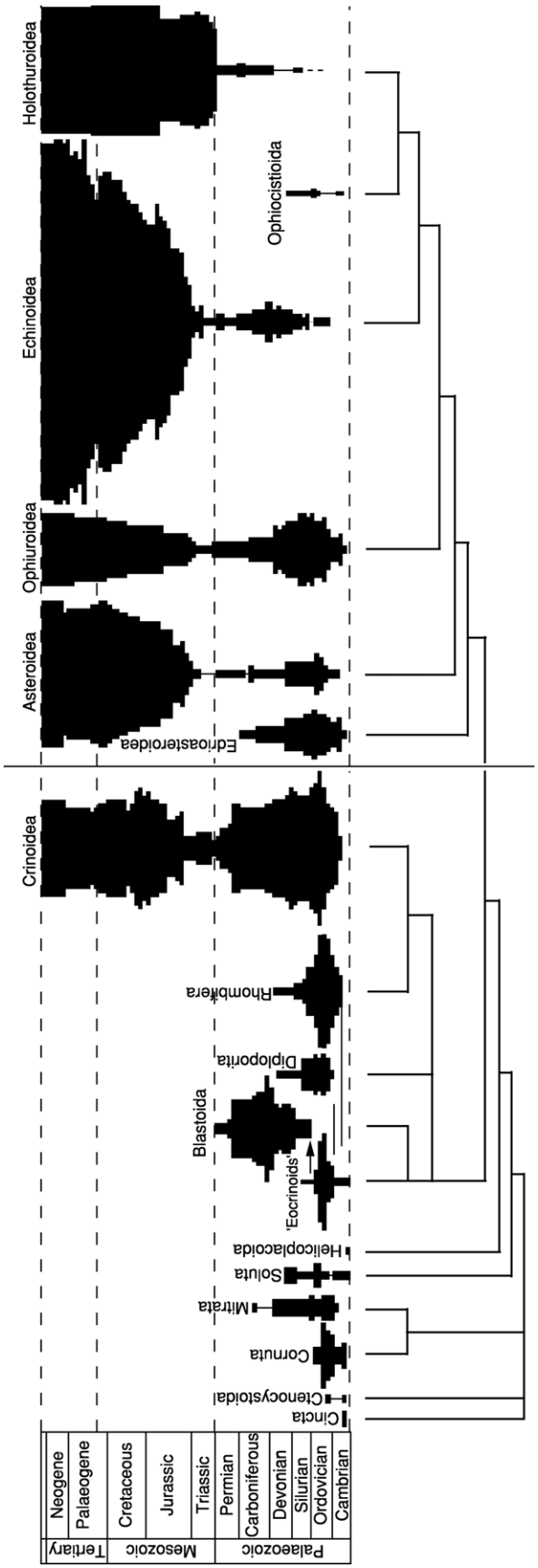


Figure 2 Stratigraphic ranges of the major groups of echinoderm. The widths of the bars are proportional to the taxic (family level) diversities. The cladogram shows their inferred phylogenetic relationships.

already taken place. However, it is not until the Early Ordovician that definite members of the five extant classes are found. The earliest crinoids, asteroids, and ophiuroids all appear close to the base of the Ordovician, while echinoids appear a little later, in the Middle Ordovician. Holothurians are not definitely known until the Late Silurian. During the Palaeozoic stemmed filtration feeders dominated benthic communities both numerically and taxonomically. Although isolated echinoderm ossicles can be present in such abundance as to be the major bioclastic constituent of some limestones, articulated echinoderm skeletons are for the most part confined to Lagerstätte deposits (*see* **Lagerstätten**) and are relatively rare.

All echinoderm groups went into decline towards the end of the Palaeozoic, and several groups became extinct. Edrioasteroids last occur in the Upper Carboniferous and blastozoans in the Permian. The groups that did survive into the Mesozoic did so only in small numbers, and indeed the modern fauna of echinoids, asteroids, crinoids, and ophiuroids were all initiated at around this time. The evolution in echinoids of a solid and robust test in the Triassic greatly enhanced their preservation potential. The history of echinoderms since the Triassic has been one of unbridled success, and the five classes have continued to diversify and expand their ecological ranges through the Mesozoic and Tertiary. Echinoderms are probably as diverse now as they have ever been.

Major Taxonomic Groups

Carpoids

Carpoids are the most primitive of the stem-group echinoderms. There are four major groups: Soluta, Cincta, Ctenocystoida, and Stylophora; Stylophora is further subdivided into the cornutes and the mitrates.The Stylophora (Figures 3C and 3D) are the most primitive and conform most closely in body plan to a basal deuterostome. Stylophora have a large anterior body with pharyngeal openings and a single posterior appendage (stele), which is muscular and bilaterally symmetrical. Cornutes (Figure 3D) are strongly asymmetric in outline, often rather boot-shaped, and have a well-developed marginal frame and a serially repeated set of external atrial openings. Mitrates (Figure 3C) on the other hand are more nearly bilaterally symmetrical with no differentiated marginal frame. Paired internal gill bars are present in at least some species. Both have a calcitic skeleton composed of stereom, and neither show evidence of possessing a water vascular system. The presence of gill openings and a motile muscular stele or tail is primitive for deuterostomes as a whole. Stylophora occur from the Middle Cambrian through to the Carboniferous and were recumbent suspension feeders that are thought to resemble tunicates in their feeding strategy (i.e. drawing water into the pharynx where it could be filtered through gill filaments).

Solutes are the most echinoderm-like of the carpoid groups. They have a sac-like body and two appendages asymmetrically arranged at opposite poles (Figure 3A). One of these appendages is an ambulacrum in the form of a single arm, so clearly this group had a tentacular water vascular system. Furthermore, the presence of a single rather than paired hydropore indicates that it was built on the echinoderm plan. The other appendage is a stalk, which is muscular near the theca and more rigid distally with a basal attachment pad. Solutes appear in the Lower Cambrian and survived until the Early Devonian. They were suspension feeders, using their ambulacrum to capture particles.

Cinctans have an ovate body with a well-developed marginal frame and a single rather rigid bilaterally symmetric appendage (Figure 3E). There is a small mouth opening through the marginal frame and to either side there is a groove that is roofed by a flexible sheet of cover plates. The right-hand groove is always the less well developed and may be lacking in some species. It seems likely therefore that cinctans possessed a pair of hydrocoel tentacles, with the left being better developed than the right. Whether these are both from the left hydrocoel or represent a hemichordate-like paired system is unknown, since hydropores have not as yet been definitely identified. At the anterior there is a large opening covered by an opercular plate, which acted as an outlet valve. This is best interpreted as an atrial opening and suggests that cinctans, like stylophorans, were active suspension feeders with some form of pharyngeal filtration basket. Like stylophorans they lived recumbent on the seafloor. Cinctans are restricted to the Cambrian.

Finally, the ctenocystoids are a small group of carpoids without a stem or tail, external gill slits, or ambulacra (Figure 3B). Their precise phylogenetic position remains uncertain, but they were free-living and possibly pharyngeal basket feeders like the Stylophora. They are found from the Middle Cambrian to the Upper Ordovician.

Helicoplacoids

The stratigraphically oldest fossil echinoderm assemblages known are dominated by helicoplacoids. Helicoplacoids are cigar-shaped echinoderms with a basal attachment at one end and a terminal anus at the other (Figure 3F). They have a lateral mouth, about two-thirds of the way up from which three ambulacra

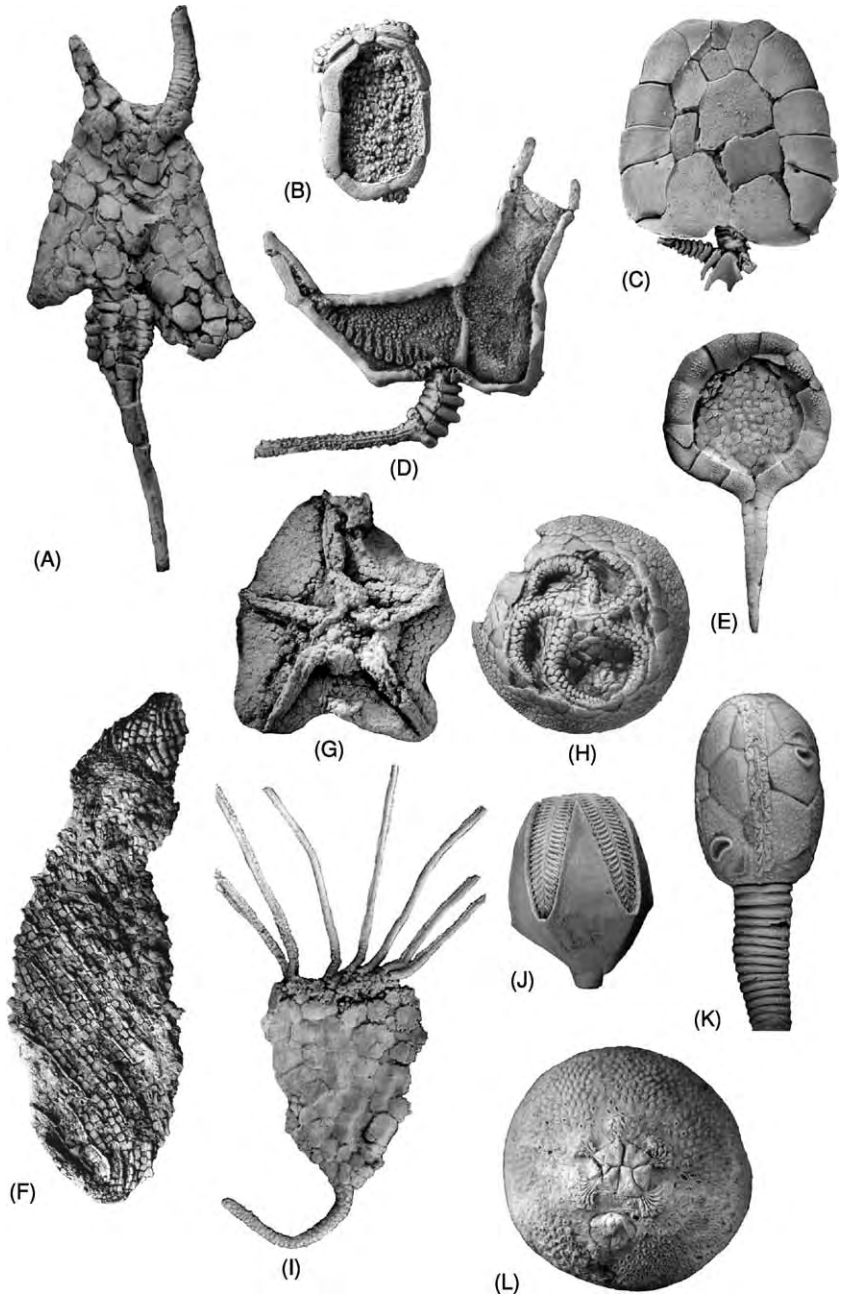


Figure 3 Representative fossil echinoderms. (A) The solute *Dendrocystoides* from the Upper Ordovician (magnification ×0.5). (B) The ctenocystoid *Ctenocystoides* from the Middle Cambrian (magnification ×3). (C) The mitrate stylophoran *Mitrocystites* from the Middle Ordovician (magnification ×1.4). (D) The cornute stylophoran *Cothurnocystis* from the Upper Ordovician (magnification ×1.4). (E) The cinctan *Trochocystites* from the Middle Cambrian (magnification ×1.4). (F) The helicoplacoid *Helicoplacus* from the Lower Cambrian (magnification ×1.4). (G) The stromatocystitid edrioasteroid *Stromatocystites* from the Middle Cambrian (magnification ×1). (H) The isorophid edrioasteroid *Carneyella* from the Middle Ordovician (magnification ×1.7). (I) The eocrinoid *Gogia* from the Middle Cambrian (magnification ×1.4). (J) The blastoid *Pentremites* from the Carboniferous (magnification ×1.4). (K) The rhombiferan *Apiocystis* from the Middle Silurian (magnification ×1.4). (L) The diploporite *Haplosphaeronis* from the Upper Ordovician (magnification ×1.4).

radiate, two spiralling upwards and one spiralling downwards. Each ambulacral series is composed of a double column of flooring plates and two flaps of cover plates. Pores indicate the positions of tube feet. The remainder of the body is covered in spirally arranged columns of rectangular elements. There is some evidence that helicoplacoids could extend and contract their body by changing the angle of spiralling. Helicoplacoids are confined to the Lower Cambrian and were suspension feeders living on algal-bound sediments. They used their tube feet to capture particles in the water column.

Blastozoans

Blastozoans are a large and morphologically diverse group of primitive stemmed echinoderms that mostly resemble crinoids in appearance and lifestyle. The plated body or theca bears a mouth, anus, and usually some form of specialized respiratory structure. The stem takes the form of a polyplated holdfast, a more regular stem comprising five series of vertically aligned ossicles, or a holomeric stem composed of disc-shaped columnals. A few blastozoans are secondarily stemless and either cemented directly to the seafloor or were free-lying. The mouth faces upwards, away from the seafloor. All blastozoans have some form of filtration fan composed of unbranched brachioles, which presumably bore tube feet and a water vascular system. However, in contrast to crinoid arms, which are formed as direct outgrowths of the body wall and carry extensions of the primary body coelom, blastozoan brachioles are side branches of the ambulacra and are not directly connected to the body coelom. Although many blastozoans show an obvious pentaradiality, diverse body forms, symmetries, and ambulacral architectures are to be found in this group. As currently constituted they are probably a polyphyletic group.

Eocrinoids are a paraphyletic assemblage of basal pelmatozoans that have an irregularly plated theca (Figure 3I). Simple sutural gaps termed epispires that notch the thecal plates are the only respiratory structures ever developed. Eocrinoids include the ancestors of all other blastozoan groups and probably of the crinoids also. Diploporites (Figure 3L) have larger and more regular plating and lack any form of ambulacral flooring plates; the brachioles arise directly from the thecal plates. Thecal plates are pierced by numerous pairs of pores (diplopores) that have a respiratory role. Rhombiferans (Figure 3K) have stout arms arising directly from around the mouth. Respiratory structures are always present and take the form of thin folded rhomboidal structures or tubes that run close to the inner or outer surfaces of the plates. Blastoids (Figure 3J) have the most regular thecal plating of any blastozoan, always constructed of three basals, five radials and five lancet plates. Pentameral symmetry is extremely well expressed. Ambulacra are well developed and form an integral part of the theca, giving rise to a dense fan of brachioles. Small openings along the margins of the ambulacra lead into an internal hydrospire system – a highly convoluted and thin-walled respiratory organ. The oldest blastozoans are eocrinoids of Lower Cambrian age. Diploporites and rhombiferans appear at the start of the Ordovician, while blastoids do not appear until the Silurian. The group finally went extinct at the end of the Permian.

Edrioasteroids

Edrioasteroids are an extinct group of sessile stem-group eleutherozoans that were discoidal, clavate, or subglobular in form (Figure 3G and 3H). They have a central mouth from which five ambulacra radiate. The ambulacra can be straight, but are commonly spiralled to provide a larger surface area, and extend to a marginal ring of plates marking the edge of the oral area. A single hydropore and gonopore open close to the mouth. Edrioasteroids lived with the mouth facing upwards and used their five ambulacra to capture organic material suspended in the water.

The earliest edrioasteroids (stromatocystitids: Figure 3G) were fully plated and appear not to have been attached to the substrate. These may represent basal eleutherozoans, ancestral to all later forms. The great majority of edrioasteroids, however, fall into a derived clade, the Isorophida (Figure 3H), that specialized as hard-ground colonizers. Isorophids lived attached to hard substrates, such as the shells of brachiopods, by their lower uncalcified surface. They were not permanently fixed but were able to realign themselves during life. Some isorophids developed the outer part of the disc as an extendible stalk, which allowed them to lower or raise the oral zone with the ambulacra, presumably in response to water flow conditions.

Stromatocystitids first appear in the Lower Cambrian, while isorophids appear a little later, in the Upper Cambrian, and survived through to the Upper Carboniferous.

Asteroids

Asteroids are stellate echinoderms with arms that merge without a break into a central body (Figure 4A). The mouth is central and faces down towards the substrate, and an ambulacral groove runs from the mouth along the lower surface of each arm. In asteroids the arms are hollow and extensions of the digestive system and gonads extend along each. Asteroids lack a jaw apparatus but are active predators, ingesting small prey whole. Some burrow after infaunal animals, while others are epifaunal hunters. One group has evolved an ingenious method of preying on bivalves. They clasp the bivalve using their tube feet and prise the two valves apart. They need to create only a small slit-like opening because they then evert their stomach into the bivalve and digest the animal. The asteroid skeleton is rarely robust and consists of a series of small elements embedded in a collagenous membrane. Consequently, they disarticulate rapidly after death and have left a rather poor fossil record. Asteroids first appear in the Early Ordovician and are never common or abundant.

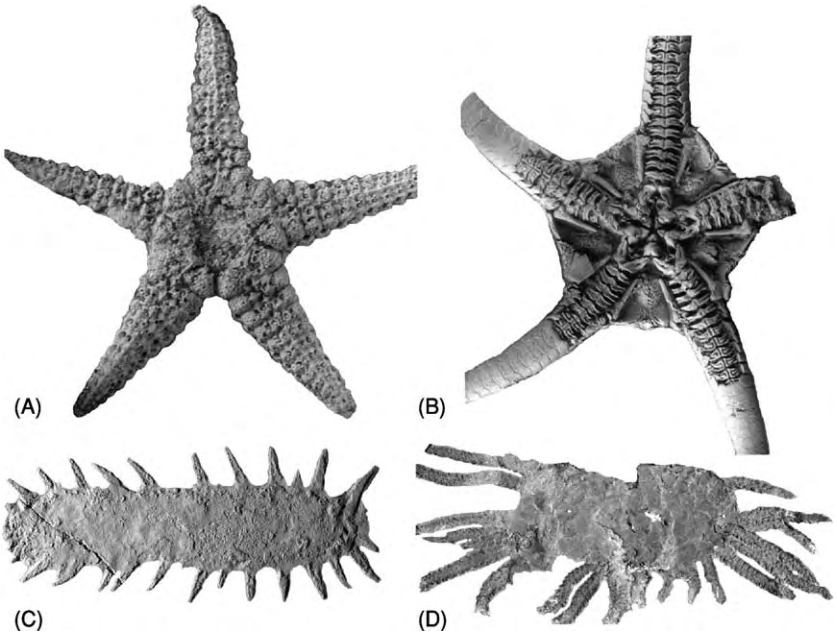


Figure 4 Representative fossil echinoderms. (A) The asteroid *Siluraster* from the Upper Ordovician (magnification ×2). (B) The ophiuroid *Palaeocoma* from the Lower Jurassic (the upper surface of the disc and the proximal parts of the arms have been lost, revealing the solid vertebrae filling the core of the arms and the central jaw) (magnification ×1). (C) The holothurian *Oneirophantites* from the Middle Triassic (magnification ×0.3). (D) The ophiocistiod *Sollasina* from the Upper Silurian (magnification ×0.7).

Ophiuroids

Ophiuroids resemble asteroids in having a stellate body plan, with five or more arms radiating from a small circular disc (Figure 4B). Unlike asteroids, however, their arms are solid, being supported by a series of internal disc-like ossicles termed vertebrae. Consequently, there are no extensions of the gonads or digestive system into the arms. The mouth is central and faces downwards. The digestive system is a simple blind sac (there is no anus), and the lower surfaces of the arms carry the water vascular system and the tube feet. All ophiuroids are carnivorous and possess a formidable jaw apparatus with strong musculature and batteries of tooth-like spines.

There are two major post-Palaeozoic groups, the euryalids and the Ophiuroida. In euryalids the arms are typically branched many times to form a filtration fan and bear short spike-like spines. Euryalids are suspension feeders that use their network of arms to ensnare small nektonic prey such as arrow worms, which are then passed to the mouth. The great majority of Ophiuroida are active predators and are able to move rapidly over the seafloor by using their highly motile arms. A few are also able to suspension feed by extending one or more arms into the water column and using their tube feet to secrete ribbons of sticky mucous to ensnare small prey.

The ophiuroid skeleton is rather fragile and readily disintegrates upon death. Consequently, like asteroids they have left a relatively sparse fossil record, starting in the Lower Ordovician.

Ophiocistioids

Ophiocistioids are a small but interesting extinct group of globular eleutherozoans with large plated tube feet (Figure 4D) that are related to both echinoids and holothurians. The mouth in ophiocistioids is central and downward facing, and there is a complex jaw apparatus identical in all important respects to the Aristotle's lantern of echinoids. They were therefore presumably active predators, like early echinoids, using their jaws to capture small benthic prey. The large plated tube feet were locomotory in function, and their water vascular system lay beneath the ambulacral plates, as in echinoids and holothurians, rather than externally, as in asteroids and crinoids. In early members the body is covered in a series of small plates. In at least one taxon, however, the body wall is reduced to microscopic spicules, which are wheel-shaped and identical to those seen in apodid holothurians. Ophiocistioids first appear in the Middle Ordovician and continue to the Early Carboniferous. They are always rare.

Holothurians

Holothurians have a cylindrical body plan with the mouth at one pole and the anus at the other (Figure 4C). The mouth is surrounded by a ring of

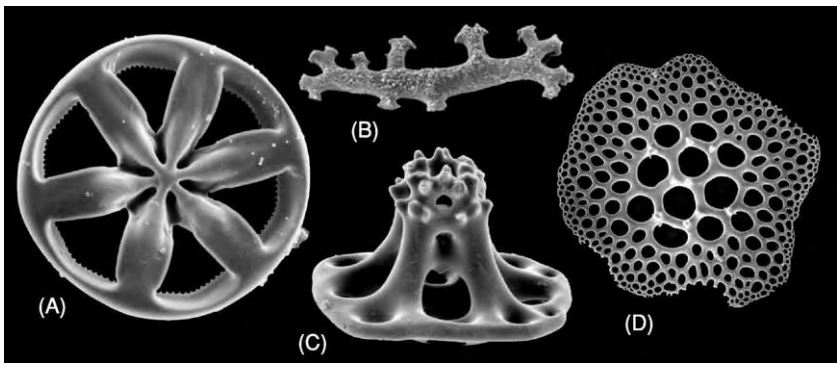


Figure 5 Characteristic spicules from the body walls of holothurians. (A) Wheel (magnification ×600). (B) Rod (magnification ×700). (C) Button (magnification ×800). (D) Plate (magnification ×50).

tentacles, which are often highly branched and are used in feeding. There is no jaw apparatus. Some holothurians have long highly branched tentacles that are extended into the water column and used for suspension feeding. Others have short finger-like tentacles that are used in deposit feeding. Today holothurians are the major deposit feeders in back-reef habitats. They live either on the surface or in shallow burrows.

Holothurians are the most diverse of the five extant classes of echinoderm, with over 2000 extant species, but they have the poorest of fossil records. This is because their skeleton is reduced to microscopic spicules. The only elements of any size are 10 ossicles that surround the mouth and provide an anchorage for the oral tentacles; these form the circumoral ring. Each holothurian has many thousands of microscopic spicules in its body wall. Many are distinctive, such as the anchor-like elements of apodians and the wheel-like elements of molpadiids (Figure 5). A few body fossils are known from Lagerstätte (Figure 4C) but almost all of our information on the history of holothurians comes from the study of their isolated spicules distributed in the sedimentary record. The first holothurian body fossil comes from the Late Silurian, but spicules probably attributable to holothurians are known from the Ordovician onwards.

See Also

Fossil Invertebrates: Crinoids; Echinoids; Graptolites. **Lagerstätten**. **Palaeoecology**.

Further Reading

Beaver HH, Caster KE, Durham JW, *et al.* (1978) *Treatise on Invertebrate Paleontology. Part S, Echinodermata 1.* Boulder, Colorado and Lawrence, Kansas: The Geological Society of America and the University of Kansas Press.

Bell BM (1976) A study of North American Edrioasteroidea. *New York State Museum, Memoirs* 21: 1 447.

Gilliland P (1993) The skeletal morphology, systematics and evolutionary history of holothurians. *Special Papers in Palaeontology* 47: 1 147.

Guensburg TE and Sprinkle J (1994) Revised phylogeny and functional interpretation of the Edrioasteroidea based on new taxa from the Early and Middle Ordovician of Western Utah. *Fieldiana Geology* 29: 1 41.

Hess H (1975) Die fossilen Echinodermen des Schweizer Juras. *Veroffentlichungen aus dem Naturhistorischen Museum Basel* 8: 1 130, pls 1 48.

Jagt JWA (2000) Late Cretaceous Early Palaeogene echinoderms and the K/T boundary in the southeast Netherlands and northeast Belgium. Part 3, Ophiuroids. *Scripta Geologica* 121: 1 179.

Jagt JWA (2000) Late Cretaceous Early Palaeogene echinoderms and the K/T boundary in the southeast Netherlands and northeast Belgium. Part 5, Asteroids. *Scripta Geologica* 121: 377 503.

Littlewood DTJ, Smith AB, Clough KA, and Ensom RH (1997) The interrelationships of the echinoderm classes: morphological and molecular evidence. *Biological Journal of the Linnean Society* 61: 409 438.

Paul CRC (1973 1997) British Ordovician cystoids. *Monograph of the Palaeontographical Society,* 1 213, pls 1 33.

Smith AB, Peterson K, Littlewood DTJ, and Wray GA (2004) From bilateral symmetry to pentaradiality: the phylogeny of hemichordates and Echinodermata. In: Cracraft J and Donoghue M (eds.) *Assembling the Tree of Life.* Oxford: Oxford University Press.

Sprinkle J (1982) Echinoderm faunas from the Bromide Formation (Middle Ordovician) of Oklahoma. *The University of Kansas Paleontological Contributions Monograph* 1: 1 369.

Waters JA and Maples CG (1997) Geobiology of echinoderms. *The Paleontological Society Papers* 3: 1 355.

Crinoids

M J Simms, Ulster Museum, Belfast, UK

Introduction

Crinoids are one of five extant classes of echinoderm (the others being echinoids, asteroids, ophiuroids and holothuroids (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids))) with a rich fossil record extending back nearly 500 Ma to the Early Ordovician. They are exclusively marine suspension feeders and, with a few exceptions, sessile benthos. Crinoids reached a peak of taxonomic diversity in the Late Palaeozoic but experienced a catastrophic decline around the Permo-Triassic boundary, from which they subsequently rediversified. More than 700 extant species and over 15 000 fossil species have been described.

Like other echinoderms, crinoids have a multielement endoskeleton of high-magnesium calcite ossicles connected by soft tissue and enclosed by a thin veneer of living tissue. Each ossicle is a single optically continuous calcite crystal with a stereom structure – a labyrinthine network of cavities permeated with soft tissue called stroma.

In the basic crinoid design the endoskeleton can be divided into three morphological sections. The stem or column is a usually slender flexible elongate structure for attachment to the substrate and elevation above it. The stem is surmounted by the cup, which is a fairly rigid structure, usually constructed of two or three circlets of plates, containing organs of digestion, movement, and nervous control. Arising from the upper circlet of the cup are elongate flexible arms involved in food gathering. Typical crinoids have a superficial similarity to flowers, giving rise to their popular name of 'sea lilies' and indeed to the name 'crinoid' itself (from the Greek *krinos* = a lily). Stalked crinoids are very much in a minority among extant taxa, with fewer than 100 species. In the remaining more than 600 species a stem is present only at the larval stage; the stemless adults are often termed 'feather stars.'

The soft parts of crinoids comprise a small and inconspicuous proportion of the total mass. They are virtually unknown in fossil crinoids, although their presence and functions can be inferred from the structure of the preserved hard parts and by analogy with extant crinoids. The digestive system is contained entirely within the cup, with both mouth and anus opening onto the upper surface. In common with all echinoderms, and unique to them, crinoids possess a network of fluid-filled tubes called the water vascular system. Branching canals extend radially into the arms from a central ring canal housed in the cup and terminate in the tube feet – small tentacle-like structures that detect and capture food particles before passing them to ciliated food grooves, which run down the arms to converge on the mouth. A further unique echinoderm character, almost certainly present in fossil taxa, is catch connective tissue, which is a type of ligament that can change its properties from pliable to rigid, enabling parts of the skeleton, such as the stem or arms, to 'lock' into position for prolonged periods with minimal energy expenditure. Catch connective tissue is noncontractile and cannot contribute to active movement. Only muscles are able to do this; they are restricted to the arms of only some taxa and are entirely absent from the stem.

Morphology and Functional Interpretations

A crinoid's morphology is determined by the interaction of three distinct factors: inherited characteristics, i.e. the morphology of its immediate ancestor; architecture, i.e. physical properties of the materials; and evolutionary selection pressure from the external environment, i.e. ecological factors.

The skeleton of most crinoids is readily divisible into three distinct sections – the stem, cup, and arms (Figure 1) – a basic morphology that was inherited from a pre-crinoid ancestor and subsequently modified within limits imposed by ancestry, architecture, and ecology. In a few specialist taxa the stem or arms may be greatly reduced or absent.

The stem's main functions are attachment to the substrate and, especially, elevation above it. Most fossil crinoids have a stem ranging from a few centimetres to perhaps a metre or so in length, similar to that of extant stalked crinoids, although in some taxa it is very much longer while in others it may be reduced or absent (particularly in free-living crinoids such as the comatulids). Typically, the stem comprises a stacked series of ossicles, or columnals, pierced by a central canal containing extensions of the coelom and nervous system. Each columnal is typically a single ossicle, but a few Early Palaeozoic taxa have what are known as meric columnals constructed from several ossicles or meres. Each columnal is connected to adjacent columnals by catch connective tissue. Short

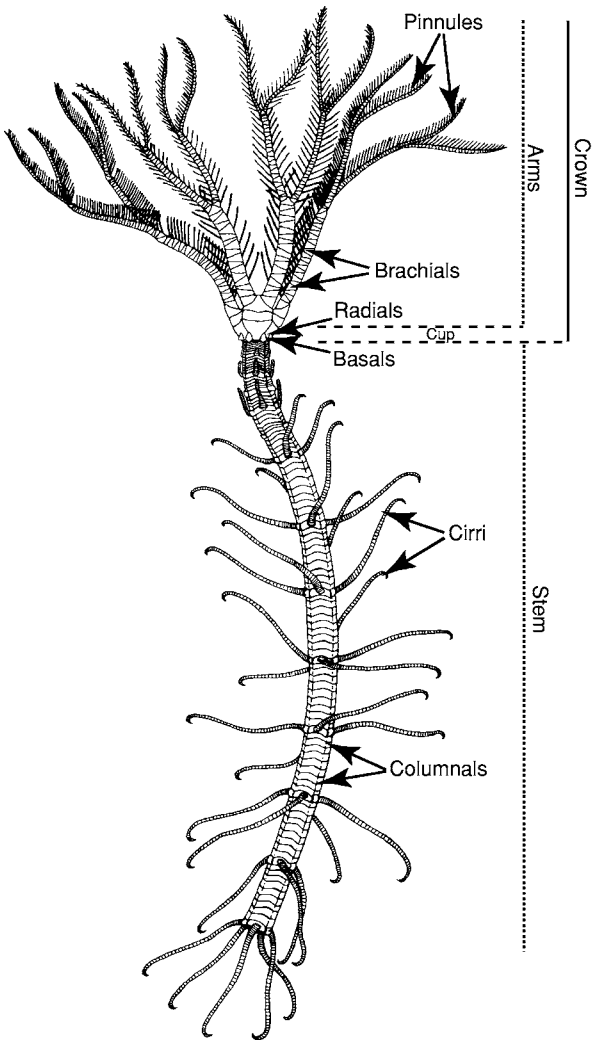


Figure 1 The main morphological features of an extant stalked crinoid (isocrinid). Some of the arms have been omitted for clarity.

ligaments connect opposing faces of adjacent columnals, with longer through-going ligaments linking groups of columnals together. Autotomy – the voluntary shedding of distal parts of the stem to allow the animal to drift to a new site – can occur at a junction between adjacent linked groups of columnals through a change in the properties of the catch connective tissue. Hence, discarded columnals can contribute to the future fossil record long before the animal itself has actually died! Articulating surfaces of adjacent columnals show a range of morphologies, most typically a radial pattern of crenulations (Figure 2A), which interlock with opposing crenulations on the adjacent columnal to resist shearing and torsional stresses. These crenulations may be arranged in more complex, often pentaradiate, configurations (Figure 2B). Another distinctive type of columnal articulation developed independently in the Silurian to Permian

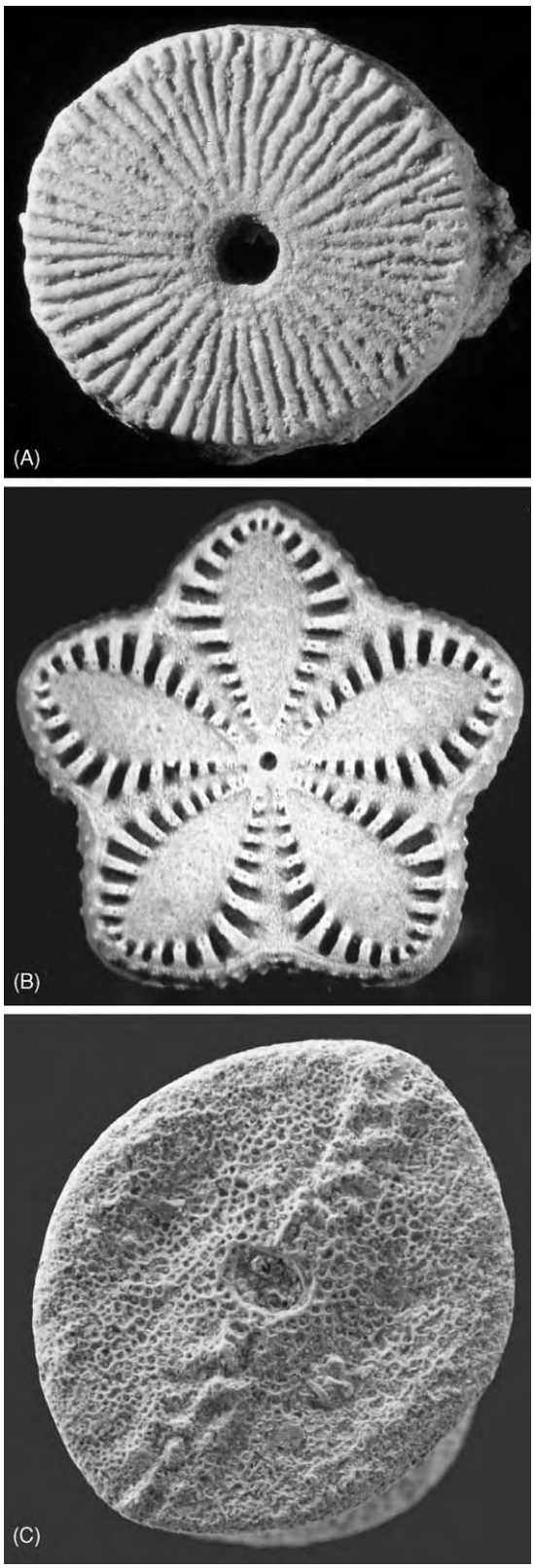


Figure 2 The main types of columnal articulations encountered in crinoids: (A) multiradiate is 8 mm across; (B) pentaradiate is 12 mm across; and (C) synarthrial. The columnal in (C) is from a larval isocrinid and is less than 0.7 mm across; the stereom is clearly visible.

platycrinitid camerates and the Cretaceous to Holocene bourgueticrinids and their allies. In these crinoids, the articulations (Figure 2C) are traversed by a prominent fulcral ridge flanked by two ligament pits, with the fulcral ridges on the proximal and distal faces of each columnal typically at an angle to each other.

The crinoid stem may be more or less permanently attached to a hard substrate by a cemented holdfast, may be rooted in sediment by irregular branches or outgrowths from the stem or may attach on a more temporary basis to hard or soft substrates using regular offshoots of the stem, known as cirri. Cirri are the most versatile form of attachment and were adopted independently by several major crinoid groups, including the dominant extant group – the stemless comatulids.

The cup or calyx, located between the stem and the arms, houses the vital organs of the digestive, nervous, and water vascular systems. Typically the cup consists of a regular series of interlocked circlets, each usually of five plates, often with one or two additional plates intercalated towards the top. In the earliest known crinoid (the Early Ordovician *Aethocrinus*) there were four circlets, but in virtually all other crinoids the plates of the lowest circlet have been lost and the cup comprises either two or three circlets (see below). Typically the plates of each circlet are offset by 36° relative to the plates of adjacent circlets, imparting rigidity even with only two circlets. The shape of the cup varies, from shallow and bowl shaped to almost globular, with the same basic designs having evolved repeatedly in different groups. The calyx incorporates all of the plates between the top of the stem and the base of the free arms, including the cup, the tegmen (which covers the oral surface), and any parts of the arms that are incorporated into the tegmen. In many crinoids the cup plates are connected only by ligaments and the tegmen is weakly constructed; hence, it disarticulates rapidly after death. However, in the Palaeozoic subclass Camerata the calyx forms a relatively rigid and often globose structure, which may remain intact for some time after death. Consequently camerate calyces are better represented as fossils than are those of the other crinoid subclasses (see below).

The food grooves of the arms converge on the mouth, which is located on (or beneath in the case of camerates) the surface of the tegmen along with the anus. The anal opening is commonly elevated on an anal tube or anal sac, forming a large and complex structure in some fossil taxa, presumably to avoid faeces entering the mouth.

The arms are the food-gathering parts of crinoids, although ultimately it is the tube feet, lining a groove on the oral side of the arms, that are directly involved in capturing food and moving it towards the mouth. The arms increase the support area for tube feet but are not essential, as evidenced by a few crinoid taxa in which they are reduced in number or even absent. Secondary functions of the arms include respiration and locomotion, but generally selection pressure is towards improvement of the food-gathering mechanism within the constraints imposed by other factors.

Flexibility of the arms is achieved in the same manner as flexibility of the stem, with brachial plates connected by ligaments and/or muscles. Muscle tissue is present in all echinoderm tube feet but appeared in brachial articulations in only one mid-Palaeozoic group, which ultimately evolved into the post-Palaeozoic subclass Articulata. As in the stem, specialised ligamentary articulations can be used to autotomize parts of the arms.

At their simplest, crinoids have five unbranched arms, each arising from one of the plates in the uppermost circlet of the cup. Increasing the effective arm length, by branching, increases the number of tube feet that can be supported and hence increases the filtration efficiency. The development of numerous small side branches, or pinnules, is one strategy that has evolved independently in several major groups. The arms themselves show a wide range of branching patterns, from simple dichotomous (isotomous) (Figure 1) through to strongly endotomous (Figure 3). Endotomous branching appears to be the most efficient pattern, in terms of expenditure of materials versus food-gathering capabilities, and an analogy has been drawn with the arrangement of roads on banana plantations. Intriguingly, only a few fossil crinoid taxa developed this pattern to a significant extent.

Phylogeny, Systematics, and Geological History

With their morphologically complex multielement skeletons, crinoids are ideal subjects for phylogenetic analysis. Each part of the skeleton – stem, cup, and arms - can provide a wealth of morphological data for use in descriptive and phylogenetic investigations. However, all too often primary descriptions of genera and species are lacking in detail, particularly for the stem, while interpretations at higher taxonomic levels have been hindered both by this primary deficiency of data and, until the mid-1990s, by a rigid adherence to traditional interpretations of skeletal homologies.

At high taxonomic levels the structure of the calyx is considered to be of major phylogenetic significance.

Figure 3 Crown, stem, and cirri of *Pentacrinites fossilis*, a pseudoplanktic crinoid from the Early Jurassic. The arms are strongly pinnulate and branch endotomously, with each of the main arms bifurcating twice and subsequent side branches arising only from the inner sides of the main branches. Crowded cirri can be seen arising from the stem in the lower part of the picture.

A few of the earliest crinoids, grouped together in the subclass Aethocrinea (named after the genus *Aethocrinus*), have four circlets of plates in the cup: lintels (at the base), infrabasals, basals, and radials (top) (Figure 4). This number is reduced in all other crinoids, although the exact circlet(s) missing may vary, with potentially any one or more of the four circlets being lost to produce a three-, two- or, in at least one species, one-circlet cup. In the initial dichotomy after the four-circlet ancestor, one group (the subclass Disparida) lost the basal plates to leave a cup composed of lintels, infrabasals, and radials. Disparid cups may display some bilateral symmetry, with the constituent plates of each circlet being of unequal, or disparate, size, giving the subclass its name. Later disparids lost some or all of the radials and, in a few taxa, even the infrabasals, reducing the cup to a single circlet of lintels. The other branch of the dichotomy lost the lintels to form a cup of infrabasals, basals, and radials. This branch comprises two distinct subclasses, the Cladida and the Camerata, each of which contributed significantly to Palaeozoic crinoid diversity alongside the disparids. In both cladids and disparids the free arms arise directly from the top of the cup, but in camerates the calyx incorporates fixed brachials and interradials in addition to the cup, so that the free arms begin above the top of the cup. Camerates also have the tegmen firmly fixed to the cup to form a rigid chamber, from which their name derives. Camerates experienced a further important dichotomy early in their history, giving rise to the three-circlet diplobathrids and the two-circlet monobathrids in which the infrabasal circlet was lost. Cladids, which derive their name from the often multiple branching of the arms, also gave rise to two subclasses – the Flexibilia and the Articulata. The morphological differences that separate articulates or flexibles from the cladids are less profound than those that separate the other four subclasses from each other. Flexibles have a three-circlet cup like that of cladids, although the infrabasal circlet is typically reduced to three plates; their name derives from the weakly sutured plates of the cup and the inrolled tips of the arms. The articulates are the only post-Palaeozoic subclass, although their origins can be traced back well into the Palaeozoic. With a few exceptions the articulate cup is reduced to two circlets – basals and radials – with the basals also greatly reduced in some taxa. The brachial articulations are predominantly muscular, a characteristic after which the subclass is named.

The precise group from which crinoids originated is unclear. A considerable diversity of noncrinoid pelmatozoan (stalked) echinoderms characterized the Early Palaeozoic, and crinoids evolved from one of these through a reduction in the complexity of the cup and the evolution of true arms carrying extensions of the coeloms and of the nervous and water

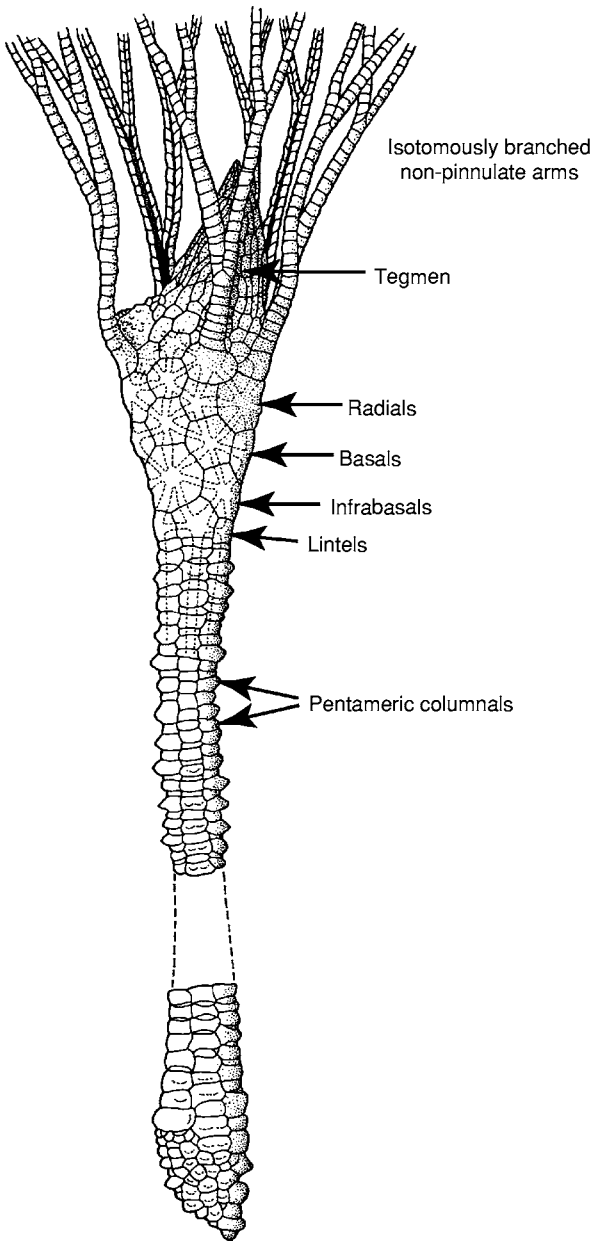


Figure 4 Reconstruction of *Aethocrinus moorei* from the Early Ordovician the earliest undisputed crinoid known. This has the primitive four circlet arrangement of plates in the cup, simple isotomous arms, and a pentameric stem. Annotated and reproduced with permission from Hess H, Ausich WI, Brett CE, and Simms MJ (1999) *Fossil Crinoids*. Cambridge: Cambridge University Press.

vascular systems. Some authorities favour an origin among the Late Cambrian or earliest Ordovician rhombiferan 'cystoids'; others have suggested, based on several crinoid-like pelmatozoans discovered in the earliest Ordovician (Tremadoc), that they are descended from an edrioasteroid ancestor.

The earliest undisputed crinoid yet discovered is *Aethocrinus moorei*, from the Early Ordovician (Tremadoc or Arenig) (Figure 4). *Echmatocrinus brachiatus*, from the Middle Cambrian Burgess Shale, was for nearly two decades regarded as a possible ancestor of all later crinoids on account of its supposedly irregularly multiplated cup, but its identity as an echinoderm is now disputed.

There was considerable diversification in the Early Ordovician such that by the start of the Llanvirn Stage nearly all of the major Palaeozoic clades (aethocrinids, disparids, cladids, diplobathrid camerates and monobathrid camerates) were represented (Figure 5). The remaining subclass, the Flexibilia, arose from among the cladids during the Caradoc Stage of the Late Ordovician. This diversification continued into the Late Ordovician, and crinoids as a whole suffered only a minor drop in diversity during the mass extinction near the close of the Ordovician, although thereafter disparids ceased to be a dominant element of crinoid faunas. The fossil record of crinoids is poor for the Llandovery Stage but those from later in the Silurian and into the Devonian show clearly that the overall diversification continued unabated, though several important families declined or disappeared as a result of the Late Devonian (Frasnian–Famennian) extinction event. Crinoid diversity reached its acme during the Late Palaeozoic; several thousand species have been described from the Lower Carboniferous, and an apparent reduction in diversity into the Permian may well be merely an artefact of collection or description failure. Crinoid debris is often abundant and diverse in late-Permian marine successions, even in the very highest strata, but this disarticulated material has not been thoroughly investigated. It is clear that, along with so many other groups, crinoids experienced a profound drop in diversity from the Permian into the Triassic, but whether crinoid diversity remained high right up to the Permo-Triassic boundary, or was declining prior to this, remains unclear.

In Early Triassic marine sediments crinoid remains are rare and form a very low-diversity or even monospecific assemblage of small crinoids, assigned to the genus *Holocrinus*, which is the common ancestor of the subclass Articulata and all post-Palaeozoic crinoids. The articulates arose from a group of morphologically advanced cladids with muscular arm articulations; all of the other cladids, along with the flexibles, disparids and camerates, appear to have been eradicated entirely by the end-Permian extinction.

The articulates diversified rapidly in the mid- to Late Triassic but experienced a significant extinction early in the Late Triassic (mid-Carnian Stage), marked particularly by the disappearance of the distinctive and diverse encrinids. Jurassic faunas were

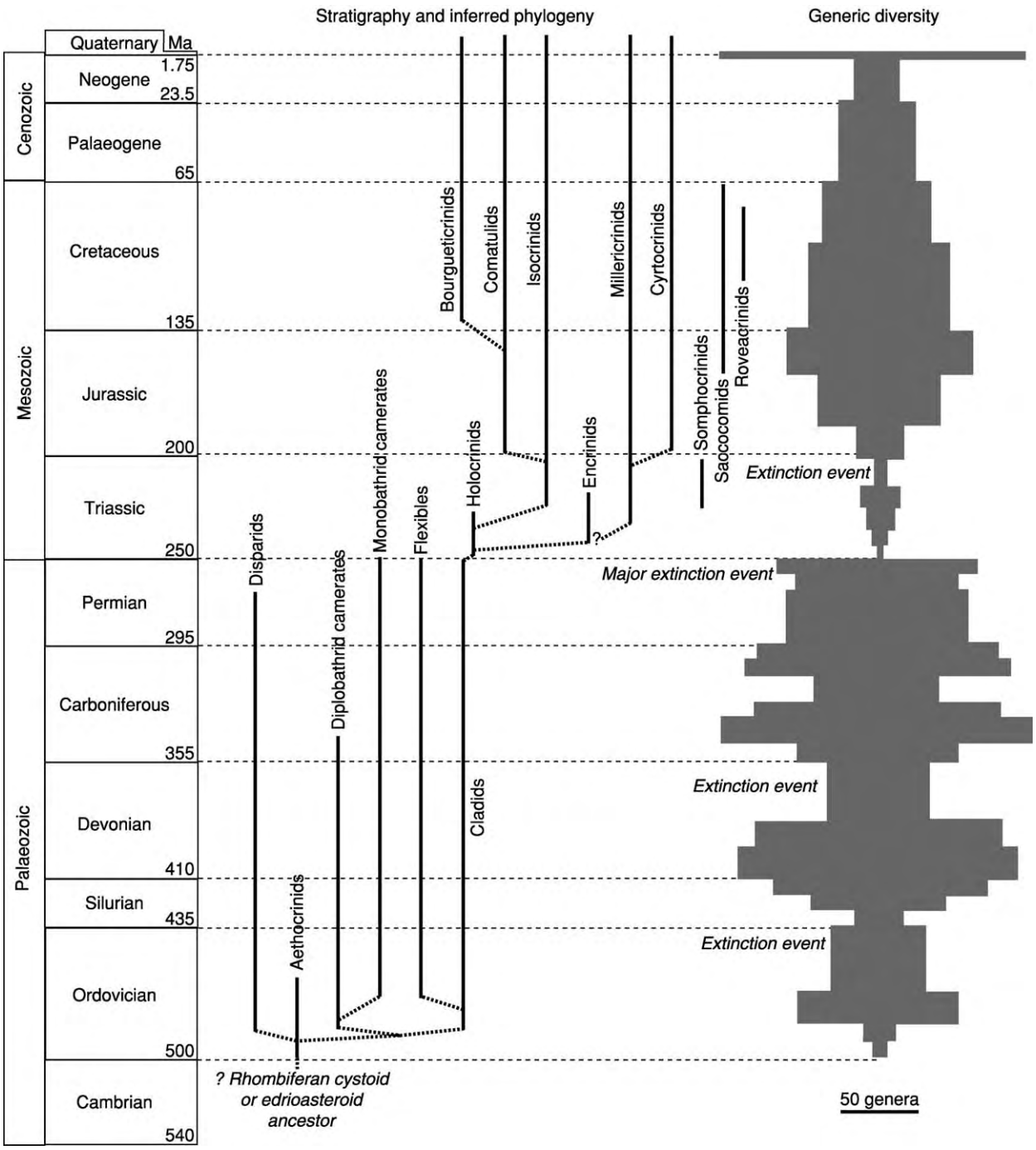


Figure 5 Stratigraphic distribution and phylogenetic relationships of the main crinoid groups, with total crinoid diversity indicated on the right.

dominated by two distinct clades – the isocrinids and the millericrinids. A stemless offshoot of the former, the comatulids, became increasingly dominant from the mid-Jurassic and now accounts for some 85% of the total crinoid taxonomic diversity.

A recurrent theme in the history of the Crinoidea is the appearance of microcrinoids, which may be broadly defined as those with an adult cup diameter of about 2 mm or less. Palaeozoic microcrinoids are known from the Silurian to the Permian and were exclusively sessile stalked benthos; most are assigned to a small number of cladid and disparid families. Mesozoic microcrinoids were quite different, lacking a stem during all known growth stages, leading to

suggestions that they were planktic in habit. Three distinct groups are recognized: the mid- to Late Triassic Somphocrinidae, the Late Jurassic Saccocomidae, and the Cretaceous Roveacrinidae. They are often grouped together as 'roveacrinids,' but each group probably evolved independently from a more 'normal' ancestor. No microcrinoids are known after the end of the Cretaceous.

Ecology and Taphonomy

All extant crinoids are fully marine, intolerant of brackish or hypersaline conditions, and have relatively high oxygen demands. Evidence from the fossil record suggests similar ecological limits for extinct taxa, and hence they are useful palaeoenvironmental indicators. Extant crinoids are exclusively passive suspension feeders, and this was probably the case for most fossil taxa, although there have been suggestions of active filtration mechanisms in some highly specialized taxa, while among the microcrinoids there may have been an element of active podial capture of plankton. For many decades it was assumed that typical crinoids adopted a 'feeding bowl' arm configuration to intercept detritus sinking through the water column. However, observations by deep submersible in the 1970s indicated that many stalked crinoids adopt a quite different strategy, that of a 'parabolic filtration fan' in which the crown is held roughly at right angles to the current to intercept material carried along by it (Figure 6). Crinoids in turbulent environments where there is no single, or tidally reversing, dominant current adopt a range of feeding configurations depending on the degree of turbulence and the number of arms, but all appear to exploit currents rather than depending on passive settling of detritus.

Elevating the feeding arms helps to place them into faster-moving currents above the benthic boundary layer (Figure 6). The development among Early Ordovician crinoids of columns substantially longer than those of contemporaneous non-crinoid pelmatozoans may well have been a major factor in the overwhelming success of crinoids, compared with other echinoderms, in the Palaeozoic. Elevation of the arms also reduced competition with other benthic suspension feeders, including other species of crinoid. During periods of high crinoid diversity, such as the Early Carboniferous, crinoid crowns occupied several distinct tiers, up to a metre or so above the sea floor, thereby partitioning the food resource brought along by horizontal currents. In some environments swept by particularly strong or turbulent currents the entire crinoid morphology may show profound ecological adaptations, sometimes with very short and

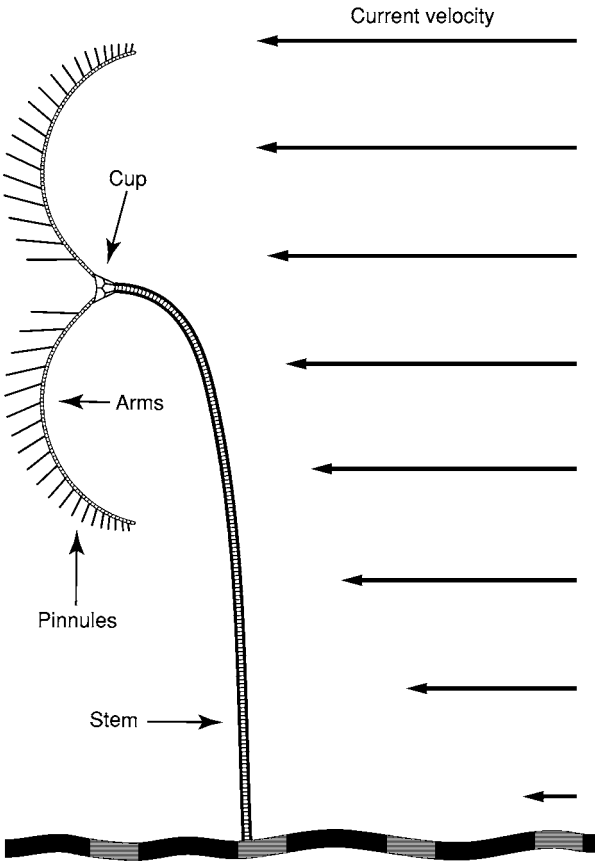


Figure 6 Diagrammatic reconstruction of a stalked crinoid in a rheophilic feeding posture, showing the position of the arms and the reduction of current flow velocity near to the benthic boundary layer.

robust stem and arms, and occasionally with the cup cemented directly to the substrate. The Jurassic to Holocene cyrtocrinids exemplify a range of such morphologies.

Despite the relatively low proportion of soft tissue in crinoids, examples of predation are well documented among extant taxa, and the spinose arms and/or calyces of some fossil taxa may have evolved as predator deterrents. However, the regenerative powers of crinoids are formidable, and entire crinoid crowns may regenerate from just a surviving basal circlet. The ability of crinoids to autotomize arms and regenerate new ones is important in this respect, and there are many fossil examples known where this has clearly happened.

The construction of the crinoid endoskeleton, with a multi-element skeleton of calcite ossicles held together by soft tissues, confers on them poor preservation potential as intact specimens. The soft tissues decay quickly after death, often within days, and disarticulation is rapid in the often turbulent and well-oxygenated environments favoured by crinoids.

Muscles decay most rapidly, with the short ligaments connecting adjacent ossicles decaying faster than the longer through-going ligaments found in the stems. Hence the postmortem disarticulation of crinoids often follows a definite sequence, with lengths of stem (pluricolumnals) held together by the through-going ligaments being the last to disarticulate into individual ossicles. This disarticulated material is common in the fossil record, in some instances attaining rock-forming abundance. Such crinoidal limestones, often termed 'encrinites' (Figure 7), occasionally reach thicknesses of tens of metres over many thousands of square kilometres. Many examples are known from the Upper Palaeozoic, but they are rare thereafter, with only a few known from the Triassic and Jurassic.

Preservation of intact crinoids indicates an unusual event. In most instances it can be attributed to obrution – the smothering of the living animals by a sudden influx of sediment sufficient both to cause death and to protect the remains against subsequent disruption by burrowing scavengers. Storm resuspension of sediment is probably the most common cause of obrution, and hence some successions, deposited in environments prone to obrution, may include numerous horizons with intact crinoids. Crinoids are also sometimes found intact in dark laminated mudstones that were clearly deposited in anoxic conditions inimical to their survival but greatly enhancing their preservation potential through the exclusion of scavengers and the slowing of decay rates. Some of these crinoids can be interpreted as allochthonous faunal elements catastrophically swept in from a more favourable adjacent environment. However, a few crinoid taxa are regularly found in such anoxic facies and appear to be poorly represented elsewhere, indicating an unusual and specialized ecology. Four such genera – the Early Jurassic *Pentacrinites* (Figure 3) and *Seirocrinus*, the Late Triassic *Traumatocrinus*, and the Late Devonian *Melocrinites* – are often intimately associated with fossil driftwood, and there is considerable evidence, such as the crinoids almost invariably being preserved beneath the wood, that in life these crinoids were pseudoplanktic. They were introduced into these anoxic environments by accident when the floating driftwood to which they were attached finally sank, perhaps under the burden of its rapidly growing crinoid passengers. Pseudoplanktic crinoids show remarkable morphological convergence, particularly in their long stems (examples of *Seirocrinus* are known with stems over 20 m long!) and strongly endotomous arm branching, suggesting intense evolutionary selection pressure in this highly specialized niche.

There have been suggestions of a planktic mode of life for several taxonomically, and morphologically, disparate groups of fossil crinoids with near cosmopolitan distributions. The small and delicate construction of the Mesozoic 'roveacrinids' has been considered to be an adaptation to a planktic or nektic habit, although this is still subject to debate. Similarly, a planktic mode of life has often been proposed for

Figure 7 A typical encrinite, or crinoidal limestone, from the Early Carboniferous of North Wales, UK. Field of view 30 cm.

two large stemless crinoid genera from the Cretaceous – *Marsupites* and *Uintacrinus* – although recent interpretations favour a benthic habit. In the Late Silurian to Early Devonian a bizarre group of large camerate crinoids – the Scyphocrinitidae – modified the distal end of the stem into a large chambered 'bulb,' which has been interpreted as a flotation structure.

See Also

Fossil Invertebrates: Echinoderms (Other Than Echinoids); Echinoids. **Mesozoic:** Triassic. **Palaeoecology**. **Palaeozoic:** Devonian; End Permian Extinctions. **Sedimentary Environments:** Carbonate Shorelines and Shelves.

Further Reading

Ausich WI (1998) Early phylogeny and subclass division of the Crinoidea (Phylum Echinodermata). *Journal of Paleontology* 72: 499 510.

Cowen R (1981) Crinoid arms and banana plantations: an economic harvesting analogy. *Paleobiology* 7: 332 343.

Guensburg TE and Sprinkle J (2003) The oldest known crinoids (Early Ordovician, Utah) and a new crinoid plate homology system. *Bulletins of American Palaeontology* 364: 1 43.

Hess H, Ausich WI, Brett CE, and Simms MJ (1999) *Fossil Crinoids*. Cambridge: Cambridge University Press.

Meyer DL and Ausich WI (1983) Biotic interactions among Recent and among fossil crinoids. In: Tevesz MJS and McCall PL (eds.) *Biotic Interactions in Recent and Fossil Benthic Communities*, pp. 377 427. New York: Plenum Publishing.

Moore RC and Teichert C (eds.) (1978) *Treatise on Invertebrate Paleontology, Part T, Echinodermata* 2, vols 1 3. Boulder, Colorado and Lawrence, Kansas: Geological Society of America and University of Kansas.

Seilacher A, Reif W E, and Westphal F (1985) Extraordinary fossil biotas: their ecological and evolutionary significance. *Philosophical Transactions of the Royal Society of London B* 311: 5 23.

Simms MJ (1986) Contrasting lifestyles in Lower Jurassic crinoids: a comparison of benthic and pseudopelagic Isocrinida. *Palaeontology* 29: 475 493.

Simms MJ (1994) Reinterpretation of thecal plate homology and phylogeny in the Class Crinoidea. *Lethaia* 26: 303 312.

Echinoids

A B Smith, The Natural History Museum, London, UK

Introduction

Echinoids are a diverse and widespread group of marine macroinvertebrates and one of the five classes of echinoderm (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids)). They are exclusively marine and benthic, living either on or within the sea floor, and with their mouth downward-facing. The group includes such well-known animals as the sand dollar, heart urchin and sea urchin. Like all echinoderms, they have a calcitic skeleton constructed of a mosaic of plates, and these are usually firmly bound together to form a rigid and robust test. This skeleton is architecturally complex and, being mesodermal, is able to provide a wealth of information about the musculature and soft-tissue anatomy. Because echinoids have a rigid skeleton, they have left the best fossil record of any echinoderm class. They are also the best understood echinoderm group from a phylogenetic perspective. This combination of a complex skeleton, good fossil record and extensively studied phylogeny has made echinoids the focus of many important palaeobiological studies in recent years.

There are about 900 extant species equally divided between regular forms ('regulars') whose anus opens in the aboral plated surface and that live epifaunally, and irregular forms ('irregulars') whose anus is displaced away from the aboral plates into the posterior interambulacral zone and that live predominantly infaunally (Figure 1). All echinoids have a well-developed system of hydraulically operated tube-feet that are important for respiration, feeding and locomotion. Echinoids also have a dense and often highly differentiated covering of calcitic spines used for locomotion and defence.

Morphology of Post-Palaeozoic Echinoids

Echinoids range in size from just a few millimetres in diameter to over 350 mm and come in a variety of body forms: globular, heart-shaped, cylindrical, hemispherical or even flattened discoidal (Figures 1 and 2). However, irrespective of their overall shape, the skeleton (also known as the 'test') in post-Palaeozoic echinoids is always constructed along the same standardized plan. At the apex of the test is a small cluster of up to 10 plates known as the apical disc (Figure 1A). In regular echinoids this series of plates surrounds the

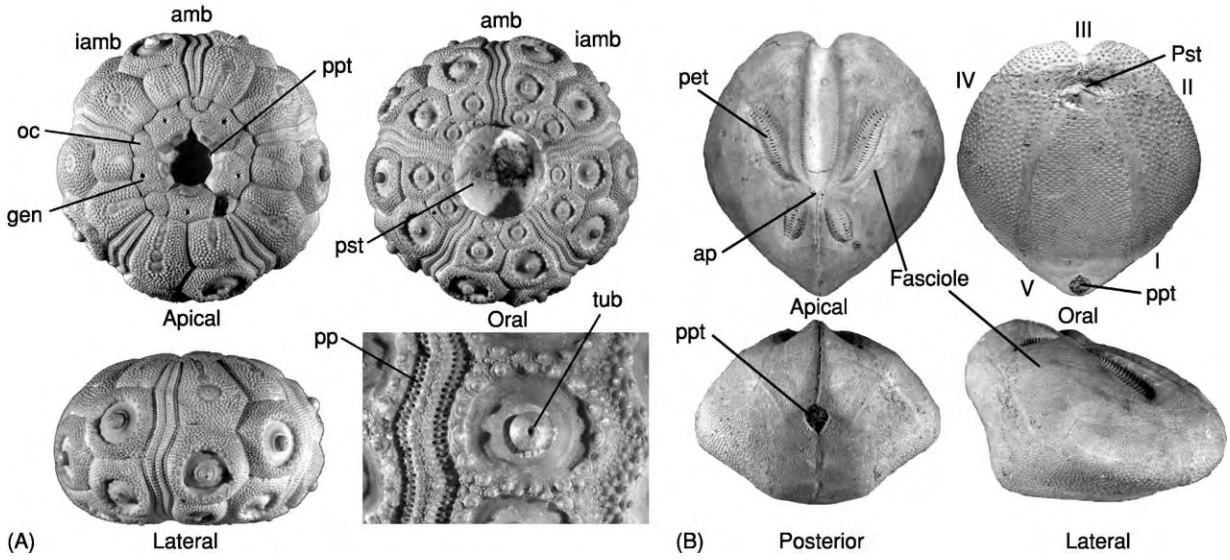


Figure 1 Morphology of (A) a regular echinoid *Stereocidaris,* an Upper Cretaceous cidaroid from England, and (B) an irregular echinoid *Schizaster*, an Eocene spatangoid from Italy. amb ambulacrum; ap apical disc; gen genital plate; iamb interambulacrum; oc ocular plate; pet petal; pp pore pair; ppt periproct; pst peristome; tub tubercle. I V numbered ambulacral zones.

anus but in irregular echinoids the anus is usually outside and to the posterior (Figure 1B). The larger genital plates are in part derived from larval skeleton and are all that remains of the aboral plated surface that is so well developed in other echinoderms. Genital plates are each perforated by a gonopore, which acts as the outlet for the gonads. One of the five genital plates is also perforated like a pepper pot, which forms the opening to the water vascular system. The smaller plates are also perforate, with their tiny opening marking where the tip of the radial water vessel reaches the exterior: these are the ocular or terminal plates. Plates are added at the distal edge of ocular plates during growth.

The remainder of the test is composed of 20 columns of plates arranged into alternate pairs of ambulacral and interambulacral elements (Figure 1). Ambulacral plates overlie the radial water vessel and are pierced by single or double pores where tube-feet connect to the interior. Interambulacral plates by contrast are generally larger and are imperforate. They arise on either side of the ocular plate. The external surfaces of both ambulacral and interambulacral plates are covered in tubercles of various sizes and these are the articulation points for spines. Regular echinoids have only a small number of large tubercles whereas irregular echinoids typically have uniform, fine, dense tuberculation (Figure 1A, B). Ambulacral and interambulacral columns of plates converge around the mouth, which lies at the opposite pole from the apical disc.

The mouth is situated in a large opening termed the peristome. This is covered in life by a flexible membrane with embedded plates that is rarely fossilized. All regular echinoids and many irregular echinoids have an internal jaw apparatus that largely fills the peristome. The lantern is a pentaradially symmetrical and highly complex in structure (Figure 3), being composed of 50 skeletal elements and 60 individual muscles. There are five teeth, each braced in a hemipyramid, and the entire structure can open and close like a grab, as well as moving in and out of the test. Because echinoids have a rigid test of fixed internal volume, having a large and active lantern that moves in and out creates problems. To compensate for changes in internal volume there are 10 expandable soft-tissue sacs around the edge of the peristome that connect directly to the interior and that accommodate any displaced body cavity fluids. Each sac passes to the exterior at a small notch in the peristome (buccal notch). Echinoids that either lack a lantern or have an entirely internal lantern lack buccal notches.

The primary muscles that move the lantern attach to the interior of the test around the rim of the peristome. Enlarged skeletal flanges mark the site of muscle attachment (Figure 3). Flanges developed from the adoral ambulacral plates are termed auricles while those developed from interambulacral plates are termed apophyses. These lantern muscle attachments together form the perignathic girdle.

Appendages that attach to the test include both spines and pedicellariae. Spines vary tremendously in size and shape according to their function (Figure 4). They can be either hollow or solid and either smooth or ornamented externally. There may be a central ligament binding the spine to its articulation ball,

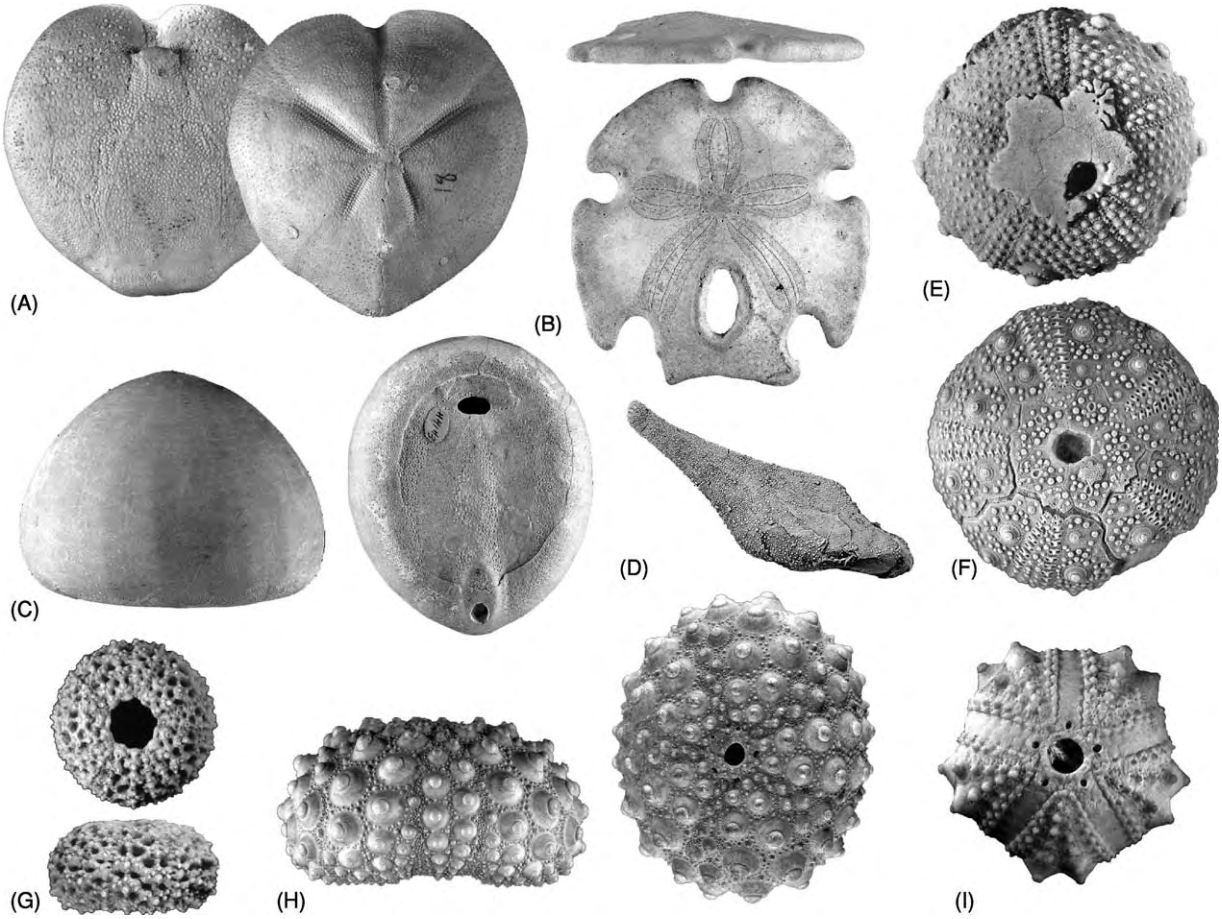


Figure 2 Representative fossil post Palaeozoic echinoids. (A) the spatangoid *Micraster* in oral and apical views, from the Upper Cretaceous of England (×0.7). (B) the clypeasteroid *Encope* in lateral and apical views, from the Pliocene of the U.S.A. (×0.4). (C) the holasteroid *Echinocorys* in lateral and oral views, from the Upper Cretaceous of England (×0.4). (D) the holasteroid *Hagenowia* in lateral view from the Upper Cretaceous of England (×2). (E) the salenioid *Novosalenia* in apical view, from the Cretaceous of the Czech Republic (×2). (F) the pedinoid *Hemipedina* in apical view, from the Lower Jurassic of England (×3). (G) the temnopleuroid *Viaudechinus* in apical and lateral views, from the Miocene of France (×2). (H) the echinoid *Heterocentrotus* in lateral and apical views from the Recent of the Indo West Pacific (×0.7). (I) the arbacioid *Murravechinus* in apical view from the Miocene of Australia (×1.5).

in which case the associated tubercle is perforate (Figure 1A) and the surrounding platform may be smooth or crenulated. Greatest functional differentiation is found in irregular echinoids where spines perform a variety of different roles. Pedicellariae are microscopic stalked, jawed appendages that resemble tulips in shape and that are used to deter small ectoparasites. They evolved from clusters of spines and are present in all echinoids from the Silurian onwards. There are many different forms of pedicellaria making them very useful species-level indicators but, unfortunately, they are rarely preserved in fossils.

Morphology of Palaeozoic Echinoids

Although Palaeozoic echinoids mostly retain a pentaradiate pattern of alternating ambulacral and interambulacral zones, the precise way their test

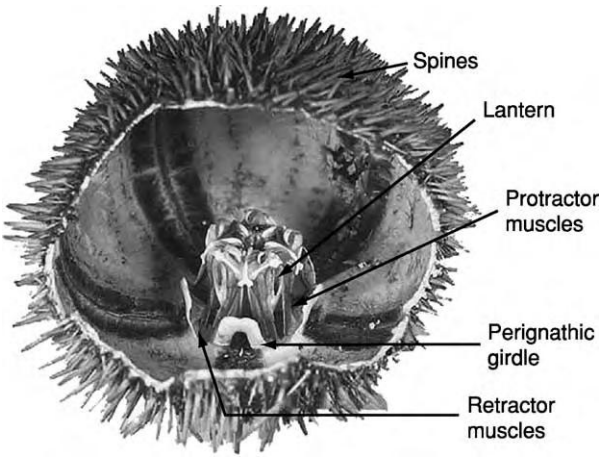


Figure 3 Test of a modern echinoid (*Echinus*) broken open to show the internal lantern and perignathic girdle.

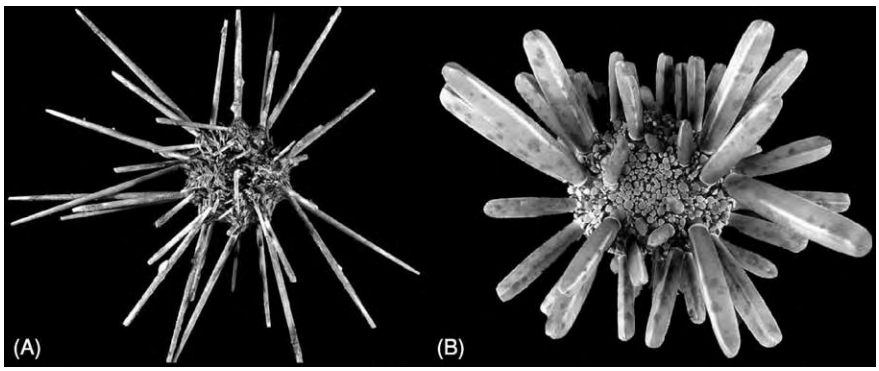


Figure 4 Modern echinoids with their complement of spines attached. (A) the cidaroid *Cidaris*; (B) the echinoid *Heterocentrotus*.

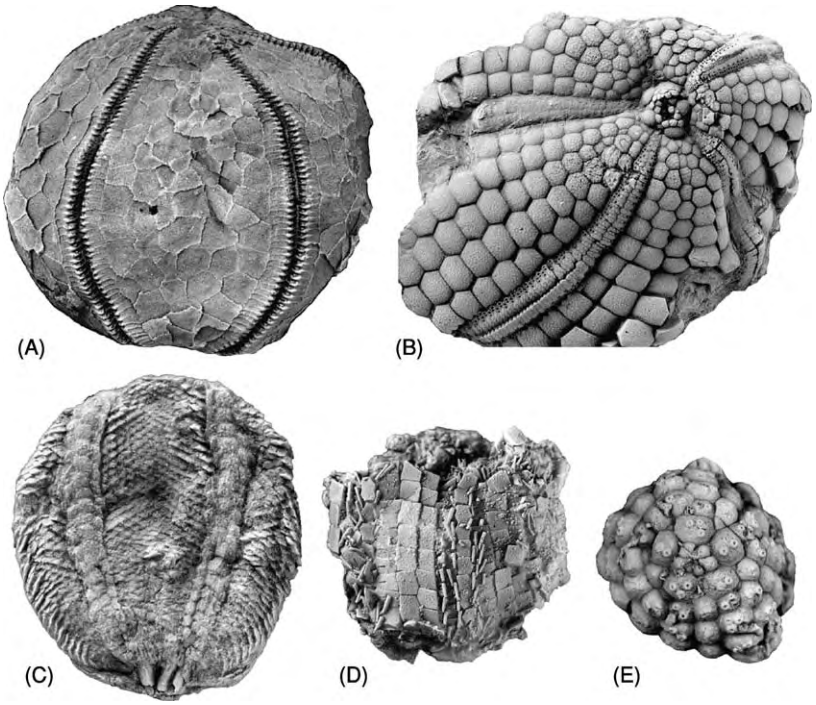


Figure 5 Representative fossil Palaeozoic echinoids. (A) *Aulechinus* (internal mould) from the Upper Ordovician of Scotland (×1). (B) *Maccoya* from the Carboniferous of England (×0.5). (C) *Lepidesthes* from the Carboniferous of the USA (×1). (D) *Bromidechinus* from the Upper Ordovician of the USA (×1). (E) *Bothriocidaris* from the Upper Ordovician of Estonia (×3).

is constructed shows a huge amount of variation (Figure 5). Only the Permian genus *Miocidaris* has a test constructed of 10 ambulacral and 10 interambulacral columns like that of post-Palaeozoic echinoids. In all others the number of columns of plates, both ambulacral and interambulacral, varies tremendously. Some, such as *Palaechinus* and *Aulechinus* (Figure 5A) have biserial ambulacra and multiple columns of non-ambulacral plates. Others, such as *Proterocidaris* and *Lepidesthes* (Figure 5C), have very wide ambulacra of up to 20 columns of plates, and narrow interambulacra. *Bothriocidaris* (Figure 5E) has no interambulacral plates at all, but there is an imperforate series of ambulacral plates. Primitive echinoids also, interestingly, have the radial water vessel enclosed within the ambulacral plates rather than being fully internal, as it is in all post-Palaeozoic forms. A lantern is present, but the teeth are very primitive in structure and no Palaeozoic echinoid has a perignathic girdle. Finally, plating is imbricate (Figure 5D) and so the test rapidly fell apart at death so that even partially articulated specimens are rare except in Lagerstätte deposits.

Echinoid Palaeobiology

Locomotion and Burrowing

Regular echinoids are vagile benthos, moving by means of their oral spines over the sea floor and using their tube-feet to climb and grip hard substrata. In forms living in high-energy, rocky, shore environments, the oral tube-feet that provide grip are highly muscular and also densely packed, an adaptation that is easily recognizable from fossil tests. Only a few regular echinoids, notably cidaroids and arbacioids, evolved specialized aboral respiratory tube-feet: for the most part tube-feet are undifferentiated. Spines would have been the primary means of locomotion in Palaeozoic echinoids, although some of the most primitive presumably used their tube-feet.

Irregular echinoids rely entirely on their spines for locomotion, and most burrow into unconsolidated sediment. This raises a number of problems: locomotion is harder, and circulation of oxygenated water and removal of waste products more difficult, particularly in finer-grained, more impermeable sediments. Infaunal echinoids have highly specialized aboral tube-feet for gaseous exchange and these regions form characteristic petals on the surface of the test (**Figure 1B**). However, only heart urchins have successfully adapted for life in fine-grained sediments. In order to do this they have specialized aboral tube-feet that are used to build and maintain an open shaft to the surface down which fresh water can be drawn. Water is drawn into the burrow by bands of highly ciliated spines, termed fascioles. These fascioles also generate a mucous coat that is held over the tips of the spines and prevents fine-grained sediment from falling between the spines. The mucous coat thus allows heart urchins to maintain a water-filled space within the burrow.

The traces of burrowing echinoids have been described in rocks from the late Lower Cretaceous onwards.

Feeding

Regular echinoids feed using their lantern. Palaeozoic echinoids had relatively weak lantern musculature and feeding must have largely involved scooping material from the sea floor. With stronger lanterns echinoids were able to utilize new food sources and could rasp encrusting organisms from hard substrata. Echinoid tooth bite marks first appear as trace fossils in the late Triassic, at about the same time that well-developed perignathic girdles and stronger lanterns evolved. During the Mesozoic the biting force that lanterns could exert was further increased through the development of teeth with a stronger T-shaped cross-section, and the stronger and more extensive bracing of hemipyramids. Mesozoic and Tertiary echinoids were mostly either algivores or predatory on sessile organisms.

Deposit feeding may have evolved in the Carboniferous, when echinoids such as *Proterocidaris* evolved large numbers of massive oral tube-feet whose most likely function was in detritus gathering. However, this clade became extinct towards the end of the Permian and deposit feeding was not adopted again by echinoids until the Early Jurassic. The evolution of irregular echinoids in the Jurassic is really the story of adaptation towards improved deposit feeding. Some irregular echinoids (cassiduloids) are bulk sediment swallowers, passing a continuous stream of sediment through the gut in order to extract the small quantities of organic mater it might contain. Others (spatangoids and holasteroids) evolved as selective deposit feeders using specialised tube-feet to pick out organic-rich detritus. Sand dollars are rather special in that they appear to be using their aboral spine canopy as a sort of sieve to separate fine organic material from amongst the sand grains. The finer particles fall between the spines and are swept towards the mouth and entrapped in mucous strings that run towards the mouth. These food grooves are easily visible even on a denuded test.

Predation and Defence

Echinoids today are preyed upon by a variety of animals (including humans), and predation has, presumably, been a major factor since the Devonian, when the first echinoids with long, lance-like spines evolved. Long, sharp spines (**Figure 4A**) are effective deterrents to many predators, but alternative strategies have also been evolved. Cidaroids have spines that are not covered in a living tissue, and that, consequently, can become heavily encrusted with algae and sessile organisms. This encrustation on flattened, paddle-shaped spines provides very effective camouflage. An alternative ploy has been to develop very thick club-shaped spines that cannot easily be snapped or broken off (**Figure 4B**). A few regular echinoids have abandoned spines in favour of poisonous pedicellariae as their first line of defence. *Toxopneustes* has a very short, uniform coating of spines and their highly venomous pedicellariae are almost as long. The aboral surface of this animal thus presents a formidable battery of dense, poisonous pedicellarial jaws.

Spines can also provide a first level of protection against impact in high-energy environments, and both in the Jurassic and in the Tertiary echinoids evolved that were short, stout and umbrella-like, forming a tessellate pavement over the surface of the test.

Reproduction

In echinoids the sexes are separate, but it is usually impossible to tell the sexes apart visually. Reproduction occurs by broadcast spawning: eggs and sperm are released into the water column where fertilization takes place. The fertilized egg then develops into a planktonic larva termed pluteus, which feeds in surface waters for a few weeks before metamorphosing into the adult form and settling to the bottom. A small number of species, most from high-latitude, do not pass through a planktonic larval stage, but develop directly. In such cases the female produces large yolk-rich eggs that are brooded on the test. Females then have special brood pouches termed marsupia where the eggs are protected, and these form sunken zones around the gonopore openings. Sexual dimorphism first evolved in the Late Cretaceous and was particularly prevalent in Australian faunas in the Early Tertiary.

Geological History

The closest living relative of the echinoids are holothurians (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids)) and their ancestry presumably lies amongst the 'asterozoan' taxa of the Early Ordovician. When echinoids first appear in the Upper Ordovician, there were already three very different body forms. Bothriocidarids have a small globular test composed of thick plates of the ambulacral series only (Figure 5E), lepidocentrids have narrow biserial ambulacra and wide zones of irregularly arranged imbricate interambulacral plates (Figure 5A), and *Bromidechinus* has biserial interambulacra and triserial ambulacra (Figure 5D). Clearly the earliest phase in the history of echinoids has yet to be uncovered.

During the Lower Palaeozoic, echinoids remained a minor component of marine benthic communties, never becoming diverse or abundant. By the Devonian one lineage had developed enlarged adoral tube-feet and appears to have specialized as a deposit feeder, and this group thrived through to the Permian before going extinct. In the Carboniferous another important group, the archaeocidarids, made their appearance. Archaeocidarids have a single large tubercle on each interambulacral plate and possessed long highly muscular spines. On functional grounds they were probably the first active predators. *Miocidaris*, the only Palaeozoic echinoid with the test architecture of post-Palaeozoic forms, evolved directly from *Archaeocidaris* by reduction of plating columns in each interambulacral zone.

The modern crown group was initiated and had already begun to diversify before the end of the Permian, with both a cidaroid and a non-cidaroid lineage passing into the Triassic. The cidaroids (Cidaroidea) developed rigid test plating in the Triassic and have been an important constituent of deeper-water, shelf communities ever since. Despite attaining a modest diversity, cidaroids have always remained very conservative in their morphology.

Small non-cidaroid forms flourished in reefal settings in the Late Triassic, but it was not until the early Jurassic that the other lineage (Euechinoidea) started to diversify in a major way. Irregular echinoids make their first appearance in the Lower Jurassic and rapidly evolved as deposit feeders. The earliest irregular echinoids still possessed a lantern like that of regular echinoids, but by the Middle Jurassic both cassiduloids and spatangoids had evolved. Cassiduloids specialized as bulk sediment feeders using dense tube-feet for handling sediment particles. Sand dollars arose in the Early Tertiary from cassiduloid ancestors and specialized for life in the near-shore. Unique amongst echinoids, they developed large numbers of microscopic tube-feet to each ambulacral plate and adopted a very singular method of sediment sifting to obtain their food. Spatangoids and holasteroids on the other hand evolved a special kind of tube-foot around the mouth for selective deposit feeding. Modern heart urchins first appear at the base of the Cretaceous and have diversified more or less constantly since then. The end-Cretaceous extinction (*see* **Mesozoic:** End Cretaceous Extinctions) affected echinoids selectively, with deposit feeders being hardest hit. Regular echinoids also thrived throughout the Mesozoic, although today's major modern groups did not become established until the Late Cretaceous.

Phylogeny and Classification

Major groups and their characteristic features are listed below and their geological ranges are shown in Figure 6:

- Cidaroida (Figure 1A) (Permian – Recent): regular echinoids with simple ambulacral plating and interambulacra with a single large primary tubercle; teeth U-shaped in cross-section.
- Echinothurioida (Jurassic – Recent): regular echinoids with hollow spines and an imbricate plated test; tubercles perforate and non-crenulated; teeth U-shaped in cross-section.
- Diadematoida (Jurassic – Recent): regular echinoids with rigid, thin-plated test, compound ambulacra and perforate and crenulated tuberculation; teeth U-shaped in cross-section.
- Pedinoida (Figure 2F) (Triassic – Recent): regular echinoids with perforate non-crenulate tubercles,

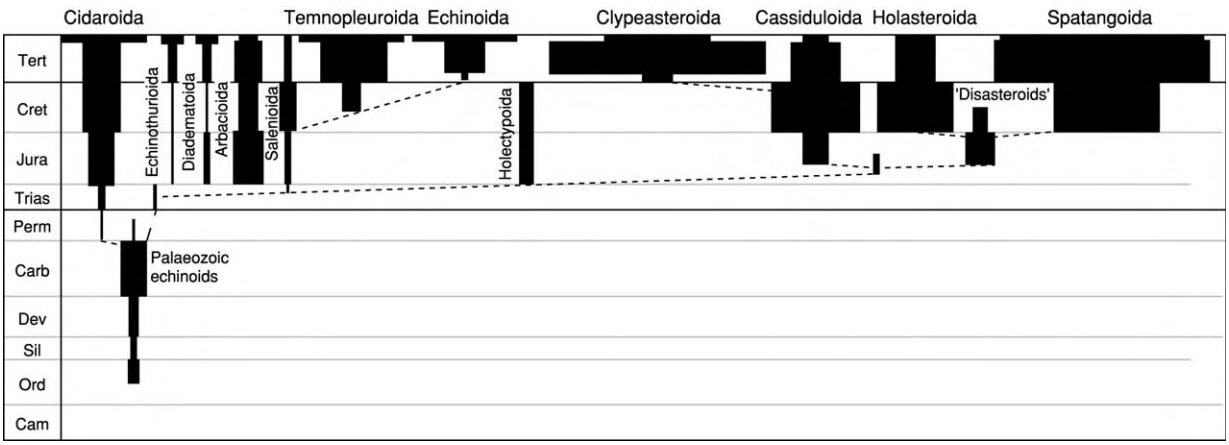


Figure 6 Stratigraphic range charts for the major groups of echinoid. Width of bars is proportional to taxic (genus level) diversity. The dashed lines show their phylogenetic relationships.

compound ambulacral plating and teeth U-shaped in cross-section.

- Arbacioida (Figure 2I) (Jurassic – Recent): regular echinoids with imperforate, non-crenulate tubercles, compound ambulacral plating and teeth T-shaped in cross-section.
- Salenioida (Figure 2E) (Jurassic – Recent): regular echinoids with imperforate tubercles, compound ambulacral plating and teeth T-shaped in cross-section. The apical disc is large and cap-like and has extra plates incorporated into it.
- Temnopleuroida (Figure 2G) (Upper Cretaceous – Recent): regular echinoids with compound ambulacral plating, imperforate tuberculation that is either crenulated or non-crenulate, and teeth T-shaped in cross-section. The test surface is often highly ornamented with pits.
- Echinoida (Figure 2H) (Upper Cretaceous – Recent): regular echinoids with compound ambulacral plating, imperforate, non-crenulate tuberculation and teeth T-shaped in cross-section.
- Holectypoida (Jurassic – Cretaceous): irregular echinoids with a large functional lantern, no petals, and peristome with buccal notches. The periproct is large and tuberculation relatively coarse.
- Cassiduloida (Jurassic – Recent): irregular echinoids with no lantern and small peristome; tube-feet concentrated around the mouth to form phyllodes; aboral ambulacra petaloid; spines fine and dense.
- Clypeasteroida (Figure 2B) (Palaeocene – Recent): irregular echinoids, often flattened and disc-like, characterized by having petals, a fully internal lantern and small peristome and very many microscopic tube-feet to each ambulacral plate.
- Holasteroida (Figure 2C, D) (Cretaceous – Recent): ovate to heart-shaped irregulars, but may be more elongate; no lantern; apical disc elongate with ocular plates II and IV meeting behind anterior genital plates.
- Spatangoida (Figures 1B, 2A) (Cretaceous – Recent): ovate to heart-shaped irregulars; no lantern; apical disc compact.

See Also

Fossil Invertebrates: Echinoderms (Other Than Echinoids); Crinoids. **Lagerstätten**. **Mesozoic:** End Cretaceous Extinctions. **Sedimentary Environments:** Reefs ('Build-Ups'). **Trace Fossils**.

Further Reading

Kier PM (1965) Evolutionary trends in Paleozoic echinoids. *Journal of Paleontology* 39: 43 465.

Kier PM (1974) Evolutionary trends and their functional significance in the post Paleozoic echinoids. *Journal of Paleontology* 48(supplement 5): 1 95.

Lewis DN and Donovan SK (1998) Fossils explained 23: Palaeozoic echinoids. *Geology Today November/December*: 235 240.

Smith AB (1984) *Echinoid Palaeobiology*, p. 199. London: George Allen & Unwin.

Smith AB and Savill JJ (2002) *Bromidechinus*, a new Middle Ordovician Echinozoa (Echinodermata), and its bearing on the early history of echinoids. *Transactions of the Royal Society of Edinburgh* 91: 137 147.

Smith AB, Littlewood DTJ, and Wray GA (1996) Compara tive evolution of larval and adult life history stages and small subunit ribosomal RNA amongst post Palaeozoic echinoids. In: Harvey PH, Leigh Brown AJ, Smith JM, and Nee S (eds.) *New Uses for New Phylogenies*, pp. 234 254. Oxford: Oxford University Press.

Smith AB (2004) *The Echinoid Directory*. World Wide Web electronic publication.www.nhm.ac.uk.

Graptolites

R B Rickards, University of Cambridge, Cambridge, UK

Introduction

The graptolites were a group of colonial animals that inhabited the Palaeozoic seas, often in very great abundance. Their remains can be found as small elongate (3 mm–1 m) skeletons of thin collagenous periderm forming simple linear or complexly branched colonies. Each branch is composed of a linear succession of interconnecting tubes (thecae) originating from a single conical individual called a sicula (Figure 1). Graptolites are of value in the study of the stratigraphy of many Ordovician and Silurian deposits the world over, because of their abundance, wide geographical distribution, and rapid evolution. Benthic graptolites appeared in the Middle Cambrian, becoming extinct in the Upper Carboniferous; planktonic graptolites ranged from the earliest Ordovician to the Middle Devonian.

The Chequered History of Graptolite Studies

Although graptolites were originally observed and illustrated as long ago as the late eighteenth century, it was not until the mid-nineteenth century that much attention was paid to them as fossils and their animal nature was fully appreciated. Even then there was considerable doubt about their zoological affinities, and they were variously allied to the plants, molluscs, corals, protozoa, and bryozoans. Modern interpretation of their morphology places them in the Hemichordata based on comparisons with living forms. Thus, their zoological position has been elevated over the years from plants, through the lower invertebrates, to the lower vertebrates.

Understanding of their morphology has progressed greatly, aided by technical developments in microscopy, especially since the late 1950s with the use of scanning electron microscopes. Consequently, there is now a tremendous amount of detailed information available on this extinct group of strange little animals that might at first appear to be, as their name (graptos, written or marked; lithos, stone) implies, little more than markings on rock.

Classification, Broad Evolution, Stratigraphy, and Mode of Life

The ancestors *Eorhabdopleura* and *Eocephalodiscus* of the living Hemichordata probably arose with the order Dendroidea from a common ancestor in the early Middle Cambrian (Figure 2). The common ancestor may have been a phoronid-like worm that evolved through one or more of the following stages of evolution: infaunal, in dense association; infaunal with loose coloniality; and finally sessile with several stolons inside one upright peridermal tube. This last stage would have been equivalent to a primitive *Mastigograptus*-like graptolite. The evolutionary relationships of the orders other than dendroids and graptoloids are unknown, and the stratigraphic record of these other orders is poor. The evolution within the Order Graptoloidea is discussed in further detail later in this article, and the distinctions between dendroids and graptoloids are discussed below.

The Sicula (or Larval) Stage

Eggs and embryos are known but are extremely rare in graptolites. However, the peridermal sheath of the larval stage, called the sicula, is known in almost all graptolites and is very common in the rock record. Eggs, embryos, and larval stages were produced by sexual reproduction, but the remainder of the colony grew by asexual budding from the sicula. In dendroid siculae (Figure 3A) a resorption foramen (or hole) formed in the prosicula, and from this grew the first asexual bud of the colony. In most graptoloids (Figure 3B) the resorption foramina were in the metasicula, but in some, the Monograptina (Figure 3C), the growth lines construct a notch, then a hole (lacuna), through which the first theca grew. The dendroid sicula shown here (Figure 3A) is of a benthonic type: a planktonic type would have a nema instead of a basal disc and may have longitudinal threads.

The Stolon System

In dendroids and tuboids the thecal tubes are attached at their bases to a black (sclerotized) stolon system; in dendroids (Figure 4) each node divides into three stolons, but in tuboids (Figure 5) the division is always into two stolons and some species have only partly sclerotized stolons (i.e. only partly hardened). In graptoloids (Figure 6) there are no sclerotized stolons,

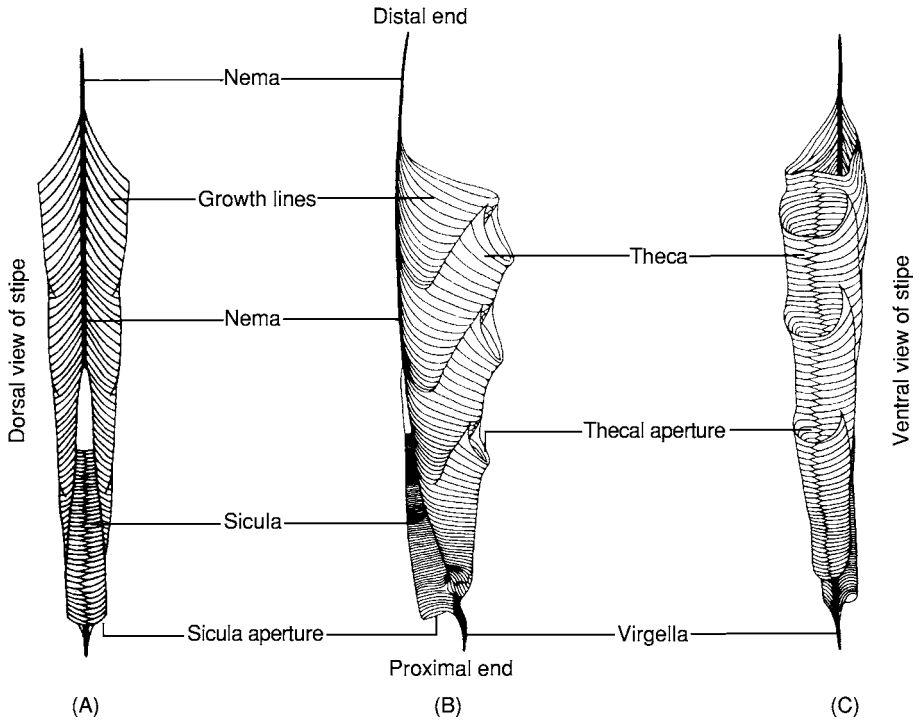


Figure 1 (A) Dorsal, (B) lateral and (C) ventral views of a graptolite stipe, showing the arrangement of the thecae.

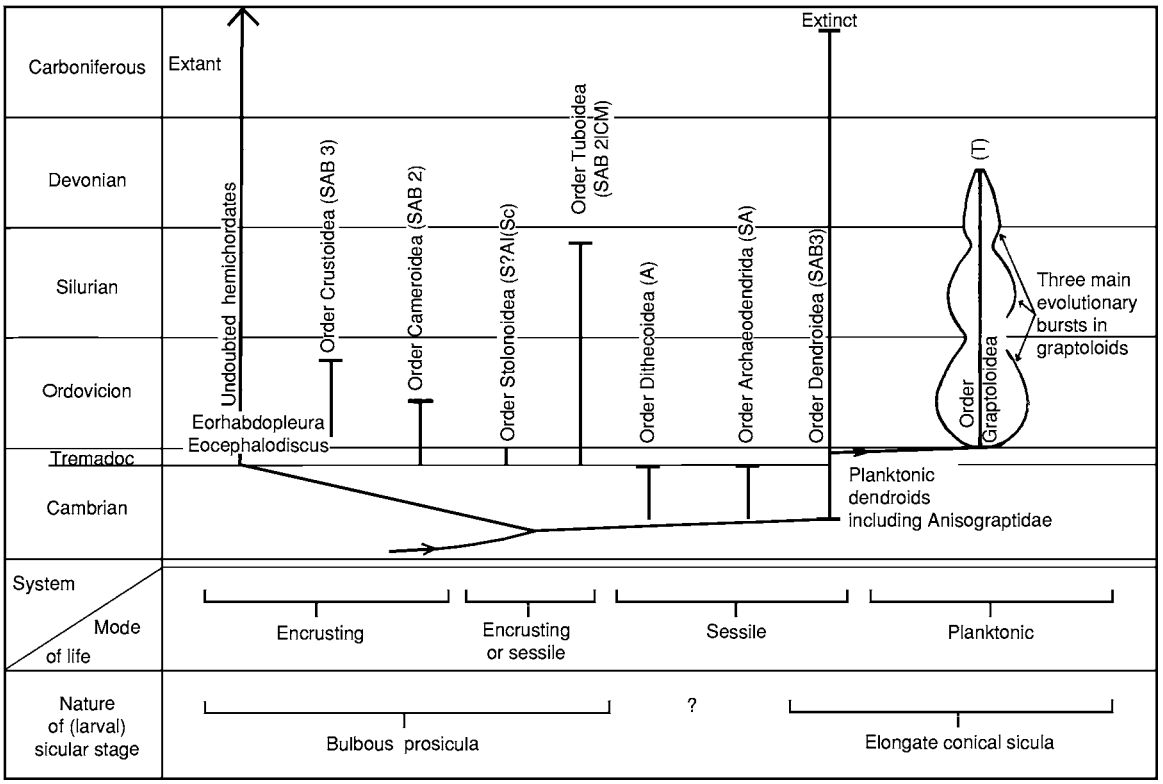


Figure 2 The evolutionary relationships of the graptolites to other early Hemichordata. The letters in parentheses following the ordinal name indicate briefly the defining characters of the order. Abbreviations are as follows: S, stolotheca; A, autotheca; B, bitheca; 2 and 3, diad and triad divisions of the stolon system, respectively; I, irregular stolons; C, conotheca; M, microthecae; I(Sc), irregular stolons, which may be partially sclerotized; T, theca.

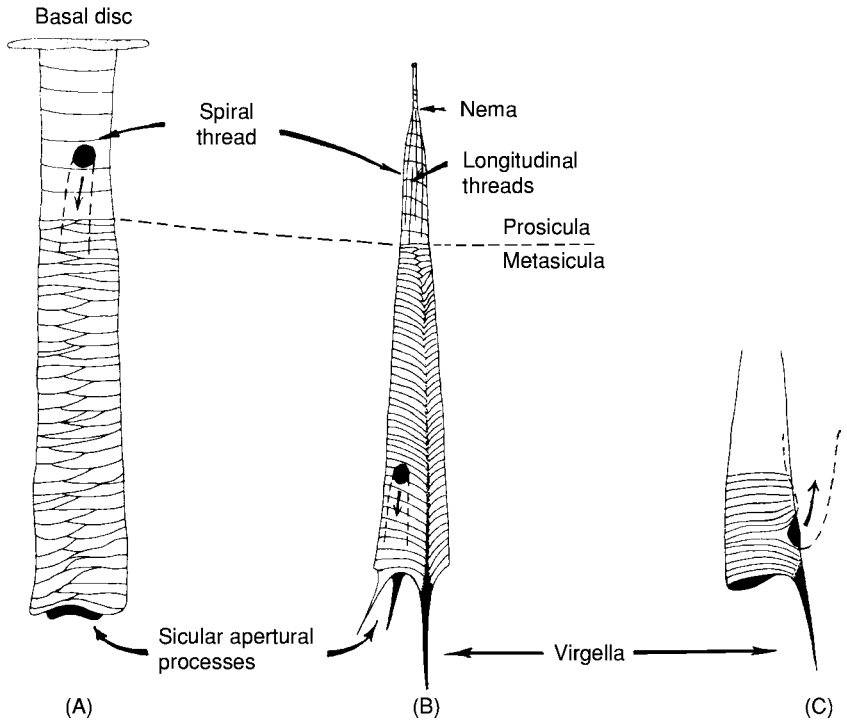


Figure 3 The sicula or larval stage, showing the resorption foramen, from which the first theca forms.

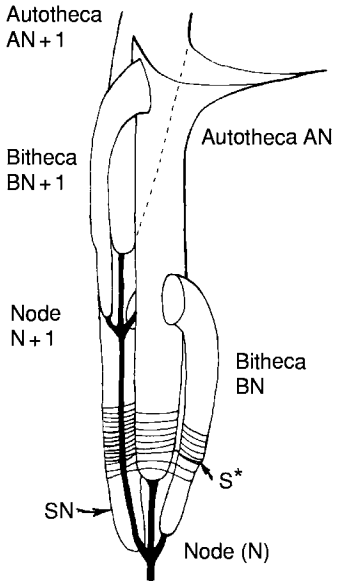


Figure 4 Schematic illustration of a dendroid, showing each node dividing into three stolons. The autotheca AN 1 is omitted for simplicity; bitheca BN would open into it, as bitheca BN + 1 opens into autotheca AN. S^* marks the distal limit of growth lines that completely encompass tubes AN, BN, and SN. Distal to this, the tubes SN and BN have their own independent fuselli. Proximal to S^* it is presumably the autothecal zooid in AN that secreted the fusellar periderm, because the growth lines below S^* are continu ous with those of AN but have an uncomfortable relationship with those of SN and BN. All tubes, whether developed at one node or at an earlier node, grow together, so that a growing end of the stipe just below the aperture of AN would have developing tubes, and hence zooids, of AN, BN + 1, AN + 1, and SN + 1.

and the individuals of the colony were probably connected by unsclerotized homologues of stolons.

The Structure of the Periderm

A modification of Kozlowski's classic interpretation is shown in Figure 7. The graptolite periderm is composed of two layers (Figure 8) – an inner fusellar layer constructed of fuselli in the form of half rings with a double zig-zag contact and an outer cortical layer comprising a bandage-like structure laid down by the zooids (Figure 9) after the tube has been built of fuselli. The bandages become more thinly layered towards the thecal apertures, at least during the early stages of thecal development. A number of graptolites have an inner, thinner, layer of cortex, which is also bandaged (Figure 10B–G, I, J).

Ultrastructure

In recent years the use of electron microscopes has increased our knowledge of detailed graptolite structure. Each fusellus has an envelope of largely granular sheet fabric within which is a sponge-like mesh of anastomosing fibrils (probably related to collagen) termed 'fusellar fabric'. Each cortical bandage, seen in transverse section in Figure 8, is also bounded by sheet fabric and comprises bundles of fibrils that are certainly fossil collagen. Within the sheet fabric are vesicular bodies, which may have been gas or

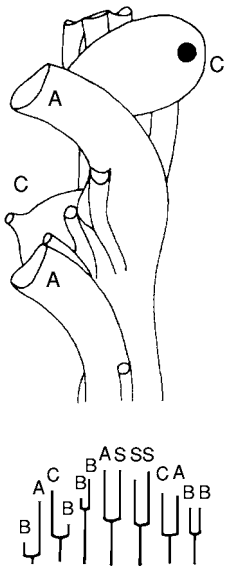


Figure 5 Schematic illustration of a tuboid, showing the div ision into two stolons and the tangle of tubes typical of some tuboids. The lower panel illustrates the diad stolon divisions that could give rise to such stripe complexity: A, autothecal stolon; B, bithecal stolon; S, stolothecal stolon; C, conothecal stolon. Note the great irregularity of stolon length compared with that of the dendroids (**Figure 4**).

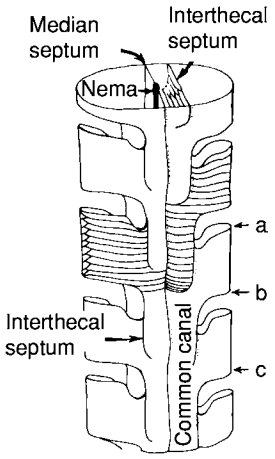


Figure 6 Schematic illustration of the graptoloid *Climacograp tus*, showing the individuals of the colony connected by unsclerotized homologues of stolons. The thecal overlap is ab/ac. The distal end of the stipe is cut to show the internal structure, and growth lines are shown on two of the thecae.

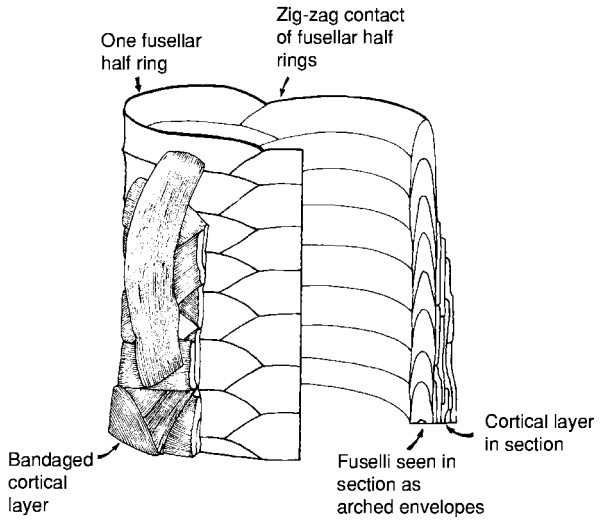


Figure 7 A modification of Kozlowski's classic interpretation of the structure of the periderm. Part of the thecal tube is cut away to show both the internal arch and overlap structure of the fuselli and the piles of bandages. The fibrous structure of the bandages is diagrammatic.

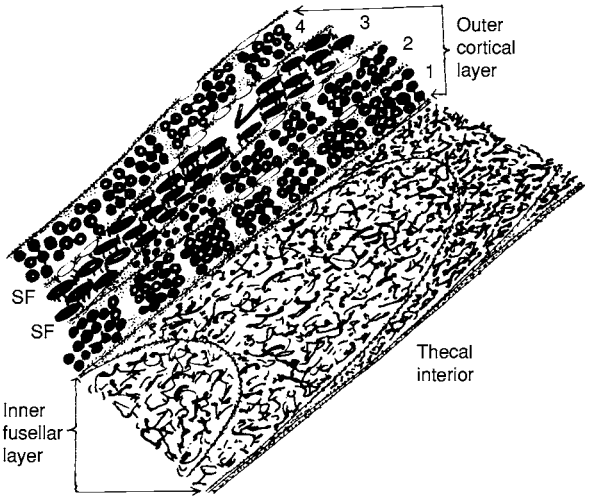


Figure 8 The graptolite periderm, showing the two layers. The minute fibrils in the outer cortical layer are parallel to the lengths of the bandages and there is a characteristic bundling of fibrils within each layer. Layer 2 shows the thinly tapering edge of one bandage. SF, sheet fabric; V, vesicular body; the magnification is about ×13 000.

fat bodies. The collagen fibrils are parallel to the long edges of the bandages (shown schematically in **Figure 9**) and in **Figure 8** are shown cut transversely in layers (or bandages) 1, 2, and 4, and obliquely in layer 3. The oblique section of layer 3 shows the typical collagen annulations, seen here as short wavy bars, which in reality are ring-like thickenings around each fibril at specific intervals.

Bandaging was almost certainly a strengthening feature and was thin or absent in forms that lived encrusted in sheltered environments, such as the insides of shells, or which needed to be light, in order to live in a planktonic environment. The graptolite zooid has only rarely been found preserved, and its structure has been deduced from living zooids known in other hemichordates. **Figure 10** summarizes this similarity in transverse section, although there

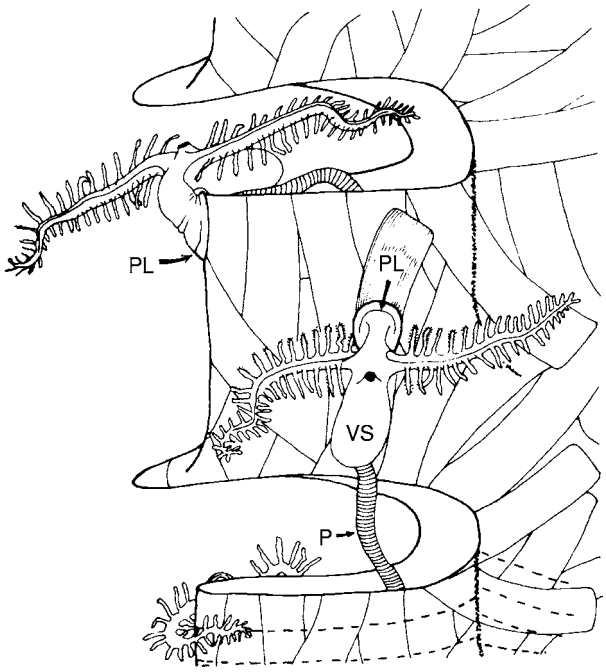


Figure 9 A zooid laying down bandages on the outside of a *Climacograptus* theca, using the pre oral lobe or disc (PL). The anus is shown as a small black foramen on top of the zooid. The mouth is underneath the visceral sack (VS), between it and the pre oral lobe. The ciliated tentacles have food gathering grooves on the upper surface, which lead down to the mouth. The zooids were probably connected by a soft pectocaulus (P). The positions of some fusellar growth lines are shown by the dashed lines, although these are not always visible if the bandaging is heavy.

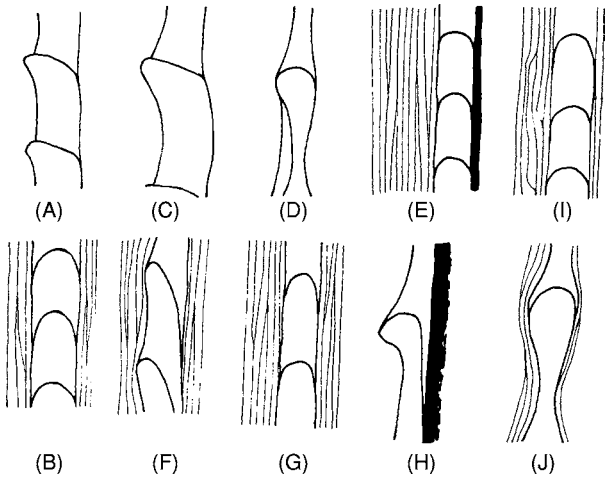


Figure 10 The similarity in transverse section between grapto lites and living hemichordates. (A, B) Distal and proximal, respect ively, fuselli and cortex of a living hemichordate. (C E) Distal, proximal, and support stem, respectively, fuselli and cortex of *Mastigograptus*, a dendroid approximately intermediate be tween graptolites and other hemichordates. (F) A tuboid (see Figure 5). (G) A dendroid. (H) A graptoblast (enigmatic encysted graptolite). (I) *Diplograptus*, showing scarped edges to the bandages. (J) *Monograptus*, showing a relatively thin cortex. Fusellar like fabric is shown in white and cortical fabric with a stipple.

are many other similarities between graptolites and living hemichordates.

Rhabdosomes and Synrhabdosomes

A complete rhabdosome of a *Monograptus* species (Figure 11A) consists of a single scandent stipe with one series of thecae (uniserial) growing upwards from the sicula and along the nema. Occasionally graptolite rhabdosomes are found in the association shown in Figure 11B, termed synrhabdosomes, which are preserved when the nemata remained fortuitously tangled. The reasons for such associations are unknown. The nemata could have been embedded in vacuolated (gas-filled) tissue and the unit as a whole could have assisted the buoyancy of the colonies. Alternatively, the association could have been related to sexual reproduction. In any event there must have been soft tissue in the centre of the synrhabdosome.

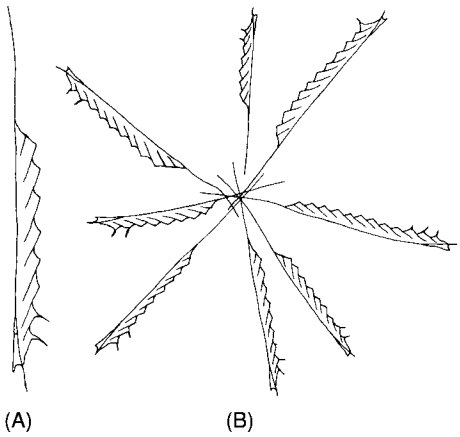


Figure 11 (A) A complete rhabdosome of a *Monograptus* species. (B) Occasionally, graptolite rhabdosomes are found in associations that are termed synrhabdosomes; these are preserved when the nemata remained fortuitously tangled.

Distinctions Between the Dendroidea and the Graptoloidea

The Dendroidea are characterized by two types of theca (autothecae and bithecae; Figure 12) linked by a black stolon system and a thick outer peridermal layer (cortex). Strengthening rods connect the numerous stipes, which each have several hundred thecae. A typical example of a dendroid is *Dictyonema* (Figures 13 and 14). In evolving from the Dendroidea, the Graptoloidea gradually lost the bithecae and stolon system, the result being colonies with one type of individual (simply termed 'thecae');

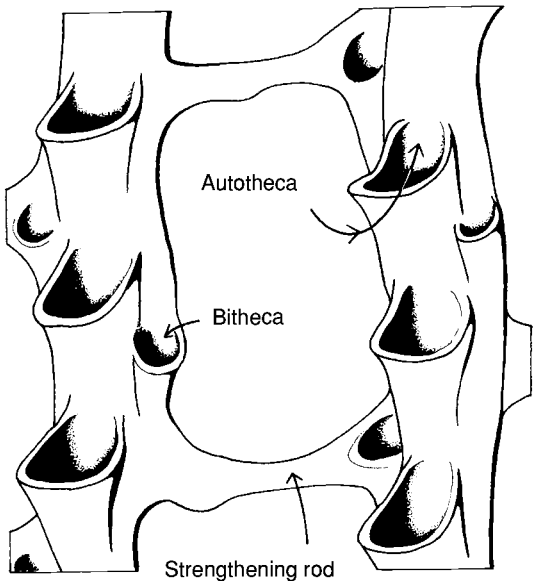


Figure 12 The structure of the Dendroidea, showing the two characteristic types of theca.

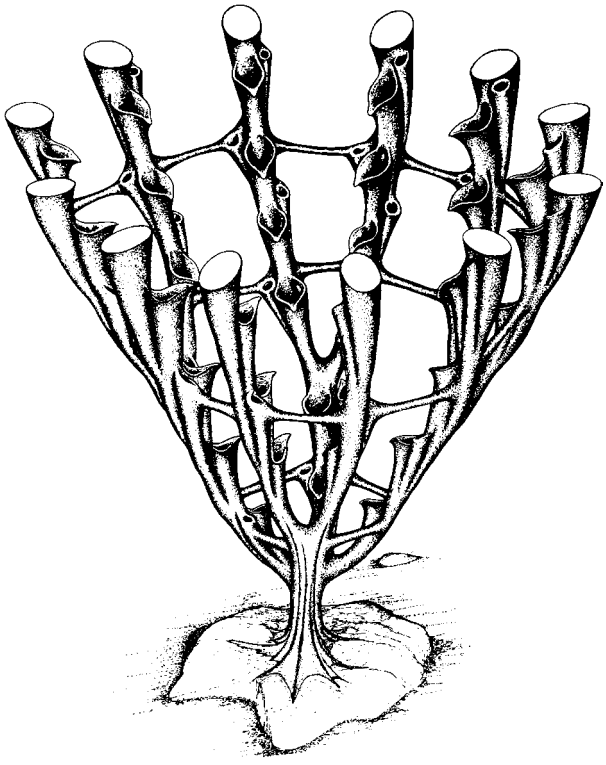


Figure 13 A typical dendroid, *Dictyonema*, which lived attached to the substrate by a basal holdfast.

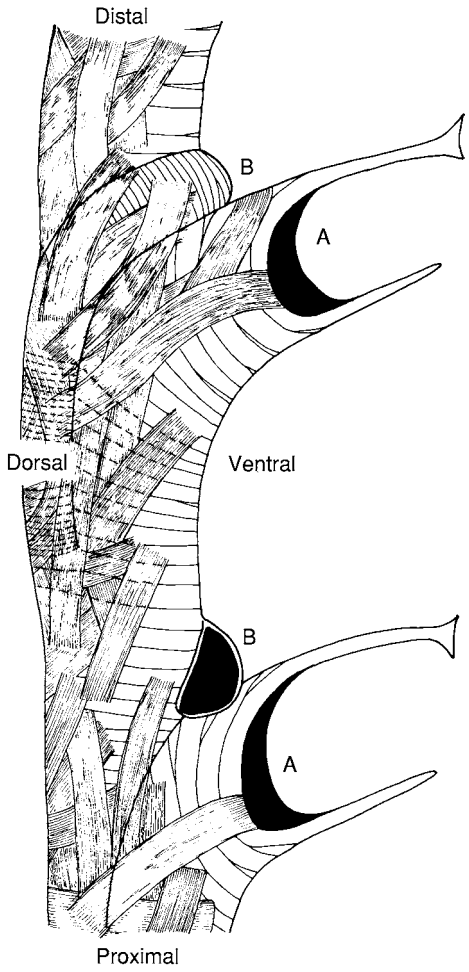


Figure 14 A section of the dendroid *Dictyonema rhinanthiforme*, showing two autothecae (A) and two bithecae (B) to illustrate the relationship of the two types of dendroid thecae. Magnification is approximately ×133. Fusellar growth lines and cortical bandages (**Figure 7**) are shown; the latter are less thickly deposited in the apertural regions of the autothecae. In most dendroids the cortical bandages are not easily seen, but the growth lines are commonly well preserved. The lower of the two bithecae grows around the far side of the stipe and opens in the angle between the two autothecae, facing the reader. The upper bitheca does exactly the opposite. The internal relationships of the bithecae and autothecae are shown in **Figure 4**. Some bandages are omitted near the ventral side of the stipe so that the fusellar growth lines can be shown. In actual specimens the growth lines can often be seen through the bandages. The positions of the fuselli are shown by dashed lines in the centre of the diagram.

the outer cortex became thinner and the colonies became lighter in weight; the stipes became fewer, but longer and more widely spread, without connecting rods; in the diplograptids the stipes came to lie back-to-back. *Rastrites* (**Figure 15**) is an extreme evolutionary product of *Monograptus* (**Figure 16**) in which the protheca is reduced to a fine thread-like tube and the metatheca is completely isolated from neighbouring thecae; there is thus no thecal overlap and no interthecal septum. The nema, virgella, and sicula are quite normal. The thecal apertures have hooks.

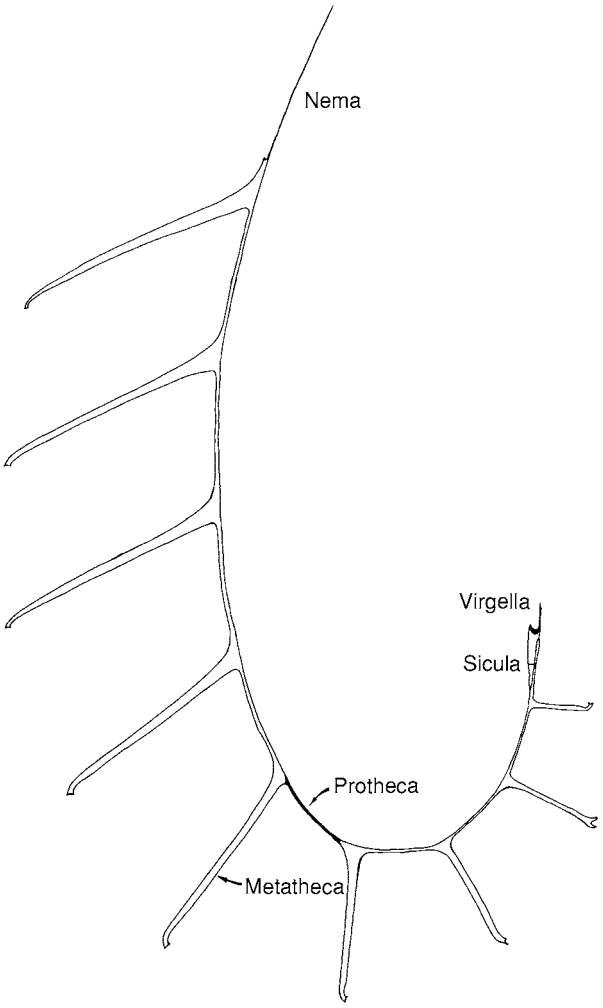


Figure 15 *Rastrites*, showing the protheca reduced to a fine thread like tube (shown in black on protheca 5) and the metatheca completely isolated from neighbouring thecae. The nema, virgella, and sicula are quite normal. The thecal apertures have hooks; theca 2 is deliberately twisted so that the two tubes at the apertures are visible; the remainder are shown in profile so that only one tube is easily visible.

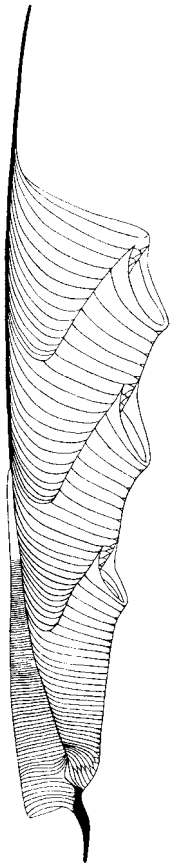


Figure 16 A typical graptoloid, *Monograptus*.

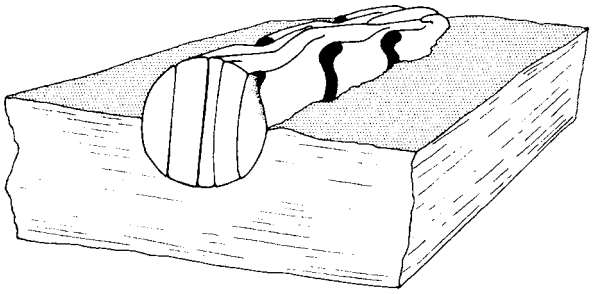


Figure 17 Preservation of a dendroid stipe in three dimensions.

Occurrence in the Rocks and Preservation

Graptolites usually occur in one of three ways in the rock record.

1. The periderm may be more or less intact and the stipes preserved in three dimensions (Figure 17). This form of preservation is most common in chemically precipitated limestones, clastic limestones, and sandstones. Dendroidea tend to be preserved most commonly in this way.
2. Much more common than (1) is the preservation of graptolites in three dimensions with an infilling of iron pyrite; the entombing rock is usually a black shale rich in organic debris. Such preservation probably occurred in oxygen-depleted conditions on the seafloor, where large amounts of decaying organic matter provided a suitable environment for bacteria that were capable of extracting iron from the seawater and precipitating it as iron pyrite.
3. Most commonly of all, graptolites may have been quickly flattened by the weight of sediment (Figure 18) such that they are now seen by the collector as films upon the bedding plane, usually white or silver on a dark background in a shale environment or as rather dark films in siltstones.

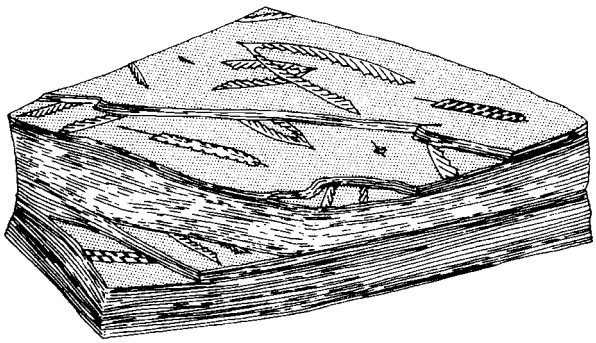


Figure 18 Graptolites flattened by the weight of the overlying sediment, to be preserved as films on the bedding plane. This is the most common mode of preservation.

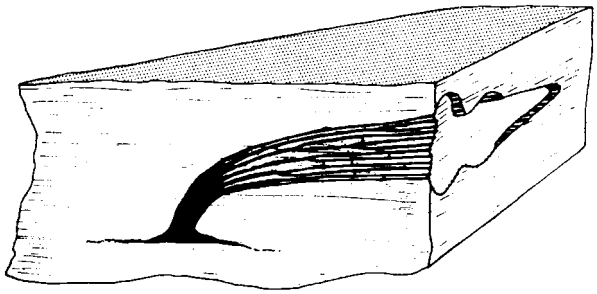


Figure 19 A dendroid colony preserved in growth position.

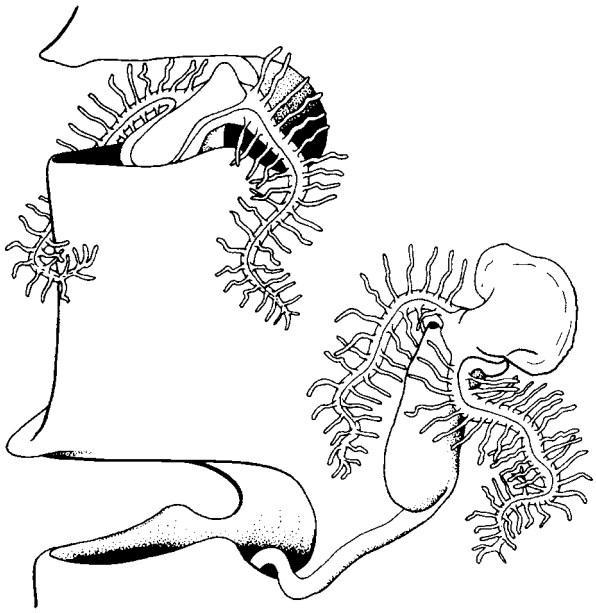


Figure 20 A hypothetical illustration depicting one zooid in the thecal tube and another extended from the thecal tube. This diagram is based upon an understanding of the detailed micro structure of the periderm or skeleton.

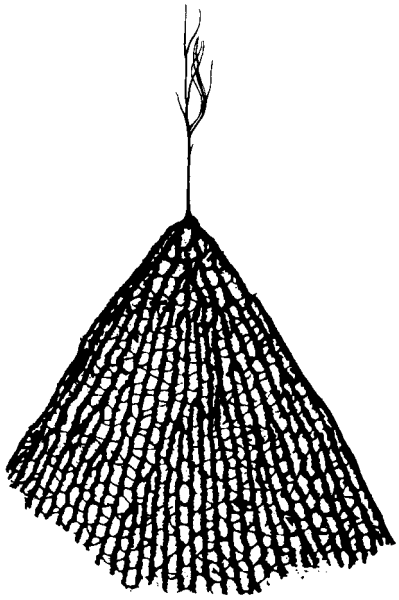


Figure 21 A planktonic species of *Dictyonema*, showing how it probably hung suspended in the surface layer of the ocean.

Figure 22 The diplograptid *Pseudoclimacograptus*, which probably hung from a dart like structure and rotated in turbulent currents.

All Graptoloidea, and the planktonic Dendroidea, occur lying flat on the bedding planes, reflecting the fact that they sank upon death from the overlying water. Sessile Dendroidea are occasionally found in positions of growth, with the basal discs or holdfasts downwards and the colony itself cross-cutting the bedding (Figure 19).

The Living Graptolite Colony

Preservation of graptolite soft parts is exceedingly rare, so that Figure 20 is largely hypothetical and is based on an understanding of the detailed microstructure of the periderm or skeleton. It is known, however, that the colonies themselves were either sessile or broadly planktonic. Figure 13 shows a species of the genus *Dictyonema* that had a basal disc or holdfast attached to the substrate of mud, sand, algal fronds, or rock or shell debris; such graptolites occurred in relatively shallow inshore environments along with a normal benthos of brachiopods, trilobites, and molluscs.

Figure 21 illustrates a planktonic species of *Dictyonema*, which probably hung suspended in the surface layer of the ocean from a termination of the nema consisting of a bundle of fibres enclosing soft tissue and tiny gas bubbles. The diplograptid *Pseudoclimacograptus* (Figure 22) was almost certainly pendant from a three-vaned dart-like structure that caused the colony to rotate in turbulent currents. Sessile graptolites have a limited geographical distribution, sometimes of only a few hundred miles, whereas the free-floating planktonic species often had worldwide distributions.

Broad Evolution of the Graptolites

Four successive, evolving, and overlapping faunas are known, leading from the Dendroidea (through the family Anisograptidae) to the Graptoloidea (Dichograptina, Diplograptina, and Monograptina). Many evolutionary lineages are known and some typical faunal elements are depicted in Figure 23. In the change from anisograptid to dichograptid we see a reduction in the number of stripes (and hence in the

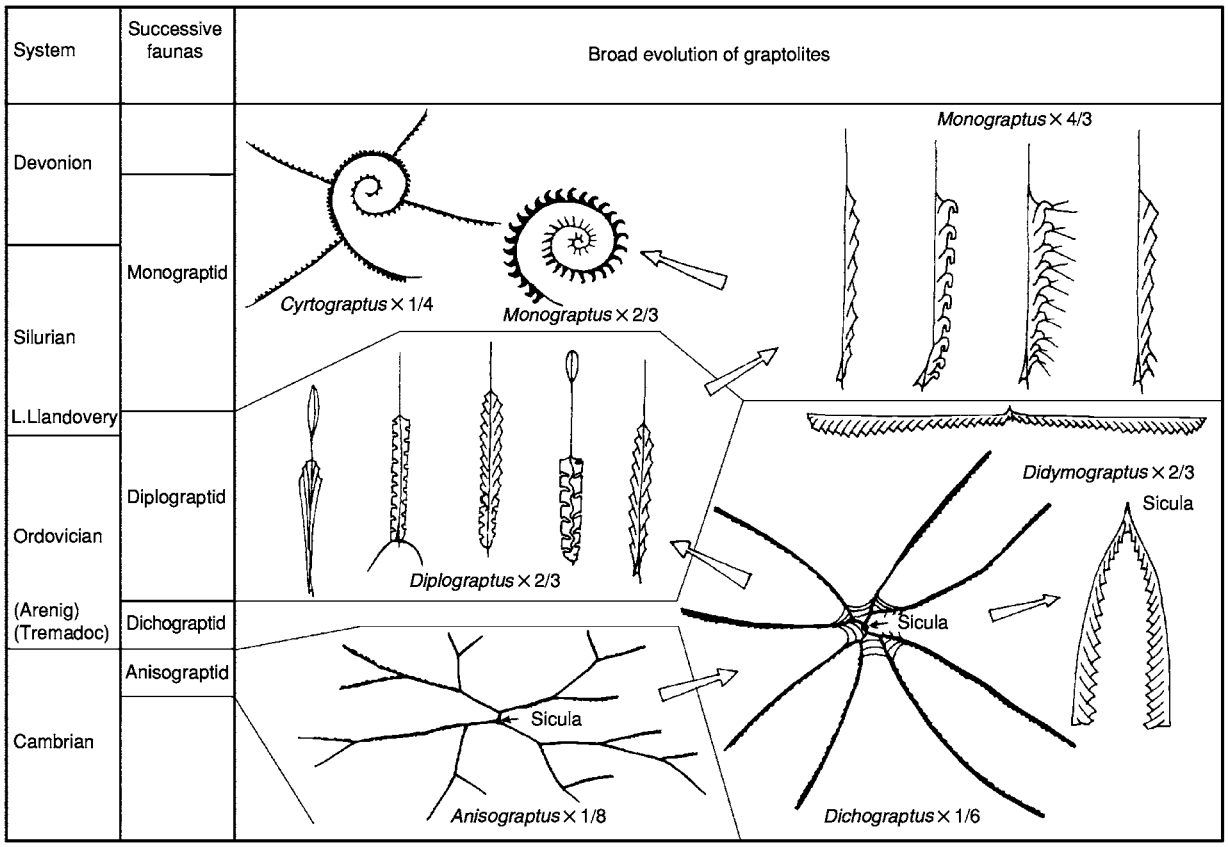


Figure 23 The broad evolution of the graptolites, from the anisograptids of the Upper Cambrian, through the dichograptids and diplograptids, to the monograptids of the Silurian and Devonian.

number of thecae), a reduction in the thickness of the cortical layer of the periderm, a loss of bithecae (typical of the dendroids; Figures 12 and 14), and the beginnings of a tendency towards a scandent position of the stipes (i.e. they tend to come upwards in later species, rather than hanging down as in earlier species). In the diplograptid fauna this last tendency reaches a peak with two stipes back-to-back, as in *Diplograptus* and *Climacograptus*, enclosing the nema for much of the rhabdosome length. In the monograptid fauna only one stipe remained in most forms, which was uniserial scandent (Figure 11A), although some did develop branches or become coiled (Figure 24).

Speciation in *Monograptus*

Further examples of speciation in the Monograptina are illustrated in Figure 24. All monograptids probably developed from a form with relatively simple thecae, which occurs first in the latest Ordovician. This row of simple thecae, uniserially arranged on a single scandent stipe, had the potential to develop thecal hooks, spines, enrolled tubes, and isolated thecae (Figure 15), whereas the rhabdosome as a whole had less potential, other than to curve, coil, or develop branches.

Stratigraphic Use of Graptolites

In the Ordovician and Silurian, and to a lesser extent in the Devonian, graptolites are perhaps the most important fossils for global correlation. In the Silurian, especially, they are particularly useful, and it has been calculated that the mean duration of a graptolite biozone is less than half a million years; in work done around the Wenlock–Ludlow boundary (mid-Silurian) the biozone duration is closer to one-quarter of a million years.

The reason for this great utility is that graptolites evolved very quickly yet often had a worldwide distribution in the tropical plankton realms. The mean duration of a graptolite species in the Silurian is under half a million years; only hominids are comparable.

Graptolites are used in several different ways in the relative dating of Lower Palaeozoic strata. The assemblage biozone depends for its recognition upon a unique association of species. The total-range biozone (acrozone) depends upon the known total temporal range of an individual species. Evolutionary lineages are also used where they are understood with some certainty: the change from one species to another occurred at a unique point in time. Finally, mathematical methods can be applied to the first and last appearances of species in different parts of the

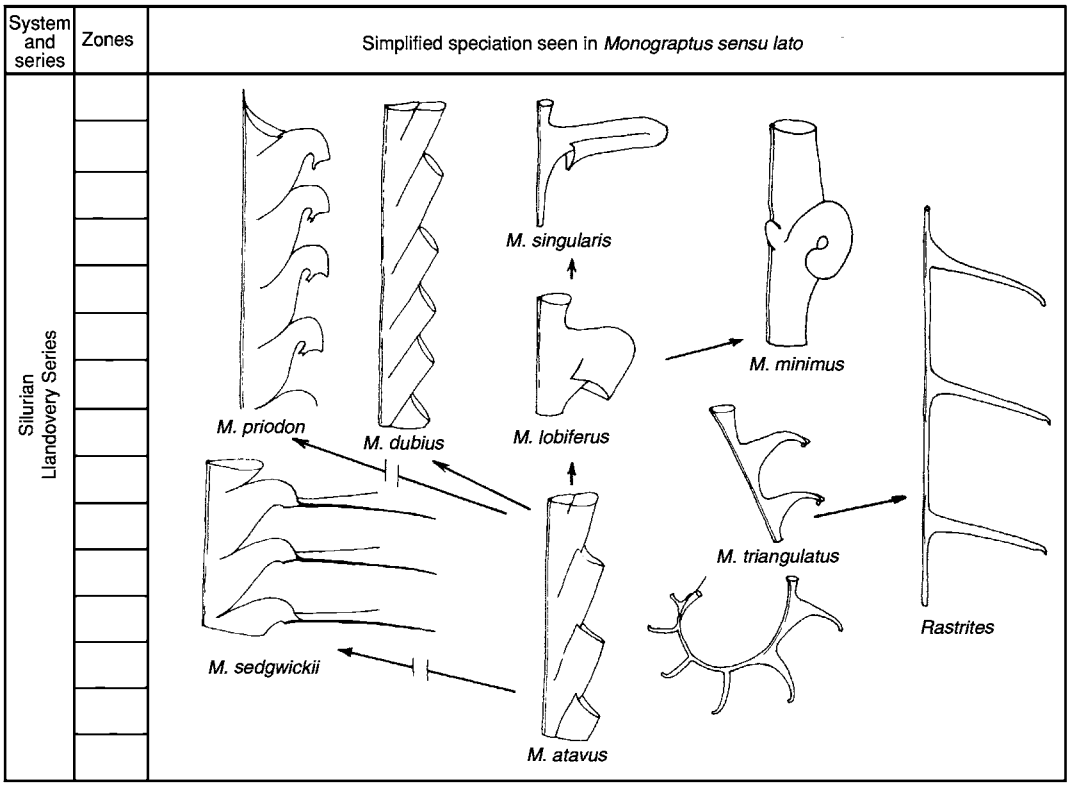


Figure 24 Examples of speciation in the Monograptina, showing very simplified sketches of some of the many known evolutionary lineages. For the purposes of simplicity, numerous morphological and stratigraphical intermediates have been omitted.

world. When these approaches are used in concert, or in conjunction with information from other groups, such as conodonts, acritarchs, and chitinozoans, an extremely powerful corrective tool results.

See Also

Biozones. **Fossil Invertebrates:** Bryozoans; Corals and Other Cnidaria. **Palaeozoic:** Ordovician; Silurian; Devonian.

Further Reading

Bulman OMB (1970) Graptolithina with sections on Enteropneusta and Pterobranchia. In: Teichert C (ed.) *Treatise on Invertebrate Paleontology,* Part V, 2nd edn, pp. 1 163. Geological Society of America and University of Kansas Press, Kansas, USA.

Elles GL and Wood EMR (1901 1918) *Monograph of British Graptolites*. Palaeontographical Society, London.

Fortey RA and Cooper RA (1986) A phylogenetic classification of the graptolids. *Palaeontology* 29: 631 654.

Kozlowski R (1949) Les graptolithes et quelques nouveaux groups d'animaux du Tremadoc de las Pologne. *Palaeon tooigca Polonica* 3: 1 235.

Palmer DC and Rickards RB (1991) *Graptolites: Writing in the Rocks*. Woodbridge: The Boydell Press.

Rickards RB (1979) Early evolution of graptolites and related groups. In: House MR (ed.) *The Origin of Major Inverte brate Groups*, pp. 425 441. Special volume 12. Systemat ics Association, London, New York, San Francisco.

Rickards RB (1995) Utility and precision of Silurian graptolite biozones. *Lethaia* 28: 129 137.

Ruedemann R (1947) *Graptolites of North America*. Memoir 19. Geological Society of America, Baltimore USA.

Urbanek A (1978) Significance of ultrastructural studies for graptolite research. *Acta Palaeontologica Polonica* 23: 595 629.

Molluscs Overview

N J Morris, The Natural History Museum, London, UK

The Mollusca is a phylum of invertebrate Metazoa that is semi-coelomate, protostomatous, and triploblastic. Molluscs inhabit marine, fresh water, and land habitats. The Mollusca are recorded from the beginning of the Cambrian Period, spreading into non-marine aqueous habitats during the latest Silurian or Early Devonian and apparently onto the land during the Carboniferous. Today they are only outnumbered, in terms of species, by the Arthropods (*see* **Fossil Invertebrates:** Arthropods).

In most modern classifications nine classes of molluscs are generally recognised, of which eight are extant. These are the Monoplacophora, the Cephalopoda (*see* **Fossil Invertebrates:** Ammonites; Cephalopods (Other Than Ammonites)), the Gastropoda (*see* **Fossil Invertebrates:** Gastropods), the Rostroconchia (now extinct), the Bivalvia (*see* **Fossil Invertebrates:** Bivalves) and the Scaphopoda making up the subphyllum Conchifera; and the Amphineura, the Solenogasters and the Caudofoveata making up the subphyllum Aculifera.

The molluscs are considered to be primitively bilaterally symmetrical, with an anterior head, a posterior anus, and a ventral foot. They have a mucus-producing skin which needs to avoid dessication. The dorsal surface of the body usually extends laterally to form an external cover (the mantle) which protects lateral pallial or mantle cavities. Primitively the mantle encloses laterally placed appendages such as the labial palps and the gills. The Mollusca need oxygen for their metabolism and usually use the blood pigment haemocyanin. However, haemoglobin is sometimes used, particularly in areas of the oceans where the oxygen concentration is low.

A characteristic feature of the Mollusca is the presence of a calcareous shell. The dorsal surface of the body and the mantle commonly secrete the shell, which has an outer organic layer, a periostracum, and an inner part made up of one or more layers consisting of complex microstructures of calcium carbonate with an organic matrix. The periostracum is secreted at the mantle margins except in the earliest growth stages and at times of shell damage when repairs may be facilitated by other parts of the mantle surface. The calcified shell is secreted across a space filled with extra-pallial fluid, by the dorsal surface and mantle margin on to the inner surface of the periostracum and or previous shell layers. The calcium carbonate of the shell may be in the form of crystals of aragonite or calcite but which are modified by their organic matrix.

The Mollusca primitively have an anterior mouth and a posterior anus with a gut that passes through the heart. They primitively have bilaterally paired kidneys and gonads. The gut is commonly coiled making it longer than the total body length in many living taxa, including the Monoplacophora and Amphineura. The gut is not coiled in the Scaphopoda and only slightly coiled in the Cephalopoda. Natural casts of stomachs and intestines are sometimes fossilised in detritus

feeding taxa, notably in some examples of the nuculoid bivalves and trochid gastropods. In all of the living classes, except the Bivalvia, feeding is assisted by a tongue like apparatus (the radula) just within the mouth, which continuously produces rows of chitinous teeth. It is generally accepted that organisms ancestral to the bivalves did have a radula. Cephalopods had developed a horny beak (superficially parrot-like) which may also be preserved as a fossil, while certain opisthobrach gastropods developed chitinous gizzard plates which have also been preserved in Cenozoic sediments.

The soft tissues of the body of Mollusca are attached to the shell by muscles which in most Bivalvia, Amphineura, Monoplacophora, and Cephalopoda are arranged in a bilaterally symmetrical pattern. In the Amphineura and the Monoplacophora this musculature is arranged in eight serially arranged blocks (metameres), giving the body a repetitive structural pattern, known as pseudometamery. In the Amphineura the musculature of the anterior two metameres is attached to the head, while the poserior six metameres are attached to the more posterior part of the body (foot). In the tryblidioidean Monoplacophora, the head makes up the three anterior metameres. This pseudometamery differs from true metamery of the Annellida in not having a discrete coelom associated with each metamere. The attachment of the muscles to the shell is by way of a particular shell microstructure, low aragonitic prisms forming the myostracum which is virtually universal in the Conchifera but not obvious in the Amphineura.

Reproduction in the Mollusca occurs primitively by fertilization of gametes released into the sea. The fertilised ovum develops into a trochophore larva similar to those of the annellids. After a short period of time, sometimes two days, the larva settles and metamorphoses into a benthonic adult which develops a shell. The larvae of many groups may extend their life by the development of a shell (secreted by a shell gland) invariably consisting of aragonite and organic material. In the larvae many bivalves and gastropods, paired lateral processes (the velar lobes) help the growing animal to support itself in seawater and provide a certain amount of locomotion. Where this stage occurs it is known as the veliger larva. Any of these stages may be modified, often within the protection of an egg capsule, which is clearly of importance in the invasion of the land habitat. Most larval shells are external, but in the scaphopods and nuculoid bivalves, the larval shell develops within the expanded skin of the trochophore, and is known as the test cell larva.

The history of larval development in the Mollusca is not firmly established. During the nineteenth century larvae with a seven-part bilaterally symmetrical chitinous shell, resembling the early development of the Amphineura, were described for the Solenogasters. These observations need to be confirmed, but if correct might indicate that the Solenogasters and Caudofoveata are secondarily without a shell. This view is contrary to that of some modern workers that the non-shelled chitons are more primitive than those with a shell and are the living representatives of a 'pre-shelled' molluscan stock.

Living Cephalopoda emerge from their eggs as (free swimming) miniatures, resembling their adults.

In many bivalves and gastropods, extension of larval life is achieved by increasing the size of the larval shell by growth at the shell margins, which often produces daily growth lines. This extension of larval life enables some achitectonicid gastropods to cross the Atlantic Ocean and may have allowed some Mesozoic inoceramid bivalves to exist in the plankton for more than sixty days, a feature coincident with the almost global distribution of some of their species.

The periostracum consists of proteinaceous material known as conchyolin and may have as many as three layers. It may be smooth or hairy and can include calcareous spicules which are sometimes fossilised. There are up to three layers of calcareous shell. The primitive conchiferan shell is considered to be made up of an external layer of aragonite prisms together with an internal layer of nacre. This shell microstructure is known to occur in the types of bivalves, gastropods, cephalopods, and monoplacophora which have other features thought to be primitive. This view is supported by evidence of shell microstructures preserved from the Ordovician. However, it has been suggested that some shells from the early Cambrian, whose inner surface structure has apparently been replicated by phosphatic material, may have consisted of calcite.

Cross-lamellar aragonite and complex cross-lamellar aragonite may develop in any of the layers and at least one of these microstructures occurs in some scaphopods, gastropods and bivalves.

'Amorphous aragonite', probably very fine-grained cross-lamellar structure, is present in some bivalves.

Outer calcite layers have evolved in some epifaunal benthonic molluscs, including some bivalves (Ostreoida, Pterioida, etc.) and gastropods (Euomphalidae, Patellidae, Buccinidae). This development may be related to the greater resistance of calcite to solution in seawater. In the oysters, bivalves in which all the shell layers have become calcitic, the muscle attachment layers (myostracum) alone remain as typical aragonitic prisms which are joined to the muscular tissue by microvilli on the basal cell membrane.

Amphineura have a dorsal integument of conchiolin, known as the girdle, which is sometimes pustulose

and rarely with calcified pustules. The shell consists of eight imbricating plates, symmetrically arranged in sequence along the centre of the girdle. The microstructure of the calcareous shellis aragonitic, and has been likened to the cross-lamellar structure of the conchifera, but the general pattern of the structure differs and is traversed by tubules which house the sensory aesthetes. The muscle insertions closely resemble those of the conchifera but they seem to have no substantial myostracum. The occurrence of fossil chitons is sporadic but their record is good back to the Early Ordovician. A number of probable chitons have been described from the earliest Cambrian (Meisuchun Stage) from Southern China.

A number of fossils of uncertain affinities may belong to the Mollusca. The hyolithids have comparable shell microstructure to some Mollusca but their appendages (helens) and the arrangement of the gut, as occasionally preserved, are quite different. Some Machaeridia may be forms intermediate between shelled and non-shelled chitons, as may be the Cambrian Halkeriids and *Wiwaxia*.

See Also

Fossil Invertebrates: Arthropods; Bivalves; Gastropods; Cephalopods (Other Than Ammonites); Ammonites.

Further Reading

Fretter V and Graham A (1994) *British prosobranch mol luscs: their functional anatomy and ecology.* (Revised Edition), vol. 144. London: Ray Society.

Moore RC (ed.) (1952 1996) *Treatise on Invertebrate Paleontology.* Parts I,K,L,N, Mollusca. 7 Vols. University of Kansas Press, Lawrence: Geological Society of America.

Natural History Museum (London) (1996) *British Palaeo zoic fossils,* edn. 4. London: HMSO.

Natural History Museum (London) (1993) *British Meso zoic fossils,* edn. 6. London: HMSO.

Natural History Museum (London) (1993) *British Caeno zoic fossils. Tertiary and Quaternary,* edn. 5. London: HMSO.

Pojeta J and Runnegar B (1976) *The paleontology of ros troconch mollusks and the early history of the phylum Mollusca.* Washington, DC: United States Geological Survey Professional Paper No. 968.

Stanley SM (1970) *Relation of shell form to life habits of the Bivalvia (Mollusca).* Boulder, Colorado: Geological Society of America. Memoir 125.

Wenz W and Zilch A (1939 1962) Gastropoda. In: Schindewolf OH (ed.) *Handbuch der Paläozoologie.* Berlin: Borntraeger.

Wilbur KM and Yonge CM (1964, 1966) *Physiology of mollusca.* 3 Vols. New York, London: Academic Press.

Bivalves

E M Harper, University of Cambridge, Cambridge, UK

Introduction

The Bivalvia, the second largest class within the phylum Mollusca, is one of the most familiar of all invertebrate taxa. Modern representatives, such as mussels, cockles, oysters, and scallops, are well known from excursions to the coast, and in many parts of the world they are important commercial species. Their generally excellent fossil record has allowed their evolutionary history to be traced back to the Early Palaeozoic and, for much of this time, they have been important components of many faunas. From rather modest beginnings, they have conquered a range of habitats from the deep sea to freshwater, exploited a wide range of life habits (from deep burrowing to swimming), and undergone a near-exponential taxonomic proliferation, a spectacular example of an adaptive radiation.

Bivalves come in all manner of shapes and sizes, from tiny, thin-shelled commensals that live in association with sea anemones, to giant clams and the extinct rudists and inoceramids which reach(ed) sizes well over 1 m. Shell morphology is extremely plastic, but all are modifications of the same basic theme. The intimacy of the shell morphology to life habit has been a great benefit in reconstructing the life habits of extinct bivalves, but has also frustrated many attempts to establish the relationships between different groups within the class. Bivalves have been proven to be good palaeoenvironmental indicators, but they have only limited use in biostratigraphy. Freshwater 'mussels' have been used to date fluvial deposits in the Carboniferous Coal Measures of Western Europe, and inoceramids have been used for Late Cretaceous deep marine settings (e.g., in New Zealand). In general, however, species of the

class are too long lived and too facies specific to be of any great value.

General Morphology

As the name implies, bivalves comprise two calcareous valves. These are arranged laterally (left and right), are joined dorsally by a partially calcified elastic ligament, and enclose the soft tissue. Each valve has clearly differentiated posterior and anterior features, i.e., inequilateral. The primitive arrangement, retained by most bivalves, was to have a plane of symmetry parallel to the commissure (the join between the two valves), resulting in valves which are mirror images of one another (i.e., equivalve). Although this symmetry is found in virtually all bivalves which live with the commissural valve perpendicular to the substrate surface (orthothetic), it has been lost in those which have adopted a pleurothetic habit where they lie on one valve (e.g., oysters, scallops). In these cases, there is a tendency for the two valves to become dissimilar (i.e., inequivalve), typically with the underlying valve becoming more bowl-like and the 'upper' one more reduced like a lid.

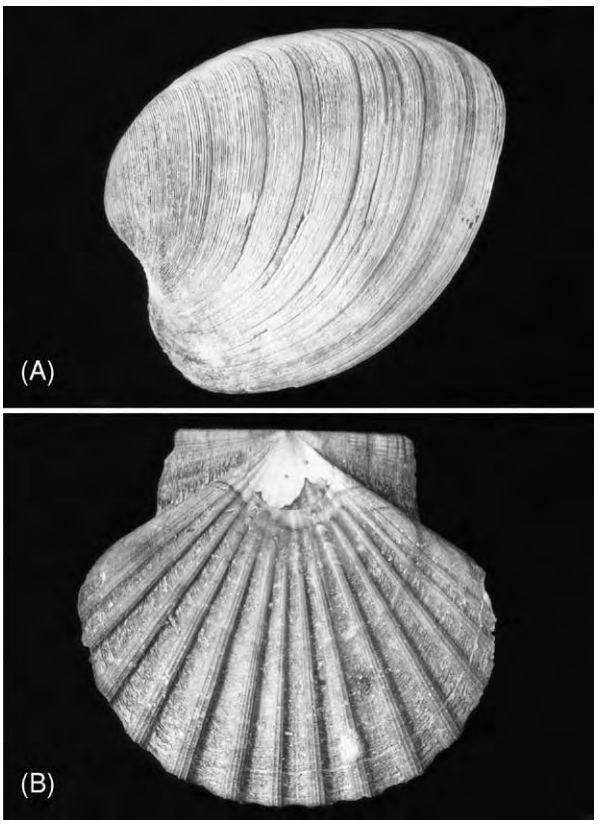


Figure 1 (A) *Mercenaria mercenaria*, a shallow burrowing bivalve from the Pliocene of Florida. Note the prominent annual growth bands. (B) *Pecten maxima*, a free living epifaunal scallop from the Holocene Atlantic.

Shell Morphology

All bivalves possess a pair of shells which may be shut to provide protection from both environmental stresses (e.g., desiccation in the intertidal habitat) and the threat of predation. Most shells are reasonably robust, which has provided the class with a generally excellent fossil record. Although shell morphology in bivalves is very variable and intimately linked to their life habits (see below), all shells are simple modifications of the basic shell secretion model used by all shelled molluscs. The shell is secreted by the mantle lobes and grows by marginal accretion, as evidenced by the growth lines on the surface of the valve (Figures 1A and 2A). These growth lines are particularly marked in bivalves from intertidal and shallow temperate habitats, where the animals experience pronounced seasonality and largely stop growing during the winter months. Bivalves which experience more equable conditions do not show such obvious or regular patterns. Inspection of the growth lines in sectioned valves shows that, although most shell material is added ventrally, the shell is also thickened during growth (Figure 2B), demonstrating that the entire mantle surface is responsible for adding material. The outermost part of the shell is an organic layer called the periostracum secreted at the mantle edge (Figure 2C). The thickness of the periostracum varies between taxa, from less than 1 μm in oysters and scallops to several hundred micrometres in some mussels. In many cases, the periostracum is lost by abrasion and decay during the life of the animal, particularly on the older parts of the shell, and there is no real prospect of it being preserved in any but the most exceptional circumstances. The primary function of the periostracum is to act as the template on which the calcareous part of the shell is deposited, but it may also provide protection from both corrosive waters and predators that dissolve the shell. It is particularly noticeable that freshwater bivalves have very thick periostraca.

The main part of the shell, however, is calcareous. It is in effect a ceramic made up of calcium carbonate crystals in an organic matrix (the latter accounting for $<5\%$ of the dry weight of the shell). The proteinaceous matrix controls both the polymorph of calcium carbonate used and the arrangement of the crystals. All bivalves contain aragonite in their shells and the vast majority are wholly so. Some taxa, however, chiefly those exploiting epifaunal life habits, also secrete calcite in their outer layers. The oysters have taken this to its extreme and the bulk of the shell is

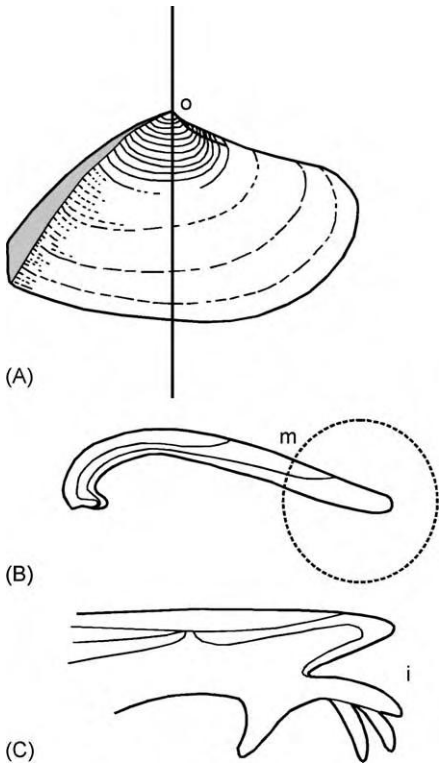


Figure 2 Details of shell formation in a generalized bivalve. (A) Marked comarginal growth lines on the shell surface. (B) Section through the shell along the line indicated in (A) showing the arrangement of growth lines within the shell. (C) The relationship between the shell and the underlying mantle edge (close up details of the circled area in (B)). i, inner mantle fold; m, middle mantle fold; o, outer mantle fold.

calcitic, with aragonite being confined to the sites of muscle attachment and the ligament.

Molluscan shell is immensely strong, in fact often much stronger than vertebrate bone. There are a number of different microstructures (Figure 3), each with different mechanical properties, and most shells are made up of two or three arranged in different layers. Different taxa show different arrangements and these are considered to be of phylogenetic significance. It is apparent that the earliest bivalves were wholly aragonitic and chiefly composed of nacre (Figure 3A), and that subsequent evolution has produced the wide array of microstructural arrangements seen today. The effect of differing crystal sizes, amount of organic material, and polymorph used has affected the preservation potential of different taxa; many of the Palaeozoic and Mesozoic taxa that were originally aragonitic are either preserved as internal moulds or are replaced by calcite.

Details of the internal features of the shell are shown in Figure 4. The hinge plate is situated dorsally and houses the ligament and teeth. The ligament is an elastic, partially calcified layer that provides a very energy-efficient opening mechanism. During valve closure, energy is stored in the ligament as it is flexed by the contracted adductor muscle(s) (Figure 4C). When the muscle is relaxed, the ligament springs the valves apart causing them to gape. This passive valve opening mechanism is the reason why many fossil bivalves are found in a disarticulated state. Although the ligament itself is seldom preserved, its position may be inferred from the presence of the ligament pits in which it is anchored (Figures 4 and 5). Most bivalves have teeth on the hinge plate which fit into corresponding sockets on the opposite valve and function to keep the valves in perfect alignment. Both ligamenture and dentition vary markedly amongst higher taxa of bivalves, and both are often used as informative characters in establishing phylogenies. Some of the range of hinge plate architectures is shown in Figure 5.

A number of attachment scars mark the locations where muscles are anchored to the shell. The most significant of these are the adductor scars (Figures 4A and 4B). If the adductor scars are paired (i.e., dimyarian), they occur posteriorly and anteriorly. If an animal is monomyarian, the single muscle (the posterior) occupies a more central position. In many taxa, there is a thin pallial line running around the shell a small distance from the ventral edges that marks the attachment of the mantle to the shell. In infaunal taxa, where the posterior mantle has been fused and elongated to form siphons, the pallial line is inflected forming the pallial sinus. The sinus represents the space into which the siphons are withdrawn when the valves are shut. Other muscle attachment scars may be more or less apparent, including the insertions of the pedal musculature (particularly in burrowers and byssate taxa).

Soft Part Anatomy

Bivalves are laterally compressed and, unlike most molluscs, there is no head or radula. The internal organs are enclosed by the two mantle lobes that are joined dorsally (Figure 4C). The chief function of the mantle is to secrete the shells, but the ventral edges of each mantle lobe are differentiated into three folds (Figure 2C), only the outermost of which is directly concerned with shell manufacture. The innermost fold controls water flow into and out of the mantle cavity, whilst the middle fold has sensory capability. In several bivalve groups (such as scallops), the middle fold is well developed with tentacles and eyes. In some taxa, the mantle is extended posteriorly and fused to form a pair of siphons through which water is directed into (inhalant) and out of (exhalant) the mantle cavity.

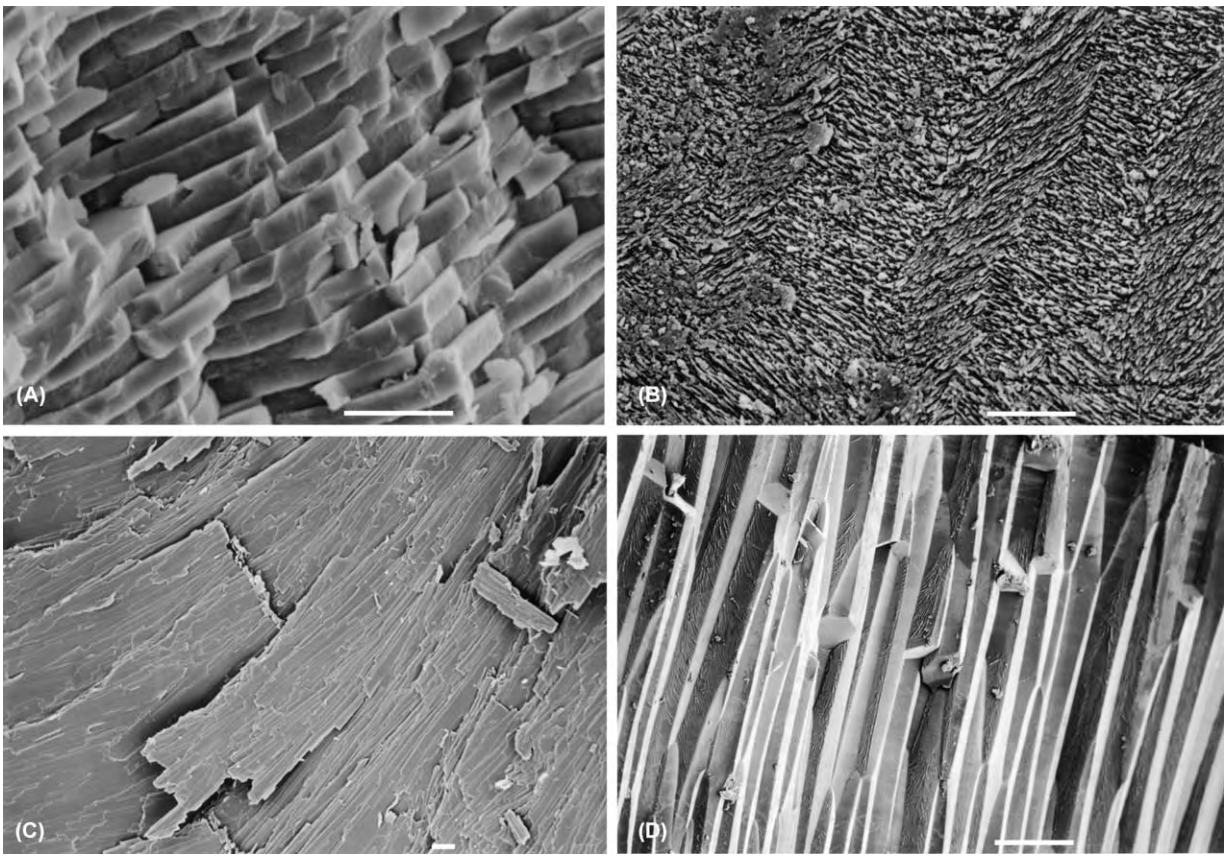


Figure 3 Scanning electron micrographs of four of the most common bivalve shell microstructures. (A) Aragonitic nacre (from the inner shell layer of *Pinna nobilis*). (B) Aragonitic crossed lamellar structure (from the outer layers of *Corbula gibba*). (C) Foliated calcite (from *Ostrea edulis*). (D) Calcitic prisms (from the outer shell layer of *Pinna nobilis*). Scale bars for (A) (C) represent 10 μm and for (D) represents 100 μm.

The mantle cavity is spanned by one or two adductor muscles that attach to the shell and act antagonistically with the ligament to close the valves on contraction (Figure 4C). A significant part of the mantle cavity is occupied by a pair of gills (the ctenidia) lying on either side of the rest of the viscera. In most bivalves, the gills are involved with both respiration and ciliary suspension feeding (filtering small particles out of the water which are then transferred to the mouth by a pair of labial palps). Recent bivalves show a number of different gill morphologies depending largely on the feeding process employed. Deposit feeders, e.g., *Nucula*, have less well-developed (protobranch) gills, whilst members of the Lucinidae augment their filter feeding by energy gained from the activities of sulphide-oxidizing chemosymbiotic bacteria living within the modified gills. The carnivorous septibranch bivalves (e.g., *Cuspidaria*, *Poromya*) suck in small prey (such as amphipods) using their modified siphons. These extraordinary bivalves have 'lost' their gills and respire over the inner surface of the mantle. Other significant organs within the mantle cavity include the gut, heart, circulatory system, and the foot. The gut runs between the anteriorly positioned mouth and the posterior anus, and includes a complex stomach which, again, has a number of configurations depending on the feeding biology of the animal. Blood is circulated throughout the animal by a three-chambered heart. A muscular foot is present in all juvenile and most adult bivalves and occupies the centre of the mantle cavity.

Naturally, the soft part anatomy of bivalves is very seldom preserved, although preservation of gill and muscle material has been reported in exceptional circumstances. Various details, however, can be inferred from the study of the internal surface of the shells. Apart from adductor muscle scars, the practised eye may pick out the attachment points of more minor muscles and impressions of radial muscles and blood vessels within the mantle.

Ecology

Modern bivalves exploit a wide range of life habits. Many burrow to varying depths within soft sediments,

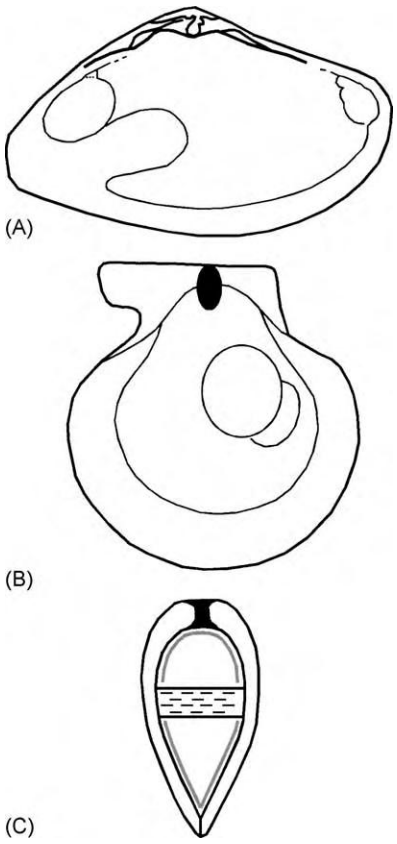


Figure 4 Various aspects of the internal morphology of bivalve shells. (A) The left valve of a generalized burrowing dimyarian. (B) The right valve of a generalized byssate scallop. (C) The relationship between the two valves and their associated mantle lobes, adductor musculature and ligament.

but others attach to or bore into hard substrates by a variety of means; others have become free living, some with the ability to swim. Most bivalves are marine, exploiting niches from the intertidal zone down into the abyssal depths, but successful groups (including the oysters) have invaded more brackish conditions and even freshwater, where modern unionid mussels cause enormous damage as biofoulers. It is clear that the most primitive bivalves were marine shallow burrowers and that other life styles evolved later. It is also apparent that many of the more specialized life habits have evolved separately in a number of different lineages (i.e., polyphyletically). Seminal work by S. M. Stanley firmly established how different aspects of the morphology of living bivalves could be related to their life habits, such that it is possible to use these characteristics of extinct taxa to reconstruct the life habits of fossil groups.

Burrowing

A large proportion of all bivalves (around 50% of all modern families) burrow into soft sediments using the foot. Most are equivalve and are isomyarian (i.e., the posterior and anterior adductor muscles are of equal size). The depth to which they burrow varies between taxa, from those which lie just under the surface with the edge of the shell virtually level with the sediment–water interface (e.g., *Cerastoderma*; Figure 6C), to depths of several centimetres (e.g., *Mya*; Figure 6B), with *Panopea* reaching spectacular depths of up to 1 m. The key to successful burrowing is maintaining contact with the seawater in order to continue both feeding and respiration. This is achieved by the siphons, snorkel-like extensions of the posterior mantle. The length of the siphons, and therefore the depth of burrowing, can be inferred from the shells by the size of the pallial sinus; deeper burrowers have more indented pallial sinuses, whereas very shallow burrowers have no sinus at all (see Figure 6). Very deep burrowers, such as *Mya*, have siphons so long that they are unable to withdraw them fully into the shell when it shuts, and have a permanent posterior siphonal gape through which they protrude. Shallow burrowers generally have strong, robust shells, often with a pronounced radial or concentric ornament that may assist the burrowing process or help the animal remain 'locked' into the sediment. Deeper burrowers tend to have thinner shells and are often smooth shelled. Although the foot is never preserved, its presence may be inferred from the pedal musculature on the inside of the valves and, in cases where the animal is a rapid and deep burrower (such as the razor shell *Ensis*), the foot may be so well developed as to require an anterior pedal gape.

It is clear from studies of the siphons of living bivalves that they are constructed in a number of different ways, suggesting that the deep burrowing habit has evolved independently in several clades.

Attachment

Almost all larval bivalves attach to the substrate, if only briefly, with tanned protein threads (the byssus) secreted by a gland at the base of the foot. In a large number of taxa, this habit has been neotenously retained into adulthood, and again it is clear that this has happened repeatedly in different groups. Byssate bivalves fall into two categories: those like *Pinna* (Figure 6A) and *Modiolus* that are orthothetic and live attached to clasts within the sediment in which they are partially buried (endobyssate), and those that are attached to the surface of hard substrates (epibyssate), either in an orthothetic (e.g., *Mytilus*; Figure 7D) or a pleurothetic (e.g., *Isognomon ephippium*; Figure 7C) orientation. Orthothetic byssate bivalves tend to be equivalve and have much reduced anteriors. This anterior reduction is reflected

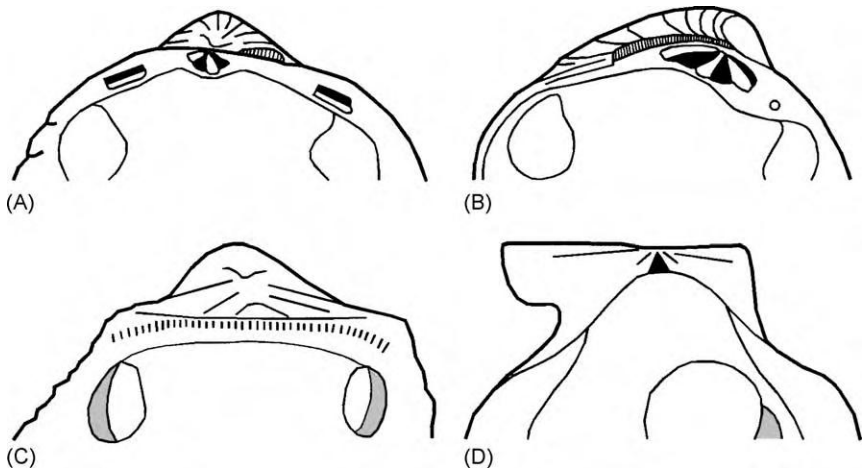


Figure 5 Selected hinge plates showing some of the variety of ligament insertion and arrangement of teeth. (A) *Cerastoderma edule*: heterodont dentition with two centrally placed cardinal teeth and two lateral teeth. (B) *Venus casina*: heterodont (similar to *Cerastoderma* but with no lateral teeth). (C) *Arca tetragona*: taxodont dentition with numerous teeth arranged in a row; the ligament forms a chevron pattern on the broad triangular area below the umbones. (D) *Chlamys varia*: two simple teeth with the internal ligament occupying a triangular pit below the umbones.

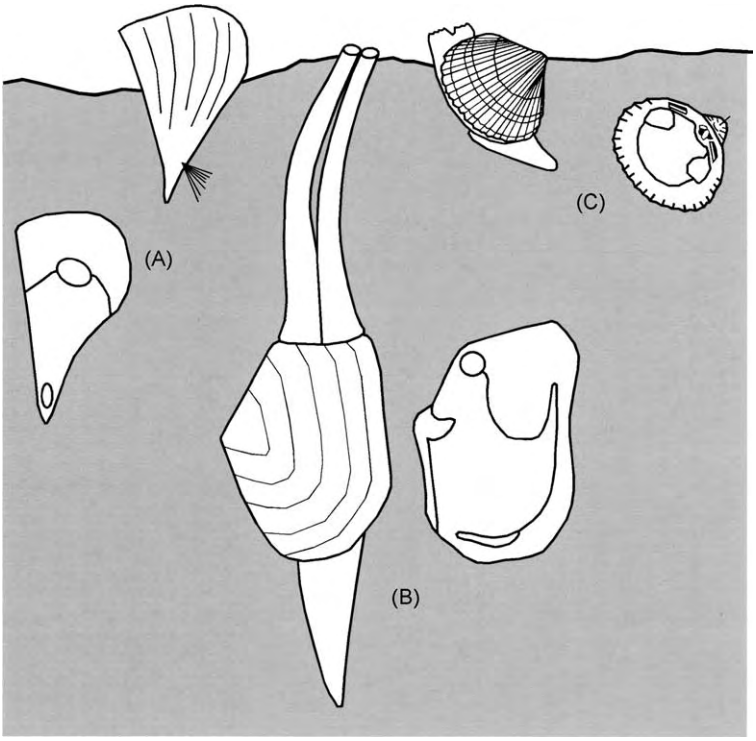


Figure 6 Morphology and mode of life of bivalves living in or partially within soft substrates. (A) *Pinna nobilis*. Not to scale. (B) *Mya truncata*. (C) *Cerastoderma edula*.

in the adductor musculature, which (although still dimyarian) is heteromyarian, with the anterior adductor much smaller than the posterior (Figure 7D). Pleurothetic byssate bivalves are often markedly inequivalve, with the 'lower' valve (which in the majority of cases is the right) often larger than the other. Although they are dimyarian early in ontogeny, the 'adults' are monomyarian, having lost the anterior muscle during ontogeny; the remaining posterior muscle is often large and centrally placed (Figure 7C). The presence of a byssus may be inferred from either a slight gape between the valves through which it passes (the byssal gape), or more obviously the byssal notch in scallops (Figure 4B).

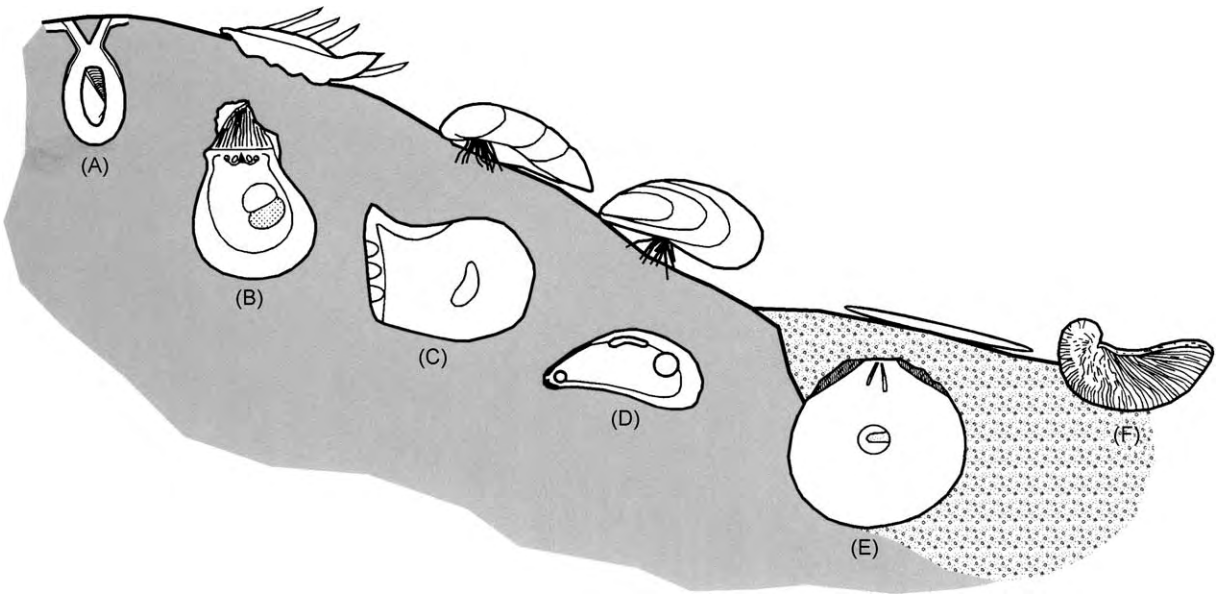


Figure 7 Morphology and mode of life of bivalves living on or boring into substrates. (A) *Spengleria rostrata*. (B) *Spondylus americanus*. (C) *Isognomon ephippium*. (D) *Mytilus edulis*. (E) *Placuna placenta*. (F) *Gryphaea arcuata*. Not to scale.

Whereas byssate attachment is flexible and also renewable, some bivalves permanently attach to hard substrata by a calcareous cement. The cemented habit always succeeds a byssate phase and it is clear that it has evolved independently in a number of different clades (e.g., oysters, rudists, and a number of pectinoids including *Spondylus*). Cemented bivalves are often easily recognizable from their irregular morphology, developed because of the requirement to conform to substratal irregularities, and are markedly inequivalve. As they tended to evolve from pleurothetic byssate stock, most are also monomyarian (Figure 7B). Most cement mainly by the right valve, but a major group, the oysters, do so by the left valve. The size of the attachment scar varies and the substrate may be instantly recognizable; for example, oysters were often attached to ammonite shells, even if the scar and substrate are no longer attached. Most cementers have thick, robust shells and may be extravagantly ornamented with spines or flanges (e.g., *Spondylus*; Figure 7B).

Free Living

A number of different taxa have independently abandoned attachment to become free living on softer sediments. Here, the challenge is not to sink into the substrate, and this has been solved in two ways. The first is by adopting a 'snow-shoe'-type morphology, i.e., resting on a large surface area, epitomized by the wafer-thin window pane shell *Placuna* (Figure 7E). Alternatively, they may be semi-submerged in the soft sediment like an iceberg. This strategy is inferred for the thick-shelled devil's toenail *Gryphaea*, common in Mesozoic clay facies (Figure 7F). These 'aberrant' oysters clearly had a cemented phase, marked by a small attachment scar at the umbo. A few free-living bivalves (notably several groups of scallops) also have the ability to swim short distances if they are threatened. These have smooth hydrodynamic shells and a well-developed posterior adductor muscle whose vigorous contraction provides the propulsion.

Boring

A number of groups, once again polyphyletically, have evolved to excavate 'burrows' in hard substrates by boring. The most successful of these, the mytilid lithophagids, do so principally by acidic secretions (which presents its own challenge of not dissolving its own shell), whilst others, e.g., *Pholas*, bore at least partly by physically rasping the substrate with small projections on the outside of the shell. Some of the most bizarre borers are the teredinids which excavate long cylindrical boreholes in wood and have the enzymatic capability of digesting the cellulose. These 'shipworms' are thought to have been part of the undoing of the ships of the Spanish Armada. Members of the boring group as a whole have a very varied morphology, but may be easily recognized because they are almost invariably fossilized within their characteristic boreholes (Figure 7A).

Taxonomy and Biological Relationships

There are around 100 families of living and a further 150 families of extinct bivalves. Traditionally, these are split into six subclasses which have been defined on the basis of a mosaic of characters, few of which are diagnostic in their own right. Table 1 gives the names and basic attributes of each of these subclasses, providing examples of each. Most modern bivalves belong to one of two subclasses: the infaunal Heterodonta and the largely epifaunal Pteriomorphia.

Despite the excellence of the bivalve fossil record, the wealth of Holocene material on which to base studies, and the general familiarity of the class, our understanding of the relationships between higher taxa is extremely unclear, and the monophyly of particular subclasses has been called into question. Even the relationship between the Bivalvia and the other molluscan classes is debatable. The principal difficulties stem from the high degree of convergent and parallel evolution. Some instances of convergence are clear; for example, the multiple evolution of byssate attachment has led to the heteromyarian 'mussel' form being adopted in unrelated taxa, such as the marine mytilids (e.g., *Mytilus*) and the freshwater driessenids. In these instances, close study of other anatomical characters clearly demonstrates that overall morphological similarity is superficial. The problem is more acute where convergence has affected much smaller groups, for example where the cemented habit has evolved repeatedly (five times) within the single family Pectinidae.

Most attempts to classify the bivalves have used either a single or very few characters. Zoologists and palaeontologists have invented different schemes because, not unnaturally, they have placed different importances on anatomical (e.g., gill or stomach architecture) and shell characters, such as hinge structure and shell microstructure, respectively. These are all subject to convergence. Modern computer-driven cladistic analyses which utilize large numbers of different characters offer a better hope of establishing phylogenetic relationships and of identifying convergent characters. Most informative of all are the new battery of molecular techniques; in particular, studies of the 18S rRNA gene have proved to be illuminating. Superimposing morphological data onto molecular phylogenies provides perhaps the ultimate means of establishing convergent characters. However, both the use of data sets combining shell and anatomical characters and the use of molecular methodologies are clearly of little help when dealing with extinct taxa. The fossil record has a critical role

Table 1 Six subclasses of bivalves

Subclass	*Examples*	*Characteristics*
Palaeotaxodonta	*Nucula*	Wholly aragonitic shells of nacre and prisms. Taxodont dentition. Equivalve. Deposit feeders with small protobranch gills. All infaunal
Cryptodonta	*Solemya*, *Cardiola*	Wholly aragonitic shells of nacre and prisms. Toothless. Infaunal. Protobranch gills, specialized to house symbiotic bacteria. Active burrowers
Pteriomorphia	*Mytilus*, *Modiolus*, *Lithophaga*, *Coralliodomus*, *Pinna*, *Ostrea*, *Gryphaea*, *Pecten*, *Spondylus*, *Isognomon*, *Imoceramus*, *Arca*	All include some aragonite within the shell (nacre, crossed lamellar microstructures), but many also have calcitic layers (prisms, foliae). A range of different hinge architectures and musculature (iso and heteromyarian). Eulamellibranch and filibranch gills. Many inequivalve. Pteriomorphs are primitively byssate, but have repeatedly evolved cementing, boring, and free living habits
Palaeoheterodonta	*Modiolopsis*, *Unio*, *Trigonia*	Wholly aragonitic nacre and prisms. Equivalve and a small number of heterodont teeth. Burrowers. Includes freshwater forms
Heterodonta	*Lucina*, *Mya*, *Corbula*, *Venus*, *Mercenaria*, *Ensis*, extinct rudists	Wholly aragonitic, mainly crossed lamellar and complex crossed lamellar microstructures. Mostly large heterodont teeth. Eulamellibranch gills. Equivalve and inequivalve forms. Majority are shallow and deep burrowers, but the extinct rudists were either cementing or free living
Anomalodesmata	*Cuspidaria*, *Poromya*	Wholly aragonitic (largely prisms and nacre). Largely toothless. Mainly isomyarian musculatures. Variety of morphologies reflecting very diverse life habits (burrowing, cementing, byssate)

to play in providing the temporal data that will help to discriminate between putative hypotheses, and provides a time-frame for establishing branching points on suggested trees.

Evolutionary History

Figure 8 shows the changing familial diversity of bivalves over the Phanerozoic. Aside from fluctuations, largely associated with mass extinction events, the overriding impression is one of almost exponential increase in diversity towards the present day. This taxonomic explosion has been accompanied by a movement away from the primitive shallow burrowing and byssate habits into more specialist habits (detailed above) in a series of adaptive radiations.

Bivalves are generally thought to have evolved from the extinct molluscan class Rostroconchia, although this is not universally accepted. The first fossils widely accepted (but again not universally) as bivalves are from Early Cambrian rocks. *Fordilla troyensis* is known from a large number of specimens from Tommotian localities in North America, Greenland, and Western Europe, whilst *Pojetaia runnegari* is known from Australia. Both *Fordilla* and *Pojetaia* are small, only a few millimetres in length, and are generally thought to have been shallow burrowers. Despite this early appearance, there is then a gap in the bivalve fossil record throughout the Middle and Upper Cambrian before their reappearance in the Early Ordovician (Tremadoc). This gap is both perplexing and frustrating; it spans some 4% of the evolutionary history of the class and covers an interval when bivalves apparently became larger and more diverse in terms of both taxa and the life habits employed. There have been several claims for Middle and Upper Cambrian bivalves, but most of these have been dismissed as misidentified inarticulate brachiopods. It is difficult to explain this gap in the bivalve record, but one plausible explanation is that the earliest bivalves may have lived in nearshore silty facies which are seldom preserved.

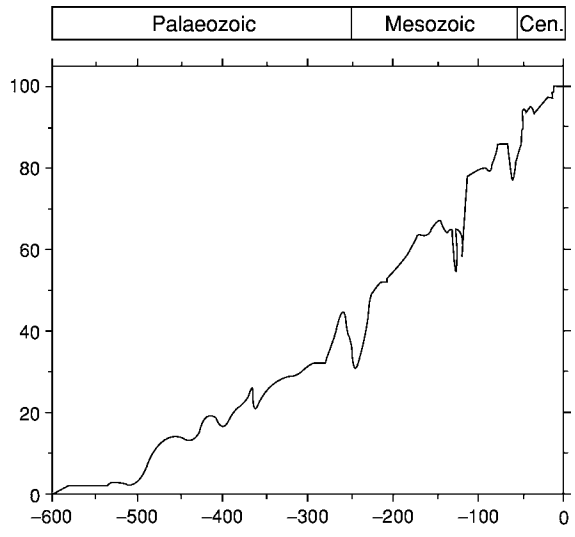


Figure 8 Familial diversity of bivalves over the Phanerozoic.

By the close of the Ordovician, bivalves were more common and abundant and all the subclasses recognized in Table 1 had appeared. There is evidence of shallow burrowers and both endo- and epibyssate forms and signs of different feeding strategies (deposit feeders, suspension feeders). Bivalve diversity continued to increase throughout the Palaeozoic, although there are small decreases associated with the Ordovician and the Devonian (Frasnian/Famennian) mass extinctions. Throughout the era most bivalves remained in 'primitive' life habit, with only one genus each of borers (*Coralliodomus*) and cementers (*Pseudomonotis*) being recognized (neither of which is widespread). Very few Palaeozoic bivalves (e.g., *Lyrodesma*) have pallial sinuses. Although the effect of the end-Permian mass extinction is clearly evident in Figure 8, losses were small in contrast with those sustained by other shelly benthos, for example the brachiopods.

The beginning of the Mesozoic saw a further combined increase in the number of taxa, with an ever increasing repertoire of life habits. Many shallow burrowers moved deeper, and large numbers of byssate taxa adopted more specialized habits, such as cementing, boring, and free living. The cementing habit is exemplified by the successful oysters and rudists. The first evidence of the adaptations for swimming in scallops are recognized in Jurassic taxa, and free-reclining gryphaeid oysters became abundant and widespread in this period. This major adaptive radiation is widely thought to have been linked to the Mesozoic Marine Revolution – the profound restructuring of shallow marine communities and, in particular, a massive increase in predation pressure at that time. Bivalves are extremely vulnerable to predators and are attacked by a wide range of important predatory groups (e.g., fish, gastropods, starfish, crustaceans, mammals, and birds) and by a variety of different methods (swallowing whole, crushing, drilling, or prising apart). It has been suggested that different groups of bivalves responded in a variety of different ways, for example by living deeper within the sediment, encasing themselves in hard substrates and becoming more robust, and, in some cases, growing formidable spines. Although the end-Cretaceous extinctions had a marked effect on the bivalves, with major groups such as the inoceramids and rudists being lost, their diversity continued to increase during the Cenozoic. The number of groups

exploiting the more specialist boring and cementing habit has continued to grow and there has been a further taxonomic explosion of the shallow burrowers at lower taxonomic levels.

See Also

Biological Radiations and Speciation. **Fossil Invertebrates:** Molluscs Overview; Gastropods; Cephalopods (Other Than Ammonites); Ammonites. **Mesozoic:** End Cretaceous Extinctions. **Palaeoecology**.

Further Reading

Beesley PL, Ross GJB, and Wells A (1998) *Mollusca: the Southern Synthesis,* vol. A. Collingwood, Australia: CSIRO Publishing.

Harper EM (1998) The fossil record of bivalve molluscs. In: Donovan SK and Paul CRC (eds.) *The Adequacy of the Fossil Record,* pp. 243 267. Chichester: John Wiley and Sons.

Harper EM and Skelton PW (1993) The Mesozoic Marine Revolution and epifaunal bivalves. *Scripta Geologica, Special Issue* 2: 127 153.

Harper EM, Taylor JD, and Crame JA (2000) *The Evolutionary Biology of the Bivalvia. Geological Society of London Special Publication 177.* London: Geological Society of London.

Johnston PA and Haggart JW (1998) *Bivalves: an Eon of Evolution. Paleobiological Studies Honoring Norman D. Newell.* Calgary: Calgary University Press.

Skelton PW, Crame JA, Morris NJ, and Harper EM (1990) Adaptive divergence and taxonomic radiation in post Palaeozoic bivalves. In: Taylor PD and Larwood GP (eds.) *Major Evolutionary Radiations. The Systematics Association Special Volume 42,* pp. 91 117. Oxford: Clarendon Press.

Stanley SM (1970) Relation of shell form to life habits of the Bivalvia. *Geological Society of America Memoir* 125: 1 296.

Taylor JD (1996) *Origin and Evolutionary Radiation of the Mollusca.* Oxford: Oxford University Press.

Vermeij GJ (1987) *Evolution and Escalation. An Ecological History of Life.* Princeton, NJ: Princeton University Press.

Vermeij GJ (1993) *A Natural History of Shells.* Princeton, NJ: Princeton University Press.

Gastropods

J Frýda, Czech Geological Survey, Prague, Czech Republic

Introduction

Gastropods are well-known animals which have been associated with humans since the dawn of civilization. Their bodies were gathered for food and their shells were used as tools, ornaments, and later as money. Their widespread occurrence is clear evidence of their successful adaptation to different environments. During a long evolution, they are the only molluscan class to have colonized the majority of marine, freshwater, and terrestrial environments. Marine gastropods occur mostly in shallow-water benthic communities; however, some gastropod species have also lived in the deep sea (e.g., faunas associated with hydrothermal vents), and others, such as holoplanktic animals, have spent their whole lives as free-swimming gastropods. The terrestrial gastropods colonized most land environments, ranging from lowlands to high mountains, and including humid to arid biotopes of tropical to subarctic areas. Such adaptive radiation is quite exceptional amongst all animal phyla and is linked to the extraordinary morphological and functional diversity of their bodies and shells. The gastropods comprise one of the most diverse groups of living animals (the second after Insecta). All these facts, together with their long and rich fossil record, make gastropods a unique animal group for evolutionary, ecological, and biogeographical investigations. There follows a brief review of gastropod anatomy, shell morphology, classification, and more than 500 million years of evolution.

Definition and General Description

The Gastropoda forms one of eight molluscan classes, and is defined by several unique anatomical features which support its interpretation as a molluscan group derived from the same ancestor (i.e., monophyly). The most characteristic feature of gastropods is torsion of their soft bodies during early larval stages, producing a crossing of their nerve connectives, bending of the intestine, and twisting of the mantle cavity (together with associated structures, including the ctenidia, anus, kidney openings, etc.) anteriorly over the gastropod head.

Anatomical Features

The exceptional morphological and functional diversity of gastropod bodies is also reflected in their anatomy. Generally, the body consists of a large foot, a

visceral mass, and a head with a mouth, tentacles, and eyes. The visceral mass is mostly enclosed, together with the mantle cavity, in a calcareous shell (Figure 1). Gastropods as soft-bodied animals use the pressure of their blood and muscles for movements of different organs. The circulatory system comprises a contractile heart with one (Caenogastropoda and Heterobranchia) or two (Patellogastropoda, Archaeogastropoda, and Neritimorpha) auricles, and a ventricle, as well as a system of arteries and veins. Gastropod blood transports oxygen using the copper-bearing pigment haemocyanin. In most marine gastropods, one or two gills (ctenidia) situated in the mantle cavity are used for respiration. Some marine and freshwater gastropods developed secondary gills after the loss of their ctenidia in previous evolution. In the terrestrial gastropods (e.g., Pulmonata), a highly vascularized internal wall of the mantle cavity (lung) is used for respiration. One or, rarely, two (Patellogastropoda and Vetigastropoda) kidneys serve for excretion through the mantle cavity. The digestive system starts with a mouth containing a tooth-bearing ribbon (radula). The organization of gastropod radulae and stomach, as well as additional parts of the digestive system, reflects their different feeding habits (herbivory, detritivory, carnivory, or parasitism). The anus opens into the mantle cavity. The gastropod nerve system includes paired ganglia which are linked with different sensory receptors by connectives and commissures. The morphology of the reproductive organs and the reproductive strategies are highly diverse. Generally, more ancient gastropod groups are gonochoristic with a simple reproductive system and external (Patellogastropoda and Archaeogastropoda) or internal (Neritimorpha) fertilization. Caenogastropoda also use internal fertilization with complex reproductive morphology, and some may be simultaneous hermaphrodites. The Heterobranchia have the most complex and variable reproductive system and are hermaphroditic.

Ontogeny

Gastropods, like all Mollusca, have a biphasic life cycle (i.e., larval and post-metamorphosis stages), and this feature is shared with closely related animal phyla (Kamptozoa, Sipunculida, Polychaeta, etc.). Like other molluscan groups, the embryonic development is characterized by a spiral cleavage, which differs slightly in the main gastropod groups. The subsequent larval stage is called the trochophore

Figure 1 Some variations in shell form of living gastropods. (A) High spired: *Mitra mitra* (Muricoidea). (B) Strombiform: *Lambis chiragra* (Stromboidea). (C) Turbiniform: *Liguus vittatus* (Orthalicoidea). (D) Convolute: *Cyprea tigris* (Cypraeoidea). (E) Spinose fusiform: *Chicoreus ramosus* (Muricoidea). (F) Fusiform: *Pleuroploca trapezium* (Muricoidea). (G) Conoidal: *Conus litteratus* (Conoidea). (H) Discoidal: *Architectonica perspectiva* (Architectonicoidea). (I) Turriculate: *Terebra* sp. (Conoidea). (J) Ovoid: *Olivancillaria gibbosa* (Olivoidea). (K) Involute: *Cypraecassis rufa* (Tonnoidea). (L) Irregularly coiled: *Siliquaria ponderosa* (Cerithioidea). (M) limpet: *Megathura crenulata* (Fissurelloidea). (A, E G, I, J) Neogastropoda (Caenogastropoda); (B, D, K) Littorinimorpha (Caenogastropoda); (C, H) Heterobranchia; (I) Sorbeoconcha (Caenogastropoda); (M) Vetigastropoda (Archaeogastropoda).

larva, and a similar larval type is developed in all molluscan groups. The trochophore larvae may be free swimming, as in the ancient gastropod groups (Patellogastropoda and Archaeogastropoda), or may occur in egg capsules, as in more advanced gastropods. The last larval stage is termed veliger, which typically bears two ciliate paddles (velum), sometimes subdivided into several lobes. If free-swimming gastropod larvae use planktic organisms for their nutrition, their development is termed planktotrophic. Marine gastropods with such development have small eggs, but numbering over half a million. Planktotrophic larvae may stay planktic for several months and thus can be carried for long distances by oceanic currents. The gastropods, however, developed another ontogenetic strategy in which their larvae were not dependent on an external food source, but on the yolk of their eggs. Gastropods with such a non-planktotrophic development (lecithotrophic) typically produce fewer eggs, which are relatively large. The larval stages end with a metamorphosis that involves anatomical and physiological reorganization of the larval body into the juvenile, post-larval body. Terrestrial and freshwater gastropods have simplified their development, and their embryonic and larval stages are fixed on egg capsules or the female body (direct development). Such ontogenetic changes considerably decreased their dispersal potential.

The Gastropod Shell

Gastropods are not only one of the most diverse animal groups, but the morphology of their shells is extremely varied (**Figures 1** and **2**). During more than 500 million years of evolution, they developed shells with various shapes and ornament, ranging in size from about 1 mm up to more than 1 m (Eocene Campaniloidea, Caenogastropoda). The shell and its ornament may be broadly linked to the mode of gastropod life (e.g., origin of limpet-shaped shells in unrelated gastropod groups). Generally, the most ornate shells occur in tropical marine environments, but freshwater and terrestrial gastropods are often less ornate.

Protoconch and Teleoconch

In shell-bearing gastropods, the shell grows during almost the whole of their ontogeny. The part of the shell formed during the embryonic and larval stages is called a protoconch (**Figure 3**), and that growing after metamorphosis is termed a teleoconch. The main gastropod groups differ in their early development, which is reflected in their protoconch morphology. The more ancient gastropod groups (Patellogastropoda and Archaeogastropoda) have the simplest shell ontogeny and their protoconchs have only an embryonic shell (protoconch I), which is followed by a teleoconch (**Figures 3B** and **3F–3H**). On the other hand, the protoconchs of more advanced gastropods (Neritimorpha, Caenogastropoda, and Heterobranchia) consist of an embryonic shell (protoconch I) and a subsequent larval shell (protoconch II). In most caenogastropods, the larval shells have different ornament from the teleoconchs (**Figures 3K** and **3L**), and both shells are coiled in the same direction (such a condition is termed homeostrophic; **Figure 4**). In contrast, in the Heterobranchia with planktotrophic development, the protoconchs are coiled in the opposite direction to the teleoconchs (**Figures 3J** and **4**). Such shells are termed heterostrophic. The Neritimorpha form typical, strongly convolute protoconchs during planktotrophic development, which are homeostrophic (**Figures 3A, 3N,** and **4**). Higher gastropods with non-planktotrophic development (some marine, freshwater, and terrestrial gastropods) have simplified their early ontogeny and thus also the morphology of their protoconchs. The latter strategy is documented from the Devonian (400 Ma).

Operculum

The majority of gastropods have a lid-like structure (operculum) to close their aperture. This operculum is present in all living gastropods during their larval stages, but is lost in some adults (e.g., limpets and the majority of terrestrial gastropods). The operculum is mostly horny (corneous) and may be tightly (multispiral) or loosely (paucispiral) coiled or concentric. Some gastropod groups have calcareous opercula, and the oldest operculum known is from the Ordovician (Macluritoidea).

Shell Structure

Most gastropod shells are composed of an outer organic layer (periostracum) and an inner, mostly much thicker, calcified layer. The colour pattern typical of many gastropod shells (**Figure 1**) is formed by different organic pigments which are limited to the periostracum and the uppermost calcified layer. This shell feature, sometimes reflecting the mode of life, has been known since the Palaeozoic (**Figure 5**). The inner layers of gastropod shells consist of minute calcium carbonate crystals (aragonite or calcite) in an organic matrix. There are over 20 structural types of gastropod shell and, in general, more ancient groups exhibit more diverse shell structures. The Patellogastropoda (Eogastropoda) had the most complex shell structure. On the other hand, the majority of the higher gastropods have developed simple

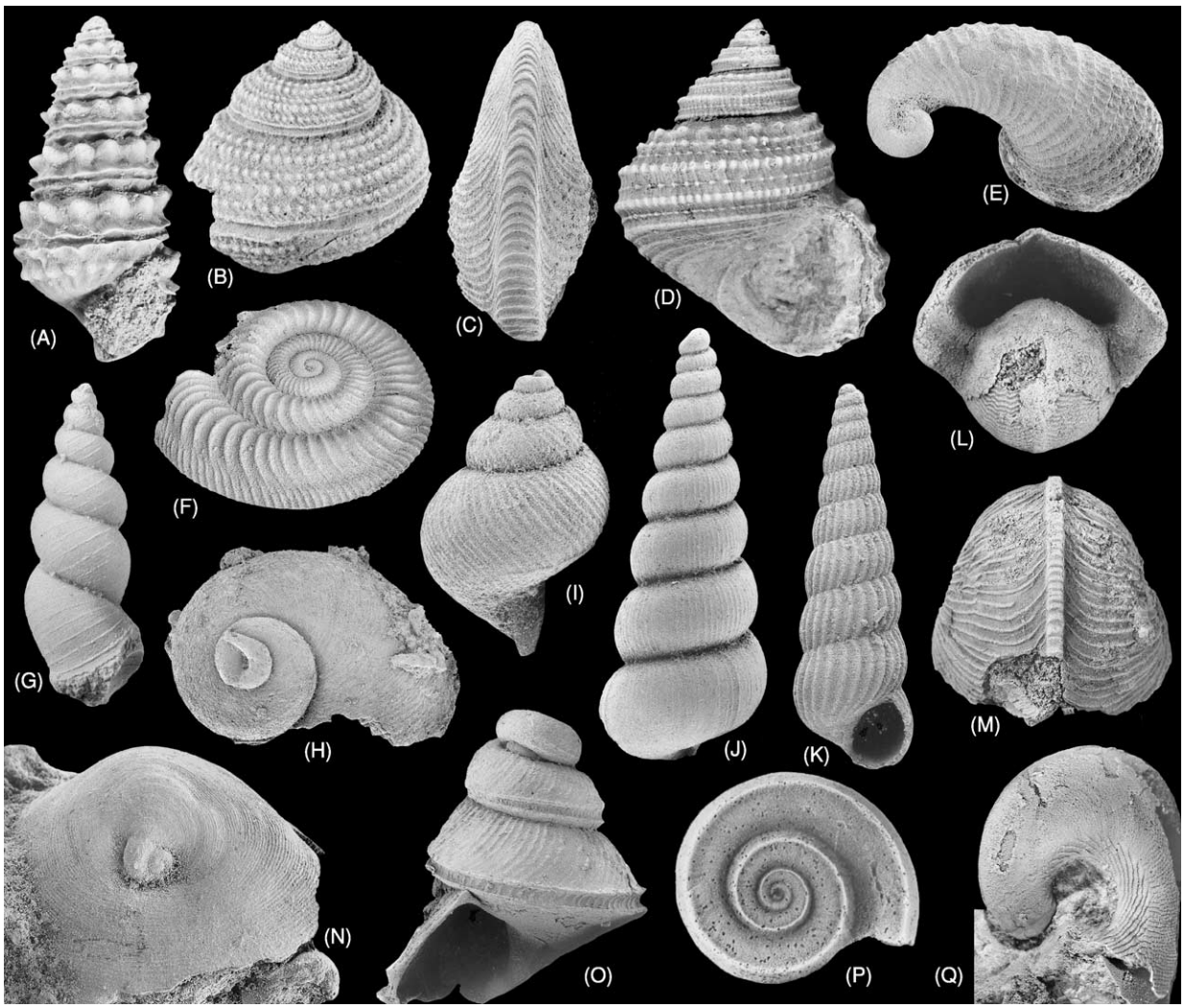


Figure 2 Some variations in shell form in the main groups of middle Palaeozoic gastropods. (A) High spired shell: *Murchisonia coronata* (Murchisonioidea). (B) Trochiform, slit bearing shell: *Devonorhineoderma orbignyana* (Eotomarioidea). (C) Bilaterally symmetrical shell with a prominent selenizone: *Kolihadiscus tureki* (Cyrtolitoidea). (D) Turbiniform shell: *Gyronema armata* (Gyronematidae). (E) Openly coiled shell: *Pragoserpulina tomasi* (Pragoserpulinidae). (F) Discoidal shell: *Stusakia pulchra*. (G) Sinistrally coiled shell: *Voskopiella barborae* (Onychochilidae). (H) Naticiform shell: *Eifelcyrtus blodgetti* (Vltavielidae). (I) Fusiform shell: *Havlicekiela parvula* (Perunelloidea). High spired shells: (J) *Pragozyga costata*; (K) *Palaeozyga bohemica* (Loxonematoidea). (L, M) Bilaterally symmetrical shell: *Bellerophon vasulites* (Bellerophontoidea). (N) Limpet: *Pragoscutula wareni* (Pragoscutulidae). (P) Sinistrally coiled shell: *Alaskiella medfraensis* (Porcellioidea). (P) Discoidal shell: *Nodeuomphalus labadyei* (Euomphaloidea); (Q) Bilaterally symmetrical shell covered by secondary shell deposits: *Branzovodiscus bajae* (Bellerophontoidea). (A, B, D, F, O) Archaeogastropoda; (C) Cyrtonellida; (G) Mimospirina; (H) Cyrtoneritimorpha; (L, M, Q) Bellerophontida; (I) Perunelomorpha; (E, J, K, N) Order uncertain; (P) Euomphalomorpha.

aragonitic shells with a crossed lamellar structure (Figures 6A and 6B). Some structural types are restricted to certain groups (e.g., nacre) and this may be used for their identification in fossils (Figure 6C). Nacreous and crossed lamellar structures have been known since the Palaeozoic.

Shell Coiling

The majority of the shell-bearing gastropods have right-handed (dextral) shells, but some have left-handed (sinistral) shells (Figures 1 and 2). Only a few gastropods have bilaterally symmetrical shells which may be uncoiled (limpets) or planispirally coiled (Figures 1M, 2C, 2L, and 2M). The limpet-shaped shells were independently developed within all main gastropod groups from the asymmetrically coiled shells of their ancestors. In contrast, planispirally coiled shells are known only in several groups, such as the Palaeozoic Porcellioidea and Bellerophontoidea (Figures 2L and 2M) or the Holocene Planorbioidea. Some gastropods may change the coiling

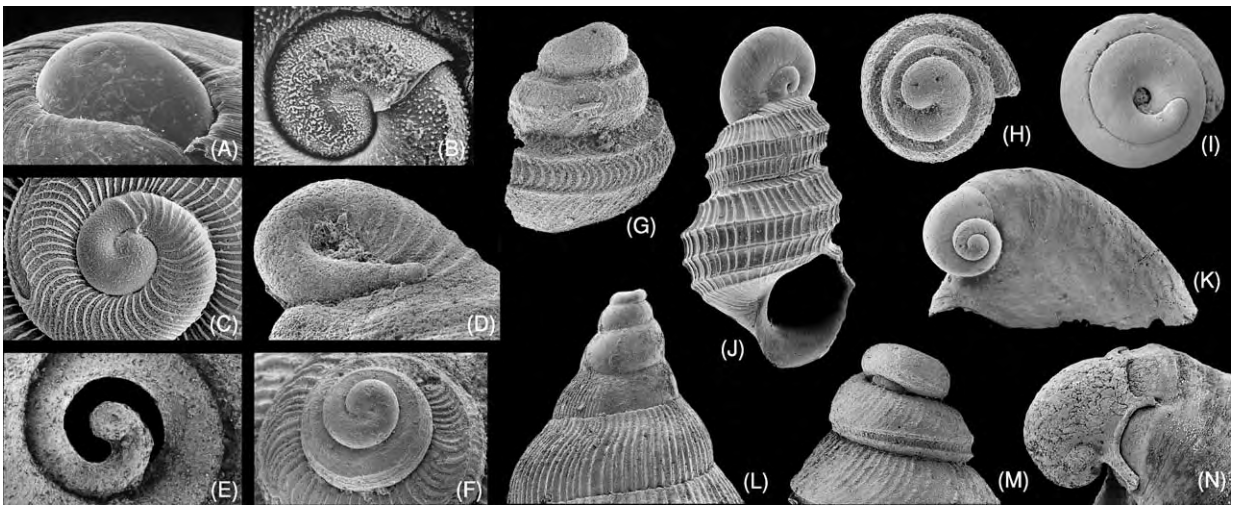


Figure 3 Variety of protoconch shape. Strongly convolute larval shell (protoconch II), Neritimorpha: (A) Holocene *Smaragdia* sp. (Neritoidea); (N) Triassic *Pseudorthonychia alata* (Pseudorthonychiidae). Embryonic shell (protoconch I) followed by teleoconch, Archaeogastropoda: (B) Triassic *Wortheniella coralliophila* (Vetigastropoda); (C) Holocene *Anatoma proxima* (Vetigastropoda); (F) Devonian *Zlichomphalina* sp. (Eotomarioidea); (G, H) Devonian *Diplozone innocens* (Murchisonioidea). (D) Openly coiled larval shell (protoconch II), Cyrtoneritimorpha, Carboniferous *Orthonychia parva* (Orthonychiidae). (E) Openly coiled early shell, Permian *Euomphalus* sp. (Euomphaloidea). (I) Openly coiled larval shell (protoconch II) of the Silurian Peruneloidea. (J) Heterostrophic larval shell (protoconch II), Jurassic *Mathilda* sp. (Architectonicoidea, Heterobranchia). Larval shell (protoconch II), Caenogastropoda: (K) Holocene *Hipponix* sp. (Vanikoroidea); (L) Devonian *Balbiniconcha cerinka* (Subulitoidea). (M) Heterostrophic early shell, Devonian *Alaskiella medfraensis* (Porcellioidea, Archaeogastropoda).

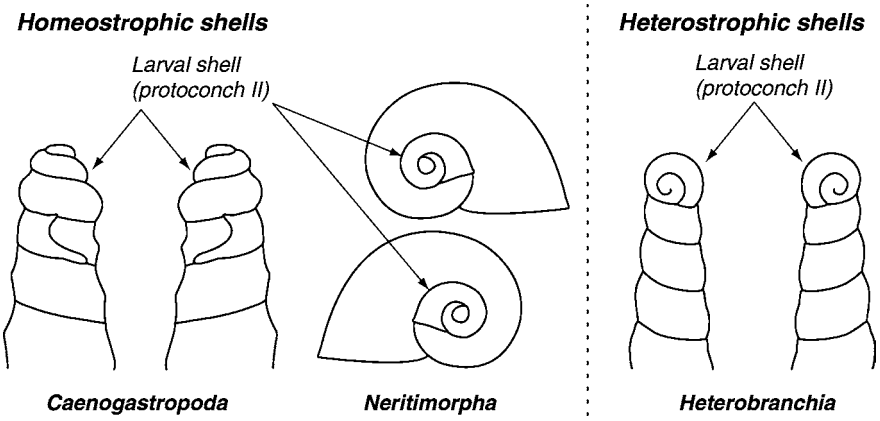


Figure 4 Schematic diagram showing the relationship between the coiling of larval (protoconch II) and post larval (teleoconch) shells in planktotrophic Caenogastropoda, Neritimorpha, and Heterobranchia. Coiling of both shells in the same direction (Caenogastropoda and Neritimorpha) is termed homeostrophic. If the handedness of the shells is opposite (Heterobranchia), the coiling is termed heterostrophic.

of their shells (Figure 4) from sinistral to dextral (dextral heterostrophy), or vice versa (sinistral heterostrophy), during ontogeny. Such a change may occur at a developmental stage, when gastropods undergo a metamorphosis from larval to post-larval stages (e.g., Heterobranchia; Figure 3J), or later (e.g., Porcellioidea; Figure 3M).

Dextrality or sinistrality of the shell is independent of the coiling of the soft body, and the asymmetrical soft body of gastropods may be dextral or sinistral. Anatomical dextrality or sinistrality may be easily recognized, even in fossil gastropods, if they developed a spiral operculum. The spiral operculum of anatomically dextral gastropods is coiled counterclockwise (viewed externally), and vice versa in sinistral gastropods. Thus, there are four possible relationships between shell coiling and body asymmetry in the shell-bearing gastropods (Figure 7). If anatomically dextral (or sinistral) animals occupy dextrally (or sinistrally) coiled shells, such a condition

is called dextral (or sinistral) orthostrophy. If the handedness of the shell and soft body is different, the term hyperstrophy is used. All four kinds of coiling (Figure 7) have occurred in gastropods, but their frequencies are very different. The great majority of living gastropods are dextrally orthostrophic, and sinistral orthostrophy is uncommon. Dextral or sinistral hyperstrophy is very rare (e.g., Ordovician Macluritoidea or some Holocene Planorbioidea).

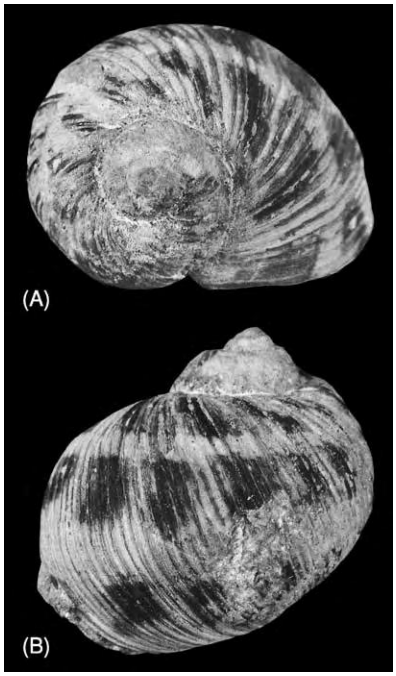


Figure 5 Apical (A) and lateral (B) views of the Middle Devonian (about 400 Ma) neritimorph, *Paffrathopsis subcostata*, showing the colour pattern.

Muscle Scars

Gastropod shells are attached to the soft body by muscles, which may leave distinct scars on the inner shell surface. The geometry of muscle scars has frequently been used as a diagnostic feature for distinction between torted (i.e., gastropods) and untorted states in the Palaeozoic molluscs (Monoplacophora, Helcionelloida, Cyrtonellida, etc.). However, new anatomical studies of living gastropods have shown that the larval muscles taking part in torsion and the post-larval muscles are developed quite independently. Thus, the muscle scar pattern sometimes observable in the fossil molluscan shells may be a good ecological indicator, but has no systematic significance.

Classification of the Gastropoda

Gastropods as an independent group of molluscs were recognized and named by the French naturalist, Georges Cuvier, more than 200 years ago (*see* **Famous Geologists:** Cuvier). Since then, scientists have tried to classify them by using different features of their bodies. However, the classification of such a numerous group with extraordinary morphological and anatomical variability of their bodies and shells has encountered many problems. During the nineteenth century, several different classifications of the Gastropoda were published, based on the shape of the shells, position of the mantle cavity, or on the arrangement of various organs (e.g., gills or head). Generally, these classification schemes used only a limited number of distinguishing characters. At the beginning of the twentieth century, the German zoologist, Johannes

Figure 6 Examples of shell structure in fossil gastropods. Aragonitic crossed lamellar structure in the Carboniferous (about 300 Ma) *Amphiscapha catilloides* (Euomphaloidea): views perpendicular to (A) and parallel to (B) the shell surface. (C) Nacreous structure (columnar nacre) in the Late Cretaceous (about 80 Ma) *Sensuitrochus ferreri* (Porcellioidea, Archaeogastropoda).

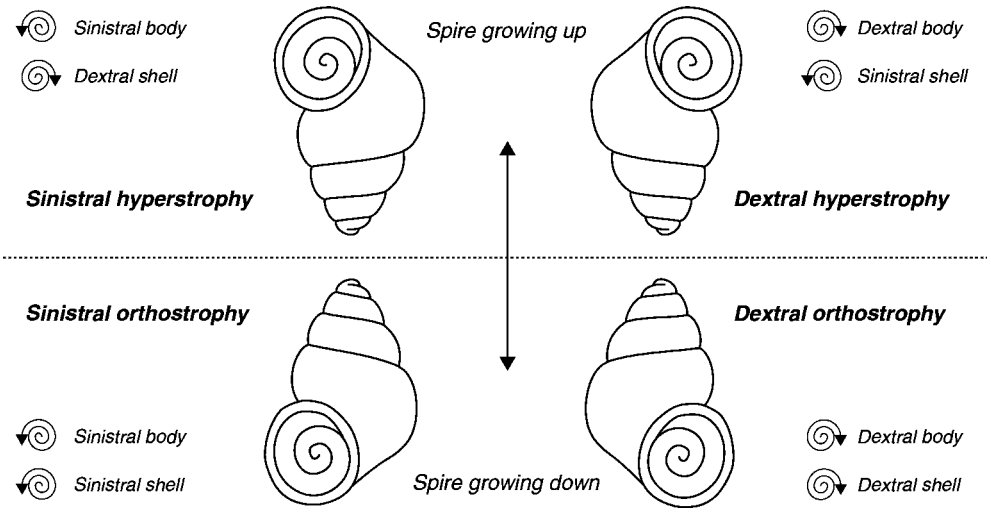


Figure 7 Schematic diagram showing the four possible relationships between shell coiling and body asymmetry. Orthostrophy means that anatomically dextral (or sinistral) animals occupy dextrally (or sinistrally) coiled shells. The term hyperstrophy is used when the handedness of the shell and soft body is different (see text for explanation).

Thiele, integrated earlier classifications and divided the gastropods into three subclasses: Prosobranchia, Opisthobranchia, and Pulmonata. In addition, the Prosobranchia were divided into three orders: Archaeogastropoda, Mesogastropoda, and Neogastropoda. Thiele's system was used by zoologists and palaeontologists for most of the twentieth century.

However, during recent decades, numerous new data on the anatomy of various gastropod groups have been accumulated, mainly by the application of new methods (e.g., transmission electron microscopy). At the same time, studies of the deep-sea faunas associated with hydrothermal vents have brought the discovery of new gastropod groups with unusual anatomical features. The evaluation of this newly gathered data in the light of the existing classification revealed a need for its revision. Recent analyses of numerous morphological and developmental characters of living gastropods have resulted in a new classification scheme (Figure 8), which has been independently supported by results from molecular studies. The placement of fossil gastropods into this classification of living gastropods has been difficult because of the lack of necessary anatomical characters.

Recent studies have revealed that Patellogastropoda (= Docoglossa, Cyclobranchia) represents the sister group to all other living gastropods. Living patellogastropods with limpet-shaped shells are exclusively marine and occur mostly on rocky shores in all continents. The Patellogastropoda and their coiled ancestors have been united into the subclass Eogastropoda. All other living gastropods and their ancestors have been placed in the subclass Orthogastropoda, comprising four main groups of living gastropods: Neritimorpha, Archaeogastropoda, Caenogastropoda, and Heterobranchia (Figure 8).

The Neritimorpha (= Neritopsina) is an ancient gastropod group with a long fossil record (Figures 2, 5, and 9), which colonized many different marine (shallow- and deep-water), freshwater, and terrestrial environments. The Palaeozoic Cyrtoneritimorpha, with openly coiled early shells (Figures 2H and 3D), may represent a closely related group. The living Archaeogastropoda unites the Vetigastropoda (Figures 3B and 3C) and several smaller groups, such as the Neomphaloidea, which occur in faunas associated with deep-sea hydrothermal vents. The Archaeogastropoda have colonized almost all marine and estuarine environments. There are also a number of extinct, mainly Palaeozoic groups (Figure 2) with uncertain relationships to living archaeogastropods. The Palaeozoic Euomphaloidea (= Euomphalomorpha; Figures 2P and 3E), known mainly from shallow-water, marine environments, may be a sister or basal group of the Archaeogastropoda.

The Caenogastropoda and Heterobranchia are sister groups which are united in the taxon Apogastropoda. Both groups are highly diverse and have colonized almost all marine, freshwater, and terrestrial environments. The Palaeozoic Subulitoidea and Peruneloidea (Perunelomorpha) (Figures 2I and 3L) may be ancestral or basal groups of the Caenogastropoda or of all Apogastropoda. The extant Caenogastropoda unites the two orders Architaenioglossa and Sorbeoconcha. Terrestrial Cyclophoroidea and freshwater Ampullarioidea form the Architaenioglossa. On the other hand, the mostly marine Sorbeoconcha represents a highly diverse group uniting

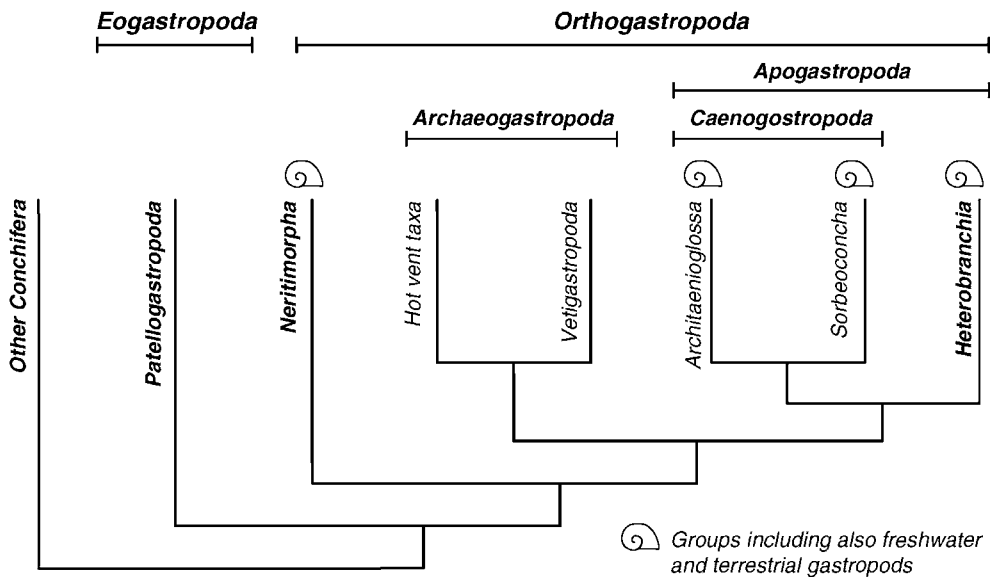


Figure 8 Recent classification scheme of living gastropods, illustrating their phylogenetic relationships and the distribution of freshwater and terrestrial groups. Based on Ponder WF and Lindberg DR (1997) Towards a phylogeny of gastropod molluscs: an analysis using morphological characters. *Zoological Journal of the Linnean Society* 119: 83 256.

more than 25 superfamilies of living gastropods (Figure 1).

The Heterobranchia encompasses the gastropod groups placed by Thiele's classification into the 'Opisthobranchia' and 'Pulmonata', as well as some 'prosobranch' groups, such as the Valvatoidea and Architectonicoidea. The Valvatoidea is an ancient group of freshwater gastropods, but the highly diverse Architectonicoidea represents a marine group (Figure 1H). The majority of the lower Heterobranchia (Opisthobranchia or sea slugs) are also marine gastropods, typically with their shells reduced or absent. They are extraordinarily variable and are divided into about 30 superfamilies of nine orders. The higher Heterobranchia (Pulmonata) form a dominant group of terrestrial gastropods, but also occur in freshwater environments. There are several classifications of the Pulmonata, which may be divided into three orders: Systellommatophora, Basommatophora, and Eupulmonata. The ancient marine Basommatophora have been separated into the Archaeopulmonata and the freshwater Basommatophora into the Brachiopulmonata. The Stylommatophora is a dominant group of terrestrial gastropods and is the most numerous group in the Eupulmonata.

The higher classification of extinct gastropods is less stable than that for living groups. The Palaeozoic, exclusively marine Pelagiellida, Bellerophontida (Figures 2L and 2M), Macluritoidea, and Mimospirina are amongst the most discussed extinct groups, and the gastropod nature of the Bellerophontida and Pelagiellida is still a frequently discussed problem. The phylogenetic relationships of the Macluritoidea and Mimospirina (Figure 2G), with sinistrally coiled shells, are uncertain and both groups may be sister groups to more advanced gastropods.

Evolution of the Gastropoda

The more than 500 million years of evolution of the Gastropoda is still poorly known. The main difficulties are that the phylogenetic positions and relationships of extinct gastropods can be inferred only from the limited number of characters observable in their fossilized hard body parts (i.e., shell and operculum). However, the number of extinct gastropod species and genera is much higher than those living. In addition, some belong to extinct higher taxa of family or order levels with unknown anatomy. Another complication is the development of similar shells in unrelated groups (homoplastic similarity) which has been documented in many living gastropods.

Origin and Early History of the Gastropoda

Since 1970, many new mollusc-like fossils from the Cambrian have been discovered (e.g., *Halkieria*, Merismoconcha, etc.). Their interpretation has given rise to different models of evolutionary relationships within the Mollusca. Even though these models are controversial, the Gastropoda has been generally accepted to be the sister group of the classes Cephalopoda (*see* **Fossil Invertebrates:** Cephalopods (Other Than Ammonites)) or Tryblidiida ('Monoplacophora'). The latter have been combined with the

Scaphopoda and Bivalvia (*see* **Fossil Invertebrates: Bivalves**) within the group Conchifera, which unites the higher Mollusca. Whether or not the Conchifera is monophyletic is uncertain.

Torsion of the soft body has been considered to be one of the main diagnostic characters of the Gastropoda. For this reason, the majority of the models of gastropod origin have been based on different interpretations of this anatomical feature in the extinct gastropod-like molluscs. The Early Palaeozoic Helcionelloida, Bellerophontida (Figures 2L and 2M), and Tryblidiida, with bilaterally symmetrical shells, as well as the Pelagiellida and Macluritida, with asymmetrically coiled shells, are the most frequently discussed groups, and have been variously interpreted as untorted or torted molluscs. However, there is no reliable method of recognizing torsion in extinct fossil molluscs. Thus, the unknown nature of the bodies in the Early Palaeozoic gastropod-like fossils has enabled controversial speculations to be made about the origin of the Gastropoda. Generally, it is accepted that the first undoubted gastropods appeared in the Late Cambrian.

Palaeozoic Era

During the Early Ordovician radiation, the diversity of gastropod groups which had appeared in the Late Cambrian (Archaeogastropoda, Euomphaloidea, Macluritoidea, Mimospirina, Peruneloidea) rapidly increased (Figure 9). The Macluritoidea with large shells, together with different groups of Archaeogastropoda, Euomphaloidea, Bellerophontida, and Mimospirina, were typical elements of gastropod faunas of the tropical regions. In contrast, higher latitude faunas were composed mainly of the Bellerophontida and Archaeogastropoda. This arrangement survived until the early Middle Ordovician, when the diversity of some groups (Macluritoidea and Euomphaloidea) decreased and some new groups appeared (the slit-lacking Archaeogastropoda, Subulitoidea, Platyceratoidea, Loxonematoidea, etc.). During the Middle Ordovician, gastropod diversity rapidly increased and, in the Late Ordovician, reached its maximum. Middle and Late Ordovician faunas consisted of members of all the main groups of Palaeozoic gastropods, except the Heterobranchia (Figure 9). The end of the Ordovician saw a dramatic decrease in gastropod diversity, as well as the extinction of the Macluritoidea.

The Silurian was a period of increasing diversity of many gastropod groups (e.g., Archaeogastropoda, Bellerophontida, and Platyceratoidea), when some gastropods in all marine communities continually increased, together with an increase in the morphological variability of their shells. This suggests an increase in their ecological adaptation to specific

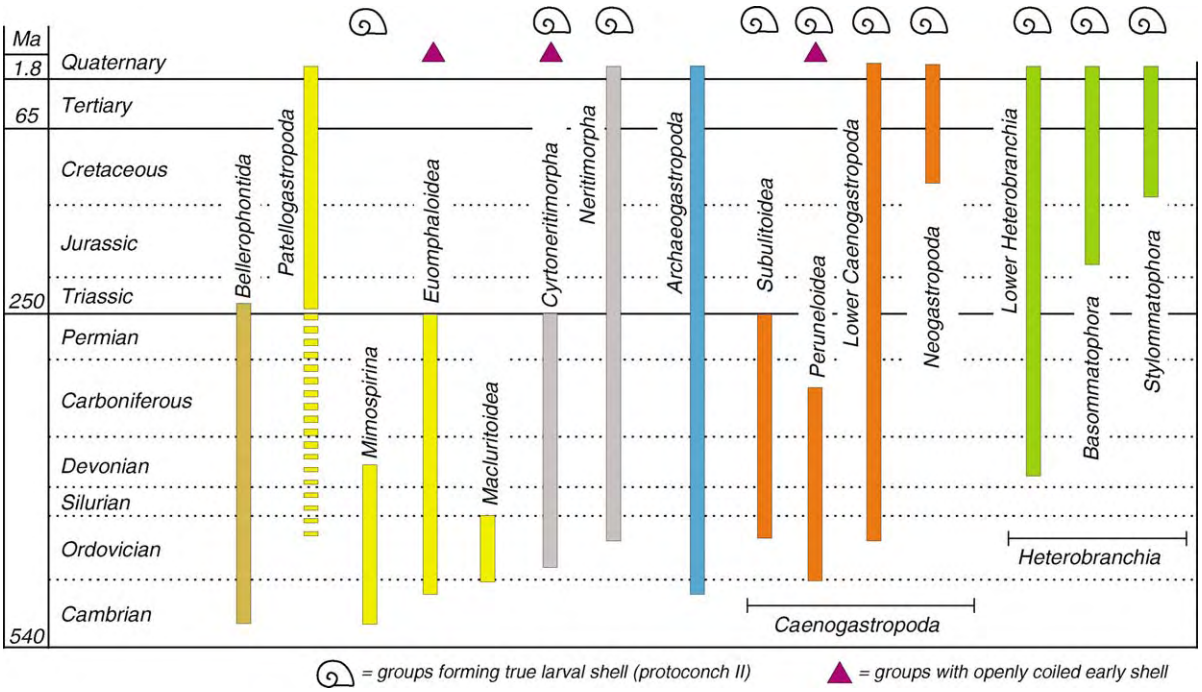


Figure 9 Diagram illustrating the evolution of the Gastropoda and the types of early shell ontogeny. The bars show the stratigraphical ranges of each gastropod group. Based on Bandel K (1997) Higher classification and pattern of evolution of the Gastropoda. *Courier Forschungsinstitut Senckenberg* 201: 57 81, and Frýda J and Rohr DM (2004) Gastropoda, 184 195 In: Webby BD, Droser ML, Paris F, and Percival IG (eds.) *The Great Ordovician Biodiversification Event*, pp. 408. New York: Columbia University Press.

environments. In comparison with the Ordovician, during the Silurian some gastropods with high-spired shells (mainly Loxonematoidea, Murchisonoidea and Subulitoidea; Figure 2) increased considerably.

The Devonian was a time of distinct changes in marine gastropod communities. Some Ordovician–Silurian groups became extinct (Mimospirina; Figure 2G), new groups appeared (Heterobranchia), and many groups underwent rapid radiation and specialization (Caenogastropoda and Neritimorpha). Thus, the Devonian faunas contained representatives of all extant gastropod orders (Archaeogastropoda, Neritimorpha, Caenogastropoda, and Heterobranchia), as well as many Palaeozoic groups (Figure 9). The Devonian was also the time when the protoconch morphology of several gastropod groups underwent considerable change. Gastropods with openly coiled protoconchs (Perunelomorpha, Cyrtoneritimorpha, and Euomphalomorpha; Figures 2H, 2P, 3D, and 3E) formed a considerable, sometimes even dominant, part of the Ordovician and Silurian gastropod communities. During the Early Devonian, their numbers rapidly decreased and none survived the Permian/Triassic extinction.

Carboniferous and Permian faunas had a similar composition of marine gastropod communities. A characteristic feature of Late Palaeozoic gastropod faunas was the fast radiation of different groups of Apogastropoda. The dominance of diverse groups of Caenogastropoda with high-spired shells (mainly Ctenoglossa and Cerithiomorpha) and Heterobranchia (Allogastropoda) was typical of shallow-water, muddy bottom communities.

Mesozoic and Cenozoic Eras

The Permian/Triassic crisis affected gastropods as well as all other marine animals. The Euomphalomorpha and Cyrtoneritimorpha (Figures 2 and 3), as well as many groups of Archaeogastropoda, Neritimorpha, and Caenogastropoda, became extinct. During the Triassic, the last members of the Bellerophontida disappeared. The Late Triassic was a time of fast radiation of neritimorphs (Neritopsoidea and Neritoidea), caenogastropods (Ctenoglossa, Cerithimorpha, Architaenioglossa, and Littorinimorpha), and heterobrachs (Allogastropoda and Archaeopulmonata). From Triassic strata, the oldest limpets of the subclass Patellogastropoda are documented. The Patellogastropoda is considered to represent the most ancient gastropod group, but their ancestors (probably bearing coiled shells) have not yet been recognized amongst Palaeozoic gastropods (Figure 9).

The composition of the Jurassic and Early Cretaceous marine gastropod faunas was roughly the same as in the Late Triassic. The characteristic feature of Mesozoic and Cenozoic gastropods was the development of more ornamented shells in most groups, as well as the lesser occurrence of openly coiled shells, by comparison with Palaeozoic gastropods (Figures 1 and 2). Both macro-evolutionary trends have been interpreted as adaptation to increasing predation activities by other animals. During the Cretaceous, more advanced caenogastropod groups (higher Mesogastropoda and Neogastropoda) appeared (Figure 9), which underwent fast radiation and diversification after the Cretaceous/Tertiary faunal crises. Both groups developed the possibility of extending their planktotrophic larval stages and, from the beginning of the Tertiary, they formed one of the dominant groups of marine gastropods. During the Cretaceous, some gastropods (lower Heterobranchia) started to reduce their shells, enabling their adaptation to holoplanktic life (e.g., pteropods). The Early Cenozoic marine gastropod faunas are very similar to extant gastropods in higher taxonomic composition.

Evolution of Freshwater and Terrestrial Gastropods

In contrast with marine gastropods, the fossil record for freshwater and terrestrial forms is less complete, limiting our knowledge of their evolution. Successful invasion to freshwater and land habitats has been closely linked with the mode of gastropod reproduction. External fertilization, which occurs in the ancient Patellogastropoda and Archaeogastropoda, limited them to marine environments. The freshwater and terrestrial environments were colonized by gastropods with egg capsules and internal fertilization (Neritimorpha, Caenogastropoda, and Heterobranchia). Even though members of these groups are known from the Early (Neritimorpha and Caenogastropoda) or Middle (Heterobranchia) Palaeozoic (Figure 9), the first freshwater and terrestrial gastropods are recorded from Late Palaeozoic strata (Archaeopulmonata). The first freshwater Basommatophora appeared during Jurassic time and, in the Cretaceous, the Stylommatophora started their invasion of the land and soon became the most diversified group of terrestrial gastropods.

Glossary

Archaeogastropoda Group of extant gastropods.
Bellerophontida Extinct group of Palaeozoic molluscs with bilaterally symmetrical shells.
Caenogastropoda Group of extant gastropods.
embryonic shell (protoconch I) Gastropod shell formed during embryonic development.

Heterobranchia Group of extant gastropods.
heterostrophic Condition of the protoconch when its whorls coil in the opposite direction to those of the teleoconch.
homeostrophic Protoconch and teleoconch whorls coil in the same direction.
hyperstrophy Condition in which anatomically dextral animals occupy sinistrally coiled shells, and vice versa.
larval shell (protoconch II) Gastropod shell formed during larval development in members of the Neritimorpha, Caenogastropoda, and Heterobranchia.
lecithotrophic Form of development in which larvae use yolk in egg for their nutrition.
Mimospirina Extinct group of Early and Middle Palaeozoic gastropods with sinistrally coiled, homeostrophic shells.
Neritimorpha Group of extant gastropods.
operculum Lid-like structure used for closing of the aperture in gastropod shells.
Opisthobranchia Gastropod subclass of Thiele's classification.
orthostrophy Condition in which anatomically dextral (or sinistral) animals occupy dextrally (or sinistrally) coiled shells.
Patellogastropoda Group of extant gastropods with limpet-shaped shells.
periostracum Outer organic layer of gastropod shells.
planktotrophic Form of development in which free-swimming larvae use planktic organisms for their nutrition.
Prosobranchia Gastropod subclass of Thiele's classification.
protoconch Gastropod shell formed during larval and/or embryonic development.
Pulmonata Gastropod subclass of Thiele's classification.
teleoconch Post-larval gastropod shell.
trochophore larva Gastropod larva formed during early larval development which may be free swimming or occurs in egg capsules.
veliger Gastropod larva formed during later larval development before metamorphosis to post-larval stages.

See Also

Biological Radiations and Speciation. **Evolution**. **Famous Geologists:** Cuvier. **Fossil Invertebrates:** Molluscs Overview; Bivalves; Cephalopods (Other Than Ammonites). **Palaeoecology**. **Palaeozoic:** End Permian Extinctions.

Further Reading

Bandel K (1997) Higher classification and pattern of evolu tion of the Gastropoda. *Courier Forschungsinstitut* 201: 57 81.

Beesley PL, Ross GJB, and Wells A (eds.) (1998) *Mollusca: The Southern Synthesis, Part B, Fauna of Australia,* vol. 5, pp. 565 1234. Melbourne: CSIRO Publishing.

Bieler R (1992) Gastropod phylogeny and systematics. *Annual Review of Ecology and Systematics* 23: 311 338.

Fretter V and Graham A (1994) *British Prosobranch Mol luscs. Their Functional Anatomy and Ecology.* London: Ray Society.

Frýda J and Rohr DM (2004) Gastropoda, 184 195. In: Webby BD, Droser ML, Paris F, and Percival IG (eds.) *The Great Ordovician Biodiversification Event,* p. 408. New York: Columbia University Press.

Knight JB, Cox LR, Keen AM, *et al.* (1960) Systematic descriptions. In: Moore RC (ed.) *Treatise on Invertebrate Paleontology, Part I, Mollusca 1,* pp. I169 I324. Law rence, KS: Geological Society of America and University of Kansas Press.

Lindberg DR and Ponder WF (2001) The influence of clas sification on the evolutionary interpretation of structure a re evaluation of the evolution of the pallial cavity of gastropod molluscs. *Organisms, Diversity and Evolution* 1: 273 299.

Peel JS (1991) The classes Tergomya and Helcionelloida, and early molluscan evolution. *Groenlands Geologiske Undersoegelse, Bulletin* 161: 11 65.

Ponder WF and Lindberg DR (1997) Towards a phylogeny of gastropod molluscs: an analysis using morphological characters. *Zoological Journal of the Linnean Society* 119: 83 256.

von Salvini Plawen L (1990) Origin, phylogeny and classi fication of the phylum Mollusca. *Iberus* 9: 1 33.

von Salvini Plawen L and Haszprunar G (1987) The Veti gastropoda and the systematics of streptonerous Gastro poda (Mollusca). *Journal of Zoology* 11: 747 770.

Taylor JD (ed.) (1996) *Origin and Evolutionary Radiation of the Mollusca.* Oxford, New York, Tokyo: Oxford University Press.

Waren A and Bouchet P (1993) New records, species, genera, and a new family of gastropods from hydrothermal vents and hydrocarbon seeps. *Zoologica Scripta* 22(1): 1 90.

Wenz W (1938 1944) Gastropoda. In: Schindewolf OH (ed.) *Handbuch der Paläozoologie,* p. 1639. Berlin: Borntraeger.

Zilch A (1959 1960) Gastropoda; Teil 2, Euthyneura. In: Schindewolf OH (ed.) *Handbuch der Paläozoologie,* p. 834. Berlin: Zehlendorf.

Cephalopods (Other Than Ammonites)

P Doyle, University College London, London, UK

Introduction

Molluscs are some of the most successful organisms on Earth, and after arthropods they are the most diverse group of invertebrates. In common with all molluscs, cephalopods have developed from a simple molluscan archetype of a conical shell, a muscular foot, a mantle cavity housing gills, and a basic feeding mechanism or radula. In evolving from this structural blueprint, cephalopods have developed a buoyant shell and a differentiated head, with well-developed brain and heightened sensory apparatus, and have modified their foot to create a set of arms, often equipped with hooks or suckers, that capture prey and draw it to a set of beak-like jaws situated at the centre of arm crown.

Cephalopods have a long geological history, which extends back into the Late Cambrian (Figure 1). The search for the earliest cephalopod has created a list of contenders, many of them remarkably like the posited molluscan archetype, but with one significant difference – the first indications of a chambered shell, creating one of the most important characteristics of the cephalopods, the ability to be buoyant or free-swimming. The first tentative formation of chambering presumably gave the nascent cephalopods an adaptive advantage, allowing the otherwise unremarkable mollusc to rise above the seabed to escape predators. This humble beginning marks the start of the radiation of all cephalopods, which were to become important predators of the Palaeozoic and Mesozoic seas, being replaced by the rise of the bony fishes only in the Late Mesozoic.

Cephalopod Morphology

Within the broad envelope of the cephalopods there is a surprising range of morphologies and in some cases body plans. For example, whereas the vast majority of living cephalopods have either internal (endocochleate) shells or no shell at all, most fossil cephalopods are known from the remains of their external (ectocochleate) shells.

Ectocochleate cephalopods have a complex shell built to withstand water pressure and act as a buoyancy aid (Figure 2). Effectively the shell is a cone, and this cone may be coiled, although numerous bizarre shapes are known. The cone-like shell is primarily aragonitic, with an external organic coating or epithelium. Some authorities have suggested that this external shell had a range of pigments in life, and this is certainly true of the surviving species of *Nautilus*, which are characterized by an ornament of reddish brown and white 'stripes'. Some fossil examples, including orthocones from the Early Palaeozoic, display complex zig-zag patterns. In other forms, a range of knobs and ribs ornament the shell.

Internally, the shell is chambered (Figure 2A). There are two basic types of chamber: buoyancy chambers or camerae, and a body chamber housing the mantle cavity, head, and arms when retracted. That part of the shell containing the buoyancy chambers is known as the phragmocone, and within it the camerae are regularly spaced and separated by chamber walls known as septa. The septa are constructed from multiple layers of aragonite, like the remainder of the shell, and at their simplest they have the strong domed shape of a watch glass. In some cases, particularly at the margins, this shape may be modified, and the junction of each septum with the inner wall of the phragmocone – the suture line – may take a variety of shapes, from a simple sinuous curve through to an extremely convoluted shape. The latter form

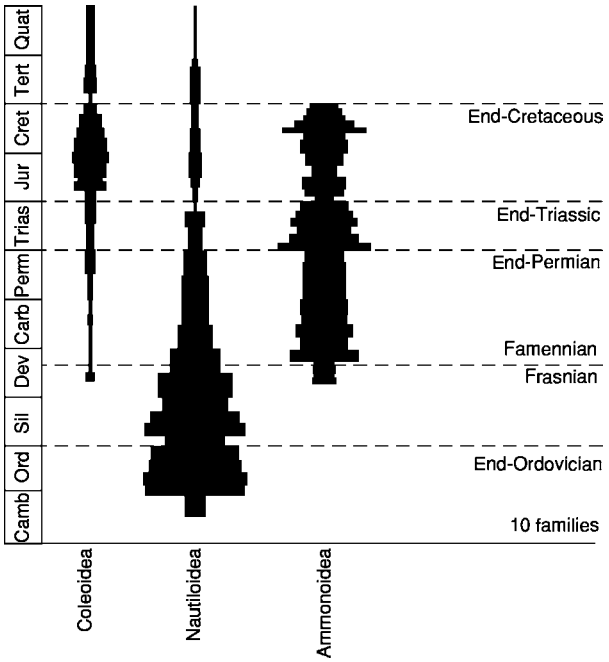


Figure 1 Ranges of the main cephalopod groups. Dotted lines indicate the main Phanerozoic extinction events. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

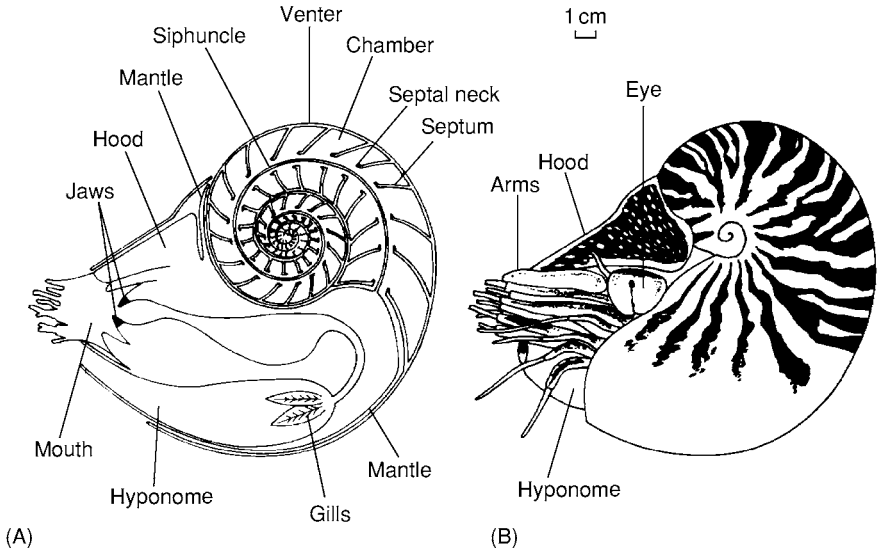


Figure 2 Form and structure of the living *Nautilus*. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

is more typical of the Ammonoidea (*see* **Fossil Invertebrates:** Ammonites); simpler sutures are found in the Nautiloidea and, where present, in the Coleoidea. Experiments have shown that the watch-glass shape imparts significant strength to the shell, and it is suggested that the water depth at which implosion will occur may be in excess of 500 m, based on experiments using living *Nautilus*.

The camerae are joined by a continuous porous pipe known as the siphuncle, which passes through each septum by means of a 'bottle-neck' arrangement, the septal neck (**Figure 2A**). This neck has many forms, most of which point forwards (prochoanitic) or backwards (retrochoanitic), with grades in between, depending on the taxon. Ultimately, the siphuncle joins all the chambers through the septal necks to the body chamber. These septal connectors are often modified with complex structures – difficult to interpret in a functional sense – which are common features of fossil nautiloids. In fact, apart from living *Nautilus*, which has a simple siphuncle, many of the higher nautiloid taxonomic groups are distinguished by such structures. Characteristically, the siphuncle is placed centrally in the living *Nautilus* and several nautiloid groups, but is marginal in other nautiloids and in ammonoids and shell-bearing coleoids.

The prime function of the camerae is to provide buoyancy, and it is the role of the siphuncle to regulate the mixture of gas and liquid within each camera in order to create a buoyant state. Regulation of buoyancy is such that during a diurnal cycle, based on recent studies, ectocochleate cephalopods are able to move through the water column. In some cases, the camerae can be modified to improve the buoyancy, with additional shell aragonite being laid down within each chamber – cameral deposits – to aid the balance of the shell. Similar deposits are found within the siphuncles of some fossil nautiloids – endosiphuncular deposits – although, again, they are absent from the relatively uncomplex *Nautilus*.

Cephalopod Classification

The cephalopods are generally treated as a class within the Phylum Mollusca (*see* **Fossil Invertebrates:** Molluscs Overview), but it is here that consensus ends. The class Cephalopoda is further subdivided into subclasses, but there is considerable debate about the relative ranks of the constituent cephalopod groups. For example, following the *Treatise on Invertebrate Paleontology* some authors propose a division of the Nautiloidea into four separate subclasses in their own right: Endoceratoidea, Actinoceratoidea, Orthoceratoidea, and Nautiloidea (in a restricted, yet still very broad, sense). A simple classification, involving subdivision into three more-or-less equal groups, is employed here.

Class Cephalopoda Cuvier, 1797 (Cambrian–Holocene). Fully marine free-swimming self-buoyant molluscs with a differentiated head and well-developed sensory apparatus.

Subclass Nautiloidea Agassiz, 1847 (Cambrian–Holocene). Cephalopods with coiled or straight (orthocone) ectocochleate shells with simple chamber walls; the siphuncle is often complex, marginal to central; living examples have multiple arms; tetrabranchiate.

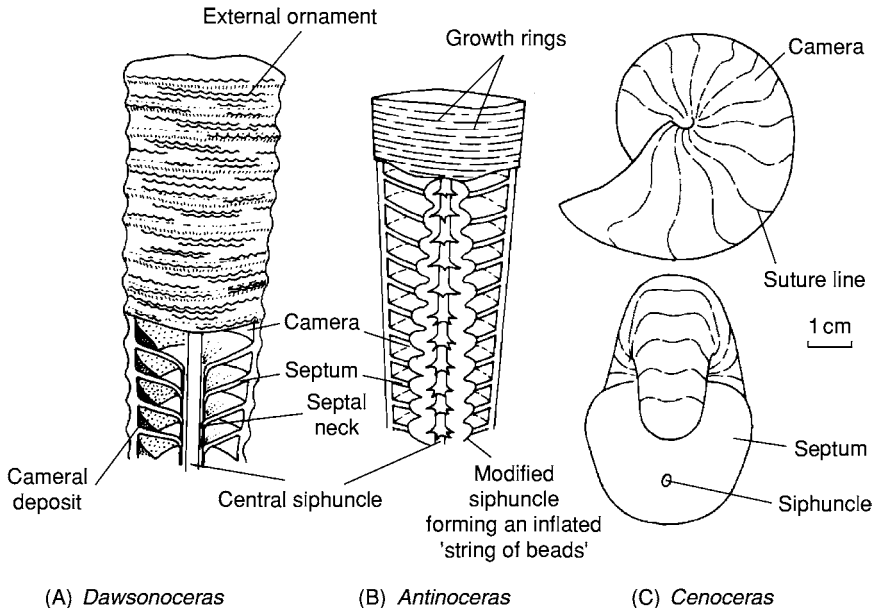


Figure 3 Fossil nautiloids: (A) the Silurian orthocerid *Dawsonoceras*, (B) the Ordovician actinoceratid *Actinoceras*, (C) and the Jurassic nautilid *Cenoceras*. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

Subclass Coleoidea Bather, 1888 (Devonian–Holocene). Cephalopods with endocochleate shells, which are chambered in some and reduced or absent in others; ten or eight arms grouped around the mouth; dibranchiate.

Subclass Ammonoidea Zittel, 1884 (Devonian–Cretaceous). Cephalopods with mostly coiled ectocochleate shells with complex suture walls; siphuncle marginal; number of arms uncertain; number of gills uncertain, presumed tetrabranchiate. This group is dealt with in detail elsewhere in this volume (*see* **Fossil Invertebrates:** Ammonites) and includes the earliest ancestors, the Bactritida.

Subclass Nautiloidea

The Nautiloidea is an important fossil group, represented today by the single living genus *Nautilus*, with its six known species or subspecies, compared with an estimated 11 000 fossil examples. This 'living fossil' is in effect the lone survivor of a group that achieved its maximum diversity in the Palaeozoic, declining through the Mesozoic to the present day (Figure 1). The Nautiloidea are representative of the first true cephalopods, and exhibit a rapid increase in diversity following widespread extinctions at the close of the Cambrian. Following the Ordovician–Silurian maximum, the nautiloids were in steady decline, such that, after the extinctions at the Triassic–Jurassic boundary, only two lineages remained extant; they survived until further extinctions at the Miocene–Pilocene boundary left the few species we know today.

The living genus *Nautilus* is one of the most powerful tools for interpreting the morphology and life history of the nautiloids, although this approach is fraught with difficulties given that *Nautilus* is the sole survivor of a rich diversity of forms from the geological past. This is particularly relevant given the range of form in such aspects as the siphuncle, particularly its size, form, and structure (Figure 3); modern-day *Nautilus*, in common with other members of its restricted group, has only simple structures, which are inadequate to explain the full diversity of the structures found in fossil representatives. The nautiloid shell is ectocochleate and is known in all forms from a straight cone (orthocone), through a curved cone (cyrtocone), to the typical tight planispiral coil of the living *Nautilus*.

Significant structural features of the nautiloid shell are the forms of the siphuncle, its septal necks, and connecting rings, with the connectors being highly variable, including everything from simple tubes (as in *Nautilus*) to highly complex examples with cone structures and endosiphuncular deposits (as in the oncocerids). The typical position of the nautiloid siphuncle is central or subcentral to the camerae, but again this is highly variable, with marginal siphuncles apparent in a number of groups (Figure 3). Its relative worth in taxonomic studies is still under active debate. Cameral deposits and endosiphuncular deposits also vary within the group.

The higher classification of the Nautiloidea is a strongly debated area, hinging mostly on the taxonomic position of the Endoceratidea, the

Actinoceratidea, the Orthoceratidea, and the Nautiloidea proper, which have been treated variously as subclasses, superorders, and orders. Here they are treated as natural but informal groups of sub-subclass rank.

Endoceratids This group comprises mostly medium-sized orthoconic shells, with some rarer cyrtoconic examples. Siphuncles are large and, unusually, marginal rather than central. Septal necks are of variable lengths, generally getting longer with time. The endoceratids typically have deposits within the siphuncle itself, often in the form of conical sheaths. The group is known from the Lower Ordovician to the Silurian, although it is commonest in Lower–Middle Ordovician deposits.

Actinoceratids Actinoceratids are generally orthoconic with short septal necks and a broad siphuncle. The siphuncle segments between the septal necks display many specialized structures, as illustrated by *Actinoceras* itself (**Figure 3**). Most actinoceratids have cameral deposits. They are known from the Middle Ordovician to the Upper Carboniferous, but are commonest in the Middle and Upper Ordovician.

Orthoceratids Orthoceratids are characterized by their orthoconic or weakly crytoconic slender shells. The genus *Orthoceras* is typical, but this group has been used by many authors to house many similar forms. The orthoceratids have narrow siphuncles, and they typically have both cameral and endosiphuncular deposits, as in *Dawsonoceras* (**Figure 3**). Orthoceratids are known from the Lower Ordovician to the Triassic, but are commonest in the Middle Ordovician–Silurian.

Nautiloids Nautiloids are diverse, with a great variety of shell forms, from orthoconic to tightly coiled. The siphuncle is variable in position, form, and size. Siphuncular and cameral deposits are similarly variable, present in some and absent in others. The earliest representatives include ellesmerocertids, which are cyrtoconic with marginal siphuncles and found in Upper Cambrian to Upper Ordovician rocks; oncocerids, which have generally short (breviconic) shells in a variety of shapes, usually with thin septa, recumbent septal necks, and expanded siphuncles, usually with endosiphuncular deposits, and are found in Middle Ordovician to Lower Carboniferous rocks; discordids, which have short stout cyrtoconic shells and thick siphuncles with endosiphuncular deposits, and are found in Middle Ordovician to Devonian rocks; tarphycerids, which consist of coiled shells with broad siphuncles and are found in Lower Ordovician to Upper Silurian rocks; and, finally, the nautilids themselves, which have a wide variety of coiled shells in which the siphuncle is narrow and cylindrical, and which range from the Early Devonian to the Holocene, some with a close similarity to *Nautilus*, such as the Jurassic genus *Cenoceras* (**Figure 3**; compare with **Figure 2**).

Subclass Coleoidea

The Coleoidea is a diverse group of cephalopods which are united by the possession of an internal or endocochleate shell. The vast majority of cephalopod species alive today are coleoids, and most of these have a greatly reduced internal shell, a 'torpedo' body shape, a well-developed head with chitinous beaks, and two gills in the mantle cavity. The geological history of the Coleoidea has been a subject of debate: the earliest definite coleoid is of Carboniferous age, although there are some doubtful records in the Devonian (**Figure 1**).

The diversification of the Coleoidea took place in the Mesozoic. Three superorders, based on aspects of the shell and soft parts, are recognized, although there is still considerable debate about the higher subdivisions of the group as a whole. The three superorders are Belemnoidea (characterized by ten subequal arms and a multi-layered conotheca), Octobrachia (characterized by ten arms in which the second pair is modified or lost), and Decabrachia (characterized by ten arms in which the fourth pair are modified as tentacles). Representatives of each of these groups have substantial internal skeletons.

The groups considered here are Superorder Belemnoidea (aulacocerids and belemnites), Superorder Decabrachia (spirulids, sepiids, and teuthids), and Superorder Octobrachia (vampyromorphs and octopods).

Superorder Belemnoidea This group comprises ten-armed cephalopods with an internal shell consisting of a multiple-layered phragmocone contained within a concentrically layered structure known as the rostrum (sometimes referred to as the telum in aulacocerids) (**Figure 4**). This group first appeared in the Devonian and became extinct at the end of the Cretaceous. It contains two major orders: the Aulacocerida, characterized by an entire body chamber and an aragonitic or organic rostrum; and the Belemnitida, characterized by a ventrally open body chamber – the pro-ostracum – and a mostly calcitic rostrum (**Figures 4 and 5**). Belemnites with soft-parts preserved have been discovered in Jurassic rocks, particularly the Lower Jurassic of southern Germany (*Acrocoelites*, *Passaloteuthis*) and the Middle Jurassic of England (*Belemnotheutis*). These specimens allow the

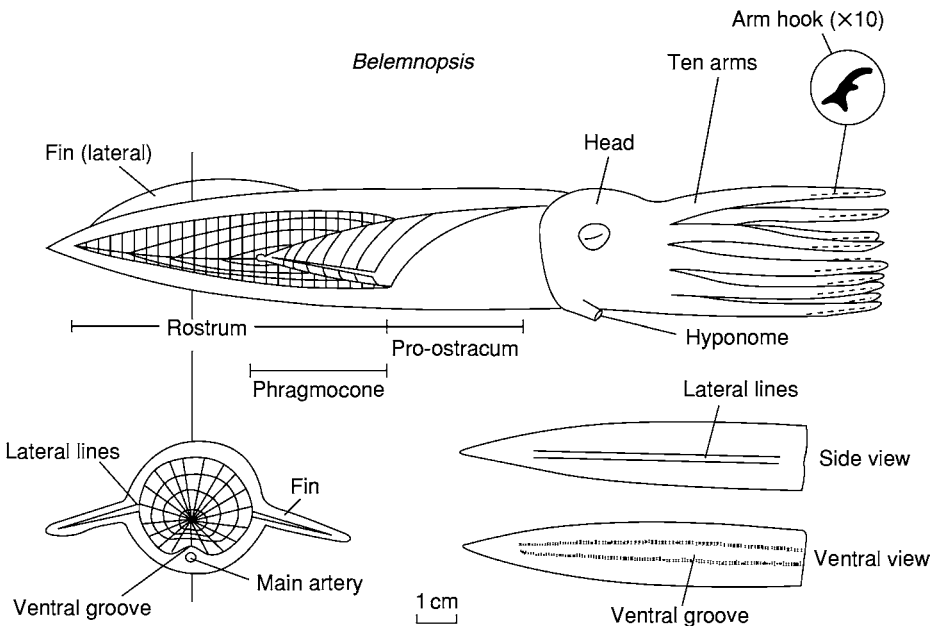


Figure 4 The belemnite shell, based on the Jurassic Cretaceous genus *Belemnopsis*. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

interpretation of the form and function of the belemnoid animal and in particular the nature of the proportional relationships between the shell and the soft body parts.

The belemnoid rostrum has a diverse structure and shell chemistry, but it is characterized by concentric incremental growth, the rostrum developing posteriorly from the first-formed primordial rostrum, associated with the protoconch. Continued growth leads to the gradual envelopment of the phragmocone in a conical cavity, the alveolus (Figure 4). The overall similarity in form, ontogeny, and function of all belemnoid rostra highlights the probable evolutionary relationship between the aulacocerid ancestors and belemnitid descendants.

The Aulacocerida is the most primitive group of the Belemnoidea. It first appeared in the Devonian or Carboniferous, and persisted until the Jurassic. Aulacocerids are often distinguished in their host rocks by the presence of large phragmocones, often with a low apical angle (belemnitids usually have a greater apical angle). Phragmocones complete with their outer wall (conotheca) display growth lines that indicate the presence of a complete body chamber, which distinguishes them from the Belemnitida, which have a ventrally open vestige of a body chamber, the pro-ostracum (Figures 4 and 5). Aulacocerid rostra are usually aragonitic and may be extensively furrowed or smooth, leading to some confusion with true belemnites.

The Belemnitida probably evolved from the Aulacocerida in the Early Jurassic, and belemnites are common fossils in Mesozoic sedimentary sequences. However, there are belemnite records from earlier Mesozoic (Triassic), Palaeozoic (Carboniferous–Permian), and Palaeogene rocks, although these are probably representative of other coleoid and even non-cephalopod groups. The belemnite rostrum typically consists of a concentric arrangement of growth rings, which alternate between organic-rich and organic-poor calcareous layers. The original mineralogy of the concentric layers has been a subject of some debate, with both low-magnesium calcite and aragonite being favoured, while some belemnites have been described with a primarily organic rostrum, but these are in need of further study. It is now generally held that the belemnite rostrum is composed of original low-magnesium calcite. However, the first-formed (primordial) rostrum of at least one taxon (*Hibolithes*) is originally aragonitic, and the sheath-like rostrum of the Jurassic belemnite *Belemnotheutis* is also known to be composed of primary aragonite.

The belemnite rostrum displays a variety of shapes and sizes, but most are based on three simple shapes: conical (e.g. *Acrocoelites*), cylindrical (e.g. *Cylindroteuthis*), and club/spear shaped (variously described as hastate, clavate, or lanceolate; e.g. *Hibolithes*), and these have influenced classification. In some rare cases belemnites appear blade-like (e.g. *Duvalia*). The rostrum displays three basic patterns of grooves and furrows (Figure 4): indistinct lateral 'lines' (thought to be the locations of fins), ventral and/or dorsal grooves (thought to be the locations of

Figure 5 Well preserved Jurassic belemnite *Cylindroteuthis*, with rostrum, phragmocone and pro ostracum. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

major blood vessels), and vascular marks (found only in belemnotellids and indicative of fine blood vessels). The primary function of the rostrum has traditionally been considered to be a counterweight, a support for musculature and/or fins, a protection for the phragmocone, or a streamlining for the backwards motion of the animal. The majority of authors favour the counterbalancing hypothesis for both belemnites and aulacocerids, and the existence of club-like rostra in particular is relevant to this hypothesis, as the club shape shifts mass posteriorly, thereby increasing the counterbalancing potential.

Superorder Decabrachia The Superorder Decabrachia comprises ten-armed coleoids in which the fourth pair of arms are modified as tentacles for the efficient capture of prey. The group is currently divided into three main orders: the Spirulida, Sepiida, and Teuthida. These encompass the living shelled cephalopods *Spirula* and *Sepia*, in which the shell is plesiomorphic, and the squids, most of which have the apomorphy of reduced shells. The geological history of the group stretches back to the Jurassic, the oldest known representative being the squid *Plesioteuthis* from the Late Jurassic, while the genera *Groenlandibelus* and *Naefia* are the oldest known representatives of the shelled spirulids.

The living genus *Spirula* is characterized by a loosely coiled chambered phragmocone, which is supported by the internal tissues of the body and which provides primary buoyancy. The life attitude of this genus approximates to vertical, with the head and arms hanging from the buoyant posterior, which contains the phragmocone. There are few traces of a rostrum or sheath, or cameral deposits, both of which would act as a counterbalance for the buoyant phragmocone. The geological record of the Spirulida extends back to the Cretaceous. Fossil representatives of the group demonstrate that the coiling of the spirulid phragmocone and its lack of a rostrum are apomorphic: the majority display only weak endogastric curvature and an aragonitic sheath-like rostrum.

The Recent genus *Sepia* has attracted much interest with regard to its mechanism for buoyancy, but fossil sepiids are less well known. The sepiid shell is entirely aragonitic and comprises a dorsal shield, which encompasses the ventrally open phragmocone; the multiple septa are closely spaced, buttressed by a series of pillars that demonstrate annular incremental growth, and the siphuncle is open to the internal tissue of the cuttlefish. The sepiid shell is fragile, and the septa are rarely preserved except in exceptional circumstances, and this has led to a number of debates about the relative antiquity of the group. Current estimates suggest that true sepiids extend as far back as the Cretaceous (with the Actinosepiidae), although claims that the Late Jurassic genus *Trachyteuthis* (**Figure 6B**) is a sepiid have recently been refuted. The evolutionary development of the sepiids can be traced to spirulid ancestors, with progressive loss of a ventral margin to the shell.

Living teuthids, encompassing many genera of squid, are characterized by an organic gladius, composed of β-crystalline chitin, which has no vestige of a phragmocone structure. The development of the gladius led to the loss of primary buoyancy: squids rely entirely upon the motion of the body for support in the water column. The geological history of the Teuthida extends back to the family Plesioteuthididae, known from the Jurassic. This fossil teuthid has a calcitic gladius, although some geochemical investigations have demonstrated a composition of primary francolite. Although some authors have suggested that this family is in fact a representative of the Vampyromorpha, other studies have indicated the homology of the fossil gladii of this family with living teuthids.

Superorder Octobrachia The Superorder Octobrachia comprises ten-armed coleoids in which the

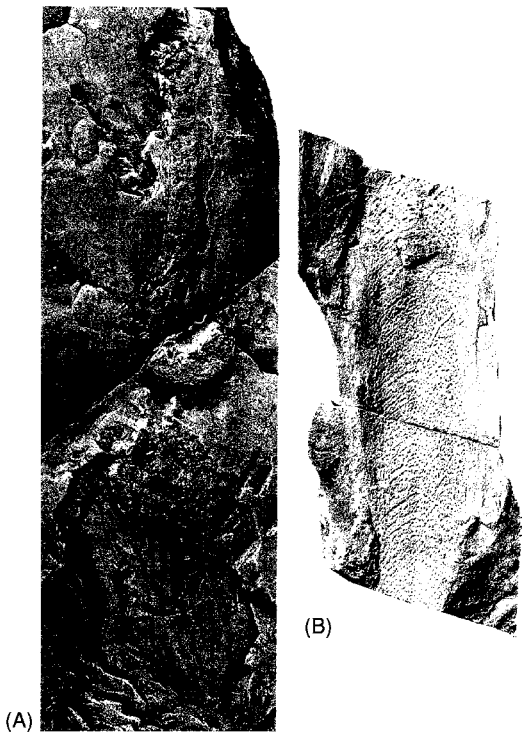


Figure 6 Jurassic vampyromorphs: (A) the Lower Jurassic loligosepiid *Jeletzkyteuthis*, complete with ink sac; and (B) the Upper Jurassic *Trachyteuthis*. (Reproduced from Doyle P (1996) *Understanding Fossils*. Chichester: John Wiley.)

second pair of arms is either modified or lost and there is no trace of a phragmocone forming a primary buoyancy aid. The Octobrachia includes the octopods, with little or no shell, the archaic order Vampyromorpha, and fossil representatives of the Loligosepiina and related families (Teudopsidae and Trachyteuthididae) (Figure 6A). These fossil groups are characterized by a calcareous shell, although some authors have argued for primary francolite and a phosphatic shell.

The geological history of the Octobrachia extends back to at least the Triassic with the first Loligosepinids. However, *Pohlsepia mazonensis* from the Carboniferous Mazon Creek Lagerstätte may, despite its ten arms, be the earliest known octopod. This view is based on the facts that it does not possess a well-defined head, it has a modified arm pair, and it has a sac-like body. *Palaeoctopus* from the Late Cretaceous of Lebanon is an undoubted octopus and has much in common with *Pohlsepia*; it has no identifiable shell, eight subequal arms (with suckers), a sac-like body, a poorly defined head, and free posterior fins.

The living genus *Vampyroteuthis* is an imperfectly known deep-sea coleoid that combines the octobrachian character of eight arms (the second pair of the normal ten-armed configuration being lost) with the essentially teuthid character of an internal gladius, which is not seen in other octobrachians. Like the teuthid gladius, the vampyromorph gladius is composed of β-crystalline chitin, and it corresponds in form to the ladle-like gladius, providing support for the visceral mass of the animal. Fossil representatives of the Teudopsidae and Trachyteuthididae have similar gladii, presumably with the same function. The Loligosepiina, also recognized as possessing vampyromorph affinities, have an arrow-head gladius, which may have had a role in positioning the head during its piston movements on swimming.

See Also

Fossil Invertebrates: Molluscs Overview; Bivalves; Gastropods; Ammonites. **Mesozoic:** Jurassic; Cretaceous.

Further Reading

Bandel K and Leich H (1986) Jurassic Vampyromorpha (dibanchiate cephalopods). *Neues Jahrbuch für Geologie und Paläontologie Monatshefte* 1986: 129 148.

Bandel K, Engeser T, and Reitner J (1984) Die Embryona lentwicklung von Hibolithes (Belemnitida, Cephalo poda). *Neues Jahrbuch für Geologie und Paläontologie Abhandlungen* 167: 275 303.

Donovan DT (1977) Evolution of the dibranchiate Cepha lopoda. *Symposis of the Zoological Society of London* 38: 15 48.

Doyle P (1993) Mollusca, Cephalopoda (Coleoidea). In: Benton MJ (ed.) *The Fossil Record 2*, pp. 229 236. London: Chapman and Hall.

Doyle P (1999) The functional morphology of the coleoid shell. In: Savazzi E (ed.) *The Functional Morphology of the Invertebrate Skeleton*, pp. 327 347. Chichester: John Wiley.

Doyle P, Donovan DT, and Nixon M (1994) Phylogeny and systematics of the Coleoidea. *University of Kansas Paleontological Contributions, New Series* 5: 1 15.

Dzik J (1984) Phylogeny of the Nautiloidea. *Palaeontolo gica Polonica* 43: 1 219.

Engeser T (1988) Fossil 'octopods' a critical review. In: Clarke MR and Trueman ER (eds.) *The Mollusca, Volume 12. Paleontology and Neontology of Cephalo pods*, pp. 81 87. San Diego: Academic Press.

Engeser T (1990) Phylogeny of the fossil coleoid Cephalo poda (Mollusca). *Berliner Geowissenschaftlichen Abhandlungen A* 124: 123 191.

Engeser T (1998) Fossil Coleoidea Page. http://userpage.fu berlin.de/~palaeont/fossilcoleoidea/welcome.html.

Engeser T (1988) Fossil Nautiloidea Page. http://userpage.fu berlin.de/~palaeont/fossilnautiloidea/fossilnautpage.html.

Engeser T and Bandel K (1988) Phylogenetic classification of coleoid cephalopods. In: Wiedmann J and Kullmann J (eds.) *Cephalopods present and past*, pp. 105 115. Stuttgart: E. Schweizerbart'sche verlagsbuchhandlung.

Fischer AG and Teichert C (1969) Cameral deposits in cephalopod shells. *University of Kansas Paleontological Contributions* 37: 1 30.
Holland CH (1987) The nautiloid cephalopods, a strange suc cess. *Journal of the Geological Society, London* 144: 1 15.
Jeletzky JA (1966) Comparative morphology, phylogeny and classification of fossil Coleoidea. *University of Kansas Paleontological Contributions, Mollusca* 7: 1 162.
King A (1993) Mollusca, Cephalopoda (Nautiloidea). In: Benton MJ (ed.) *The Fossil Record 2*. London: Chapman and Hall.
Kluessendorf J and Doyle P (2000) *Pohlsepia mazonensis*, an early 'octopus' from the Carboniferous of Illinois, USA. *Palaeontology* 43: 919 926.
Macleod N (ed.) (2003) *Paleobase Macrofossils Part 2.0, Ammonoids, Bivalves, Coleoids, Gastropods and other Mollusca*. Oxford: Blackwell.
Riegraf W and Hauf R (1983) Belemnitenfunde mit Weichkörper, Fangarmen und Gladius aus dem Unter toarcium (Posidonienschiefer) und Unteraalenium (Opalinuston) Südwestdeutschlands. *Neues Jahrbuch für Geologie und Paläontologie Abhandlungen* 165: 466 483.
Sælen G (1989) Diagenesis and construction of the belem nite rostrum. *Palaeontology* 32: 765 798.
Teichert C (1988) Main features of cephalopod evolution. In: Clarke MR and Trueman ER (eds.) *The Mollusca, Volume 12. Paleontology and Neontology of Cephalo pods*. London: Academic Press.
Yochelson EL, Flower RH, and Webers GF (1973) The bearing of the new later Cambrian monoplacophoran genus *Knightoconus* on the origin of the Cephalopoda. *Lethaia* 6: 275 310.

Ammonites

G E G Westermann, McMaster University, Hamilton, ON, Canada

Introduction

The term 'ammonite' (from genus *Ammonites*, Latin = lover/team of Ammon) was derived from *cornu Ammonis* (Latin = Horn of Ammon) because of its resemblance to the curved horns of the ram-headed sphinxes at the Roman temple of Jupiter Ammon in the Ammon oasis of Libya, a latinized form of the much older Egyptian temple Amun-Re at Thebes.

'Ammonites' is the popular name for the shells of the extinct Subclass Ammonoidea (Class Cephalopoda), which lived from Devonian to Cretaceous times. They are the most important fossils for Mesozoic biostratigraphy and biogeography. The animal that secreted the protective, external shell presumably had ten arms, two of which may have been modified into tentacles, but there is no direct evidence for the structure of the soft parts. The body chamber ('living chamber') housed the soft parts, whereas the chambered, essentially empty phragmocone provided the uplift of the neutrally buoyant organism. Some ammonites resemble the shell of the extant *Nautilus*, but all differ from the Subclass Nautiloidea by their internal structure, i.e., folded septa and a marginal, thin siphuncle. *Nautilus* shell is structurally very similar to ammonite shell.

This article treats, in sequence, ammonite shape and architecture, phylogeny, growth and sexual dimorphism, buoyancy and poise, functional morphology, and ecology.

Shape and Architecture

Shell size is usually from 5 to 30 cm, but ranges from 1 to 150 cm. The most common shape is ammonitocone, a closed planar spiral with involute to evolute coiling and varying whorl section (Figures 1A–C). Later ammonoids included a great variety of heteromorphs (= other shapes), including orthocones (straight), ancylocone (hooked), gyrocones and cyrtocones (open-curved), and helicones (snail-like). All may be smooth or ornamented with ribs (costae/plicae), tubercles, or spines, often changing with growth. Prominence of ornament tended to be higher in inflated than in compressed forms and increased during ammonoid evolution. Ribs may be single, branching, intercalating, or fasciculating (bundled). Nodes are rounded and spines prominent. Constrictions are periodic, deep folds.

The shell consists of high-strength nacre (mother-of-pearl), a bicomponent material of aragonite platelets separated by organic sheets. Fully grown, adult shells are recognized by the shape of body chamber and peristome. The former is often modified by partial uncoiling and inflation, both lowering the centre of mass; the peristome usually thickens with age, and may carry lateral lappets or a ventral rostrum. Other ammonoids grew until death.

Internally, the shell is divided into the phragmocone providing buoyancy and the body chamber protecting the animal. The phragmocone has transverse

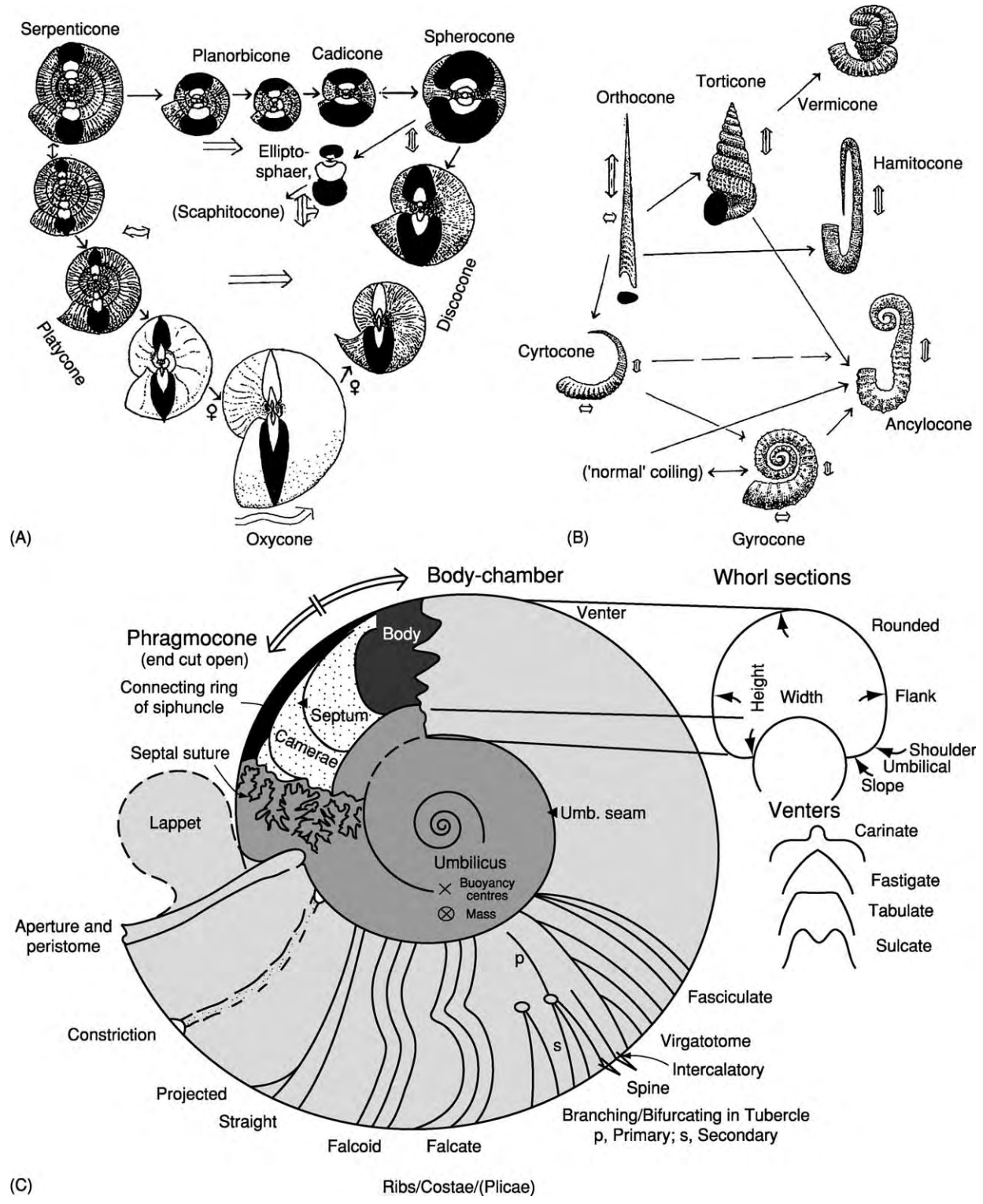


Figure 1 Ammonite morphology. (A) Planispiral shell shapes, illustrating the morphologic continuum (adapted from Westermann (1996)). (B) Major types of heteromorphs (adapted from Westermann (1996)). (C) Terminology of the ammonite shell; cut open at upper left to show interior of phragmocone.

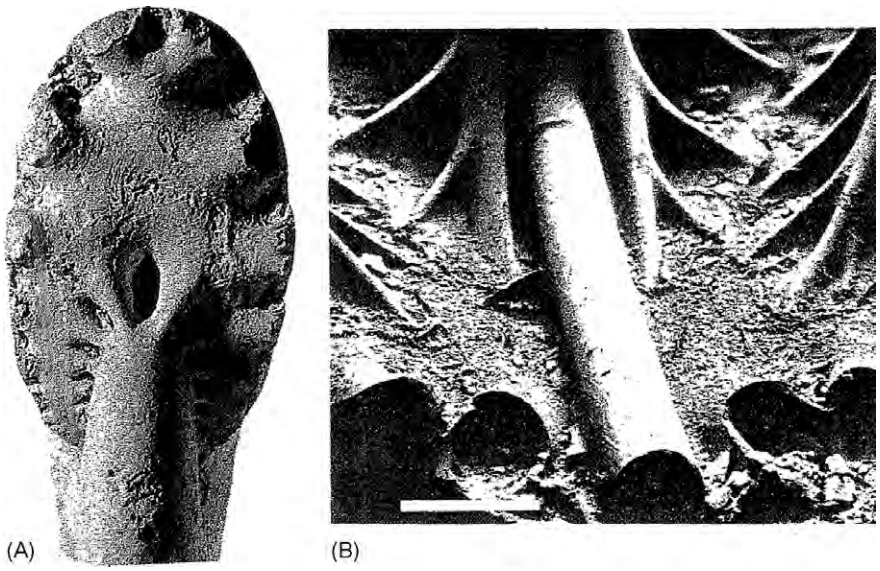


Figure 2 Ammonite septum and siphuncle. (A) Anterior view of the last septum of a Jurassic phylloceratid at natural size; from the central, flat region the flutes increase in prominence towards the margin but decrease on the narrow whorl overlaps; the positive flutes are the saddles, the negative ones the lobes. (B) Frontal view into an empty chamber of another Jurassic phylloceratid, scale bar 1 mm; shown are the siphuncle and 3rd and 4th order 'frills' of two septa that attach to the wall, forming the medium part of the external lobe. (Adapted from Westermann and Tjujita (1999).)

septa that separate the camerae. As in all living cephalopods, these contained gas at less than atmospheric pressure, together with small amounts of water (cameral liquid). The ventro-marginal siphuncle (but dorso-marginal in Clymeniida), with phosphatised organic connecting rings, connected all camerae to the body and once contained the siphon.

Ammonoid septa are folded into a series of anticlastic, concavo-convex arches or 'flutes' (Figure 2), which abut the wall in the sutures ('suture lines') consisting of a series of adorally concave lobes separated by convex saddles (Figure 3). Sutures with undivided lobes and saddles are called goniatitic and usually are of either Z type (for 'zigzag') or M type (for 'meander'). Both reduced the free (unsupported) wall space, the Z type by large, subtriangular lobes and saddles that telescope with neighbouring sutures; the M type by more numerous lobes and saddles with parallel sides, without overlap. Ammonitic sutures developed when the original (primary) flutes formed secondary flutes, called lobules and folioles, during evolution or ontogeny; further (3rd- and 4th-order) subdivision of lobules and folioles are typical of later ammonoids. This resulted in 'complex' sutures with fine 'frills'. Significantly, size and spacing of the sutural elements are highly organised, providing improved wall support compared with goniatitic sutures. Sutures have been used in classification for two centuries, and their function is much debated (see below).

The only commonly preserved structures not attached to the shell are the aptychi (Figure 4), which are strongly calcified lower jaws. Aptychi are found mainly in Late Mesozoic ammonites and consist of paired plates hinged together by the organic 'horny' layer forming the ancestral jaw. They more or less fitted the whorl cross-section near the aperture and acted as opercula or protective lids. During active periods they could be folded and withdrawn for possible duplicate use as lower jaws.

Phylogeny

The evolution of the Subclass Ammonoidea (Figure 5) began with slender orthocones and cyrtocones, the long-ranging (Silurian-Triassic) Bactritoidea, which evolved into Palaeozoic nautiloids and lived as plankton with a vertical poise. From them evolved the Devonian Anarcestida, which retained orthoconic innermost whorls before coiling and had sinuous to goniatitic-Z sutures. The Goniatitida were mostly smooth spherocones and discocones with goniatitic-Z sutures that became goniatitic-M and ammonitic near their Permian extinction. The short-lived, Late Devonian Clymeniida, with goniatitic-Z sutures and diverse shapes including subtriangular coiling, were unique in their dorso-marginal siphuncle. The Prolecanitida were commonly smooth, platyconic, and discoconic brevidomes, suggesting nektic habitats, with goniatitic-M or ceratitic sutures. They were ancestral

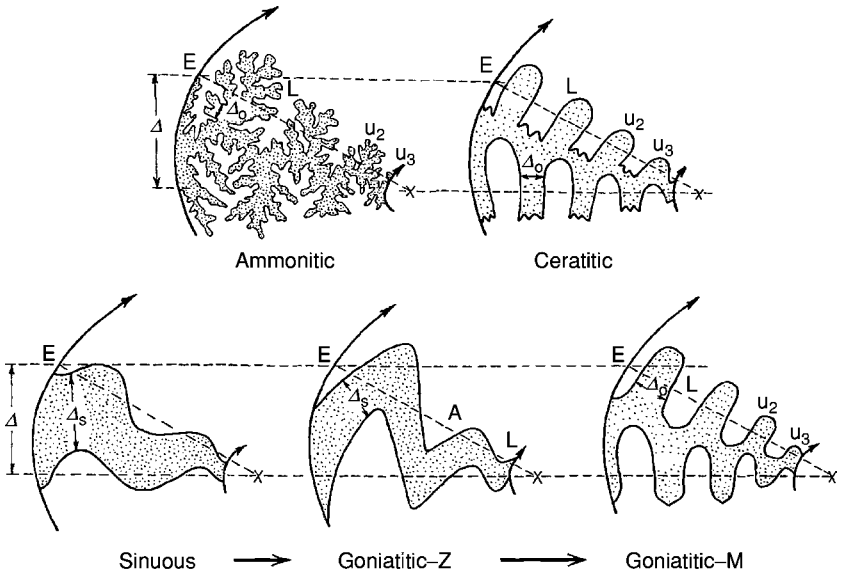


Figure 3 Schematic illustration of the major sutural types according to their function as wall supports, i.e., minimizing unsupported shell wall (between sutures Δ_s; between parts of same suture Δ_o), with septal spacing (Δ (Δ)) constant. Arrows indicate evolutionary trends. The lobes from venter to umbilical seam are: E, external; L, lateral; U, umbilical; A, adventive. Saddles are named after the lobes forming them, e.g., E/L saddle. (Adapted from Westermann (1975).)

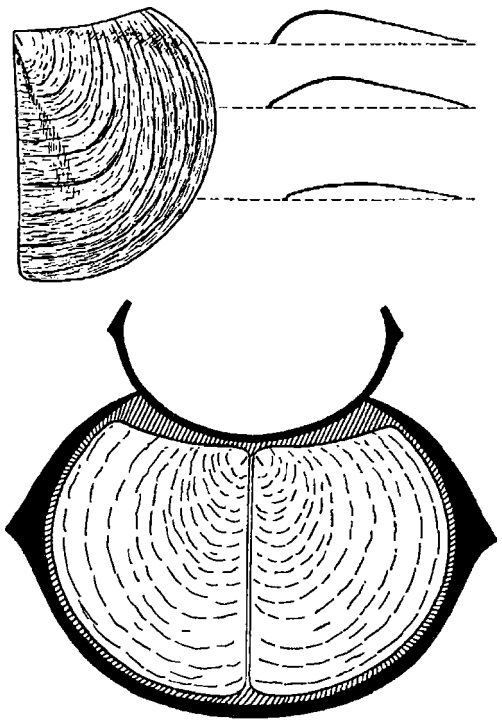


Figure 4 Jurassic aptychus (*Praestriaptychus*) and its recon struction as a hinged pair in the aperture of the associated ammonite (*Stephanoceras*, Ammonitida); note excellent fit. Space at dorsal overlap presumably enabled water intake (3X).

to all Mesozoic ammonoids, beginning with the mainly Triassic Ceratitida, which had ceratitic to ammonitic sutures, closely followed by the long-ranging Phylloceratida, with broadly rounded saddles or folioles. In the earliest Jurassic, the Prolecanitida gave rise to the Lytoceratida and Ammonitida; the former distinguished by bipartite (not tripartite) saddles and lobes; the latter being the most common, coiled and often ornamented ammonites with complex sutures. Near the end of the Jurassic arose the last ammonoid order, the Ancyloceratida, probably also from the Lytoceratida. They included a host of heteromorphs and, remarkably, began with uncoiled forms. All four orders became extinct at the great end-Cretaceous mass-extinction.

Growth, Longevity, and Sexual Dimorphism

Growth of the shell began with the ammonitella in the minute egg (1–2 mm). After hatching, the shell grew by terminal secretion at the persistome, so that all growth stages are preserved in the fully grown shell. During periodic growth, the posterior mantle secreted cameral liquid in front of the last septum, followed firstly by a membrane and secondly by the nacreous septum. On completion, the (incompressible) liquid, possibly a gel as in the cuttlefish *Sepia*, was withdrawn through the siphuncle into the body. Salinity differences in the epithelium of the vascular siphon generated sufficient osmotic pressure to transport the fluid through the porous connecting rings into the body cavity, against the hydrostatic pressure of the ambient seawater. The new camera was at first

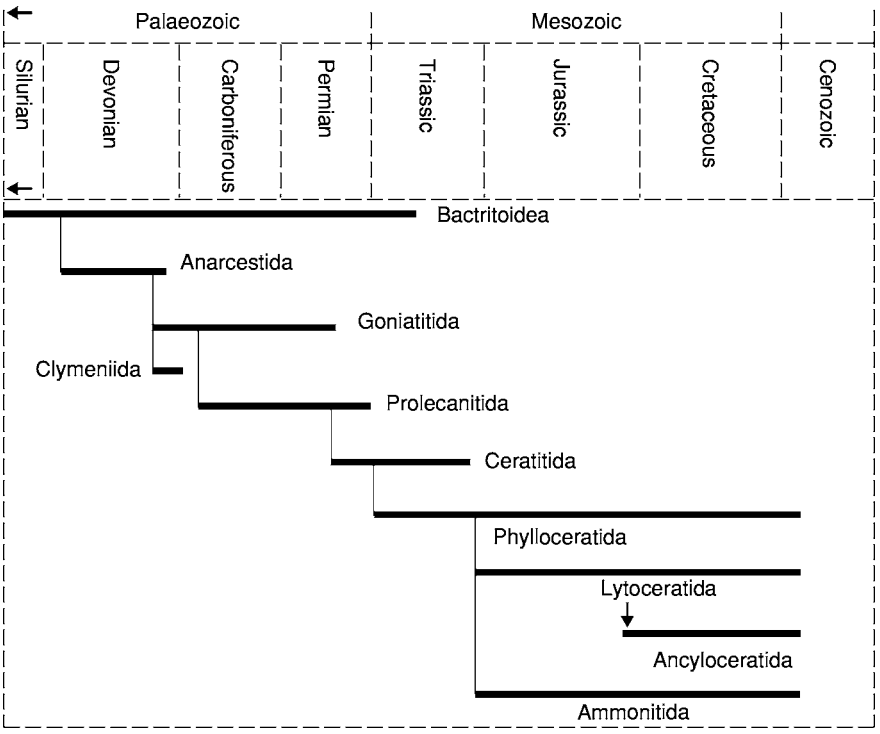


Figure 5 Phylogeny of the Ammonoidea orders.

under vacuum which was later partially replaced by gas at up to 1 atmosphere. Thus, the entire ambient pressure acted (directly) on the outer phragmocone wall and (indirectly through the soft body) on the latest/adoral septum risking its implosion, as well as on the organic connecting rings risking their explosion.

Growth of the pre-adult shell was mainly exponential (secretion time of each whorl increasing with the expansion rate W), followed by a slowing down, as indicated by the more closely spaced (approximated) last several septa. The total duration of growth depended on: (i) size; (ii) shell thickness, especially of the septa; (iii) diameter and wall thickness of the connecting rings: liquid transport depends on surface area and permeability; (iv) habitat depth: ambient pressure slowed cameral emptying; and (v) temperature: shell secretion is more rapid in warm than in cold water. The oceanic *Nautilus*, living at 200–400 m depth, is useful for comparison: at 20–25 cm adult diameter, the thickness of wall and septa are about 1 mm, and the connecting rings are narrow and thick-walled (i.e., strong); growth takes 10–15 years; adult/gerontic life is several more years. Most ammonoids were smaller or similar in size, had thinner septa and shell, broader and thinner-walled (i.e., weaker) connecting rings, and lived in mostly warm epeiric seas or as oceanic epiplankton at 30–150 m; growth took from 2 to 7 years. At the other extreme, some oceanic ammonoids of similar size lived at 250–500 m depth (e.g., mesopelagic lytoceratids, phylloceratids, desmoceratids), resembled *Nautilus* in shell and siphuncle properties, and presumably grew at similar slow rates. Some ammonoids, however, were much larger and became proportionately older, perhaps 50 years or more. Several more years of adult/gerontic life are indicated for ammonites with strongly modified peristomes. But in some ammonoids, especially among oceanic forms, growth was indeterminate up to death and their size became gigantic ('megaconchs').

For a century and a half, consistently associated pairs of distinctly modified shells have been observed in many Jurassic and Cretaceous Ammonitida and interpreted as sexual dismorphism. In extreme cases, the two forms, called antidimorphs, differed so strongly that they have been placed in different families. The most obvious distinction is in adult size, with the larger shell or macroconch usually two to three times larger than the smaller shell or microconch (Figure 6). Because in living animals with size ratios between the sexes of at least 2, the larger one is always the female; the macroconch is considered the female shell and the microconch the male shell. Furthermore, egg capsules have been found in macroconchs only. In the classic sexual dimorphism of Ammonitida, the peristomes differ in that only the

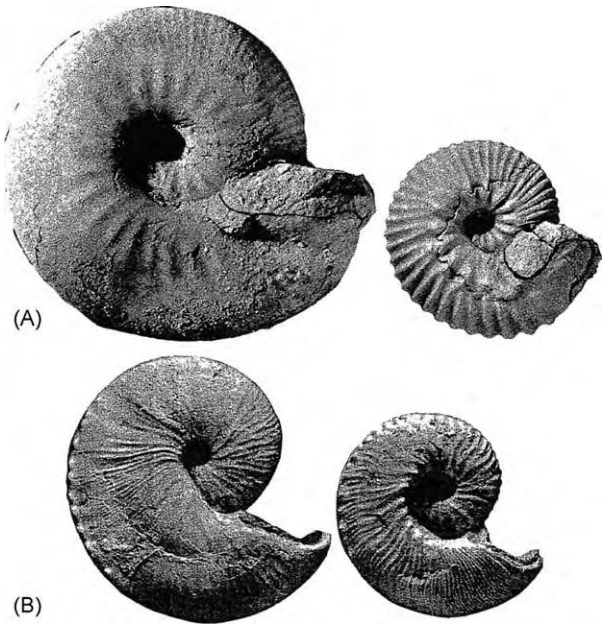


Figure 6 Sexual dimorphism. (A) Difference in size, body chamber and aperture in Jurassic Ammonitida (*Emileia giebeli* ♂ and ♀, peristomes traced). (B) Difference in size and body chamber in Cretaceous Ancyloceratida (*Hoploscaphites nicolletii*). (Adapted from Davis *et al.* (1996).)

microconchs or males have lateral extensions called lappets. In other Mesozoic ammonoids (e.g., Ancyloceratida), dimorphism is developed only in size and major to minor body-chamber modification, whereas dimorphism was rare in Palaeozoic ammonites and less clearly developed. Antidimorphs are recognized as conspecific because the inner whorls are identical and the animals have the same stratigraphic and geographic distributions. However, their numbers are often grossly mismatched; in many cases, the males are 10 (and up to 100) times more frequent, in others the females. The most probable causes for the mismatches are sexual segregation except during mating periods, and differences in postmortem drift (see below). Although the existence of sexual dimorphism is now universally acknowledged, at least for the Ammonitida, disagreement remains in classification; some ammonitologists still place all antidimorphs in different genera, even naming new ones; most place them in the same genus when specific pairing is impossible, and in the same species when pairing is highly probable. In the latter case, the sex symbols or 'macroconch (M)' and 'microconch (m)' are appended. The choice of classification is obviously of great consequence to phylogenetic reconstruction, biogeography, etc.

Disputed Functions of Ammonitic Septa and Complex Sutures

The main function of complex ammonitic sutures remains the most controversial topic among ammonitologist. The oldest ('orthodox') and still prevalent hypothesis resembles that of the Oxford professor William Buckland, who in 1827 concluded that this structure strengthened the shell (phragmocone) against implosion by ambient water pressure. *Nautilus*, which lives deeper than most ammonites did, has only simple, sub-hemispherical septa, but these are placed within well rounded, ovoid whorls with thick walls. Many typical ammonites, on the other hand, had flattened thin-shelled flanks with little (membrane) strength of their own, for example, platycones, oxycones, and compressed discocones. Their sutures were a functional improvement over goniatitic sutures by providing more closely spaced and somewhat elastic support for the weak, flat parts of the phragmocone wall (Figure 3), as well as some protection against predators. This prevented the shell from bending and fracturing (under tension) over stiff, distant sutures and from fracturing from tooth bites. Whorl section (e.g., laterally compressed for 'streamlining') was, therefore, the dominating factor, rather than habitat depth. Among related taxa with similar whorl shape, however, epeiric species often had simpler sutures (and weaker septa) than their oceanic relatives. Computer modelling and analysis (Finite-Elements) have shown that, with increasing complexity, the septa became stronger against the hydrostatic pressure transmitted through the body (as earlier theory had predicted). Recent claims to the contrary were based on faulty modelling.

Other hypotheses attempting to explain sutural complexity, include: (i) improved muscle attachment in the lobules and 'frills' for (a) better attachment of the mantle and/or muscle fibres, or (b) for a pre-septal gas bladder that could change overall buoyancy by expansion and contraction; (ii) improved liquid transport out of and/or into the chambers; and (iii) the retention of cameral liquid in the microcavities created by the lobules, folioles, and 'frills'. Hypotheses (i) and (ii) are not feasible because: (ia) The principal muscles were attached elsewhere; (ib) a preseptal gas bladder was inoperable because of basic gas laws and limitation of muscular forces; (ii) improved liquid transport would be ill served by marginal fluting because of difficult transport to the siphuncle, creating a 'bottleneck'; (iii) liquid storage was probably a *secondary* function of marginal fluting, keeping ballast water safe from swishing around.

Hydrostatics and Hydrodynamics: Reconstructing the Living Organism

Buoyancy, poise, stability, and bathymetry can be calculated for the shell and approximated for the living ammonoid, although the soft parts remain essentially unknown. The arms, which were presumably 10 in number, could not have been very muscular as in extinct coleoids (squid, belemnites, sepiids) which are well known from bituminous shales, and they must have been small to fit into the bodychamber (e.g., behind the aptychi). The head, arms, and hyponomes of ammonoids can therefore only be conjectured. They are here shown adapted to diverse feeding strategies (**Figure 7**).

The neutral buoyancy necessary for a free-floating organism, and which is present in all living cephalopods, requires strict correlations between shell thickness and the volumes of phragmocone and body chamber. The shell may be envisaged as a coiled cone: assuming similar shell thickness, the relative volumes and, hence, the relative lengths of phragmocone and body chamber must remain constant, whether the cone is slender or thick. When coiled into a logarithmic spiral, the slender cone (longicone) is a multiwhorled 'serpenticone' with long, thin body chamber (longidome; low expansion rate W *ca*. 1.5); the thick cone (brevicone) expands more rapidly (W *ca*. 2.5), and phragmocone and body-chamber become shorter (brevidome). In the longidomes, the body chamber was 1 to 2 whorls long; about 3/4 whorls in the abundant mesodomes (W *ca*. 2, i.e., shell diameter doubles with each whorl); and only about 1/2 of a whorl in the brevidomes, which resemble *Nautilus* (W 3–3.5).

Poise, i.e., the orientation of the aperture, and stability were obviously of great importance to the animal. Both were controlled by the positions of the centres of buoyancy and mass. The centre of buoyancy is the three-dimensional centre of the seawater displaced by the entire organism; the centre of mass is that of *all masses* within the same volume, with the principal variable being body-chamber length (measured in whorls). The buoyancy centre lies above the mass centre, and the distance between them determines the degree of stability, i.e., against the torque produced by the hyponome that threatens to rotate the organism. Stability limits the force of jet propulsion, depending on jet direction (vector force) relative to the centre of rotation (**Figure 7**). The densities of the phragmocone (*ca*. 0.2 kg/l) and body chamber with body (*ca*. 1.2 kg/l) differed greatly, so that brevidomes, with their body chamber mainly below the phragmocone, were more stable than mesodomes, and longidomes, with the body chamber completely surrounding the phragmocones, were highly unstable. Stability sufficient for jet propulsion existed only in brevidomes and mesodomes. But torque could be prevented only in mesodomes, when the jet force would pass through the centre of rotation – but the animal was limited to swimming backward, as well as up- and downward. Forward swimming explains the apparently wasteful 'rocking' of *Nautilus*. Its extremely brevidomic shell places the hyponome so low that it can curve backward below the shell, although this creates a torque; *Nautilus* is able to swim forward as well as backward and upward simply by curving the hyponome. 'Rocking' results from jet pulsation: the power phase rotates the animal backward and the inherently high static stability makes its return to rest position.

Backward swimming, however, required good steerage for maneouvrability; spherocones could, at most, have used long, trailing tentacles as rudders, whereas in oxycones and keeled platycones the shell allowed good steerage.

Hydrodynamic potential varied greatly among ammonoids. Involute, compressed shells (oxycones, platycones, discocones), more or less smooth, are obviously 'streamlined', i.e., they produce relatively little drag and friction during locomotion; brevidomes, sometimes with their body chambers slightly uncoiled and inflated to lower the centre of mass, were among the best swimmers among ammonoids. Surface roughness provided by fine riblets (the golf-ball effect) may have further reduced drag in some cases. Like *Nautilus*, brevidomes could swim forward and backward. Velocity increased with size, but swimming was sluggish in all externally shelled (ectocochliate) cephalopods because the body-chamber limited contraction of the mantle cavity for pumping. Only some oxyconic brevidomes among the Ammonitida (e.g., *Aconeceras*, *Quenstedtoceras*) show large muscle scars that indicate the presence of effective head retractors as in *Nautilus*. These were either nektic hunters or demersal ('bentho-pelagic'), feeding from the seafloor. At the other extreme were the serpenticones and longidomic sphaerocones, ribbed or smooth, which were unable to swim because the slightest jet force would have rotated the unstable, high-drag shells. They belonged to the mega-plankton and, according to their shell strength, were either limited to drifting near the surface or able to dive into deep water, presumably diurnally. The many intermediate forms, i.e., the abundant planorbicones and cadicones, usually mesodomic and with prominent ornamention, also tended to increase stability at maturity. They were probably very sluggish

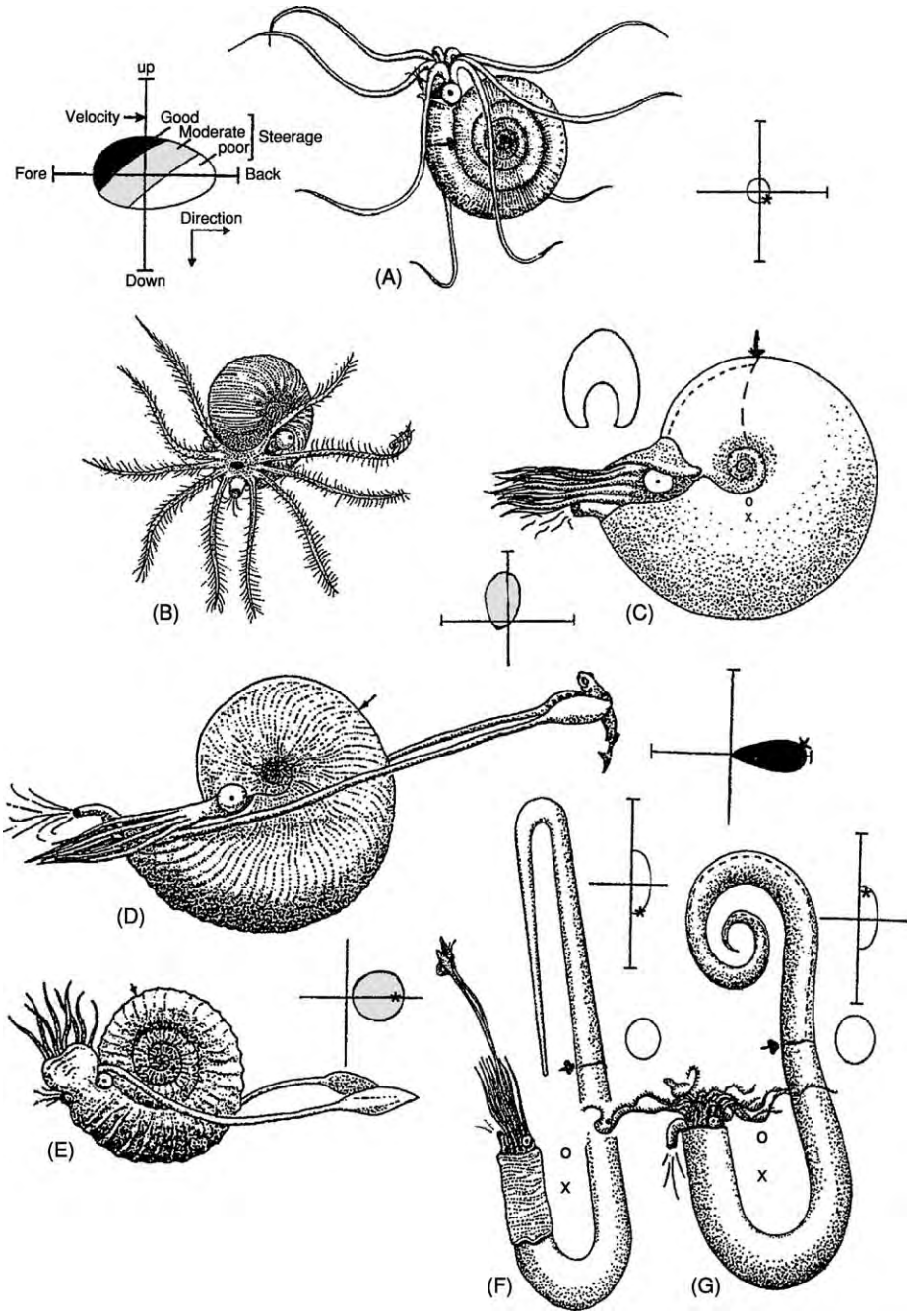


Figure 7 Reconstructions of poise and swimming potential according to shell form, body chamber length, and stability based on estimated centres of mass (x) and buoyancy (o); head, arms, and hyponome conjectured for various feeding strategies. (A) serpenticone; (B) longidomic spherocone; (C) mesodomic discocone; (D) brevidomic oxycone; (E) mesodomic platycone, microconch; (F) hamitocone, with mantle overlap; (G) ancylocone. Small arrows indicate end of body chamber. (Adapted from Westermann (1996) and Westermann and Tsujita (1999).)

backward swimmers and/or vertical migrants, mostly in shallow water, where their coarse ribs and/or spines would have provided some protection from predators and, perhaps, camouflage. Ellipto-spherocones possessed an unstable immature stage followed by a stable adult stage. Extension of this trend led to the heteromorph scaphitocones. The result was greatly improved stability, with the aperture turned upward, i.e., an adaptation for vertical migration (Figure 7). Whereas the open coils of gyrocones lie in one plane,

the wider coils of cyrtocones are irregularly helicoidal. The former grew large and may have been demersal; the latter were small and perhaps lived as pseudoplankton coiled around floating seaweed (e.g., *Sargasso* Community). Hamitocones, with a series of parallel segments, must have rotated during growth. The adults had stable near-vertical poise with upward directed apertures, implying diving habits as megaplankton. Orthocones also had stable near-vertical poise, but with the aperture downward, implying that they were sluggish benthos feeders with rapid escape potential. The torticones were also (diurnal) vertical migrants, whereas the rare vermicones, with a complex growth programme, were planktic. Buoyancy calculations have shown that all heteromorphs were neutrally buoyant.

Habitat depth limits can be calculated from the strength of septa, phragmocone wall, and connecting rings, because cameral gas pressure was maximally that of the atmosphere. Physical properties of ammonoid shell (test) are calibrated on *Nautilus*. Ammonoid connecting rings resemble the chitinous inner tubes of *Nautilus*, but ammonoid septa differ greatly in their architecture, being folded and not basically hemispheres. Stresses in ammonite septa are therefore calculated using curvature and thickness of lobes or lobules. Wall strength between sutures has also been calculated (Figure 8). The resulting depth limits range from 50 to over 1000 m. Actual habitat depth was at about one-third to two-thirds of the limit.

Ecology: Food, Predators, and Habitat

Feeding habits are known only for a few ammonoids, but were presumably as diverse as their shells. The general presence of weak jaws and radulae and the apparent absence of muscular arms suggest a number of possibilities, such as: (i) filter-feeding in mid-water on passing microplankton and detritus; (ii) predation on macroplankton and/or sluggish nekton, including smaller or young ammonoids, by pursuit or ambush; and (iii) preying on benthos while hovering. The radula aided in swallowing soft organisms whole. The known rare records of crop or stomach contents are of pelagic ostracods, microcrinoids, larger crustacean appendices, and juvenile aptychi indicating cannibalism.

Many ammonite shells suffered broken body chambers and some of these injuries were healed by the mantle before embedding. The predators are usually not identified, but large crustaceans, fish, and marine reptiles are strongly suspect. The best evidence comes from large, shallow-water oxycones (*Placenticeras*) that lived in the murky surface waters of the Cretaceous Western Interior Seaway of North America. These shells are perforated by numerous round holes that sometimes have the exact spacing and jaw angle of mosasaurs, large marine reptiles with conical teeth that are known from the same formation. Some authors believe that the holes are the diagenetically collapsed home scars (resting places) of limpets grazing on the empty shell, but this is improbable because of the shape and orientation of the holes.

The minute eggs and early juvenile (neanic) growth stages are rare in bituminous shales, which were deposited in lethal, oxygen-starved environments. This suggests that eggs (in gel?) and hatchlings floated in mid-water, where they were protected from active predators by the slight oxygen deficiency; they died and sank to the floor when dysoxic bottom waters rose periodically (Figure 9A). As in living cephalopods, some ammonoids may have spawned on oxygenated seafloors, but their minute aragonitic shells would have dissolved. Juvenile and immature ammonites are much scarcer than expected and the adults tend to range through a variety of lithofacies of epeiric seas upto 150–300 m deep. Most ammonoids did not depend on the seafloor and were pelagic: only a few species were demersal bottom feeders. A single case of brackish habitat has been documented by stable isotopes for the Late Cretaceous *Placenticeras*. Deep-water, oceanic sediments did not preserve ammonites, although epipelagic and some mesopelagic ammonoids lived in most oceans (Figure 9B). Most Jurassic-Cretaceous Phylloceratida and Lytoceratida have long been recognized as mostly deep-water ocean dwellers; planktic Ammonitida drifted and heteromorphic Ancyloceratida dived in the upper waters, and were dispersed far and wide by surface and subsurface currents.

Life Versus Death Assemblages: Migration and Post-Mortem Drift

There are two main reasons for numerical mismatches between sexual partners and for the common scarcity or even absence of immature shells from most fossiliferous deposits; spawning migrations and post-mortem transport. Most living cephalopods feed in open, deeper waters, but breed and die in shallow, warm waters. This often involves segregation of the sexes before and/or after mating, causing frequent mismatches (see Dimorphism above). The second reason for uneven sex ratios, as well as for the scarcity of immature shells, is selective surface drift of the empty shells. A strong natural bias for large shells of deep-water ammonoids over small, shallow-water species was caused by size and ambient pressure. Reflooding of the phragmocone causing sinking was much slower in large shells than in small ones,

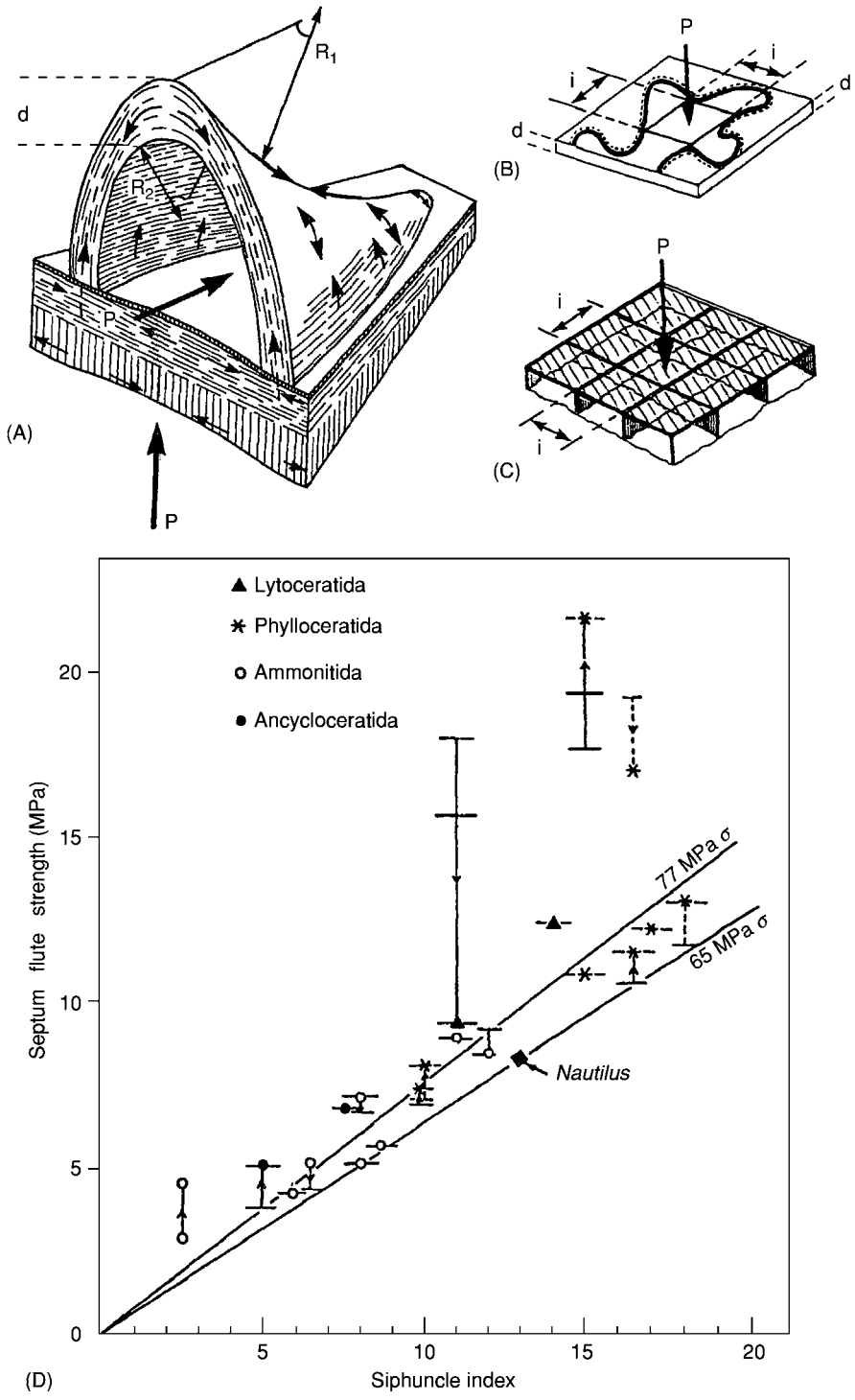


Figure 8 The shell parameters used for bathymetry, based on the ammonitic septum, phragmocone wall, and siphuncle. (A) Lobule of last septum with anticlastic curvature used to calculate Septum Flute Strength; R_1 and R_2, orthogonal radii; P, hydrostatic pressure; d, septum thickness; arrows, compressive and tensile stresses. (B) Wall with lobule to calculate wall strength, the implosion depth based on the unsupported wall distance; i, length of largest unsupported square; d, wall thickness. (C) Connecting ring of siphuncle, cross section (Siphuncle Strength Index 100X d_s/r_i). (D) Habitat limits based on septum and siphuncle of adult ammonoids and *Nautilus* (ontogenies indicated): correlation is good except for some compressed, deep water Phylloceratida, which have thicker septa that functioned significantly as lateral braces; depth limit calibrated on *Nautilus*. (Adapted from Hewitt (1996).)

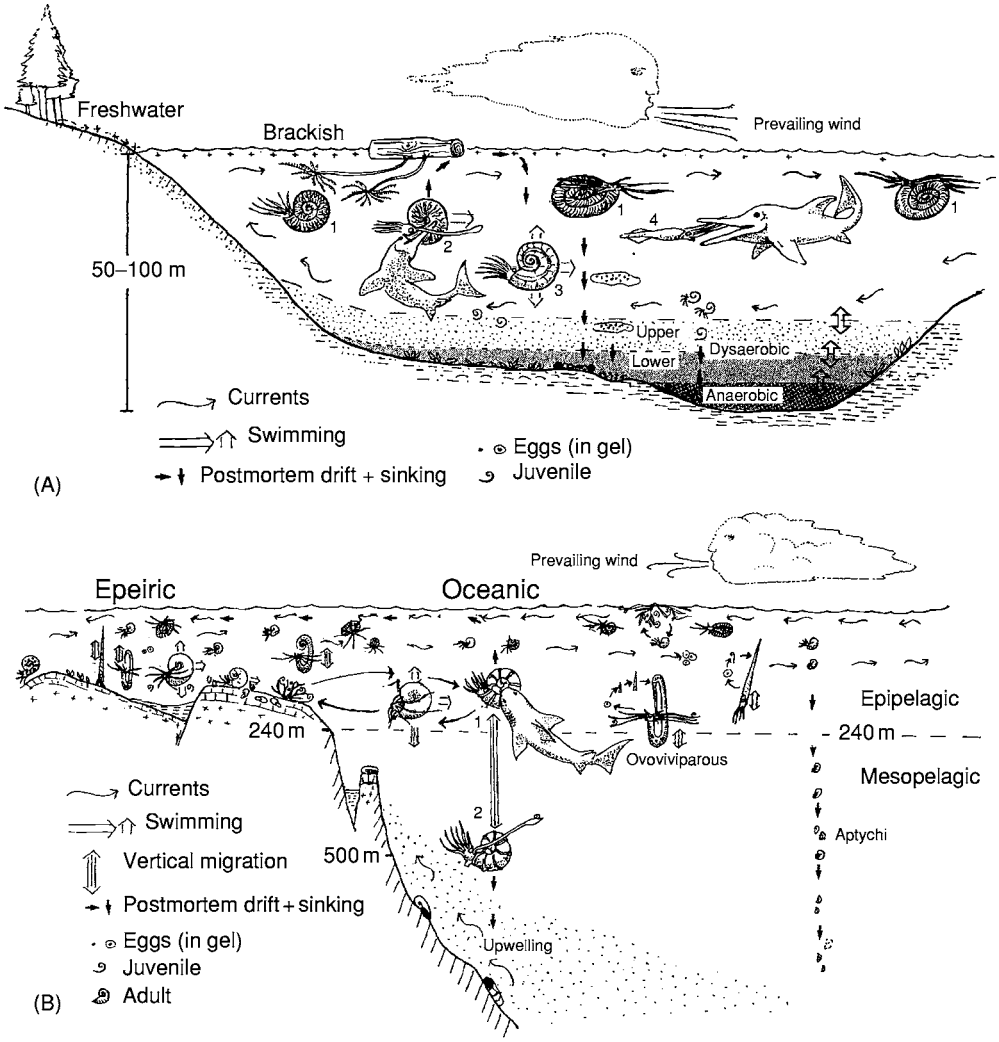


Figure 9 Possible scenarios for Jurassic Cretaceous ammonoid habitats, with life cycles and post mortem drift. (A) Epeiric basin, black shale facies, e.g., Early Jurassic *Posidonia* Shale or Jet Rock; 1, serpenticonic, epiplanktic drifters (*Dactylioceras*); 2 and 3, discoconic predators (*Harpoceras, Hildoceras*), prey of ichthyosaurs; 4, belemnite animal. (B) Ocean, slope, shelf, and epeiric sea; serpenticonic Ammonitida drift and heteromorphic Ancyloceratida dive in both biomes; most Ammonitida are restricted to epeiric seas; typical Phylloceratida (1) were deep water swimmers and typical Lytoceratida (2) among the deepest vertical migrants; only calcitic aptychi reach the deep sea floor. (Adapted from Westermann (1996).)

especially those with thick-walled, narrow connecting rings (siphuncle surface grows with the square of diameter, chamber volume with the cube) – just as cameral emptying rates in growing shells. Similarly, the reflooding rate depended on the depth at death: it was rapid only under high ambient pressure, so that the shell sank immediately; during slower refilling in shallower water the shell would rise to the surface (Figure 9) before sinking after weeks or months of drifting. Drifting distance has been much debated, but it was probably rarely more than a few tens of kilometres in epicontinental seas. An exceptional case, however, was observed in the Andes at the edge of the palaeo-Pacific: several gigantic lytoceratids of almost certainly meso-pelagic habitat were found in a nearshore assemblage. The sinking aragonitic shells of dead oceanic ammonoids were dissolved below compensation depth, but their calcitic aptychi have been found in red deepsea clay.

Summary

Most ammonoids were pelagic and a few demersal between 30 and 150 m depth in epeiric seas; others were epi- to mesopelagic in the oceans; and some lived in epeiric seas as well as oceans, for example,

cosmopolitan longidomes. Their principal food was micro- and mesoplankton, including young ammonoids, presumably caught with weak arms or tentacles, and sinking organic detritus netted with an umbrella-like arm crown, as well as benthos. Their predators were fish, reptiles, and larger ammonoids. The most diverse ammonoid assemblages are found in sediments of warm tropical and subtropical seas and typically include mainly adults. Gray and black shales that originated in temperate or basinal waters, on the other hand, tend to have low-diversity faunas that may include ammonitellas and juveniles; single, highly variable species dominate, sometimes ranging from ribbed spherocones to smooth oxycones or platycones. Their contrasting hydrodynamic properties did not apparently function; these ammonoids were mega-plankton without significant locomotion.

Ocean currents, rather than swimming, were the main means of wide dispersal, because planktic species tend to be more cosmopolitan than nektic species. Prime examples are longidomic spherocones (e.g., arcestid Ceratitida) and serpenticones (e.g., dactylioceratid and psiloceratid Ammonitida) as well as many heteromorphs (Ancyloceratida), which lived in the shallow waters of all seas. The mainly oceanic Phylloceratida and Lytoceratida evolved much more slowly, which may account for their wide species distributions. The importance of most ammonites in biostratigraphy (*see* **Biozones**) and biogeography resulted from rapid evolution combined with high potential for planktic dispersal.

See Also

Biozones. **Fossil Invertebrates:** Molluscs Overview; Cephalopods (Other Than Ammonites).

Further Reading

Becker RT and Kullmann J (1996) Paleozoic ammonoids in space and time. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 711 754. New York: Plenum Press.

Bucher H, Landman NH, Klofak SM, and Guex J (1996) Mode and growth in ammonoids. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 408 462. New York: Plenum Press.

Davis RA, Landman NH, Dommergues J L, Marchand D, and Bucher H (1996) Mature modifications and di morphism in ammonoid cephalopods. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 463 539. New York: Plenum Press.

Doguzhaeva LH and Mutvey H (1996) Attachment of the body to the shell in ammonoids. In: Landmann, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 44 63. New York: Plenum Press.

Hewitt RA (1996) Architecture and Strength of the Ammo noid Shell. In: Landman, *et al.* (eds.) *Ammonoid Paleo biology,* pp. 297 339. New York: Plenum Press.

House MR and Senior JR (eds.) (1981) The Ammonoidea. *The Systematics Association*, Spec. Vol. 18. London: Aca demic Press.

Jacobs DK and Chamberlain JA, Jr. (1996) Buoyancy and hydrodynamics in ammonoids. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 169 224. New York: Plenum Press.

Landman NH, Tanabe K, and Davis RA (eds.) (1996) *Ammonoid Paleobiology.* New York: Plenum Press.

Page KN (1996) Mesozoic ammonoids in space and time. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 755 794. New York: Plenum Press.

Tanabe K and Fucuda Y (1996) Morphology and function of cephalopod buccal mass. In: Savazzi E (ed.) *Functional Morphology of the Invertebrate Skeleton,* pp. 245 262. Chichester: Wiley.

Westermann GEG (1971) Form, structure and function of shell and siphuncle in coiled Mesozoic ammonoids. *Royal Ontario Museum, Life Sciences Contributions* 78: 1 39.

Westermann GEG (1975) Model for origin, function and fabrication of fluted cephalopod septa. *Palaeontolo gische Zeitschrift* 49: 235 253.

Westermann GEG (1996) Ammonid life and habitat. In: Landman, *et al.* (eds.) *Ammonoid Paleobiology,* pp. 607 707. New York: Plenum Press.

Westermann GEG and Tsujita CJ (1999) Life habits of ammonoids. In: Savazzi E (ed.) *Functional Morphology of the Invertebrate Skeleton,* pp. 299 325. Chichester: Wiley.

Porifera

J K Rigby, Brigham Young University, Provo, UT, USA

Introduction

The phylum Porifera, or sponges, includes simple or primitive multicellular sedentary organisms, and is among the most diverse and successful of extant phyla. Living sponges are filter feeders and are characterized by body plans of three structural grades (Figure 1), which have been recognized also in fossil forms. Water is pumped in through numerous smaller inhalant openings, or ostia, by the uncoordinated beating of flagella of choanocytes, or collar cells. These cells are located in interior chambers within the sponge, and cause the water to circulate via numerous canals throughout the sponge body, and to exit through larger exhalant openings, or oscula. Food particles are filtered from the inhalant water virtually on a cell-by-cell basis, whilst waste products are emitted into the outflowing exhalant currents. Soft parts or tissues of sponges are supported by internal skeletons made of organic fibres, mineralized needle-like or multirayed spicules, or a combination of fibres and spicules. It is primarily the mineralized skeletons that make up most of the geological record of the group. The nature, shapes, and interrelationships of spicules and other structures within their skeletons are used to classify fossil sponges.

Skeleton elements of living sponges are secreted by specialized cells, termed spongocytes, which secrete the spongin fibres, or sclerocytes, which secrete the mineralized spicules. These spicules are composed of opaline silica or crystalline to microgranular calcium carbonate. Their compositions and shapes are used to differentiate taxa at several levels. Calcium carbonate may also occur as layered, granular to crystalline aragonite or calcite in the Calcarea, Archaeocyatha, and Sclerospongiae. Intermixed siliceous and carbonate skeletal elements are known in the Sclerospongiae and in hypercalcified sponges in two subclasses of the Demospongea.

The nomenclature of spicules is based upon their size and the numbers of rays or axes that they have. One sponge may have several kinds of spicules, and the same kind of spicule may occur in several sponges. Megascleres (Figure 2) are the large principal skeletal elements and range from monaxons, with a single axis, to triaxons or tetraxons, with three or four axes of ray growth. Monaxial spicules may be termed monactine or diactine, depending upon whether there are one or two directions of growth. Similarly, triaxons may have three directions of growth and be triactines, which are generally of calcareous composition, or they may be hexactines with six directions of ray growth and of siliceous composition. Tetractines have four rays and directions of growth. Octactines are specialized calcareous tetractines with eight rays, and they characterize the Heteractinida, a class of Palaeozoic fossils. Some sponges have spicules with many rays and axes of growth. Such spicules are termed polyactines or polyaxons, or sphaeractines.

Microscleres are small spicules used in the classification of modern sponges (Figure 3), but they are only occasionally preserved in fossil sponges, and are consequently of minor use in palaeontology.

Living sponges are aqueous organisms and are dominantly marine, although they do occur in a few freshwater lakes and streams. Marine forms occur from polar to tropical environments, and from shallow tidal environments to abyssal depths. They are characteristically bottom dwellers. Fossil sponges occupied the same general environments, based on their occurrences in various kinds of rock in the geological record.

Sponges are relatively minor fossils when the total palaeontological spectrum is examined, but they played major roles locally, as in the construction of the famed Permian reefs of western Texas and New Mexico in the USA, or the Jurassic reefs of Europe, for example. Following death, some forms with unfused skeletons broke up and their dissociated mineralized spicule elements locally accumulated to form sedimentary deposits, known as spiculites or spiculitic chert.

Classification

Fossil sponges are classified on the basis of the composition and forms of their spicules, their canal systems, and their structural grades. Traditionally, five classes have been recognized in the phylum, including the Demospongea, Calcarea, Hexactinellida, Sclerospongia, and the extinct Heteractinida. To these have recently been added the Archaeocyatha, which had been considered as a separate phylum for many years, and the Stromatoporoidea and possibly related groups, which have been grouped with the Coelenterata for many years.

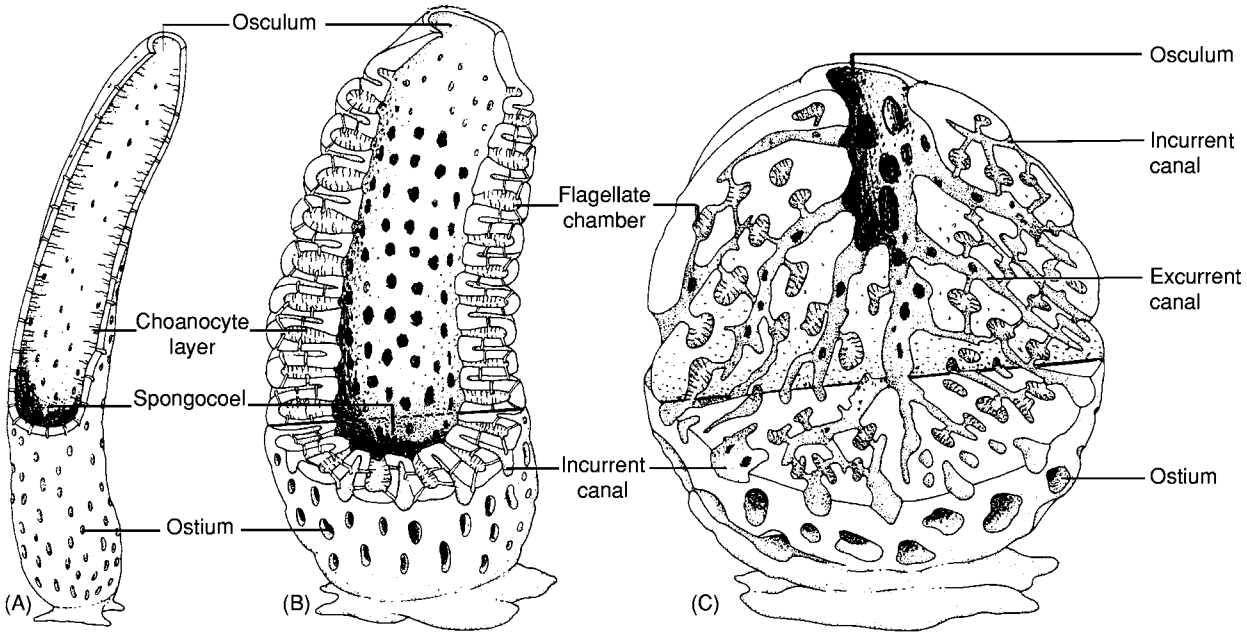


Figure 1 Structural grades of living and fossil sponges. (A) Ascon grade sponges have thin walls and a large spongocoel, or central cavity, lined by flagellate choanocytes. (B) Sycon grade sponges have somewhat thicker walls that contain distinct chambers of flagellate choanocytes, chambers which open directly into the spongocoel. (C) Leucon grade sponges are the most complex and have isolated flagellate chambers that are interconnected with inhalant and exhalant canals by numerous small canals. Not to scale. Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

Demospongea

Sponges of the Class Demospongea are characterized by skeletons composed of organic spongin fibres, siliceous spicules, or mixed spongin and siliceous spicules. Those with spongin skeletons have left a poor record and essentially all recognizable fossil demosponges are characterized by their shapes, canal patterns, and spicule shapes and relationships (Figure 4). Spicules of the class range from loose monaxons to tetractines whose rays do not join at right angles, or to irregular root-like forms (Figure 2). Many Palaeozoic and Mesozoic demosponges belong to the Order Lithistida, whose skeletons are formed of fused and cross-braced spicules that have articulated ray tips or rough exteriors with nodes or spines that united to form rigid, resistant skeletons that, once buried, were commonly preserved.

Demosponges first appear in the geological record in the Late Precambrian, and diversified moderately quickly so that, by the Middle Cambrian, several genera are known that represent the initiation of important distinct major Palaeozoic lineages. During the Lower and Middle Ordovician, they underwent great diversification and, for example, became major reef builders and also occupied diverse environments. Their Silurian record shows that they remained common and diversified, so that they are relatively abundant fossils in both carbonate and clastic sequences that accumulated in both shallow and deeper marine environments. Although locally abundant, they became less dominant in later Palaeozoic sequences. Demosponges again played major roles in the Jurassic and Cretaceous, when they reached a second peak of diversity and abundance in a variety of environments.

Demosponges are the dominant class of living sponges, where several thousand species have been described. Although mainly shallow marine organisms, the class also includes the only known living and fossil freshwater sponges. To date, nearly 500 genera of fossil demosponges have been described.

Class Demospongea Sollas, 1875, Precambrian–Holocene
- Subclass Clavaxinellida Lévi, 1956, Precambrian–Holocene
- Subclass Choristida Sollas, 1880, Upper Ordovician–Holocene
- Subclass Tetractinomorpha, Lévi, 1953, Middle Ordovician–Holocene

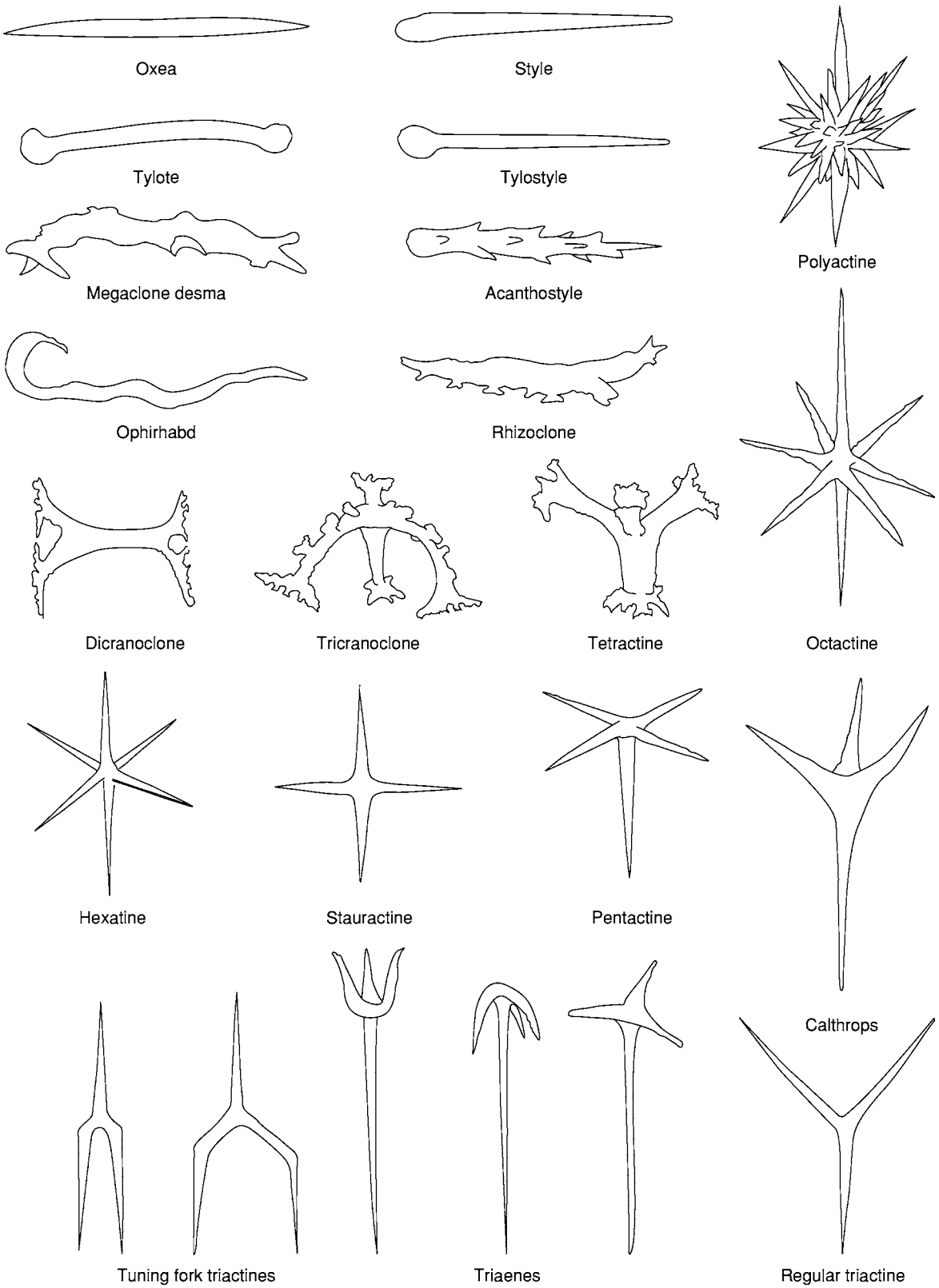


Figure 2 Nomenclature of common megascleres or main body spicules, which may range from a fraction of a millimetre to several millimetres in length. Shapes and compositions of megascleres are important elements in classification. Triactines, for example, are commonly calcareous and representative of the Calcarea, and hexactines and derivative spicules, such as stauractines and pentactines, are siliceous and characterize the Hexactinellida.

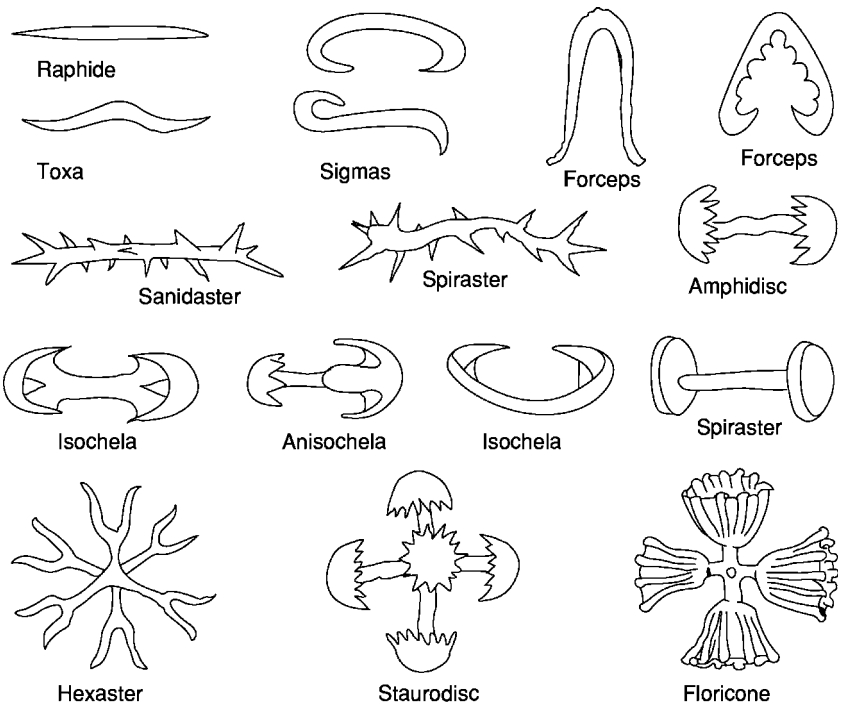


Figure 3 Representative microscleres of the demosponges and hexactinellid sponges. They are considerable smaller than associated megascleres and are usually only a fraction of a millimetre in length.

Subclass Ceractinomorpha Lévi, 1953, Middle Cambrian, Middle–Upper Ordovician, ? Pennsylvanian, Holocene
Subclass Lithistida Schmidt, 1870, Cambrian–Holocene

Hexactinellida

The Class Hexactinellida is characterized by skeletons composed of siliceous hexactine-based spicules, and has a geological range that extends from the Late Precambrian to the Holocene. It was, and is, an exclusively marine class of sponges. Two major Palaeozoic groups are documented in the geological record (Figure 5): the asconoid protosponges of the Cambrian and Ordovician, which have thin walls of unfused hexactines that are regularly arranged, and the syconoid lyssacinosid dictyosponges, which have thicker walls of quadrangular arranged to irregularly arranged hexactines with unfused to weakly fused overlapping rays (Figure 6A). The latter group is particularly well known in Devonian to Carboniferous rocks from New York and Pennsylvania to Indiana (Figure 5B). By the latest Permian, and certainly by the Triassic, the Order Hexactinosa developed with rectangularly based, rigid, fused, hexactine skeletons. These reticular skeletons (Figure 6B) were produced where overlapping rays of spicules were fused by cementation, by enclosure of overlapping rays in a siliceous coating, by interconnecting small nodes, or, more rarely, by tip-to-tip fusion. The Hexactinosa reached their peak of development in the Cretaceous. More or less concurrently, sponges of the Order Lychniscosa developed lantern-like, cross-braced, nodal octahedra at spicule centres (Figure 6C) in their more or less rectangular fused hexactinellid skeletons. That order reached its maximum development in the Jurassic and Cretaceous of Europe (Figure 5D).

Class Hexactinellida Schmidt, 1870, Precambrian–Holocene
Subclass Amphidiscophora Schulze, 1887, Precambrian–Holocene
Order Amphidiscosa Schrammen, 1924, Ordovician–Holocene
Order Reticulosa Reid, 1958, Precambrian–Upper Permian
Order Hemidiscosa Schrammen, 1924, Upper Pennsylvanian–Cretaceous
Subclass Hexasterophora Schulze, 1887, Ordovician–Holocene
Order Lyssacinosa Zittel, 1877, Ordovician–Holocene
Order Hexactinosa Schrammen, 1903, Permian–Holocene
Order Lychniscosa Schrammen, 1903, Upper Triassic–Holocene

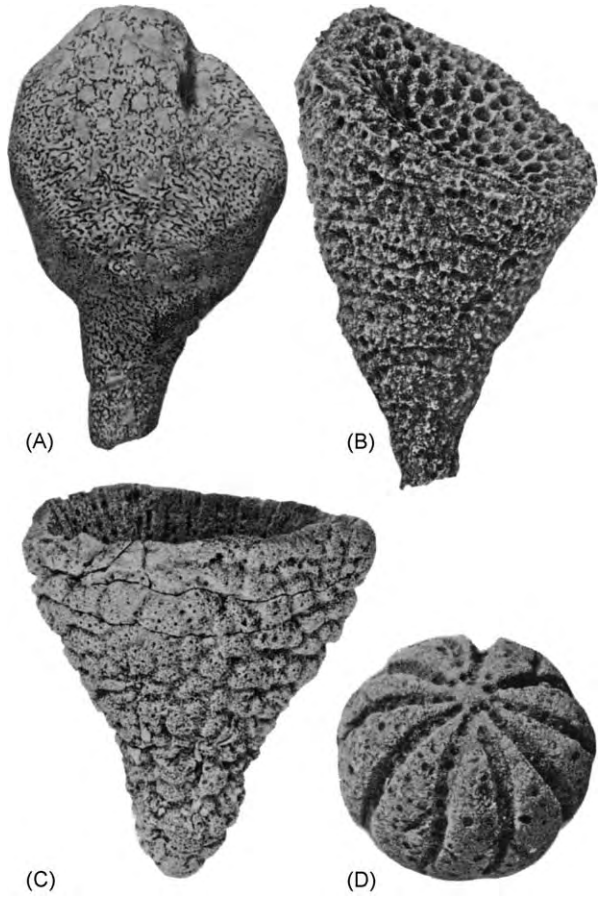


Figure 4 Representative fossil demosponges. (A) *Phymatella*, Cretaceous, Germany. (B) *Aulocopoides*, Devonian, Western Australia. (C) *Camellaspongia*, Ordovician, Minnesota, USA. (D) *Caryospongia*, Silurian, Tennessee, USA. All approximately normal size. Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

Hexactinellid sponges are exclusively marine forms and, in modern seas, they are most common on seafloors from 200 to 2000 m deep, although many species have been reported from lower bathyal depths. They also occur, but are less abundant, in hadal depths over 6000 m deep, where they may form dense reef-like clusters. On the other hand, they are known to range up into water as shallow as 25–30 m off the south-western Canadian Pacific Coast, where they also form distinct reef-like structures.

Calcarea

The Class Calcarea includes sponges with calcareous skeletal elements that range from those with distinct three-rayed spicules of calcite or aragonite (Figure 2), to those with rigid skeletons of fused polygonal elements or imbricate calcitic plates. Living calcareous sponges are exclusively marine and are most common in shallow tropical environments. Those included in the Orders Murrayonida and Lithonida are the traditional 'pharetronids' and the other orders are the traditional 'inozoids.' The class has a geological record ranging questionably from the Late Precambrian, but certainly from the Lower Cambrian to the Recent or Holocene.

Class Calcarea Bowerbank 1864, Cambrian–Holocene
- Subclass Calcinea Bidder, 1898,?Precambrian, Cambrian–Holocene
  - Order Clathrinida Hartman, 1958, Holocene
  - Order Murrayonida Vacelet, 1981,?Precambrian, Cambrian–Holocene
- Subclass Calcaronea Bidder, 1898, ?Cambrian,? Triassic, Jurassic–Holocene
  - Order Leucosoleniida Hartman, 1958, Holocene
  - Order Sycettida Bidder, 1898, Holocene
  - Order Sphaerocoeliida Vacelet, 1977, Cretaceous
  - Order Lithonida Doederlein, 1892, Jurassic–Holocene

Many genera of chambered fossil sponges with calcareous skeletons (Figure 7) had been included in the Calcarea, in the Order Sphinctozoa, until recently, when the polyphyletic origins of the sphinctozoans and their development principally as a structural grade were recognized. They are now treated as a group of 'hypercalcified' sponges drawn from other classes. Sphinctozoan grade sponges appeared in the Middle Cambrian, but played a minor role until the Carboniferous and Permian, when they helped to produce massive reefs, such as those in the Guadalupe Mountains in western Texas and south-eastern New Mexico, and in Tunisia and China, and during the Triassic, when they were similarly involved in the construction of reefs in what is now the Alpine region of southern Europe. They were thought to have become extinct by the end of the Cretaceous, but a living form was discovered in 1977.

Their chambered construction, in which 'living space' was limited by an outer rigid skeleton, required that new chambers be added for growth of the sponge. Such chambered skeletons developed at various times in several groups, so that a few calcareous and hexactinellid genera and many demosponge genera all had 'sphinctozoan' skeletal patterns. Chambered hexactinellid 'sphinctozoans' have been recognized, for example, because their chambered skeletons contain reticulate hexactines. Chambered demosponge sphinctozoans with rigid skeletons of aragonite or high-magnesium calcite are similarly included here. Skeletons of these latter

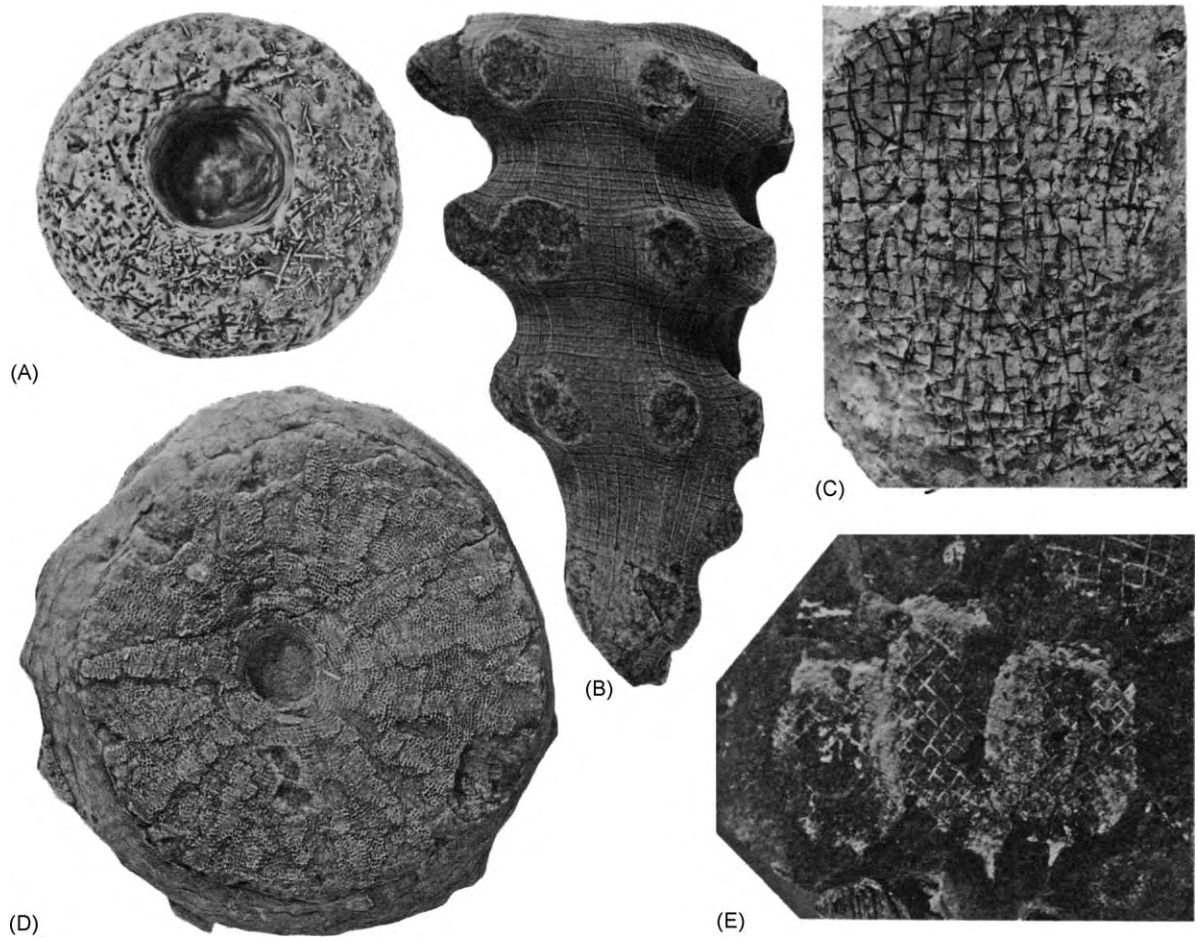


Figure 5 Representative hexactinellid sponges. (A) *Twenhofelella*, Silurian, Quebec, Canada. (B) *Hydnoceras*, Devonian, New York, USA. (C) *Mattaspongia*, Devonian, Alberta, Canada. (D) *Coeloptychium*, Cretaceous, France. (E) *Diagoniella*, Cambrian, Utah, USA. All are natural size. Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

forms may or may not contain primary siliceous monaxon spicules, or pseudomorphs of such spicules, within their calcareous rigid elements.

Taxa within the polyphyletic group are differentiated by their shapes and chamber arrangements, their spicule content, as well as by the mineralogical compositions of their skeletons, internal segmentation, canal systems, and filling structures. They range from simple beaded moniliform sponges to polyglomerate cylindrical forms, or to mound-like or stratiform multichambered sponges (Figure 7). Filling structures within chambers may include septae, vesiculae, spore-like elements, or fine reticular structures, amongst others, or the chambers may have been hollow and lacked such skeletal structures.

Heteractinida

The Class Heteractinida, considered by some palaeontologists to be an order of the Calcarea, is characterized by skeletons made of large calcareous octactine spicules. The small class has a geological record that extends from the Lower Cambrian to the Lower Permian, and is one of only two classes of sponges to have become extinct. The Heteractinida includes such forms as the broad saucer-shaped octactinellid *Astraeospongium* (Figure 8), which is a common genus from the Silurian. The class is particularly common in Silurian and Carboniferous rocks of Europe and North America.

Class Heteractinida de Laubenfels, 1955, Lower Cambrian–Lower Permian
Order Octactinellida Hinde, 1887, Lower Cambrian–Lower Permian
?Order Hetairacyathida Bedford and Bedford, 1937, Lower Cambrian

Sclerospongiae

The Class Sclerospongiae was proposed to include a few living sponges, and is now thought by some

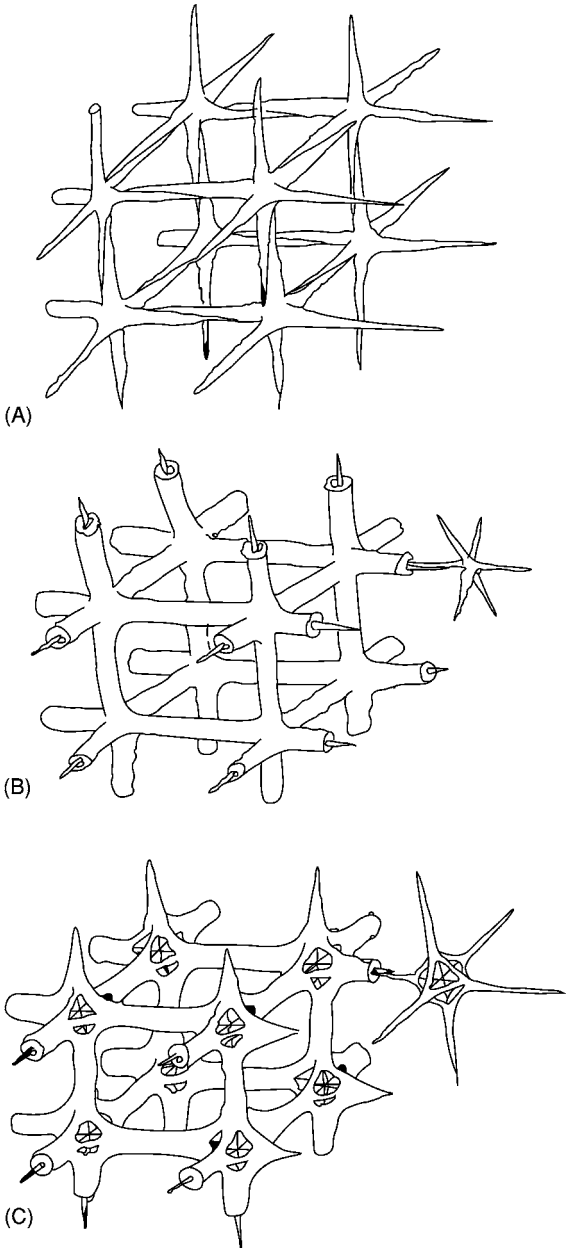


Figure 6 Diagrams of enlarged hexactinellid skeletal structures. (A) Lyssacid grade reticulate skeletal structure with unfused or weakly fused overlapping hexactine spicule rays. (B) Dictyid or hexactinosan grade with overlapping spicule rays fused into a rigid, three dimensional, quadrangular skeletal meshwork. (C) Lychniscoid grade with overlapping rays fused into a rigid skeletal meshwork and with diagnostic octahedral 'lanterns' at the centre of each spicule.

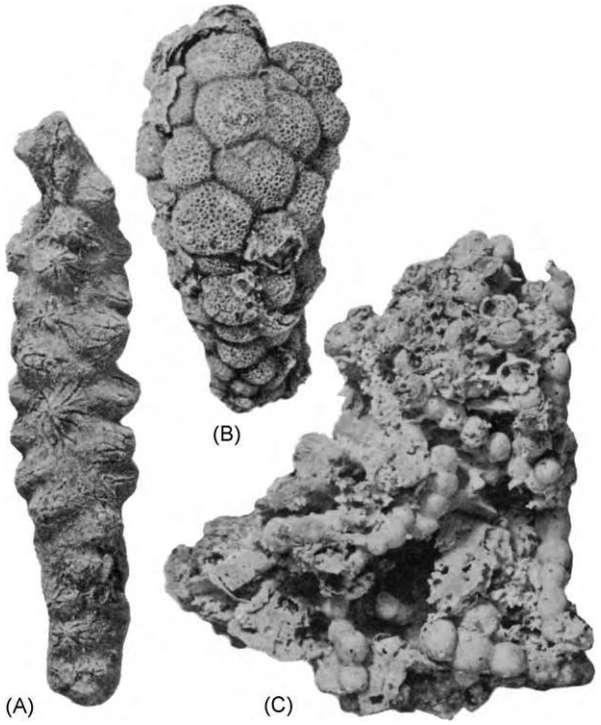


Figure 7 Representative calcareous and 'sphinctozoan' sponges. (A) *Stellispongia*, Permian, Tunisia, an inozoan calcareous sponge. (B) *Cystothalamia*, 'sphinctozoan' hypercalcified demosponge, Permian, Texas, USA. (C) *Girtyocoelia*, 'sphintozoan' hypercalcified demosponge, Permian, Texas, USA. All are natural size. Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

workers to also include the taxonomically limited fossil chaetetids and the diverse stromatoporoids. Stromatoporoids had long been considered as coelenterates related to hydrozoans, but recently they have been moved to the Porifera. These sponges have a laminated calcareous skeleton that is perforated by astrorhizal canals, and the living animal is thought to have occupied essentially only the upper surface. If the Stromatoporoidea are included, the class Sclerospongiae has a stratigraphical range from the Cambrian to the Holocene, with a major period of development in the Silurian and Devonian, when these organisms were important reef formers. The class became a minor marine element in the Late Palaeozoic, however, and remains so in modern marine faunas.

Stromatoporoids have domal, tabular, branching, or bulbous growth forms. They are commonly laminated to vertically tubular or pillared, and have generally aspiculate, calcareous skeletons in early forms and more cystose and vertically tubular skeletons in later forms. Laminate forms commonly show cyclicity, or growth interruptions, with thicker latilaminae separated by several thinner sheet-like porous laminae (Figure 9). These laminae are interconnected or supported by vertical pillars that may extend only between adjacent laminae or may be superimposed through several laminae. They range from long and

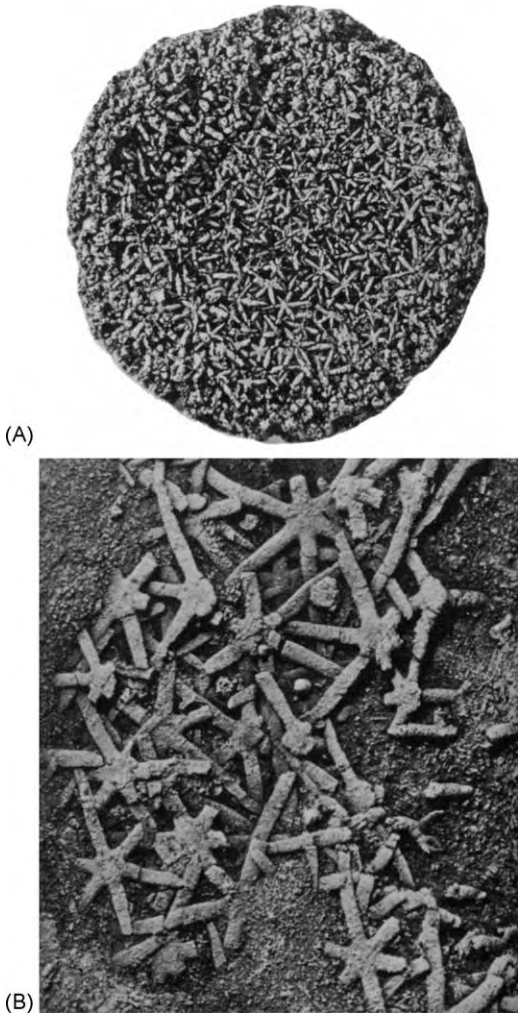


Figure 8 Representative heteractinid sponges. (A) *Astraeospongium*, Silurian, Tennessee, USA, coarse spiculed, broad, saucer shaped sponge as seen from above, with large octactine spicules, natural size. (B) *Ensiferites*, Devonian, New York, USA, large spicules with six horizontal tangential rays, (×5). Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Board man RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

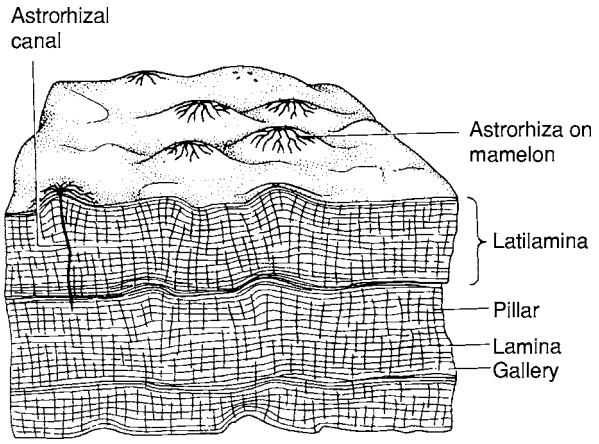


Figure 9 Block diagram of the stromatoporoid, *Actinostroma*, showing internal horizontal thin laminae and coarser latilaminae, with vertical pillars that combine with laminae to define openings (galleries) in the skeleton. Mound like mamelons on the surface are marked by convergent canals at the upper ends of astrorhizal canals (×5). Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

continuous to more spool-shaped and irregular, or they may be rod-like in an interconnected network. Laminate skeletons are commonly perforated by vertical astrorhizal canals with transverse or horizontal stellate tributary canals which show best as interruptions in laminae. Several different types of microstructure are present in the calcareous skeletal structures, apparently a result of variations in the secretion of calcium carbonate skeletons, and in their subsequent diagenesis.

Stromatoporoids first appeared in the geological record in the Ordovician, and by the Middle Devonian they occurred worldwide in sediments of shallow marine environments, where they were significant contributors to mid-Palaeozoic reefs. They left a discontinuous Carboniferous record, but became more common in the Permian and thrived during the mid-Mesozoic, when they again formed reefs. They became extinct at the end of the Cretaceous.

The possibly related chaetetids are a moderately small group of colonial organisms that were included with tabulate corals for many years, and only recently have been considered to be sclerosponges, once spicule pseudomorphs and astrorhizal canals were observed in Palaeozoic representatives of the group. The chaetetids were shallow-water marine organisms. They range stratigraphically from the Ordovician to the Miocene, and are locally abundant organisms in Pennsylvanian formations. For example, they form small mound-like reefs in Middle Pennsylvanian rocks of Kansas.

They have compound skeletons that range from plate-like or domed to columnar (Figure 10), and are composed of clustered narrow tubes, or calicles, that are less than 1 mm in diameter and commonly polygonal to irregular in cross-section. These calicles lack true septae, but have common horizontal tabulae, and their shared or joined walls are imperforate. The colonies appear to have enlarged by budding or longitudinal fission of calicles. The growth form, skeletal structure, and the rare occurrence of spicules indicate a strong relationship to other sclerosponges.

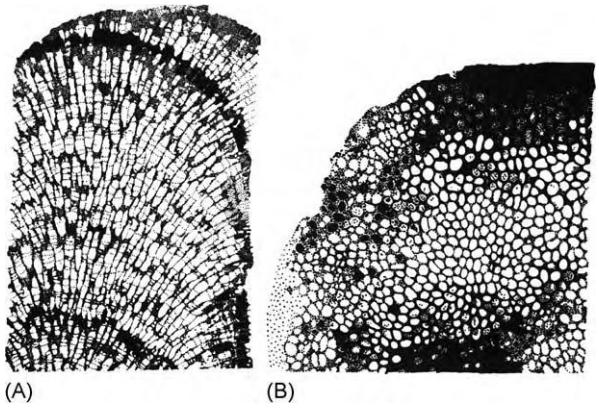


Figure 10 Longitudinal (A) and transverse (B) sections of *Chaetetes ?septosus* Fleming, Lower Carboniferous, Europe, showing narrow calicles, or tubes, with polygonal transverse sections, that become rounded where growth was inhibited, and perforate horizontal tabulae (×4). Adapted from Hill and Stumm, (1956), Treatise on Invertebrate Paleontology, Part F, with permission from, The Geological Society of America and The University of Kansas Press.

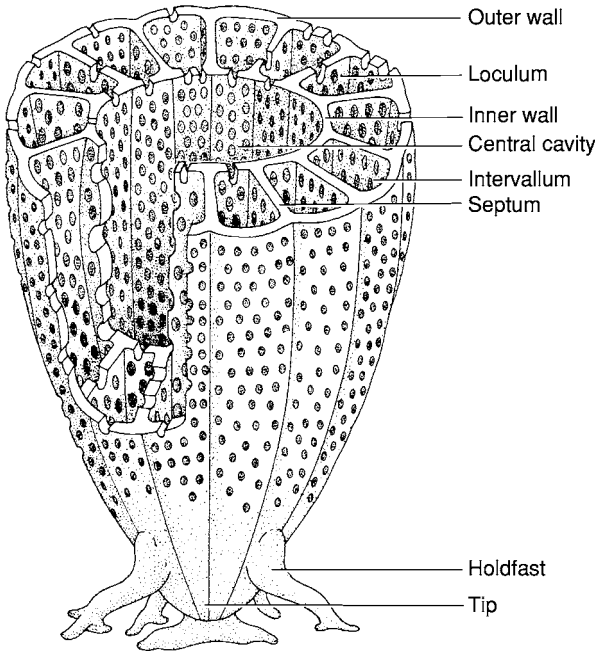


Figure 11 Generalized archaeocyathan showing characteristic skeletal elements of a double walled, steeply obconical, form (×2). Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

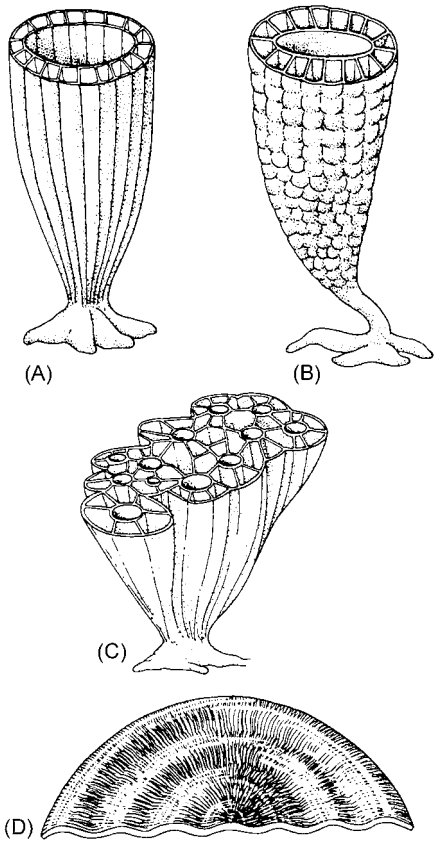


Figure 12 Range in growth forms of representative solitary and colonial archaeocyathans. (A) *Ajacicyathus*, a simple obconical erect form. (B) *Kotuyicyathus*, sculptured obconical form. (C) *Ajacicyathus*, massive colonial form. (D) *Okulitchicyathus*, wavy, discoidal flabellate form. All approximately natural size and all of Lower Cambrian age. Adapted with permission from Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates*. Palo Alto, Oxford, London: Blackwell Scientific Publications.

Archaeocyatha

Fossils of the Class Archaeocyatha are known essentially from the Lower Cambrian, where they became relatively abundant as the first metazoan reef formers, and from which over 300 genera have been described. They appear in the geological record near the base of the Cambrian, but were extinct by Middle Cambrian time, except in Russia and Antarctica where two genera persisted into the Early–Middle Cambrian.

Their skeletons are typically solitary conical to branching structures and, in early simple forms, are structures in which a porous inner wall surrounds the central cavity and is separated from a porous outer wall by the intervallum, which may be divided into longitudinal openings by radial septa that bridge between the two walls (Figure 11). Basal tips may have holdfasts, which are root-like attachment structures. Simple single-walled forms and more complex thalamid chambered forms are also known, as are less common fan-shaped and bowl-shaped genera and branched colonial forms with moderately complex walls (Figure 12).

Class Archaeocyatha Bornemann, 1884
- Order Monocyathida Okulitch, 1935, Lower Cambrian
- Order Ajacicyathida Bedford & Bedford, 1939, Lower Cambrian
- Order Archaeocyathida Okulitch, 1935, Lower–Middle Cambrian
- Order Capsulocyathida Zhuravleva, 1964, Lower Cambrian
- Order Kazachstanicyathida Konyushkov, 1967, Lower Cambrian
- Order Tabulacyathida Vologdin, 1956, Lower Cambrian

The archaeocyathans were an exclusively marine group of organisms, and are most common and diverse in argillaceous and carbonate rocks that appear to have accumulated in normal marine waters up to 20–30 m deep.

See Also

Mesozoic: Jurassic. **Palaeoclimates**. **Palaeozoic:** Cambrian; Permian. **Sedimentary Environments:** Reefs ('Build-Ups').

Further Reading

Bergquist PR (1978) *Sponges*. Berkeley and Los Angeles: University of California Press.

Broadhead TW (ed.) (1983) *Sponges and Spongiomorphs*, Notes for a short course, organized by J. K. Rigby and C. W. Stearn, *University of Tennessee Studies in Geology* 7. Knoxville: University of Tennessee. (Contains several chapters concerning various groups now included within the Porifera, each written by an authority on the group.)

De Laubenfels MW (1955) Porifera. In: Moore RC (ed.) *Archaeocyatha and Porifera,* Part E, *Treatise on Invertebrate Paleontology,* pp. E21 E112. Boulder, Colorado and Lawrence: Geological Society of America and University of Kansas Press.

Finks RM (1960) Late Paleozoic sponge faunas of the Texas region: the siliceous sponges. *American Museum of Natural History Bulletin* 120.

Hall J and Clarke JM (1899) A memoir of the Paleozoic reticulate sponges constituting the family Dictyospongidae. *New York State Museum Memoir* 2.

Hill D (1972) Archaeocyatha. In: Teichert C (ed.) *Treatise on Invertebrate Paleontology,* Part E (revised), 2nd edn, pp. E1 E158. Boulder, Colorado and Lawrence: Geological Society of America and The University of Kansas.

Hill D and Stumm EC (1956) Tabulata. In: Moore RC (ed.) *Coelenterata,* Part F, *Treatise on Invertebrate Paleontology,* pp. F444 F475. Boulder, Colorado and Lawrence: Geological Society of America and University of Kansas Press.

Hooper JNA and Van Soest RWM (eds.) (2002) *Systema Porifera, A Guide to the Classification of Sponges*. New York: Kluwer Academic/Plenum Publishers (This major work has many authoritative chapters, written by a variety of authors, on the spectrum of groups currently considered within the Porifera, but with an emphasis on living forms).

Pisera AA (1997) Upper Jurassic siliceous sponges from the Swabian Alb: taxonomy and paleoecology. *Palaeontologie Polonica* 57: 3 219.

Reid REH (1958 1964) A monograph on the Upper Cretaceous Hexactinellida of Great Britain and Northern Ireland. *Palaeontographical Society (London) Monograph* 1958, Part 1, pp. i xlvi; 1959, Part 2, pp. xlvii xlviii, 1 26; 1961, Part 3, pp. 27 48; 1964, Part 4, pp. xlix cliv.

Rigby JK (1986) Sponges of the Burgess Shale (Middle Cambrian) British Columbia. *Palaeontographica Canadiana* 2.

Rigby JK (1987) Phylum Porifera. In: Boardman RS, Cheetham AH, and Rowell AJ (eds.) *Fossil Invertebrates,* pp. 116 139. Palo Alto, Oxford, London: Blackwell Scientific Publications.

Rigby JK (1991) Evolution of Paleozoic heteractinid calcareous sponges and demosponges patterns and records. In: Reitner J and Keupp H (eds.) *Fossil and Recent sponges,* pp. 83 101. Berlin and Heidelberg: Springer Verlag.

Rigby JK and Senowbari Daryan B (1996) Upper Permian inozoid, demospongid, and hexactinellid sponges from Djebel Tebaga, Tunisia. *The University of Kansas Paleontological Contributions, New Series,* 7.

Rigby JK, Senowbari Daryan B, and Liu H (1998) Sponges of the Permian Upper Capitan Limestone, Guadalupe Mountains, New Mexico and Texas. *Brigham Young University Geology Studies* 43: 19 117.

FOSSIL PLANTS

Contents

Angiosperms

P Kenrick, The Natural History Museum, London, UK

Introduction

The angiosperms, or flowering plants, have been the dominant vegetation on land for over 60 million years. The group encompasses an enormous diversity of forms, which reflects the variety of habitats that they occupy as well as their interactions with other organisms – in particular, animals. Conservative estimates indicate that there are in excess of 220 000 known living species of flowering plant, which is much more than all other land plants combined. Despite the ecological supremacy of this group and the huge diversity of living forms, in geological terms the angiosperms are comparative newcomers. Fossil evidence shows that angiosperms first appeared and began to diversify towards the end of the Mesozoic Era, about 130 million years ago. Whereas the early diversification of angiosperms is increasingly well documented, the origin of the group is a subject that has been shrouded in mystery and much debated. For many years, investigation has been hampered by a seemingly uninformative early fossil record, uncertainty about relationships among living species, and apparently insuperable gaps between certain aspects of flowering plant morphology and that of the closest relatives, the gymnosperms. Recently, there have been significant advances in our understanding of the relationships of flowering plants. Of particular importance has been the application of cladistic methods to the time-honoured approach of comparative morphology and the increasing use of molecular methods. From a geological perspective, the use of maceration and sieving techniques on Cretaceous clays has brought to light a wealth of minute but exquisitely preserved fossil flowers, documenting new early evidence on floral diversity. Together with analyses of the fossil record of pollen, these new data are providing a much clearer picture of the early evolution of flowering plants and their rapid rise to dominance.

Characteristics

Angiosperms are, for the most part, an easily recognizable group of plants: they are characterized by the production of flowers. The flower consists of four basic elements arranged in series around a central axis. These include (1) the sepals, which are an outer series of modified bracts, (2) the petals, an inner series of modified bracts that are often coloured, (3) the stamens, which contain the pollen, and (4) the carpels, which contain the ovules (seeds when fertilized). Not all of these units are present in all flowers. For example, within the magnoliids, which are a basal assemblage within angiosperms, there are both large multipart flowers (e.g., *Magnolia*) as well as minute, simple flowers comprising a single carpel and two stamens (e.g., *Piper*) (**Figure 1**). The enormous complexity of form, colour, and scent in flowers is linked to the evolution of animal pollinators, especially insects. A further major characteristic of angiosperms is that the ovules are enclosed within an ovary. Fertilization takes place when pollen lands on a special surface (the stigma) and germinates, producing a pollen tube that grows into the ovary and delivers the male nuclear material to the ovules. Angiosperms come in both woody and herbaceous forms. They possess a type of pollen with a distinctive cell wall. Broad leaves with reticulate venation are typical of many members of the group. The wood is also typically diagnostic. Seeds, wood, leaves, and pollen are frequently preserved as fossils (**Figures 2 and 3**), and, together, serve as principal sources of evidence on the geological history of the group. Flowers are more rarely preserved (**Figure 4**), because they are more likely to fragment into their constituent pieces.

Classification

Classifications of angiosperms based on comparative morphology are still widely used, but these long-standing and familiar systems are now under major revision due to the impact of molecular systematics.

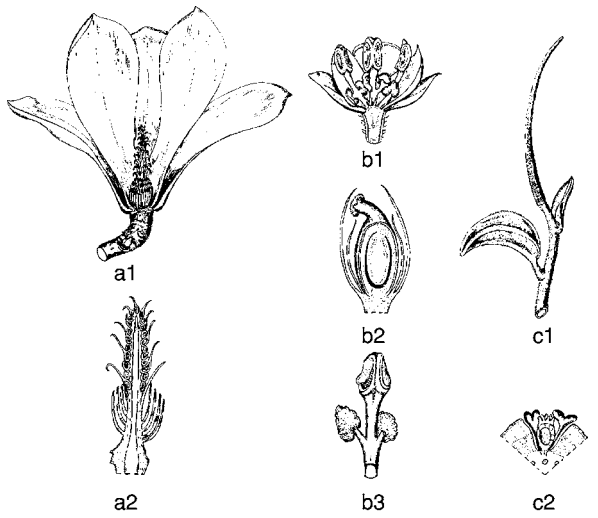


Figure 1 Floral diversity in magnoliid angiosperms. Left: *Magnolia* (Magnoliaceae). (a1) Section through flower showing petals and central receptacle bearing numerous stamens and ovules. (a2) Section through receptacle showing details of ovules in upper part and stamens in lower part. Middle: *Laurus* (Lauraceae). (b1) Male flower, showing petals and stamens. (b2) Female flower, showing ovary with one ovule. (b3) Single stamen detached from flower, showing flaplike valve at apex and two glands near the base. Right: *Peperomia* (Piperaceae). (c1) Elongate terminal inflorescence composed of numerous closely set simple flowers. (c2) Transverse section through inflorescence, showing single flower composed of central ovule and two stamens. Reproduced with permission from Hickey M and King C (1981) *100 Families of Flowering Plants*, p. 567. Cambridge: Cambridge University Press.

The family unit is the focal point of higher level classifications, and from 320 to 590 or more families are recognized, depending on taxonomist and system. Orders and higher level groupings have until recently been of little importance. With a few exceptions, these are difficult to recognize on morphological criteria, and the monophyletic status of such groupings is frequently questioned.

It is no exaggeration to say that the application of gene sequencing technology is revolutionizing the understanding of relationships among angiosperms at all levels. The use of molecular methods is more advanced in plants than in any other major group of organisms, and new classifications are based on a molecular systematic framework. A recent classification published by the Angiosperm Phylogeny Research Group recognized 462 families. These have been classified into 40 monophyletic orders, which are further categorized into a number of informally named higher groups (Figure 5). The outline classification illustrated in Figure 5 represents a major advance in knowledge of the relationships among angiosperms, and much of this has been worked out since 1990. Conservative estimates indicate that at least 10% of families are not monophyletic groups. Furthermore, although many families can be placed within higher groups, their relationships within these are still unclear. Many families are not yet included within orders. Much work remains to be done in fleshing out the details within and between families and in some cases orders, but the currently broad picture is unlikely to change in a radical way.

The root of the angiosperm phylogenetic tree lies within a basal grade of organization that has been dubbed the magnoliid dicots. Included in the magnoliids are families such as Amborellaceae, Austrobaileyaceae, Canellaceae, Chloranthaceae, Hydnoraceaceae,

Figure 2 *Acer trilobatum* (Miocene; Oeningen, Baden, Germany). The leaves of angiosperms are common in the fossil record beginning in the mid Cretaceous.

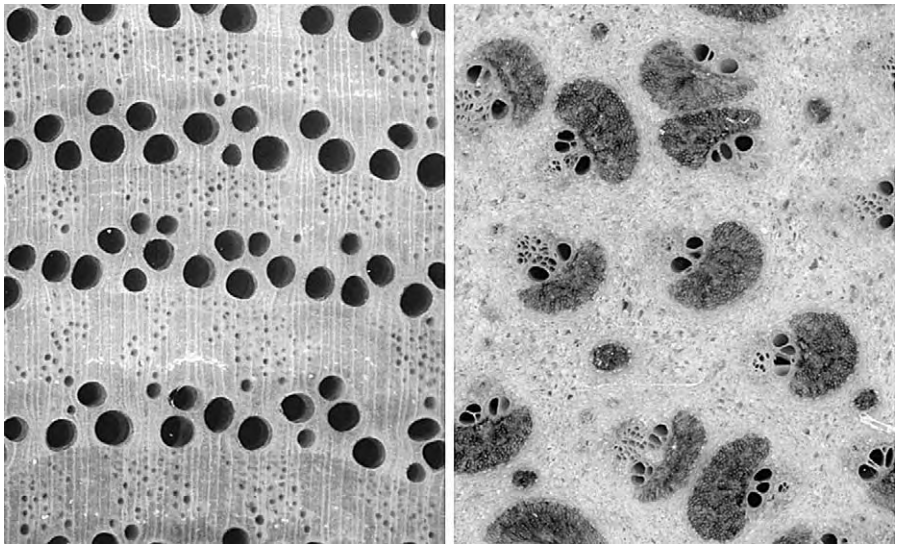


Figure 3 Angiosperm wood with its distinctive cell structure is first recorded in sediments of the mid Cretaceous (Albian stage). The images show two transverse sections through fossilized angiosperm wood. Left: *Quercus* (oak, Miocene; Saddle Mountain, Washington, USA). Right: *Palmoxylon* (palm, Oligocene; Antigua, Eastern Caribbean).

Figure 4 Fossil flower. *Porana oeningensis* (Miocene; Oeningen, Baden, Germany).

Illiciaceae, Nymphaeaceae [+ Cabombaceae], Rafflesiaceae, Schisandraceae, Trimeniaceae, and Winteraceae. The group also encompasses the orders Ceratophyllales, Laurales, Magnoliales, and Piperales. Flowers are basically simple in the magnoliids. The large, complex flowers of magnolias and water lilies are exceptions (Figure 1). The monotypic Amborellaceae is probably the most basal named taxon and sister group, compared to all other angiosperms. The interrelations among the basal orders of angiosperms remain elusive. These groups include the magnolias, peppers, and laurels. The informally named 'monocots' (monocotyledons) is one of the few higher groups that has been widely recognized on morphological and molecular criteria (Figure 5). The monocots are clearly monophyletic, and this large clade comprises ~22% of living species arrayed in 10 orders, including familiar groups such as the lilies, palms, and grasses. All other flowering plants are in the informally named 'eudicots' clade (Figure 5). Within eudicots there is increasing support for a large subgroup with predominantly five-part flowers, termed 'core eudicots'. This group contains two large subclades, the rosids and asterids. Over 75% of living species diversity is contained within the eudicots.

Angiosperm Origins

The origin of flowering plants is one of the most widely discussed and enduring mysteries in evolutionary botany. This whole area has remained controversial, and it is plagued by competing and contradictory ideas. There are several reasons for this. First, until recently there was no clear consensus about which of a diverse group of primitive living families is actually basal within angiosperms. In other words, the root of the angiosperm family tree was unknown, and therefore what constitutes the primitive condition in key

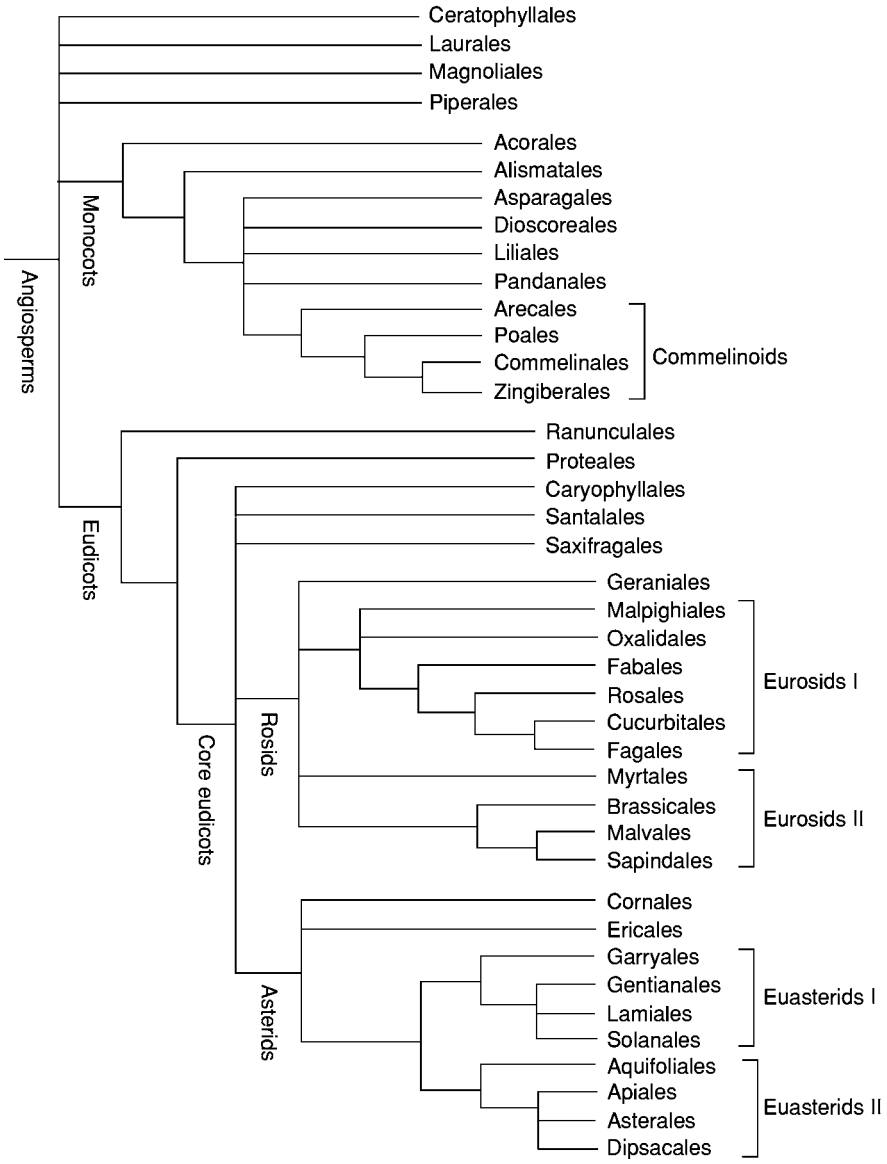


Figure 5 The 1998 Angiosperm Phylogeny Group classification of living angiosperms at the ordinal level. The recognized groupings are based on both molecular and morphological characteristics. Some families (not shown) are not placed within orders.

features, such as the flower, was difficult to assess. Second, although the fossil record is replete with information on pollen, leaves, fruit, and wood, the record of flowers is much sparser. First-hand information on the nature of early Cretaceous flowers has been hard to find. This means that one of the crucial features of flowering plants has remained poorly characterized in early members of the group. Also, because most fossils represent organs or plant fragments rather than whole individuals, their implications for angiosperm origins are not always easily interpreted. Third, although it is clear that angiosperms arose from within the gymnosperms, there are conflicting ideas on which living or extinct groups of gymnosperms constitute the closest relatives. Furthermore, living angiosperms have diverged substantially from related plants in a number of ways. This makes it difficult to draw comparisons between, for example, aspects of the angiosperm flower and putative equivalent structures in gymnosperms. In other words, angiosperms are divergent, and this obscures their relationships to other groups of plants. Fourth, 'angiosperm origins' is really a discussion about the origins of two groups, not one, and this can lead to confusion and argument (Figure 6). There is the issue of the origin of the crown group, which is the group that contains all living species and their most recent common ancestor. Members of this group typically

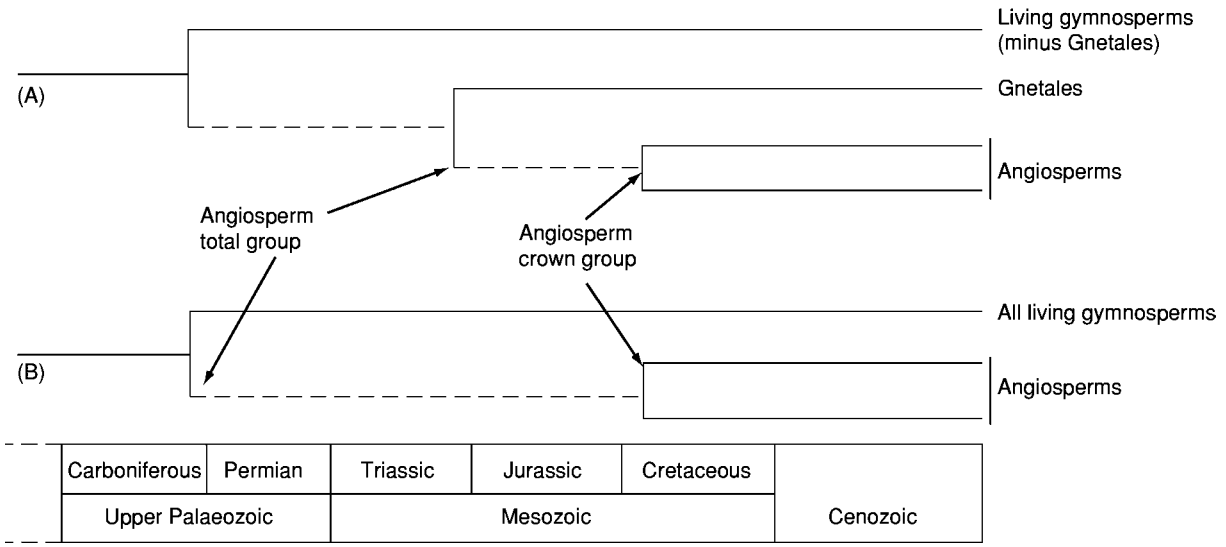


Figure 6 Two hypotheses (A and B) of angiosperm origins. In both hypotheses, the angiosperm crown group (the group that contains the most recent common ancestor of all living species) originated and diversified during the Early to mid Cretaceous. The two alternatives have implications for the origin of the angiosperm total group. (A) Anthophyte hypothesis. Gymnosperms in the Gnetales are the living sister group to angiosperms. Gnetales are documented in the fossil record long before unequivocal angiosperms. This hypothesis implies an angiosperm ghost lineage extending into the Late Triassic. (B) Gne pine hypothesis. Living gymnosperms are a monophyletic group the Gnetales (not shown) are most closely related to conifers, not angiosperms. This whole gymnosperm clade is sister group to angiosperms. This hypothesis implies an angiosperm ghost lineage extending back as far as the Carboniferous Period. Note that direct fossil evidence of angiosperm ghost lineage members is scant. However, there are a number of candidates among Mesozoic gymnosperms.

possess all of the familiar features that we associate with modern angiosperms. There is also the question of the origin of the total group, which contains the crown group plus closely related extinct plants that possess some, but not all, of the attributes of modern angiosperms. This is the group that formed when the angiosperm lineage split from the lineage leading to its sister group in living gymnosperms. It is important to understand that these are two different but related issues, and that although both group concepts encapsulate aspects of the origin of angiosperms, they do not mean quite the same thing. Recent advances in phylogenetic research and new discoveries in the fossil record have gone a long way towards addressing some of these problems.

One of the unequivocal findings of recent phylogenetic research is that angiosperms are a monophyletic group. This means that it is correct to dispense with earlier theories of multiple independent origins of flowering plants. There is no longer any doubt that the group had a single origin: all angiosperms share a common ancestor. Molecular phylogenetic analyses have identified the water lilies (Nymphaeales: Cabombaceae and Nymphaeaceae), together with four other small groups of flowering plants (Amborellaceae, Illiciales, Trimeniaceae, and Austrobaileyaceae), as the first diverging lineages from the main branch of the angiosperm phylogenetic tree. This means that researchers are now very close to understanding exactly where the root of angiosperms lies within the vast number of living species. From the perspective of the fossil record, the most interesting data have come through the application of a new method of looking for early fossil evidence for flowers. This involves the disaggregation and sieving of Cretaceous clays to separate mineral from organic constituents. This approach has yielded remarkable new information on microscopic fossilized flowers preserved either as charcoal or in a mummified state. These data provide clear evidence of the Chloranthaceae and the Nymphaeales in Early Cretaceous sediments. Both groups are thought to be basal based on phylogenetic research. Together with evidence from the fossil record of dispersed pollen, a conservative minimum age estimate for the origin of the angiosperm crown group would be Early Cretaceous (Valanginian). Furthermore, the absence of distinctive triaperturate pollen (diagnostic of a large subclade within angiosperms: eudicots) in well-documented Triassic and Jurassic palynofloras argues against a long period of cryptic evolution of the crown group implied by estimates based on molecular clocks.

The origin of angiosperms also encompasses the dating of the age at which the lineage leading to

flowering plants split from living gymnosperms (i.e., the total group) as well as the discovery and description of fossils of an intermediate nature (Figure 6). Dating the origin of the total group provides only minimum age estimates. Counterintuitively, this does not depend entirely on finding early angiosperm-like fossils. Age estimates of the angiosperm total group can also be made by direct inference from the stratigraphic range of the angiosperm sister group (Figure 6). To do this, it is essential to know how angiosperms are related to gymnosperms (*see* **Fossil Plants:** Gymnosperms), but this is one of the most controversial areas in systematic botany. There are two main ideas on relationship and these have radically different implications for the origin of angiosperms and for the relevance of extinct Mesozoic gymnosperms (Figure 6). Comparative morphology indicates that angiosperms are most closely related to living gymnosperms in the Gnetales and to extinct gymnosperms in the Pentoxylales and the Bennettitales. This has been termed the 'anthophyte hypothesis'. Molecular data tell a very different story: they place Gnetales within Coniferales as sister group to Pinaceae. Angiosperms emerge as sister group to a monophyletic group comprising all living gymnosperms. This has been termed the 'gne-pine hypothesis'. Neither hypothesis, though, is particularly strongly supported by the current data. Furthermore, because molecular data cannot deal with the fossil groups, it is, of course, unclear where in molecular phylogenetic trees the fossils would in fact fall out.

The anthophyte and the gne-pine hypotheses have very different implications for the origin of angiosperms and for the relevance to the question of extinct Mesozoic gymnosperms (Figure 6). The anthophyte hypothesis indicates that the lineage leading to angiosperms originated in the mid-Triassic, because this is when fossils attributable to living gymnosperms in the Gnetales first appear. If this hypothesis is correct, the search for angiosperm relatives should be confined to extinct gymnosperms of Late Triassic and Jurassic ages. The gne-pine hypothesis, however, removes Gnetales as sister group to angiosperms and replaces it with the entire gymnosperm crown group. Living gymnosperms are much more remote in a phylogenetic sense from angiosperms. This would indicate a much more ancient origin of the angiosperm lineage, which could reach back as far as the Late Carboniferous. Under this hypothesis, the search for angiosperm relatives should cast a wider net that encompasses gymnosperms of Early Mesozoic as well as Late Palaeozoic ages. Even though it is likely that the lineage leading to angiosperms diverged from other plants during the Triassic or earlier, current knowledge of Mesozoic floras makes it unlikely that angiosperms were diverse or abundant before the Cretaceous.

The search for early angiosperm-like fossils has to confront the problem of recognition. In most respects, early fossils are likely to be more similar to gymnosperms of one sort or another than what might currently be considered as a typical member of the angiosperm crown group. This is undoubtedly one reason why such fossils have proven to be so elusive. There are several plausible candidates among the Mesozoic gymnosperms, but questions remain in all cases due to missing data on key organs. The most recent fossil to emerge as potential sister group to the angiosperm crown group is *Archaefructus* (Figure 7). This plant comes from the famous Early Cretaceous Jehol Biota of north-eastern China. *Archaefructus* is thought to have been a herbaceous aquatic. It had thin stems and finely dissected compound leaves. The 'flower' presents a unique combination of features. It bore clusters of structures interpreted as carpels at the shoot apex. These were subtended by a zone that bore stamens. The pollen was monosulcate, resembling common forms found dispersed in Cretaceous sediments. Unlike most modern angiosperms, though, the 'flower' of *Archaefructus* did

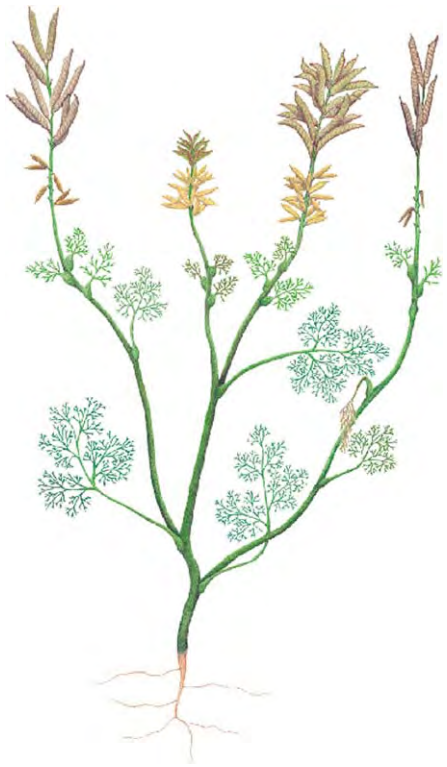


Figure 7 *Archaefructus liaoningensis.* An Early Cretaceous plant thought to be an early angiosperm or close relative of the angio sperms. Reproduced with permission from Sun G, Ji Q, Dilcher DL, *et al.* (2002) Archaefructacea, a new basal angiosperm family. *Science* 296(5569): 899 904.

not have petals and sepals. On the face of it, *Archaefructus* would seem to fall within the angiosperm total group but outside the crown group. As such, it could provide the first clear evidence of a fossil intermediate between living gymnosperms and living angiosperms. Some aspects of the interpretation of *Archaefructus* are, however, controversial, and the precise nature of its relationship with angiosperms is still in question.

An important source of evidence on the origin of angiosperms is the fossil record of pollen (Figure 8). Because pollen is produced in vast quantities and is dispersed over wide areas, the pollen 'footprint' of plants is relatively large. Pollen could therefore provide the most precise constraint on age as well as providing a window onto floral changes through time. The disadvantage of pollen is that its taxonomic resolution is generally low, due to the small suite of available characters. Whereas some pollen seems to be diagnostic of higher level groups, other pollen is more difficult to place within families or orders. Furthermore, it is not always easy and may not be possible in some cases to distinguish clearly the angiosperm crown group or total group from closely related extinct gymnosperms based on pollen alone. Angiosperm-like pollen has been reported from sediments of Triassic and Jurassic ages. The so-called Crinopolles group (Late Triassic) has a very angiosperm-like monosulcate aperture combined with a reticulate–columellar exine. Analysis of exine structure using transmission electron microscopy indicates some differences from angiosperms. A relationship between Crinopolles pollen grains and angiosperms is possible, but it is unlikely to be indicative of the crown group.

Angiosperm Diversification

The early diversification of angiosperms was dominated by plants related to a phylogenetically basal grade (the magnoliid dicots; Figure 1). These are represented in Early Cretaceous sediments by leaves, pollen, and, increasingly, fossilized floral organs, fruits, and seeds. The maceration of clays from Early Cretaceous localities in Portugal and eastern North America, followed by sieving of the organic component, has been a key methodological breakthrough. This technique, combined with scanning electron microscopy of the organics, is revealing a huge diversity of plant organs attributable to charcoal-preserved or mummified angiosperms. One striking feature of the plant assemblages recovered in this way is that all of the early flowers are small – usually less than 2 mm long (Figure 9). It seems unlikely that this is due to depositional biases such as current sorting, because other larger plant parts belonging to conifers are usually found in association. Early fossil flowers are simple, being composed of relatively few parts and an undifferentiated perianth. Stamens are small and the pollen sacs typically disperse their pollen through a hinged flap or valve. The

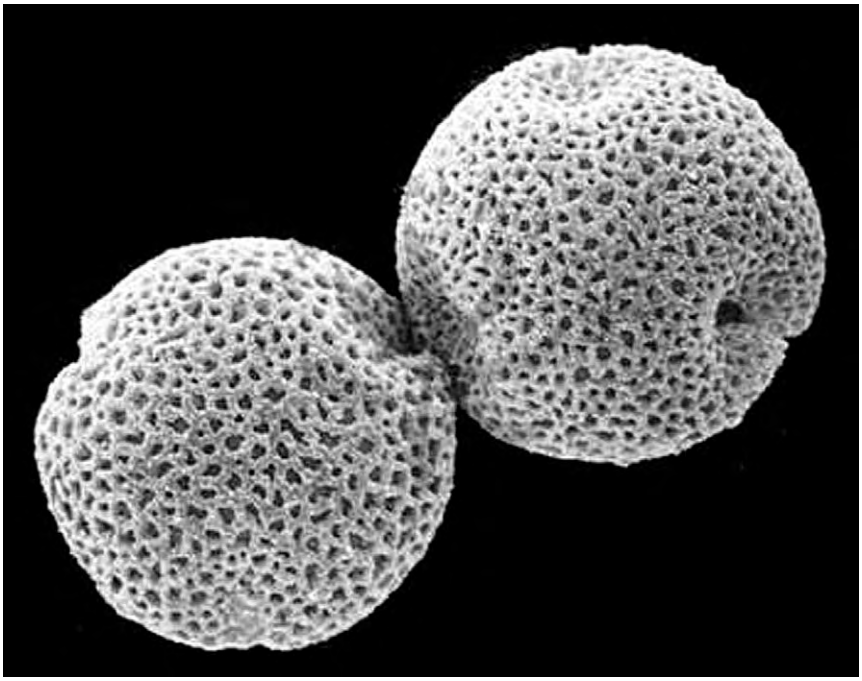


Figure 8 Pollen of living ash (*Fraxinus*).

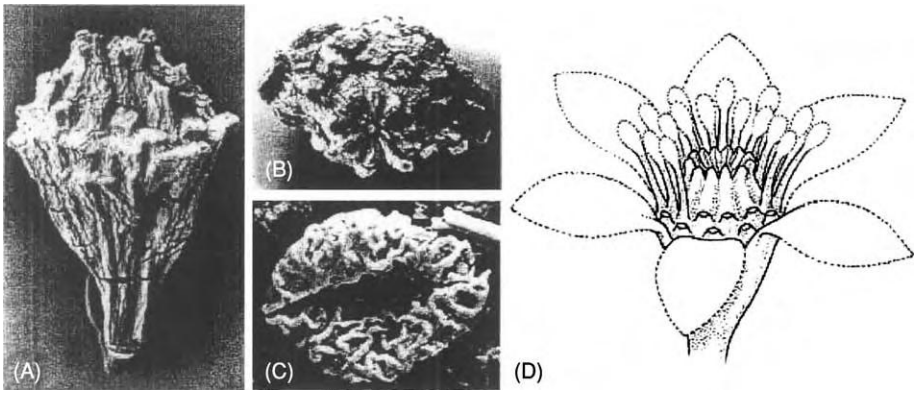


Figure 9 Fossil water lily from the Early Cretaceous (Barremian or Aptian) of Portugal. Scanning electron micrographs of flower (A, B) and pollen (C), and reconstruction of flower (D). Reproduced with permission from Friis EM, Pedersen KR, and Crane PR (2001) Fossil evidence of water lilies (Nyphaeales) in the Early Cretaceous. *Nature* 410: 357 360.

Figure 10 Water lilies, one of the most basal lineages of flowering plants.

stigmatic surface of the carpels is poorly differentiated. Early Cretaceous flowers and fruits are rarely assignable to living families or orders. They show a blend of features of the general magnoliid type.

The earliest fossils assignable to modern families have been recovered from the Barremian–Aptian stages of the Early Cretaceous. These floral organs, fruits, and seeds are attributable to water lilies (Nymphaeaceae) (Figures 9 and 10) and to Chloranthaceae. In marked contrast to the magnoliid dicots, the Cretaceous fossil record of monocots is depauperate. This is surprising in view of the basal phylogenetic position of the monocots, and it may reflect a representation bias and perhaps also recognition bias. Monocots are clearly present by the Late Cretaceous. The earliest fossils include fruits of the ginger family (Zingiberales) and leaves and stems of palms (Arecales). Many modern families of monocots are recognizable by the Early Cenozoic. The earliest fossil evidence for the eudicots comes from grains of triaperturate pollen dispersed in sediments of the Barremian–Aptian stages. The origins of eudicots therefore marginally predate the major diversification and ecological radiation of angiosperms.

Among the first modern families to appear during the Albian stage are the plane and buttonwood trees (Platanaceae) and probably also the box family (Buxaceae). Many more living families are recognizable by the Turonian stage. These include the witch hazels and sweet gums (Hamamelidaceae) and the saxifrages (Saxifragales). The rosids also first appear in the Turonian stage. There is evidence for the capers (Capparales), and somewhat later in the Santonian stage are records of the Myrtales; walnuts, hickories, and pecan nuts group (Juglandales); Myricales; and the birches, alders, beeches, and oaks group (Fagales). The main asterid lineage is also present by the Turonian stage. These are represented by flowers of ericalean affinity, and members of the Hydrangeaceae have been recorded in the Coniacian–Santonian stages. Nearly all of the main eudicot clades are represented by at least one member of their lineage in the Upper Cretaceous. Some notable groups do not, however, appear until the Maastrichtian stage or the Early Cenozoic. These include the pea or bean family (Fabaceae), the sunflower family (Asteraceae), the Lamiales, and the Gentianales, which together comprise 45% of living species diversity in eudicots. The three subfamilies of pea are well documented by flowers and fruits in the Eocene, but pollen grains of Caesalpinioideae have been recorded from the Maastrichtian stage of the Late Cretaceous. The earliest unequivocal records of the Asteraceae, Lamiales, and Gentianales come from the Palaeogene. The fact that most species-rich groups are known from relatively young fossils indicates that a significant proportion of eudicot diversity is the

product of relatively recent radiations that occurred during the second half of angiosperm evolution. The evolutionary basis for the rapid diversification of specific eudicot clades remains unknown.

The rapid rise to prominence of angiosperms during the mid-Cretaceous marks the transition from Mesozoic ecosystems dominated by ferns and gymnosperms to those of the Late Cretaceous and Cenozoic, when flowering plants predominate. The record of fossil pollen in the northern hemisphere shows that the initial increase in angiosperm diversity occurred at low palaeolatitudes (**Figure 11**). During the Early Cretaceous, low-latitude areas experienced semiarid or seasonally arid climates. These conditions may have promoted a weedy life history with precocious reproduction favouring herbaceous annuals rather than long-lived woody perennials. There is some fossil evidence to support this. With the exception of one disputed record, angiosperm wood has not been recorded from rocks that predate the Albian. Angiosperm leaves are also extremely rare in these older sediments. Furthermore, it seems likely that some of the angiosperms from other Lower Cretaceous assemblages (e.g., Crato Formation, Brazil, and Portugal) were aquatic. These data indicate that the earliest angiosperms were predominantly herbaceous and that there may have been considerable diversity in aquatic habitats.

By the end of the Cretaceous period, flowering plants represented about 50–80% of land plant species diversity, but pockets of gymnosperm- and fern-dominated vegetation persisted. This notable change in floral composition predates the end of the Mesozoic Era by some 30 million years and was more radical than that which occurred at the Cretaceous–Tertiary boundary. Some groups of flowering plants that are ecologically prominent today were completely absent from the Cretaceous Period. Grasses, for example, did not become abundant and widespread until the Late Oligocene and Miocene. Ferns may well have played a more prominent role as herbaceous ground cover in prairie vegetation prior to the mid-Cenozoic.

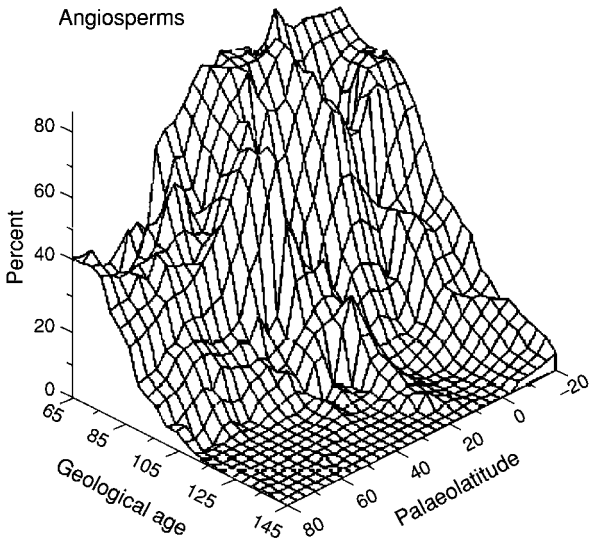


Figure 11 Analyses of dispersed pollen grains indicate that the diversification of angiosperms began at low palaeolatitude. Reproduced with permission from Crane PR and Lidgard S (1989) Angiosperm diversification and palaeolatitudinal gradients in Cretaceous floristic diversity. *Science* 246: 675 678.

The diversification of angiosperms has often been linked to the evolution of various animal groups. But proving cause and effect in the fossil record is often very difficult. For example, the rapid diversification of angiosperms during the mid-Cretaceous coincides broadly with a transition from sauropod-dominated to ornithopod-dominated dinosaur faunas (*see* **Fossil Vertebrates:** Dinosaurs). It is difficult to show whether these changes are in some way linked or just coincidental. A stronger case can be made for a link between insects (*see* **Fossil Invertebrates:** Insects) and flowering plants. The diversification of pollen- and nectar-collecting insects has been linked to flowering plants on the supposition that insect pollination provides new possibilities for reproductive isolation and therefore elevation of speciation rates. Compared to wind pollination, which is widespread in gymnosperms, insect pollination may permit more effective out-crossing at lower population densities and in a greater range of environments, thereby reducing extinction rates. Pollination biology may provide a plausible explanation for spectacular diversity in some angiosperms families (e.g., orchids), but the extent to which this provides a general explanation for the mid-Cretaceous rise of angiosperms is uncertain. There is evidence that insect pollination was an aspect of the biology of some extinct gymnosperms (e.g., Bennettitales) and some living gymnosperms (e.g., Gnetales). Early angiosperms may therefore have had the opportunity to co-opt pollinators from previously established relationships with other groups of seed plants.

The fossil record of flowers from the Early Cretaceous clearly indicates that early angiosperms were insect pollinated. Many of the features that have been documented in early fossil flowers are comparable to those found in modern insect-pollinated groups. The anthers show that pollen output was low, and the pollen grains are often smaller than the most effective size for wind dispersal. Release of pollen from anthers occurred via hinged flaps, and the individual grains are often covered by a pollenkitt-like substance. The stigmatic surfaces of the carpels are generally unelaborated. These features are consistent with

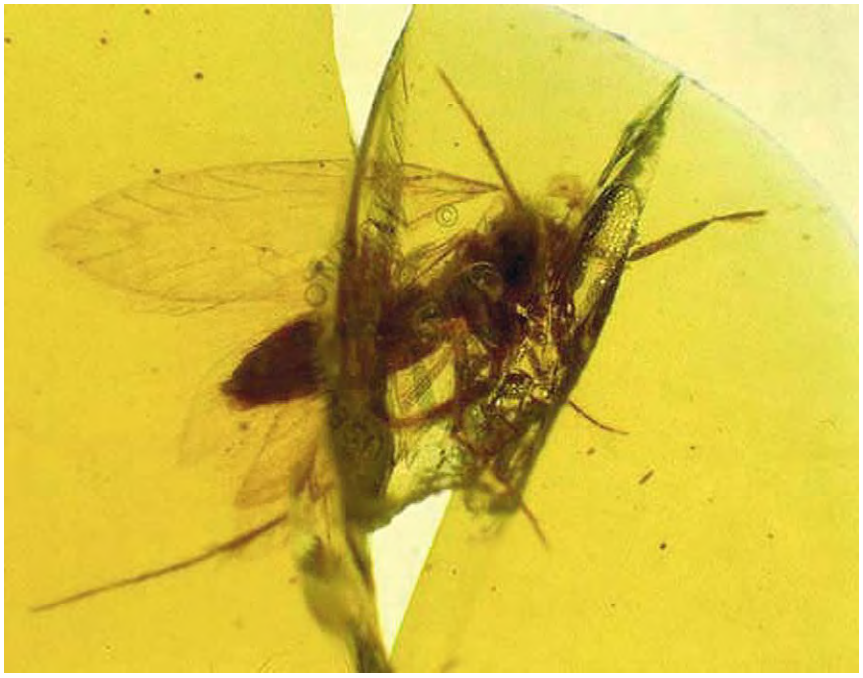


Figure 12 Moth belonging to the living pollen feeding family Micropterygidae, in Burmese amber of Late Cretaceous age.

pollination by pollen-collecting or pollen-eating insects (Figure 12). Flowers pollinated by nectar-collecting Hymenoptera and Lepidoptera occur in more derived groups of angiosperms and appear later in the fossil record. Even though early angiosperms may have been very similar to their living relatives in pollination syndrome, modes of fruit and seed dispersal were probably very different. Cretaceous fruits and seeds are generally very small compared to their modern relatives, and there is no evidence of specialized mammal or bird dispersal of fruits and seeds. Among basal groups of angiosperms, the evolution of fleshy fruits, arillate seeds, and other adaptations for dispersal by animals seems to be correlated with the evolution of fruit- and seed-eating birds and mammals, perhaps during the latest Cretaceous but most strikingly during the Early Tertiary.

See Also

Fossil Invertebrates: Insects. **Fossil Plants:** Gymnosperms. **Fossil Vertebrates:** Dinosaurs.

Further Reading

Angiosperm Phylogeny Research Group (APG) (1998) An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85: 531 553.

Chase MW, Fay MF, and Savolainen V (2000) Higher level classification in the angiosperms: new insights from the perspective of DNA sequence data. *Taxon* 49(4): 685 704.

Crane PR, Friis EM, and Pedersen KR (1995) The origin and early diversification of angiosperms. *Nature* 374: 27 33.

Crane PR and Lidgard S (1989) Angiosperm diversification and palaeolatitudinal gradients in Cretaceous floristic diversity. *Science* 246: 675 678.

Friis EM, Chaloner WG, and Crane PR (eds.) (1987) *The Origin of Angiosperms and Their Biological Consequences*. Cambridge: Cambridge University Press.

Friis EM, Pedersen KR, and Crane PR (1999) Angiosperm origin and radiation. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology II*, pp. 97 102. Oxford: Blackwell Science.

Herendeen PS and Crane PR (1995) The fossil history of the monocotyledons. In: Rudall PJ, Cribb PJ, Cuttler DF, and Humphries CJ (eds.) *Monocotyledons: Systematics and Evolution*, pp. 1 21. Kew: Royal Botanical Gardens.

Magallón S, Crane PR, and Herendeen PS (1999) Phylogenetic pattern, diversity, and diversification of eudicots. *Annals of the Missouri Botanical Garden* 86: 297 372.

Sun G, Ji Q, Dilcher DL, *et al.* (2002) Archaefructaceae, a new basal angiosperm family. *Science* 296(5569): 899 904.

Taylor DW and Hickey LJ (eds.) (1995) *Flowering Plant Origin, Evolution Phylogeny*. New York: Chapman & Hall.

Wikstrom N, Savolainen V, and Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London, Series B, Biological Sciences* 268(1482): 2211 2220.

Calcareous Algae

J C Braga, University of Granada, Granada, Spain
R Riding, Cardiff University, Cardiff, UK

Introduction

Algae are anatomically simple photosynthetic plants, lacking vascular tissues, roots, and leaves. They constitute a heterogeneous group comprising many divisions (*see* **Palaeontology**) (botanically equivalent to phyla). In addition, it is convenient to consider some photosynthetic bacteria (*see* **Biosediments and Biofilms**), such as Cyanobacteria (blue-green algae) along with eukaryotic algae. The algal body (thallus) can be unicellular, or multicellular with a certain degree of complexity due to cell differentiation. Although most algal groups are aquatic, some inhabit soils and subaerial environments.

Calcareous algae precipitate $CaCO_3$ during life at particular sites on or within their thalli. Calcification is a cross-systematic feature, occurring in diverse divisions, in freshwater and marine environments, and in planktic and benthic forms. However, fewer than 10% of extant benthic marine algae calcify, and they principally belong to red and green algae. Only one extant calcified genus of brown algae (*Padina*) is known, and cyanobacteria rarely calcify in marine environments at the present day although they did so at times in the past. In freshwater, charaleans and a few other chlorophytes calcify, and cyanobacterial calcification is common in calcareous streams and lakes. Photosynthesis raises pH and increases saturation state with respect to $CaCO_3$ minerals. However, the degree of biological control over calcification depends mainly on the location of sites for precipitation within the organism. In general, cyanobacterial calcification is environmentally controlled, and calcification by chlorophytes is also environmentally dependent. In contrast, coralline red algae and coccolithophores calcify intracellularly and relatively closely control their calcification. This is reflected in their wider latitudinal distribution in marine environments.

Fossils that may be calcified algae have been reported from the Neoproterozoic but the substantial record of calcified cyanobacteria and algae is mainly Phanerozoic (Figure 1). Their long geological history and wide environmental distribution result in calcified bacteria and algae being significant producers of calcium carbonate sediment in marine and freshwater deposits of many ages. Sedimentary products range from in-place reefal masses (*see* **Sedimentary Environments:** Reefs ('Build-Ups')), through nodules to coarse and fine bioclastic fragments. Postmortem disintegration of a wide variety of calcified algae and cyanobacteria has probably extensively produced mud-sand-gravel sediment, although often only coarser components are readily recognizable.

Rhodophyta (Red Algae)

Corallinales

Coralline algae are the major extant group of calcified marine red algae. Their thalli consist of branched cell filaments with a coherent unified, pseudoparenchymatous, organization. Growth is achieved by addition of new cells through cell division at the tip of each filament. Most present-day corallinaleans are heavily calcified by high magnesium calcite precipitated in the cell wall (Figure 2). This consists of a layer of crystals parallel to the wall from which radial crystals grow towards the cell interior. In non-geniculate (non-articulated) corallines all vegetative cells,

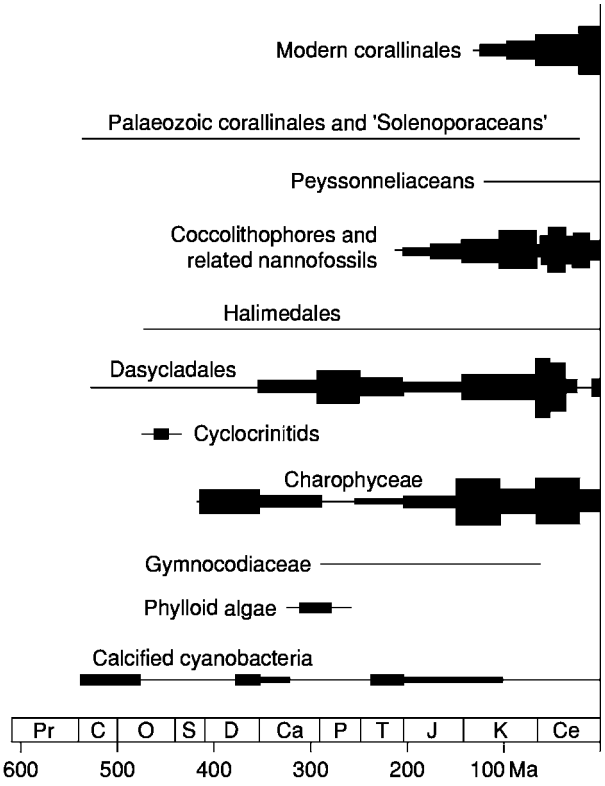


Figure 1 Stratigraphic range of the major groups of calcareous algae. Width of bars indicates the relative diversity of the group in each period. Pr: Proterozoic, C: Cambrian, O: Ordovician, S: Silurian, D: Devonian, Ca: Carboniferous, P: Permian, T: Triassic, J: Jurassic, K: Cretaceous, Ce: Cenozoic.

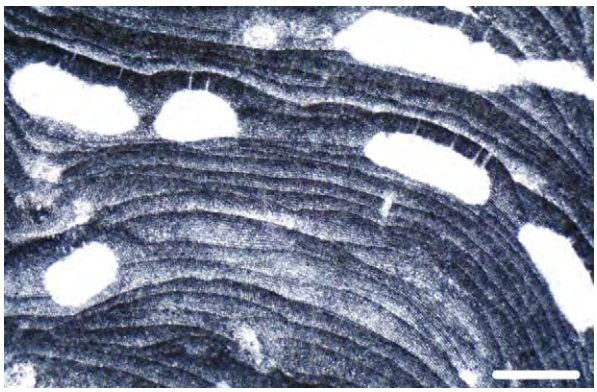


Figure 2 Section of the coralline alga *Lithothamnion*. The plant body consists of closely packed cell filaments with calcitic walls. Reproductive structures (conceptacles, multiporate in this example) remain empty after release of the spores. Dark and lighter areas represent seasonal growth zones. Upper Miocene, south eastern Spain. Scale bar 500 microns.

except those at the surface (epithallial cells), are calcified. Consequently, the entire plant has a rigid structure with high preservation potential. Non-geniculate corallines can grow on one another, and on firm substrates, including other plants, or live unattached on mobile sediments. They can form thick crusts and develop long protuberances and branches. Individual plants can reach several centimetres in size. In geniculate (articulated) species, calcified segments of the thallus are separated from one another by non-calcified joints (genicula). These articulated corallines form erect branching thalli usually attached to the substrate by an encrusting holdfast. They have a bushy appearance and grow up to 10 cm in height. After death, articulated corallines disintegrate and calcified segments are shed as sand and gravel particles.

Sexual and asexual reproduction both occur in the life cycle of most extant species. Gametes (male and female) and spores are produced in separate plants that otherwise exhibit similar vegetative morphology. Gametes and spores are produced in conceptacles within the thallus. The shape and size of conceptacles producing male gametes (and sometimes those producing female gametes too) differ from those in which spores develop. These differences can be recognised in fossil plants (Figure 2). The nature and number of openings (pores) of spore-producing conceptacles are key characters in corallinalean taxonomy.

Coralline algae constitute a cosmopolitan marine group with tropic to polar distribution and from the intertidal zone to depths of 270 m. In addition to this wide environmental range, they can tolerate very variable levels of nutrient supply and turbulence, as well as hypersaline to brackish coastal settings. Corallines are significant components of tropical shallow-water reefs (*see* **Sedimentary Environments:** Reefs ('Build-Ups')) but are also common on seafloors with low terrigenous supply at depths to 100–120 m in all climatic belts. In both shallow and deeper settings, they build rigid frameworks or lie unattached on loose substrates where they form nodules called rhodoliths.

Despite the wide geographic and environmental distribution of corallines as a whole, individual families and sub-families exhibit substantially differing depth-temperature distributions. Sporolithaceans are almost entirely restricted to low latitudes where they mainly occupy deep-water or cryptic habitats. Melobesioids dominate deep-water coralline assemblages in low–mid latitudes and occupy shallow water in high latitudes. Lithophylloids and mastophoroids live mainly in shallow water in low–mid latitudes, mastophoroids predominate in the tropics, and lithophylloids are more common in sub-tropical and warm temperate conditions. These habitat preferences of different taxa, together with variations in growth morphology according to levels of turbulence and light intensity, make fossil corallines valuable palaeoenvironmental indicators.

Coralline-like algae are known in the Ordovician (e.g., *Arenigiphyllum*) and Silurian (e.g., *Graticula*). The evolutionary relationships of these fossils, and of Late Palaeozoic fossils that have been called ancestral corallines (e.g., *Archaeolithophyllum*, see Phylloid Algae), to younger examples are still being elucidated. The continuous well-documented subsequent history of corallines commenced in the Early Cretaceous, and the group diversified throughout the Late Cretaceous and Cenozoic (Figure 1). About 30 extant genera are recognized.

'Solenoporaceae'

Solenoporaceans are traditionally regarded as an extinct family of red algae, morphologically similar to but simpler than corallines, ranging from the Cambrian to the Palaeogene. The superficial similarities that unite the 'Solenoporaceae' are simple, nodular form and an internal structure composed of narrow, juxtaposed branching tubes with diameters up to 100 microns (Figure 3). Details that indicate systematic differences include the cross-sectional shape of the tubes and the presence or absence of cross-partitions. However, the group is heterogeneous and contains some sponges as well as red algae, together with fossils resembling cyanobacteria. The type species of *Solenopora* is a chaetetid sponge. Some Ordovician and Silurian fossils formerly attributed to *Solenopora*, such as *Graticula*, closely resemble the extant coralline *Sporolithon* (see Corallinales above).

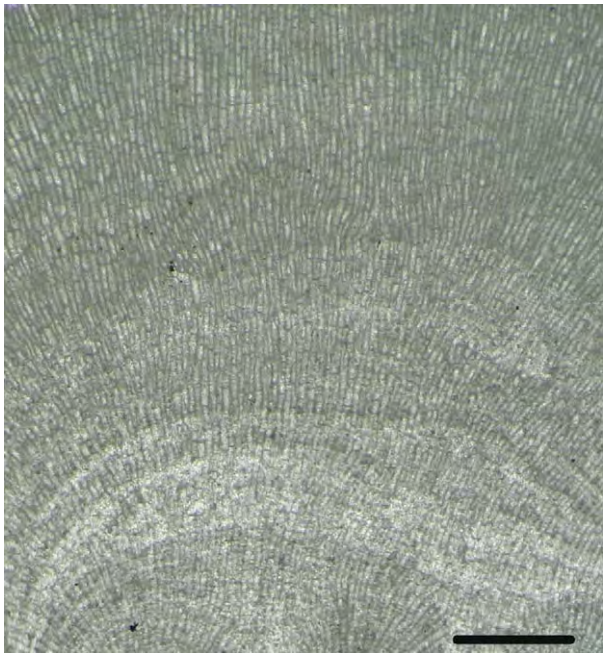


Figure 3 Section of *Solenoporella*, fossil attributed to 'solenoporaceans'. The densely packed cell filaments with transverse cell partitions suggest an affinity with red algae similar to Corallinales. Upper Jurassic, southern England. Scale bar 1 mm.

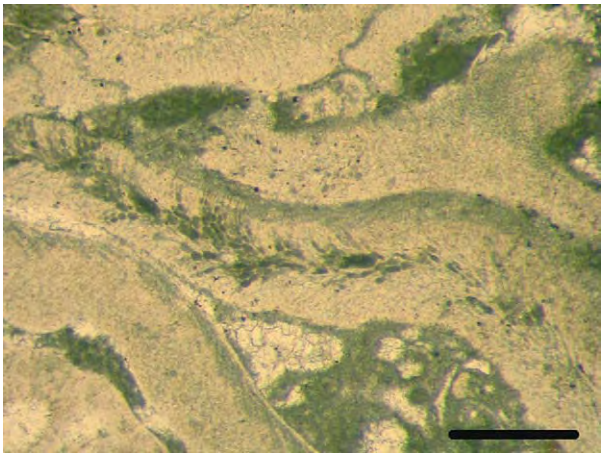


Figure 4 Section of *Polystrata*, a peysonneliacean alga. The thalli consist of cell filaments. The aragonitic nature of the original cell walls usually results in poor preservation of the internal structure. Paleocene, western Pyrenees, Spain. Scale bar 500 microns.

Peyssonneliaceae

Calcification in peyssonneliaceans (formerly squamaricaeans) is more variable than in corallines and is aragonitic, but otherwise peyssonneliaceans and corallines are broadly similar in consisting of branched cell filaments, forming a coherent thallus structure, and having intracellular calcification (Figure 4). Peyssonneliaceans occur as encrusting and leaf-like thalli up to 2 mm thick. Reproductive structures are rarely preserved in fossils. Aragonite may be densely or weakly deposited in the cell wall. Consequently, geological preservation of peyssonneliaceans ranges from good, with faithful preservation of the multicellular thallus, to very poor. The fossil record of peyssonneliaceans is consequently less well known than that of corallines, but they are common in the Palaeogene to present-day carbonate deposits (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves), and also build rhodolith nodules. External basal cement-like crusts commonly form, similar to those seen on crustose corallines. The confirmed range of peyssonneliaceans is Early Cretaceous to present-day, but they show similarities with some Carboniferous phylloids (see Phylloid Algae below) such as *Archaeolithophyllum* (see Corallinales above). Extant genera include *Polystrata* (= *Ethelia*) and *Peyssonnelia*. The depth and temperature ranges of present-day peyssonneliaceans are broadly similar to those of coralline algae.

Haptophyta

Coccolithophorales

Coccolithophores are a major group of nannoplankton (size 60 microns or less) and are the most important extant primary producers precipitating $CaCO_3$. They are mostly unicellular coccoid (non-motile) algae, although some species possess life stages of motile flagellate cells, non-motile colonies, or filaments. Most coccolithophores have one or more layers of organic (polysaccharide) scales surrounding the plasma membrane. Calcified scales, called coccoliths, form a cover termed the coccosphere (Figure 5) that is external to any additional organic scales that are present. The coccoliths are low-magnesium calcite in an organic matrix and range 1–25 microns in size. In different species, coccoliths differ widely in shape and arrangement, generating distinctive coccospheres composed of 10–100 coccoliths. Vegetative (fission) and sexual reproduction both occur. Life cycles are diverse among coccolithophoraleans but include a non-motile stage alternating with one or more motile phases. The non-motile stage can be benthic and colonial. Both motile and non-motile phases can bear coccoliths.

There are two basic types of coccolith: heterococcoliths composed of crystal elements of diverse size and shape (Figure 5), and holococcoliths made up of small identical crystals. Individual coccoliths range from simple disks, with various sculpture including

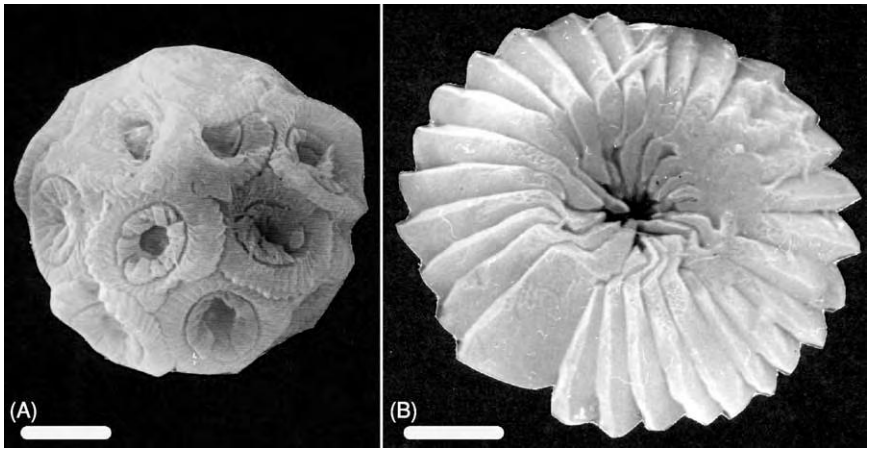


Figure 5 Coccolithophoralean coccosphere (A) and individual coccolith (B). Calcitic coccoliths are a major source of pelagic calcium carbonate sediment in modern oceans. Eocene, southern Spain. Scale bars 2 microns in (A) and 1 micron in (B). (Courtesy of B. El Mamoune.)

relatively long projections, to two plates joined by a central column. Many coccolithophores have different types of coccoliths on different parts of the cell surface. Others possess different coccoliths in different life stages.

Present-day coccolithophores are important phytoplankton, both in open oceans and in low-salinity seas such as the Black Sea. They extend from tropical to subpolar regions, although species assemblages differ according to latitude. Assemblages also show seasonal changes in relative abundance and species composition. Diversity is highest in warm oligotrophic waters and tends to decrease in coastal and restricted seas. Coccolithophores mostly live in the photic zone. This is deeper in low latitudes, particularly in the subtropics. Consequently, coccolithophore depth assemblages are well-developed in subtropical regions where shallow assemblages dominated by species bearing holococcoliths can be distinguished from deeper assemblages down to 220 m. These assemblages have differing temperature and nutrient, as well as light requirements. The high surface-to-volume ratio of individual coccospheres and coccoliths results in slow settling through the water column that allows dissolution of the calcite crystals. Faster settling is accomplished in zooplankton faecal pellets and other aggregates. Selective dissolution modifies the composition of sedimented coccolith assemblages and their diversity tends to reduce with increasing water depth.

Coccolithophores are first known from the Late Triassic. They gradually increased to a diversification maximum in the Late Cretaceous. Following major extinction at the Cretaceous/Tertiary boundary, the group regained high diversity during the Eocene but this has since decreased (Figure 1) (*see* **Palaeontology**). Several other types of nannofossils of uncertain affinities co-occur with coccoliths in pelagic oozes. Spherical tests made up of stacked calcite crystals, e.g., *Prinsiosphera*, *Schizosphaerella* and *Thoracosphaera*, are among the first nannofossils recorded in the Late Triassic. Nannoconids occur as tubes of calcite plates in the Late Jurassic to Cretaceous. Discoasterids have star or rosette shape and range from Paleocene to Late Pliocene.

Coccolithophores and related nannofossils have important applications in biostratigraphy and palaeo-oceanography. Nannofossil-based biostratigraphic zonations exist for the Jurassic to the present day. Most of these zonations consist of interval zones with boundaries characterized by first appearance or extinction datums of nannofossil species that are considered to be synchronous. Relative to other biostratigraphic markers in pelagic deposits, some nannofossils possess a cosmopolitan distribution and are regarded as relatively independent of water mass characters (i.e., temperature, nutrient content, salinity). Their first appearances and extinctions are therefore used for long-distance correlation. Nannofossil zones have good time resolution, in particular from the Late Cretaceous onwards, ranging from a few hundred thousand to a few million years. Correlation of calcareous nannoplankton zones to standard chronostratigraphic units is well established, facilitating widespread use of these fossils in biostratigraphy. Coccolith assemblages preserved in seafloor sediments in present-day oceans reflect the communities living in overlying near-surface waters. Since calcareous nannofossil assemblages vary with temperature over latitudinal gradients modified by

currents and other oceanographic factors, they can be used as proxies for palaeo-oceanographic conditions in ancient pelagic sediments.

Chlorophyta (Green Algae)

Halimedales

This order (formerly included in the Codiaceae or Udoteaceae) comprises marine algae of siphonous organisation, i.e., the plant is a single multinucleate cell forming a branching tube or siphon (the parts of the tube resulting from branching are also called siphons). In *Halimeda* the thallus is a bundle of interwoven, subparallel siphons that branch into swollen vesicles (utricles) to form the outer surface. The thallus is composed of segments encrusted with aragonite needles that form in the spaces between siphons and vesicles (Figure 6). Thalli, typically a few decimetres in size, can reach 1 m or more in length. They are articulated by having uncalcified connections between calcified segments. *Halimeda* attaches to hard substrates or uses a large bulbous holdfast (aggregate of rhizoidal siphons) to anchor itself in mobile sand-gravel substrates. Plants can be erect or, especially when larger, sprawling with several points of attachment. Male and female gametes are produced in separate plants in branched stalks at the upper margins of segments. In sexual reproduction the cell contents of the entire plant are transformed into gametes (mass spawning), leaving the thallus empty and dead.

Halimeda occurs in tropical to warm temperate seas. In temperate waters such as the Mediterranean Sea the thallus is weakly calcified. In tropical carbonate environments *Halimeda* is common in coral reefs and lagoons from depths of less than 1 m to 150 m. Its fast growth produces copious quantities of gravel-size segments that are shed both during life and after death. *Halimeda* sediment production is particularly important in relatively deep water, at depths of 20–50 m. *Halimeda* mounds at this depth cover areas of hundreds of square kilometres on the outer Queensland shelf of the Great Barrier Reef. In the Florida Keys, *Halimeda* is the single most important component of carbonate sand and gravel, rivalling coral in abundance.

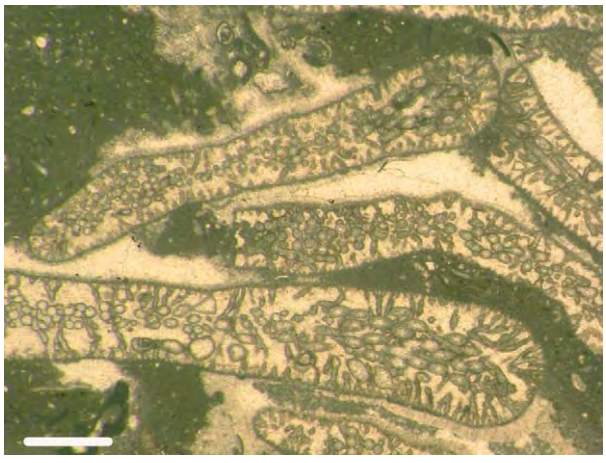


Figure 6 *Halimeda* segments. Aragonitic crystals (now calcite) form a cast of the complex, interwoven cell tubes of the alga, which remain as voids later filled with cement. Upper Miocene, south eastern Spain. Scale bar 1 mm.

Halimeda-like fossils are present in the Palaeozoic, e.g., the Ordovician *Dimorphosiphon* and the Devonian *Litanaia*. The group diversified in the Mesozoic with *Boueina* (Late Triassic–Late Cretaceous) and *Arabicodium* (Mid-Jurassic–Early Tertiary). *Halimeda* ranges from Early Cretaceous to present. Other extant halimedaleans, such as *Udotea*, *Penicillus*, *Rhipocephalus*, and *Tydemania*, are also important carbonate producers in tropical shallow water environments. *Udotea* has a fan-shaped thallus with a simple stalk. *Penicillus* and *Rhipocephalus* have brush-like thalli anchored by rhizoidal siphons. *Tydemania* is globular. Greater post-mortem skeletal disaggregation than in *Halimeda* limits the preservation potential of these genera as recognizable fossils.

Dasycladales

As in halimedaleans, the thallus is siphonous, consisting of a single multinucleate cell with a large central vacuole surrounded by cytoplasm. Plants range from a few millimetres to 20 cm in height, and are anchored by a rhizoidal holdfast. They have a single erect axis (dichotomously divided in a few species) with lateral branches. Gametes (zooids, no sexual differentiation occurs) are produced in fertile laterals. Dasycladalean taxonomy is complex. Six major families can be recognized, four of which (Seletonellaceae, Bereselleaceae, Diploporaceae, Triploporellaceae) are extinct. Dasycladaceae and Polyphysaceae (formerly Acetabulariaceae) are extant.

Present-day Dasycladaceae (six genera) occupy shallow marine environments often less than 5 m in depth, in sheltered areas in tropical and subtropical waters. Spaces between the main axis and lateral branches are filled to varying degrees by aragonite. The resulting calcareous skeleton forms a cast of the plant that preserves the shapes of the main axis and laterals as voids (Figure 7). Degree and site of calcification vary with the species. This strongly influences its potential to preserve details of thallus anatomy. Calcification generally decreases with illumination and temperature. A few poorly calcified species grow in warm-temperate seas. *Batophora*, which does not calcify, can also be found in brackish and fresh water.

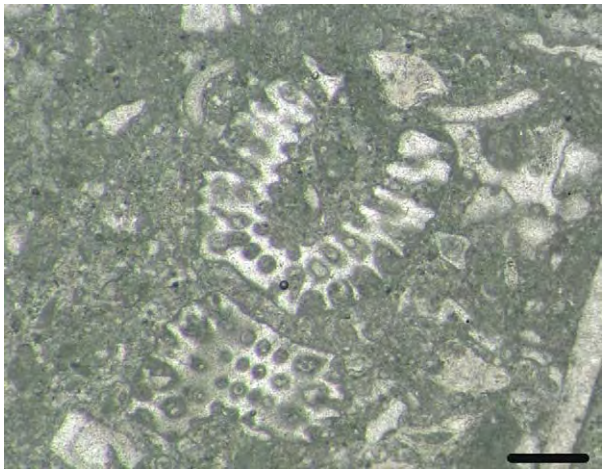


Figure 7 Dasycladalean sections. The aragonitic skeletons (now calcite) form a cast of the cell axis and lateral branches, which remain as matrix filled voids. The oblique section (upper) partly shows the size and shape of axis and laterals whilst the tangential section (lower) indicates the arrangement of the laterals. Paleocene, western Pyrenees, Spain. Scale bar 500 microns.

Present-day Polyphysaceae (five genera) also occupy shallow tropical to warm temperate seas. Some species tolerate high salinity fluctuations in shallow-water lagoons. Members of this family are characterised by the presence of a cap of gamete-producing laterals at the top of the plant as in *Acetabularia*, or alternating with sterile laterals in *Halycorne*. Calcification mostly takes place in the cap.

Possible dasycladaleans (e.g., *Yakutina*) are known from the Cambrian, but the earliest firm records are Ordovician. Dasycladaleans are common in platform carbonate deposits since the Carboniferous (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves). They reached high diversity in the Late Jurassic to Early Cretaceous and Eocene. The species richness of the group decreased after the Eocene, and extant species can almost be considered 'living fossils' (Figure 1). Dasycladaleans are used in biostratigraphy of shallow-water carbonates from the Late Palaeozoic to the Palaeogene. Zonations based upon dasycladalean genera and species have relatively poor time-resolution and their relation to standard chronostratigraphic scales is imprecise. However, together with benthic foraminifers, dasycladaleans are often the only fossils of biostratigraphic value found in extensive shallow platform carbonates.

Cyclocriniteae

Cyclocrinitids, or Cyclocriniteae, are a group of probable dasycladaleans (see Dasycladales above) mainly of Mid-Ordovician to Early Silurian age. They are characterized by an apically-inflated main axis giving rise to a cluster of radiating lateral branches. The tips of the laterals expand into hexagonal facets that unite to form an unbroken outer globular surface. The resulting swollen thallus is usually several centimetres in diameter. Common genera, including *Apidium*, *Coelosphaeridium*, and *Cyclocrinites*, are locally abundant. The affinities of Cyclocriniteae have been complicated because they have also been compared with, and sometimes placed in, receptaculitids – a problematic Palaeozoic group that shows some similarities with sponges. However, although cyclocrinitids were originally regarded as animals, their morphology is similar to that of the modern weakly calcified dasycladalean *Bornetella*.

Charophyceae

These chlorophytes are the closest algal relatives of embryophytes (mosses and vascular plants). Living charophyceans are grouped in the order Charales. Two additional extinct orders occur in the Palaeozoic. The group is characterized by complex bushy thalli, up to many decimetres in height, consisting of alternating nodes and internodes. The internodes are mostly made up of elongate cells. The nodes exhibit whorls of short branches ('leaves', and spine cells in some species) and branches of unlimited growth (Figure 8). Most charophyceans are anchored in soft sediment by rhizoids and have the appearance of subaqueous shrubs. The stem and branches can be encrusted by low-magnesium calcite and, more rarely, aragonite. Reproductive structures occur at nodes. Male gametes (spermatozoids) are produced in spherical structures called antheridia. Female gametes form in an ovoid oogonium that consists of an egg cell surrounded by sterile cells (Figure 8). In some species, following fertilization of the egg the surrounding cells become encrusted by calcite, while the remainder of the oogonium decays. The resulting calcified structures, gyrogonites, are the most common fossil remains of charophyceans (Figure 9).

Gyrogonites are made up of vertically elongate cells in primitive members of the Sycidiales, ranging from Late Silurian to Early Carboniferous; of dextral spiral cells in Trochiliscales, ranging latest Silurian-Permian, and of sinistral spiral cells in the extant order Charales, first recorded in the Devonian. The number of cells and the type of opening in gyrogonites are family-level taxonomic characters. Members of the Mesozoic family, Clavatoraceae possess gyrogonites with an external calcified cover (utricle) of various shapes. The Charales achieved greatest diversity in the Cretaceous to Palaeogene. It is a small, almost relict, group at the present day (Figure 1).

Charaleans grow in freshwater streams, ponds, and lakes and are especially abundant in calcareous

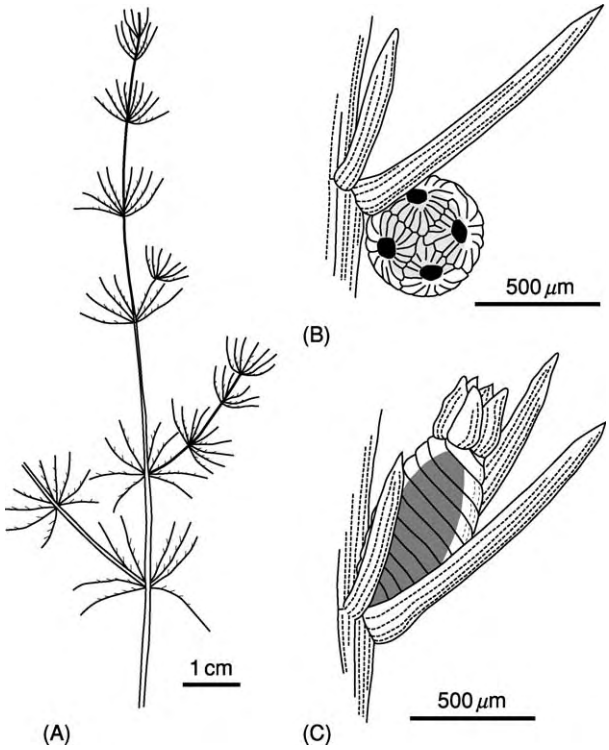


Figure 8 Charalean plant (A), showing the male reproductive structure (antheridium) at a node (B), and a female reproduct ive structure (oogonium) (C). The spiral cells in the oogonium are encrusted by calcite after fertilization. The resulting calcified structures, gyrogonites, are the most common fossil remains of Charales. (Adapted from Engler and Prantl (1898 1900).)

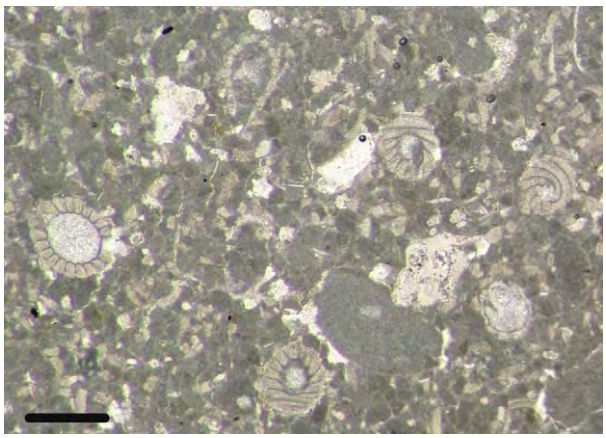


Figure 9 Sections of charalean gyrogonites. Lower Cret aceous, Transylvania, Romania. Scale bar 500 microns.

oligotrophic waters. Some species live in brackish lagoons. Fossil gyrogonites are used in biostratigraphy of non-marine deposits, especially from the Early Cretaceous to the Palaeogene. The biostratigraphic resolution of charalean zones is relatively low and correlation with marine chronostratigraphy is imprecise. Nonetheless, charalean fossils permit stage, or even sub-stage, level dating of deposits that often lack other biostratigraphic fossils.

Groups of Uncertain Affinity

Gymnocodiaceae

In vegetative construction, gymnocodiaceans are similar to halimedaleans (see Halimedales above), but rounded cavities within the cortex of some specimens resemble reproductive structures and suggest a red algal affinity. Gymnocodiaceans have been compared with the modern weakly calcified chaetangiacean *Galaxaura*. This enigmatic group, including *Gymnocodium, Permocalculus, Nanjinophycus*, and possibly *Succodium*, is most widespread and diverse in the Late Permian. *Permocalculus* also occurs in the Cretaceous and has been reported from the Paleocene and Miocene. Gymnocodiaceans appear to be shallow tropical marine in distribution, and locally are important bioclastic sediment producers.

Phylloid Algae

The term 'phylloid algae' describes fossils with flattened leaf-like shape that are especially common in Upper Carboniferous and Lower Permian reefs. Their recognition as a group is practical because the internal structure of these fossils is commonly poorly preserved, but it overlooks important morphological and systematic variations, and phylloids are a highly heterogeneous group that probably includes both green and red algae. Their internal structure is gradually being elucidated by discovery of well-preserved specimens and by use of cathodo-luminescence. Two morphological subgroups can be distinguished: (i) erect blades with medullary and cortical structure resembling that of halimedalean green algae, e.g., *Anchicodium, Eugonophyllum*, *Ivanovia* and, doubtfully, *Calcifolium* (see Halimedales); (ii) prostrate crusts with internal cells and possible conceptacles, e.g., *Archaeolithophyllum* that has differentiated multicellular filaments together with surficial conceptacle-like structures (**Figure 10**). This latter subgroup is compared with red algae, particularly peyssonneliaceans (see Peyssonneliaceae above). Phylloid algae formed skeletal frame reefs in the Late Carboniferous to Early Permian.

Calcified Cyanobacteria

Originally named blue-green algae, the prokaryotic cell organization of this group clearly allies it with bacteria. In natural environments, cyanobacteria extrude polymeric substances to form protective

Figure 10 Sections *Archaeolithophyllum*, a phylloid alga. Despite recrystallization, the multicellular structure of the crusts can be discerned. Upper Carboniferous, Cantabrian Mountains, northern Spain. Scale bar 500 microns.

sheaths. These can create diffusion limited sites in which photosynthetic uptake of HCO_3 generates pH gradients that can result in $CaCO_3$ precipitation. However, calcification only occurs if environmental conditions favour precipitation. Since cyanobacterial calcification is localised in the external mucilaginous sheath, the resulting microfossils are simple in form and include bush-like forms (e.g., *Angusticellularia*), tube-like filaments (e.g., *Girvanella*, *Hedstroemia*, *Cayeuxia*) (Figure 11), and chambered clusters (e.g., *Renalcis*).

Whereas some fossils are readily recognized as calcified cyanobacteria, others are more problematic. These include sedimentologically important Cambrian and Late Devonian fossils such as *Epiphyton* and *Renalcis* that have some resemblances to cyanobacteria but are not identical to modern examples, and are generally referred to as calcimicrobes. *Renalcis* and *Epiphyton*, together with the cyanobacterium *Angulocellularia*, were major reef builders in the Cambrian.

Cyanobacteria have a geological record from the Archaean, but are not conspicuous as calcified fossils in marine environments until the Phanerozoic and even then their secular distribution is episodic. This could reflect variation in the saturation state of seawater over geological time. Calcified cyanobacteria assumed major importance in the Cambrian and Early Ordovician. They reappeared in abundance in the Late Devonian, and were common until the mid-Cretaceous, but are scarce or absent in marine environments during the Cenozoic (Figure 1). During episodes of abundance, calcified cyanobacteria were major reef components and in the Palaeozoic they often rivalled calcareous sponges such as archaeocyaths and stromatoporoids in importance.

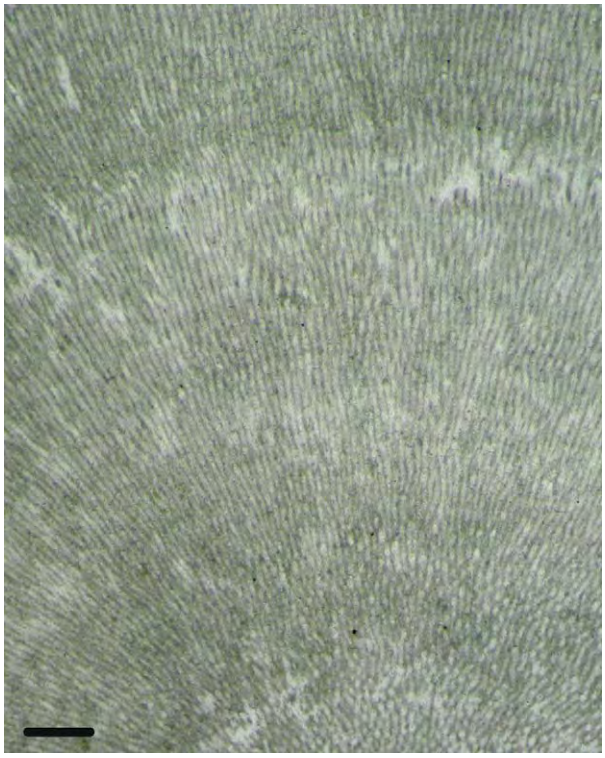


Figure 11 Section of *Cayeuxia*, a calcified cyanobacterium. The branching tubes are the result of calcification of the external mucilaginous sheath surrounding cell filaments. Middle Triassic, Betic Cordillera, southern Spain. Scale bar 500 microns.

In present-day lakes and rivers, calcified cyanobacteria can form thick tufa bioherms and dams, together with oncoids (spherical stromatolites).

Small unicellular picoplanktic cyanobacteria, such as *Synechococcus*, form seasonal blooms in lakes along with diatoms and other planktic algae. Their photosynthesis can stimulate water column precipitation of small calcite crystals (whitings). Marine whitings in tropical seas may have a similar origin and potentially account for abundant lime mud production on ancient carbonate platforms.

See Also

Biosediments and Biofilms. **Palaeontology**. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Rocks:** Limestones.

Further Reading

Adey WH (1986) Coralline algae as indicators of sea level. In: Van de Plassche O (ed.) *Sea level research: a manual for the collection and evaluation of data,* pp. 229 279. Amsterdam, The Netherlands: Free University of Amsterdam.

Bassoullet JP, Bernier P, Deloffre R, Génot P, Poncet J, and Roux A (1983) Les algues udoteacées du Paléozoique au Cénozoïque. *Bulletin des Centres de Recherche, Exploration Productionn, Elf Aquitaine* 7: 449 621.

Berger S and Kaever MJ (1992) *Dasycladales. An Illustrated Monograph of a Fascinating Algal Order.* Stuttgart, Germany: Georg Thieme.

Braga JC, Bosence DWJ, and Steneck RS (1993) New anatomical characters in fossil coralline algae and their taxonomic implications. *Palaeontology* 36: 535 547.

Flügel E (ed.) (1977) *Fossil Algae. Recent Results and Developments*. Berlin, Germany: Springer.

Graham LE and Wilcox LW (2000) *Algae*. Upper Saddle River, NJ: Prentice Hall.

Jordan RW (2002) Environmental applications of calcareous nannofossils. In: Haslett SK (ed.) *Quaternary Environmental Micropalaeontology,* pp. 185 206. London, UK: Arnold.

Martín Closas C and Schudack M (1991) Phylogenetic analysis and systematization of post Paleozoic charophytes. *Revue de la Societé Botanique de France* 138: 53 71.

Riding R (ed.) (1991) *Calcareous Algae and Stromatolites*. Berlin, Germany: Springer.

Tappan H (1980) *The Paleobiology of Plant Protists*. San Francisco: Freeman. W.H.

Toomey DF and Nitecki MH (eds.) (1985) *Paleoalgology. Contemporary Research and Applications*. Berlin, Germany: Springer.

Van den Hoek C, Mann DG, and Jahns HM (1997) *Algae. An Introduction to Phycology.* Cambridge: Cambridge University Press.

Winter A and Siesser WG (eds.) (1994) *Coccolithophores*. Cambridge University Press: Cambridge.

Woelkerling WJ (1988) *The Coralline Red Algae: An Analysis of the Genera and Subfamilies of Nongeniculate Corallinaceae*. Oxford: Oxford University Press.

Fungi and Lichens

T N Taylor, University of Kansas, Lawrence, KS, USA
M Krings, Bayerische Staatssammlung für Paläontologie und Geologie, Geo-Bio Center, Munich, Germany

Introduction

Our understanding of the ecology of ancient ecosystems has progressed rapidly in recent years, so it is now possible to examine highly complex interactions involving previously understudied organisms, such as fungi and lichens. Based on the biological and ecological significance of fungi and lichens in many ecosystems today, a detailed knowledge of their fossil record, evolution, and the roles they played in biological and ecological processes in the past is important in understanding the evolutionary history of ancient and modern ecosystems. However, the fossil record of fungi and lichens remains incomplete. This article surveys the information gathered on fossil fungi and lichens. In some instances, preservation of the specimens is so extraordinary that it is possible to make direct comparisons with modern organisms and to examine interactions between fungi and other saprophytic, parasitic, and mutualistic organisms. Examples of enigmatic fossils, in which the fungus or lichen identity is still being debated, are also presented.

Fungi

Fungi are a unique group of eukaryotic organisms that are typically constructed of delicate filaments (hyphae), which may be loosely arranged or tightly packed to form complex structures. Fungi are heterotrophic, which means that they lack the ability to manufacture their own food and must rely on external sources to obtain nourishment. Many fungi cause serious diseases in plants, animals, and humans, whereas others are useful in their ability to produce antibiotics and various organic substances and to metabolize sugars into alcohol. Moreover, fungi can serve as potential agents for clearing toxic substances from the environment. Today, fungi and bacteria are the primary decomposers in the biosphere: they break down organic materials, such as the lignin in wood, and make the carbon, nitrogen, and phosphorous available to other organisms. Despite the fact that an estimated 1.5 million different types of fungi exist on Earth today, less than 100 000 species have been described and named to date.

Information about the geological history of fungi has been slow to accumulate because it has generally been assumed that the delicate filamentous bodies of most fungi (the mycelium) would not be adequately preserved in the fossil record. In addition, there is a gulf between palaeobotanists, who collect fossil fungi, and mycologists, who possess the experience to describe their structure and organization adequately. Finally, there is an inherent bias against collecting fossil fungi because the symptoms that they typically cause (e.g. rot, decay, disease) make the host specimens unattractive and often, therefore, not collected. Despite these obstacles, there is an ever-increasing awareness that fungi are adequately preserved in certain rocks and thus are a critical component in

understanding the evolution of other organisms and the ecosystems in which they lived.

The extant fungi are classed into four major groups: chytridiomycetes, which are microscopic forms that occur in both terrestrial and aquatic habitats and produce motile cells; zygomycetes, which form thick-walled spores as a result of sexual reproduction and are constructed of coenocytic hyphae (included in this group are the mycorrhizae, which have formed intricate symbiotic associations in the roots of most terrestrial plants); ascomycetes, which have specialized cells (asci) that produce internally formed spores (ascospores); and basidiomycetes, which produce spores externally on a specialized cell (basidium). Although other structural, biochemical, and physiological characters are used to subdivide these major groups further, the recognition of fossil fungi is based on certain morphological features and, where sufficiently preserved, the symptoms they cause.

Fossil Fungi

Although evidence of life extends well back into the Proterozoic, the first unequivocal fungal remains come from the Palaeozoic. These consist of branched tubular hyphae with specialized cross-walls and multicelled spores recovered from Lower and Middle Silurian rocks. Based on similarities to structures seen in modern fungi, the Silurian remains have been suggested to have affinities with the ascomycetes. These early reports are based on fossils that were discovered after macerating rock fragments, and thus nothing is known about the role that the fungi played in the ecosystem in which they lived. Fungal remains in the form of spores have also been reported from the Ordovician; however, these spores look so modern that there is some question as to whether they may be contaminants of living fungi.

The Early Devonian Rhynie Chert ecosystem in Scotland has provided the most complete view of Palaeozoic fungi to date. Because this freshwater hot-springs ecosystem is nearly completely preserved *in situ* in a siliceous matrix, it contains a large number of different types of fungi, which were present at the time of fossilization; many of these fungi demonstrate a wide variety of interactions with other organisms. The most diverse group of fungi in the Rhynie Chert are the chytridiomycetes, which were probably the principal decomposers of organic matter in the Early Devonian. Like their modern equivalents, the chytridiomycetes from the Rhynie Chert include forms that are found on the surfaces of other organisms as well as inside cells. Because of their small size and great abundance, it has been possible to detail not only the morphology but also the life history of several forms. Other Rhynie Chert fungi (members of the Zygomycetes) formed symbiotic associations with a number of land plants as vesicular–arbuscular mycorrhizae (**Figures 1E** and **1F**). In these associations, the fungus (*Glomites* sp.) has the ability to penetrate certain cells of the plant to form highly branched absorbing structures termed arbuscules (**Figure 1F**), which act as physiological exchange sites where the fungus obtains carbon. The host plant benefits from the greater ability of the fungus to obtain nutrients and water. Many believe that this fungus–land-plant relationship allowed plants to become established in the terrestrial realm, since it is estimated that today more than 90% of all land plants share a beneficial relationship with certain fungi.

A large number of extant fungi are parasites and pathogens of other organisms, and some of these associations are represented in the Rhynie Chert. Plant-parasitizing fungi in the Rhynie Chert ecosystem typically invaded the softer tissues of the plants, and it is possible to observe areas where, as a result of fungal infection, cells are broken down, sometimes resulting in easily observable necroses. Other parasitic fungi (predominantly chytridiomycetes) in the Rhynie Chert ecosystem colonized the spores of land plants (**Figure 2B**); as a result, the spores lost their capability to germinate and produce a gametophyte. Still other fungi were mycoparasites, i.e. they obtained nourishment from other fungi and/or fungal spores (**Figures 1A** and **1B**). In some instances, when a plant-parasitizing fungus attempted to enter a cell, the plant responded by increasing the size of the cell, increasing the number of cells in the immediate region, and/or producing chemical deterrents in order to encapsulate or ward off the infection. Responses of this type are common in plants today and indicate not only that the Rhynie Chert plants were alive when the fungi attacked but also that the signalling mechanisms between these two organisms were well-established more than 400 Ma ago.

To date, three of the four major groups of fungi have been found to be represented in the Rhynie Chert; only the basidiomycetes are not present or have not yet been discovered. Especially noteworthy is the occurrence of an advanced saprophytic or parasitic perithecial ascomycete. Beneath the epidermis of upright stems and rhizomes of the land plant *Asteroxylon* (an early representative of the lycophytes) occur spherical to flask-shaped opaque structures (**Figure 1C**), which represent the reproductive structures (perithecia) of this fungus. The perithecia are characterized by short ostiolate necks that typically protrude from the epidermis of the host plant through stomatal openings. Lining the interior of the perithecium are elongate thin-walled hairs interspersed with

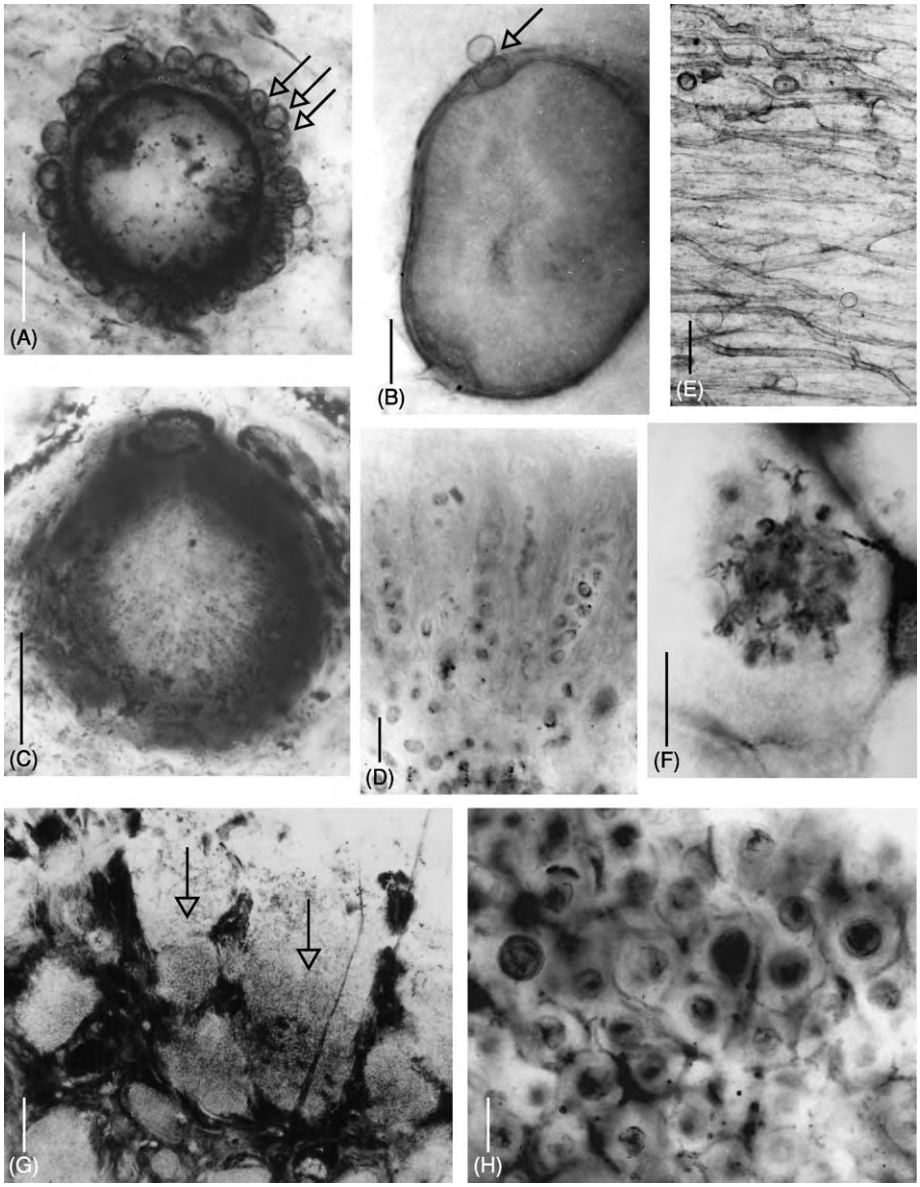


Figure 1 (A) Fungal spore from the Early Devonian Rhynie Chert, with its outer surface densely covered by parasitic chytridiomy cetes (arrows); scale bar: 50 μm (From Taylor TN, Remy W, and Hass H (1992) Fungi from the Lower Devonian Rhynie Chert: Chytridiomycetes. *American Journal of Botany* 79: 1233 1241.) (B) Parasitic chytridiomycete (arrow) that has developed between the wall layers of a fungal spore from the Rhynie Chert; scale bar: 10 μm. (C) Longitudinal section through a perithecium of a Rhynie Chert ascomycete; scale bar: 100 μm. (D) Detail of **Figure 1C**, showing part of the perithecium with several asci containing ascospores; scale bar: 15 μm. (E, F) Vesicular arbuscular mycorrhiza in the Rhynie Chert land plant *Aglaophyton major*; (E and F adapted from Taylor TN, Remy W, Hass H, and Kerp H (1995) Fossil arbuscular mycorrhizae from the Early Devonian. *American Journal of Botany* 87: 560 573.) (E) hyphae with vesicles of the mycorrhiza fungus *Glomites* sp.; scale bar: 100 μm; and (F) *Glomites* arbuscule within a cortical cell from an aerial axis of *A. major*; scale bar: 10 μm. (G) Section of the thallus of the Rhynie Chert lichen *Winfrenatia*, showing depressions that contain cyanobacteria (arrows) surrounded by more a opaque zone that represents the mycobiont; scale bar: 50 μm. (H) Detail of **Figure 1G**, showing part of a single depression containing unicellular cyanobacteria and fungal hyphae that form a net like structure; scale bar: 25 μm.

sacs (asci) that contain unicellular to up to five times septate sexual spores (ascospores) (**Figure 1D**).

Plants with abundant woody tissues first appear in the Middle to Late Devonian, and it is at this time that the first wood-decomposing fungi – basidiomycetes – appear. The earliest fossil evidence for the possible existence of basidiomycetes occurs in the form of branching hyphae of varying diameters

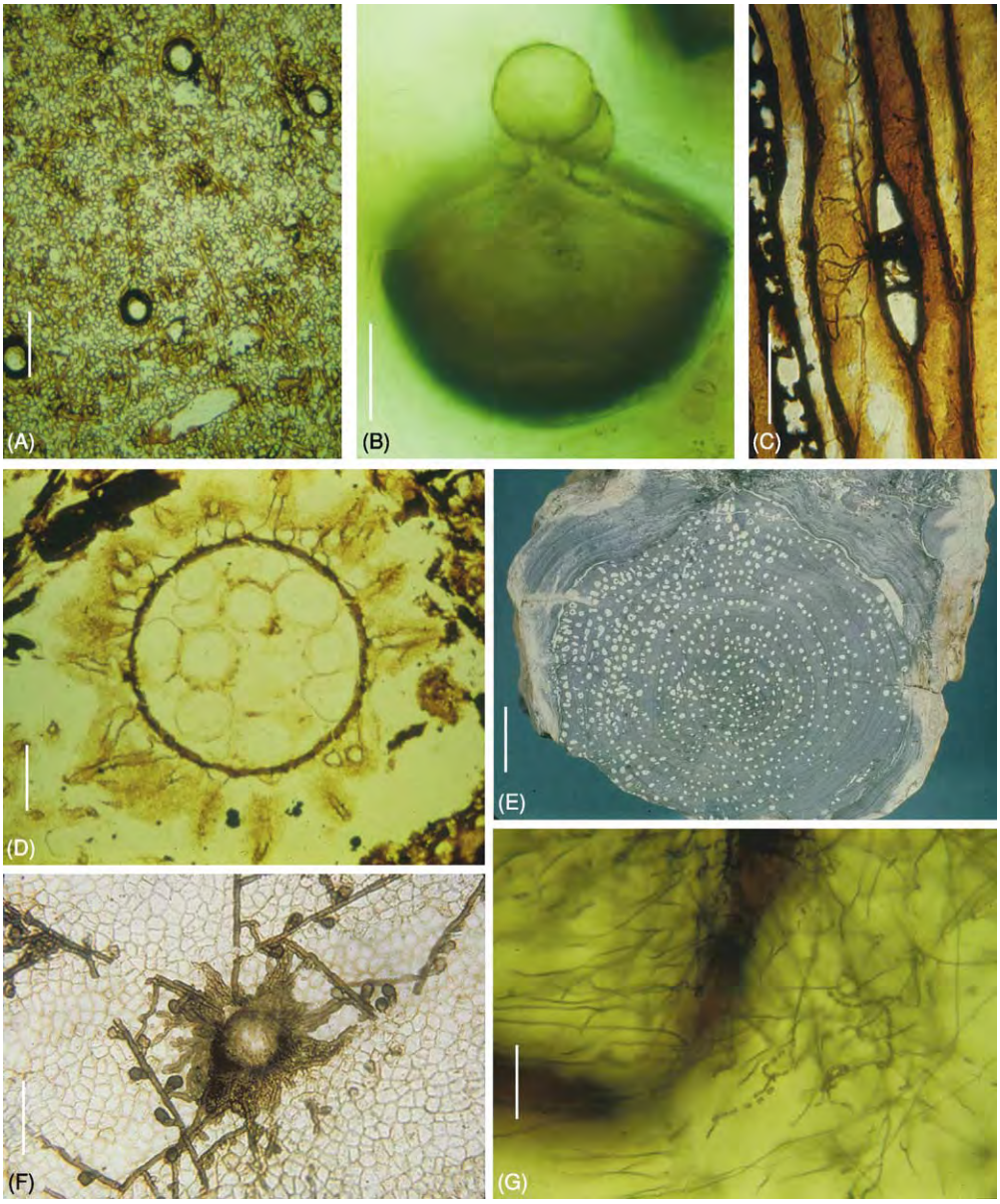


Figure 2 (A) *Prototaxites*; transverse section showing the internal anatomy, which is entirely constructed of interwoven septate tubules (?hyphae); scale bar: 100 μm. (B) Parasitic Chytridiomycetes on a spore of the Rhynie Chert plant *Aglaophyton major*; scale bar: 10 μm. (From Taylor TN, Remy W, and Hass H (1992) Fungi from the Lower Devonian Rhynie Chert: Chytridiomycetes. *American Journal of Botany* 79: 1233 1241.) (C) Radial section of *Callixylon newberryi* (progymnospermous wood) from the Upper Devonian of Indiana, showing branching fungal hyphae in one of the tracheids; scale bar: 35 μm. (D) *Traquairia williamsonii*, an ornamented sporocarp from the Upper Carboniferous of Kentucky, containing several thin walled spores; scale bar: 80 μm. (Adapted from Stubblefield S and Taylor TN (1983) Studies of Paleozoic fungi.1. The structure and organization of *Traquairia* (Astomycota). *American Journal of Botany:* 70: 387 399.) (E) Cross section through an *Araucarioxylon* (gymnospermous wood) stem from the Triassic of Antarctica, showing numerous pockets of decay (lighter areas); scale bar: 1.2 cm. (From Stubblefield S and Taylor TN (1986) Wood decay in silicified gymnosperms from Antarctica. *Botanical Gazette* 147: 116 125.) (F) Leaf colonizing fungus on an angiosperm leaf from the Eocene (Tertiary) of the USA; scale bar: 60 μm. (G) *Geotrichites glaesarius*, a saprophytic fungus from the surface of a spider that is preserved in Tertiary amber from the Dominican Republic; scale bar: 60 μm.

(e.g. Figure 2C), some possessing rounded knobs (so-called clamp-connections), and spores; indirect evidence of the presence of these fungi includes solution troughs along the inner surfaces of water-conducting cell walls. Because modern basidiomycetes are the primary degrading agents of lignin and cellulose, dissolution patterns in fossil wood that resemble those seen in modern wood indicate the existence

of wood-degrading fungi, even though there is no direct evidence of the fungus. In Mesozoic fossils from Antarctica and other southern-hemisphere continents, there is another type of pathology caused by fungi in plants. When certain fossil logs are examined in cross-section, the wood contains numerous circular areas that are devoid of cells (Figure 2E). These areas, usually 1–2 mm in diameter, represent spindle-shaped cavities that were formed by the enzymatic activities of a fungus. When examined in detail, the cells at the margins of these areas show evidence of selective dissolution of the cell-wall components. Patterns of this type, termed white rot and white pocket rot, are common in woody plants today, where they are caused by several basidiomycetous fungi. What is interesting is that the trees attacked during the Mesozoic are all now extinct, but the fungus has survived and produces exactly the same symptoms today in different hosts. Another example of fungal attack in woody plants comes from the Triassic Petrified Forest in Arizona. Sections of wood show regions of highly disrupted cells. Although evidence of the fungus responsible for this degradation is generally lacking, the large number of trees with the same symptom has led to the suggestion that perhaps much of this forest was attacked by a fungus, which caused a level of devastation comparable to that seen in modern trees infected with Dutch elm disease. Pathogenic fungi are also known from several types of seed, in which the fungus has greatly altered the internal tissues. For example, the disruption of nutritive cells in the seed and the formation of fungal reproductive structures in their place means that the seed was no longer capable of producing a new plant.

Fungi can also be found on the leaf surfaces of many different fossil plants (e.g. Figure 2F). Excellent examples are common on Eocene flowering plants, where the fungus produces specialized structures that penetrate the leaf surface to obtain nutrients. Based on the types of plants and the fungi on the leaves, these fossil fungus–plant interactions provide information about fungal host preferences and can perhaps also be used as a proxy record for some climate variables.

Fungi, like all organisms, must be capable of producing large numbers of spores and other propagules that can be disseminated and serve as reinfecting agents. One interesting fossil example comes from the Eocene Princeton Chert of British Columbia, in which a fungus invaded the normal pollen-producing structures of a flowering plant. This fungus, which is most like a modern smut fungus, replaced the pollen with its own spores and thus was able to colonize new host plants effectively because the spores were distributed by pollinating insects that visited the flowers in search of nectar.

Today several major groups of fungi have successfully exploited animals, including humans, as hosts. Modern fungi are found in the hind guts of various arthropods, and at least one record suggests that this association extends back to the Triassic. In this association, the fungus is able to exploit new niches through the mobility of the host animal; it is not fully understood, however, whether there was a benefit for the host. Some animals use fungi as a food source (fungivory). Aggregations of fungal hyphae in the form of coprolites date from the Middle Silurian onwards and indicate that fungi were a principal component of the diet of some early animals. Coprolites are a potential new source of information about fungus–animal associations in the fossil record. Another excellent preservation matrix for fossil fungi is plant resin (amber), which dates back to the Carboniferous. Fungus–animal associations are especially well documented from Tertiary amber and include fungi on a partially decomposed spider from the Upper Oligocene–Lower Miocene (Figure 2G) and pathogenic fungi on a termite, an ant, and within the body cavity of a nematode.

Fungi with Uncertain Affinities

Throughout the Carboniferous a variety of conceptacle-like structures are found that have been interpreted as sporocarps. They occur singly or in clusters and are common in the permineralized peat that forms Carboniferous coal balls. The walls of these circular (up to 1 mm in diameter) structures are composed of interlaced hyphae, and the outer surface may be smooth or ornamented by processes (Figure 2D). Some contain thin-walled spores, which in turn contain additional spores. These sporocarps have been assigned to the Ascomycetes by some, whereas an alternative hypothesis places them within the Zygomycetes. Another interesting fungus from the Carboniferous with problematic affinities combines features of the Ascomycetes with those of perhaps a basidiomycete. While this interesting fungus may be a form that shares features of two groups, another idea is that one fungus has parasitized the other. Perhaps the most unusual life form considered by many scholars to be a fungus is the Nematophyte *Prototaxites*, which existed until the Late Devonian. This enigmatic trunk-like organism grew up to 1 m in diameter and its internal anatomy was entirely constructed of interwoven septate tubules (?hyphae) of three different sizes (Figure 2A); the arrangement of tissues suggests a periodicity in growth. The reproductive parts of *Prototaxites* remain equivocal; however, there is some suggestion that its affinities lie with the Basidiomycetes. Whether *Prototaxites* was a saprophytic fungus, as has been hypothesized, or

some other type of life form with an unusual combination of features, perhaps not even remotely related to any modern organism, it must have presented an imposing structure, extending more than 8 m above the very small plants that made up the Early Devonian landscape.

Lichens

Lichens may be interpreted as 'dual organisms' since they do not represent a single organism, in the way that most other life forms on Earth do, but rather a combination of two or three entirely different organisms: a fungus (mycobiont) that lives in an obligate mutualistic association with a green alga or cyanobacterium or both (photobiont). The lichen symbiosis is unique since it represents a new 'hybrid' life form with a distinct body plan (lichen thallus) and structural and physiological properties that are different from those of either of the partners. An estimated 13 000–17 000 species of lichen exist today, extending from the tropics to the polar regions and growing on a highly diverse array of substrates, including soil, bark, bare rock surfaces, leaves of vascular plants, barnacle shells, and other lichens. The vast majority of lichen-forming (lichenized) fungi are members of the Ascomycetes, but some 20 species of basidiomycete and one zygomycete are also known to enter into lichen symbioses. In contrast to the many thousands of lichenized fungi, there are only about 100 different photobionts, most of which are unicellular green algae (Chlorophyta).

It is still not fully understood how this unique assemblage of organisms evolved. The failure to resolve more fully the evolutionary history of lichens is due primarily to a meagre fossil record. Moreover, in order to establish the existence of a fossil lichen, it is necessary to demonstrate not only the individual partners but also that there is some degree of interdependence between the separate organisms. Thus, even if an entire association is well preserved, it may be difficult to determine whether it is a lichen or simply an asymptomatic aggregate of algae or cyanobacteria and fungi that are closely associated but do not interact. Alternatively, the aggregation may represent a fungus that has conveniently parasitized an alga or cyanobacterium. Molecular and genetic studies of extant lichens suggest that the lichen symbiosis has arisen several times over the course of geological time.

Fossil Lichens

It has been hypothesized that one of the first steps in the colonization of land by eukaryotes may have been the formation of a lichen symbiosis. However, compelling fossil evidence supporting this hypothesis remains elusive. The earliest fossil (from the Early Proterozoic Witwatersrand Group, South Africa) that has been considered to be a lichen is a thallus-like structure comprising erect columns made up of branched and apparently septate filaments. The lichen nature of this fossil is, however, questionable since remains of the photobiont are not observed. Moreover, the filaments are much smaller than most fungal hyphae and are in the size range of certain filamentous bacteria. Some believe that the filaments are not fossils because similar objects can be produced artificially in the laboratory. Other Precambrian life forms, which have been suggested to be lichens, are the so-called 'Ediacara fossils' or 'Vendobionta' (*see* **Precambrian:** Vendian and Ediacaran). These unusual 600 Ma old organisms, which presumably lived on soft mud or sands at the bottom of shallow coastal seas, are usually considered to be primitive animals, but one hypothesis considers them to be lichens or lichen-like based on the fact that they were relatively resistant to being compressed after burial. This idea has not received much support from either the palaeontological community or the lichenological community.

Several groups of Palaeozoic fossils have variously been interpreted as lichens or lichen-like associations. One group is the Nematophytes, which occurred from the Late Ordovician to the Late Devonian. The most impressive Nematophyte, *Prototaxites* (**Figure 2A**), is usually considered to be a giant fungus (see above), but some workers have speculated that it may have been a lichen-like association of two or more different kinds of organism. A second member of the Nematophytes, *Nematothallus*, was a small foliose organism that consisted of intermixed large and small tubules, covered by a thick cuticle-like layer. *Nematothallus* possesses a number of striking features that are also found in certain extant lichens. The cuticle-like surface layer, for example, displays impressions of a distinct cell pattern, which is virtually indistinguishable from that seen on the thallus surface in certain extant cyanolichens. Unfortunately, as is the case with other early lichen-like fossils, the affinities of *Nematothallus* cannot be established with certainty because a photobiont partner has not yet been discovered. Another example of a Palaeozoic life form, the identity of which is still being debated, is *Spongiophyton* (Middle Devonian). This thalloid organism is typically preserved as a flattened cuticle-like sheath with circular pores on one surface. Some have suggested that *Spongiophyton* was an early bryophyte (perhaps a primitive liverwort), whereas others speculate that it may have been a lichen.

The most convincing fossil example of a Palaeozoic lichen comes from the Early Devonian Rhynie Chert.

Winfrenatia (Figures 1G and 1H) consists of a mycelial mat, constructed of interwoven aseptate hyphae. Along the upper surface of the mat are numerous shallow relatively uniform depressions (Figure 1G). Within many of these depressions are coccoid unicells that are morphologically similar to certain extant cyanobacteria; also present are clusters of the same cells within mucilaginous investments that are interpreted to be stages in the life history of the cyanobacterium. Hyphae of the fungus extend into the depressions and become intertwined with the cyanobacteria (Figure 1H). The range of features observed in *Winfrenatia*, including the size of the thallus and the number of depressions on the surface, have allowed speculations as to the life-history strategy, which included the production of new cells of the photobiont to maintain the symbiosis and support the fungus and, at the same time, an increase in the size of the fungal mycelial mat. It has been suggested that *Winfrenatia* is not a true lichen but rather an unstable association in which a fungus parasitized a cyanobacterial colony. While the physiological stability of a symbiotic association in the fossil record may never be fully deciphered, the definition of a lichen as a controlled parasitism is perhaps more accurate, as it underscores the varying degrees of lichenization that appear in modern ecosystems. In that context, *Winfrenatia* may quite accurately be regarded as an Early Devonian lichen. The systematic position of the fungal partner of *Winfrenatia* has not been conclusively identified. The presence of aseptate hyphae and certain thick-walled sculptured spores associated with the mycelial mats suggest that the affinities of the fungus may lie close to the Zygomycetes. This is especially noteworthy since there is only one extant lichen with a zygomycetous mycobiont and a cyanobacterial photobiont, *Geosiphon pyriforme*. In this unique endosymbiotic association, cyanobacteria of the genus *Nostoc* become encapsulated in pear-shaped bladders formed by the fungus.

A few fossil lichens have been reported from the Cenozoic, most of which are preserved in Tertiary amber. That these fossils represent lichens is relatively easy to demonstrate, since their morphology closely resembles that of extant lichens; some forms can even be assigned to modern families and genera. In addition to these lichens, there are a few reports of adpressed lichen thalli, including a fossil *Lobaria* from Miocene rocks at Redding Creek (California) and the foliicolous (leaf-colonizing) lichen *Pelicothallos*, which was discovered on an angiosperm leaf from the Eocene of Tennessee.

It remains puzzling why lichens are so rare in the fossil record, or have been so infrequently described. One reason may be that ancient lichens were quite different in appearance from those seen today, and, as a result, it is difficult to recognize these organisms in the fossil record. Moreover, some workers have pointed out that many lichen-dominated habitats, such as tundra and mountains, have a relatively low potential for yielding fossils. However, lichens today also thrive in many environments that could lead to preservation, including swamp forests. Still, even in the extensive Carboniferous coal swamp forests, which have been studied extensively for more than 150 years, not a single compelling lichen fossil has ever been recorded.

Concluding Remarks

As is the case in many areas of palaeontology, the discovery and recognition of 'unique' fossils requires an intersection between the vagaries of preservation and the necessary expertise to interpret the fossil. Historically, relatively little attention has been directed at fossil bacteria, algae, fungi, and other miscellaneous organisms in which taxonomic placement has been difficult. As palaeontologists strive to understand the complexities of the biological and physical world of the past, it is becoming increasingly clear that organisms such as fungi and lichens were not only present but played pivotal roles in shaping the ecosystems of the past.

Glossary

Clamp-connections Outgrowths of fungal hyphae that, at cell division, make a connection between the resulting two cells by fusion with the lower

Coal balls Preservation type in which mineral matter has filled the cell lumens and intercellular spaces but has not replaced the cell walls.

Coprolites Fossilized faecal material.

Hypha (plural hyphae) Tubular filament that is the structural component of the majority of fungi and may be coenocytic or septate.

Mycelium (plural mycelia) The total mass of hyphae that constitutes the vegetative portion or thallus of a fungus.

Mycoparasite A fungus that is parasitic on another fungus.

Mycorrhiza Mutualistic association of a fungus with (the roots of) a higher plant.

Necrosis Localized death of living cells due to infection or injury.

Parasite An organism that obtains nourishment from another living organism.

Pathogen Any organism capable of causing a disease.

Saprophyte An organism that obtains nutrients from dead organic material.

Sporocarp A closed body or conceptacle containing one or more masses of spores or sporangia.

Symbiosis A form of interaction where one organism lives in an intimate association with another.

Thallus The vegetative body of a plant or lichen that is not differentiated into organs such as leaves and stems.

See Also

Biosediments and Biofilms. **Carbon Cycle**. **Fossil Plants:** Calcareous Algae. **Palaeozoic:** Devonian. **Precambrian:** Vendian and Ediacaran. **Sedimentary Rocks:** Chert. **Tectonics:** Hydrothermal Activity. **Tertiary To Present:** Eocene.

Further Reading

Hallbauer DK and van Warmelo KT (1974) Fossilized plants in thucholite from Precambrian rocks of Witwatersrand, South Africa. *Precambrian Research* 1: 199 212.

Hass H, Taylor TN, and Remy W (1994) Fungi from the Lower Devonian Rhynie Chert: mycoparasitism. *American Journal of Botany* 81: 29 37.

Heckman DS, Geiser DM, Eidell BR, *et al.* (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science* 293: 1129 1133.

Hueber FM (2001) Rotted wood alga fungus: the history and life of *Prototaxites* Dawson 1859. *Review of Palaeobotany and Palynology* 116: 123 158.

Poinar GO (1992) *Life in Amber.* Stanford, CA: Stanford University Press.

Purvis W (2000) *Lichens.* Washington, DC and London: Smithsonian Institution Press and The Natural History Museum.

Stubblefield S and Taylor TN (1986) Wood decay in silicified gymnosperms from Antarctica. *Botanical Gazette* 147: 116 125.

Stubblefield S and Taylor TN (1988) Recent advances in palaeomycology. *New Phytologist* 108: 3 25.

Taylor TN (1993) Fungi. In: Benton MJ (ed.) *The Fossil Record,* vol. 2, pp. 9 13. London: Chapman & Hall.

Taylor TN and Taylor EL (1993) *The Biology and Evolution of Fossil Plants.* Englewood Cliffs, NJ: Prentice Hall.

Taylor TN and Taylor EL (2000) The Rhynie Chert ecosystem: a model for understanding fungal interactions. In: Bacon CW and White JF (eds.) *Microbial Endophytes,* pp. 31 47. New York: Marcel Dekker, Inc.

Taylor TN, Remy W, and Hass H (1992) Fungi from the Lower Devonian Rhynie Chert: Chytridiomycetes. *American Journal of Botany* 79: 1233 1241.

Taylor TN, Hass H, and Kerp H (1997) A cyanolichen from the Lower Devonian Rhynie Chert. *American Journal of Botany* 84: 992 1004.

Tiffney BH and Barghoorn ES (1974) The fossil record of the fungi. *Occasional Papers of the Farlow Herbarium of Cryptogamic Botany* 7: 1 42.

Waggoner BM (1995) Ediacaran lichens: a critique. *Paleobiology* 21: 393 397.

Gymnosperms

P Kenrick, The Natural History Museum, London, UK

Introduction

Gymnosperms are an ancient group of plants that include some familiar living forms such as conifers, cycads, and ginkgos as well as a great variety of far less well-known and long-extinct kinds. The modest number of living species (approx. 750) belies the geological importance of the group. Gymnosperms dominated the Mesozoic era. It was during this time that many of the major subgroups originated, diversified, and finally succumbed to their ultimate fate, extinction. The evolution of gymnosperms and in particular their relationships with the angiosperms (flowering plants) is also the subject of much debate. Phylogenetic studies clearly show that these two groups are close relatives, but the exact nature of their kinship is still unclear. Knowing how these two groups are related is essential to developing a family tree of gymnosperms and to understanding the origin of the flowering plants. Because much of the diversity of gymnosperms is extinct, the fossil record continues to play a crucial role in elucidating evolutionary patterns. Recently, molecular systematics has introduced an enormous amount of new comparative data. This has brought unprecedented precision in our ability to reconstruct family trees, shedding new light on the evolution of gymnosperms and their lengthy and chequered fossil history.

Characteristics

Gymnosperms are a varied assemblage of plants whose principal botanical characteristic is reproduction by means of ovules, or seeds as they are called in their fertilized state. This is an aspect of their biology that they share with the angiosperms. One of the main differences between gymnosperms and angiosperms relates to the appendages associated with the ovules. In the angiosperms, ovules are part of the flower. One or more ovules is wrapped within a layer of tissue, the whole structure being called the carpel. Flowers usually comprise carpels and additional organs such as pollen-bearing sacs, petals, and sepals. Gymnosperms do not have flowers. The ovules are borne in cones or singly on more or less modified branches. Furthermore, the ovules are not enclosed within a carpel, they are 'naked', hence the origin of the term gymnosperm (Greek 'gumnospermos': 'gumnos', naked + 'sperma', seed). Modern gymnosperms are generally also woody shrubs and trees, but some have a vine-like growth form. A more diverse array of growth forms, including groups with fern-like leaves, is known from the fossil record. Ovules and seeds, wood, leaves, and pollen are frequently preserved as fossils, and together these provide our principal sources of evidence on the geological history of the group.

Classification

Recent classifications of gymnosperms are based upon phylogenetic studies. Because many aspects of gymnosperm phylogeny still remain unclear, modern classifications at the family or ordinal level must be regarded as very tentative. There are problems in placing some extinct groups because they are still very poorly characterized. In other instances, there are conflicts between the phylogenetic trees generated from comparative morphology and those produced from molecular systematic studies. Despite these problems there is reason to be optimistic in the longer term. The history of gymnosperm classification shows that better characterization of problematic fossils tends to lead to more stable ideas on relationships, and in the molecular world there is a seemingly inexhaustible supply of new data that should lead ultimately to a firm backbone phylogeny for the living groups. Both approaches should converge on a mutually consistent and stable family tree.

One of the findings of recent phylogenetic studies is that some groups of gymnosperms are more closely related to angiosperms than they are to other gymnosperms (Table 1). In technical terms, gymnosperms are paraphyletic to angiosperms. The precise nature of this relationship, however, remains unclear, and this is one major source of conflict between family trees based on comparative morphology and those derived from molecular data. Comparative morphology indicates that angiosperms are most closely related to living gymnosperms in the Gnetales and extinct gymnosperms in the Pentoxylales and the Bennettitales. This has been dubbed the 'anthophyte hypothesis' (Figure 1). Molecular data tell a very different story. They place Gnetales within Coniferales. Angiosperms emerge as sister group to a monophyletic group comprising all living gymnosperms. This has been dubbed the 'gne-pine' hypothesis (Figure 1). Neither hypothesis though is particularly strongly supported by the current data. Furthermore, because molecular data cannot deal with the fossil groups, it is of course unclear where in molecular

Table 1 Classification of seed plants (Spermatophyta)

Spermatophyta	
'Calamopityales' *incertae sedis*[a, b]	
Hyrdaspermales[a, b]	
'Lyginopteridales'[a, b]	
Medullosales[a, b]	
Euspermatales (taxa *sedis mutabilis*)	
Cycadales	
Callistophytales[a, b]	
Coniferophytales	
Cordaitiales[b]	
Coniferales (Pinales)	'Gymnosperms'
Glossopteridales[a, b]	
Czekanowskiales[b]	
Ginkgoales	
'Peltaspermales'[b]	
'Corystospermales' ('Umkomastales')[b]	
Caytoniales[b]	
Pentoxylales[b]	
Bennettitales[b]	
Gnetales	
Archaefructaceae[b]	? Angiosperm
Angiospermopsida	Angiosperms

Grouping is indicated through the use of indentation and the sequence a name occupies in the list. Commonly applied names for individual groups are used without regard to Linnean rank order convention. Questionably monophyletic groups are marked with quotes. *Sedis mutabilis* indicates that groups which follow at the next level of the hierarchy are of uncertain relationship (i.e., form a polytomy or conflict strongly in different phylogenetic analyses). *Incertae sedis* indicates that the taxon bearing this epithet is of uncertain relationship to those at the same level. Gymnosperms is a paraphyletic group. It is equivalent to seed plants minus flowering plants (i.e., Spermatophyta minus Angiospermopsida). Modified after Crane and Kenrick (1997) Problems in cladistic classification: Higher level relationships in land plants. *Aliso* 15: 87 104.
[a]Groups typically included as Pteridosperms, or seed ferns, which are a heterogenous assemblage of gymnosperms with fern like foliage.
[b]Extinct group.

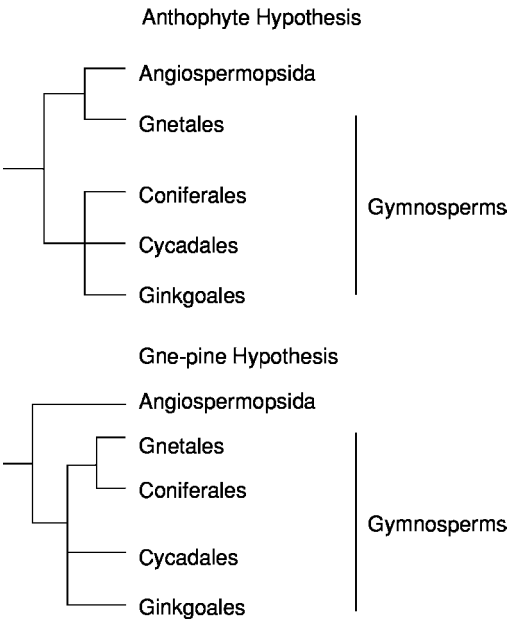


Figure 1 Two hypotheses of how living gymnosperms are related to angiosperms (flowering plants). Morphological data favour the anthophyte hypothesis, whereas molecular data favour the gne pine hypothesis.

Figure 2 Frond of the progymnosperm *Archaeopteris hibernica* (Upper Devonian, Kiltorkan, Kilkenny, Ireland).

phylogenetic trees the fossils would in fact fall out. However, it seems likely that whichever hypothesis ultimately wins out, Early Palaeozoic gymnosperms such as Calamopityales, Hydraspermales, Lyginopteridales, and possibly Medullosales (Table 1) will remain basal in a phylogenetic sense to both angiosperms and to the more derived living gymnosperms. Also, it is plausible that extinct Mesozoic groups such as Pentoxylales, Bennettitales, and possible Caytoniales are more closely related to angiosperms than they are to other gymnosperms. From a phylogenetic perspective, angiosperms are in reality little more than a morphologically divergent group of gymnosperms.

Gymnosperm Origins

Gymnosperms originated from a grade of Late Devonian plants called the progymnosperms. The best-known exemplar is *Archaeopteris*. This plant had large fern-like fronds, some of which are known to have been well over 1 m in length (Figure 2). Like ferns, *Archaeopteris* produced spores rather than seeds. However, unlike true ferns, the fronds were borne on woody branches. In this respect, *Archaeopteris* resembled gymnosperms. Some species of *Archaeopteris* were undoubtedly large trees, as trunks of the distinctive wood (*Callixylon*) are known to exceed several metres in diameter. This extraordinary extinct plant therefore possessed a unique combination of characteristics that are intermediate between gymnosperms and early relatives of the ferns. It shows that the woody architecture that characterizes gymnosperms evolved before other defining features, such as the ovule.

The earliest gymnosperms are found in Upper Devonian sediments of Europe and North America. These are known from plant fragments such as dispersed seeds and seed-bearing branches (Figure 3) as well as a handful of more completely known plants. *Elkinsia polymorpha* (Hydraspermales) from the Upper Devonian of West Virginia, USA, is one of the more completely characterized forms. Petrified ovules in which cellular anatomy is preserved show gymnosperm features alongside other features that differ from modern forms. Specifically, many early ovules lacked a completely formed integument (enveloping tissue layer), so they are sometimes called 'preovules'. Some also had additional specialization to aid fertilization (pollen capture). In many, ovules were borne on the ends of branches in 'cupules'. The earliest gymnosperms were small woody shrubs that would have had a fern-like appearance (Figure 4).

Figure 3 Ovule (seed) bearing cupules of the early gymnosperm *Xenotheca devonica* (Upper Devonian, Devon, England).

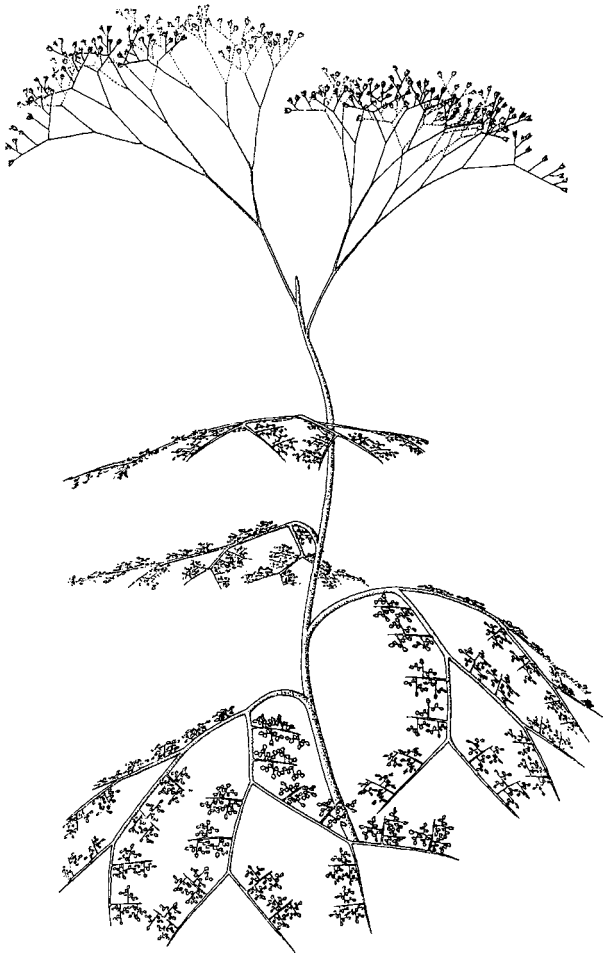


Figure 4 Reconstruction of the early gymnosperm *Elkinsia polymorpha* (Upper Devonian, USA). After Serbet and Rothwell (1992). *International Journal of Plant Sciences* 153: 602 621.

This fern-like theme in gymnosperm evolution continued into the Carboniferous period. Much of the fern-like foliage of coal swamp forests is actually attributable to extinct gymnosperms. These are the so-called 'seed ferns' or Pteridosperms (Figure 5). As the name implies, these were seed-bearing plants that possessed fern-like leaves, which is a combination of characteristics seen only in extinct gymnosperms. Groups common but not exclusive to the Carboniferous period include Calamopityales, Hydraspermales, Medullosales, Lyginopteridales, and Callistophytales. For the most part, these were understory shrubs and small trees. Some are also known to have been vine-like climbers. The earliest gymnosperms of more modern aspect were an extinct group of conifers called the Voltziales (Figure 6). This group first appeared in the Late Carboniferous and the plants are thought to have resembled in general habit and leaf morphology living conifers such as *Araucaria heterophylla* (Norfolk Island Pine).

During the Late Palaeozoic, gymnosperms were an important component of terrestrial floras, but there was a significant increase in species diversity during the Early Mesozoic. Numerous petrifactions of stem and trunk show that most Mesozoic gymnosperms, like their modern relatives, were woody shrubs and trees. Evidence from fossil forests deposits reveal landscapes dominated by gymnosperms, which formed the main canopy element (Figure 7). In particular, the conifers underwent a major radiation during the Triassic period. Studies of the pollen record show that the living Gnetales diversified alongside angiosperms during the mid-Cretaceous, but then underwent a decline towards the end of this period.

Molecular phylogenetic studies provide an alternative perspective on the fossil history of some groups of gymnosperms. These are beginning to yield an unprecedented level of phylogenetic resolution in the most species-rich living groups, and they are providing additional insights into long-standing phylogenetic conundrums. Molecular data support the hypothesis that conifers had a single origin, clearing up a long-standing controversy over the phylogenetic position of the Taxaceae (Yews) (Figure 8). Also, within conifers, the pines (Pinaceae) come out as sister group to all other living species. As long suspected, the monkey puzzle family Araucariaceae is closely related to the Yellow-Wood family, Podocarpaceae. In addition to

Figure 5 Fern like frond of the extinct gymnosperm *Neuropteris heterophylla* (Carboniferous, Clay Cross, Derbyshire, England).

Figure 6 Leafy branch of *Walchia piniformis* (Permian, Montpellier, France).

providing phylogenetic information, one can estimate the divergence times of these groups based on a molecular clock approach. Despite some shortcomings, this approach provides an additional perspective on the evolutionary history of gymnosperms. The results of these gene-based calibrations of various family trees for living gymnosperms can and should be compared against direct evidence from the fossil record. Within cycads, family trees calibrated using the chloroplast gene *rbc*L indicate significantly later origins of genera and other major groupings than fossil evidence would support. This has been used to suggest that the disjunct distributions of modern genera are the result of Miocene or later dispersals. The fossil record indicates that this late dating of groups is unlikely because of the wider geographic distributions of modern genera and their longer histories through the Tertiary.

Major Gymnosperm Groups

Hydraspermales (Upper Devonian to Lower Carboniferous)

This extinct group contains the earliest known gymnosperms. The Hydraspermales are characterized by a very particular suite of morphological characteristics associated with their seed-like reproduction and pollination biology. The ovules of Hydraspermales are technically called preovules because they are not completely enclosed within an integument—the outer coat of later true ovules. Detailed information on the morphology and anatomy of these preovules is providing insights into the early evolution of the seed. The preovules of Hydraspermales had a unique structure termed the lagenostome, which takes the form of a funnel, ring, or cup situated at the apex. This structure was intimately involved with the capture of pollen. Preovules were borne in clusters in lobed cupulate structures. Many Hydraspermales are known only as plant parts, such as isolated seeds or cupules (Figure 3). One of the most completely characterized plants is *Elkinsia polymorpha* from the Upper Devonian of West Virginia, USA. This was a small, slender shrub that bore delicate fronds of the *Sphenopteridium* type (Figure 4).

Calamopityales (Upper Devonian to Lower Carboniferous)

This is an extinct group of rather poorly characterized small- to medium-size shrubs. The concept of Calamopityales is based largely on fragments of petrified stem, and because of this our knowledge of the whole

Figure 7 210 My old silicified tree trunks from the Petrified Forest National Park, Arizona, USA. Most of the wood is attributable to the conifer *Araucarioxylon* (Late Triassic, Chinle Formation, Arizona, USA). © Paul Davis.

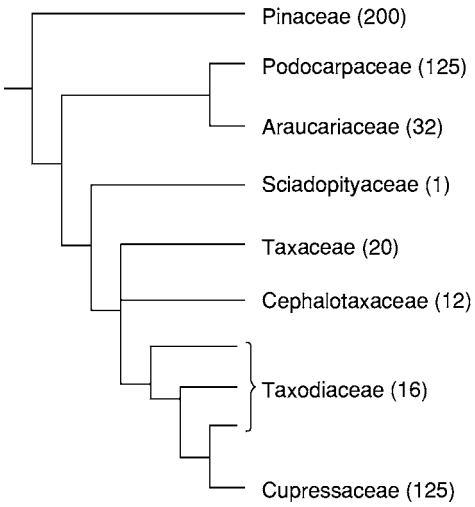


Figure 8 Phylogeny of the conifers inferred from partial 28S rRNA gene sequences. Approximate numbers of living species are given in parentheses. After Stefanovic *et al.* (1998) Phylogen etic relationships of conifers inferred from partial 28S rRNA gene sequences. *American Journal of Botany* 85: 688 697.

plants is poorly developed. The seeds, pollen organs, and foliage of Calamopityales are for the most part unknown. On the basis of frequent association, it seems likely that stems called *Stenomyelon* bore seeds known as *Lyrasperma*, which are found dispersed in sediments. Also, it is probable that Calamopityales foliage was of the general fern-like frond variety, as exemplified by form genera such as *Triphyllopteris*, *Adiantites*, and *Sphenopteridium*.

Lyginopteridales (Mid- to Late Carboniferous)

This extinct group contained small slender shrubs and possibly also trees of substantial size. The group is known mainly from petrifactions documented in coal balls, and many of the defining characters come from the often exquisitely preserved anatomy of the stem. *Lyginopteris* was probably a small shrub or a vine-like plant. It bore large fronds of the *Sphenopteris* or *Pecopteris* type. Ovules were borne in lobed cupules, and pollen developed within cylindrical sacs that were borne fused in clusters. Other possible member of the group is *Pitys*, which was a large tree with trunks in excess of 1 m diameter and 5 m in height.

Medullosales (Lower Carboniferous to Permian)

The Medullosales were among the largest of the seed ferns. This extinct group comprised small- to medium-size trees, some of which exceeded 3.5 m in height. In overall appearance, *Medullosa* bore a superficial resemblance to a modern tree fern. The trunk bore a crown of large fern-like leaves (Figure 5). Large seeds—up to 7 cm long—were borne on the underside or at the tips of fronds (Figure 9). The Medullosales produced large characteristic pollen grains in morphologically complex spore-bearing organs. Some species may have occupied mangrove-like habitats.

Cycadales (? Carboniferous to Recent)

This group contains about 140 living species classified into 11 genera and up to 4 families (Cycadaceae, Zamiaceae, Boweniaceae, Stangeriaceae) (Figure 10).

Figure 9 Large nutlike seeds (about 2.5 cm long) of *Trigonocarpus parkinsoni* (Upper Carboniferous, England).

Figure 10 Living cycad in the Fairchild Botanical Garden, Florida, USA.

The group has a predominantly tropical to subtropical distribution. Some species reach 18 m in height, whereas others are small with tuberous stems. Cycads bear one or more crowns of large compound leaves. Seeds and pollen are borne in cone-like structures on separate plants. Large species of cycads bear a superficial resemblance to palm trees with stout stems. They are also closely similar in overall appearance to extinct Mesozoic gymnosperms in the Bennettitales. Significant differences in the structure of the seed and pollen-bearing organs indicate that the two are only distant relatives. Cycads probably originated in the Late Palaeozoic. Plants such as *Archaeocycas* are known from the Lower Permian, and the ancestors of modern cycads are thought to extend back into the Carboniferous. The group reached its maximum diversity during the Mesozoic era.

Callistophytales (Late Carboniferous)

This small extinct group was first characterized in detail in 1975. The best known species is *Callistophyton poroxyloides*. The stem anatomy of the group resembles that of Lyginopteridales, whereas the reproductive structures are closer to those of Cordaitales. *Callistophyton* is portrayed as a small shrubby plant with stems up to 3 cm diameter, bipinnate fronds, and a scrambling habit. The leaves are similar to fossil foliage of *Medullopteris* type. Ovules and pollen organs were borne on the underside of unmodified leaves. Pollen was monosulcate with a bladder, resembling the pollen of conifers.

Cordaitales (Late Carboniferous to Early Permian)

This distinctive extinct group had a cosmopolitan distribution. Some species were large trees that attained heights of as much as 30 m, whereas at least one species is known to have been a prostrate scrambler some 2 m in height. The leaves of many species were large and strap-like, bearing a striking resemblance to the leaves of flowering plant monocotyledons such as *Hippeastrum* (*Amaryllis*) or living conifers in the genus *Agathis*. Pollen- and ovule-forming structures were borne on shoots scattered among the leaves. The wood was similar in structure to that of *Araucarioxylon*. The centre of the stems contained a large pith cavity with transverse septa. Fossilized infillings of this pith (*Artisia*) are commonly found. The reproductive structures are compound fructifications comprising cones subtended by bracts borne along a central axis.

Voltziales (Late Carboniferous to Early Jurassic)

This extinct group of conifer-like plants comprises three families: Utrechtiaceae (Lebachiaceae), Emporiaceae, and Majoniaceae (Figure 6). On the whole, these are thought to resemble in general habit and leaf morphology living conifers such as *Araucaria heterophylla* (Norfolk Island Pine). Some species had bisaccate pollen like that of living Pinaceae. The structure of the wood also resembles that of living conifers in the Araucariaceae. The fructifications of Voltziales

resemble cones, and they were borne at the tips of leafy branches. The major difference between the cones of true conifers lies in the nature of the structure that bore the ovules. The ovule-bearing scales are attached to a short shoot rather than directly to the bract as in conifers. This condition is interpreted as intermediate between the fructifications of Cordaitales and the cones of Coniferales, to which Voltziales are clearly closely related.

Coniferales (Triassic to Recent)

This major group of gymnosperms contains some 537 living species classified into 69 genera and 9 families (Pinaceae, Podocarpaceae, Araucariaceae, Phyllocladaceae, Sciadopityaceae, Cephalotaxaceae, Cupressaceae, Taxaceae, Taxodiaceae). Most living members are medium to large evergreen trees, but the group contains shrubs and one rare parasitic species (*Parasitaxus*, Podocarpaceae). Several extinct family level units are recognized, of which the best known and most diverse is the Mesozoic Cheirolepidiaceae. The leaves of conifers are on the whole simple, needle-shaped or flattened and strap-shaped and more rarely ovate-lanceolate-elliptical. Pollen and ovules are produced in cones, except for Taxales in which single ovules terminate branches. Araucariaceae (~32 living species) have a tropical–subtropical Old World distribution, but they are absent from Africa. This is an ancient family that probably originated during the Late Permian. Araucariaceae had a much wider geographic distribution during the Mesozoic in both southern and northern hemispheres (Figure 11). Cephalotaxaceae (~12 living species) are plants of temperate to subtropical south-east Asia. They have a fossil record that extends possibly as far back as the Triassic. Cupressaceae (~125 living species) is a cosmopolitan group. Leaf shoots assignable to this family have been described from the Early Jurassic. The extinct Cheirolepidiaceae range from Triassic to Mid-Cretaceous. These were probably mostly trees with scale-like foliage of the Cupressaceae type (Figure 12). They produced a distinctive type of pollen called *Classopollis*. Pinaceae (~200 living species) are almost exclusively northern hemisphere. This is the largest living family of Coniferales. The fossil record of Pinaceae might extend back into the Triassic, but it seems likely that the crown group began to diversify comparatively late during the Cretaceous period. Podocarpaceae (~125 living species) are tropical–subtropical mostly montane conifers of the southern hemisphere. This ancient family is first documented in the Late Triassic. The monotypic Sciadopityaceae is confined to temperate south-east Asia. A Late Cretaceous or Palaeocene origin of this group seems likely. Taxaceae (~20 living species) occur mainly in the northern hemisphere. The earliest fossils are Early Jurassic. Taxodiaceae (~16 living species) are mainly northern temperate—subtropical. This family is known from the Jurassic (Figure 13). Fossils leaves in the form genus *Elatides* are most like living *Cunninghamia*. Much fossil wood of Mesozoic age is assignable to the Coniferales.

Glossopteridales (Permian to Triassic)

This is an extinct group of Gondwanan gymnosperms. Over 50 species have been described based on leaf morphology. Many of these are probably variants of

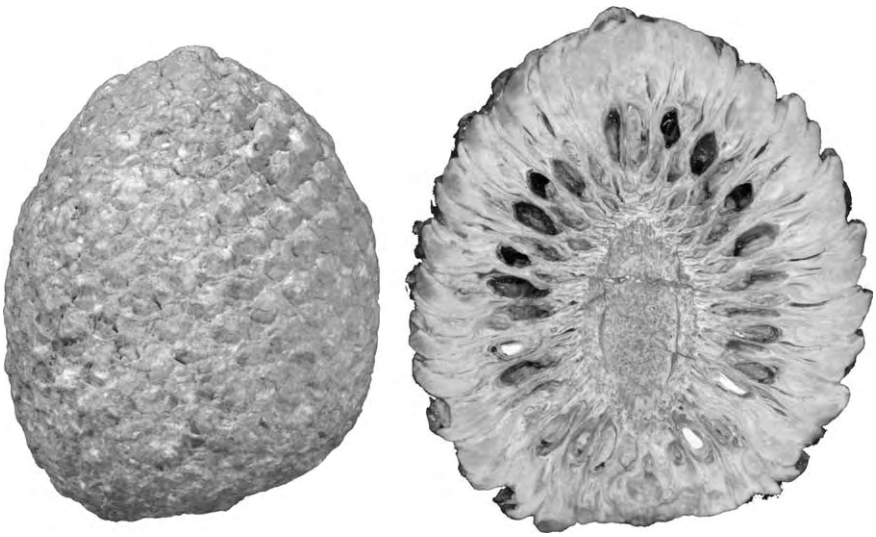


Figure 11 Silicified cones of *Araucaria mirabilis* (Upper Jurassic, Cerro Alto, Santa Cruz, Argentina). Specimen on right cut and polished to reveal inner structure.

Figure 12 Thick, robust leaves of the extinct conifer *Pagiophyllum peregrynum* (Upper Jurassic, Lyme Regis, Dorset, England). Foliage of this type could belong to several families of conifers including Araucariaceae and Cheirolepidiaceae.

Figure 14 Tongue shaped leaves of *Glossopteris* are characteristic of the Permian and Triassic rocks of Gondwana: *G. browniana* (Permian, Nagpur, India).

Figure 13 Foliage of the Dawn Redwood *Metasequoia* was documented in the fossil record long before living specimens of these large conifers were discovered in the Chinese province of Szechuan in 1944 (Eocene, Driftwood Canyon, Smithers, British Columbia, Canada).

the same species or individuals. Some are known to have been large trees. The leaves of Glossopteridales are highly characteristic. They are large and tongue-shaped with a conspicuous midrib and a network of veins (Figure 14). Pollen organs and ovules were borne on special branches on the upper surfaces of some leaves. Some fructifications are foliar, whereas others are thought to resemble cones.

Czekanowskiales (Jurassic to Cretaceous)

This is an extinct and poorly understood group. The leaves of Czekanowskiales are distinctive. They are highly dissected, in some species almost like a bundle of fibres. Ovules are thought to have been borne in cones composed of capsulate appendages. The habit of the plant remains poorly understood.

Ginkgoales (? Permian to Recent)

This group contains one living species *Ginkgo biloba*, which is a large tree that bears distinctive fan-shaped leaves. Pollen organs are loose catkin-like cones, whereas ovules are in pairs on stalks. They are borne on short leafy shoots on separate plants. The natural distribution of modern *Ginkgo* is restricted to a small area of China, but leaves of *Ginkgo* type are

widespread in the fossil record of the Mesozoic and Tertiary eras. The leaf morphology ranges from highly dissected resembling leaves of Czekanowskiales to forms very similar to those of the modern species (Figure 15). Even though ovules attributable to the Ginkgoales were common in their dispersed form during the Mesozoic very few intact reproductive structures are known. Ovulate structures and catkin-like pollen cones have been reported from the Jurassic of Yorkshire, England.

Peltaspermales (Permian to Late Triassic)

The concept of Peltaspermales is based on leaves, pollen organs, and ovulate organs. Knowledge of whole plants is scant, but organ associations and characteristic cuticular features have been used to piece together some aspects of whole plant morphology. Clusters of pollen sacs and ovule-bearing structures were borne in a pinnate arrangement on branches. Plants bore large fern-like leaves, which were probably deciduous. This is consistent with a woody habit.

Corystospermales (Mid-Triassic to Mid-Jurassic)

The concept of this group is based on leaves, pollen organs, and ovulate organs. Knowledge of whole plants is scant, but organ associations, characteristic cuticular features, and the association of pollen with ovules have been used to piece together some aspects of whole plant morphology. Woody stems bore fern-like foliage. Ovules were borne in cupules, whereas pollen was borne in elongated sacs on the underside of oval laminae. Woody stems bore fern-like foliage.

Caytoniales (Upper Triassic to Lower Cretaceous)

This extinct group is widely recognized on the basis of distinctive leaves. The leaves (*Sagenopteris*) are palmate, composed of three to six lanceolate blades attached to a petiole (Figure 16). Each blade has an anastomosing network of veins. Knowledge of whole plants is scant, but the concept of the group has been built from well-established organ associations. Ovules were borne in cupules along a branch (*Caytonia*). The pollen was produced in elongated anther-like structures (*Caytonanthus*). *Caytonia* is though to have been a small tree with a woody trunk.

Pentoxylales (? Jurassic to Cretaceous)

This extinct group is known from India, New Zealand, and Australia. Knowledge of whole plants is scant, but the stems (*Pentoxylon*) are known to have borne a distinctive type of wood, resembling that of conifers. The leaves were large (up to 20 cm long), strap-shaped, and borne on short lateral shoots.

Figure 15 Leaf of the Maidenhair tree, *Ginkgo gardneri* (Palaeocene, Isle of Mull, Scotland).

Figure 16 The extinct gymnosperm *Caytonia* bore palmate leaves: *Sagenopteris phillipsi* (Jurassic, Cayton Bay, Yorkshire, England).

Figure 17 Stout, barrel shaped trunk of the extinct Bennettitalean gymnosperm *Cycadeoidea microphylla* (Jurassic, Isle of Portland, Dorset, England).

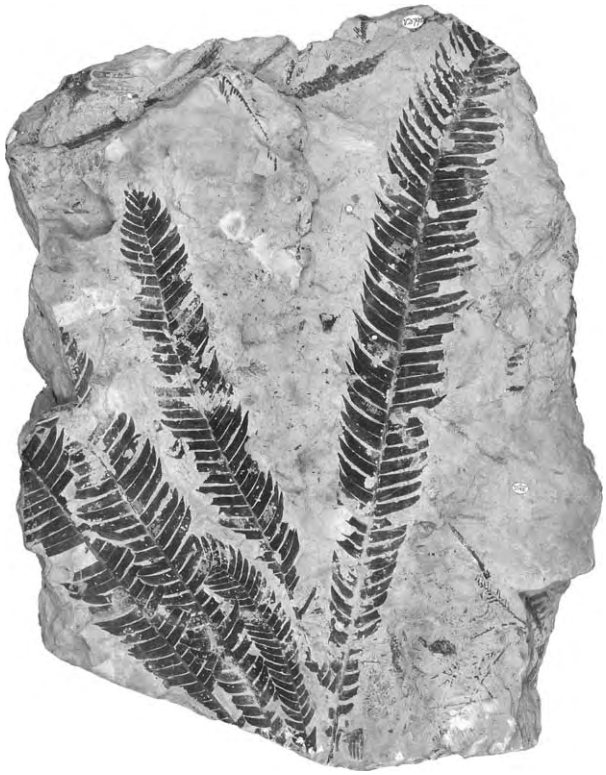


Figure 18 Leaves of the extinct cycad like Bennettitalean gymnosperm *Zamites gigas* (Jurassic, Yorkshire, England).

Ovules were borne in clusters of compact cones. Pollen organs were borne in clusters on a receptacle. These plants were probably small trees or shrubs.

Bennettitales (Triassic to Upper Cretaceous)

This extinct group had a cosmopolitan distribution. In habit, Bennettitales were shrubs or small trees with stout, sparsely branched trunks or squat, unbranched, and barrel-shaped trunks (Figure 17). The cones of Bennettitales were highly distinctive flower-like structures. Cones in the genus *Williamsoniella* comprised a central ovule bearing receptacle subtended by a whorl of rather broad, leaf-like pollen-bearing segments. Whorls of leaf-like bracts subtend the pollen and ovule-bearing parts. Cones were borne either along branches (Williamsoniaceae) or along the squat trunks embedded and partly protruding from a thick layer of old leaf bases (Cycadeoidaceae). In some, pollen- and ovule-bearing parts were in separate cones (e.g., *Williamsonia*), whereas in others they were borne within the same cone (e.g., *Williamsoniella*). Leaves were large and either pinnate or flat bladed (Figure 18). The striking superficial similarity to modern cycads masks fundamental differences in anatomy and especially the structure of the cones.

Gnetales (Late Triassic to Recent)

This group of gymnosperms contains approximately 70 living species classified into three genera and families (Ephedraceae, Gnetaceae, Welwitschiaceae). Living members are predominantly shrubs or vines, but the group contains some small trees. Of the living groups of gymnosperms, the fossil record of gnetales is most poorly understood. Macrofossil remains are sparse, but there is a lengthy pollen record that begins in the Late Triassic. Fossil pollen resembles that of living Ephedraceae and Welwitschiaceae. The record of dispersed pollen indicates that living Gnetales are relicts of a group that was once more widespread and much more diverse. Maximum diversity was achieved during the mid-Cretaceous. Some putative early Gnetales macrofossils resemble conifers, raising the possibility that the poor fossil record might in part be explained by misidentification (Crane, 1996).

See Also

Dendrochronology. **Fossil Plants:** Angiosperms. **Mesozoic:** Triassic; Jurassic; Cretaceous. **Microfossils:** Palynology. **Palaeozoic:** Devonian; Carboniferous. **Tertiary To Present:** Oligocene.

Further Reading

Beck CB (ed.) (1988) *Origin and Evolution of Gymnosperms*. New York: Columbia University Press.

Crane PR (1996) The fossil history of the Gnetales. *International Journal of Plant Sciences* 157: S50–S57.

Crane PR and Kenrick P (1997) Problems in cladistic clas sification: higher level relationships in land plants. *Aliso* 15: 87 104.
Donoghue MJ and Doyle JA (2000) Seed plant phylogeny: demise of the anthophyte hypothesis. *Current Biology* 10: R106 R109.
Kramer KU and Green PS (eds.) (1990) *The Families and Genera of Vascular Plants. I. Pteridophytes and Gymno sperms*. Berlin: Springer.
Meyen SV (1984) Basic features of gymnosperm systematics and phylogeny as evidenced by the fossil record. *Botan ical Review* 50: 1 111.
Rydin C, Kallersjo M, and Friis EM (2002) Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: Conflicting data, rooting problems, and the monophyly of con ifers. *International Journal of Plant Sciences* 163(2): 197 214.
Stefanovic S, Jager M, Deutsch J, Broutin J, and Masselot M (1998) Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. *American Jour nal of Botany* 85: 688 697.
Stewart WN and Rothwell GW (1993) *Paleobotany and the Evolution of Plants*. Cambridge: Cambridge University Press.
Treutlein J and Wink M (2002) Molecular phylogeny of cycads inferred from rbcL sequences. *Naturwissenschaf ten* 89(5): 221 225.

FOSSIL VERTEBRATES

Contents

Jawless Fish-Like Vertebrates

D K Elliott, Northern Arizona University, Flagstaff, AZ, USA

Introduction

Jawless fishes (agnathans) have a fossil record that spans almost 500 million years, now extending back to the Early Cambrian of China (*see* **Palaeozoic:** Cambrian). They were particularly prominent during the Late Silurian and Devonian (*see* **Palaeozoic:** Devonian; Silurian) when they radiated to form a variety of diverse groups, most of them characterised by the presence of bony armour plating over the head and body. They are valuable in biostratigraphy, helping to determine the placement (*see* **Stratigraphical Principles**) of continents during the Palaeozoic and they chart the development of many features important in vertebrate history such as the appearance of bone, paired fins, lateral line systems, and an inner ear with two semicircular canals. During the last decade there has been a dramatic increase in the amount and quality of information about the origin of vertebrates and the diversification of the earliest forms. This has stemmed partly from new finds that have provided us with examples of early forms and also from new

methods of analysis, which have enabled the development of a better understanding of relationships (*see* **Micropalaeontological Techniques**).

Origin and Relationships of Vertebrates

Vertebrates belong to the phylum Chordata, which also includes the marine cephalochordates and tunicates. Vertebrates have the basic chordate characters of a notochord (a tough flexible rod that runs down the back) and V-shaped muscle blocks or myotomes along the length of their bodies but they also have the hard tissue bone and elements of the vertebral column, characters not shared by the other chordates. It is generally considered that the cephalochordates, small eel-like marine filter feeders exemplified by amphioxus, are the closest relatives of the vertebrates while the tunicates, attached marine suspension feeders, are considered to be less closely related. Within the vertebrates the significance of the development of the jaw is recognised by the classification of vertebrates into gnathostomes ('jaw mouths') and agnathans ('without jaws'). The jawless vertebrates are considered to be the most primitive and were originally separated as the Agnatha by Cope in 1889. At that time the Agnatha included the modern lampreys and hagfishes, together with a number of fossil groups often termed ostracoderms ('shell skin'). The modern forms share a similarity in their primitive eel-like appearance and their feeding habits (lampreys are parasites on other fish, sucking their blood and rasping their flesh, while hagfish eat their way through dead or dying fishes). The absence of jaws also assigns the ostracoderms to the Agnatha, however, it is clear that many of them also share characters with the gnathostomes, thus leading to the problem of their relationship to modern vertebrates. Developing an understanding of these relationships has been difficult because of the generally poor preservation of the internal anatomy of ostracoderms and because careful studies of their anatomy were not carried out until recently. However, analysis of some forms using serial grinding techniques in the early 1900s, together with recent discoveries of many new kinds of ostracoderms and the development of new acid preparation techniques have greatly improved our knowledge of these organisms.

It has become clear more recently that the lampreys share a large number of characters with the gnathostomes and they are now considered to have a sister-group relationship with them, the hagfishes being the sister group of lampreys plus vertebrates. The grouping of lampreys plus gnathostomes is now more correctly referred to as the Vertebrata because elements of the vertebral column are present while these are missing in the hagfishes. The more inclusive group including hagfishes is termed the Craniata because they all have a skull but frequently the two terms are used interchangeably and in this article I will use Vertebrata.

Early Chordates

The earliest chordates are known from the Chengjiang and Burgess Shale lagerstätte which are Early and Middle Cambrian in age, respectively (*see* **Palaeozoic:** Cambrian). Of these the best known is *Pikaia gracilens* which is a leaf-shaped organism with clearly defined sigmoidal muscle blocks and a notochord. There is some doubt about its cephalochordate affinities, however, as there is no evidence for any gill-like structures. A similar animal from the Chengjiang fauna, *Cathaymyrus*, is also thought to be a cephalochordate and both have been compared with the modern cephalochordate amphioxus. Thus, there may be evidence of the closest relatives of vertebrates as far down as the Early Cambrian. Very recently however, two additional species have been reported from Chengjiang, *Myllokunmingia* and *Haikouichthys*, both of which are described as vertebrates. Both show zig-zag shaped muscle blocks, relatively complex and presumably cartilaginous skulls, gill arches, heart, and fin supports but the presence of serially arranged gonads and a dorsal fin with rostrally tilted radials casts doubt on their inclusion within vertebrates.

Prior to the discovery of the Chengjiang vertebrates, the evidence for the presence of vertebrates in the Cambrian rested on the preservation of fragments of phosphatic hard tissue. Of these *Anatolepis* has been reported from the Late Cambrian although initially described from the Early Ordovician of Spitsbergen (*see* **Palaeozoic:** Ordovician). It consists of microscopic plates and spines with scale-like ornamentation and a layered internal structure. Although there has been some controversy over the vertebrate attribution of this material, recent studies of the histology of these fragments show that they do contain the characteristic vertebrate hard tissue dentine. Although fragments of purported vertebrate hard tissue have also been recently reported from the Late Cambrian of Australia there is some doubt as to their affinity as thin sections show a resemblance to some arthropod cuticles.

Conodonts

Conodonts are small tooth-like structures made of apatite and have been known since the mid-1800s.

Although extremely abundant in marine rocks and used extensively in biostratigraphy their vertebrate affinities have only recently been understood (*see* **Microfossils:** Conodonts). The hard tissue of the crown consists of an enamel-like material while the underlying basal body includes the mineralised tissue dentine. The discovery of complete conodont animals from the Mississippian Granton Shrimp Beds of Scotland in 1983 demonstrated that they were small and eel-like (22–25 mm long) with a feeding basket composed of the conodont elements that functioned in seizing prey and cutting it up (Figure 1). Additional specimens from the Silurian of Wisconsin and the Upper Ordovician (*see* **Palaeozoic:** Silurian; Ordovician) of South Africa are larger (up to 40 cm long) and have demonstrated the presence of a number of chordate characters including a notochord, a dorsal nerve cord, the presence of myotomes or muscle blocks, the presence of a midline tail fin, and a bilaterally symmetrical body. Further characters that place the conodonts within the vertebrates include a head lying in front of the notochord, a caudal fin with radial supports, extrinsic eye musculature, and an apatitic skeleton. Interestingly, the presence of hard tissues in conodonts indicates that they have more advanced characters than either the hagfishes or the lampreys.

Agnathan Diversity

At least 600 species of ostracoderm are known and most species were covered in a bony armour. This armour may have developed as a storage area for calcium and phosphate or to prevent loss of water from the body tissues, or it may have developed as a protection from predators. In early vertebrates two principal types of bone occur: acellular bone (aspidin) which was laid down by the dermis and is often laminated, and cellular bone in which there are cell spaces for osteocytes (bone-forming cells) throughout the bone. The ostracoderms show a bewildering variety of types of bone and plate and tail shapes. In spite of this variety though most ostracoderms can be classified into one of a small number of groups.

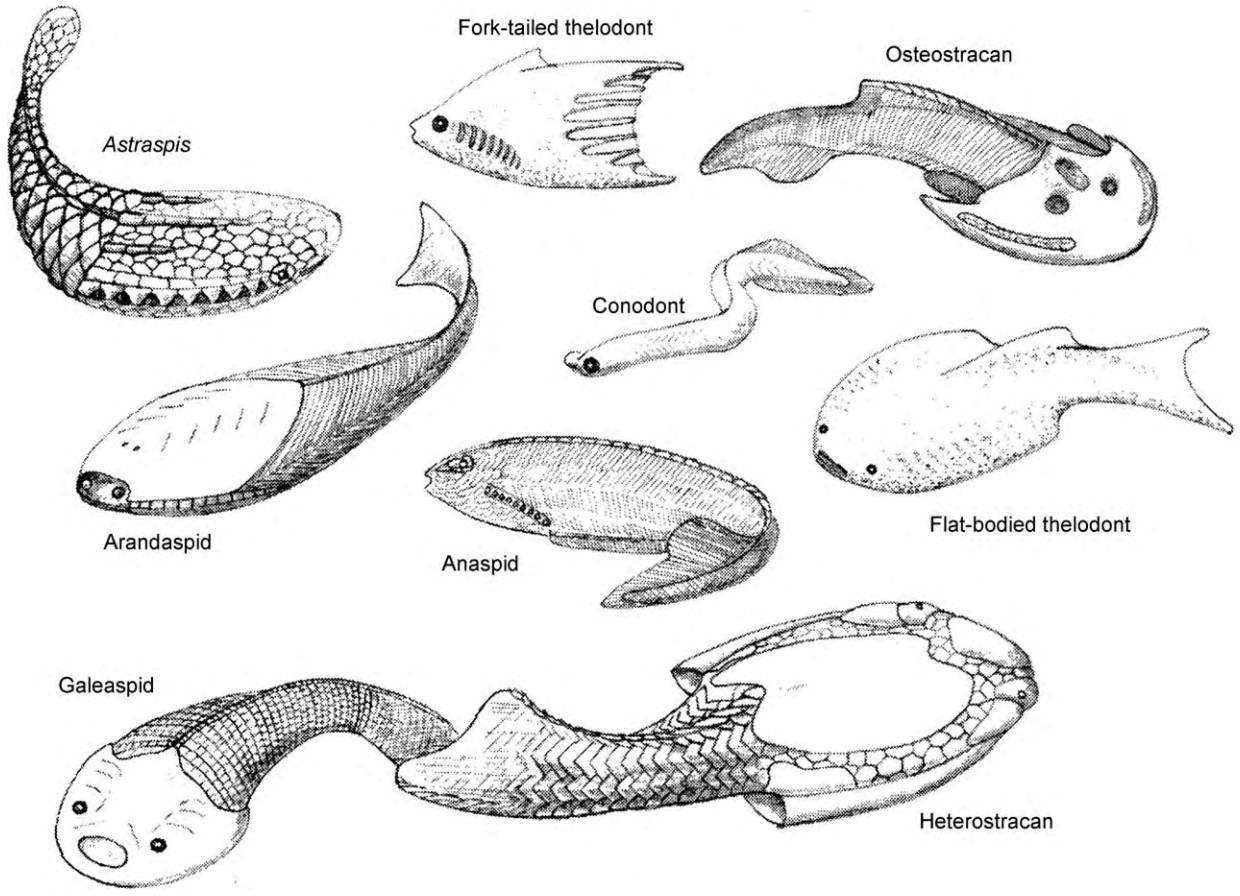


Figure 1 Reconstructions of some of the main groups of jawless vertebrates. (Based partly on Elliott, 1987 and Janvier, 1998).

Ordovician Ostracoderms

The North American fishes *Astraspis* and *Eriptychius* were long known as the earliest fossil fish, having been initially described by Charles Walcott in the late nineteenth century. *Astraspis* (Figures 1–3) is known from a number of localities flanking the Ordovician Transcontinental arch and was covered by tesserae (small bony plates) with a three-layered structure of honeycombed sheets of aspidin covered by tubercles formed from dentine with an enameloid cap. There were eight pairs of gill openings behind the laterally positioned eyes, no paired or dorsal fins, and a tail covered by large overlapping scales. The structure of the anterior of the headshield and of the tail termination are currently unknown. *Eriptychius* is known only from isolated tesserae and a single partially articulated head, which demonstrates the presence of an endoskeleton formed of globular calcified cartilage. There is little information on *Eriptychius* and its affinities have yet to be resolved.

The arandaspids are known from the articulated remains of two genera, *Arandaspis* (Australia) and *Sacabambaspis* (Bolivia) (Figure 1). Both animals have headshields formed from roughly oval dorsal and ventral plates ornamented with oak-leaf or teardrop shaped tubercles. The gill openings are numerous and form a sloping row, covered by square branchial plates, between the two main plates. The eyes are anterior and paired pineal and parapineal openings are present, a feature unique among vertebrates. The trunk scales are narrow and elongated. Although several other genera are known from scales or tesserae from the Early Ordovician of central Australia their affinities are currently unknown.

Silurian and Devonian Ostracoderms

The ostracoderms flourished during the Late Silurian and Devonian during which period they were found worldwide. The osteostracans (Figure 4) were a successful group of ostracoderms that lived in North America, Europe, Siberia, and central Asia. A semicircular headshield encased the dorsal surface of the skull while the undersurface was covered with small tesserae that surrounded the gill openings and mouth. The brain was surrounded by perichondral bone (present in many gnathostomes but not lampreys and hagfishes) and its details are thus well known. The eyes were situated on top of the head and were separated by openings for the pineal and nasohypophysial openings. Central and lateral 'sensory fields' on the head are characteristic of this group and were probably a specialised part of the lateral line system that enables fishes to detect vibrations in the water. Paired pectoral fins and a flexible tail made these ostracoderms very maneuverable.

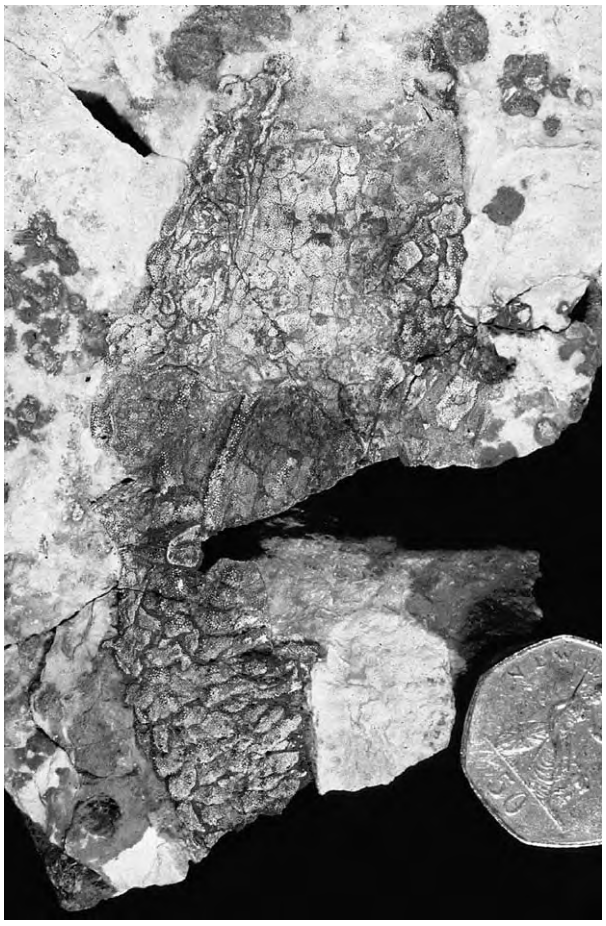


Figure 2 An articulated specimen of *Astraspis desiderata* from the Middle Ordovician Harding Sandstone of Colorado. The animal is exposed on the dorsal surface and the orbits can be seen on either side at the top of the preserved part of the specimen. The anterior part of the head is lost to erosion and the tail is also missing. The specimen is 120 mm long.

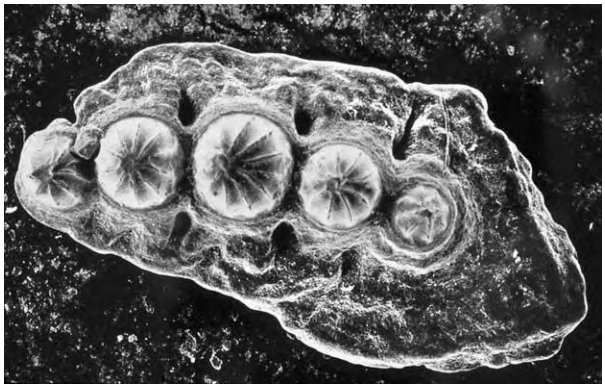


Figure 3 SEM photograph of a single tessera (4 mm long) from the headshield of *Astraspis desiderata* showing the stellate tubercles that give it its name.

Figure 4 Lateral view of the headshield of the Early Devonian osteostracan *Cephalaspis utahensis* showing the location of the orbits and the specialized sensory areas. The headshield is 130 mm long.

Figure 5 Articulated specimen of the Early Devonian hetero stracan *Errivaspis waynensis* in ventral view. the oral area is formed of finger like oral plates whose function is currently unknown. The animal is 140 mm long.

Figure 6 A dorsal shield of the Early Devonian heterostracan *Tuberculaspis elyensis* showing the single pair of branchial ducts opening posteriorly and the posterior dorsal spine. The shield is 65 mm long.

A related group of ostracoderms, the galeaspids (Figure 1), inhabited parts of China and North Vietnam during the Late Silurian and Devonian. Although similar to osteostracans in general appearance they differ in having a central opening in the headshield that was used to take water into the gill chamber. Galeaspids also lacked paired fins.

A small and enigmatic group of agnathans from the Middle Devonian of Australia have been named the pituriaspids ('hallucinatory shield'). These animals had an elongated headshield with small, paired areas similar to the sensory areas in osteostracans. They also appear to have had paired pectoral fins that projected through openings in the headshield. Their relationship to other agnathans is unclear, although they are generally placed close to osteostracans and galeaspids.

Anaspids lived at the same time as the osteostracans and galeaspids in North America and Europe, and had fusiform bodies with downturned tails (Figure 1). They may have had ribbon-like paired fins behind the row of gill openings that allowed them to manoeuvre slowly forward and back. Some anaspids were scaled while others lacked scales and the latter type sometimes had circular cartilage surrounding the mouth together with large eyes and a basket-like branchial skeleton. Although these features have been cited in the past as suggesting a relationship with lampreys they are now generally considered to be convergences.

Heterostracans were among the first of the armoured agnathans to be discovered although it was not until 1858 that Thomas Huxley recognised them to be vertebrates. They are a group of armoured agnathans that are characterised by the presence of a common external branchial opening on either side of the head (Figures 1, 5 and 6). The armour consists of apidin with a well-developed honeycomb structure and a surface of dentine ridges or tubercles. The earliest forms are Early Silurian in age and the youngest are found in the Late Devonian (Late Frasnian) where they reached 1.5 meters in length making

them the largest known agnathans. Heterostracans are usually found as disarticulated plates and scales but some complete animals are known from the Late Silurian and Early Devonian of northern Canada and from the Early Devonian of Germany and the Welsh Borders. None of these specimens shows any trace of the endoskeleton beyond impressions on the internal surface of some of the dermal plates and it is assumed to have been cartilaginous, however, the impressions have enabled some reconstruction of the internal anatomy to be made. Because of this lack of internal information the arrangement and structure of the dermal armour is the main feature used in discussions of classification, phylogeny, or affinities.

The head armour of heterostracans consists of a series of large plates that also contain the canals of the sensory canal system. These plates are almost always arranged to form a dorsal and a ventral shield. The dorsal shield generally consists of a series of median and lateral plates while the ventral shield is generally a single plate. There are variations to this arrangement as some heterostracans have a ventral shield composed of small platelets while other forms have a headshield in which all the plates are fused into a continuous capsule. The body is covered by large scales which become smaller on to the caudal fin where they are arranged in radial rows. They seem to have been animals that lived in brackish and freshwater environments perhaps feeding on algal layers or microorganisms suspended in the water column.

Thelodonts (Figure 1) are an interesting group of agnathans that occur from the Ordovician through the Devonian worldwide. They are rarely found complete but the tiny scales that covered their bodies (Figure 7) can be extremely numerous and are valuable tools in biostratigraphic correlation. Rare complete specimens show that most were flattened fish with slanting gill openings and broad pectoral areas. However, new material from the Mackenzie Mountains in Canada shows that some were laterally compressed with large forked tails and small triangular dorsal fins. These Furcacaudiformes have a recognisable stomach which is an organ previously thought to be absent in jawless fishes as it is lacking in the living forms. The structure of thelodont scales is in some cases very similar to the placoid scales of sharks and this together with the presence of a stomach has given rise to the suggestion that some thelodonts may be closely related to gnathostomes.

Figure 7 Scales of thelodont *Thelodus macintoshi* in cross section and showing the central pulp cavity surrounded by lam inar aspidin and capped by a clear crown of dentine. The large central scale is about 1 mm across.

The Last Ostracoderms

Although lampreys and hagfishes have a very poor fossil record, they are known from the Pennsylvanian and thus must have continued through the Palaeozoic and Mesozoic to the present. The armoured ostracoderms, however, had almost all become extinct by the start of the Late Devonian. At that point one osteostracan and three anaspids are known from Canada, one galeaspid from northern China, a single Australian thelodont, and some of the last heterostracans from Europe. The decline is attributed mostly to the rise of the gnathostomes as all the major groups of jawed fishes evolved during the Devonian and by the Middle Devonian many had reached a peak of diversity. It is clear that in many cases the niches once inhabited by agnathans were taken over by gnathostomes as they succumbed to increasing predation pressure.

Agnathan Relationships

The enormous range of features in the agnathans makes it difficult to sort out the relationships between them. Because of this a variety of schemes have been proposed, starting with the idea that the lampreys and hagfishes were most closely related to each other (the 'cyclostomes') because both were scale-less and elongated and had pouch-like gills and median nostrils. By the 1970s, however, it became clear that lampreys shared a suite of characters with gnathostomes including the presence of neural arches along the notochord, large eyes with associated muscles, nervous control of the heart, osmotic regulation of body fluids, and a brain with a cerebellum and other advanced features. It was proposed at this time that lampreys were more closely related to the gnathostomes than either was to the hagfishes. Although molecular data has now been brought to bear on this problem, it only supports the close relationship

of lampreys and gnathostomes if the cephalochordates are used as the group for comparison. If the tunicates are used then the cyclostome relationship is supported. These results suggest that more molecular analyses are necessary to clarify the historical message. Fossil lampreys and hagfishes do not help with this problem as they are virtually similar to the modern forms.

Analysis of the fossil agnathans has depended very much on the methodology termed 'phylogenetic systematics' or 'cladistics'. (*see* **Palaeontology**) This method was first introduced in the 1960s by the German entomologist Willi Hennig and has been improved most recently by the introduction of computer algorithms. The goal of cladistics is to recognise sister-group relationships as in the statement that taxon A is more closely related to taxon B than to taxon C. These relationships are recognised by the presence of derived (advanced) characters that are unique to the sister-group pair and so are assumed to have arisen in the most recent common ancestor of both. Primitive characters present in taxa other than the sister-group are thus not relevant to the relationship. Analysis of the distribution of characters results in groupings that are most parsimonious, i.e., groups in which the number of changes is kept to a minimum.

An analysis of features within modern and fossil agnathans confirms that among living forms lampreys and gnathostomes share a sister-group relationship (Figure 8). The hagfishes are the most primitive known vertebrates and therefore many hagfish characters must indicate the primitive condition including the lensless eye, simple ear, lack of nervous regulation of the heart, the absence of lateral line system, electroreceptors, and cerebellum, and an inability to regulate internal ionic concentrations. Among the ostracoderms, the osteostracans are most closely related to the gnathostomes while conodonts and anaspids are the most distantly related. That the dermal skeleton

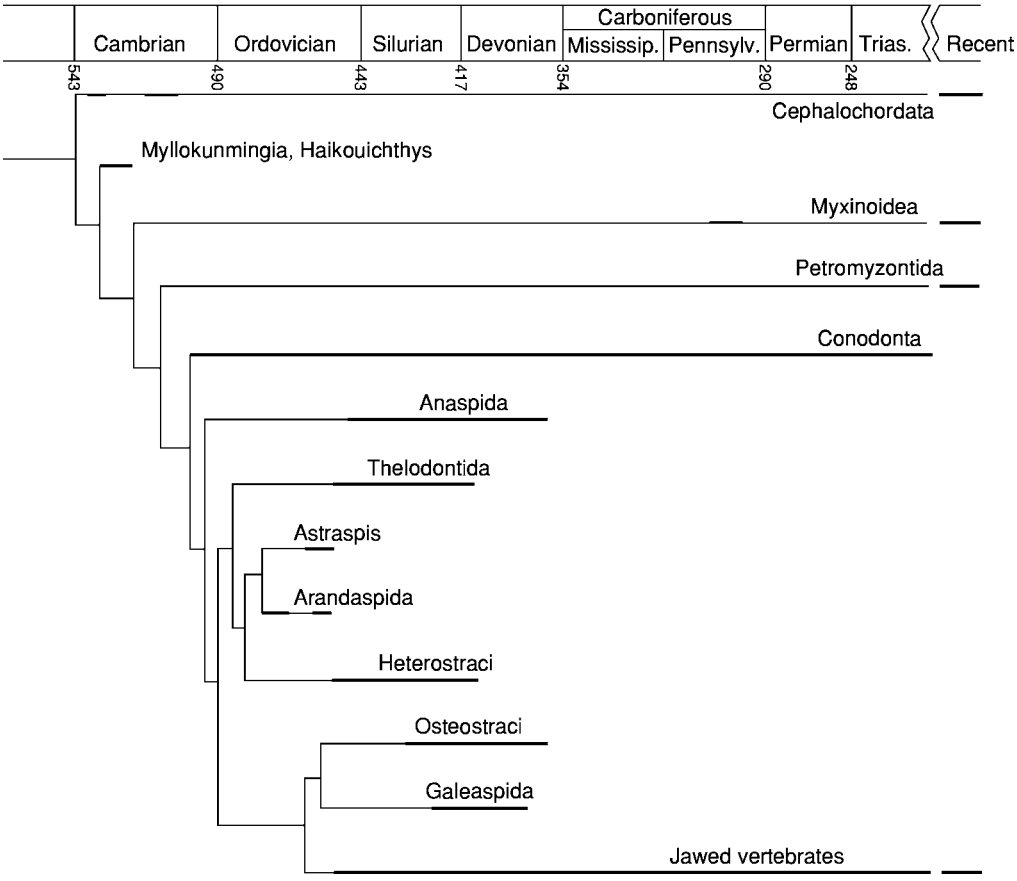


Figure 8 The evolutionary relationships of jawless vertebrates are here superimposed on a stratigraphic scheme. Broad lines indicate the fossil record of each lineage. The relationships of the Conodonta and of the Cambrian genera *Myllokunmingia* and *Haikouichthys* are a matter of some debate at present. Although the earliest accepted gnathostomes are Silurian acanthodians there are a number of scales of chondrichthyan, acanthodian, and placoderm type recognized from the Harding Sandstone. (Compiled from Donoghue *et al.*, 2000; Forey, 1995; Janvier, 2003).

arose later is suggested by the fact that all ostracoderms seem to be more derived than lampreys or hagfishes. Initially it would have consisted of dentine and acellular bone (aspidin) with cellular bone appearing later. A calcified internal skeleton also appeared in the ancestors of gnathostomes after the divergence of lampreys from hagfishes. Paired fins, characteristic of gnathostomes, were presumably present in the ancestors of osteostracans and possibly pituriaspids as they were the only other fossil group that might have possessed them.

Following this analysis of relationships it is possible to reconstruct the evolution of some of the vertebrate characters. An example is the lateral line system, electroreceptors, and the labyrinth of canals in the inner ear used in balance. Hagfishes lack both electroreceptors and neuromasts (lateral line cells) while the lampreys have neuromasts in short lines on the head and electroreceptors scattered over the head and body. In the extinct arandaspids and *Astraspis* from the Ordovician grooves in the surface of the headshields are thought to have held neuromasts while the heterostracans, thelodonts, osteostracans, and galeaspids have a lateral line system that consists of a complex of canals that runs within the plates and scales and opens to the surface through pores. This system is similar to that found in gnathostomes and so development seems to have been from isolated neuromasts to neuromasts in grooves on the head and body to neuromasts embedded in the plates and scales. Similarly the lampreys, osteostracans, and galeaspids have two vertical canals in the inner ear, as do the heterostracans, while the gnathostome ear has two vertical semicircular canals together with a horizontal canal. Unfortunately it is unknown what the brains of arandaspids, thelodonts, and anaspids were like.

Many questions concerning the early development of vertebrate structures still remain unanswered but it is clear that the ostracoderms have provided information unobtainable in any other way. New information will certainly be gained as further searches are made in the Cambrian and Ordovician rocks that are now yielding early vertebrates and perhaps into rocks below the Cambrian that may contain even more ancient forms.

See Also

Fossil Vertebrates: Fish; Palaeozoic Non-Amniote Tetrapods. **Microfossils:** Conodonts. **Micropalaeontological Techniques**. **Palaeoecology**. **Palaeontology**. **Palaeozoic:** Cambrian; Ordovician; Silurian; Devonian. **Stratigraphical Principles**. **Time Scale**.

Further Reading

Blieck A, Elliott DK, and Gagnier P Y (1991) Some questions concerning the phylogenetic relationships of the heterostracans, Ordovician to Devonian jawless vertebrates. In: Chang MM, Liu YH, and Zhang GR (eds.) *Early Vertebrates and Related Problems of Evolutionary Biology*, pp. 1 17. Beijing: Science Press.

Donoghue PCJ, Forey PL, and Aldridge RJ (2000) Conodont affinity and chordate phylogeny. *Biological Review* 75: 191 251.

Elliott DK (1987) A reassessment of *Astraspis desiderata*, the oldest North American vertebrate. *Science* 237: 190 192.

Elliott DK, Blieck A, and Gagnier P Y (1991) Ordovician Vertebrates. In: Barnes CR and Williams SH (eds.) *Advances in Ordovician Geology*, Geological Survey of Canada Papers 90, pp. 93 106. Canada: Ottawa.

Forey PL (1995) Agnathans recent and fossil, and the origin of jawed vertebrates. *Reviews in Fish Biology and Fisheries* 5: 267 303.

Janvier P (1998) *Early Vertebrates*. Oxford: Clarendon Press.

Janvier P (2003) Vertebrate characters and the Cambrian vertebrates. *C. R. Paleovol* 2: 1523 1531.

Long JA (1993) *Palaeozoic Vertebrate Biostratigraphy and Biogeography*. London: Belhaven Press.

Sansom IJ and Elliott DK (2002) A thelodont from the Ordovician of Canada. *Journal of Vertebrate Paleontology* 22: 867 870.

Sansom IJ, Smith MP, Smith MM, and Turner P (1997) *Astraspis* the anatomy and histology of an Ordovician fish. *Palaeontology* 40: 625 643.

Sansom IJ, Smith MM, and Smith PM (2001) The Cambrian origin of vertebrates. In: Ahlberg PE (ed.) *Major Events in Early Vertebrate Evolution*, Systematics Association Special Volume Series 61, pp. 67 84.

Sansom IJ, Smith MM, and Smith PM (2001) The Ordovician radiation of vertebrates. In: Ahlberg PE (ed.) *Major Events in Early Vertebrate Evolution*, Systematics Association Special Volume Series 61, pp. 156 171.

Fish

J A Long, The Western Australian Museum, Perth, WA, Australia

Introduction

Fishes, loosely defined as free-swimming aquatic vertebrates, have a fossil record spanning some 540 million years (Figure 1). The fossil remains of early fishes are useful for elucidating the major anatomical stages of vertebrate evolution, from the origins of bone, jaws, and teeth through to the development of powerful limbs leading to the tetrapod invasion of land. Some Palaeozoic fish remains are biostratigraphically useful, and in some instances can provide valuable palaeoecological information.

The earliest fishes are boneless forms like *Haikouichthys* and *Myllokungmingia* from the Early Cambrian Chengjiang fauna, Yunnan, China. *Haikouichthys* is regarded as the first fish as it has a fusiform body with notochord, gill slits, V-shaped muscle bands, and fin-rays supporting a median fin and tail fin. These are now recognized as the first true fishes even though they lack bone.

The evolutionary development of organically mineralized tissue which supports and protects the soft tissues and allows for more efficient attachment of muscles marks the beginning of a great diversity of the first jawless fishes, the agnathans.

The oldest vertebrates remains of scales composed of bone-like tissues come from the Late Cambrian Georgina Basin of central Queensland, Australia. Early Ordovician fish remains from central Australia include several forms known from scales, such as *Areyongia* and *Apedolepis*. The first well-preserved impressions of complete jawless fish armours occur in the Middle Ordovician of Australia. *Arandaspis* (Figure 2) had well-developed dorsal and ventral bony shields, with distinct rows of squarish branchial plates over the gills. It was a precursor to the more successful heterostracan fishes, such as *Pteraspis*, which had a single branchial plate covering the gills. Such fishes were prevalent in Silurian and Early Devonian seas, and to a lesser extent, freshwater habitats, largely occupying the niches of benthic detrital and filter feeders. The bone of heterostracans lacked bone cells, and has been termed 'aspidin'.

Agnathan Diversity

Many kinds of jawless fishes evolved throughout the Silurian and Devonian. These include armoured forms with a scoop-shaped solid bony shield such as the osteostracans (e.g., *Cephalaspis*), the naked lamprey-like anaspids (e.g., *Jaymoytius*), and the heavily-scaled thelodonts (e.g., *Thelodus* (Figure 2)). Finds from Canada indicate that some thelodonts, the Furcacaudiformes, had laterally compressed bodies with well-developed stomachs. Thelodonts, in general, had an internal soft anatomy (evidenced by impressions in *Thelodus*) suggesting close affinity to jawed fishes (gnathostomes). All early agnathans had well-developed sensory-line systems, which are seen as lines of pores or grooves in the dermal bones. Osteostracans possessed complex fields of sensory organs along the sides and top of their shields.

Osteostracan fishes appear to be the closest group to the gnathostomes because they share perichondral bone enveloping the braincase, a large median head vein, and true bone with cell sites. The head shields of osteostracan fishes indicates that the first gill arches were well forward of the eyes, immediately above the mouth. Osteostracans presumably had cartilaginous gill arch supports, as shown by impressions of gill structures on the visceral surface of their shield. Primitive agnathans, like heterostracans, lacked pectoral and pelvic fins. These are found in anaspids as extensions of the fin-folds, but only osteostracans had muscular pectoral fins which were internally attached to a cartilaginous ossification, the scapulocoracoid. A recent observation that osteostracans, acanthodians, and possibly chondrichthyans all had similar inner ear physiology with open endolymphatic ducts further reinforces a close phylogenetic relationship between osteostracans and gnathotomes.

The Appearance of Jaws

Primitive jawless fishes had many more paired gill arches than the jawed fishes; some like the anaspid *Legendrelepis*, possessed twenty or more paired gill pouches. Jaws may have first originated in fishes by modification of the front gill arch support bones. The first primitive set of jaws and teeth could have formed when dermal scales, with a tooth-like structure and shape, invaded the pharynx and developed as mucous membrane denticles. Some agnathans had tooth-like scales lining the buccal cavity (e.g., thelodonts) which could have acted as 'prototeeth' in the reduction of food. Furthermore, the dermal scales of most jawless fishes are constructed of dentinous tissues underlying the enameloid crown, with a bony base. Thelodonts even have a pulp cavity in the base of the scales,

making these scales resemble 'primitive teeth' in terms of their histology.

The appearance of jaws should naturally herald the abundant appearance of teeth in the fossil record at the same time. Gnathostomes (jawed vertebrates) include the first fishes with jaws and teeth: chondrichthyans, acanthodians, placoderms, and osteichthyans, and their evolutionary descendants, the tetrapods (amphibians, reptiles, birds, mammals). Sharks, for example, have hundreds of teeth in their mouths which they grow and shed continuously through life. The average modern shark may shed up to 20 000 teeth into the sediment, although Palaeozoic sharks may not necessarily have shed as many teeth as modern forms as they often exhibit fewer tooth rows. Shark skin contains thousands of tiny placoid scales which are also shed into the sediment after death, and readily identify the fossils.

There has been much recent debate over the appearance of true teeth in certain placoderms, mainly as some workers see 'teeth' as having arisen from patterns of dermal denticles whereas others think that placoderm teeth arose from dental lamina, and are not related to the patterning of dermal denticles seen on the plates.

Early Sharks and Their Relatives

The oldest shark-like scales come from the Late Ordovician Harding Sandstone of North America and the Early Silurian of Mongolia, yet the oldest shark teeth are of lowest Devonian age, some 30 million years later. As the numbers of teeth and scales are both high per individual shark, the absence of shark teeth in Silurian strata would seem to be a real observation, suggesting that the first sharks had scales of recognizable morphology, akin to those of modern sharks, yet they probably lacked teeth. The appearance of the first teeth most likely correlates with the evolution of jaws in sharks.

The most complete early shark is *Doliodus problematicus*, from the Early Emsian (*ca*. 400 mya) of New Brunswick, Canada. It shows tooth families *in situ*, a typical early shark braincase and the pectoral fins have paired spines preceding them, a feature previously only seen in acanthodians, placoderms, and the basal osteichthyan *Psarolepis*. This implies that paired pectoral spines may be a primitive character of all jawed fishes.

Sharks possibly first arose by the Early Devonian in Laurentia (*Doliodus*), and could have spread via

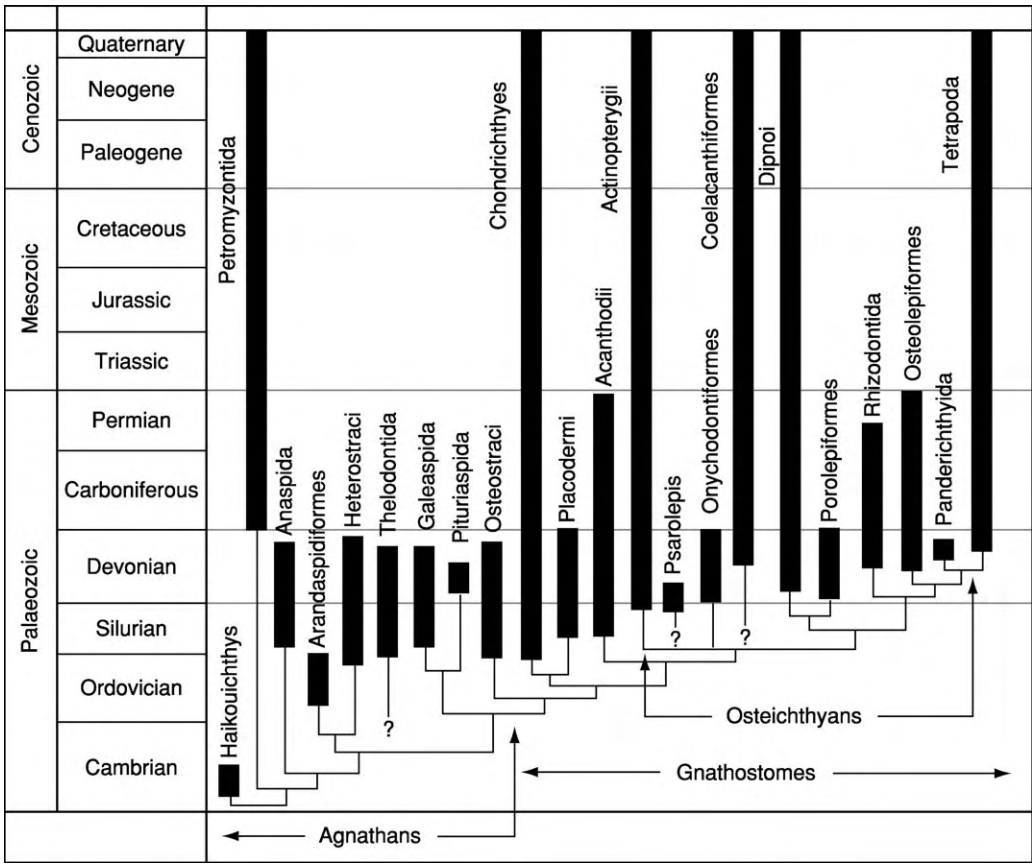


Figure 1 Major radiations of fish groups through time.

shallow seaways into northern Gondwana, as their remains are commonly found in late Early Devonian and Middle Devonian deposits in Spain, Australia, Antarctica, South Africa, and South America. The Middle Devonian Aztec Siltstone of Antarctica has an unusually high diversity of sharks within one stratigraphic unit in both species' numbers and the size of teeth, suggesting that Gondwana may well have been the place where toothed sharks underwent their first major radiation. By the Middle Devonian, sharks teeth are commonly found throughout the world, and by the Late Devonian more than forty species are known. The first of the modern sharks, or neoselachians, may date back as far as the Middle Devonian, represented by teeth of *Mcmurdodus* which have multilayered enameloid. *Hamiltonichthys*, from the Lower Carboniferous, was the oldest of the hybodontid shark lineage, a group prevalent in the Mesozoic Era. In the Lower Carboniferous some very large sharks had evolved, such as *Edestus giganteus*, from North America, which possessed serrated teeth up to 7–8 cm high on a continuous whorl. Such mega-predators would have reached 6 metres or more in length. Another major chondrichthyan group, the holocephalomorphs, represented today by the chimaerids and rabbitfishes, appeared in the Early

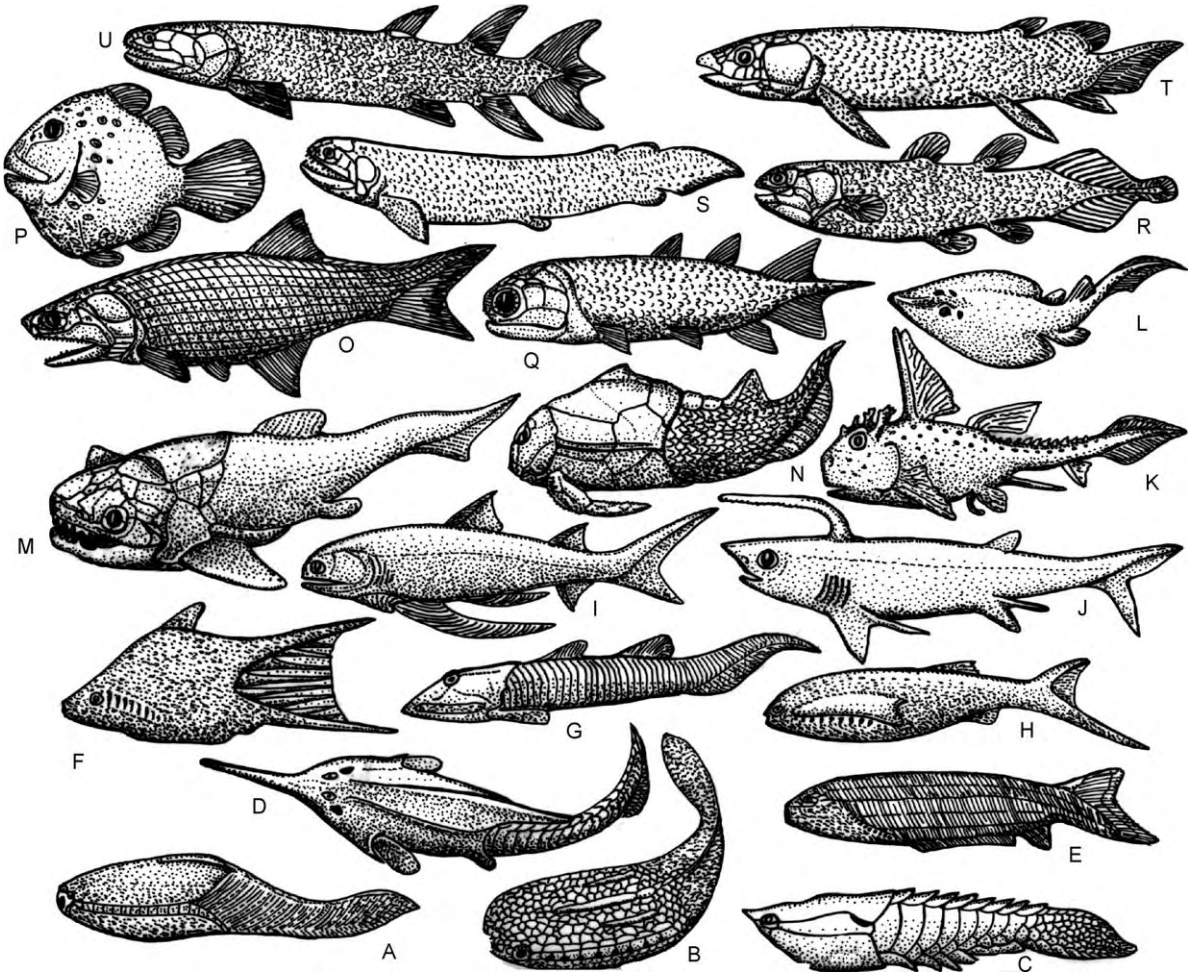


Figure 2 Diversity of major fish groups through time. (A H, agnathans; I U, gnathostomes). A, *Arandaspis* (Arandaspidiformes; Ordovician). B, *Eriptychius*, (Astraspidiformes; Ordovician). C, *Anglaspis* (Heterostraci; Silurian). D, *Pituriaspis* (representing Pituriaspida and Galeaspida; Devonian). E, *Birkenia* (Anaspida; Silurian). F, *Furcacauda* (Furcacaudiformes; Devonian). G, *Ateleaspis* (Osteostraci; Silurian). H, *Thelodus* (Thelodontida; Silurian). I, *Gyracanthides* (Acanthodii, Climatiiformes; Devonian Carboniferous). J, *Falcatus* (Chondrichthyes, Selachii; Carboniferous). K, *Echinochimaera* (Chondrichthyes, Holocephalomorphi; Carboniferous). L, *Rhinobatis* (Chondrichthyes, Selachii, Batoidea; Jurassic Recent). M, *Mcnamaraspis* (Placodermi, Arthrodira; Devonian). N, *Sherbonaspis* (Placo dermi, Antiarchi; Devonian). O, *Mimia* (Osteichthyes, basal Actinopterygii; Devonian). P, *Eoplectus* (Osteichthyes, Actinopterygii, Teleostei; Eocene). Q, *Strunius* (Osteichthyes, Onychodontiformes; Devonian). R, *Hadronector* (Osteichthyes, Coelacanthiformes; Carboniferous). S, *Strepsodus* (Osteichthyes, Rhizodontiformes; Carboniferous). T, *Dipterus* (Osteichthyes, Dipnoi; Devonian). U, *Eusthenopteron* (Osteichthyes, Osteolepiformes; Devonian).

Carboniferous (Figure 2). Some sharks of this time, like the petalodonts, had unusual crushing or nipping teeth, possessing few sets in the jaws.

Sharks radiated and diversified throughout the Late Palaeozoic and Mesozoic. The largest Palaeozoic predatory sharks may have been the edestids, such as *Helicoprion*, whose coiled tooth whorls occur in the Middle Permian. If these large whorls sat on the tip of the lower jaws, estimated sizes of 10 metres would not be unlikely for these sharks. Huge lamnid sharks evolved in the mid-Tertiary, such as *Carcharocles megalodon* with teeth 18 cm high, suggesting a total body length of about 15 metres.

The rays originated from benthic sharks back in the Late Jurassic and by the end of the Mesozoic were a widespread and diverse group.

Acanthodians

The oldest fishes with jaws and teeth preserved intact are the acanthodians, which date back to the Early Silurian. They are often represented in microscopic residues from dissolved limestone as isolated teeth, scales, and fin spines dating back to the Early Silurian. Acanthodians have been likened to 'spiny sharks' in past literature although their affinities appear to lie closer to the higher jawed fishes, such as osteichthyans (true bony fishes) because of scale structure and gross morphology of their braincase. The oldest acanthodians include forms with distinct gnathal bones on to which strong teeth are ankylozed. This group, the ischnacanthids, were moderately large predators in the Early Devonian seas; like *Xylacanthus grandis* from Spitsbergen, reached sizes of 2–3 metres. The climatiforms (e.g., *Climatius* *Gyracanthides* (Figure 2)) had elaborate dermal shoulder girdle armour and some had many additional spines along the ventrolateral ridge of the body. Acanthodiforms, such as *Acanthodes*, were filter-feeding forms which lacked shoulder girdle armour and had extensive gill rakers for sifting food, plus possessed otoliths (statoconia, ear stones) presumably for refining balance and orientation during swimming manoeuvres. The group survived until the end of the Permian Period.

Placoderms

The most successful of all the Middle Palaeozoic jawed fishes were undoubtedly the placoderms, which appeared in the Early Silurian and reached a peak of diversity during the Middle–Late Devonian. Their name means 'plated skin', referring to their characteristic mosaic of overlapping dermal plates that enveloped the head and trunk region (Figure 3). They possessed shark-like heterocercal tails and in their overall anatomy were very much akin to chondrichthyans. Placoderms have alternatively been placed as possible relatives of the bony fishes (osteichthyans) because of their dermal bones and internally ossified braincases. However, the teeth which developed on placoderm jaw bones were not real teeth with roots set into a discrete jaw bone, like those of the osteichthyan fishes, but well-developed pointed cusps that protruded from the jaw bone itself, with semidentine present. Placoderms exhibit a multitude of dentition types, revolving around rows of pointed cusps on gnathal bones, or clusters of small denticles or pointed cusps on sheets of bone. As discussed above, there is ongoing debate as to whether

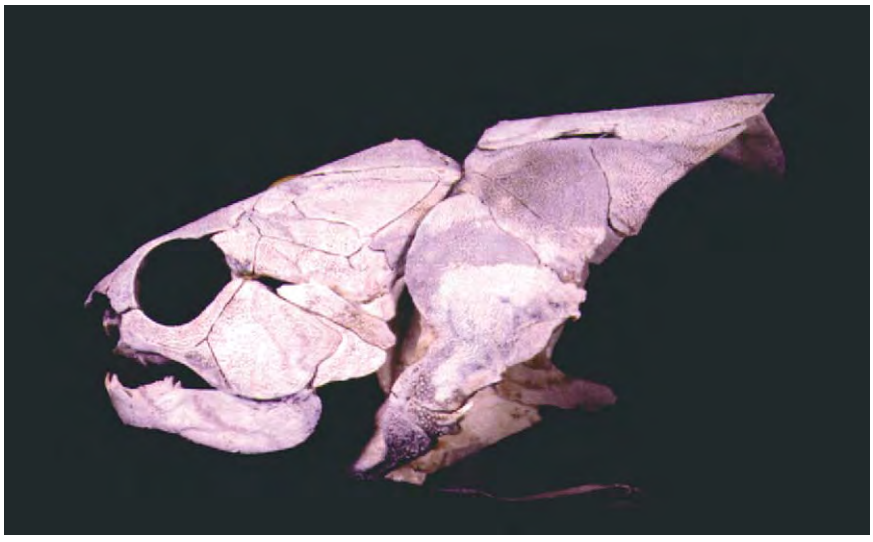


Figure 3 The armour of a Late Devonian placoderm fish *Mcnamaraspis kaprios*, from the Gogo Formation of Western Australia.

placoderm have true teeth or whether they have 'teeth' derived from dermal denticles.

The largest group of placoderms were the arthrodires, which includes the biggest forms, the dinichthyids (e.g., *Gorgonosteus*) reaching 6 metres or more. The small antiarchs had external bone-covered props for pectoral fins (e.g., sherbonaspis, Figure 2). One of these, *Bothriolepis*, was ubiquitous in the Middle–Late Devonian, and is represented by over one hundred species worldwide. Other groups include the flattened ray-like rhenanids and their heavily armoured relatives, the acanthothoracids, the petalichthyids, and the flattened phyllolepids. Placoderms can often be identified from isolated remains because of their unusual plate shapes and dermal ornamentation, making them useful in Devonian biostratigraphic and biogeographic studies. Placoderms inhabited both marine and freshwater environments, with some genera found in both fluvial and marine deposits (e.g., *Bothriolepis, Remigolepis, Groenlandaspis*). The placoderms became extinct at the end of the Devonian Period, possibly being displaced by the steadily increasing radiation of chondrichthyans.

Osteichthyans

The osteichthyans (true bony fishes) appeared by the Late Silurian, represented by isolated scales, teeth, and rare isolated bones of small ray-finned and lobe-finned fishes (Figure 2). The Osteichthyes comprise the ray-fins (Actinopterygii) which are more than 99% of all living fish species and are represented by over 23 000 species (e.g., trout, salmon, goldfish); and the Sarcopterygii, which have muscular, fleshy lobed fins and include the lungfishes (Dipnoi), and what was commonly grouped as the 'crossopterygians' (now recognized as a paraphyletic mixture of groups). In palaeontological terms, this group is better refered to by Order, and includes mostly extinct groups apart from one surviving species, the coelacanth *Latimeria chalumnae* (within the Order Actinistia or Coelacanthiformes). The many extinct groups of non-dipnoan sarcopterygians that diversified during the Middle Palaeozoic include the orders Onychodontiformes, Porolepiformes, Rhizodontiformes and Osteolepiformes.

Actinopterygians

The ray-finned fishes were represented in the Devonian by primitive palaeoniscoids such as *Cheirolepis* and *Moythomasia*. The group radiated rapidly by the beginning of the Carboniferous, and over fifty families are known during the Late Palaeozoic, typified by forms such as *Palaeoniscus, Elonichthys*, and *Gonatodus*. Most of the heavily scaled primitive forms gave way to those having thinner round scales, with less rigid cheek and jaw bone arrangements; this enabled more efficient feeding and respiratory mechanisms, such as the buccal-pump system, to evolve. This line of evolution saw the rise of holostean

Figure 4 *Leptolepis koonwarri*, an Early Cretaceous teleost fish from Australia.

and subholostean fishes in the Late Palaeozoic (today represented by gars and bowfins, e.g., *Amia, Lepisosteus*) and the first of the modern ray-fins, the teleosteans, by the Late Triassic. Teleosteans have advanced tail fin skeletons with many supporting bones, plus mobile cheeks and detached jaw bones. Their first gill arch (hyoid), which supports the jaws, is often vertically or forwardly inclined. Many of the modern families of fishes had appeared by the close of the Cretaceous Period (Figure 4), and the first teleosteans adapted specifically to life on reefs had appeared by the Eocene Period, represented by species from Monte Bolca, Italy (e.g., Eoplectus, Figure 2).

Sarcopterygians

In the Devonian Period, the most diverse group of bony fishes were the sarcopterygians (over one hundred species); the ray-fins were only a minor component of fish faunas (about ten species). The earliest sarcopterygians, such as *Psarolepis* from China had a shiny tissue, cosmine, that covered the dermal bones and scales and was lost in more advanced lineages. Cosmine may have functioned as an electrosensory tissue.

One of the basal groups of sarcopterygians is the dagger-toothed fishes or Onychodontiformes. These are a poorly known group whose best representative is *Onychodus* from the Late Devonian Gogo Formation of Western Australia. *Onychodus* is known from isolated bones and tooth whorls in the Middle–Late Devonian worldwide, and was the largest of all Devonian osteichthyans, *O. sigmoides*, reaching estimated lengths of 4 metres.

The ability to gulp air and transgress environmental boundaries evolved within lungfishes during the Middle Devonian (e.g., *Dipterus*), becoming prominent in the Late Devonian. This gave them an edge on other groups, with later forms developing a further enhanced survival strategy, the ability to aestivate (e.g., *Gnathorhiza*). Lungfishes diversified into forms with crushing tooth plates (e.g., *Chirodipterus*) and others with denticulated palates (e.g., *Griphognathus*). Post-Devonian lungfishes were largely freshwater forms. Today there are three living lungfish genera, the most primitive of which is *Neoceratodus forsteri* from Australia, which dates back to the Cretaceous Period.

Some of the advanced osteolepiform fishes appear to have developed the ability to breathe air independently, as they had evolved a palatal nostril (a choana) by the Middle Devonian (e.g., *Gogonasus*). Osteolepidids were a common component of most Devonian fish faunas, in many cases taking the role of top predator. More advanced osteolepiforms, like *Eusthenopteron* (Figure 2), from the Late Devonian of Canada, lost the cosmine layer and had rounded scales. The largest of these predatory lobe-fins were the rhizodontiforms, which dominated the lakes and rivers of the Carboniferous Period. Forms like *Rhizodus* may have reached 6–7 metres in length.

The panderichthyid fishes are morphologically closest to the elginerpetonid amphibians, known from Scotland and Russia, in having identical skull-roof patterns, pectoral and pelvic girdles, and endoskeletons which approach the morphology of the basic tetrapod limbs more closely than any other fish. A primitive rhizodontid pectoral skeleton (*Sauripterus*), with eight fin radials supporting the pectoral fin, has been described from North America. This matches the eight digits occurring in the limbs of the Devonian amphibian *Acanthostega*. The first tetrapods evolved by the end of the Frasnian stage of the Late Devonian, and at least nine genera had appeared by the close of the Devonian Period.

See Also

Fossil Vertebrates: Jawless Fish-Like Vertebrates.

Further Reading

Campbell KSW and Barwick RE (1987) Palaeozoic lung fishes a review. *Journal of Morphology, Supplement* 1: 93 132.

Forey PL and Janvier P (1993) Agnathans and the origin of jawed vertebrates. *Nature* 361: 129 134.

Jarvik E (1980) *Basic structure and evolution of verte brates*. New York: Academic Press.

Long JA (1993) Cranial ribs and the origin of air breathing in lungfishes. *Memoirs of the Australisian Association of Palaeontologists* 15: 199 210.

Long JA (1995) *The rise of fishes 500 million years of evolution*. University of New South Wales: Sydney, Johns Hopkins University Press: Baltimore.

Long JA, Campbell KSW, and Barwick RE (1997) Oste ology and functional morphology of the osteolepiform fish *Gogonasus andrewsae* Long, 1985, from the Upper Devonian Gogo Formation, Western Australia. *Records of the Western Australian Museum, Supplement* 53: 1 89.

Miller RF, Cloutier R, and Turner S (2003) The oldest artriculated chondrichthyan from the Early Devonian Period. *Nature* 425: 501 504.

Sahney S and Wilson MVH (2001) Extrinsic labyrinth infillings imply open endolymphatic ducts in Lower Dev onian osteostracans, acanthodians, and putative chon drichthyans. *Journal of Vertebrate Palaeontology* 21: 660 669.

Shu D G (2003) A palaeontological perspective of vertebrate origin. *Chinese Science Bulletin 2003,* 48: 725 735.

Shu D G, Conway Morris S, Han J, *et al.* (2003) Head and backbone of the Early Cambrian vertebrate *Haikouichthys*. *Nature* 421: 526 529.

Vorobyeva E and Schultze H P (1996) Description and systematics of panderichthyid fishes with comments on their relationship to tetrapods. In: Schultze H P and Trueb L (eds.) *Origins of the higher groups of tetrapods: controversy and consensus*, pp. 68 109. Cornell: Cornell University Press.

Wilson MVH and Caldwell MW (1998) The Furcacaudi formes: a new order of jawless vertebrates with thelodont scales, based on articulated Silurian and Devonian fossils from northern Canada. *Journal of Vertebrate Palaeon tology* 18: 10 29.

Young GC (1997) Ordovician microvertebrate remains from the Amadeus Basin, central Australia. *Journal of Vertebrate Paleontology* 17: 1 25.

Young GC, Karatajute Talimaa VN, and Smith MM (1996) A possible Late Cambrian vertebrate from Australia. *Nature* 383: 810 812.

Zhu M, Yu X B, and Janvier P (1999) A primitive fossil fish sheds light on the origin of bony fishes. *Nature* 397: 607 610.

Palaeozoic Non-Amniote Tetrapods

J A Clack, University of Cambridge, Cambridge, UK

Introduction

The Late Palaeozoic saw the emergence of land vertebrates from their aquatic ancestors. Vertebrates with legs and digits are usually referred to as 'tetrapods', meaning 'four-legged'. In fact, the term not only refers to ancient extinct forms, but also embraces all modern forms which either possess legs (or arms, wings, or flippers) or whose ancestors possessed them (such as snakes and others that have lost them). This great 'family tree' of animals has its roots in the Late Palaeozoic, specifically the Late Devonian, when the first legged vertebrates evolved. Because of the nature of evolution, it is, in practice, sometimes difficult to draw the distinction between a 'fish' and a 'tetrapod', especially when there are no limbs or fins preserved in the fossils we have; therefore, to understand the emergence of tetrapods, it is also necessary to look at the 'fish' relatives from whom the 'tetrapods' evolved. This article therefore starts by describing a few 'tetrapod-like fish' and their relationships to the 'fish-like tetrapods' that evolved from them, before going on to see what kinds of animals emerged from the swamps and began their colonization of the land during the Carboniferous and Permian periods.

It was during the Late Palaeozoic that the foundations were laid for fully terrestrial living vertebrates, when they adapted their skeletons and physiologies for a very different environment from that in which they first evolved. Breathing mechanisms, feeding mechanisms and strategies, sensory systems, reproductive modes, and locomotory techniques all had to be modified during this greatest of evolutionary transitions, and many of these changes are reflected in skeletal changes that can be picked up in the fossil record. The differing means by which this was achieved is also reflected today in the legacy of modern tetrapod anatomy and physiology.

The general reader may be aware that modern tetrapods fall into a number of major groups, namely the amphibians, reptiles, birds, and mammals, but in fact these can be clumped into two coherent but biologically divergent groups, with the amphibians (known as 'lissamphibians') forming one lineage and the rest forming the other, known as the amniotes. Amphibians and amniotes each trace their origins to the Late Palaeozoic and, within the amniotes, each of the subgroups can also be traced back to a Carboniferous or Permian origin. By contrast, one of the least understood episodes of tetrapod evolution is the origin of modern lissamphibian groups (of which there are only three: frogs, salamanders, and caecilians). None of these can be found in the fossil record before the Early Jurassic, and the earliest stem member is *Triadobatrachus*, an animal with both frog- and salamander-like features from the Late Triassic. To be called an amphibian in the strict sense, an animal would be more closely related phylogenetically to these modern forms than to any amniote group.

Amongst Palaeozoic forms were also many lineages that have left no modern descendants, and whose origins do not seem to fall into either the amphibian or the amniote lineages, and these simply have to be called 'early tetrapods'. (They may have had an 'amphibious' life style, laying eggs in water and emerging onto land only as adults, but this describes their mode of life, not their relationships, and in many cases there is no evidence of what their reproductive mode might

have been.) This article looks mainly at these early tetrapod groups and the amphibian lineage; although the earliest amniotes were first found in the Late Carboniferous, they largely fall outside the scope of this article, but are dealt with in more detail in **Fossil Vertebrates:** Reptiles Other Than Dinosaurs.

'Tetrapod-Like Fish'

Tetrapods and their closest fish relatives belong to the 'lobe-finned vertebrates' or sarcopterygians. Other modern sarcopterygians are the lungfishes (dipnoans) and the coelacanths (actinistians), but during the Devonian there were many representatives of a lineage now known as the tetrapodomorphs (various subsets of these have been referred to as 'crossopterygians' or 'rhipidistians', but these terms can be misleading and are not now in general use). The best known tetrapodomorph is a fish called *Eusthenopteron*, mainly from the Frasnian of Canada and the Baltic states. It used to be pictured as the 'ancestor' of tetrapods, because it shares some fundamental features with tetrapods, including the form of its nose and the shape of some of the bones in its pectoral fin skeleton (Figure 1). More recently, other tetrapodomorph fish have been discovered which are more closely related to tetrapods than is *Eusthenopteron*, including the Frasnian genera *Panderichthys* (Figure 1) from the Baltics and Russia, and *Elpistostege* from Canada. Knowledge of their anatomy is helping to establish the timing and sequence of acquisition of some key tetrapod characters in skull, jaw, and limb construction. More recently still, tantalizing fragments, such as the lower jaw of an animal called *Livoniana* (Figure 1), have been found, hinting that, at the cusp of the origin of tetrapods, a wide range of animals existed that exploited different ecological niches. *Livoniana* sported at least seven rows of teeth on parts of its lower jaw, but also showed some subtle but key features marking it as close to tetrapods themselves. Figure 1 shows a diagram of the relationships of some of these early groups to one another, as well as the time around which they existed. The diagram implies that tetrapods originated around the Middle to Late Devonian boundary.

The Earliest Tetrapods

The earliest known limb or girdle elements belong to a tetrapod called *Elginerpeton* from the Frasnian of Scotland, associated with lower jaw and skull fragments. These establish a minimum date for the appearance of tetrapods. After that, during the Famennian, the fossil record expands, so that, at present, at least seven genera of tetrapods are recognized from that epoch. Three of them are known from articulated limb and digit material: *Acanthostega* (Figure 2), *Ichthyostega* (Figure 2), and *Tulerpeton*, the first two from East Greenland and the third from Russia. *Ichthyostega* was first discovered in 1929, and numerous skull and postcranial skeletons exist. Some parts of the anatomy have been described in detail, while other parts have only been sketchily portrayed, and the formal systematic nomenclature has been problematic until recently. Present studies show that there are two main species, associated with two distinct geological formations, an example of micro-evolution that is rarely possible to establish in early tetrapod material because of its rarity. *Ichthyostega* is, in many ways, an unlikely early tetrapod because some of its anatomy is highly modified and unlike that of any other tetrapod, ancient or modern. It has very broad overlapping ribs with a relatively huge shoulder girdle. Its recurved, keeled teeth are those of an undoubted predator, in a solidly constructed, massive skull. Its ear region and braincase suggest that it was specialized for underwater hearing, and its finned tail and paddle-like hind limbs support the notion of a largely aquatic animal.

Acanthostega, likewise, was an aquatic animal, but in a much more committed fashion. Its finned tail was deeper and longer than that of *Ichthyostega*, and both limbs appear to have been paddles rather than walking legs. Its ribs were short, and its shoulder girdle displayed some primitive fish-like features. Furthermore, grooved gill bars imply that functional internal gills were still present. Its generally primitive appearance suggests that it was primarily, rather than secondarily, aquatic, and that it gives us a good view of what a very primitive early tetrapod was like. *Acanthostega* is known from many specimens, although most of the best derive from a single small exposure that may represent a point bar against which a number of carcasses were swept in a flood.

Tulerpeton is known mainly from a single specimen that shows a complete fore and hind limb (Figure 3), some shoulder girdle elements, and part of a snout. Other fragments from the same locality may belong to it. The sediments from which it derives appear to have been marginal marine in origin, possibly a shallow, warm lagoon, but probably quite far from the nearest landmass. However, its limbs appear more like those of a terrestrial animal than those of the Greenland forms.

These three animals have something significant in common: they all possessed more than five digits on the limbs that are preserved (Figure 3). In *Ichthyostega*, the hind limb bore seven digits in a remarkable

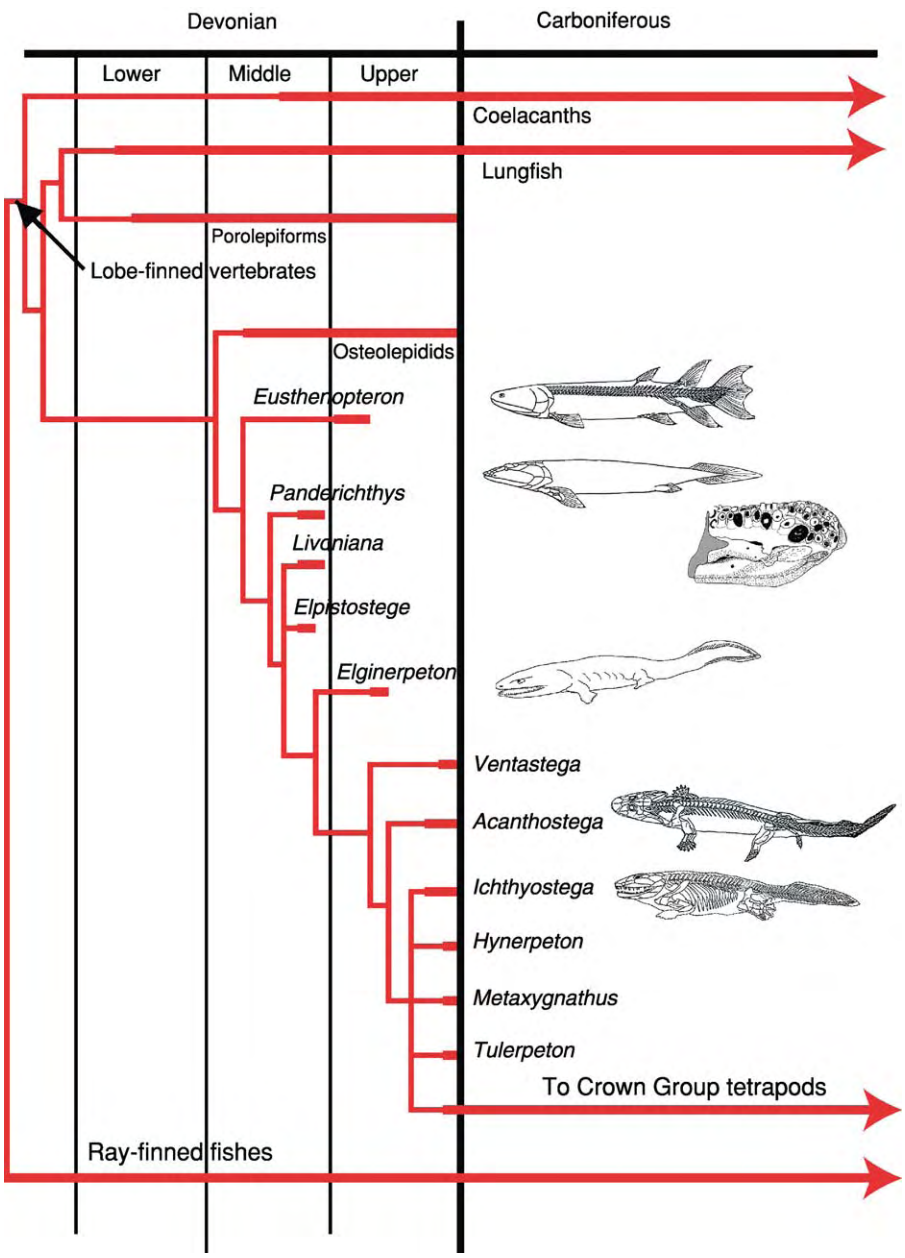


Figure 1 Cladogram of lobe finned fish and tetrapodomorphs, set on a time scale showing their occurrence. Note the loss of midline fins from *Panderichthys* onwards, and the retention of the tail finning in *Acanthostega* and *Ichthyostega*. *Livoniana* is known only from two jaw fragments, of which one is shown here. *Panderichthys* and *Ichthyostega* grew to over a metre in length, while *Eusthenopteron* and *Acanthostega* were slightly smaller, and *Elginerpeton* probably somewhat larger. The animals are not shown to scale. Arrows show that the group or its descendants survived to the present day.

pattern – at the front edge, two small followed by one extremely small digit formed a leading support for the paddle, with four stout digits forming the rest of the paddle (Figure 3). There are at least three specimens which show this. To be preserved in this way suggests that they were held in strong tissue in life. In *Acanthostega*, both the fore and hind limbs had at least eight digits, although there is only one specimen of each (Figure 3). In *Tulerpeton*, certainly the fore limb and probably the hind limb each had six digits. These discoveries have stimulated a rethink about how limbs with digits evolved and have provoked interchange with ideas coming from developmental biology and genetics. They also suggest that limbs with digits arose initially for swimming or paddling in water, not for walking on land.

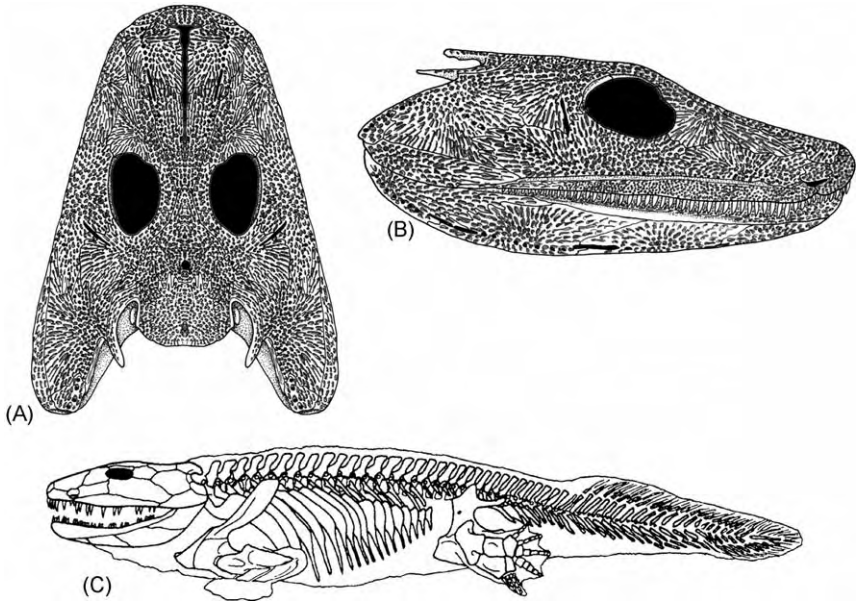


Figure 2 (A, B) Skull roof of *Acanthostega* in dorsal (A) and right lateral (B) views. The skull grew to about 200 mm in length. (C) Reconstruction of the skeleton of *Ichthyostega* based on recent information. However, study in progress indicates that further revisions will be necessary.

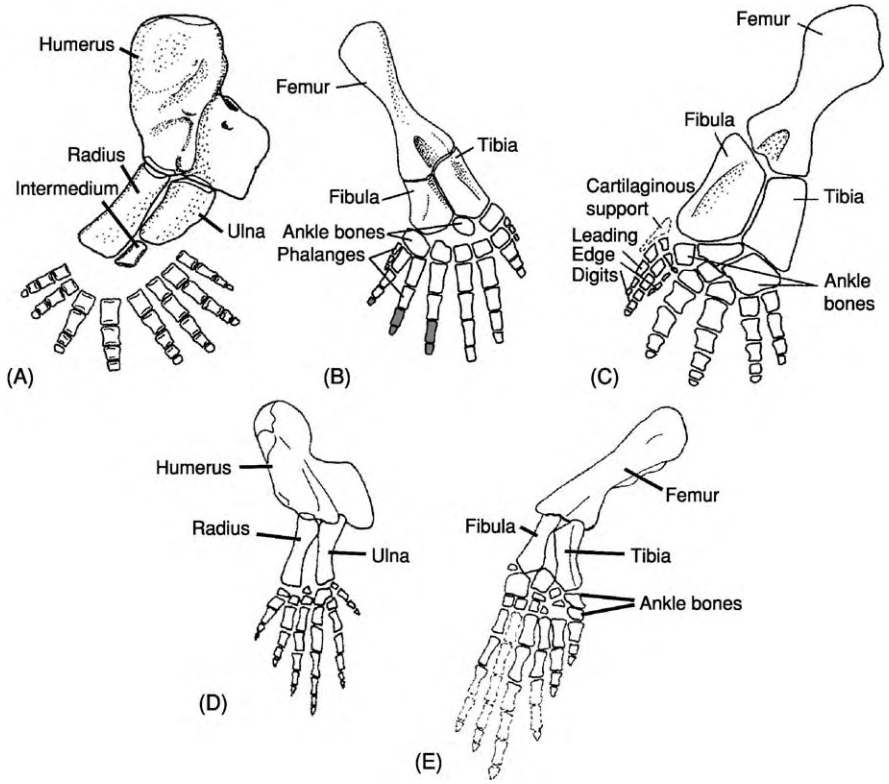


Figure 3 Limbs of Devonian tetrapods. (A) Fore limb of *Acanthostega*. (B) Hind limb of *Acanthostega*. (C) Hind limb of *Ichthyostega* (the manus is unknown in this form). (D) Fore limb of *Tulerpeton*. (E) Hind limb of *Tulerpeton*. Not to scale. Grey shading in *Acanthostega*, reconstructed elements. Broken lines in *Tulerpeton*, reconstructed elements. Adapted with permission from Clack JA (2002) *Gaining Ground: the Origin and Evolution of Tetrapods*. Bloomington, IN: Indiana University Press.

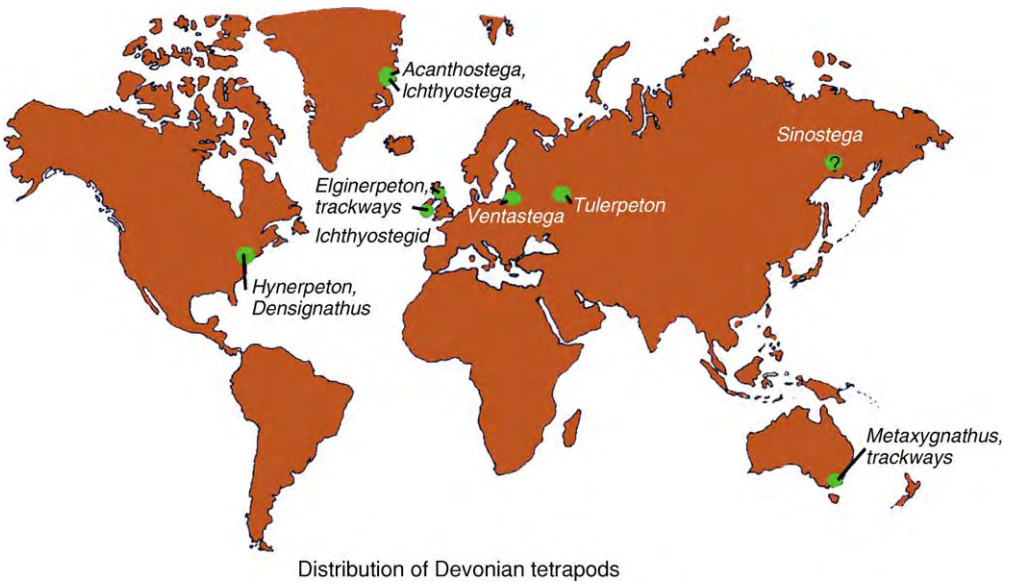


Figure 4 Map of the world showing the distribution of Devonian tetrapods worldwide.

Other Devonian tetrapods discovered over the last 10 years or so include *Hynerpeton*, known from a lower jaw and pectoral girdle material, and *Densignathus*, known from a lower jaw, both from Pennsylvania, USA. *Ventastega,* represented by excellent skull and pectoral girdle material, is from Latvia and, most recently, new taxa have been found in Greenland (as yet un-named), China (*Sinostega*) and Belgium (un-named). *Metaxygnathus*, from New South Wales, Australia, has been reconfirmed as a tetrapod in the light of work on *Acanthostega*. *Ventastega* is the best known of these and, interestingly, although it shows many similarities to *Acanthostega*, it also shows some more primitive and some unique features. Figure 4 shows the distribution of Devonian tetrapod finds worldwide.

Tetrapods of the Early Carboniferous

After the end of the Devonian, there is a long hiatus in the fossil record of tetrapods, informally known as 'Romer's Gap' (Figure 5). There may have been a mass extinction that correlates with the end-Devonian, but, whether sudden or gradual, there appears to have been a climate change affecting plants and animals alike for the next 15–20 million years. Unfortunately, it was during this period that tetrapods became fully terrestrial, but evidence to document this crucial period in the history of life on Earth is extremely scarce. Only a few specimens represent the Tournaisian and early Viséan epochs in tetrapod history, whereas the record of the late Viséan and Namurian is now much fuller than it was even two decades ago.

Pederpes finneyae is the only articulated tetrapod specimen currently known from Romer's Gap (Figures 5 and 6). It was found near Dumbarton in Scotland, in marginal marine cementstone sequences that have not otherwise yielded many fossils. It shows the earliest evidence of a hind limb that was used for walking and that was functionally pentadactyl, although what little is known of its fore limb suggests that it might have had relict accessory toes like those of the Devonian forms. Its closest relative, called *Whatcheeria deltae*, is from the Viséan of Iowa, USA. It is also possibly related to a mid-Viséan form, *Ossinodus*, from Queensland, Australia, and another possibly Tournaisian or Viséan form from Northern Ireland. They each show some primitive features, but share some more derived ones with later Carboniferous forms, and appear to form a clade (or possibly a grade) of Early Carboniferous forms that were widely distributed in time and space. Apart from these, a few isolated limb and girdle elements, and some large footprints from Horton Bluff in Nova Scotia, are all that represent the Tournaisian and early parts of the Viséan.

The tetrapods that have been discovered in the period following Romer's Gap show a range of sizes and body forms greater than that seen in almost any vertebrate group before (Figures 5 and 6). Certainly the tetrapods of the Late Devonian are quite conservative in shape and size – all are around a metre long with flattened and approximately spade-shaped heads. Those of the mid-Carboniferous range from mouse-sized to several metres in length with evidence of diversity in locomotory and feeding adaptations.

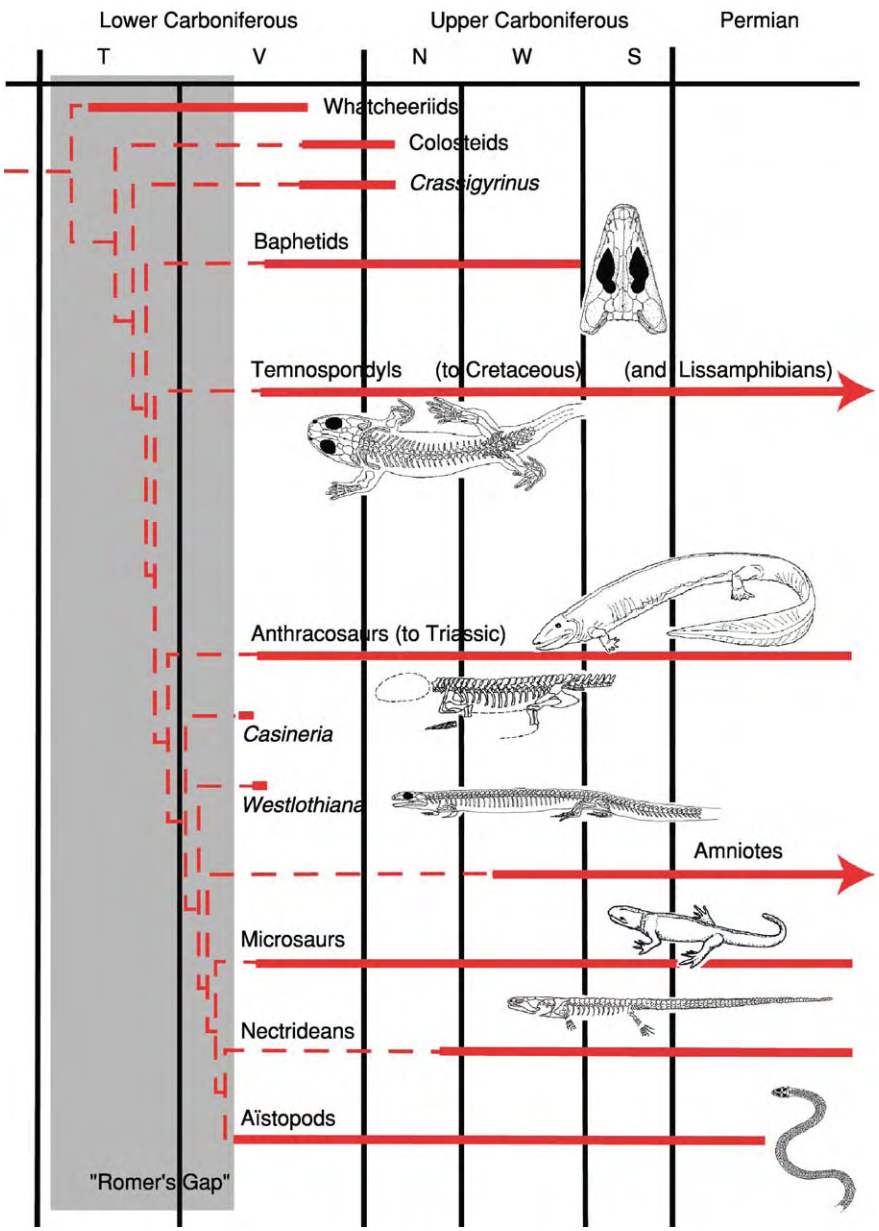


Figure 5 Diagram to show the distribution in time of some Late Palaeozoic tetrapod groups, with one view of their relationships. Romer's Gap is shown as a shaded band, and this time period probably represents that during which these major groups diversified. Representative animals are shown on some of the branches, but these are not to scale. Baphetid skulls, such as *Megalocephalus* figured here, reached over 300 mm in length. *Balanerpeton*, shown representing temnospondyls, grew to a total body length of between 300 and 400 mm. The embolomere, *Pholiderpeton*, shown representing anthracosaurs, probably reached a length of over 3 m. *Casineria* was about the size of a mouse, and *Westlothiana* that of a small lizard. The microsaur, *Pelodosotis*, featured here was only about 30 mm in length, but others grew to around 300 mm. *Keraterpeton*, representing the nectrideans, reached about 300 mm in total, although later ones grew much larger (see **Figure 8**). The aistopod, *Lethiscus*, seen here had a skull length of about 30 mm, but the full length of its body is not known. Arrows show that the group or its descendants survived to the present day. T, Tournaisian; V, Viséan; N, Namurian; W, Westphalian; S, Stephanian.

In the later Viséan, three localities, all near Edinburgh, Scotland, have yielded significant tetrapod specimens. Two, close in both time and space, have each yielded a single but intriguing specimen, while the third has yielded a whole flora and fauna of terrestrial organisms. *Casineria kiddi* (**Figure 5**), from Cheese Bay on the Firth of Forth, is the earliest fully terrestrial tetrapod known. It is small, unusual in such an early tetrapod, and has a clearly pentadactyl hind foot that could grasp the substrate. Despite the specimen not preserving a head, this animal appears to be the earliest representative of the amniote stem

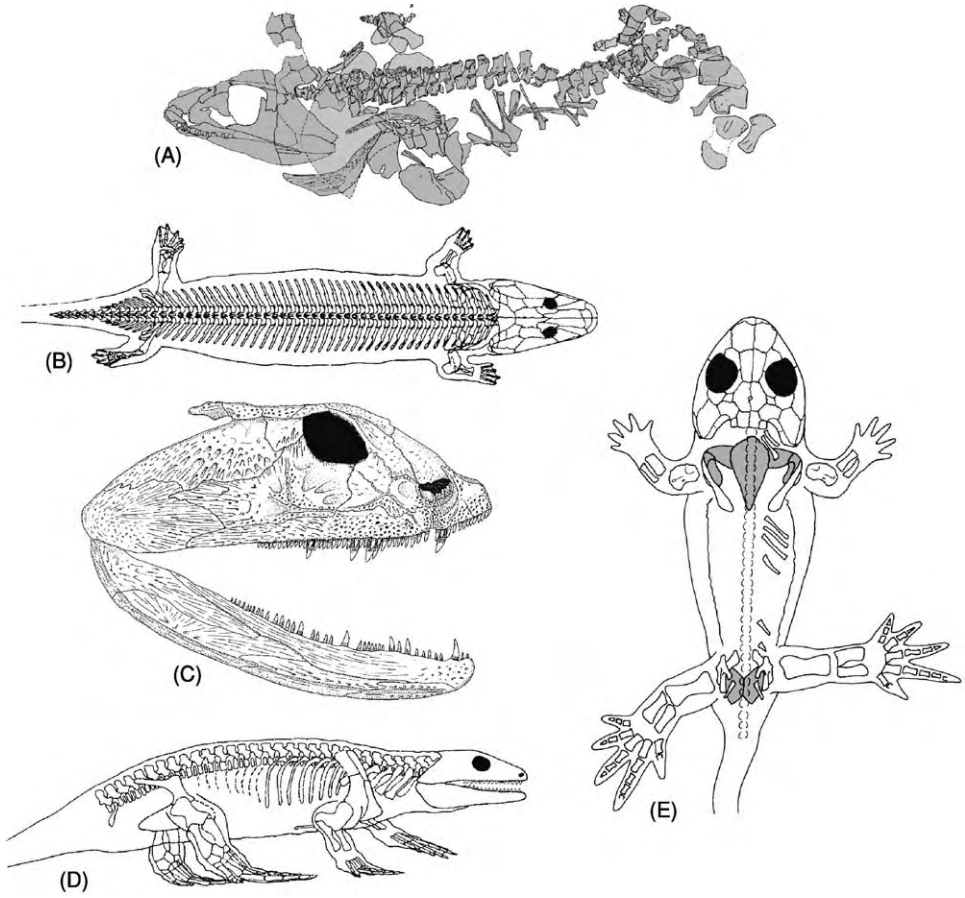


Figure 6 Early Carboniferous tetrapods. (A) Diagram of the (only) specimen of *Pederpes*, about 650 mm in length as preserved. (B) Reconstruction of the colosteid, *Greererpeton*, with a skull up to 180 mm in length. (C) Skull of *Crassigyrinus*. With its elongate body and tail, this animal probably reached between 2 and 3 m in length. (D) *Eldeceeon*, an early anthracosaur, about 300 400 mm in length. (E) *Eucritta*, probably a basal baphetid, with a maximum known skull size about 80 mm in length. (B) (E) adapted with permission from Clack JA (2002) *Gaining Ground: the Origin and Evolution of Tetrapods*. Bloomington, IN: Indiana University Press.

lineage. *Lethiscus stocki* (Figure 5), from Wardie, also on the Firth of Forth, is by contrast an aïstopod, an animal with a highly specialized skull structure and which has lost all its limbs and girdles, elongated its vertebral column and become a snake analogue. These two specimens hint at a radiation of tetrapods that took place soon after the end-Devonian, in which they began to explore and exploit a range of terrestrial niches, but of which we know virtually nothing.

East Kirkton, to the west of Edinburgh, has yielded at least six named genera of tetrapods, most of which seem to have been terrestrial. Its unusual volcanigenic geology has preserved plants, invertebrates, and tetrapods in some detail and diversity. The fauna includes the earliest representatives of at least four major groups of tetrapod that went on to radiate later in the Carboniferous, and some of which ultimately founded the lineages leading to modern amphibians and amniotes. The most common tetrapod from East Kirkton belongs to the group known as temnospondyls and is called *Balanerpeton woodi* (Figure 5). Temnospondyls, according to most analyses, are the group from which at least some modern amphibians (frogs and salamanders) arose. *Balanerpeton*, although the earliest, is apparently not the most primitive, and it shows many features characteristic of the group. *Silvanerpeton miripedes* and *Eldeceeon rolfei* (Figure 6) are the earliest members of the group known as the anthracosaurs, a group that traditionally has been seen as related to the amniotes, although the relationship may not be as close as suggested at one time. *Eucritta melanolimnetes* (Figure 6), an initially puzzling form that combines features of both temnospondyls and anthracosaurs, may be the earliest member of the baphetids, a group whose relationships to other tetrapods are still problematic. *Westlothiana lizziae* (Figure 5) is notable for having

been considered the earliest stem amniote for a while, although it has recently been supplanted by *Casineria*. Its elongate body and short limbs may suggest, alternatively, that it belongs to a further group, the microsaurs, but its poor skull preservation does not yet allow the resolution of this question.

Most other late Viséan forms actually span the boundary into the early Namurian, but deserve to be treated as Early Carboniferous forms. A gap in the fossil record in the mid–late Namurian seems to represent some kind of faunal turnover, as there is only one genus that has a representative both before and after the interval. A distinctively Late Carboniferous fauna emerges during the later Namurian.

The mid-Carboniferous forms include the bizarre aquatic predator, *Crassigyrinus scoticus* (Figure 6), which features an enormous gape, huge palatal teeth, and a snout with a large midline foramen at the front, whose function is obscure. *Crassigyrinus* was an elongate animal with disproportionately small fore limbs, and was possibly an analogue of a modern moray eel. It is known from a few localities in Scotland. The colosteids form a group that had members in Scotland and the USA. They were permanently aquatic predators that retained a number of primitive features in their anatomy, and probably had persistent gillbars. Members include *Pholidogaster*, *Greererpeton* (Figure 6), and *Colosteus*. Their relationship with other Carboniferous groups is disputed, but that is true both of *Crassigyrinus* and another group that appears in the mid-Carboniferous, the baphetids (formerly known as loxommatids).

Tetrapods of the Late Carboniferous

Temnospondyls, anthracosaurs, and baphetids that originated in the mid-Carboniferous survived to radiate in the Late Carboniferous, producing large crocodile-like predatory forms characteristic of the coal swamps of the time. Baphetids are striking for their curious keyhole-shaped eye-sockets (Figure 5); the purpose of the forwardly directed extension is obscure. Very little is known about their postcranial skeletons, because, although their skulls are well known, only two have postcranial material associated with them. Anthracosaurs (Figures 5 and 7) retain many primitive characters of the skull, but these are shared with later amniote-like forms, and they also share their type of vertebral construction (called gastrocentrous) with more terrestrially adapted forms, including amniotes. By contrast, most early temnospondyls retained a primitive style of vertebral construction (called rhachitomous), associated with an aquatic life style and seen in *Acanthostega*. Their skulls, however, share features later seen in modern amphibians; large vacuities in the roof of the mouth (although covered with skin in life) might have been part of a breathing and hearing mechanism similar to that found in modern amphibians today. Temnospondyls also show other similarities to frogs and salamanders in their life histories. Several families are known to have had aquatic larvae with external gills and that metamorphosed into more terrestrial adults, and some, it is suspected, had members that remained as permanently aquatic larval-like forms, as in some modern salamanders, such as the axolotl. These larval temnospondyls are collectively known as 'branchiosaurs' (Figure 7), but the term embraces several different families. By contrast, some temnospondyls became much more fully terrestrial. One group, the dissorophoids, produced some of the most terrestrial of temnospondyls, while, at the same time, including forms that are apparently the most closely related to frogs (Figures 7 and 8).

The three groups described above (together with colosteids and a few others) were, at one time, lumped together as 'labyrinthodonts', but in fact these animals may not be particularly closely related to each other – indeed they may belong to opposing sides of the dichotomy leading to the two modern tetrapod groups. These groups filled the large-predator niche at the same time as several other groups filled the role of small aquatic or terrestrial nibblers of invertebrates. These forms, often collectively known as 'lepospondyls', include the microsaurs (both aquatic forms with elongate bodies and small short-bodied terrestrial forms that appear superficially amniote-like), the adelogyrinids and aïstopods (Figure 5) (elongate and almost or completely lacking limbs), and the nectrideans (united by specialized vertebrae in the tail that formed a vertically flattened and stiffened but laterally flexible and powerful organ for swimming). The skulls of all these groups are specialized and difficult to compare with each other or with non-lepospondyls. This is one reason why their relationships to each other and to non-lepospondyls are still highly controversial. Some studies have suggested that these forms are in fact closer to modern amphibians and amniotes than are temnospondyls or anthracosaurs.

Tetrapods of the Late Carboniferous are found almost universally distributed over a wide area of Central and Western Europe and eastern North America, representing the known extent of the coal forest. Although genera may differ, families produced their equivalents in widely spread localities, suggesting a uniformity of habitat and climate over many millions of years. Finds have usually been found in

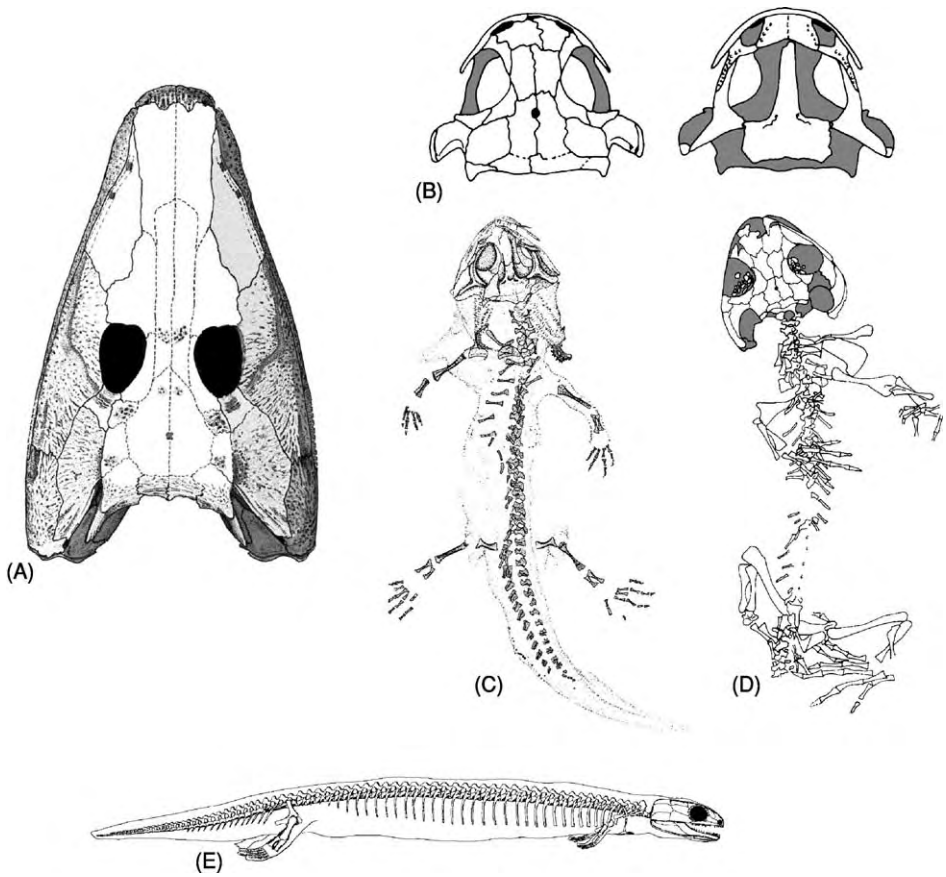


Figure 7 Late Carboniferous tetrapods. (A) Skull of the embolomerous anthracosaur, *Pholiderpeton*, in dorsal view, about 300 mm in length. (B) Skull of the branchiosaur, *Apateon*, in dorsal (left) and palatal (right) view. (C) Image of a branchiosaur fossil, about 50 mm in total length. (D) The dissorophoid, *Platyrhinops*, with a body length of about 120 mm. (E) The microsaur, *Microbrachis*, about 140 mm in length. (A) (D) adapted with permission from Clack JA (2002) *Gaining Ground: the Origin and Evolution of Tetrapods*. Bloomington, IN: Indiana University Press; (E) adapted with permission from Carroll RL, Bossy KA, Milner AC, Andrews SM, and Wellstead CF (1998) *Handbook of Palaeoherpetology*, Part 1, *Lepospondyli*. Munich: Verlag Friedrich Pfeil.

localities representing shallow channel-fills or deeper lake habitats, but few represent terrestrial environments for the Late Carboniferous. This may explain why the origins of amniotes and of modern amphibians are still poorly understood, in that their early members may have lived in environments that have not left a good fossil record. However, it is certain that the earliest true amniotes are found in the Late Carboniferous.

Tetrapods of the Early Permian and Their Descendants

Essentially, tetrapods of the Early Permian represent a continuity of lineages from the Late Carboniferous; this may be an artefact of stratigraphy in that many localities usually considered as earliest Permian may turn out to be latest Carboniferous after all. As the Carboniferous ended and the Permian began, the climate of the former coal swamp areas became gradually drier and the coal forests dwindled. Many localities from the Carboniferous–Permian boundary are formed from red-bed deposits characterized by evaporites, and there is a much greater representation of terrestrial environments. Even so, it has become clear recently that similar faunas existed over the areas of eastern North America and Western Europe.

Baphetids did not survive beyond the end of the Carboniferous, but anthracosaurs lived on into the Permian, with the long-bodied *Archeria crassidisca* of North America being the best known. The last members lingered on into the Late Permian and Early Triassic in Russia. Temnospondyls, by contrast, continued to dominate aquatic habitats, and produced a wide range of morphologies and life styles throughout the Permian. Although they were hit

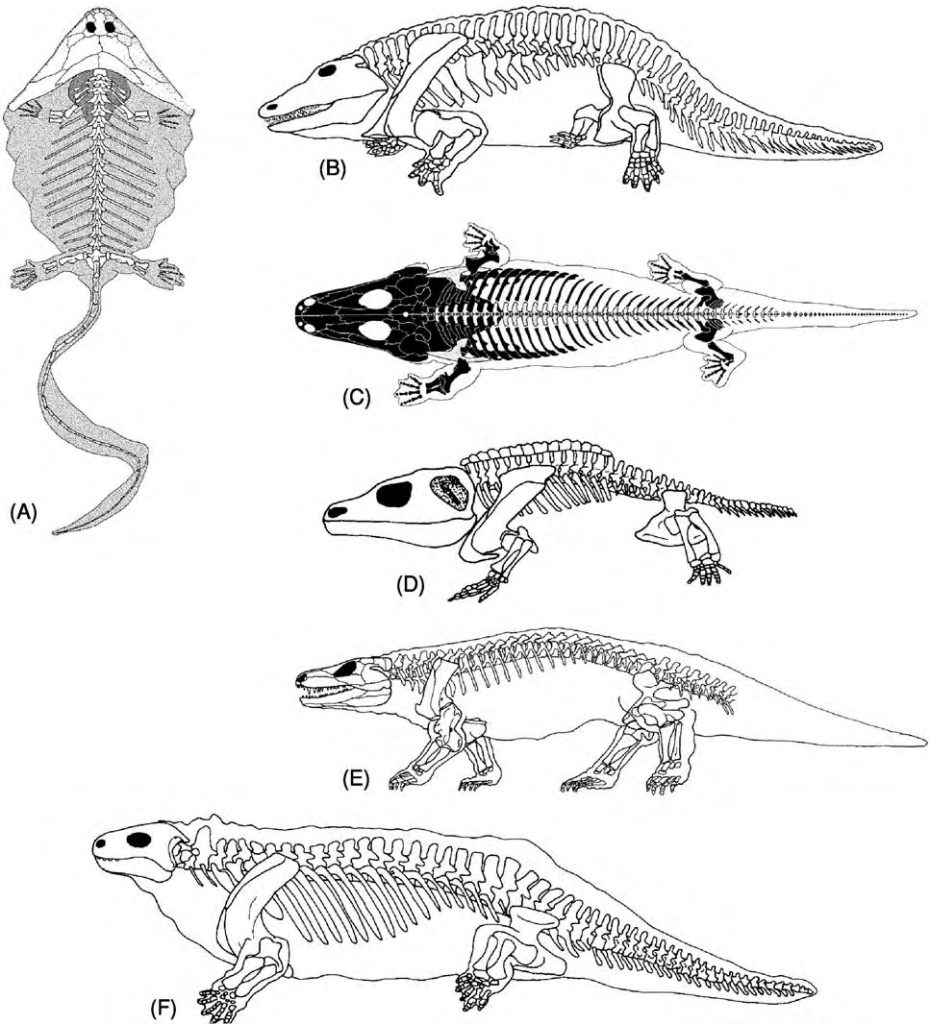


Figure 8 Early Permian tetrapods. (A) The nectridean, *Diplocaulus*, known from a range of sizes with skulls up to 147 mm in length. The body outline, suggested by trace fossils, is included in grey, although it is not clear whether the lateral skin flaps extended down the thighs. (B) The temnospondyl, *Eryops*, about 2 m in length. (C) The capitosauroid temnospondyl, *Mastodonsaurus*, an aquatic form up to 6 m in length. (D) The dissorophid temnospondyl, *Cacops*, a terrestrial form with a skull about 150 mm in length. (E) The seymouriamorph, *Seymouria*, about 1 m in length. (F) The diadectomorph, *Diadectes*, about 2 m in length. (B) and (D) (F) adapted with permission from Clack JA (2002) *Gaining Ground: the Origin and Evolution of Tetrapods*. Bloomington, IN: Indiana University Press; (A) adapted with permission from Carroll RL, Bossy KA, Milner AC, Andrews SM, and Wellstead CF (1998) *Encyclopaedia of Palaeoherpetology*, Part 1, *Lepospondyli*. Munich: Verlag Friedrich Pfeil; (C) adapted with permission from Schoch RR and Milner AR (2000) *Handbook of Palaeoherpetology*, Part 3B, *Stereospondyli*. Munich: Verlag Friedrich Pfeil.

heavily by the end-Permian extinction, after it they re-radiated into a range of aquatic forms, some extremely large, and at least one lineage survived into the Early Cretaceous. The best known of the Early Permian temnospondyls is *Eryops* (Figure 8) from south-eastern North America. It was thought of for many years as a 'model' early tetrapod, but this idea has turned out to be misleading. It is atypical for a temnospondyl in being a large semi-terrestrial form, when most were either large and fully aquatic or small and terrestrial. Dissorophids, for example, produced small terrestrial desert dwellers with nasal adaptations for retaining moisture (Figure 8), whereas capitosaurs were flat-headed aquatic giants (Figure 8).

Three groups that originated in the Late Carboniferous came to dominate the terrestrial habitats of the Early Permian. These are the seymouriamorphs, the diadectomorphs, and the synapsids. The latter were true amniotes and their lineage gave rise to mammals, but the other two, although fully terrestrial, are not thought to be fully amniote. Indeed, seymouriamorphs are known from some exquisite

material from Slovakia that includes growth series, and this shows that they grew from aquatic larvae with external gills, like the branchiosaurs. As they grew, the land-going features of their skeletons, such as the attachment of the hip girdle to the vertebral column, became better developed. *Seymouria* (**Figure 8**), well known from Texas, Oklahoma, and German deposits, had a uniquely specialized ear region, suggesting terrestrial audition. *Diadectes* (**Figure 8**) was a large barrel-bodied form with crushing dentition, surely a herbivore eating tough terrestrial plants.

A few of the lepospondyl groups persisted into the Early Permian. Nectrideans produced some bizarre forms with heads shaped like advanced hang-gliders, such as *Diplocaulus* (**Figure 8**). The laterally pointing 'horns' are known from trace fossils to have been attached by broad flaps of skin to the body, and though nectrideans were undoubtedly aquatic, their precise mode of life is not known (**Figure 8**). Lysorophids, elongate and almost limbless derivatives of microsaurs, are known to have formed burrows in which to aestivate during dry seasons, as some lungfish do today. All these lepospondyl groups, however, had died out by the end of the Early Permian.

The Permian was the time when the land came to be dominated by the amniotes, as adaptations to land dwelling improved with changes to feeding and breathing mechanisms for dealing with terrestrial conditions. The story of non-amniotes is largely confined to temnospondyls until the mid-Mesozoic, when fossils representing modern amphibian groups are first picked up.

See Also

Fossil Vertebrates: Jawless Fish-Like Vertebrates; Fish; Reptiles Other Than Dinosaurs; Mesozoic Amphibians and Other Non-Amniote Tetrapods; Mesozoic Mammals. **Palaeozoic:** Devonian; Carboniferous; Permian; End Permian Extinctions

Further Reading

Ahlberg PE and Milner AR (1994) The origin and early diversification of tetrapods. *Nature* 368: 507 514.

Benton MJ (2000) *Vertebrate Palaeontology,* 2nd edn. Oxford: Blackwell Science.

Bolt JR and Lombard RE (2000) Palaeobiology of *Whatcheeria deltae*. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeontology,* pp. 1044 1052. Chipping Norton, NSW: Surrey Beatty.

Boy JA and Sues H D (2000) Branchiosaurs: larvae, meta morphosis and heterochrony in temnospondyls and sey mouriamorphs. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeontology,* pp. 1150 1197. Chipping Norton, NSW: Surrey Beatty.

Carroll RL (1988) *Vertebrate Palaeontology and Evolution.* New York: W. H. Freeman and Co.

Carroll RL (2000) Lepospondyls. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeon tology,* pp. 1198 1269. Chipping Norton, NSW: Surrey Beatty.

Carroll RL, Bossy KA, Milner AC, Andrews SM, and Wellstead CF (1998) *Handbook of Palaeoherpetology.* Part 1, *Lepospondyli.* Munich: Verlag Friedrich Pfeil.

Clack JA (2000) The origin of tetrapods. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palae ontology,* pp. 979 1029. Chipping Norton, NSW: Surrey Beatty.

Clack JA (2002) *Gaining Ground: the Origin and Evolu tion of Tetrapods.* Bloomington, IN: Indiana University Press.

Clack JA and Carroll RL (2000) Early Carboniferous tetra pods. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeontology,* pp. 1030 1043. Chipping Norton, NSW: Surrey Beatty.

Holmes R (2000) Palaeozoic temnospondyls. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeontology,* pp. 1081 1120. Chipping Norton, NSW: Surrey Beatty.

Laurin M (2000) Seymouriamorphs. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeon tology,* pp. 1064 1080. Chipping Norton, NSW: Surrey Beatty.

Milner AR (1993) Biogeography of Palaeozoic tetrapods. In: Long JA (ed.) *Palaeozoic Vertebrate Biostratigraphy and Biogeography,* pp. 324 353. London: Belhaven Press.

Rocek Z and Rage J C (2000) Anatomical transformations in the transition from temnospondyl to proanuran stages. In: Heatwole H and Carroll RL (eds.) *Amphibian Biol ogy,* vol. 4, *Palaeontology,* pp. 1274 1282. Chipping Norton, NSW: Surrey Beatty.

Ruta M, Coates MI, and Quicke DLJ (2003) Early tetra pod relationships revisited. *Biological Reviews* 78: 251 345.

Schoch RR and Milner AR (2000) *Handbook of Palaeoher petology.* Part 3B, *Stereospondyli.* Munich: Verlag Fried rich Pfeil.

Smithson TR (2000) Anthracosaurs. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology,* vol. 4, *Palaeon tology,* pp. 1053 1063. Chipping Norton, NSW: Surrey Beatty.

Reptiles Other Than Dinosaurs

R R Reisz, University of Toronto at Mississauga, Mississauga, ON, Canada

Introduction

Who are the reptiles? The term 'reptile' can be readily associated with a series of extant animals, the turtles, crocodiles, lizards, and snakes, and the tuatara, but in a historical context this term also refers to a bewildering array of extinct forms that include the ancient Palaeozoic anapsids, flying reptiles, aquatic reptiles, mammal-like reptiles, and, of course, dinosaurs. In order to understand the evolutionary history of reptiles we must consider the wider implications of this name. In effect, we must start with the group Amniota, vertebrate tetrapods that are characterized by the presence of the embryological structure amnion and three other extraembryonic membranes that are critical to the survival and development of the embryo. All these animals lay an amniotic egg or did so primitively. Even the embryos of viviparous amniotes (therian mammals and some snakes) retain the extraembryonic membranes found in the amniotic egg (the allantois, yolk sac, chorion, and amnion). The allantois is a bladder storing metabolic wastes; it absorbs water from the environment because of the high osmotic pressure of its fluids and helps gas exchange. The yolk sac stores nutrients, the chorion provides a surface for gas exchange with the outside, while the amnion maintains a liquid environment around the embryo. This is in strong contrast to amphibians (frogs, salamanders, and apodans) which lack this complex set of embryological structures.

However, Amniota encompasses not only fossil and living reptiles but also mammals and birds, and this complicates the issue because the distant, ancient relatives of mammals were biologically reptile-like, and birds are technically reptiles because of their close association with dinosaurs. The early diversification of amniotes is probably the most exciting chapter in higher vertebrate evolution, and it is this section of the fossil record that I will discuss in some detail (**Figure 1**).

Early Amniotes

Amniotes can be divided into two large groups: synapsids and reptiles. Synapsida includes mammals and all fossil amniotes more closely related to mammals than to reptiles. The Palaeozoic synapsids are the so-called mammal-like reptiles that represent the most spectacular radiation of amniotes during the Carboniferous and Permian Periods. Reptilia includes extant lizards, snakes, crocodiles, collectively called diapsids because ancestrally they all had two temporal fenestrae behind the orbit, but also includes the anapsid testudines, and all the fossil amniotes more closely related to them than to mammals.

Many early reptiles were previously viewed as anapsids, referring to the absence of any temporal fenestrae behind the orbits. These include mesosaurids, parareptiles, captorhinids, and the so-called protorothyridids. The taxon Anapsida does not have any phylogenetic significance, being largely based on the absence of a series of derived characters. They are, however, part of the initial diversification of amniotes.

Mesosaurs were the first fully aquatic amniotes. They are known from the Permian of southern Africa and eastern South America. Their presence on both sides of the Atlantic has been used to support the theory of continental drift. Mesosaurs are highly specialised amniotes, characterized by numerous autapomorphies that are found in most parts of their skeletons. Their snout was very long, mainly composed of long, slender premaxillae and dentaries, and their unusually long teeth extended anteriorly (near the tip of the snout), laterally (on the anterior half of the sides of the snout), and ventrolaterally (on the posterior half of the sides of the snout) from the upper jaw. The dentition seems to have formed a basket for filtering water and capturing small invertebrates. The neck is long, an adaptation of several aquatic amniotes. The tail was long and laterally compressed, and was probably used for swimming. The trunk ribs are pachyostotic; they were very thick and heavy and served as ballasts. The scapular blade was very short, as in several other aquatic amniotes. Contrary to previous interpretations, there is no evidence for a lateral temporal fenestra in these forms.

Mesosaurs are now considered to be part of the Parareptilia, a group of Palaeozoic and Early Mesozoic reptiles that may include turtles and their close relatives, and also includes four main groups; millerettids, pareiasaurs, procolophonids (**Figure 2**), and turtles. They are united by the several synapomorphies, including a dorsally enlarged quadratojugal bordering the quadrate emargination. Millerettids appear at the beginning of the Upper Permian. They were small, insectivorous animals superficially similar to lizards. Several millerettids have either a lower temporal fenestra or a deep cheek emargination. They

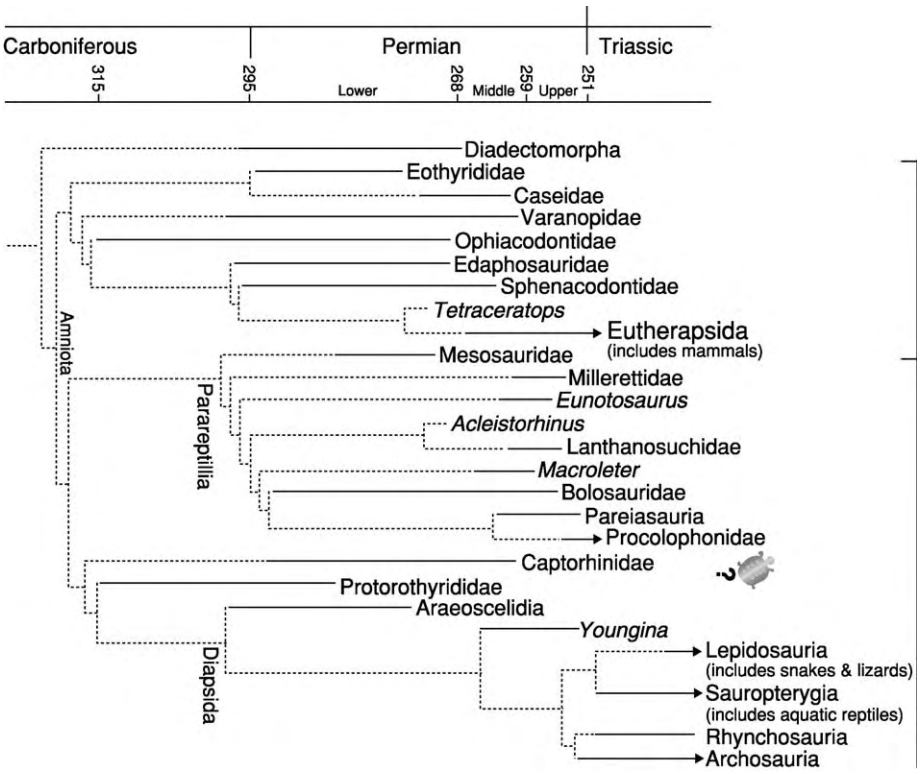


Figure 1 Early amniote phylogeny. The pattern of solid and dotted lines represent evolutionary relationships of the various taxa known from the fossil record. The solid lines with the taxonomic units to their right represent the known fossil record of those particular taxa. For example the fossil record of the Diadectomorpha extends from the Permo Carboniferous boundary throughout Lower Permian strata. The dotted lines represent reconstructed histories of particular taxa (also called ghost lineages). These are based on the combination of the known earliest appearance of certain lineages, and the proposed pattern of evolutionary relationships. For example, the known record of Eothyrididae extending to the Permo Carboniferous boundary automatically creates a long ghost lineage for Caseidae, whose record only starts towards the top or the Lower Permian.

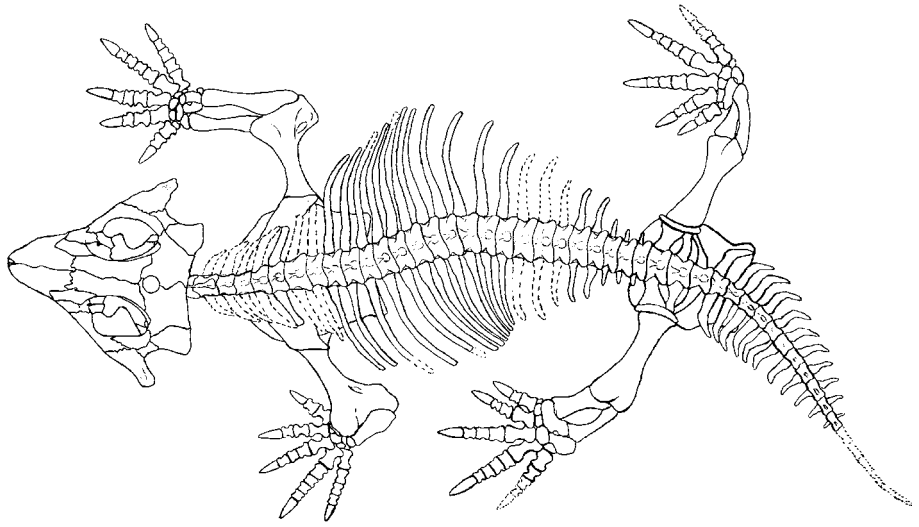


Figure 2 *Procolophon*. This small parareptile is known from the Early Mesozoic, Lower Triassic sediments of Gondwana.

have a quadrate emargination supporting a tympanum. Millerettids appear to be the first amniotes to acquire a tympanum. This tympanum may be homologous to the tympanum of turtles but it is not homologous to the tympanum found in diapsids, in mammals, or in frogs and temnospondyls. Pareiasaurs were large, herbivorous amniotes that lived in the Upper Permian in Africa, Western Europe, Russia, and China, one of the first group to achieve nearly global distribution. They were also among the largest Palaeozoic amniotes, reaching a length of 3 m. They were covered by an extensive armour of osteoderms (dermal bony scales). Their skull was massive, broad, and had strange bony processes in the cheek and lower jaw. They probably had a parasagittal, semi-erect posture. They had a reduced phalangeal formula in the foot, a character advantageous for an animal having a parasagittal posture. Procolophonids were small and primitively insectivorous, as millerettids were, but late procolophonids have transversely broad teeth that may indicate a herbivorous diet. Procolophonids appear in the fossil record at the very end of the Upper Permian and became extinct at the end of the Triassic. The orbit is expanded posteriorly, especially in late procolophonids. The last procolophonids had strange bony processes and their orbit was much longer posteriorly than the orbit of early procolophonids.

Testudines may be the only surviving parareptiles. They appear in the fossil record in the Upper Triassic, but they must have been present earlier, but are not preserved in the fossil record. Turtles have also been associated with diapsid reptiles, but their origins remain controversial.

Captorhinids were one of the most successful groups of early amniotes and are part of the group called Eureptilia. They appear in the fossil record in the Lower Permian and became extinct in the Upper Permian. Captorhinids are known mainly from sediments of North America, but a few individual taxa have also been recovered from North and Central Africa, India, and Russia, making it the earliest known group of reptiles to have a geographically widespread distribution. Early captorhinids were relatively small, and had a single row of marginal teeth, but soon after the appearance of the group, several captorhinids developed multiple marginal tooth rows in the upper and lower jaws. These tooth rows are restricted to the maxilla in the upper jaw, and the dentary in the lower jaw. Some of the more derived taxa, such as *Moradisaurus*, had up to 12 rows. In this derived, large captorhinid from the Late Permian of Africa, the cheeks are greatly expanded. The limbs were relatively short and massively built, with broad hands and feet.

The morphology of small captorhinids is known best from a vast amount of material collected from a Lower Permian fissure-fill locality near Fort Sill, Oklahoma. The overwhelming majority of the bones in these fissure fills are attributable to a single taxon, the small, multiple tooth-rowed *Captorhinus aguti* (Figure 3).

Protorothyridids are a group of small (100 mm snout-vent length), lizard-like eureptiles. We are not certain that they are a monophyletic group, because the interrelationships of members of this group have not been analysed, all previous studies having considered them only in terms of a generalized basal amniote pattern. They are closely related to diapsids because both groups have long and slender hands and feet (they are broad and robust in other amniotes).

The fossil record of protorothyridids is restricted to the Carboniferous and Early Permian of North America and Europe, and all the known taxa are represented either by a single, incomplete skeleton, or at most by a handful of partially preserved remains. *Hylonomus*, the oldest known amniote, has been recovered from the Middle Pennsylvanian (Westphalian A) of Joggins, Nova Scotia. All the specimens have been collected from inside upright *Sigillaria* stumps. Although known from several partial skeletons, all the specimens are completely disarticulated, and poorly preserved. A better known member of this group is *Paleothyris*, represented from three nearly complete skeletons from the Middle

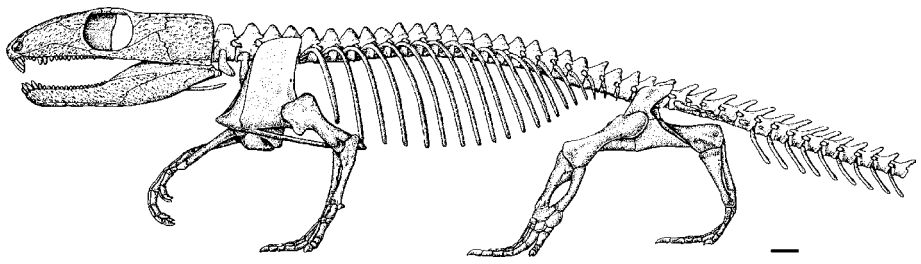


Figure 3 *Captorhinus*. This small eureptile is known from the Early Permian of North America. Members of this group, the Captorhinidae dispersed widely, and had a nearly cosmopolitan distribution in the Middle and Late Permian.

Pennsylvanian (Westphalian D) of Florence, Nova Scotia. Although these specimens were also collected from upright *Sigillaria* stumps, their preservation is superior to that of *Hylonomus*, allowing for a more detailed study. *Protorothyris* is a slightly larger form from the Early Permian of Texas. This form is known from a few skulls and a partial skeleton.

The meager evidence provided by the known protorothyridids give the general impression that they resembled early captorhinids, but had more slender limbs and narrow neural arches. This general impression is reinforced by the similarities between such taxa as *Protorothyris* and the recently described *Thuringothyris*, and captorhinids. Recent phylogenetic analyses indicate that protorothyridids are not an ancestral group of amniotes, as previously suggested, but are the sister-group, or closest known relative of diapsids.

Early Diapsid Reptiles

Diapsids are the most diverse and numerous eureptiles, and include all modern reptiles (except perhaps for turtles). They also include birds and several groups of extinct marine, flying, and terrestrial reptiles such as dinosaurs. The early history of this highly successful group extends into the Early Permian and the Late Pennsylvanian (latest Carboniferous), represented by the early diversification of the Araeoscelidia. Three well-known members of this group are now known, *Petrolacosaurus*, *Araeoscelis*, and *Spinoaequalis*.

The oldest known diapsid is *Petrolacosaurus* from the Upper Pennsylvanian of Kansas (Figure 4). Numerous specimens have been recovered from sediments that have filled in a Late Carboniferous tidal channel and preserved the most diverse amniote assemblage known for that period. *Petrolacosaurus* is the most commonly found amniote at this small site, other skeletons representing a wide variety of endemic synapsids. This small diapsid is characterised by an exceedingly delicate skull, which carries numerous, delicately built marginal teeth. *Petrolacosaurus* is unique among Late Pennsylvanian terrestrial vertebrates in having an unusually elongate neck, a long, slender tail, and slender, elongate limbs. Particularly striking among these is the delicately constructed forelimb, with long, slender fingers. The well preserved skeletal anatomy provides clear evidence that the oldest known diapsid was the most agile amniote of its time.

Araeoscelis from the Lower Permian of Texas is also a small diapsid, with a lightly built, slender skeleton, but its skull is much more massively built than that of *Petrolacosaurus*, and its marginal dentition is bulbous, and transversely expanded. *Araeoscelis* is a particularly fascinating early diapsid because there is strong evidence to indicate that it exhibits a secondarily closed lower temporal fenestra. Both of these diapsids are represented by several good skeletons, but the third, more recently described araeoscelidian, *Spinoaequalis*, from the Late Pennsylvanian of Kansas, is known from a single partial skeleton. This small reptile is only slightly younger than *Petrolacosaurus*, lacks the elongated neck, but has elongated limbs, as seen in the other two members of this clade. The appearance of *Spinoaequalis* in the fossil record, so soon after *Petrolacosaurus*, provides direct evidence for the diversification of diapsids within the Carboniferous. Of particular interest is the discovery

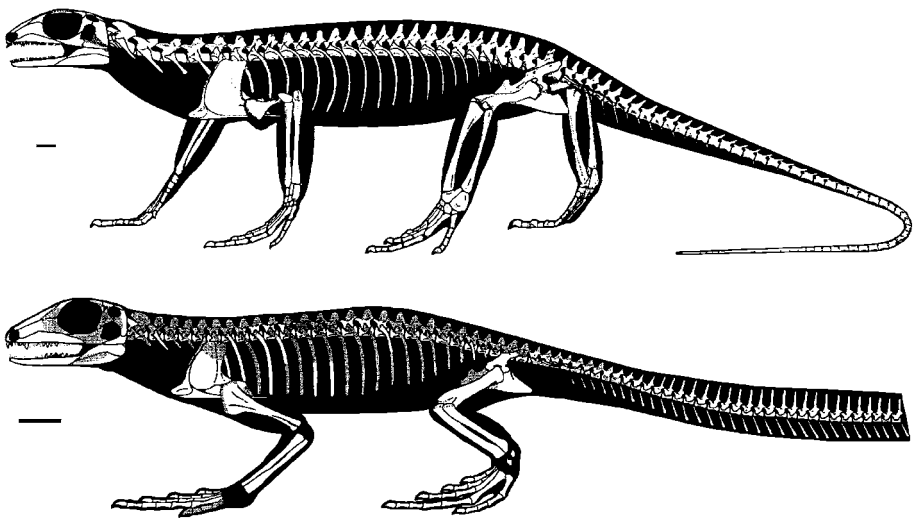


Figure 4 *Petrolacosaurus* and *Spinoaequalis*, the oldest known diapsid reptiles.

of a number of skeletal features in *Spinoaequalis* which display evidence for aquatic specialisations. Most striking among these is the presence of a dorsoventrally expanded tail. Araeoscelidians also share the typical diapsid characters with more recent diapsids: two temporal fenestrae, an upper fenestra between postorbital, parietal, and squamosal, and a lower fenestra between the postorbital, squamosal, and jugal (the quadratojugal sometimes enters the edge of the fenestra); a suborbital fenestra between the maxilla, palatine, and ectopterygoid.

Younginiforms are an important group of fossil diapsids from the Upper Permian and Lower Triassic of South Africa, eastern Africa, and Madagascar. *Youngina* is the best-known member of this group. Younginiforms and modern diapsids form a monophyletic group called the Neodiapsida. They are united by the presence of a ventromedial flange of the parietal on which the jaw musculature originated.

All other, younger diapsids belong to two major groups: the Lepidosauromorpha and the Archosauromorpha. Together, these two taxa include the forms that epitomise the Age of Reptiles, the Mesozoic, and also include most extant reptiles, with the possible exception of turtles (if the latter are parareptiles).

Lepidosauromorpha include mainly small- to medium-sized reptiles, both fossil and extant, and are represented today by *Sphenodon*, lizards, and snakes. These forms are characterized by an incomplete lower temporal bar. The modern sphenodontid *Sphenodon* has reverted to the primitive diapsid condition and has a complete lower temporal bar. Sphenodontids are represented only by *Sphenodon* in the modern fauna, but they were once a moderately diverse group. They first appear in the fossil record in the Upper Triassic. *Sphenodon* lacks a tympanum, but its fossil relatives probably had one. Sphenodontids have a strong jaw musculature, so their temporal arches are bowed laterally to accommodate it. Their dentition is acrodont (the teeth are fused to the surface of the jaw).

The Squamata is one of the largest group of living reptiles. It includes 6000 species of lizards and snakes. Snakes appear in the Cretaceous and lizards are first found in the Upper Jurassic. Several Upper Permian and Triassic lizards have been identified, but these are probably not lizards. Lizards are not a natural group because they include the ancestors of snakes. Most squamates have a pleurodont tooth implantation. This means that the teeth are attached to the medial surface of the jaws. There are very shallow tooth sockets, when they are present, and the alveolar shelf is strongly inclined (far from the horizontal). Early tetrapods have a subpleurodont or subthecodont implantation, in which the teeth are set in shallow sockets and the alveolar shelf is gently inclined. A few large squamates such as the Cretaceous mosasaurs, have teeth set in deep sockets (thecodont implantation). In a few squamates, such as agamids and chamaeleontids, the dentition is acrodont (fused to the surface of the jaw), as in sphenodontids.

Among squamates, mosasaurs were a large group of large aquatic lizards from the Upper Cretaceous. Almost 20 genera are known; the largest one, *Hainosaurus*, reached a length of 45 feet. Mosasaurs had a long body (29 to 51 presacral vertebrae) and small limbs with a reduced number of elements. They swam by anguilliform motion, which is rather slow, and the limbs were used only for steering. Mosasaurs lived in shallow, coastal marine waters. Some mosasaurs seem to have fed on ammonites (large cephalopods resembling the nautilus), because we found several ammonite shells with tooth marks that look like they were made by a mosasaur.

Snakes are one of the most recent and most rapidly evolving groups of reptiles. Most modern genera belong to families that appeared at the beginning of the Miocene or in the Oligocene. Most snakes have no limbs and there is not even a trace of the pectoral girdle, but several genera retain vestiges of the pelvic girdle and posterior limbs. Recent evidence indicates that snakes are probably related to mosasaurs, which in turn are related to varanoid lizards. Snakes are, therefore, nested within the clade of lizards. *Dinilysia* is the oldest known snake, dating from the Upper Cretaceous.

Aquatic Reptiles

As exemplified by the ancient Permian mesosaurs and the Cretaceous mosasaurs, return to the water is a common phenomenon throughout reptilian history. It has been argued that because of the relatively low metabolic rates, tolerance to anoxia, and generally low body temperatures, reptiles can make the transition from a terrestrial lifestyle to an aquatic one without major difficulties. It has been shown experimentally, in the modern marine iguana that the metabolic cost of swimming is only one-quarter that of walking. However, a return to the water to catch an easy meal is not the same as life in the water. Many reptiles today are amphibious and at home in the water but they do not live their entire lives there. Back in the Palaeozoic and more notably in the Mesozoic, many reptiles did evolve a fully aquatic lifestyle.

The relationships between Palaeozoic and Mesozoic marine reptiles and their terrestrial relatives are difficult to assess, but the following groups are generally considered to be derived from diapsids, probably lepidosauromorphs, with the lower temporal fenestra

being without a lower temporal bar, and only the upper temporal fenestra being retained in its original form.

Nothosaurs are medium-sized aquatic reptiles restricted to the Triassic. They are one of the two major groups of sauropterygians (the other one being plesiosaurs). They range in length from 20 cm to 4 m in body length. Their skull is relatively small, and the lower temporal bar has been lost. The slender stapes and quadrate emargination suggest that they may have had a tympanum. The interpterygoid vacuity is closed. This type of palate links nothosaurs with plesiosaurs. The ribs of nothosaurs were often pachyostotic. Their limbs and girdles are poorly ossified, but they have no striking aquatic adaptations. Nothosaurs probably swam relatively slowly by lateral undulation. The limbs were probably used for steering, although some people think that the forelimbs may also have been used for propulsion, because the coracoid is relatively large.

Plesiosaurs, a group of aquatic diapsids closely related to nothosaurs, appeared in the Jurassic and became extinct at the end of the Cretaceous. They had a short trunk and a short tail. Their coracoid, and to a certain extent their scapula, are greatly expanded ventrally. Young plesiosaurs had a shoulder similar to nothosaur's girdle. Plesiosaurs have massive ventral scales called gastralia, which may have strengthened the trunk. This would have been useful, because plesiosaurs used their limbs to swim. The exact way in which the limbs were used is still debated. They may have moved in the water much like the modern leatherback turtle (by subaqueous flight) or rowed like sea-lions. According to some, the short scapular and iliac blades implies that the muscles elevating the limbs were weak and poorly suited for subaqueous flight. On the other hand, perhaps the origin of these muscles had just shifted medially, closer to the vertebral column. The massive coracoids would, in any case, have provided a large surface for attachment of the forelimb retractor muscles. So, this question remains unsolved.

Placodonts are Triassic aquatic reptiles with robust tooth plates that strongly suggest that they had a durophageous diet; they may have fed on molluscs and arthropods. They had a short, stout body, and an extremely robust skull. The internal naris (choana) is a single, median opening. Their anterior dentary and premaxillary teeth are procumbent and spatulate; they were probably used to grab the prey. The interclavicle has a very short posterior stem and it is ventral to the clavicle, as in nothosaurs. The endochondral shoulder girdle is poorly ossified. Some genera had a carapace composed of polygonal dermal bones covered by epidermal scutes. In the most advanced members of this group the shoulder girdle has moved inside the armor and rib-cage, in a similar fashion to that seen in turtles. Some armored placodonts had lost their anterior teeth and may have had a horny beak instead.

Ichthyosaurs were similar to dolphins and sharks in size and shape. They appeared in the Lower Triassic, were numerous in the Jurassic, and became extinct before the end of the Cretaceous. Of all the Palaeozoic and Mesozoic marine diapsids, they were the most highly adapted to a marine existence. Typical ichthyosaurs (from the Jurassic and Cretaceous) had a relatively short, fusiform trunk, a dorsal fin, and a high, lunate caudal fin. We know about the dorsal fin because some specimens have the body outline preserved as a carbonaceous film. These features suggest that ichthyosaurs could swim rapidly. We know that ichthyosaurs were viviparious, because some of them died while giving birth and were preserved with the young ichthyosaur still in the pubic canal.

The orbit was large and the eye was protected by a well ossified sclerotic ring. The snout was long and bore numerous homodont, sharp teeth (in most genera). The quadrate was not notched posteriorly and the stapes was relatively massive and articulated with the quadrate distally, so there was probably no tympanum (it would not have been useful under water). The vertebrae of ichthyosaurs were short, wide amphicoelous disks. The neural arches never fused to the centra and there were no transverse processes.

Early Archosauromorphs and Crocodiles

Archosauromorphs include the most spectacular reptiles of the Mesozoic Era, were the dominant terrestrial and aerial vertebrates of that era, and include crocodiles and crocodile-like reptiles, dinosaurian relatives, flying reptiles, dinosaurs, and birds. The oldest known members of this large group show up near the end of the Permian, but the main radiation of this group starts during the Triassic.

Rhynchosaurs were the most common and widespread early archosauromorphs. They are found throughout the Triassic. They have been used for biostratigraphic correlations. They have a single, median external naris and a premaxillary beak. *Mesosuchus* is a Lower Triassic rhynchosaur, more primitive than later rhynchosaurs in retaining premaxillary teeth and in having a single row of maxillary and dentary teeth. Later rhynchosaurs have multiple tooth rows and a broad skull. Rhynchosaurs were among the most abundant large herbivores in the Middle and Upper Triassic. Their extinction at the end of the Triassic

may support the existence of a mass extinction event at that time.

Perhaps the most spectacular feature of the archosaurian skull is the presence of a new opening in the skull in front of the orbit, the antorbital fenestra. This was probably just a pneumatic structure, as in modern birds, making the skull lighter, but this feature readily distinguishes these forms from other diapsids. With the exception of a few other basal members of this large group, like *Proterosuchus* and *Euparkeria*, all other archosauromorphs belong to either the Crurotarsi, which includes crocodiles and their fossil relatives, and the Ornithodira, which includes flying reptiles, dinosaurs, and their closest relatives, and birds.

The Crurotarsi include some very spectacular relatives of crocodiles, but the earliest crocodile-like forms belong to the Sphenosuchidae. Sphenosuchids are known primarily from the Upper Triassic of Europe, North and South America, and the Lower Jurassic of South Africa. In contrast to true crocodiles, Sphenosuchids were relatively lightly built and had an erect, bipedal posture. True crocodiles (crocodyliforms) first appeared in the Upper Triassic. They are divided into three major groups, each representing an important adaptive radiation. These three groups are the Protosuchia, Mesosuchia, and Eusuchia. The Protosuchia and Mesosuchia are probably paraphyletic, so they are not real taxa. The skull of true crocodiles is large, massively built, and low. The skull roof is flat and has heavy pitting on the surface. The antorbital fenestra is small in primitive crocodiles and absent in living crocodiles. Living crocodiles have a long secondary palate to separate the passage of air from the mouth, as in mammals.

Palaeozoic Synapsids, the Mammal-Like Reptiles

No evolutionary history of reptiles is complete without some consideration of the so-called mammal-like reptiles, the Palaeozoic synapsids. Palaeozoic synapsids (also called pelycosaurs and early therapsids) occupy a central position in early amniote evolution. Palaeozoic synapsids include some of the oldest known amniotes, and their fossil remains record the earliest successful adaptations of terrestrial vertebrates to herbivorous and active predatory modes of life. During the Palaeozoic, synapsids diversified greatly and became the most conspicuous terrestrial vertebrates of their time. Although both pelycosaurs and early therapsids are paraphyletic taxa at the base of the clade that includes advanced therapsids and mammals, these Palaeozoic reptiles have played a pivotal role in considerations of synapsid evolution, including the origin of mammals. The oldest known mammal-like reptiles have been found in sediments of Early and Middle Pennsylvanian (Late Carboniferous) age in Nova Scotia. The remains of *Protoclepsydrops haplous*, retrieved from upright *Sigillaria* stumps from Joggins, Nova Scotia are too fragmentary to provide any useful information; even the identification of these remains as a pelycosaur has been questioned. By the Late Pennsylvanian, pelycosaurs become the most common amniotes in Laurasian sediments. The fossil record of this group is most extensive during the Early Permian, representing nearly 70% of all amniote finds of this time; by the Late Permian, however, pelycosaurs disappear from the fossil record. Instead, the early therapsids become the most diverse and common amniotes. Only during the Triassic do true reptiles become the most diverse and common amniotes.

As primitive synapsids, pelycosaurs are characterized by the presence of an anteriorly inclined plate-like occiput, and a lateral temporal fenestra that is bordered by jugal, squamosal, and postorbital bones. This is in contrast to the condition in diapsids, where the lateral temporal fenestra is bordered by a fourth bone, the quadratojugal primitively, and the occiput has large postemporal fenestrae. Recent phylogenetic studies of pelycosaurs have proposed that within this group, six major clades (families) can be recognized: small insectivores (Eothyrididae), two different types of large, bulky herbivores (Caseidae and Edaphosauridae), and three different types of medium to large-sized carnivores (Varanopidae, Ophiacodontidae, and Sphenacodontidae).

Eothyrididae

This family was erected as a provisional group for the small pelycosaur *Eothyris*, known from a single skull from the Lower Permian of Texas, and fragmentary remains of three large pelycosaurs. Other fossils were subsequently placed within this family, but at present only two genera, *Eothyris*, from the Early Permian of Texas, and *Oedaleops*, from the Early Permian of New Mexico, can be assigned with confidence into the Eothyrididae. *Eothyris* is known from a single skull, whereas *Oedaleops* is represented by three partial skulls and some limb elements. These specimens, although fragmentary, have revealed that eothyrids were small carnivorous amniotes.

Caseidae

The casseids, closely related to the eothyrids, were a widespread and diverse group of herbivorous pelycosaurs, with a fossil record that extended from the Early Permian of North America and western Europe

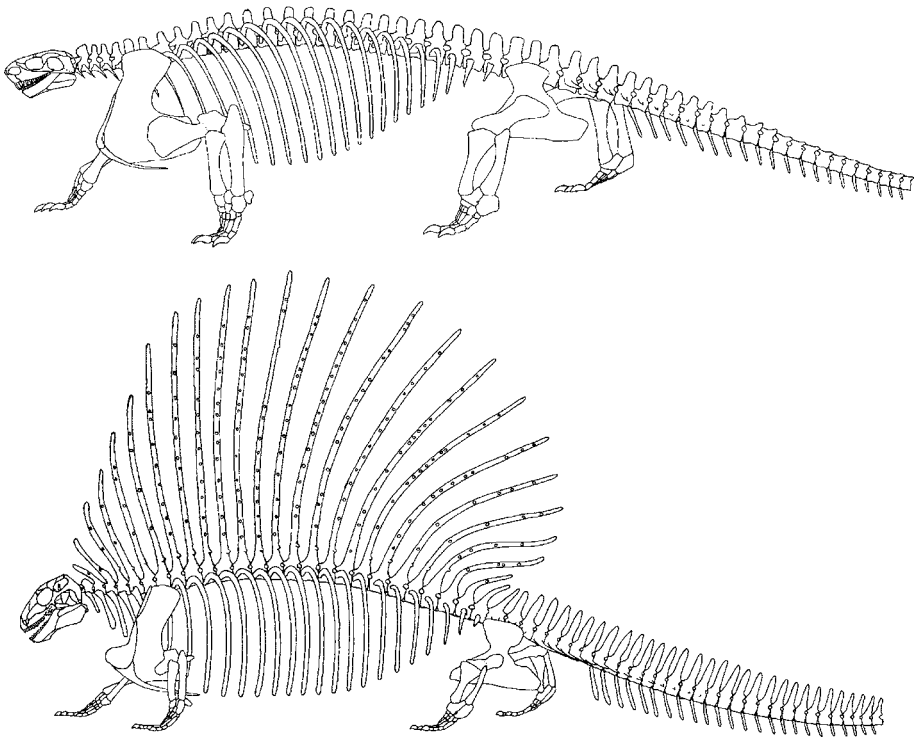


Figure 5 *Corylorhynchus* and *Edaphosaurus* two large Early and Middle Permian herbivorous synapsids (''mammal like reptiles'').

to the Late Permian of eastern Europe. Although ranging in size from 1 to nearly 4 meters in body length, caseids are surprisingly conservative in their skeletal anatomy and body proportions. The skulls are distinctive in the presence of relatively large temporal fenestrae, enormous external nares, and large pineal foramen. The snout overhangs dramatically the tooth row to form an anteriorly projecting rostrum, as in eothyridids. The external surface of the skull has unusual rounded deep pits and sometimes crevice-like depressions that form a distinctive sculpturing pattern. The marginal teeth have spatulate crowns and are serrated along the edge, in some respects similar to the teeth of pareiasaurs. The largest teeth in the marginal series is located anteriorly, on both the premaxilla and dentary. All caseids, whether modest in size, or enormous, are characterised by small cervical vertebrae, bulky, barrel-shaped bodies, and relatively massive limbs. There is reduction in the phalangeal formula, ranging from 2-3-4-4-3 to 2-2-2-3-2. Two distinct types of terminal phalanges can be found in caseids. In *Corylorhynchus*, the claw bearing element is very large and broad, with apparently sharp edges on the sides. In *Angelosaurus*, the claw bearing phalanges are also massive, but more conservative in morphology, roughly triangular in outline.

Caseids and eothyridids share a number of derived characters associated with the morphology of the snout and external naris that support the hypothesis that these primitive pelycosaurs form a clade, the Caseasauria. All other pelycosaurs can be placed in the Eupelycosauria, a clade that also includes therapsids and mammals (Figure 5).

Varanopidae

These faunivorous synapsids of small to moderate size are all characterized by slender, lightly built skulls with highly specialized marginal dentition (Figure 6). All varanopids have mediolaterally flattened teeth with anterior and posterior cutting edges, and are strongly recurved. Both the premaxilla and maxilla are also highly modified, the tooth bearing portion of the premaxilla being enlarged, and the maxilla being greatly elongated in association with the posterior extension of the marginal dentition. Other diagnostic fedatures of varanopseids includes enlargement of the temporal fenestra and the reduction of the subtemporal arch to a narrow bar. The lower jaw is slender, and the mandibular symphysis is reduced in size. The postcranial skeleton is also lightly built, giving varanopids a sleek appearance, superficially similar to modern monitor lizards. It is, therefore, reasonable to suggest that varanopid pelycosaurs were probably the most agile predators of their time. *Mycterosaurus longiceps* is the most basal member of the group. The larger varanopids, *Varanops, Aerosaurus*, and

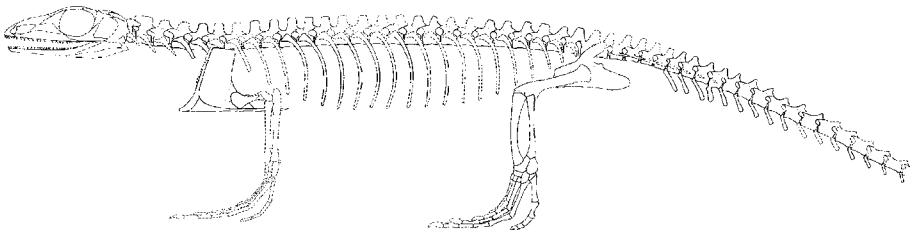


Figure 6 *Archaeovenator*, a Late Carboniferous varanopids synapsid (''mammal lime reptile'') from North America.

Varanodon, are all characterized by the presence of a massive external maxillary buttress above the caniniform region, a dramatic anterior tilt of the occiput, and anteroposterior enlargement of the temporal fenestra.

The hypotheses of phylogenetic relationships of this group of pelycosaurs has changed dramatically as our knowledge of primitive synapsids has increased. In 1940 varanopids were associated with the advanced sphenacodontids as a group of primitive carnivorous sphenacodonts, and it was even suggested that *Varanops* may have been close to the ancestry of sphenacodontids. The most recent hypothesis places varanopids within the Eupelycosauria, as its most primitive member. Although retaining many primitive cranial and postcranial features that appear in derived form in edaphosaurs, ophiacodontids and sphenacodontids, it is obvious that varanopids developed a number of autamorphies, and the advanced members of this group outlasted all other carnivorous pelycosaurs as highly specialised, agile predators. The large, predatory *Varanodon* is one of the youngest known pelycosaurs in North America, but two small varanopids, *Elliotsmithia* from the Upper Permian strata of South Africa, and *Mesenosaurus* from the Middle Permian of Russia extend significantly the fossil record of this group.

Ophiacodontidae

The oldest known pelycosaurs have been included in this family. *Protoclepsydrops* from the Middle Pennsylvanian of Joggins, Nova Scotia, has been placed within the Ophiacodontidae, but its identity and taxonomic status is uncertain, partly because the known skeletal remains are fragmentary.

Archaeothyris, from slightly younger sediments near Florence, Nova Scotia, is the oldest known diagnosable ophiacodontid.

Ophiacodontids include small to very large carnivorous pelycosaurs that have unusually long snouts formed by the elongation of the nasal, maxillary, and lacrimal bones. There is also a well developed lateral mandibular fenestra in a notch of the dentary. Among ophiacodontids there is a tendency towards increase in the skull to trunk ratio, larger taxa having unusually large, awkward looking, elongate skulls. In the postcranium, ophiacodontids have relatively large, massive shoulder girdles, with anteroposteriorly broad scapulocoracoids, but small headed interclavicles and clavicles. *Ophiacodon* is the best known member of the family, represented by an extensive fossil record in North America, that has been subdivided, somewhat arbitrarily (based largely on stratigraphic position and size) into six species. There appears to be extensive variation on morphology among the known specimens, even to the presence of two lateral temporal fenestrae on each side in larger specimens. In one particular specimen of *Ophiacodon retroversus*, the second opening is apparently present on one side of the skull but not on the other. A particularly puzzling characteristic of the postcranial skeleton is its reduced level of ossification, especially in the larges species, and the lack sharp claws as indicated by the truncated tip of the terminal phalanges. It is largely because of these features that *Ophiacodon* has been considered as being amphibious. This hypothesis is supported by the location of the orbit high on the side of the face, but it is nevertheless difficult to imagine how this animal, with a tall, slender skull, could be an effective aquatic predator.

Edaphosauridae

One of the most striking features exhibited by several pelycosaurs is the great elongation of the neural spines. The poorly known Middle Pennsylvanian *Echinerpeton*, the Lower Permian pelycosaur *Lupeosaurus*, at least three distinct genera of sphenacodontids, as well as all the edaphosaurs, have large sails supported by tall neural spines above the presacral portion of the vertebral column. Among these, edaphosaurs have the most spectacular sail because the greatly elongated neural spines have lateral projections or tubercles along each side. It is generally accepted that this great elongation of neural spines must have occurred independently at least three times among pelycosaurs. The presumed function of such elongation is for support of a membrane that spanned the space between successive spines. This membrane probably facilitated

temperature control, but the sail and its associated spines may have also served in display behaviour, and in edaphosaurs it may have also served as protection against predators. This last hypothesis is supported by the presence of the lateral tubercles, as well as by the orientation of the neural spines; anteriorly tilted cervical and thoracic spines and posteriorly tilted lumbar spines probably provided protection in the neck and thigh region, respectively.

Edaphosaurus and *Ianthasaurus* are the only pelycosaurs that preserved the highly specialized lateral tubercles on the spines. Both are characterized by the presence of greatly elongated neural spines that are rounded in traverse section, and by the presence of remarkable lateral tubercles. The arrangement of these tubercles along the height of the spines is similar in the two taxa, the proximal lateral tubercles being not only the largest of the set, but also paired. *Ianthasaurus* is a small edaphosaur from the Upper Pennsylvanian that lacks many of the spectacular specialisations seen in *Edaphosaurus*. For example, the marginal dentition of *Ianthasaurus* is similar to that of insectivorous reptiles, with slender conical teeth which are slightly recurved at the tips, and there is a slight development of a caniniform region. The palatal and mandibular dentition is unspecialised, and there are no batteries of teeth for crushing of plant materials. This is in contrast to the cranial morphology of *Edaphosaurus*, where the palate has been greatly modified in order to carry a massive array of closely packed teeth that occluded against a similarly developed set of teeth on the medial surface of the mandible. The two edaphosaurs are also quite distinct from each other in their postcranial morphology. Whereas *Ianthasaurus* appears lightly built and was probably quite agile, *Edaphosaurus* possesses a combination of features that suggest that this animal was a heavy, relatively slow herbivore. For example, the cervical vertebrae are reduced in length in association with the reduced size of the skull, the dorsal vertebrae are massive, the tail is deep, the limbs are short, and robust and the ribs form a wide ribcage. Nine species of *Edaphosaurus* have been described, ranging in size from small to very large, bulky animals. The largest species, *Edaphosaurus cruciger* and *Edaphosaurus pogonias*, have modified their cervical and anterior thoracic spines into massive club-like processes. It is not unreasonable to suggest that the small, insectivorous *Ianthasaurus* represents the primitive edaphosaur pattern from which the larger herbivorous *Edaphosaurus* may have been derived.

Sphenacodontia

Haptodus is a relatively small pelycosaur that is considered to be a sister taxon to the sphenacodontians. *Haptodus* has been recovered from the Late Pennsylvanian and Early Permian of North America and Europe. Their morphology suggests that members of this genus were probably effective predators, being able to feed on both arthropods and small vertebrates. A number of haptodontine-like forms appear to represent progressively closer sister-taxa to the most spectacular pelycosaurs, the sphenacodonts. The Sphenacodontidae include the large carnivores *Dimetrodon*, *Sphenacodon*, *Ctenospondylus*, and *Secodontosaurus*. These forms were the dominant predators of their time, and their fossil record extends from the Late Pennsylvanian throughout the Early Permian. The significance of this group, however, lies with its phylogenetic relationships, because sphenacodontids are generally considered to be the nearest pelycosaurian relatives of therapsids. Sphenacodonts have a supracanine buttress on the medial surface of the maxilla, the first premaxillary and second dentary teeth are greatly increased in size, similar in length to the enlarged canine, the ventral edge of the premaxilla

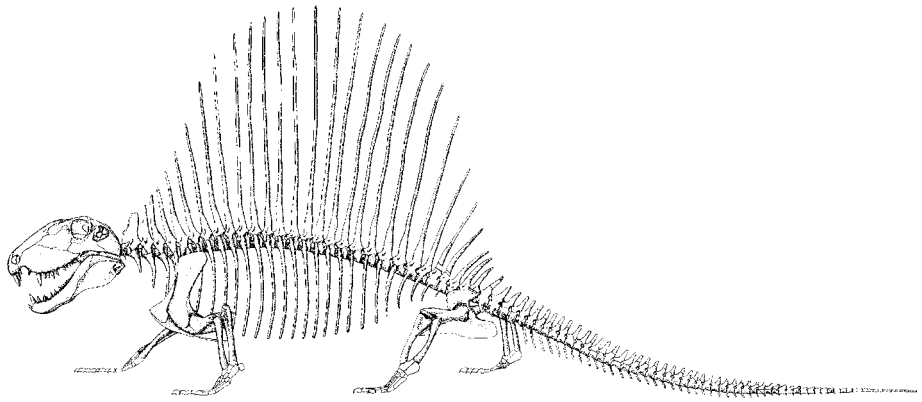


Figure 7 *Dimetrodon*, one of the earliest known top predators, an amniote synapsid (''mammal like reptile'') from the Permian of North America.

slopes anteroventrally, giving the snout a massive appearance, the supraorbital shelf forms a hood over the orbit and terminates posteriorly on the concave lateral surface of the postorbital, and the pineal foramen is located on a slightly raised area of the parietal, with a distinct ridge surrounding the foramen.

Sphenacodontids (*Dimetrodon*, *Sphenacodon*, *Secodontosaurus*, and *Ctenospondylus*) are all large to very large predators that have tall neural spines, strongly excavated centra and neural arches, and long limbs making them relatively agile, fast moving animals, especially when compared to their slower, bulkier herbivorous relatives. *Dimetrodon*, *Sphenacodon*, and *Ctenospondylus* all have similarly constructed, massive skulls with extremely large anterior incisors and canines (Figure 7). *Sphenacodon* and *Ctenospondylus* have blade-like neural spines, whereas *Dimetrodon* has greatly elongated spines that are rounded in transverse section. *Secodontosaurus* is greatly modified from the pattern seen in the other large sphenacodontids by reducing the width and height of the skull roof, while retaining the neural spine morphology of *Dimetrodon*. The cranial modifications of *Secodontosaurus* can only be interpreted as an adaptation to specialised feeding strategies, perhaps preying upon burrowing animals, or feeding on tetrapods that tried to evade capture by hiding in crevices.

Origin of Therapsids

A major morphological gap exists between pelycosaurs, which are mostly Pennsylvanian and Early Permian synapsids from North America and Europe, and therapsids, which are Late Permian and Triassic advanced synapsids from South Africa and Russia. Previous attempts at bridging this gap have not been successful. Among pelycosaurs the haptodontines and sphenacodontids have been proposed as the most likely candidates for sister-group status to therapsids. In 1940 it was proposed that of all the known sphencodontids *Haptodus* is the most likely ancestor of therapsids. More recently, in 1992, it was shown that the pattern of distribution of synapomorphies in haptodontines, sphenacodontids, and therapsids supports the hypothesis that sphenacodontids are the sister group of therapsids. The morphology of the reflected lamina of the angular provides the strongest evidence for the latter hypothesis: in both primitive therapsids and sphenacodontids, the reflected lamina is a prominent feature of the lower jaw, whereas in haptodontines the reflected lamina, if present at all, is a modest, slightly developed ventral projection with little or no lateral displacement.

Recent studies of *Tetraceratops* indicate that this enigmatic fossil from the Early Permian of Texas bridges the morphological and temporal gaps between the pelycosaurian and basal therapsid synapsids. As the oldest known therapsid, this form shares with other early therapsids a number of derived cranial features, but it also has numerous primitive features that appear in derived form in all other therapsids. *Tetraceratops* is, therefore, the sister taxon to all other therapsids.

Further Reading

Boy JA and Martens T (1991) Ein neues captorhino morphes Reptil aus dem thüringischen Rotliegend (Unter Perm; Ost Deutschland). *Paläontologische Zeits chrift* 65: 363 389.

Carroll RL (1963) The earliest reptiles. *Zoological Journal of the Linnean Society* 45: 61 83.

Carroll RL (1988) *Vertebrate Paleontology and Evolution.* New York: W.H. Freeman and Company.

Carroll RL and Baird D (1972) Carboniferous stem reptiles of the Family Romeriidae. *Bulletin, Museum of Comparative Zoolology* 143: 321 364.

Clark J and Carroll RL (1973) Romeriid reptiles from the Lower Permian. *Bulletin, Museum of Comparative Zoolology* 144: 353 407.

deBraga M and Reisz RR (1995) A new diapsid reptile from the uppermost Carboniferous (Stephanian) of Kansas. *Palaeontology* 38: 199 212.

Dodick JT and Modesto SP (1995) The cranial anatomy of the captorhinid reptile *Labidosaurikos meachami* from the Lower Permian of Oklahoma. *Palaeontology* 38: 687 711.

Dilkes DW and Reisz RR (1996) First record of a basal synapsid ('mammal like reptile') in Gondwana. Pro ceedings of the Royal Society of London, *Series B* 263: 1165 1170.

Laurin M (1993) Anatomy and relationships of *Haptodus garnettensis*, a Pennsylvanian synapsid from Kansas. *Journal of Vertebrate Paleontology* 13: 200 229.

Laurin M and Reisz RR (1990) *Tetraceratops* is the oldest known therapsid. *Nature* 345: 249 250.

Laurin M and Reisz RR (1995) The osteology and relationships of *Tetraceratops insignis*, the oldest known therapsid. *Journal of Vertebrate Paleontology* 16: 95 102.

Laurin M and Reisz RR (1995) A reevaluation of early amniote phylogeny. *Zoological Journal of the Linnean Society* 113: 165 223.

Modesto SP and Reisz RR (1990) A new skeleton of *Iantha saurus hardestii*, a primitive edaphosaur (Synapsida: Pelycosauria) from the Upper Pennsylvanian of Kansas. *Canadian Journal of Earth Sciences* 27: 834 844.

Modesto SP (1995) The skull of the herbivorous synapsid *Edaphosaurus boanerges* from the Lower Permian of Texas. *Palaeontology* 38: 213 239.

Modesto SP (1996) A basal captorhinid reptile from the Fort Sill fissures, Lower Permian of Oklahoma. *Oklahoma Geology Notes* 56: 4 14.

Olson EC (1991) An eryopoid (Amphibia: Labyrinthodontia) from the Fort Sill fissures, Lower Permian, Oklahoma. *Journal of Vertebrate Paleontology* 11: 130 132.
Reisz RR (1977) *Petrolacosaurus*, the oldest known diapsid reptile. *Science* 196: 1091 1093.
Reisz RR (1997) The origin and early evolutionary history of amniotes. *Trends in Ecology and Evolution* 12: 218 222.
Reisz RR, Heaton MJ, and Pynn BR (1982) Vertebrate fauna of Late Pennsylvanian Rock Lake Shale near Garnett, Kansas: Pelycosauria. *Journal of Paleontology* 56: 741 750.
Reisz R (1972) Pelycosaurian reptiles from the Middle Pennsylvanian of North America. *Bulletin of MCZ* 144: 27 61.
Reisz RR (1986) Pelycosauria. vol. 17 A. *Handbuch der Paläoherpetologie*. Wellenhofer P (ed.) Stuttgart, New York: Gustav Fischer Verlag, 1 102.
Reisz RR, Berman DS, and Scott D (1992) The cranial anatomy and relationships of *Secodontosaurus*, an unusual mammal like reptile (Synpasida: Sphenacodontidae) from the early Permian of Texas. *Zoological Journal of the Linnean Society* 104: 127 184.
Reisz RR and Modesto SP (1996) *Archerpeton anthracos* from the Joggins Formation of Nova Scotia: a microsaur, not a reptile. *Canadian Journal of Earth Sciences* 33: 703 709.
Romer AS and Price LW (1940) Review of the Pelycosauria. *Geologica Society of America Special Papers* 28: 1 538.

Dinosaurs

A M Yates, University of the Witwatersrand, Johannesburg, South Africa

Introduction

Dinosaurs were the dominant terrestrial vertebrates of the Mesozoic Era. All large terrestrial animals from the end of the Norian Stage (Late Triassic) to the close of the Cretaceous were dinosaurs. This is not to say that any large Mesozoic animal is automatically referred to as a dinosaur. Membership of the group can be diagnosed by the presence of numerous evolutionary novelties (Figure 1), which indicate that dinosaurs are a natural evolutionary group or clade.

On the evolutionary tree of terrestrial vertebrates, dinosaurs nest deeply within the Reptilia. Within the reptiles they belong to the Diapsida (reptiles with two openings, or fenestrae, for jaw muscles in the temporal regions of their skulls), within the diapsids they belong to the Archosauria (diapsids with an additional fenestra in the skull behind the nostril and in front of the orbit and that lay calcite-shelled eggs), and within the archosaurs they belong to the Avemetatarsalia (archosaurs with a simple hinge-like, or mesotarsal, ankle joint and elongate bunched metatarsal bones in the foot).

The Dinosauria contains two subgroups (Figure 2), the Ornithischia ('bird-hipped' dinosaurs) and the Saurischia ('lizard-hipped' dinosaurs). These subgroups were long thought to have evolved from different reptiles, but modern cladistic work has shown that they almost certainly form a clade that excludes all other archosaurian groups.

Diagnostic Characters of the Dinosauria

A great number of derived characteristics have been proposed for the diagnosis of the Dinosauria. However, many have since been shown to be present in other archosaurs (e.g. an s-shaped neck) or to be independent acquisitions in different dinosaur groups (e.g. three or more sacral vertebrae). The following list includes the characters that have been well documented and are likely to remain diagnostic of the dinosaur clade.

1. Loss of the postfrontal bone from the skull roof. In other archosaurs there is a postfrontal bone between the frontal and postorbital bones.
2. A contribution of the frontal bone to the depressed area that surrounds the upper temporal fenestra. In life this fossa would have served as an attachment site for jaw closing muscles. In other archosaurs the frontal bone is excluded from this fossa.
3. Loss of the lateral sheet of the squamosal bone that covers the dorsal head of the quadrate bone. The dorsal end of the reptilian quadrate bone (which forms the upper part of the jaw joint) forms a smooth, rounded head that fits into a socket under the squamosal bone. In other archosaurs a lateral sheet of the squamosal bone obscures the quadrate head in lateral view but in dinosaurs the quadrate head is plainly visible.
4. The sternum consists of paired ossified plates. In early tetrapods the sternum was composed of a single median interclavicle situated behind the paired clavicles ('collar bones' in humans). The interclavicle was lost early in the history of archosaurs, while the clavicles were strongly reduced in size.

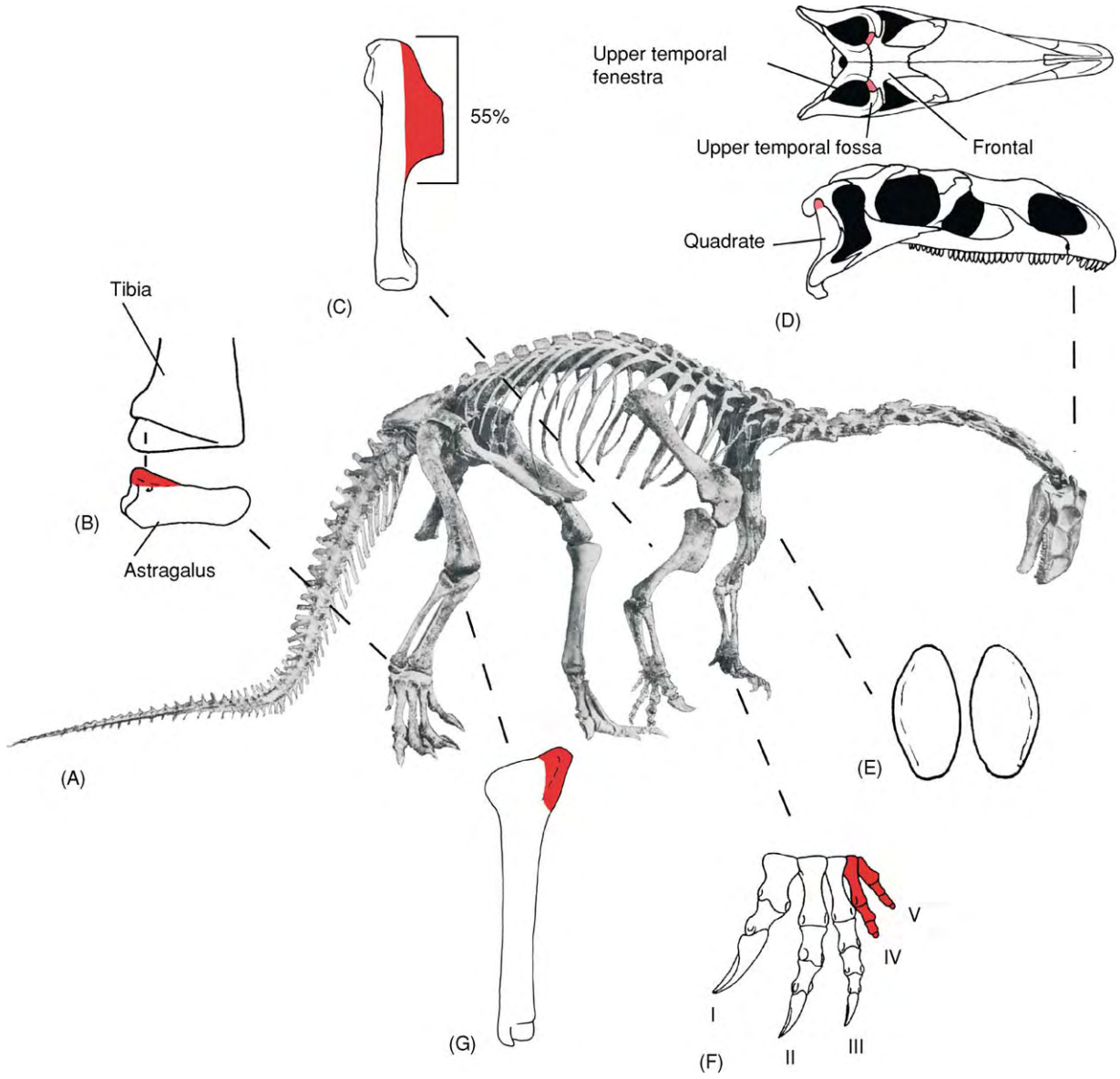


Figure 1 The diagnostic characters of the Dinosauria as shown by (A) the sauropodomorph *Plateosaurus*. (B) Distal part of the tibia and astragalus in anterior view, showing the ascending process (red) and the notch it fits into. (C) Right humerus in lateral view, showing the enlarged (over 30% of humeral length) and rectangular deltopectoral crest (red). (D) Skull in dorsal (top) and lateral (bottom) views, showing the exposure of the frontal bone in the upper temporal fossa (top) and the lateral exposure of the quadrate head (bottom) in red. (E) Paired sternal plates. (F) Left hand in dorsal view, showing the reduced and clawless digits IV and V (red). (G) Right tibia in lateral view showing the cnemial crest (red).

Dinosaurs evolved two new plates of bone to cover the anterior end of the chest. These plates are paired and lie on each side of the midline (Figure 1E).

5. The deltopectoral crest of the humerus is an elongated rectangular crest that extends for more than 35% of the length of the humerus (Figure 1C).
6. The outer (fourth and fifth) digits of the hand are short and clawless. This character is particularly unusual and is strong evidence that dinosaurs are indeed a natural group. Reduction or loss of digits is a common event in the evolutionary history of terrestrial vertebrates; however, it almost always occurs in a symmetrical manner about the central (third) digit. Furthermore, the first digit to be reduced is the innermost (in the hand this is the thumb or pollex) because it is the last digit to form during the development of the embryo. The asymmetrical hand of dinosaurs with the area of reduction centred upon the outermost digits (the first to form during

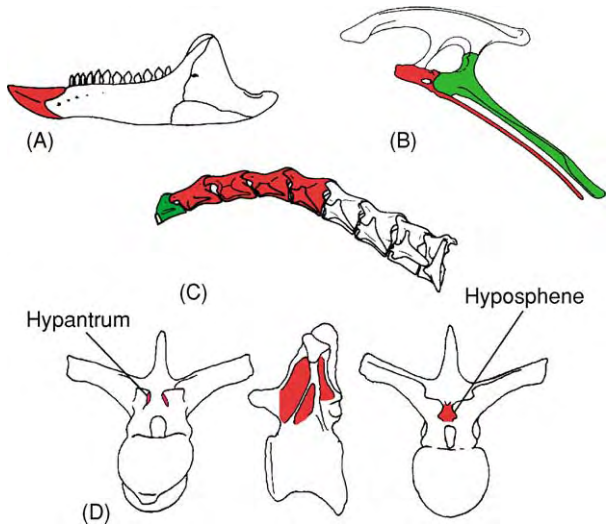


Figure 2 The diagnostic characters of (A, B) the Ornithischia and (C, D) the Saurischia. (A) Left lower jaw of an ornithischian in lateral view, showing predentary bone (red). (B) Pelvis of an ornithischian in left lateral view. The shaft of the pubis (red) has rotated backwards so that it lies parallel to the ischium (green). (C) The cervical series of a saurischian in left lateral view. The first four postaxial vertebrae (red) are elongate and exceed the axis (green) in length. (D) A trunk vertebra of a saurischian in (from left to right) anterior, left lateral, and posterior views. Red areas show the accessory articulations in the anterior and pos terior views, while the lateral view shows the large pneumatic fossae separated by thin laminae. Figures not to scale.

development) is unique among digit-bearing vertebrates (Figure 1F).

7. There is a cnemial crest. This is a vertical crest of bone that grows out of the anterior face of the proximal end of the tibia, just below the knee joint. It supported various lower-leg muscles and seems to be correlated with enhanced running ability. No other archosaurs have a cnemial crest, but other terrestrial vertebrates, such as cursorial mammals, do (Figure 1G).
8. The astragalus has an ascending process on its anterior edge. The astragalus is the larger of the two bones in the proximal row of the tarsus (ankle joint). Dinosaurs have a mesotarsal ankle, where there is a simple straight hinge running between the proximal and distal rows of bones in the tarsus. The ascending process is a triangular block of bone that fits into a corresponding notch in the distal tibia, thus preventing any movement in this part of the ankle (Figure 1B).

Origin of the Dinosauria

The closest known relatives of the dinosaurs are small slender avemetatarsalians from the Middle Triassic of South America (Figure 3A). These were relatively unspecialized and form good models for a dinosaurian precursor. *Marasuchus* is the best represented and most thoroughly studied of these animals. It was exceptionally small compared with its dinosaurian cousins, standing no more than 150 mm high at the hip and weighing no more than a few hundred grams. It was an agile long-legged carnivorous biped.

By the end of the Carnian Stage of the Late Triassic we have the first known remains of true dinosaurs. They are scattered around the world, but the best specimens come, once again, from South America. Though much larger than the earlier avemetatarsalians (*Herrerasaurus* weighed around 200 kg), the early dinosaurs were neither the largest terrestrial animals of their time nor particularly abundant in terms of numbers of species or individuals. Nevertheless, a modest radiation had occurred with carnivores (*Eoraptor, Staurikosaurus*, and *Herrerasaurus*), omnivores (*Saturnalia*), and herbivores (*Pisanosaurus*) all represented. Some palaeontologists have proposed that a small but significant extinction event hit many of the dominant groups of terrestrial vertebrates at the end of the Carnian, leaving many of the large animal niches vacant in the succeeding Norian Stage, but this hypothesis remains controversial. For whatever reason, dinosaurs dominated the large-herbivore niches around the globe in the Norian, and smaller carnivorous dinosaurs were common. Nevertheless, other tetrapod groups were still significant components of the large terrestrial faunas. These groups included the last of the dicynodont synapsids and various non-dinosaurian archosaurs, including the rauisuchians, which were the top predators of the time. However, by the beginning of the Jurassic Period these groups had become extinct and the ascendancy of the dinosaurs was complete. There may have been another mass extinction at the Triassic–Jurassic boundary, and there is some evidence, from eastern North America, for an extraterrestrial bolide impact at that time.

Dinosaur Subgroups

Ornithischia

The Ornithischia are a distinctive group, and there is no doubt that they are clade. Even the earliest ornithischians were specialized towards herbivory (although some of the earliest members may well have been somewhat omnivorous), and all subsequent species seem to have been herbivorous. Two of the most distinctive derived characters are the arrangement of the pelvic bones (Figure 2B) and the presence of an extra bone, the predentary (Figure 2A), in the lower jaw. In ornithischians the symphysis between

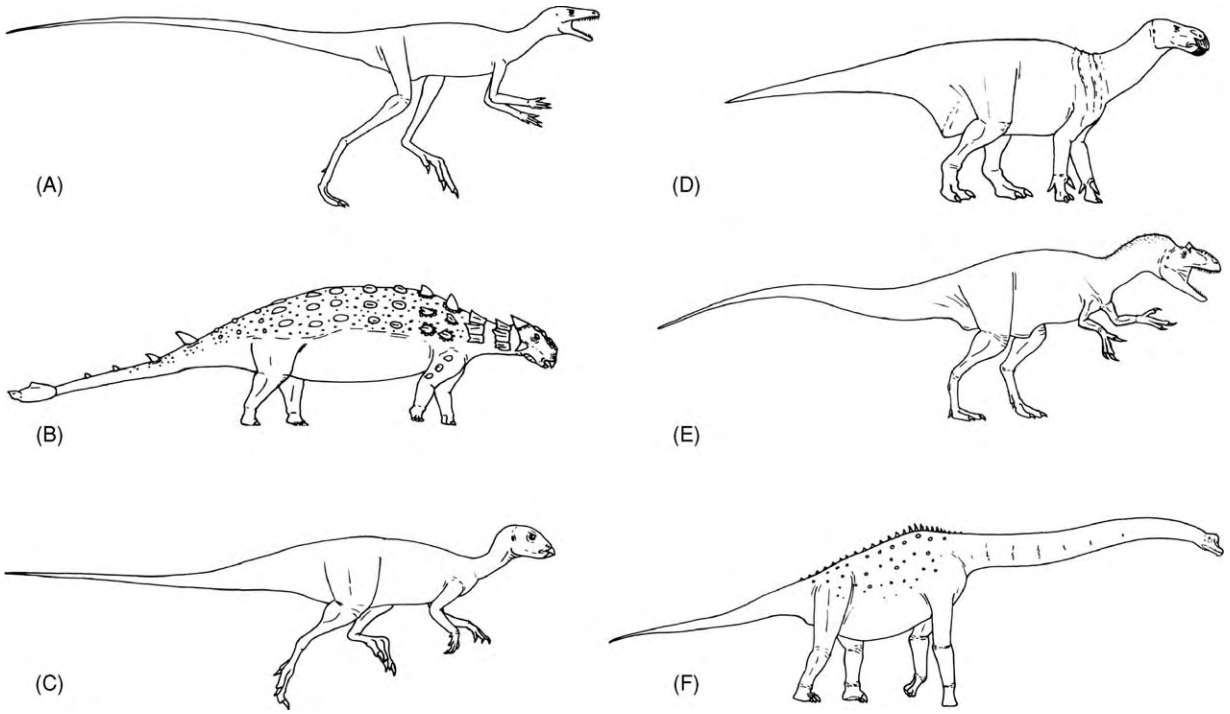


Figure 3 A sample of the diversity of dinosaurs and their closest relatives. (A) *Marasuchus*, a dinosaurian precursor (total length of 0.5 m). (B) *Saichania*, an ankylosaurid thyreophoran (total length of 6.5 m). (C) *Hypsilophodon*, a neornithischian that retained the primitive ornithischian body plan (total length of 1.5 m). (D) *Iguanodon*, a derived neornithischian (total length of 11 m). (E) *Allosaurus*, a typical theropod (total length of 8 m). (F) *Brachiosaurus*, a sauropod (total length of 21 m). Figures not to scale.

the pubic bones is interrupted and each pubis is rotated backwards to extend parallel to the ischia. This may have allowed the gut to expand backwards into the space between the legs, thus increasing its length (to deal with poorly digestible plant material) without shifting the centre of gravity forwards. The position of the centre of gravity is important as the early ornithischians retained the bipedal locomotion of their ancestors. Another feeding adaptation was the development of a predentary bone at the front of the mandible. This acted as a stable single-piece platform to support a horny beak while freeing the two mandibular rami to rotate slightly around their long axes, thus increasing the amount of tooth-on-tooth contact during chewing.

Ornithischians were a rare group during the Triassic. It was only in the Jurassic that they became abundant and diverse. These more successful ornithischians comprise two large clades, the Thyreophora and the Neornithischia.

Thyreophora Thyreophorans abandoned a bipedal cursorial lifestyle for a slower quadrupedal existence. They retained a primitive simple set of teeth and jaws and clearly relied heavily on their capacious guts to deal with their plant diet. Instead of using speed to escape predation, thyreophorans evolved a protective coat of keratin-covered dermal bones, or osteoderms, arranged in longitudinal rows. In early forms the osteoderms were simple studs, but they became highly elaborated in the two main thyreophoran groups: the stegosaurs and the ankylosaurs. In the stegosaurs the rows of osteoderms on each side of the dorsal midline were enlarged into a series of large plates and spines at the expense of the rest of their armour. In contrast, ankylosaurs increased the number of osteoderms to create a protective pavement that covered their dorsal surface (**Figure 3B**). One Late Cretaceous group, the Ankylosauridae, developed a large bony club at the end of the tail, which could be wielded like a mace.

Neornithischia Unlike the thryreophorans, the neornithischians evolved a number of specializations of their jaws and teeth for improved oral processing of plant matter. These included a larger cropping beak at the front of the jaw (indicated by an enlarged predentary bone), a diastema between the beak and the cheek teeth, and a thick layer of enamel covering one side of the 'cheek' teeth. This last feature creates a ridge of hard enamel that juts above the dentine when the tooth is worn down. Thus, a rougher file-like chewing surface is developed for efficient processing of plant matter. Apart from these features, the

neornithischians display a great diversity of body plans. Early forms were small and bipedal, superficially resembling basal ornithischians (Figure 3C). Some primitive taxa (e.g. *Thescelosaurus*) kept this conservative body plan right through to the end of the Cretaceous. Two separate clades, the Neoceratopsia and the Iguanodontia (Figure 3D), continued to improve the oral processing of their food and culminated in the Ceratopsidae and the Hadrosauridae, respectively. Members of these clades became important herd-forming herbivores in the Late Cretaceous. Both developed batteries of hundreds to thousands of continuously replaced teeth to grind their food efficiently. They also developed large apparently species-specific display structures (frills and horns in the ceratopsids, expanded nasal crests in the hadrosaurids), which may have been significant for maintaining social structure within their herds.

Saurischia

At first sight saurischians are less obviously a natural group than the ornithischians. The 'lizard-like' hip is a primitive feature that is also found in many other reptiles and is not evidence of a close relationship. Nevertheless, there are several derived characters that unite the saurischians together as a clade. Many of them are modifications of the vertebral column. These include a lengthening of the anterior cervical (neck) vertebrae (Figure 2C), the presence of extra articulations between the vertebrae of the trunk (called the hyposphene–hypantrum system; Figure 2D), and the excavation of the vertebrae by large hollows that are bounded by thin webs or struts of bone (Figure 2D). The additional articulations in the trunk vertebrae stiffen the vertebral column and reduce flexibility, while the longer cervical vertebrae lengthen the neck. Air-filled outpockets of the lungs occupy the vertebral excavations in the only surviving saurischian group (birds), and it is reasonable to assume that this was also the case in extinct saurischians.

Two large clades arose from within the Saurischia: the largely carnivorous Theropoda and the predominantly herbivorous Sauropodomorpha. There is much debate about whether some of the very earliest saurischians (Herrerasauridae and *Eoraptor*) were early examples of the Theropoda or represent early branches of the saurischian tree that diverged before the split between Theropoda and Sauropodomorpha.

Sauropodomorpha Sauropodomorphs were saurischians characterized by long necks, small heads, and leaf-shaped teeth. These adaptations indicate that they were eating significant amounts of plant matter. However, early examples (which are sometimes placed in a clade of their own, the Prosauropoda) lack specialized chewing mechanisms, tooth-on-tooth contact, and large guts to deal with a strict diet of foliage. Thus, it seems likely that they fed on softer highly nutritious parts of plants (new shoots, fleshy reproductive parts, etc.), perhaps supplemented with small prey items and carrion. These early sauropodomorphs also had a generalized skeleton that could accommodate both bipedal and quadrupedal locomotion. In contrast, the Sauropoda, which first appeared in the Late Triassic but did not rise to prominence until the Middle Jurassic, consisted of strictly herbivorous obligate quadrupeds (Figure 3F). The Sauropoda are justly famous for producing the largest land-dwelling animals of all time. Even the smallest species weighed several tons and reached 10 m in length. The largest, for which we have only a few bones, may have reached 40 m in length and weighed up to 100 tons. They had massively constructed pillar-like limbs and an exceptionally small head relative to body size, which was perched atop an extremely elongated neck. They are often characterized as having very weak dentition, but their oral processing was more extensive than that of prosauropods, with broad jaws, precise tooth-on-tooth contact, and improved jaw musculature (shown by the development of a tall coronoid process in the lower jaw). Sauropod diversity in the northern continents declined in the Early Cretaceous, whereas they appear to have remained the dominant large herbivores in South America and Africa. By the Late Cretaceous titanosaurian sauropods were the dominant large herbivores in most parts of the world except North America and Asia. These were robust, and sometimes armoured, sauropods with distinctive articulations between the tail vertebrae. The group includes both the largest and some of the smallest sauropods.

Theropoda Most Mesozoic species retained the predatory bipedal way of life that they inherited from the earliest dinosaurs. Most were long-legged hunters with large skulls and serrated blade-like teeth (Figure 3E). Despite their outwardly conservative body form, the group shares a great number of anatomical specializations that make it clear that they are a clade. These include a deep, probably pneumatic, recess on the ventral surface of the braincase, a distally stiffened tail (forming a dynamic stabilizer), fusion of the clavicles into a furcula ('wishbone'), complete loss of the fifth finger, and a splint-like first metatarsal that carries a functional toe but does not reach the ankle joint. Although it is clear that almost all theropods were carnivorous, it is difficult to find conclusive evidence that they were predators as opposed to scavengers of carrion. Nevertheless, all modern terrestrial

carnivores include the hunting and killing of prey in their behavioural repertoire, and it is most unlikely that theropods did not. The strong build of their skulls and their grasping hands seem adapted for grappling with powerful struggling prey, supporting this idea. One clade of derived theropods, the Coelurosauria, produced a range of divergent body plans. This radiation probably began in the Middle Jurassic, but coelurosaurs did not become abundant until the Cretaceous. The group included gigantic carnivores (tyrannosaurids), toothless presumed herbivores (ornithomimosaurs and oviraptorosaurs), and long-necked herbivores (therizinosaurs). The radiation also produced the sickle-clawed deinonychosaurs and their close relatives the flying dinosaurs or birds.

Bird Origins

Dinosaurs did not become entirely extinct at the end of the Cretaceous. One highly specialized group of theropods, the birds, survived and are still a major part of modern vertebrate faunas (*see* **Fossil Vertebrates:** Birds). The great preponderance of evidence from skeletal anatomy, eggshell microstructure, and integumentary structures indicates that the origin of birds lies deep within the theropod clade, with the Deinonychosauria being their closest relatives. These two groups share a number of significant similarities, including a decoupling of the tail from the muscular system driving the hind legs (shown by the extreme mobility of the articulations of the proximal tail vertebrae and the reduction of the bony attachment sites for the caudofemoralis muscle), retroversion of the pubis (similar to the condition in Ornithischia), and a laterally facing shoulder socket that allows a flapping motion of the forelimb. The caudofemoralis muscle is the main leg-retracting muscle in modern reptiles and extends from the posterior surface of the femur to the side of the tail, keeping the two organs tightly coupled during locomotion.

The most compelling evidence for the dinosaur ancestry of birds is the discovery of numerous species of non-avian coelurosaur that are preserved with a covering of feathers in the Early Cretaceous rocks of the Liaoning Province, China. These range from simple filaments in the primitive *Sinosauropteryx* to complex vaned feathers (with a central rachis and lateral barbs) in more derived species such as *Caudipteryx*, *Protarchaeopteryx*, *Sinornithosaurus*, and *Microraptor.*

The environment in which birds evolved and the possible behaviours and selection pressures that led to the origin of flight are more difficult to elucidate. Critics of the dinosaur hypothesis for the origin of birds point to the improbability of flight evolving in a ground-dwelling runner. Nevertheless there is no logical reason why some dinosaurs could not have adopted a scansorial (climbing) existence. Indeed, some of the newly discovered small deinonychosaurs from Liaoning seem to show some scansorial adaptations.

Physiology

No other area of investigation into dinosaur biology has produced as much controversy as the issue of their metabolic and thermal physiology. Although the debate is often portrayed as a simple dichotomy between 'warm-blooded' and 'cold-blooded' dinosaurs, the issue is more complicated.

Few doubt that dinosaurs were capable of sustained aerobic exercise and that they maintained at least some degree of homeothermy (stable body temperatures); these are not typical characteristics of modern ectotherms (animals that require an external heat source to reach their optimum body temperature). Evidence for this comes from their erect gaits, their frequent cursorial adaptations, and the fibrolamellar microstructure of their bones (indicating fast growth). The question remains, did they achieve homeothermy through an elevated resting metabolic rate (tachymetabolism) and the internal heat that it generates (endothermy), or did they use other means, such as the thermal inertia resulting from their large size and the fairly equable climate of the Mesozoic? Most modern ectotherms cannot maintain homeothermy (a condition known as poikilothermy) but there are exceptions, such as the leatherback turtle (*Dermochelys*).

Most of the evidence that has been used to support endothermy in non-avian dinosaurs either implies homeothermy, but not necessarily endothermy, or simply cannot be trusted owing to the distortions caused by the processes of taphonomy and diagenesis (e.g. predator–prey ratios and stable isotope signatures in bone). The evidence for ectothermy is also weak, although it has been claimed by some researchers that the absence of nasal passages large enough to house respiratory turbinate bones is the 'Rosetta Stone' that demonstrates dinosaurian ectothermy. However, the recent discovery of small coelurosaurs that were insulated by feathers means that it is almost certain that at least these dinosaurs were endotherms. Since coelurosaurs are included amongst the dinosaurs that have been claimed to be ectotherms by virtue of their small nasal passages, it would now seem that nasal-passage size and thermal physiology are not as tightly correlated as the proponents of this hypothesis contend.

Reproduction and Growth

Like all other archosaurs, dinosaurs laid calcitic-shelled eggs. Eggs have been found that (owing to the presence of embryos) can be attributed to theropods, neornithischians, titanosaurian sauropods, and prosauropods. Many other egg types have been found, and these probably represent all the major dinosaur groups. However, matching an egg type to a producer without direct evidence from embryonic remains is fraught with difficulty. An intriguing feature of non-avian dinosaur eggs is their relatively small size. The largest non-avian dinosaur eggs are about 4 l (laid by gigantic sauropods) and are actually smaller than the largest bird eggs, which are about 9 l (laid by the recently extinct elephant bird). A further difference is the average number of eggs laid in a single clutch. Fossilized non-avian dinosaur clutches usually contain tens of eggs (up to 40), but the actual average was probably higher, as fossil clutches are almost always partly destroyed by erosion when discovered. Avian clutch sizes can be as low as one and are usually less than ten. Further differences are that non-avian dinosaurs partly buried their eggs and apparently did not turn them once laid. However, oviraptorosaur skeletons found on top of their nests show that they did assume a bird-like brooding posture over their eggs. Furthermore, hadrosaurids show evidence of post-hatching parental care. The degree of ossification of the articular ends of the baby limb bones found inside nest structures indicates that they were nest-bound and must have relied on food bought back to them by their parents.

Owing to the great size difference between adult and hatchling dinosaurs (three, perhaps as much as four, orders of magnitude), young dinosaurs must have grown rapidly, which is borne out by the fast-growth-style microstructure of their bones. Estimates of actual growth rates have been obtained by matching the histological type of fossil taxa to modern taxa of known growth rate or by counting lines of arrested growth (LAGs), which are assumed to be seasonal. The range of growth rates that can produce each particular type of bone histology has not been adequately explored, and recent evidence suggests that the range is broader than previously suspected. Thus, estimates based on the first method are suspect. LAGs have proved even more unreliable, with the number of LAGs varying widely between different bones of a single individual. Experimental work has shown that LAG formation is not necessarily correlated with seasonal hardship. Until the biological meaning of LAGs is better understood, they should not be used to determine growth rates in dinosaurs.

See Also

Fossil Vertebrates: Reptiles Other Than Dinosaurs; Birds; Flying Reptiles. **Mesozoic:** Triassic; Jurassic; Cretaceous; End Cretaceous Extinctions.

Further Reading

Benton MJ (1983) Dinosaur success in the Triassic: a non competitive ecological model. *Quarterly Review of Biology* 58: 29 55.

Farlow JO and Brett Surman MK (1997) *The Complete Dinosaur*. Bloomington: Indiana University Press.

Farlow JO, Dodson P, and Chinsamy A (1995) Dinosaur biology. *Annual Reviews of Ecology and Systematics* 26: 445 471.

Horner JR (2000) Dinosaur reproduction and parenting. *Annual Reviews of Earth and Planetary Science* 28: 19 45.

Ji Q, Currie PJ, Norell MA, and Ji S A (1998) Two feathered dinosaurs from northeastern China. *Nature* 393: 753 761.

Olsen PE, Kent DV, Sues H D, *et al.* (2002) Ascent of dinosaurs linked to an iridium anomaly at the Triassic Jurassic boundary. *Science* 296: 1305 1307.

Padian K, de Ricqlès AJ, and Horner JR (2001) Dinosaurian growth rates and bird origins. *Nature* 412: 405 408.

Sereno PC (1997) The origin and evolution of dinosaurs. *Annual Reviews of Earth and Planetary Science* 25: 435 489.

Sereno PC (1999) The evolution of dinosaurs. *Science* 284: 2137 2147.

Starck JM and Chinsamy A (2002) Bone microstructure and developmental plasticity in birds and other dinosaurs. *Journal of Morphology* 254: 232 246.

Weishampel DB, Dodson P, and Osmolska H (1990) *The Dinosauria*. Berkeley: University of California Press.

Birds

G J Dyke, University College Dublin, Dublin, Ireland
L M Chiappe, Natural History Museum of Los Angeles County, Los Angeles, CA, USA

Introduction

The evolutionary history of birds – class Aves – and their relationships with other groups of fossil and living vertebrates have been debated for more than 100 years. Now, however, the overwhelming consensus among palaeontologists and zoologists is that all living birds and their fossil cousins are the descendants of carnivorous theropod dinosaurs; indeed, evolutionary hypotheses place Aves as a nested group within this subdivision of the dinosaurian family tree (Figure 1). As a group, birds are characterized by their active flapping flight and have an almost uninterrupted fossil history from the Late Jurassic (140 Ma) to the Holocene. The earliest phases of their evolution occurred in the Cretaceous, possibly including the origination of modern birds (Neornithes), whose extant orders and families are first recorded in the latest Cretaceous–earliest Tertiary, some 60–70 Ma ago. Throughout the Cenozoic birds radiated (diverged and diversified) to produce the approximately 10 000 species alive today. In this article, we discuss many of the key fossil taxa that have aided our understanding of avian evolution throughout the Mesozoic, outline the geological environments in which they are preserved, and present arguments for the pattern of the evolutionary radiation of modern birds. Although birds are among the most familiar and immediately recognizable of the major living vertebrate groups, very little was known about their early evolutionary history until just a few years ago – recent times have been very exciting for Mesozoic avian palaeontologists: in just the last five years, more fossil bird taxa have been discovered and described from the Cretaceous alone than were known for much of the preceding century.

Birds from the Mesozoic: Not just *Archaeopteryx*

Although it was described nearly 150 years ago from the lagoonal limestones of Solnhofen in Bavaria (Germany), *Archaeopteryx* is still the oldest recognized bird. Eight skeletal specimens and a feather are all that is known of this primitive taxon; all have been found in the 140 Ma old Solnhofen limestone deposits, which were formed in a poorly oxygenated hypersaline shallow tropical lagoon, thus facilitating the exquisite preservation of these specimens. A great deal has been written about *Archaeopteryx*. – How well was it able to fly? Was it an arboreal or a terrestrial animal? Do all known specimens come from one or several closely related species? Since its initial description in the 1860s, one thing has remained clear: *Archaeopteryx* provides a tantalizing glimpse of the earliest stages of avian evolution. Having a long bony tail, sharply clawed forelimbs, and a primitive pelvis, this toothed bird is in many ways anatomically more similar to theropod dinosaurs than to today's birds. At one time *Archaeopteryx* was one of very few fossils from the Mesozoic upon which solid evolutionary inferences could be based; however, over the course of the last two decades, more and more fossil information has come to light, revealing an unexpectedly large diversity of primitive birds that existed throughout the last half of the Mesozoic Era. These fossil birds vary vastly in age, degree of preservation, and the environments they inhabited. Although none of these newly discovered lineages have left living descendants, the many exquisite fossils allow us to reconstruct patterns of morphological change, including the origin and development of modern avian wing anatomy. They also furnish evidence documenting the pattern of genealogical relationships of early birds and help us better understand the radiation of extant species.

In addition to *Archaeopteryx*, new fossils that have come to light over the last few years include several of the other most primitive birds known, *Jeholornis* and *Zhenzhoraptor* (which are perhaps synonymous taxa) and *Rahonavis*, from the Early and Late Cretaceous of China and Madagascar, respectively. Although both these taxa retain the long bony tails of *Archaeopteryx* and other non-avian theropods, they show that more advanced shoulder girdles and wings (which characterize more derived birds) had begun to develop by the earliest Cretaceous (*ca*. 125 Ma ago) (Figure 2). *Rahonavis* is remarkably theropod-like in its anatomy, even retaining the enlarged sickle claw seen on the feet of theropods such as dromaeosaurids. Even though *Rahonavis* is much younger than *Archaeopteryx* (more than 60 Ma younger), the two birds may be closely related.

Fossil discoveries of birds at the base of the avian evolutionary tree, close in their anatomy to *Archaeopteryx*, have characterized some, but not all, of the

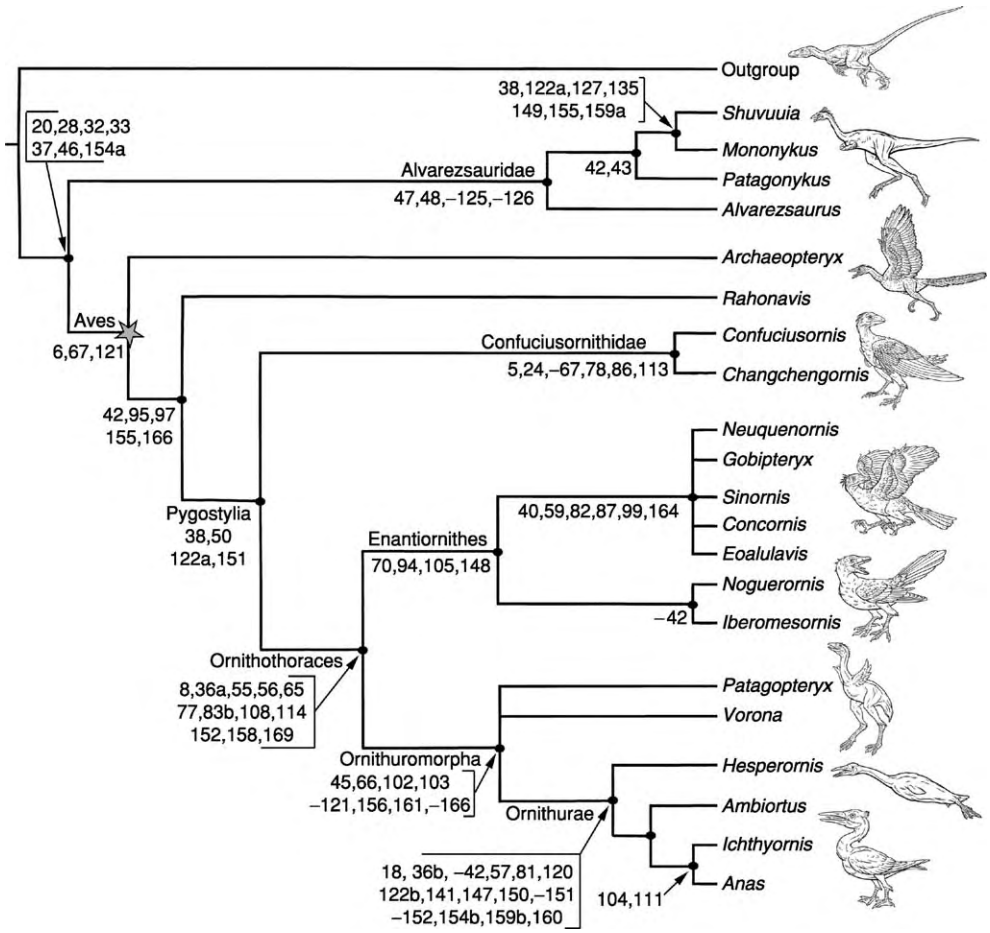


Figure 1 Cladogram of Mesozoic bird relationships showing theropod dinosaurs at the base of the tree. Many of the taxa seen at the tips of this tree have been described since 1990. Character numbers and names of nodes are taken from Chiappe (2001).

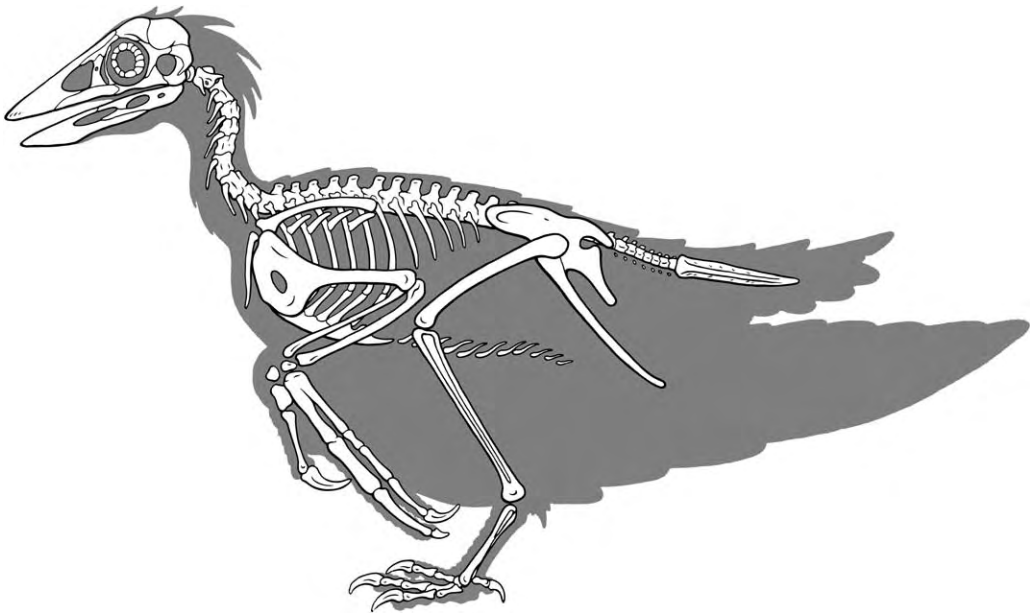


Figure 2 Artist's rendering of *Confuciusornis*, a primitive fossil bird from the Cretaceous of China.

advances in the study of avian evolution. Arguably the most significant development in this area came in the early 1980s with the discovery of an entirely new group of Cretaceous birds, the subclass Enantiornithes. More than 20 genera of Enantiornithes have now been discovered and described from all kinds of environments and from throughout the Cretaceous (Table 1). It is important to remember, however, that these birds were recognized only in 1981, from the Late Cretaceous of Argentina. The Enantiornithes were by far the most diverse group of Mesozoic birds, showing a good deal of anatomical variation, as well as diversity in their feeding adaptations, flight apparatus, life styles, and patterns of skeletal development. Representatives of this lineage are now particularly well known from the Early Cretaceous of Spain and China. Although most were toothed (e.g. *Sinornis*, *Eocathayornis*, and *Logipteryx*), some did not have teeth (*Gobipteryx*), and they range in size from species that were the size of a sparrow (*Iberomesornis*, *Eoalulavis*) to taxa with wingspans of almost a metre (*Enantiornis*). Their anatomy demonstrates that Enantiornithes were proficient fliers (probably very similar to living birds) and, interestingly, provides some of the earliest evidence for perching (based on the morphology of their feet) that has been found in avian evolution. Although these birds have mostly been recovered from inland deposits, enantiornithines also occupied littoral and marine environments, even extending into the polar regions. Enantiornithes are an anatomically distinctive group (having unique morphological specializations in their forelimbs, hindlimbs, and pectoral girdles) but are not closely related to extant birds (Figure 1); these taxa comprise a basal series of lineages that did not survive the end-Cretaceous extinction event.

In addition to these diverse archaic birds, a number of other currently less well-known taxa (in terms of numbers of described fossil specimens) comprise a series of evolutionary intermediates that fits neatly in between the enantiornithine radiation and the divergence of a second major avian group, the Ornithuromorpha (Figure 1). Most importantly, this clade contains the immediate relatives of all living birds. Members of the Ornithuromorpha are much more modern in their morphology than are members of the other Mesozoic groups – their earliest records are from the Early Cretaceous of China. Just as in the Enantiornithes, there has been something of an explosion in fossil discoveries of ornithuromorphs in recent years. The exceptionally well-preserved *Apsaravis* from the Late Cretaceous of the Gobi Desert in Mongolia is one example (Figure 3). The bulk of the early fossil record of these birds was limited to the flightless loon-like *Hesperornis* and its kin (Hesperornithiformes) and the more modern looking *Ichthyornis* (Ichthyornithiformes) from Late Cretaceous marine sediments in the northern hemisphere. Although much more derived in their anatomy, these birds were described soon after *Archaeopteryx* in the 1870s from rocks formed in the Western Interior Seaway, a shallow tropical sea that bisected North America during the Late Cretaceous. Both are toothed, and *Hesperornis* is known to have been a specialized foot-propelled diver with extremely abbreviated forelimbs (about the size of an emperor penguin), whereas *Ichthyornis* was much smaller and able to fly.

The Radiation of Modern Birds: Bursting into the Cenozoic

Neornithes is the group that includes the 10 000 or so living species of bird. Today, they are a diverse and cosmopolitan group, but their early evolutionary history remains far from well understood. Although *Ichthyornis* is the best-known close relative of the Neornithes, several other recently described taxa might be closer. These taxa, however, are known only from extremely fragmentary remains, thus rendering problematic our understanding of their genealogical relationships to the Neornithes. In spite of this, debates have centred on the timing of the origination of the major lineages of Neornithes (the extant orders and families), and specifically on whether or not these taxa differentiated as a group prior to the end-Cretaceous extinction event 65 Ma ago and, if so, how deep does the history of modern birds extend into the Mesozoic?

Evidence to support or refute hypotheses for the time of divergence of the Neornithes is entirely geological, but the fossil record of putative modern birds from the Cretaceous remains scanty. Just a handful of fossil specimens have been described from before the Creataceous–Tertiary boundary and classified within modern groups; they are mostly single bones that lack clear diagnostic features of extant lineages. It is in the earliest Tertiary that the known fossil record of Neornithes improves dramatically: hundreds of fossils (in many cases complete skeletons, often with feathers and other soft-tissue impressions) are known from a series of localities of Paleocene and Eocene age (60–55 Ma), particularly in Europe and North America, and from a range of environments. This abundance of fossil birds that are modern in their anatomy has led some workers to propose that the bulk of their evolutionary radiation occurred rapidly, immediately after the Cretaceous–Tertiary extinction event (Figure 4). Related to this, there is a further hypothesis that perhaps the extinction itself allowed

Table 1 Fossil Mesozoic birds, their ages and environments (Chiappe and Dyke, 2002)

Taxon	*Stratigraphical age*	*Material*	*Depositional environment*	*Geographical distribution*	*Year described*
Alexornis antecedens	Late Cretaceous	Single specimen	Inland	Mexico	1974
Alvarezsaurus calvoi	Late Cretaceous	Single specimen	Inland	Argentina	1991
Ambiortus dementjevi	Early Cretaceous	Single specimen	Inland	Mongolia	1982
Apatornis celer	Late Cretaceous	Single specimen	Marine	USA	1876
Apsaravis ukaani	Early Cretaceous	Single specimen	Inland	Mongolia	2001
Archaeopteryx lithographica	Late Jurassic	Several specimens	Near shore	Germany	1861
Archaeovolans repatriatus	Early Cretaceous	Single specimen	Inland	China	2002
Avisaurus archibaldi	Late Cretaceous	Single bone	Inland	USA	1985
Avisaurus gloriae	Late Cretaceous	Single bone	Inland	USA	1995
Baptornis advenus	Late Cretaceous	Several specimens	Marine	USA	1876
Boluochia zhengi	Early Cretaceous	Single specimen	Inland	China	1994
Changchengornis hengdaoziensis	Early Cretaceous	Single specimen	Inland	China	1999
Chaoyangia beishanensis	Early Cretaceous	Single specimen	Inland	China	1995
Concornis lacustris	Early Cretaceous	Single specimen	Inland	Spain	1995
Confuciusornis sanctus	Early Cretaceous	Hundreds of specimens	Inland	China	1996
Enaliornis barretti	Early Cretaceous	Several specimens	Marine	UK	1876
Enantiornis leali	Late Cretaceous	Several specimens	Inland	Argentina	1976
Eoalulavis hoyasi	Early Cretaceous	Single specimen	Inland	Spain	1992
Eoenantiornis buhleri	Early Cretaceous	Single specimen	Inland	China	1999
Gansus yumensis	Early Cretaceous	Isolated bones	Near shore	Canada	1997
Gobipteryx minuta	Late Cretaceous	Several specimens	Inland	Mongolia	1996
Halimornis thompsoni	Late Cretaceous	Single specimen	Near shore	USA	2001
Hesperornis regalis	Late Cretaceous	Several specimens	Marine	USA	1876
Iberomesornis romerali	Early Cretaceous	Single specimen	Inland	Spain	1993
Ichthyornis dispar	Late Cretaceous	Several specimens	Marine	USA	1873
Lectavis bretincola	Late Cretaceous	Single bone	Inland	Argentina	1993
Liaoningornis longidigitus	Early Cretaceous	Single specimen	Inland	China	1996
Limenavis patagonica	Late Cretaceous	Single specimen	Inland	Argentina	2001
Longipteryx chaoyangensis	Early Cretaceous	Single specimen	Inland	China	2001
Mononykus olecranus	Late Cretaceous	Several specimens	Inland	Mongolia	1993
Nanantius eos	Early Cretaceous	Single bone	Inland	Australia	1986
Neuquenornis volans	Late Cretaceous	Single specimen	Inland	Argentina	1994
Noguerornis gonzalezi	Early Cretaceous	Single specimen	Inland	Spain	1992
Otogornis genhisi	Early Cretaceous	Single specimen	Inland	China	1992
Parahesperornis alexi	Late Cretaceous	Single specimen	Marine	USA	1989
Parvicursor remotus	Late Cretaceous	Single specimen	Inland	Mongolia	1996
Patagonykus puertai	Late Cretaceous	Single specimen	Inland	Argentina	1996
Patagopteryx deferrarisi	Late Cretaceous	Single specimen	Inland	Argentina	1992
Protopteryx fengningensis	Early Cretaceous	Single specimen	Inland	China	2000
Rahonavis ostromi	Late Cretaceous	Single specimen	Inland	Madagascar	1998
Shuvuuia deserti	Late Cretaceous	Several specimens	Inland	Mongolia	1998
Sinornis santensis	Early Cretaceous	Several specimens	Inland	China	1992
Soroavisaurus australis	Late Cretaceous	Single bone	Inland	Argentina	1993
Vorona berivotrensis	Late Cretaceous	Single specimen	Inland	Madagascar	1996
Yanornis martini	Early Cretaceous	Single specimen	Inland	China	2001
Yixianornis grabaui	Early Cretaceous	Single specimen	Inland	China	2001

the repatriation of avian ecological niches that were occupied by more archaic birds during the Cretaceous. These ideas are based on literal readings of the known fossil record of Neornithes – small numbers of taxa (perhaps misidentified) from the Cretaceous, compared with a much larger diversity of definitive Neornithes from the Early Tertiary. Can this approach alone resolve the age-old issue of modern-bird evolutionary dynamics?

Literal interpretations of the fossil record to explain the pattern of the radiation have, however, been challenged by a number of other lines of evidence, notably the genealogical interpretation of living birds based on genetic data. By considering the numbers of

Figure 3 The single known specimen of the well preserved ornithuromorph *Apsaravis* from the Late Cretaceous of the Gobi Desert, Mongolia (reproduced with the kind permission of Mark Norell and the American Museum of Natural History, New York).

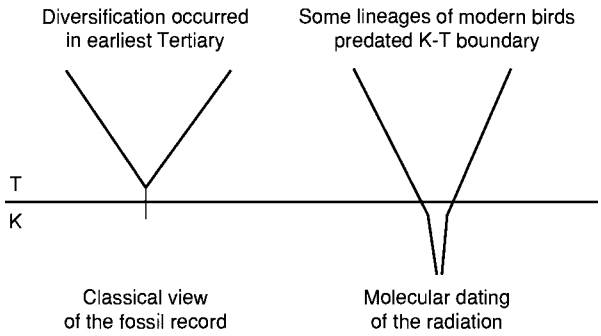


Figure 4 Two competing (alternative) hypotheses for the pat tern of the radiation of Neornithes (modern birds). The one on the left is based on fossil data; the one on the right is based on molecular data.

differences in the aligned DNA sequences (both nuclear and mitochrondrial) of living birds and by using either a fossil of known systematic position or a well-dated continental split as a calibration point, new ideas about the timing of the divergence of the major lineages of living birds have been developed, based on the so-called 'molecular clock'. Divergence estimates have been made for a number of groups of living birds and in all cases have indicated that the radiation of the Neornithes occurred much earlier than has been inferred from the known fossil record (Figure 4). 'Molecular clock' estimates vary, but have reached a consensus that the majority of the major lineages of living birds did originate sometime in the Cretaceous. It has been argued that these hypotheses, although currently receiving little support from the fossil record, agree with what is known about the pattern of breakup of the continental landmasses during the Mesozoic. A causal correlation between the breakup of landmasses and the differentiation of certain groups of extant birds (i.e. the occurrence of endemic southern-hemisphere taxa towards the base of the neornithine tree) has been proposed, and there is a general agreement among molecular systematists that the initial radiation of the Neornithes probably occurred in the southern hemisphere during the Mesozoic.

There is thus a discrepancy between the results of studies founded on molecular data and the apparent pattern seen in the fossil record (Figure 4). Despite a large number of exceptional fossil-bearing deposits from the Late Cretaceous (especially in the northern hemisphere), convincing remains of modern birds have yet to be found. Opinions about the apparent absence of these birds vary: while some have proposed that it probably reflects a real evolutionary pattern (fossils of other small vertebrates, such as mammals and lizards, are well-known from the Cretaceous), others have suggested that perhaps palaeontologists have not been looking in the right place (relatively little collecting effort has been made in the Cretaceous rocks of the southern hemisphere). What is clear is that the existing records of Cretaceous modern birds should be treated with caution – the earliest neornithine birds that are complete enough to be informative for cladistic analyses, and hence potentially informative for estimating the temporal divergence of the extant lineages, come from rocks that are roughly 55–60 Ma old, deposited some 10 Ma after the end of the Cretaceous.

Where From Here?

Thus, palaeontological evidence accumulated over the last decade has cemented the notion that birds are the living descendants of small carnivorous theropod dinosaurs. Exceptional new finds of Cretaceous fossils, especially from China, Spain, Argentina, and Mongolia, have documented a remarkable diversity of avian lineages that thrived and diversified during the latter stages of the Mesozoic. These discoveries have filled substantial anatomical and genealogical gaps in the early history of birds. Yet the evolutionary dynamics (time of divergence and radiation patterns) of the major lineages of extant birds remains an open question that can be resolved only by more well-preserved fossils and a better understanding of their evolutionary relationships.

See Also

Biological Radiations and Speciation. **Evolution**. **Fossil Vertebrates:** Dinosaurs; Flying Reptiles. **Mesozoic:** Cretaceous; End Cretaceous Extinctions.

Further Reading

Chatterjee S (1997) *The Rise of Birds*. Baltimore: Johns Hopkins University Press.

Chiappe LM (2001). Phylogenetic relationships among basal birds. In: Gauthier JD and Gall LF (eds.) *New Perspectives on the Origin and Early Evolution of Birds: Proceedings of the International Symposium in Honor of John H. Ostrom*. Yale Peabody Museum: New Haven.

Chiappe LM and Dyke GJ (2002) Mesozoic radiations of birds. *Annual Review of Ecology and Systematics* 33: 91 127.

Chiappe LM and Witmer LM (2002) *Mesozoic Birds: Above the Heads of Dinosaurs*. Berkeley: University of California Press.

Feduccia A (1999) *The Origin and Evolution of Birds*, 2nd edn. New Haven: Yale University Press.

Paul G (2002) *Dinosaurs of the Air: the Evolution and Loss of Flight in Dinosaurs and Birds*. Baltimore: Johns Hopkins University Press.

Swimming Reptiles

G W Storrs, Cincinnati Museum Center, Museum of Natural History and Science, Cincinnati, OH, USA

Introduction

Fossil reptile remains may be found in a variety of ancient depositional environments, but among the best preserved are those from marine settings. Consequently, fossil marine reptiles have occupied the interest of scientists, scholars, and naturalists from the earliest days of vertebrate palaeontological study. The first discoveries of such fossils were in the Mesozoic rocks of Europe, particularly in England, France, and Germany, in the late eighteenth and early nineteenth centuries. The occurrences of these fossil remains highlighted the relatively new concept of extinction and enabled the first palaeoenvironmental reconstructions, although the fossil discoveries predated the publication of both the theory of natural selection by Charles Darwin, in 1859, and the name Dinosauria, credited to Sir Richard Owen, in 1842. In spite of current popular belief, however, these marine-adapted reptiles were not dinosaurs. The major lineages of dinosaurs were, to the best of our knowledge, exclusively terrestrial in habit until the radiation of avian theropods (birds). Rather, marine reptiles are known to represent a variety of distinct adaptations to habitual life in the sea, and thus ocean-going lineages arose numerous times during the Mesozoic.

Marine Reptile Groups

A wide variety of fossil marine reptiles are known, most of which apparently originated in the Mesozoic. Indeed, most fossil marine lineages are restricted to the Mesozoic, excepting sea turtles and some minor faunal components such as sea snakes. However, a few swimming reptiles from freshwater and lagoonal environments are known from the Permian. *Claudiosaurus* of Madagascar and its relatives are notable for the controversial suggestion that they have an ancestral (or sister-group) relationship to the marine Sauropterygia. The Permian mesosaurs are famed for their role as palaeogeographical indicator fossils and evidence of 'continental drift' and a formerly united Gondwana (Figure 1). All marine reptiles, whatever their lineage, had terrestrial forebears, and thus their occupation of an aquatic environment is a return to the sea, or a secondary evolutionary adaptation. In returning to the sea, each species was faced with new physical constraints and functional needs requiring new adaptational solutions relative to their former life on land. Not the least of these adaptations was associated with the problems of aqueous locomotion, buoyancy control, food acquisition, and reproduction. How these problems were overcome often defines the disparate marine reptile groups.

Active subaqueous locomotion by vertebrate animals can be divided into two basic types: axial propulsion and paraxial propulsion. The former type is by far the most common and is characterized by sinuous flexion of the vertebral column. Constrained by evolutionary history, flexion is in the lateral plane in all reptiles that utilize this style of movement, as opposed to the vertical flexion of mammalian axial propulsors (e.g., cetaceans). Paraxial propulsion is accomplished through the use of the animal's limbs as the primary means of locomotion. Among Mesozoic axial propulsors are the ichthyosaurs, thallatosaurs, thalattosuchians, and mosasaurs. Paraxial swimmers include the sea turtles and the 'reptile wings' or Sauropterygia, i.e., the plesiosaurs and their kin, as defined by Richard Owen in

1860. The great majority of these animals functioned as 'habitual' swimmers; they were restricted to their role as oceanic creatures and few could return to the land, even for brief periods. As such, most probably gave birth to live young (i.e., were viviparous or ovoviviparous), having evolved beyond the need to lay a hard-shelled egg on land. Most of these animals ate fish or other reptiles, although some were apparently molluscivorous. Herbivory is an advanced dietary specialization and among marine reptiles is seen only in the turtles.

Figure 1 The Upper Permian aquatic reptile *Mesosaurus* of Brazil.

Axial Swimmers

The dominant marine group of the Triassic and Jurassic, at least in terms of numbers, was the Ichthyosauria, or 'fish lizards'. These species were also the most highly transformed from a terrestrial condition. Fossils dating to the Early Triassic indicate small, elongate, eel-like animals; by the Jurassic and Cretaceous, they had become powerful, streamlined swimmers with extremely derived, fish-like shapes (pisciform/thunniform). They ultimately developed a high dorsal fin and a symmetrically fluked tail supported by their down-turned, heterocercal vertebral column. Early discoveries in the English Liassic Series (Lower Jurassic) suggested to Owen in 1840 the presence of such caudal flukes. The flukes were later confirmed by discoveries of skeletons with surrounding organic films from the German Liassic (the famed Lagerstätte of Holzmaden and environs) (Figure 2). The limbs of ichthyosaurs were also much modified, ultimately becoming both hyperphalangic and polydactylous (i.e., with extra finger joints and digits). Used in manoeuvering, these paddles or fins became inflexible, and either long ('longipinnate') or broad ('latipinnate'). Ichthyosaurs had clearly lost their ability to leave the water early in their history.

Ichthyosaurs are first known from the Lower Triassic of Asia and survived until the Cenomanian (lower Upper Cretaceous). The Triassic mixosaurs were small and included animals with durophagous dentition (e.g., *Mixosaurus*), apparently for crushing nektonic cephalopods. The shastasaurids and cymbospondylids were giant, whale-like animals. As reported by Richard Ellis in 2003, a monstrous, 20-m-long ichthyosaur has been discovered in the Triassic of British Columbia, Canada. The Jurassic *Stenopterygius* and *Ophthalmosaurus* were among the most derived. The last records include fragmentary

Figure 2 The Lower Jurassic ichthyosaur *Stenopterygius* of Germany.

remains from Europe and North America when ichthyosaur abundance was seemingly a shadow of its former self.

Notable functional aspects of ichthyosaur anatomy include their obvious great speed and the possibility of an elevated physiology, and their well-developed eyes and vision, perhaps for finding prey at depth. Discoveries at Holzmaden of pregnant ichthyosaur (*Stenopterygius*) females, along with those killed in the act of giving birth, have demonstrated that these animals at least were viviparous. Ichthyosaurs in general are a classic example of convergent evolution, as is often cited in textbooks, whereby their habitus, or body shape, mimics that of fast teleost fish, sharks, and cetaceans.

The Thalattosauria ('sea reptiles') were a minor group of Triassic axial propulsors with rather more traditional reptilian proportions. *Askeptosaurus*, for example, from the Middle Triassic of the Monte San Giorgio region of Switzerland, had an elongate body and long, laterally compressed tail. Sinuous waves of the tail apparently propelled the animal through the water. The head had a pointed, elongate snout or rostrum, with numerous, needle-like teeth for capturing and holding slippery, struggling prey. The limbs were little modified from a terrestrial condition save for probable webbing of the feet. Similar animals are known from the west coast of North America. Aberrant, possibly bottom-dwelling forms such as *Clarazia* and *Heschelaria* may be related. All were near-shore, shallow-water animals, perhaps retaining some amphibious ability.

Thalattosuchians ('sea crocodiles') are derived Mesozoic crocodylomorphs (broadly, 'crocodilians', in older terminology). As catalogued by CW Andrews in 1913, like thalattosaurians, they also retained a broadly reptilian shape with elongate tail, body, and pointed rostrum. The powerful tail was the main propulsive organ and the limbs acted as directional controls. The Jurassic and Early Cretaceous teleosaurids (e.g., *Steneosaurus*) and metriorhynchids (*Metriorhynchus*, *Dakosaurus*) were extremely well adapted to life at sea. Indeed, at the height of their success, advanced metriorhynchids such as *Geosaurus* lost the bony armor of typical crocodyliforms and displayed modified, paddle-like limbs and heterocercal caudal tails, much in the manner of ichthyosaurs. Other crocodyliform groups such as the pholidosaurs (the Cretaceous *Teleorhinus* of North America) and the Cretaceous–Tertiary dyrosaurs (e.g., *Dyrosaurus*, *Hyposaurus*) dabbled with life in the ocean, but with less extreme morphological adaptation, retaining typical "crocodilian" limbs, for example. All possessed elongate, fish-catching rostra. The only crocodilian today that is able to tolerate fully oceanic waters, *Crocodylus porosus*, the estuarine or saltwater crocodile of Australasia, can hardly be called a true marine reptile because it is just as likely to be found well inland.

A very significant group of marine reptiles to utilize axial propulsion was the Mosasauroidea (**Figure 3**). Named for the Meuse River of north-western Europe, these animals were extremely numerous in the Late Cretaceous, to which they were restricted. Worldwide

Figure 3 The Upper Cretaceous mosasaur *Tylosaurus* of Kansas, as painted by Charles Knight.

in distribution, the mosasaurs were squamates related to today's lizards and snakes. Indeed, they appear to be super varanids – voracious predators that had abandoned life on land to dominate the seas. Mosasaurs, too, developed their fore- and hindlimbs into broad, webbed paddles to complement a tall, laterally compressed tail. Small, Early Cretaceous reptiles such as *Coniasaurus* and *Dolichosaurus* of Texas and England appear to be antecedents to the group. By the Late Cretaceous, giant animals such as the 10-m-long *Tylosaurus* and *Mosasaurus* devoured fish, cephalopods, other reptiles, and themselves. Mosasaurs are best known from the Smoky Hill Chalk beds (Niobrara Formation) of Kansas and adjacent states, but have been collected also from Europe, New Zealand, Africa, and Antarctica, for example.

An interesting functional characteristic of mosasaurs is their highly kinetic skull and unique jaw articulation. Hinged in the middle, the lower jaws were able to flex outwardly while rolling backwards on a movable skull/jaw joint (quadrate/articular). This apparently allowed large prey items to be systematically forced into the gullet even as they may have struggled to free themselves from the predator's grip. Escape of prey was generally prevented by a series of backwardly curving palatal teeth (pterygoid dentition). Interestingly, the V-shaped pattern of the marginal teeth and the curved lines of the palatal teeth have both been observed as a series of punctures on the fossilized shells of placenticerid ammonites, suggesting that these nektonic molluscs formed part of the mosasaurian diet. The crushing teeth of *Globidens* may have been a specialization for molluscivory. Mosasaurs probably bore live young.

A few living squamates may also be considered marine. The Galapagos Marine Iguana is physiologically well adapted to ocean waters, but is little transformed from terrestrial relatives and does not venture far from shore. It spends much of its time basking on coastal rocks but dives and forages underwater for its diet of sea grasses. Sea snakes (Hydrophiidae) do not come ashore and thus are marine reptiles in the truest sense. Extremely venomous and colourful, again with a vertically expanded, laterally compressed tail, they are a minor component of reptilian diversity and have no fossil record.

Paraxial Swimmers

Turning to paraxial swimmers, the familiar sea turtles have a long history from the Cretaceous to the Present. Clearly successful in their adaptation to oceanic life, they come ashore only to lay eggs in sandy beaches of the tropics. Ungainly and lumbering on land, they are graceful and far-travelled in water. Though most swimming reptiles have retained the sinuous lateral body movements of fish and early tetrapods, turtles have by necessity resorted to the dominant use of their forelimbs (paddles) for locomotion (**Figure 4**). Obviously, the stiffened thorax created by the presence of a hard, skeletal shell has constrained the direction of evolution in turtles. Prevented from flexing their bodies in any direction, sea turtles must swim with enlarged paddles, with which they essentially fly through the water. Their limb movements are akin to the lift-generating wing beats of birds, only being applied in a more viscous fluid medium, i.e., water, as opposed to air. Fossil sea turtles are relatively common in Cretaceous and Tertiary rocks, in particular. Continental deposits such as the shales and chalks of the North American Western Interior Seaway, and those of Europe and Africa, preserve large protostegids (e.g., *Protostega*, *Archelon*) and a variety of lesser sized cheloniids, osteopygids, toxochelyids, and others. They are all similar in appearance to living forms and are unmistakable as turtles. The living leatherback turtle (*Dermochelys*), at 2.5 m, is among the largest of all turtles. All living sea turtles are threatened or endangered.

Figure 4 Skeleton of the modern sea turtle *Dermochelys*.

The Mesozoic Sauropterygia contains a wide variety of paraxial propulsors. The earliest representatives are the Triassic pachypleurosaurus, 'nothosaurs' (i.e., stem-group Eusauropterygia) and placodonts. They share with the later plesiosaurus a 'euryapsid' skull condition and a reversed relationship of dermal and endochondral bones of the shoulder girdle, among other features. All typically also possess elongate necks, allowed by the development of limb-dominated swimming, whereas a short neck is a physical requirement for efficient axial propulsion. Pachypleurosaurs are exclusively known from the Alpine and Germanic basins of Europe and the east Tethyan province of China. They were small animals, generally much less than a metre in length, and had not advanced far beyond the terrestrial morphology of their ancestors. *Neusticosaurus*, *Keichousaurus*, and the others had long tails, limbs with obvious digits, and small heads with many, needle-sharp teeth. However, initial indications of paraxial propulsion are given by their lengthened necks, stiff thoraces, and broad and inflexible forelimbs. Pachypleurosaurs were near-shore (littoral) animals, but it is unknown whether they retained the ability to return to land. They probably ate small fish and invertebrates. Dense, pachyostotic vertebrae and ribs for use in buoyancy control were common features in pachypleurosaurs.

The Placodontia were bottom-dwelling, molluscivous forms of the Middle to latest Triassic, ranging in length up to 1.5 m. As far as is known, they were restricted to the Tethyan and adjacent provinces of Europe and the Mediterranean. Placodonts are generally separated into two subdivisions, the nonarmoured Placodontoidea (e.g., *Placodus* and *Paraplacodus*) and the armored Cyamodontoidea (e.g., *Cyamodus*, *Psephoderma*, and *Henodus*). All had short heads, a broad body, and some degree of limb-dominated propulsion. A variety of osteoderms or bony plates were present on the carapace and sometimes also fused to the back of the skull of cyamodontoids. The placodont skull was stout and akinetic, with a broad, strong cheek region in response to increased stresses related to molluscivory. Their diet is obvious from the durophagous, anvil-like crushing teeth of the palate and usually the maxilla and dentary. The Placodontia occupied shallow, near-shore, perhaps lagoonal, palaeoenvironments where the invertebrate benthos was abundant.

The remaining Triassic sauropterygians are relatively closely related to the later plesiosaurs but may not form a natural phylogenetic group. The Old World 'nothosaurs' were often large (up to 3 m) predators with a conservative, elongate body plan. Their swimming style was probably similar to that of the pachypleurosaurs. The flattened heads of 'nothosaurs' were often elongate and usually contained large teeth of unequal length (anisodont dentition). *Nothosaurus* of the Germanic Muschelkalk and *Lariosaurus* of the Alps were seemingly formidable fish eaters. Germany's *Simosaurus* was remarkable in having a rather shorter face with small, spatulate teeth. The behavioral significance of this morphology is unknown. Other taxa, including the pistosaurs (*Pistosaurus* of the Muschelkalk and *Augustasaurus* from the Middle Triassic Star Peak Group of Nevada), retained the primitive (plesiomorphic) nasal bone of 'nothosaurs', yet exhibited an open palate as in Plesiosauria. *Corosaurus* of the Wyoming Alcova Limestone may have had a sister-taxon relationship to pistosaurs.

The Plesiosauria (e.g., *Plesiosaurus*, *Pliosaurus*, and *Thalassiodracon*) were the most advanced and successful of the sauropterygians (**Figure 5**). They ranged from the latest Triassic to the Late Cretaceous when, like mosasaurs, they died off at the terminal Cretaceous mass extinction event. First known in England, they enjoyed a worldwide distribution by the Jurassic. Diverse, fully nektonic, and far ranging, they had no ability for terrestrial locomotion as evidenced by their wing-like limbs and separate ventral and dorsal skeletons. A viviparous or ovoviviparous reproductive style may be inferred. The adoption of paraxial locomotion using two symmetrical sets of hyperphalangic limbs remains unique to plesiosaurs. Also notable was their possession of elongate necks that, in the Cretaceous elasmosaurs (e.g., *Styxosaurus*), could include nearly 70 vertebrae. Once viewed as surface-going 'rowers' with flexible, swan-like necks, it is now known that plesiosaurs were subaqueous 'flyers' with relatively stiff, horizontally held necks. They were agile, mid-column swimmers, seemingly with well-developed underwater vision and perhaps olfaction. The development of sauropterygian 'underwater flight', perfected in the plesiosaurs, appears to have been constrained by stiff ventral baskets of dense gastralia, or ventral ribs. These heavy baskets were probably expanded to act as ballast in the skeletons of ancestral sauropterygians. Many plesiosaurs, particularly the elasmosaurs, swallowed 'stomach stones' (gastroliths) to aid in the production of neutral buoyancy.

Though probably not representing a natural phylogenetic division, two basic plesiosaur morphotypes paired very elongate necks with small heads ('plesiosauroids'), versus large heads with relatively short necks (pliosauroids). Intermediate combinations are also known. In general, small-headed forms with thin sharp teeth are thought to have been piscivorous, whereas those with large heads and teeth were carnivorous, rapid-pursuit predators. Known examples of bite marks indicate predation of the large on the

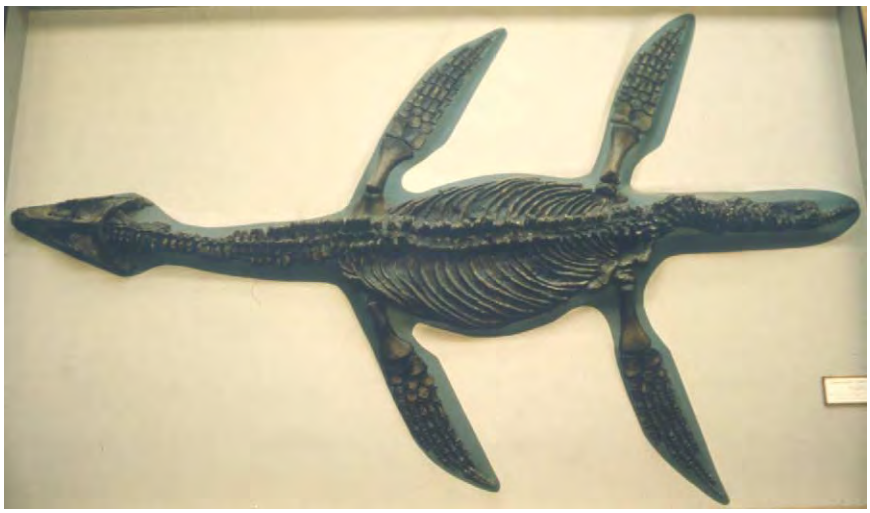


Figure 5 The Lower Jurassic plesiosaur *Rhomaleosaurus* of England.

small. Small, longirostrine pliosauroids, the polycotylids (e.g., *Trinacromerum*), are probably related to elasmosaurs.

Finally, a brief word may be said about the variety of ocean-living birds that might also be considered advanced marine reptiles. Notable among these is the Sphenisciformes, or penguins (Eocene to the Present). Advanced paraxial swimmers, they are far more at home in water than on land. As part of the radiation of archosaurs, the 'ruling' reptiles, penguins may legitimately be considered marine dinosaurs. Excepting birds, the extremely diverse dinosaurs never exploited the marine realm. Uniquely, dinosaurs adopted an upright stance or posture, as opposed to the more or less sprawling stance of typical reptiles. Most swimming reptiles adopted a sinuous undulation of their bodies and tails for locomotion. Others (turtles and plesiosaurs, for example) were prevented from doing so by a stiff shell or body. The upright stance of ground-dwelling dinosaurs may have prevented a viable swimming style. Because birds are unique in having two independent locomotory systems – legs for walking/running and wings for flight – the development of wings was essentially a pre-adaptation allowing underwater flight. Perhaps dinosaurs could effectively enter the sea only after abandoning the ground for the air.

See Also

Fossil Vertebrates: Reptiles Other Than Dinosaurs. **Lagerstätten**. **History of Geology From 1780 To 1835**. **Mesozoic:** Triassic; Jurassic; Cretaceous. **North America:** Continental Interior.

Further Reading

Andrews CW (1913) *A Descriptive Catalogue of the Marine Reptiles of the Oxford Clay. Part II*. London: British Museum (Natural History).

Bardet N (1994) Extinction events among Mesozoic marine reptiles. *Historical Biology* 7: 313 324.

Bell GL (1997) A phylogenetic revision of North American and Adriatic Mosasauroidea. In: Callaway J and Nicholls E (eds.) *Ancient Marine Reptiles*, pp. 293 332. San Diego: Academic Press.

Carpenter K (1997) Comparative cranial anatomy of two North American Cretaceous plesiosaurs. In: Callaway J and Nicholls E (eds.) *Ancient Marine Reptiles*, pp. 191 216. San Diego: Academic Press.

Cruickshank ARI, Small PG, and Taylor MA (1991) Dorsal nostrils and hydrodynamically driven underwater olfac tion in plesiosaurs. *Nature* 352: 62 64.

Darwin C (1859) *The Origin of Species by Means of Nat ural Selection*. London: John Murray.

Ellis R (2003) *Sea Dragons: Predators of the Prehistoric Oceans*. Lawrence, KS: University of Kansas Press.

Hua S and Buffetaut E (1997) Crocodylia: introduction. In: Callaway J and Nicholls E (eds.) *Ancient Marine Rep tiles*, pp. 357 374. San Diego: Academic Press.

Massare JA (1988) Swimming capabilities of Mesozoic marine reptiles: implications for method of predation. *Paleobiology* 14: 187 205.

Moody RTJ (1997) The paleogeography of marine and coastal turtles of the North Atlantic and Trans Saharan regions. In: Callaway J and Nicholls E (eds.) *Ancient Marine Reptiles*, pp. 259 278. San Diego: Academic Press.

Motani R, You H, and McGowan C (1996) Eel like swim ming in the earliest ichthyosaurs. *Nature (London)* 382: 347 348.

Motani R (2002) Swimming speed estimation of extinct marine reptiles: energetic approach revisited. *Paleobiol ogy* 28: 251 262.

Nosotti S and Pinna G (1996) Osteology of the skull of *Cyamodus kuhnschnyderi* Nosotti & Pinna 1993 (Rep tilia, Placodontia). *Paleontologia Lombarda* 6: 1 42.

Nosotti S and Rieppel O (2002) The braincase of *Placodus* Agassiz, 1833 (Reptilia, Placodontia). *Memorie Società Italiana di Scienze Naturale Milano* 331: 1 18.

Owen R (1840) Note on the dislocation of the tail at a certain point observable in the skeleton of many Ichthyo sauri. *Transactions of the Geological Society of London* 5: 511 514.

Owen R (1842) Report on British fossil reptiles. *Report of the British Association for the Advancement of Science* 1841: 60 204.

Owen R (1860) On the orders of fossil and Recent Reptilia, and their distribution in time. *Report of the British Asso ciation for the Advancement of Science* 29: 153 166.

Rieppel O (1994) Osteology of *Simosaurus gaillardoti* and the relationships of stem group Sauropterygia. *Fieldiana: Geology* 28: 1 85.

Rieppel O, Sander PM, and Storrs GW (2002) The skull of the pistosaur *Augustasaurus* from the Middle Triassic of north western Nevada. *Journal of Vertebrate Paleon tology* 22: 577 592.

Schuchert C (1928) The continental displacement hypoth esis as viewed by Du Toit. *American Journal of Science* 16: 266 274.

Storrs GW (1993) Function and phylogeny in sauroptery gian (Diapsida) evolution. *American Journal of Science* 293A: 63 90.

Storrs GW (1997) Morphological and taxonomic clarifica tion of the genus *Plesiosaurus*. In: Callaway J and Nicholls E (eds.) *Ancient Marine Reptiles,* pp. 145 190. San Diego: Academic Press.

Taylor MA (1997) Before the dinosaur: the historical sig nificance of the fossil marine reptiles. In: Callaway J and Nicholls E (eds.) *Ancient Marine Reptiles,* pp. xix xlvi. San Diego: Academic Press.

Flying Reptiles

D Naish and D M Martill, University of Portsmouth, Portsmouth, UK

Introduction

Reptiles, the group of amniotes that encompasses turtles, lepidosaurs (lizards, snakes, and amphisbaenians), crocodilians, and all of their extinct relatives, appeared around 320 million years ago as small, quadrupedal terrestrial predators. During the Palaeozoic and Mesozoic, reptiles burgeoned into a phenomenal diversity, and a number of Permian and Triassic groups took to the air, often as superficially lizard-like gliders whose extendible wings were supported by ribs or rib-like structures. Some modern lizards and snakes glide using flaps of skin along the flanks and limbs, webbing between their toes, or expandable connected ribs. It should also be pointed out that, from the phylogenetic point of view, birds are reptiles because they are part of the Dinosauria. Flighted birds are thus also flying reptiles.

Why flight evolved in reptiles is an interesting question that may not have a single answer. Flying reptiles may have been able to pursue insects, avoid predators, and cover difficult terrain better than their grounded relatives, and wing membranes may have initially evolved as signalling devices.

Pterosaurs – Actively Flying Reptiles of the Mesozoic

Except in birds, the only reptiles to evolve true flapping flight were the pterosaurs ('wing lizards') or pterodactyls ('wing fingers'), a diverse and unique group from the Triassic, Jurassic and Cretaceous periods of the Mesozoic Era. Pterosaur anatomy, flight mechanics, and palaeobiology have proved controversial and, while scientific study of the group began as early as the late 1700s, the last three decades have seen a major increase in the study of the group.

The earliest known pterosaurs are from the Norian stage of the Late Triassic, but already they were true pterosaurs. The lack of a definite 'proto-pterosaur' has made it difficult to understand pterosaur origins, although some possible contenders have been described (see below). An array of primitive pterosaurs had mostly died out by the end of the Jurassic but, by this time, the advanced pterosaur group, the Pterodactyloidea, had evolved. These were abundant and widespread throughout the Early Cretaceous but, towards the end of the Late Cretaceous, their diversity waned and only one or two groups persisted until the end of the Maastrichtian. Why this decline occurred is unknown, but the Cretaceous diversification of birds may have decreased the ecological opportunities available to pterosaurs.

History of Discovery

The first pterosaur was discovered in the Late Jurassic Solnhofen Limestone of Eichstätt, Germany, and was described by Cosimo Collini (1727–1806) in 1784. Collini concluded that it was a possible sea creature of unknown affinity, although he did note bat-like features. In 1801, the great French anatomist Georges Cuvier (1769–1832) recognized that the creature was a reptile and that its elongated digits must have supported flight membranes. Cuvier was thus the first to recognize pterosaurs as flying reptiles and, in 1809, he coined the name 'Ptero-Dactyle'. This later became the generic name *Pterodactylus* (Figures 1 and 4).

In the decades that followed, a succession of further pterosaurs from the Solnhofen Limestone was announced, many in a spectacular state of preservation and some with their wing membranes intact. The first recognized British pterosaur, a specimen of the deep-skulled *Dimorphodon*, was discovered by Mary Anning (1799–1847) in 1827 in Lower Jurassic rocks of Lyme Regis, Dorset. We now know that Gideon Mantell (1790–1852), best known for the discovery of *Iguanodon*, found pterosaur remains before this in the Early Cretaceous Wealden strata of Sussex, but had thought that these were from birds. North America yielded its first pterosaur to the prolific palaeontologist O. C. Marsh (1831–1899) in 1871 and, by 1876, Marsh had recognized it as a new, distinctive genus he named *Pteranodon* (meaning 'winged and toothless'). With an estimated wingspan of 6 m, *Pteranodon* was huge compared to most earlier discoveries.

While these discoveries and others were being made, varied opinions on the nature and life style of pterosaurs were appearing, and they were variously depicted as swimming creatures, as bats, marsupials, or as kin of birds. By the early 1900s, it was generally agreed that pterosaurs were bat-like flying reptiles and, in 1901, Harry Seeley (1839–1909) published *Dragons of the Air*, the first book devoted to pterosaurs.

South American Cretaceous pterosaurs have proved to be among the most important in the world, but not until 1971 was the first pterosaur from the now famous Santana Formation of Brazil discovered. Since then a significant number of new kinds from around the world (around 70 genera are presently recognized) have revealed previously unimagined morphologies and maximum sizes. Until 1971, *Pteranodon sternbergi*

Figure 1 Life restoration of the Late Jurassic pterodactyloid, *Pterodactylus*, from the German Solnhofen Limestone in a quadrupedal stance. Note the presence of body hair and the soft tissue head crest. Reproduced with permission from Dino Frey. Buffetaut E and Mazin J M (2003) *Evolution and Palaeobiology of Pterosaurs. Geological Society Special Publication 217*. London: The Geological Society of London.

(wingspan, 9 m) was the largest known flying animal, but the discovery in Texas of *Quetzalcoatlus* revealed that the biggest pterosaurs achieved wingspans of 11 m. Related pterosaurs of similar or larger size were discovered in the 1990s in Spain and eastern Europe.

The Pterosaur Skeleton

The pterosaur skeleton was highly modified for flight, and the most obvious features are the huge size of the skull compared with the body and the extreme elongation of one of the fingers. Like birds, most pterosaurs had hollow bones with foramina (small openings), indicating that they contained air sacs connected to the lungs. Pterosaur bones were supported internally by struts, and the bone walls themselves, usually no thicker than 2 mm, are composed of multiple overlapping layers and thus combine lightness with strength.

Pterosaur skull morphology is varied, although the majority had long, slim, shallow jaws and all had large orbits (eye-sockets). In basal pterosaurs, the external nostril was separate from an opening in front of the orbit called the antorbital fenestra. In pterodactyloids, these two openings merged into a single one called the nasoantorbital fenestra. Pterosaur teeth were extremely variable. Widely spaced pointed teeth, were widespread and from ancestors with teeth like these evolved species with fang-like teeth at the jaw tips and the unique *Istiodactylus* with its short petal-shaped teeth. The Late Triassic *Eudimorphodon* and *Austriadactylus* possessed multicusped teeth while elongate, slender teeth numbering in the hundreds evolved in the ctenochasmatoids. Toothlessness evolved several times. Some pterosaurs skulls sport bony crests at the jaw tips, along the midline or at the back of the skull.

Unlike birds and bats, the main wing spar in pterosaurs was formed by a hypertrophied digit (Figure 2). This 'wing finger' is generally considered to be the fourth because the digital formulae of the pterosaur hand best matches that of digits one to four in the hands of other reptiles. However, a rod-shaped bone projecting from the pterosaur wrist, called the pteroid bone, has at times been argued to represent the first hand digit. This is a minority view today but, if it is correct, then pterosaurs have five hand digits and the wing finger is the fifth. Although most pterosaur fossils show the pteroid pointing towards the shoulder, some workers suggest that it pointed forwards parallel to the neck. Regardless, the pteroid was probably mobile and used to control the attitude of the propatagium (Figure 3).

The pterosaur pectoral girdle includes a (normally fused) scapula and coracoid that meet at an acute angle and, as expected for flying animals, the socket for the humerus faces sideways and slightly upwards. The coracoids attach to an enlarged keeled sternum that anchored most of the major flight muscles. The bones of the pterosaur pelvis were short, usually fused together, and with a closed hip socket. Pterosaurs have a pair of unique rod- or plate-like bones called the prepubes projecting forwards from the bottom of the pelvis. Like the gastralia ('belly ribs') that all pterosaurs possessed, they may have helped support the gut or keep the abdomen rigid.

The vertebral column in pterosaurs can clearly be differentiated into cervical, dorsal, sacral, and caudal portions. The number of vertebrae is variable and pterosaurs have 7–9 cervical, 11–16 dorsal, 3–10 sacral, and 11–40 caudal vertebrae. The many caudal vertebrae of basal pterosaurs are encased in long bony processes that make the tail stiff and rod-like. In derived Cretaceous pterosaurs, most of the dorsal

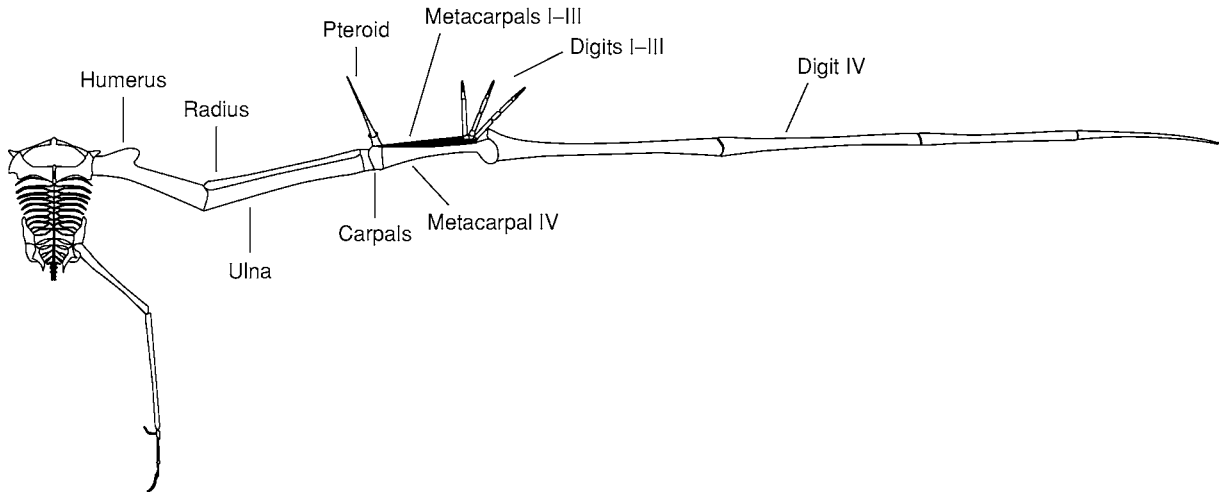


Figure 2 Wing skeleton of an ornithocheiroid pterodactyloid.

vertebrae are fused together forming a structure called the notarium.

Compared with the other wing segments, the pterosaur humerus is short, although generally with massive crests for muscle attachment. The ulna is always larger than the radius and both are attached distally to block-shaped carpal bones. Projecting from one of these is the unique pteroid. Pterosaurs had four metacarpals, the first three of which were slim and, with the exception of *Nyctosaurus* from Late Cretaceous North America, attached to short, clawed fingers. Why *Nyctosaurus* lacked clawed fingers is unknown, but in all other pterosaurs these digits may have served important functions. Trackways show that they were used in walking, and it is also possible that they were employed in grooming or climbing. The fourth metacarpal was robust and tipped with a twisted, roller-like distal end to which was attached the massive wing finger. This consists of four long straight bones, excepting a few genera where there were only three. Because of the twisted end of the fourth metacarpal, the wing finger would have lain parallel to the body's long axis when the wing was folded up.

The pterosaur hind limb is lightly built and the head of the femur is only slightly offset from the long axis of the shaft. Pterosaur hind limbs seem to have been quite flexible, but mostly sprawled to the sides. During flight, the hind limb was probably held in a bat-like orientation and could have been used to control the shape of the wing membranes. Basal pterosaurs are five-toed, with a prominent curving fifth digit that is hooked towards the tail. In pterodactyloids, the fifth digit is either absent or present as a tiny stub.

Because some articulated fossils indicate that the foot could assume a 90° angle relative to the tibia (and there is little evidence for much motion at the metatarsophalangeal joints), pterosaurs have generally been regarded as plantigrade (placing the whole length of the foot on the ground when walking). In 1983, Kevin Padian argued that this was not the case for *Dimorphodon* and that it may instead have been digitigrade (walking only on the toes). This was later inferred for all pterosaurs. An articulated *Dimorphodon* foot shows, however, that only limited motion was possible at the metatarsophalangeal joint, thus supporting a plantigrade posture. This is in agreement with probable pterosaur tracks preserved as trace fossils.

Soft Tissue, Integument, and Pterosaur Life Appearance

Many aspects of pterosaur life appearance remain unknown or controversial, although a number of exceptional fossils have provided some surprising details. Pterosaur body hair was reported as early as 1831 and described for various Jurassic pterosaurs between the 1920s and 1970s and today it is clear that pterosaurs had bristle-like hairs covering their necks and bodies (Figure 4). The active flapping flight and body hair of pterosaurs suggest that they had an elevated metabolism.

Other exceptional fossils show that some pterosaurs possessed a throat pouch, webbing between the toes, and scales on the soles of the feet. Soft

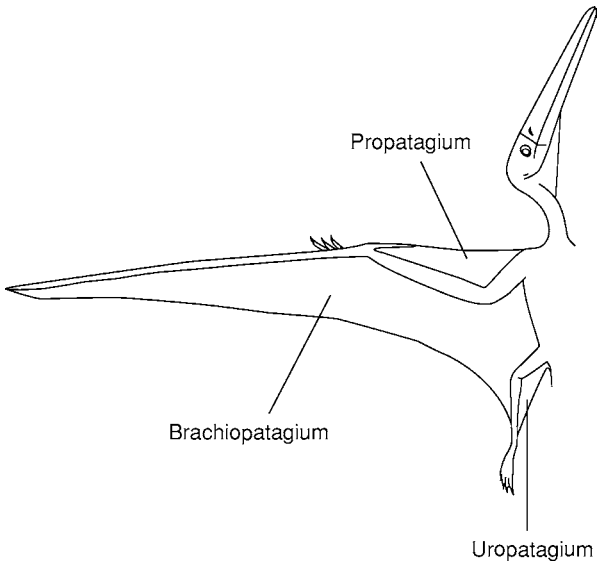


Figure 3 Schematic representation of the flight membranes in a generalized pterodactyloid pterosaur.

Figure 4 An exceptionally well preserved skeleton of the Late Jurassic pterodactyloid *Pterodactylus* from the German Solnhofen Limestone. This specimen preserves parts of the flight membranes, a throat pouch, and hairs on the neck and back.

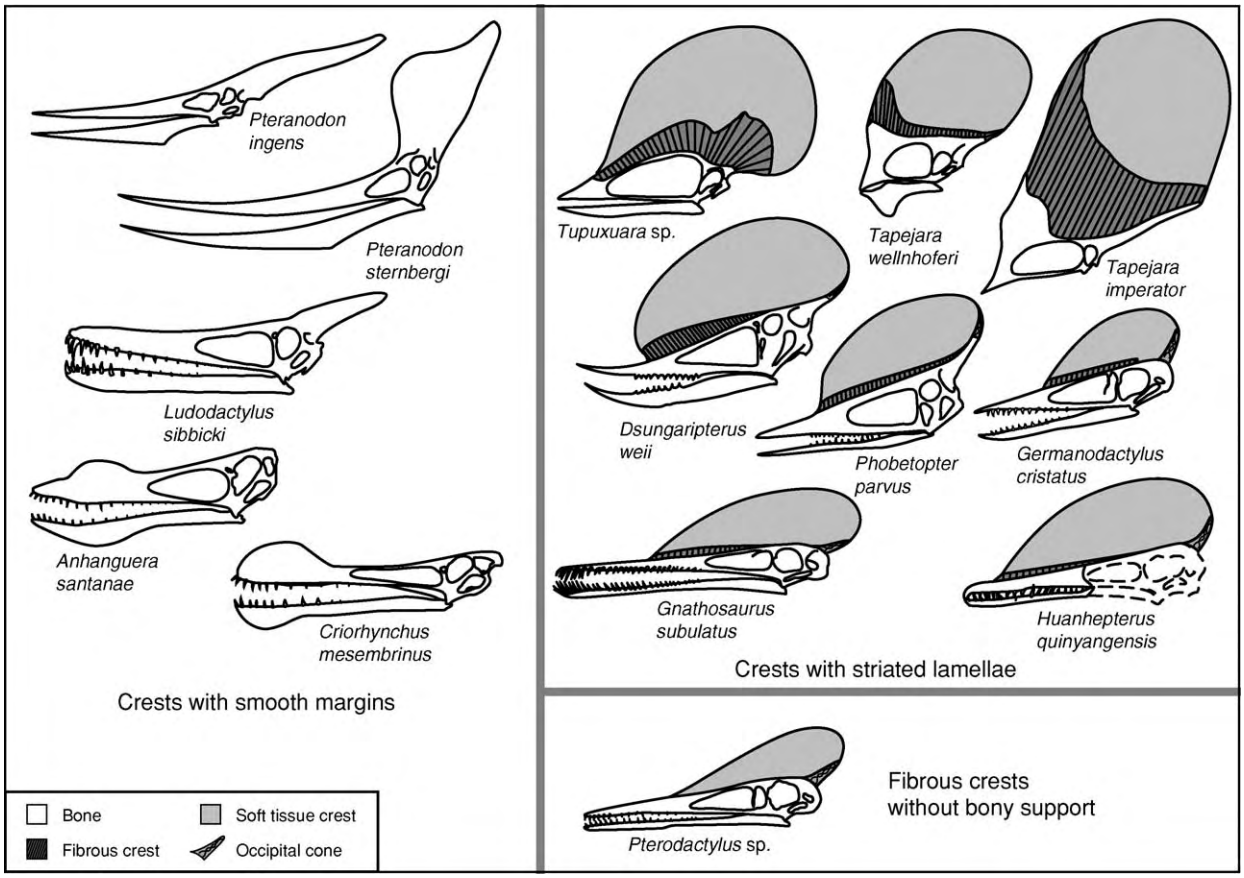


Figure 5 Variation in skull crest morphology in pterodactyloids. Soft tissue crests are now known for a wide diversity of pterodactyloids. Reproduced with permission from Dino Frey and Marie Celine Buchy. Buffetaut E and Mazin J M (2003) *Evolution and Palaeobiology of Pterosaurs. Geological Society Special Publication 217*. London: The Geological Society of London.

tissue skull crests connected to the underlying bony crests have proved to be widespread and appear to have doubled the size of the bony crests (Figures 5 and 6). An unexpected discovery is a soft tissue crest in *Pterodactylus*, a genus that lacks a bony crest (Figure 1). The presence of a distinctive bone texture on the pterosaur snout, jaw, and palate indicates that pterosaurs were beaked.

Pterosaur wing membranes are known from well-preserved specimens from the Solnhofen Limestone and the Early Cretaceous Brazilian Crato and Santana formations. A membrane called the propatagium extended from the shoulder to the pteroid and perhaps distally to encompass the first three fingers. The main flight membrane, the brachiopatagium (also called the cheiropatagium), extended from the tip of the wing finger to the hind limb, extending as far distally as the knee, shin, or ankle. Another membrane, the uropatagium, was present between the hind limbs (Figure 3). The wing membrane appears to have been complex, with a thin epidermis, a layer of vascular tissues, a layer of stiffening fibres called aktinofibrils, a thin sheet of muscle, and a

Figure 6 Skull of the tapejarid pterosaur *Topejara navigaus* from the Early Cretaceous Crato Formation of Brazil with bony and soft tissue skull crest. Buffetaut E and Mazin J M (2003) *Evolution and Palaeobiology of Pterosaurs. Geological Society Special Publication 217*. London: The Geological Society of London.

blood capillary network. *Rhamphorhynchus* and probably other long-tailed pterosaurs possessed a vertical diamond-shaped membrane at the tail tip.

With skin membranes connecting the wings, body, and legs, pterosaurs may have been superficially bat-like but, because bats are mostly dark-coloured nocturnal animals, it is doubtful that the similarities were strong. Pterosaurs mostly seem to have been ecological analogues of sea- and water-birds, and it might be that they were patterned in whites, blacks, and greys, although bright colours presumably decorated their crests.

The Affinities and Origin of Pterosaurs

Historically, pterosaurs have been allied with Mesozoic marine reptiles, bats, marsupials, and birds (*see* **Fossil Vertebrates:** Dinosaurs; Birds). However, major improvements in the understanding of vertebrate evolution allowed the palaeontologists of the nineteenth and twentieth centuries to realize that pterosaurs were related at least vaguely to dinosaurs (*see*) and their allies.

Although it is clear that pterosaurs are part of the major reptile assemblage known as the Diapsida, their affinities within this group are controversial. The presence of an antorbital fenestra has conventionally meant that pterosaurs have been regarded as archosaurs, the so-called ruling reptile group that incorporates crocodilians, dinosaurs, and kin. Among archosaurs, pterosaurs share a simple hinge-like ankle joint with dinosaurs and consequently have been regarded as close relatives of dinosaurs in most studies. This view was developed at a time when some workers thought that pterosaurs originated from terrestrial bipedal ancestors and that pterosaurs themselves were bipedal and digitigrade. A small bipedal, long-legged archosaur from Late Triassic Scotland, *Scleromochlus*, was argued to be a pterosaur ancestor, but recent studies refute this idea. The idea that pterosaurs might be close relatives of dinosaurs can certainly be regarded as the 'mainstream' view in vertebrate palaeontology today. However, several recent studies have questioned the evidence for this proposed affinity.

An alternative hypothesis argues that pterosaurs belong instead to a group of archosaur-like diapsids, the Prolacertiformes. Most prolacertiforms were superficially lizard-like, but *Sharovipteryx* from Late Triassic Kyrgyzstan appears to be intermediate between conventional prolacertiforms and pterosaurs. It has pterosaur-like hind limbs and vertebrae and membranes between its hind limbs and tail.

Some other models for pterosaur ancestry have been proposed. In 1996, S. Christopher Bennett argued that pterosaur hind limbs are only superficially similar to those of dinosaurs, and that re-analysis favoured a position for pterosaurs outside of the crocodilian–dinosaur group. Rather more heterodox recent ideas include the suggestion that pterosaurs are the closest relatives of birds and that pterosaurs are part of the Dinosauria.

Several different models have been proposed for the origin of pterosaurs, but the presence in basal pterosaurs of climbing features and of various details in the hind limb and pelvis indicative of a leaping ability suggest that pterosaurs first evolved as tree-climbing leapers.

Pterosaur Diversity and Phylogeny

It was recognized in 1901 that pterosaurs could be divided into two groups: the toothed, mostly long-tailed Rhamphorhynchoidea, and the short-tailed Pterodactyloidea (including both toothed and toothless kinds). Today, it is clear that rhamphorhynchoids include the ancestors of pterodactyloids and, consequently, Rhamphorhynchoidea is a grade and not a clade. Pterosaurs previously referred to as rhamphorhynchoids are nowadays termed basal pterosaurs or non-pterodactyloids. Although basal pterosaurs were diverse, it is notable that they were small compared with the majority of Cretaceous pterodactyloids.

The evolutionary relationships of pterosaurs are relatively understudied and only recently has pterosaur phylogeny been analysed. Although some areas of consensus have emerged, authors disagree on the details. We follow the phylogeny proposed by David Unwin of the Museum für Naturkunde in Berlin (Figure 7).

Perhaps the most basal pterosaur is *Preondactylus* from the Late Triassic of Italy. This form has a shorter coracoid and humerus and longer legs than other pterosaurs. Dimorphodontids, which include *Dimorphodon* from Early (and perhaps Middle) Jurassic

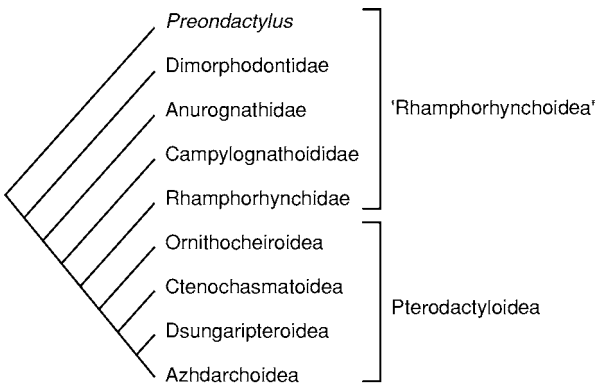


Figure 7 Cladogram depicting the relationships of all the major pterosaur groups. Reproduced from David Unwin.

England and Mexico, are basal pterosaurs with deep skulls superficially like those of puffins, whilst anurognathids were unusual in having short, broad snouts and abbreviated tails. A surprising recent discovery is the persistence of anurognathids into the Early Cretaceous. Two basal pterosaurs, *Eudimorphodon* from Late Triassic Italy and Greenland and *Campylognathoides* from Early Jurassic Germany and India, are united in the Campylognathoididae based on a distinctive lower jaw in which two pairs of large conical teeth are followed by multiple smaller ones.

Rhamphorhynchids were successful Jurassic pterosaurs known from Eurasia, North America, and Africa. *Rhamphorhynchus* from Late Jurassic Europe exhibits a laterally compressed, ventrally directed lower jaw tip and an array of forward-pointing teeth. It probably used these to grab fish and other small prey from the water. Another rhamphorhynchid lineage, the scaphognathines, had deeper skulls with teeth perpendicular to the jaw margins.

The Pterodactyloids

Pterodactyloids, the advanced short-tailed pterosaurs, originated from a rhamphorhynchid-like ancestor during the Middle Jurassic. The pterodactyloid radiation consisted of four major groups: the robust-jawed ornithocheiroids, the slim-jawed ctenochasmatoids, the low-crested dsungaripteroids, and the long-necked, crested azhdarchoids. A fifth group, the lonchodectids from Early Cretaceous England, are of uncertain affinity. Lonchodectids were small (wingspan, 1–2 m) with long, dorsoventrally flattened jaws with small teeth, each of which was supported by a low bony collar at its base.

Ornithocheiroids were large predatory pterosaurs (wingspan, 2–9 m) with robust beaks, often housing recurved, fang-like teeth at their tips (Figure 8). Their jaws frequently possessed keel-like dorsal and ventral crests, and some forms also possessed crests on the back of the skull. The toothless pteranodontids and nyctosaurids appear to be members of this group. The earliest known ornithocheiroids appear at the start of the Cretaceous, while nyctosaurids survived to the very end of the Cretaceous.

Ctenochasmatoids had needle-like meshes of teeth set in long, thin jaws. In *Pterodaustro* from late Early Cretaceous Argentina, the upturned lower jaw contains approximately 1000 bristle-like teeth. These were surely used for filtering small organisms from the water. Unlike ornithocheiroids, ctenochasmatoids had elongate cervical vertebrae and were generally small (wingspan, 50 cm to 2 m), although *Cearadactylus* from Early Cretaceous Brazil was a giant with a wingspan of 5.5 m.

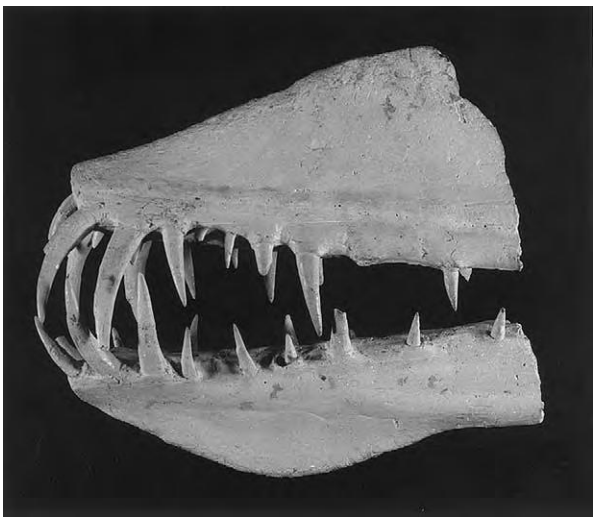


Figure 8 Jaw tips of the Cretaceous ornithocheiroid pterosaur *Coloborhynchus*. Note the massive fang like anterior teeth and the low keel like crests on both jaws. This specimen is from the Santana Formation of Brazil and would have belonged to an animal with a skull of approximately 1 m in length.

Members of the Dsungaripteroidea are known from the Late Jurassic and Early Cretaceous of Eurasia, Africa, and South America. Like some ctenochasmatoids, dsungaripteroids had a midline crest on the top of the skull. Their beaks often had toothless tips and they may have been predators of molluscs and crustaceans.

Finally, the strangest pterodactyloids must be the azhdarchoids. These include the long-necked azhdarchids and the crested tapejarids. Azhdarchids may have exceeded wingspans of 11 m and were widely distributed in the Late Cretaceous. They may have been ecological generalists akin to storks, and were probably not specialist carrion feeders or mud probers as has been proposed. Determining the life style of the tapejarids is more difficult. The vaguely parrot-like skull of *Tapejara* (Figure 6) from Lower Cretaceous Brazil led some workers to propose that it was a fruit eater, but it might better be imagined as an auk analogue. Recently, it has been suggested that *Thalassodromeus*, also from Lower Cretaceous Brazil, was a fish eater that trawled its blade-like lower jaw through the water.

Pterosaur Palaeobiology

Because pterosaurs are unique and extinct, reconstructing their palaeobiology is difficult and nothing is known about several aspects of their lives. Limited evidence does allow us, however, to reconstruct their sensory abilities, feeding behaviours, and styles of locomotion.

The large orbits of pterosaurs show that they had large eyes, and the abundance of visual display

structures in the group implies that vision was important. Given that living reptiles see in colour, pterosaurs probably did as well, and pterosaur brain casts reveal enormous optic lobes suggesting excellent eyesight. The many skull crests seen in pterosaurs were therefore probably used in sexual display, and differing crest shapes seen in various genera, most notably *Pteranodon*, are suggestive of sexual dimorphism. Individuals with large, long crests are inferred males and those with small, short crests are inferred females. This is supported by the pelvic structure, as small-crested individuals have wider pelvic canals than large-crested individuals.

Pterosaur crests have also been suggested to have functioned in muscle attachment, thermoregulation, or aerodynamics. Keel-like ornithocheiroid crests may have functioned as cutwaters or stabilizers when the jaws were pulled through the water, and it is possible that the tall soft tissue crests of tapejarids and other pterodactyloids functioned as sails.

Little direct evidence for pterosaur diet is known. In general, pterosaurs seem to have been predators of fish and other aquatic prey, although insects and small terrestrial vertebrates could have been captured by some kinds. A Solnhofen Limestone dragonfly preserves what appears to be a pterosaur bite mark on one of its wings. Stomach contents are known for a handful of pterosaurs and show that *Eudimorphodon*, *Rhamphorhynchus*, *Pterodactylus*, and *Pteranodon* ate fish. Pterosaurs themselves were sometimes preyed on; the bones of a small Triassic pterosaur have been discovered in a gastric pellet probably produced by a fish, and pterosaur bones have been discovered among plesiosaur stomach contents. An azhdarchid wing bone gnawed on by a predatory dinosaur has been described from Late Cretaceous Canada.

Some pterosaur specimens reveal evidence of accidents and disease. Fractures have been reported for jaws, ribs, and wing finger and toe bones in *Rhamphorhynchus*, *Pterodactylus*, *Anhanguera*, and *Pteranodon*; evidence for secondary infection but an absence of evidence for healing in most of these cases indicates that the animals died shortly after receiving these injuries, although one *Pterodactylus* specimen lived long enough for its fractured femur to heal. Osteoarthritis and evidence for infected facial lesions have been reported for ornithocheiroids.

Virtually nothing is known about pterosaur reproduction, and eggs and nests are unknown. We infer that pterosaurs laid eggs on the basis of the egg-laying behaviour of their living relatives. Juveniles of a few species are known and suggest that at least some species were able to fly soon after hatching. Unlike birds, these pterosaurs therefore underwent much of their growth while they were capable fliers. According to some experts, large samples of the same species reveal all individuals ceasing growth at the same size and consequently it has been argued that some pterosaurs exhibited determinate growth like most birds and mammals. There are some indications, though, that basal pterosaurs and azhdarchids may have continued growing for their whole lives.

Pterosaur Locomotion

Pterosaur locomotion, both aerial and terrestrial, has been investigated intensively. Although pterosaurs were imagined in the past as weak gliders, largely unable to flap, the evidence for substantial wing muscles attached to the sternum, scapulocoracoid, and dorsal vertebrae indicates that even the biggest forms were powerful flappers. Their muscles mean that, although they were light for their size, the biggest pterosaurs were not as low in weight as often stated, and an azhdarchid with an 11 m wingspan probably weighed in excess of 100 kg. A comparison of pterosaur wing shapes with those of birds and bats indicates that, as a generalization, pterosaurs were slow, manoeuvrable fliers equivalent to large sea-birds. Large pterosaurs presumably exploited rising air and winds to soar, and various studies have shown that they had low gliding and stalling speeds. This would be ideal for animals that caught their prey from the water's surface. Pterosaurs probably controlled their rate of ascent and descent by using the propatagium to alter the flow of air over the wing's upper surface. How pterosaurs took off is mysterious given that they were probably not proficient runners. Small forms may have leapt from trees, rocks, or cliffs, while larger kinds may have relied on headwinds and updraughts.

Pterosaurs could probably swim and there is no good reason why some of them could not have dived beneath the water surface. The toe proportions and claw shapes of small basal pterosaurs suggest that they may have been good tree climbers.

The terrestrial ability of pterosaurs is one of the most controversial areas in pterosaur study. Trackways now thought to have been produced by pterosaurs were first reported from Late Jurassic rocks in the USA in 1957 and reveal a three-fingered hand and a long, four-toed foot, the proportions of which match the pterosaur foot skeleton. Dubbed *Pteraichnus*, these tracks support the view of pterosaurs as plantigrade quadrupeds. *Pteraichnus* tracks are now known from many Late Jurassic and Early Cretaceous North American and Eurasian deposits and were probably made by ctenochasmatoids. Quadrupedal trackways that appear to have been produced by large (estimated wingspan, 6 m) ornithocheiroids and

basal pterosaurs have also been reported. However, although at least some pterosaurs seem to have been quadrupeds, it is still not certain that this was the case for all of them.

See Also

Fossil Vertebrates: Reptiles Other Than Dinosaurs; Dinosaurs; Birds; Swimming Reptiles; Mesozoic Amphibians and Other Non-Amniote Tetrapods.

Further Reading

Buffetaut E and Mazin J M (2003) *Evolution and Palaeobiology of Pterosaurs. Geological Society Special Publication 217*. London: The Geological Society of London.

Naish D and Martill DM (2003) Pterosaurs a successful invasion of prehistoric skies. *Biologist* 50(5): 213 216.

Wellnhofer P (1991) *The Illustrated Encyclopedia of Pterosaurs*. London: Salamander.

Mesozoic Amphibians and Other Non-Amniote Tetrapods

A R Milner, Birkbeck College, London, UK

Introduction

In the Mesozoic, the various amniote groups (reptiles, birds, mammals, and their earlier relatives) came to dominate terrestrial niches; the fossil records of this era depict a general decline of the archaic amphibian groups and diversification of the modern amphibian groups. In this article, the effect of the Permo–Triassic extinction event on amphibians is considered and the archaic amphibians of the Triassic and the few relicts that survived into the later Mesozoic are reviewed. This is followed by a consideration of the modern amphibian groups – the frogs, salamanders, and caecilians – usually combined as the Lissamphibia, and what is known of their evolution through the Mesozoic.

Amphibians and the Permo–Triassic Event

The Permo–Triassic (P–T) extinction event (*see* **Palaeozoic:** End Permian Extinctions) appears to have had a dramatic effect on amphibian turnover, although the Late Permian record is too patchy for there to be certainty that many of the lineages died out during the event and not before. It is clear that an immense diversity of new types of amphibians appeared in the Early Triassic, most belonging to the Stereospondyli. This has been interpreted as a direct response to the extinction event, but it has recently been suggested that it represents a Late Permian diversification in Gondwana that appeared suddenly in Laurasia as the result of dispersal from the south after the P–T extinction. The Lissamphibia (modern groups such as frogs and salamanders) also make their first appearance in the fossil record after the P–T event. They may represent a major diversification following the P–T extinctions, but the Triassic record is too poor for such a pattern to be seen and it can only be inferred. As well as the Stereospondyli and Lissamphibia, a few other lineages of temnospondyls and of chroniosuchian anthracosaurs survived the P–T event and persisted as relicts during the Early and Middle Triassic, without further diversification.

Mesozoic Assemblages

Most Triassic amphibian assemblages are composed of large temnospondyls in red-bed floodplain or coastal lagoon situations. Much of the fossil material is disarticulated and many forms are known only from the massively constructed skulls. The temnospondyls appear to have filled the niche later occupied by parasuchians and crocodiles and were gradually replaced by them during the Triassic. Only a very few Triassic localities have produced evidence of small amphibians, either juvenile temnospondyls or early representatives of the modern amphibian groups.

In the Jurassic and Cretaceous, fossils of the few remaining large amphibians occur in lake-bed or floodplain deposits. The increasingly abundant frog and salamander fossils occur either as intact skeletons in lake deposits (localities, or 'lagerstätten', where prevailing conditions have produced highly conserved or diverse fossils, termed 'konservat-lagerstätten') (*see* **Lagerstätten**) or as isolated elements in microvertebrate bone accumulations. The latter are common and produce abundant fossil material, although large samples are needed to make associations between isolated bones. Notable Cretaceous konservat-lagerstätten for frogs and salamanders are Las Hoyas, Spain and Liaoning, China.

Temnospondyls in the Triassic

Stereospondyli

The stereospondyls comprise one of the largest subgroups of the amphibian group Temnospondyli. They were mostly 1- to 2-m-long aquatic and amphibious carnivores filling a broadly crocodile-like niche. The 'stereospondylous' type of vertebra, after which the group is named, is one in which the anterior central element, the intercentrum, enlarges to form the entire central body (Figure 1A). In fact, this structure characterizes only a few genera (including some of the first to be discovered, hence its original use as a characteristic). One of the most obvious, though not unique, characteristics of the group is that the braincase is broadly sutured to the lateral palatal bones (pterygoids) on either side.

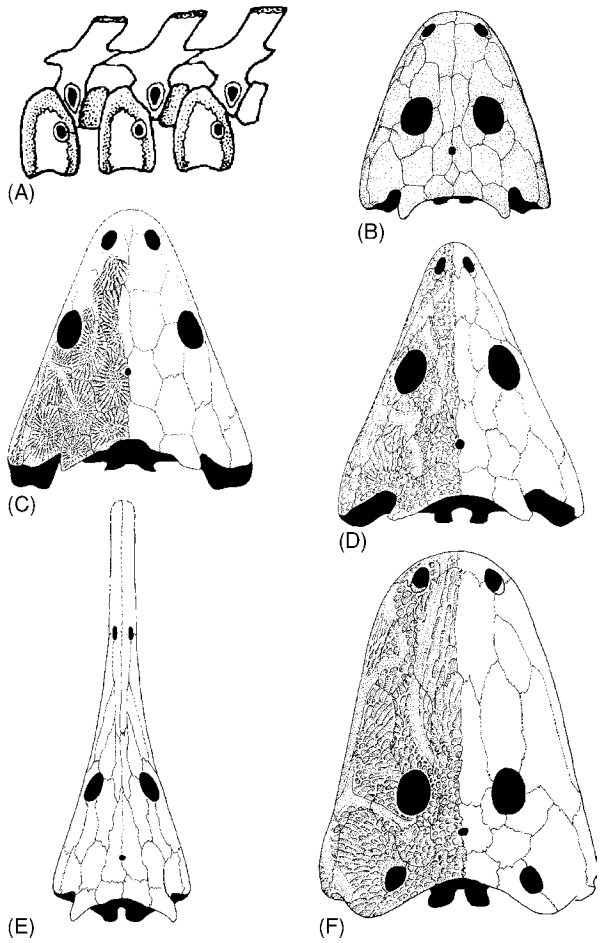


Figure 1 Stereospondyli. (A) *Mastodonsaurus* vertebra, demonstrating the stereospondyl condition, with an enlarged intercentrum and rudimentary pleurocentra; (B) skull of the lydekkerinid *Lydekkerina huxleyi*; (C) skull of the rhytidosteoid *Peltostega erici*; (D) skull of the trematosauroid *Lyrocephaliscus euri*; (E) skull of the trematosauroid *Aphaneramma rostratum*; (F) skull of the capitosauroid *Cyclotosaurus robustus*. All reproduced with permission from Schoch RR and Milner AR (2000) Stereospondyli. In: Wellnhofer P (ed.) *Handbuch der Paläoherpetologie,* part 3B. Munich: Pfeil.

The stereospondyl radiation had commenced diversification in the Late Permian but is one of the characteristic elements of continental and coastal Triassic faunas. There is not full agreement on the classification of the stereospondyls, but three major groups, Rhytidosteoidea, Trematosauroidea, and Capitosauroidea, have recognized by several workers, together with three families of controversial relationships.

Rhytidosteoidea

All of the members included with certainty in the Rhytidosteoidea are restricted to the Late Permian and Early Triassic. The basal forms are in the family Lydekkerinidae; these small forms are less than 1 m long and have orbits set in the middle of a blunt alligator-like skull (Figure 1B). They appear to have been amphibious animals. The more advanced Rhytidosteidae were more specialized aquatic forms, with small laterally placed orbits and either very triangular skulls (Figure 1C) or flattened rounded skulls. This appears to be a fundamentally Gondwanan group, with a few lineages penetrating into Laurasia. Most lydekkerinids and rhytidosteids are from the Lower Triassic of South Africa, Madagascar, India, Australia, and Antarctica, with fossil material recently reported from Brazil. Three genera are known from Greenland, Svalbard, and North Russia.

Trematosauroidea

The trematosauroids are an entirely Triassic clade of temnospondyls, first appearing after the P–T extinction and last appearing in the mid-Norian. The stem of the group is the Benthosuchidae, known largely from fossil material from Russia. Benthosuchids were a lineage of crocodile-like temnospondyls with elongate triangular skulls, and the most well-known taxon, *Benthosuchus* (Figure 2), is known from many hundreds of specimens preserved in three-dimensional form from the Sharzhenga River region. This taxon was described fully in 1940 by Bystrow and Efremov in a monograph that still serves as a standard for descriptions of this group.

The group adaptively radiated in the Early Triassic as the family Trematosauridae. The skull shapes within this family vary from an almost isosceles triangle shape (*Lyrocephaliscus*, from Svalbard; Figure 1D) to an extreme elongation of the snout that results in a gharial-like head (*Aphaneramma*, also from Svalbard; Figure 1E). All group members are characterized by a distinctive pattern of elongation of the back of the skull roof, small, laterally placed orbits, a braincase

floor that underplates the occipital region, and a parasphenoid (central strut in the palate) that is extremely narrow. Most had poorly ossified postcranial skeletons and small limbs and can be assumed to have been aquatic predators. Early in the Late Triassic, the trematosaurids vanish from the fossil record, but two descendant groups appear, the Almasauridae and the Metoposauridae (**Figure 3**). Both have long skulls like the trematosaurids had, but the snouts are short and the orbits are quite near the front of the head. Almasaurids were small animals, no more than a metre in length, whereas metoposaurids grew to 2 m and had extremely flattened skulls.

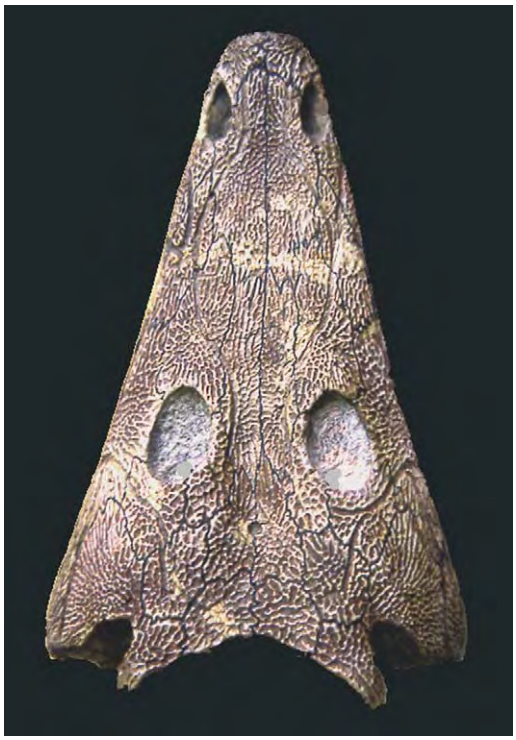


Figure 2 Stereospondyli. Skull of the primitive trematosauroid *Benthosuchus sushkini*, from the Lower Triassic, Russia. Specimen at the Natural History Museum, London. © Andrew Milner.

The Trematosauroidea appears to have been a basically Laurasian group, with successful forms extending into Gondwana. Evidence of the basal Benthosuchidae and primitive trematosaurs such as *Thoosuchus* and *Angusaurus* is found only in continental deposits in the Lower Triassic of Russia. The more advanced trematosaurids are mostly Early Triassic, but one relict genus, *Hyperokynodon*, survived up to the Late Triassic in Germany. Records of the more advanced trematosaurids are also found in Laurasia from Arizona to Russia, but evidence of two subfamilies exists among all of the major Gondwanan temnospondyl faunas. Most are in freshwater deposits, but in Svalbard, a range of trematosaurid fossils is found in unambiguously marine deposits, providing one of the few examples of a marine amphibian fauna. The Late Triassic almasaurids and metoposaurids appear to have been restricted to the floodplains associated with the developing rift valley systems across north and central Pangaea, and most fossil materials are found in North America, Europe, North Africa, and India.

Capitosauroidea

The capitosauroids (also known as mastodonsauroids) were a large, successful, and comparatively structurally uniform group occurring globally throughout the Triassic, but not surviving into the Jurassic. Most adults were 2–3 m long (**Figure 4**), but individuals in a few taxa (e.g., *Mastodonsaurus*) grew to a total length of 5–6 m. The skull was superficially alligator-like in most genera, with a long and broad snout. One capitosauroid family of note is the Cyclotosauridae, a group characterized by having the tympanic (eardrum) region of the skull surrounded by outgrowths of the neighbouring skull bones (**Figure 1F**). Cyclotosaurids first appear as a minor group in the early Triassic but are seen to be the predominant capitosauroid group in the Late Triassic. *Mastodonsaurus* from the Middle Triassic of Germany is noteworthy as the first

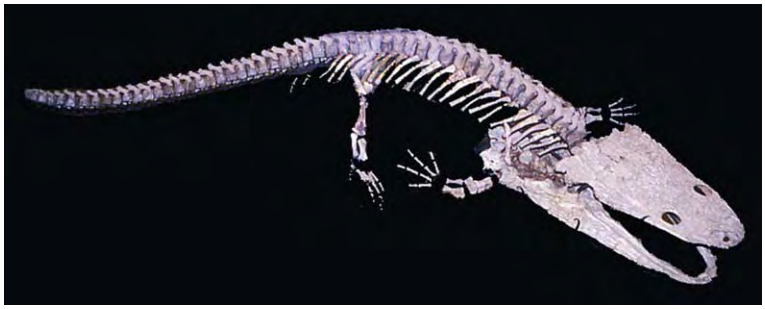


Figure 3 Stereospondyli. Skeleton of the metoposaurid *Buettneria perfecta*, from the Upper Triassic, North America. Specimen at the American Museum of Natural History, New York. © Andrew Milner.

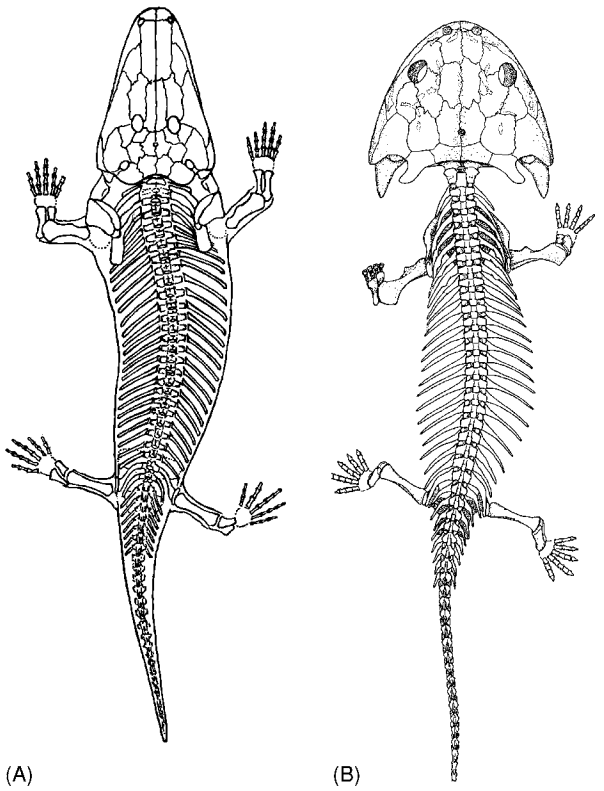


Figure 4 Stereospondyli. (A) Reconstruction of the capitosaurid *Paracyclotosaurus davidi*, from the Middle Triassic, New South Wales, Australia; (B) reconstruction of the chigutisaurid *Siderops kehli*, from the Lower Jurassic, Queensland, Australia. Reproduced with permission from (A) Watson, DMS (1958) A new labyrinthodont (*Paracyclotosaurus*) from the Upper Trias of New South Wales. *Bulletin of the British Museum (Natural History) Geological Series* 3, 233 263; (B) Warren AA and Hutchinson MN (1983) The last labyrinthodont? A new brachyopoid (Amphibia, Temnospondyli) fron the early Jurassic Evergreen Formation of Queensland, Australia. *Philosophical Transactions of the Royal Society, London Series B* 303, 1 62.

archaic fossil amphibian to be recognized and described (in 1828). Capitosauroidea may have originated in Gondwana, where the most primitive genus, *Watsonisuchus*, is known to have existed, but much of the record of this group suggests a uniform presence across Pangaea. The only subgroup that may be local in distribution is the Mastodonsauridae (*sensu stricto*), which has not yet been recognized outside Eurasia.

Enigmatic Paedomorphic Families

Three other families of temnospondyls represent very specialized aquatic forms that have apparently evolved by paedomorphosis (maintaining the larval body form throughout growth to a large size). The relationships between these three groups (Plagiosauridae, Chigutisauridae, and Brachyopidae) are

Figure 5 Stereospondyli. Skull of the plagiosaurid *Gerrothorax rhaeticus*, from the Upper Triassic, Germany. Specimen at the Staatliches Museum fur Naturkunde in Stuttgart. © Andrew Milner.

controversial and they are treated here as three distinct groups of uncertain affinities.

The Plagiosauridae were one of the strangest-looking groups of temnospondyls; they were extremely flat, having a very wide and shallow gape, massive orbits (Figure 5), and vertebrae suggestive of a vertically undulating back, and they were covered in a chain-mail-like armour both dorsally and ventrally. The branchial (gill) arches were large and well-ossified and they clearly had a large internal gill apparatus. Most were less than 50 cm long, but one large fossil specimen from Svalbard is at least 2 m long. They are believed to be a paedomorphic lineage evolved from a more terrestrial stereospondyl group, and two quite different terrestrial genera, *Peltobatrachus* and *Laidleria*, have both been proposed as the terrestrial relatives of the plagiosaurs. Almost all plagiosaurids are known from Eurasia and Greenland with some fragments having been found in Thailand and Australia.

The Chigutisauridae were large stereospondyls with semicircular skulls bearing a suction–gulping feeding apparatus and small, anteriorly placed orbits (Figure 4B). They may be paedomorphic relatives of the Rhytidosteidae. All Triassic chigutisaurids are from the Upper Triassic of Gondwana. The Brachyopidae are similar (Figure 6), but much of the resemblance may be convergence and they have some similarities to the Tupilakosauridae (see later), which are certainly not stereospondyls. Most Triassic brachyopids are from Gondwana, with one problematic genus (*Virgilius*) from Arizona. The chigutisaurids and brachyopids occur at low taxonomic diversity through their history.

Other Temnospondyls

A few non-stereospondyl lineages of temnospondyls are known to have survived the P–T event and persisted as relicts for a short period in the Lower

Triassic. Best known is the genus *Micropholis* (Figure 7), which is locally common in the Karoo of South Africa. It is a late survivor of the Carboniferous–Permian family Amphibamidae, resembling long-legged salamanders and believed to be relatives of the modern amphibians by some workers. The genus *Tungussogyrinus* from the Tungus Basin in Siberia is a tiny gill-bearing temnospondyl, first described as a larva but recently argued to be a late member of the Branchiosauridae, small paedomorphic temnospondyls best known from Permo–Carboniferous lake deposits in central Europe. Finally the Tupilakosauridae known, from Greenland, Russia, and South Africa, were probably found throughout Pangaea and were long-bodied, almost eel-like, forms (Figure 8).

Figure 6 Skull of the brachyopid *Batrachosuchus haughtoni*, from the Middle Triassic, South Africa. Specimen at the Natural History Museum, London. © Andrew Milner.

Figure 7 An aggregation of five individuals of the temnospondyl *Micropholis stowii*, from the Lower Triassic, South Africa. Specimen at the Bayerische Staatsammlung fur Geologie und Palaontologie, Munich. © Andrew Milner.

Chroniosuchians

Anthracosaurs, a significant group of Carboniferous and Early Permian large carnivores, are widely believed to be stem-amniotes, i.e., amphibian-grade organisms closer to the origin of reptiles, birds, and mammals than to modern amphibians. One relict anthracosaur taxon, the chroniosuchians, survived in the Late Permian of Russia (*see* **Palaeozoic:** End Permian Extinctions). They were small (50 cm–1 m) terrestrial carnivores superficially resembling monitor lizards. Until the early 1990s, they were assumed to have been wiped out by the P–T event. Recently, however, isolated characteristic armour-bearing vertebrae of chroniosuchians dating to the Middle Triassic have been found in Russia and Germany.

Post-Triassic Temnospondyls

Until 1977, it was generally assumed that all remaining temnospondyl lineages became extinct in the Norian during the extensive faunal turnover that occurred in the Late Triassic. This still may be true for central and western Pangaea, but two families of aquatic temnospondyl survived in the eastern peripheries of the continent until much later. The Brachyopidae are now known from the Middle–Upper Jurassic of Mongolia (*Gobiops*) and China (*Sinobrachyops*), and the similar Chigutisauridae have been described from the Lower Jurassic (*Siderops*; Figure 4B) and mid-Cretaceous of Australia. The latter record, *Koolasuchus* from the Aptian Stretzlecki Formation, is the latest known temnospondyl.

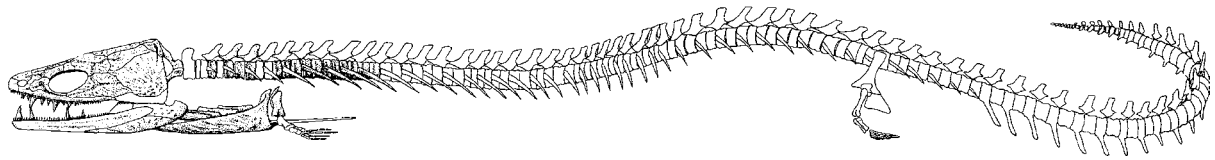


Figure 8 Reconstruction of the temnospondyl *Thabanchuia oomie*, from the Lower Triassic, South Africa. Reproduced with permission from Warren AA (1999) Karroo tupilskosaurid: a relict from Gondwana. *Transactions of the Royal Society of Edinburgh: Earth Sciences 89, 145 160.*

The Lissamphibian Groups

The taxon Lissamphibia applies to the living frogs, salamanders, and caecilians and their immediate relatives. They are believed by most workers to be a single adaptive radiation that started to diversify in the Permian, with the frogs demonstrably present in the Lower Triassic. The other groups first appear in the Jurassic and are presumed to be as yet undiscovered in Triassic rocks.

Frogs

Frogs occur on all non-polar continents and on many islands, particularly those that are relicts (New Zealand, Madagascar, and the Seychelles) of larger

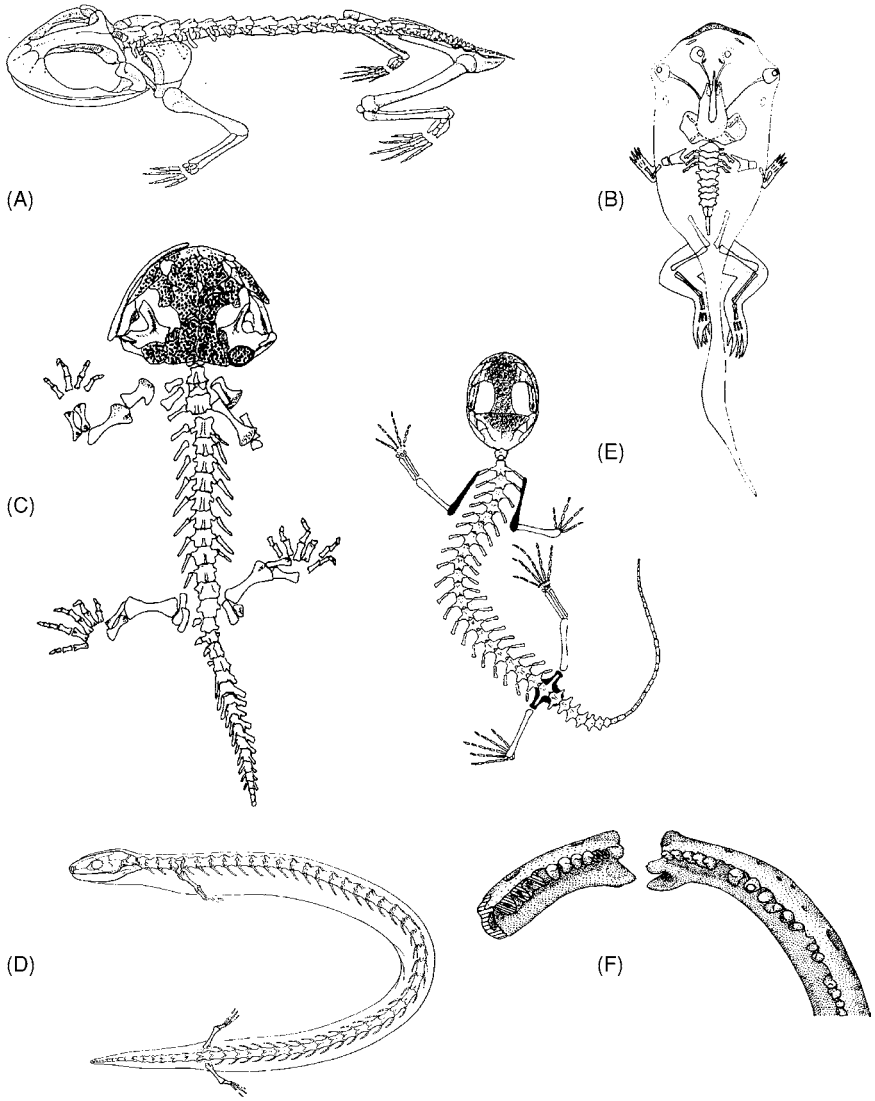


Figure 9 Lissamphibia. (A) Reconstruction of the protofrog *Triadobatrachus massinoti*, from the Lower Triassic, Madagascar; (B) the pipid tadpole *Shomronella jordanica*, from the Lower Cretaceous, Israel; (C), reconstruction of the primitive salamander *Karaurus sharovi*, from the Upper Jurassic, Kazakhstan, (D) reconstruction of the primitive caecilian *Eocaecilia micropodia*, from the Lower Jurassic, Kayenta Formation, Arizona; (E) reconstruction of the albanerpetontid *Celtedens ibericus*, from the Lower Cretaceous, Las Hoyas, Spain; (F) mandibles of the albanerpetontid *Celtedens*, showing the unique asymmetrical symphysial articulation. Reproduced with permission from (A) Rage, J C and Roček, Z (1989) Redescription of *Triadobatrachus massinoti* (Piveteau, 1936) an anuran amphibian from the early Triassic. *Palaeontographica A* 206: 1 16; (B) Estes, R, Špinar, ZV and Nevo, E (1978) Early Createceous pipid tadpoles from Israel (Amphibia: Anura) *Herpetologica* 34: 374 393; (C) Milner, AR (1994) Chapter 1. Upper Triassic andJurassic amphibians, fossil record and phylogeny. 5 22. In: Fraser, N and Sues, H D (eds), *In the Shadow of the Dinosaurs: Early Mesozoic Tetrapods*. Cambridge: Cambridge University Press; (D) Jenkins, FA and Walsh, D (1993) An early Jurassic caecilian with limbs. *Nature* 365: 246 250; (E) McGowan, G and Evans, SE (1995) Albanerpetontid amphibians from the Cretaceous of Spain. *Nature* 373: 143 145; (F) Fox, RC and Naylor, BG (1982) A reconsideration of the relationships of the fossil amphibian *Albanerpeton*. *Canadian Journal of Earth Science*. 19: 118 128.

continents. Frog skeletons are readily recognizable; the skeleton is highly modified for jumping, the backbone is very short (seven to nine vertebrae), and the tail vertebrae are fused as a rodlike urostyle. In the pelvis, the ilium is elongate, and in the hindlimb, the tibia and fibula are fused and the ankle bones are elongated. Fossils with some of the characteristics of frogs appear in the Early Triassic as *Triadobatrachus* (Madagascar) (**Figure 9A**) and *Czatkobatrachus* (Poland). *Triadobatrachus*, known from a near-complete skeleton, had the long ilium and modified ankles but had not yet acquired the other features. There are no other Triassic frogs or frog relatives, and the first frogs with the full suite of jumping adaptations are the Early Jurassic genera *Prosalirus* from Arizona and *Vieraella* from Argentina. By the Upper Jurassic, there are frogs attributable to the primitive modern family Discoglossidae in Laurasia, and it appears that the diversification to the modern families was under way.

Cretaceous frogs are not well documented, but the limited record suggests that much of the diversification to extant families had taken place. Discoglossids, pelobatids, rhinophrynids, and palaeobatrachids are known from Laurasia; leptodactylids and pipids are known from South America; and pipids are known from Africa. The genus *Gobiates*, sometimes elevated to the family Gobiatidae, is abundant in the mid-Late Cretaceous of central Asia but disappears in the Campanian and may be the only major type of Cretaceous frog not to survive into the Cenozoic. Rich assemblages of tadpoles of pipid frogs are known from the Lower Cretaceous of Makhtesh Ramon in Israel (**Figure 9B**).

Salamanders

The major diversity of living salamanders occurs in North America, central America, and Eurasia. A few genera are found in South America and North Africa, and these are believed to represent Neogene range extensions. Most of the Mesozoic record of salamanders is also restricted to Laurasia and the group appears to have been endemic to this region for much of its history. Diagnostic remains of salamanders first appear in the Bathonian of England and central Asia, and by the Late Jurassic they were widespread throughout Eurasia and North America. The known Jurassic salamanders are more primitive than most, if not all, living forms. A single articulated skeleton (*Karaurus*) is known from the Kimmeridgian of Kazakhstan (**Figure 9C**).

During the Cretaceous, the group diversified and basal members of some of the modern lineages appeared, and by the Maastrichtian, most modern families are represented in the record or can be inferred to have been present. Articulated skeletons are known from the konservat-lagerstätten at Las Hoyas, Spain (**Figure 10**), and Liaoning, China, and there are large numbers of salamander-producing microvertebrate assemblages in Eurasia and North America. Almost all salamanders were still restricted to the northern continents, but some reached the African region of Gondwana and are known from Israel (*Ramonellus*), Sudan, and Niger (*Kababisha*).

Figure 10 The salamander *Valdotriton gracilis*, from the Lower Cretaceous, Las Hoyas, Spain. Specimen at Las Hoyas Museum, Spain. © Andrew Milner.

Caecilians

The caecilians are tiny, burrowing amphibians now found in moist soils and leaf litter in South America, Africa, the Seychelles, India, and South-east Asia. They are superficially earthworm-like in shape, with no limbs or tail and with rudimentary eyes, and a scale distribution that gives them a segmented appearance. They are poorly represented in the Mesozoic, with just three records. The Lower Jurassic *Eocaecilia* from Arizona is known from articulated material and retained small limbs (**Figure 9D**). The two Cretaceous

records from Morocco and Sudan are both based on isolated bones.

Albanerpetontids

The albanerpetontids are the extinct fourth group of lissamphibians; first recognized in 1976, subsequent discoveries have proved that they were widespread in the northern continents during the mid- to Late Mesozoic. They were small, superficially salamander-like forms (**Figure 9E**), uniquely characterized by the two halves of the lower jaw articulating anteriorly by means of an asymmetrical ball-and-socket joint (**Figure 9F**). Even small fragments of such jaws are readily recognized in microvertebrate assemblages. Articulated albanerpetontids are known from the konservat-lagerstätte at Las Hoyas, Spain, and this material is the key to understanding their morphology. Albanerpetontids first appear with the full suite of characteristics in the Bathonian, and they appear to have been present throughout Eurasia and North America in the Upper Jurassic and Cretaceous, with a recent report from the Lower Cretaceous of Morocco representing the first Gondwanan record.

See Also

Fossil Vertebrates: Palaeozoic Non-Amniote Tetrapods; Cenozoic Amphibians. **Lagerstätten**. **Palaeozoic:** End Permian Extinctions.

Further Reading

Bystrow AP and Efremov IA (1940) *Benthosuchus sushkini* Efr. A labyrinthodont from the Eotriassic of Sharjenga River. *Trudy Paleozoologicheskogo Instituta Akademii Nauk SSSR* 10: 1 152.

Holman JA (2003) *Fossil frogs and toads of North America*. Bloomington and Minneapolis: Indiana Press.

Milner AR (1990) The radiations of temnospondyl amphib ians. In: Taylor PD and Larwood GP (eds.) *Systematics Association Special Volume 42. Major Evolutionary Ra diations*, ch. 15, pp. 321 349. Oxford: Clarendon Press.

Milner AR (2000) Mesozoic and Tertiary Caudata and Albanerpetontidae. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palaeontology: The Evolutionary History of Amphibians*, ch. 18, pp. 1412 1444. Chipping Norton, NSW: Surrey Beatty.

Roček Z (2000) Mesozoic anurans. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palae ontology: The Evolutionary History of Amphibians*, ch. 14, pp. 1295 1331. Chipping Norton, NSW: Surrey Beatty.

Sanchíz B (1998) Salientia. In: Wellnhofer P. (ed.) *Hand buch der Paläoherpetologie*. part 4. Munich: Pfeil.

Schoch RR and Milner AR (2000) Stereospondyli. In: Wellnhofer P (ed.) *Handbuch der Paläoherpetologie*. part 3B. Munich: Pfeil.

Warren AA (2000) Secondarily aquatic temnospondyls of the Upper Permian and Mesozoic. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palae ontology: The Evolutionary History of Amphibians*, ch. 8, pp. 1122 1149. Chipping Norton, NSW: Surrey Beatty.

Cenozoic Amphibians

A R Milner, Birkbeck College, London, UK

Introduction

The fossil record suggests that archaic amphibian groups had entirely vanished by the Cenozoic, and that at least the majority of extant families were already established, many in the regions in which they still occur. In this article, the Cenozoic fossil record of each of the modern groups and of the extinct albanerpetontids is reviewed.

Amphibians and the Cretaceous and Tertiary Boundary

By the Late Cretaceous, the fossil record indicates that the only amphibians were the three modern lissamphibian groups (frogs, salamanders, and caecilians) and the enigmatic albanerpetontids. Such evidence as exists suggests that the Cretaceous and Tertiary (K–T) boundary event (*see* **Mesozoic:** End Cretaceous Extinctions) had no effect on the diversity of amphibians. In Montana, United States, and Alberta, Canada, there are freshwater faunas bracketing the Maastrichtian to Paleocene transition, and the frog and salamander assemblages are unchanged in diversity and taxonomic content through this time. Evidence from the rest of the world is consistent with this, but at too poor a resolution to be convincing in its own right.

Cenozoic Assemblages

Cenozoic amphibians are represented either as complete skeletons in assemblages in a few konservat-lagerstätten (*see* **Lagerstätten**) (localities, or 'lagerstätten', where prevailing conditions have produced highly conserved or diverse fossils) or, more commonly, as isolated elements in microvertebrate

accumulations. Notable lagerstätte assemblages are the Eocene brown coals from Messel and Geiseltal in Germany, and the Oligocene–Miocene freshwater limestones from Oeningen in Switzerland and from Bechlejovice in the Czech Republic. These produce large numbers of complete, articulated frogs and tadpoles and a smaller number of salamanders. Modified body tissues may be preserved, but the skeletons are generally crushed flat. In contrast, microvertebrate assemblages may contain few to abundant amphibian bones that are three-dimensionally preserved but are not associated, resulting in problems of association unless there is significant similarity to living taxa.

The Cenozoic record of caecilians is negligible and that of albanerpetontids is too patchy to evaluate in the absence of living descendants. The record of frogs and salamanders predominantly derives from North America and Eurasia, with frogs also well represented in South America. For salamanders, this 'Laurasian' range may represent a genuine pattern, because most living genera of salamander are found in the same area, but for frogs it is undoubtedly an artefact of the distribution of microvertebrate collecting. There is ample morphological and molecular evidence that the Neobatrachia or higher frogs originated in Gondwana and that they diversified during the breakup of that supercontinent. It can be anticipated that a substantial record will eventually be found in Africa, India, and Australia as well as in South America.

Aquatic amphibians living in lakes and ponds are preserved most readily and so are disproportionately well represented in the fossil record. The frog families Pipidae and Palaeobatrachidae (Figure 1) and the salamander family Cryptobranchidae fall into this category. At the other extreme, stream-dwelling amphibians and those living in damp terrestrial environments are hardly represented at all. The salamander family Plethodontidae, for example, comprised mainly of such forms, is the largest family of living salamanders but has one of the poorest fossil records.

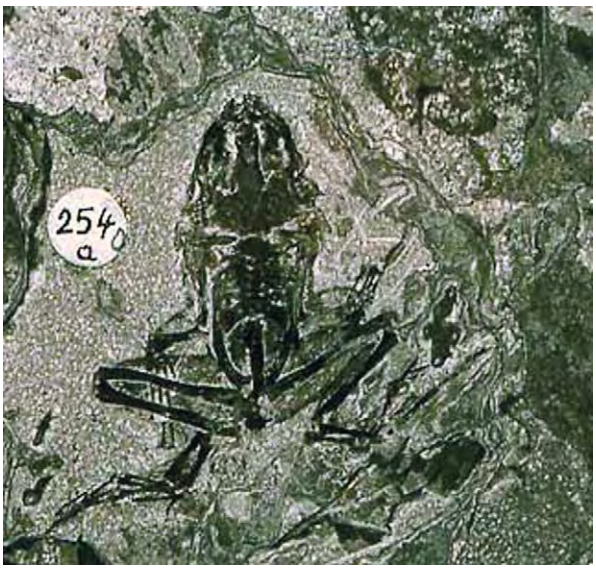


Figure 1 The palaeobatrachid frog *Palaeobatrachus grandipes*, from the Oligocene, Bechlejovice, Czech Republic. Specimen at the National Museum, Prague. © Andrew Milner.

Frogs

Frogs occur on all non-polar continents and on many islands, particularly those such as New Zealand, Madagascar, and the Seychelles, which are relicts of larger continents. Early in Mesozoic frog evolution, there appears to have been a major dichotomy between Laurasian and Gondwanan frogs, with the discoglossids, pelobatids, pelodytids, and palaeobatrachids evolving in Laurasia, and the pipids and neobatrachians (~20 families) evolving in Gondwana. By the Cenozoic and up to the present day, this division is still recognizable in the distribution of frog families. However, during the Cenozoic, several successful Gondwanan families, the ranid frogs (Figure 2), bufonid toads, and hylid tree frogs, extended their ranges into Laurasia, diluting the original pattern. First appearances of fossils of these families in the northern continents show this to have happened in the Paleocene and Eocene, for ranids and bufonids.

Much of the Cenozoic fossil record of frogs comprises earlier representatives of the extant frog faunas of each continent, e.g., leptodactylids in South America, pipids and ranids in Africa, myobatrachids in Australia, and discoglossids and pelobatids in Europe. Occasionally, however, frog fossils reveal that past faunas had distinctive components. *Latonia gigantea*, for example, a discoglossid from the Oligocene–Pliocene of Europe, was twice as large as any modern discoglossid and had a heavily sculptured skull.

Salamanders

The major diversity of living salamanders occurs in North America, central America, and Eurasia. A few genera are found in South America and North Africa, and these are believed to represent Neogene range extensions. The Cenozoic record is consistent with this, with much of the fossil material belonging to the same families and genera that occur in each continent today. Most of the European fossil record is of salamandrids and proteids; the Asian record comprises salamandrids, hynobiids, and cryptobranchids; and the North American record is of cryptobranchids, sirenids, amphiumids, proteids, ambystomatids, and

Figure 2 The ranid frog *Rana ridibunda* (*'Rana pueyoi'*), from the Miocene, Libros, Spain. Specimen at the Natural History Museum, London. © Andrew Milner.

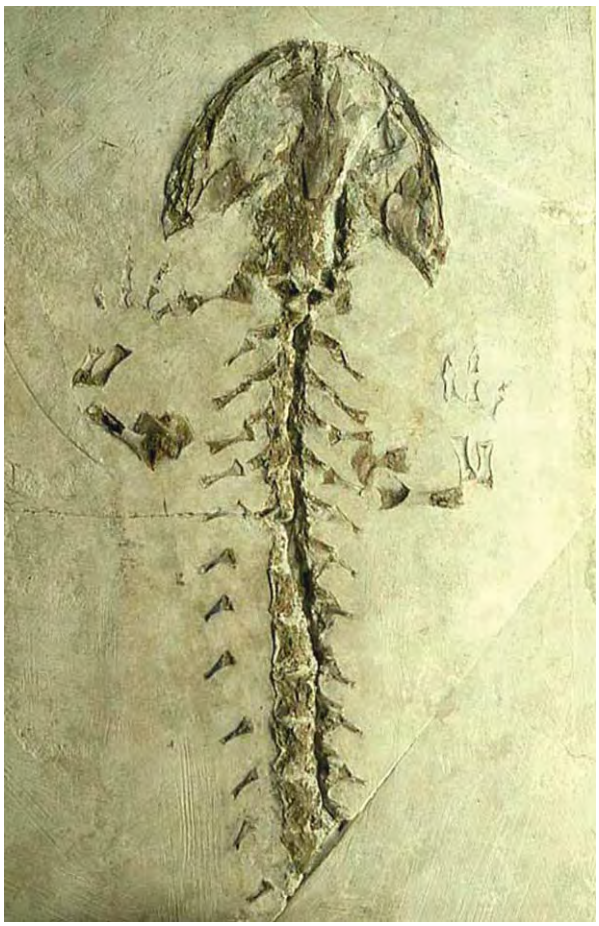


Figure 3 The cryptobranchid salamander *Andrias scheuchzeri,* from the Oligocene, Oeningen, Switzerland. Specimen at the Museum fur Naturkunde, Humboldt University, Berlin. © Andrew Milner.

plethodontids. Some records fall outside this pattern. A large Paleocene–Eocene salamander, *Piceoerpeton* (of uncertain relationships), not only occurs within the modern range of salamanders in North America but has been found as far north as Ellesmere Island in Eocene beds, demonstrating a major Paleogene range extension for the group. The large aquatic cryptobranchid salamanders appear to have originated in Asia, where they occur as Paleocene fossils and the living *Andrias*. *Andrias* (Figure 3) first appears in Europe in the Late Oligocene, after the 'Grand Coupure' event, a combination of climate change and faunal dispersal between Europe and Asia. *Andrias* fossils are abundant from the Late Oligocene to the Pliocene, after which the species became extinct in Europe.

Caecilians

Caecilians are now found in South America, Africa, the Seychelles, India, and south-east Asia. Phylogenetic evidence from living forms indicates that the first four areas represent relicts of an original west–central Gondwana range, whereas the south-east Asian taxa represent a range extension out of India after its contact with Asia. Cenozoic fossils might be expected in any of these areas, but, in practice, caecilian bones are so rare and so tiny that the only discoveries to date are two single vertebrae from the Paleocene, one from Brazil and one from Bolivia.

Albanerpetontids

The Albanerpetontidae are a group of Jurassic–Pliocene amphibians; they were small salamander-like animals but are structurally distinct from any of the living groups. A unique feature is that the two halves of the lower jaw articulate anteriorly by means of an asymmetrical ball-and-socket joint. Albanerpetontid bones had been misassociated with those of salamanders for many years, but were finally recognized in 1976 from Miocene microvertebrate material from La Grive St Alban in France, this material forming the original *Albanerpeton inexpectatum* (Figure 4). Most subsequent discoveries have

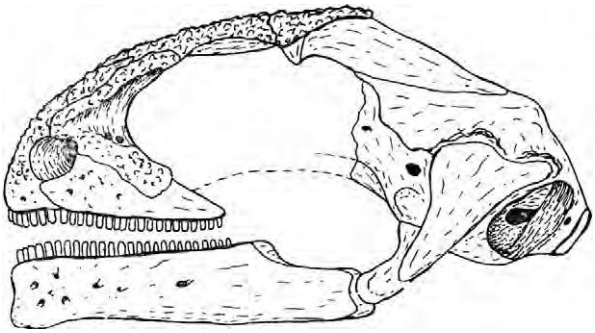


Figure 4 The skull of the Miocene amphibian *Albanerpeton inexpectatum*, from La Grive St Alban, France. Reproduced with permission from Estes R and Hoffstetter R (1976) Les urodèles du Miocène de la Grive Saint Alban (Isère, France). *Bulletin du Muséum National d'Histoire Naturelle, Sciences de la Terre* 57: 297 343.

been Mesozoic, but further finds have been made in Paleocene material of North America, Miocene material of Germany, and Pliocene material of Romania. At present, it appears that a few lineages survived the K–T event in the northern continents, making the last appearance of these species in North America in the Paleogene, but lingering on in Eurasia until the Pliocene, when they appear to have died out during the glaciations.

Amphibians and the Pleistocene Glaciations

The sequence of Pleistocene glaciations appears to have had a dramatic effect on the amphibian fauna of Europe. In part, this is undoubtedly because the southward movement of aquatic vertebrates was constrained by the Pyrenees, the Alps, the Carpathians, and, more profoundly, the Mediterranean Sea. Several genera of amphibians, widespread as fossils over Early Neogene Europe, are now restricted to various permutations of the Iberian Peninsula, Italy, and the Balkans (the salamander genera *Pleurodeles* and *Chioglossa* and the frog genus *Discoglossus*). More dramatically, four families became extinct in Europe, the hynobiid and cryptobranchid salamanders surviving elsewhere, but the palaeobatrachid frogs and albanerpetontids dying out entirely, Europe apparently being their last refugium.

See Also

Lagerstätten. **Mesozoic:** End Cretaceous Extinctions.

Further Reading

Báez AM (2000) Tertiary Anura of South America. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palaeontology: The Evolutionary History of Amphibians* ch. 15, pp. 1388 1401. Chipping Norton, NSW: Surrey Beatty.

Holman JA (1995) *Pleistocene Amphibians and Reptiles in North America. Oxford Monographs on Geology and Geophysics, 32*. New York, Oxford: Oxford University Press.

Holman JA (1998) *Pleistocene Amphibians and Reptiles in Britain and Europe. Oxford Monographs on Geology and Geophysics, 38*. New York, Oxford: Oxford University Press.

Holman JA (2003) *Fossil Frogs and Toads of North America*. Bloomington and Minneapolis: Indiana Press.

Milner AR (2000) Mesozoic and Tertiary Caudata and Albanerpetontidae. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palaeontology: The Evolutionary History of Amphibians* ch. 18, pp. 1412 1444. Chipping Norton, NSW: Surrey Beatty.

Rage J C and Roček Z (2003) Evolution of anuran assemblages in the Tertiary and Quaternary of Europe, in the context of palaeoclimate and palaeogeography. *Amphibia Reptilia* 24: 133 167.

Roček Z and Rage J C (2000) Tertiary Anura of Africa, Asia, Europe, North America and Australia. In: Heatwole H and Carroll RL (eds.) *Amphibian Biology, Volume 4, Palaeontology: The Evolutionary History of Amphibians*. ch. 15, pp. 1334 1389. Chipping Norton, NSW: Surrey Beatty.

Sanchíz B (1998) Salientia. In: Wellnhofer P (ed.) *Handbuch der Paläoherpetologie*, part 4. Munich: Pfeil.

Mesozoic Mammals

Z-X Luo, Carnegie Museum of Natural History, Pittsburgh, PA, USA

Introduction

Two-thirds of mammalian history occurred during the Mesozoic, when dinosaurs dominated the land. During their long history in the Mesozoic, mammals diversified into many branches of a prolific evolutionary tree and underwent an enormous morphologic evolution. To date, more than 280 genera of Mesozoic mammals and mammaliaforms are known to science, and these belong to more than 25 distinctive family-level lineages, each of which has unique dental characteristics. Only a tiny fraction of these diverse Mesozoic mammalian groups survived to give rise to the three modern mammalian groups: the egg-laying monotremes, the pouched marsupials, and the placentals, with an elaborated placental reproductive structure.

Mesozoic mammals represent the trunk and the basal branches of the entire mammalian family tree, and they provide evidence of the ancestral condition from which modern mammals have evolved. Humans are primates, primates are placentals, placentals are eutherians, and eutherians are rooted in the evolutionary past of all Mesozoic mammals. The fossil record of Mesozoic mammals, therefore, is indispensable for understanding the deep history that gave rise to extant mammals, including the placental lineage to which humans belong.

Mammals are part of a more inclusive evolutionary lineage known as the synapsids, which are distinguishable from other non-synapsid vertebrates by the presence of a lower temporal fenestra in the skull. The earliest known fossils of the synapsid lineage are from the Late Carboniferous (300 million years ago) in the Palaeozoic. Cynodonts are a derived subgroup of synapsids; they emerged in the Late Permian, around 250 million years ago, and thrived in the Triassic, with some relicts surviving into the Early Cretaceous. Cynodonts are distinguished from precynodont synapsids in having a better developed secondary bony palate that separates the nasal passage (for breathing) from the mouth cavity (for feeding), an enlarged temporal fenestra, and a larger coronoid process on the mandible for the better developed jaw adductor muscles, plus a functional middle ear in the posterior part of the mandible.

Mammaliaforms are a subgroup of cynodonts (Figure 1), and they are distinguishable from premammaliaform cynodonts in three ways: by the presence of a derived jaw hinge, formed by the dentary bone of the mandible and the squamosal bone of the cranium; by the more derived features of the ears, such as the promontorium, which encloses the inner ear cochlea; and by precise dental occlusion. Mammaliaforms are close relatives of modern mammals, with many very mammal-like features; but they are not advanced enough to be placed within the evolutionary lineage of modern mammals. Many mammaliaforms are known from the fossil record of the Late Triassic through the Late Jurassic.

Modern Mammalia are a subgroup of mammaliaforms, and they are defined as an evolutionary lineage consisting of the ancestor common to all three living mammalian groups, plus those fossil taxa that can be placed within these living mammal groups by comparative morphological evidence in the bone and dental structures.

Origins of Mammalian Features

Modern mammals are more derived than most non-mammalian vertebrates are, in five aspects: (1) the jaw hinge formed by the dentary condyle and squamosal glenoid (instead of the quadrate and articular bones), (2) the middle ear suspended in the base of the cranium (and detached from the mandible), (3) the inner ear with elaborated cochlear canal, a part of the auditory bony labyrinth structure, (4) the enlarged brain endocast and a series of braincase structures to support and protect the brain, and (5) a diphyodont (two generations) dental replacement (instead of multiple dental replacements as in toothed modern reptiles). Some precursor conditions to these derived evolutionary features of modern Mammalia can be traced to the mammaliaforms in the Late Triassic to Early Jurassic (Figure 1); other mammal-like precursor features can be seen in the advanced cynodonts. By mapping the patterns of distribution of various precursor conditions of mammalian features on the successive hierarchies of mammals, mammaliaforms, and premammaliaform cynodonts on the evolutionary tree (Figure 2), it is possible to infer the historical sequence through which the derived mammalian features arose during the transition from premammaliaform cynodonts to the earliest mammaliaforms, and then from mammaliaforms to mammals. Mammals underwent significant anatomical evolution in their early history; the fossil evidence of this evolution is so extensive that the origins of mammals

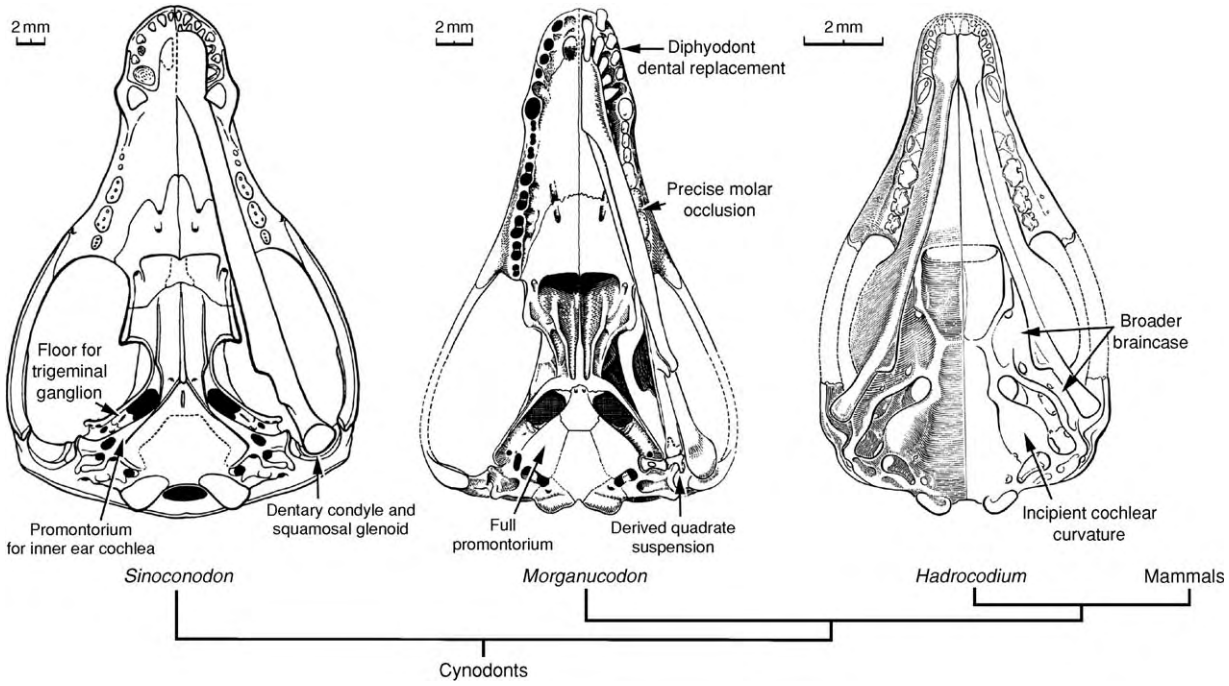


Figure 1 Mammaliaforms (skulls in ventral view) of the Late Triassic and Early Jurassic, showing the derived (diagnostic) mammalian characteristics that evolved with these transitional mammaliaforms. *Sinoconodon*, from the Lower Jurassic, is one of the most primitive mammaliaforms known; it has a jaw hinge of dentary condyle and squamosal glenoid, a petrosal promontorium for the cochlear canal, and a braincase floor for the trigeminal cranial nerve ganglion. *Morganucodon*, from the Late Triassic and Early Jurassic, is the earliest known mammaliaform with precise occlusion of the upper and lower molars and a diphyodont dental replacement pattern. *Hadrocodium* from the Early Jurassic, is the most derived among mammaliaforms in having a more enlarged braincase, an enlarged promontorium with curved cochlea, and the absence of primitive mandibular structures for attaching the middle ear. (Morganucodon modified from KA Kermack *et al.* (1981). *Zoological Journal of Linnean Society*. London 71: 1 158.)

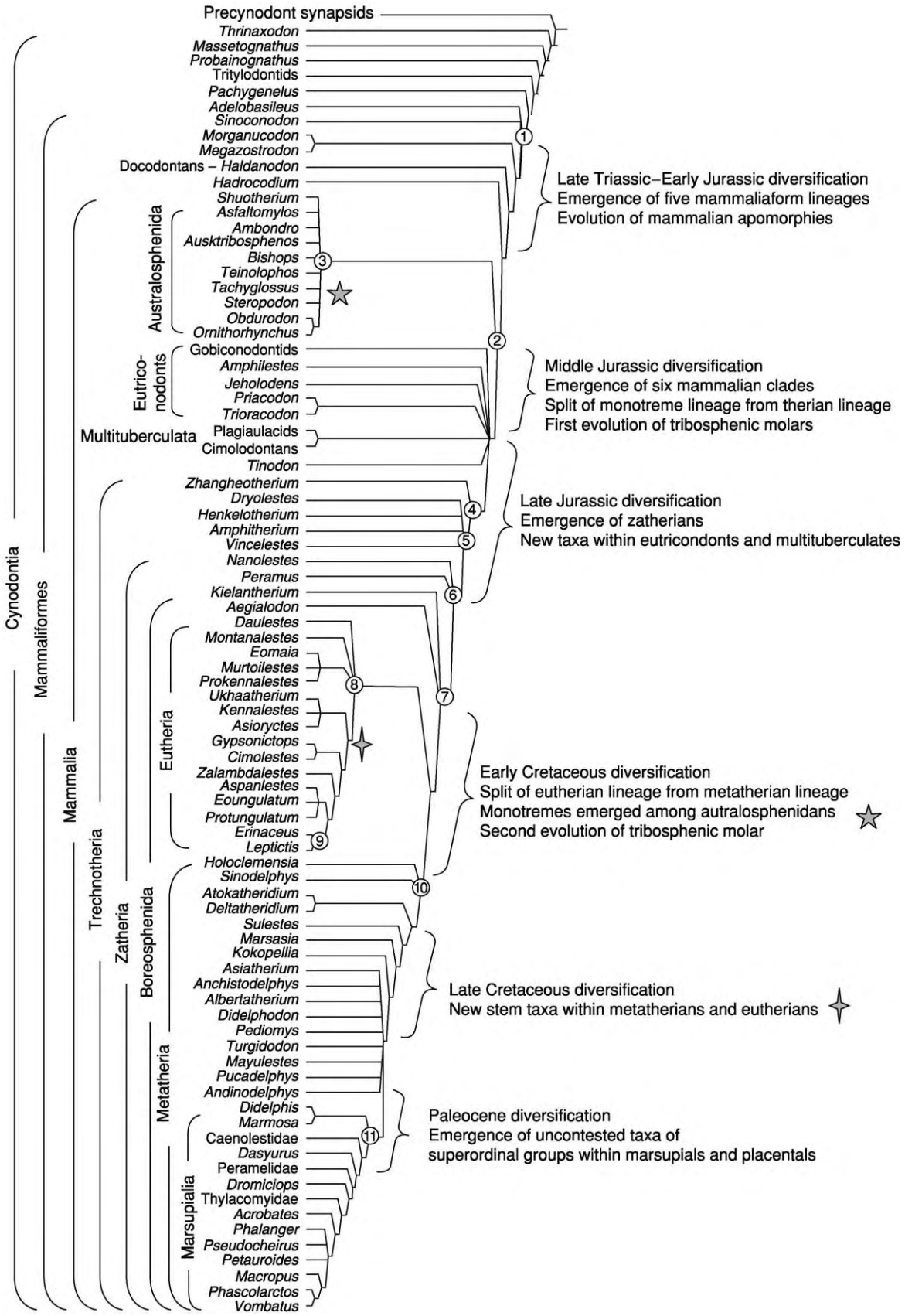
from premammalian cynodonts is one of the best documented cases for the origin of a major vertebrate group.

Jaw Hinge and Middle Ear

The earliest mammaliaforms are characterized by a 'double joint' in their jaw hinge (Figure 2: *Morganucodon*). They developed the derived (mammalian) dentary condyle to squamosal glenoid in the jaw joint, but also retained the primitive (cynodont) features of an articular-quadrate jaw joint. The earliest mammaliaforms are similar to premammaliaform cynodonts in that the middle bones (also known as the postdentary bones) and the ossified Meckel's cartilage are attached anteriorly to the mandible, in contrast to extant mammals, in which the embryonic Meckel's cartilage disappeared and the middle ear bones are separated from the mandible in adults. The quadrate (homologue to the incus of the mammalian middle ear) has developed contact with the petrosal bone in some mammaliaforms, as in modern mammals. In other words, the earliest mammaliaforms have achieved the diagnostic mammalian features (the dentary/squamosal jaw joint and the way in which the middle ear is suspended by the cranium via the quadrate) (Figure 2).

Other features of the jaw hinge and the middle ear in these Jurassic mammaliaforms are primitive features shared by more distant premammaliaform cynodonts. In the more derived eutriconodont mammals, the middle ear elements, such as the angular (homologue to the tympanic in mammals), achieved some

Figure 2 Evolutionary relationships of Mesozoic mammals, mammaliaforms, and cynodonts. The phylogenetic ranks of Mesozoic mammals and mammaliaforms correspond approximately to the temporal sequence of their diversifications. Taxic macroevolution of Mesozoic mammals is characterized by several successive episodes of diversifications and emergence of new clades with new dental features. Numbers represent nodes for successive ranks or hiearchies of Mesozoic mammalian phylogenetic relationships: 1, Mam maliaformes; 2, Mammalia; 3, Australosphenida; 4, Trechnotheria; 5, Cladotheria; 6, Zatheria; 7, Boreosphenida; 8, Eutheria; 9, Placentalia; 10, Metatheria; 11, Marsupialia. (Cladogram modified from Z X Luo *et al.* (2003) *Science* 302: 1934 1940.)


Precynodont synapsids
Thrinaxodon
Massetognathus
Probainognathus
Tritylodontids
Pachygenelus
Adelobasileus
Sinoconodon
Morganucodon
Megazostrodon
Docodontans – Haldanodon
Hadrocodium
Shuotherium
Asfaltomylos
Ambondro
Ausktribosphenos
Bishops
Teinolophos
Tachyglossus
Steropodon
Obdurodon
Ornithorhynchus
Australosphenida
Gobiconodontids
Amphilestes
Jeholodens
Priacodon
Trioracodon
Eutrico-nodonts
Multituberculata
Plagiaulacids
Cimolodontans
Tinodon
Zhangheotherium
Dryolestes
Henkelotherium
Amphitherium
Vincelestes
Nanolestes
Peramus
Kielantherium
Aegialodon
Daulestes
Montanalestes
Eomaia
Murtoilestes
Prokennalestes
Ukhaatherium
Kennalestes
Asioryctes
Gypsonictops
Cimolestes
Zalambdalestes
Aspanlestes
Eoungulatum
Protungulatum
Erinaceus
Leptictis
Eutheria
Holoclemensia
Sinodelphys
Atokatheridium
Deltatheridium
Sulestes
Marsasia
Kokopellia
Asiatherium
Anchistodelphys
Albertatherium
Didelphodon
Pediomys
Turgidodon
Mayulestes
Pucadelphys
Andinodelphys
Didelphis
Marmosa
Caenolestidae
Dasyurus
Peramelidae
Dromiciops
Thylacomyidae
Acrobates
Phalanger
Pseudocheirus
Petauroides
Macropus
Phascolarctos
Vombatus
Marsupialia
Metatheria
Boreosphenida
Zatheria
Trechnotheria
Mammalia
Mammaliformes
Cynodontia
1
2
3
4
5
6
7
8
9
10
11
Late Triassic–Early Jurassic diversification
Emergence of five mammaliaform lineages
Evolution of mammalian apomorphies
Middle Jurassic diversification
Emergence of six mammalian clades
Split of monotreme lineage from therian lineage
First evolution of tribosphenic molars
Late Jurassic diversification
Emergence of zatherians
New taxa within eutricondonts and multituberculates
Early Cretaceous diversification
Split of eutherian lineage from metatherian lineage
Monotremes emerged among autralosphenidans
Second evolution of tribosphenic molar
Late Cretaceous diversification
New stem taxa within metatherians and eutherians
Paleocene diversification
Emergence of uncontested taxa of
superordinal groups within marsupials and placentals

degree of medio-lateral separation from the mandible, whereas the middle ear bones are connected anteriorly to the mandible via the Meckel's cartilage, even in adults. The best available evidence suggests that loss of the Meckel's cartilage may have occurred separately in modern monotremes and in the lineage of placentals plus marsupials, after the ancestor of the latter groups had split from eutriconodontans.

Elaborated Structure of Inner Ear Cochlea

Mammals are most derived among living vertebrates in their complex inner ear cochlea and related hearing adaptation. The petrosal bone has developed an enlarged bony housing for the cochlear part of the inner ear; this is crucial for sensitive hearing function, especially for high-frequency sound. Development of the derived inner ear and its bony housing occurred in several evolutionary steps in the transition from cynodonts and early mammals, as evidenced by the precursor conditions of these structures in mammaliaforms.

Premammalian cynodonts have a cochlear cavity for the auditory function in the inner ear. This cochlear cavity is small and globular; it does not extend anterior to the fenestra vestibuli in the reconstructed inner ear endocast. The bony housing for the inner ear is formed by a mosaic of multiple bones. Two of these bones, the prootic and the opisthotic bones, correspond to the homologous bones in extant nonmammalian vertebrates, which are developed from ossification of the embryonic auditory capsule in modern vertebrates. None of the bones surrounding the inner ear are fused. The bony housing of the inner ear in adults of the cynodont *Thrinaxodon* is formed by a mosaic of several bones, which can be either endochondral or intramembranous in embryonic origin among extant vertebrates.

By comparison, the inner ear of Early Jurassic mammaliaforms has an elongate bony cochlear canal that is much better developed compared to the cochlear cavity of premammaliaform cynodonts. The elongate bony cochlear canal may indicate a greater sensitivity to high-frequency sound, which is very important in the hearing function of all extant mammals and was probably important for at least some of the earliest mammals. The bony housing of the inner ear in stem taxa of mammals is formed exclusively by the petrosal, which is the single bone homologous to the fused prootic and opisthotic elements in premammaliaform cynodonts. The petrosal is not only much larger than the prootic and opisthotic, but it also forms a distinctive structure of the promontorium. The enlarged petrosal excludes the other cranial bones, such as the basisphenoid complex, the basioccipital, and the exoccipital, from the bony housing for the inner ear. The mosaic of multiple bones for the inner ear housing of cynodonts is replaced by a single bone in the derived mammaliaforms, including modern mammals.

The differences between mammals and cynodonts in the inner ear and its bony housing are the result of their correlated structural transformation. Enlargement of the promontorium is correlated with elongation of the cochlear canal. The enlarged promontorium displaced the neighbouring sphenoid complex and basioccipital bone. Inflation of the bulbous promontorium in the mammalian crown group is associated with coiling of the cochlear canal. If the precursor condition in mammaliaforms is mapped on the cynodont-mammal evolutionary tree, there is a clear pattern of incremental evolution of mammalian characteristics in the inner ear cochlear canal and the inner ear bony housing.

The Larger Brain

In early mammalian evolution, there is a clear tendency towards a larger volume of brain endocasts, and their morphological features are also better differentiated in the more derived taxa. The relative brain sizes of the transitional mammaliaforms of the Early Jurassic are intermediate between the larger brain size of the derived Cretaceous mammals and the smaller brain size of the Triassic cynodonts. Larger brain capacity of mammals compared to that of nonmammalian vertebrates indicates a larger volume of metabolically expensive neural tissues of the mammalian brain. This can be further correlated to the development of a more elevated metabolism in early mammaliaforms than in premammaliaform cynodonts. It could also be correlated with better sensory perception and elaborated neural control of the skeleto-muscular system for mastication and locomotion.

In the larger brain endocasts in the successively more derived mammaliaforms and mammals, the cerebral hemispheres are much better developed in the successively more derived groups. The posterior part of the brain endocast tends to show a better differentiation of the mid-brain from the cerebellar structure in the mammalian crown groups: metatherians, eutherians, and monotremes. Related to the tendency of increasing brain size, new structures in the braincase in the derived mammaliaforms are absent in more primitive cynodonts. For example, the braincase of mammaliaforms has a bony floor to enclose the trigeminal ganglion of the cranial nerve V into the braincase (**Figure 1**: *Sinoconodon*). The enlarged brain is correlated with a posterior shift of the braincase relative to the jaw hinge and other skull structures. In most premammaliaform cynodonts, the

floor to the narrow anterior braincase was not ossified and the related orbital structure of the skull is also absent. The floor for the anterior part was presumably formed by the cartilaginous structure of the cranium in life. In mammaliaforms, by comparison, the floor of the anterior braincase is fully ossified, as are most of the orbital structures.

Diphyodont Dental Replacement

The diphyodont dental replacement in modern placental mammals is characterized by a single tooth replacement for incisors, canines, and premolars, but the molars have only one generation and are never replaced. Marsupial mammals show a more reduced dental replacement compared to the typical diphyodont replacement of placentals. The number of replacements per tooth locus in modern mammals is far smaller than in the polyphyodont ('multiple generations') replacement of nonmammalian vertebrates with teeth, as in crocodiles, dinosaurs, and premammaliaform cynodonts.

Evolution of reduced dental replacement rates in mammals is correlated with lactation. The reduction of dental replacement is feasible because lactation after birth makes it possible for mammalian neonates to achieve fast cranial growth without teeth, resulting in a delay in eruption of the deciduous teeth. Also, because relatively fast early cranial growth of the mammalian skull during lactation will slow after weaning, skull growth terminates relatively early in the life of individuals, accompanied by an early termination of dental replacement. The end of skull growth (determinate pattern) usually coincides with the eruption of the last molar. Therefore, diphyodont dental replacement is a crucial and derived feature that indicates the life history characteristics of modern mammals.

Some precursor conditions to modern mammalian dental replacement are seen in *Sinoconodon,* one of the earliest mammaliaform lineages to undergo a split. *Sinoconodon* has mammalian characteristics in the jaw hinge and in the ear region (**Figure 1**), but has retained some cynodont-like characteristics of the dental replacement, and a 'reptilian' pattern of skull growth. The incisors and canines in *Sinoconodon* were replaced at least three times in an alternating pattern, as seen in many cynodonts and other toothed non-mammalian vertebrates. The posterior molars of *Sinoconodon* have one replacement in the larger (presumably older) individuals, in a manner that is similar to the replacement pattern of some cynodonts, such as diademodontids.

Replacement of the postcanines (premolars and molars) in *Sinoconodon* is sequential in the anteroposterior direction, a mammal-like feature. But after the replacement of the premolars, the premolars and anterior molars are lost, resulting in a dental diastema (gap in the dental series) that becomes increasingly larger in older individuals. Coupled with successive addition of the newly erupted molariforms at the posterior end of the tooth row, there is a posterior shift of the functional tooth rows in the jaws. These primitive features are present in some cynodonts and also in other mammaliaforms to a lesser extent.

The currently available sample of *Sinoconodon* specimens shows a large range of growth, from the smallest individual, with an estimated body mass of about 13 g and already showing the canine replacement, to the largest, with an estimated body mass of more than 500 g. During this growth, the posterior molariforms were being replaced while the upper and lower jaws continued to lengthen in the successively older individuals. This suggests that *Sinoconodon* had indeterminate growth at least in its skull, associated with continuous tooth replacement, as in premammaliaform cynodonts. Given its basal position in the mammalian family tree, the characteristics of the dental replacement and skull growth of *Sinoconodon* can be regarded as an intermediate stage in the evolution from the primitive pattern of polyphyodont replacement seen in most cynodonts to the derived diphyodont replacement of mammals. Reduction in premolar and molariform replacement in the posterior dentition preceded reduction of the replacement of incisors and canines in the anterior dentition.

Interestingly, the mammaliaforms (e.g., *Morganucodon* and docodontans) that are closer to modern mammals all have the typical diphyodont replacement of modern mammals and a narrower range of skull growth similar to that of modern mammals. When dental replacement characteristics are mapped on the mammalian evolutionary tree, it helps to show that the transition from reptile-like skull growth and dental replacement to the typical mammalian determinate skull growth and diphyodont dental replacement occurred in the Late Triassic through Early Jurassic. Related to diphyodont dental replacement, derived mammaliaforms *Morganucodon*, *Hadrocodium*, and docodontans also evolved precise occlusion of molars, in which the upper and lower molars developed wear facets associated with individual molar cusps. *Sinoconodon* lacked precise molar occlusion because it did not have a consistent pattern of opposition between upper and lower molars, as a consequence of the partial replacement of the posterior molariforms and the successive posterior shift of the functional tooth row, as part of the indeterminate growth pattern of the skull. However, *Morganucodon* and other taxa more closely related to modern mammals than *Sinoconodon* have developed the

precise cusp occlusion between the upper and lower molars for better and effective mastication of food.

Successive Diversifications of Mesozoic Mammals

Compared to other small vertebrates in the terrestrial biota, Mesozoic mammals were never numerically abundant. Most taxonomic records of Mesozoic mammals are based on teeth and jaw fragments. Nonetheless, sufficient numbers of their fossils have been collected for palaeontologists to know that there were several episodes of diversification of Mesozoic mammals. The first burst of diversification occurred as several mammaliaforms emerged during the Late Triassic to Early Jurassic. Although these mammaliaforms have almost identical mandibular design, their dentitions are different. The five main groups of mammaliaforms in the Triassic and Early Jurassic have developed three distinctive types of specialized teeth. Haramiyidans have multirow, multicusped, or 'multituberculate-like' molars; *Sinoconodon*, morganucodontans, and *Hadrocodium* have 'triconodont-like' molars showing three main cusps in alignment. Kuehneotheriids have triangulated, 'symmetrodont-like' molars. The phyletic diversification is characterized most prominently by the differentiation in dental morphology.

The Middle Jurassic diversification saw the emergence of six more order- or family-level evolutionary lineages: amphilestids, with triconodont-like molars; eleutherodontans, with multituberculate-like molars; spalacotheriids, with 'symmetrical' and triangulated molars; and amphitheriids and peramurids, with a triangulate trigonid plus a talonid heel on the lower molars. These groups are known mostly from the Laurasian continents during the Middle Jurassic. A major morphological innovation in mammalian evolution of this time was the molar structure for grinding function, which evolved in at least three separate Middle Jurassic lineages: docodontans, with complex shearing and grinding surfaces on the molars; shuotheriids, with an anterior grinding basin ('pseudotalonid'); and australosphenidans from Gondwana continents, with the tribosphenic molars capable of the shearing ('sphen') and the mortar–pestle crushing and grinding ('tribo'). The tribosphenic lower molar has a posterior talonid basin, functioning as a mortar that receives the inner main cusp (protocone) of the upper molar that is analogous to a pestle. These advanced features of crushing and grinding of tribosphenid mammals are developed in addition to the primitive shearing features on the molars that are also shared by their pretribosphenic relatives. It is obvious that the adaptive molar structures for grinding functions are homoplastic among these groups.

Consistent with this theory of homoplastic evolution of the grinding functions of molars is the fact that the derived grinding features of multifunctional molars in docodontans, shuotheriids, and at least some Middle Jurassic australosphenidans co-existed with the primitive features of the postdentary trough on the mandibles of these groups. The combined analyses of these primitive mandibular features and the derived molar features in the Jurassic southern mammals suggest that the australosphenidans or the southern tribosphenic mammals are not closely related to boreosphenidans, or the northern tribosphenic mammals. The latter group evolved some 25 million years later in the Laurasian continents, as a separate lineage with a convergent molar design, but far more derived mandibular features.

The third episode of diversification among early mammals occurred in the Late Jurassic, with five newly evolved order-or family-level lineages: multituberculata (*sensu stricto*), triconodontids, spalacotheriids, tinodontids, and dryolestoids. The most significant apomorphic feature of these emergent groups of the Late Jurassic is the absence of the postdentary trough on the mandible, which is a primitive feature retained by almost all Early Jurassic and Middle Jurassic lineages (except for amphilestids and *Hadrocodium*). Other striking similarities are in the structure of the dentary. Despite the fact that eutriconodontans, multituberculates, and spalacotheriids have very different dentitions, these three groups all have fairly similar characteristics in the posterior part of the mandible – a rounded 'angular' region grades into the dentary condyle, and there is often a prominent medial pterygoid crest ('shelf') along the ventral border of the mandible.

The last major episode of diversification occurred in the Early Cretaceous when the stem taxa of metatherians, eutherians, and basal boreosphenidans appeared in Laurasia, while toothed monotremes and some of their australosphenidan relatives appeared in Gondwana. The only new major lineage to appear in the Late Cretaceous was that of gondwanatherians from several southern landmasses. The existing lineages of eutherians, metatherians, and multituberculates greatly increased both in generic diversity and in numerical abundance throughout the Cretaceous and eventually survived (albeit with varied success) the mass extinction at the end of the period. Dryolestoids survived into the Cenozoic in South America. However, various mammaliaform clades, spalacotheriids, eutriconodontans, and gondwanatherians had declined and were extinct by the end of Cretaceous.

Rise of Modern Monotremes, Marsupials, and Placentals

Mammals now thrive throughout the world, with about 4600 species. Only the three extant species of egg-laying monotremes survive today, and all are endemic to Australia. Although extant adult monotremes lack teeth, their extinct relatives from the Miocene of Australia have teeth, and their dental fossils have helped to trace the history of monotremes to three earliest known toothed monotremes in Australia, from 110 million years ago. One toothed monotreme was also discovered from the Paleocene of Argentina, suggesting that monotremes had a wider distribution on Gondwana landmasses during the Mesozoic.

The earliest monotremes shared some derived premolar and molar features with the southern tribosphenic mammals or australosphenidans that existed on the Gondwana landmasses from the Middle Jurassic to Early Cretaceous. The earliest monotremes and other australosphenidans also bear resemblance in their primitive mandibular features. It has been suggested that monotremes are a surviving lineage of the more diverse and more widely spread australosphenidans endemic to the Gondwanan landmasses, when the Gondwanan continents were more widely separated from Laurasian continents during the Middle Jurassic through the Early Cretaceous. This hypothesis of monotreme evolution is contingent on morphological evidence in support of a close relationship between monotremes and the southern tribosphenic mammals. Although corroborated by a recent discovery of an australosphenidan mammal from the Middle Jurassic of Argentina, the putative close relationship of monotremes to other australosphenidan mammals has been questioned. According to the hypothesis, the australosphenidan lineage including modern monotremes had split from the rest of Mesozoic mammals before 165 million years ago (**Figure 2**: Middle Jurassic diversification).

Placentals and marsupials dominated the world's terrestrial biotas for 65 million years in the Cenozoic, culminating in a vast array of species that make up 99.9% of all mammals today. Modern placentals have a worldwide distribution, but modern marsupials are mostly restricted to the Australian region and South America, with only one species in North America. In the Mesozoic, their geographic distribution was much different. The uncontested Cretaceous relatives to modern placentals and marsupials are all from Laurasia, as are the stem boreosphenidans, the nearest relatives of the common ancestor of modern marsupials and placentals.

The earliest fossils that can be reliably attributed to the placental lineage (eutherians) and the marsupial lineage (metatherians) are from the Yixian Formation of north-eastern China, dated to 125 million years. The earliest known eutherian, *Eomaia scansoria*, and the earliest known metatherian, *Sinodelphys szalayi* (**Figure 3**), have provided evidence on the earliest skeletal evolution, during the divergence of the placental and marsupial lineages. By the Late Cretaceous, there are dozens of taxa that can be unambiguously

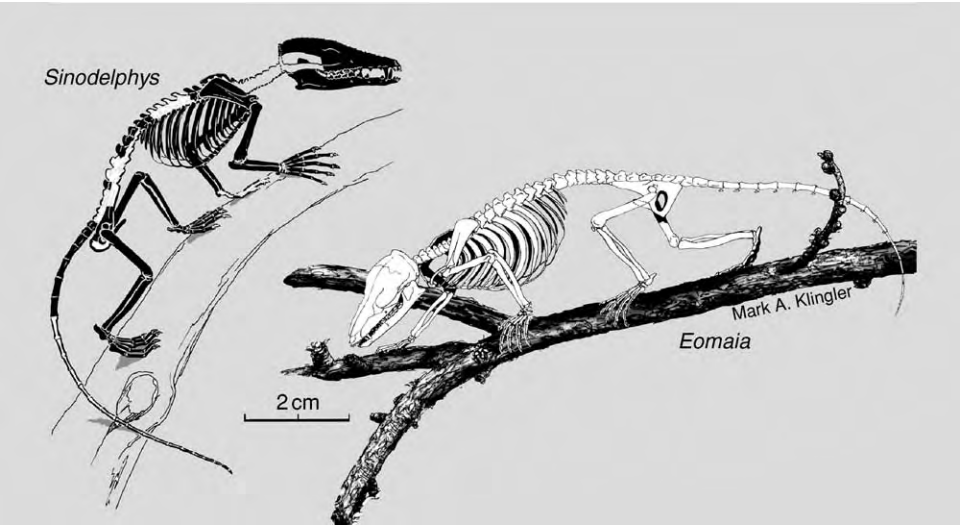


Figure 3 Restoration of the earliest known eutherian *Eomaia scansoria* and the earliest known metatherian *Sinodelphys szalayi*. *Eomaia* skeletal reconstruction is a composite from two specimens. *Sinodelphys* reconstruction is based on a partial skeleton, with preserved parts in black. Available skeletal evidence from these fossils shows adaptations that would allow climbing on uneven substrates and walking on branches. Illustration by M A Klingler, Carnegie Museum of Natural History, with permission.

assigned to metatherians and eutherians. Eutherians and metatherians can be distinguished by many characteristics of molars, anterior dentition, ear region, ankle bones, and wrist bones. The Late Cretaceous metatherians can be recognized by the presence of three premolars and four molars (seven positions of postcanines), with a single replacement at the last premolar, an inflected angular process of the mandible, and its posterior shelf, the masseteric muscle fossa. Other diagnostic features are from the wrist bones or carpals, such as enlargement of three carpal bones (the hamate, the triquetrum, and the scaphoid), enabling a more forceful hand grip. These marsupial-like characteristics are supplemented by features of ankle bones or tarsals. Metatherians can be recognized by features of the astragalar bone (or talus) and the heel bone (or calcaneus). The derived marsupial-like features in the ankle are correlated to a greater mobility between the tarsal bones and tibia and fibular at the upper ankle joint, and among the tarsals at the trans-tarsal joint. These features contributed to a greater range of the medial and lateral movement of the hindfoot. By contrast, the Early Cretaceous eutherians can be recognized by the presence of five premolars and three molars (eight positions of postcanines) and the presence of certain features related to the internal carotid artery and its branches in the petrosal bone of the ear region. The ankle bones of the Cretaceous eutherians (with a narrow head and pulley-like upper joint in the astragalar bone) are different from those of all non-eutherian mammals.

The newly discovered *Eomaia* and *Sinodelphys* help to establish the ancestral anatomical conditions from which modern marsupials and placentals could have evolved, and provide evidence for the sequence of evolutionary acquisition of the marsupial-like features in the metatherian lineage and the acquisition of placental-like characters in the eutherian lineage. Current evidence suggests that the foremost phylogenetic distinctions between metatherians (including marsupials) and eutherians (including placentals) reside in wrist and ankle anatomy, followed by the diagnostic characteristics of marsupial and placental dentitions, such as the reduced dental replacement related to specialized marsupial life history pattern, plus many characteristics in placental and marsupial molar crown morphology. Cretaceous metatherians and eutherians also show remarkably diverse locomotory adaptations in the features of hands and hindfeet. The phalangeal proportion of the finger bones and the shape of the claws suggest that both *Eomaia* and *Sinodelphys* were capable of branch walking and climbing on uneven substrates (**Figure 3**). The skeletons of many Late Cretaceous eutherians clearly show that these were terrestrial mammals. Locomotory skeletal structures of the Cretaceous eutherians and metatherians indicate that their locomotory adaptations were diverse, and this may have facilitated diversification of early eutherians and metatherians into different ecological niches.

See Also

Fossil Vertebrates: Palaeozoic Non-Amniote Tetrapods; Placental Mammals. **Mesozoic:** End Cretaceous Extinctions.

Further Reading

Allin EF and Hopson JA (1992) Evolution of the auditory system in Synapsida ("mammal like reptiles" and primi tive mammals) as seen in the fossil record. In: Webster DB, Fay RR, and Popper AN (eds.) *The Evolutionary Biology of Hearing*, pp. 587 614. New York: Springer Verlag.

Cifelli RL (2001) Early mammalian radiations. *Journal of Paleontology* 75: 1214 1226.

Hopson JA (1994) Synapsid evolution and the radiation of non eutherian mammals. In: Spencer RS (ed.) *Major Features of Vertebrate Evolution*, pp. 190 219. Knox ville, TN: The Paleontological Society.

Kemp TS (1982) *Mammal like Reptiles and the Origin of Mammals*. London: Academic Press.

Kielan Jaworowska Z, Cifelli RL, and Luo Z X (2004) *Mammals from the Age of Dinosaurs: Origins, Evolution and Structure*. New York: Columbia University Press.

Luo ZX, Cifelli RL, and Kielan Jaworowska Z (2001) Dual origin of tribosphenic mammals. *Nature* 409: 53 57.

Luo ZX, Crompton AW, and Sun AL (2001) A new mammal from the Early Jurassic and evolution of mammalian characteristics. *Science* 292: 1535 1540.

Luo ZX, Kielan Jaworowska Z, and Cifelli RL (2002) In quest for a phylogeny of Mesozoic mammals. *Acta Palaeontologica Polonica* 47: 1 78.

Luo ZX, Ji Q, Wible JR, and Yuan CX (2003) An Early Cretaceous tribosphenic mammal and metatherian evolution. *Science* 302: 1934 1940.

Novacek MJ (1992) Mammalian phylogeny: shaking the tree. *Nature* 356: 121 125.

McKenna MC and Bell SK (1997) *Classification of Mammals Above the Species Level*. New York: Columbia University Press.

Rowe TB (1993) Phylogenetic systematics and the early history of mammals. In: Szalay FS, Novacek MJ, and McKenna MC (eds.) *Mammal Phylogeny: Mesozoic Differ entiation, Multituberculates, Monotremes, Early Therians, and Marsupials*, pp. 129 145. New York: Springer Verlag.

Szalay FS (1994) *Evolutionary History of the Marsupials and an Analysis of Osteological Characters*. Cambridge: Cambridge University Press.

Wible JR, Novacek MJ, and Rougier GW (2004) New data on the skull and dentition in the Mongolia Late Cret aceous eutherian mammal *Zalambdalestes*. *Bulletin of the American Museum of Natural History* 281: 1 144.

Placental Mammals

D R Prothero, Occidental College, Los Angeles, CA, USA

Introduction

The Eutheria, or placental mammals, are the dominant group of vertebrates on the Earth today, and they have ruled the planet ever since the extinction of the dinosaurs at approximately 65 Ma. At the latest count, there are over 4400 living species of mammal, classified into over 1000 genera, 140 families, and 18 orders. However, the number of extinct mammals is at least five times as great. Most living mammals are terrestrial, including large beasts, such as elephants, rhinoceroses, hippopotamuses, and giraffes, as well as a great diversity of smaller land animals. The largest known land mammal is the extinct 20 tonne hornless rhinoceros *Paraceratherium*. Many groups of mammals evolved from land-dwelling ancestors to life in the water. These include manatees and dugongs (which are distantly related to elephants), otters (which are related to weasels), seals, sea lions, and walruses (which are distantly related to bears), and whales (which are distantly related to even-toed hoofed mammals), as well as numerous extinct groups. The living blue whale (at 30 m in length and 150 tonnes) is by far the largest animal that has ever lived, surpassing even the largest dinosaurs. Mammals have also taken to the air, with almost 1000 living species of bat, as well as numerous gliding forms, such as the 'flying squirrels' and 'flying lemurs' or colugos. Mammals are even more successful when their body sizes are small, with hundreds of small species of rodent, rabbit, and insectivore. The smallest living mammal, the 1.5 g Kitti's hog-nosed bat, is at the lower limit of possible body size for mammals, since the physiology and anatomy of mammals prevent them from thriving in the tiny-body-size niche inhabited by insects and other arthropods.

Anatomy and Physiology

Mammals are distinguished from all other animals by the possession of a number of unique characteristics. These include a body covered in hair or fur (secondarily reduced in some mammals, particularly aquatic forms); mammary glands in the females for nursing their young; a jaw composed of a single bone, the dentary; and three middle-ear bones, the incus, malleus, and stapes. All mammals maintain a constant body temperature through the production of metabolic heat. They have a four-chambered heart (two ventricles and two auricles), which keeps the circulation of the lungs separate from that of the rest of the body, resulting in more efficient oxygen transport to the body tissues. They have many other adaptations for their active lifestyle, including specialized teeth (incisors, canines, molars, and premolars) for biting, tearing, and grinding up their food for more efficient digestion. These teeth are replaced only once in the lifetime of the animal (rather than continuously as in other toothed vertebrates). Mammals have a unique set of jaw muscles, which allow the jaw to move in many directions for chewing and for stronger bite force. Their secondary palate encloses the internal nasal passage, and allows them to breathe while they have food in their mouth. Ribs (found only in the thoracic region) are firmly attached to the breastbone, so that expansion of the lung cavity is accomplished by movement of a muscular wall in the abdominal cavity called the diaphragm.

All living mammals have relatively large brains for their body size. Most mammals have excellent senses, and some have extraordinary senses of sight, smell, and hearing. To accommodate their larger brains and more sophisticated development, most mammals are born alive (rather than from eggs) and may require considerable parental care before they are ready to fend for themselves. Juvenile mammals have separate bony caps (epiphyses) on the long bones, which are separated from the shaft of the bone by a layer of cartilage. This allows the long bones to grow rapidly while still having a strong, bony articulation at the end. When a mammal reaches maturity, these epiphyses fuse to the shaft, and the mammal stops growing (in contrast to other vertebrates, which grow continuously throughout their lives).

Reproduction and Classification

Living mammals are divided into three major groups: the monotremes (platypus and echidna), which still lay eggs, retain a number of reptilian bones in their skeletons, and have other primitive features in their anatomy and physiology; the marsupials (opossums, kangaroos, koalas, wombats, and their relatives), which give birth to an immature embryo that must crawl into its mother's pouch (marsupium), where it finishes its development; and the placentals (the rest of the living mammals), which

carry their young through a long gestation until they give birth to relatively well-developed progeny. This last group derives its name from the placenta, the membrane that surrounds the developing embryos within the uterus of the female. The placenta is formed by the fusion of the two other embryonic membranes, the chorion, which surrounds the embryo, and the allantois, which holds the embryonic wastes in egg-laying vertebrates. The placental membrane attaches the embryo to the wall of the uterus, and exchanges nutrients, gases, and waste products with the circulatory system of the mother, so that no hard-shelled egg (or its associated features, such as the allantoic waste sac or the large yolk sac) is needed. The placenta allows the embryos to undergo more development within the uterus, and to be born more mature than in other groups of mammals. By contrast, in marsupials the lack of a placenta protecting the embryo means that, when the immune system develops, the embryo must be born prematurely and move to the pouch, otherwise the mother's immune system would reject the embryo as a foreign object within the uterus.

In addition to these three living groups, there were many other major groups, such as the squirrel-like multituberculates, which are now extinct. The most recent classification of the mammals can be summarized as follows:

Class Mammalia
- Subclass Prototheria (monotremes)
- Subclass Theriiformes
  - Infraclass Holotheria
    - Cohort Marsupialia (marsupials or pouched mammals)
    - Cohort Placentalia (placentals)
      - Magnorder Xenarthra (sloths, anteaters, armadillos)
      - Magnorder Epitheria
        - Grandorder Anagalida (= Glires) (rodents, rabbits, elephant shrews)
        - Grandorder Ferae (carnivores, pangolins, and many extinct groups)
        - Grandorder Lipotyphla (hedgehogs, shrews, moles, tenrecs, and kin)
        - Grandorder Archonta
          - Order Chiroptera (bats)
          - Order Primates (lemurs, monkeys, apes, humans)
          - Order Scandentia (tree shrews)
        - Grandorder Ungulata (hoofed mammals)
          - Order Tubulidentata (aardvarks)
          - Order Artiodactyla (even-toed hoofed mammals: pigs, hippopotamuses, camels, deer, antelopes, cattle, giraffes, pronghorns, and their relatives)
          - Order Cete (whales and their extinct land relatives)
          - Order Perissodactyla (odd-toed hoofed mammals: horses, rhinoceroses, tapirs, and extinct relatives)
          - Order Hyracoidea (hyraxes)
          - Order Tethytheria (elephants, manatees, and extinct relatives)

This classification does not list all the extinct groups, which include at least a dozen more ordinal-level taxa. It is a considerable improvement on previous mammalian classifications, which listed over 30 different orders with no indication of how they were related to one another. Although there are more ranks in this classification than is traditional, this nesting of groups within groups accurately reflects their evolutionary branching sequence (**Figure 1**).

In the past few years, molecular studies have suggested an alternative arrangement of placental relationships, which conflicts strongly with that indicated by the morphology and the fossil record. This classification groups the orders of placental mammals as follows:

Subclass Placentalia
- Superorder Xenarthra (edentates)
- Superorder Afrotheria (African insectivores, such as the golden moles and tenrecs, plus elephant shrews, aardvarks, and the tethytheres, including elephants, sirenians, hyraxes)
- Superorder Euarchontoglires
  - Infraorder Archonta (primates, colugos, tree shrews)
  - Infraorder Glires (rodents, rabbits)
- Superorder Laurasiatheria
  - Infraorder Cetartiodactyla (whales plus artiodactyls)
  - Infraorder Ferungulata (perissodactyls, carnivores plus pangolins)
  - Infraorder Chiroptera (bats)
  - Infraorder Eulipotyphla (shrews, moles, hedgehogs)

Advocates of this molecular classification argue that it makes biogeographical sense for the African mammal groups (the 'Afrotheria') to be closely related, since Africa was an island continent isolated from the rest of the world during the Early Cenozoic. Likewise, the only placentals from the island continent of South America, the edentates, are in a separate group. The rest of the placental mammals, which have lived on the northern ('Laurasian') continents through most of their history, also cluster together. Critics of this classification point to a huge number

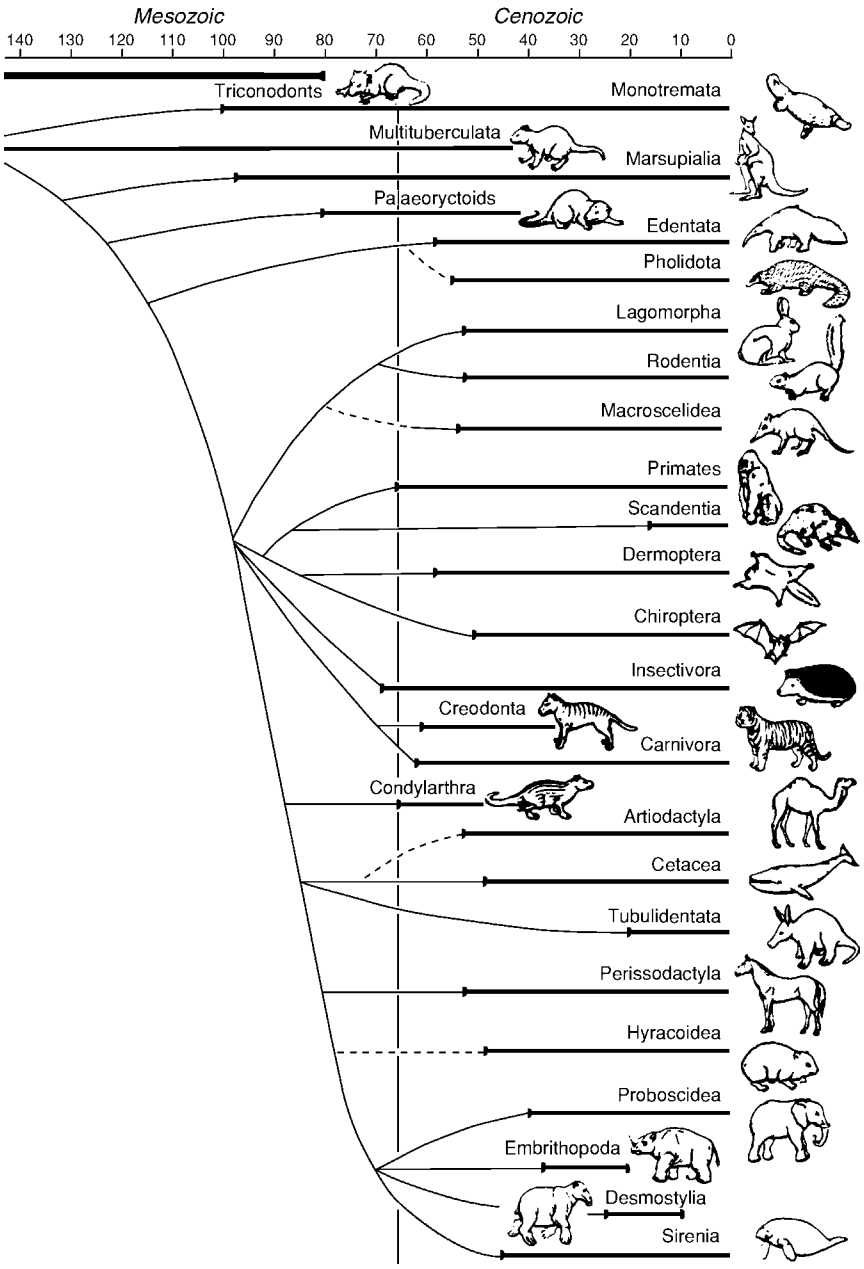


Figure 1 Evolutionary relationships and history of the major groups of mammals (modified from Novacek MJ (1994) The radiation of placental mammals. In: Prothero DR and Schoch RM (eds.) *Major Features of Vertebrate Evolution*, pp. 220 238. Paleontological Society Short Courses in Paleontology, no. 7. Paleontological Society, Lawrence, Allen Press).

of conflicts with well-established morphological evidence. For example, it breaks up the hoofed mammals by placing elephants and hyraxes in the Afrotheria, whereas the morphological evidence and the fossil record strongly suggest that they are related to the rest of the hoofed mammals (perissodactyls and artiodactyls plus whales). Likewise, good morphological evidence links at least some of the African insectivores with the true lipotyphlan insectivores, so most palaeomammalogists view the new molecular phylogenies with scepticism. They have good reason to do so, for molecular biologists have been claiming that they have 'the answer' to placental phylogeny in each new molecular system they examine, but, over the years, they have produced wildly varying results, so their classifications are not stable or reproducible enough yet to justify reordering the entire taxonomy of placental mammals.

Evolution

Mammals evolved from the Synapsida, an early branch of the terrestrial amniotes that has been erroneously called the 'mammal-like reptiles'. This name is inappropriate, because synapsids were never reptiles. Synapsids and reptiles originated independently about 320 Ma, and have evolved separately ever since. Early synapsids, such as the finback *Dimetrodon*, show relatively few mammalian characteristics, but, as their evolution progressed through the Late Palaeozoic, synapsids became progressively more and more mammal-like.

The first undoubted mammals appeared in the Late Triassic (about 210 Ma), and were tiny insectivorous forms much like living shrews. A number of different groups evolved over the next 145 Ma of the Jurassic and Cretaceous. Most remained tiny shrew-like animals, hiding from the dinosaurs in the underbrush and coming out mostly at night. The first two-thirds of mammalian history had passed before the dinosaurs became extinct 65 Ma, allowing mammals to emerge from their shadow.

Opossum-like marsupials are known from the Early Cretaceous (about 110 Ma), and they were more common than placentals just before the end of the age of dinosaurs. The first probable placental is the beautifully preserved skeleton (complete with hair impressions) from the Early Cretaceous of China known as *Eomaia scansoria*, which is about 125 Ma old. Although it is very primitive in most features, it had teeth for an insectivorous diet and limb proportions that suggest it was a tree dweller. By the Late Cretaceous (75 Ma), very primitive hoofed mammals (zhelestids) are known from Uzbekistan, and primitive members of the insectivore, carnivore, and primate lineages are known from the latest Cretaceous in both North America and Asia (65–70 Ma).

Although these early placentals are extremely primitive and hard to distinguish from early marsupials based on the limited evidence of their teeth and jaws, they can still be recognized by the trained palaeontologist. Early placentals had three upper and lower molars on each side of the jaw, and no more than five premolars (which reduces to four in later placentals). By contrast, early marsupials had four molars and usually only three premolars. In placentals, the last premolar has a tendency to be shaped like a molar (even though, as a premolar, it replaces a deciduous tooth, while a molar is never replaced). Primitive marsupial upper molars have a large expanded region on the outer edge of the tooth crown, whereas the upper molars of early placental mammals have little area outside the main cusps and instead are wider on the inside edge of the tooth. In the lower jaw of marsupials, the projection below the hinge (the angular process) flexes towards the midline, but in placentals it points straight back.

Between 65 and 55 Ma, a rapid adaptive radiation (**Figure 1**) yielded all the living orders of placental mammals and many extinct forms as well. The Xenarthra, or edentates, was the first group of placentals to branch off. Although the name 'edentate' means toothless, this is true only of anteaters; sloths and armadillos have simple peg-like teeth that lack enamel. The archaic nature of edentates among the placentals is shown by a variety of characters, including a uterus simplex (divided by a septum and lacking a cervix), a slower less well-regulated metabolism, retention of several reptilian bones lost in all other placentals, and a primitive rod-like stapes in the middle ear. Edentates evolved in isolation in South America throughout most of their history, developing into a variety of sloths (both tree sloths and huge ground sloths weighing up to 3 tonnes), anteaters, and armadillos (including the giant glyptodonts, which were 2 m long and had 400 kg of body armour and a spiked club on the tip of the tail). The long period of isolation in South America ended about 3 Ma ago, and ground sloths and glyptodonts migrated to Central America and parts of North America, disappearing at the end of the last Ice Age.

The remaining (non-edentate) placentals, or epitheres, diversified primarily in Eurasia and North America, and spread throughout the world in the Early Cenozoic. The true lipotyphlan insectivores (represented by shrews, moles, and hedgehogs) continued to diversify throughout the Cenozoic all over the northern continents. Most remained small in body size and ate insects, worms, and other small animals, although the extinct giant hedgehog *Dinogalerix* was the size of a large dog and killed sizable prey. The smaller lipotyphlan insectivores, such as the shrews, are among the smallest living placentals. Their body size is so tiny that they are constantly losing body heat because of their large surface area relative to their tiny mass. Consequently, they must eat almost continuously, or they will starve to death in a matter of days. This makes them extremely active and voracious predators, attacking not only worms and arthropods, but also animals much larger than themselves. They are fearless and will fight much larger animals when cornered.

The archontan radiation began with an enormous expansion of primitive lemur-like primates in the Early Cenozoic when the world had dense jungle vegetation all the way to the poles. Primates declined

in the Oligocene, when their forest habitats disappeared, and became restricted to Africa (Old World monkeys and apes) and South America (New World monkeys). From the Old World monkeys evolved the great apes. About 15 Ma, apes were more diverse than monkeys in the Old World. Today only four groups of living great apes are known: the gibbons, the orang-utans, the gorillas, and the chimpanzees. They diverged from our own family, the Hominidae, about 7 Ma. This timing was originally established by molecular evidence, but recently a hominid fossil of that age, *Sahelanthropus tchadensis*, has been described from rocks 6–7 Ma in Chad in western Africa. In the past 5 million years, there have been dozens of species of hominid, divided into many genera: *Orrorin, Ardipithecus, Paranthropus, Australopithecus*, and our own genus, *Homo*, which appeared about 2.5 Ma. Fossils that are recognizably members of our own species, *Homo sapiens*, are known from South Africa in rocks about 100 000 years old.

The earliest bats (known from about 50 Ma) already had fully developed wings. They are the second most diverse group of mammals after the rodents, with almost 1000 living species and many more fossil species. They are divided into two main groups, the insectivorous Microchiroptera (which use echolocation to find their prey on the wing) and the fruit bats, or Megachiroptera (which fly during the day, eating fruit in the trees of the tropics, and do not echolocate).

The radiation of the Glires began in the Paleocene of Asia, where numerous primitive relatives of rodents and rabbits are found. In the Eocene both groups migrated to Europe and North America, where they soon took over the niche of small-bodied fruit, seed, and nut eaters that had been occupied by multituberculates and primitive primates. Rodents and rabbits are both characterized by chisel-like ever-growing front incisors that are used in gnawing. These incisors have a band of enamel only on the front edge; the rest of the tooth is made of softer dentin. The rodent keeps the teeth sharp by continuous gnawing, so that the more resistant enamel edge is worn and kept sharper than the dentin behind it. If the incisors are not worn down, but are misaligned by malocclusion, they will grow around in a curve until they pierce the skull.

Although rodents and rabbits are closely related, they form two different orders within the Glires and can be easily distinguished. Rodents have only a single pair of incisors, while rabbits have two. The enormous diversification of the Rodentia since the Eocene has given rise to over 1700 species (about 40% of the Mammalia), with forms ranging in size from the pig-sized capybara down to the many tiny mice and voles.

Predatory mammals (the Ferae) include the extinct creodonts (an archaic group that were the dominant predators and scavengers of the Early Cenozoic) and the living order Carnivora (cats, hyenas, mongooses, civets, dogs, bears, weasels and their kin, seals and sea lions, raccoons, and many extinct groups). All Carnivora are distinguished by their distinctive shearing teeth, the carnassials, developed between the last upper premolar and the first lower molar. True Carnivora began as weasel-like forms in the Early Eocene, but by the Oligocene they had taken over most of the predatory niches from the creodonts. By the Miocene, the ancestors of seals and sea lions had evolved from bear-like ancestors. Carnivorous mammals show remarkable convergence on a limited number of body forms. For example, sabertoothed forms evolved four times, once in the creodonts, once in the true cats, once in the extinct cat-like nimravids, which are related to dogs, and one extinct sabertoothed marsupial that has been found in South America. In North America, the borophagine dogs converged on hyenas, with similar bone-crushing teeth.

The hoofed mammals, or ungulates, are first known from about 85 Ma in central Asia. In the latest Cretaceous and the Paleocene, archaic hoofed mammals ('condylarths') were among the most common forms in North America and Asia. From these roots, numerous orders evolved. The first to branch off were the even-toed Artiodactyla, which have two or four toes on each foot and a distinctive ankle structure. First appearing in Pakistan in the earliest Eocene, artiodactyls quickly diversified into a number of different groups. Today there are over 190 living species of artiodactyl, and at least twice that number of fossil species are known. Artiodactyls include the suoids (pigs, peccaries, hippopotamuses), tylopods (camels and their extinct relatives), and ruminants (deer, giraffes, pronghorns, cattle, sheep, goats, antelopes). With their four-chambered stomachs for more efficient digestion, the ruminants became the dominant group of large herbivorous mammals as global climates became drier in the later Cenozoic and grasslands expanded.

The ancestors of whales were large hoofed predators known as mesonychids. Recently, transitional forms between mesonychids and primitive whales have been found in the Eocene of Pakistan. However, some of these fossils also show the 'double pulley' ankle bones found in all artiodactyls, which lends weight to the suggestion by molecular biologists that

whales are descended directly from artiodactyls, rather than from mesonychids. By the Oligocene, whales had diversified into the predatory toothed whales (dolphins, orcas, sperm whales) and the filter-feeding baleen whales (blue, right, humpback, grey and many other whales).

The Perissodactyla, or odd-toed ungulates, have one or three toes on each foot. Today they include horses, rhinoceroses, and tapirs, but they were much more diverse in the past, with huge two-horned brontotheres and bizarre clawed chalicotheres. Closely related to perissodactyls are the hyraxes or conies and the tethytheres (elephants, sea cows, and their relatives). Tethytheres are so named because they originated from Late Paleocene ancestors that once lived along the Tethys seaway (which stretched from Gibraltar to Indonesia). Although sea cows had spread around the world by the Eocene, their earliest evolution was a mystery. Then a nearly complete specimen of *Pezosiren* was described from the Eocene of Jamaica, which showed what their earliest evolutionary transitions looked like. This fossil has a fully sirenian skull and teeth, and thick ribs for ballast, but retains fully functional legs and feet, rather than flippers.

The remaining tethytheres include the elephants, the huge two-horned extinct arsinoitheres, and the hippopotamus-like desmostylians from the Miocene of the Pacific Rim. The earliest relative of the elephant is *Paschatherium*, known from teeth from the Paleocene of Morocco. By the Eocene, the elephant family was represented by *Moeritherium*, which looked like a small pig or hippopotamus, although it had the beginning of a trunk or proboscis. The evolution of most of these groups was restricted to Africa until the Middle Miocene, when both groups (proboscideans and arsinoitheres) spread to Eurasia, and mastodonts even reached North America.

See Also

Evolution. **Fossil Vertebrates:** Mesozoic Mammals; Hominids. **Tertiary To Present:** Paleocene; Eocene; Miocene.

Further Reading

Benton MJ (2000) *Vertebrate Palaeontology,* 2nd edn. Oxford: Blackwell.

Corbet GB and Hill JE (1991) *A World List of Mammalian Species,* 3rd edn. Oxford: Oxford University Press.

Eisenberg JF (1981) *The Mammalian Radiations*. Chicago: University of Chicago Press.

Ji Q, Luo Z X, Yuan C X, *et al.* (2002) The earliest known eutherian mammal. *Nature* 416: 816 822.

Macdonald D. (1984) *The Encyclopedia of Mammals*. New York: Facts on File Publications.

McKenna MC and Bell SK (1997) *Classification of Mammals Above the Species Level*. New York: Columbia University Press.

Murphy WJ, Eizirik E, O'Brien S, *et al.* (2001) Resolution of early placental mammal radiation using Bayesian phylogenetics. *Science* 294: 2348 2351.

Novacek MJ (1992) Mammalian phylogeny: shaking the tree. *Nature* 356: 121 125.

Novacek MJ (1994) The radiation of placental mammals. In: Prothero DR and Schoch RM (eds.) *Major Features of Vertebrate Evolution,* pp. 220 238. Paleontological Society Short Courses in Paleontology, no. 7. Paleonto logical Society. Lawrence, Allen Press.

Nowak RM (1991) *Walker's Mammals of the World*. Balti more: Johns Hopkins University Press.

Pough FH, Janis CM, and Heiser JB (1999) *Vertebrate Life*. Upper Saddle River: Prentice Hall.

Prothero DR (1994) Mammalian evolution. In: Prothero DR and Schoch RM (eds.) *Major Features of Vertebrate Evolution,* pp. 238 270. Paleontological Society Short Courses in Paleontology, no. 7. Paleonto logical Society. Lawrence, Allen Press.

Prothero DR and Schoch RM (2002) *Horns, Hooves, and Flippers: The Evolution of Hoofed Mammals*. Baltimore: Johns Hopkins University Press.

Savage RJG and Long MR (1986) *Mammal Evolution: An Illustrated Guide*. New York: Facts on File Publications.

Szalay FS, Novacek MJ, and McKenna MC (eds.) (1993) *Mammal Phylogeny*. New York: Springer Verlag.

Vaughn TA, Ryan JM, and Czaplewski NJ (2000) *Mammalogy,* 4th edn. Fort Worth: Saunders College Publishing.

Wilson DE and Reeder DM (1993) *Mammal Species of the World*. Washington, DC: Smithsonian Institution Press.

Young JZ (1981) *The Life of Mammals,* 3rd edn. Oxford: Clarendon Press.

Hominids

L R M Cocks, The Natural History Museum, London, UK

Introduction

Humans and their history and activities are dealt with in a large number of enormous reference tomes, and thus the editors of this Encyclopedia have decided to refer to them only very briefly in this work. The studies of humans as organisms can be grouped into two, anthropology and medicine. The latter, dealing only with our ailments, receives no further mention here, but anthropology is divided into Physical Anthropology, which considers us as animals, and Social Anthropology, which examines how we interreact both with other humans and with our environments (a discipline which would be termed Ecology or Palaeoecology (*see* **Palaeoecology**) for other organisms elsewhere in this Encyclopedia).

Humans are placental mammals and are within the Primates, which also include such creatures as lemurs and lorises. Higher primates include monkeys and apes, and in particular gorillas, chimpanzees, and hominids (the latter systematically classified as the Family Hominidae). From genetic data, gorillas are thought to have diverged from the rest in Early Miocene times (at ~10 Ma), and hominids from chimpanzees at about 6.5 Ma. In the nineteenth century, much was made of the 'missing link' which apparently separated humans from their ancestors; however, that missing link has now vanished due to the large numbers of hominid specimens that have been found, particularly in East Africa since the Second World War. Because hominids are land animals, the preservation and fossilization of specimens was and is very chancy and they are best preserved in cave deposits and lake sediments. Although objective bone measurements can be and are made and used to biometrically distinguish between populations, the material is often sparse in numbers of specimens, and thus the resulting systematic classification somewhat subjective. The key features which physical anthropologists have found important are the braincase size and shape (including the development of brow ridges over the eye sockets), the teeth, the relative proportions of arms and legs, and the differing shapes of the ankle, knee, and elbow joints. Various DNA studies have been made, which have been differently interpreted in detail, but which in general support the story of hominid evolution and history outlined below.

Our own genus is termed *Homo*, and our species *sapiens*, and this brief article is divided into three: hominids other than *Homo*; *Homo* species other than *sapiens*; and *Homo sapiens*.

Hominids other than *Homo*

All hominids apart from *Homo* are known only from East and South Africa. A key feature of hominids is the development of bipedalism, which of course leaves the hands free for other activities, such as the gathering of food or the use of tools and other implements. Three genera with characters intermediate between chimpanzees and hominids, but known only from rather fragmentary fossils, are *Sahelanthropus*, found in Chad and dated to nearly seven million years ago, *Ardipithecus*, which lived in Ethiopia between about 6.0 and 4.4 million years ago, and *Orrorin* from Kenya, which lived at about 6 Ma. However, the earliest well-known hominid is their probable descendant *Australopithecus*, which lived between 4.2 and 2.4 million years ago (Figure 1). Possibly the best-known relatively complete specimen of *Australopithecus* is the one named 'Lucy', which was found at Hadar in Ethiopia and dated to 3.2 Ma. Modern chimpanzees have an average brain size of 390 c.c., Lucy's species a size of about 400 c.c. and modern humans about 1,300 c.c., and Lucy would have looked more like an ape than a human, walking upright, but with an ape-shaped body. Other hominid genera which have been named are *Paranthropus* (2.6 to 1.4 Ma) and *Kenyanthropus* (2.4 to 1.9 Ma).

Homo other than *sapiens*

The oldest representative of our genus is *Homo habilis* (and the similar and possibly synonymous *H. rudolfensis*), which lived in East and South Africa from 2.4 to 1.6 million years ago, with the best-known specimens, dated at about 2 Ma, being found by the Leakey family at Olduvai Gorge, Tanzania (Figure 2). *H. habilis* had a brain size of about 600 c.c. It made primitive stone tools from volcanic rock and chert, and was still somewhat chimpanzee-like, with long arms, short legs, and a thick waist. It was sometime around two million years ago that our ancestors made the important change from a largely vegetarian diet to one which also included meat. From then on they may be classified as hunter-gatherers, and that extension of activities into hunting might well have involved the combination into pack groups to make the killing of prey more effective.

Figure 1 Reconstruction by the late Maurice Wilson of *Australopithecus afarensis*, based on the 3.2 Ma specimens of 'Lucy' and other individuals from Ethiopia.

The invention of roasting food by fire appears to have been developed by about 1.6 Ma.

However, it was the successor of *habilis* which made much greater impact. This was *Homo erectus* (earlier versions of which are sometimes termed *H. ergaster*), which was relatively slim-waisted and, in some cases, just as tall as modern humans (Figure 3). It had a relatively larger cranium, but still had an extensive brow ridge. The brain case size had much increased; however, its average size at about 900 c.c. was only that of a one-year-old human child today. All these changes much improved mobility, and it was this species that expanded widely from the ancestral areas of East and South Africa for the first time. It has been postulated that these migrations would have been facilitated by climate changes related to the Plio-Pleistocene glaciation (*see* **Tertiary To Present:** Pleistocene and The Ice Age) near the Poles, which would have led to the expansion of more easily-traversed savannah grasslands and a reduction in the areas of tropical jungles. *H. erectus* existed from about 2.0 million to as recently as about 100 000 (100 Ka) years ago, and some of the best-known early material is from Koobi Fora and Lake Turkana, Kenya. By about 1.8 Ma the species (site names in brackets) had reached Georgia (Dmanisi) in Central Asia, soon afterwards Indonesia (Java) and subsequently China ('Peking' Man), and by 800 Ka it had spread over southern Europe to Italy (Ceprana) and Spain (Atapuerca). By 500 Ka years ago a descendant species, *H. heidelbergensis* (Figure 4), which lived from 800 to 300 Ka, had reached southern England (Boxgrove, Sussex).

A later species was *H. neanderthalensis*, commonly known as Neanderthal Man, which lived from around 400 000 years ago to as recently as about 30 000 years ago in Croatia (Figure 5). Neanderthals (named after their first-known nineteenth-century site in the Neander Valley in Germany), were relatively large-brained, big-faced but with a low skull and little chin, and most of their dietary protein was from animals. *Homo sapiens* lived in Europe together with Neanderthals between 40 and 30 thousand years ago: it is uncertain whether the latter became extinct through competition with humans or by a reduction of their habitats. It is also uncertain whether or not Neanderthals and modern humans could or did interbreed.

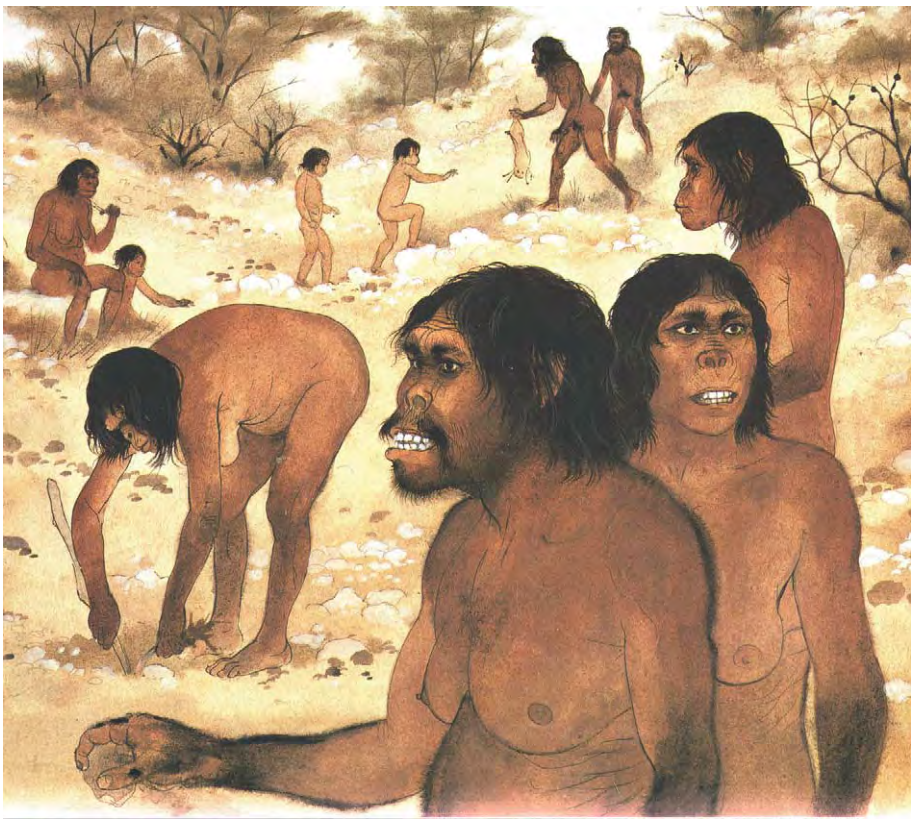


Figure 2 Reconstruction by the late Maurice Wilson of *Homo habilis*, based on 1.9 Ma individuals from Kenya and Tanzania.

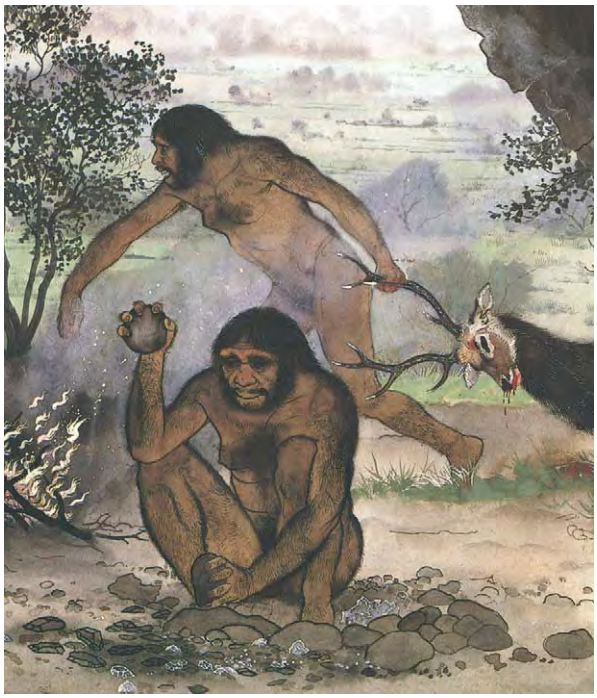


Figure 3 Reconstruction by the late Maurice Wilson of *Homo erectus* from 450 Ka individuals at Zhokoudian Cave, China.

Homo sapiens

Homo sapiens probably evolved from *H. heidelbergensis* at about 300 Ka, somewhere in eastern Africa. It is difficult to be sure whether the transition between *H. sapiens* and its ancestors was sudden or took place over a relatively long time period. They differ from their ancestors in having a much larger brain size and in the virtual absence of a brow ridge. They spread into the Middle East at about 100 Ka, and on into southern Europe relatively quickly after that. Migration to northern Asia appears to been have later (the earliest humans in China date from about 50 Ka). The oldest human remains in Australia also appear to date from about 50 Ka. Much later, humans colonised the Americas via the Siberian land bridge during a sea-level low stand before 13 Ka. Once in America they spread quickly, reaching Patagonia in southernmost South America at about 11 Ka. Differentiation between the various human racial types (African, Caucasian, etc.) occurred through separation of the breeding populations, but the process has not gone so far that we have ceased to be a single biological species.

A key change in the behaviour of humans was the transition, which was progressive over a long time

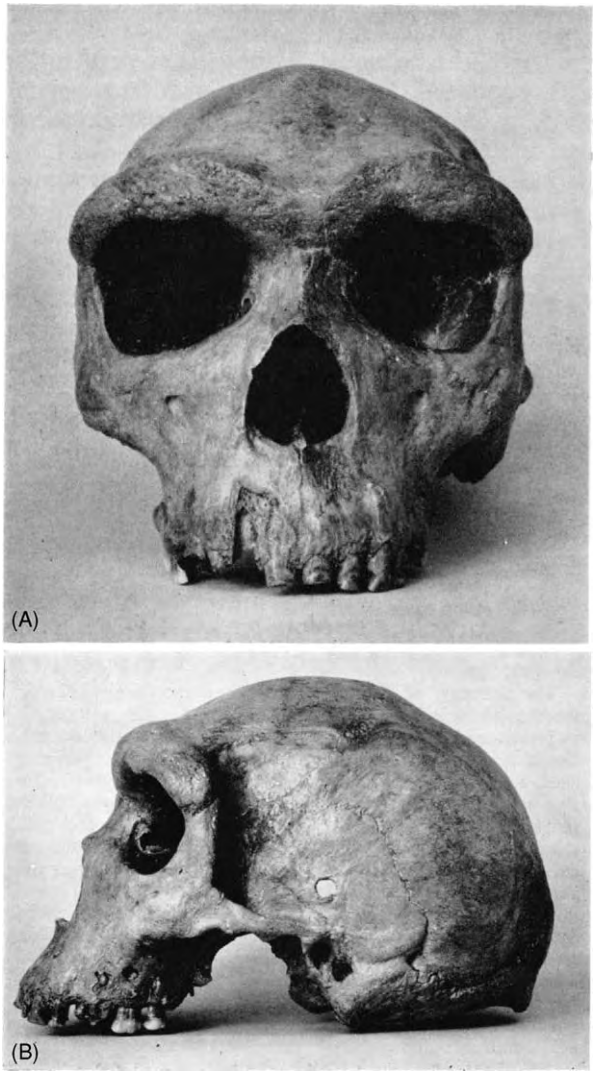


Figure 4 *Homo heidelbergensis*, a 300 Ka skull from Broken Hill, Zambia. Note the brow ridges.

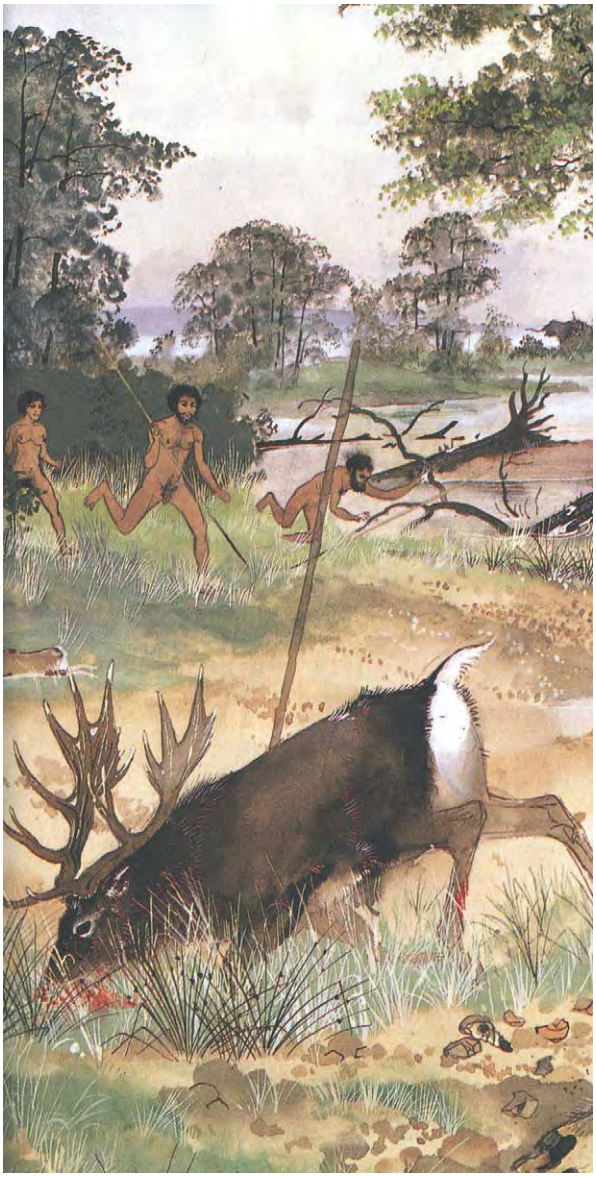


Figure 5 Reconstruction by the late Maurice Wilson of a hunting scene with *Homo neanderthalensis* at Swanscombe, Kent, England, at about 400 Ka.

period, from being hunter-gatherers to the planting of crops and the advent and development of agriculture. The latter implies a far more static existence, which led to more permanent dwellings and eventually into organized settlements, progressively villages, towns, and cities. The earliest organized human settlements yet known are those dating from about ten thousand years ago in Turkey and Syria. The more than minimal quantity of food produced by agriculture resulted in the introduction of both leisure time and also in the division of human activity into separate occupations; a process which led inexorably to the civilization and cultural divisions which we know today. It is not known when hominids developed speech and language, but the initiation of those facilities may have taken place as early as *H. habilis* times, with progressive development as the advantages of improved communications became self-evident.

Modern humans have much affected the geology on a global basis, not only in the extraction of many varied building stones, aggregates, minerals, hydrocarbons, and other resources, but also in the management of water, construction of buildings, and the varied activities which have led to the global warming which directly affects sedimentation and other geological processes both now and progressively more so in the future. Some anthropogenic details of these effects in historical times are outlined in the article on the Holocene of Europe (*see* **Europe: Holocene**).

See Also

Europe: Holocene. **Fossil Vertebrates:** Placental Mammals. **Palaeoecology**. **Tertiary To Present:** Pliocene; Pleistocene and The Ice Age.

Further Reading

Johanson D and Edgar B (1996) *From Lucy to language*. London: Weidenfeld and Nicholson.

Klein RG (1999) *The Human Career*. Chicago: University of Chicago Press.

Lewin R (2004) *Human Evolution: an Illustrated Guide*. Oxford: Blackwell Scientific Publications.

[no author] (2003) New look at Human evolution. *Scientific American* (special edition). June: 100.

Stringer CB (2002) Modern Human origins – progress and prospects. *Philosophical Transactions of the Royal Society, London B* 357: 563–579.

Stringer CB and Gamble C (1993) *In search of the Neanderthals*. London: Thames and Hudson.

ENCYCLOPEDIA OF
GEOLOGY

ENCYCLOPEDIA OF GEOLOGY

EDITED BY

RICHARD C. SELLEY
L. ROBIN M. COCKS
IAN R. PLIMER

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

GAIA

G J H McCall, Cirencester, Gloucester, UK

Introduction

The Gaia concept has evolved in the 30 years since it was first introduced by James Lovelock, an independent scientist and inventor. It was initially a rather vague model relating to the climate and diversity of the planet Earth, though living organisms were critical to it. The workings of the model were initially unspecified. The concept, however, was one of a 'superorganism' operating to 'regulate' the planet, especially its surface temperature, yet lacking the 'foresight' possessed by intelligent animals. Lovelock updated his work in 2000, publishing *The Ages of Gaia*. The discussion here is based on this later book, and it is quite unavoidable to echo much of what Lovelock has said, because he is the only and definitive source.

Gaia in the Twenty-First Century

Gaia is essentially about life, because life is seen to combine with inanimate processes on Earth, affecting and even regulating the physical state of the biosphere. Lovelock found it surprisingly difficult to find a good definition of 'life'. Of the definitions found in *Webster* – "the property of plants and animals (ending in death and distinguishing them from organic matter) which makes it possible for them to take in food, get energy from it, grow etc." – and in *Oxford* – "the property which differentiates a living animal or plant or a living portion of organic tissue, from dead or non-living matter; the assemblage of the functional activities by which this property is manifested" – neither is satisfactory, and the second is tautological in the extreme. To the first might be added, before "etc.", the words "and move, in the case of animals". Lovelock added to his definition of life that "living things use energy directly from the Sun and indirectly from food" (*see* **Origin of Life**).

There is no difficulty in accepting that advanced living animal organisms, such as humans, are made up of intricate communities of connecting cells, and, as Lynn Margolis has shown, that cells are derived from micro-organisms that once lived free (*see* **Precambrian:** Prokaryote Fossils; Eukaryote Fossils). Larger entities, such as ecosystems, are also accepted, and space exploration has contributed to this understanding by allowing the entire planet Earth to be viewed from space. The Gaia concept likewise involves envisaging the entire globe as an integrated system, with the atmosphere, the seas, the rivers, and the rocks interacting to modulate the planet's physical state and thus the environment in which life can exist, with the presence of life contributing significantly to the interactions. Gaia is thus not a synonym of 'biosphere' or 'biota': it is a much larger entity. When the Gaia model was originally proposed in the 1970s, it was considered that the atmosphere, oceans, climate, and crust of Earth were regulated to maintain a comfortable state for life to exist, by and for the biota. Temperature, oxidation state, acidity, and certain aspects of the rocks and waters were kept, at any time, constant, maintained in homeostasis, by the organisms at Earth's surface. This concept is now seen to have been incorrect, because both life, which is continually evolving, and the geological environment are in a state of constant change, and the conditions needed to maintain life change very rapidly, with the changing needs of the biota, so homeostasis cannot be maintained for more than very brief periods in Earth history. The Earth is thus seen as being like a helicopter, which is, unlike a fixed-wing plane, never in stable flight. The changing and evolving needs of the biota require that the brief periods of homeostasis are quickly overtaken by new requirements. The concept now is of a superorganism in which the active feedback processes operate automatically, so that solar energy maintains comfortable conditions for life.

Molecular Biology: The Laws of the Universe

Lovelock regarded the emergence of the science of molecular biology – embodied in the information-processing chemicals that underpin the genetic basis of most life on Earth – as having taken life science out of a routine classificatory and descriptive pursuit into a new and exciting study of how all the components in life are related. Equally important are physiology, the study of organisms seen holistically, and thermodynamics, a branch of physics dealing with time and energy, connecting living processes with universal laws. Two fundamental universal laws of physics are that (1) energy is conserved, however much it is dispersed, and (2) energy is always abating. Hot objects cool, but cool objects do not heat up spontaneously; water flows downhill, but not uphill. Once used, energy cannot be recovered. Natural processes always move towards an increase in disorder, which is measured by entropy; entropy expresses the tendency to burn out. Looking at the relationship between life and entropy, Lovelock

referred to Erwin Schrödinger's conclusion that life has the ability to move upstream against the flow of time, apparently paradoxically and contrary to the second universal law. In fact, what is operating is a tightly coupled system to favour survival; energy is taken in (e.g., oxygen from the atmosphere is breathed), converted (e.g., stored body fats and sugars are transformed), and then excreted (e.g., waste products such as carbon dioxide are released back into the atmosphere). If the entropy of excretion is larger than the entropy of the oxygen consumed, life continues, despite the second universal law.

The Superorganism Concept

There is difficulty in envisaging an eruptive planet with a molten core and other complex inorganic processes as a living superorganism. However, the inspiring 'whole-planet' image of Earth as seen from space and the contrast between the environment on Earth and the environments on the moon, Mars, Venus, and Mercury have focused research on considering how the significant planetary differences arose, and in particular on the question of how and why the atmospheres differ. The atmospheres of the Moon, Mars, Venus, and Mercury are a good starting point for comparisons to Earth, because the atmospheres are the least complex and most accessible of the zones of all these planets; indeed, the atmospheric compositions on other planets were known before space exploration commenced (*see* **Solar System:** Mars; Moon; Mercury; Venus).

The Earth has an atmosphere of N and O, with traces of carbon dioxide, methane, and nitrous oxide, not in equilibrium, whereas the atmospheres of Mars and Venus are dominated by carbon dioxide and are in equilibrium. If the atmospheres of Mars and Venus were heated, there would be no reaction with the surface materials, whereas heating Earth's atmosphere would produce reactions leading to a carbon dioxide-dominated atmosphere. Lovelock concluded that the improbable atmosphere of Earth "reveals the invisible hand of life". The atmosphere contains oxygen and methane, which should react to form water vapour and carbon dioxide: that this does not occur, and that constant atmospheric compositions of these gases are maintained, reveal, Lovelock believes, that there is regulation by life (*see* **Atmosphere Evolution**).

Scientists as early as Eduard Suess and Vladimir Vernadsky accepted that there was continuous interaction between soils, rocks, oceans, lakes, rivers, the atmosphere, and life. Much later, Stephen Jay Gould stated that "organisms are not billiard balls, struck in a deterministic fashion and rolling to optimal positions on life's table". Living things influence their own destiny in an interesting and complex, but comprehensible, way. Thus the sum total of the physical state of a planet, with life, is a combination of the inanimate processes and the effects of life itself. JZ Young said that the entity that is maintained intact, and of which we all form part, is the whole of life on the planet. This statement really provided the link between theory and consensus, on the one hand, and Gaia concept, on the other, expressing as it does the view that the entire spectrum of life on the planet has to be considered alongside the geological and inanimate physical processes, if we are to understand how the planet works. This, of course, has led to the present preoccupation in educational circles with 'Earth System Science' (*see* **Earth System Science**). Gaia goes further than Earth System Science, which is purely a holistic educational approach, in requiring a global system that has the capacity to regulate the temperature and composition of Earth's surface, hydrosphere, and atmosphere, keeping it comfortable for living organisms.

Criticism of the Gaia Concept

Criticism of the Gaia concept, once advanced, was by no means slight, and the Gaia model was not taken seriously by scientists, at all, until the early 1970s. Fred Doolittle came out with the belief that "molecular biology could never lead to altruism on a global scale" – altruism by living organisms being apparently inherent to the concept. Richard Dawkins in 1982 supported him: "the selfish interests of living cells could not be expressed at the distance of the planet". It was also remarked that Gaia lacked a firm theoretical basis. Heinrich D Holland considered that biota simply react to change in the state of Earth's near-surface environment and processes, geologically produced, and those that adapt better survive: the rest do not. Many scientists saw Gaia as a teleological concept, requiring foresight and planning by organisms, something that the model surely never represented. However, a major step was taken at the Chapman conference of the American Geophysical Union in 1988, when numerous papers on Gaia were presented: the question of the scientific testability of the Gaia hypothesis was raised.

Holland's statement was really an oversimplification, because the environmental constraints to which an organism adapts can in no way be entirely inorganic in origin – geological processes are a combination of the inorganic and the organic. Lovelock stated this when he objected that "life cannot have adapted solely to an inert world determined by the dead hand of chemistry and physics". The two main objections to Gaia were, first, the teleological one

(forecast or clairvoyance seemed to be needed for Gaia to be true) and, second, the fact that ecological regulation by life could only be partial. What must be operating is a combination of regulation by the living and the inorganic, even though in the long term, a wholly inorganic interference, such as an asteroidal impact, would be reacted to by the biota, much of it by dying off. However, some of the biota may survive, regenerating and expanding to again affect the physical environment of the life forms continuing to exist at or near the surface.

Other objections to the Gaia concept have been a creationist argument, based on Schrödinger's conclusions on 'life', that some organisms do not need the sun, and also a suggestion that Lovelock's image of Earth as a 'spaceship' ignores the fact that 40 000 tonnes year ^{1} of extraterrestrial dust enters Earth's atmosphere and is deposited on Earth's surface.

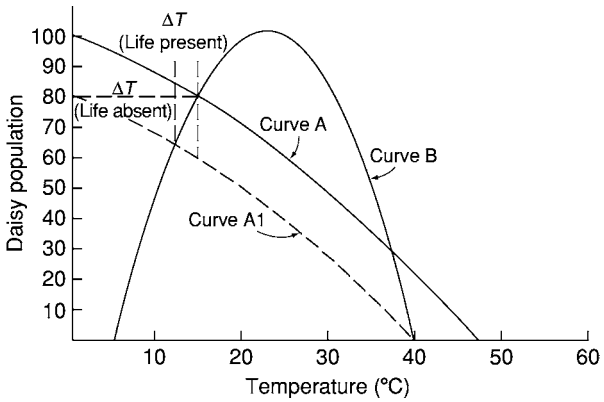


Figure 1 Regulation by white daisies. The helmet shaped curve (B) depicts the response of daisies to temperature. Curves A and A1 depict the responses of planetary temperature to the area covered by daisies, curve A1 being for a lower heat input by the planet's star. In the absence of daisies, the change in planetary temperature (ΔT) would be nearly 15°C, whereas in their presence ΔT is only about 3°C. Reproduced with permission from Lovelock J (2000) *The Ages of GAIA: A Biography of Our Living Earth*. Oxford: Oxford University Press.

Lovelock's Hypothesis

The 'Daisyworld' Model

Lovelock saw that he needed a simple model to illustrate his point. The effect of snow cover on the ground, a purely inorganic change, had already been modelled to show how snow cover changes the albedo and thus the cooling in the atmosphere. Lovelock developed a simple model, the 'Daisyworld' model, illustrating the way life could have a similar effect.

The Daisyworld parable of 1982 proposes a planet like Earth in size, mass, and orbital distance from a star that is like Earth's sun in mass and density; like Earth's Sun, the Daisyworld Sun increases its output as it ages (the nature of the H/He reaction means that our Sun was 30% cooler at its beginning, but will eventually heat up so as to consume Mercury, Venus, and possibly Earth). The Daisyworld planet has more land than sea, compared to Earth, is well watered, and plants can grow anywhere on the land surface if the climate is right. The sole plant is a daisy, which may be dark, neutral, or light in colour shade. A single parameter, variable temperature, controls whether the daisy can grow – 5°C is the growth threshold, 20°C is the optimum, and 40°C is the upper limit. The mean temperature of the planet is a simple balance between the heat received from the sun and the heat lost to the cold depths of space in the form of long-wave radiation. The complication on the real Earth of reflection upwards and blanketing downwards of heat by clouds is avoided, by having all the rainfall on the Daisyworld planet occur at night and having no clouds present in daytime. There is just enough carbon dioxide in the atmosphere for the plants to grow. The mean temperature will be determined by the average shade of colour of the daisy coverage. If the average is the dark colour shade, the albedo is low, more heat is absorbed, and the surface is warmed; if the average is the light shade, 70–80% of heat is reflected and the surface is colder. In the scale 0–1, dark daisies will have an albedo of, say, 0.2; light daises will have an albedo of 0.7 and the bare ground, 0.4. The initial weak Sun will increase until 5°C is attained and that will favour the dark daisies, because they have greater absorption of sunlight and will warm up the temperature just beyond 5°C. The light daisies will be disadvantaged and will fade and die. Next season, there will be more seeds of the dark strain and it will become warmer and so on, until most of the planet is colonized by the dark strain. However, when the upper limit temperature is approached, the light strain, because of their high albedo, will keep the temperature in check and so will take over from the dark strain. The growth of the star's heat flux will eventually be so great that nothing can keep the temperature below 40°C and all the daisies will die off. Flower power will no longer be enough. The planet will be barren and there will be no way back. Figure 1 reproduces Lovelock's most important Daisyworld diagram. The Daisyworld parable was devised to counter the 'teleological' criticism.

The Four Components of Gaia

Lovelock described Gaia as an automatic, non-purposeful, goal-seeking system, and considered it to be made up of four components:

1. Living organisms that grow freely, exploiting any environmental opportunities that open up.
2. Organisms that are subject to the rules of Darwinian natural selection: the species that leave the most progeny survive.
3. Organisms that affect their physical and chemical environment. Thus, animals change the atmosphere by breathing, taking in oxygen and breathing out carbon dioxide. Plants and algae do the reverse. In numerous ways, all forms of life incessantly modify the physical and chemical environment.
4. The existence of constraints or bounds that establish the limits of life. It can be too hot or too cold; there is a comfortable warmth in between, the preferred state. It can be too acid or alkaline; neutrality is preferred. For almost all chemicals, there is a range tolerated or needed by life. For many elements, such as iodine, selenium, or iron, too much is a poison and too little causes starvation. Pure, uncontaminated water will support little, but neither will the saturated brine of the Dead Sea.

Definitions and Criticisms

Lovelock advanced a set of definitions:

1. Life is a planetary phenomenon with a cosmological lifespan. On this scale, it is nearly immortal and has no need to reproduce.
2. There can be no partial occupation of the planet by living organisms. Such a condition would be as impermanent as half an animal. Ineluctable physical and chemical forces would soon render (the planet) uninhabitable. The presence of a sufficient number of animals on the planet is needed for the regulation of the environment.
3. Our interpretation of Darwin's great vision is altered. Gaia draws attention to the fallacy of the concept of adaptation. It is no longer sufficient to say that "organisms better adapted than others are likely to leave offspring". It is necessary to add that the growth of the organism affects its physical and chemical environment; the evolution of the species and, therefore, the evolution of the rocks are tightly coupled in a single, indivisible process.
4. Theoretical ecology is enlarged. By taking the species and the physical environment together as a single system, we can, for the first time, build ecological models that are mathematically stable, and yet include large numbers of competing species. In these models, increased diversity among the species leads to better regulation.

There are criticisms of Lovelock's definitions. Of definition 1, it is perhaps better to admit that there remain possibilities of abrupt extinction of life on Earth through physical disasters. For example, the entire planetary biota might have been extinguished in the latest Proterozoic; if glaciation, a complex phenomenon attributed to Milankovitch cycles in the Solar System, had covered the globe in the 'snowball Earth' episode, it might well have extinguished all life. Definition 2 begs the question, "what about the long early period in the Archaean, when only very primitive forms of unicellular life apparently existed?" The planet in fact got on very well through a vast period of time when it was only partially inhabited by life, and that being very primitive, compared with now. Regarding definition 3, surely to the word 'rocks' the words 'hydrosphere and biosphere' should be added. Finally, in definition 4, the statement that "increased diversity leads to better regulation" is questionable. The evidence for this does not seem to be to be given by Lovelock: the word 'regulation', which is fundamental to Gaia concept, seems to be a problem. The concept appears to be tantamount to accidental regulation: in the real world, as opposed to the Daisyworld, the reality would seem to be that the biota make a major contribution to the physical and chemical environment of the biosphere, and that, throughout Earth history since life appeared on the planet, the bounds of livability of the planetary biota as whole have never been exceeded, despite several mass extinctions – yet total extinction could happen at any time. Also, like the inorganic physical and chemical (geological) changes, changes induced by the biota can be benign or adverse. The biota can react to change to counter adverse changes; initially benign changes can lead to adverse changes in the long run (e.g., overpopulation). There is no law of the sum total planetary biota adjusting to regulate the environment, keeping within bounds, but so far (luck?), the bounds have not been overridden in the case of the overall biota, not at any time in geological history, since life first appeared.

Gaia and the Geological Record

Lovelock, in his second book, considered Gaia in terms of the geological record, discussing the Archaean, the Middle Ages, and the contemporary environment. Before concluding, it seems apposite to refer to a recent contribution to the discussion by Euan Nisbet.

Nisbet's Essay

Euan Nisbet, in a wide-ranging Fermor Lecture to the Geological Society in 2002, explored the question whether the presence of life or inorganic processes

had constituted the dominant factor in shaping the physical development and in controlling the physical conditions at or near Earth's surface since life appeared on the planet more than 3500 million years ago. He found that life had a dominant role in controlling the condition of the atmosphere through this immensely long period, and that, in turn, the surface temperature had a very significant control on the tectonic evolution, especially plate tectonics. Even so, Nisbet could not rule as to which processes, biological or inorganic, had been the dominant factor – the two kinds of processes had operated together to control the physical conditions. That life had exerted a significant effect on the ongoing physical evolution could not be argued against.

Nisbet based his conclusions on inorganic and organic models, and in them not all the assumptions made are necessarily correct. For instance, Nisbet tended to extend plate tectonics back through the Archaean, whereas there are strong arguments against this (e.g., see the publications by Hamilton in 1998, by Bleeker in 2002, and by McCall in 2003). Considering the real world, rather than models, Nisbet could not decide which had been in the driving seat, and he believed that the question was possibly not quantifiable. The atmosphere was, however, largely the product of the existence of life at or near Earth's surface.

Beyond Gaia

Gaia, at the least, is a brilliant concept, and Lovelock has provided a stimulating basis for looking at the entire globe and for integration of the effects of the progressive development of life and purely inanimate geological development through time. This concept also encompasses the maintenance throughout time of conditions suitable for life to continue to exist throughout 3500 million or more years. However, at this minimal assessment, it is no more than Earth System Science under another name, but emphasizing the special contribution of life to the changing state of the atmosphere, hydrosphere, and geological processes, which in sum determine the physical conditions under which life of some sort can continue to 'operate'. The Gaia concept, however, is claimed to be more than this, and the critical word is 'regulate'. Lovelock never intended Gaia to have a teleological significance, but even so, to regulate something would seem to imply an element of design. It may be that when an extreme condition occurs, such as an extreme greenhouse condition (Cretaceous) or a 'snowball Earth' approximation (end Proterozoic), the reactions of the living plant and animal populations that survive may be beneficial, but equally they could actually be adverse and eliminate even more of the global biota. One cannot escape the fact that the physical state of the near-surface zones of the planet has, for more than 3500 million years, been sufficiently benign for some life to continue to exist; throughout the development of life, from protozoans to destructive humans, the physical state of Earth's surface has never led to extinction of the sum total of life on the planet. Yet the history has surely been one of reactions by life to changes, including extremely critical situations, both inorganically and organically triggered; changes have never reached the point of driving life over the edge to extinction, but even so, this is not regulation, because there is no rule involved. The irretrievable end-point situation has just not happened. In fairness to Lovelock, in his 1991 book he did include a chapter entitled "The People Plague", in which he argued that humans, having no predators, had in effect become a plague on the planet, and could well take the life on the planet to that irretrievable end-point situation. Of course, one could argue that humans do have predators, even if minute ones – viruses and microbes that could blot them out forever and leave behind an Earth smiling with other life forms, until the Sun fulfils its destiny and incinerates the planet.

The influence of the existence of life on the physics and chemistry near the surface of the planet, and the feedback to life from that – the essence of Gaia – is well understood, but has obviously become more complex as the spectrum of life on Earth has become more complex, starting with the primitive prokaryotes in the Early Archaean. For example, a plant population limited to algae must have left bare rock surfaces in the Archaean, whereas once higher plants appeared, there must have been a much more complex reaction of the land surface to the Sun's radiation. In fact, these relationships, though complex and possibly unquantifiable at the present state of the planet, are less obscure than are other processes. For example, how does an animal obtain information that another animal exists and has a character or activity to which the first animal can adapt, to obtain benefit? How does the information get into the genetic process? The present author, in an unpublished work, *The Vendian (Ediacaran) in the Geological Record*, has called this mysterious process 'cognizance'. An example of this is the case of eoforaminifera found in Uruguay, in latest Proterozoic rocks. These organisms agglutinated fine mineral particles on their surface, and it has been glibly said that they did this to make themselves less palatable to predators. But how did they know that their mates were being eaten? It would seem that this aspect of genetics is what scientists should now be concentrating on, rather than 'regulation' by the sum total biota.

See Also

Atmosphere Evolution. **Biodiversity**. **Earth Structure and Origins**. **Earth System Science**. **Evolution**. **Origin of Life**. **Palaeoclimates**. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils. **Solar System:** Mars; Moon; Mercury; Venus. **Trace Fossils**.

Further Reading

Bleeker W (2002) Archaean tectonics a review. In: Fowler CMR, Ebinger CJ, and Hawkesworth CJ (eds.) *The Early Earth: Physical, Chemical and Biological Development. Special Publications 199*, pp. 151 181. London: Geological Society.

Gaucher C and Sprechmann P (1999) Upper Vendian skeletal fauna of the Arroyo del Soldado Group, Uruguay. *Beringeria* 23: 55 91.

Hamilton WB (1998) Archaean magmatism and deformation were not products of plate tectonics. *Precambrian Research* 91: 131 175.

Lovelock J (1978) *GAIA: A New Look at Life on Earth*. Oxford: Oxford University Press.

Lovelock J (1991) *GAIA: The Practical Science of Planetary Medicine*. London: Gaia Books.

Lovelock J (2000) *The Ages of GAIA: A Biography of Our Living Earth*. Oxford: Oxford University Press.

McCall GJH (2003) A critique of the analogy between Archaean and Phanerozoic tectonics based on regional mapping of the Mesozoic Cenozoic plate convergent zone on the Makran, Iran. *Precambrian Research* 127(1 3): 5 17.

Myers N (1984) *Gaia: An Atlas of Planet Management*. New York: Doubleday.

Nisbet E (2002) The influence of life on the face of the Earth. In: Fowler CMR, Ebinger CJ, and Hawkesworth CJ (eds.) *The Early Earth: Physical, Chemical and Biological Development. Special Publications 199*, pp. 275 307. London: Geological Society.

GEMSTONES

C Oldershaw, St. Albans, UK

Introduction

There are almost 4000 minerals known, of which only about 50 are commonly used as gemstones. Those that form crystals of sufficient size and quality to be cut and fashioned as gems are referred to as 'gem quality' or 'cuttable' pieces; other minerals or rocks with particularly attractive features (colour, texture, or pattern) may be called 'decorative' pieces. Crystals are usually faceted (cut and polished) to give a gemstone with a number of flat faces, while decorative stones are mainly tumbled or polished to produce pieces for personal adornment or *objets d'art*.

Gemstones are formed in each of the three main rock types: igneous, sedimentary, and metamorphic. Mining methods depend on the type of gemstone and its optical and physical qualities and on whether it is being mined from the rock in which it was formed or retrieved from secondary (placer) deposits produced by weathering and erosion.

A study of modern-day gemstone mining and retrieval covers every mining method, from the traditional searches in streams and rivers using a pan or sieve to the ultra-high technology and research models used in diamond mines deep underground. Ultimately, any source of gemstones will be mined and exploited only if is financially viable.

Quantifying Gemstone Mining

While mineral exploration is generally well documented and statistics are available, providing figures for production for most minerals on a country-by-country basis, this is generally not the case with gemstones (except perhaps for diamonds), which have always been difficult to quantify. Much gemstone mining is carried out in remote places, with secrecy and security to protect the interests (and sometimes the lives) of the owners (Table 1). In some areas, families or groups of villagers purchase a licence to mine; in other areas there is little regulation, and mining may be controlled by local leaders or warlords, with all the problems associated with regions of conflict and political unrest.

Over large regions it is difficult to assess the production figures for particular gemstones, although general trends in availability and value (which could be as much a product of fashion as of supply) give some indication of production figures. On a smaller scale, a specialist gemstone buyer (such as a ruby or sapphire buyer) will, by building relationships with miners, mine owners, and local traders, gain expertise and some knowledge of the volume and quality of gemstones being mined in that region. As new localities are identified and new gemstones reach the markets, previously known localities or gemstones may lose their appeal or a region may become unproductive – so the market can be quite changeable.

Table 1 The main localities of some of the best known gemstones (diamond, ruby, sapphire, emerald, aquamarine, chrysoberyl, topaz, tourmaline, peridot, garnet, pearl, opal, spinel, zircon, turquoise, nephrite jade, and jadeite jade)

USA
- Aquamarine
- Emerald
- Jadeite jade
- Nephrite jade
- Peridot
- Ruby
- Sapphire
- Topaz
- Tourmaline
- Turquoise

Canada
- Diamond
- Nephrite jade

Mexico
- Opal
- Topaz
- Tourmaline
- Turquoise

Honduras
- Opal

Colombia
- Emerald

Brazil
- Chrysoberyl
- Diamond
- Emerald
- Opal
- Topaz
- Tourmaline

Germany
- Topaz

Italy
- Tourmaline

Former Czechoslovakia
- Garnet
- Opal

Egypt
- Emerald
- Peridot
- Turquoise

Nigeria
- Aquamarine
- Sapphire
- Spinel
- Topaz

Zaire, Angola, and Namibia
- Diamond

Zambia
- Chrysoberyl
- Emerald

Botswana
- Diamond

South Africa
- Diamond
- Emerald
- Peridot
- Ruby
- Tourmaline

East Africa
- Aquamarine
- Diamond
- Emerald
- Ruby
- Sapphire
- Tanzanite
- Tourmaline

Madagascar
- Aquamarine
- Chrysoberyl
- Topaz
- Tourmaline

Russia
- Demantoid garnet
- Diamond
- Emerald
- Nephrite jade
- Topaz
- Tourmaline

Afghanistan
- Aquamarine
- Ruby
- Spinel

Pakistan
- Aquamarine
- Emerald
- Grossular garnet
- Ruby
- Spinel
- Topaz

India
- Almandine garnet
- Aquamarine
- Chrysoberyl
- Diamond
- Emerald
- Ruby
- Sapphire

Sri Lanka
- Chrysoberyl
- Garnet
- Ruby
- Sapphire
- Spinel
- Topaz
- Tourmaline
- Zircon

Myanmar (*formerly Burma*)
- Chrysoberyl
- Jadeite jade
- Peridot
- Ruby
- Sapphire
- Spinel
- Topaz
- Tourmaline
- Zircon

Thailand
- Almandine garnet
- Ruby
- Sapphire
- Zircon

China
- Aquamarine
- Diamond
- Nephrite jade
- Peridot
- Ruby
- Sapphire
- Turquoise

Japan and Taiwan
- Jadeite jade
- Topaz

Australia
- Diamond
- Emerald
- Nephrite jade
- Opal
- Sapphire

New Zealand
- Nephrite jade

Guatemala
- Jadeite jade

Iran
- Turquoise

Gem-testing laboratories, museums, and specialist gemstone collectors are often the first to hear of a new find. Some gemstones will be mined in sufficiently large volumes to reach the marketplace worldwide, for example tanzanite (the blue variety of the mineral zoisite), which was discovered in 1967 in Tanzania and is now available through many retail outlets. Others may be mined out fairly quickly, with most of the gemstones going to a few specialist buyers.

Some particularly rare finds are found only in museums or private collections. Other gems are cut for collectors but, because of their rarity or physical properties (for example they may be too soft for everyday wear), do not reach the retail market. New localities and new gemstone finds of interest, such as a particularly fine-coloured gemstone or a mine that is producing particularly clean large specimens of a gemstone, are usually reported in gemmological journals such as the *Journal of Gemmology*, *Gems and Gemmology*, and *Australian Gemmologist*.

Alluvial and Eluvial Deposits

Gems formed in igneous or metamorphic rock may be subject to weathering and erosion, breaking down the rock and releasing the gemstones (*see* **Weathering**). The gems may remain where they are or be transported by ice, wind, or water and deposited elsewhere. Eluvial gem deposits are formed as a result of the weathering of rock that has remained in the same place. Alluvial deposits are deposited by flowing water (*see* **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings).

Gemstones are generally harder and heavier than the surrounding minerals. They are not carried as far

by flowing water and tend to sink faster, becoming concentrated in pockets or areas along riverbanks or within gravels (gem gravels). Therefore, alluvial gemstone deposits contain gemstones that are sufficiently hard and durable to withstand the conditions without breaking, rather than those that are heavily included or prone to fracture or cleavage.

Because nature has already partially eliminated the weaker specimens, the percentage of gem-quality gemstones in gem gravels is usually high, with the result that more gemstones are retrieved from gem gravels than from any other type of deposit. Alluvial deposits, such as the gem gravels of Sri Lanka and Myanmar (formerly Burma), contain a wide range of gemstones including ruby, sapphire, spinel, chrysoberyl, topaz, tourmaline, and garnet.

Because gemstones associated with gem gravels are often found together, the discovery of one type of gemstone from a gem association (sometimes called a tracer gem) can be used by exploration teams and prospectors to 'trace' potential gemstone mining areas. Another technique is to map the courses of ancient river beds or present-day rivers and streams and then follow tracer gems downstream to find areas where the gemstones are present in large enough concentrations to be retrieved.

The oldest and most traditional mining methods are still practiced in areas where gemstones are near the surface and relatively easy to find and retrieve and where labour is cheap. In Indonesia, Malaysia, Sri Lanka, and India, for example, local people search rivers and streams and excavate the gravels and sediments from now-buried riverbeds.

Panning for gems works on the principle that the gems are generally heavier than the surrounding mud, pebbles, or rock fragments. As the pan of water and sediment is 'jiggled', the gemstones settle towards the bottom of the pan while the lighter constituents and water are washed over the pan's edge (Figure 1). The heavier concentrate may be sieved to separate larger gems or it may be spread out on tables or cloths to be hand-sorted.

Alluvial deposits are generally mined by similar methods wherever they are found in the world, usually by digging shafts or pits or by collecting the alluvium from rivers and streams using sieves and pans. Dams may be constructed to manage the water flow, and water may be diverted to wash the gem gravels (Figures 2 and 3). Final sorting is done by hand.

Often the mining and retrieval of gems in these traditional ways is a family project, varying only slightly from country to country. The men and boys may work in the rivers or dig the pits, while women and girls sieve or sort the gems (Figure 4).

Figure 1 A lady river panning, washing gravel in search of chrysoberyl cat's eyes (Kerala, India). Printed with permission from Alan Jobbins.

Ruby and Sapphire Deposits

The rubies and sapphires of Thailand and Cambodia are found in alluvial and eluvial gravels, derived from highly alkaline basalt (Figure 5). The main mining area is near the border between Thailand and Cambodia, around Chantaburi.

Sri Lanka is a source of alluvial sapphires, which are derived from pegmatite and gneiss and are found in a wide range of colours, including the pinkish-orange padparadscha. Padparadscha is particularly rare and is named after the sinhalese for lotus flower, whose colour it resembles.

Rubies and sapphires are members of the corundum family, and different trace elements within them give rise to the colour range. Rubies are red (coloured by chromium and possibly vanadium); blue sapphire is coloured by iron and titanium. Other sapphires, such as green sapphire, pink sapphire, and mauve sapphire, result from the addition of various trace elements during the gem's formation. Colourless (white) sapphire has the simple chemical composition aluminium oxide (Al_2O_3).

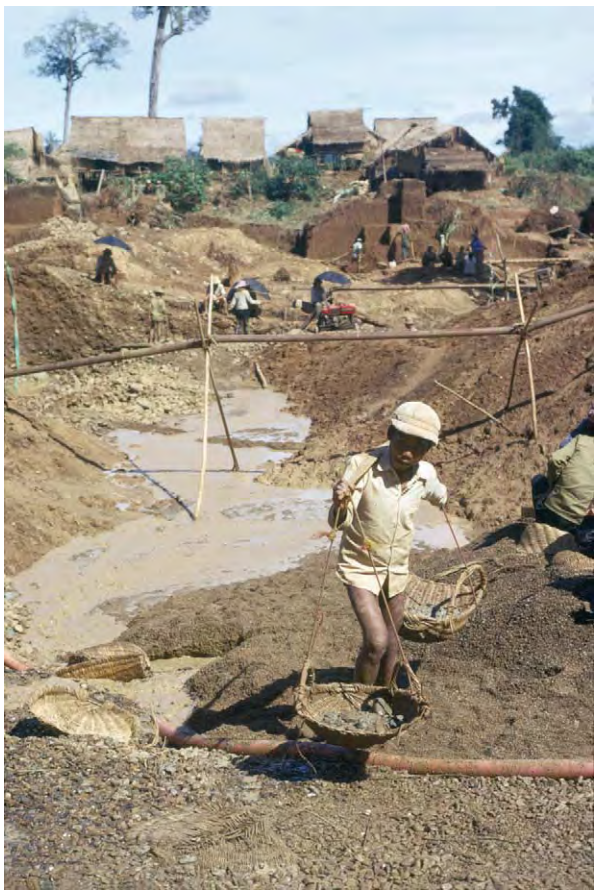


Figure 2 A young worker carries the gemstone concentrate called 'pay dirt' or 'rai' in 'Burmese style' baskets (Pailin, Cambodia). © Alan Jobbins.

Figure 3 Washing gemstone concentrate in a small pond (Pailin, Cambodia). © Alan Jobbins.

The colour does, to some extent, affect the mining, as the more popular and valuable colours are more sought after. It should be noted, however, that some gemstones, including corundum, may have their colour changed or enhanced with heat treatment, irradiation (electrons, X-rays, etc.), or other techniques.

Kashmir sapphires, which are famous for their cornflower colour, are mined in Sanskar, an inhospitable mountainous area often covered by snow, northwest of the Himalayas. The sapphires are found in feldspar pegmatites or in gravels derived from the pegmatites. There has been intermittent mining since the 1920s; however, little is known about present mining because of the political situation.

Mogok Rubies

The most famous rubies are the 'pigeon blood'-coloured rubies from the Mogok Stone Tract area of northern Myanmar, about 110 km north of Mandalay. Documentary evidence shows that mining has been carried out since the sixteenth century, but there have been finds of prehistoric tools that suggest that mining has been carried out for many thousands of years.

Traditional methods involve digging a narrow pit or vertical shaft, just wide enough for the miner to be lowered down to dig out the layer of weathered metamorphosed crystalline limestone (marble) that contains the rubies. The miner removes rock (often 2–3 m below the surface) and places it in a basket to be hauled to the surface and sorted.

Following the establishment of the Burma Ruby Mine Company in Victorian times, some mining techniques were modernized. Water pumps were introduced to speed the washing process, and intricate channels were constructed from wooden planks, along which water was sluiced to separate the rubies from rock fragments. As a result of the collapse of the market following the introduction of synthetic rubies and sapphires in the early twentieth century, the British Company failed and traditional methods were reintroduced. In the 1990s new mines were established in the Mong Hsu area, about 250 km east of Mandalay.

Figure 4 Girls sorting gemstone concentrate (Pailin, Cambodia). © Alan Jobbins.

Figure 5 Eluvial workings in weathered (decomposed) basalt above basalt lava (Pailin, Cambodia). © Alan Jobbins.

Gemstones in Igneous Rocks

Where rock is harder, picks and drills may be sufficient to prise loose the gemstones from the parent rock (the host); otherwise the rock has to be mined, crushed, washed, and sorted to retrieve the gemstones. Diamonds are mined on a larger scale and with more highly mechanized methods than any other gemstone.

Extrusive Igneous Rocks

Alluvial gem deposits derived from basaltic lavas have been mentioned above. The basaltic lavas (extrusive igneous rock, where magma has been erupted from volcanoes as lava) may also be mined. Peridot is commonly formed in basaltic lava. The peridot generally forms as small crystals or within vesicles (small bubbles) and voids in the lava. Other gems that crystallize from lavas as they cool include zircon, ruby, and sapphire.

Intrusive Igneous Rocks

Some of the largest crystals form where the magma is not erupted but rises sufficiently in the Earth's crust to cool slowly. The slower the cooling, the larger the crystals that can form. Igneous rocks formed in this way are said to be intrusive.

Gems may form in dykes (small igneous intrusions) intruded into surrounding rocks, which may crop out at the surface as a result of weathering. Mining

of weathered rock on the surface and the retrieval of gems from these eluvial deposits may lead to mining of the gem-bearing rock beneath. Different mining methods will be used as conditions and rock types change; for example, mining of the Montana sapphires of North America began as surface mining and dyke digging. The harder rock of the dykes – the intrusive igneous rocks that held the rubies – was mined in preference to the non-gem bearing rock on either side. As the gem-bearing rocks were followed deeper, mining changed from surface and dyke digging to underground mining.

Granite is a common intrusive igneous rock. Gemstones found in granite include quartz, feldspar, and tourmaline, but often the crystals are not of sufficient quality to be cuttable or may be too difficult to remove from the rock without damage. In this instance, the rock may be mined for use with the gemstones still *in situ* and used as a decorative stone, for example as a work-surface in a kitchen, a stone floor in a shopping mall, or a façade on a building (*see* **Building Stone**).

Pegmatites Pegmatites are intrusive igneous rocks. They produce a greater range of gemstones than any other rock type and have also been the source of some of the largest gemstones ever mined. The large crystals form as the water-rich portion of a granite-like molten rock is put under increased pressure as it is squeezed into fractures in the surrounding rock. As the molten rock begins to solidify, the elements that it contains begin to crystallize. The largest gemstones and some quite rare gemstone varieties are found in gem pockets at the centre of the pegmatite, where they have formed from the hot concentrated mineral-rich fluid that was the last to crystallize (Figure 6).

Pegmatites occur throughout the world, but the largest gem-producing pegmatites are those of Minas Gerais, Brazil. Other pegmatite areas include the Pala area of San Diego County in California, the Nuristan area of Afghanistan, the Sverdlovsk region of the Urals, and the Altai region of north-western China. Gems formed in pegmatites include topaz, tourmaline, kunzite, and members of the beryl family, such as the blue aquamarine and the pink morganite.

Hydrothermal gemstones Gems are also formed from hydrothermal fluids (from 'hydro' meaning water and 'thermal' meaning hot), which escape from magmas and may contain rare elements such as fluorine and beryllium. As the hydrothermal fluids move away from the magma along fractures and fissures in the surrounding rock, they solidify and form mineral veins (*see* **Mining Geology:** Hydrothermal Ores). Close to the surface, hydrothermal veins may also include elements carried by groundwater and other near-surface waters. Amethyst, topaz, benitoite, and emerald are examples of gemstones found in hydrothermal veins.

Figure 6 Workings in granite with cross cutting pegmatite, which is a source of aquamarine and chrysoberyl (River Marambio Baia, Minas Gerais, Brazil). © Alan Jobbins.

In April 2000, while blasting for a new tunnel in the Naica silver and lead mine of central Chihuahua, Mexico, miners found two underground caverns containing crystals of selenite, some of which were more than 15 m in length. Hydrothermal fluids deposited gold, silver, lead, and zinc near the surface, but groundwater in these caves formed the huge crystals. Although the caverns were kept secret, for fear of vandalism or theft, the conditions within were sufficiently hostile to keep most people out. With temperatures of more than 50°C and humidity above 85%, visitors became uncomfortable and disorientated in less than 10 min. Mining of such large gemstones has generally been carried out only to supply museums and to satisfy collectors of spectacular specimens.

Gemstones in Metamorphic Rocks

All rocks, including sedimentary and igneous rocks, may be altered by pressure and/or temperature to form metamorphic rocks. Metamorphic rocks include marble, gneiss, and schist. The gemstones they contain will depend on the composition of the rock before metamorphism and on the conditions of pressure and temperature to which they have been exposed.

Ruby and Sapphire

The area affected by metamorphism may cover only a few metres (usually as a result of local faulting or folding) or hundreds or thousands of kilometres (owing to larger regional tectonic events such as mountain building). For example, the rubies of Myanmar are associated with the mountain-building episode that formed the Himalayas more than 65 Ma ago, when the continent of India moved north, colliding with the mainland and pushing up the highest mountains in the world. The Myanmar rubies and sapphires (Mogok), Pakistan rubies (Hunza Mountains), Tanzania rubies (Morogoro), and Vietnam rubies (Luc Yen deposit and Quy Chau deposits) occur in marble formed by the metamorphism of limestone.

Kenya rubies (Mangari area of the Tsavo West National Park) are found in micaceous metamorphic rocks; the main mines are the Penny Lane and John Saul mines. The Penny Lane ruby mine is worked as an open cast pit and, along with the John Saul mine, is a modern and highly advanced plant. Large construction vehicles follow tracks down into the pit to collect the rock fragments that have been blasted from the rock face. These are then taken to the washing and sorting plants, where the rubies are retrieved.

Mica schist, formed from metamorphosed shale and slate, may contain gems such as garnet, andalusite, kyanite, and iolite (cordierite). The best jadeite jade is from Myanmar. Formed by high-pressure metamorphism, it is found as river boulders or mined from dykes in serpentinized rocks.

Emerald

In Colombia, the Muzo emeralds are found in thin layers of white limestone in soft black carbonaceous shales. Explosives are used and enormous trucks transport the soft shales to washing plants, where the harder limestone is separated from the shale and the emeralds are retrieved. Landscapes are altered as hillsides are removed. In addition to the large-scale removal of the rocks, groups of people also work on a far smaller scale. As the remaining shales are weathered, the gems are washed out and transported downhill into the valleys below. Local people search the riverbeds and sediments for emeralds, often using spades or just their hands.

The Muzo and Chivor regions are the two main Colombian emerald-mining regions. Fine-coloured emeralds are produced from both regions, though those from the Muzo tend to have a more yellowish-green colour than the bluish-green Chivor emeralds. Private companies with government supervision run the mines at Muzo. The main Chivor mine is privately run. The government issues short-term permits, usually lasting only 5 years, then issues new ones.

The Colombian emeralds have a long mining history. They were used for decorative and ceremonial purposes before their discovery by the Spanish conquistadors, who introduced them to Europe. After this they became the source of most emerald jewellery until the end of the nineteenth century.

During Roman times emeralds were mined from sites in the Habachtal area of the Austrian Alps, about 3000 m above sea-level. The oldest known source of emeralds is the famous Cleopatra Mine of Ancient Egypt, situated on the Red Sea, where mining is thought to have been carried out as long ago as 4000 years BP. Mining was a labour-intensive practice, and it is thought that thousands of miners worked below ground in the labyrinth of tunnels.

More recently emeralds have been discovered in other countries, including Brazil, Australia, Pakistan (Swat Valley), Afghanistan (Panjshir Valley), Russia (near Sverdlovsk, Ural Mountains), South Africa, Zimbabwe (Sandawana Valley), Tanzania, and Zambia. They occur in hydrothermal veins, in schists (including mica schist and weathered talc schist), and in metamorphic limestones (including dolomitic marble). In some areas they are associated with pegmatite intrusions.

The quality and colour of the emeralds varies between countries and from mine to mine, but many compete well with the best Colombian emeralds, producing high-quality gemstones. Production has varied, often depending upon the political situation in a country rather than the potential of the mine.

Specimens may have characteristic inclusions that can be used to confirm their country of origin and even sometimes the mine from which they came (*see* **Fluid Inclusions**). For example, emeralds from Chivor can contain inclusions of small pyrite crystals, while those from Muzo may contain calcite crystals. Both the Chivor and the Muzo emeralds have distinctive three-phase inclusions, comprising a gas, a liquid, and a solid. Sandawana emerald is famous for its inclusions of the mineral tremolite, which look like a series of fine bars; they may also contain two-phase inclusions (containing a liquid and a gas bubble). Some Russian emeralds contain inclusions

of the mineral actinolite and shiny yellowish-brown mica flakes. Indian emeralds have a characteristic two-phase inclusion that resembles a comma with a jagged outline.

Other Sedimentary Gemstone Deposits

In addition to alluvial and eluvial gemstone deposits, gemstones may also form in veins or cavities within rocks or as crusts on their surfaces. Following transportation, they react with other elements to produce new minerals or may be deposited directly from solutions. New minerals may also form deposits when water cools, evaporates, or changes level, leaving a mineral residue. Examples of gemstones formed in rocks as a result of the evaporation, cooling, or transportation of mineral-rich fluids include turquoise, malachite, rhodochrosite, agate, and amethyst.

Turquoise usually forms as a thin layer or vein within the rock. The rock is mined and the turquoise removed for polishing or fashioning, often as beads or small pieces for use as inlays. Malachite can occur as large botryoidal (rounded) masses or as thin layers and is often associated with azurite. Rhodochrosite can occur as large outcrops, which are mined using heavy machinery and cut and polished as smaller decorative pieces.

Agates are fairly common and are usually found as harder geodes within rock faces of igneous origin. Underground mining, following rock layers containing the geodes, results in a labyrinth of tunnels and mineshafts. The agate mines in Idar-Oberstein (Germany), which have been mined for agates since Roman times, are now popular with tourists and groups of visiting gemmologists.

Opal is formed in sedimentary rock where organic remains have decomposed to leave minerals within the rock or cavities into which minerals have been transported by fluids. Large specimens, such as fossilized dinosaur bones, may be replaced by opal, and small irregular voids within a rock can be filled.

Australia is the best-known and most important source of opal. The main mining areas are Lightning Ridge, Coober Pedy, White Cliffs, Quilpie, and Andamooka. Tourists can obtain licenses to mine, checking the ground and riverbeds for small rock fragments containing opal and panning rivers during their stay. Mining companies work on a larger scale with mechanized vehicles and diggers. In order to escape the extreme heat, some miners live in subterranean houses, where only the rooftops show above the surface.

Conclusion

The mining method used depends on whether the gemstones are found *in situ* (in the host rock in which they formed) or as secondary placer deposits eroded from or washed out of the host rock and accumulated elsewhere. The locality, its accessibility, the availability of a local workforce, and the political situation are also factors. The specific properties of gemstones – their beauty, rarity, durability, and value, which is often a result of fashion as much as size or clarity – will also affect the mining method used. The method chosen should not damage fragile or particularly valuable specimens, as the aim is to maximize production and, ultimately, profit.

See Also

Building Stone. **Fluid Inclusions**. **Igneous Rocks:** Kimberlite. **Mining Geology:** Exploration; Mineral Reserves; Hydrothermal Ores. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings. **Weathering**.

Further Reading

Hall C (1993) *Gems and Precious Gemstones*. London: Apple Press.

Hall C (1994) *DK Eyewitness Handbook Gemstones*, 2nd edn. London: Dorling Kindersley.

Hughes RW (1997) *Rubies and Sapphires*. Boulder: R W H Publishing.

Mumme I (1988) *The World of Sapphire*. Port Hacking: Mumme Publications.

Oldershaw C (2003) *Philip's Guide to Gems*. London: Philips.

Oldershaw C, Woodward C, and Harding R (2001) *Gemstones*, 2nd edn. London: Natural History Museum Publications.

Sevdermish M and Mashiah A (1996) *The Dealer's Book of Gems and Diamonds*, 2nd edn. Gemmology (A M) Publishers Ltd.

Sinkankhas J (1981) *Emeralds and Other Beryls*. Radnor, PA: Chilton Press.

Webster R (1994) Read PG (ed.) *Gems: Their Sources, Descriptions and Identification*, 5th revised edn. London: Butterworths.

GEOARCHAEOLOGY

L Joyner, Cardiff University, Cardiff, UK

Introduction

The interface and overlap between the Earth sciences and archaeology has developed into the subject called geoarchaeology or archaeological geology, where geological concepts and methods are applied to archaeology. The definition of exactly what geoarchaeology encompasses is still open to discussion. The emergence of this new sub-discipline really took off in the mid-twentieth century when geological techniques began to be applied to archaeological sites and materials, although the origins of the use of geological principles and techniques in archaeology date back to the nineteenth century when the two disciplines were developing. The emergence of geoarchaeology coincided with the development of the 'new archaeology' which embraced a much more scientific approach to the study of antiquity. The first approach of geoarchaeology was the application of geological techniques to archaeology, but has since developed into a more integrated interdisciplinary approach. Geoarchaeology encompasses both field and laboratory techniques and includes stratigraphy for determining the succession of occupation levels, sedimentological studies related to archaeological site formation, geomorphology of archaeological landscapes, pedology, geochronology, petrology, geochemistry of sediments and artefacts to determine composition, technology and provenance, and geophysics used to detected subsurface archaeological features. Thus geoarchaeology deals with both natural materials and processes, and artificial 'manmade' materials and landforms. Some definitions of geoarchaeology only include stratigraphy, site formation processes, and geomorphology, and omit the analysis of artefacts and geophysical techniques. This overview with cover the broader definition of geoarchaeology.

Stratigraphy

The principles of geological stratigraphy are equally applicable to archaeological stratigraphy, a fact that has long been accepted in archaeology. Geoarchaeologists are able to interpret the nature of the stratigraphic units; whether they are natural deposits or are manmade. Most archaeological deposits can be treated as special kinds of sediments, and sedimentological principles and interpretations applied. A thorough knowledge of postdepositional processes is required, as these can significantly affect the original structures and textures, as well as the stratigraphy. Techniques of stratigraphy have been used to date artefacts found in specific layers, either directly through annual laminae, or indirectly through dating of, for example volcaniclastic deposits.

Soils and Sediments

The study of soils and sediments in and around archaeological sites provides much evidence for the landscape, environmental setting, as well as for human activities and natural events. An assessment is usually made of the types of deposits encountered including composition, texture, chemistry, particle sizes and shapes, sediment structures, and colour. These details can be used to identify the type of soil or sediment, and assess how the deposit accumulated through time. Soil micromorphology has been successfully used to identify areas within sites that were used for specific activities such as cooking areas, smelting of metals, and animal enclosures.

Site Formation Processes

Site formation processes often involve both natural and anthropogenic processes of accumulations or depletions of sediments and/or soils through time and space. The study of these sediments and soils involves identifying the type and composition of the deposit and its source, as well as the identification of any weathering, transportation, environmental, or postdepositional processes involved. Much experimental and ethnographic research has been conducted in order to understand the degradation processes operating at archaeological sites, and the rate and nature of sediment accumulation over time and space. For example, experimental studies have examined the way in which ditches silt up, and ethnographic studies have been concerned with the rate of degradation of mudbrick (adobe) structures.

Geomorphology

Changes occur in the geomorphology of the landscape over time, such as the position of the coastline and the courses taken by rivers. The study of these changes can reveal the shape of the landscape at a particular time in history. The shape of the landscape

would have affected the way in which ancient peoples interacted with their environment. Therefore, the study of the geomorphology of the area around archaeological sites can provide valuable evidence about the topography in antiquity and how ancient peoples used the landscape in particular periods. It can provide information on the agricultural potential, the availability of water supplies and raw materials, the communication routes available, and the defence potential of a site. Many archaeological sites that were formerly by the sea or a river are now several kilometres away from the modern shoreline or river bank. It can also reveal evidence for catastrophic events that may have affected past civilisations. For example, tephra deposits on Santorini and Crete in Greece provide evidence for the 1628 BC eruption of the volcano on Santorini (Thera) which is thought by some archaeologists to signify the decline of the Minoan civilisation on Crete. Geomorphological studies can also reveal evidence for manmade structures, for example the Iron Age Hillforts that are common in England are manmade earthworks. The geomorphological study of archaeological sites through the time periods they were occupied is very important. Past landscapes may have been very different to the pattern of the landscapes seen today, having changed through time. The interpretation of how a site operated could thus be drastically different and therefore wrongly interpreted, if the modern landscape was assumed to be identical to the ancient one. For example, the ancient Greek city of Thermopylae was located on the coast in 480 BC when the Battle of Thermopylae was fought between the Spartans and the Persians. Today the coastline is some 5 km away from this site (Figure 1).

Geomorphological studies can also provide information on how an archaeological site has changed through time and the impact this will have had on the preservation of a site, whether it has been preserved by a covering of sediment or partially destroyed by erosional processes. Geomorphological studies will provide information on environmental changes in areas around archaeological sites, and will thus aid environmental reconstructions.

The exploitation of natural resources can significantly affect the landscape. For example, the development of quarrying and mining activities leaves behind visible signs of human activity.

Geomorphological surveys involve mapping topographic features seen in the field and noting the drift geology to produce geomorphic maps. The data collected is supplemented by the use of aerial photography and satellite images, as well as by analysis of core drills

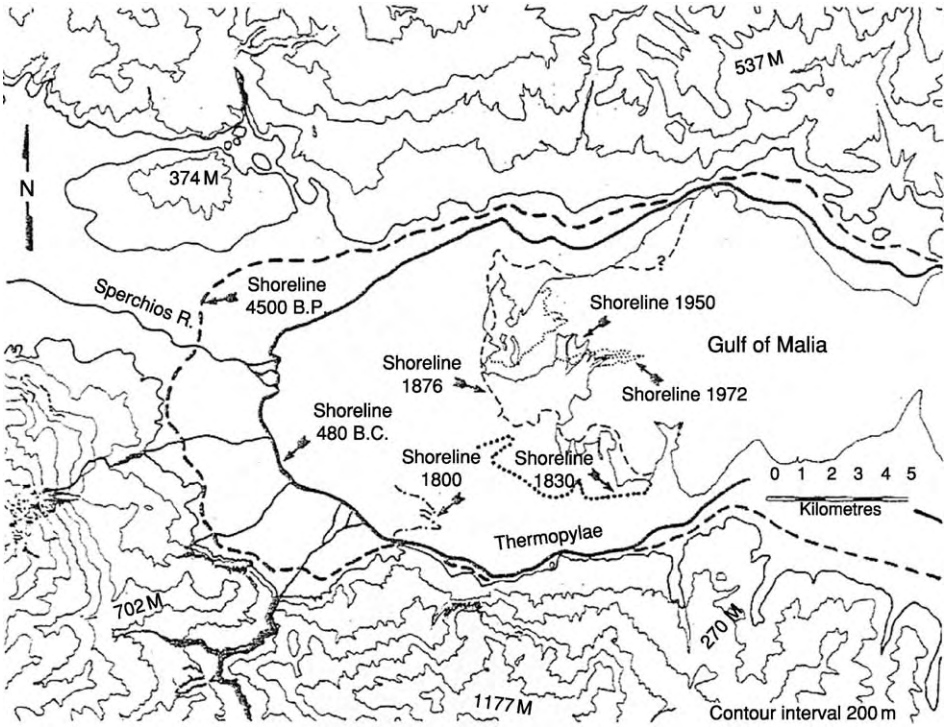


Figure 1 Map showing where the coastline has moved from 480 BC to the present day. The ancient Greek city of Thermopylae was located on the coast in 480 BC when the Battle of Thermopylae was fought between the Spartans and the Persians. Today the coastline is some 5 km away from this site. (Kraft *et al.* (1987) *Journal of Field Archaeology* 14: 181 198. Reproduced from the Journal of Field Archaeology with permission of the Trustees of Boston University. All rights reserved.)

taken from specific areas of a site. Core drilling has been used in geomorphological studies to ascertain the nature, stratigraphy, and extent of human occupation. This is a relatively rapid and cost-effective method which is potentially minimally destructive.

Aerial Photography

Aerial photography has been used since the beginning of the twentieth century to identify archaeological features not usually visible at ground level, or those that are buried. Both vertical and oblique photographs can be used, as each can accentuate different features. Stereoscopic pairs of aerial photographs are often used as they give a three-dimensional image of the topography. Some buried features will be visible through differential crop markings which result in contrasting crop growth over the buried feature. For example, ditches and pits are shown up by superior crop growth.

Geophysical Methods

A number of geophysical techniques have been adapted and applied to archaeology, since the mid-twentieth century. These include electrical resistivity, magnetometry, gravitometry, and ground penetrating radar. These techniques allow the identification of archaeological sites and their layouts thus aiding the interpretation of ancient settlements and manmade structures prior to potential excavation. These techniques are often used on their own, with no further need for excavation. They are used to locate and detect the extent of archaeological sites, as well as to map the distribution of features within known archaeological sites (Figure 2). In addition to being used to gain information about the spatial distribution of features, geophysical methods can be used to provide information on features buried at variable depths thus giving information on site development and the stratigraphic sequence of a site. The geophysical prospection of urban sites has become increasingly common with the redevelopment of 'brownfield' sites within large cities which require an archaeological survey to be carried out before development is possible. However, electrical and magnetic methods can prove problematic when used in modern urban areas.

Magnetometry identifies differences in the magnetic properties of soils which may have been caused by human activities such as burning, humic decomposition, compaction, and by building of structures. Magnetic surveying is difficult if it is carried out in an area with nearby interfering magnetic sources such as near modern buildings and power lines, which tend to obscure the weaker signals from archaeological features. Electrical resistivity detects differences in the electrical conductivity of the soil which may have resulted from the construction of ditches and mounds, and compaction. The soil moisture content is important for electrical resistivity surveying, as it cannot be undertaken if the soil is too dry.

Geohysical methods are non-destructive and more cost-effective than test excavations. A geophysical survey will provide valuable information which can be used to plan an excavation programme effectively and efficiently, as well as providing site information without the need for excavation.

Magnetic Susceptibility

Magnetic susceptibility surveys have been used to identify areas of human activity, as these can change the magnetic character of soils and sediments at the site of these activities such as cooking, metal processing, and other activities involving heating. This technique is usually used in conjunction with magnetometry for optimum results.

Archaeoseismology

Archaeoseismology is the study of the effects of earthquakes in antiquity. The devastating effects of sizeable earthquakes can be recorded in the archaeological record. The way in which buildings and other structures collapse can give information on the earthquake event, and can be used to date destruction layers at some sites. The Modified Mercalli scale of earthquake intensity is used to interpret the size of the earthquakes that caused damage at archaeological sites. Stone buildings tend to be more resistant to earthquake damage than wooden buildings, and mud-brick structures are most susceptible to earthquake damage. In addition, the type of deposit that an archaeological site is built upon will also affect the intensity of damage caused by an earthquake. Most damage is caused where buildings are constructed on unconsolidated sediments, whereas a harder underlying rock will reduce the severity of earthquake damage. The interpretation of archaeoseismic events can be confused with other events which produce similar patterns of destruction, such as landslides and poor construction techniques of buildings.

Palynology

Palynology, the study of pollen, can reveal evidence for ancient cultigens, dietary information, and the seasonal occupation of sites. Changes in the types of vegetation on a site may be significant and can be used to indicate changes of use of a site or climate

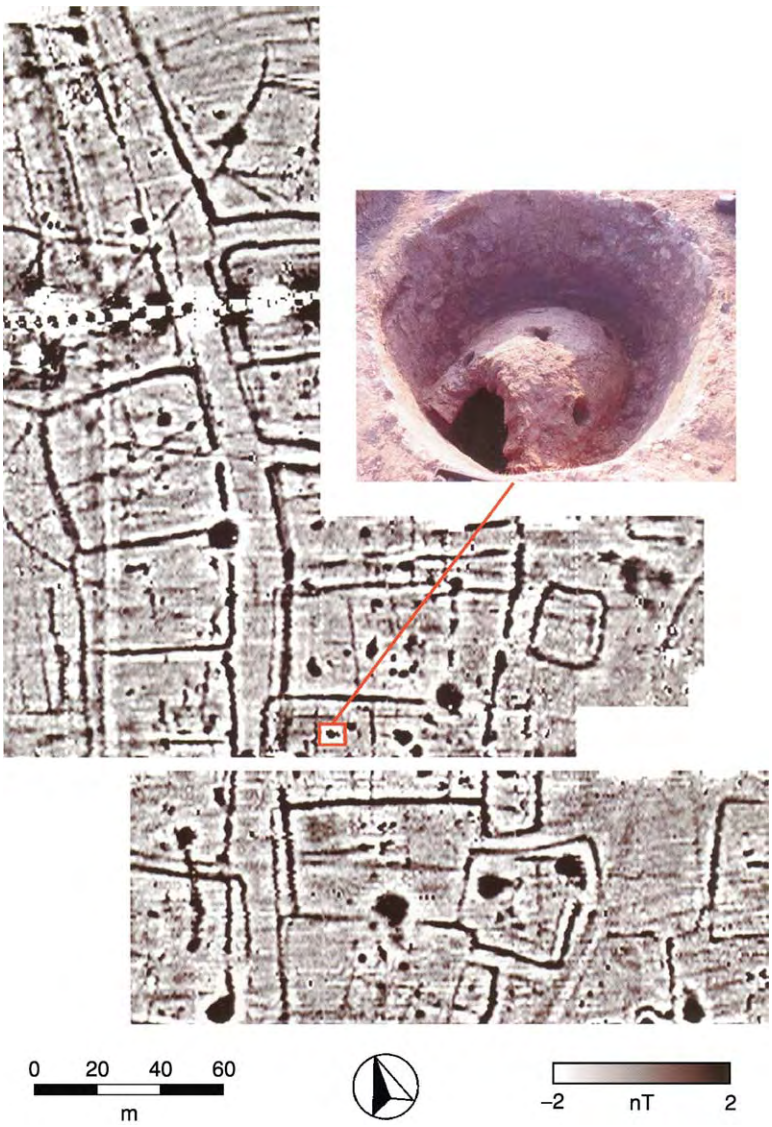


Figure 2 Sedgefield, County Durham. Fluxgate gradiometer data. 1 × 0.5 m. Romano British settlement straddling a road. Excavation by Time Team revealed a complex of workshops and a small kiln as highlighted in the inset photograph. (Reproduced from Gaffney C and Gater J (2003) *Revealing the Buried Past: Geophysics for Archaeologists*, Tempus Publishing Ltd (Plate 19).)

change. Palynology has been extremely important in aiding environmental reconstructions from archaeological evidence, particularly with the identification of native and non-native pollen species types.

Characterization Studies

Characterization studies seek to address questions about the raw material constituents of artefacts, their provenance, ancient trade, or distribution routes, their technology of production and manufacture, and patterns of consumption. Various geochemical, mineralogical, and imaging techniques have been applied to archaeological materials to answer these questions. These techniques may involve taking samples from artefacts. However, the recent trend has been towards micro-sampling and non-destructive analysis.

Geochemistry of Archaeomaterials

The geochemistry of a range of archaeomaterials has been carried out using a number of different techniques which are also used in the Earth sciences. Major, minor, and/or trace elements are detected by a number of analytical techniques. The range of archaeomaterials include ceramics, metals, glass, lithics (particularly obsidian), and organic materials. The geochemical signature of these archaeomaterials has been used to answer a number of archaeological

questions. The chemical characteristics of particular artefacts are used to predict where the artefacts were made, provenance their source of production, as well as to source the raw materials used. This is done by comparing the geochemical signatures of unprovenanced artefacts with the geochemical signatures of locally available raw materials, or the geochemical signature of artefacts that have been produced in known localities (Figure 3). The multi-element chemical data obtained from known and unknown artefacts is usually processed by various multivariate statistical techniques. These techniques show how the samples group together, thereby suggesting a close association and possible provenance. Many artefacts of material culture are artificially produced, so cannot be directly compared with local raw materials sources. These include metal alloys, ceramics, glass, and faience. Lithics may be compared directly with rock sources. The chemistry of material artefacts has also helped to deduce what they are made of, and to suggest the technological processes involved in production. The geochemical techniques used in archaeological studies have included optical emission spectroscopy (no longer used), which was superseded by atomic absorption spectroscopy (commonly used in the 1970s to the 1990s), and then more recently by inductively coupled plasma atomic emission spectroscopy and mass spectroscopy (favoured techniques since the 1990s). Instrumental neutron activation analysis has been used for trace element determination, but this technique is declining in use as reactors are closed down. Surprisingly, wavelength dispersive X-ray fluorescence spectrometry has not been widely used in archaeological science. However, air-path energy dispersive X-ray fluorescence spectrometry has been used regularly, especially for metals, but also for other materials, as it is a non-destructive technique. Proton induced X-ray emission spectrometry and proton induced gamma ray emission spectrometry analysis have also been used, as has electron probe microanalysis. Chromatographic techniques have been applied to the study of organic archaeomaterials and these include gas chromatography and gas chromatography mass spectrometry.

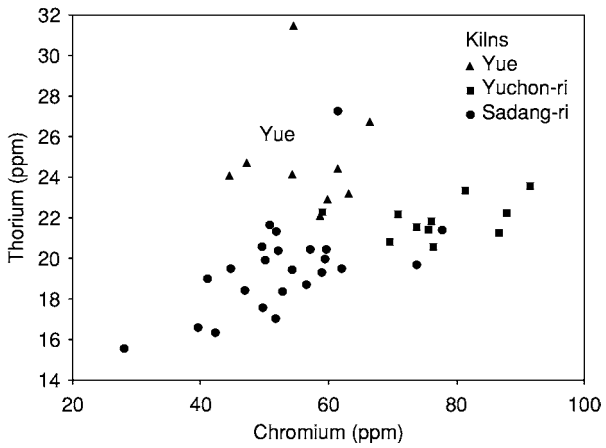


Figure 3 Geochemical grouping of trace element data from Korean Celadon ceramics dating to the Koryo dynasty (twelfth century AD). The bivariate plot shows the concentrations of Cr (ppm) plotted against Th (ppm) (determined by instrumental neutron activation analysis) for ceramic pots manufactured at different kiln sites in Korea and Chinese Yue ceramics for comparison. The trace element data shows that the geochemical signature of the clays used in these two Korean kiln sites are distinguishable from each other and the Chinese Yue ceramics. (Reproduced from Hughes M and Joyner L (2000) In: Portal J (ed.) *Korea: Art & Archaeology*, British Museum Press. Courtesy of the Trustees of the British Museum.)

Postdepositional processes can also affect the geochemistry of artefacts, substantially altering their geochemical signature. This particularly affects ceramics. Recycling, such as the addition of scrap metals in metal production or the addition of cullet in glass production, obscures the chemical signatures restricting the use of geochemistry in provenancing artefacts made of metals and glass.

Metals and Ores

Metals such as gold, silver, lead, tin, copper, iron, and arsenic were used in antiquity. Some were used in their raw state, such as gold and copper, and others were alloyed, such as copper and tin to produce bronze. Geoarchaeologists are concerned with all aspects of metal production, from ore sources to extraction and mining, to roasting of ores, smelting, melting, refining, alloying, casting, and the use and trade of metal artefacts (Figure 4).

Stable Isotopes

The use of stable isotope analyses in artefact provenance has increased rapidly in recent years. Carbon, oxygen, sulphur, strontium, and lead have all be used to provenance materials such as lithics, metals, and glass. Marble provenancing was one of the first studies to use stable isotopic data to discriminate between sources in the Aegean region. A large database of the isotopic signatures of $\delta^{13}C$ and $\delta^{18}O$ for marble quarries in Greece and Turkey has been built up which has been used for provenancing artefacts and for associating broken pieces of artefacts (Figure 5).

Mineralogy of Archaeomaterials

Many artefacts of material culture are made from geological raw materials such as rocks for building and statuary, clays for ceramics, sand for glass

production, limestone and gypsum for plaster production, gemstones for decorative purposes, minerals for pigments, and metals for tools. Mineralogical techniques that are commonly employed in the Earth sciences have been applied to the study of ancient artefacts. These include optical microscopy, polarising light microscopy, and X-ray diffraction.

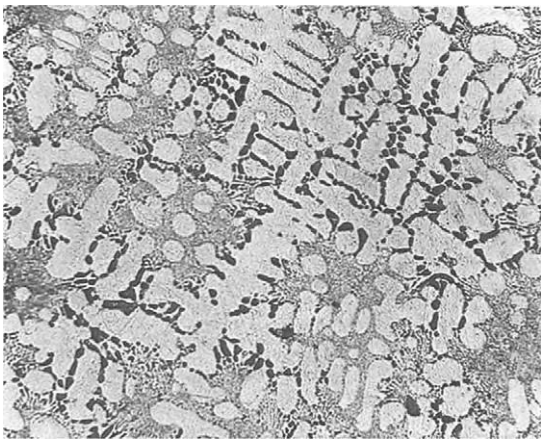


Figure 4 Structure of the cast silver copper alloy of an Egyptian silver figure of Nefertum (EA66818). Polished and etched sample. Field of view 0.5 mm across. (Reproduced from Cowell M and La Niece S (1991) In: Bowman S (ed.) *Science and the Past*, British Museum Press. Courtesy of the Trustees of the British Museum.)

Other techniques, which are less commonly used in the Earth sciences, have also been applied to the study of ancient artefacts, and these include infrared spectroscopy and Raman microscopy.

Lithic Petrology

In archaeology, the term lithic or stone is used to describe rock that has been utilized by humans. The use of lithics in antiquity dates back to before the Paleolithic period and includes building stone, statuary, tools, vessels, sarcophagi, jewellery, and pigments. Petrological examination of lithic artefacts has allowed potential sources of these raw materials to be suggested. This allows suggestions to be made for potential trade/exchange routes that may have been taken to import the lithic materials/artefacts to a site where they were excavated, as well as highlighting contacts between past groups of peoples. For example, the identification of Preseli Mountain stone at Stonehenge has raised the possibility that the stone was transported from South Wales to Wiltshire specifically to be used in the construction of Stonehenge.

Ceramic Petrology

Petrological techniques have been applied to pottery to determine clay paste compositions, the possible

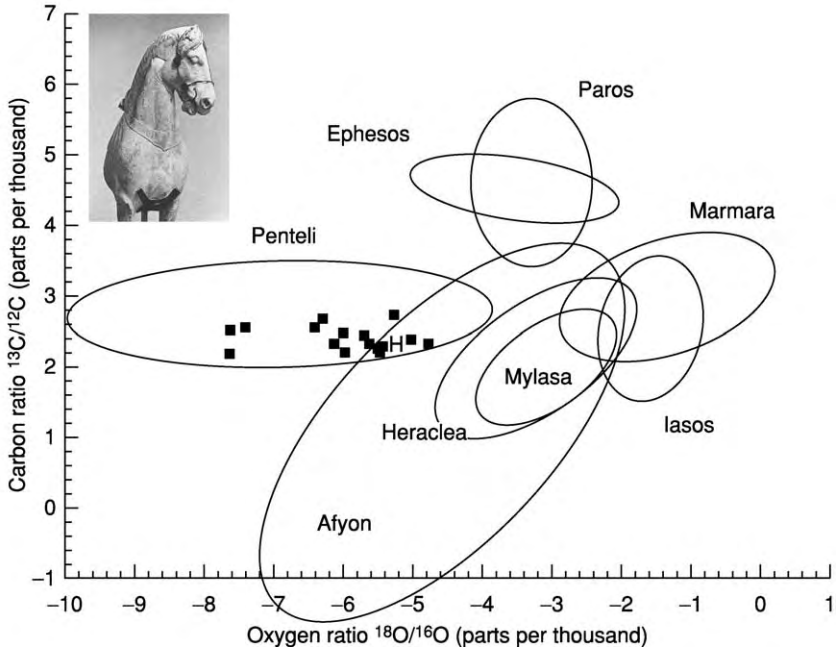


Figure 5 Stable isotope analysis for carbon and oxygen of the marble sculptures from the mausoleum of Halicarnassus including the horse, denoted by H (see inset), superimposed on the data for Classical marble quarries. The ellipses represent 90% confidence limits for the samples from the named quarries. (Reproduced from Hughes M (1991) In: Bowman S (ed.) *Science and the Past*, British Museum Press. Courtesy of the Trustees of the British Museum.)

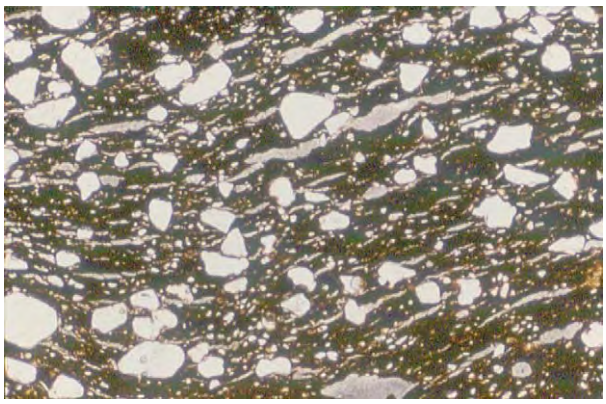


Figure 6 Thin section of an Early Byzantine pottery sherd from a cooking pot excavated from the Sanctuary of St Lot, Deir 'Ain Abata, Jordan. Plane polarized light. The fabric shows the clay micromass (brown) in which rounded quartz sand grains (white) have been intentionally added (temper). The thin elongate structures are voids which are naturally incorporated into the clay as it is worked (wedging) prior to vessel formation. This vessel has been wheel thrown and has been well fired in an oxidising atmosphere. Field of view is 3 mm across. (Reproduced from Joyner and Politis (2000) *Internet Archaeology* **9**. http://intarch.ac.uk/journal/issue9/daa_toc.html)

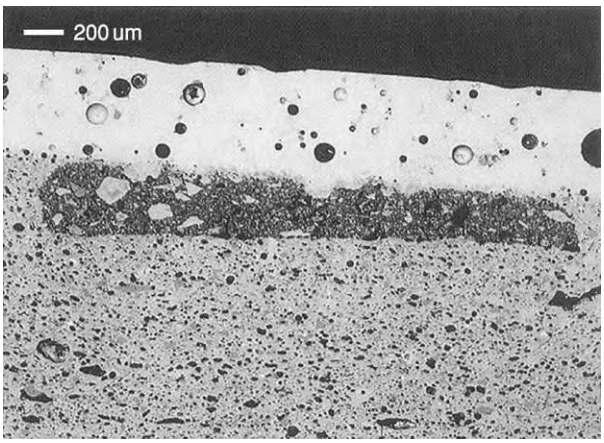


Figure 7 Back Scattered Electron Microscope image of a Korean Celadon ceramic dating to the Koryo dynasty (twelfth century AD) showing the ceramic body (grey), white inlaid decor ation composed of kaolinite (rectangular dark area) and alkaline glaze (white). Note the scale bar. (Reproduced from Hughes M and Joyner L (2000) In: Portal J (ed.) *Korea: Art & Archaeology*, British Museum Press. Courtesy of the Trustees of the British Museum.)

addition of temper (material such as sand, shell, or crushed pottery or rock added to the clay to improve its workability and/or firing properties), provenance, and firing conditions (Figure 6). The use of ceramic petrology has revealed the complexity of pottery making and the long distance trade of pottery, either for itself or for the contents the vessels may contain. The study of prehistoric pottery in the American Southwest in the 1940s revealed a widespread trade which was unexpected in such an early period. Petrological techniques have also been applied to bricks, mud bricks, plasters, cements, mortars, and concretes.

Technology Studies

Various analytical techniques have been employed to deduce how artefacts were made. Some of these techniques include the chemical and mineralogical techniques already mentioned. In addition, imaging techniques are frequently used to elucidate the sequence of manufacture of artefacts of different materials. Microscopy, scanning electron microscopy, and radiography are the techniques that are frequently used (Figure 7). For example, in pottery studies, polarising light microscopy is used to deduce the clay paste recipes used, the scanning electron microscope is used to estimate the approximate temperatures achieved during the firing process, and radiography is used to look at the forming techniques used to construct pottery vessels.

Geochronology – Dating Techniques

Before the advent of modern dating techniques, archaeologists relied upon stratigraphy to determine the relative ages of cultural deposits. Since the middle of the twentieth century, a number of scientific dating methods have been applied in the archaeological sciences which have provided absolute ages, which are independent of the stratigraphy. These include radiocarbon dating (or ^{14}C dating), 40Ar-39Ar dating, Uranium series dating (which involves the radioactive decay of ^{238}U, ^{235}U & ^{232}Th to lead), electron spin resonance dating, thermoluminescence dating (releases accumulated energy by application of heat), obsidian hydration (the rate of absorption of water is dependent on glass chemistry and temperature, resulting in an hydration rim whose thickness can be used to estimate the time elapsed since the obsidian artefact was created), and dendrochronology (tree rings).

Conclusions

Geoarchaeology combines both the Earth sciences and archaeology in the quest for interpreting the material culture of past civilizations. It has become a broad but distinct subdiscipline within archaeology requiring a firm knowledge of both geological techniques and processes, and archaeology. It is hoped that this broad, though not necessarily comprehensive, review of geoarchaeology will give a feel for the range

of work being conducted in archaeology which requires a geological understanding of the environment and ancient materials. It is a dynamic discipline in which there is a continual interaction and cross-over of the two subject areas.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray); Mineral Analysis.

Further Reading

Bowman S (ed.) (1991) *Science and the Past*. London: British Museum Press.

Davidson DA and Shackley ML (1976) *Geoarchaeology: Earth Science and the Past*. London: Duckworth.

French C (2003) *Geoarchaeology in Action: Studies in Soil Micromorphology and Landscape Evolution*. London: Routledge.

Geoarchaeology. Journal published by Wiley Interscience.

Goldberg P, Holliday VT, and Ferring CR (2001) *Earth Sciences and Archaeology*. New York: Kluwer Academic/Plenum Publishers.

Herz N and Garrison E (1998) *Geological Methods for Archaeology*. Oxford: Oxford University Press.

Leute U (1987) *Archaeometry: An Introduction to Physical Methods in Archaeology and the History of Art*. Weinheim, Cambridge: VCH.

Rapp GR Jr and Hill C (1998) *Geoarchaeology: The Earth Science Approach to Archaeological Interpretation*. New Haven: Yale University Press.

Renfrew C and Bahn P (2004) (Fourth Edition) *Archaeology: Theories, Methods and Practice*. London: Thames and Hudson Ltd.

Stiros S and Jones RE (eds.) (1996) *Archaeoseismology*, Fitch Laboratory Occasional Paper 6, British School at Athens.

Waters MR (1992) *Principles of Geoarchaeology: A North American perspective*. Tuscon: University of Arizona Press.

GEOCHEMICAL EXPLORATION

E M Cameron, Eion Cameron Geochemical Inc., Ottawa, ON, Canada

Introduction

Mineral deposits are small in size, ranging in width from several metres for a gold vein to a few kilometres for a large porphyry copper deposit. The deposits that were not found during the early decades of exploration are usually concealed beneath younger geological formations. All deposits can be discovered by grid drilling, but this would be enormously expensive. Thus, a body of exploration techniques has been developed to narrow the search to promising targets. The exploration geologist searches for rocks and geological features that are known to be associated with mineral deposits. The geophysicist looks from the air or ground for patterns that may indicate mineralization. The geochemist seeks element signatures that have dispersed outwards and upwards from a deposit. The aim is not to be constrained by the small physical size of a deposit, but to look for evidence that will enlarge the target and make it easier to locate. Much of the globe has been assessed by geochemical surveys guided by materials such as soils, stream and lake sediments, waters, and plants. Originally, the sole purpose of these surveys was to identify targets for mineral exploration (*see* **Mining Geology:** Exploration), but information on the geochemistry of the environment is becoming an increasingly important objective.

Near-Surface Dispersion of Elements

One of the most frequently used geochemical approaches for expanding target size is the sampling of stream sediments. Where a stream cuts across a partially exposed deposit, it removes mineral fragments and disperses these downstream. Several kilometres downstream, there may be minerals that are characteristic of the deposit. These may be detected by chemical analysis of the stream sediment or by physically separating and identifying minerals. Most mineral deposits contain sulphides (*see* **Minerals:** Sulphides). These are not stable when in contact with air, and they oxidize, releasing the constituent elements. Some elements, such as zinc, which is released during the oxidation of sphalerite, are soluble in water and move downstream in solution. Other elements are less soluble and form secondary minerals that are precipitated close to the source or move as solids suspended in stream waters; lead in galena, for example, forms the secondary mineral lead sulphate. Thus there may be 'clastic' dispersion, which is the movement of solids, and 'hydromorphic' dispersion, which is the movement of dissolved elements. Elements in solution can be measured by analysing the

water in which they are dissolved, but they tend to precipitate as the water moves downstream; analysis of sediments can therefore identify elements that have been dispersed in both clastic and hydromorphic form. In areas where streams are scarce and lakes are abundant, such as in the Canadian and Fennoscandian shields, both lake sediments and waters are used in analyses. Glaciers erode mineral deposits and disperse the minerals. After melting of the continental-wide glaciers of the northern hemisphere, vast areas of till and related glacial sediments were exposed. Evidence of 'dispersion trains' or 'fans' leading out from mineral deposits can be obtained by analysis of the tills or the soils developed on the tills. Also, because sulphide minerals may have been dispersed with the tills, oxidation of the tills and the included sulphides may lead to secondary hydromorphic dispersion from the fans into streams and lakes (Figure 1).

Once a target is defined by stream or lake sediments, soils are usually collected and analysed to define the target more closely. Plants can also be sampled and analysed. Plants have the advantage of extending their roots over an area and depth that is greater than the volume comprising a soil sample, and thus may give a more representative estimate of trace elements present. In arid regions, some plant roots extend to great depth to reach water, and thus may be mineralized near a deposit. A few species of plants grow in abundance on soils with a high content of a particular metal and can be used to identify the metal-rich area.

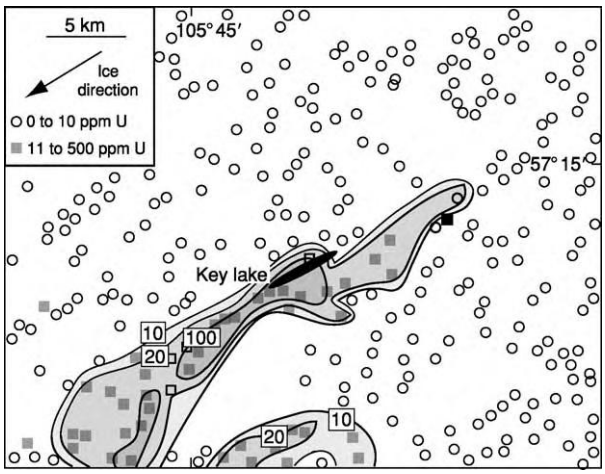


Figure 1 Uranium content of lake sediments from the uranium nickel deposit around Key Lake, Saskatchewan. Contours at 10, 20, and 100 ppm U. Ice has dispersed fragments of the mineralization to the south west in glacial tills. Following retreat of the ice, the tills and the uranium and nickel minerals were weathered, then uranium was dissolved as an anion and moved into the lakes. Within the lakes, the uranium precipitated in organic rich centre lake sediment. Anomalous contents of uranium in waters were first found in 1969. Subsequent detailed geochemistry, including lake sediment sampling and boulder tracing, led to the discovery of the deposit in 1975. Modified with permission from Maurice YT, Dyck W, and Strnad JG (1985) Secondary dispersion around the uranium nickel deposit at Key Lake, northern Saskatchewan. In: Sibbald TII and Petruk W (eds.) *Geology of Uranium Deposits. Canadian Institute of Mining Metallurgy Special Volume,* pp. 38 47. Montreal: Canadian Institute of Mining, Metallurgy and Petroleum.

The 'mobility' of an element is a critical aspect of its usefulness in geochemical exploration. Mobility is an indication of how far an element may travel dissolved in water and thus broaden the signal derived from the mineral deposit. Some metals dissolve as cations, others dissolve as anions. Cations are most soluble at low pH. Sulphide deposits that oxidize at the surface may produce waters with a pH of 3 or even lower. In acidic waters, most metal cations are soluble. As waters derived from such a deposit flow from the deposit, they mix and become diluted by other waters and their pH increases. As the pH increases, cations are precipitated in a sequence that corresponds to their different chemical properties. In water that originally contains lead and zinc, lead will be precipitated first, at a pH lower than the pH that precipitates zinc. Elements that dissolve as anions behave differently, because these are most soluble at neutral and high pH; having a negative charge, they are not adsorbed by colloids with a negative surface charge. Because most surface waters and groundwaters are not acidic, elements dissolving as anions, such as uranium, molybdenum, arsenic, and rhenium, are most useful as wide-travelling indicators.

Diamond Exploration

Diamonds are a high-value commodity. Raw diamonds worth six billion dollars are sold each year; a single mine may produce several billion dollars worth of diamonds over the life of the mine, and one-half billion dollars is expended each year on exploration. Diamonds come from the upper mantle. Fragments of diamond-bearing rock are picked up by kimberlite magma (*see* **Igneous Rocks:** Kimberlite) rising from the deeper mantle en route to the surface. The magma has a high content of volatiles and moves upwards at a speed that has been likened to that of an express train. When the kimberlite reaches the surface, the sudden release of magmatic gas and interaction with groundwater causes explosions, producing pipe-shaped diatremes filled with fragmented kimberlite and country rock. Most pipes have a small surface area, usually 50 to 500 m in diameter, but occur in clusters. Kimberlite is soft and easily eroded, forming depressions that become lakes or are covered by overburden. The ease of erosion, which leads to their

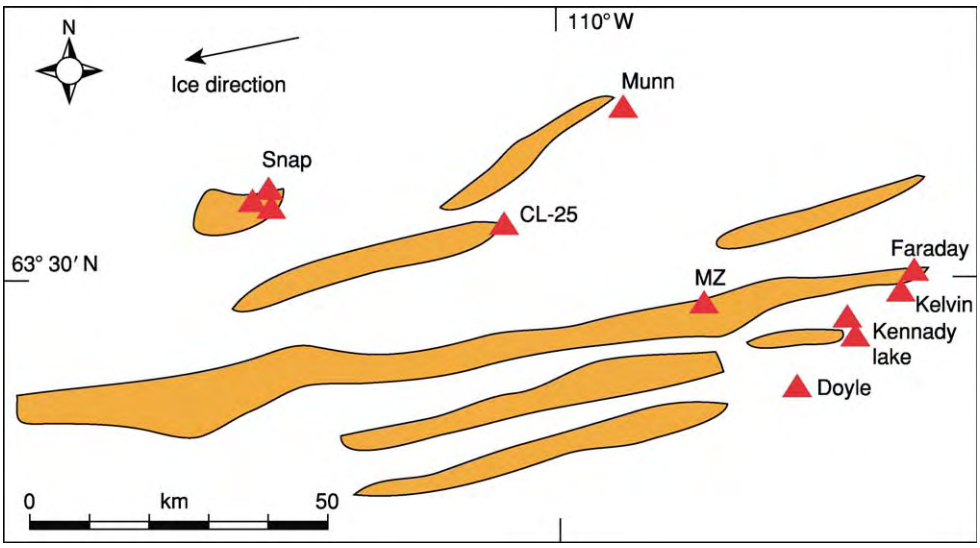


Figure 2 Dispersal trains of kimberlite indicator minerals in glacial tills, south east Slave province, Canadian Shield. Triangles indicate locations of known kimberlites. Indicator minerals first found in the Mackenzie valley, 600 km to the west, led Charles Fipke and Stewart Blusson to the discovery of this rich diamond province. Original figure by John Armstrong.

concealment, also provides a key to their discovery: their mineral constituents may be dispersed far by streams, glaciers, or wind.

Being directly derived from the mantle, the minerals forming kimberlites are different from those of most other rocks exposed at the surface and are of higher specific gravity. When these minerals (for example, pyrope garnet, chrome diopside, chromite, ilmenite, and magnesian olivine) are identified in a heavy mineral concentrate of stream or river sediment or till, they may be followed to the source. In the Slave diamond province of northern Canada, slim, pencil-like glacial dispersal trains of kimberlite indicator minerals (KIMs) in glacial tills can extend far from their sources (Figure 2). Once the approximate sources of the KIMs are determined, other techniques, mainly geophysical, can be used to locate the kimberlites. Diamond, being so hard, is well preserved during erosion and transport, but its concentration is so low in the kimberlites – less than a part per million – that it is rarely seen in KIM concentrates (*see* **Igneous Rocks:** Kimberlite).

Exploration for diamonds would be simple if every kimberlite contained economical amounts of the mineral. But that is not the case; less than 1% of kimberlite intrusions is worth mining. So, how can KIMs that relate to diamondiferous kimberlites be identified? The clue to the answer was first provided by the electron microprobe analyser. Diamonds often contain small inclusions of other more common minerals, derived from the same source region in the mantle. Using the microprobe, these can be identified and their compositions determined. Thus, it was found that inclusions of pyrope garnet in diamonds have lower contents of calcium than do similar garnets from elsewhere, and chromites in diamonds have unusually high chromium contents. Ilmenite has high magnesium and low ferric iron contents. With the knowledge gained from inclusions, kimberlite minerals with specific compositions, derived from the part of the mantle where diamonds coexist, but much more abundantly than diamonds, could be sought in the KIM concentrate. Ilmenite compositions have an additional significance. Diamond is a reduced substance and can be destroyed if its host region in the mantle becomes more oxidized, or during transport within a kimberlite magma. Ilmenite, with high magnesium and low ferric iron contents, is thus indicative of reduced conditions that preserve diamond. Figure 3 shows four different fields containing levels of MgO and Fe_2O_3 for ilmenite from four kimberlite localities in southern Africa. In the most oxidized field (1), no diamonds are preserved. In the most reduced field (4), diamond crystals are preserved and show late-stage overgrowths.

Deep-Penetrating Geochemistry

Most undiscovered mineral deposits lie beneath some form of cover, which may range up to several hundred metres in thickness. The cover may be laterite, the product of severe modification of the underlying rocks during tropical weathering. In this case, there may be sufficient geochemical clues retained from the parent material that analysis of the lateritic soils can identify the presence of a buried mineral target.

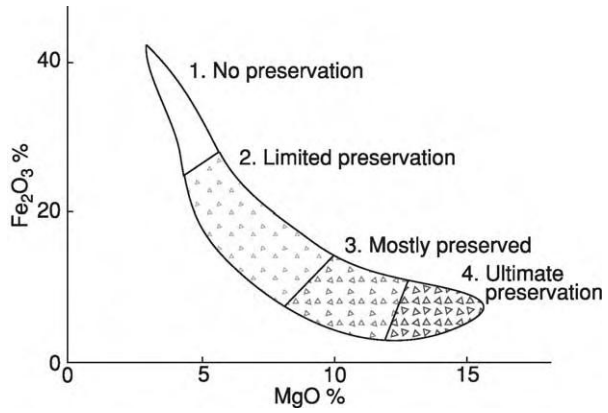


Figure 3 Composition of ilmenite as a predictor of the preservation of diamond in the host kimberlite. Figure based on samples from kimberlites in southern Africa. See text for discussion. Modified with permission from Gurney J and Zweistra P (1995) The interpretation of major element compositions of mantle minerals in diamond exploration. *Journal of Geochemical Exploration* 53(1 3): 293 309.

However, in many cases, the cover is 'exotic': it has been derived from elsewhere and has no relationship to the underlying rocks. Examples are piedmont gravels, desert sands, alluvial deposits, and most types of glacial deposits. Exotic cover makes it more difficult to locate the buried target; identification must depend on the migration of elements from the deposit, through the cover, to the surface. The indicator elements must migrate either as a gas (for example, radon or sulphur gases) or as metals dissolved in water. This involves movement through a zone that is unsaturated in water, the vadose zone, which can be tens to hundreds of metres thick in arid climates. In the vadose zone, the mineral grains are surrounded by a thin film of water, which moves slowly downward to the water table. Migration may take place by diffusion of ions through the water film or as gases diffusing through air space in the vadose zone. An alternative mechanism to diffusion is advection, which is the movement of an entire mass of air or groundwater containing the gas or dissolved ion.

The exploration geochemist uses a variety of gases in soils, including hydrocarbons and sulphur gases, to explore for buried deposits. For the migration of gases from depth, an interesting experiment was performed by CR Carrigan and colleagues at the Nevada Nuclear Test Site. Chemical explosives equivalent to a 1-kt nuclear charge were detonated at a depth of 400 m in bedded tuff. Two bottles of gas were placed near the charge, one containing ^{3}He, the other containing SF_6. Sampling sites were established at the surface to detect these gases. SF_6 was the first gas to be detected, after 50 days, along a fault, during a strong barometric depression. ^{3}He was first detected at the surface 325 days later. These results do not fit with gaseous diffusion; compared with SF_6, ^{3}He has a much higher diffusivity and should reach the surface long before SF_6 does. The diffusivity of SF_6 is such that it would require tens to hundreds of years to reach the surface, yet it has happened in days. Why? The reason is that gaseous diffusion is overridden by a much faster advective mechanism, barometric pumping. High barometric pressure forces air down fractures and into pore space in the rock, around the fractures. Gases within the rock mix with air in the pores. When the barometric pressure drops, air in the porous rock, now containing the gases, returns to the fracture and, after several cycles of high and low pressure, reaches the surface. Pumping occurs because the volume of air entering rock porosity is much greater, compared to the volume of air present in the fractures. Pumping provides the "breathing volume" that permits large vertical movements during high-pressure 'inhalation' and low-pressure 'exhalation'. Barometric pumping can withdraw gases from depth several orders of magnitude faster than is possible by molecular diffusion. But why did it take much longer for ^{3}He, compared to SF_6, to reach the surface? This is because a gas with high molecular diffusivity can more readily diffuse out of the upward-moving air–gas mixture in the fractures, into porous wall rock.

In the search for buried deposits, great attention has been given to the detection in soils of elements that have been transported to the surface as dissolved constituents in water. Initial studies focused on chemical diffusion of ions in water. But rates of diffusion are very low and the water films around grains in the vadose zone move downwards at a rate that can be several orders of magnitude faster than the rate at which ions diffuse upwards. It is analogous to trying to walk up a very fast down escalator. As with gases, the most effective way to bring dissolved ions to the surface is by advection, the upward transport of mineralized groundwater. Capillary migration is one such mechanism for advective transport, but it is effective only for depths below the surface not much greater than 10 m. A different process, earthquake-induced surface flooding, or seismic pumping, has been identified as being effective over much greater depths. This process has long been known to seismologists. Surface flows of groundwater occurred along fault lines in desert areas of Iran during earthquakes in 1903, 1923, and 1930, and more recent examples have been observed in Montana and California. During earthquakes, stress fields can become compressional, closing fractures in basement rocks and forcing groundwater to the surface, up faults.

Because many mineral deposits occur along faults, mineralized groundwaters bathing a buried deposit can become incorporated in the upward flow. The Atacama Desert, a region in northern Chile of high seismicity and hyperaridity, produces the world's greatest quantity of copper. The principal deposits were found exposed at the surface, but large areas of prospective ground are covered by piedmont gravels of Miocene age. EM Cameron and colleagues have found soils with elevated amounts of various porphyry indicator elements (copper, molybdenum, rhenium, and selenium), above porphyry deposits buried beneath gravels. These anomalies lie along fracture zones in the indurated gravels, which appear to have formed during the reactivation of older basement faults that were involved in the formation of the deposits (Figure 4). These soils are more saline than are soils away from the fracture zones. Mineralized, saline groundwaters have been pumped to the surface during earthquakes; these have evaporated, then the constituents have been redistributed by the infrequent rains. In the case of the Spence copper deposit in northern Chile, the proportions of elements found in the soils correspond to those found in groundwater 80 m below.

A very different environment is found in the Abitibi belt of Ontario, Canada. Here world-class gold and base metal deposits of Archaean age have been mined for over 100 years. The first deposits were found exposed in areas free of glacial cover. However, large parts of this highly prospective region are covered by thick, impervious lacustrine clays deposited during retreat of the ice sheet 10 000 years ago. Recent work by SM Hamilton and co-workers has provided encouraging evidence that geochemistry can 'see' through the clays. Hamilton's group has identified reduced (i.e., low-eH) columns in the clays; the columns extend from the sulphide subcrop to the water table and represent an upward flux of reduced substances (e.g., Fe^{2+}) derived from oxidation of the sulphides. On reaching the water table, these substances react with atmospheric oxygen to produce acid: $Fe^{2+} + O_2 + 2H_2O\ Fe(OH)_3 + H^+$. The acid dissolves carbonate originally contained in the clays,

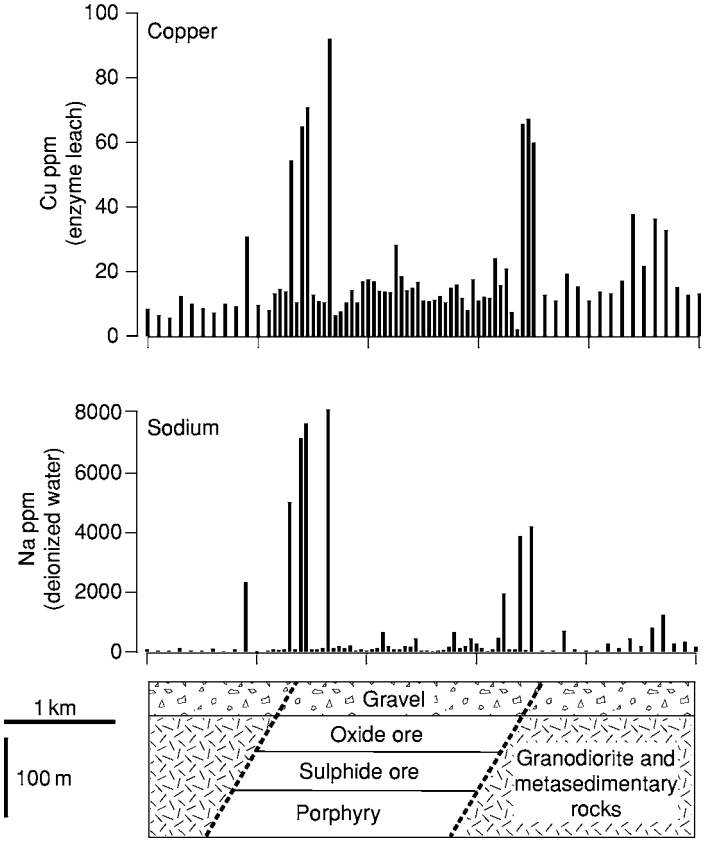


Figure 4 Analyses of soil samples for sodium and copper along a 5 km traverse across the Gaby Sur copper porphyry deposit, northern Chile. The deposit contains 400 million tonnes of 0.54% Cu and is overlain by piedmont gravels of Miocene age. Sodium was extracted by deionized water and copper was extracted by enzyme leach. Groundwater is found only in the basement below the gravel where drill holes intersect fracture zones, such as the boundary faults (shown by dashed lines). The anomalies for sodium and copper in the soils are interpreted to represent mineralized groundwater pumped to the surface during seismic activity.

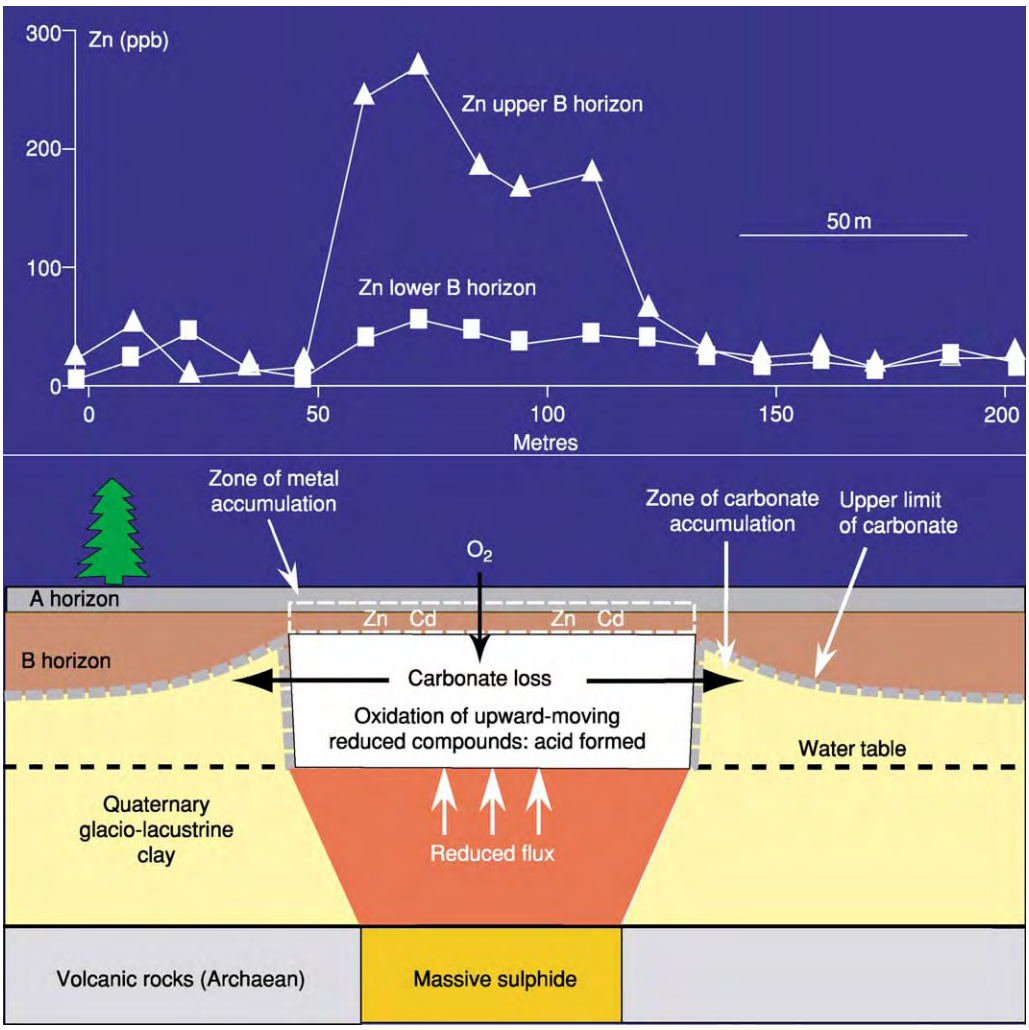


Figure 5 Diagram showing effects of an upward flux of reduced material from a subcropping massive sulphide body to the surface through impermeable clay. The oxidation of the reduced material by oxygen above the water table causes formation of acid, which removes the carbonate from the clay. Zinc and other base metals from the massive sulphide are precipitated in a thin zone of soils at the top of the B horizon. Zinc analyses of soil after extraction were by enzyme leach.

forming a zone of low pH and low carbonate in soils directly above the sulphides (Figure 5). Other elements from the sulphides, such as zinc and lead, also move to the surface and are precipitated in a thin zone, which may be the upper B horizon or Ae horizon, immediately below organic-bearing soils. Isotopic measurements of lead in this horizon show that it is similar in composition to the lead of sulphide mineralization and much different from the more radiogenic lead of the clays. Chemical diffusion is much too slow a process to explain the upward flux of these elements through 30 m or more of clay in a period of only 10 ky; other processes must be involved, and these are currently being investigated. This observation may be relevant to the use of clays to cover buried nuclear waste.

Analysis of Samples

Much of the progress in geochemical exploration has come as a result of rapid advances in analytical techniques. Today, analysts can measure a larger number of elements, to lower concentrations and at lower cost, than was possible even a decade ago. Most analyses are carried out using inductively coupled plasma mass spectrometry (ICPMS). Commercial laboratories routinely analyse samples for 50 elements at a cost of about $15 per sample. Some of the elements (rhenium, for example) are measured at concentrations below 1 part per billion (ppb). Rhenium occurs in low concentrations in molybdenite, which is present in most copper porphyry deposits. Rhenium serves as a far-travelling indicator of these deposits,

because it dissolves as an anion in groundwater. In soil from the porphyry belt of northern Chile, anomalously high concentrations of rhenium, indicative of porphyries, can be as low as 0.5 ppb, but this is still above the limit of detection when samples are analysed by a weak leach (see later) of 0.05 to 0.10 ppb Re.

As mineral deposits become harder to find, much effort has gone into amplifying the geochemical signal from the deposit, over the background noise. One of the first approaches to this was the collection of heavy minerals, as described in the preceding discussion on diamonds. Panning for gold is another example. A chemical method for amplification is selective removal or leaching of the mobile component of a soil or sediment, then measuring this component by ICPMS. 'Selective leach' reagents remove only a small fraction of the metal that might otherwise be dissolved by strong acids or by total dissolution, in the expectation that this fraction represents a more readily dissolved mobile phase derived from an ore deposit. The more abundant element fraction that comes from primary minerals forming the soil is called the endogenic phase, whereas that from external sources, including a mineral deposit, is the exogenic phase. The exogenic phase is initially introduced into the soil in water-soluble form and, as a result of soil-forming processes, is incorporated into secondary minerals, such as carbonates or iron and manganese oxides. Because the exogenic phase enters in water-soluble form, one approach is to use a weak leach that does not attack any of the minerals, but rather dissolves water-soluble salts and elements loosely adsorbed to mineral surfaces. Such is the 'enzyme leach' used to obtain the data for copper in Figure 4 and for zinc in Figure 5. For secondary minerals, ammonium acetate is commonly used to dissolve carbonate, and hydroxylamine hydrochloride at different concentrations and acidities is used separately to dissolve manganese oxides and iron oxides. Other reagents are used to dissolve the organic material that often accumulates with metals of exogenic origin.

From Regional Exploration Geochemical Surveys to Environmental Geochemical Mapping

Regional geochemical surveys using a variety of media (stream and lake sediments and waters or soils) have evolved in scale over time from covering hundreds of square kilometres to surveys that cover more than 100 000 km^2 in a field season. In the 1970s, 65% of the United States was surveyed to identify areas of potential for uranium and other minerals. Aircraft were key to this increase in productivity. Helicopters can ferry crews to sites where soils or sediments are sampled, and geochemists in low-flying aircraft can map the distribution of the radioactive series, i.e., uranium, thorium, and potassium. A few perceptive geochemists, notably JS Webb, saw early on that these surveys could reveal more information than had been mandated as their primary purpose (to identify mineral potential). In 1978, Webb and co-workers published a geochemical atlas of England and Wales based on stream sediment sampling. This showed trace element distributions that affected human and animal health. For example, areas underlain by molybdeniferous marine shales gave rise to pasture that caused molybdenum toxicity in cattle and a molybdenum-induced inability to absorb the essential trace element, copper. This problem was already known to veterinarians working in the most affected areas, but the survey revealed other areas where cattle had subclinical symptoms. Addition of copper to the cattle feed produced an increase in animal weight. Similarly, trace elements, either by their deficiency or their excess in soils and waters, can affect human health. Excess materials in the environment can have natural causes or may be the result of pollution. (*see* **Environmental Geochemistry**).

As the importance of environmental geochemical mapping became increasingly apparent, geological survey organizations began to include this purpose, in addition to resource evaluation, as a rationale for surveys. Today, environmental information has become the primary purpose of many surveys. This has resulted in modifications in the techniques applied, including the nature of the sampling media and the sampling density. Sampling of moss is now used to identify airborne pollution, for example, to map the emissions from nickel smelters in the Kola Peninsula. Early surveys for resource purposes were generally carried out at a high sampling density, with one sample per square kilometre being typical, but it became apparent that useful information could be obtained more rapidly and economically by sampling at lower densities. In 1972, the first survey of Canada's National Geochemical Reconnaissance program collected lake sediments and waters at a site density of 1 sample per 23 km^2, permitting 93 000 km^2 to be sampled in 6 weeks. There has been a progression of this trend to lower density sampling, particularly where surveys are being carried out on a multinational or global scale for environmental purposes. The Global Geochemical Baselines project collects samples on the basis of 160-km by 160-km cells. The mapping of Europe at this scale was completed in 2004 (Figure 6). Standardization of sampling and

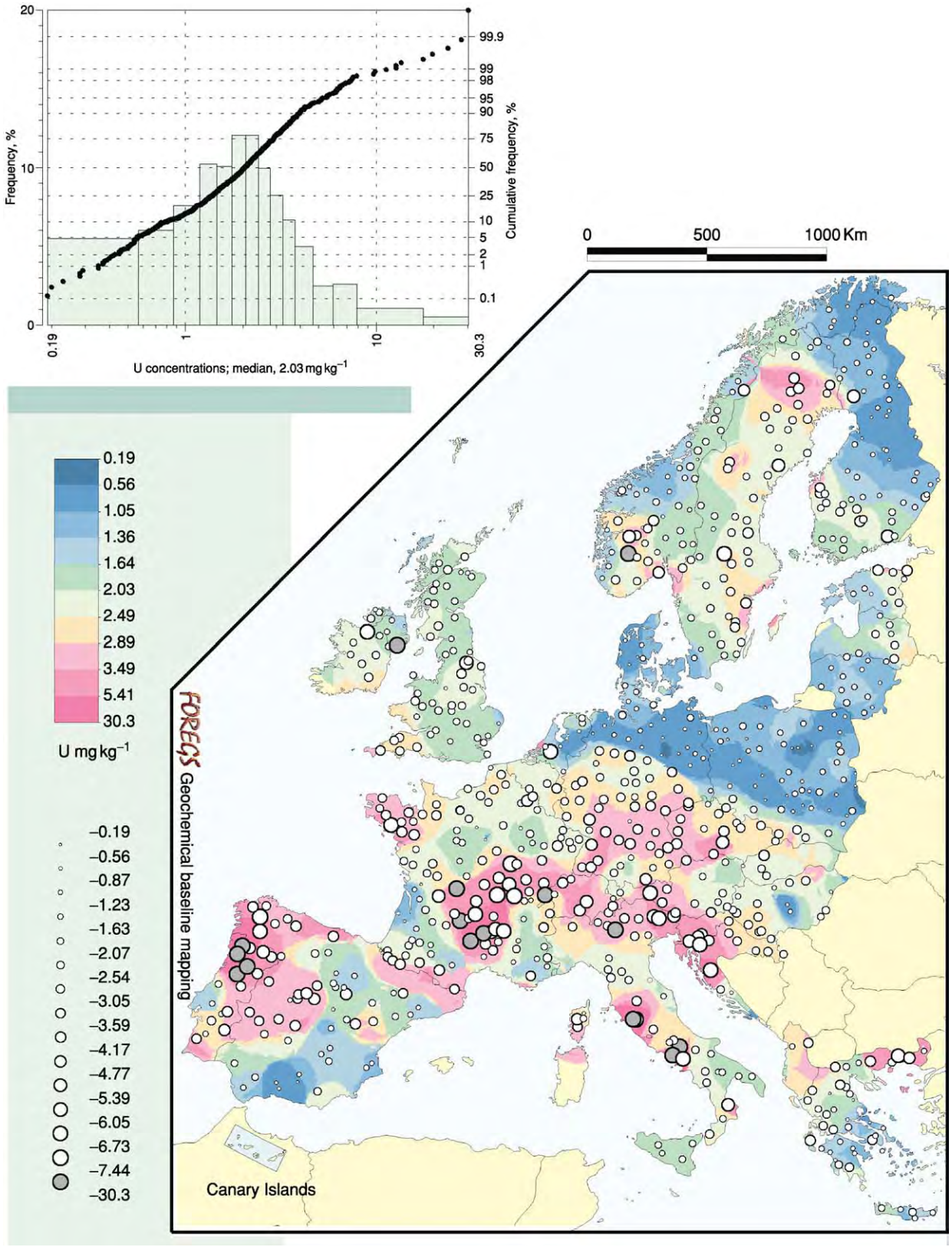


Figure 6 Geochemical map of Europe, showing the distribution of uranium in soils. Uranium concentrations of 764 samples were determined using inductively coupled plasma mass spectrometry (detection limit, 0.1 μg kg ^{1}). Courtesy of Reijo Salminen, Geological Survey of Finland. Survey sponsored by Forum of European Geological Surveys.

analytical techniques and quality control of the data were not always given full consideration in early surveys, but are now mandatory where comparisons are to be made on a global basis. Just as the mobile component of elements has been found to be most useful for identifying mineral deposits, for environmental surveys, emphasis must be given to measuring the fraction of the element that is bioavailable. Thus, selective leaches are used and, where waters are sampled, speciation calculations can be made to estimate the amounts of potentially toxic elements that are bioavailable.

See Also

Environmental Geochemistry. **Gold**. **Igneous Rocks:** Kimberlite. **Minerals:** Sulphides. **Mining Geology:** Exploration. **Sedimentary Processes:** Glaciers. **Soils:** Modern.

Further Reading

Brooks RR (1983) *Biological Methods of Prospecting for Minerals*. New York: Harper and Row.

Cameron EM, Hamilton SM, Leybourne MI, Hall GEM, and McClenaghan BE (2004) Finding deeply buried deposits using geochemistry. *Geochemistry: Exploration, Environment, Analysis* 4: 1 26.

Carrigan CR, Heinle RA, Hudson GB, Nitao JJ, and Zucca JJ (1996) Trace gas emissions on geological faults as indicators of underground nuclear testing. *Nature* 382: 528 531.

Griffin WL (1995) Diamond exploration into the 21st century. *Journal of Geochemical Exploration* 53: 1 367.

Gurney J and Zweistra P (1995) The interpretation of major element compositions of mantle minerals in diamond exploration. *Journal of Geochemical Exploration* 53: 293 309.

Hale M and Plant JA (eds.) (1994) *Drainage Geochemistry*. Amsterdam: Elsevier.

Hall GEM (1998) Analytical perspectives on trace element species of interest in exploration. *Journal of Geochemical Exploration* 61: 1 20.

Maurice YT, Dyck W, and Strnad JG (1985) Secondary dispersion around the uranium nickel deposit at Key Lake, northern Saskatchewan. In: Sibbald TII and Petruk W (eds.) *Geology of Uranium Deposits. Canadian Institute of Mining Metallurgy Special Volume*, pp. 38 47. Montreal: Canadian Institute of Mining, Metallurgy and Petroleum.

Plant J, Smith D, Smith B, and Williams L (2000) Environmental geochemistry at the global scale. *Journal of the Geological Society, London* 157: 837 849.

Rose AW, Hawkes HE, and Webb JS (1979) *Geochemistry in Mineral Exploration*. London: Academic Press.

Zeegers H and Butt CRM (eds.) (1992) *Regolith Exploration Geochemistry in Tropical and Subtropical Terrains*. Amsterdam: Elsevier.

GEOLOGICAL CONSERVATION

J E Gordon, Scottish Natural Heritage, Edinburgh, UK

Introduction

The long and varied geological history of planet Earth is reflected in the great diversity of its rocks, minerals, fossils, landforms, soils, and active geomorphological processes. In many parts of the world this diversity is under threat from a range of human activities. Traditionally, geological conservation has involved the protection, management, and interpretation of specific sites, or 'monuments', recognized to be of national or international importance on scientific grounds because of their geological and geomorphological features. More recently, a broader approach has centred on the idea of 'Earth heritage' (Earth heritage conservation) and latterly on the concept of 'geodiversity' (geoconservation). Geodiversity is the variety of geological environments, rocks, minerals, fossils, landforms, unconsolidated deposits, soils, and active geomorphological processes in a defined area. Geodiversity provides the foundation for life on Earth and for the diversity of natural habitats and landscapes, as well as for many aspects of cultural landscapes and built environments. It is therefore a valuable environmental, scientific, educational, cultural, and economic resource.

It is now increasingly recognized that proper conservation management of the non-living parts of the natural world is crucial for sustaining living species and their habitats. This leads to a broader role for geological conservation, incorporating not only the protection of geological and geomorphological features of scientific and educational value but also the management of natural environments and processes that support habitats and species. Such a role requires more integrated approaches to nature conservation and the management of sites and landscapes, which recognize the dependencies between biodiversity and geodiversity. It also embraces the involvement of people and society and their

interactions with geology, landforms, soils, natural processes, and landscapes. Geological conservation is therefore concerned with the protection and management of geodiversity both in designated sites and in the wider landscape. It includes promoting sustainable management of the natural world based on understanding Earth system processes, as well as influencing policy makers, planners, and land and water managers, and raising public awareness.

The Value of Geodiversity

The value of geodiversity ranges from the narrow scientific interest for geological research and education to the much broader economic and societal value of mineral resources and construction materials and the aesthetic and cultural benefits arising from activities such as geotourism. Geodiversity also has a key functional role in underpinning aspects of biodiversity.

Rocks, fossils, landforms, and soils are valuable resources for scientific research and education. Geology and its allied disciplines are fundamentally field-based, so that exposures of rocks and landforms are essential for continued research and education. As part of the process of scientific advancement, there is a continuing need for these sites to be available for further research in order to test and develop new theories and for demonstration and reference purposes. On economic and social grounds, such sites are essential for the training of new generations of geologists to help locate and develop the oil, gas, mineral, and aggregate resources that modern society ultimately depends upon. The applied value of geoscience is also apparent in the use of land for agriculture, forestry, mining, quarrying, building, and infrastructure, all of which are closely related to the underlying geology, landforms, and soils. Rocks, landforms, and sediments also record climate change and the dynamic history of the landscape. These environmental records allow present changes to be set in the context of past natural variations. This is important since past landscape changes provide a basis for understanding present changes and predicting future events and risks, such as those associated with flooding, earthquakes, volcanic activity, and coastal erosion.

From an Earth heritage perspective, the geological record contains a remarkable story of the Earth's history, including the evolution of life. This story embraces plate tectonics, continental drift, mountain building, volcanism, oceans opening and closing, dynamic surface processes, climate change from icehouses to hothouses, and a record of life through evolving habitats, changing biotas, and species extinctions. It is a story that deserves to be more widely appreciated as a fundamental part of the natural world in which we live. On a worldwide scale, many geological sites are unique or contain exceptional features or landforms (e.g. the Grand Canyon, the Burgess Shale, the Channelled Scablands, Giant's Causeway, and Dinosaur Provincial Park). Many sites and areas also have great historic and cultural value because of their crucial role in the development of geoscience and their association with key historical figures. For example, many important geological principles have been developed in Scotland, through the work of James Hutton (*see* **Famous Geologists:** Hutton; Lyell), Archibald Geikie, James Geikie, Hugh Miller, Benjamin Peach, and John Horne, and applied worldwide. Many of the names for periods of geological time are derived from specific geographical localities (e.g. Cambrian, Ordovician, and Silurian) and form part of the local and national heritage. At a local community level, geodiversity provides the basis for landscape heritage and sense of place. Many sites have local-amenity, aesthetic, historical, industrial-heritage, and cultural values. Urban landscapes, too, can have a close affinity with geology; this is exemplified by the volcanic hills of Edinburgh, but is also apparent in the use of local building stone (*see* **Urban Geology**).

Rocks and landforms are the basis of landscapes and scenery, which are highly valued for aesthetic and practical reasons for tourism and a range of recreation and leisure opportunities. Conservation of geodiversity therefore helps to maintain an economic resource for geotourism and recreation activities.

Geodiversity also has much wider value through its functional links with other parts of the natural heritage. Rocks, landforms, soils, and geomorphological processes provide the basis for the diversity of valued habitats and species. Management of sites for the purpose of conserving biodiversity therefore requires an understanding of their geological and geomorphological settings and current process dynamics. For example, many internationally important habitats owe their origins to geological and geomorphological processes (e.g. species-rich chalk grasslands, coastal sand-dune systems, and the wintering sites of internationally important goose populations in the estuaries and salt marshes of the UK). Active geomorphological processes maintain dynamic habitats and ecosystems through sediment and water flows, nutrient cycling, and hydrology. For example, Atlantic salmon rely on the availability of river-channel features such as pools, riffles, and glides, which in turn depend on the underlying geology and ongoing geomorphological processes of floods, erosion, and deposition. Maintaining natural process systems is a key part of conserving biodiversity as

well as of maintaining the natural landscape. Degradation of landforms or soils or the interruption of natural processes through the building of river or coast defences or the construction of dams can have adverse impacts on biodiversity.

Soil is an often overlooked component of geodiversity but provides the medium for plant growth, agriculture, forestry production, and habitat development. It also performs key environmental functions (e.g. water filtering and storage, carbon storage, acting as a reservoir for biodiversity) and records climate and land-use changes.

Pressures on Geodiversity

Pressures on geodiversity are many and varied and arise principally from urban, industrial, and infrastructure developments and changes in land use. Geological sites typically occur as natural or man-made exposures. Some sites are therefore created and maintained by natural processes; others are created by human activities such as quarrying. While there is a view that rocks and landforms are robust, significant damage and loss of key sites have occurred in the past and are ongoing. The main impacts include physical damage, destruction or removal of the focus of interest, loss of visibility and access to exposures through burial by landfill or concealment by vegetation, damage to site integrity through fragmentation of the focus of interest, and disruption of active geomorphological processes. For example, exposures in disused quarries can be lost through landfill, prime glacial landforms can be destroyed by quarrying for sand and gravel, and key exposures can be sealed and natural processes disrupted by coastal protection measures, river-bank protection, and flood defences.

Mineral extraction can have positive and negative impacts. Quarries and gravel pits are a significant geological resource, particularly in areas where natural exposures are poor or scarce. Quarrying may reveal new sections of value, and many important sites are in former quarries, where the focus of the geological interest would not otherwise have been exposed. In some cases, quarrying may pose a direct threat to particular landforms; for example, limestone quarrying may destroy parts of cave systems and limestone pavements. In other cases, the loss of the surface landform and morphological integrity through quarrying, for example of an esker system, may need to be balanced against the potential value of new sections revealing the three-dimensional architecture of the constituent deposits.

Planning conditions usually require the restoration and landscaping of quarries, frequently involving landfill. The economic value of landfill space often competes against the value of conserving geological exposures. It is therefore crucial that potential geological interests are identified as far as possible at an early stage and incorporated into site restoration schemes. Road cuttings can also provide important geological exposures. Landscaping usually involves covering the cuttings with soil, seeding them with grass, and planting trees, leading to the loss of potentially valuable exposures.

A range of agricultural and forestry operations may affect geological sites and landforms. Landforms are often concealed beneath blanket commercial afforestation and damaged by extraction haul roads. Soils are under pressure from land-use practices and contamination, intensification of agriculture, afforestation, waste disposal, acid deposition, and urban expansion.

Traditional approaches to protecting coasts and rivers from erosion and flooding typically involve large-scale heavy engineering, which seals key exposures behind concrete seawalls, rock armour, or gabions. Natural processes of sediment supply and movement are disrupted, usually displacing the problem elsewhere.

There is a consensus that responsible fossil collecting can promote the science of palaeontology, providing that a code of good practice is followed. However, the irresponsible collecting of rare fossil and mineral specimens, often through the excavation of key sites for commercial gain, represents a significant loss to science as the context of the specimens is not recorded. Irresponsible collecting can also damage exposures and result in the loss of other specimens. Mechanical excavators, explosives, crowbars, and rock saws have all been used to remove fossil material, in a search for high-quality commercially saleable specimens.

Practical conservation therefore requires a combination of statutory protection and management for key sites and raised awareness of the value of geodiversity.

Conservation of Key Sites

Site Assessment

The identification and protection of key localities for research and education lies at the core of geological conservation. This approach is based on the selection of special or representative sites using scientific criteria and has been implemented in different ways in different countries through a variety of measures and instruments, including national parks, natural monuments, and other categories of protected site. It is particularly well illustrated by the British system of

national assessment, documentation, and protection of geological and geomorphological sites. Historically, this type of approach dates back to the mid-nineteenth century, early examples being the enclosure of the stumps of a former forest of Carboniferous lycopods at Fossil Grove in Glasgow in 1887 and the listing of erratic boulders in Scotland in the 1870s. The formal identification of key sites began in the 1940s with the compilation of a series of site lists, which were then added to in an *ad hoc* way. This process was superseded by the Geological Conservation Review (GCR), a major programme of systematic assessment of the conservation value of geological and geomorphological sites throughout Great Britain. Site assessment was undertaken between 1977 and 1990 and was the most comprehensive review of sites in any country. It was designed to reflect the full diversity of Earth heritage in Great Britain, spanning all the major time periods from the Precambrian to the Quaternary. Publication of the results in a series of 42 scientific volumes is now nearing completion. These describe the interests of individual sites and provide the scientific justification for their selection. Over 3000 individual localities were identified and form the basis of a network of Sites of Special Scientific Interest. These are accorded a measure of legal protection, which includes a requirement for consultation with the statutory conservation agencies over developments requiring planning consent and other notifiable activities.

The aim of the GCR was to identify sites of national and international geoscientific importance in Great Britain, based on a set of site-selection criteria and guidelines and extensive consultations within the geoscience community. Site selection is based on the concept of networks of sites representing the main features and spatial variations of geological events and processes during the main time periods. Three categories of site have been identified: sites of international importance, exceptional features, and representative features.

Many sites are of fundamental importance as international reference sites (stratotypes, type localities for biozones and chronozones, and type localities for rock types, minerals, or fossils), providing the building blocks for stratigraphy and the essential reference standards for the global correlation of rocks (e.g. Dob's Lin in the Scottish Borders is the boundary stratotype between the Ordovician and the Silurian). From a historical perspective, many sites are also internationally important classic localities in the development of geoscience, where features were first recognized or key concepts developed. In Scotland, the Northwest Highlands, Glen Coe, the island of Rhum, and Siccar Point have all provided crucial evidence for interpreting geological processes of global significance – respectively, the Moine Thrust, cauldron subsidence, magmatic processes and the origins of layering in igneous rocks, and a classic unconformity that provided the crucial evidence on which James Hutton developed the foundations of modern geology.

Some sites demonstrate unique or exceptional features (for example, the Rhynie chert in Scotland contains some of the oldest known fossils of plants and insects), while other sites demonstrate classic landforms or textbook examples of particular features, such as the Parallel Roads of Glen Roy and Chesil Beach. Many more sites contain nationally important representative examples of particular geological processes, environments, or events, which are essential for teaching and demonstration purposes and fundamental to understanding the geological history of Great Britain. Sedimentary rocks provide a valuable record of past environmental changes, and many contain valuable fossil remains that have helped to elucidate patterns of the evolution of life on Earth.

At a local level, geological conservation is pursued through the voluntary sector and the Regionally Important Geological/Geomorphological Sites (RIGS) movement. Sites of local importance are selected on the basis of their scientific and educational importance, historic interest, and aesthetic and cultural values, reflecting local rather than national values. Although these sites do not have statutory protection, many local authorities now have conservation policies for RIGS as well as other local wildlife sites. An important recent initiative has been the preparation of Local Geodiversity Action Plans in some areas. These should help to ensure greater protection for geodiversity as well as encouraging local awareness and involvement.

At an international level, many individual countries have compiled lists of geosites, particularly in Europe where there is a strong lead from ProGEO, the European Association for the Conservation of the Geological Heritage. Work is also in progress to develop international lists of sites under the auspices of the International Union of Geological Sciences, including a European initiative by ProGEO. In North America and Australasia, many geological and geomorphological features are protected by a variety of existing designations, including national and state parks, National Natural Landmarks, and provincial parks and nature reserves, although, with a few exceptions (e.g. Ontario, Tasmania, New Zealand), geological features have not been systematically assessed. A number of World Heritage Sites are so designated because of their geological features or

have significant geodiversity interests, but the list is not comprehensive or representative. Moreover, there are no international conventions or regional instruments for geodiversity comparable to those for biodiversity (e.g. the Convention on Biological Diversity or the EU Habitats Directive).

Site Management

Conservation is not only about the selection of sites but also about their safeguard. This requires the development of clear management objectives and periodic monitoring. In conservation management, there is an important distinction between 'integrity' and 'exposure' sites. Integrity sites include finite or relict features, which, if damaged or destroyed, cannot be reinstated or replaced because they are unique (e.g. fossil beds) or because the processes that created them are no longer active (e.g. Pleistocene glacial landforms). The prime management objective for integrity sites is to protect the resource. There are usually fewer options for compromise in reconciling conservation and development by employing practical management approaches such as the design of alternative solutions (e.g. excavation of a replacement exposure nearby). Conservation of active geomorphological sites depends on maintaining the freedom of the natural processes to operate across most or all of their natural range of variability.

Exposure sites are those in rock units or sediments that are spatially extensive below ground level, so that if one site or exposure is lost, another could potentially be excavated nearby. They include exposures in active and disused quarries, coastal and river cliffs, foreshore exposures, and natural rock outcrops inland. The principal management objective for such sites is to preserve exposure, but the precise location may not be crucial. Exposure sites are not usually damaged by quarrying or erosion, but are susceptible to landfill and other developments that obscure the sections.

In Great Britain, generic conservation management principles have been developed for different types of site, and site-condition monitoring is now a statutory requirement. Where natural exposures are poor, the appropriate restoration of disused quarries is crucial. However, the minerals industry now increasingly recognizes the importance of geological conservation afteruse. Through dialogue and partnerships between the statutory conservation agencies and the industry's trade organizations, good practice guidelines have been developed and practical conservation measures have been implemented for the retention of conservation sections and the provision of access to key geological exposures. Soft sediment exposures, however, continue to present long-term conservation problems, and, where the resource is finite, it may be necessary to keep the features buried and to excavate only for research purposes.

Practical conservation approaches and methods adopting low-impact less interventionist solutions have also been developed for river and coastal management. With more strategic assessment of environmental impacts, planning controls on floodplain and coastal developments, and the implementation of catchment plans and shoreline management plans, this should lead to outcomes that are more in harmony with natural processes.

Guidelines and codes of conduct for responsible collecting at fossil sites have been developed and applied. Commercial collectors are encouraged to work with specialists and museum curators to ensure that material is recorded and studied.

Sustainable Management of Natural Systems

In a wider context, the conservation of geodiversity is an integral part of sustainable management of natural systems. It is becoming clear that the management of sites for habitats and species cannot succeed without reference to the underlying geology, soils, and geomorphological processes. Understanding the links between geodiversity and biodiversity is crucial for conservation management in dynamic environments where natural processes (e.g. floods, erosion, and deposition) maintain habitat diversity and ecological functions. There is a strong case for more integrated approaches to conservation that would benefit both biodiversity and geodiversity, and this is now recognized by conservation agencies, notably in Great Britain, Tasmania, and Ontario, and also by bodies such as ProGEO.

The geological record clearly demonstrates how the Earth's natural systems have evolved in the past and how they might behave in the future. However, human activity is now an important 'geological force', reshaping the surface of the Earth through the movement of rock and soil, the building of cities, motorways, and dams, the fixing of the coast with concrete barriers, deforestation and soil erosion, and causing the extinctions of species. Not only are many resources being used at a greater rate than that at which they are being replenished, but also there are many uncertainties about the long-term environmental effects of the disposal of waste from human activities. Human activity is also having a potentially significant impact on global climate. A major challenge is to use an understanding of the Earth's processes to mitigate future human impacts, to contribute to the restoration of areas

already damaged by human activities, and to work with others to develop strategies for the sustainable use of the Earth's resources.

Geomorphological processes frequently impinge on human activity (e.g. through flooding, coastal erosion, and soil erosion), with resultant economic and social costs. Management responses often result in locally engineered solutions, such as riverbank and coastal protection measures, that are unsuccessful or simply transfer the problem elsewhere. Typically, management timeframes are based on human experience and are not sufficiently informed by the longer-term geological perspective. However, it is this perspective that is vital in assessing natural hazards and implementing sustainable management of natural resources.

Sustainable management of natural systems therefore depends on the effective application of earth science knowledge as part of the development of more integrated approaches; for example the maintenance of sediment transport at the coast or of natural flow regimes in rivers. Various guiding principles have been proposed.

- The inevitability of natural change should be recognized.
- Any management or intervention should work with, rather than against, the natural processes.
- Natural systems should be managed within the limits of their capacity to absorb change.
- The sensitivity of natural systems should be recognized, including the potential for irreversible changes occurring if limiting thresholds are crossed.
- Natural systems should be managed so as to maintain natural rates and magnitudes of change and their capacity to evolve through natural processes.
- Natural systems should be managed in a spatially integrated manner (e.g. at a catchment or coastal zone level).

Many of these principles are now being applied through a range of approaches or programmes in different countries, recognizing that landscapes are mosaics of geological, natural, and cultural features that need to be managed and interpreted in an integrated fashion. This is well exemplified by shoreline management plans and integrated river-catchment management and by the comprehensive ecosystem-based planning frameworks developed for the Oak Ridges Moraine and Niagara Escarpment in Ontario. The ecosystem approach, in particular, has been adopted as a primary framework for action under the Convention on Biological Diversity and provides a means for the closer integration of geodiversity and biodiversity on a wider scale. The importance of soil conservation is also gaining recognition. For example, in response to concerns about the degradation of soils in the European Union, the European Commission is developing a Thematic Strategy for Soil Protection that recognizes the value of the functions that soils perform and the need for soil protection to be integrated with other environmental policies.

Raising Awareness of Geodiversity

Raising wider awareness and involvement is a key part of geological conservation. At one level, as part of an integrated approach to management of the natural heritage, sustainable development, and landscape management, public bodies, industry, land and water managers, planners and policy makers, and their advisors should be aware of the value of geodiversity and its conservation, as well as having an understanding of natural processes, so that informed decisions can be made. At another level, there is a need to raise awareness of the value of geodiversity with the general public and in schools to help form the basis of a wider constituency of support for geological conservation. If communities value and take pride in their local geodiversity, they are more likely to support its stewardship and conservation, as well as contributing to public debate on the wider issues. Effective geological conservation will ultimately depend on better public awareness, understanding, and support.

Interpretation of geodiversity and geology-based tourism (geotourism) are not new, as demonstrated by the appeal of show caves, glaciers, and other natural wonders. Traditional geological interpretation, however, was often conveyed by interpretation boards using detailed and overly technical language, and providing information rather than interpretation. Recent developments have seen more effective communication, resulting in the production of more appropriate materials that are presented in stimulating ways using a range of media and based on best interpretive practices and sound educational principles. Such developments may involve more integrated landscape interpretation, linking geology, landscape, human activities, and industrial archaeology, for example. They include the designation of Geoparks and the promotion of geotourism (e.g. through the European Geoparks programme and the 'Landscapes from Stone' initiative in Ireland). As well as helping to raise awareness of Earth heritage, these activities have an economic dimension. Urban geology is also an important vehicle for raising public awareness through the exploration of the links between geology, use of building and paving stones, and architectural heritage.

Conclusion

Geodiversity is an integral part of our natural and cultural heritage that deserves to be better known and conserved for the benefit of future generations. However, in comparison with the conservation of biodiversity, the conservation of geodiversity has received much lower priority in national and international conservation programmes. This reflects, in part, a traditional focus on site-based protection for geological research and, in part, a lack of awareness at all levels of the wider significance and value of geodiversity. This has meant that geological conservation has lagged behind developments in mainstream conservation. However, there is now growing recognition of the functional links between geodiversity and biodiversity, and there are calls for more unified approaches, which will hopefully see better integration of geodiversity into conservation policy, protected area management, and sustainable management of natural systems. A key priority is to raise awareness among decision makers, policy makers and their advisors, planners, land and water managers, and the education sector, and at the same time to stimulate greater public interest and involvement.

See Also

Building Stone. **Environmental Geology**. **Famous Geologists:** Hutton; Lyell. **Quarrying**. **Soils:** Modern. **Urban Geology**.

Further Reading

Barettino D, Vallejo M, and Gallego E (1999) *Towards the Balanced Management and Conservation of the Geological Heritage in the New Millenium*. Madrid: Sociedad Geológica de España.

Bennett MR, Doyle P, Larwood JG, and Prosser CD (1996) *Geology on Your Doorstep. The Role of Urban Geology in Earth Heritage Conservation*. Bath: The Geological Society.

Commission of the European Communities (2002) *Towards a Thematic Strategy for Soil Protection*. COM(2002) 179 final. Brussels: Commission of the European Communities.

Doyle P and Bennett MR (1998) Earth heritage conservation: past, present and future agendas. In: Bennett MR and Doyle P (eds.) *Issues in Environmental Geology: a British Perspective*, pp. 41 67. Bath: The Geological Society.

Ellis NV, Bowen DQ, Campbell S, *et al.* (1996) *An Introduction to the Geological Conservation Review*. Peterborough: Joint Nature Conservation Committee.

Glasser NF (2001) Conservation and management of the Earth heritage resource in Great Britain. *Journal of Environmental Planning and Management* 44: 889 906.

Gordon JE and Leys KF (2001) *Earth Science and the Natural Heritage. Interactions and Integrated Management*. Edinburgh: The Stationery Office.

Gray JM (2003) *Geodiversity. Valuing and Conserving Abiotic Nature*. Chichester: Wiley.

Hooke J (1998) *Coastal Defence and Earth Science Conservation*. Bath: The Geological Society.

Johansson CE (2000) *Geodiversitet i Nordisk Naturvård*. Copenhagen: Nordisk Ministerråd.

O'Halloran D, Green C, Harley M, Stanley M, and Knill J (1994) *Geological and Landscape Conservation*. Bath: The Geological Society.

Parkes M (2004) *Natural and Cultural Landscapes The Geological Foundation*. Dublin: Royal Irish Academy.

Sharples C (2002) Concepts and principles of geoconservation. Published on the Tasmanian Parks and Wildlife Service website at http://www.dpiwe.tas.gov.au/inter.nsf/WebPages/SJON 57W4FD?open

Stevens C, Gordon JE, Green CP, and Macklin MG (1994) *Conserving Our Landscape. Evolving Landforms and Ice Age Heritage*. Peterborough: English Nature.

Taylor AG, Gordon JE, and Usher MB (1996) *Soils, Sustainability and the Natural Heritage*. Edinburgh: HMSO.

Wilson RCL (1994) *Earth Heritage Conservation*. London and Milton Keynes: The Geological Society and the Open University.

GEOLOGICAL ENGINEERING

A K Turner, Colorado School of Mines, Colorado, USA

Introduction

This section explores the character of 'geological engineering' in the context of 'engineering geology'; the two distinct, but closely related professional fields. It also considers the relationships between geological engineering and several other specializations, including geotechnical engineering (*see* **Geotechnical Engineering**), ground engineering, environmental geology (*see* **Environmental Geology**), and hydrogeology (*see* **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills). The development of these

various specialized fields reflects the complexity of modern engineering design and construction, especially those designs involving the interface between naturally occurring earth materials and the engineered structure, or the use of naturally occurring materials within the constructed facility.

Geological engineering is primarily a reflection of legal and technological conditions within the USA. Technological developments in Canada, western Europe, and elsewhere generate very similar demands for individuals with appropriate technical skills, but without the same pressure to also meet professional engineering registration standards. A brief historical review of the relationships between engineers and geologists over the past 200 years provides some insight into the current situation surrounding the accepted professional stature and roles for geologists and engineers.

Individual practitioners are increasingly likely to become involved in litigation, professional liability has become an important concern for many professions in many countries, and geological engineers and engineering geologists are not immune from this condition. These concerns have led to increased professional registration options for both geologists and engineers, although the exact methods of achieving this vary from country to country.

Geological engineering has developed as a relatively small and unique specialization within the broader engineering profession. The skills of a geological engineer are becoming more desirable than ever as the technologies involved in construction continue to evolve. So the future of geological engineering would appear bright, except for the financial pressures faced by many universities, which are restricting the growth of 'high-cost' fields that attract relatively low student numbers. These trends suggest that it may be difficult for adequate numbers of geological engineers to be trained in order to maintain a viable cadre of professionals.

What is in a Name?

At first glance the terms 'geological engineer' and 'engineering geologist' appear synonymous. Because the two terms employ essentially the same two words, 'geology' and 'engineering'; although in reverse order, the opportunity for confusion is great. The word choices may be unfortunate, but the two terms represent distinct, although related, concepts concerning educational and professional endeavours.

The following attempts to differentiate a 'geological engineer' from an 'engineering geologist'. Before exploring the details, the following points should be understood:

- The term 'geological engineer' was developed in the USA in response to a combination of technical opportunities and the established legal processes for obtaining professional engineering registration in the USA.
- The initial demands for geological engineers came from the minerals and petroleum industries; the demand for significant numbers of specialists to work with civil engineers (in engineering geology) only developed in the latter half of the twentieth century.
- The term 'geological engineer' is thus most widely used in the USA. The term has been used only to a limited extent in Canada, which has a different professional registration structure, and only to a very limited extent and very recently in the UK, western Europe, and elsewhere.
- The term 'engineering geology' is widely used throughout the world in two contexts: to describe the application of geological principles relevant to engineering works, environmental concerns, and societal concerns, and to define specialist geologists ('engineering geologists') who are involved in such studies.
- In contrast to the 'geological engineer', who is trained as an engineer with additional geological knowledge, the 'engineering geologist' remains a scientist. This difference has ramifications for professional licensure and legal authority.
- Because engineers and scientists may be equally held liable for public safety and welfare issues; issues of certification, licensure, or registration increasingly affect the field of engineering geology, and the geological engineers, engineering geologists, and others involved in major engineering works. These issues are resolved in many forms in different parts of the world.

Defining 'Geological Engineering'

A geological engineer is trained as an engineer, but an engineer with a broad understanding of applied geological science. The concept of geological engineering originated in the USA in response to demands by North American industries for individuals who could apply both geological and engineering principles to the evaluation and design of projects involving earth materials, structures, and forces.

The legal circumstances in the USA related to professional practice pertaining to evaluation, design, approval, and operation of projects – in particular, the requirements enforced by official engineering registration boards in each state – demanded that these individuals have the academic credentials suitable to meet professional engineer (PE) registration criteria.

These demands developed in the early twentieth century as advances in both technology and engineering made larger and more complex engineering works feasible. New branches of engineering, such as petroleum engineering, developed in response to technological advances. Existing engineering disciplines, such as civil engineering, increased demands for new specialties to design ever larger and more complex dams, tunnels, and transportation systems. In particular, the minerals and petroleum industries required increasing numbers of exploration and production specialists and administrators, roles for which engineering training combined with geological knowledge were the basic requirements.

Consequently, a number of universities and mining schools in the western USA began to offer engineering programmes leading to a degree in 'geological engineering'. Graduates from these programmes were hired by petroleum and minerals exploration and production companies, and placed in positions where their combined geological and engineering training made them uniquely qualified. Subsequently, as professional engineering registration procedures became codified, these geological engineering programmes became accredited, allowing their graduates to later achieve the status of PE. Geological engineers that obtained PE status could legally approve designs for engineering works, an important consideration in some situations.

The term 'geological engineer' developed in the USA in response to both technological demands and to legal professional engineering registration procedures. Although the Canadian economy was even more dependent on the petroleum and mineral extraction industries, and thus experienced a similar demand for individuals with combined geological and engineering knowledge, the 'geological engineering' term was not adopted. Rather, in response to the demand for qualified graduates, many geology departments at Canadian universities either partially or entirely joined the faculties of applied science (in other words: engineering). This allowed many, if not all, of their graduates to achieve registration as professional engineers, and many individuals did so. The issues of professional engineering registration and legal liability issues are discussed more fully in subsequent sections.

The reader should note that most of the early geological engineers did not work on civil engineering projects. They were more likely to work on minerals exploration and exploitation projects with mining engineers, or on petroleum exploration and production projects with petroleum engineers. The employment situation was similar in both Canada and the USA. In the latter half of the twentieth century, major civil engineering projects following World War II placed new demands for specialists to work with civil engineers. In response to this new demand, many of the geological engineering academic programmes in the USA, as well as the Canadian geological programmes, began to provide 'options', usually three, with titles such as 'Petroleum Exploration', 'Mineral Exploration' and 'Engineering Geology'.

The majority of recent geological engineering graduates are now employed in practices focused on engineering geology (civil engineering applications), and ground water (hydrogeology) projects. Graduates may continue to specialize with more advanced degrees in such areas as geotechnical engineering, rock mechanics (*see* **Rock Mechanics**) (for tunnelling and underground construction), hydrogeology, contaminant transport (to evaluate ground pollution issues), or various geohazard mitigation studies (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards) (landslides, earthquakes, or floods).

In all cases, the geological engineering programs in the USA retained their accreditation as engineering programmes, and their graduates, along with their Canadian counterparts, therefore, continued to have the right to achieve professional engineering registration.

Defining 'Engineering Geology'

In contrast to the geological engineer, the engineering geologist remains a scientist, albeit a rather applied one. Engineering geology uses geology to create more efficient and effective engineering works, to assess and allay environmental concerns, and to promote public health, safety, and welfare. The concept of geologists advising on engineering works dates from the earliest period of the science of geology, the historical evolution of relationships between civil engineering, and geology is discussed briefly in a subsequent section.

The term 'engineering geology' became widely accepted only as the previously noted demand for geological specialists to advise civil engineers developed in the latter half of the twentieth century. While the geological engineering programmes adjusted their focus to accommodate such demands, at the same time other universities in the USA and elsewhere began to offer an 'engineering geology' option within their science-oriented geological degree programmes. In general, the graduates from these science-oriented programmes cannot easily achieve professional engineering registration, although they often become members of multi-disciplinary teams undertaking a variety of construction projects and environmental evaluations.

The engineering geologist applies geological knowledge and investigative techniques to provide

quantitative geological information and recommendations to engineers for use during design and construction of engineering works, and in related professional engineering practice. Through cooperation, the engineering geologist and the civil engineer share the responsibility for ensuring the public health, safety, and welfare associated with geological factors that may affect or influence engineering works. In most cases, the public demands that the professional engineers be held responsible for the safety and integrity of their works. Thus, the engineering geologist may be considered as a specialist advisor to the design team, and may hold a position similar to an architect or other design specialist.

In this role, the engineering geologist often cannot provide legally binding approval of a design for an engineering project, many laws require a PE to make such judgments. In the USA, only a few individuals calling themselves 'engineering geologists' hold PE registration credentials; such persons are more likely to use the designation 'geological engineer'. However, there is a growing requirement for establishing a separate professional registration of geologists, especially for those individuals undertaking engineering geological investigations. In Canada, the situation is slightly different, with several Provinces having a single entity that supervises the professional registration of both engineers and geoscientists. The issues of professional registration and legal liability issues, for both geological engineers and engineering geologists, are discussed later.

Regardless of the existence of professional registration requirements, all engineering geologists carry a serious responsibility when applying specialist geological training and experience in communications with engineers. They must be aware of the requirements of the engineers during the design, construction, or operation and maintenance of facilities. They may be called upon to provide judgmental recommendations as well as quantitative data. The potential for misunderstandings must always be recognized; and a very high standard of written, graphical, and oral reporting is crucial to a successful engineering geologist (and geological engineer!).

In recent years the scope of engineering geology practice has expanded beyond its original closely defined connection with civil engineering. Many engineering geologists currently work closely with land-use planners, water resource specialists, environmental specialists, architects, public policy makers, and property-owners, both public and private, to prepare plans and specifications for a variety of projects that are influenced by geological factors, involve environmental modifications, or require mitigation of existing or potential effects on the environment.

Relationship of Geological Engineering to Associated Fields

The environmental movement has impacted on both geology and engineering, and on how these disciplines relate to the demands of society. In many western countries, resource extraction has become viewed with suspicion. At the same time, environmental concerns and demands for new and renovated infrastructure to support increasingly large urban populations (transportation, community expansion, water supplies, and waste disposal) have led to demands for new thematic products. National geological surveys continue to grapple with policies to define an appropriate response to these demands. Many now host important 'engineering geology' sections.

One result of the increased environmental awareness has been the development of a subject called 'environmental geology', which most engineering geologists consider to be largely within their field of expertise. Entirely new types of technical reports have been developed, aimed at the non-specialist and a broader public audience. Another aspect of the increased environmental awareness is the demand for assurances on safe and clean water supplies on the one hand, and their protection through the careful disposal of wastes on the other. Water is becoming the dominant factor for development throughout the world as populations increase and demands are placed on diminishing supplies. As a consequence, many geological engineers and engineering geologists are specializing in water-related topics, commonly considered the realm of 'Hydrogeologists'.

In response to the need for environmental protection on the one hand, and technological advances and economic forces on the other, engineering projects have become much more complex; bridges and tunnels have become longer, and high-speed transportation links have become common. Population growth has pushed developments into more complex geological locations where site conditions are less than optimal and geohazards more likely. These trends have led to an increase in the need for engineering specializations. 'Geotechnical engineering' (*see* **Geotechnical Engineering**) is a specialty that deals with the solution of civil, environmental, and mining engineering problems related to the interaction of engineering structures with the ground. Geotechnical engineers typically have expertise in soil mechanics and rock mechanics, and relatively little knowledge of geological science. They are predominantly civil engineers and are trained to design structures for foundations in soil or rock. For some projects their training limits their ability to account for the natural heterogeneity or complexity of naturally occurring

geological features. Under such circumstances, a consultation with engineering geologists or geological engineers is desirable. In many countries the term 'ground engineering' is used whenever soil mechanics or rock mechanics principles are employed in actions that modify the properties of naturally occurring materials. Usually these actions are directed toward making the materials stronger (i.e. capable of supporting larger structural loads) or reducing the permeability (e.g. to reduce the inflow of water into an excavation).

These disciplines (geological engineering, engineering geology, environmental geology, geotechnical engineering, and ground engineering) are frequently involved with construction sites, especially those for large and highly visible projects, and within or near large urban centres. Such applications lead to the use of two additional terms related to these typical locations: 'construction geology' and 'urban geology' (*see* **Urban Geology**).

In summary, geological engineering refers to a particular style of engineering that is predominantly engineering by training and experience, but utilizes special additional knowledge of geology. Geological engineers can perform a variety of tasks in the resource exploration and production fields, but in recent times most individual geological engineers are employed on civil engineering and environmental projects. They frequently work closely with engineering geologists, hydrogeologists, and geotechnical engineers, and there is in fact a continuum in training, experience, and background among individuals following these career paths. All such individuals may be employed in environmental geology, hydrogeology, or ground engineering projects and may further specialize in particular types of project that may be referred to as 'construction geology' or 'urban geology'.

Historical Interactions Between Civil Engineering and Geology

Geologists and engineers have interacted with varying degrees of support and antagonism for over 200 years. The industrial revolution demanded the transport of large quantities of heavy goods, first by canals and then by railways. When constructing civil works by large gangs of men using little more than picks, shovels, and wheelbarrows, the engineers had a considerable interest in exactly what they would be excavating, and how difficult the work might be.

William Smith (1769–1839), during his surveys for and construction of the Somerset Coal Canal near Bath in southwestern England, noted the regular succession of strata, and was able to correlate them with those in other locations by the use of fossils (*see* **Famous Geologists:** Smith). In 1799 Smith coloured his geological observations on a map of the Bath area: the oldest geological map in existence! In the same year Smith wrote a document, *Table of Strata near Bath*, and for this he became known as 'The Father of English Geology', although he continued to refer to himself as a civil engineer. In 1801, and subsequently in 1815, Smith produced further geological maps of England and Wales. He also conducted numerous civil engineering and geological investigations, and continued to assert the importance of geology to engineering.

Alexandre Collin, a French engineer responsible for the construction of several canals, conducted extensive field surveys to determine the characteristics of slope failures of cuttings and embankments. His 1846 treatise on the stability of clay slopes recognized that the characteristic circular failure surfaces were the result, not the cause, of the landslide movements. Even more importantly, he undertook laboratory experiments to determine the shear strength of these materials. He basically invented the subject of soil mechanics long before it became a popular civil engineering discipline. Unfortunately, his writings were poorly distributed and were largely ignored until about 100 years later.

The railways also required considerable earthworks and encouraged an ongoing close cooperation, and friendly relationship, between civil engineers and geologists during the latter half of the nineteenth century. The minutes of many meetings and the subjects of public lectures provide ample evidence of this collaboration – as, for example, the following quotation from an 1841 meeting of the Institution of Civil Engineers in London: "Mr. Sopwith called the attention of the meeting to the valuable Geological Sections presented by the railway cuttings . . . the crops of the various seams of coal, with the interposing strata, were displayed in the clearest manner, developing the geological structure of the country which the railway traverses."

Beginning in the 1890s, the former close association of geology (and geologists) and civil engineering (and civil engineers) broke down. The introduction of powered machinery began to change the perceptions of many engineers, any job became feasible and 'successes', such as the completion of the Panama Canal by American engineers using much more powerful equipment (and also superior medical knowledge concerning tropical diseases) after the earlier failure of the French, merely served to increase this 'can do' attitude. A minority of engineers continued to strongly recommend that their colleagues seek geological advice. Chief among them in this period was Karl Terzaghi, 'The Father of Soil Mechanics' who strongly supported the linkage between of geology and engineering.

The attitudes of geologists during this period also contributed to the breakdown in the previously close working relationships. Increasingly, complex terminology was applied throughout the science, often based on Greek or Latin vocabularies. Engineers did not take kindly to this new language of the geologist. Contributions by geologists were increasingly considered to be irrelevant to the engineering design and construction processes. This separation of disciplines, and breakdown in communication and mutual respect, was most unfortunate, for the first half of the twentieth century witnessed the construction of major public works in many parts of the world. Yet, only the largest projects called for any significant geological consultation. Some spectacular failures were the result.

The introduction, debate about, and ultimate acceptance of the plate tectonics concept in the late 1960s completely revolutionized geology, providing a coherent underlying theory for evaluating descriptive geological observations. Engineering geological studies were able to make better predictions of subsurface conditions, especially in complex geological situations and on a regional basis. At about the same time the availability of powerful computers caused a complete change in engineering design procedures. No longer were dams and other structures designed using slide rules. Today modelling and optimization are required. These, in turn, place new demands for accurately predicting how geological materials will interact with the new engineered structure. At the same time, computers provide geologists with numerical analysis tools, including the ability to create and evaluate complex 3-D models of the subsurface.

As a consequence, in recent decades, the 'estrangement' of geologists and engineers has largely dissipated and the field of engineering geology, and the employment of geological engineers and engineering geologists on many projects, has begun to expand throughout the world.

Professional Liability Concerns

Liability is the legal responsibility for any loss or damage from ones actions, performance, or statements. Malpractice refers to improper, negligent, or unethical conduct or practice that results in damage or injury. Negligence claims are thus basically claims of malpractice. Court decisions are usually based on a legal responsibility for individuals to practice according to 'state of the art', 'best practice', or 'standard of practice' criteria.

Prior to the twentieth century, geologists were infrequently involved or concerned with legal matters. An exception was the involvement in 1839 of James Hall (*see* **Famous Geologists:** Hall) of the New York Geological Survey in assessing the conditions encountered during excavations to enlarge the Eire Canal locks at Lockport, New York. The contract specified a unit price for 'solid rock' and a lower price for 'slate rock and shale'; these classifications were subject to dispute. Hall was asked to provide expert testimony.

Today, geologists and geological engineers are increasingly likely to be involved in some aspect of litigation, as members of a team or as individuals, often as expert witnesses serving a client who is either a plaintiff or defendant. The probability of a geologist being sued for negligence or malpractice as an individual varies enormously depending on the nature of the work.

Geologists and geological engineers advising or employed by geological or geotechnical engineering consulting firms are most vulnerable to charges of malpractice or negligence. These individuals must abide by requirements for certification, registration, or licensing before working in any location. Those not meeting such legal requirements may be found guilty of breaking the law, may find their work defined as unqualified or unacceptable, and thereby be subjected to malpractice litigation.

There are differing legal opinions concerning responsibility whenever geological problems arise. In large firms, many junior geologists and engineers work under a principal who is registered, and in such cases the principal or the firm is responsible for the actions of the staff members. Individuals working on smaller projects sometimes seek to gain some protection from their clients by having them agree, by contract, to specifically indemnify them for consequences of their consulting services. Such actions are not always successful. Since geologists and geological engineers are increasingly involved in nearly all aspects of construction on or within the earth, or with earth materials, their degree of liability is growing.

Professional Registration and Certification Issues

The topic of professional registration and certification is complex, and is undergoing fairly rapid evolution in both North America and Europe. The pattern and procedures of registration of both engineers and geologists is quite different in the USA, Canada, and Europe. The following sections briefly consider them.

Geological Engineering Professional Registration in USA

Legal responsibility for the professional registration of engineers of all disciplines is delegated to 'Professional

Engineers Registration Boards' in each State. Although requirements and regulations do vary by state, most state boards require applications for 'registration' as a PE, to have: (1) obtained a university education from an engineering program accredited by the Accreditation Board for Engineering and Technology (ABET), (2) passed a 'Fundamentals of Engineering (FE)' exam, and (3) after several of years of experience, to have passed a PE exam. ABET has representatives from all engineering disciplines; it 'accredits' (that is reviews and approves) engineering programmes at universities and supervises the administration of the FE and PE exams. The Society of Mining, Metallurgy and Exploration (SME) represents the interests of geological engineering within the ABET organization.

State laws govern most environmental and hazard investigations, and many state laws require design documents for such projects to be signed by a PE. This requirement exists in spite of the fact that some projects involve geological aspects that an engineer is not necessarily competent to evaluate! The situation also means that a fully qualified and registered geological engineer can carry legal responsibilities that their engineering geologist brethren cannot. The laws reflect a reality; that the development and acceptance of an equivalent process to provide 'professional geologist' registration is occurring slowly and sporadically. However, some states, California being the earliest and perhaps the best example, do have considerable legal requirements for the registration of geologists.

Competing Approaches to Geologist Registration in the USA

Professional registration of geologists within the USA has been debated for about 20 years. Geologists employed in petroleum and mineral exploration have generally been opposed to calls for registration, while geologists involved in engineering, hydrogeology, and environmental projects, where public health and safety issues are readily apparent, have generally favoured registration efforts. State-by-State registration of geologists, following the engineering 'ABET' model appears to be the generally accepted method. California was the first state to legislate registration of geologists. Currently, 26 out of the 50 states require registration of geologists. State boards of registration, independent of the engineering boards, supervise the registration procedures in their state, and these boards cooperate through the National Association of State Boards of Geology (ASBOG).

ASBOG supervises the development and scheduling of two examinations: a 'Fundamentals of Geology (FG)' exam and a 'Practice of Geology (PG)' exam, which are administered by the state boards. The distribution of ASBOG examination questions reflects the importance of tasks performed by engineering geologists/hydrogeologists. This may be the natural result of the importance of engineering geology to ensuring public safety, this being the primary legal justification for professional registration.

A disturbing result of this examination process has been a very low pass rate for these exams, over the past decade only about 57% of candidates passed the FG exam and only about 68% of the candidates passed the PG exam. The percent passing has not materially changed from year to year over this period. These results suggest that many university graduates do not have an adequate grasp of the necessary geological skills. This has been used as an argument in support of registration of geologists in other states that currently do not require registration.

The American Institute of Professional Geologists (AIPG) also provides a 'certification' of geologists, giving those who are approved access to the title 'Certified Professional Geologist' (CPG). This certification is conducted by peer review of credentials without any examination. It has no legal standing in those states requiring registration, but does provide individuals with some 'national' credentials that may assist them when providing expert testimony and in similar situations.

Professional Registration Approaches in Europe

In Europe, the idea of 'professional registration' is quite different to North America. The professions generally are more self-regulating and professional credentials are tied to membership of professional societies within each country, and these in turn require completion of university degrees. In the UK, for example, civil engineers may become 'Members' of the Institution of Civil Engineers, thereby gaining the ability to use the designations 'MICE'. As a result, they qualify for registration with the Engineering Council as a Chartered Engineer (CEng).

With the creation of the European Union, Europe-wide credentials have evolved. European engineers can obtain approval to use the title 'European Engineer' and the designation 'Eur.Ing' from the European Federation of National Engineering Associations (FEANI), while in a similar fashion European geologists can obtain approval to use the designation 'Eur.Geol.' from the European Federation of Geologists (EFG). However, at the present time, whereas the designation 'Eur.Ing.' provides some specific legal standing, the 'Eur.Geol.' designation does not. Many European practitioners in the field of engineering

geology advertise their membership of the International Association of Engineering Geology and Environment (IAEG) as a credential showing competence, although it does not carry legal weight.

Professional Registration in Canada

Canadian registration procedures follow a path that lies between American and European registration practice. European procedures are largely related to membership of national professional societies. Canadian provinces have enacted registration laws in a similar manner to the American state legislatures. However, these laws generally designate appropriate provincial professional associations or societies as having the power of 'self-governance'. In the majority of provinces, a joint association supervises the registration of engineers and geoscientists. However, Ontario and Quebec have separate registration procedures; in the case of Ontario this developed when the creation of a joint registration procedure was stopped by the action of a group of engineers. In spite of some occasional evidence of friction between geologists and engineers, interactions are generally good and some individuals hold dual registration.

Conclusions Concerning Professional Registration

Several competing registration approaches have developed in the USA, Canada, and Europe. In North America, the professional registration of engineers has been legislated at the state/provincial level since the early twentieth century and has been accepted as necessary to protect the public interest. The case for an equivalent registration of geologists has not been so clearly made, and in fact there has been considerable opposition to such registration by many geologists.

In the USA, procedures to register engineers and geologists are administered quite independently by distinct official boards of registration. Whereas all states have engineering boards, only about one-half the states have geology boards. In Canada, the provincial legislatures delegate the registration process to professional associations, and in the majority of the provinces a single association supervises the registration of both engineers and geologists.

In the light of these developments, it is perhaps not surprising that the concept of 'geological engineering' should have arisen first in the USA, allowing engineers with specific geological knowledge and skills to become registered as engineers, while in Canada those geologists desiring registration and having the requisite skills and experience could obtain registration as geologists, engineers, or both.

The concept of professional registration is evolving in Europe. Once again, geologists are tending to lag behind engineers in embracing the need for professional registration.

International trends, especially the increased globalization of markets for consultation services as well as goods, have placed new pressures on the existing professional registration procedures. The requirements to have multiple registrations in several states or provinces in North America in order to undertake projects at several locations impose time and cost constraints on individual engineers and geologists, and their employers. Only limited reciprocal arrangements exist between Canada and the USA, in spite of the regulations embodied in the North American Free Trade Agreement (NAFTA). Similar trends within the European Union have led to new developments that promote European designations.

The concept of professional registration for geologists is still relatively young. Major constraints are the lack of public acceptance of the need for registration, the lack of 'official' legal standing, the objections of many geologists who see registration as restricting their mobility and freedom to conduct studies, objections by other professions, and competition among professional societies for authority to provide and supervise such registrations.

A Look to the Future

Geological engineering appears to be at a crossroads. Demand for geological engineering expertise to solve society's needs and desires for a better living environment points toward a bright future. Certainly, new and ever more challenging environmental issues will make the design and construction of new transportation and other facilities depend even more on an accurate prediction of geologic conditions. The increasingly sophisticated designs depend for their success on the involvement and acceptance of the geological engineer and engineering geologist.

Yet the entire capacity to educate and train fresh geological engineers is quite limited. The majority of the academic programmes in the USA have relatively small recruitment. The economic pressures facing many universities encourage the elimination of smaller 'specialist' or 'elitist' and high-cost programmes and departments. The establishment of geological engineering educational programmes beyond the USA has been, and continues to be limited because the same combination of technological and professional registration procedures that encouraged the establishment of geological engineering in the USA does not occur.

In western Europe and North America, the enrollment of students in engineering and science, especially geoscience, has been falling for several years. Many talented students are not selecting such 'tough' courses of study demanded by engineering and science fields. Topical areas perceived as narrow specialties, such as geological engineering, are apparently at a further disadvantage when attracting new recruits.

Geological engineering does not have the advocacy within the larger established professional societies to ensure its growth or even survival as a designated independent engineering specialization. Even the wider field of engineering geology practitioners, encompassing both geological engineers and engineering geologists, is facing a similar identity crisis. This is occurring in spite of expanding employment opportunities and the recognition of the need for such specialists by potential employers.

See Also

Engineering Geology: Codes of Practice; Natural and Anthropogenic Geohazards; Ground Water Monitoring at Solid Waste Landfills. **Environmental Geology**. **Famous Geologists:** Hall; Smith. **Geology, The Profession**. **Geotechnical Engineering**. **Rock Mechanics**. **Urban Geology**.

Further Reading

Kiersch GA (ed.) (1991) *The Heritage of Engineering Geology; The First Hundred Years*, Vol. 3, *Centennial Special*, p. 605. Boulder, Colorado, USA: Geological Society of America.

Legget RF (1962) *Geology and Engineering*, 2nd edn., p. 857. New York: McGraw Hill.

Legget RF (1973) *Cities and Geology*, p. 578. New York: McGraw Hill.

Paige S (ed.) (1950) *Application of Geology to Engineering Practice*. Berkey Vol., p. 327. Boulder, Colorado, USA: Geological Society of America.

Accreditation Board for Engineering and Technology, Inc.(ABET) provides, operates and maintains an independent and objective accreditation system for applied science, computing, engineering, and technology education. http://www.abet.org/.

National Association of State Boards of Geology (ASBOG) serves as a connective link among the individual state geologic registration licensing boards. http://www.asbog.org/.

National Council of Examiners for Engineering and Surveying (NCEES) a national non profit organization composed of engineering and land surveying licensing boards representing all U.S. states and territories. http://www.ncees.org/.

Canadian Council of Professional Engineers (CCPE) provides links to all provincial/territorial engineering associations. http://www.ccpe.ca/.

Canadian Council of Professional Geoscientists (CCPG) provides links to all provincial/territorial professional associations. http://www.ccpg.ca/.

Association of Engineering Geologists (AEG). http://www.aegweb.org/.

Geological Society of America (GSA). http://www.geosociety.org/.

International and European Societies.

International Association for Engineering Geology and the Environment (IAEG). http://www cgi.ensmp.fr:88/.

Geological Society of London (GSL). http://www.geolsoc.org.uk/.

Institution of Civil Engineers (ICE). http://www.ice.org.uk/.

European Federation of Geologists (EFG). http://www.eurogeologists.de/.

European Federation of National Engineering Associations (FEANI). http://www.feani.org/.

GEOLOGICAL FIELD MAPPING

P Garrard, Imperial College London, London, UK

Introduction

Geological field mapping provides the fundamental scientific basis for most geological maps. It can range from large area reconnaissance to detailed mapping of areas only a few metres across, and sometimes has a specific aim such as mineral exploration. All-purpose, 'survey-type' mapping involves a geologist in the field examining rocks in their natural location, plotting data onto a base map, and recording details in a field notebook. Progressively, the field map displays the distribution of rocks and superficial deposits at the Earth's surface, their orientations, and the nature of the contacts between them. Fossils and rock samples may be collected for laboratory investigation. Analysis of structures indicates the tectonic development. From such information, it is possible to draw sections in any direction across the area, predicting geological relationships at depth, and to compile a report detailing the geological history. Other types of investigation, e.g., geophysical, geochemical, and remote sensing, can provide supplementary information.

National Geological Surveys publish district geological maps at scales of 1 : 50 000 or 1 : 25 000, based on field mapping typically performed at larger scale, for instance 1 : 10 000 in the case of the British Geological Survey (BGS). District maps hold a large amount of detail from routine fieldwork, plus data from boreholes, water wells, underground workings, and similar. Compilations give rise to regional and national maps, and the maps are also used by professionals requiring information on soils, economic deposits, land use, water supply, and hazard potential. Hence the field geologist needs to record all aspects of the geology.

Two points require emphasis. First, field investigations must be rigorous, making careful, comprehensive examination of the geology and recording the results accurately at the correct location. Inadequate or wrongly positioned data will result in incomplete or incorrect deductions. Second, a geological field map is partly interpretative. It contains 'factual' data, studied and measured in the field, but incomplete exposure means that the nature, position, or even existence of various boundaries depends on judgement by the geologist in the light of field observation, experience, and existing concepts. Further information may dictate the need for revision.

Basic equipment comprises a mapping board with plastic cover, base map, compass, clinometer, notebook, black and coloured pencils, sharpener, and eraser. Air photographs can be extremely useful. A hammer and hand lens enable fresh rock to be obtained and studied. A fine-line pen with waterproof black ink will be needed to ink-in the map and notebook, and it is helpful to record some map features in coloured ink.

Base Map

When topographical base maps are obtainable at different scales, the choice will be guided by the nature of the investigation. Scales of 1 : 100 000 to 1 : 50 000 are appropriate for regional exploration and reconnaissance. More detailed mapping is generally performed on scales of 1 : 25 000 to 1 : 5000. Very detailed work in connection with mines, quarries, and engineering sites may require scales of 1 : 1000 to 1 : 50, often available as company plans. The 1 : 10 000 Ordnance Survey sheets used by the BGS show contours at 5 or 10 m intervals, the positions of streams, buildings, fences, and roads, and a numbered grid. It is relatively easy to locate a position on the map, record it as a grid reference, and identify morphology by 'reading' the contours. Accurate contoured maps can be made from air photographs using specialized photogrammetric techniques.

Where suitable topographical maps have not been produced, have poor accuracy, or are unavailable because of political or military restriction, the geologist must find an alternative. Small-scale maps can be enlarged photographically, or scanned or digitized and redrawn by computer, but only have the detail of the original.

Satellite images, obtainable for all the Earth's surface, can be processed to give a coarse resolution base map suitable for reconnaissance work. However, the creation of contoured maps is restricted to satellites which 'see' in three dimensions, through overlapping images.

Air Photographs

The most commonly used air photographs are those made on black and white panchromatic film from level flight with the camera axis vertical. Approximate scales of contact prints are generally on the order of 1 : 8000 to 1 : 40 000. Typically, there is 60% photograph overlap along the line of flight and 30% overlap of adjacent lines, allowing the area to be viewed as a three-dimensional image through a stereoscope. As a result of spacing, the vertical scale is markedly exaggerated, enhancing topographical features that might not otherwise be obvious.

Air photographs are not maps. Only the centre (principal point) is viewed from directly above, comparable to a map. Away from the centre, the top of a vertical object appears displaced radially outwards relative to its base. The scale varies according to terrain height relative to flying height and to camera tilt.

Despite such limitations, air photographs can be immensely useful for locating position and tracing boundaries. Data recorded on overlays in the field are subsequently transferred to the base map. In three-dimensional view, photographs display relationships between terrain, drainage, and geology over a broader area than can be seen from a ground position. Variations in tone and texture, plus the dip of strata and patterns of fractures, can be interpreted to yield a photogeological map showing solid and superficial units and the size, shape, and orientation of the principal structures. Field investigation is essential to identify rock types and measure structures, but photo-interpretation is strongly recommended prior to fieldwork and each evening to give a rapid, detailed overview and allow optimum planning of traverses. Methods of interpretation are described in relevant books. Photogeological information is recorded on the base map in a colour that distinguishes it from field data.

Equipment

Map Board/Case

The simplest combination is an A4 clipboard or plywood, plastic, or aluminium sheet of about 33 cm × 23 cm, plus elastic bands to hold the field maps and a large polythene bag for weather protection. More complicated map cases may have a lid and pockets and slots to hold pencils, notebook, and other items. Field sheets can be held by spring clips, but as these affect a nearby compass it is better to use elastic bands. Cut the base map to suitably sized field sheets, say 18 cm × 25 cm, number all grid lines, and mount or photocopy onto thin card leaving a blank protective margin.

Pencils

Sharp graphite pencils of H or 2H hardness are recommended for map work because softer grades tend to smudge. Structural readings and map notes should be made in pencil in the field, not directly in ink. Pencil notes can be repositioned if their map space is needed, but ink errors are difficult to remove. At the end of the day's work, however, all pencilled data must be inked. Coloured pencils for recording lithologies should be good quality, thin-leaded, and waterproof. About 12 contrasting colours serve most purposes. Carry spare pencils, plus a sharpener and eraser.

Mapping Pens

Waterproof black and coloured inks (red, blue, and green) are used to make a permanent record of the day's work on the map and in the notebook. Barrel-type technical pens giving a line width of 0.15–0.2 mm are a common preference for maps and 0.2–0.3 mm pens for notebooks. Fine-line, waterproof, fibre- or ceramic-tipped pens of nominal 0.1 mm line size are an economical option for the colour pigments.

Compass and Clinometer

A compass is used to take bearings on distant objects to establish one's map position and to determine the orientation of structures. A clinometer measures the inclination of a planar or linear feature relative to a horizontal reference plane.

Several instruments combine both functions. A popular unit, the Silva Ranger Type 25TDCL, has a liquid-filled compass cell, graduated 0–360°, which can be rotated in its rectangular baseboard. The hinged lid has a sighting line and a mirror to view an object and the compass simultaneously. The transparent cell base has edge markings around half the circumference, with 0° in the centre and 90° at each end. When the ends are aligned with the baseboard edge and the instrument is held in a vertical plane, a centrally mounted, free-swinging arrow registers the angle of inclination. The cell base is also inscribed with parallels for alignment with the map grid and an arrow shape to register the position of the compass needle. An adjusting screw allows the arrow to be turned relative to the parallel lines, thus correcting for the angle between magnetic north and grid north. The use of such an instrument is illustrated in Figure 1. Readings taken from the rock are plotted immediately on the map. The orientation of distant or inaccessible surfaces can be found by moving to a position along strike (sharpest edge view), taking a compass reading, and then tilting the clinometer in mid-air to match the dip.

Hand Lens

A lens is essential and it is worth paying for optics which give a flat field and sharp undistorted image to the edge. A magnification of ×10 is the most useful and the lens should be worn on a cord around the neck so that it is constantly at hand to examine rock textures, mineral grains, and the like.

Hammer and Chisel

Geological hammers are used to break off a fresh surface of rock for examination or sampling, and are available in a range of sizes, shapes, and materials. For general use, a weight of 0.7–0.9 kg is suitable. Heavier ones (1.1–1.8 kg) are better suited for hard igneous and metamorphic rocks. One end of the head has a square face; the other may be chisel shaped, useful for layered rocks, or pick shaped, for prising cracks or digging into soil. Handles can be of wood, fibreglass, or steel.

Because of the hardened steel face, one must never use a hammer as a chisel and hit it with another, as metal shards can fly off and cause injury. Instead, use a soft steel chisel. Safety goggles should be worn when hammering or chiselling rocks.

Field Notebook

The general preference is for a pocket-sized notebook of around 12 cm × 20 cm. The paper must be of high quality, still usable after several cycles of soaking and drying, and securely bound between hard covers. Books designed for geologists are stocked by relevant dealers. Ordinary stationers' notebooks are not suitable. Notebooks can have an end or side hinge, and the pages may be plain, lined, gridded, or mixed. A 5 mm square grid is useful for recording measured stratigraphical sections and scaled plans.

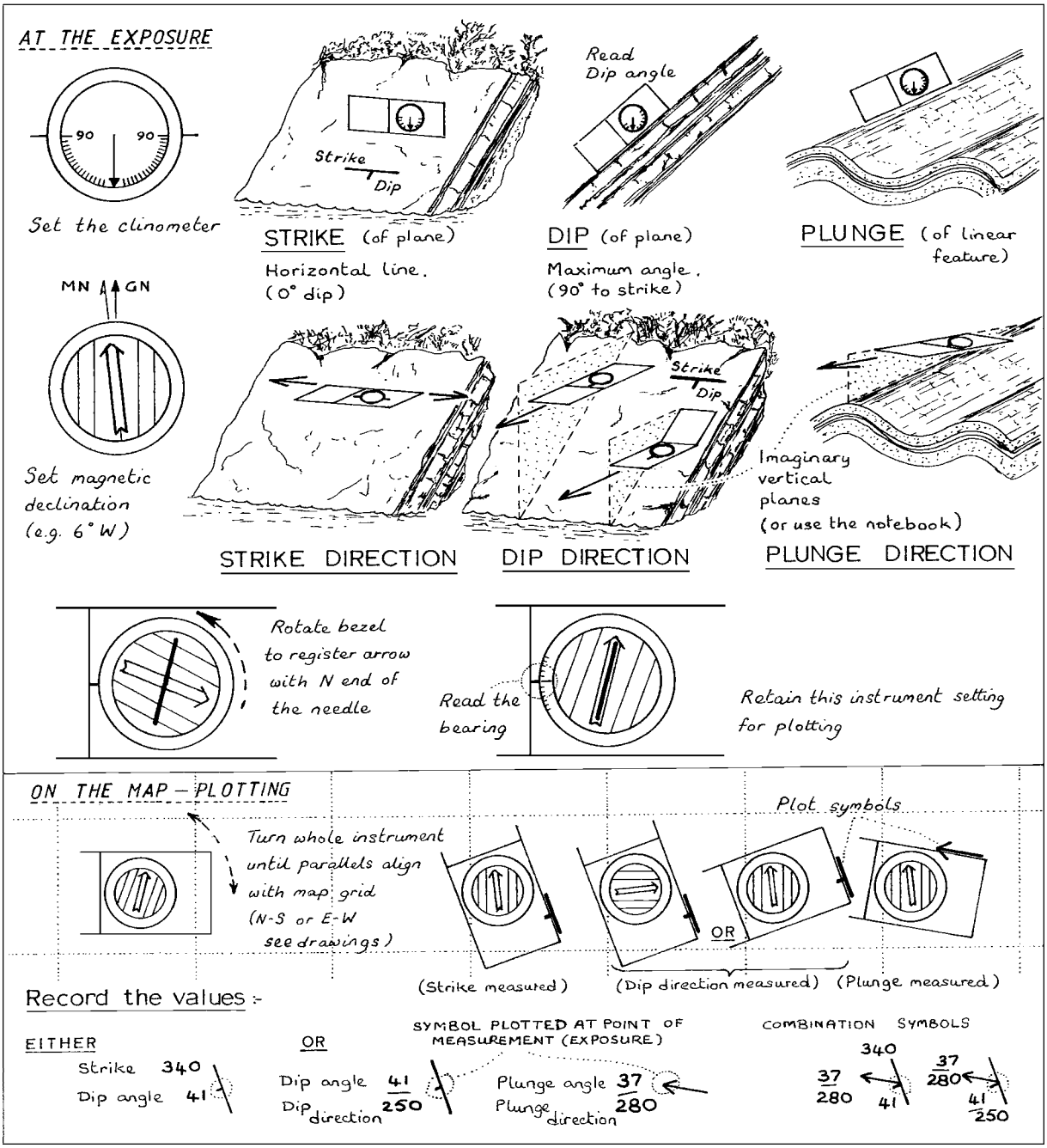


Figure 1 Use of compass clinometer.

Other Equipment

For general field mapping, the following are often or occasionally used.

- Small dropper bottle containing 10% hydrochloric acid. A drop causes calcite to effervesce, but shows no reaction with dolomite.
- A 3 m steel measuring tape to determine bed thicknesses, and perhaps a 30 m linen tape for distances.
- A scale rule and protractor for measuring distances and angles on the map.
- A Global Positioning System (GPS) instrument. Although rather expensive, this is increasingly being seen as a standard field item for its ability to locate position to within about 10 m.
- An altimeter can assist positioning if the field map has reliable contours, but is less useful than GPS and subject to weather-related pressure changes.

- A pocket stereoscope is important if mapping involves the field use of air photographs.
- A stereographic projection net assists in understanding structural relationships.
- Binoculars allow the study of inaccessible rock faces and can help in locating exposures and routes.
- A camera, preferably with the ability to take close-up as well as distant shots.
- A penknife for conducting scratch tests, amongst other uses.
- Sampling bags.

Field Clothing

Clothing appropriate to the conditions allows the geologist to work comfortably throughout the day and not to be stressed by heat, cold, or wetness to the point at which concentration lapses and safety is at risk. Typically, a day rucksack is carried, large enough to take spare clothing, food, water, mapping equipment, safety items, reference literature, and collected specimens. In cold wet climates, the aim is to be warm and weatherproof by wearing items such as an undervest, shirt, fleece or sweater, and waterproof hooded anorak, plus a warm hat, scarf, and gloves. Trousers are best made of quick-drying, close-weave cotton, not jeans, with further protection from waterproof overtrousers. Well-cleated strong leather boots giving firm ankle support are recommended for most fieldwork, with wellington boots for wading activities.

Duvet clothing, warm padded footwear, and windproof outer garments are used in very cold dry climates. In hot dry conditions, protection against direct sun radiation is best given by a loose, long-sleeved cotton shirt, long trousers, and a wide-brimmed sun hat, but short trousers and sleeves are preferred by many. Corresponding footwear is lightweight whilst still providing support and grip.

Health and Safety

A doctor can advise on what inoculations, vaccinations, and other medical precautions are needed for work in different parts of the world. A first-aid kit should always be part of the field geologist's equipment. Items which come under the general heading of safety include goggles, a hard hat, torch, whistle, phone/radio, and survival bag. Brightly coloured field clothing will help a search party, should the need arise.

Mapping Preliminaries

Preparatory work before entering the field enables the survey to be conducted more efficiently. Administrative aspects include obtaining permits, establishing how to contact medical and emergency services, and arranging that others know your daily whereabouts. Relevant literature, maps, and borehole records will indicate what is known about the rocks and the nature of uncertainties. Air photograph interpretation and the production of a photogeological map will identify the type of terrain, the range of solid and superficial units, and the principal structures.

One or more days will be needed for reconnaissance before mapping starts to provide first-hand knowledge of the rock types, mappable formations, nature of exposure, and style of structures. Notes are made on access routes, areas of good and poor exposure, and parts that are inaccessible or hazardous. An important act is to make contact with local people to explain the work and confirm permission.

Following the field visit, decisions can be made on mapping methods. Often the most productive traverses are along streams which flow across the general strike and have cut down to bedrock. Parts of an area may be best mapped by tracing the outcrop of a marker unit. Broad area reconnaissance may dictate widely spaced traverses on parallel compass lines, whereas a mineral prospect might need enlarged scale mapping controlled by plane-tabling or compass and tape. Measured stratigraphical sections and key structural areas can be drawn at large scale in the notebook. The overall aim is to measure and record as many exposures as possible to cover the area within the allotted time, constantly remembering that the map and notebook must be easily read, understood, and followed by another geologist. Hence exposures must be precisely located, symbols drawn neatly and accurately in conventional forms, numbers legible, and notes comprehensible. Map and notebook must both distinguish between observation and inference. Various symbols, colours, and abbreviations will be used to depict the geology. Prepare a 'key' on the map to explain these, with formations in stratigraphical order.

Symbols

There is no universal set of mapping symbols, but published lists have broad similarity and an example set is given in Figure 2. Planar structures are generally depicted by a bar in the strike direction and a tick or other shape in the dip direction. Linear structures are shown as arrows, with various head, shaft, and tail ornament. Orientation measurements for lineations are given as an inclination angle (plunge) and its bearing (plunge direction) in the form 20/232. Two methods are in use for the orientation of planes. One

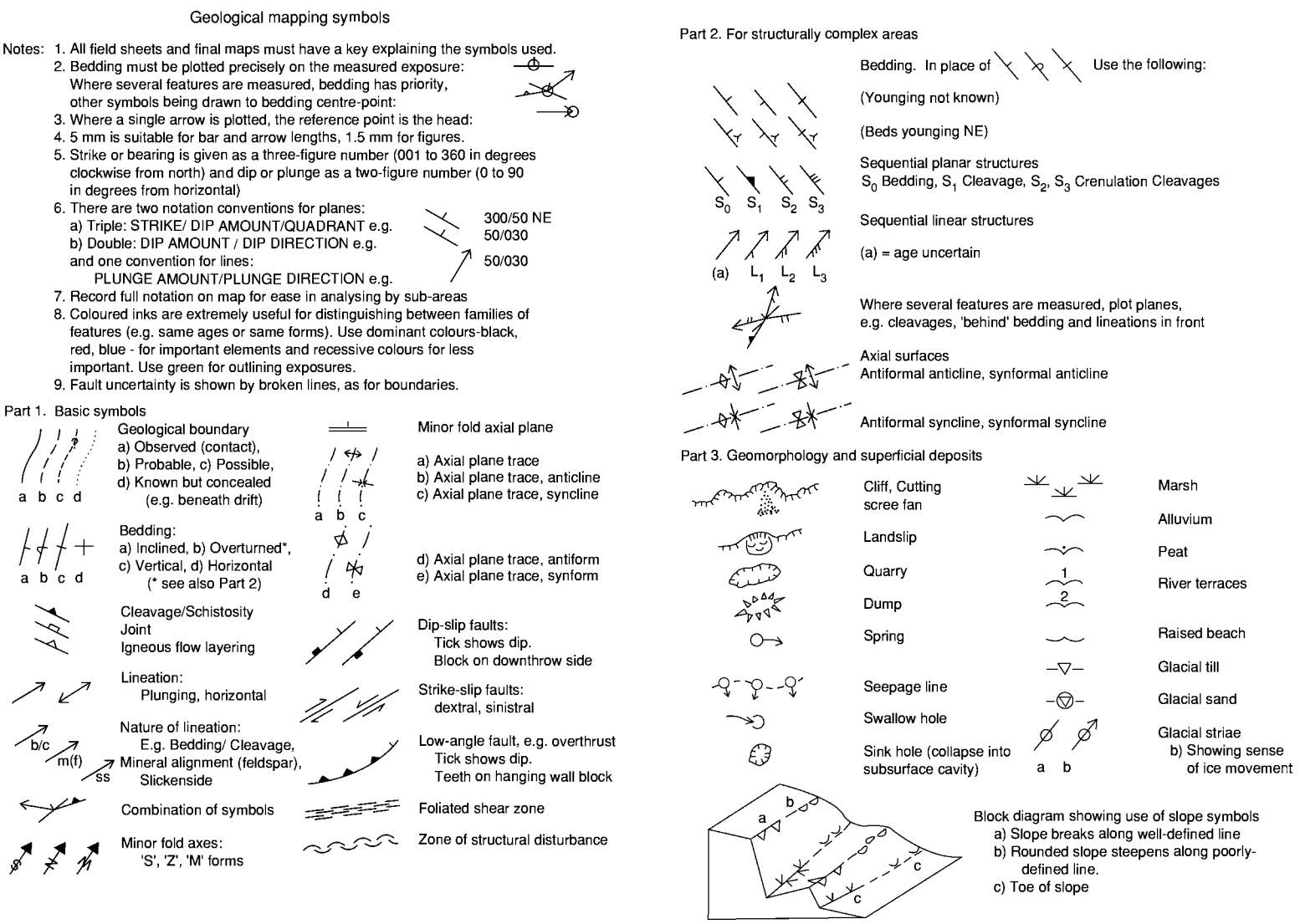


Figure 2 Geological mapping symbols.

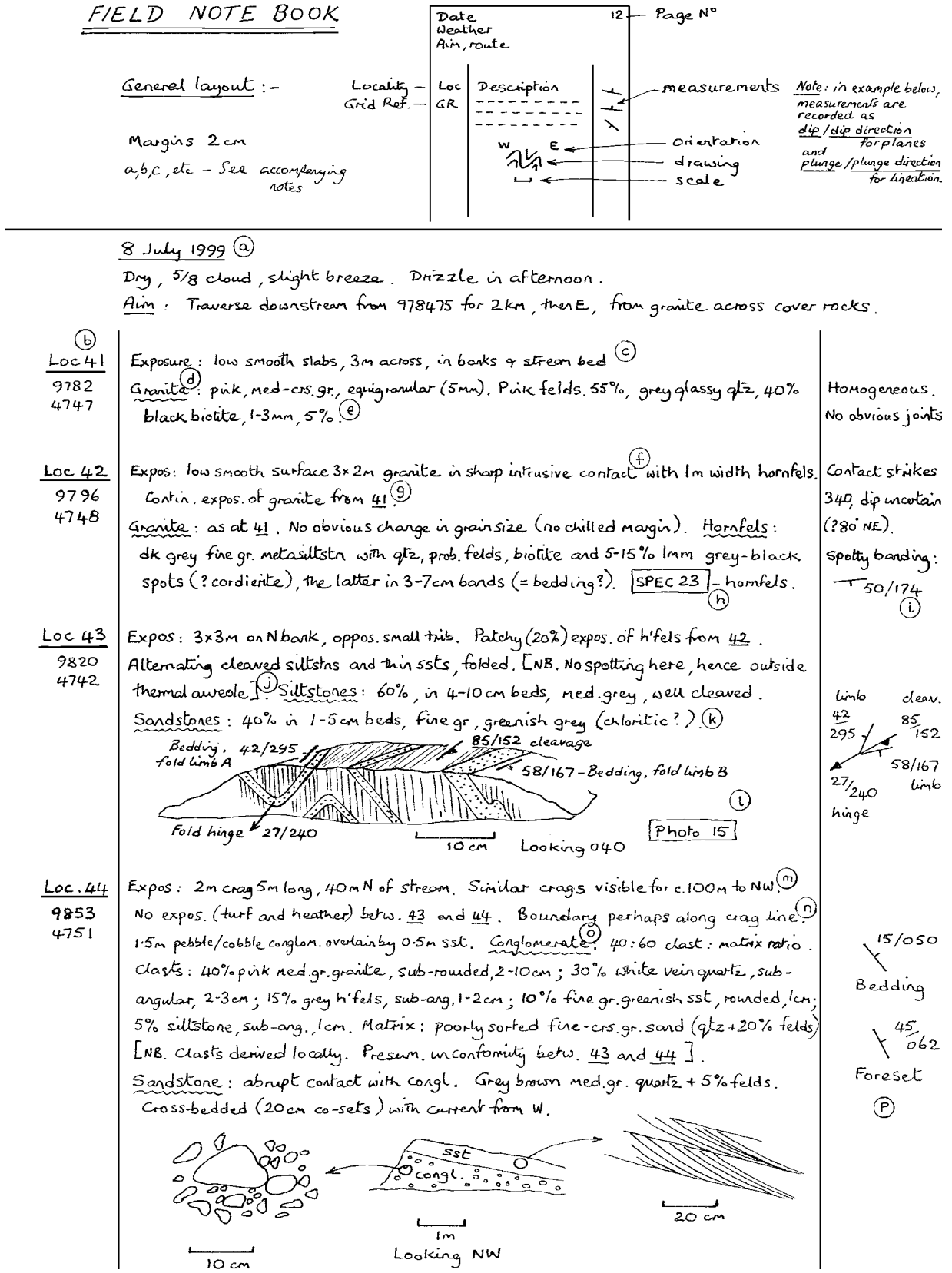


Figure 3 Field notebook. (a) The daily heading gives the date, weather conditions, general aim, and route. (b) Number the localities consecutively. Do not restart each day. An eight figure grid reference (eastings then northings) locates the position to 10 m. (c) Describe the nature of the exposure, particularly its size. A 1 m exposure of sandstone might be part of a shale formation, whereas a

Continued

records strike / dip angle / quadrant, as in 330/30 SW. The other gives dip angle/dip direction, as in 30/240. The latter is advised. It has fewer components, is less prone to error, and is in the same form as lineation data.

Use of Field Notebook

The notebook is the essential constant companion to the field map. While the map depicts the important readings, boundaries, and formations in spatial relationship, the notebook holds further readings, detailed descriptions, drawings, and much extra information. The order of use differs amongst geologists, but a common practice is to make a thorough examination of the exposure, record it in the notebook, and then plot the salient features on the map. An example of notebook layout and entries appropriate to a mapping survey is shown in **Figure 3**. Corresponding explanatory notes are also given, extended in a few cases to cover other geological situations. Should the field map be lost, it would be possible to reconstruct much of it from the notebook, although not boundaries drawn directly from field observation. Drawings should always be made in pencil, to be inked subsequently when the correct line has been achieved. Notes can be in pencil or ink.

Use of Field Map

Mapping involves placing a record on the map of: (1) the presence of rocks at the surface (exposures); (2) the type of rock(s) exposed; (3) the orientation of bedding, contacts, and tectonic structures; (4) the nature of superficial cover; and (5) landform features, such as river terraces. Information from various sources is combined to allow: (6) boundaries to be drawn on the map, thus giving the 'outcrops' of mappable units.

At each location, proceed as follows.

1. Determine the position on the map by GPS, by reference to nearby features and contours, or by resection using two or three compass bearing lines 60°–90° apart.
2. Outline the limits of the exposure.
3. Identify the rock type(s) and formation, and colour the exposure on the map.
4. Record the locality, grid reference, rock descriptions, measurements, drawings, etc., in the notebook.
5. If a formation contact is present, plot it on the map. Scan surrounding country for clues to extend the line.
6. Measure the orientation of bedding and plot the symbol on the map, precisely where the reading was taken. Record measurements in 1.5 mm figures.
7. Plot other features, such as cleavage, drawing the symbols to the centre of the bedding bar. In complex areas, plot the most important structures, leaving others in the notebook. No reading should be plotted in a position which was not the point of measurement.
8. Add locality number and relevant notes, particularly for key stratigraphical, fossil, and structural localities. Notes can be written on unused areas, such as the sea, and linked to the exposure by a dotted line.
9. Move to another locality, observing continually. If unexposed ground is thought to be of the same formation, shade it more lightly with the same colour.

If the next exposure is of an entirely different formation, there must be a boundary somewhere in the intervening ground. Whether it is sedimentary, igneous, or tectonic is commonly inferred from its geological setting. Its position is judged after assessing

30 m cliff of sandstone clearly belongs to a sandstone formation. (d) Underline key words. (Use the formation colour employed on the map.) (e) Give size, shape, colours, and proportions of the different rock constituents, not just a rock name or mineral list. (f) Contacts demand particular attention. Is the contact sedimentary, igneous, or tectonic, sharp, gradational, sheared, discordant, etc.? What are the age relationships? With igneous contacts look for features such as chilling, baking, xenoliths, and veins. (g) Note the nature of ground between localities, e.g., continuous, patchy or no exposure, type of vegetation. (h) Record specimen numbers. Collect typical rocks: abnormal ones are additional. Make notes if a specimen has been oriented (marked in the field such that it can be oriented identically in the laboratory). (i) Plotting symbols and figures in the margin as they will appear on the map provides a check against possible map errors, allows easy data retrieval, and provides a visual record of changes in orientation. (j) Include deductions or implications arising from the geological observations. (k) Proportions and thicknesses are more informative than descriptions such as 'thinly bedded sandstones and siltstones'. If fossils are present, record their identity, size, state, distribution, and abundance. (l) In deformed areas, record all planar and linear elements (bedding, cleavages, axial planes, fold axes, lineations). Always draw folds, because readings alone can never convey shape. Draw accurately, label the rock types, and avoid ornamental lines that might be mistaken for geological fabric. Incorporate a scale and an orientation (N arrow for plan views, viewing direction for sections). It helps to show readings on the drawing. (m) Note geological features in view, as well as those at the exposure. (n) Record any evidence, or lack of it, for a deduced boundary. (o) Conglomerates indicate which rocks were exposed to erosion at a particular time. The proportions and shapes of the constituents provide evidence on transport and depositional environment. (p) The measurement of bedding and truncated foresets shows younging of the sequence and gives information to calculate the current flow vector. Other sedimentary structures have similar potential.

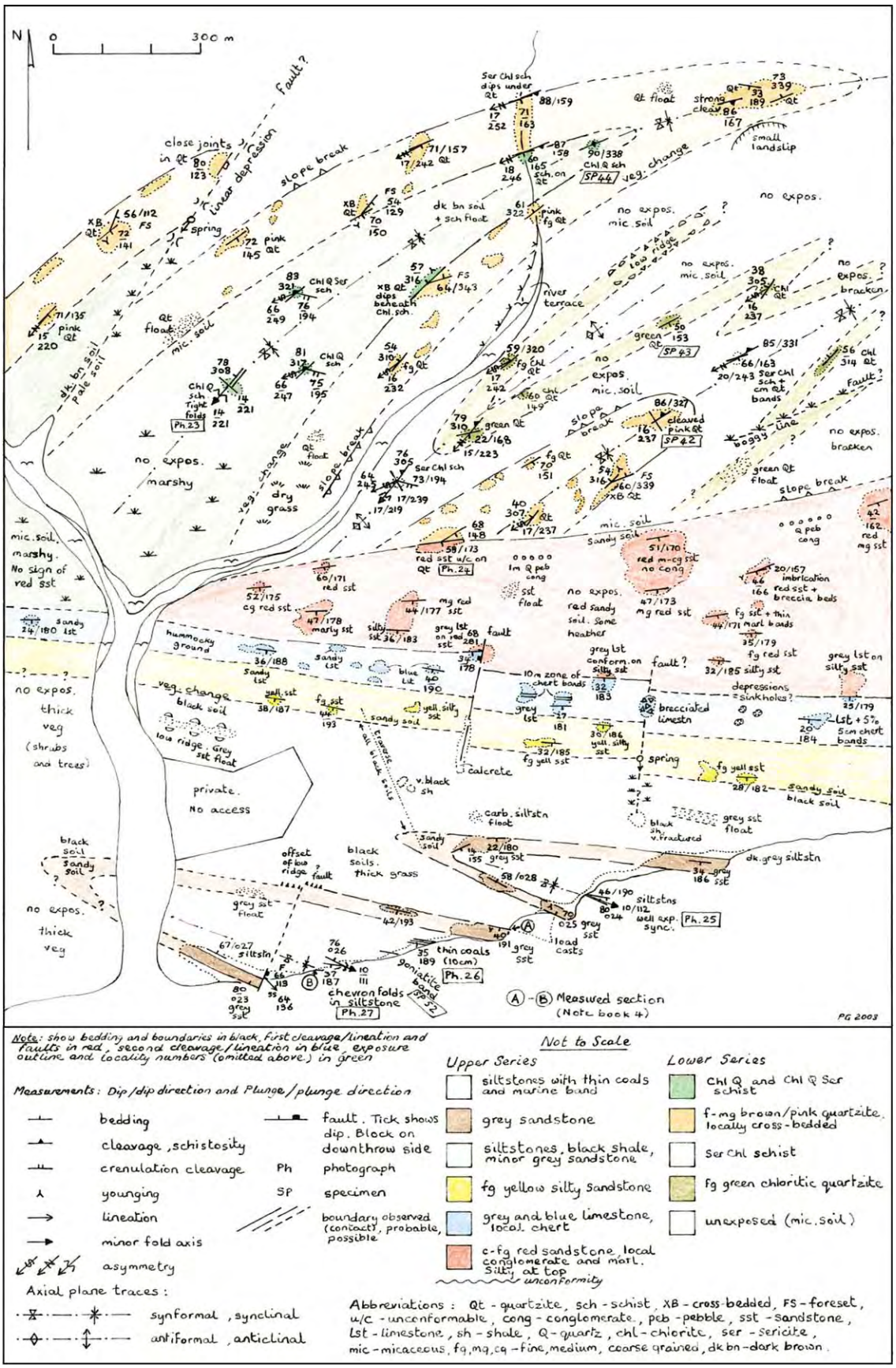


Figure 4 Example of field map entries. Map grid, contours, and locality numbers have been omitted for clarity. The map symbols would be in coloured ink. Exposures would be outlined in a continuous, not dotted, green line.

field evidence, such as a change in soil composition, soil colour, slope change, topographical feature, spring line, and change in vegetation, and recording these on the map. Study of air photographs may help. A boundary is pencilled on the map in a position which fits the evidence best, drawn as a full line if well constrained, or long or short dashes for decreasing confidence.

It is clear that geological field mapping means making a map in the field whilst the evidence is in sight, not placing locality numbers on the map to plot the data later and draw arbitrary lines between rock changes. An example of field map entries is given in Figure 4.

At the end of the field day, the map and notebook are inked in to make a permanent record, using coloured ink on the map as appropriate (Figure 4). Ink the notebook drawings accurately to preserve the lines made whilst the object was in view. Analyse structural data on a stereonet and construct trial cross-sections. Never leave a field area without drawing at least one geological cross-section. Specimens collected during the day are given permanent labels and laid out for further reference. It is helpful at this stage to summarize the day's findings, place it in context, and make notes for the final report.

Superficial Deposits

Some areas are deeply weathered and covered by residual soils or regolith through which only the most resistant rocks emerge at the surface. Others are blanketed by transported materials, such as windblown sands or glacial deposits. For residual soils, information can be obtained by a study of auger samples, recording parameters such as soil colour, composition, and consistency. Look also for fragments ('float') brought to the surface by fallen trees, ploughing, or burrowing animals. Air photographs can prove to be useful in indicating boundaries, particularly where vegetation is well controlled by soil type and has been little disturbed by human activity.

For transported materials, the geologist aims to outline the deposit and classify the type. The 'drift' maps of the BGS, for example, show various kinds of gravel, sand, and clay deposits formed from standing, moving, or melting ice, plus river terraces, fluvial and estuarine alluvium, peat, and others. Some can be clearly seen and mapped from air photographs or a high vantage point. Relevant observations are placed on the field map by a selection of colours, overprint symbols, and notes. Areas of slope failure should be marked.

Ancient soils are not commonly preserved in the stratigraphical column but may have left traces, such as weathered and reddened surfaces; here, a knowledge of soil processes helps in assessing the palaeoclimate.

See Also

Engineering Geology: Geological Maps; Natural and Anthropogenic Geohazards; Site and Ground Investigation; Ground Water Monitoring at Solid Waste Landfills. **Environmental Geology**. **Geochemical Exploration**. **Geological Maps and Their Interpretation**. **Geological Surveys**. **Geomorphology**. **Remote Sensing:** GIS. **Seismic Surveys**. **Soil Mechanics**. **Soils:** Modern.

Further Reading

Allum JAE (1966) *Photogeology and Regional Mapping*. Oxford: Pergamon.

Barnes JW (1991) *Basic Geological Mapping,* 2nd edn. *Geological Society of London Handbook,* Milton Keynes: Open University Press.

Boulter CA (1989) *Four Dimensional Analysis of Geological Maps*. Chichester: Wiley.

Compton RR (1985) *Geology in the Field*. New York: Wiley.

Lattman LH and Ray RG (1965) *Aerial Photographs in Field Geology*. New York: Holt, Rinehart and Winston.

Lillesand TM and Kiefer RW (1979) *Remote Sensing and Image Interpretation*. New York: Wiley.

Maltman A (1990) *Geological Maps: An Introduction*. Milton Keynes: Open University Press.

McClay KR (1987) *The Mapping of Geological Structures*. *Geological Society of London Handbook*. Milton Keynes: Open University Press.

Moseley F (1981) *Methods in Field Geology*. Oxford/San Francisco: Freeman.

Ramsay JG and Huber MI (1987) *The Techniques of Modern Structural Geology,* vol. 2, *Folds and Fractures*. London: Academic Press.

Ray RG (1960) *Aerial Photographs in Geologic Interpretation and Mapping*. *United States Geological Survey Professional Paper 373*. Washington: US Government Printing Office.

Tucker ME (1982) *The Field Description of Sedimentary Rocks*. *Geological Society of London Handbook*. Milton Keynes: Open University Press.

GEOLOGICAL MAPS AND THEIR INTERPRETATION

A Maltman, University of Wales, Aberystwyth, UK

Introduction

Geological maps at their simplest are portrayals of the distribution of geological materials at the Earth's surface (*see* **Geological Field Mapping**). Any covering, say of agricultural soil or concrete, is ignored. Such a picture of areal distributions is itself highly useful, but the real power of geological maps arises from the further interpretations that can be made from them – into the third-dimension below the ground and back into geological time. Because of this breadth of possibilities, geological maps are a fundamental tool of both academic and applied geologists, and the governments of most developed countries underwrite a national body responsible for the geological surveying of the nation. Indeed, geological maps have been referred to as 'the visual language of geologists'. Today, yet new possibilities are blossoming as GIS (*see* **Remote Sensing:** GIS), spatial databases, and digital imagery are applied to geological map methods. How best to harness all this tremendous potential is the subject of much current dialogue. Fundamental map principles, however, are unchanging, and it is these that are summarized here.

The Nature of Geological Maps

The distributions of the different kinds of rocks are plotted onto some kind of topographic base image, traditionally a topographic map. Scales vary from very small, such as 1:1 000 000 or less (which may summarize the geology of entire regions), through medium scales such as 1:100 000 to 1: 25 000 (showing particular areas in reasonable detail), to large-scale maps at scales of 1:500 or even greater (typically used for specific sites of some commercial activity). Small-scale maps commonly have to neglect topographic relief, but at larger scales topographic contours can be shown, and the interaction between the physiography of the land surface and the underlying geology can be important in three-dimensional interpretations. The distribution plots are increasingly being linked into Digital Elevation Models, to aid visualisation of the interplay between geology and topographic relief.

The areal extents of the various materials are shown by different ornaments in the case of black and white maps, or by different colours in the case of most published maps, sometimes with additional letter or number symbols to aid distinction. The geological ornaments/colours and symbols are explained in an accompanying key or legend. While in some cases the Earth materials in a given area fall naturally into groups that are convenient for depiction at a particular map scale, in many cases the surveyor has to judge how best to make appropriate sub-divisions. Small-scale maps commonly portray divisions based on the geological period in which the material formed; some more specialized geological maps may show divisions according to the fossils contained in deposits, or some other particular characteristic.

The majority of geological maps, however, are litho-stratigraphic: the divisions are based on the type of Earth material and their stratigraphic position. The various divisions of material types are, conventionally, arranged in the map key in order of their genesis, upwards from oldest to youngest. The actual geological ages may or may not be added. On medium- and larger-scale maps, each division is commonly referred to as a map 'unit' or 'formation'. It is important to understand that although a particular formation may be coloured boldly on the map, and given some imposing name, it has been defined subjectively by the surveyor. It may be quite indistinct on the ground; another worker may have divided the units differently. It is also the convention to make a major distinction between solid bedrock and any overlying, unlithified, geological materials. Some geological maps ignore the latter, except where they are especially significant, and are thus referred to as bedrock or – in the UK – solid maps. On the other hand, maps that emphasize them are variously referred to as surficial, superficial, Quaternary or – again in some UK usage – drift maps.

Many published geological maps, which by definition depict things in the horizontal plane of the Earth's surface, are accompanied by geological cross-sections. These portray the arrangement of the material in the vertical. It is this complementarity of maps and sections that underscores the three-dimensional aspects inherent in geological maps, a topic expanded in the following section.

Mention was made above of the subjectivity of the units into which the Earth materials are divided for the purpose of a map. Also intrinsic in a geological map is the interpretive nature of the distributions shown. Unlike almost all other kinds of maps, much of the information shown is interpretive. Except in

the most arid parts of the land surface, the covering of soil, etc., means that the surveyor has had to judge the areal extent of the underlying Earth materials, from scattered exposures, in cliffs, river banks, road cuttings and the like, and other, indirect, evidence. Geological maps, therefore, are continually being refined as additional evidence becomes available.

The Interpretation of Geological Maps

Many of the uses of geological maps involve dealing with geological information in three dimensions. New IT techniques are becoming available to facilitate this but much work still relies on interpreting the third-dimension from the flat map – the sheet of paper or computer screen. Sedimentary rocks, being arranged in approximately tabular strata, lend themselves to this kind of treatment. The strata comprise a three-dimensional configuration, originally a horizontal succession but in the case of marine deposits that have been uplifted to form land – the most widespread situation – now in some combination of horizontal, tilted, folded, and faulted layers. The present-day land surface is an eroded slice, roughly horizontal, through this array. Properly, therefore, a geological map is a plot of the intersections or outcrops of the strata with the present land surface, projected onto the horizontal plane of the map. Note incidentally, in this context, the geological terms exposure and outcrop differ in meaning. An outcrop of a formation, i.e., where it intersects with the land surface, and hence what is depicted on a geological map, may or may not actually be exposed, i.e., lacking modern cover and be directly visible. From the patterns made by the outcrops on the map, a trained geologist can infer the three-dimensional configuration of the formations, that is, how they were arranged above the land surface before their erosion and, more importantly, how they are disposed below the ground, in the 'subsurface'. The geologist can do this in a quantitative or a non-quantitative way, the latter by simply assessing the geological patterns on a map and mentally visualizing – perhaps while sketching rough cross-sections – the spatial arrangements. Some examples follow.

Visual Assessment

A sequence of map formations passing progressively from older to younger implies: (i) that the succession has been tilted (if they were still horizontal only the uppermost division would outcrop), and (ii) the direction of dip (Figure 1). Even with such a preliminary interpretation it becomes possible to gauge what lies below the land surface in various places, apart from those sites where the lowest, oldest known formation is outcropping. This simple principle, that

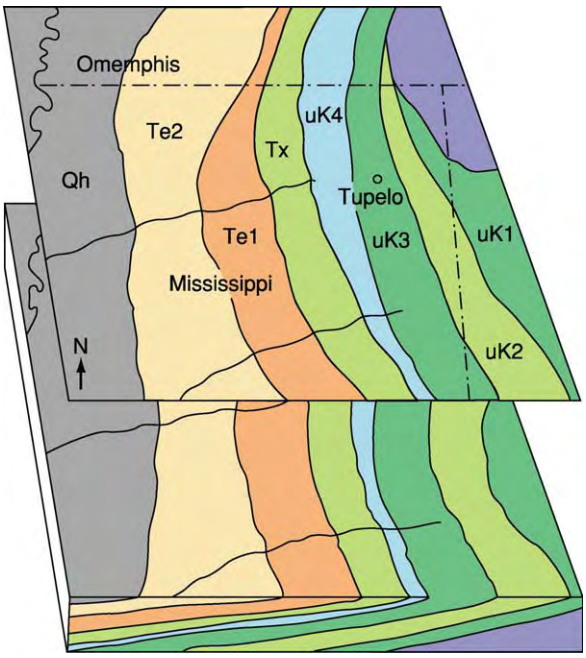


Figure 1 Block diagram and map to show how formations dip towards the direction in which they become younger (ignoring, as may be reasonable on a small scale map, topographic relief). In this example, the geological map shows that a journey west wards across the State of Mississippi would encounter progres sively younger formations. The simplest and much the most likely interpretation is that the formations are dipping westwards, as evident from the front panel of the block diagram. uK1, uK2, uK3, and uK4 are progressively younger Cretaceous formations, Tx, Te1, and Te2 are progressively younger Tertiary formations, and Qh is the youngest formation, of Quaternary age. Air photographs and satellite images may be helpful; geological maps of the sea floor commonly involve seismic surveying.

outcropping formations pass from older to younger in the direction towards which they are dipping, needs care with larger-scale (usually contoured) maps of rugged country, where the topographic relief may interfere with and complicate the outcrop patterns. However, further simple principles can be applied in such cases. A dissected landscape developed in horizontal formations can give what at first glance is a bewildering geological map, but the outcropping formations, discerned by their boundary traces, will be exactly parallel to the topographic contours (because they, too, of necessity are horizontal; Figure 2). The outcrop patterns progressively deviate from this parallelism as the dip angle increases until, at the other extreme, vertical formations outcrop without any influence of topography. Their outcrops cross topographic contour lines to give straight traces. The more general case of dipping formations is best interpreted by assessing the direction and shapes of the V-shape outcrops made where they intersect with topographic valleys (Figures 3 and 4). Note that these are geometric principles arising from the intersections

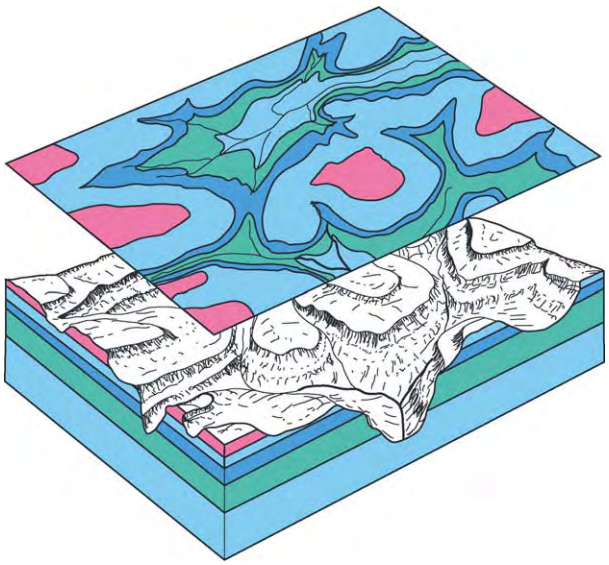


Figure 2 Block diagram and map to show how outcrop traces of horizontal formations parallel topographic contours (as they must also be horizontal). In the example shown here, the geo logical map is deceptively intricate. The outcrops are irregular, despite the simple, horizontal arrangements of the strata, because the landscape is highly dissected.

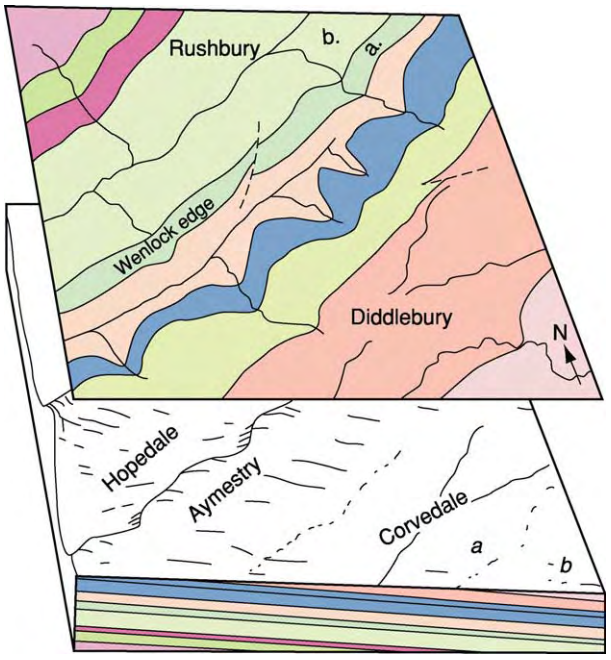


Figure 3 Block diagram and map to show the V shape made by the outcrop of a dipping formation crossing a valley. The V in map view points in the direction of dip, here towards the southeast. Note also the influence of land slope on outcrop width. For example, the outcrop of formation *a*, at the steep escarpment of Wenlock Edge, is much narrower than that of formation *b*, at the flat land of Hopedale, even though they have similar thicknesses.

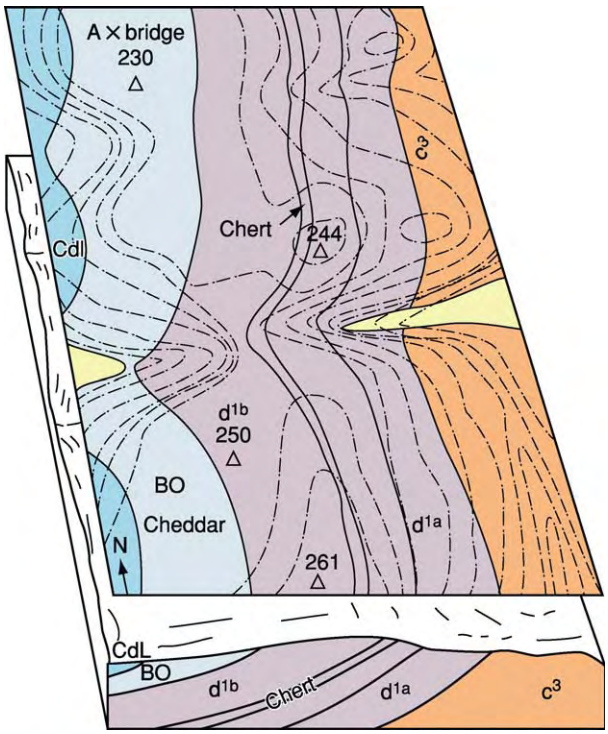


Figure 4 Block diagram and map to show how outcrops cross valleys with a more open shape where the formation dip is greater. As in **Figure 3**, the V in map view points in the direction of dip, here towards the W. The outcrop of more steeply dipping formations such as c^3 makes a more open V than move gently dipping units such as BO. The superficial deposits (yellow) are approximate horizontal and make the tightest V shape, parallel to topographic contours.

of surfaces – any geological surfaces, whether or not they are sedimentary strata. Thus geological faults (*see* **Tectonics:** Faults) commonly appear on maps as more or less straight traces, not because they are faults but because fault surfaces are commonly very steeply dipping or vertical. Gently dipping thrust faults, on the other hand, will give traces that are close to parallelism with the topographic contours (Figure 5).

The relative widths of outcrops on a map of a given scale reflect the thicknesses of the formations. Hence the outcrop width will vary if the material was deposited in different thicknesses at different places. Other things being equal, thicker units give wider outcrops. On larger-scale maps, however, the dip angle of the formation and the topographic slope will further influence the outcrop width. In general, steeper dips give narrower outcrops, as do steeper topographic gradients (Figure 3).

Symmetrical, roughly mirror-image repetitions of outcrops indicate folding (*see* **Tectonics:** Folding). Older formations symmetrically flanking younger ones imply that the strata have been downwarped

into the structure known as a syncline; younger formations flanking older indicate an upwarp or anticline (Figure 5). In addition, the shape of the symmetrical pattern contains information on the three-dimensional arrangement of the fold. Outcrop sequences that repeat in a translation pattern rather than a mirror-image reflection (Figure 5) imply the presence of a fault. The effects of faults on outcrop patterns are diverse, depending on the nature of the fault and its orientation with respect to the displaced strata. On published maps, faults are normally marked by a particular symbol – commonly a heavier line – indicated in the map key. The actual nature of the fault displacement is commonly impossible to interpret from a map, because the same outcrop patterns can be generated by different kinds of fault motion. Additional, often field-based, evidence is normally needed. Unconformities are recognised on maps by the absence of strata representing a significant length of geological time, and also by discordant outcrop patterns in the case of an angular unconformity (*see* **Unconformities**). The latter is important in three-dimensional map interpretations because the geometric arrangements will differ above and below the surface of unconformity.

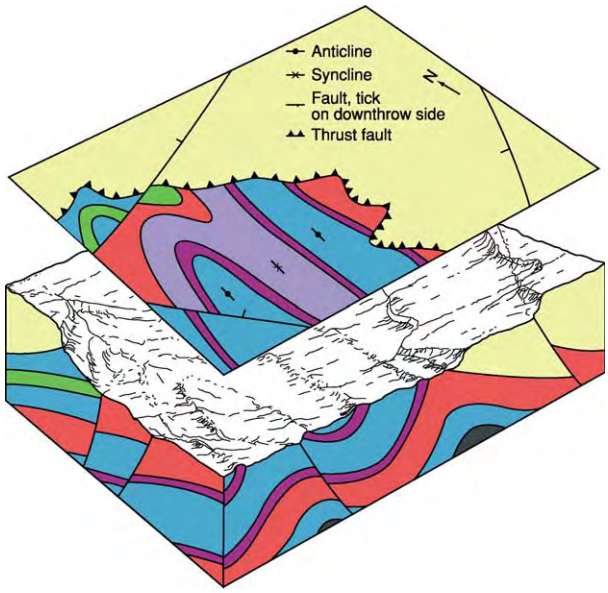


Figure 5 Block diagram and map to show outcrop patterns typical of some structures. The thrust fault is gently dipping and hence its trace is closely influenced by topography; steeply dipping faults crop out with relatively straight traces. The translatory repetition of outcrops is a typical effect of faults. The mirror image repetition of outcrops is characteristic of folds.

Quantitative Treatments

In practical applications it is commonly not enough just to visualize or sketch the subsurface arrangements; the actual subsurface depths of particular formations and their thicknesses are often required. Quantitative answers to such questions are usually arrived at through some combination of three methods. First, a scaled geological cross-section can be constructed through the site of interest and the required data measured directly from the diagram. Larger-scale maps commonly provide symbols indicating the dip angles measured at exposure by the surveyor, and extrapolations of surfaces to depth are based on these values. However, with cross-sections there can be construction inaccuracies or even mistakes – irrespective of whether the work is done manually or by computer – though this pictorial approach has the advantage of making errors more readily apparent. A second approach utilises trigonometric calculations (Figure 6). Both these approaches rely on assumptions, such as the dip angle as measured at exposure being maintained at depth, and they

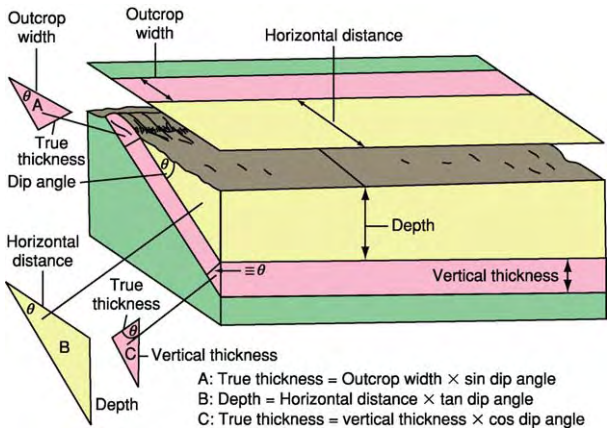


Figure 6 Schematic block diagram and map to illustrate how the thickness and subsurface depth of dipping formations can be calculate trigonometrically from map information.

only provide the quantitative information along the specified line, in the case of a cross-section, or, with the trigonometric approach, at the specified point. Computer-generated visualisations can be deceptive in their visual attractiveness: as with traditional methods they involve assumptions and subjective interpretations (and simplified algorithms).

Much more powerful is a third approach, using structure contours (Figure 7). Exactly analogous to topographic contours of the land surface, structure contours portray quantitatively the form of some specified surface below the ground. The surface might be the upper boundary of a formation of particular interest, perhaps an aquifer or a mineral-bearing horizon; it could be an unconformity or a fault surface. The contour values can represent the depth below ground of a surface or, more usefully where the land is not flat, the elevation of the geological surface, with respect to sea-level or some local datum. The numerical difference in altitudes between topographic contours for the land surface and the structure contours gives the depth of the contoured geological surface, at any desired location in the map area. Values for locations intermediate between contour lines can be interpolated. The numerical difference between contours for the upper and lower boundaries of a formation gives the vertical thickness of the material. For dipping formations this value will differ from the true thickness, at right-angles to the bounding surfaces, but the trigonometric corrections are straightforward. Contours that directly portray thicknesses are known as isopachytes.

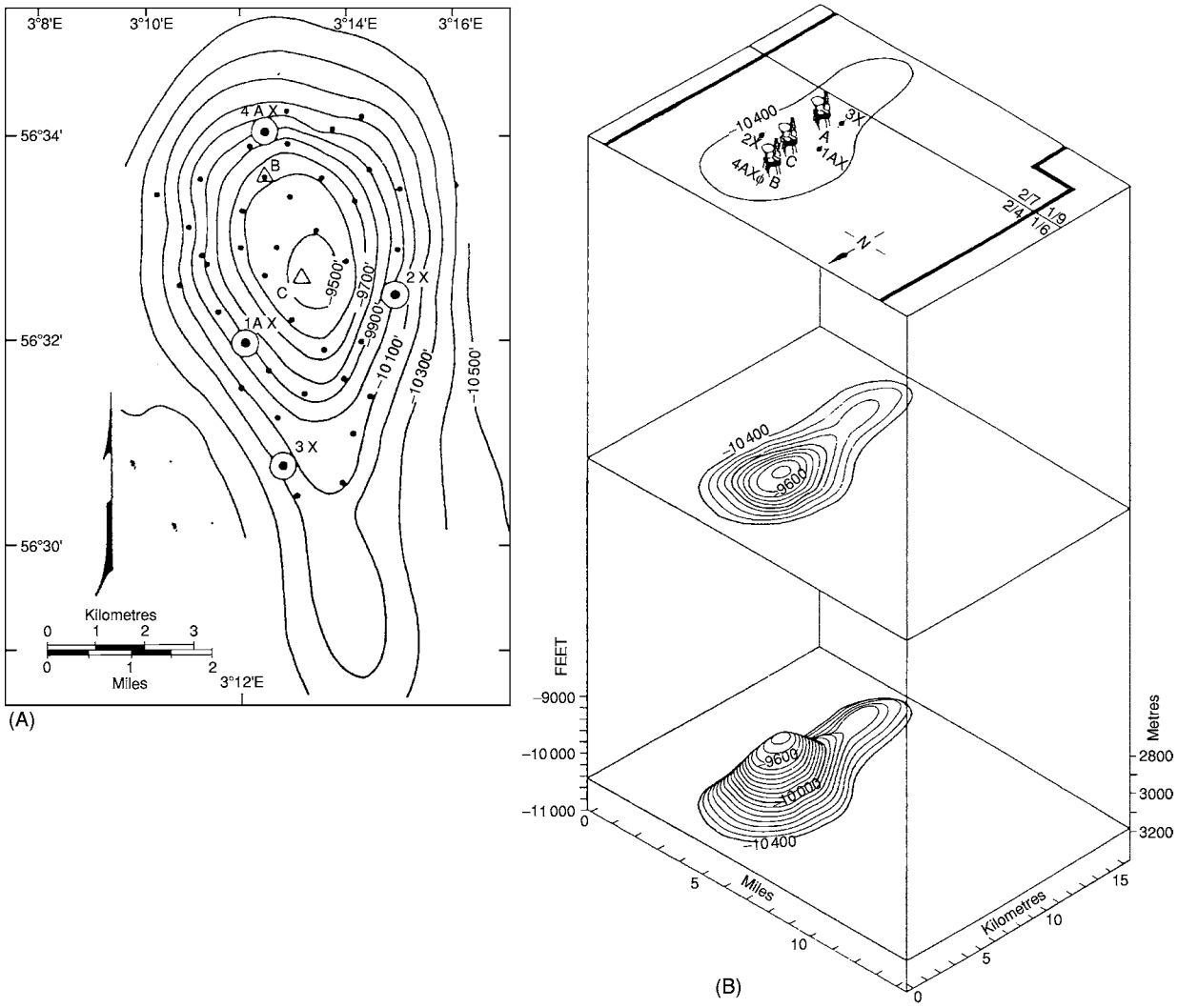


Figure 7 An example of a structure contour map: the Ekofisk oilfield, North Sea. (A) Structure contour map of the top of the oil bearing formation. (B) Oblique view of the form of the contoured surface (bottom level of drawing) and structure contours (drawn at an arbitrary level) and the sea bed (top level). Reproduced with modification from van der Bark and Thomas, 1980, American Association of Petroleum Geologists; used by permission.

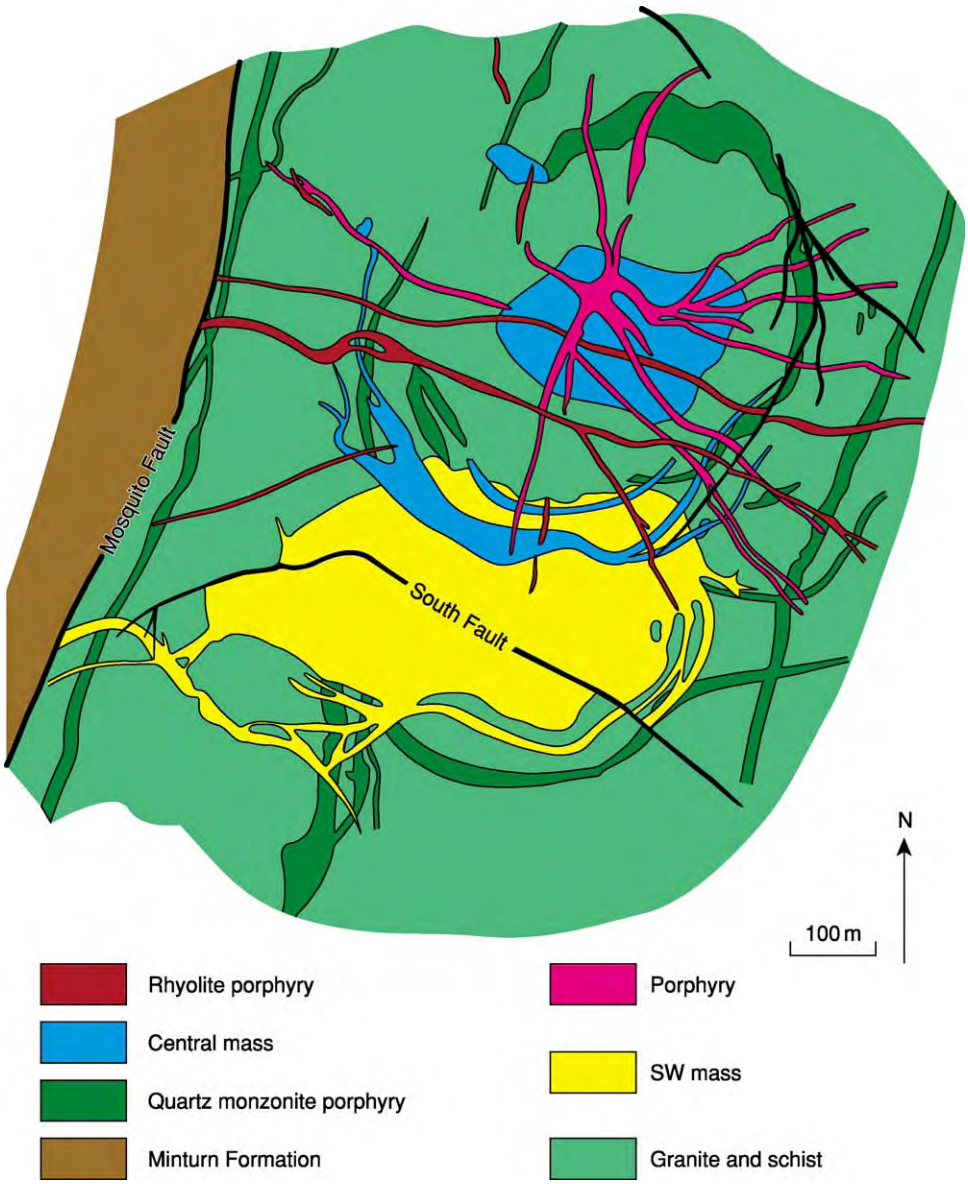


Figure 8 Map of the Climax molybdenum mine, Colorado, to illustrate the use of cross cutting relationships in determining geological histories from maps. The Mosquito Fault displaces all the rock units and therefore appears to represent the youngest geological event. The dykes of rhyolite porphyry represent the youngest rocks E of the Mosquito Fault as they cross cut all the units there; their relationship with the Minturn Formation, only seen to the W of the fault, is unclear from the map. Further cross cutting relationships indicate that the rocks are progressively older in the following sequence: porphyry, the ''central mass'' of the Climax igneous complex, the SW mass, and quartz monzonite porphyry. The latter two units are displaced by the South Fault, which must therefore be younger than the time of their intrusion. All the igneous rocks are intruded into granite and schist, which is therefore the oldest unit of these units. Because no igneous rocks are shown in the contiguous Minturn Formation, interpretation of the map suggests it is the youngest unit. (In fact the Minturn Formation does contain related igneous rocks, and these have been dated as Tertiary. The Minturn Formation is Pennsylvanian (Upper Carboniferous) and the granite/schist host rocks are Precambrian in age.) Adapted from a map by Wallace, Dahl and others, Graton Sales Volume, AIMME, 1968, used by permission.

Of course, the structure contours themselves have to be interpreted from control points such as boreholes or any sites where the formation of interest is exposed (at a known altitude), and their reliability varies with the density of control points available. If such control data can be derived from a geological map (e.g., by noting where outcrop traces cross topographic contours) it is possible for the map reader to construct structure contours for the surface of interest. A minimum of three points is needed. The construction is common on maps produced for specialised purposes – depth to bedrock, for example,

in engineering applications (*see* **Engineering Geology:** Geological Maps), or gravel thickness in the assessment of aggregates – but some general geological maps include structure contours for selected underground surfaces, to give a more quantitative picture of the subsurface arrangements. Developed oilfields or mineral deposits are likely to have a high density of well information and hence can yield structure contours of great accuracy.

The principles outlined above, utilising geometric surfaces, apply best to tabular sedimentary materials rather than other forms of rocks. Even so, extrapolations of depths and thicknesses of sediments have to be handled cautiously where the values are known to vary laterally, and this is particularly true of superficial sedimentary deposits. Because many such materials formed on land, where the depositional processes are much more localized than submarine sedimentation and where they are soon vulnerable to erosion, they characteristically lack lateral persistence. Igneous materials of tabular form, such as lava flows, sills, and dykes, can be interpreted using the basic geometric ideas, as can tabular mineral deposits. Less regular bodies will be difficult to interpret geometrically. Metamorphism does not itself affect application of the three-dimensional principles, but because it is commonly accompanied by substantial deformation, the configurations of metamorphic rocks can be complicated and more difficult to interpret from maps.

Geological Histories

In addition to the sub-surface extrapolations outlined above, another important interpretive aspect of geological maps is the reconstructing of geological histories. From the relative ages of formations normally indicated in the key, together with principles such as that a feature cross-cutting another must be the younger of the two, it is possible to erect a sequence of the geological events in the history of an area (Figure 8). By then linking in the lithologies and the conditions under which they must have formed, it becomes possible to reconstruct past geological circumstances. In the case of sedimentary rocks, for example, past depositional environments and their evolution through time can be deduced. It may even be possible to create a geological map for the land surface as it was at some past time – what is properly called a palaeogeological map. Such historical aspects are not only of academic interest but find applications in the petroleum industry, for example, where the linking of geological with palaeogeological maps, together with cross-sections, structure contour and isopachyte maps, is routinely used in hydrocarbon exploration and extraction. Such manipulations are these days largely carried out by computers, but a geologist familiar with the principles of geological maps still has to evaluate the output, and to be alert to shortcomings of the software and any consequent unnatural results.

Applications of Geological Maps

In ways mentioned above, geological maps find applications in different spheres of academic and commercial geology, and a host of related disciplines. Geological maps are relevant, to mention just a few examples, in civil engineering, hydrogeology, and archaeology. They help understand the landscape around us, both physical aspects such as landforms, slope instabilities, and building materials, and more human aspects, such as the soils derived from the bedrock and hence the crops that are viable, and settlement characteristics – in some cases even the place names. A recent development is the rise of geological maps for environmental planning, in which the geological aspects of an area relevant to, say, subsidence problems, landfill sites, contaminated ground, and radon concentrations are highlighted. Some national surveys (*see* **Geological Surveys**) now produce folios of such maps, each of which emphasises a particular environmental parameter, in an effort to foster sensible forward planning. Finding such wide usage, geological maps are so much more than mere plots of areal distributions.

See Also

Engineering Geology: Geological Maps. **Geological Field Mapping**. **Geological Surveys**. **Remote Sensing:** GIS. **Tectonics:** Faults; Folding. **Unconformities**.

Further Reading

Maltman A (1998) *Geological Maps: an Introduction*, 2nd edn. Chichester: John Wiley and Sons.

Soller DR (ed.) (2001) *Digital Mapping Techniques '01 Workshop Proceedings*. US Geological Survey Open File Report 01 223. http://pubs.usgs.gov/of/2001/of01 223/

United States Geological Survey (2002) *Three Dimensional Geologic Maps and Visualization*. http://3d.wr.usgs.gov/docs/wgmt/3d/pp02.html

GEOLOGICAL SOCIETIES

G L Herries Davies, University of Dublin, Dublin, Ireland

Introduction

The development of modern geology has brought into being many types of geological institutions, ranging in character from geological surveys and mineral exploration companies to geological museums and departments of geology within seats of higher learning. But among all these types of institutions, a very special and significant niche is occupied by the world's geological societies. The seventeenth and eighteenth centuries saw the establishment of many a society dedicated to the advancement of what had been termed 'the New Learning'. Of these, the most famed was – and is – the Royal Society of London for the Promotion of Natural Knowledge, which received its charter from King Charles II on 15 July 1662. Phenomena such as rocks, minerals, fossils, landforms, and Earth history were all subjects that engaged the eager attention of the devotees of the New Learning. In the Royal Society's *Philosophical Transactions* number 52 (17 October 1669) is an account of minerals found following an eruption at Mount Etna, and in number 76 of the same periodical (22 October 1671), Martin Lister (1638?–1712) writes from York on the subject of local fossils. Similarly, the *Histoire de l'Académie Royale des Sciences* for 1771 (published in 1774) contains a paper on the volcanoes and basalts of the Auvergne by Nicolas Desmarest (1725–1815). In 1788, the Royal Society of Edinburgh (founded 1783) published in its *Transactions* the famed 'Theory of the Earth; or an Investigation of the Laws Observable in the Composition, Dissolution, and Restoration of Land upon the Globe', by James Hutton (1726–97). In 1799, the American Philosophical Society Held at Philadelphia for Promoting Useful Knowledge (founded 1769) published within its own *Transactions* a paper on fossil bones from Virginia by its president, Thomas Jefferson (1743–1826), who shortly was to become the third president of the United States.

Similar eighteenth-century societies made various other types of contributions to the development of the Earth sciences. Within the British Isles, the Dublin Society for Improving Husbandry, Manufactures, and other Useful Arts and Sciences (founded 1731; it assumed the title 'Royal' in 1820), between 1801 and 1832, published statistical surveys of 23 of Ireland's 32 counties. Several of these surveys, especially the 1802 survey of County Kilkenny by William Tighe (1766–1816), contain comprehensive essays in regional geology. In 1808, the same society offered a premium of £200 (it was never claimed) for a geological map of County Dublin, and in 1809, the society commissioned a geological survey of the Leinster coalfields. In London, the Society for the Encouragement of Arts, Manufactures, and Commerce (founded 1754; styled the Royal Society of Arts after 1908) offered a gold medal or a premium of 50 guineas for the first mineralogical map of England and Wales, of Scotland, and of Ireland. The offer was readvertised annually until February 1815, when William Smith (1769–1839) claimed the premium for his pioneering geological map of England and Wales. Again in London, at the Royal Institution (founded 1799), the professor of chemistry, (Sir) Humphry Davy (1778–1829), in 1805 and in subsequent years, delivered a course of lectures in geology; in support of those prelections, the Institution established a comprehensive geological museum.

The foundation of learned societies of a generalist nature continued in Britain down to the close of the eighteenth century. The Lunar Society of Birmingham was founded in 1775; the Manchester Literary and Philosophical Society, in 1781; the Royal Irish Academy, in 1785; and the Literary and Philosophical Society of Newcastle-upon-Tyne, in 1793. All of these societies brought geology with their purview, but as science progressed, there came into being specialists who were interested in the foundation of societies dedicated exclusively to individual sciences. In London, for instance, there were founded the Medical Society of London (1773), the Linnean Society of London (1788), the short-lived British Mineralogical Society (1799), the Horticultural Society (1804; it was granted the title 'Royal' in 1861), the Institution of Civil Engineers (1818), the Astronomical Society (1820; it received the title 'Royal' with its Charter of 1831), and the Zoological Society of London (1826).

The British Mineralogical Society was dissolved in December 1806 after a life of less than 8 years. One of the society's members had been the eminent London physician William Babington (1756–1833), and following the society's demise, a group of its former members, together with other gentlemen, began to hold informal scientific meetings at Babington's London home, which, from 1807, was at 17 Aldermanbury. Most of the meetings were held over breakfast, before the start of Babington's busy

daily medical rounds, but when the distinguished Humphry Davy joined the group – he termed his companions 'the Geophilists' – he urged that future meetings should be held following a dinner arranged within the convivial atmosphere of some favourite hostelry. His proposal found favour. A dinner was arranged to be held at the Freemasons' Tavern, in London's Great Queen Street, on 13 November 1807. Eleven gentlemen attended. One of them was Davy. He expected to see the Geophilists transformed into a properly constituted dining club (what he termed 'a little talking Geological Dinner Club'), but what he actually witnessed that evening was an event of a very different order of magnitude. He saw founded the Geological Society of London. It was the world's first society to be devoted exclusively to the science of geology. The resolution adopted that November evening read as follows:

> That there be forthwith instituted a Geological Society for the purpose of making Geologists acquainted with each other, of stimulating their zeal, of inducing them to adopt one nomenclature, of facilitating the communi cation of new facts, and of ascertaining what is known in their science, & what remains to be discovered.

Sir Joseph Banks (1743–1820), the president of the Royal Society, asked to be enrolled within the new body, but he soon had second thoughts. He had believed himself to be joining a geological dining club; he now discovered that the former Geophilists were aspiring to become a full-blown scientific society. For over a 100 years, geology had been a constituent of the Royal Society's empire of science. Now this upstart society seemed to be laying claim to geology as its own special preserve. Further, for a loyal Fellow of the Royal Society to become a member of the Geological Society was then seen as akin to being guilty of scientific bigamy. Banks, Davy, and several others who in all innocence had joined the new society promptly submitted their resignations.

Geological Society of London

Despite so inauspicious a beginning, the Geological Society of London prospered. Geology flourished: it was the most 'popular' science of the greater part of the nineteenth century. And for much of that century, the Geological Society of London was a cynosure for the world's geological eyes. The society had reached a membership of 450 by 1825, the year in which the society was granted the Royal Charter of Incorporation, under which its members were to be transformed into 'Fellows'. Honorary members were elected from all over the British Isles during the society's early years; from 1814, overseas geologists were elected as foreign members, and in 1863, there was added the additional category of foreign correspondent. The society first acquired its own premises in 1808; in 1828, the government provided the society with rent-free accommodation, first at Somerset House (1828–74) and then at Burlington House (1874–present). By its centenary in 1907, the society possessed a total of 1356 Fellows and overseas members.

From the outset, a prime function of the society was the holding of regular meetings at which papers were to be read and discussed. The first two papers, by Richard Knight (1768–1844) and Jacques-Louis, Comte de Bournon (1751–1825), were read following a dinner held at the Freemasons' Tavern on 4 December 1807, and since then tens of thousands of papers have been presented before the society. On 5 March 1824, the society received its first communication from a woman, Mrs Maria Graham (later Lady Maria Callcott; 1785–1842), and from 1887, the society received many papers from female authors. It was nevertheless March 1919 before the society finally resolved to open its fellowship to women geologists. The first woman to hold the presidency of the society, between 1982 and 1984, was Janet Vida Watson (1923–85).

The young society set about the collection of geological information out of which, it was hoped, there might emerge a true theory of the earth. To assist geological observers, the society in 1808 compiled and circulated a questionnaire entitled *Geological Inquiries*. This 20-page pamphlet was widely noted and was even reprinted in the United States within the *American Mineralogical Journal*. As a further manifestation of its desire to assemble and use geological information, the society in 1808 began the compilation of a geological map of England and Wales. This task was carried out under the direction of George Bellas Greenough (1778–1855), the society's first president (1807–13). The map was published, at a scale of 1 inch to 6 miles (1:380 160), on 1 May 1820 (not 1 November 1819, as recorded on the map), and subsequent editions were published by the society in 1840 and 1865.

The society in 1808 founded a museum to contain rocks, minerals, and fossils from around the world, but unsolicited gifts eventually converted the museum into an incubus. Inadequately housed and curated, the collection was disbanded in 1911, part of it being given to the Museum of Practical Geology in London's Jermyn Street (now the Museum of the Geological Survey at Keyworth, Nottinghamshire) and the remainder going to the British Museum (Natural History) (now the Natural History Museum in London's Cromwell Road). Happier is the story

of the society's library. Coeval with the society, by 1826, the library contained 1072 volumes and many pamphlets and maps, and substantial annual growth has continued to the present day. Today the library houses one of the world's most comprehensive collections of geological literature. The library is much consulted by the Fellows and by a wide range of individuals having need of geological information. The library is the society's proudest possession.

The great majority of papers read to the society have been published in one or other of the society's periodicals. These include the *Transactions of the Geological Society of London* (1811–56), the *Proceedings of the Geological Society of London* (1826–45), the *Quarterly Journal of the Geological Society of London* (1845–1970), and, since 1971, the *Journal of the Geological Society*, which appears bimonthly. Among the thousands of papers published by the society, highlights include the work by William Henry Fitton (1780–1861) on the geology of the opposed coasts of France and England (1826); (Sir) Henry De La Beche (1796–1855) on the geology of Jamaica (1827); Louis Agassiz (1807–73) on the glacial theory (1840); Henry Clifton Sorby (1826–1908) on the microscopical structure of Yorkshire grits (1850); Joseph Beete Jukes (1811–69) on the geomorphology of the south of Ireland (1862); Charles Lapworth (1842–1920) on the Moffat Series (1877); Benjamin Neeve Peach (1842–1926) and colleagues on the northwest highlands of Scotland (1888); Arthur Vaughan (1868–1915) on the stratigraphy of the Bristol region (1904); (Sir) Edward Battersby Bailey (1881–1965) and colleagues on the cauldron subsidence of Glen Coe (1909); Owen Thomas Jones (1878–1967) on the evolution of a geosyncline (1938); and John Frederick Dewey (1937–present) on the development of the South Mayo Trough (1962).

Since 1831, the society has added another important spoke to the wheel of its activities, by assuming a leading role within the reward system of the international community of the earth sciences. This was first possible when, on 10 December 1828, the society's council learned that William Hyde Wollaston (1766–1828), a mineralogist and a member and Fellow of the society since 1812, had left to the society the sum of £1000, the income from the investment to be used "in aiding or rewarding the researches of any individual or individuals, of any country". Further, Wollaston enjoined the society not to hoard the income, but to strive to make an award every year. The council resolved to bring into being a medal to be struck in gold (the medal has also been struck in palladium, a metal discovered by Wollaston), and the first of the long line of Wollaston Medals was awarded to William Smith on 18 February 1831. Since 1835, one or more Wollaston Medals have been awarded every year, and the medal is today recognized as the premier award within the world of geology. The roll of the medal's recipients glitters with distinction, as the following sample of recipients reveals: Louis Agassiz, 1836; Charles Darwin, 1859; Sir Charles Lyell, 1866; James Dwight Dana, 1872; Eduard Suess, 1896; Grove Karl Gilbert, 1900; Albert Heim, 1904; Baron Gerard Jacob de Geer, 1920; Reginald Aldworth Daly, 1942; Arthur Holmes, 1956; Alfred Sherwood Romer, 1973; John Tuzo Wilson, 1978. When Sir Roderick Impey Murchison (1792–1871) died, he left to the society a bequest sufficient to endow the annual award of a Murchison Medal, and at the death of Sir Charles Lyell (1797–1875), a similar bequest allowed the endowment of a Lyell Medal. Today the Wollaston, Murchison, and Lyell medals, together with sundry other younger awards placed within the gift of the society, are all made at the President's Evening of the society held early in May each year.

Other Geological Societies

The young Geological Society of London aspired to a standing that was national, if not international, but it speedily became the prototype for other more local geological societies that soon began to arise in many parts of the British Isles. Several of these societies of lesser ambition were rooted in regions where a nearby mining industry imparted to geology an especial significance. The following represent the principal British and Irish geological societies founded during the nineteenth century, in imitation of the Geological Society of London: The Royal Geological Society of Cornwall (1814), the Geological Society of Dublin (1831; restyled the Royal Geological Society of Ireland in 1864), the Edinburgh Geological Society (1834), the Geological and Polytechnic Society of the West Riding of Yorkshire (1837; after 1877, the Yorkshire Geological and Polytechnic Society; after 1905, the Yorkshire Geological Society), the Manchester Geological Society (1838; after 1903, the Manchester Geological and Mining Society), the Dudley and Midland Geological Society (1842; refounded as the Dudley and Midland Geological and Scientific Society and Field Club in 1862), the Geologists' Association (1858), the Geological Society of Glasgow (1858), the Liverpool Geological Society (1859), the Norwich Geological Society (1864), and the Hull Geological Society (1888). One other British foundation merits mention. At York in 1831, there was established the British Association for the Advancement of Science, based, somewhat, on the model of the German Gesellschaft Deutscher

Naturforscher und Ärzte (1822). Each year since 1831, with some interruptions during the two world wars, the 'B.A.', or the 'British Ass', has held its annual peripatetic gatherings in cities throughout the British Isles and in Commonwealth locations such as Montreal (1884), Toronto (1897, 1924), South Africa (1905, 1929), Winnipeg (1909), and Australia (1914). The Geological Section – today Section C – of the Association has always been one of the most important, successful, and popular of the Association's numerous sections.

The geological societies named so far have all been bodies dedicated to the earth sciences in general, but since the middle decades of the nineteenth century, increasing specialization within the earth sciences has encouraged the development of a new generation of societies dedicated to just one field within the earth sciences. In Britain, the earliest of these more narrowly focused societies was the small London Clay Club, established in 1838 for the collection, description, and illustration of the local Eocene mollusca. Considerably more important is the Palaeontographical Society, founded following a meeting of the Geological Society of London held on 3 February 1847, and dedicated to the publication of monographs devoted to British fossils. Other specialized British bodies founded within the field of the earth sciences include the Mineralogical Society (1876), the Institution of Mining and Metallurgy (1892), the Institution of Petroleum Technologists (1913), the Palaeontological Association (1957), the British Geomorphological Research Group (1961), the British Micropalaeontological Society (1970), and the Geological Curator's Group (1974). Within the Geological Society of London there are now specialist groups, of which the following organizations serve as a sample: the British Geophysical Association, the British Sedimentological Research Group, the Environmental and Industrial Geophysics Group, the Geochemistry Group, the Geological Remote Sensing Group, the Geoscience Information Group, the Hydrogeological Group, the Marine Studies Group, the Metamorphic Studies Group, the Petroleum Group, the Tectonic Studies Group, and the Volcanic and Magmatic Studies Group. The *Quarterly Journal of Engineering Geology* (now the *Quarterly Journal of Engineering Geology and Hydrogeology*) has been published by the Geological Society since September 1967.

During the past 150 years, the pattern of geological society foundation, evident in Britain since 1807, has been mirrored the world over as the global population of earth scientists has undergone dramatic expansion. The increasing interest in higher education, the growth of the hydrocarbon industry, the bourgeoning demand for industrial minerals, the incessant call for building materials, and the world's never-to-be-assuaged thirst for water have all proved to be the detonators of a geological population explosion. It has been an explosion such as could never have been imagined by those 11 founders of the Geological Society of London who dined together at the Freemasons' Tavern on 13 November 1807. In six continents, thousands of geologists have combined within the convenient and comfortable ambiance of geological societies. There geologists have sought intellectual companionship, there they have relished the inspiration of a geological milieu, there, for their research discoveries, they have found both fora for discussion and channels for communication, and there, by their labours, they have earned for their science that public prestige and influence that comes from nicely modulated collective activity. Further, in recent years, large numbers of enthusiastic amateurs have been attracted into geological societies. Never since the furore surrounding Darwin's *Origin of Species* (1859) have the earth sciences attracted as much attention as is theirs today. Plate collision, mass extinction, *Jurassic Park*, Creationist claims, lunar rocks, Martian images – these have all served as eye-catching billboards for geology. As a result, internationally, droves of amateurs have joined the ranks of the world's geological societies, there to make absorbing contact with a fundamental element of the human environment.

The Republic of Ireland has a population of only 3.9 million people, but it affords a microcosm of the type of recent institutional development that has characterized geology throughout the world. The Royal Geological Society of Ireland, dating back to 1831, died in 1894, partly as a result of the completion of the primary geological survey of Ireland in 1890, and partly as a result of the collapse of the small Irish mining industry as new and more profitable mines were opened overseas. With the second half of the twentieth century came revival. The Irish Mining and Quarrying Society was established in 1958 for those involved in the renewed and expanded extractive industries. The Irish Geological Association was launched in 1959 to offer a regular programme of events of interest to both the professional and the amateur. The Irish Association for Economic Geology was founded in 1973 to serve all those with a professional involvement in the economic aspects of the earth sciences. In 1978, there was founded a new journal, the *Journal of Earth Sciences Royal Dublin Society* (today the *Irish Journal of Earth Sciences* of the Royal Irish Academy). The Cork Geological Association was born in 1992 out of the enthusiasm of some Cork citizens

who had just attended an evening diploma course at the local University College. Finally, in 1996, there was established a Mining History Society of Ireland (since 2000, the Mining Heritage Trust of Ireland) "for all those persons interested in Ireland's historical mining industry".

The Irish societies are little more than dust on the international scene of geology. But on that same scene, other societies are massive landmark boulders, visible from all quarters of the geological world. These are the major societies that have played no little part in giving to the earth sciences their present character. Among those societies, pride of place must be accorded to the Geological Society of London. As the world's senior such society, it sometimes emphasizes its standing by strengthening the definite article and by dropping the geographical designation, to become simply 'the' Geological Society. It has been hugely influential within geology. At its centenary celebrations in 1907, the geological world paid a tribute to the mother of all geological societies, and many another tribute will doubtless be proffered at the society's approaching bicentenary in 2007.

Among the world's other notable societies are the Société Géologique de France (1830), the Società Geologica Italiana (1881), the Geological Society of South Africa (1895), the Palaeontological Society (1909), the American Association of Petroleum Geologists (1917), the International Quaternary Association (1928), the Geological Society of Australia (1952), the Sociedad Venezolana de Geólogos (1955), the Geological Society of India (1958), and the Nepal Geological Society (1980). Finally, among these leading societies, there is the Geological Society of America (1888). With a membership of around 20 000 geologists, it is the world's largest such society. The *Bulletin of the Geological Society of America* (1890) is one of the world's most significant geological journals, and the annual conference of the society is a major highlight in the calendar of international geology. The headquarters of the society was built in 1972 in Boulder, Colorado, at 3300 Penrose Place, named after Richard Alexander Fullerton Penrose (1863–1931), who was a major benefactor of the society. In Boulder, a fascinating location, both the grounds and headquarters building are replete with features of geological interest. Even the handles on the main doors are shaped in Scandinavian labradorite!

Although not a geological society in the strictest sense, one other geological institution, the International Geological Congress, does here merit mention. The story of the Congress begins in the United States. In 1876, the American Association for the Advancement of Science was meeting in Buffalo, New York, when there crystallized the notion that regular international gatherings of geologists might prove of benefit to their science. More specifically, the Buffalo resolution referred to such gatherings as being "for the purpose of getting together comparative collections, maps, and sections, and for the settling of many obscure points relating to geological classification and nomenclature". A successful congress was held in Paris in 1878, during the Paris Exposition Universelle, with 312 geologists present from 22 different countries. There followed seven other nineteenth-century congresses, at Bologna (1881), Berlin (1885; a congress planned for 1884 was abandoned because of an outbreak of cholera in southern Europe), London (1888), Washington (1891), Zurich (1894), St Petersburg (1897), and Paris (1900). Since 1900, 24 additional congresses have been held, although sometimes international events have interfered with the measured regularity of the congress procession. Between 1913 (Toronto) and 1922 (Brussels), World War I prevented the holding of any congresses, and the 1940 congress, scheduled for London, was cancelled on the outbreak of World War II in 1939. Most dramatic of all, the 1968 congress in Prague had to be abandoned on its second day as a result of the Soviet invasion of Czechoslovakia on the evening of 20 August 1968. That congress, like its predecessor in New Delhi (1964) and all subsequent congresses, was organized under the aegis of the International Union of Geological Sciences founded during the course of the Norden congress of 1960.

Further Reading

Eckel EB (1982) *The Geological Society of America: Life History of a Learned Society.* Boulder, CO: Geological Society of America.

Schneer CJ (1995) The geologists at Prague. History of the International Union of Geological Sciences. *Earth Sciences History* 14(2): 172 201.

Société Géologique de France (1930) *Centenaire de la Soci été Géologique de France: Livre Jubilaire 1830 1930.* Paris: Société Géologique de France.

Woodward HB (1907) *The History of the Geological Society of London.* London: Geological Society.

GEOLOGICAL SURVEYS

P M Allen, Bingham, Nottingham, UK

Introduction

'Geological survey' is the generic term used for the government organizations that carry out geological mapping; in addition a survey may also provide other basic geoscientific information and services to government, industry, commerce, and the general public. In the English-speaking world, survey organizations are often named simply 'Geological Survey', with country name as some part of the formal organization title. Elsewhere, the literal translation of the organization name into English may also be 'Geological Survey'; for example, the Sveriges Geologiska Undersökning (SGU) is the Swedish Geological Survey. Worldwide, however, there is an enormous variety in organization names (Table 1) and most geological entities have undergone several name changes during their history.

The first record of a geologist being employed by government was in late sixteenth-century Russia. In the late eighteenth and early nineteenth centuries, the French government and some of the states in the USA employed geologists specifically to make geological maps. This was stimulated by a need for information to support agriculture and mineral exploration, but when the first modern geological survey, the Geological Survey of Great Britain (GSGB), was founded in 1835, their remit had broadened. Now, most countries and many states, provinces, and regions have an organization that is recognized by international bodies as the representative governmental geological survey group. In some poor countries, particularly in Africa, the geological surveys are effectively moribund or are sustained by foreign aid. Nearly all geological surveys are funded by and are part of government. There is a common core element to their work. For most, their main purpose has been and remains to support the minerals industry, but among the more mature surveys this role is reducing as issues relating to the environment become more important in society.

The Work of Geological Surveys

The principal objective of geological surveys is to collect and interpret geoscientific data and put the information into the public domain. The work falls within five essential categories:

- Geological mapping on land and the continental shelf. In many countries, geochemical, geophysical, and hydrogeological (groundwater) mapping projects are also carried out by the survey.
- Appraisal and assessment of mineral, energy, water, and land resources.
- Baseline studies on geohazards (volcanic, seismic, and ground stability) and the environment.
- Maintenance of a geoscientific archive or database.
- Publication of maps, books, reports, and data packages about the survey work.

As a secondary objective, some geological surveys carry out work commissioned by government departments outside the parent organization, or by the private sector, which will pay for it at the market price.

In the USA, Canada, and several colonial African countries, among others, many geological surveys had to make topographic maps as a base for their geological maps. In most cases, topographic mapping is now done in another part of government, but the United States Geological Survey (USGS), still retains responsibility for it as well as for biology, hydrology, and hydrogeology. All geological surveys do geological mapping, but not all carry out work in all the other categories. Responsibility for groundwater resources is often vested in a separate body within government. Where hydrocarbons are important to the economy, they are sometimes excluded from the remit of the geological survey. A typical example is the Division of Geological and Geophysical Surveys (DGGS) in Alaska's Department of Natural Resources. It has no responsibility for oil and gas or even basin analysis, which are covered by another division in the state government.

Many geological surveys carry out activities beyond those regarded as essential. Some operate a minerals bureau, both issuing exploration licences and collecting production statistics. They carry out research in mining, mineral dressing, and metallurgy and may provide support for small-scale mining activities. Others, particularly state surveys in the USA and Italy, have an educational function. The Geological Survey of Zambia has a gemmology unit. Geological surveys in Indonesia, New Zealand, Japan, the USA, Italy, central America, and the Andean countries all have specialist volcanological units and carry out research in earthquake hazard. The British Geological Survey is one of a small number of surveys that does research in geomagnetism. Several surveys carry out overseas aid programmes and some, such as the Geological Survey of

Table 1 National and federal geological surveys founded in the nineteenth century

Country	*Present name of geological survey*	*Year founded*	*Total staff* [a]
Austria	Geologische Bundesanstalt	1849	80
Belgium	Belgian Geological Survey	1896	17
Bulgaria	Directorate of Geology and Protection of Substrate	1880	25
Canada (federal)	Geological Survey of Canada	1842	667
Denmark	Geological Survey of Denmark and Greenland	1888	354
Egypt	Egyptian Geological Survey and Mining Authority	1896	2600
Finland	Geological Survey of Finland	1885	671
France	Bureau de Recherches Géologique et Minière	1868	848
Germany (federal)	Bundesanstalt fur Geowissenschaften und Rohstoffe	1873[b]	660
Hungary	Geological Institute of Hungary	1869	143
India	Geological Survey of India	1851	2900
Indonesia	Geology Research and Development Centre, Directorate General of Geology and Mineral Resources	1850[c]	N/A
Ireland	Geological Survey of Ireland	1845	51
Italy	Agency for Environment Protection and for Technical Surveys	1867	86
Japan	Geological Survey of Japan/AIST	1882	c300
New Zealand	Institute of Geological and Nuclear Sciences Limited	1867	258
Norway	Geological Survey of Norway	1858	198
Philippines	Lands Geological Survey Division, Mines & Geosciences Bureau	1886	394[d]
Portugal	Instituto Geologico e Mineiro	1857	290
Russia[e]	Geological Institute, Russian Academy of Sciences	1883	N/A
Spain	Instituto Technológico GeoMinero de España	1849	335
Sweden	Swedish Geological Survey	1858	268
Switzerland	Swiss National Hydrological and Geological Survey	1872	16
United Kingdom	British Geological Survey	1835	815
United States of America	The United States Geological Survey	1879	1600[f]

[a] In 2002, unless noted otherwise.
[b] The Prussian Commission of Surveying, the predecessor organization of the Prussian (then German) Geological Survey, was founded in 1841.
[c] The first government funded geological research was initiated by the Dutch colonial government in 'Dienst van het Mijnwezen' based in Bogor, in 1850. The organization moved to Jakarta in 1869 and to Bandung in 1924, undergoing name changes each time.
[d] Total staff in 1994.
[e] The current organization in Russia, founded in 1883, can trace its predecessor organizations back to 1584.
[f] In July 2003.

India and the USGS, have work programmes in Antarctica. The Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) in Germany runs a research ship and carries out geoscientific research in the world's oceans.

Types of Geological Survey

Geological surveys are conducted at three levels, national, state, and federal.

National Geological Surveys

National geological surveys are solely responsible within their country for the national geological mapping programme. In most countries, the national geological survey will be the only geoscientific institution at any level of government that carries out geological mapping. Countries with strong regional, state, or provincial governments may have geological surveys that are administered at that lower level of government, but, except in certain federal nations, overall control of the national geological mapping programme is still retained at the national level. This applies in Brazil, where some provinces have their own geological surveys, but the national survey has offices in all the provinces and controls the mapping programme centrally. In Italy, all the regions have their own geological survey as a branch of their administration and fund them independently of the central government, but the national geological survey, the Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici (APAT), coordinates and provides funding for the national mapping programme, even though the work is done by the regional surveys or universities.

In some countries where there is a single, national geological survey, not all of the full range of survey activities are carried out by the main organization. In France, where the modern Bureau de Recherches Géologique et Minière (BRGM) did not come into being until 1959, other government bodies carry out geophysics and marine geoscience, and engineering geology is almost entirely done within universities.

In countries that were once part of the Soviet Union, or influenced by it, government-funded geoscientific research was always dispersed among several institutes, often in different academies of science, and it is not uncommon for geophysics and geochemistry to be done in separate institutes to geological mapping.

British Geological Survey The Geological Survey of Great Britain (GSGB), founded in 1835 as a branch of the Ordnance Survey, acquired independent status within Government and the right of access to private and public land for the purposes of making geological maps through Act of Parliament in 1845.

The Survey's aim, now nearly achieved, was to produce a uniform series of geological maps at the scale of one inch to the mile (now 1:50 000) for the whole of Great Britain. Coalfields and other areas of economic importance were resurveyed several times. From the early 1860s Survey geologists used maps at the scale of six inches to the mile (now 1:10 000) for field recording.

A core of field geologists was supported by petrographers, mineralogists, palaeontologists and chemists. The Museum of Practical Geology was part of the GSGB until 1984. The Survey first addressed matters relating to groundwater in 1872 and used geophysics to aid mapping from 1926. Staff in the two decades after 1945 was around 140.

In 1965 the GSGB was merged with the Overseas Geological Surveys to create the Institute of Geological Sciences (renamed the British Geological Survey in 1984), within the Natural Environment Research Council. Disciplines brought in from OGS included photogeology, isotope geology, applied mineralogy, mineral intelligence, economics and statistics, modern chemical laboratories, applied geophysics and responsibilities for mapping overseas. Geomagnetism and seismology units were also added at this time.

By 1980 major new programmes included mapping the continental shelf, bulk and metalliferous mineral resource evaluation, national geochemical and geophysical surveys, geothermal energy, offshore hydrocarbons assessment, deep geology, environmental pollution, radioactive waste disposal, geohazards and urban geology.

In 1971 the Rothschild Report, recommending that direct funding should be partly replaced by a customer/contractor relationship with government departments, was implemented. Direct funding fell sharply and the Survey's mapping programme, which no single department would fund, was reduced to a minimum.

The Butler Enquiry (March 1987) restored balance to the work programme. From 1990 direct funds were made available for a core programme of geoscientific surveys overseen by a Programme Board made up of representatives of the user community. Subsequent government enquiries led to progressive commercialisation of the BGS. By 2004 its income came from direct Government grants, research commissioned by Government departments, the European Union, research grants and commercial activity.

The Survey complement grew from 501 in 1965 to 1200 in 1984, declining to around 800 in 2004. Gross turnover exceeds £30 million annually.

State Geological Surveys

At state level, geological mapping and many of the other activities of a geological survey are carried out exclusively within the boundaries of a state, province, or region. Usually, state geological surveys are part of the administration of state, provincial, or regional levels of government and are funded by them. State-level bodies are found in countries that have a national geological survey, such as Italy and Brazil, as well as in countries with federal systems of government, such as Australia, Canada, Germany, and the USA, in which nearly all states and provinces have their own geological surveys.

Federal Geological Surveys

Federal geological surveys are funded by the federal government to provide a variety of umbrella functions, including the generation of nationwide coverage of small-scale geological maps, but they are not responsible for the systematic, nationwide, medium-scale geological mapping programme. The mapping is done by the geological surveys at the lower level of government. There are four federal geological surveys: the United States Geological Survey, Geoscience Australia, the Geological Survey of Canada, and the Bundesanstalt für Geowissenschaften und Rohstoffe in Germany. Typically, federal bodies provide specialist laboratory facilities and take responsibility for hydrocarbons research, maritime surveys in home and foreign waters, overseas activities, research that has a nationwide relevance, and the compilation of national, small-scale, overview maps.

United States of America Established within the Department of the Interior, the United States Geological Survey (USGS) was founded in 1879. Its remit included, "classification of public lands, and examination of the geological structure, mineral resources and products of the national domain." To carry it out the USGS then, as now, had to do both topographical and geological surveys.

The USGS started with a staff of 38 and a budget of $106 000. In its centenary year the available budget

was $765 million and there were 12 000 staff. Funding throughout its history has come from direct Government appropriation.

The Survey's programme was always practically based, but in the 1890s Director Charles D Walcott broadened its remit to include basic scientific research.

Originally concerned with mapping publicly owned mineral lands, the Survey's programme grew by responding to national and international economic and political pressures. It began its hydrographic programme in 1888 after a severe drought. The first staff sent overseas went to Nicaragua in 1897. Groundwater became a Survey interest in 1894; oil in 1901; coal and iron in 1905; strategic minerals in 1914. After a period of decline from 1920 to 1939, the USGS began a major period of expansion, partly stimulated by the Cold War. Its focus for geological mapping shifted to the search for radioactive minerals and to support engineering projects in 1950. By then there were geological maps for less than 10% of the country and the responsibility for systematic geological mapping was taken over by the State surveys. In 1956 the USGS began evaluating the impact of underground nuclear explosions. Extraterrestrial studies began in 1959 with a photogeological map of the moon. Antarctic exploration began in 1958; marine studies in 1962 and earthquake research after the Alaska earthquake in 1964, adding to existing research on volcanological and other natural hazards. Geothermal research began in 1974 after the oil crisis. From 1970 the USGS became increasingly involved in all aspects of environmental geology.

The USGS has suffered and benefited from political interference. There have been many reorganizations, culminating in major changes in the late 1990s following a threat to its existence. Now there are four major divisions: Biology; Geography; Geology and Water. The Survey's headquarters, in Washington D.C. up to 1973, are in Reston Virginia. There are regional offices at Menlo Park, California and Denver, Colorado; Earth Science Information Centers at Rollo and Sioux Falls, and representatives in all states.

Australia Government-appointed geologists were working in western Australia in 1847, but the first state survey in Australia was the Mines Department in Victoria, which by 1861 was producing geological maps. The federal survey, called the Bureau of Mineral Resources (BMR), based in Canberra, was not established until 1946. It formed by merging the Aerial Geological and Geophysical Survey of northern Australia (1935) and the Mineral Resources Survey (1942) with the aim to increase Australia's metal output. It took responsibility for geological mapping in northern Territory, Papua-New Guinea, and Canberra, where there were no geological surveys, and worked in conjunction with the geological surveys of Queensland and western Australia on their mapping programmes. The BMR subsequently changed its name to the Australian Geological Survey Organisation and later to Geoscience Australia. Both northern Australia and Papua-New Guinea now have their own geological surveys.

Canada The Geological Survey of Canada (GSC) was founded in 1842 to carry out a reconnaissance geological survey of Canada. It made the first topographic maps and carried out biological surveys. In 1934, restructuring made the GSC concentrate essentially on geological surveying. This task has now passed to the provincial surveys, though the GSC carries out geological mapping in provinces that do not have their own geological survey. The provincial surveys tend to work closely with the mining industry. Many were originally closely associated with the provincial universities, sharing staff with them. Among the first to form, in 1864, was the Geological Survey of Newfoundland. The youngest is the Yukon Geological Survey, founded in 2003.

Germany The German Geological Survey arose out of the Prussian state geological survey, which was founded by merging the Prussian Commission of Surveying with the Mining Academy and the Museum of Economic Geology in 1873. The Commission of Surveying had been carrying out geological mapping since 1841 and had established a standard scale of 1:25 000 for its work in 1866. The Mining Academy was removed from the survey in 1916. After the First World War, the German Geological Survey became more concerned with economic geology, especially oil exploration. From 1939, all the German state geological surveys came under central control, and throughout the war, the German Geological Survey attempted to coordinate geological surveying and oil exploration in all of the occupied countries. The structure collapsed in 1945, but was reconstituted by the British military government as the Geological Survey of the Federal Republic. Headquarters were established in Hannover, shared with the Geological Survey of Lower Saxony. Now it is called the Bundesanstalt für Geowissenschaften und Rohstoffe (the Federal Institute of Geoscience and Raw Materials). Before reunification, western Germany consisted of nine states, each with its own geological survey. The work at state level is distinctly different from that at the BGR, being concerned with strictly local geological matters, including geological mapping. In the German

Democratic Republic, there was a single geological survey, the Zentrales Geologisches Institut (ZGI), until reunification. Now there are five state geological surveys in the former East Germany.

The Start of Geological Surveys

The history of government funding for geology starts in 1584, when the Russian government set up a 'stone department' (*Kamenay prikaz*) to regulate and encourage private enterprise in prospecting for mineral deposits as it colonized Siberia. In 1700, Peter the Great transformed that body into the Department of Mining and later (1719), the Ministry of Mining. This ministry continued to organize geological research and mapping and in 1883 established a 'Geological Committee', modelled on the Geological Survey of Great Britain, to stimulate a national geological mapping programme. Under various names, this organization has continued to the present day. In France, government charged the Corps des Mines to gather geological information about the country in 1794 and authorized the preparation of a geological map of the country in 1822. It was not until 1868 that the Service de la Carte Géologique de la France, the precursor to the present BRGM, was founded to oversee the preparation of a national geological map.

In the United States, federal funding for geological work was made available in 1804 in connection with military expeditions to unexplored territories. In 1810, the Academy of Science in Connecticut approved the appointment of a geologist to map the state. Work began in 1820. Geological surveys were established in the Carolinas in 1823 and 1825 and in 14 other states between 1830 and 1839. In many cases, the stimulus for the foundation of the geological bodies was to provide support for agriculture, the mainstay of the US economy at that time. In 1834, the United States Congress attempted to establish a federal geological survey by permitting the Topographical Bureau of the US Army to spend money on a geological map of the United States. This was abandoned after 2 years and it was not until 1879 that the United States Geological Survey was founded.

In Great Britain, a geologist was appointed to the Trigonometrical Survey (Ordnance Survey) in England in 1814, and in 1826 in Ireland. As early as 1815, William Smith, in the memoir accompanying his geological map of England, listed beneficiaries ranging from agriculture to water supply to mining and quarrying. When the Geological Survey of Great Britain was founded in 1835, then, there was no single justification for it. The GSGB differed from all other similar institutions at that time. Small-scale geological maps of Great Britain, created independently of any government funding, were already in existence. Thus, the GSGB founder, Henry de la Beche, proposed to map the whole country systematically at the comparatively large scale of 1 inch to 1 mile. He also started a Museum of Practical Geology, a School of Mines (now the Royal School of Mines), an educational programme for the common people, and a Mining Records Office. None of these now remain within the field of responsibility of the British Geological Survey, but de la Beche's vision provided a model that was copied worldwide.

Later Developments

De la Beche's broad view had appeal throughout Europe and directly stimulated many countries to institute their own geological surveys on the GSGB model (Table 1). The German states of Prussia, Westfalia, and Rhine all started geological mapping programmes in 1841; Ireland followed in 1845, and Austria and Spain, in 1849. By the end of the nineteenth century, more than half the European countries had national geological surveys. Today, all except the small states (Andorra, San Marino, and Liechtenstein) have them, the most recent being Latvia, in 1995.

Throughout the nineteenth century, the demands of the British manufacturing economy and the developing industrial economies of Europe and North America led to many geological surveys being established with the purpose of providing geological maps for use in the exploration for natural resources for export. In the British Empire, geological surveys were founded in Canada in 1842, India in 1851, the state of Victoria in 1861, and the Cape of Good Hope, South Africa in 1895. Elsewhere in the British Empire, government-funded geological reconnaissance was carried out throughout the second half of the nineteenth century by geologists seconded from the GSGB. Geological surveys were established in several parts of the Empire in the first quarter of the twentieth century (*see* **Colonial Surveys**), but most came in a second wave in the 1940s and 1950s during the post-World War II drive to find resources for reconstruction. There is a similar picture among the French colonies. In Indochina, three organizations were established to provide geological information for what became Vietnam, Laos, and Cambodia: Service des Mines de la Cochinchine (1868), Service des Mines de l'Indochine (1884), and the Service Géologique de l'Indochine (1898). In Africa, however, the French colonial geological surveys were either small or nonexistent until after the Second World War, when considerable expansion took place for essentially the same reasons as in the

British Empire. The Dutch, Spanish, and Portuguese administrations all started geological surveying in their colonial possessions. In Indonesia, it was in 1850, 53 years before a national geological survey was established in The Netherlands. Outside Europe and its colonial possessions, most countries were late to found their own geological surveys. Japan (1882) was among the earliest; China's was set up in 1913. In central America, South America, and the Caribbean, where there are now over 20 geological surveys, those in Argentina (SEGEMAR) and Columbia (SERGEOMIN) were founded in 1904 and 1916, respectively, but nearly all the rest were established after 1940.

By the end of the twentieth century, nearly all countries had a geological survey and, outside Europe, many are still primarily concerned with supporting the mining industry. Though geological mapping remains an essential activity for all surveys, the focus of attention in the more mature surveys has now shifted away from providing maps and other information to support mineral exploration and exploitation to environmental issues such as groundwater supply and quality, rehabilitation of 'brown' land, and on providing information on undermined areas, ground stability, and geohazard.

Geological Mapping

There is a natural progression in the way in which geological mapping (*see* **Geological Field Mapping**) is carried out in geological surveys:

- Stage 1 is reconnaissance, to find out what is there and to construct a small-scale map for the whole territory. Some of the early state surveys in the United States were closed down after this stage was completed. The scale of the reconnaissance map is small. The first map of the whole of the USSR, completed in 1956, and in China, completed in 1999, was 1:1 million. In France, the first geological map, published in 1841, was at 1:500 000, as was the first in Portugal, in 1876. In Czechoslovakia, the first map, completed in 1960, was scaled at 1:200 000.
- Stage 2 is the first, detailed, systematic survey. Its purpose is to generate medium-scale geological maps that fit into a regular grid pattern covering the whole of the territory. The scale chosen for this stage was 1:200 000 in China and 1:100 000 in Argentina. Most countries, even some large ones such as India, Pakistan, and South Africa chose a scale of 1:50 000 for complete, detailed national cover. This is the preferred scale for most European countries. Some state surveys, including the Prussian Geological Survey and the Ontario Geological Survey, have chosen 1:25 000 or a combination of 1:25 000 and 1:50 000. In some American states, medium-scale maps are drawn at 1:24 000.
- Stage 3 is a systematic, but local, large-scale survey. Because this mapping is demand driven, it will not normally lead to full, national map cover, but will be confined to areas of particular economic importance. In Great Britain, the chosen scale is 1:10 000; in France, 1:20 000; in Russia, either 1:50 000 or 1:25 000; and in India, 1:25 000. In many countries, urban areas are mapped at 1:5000.

The Size of Geological Surveys

Among the smallest geological surveys are the Service Géologique du Luxembourg and the Planning Directorate in Malta, both with a staff of eight. The largest was, and remains, the China Geological Survey, which in 1994 employed 1.1 million; in 1999, it was reorganized and staff was reduced to 6500. Much of the work the previous organization carried out was devolved to commercial enterprises in the provinces. After the breakup of the Soviet Union, many large eastern European geological surveys underwent drastic staff reductions, mostly accomplished by reducing the numbers of non-scientific staff. The most savage reduction was in the Bulgarian geological survey, which in 2001 was 25% of its size in 1996. Within Europe, the large geological surveys (Table 2) employ 600 to 850 persons; medium-sized surveys employ around 300 and small ones have 100 or fewer staff. Outside Europe, the ratio of scientific to non-scientific staff continues to vary considerably, so that comparisons of size using total staff numbers are not meaningful. Turkey employs 3456; the much larger India has a staff of only 2900. Because the prime function of any geological survey it to make geological maps, a useful measure of whether the survey is appropriately sized for its country is to divide the area of the country (in square kilometres) by the number of permanent staff. Another measure is to relate the size of the survey to population (Tables 2 and 3).

Relationships with Government

Most geological surveys are embedded within the structure of government as a division within a department or ministry. The home department will be most commonly one with responsibility for mines, minerals, or natural resources; energy; the environment; science and technology; or industry. Some geological surveys are publicly funded research institutes standing apart from government, but with a government department acting as a supervisory body. Among these, the Geological Survey of South Africa, called the Council for Geoscience since 1993, is a

Table 2 Ratios of area and population to total staff for 30 national geological surveys[a]

Country	*Ratio of area (km^2) to staff*	*Total staff*	*Ratio of population (in thousands) to staff*	*Rank of country in terms of population ratios*
Netherlands	130	315	50.1	12
Greece	159	827	12.9	2
Turkey	225	3456	19.3	5
United Kingdom	300	815	72.2	17
Luxembourg	323	8	53.9	14
Malta	323	8	48.6	11
Egypt	385	2600	26.3	8
Estonia	417	108	12.9	3
Finland	502	671	7.7	1
Lithuania	567	115	31.9	9
France	641	848	69.7	15
Hungary	650	143	70.2	16
Pakistan	718	1120	139.7	23
New Zealand	1041	258	15.0	4
Austria	1048	80	102.6	20
India	1092	2900	349.5	27
Japan	1232	300	442.4	28
Ireland	1350	51	73.1	18
Poland	1431	725	53.5	13
China	1475	6500	196.5	25
Spain	1507	335	118.3	21
Norway	1636	198	22.6	7
Sweden	1672	269	33.1	10
Belgium	1795	17	597.2	29
Iran	2354	700	96.7	19
South Africa	3822	310	130.2	22
Bulgaria	4436	25	329.0	26
Brazil	7170	1187	143.3	24
Ukraine	9432	64	788.4	30
Namibia	9931	83	20.8	6

[a]Staff numbers taken from various sources dated 2001 to 2003. Population figures and land areas are from the GeoHive Global Data website at www.geohive.com.

Table 3 Ratios of area and population to total staff for some state and provincial geological surveys[a]

State or province	*Ratio of area (km^2) to staff*	*Total staff*	*Ratio of population (in thousands) to staff*	*Rank in terms of population ratios*
Lower Saxony (Germany)	207	230	34.3	4
Delaware (USA)	353	15	53.8	6
Indiana (USA)	1089	86	71.6	7
Alabama (USA)	2526	53	84.7	8
Tasmania (Australia)	2974	23	20.6	3
Ohio (USA)	3148	34	335.9	11
Victoria (Australia)	5685	40	121.4	9
Texas (USA)	5906	117	186.2	10
South Carolina (USA)	7326	11	373.4	12
Alberta (Canada)	11 030	60	51.9	5
Western Australia	19 022	133	14.4	1
Alaska (USA)	40 281	38	16.9	2

[a]Staff levels are taken from geological survey websites in 2003. Population and land area figures are from the GeoHive Global Data website at www.geohive.com.

free-standing science council, and the BGS is part of the Natural Environment Research Council, a non-ministerial government department. The most extreme development of this kind is the Institute of Geological and Nuclear Sciences Ltd in New Zealand, which is a limited liability company. Some of the state surveys in the United States have no organizational links with government, but are attached to universities.

Funding

Most geological surveys are funded entirely or mainly by a direct grant from government, and some are prohibited by law from being involved in commercial activity. Others, particularly in Europe, are semicommercial in their mode of operation. Until 1994, The BRGM in France owned and operated a mining company fully commercially, though no funding was transferred from mining to the survey activity. Since 1973, the British Geological Survey has derived most of its income from commissioned research for government departments other than the supervisory one, and in the past decade, by an increasing amount of commercial activity. This pattern is repeated in many other countries, and by 2001 nine European surveys received less than 70% of their income as a direct grant from government. The lowest percentage grant from government: the lowest being Iceland, with only 30%. The balance in these surveys is made up of income earned from the European Union research grants, some from the private sector, but mostly on commission from various government departments. The Council for Geoscience in South Africa is divided into business units and its commercial freedom extends to allowing it to own and run hotels.

Associations and Resources

Geological surveys throughout the world have banded together to form loose, common-interest groups. Among these are the International Consortium of Geological Surveys (ICOGS); the Commonwealth Geological Surveys Forum (for countries in the British Commonwealth); the Forum of the European Geological Surveys Directors (FOREGS), which arose out of the Western European Geological Surveys (WEGS, founded in 1973); the Association of the Geological Surveys of the European Union (EuroGeoSurveys), a lobbying group to act on behalf of the geological surveys within the European Union; the Asociación de Servicios de Geología y Minería Iberoamericanos (ASGMI), which consists of the geological surveys of Spain, Portugal, and Latin America; and the Coordinating Committee for Offshore and Coastal Geoscience Programmes in South-east Asia (CCOP), which contains all the geological surveys of the region, though membership is not exclusive to them.

Information about geological surveys is dispersed. Some geological surveys have their own written histories, which can be acquired directly from them or through a library. In most cases, the best information is obtained through survey/association websites. Three groups maintain gateway sites that can be used to access geological survey information: the British Geological Survey, at www.bgs.ac.uk/geoportal; the Open Directory Project, a citizen-editor/contributor database project, at www.dmoz.org/Science/Earth Sciences/Geology/Organizations; and McCully Web, at www.mccullyweb.com. Other sources include newsletters of the International Consortium of Geological Surveys and publications of the Forum of the European Geological Surveys Directors.

See Also

Colonial Surveys. **Geological Field Mapping**. **Geological Societies**.

Further Reading

The report on the symposium meeting on the Organization of Geology Overseas, in the *Proceedings of the Geological Society of London, No. 1633, Sept 1966*, gives a good account of the origin of geological surveys in Australia, Canada, China, Czechoslovakia, France, the United States, and the former USSR.

GEOLOGY, THE PROFESSION

G L Jones, Conodate Geology, Dublin, Ireland

Introduction

Geology deals with the Earth and its processes. The variety of ways in which geologists work has been evolving for millennia. It is now the most complex and diverse profession in society, so that geologists could be regarded as modern-day polymaths. The profession continues to evolve, with many areas of expertise, ranging from palaeontology to geophysics and from volcanology to environmental geology. Along with other modern professions, it has adopted the professional qualification system, which is based on the four pillars of academic training, professional experience, a code of ethics, and continuing professional development (Figure 1).

Ancient History

Geology as a profession can be traced back to the birth of civilization. At the moment that our hunter–gatherer ancestors picked up a stone as an implement, the profession of geology was born.

Soon, the discerning, evolving, human being realized that some rocks have better properties than others depending on the needs at hand. Some were harder and did not shatter when used as a hammer; others could be worked to produce a sharp edge and wielded as axes, arrows, or spears. As an interest in rocks developed because of their use as building materials, metal sources, agricultural ingredients, pharmaceuticals, jewellery, and wealth, the *Homo sapiens* with a geological eye became valued by society.

The specialized quarryman, ore smelter, or miner was representative of geologists for thousands of years. Their keen observation and skill was exemplified by their ability to hue megaliths and other objects for use in many aspects of Mesopotamian, Egyptian, Greek, Roman, Aztec, and Inca life. Their constructions produced such marvels as elegant obelisks, the impressive Parthenon, dramatic amphitheatres, and pyramids – to mention only a few. These men knew their geology, and they understood the nature of bedded rocks and how joints broke up the beds into usable blocks. So the profession was divided for a long while into miners, quarrymen, masons, brick makers, etc.

However, even in ancient times people thought about the causes of phenomena such as earthquakes and volcanoes. Philosophers such as Aristotle, Lucretius, Herodotus, and especially Avicenna made astute observations based on meagre information, and these data were preserved through the dark ages only by Arab intellectuals. Then, towards the end of the fifteenth century, Leonardo da Vinci observed that fossils had once been living organisms and that the land had once been covered by the sea. So it was with the development of the Enlightenment that ancient ideas were re-examined and new thoughts advanced by such Italians as Vallisneri and Moro. They followed Bishop Nicholas Steno, the founder of modern geology, who developed the ideas of stratigraphy in Florence in 1669. Then, in the late seventeenth century, Descartes in France followed by Leibnitz in Germany formulated the concept of the development of the planet from vapour through molten rock to the solid surface.

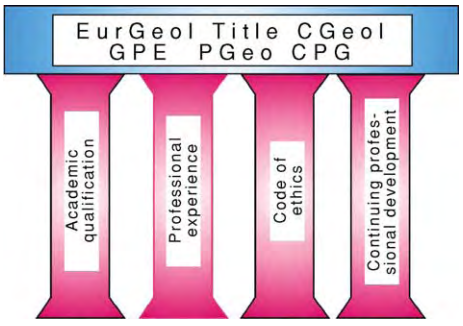


Figure 1 The four pillars of professional geological qualifications. EurGeol: European Geologist (Europe), CGeol: Chartered geologist (UK), GPE: Geologo Profesional Especialista (Spain), PGeo: Professional Geologist (Ireland) and Professional Geoscientist (Canada), CPG: Certified Professional Geologist (USA).

More Recent History

In the eighteenth century geology became the preserve of the gentleman polymath. These men of substance, quite a number of them questing clergymen or surgeons, found that they could read the rocks like a book and could see that certain layers always appeared above others and that they contained distinctive fossils. Gradually they put together the sequence of deposition of the rock layers.

In France, the Académie Royale des Sciences was the seat of the discussions that crystallized French learning about the Earth, such as Desmarest's 1771 memoir, in which he presented his theory of the

volcanic origin of basalts. The founding fathers of modern geology included men such as James Hutton (*see* **Famous Geologists:** Hutton), who published his '*Theory of the Earth*...' in 1795, and Charles Lyell (*see* **Famous Geologists:** Lyell), whose *Principles of Geology* was published in 1833. In England gentlemen organized themselves into the Geological Society of London in 1807; this was initially a discursive dining club, but it eventually became the home of geology (*see* **Geological Societies**).

Geological maps were being drawn at that time in many countries and included Tighe's Kilkenny map of 1802 in Ireland and Mantell's 1822 map of the South Downs in England. However, one man who was involved with the practicalities of the industrial revolution made enormous strides. A blacksmith's son working on the development of canals throughout the UK drew detailed maps of the changes in the rocks across the countryside. So William 'Strata' Smith (*see* **Famous Geologists:** Smith) drew the first geological map of most of Britain in 1815.

Thus during these times geological learning evolved from a restricted interest in mining or mineralogy to broader geology. For a long time, in many parts of Europe, geology was seen as a subservient branch of engineering, and amazingly in some countries a geologist still cannot sign his own professional report, but must ask an engineer to do it for him!

The Breadth of Geology

It can be seen that the practice of geology developed and enlarged over the millennia and embraced many disciplines. It has burgeoned into perhaps the broadest discipline to be found in society.

Geologists include palaeontologists and palynologists who specialize in zoological and botanical sciences; their understanding of the processes of evolution is central to their work. Sedimentologists image past processes and compare them to modern ones. Mineralogists and crystallographers deal with minerals and crystals. Metamorphic geologists look at the changes that take place in rocks under extreme heat and pressure. Hydrogeologists deal with the crucial area of the movement of groundwater and our ability to provide clean water to society; a subgroup specialize in the disposal of society's waste in a safe, clean manner. Geophysicists, through mathematics and physics, use diverse techniques, such as the micromeasurement of gravity, the response of the ground to electrical currents, the reflection of radar waves, and, of course, the measurement of seismic waves, whether produced naturally by earthquakes or by artificial sources. Marine geologists look at geology below the surface of the sea, imaging the seafloor and working with their geophysicist colleagues to see deep into the rock strata. Petroleum geologists understand the formation of oil and gas and how these vital commodities are caught in traps in the rocks; they help to deliver these resources to society. Volcanologists or igneous geologists deal with active volcanoes and igneous processes and with their now fossilized equivalents in the geological record. Geochronologists use their knowledge of the radioactive decay of some elements as a crucial tool in the absolute dating of rocks and sediments, for example by using the decay of uranium into lead to date ancient intrusive igneous rocks that are between 1 Ga and 400 Ma old, the decay of potassium-40 into argon do date rocks formed during the last 500 Ma, or the decay of carbon-14 into nitrogen to date more recent organic sediments that are a few tens of thousands of years old. Geochemists specialize in the mineral constituents of rock, following on from the smelting of ores for thousands of years, and this was an important aspect of the development of the science of chemistry. Structural geologists apply mathematical and geometrical knowledge to the way that rocks behave under varying conditions of pressure and temperature. Engineering geologists must know about the technical properties of rocks and sediments and be able to communicate this information to their engineering colleagues; they also deal with natural hazards such as danger from landslips, etc. Environmental geologists have become central to the management and development of society; their input into land-use planning improves the quality of life for the inhabitants and ameliorates the risks posed by geohazards. Planetary geologists learn from our planet and its moon and work with their astronomical colleagues in investigating distant bodies. Remote-sensing geologists use satellite imagery to examine the surface of our planet.

Our science overlaps into a plethora of other disciplines.

Academic Education

As geology gradually evolved, existing university natural history courses began to include geological options. In the five years after Waterloo in 1815, the universities began to formalize matters: Buckland was taken on as reader in Oxford, and Sedgwick (*see* **Famous Geologists:** Sedgwick) received the chair in Cambridge. These chairs were in natural history, mineralogy, geology, or a combination, but the course was becoming a geological one. The degree awarded was still a Bachelor of Arts, of course.

This trend continued, and colleges around the world began teaching geology. Degree courses

evolved until 3 years of education were required to receive a Batchelor of Science degree, or 4 years with Honours. Most universities now offer postgraduate Master of Science degrees, involving 1–2 years work, through thesis and/or examination, whilst Doctor of Philosophy (PhD) degrees may take between 3 and 5 years of research to complete. In Europe, this may take up to 10 years. In general it is considered that most geologists need to complete at least 5 years of education before becoming suitable for a long-term geological career.

The Learned Societies

In order to cater for all the interests of these geologists, various bodies came into existence. They fall into the categories of learned and technical societies, but a new function or even a new type of society has evolved with the recent birth of the professional societies (*see* **Geological Societies**).

We see that the major national societies such as the Geological Society of London, the Union Français des Geologues, and the Ilustre Colegio Oficial de Geólogos cater for many areas of geological life and interest, especially when they are able to set up special interest groups. The specific concerns of large groups of specialized geologists, especially when these are of a global nature, have lead to the establishment of large organizations such as the International Association of Engineering Geologists, the European Union of Geological Sciences, the Petroleum Exploration Society of Great Britain, the Micropalaeontological Society, the International Association of Hydrogeologists, and the Society of Exploration Geophysicists. All cater for a particular area of geological interest, and most produce a journal carrying technical peer-reviewed articles.

The Professional Bodies

The evolution of the professional geologist, working in academia, government, or industry, has lead to the need for organizations to look after them. Some learned societies, such as the Geological Society of London, have evolved easily to cater for this new function. In Ireland it was necessary to create a new independent body – the Institute of Geologists of Ireland – to look after Irish professional geologists, and in the USA the American Institute of Professional Geologists took on this role. The European Federation of Geologists (Figures 2 and 3) took on the continental role of representing the national associations, some of which were professional bodies and others learned societies. Similarly, in Canada the Canadian Council of Professional Geoscientists represents the professional interests of the provincial bodies (Table 1).

All of these bodies are striving towards the same goal of ensuring professional standards and representation. They are establishing mutual recognition and other agreements to benefit their members worldwide.

The Profession

The practice of geology occurs in three principle areas: academic, governmental, and industrial.

Clearly, the teaching of geology is crucial, and it is amazing that so little is taught in primary or secondary schools; geology is simply seen as a small aspect of geography! It is in tertiary education that geology comes into its own, and universities develop courses that give the neophyte geologist a grounding in so many disciplines. So, most geologists working in academia are based in university geology departments, though some may be found in the allied disciplines of geography, archaeology, etc. Some can also be found in Schools of Mines or Technical Institutes, where they instruct mining geologists, geosurveyers, and geological technicians.

In Government, the traditional area of practice has been in geological surveys. These have long been seen as the providers of modern geological maps, and this is indeed one of their main functions. It has been interesting to see the variation in the response of geological surveys around the world to modern changes in such areas as digital data management. Those surveys that have remained traditional have come under great pressure, and some have had to close. Others that understand the demands of the digital era have been quick to adjust to being data managers and providers, not just map makers, and so have become a necessary organization rather than an 'appendix'. Consequently, geologists working in these areas have also developed their geological skills, often into new areas such as heritage or marine. Other bodies have also created opportunities for geological employment, and in the last few years departments such as Environmental Protection Agencies have sprung up to look after the world we live in. Geology naturally also has a role to play in the supply of renewable energy, allowing countries to be less reliant on fossil fuels.

Industry is a huge area of employment for geologists. The supply of fossil fuels or mineral resources depends on the geological setting and on the location of the deposit to be exploited. So a geologist must be mobile and able to travel to any part of his or her country and even to any part of the world if necessary. Consequently, an ability with foreign languages can be highly advantageous at times. However, in the area

Figure 2 Mission and objectives of the European Federation of Geologists.

of industrial minerals, geologists deal with the supply of basic raw materials such as limestone, sand and gravel, and aggregates; many of these materials are supplied very close to their extraction source since transport is the major cost.

The birth of new areas of geology, taking over to some extent from the older ones, has lead to more home-based occupation, including hydrogeology, waste disposal, pollution remediation, environmental geology, geotechnical geology, renewable energy geology, and geophysics.

Professional Qualifications

Professional qualifications have been developed during the last half of the twentieth century. This profession has set up requirements that geologists must demonstrate not only that they have a good academic education, but also that they have achieved a high standard of professional expertise over a number of years and that they can demonstrate this in front of their peers by examination of both their work and their knowledge. Another part of this process has been the signing by geologists of agreements to obey codes of ethics demonstrating that they will work to high ethical standards. These codes are backed up by disciplinary committees. A further aspect that has become formalized is the agreement to demonstrate that continuing professional development is being carried out by the geologist.

Around the world this movement has proceeded and the standards that have been set up are very similar in most countries, so that across Europe there is a single standard to be achieved, and this is recognized to be equivalent to the standard achieved in other continents, with reciprocal rights and recognitions existing between different international groups, allowing geologists to travel and practice around the world.

Also of importance has been the acceptance by government and business bodies of this professional standard. For them, this quality-assurance mark is a

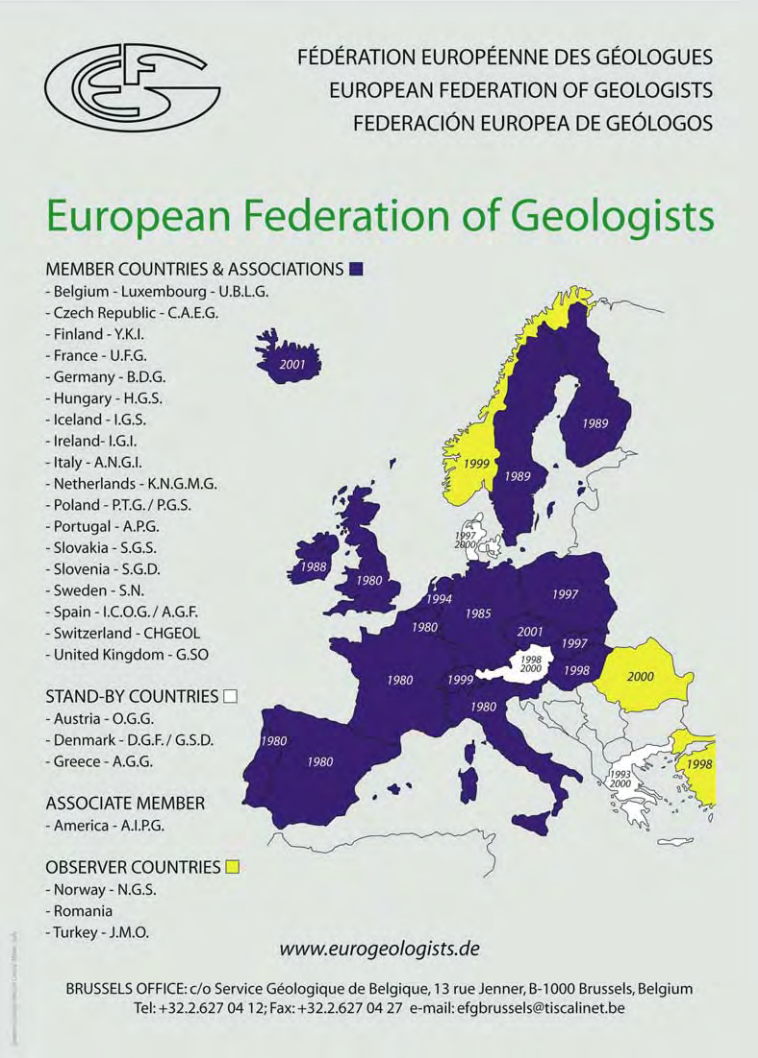


Figure 3 Members of the European Federation of Geologists. Single dates are years that countries became observers or members. Double dates are periods for which countries were members.

Table 1 Professional websites

European Federation of Geologists	www.eurogeologists.de
Geological Society of London (UK)	www.geolsoc.org.uk
Illustre Colegio Oficial de Geólogos (Spain)	www.icog.es
Institute of Geologists of Ireland	www.igi.ie
American Institute of Professional Geologists	www.aipg.org
Canadian Council of Professional Geoscientists	www.ccpg.ca
Berufsverband Deutscher Geologen, Geophysiker und Mineralogen EV (Germany)	www.geoberuf.de

way of ensuring high standards in the geological input to projects.

Part of the motivation for this movement has been the necessity to demonstrate high standards to local, national, and continental governments. Especially in this litigious age, the use of appropriately highly qualified people in all aspects of life is becoming a necessity.

Another powerful motivation has been the recent requirement by financial bodies such as the Canadian, Australian, British, and Irish stock exchanges that all sections of bankable reports on mineral

deposits are signed by competent people. The holders of these professional geological qualifications, available in Europe and North America, are accepted by these bodies as competent people in their fields of geological expertise. Government departments are also beginning to recognize that these qualifications help to ensure the quality of reports submitted to them by professional geologists.

In Europe, the European Federation of Geologists is the representative body for geologists, and it is composed of the national geological associations of each of the nineteen member countries (Figure 3). It administers the professional title of European Geologist, which is equivalent to the professional titles of Professional Geologist in Ireland, Chartered Geologist in the UK, and Titulo de Geologo Profesional Especialista in Spain. For these countries and for those that have no national professional qualification it provides a European route to a continentally recognized qualification.

Regulation

In some countries the State has taken upon itself to regulate the profession. In Italy and in Spain, in order to practice, a geologist must be a member of the body recognized by the state – the Consilio Nazionale dei Geologi and the Ilustre Colegio Oficial de Geologos, respectively. In the USA, many states require a geologist to be licensed in order to practice. The National Association of State Boards of Geology coordinates this licensing system. The role of the professional body, the American Institute of Professional Geologists, then becomes the organization of a high professional standard. In Canada the Provincial Geological Bodies have come together to form the Canadian Council for Professional Geoscientists, which coordinates the professional standards organized by each state and formalizes their mutual recognition.

The Future

The profession is still changing rapidly. Will the trend towards calling ourselves Earth Scientists instead of Geologists hold sway? I hope not! It appears to be a fad with no practical basis except to persuade academic bodies that real change is going on.

The geological profession has a clearly defined role in caring for Earth resources – clean water, the environment, renewable and finite energy, geohazards, and heritage – and in observing the continually changing process that is our Earth.

See Also

Engineering Geology: Codes of Practice. **Famous Geologists:** Hutton; Lyell; Sedgwick; Smith. **Geological Maps and Their Interpretation**. **Geological Societies**. **Geological Surveys**. **History of Geology Up To 1780**.

Further Reading

Cadbury D (2000) *The Dinosaur Hunters*. London: The Fourth Estate.

Geikie A (1987) *The Founders of Geology*. London: Macmillan.

Herries Davies GL (1983) *Sheets of Many Colours*. Dublin: Royal Dublin Society.

Regueiro M and Jones GL (eds.) (2000) *European Geologist 10, Proceedings of the First International Professional Geology Conference, July 2000. Alicante* Brussels: European Federation of Geologists.

Selley RC (1996) What on Earth is a Geologist? *European Geologist,* p. 3 4.

Winchester S (2001) *The Map that Changed the World*. London: Penguin.

GEOLOGY OF BEER

S J Cribb, Carraig Associates, Inverness, UK

Introduction

The fermentation of an extract of cereal grains is probably one of the oldest processes known to man, with techniques stretching back to Ancient Egypt. In the UK and parts of western, central, and eastern Europe a wide variety of fermented barley beverages are produced, known generally in English as beer or ale. Much of the variety is a direct result of the chemical composition of the water used to extract the sugars from the barley prior to fermentation. This variation derives from the geology of the water sources.

The pleasures of beer consumption are encapsulated in the following quotation, from an unknown source.

> The enjoyment of a glass of beer may be received by many senses: the sight may be attracted first by the clarity of a pale ale or the rich creamy head of a stout. As the glass is raised to the lips the aroma of the bever age, possibly the bouquet of the essential oils of the hops, may excite the nostrils. Then, as the liquid flows over the taste buds at the back of the mouth, and further volatile products diffuse into the back of the nose, the flavour of the beverage is perceived. Finally, the beer enters the body, where the alcohol is rapidly absorbed into the bloodstream and exerts its well known physiological and psychological effects.

In more formal terms, in any particular brew the brewer is seeking a combination of five characteristics: flavour (including both taste and aroma), alcohol content, colour, head retention, and clarity. The chemical nature of the water used in the brewing process exerts a strong control over all of these.

The Brewing Process

There are several stages in the brewing process, and these are summarized below.The first stage is malting, which is the conversion of barley grains into fermentable malt. The grains are steeped in water and spread out to germinate. This initiates enzyme processes, which start to break down complex carbohydrates into sugars. Kilning, heating to around 100°C, arrests these processes before they go to completion.

The malt is milled (ground) to form grist and then mixed (mashed) in a mashing vessel or tun with hot brewing water (liquor). Here the processes started during malting continue.

The liquor and malt (wort) are then boiled in the presence of hops, and bitter acids are extracted from them and converted into resins. The wort is then cooled by water-powered coolers, at which time proteins and tannins separate from the hot resins. Yeast is then added and fermentation takes place. Fermentation adds fruitiness and fullness to the liquid and converts the sugars into alcohol. Yeasts may ferment on the surface, something that is characteristic of most British brewing, or at the bottom of the fermentation vessel, as is more common in Europe. A problem with top fermentation has always been the susceptibility of the yeast to breeding with wild airborne yeasts. This often meant that brewing was difficult or impossible in the summer. A solution was found by brewers in Europe in the late 1880s, who developed a yeast that sank to the bottom of the fermentation vessel. Bottom fermentation is generally thought to produce a clean, soft, non-fruity beverage.

The Importance of Water

Great volumes of water are involved in all stages of the brewing process, which is why so many breweries were originally sited near major rivers, though it will be a great relief to many to know that, despite suggestions to the contrary, river water was not necessarily used in the beer itself.

It is possible to brew beer anywhere, from any type of grain and water, but historically the classical area for brewing is the Beer Belt of Europe. This belt essentially comprises six countries – Ireland, the UK, Belgium, Germany, the Czech Republic, and Slovakia – but also includes, to the south, parts of Austria (Vienna), Switzerland, and northern France and, to the north, the Netherlands, Denmark, and Sweden.

The greatest contribution to the chemistry of the brewing liquor comes from the brewing water, which forms a vital and integral part of the nature of the final product.

There are four anions that are particularly significant, of which Calcium is by far the most important. It has three major effects. It stabilizes the enzyme α-amylase and helps the breakdown of starch from the malt in the mash tun and in later processes. It precipitates phosphate and thus increases the acidity of the wort, which is important because acidity influences the strength and character of the fermentation and the microbiological stability of the enzyme processes. Lastly, it promotes flocculation (clumping together) of the yeast during fermentation. Magnesium produces a sour to bitter taste, but retards phosphate precipitation, which in turn stops the required drop in pH level. Sodium in small amounts gives a salty to sour taste, and potassium, whilst also contributing a salty flavour, can be particularly laxative above 10 ppm.

Although, as just described, it can be seen that the anions are important, it is really the cations that have the major influence, and this will be illustrated, first, using examples from the UK, before extending the logic to cover the rest of Europe.

Brewing in the UK

Brewing was established in Burton-on-Trent in the English Midlands in the sixth century, when beer was brewed by the monks of the local abbey. They drank beer rather than the more suspect river water. The brewing waters at Burton-on-Trent (Table 1) are characteristically very high in sulphate (SO_4), derived from the Triassic gypsiferous marls in the area. Sulphate is very important for the brewing of bitter beers because it helps in the degradation of proteins and

Table 1 Brewing water analyses

TDS	*Calcium*	*Magnesium*	*Potassium*	*Sodium*	*Sulphate*	*Chloride*	*Carbonate*	*Location*
1401	283	90	0	29	725	54	171	Burton on Trent
670	111	38	2	83	168	62	162	Wolverhampton
428	109	5	5	29	69	37	128	London
305	100	16	0	0	17	17	150	Dublin
800	140	36	0	92	231	60	210	Edinburgh
1011	260	23	0	69	283	106	270	Dortmund
273	80	19	0	1	5	1	164	Munich
31	7	1	0	3	6	5	9	Pilsen

All figures are in parts per million.
TDS, total dissolved solids.

starch and allows the full extraction of bitter oils from the hops. It also reacts with magnesium carbonate to give magnesium sulphate, which is in itself bitter. The high proportion of hop oil acts as a preservative, as well as a flavouring, and consequently it was this type of beer that was exported to the colonies in the nineteenth century as India Pale Ale.

To the west of Burton-on-Trent, in the Birmingham and Manchester areas in particular, the predominant evaporite deposit in the Triassic red beds is halite rather than anhydrite, and the composition of the ground water reflects this difference (Table 1). The salty to sour taste provided by the sodium ion combines with particular chloride sweetness giving, at levels under 200 ppm, what brewers term palate fullness. The brews are still reasonably high in sulphate, though, as it is lower than in the Burton waters, they cannot be as highly hopped. The resulting sweeter beer, which still retains an amount of bitterness, is the classic mild ale of these areas.

In southern and eastern England, where Cretaceous chalk is the main aquifer, the groundwater has a very different composition. It is, as would be expected, dominated by carbonate (Table 1). Clearly the use of water of this composition for brewing presents problems, primarily with malting and fermentation, because of the difficulty of keeping the pH low. Additionally, the lower sulphate means that the extraction of oils from the hops is poor, giving a less flavoured brew. A brew was developed in the 1770s that was suited to these waters, and this was a sweet, dark beer in which the dominant flavour came from roasted malt rather than hops. This drink was known as porter because of its popularity in the London markets of Billingsgate and Covent Garden.

Brewing in the Rest of Europe

In Ireland at about the same time attempts to brew bitter beers had failed, though it was found that the waters used in brewing there were ideal for beers of the porter type. The reason is that most of the waters used in Dublin and Cork in particular are derived from the Carboniferous limestones in the centre of the country and are thus dominated by carbonate (Table 1). The most famous of these brews was developed by Arthur Guinness in Dublin, who produced his extra stout (meaning thick) porter, which on contraction introduced the word 'stout' into the language of brewing.

Many countries in the European Beer Belt use mineralized groundwater for their characteristic brewing, and this is highlighted by the following examples. Two of the classic brewing areas of Germany are Dortmund and Munich, where mineralized groundwater determines the brew type. In Dortmund (Table 1) much of the water is derived from the Coal Measures, where waters similar to those in Burton-on-Trent are found, though they are not as strongly mineralized and have a higher carbonate content. The classic 'Dortmunder' is bottom-fermenting, as one would expect in Europe, has enough sulphate to be well hopped, and has a high enough total mineral content to promote good fermentation and thus produce a clean, dry beer. Munich on the other hand draws water from more recent rocks, which, as can be seen from Table 1, is closer to the water of Dublin in chemical composition. The classic Munich beer is sweet, full, dark, and brown, to all intents and purposes the equivalent of stouts and porters. Denmark, particularly in Jutland, has an essentially Cretaceous geology overlain by Tertiary deposits. The naturally occurring groundwater is similarly rich in calcium carbonate and the response to this is the brewing of sweet, dark beers termed Stowts.

However, in much of Europe the groundwaters, whether derived from surface run-off, sandstones, or impermeable lower Palaeozoic or Precambrian strata, have very low levels of both calcium and sulphate. In these areas the beers that are produced are alcoholically strong but lightly hopped and texturally thin. This type of beer was developed initially in

the town of Pilsen (Table 1) in Bohemia (now in the Czech Republic). A critical process was the storage of the fermented beer in cold caves – lagering – and thus two new terms were introduced into the brewing vocabulary, pils and lager. This direct response to brewing with waters of low ion content spread rapidly in the 1880s throughout those areas of northern Europe where such waters are common. This included the Netherlands, Germany, and Denmark, where to this day lager is recognized as a major beverage.

So far, no mention has been made of another important brewing nation, and that is Belgium. The country is fascinating, as geologically it is not always very promising as far as brewing waters are concerned, but it illustrates a classic example of a unique response. Simply, in areas where there are useful mineralized waters, use them to their best effect; otherwise, ignore the rules. In East Flanders, particularly around Oudenaard, there are carbonate-rich waters, and the characteristic Brown Beers are top-fermenting, dense, and smooth. These contrast markedly with the Red Beers of West Flanders, where the less mineralized waters lead to the brewing of light and not very bitter beers. There are four monastic Trappist breweries in the Ardennes, where the waters are highly enriched in calcium carbonate and as a result the typical brews are top-fermenting, dark brown, and sweet. There is a fifth brewery at Westmalle, north of Antwerp, where the water is hard and a brew much closer to a bitter ale is produced. Where the waters are poorly mineralized there are only two options. First, there is the classical approach into pils brewing, typically around Louven, Alken, and Lindberg where the groundwater is derived mainly from superficials and surface run-off. Second, there is a much more idiosyncratic Belgian approach, which is the production of Lambic, traditional in the Brussels area. In this style wheat and malted-barley beers are produced using naturally occurring wild yeasts, which ferment completely and are very dry. Hops are used not for flavour but purely to protect the beers from unwanted infection. Fruit flavourings are often introduced, the best known being black cherry (Kriek) and strawberry (Framboise).

Scandinavian brewing is dominated by lager-style beers, as would be expected from poorly mineralized waters derived mainly from Precambrian rocks, though occasionally a Belgian style approach is attempted, as with the juniper-flavoured beers of Finland.

Modern Brewing

Many breweries are turning to the use of mains water that is deionized and then reconstituted to a given composition for a particular brew. Commercially there are advantages in this procedure, specifically related to purity and continuity of supply. Where a water grid system is in place a major disadvantage is that the composition of the water entering the brewery can change very rapidly as sources are changed, and this completely alters the nature of the brew in progress. Wells and springs are still used by some breweries, but the problems of microbiological pollution, increasing nitrate content, and, near coasts, overpumping causing seawater intrusion are gradually reducing this number.

An interesting thought provided by one brewer is that, despite all that has been written about water type, the siting of breweries initially may have been controlled not by the analysis of the water but by the presence of a source of abundant water at a constant temperature in the days before thermometers were in constant use.

See Also

Geology of Whisky. **Geology of Wine**. **Minerals:** Carbonates; Sulphates. **Palaeozoic:** Carboniferous. **Sedimentary Rocks:** Chalk; Limestones.

Further Reading

Jackson M (2000) Great Beer Guide. London: Dorling Kindersley.

GEOLOGY OF WHISKY

S J Cribb, Carraig Associates, Inverness, UK

Introduction

The distillation of liquors produced by the fermentation of cereals is a process that goes back into the mists of history. There is historical evidence from Scotland dating the distillation of malted barley from the middle of the sixteenth century. The distillate was known originally by its Gaelic name *uisge beathe* ('water of life'), which has become the present-day word 'whisky' (spelt variously in many countries with or without an 'e'). The distilling process was legalized in Scotland in the 1830s, and since then the manufacture of this special liquid, whisky, has been developed to a fine degree, as is evident in the array of unique products available today.

The manufacture of whisky occurs in many countries, but the pinnacle of success with this product is found in Scotch whisky, a specific product of Scotland. To qualify as Scotch whisky, the spirit must be derived from malted and/or unmalted cereals and be matured in oak barrels in Scotland for at least 3 years. Whiskies fall into three categories: grains, malts, and blends. Grain whisky is manufactured generally by a continuous industrial process from malted and unmalted cereals. Malt whiskies are made totally from malted barley. Blended whiskies comprise a base of between 40 and 60% grain whisky overlain by up to 30 malts. Malt whisky represents probably the pinnacle in the development of the whisky-making process and of the 100 or so distilleries in Scotland, around 75 are producing at any one time. Each of these distilleries produces a unique spirit characterized partly by the unique source of water for its production, and it is in the chemistry of this that the varying geology of the water sources comes into play.

Whisky Production

The whisky-making process essentially comprises a brewing phase, followed by the distillation of the resulting liquor. In fact, the result of the first stage is the production of the equivalent of a strong beer, though unflavoured by hops or other additives. The first part of this process is the malting of barley, whereby the grain is encouraged to germinate for several days to initiate the enzyme processes that break down starch to sugars, before the process is stopped by heating. Traditionally, where the malting takes place at a distillery site, a certain amount of peat smoke is allowed to permeate the malt during the heating process. The level of permeation is reflected in the smokiness of the final product. Nowadays there are larger scale centralized maltings serving several distilleries, but the process of peating is essentially the same. The second stage is the mixing of the malted barley with fresh potable water to extract the sugars and produce a liquor ready for fermentation. Fermentation takes place using a combination of beer and wine yeasts to raise the alcohol content to around 8%. The resulting liquid is then heated in the distinctive onion-shaped copper stills, normally in a two-stage process that raises the alcohol content, first to 25% and second to the final spirit alcohol content of between 55 and 65% alcohol. The raw spirit is stored in oak barrels of various types: new wood, European oak barrels generally from the sherry industry in Spain, and American oak bourbon barrels. Thus the four crucial factors that combine to produce the wide variety of malt whiskies are the degree of peating of the malt, the shape and type of still, the types of barrels used for maturation, and the chemistry of the process waters. It is with the latter step that the geological influence on whisky type is exerted.

Water chemistry affects three areas of the production process: the extraction of sugars from the malt, the nature and type of fermentation, and the chemistry of the distillation process (which is largely unstudied). There is an oft-quoted maxim that the best type of water for the production of malt whisky is soft water through peat over granite. Even a cursory comparison of the siting of malt whisky distilleries with the geological map of Scotland will show that only in comparatively rare instances is it likely that such water will be available. In fact, only 20 distilleries, most of which are centred in the Speyside area of north-east Scotland, have such water. The remaining rock types from which water is sourced cover the whole spectrum from sandstones and schists to quartzites and shales. Clearly the chemistry of the waters derived from such a range of sources will show significant variations up to and including significantly heavily mineralized (hard) waters.

Distilleries

Distilleries come in all shapes and sizes and they are situated on the coast, inland, on hills, in valleys, by mighty rivers, near small streams, and on islands

Figure 1 Clynelish Distillery at Brora in Sutherland.

Figure 2 The Tarlogie Spring; the source of hard water used by the Glenmorangie Distillery near Tain in Ross shire.

(Figure 1). They use water from boreholes, rivers, springs, and lochans (small lakes), and every water source is unique in terms of its chemical make-up; distillery waters vary significantly not only in the chemistry of the major elements, but even apparently similar waters have markedly different trace element profiles (Figure 2). The study of malt whisky is in a very real sense a study of the geology of Scotland.

Starting with the Speyside district of north-eastern Scotland, centred around the villages of Dufftown and Rothes, there are over 20 distilleries in an area of approximately 100 miles^2. In this area are some of the most famous distilleries of all. The whiskies fall easily into three groups, based on taste and aroma, and these groups correspond very closely with the geology of the water sources. The first group taking its water from the granites to the west of Dufftown include Glenfiddich, Balvenie, and Aberlour. The second group, in the west and north-west of the region, draw spring waters from the well-bedded, light grey, Dalradian quartzites of the Grampian group, and include Glen Grant and Knockando. The third group takes water from a wider range of Dalradian quartzites, phyllites, and schists. Among these are Glenlivet and Strathisla.

Moving to the west of Speyside, in the area called the Grampian Highlands, which lies between the Highland Boundary Fault to the south and the Great Glen Fault to the north, is an area that stretches from the islands of Islay and Jura in the south-west to the Buchan coast in the north-east. Within the area there are many igneous intrusions of granitic, basic, and ultrabasic composition. The mountain of Lochnagar provides water for the Royal Lochnagar distillery, which is situated behind Balmoral Castle. At Knock Hill, east of Keith, the water seeps out between the Dalradian schists and the underlying gabbro. The whisky produced has the Gaelic name An Cnoc. To the south around Insch, 17 springs provide the ultrabasic water source for Ardmore, an unusual whisky.

A drive along the main road north from Perth to Inverness highlights not only the change in geology but also the variation of spirit and water source, in a comparison of five distilleries. Starting in the rich agricultural land just north of Perth, underlain by the Old Red Sandstone, the Highland Boundary Fault is soon crossed at Dunkeld and immediately the scenery becomes mountainous as the Dalradian succession is encountered. North of Dunkeld is the town of Pitlochry, home to the Blair Atholl distillery, and close by is the small distillery of Edradour. Both of these draw their water from high in the schistose hills around Ben Vrackie. Northwards through the pass of Drumochter, the highest distillery in Scotland is reached at Dalwhinnie; the water source is the clear blue sparkling mountain waters of the Lochan na Doire-uaine. Further north, near the town of Newtonmore, is the distillery of Drumguish, and further still, near Carrbridge, is the huge distillery of Tomatin, taking its water from the Allt na Frith, a stream that bounds down a rocky juniper-filled valley, flowing through the central Highland migmatites.

To the far south-west of Scotland lie the Argyll islands of Islay and Jura, known particularly for the very heavily peated malts of Laphroig, Lagavulin, and Ardbeg. However, there are other whiskies that, being less peated, reflect more closely the differences in geology of their water sources. The furthest north is Bunnahabhain, the only distillery on the island to take its water from underground springs rather than from surface streams. These waters, rising as they do through Dalradian dolomites, are hard, containing substantial amounts of calcium and magnesium carbonates. This type of hard water, combined with the use of totally unpeated malt, produces a markedly unique whisky, and the water composition is considered a major influencing factor. In the far west of Islay is Bruichladdich, the most westerly of all distilleries in the United Kingdom, taking its water from the dark brown sandstones of the Colonsay group, a group for which the affinities are not yet completely understood.

An area of Old Red Sandstone that runs along the north-east coast of Scotland, taking in Inverness, Dingwall, and Dornoch, includes the Caithness towns of Wick and Thurso and the Orkney Islands in the far north. The waters from these sandstones generally have a high calcium carbonate hardness, something that the distilleries in the area claim has a significant effect on the nature of the whisky that they make. The red sandstones can, however, create problems, because they break down to form excellent farmland, and where there is intensive agriculture there is often a problem with the high nitrate content of groundwaters. Indeed, several distilleries within the area have closed particularly because of water problems. The most famous distilleries in the region are Glenmorangie in Tain, Clyneleish in Brora, Old Pulteney at Wick, and the most northerly distilleries of Highland Park and Scapa on Orkney.

Far to the south, beyond the Highland Boundary Fault, lies the Midland Valley of Scotland. At the northern edge, close to the Highland Boundary Fault, there are Old Red Sandstones similar to those in the north of Scotland. The distilleries in this area may be divided into two groups, based on their water sources. First, there are those that are situated in the Lowlands, taking their water from either across the Highland line or from the red sandstones just to the south, such as Fettercairn. Second, there are those that utilize a variety of water sources from rocks within the Midland Valley, such as Auchentoshan on the Clyde (Carboniferous volcanics), Glengoyne (Devonian volcanics), and Glenkinchie south of Edinburgh, which takes its water from south of the Southern Uplands Fault from a reservoir in Ordovician volcanics. In western and north-western Scotland, the

lavas of the Tertiary Volcanic Province provide surprisingly hard waters; these are seen in Talisker on Skye and Tobermory on Mull and across the water at Bushmills in the plains of Antrim in Northern Ireland. Rarely does a new distillery open, but the latest to do so was built in 1995 at Lochranza in the north of the Isle of Arran. The Isle of Arran single-malt whisky is well into production and in one sense closes the circle with regards to a water source, in that it utilizes soft waters from the Easan Biorach, which flows over granite, having passed through thick peat beds.

Waters used in whisky making, though more esoteric than brewing waters used for beer, are certainly less studied, and a full understanding of the effects of water chemistry on whisky making has still to be achieved, but there can be no doubt that the effects are significant and demonstrable.

See Also

Engineering Geology: Ground Water Monitoring at Solid Waste Landfills. **Geology of Beer**. **Geology of Wine**.

Further Reading

Cribb S and Cribb J (1998) *Whisky on the Rocks*. Keyworth, Nottingham: British Geological Survey.

GEOLOGY OF WINE

J M Hancock[†], Formerly Imperial College London, London, UK

Introduction

Geology can be important for the growers of grapes for wines; it can be ignored by the consumer. Most of the arguments for the importance of '*terroir*', which includes geology, have been along the lines: "This wine has one flavour, that wine has another flavour, yet they are made from the same grape variety". The difference must be something in the ground, i.e., it depends on the geology. Geologists have known for a long time that this is a misguided simplification. In the mid-nineteenth century, a distinguished French geologist, Henri Coquand, played a joke on the wine producers by stating that the quality of cognac (which is a distilled product anyway) was directly related to the quantity of chalk in the ground. The zones of quality for cognac are arranged in circles centred on Cognac: the strata have a roughly linear strike NW–SE. The best cognac comes from an area in which chalk is largely absent. In spite of the obvious nonsense of a relationship between chalk and the quality of cognac, the idea is still being quoted today in some books.

As with any agricultural product, there are many factors that control the quality of wine. Geology can play a part in three of these: the temperature around the vines in general and the bunches of grapes in particular; variations in porosity and permeability around the roots of the vine, which will affect both the supply of moisture and the rate at which the vine can take up nourishment through its roots; and variations in the composition of the ground, which will control the availability of nourishment supplied through the roots.

Temperature

It was known to Lucius Columella in the first century AD that the quality of grapes depends on the temperature around the vines. All modern work on the relationship between heat and grapes originates from Amerine and Winkler in California, who showed that each grape variety requires its own heat regime to bring out the best of its qualities. Amerine and Winkler worked in California where, at that time, vines were kept around 1.1 m above nearly level ground. This meant that it was the general ambient temperature of the district which was the controlling factor. However, in many vineyard regions, the geology affects the mesoclimate, principally by its control of the topography.

1. *Shelter from cold winds*. In more northerly vineyards (or more southerly vineyards in the southern hemisphere), the better vineyards are located on slopes which face south or south-west. Vineyards are not extended to the top of the hill, so that the summit can act as a break from cold winds from the north and north-east. It is even better if there is a clump of trees on the summit. This sheltering effect is of value even if the local slopes in the vineyard are to the east or north.
2. *Thermal belt on slopes*. With a layer of cold air at the bottom of a valley and a layer of cold air near the ground over the plateau above the slope, an

[†]Deceased

intermediate zone, known as the thermal belt, is developed on the slope. Although somewhere in the middle of the slope, its actual height is highest in winter and lowest in summer. Its position on any one hillside is remarkably constant. Sites only 2 or 3 km apart horizontally may show differences of 8°C over height differences of less than 100 m. In vineyards, the main importance of a thermal belt is often the maintenance of slightly higher temperatures during the night.

3. *Radiant heat from the sun* (Figure 1). The radiant heat that a plot of ground receives from the sun depends on the angle it presents to the sun. In simplified form, the relationship is:

$$I = k \sin(\alpha + \beta)$$

where I is the intensity of radiation received on the slope, k is a constant, α is the angular elevation of the sun, and β is the angle of inclination of the slope to the horizontal along a meridian, i.e., to the south in the northern hemisphere, and to the north in the southern hemisphere.

Thus, the steeper the slope, the more radiant heat the vineyard will receive (Figure 2). As the direction of slope is angled away from the south (in the northern hemisphere), the less radiant heat the vineyard will receive. On a cloudy day in January in central Europe, a slope of 20° facing south receives roughly twice as much radiant heat as a flat surface.

This effect is well illustrated in the Mittelmosel. Thus, near the village of Piesport, the best vineyards on the steep, south-facing slopes of up to 30–35° on phyllites are on the north flank of the river. In an area of less than 3 km × 1 km, there are seven Einzellagen (individually named vineyards and wines), distinguished by modest differences of angle and direction of slope, but each of high quality (the most famous Einzellage – 'Goldträpfchen' – is not on the steepest part of the slope because of the position of the thermal belt). The whole of the flat area south of the river is lumped into a single Einzellage, Treppchen, producing quantities of undistinguished wine, with lower concentrations of alcohol and less flavour.

4. *Re-radiation of heat from the ground*. Because the ground is kept clear of weeds in most vineyards, the effects of re-radiation of heat can be considered in terms of bare soil. The albedos (diffuse reflectivity) of most rocks, including soils, are in the range 0.1–0.3 (i.e., 10–30% of radiation from the sun is reflected back). Sandy soils have higher albedos than clays; dry ground has a higher albedo than damp ground. Hence, a dry sandy soil reflects back about twice as much radiation as damp clay. Dark rocks absorb radiation more easily than light-coloured rocks, and then re-radiate when the sun sets or is hidden by cloud. The importance of this re-radiation process is enhanced by the fact that the re-radiation has a longer wavelength, i.e., is more warming, than the incoming radiation.

The value of re-radiation is well known to vineyard managers. In the Schlossböckelheim district of the Nahe, the growers pile pebbles and boulders of the feldspar porphyry beneath the rows of vines. In the Mosel, the dark phyllite, with a particularly high albedo, has long been used to enhance this effect.

However, it should be noted that the strength of re-radiation is inversely proportional to the square of height above the ground, i.e., doubling the height of the grapes quarters the energy received by them. In Chateauneuf du Pape, where many vineyards are covered with pebbles of quartz, many of the bunches of grapes are no more than 0.2 m above the ground. This practice is even more marked for the sugar-rich grapes in the Montilla-Morilles district of south-west Spain.

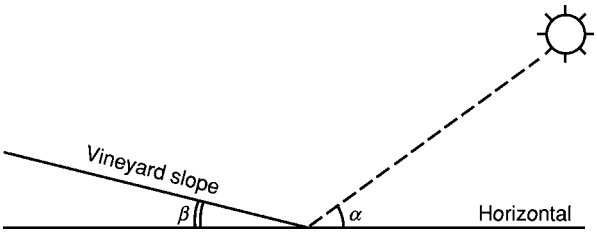


Figure 1 I $k \sin(\alpha + \beta)$, where I is the intensity of radiation received in the vineyard, α is the angular elevation of the sun, β is the angle of inclination of the vineyard to the horizontal along a meridian, and k is a constant. © Jake Hancock.

Figure 2 Vineyard of Schloss Bockelheimer Kupfergrube, north of Oberhausen, southern Nahe, Germany. An example of a European planned vineyard, only developed in the early twen tieth century, on a steep south facing slope to gain maximum advantage of direct radiation from the Sun in a relatively cool climate. © Jake Hancock.

Factors 1–4 can enhance the temperature during the growing season. There is also a slope factor during spring budding.

5. *Drainage of cold air.* The woody stems of vines can withstand extremely low temperatures but, once buds start to break from the stem, the vine is vulnerable to damage by even modest frosts. Bud-break is usually taken to be at 10°C but, depending on the variety of vine and the amount of nutriment stored from the previous season, budding may start at 3.5–7°C. Keeping in mind that the dates of the 'ice-saints', 11–14 May, are long after bud-break in northern Europe, late frosts can be a serious problem. The slope needed to provide a natural protection from still cold air depends on how severe the frost will be. In the Napa Valley in California, a slope as little as 2.5° may be sufficient to allow the freezing air to drain away. In Chablis, in northern France, a slope twice as steep as this may not be adequate to cope with the more severe frosts.

Water Balance

Like all plants, vines need water in the ground. However, without actually being a species of arid climates, its need for water is small. As a result, vines will grow almost anywhere. Nevertheless, to produce a good crop of grapes, the plant needs a small steady supply of moisture from the ground.

Without artificial supplies of water, e.g., by irrigation, the ideal drainage involves a high porosity for storing water, a low matrix permeability to stop it draining away, and a high mass permeability to ensure that excess rainwater drains away fast. A facies with ideal drainage is chalk, a distinctive, very fine-grained limestone, which occurs in much of Champagne. Typical chalks have a porosity of 35–45%, a matrix permeability of 2–6 mD, and a very high, but variable, mass permeability (normally more than 150 mD and can be several thousand millidarcy, i.e., more than a darcy).

For the vine, such a high porosity and low matrix permeability mean that plenty of water is held in the pores of the sediment, which cannot escape under gravity because the pores are less than 30–60 μm. However, plants can exert suctions equivalent to 100 m or more of water, i.e., the vines can extract water which cannot escape under gravity. In addition, the low matrix permeability means that the chalk does not easily dry out.

However, when there is very heavy rain, the mass permeability becomes valuable: it ensures that the large volume of rainwater drains away without leaving the sediment, and hence the roots of the vine, waterlogged (**Figure 3**).

For simple survival of a vine, waterlogging is a greater danger than a shortage of water, because the limited root surface area can be poisoned from a shortage of oxygen in solution. If the vine has used all the oxygen in stagnant water, the root metabolism starts to form poisonous alcohols. Clearly, it is essential that even the lowest roots of the vine are well above the top of the water table. As the roots normally penetrate downwards to several metres, and can penetrate to 20 m, vineyards can only be successfully established many metres above the normal local water table. In the Médoc, land drainage is used in vineyards which are closer to the Gironde Estuary. In the Barossa, in South Australia, where low rainfall makes drip-irrigation helpful, it is essential that open reservoirs are many metres below the height of the vineyards themselves.

Figure 3 Muscovite rich silty sand in a vineyard of Michel Torino, Cafayte, south south west of Salta, Argentina. A reliable soil for vines: medium porosity with relatively high permeability, holding water for roots, but allowing it to drain away before all dissolved oxygen has been lost. It will need some nitrogen for optimal growth. The muscovite, although not a direct source of potassium itself, probably indicates that adequate potassium is present from the weathered muscovite and illite. © Jake Hancock.

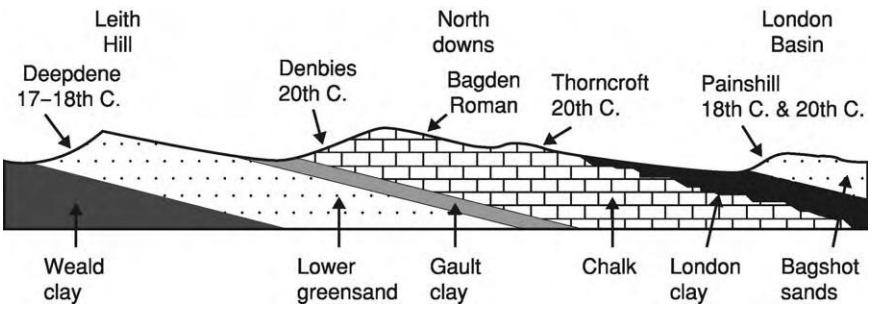


Figure 4 South north cross section from the Wealden anticline into the London basin of Surrey, south east England, showing the location of vineyards ancient and modern. Note that though they planted on rocks of different ages and types, vineyards have been located on well drained sunny south facing slopes. From Selley (2004). Courtesy of Petravin Press.

There is another complication with the drainage for vines because of their long roots. In many vineyards, most of the nourishment is obtained in the top 0.5 m, i.e., from the soil. Moisture for everyday metabolism can be obtained from much lower levels, possibly from superficial sediments, but commonly also from bedrock. Each of these three may have their own values of porosity and permeability. Of course, the roots take in nourishment and water from all levels, but the soil is likely to have the most nourishment available, and the bedrock may have the most steady and reliable supply of simple water. The control of geology on vineyard location is demonstrated in northern Surrey, England. Here vineyards have been planted over two millennia on rocks of various ages and types, but always on well-drained sunny south-facing slopes (Figure 4).

Vine Nourishment

Many of those who believe in geological '*terroir*' feel that it must be something in the nourishment from the ground which gives character to the wine. However, of the three geological influences, nourishment is the least important. Only the three major elements required by plants will be considered here.

1. *Nitrogen*. A nitrogen-rich soil makes for a flabby wine. In practice, a deficiency in nitrogen is the most widespread problem in the nourishment of vineyards. However, geology has little to contribute. There is some nitrogen in primary rocks, typically 10–12 $\mu g\ g^{-1}$, but almost all the nitrogen for plants is supplied as NH_4^+, NO_2, and NO_3, formed in a complex series of actions by heterotrophic bacteria and fungi in the soil.
2. *Phosphorus*. The most common phosphorus mineral is fluorapatite, usually just named apatite: $Ca_5(PO_4)_3F$. Biogenic apatite usually contains some OH^- and/or CO_3^{2-}. The solubility is very low: a typical soil contains 0.3 mg of phosphorus per litre of soil solution, most effectively as H_2PO_4; the double negative charge of HPO_4^{2-} makes it less effective because of the negative charge on the roots.

   Vines need very little phosphorus and a phosphorus deficiency is almost unknown. Nevertheless, the relatively small surface area of the roots means that vines would have difficulty in obtaining sufficient phosphorus, but for the fact that they belong to the group of plants which live symbiotically with mycorrhizal fungi. The numerous hyphae of the fungus penetrate the cortical cells (but not the protoplast) and form complex branches which increase the absorption area of the plasma membrane. This allows the root to absorb relatively large amounts of phosphorus. In return, the vine gives the fungus carbohydrates and vitamins. The effect is enhanced still further by the concentration of phosphorus in the fungi themselves, which is about three times the amount of phosphorus in the vines.
3. *Potassium*. Geology is a major controlling factor for potassium in vines. The plant takes up the element as the ion K^+. The vine seems to have no control on how much K^+ it absorbs: if there is a shortage of K^+, the vine suffers; if there is an excess of K^+ available, the roots simply absorb it.

   The most common potassium minerals are: muscovite, $K_2Al_4[Si_6Al_2O_{20}](OH,F)_4$; biotite, which differs in basic composition from muscovite only in containing Mg_6 instead of Al_4; K-feldspar (e.g., orthoclase and microcline), $K(AlSi_3O_8)$; and the clay mineral illite, $K_yAl_4(Si_{8-y}Al_y)O_{20}(OH)_4$ where $y < 2$.

   Muscovite is resistant to weathering, biotite is quantitatively insignificant, K-feldspar is important in some vineyards (see below), and illite is widespread and can yield K^+ by direct proton replacement. However, in most vineyards, as with

most agricultural products, K^+ comes not from the weathering of primary minerals, but from ions held by electrostatic attraction to the negatively charged colloidal-sized particles of clay minerals. The ease with which a clay mineral can release cations back into solution for plant roots is known as the cation exchange capacity (CEC). The CEC of the smectite group clay minerals is high (47–162 $cmol_c\ kg^{-1}$), and this can be valuable in some vineyards, e.g., on the weathered microgranites of Beaujolais, and possibly in soils derived from weathered volcanic rocks in California. The CEC of the illite group is only 20–40 $cmol_c\ kg^{-1}$, but illite is more important in most vineyards because of the high illite content in the clays and marls of districts such as Sancerre and the Côte d'Or of Burgundy. Moreover, as the surface K^+ is removed by the roots, some K^+ can be replaced from the illite itself.

There is a common myth that, because K^+ has a high solubility, it is easily washed out of a soil. This ignores the power of illite and smectite particles to seize K^+. A single dose of a potassium fertilizer can affect the fertility of the ground 50 years after it has been applied.

Although cationic exchange from illite is probably the main source of K^+ in vineyards, a remarkable number of high-quality wines come from vineyards which may obtain some K^+ from the weathering of K-feldspar: the Méric Conglomerate in the Médoc; the granite of Hermitage; the feldspathic sandstones of the Permian Rotliegend west of Nierstein in the Rheinhessen, and some of the grands crus vineyards in Alsace; the feldspar porphyry of Schlossböckelheim in the Nahe; the alluvial sediments in the Napa valley derived from volcanic and pyroclastic rocks lining the sides of the valley; and the lateritic soils developed on granite–gneiss in the Margaret River country of western Australia.

Summary

The fact that vines grow on rocks of all ages and types may superficially suggest that geology has no role to play in viticulture. Nothing could be further from the truth. The interplay of geology and climate determines the landscape within which a vineyard stands, and the soil on which it grows (Figure 5).

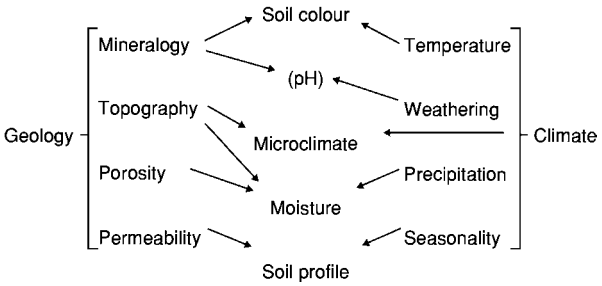


Figure 5 Soil, referred to as 'the soul of the vine' (Wilson, 1998) results from the complex interplay of geology and climate. Note that in many vineyards it is the superficial deposits, such as alluvium, eolian sand or boulder clay that determine the charac ter of the soil, as much as the solid rock beneath. From Selley (2004). Courtesy of Petravin Press.

See Also

Geology of Beer. **Geology of Whisky**. **Sedimentary Rocks:** Chalk; Clays and Their Diagenesis; Phosphates; Sandstones, Diagenesis and Porosity Evolution. **Soils:** Modern.

Further Reading

Amerine MA and Winkler AJ (1944) Composition and quality of musts and wines of Californian grapes. *Hilgardia* 15: 493 675.

Asselin C, Flanzy C, Sapis JC, and Martin JC (eds.) (1997) *Les Terroirs Viticoles Concept, Produit, Valorisation*. Angers: Institut National Recherches Agronomiques.

Coomb BG and Dry FR (eds.) (1988) *Viticulture, 1. Resources*. Adelaide: Winetitles.

Geiger R (1961) *Das Klima der Bodennahen Luftschicht* (translated as *The Climate Near the Ground*). Cambridge: Harvard University Press.

Gladstones J (1992) *Viticulture and Environment*. Adelaide: Winetitles.

Hancock JH and Selley RC (2003) *Coquand's joke*. London: Geoscientist. 13.10.17.

Jackson RS (2000) *Wine Science*, 2nd edn. San Diego: Academic Press.

Pomerol C (1989) *Terroirs et Vins de France*, 2nd edn. (translated as *The Wines and Winelands of France*). London: Robertson McCarta.

Selley RC (2004) *The Winelands of Britain: past, present & prospective*. Dorking: Petravin Press.

Unwin T (1991) *Wine and the Vine*. London: Routledge.

Wild A (ed.) (1988) *Russell's Soil Conditions and Plant Growth*, 11th edn. Harlow: Longman.

Wilson JE (1998) *Terroir*. London: Mitchell Beazley.

GEOMORPHOLOGY

P H Rahn, South Dakota School of Mines and Technology, Rapid City, SD, USA

Introduction

Geomorphology is the study of landforms and of the processes that act on Earth's surface to produce these landforms. This article emphasizes the geomorphologic processes that form different types of surficial deposits. For the current purposes, the range of surface processes can be considered with reference to the following four topics: fluvial processes/floods, mass wasting/landslides, tectonic terrains, and glaciation.

Fluvial Processes/Floods

Streams play an important role in erosional processes. The surface of Earth, over geological time, is continually and gradually eroded, and many landforms are a result of stream erosion or deposition. During the erosion process, when a stream reaches base level it begins to meander (Figure 1) and carves out a flood plain, an area that is subject to flooding (Figure 2) (*see* **Sedimentary Processes:** Fluvial Geomorphology). Only a small percentage of Earth's surface is subject to flooding. The degree of inundation depends on subtle topographic features such as natural levees, oxbows, local alluvial fans, low terraces, and mass-wasting accumulations from adjacent hill slopes. Identification of these features requires detailed geomorphological scrutiny using aerial photographs and topographic maps.

A stream terrace is an abandoned flood plain that results from a lowered base level. Terraces may be paired or unpaired. The continuity of a terrace along a valley, paired on both sides, may be correlated to a former graded condition. On the other hand, progressive stream down-cutting will leave isolated unpaired terraces along the valley sides. The surface elevation of terraces cannot be correlated because the terraces were never portions of a single, continuous surface. Unpaired terraces imply continuous down-cutting accompanied by lateral erosion. The stream shifts back and forth from one side of the valley to the other, and the valley floor is gradually lowered.

Figure 1 Meanders and flood plain of the Wind River near Dubois, Wyoming.

Terraces are of special interest as flood hazards because some very low terraces may be inundated during floods. For example, alluvial stratigraphy and geomorphology were used to evaluate the inundation of low terraces along the Connecticut River during a flood that occurred in 1936. It was shown that the riverine terraces were hazardous but that the higher kame terraces were not inundated. Another example is the Missouri River. A Landsat image of Glasgow, MO in September 1992 (Figure 3A) shows the Missouri River meandering on the flood plain and the terraces between the confining bluffs. A later image (Figure 3B) was taken in September 1993, during severe flooding. The entire flood plain and some low terraces are inundated; though one terrace on the north side was above the water (*see* **Sedimentary Processes:** Catastrophic Floods).

The identification of flood-prone terraces may be complicated by changes in aggradation or degradation. Such changes can be natural processes or can be caused by humans. Cyclic episodes of aggradation or degradation can be related to overgrazing or climatic change. A low terrace may be practically indiscernible from the flood plain, and, unfortunately, geological maps rarely discern these geomorphological features. Instead, the maps simply classify bottomland as Quaternary alluvium (Q_{al}). Rivers can have complex terrace systems. For example, the active flood plain width of the lower Mississippi River is approximately 16 km, lying within a terraced valley floor varying from 40 to 200 km wide. The terraces along the lower Mississippi River were formed by degradation accompanying sea-level lowering during glacial maximum, whereas the terraces along the upper Mississippi river were formed by degradation that accompanied reduced sediment loads during interglacial time.

Planners and engineers responsible for construction projects often overlook the geomorphology and interpretation of surficial deposits. For example, the

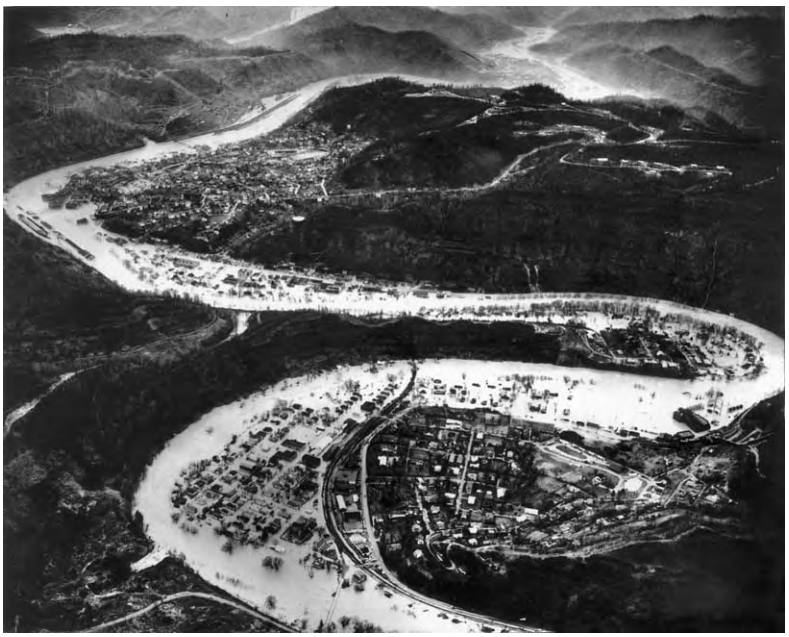


Figure 2 Oblique aerial photograph of the Kentucky River during the 1963 flood at Hazard, Kentucky. There is a clear division between the flood plain inundated by the 1963 flood and the adjacent hills. Courtesy of the Geological Society of America.

Figure 3 Landsat images of Glasgow, Missouri. (A) Image taken in September 1992. (B) Image taken in September 1993, during the flood peak. Courtesy of the United States Geological Survey.

United States Federal Emergency Management Agency requires communities to identify the so-called floodway based on the estimated 100-year flood discharge. Engineering firms use computer programs to estimate roughness and stage for this flood, and superimpose this stage ('base flood elevation') on topographic maps. The estimate of the 100-year flood discharge, the quality of the topographic maps, and the other assumptions of the computer programs limit this hazard evaluation. Further, this process does not protect flood-prone areas just above the 100-year flood stage.

An example of the usefulness of geological maps showing surficial deposits can be seen in Figure 4. In

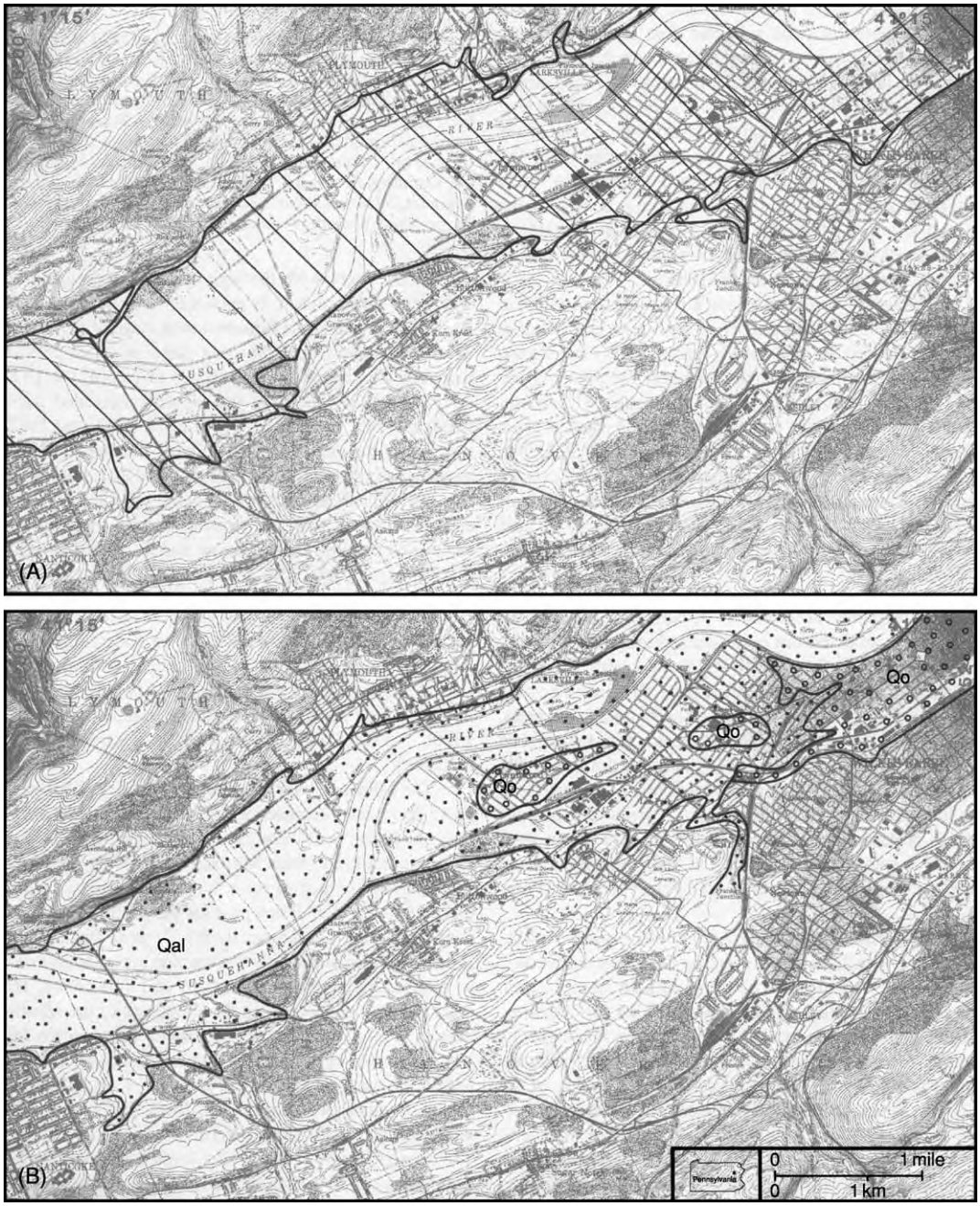


Figure 4 Maps of the area around Wilkes Barre, Pennsylvania. (A) Portion of a Wilkes Barre topographic map. The topography of the area has been affected by coal mining; culm banks more than 30 m high and subsidence pits are abundant. The area inundated by the June 1972 flood is indicated by diagonal lines. Courtesy of the United States Geological Survey. (B) Geological map of the same area. Quaternary alluvium (Q_{al}) and outwash (Q_o) are indicated. Courtesy of the Pennsylvania Geological Survey.

1972, at Wilkes-Barre, PA, the Susquehanna River crested at 12.4 m and the downtown area (Figure 4A) was inundated. The 1972 floods, resulting from Hurricane Agnes, caused over $3 billion damage in the northern Appalachian Mountains region. A geological map of Wilkes-Barre (Figure 4B) shows glacial outwash overlain by Holocene alluvium within the alluvial plain of the Susquehanna River valley. The coincidence of the two maps shows the utility of good topographic and surficial geological maps.

Mass Wasting/Landslides

Mass wasting includes several gravity-driven processes that act on Earth's surface. These hill-slope processes deliver material to a stream that ultimately carries the eroded debris to the sea. Mass-wasting processes include debris flow, landslides, and creep. Debris flows typically originate in semi-arid areas where thick soils or surficial deposits begin moving as a landslide and become more mobilized as they move into a stream channel. Eventually, they flow like wet concrete. Many alluvial fans include areas covered by prehistoric debris-flow deposits and hence are susceptible to continued activity. The levees associated with these deposits may get overtopped and exceed the stage of a 100-year flood as determined by stream discharge probability. Areas susceptible to debris flow require detailed geomorphologic analysis to predict the degree of hazard. A rock glacier is a form of debris flow; it is typically found in a glacial cirque and may have an ice core. Figure 5 is a radar image 20 km east of Seattle, WA. High-resolution light-detection and ranging (lidar) topography is a technology utilizing aircraft rangefinders to determine distance with a laser pulse. Topographic features are enhanced, even in forested terrain.

Landslides can occur within minutes, but many move very slowly, even over a period of years. Figure 6 shows some of the many landslides that occurred 130 km from the epicentre of the 1964 Alaska earthquake (M = 8.6). The landslides developed in the Bootlegger Cove Clay, a Quaternary marine clay. Careful mapping of surficial deposits is an important job for the engineering geologist in earthquake areas. Landslides are particularly common on steep hill slopes with clayey soils, or where a hill slope is oversteepened by shoreline processes or a river meander. Above a wave-cut cliff, an entire hill slope may be a slowly moving complex of landslides undergoing progressive failure. Translational landslides typically occur along bedding planes of moderately to steeply dipping sedimentary rocks. Geological maps, aerial photographs, and detailed fieldwork are needed to identify landslide hazards (*see* **Sedimentary Processes:** Landslides).

Gravity continuously acts on a hill slope, and weathered debris is dragged down slowly, forming colluvium. In humid areas, colluvium practically covers the whole land surface, obscured only where an occasional outcrop protrudes or a surficial deposit is present. Some landslides develop in colluvium, so the processes of creep and landslide are intermingled. Engineering geologists can help evaluate the stability of hill slopes for catastrophic landslides or minor damage from creep. The classic landslide is spoon shaped and has a bulging toe and a scarp at the top. Ancient landslides may still show this distinct topographic pattern.

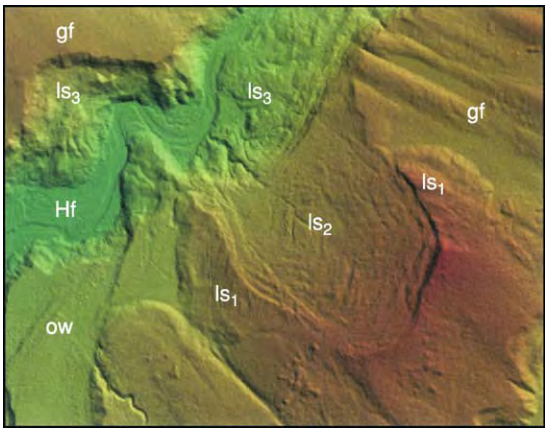


Figure 5 Lidar topography showing landslides near Seattle, Washington. The view is 3.5 km wide. Symbols: ow, outwash; ls, landslides (1 youngest, 3 oldest); Hf, Holocene flood plain; gf, glacially fluted surface. Courtesy of the Geological Society of America (see *GSA Today*, November 2003, pp. 4 10).

Figure 6 Landslides at Anchorage, Alaska, caused by the 1962 Alaska earthquake. Courtesy of the United States Geological Survey.

Residual soils develop *in situ* from the weathering of the parent bedrock (*see* **Soils:** Modern). Saprolitic deposits develop soil-like attributes and can accumulate to great thickness in humid areas such as the south-eastern United States. Transported soils, on the other hand, are simply surficial deposits that were moved and deposited by glaciers, running water, wind, or coastal processes. These geomorphologic processes produce deposits that have distinct attributes. For example, loess (wind-blown silt) is noted for its property of standing in vertical excavations. Till, unsorted debris deposited directly by a glacier, covers vast areas of northern Europe and North America. Till is noted for its poor drainage and often forms swamps. Its low permeability is well known, making it useful for ameliorating groundwater contamination plumes. Engineering geologists who work in glacial terrain need to understand the origin and characteristics of glacial deposits. A good working knowledge of the Quaternary deposits in a glaciated region is important in engineering endeavors and a valuable asset to a geotechnical firm.

Tectonic Terrains

Much of Earth's surface is the result of active tectonism. Mountains such as the Alps and Himalayas have been formed by tectonic uplifts . Normal faults in the western USA have produced a distinctive 'basin and range' topography; alluvial fan complexes form in the grabens adjacent to the uplifted mountains. Active strike–slip faults in California have resulted in diverse topographic features. Figure 7 is an oblique aerial photograph showing "basin and range" topography at the Wasatch Mountains near Salt Lake City, UT. A geomorphologist would quickly grasp the significance of this remarkable photograph. It shows that the terminal moraine of a mountain glacier has been offset by a fault. The moraine obviously is late Pleistocene in age, probably deposited around 20 000 years ago. Thus the fault displacement is relatively recent, and the fault must be considered active. The area along the fault will be offset again. The damage from the earthquake will not be limited by the shearing of any road or structure directly on the fault trace, but will also affect nearby areas. The response of earth materials to ground shaking varies. Figure 6, for example, shows the response of clay deposits at a great distance from the epicentre of the 1964 Alaska earthquake: the marine clay broke into hundreds of landslides.

Figure 7 Oblique aerial photograph of a faulted terminal moraine, Salt Lake City, Utah. Courtesy of the Utah Geological Survey.

Geological deposits have a wide range of responses to earthquakes. Surficial deposits are especially susceptible to shaking. For example, in the 1989 Loma Prieta earthquake (M = 7.1) in the San Francisco Bay area, damage to houses was widespread where thick residual soils occur in the mountainous terrain. Damage to the Oakland waterfront was related to infilling along the wharves. Condominiums collapsed in the Marin District of San Francisco in areas underlain by artificial fill. Another example is at Kobe, Japan, where liquefaction of alluvium and lagoonal clay occurred during the 1995 earthquake (M = 7.2). Clearly, an important aspect of engineering geology in earthquake-prone regions is the identification of sensitive surficial deposits.

Glaciation

There are two general types of glaciers: mountain glaciers and continental glaciers (*see* **Sedimentary Processes:** Glaciers). Mountain glaciers form distinctive erosional and depositional topographic features. Terminal moraines mark the terminus of Pleistocene mountain glaciers (Figure 7). Vast continental icesheets covered much of the land of the northern hemisphere and have left surficial deposits, including till and stratified drift. The best aquifers in glaciated terrain are stratified drift. Figure 8 shows an area of eastern Connecticut where stratified drift primarily occurs in the form of kame terraces. Where saturated, these constitute excellent aquifers, typically yielding over 500 gallons minute ^{1} (2500 m^3 day ^{1}) to wells. Much of the uplands in Figure 8 are bedrock (schist and gneiss) that is thinly veneered with glacial till. Drumlins are found throughout much of the till-covered areas in New England. The yield of wells in these upland areas is very low, on the order of 1 gallon minute ^{1} (5 m^3 day ^{1}). The kame terraces, primarily those along the Willimantic River valley, are excellent aquifers.

Other surficial deposits are associated with glaciation, including loess (wind-blown silt) and deltaic deposits. Figure 6 shows a large mass of varved clay that was deposited below sea-level during the

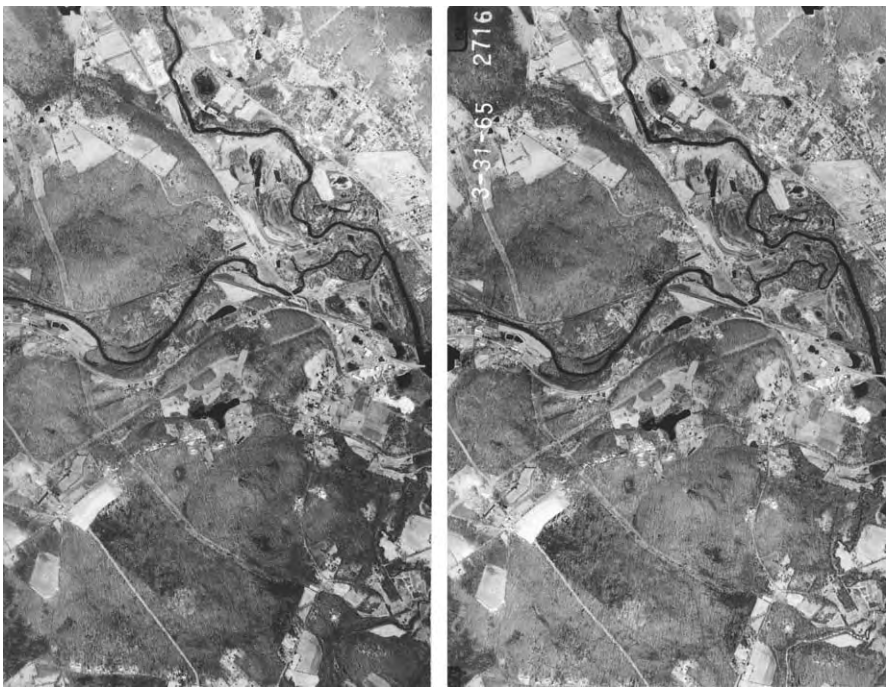


Figure 8 Stereo aerial photographs of glaciated terrain near Willimantic, Connecticut. Stratified drift, primarily kame terraces, are found in the Willimantic River valley; some are being quarried for sand and gravel. Best viewed with a pocket stereoscope.

Pleistocene Epoch. Much of Anchorage, AK, was destroyed when the 1964 Alaskan earthquake (M = 8.3) shook these deposits and they broke into hundred of landslides.

See Also

Engineering Geology: Geomorphology; Liquefaction. **Sedimentary Processes:** Catastrophic Floods; Fluvial Geomorphology; Glaciers; Landslides. **Soils:** Modern. **Tectonics:** Earthquakes.

Further Reading

Coates DR (1976) *Geomorphology and Engineering*. Stroudsburg, PA: Dowden, Hutchinson and Ross.

Costa JE and Baker VR (1981) *Surficial Geology, Building with the Earth*. New York: John Wiley & Sons.

Eckel EB (1970) *The Alaskan Earthquake, March 27, 1964, Lessons and Conclusions. US Geological Survey, Professional Paper 546*. Washington, DC: US Geological Survey.

Fookes PG and Vaughan PR (eds.) (1986) *A handbook of Engineering Geomorphology*. New York: Chapman and Hall.

Jahns RH (1947) *Geologic Features of the Connecticut Valley, Massachusetts, as Related to Recent Floods. US Geological Survey, Water Supply Paper 996*. Washington, DC: US Geological Survey.

Kiersch GA (ed.) (1991) *The Heritage of Engineering Geology; the First Hundred Years. Geological Society of America, The Decade of North American Geology, Centennial Special Volume 3*. Boulder, CO: Geological Society of America.

Leopold LB (1994) *A View of the River*. Cambridge, MA: Harvard University Press.

McNutt SR and Sydnor RH (1990) *The Loma Prieta (Santa Cruz Mountains), California, Earthquake of 17 October 1989. California Division of Mines and Geology, Special Publication 104*. Sacramento, CA: California Geological Survey.

Rahn PH (1996) *Engineering Geology, an Environmental Approach*. Upper Saddle River, NJ: Prentice Hall.

Ritter DF, Kochel RC, and Miller JR (1995) *Process Geomorphology*. Dubuque, IA: William C Brown.

Royster DL (1973) *Highland Landslide Problems along the Cumberland Plateau in Tennessee. Association of Engineering Geologists, Bulletin 10*, pp. 255 287. Boulder, CO: Association of Engineering Geologists.

Underwood JR and Guth PL (eds.) (1998) *Military Geology in War and Peace. Reviews in Engineering Geology*, vol. 13. Boulder, CO: Geological Society of America.

Verstappen HT (1983) *Applied Geomorphology*. Amsterdam: Elsevier.

Waltham AC (1994) *Foundations of Engineering Geology*. London: Blackie Academic and Professional.

West TR (1995) *Geology Applied to Engineering*. Englewood Cliffs, NJ: Prentice Hall.

GEOMYTHOLOGY

A Mayor, Princeton, USA

Introduction

Geomythology (also called 'legends of the Earth', 'myths of observation', 'natural knowledge', and 'physico-mythology') is the study of etiological oral traditions created by pre-scientific cultures to explain, in poetic metaphor and mythological imagery, geological phenomena such as volcanoes, earthquakes, floods, fossils, and other natural features of the landscape. (A related field, 'archaeoastronomy', studies pre-scientific knowledge of sky phenomena.)

Two types of geomyth have been identified: folk explanations of notable geological features, and often-garbled descriptions of catastrophic geological events that were witnessed in antiquity. In the case of geomorphic events that occurred in the pre-human past, observation and imagination led to mythic explanations that were handed down over millennia. In the case of natural cataclysms within living human history, descriptions were transmitted over generations, often accruing supernatural details. Because of the mythological language of oral folklore, scientists and historians have often missed the kernels of truth and rational concepts embedded in geomythological narratives. Some geomyths are simply fanciful stories based on imagination or popular misconceptions, such as tales of creatures or humans that were magically transformed into rock to explain the shapes of landforms. Many geomyths, however, contain surprisingly accurate insights into geological processes, as well as important eyewitness data from the distant past. Modern scientific investigations have revealed that much ancient folklore about the Earth was based on rational speculation and understandings grounded in careful observations of genuine, but extraordinary physical evidence over time.

Geomythology in Classical Antiquity

Although the term 'geomythology' was coined in 1968 by the geologist Dorothy Vitaliano, and it is often considered a new field of study, the concept was known and applied since antiquity. Euhemerus, a Greek philosopher (*ca.* 300 BC), held that myths about divinities and their activities were poetic accounts of real people and events. His approach, later called 'euhemerism', was taken up by other classical scholars who rationalized myths by stripping away supernatural and impossible details to reveal an underlying core of facts. Some of the rationalizing deconstructions of hero and monster myths by the Greek euhemerist Palaephatus (fourth century BC) may seem contrived, but others, such as his interpretation of the myth of Cadmus sowing the dragon teeth, are quite sophisticated. Palaephatus suggested that the tale represented an ancient misunderstanding of fossil elephant molars, which were frequently found in the ground and treasured by kings in archaic Greece before knowledge of elephants was brought back from India by Alexander the Great in the fourth century BC.

Many other classical Greek and Latin writers identified archaic myths about creation and primal creatures, first written down in the eighth century BC by the epic poets Homer and Hesiod, as symbolic ways of describing actual events and processes in deep time or within human memory. In his account of Atlantis and in other works, for example, the philosopher Plato (fourth century BC) correctly described large-scale changes in prehistoric land masses and coastlines in the Aegean. In the first century BC, the Latin poet Ovid expressed accurate conceptions of geomorphology and the process of petrifaction in his *Metamorphoses*; verses about the transformations of mythic beings. Around the same time, the Greek geographer Strabo regarded traditional geomyths as cryptic historical records, observing that "the ancients expressed physical notions and facts enigmatically by adding mythical elements".

In his fifth century BC tragedy *Prometheus Bound*, the Greek dramatist Aeschylus recounted the myth of Zeus burying Typhoeus (a monstrous, many-headed dragon with many voices that embodied primal chaos) under Europe's largest and most active volcano, Mount Etna in Sicily. The etiological account explains Etna's eruptions as Typhoeus's struggles to escape the subterranean prison: his roars and hisses were the auditory features of the volcano, and his fiery breath was supposed to melt rock, creating the periodic lava flows that endangered towns on the volcano's slopes. In the first centuries BC and AD, the Latin epic poet Virgil and the natural historian Pliny the Elder suggested that the one-eyed giant Cyclopes described by Hesiod in his *Theogony* (*ca.* 750 BC) personified other active volcanoes in the Mediterranean.

Contributions of Geomythology to Modern Science

Robert Hooke was the first modern scientist known to have employed classical geomythology as historical evidence. To support his theory of vast geological alterations over time, in lectures presented in the period 1667–1688 Hooke drew on geomythology of landform changes and earthquake-related lore recorded by Plato, Pliny, Strabo, Virgil, and Ovid (as well as biblical scriptures). Hooke sought the empirical evidence of mythic narratives to show that great floods and earthquakes had repeatedly transformed the Earth in the deep past, and could justify the presence of petrified marine remains on mountaintops. His explanatory hypothesis was based on the idea of pole wandering (to be tested astronomically), which was not well received by some of his contemporaries. It was for this reason that he collected evidence from myths in support of his ideas about Earth history.

Another early geomythologist was the founder of comparative anatomy, Georges Cuvier (*see* **Famous Geologists:** Cuvier). He compiled a collection of ancient Greek accounts and North and South American Indian traditions about the discoveries of petrified bones of remarkable size to demonstrate the worldwide distribution and longstanding observations of the fossilized remains of immense creatures, which Cuvier identified as extinct elephant-like creatures. Edward Burnet Tylor (1865) was another early pioneer of geomythology. He called traditional legends about natural history 'myths of observation' to emphasize that they were reasonable efforts to account for mysterious physical evidence.

Vitaliano was the first modern geologist to systematically match the insights contained in various cultures' myths about geology to modern scientific knowledge, thereby giving the study of folklore a disciplinary status. As the first stirrings of geological observation and hypothesis-forming that would later evolve into the Earth sciences, geomyths are significant milestones in the history of science. As Vitaliano and others have noted, scientific theories themselves are analogous to etiological geomyths in that both are efforts to explain mysterious observed facts. The ability to link traditional descriptions with present-day science is a notable contribution to scientific knowledge. Geomyths can provide previously unknown additional data for studying geological events that were actually witnessed in the pre-scientific past. The study of geomythology also helps reveal the processes of transmitting cultural memories over many generations and the origins and functions of oral mythopoesis.

Scientific analysis of geomyths can verify the historical foundations of many myths previously viewed as imaginary products of creative storytelling. For example, in 1999, the frozen mummy of Kwaday Dan Sinchi (meaning 'long ago person found') was discovered in a melting glacier between Yukon Territory and north-western British Columbia, Canada. Radiocarbon dating showed that the young man had lived in the 1400s. He carried a waterproof hat woven of roots from the Pacific coast, a leather bag of dried fish and plants, tools made from both coastal and inland trees, and a cloak made from gophers that live far inland. These finds confirmed the ancient oral traditions of the local Champagne and Aishihik First Nations, which describe their ancestors using the glaciers as trade routes to travel between the interior and coast.

Examples of Geomythology

Geomyths from around the world explicate the gamut of geological phenomena, from seismic and volcanic events to fossil deposits, such as shells and marine creatures stranded far from the sea and strange, oversized, unfamiliar skeletons embedded in rock. Extraordinary landmarks and the sudden disappearance or appearance of islands; climatic changes; great floods and changing watercourses; natural petroleum fires and deadly gases emitted from the Earth; the formation of minerals and gems underground; and myriad other large and minute natural features of the landscape are all featured in geomythology.

For example, legends of deadly miasmas and 'birdless places' often arose in regions where toxic natural gases are released from vents in the Earth, affecting plants and wildlife. The ancient image of the cave-dwelling, fire-breathing monster, the Chimera, which supposedly dwelled in what is now modern Turkey was no doubt influenced by observations of spontaneously burning natural gas wells in Asia Minor. In 2002, a team of archaeologists and geologists confirmed a long-discounted classical Greek tradition that the priestess possessed by the god Apollo at the Oracle of Delphi was inspired by fumes emanating from a crack in the Earth. The team discovered that intoxicating methane and other gases escape from fissures at the ancient site of the Oracle.

Volcanoes and earthquakes are well represented in geomyths. The spectacular volcanoes of Hawaii inspired legends of the fire goddess Pele digging a series of great fire-pits as she traveled across the islands. Geologists point out that the legend reflects an ancient awareness that the volcanic activity from northwest to south-east was progressively younger.

The Greek myths of the cosmic wars between Zeus and the Titans, Cyclopes, and Typhoeus as described in Hesiod's *Theogony* were scientifically analyzed in 1992 by geology historian Mott T. Greene. Hesiod's poem contains some ancient oral stories that date to the second millennium BC. Greene demonstrated that the violent battle with the one-eyed Cyclops can be matched to volcanic phenomena associated with the solitary Mount Vesuvius and the solfataric gas emissions in the 'fields of fire' near Naples, Italy. In contrast, the god's conflict with Typhoeus and the Titans represents Mount Etna's multiple cones, and the hissing and roaring features, lava flows, and deep tectonic earthquakes. Moreover, details of Hesiod's poem suggest that it forms a chronological record of datable major eruptions of Etna in about 1500 BC and in 735 BC, and the Plinian eruption in 1470 BC on the island of Thera–Santorini, which destroyed the Minoan civilization.

Ancient Greek earthquake lore distinguished between local, weak volcanic tremors, attributed to struggling giants imprisoned by Zeus in the Earth, and large-magnitude tectonic quakes with associated tsunamis, which were attributed to Poseidon, the Earth-shaking god of the sea. In West Africa, where tectonic shocks typically emanate from the west, the natives imagined a giant who tires of facing east clumsily shifting position. Highly seismic Japan has elaborate quake myths, with some tales attributing the earthquakes and tsunamis to the movements of a colossal serpent-dragon surrounding the island. According to traditions in India and many other cultures, earthquakes were caused by giant creatures burrowing underground.

In Samos, an Aegean island with rich Miocene mammal fossils and subject to severe earthquakes, ancient Greeks devised an ingenious myth (fifth century BC) to explain both features. Before the era of humans, it was said that the island was populated by enormous monsters called Neades, whose deafening shrieks caused the very Earth to collapse upon them. According to ancient writers, their huge bones were displayed *in situ*, and archaeologists have found fossil relics in the ruins of the Temple of Hera on the island. By the first century AD the great bones were identified as the remains of Indian elephants brought to Greece by the god Dionysus. Not only are earthquakes on Samos distinguished by loud roaring, but many of the fossils on Samos are those of large, extinct elephant ancestors and the bone deposits are often found trapped underneath earthquake-faulted blocks.

Fossil remains generated a variety of geomyths speculating on the creatures' identity and cause of their destruction. Many ancient cultures, from China and India to Greece, America, and Australia, told tales of dragons, monsters, and giant heroes to account for fossils of animals that they had never seen alive. Some scenarios of their destruction in 'deep time' anticipated catastrophic extinction theories, first suggested scientifically by Cuvier, while other ancient accounts leaned toward more gradualist theories.

For the ancient Greeks, the Gigantomachy, the cosmic wars in which the gods destroyed giants and monsters and buried them underground, accounted for abundant deposits of the fossil skeletons of enormous, extinct Tertiary mammals found in 'giants battle fields' all around the Mediterranean. In the Siwalik foothills of the Himalayas, ancient Greek travellers reported that Indians displayed bizarrely horned dragons with sparkling gems embedded in their skulls. The origin of the myth came to light when nineteenth-century palaeontologists discovered rich fossils of giant giraffids and curiously tusked proboscids encrusted with calcite crystals. In China, the 'dragon' bones collected and ground into medicine turned out to be the fossils of extinct mammals and dinosaurs. In central Asia, the legend of the gold-guarding griffin, a creature with the body of a lion and beak of a raptor, arose as nomadic prospectors on their way to gold deposits came upon conspicuous fossils of beaked quadruped dinosaurs in the Gobi desert.

In Europe, observation of dinosaur tracks in Triassic sandstones in the Rhine Valley probably influenced the legend of the slaying of the dragon Fafnir there by the Germanic hero Siegfried. For Aborigines near Broome, north-west Australia, the footprints of Cretaceous carnosaur and stegosaur dinosaurs in sandstone near Broome, form a 'song-line' from dream-time. The trackways are considered the trail of a giant 'Emu-man' of myth. Where tracks head out to sea and back to shore, legend tells of him wading into the ocean and returning. Wherever he rested, Emu-man's feathers made impressions in the mud, a logical interpretation of fern fossils in sandstone, which resemble large feathers. According to an Aztec legend preserved by sixteenth-century Spanish explorers, the great feathered serpent-god Quetzalcoatl left his hand and seat prints in stone near Mexico City, one of the earliest geomyths recorded in America. Pleistocene fossils of large mammals are in fact abundant around ancient Aztec sites near Mexico City and the tracks of these probably account for the myth.

Myths of a devastating flood are nearly universal. Recently, a team of geologists studied sediments in the Black Sea region and concluded that Mesopotamian and Biblical flood myths originated when the rising Mediterranean suddenly broke through the Bosphorus,

inundating the populous farmlands of the Black Sea basin about 6000 years ago. The awareness that a vast sea once covered the American south-west in Cretaceous times is evident in Zuni Indian traditions about the bizarre huge marine monsters of a long-past era before humans evolved, whose remains are found along with shells and ripple marks in the desert bedrock. The Zuni creation myth describes how the sea was dried out by a great conflagration. Indeed, giant marine reptiles of the Cretaceous are found in Zuni lands, along with the burned stumps of great prehistoric forests in the desert.

Remarkable landforms have long elicited folk etiologies to explain their origins and notable features. Devil's Tower in Wyoming, a prominent volcanic formation in the American West with distinctive grooves and facets, was said by Native Americans to have been formed when a gigantic bear clawed at the rock in an attempt to reach children trapped on the top. Notably very large dinosaur claws are found in the region. The unusual Cuillin Mountains of Skye, Scotland, were fabled to have been formed when the Sun hurled his fiery spear into the ground. Where it struck, a huge blister or boil appeared and grew, swelling until it burst and discharged molten, glowing material that congealed to form mountains perpetually covered in snow. Geologists have remarked that the legend accurately recounts the formation of a volcanic dome, which grows, bursts, and spews glowing-hot magma. The Cuillins consist of gabbro, crystallized molten matter, and the adjacent mountains of granite, the Red Hills, are indeed snow-capped in contrast to the steeper Cuillins. However, the Cuillins are much too old to have been observed being formed by humans.

Controversies and Future Directions

In recent years, the horizons of geomythology have been expanded by Native American scholars who relate Amerindian geological traditions to modern scientific knowledge. These scholars also grapple with the controversial questions raised by geomyths: how far can human memory, perpetuated in spoken traditions over generations, extend back in time? Strabo was the first to address this issue in the first century BC. The very magnitude of time encompassed in folk memories makes legends seem incredible, he wrote, yet it is worth trying to decipher what the ancients understood and witnessed.

In the Renaissance, naturalistic approaches to the meaning of ancient geomyths vied with moralistic interpretations; and debates over the validity of traditional folk knowledge continued into the scientific era. For example, debate raged in the nineteenth century over Native American legends that seemed to contain ancestral memories of mastodons hunted to extinction in the last Ice-Age about 10 000 years ago. The possibility was supported by twentieth-century archaeological discoveries of mammoth kill sites that matched local tribal lore about elephant-like monsters.

One method of testing the reliability of oral geology traditions is to examine traditions about geomorphic events in specific geographic areas with datable chronologies. Greene's analysis of Hesiod's ancient volcano data provides an example of this approach. In 2003, geologists found that Homer's description of the landforms around Troy, now radically changed, is consistent with the way the region probably looked 3 millennia ago. Another convincing geomyth of surprising antiquity is the Klamath Indians' oral tradition about the largest Holocene eruption in North America, the volcanic explosion of Mount Mazama in the Cascades Range of southern Oregon. About 7500 years ago, the spectacular eruption blew off the top of the mountain and rained ash over a half million square miles. The resulting caldera formed Crater Lake. Surviving palaeo-Indian witnesses created a detailed oral tradition of the violent event, expressed in a mythological story that has been transmitted in the original Native American language over some 250 generations. The Klamath myth contains geological facts about the eruption and collapse of the mountain that were unknown to scientists until the early twentieth century.

Future directions in geomythology will continue the search for evidence of very early geological knowledge, based on logical reasoning or firsthand observations of natural phenomena, embedded in ancient mythologies. Some of the most interesting new research trends in geomythology are the investigations by anthropologists, psychologists, and folklorists into the evolutionary psychology of the oral mythmaking process itself, to learn more about the mechanisms of preserving and perpetuating ancestral human memories over millennia. Such studies may help solve an urgent geological dilemma that requires today's scientists to think geomythologically: the permanent and safe geological disposal of thousands of tons of highly radioactive nuclear waste from reactors and weapons production. The transuranic waste is expected to remain radioactive for 100 000 years. The plan is to bury the dangerous materials very deep in the Earth and guarantee that the sites remain undisturbed for at least 10 000 years. The vast geological and chronological scale of the project means that warnings to succeeding generations must survive in meaningful form until the year AD 12 000.

Scientific proposals for ensuring that inadvertent human intrusion will not occur at such burial sites in the far distant future have called for the creation of new, long-lasting geomythological 'traditions', with written and visual markers of menacing design, to indicate to future generations the grave perils of what lies buried underground.

See Also

Biblical Geology. **Famous Geologists:** Cuvier. **History of Geology Up To 1780**. **Tectonics:** Earthquakes. **Volcanoes**.

Further Reading

Barber E and Barber P (2005) *When They Severed Earth from Sky: How the Human Mind Shapes Myth*. Princeton: Princeton University Press.

Birkett K and Oldroyd D (1991) Robert Hooke, physico mythology, knowledge of the world of the ancients and knowledge of the ancient world. In: Gaukroger S (ed.) *Uses of Antiquity*, pp. 145 170. Dordrecht: Kluwer Academic Publishers.

Cataldi R, Hodgson SF, and Lund JW (eds.) (1999) *Stories from a Heated Earth: Our Geothermal Heritage*. Sacra mento: Geothermal Resources Council, International Geothermal Association.

Clark E (1952) *Indian Legends of the Pacific Northwest*. Berkeley: University of California Press.

Deloria V (1997) *Red Earth, White Lies*. Golden (CO): Fulcrum.

Echo Hawk RC (2000) Ancient history in the New World: integrating oral traditions and the archeological record in deep time. *American Antiquity* 65: 267 290.

Greene MT (1992) *Natural Knowledge in Preclassical Antiquity*. Baltimore: Johns Hopkins University Press.

Hale JR, de Boer JZ, and Chanton J (2001) New evidence for the geological origins of the ancient Delphic oracle (Greece). *Geology* 29: 707 710.

Kraft J C, Rapp G, Kayan I, and Luce JV (2003) Harbor areas at ancient Troy: sedimentology and geomorphology complement Homer's Iliad. *Geology* 31: 163 166.

Mayor A (2005) *Fossil Legends of the First Americans*. Princeton: Princeton University Press.

Mayor A (2000) *The First Fossil Hunters: Paleontology in Greek and Roman Times*. Princeton: Princeton University Press.

Mayor A and Sarjeant WAS (2001) The folklore of foot prints in stone: from classical antiquity to the present. *Ichnos* 8(2): 1 22.

Rappaport R (1997) *When Geologists were Historians, 1665 1750*. Ithaca: Cornell University Press.

Ryan W and Pitman W (1998) *Noah's Flood: The New Scientific Discoveries about the Event that Changed History*. New York: Simon & Schuster.

Tylor EB (1865) *Researches into the Early History of Man kind*. London: John Murray (reprinted 1964, Chicago: University of Chicago Press).

Vitaliano D (1973) *Legends of the Earth: Their Geological Origins*. Bloomington: Indiana University Press.

GEOPHYSICS

See **EARTH: Orbital Variation (Including Milankovitch Cycles); EARTH SYSTEM SCIENCE; ENGINEERING GEOLOGY: Seismology; MAGNETOSTRATIGRAPHY; MOHO DISCONTINUITY; PALAEOMAGNETISM; PETROLEUM GEOLOGY: Exploration; REMOTE SENSING: Active Sensors; GIS; Passive Sensors; SEISMIC SURVEYS; TECTONICS: Seismic Structure At Mid-Ocean Ridges**

GEOTECHNICAL ENGINEERING

D P Giles, University of Portsmouth, Portsmouth, UK

Introduction

Geotechnical engineering has recently (1999) been formally defined in a Memorandum of Understanding on the proposed unification of the British Institution of Civil Engineers Ground Board and the British Geotechnical Society. Appendix A of that memorandum, establishing the British Geotechnical Association, sets out the following definition:

> Geotechnical engineering is the application of the sci ences of soil mechanics and rock mechanics, engineering geology and other related disciplines to civil engineering construction, the extractive industries and the preserva tion and enhancement of the environment.

Geotechnical engineering plays a key role in all civil engineering projects, since all construction is built on or in the ground. In addition it forms an important part of extractive industries, such as open cast and underground mining and hydrocarbon extraction, and is essential in evaluating natural hazards such as earthquakes and landslides.

The use of natural soil and rock makes geotechnical engineering different from many other branches of engineering: whereas most engineers specify the materials they use, the geotechnical engineer must use the material existing in the ground and in general cannot control its properties.

In most cases the complexity of the geology means that the geotechnical engineer is dealing with particularly complicated and variable materials; their mechanical properties usually vary with time and are critically dependent on the water pressures in the ground, which can often change.

The geotechnical engineer does sometimes have the opportunity to specify certain properties or treatment of soils, rocks and other materials used in construction.

Geotechnics can thus be primarily considered as the science of the engineering properties and behaviour of rocks and soils. Geotechnical engineering can be considered as the professional practice and implementation of that knowledge contributing principally to the design of engineered structures in and on the ground (Figure 1).

Fundamental to geotechnical engineering are the study and practice of engineering geology, geomechanics (rock mechanics and soil mechanics), the design of foundations, the stabilization of slopes, the improvement of ground conditions, the excavation of tunnels and other underground openings, the analysis of ground behaviour, and the assessment of ground movements.

Soil Mechanics and Rock Mechanics

Soil mechanics (*see* **Soil Mechanics**) and Rock Mechanics (*see* **Rock Mechanics**), together known as 'geomechanics', involve the study and understanding of the physical properties and behaviour of rocks and soils. These properties will include material strength (in tension, compression, and shear), moisture content, porosity and permeability, and a description in engineering terms (including a description of the weathered state and of the rock and soil

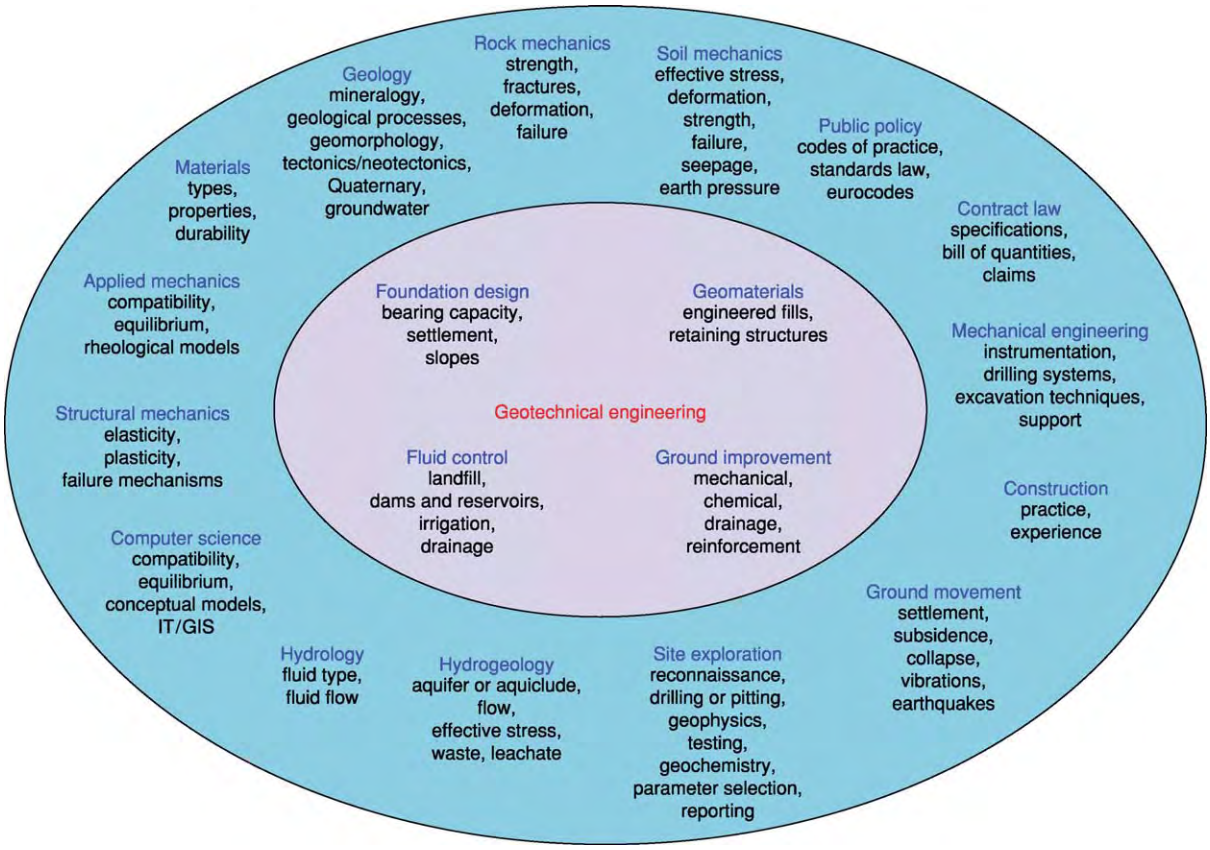


Figure 1 The practice of geotechnical engineering encompasses a wide variety of skills. Modified from the British Geotechnical Association (1999) Memorandum of Understanding, Appendix A, Definition of Geotechnical Engineering. *Ground Engineering*, Nov. p. 39 EMAP, London.

Table 1 Some typical geotechnical properties of engineering rocks

Rock type	*Unit weight (kN m ^{3})*	*Porosity (%)*	*Dry UCS range (MPa)*	*Dry UCS mean (MPa)*	*Saturated UCS (MPa)*	*Modulus of elasticity (GPa)*	*Tensile strength (MPa)*	*Shear strength (MPa)*	*Friction angle (φ°)*
Granite	27	1	50 135	200		75	15	35	55
Basalt	29	2	100 350	250		90	15	40	50
Greywacke Ordovician	26	3	100 200	180	160	60	15	30	45
Sandstone Carboniferous	22	12	40 100	70	50	30	5	15	45
Sandstone Triassic	19	25	5 40	20	10	4	1	4	40
Limestone Carboniferous	26	3	50 150	100	90	60	10	30	35
Limestone Jurassic	23	15	15 70	25	15	15	2	5	35
Chalk Cretaceous	18	30	5 30	15	5	6	0.3	3	25
Mudstone Carboniferous	23	10	10 50	40	20	10	1		30
Shale Carboniferous	23	15	5 30	20	5	2	0.5		25
Clay Cretaceous	18	30	1 4	2		0.2	2	0.7	20
Coal Carboniferous	14	10	2 100	30		10	2		
Gypsum Triassic	22	5	20 30	25		20	1		30
Salt Triassic	21	5	5 20	12		5			
Hornfels	27	1	200 350	250		80			40
Marble	26	1	60 200	100		60	10	32	35
Gneiss	27	1	50 200	150		45	10	30	30
Schist	27	3	20 100	60		20	2		25
Slate	27	1	20 250	90		30	10		25

UCS, uniaxial compressive strength.

Courtesy of Waltham AC (2001) Foundations of Engineering Geology. 2nd Edition. Spon Press. London.

en masse as well as material, a description of any discontinuities – such as fissures, fractures, joints, faults, and shears – and any other geological aspect of the mass fabric).

Geotechnical Engineering

Geotechnical engineering encompasses the civil engineering of the ground and deals with the interaction of that ground with engineered structures. The principal works undertaken are foundations, slopes, ground improvement, and underground excavations.

Foundations

The design and construction of foundations involves the calculation of bearing loads and capacities, the selection of the optimum foundation type (e.g. footing or pile), embankment design, retaining wall design, and seepage analysis. The geomechanical investigations required for foundation design will include evaluating soil stiffness, lateral earth pressures, consolidation coefficients, and material strengths, together with an analysis of the effective stress conditions.

Slopes

The study of slopes and landslides requires the design of cuttings, analysis of slope stability (in rock, soil, or debris), and consideration of slope stabilization measures (which may include slope drainage schemes, slope reprofiling, ground anchor and pinning systems, and other physical support methods).

Ground Improvement

Ground improvement includes mechanical and reinforcement measures, chemical treatment, drainage, bioengineering, and the use of geotextiles.

Underground Excavations

The design of tunnels and underground excavations involves the consideration of rock mass strength, the design of suitable excavation techniques (such as drill and blast, use of tunnel boring machines, use of roadheaders), and the design of suitable support systems (such as sprayed concrete (shotcrete), steel sets or supports, concrete segments, rock bolts). Arrangements must also be made to monitor any ground settlement, closure, or overstressing during the excavation operations.

Risk Analysis

The identification of geohazards and their incorporation in the risk management process requires hazard assessment and risk analysis (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards). This work underpins all aspects of modern geotechnical engineering practice. The geotechnical engineer will be responsible for compiling a risk register, which includes an analysis and assessment of all ground-related risks that could affect the engineering project together with a programme of suitable responses to prevent such risks occurring or, should such risks be realized, suitable mitigation and management measures that should be implemented if the event occurs.

Ground Investigation and Characterization

Geotechnical engineering is underpinned by the assessment and characterization of ground conditions using site and ground investigation data (*see* **Engineering Geology:** Site and Ground Investigation). Geotechnical engineers therefore need to work closely with engineering geologists to develop conceptual models appropriate to the ground conditions, including the geological setting of the site and the geotechnical variability (Tables 1–4). Detail is provided by the site and ground investigation process, building on the background knowledge provided by earlier investigations, research, and the records held by local and regional authorities (notably the British Geological Survey). Site-specific data is used to generate factual and interpretive accounts of the ground conditions, using both *in situ* data from intrusive test programmes and tests conducted on sample data obtained from extensive laboratory testing programmes. These data are used to create a conceptual model from which expected ground conditions can be anticipated. This data analysis provides the key geotechnical parameters and values that are used in the subsequent geotechnical design.

Table 2 Some typical geotechnical properties of engineering soils

Soil type	*Grain size (mm)*	*Liquid limit (%)*	*Plasticity index (%)*	*Friction angle (φ°)*
Gravel	2 60	N/A	N/A	>32
Sand	0.06 2	N/A	N/A	>32
Silt	0.002 0.006	30	5	32
Clayey silt	0.002 0.06	70	30	25
Clay	<0.002	35	20	28
Plastic clay	<0.002	70	45	19
Organic	Generally amorphous	>100	>100	<10

N/A, not appropriate.

Courtesy of Waltham AC (2001) Foundations of Engineering Geology. 2nd Edition. Spon Press. London.

Table 3 Some typical geotechnical properties of typical clay soils

State	*LI (%)*	*SPT (N)*	*CPT (MPa)*	*C (kPa)*	*m_v (m^2 MN^{-1})*	*ABP (kPa)*
Soft	>0.5	2 4	0.3 0.5	20 40	>1.0	<75
Firm	0.2 to 0.5	4 8	0.5 1	40 75	0.3 1.0	75 150
Stiff	0.1 to 0.2	8 15	1 2	75 150	0.1 0.3	150 300
Very stiff	0.4 to 0.1	15 30	2 4	150 300	0.05 0.1	300 600
Hard	<0.4	>30	>4	>300	<0.005	>600

LI, liquidity index; SPT, standard penetration test; CPT, cone penetration test; C, cohesion; m_v, compression coefficient; ABP, acceptable bearing pressure.
Courtesy of Waltham AC (2001) Foundations of Engineering Geology. 2nd Edition. Spon Press. London.

Table 4 Some typical geotechnical properties of typical sand soils

Packing	*Relative density*	*SPT (N)*	*CPT (MPa)*	*Friction angle (φ°)*	*SBP (kPa)*
Very loose	<0.2	<5	<2	<30	<30
Loose	0.2 0.4	5 10	2 4	3 32	3 80
Medium dense	0.4 0.6	11 30	4 12	32 36	8 300
Dense	0.6 0.8	31 50	12 20	36 40	3 500
Very dense	>0.8	>50	≥20	≥40	≥500

SBP, safe bearing pressure; CPT, cone penetration test end resistance; SPT, standard penetration test corrected N value.
Courtesy of Waltham AC (2001) Foundations of Engineering Geology. 2nd Edition. Spon Press. London.

Hydrology and Hydrogeology

Geotechnical engineering requires an understanding of and the ability to control fluids. Knowledge of both fluid type and fluid flow (hydrology) is necessary and builds on a knowledge of the geological controls (hydrogeology (*see* **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills)). This enables the design and construction of landfills, dams for impounding reservoirs, irrigation, and abstraction for water supply.

Geotechnical Modelling

Geotechnical modelling generally involves the construction of numerical and computational models of the ground conditions, permitting an evaluation of alternative engineering solutions for the design and construction of the proposed works. Computer-based techniques may use finite element, finite difference, distinct element, or particulate codes. These techniques are used to model, analyse, interpret, and visualize the variety of geotechnical processes and designs.

Geotechnical engineering is therefore a multidisciplinary subject that critically involves the understanding and application of soil mechanics, rock mechanics, and hydrogeology and is linked to the study of engineering geology. It is applied to civil engineering construction, mineral extraction, and the improvement of our environment.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards; Site and Ground Investigation; Ground Water Monitoring at Solid Waste Landfills. **Environmental Geology**. **Rock Mechanics**. **Soil Mechanics**.

Further Reading

Anon (1999) Memorandum of Understanding, Appendix A, definition of geotechnical engineering. *Ground Engineering* 32: 35 39.

Attewell PB and Farmer IW (1976) *Principles of Engineering Geology*. London: Chapman & Hall.

Barnes EE (2000) *Soil Mechanics: Principles and Practice*. Basingstoke: Palgrave Macmillan.

Bell FG (ed.) (1987) *Ground Engineer's Reference Book*. London: Butterworth Heinemann.

Brady B and Brown ET (1992) *Rock Mechanics for Underground Mining*. London: Chapman & Hall.

British Geotechnical Association (1999) Memorandum of Understanding, Appendix A, Definition of Geotechnical Engineering. *Ground Engineering*, Nov. p. 39. Emap, London.

Bromhead EN (1992) *The Stability of Slopes*. Glasgow: Blackie.

Clayton CRI (2001) *Managing Geotechnical Risk: Improving Productivity in UK Building and Construction*. London: Thomas Telford.

Clayton C, Matthews M, and Simons N (1995) *Site Investigation: A Handbook for Engineers*. UK: Blackwell.

Craig RF (1997) *Soil Mechanics*. London: Spon and Chapman & Hall.

Fetter CW (2001) *Applied Hydrogeology*. New Jersey: Prentice Hall.
Goodman RE (1989) *Introduction to Rock Mechanics*. New York: John Wiley & Sons.
Goodman RE (1993) *Engineering Geology*. New York: John Wiley and Sons.
Hoek E and Bray J (1981) *Rock Slope Engineering*. London: Spon Press.
Hoek E and Brown ET (1981) *Underground Excavati ons in Rock*. London: Institution of Mining and Metallurgy.
Hoek E, Kaiser PK, and Banden WF (1995) *Support of Underground Excavations in Hard Rock*. Rotterdam: Balkema.
Moseley MP (1992) *Ground Improvement*. London: Spon Press.
Muir Wood D (2004) *Geotechnical Modelling*. London: Spon Press.
Potts DM and Zdravkovic L (2001) *Finite Element Analysis in Geotechnical Engineering: Theory and Application*. London: Thomas Telford.
Smith GN (1998) *Elements of Soil Mechanics*. London: Blackwell.
Smith IM (1994) *Numerical Methods in Geotechnical Engineering*. London: Thomas Telford.
Tomlinson MJ (2001) *Foundation Design and Construc tion*. London: Prentice Hall.
Wyllie DC (1992) *Foundations on Rock*. London: Spon.

GEYSERS AND HOT SPRINGS

G J H McCall, Cirencester, Gloucester, UK

Introduction

Geysers and hot springs that emerge at openings on Earth's surface are primarily found in regions of senescent or dormant volcanic activity. A second type of activity unrelated to volcanism also produces thermal waters; these 'meteoric' waters, the product of rain and snowfall, after descending deep into the ground through rock fissures and pores, have elevated temperatures at depth due to the global geothermal gradient. Both types of geothermal waters, convective and conductive, respectively, serve as geothermal energy resources and as resources for other purposes, but convective waters are more commonly exploited at present.

A unique ecology exists at geothermal sites. Under the sea, black smokers that emanate from geothermal vents relate closely to valuable sulphide mineral deposits found in ancient rocks; these sites also harbour extraordinary tube worm populations. On land, various minerals are deposited at sites of geothermal activity, micro-organisms thrive in the surrounding heated environment, and the warm waters have long been used as health spas. In addition to their popular use for health and medicinal purposes, geysers and hot springs serve as indicators of below-surface events. Increased geothermal activity provides a valuable warning of renewed volcanic activity. The entire spectrum of mild eruption as a volcanic process has an extension into the energy field, into the field of geomedicine, and into commercial enterprise. Geothermal energy is used in a number of countries worldwide, and some of the geothermal mineral deposits (for example, sulphur and borates) and associated rock formations (travertine and tufa) are of commercial value.

Geothermal Systems

The rocks of the uppermost 2–4 km of Earth's crust are generally porous and/or fissured and may be aquifers, filled with groundwater that has percolated down from rain and snow falling onto the surface. Those rocks that are not porous or fissured, i.e., are impermeable, form impermeable aquacludes, which separate or cap groundwater bodies. Heat within Earth is of two types: heat produced by gravitation on accretion and radiogenic heat from radioactive mineral decay. The transport of heat to Earth's surface is manifested in two ways. The first of these is a convective process, associated with the rise of hot magma, generated by partial melting in the mantle or crust. The molten rock cools in magma chambers, and this is the main source of geothermal heat as manifested in geysers and hot springs. The heat is passed from the cooling magma in the chamber to the surrounding country rocks, especially those above the chambers. Where these chambers contain groundwater bodies, the water expands and rises buoyantly, to be replaced by cold water flowing in from the sides, which is in turn heated and rises. This process establishes a geothermal circulatory system, which cools the

magma. Heat is initially brought up by the rising magma and then by the geothermal system as water or steam (the latter, if the water boils, which is a factor of temperature or pressure). Volcanoes, especially those with caldera development, are typically underlain by shallow magma chambers, which are fed from other larger reservoirs deeper in the crust or mantle. Thus geothermal systems are characteristic of regions with active, senescent, or dormant volcanoes. Most active volcanoes have been in action for less than a million years, so geothermal systems are geologically very young systems (*see* **Volcanoes**).

The second type of geothermal system is not volcanic related, but is related to the conductive transfer of heat outward through the crust. Because there is a temperature gradient in the crust, increasing with depth from the surface towards the mantle and core, groundwater taken in by recharge from rain and snowfall can be heated simply by being carried down to depths of a few hundred metres. These heated waters may then return to the surface as hot springs or geysers, or else remain as contained, heated groundwater or steam bodies in the subsurface. Heated water bodies below the surface are essentially unstable if they are overlain by cold water bodies, and will tend to move upward if not confined. Some hot water bodies may be in artesian conditions and flow out naturally in wells and bores. Natural outflows can be used for spa purposes. Hot water bodies in the subsurface that are confined by aquacludes can be tapped by boreholes and utilized for energy purposes and space heating.

By far, a majority of geothermal occurrences are related to volcanic activity and thus occur in volcanic provinces, but some (for example, in Bath, England) have no such obvious relationship. The entire Paris Basin, a structure analogous to the Cretaceous chalk-underlain London Basin, is underlain by heated groundwater bodies that can be exploited to a limited extent for residential heating and other purposes.

Mild Eruption (Volcanic-Related Geothermal Processes)

Hydrothermal

Hydrothermal relationships on Earth have been treated concisely by Alwyn Scarth in the book *Savage Earth*. On land, mild eruptions are limited to emissions of gas (mainly carbon dioxide), steam, and hot water and to formation of sulphurous fields, bubbling mud pools, hissing holes, and fissures. Such developments are associated with dormant or dying volcanoes that have been much more violent in the past and may become so in the future. Where such manifestations are present, it is impossible to say that the volcano is extinct. Lassen Peak (California, USA), Tiede volcano (Tenerife), and Taftan Peak (Southern Iran) are examples of dormant volcanoes. In the Gregory Rift Valley of Kenya, throughout several hundred kilometres, no volcano is presently active, but geothermal indications are ubiquitous. There is a geyser accompanied by hot pools at Lake Bogoria and a hot-water waterfall south of the lake, at Kapedo (Figure 1). In volcanic provinces, to augment seismic records, ground movement measurements, and other monitoring systems, geothermal hot springs and geysers should be monitored, because any increase in activity provides a valuable indicator of likely renewed volcanic eruptivity.

Figure 1 Hot water waterfall, Kapedo, Kenya. There are a number of hot springs immediately above the fall, the headwater of the Suguta River, and the water is believed to travel underground from Lake Baringo, ~50 km to the south. Photograph by GJH McCall.

Geothermal waters can be analysed to determine their origin. Isotopic ratios can differentiate between magmatic water ('juvenile' water, or water that has been brought to Earth's surface by upward movement of rocks) and meteoric water (groundwater); geothermal waters are almost completely meteoric. Meteoric water is water from rain and melting snow that has percolated downward. In many volcanic areas, hot magma still lies close to the surface, long after lava and ash eruptions have ceased, and the rocks above such residual magma chambers remain hot. As groundwater percolates down, heated rocks increase the groundwater to temperatures of more than 200°C, without boiling, because of the pressure. In the simplest case, this water returns to the surface to form hot water springs and pools, hissing and bubbling up, or it is tapped in wells and boreholes. Such developments at Rotorua, New Zealand, are well known tourist and health spa attractions.

Solfataras, Fumaroles, and Mofettes

Vent openings in the ground allow escape of heated waters and vapours. Escaping steam may mix with sulphurous gases and will deposit sulphur when its emerges at the surface. Where sulphur dioxide and hydrogen sulphide are emitted, the hissing escape channels are ringed at the surface with yellow sulphur deposits. Vent openings where such emissions occur are termed 'solfataras' (after La Solfatara volcano, Pozzuoli, Italy) and 'fumaroles'. Vents that are sources of toxic gases are termed 'mofettes'. In addition to carbon dioxide and sulphurous gases, inert gases such as nitrogen and argon can also be emitted. These gases may be almost undetectable and are in such cases dangerous. The principal gas emitted from most vents, carbon dioxide, can be lethal in high concentrations. It was the cause of a disaster at Lake Nyos, Cameroon, in 1986. Lake Nyos, the mouth of an ancient volcano, released a lethal cloud of CO_2 that asphyxiated more than 1700 people in nearby villages.

Geysers

Where hot water cannot circulate freely below the surface, it will gush out intermittently as pressure builds up. The water may appear muddy if it is mixed with material derived from the buried rock mass in which it is contained. In the extreme case, the phenomenon of the geyser (named after Geysir, a locality in Iceland) occurs, and water and steam spurt from the surface intermittently in a high-reaching fountain. 'Old Faithful' at Yellowstone, Wyoming, USA (Figure 2) reaches a height of 20 m, erupts every 60 minutes, and the eruption lasts for 5 minutes. It erupts so regularly that timetables have been created

Figure 2 'Old Faithful' geyser, Yellowstone National Park, Wyoming, USA. Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, Berlin: Springer Verlag.

announcing the schedule of eruption. The Nordic version of Old Faithful is Strokkur, in Iceland (Figure 3). Geysers have individual schedules: a geyser at the south end of Lake Bogoria (formerly Hannington), Kenya, erupts about every 10 minutes, whereas Beehive geyser, in Yellowstone, erupts only once a year, and a geyser at Rotorua, New Zealand, erupts four times a week. Changes in rainfall patterns can affect geysers; the Lake Bogoria geyser can temporally cease activity if the lake level gets too high, and the Great Geyser in Iceland, though active for 8000 years, now has to be stimulated by an injection of soap powder.

Figure 3 Strokkur geyser, Iceland. Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, Berlin: Springer Verlag.

Sinter and Travertine Terraces

Hot waters dissolve various chemicals as they traverse underground rock channels. When hot waters exit at surface openings, the dissolved chemicals are precipitated and form cowls or cones around the exit fissure or aperture. If the waters stream downslope, they may form spectacular terraces. The pink terraces at Rotorua, New Zealand, were long famous but were destroyed by earthquake action, and the best example now extant is the magnificent organ-pipe terraces at Mammoth Springs, Yellowstone, Wyoming (Figure 4). Such terraces are commonly composed of calcium carbonate and are referred to as tufa, where they are carious, or as travertine, where they occur in massive layers. Where such deposits are siliceous, they are sometimes referred to incorrectly by these terms, and are best referred to as 'siliceous sinter' (calcareous deposits are also sometimes referred to as 'calcareous sinter').

Figure 4 Mammoth Hot Springs terraces, Yellowstone National Park, Wyoming, USA. Reproduced from Science Photo Library.

Bioherms and Stromatolites

The mounds around hot springs may accommodate living creatures, forming bioherms. Stromatolites are layered, built-up structures found near hot springs; they are composed of populations of unicellular cyanobacteria and algae, organisms that have been dated to as far back as 3500 million years ago (*see* **Biosediments and Biofilms, Minerals:** Carbonates). At Lake Bogoria, Kenya, there are intermittent developments of stromatolites in the highly saline shallows of the lake close to the cluster of hot springs, pools, and geyser. Such stromatolites and algal fossils are commonly preserved in travertine that is used for ornamental building stone or cladding.

Exploitation of Volcanic-Related Geothermal Heat

In Reykjavík, Iceland, the volcanic-sourced heat is used to meet energy needs for general urban purposes. This local utilization is similar to that applied to non-volcanic-related geothermal heat in the Paris Basin (see later). The derived energy is

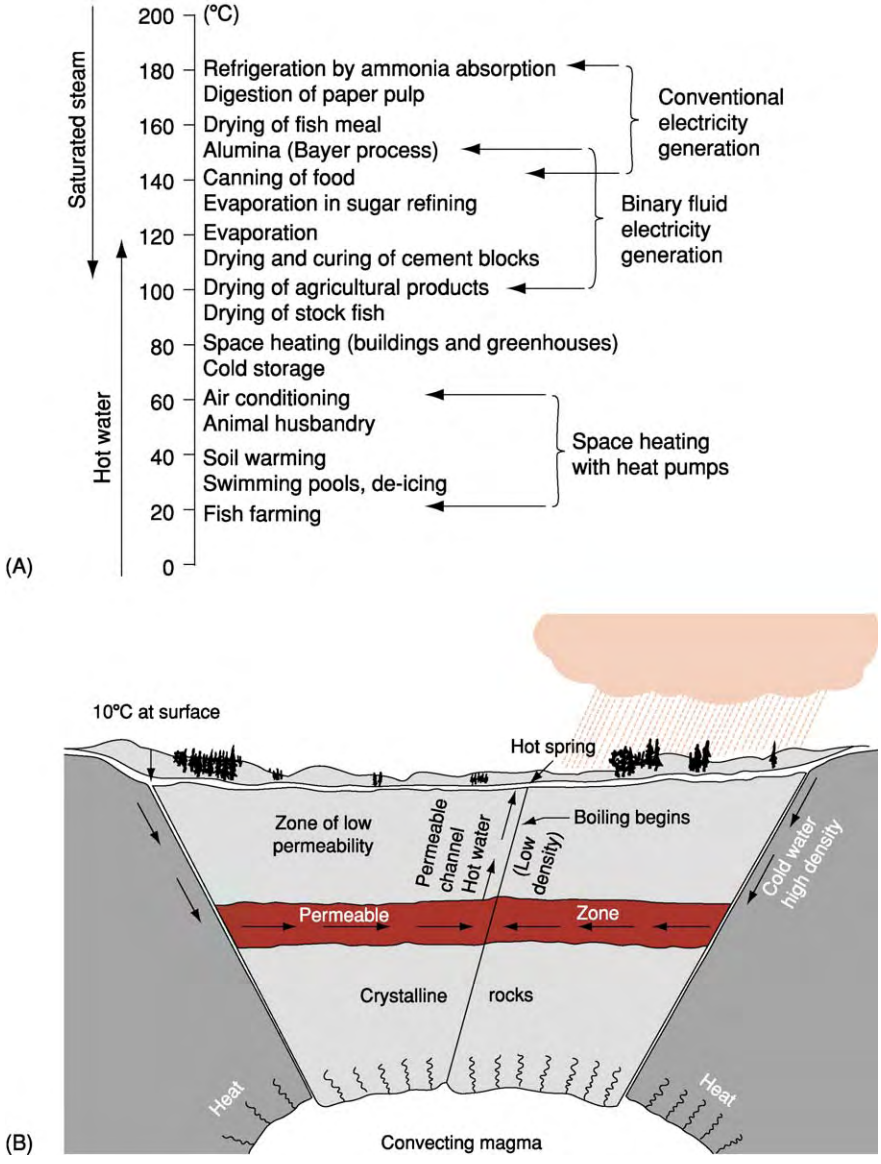


Figure 5 (A) Temperature requirements of geothermal fluids for various uses. Reproduced with permission from Arnorsson S (2000) Geothermal energy. In: Hancock PL and Skinner PJ (eds.) *Oxford Companion to the Earth*, pp. 437 440. New York: Oxford University Press. (B) Diagram showing a vapour dominated geothermal field. Reproduced with permission from Cargo N and Mallory BF (1977) *Man and His Geologic Environment*. Reading, MA: Addison Wesley.

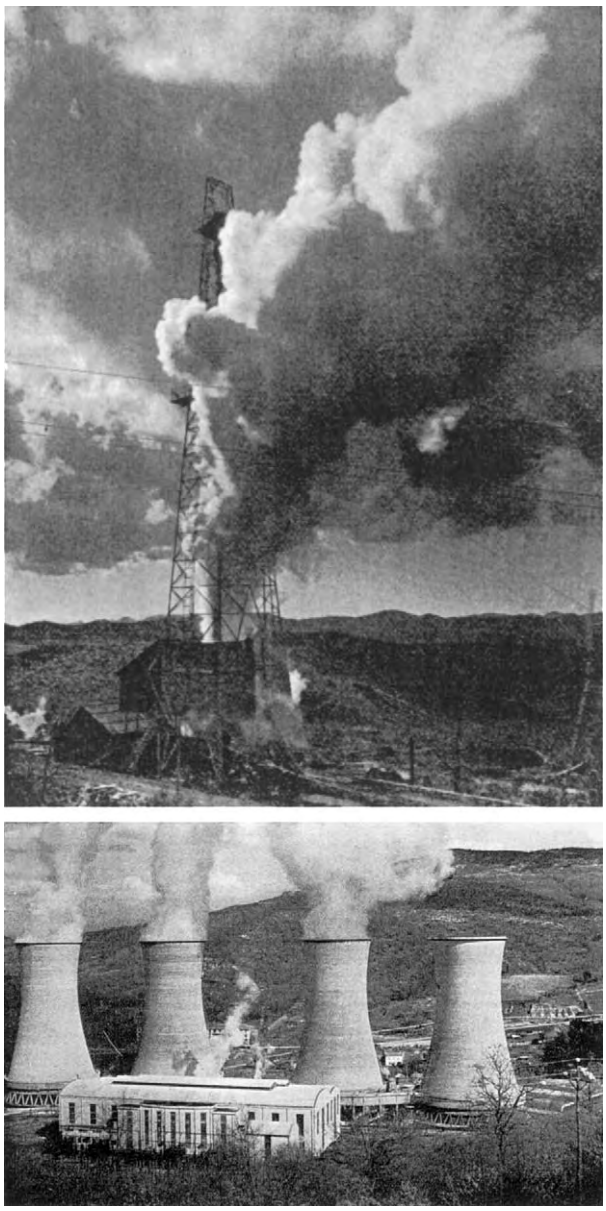


Figure 6 (A) 'Soffioni' blasting out at Larderello. (B) Power station installation at Larderello. Reproduced from Mazzoni A (1954) *The Steam Vents of Tuscany and the Larderello Plant*. Bologna: Arti Grafiche Calderini.

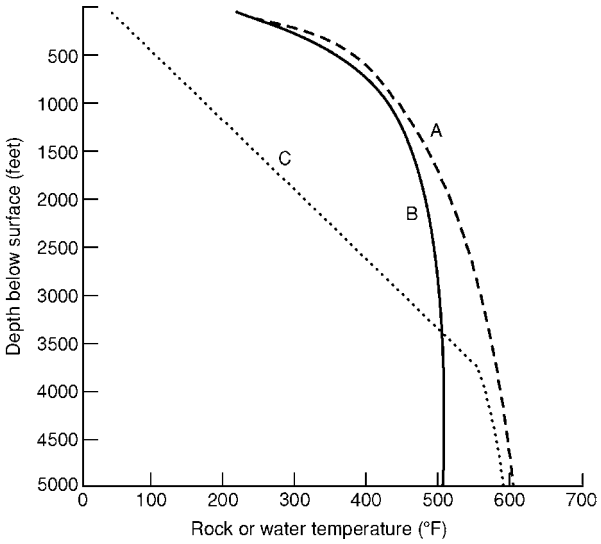


Figure 7 Water temperature variations with depth. Curve A: normal increase in boiling point. Curve B: increase in boiling point in hot spring region; magma or hot rocks near surface. Curve C: increase in boiling point in region where impermeable rocks prevent upward leakage; magma or hot rocks below imper meable rocks. Reproduced with permission from Cargo N and Mallory BF (1977) *Man and His Geologic Environment*. Reading, MA: Addison Wesley.

suitable for space heating of homes and greenhouses and for use in local industry. The minimum temperatures of geothermal waters for various utilizations are given in **Figure 5A**.

Major exploitation of geothermal energy has been so far restricted to volcanic-related occurrences. For such utilization, a reservoir temperature of more than 200°C is desirable. The Larderello field, south of Pisa, Italy, initiated in 1906, is a 'dry' or 'vapour-dominated' steam field in that the deep resource is under high temperature but low pressure, not much more than atmospheric (**Figure 5B**). The water boils underground and the generated steam that is trapped in pores within the underground rock can be tapped by boreholes. A 'soffioni' (vent) at Larderello can emit a burst with immense power in a spectacular fashion; the steam reservoir there is capped by impermeable rock such as schist, and boring is hazardous in that equipment and personnel can be injured when the borehole reaches a steam pocket (**Figure 6A**). The steam is superheated and may be corrosive, especially related to boric acid, which was formerly extracted from Larderello for commercial use. The energy output of the several power stations at Larderello (**Figure 6B**) is very large, and most of the railways in northern Italy were at one time reported to utilize it. The commercial geothermal plant at the Geysers, north of San Francisco, California, is similar to the Larderello plant; it is the only steam-dominated geothermal field in the United States, and the output has now gone well beyond the original 12 500-kW capacity.

The 'wet', or liquid-dominated, geothermal plants tap reservoirs under high pressure, where the temperature may reach 350°–700°F without boiling (**Figure 7**). Under these pressures and temperatures, water flashes to steam up the borehole.

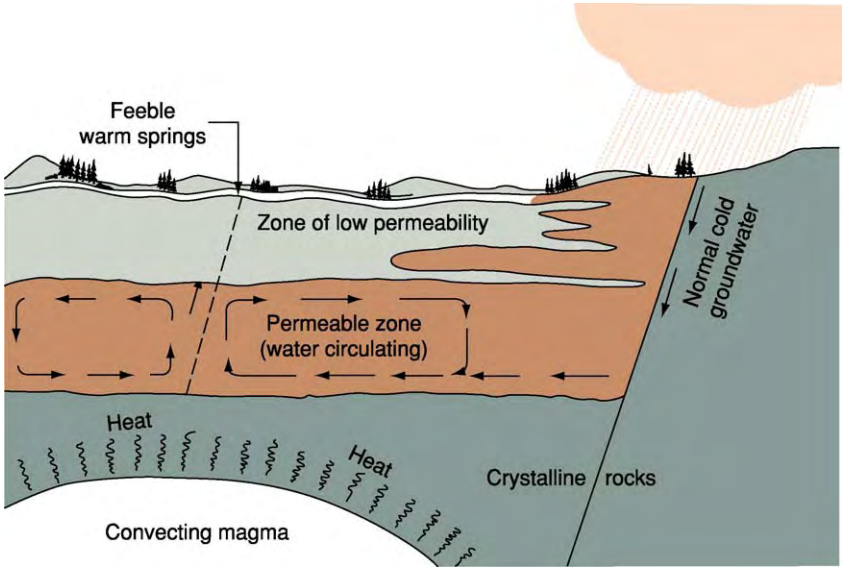


Figure 8 Diagram showing a liquid dominated geothermal field. Reproduced with permission from Cargo N and Mallory BF (1977) *Man and His Geologic Environment*. Reading, MA: Addison Wesley.

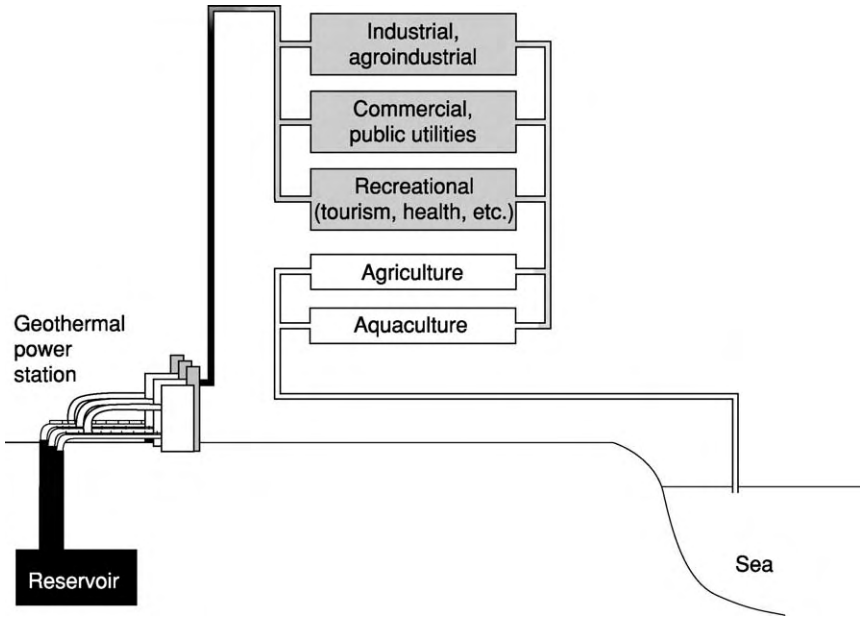


Figure 9 Diagram showing cascading relationships adopted at Reykjavík, Iceland. Reproduced with permission from Albu M, Banks D, and Nash H (1997) *Mineral and Thermal Groundwater Resources*. London: Chapman and Hall.

The geothermal installation at Wairakei, New Zealand is of this type (Figure 8). Other such installations are situated at Tongonan, Philippines, and Cerro Prieto, Mexico.

The most efficient use of volcanic-related thermal water is in Iceland, where 'cascading' applications have been developed, i.e., each step in the process uses a successively lower temperature (Figure 9). Chemicals and other raw materials may be extracted during the sequence of operations. It is sobering to realize how small a contribution geothermal energy makes towards meeting worldwide energy requirements. Table 1 lists

Table 1 Use of geothermal energy worldwide[a]

	Electric production		*Direct use*	
Country	*Power (MW)*	*Annual use (GWh)*	*Power (MW)*	*Annual use (GWh)*
Algeria			1	5
Argentina	0.7	3.5		
Australia	0.4	0.8		
Austria			21.1	84
Belgium			3.9	19
Bosnia Herzegovina			33	230
Bulgaria			95	346
Canada			3	13
China	32	175	1 914	4 717
Costa Rica	120	447		
Croatia			11	50
Czech Republic			2	15.4
Denmark			3.2	15
El Salvador	105	486		
France	4	24	309	1 359
Georgia			245	2 145
Germany			307	806
Greece			22.6	37.3
Guatemala	5			
Hungary			750	3 286
Iceland	140	375	1 443	5 878
Indonesia	590	4 385		
Ireland			0.7	1
Israel			42	332
Italy	768	3762	314	1 026
Japan	530	3530	1 159	7 500
Kenya	45	390		
Macedonia			75	151
Mexico	743	5 682	28	74
New Zealand	345	2900	264	1 837
Nicaragua	70	250		
Philippines	1848	8000		
Poland			44	144
Portugal	11	52	0.8	6.5
Romania	2	?	137	528
Russia	11	25	210	673
Serbia			86	670
Slovakia			75	375
Slovenia			34	217
Sweden			47	351
Switzerland			190	420
Thailand	0.3	2	2	8
Tunisia			70	350
Turkey	20	71	160	1 232
Ukraine			12	92
United States	2850	14 600	1 905	3 971
Europe	936	4309	4 368	20 505
America	3883	21 529	1 908	3 984
Asia	3031	16 092	3 075	12 225
Oceania	345	2901	264	1 837
Africa	45	390	71	355
Total	8240	45 220	9 686	38 906

[a]Reproduced from Arnorsson S (2000) Geothermal energy. In: Hancock PL and Skinner PJ (eds.) *Oxford Companion to the Earth*, pp. 437 440. New York: Oxford University Press. Data for megawatts (MW) and gigawatt hours (GWh) are valid for 1998.

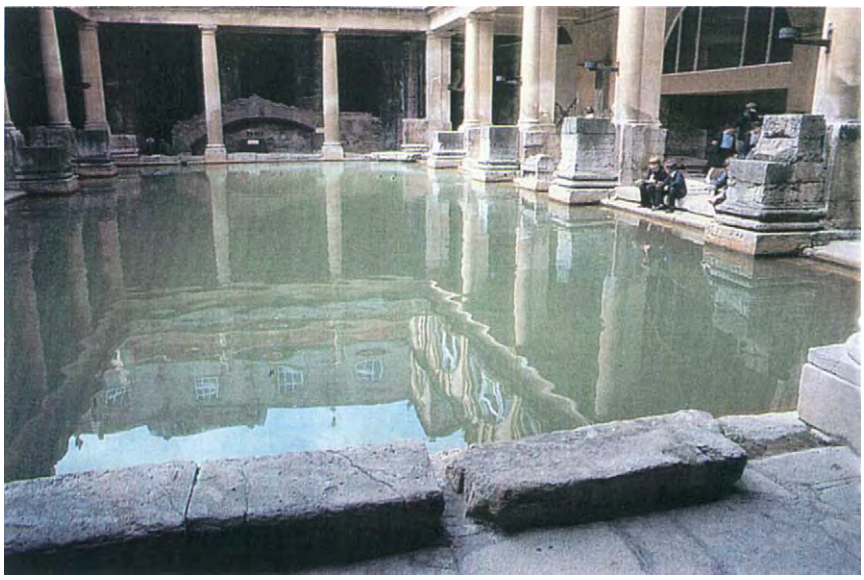


Figure 10 The Roman bath at Bath, England, fed by hot water from King's Spring. Reproduced from Kellaway GA (1991) *Hot Springs of Bath: Investigation of Thermal Waters of the Avon Valley*. Bath, UK: Bath City Council.

Table 2 Total heat output from hot springs at Bath, England, November 1990

Source	*Water output (m^3 day^{-1})*	*Temperature excess (above 10°C)*	*Heat output (Mcal day^{-1})*
King's Spring	1180	35°	41 300
Hetling Spring	75	37°	2775
Cross Bath Spring	35	31°	1085
Total	1290		45 160[a]

[a]Adding 20% for losses gives a total heat output of 54 192 Mcal day^{-1}.

Table 3 Basic geochemistry of geothermal waters of Bath, England

Variable[a]	*King's Spring and Stall Street borehole*	*Hetling Spring*	*Cross Bath Spring*	*Kingsmead borehole*	*Sports Centre borehole*	*Hot Bath Street borehole*	*Weston borehole*	*Batheaston coal shaft*
Temperature	44°C	47°C	41°C	35°C	20°C	41°C	13°C	17°C
Na	187	195	183	300	550	200	420	476
Ca	390	358	380	360	296	367	110	261
SO_4	1010	1015	1050	1060	1000	1020	795	940
Cl	286	340	288	440	770	350	235	964
HCO_3	199	193	189	154	198		245	81
Mg	53	57	54	56	71		35	65
Si	21	21		17	7	13		10
Fe	1	0.5	0.2	1		2	0.2	1.5

[a]Mineral content is given in milligrams per liter.

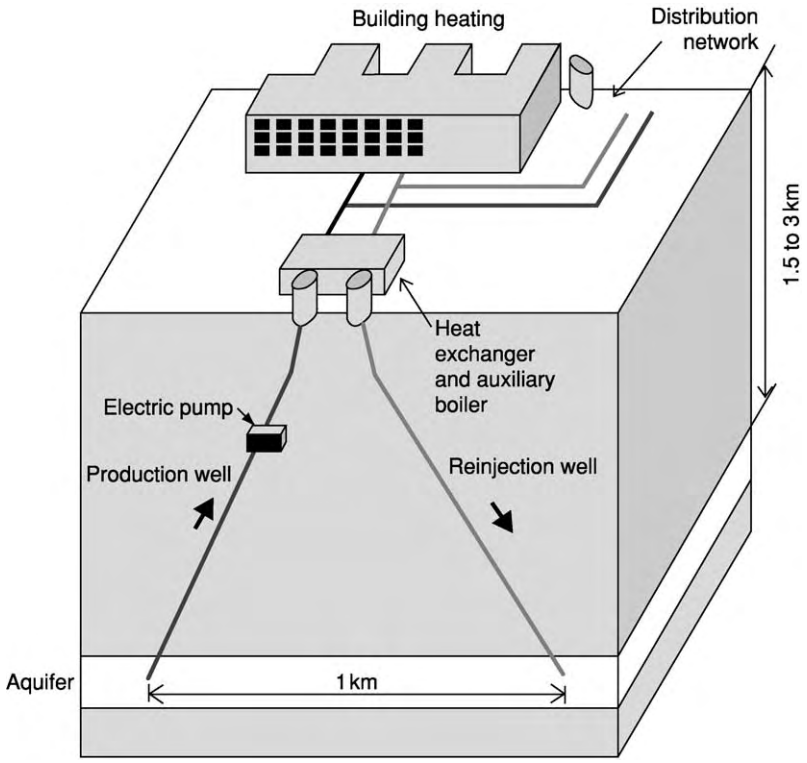


Figure 11 Doublet system with up and down circulation as used in the Paris Basin. Reproduced with permission from Albu M, Banks D, and Nash H (1997) *Mineral and Thermal Groundwater Resources.* London: Chapman and Hall.

the countries that utilize geothermal resources, and the sum total of electrical power produced is less than 10% of the electricity consumption of Great Britain.

Non-Volcanic-Related Geothermal Processes

Geothermal processes unrelated to volcanic activity are quite widespread. Because Earth's interior has a geothermal gradient, increasing in temperature with depth, when precipitation in the form of rainfall and snowfall descends deep enough through subsurface pores and passages, it is heated. Subsurface heated waters eventually rise to the surface under pressure, naturally as hot springs or artificially when tapped by boreholes. The first case is exemplified by hot springs in Bath and Bristol, England; the Paris Basin in France is an example of a region where 'heat mining' activities have been developed.

Bath and Bristol Hot Springs

The hot springs in western England (Figure 10) were utilized for spa purposes by the Romans, and possibly earlier, by the Celts; after the end of Roman rule, the baths fell into disrepair, but were rejuvenated during the Elizabethan Era and were an attraction in the Bath region from the seventeenth century onwards. The heated waters in the natural springs of Bristol and Bath have no known volcanic association and are widely believed to relate to rainfall recharge in the Mendip Hills, about 40 km away. In this process, rainwater is routed downward through fissures to the Carboniferous limestone at 600–900 m below sea-level, and beyond that, to the Lower Palaeozoic rocks. There is up to 4600 m of section above the base of the Ordovician, and calculations show that the water could be heated in depth to 121°C. The water returning to the surface under Bath and Bristol (Hotwells), through a series of fissures and aquifers, reaches the surface (artesian conditions are locally operative) still hot, though having lost much of its heat. The passage to Bath from the Mendip recharge area is estimated to take 3000 years. Flow measurement, temperature, and heat output and the basic chemistry of the Bath waters are given in Tables 2 and 3. Radiogenic elements and dissolved

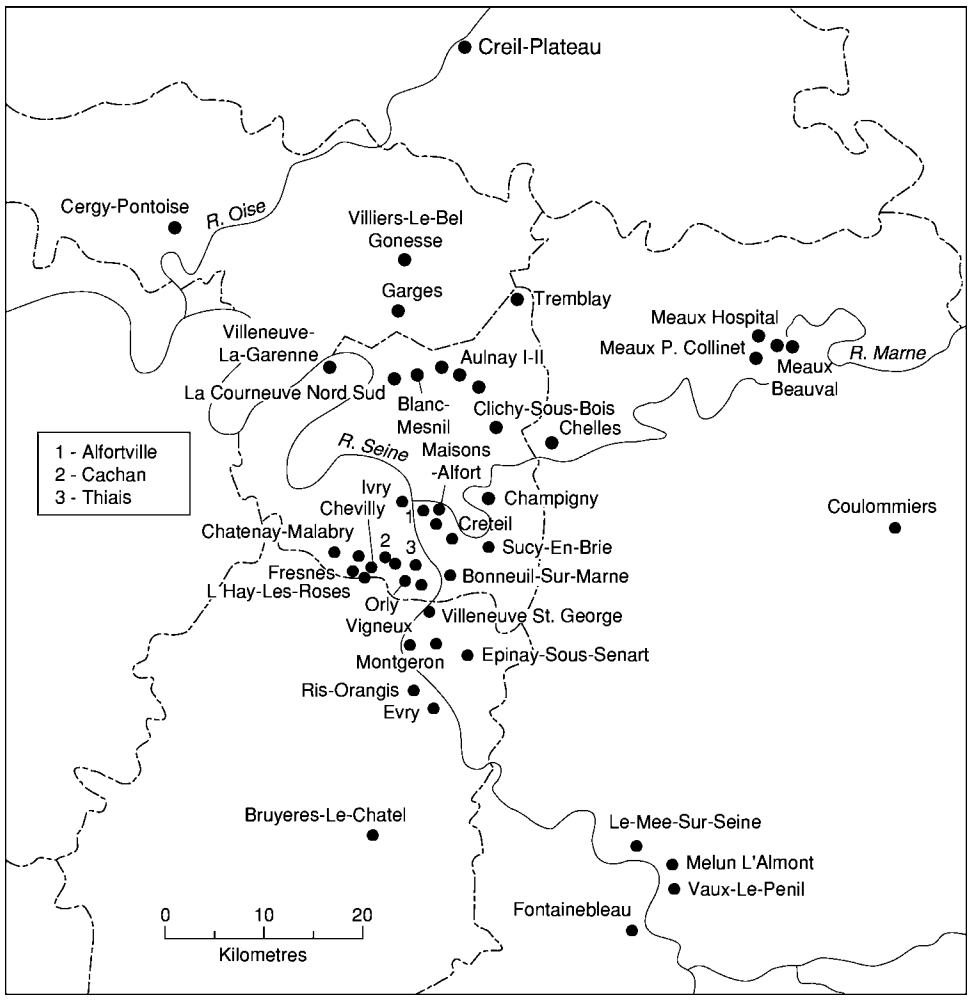


Figure 12 Geothermal utilization in the Paris Basin. Reproduced from McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam and Brookfield, VT: Balkema.

noble gases in the waters are believed to come from Lower Palaeozoic rocks, being collected by the water on its passage through the substrata.

The Paris Basin

Rainwater in the urban outskirts of Paris, France, is carried down to deep aquifers in the same way as occurs at Bath and Bristol, but instead of using the heated waters for spa purposes, the practice in the Paris Basin is to use the heat mainly to service large estates of apartments. Borehole pairs form up and down doublets (Figure 11), and the extracted geothermal water is utilized as a direct heat source. The waters have temperatures of 60–70°C. As shown in Figure 12, the French have developed this type of low-level geothermal utilization widely. A cascading process is utilized here as in Iceland. Geothermal heat energy is a very local resource when used in this way. Thermal resources further south in the Auvergne are volcanic related, but the Paris Basin occurrences are not.

Seafloor Geothermal Activity

Hot springs emanate in volcanic parts of the ocean floor. Volcanism occurs in mid-ocean ridges, island arcs, and ocean islands (hotspots), but the mid-ocean ridges have become famous for their black smokers, regions where clouds of black mineralized smoke accompany thermal water emissions and strange, blind tube worms cluster feed on the emissions (*see* **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges). These occurrences are closely related to the formation of ancient volcano-related sulphide mineral deposits. In the Red Sea, such submarine

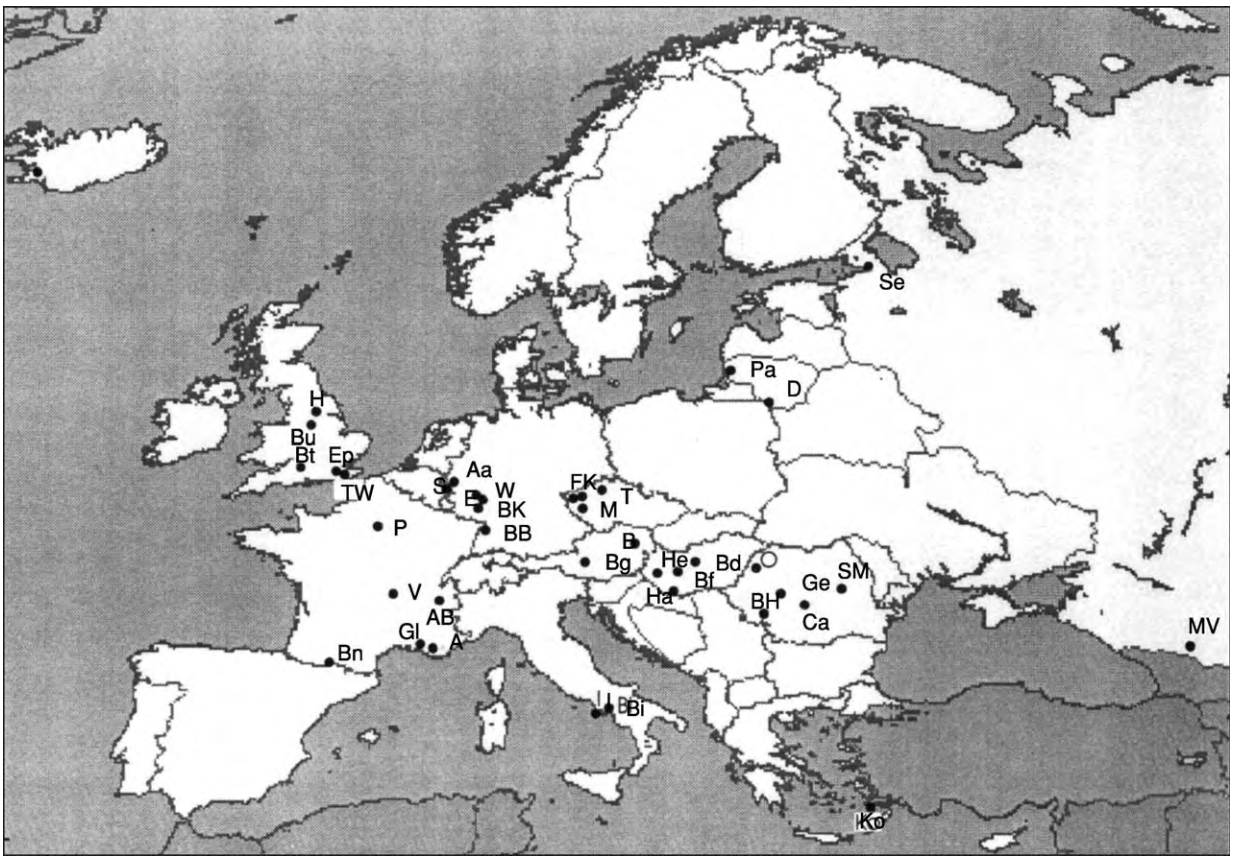


Figure 13 Sites where spas were once located throughout Europe. Reproduced with permission from Albu M, Banks D, and Nash H (1997) *Mineral and Thermal Groundwater Resources*. London: Chapman and Hall.

thermal springs are associated with zinc mineralization. There is probably little potential for economic recovery of minerals from such hot springs, though the hydrothermal deposits of gold in the Western Pacific might prove of economic significance.

Commercial Applications

There has been much research in Cornwall, England into utilizing the deeply buried hot rocks as an energy source, but this has so far not yielded any practical results. Health spas, however, are of economic interest. Spas were once numerous throughout Europe (Figure 13), and most utilized thermal waters. The Bath spa dates from Roman times, and the Roman thermal bath there (the King's Spring) is still in existence (Figure 10). Good accounts of it were written in the seventeenth century by Celia Fiennes, John Evelyn, and Samuel Pepys, and a recent account edited by GA Kellaway examines geophysical evidence for the origin of the hot springs. There were five baths originally, according to Celia Fiennes: Hot Bath, Lepours, Cross Bath, Kitchings, and King's Bath, but not all have survived. In 1978, after bathing in the waters at Bath, a child died from meningitic infection by *Naegleria fowleri*, a free-living amoeba found in hot spring water, and the spa was closed. This danger has been eliminated following a major investigation and restoration; the basis of success was correct assessment of the geological structure of the springs and understanding the ecological requirements of *Naegleria*. *Naegleria* requires oxidized water for sustained growth, thus the baths must be replenished by drilling for amoeba-free water, preventing oxidation of the water before it reaches the surface, and isolation of this water from oxidized water. The Bath spa is scheduled to reopen in 2004 or 2005.

Travertine and tufa, widely used as building stones and cladding, are another important economic contribution of geothermal features. Travertine is particularly widely used for cladding McDonald's restaurants. The cream-coloured stone comes from Tivoli, Italy; it often contains cavities made by

Figure 14 Travertine cladding on a McDonald's Restaurant, Hereford, England. Photograph by GJH McCall.

dissolution of reeds, and these have to be filled artificially, after which the stone makes a striking cladding, showing sinuous algal banding (Figure 14).

See Also

Biosediments and Biofilms. **Earth:** Mantle; Crust. **Minerals:** Carbonates. **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Albu M, Banks D, and Nash H (1997) *Mineral and Thermal Groundwater Resources*. London: Chapman and Hall.

Arnorsson S (2000) Geothermal energy. In: Hancock PL and Skinner PJ (eds.) *Oxford Companion to the Earth*, pp. 437 440. New York: Oxford University Press.

Cargo N and Mallory BF (1977) *Man and His Geologic Environment*. Reading, MA: Addison Wesley.

Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, Berlin: Springer Verlag.

Kellaway GA (1991) *Hot Springs of Bath: Investigation of Thermal Waters of the Avon Valley*. Bath: Bath City Council.

Mazzoni A (1954) *The Steam Vents of Tuscany and the Larderello Plant*. Bologna: Arti Grafiche Calderini.

McCall GJH (1967) *Geology of the Nakuru Lake Hannington Thomsons Falls Area. Geological Survey of Kenya, Report No. 78*. Nairobi: Ministry of the Inter ior, Geological Survey of Kenya.

McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam and Brookfield, VT: Balkema.

Scarth A (1997) *Savage Earth*. London: Harper Collins.

GLACIERS

See **SEDIMENTARY PROCESSES: Glaciers**

GOLD

M A McKibben, University of California, CA, USA

Characteristics and Uses

Gold is a rare heavy metal that is soft, malleable, ductile, and bright sun yellow in colour when pure (Figure 1). The last property is reflected by its chemical symbol, Au, which comes from the Latin word *aurum* ('shining dawn'). Gold resists chemical attack and corrosion, has excellent electrical conductivity, and reflects infrared radiation.

Humans have made use of gold for more than 40 000 years. Because of its inertness and value, most of the gold that has ever been mined (about 130 000 metric tons) is still in use! Most gold is hoarded, in the form of bullion, coins, and jewellery. This usage stems from the metal's aesthetics, its role as a medium of exchange among banks and governments, and its perceived value by individuals and families as a hedge against economic uncertainty. Much like diamonds, gold has an emotional (and sometimes irrational) appeal that drives its free-market value to levels far above those that would otherwise be justified solely by its practical uses in technology.

Practical uses of gold take advantage of its chemical inertness, electrical conductivity, malleability, and ductility. Gold's high electrical conductivity and solderability make it excellent for creating reliable electrical contacts; it can be drawn into wires thinner than a human hair and pounded into sheets thin enough to pass light. It is therefore used in electronics (circuit boards, connectors, contacts, thermocouples, potentiometers), corrosion-resistant processing equipment (acid vats), infrared reflectors (windows of high-rise buildings, astronaut's helmet visors), and dental applications (fillings, crowns, etc.).

Pure gold is too soft, malleable, and ductile for uses that require physical endurance, particularly jewellery, so instead mixtures or alloys of gold with other metals (copper, nickel, silver, platinum, etc.) are used to enhance an object's durability. The gold content of such alloys can be defined in terms of carats, or parts of gold per 24 parts of total metal by weight. Durable jewellery is commonly made from 14 carat gold (containing 58.3% gold). Knowing the remaining metals used in gold alloys can be important to people who are allergic to specific metals in jewellery, such as nickel, particularly when the skin is pierced (earrings, nose rings, etc.).

The term 'gold-filled', which is used in jewellery making, means a layer of gold alloy that is placed over a less valuable core of base metal – in a few countries such layers are required by law to be at least 10 carats. 'Rolled gold plate' may be applied to layers that are less pure. Gold 'electroplate' jewellery must have at least seven millionths of an inch of gold overlaid, otherwise terms such as 'gold flashed' and 'gold washed' must be used.

Ironically, during the Spanish exploitation of the New World, comparatively 'worthless' platinum objects from South America were sometimes plated over with gold and passed off as pure. In modern times the relative values of these two metals are usually reversed, and the opposite strategy would be more lucrative. The consumer must be cautious at all times!

Another scale for expressing the purity of gold (mainly in bars, ingots, and coins) is fineness, based on parts of gold per 1000 parts of total metal. A gold bar that is '995 fine' thus contains 99.5% gold by weight.

The most common unit of weight for gold is the troy ounce, which is equal to 1.097 avoirdupois ounces or 31.10 g. A typical bar of gold (such as held by central banks and governments) weights about 400 troy ounces, or about 27.5 pounds (= ~12.4 kg). A troy pound contains 12 troy ounces, and a troy ounce contains 20 troy pennyweights, the latter being the unit of weight most often used in jewellery. A troy pennyweight (dwt) equals 1.555 g.

Mineralogy, Geochemistry, and Natural Concentration of Gold

Gold, silver, and copper are often associated in nature, because each has a lone outer S-orbital electron and belongs to the IB transition-metal subgroup of the periodic table. Other elements commonly associated with gold include arsenic, antimony, bismuth, iron, lead, and zinc.

In addition to the zero-valent elemental form, gold has two oxidation states: Au^{+} (aurous) and Au^{3+} (auric). Most gold minerals contain zero-valent or aurous gold (Table 1). About 40 natural gold minerals are known; native gold and electrum are the most common forms, tellurides are rare, and other forms are very rare. Because most gold ore occurs as

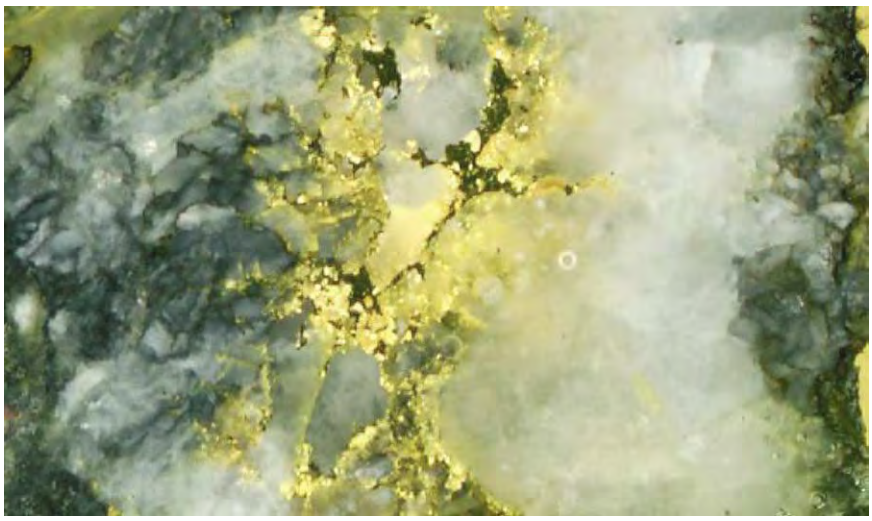


Figure 1 Polished slab of hydrothermal gold quartz vein ore from the Hollinger Mine, Porcupine District, Ontario, Canada. This occurrence is typical of orogenic lode (greenstone belt) gold. The width of the image is approximately 1 cm.

Table 1 Examples of gold minerals

Native gold	Au
Electrum	(Au,Ag)
Cuproauride	(Au,Cu)
Rhodite	(Au,Rh)
Amalgam	Au_2Hg_3
Calaverite	$AuTe_2$
Krennerite	$(Au,Ag)Te_2$
Sylvanite	$(Au,Ag)Te_4$
Petzite	Ag_3AuTe_2
Nagyagite	$Pb_5Au(Te,Sb)_4S_{5-8}$
Uytenbogaardtite	Ag_3AuS_2
Aurostibite	$AuSb_2$
Fischesserite	Ag_3AuSe_2

the relatively pure native form, smelting or roasting of gold ore is sometimes not required.

Average gold concentrations in upper-mantle and crustal rocks vary by about an order of magnitude, typically between 1 ppb and 10 ppb. Mafic rocks generally contain higher concentrations than felsic rocks, and some minerals such as magnetite and pyrite may act as natural 'concentrators' to enrich gold in rocks. Continental crust averages about 4 ppb gold, whereas seawater contains only about 0.01 ppb gold.

Gold can be mobilized and concentrated into economic deposits by many geological processes, including igneous melting and crystallization, hydrothermal leaching and precipitation, and fluvial hydraulic sorting. In silicate melts, gold tends to follow other chalcophile elements and is often concentrated in metallic sulphide and oxide phases. When magmas ascend to shallow levels in the crust, gold may also partition into chlorine- and sulphur-rich aqueous fluids, which are expelled into the surrounding wall-rock as the magma crystallizes. Vapour-phase transport of gold chlorides has been observed in some volcanic emissions. Metamorphism and consequent dewatering of mafic rocks in orogenic zones can also lead to gold remobilization and precipitation in veins, along with quartz and carbonate minerals. In shallow crustal settings where hot aqueous fluids circulate and sometimes boil within permeable volcanic and sedimentary rocks (hot-spring or epithermal systems), gold seems to be preferentially transported and precipitated with silica (quartz) in veins and pores.

Most major gold deposits have resulted from hydrothermal processes (Figure 2) or from subsequent erosion and hydraulic sorting of rock grains derived from hydrothermal deposits (*see* **Tectonics:** Hydrothermal Activity; **Mining Geology:** Hydrothermal Ores). Thus, the simplistic terms lode (vein) and placer (sediment) for the origin of gold dominate the rich history and legal framework of gold exploration and mining. Lodes form when hydrothermal fluids pass through fractured permeable rock and deposit gold and other minerals as tabular veins and adjacent pore fillings. Placers form where surface waters (mainly fluvial) experience a change in hydraulic conditions (especially velocity), forcing them to drop suspended sediment grains of a certain density, size, or shape. Tributary intersections, meander loops, and canyon mouths are examples of locations where placer gold can accumulate in rivers.

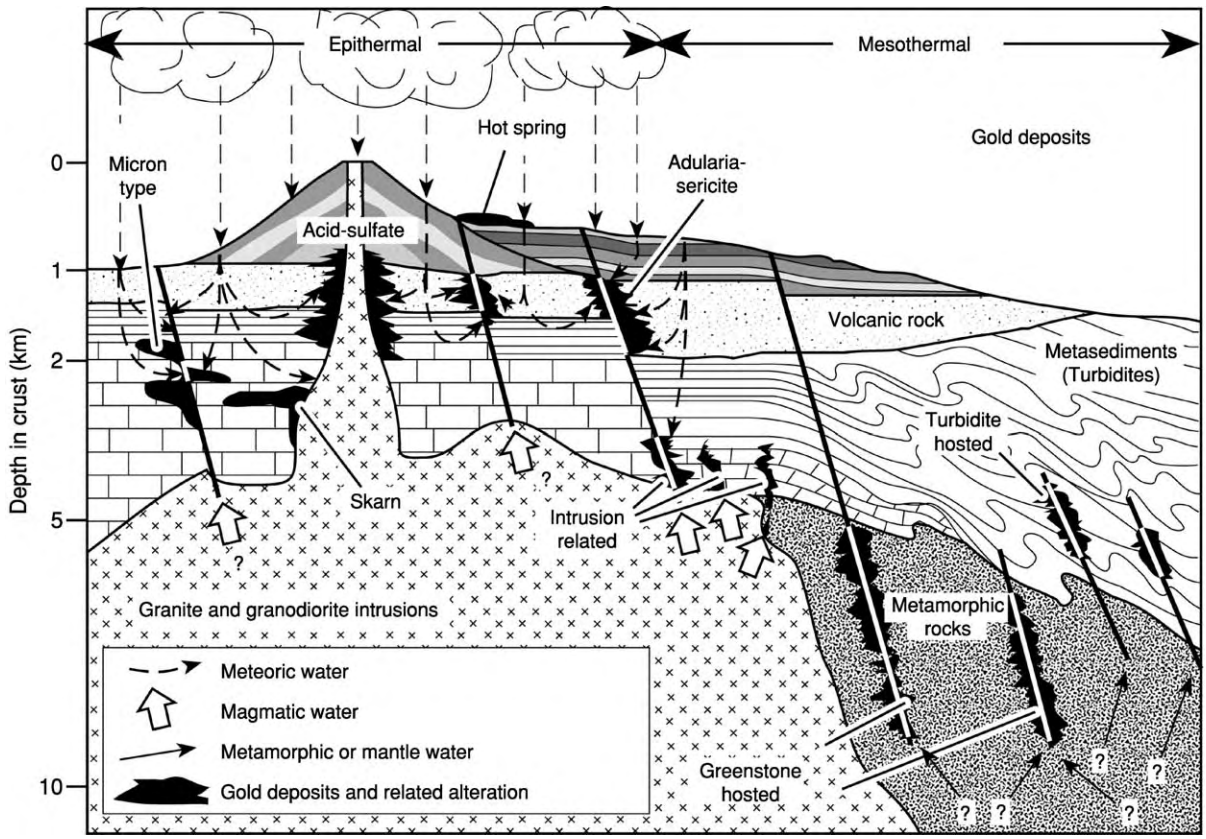


Figure 2 Schematic illustration of the wide variety of occurrences of hydrothermal gold deposits in continental crust. The terms epithermal and mesothermal refer to shallow crustal and mid crustal level deposits, respectively. Reproduced from Kesler SE (1994) *Mineral Resources, Economics and the Environment*. New York: MacMillan.

In most hydrothermal fluids, gold is transported in the form of aurous aqueous complexes with hydroxide (OH^-), chloride (Cl^-), and bisulphide (HS^-) anions. Therefore, gold may be precipitated from such fluids via geochemical processes that destabilize the aqueous complexes and reduce the aurous gold to the elemental state, for example:

$$Au(HS)_2^- + H^+ + 0.5H_{2(g)} = Au_{(S)} + 2H_2S.$$

Environmental changes that can destabilize such complexes include cooling, pH changes, and changes in fluid oxidation state that accompany boiling, mixing, and chemical reaction of migrating hydrothermal fluids with wall rock and other crustal waters. Many gold deposits are associated with minerals, such as silica (quartz), carbonate, sulphates, and clays, that form during such geochemical changes.

Common host minerals for gold include iron sulphides, oxides, and hydroxides, and such phases may thus play a role in the local reduction of aurous gold. For example, the oxidation of ferrous to ferric iron can facilitate the localized reduction of aurous gold from solution in the presence of ferrous mineral phases:

$$AuCl_2^- + Fe^{2+} = Au_{(s)} + Fe^{3+} + 2Cl^-$$

The occurrence of tiny grains of gold within larger mineral hosts (pyrite, oxides, hydroxides, etc.) and the presence of encapsulating and metallurgically recalcitrant phases such as silica can be significant factors influencing the degree to which mined gold ores must be ground, leached, and otherwise processed to release the gold. Failure to examine and characterize adequately the distribution and texture of fine gold occurrences within ores has led to some spectacular (and very expensive) mistakes in designing mineral processing facilities at some gold mines.

In the past few decades, the advent of inexpensive bulk-mining and cyanide heap leaching methods has made rocks containing as little as 1 ppm gold (0.3 troy ounces per ton) rich enough to be ores profitable for mining. Many of these rocks occur in the feebly mineralized (and previously ignored) areas surrounding now mined-out richer vein deposits.

Though likewise hydrothermal in origin, such bulk-disseminated low-grade gold ores do not fall conveniently into the traditional 'lode' category, leading to some interesting legal ambiguities in their status. Rigid governmental regulations for the dimensions of claim blocks, designed originally for idealized tabular 'lode' ore bodies, cannot be easily applied to more disseminated and irregular forms of mineralization.

Gold Ore Deposits

Being so valued a metal, gold has been actively sought out and mined in a wide variety of geological settings, wherever it has occurred. Only some of the major geological types of gold deposit will be briefly described here; some of the articles in the further reading section provide a more exhaustive catalogue of world gold-deposit types.

Archaean Gold–Quartz Conglomerates ('Palaeoplacers')

A majority of the world's gold has been mined from what were traditionally viewed as ancient placer deposits (Figure 3). While many people might envision a wizened prospector panning or sluicing for loose gold nuggets along a modern river bank, the most important of these ancient deposits are in hard lithified metamorphosed coarse sediments of Archaean age (2.3–2.8 Ga). The Witwatersrand ('Wits') gold fields of South Africa have produced more than one-third of the gold mined on Earth since 1886, and one single Wits mine (the Vaal Reefs mine) has produced almost as much gold as all of the gold mines in the USA and Australia combined. Similar ancient 'palaeoplacer' districts include the Elliot Lake–Blind River district in Canada. The gold in Wits ore is typically associated with detrital pyrite, uraninite, and carbon in quartz pebble conglomerates and quartzites that have been subjected to peak metamorphic temperatures of 300–400°C.

Placer and hydrothermal-replacement models for Wits gold have both been advanced, but most recent evidence favours a hydrothermal origin for the gold in these ancient sediments. The emplacement of the gold seems to have occurred at the time of peak metamorphism, via low-salinity fluids channelled along fault structures, unconformity surfaces, and bedding planes. Reaction with carbon- or iron-bearing (pyritic) rocks located just above the unconformity surfaces caused gold precipitation. The ultimate origin of the gold was probably underlying mafic rocks, rather than eroded lode quartz veins as traditionally implied by sediment clasts and the placer model.

In addition to their significance as major sources of gold, the Wits palaeoplacers provide strong evidence that the Earth's Archaean atmosphere was very different from our modern atmosphere. Detrital grains of pyrite and uraninite cannot survive significant fluvial transport in our oxygen-rich modern atmosphere; their abundance in the Wits placers indicates that the ancient atmosphere was less oxygenated and perhaps more

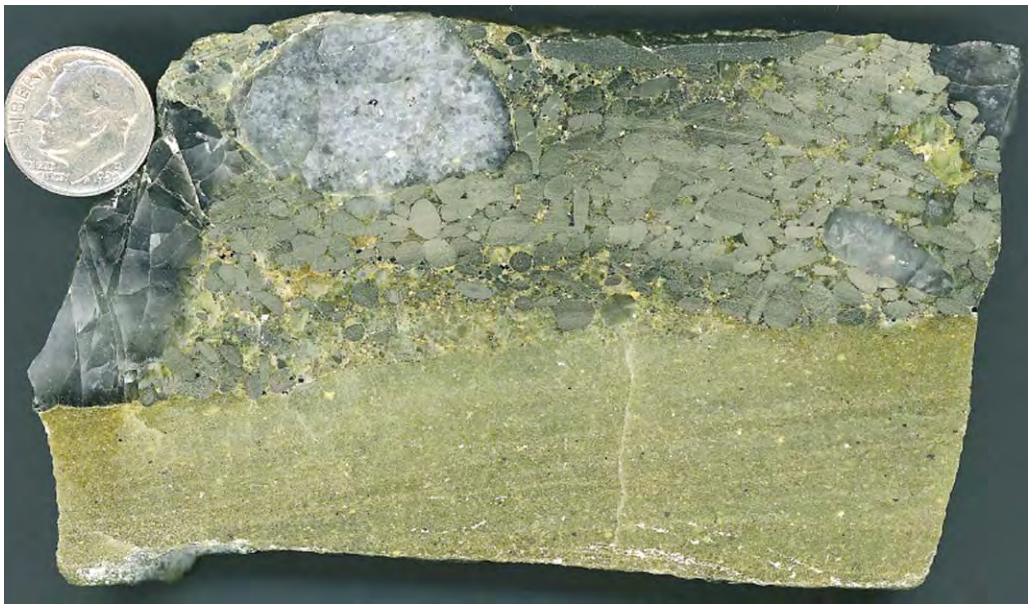


Figure 3 Polished slab of Archaean quartz pebble conglomerate gold ore from the Vaal Reefs Mine, Witwatersrand District, South Africa. Rounded detrital pyrite is clearly visible. Curved marks across the pyrite grains are saw blade artefacts. A US 10 cent coin is shown for scale.

enriched in carbon dioxide and/or methane than the modern atmosphere (*see* **Atmosphere Evolution**).

Orogenic Lode Gold

The Mother Lode District of California may be the most famous of a class of lode gold-ore deposits that are found in metamorphic terranes associated with accretionary plate tectonics. They have also been called greenstone-belt and mesothermal lode ores (Figure 1). The discovery in 1848 of placer gold at Sutter's Mill, downstream from this district, led to the California Gold Rush of 1849, one of the most significant New World social migration and settlement events in recent history. To this day many people still wear a popular miner's legacy from this gold-driven rush: Levi's blue jeans.

Orogenic lode gold deposits range in age from Archaean (such as the Barberton district in South Africa) to Mesozoic–Tertiary (such as the Mother Lode district). There remain a number of theories for their origins, but all seem to be associated with major accretionary orogenic events (plate collisions) and occur near major translithospheric structures (shear zones) bounding metamorphosed terranes. Typical conditions of formation are 250–400°C and 1–3 kbar.

The mineralization is typically of syn- to post-peak metamorphic age, associated largely with mid-crustal-level greenschist facies (gold plus quartz plus carbonate) mineral assemblages. The deposits are structurally controlled, often along splays associated with high-angle faults; the gold is syn-kinematic and localized near the brittle–ductile crustal transition. Fluid pressures range from superlithostatic to sublithostatic within the brittle–ductile shear zones. Within individual deposits, the quartz–carbonate vein systems are sometimes more than 2 km in vertical extent. The most common alteration minerals associated with the gold are quartz, carbonate, white mica, chlorite, and pyrite; both gold and silver are enriched in the deposits, while the abundance of base metals is relatively low. Fluid-inclusion data indicate that the aqueous fluids that formed the ores were of low salinity (salts) and rich in carbonic components (carbon dioxide, methane) (*see* **Fluid Inclusions**).

These hydrothermal or metasomatic systems thus seem to have developed during plate collisions as deep metamorphic fluids were expelled during shearing and deformation and transported into structurally favourable settings where they deposited the gold.

Epithermal Gold

As the name implies, epithermal gold deposits form in shallow (1–2 km depth) anomalously hot (100–300°C) crustal settings. Most develop around areas of volcanic activity, with distal to proximal magmatic heat flow driving the convective circulation of groundwaters in brittle fractured rock. Boiling hot springs and geysers near active volcanoes are modern examples of the surface manifestations of such systems. Some epithermal systems may grade with depth into magma-hydrothermal systems that are more characteristic of porphyry systems. The terms 'low-sulphidation state' and 'high-sulphidation state' are used to describe epithermal systems dominated by meteoric and magmatic fluids, respectively.

Epithermal deposits form most often in volcanic arc settings above subduction zones, where sustained high heat flow, active tectonism, and meteoric precipitation encourage fluid heating and circulation. The vein and disseminated gold ores in epithermal systems are typically associated with silica, alkali feldspar, and other secondary minerals (clays, micas, sulphates) formed by acid alteration of the host rock.

Examples of major epithermal gold deposits include Hishikari (Japan) (Figure 4), El Indio (Chile), Round Mountain (Nevada, USA), and Cripple Creek (Colorado, USA).

Carlin-Type Gold

The term 'Carlin-type' is applied to enigmatic mid-Tertiary gold deposits in western North America

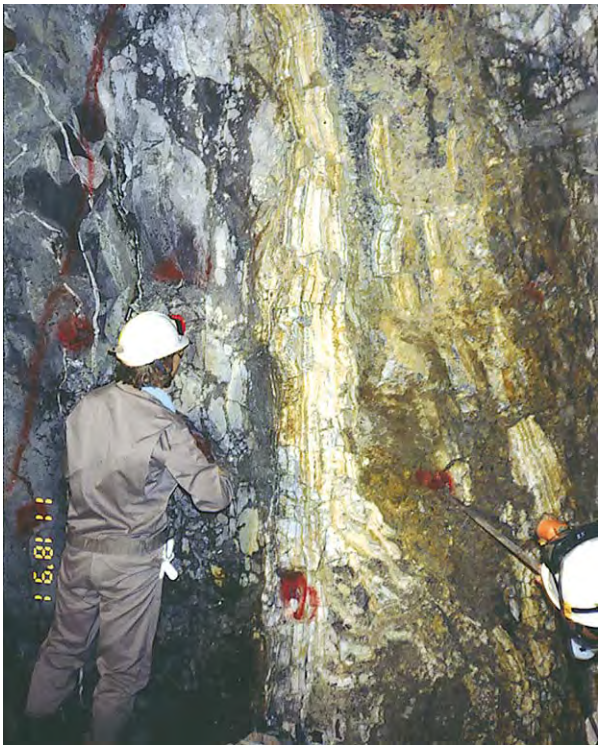


Figure 4 Banded crustiform epithermal veins of quartz in brecciated rock comprise the main ore type at the underground Hishikari Mine in Japan, one of the world's richest gold mines.

(epitomized by the Carlin district in Nevada) whose origin seems to fall somewhere between the epithermal and orogenic lode gold types. The epigenetic gold in these deposits is finely disseminated and hosted by Palaeozoic calcareous sedimentary rocks. Arsenical pyrite, marcasite, and arsenopyrite are the main host minerals. The host rocks exhibit evidence of carbonate dissolution, argillic alteration, silicification, and sulphidation. It has been proposed that reduced hydrogen sulphide-rich auriferous fluids migrated up pre-existing fault structures and spread laterally into permeable reactive carbonate rocks, depositing gold at temperatures of 150–250°C at depths of more than 2 km. The direct role (if any) of crustal magmatic processes in driving circulation and contributing components to the fluids is not clear. More recent models favour a metamorphic origin, driven by heat from a mantle plume underlying the region during crustal extension.

Though Carlin-type gold deposits are not as widely distributed in space and time as the other types of gold deposit described above, their more recent discovery was instrumental in driving the development of inexpensive cyanide heap leaching technology in the late twentieth century, making these and other lower-grade disseminated gold deposits economic and competitive with the traditionally dominant Witwatersrand gold deposits in South Africa.

Gold as a By-Product in Other Metal Deposits

Gold is frequently mined as a by-product from ore deposits dominated by base-metal production (Cu, Pb, Zn). Such deposits include porphyry, skarn, sediment-hosted, and volcanogenic massive sulphide types. In some cases, the addition of by-product gold makes the base-metal deposit economic. There are also deposits of these types that are mined only for gold.

Examples of major base-metal deposits with significant by-product gold include Grasberg (Indonesia) (porphyry and skarn copper), Bingham (Utah, USA) (porphyry copper), Red Dog (Alaska, USA) (sediment-hosted zinc), Rammelsberg (Germany) (sediment-hosted zinc and lead), and Mount Morgan (Queensland, Australia) (volcanogenic massive sulphide copper).

Gold Mining

Placer Mining

Mining of unconsolidated placer gold at the Earth's surface has taken place for centuries, ranging from the simple panning and sluicing techniques still used by weekend 'hobby' prospectors to the far more massive dredging and hydraulic techniques used by mining companies. All of these techniques involve the use of flowing water to wash away (sluice) less dense grains (mainly silicates) from the heavier gold grains contained in sediment. Dredges are barges with automated conveyor-belt mechanisms for scooping up sediment at the front end, processing it on board to extract the gold, and then dumping the worthless tailings out the back end. The ecologically destructive effects of dredges on river systems can be significant, not only locally, but also far downstream because of silting.

Following the 1849 California Gold Rush, unconstrained hydraulic mining of placers with large water cannons (monitors) and river dredges wreaked environmental havoc on scenic landscapes and waterways from the Sierran foothills to San Francisco Bay, prompting John Muir to form the Sierra Club and ultimately forcing government legislation banning most such practices in the USA. However, such practices still continue in less-developed countries, particularly in Latin America and the Pacific Rim. Offshore (marine) placers, where rivers have discharged their sediment loads into the oceans, are also mined for gold.

Relying solely on hydraulic techniques to separate gold from sediment is not 100% effective, especially for fine gold grains. Amalgamation, or dissolving gold grains in liquid mercury, is far more effective and was used extensively until the development of cyanidation in the 1890s. Copper-lined tubs were coated (amalgamated) with mercury and the placer concentrate was placed in the tub along with more liquid mercury. The tub was then agitated until the gold amalgamated and adhered to the copper amalgam. The amalgam was then scraped off and the gold was reconstituted to a solid state via evaporation (distillation) of the mercury. Unless it was captured by retorts that recondense the vapour, the mercury was released into the environment and entered the food chain (and humans) mainly via fishes. The historic and modern use of amalgamation to recover gold from placers continues to cause a serious health problem in some areas. Areas downstream of the Mother Lode district still suffer from mercury contamination, more than a century after mining has ceased. Amalgamation of placer gold is still used in remote areas of Brazil by native miners who have no other means of employment and cannot afford retorts or cyanidation; their communities suffer from mercury poisoning as a consequence.

Hard-Rock Mining

Upon rapid depletion of the most easily found New World placer deposits in the late nineteenth century, gold miners' attention shifted to hard-rock mining

with the discovery of South Africa's Witwatersrand deposits in 1886. These enormous deposits have since been the world's major gold source, yielding over 30 000 metric tons of the metal. Hard-rock mining involves drilling, blasting, digging, and scraping, which increase the costs far above those of placer mining. Open pits and underground workings need to be shored up and stabilized, and tailings disposed of properly. For underground operations, engine exhaust, dust, and rock gas must be vented and fresh air continuously injected for workers to survive. Those mines reaching below the water table require constant pumping.

Surface mining of hard-rock gold ore is feasible even at modest grades, but costs rise rapidly with depth, and so underground mining is possible only if the grades and tonnages of ore are high. Some mines start out as open-pit operations and then convert to underground workings as the shallower ore is mined out. Though still quite rich in ore, many of the Witwatersrand mines are now quite deep and hot (3–4 km and up to 65°C) adding greatly to costs and reducing labour productivity. Dramatic increases in gold price in the late twentieth century fuelled global exploration and led to the discovery of many new gold districts, some of which are amenable to inexpensive cyanide heap leaching techniques. These competitive trends caused gold production from South African mines to peak in the 1970s and steadily decline thereafter. An ounce of gold recovered from Wits mines now costs roughly twice as much as one recovered from typical surface mines in the western USA. Adding further to the woes of the South African mining industry is the problem of a workforce that will be significantly affected by the AIDS virus, both in terms of reduced labour productivity and long-term healthcare needs.

Gold Ore Processing

Once hard-rock gold ores have been blasted and taken from the ground, their processing can involve several steps. If the gold is 'locked up' in phases such as tellurides or sulphides, then efforts must be made to separate these minerals from the rest of the rock. Crushing and grinding are accomplished by brute force, using steel jaws or cone crushers followed by ball or rod mills (giant tumblers containing steel balls or rods). Next comes classification, or segregation of ground mineral grains into fractions of different sizes. This step is accomplished by using screens of different mesh sizes, much like sifting flour. Following that, the different minerals are separated from one another by taking advantage of their differing densities or other physical properties. For example, an air-driven cyclone separator can effectively segregate dense gold and sulphide from less dense silicate minerals. These steps together constitute beneficiation and result in an ore concentrate that is separated from the worthless minerals or tailings. These steps typically take place as close to the mine as possible.

Finally there is the finishing step, in which the metal is extracted from the concentrate, purified, and readied for market. The precise steps taken to extract the gold depend strongly on the nature of the gold in the ore.

Concentrates containing only native gold can be treated by amalgamation (as described above). Aqueous solutions containing both an acid and an oxidant can also be used to dissolve gold. For example, aqua regia ($HCl + HNO_3$) dissolves gold as auric chloride

$$Au_{(s)} + 4HCl + HNO_3 = HAuCl_4 + 2H_2O + NO.$$

Since the 1890s, however, the most important method used to dissolve gold from concentrate has been cyanidation with oxygen

$$4Au_{(s)} + 8NaCN + O_2 + 2H_2O = 4NaAu(CN)_2 + 4NaOH.$$

The concentrate is placed in vats of cyanide solution and agitated to increase the contact time for the reaction between the grains and the solution. The pregnant (gold-bearing) liquor is then filtered from the solids and further processed to recover the gold, sometimes by reacting it with powdered zinc metal in a process known as cementation

$$Zn_{(s)} + NaAu(CN)_2 + H_2O + 2NaCN = Au_{(s)} + Na_2Zn(CN)_4 + NaOH + 0.5H_2.$$

Another method of enhancing gold recovery uses activated carbon (such as found in aquarium filters) to adsorb gold cyanide complexes from the pregnant liquor. Once loaded up with adsorbed gold (up to 250 troy ounces of gold adsorbed per ton of carbon) the carbon is leached with acid solutions at high temperature and pressure, releasing the gold to be recovered by cementation or electrowinning. Electrowinning involves passing an electrical current through steel-wool electrodes that are immersed in the pregnant leachate, causing the gold to plate out in place of the iron via an electrochemical reaction. The high surface area of the inexpensive steel wool makes the process very efficient and cost-effective.

Cemented or electrowon gold is impure (typically 70–75% gold) because it contains residual iron or zinc along with non-aurous metals derived from the ore, such as silver, copper, and mercury. Such impure material is refined into nearly pure gold by heating it

in crucibles in a furnace to more than 1000°C in the presence of silica, oxidizers, and fluxes. The impurities oxidize and melt to form a separate silica-rich slag, which floats on top of the nearly pure molten gold. The slag is skimmed off and the gold (containing more than 95% gold at this stage) is called dore. Molten dore is further refined to in excess of 99% purity by injecting it with chlorine, which picks up the remaining metal impurities and removes them as volatile chlorides and solid chloride salts. A final sequence of redissolution and electrowinning (or precipitation) improves the purity to more than 99.9% gold, which is then remelted and cast into gold bars for market.

When gold ores yield less than 80% of their gold after normal grinding and cyanidation, they are termed refractory – such ores typically contain gold locked up in sulphides, tellurides, and arsenides. Such ores require more expensive (finer) grinding and sometimes roasting methods to free up the gold prior to cyanidation. Roasting such ores in a furnace (to break the host phases down and release the gold) has environmental impacts because of the release of sulphur dioxide (which contributes to acid rain) and toxic arsenic gases.

Besides complicating the freeing up of gold during beneficiation, host phases can interfere with the effectiveness of cyanidation by having low solubility in the cyanide solutions, consuming oxygen or cyanide, or competing with gold to form aqueous cyanide complexes. Other phases in the ore, such as coatings of carbon or silica, or the presence of clay minerals can inhibit the reactions and adsorb gold and other ionic constituents, causing gold recoveries to be less than ideal.

High-pressure and high-temperature oxidation in strong acids is sometimes used to process refractory ores, but this adds substantially to costs. Other more recent (and less costly) methods include microbiological approaches, in which bacteria are used to break down refractory gold-bearing phases prior to cyanidation. Species such as *Thiobacillus ferrooxidans* catalyze the oxidative destruction of pyrite and arsenopyrite (*see* **Minerals:** Sulphides), freeing up gold grains that are locked up in such minerals.

Cyanide Heap Leaching

In many modern low-grade gold mines, particularly those of the western USA, the gold is finely disseminated in fractured near-surface rock that can be easily mined by open-pit methods and piled into heaps after only modest blasting and beneficiation. The clever and inexpensive technique of cyanide heap leaching is used to extract gold very effectively from such ore heaps, avoiding the costs associated with grinding and smelting more traditional high-grade hard-rock ores.

The ore heaps are piled on top of thick plastic mats, and a sprinkler system is installed on top of the heaps (Figure 5). A cyanide solution is sprinkled on the

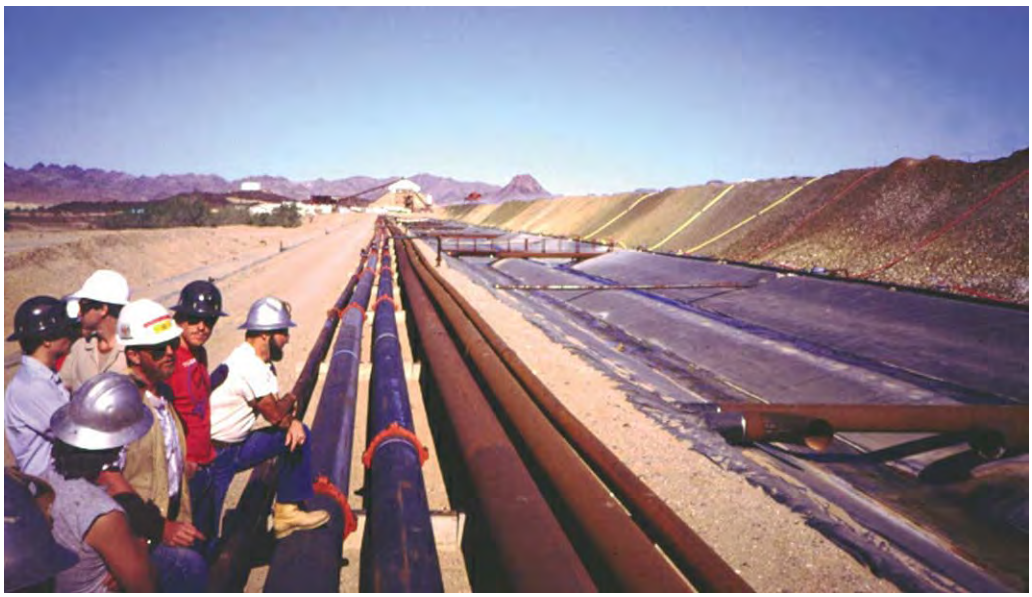


Figure 5 Heaps of low grade epithermal gold ore are piled on black plastic mats and sprinkled with sodium cyanide solution at the open pit Mesquite Mine, Imperial County, California, USA. One set of steel pipes delivers fresh solution to the sprinkler systems atop the heaps, while the other set carries away 'pregnant' gold laden cyanide liquor that drains from the heaps and accumulates in the trenches.

heaps and trickles down through them, dissolving out the gold and accumulating on the mat surfaces. Drains and pipes carry the pregnant effluent from the mats to an activated-carbon circuit, where the gold is adsorbed and recovered; the stripped carbon and barren cyanide solution are then reused. Such technological innovations have made low-grade near-surface gold ores economically competitive to mine in recent decades, especially in the western USA. Some mines have tried combining cyanide heap leaching and microbiological approaches to reduce costs further. Gold recovery via cyanide heap leaching often exceeds 95%.

Though innovative and cost-effective, the use of open cyanide circuits in close proximity to wildlife and water resources has led to public concerns and calls for increased regulatory scrutiny of such mining operations. After some mis-steps in the 1970s and 1980s, the mining industry has responded to such concerns. Most operations now take pre-emptive measures to keep birds and other wildlife away from the heaps and cyanide ponds, such as using net covers and installing propane cannons that discharge at 30 s intervals. Although natural bacteria in the soil probably decompose most escaping cyanide before it can percolate into groundwater, the name 'cyanide' still carries a very negative connotation and therefore care must be taken to design a leaching circuit that is as leak-free as possible.

Gold Assaying

While visible nuggets of gold found near Sutter's Mill were indisputable evidence of rich placers and got the '49ers' of the California Gold Rush very excited, most modern gold ores are far less dramatic and typically contain gold that is very fine-grained and often invisible, even under some microscopes. Much of this gold is in solid solution in sulphide and arsenide minerals, and only an assay (chemical analysis) or the use of far more costly electron microscopy can prove its presence.

Gold can be qualitatively detected in samples by dissolving them in aqua regia and using a simple colorimetic test based on the addition of tin salts. When stannous chloride ($SnCl_2$) is added to the solution, stannic ions are formed and hydrolyze to a stannic hydroxide flocculant. As the tin is oxidized to the stannic state, gold is simultaneously reduced and adsorbed on the stannic hydroxide, turning the solution a brilliant purple colour called the 'Purple of Cassius'. This is a very sensitive test for gold, detecting as little as one part in a hundred million (10 ppb).

The most common and traditional assay technique for precisely quantifying gold content is the gravimetric lead fire assay. A powdered ore sample is mixed with carbon and lead oxide and fused at 900–1100°C for at least an hour. The carbon reduces the lead in the oxide to molten lead metal, and any metals present in the powder dissolve (amalgamate) into the molten lead. After cooling, the solid lead 'button' is placed in a special cup called a 'cupel' (made of bone, ash, or magnesia) and heated to 950–1000°C; the cupellation forces any remaining lead oxide and base metals in the remelted button to be absorbed by the cupel. After cooling there remains a metal 'bead' containing only silver, gold, and platinum-group elements. Treating the bead with nitric acid dissolves out the silver, leaving a bead composed only of gold and platinum-group elements to be accurately weighed.

Other methods used for gold determination include volumetic titration by iodine, spectrophotometric absorbance by gold bromide, and atomic emission and absorption methods.

Making judgments of ore grades based on limited assays of small samples can be risky because of the nugget effect, whereby the presence of one large gold grain in a small rock or drill-core sample can give the misleading impression that the entire mineralized area has high grades of gold. History is replete with cases of such high apparent assays (or assays of 'salted' barren rock) being used to sway and cheat naive investors. Assaying is most accurate when performed by trustworthy individuals on multiple large samples of mineralized rock.

Knowledge of the local geology can also help in recognizing fraudulent or misleading assay claims. In an undercover news investigation, the author once witnessed scam artists impressing naive potential investors with 'representative' vials of coarse gold nuggets – but the nuggets were far too large and rounded to have come from the local fine-grained dry lake beds that were being touted as the site of rich placer gold.

Gold Markets and Economics

Three events in human history have lead to dramatic increases in global gold production: the discovery of the New World in 1492; the California Gold Rush of 1849; and the end of government-regulated gold prices in 1968.

Governments and individuals have long valued gold as a measure of wealth and a hedge against economic uncertainty. Many governments back their paper currency with gold reserves held in central banks, and some issue gold coinage – the Canadian Maple Leaf and the South African Krugerrand are current examples. Gold (and other precious metal)

prices are quoted in dollars per troy ounce for 900 fine metal.

For many years the 'official' value of gold was set by the USA at $20.67 per troy ounce. The St Gaudens Double Eagle, a $20 US coin, was minted from 1907–1933 and weighed 34 g. In 1934 the value of gold in the USA was increased to $35.00 per troy ounce, to build up reserves and stabilize the dollar after the Great Depression. Although this had a stabilizing effect on the currency markets, it gradually devalued gold over the next 34 years as inflation caused other commodity prices to rise slowly. American citizens were also prohibited from owning gold bullion, whereas foreign parties could purchase gold from the US government at the fixed price.

In 1968, partly because of the growing gap between the 'real' value of gold and the fixed price, a two-tiered pricing system was briefly established whereby gold was still transferred among governments at the 'official' price of $35.00 per troy ounce while the price on the private market was allowed to fluctuate. This tiered price system soon failed (mainly because of South African government sales at the higher market price) and in 1971 the US government finally abandoned the gold monetary standard, allowing gold prices to float freely. In 1974 American citizens were once again allowed to own gold bullion.

With the global oil crises of the 1970s, oil and gold prices spiralled upwards dramatically, the latter peaking at about $850 per troy ounce in 1980. Americans were now free to invest in gold bullion and gold futures markets, which expanded dramatically. The increased prices and influx of investment capital stimulated a new wave of global exploration for gold deposits. It was during this period that South Africa's century-long domination of global gold production from the Wits deposits began to wane. Many gold districts in the USA, Australia, and other nations were discovered or expanded, their financial bottom lines aided by cheap cyanide heap leaching technology.

Increases in the price of gold in the 1980s prompted a wave of gold investment scams, with crooks setting up temporary field operations and fancy offices that were designed to lure investors into buying 'future' gold ore in lots of $5000–$15 000. Enticements of free food, alcohol, and female companionship were often used in such 'investment seminars'. Apparently some investors were hesitant to report losses from such scams because they were using money that they did not want tax officials or former spouses to know about.

Being highly influenced by emotion, greed, international conflict, and economic uncertainty (rather than a more rational basis of technological usefulness), trends in gold price have never lent themselves to accurate prediction. Recent years have seen downward pressure on prices as some western governments sell off some of their central banks' gold reserves, while continued fears of economic and political conflicts in Asia, Europe, and the Middle East have pushed prices upwards. The net result is, as always, erratic and unpredictable trends.

More than any other Earth resource, the concept of 'caveat emptor' applies to gold in all of its manifestations. The most successful investors in gold are probably those who buy stock in reputable gold-mining companies with adequate reserves and low overheads.

See Also

Atmosphere Evolution. **Economic Geology**. **Fluid Inclusions**. **Minerals:** Native Elements; Sulphides. **Mining Geology:** Hydrothermal Ores. **Tectonics:** Hydrothermal Activity.

Further Reading

Amey EB (2002) Gold. In: *US Geological Survey Minerals Year book 2002*, vol. 1, pp. 1 6. US Geological Survey, Washington, DC: Govt. Printing Office.

Boyle RW (1987) *Gold: History and Genesis of Deposits*. New York: Van Nostrand Reinhold.

Craig JR, Vaughan DJ, and Skinner BJ (2001) *Resources of the Earth: Origin, Use, and Environmental Impact*, 3rd edn. Prentice Hall, Upper Saddle River.

Foster RP (ed.) (1991) *Gold Metallogeny and Exploration*. Glasgow: Blackie.

Gasparrini C (1993) *Gold and Other Precious Metals: From Ore to Market*. New York: Springer Verlag.

Hagemann SG and Brown PE (eds.) (2000) *Gold in 2000*. Reviews in Economic Geology 13. Littleton: Society of Economic Geologists.

Kesler SE (1994) *Mineral Resources, Economics and the Environment*. New York: MacMillan.

Kirkemo H, Newman WL, and Ashley RP (2000) *Gold*. Factsheet. US Geological Survey.

Yannopoulos JC (1991) *The Extractive Metallurgy of Gold*. New York: Van Nostrand Reinhold.

GONDWANALAND AND GONDWANA

J J Veevers, Macquarie University, Sydney, NSW, Australia

Introduction

Gondwanaland existed from the 600–500 Ma accretion of the African and South American terranes to Antarctica–Australia–India, through the 320 Ma merging with Laurussia to form Pangaea, until breakup between 180 Ma and 100 Ma (Figures 1 and 2).

The name Gondwanaland was introduced in 1885 by Eduard Suess for the regions with the *Glossopteris* flora, in particular the Gondwana System of peninsular India. In 1912 Alfred Wegener interpreted Gondwanaland as the supercontinent with Gondwanan floras. Later there was a view that Gondwanaland should be replaced by Gondwana, which, interpreted as 'Land of the Gonds', included one 'land' already. Semantic confusion of supercontinent and stratigraphical system was averted when 'wana' was found to stand for 'forest', so that Gondwanaland means the supercontinent and Gondwana means the Indian kingdom of forest dwellers. The distinction is valuable because the Gondwana facies started only after Gondwanaland merged with Laurussia to form Pangaea. However, not all workers use this terminology.

This account of Gondwanaland is told through a set of maps that stretch from assembly, through the merger with Laurussia to form Pangaea, to breakup.

Early–Middle Cambrian (530–500 Ma)

In the Early–Middle Cambrian, Gondwanaland was bounded to the north (at the modern coordinates of Africa) by terranes now in Laurentia, Europe, and Asia, and to the west and south by a trench (Figures 3 and 4). The interior was crossed by fold belts generated during the terminal Pan-Gondwanaland (600–500 Ma) deformation, which endowed Gondwanaland with a thick buoyant crust and lithosphere, and nonmarine siliciclastic deposits behind a peripheral shoreline.

Between 650 Ma and 570 Ma, stress 1 (Figure 3) was generated from the oblique collisions of, first, Avalonia–Cadomia with the West African Craton and, second, West Gondwanaland with East Gondwanaland during the closure of the Mozambique Ocean. The West African Craton, which was rotated counterclockwise, imparted clockwise rotation to the Amazonia Craton, which, in turn, rotated the Congo Craton counterclockwise. Between 550 Ma and 490 Ma, stress 2 was generated first by oblique subduction of the Palaeo-Pacific Plate beneath Antarctica and second by transcurrence beneath Avalonia–Cadomia and the West African and Amazonia Cratons. The India–East Antarctica–West Australia Craton was driven into counterclockwise rotation, which imparted clockwise rotation to the North Australia Craton, modelled by dextral shear along small circles about a pole in the Pacific. A fold belt was extruded between the West African and Congo Cratons, and another between the Congo and Amazonia Cratons. The cycle ended at about 500 Ma with final convergence along the Palaeo-Pacific margin and uplift and cooling in Gondwanaland. The heat emitted during convergence, added to that generated during the Pan-Gondwanaland cycle, built buoyancy into Gondwanaland by underplating the lower crust with mafic magma to promote isostatic uplift and concomitant downwearing.

The shoreline alongside Antarctica, through Australia and north-west India, continued along the northern margin to the Levant and north-east Africa, with marine sediment deposited in belts of terrigenous, mixed, and carbonate facies across West Africa and southwards (in modern coordinates) into South America (past the north pole) and then across the Damara fold belt, to link with carbonate in the Transantarctic Mountains, and over flood basalt on the Australian platform. In the south-east, a newly generated marginal basin started to close by north-eastward-directed subduction beneath a volcanic arc.

Above subducting slabs, the margin includes granite in the Suwannee terrane of Florida, granite in Argentina and beside a rift projected towards the Transantarctic Mountains, the string of Ross plutons, the Delamerian granites, and granites in north-eastern Australia. The Prydz-Leeuwin Belt and Mozambique Orogenic Belt were metamorphosed.

Early Ordovician (490–458 Ma)

In the Early Ordovician the north pole lay in the Sahara (Figure 5), but ice did not appear until 444 Ma. Australia was crossed by the (Larapintine) sea behind a magmatic arc generated by westwards-directed subduction of ocean floor that was flooded by fans of quartzose sediment from Antarctica.

Areas of Pan-Gondwanaland deformation had cooled. The shoreline lapped the Beardmore Shelf and Table Mountain Shelf. In the ancestral Paraná

Basin, north-east-trending rift basins filled with nonmarine detritus, including rhyolitic volcaniclastics. Beach sand was deposited in Florida, and carbonate was deposited around the ancestral Sierra Pampeanas, which were intruded by granite.

The Avalonian terranes drifted away from Africa at 470 Ma in the first of many transfers of material from Gondwanaland to the 'northern' continents.

Late Ordovician (458–443 Ma)

The end-Ordovician glaciation affected Saharan Africa and southern Africa–South America (Figure 6). Glacial advances and retreats climaxed with a big advance at the end of the Ordovician (444–443 Ma). In the Sahara, directional structures indicate uplands including Sudan and Arabia. Interglacial marine incursions swept over nonmarine glacigenic landforms and sediment. Distal glacimarine deposits extend over much of the northern margin of Africa–Arabia, including the marginal terranes. Other ice centres on uplands lay 3000 km away along the Pacific margin, suggesting that proximity of uplands to the sea was a factor. Nonmarine glacigenic deposits are found in Bolivia and Venezuela.

The interglacial sea received marine glacigenic sediment in the Volta and Bowé Basins, Iberia and Cadomia, North Africa, southern Turkey, and Arabia, which were all fringed by nonmarine basins and uplands and scattered anorogenic igneous complexes. The shoreline crossed northern India and continued alongside Western Australia and the North China Shelf. The intermittent Larapintine Seaway produced thick halite evaporites in the west. In the east, the sea opened onto a convergent margin and magmatic arc. The shoreline passed the Beardmore Shelf to the embayments of the Table Mountain Shelf and Don Braulio area.

Early Silurian (443 Ma)

At the beginning of the Silurian, the ice-sheets contracted to a strip in North Africa–Arabia and advanced over Brazil. In transgressing North Africa, the post-glacial sea accumulated glacigenic sediment, including 'hot shales' (which are sources of petroleum) with graptolites in the Fort Polignac Basin, shed from surrounding uplifts, including the Sudan upland, which was intruded by anorogenic complexes (Figure 7). Uplifts in South America likewise accumulated ice. Glacimarine diamictites and shales were deposited in the Paraná Basin, including the Iapó Formation, and in the Parnaíba Basin. Nonmarine diamictites were deposited in the eastern Paraná Basin and the Jatobá Basin. The marine diamictite of the Amazonas Basin crops out in a 1500 km long belt in the eastern Andes. Farther south, the uppermost Don Braulio Formation is not glacigenic. The Malvinokaffric (zoogeographical) Realm occupied South America, southern Africa, and Antarctica before it disappeared in the Middle Devonian.

In Antarctica, the nonmarine Crashsite Group in the Ellsworth Mountains lay behind a zone of deformation and metamorphism that extended through Marie Byrd Land and New Zealand into Eastern and Central Australia, including the Melbourne Terrane and Benambran Highlands (BH). Evaporites were deposited in the Bonaparte Basin.

Early Devonian (418–394 Ma)

In the Early Devonian the south pole lay off Patagonia but glacigenic deposits are unknown. Laurentia was about to make contact (Figure 8). Shelves in Arabia and North Africa widened during a marine transgression from the Lochkovian (415 Ma) to the Emsian (400 Ma) that formed a wide embayment with its head at Accra. Carbonate sediment indicates warm to moderate water, consistent with the tropical latitude. The Sudan upland, which was peppered with ring complexes, shed sand into nonmarine basins, including the Kufra Basin, which lay inland of the marine Ghazalat Basin. Past Arabia, the shoreline continued through the Chitral area of Pakistan and northern India to north-west Australia.

The Emsian shoreline passed the Lolén Formation and Bokkeveld Group. The rest of South America was crossed by the arms of a shallow sea between large islands that faced the Pacific margin on the west. Granite was emplaced in the Sierra Pampeanas.

The shoreline passed the Ellsworth Mountains and Ohio Range and was backed by nonmarine sandstone in the Pensacola Mountains, Beardmore Glacier area, and South Victoria Land. In New Zealand, marine sediment was deposited on terranes that amalgamated at 415 Ma. Amalgamation of north-east and west Tasmania followed at 400 Ma. In south-east Australia, rivers deposited sediment in the wide back-arc region of granite intrusion behind a volcanic arc and subduction complex. Beyond a group of nonmarine deposits in the centre, evaporites were deposited in the Bonaparte Basin, aeolian and playa deposits in the Canning Basin, and nonmarine redbeds in the Carnarvon Basin.

Late Devonian (382.5–362 Ma)

In the Late Devonian (Figure 9), the sea, initially (in the Frasnian) still in North Africa and South America and accumulating black shale, retreated in the Strunian

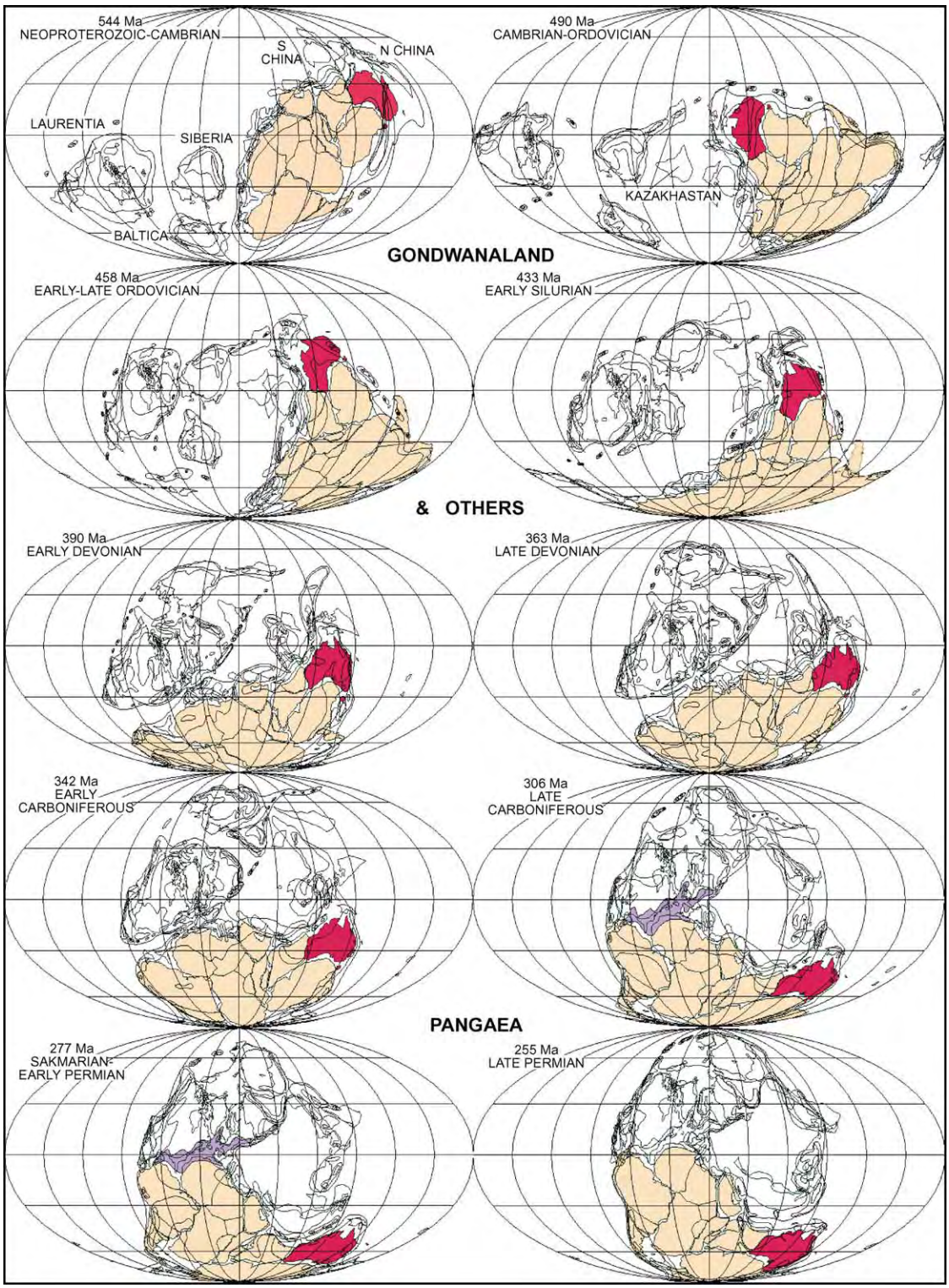


Figure 1 Australia (red), associates in Gondwanaland (brown), and other continents and terranes (Laurentia, Baltica, Siberia, Kazakhastan, Cimmeria, North and South China) through ten stages of the Palaeozoic. Reproduced with permission from Veevers JJ (2001) *Atlas of Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press and C. R. Scotese.

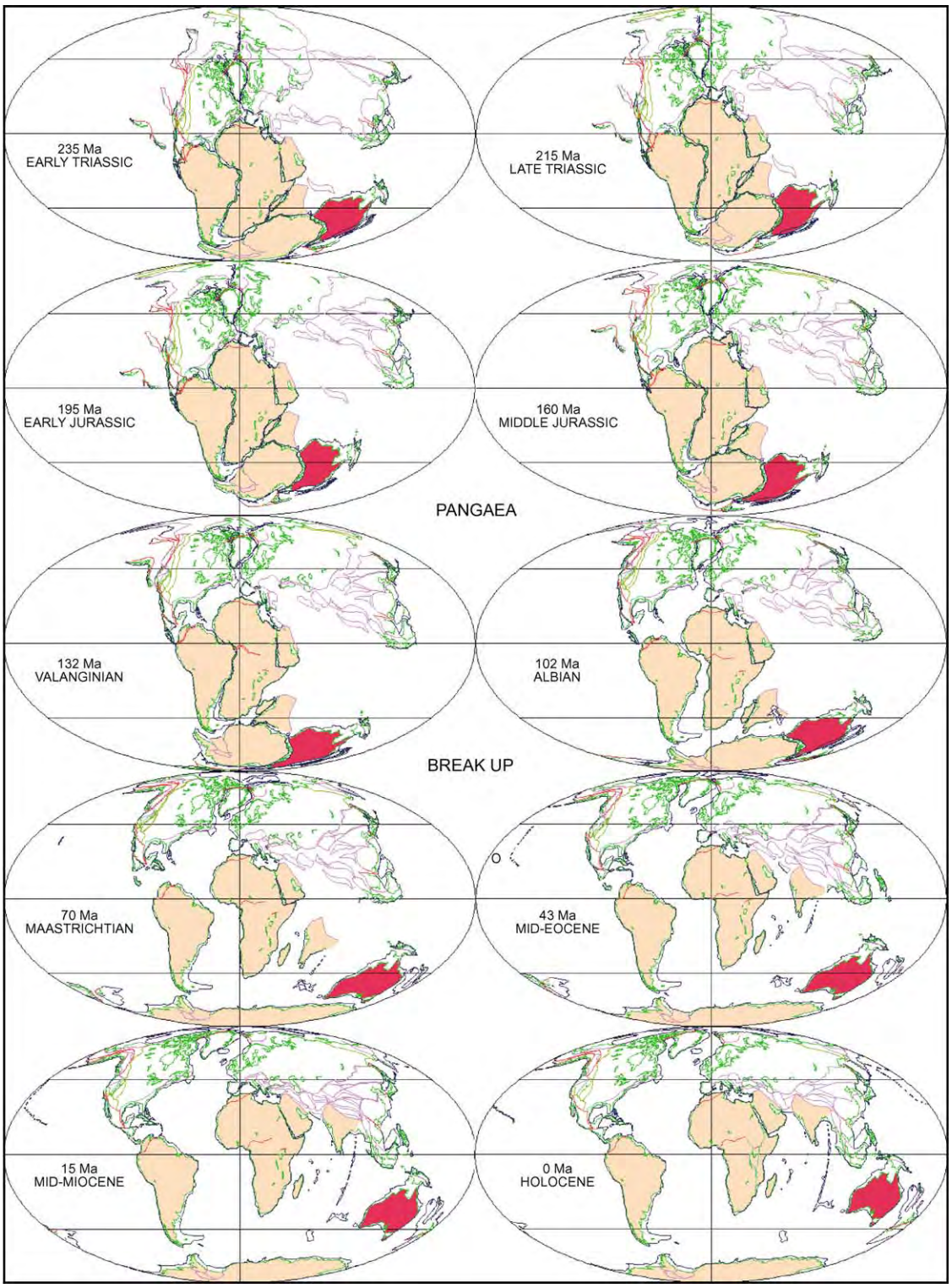


Figure 2 Australia (red), associates in Gondwanaland (brown), and other continents and terranes through ten stages in the Mesozoic and Cenozoic. Reproduced with permission from Veevers JJ (2001) *Atlas of Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press and C. R. Scotese.

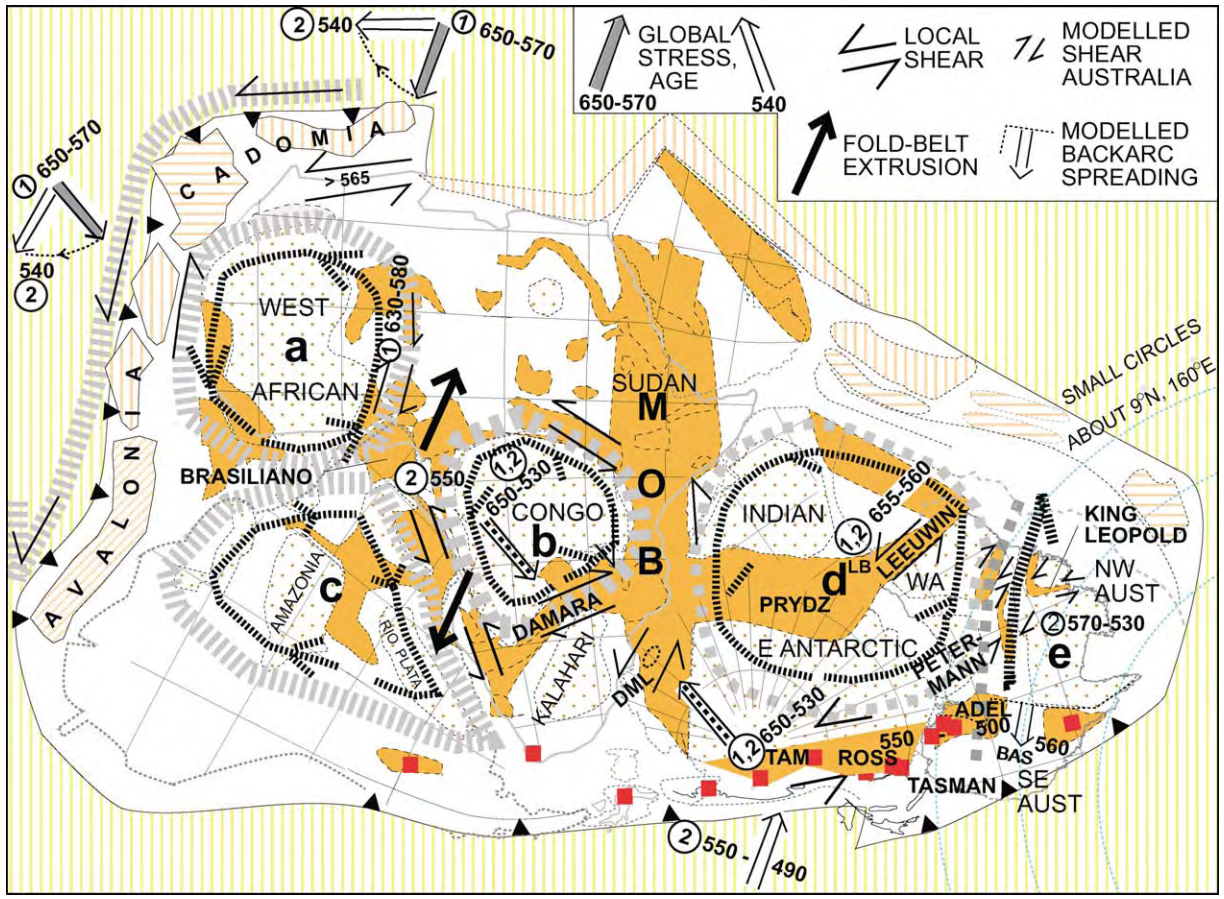


Figure 3 Early Cambrian Middle Cambrian (530 500 Ma) tectonics of Gondwanaland, showing cratons and sense of rotation (narrow broken lines and arrows) grouped into composites (enclosed by wide broken line): a, West African Craton; b, Congo Craton; c, Amazonia Rio de la Plata Craton; d, India East Antarctica West Australia Craton, bisected by the Prydz Leeuwin belt; e, North Australia Craton. Pan Gondwanaland (600 500 Ma) fold belts are orange. Ages are given in Ma. Circled numbers indicate stresses 1 and 2. DML, Dronning Maud Land; LB, Leeuwin block; MOB, Mozambique orogenic belt; TAM, Transantarctic Mountains; WA, Western Australian Craton; BAS, back arc spreading. Reproduced with permission from Veevers JJ (2003) Pan African is Pan Gondwanaland: oblique convergence drives rotation during 650 500 Ma assembly. *Geology* 31: 501 504.

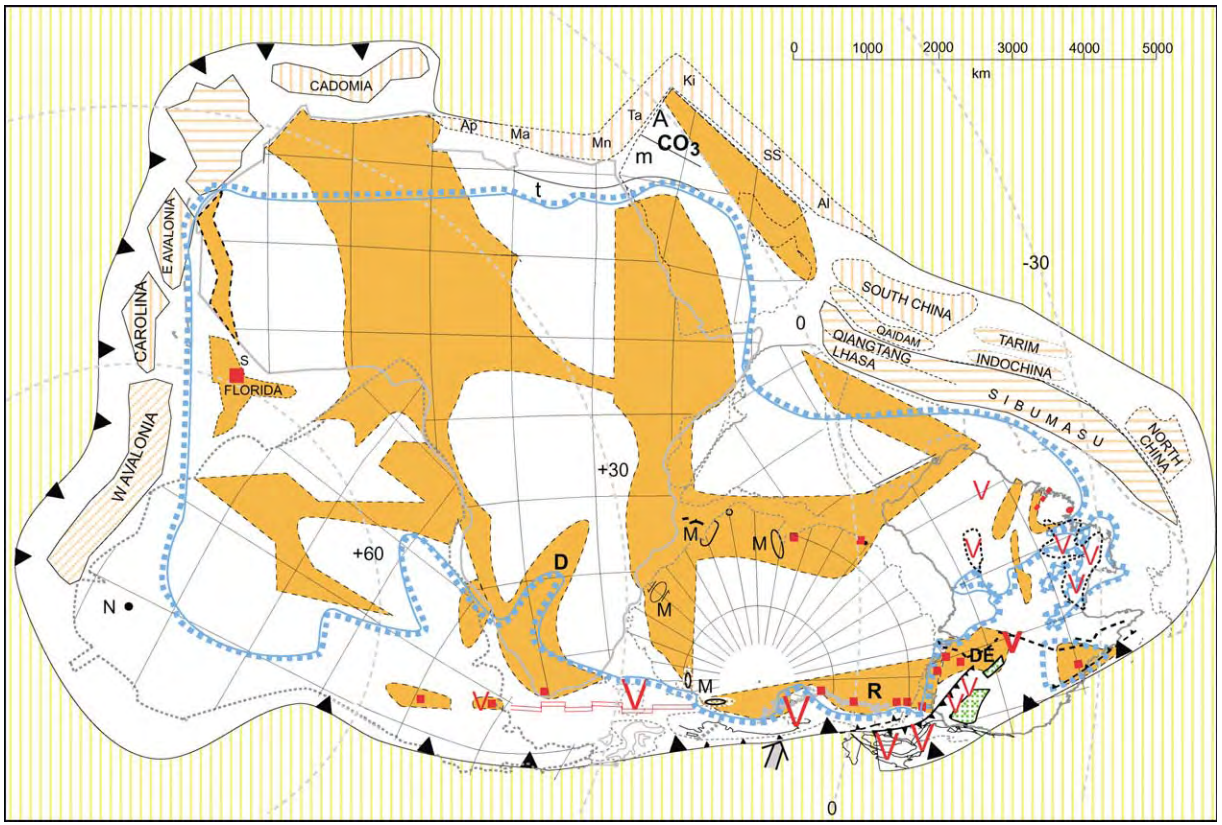


Figure 4 Early Cambrian Middle Cambrian (530 500 Ma) palaeogeography, with shorelines (blue line dots away from the land) skirting fold belts (orange). Prospective terranes run from West Avalonia to Cadomia and from Apulia to Sibumasu/North China. From west to east, the terranes on the north include: Ap, Apulia; Ma, Mani; Mn, Menderes; Ta, Taurus; Ki, Kirshehir; SS, Sanandaj Sirjan; Al, Alborz. Facies indicated as: t, terrigenous; m, mixed; and CO_3, carbonate (Red Vs indicate volcanics, red squares granite). Other abbreviations: D, Damara fold belt; S, Suwannee terrane; DE, Delamerian granites; R, Ross plutons and M, metamorphism. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

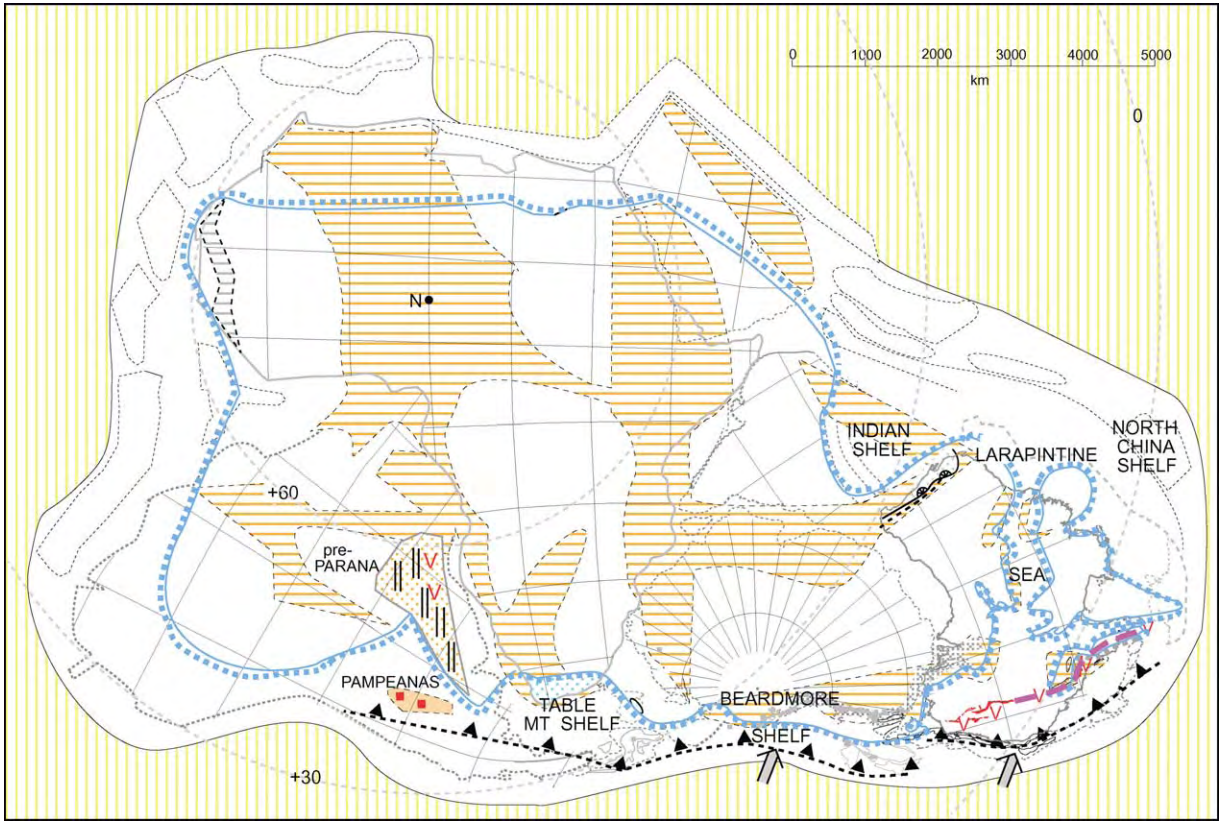


Figure 5 Early Ordovician (490 458 Ma) palaeogeography. N, position of north pole; orange horizontal shading, areas of Pan Gondwanaland deformation. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

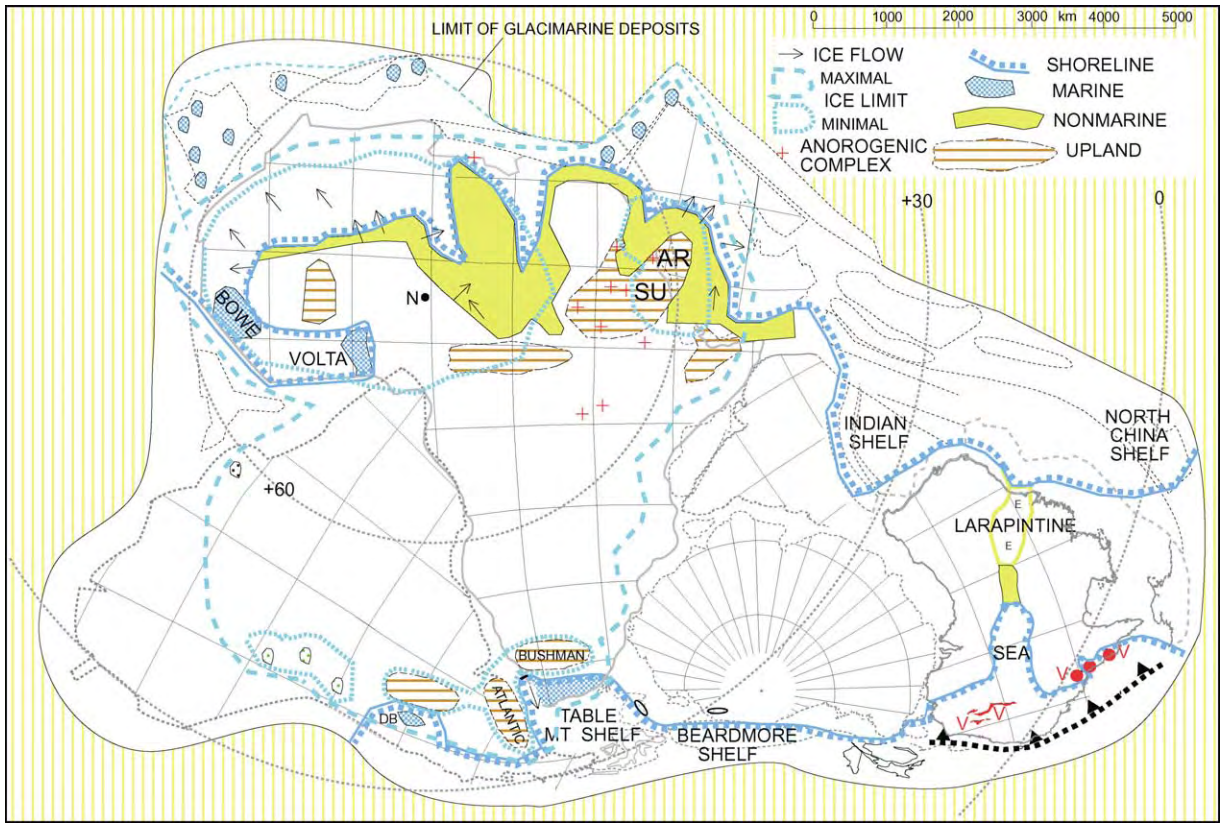


Figure 6 Late Ordovician (458 443 Ma) palaeogeography: DB, the glacimarine Don Braulio Formation; AR, Arabian upland; SU, Sudan upland; E, area of evaporite deposition; + anorogenic complex. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

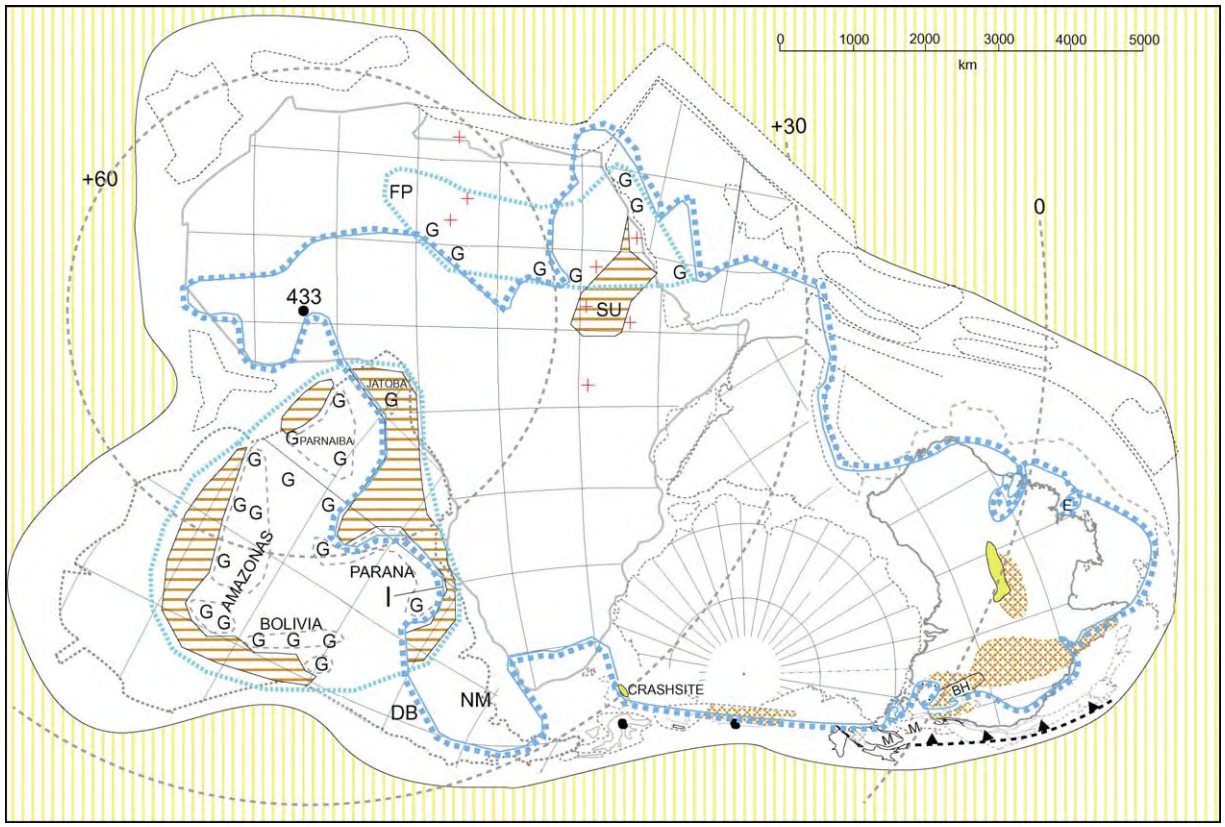


Figure 7 Early Silurian (443 Ma) palaeogeography. Shading as in **Figure 6**, with brown cross hatching representing deformation and metamorphism. NM, nonmarine sediment in the Sierra de la Ventana of Argentina; G, glacigenic sediment; FP, Fort Polignac Basin; SU, Sudan upland; red crosses, anorogenic complexes; I, Iapó Formation; DB, Don Braulio Formation; M, Melbourne Terrane; BH, Benambran Highlands; E, evaporites. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

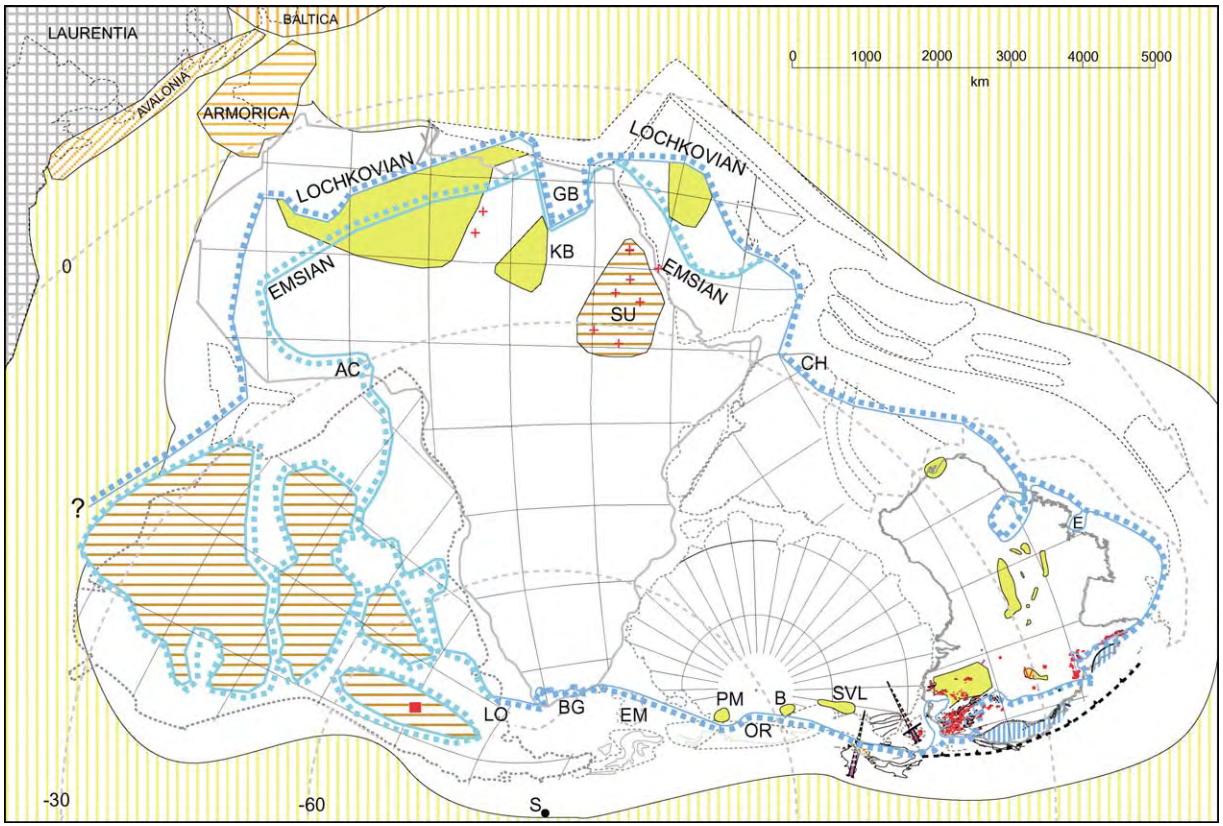


Figure 8 Early Devonian (418 394 Ma) palaeogeography: AC, Accra; SU, Sudan upland; KB, Kufra Basin; GB, Ghazalat Basin; CH, Chitral; LO, Lolén Formation; BG, Bokkeveld Group; EM, Ellsworth Mountains; OR, Ohio Range; PM, Pensacola Mountains; B, Beardmore Glacier; SVL, South Victoria Land; E, evaporites; horizontal orange shading, large islands separated by shallow sea; screw symbol, amalgamation of terranes. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

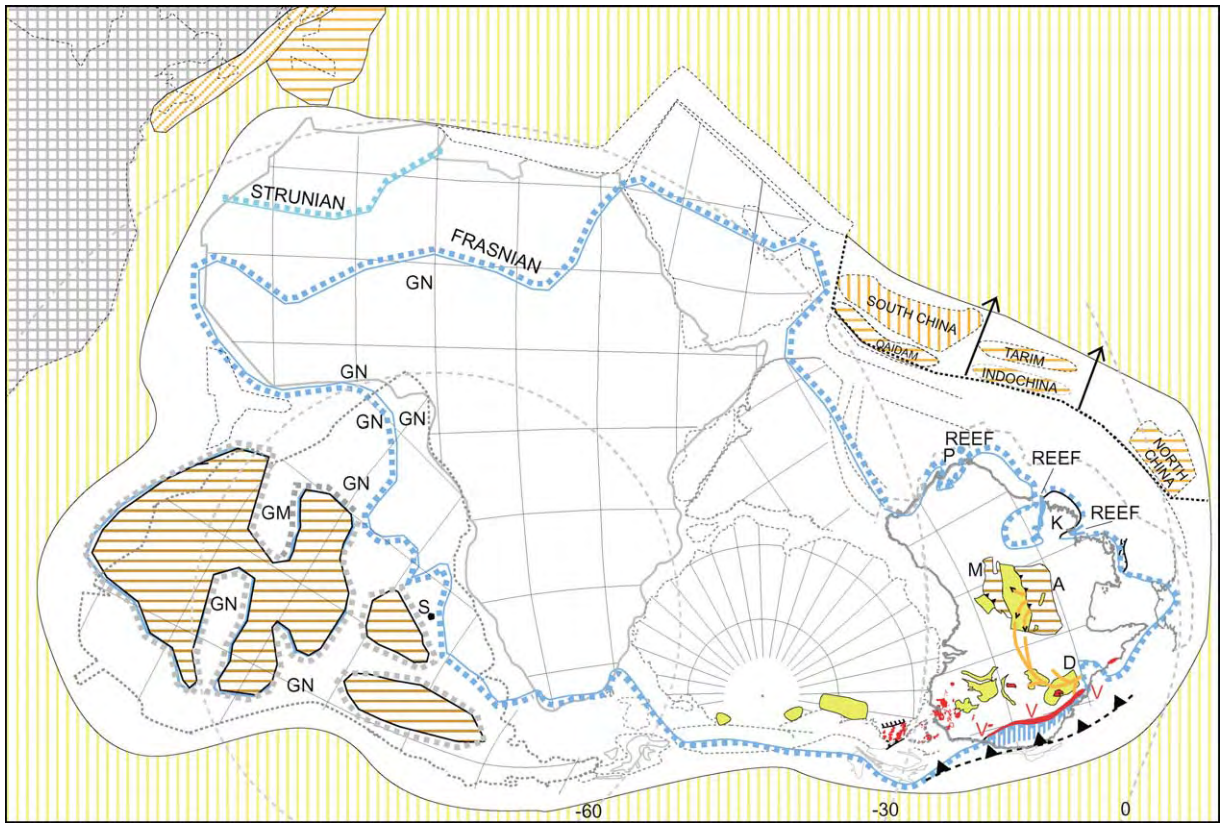


Figure 9 Late Devonian (382.5 362 Ma) palaeogeography: GN, glacigenic nonmarine sediment; GM, glacigenic marine sediment; A, Arunta block; M, Busgrave block; D, Drummond Basin; K, Kimberley block; P, reef facies. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

(latest Devonian). Terranes started to break off in the north-east leading to the generation of Palaeo-Tethys. The Centralian Superbasin, which had been subsiding since 840 Ma, was initially dismembered during crustal shortening, and sediment flowed into the convergent eastern Australia. Laurussia touched Gondwanaland.

In South America, black shale was deposited during the Frasnian, and glacigenic nonmarine sediment (glacimarine sediment in the Amazon Basin) was deposited during the end-Famennian regression. Glacigenic sediment extended into adjacent Africa. In North Africa–Arabia, the depositional facies changed from detrital in the south to carbonate in the north. In southern Africa and Antarctica, the shoreline and nonmarine depocentres remained in the same place as before. During continuing westward-directed subduction, granite was intruded in North Victoria Land and northwards, where volcanic rifts opened after the contractional Tabberabberan Orogeny. Farther north a magmatic arc and subduction complex developed behind the trench.

In central Australia, the overthrusting Arunta block disrupted the wider Amadeus Basin and culminated in 5 km of uplift reflected in extensive alluvial fans. Similar sediments were shed southwards from the overthrusting Musgrave block. In the Amadeus Basin, eastward flow joined the northward flow in the Drummond Basin to debouch on the margin. In the Bonaparte Basin, alluvial fans were succeeded by a reef complex. Likewise in the Canning Basin, gravel fans from the Kimberley block mingled with a reef complex on a platform in front of a deep axis; in the south, paralic sediment was deposited in an arm of the sea. In the Carnarvon Basin, shelf limestone gave way to reef facies.

Early Carboniferous, Visean (335 Ma)

In the Early Carboniferous, the Moroccan salient of Gondwanaland collided with the Armorican salient of Laurussia. Palaeo-Tethys further separated the Chinese blocks (Figure 10).

In North Africa–Arabia, the shoreline, bounded by nonmarine sediment, penetrated almost to Nigeria. In the east, uplifts include the Sudan Arch, the Central Arabian Arch, and the Summan Platform; igneous centres include rhyolite.

In South America, an Early Carboniferous contraction with arc magmatism (not shown) led to deformation and uplift that drove out the sea. Floodplain deposits with coal are found in the Lake Titicaca area. Glacigenic nonmarine sediment was deposited in the Pimenta Bueno, Jaurú, Solimões, central Amazonas, and Parnaíba–Itacaja areas. The pole was in north-east Africa, so the tillitic beds in North Africa were within 30° of the pole, but the glacigenic localities of South America, between 25° and 60° from the pole, all nonmarine, reflect alpine glaciation.

South-eastern Australia was subjected to the Kanimblan east–west contraction, followed by widening of the volcanic arc and finally north–south contraction in megakinks. Similar events took place in northern Queensland. In between, nonmarine deposition in the Drummond Basin continued until 330 Ma, when it was terminated by gentle folding, part of the Alice Springs terminal folding and thrusting that dismembered the Centralian Superbasin, a distant effect of the collision in north-west Africa. The youngest preserved sediment in the foreland basins of Central Australia is the 330 Ma Mount Eclipse Sandstone.

Mid-Carboniferous, Namurian (327–311.5 Ma)

In the Namurian, Laurussia and Gondwanaland merged to form Pangaea by definitive right-lateral contact along the Variscan suture, and the collisional stress and subsequent uplift was felt as far afield as Australia (Figure 11). The south pole had moved 60° since 335 Ma to a location in Marie Byrd Land, and ice sheets developed on the tectonic uplands south of 25° S in South America and south of 45° S elsewhere. The collisional uplands of equatorial north-west Africa and adjacent Europe were subjected to intense rainfall and backwearing.

Rapid uplift and concomitant downwearing must have produced copious sediment, yet the depositional record over the Australian platform is blank. Where could the sediment have gone? The paradox can be resolved by postulating that uplift combined with rapid polar movement triggered a continent-wide glaciation so that sediment shed from the nunataks of the central uplifts and from the eastern cordillera was carried away in the ice-sheet. Only in the east was nonmarine glacial sediment deposited by glaciers that broke through the eastern cordillera. During its retreat in the earliest Permian the ice released its load of sediment. Ice-sheets continued across Antarctica through the Ross and Gamburtsev areas, and into adjacent India and southern Africa, through uplands called Windhoek, Cargonian, and proto-fold belt. A separate ice-sheet covered uplands in South America, which shed glacimarine sediment in front of a volcanic arc.

Now part of Pangaea, Gondwanaland underwent a (Pangaean) cycle of tectonic and climatic events.

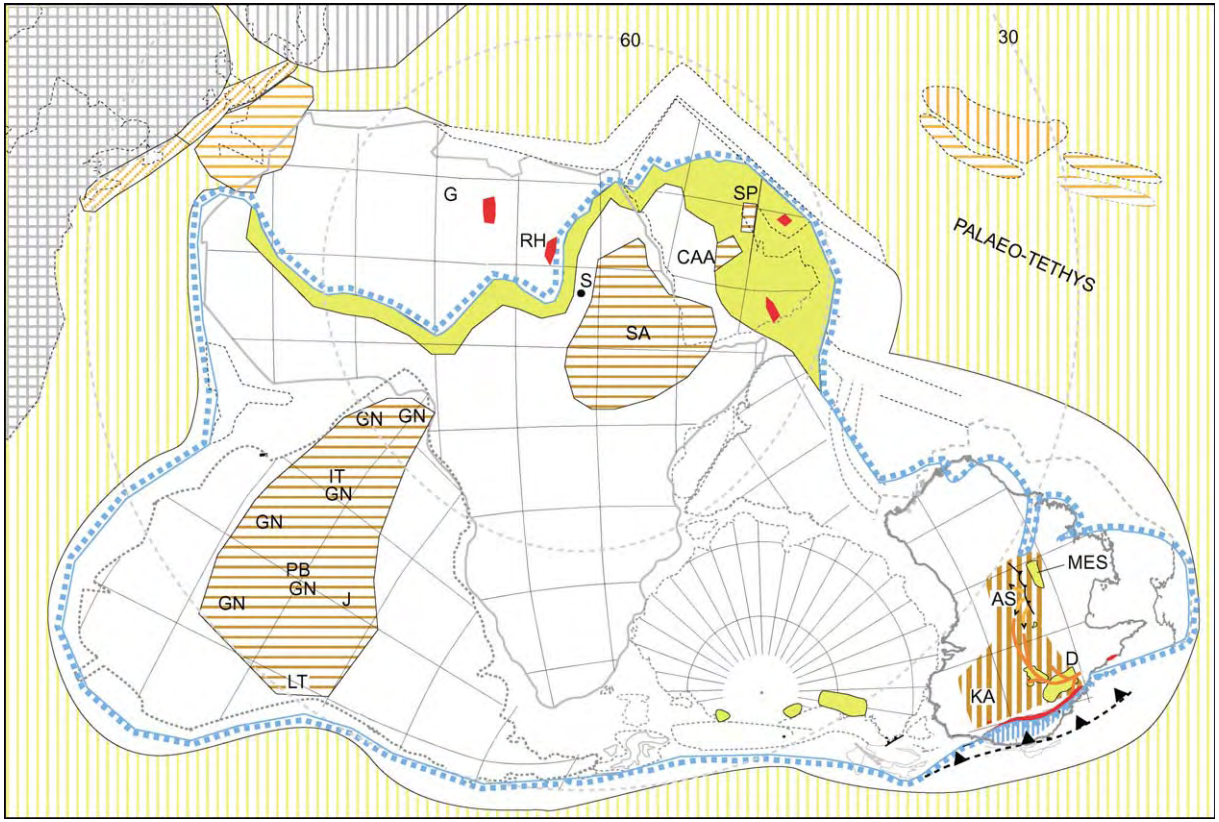


Figure 10 Early Carboniferous (335 Ma) palaeogeography: SA, Sudan Arch; CAA, Central Arabian Arch; SP, Summan Platform; RH, rhyolite; LT, Lake Titicaca; GN, glacigenic nonmarine sediment; PB, Pimenta Bueno; J, Jaurú; IT, Itacaja; G, tillitic beds; KA, Kanimblan contraction; D, Drummond Basin; AS, Alice Springs; MES, Mount Eclipse Sandstone; red shading, igneous centres; orange upland; lime green nonmarine fringe behind shoreline. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

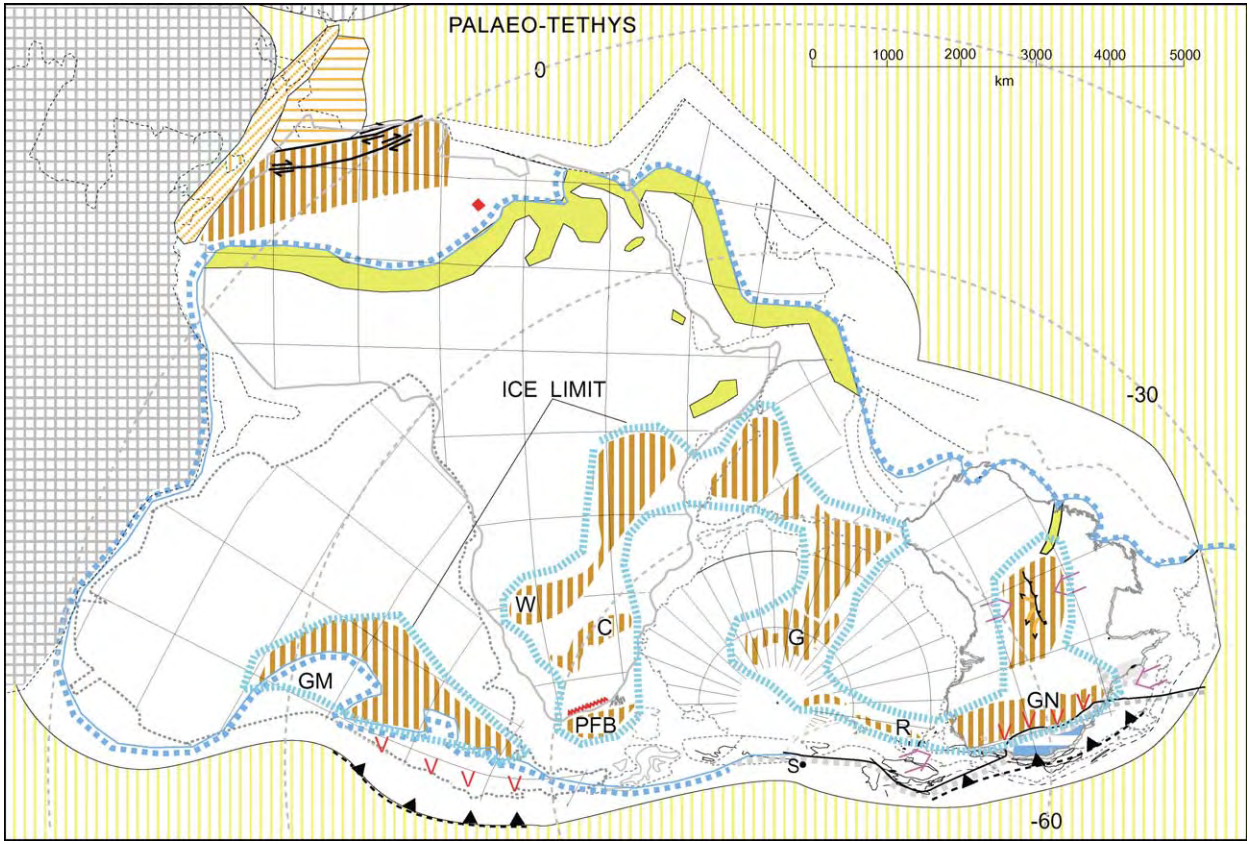


Figure 11 Mid Carboniferous, Namurian (327 311.5 Ma), palaeogeography. Pink arrows in central and eastern Australia indicate farfield stress. GN, glacigenic nonmarine sediment; R, Ross area; G, Gamburtsev area; W, Windhoek upland; C, Cargonian upland; PFB, proto fold belt; GM; glacimarine sediment; vertical orange shading, collisional stress and uplift. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

Tectonic–Climatic Model of the Pangaea-to-Dispersed-Continents Cycle

The heightened radioactivity of continental rocks led to the idea that a system of ascending convection ('monsoonal') currents beneath an insulating supercontinent would spread out at the top and eventually break the supercontinent into individual continents. Changes in the pattern and vigour of mantle convection would increase atmospheric carbon dioxide, creating a greenhouse effect. Heat was released in five stages (Figure 12).

- Stage 1 (platform lacuna). The Earth comprises the single continent (Pangaea) and ocean (Panthalassa). The amounts of spreading and subduction are minimal. Minimal turnover of mantle material leads to minimal venting of carbon dioxide and an icehouse climatic state. Maximal continental freeboard is due to, first, the short and narrow mid-ocean ridge displacing less water, second, the continental crust having a maximal mean thickness because its ocean frontage of thin (rifted) crust is minimal, and, third, the self-induced Pangaean heat store and accelerated mantle plumes generating a geoid high. The Earth is dominated by dry land.
- Stage 2. The heat impounded beneath Pangaea soon leads to localized thinning of the Pangaean crust and lithosphere, initially by sagging of cratonic basement and rifting of orogenic basement (extension I).
- Stage 3. Continued crustal thinning leads to rifting between the incipient continents (extension II).
- Stage 4. Pangaea breaks up by spreading of intra-Pangaean rift oceans to form dispersed continents and oceans. The mid-ocean-ridge spreading and subduction are maximal, leading to maximal carbon dioxide venting and greenhouse conditions. Low continental freeboard arises from, first, the mid-ocean ridges displacing more water, second, the crust of the continents having a minimal average thickness because its ocean frontage of thin crust is maximal, and, third, the rapidly depleting Pangaean heat store and decelerated mantle plumes supporting a lower geoid. The Earth is dominated by ocean.
- Stage 5. The depleted heat store leads to slower spreading and subduction and preferential closing of the rift oceans, so that eventually the continents reform Pangaea and the oceans reform Panthalassa in a return to Stage 1.

Earliest Permian (302–280 Ma)

Following a lacuna (Pangaean stage 1), the heat beneath Pangaea drove differential subsidence of the Gondwanaland platform (stage 2, extension I) to trap sediment released from the ice (Figure 13). Terranes left the northern margin, and rift zones penetrated East Africa and between India and Australia. Granite moved into transtensional rifts, driven by the right-lateral shear between Laurussia and Gondwanaland. Coal with the *Glossopteris* flora succeeds glacigenic sediment in all parts of Gondwanaland except tropical South America and Africa. A magmatic and orogenic zone along the Panthalassan margin (Alexander Du Toit's Samfrau Geosyncline) developed in South America.

The Sakmarian (288 Ma) postglacial shoreline ran along the margin in North Africa and made broad indentations across Arabia and India (the India–Australia Rift Zone) and narrow indentations across north-west Australia. It lapped the magmatic zone of eastern Australia and enclosed a gulf between Australia and Antarctica and another between Antarctica and South America.

The Gamburtsev upland shed glacial sediment into southern Africa, India, and south-west Australia, and the Beardmore-Ross upland continued to shed sediment into south-east Australia and New Zealand. Ice extended across the Congo Basin into North Africa and Arabia, with outwash material along the margin.

The inception of Gondwanan glaciation is linked to the lowered input of carbon dioxide in Pangaean stage 1, the removal of atmospheric carbon dioxide during accelerated erosion and weathering of the uplands, and the blocked oceanic circulation at the equator.

Glossopteris of the Gondwana palaeobotanical province made voluminous coals during the entire Permian (Figure 14). The Euramerica province extended into northern South America and Africa, and all three provinces – Euramerica, Gondwana, and Cathaysia – were juxtaposed in Arabia. The boundary between the Euramerica and Gondwana provinces approximates the ice limit (Figure 13). The Variscides also contain coal but without *Glossopteris*.

Late Permian (255–250 Ma)

In the Late Permian the sea retreated to the margins except in the north-east (Figure 15). The magmatic orogen (Gondwanides I) had propagated past northern Queensland. It was backed by the Karoo foreland basin, which was occupied by a vast lake with endemic nonmarine bivalves (preserved in the Waterford Formation, which extends into the Estrada Nova Formation of South America). Smaller lakes crossed the rest of southern and central Africa.

In eastern Australia, volcanogenic sediment flowed across a foreland basin and forebulge into

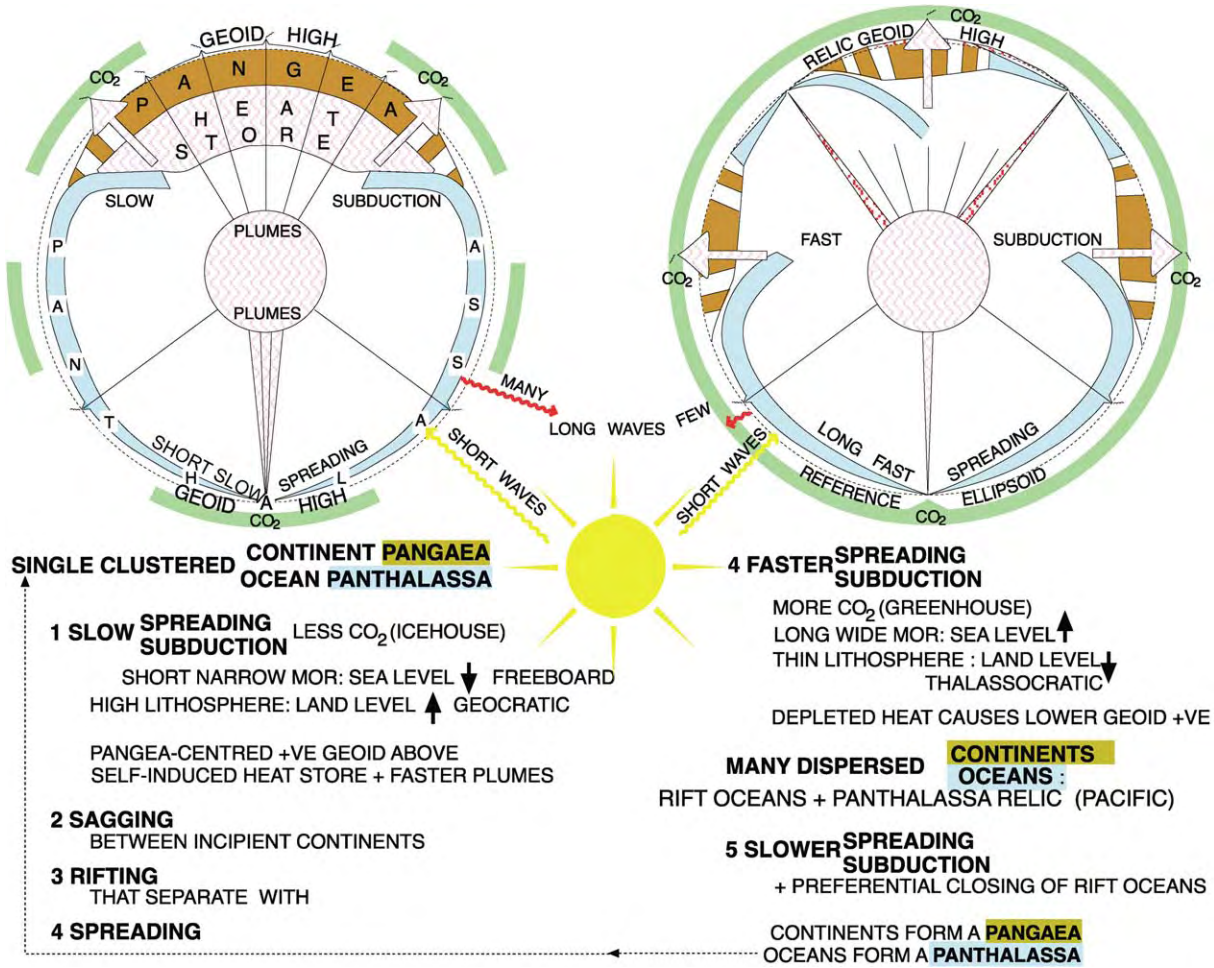


Figure 12 Diagrammatic model of the single continent (Pangaea) and ocean (Panthalassa) alternating with dispersed continents and oceans through five stages. Reproduced with permission from Veevers JJ (2001) *Atlas of Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

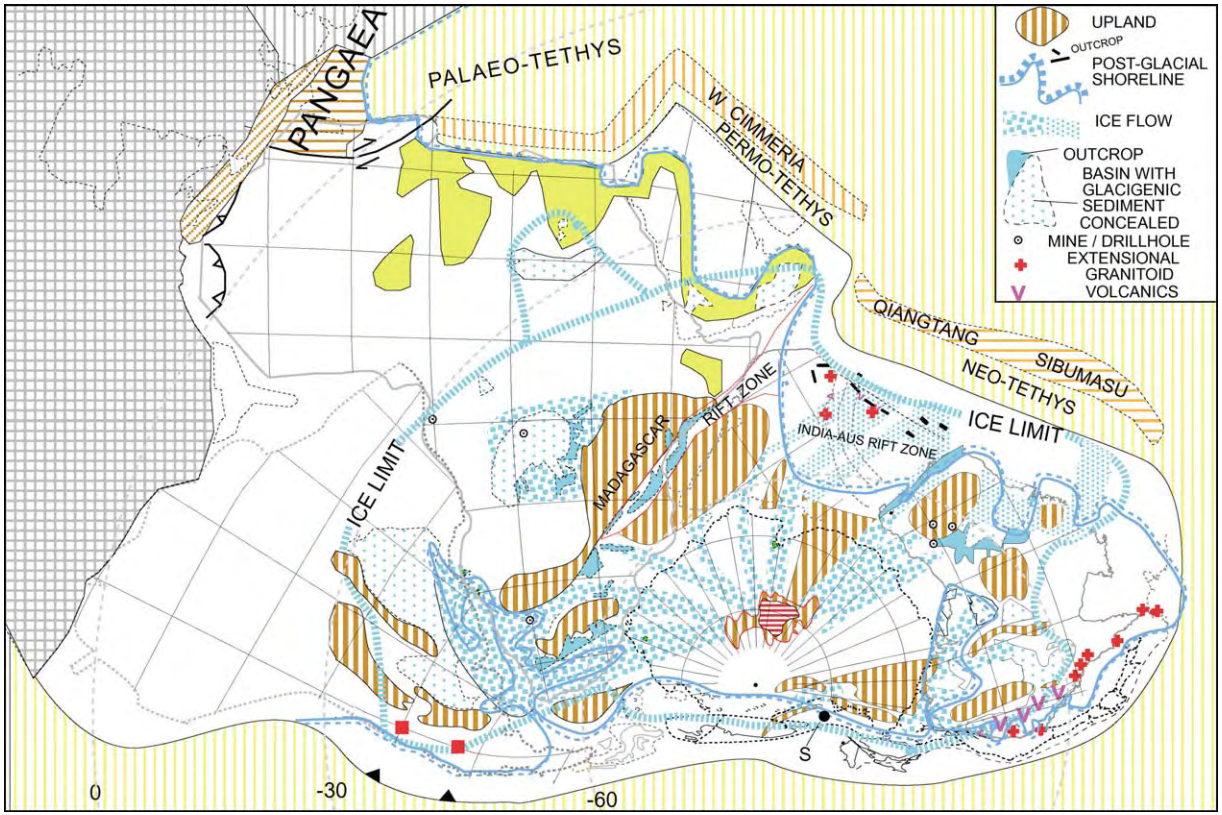


Figure 13 Earliest Permian (302 280 Ma) palaeogeography. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

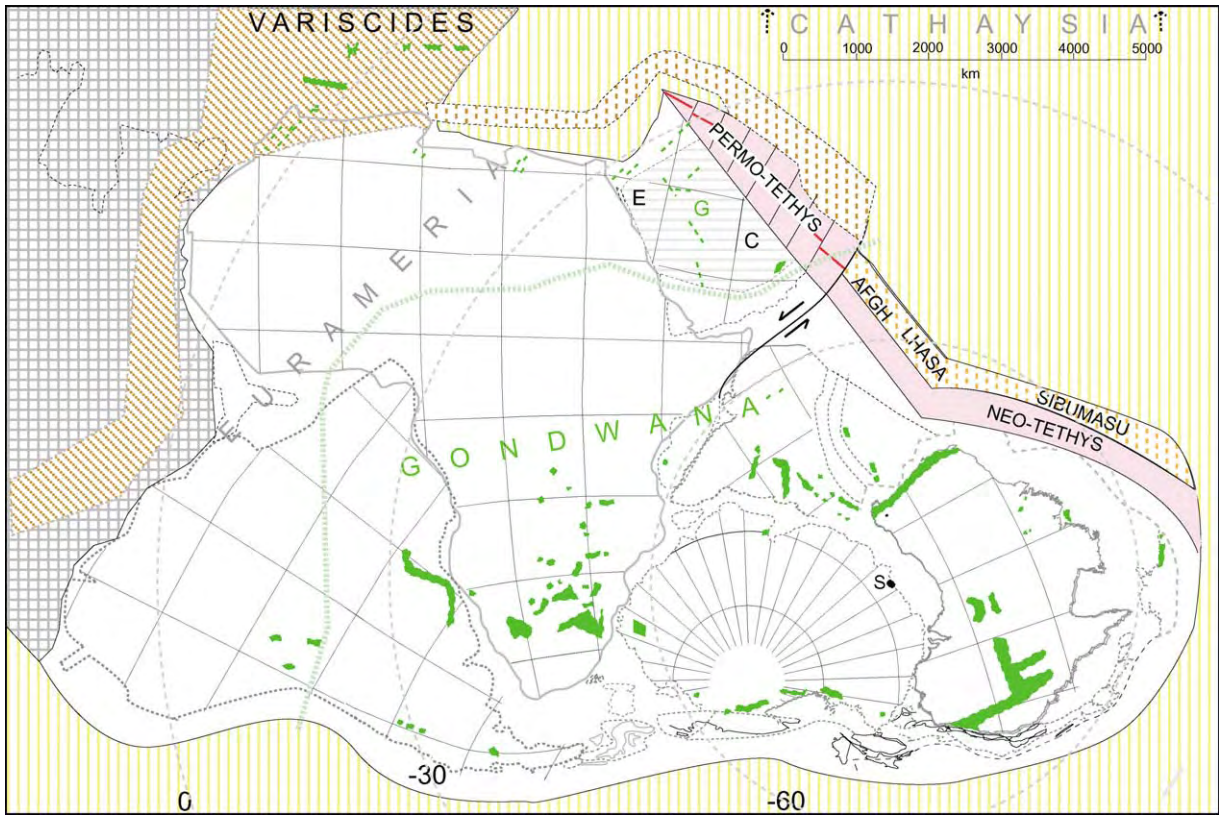


Figure 14 Distribution of Permian coal. The Permo Tethys configuration pertains to the mid Permian. E, Eurameria; G, Gondwana; C, Cathaysia. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

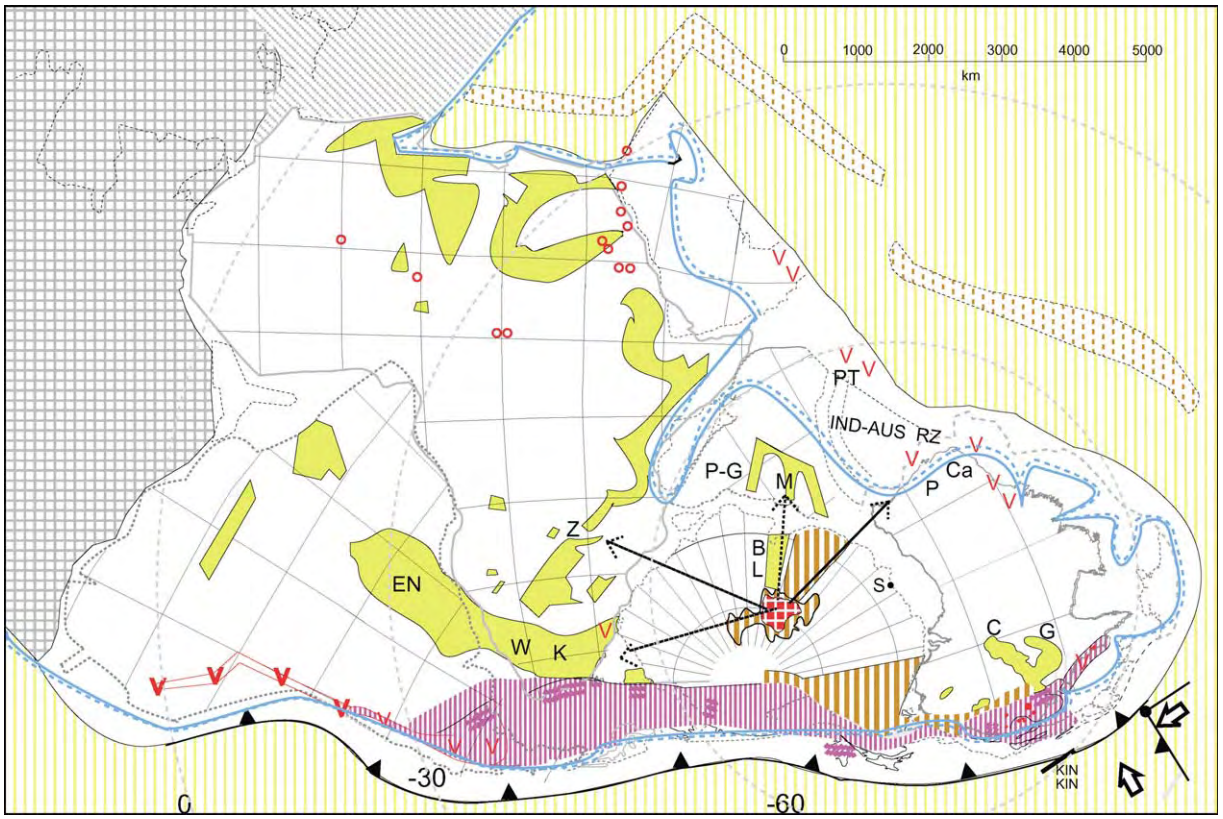


Figure 15 Late Permian (255 250 Ma) palaeogeography: The pink vertical shading indicates the magmatic orogen of Gondwanides I, the red open circles anorogenic igneous complexes, the red square pattern the central Gamburtsev upland. K, Karoo foreland basin; W, Waterford Formation; EN, Estrada Nova Formation; G, Galilee Basin; C, Cooper Basin; P, Perth Basin; Ca, Carnarvon Basin; BL, Beaver Lake Basin; M, Mahanadi Basin; P G, Pranhita Godavari Basin; Z, Zambezi Basin; PT, Panjal Traps, IND AUS RZ, India Australia Rift Zone. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

the epicratonic Galilee and Cooper Basins, and the Kin Kin Terrane docked. The Gamburtsev upland continued to send sediment into the Perth and Carnarvon Basins, through Beaver Lake to the Mahanadi Basin, and into the Pranhita-Godavari, Zambezi, and Karoo Basins. Broad fluviolacustrine basins covered North Africa. Volcanics, including the Panjal Traps, erupted in the India–Australia Rift Zone and in Arabia. Anorogenic complexes erupted in North Africa.

Early Triassic (250–235 Ma)

The end of the Permian saw a much lowered sea-level, and *Glossopteris* vanished as part of the end-Permian extinction event (*see* **Palaeozoic:** End Permian Extinctions). The Triassic saw global warming and the deposition of redbeds. The tectonic situation was unchanged. The Triassic sea returned to its former level, and nonmarine deposition resumed in most of the previous areas (Figure 16).

Mid-Triassic (234–227 Ma)

In the mid-Triassic the Panthalassan margin was terminally deformed (Gondwanides II). Deposition continued in South America and in North Africa-Arabia, which was intruded by anorogenic magmas (Figure 17).

Late Triassic (227–206 Ma)

During the late Triassic, the platform relaxed in Pangaean stage 3 (Extension II), and, following the Early and Middle Triassic coal gap, plant diversity and peat thickness recovered. Coal measures found accommodation space in Gondwanides II as the Molteno Coal Measures, the Lashly Formation, the Topfer Coal Measures, the New Town Coal Measures of Tasmania, and the Ipswich Coal Measures of Queensland (Figure 18). On the craton, the Leigh Creek Coal Measures were deposited on Neoproterozoic basement, and the carbonaceous sediment of the Peera Peera Formation filled the initial Eromanga Basin. Carbonaceous material is found in the McKelvey Member and the Dubrajpur and Colorados formations. Northern South America was crossed by volcanics, and North Africa was dotted with anorogenic magmas.

An arm of the sea crossed North America–Africa to produce evaporites; along its edge, grabens (Extension II) from Texas to Nova Scotia and Morocco filled with sediment that contained coal, as in the Productive Coal Measures of Maryland and Virginia.

Early Jurassic (200–184 Ma)

The vast 200 Ma Central Atlantic magmatic province of tholeiitic flows, dykes, and sills preceded the 190–180 Ma breakup of Pangaea (stage 4; spreading of intra-Pangaean rift oceans) by seafloor spreading (Figure 19). Another vast province of tholeiitic flows and sills was erupted between 184 Ma and 179 Ma in the back-arc region between southern Africa and south-eastern Australia, and southernmost South America was covered by felsic volcanics that have been dated at 187 Ma and younger.

End-Jurassic (145 Ma)

By the end of the Jurassic, the Central Atlantic had reached a width of 1000 km and was continuous through the Straits of Gibraltar with Neo-Tethys and its complex of marginal basins (Figure 20). The north-eastern Indian Ocean had opened at 156 Ma by seafloor spreading, wedging Argo Land off Australia, and the western Indian Ocean had begun to open at 150 Ma by seafloor spreading in the Natal and Somali basins.

The Chon Aike volcanics and a granite in Chile provide evidence of continuing subduction. The Antarctic margin was rifted in the Explora Wedge and Byrd Subglacial Mountains. Scattered magmatism continued in North Africa–Arabia. Nonmarine sediment was deposited in large areas of South America and Africa, on the north-western margin of India, in the rifts at the triple junction between India, Antarctica, and Australia, and in north-west and eastern Australia.

Mid-Cretaceous, Albian–Cenomanian (*ca.* 100 Ma)

Mid-Cretaceous shorelines, nonmarine basins, and igneous rocks are shown here on a pre-breakup base to maintain the same scale as previous figures; the space occupied by the oceans is denoted by a red broken line except between Antarctica and Australia, which were about to break up (Figure 21).

By the mid-Cretaceous, Gondwanaland had split into four pieces, with one of these about to split into two. Earth had entered the state of dispersed continents. The 94 Ma end-Cenomanian shoreline made its maximum penetration into North Africa and northern South America and its minimum into Australia. Earlier, in the Aptian (115 Ma), two-fifths of Australia had been covered by an epeiric sea; in the Cenomanian, the shoreline retreated to the present coast and beyond because uplift outpaced the eustatically rising sea; other continents sank passively beneath the rising sea. Australia's behaviour was caused by a

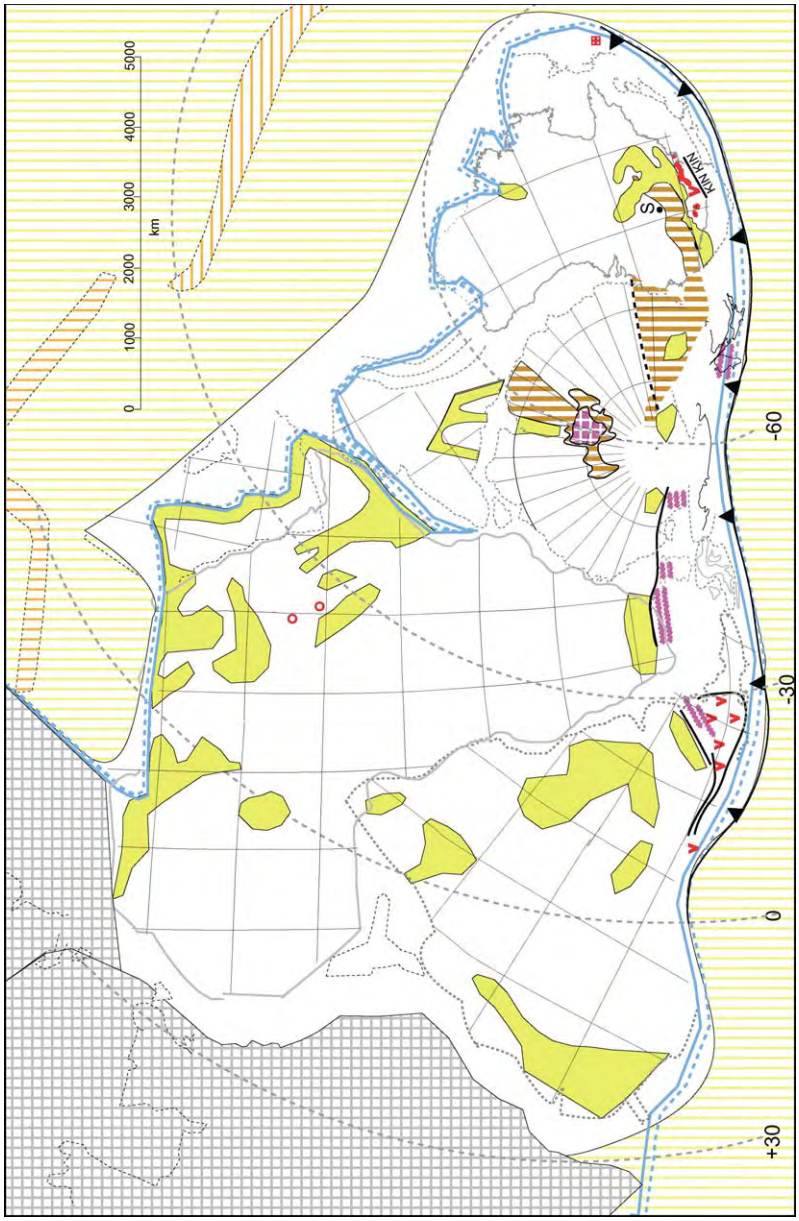


Figure 16 Early Middle Triassic (250–234 Ma) palaeogeography. The orange vertical shading indicates uplands, the purplish wiggly lines zones of deformation, and red dots granitic plutons. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650–500 Ma assembly through 320 Ma merger in Pangaea to 185–100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

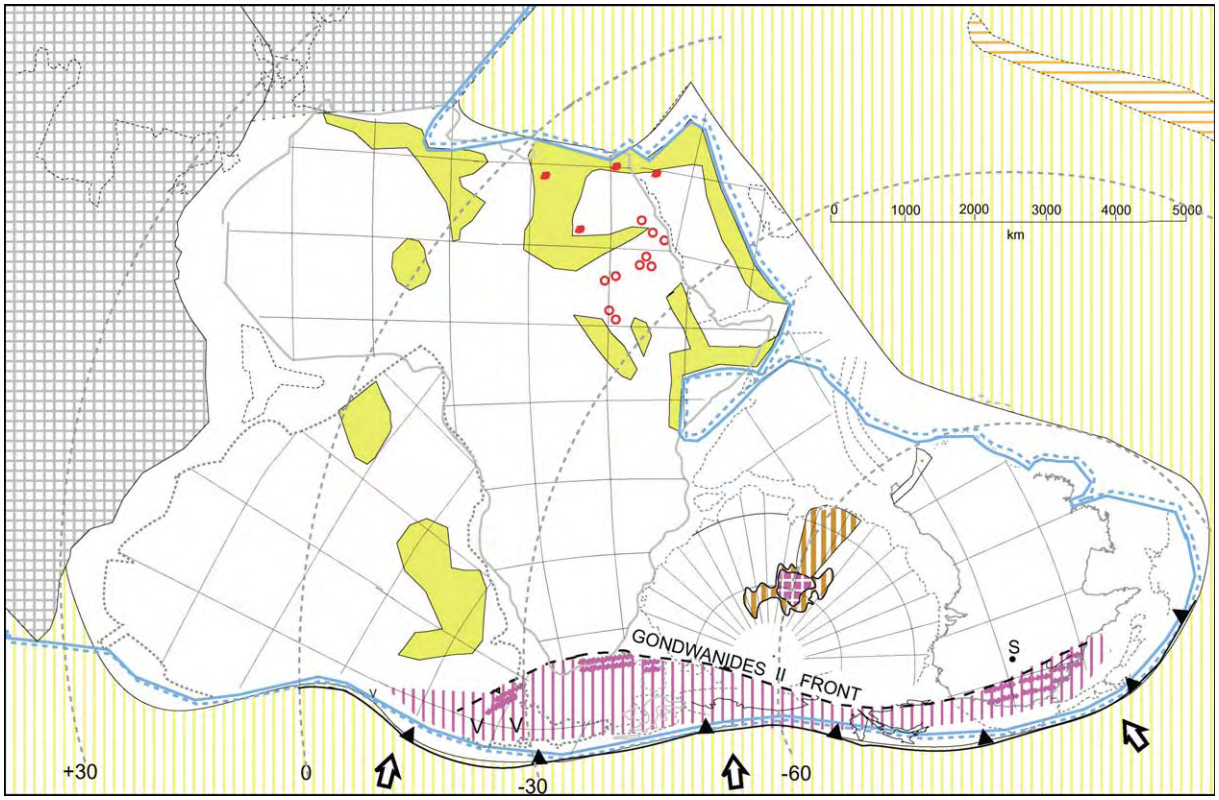


Figure 17 Mid Triassic (234 227 Ma) palaeogeography. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

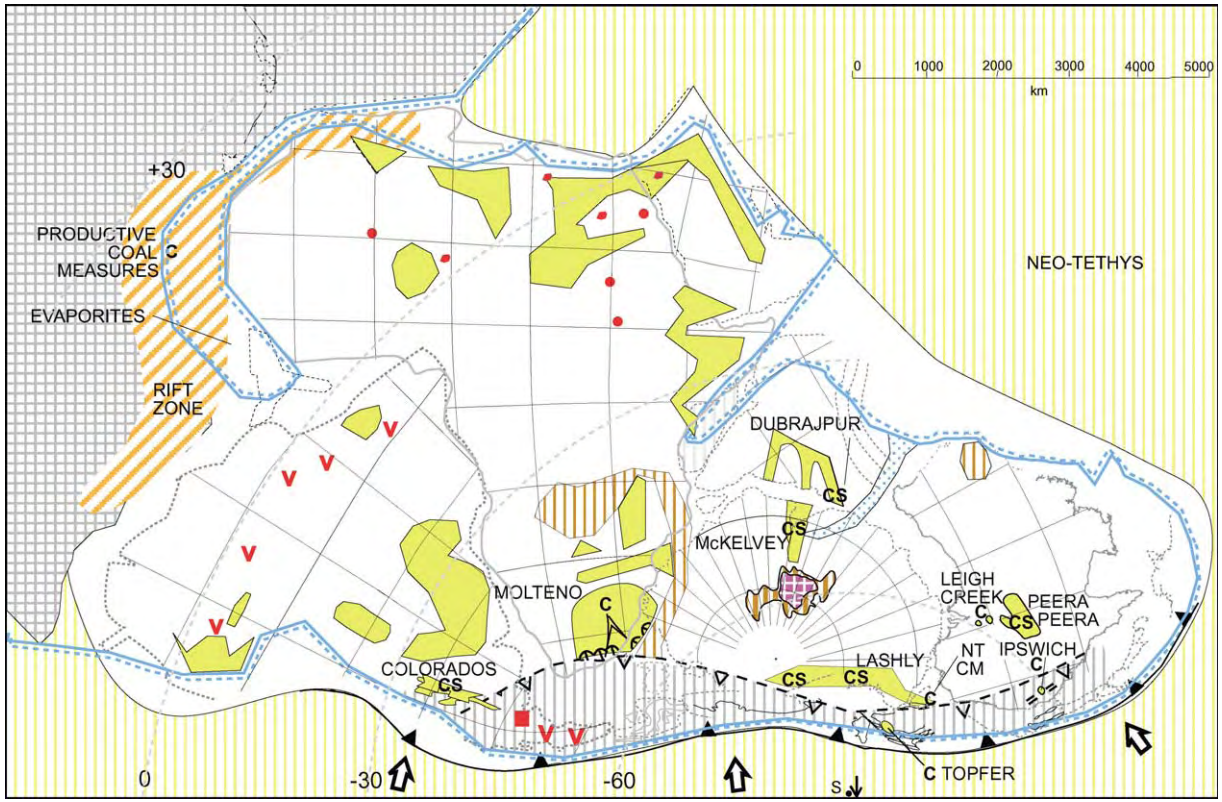


Figure 18 Late Triassic (215 Ma) palaeogeography: C, coal; CS, carbonaceous sediment; NTCM, New Town Coal Measures. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

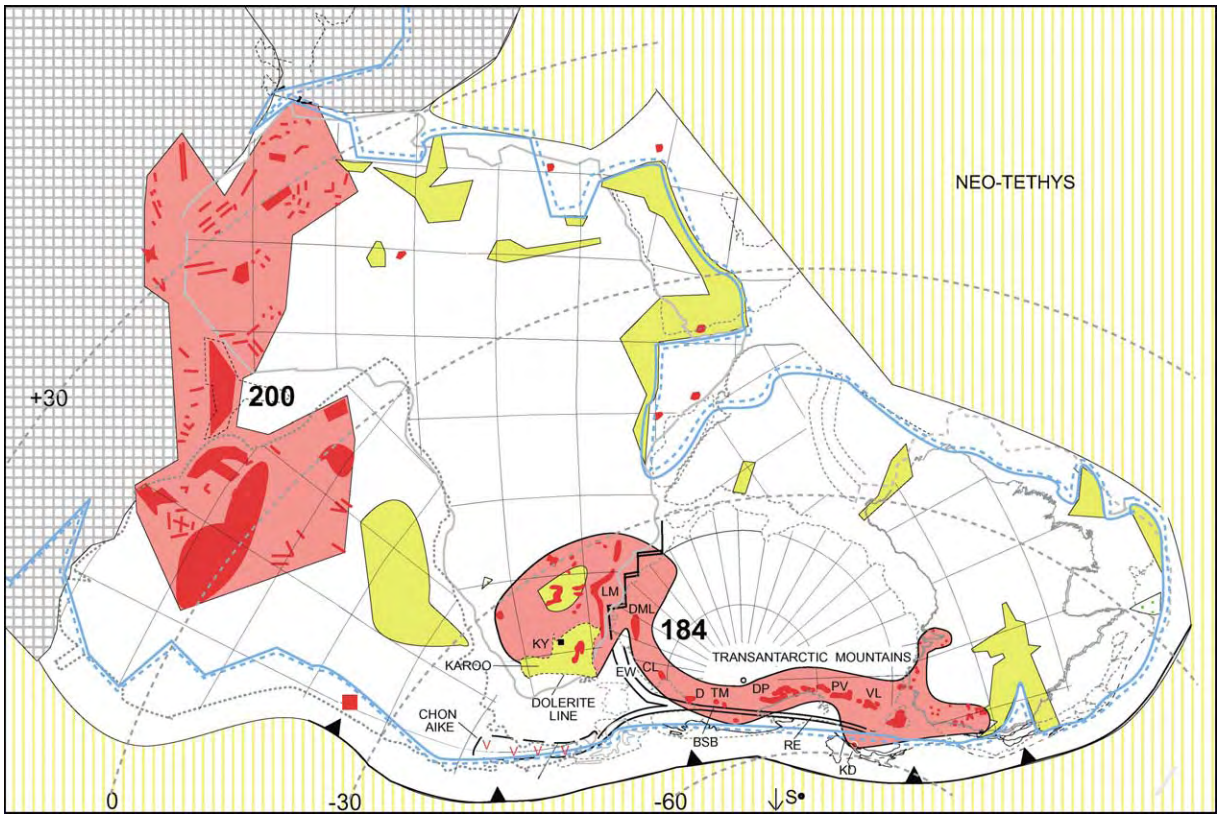


Figure 19 Early Jurassic (200 184 Ma) palaeogeography (ages are given in Ma). Areas covered by tholeiitic flows and sills (shown in pink, with outcrops in red) include the 200 Ma Central Atlantic magmatic province and another behind the Panthalassan margin that stretches from the Karoo through the Transantarctic Mountains, Tasmania, and New Zealand, to south east Australia; individual localities, from left to right, are KY, Kimberley; LM, Lebombo monocline; DML, Dronning Maud Land; EW, Explora Wedge; CL, Coats Land; D, Dufek intrusion; BSB, Byrd Subglacial Mountains; TM, Thiel Mountains; DP, Dawson Peak; RE, Ross Embayment; PV, Pearse Valley; VL, Victoria Land; KD, Kirwans Dolerite; orange upland; lime green nonmarine fringe behind shoreline. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

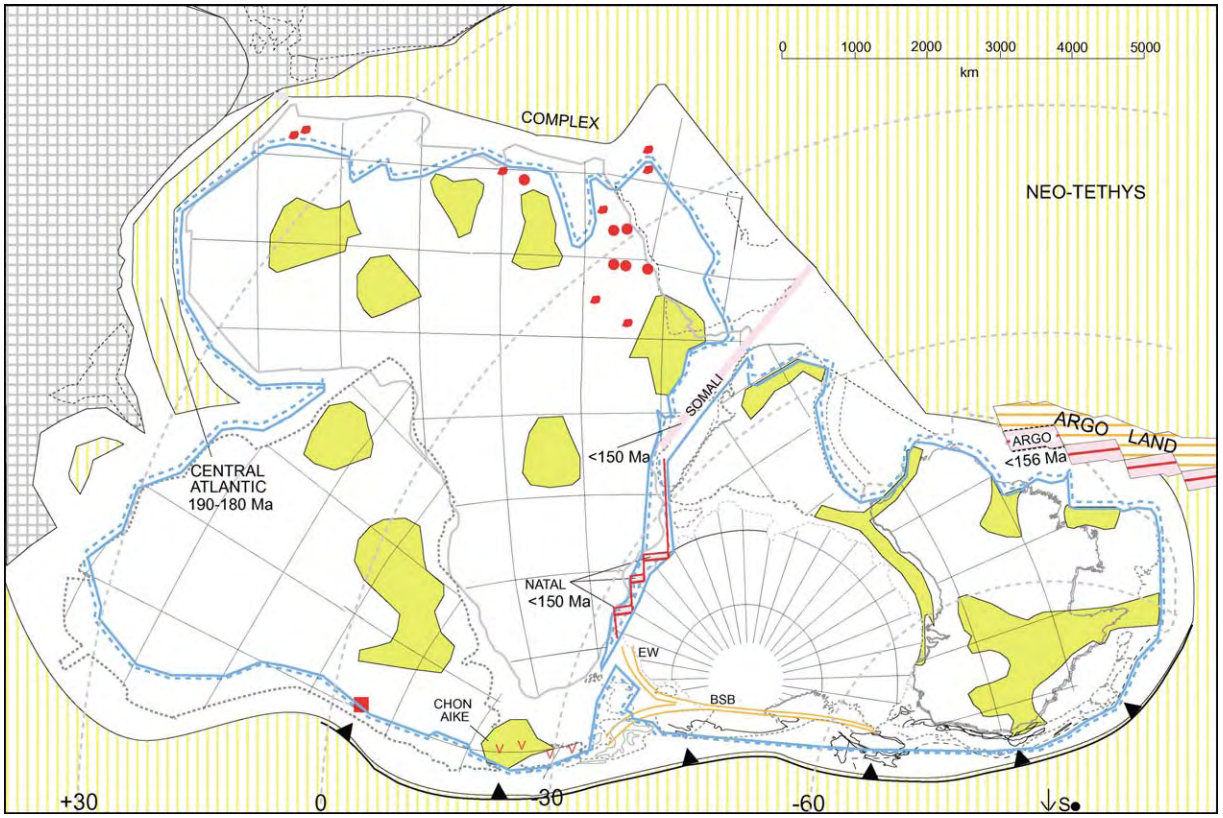


Figure 20 Late Jurassic (150 Ma) palaeogeography: EW, Explora Wedge; BSB, Byrd Subglacial Mountains. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

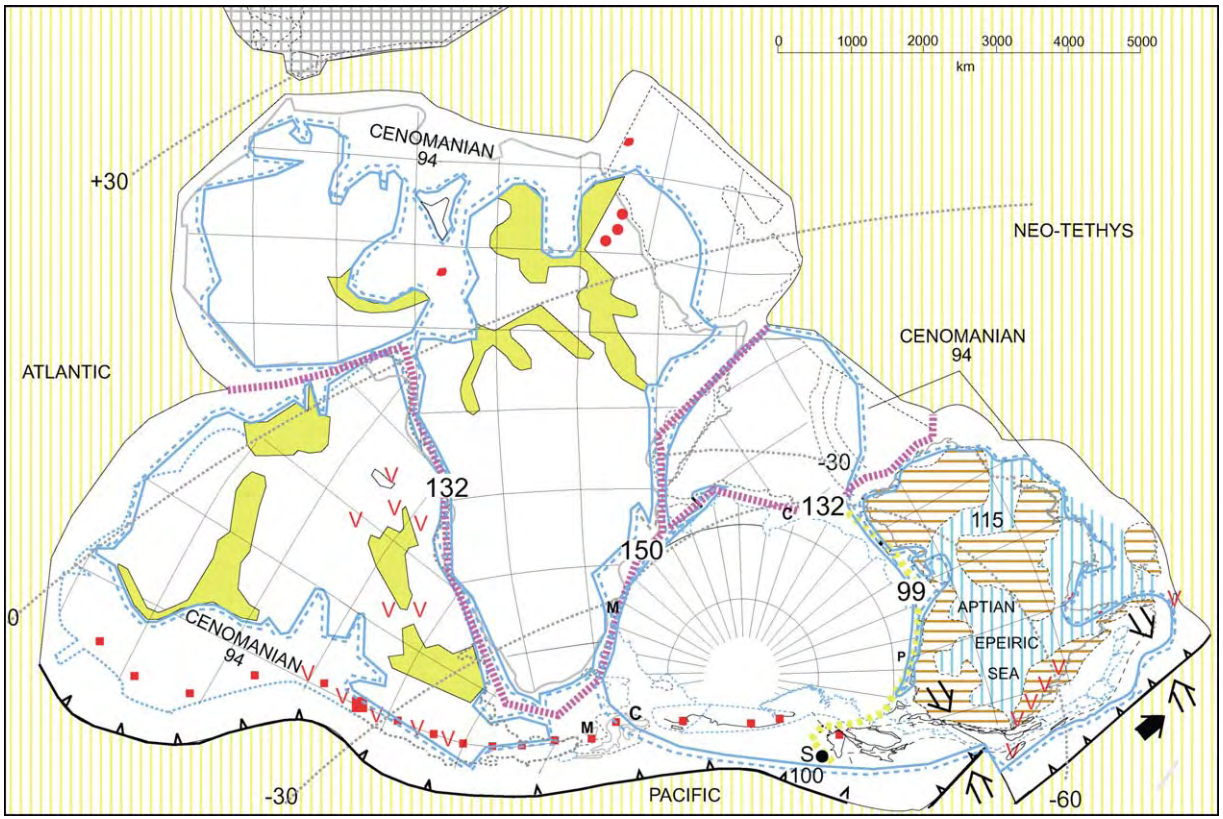


Figure 21 Mid Cretaceous (*ca.* 100 Ma) palaeogeography of the individual continents, which are now dispersed, as shown in Figure 2, but which are shown together for cartographic convenience (the space occupied by the oceans is denoted by a red broken line). Ages of breakup are given in Ma. Also shown is the Aptian epeiric sea on Australia between land (blue pattern). Palaeolatitudes pertain to the conjoined Australia Antarctica, and to the separate South America, Africa, and India. C, coal; P, nonmarine sediment with Aptian palynomorphs; M, marine deposits. Reproduced with permission from Veevers JJ (2004) Gondwanaland from 650 500 Ma assembly through 320 Ma merger in Pangaea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews*

change in the vector of the Pacific Plate with respect to Australia, from head-on collision in the Aptian to transcurrence or side-swipe in the Cenomanian, with concomitant cessation of subduction-related volcanism and uplift.

In Antarctica, coal is found on Alexander Island and in Prydz Bay; nonmarine sediment with Aptian palynomorphs is found in offshore Antarctica, with marine deposits on the opposite side of the continent. The Pacific arc glowed. Magmatic activity elsewhere was confined to volcanism in Brazil and anorogenic intrusion in North Africa–Arabia.

Cenozoic Aftermath (65–0 Ma)

During their dispersal (Figure 2), the Gondwanaland continents made new connections: South America with North America, India and Africa–Arabia with Eurasia, and Australia (almost) with south-east Asia. Antarctica alone remained unattached. Microcontinents (terranes) from the northern margin of Gondwanaland crossed successive generations of Tethys to lodge in Eurasia; New Zealand and New Caledonia remained isolated in the southern Pacific Ocean.

See Also

Africa: Pan-African Orogeny; North African Phanerozoic. **Antarctic**. **Argentina**. **Australia:** Phanerozoic; Tasman Orogenic Belt. **Brazil**. **New Zealand**. **Palaeozoic:** End Permian Extinctions. **Pangaea**.

Further Reading

Crowell JC (1999) *Pre Mesozoic Ice Ages: Their Bearing on Understanding the Climate System*. Geological Society of America Memoir 192. Boulder: Geological Society of America.

Franca AB, Milani EJ, Schneider RL, *et al.* (1995) Phanerozoic correlation in southern South America. In: Tankard AJ, Suarez Soruco R, and Welsink HJ (eds.) *Petroleum Basins of South America,* pp. 129 161. American Association of Petroleum Geologists Memoir 62. Tulsa: American Association of Petrol eum Geologists.

Gaetani M (ed.) (2003) Peri Tethys Programme. *Palaeogeog raphy, Palaeoclimatology, Palaeoecology* 196: 1 263.

Nance RD and Thompson MD (eds.) (1996) *Avalonian and Related Peri Gondwanan Terranes of the Circum North Atlantic*. Special Paper 304. Boulder: Geological Society of America.

Schandelmeier H and Reynolds P O (eds.) (1997) *Palaeo geographic Palaeotectonic Atlas of North Eastern Africa, Arabia, and Adjacent Areas*. Rotterdam: Balkema.

Selley RC (1997) The sedimentary basins of northwest Africa: stratigraphy and sedimentation. In: Selley RC (ed.) *African Basins,* pp. 3 16. Amsterdam: Elsevier.

Stampfli G, Borel G, Cavazza W, Mosar J, and Ziegler PA (eds.) (2001) *The Paleotectonic Atlas of the PeriTethyan Domain*. CD ROM. Katlenburg Lindau: European Geophysical Society.

Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Veevers JJ and Powell CMcA (eds.) (1994) *Permian Triassic basins and foldbelts along the Panthalassan margin of Gondwanaland*. Geological Society of America Memoir 187. Boulder: Geological Society of America.

Veevers JJ and Tewari RC (1995) *Gondwana Master Basin of Peninsular India between Tethys and the interior of the Gondwanaland Province of Pangea*. Geological Society of America Memoir 187. Boulder: Geological Society of America.

Veevers JJ (2001) *Atlas of Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press.

Veevers JJ (2004) Gondwanaland from 650 500 Ma assem bly through 320 Ma merger in Pangea to 185 100 Ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. *Earth Science Reviews* [submitted November 2003].

GRANITE

See **IGNEOUS ROCKS: Granite**

GRENVILLIAN OROGENY

R P Tollo, George Washington University, Washington, DC, USA

Introduction

Orogenies are defined by extended periods of mountain building, usually resulting from convergence of tectonic plates. Such episodes in Earth's history typically involve a series of geological environments that reflect changes in the tectonic setting as convergence proceeds. The Grenvillian Orogeny is named after the village of Grenville in Québec, and the term is widely used to refer to a range of Mesoproterozoic tectonic events that occurred between 1.3 and ~1.0 Ga, resulting in development of a series of orogens that may have stretched across the globe for nearly 10 000 km. Within the Grenville Province of south-eastern Canada, which is the most thoroughly studied portion of this composite orogen, this period of orogenesis included (1) an early accretionary stage at 1.3–1.2 Ga, (2) an interval of widespread magmatism at 1.18–1.08 Ga, (3) and a period of continent–continent collision at 1.08–0.98 Ga that was rapidly followed by uplift and exhumation of the orogenic core. Development of the widespread Grenville orogen was the last major tectonic event to affect the Precambrian core of Laurentia, and marked the final stage in assembly of the Mesoproterozoic supercontinent of Rodinia. In North America, this widespread tectonism is recorded by a broad swath of igneous and metamorphic rocks extending 2000 km from the Atlantic coast of southern Labrador to Lake Huron in Canada and the Adirondacks in the United States (Figure 1). The belt of affected rocks continues south-westward for another 1500 km, mostly in the subsurface, to the Mississippi embayment in the United States, reappearing to the west in Texas and Mexico. Igneous and metamorphic rocks of similar age and tectonic affinity also occur in a series of internal and external massifs associated with the Appalachian orogen in the United States (Figure 2). Grenvillian rocks also constitute the Sveconorwegian Province of southern Norway and Sweden and are recognized as inliers within the Caledonides of Northern Ireland, Scotland, and Norway. Recent palaeogeographic reconstructions suggest that fragments of the dismembered orogen are also present in Antarctica, South America, and Australia.

Major geological events such as the Grenvillian Orogeny that result in assembly of supercontinents are relatively rare events in Earth's history, and the causal mechanisms are not yet well understood. Such events may be driven by global-scale geodynamic mechanisms, such as mantle downwelling, or may involve periodic random amalgamation of cratons resulting from subduction of the oceanic lithosphere. Geological studies of orogens and of the orogenies that produce them are typically spurred by economic factors, because orogenic belts contain much of the world's metallic mineral resources. Because orogens and orogenies are manifestations of large-scale Earth processes, enhanced understanding of the geological factors involved in their genesis also provides important evidence bearing on the mechanisms of plate tectonics and the physical evolution of the planet. The Grenvillian Orogeny represents an episode of unusually widespread tectonism that profoundly affected Earth's palaeogeography; like other major orogenies, such as the Permo-Carboniferous Appalachian Orogeny, the Grenvillian Orogeny marked the end of a major era in the geological time-scale of Earth's history.

Definition of the Grenvillian Orogeny

Geologists working in south-eastern Ontario and western Québec recognized in the early part of the twentieth century that the Grenville Province was structurally distinct from the rest of the Canadian Shield. Advances resulting from detailed field mapping, structural studies, and application of isotopic dating techniques in the 1960s and 1970s led to a more comprehensive understanding of the internal geology of the province and the timing of the Grenvillian Orogeny. During this time, researchers recognized that many areas of the orogen preserved tracts of older recycled crust that had not been completely overprinted by the effects of Grenvillian orogenesis. By 1980, geologists had determined that deposition of the Flinton Group, a Precambrian succession of metamorphosed clastic and carbonate rocks, was constrained to the interval 1080–1050 Ma, and that this deposition occurred during a tectonic hiatus separating two major episodes of orogenesis: (1) a pre-1080-Ma period of arc-related magmatism, uplift, and erosion and (2) a post-1050-Ma period of widespread regional metamorphism (see column A in Figure 3). The earlier episode of orogenesis was referred to as the Elzevirian Orogeny; the later episode is known as the Ottawan Orogeny. Both periods of orogenesis were considered part of the Grenville

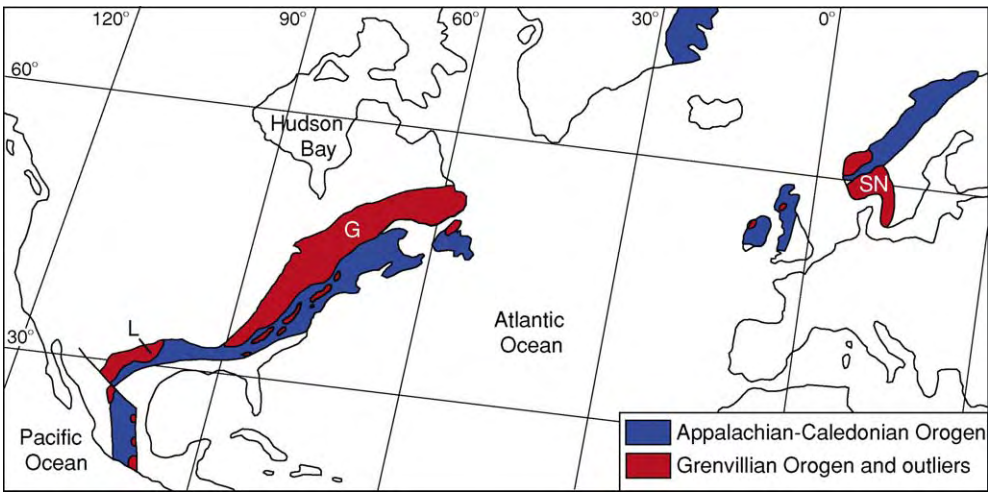


Figure 1 Map showing the location of the Mesoproterozoic Grenvillian Orogen (red) and outliers and the Palaeozoic Appalachian Caledonian Orogen (blue). Abbreviations: G, Grenville Province of Canada; SN, Sveconorwegian Province of southern Norway and Sweden; L, Llano uplift. Modified with permission from Davidson A (1998) An overview of Grenville Province geology, Canadian Shield. In: Lucas SB and St Onge MR (eds.) (coordinators) *Geology of the Precambrian Superior and Grenville Provinces and Precambrian Fossils in North America: Geological Survey of Canada, Geology of Canada, No. 7*, chapt. 3, pp. 205 270. Dartmouth, NS: Geological Survey of Canada.

Orogenic Cycle, which is widely defined as encompassing the interval 1.3–1.0 Ga. Definition of the Grenvillian Orogeny has evolved as new information has become available and as new perspectives have been presented (Figure 3). For example, in the Adirondack massif, where the age brackets on orogenesis closely resemble limits originally defined in the Canadian Province, the time interval between Elzevirian and Ottawan orogenesis did not involve widespread deposition of sediments; instead, this time interval was dominated by intrusion of large volumes of igneous rocks that collectively define the anorthosite–mangerite–charnockite–granite (AMCG) suite (see column B in Figure 3). Such rocks of similar age also occur in the Canadian Province. More recently, researchers recognized that evidence of Elzevirian orogenesis is limited to only a relatively small part of the Canadian Grenville Province and is indeed absent from much of the remainder of the Grenvillian orogenic belt. As a result, and because post-1080-Ma orogenesis ('Ottawan Orogeny' of the original definitions) is now understood to have been primarily responsible for the present geological configuration of the Grenville Province, this later period is referred to by many geologists as the 'Grenvillian Orogeny', although the time-frame for this activity and the existence of pulses of crustal shortening remain subjects of debate (compare columns C and D in Figure 3).

The evolution in the availability and interpretation of data has led to the term 'Grenvillian Orogeny', conveying different meanings to different geologists. To some, it refers to the *ca.* 1.3- to 1.0-Ga period of multistage orogenesis that ultimately resulted in amalgamation of the Grenville Province and creation of Rodinia; to others, it refers only to the culminating (typically post-1080 Ma) events of this large-scale process. In the following discussion of the tectonic events associated with creation of the Grenville Orogen, the 1.3- to 1.0-Ga interpretation is used because it is necessary to describe the sequence of convergent tectonic events that took place during closure of the ocean that bordered this part of Rodinia and to document more fully the nature of rocks constituting the orogenic belt. In this usage, the Grenvillian Orogeny (or Grenville Orogenic Cycle) includes two or more major periods of orogenic activity that have been given status as both 'pulses' and 'orogenies' by various authors (Figure 3). The terms 'Grenvillian Orogeny' (*sensu lato*) and 'Grenville Orogenic Cycle' are widely used in the geological literature to encompass the 1.3- to 1.0-Ga time-span of multiphase orogenic events. However, use of the latter term should be regarded with caution because orogenies are now understood to be lineally progressive phenomena in which crust evolves geologically from one state to another. In this way, orogenies differ from the more cyclic evolutionary history of ocean basins, which may involve successive periods of opening, closure, and reopening. Although unidirectional in modifying crust, orogenic tectonic activity is nevertheless typically diachronous along the length of an orogen, as illustrated by geological evidence from both

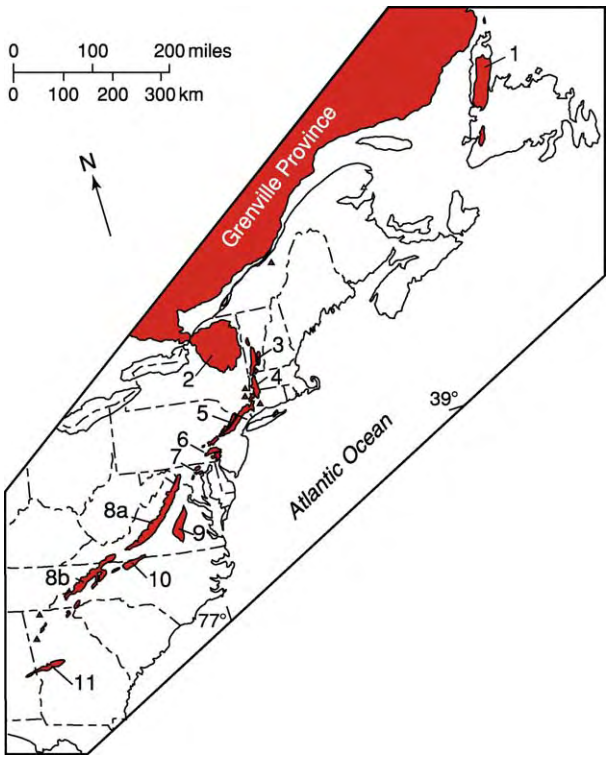


Figure 2 Map showing locations of major occurrences of the Grenville age basement (shaded areas) in eastern North America, including inliers within the Appalachian orogen. 1, Long Range massif; 2, Adirondack massif; 3, Green Mountains massif; 4, Berkshire massif; 5, Reading Prong New Jersey Highlands Hudson Highlands; 6, Honey Brook upland; 7, Baltimore Gneiss antiforms; 8, Blue Ridge province, including Shenandoah (8a) and French Broad (8b) massifs; 9, Goochland terrane; 10, Sauratown Mountains anticlinorium; 11, Pine Mountain belt. Modified with permission from Rankin DW, Drake AA Jr., and Ratcliffe NM (1990) Geologic map of the U.S. Appalachians showing the Laurentian margin and Taconic orogen. In: Hatcher RD Jr., Thomas WA, and Viele GW (eds.) *The Appalachian Ouachita Orogen in the United States*: *The Geology of North America*, vol. F 2, plate 2. Boulder, CO: Geological Society of America.

the Palaeozoic Appalachian and Mesoproterozoic Grenville belts.

Tectonic Evolution of the Grenvillian Orogeny

The most complete geological record of the Grenvillian Orogeny is preserved in the Grenville Province of Canada and the adjacent Adirondack massif in New York (Figure 2). Tectonic events in this region occurring within the 1.3- to 1.0-Ga timeframe record subduction of oceanic crust, closing of an ocean basin, accretion of magmatic arcs (Elzevirian Orogeny), sedimentation (Flinton Group and related metasedimentary successions) and AMCG magmatism, Himalayan-style continent–continent collision (Ottawan Orogeny), and uplift of the orogenic belt.

Elzevirian Orogeny

The culminating sequence of tectonic events that resulted in closure of an ocean bordering proto-North America and, ultimately, in creation of the Grenville orogen in its present form began between 1250 and 1230 Ma with a period of orogenesis termed the 'Elzevirian Orogeny'. (In this article, 'proto-North America' refers to the cratonic core of the eventual supercontinent Rodinia, prior to accretion of the Grenville orogen. 'Laurentia' came into existence following addition of the Grenville orogen.) Evidence for this orogeny is limited in Canada to the south-western portion of the province, and is unequivocally present in the United States only in the vicinity of the Llano Uplift in Texas (Figure 1). The Elzevirian Orogeny was preceded by subduction of oceanic crust and development of associated, possibly extensive, magmatic arcs that constitute the Composite Arc Belt in Canada and parts of the Adirondack Highlands (Figure 4) and Green Mountains in the United States (Figure 2). The Composite Arc Belt (Figure 4), where the most diverse assemblage of metamorphosed arc-related lithologies is recognized, consists mostly of <1300-Ma volcanic, plutonic, volcaniclastic, carbonate, and siliciclastic rocks derived from arcs and marginal basins. The inferred origin of many of these metamorphic lithologies as arc-related igneous rocks is based on the unique assemblage of lithologic types and chemical similarities between these rocks and calc-alkaline andesitic to tonalitic suites from modern arc terranes such as the western Americas, Aleutians, and Japan.

Prior to amalgamation with proto-North America, a composite Elzevirian volcanic-arc assemblage was separated from the main proto-North American craton by a back-arc basin and bordered by a subduction zone that dipped beneath the arc terrane and was responsible for production of the characteristic calc-alkaline magmatism (Figure 5A). This tectonic scenario bears similarities to the present Taiwan region in the western Pacific. The cessation of arc and back-arc magmatism at *ca.* 1230 Ma is interpreted to mark the termination of local subduction of oceanic lithosphere, an event that was followed by the Elzevirian Orogeny, which, by most recent estimates, occurred at 1230–1180 Ma, and is associated in the Canadian region with closure of the back-arc basin and accretion of the outboard Frontenac–Adirondack–Green Mountains crustal block (Figure 5B). Possible slices of the oceanic crust that previously separated the arc terrane from the continent are present as ophiolites within the accreted arc block. An important

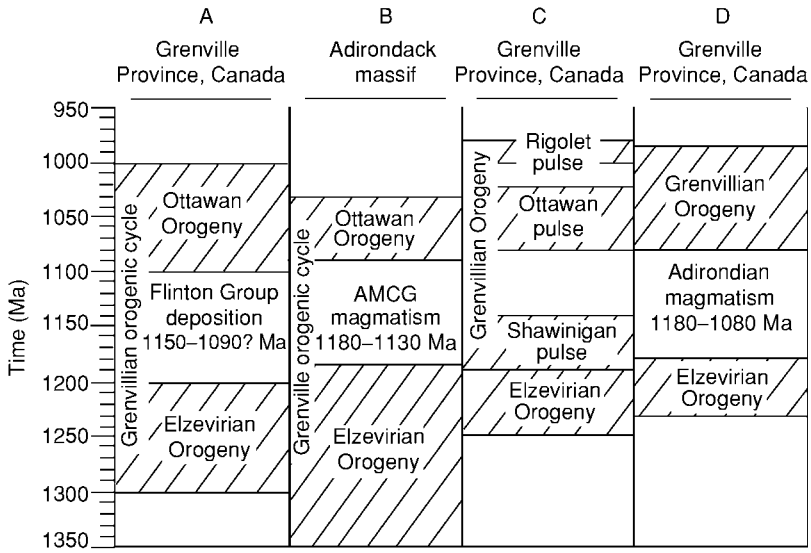


Figure 3 Comparative nomenclature for Late Mesoproterozoic orogenesis in the Grenville Province of Canada and the Adirondack massif, as published in the geological literature from 1980 to 2002. (Nomenclature was derived from the 1980, 1996, 1997, and 2002 work by Moore and Thompson, McLelland and co workers, Rivers, and Gower and Krogh, respectively.) Diagonal pattern indicates time intervals of orogenesis. The various definitions of orogenic periods reflect differences in study areas, types of geological data compiled, and the type and quality of geochronological information available. AMCG, Anorthosite mangerite charnockite granite.

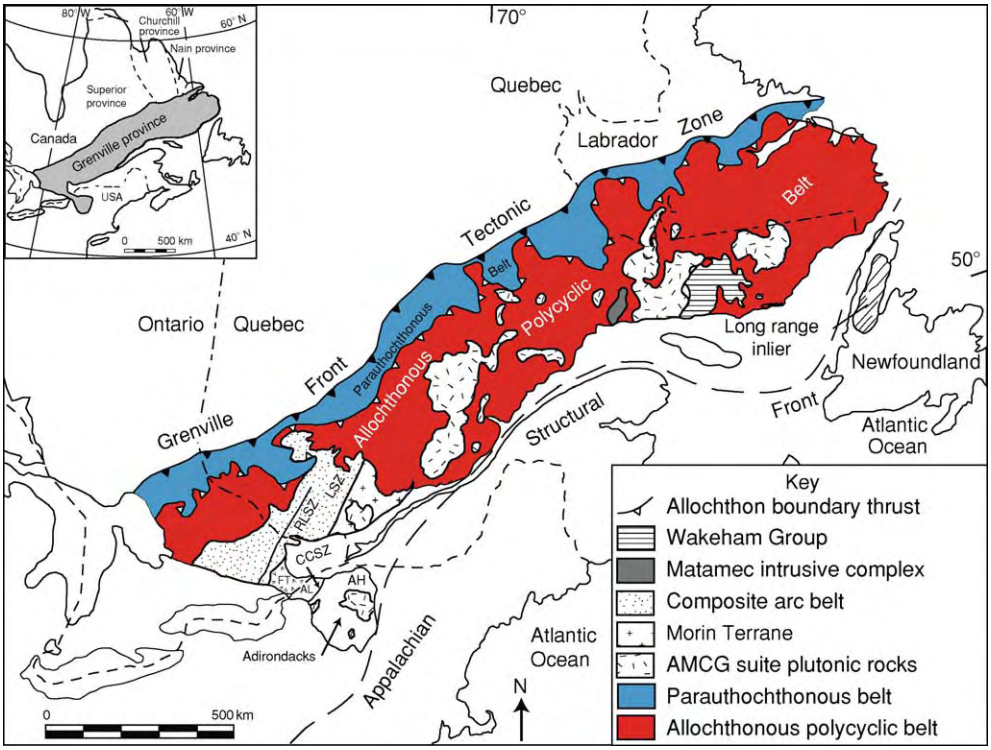


Figure 4 Map of the Grenville Province in Canada. Abbreviations: RLSZ, Robertson Lake shear zone; LSZ, Labelle shear zone; CCSZ, Carthage Colton shear zone; FT, Frontenac Terrane; AL, Adirondack Lowlands; AH, Adirondack Highlands; AMCG, anortho site mangerite charnockite granite. Nomenclature for principal lithotectonic belts of the Canadian Grenville Province adopted from the work of the Carr group and Rivers and co workers. Modified with permission from Davidson A (1998) An overview of Grenville Province geology, Canadian Shield. In: Lucas SB and St Onge MR (coordinators) *Geology of the Precambrian Superior and Grenville Provinces and Precambrian Fossils in North America: Geological Survey of Canada*, *Geology of Canada, No. 7*, chapt. 3, pp. 205 270. Dartmouth, NS: Geological Survey of Canada.

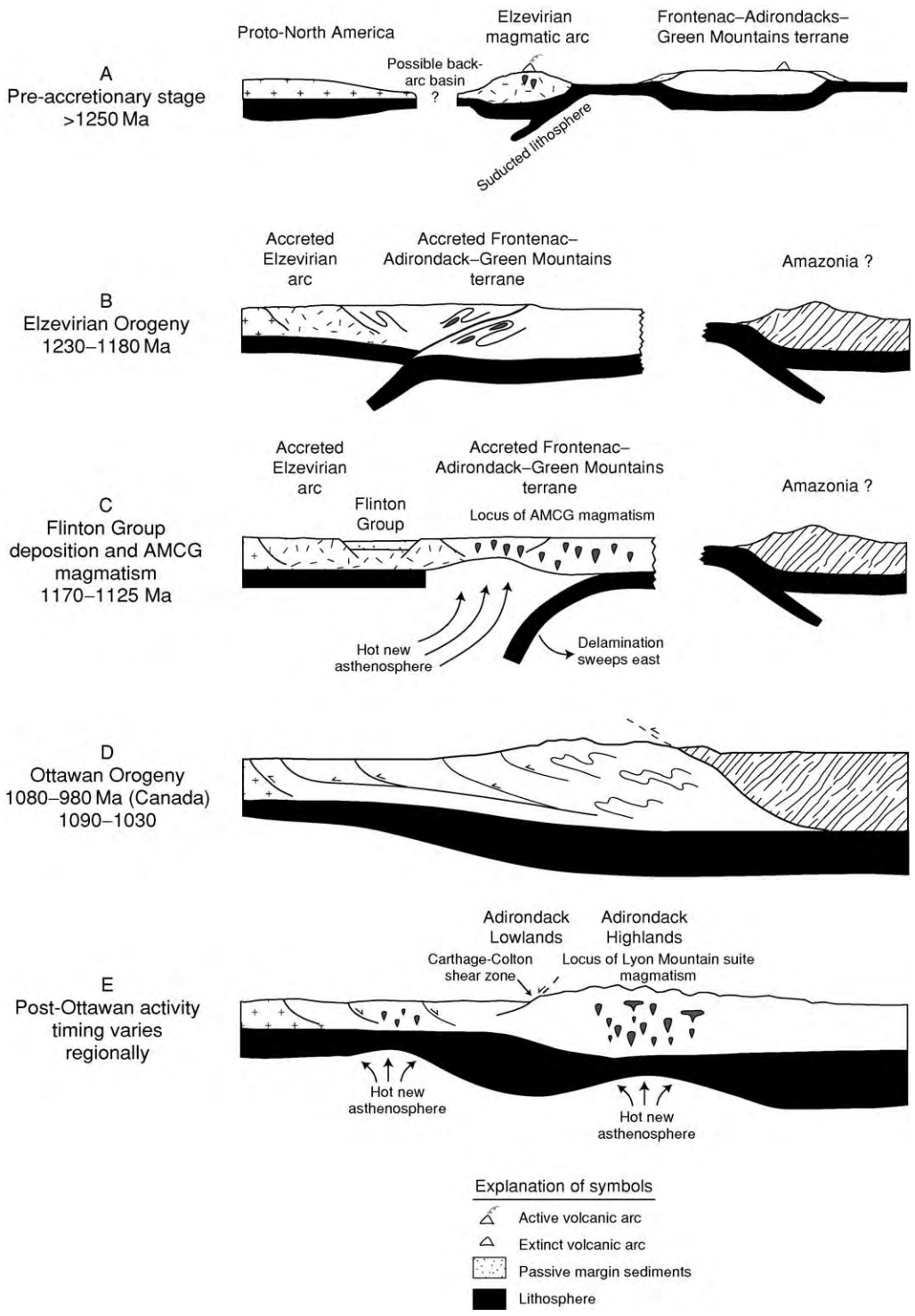


Figure 5 Summary of major events involved in Grenvillian orogenesis during the interval from 1.3 to 1.0 Ga in the southern Grenville Province of Canada and the Adirondack massif. Modified with permission from McLelland J, Daly S, and McLelland JM (1996) The Grenville Orogenic Cycle (*ca.* 1350 1000 Ma): an Adirondack perspective. *Tectonophysics* 265: 1 28.

consequence of closure of the back-arc basin was the shifting of the zone of active subduction outboard of the newly accreted crust and the establishment of a dominantly compressive tectonic setting; these events resulted in crustal thickening through thrust imbrication and the likely formation of a fold–thrust belt along the eastern (present coordinates) portion of the Composite Arc Belt, where rocks were locally

metamorphosed at amphibolite-facies conditions during Elzevirian orogenesis. A similar sequence of arc accretion accompanied by fold-and-thrust tectonism and metamorphism occurred during the Taconian orogenesis that affected much of the Appalachian region during the Ordovician. Such accretion of subduction-related arcs commonly occurs during the early stages in the tectonic evolution of collisional orogens involving closure of oceans. The regional scope of the Elzevirian Orogeny is a matter of debate because geological evidence of similar terrane amalgamation is generally lacking in the eastern Canadian Grenville Province. A possible arc-related assemblage of rocks is recognized in the Mesoproterozoic New Jersey Highlands (Figure 2), but is undated by modern isotopic techniques. The existence of a volcanic arc similar in age to the Elzevir Arc is documented in Texas, where it is believed to have docked with proto-North America at about 1150 Ma, producing polyphase deformation and high-pressure metamorphism within the Llano uplift.

Post-Elzevirian Activity

The period following Elzevirian orogenesis from 1180 to 1080 Ma is marked only by penecontemporaneous local sedimentation and widespread AMCG magmatism. In Canada, three small and isolated sedimentary successions of fluvial to shallow marine origin preserve evidence of crustal extension occurring across a broad area of the south-western Grenville Province. These rocks include the Flinton Group in eastern Ontario (Figure 5C), which is recognized as originally deposited unconformably on older basement rocks of Elzevirian affinity and subsequently isoclinally folded and metamorphosed to varying grade with the basement. Studies of detrital zircons indicate that the Flinton Group rocks, consisting of quartz-pebble conglomerate, pelite, quartzite, siltstone, and carbonate units, were deposited after 1150 Ma in a local fault-bounded basin. Studies of similar rocks exposed north-west of the Flinton Group indicate a likely comparable age of deposition.

The association of crust thickened by orogenic processes, deposition of sediments within encratonic basins, extensional faulting, and emplacement of mantle-derived AMCG igneous rocks led researchers to propose that subcontinental lithosphere beneath the Grenville orogen was probably replaced by asthenosphere during crustal shortening. Such juxtaposition of hot asthenosphere against thinned continental lithosphere would have profound effects throughout the overlying crust, and could have resulted from either lithospheric delamination or convective thinning of the lithosphere. These models produce similar consequences, including (1) establishment of steep temperature gradients in the continental lithosphere, (2) increases in surface elevations, and (3) creation of a thermal anomaly in the extended crust. In either case, magmatic activity such as represented by the Grenvillian AMCG suites is a likely result. In the southern Grenville Province and adjacent Adirondacks, where AMCG rocks are abundant, trends in crystallization ages of plutons indicate that the delamination model is more likely to account for this important period of igneous activity. The combination of overthickened crust and lithosphere resulting from Elzevirian orogenesis, and subsequent relaxation of contractural strain throughout much of the Grenvillian region, established conditions conducive to local delamination of upper mantle lithosphere (Figure 5C). Such delamination, which appears to have been concentrated in the southern Grenville Province–Adirondacks–Green Mountain region, resulted in crustal rebound and local structural collapse, forming isolated, fault-bounded basins in which the Flinton Group and related sedimentary successions were deposited. As delamination proceeded towards the south-east, the direction indicated by crystallization ages of related plutonism, high-temperature asthenosphere rose into the delaminated zone, bringing relatively hot, mafic melt towards the base of the crust. Tholeiitic gabbroic magmas produced by depressurization melting of the rising asthenosphere ponded at the base of the crust and underwent fractional crystallization within a relatively quiescent, nonorogenic environment, resulting in the production of plagioclase-rich crystal mushes through flotation and concomitant sinking of olivine–pyroxene-rich cumulates (Figure 6). Heat from the gabbroic magmas resulted in partial melting of the lower crust, producing relatively anhydrous felsic magmas that intruded to shallow levels to form plutons of syenitic, monzonitic, and granitic composition, which, on crystallization of orthorhombic pyroxene, locally yielded mangerites and charnockites. Ultimately, plagioclase-rich magmas originally produced through the aforementioned crustal melting used crustal fractures and other weaknesses to ascend to the level of the felsic melts, producing anorthosites and the related AMCG suites that are characteristic of the Grenville Province. Some of the anorthositic complexes form very large, petrologically diverse plutonic bodies, such as the Lac St. Jean AMCG suite in the central Grenville Province and the Marcy Massif in the Adirondacks. Further fractional crystallization of the plagioclase melts may have resulted in the formation of mafic-rich residual liquids represented by ferrodiorites (also referred to as ferrogabbros and jotunite), which, through immiscibility, may have in

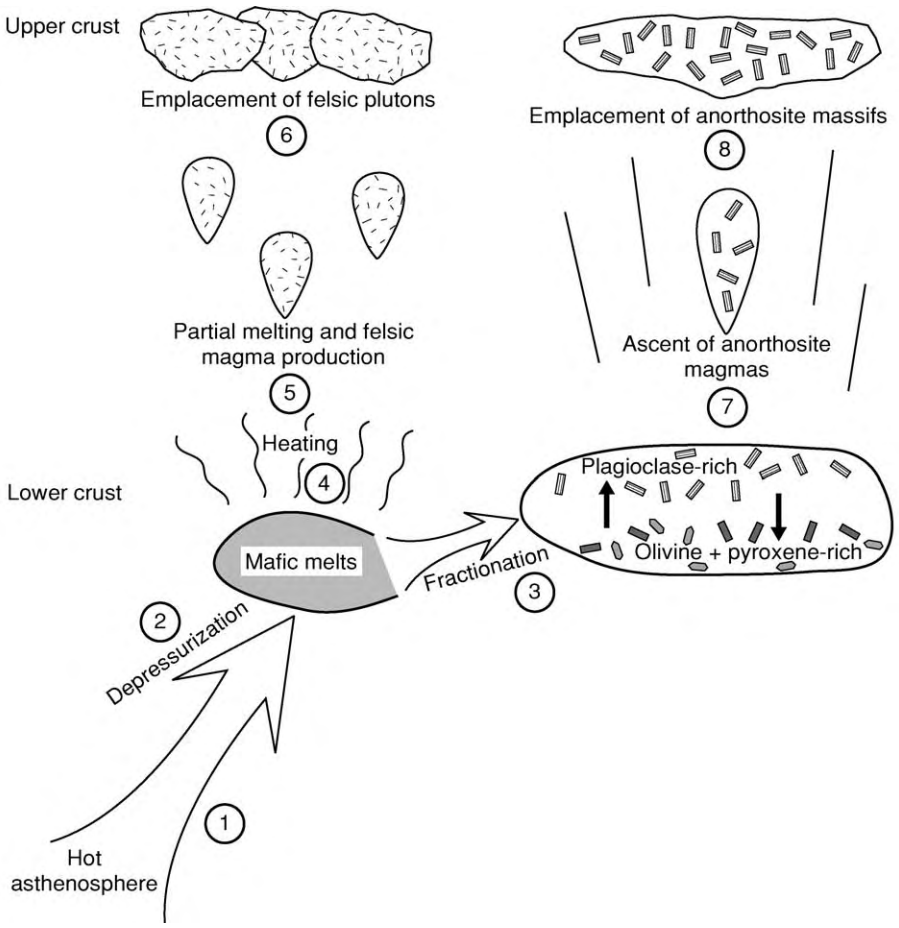


Figure 6 Schematic summary of a possible model for generation of Grenvillian anorthosite mangerite charnockite granite suites. Following either lithospheric delamination or convective thinning of the lithosphere resulting from crustal thickening related to Elzevirian orogenesis, the hot asthenosphere rises towards the base of the intact crust (1), where depressurization reactions result in production of mafic magmas (2). These mafic magmas undergo crystallization under relatively quiescent, nonorogenic conditions, then fractionate to produce plagioclase rich melts, through flotation, and olivine + pyroxene rich melts, through gravitational settling (3). The mantle derived mafic melts also provide significant heat transfer to the lower crust (4); this results in partial melting and production of felsic magmas (5) that rise buoyantly towards emplacement in the upper crust as plutons composed of granite, charnockite (syenogranite containing orthorhombic pyroxene), and mangerite (monzonite containing orthorhombic pyroxene) (6). The plagioclase rich differentiates of the mantle derived gabbroic magma use fractures and other crustal weaknesses to ascend within the crust (7) towards ultimate emplacement as petrologically complex massifs containing anorthosite and related anorthosite mangerite charnockite granite rocks.

turn been responsible for the production of associated ilmenite–magnetite ores.

AMCG magmatism was widespread throughout the south-western Grenville Province and the Adirondacks during the 1180- to 1080-Ma interval, and marked an abrupt change in the nature of plutonic activity within the region. Igneous activity related to the Elzevirian magmatic arc was calc-alkaline and thus characteristic of rocks associated with modern convergent plate margins. In contrast, the AMCG suites are dominantly tholeiitic and are similar to igneous rocks associated with modern within-plate tectonic environments. Such contrasts in composition reflect the differing nature of the sources involved in petrogenesis. These differences underscore the usefulness of geochemical studies of ancient igneous and meta-igneous rocks to decipher both the modes and the environments of origin. Massif anorthosites and associated mangerites and granites have traditionally been considered to constitute an 'anorogenic trinity', the formation of which is indicative of crustal conditions in which orogenic effects are absent. However, considering the entire Grenville Province, it is clear that metamorphism and contractural deformation occurred in some parts of the province at the same time that AMCG suites were emplaced in other parts. Such contrasting styles of tectonic activity should perhaps be expected in an area that is as

large and complex as the Grenville Province is. The preponderance of available data indicate that such suites are a likely consequence of orogenesis and result from deep-seated lithospheric mechanisms that originated in response to previous orogenic processes. Viewed in this light, the Grenvillian AMCG suites represent postorogenic magmatism with origins that relate directly, albeit locally, to Elzevirian orogenesis.

Ottawan Orogeny

Final closure of the ocean bordering proto-North America occurred as a result of continent-to-continent collision beginning at 1080 Ma, involving a series of events referred to as the Ottawan Orogeny (also as the 'Grenvillian Orogeny' by some geologists; see Figure 2). Compressional tectonic forces resulted in telescoping of the continental margin and the formation of large thrust slices and ductile folds with dominant direction of transport towards the north-west (Figure 5D). Multiple lines of evidence indicate that the orogen propagated north-westward, with the youngest faults located in the vicinity of the Grenville Front Tectonic Zone (Figure 4). Metamorphic assemblages in the Grenville Province generally indicate recrystallization at granulite-facies conditions in the hinterland (region located towards the south-east), where large-scale, recumbent nappes were developed and most plutonism occurred. Thrusting was more common in the foreland (region located towards the north-west), where metamorphic conditions were more variable, ranging from greenschist to high-pressure amphibolite facies, with eclogite-facies conditions developed locally. Many of the north-west-directed thrusts were of large scale, including the Allochthon Boundary Thrust (Figure 3) that is associated with high-pressure mineral assemblages. Results from many studies throughout the province indicate that metamorphism and deformation were ongoing across the entire province during the interval from 1080 to 980 Ma. However, disagreement exists regarding the nature of these tectonic processes. Some researchers suggest that metamorphism and deformation occurred in at least three discreet pulses (see column C in Figure 2) separated by intervals of extension. The first two pulses of crustal shortening were concentrated in the hinterland, whereas the latest pulse caused north-westward propagation of the orogen into its foreland. Abundant magmatism, including emplacement of anorthosite complexes, occurred during the intervening extensional periods. A different interpretation of the tectonic evolution of the eastern Grenville Province suggests that orogenesis peaked at different times in different places, and that, as the crust yielded through thrusting in one location, stresses were transferred elsewhere. Thus, instead of separate orogenic pulses, this model proposes that the 'pressure point' responsible for compressional tectonism shifted periodically as the crust yielded locally to the forces of deformation.

The effects of the Ottawan Orogeny were generally more intense in the Adirondack massif, where evidence of ductile structural fabrics and high-grade metamorphism is widespread. Studies of metamorphic rocks indicate that the crust beneath the Adirondack Highlands nearly doubled in thickness as a result of Ottawan orogenesis. Such extreme crustal thickening, considered together with the general lack of Ottawan-age calc-alkaline rocks of possible arc affinity in the Adirondacks and contiguous Grenville Province, constitutes important evidence in support of continent–continent collision as the primary cause of the orogeny. The Adirondack Highlands represent one of the world's classic granulite-facies terranes, where peak metamorphic conditions during the Ottawan Orogeny reached temperatures of 750–800°C at pressures of 6–8 kbar. Metamorphism was associated with recumbent isoclinal folding and development of intense penetrative fabrics that typically overprinted features developed during the Elzevirian orogenesis throughout the Adirondacks (Figure 7A and B). However, the anorthosite complexes in this area, all of which were emplaced at about 1150 Ma, are typically not structurally overprinted (Figure 7C) because these generally anhydrous, dominantly monomineralic plutons acted as rigid bodies that deflected stresses during Ottawan orogenesis. Such kinematics that result in plutonic rocks that appear to be relatively undeformed when, in fact, they predate deformation, must be considered carefully by field geologists interested in deciphering the sequence of local structural events. Field relations and isotopic ages of plutonic rocks indicate that Ottawan orogenesis occurred from 1090 to 1030 Ma in the Adirondacks. The occurrence of thick allochthons and associated high-pressure belts in the Canadian Grenville Province stands in contrast to the dominantly ductile deformation and large recumbent fold nappes of the Adirondacks. Such differences are attributed to variations in style, metamorphic grade, and ductility between the mobile interior of the orogen to the south-east and the foreland to the north-west, as well as to the thermal softening effects imparted on the Adirondacks by intrusion of the Hawkeye granitic suite at 1100 Ma.

Post-Ottawan Activity

Tectonic activity associated with the Ottawan Orogeny did not terminate simultaneously throughout

Figure 7 Deformation features developed during Ottawan orogenesis in the Adirondack Highlands. (A) Refolded isoclinal fold developed in Elzevirian age migmatitic metapelite. Dashed line outlines the fold, which is about 1 m in length. (B) L Tectonite (also referred to as 'straight gneiss'), developed through extreme grain size reduction accompanying ductile deformation in migmatitic metapelite located between the Piseco anticline and Glen Falls syncline in the south eastern Adirondacks. This is the deformed equivalent of the rock shown in A. Pen is oriented parallel to the tectonic transport direction and is 14 cm in length. (C) Inequigranular anorthosite from the southern edge of the Oregon dome in the south eastern Adirondacks. Large blue gray andesine crystals are typical of Grenvillian massif anorthosites. Pen is 14 cm in length.

the Grenville orogen, but rather ended with a series of sporadic events. Within Canada, north-west-directed thrusting and shearing along the Grenville Front Tectonic Zone (Figure 4) occurred as late as 1000–980 Ma, and weakly deformed 980-Ma AMCG plutons document the waning stages of orogenesis in the north-eastern Grenville Province. In contrast, undeformed 975-Ma granitic intrusions that spread across the southern part of the eastern Grenville Province appear to be postkinematic, signalling a likely end to contractional tectonics. Unmetamorphosed and undeformed 1045-Ma fayalite granite, part of the late synkinematic to postkinematic Lyon Mountain granite suite, provides a minimum age for the end of widespread Ottawan activity in the Adirondacks, although localized high-temperature events occurred until about 1015 Ma. Rapid uplift followed Ottawan orogenesis in the Adirondacks and neighbouring Grenville Province, possibly caused by a second cycle of lithospheric delamination (Figure 5E). This uplift was accompanied by extensional collapse, especially along the Carthage–Colton shear zone, where amphibolite-facies rocks of the Adirondack Lowlands were displaced downwards to the north-west to become juxtaposed against the granulite-facies Adirondack Highlands.

Appalachian Inliers

Grenville-age rocks within the Appalachians preserve evidence of geological processes that are grossly similar to those documented in the Canadian Grenville Province and Adirondacks. Rocks of AMCG affinity are present in some massifs, and a major period of metamorphism and deformation that is roughly contemporaneous with Ottawan orogenesis in south-eastern Canada has been documented in most areas. However, direct evidence of accreted arc terranes and orogenesis that might have been contemporaneous with Elzevirian activity in Canada is lacking in most massifs. Indeed, some researchers have used results from Pb and Nd isotopic studies to propose a non-Laurentian origin for the Appalachian inliers, in spite of the similarities in lithologies and tectonic evolution observed between these massifs and areas of the Adirondacks and Canadian Grenville Province.

Plate Tectonic Setting

The tectonic truncation of Precambrian geological belts in the south-western United States led to the search for possible continuations of the Grenville Orogen on other continents. Such investigations ultimately resulted in palaeogeographic reconstructions of a possible supercontinent named Rodinia (from the Russian verb 'rodit', meaning 'to beget' or 'to grow'), the amalgamation of which was marked in part by the collisional Grenvillian Orogeny. Two such reconstructions, the south-west United States–East Antarctic (SWEAT) and Australia–south-west United States (AUSWUS) hypotheses, extend the Precambrian geology of the south-western United States to the Australia–Antarctic shield region, but differ in the proposed position of the Australian craton relative to the conjugate margin. Another reconstruction, based largely on sedimentological patterns, proposed a Proterotozoic connection between Siberia and western Laurentia. Such palaeogeographic reconstructions, which depend on many lines of multidisciplinary geological evidence, are constantly refining geologists' view of Mesoproterozoic Earth.

Figure 8 Plate tectonic reconstruction of Rodinia at about 1000 Ma, with Laurentia located in its present day North Ameri can coordinates. Orogenic belts are depicted in random line pattern and include the Grenville belt (1), Llano belt (2), Namaqua belt (3), Albany/Fraser belt (4), and East Ghats belt (5). Labelled cratons include Baltica (B), Amazonia (AM), Rio de Plata (RP), Siberia(S,), Australia (AUS), East Antarctica (EANT), India (IND), Madagascar (M), Kalahari (K), and Congo (C). Modified with permission from Dalziel IWD, Mosher S, and Gahagan LM (2000) Laurentia Kalahari collision and the assembly of Rodinia. *The Journal of Geology* 108: 499 513.

The sequence and style of structural and magmatic-thermal events, extended time-span of development, and considerable degree of crustal thickening that characterized Ottawan orogenesis are consistent with the case histories of other major mountain belts, including the Himalayas. Although the cause of the Ottawan Orogeny is not yet rigorously identified, there is general agreement that it involved collision with a large outboard continental block located to the south-east. Recent palaeogeographic reconstructions indicate that Grenville orogens defined a segmented belt that resulted from convergence of multiple cratons at 1.2–1.0 Ga. The type-Grenville belt of eastern Laurentia and its possible southern continuation, the Llano-Namaqua Orogen of Texas and south-west Africa, formed a major part of this orogenic global network (Figure 8), resulting from accretion of the Amazonian and Rio de Plata cratons to the large nucleus of Rodinia (consisting of Laurentia, East Antarctica, and the Grawler Craton of south-eastern Australia) at *ca.* 1.1 Ga. The Llano-Namaqua Orogen formed nearly synchronously as a result of collision with the Kalahari Craton (Figure 8). These tectonic events were part of a series of cratonic collisions that took place on a global scale at 1.2–1.0 Ga during a relatively short period in which most of most of Earth's continental lithosphere was amalgamated to form the supercontinent of Rodinia, the continental shelves of which were later to give rise to many of Earth's earliest organisms and the eventual dismemberment of which produced all subsequent continents.

See Also

Europe: Scandinavian Caledonides (with Greenland). **Igneous Processes**. **Metamorphic Rocks:** PTt - Paths. **North America:** Precambrian Continental Nucleus; Northern Appalachians; Southern and Central Appalachians. **Plate Tectonics**. **Precambrian:** Overview. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Carr SD, Easton RM, Jamieson RA, and Culshaw NG (2000) Geologic transect across the Grenville orogen of Ontario and New York. *Canadian Journal of Earth Sciences* 37: 193 216.

Condie KC (1997) *Plate Tectonics and Crustal Evolution.* Oxford: Butterworth Heinemann.

Dalziel IWD, Mosher S, and Gahagan LM (2000) Lauren tia Kalahari collision and the assembly of Rodinia. *The Journal of Geology* 108: 499 513.

Davidson A (1998) An overview of Grenville Province geology, Canadian Shield. In: Lucas SB and St Onge MR (eds.) (coordinators) *Geology of the Precambrian*

Superior and Grenville Provinces and Precambrian Fossils in North America: Geological Survey of Canada, Geology of Canada, No. 7, chapt. 3, pp. 205 270. Dart mouth, NS: Geological Survey of Canada. (Also pub lished as *The Geology of North America* (1998), vol. C 1. Washington, DC: Geological Society of America.)

Gower CF and Krogh TE (2002) A U Pb geochronological review of the Proterozoic history of the eastern Gren ville Province. *Canadian Journal of Earth Sciences* 39: 795 829.

Hynes A and Ludden JN (eds.) (2000) *Canadian Journal of Earth Sciences, Special Issue: The Lithoprobe Abitibi Grenville Transect*, vol. 37, nos. 2 and 3. Ottawa: NRC Research Press.

McLelland J, Daly S, and McLelland JM (1996) The Gren ville Orogenic Cycle (ca 1350 1000 Ma): an Adirondack perspective. *Tectonophysics* 265: 1 28.

Moore J and Thompson P (1980) The Flinton group: a late Precambrian metasedimentary sequence in the Grenville Province of eastern Ontario. *Canadian Journal of Earth Sciences* 17: 1685 1707.

Moores EM and Twiss RJ (1995) *Tectonics*. New York: WH Freeman.

Rankin DW, Drake AA Jr., and Ratcliffe NM (1990) Geo logic map of the U.S. Appalachians showing the Lauren tian margin and Taconic orogen. In: Hatcher RD Jr., Thomas WA, and Viele GW (eds.) *The Appalachian Ouachita Orogen in the United States: The Geology of North America,* vol. F 2, plate 2. Boulder, CO: Geological Society of America.

Rivers T (1997) Lithotectonic elements of the Grenville Province: review and tectonic implications. *Precambrian Research* 86: 117 154.

Wardle RJ and Hall J (eds.) (2002) *Canadian Journal of Earth Sciences, Special Issue: Proterozoic Evolution of the Northeastern Canadian Shield: Lithoprobe Eastern Canadian Shield Onshore Offshore Transect*, vol 39, no. 5. Ottawa: NRC Research Press.

HERCYNIAN OROGENY

See **EUROPE: Variscan Orogeny**

HIMALAYAS

See **INDIAN SUBCONTINENT**

HISTORY OF GEOLOGY UP TO 1780

O Puche-Riart, Polytechnic University of Madrid, Madrid, Spain

Introduction

Ancient civilisations in contact with nature inquired about their origins and about particular geodynamic phenomena. In most cases they satisfied themselves with empiric explanations; they even used deities in order to understand inexplicable situations.

Little by little humans learnt how to observe their environment and arrange processes. During the Renaissance the first geologic principles were born and this knowledge spread rapidly. Natural phenomena were understood in terms of dynamic cause–effect, although many dogmatic and magic interpretations persisted.

Many authors agree that geology, began to be structured as a science in the second half of the eighteenth century with Abraham Gottlob Werner (1749–1817), father of Neptunism (Figure 1). However, some geologic paradigms such as diluvialism existed before neptunism; all of them contained countless mistakes and ambiguities.

This article outlines the period up until 1780, which thus incorporates the work of James Hutton (*see* **Famous Geologists:** Hutton). His ideas were important in the development of geology, more specifically relating to the origins and dating of rocks. Geology was not completely defined till the birth of Stratigraphy at the end of the eighteenth century and Palaeontology at around 1830.

The Dawn of Geology

Thinking about the Earth first occurred when man, faced with natural phenomena such as earthquakes and volcanoes, posed questions about such phenomena and sought to provide answers in naturalistic terms. Practical matters, such as the task of prospecting for mineral resources, also stimulated interest in the Earth.

Figure 1 Abraham Gottlob Werner (1749 1817). Father of Neptunism.

The Greek philosophers thought that the universe was governed by unchanging principles and with intelligible and discoverable natural laws. This contrasted with the mythopoeic or magical explanations of nature found more generally in the ancient world and in non-scientific cultures today.

In his *Histories*, Herodotus of Halicarnassus (Ca 484–425 BC) spoke of the sedimentary loads of the Nile and of the slow growth of its delta. This was perhaps the first recorded statement based on observation indicating an awareness of the magnitude of geological time. But myth and naturalistic explanation were intertwined in Greek thought.

Plato (427 or 428–348 BC) (in *Phaedo*, 111–112) described the Earth as having internal passages carrying "a vast tide of water, and huge subterranean streams of perennial rivers, and springs hot and cold, and a great fire, and great rivers of fire, and streams of liquid mud, thick or thin", as well as a great internal chasm, Tartarus. Water moved with a 'see-saw' motion within the Earth, like the tides, and produced springs that fed rivers and streams, and returned to the sea and thence to Tartarus. The surging waters also generated great winds inside the Earth. Volcanoes were produced by the escape of rivers of fire from within. These speculations were naturalistic, but also explicitly said by Socrates to be 'myth'. Such ideas were to endure until the eighteenth century.

Along similar lines Aristotle (384–322 BC), a pupil of Plato, suggested that earthquakes were caused by subterranean winds passing through cavities within the Earth. Fossils were nature's failed attempts in the creation of living beings (the theory of *vis plastica*).

Although some authors consider Theophrastus of Ephesus (*ca* 371–*ca* 287 BC) to have written the first mineralogical treatise, *Perilithon*, there are references to a work, now lost, written by his teacher Aristotle. In the surviving *Meteorologica*, Aristotle ascribed the origin of minerals and metals to dry/smoky or moist/vaporous exhalations from within the Earth.

Minerals' curative purposes were considered in Dioscorides' *De materia medica* (*ca* 77 AD). Processes such as saline crystallization or exfoliation were remarked upon, and origins of substances of supposed medicinal value were mentioned. This medical tradition was to continue in the attempted mineral/chemical cures advocated by Paracelsus in the sixteenth century.

The Romans were less interested in abstract knowledge than were the Hellenes, but were practical, and skilled in the use of stone for building. The most notable Latin 'scientific' text was Pliny the Elder's *Historia naturalis* (first century), consisting of 37 books, uncritically compiled from 2000 works of antiquity. The last five books dealt with the mineral kingdom, with mining and smelting practices, and with the characters, occurrences, and uses of many mineral substances.

On observing sea shells in the mountains, Ovid (43 BC–17 or 18 AD) inferred that those lands had formerly been covered by the sea. He also realised how fluvial valleys could be formed and how water gradually reduced relief. The materials swept along would be deposited, lower down, in flooded areas, where on drying and hardening they would become rocks. We have for the first time the pattern: erosion, transportation, sedimentation, and lithification.

The idea of the regeneration of minerals and ores in mines was advanced by Pliny's teacher, Papirio Fabiano, an idea still maintained in the seventeenth century, as in the case of Alvaro Alonso Barba in *El arte de los metales* (1637). In this work, the Earth supposedly had the ability to 'reproduce', as envisaged in antiquity.

During the Dark and Middle Ages, Aristotle's influence continued in the West, but linked with Christian viewpoints. Thus, for example, St Isidore of Seville's (560–636), in *Etymologies* (a work considered to be the first encyclopedia), pointed to the organic origin of fossils, but connected them with the Flood.

Alchemy coming from Persia (eighth–nineth centuries) influenced the works of Ibn Sina (Avicenna) (930–1037) and subsequently Christian authors like Alfonso X (1221–1284), Raymond Lully (1235–1315), Arnaldo Vilanova (*ca* 1238–1311), Ulisse Aldrovandi (1527–1605), Andreas Libavius (1560–1616), and Alonso Barba (1569–1662). There developed the so-called theory of the opposites whereby things combined or repelled one another according to their 'sympathies' or 'antipathies'. Some spoke of the gender of minerals. For example, the word 'arsenic' derives from the Greek word for male. Minerals supposedly formed from the appropiate combinations. Alchemy was the forerunner of inorganic chemistry.

Another feature of the Middle Ages was the proliferation of 'lapidaries': list of stones, etc., with descriptions of their properties, uses, etc. Ibn Sina wrote *De lapidibus*, in which minerals were classified according to the quadrichotomy: stones/earths; metals ('fusibles'); sulphurous fossils (combustibles); salts ('solubles'). Ahmad Al Biruni (973–after 1050) mentioned more than 100 minerals and metals in his treatise on gems (*Kitab-al-Jamahir*), and accurately determined specific gravities for several types. Also Alfonso X of Castile (1221–1284) (Alfonso the Wise), translated numerous Arab lapidaries, where the properties of minerals supposedly varied according to the positions of the heavenly bodies. Al Biruni, born in Uzbekistan and a great traveller, was also notable for his studies of rivers. He recorded evidence for

changes in the course of the Amu Darya River, and the decrease of sediment size down the Ganges. So (anachronistically) he could be called a fluvial geomorphologist.

The greatest Mediaeval author on the mineral kingdom was Albertus Magnus (St Albert of Cologne), Bishop of Ratisbon and doctor of the Church (1193–1280). Anticipating Renaissance authors, he stated that experience alone was the source of knowledge of physical things. He tried to link faith and reason when he pointed out that the sea could never have covered the whole Earth by natural causes. In *De mineralibus* he recognised about 100 mineral species. Both minerals and rocks were thought to have formed from molten masses.

The First Geological Principles: The Observation Phase

With the Renaissance, the geocentric Aristotelian and Thomist universe collapsed in the 'Copernican Revolution', and observation rather than 'authority' became central to science. For example, Bernard Palissy (1510–1590), pointed out "I have never had any other books than the skies and the Earth whose pages are open to all". Systematic ordering of the observations facilitated the establishment of the first geological principles. Information also spread faster thanks to the printing press. This was particularly true of great natural catastrophes, such as the eruption of Vesuvius in 1538, which prompted interest in the Earth.

Leonardo da Vinci (1452–1519) visited the Alps and realised that the geological structure was the same on both sides of the fluvial valleys. The rivers carried away materials to the sea, where they might bury shells. When land rose up it formed hills that, on being cut by rivers, reveal layers or strata. Shells in such strata were not carried there by the Flood. We have one of the first visions, albeit incomplete, of the geological cycle.

Palissy showed that what are today called rudist lamellibranchs are 'lost' species. This recognition of extinction was an important contribution towards recognition of the Earth's antiquity. Interest in fossils developed little by little, as when Father Jeronimo Feijoo y Montenegro (1676–1764) also cited discoveries of lost species. There was still a long way to go to before fossils were used to determine the relative chronology of the landscape.

In his *Principia philosophiae* (1644) the French philosopher René Descartes (1596–1650) considered the Earth as an old cooling star. There was incandescent material in its interior, around which there was a layered structure (metallic, heavy material, air–water, and outer crust). As the globe cooled, the crust cracked and collapsed, thus creating mountains and seas. In this speculative theory we have the first attempt to explain the internal structure of the eEarth in mechanical terms (i.e., in terms of the 'mechanical philosophy' according to which all natural phenomena were explained in terms of matter and motion). Descartes also saw the planet as a great 'still', heated by its internal material. So sea water penetrating into the Earth was distilled in the interior, leaving the salt there.

Descartes' theory of a central heat re-appeared in the work of the Jesuit Athanasius Kircher (1601–1680), *Mundus subterraneus* (1665), which proposed a great central *pyrophylacium* or repository of heat, linked also with ideas going back to Plato. The main repository was connected by channels to other lesser fires, and the network of interconnected channels served as conduits for volcanoes at various places on the surface. In addition to the *pyrophylacium*, there were *aerophylacia*, through which circulated the subterranean winds that supposedly caused earthquakes; and *hydrophylacia*, or water-containing caverns, which were fed from the sea and sustained springs. (The model had similarities to that in Plato's *Phaedo*.) Earthquakes gave rise to the formation of mountains. Kircher also revived the organicist theories, speaking about the uterus of the globe, and *vis petrifica* and *vis seminalis* (petrifying and seminal powers). The Earth was a living organism with a capacity for reproduction and the other functions of a living being (so inside the Earth salt water becomes fresh through a quasi-'metabolic' process). Thereby both external and internal 'geodynamic' phenomena were explained.

Niels Stensen (1638–1686) (Nicolaus Stenonis or Steno) (*see* **Famous Geologists:** Steno), a Danish physician in the service of the Medici family in Florence, was less speculative and more original. He authored *De solido intra solidum de naturaliter contento dissertationis prodromus* (1669), in which, from the study of quartz crystals, the law of the constancy of interfacial angles was first recognized. With Steno, we also have what might be called the first 'stratigraphic diagram'. Sediments accumulate, forming horizontal layers in which marine or terrestrial fossils were buried, the oldest layers being below and the younger ones above. These layers could be undercut by erosion, fracturing and collapsing. Then new horizontal layers were deposited, at an angle to the earlier ones. One of geology's main problems, to establish a chronological order of events, had begun to be resolved. (Steno's principle of superposition was relatively trivial: the lowest layer of bricks in a wall is put in place before the upper ones. But it required imagination to apply this idea to the easily observed layered rocks.)

But stratigraphy had still to be put together. Giovanni Arduino (1714–1795) made the first chronostratigraphic division (with geological plan and section included) when he divided the rocks of the Alpine landscape into: 1) Primary: formed by quartzites, and slates; 2) Secondary: formed by limestones, sandstones, and shales; 3) Tertiary: formed by limestones, sandstones, gypsums, and clays; and 4) Alluvium. The idea was set forth in two letters addressed to Antonio Vallisnieri (published 1760).

The seventeenth century was also characterized in the West by attempts to reconcile observations of natural history with the Bible, aligning 'faith and reason', in what was called 'physico-theology', or the attempted interweaving of natural philosophy (science) and religion. Such work continued well into the nineteenth century, and even to the present.

Thomas Burnet's (1636–1715) work *Telluris theoria sacra* (1681) provides a good example. For this Anglican cleric, the Earth's initial chaotic material was ordered by gravity, with the heaviest parts in the centre and the lightest parts at the surface. The result was a concentric structure: 1) a Kircherian or Cartesian igneous core; 2) liquid; 3) an oily layer; and 4) an outer crust hardened by the sun (*ossatura telluris montium*). When the central vapours acted on the outer crust it cracked and broke, giving rise to the Flood ("all the fountains of the great deep [were] broken up": Gen. 7, 11). If the Earth had not been flat, it could not have been covered by the waters (here reason and design were introduced). After the Flood, the waters supposedly withdrew, taking much with them, thus causing the Earth's relief. Such theories, connecting the Noachian Flood with geological observations, came to be called 'diluvialist'. An antecedent of 'classical' diluvialism was perhaps the Spaniard José González Salas (1588–1651) who, in 1650, stated that the Flood changed the face of the Earth.

For Isaac Newton's successor at Cambridge, William Whiston (1671–1752), a comet caused water escape from the Earth's interior, while for John Woodward (1665–1728), the waters supposedly dissolved the Earth, which was then converted to its present layered state as matter separated out according to the law of gravity (*Essay Toward a Natural History of the Earth*, 1695).

The age of the Earth was calculated in accordance with the biblical records, as Alfonso X the Wise had done in his *General History*. This was likewise done by the Cambridge classicist John Lightfoot (1642–1644); the Anglican Primate of Ireland, James Ussher (1650); and William Lloyd, Bishop of Worcester (1701). They arrived at various values between 3928 and 5199 years old. These authors, who today may seem detached from reality, were in fact careful scholars who sought to reconcile Jewish, Christian, and pagan historical records. Ussher's date for the Earth's creation (4004 BC) became the best known, as printed in the margins in Lloyd's edition of the 'King James' Bible (1701).

Outside the religious arena, Gottfried Liebniz's (1646–1716) *Protogea* (1684) proposed that rocks had been formed by two processes: 1) the cooling of fused material to give an Earth with a 'glassy' surface; 2) the action of waters on this hard surface and the concretion of solid elements contained in aqueous solution. Leibniz's ideas thus anticipated the late 18th century debate between 'Plutonists/Vulcanists' and 'Neptunists'.

José Vicente del Olmo (1611–1699), in his *New Description of the Orb* (1681), stated that the mountains were raised up due to internal exhalations. The elevated areas were then eroded by rain, wind, and river floods. Thus a balance was established in nature, rather than a single progression of change such as Burnet envisaged. But let it be remembered that Seneca (*ca* 3 BC–65 AD) and the Epicureans had long before envisaged a 'balance of nature'.

Geology as a Science is Born

The eighteenth century was characterized by the economic development of the western nations and the development of democratic ideals. Inspired by the accomplishments of science and technology, the Enlightenment world-view, which saw things as essentially intelligible with problems being capable of solution by rational beings with minds unclouded by superstition, was to be driven on by the idea of progress, which was born with it. New centres of teaching such as Göttingen University were founded, and scientific publications for the technical and educated bourgeoisie, such as Denis Diderot's *Encyclopédie*, appeared. Experimentation also acquired greater importance, even in the study of the Earth. It was a period of glorification of the rational, where it seemed that the only things that mattered were those that could be counted, measured, weighed, or rationally calculated.

The end of the eighteenth century coincided with the Industrial Revolution (England) whose foundations were iron, coal, steam, and textile manufactures. The need for additional natural resources boosted mining, and between 1766 and 1788 the mining academies of Freiberg, Chemnitz (now Banksa Stiavnica), St Petersburg, Almaden, and Paris were founded in that order. 'Subterranean geometry' and mineralogy were taught and mineralogy began to develop into petrography, stratigraphy, palaeontology (later), and, eventually, geology (around the

end of the eighteenth century). Curiously, Britain was backward in such centralized technical education.

Abraham Gottlob Werner (1750–1817) was appointed to the Mining Academy of Freiberg in Saxony in 1775 where he developed his Neptunist theory (1777), which proposed that all rocks, even basalts, were formed by chemical precipitation from a primordial ocean or *allgemeines Gewaesser*. (There were, however, precursors of this theory, such as the Frenchman Benôit de Maillet [1755].) According to Werner, by successive sedimentation onto an irregular terrestrial core, four types of formations were supposed to be deposited: 1) Primitive: crystalline rocks such as granite and gneiss; 2) Transitional: limestones, slates and quartzites; 3) *Flöetz*: formed from what we consider today to be the layered rocks from the Permian to the Cenozoic; and 4) Alluvial: (superficial) deposits. (The 'Transition' category was absent from initial exposition of Werner's theory.) These 'chronostratigraphic' divisions had previously been adumbrated in Germany by others such as Johann Lehmann (1719–1767) and Georg Christian Füchsel (1722–1773), and also by the German traveller, Peter Simon Pallas (1741–1811), in the Urals (1768). The Primitive formations would be found in the central parts of mountain ranges, from which the water would have withdrawn first.

Werner's theory gave an approximation to the order of rocks observed in the field. But there were questions that Werner's theories couldn't solve:

1. Where did the water of the supposed primordial ocean go to?
2. Is the Earth inactive? (For Werner, sloping strata corresponded to margin sedimentation.)
3. How were rocks such as basalt, found on the tops of hills, to be explained? (Werner thought that basalt was also precipitated from his ocean, the level of which supposedly rose again for some unexplained reason.)
4. How were mineral veins and dykes to be accounted for? (Werner thought that material might have precipitated from above filling rents in the crust.)
5. How were volcanoes to be explained? (Werner thought that they might be due to the combustion of subterranean coals, etc.)
6. How could the universal ocean dissolve so much siliceous matter? (This question was never answered satisfactorily, though the occurrence of siliceous springs and quartz veins in some rocks suggested precipitation from solution.)

Not everything that came from Werner was wrong. He praised observation and the use of scientific method and he assisted into the emergence of geognosy or geology, 'oryctognosy' or mineralogy. With Werner (and before him in Russia with Lomonosov [1711–1765]), mineralogy acquired its own body of doctrine. He classified minerals according to their external characteristics (physical properties), as Linnaeus (another important Enlightenment figure) had done with plants and animals, between 1735 and 1760. The observation of crystalline forms was to lead to the birth of crystallography. This science had taken its first steps beyond Steno thanks to the Swiss naturalist Moritz Anton Capeller (1685–1769) with his *Prodomus cristallographie* (1723) and the Frenchman Jean Baptiste Louis Romé de l'Isle's (1736–1790) *Essai de cristallographie* (1772), soon to be developed further by René-Juste Haüy.

Werner was to have many disciples who would write important pages in the annals of geology during the nineteenth century, such as Guyton de Morveau (1737–1816), Horace Bénédict de Saussure (1740–1799), Déodat Gratet de Dolomieu (1750–1801), Juan José Elhuyar (1754–1896), Fausto Elhuyar (1755–1833), Andrés Manuel del Río (1765–1849), Alexander von Humboldt (1769–1859), Leopold von Buch (1774–1853), Robert Jameson (1774–1854), etc. They tried to use his stratigraphic order, worked out in Saxony, as a 'paradigm' for examining and interpreting rocks in other parts of the world.

Another notable eighteenth-century authority was the keeper of the *Jardin des Plantes* in Paris: Georges Louis Leclerc, Comte de Buffon (1707–1788), author of a great 36-volume *Histoire naturelle*. In a supplement of this work entitled *Époques de la nature* (1778) he put forward three basic ideas: 1) a longer duration of geological time (compared with the Biblical account); 2) organic evolution, preparing the way for transmutationism and evolutionism; and 3) palaeogeography.

Like Descartes, Buffon thought that mountains were formed by contraction during the Earth's cooling. He also examined the problem of the age of the Earth *experimentally*, heating spheres of different sizes and measuring how long they took to cool until they could be touched; and by analogy he estimated the possible age of the Earth. He arrived at the conclusion that it would have taken 74 832 years to have cooled to its present temperature (and privately speculated on the possibility of a much greater age).

Through further experimentation Buffon obtained silicates by melting clays. Nevertheless, he held to some older ideas, such as the view that earthquakes were caused by explosions of gases in the Earth's cavities or that volcanoes were produced by the combustion of sulphur and bitumen.

The hydrological cycle was also quantified, in accordance with the calculations of Edmé Marriotte

(1690), Pierre Perrault (1674), and the suggestion of Edmund Halley (1714–1716) that one could measure the rate of increase of salinity in lakes that had no discharge rivers, and then gauge how long it might have taken for the oceans to acquire their salinity. Nevertheless, eighteenth-century geology was not obviously an experimental science. It was not then possible to reproduce variables such as pressure, temperature, or time, which reach very high values in many of the processes occurring in nature.

In the eighteenth century there were still authors who regarded fossils as 'figured stones', as did the French physician Pierre Barrère, author of *Observations sur l'origine et la formation des pierres figurées* (1746). The diluvialist school was also active. Thus, for the Spaniard, Father Antonio Torrubia, in his *Aparato para la historia natural* (1754), fossils were represented as remains of the Flood. Nevertheless, a significant rejection of diluvialism occurred in the mid-eighteenth century, mainly in central Europe, with authors, such as the Göttingen professor Samuel Christian Hollmann (1753), while the Swiss cleric Johan Georg Sulzer (1762), pointed out the marine origin of fossils. Numerous examples were described and the natural history cabinets were filled with specimens, but without an agreed system for their cataloguing. Although the influential Werner rejected fossils as the basis for the study of stratification, they began to gain in importance, and increased knowledge began to pave the way for the birth of stratigraphy, at the end of the century, and of scientific palaeontology, which entered at the end of the eighteenth century.

Practical matters were also important in the Enlightenment. Between 1778 and 1782, Jean Étienne Guettard (1715–1786) and Inspector General of Mines Antoine Grimoald Monet (1734–1817) jointly published their *Atlas minéralogique de la France*, which showed the distribution of deposits of economic significance across their country. In Sweden, the chemist Torbern Bergman (1777) initiated general methods of mineral analysis in the 'humid' way, bringing mineral substances into solution by the action of acids or alkalis and then identifying components by a sequence of precipitation reactions. Prospecting for coal was enhanced by boring techniques, but without palaeontological control the results were not always useful through misidentification of strata.

The great catastrophe of the Lisbon earthquake on 1 November, 1775, sowed pessimism in the scientific world. There were many, including Buffon, who thought of the progressive degradation of the cooling globe. But in Spain, the naturalist Brother Benito Feijoo y Montenegro, in his *Cartas eruditas y curiosas* (1760), tried to calm things down by pointing out the greater the force of the previous one and that repetitions of earthquakes are less likely. In Germany, Immanuel Kant argued that earthquakes had natural causes and had nothing to do with the moral condition of mankind. But they could remind us not to try to find happiness in worldly goods. Old earthquake myths endured nevertheless, and it was only at the end of the nineteenth century that geologists began to suspect the main causes of tremors.

See Also

Biblical Geology. **Famous Geologists:** Hutton; Steno. **Geomythology**. **History of Geology From 1780 To 1835**. **Minerals:** Definition and Classification. **Stratigraphical Principles**.

Further Reading

Ellenberger F (1988) *Histoire de la Geologie Tome 1 Des Anciens à la première moitié du XVIIe siècle*. Paris: Technique et Documentation Lavoisier.

Ellenberger F (1994) *Histoire de la Geologie Tome 2 La grande éclosion et ses prémices 1660 1810*. Paris, London, and New York: Technique et Documentation Lavoisier.

Faul H and Faul C (1983) *It Began with a Stone: A History of Geology from the Stone Age to Age of Plate Tectonics*. New York, Chichester, Brisbane, Toronto and Singapore: John Wiley & Sons.

Gaudant G and Bouillet G (2000) La genèse et l'interpréta tion des 'fossiles' dans la science classique: de la Renais sance aux Lumières. *Bulletin de la Société Géologique de France* 171: 587 601.

López Azcona JM (1985) Los jheólogos. *Revista de Materiales y Procesos Geológicos* 3: 179 187.

Mather KF and Mason SL (1939) *A Source Book in Geology*. New York and London: McGraw Hill.

Oldroyd DR (1996) *Thinking about the Earth: A History of Ideas in Geology*. Cambridge (Mass): Harvard University Press.

Pelayo F (1996) *Del Diluvio al megaterio. Los orígenes de la paleontología en España*. Madrid: Consejo Superior de Investigaciones Ciéntificas.

Rappaport R (1997) *When Geologists were Historians, 1665 1750*. Ithaca and London: Cornell University Press.

Rossi P (1984) *The Dark Abyss of Time: The History of the Earth & the History of Nations from Hooke to Vico*. Chicago and London: The University of Chicago Press.

Sequeiros L (2003) Las raíces de la Geologia: Nicolas Steno, los estratos y el diluvio universal. *Enseñanza de las Ciencias de la Tierra* (10,3): 217 242.

Wagenbreth O (1999) *Geschichte der Geologie in Deutsch land*. Stuttgart: Georg Thieme Verlag.

Wendell E Wilson (1994) The history of mineral collecting. *Mineralogical record* 25(6): 1530 1799.

HISTORY OF GEOLOGY FROM 1780 TO 1835

D R Oldroyd, University of New South Wales, Sydney, Australia

Introduction

The years 1780–1835 mark the period when geology emerged as a science *sui generis*, distinct from mineralogy. It began to have an institutional base in the form of scientific societies and began to be taught at universities (though there was earlier tuition in cognate subjects at mining academies in Europe). 1835 saw the establishment in England of the world's first national geological survey (though there were a few earlier 'private' surveys, or national 'mineral surveys'). The period concerning us here has been called the 'hinge of history', with its shift from the eighteenth century Enlightenment to the Romantic Movement, the Gothic Revival, and extensive industrialization. It was also a time of political upheaval, with the French Revolution and the Napoleonic Wars. There was a great expansion of European horizons, both in space and time. It included notable explorations, such as those of Flinders and Baudin around Australia, d'Orbigny and von Humboldt in South America, and Darwin's *Beagle* voyage. America was extending westwards, and Russians were consolidating their hold in Siberia. Much of India was under British rule and exploration was getting underway in Africa. These explorations had economic and military imperatives, but were also of great scientific significance. By 1835, the bounds of time of earlier theological constraints were truly burst by geologists, and during this 'hinge period' the Earth began to be seen as an object whose history could be revealed by empirical examination of rocks, fossils, and strata. The cultural movement known as 'historicism' came to the fore, according to which things (like the principles of law) could supposedly be best understood by studying their history. In this sense, geology fitted into the 'spirit of the times'. Understanding of the Earth became 'historicized'.

That is one way of looking at the period of geology's emergence. But geology can also be seen as having emerged within the context of the Industrial and Agricultural Revolutions: it was the practical men concerned with mining, quarrying, surveying, agriculture, etc., who were responsible for many of the discoveries on which rested the intellectual achievements of the 'geological elite', who established learned societies and developed high theory. The 'little men', on this view, did as much to found the science as did the 'gentlemanly geologists' whose writings have since become well known to historians.

The period 1780–1835 is also interesting for the appearance of important rival geological theories, and the contests between supporters of the competing doctrines. The so-called Neptunist and Vulcanist theories, associated particularly with the names of Abraham Gottlob Werner and James Hutton (*see* **History of Geology Up To 1780, Famous Geologists:** Hutton), offered radically different geological theories and effectively functioned as competing paradigms. Likewise, there were strong differences of opinion (amounting to substantial philosophical differences), summarized by the terms 'catastrophism' and 'uniformitarianism'. Debates on these matters ranged into the issue of the age of the Earth; the question of whether geological discoveries could or could not be reconciled with, or support, theological beliefs; and questions about the appropriate methodology for the geosciences. Rocks and minerals were first clearly distinguished by Alexandre Brongniart (1827).

The 'Little Men' and a Geological Map

The name of William Smith has long been remembered for his discovery that different strata could be recognized and discriminated by means of their fossil contents (*see* **Famous Geologists:** Smith). Smith, who came from south-west England, was a surveyor and engineer who worked on road and canal projects, land drainage schemes, coal and ore prospecting, etc. Living near Bath, he came to realize that there were two kinds of oolitic limestones with similar appearances but different fossils. Further afield, chalks and greensands showed similar distinctions. Accordingly, as his work gave him ample opportunity for observing and collecting, he began to assemble and arrange fossils according to what we would call their stratigraphic horizons. He drew up a table (1799) listing the lithologically different strata of southern England (from coal to chalk), along with their characteristic fossils. He then embarked on the huge project of trying to identify and enter on a coloured map the strata of the whole of southern Britain. This single-handed, hand-painted, map was issued in 1815. But the intellectual elite of the Geological Society (founded 1807) hardly gave credit where credit was due, and set about making their own map, which was issued in

1819, making some use of Smith's results. But in 1831 Smith was dubbed the 'Father of English Geology' by Adam Sedgwick (*see* **Famous Geologists:** Sedgwick).

Apart from his map (and sections), Smith should be recognized for his practical and economic contributions. On the basis of his biostratigraphic knowledge, he knew where coal might or might not be expected to be found and could advise entrepreneurs accordingly. He was not much interested in *why* the rocks and fossils were distributed as they were or how old they were. Smith was only one of the many practical men who contributed to the emergence of geology, but most of them left few historical traces and are little known. There were also skilled artisan fossil collectors, the best known of whom was Mary Anning working on the south coast of England, who extracted specimens of ichthyosaurs and lesser fossils for sale to connoisseurs and museums.

Neptunism

In Germany, lithostratigraphy was chiefly deployed during the late eighteenth century, according to the system of Abraham Werner, which was taught at the Freiberg Mining Academy in Saxony. It postulated the existence of a standard 'formation sequence' of rocks: Primitive rocks (especially granite); Transition rocks (schists, greywackes, etc.); *Flöetz* (layered) rocks; Basalt; and Alluvial deposits. His observational work was chiefly undertaken in Saxony and Silesia, and through the influence of his many distinguished pupils this area of Europe came to be regarded as an exemplar (paradigm) for studies further afield. Werner was a gifted mineralogist and teacher, but his lithostratigraphic theory was based on the misconception that vast amounts of rock could formerly have been dissolved in a hypothetical universal ocean. He thus denied the igneous origin of rocks like granite or basalt, and because basalt caps were found overlying sand beds on hills near Freiberg he imagined that basalt could have been deposited from water (an idea perhaps encouraged by its frequent columnar jointing). This meant that the supposed mineral-bearing ocean would have had to have risen to deposit the hilltop basalts. So the theory of a falling universal ocean was implausible, both chemically and physically. Nevertheless, Werner's mineralogical exactitude and lithostratigraphy were appealing and influential.

But already in the eighteenth century visitors to central France such as Jean-Étienne Guettard had recognized the existence there of former volcanoes, and their associated basalts were interpreted as igneous products by Nicholas Desmarest, who deduced the historic sequence of lava flows in the area, the older ones being substantially eroded while the younger ones were fresh and had run down previously formed valleys. Pupils of Werner such as Leopold von Buch started off their careers as Neptunists, but turned against their master's ideas when they reached the Auvergne.

From about 1820, there were few who still advocated Neptunism, but later in the nineteenth century there was renewed interest in the role of water in the formation of rocks, and again in the migmatist/magmatist debates of the twentieth century: Werner's ideas had intellectual descendants. In Edinburgh, they had a strong advocate in the Professor of Natural History, Robert Jameson.

Vulcanism

Vulcanism (to be distinguished from the modern term volcanism) was the sobriquet for the theory that ascribed primacy to the agency of heat in geological processes. Such ideas were strongly argued by the Scottish Enlightenment geologist, James Hutton. As a farmer, Hutton saw soil being washed into the sea, and he supposed that if weathering and erosion continued indefinitely all the fertile land would be lost, and what he regarded as a world 'devised in wisdom' for human habitation would eventually be destroyed (he took a very long view of geological history). He therefore contemplated a cyclic theory that would allow for replenishment of soil.

In Werner's theory, volcanoes were 'weakly' explained by the combustion of underground coal deposits. But Hutton advocated a great central source of heat, in a way similar to Kircher's '*Pyrophylaciorum*' (*see* **History of Geology Up To 1780**), and analogous to the fires of the machines of the Industrial Revolution, which made the wheels of industry turn. Hutton envisaged the Earth as a kind of great machine. After his death, his 'Plutonist' theory achieved support from the experimental work of Sir James Hall, and from measurements of temperature gradients within mines, such as those made by Louis Cordier, which showed increased temperatures with depth supporting the idea of the Earth having a hot interior.

The Vulcanist–Neptunist Dispute

This controversy was prosecuted with greatest vigour in Edinburgh, between supporters of Hutton and Jameson. From today's perspective, it would seem that Vulcanism would be the easy winner, but the results of Hall's attempted experimental vindication of Hutton's doctrine were somewhat ambiguous. When granite was fused and slowly cooled the product did not look exactly like the starting material.

It is interesting to remark that a place like the Isle of Arran, near Glasgow, had exposures that were broadly compatible with both the Huttonian and Wernerian theories (see Figure 1). But unconformities such as the famous one examined by Hutton at Siccar Point (see Figure 2) could only be well explained in Huttonian terms. And granitic veins, such as those famously examined by Hutton at Glen Tilt, and anastomosing mineral veins, could hardly be accounted for by the Wernerian theory.

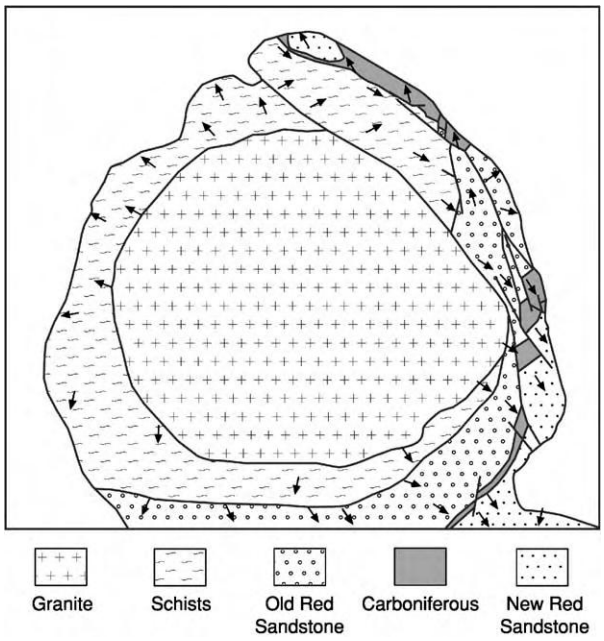


Figure 1 Sketch of the geology of the northern part of Arran. The exposures can be accounted for in terms of either the deposition of 'transitional' rocks (schists) and *Floëtz* sediments round a central granitic core; or the intrusion of granite into sediments, forming a dome structure with partial metamorphism round the granite.

Reconstruction of Past Environments

The chemist Antoine Lavoisier accompanied Guettard during his travels for the purposes of compiling a mineralogical map of France and acquired some knowledge of geology. Lavoisier became interested in the conditions that might be expected to occur at the bottom of the sea, at different depths and different distances from the shore, with cobbles near the edge of the sea, sand further out, and fine sediment some distance offshore. An idealized profile was published (1789) showing the occurrence of pelagic and littoral deposits. Such different sediments could also be seen in the strata observed inland, and the analogies were clearly understood. Thus we have the beginnings of attempts to reconstruct past environments on the basis of what was visible in the strata. Work of a similar kind was undertaken in the Paris basin in the early nineteenth century by Alexandre Brongniart and Georges Cuvier (see below). They recognized marine and fresh-water sediments on the basis of the fossil shells they contained (fresh-water forms generally having thinner shells). The arguments were based on analogies made with modern forms. Later, Charles Lyell (see below) realized that the processes of lime formation in freshwater lakes in his property in Scotland resembled similar depositions occurring in modern France. Examination of coal deposits suggested hot and humid swamp conditions. Or desert sandstones could be recognized for what they were. The beginnings of palaeoecology occurred during the 'hinge of history'.

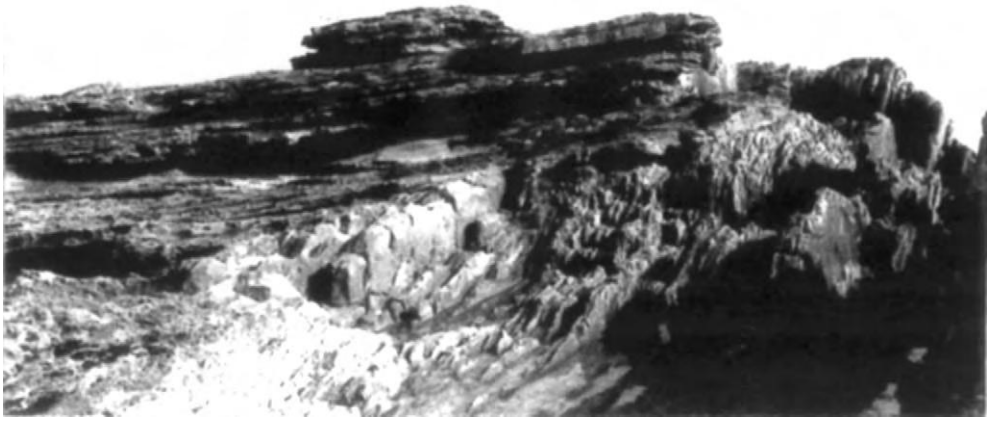


Figure 2 Hutton's Unconformity at Siccar Point, south east of Edinburgh. 'Old Red Sandstone' with basal conglomerate lying almost horizontally on upright 'schistus' (Silurian greywackes). After Secord JA, *Controversy in Victorian Geology*. Copyright © 1986 by P. U. Press. Reprinted by permission of Princeton University Press.

Palaeontology and Comparative Anatomy

Organic fossils had been known for centuries and by the eighteenth century few doubted that they were former living organisms. Fossil collecting became a popular, and even lucrative, pastime, especially when large specimens of ammonites, ichthyosaurs, ancient fishes, mammoths, etc., were discovered. Many collections, in private or public hands, were initiated. It was Georges Cuvier, Professor of Zoology at the *Muséum d'Histoire Naturelle* in Paris, who began to establish some fundamental principles for the study of fossils (*see* **Famous Geologists:** Cuvier). He started by trying to reassemble collections of mastodon bones from America. They were manifestly creatures rather similar to modern elephants, so he used the known arrangement of elephant bones for purposes of comparison, in order to reconstruct the mastodon skeleton. In this work, he recognized that the total structure must be one in which all the parts fit together to form a 'working whole'; and the structures must be such as to enable the creatures to live successfully in their 'conditions of existence'. This 'goodness of fit' necessarily applied to the lost soft parts as well as those that had become fossilised. So Cuvier enunciated his principle of 'co-ordination of parts'.

However, most fossils were shells, etc., and it was Cuvier's colleague Jean-Baptiste Lamarck who was responsible for collecting and studying these. He was the first to make a formal distinction between vertebrates and invertebrates and state the major divisions of the latter. Only then could invertebrate palaeontology make a proper start. It was less easy to make comparative anatomical studies for fossil invertebrates than for vertebrates, though Charles Darwin (*see* **Famous Geologists:** Darwin) later famously did so with barnacles.

In examining the fossil record, Cuvier found that there were substantial and apparently sudden changes in the fossil record as one ascended the stratigraphic column. This finding largely holds to this day for macrofossils, though smooth trends are known for foraminifera. The biostratigraphic 'breaks' were, in fact, convenient for stratigraphic investigations and mapwork such as that conducted by William Smith, but their theoretical explanation was difficult, using modern analogies as Cuvier had done for his mastodon work. Cuvier's explanation was couched in terms of sudden and catastrophic events (which we might call mega-tsunamis), of unknown cause, perhaps of global extent (Cuvier was inconsistent in his statements as to the universality or otherwise of his catastrophes). Famously, he wrote that in these supposed events "the thread of operations is here broken; the march of Nature is changed". His doctrine was later dubbed 'catastrophism' by William Whewell, and there was the suggestion that the very laws of nature were broken or suspended during such catastrophes; or past events were quite unlike anything occurring today.

Seventeenth- and eighteenth-century writers such as Steno (*see* **Famous Geologists:** Steno) had sometimes propounded 'catastrophist' theories of the Earth because they had a greatly compressed timescale for Earth history, supposedly in line with the 'Biblical' age of the Earth, of some 6000 years. This was not Cuvier's view. He wrote: "Genius and science have burst the limits of space;... Would it not also be glorious for man to burst the limits of time, and, by means of observations, to ascertain the history of this world, and the succession of events which preceded the birth of the human race?". This was the voice of modern geology. His theory envisaged geological direction and progress, taking place over unspecifiably great periods of time. However, he was unable to account for the origin of new living forms (unlike Lamarck, he was no evolutionist). He thought there was a catastrophe some 5000–6000 years ago.

Geology and Religion/Theology

It is easy to overplay the connections between geology and religion in the early years of the nineteenth century. But they certainly existed and were interesting and important. One line of approach, taken by the Swiss naturalist Jean-André de Luc, who was greatly concerned to 'harmonise' geology and Biblical history, was to suppose that there had been a great catastrophe, perhaps 5000 years ago, which separated history into antediluvian and postdiluvian eras, the division being associated with the Noachian Flood. The period since the Flood could be estimated by the rate of lake infilling, or thickness of peat accumulations. Such evidences almost allowed the date of the Flood to be quantified empirically. Antediluvian time could be regarded as virtually limitless, the 'days' of Creation being of unspecified duration. De Luc's empirical work was, in fact, leading him towards a date corresponding with what we regard as the end of the last ice age. Nature was playing him a cruel trick!

At Oxford, there had been lectures in geology for many years. The field fell into desuetude in the late eighteenth century but the Reverend William Buckland gave popular lectures there from 1814 to 1849, and particularly in his early years he endeavoured to show linkages between Cuvierian geology and Biblical history, identifying Cuvier's last catastrophe with the Noachian Flood. Buckland specially interested himself in cave deposits, supposing that the animal bones found therein might be the remains of animals that had sheltered in the caves or had been washed

into them. His lectures were popular, in that they seemed to show that Biblical history could find support in the latest findings of science. The convenient occurrence of coals, limestones, and ore bodies supposedly evinced divine wisdom. Regarding the Flood, Buckland believed that it was an historical event of divine origin and humans were of recent origin, as described in Mosaic history (though, following Cuvier, he denied that there were any evidences of fossilized antediluvian humans, such as might be expected if Mosaic history were correct). But, along with this 'natural theology', Buckland was a dedicated scientist, doing much fieldwork. He showed that the marks on his cave bones were compatible with those left by the teeth of modern hyenas. He even let a tortoise walk across pastry, claiming that the marks it made were comparable to those found on sedimentary rocks from Scotland, suggesting that the latter were imprints of an ancient reptilian analogue. This was a kind of use of Cuvier's 'comparative method'. Buckland was one of many who made geology a highly popular science in the early nineteenth century.

Lyell and Uniformitarianism

But Cuvier's and Buckland's geology was an anathema to the followers of Hutton, notably Charles Lyell, in his *Principles of Geology* (1830–1833) (*see* **Famous Geologists:** Lyell). Lyell, a lawyer turned writer and geologist, envisaged an open-ended time-scale, without direction to geological change. Land could rise and fall in an irregular fashion, with consequent climatic changes, as high land happened to be chiefly near the poles or the equator. With changing conditions of existence, new species would be formed (by an unknown process – perhaps God-caused?) or become extinct, the forms approximately tracking the changing conditions. Hence, the proportions of extant forms would gradually decrease into the past, so that the percentage of present types could be used to provide divisions of Tertiary time. Cuvier's great palaeontological breaks were thought by Lyell to be due the loss of strata by erosion; or not all living forms having been preserved. Lyell was a gradualist and a non-progressionist. He was also dubbed a 'uniformitarian' or 'actualist', i.e., he thought that the proper mode of reasoning in geology was to explain the past by using analogies with the knowable circumstances and constant laws of nature of the present. (Use analogies from knowledge of circumstances as they *actually* are; or use 'the present as the key to the past'—an aphorism not coined by Lyell himself.)

An important example of Lyell's actualist geology is provided by his examination of Etna to argue for the great age of the Earth. He could observe recent eruptions and lava flows and historical records of eruptions. Hence he had an idea of the rate of accumulation of lava and the build-up of the volcano. Knowing the height of the mountain and its approximate rate of formation (*assumed* to be constant), he could estimate its approximate age: some hundreds of thousands of years, which was very ancient in human terms. But the lava flows lay on sediments containing fossils of *recent* appearance, similar to organisms in today's Mediterranean. Therefore, the sediments at the *top* of the stratigraphic column were very *old* in human terms. Therefore, the Earth itself must be exceedingly ancient. In this manner Lyell argued for the Earth's enormous antiquity. Geological time had been 'proved', on the basis of uniformitarian argument. It should be remarked that Cuvier, Buckland, and Lyell all used 'actualist' arguments, even though the first two were 'catastrophists' while Lyell was a 'uniformitarian'. Thus, there was some unity within geological thinking, even though many have thought that the early nineteenth century was divided into two hostile camps (as was undoubtedly true to an extent). Lyell was a religious man, and believed that humans were divinely created. But he had no truck with such ideas as universal catastrophes or the 'thread of operations' being broken.

Mountain Building

Lyell was a Huttonian, and accepted the idea of a hot interior to the Earth, and Hutton's cyclic geohistory. But he did not have much to say about the formation of mountains other than the build-up of lava-cones such as Etna. In 1829–1830, however, an influential theory of mountain formation was published in France by Léonce Élie de Beaumont ('*Recherches sur quelques-unes des revolutions de la surface du globe*'). He envisaged a slowly cooling and contracting Earth, with the contraction of the fluid interior leading to occasional bucklings of the solid external crust (like the wrinkles on the skin of a drying apple, we might say), and hence the formation of linear mountain ranges and ocean trenches. It was suggested that the several mountain ranges, running in different directions, might represent different 'wrinklings' of the crust (or orogenies), occurring at different times. The idea was that parallel mountain ranges might have formed at about the same time. Consequently, one could develop a kind of 'tectonostratigraphy'. The theory was subsequently developed, with the idea of a vast geometrical pattern of foldings eventually forming around the globe. It became especially influential in France, and was the forerunner of tectonic theories subsequently developed in the nineteenth century, such as that of Eduard Suess (*see* **Famous Geologists:** Suess), which envisaged

cooling and contraction as the major cause of mountain formation. In essence, cooling/contraction theories held the field until the arrival of 'drift' and plate tectonic ideas in the twentieth century.

Minerals, Rocks, and Crystals

Werner and his school concentrated on the study of the external features of minerals, and developed elaborate schemes for their description and classification. The late eighteenth century saw the emergence of chemistry successfully applied to minerals. The older methods of pyro-analysis with the help of the blowpipe, though useful in the field, could give little quantitative information. But in Sweden Torbern Bergman (1784) published a general method for the chemical analysis of gems. They could be brought into solution by fusion with alkali and then, by a sequence of precipitation reactions, and heating and weighing the several products, the different constituent 'earths' (silica, magnesia, alumina, lime, etc.) could be ascertained as percentages. Bergman's results were inaccurate, but the principles of his procedure were valid and were soon applied more successfully by chemists such as Richard Kirwan, Nicholas Vauquelin, Martin Klaproth, and Jons Jacob Berzelius, both to minerals and rocks. Aided by Lavoisier's theory of elements as simple substances, obtained as the last terms of chemical analysis, mineralogy had a satisfactory theoretical and practical basis for chemical understanding. But old problems remained. A substance of one chemical composition could have many different mineral forms and substances of similar crystalline form could have numerous different chemical compositions. The question of the best way to characterize mineral species remained contentious. Geology *per se* did not take a great leap forward through the progress in chemical mineralogy before 1830.

In petrology, the distinction between bedding and cleavage was understood by the English geologist Adam Sedgwick by the 1820s, but he probably learnt it from quarrymen. Following Ami Boué (1819), the category of metamorphic rocks was introduced by Lyell (1833): '*altered* stratified' rocks—the alteration being due to heat and pressure. He referred to 'hypogene' (formed-at-depth) rocks, instead of 'primary' or 'primitive', and he divided them into those that were 'unstratified' (plutonic, e.g., granite) and 'stratified' (metamorphic, e.g., gneisses or schists).

In crystallography, the most important contributions came from the Frenchman René-Juste Haüy, the Englishmen William Wollaston, William Whewell, and William Miller, and the German Eilhard Mitscherlich. Haüy (1784, 1801, 1822) supposed that crystals were made up of a small number of fundamental '*molécules intégrantes*' (tetrahedron, triangular prism, and parallelipipedon), which could be revealed by crystal cleavage and the 'conceptual analysis' of crystals. From these starting points, he hypothesized the 'building' of many different crystals forms from similar basic building blocks, according to assumed rules of decrement for the addition of the 'integrant molecules'. His reasoning was in part circular, but it gave intelligibility to crystallography. Haüy's 'integrant molecule' foreshadowed the modern chemical concept of molecule.

Haüy used contact goniometers, which were of limited accuracy. Wollaston (1809) devised the more accurate reflecting goniometer, and its increased precision led him to question Haüy's methods and results. But Whewell (1824), developing Haüy's concepts, was able to use co-ordinate geometry to describe crystals, arriving at the equations $x/h + y/k + z/l = 1$ or $px + qy + rz = m$ to represent crystal faces, all coefficients being integers. The indices p, q, and r are now known as the Miller indices, being reciprocally related to the co-ordinates of a vector perpendicular to the plane of a crystal face. By such analysis, crystallography could become mathematized and quantifiable, while geology remained in an 'historical' and largely qualitative mode. Mitscherlich was responsible for introducing the concepts of isomorphism, dimorphism, and polymorphism, which assisted understanding of the complexities of empirical mineralogy.

Volcano Theory

Chemistry also offered ideas about the Earth's internal heat. With the discovery of the alkali metals by Humphry Davy (1807), the suggestion was made that the heat might be generated by the action of water penetrating into subterranean stores of these metals, sufficient to produce volcanic eruptions. This accorded with the idea that volcanoes might be produced by the expansion of gases within the Earth, causing localized 'swellings' of the crust (theory of 'craters of elevation' as advocated by Alexander von Humboldt and Leopold von Buch). There was extensive controversy concerning this issue, but Lyell's theory of volcanoes being produced by successive accumulation of lava flows (or ash emissions) eventually prevailed. Chemical theories of the Earth's heat gradually declined in the nineteenth century, but improved suggestions were not really forthcoming until the twentieth century.

See Also

Biblical Geology. **Famous Geologists:** Cuvier; Darwin; Darwin; Hutton; Lyell; Sedgwick; Smith; Steno; Suess. **History of Geology Up To 1780**.

Further Reading

Ellenberger F (1994) *Histoire de la géologie Tome 2: La grande éclosion et ses prémices 1660 1810*. London and New York: Technique et Documentation (Lavoisier).

Laudan R (1987) *From Mineralogy to Geology: The Foun dations of a Science, 1650 1830*. Chicago & London: The University of Chicago Press.

Lewis CLE and Knell SJ (eds) (2003) *The Age of the Earth: From 4004 BC to AD 2002*. London: The Geological Society.

Oldroyd DR (1979) Historicism and the rise of historical geology. *History of Science* 17: 191 213, 227 257.

Oldroyd DR (1996) *Thinking about the Earth: A History of Ideas in Geology*. London: Athlone Press; Cambridge (Mass): Harvard University Press.

Oldroyd DR (1998) *Sciences of the Earth: Studies in the History of Mineralogy and Geology*. Aldershot, Brooklands, Singapore and Sydney: Ashgate Variorum.

Rudwick MJS (1969) Lyell on Etna, and the antiquity of the earth. In: Schneer CJ (ed.) *Toward a History of Geology*, pp. 288 304. Cambridge (Mass) & London: MIT Press.

Rudwick MJS (1997) *Georges Cuvier, Fossil Bones, and Geological Catastrophes: New Translations & Interpret ations of the Primary Texts*. Chicago & London: The University of Chicago Press.

Torrens HS (2002) *The Practice of British Geology 1750 1850*. Aldershot and Burlington: Ashgate Variorum.

HISTORY OF GEOLOGY FROM 1835 TO 1900

D R Oldroyd, University of New South Wales, Sydney, Australia

National Geological Surveys

The year 1835 is important in the history of geology, being the date when the first national survey was initiated in England and Wales by Henry De la Beche. There had been earlier 'private' surveys, such as that of John Macculloch in Scotland and Richard Griffith in Ireland, but De la Beche's was the first national survey to receive direct government funding. So with De la Beche, geology (as opposed to mining engineering) became a paid profession rather than a gentlemanly pastime or learned avocation. His enterprise was possible because by 1835 Britain had a good set of 6-inch 'Ordnance Survey' maps, produced for military purposes. Mapping was extended to Scotland in the 1860s, a Scottish Branch being established there (1867) under Archibald Geikie (*see* **Geological Surveys**).

The early British survey was an almost single-handed effort, but by the 1840s new staff were being taken on and the band of 'hammerers' gradually spread their work across the country from the south-west. By the end of the century, England, Wales, and Ireland had all been covered, and much of Scotland. The objective was to identify different divisions of the stratigraphic column and represent them on cut-up portions of the ordnance maps (field-slips), and then enter the information on full-sized maps, before reduction to 1-inch maps, which were issued hand coloured until the end of the nineteenth century. The survey thus had the task of *standardizing* nomenclature and colour symbolism for the nation's geology. Large collections were amassed and specimens exhibited at the fine Geological Museum in Jermyn Street, London, with which was associated the Royal School of Mines.

By 1900, most American states had established surveys, as had the main British colonies and the leading countries of Europe, Argentina, and Japan. The US Federal Survey was established in 1879. To an extent, stratigraphy worldwide was dominated by the ideas of British geologists, building on the work of William Smith (*see* **Famous Geologists:** Smith). However, there was confusion in nomenclature and difficulty in international correlations, for there was no reason in principle why, if Lyellian geology were correct, the stratigraphic columns should correspond in different parts of the world. The International Geological Congresses, the first of which was held in Paris in 1878, had as one of their main goals the rationalization and co-ordination of international stratigraphic nomenclature. But this project was hardly successful in the nineteenth century.

Stratigraphy

The major subdivisions (eras) of the stratigraphic column (Palaeozoic, Mesozoic, Kainozoic) were proposed by John Phillips (1840). The periods were proposed as follows: Carboniferous (Conybeare/Phillips in 1822); Cretaceous (d'Omalius d'Halloy in 1822); Eocene, Miocene, and Pliocene (Lyell in 1833);

Triassic (Alberti in 1834); Silurian (Murchison in 1835); Cambrian (Sedgwick in 1835); Devonian (Sedgwick and Murchison in 1839); Jurassic (von Buch in 1839); Pleistocene (Lyell in 1839); Permian (Murchison in 1841); Oligocene (von Beyrich in 1854); and Ordovician (Lapworth in 1879). The Precambrian was suggested by Jukes in 1862. The 'Tertiary' (Arduino in 1760) survived for the units Eocene–Pliocene. Some of the units (e.g. Cretaceous) had previously been recognized by their lithologies, but were not formally 'introduced', with palaeontological criteria, before the foregoing dates.

The introduction of several of the Periods involved well-known geological controversies. Notably, there was a battle between the 'professional' De la Beche on the one hand, and the 'amateur' gentlemanly geologists Adam Sedgwick (*see* **Famous Geologists:** Sedgwick) and Roderick Murchison (*see* **Famous Geologists:** Murchison) on the other, over the establishment of the Devonian. The Old Red Sandstone (ORS) had long been recognized as a distinctive red sandstone unit, but there were marine rocks in Devonshire that seemingly had a similar age. De la Beche argued on lithological and structural grounds that the Devon rocks were not relatable to the ORS, but his opponents successfully argued otherwise using palaeontological criteria.

But then Sedgwick and Murchison fell out, with even greater rancour, over the Welsh strata. Sedgwick studied the rocks of north-west Wales and envisaged a Cambrian system there. It was not, however, well characterized by distinctive fossils. Murchison started from the fossiliferous Welsh Border region and worked towards Sedgwick's territory. They failed, however, to establish a clear section, or boundary, between the two systems and Murchison began to extend his Silurian downwards, eventually extending it to the time when shelly fossils first appeared. Sedgwick, on the other hand, maintained the integrity of his Cambrian and tried to extend his Cambrian upwards into the Silurian domain. The controversy was only resolved after their death by the schoolmaster geologist, Charles Lapworth, who proposed (1879) a threefold subdivision of the Palaeozoic, by analogy with a similar threefold division described for Bohemia by the palaeontologist Joachim Barrande. The 'debatable' ground in the Cambrian and Silurian became the new System, the Ordovician.

Finer biostratigraphic subdivisions (stages) were established by Alcide d'Orbigny in France and Albert Oppel in Germany. D'Orbigny was a grand traveller in France and South America, and amassed huge collections; while Oppel particularly made a detailed survey of the available literature to establish what fossils occurred in which parts of the various systems, notably the Jurassic. It became evident that the Jurassic (for example) could be subdivided into stages, each with its own characteristic suite of fossils; with even finer subdivision possible into zones. D'Orbigny's 27 stages were named according to the localities chosen to 'define' them (by standard sections). Zones, with their characteristic fossil assemblages, were named by Oppel according to their most characteristic (or index) fossil species. Ideally, index fossils should be of wide geographical and short temporal range. Ammonites served this purpose well for the Jurassic.

The work of d'Orbigny and Oppel was fundamental for stratigraphy (and the aforementioned mapping), but it depended on the existence of 'breaks' in the stratigraphic sequences, such as Georges Cuvier had previously envisaged (*see* **Famous Geologists:** Cuvier). In fact, d'Orbigny willingly accepted this 'catastrophist' stratigraphy, but did not attribute metaphysical significance to the 'breaks'. They were simply useful for the practical purpose of stratigraphic subdivision and delineation of strata. It did seem, however, that the fossil record manifested some kind of 'progress', with organisms in the stratigraphic column gradually becoming more like those found alive today.

Darwinism and Evolution

As is well known, Darwin published his theory of evolution in 1859 (*see* **Famous Geologists:** Darwin). It explained the nature of the fossil record well in some respects ('progress', extinction, and the appearance of new forms). But it did not lead one to expect that fossil transitions would be quite 'jerky'. To account for the observed fossil record, Darwin took the view that there were innumerable gaps, where organisms had not been preserved in the first place, or had been lost by subsequent erosion. This point was neither provable nor disprovable, and thus to some degree the catastrophist/gradualist distinction remained metaphysical in character. Even now, smooth 'trends' in fossil forms are rare, but have been found for foraminifera, where numerous specimens may be found in small thicknesses of sediment. (D'Orbigny made important collections of forams but did not use them to discuss this theoretical issue.) In his later work, Darwin increasingly deployed Lamarckian ideas, supposing that the environment 'caused' increased variation (as apparently occurred under domestication), and that acquired characters were inherited. In the late nineteenth century, a significant number of writers turned away from 'classical' natural selection theory and supposed that the stratigraphic record revealed 'directedness' towards apparent goals (e.g., greater size) in a process that apparently occurred independently of natural selection. Where organisms were

seemingly 'trying' to adapt themselves to the conditions of existence, as it were taking charge of their own destinies, this could be construed as a version of Lamarckism. So there were neo-Lamarckians, such as the American palaeontologist Edward Drinker Cope. A variant of their theory was 'orthogenesis' (straight-line evolution), advocated by the German biologist Theodor Eimer. It suggested that evolutionary trends might continue until they became maladaptive, as perhaps in the evolution of the Irish elk to the point that its horns became so heavy as to drive the species to extinction. The empirical background to such ideas was the discovery of gigantic dinosaur skeletons in America by (among others) Cope and his bitter rival, Othniel Marsh.

Glacial Theory

Lyell's theory allowed for periods of terrestrial warming or cooling according to where the highest mountains happened to be on the Earth's surface at any given time. But there was no 'direction' to the process. In the early nineteenth century, some little-known Swiss men, observing moraines, etc., drew attention to the former extent of the Swiss glaciers. This suggestion was picked up by the Neuchâtel professor Louis Agassiz (1837) (*see* **Famous Geologists:** Agassiz). His idea was that the Earth was cooling, but did so in such a way as to fall, at times, below the temperature of an ordinary cooling curve, and then reverted to the 'normal' temperature of a cooling body. So not long before the present (in geological terms) the temperature could have been significantly lower than today, low enough to produce an 'ice age', with widely extended glaciers. The hypothesis could explain many of the curious phenomena of the superficial deposits of northern Europe: the spread of 'boulder clay' over the northern plains; vast gravel deposits in the valleys running north from the Alps; scratches on rock surfaces now without ice cover; boulders distant from places where such rocks occurred *in situ*, etc. So Agassiz promoted the idea of an *Eiszeit* or Ice-Age, which could account for phenomena formerly explained by the action of Noah's Flood. Agassiz attended a British Association meeting in 1840 and attracted some converts, notably William Buckland and even (temporarily) Lyell.

But the land-ice theory did not receive immediate acceptance. Icebergs were also proposed to transport 'erratic' boulders. So arose the 'glacial submergence' theory: that there was simultaneous global cooling and lowering of land surfaces *or* rise in sea levels. The iceberg hypothesis gave rise to the notion of 'drift' deposits. Loose marine shells were found atop some hills in North Wales, and some submergence seemed necessary for them to have got there. Agassiz's land-ice theory seemed incredible and incompatible with uniformitarian doctrine.

However, the land-ice theory began to make more progress after about 1860 when it was taken up by the British Surveyor Andrew Ramsay. For several years it contended with the glacial-submergence doctrine, eventually winning out over the latter. In Switzerland, the stratigrapher Adolf Morlot (1856) noted what appeared to be multiple glacial deposits near Lausanne, occurring in the 'Quaternary' (so named by Paul Desnoyers in 1829 for deposits of the Seine Basin, thought to be younger than Tertiary). The idea of four major glaciations (Günz, Mindel, Riss, and Würm) became almost paradigmatic through the publications of Edouard Brückner and Albrecht Penck (1901–1909) on the outwash gravels of the Alps and the Pyrennees, having been given an attractive astronomical explanation by James Croll (1875) in terms of the changing ellipticity of the Earth's orbit and precessional motion. The fourfold Quaternary glaciation was repudiated in the twentieth century, to be replaced by the more complicated theory of Milankovich cycles. But already in the nineteenth century glacial theory had largely solved the riddle of 'diluvial' deposits and put paid to a *global* Flood as a geological agent.

Geomorphology and Landforms

Attention to landforms and river patterns in the nineteenth century, coupled with ideas about slow land elevations and denudations, allowed explanation of many geomorphological peculiarities. Joseph Jukes in Ireland (1862) emphasized the role of rivers (rather than the sea) in eroding away its former Carboniferous cover, producing river patterns that could not have been caused by the sea (though the rivers supposedly eroded a surface that had undergone marine peneplanation). He also enunciated the principle that rivers transverse to a geological structure are generally older than their longitudinal or strike side branches. Rivers excavate their own valleys, but adjust their courses to fit underlying geological structures. At about the same time Ramsay successfully argued that glaciers could have excavated rock basins in relatively soft strata, forming lakes that are presently being filled with sediment after the end of the glacial epoch.

Notable advances in geomorphological understanding were made by the geological explorations of the American West by the likes of J.W. Powell, G.K. Gilbert, C.E. Dutton, and W.M. Davis. Powell in 1875 had the idea of 'base-levelling', and that drainage patterns could be older than the mountains through which the rivers run. Thus rivers could cut down at a rate equivalent to that at which land was

rising, explaining otherwise anomalous drainage patterns. Gilbert introduced the idea of graded rivers and tried to educe laws governing the sculpture of the Earth's surface. Dutton in 1882 drew attention to the influence of differences in the hardnesses of strata on the cross-sections of river valleys and hence the development of canyon profiles. Davis, considering evidence from the Appalachians (1889), wrote about the 'life-cycles' of land-forms, from 'youth', through 'maturity', to 'old age'. But rivers could be 'rejuvenated' by land elevation. He spoke of 'antecedent drainage' and in general considered the *evolution* of landscapes almost as if they were living entities.

In 1859, back in England, Darwin had given greater emphasis to marine erosion, proposing an excessive estimate of the age of the Weald valley near his home on the assumption that it was cut by the sea. Earlier, during his *Beagle* voyage, he had successfully explained the origin of coral atolls by supposing that coral could grow upwards at about the rate that land was subsiding, hence explaining the peculiarities of fringing reefs.

Mountain Formation and Isostasy

The idea of the major features of the Earth's surface being due to cooling, contraction, and wrinkling of its crust dominated the nineteenth century from the work of Léonce Élie de Beaumont and Eduard Suess particularly. But the work of the American James Dwight Dana was probably more influential worldwide than that of Élie de Beaumont. Well travelled through his participation in the Wilkes expedition, in 1847 Dana recognized a fundamental difference between continents and ocean basins, thinking that they formed early in the planet's history. North-west and north-east trending island chains supposedly marked 'cleavage lines' originating back in Archaean (primaeval) times, which still influenced the evolution of the crust. The continents cooled and solidified first, whereas ocean basins were situated where subsequent cooling and contraction were concentrated and where volcanoes were still chiefly active. Basin subsidence caused lateral pressure, folding, and uplift of the continental margins.

In 1856 Dana envisaged growth of the North American continent, starting from the V-shaped ancient core or Azoic nucleus of the metamorphic rocks of the Hudson Bay region, to which additions were successively made from the south-east and south-west. (This was the forerunner of the twentieth century concept of 'cratons'.) Further, he thought of continental interiors as relatively stable, so that folding and faulting were concentrated at their margins, as exemplified by the Appalachian range. The whole process of contraction and accession of new land was supposedly divinely guided or teleological.

Dana was challenged by his countryman in 1859, James Hall (*see* **Famous Geologists:** Hall), who saw vertical movements of the crust as responses to gravitational loading. Sediment could be deposited in long trenches (later called geosynclines), parallel to the continental margins. Global contraction caused crumpling of the upper sediments, while the downward-bulging trench bottoms would be fractured and intruded by igneous matter. Linear mountain ranges such as the Appalachians might have accumulated their sediments in geosynclinal structures, but the process of their uplift following sediment deposition was obscure. Élie de Beaumont's lines of mountain elevation were Hall's lines of original accumulation. But, as Dana in 1866 complained, Hall offered a "theory of the origin of mountains, with the origin of mountains left out".

In 1873 Dana coined the term 'geosynclinal' (later geosyncline), and its complement 'geanticlinal'. The evolution of the two, and concomitant growth of a continental margin as envisaged by Dana, is illustrated in Figure 1.

Meanwhile, in Switzerland, and especially in the Glarus Canton, geologists such as Arnold Escher had, since the 1840s, been finding substantial evidence for lateral earth movements, and seeming inversions of the usual order of strata, according to the palaeontological evidence. In 1878 Albert Heim proposed a great double fold for the Glarus region with two mountain masses moving together. He was addressing a real structural problem, but his solution was mechanically implausible. The great synthesis for the Alps, and, indeed, worldwide, was provided by the Austrian geologist Suess's *Das Antlitz der Erde* (1883–1909) (*see* **Famous Geologists:** Suess).

By Suess's theory, global contraction gave rise to subsidence in parts of the Earth, with generation of tangential forces, manifested as thrust faults. Thus, the Alps might thrust northwards over the 'foreland' region of Germany, while in the 'backland' of the Mediterranean and Adriatic seas there could be further collapse and volcanic activity. Likewise the Carpathians could ride over the Russian foreland. But in China the lateral movement was southward, as in Yunnan. Suess's theory was also linked to stratigraphy. Collapses of ocean floor regions would cause worldwide marine *regressions*. But an oceanic collapse would stimulate erosion of the more exposed land surfaces, and the increased sediment supply would fill up the basins and produce marine *transgressions*. So there would be cycles of erosion and deposition, and since all the oceans interconnected the

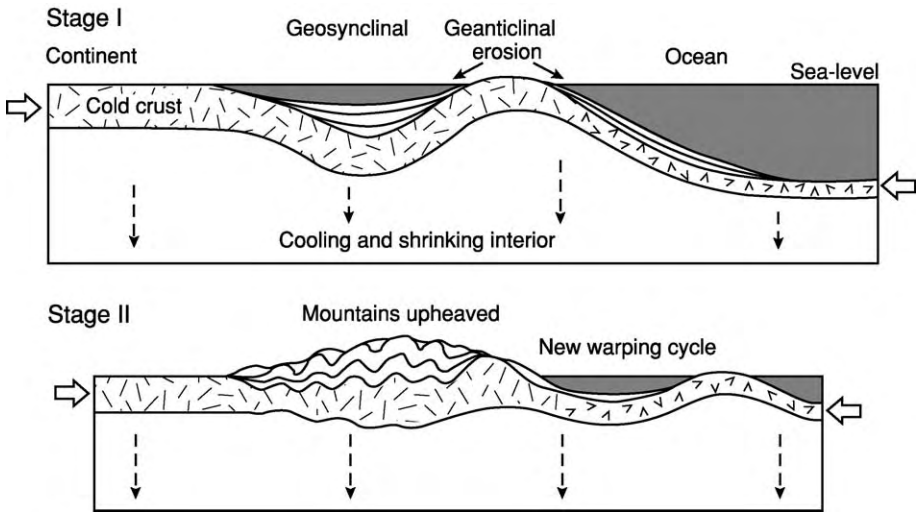


Figure 1 Representation of Dana's 1873 theory of mountain building, as reconstructed by Robert H. Dott (1997, p. 299). Reproduced by courtesy of the *American Journal of Science*.

worldwide nature of the major divisions of the stratigraphic column could be explained. In 1888, Suess thus introduced the notion of 'eustatic movement' or eustasy.

Additional ideas were furnished by the trigonometric survey conducted in India in the nineteenth century, when it was discovered that there was a discrepancy between the results of astronomical and geodetic determinations of latitude along the line of meridian being determined. In 1855, the mathematical Archdeacon of Calcutta, John Henry Pratt, ascribed this to the Himalayas' attraction of the plumb-line used in the astronomical determinations. The point was noticed by the Astronomer Royal, George Airy, who thought that mountain ranges might have underlying 'roots' in the Earth's fluid interior, with mountains being held up rather like floating icebergs. Pratt also assumed a balance between Archimedean upthrust and the weight of crust, but the two models were different. Pratt assumed a common 'depth of compensation', with the less dense parts of the crust standing higher, while Airy's model had deeper roots for the higher mountains.

The general idea was taken up by Dutton in 1871 and in 1892, who introduced the term 'isostasy' or 'equal standing'. An inhomogeneous spinning earth would not be perfectly spheroidal but would have depressions where the crust was more dense and bulges where it was less dense. The form of the Earth (geoid), determined for North America by the US Coast and Geodetic Survey and eventually published early in the twentieth century by John Hayford, suggested that the Earth was generally in a state of isostatic balance. But the Pratt model was assumed for ease of computation, and it entered the thinking of William Bowie (who succeeded Hayford) and subsequently stood in the way of acceptance of continental drift theory.

The Formation and Age of the Earth

After Lyell, most geologists were willing to accept an indefinitely large age for the Earth, and the Biblical age was only upheld by theologians and non-geologists. Darwin thought he could assume as much time as he required for the processes of erosion and deposition and organic evolution. His figure of 1859 for the time taken for the marine erosion of the valley of the Weald was some 300 million years, a figure with which he and other geologists were comfortable. However, in 1862 the physicist William Thomson (Lord Kelvin) argued that the sun acquired its initial heat from meteoric impacts, and had subsequently been cooling. It might be somewhere between 10 and 500 million years old. He estimated that a cooling Earth with temperature gradient and heat loss as at the present could be 20 to 400 million years old. In 1871 he also suggested that if the Earth acquired its spheroidal shape when it was still molten, one could estimate its rotation rate at the time of solidification. Subsequently, it had been slowing its rotation, due to tidal friction, to the present value. By estimation of the decrease of rotation rate and thermal contraction it seemed that the age might be of the order of 100 million years. This was just about acceptable to evolutionists and geologists, but both would have preferred more time, while being unable to argue with Kelvin's calculations. The matter rested thus until 1909, and the work of John Joly, where the effect of radiometric heat was considered and the

supposed rate of terrestrial cooling was seen to need revision. Thus, physical objections to the great age of the Earth, which seemed substantial in the second half of the nineteenth century, subsequently receded.

The eighteenth-century theory of the Earth's origin was that of Laplace and Kant, which supposed a coalescence of matter under gravity from a primaeval spinning 'cloud' of nebular matter. Thus, the Earth originated as a spheroid of hot gas, which cooled to a liquid and then to a solid state. This had dynamical problems, and at the end of the nineteenth century the influential American geologist, Thomas Chamberlin, in collaboration with the astronomer Forest Moulton, proposed their theory of 'planetesimals' (miniscule planets), according to which matter dragged out from the spinning sun by some other star could have accreted to form small solid bodies, which could have collided to form the several planets. This was the beginning of planetary geology, or the conflation of geology, cosmology, and astronomy.

Rocks and their Formation

There were few major contributions to sedimentary petrology in the period here under review. The Irishman Patrick Ganly's 1830s discovery of the 'wayupness' criterion offered by current bedding was not utilized until the twentieth century. In 1839 Christian Ehrenberg published his microscopic studies of chalks and limestones. Formal distinction of sedimentary, igneous and metamorphic rocks was made by Henri Coquand in 1857.

Various igneous rocks such as basalts, granites, gabbros, syenites, or porphyries had been recognised since antiquity and many classifications were proposed in the nineteenth century, according to chemical and/or mineral composition, texture, or supposed mode of formation, but the field was confused. Examination of rocks in thin section by Henry Sorby assisted in a sense, but the proliferation of information also added to the confusion. While Huttonian theory was triumphant as regards Werner's original theory, there was continued interest in the role of water in the formation of igneous and metamorphic rocks. Notably, in 1857 and 1859 Gabriel Auguste Daubrée of the French Mines Department subjected materials to high temperatures and pressures, with or without water, and concluded that new minerals could crystallize without wholesale melting. He thought that past conditions could have been radically different from those at present and that there might have been an *ocean primitif*. Granite was thought to be produced by 'aqueous plasticity', not igneous melting. At Freiberg, Bernhard von Cotta held analogous views. But in 1860 Daubrée thought that foliations were due to pressure during 'regional metamorphism'.

The variety of igneous rocks raised the question whether there were different kinds of subterranean magma, or whether some processes of differentiation occurred from essentially the same starting material. In 1844 Darwin had the idea of differentiation of magma by gravity settling of first-formed crystals; and in 1846 he distinguished cleavage, foliation, and stratification (while regarding gneisses as stratified rocks). Dana thought that differentiation of magma might precede crystallization. By contrast, on the basis of observations in Iceland, in 1851 and 1853 the chemist Robert Bunsen proposed that there were two separate magma chambers under the island, producing 'trachytic' and 'pyroxenic' rocks or intermediate mixtures. In 1853 Wolfgang Sartorius von Waltershausen hypothesized the existence of different subterranean zones; and the eruption of more siliceous types preceded the more basic, thus relating igneous compositions to age. In 1857 Joseph Durocher asserted the liquation model and priority for the idea over Bunsen. Metallic lodes were ascribed to '*émanations*'. Following work in Hungary and California, Ferdinand von Richthofen envisaged a succession of magmas, the earliest being more siliceous, the great outpourings of Tertiary basalts being due to earlier depletion of siliceous magma. In 1878 in America, Clarence King thought pressure release could facilitate fusion. In 1880 Dutton suggested that fusion could follow pressure release, local temperature elevation, or water absorption.

The master petrologists of the period were Harry Rosenbusch in Strasbourg, Ferdinand Zirkel in Leipzig and Ferdinand Fouqué and Auguste Michel-Lévy in Paris, who specialized in the study of feldspars. All were adept with the use of the petrographic microscope. In 1873 Rosenbusch published a catalogue of all then known magmatic and metamorphic rock types. Rosenbusch's 1877 study of metamorphism around the Barr–Andlau granite in the Vosges was important for his recognition of zones of contact metamorphism (schists, knotted schists, hornfelses), seemingly without feldspars. But Michel-Lévy found feldspars in the contact aureole of the Flamandville granite in Normandy and he and Fouqué thought there was no fundamental distinction between contact and regional metamorphism. Throughout this period Continental petrologists continued, in the Wernerian tradition, to try to find relationships between age and 'hard-rock' composition, whereas their British counterparts chiefly concerned themselves with biostratigraphy.

However, in 1893 the British surveyor George Barrow, working on metamorphic rocks in the southern Scottish Highlands, found characteristic

metamorphic minerals (sillimanite, kyanite, and staurolite) around a granitic mass, and this gave him mappable subdivisions of the region. These were developed in the twentieth century as 'Barrovian zones', but the useful idea was not initially followed up.

Experimental petrology was undertaken in the nineteenth century, but until high-pressure and pressure techniques were developed in the twentieth century for simulating rock formations, work on phase diagrams was developed, and ideas about the structure of the Earth's interior could be pursued through seismology, petrological understanding remained speculative and somewhat at the level of natural history.

See Also

Famous Geologists: Agassiz; Cuvier; Darwin; Hall; Lyell; Murchison; Sedgwick; Smith; Suess. **Geological Maps and Their Interpretation**. **History of Geology From 1780 To 1835**. **Metamorphic Rocks:** Facies and Zones. **Sedimentary Processes:** Glaciers. **Stratigraphical Principles**. **Time Scale**.

Further Reading

Berry WBN (1968) *Growth of a Prehistoric Timescale Based on Organic Evolution*. San Francisco and London: W.H. Freeman & Co.

Bowler PJ (1976) *Fossils and Progress: Paleontology and the Idea of Progressive Evolution in the Nineteenth Century*. New York: Science History Publications.

Buffetaut E (1987) *A Short History of Vertebrate Palaeontology*. London, Sydney and Wolfeboro: Croom Helm.

Burchfield JD (1975) *Lord Kelvin and the Age of the Earth*. New York: Science History Publications.

Cross W (1902) The development of systematic petrography in the nineteenth century. *The Journal of Geology* 10: 331–376.

Davies GH (1969) *The Earth in Decay: A History of British Geomorphology 1578–1878*. London: Macdonald Technical and Scientific.

Dott RH (1997) James Dwight Dana's old tectonics—global contraction under divine direction. *American Journal of Science* 297: 283–311.

Gohau G (1997) Évolution des idées sur le métamorphisme et l'origine des granites. In: Bonin B, Dubois R, and Gohau G (eds.) *Le métamorphisme et la formation des granites evolution des idées et concepts actuels*, pp. 9–58. Paris: Nathan.

Greene MT (1982) *Geology in the Nineteenth Century: Changing Views of a Changing World*. Baltimore: Johns Hopkins University Press.

Nieuwenkamp W (1977) Trends in nineteenth century petrology. *Janus* 62: 235–269.

Oldroyd DR (1996) *Thinking about the Earth: A History of Ideas in Geology*. London: Athlone Press; Cambridge (Mass): Harvard University Press.

Rudwick MJS (1985) *The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists*. Chicago and London: The University of Chicago Press.

Secord JE (1986) *Controversy in Victorian Geology: The Cambrian Silurian Dispute*. Princeton: Princeton University Press.

Yoder HS (1993) Timetable of petrology. *Journal of Geological Education* 41: 447–489.

Young DA (2003) *Mind over Magma: The Story of Igneous Petrology*. Princeton and Oxford: Princeton University Press.

Zittel KA von (1901) *History of Geology and Palaeontology to the End of the Nineteenth Century*. Translated by Maria M. Ogilvie Gordon. London: Walter Scott.

HISTORY OF GEOLOGY FROM 1900 TO 1962

D F Branagan, University of Sydney, Sydney, NSW, Australia

Introduction

The period prior to the plate tectonics revolution of the 1960s has been said by some historians of science to have been a time of stagnation for geology. This supposed stagnation is based on the idea, then largely held, of the fixity of the continents and oceans, which some have extended to suggest that geologists in the main remained rather fixed in their ideas and were concerned only with mundane geological matters. Was this so? It might be partly true, in that only a few people were attending to 'large questions'. However, many unsolved geological problems were studied, and one could argue that these had to be tackled before fundamental concepts could be challenged. The first half of the twentieth century was marked by two world wars and the disruption of scientific contact for much longer than just the war years. However, even in these years two things happened that would benefit geology: techniques were developed that allowed the quantification of many aspects of geology; and geology was increasingly applied to engineering problems.

Aerial photography is a prime example. It led to more rapid geological mapping, and by the mid-1920s it was applied to the search for gas- or oil-yielding structures. Later, systematic photographic coverage contributed to the understanding of regional, and even supra-regional, problems. Likewise the geophysical instruments used to detect spatial changes in rock strata were improved through research during the First World War. Geologists during that war introduced environmental geological maps, the forerunners of many variations of the geological map later developed by engineering geologists. Attention was also given to locating strategic minerals, especially during the Second World War. In the interwar period, explosion seismology was used in oil prospecting (first in Oklahoma in 1921), and other geophysical techniques were brought into use. Although modern ocean research had begun in the 1930s, the development of radar during the Second World War quickly produced the first significant information about the ocean floors, seamounts, and deep trenches. The accumulation of data required more technical expertise, and in general the 1950s saw the rise of 'team' efforts and multiauthored publications. To some extent, this heralded the demise of the geological polymath, and few people attempted to generalize from the new information that was forthcoming, preferring to be one of a consensual group.

The Age of the Earth, and its Subdivisions

Determining the age of the Earth was perhaps the most significant achievement of geological research in the early twentieth century. The geological significance of radioactivity was recognized in 1903, when Pierre Curie and his researchers found that radium salts release heat constantly, and Ernest Rutherford and Frederick Soddy saw that energy was released by radiation from radioactive materials. These findings implied that the Earth was not necessarily cooling. Rutherford also noticed that helium was trapped in radioactive minerals and thought that measuring the content of this gas might be used to determine geological ages. Lord Rayleigh and Bernard Boltwood were the first to study the radioactivity of rocks, and in 1905 Boltwood noted that lead was invariably associated with uranium and might be an end-product of the radioactive decay of uranium. Experimentation, mostly with 'home-made' equipment, indicated that there were two uranium isotopes, which decayed at different rates and produced different lead isotopes while releasing helium. By 1907 Boltwood was working on the uranium–lead ratios, while Rayleigh dated minerals using the helium produced. However, the gas could escape, leading to errors in the ages determined. In 1913, Arthur Holmes published the first full review of the methods and became personally involved in the experimental work.

Not all geologists liked the idea that the Earth might be billions of years old, and Holmes faced opposition to his conclusions about the Earth's age. Only in the late 1920s did his work begin to be accepted, when he calculated that the Earth was about 3300 Ma old. One of the problems attending radiometric age determinations was the uncertainty as to whether the rate of breakdown was constant, an issue raised by Joseph Barrell in 1917. In 1919–1920, Frederick Aston built the first mass spectrometer, which separated atoms according to their weight, but it was nearly 20 years before consistent results were obtained, using a machine built by Alfred Nier, with newly developed vacuum pumps. In 1956 Clair Patterson calculated the generally accepted age of the Earth (4.55×10^9 years) and the age of the solar system, based on lead-isotope ratios measured from iron meteorites. Later workers took up the study of various other isotopic relations, such as rubidium–strontium (mainly in the 1950s and 1960s), potassium–argon, and potassium–calcium (between 1920 and 1943). Different methods proved suitable for determining the ages of different parts of the stratigraphical time-scale. For the most recent 20 000 years or so, the radiocarbon (^{14}C isotope) method, developed by Walter Libby in 1952, has proved to be an effective dating tool. Radiocarbon is formed when atmospheric nitrogen is bombarded with neutrons and taken into organic material by photosynthesis, radioactively decaying after death. By the late 1950s the field of geochronology was well established. However, the expensive equipment required made it a somewhat exclusive field, and it has remained so.

Petrology (Igneous and Metamorphic)

The diversity of igneous rocks led to many attempts to explain the varieties and to classify the rocks into meaningful groups. Experimental petrology as a special branch of geology began in the 1890s, as the chemical analysis of rocks and minerals became more precise and more closely linked to the identification of the relationships of rocks using the petrological microscope. Controlled high-temperature and high-pressure studies commenced at the Geophysical Laboratory of the Carnegie Institution, established in Washington in 1905. Classifications were proposed on the basis of mineral content and texture, and more particularly chemical content, as assay methods became easier and cheaper. The so-called 'CIPW normative classification', the result of research by Charles Whitman Cross, Joseph Iddings, Louis Pirsson, and Henry Washington

in 1902, was one of the first classifications to be widely accepted. It depended on the recalculation of the 10 or so oxides most commonly found in igneous rocks by analysis, and assigning the values to particular minerals. It was particularly valuable for glassy rocks, where the minerals were not easily identified. The 'norm value' contrasted with the 'mode' or actual mineral composition. Although it was an artificial system, norm calculations allowed ready comparison, based on the rocks' chemistry, and grouping and subdivision without the distraction of minor mineral or textural variations. It was thought to give clues about the order of mineral crystallization in magmas.

Variations of the normative approach were proposed by researchers such as Samuel Shand, who suggested a classification of igneous rocks based on the proportion of silica present. While this was a 'logical' classification, it did not explain the origins of such variations. Classifications were also proposed based on rock textures, such as flow-banding, porphyritic content, and crystallinity.

There was widespread acceptance that the compositional and textural variations were caused by magmatic differentiation. This was thought to result from the presence of immiscible liquids within the magma, or from the separation of minerals in crystallization order by, for example, the sinking of heavier early-formed minerals. Others placed more emphasis on crystal–liquid fractionation (which was subsequently regarded as more important). A major influence from 1910 was the experimental work of Norman Bowen at the Carnegie Laboratory; one of his results is encapsulated in Figure 1. His classic book *The Evolution of the Igneous Rocks* (published in 1928), which summarized much of his experimental work to that time, influenced several generations of petrologists.

In the 1940s, Herbert Read reintroduced the idea of granitization, rather than magmas, as the major source of 'igneous' rocks. He argued that there was a space problem in the emplacement of large batholiths and suggested that igneous rocks were often earlier-formed rocks that had been altered by the action of active fluids that caused recrystallization. Read suggested that there was a 'continuum' from regional-metamorphic rocks to igneous rocks through what he called 'migmatitic' rocks (a term introduced by Jakob Sederholm). The role of volatiles consequently began to attract more attention, and Bowen, with Orville Tuttle, studied partial melting (hydrothermal activity) in addition to magma fractionation. Laboratory work, using pressure cells, contributed to the study of metamorphism, simulating conditions of pressure and temperature deep within the Earth. So, work on metamorphism proceeded hand-in-hand with the studies of igneous rocks.

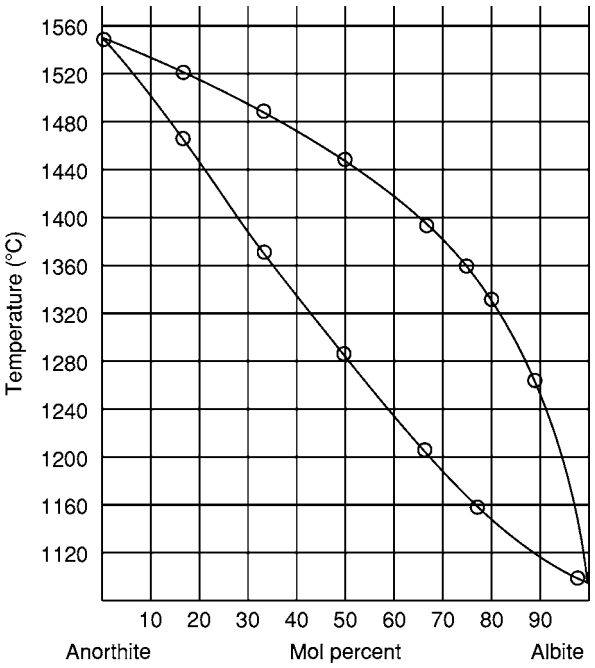


Figure 1 Phase diagram for plagioclase feldspar at one atmos phere. (Reproduced from Bowen N (1913) The melting phenom ena of the plagioclase feldspars. *American Journal of Science, Series 4* 35: 577 599 [Figure p. 583].)

In 1912, Victor Goldschmidt recognized that metamorphism essentially depended on two physical variables, pressure and temperature, and that most changes occurred while the original rocks remained essentially solid. A few years earlier Friedrich Becke had clearly described different metamorphic fabrics (distinctive patterns of minerals caused by different degrees of metamorphism). While Goldschmidt's ideas were valid in most cases, there were exceptions. In 1936, Dmitrii Korzhinskii and others showed that some melting could occur and that the mobility of some fluids, particularly in low-temperature hydrous activity, was important, although it had been dismissed at the beginning of the century. Related to this was the study of diagenesis, or very low-temperature metamorphism, which occurs as sediments are converted to sedimentary rocks. Serious study of this topic began in the 1920s, when certain mineral associations were recognized as indicators of the changes taking place through time. The work of Douglas Coombs in New Zealand in the 1950s played a major role in understanding these processes.

Geochemistry

The study of geochemistry began in the second half of the nineteenth century with the collecting of analytical data from rocks, minerals, and mineral waters by

a few chemists and geologists, including, at the turn of the century, Frank Clarke. In the 1920s, attention turned to attempting to understand the distributions and abundances of the elements in the Earth's crust, led by Goldschmidt in Germany and Vladimir Vernadsky in Russia. Goldschmidt's *Geochemistry* (published in 1957) brought together the essentials of the subject, although it had already moved on with the studies of isotopes begun at the University of Chicago, immediately after the Second World War, by Harold Urey and his co-workers. An international Geochemical Society was formed in 1955, and important seminars were held in 1957 and 1958 at the Carnegie Institution. The study of organic geochemistry also developed, with practical work related to deleterious substances in coal, mine dusts, and petroleum, and their dispersal into the atmosphere.

Stratigraphy and Sedimentology

Stratigraphy and the attempt to write the history of the Earth's crust had been the essence of geology since the late eighteenth century. But the rock layers and their enclosed fossils called for further study. During the first part of the twentieth century, the field mapping of the late nineteenth century was extended to many parts of the Earth that had been previously untouched. A major improvement in the maps was the attention to structural and sedimentological detail. Around 1920, it was shown that features such as cross-bedding and graded bedding could indicate the order of superposition, particularly in vertical and highly folded non-fossiliferous strata. The difference between cleavage and bedding in metamorphic rocks also had to be recorded. Thus previous interpretations of important regions, such as the Scottish Highlands, had to be re-examined. The structural geologist Marland Billings wrote "one clear contact or key bed is more valuable than a hundred petrofabric diagrams".

To make sense of the observable history worldwide, international agreement was needed about naming certain features and presenting them on maps and cross-sections. Stratigraphical commissions were established to set down rules for such things as identifying 'type localities'. The study of facies – the total nature of a volume of strata (rock composition, fossil content, type of bedding, sedimentary structures, etc., usually reflecting the conditions of origin) – was developed by Johannes Walther and others. Walther recognized that facies relations were dynamic. He stated that only rock types that can be deposited side-by-side can overlie each other directly in a vertical sequence. This led to the study of time-stratigraphical units, with the recognition that rock boundaries did not necessarily correspond with time. In this work, and in sedimentation studies in general, Amadeus Grabau was an acknowledged leader. Statistical analysis, long neglected by geologists and first used in petrology by Paul Niggli in 1924, became widely applied in stratigraphy, with information about rock-layer thickness (isopach) and lithofacies variation (e.g. sand–shale ratios) being presented in graphical form.

Closely related to stratigraphical studies was the examination of the materials of sedimentary rocks. Pioneers were Johann Udden and Chester Wentworth, who, in about 1920, studied size distributions and the shapes of grains, devising a quantification chart that was useful in the interpretation of sediment histories. Ralph Bagnold's study of the physics of blown sand in 1941 brought together considerable earlier research. Between 1900 and the 1950s, the heavy accessory detrital minerals of sedimentary rocks were widely used for stratigraphical correlation by William Rubey, Percy Boswell, and others. Laboratory work began to supplement field studies. From the 1930s flumes were widely used by Henry Milner, Paul Krynine, and others to examine the behaviour of sedimentary materials under different conditions. One problem, not always adequately addressed but worked on by Francis Shepard, one of the first US marine geologists, and M. King Hubbert, was the scaling from the actual geological dimensions to the laboratory dimensions.

In the early 1940s, Krynine attempted to set up a sediment research laboratory in the USA, bringing together the skills of oil companies, academics, and government geologists. This did not eventuate, and separate laboratories continued. In 1950 experimental flume work by Phillip Kuenen and Carlo Migliorini showed how 'turbidity currents' occurred, carrying materials rapidly downslope from a shallow source. This led to a 'turbidite revolution' in interpreting many types of stratigraphical occurrences, particularly greywackes, graded bedding, and other sedimentary features, and to an understanding of how some submarine cables might have been cut on continental shelves following earthquakes.

Palaeontology

The study of invertebrate fossils continued apace during the first half of the twentieth century. Initially palaeontologists were concerned mostly with taxonomy and classification, but additional fields opened up with the study of the evolution of particular groups. The Jurassic ammonites provided an example of a rapidly evolving fauna, which helped to pin down time zones, while the Ordovician and Silurian graptolites facilitated correlation between sedimentary rocks from these periods in widely separated parts of the

world. Palaeoecology began to develop, with the consideration of relations between fossil groups and sedimentary facies, and researchers began to lean heavily on the study of present-day environments.

From the 1920s, the importance of previously neglected organisms, such as microfossils, in identifying sections of the Tertiary epochs with the potential for oil productivity was recognized. Palynology, the study of plant spores, was also taken up, along with the parent study of palaeobotany. The period from the late nineteenth century to 1930 has been called the 'heroic period' of vertebrate palaeontology. Led by Henry Osborn, there were major discoveries on all continents. In 1910 Osborn proposed that Central Asia was the cradle of mammalian evolution, and much research was devoted to testing this theory. Asiatic discoveries expanded the fossil vertebrate (especially mammalian) record back to the Permian. Of particular significance was the discovery of human remains near Peking in 1926. Important work in palaeoanthropology was also carried out in Europe, Africa, Indonesia, Australia, and elsewhere.

Structural Geology

Structural geology made considerable progress in the first half of the twentieth century, often as a result of studies of metamorphic rocks. In the early 1930s, Bruno Sander and Walter Schmidt initiated 'petrofabric analysis': the study of spatial relations, including those between the individual minerals making up a rock, and the movements that could have produced these relations. The methods were used to investigate rock deformation and were taken even further to consider the genesis of both sedimentary and igneous rocks. Thus a special field of structural petrology was born, dealing with deformed rocks and their tectonic history.

Analysis of thin-sections using the 'universal stage' allowed the determination of the three-dimensional orientations of mineral grains relative to the original positions of rock specimens recorded in the field, and their representation on stereograms. This led to the recognition of various phases of deformation by workers such as Coles Phillips and Lamoral de Sitter. In structural geology, bedding planes, joints, and foliations could be represented graphically, and data could be averaged by the contouring of data points on the stereograms; hence polyphase deformations could be revealed. However, there were controversies about the significance and order of particular deformation events thus interpreted. Broad aspects of folding and even mountain building were studied in laboratory experiments by Rollin Chamberlin, Bailey Willis, and David Griggs, among others, using theories of scaling from engineering, but this pressure-box work suffered to some extent from the use of unsuitable materials and scaling problems.

Geomorphology

The development of landforms was widely studied in the early part of the twentieth century. In the first half of the twentieth century William Davis's erosion-cycle concept was widely accepted. This was the idea that the landscape tended to be worn down, but with decreasing speed, to a 'base level': the peneplain. The concept of Davis's stages – youthful, mature, and old-age landscapes (followed by rejuvenation by uplift) – became widely accepted, and various topographical levels were thought to represent the end points of separate cycles of peneplanation, recognizable back to the Mesozoic.

However, workers such as Albrecht Penck (1924) suggested other possible methods of surface evolution, including scarp retreat, the preservation of original depositional surfaces, and the formation of pediplains by the coalescence of pediments below scarps. It was not until the 1940s that Davis's concepts were significantly challenged. In 1945, Robert Horton published an article on the development of drainage networks, which marked the beginning of studies concentrating on the mechanisms of denudation rather than description. Arthur Strahler encouraged 'dynamic methods' of study, in which landscape units were treated as open systems in equilibrium, where a change in the system caused an adjustment to offset the effects of the change. Measurement of modifications in river profiles, slopes, and runoff began to provide evidence of landscape evolution. A decade of research on river-channel patterns followed, by workers including Luna Leopold and Markley Wolman (1957), using methods borrowed from engineering (such as fluid mechanics). This work was linked to experimental studies in sedimentology by workers such as John Allen, Alan Jopling, and notably William Krumbein in 1963, who treated beach morphology and processes as part of a system, whose changes could be computed.

Glaciation, Climate, and Palaeogeography

While the question of the existence of an Ice Age was resolved in the nineteenth century, studies during the twentieth century in Europe, North America, and New Zealand showed that it was an event with considerable variations, involving a series of intense glaciations with intervening interglacial periods. Of perhaps greater significance were the studies of earlier glaciations, particularly the Late Palaeozoic event,

which was first recognized in India, then in Australia and South Africa, and a little later in South America. These discoveries gave considerable support to the concept of continental drift. Evidence of Late Precambrian glaciation was also recognized early in the century in China, Norway, and Australia. These studies encouraged research into climate change through geological time.

Palaeogeography was developed in the early 1900s by Auguste de Lapparent. He published maps of France, Europe, and the world. Preparation of the maps required knowledge of the three-dimensional extents of rock units and their environments of deposition. The point of time chosen depended on identification of fossils. A particular exponent was Charles Schuchert. However the maps were general and usually covered too much time (because of a lack of secure age determinations). Nevertheless they were studied by petroleum geologists, and, thanks to the detailed stratigraphical studies that resulted, in return, from oil-search drilling, maps of specific 'slices of time' became possible.

Petroleum Geology

By 1900, some of the basic concepts of petroleum geology were understood. Oil was clearly of organic origin, from both vegetable (dominantly) and animal sources, and had formed at normal rock temperatures. It occurred in reservoirs within sandstones and limestones, and permeability was important, with the limestones usually being 'tighter'. A relatively impervious roof of rock, such as shale, was a primary requisite for oil to accumulate. Many reservoirs showed a separation by gravitation into three layers: gas, oil, and saltwater. A particular problem demanding explanation was the enormous pressures sometimes encountered in oil fields. Edward Orton was one of the first to relate these pressures to artesian conditions.

It took many years to determine how the organic material was transformed into oil. In the 1930s Parker Trask argued that only certain organic material could produce petroleum, while J. M. Sanders thought that almost any organic material could be converted to petroleum, given the right conditions. It seemed that specific sedimentary deposits, such as dark marine shales, were likely source beds. Studies showed that petroleum formed slowly, with solid matter being converted to heavy and viscous fluids. In time these thick fluids were changed by heat and pressure into lighter oils. Oil 'pools', containing immense accumulations of oil, contrast with the disseminated nature of the oil forming in the source beds. Thus geologists began to study the migration from source beds to reservoir rocks. Differential pressure was recognized as the essential cause of the movement of the fluids through porous beds, and permeability was an essential condition.

A wide variety of oil 'traps' was described: structures such as anticlines or domes were important, but there were also depositional traps in which particular favourable beds thinned out between impermeable beds or were partly eroded and covered by 'tighter' beds. Faulting could also cause traps to form. Many traps were produced as a result of a combination of various causes. After the Second World War oil companies devoted considerable attention to the study of modern environments (especially deltas) to elucidate many of the fundamental aspects of the accumulation of organic materials and their conversion to oil, while also investigating the complexities of oil structures and the pressure–temperature regimes that contributed to the variation in types of oil or gas accumulations.

Exploration Geophysics

In parallel with developments in petroleum geology, there were major advances in exploration geophysics during the first half of the twentieth century, particularly between 1925 and 1929. Although concentrating on oil and mineral prospecting, some of these practical advances also contributed to the broader studies of the Earth's interior, to seismology, terrestrial magnetism, hydrology, geodesy, and meteorology. Four geophysical methods were widely used in mineral and oil exploration: seismic, gravimetric, electric/electromagnetic, and magnetic. After the Second World War other methods, such as radiometric, also began to be used. Each of these has been applied in different ways. The seismic methods used artificial shock waves: seismic reflection records the rebound from a reflecting surface, while seismic refraction records the path of waves refracted along high-speed layers. Although seismic reflection is simpler in theory, seismic refraction proved to be more useful in practice. Gravity methods used an Eötvos torsion balance or a gravimeter. Electrical methods measured natural or artificially induced earth currents. The electrical methods were combined and extended to logging the variations in resistance and electric potential down a drill hole. The resistance gave an indication of the type of bed (high-resistance oil sandstones and coal beds contrast with salt-water-bearing sandstones and shales). These electrical tests often gave consistent results within particular beds, allowing the correlation of both stratigraphy and structure (**Figure 2**). Two magnetic methods were used: one measured anomalies caused by ore bodies containing significant quantities of magnetic minerals; the other measured

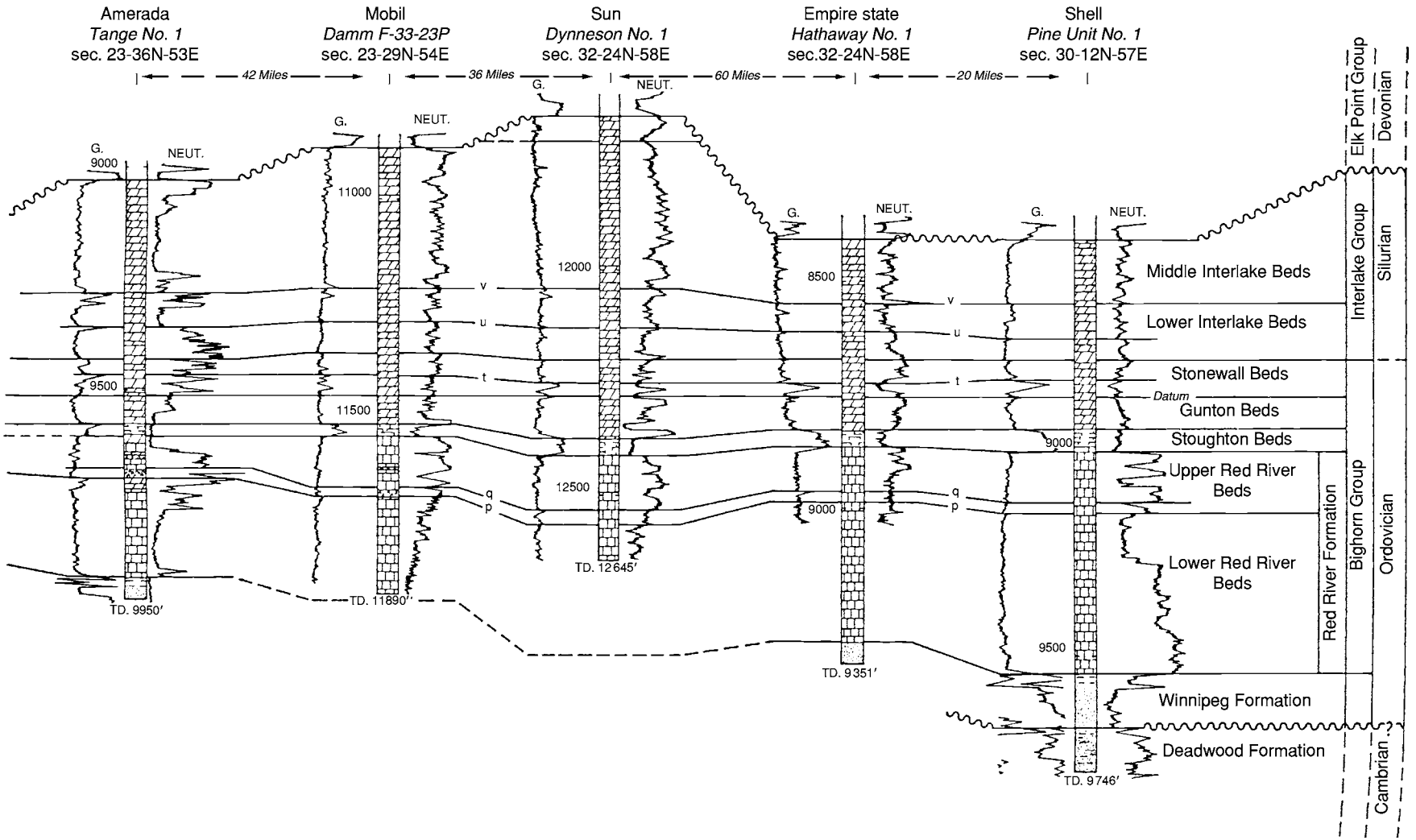


Figure 2 Williston Basin, Eastern Montana, Subsurface section (approximately 150 kms length), showing correlation achieved using Schlumberger G(amma) Neut(ron) well logs. (Reproduced from Krumbein W and Sloss L (1963) *Stratigraphy and Sedimentation*, 2nd ed., p. 382. San Francisco, W.H. Freeman & Co.)

slight variations in the magnetic properties of common rocks (basalt is normally more magnetic than other rocks) to construct structural maps, which had the potential to locate salt domes and volcanic plugs. A pioneer of geophysical prospecting was Conrad Schlumberger, who developed his electric methods between 1908 and his death in 1936. Surveying for structure using the Eötvos balance began in 1921 when the Anglo-Persian Company tested it in Hungary. The Royal Dutch Shell Oil Company took up the method the following year in Egypt and later in Texas and Mexico. There was similar interest in Germany and Russia at about the same time. The formation of the American Association of Petroleum Geologists in 1917 led to increased communication between geologists and spread their enthusiasm for testing geophysical methods.

Economic Geology

The considerable close study of metalliferous mineral deposits and related geological aspects early in the century led to the formation of the Society of Economic Geologists and the *Journal of Economic Geology* in 1905. Mineral deposits were often associated with igneous bodies, and Waldemar Lindgren's (1913) hydrothermal classification (hypothermal, mesothermal, and epithermal: that is, deposits formed from high-, moderate-, and low-temperature/pressure fluids emanating from igneous bodies) was widely accepted and applied and in many cases was an effective tool in the recognition of variations within an ore body and was useful in the search for new ore bodies. Important work was done by William Emmons on the secondary enrichment and zonation of ore bodies. There was general acceptance that deposits showing evidence of original sedimentary bedding were the result of 'replacement' of the bedding by later-introduced ore-bearing fluids from a nearby igneous source. In the 1950s the 'replacement' theory of the formation of certain ore bodies was challenged by Haddon King, who with his associates studied the massive Broken Hill ore body in Australia. King proposed that the lead and zinc layers were deposited as sediments and that the only major change had been later folding. This idea has become widely accepted for many of the world's largest base-metal deposits and the major iron-ore bodies. The significance of micro-organisms and their ability to concentrate and deposit metals were uncovered by laboratory experiments such as those by Lourens Baas-Becking. The use of reflectance microscopy (mineragraphy) to identify opaque ore minerals, study the relationships between such minerals, and find clues about the deposition of ores began with J. der Veen in the 1920s. It reached a high level in the 1940s and 1950s through the work of Hans Schneiderhöhn and Paul Ramdohr in Germany, and Frank Stillwell and Austin Edwards in Australia.

Engineering Geology

The failure of the St Francis Dam in California in 1928 drew attention to the need to assess foundation conditions and rock quality. After the Second World War there were numerous attempts to quantify, for example, the rate of weathering of stone, joint distributions, and other weaknesses within rock masses. The field of rock mechanics grew from these developments through the efforts of workers such as David Griggs, Karl Terzaghi, Charles Berkey, John Jaeger, J. Talobre, and Robert Legget as well as researchers in South African mines and Australian hydroelectric schemes.

In the late 1930s geophysical methods (electrical and seismic) were adapted by engineering geologists to determine the depth to solid rock at dam sites and elsewhere, while electrical methods were used to determine the depth to the water table in arid areas. Such uses continued to develop after the Second World War.

World Views

Despite the growth of specialization, there were, of course, attempts to develop global theories, and global structural patterns were particularly discussed in the early twentieth century. The most famous of these studies was Eduard Suess's *Das Antlitz der Erde*. It became influential in the English-speaking world following its translation in the early 1900s. A particular aspect of Suess's theory, based on the study of major features, such as mountain ranges and the patterns of coastlines (Atlantic and Pacific types; the former 'fractured', the latter with fold mountains parallel to the coast), was the idea of an Earth that had been contracting since its formation. Suess rejected the idea that the present continents and oceans had existed from earliest times, believing that the Pacific Ocean was the oldest, possibly formed when the Moon separated from the Earth. Hans Stille also believed in a contracting Earth, but majority opinion held to a fixist concept and the permanence of the continents and oceans. This was questioned, albeit cautiously, by some. Reginald Daly was more adventurous in *Our Mobile Earth* (published in 1926) and was supported by the seismologist Beno Gutenberg in proposing mobility of the Earth's crust, even though the processes were unknown.

The associated concepts of geosyncline (mobile fold-belt) formation and isostasy to explain the accumulation of enormous thicknesses of (mainly)

shallow-water sediments and mountain building, although formulated in the second part of the nineteenth century, continued to be developed, and were widely applied in the period up to 1950, particularly in North America. There were arguments as to how the geosynclines had contributed to continental growth. Was it by accretion on the oceanic edges of supposedly stable continental cratons? Or was a geosyncline formed within a continental mass when one side had subsided? Such subsidences were often invoked to explain the cutting off of routes of animal and plant migration between continental masses. While the geosynclinal concept as the basis of a tectonic theory has, since the acceptance of plate tectonics, been dismissed by many, the descriptive aspects of geosynclinal sedimentation are still useful (*see* **Famous Geologists:** Hall; **History of Geology From 1835 To 1900**).

The meteorologist Alfred Wegener published his famous book, Die Enstehung der kontinente und Ozeane (*The Origin of Continents and Oceans*) in 1915, proposing a tectonic theory based on lateral movements of the Earth's continental crust. Others, such as Frank Taylor, had previously made similar suggestions, but had not set out their ideas as fully as Wegener did. Wegener's theory became more widely known when his book was published in English in 1924. In the following 30 years or so, geophysicists, led by Harold Jeffreys, claimed that the forces postulated by Wegener were insufficient to cause horizontal movements of the continents, and, therefore, the theory should be forgotten. However, it was taken up by some European geologists in the 1920s, one supporter being the Swiss geologist Emile Argand, who made a 'spirited defence' at the first post-war International Geological Congress in 1922, discussing the 'Tectonics of Asia'. Argand's nappe theory, which was generally accepted by then (although general 'mobilism' was not), indicated that large horizontal movements (extending through a considerable thickness of crust) caused by strong lateral compression had occurred in parts of Europe and North Africa. However, he could not suggest how the forces required to cause these movements, or by the theory of continental drift, could be generated. He was unaware that a solution for the Alpine deformations had been suggested by Otto Ampferer in 1906 – the action of massive convection undercurrents within the upper mantle causing what Ampferer called 'subduction'.

In 1925 Ampferer told Wegener that these currents must be contributing factors in the mechanism of continental drift. Although it was accepted only in about 1950, Arthur Holmes, in the late 1920s, essentially explained how massive slow-moving convection currents could operate within the upper mantle (Figure 3). Holmes recognized that the radioactivity of continental rocks was generally greater than that of oceanic rocks, causing higher temperatures below the continents. This unequal heating of the mantle would cause ascending currents under a continental region, which would spread out at the top in all directions towards the cooler peripheral regions. The downward currents would be strongest beyond the continental edges. Thus the continental block could be ruptured, and portions could be carried apart on the backs of the currents.

While supported by some European geologists, Wegener's idea was rejected out-of-hand by the majority of North American geologists, and some

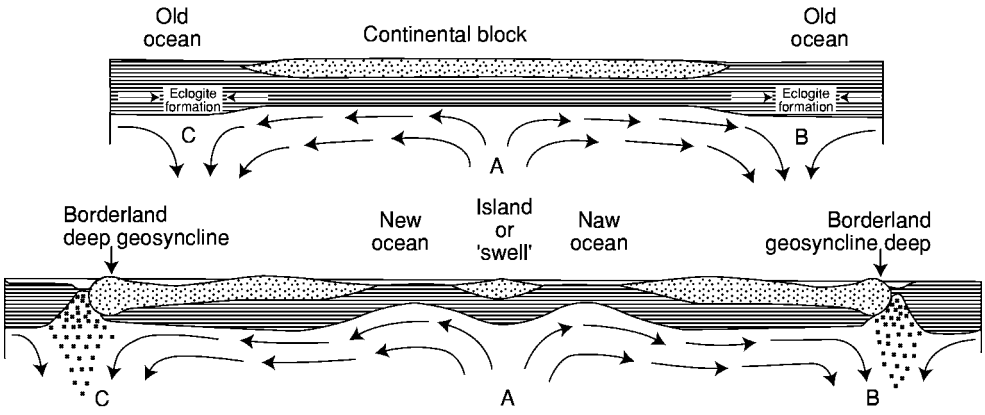


Figure 3 Holmes's convection model. The top figure shows a postulated system of convection currents in the mantle rising at A beneath continental crust and moving horizontally to edges of the continental block before descending at B and C (incidentally causing the formation of eclogite, because of high temperature and pressure). The lower figure shows the consequence of this movement. The horizontally moving currents have split the old continental block into three, forming two new continents (and a minor remnant). The outer edges of the new continents are deformed during down dragging. (Reproduced from Holmes A (1931) Radioactivity and earth movements. *Transactions of the Geological Society of Glasgow for 1928 29* 18: 559 606 [Figure p. 579]).

German geologists, particularly Stille, also opposed Wegener. It was left to southern-hemisphere geologists, particularly the South African Alexander du Toit, to accumulate evidence, such as the similar widespread Late Palaeozoic glaciations, Permian coals, and Mesozoic continental successions in the southern continents (and India), that substantiated Wegener's views, but this work interested few northern-hemisphere geologists. The idea of continental drift was kept alive into the 1950s by the South African Lester King and the Australian Warren Carey, who were not taken seriously, in the early 1960s, by many North American geologists. However Carey's symposium on continental drift, held in Tasmania in 1958, attracted attention at a time when ideas in many fields of geology were fomenting.

Ironically, one of the American scientists opposed to Wegener's ideas, William Bowie, was responsible for getting one of the pioneers of plate tectonics, Maurice Ewing, into ocean-floor exploration. In 1934, Bowie (with Richard Field) asked Ewing if he could study the outer edge of the continental shelf. Was it a basic geological feature, such as a fault, or just superficial – the result of outbuilding of sediment? Ewing said that the problem could be studied using seismic-refraction geophysical methods and proceeded to do so. Ewing's work in the ocean basins was to be crucial in resolving the continental-drift controversy.

Around 1900 the phenomenon of palaeomagnetism, and field reversals preserved in lavas, were recorded by Bernard Brunhes, but their significance was ignored until the 1920s. A key to change came with the publication of the 'pole-wander' studies of John Graham in 1949. These results were dismissed as 'impossible' by prominent mathematicians, but by 1955 Keith Runcorn and Paul Blackett provided evidence that the poles had indeed wandered during geological time. Runcorn maintained that movement of the magnetic poles relative to the crust had occurred, in contrast to Blackett's idea that portions of the crust had moved many thousands of kilometres. But there was a problem: the studies on different continents gave different sets of results. These sets could, however, be reconciled by rotating the continents, bringing the various sets into coincidence. There had to be something in the idea of continental drift. The resolution of the problem is discussed in **History of Geology Since 1962**.

The Inner Earth

An important breakthrough in determining the constitution of the Earth's interior and particularly the size of its proposed liquid core was made by Richard Oldham in 1906, using data from various earthquake records. Oldham showed that there were three separate records from a major earthquake, caused by three distinct forms of wave motion, propagated at different rates and along different paths, forming three distinct phases in the record of the earthquake at a distant point. The latest record was due to waves propagated on or near the Earth's surface. The other two were due to waves that had travelled through the Earth. Oldham indicated that the crust was thin and non-homogeneous and therefore did not transmit mass waves. Below this crust, the records suggested that there were two distinct zones, the outer three-fifths (mantle) differing from the inner two-fifths (core). 'The core appeared to be liquid. . . .shadow' and: 'The S waves (Figure 4) were delayed in arriving on the opposite side of the Earth'. Oldham, at the time thought that they travelled more slowly through the core for some unknown reason, but recognized that if the core were liquid this would prevent the S waves travelling through and they would be diverted

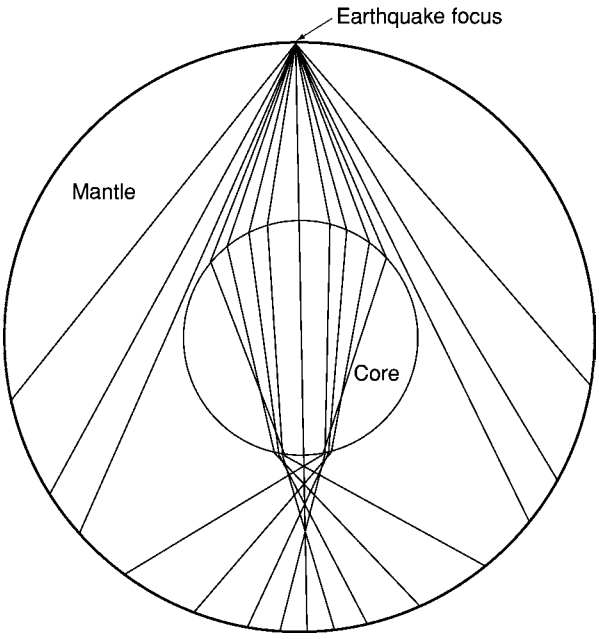


Figure 4 Richard Oldham's simplified figure showing the transmission of seismic waves through the Earth. The figure represents the deducted paths of what Oldham called second phase waves (i.e. S waves) travelling through a central core occupying. 4 of the radius of the earth at half the original speed of propagation. He pointed out that 'there should be a zone, at about 140° from the origin where the second phase waves would be so dispersed, and consequently so feeble, that it would amount to a shadow, and the second phase should be absent in records from this distance, or [only] feebly marked.' (Reproduced from Oldham R (1906) The constitution of the earth as revealed by earthquakes. *Quarterly Journal of the Geological Society of London* 62: 456 475 [Figure p. 468]).

(refracted) around the core (and thus delayed). It was not until 1913 that he accepted the idea of a liquid core which caused the 'shadow' for S waves if the angular distance from the focus of the earthquake were more than 120°. Oldham and most other researchers were unaware that the Russian Leonid Leybenzon had suggested the same idea in 1911.

There then followed the recognition of the 'Moho discontinuity', discovered by Andriya Mohorovicic in 1909, separating, he thought, the crust from the mantle. However, what was the nature of this discontinuity? Was it the result of chemical or merely physical differences? Were the concepts of two separate crustal layers, the heavier 'sima' (silica and magnesium) and the lighter 'sial' (silica and aluminium), introduced by Suess, still valid? Ideas and terms remained in turmoil, as some, such as Bailey Willis, equated 'crust' with sial. Did such crustal divisions explain the major differences in magmas and the resultant intrusions and extrusions that could be observed at the surface, or were deeper levels involved in the evolution of surface geology?

Other workers suggested different models of the Earth's interior, and in about 1914 Joseph Barrell introduced the concept of the 'lithosphere' as a thin crust, underlain by the 'asthenosphere', a relatively weak 100 km or so, with an even weaker layer of unknown thickness beneath, in which adjustments to isostasy took place. Between 1940 and 1942 Keith Bullen divided both the mantle and the core into three separate zones, but there were still unresolved questions. In 1954, a conference 'The Crust of the Earth' was convened at Columbia University, New York, and many prominent geologists and geophysicists contributed, reaching some agreement on terms and directions for future research into the crust and mantle. However, the question of the Earth's inner composition remained largely unsolved. The increasing sophistication of geophysical equipment, particularly seismic, showed that the 'Moho' was not a continuous feature, but was 'replaced' in places, such as below mountain belts, by a disturbed 'mixed' zone. In the early 1950s, evidence from marine surveys showed that the ocean floors were covered in many areas by only a thin layer of sediments. With the increasing ability to drill deep within the Earth it was suggested in the late 1950s that some 'real evidence' of the nature of the Moho discontinuity could be obtained by drilling a hole 5 or 6 km through the crust on the seafloor. The 'Mohole project' was taken up with enthusiasm, but eventually failed because the ideas of the geologists involved differed from the approach advocated by the preferred drilling company, and the project was terminated in 1966.

While the records of earthquakes were used to study the Earth's interior, seismologists also devoted considerable efforts to predicting and classifying earthquakes. In the 1930s a number of seismologists began work at the California Institute of Technology. They included Harry Wood, Beno Gutenberg, Hugo Benioff, Frank Press, and Charles Richter, who were essentially the founders of modern seismology, although there were important researchers such as Fiusakichi Omori in Japan. Nevertheless the accurate prediction of earthquakes has not yet been achieved (*see* **Seismic Surveys**).

Impact Craters

While actualism (uniformitarianism) ruled during the early part of the twentieth century, there was still a suspicion that 'catastrophes' might have played a part in shaping the Earth. Late in the nineteenth century Grove Karl Gilbert argued for impacts as the origin of most of the Moon's craters. However, it was a long time before such features on the Earth's surface were accepted as being due to meteorites. Specimens from small craters at Henbury in Central Australia were described in 1932 and accepted as meteorites, but it was not until the 1950s that the larger famous Barringer (Meteor) Crater in Arizona was accepted as an impact crater, despite information obtained from drilling in the early 1900s. As early as 1942, Harvey Nininger argued that large impacts might have influenced the evolution of life. This idea was not really taken up until the 1980s, although when the Yucatán Puerto Chicxulub crater was investigated for oil in the 1950s the drill cores showed impact melt material, and, in 1970, Digby McLaren also talked about the idea. In the mid-1950s research on craters began in earnest, initially in two independent fields: photographic and field studies investigated the origin of craters and the possible significance of accretion of impact bodies in the formation of heavenly bodies; and the mode of formation of craters was investigated by nuclear and chemical explosions and detailed studies of shock-wave propagation. Important researchers in the former (geological) field included Ralph Baldwin, Robert Dietz, and Eugene Shoemaker; Donald Gault's experimental work on hypervelocity impacts was typical of the latter. By the late 1950s there was considerable interaction between these two groups, and the first international impact symposium was held in 1961. Pioneering work in the 1940s and 1950s by George Baker and others on the glassy meteorites (tektites), notably their shape, proved important in the study of the problem of re-entry by spaceships into the Earth's atmosphere.

The International Geophysical Year, 1957–1958

The years after the Second World War saw a rapid increase in air travel, which enabled researchers to meet, face-to-face, other workers with similar interests but different backgrounds, and a considerable amount of international collaborative work began. The International Geophysical Year, beginning in 1957, was a significant event. Following international cooperation in 1882–1883 and 1932–1933 in gathering data about the polar regions, it was first suggested that 1957–1958 should be the Third Polar International Year. However, it became much larger, as a result of the encouragement of the UN-supported International Union of Scientific Unions. Between July 1957 and December 1958 (a period of high solar activity) 13 scientific programs investigated the Earth, the atmosphere, and space, and their interrelations. At least 60 nations participated, and thousands of scientists were involved, most undertaking national problems, but there was, as in the earlier 'Years', a concentration on Antarctica and the Arctic, with new work along the equator and on selected meridians. Tuzo Wilson, President of the International Union of Geodesy and Geophysics, played a considerable part in helping to coordinate the results of the research. One of the best-known events of the International Geophysical Year was the launching of *Sputnik*, which produced significant scientific data and also inspired the 'space race'. A second, and happier, result was the Antarctic Treaty to protect that continent. The 18-month 'year' provided the impetus for work that has continued to the present.

Summary

The period considered here saw many advances. While broad concepts were initially tied to the idea of a generally ordered Earth with fixed continents and oceans, facts that were accumulating hinted at the need for major changes in thought. The increased ability to measure, thanks to much improved technology, meant that rapidly acquired data could be quickly analysed and disseminated through more easily accessible scientific media. Coming towards the end of the period, the International Geophysical Year was the culmination of a long phase of gestation. The plate-tectonics revolution could not have occurred until many years later were it not for the accumulation and analysis of more than 60 years of data. There was conservatism, and new ideas took time to be accepted. Nevertheless, much was achieved.

See Also

Analytical Methods: Geochronological Techniques; Gravity. **Diagenesis, Overview**. **Famous Geologists:** Du Toit; Walther; Wegener. **History of Geology Since 1962**. **Microfossils:** Palynology. **Seismic Surveys**. **Time Scale**.

Further Reading

Davis WM (1954) *Geographical Essays*. New York: Dover Publications.

Ginsburg RN (ed.) (1973) *Evolving Concepts in Sedimen tology*. Baltimore: Johns Hopkins University Press.

Good G (1998) *Sciences of the Earth: An Encyclopedia of Events, People and Phenomena*. New York: Garland.

Holmes A (1944) *Principles of Physical Geology*. London: Thomas Nelson.

Lewis CLE and Knell SJ (eds.) (2001) *The Age of the Earth: From 4004 BC to AD 2002*. London: The Geological Society.

Mather KF (ed.) (1967) *Source Book in Geology 1900 1950*. Cambridge, MA: Harvard University Press.

Muir Wood R (1985) *The Dark Side of the Earth*. London: Allen and Unwin.

Müller DW, McKenzie JA, and Weissert H (eds.) (1991) *Controversies in Modern Geology: Evolution of Geo logical Theories in Sedimentology, Earth History and Tectonics*. London: Academic Press.

Oldroyd DR (1996) *Thinking about the Earth: A History of Ideas in Geology*. London: Athlone.

Oldroyd DR (ed.) (2002) *The Earth Inside and Out: Some Major Contributions to Geology in the Twentieth Century*. London: The Geological Society.

Pettijohn FJ (1984) *Memoirs of an Unrepentant Field Geologist*. Chicago: University of Chicago Press.

Umbrove JHF (1947) *The Pulse of the Earth*. The Hague: Nijhoff.

Wertenbaker W (1974) *The Floor of the Ocean: Maurice Ewing and the Search to Understand the Earth*. Boston: Little, Brown and Company.

White JF (ed.) (1962) *Study of the Earth: Readings in Geological Science*. Englewood Cliffs: Prentice Hall.

HISTORY OF GEOLOGY SINCE 1962

U B Marvin, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

Introduction

The theory of plate tectonics postulates that the Earth's outermost layer, the lithosphere, which consists of the crust (*see* **Earth:** Crust) and the uppermost mantle, is about 100 km thick and is broken into rigid plates that slowly move and change their configuration in response to thermal instabilities in the mantle. The theory holds that the ocean floors are youthful due to their continual creation at spreading ridges and destruction as they plunge back deep into the mantle (*see* **Earth:** Mantle). The continents, in contrast, are buoyant sialic blocks with components of all ages, which ride passively on the surfaces of the plates. This theory, established in 1968, has unprecedented power for making quantitative calculations of past crustal motions and predictions of future ones. The immediate precursor of plate tectonics was the hypothesis of seafloor spreading proposed separately by two scientists in 1960 and 1961. Seafloor spreading was based primarily on information provided by geophysical explorations of the ocean basins during and after World War II. This article reviews the discoveries that led to plate tectonics, and the changing climates of opinion that led to a new era in the geosciences.

Post-war Explorations of the Ocean Basins

The geology of the ocean basins, which occupy more than 60% of the Earth's surface, remained largely unknown until after World War II. Despite the early hypothesis of continental drift, a majority of geologists regarded the ocean basins, particularly the Pacific Basin, as primordial features, formed when the Earth's crust first cooled. In 1946 the Dutch geologist, Philip Kuenen, calculated that if the Earth is 3 billion (3×10^9) years old, the ocean floors should be covered with a layer of sediments 5 km thick. In 1956, when radiometric dating showed the Earth to be ~4.55 billion years old, many expected the sediments to be much thicker. Indeed, as late as 1958, some of the scientists who were planning Project Mohole, to drill to the Mohorovičić discontinuity through the Pacific floor, anticipated that the cores would reveal the entire sequence of sediments deposited during the Lipalian interval from the end of the Precambrian to the beginning of the Cambrian Period – a time of severe erosion on the continents.

In the latter 1940s, the United States Navy provided generous funding for oceanographic research including ships, state-of-the-art equipment, laboratories, and scientists. Two of the four institutions credited with founding plate tectonics began sea-going explorations: the Scripps Oceanographic Institute at the University of California in San Diego, and the Lamont Geological Observatory (now the Lamont-Doherty Earth Observatory) at Columbia University in New York. The other two institutions contributing to the establishment of plate tectonics were the Department of Geophysics at Cambridge University in England, and Princeton University.

Preeminent scientists directed both of the oceanographic institutes: Roger Revelle at Scripps, and Maurice Ewing at Lamont. Both began intensive programmes of mapping the submarine topography by echo depth soundings, displaying the layering of the ocean floors by seismic profiling, and collecting samples by piston coring. Scientists aboard Scripps vessels also measured heat-flow values, while those from Lamont focused more on measuring gravity at sea. Both organisations towed magnetometers behind their ships. At the beginning, these scientists were not testing hypotheses; they were performing the first comprehensive investigations of the last unknown domain of the Earth. There were many surprises in store for them.

The ship *Horizon*, which left Scripps in July 1950, with Revelle and several other now-famous scientists aboard, made its first discovery 300 miles south-west of San Diego. There, in deep water, its companion Navy ship fired the first explosives and *Horizon*'s seismic profiler showed that the Pacific sediments were not 5 km or more thick: they were only 260 metres thick. The profiler detected three layers in the ocean floor: Layer 1, consisting of sediments ~0.25 km thick; Layer 2, of consolidated sediments or volcanics or both, 1–2 km thick; and Layer 3, the authentic bedrock of the ocean floor, which ultimately was shown to maintain a uniform thickness of ~4.5 km throughout the ocean basins.

From the beginning, the echo soundings showed the ocean floor to be unexpectedly crowded with hills, 10–30 or more km long, 2–5 km wide, and 50–500 m high. Further explorations showed that these so-called 'abyssal hills' occupy more than 30%

of the floors of all oceans, placing them among the Earth's most abundant topographic features.

Revelle and Arthur Maxwell measured heat-flow from the ocean floor and observed, to their astonishment, that the values averaged within 10% of those from the generally granitic and consequently more radioactive continents. For want of other possibilities, they suggested that the heat in the ocean floor must be carried there by rising limbs of convection cells in the mantle, an idea compatible with the earlier 'drift' model proposed by Arthur Holmes in the1920s.

In the mid-Pacific, *Horizon* dredged the surfaces of guyots – volcanic cones with flat summits, submerged 1–2 km below sea-level. These curious features had been discovered and named by Harry Hess, of Princeton University, who located 160 of them while he was on naval duty during World War II. At first, Hess assumed they were Precambrian islands, erupted long before reef-building corals evolved, that had been eroded at sealevel and then drowned by the rising of water due to the deposition of 2–3 km of oceanic sediments. The rare opportunity to study and describe the dredged materials fell to Edwin Hamilton, to whom Robert Dietz, on the *Horizon*, had assigned this topic for his PhD thesis. The hauls brought up fragments of Cretaceous corals, only about 100 my old. Suddenly, the 'ancient' ocean basins were seen to be much younger than the continents.

In the final days of the cruise, the *Horizon* steered a course to allow the youthful Henry Menard, of the US Naval Electronics Laboratory, to determine the nature of the Mendocino escarpment, a long narrow feature, striking westward from Cape Mendocino in California. On a S–N traverse of ~60 km, the echo-sounder showed that, starting in deep water the ship passed over a low ridge tilted to the south, a narrow trough, a steep scarp 18 km high, a high ridge tilted to the north, a low swale, and then it arrived in relatively shallow water. This 'fracture zone' was a new class of submarine structure of which Menard and others would find several similar examples farther south in the eastern Pacific Ocean. They appeared to be strike-slip faults with a component of normal faulting (*see* **Tectonics:** Faults), but they terminated before slicing into the American continents. Eventually, fracture zones, long and short, would be found in all the ocean basins.

Meanwhile, Lamont ships had begun producing the earliest detailed topographic charts of the mid-Atlantic Ridge and its linkage to what would prove to be a world-encircling succession of submarine mountain ranges, with various spurs, nearly 60 000 km long, 1–3000 km wide at the base, and 2 km high, with peaks rising to 4 km above the ocean floor. In 1956, Maurice Ewing and his colleague, Bruce Heezen, reported finding a deep rift valley (*see* **Tectonics:** Rift Valleys) occupying the ridge crests. Subsequently, the rifts proved to be discontinuous but, where present, they were zones of shallow-focus seismicity, strong magnetic anomalies, and higher than average values of heat-flow. The ridges are offset horizontally into segments by transverse fracture zones, which were assumed to be strike-slip faults.

The Navy prohibited publication of contoured bathymetric charts, but two Lamont scientists, Heezen and Marie Tharp, began drawing 'physiographic diagrams' showing how the ocean floors would look with the waters drained away. These diagrams, published in the 1960s, astonished both non-scientists and scientists, who found them remarkably useful for interpreting the structures and history of the ocean basins. Heezen and Tharp both won gold medals for their work.

The East Pacific Rise (or Ridge) is unique in being a huge bulge (nearly 13 000 km long, 3000 km broad, and nearly 4 km high) with no sharp crest, no central rift, and no mid-ocean position – it lies close to the margins of the American continents. In 1960, Menard proposed that convection currents had moved the northern portion of the Rise beneath western North America where its presence would account for the high plateaus of Mexico and Colorado and the Tertiary Basin and Range topography of Utah and Nevada. It followed that the floor of the eastern Pacific Ocean, cut by the great E–W fracture zones, was the western flank of the buried Rise.

Many of the observations, made in the 1950s, suggested youthful, mobile ocean floors, but contrary explanations were offered in each case, puzzling scientists who were trying to piece together the dynamic history of the Earth. Then in 1960 and 1961, Hess and Dietz independently but simultaneously, proposed hypotheses that addressed global dynamics in terms of moving seafloors driven by convection in the mantle.

Sea-floor Spreading: 1960 and 1961

In 1960, Hess circulated the preprint of a book chapter in which he described the ocean floors as the exposed surface of the mantle. He proposed that large-scale convection cells in the mantle create new ocean floor at the ridges, where the rising limbs diverge and move to either side until they cool and plunge down into the mantle at the trenches (*see* **Tectonics:** Ocean Trenches) or beneath continental margins (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges). He now explained guyots as volcanic peaks eroded on oceanic ridge crests and carried to the depths by the moving seafloor. Where

limbs rise beneath continents they split them apart and raft the granitic fragments in opposite directions until grounding them over zones of down-welling. Hess suggested that inasmuch as granite is too buoyant to sink into the mantle, some fragments of continents have survived since the beginning of geologic time, while the ocean floors have been swept clean and replaced by new mantle material every 300–400 my.

Hess further theorized that the mantle consists of peridotite, an olivine-rich rock that becomes hydrated to serpentinite at the ridges by reaction with heated waters released from depth. He favoured serpentinite partly because of its ease of recycling and partly because he believed it would be impossible to achieve the uniform 4.5 km thickness of the ocean floors with dykes and lava flows. Hess acknowledged that mantle convection was regarded as too radical an idea to be widely accepted by geologists and geophysicists, but he pointed out that his model would account for many phenomena in a coherent fashion: the formation of the ridge-rift system and the trenches, the youth and uniform thickness of the ocean floors, the thinness of pelagic sediments, and moving continents. Hess called his chapter an essay in 'geopoetry'.

In 1961, Dietz published a three-page article in *Nature* in which he, too, argued that convection in the mantle, moving at a few centimetres per year, could produce the overall structure of the ocean basins: the ridges form over sites of rising and diverging limbs; the trenches form at sites of converging and down-welling limbs, and the fracture zones are shears between regions of slow and fast creep. Dietz suggested that the mantle consists of eclogite, a dense pyroxene-garnet rock, and that the ocean floors are built of basaltic dykes and pillow lavas, formed by the partial melting of the eclogite at the ridges. But he added that the actual rock compositions were of less importance than the fact that the ocean floors must be recycled mantle material. Dietz pointed out, as had Hess, that rising convection currents rift continents apart and carry the sialic fragments *en bloc* to sites of down-welling and compression, where folded mountain ranges form on their margins. He added that the pelagic sediments ride down into the depths on the surfaces of the plunging oceanic slabs where they are granitized and welded onto the undersides of the continents – thus contributing to the persistent continental free-board, despite steady erosion of the continental surfaces toward base level.

Dietz called this process 'sea-floor spreading'. This concept, he argued, requires geologists to think of Earth's outer layers in terms of their relative strengths, so we should begin referring to the 'lithosphere', a historic name for Earth's outermost layer, which is relatively strong and rigid to a uniform depth of about 70 km (now generally taken as 100 km) under both continents and oceans. Beneath the lithosphere lies the weaker, more yielding, 'asthenosphere' on which the lithosphere moves in response to convection currents. The asthenosphere had previously been hypothesized on the basis of seismic evidence.

Dietz made the first suggestion (later confirmed) that the oceanic abyssal hills are a chaos topography, developed when strips of juvenile sea floor have ruptured under stress as the floors move outward. Finally, he referred to two papers *in press*, one by Victor Vacquier *et al*., and one by Ronald Mason and Arthur Raff, reporting the discovery of linear magnetic anomalies on the Pacific floor. Some of the magnetic lineations appeared to be offset by up to 1185 km along the Mendocino fracture zone. Dietz suggested that the lineations are developed normal to the direction of convective creep of the ocean floor. Noting that neither the fracture zones nor the magnetic lineations impinge upon the continental margins, he suggested that both lost their identity as the Pacific floor slipped beneath the American continents.

Rarely are ideas subsequently seen as basic to a grand new system of thought in science, appreciated at full value when they first appear. That was the case with sea-floor spreading, which failed to catch the attention of more than a few readers at a time when most geologists and geophysicists were unprepared to take seriously the idea of mantle convection, much less that of the new notion of spreading seafloors. However, Dietz's paper elicited a favourable letter to *Nature* from J Tuzo Wilson at the University of Toronto, and it inspired Ewing to redirect a large portion of Lamont's research into testing the sea-floor spreading hypothesis. Ewing outfitted two ships with upgraded seismic reflection profilers to measure the thickness of pelagic sediments across the oceans. He was to find virtually no sediments on the ridges and a modest thickening of the layers towards the edges of the continental shelves. Ewing took the earliest photographs of deep sea sediments and observed ripple marks, which, until then, had been used as diagnostic of shallow waters.

In the next few years, as data favourable to sea-floor spreading accumulated, a regrettable negative reaction developed toward Dietz. Even though his model differed from that of Hess, the belief spread that he had 'stolen' Hess's basic idea and rushed it into print under his own name. This impression still persists to some degree, even though Dietz's choice of basaltic rather than serpentinized ocean floors is the one universally accepted today.

In 1986, Menard reviewed this controversy in his *The Ocean of Truth*, a personal account of the

sequence of events that led to plate tectonics. Menard, who had refereed the pre-publication manuscripts of both Hess and Dietz, wrote that he felt certain – and said so at the time – that Dietz was unaware of Hess's preprint when he wrote his own paper. Nevertheless, Menard believed that only one person can have priority for an idea and it should be Hess, so he urged Dietz, if only for appearances sake, to add a footnote to his next publication on seafloor spreading conceding credit for the idea to Hess. Dietz did so, and the two men made their peace in print. Subsequently, Menard came to realise that, in fact, more than one person can be struck with the same idea, particularly at a time when new data are coming in and are being freely discussed. He cited several additional examples that took place during the race towards plate tectonics.

Finally, Menard remarked that the priority for this idea probably should go to Arthur Holmes in Britain, who had favoured convection in the mantle as a ruling factor in global tectonics from the late 1920s until he died in 1965. Holmes described a basaltic layer which becomes a kind of endless travelling belt as it moves from ridges to trenches carrying continental fragments along with it. Over the years, Holmes changed his diagrams somewhat, but they all show the limbs of a convection cell rising under a continental slab, which is stretched and pulled apart into two fragments that are rafted to either side leaving behind new ocean floors. Holmes did not depict the creation of new ocean floor at a spreading oceanic ridge; he showed the ridges as being wholly or partially sialic. In that important respect his idea differed from those of Hess and Dietz. Nevertheless, Holmes's basic model was, without question, a predecessor of sea-floor spreading.

History books do not necessarily aid us in resolving disputes. In 1973, Allan Cox of the US Geological Survey omitted Dietz's article from his collection of the landmark papers that led to plate tectonics. And in the entry on Hess in Volume 17 of *The Dictionary of Scientific Biography*, published in 1990, we read that his hypothesis of sea-floor spreading was the most important innovation leading to plate tectonics, but that it was given its name by Dietz who, "with Hess's preprint in hand", published the first article on it in 1961. Fortunately, others knew better. In 1966, the Geological Society of America would present its highest honour, the Penrose Medal, to Hess for his research on the petrology of ultramafic rocks and for his provocative tectonic hypotheses including that of the spreading ocean floor. In 1988, the Penrose Medal went to Dietz for his "world-class, innovative contributions in three divisions of the geosciences: sea-floor spreading, recognition of terrestrial impact structures, and the meteorite impact of the Moon's surface".

Magnetic Anomalies on the Ocean Floors: 1961, 1963

In August, 1961, the paper by Mason and Raff, of which Dietz had seen a preprint, was published by the Geological Society of America along with a companion paper by Raff and Mason that included a particularly striking diagram showing long, narrow, mostly vertical black and white 'zebra stripes' of alternating high and low magnetic intensities measured on the floor of the north-eastern Pacific Ocean (**Figure 1**). The authors suggested that the contrasting intensities might reflect structural ridges and troughs or a system of sub-parallel dykes, but subsequent topographical and gravity surveys failed to detect either one. This research project, which proved to be momentous, was undertaken when Mason, a visitor to Scripps, casually asked a seismologist over morning coffee if anyone had thought of towing a magnetometer behind a ship to gather data while the ship was engaged in other operations. Revelle overheard the question and offered Mason the assignment, then and there. Mason soon learned that Lamont already had towed a magnetometer behind a ship in the Atlantic, and he could borrow it pending acquisition of an instrument by Scripps.

One reader, Lawrence W Morley of the Geological Survey of Canada, who had conducted extensive aeromagnetic surveys over lands and seas, remained mystified by this pattern for nearly two years until he discovered Dietz's paper on sea-floor spreading. Morley immediately wrote a short article and submitted it to *Nature* in February 1963, proposing a test of sea-floor spreading. He argued that the magnetic stripes form at the crests of spreading ridges where the erupting lavas acquire the magnetization of the Earth's ambient field (a detail Dietz had not specified). Today, the Earth's magnetic field is north-seeking, but in the 1950s it had become clear that some rocks have cooled during periods of south-seeking polarity, so Morley proposed that the parallel stripes on the moving sea-floor record the periods of normal and reversed polarity. And, inasmuch as Cretaceous rocks were the oldest yet recovered from the ocean floors, he speculated that such stripes have recorded the history of the ocean basins for the past 100 my or so. Morley tentatively calculated the rates of spreading and lengths of reversal periods. *Nature* rejected his paper saying they had no space available. Morley then sent it to the *Journal of Geophysical Research*, which kept it for some time and then rejected it, with a message from one referee saying that such speculation was more appropriate to cocktail party chatter.

On 7 September 1963, *Nature* published the now famous paper by Frederick Vine and Drummond

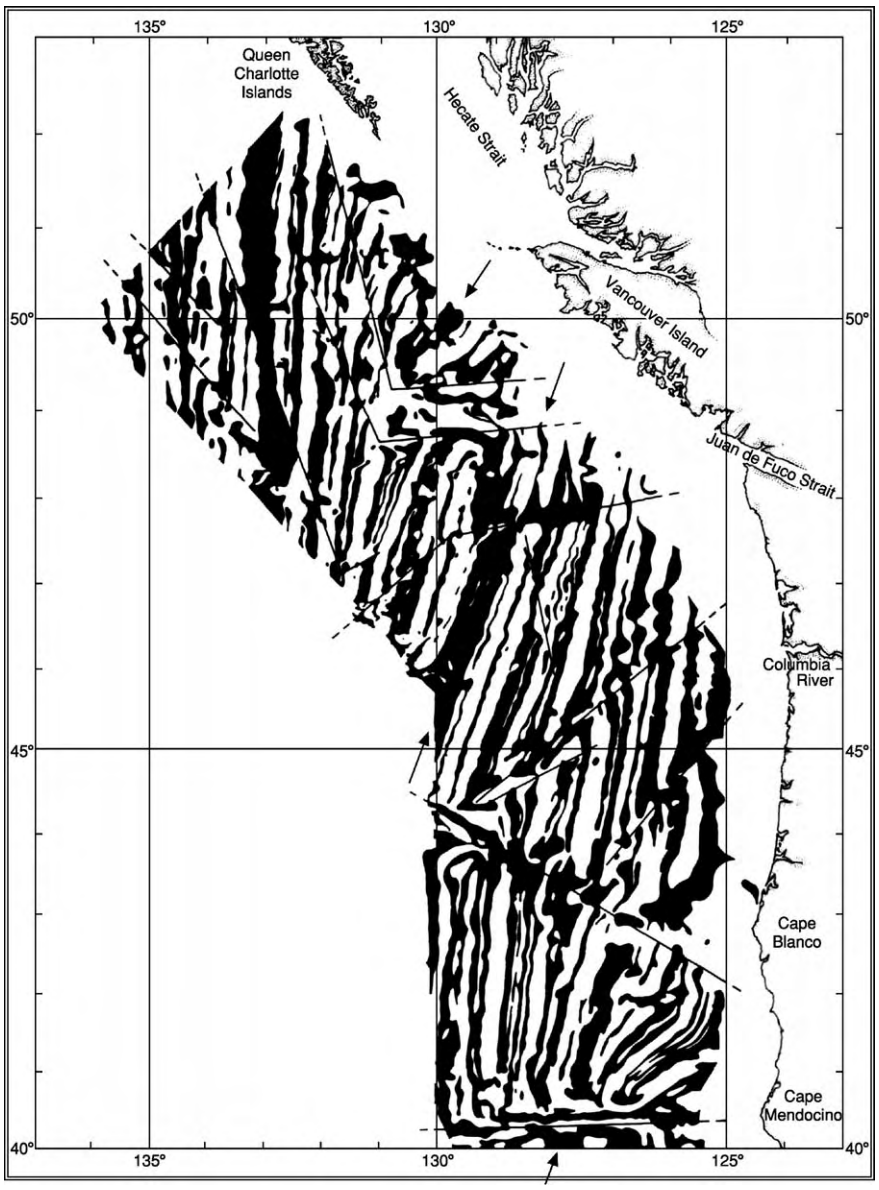


Figure 1 Magnetic anomalies offshore from Cape Mendocino, California, to the north of Vancouver Island, British Columbia. In 1961, Raff and Mason interpreted these stripes as most likely due to high and low magnetic intensities in linear patterns of mafic oceanic rocks. They drew the lettered lines as fault traces. In subsequent years, such 'zebra stripes' would be seen as evidence of magnetic field reversals with normally magnetized rocks (black) alternating with reversely magnetized ones (white). (Reprinted by courtesy of Raff and the Geological Society of America.)

Matthews of Cambridge University: 'Magnetic Anomalies over Oceanic Ridges'. Like Morley, Vine and Matthews assumed that convection, sea-floor spreading, and reversals of magnetic polarity all occur, and that the ocean floors consist of basalt that becomes strongly magnetized at the ridge crests. Today, the Vine–Matthews paper is widely seen as the founding paper of plate tectonics, but at the time it was poorly received and largely ignored for the next three years. It was not even included by their Department of Geophysics at Cambridge University in its list of important contributions for 1963! Years later, when attitudes changed, and Morley's story came out, the Earth science community began to speak of the Vine–Matthews–Morley (VMM) hypothesis. Morley then realized that he had gained more fame by having his paper rejected than he would have by its publication.

In 1963, all three of the basic assumptions listed by Vine and Matthews were suspect to geoscientists. Professor Harold Jeffreys, at Cambridge University, argued in every edition of his book, *The Earth*, beginning in 1924, that the mantle of the contracting Earth

is too stiff to allow for convection. In his sixth and final edition of 1976, he added several pages arguing against plate tectonics. In 1963 and again in 1964, Gordon MacDonald, of the Institute of Geophysics and Planetary Physics of the University of California, published a detailed paper, 'The Deep Structure of the Continents', in *Reviews of Geophysics* and a short version of it in *Science*. He intended this paper as a death-blow to continental drift and sea-floor spreading. MacDonald argued that isostasy prevails; thousands of measurements made on the surface and by satellites show that gravity over continents is equal to that over oceans – despite marked difference in their compositions and densities. This, together with the equal values of heat-flow determined worldwide indicated that the continents must have formed by vertical segregation of the mantle directly beneath them, which, therefore, must differ from the mantle under ocean floors to depths of 400 to 700 km. MacDonald concluded that no significant horizontal motion, due to convection or any other process, has occurred.

Magnetic Field Reversals, Isotopically Dated: 1964

Since early in the twentieth century, many advances had been made in techniques of measuring the remanent (permanent) magnetization of rocks and interpreting the results. In the 1950s, polar wandering curves had yielded evidence in support of continental drift. Nevertheless, inasmuch as Earth's magnetic field varies in alignment, intensity, and polarity, many geoscientists still viewed palaeomagnetism as an incomprehensible property, studied by 'black box' techniques, and scarcely to be trusted. However, by the early 1960s, measurements had begun to show that whereas recent basaltic lavas from widely spaced sources behave as north-seeking compasses, those of Early Pleistocene age behave as south-seeking compasses. Clearly, reversals of the magnetic field have taken place and the search was on for reliable means of dating them.

In June 1964, Allan Cox, Richard Doell, and Brent Dalrymple of the US Geological Survey at Menlo Park, California, published a paper entitled, 'Reversals of the Earth's Magnetic Field', in which they reported that they had dated rocks of normal and reversed magnetism by the K/Ar method. Their results documented the occurrence of two epochs of normal polarity and one of reversed polarity during the past 3.5 my. They named the epochs in honour of pioneers of palaeomagnetic studies (*see* **Palaeomagnetism**): the Bruhnes normal epoch, from the present to 1.0 my ago; the Matuyama reversed epoch from 1.0 to 2.5 my ago; and the Gauss normal epoch from 2.5 to 3.4 my ago. Both the Matuyama and the Gauss epochs were interrupted by short 'events' of opposite polarities – named, respectively, the Olduvai and the Mammoth events – each lasting 250–300 000 years. Later, in February 1966, Doell and Dalrymple revised this time-scale by shortening the present Bruhnes normal epoch to 700 000 years and adding the newly discovered 'Jaramillo normal event', which occurred about 0.85 my ago in the early part of the Matuyama reversed epoch. Since then, additional epochs and events have been dated for the past 180 my, back to the mid-Jurassic Period.

Cox and his group dated their basaltic samples using a mass spectrometer of the type designed in the late 1950s by John Reynolds of the University of California at Berkeley, which opened a new age in isotopic geochemistry. Reynolds' all-glass instruments could be heated to evacuate all traces of atmospheric argon in order to yield accurate measurements of the small amounts of argon produced by the radioactive decay of potassium. The work by Cox and his colleagues persuaded many scientists that palaeomagnetism must be taken seriously. Once again, however, another group was doing the same research. Within weeks of its appearance, *Nature* published 'Dating Polarity Geomagnetic Reversals' by Ian McDougal and Donald Tarling, who had established a palaeomagnetism laboratory at the Australian National University, and were friendly rivals of the group at Menlo Park.

Transform Faults: 1965

In June 1965, *Nature* published a paper by Wilson entitled: 'A New Class of Faults and their Bearing on Continental Drift', in which he proposed what he called 'transform faults' as a test of sea-floor spreading. These are faults with large horizontal movements that appear, along with their seismic activity, to terminate abruptly. He wrote that instead of just terminating, these faults are 'transformed', at each end, into mid-ocean ridges (*see* **Tectonics:** Mid-Ocean Ridges), mountain ranges, trenches, or island arcs (*see* **Tectonics:** Ocean Trenches), which together make up the network of mobile belts that divide the Earth's surface into large, rigid plates. Wilson cited the San Andreas as a transform fault on land with its southern end connected to the oceanic ridge in the Gulf of California and its northern end to one north of Cape Mendocino. But his clearest examples of transform faults were those that connect offset segments of spreading oceanic ridges. As shown in Figure 2, all the slabs of ocean floor on each flank of a ridge move side-by-side down-slope and across the seafloor. Nevertheless, along the transform fault plane connecting the ridge crests, and only there, the

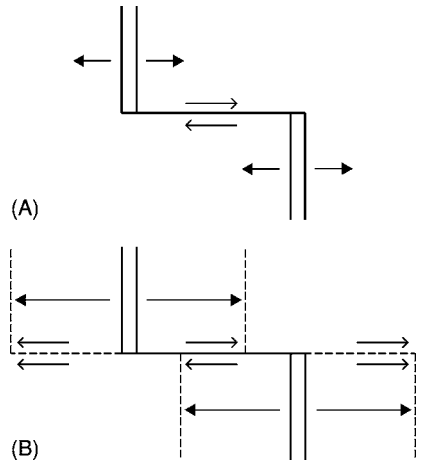


Figure 2 An evolving transform fault (A) Two expanding ridges are connected by a transform fault along which rocks moving down the eastern slope of the upper segment move in the opposite direction from those moving down the western slope of the lower segment. (B) After a lapse of time, spreading has advanced but the offset of the ridge has not changed, and movement along the fault still is confined to the ridge ridge section. The dashed extensions of the fault have become aseismic fracture zones. Note that along a transform fault, the hot eruptive zones on the ridge crests lie adjacent to cold walls of older, subsiding rocks on the flanks. (After Wilson (1965) combined by Mason in Oreskes, 2001; reproduced by permission of Oreskes to Mason and Westview Press.)

rock walls move in opposite directions. Wilson noted that in 1963 Lynn Sykes at Lamont had shown that along the fracture zones cross-cutting the ridges, seismic activity is strictly limited to these transform fault planes. Wilson attributed this to the relative motions of the rock walls and also to the marked differences in temperature, elevation, and age of the rocks across the fault – where hot rock beneath a ridge is juxtaposed with a cold wall of much older seafloor. How did the ridges become offset into segments? Wilson and Sykes both argued that the offsets formed in place when the lithosphere first split open at the beginning of the present cycle of ridge building. No strike-slip faulting separated them, nor has there been any along the fracture zones, which now were recognized as topographic continuations of transform faults that became aseismic when spreading carried them beyond the ridge crests.

Wilson's paper led to a major change in the history of structural geology. Its reception was mostly positive as geoscientists looked at their problems from the new perspective. But in March 1965, Alan Coode, at Newcastle University, had sent a short note illustrated with a block diagram of ridge segments connected by a transform fault (although he gave no new name to it) to the *Canadian Journal of Earth Sciences*. That journal appears quarterly, so Coode's article was published, to little notice, a week or two after Wilson's paper appeared in *Nature*. Great honours would be showered on Wilson; none on Coode, although he had submitted his paper first.

The *Eltanin* Profile: 1966

Sea-floor spreading got off to a slow start partly because neither the Vine–Mathews nor the Mason–Raff papers showed obvious symmetry of linear magnetic anomalies across oceanic ridges. Early in 1966, James Heirtzler, Xavier Le Pichon, and J G Baron at Lamont published the results of an aeromagnetic survey over the Reykjanes Ridge, south of Iceland, which yielded a beautifully symmetrical pattern of magnetic stripes across a spreading ridge. They computed the spreading rate at slightly less than 2 cm per year of separation.

In December 1966, Walter Pitman III, and Heirtzler, at Lamont published the magnetic profiles recorded by the research vessel, *Eltanin*, during four passes over the Pacific–Antarctic Ridge, south of Easter Island. All four profiles gave similar results but the most southerly one, *Eltanin*-19, was so spectacularly symmetrical, both in the topography of the ridge and its record of magnetic reversals that it seems to have prompted mass conversions to sea-floor spreading, first at Lamont and then elsewhere. The profile showed each of the dated magnetic epochs of the past 3.4 my, for which it yielded a computed spreading rate of 9 cm per year of separation. Assuming a constant spreading rate within 500 km on either side of the ridge, the profile made it possible to extend the series of geomagnetic reversals back to 10 my. Pitman and Heirtzler documented a good match between the anomalies in the South Pacific and those on the Reykjanes Ridge, adjusted for the slower spreading rate in the Atlantic. Today the *Eltanin*-19 profile is ranked as one of the most important pieces of evidence in the history of geophysics. And further confirmation was at hand. Approaching the problem by an independent method, Neil Opdyke and his colleagues at Lamont, plotted the magnetic polarities of fossiliferous strata from deepsea cores of the South Pacific floor. Early in 1966, they found a definitive match with the dated record in the basaltic oceanic bedrock.

In February 1966, Vine visited Lamont where Heirtzler gave him a copy of the *Eltanin*-19 profile. Vine incorporated it into his paper 'Spreading of the Ocean Floor: New Evidence', which appeared in *Science* the following December, shortly after one by Pitmann and Heirtzler. In it, Vine presented six symmetrical profiles of magnetic anomalies across

ridges in the Atlantic, Indian, and Pacific Oceans, and showed that the linear anomalies on the East Pacific Rise match those on the Juan de Fuca Ridge offshore from British Columbia, even though they lie 11 000 km apart. His computed spreading rates for all the ridges ranged from about 2.0 to 3.0 cm of separation per year in the Atlantic and Indian Oceans to 8.8 cm per year across the East Pacific Rise. Vine speculated that the whole history of the ocean basins in terms of ocean-floor spreading must be 'frozen-in' as paired magnetic anomalies in the oceanic crust. Meanwhile, Vine had given a summary of his results in November at the annual meeting of the Geological Society of America, where it startled many geologists with their first serious introduction to sea-floor spreading.

A Matching of Continents: 1966

At the same GSA meeting, Patrick Hurley, a geochronologist at the Massachusetts Institute of Technology, presented a paper, with nine co-authors, offering radiometric evidence for continental drift. After attending the Symposium on Continental Drift, organised in 1964 for the Royal Society by Patrick Blackett, Edward Bullard, and Keith Runcorn, Hurley said: "I went to London a fixist and came home a drifter". Bullard had displayed a map that had been programmed by a computer to apply Euler's theorem to find the best fit of the continents across the Atlantic. The map showed an especially close fit between Brazil and Ghana, where Hurley knew that a sharp contact had been mapped between two rock provinces that were 600 million years old and 2000 million years old, respectively. If Bullard's map were applicable, the same two provinces should occur near São Luis at the easternmost tip of Brazil. Hurley gathered a group of Brazilian and other collaborators who obtained rock samples from the region of São Luis and dated them. The results showed an excellent match between the petrology and dates of the two rock provinces now on opposite sides of the Atlantic. Hurley's research was crucial in persuading many American geologists to accept continental drift *via* seafloor spreading. In America, the doctrine of the permanence of continents and ocean basins, founded by James Dwight Dana at Yale in 1846, finally expired in 1967.

Plate Tectonics: 1967–1968

Of the many papers published during the emergence in 1967 and 1968 of plate tectonics from sea-floor spreading, four of them often are cited as milestones. The earliest of the four, 'The North Pacific: An Example of Tectonics on a Sphere', by Dan MacKenzie and David Parker, applied the concept of transform faults to motions on a sphere on which aseismic areas move as rigid plates. They, too, applied Euler's theorem (with which Wilson acknowledged he had been unfamiliar when he wrote of the movement of rigid plates) to the effect that any displacement of a plate on a spherical surface may be considered as a rigid rotation about a fixed vertical axis. It follows that the slip vectors of moving plates must describe small circles around the Euler pole. McKenzie and Parker used fault-plane solutions for earthquakes in the North Pacific to work out the actual directions of the slip vectors and found that they wholly conformed to the sea-floor spreading model.

The remaining three papers, all of which review the evidence in broad fields, appeared early in 1968, in Volume 73 of the *Journal of Geophysical Research*. In his paper entitled 'Rises, Trenches, Great Faults, and Crustal Blocks', Jason Morgan at Princeton presented the first depiction of the entire surface of the Earth divided into rigid blocks, each having three types of boundaries: ridges where new crust is formed; trenches or mountain ranges where crust is destroyed by sinking or shortened by folding; and great faults. His diagram showed twenty rigid blocks, large and small, identified by the seismicity at their boundaries.

Morgan's interpretation required that the seafloors must move as rigid blocks, for which the thin oceanic crust lacks the required strength. Therefore, he envisioned the outer layer of the Earth as a rigid 'tectosphere' about 100 km thick, sliding over the weak asthenosphere. He argued that the location of an oceanic ridge is determined, not by the activity of some deep-seated system of thermal convection, but by the motions of the blocks themselves: wherever tensional forces fracture the tectosphere to a depth of 100 km, hot mantle material wells up between the blocks and serves as a zone of weakness. Subsequently, each new intrusion of mantle material is injected into the centre of the most recent one, thus maintaining a median position for a spreading ridge in most cases. He argued that the emplacement of new ocean floor would be reflected in the symmetry of its magnetic anomalies and from these he charted the history of motions that brought the blocks to their present positions.

In an article titled: 'Sea-floor Spreading and Continental Drift', Xavier Le Pichon, then a visitor at Lamont, asked whether large crustal blocks do, in fact, move for significant distances without any deformation of the sea-floor sediments or the continental interiors. He adopted a simple model of six large

blocks (rather than Morgan's twenty), analysed their slip vectors, and found the results in reasonable agreement with the globe's physiographic, seismic, and geological data. Le Pichon then extended his analysis back over the past 65 my, looking for patterns of continental drift and sea-floor spreading. He reported evidence for three major episodes of spreading – one in the Late Mesozoic, one in the Early Cenozoic, and one in the Late Cenozoic – with a major reorganization of the global pattern occurring at the beginning of each cycle. He correlated the slowing down of each cycle of spreading with the onset of mountain-building. Le Pichon's analysis of surface motions provided a coherent and intelligible history of interrelated plate motions over the entire globe.

Lamont's group of seismologists contributed 'Seismology and the New Global Tectonics', by Bryan Isacks, Jack Oliver, and Lynn Sykes. Sykes previously had shown that not only was seismicity limited to transform fault planes but that first motions on spreading ridges also confirmed Wilson's predictions for transform faults. Now the group reviewed the full range of seismological evidence that supports plate tectonics. The spreading oceanic ridge-rifts are loci of shallow-focus normal faulting. The East Pacific Rise, which has no rift and few earthquakes, appears to be spreading rapidly by ductile flow. Deep-focus earthquakes occur only on planes dipping beneath continents and into trenches. At first the group had been surprised to observe that these quakes generate high-frequency seismic waves which travel faster and more efficiently along the seismic zones than had been expected. This led to their realisation that the quakes are generated within slabs of rigid lithosphere as they descend into the yielding asthenosphere – the process called subduction (*see* **Tectonics:** Ocean Trenches). These seismologists concluded that the sinking of cold, heavy plates plays a dominant role in generating plate motion. Their task of analysing earthquakes globally was aided immeasurably by the establishment by the US Coast and Geodetic Survey of the World Wide Standardized Network (WWSSN) of 125 stations from which data was continually transmitted to a central location for digitising and archiving. By the mid-1960s, the data were available on microfilm. The original purpose had been to distinguish between earthquakes and underground nuclear explosions, but the Network was an invaluable source of high-quality data for seismologists. The mid-1960s also saw the development of high-speed computers and of programs, capable of reducing reams of data from continuous recordings of many kinds. These, along with other technical advances, played an essential role in developing the theory of plate tectonics.

Plate Tectonics Today

Plate tectonics rapidly became the dominant model of geoscience. It simplifies our understanding of the three principal igneous rock types that make up the Earth's crust: the basalts of the ocean floors are derived by partial melting of the peridotite mantle and erupted at the spreading ridges; the more siliceous andesites are derived by partial melting plus dehydration of descending slabs of lithosphere and are erupted over subduction zones, where they continuously contribute to island arcs and continental margins; the still more silica-rich continental rocks arose from the growth of andesitic landmasses and their subsequent reworking by metamorphism and metasomatism to form granitic plutons and rhyolitic volcanics, with no contribution from the mantle. Beginning with the first andesitic volcanoes, the buoyant continents have grown in area throughout geologic history while the oceans floors have lost area to subduction. By one estimate, the rapidly moving Pacific floor may disappear about 200 million years from now, when North America will collide with Asia.

Our understanding of the mechanism driving plate tectonics remains incomplete. Many scientists have abandoned the idea of rolling convection cells, which would not form a coherent system beneath the segmented oceanic ridges. However, we know that the asthenosphere is close to the melting point, so some scientists argue that global stresses rent the lithosphere into the pattern of mobile belts resembling the seam on a tennis ball. To them, the cracks came first and the basalts of the spreading ridges arose from decompressional melting of the mantle. They also see the gravitational pull of the cold, heavy slabs of sinking lithosphere as a prime factor in maintaining plate motion.

But the mantle is by no means a quiescent domain. It includes very deep-seated 'hot spots' from which vertical plumes arise and erupt as volcanoes. Each hot spot persists for millions of years. A dramatic example is the hot spot beneath the big island of Hawaii, which has remained fixed-in-place for 60 my while the Pacific plate has moved over it carrying a long line of islands and seamounts with ages steadily increasing toward the north-west. There is even a bend in the line showing that about 46 my ago the plate shifted its motion from northwest to almost due north. We now know of more than 100 mantle plumes under continental areas, oceanic ridges, and islands (**Figure 3**). Surely these long-lived sources of magma make a powerful contribution to the thermal instabilities involved in plate motions. Today, with or without our having to specify details of the causal mechanism, the theory of plate tectonics seems secure.

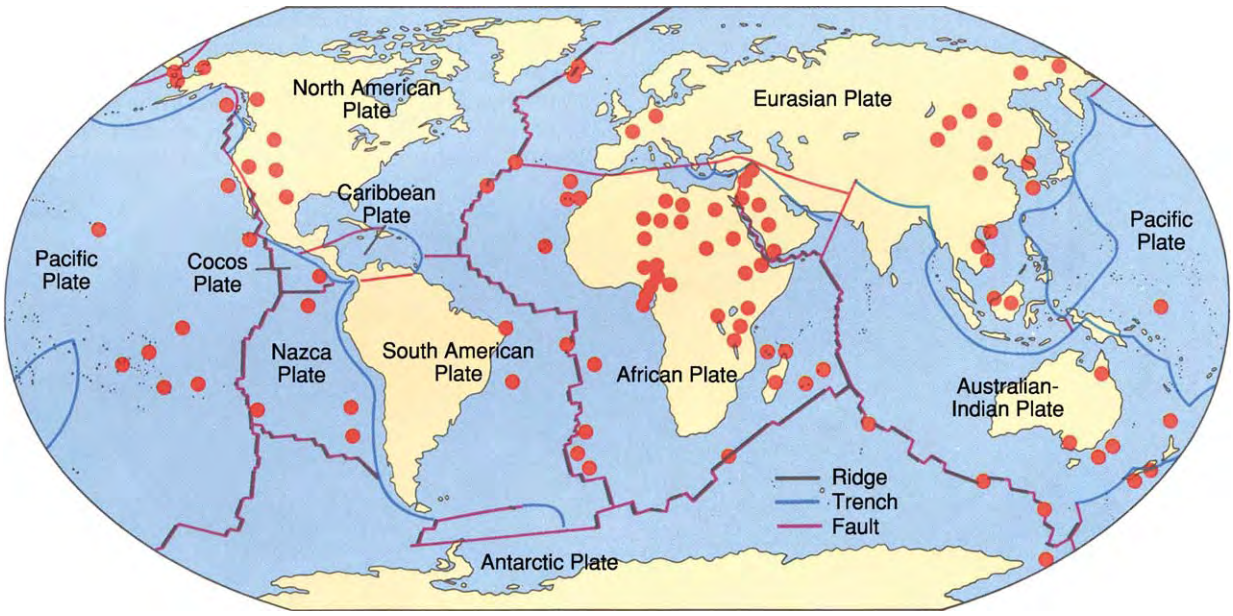


Figure 3 Sketch map showing the Earth's seven major and three minor plates bounded by oceanic ridges, trenches, and faults, and 101 fixed hot spots from which plumes of magma arise from deep within the mantle. Note that Africa has an unusually high number of hot spots for a continental area. This, together with the fact that Africa is almost surrounded by spreading ridges, suggests that this continent has remained fixed in place for a very long time. (From Skinner BJ and Porter SC (1995) *The Blue Planet*; reproduced by permission of Skinner and John Wiley and Sons.)

Is Plate Tectonics Unique to the Earth?

Only two of the more than 100 planets that we have discovered outside our solar system have actually been seen as black spots passing in front of their suns. All the rest are inferred from irregularities in the orbits of their suns. Therefore, a search for plate tectonics outside the Earth must be limited to our four rocky neighbours in the inner Solar System – our Moon, Mercury, Venus, and Mars – and some of the moons of the outer planets. We have analysed samples collected on the Moon, and analysed those rare meteorites that have been projected to Earth by impacts on the Moon and Mars. In addition, we have landed analytical instruments on the surfaces of the Moon, Venus, and Mars. All our results show that the crusts of these bodies consist of basalts and related rocks that are akin to those of our ocean floors. As yet we have found no evidence of granitic continents.

Scientists have searched in vain on images of these bodies for patterns indicative of plate boundaries (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges): long ridges like spreading zones, chains of folded mountain ranges, or lines of volcanoes. Our Moon and Mercury are volcanically dead and pockmarked by impact craters. Venus, beneath its thick layer of clouds, appears to have been resurfaced with fresh basalts about 5 my ago, but it shows no linear pattern of features resembling ridges or subduction zones. Mars has the largest volcanoes in the Solar System, but they are scattered over a large area, and we see no evidence there of linear mountain ranges or trenches. Even less promising are the icy moons of the giant planets. We conclude that plate tectonics is unique to the Earth.

In this, we are fortunate, because plate tectonics is one of the essential factors that enable our planet to support complex forms of life. Volcanic eruptions continually renew Earth's waters and its atmosphere, including ozone, that shields us from lethal solar radiation, and the greenhouse gases that moderate the temperature and maintain it within a livable range for mammals. The creation of new islands and splitting and rearranging of continental areas generate a great variety of habitats which, in turn, minimizes the effectiveness of mass extinctions from whatever cause. Were plate motion to cease, mountain building (*see* **Tectonics:** Mountain Building and Orogeny) would cease, and uninterrupted erosion would begin to wear down the continents. Ultimately, the deposition of sediments in the oceans would presumably raise sea-level enough to flood much of the land surface. We literally owe our lives to plate tectonics.

See Also

Earth: Mantle; Crust. **Famous Geologists:** Du Toit; Wegener. **Gondwanaland and Gondwana**. **History of Geology From 1900 To 1962**. **Palaeomagnetism**. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Faults; Mid-Ocean Ridges; Mountain Building and Orogeny; Ocean Trenches; Rift Valleys.

Further Reading

Blackett PMS, Bullard E, and Runcorn SK (1965) A symposium on continental drift. *Philosophical Transactions of the Royal Society* London, Series A, 258, No. 1088.

Cox A (ed.) (1973) *Plate Tectonics and Geomagnetic Reversals*. San Francisco: WH Freeman and Company.

Marvin UB (1973) *Continental Drift: The Evolution of a Concept*. Washington, DC: Smithsonian Institution Press.

Menard WH (1986) *The Ocean of Truth: A Personal History of Global Tectonics*. Princeton: Princeton University Press.

Oliver JE (1966) *Shocks and Rocks: Seismology in the Plate Tectonics Revolution*. Washington, DC: American Geophysical Union.

Oreskes N and Le Grand HE (eds.) (2001) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Boulder: Westview Press.

Skinner BJ and Porter SC (1995) *The Blue Planet: An Introduction to Earth System Science*. New York: John Wiley and Sons, Inc.

Ward PD and Brownlee D (2000) *Rare Earth: Why Complex Life is Uncommon in the Universe*. New York: Copernicus, Springer Verlag.

IGNEOUS PROCESSES

P D Asimow, California Institute of Technology, Pasadena, CA, USA

Introduction

All phenomena that involve rock in the liquid (molten) state are igneous processes. Molten or partially molten rock at depth in the earth is called magma; if erupted onto the surface at a volcano, it is then called lava (*see* **Lava**). Igneous processes involve making magma from previously solid source rocks, moving magma from one place to another, and freezing magma to produce igneous rocks. Although all of these steps may overlap, it is convenient to separate a discussion of igneous processes into three separate processes: melting processes, such as decompression melting, flux melting, conductive heating, and impact melting; magma transport processes, such as porous flow, dike injection, pluton emplacement, and volcanic eruption; and differentiation processes, such as fractional crystallization, magma mixing, and assimilation.

Igneous processes are distinguished from metamorphic processes, in which minerals react to form new rocks but melting does not occur, and from aqueous or hydrothermal processes, in which the composition of the fluid phase is dominated by water or other highly volatile components rather than by molten rock-forming minerals, although both metamorphism and hydrothermal activity are closely associated with igneous activity. Igneous processes are an essential part of the rock cycle; they lead to the formation of the wide diversity of observed igneous rocks and they constitute the fundamental process by which Earth and the planets differentiate into crust and mantle. The surface of the Earth reflects sources and processes at depth; by delivering material to the surface, igneous processes provide geologists and geochemists with essential information from which to constrain the compositional, mineralogical, and thermal character of Earth's interior. On Earth, igneous processes are closely related to plate tectonics, in that igneous activity is concentrated at divergent and convergent plate boundaries, although intraplate magmatism occurs as well. The types of igneous processes that occur in particular tectonic environments and the types of igneous rocks that result are generally characteristic of those environments. Therefore, distinctive associations of ancient igneous rocks, together with distinctive structural, metamorphic, and sedimentary features, can be used to identify ancient tectonic environments by comparison to modern examples.

Igneous processes are closely linked to the igneous rocks that are their final products. Basaltic magmatism (creating rocks such as basalt, gabbro, and komatiite) results from partial melting of planetary mantles of ultramafic composition; on Earth, this is seen at mid-ocean ridges, back-arc basins, large igneous provinces, continental rifts, intraplate hotspots, and subduction zones. Intermediate and silicic magmatism (creating rocks such as andesite, diorite, rhyolite, obsidian, and granite) is restricted to continental crust and is most common above subduction zones (*see* **Igneous Rocks:** Granite; Obsidian). Many economically significant ore deposits are associated with silicic magmatic activity; these are known as magmatic ores (*see* **Mineral Deposits and Their Genesis**).

Melting Processes

An important aspect of the modern understanding of igneous processes is that the interior of Earth is predominantly solid. This has not always been recognized; early geologists supposed that volcanic eruption of lava resulted from tapping of some permanently molten or glassy layer within Earth. Seismology, however, shows that Earth's crust and mantle transmit shear waves, which requires them to have rigidity. It is also known that the source regions of magma in the interior of Earth are too hot for glass to persist over geologic time. The shallow part of the mantle magma is less dense than solid rock and will rapidly separate and ascend rather than be stored for long times, whereas the pressure in the deeper part of the mantle is high enough to keep rocks in the solid state despite their high temperatures. Therefore, regions of volcanic activity are not simply areas where magma is tapped, but, first and foremost, are areas where magma is being created by melting of previously solid rocks.

In common experience, a solid material (e.g., an ice cube) is melted by first increasing its temperature to the melting point and then supplying the latent heat of fusion to transform the solid into a liquid. Intuitively, then, it might be imagined that melting of rocks is accomplished by increasing the temperature of the source rocks or by heat flow into the melting region. In this case, intuition is misleading; the dominant melting processes inside Earth are decompression melting and flux melting, both of which occur without increase in temperature or flow of heat, because the energy necessary to drive melting under the right circumstances is

already present in the source material. Furthermore, the phase relations of rocks are somewhat more complicated than can be understood by thinking of a single melting point, such as that of ice at ambient conditions. First, even for a pure substance such as ice, or for certain model mineral systems, the melting temperature is a function of pressure; for most substances, the melting temperature increases with pressure. Also, natural rocks are not pure substances but are mechanical mixtures of several minerals, each of which is a solid solution of many chemical components. This causes the melting temperature to widen into a range of partial melting temperatures, even at constant pressure. The point at which melting of a given rock begins with increasing temperature is called the solidus; the locus of such points at various pressures is the solidus curve. The point at which melting is complete and the last crystal disappears with increasing temperature is called the liquidus; the locus of such points at various pressures is the liquidus curve.

The solidus and liquidus curves for the most common composition in Earth's mantle are shown in Figure 1. These curves provide a context for much of the discussion of melting, because they show the pressure and temperature conditions at which mantle rocks partially melt, as well as the temperatures that actually occur in Earth's upper mantle in various tectonic regions. The gradient of temperature over the first several hundred kilometres in depth below a given place on Earth is divided into two intervals. The lithosphere generally encompasses the entire crust and that part of the upper mantle that is cold enough to act as a rigid material. The lithosphere can deform by faulting, but vertical motions are limited and hence heat transport through the lithosphere is dominated by conduction. Transportation of the geothermal heat flow by conduction causes the lithosphere to sustain a large thermal gradient; temperature increases with depth in the lithosphere by $500\,\mathrm{K\,GPa^{-1}}$ ($16^{\circ}\mathrm{C\,km^{-1}}$) or more. With increasing temperature, the plastic strength, or viscosity, of mantle rocks decreases to the point at which it becomes more efficient to transport geothermal heat flow by convection (in other words, by upwards vertical flow of hot material and downwards vertical flow of colder material at plate-tectonic rates of up to tens of centimetres per year). Heat transport by convection tends to drive the vertical temperature gradient towards the value that a parcel of material follows when its pressure changes without addition or removal of heat, the adiabatic gradient, which for mantle minerals is about $10\,\mathrm{K\,GPa^{-1}}$. The part of the mantle dominated by convective transport and displaying a nearly adiabatic temperature gradient is called the asthenosphere.

The temperature within the asthenosphere is characterized by a quantity called potential temperature, which is the temperature obtained by extrapolating the adiabatic gradient to the surface as if there were no lithosphere (and no melting). The potential temperature of the upper mantle sampled by mid-ocean ridges is thought to vary between about 1250° and 1400°C, whereas hotspots such as Hawaii may be underlain by hot plumes with potential temperature up to 1500°C. The temperature gradient and heat flow in the lithosphere are governed by the age and thickness of the lithosphere. For a normal potential temperature of 1350°C, Figure 1 shows a mid-ocean ridge thermal profile, wherein the zero-age lithosphere is only 10–30-km thick and the asthenosphere temperature profile extends nearly to the surface. Note that this temperature profile crosses the dry solidus of peridotite. Also plotted are an old oceanic thermal profile with a roughly 100 km-thick lithosphere and a stable continental thermal profile with a 150-km-thick lithosphere, both of which cross the water-saturated solidus of peridotite but stay well below the dry solidus.

Decompression Melting

Decompression melting drives basaltic volcanism at mid-ocean ridges, hotspots, back-arc basins, and may contribute to some extent in subduction-related island arcs and continental arcs as well. To understand decompression melting, it is necessary to consider the temperature structure of Earth, the phase relations of mantle rocks, and the existence of vertical flows in the asthenosphere that are fast enough to move material around without significant heat exchange with the surroundings.

The surface of Earth is cold, too cold for molten rock to be stable. At depths of more than a few hundred kilometres, the interior of Earth is everywhere at high enough temperature that, at ambient pressure, mantle rocks would be at least partially molten. However, at these depths, the pressure is high enough to raise the solidus temperature well above modern asthenosphere temperatures. In other words, although the temperature increases with increasing pressure in the asthenosphere by about $10\,\mathrm{K\,GPa^{-1}}$, the dry solidus temperature of mantle rocks increases by about $130\,\mathrm{K\,GPa^{-1}}$. Conversely, then, if solid material in the asthenosphere can flow upwards more quickly than heat can be conducted out of it, it will cool by only $10\,\mathrm{K\,GPa^{-1}}$ while the solidus temperature decreases by $130\,\mathrm{K\,GPa^{-1}}$ and the upwelling material will intersect the solidus and begin melting. Therefore, the expectation is to find basaltic magma anywhere that horizontal extension of the lithosphere allows the asthenosphere to

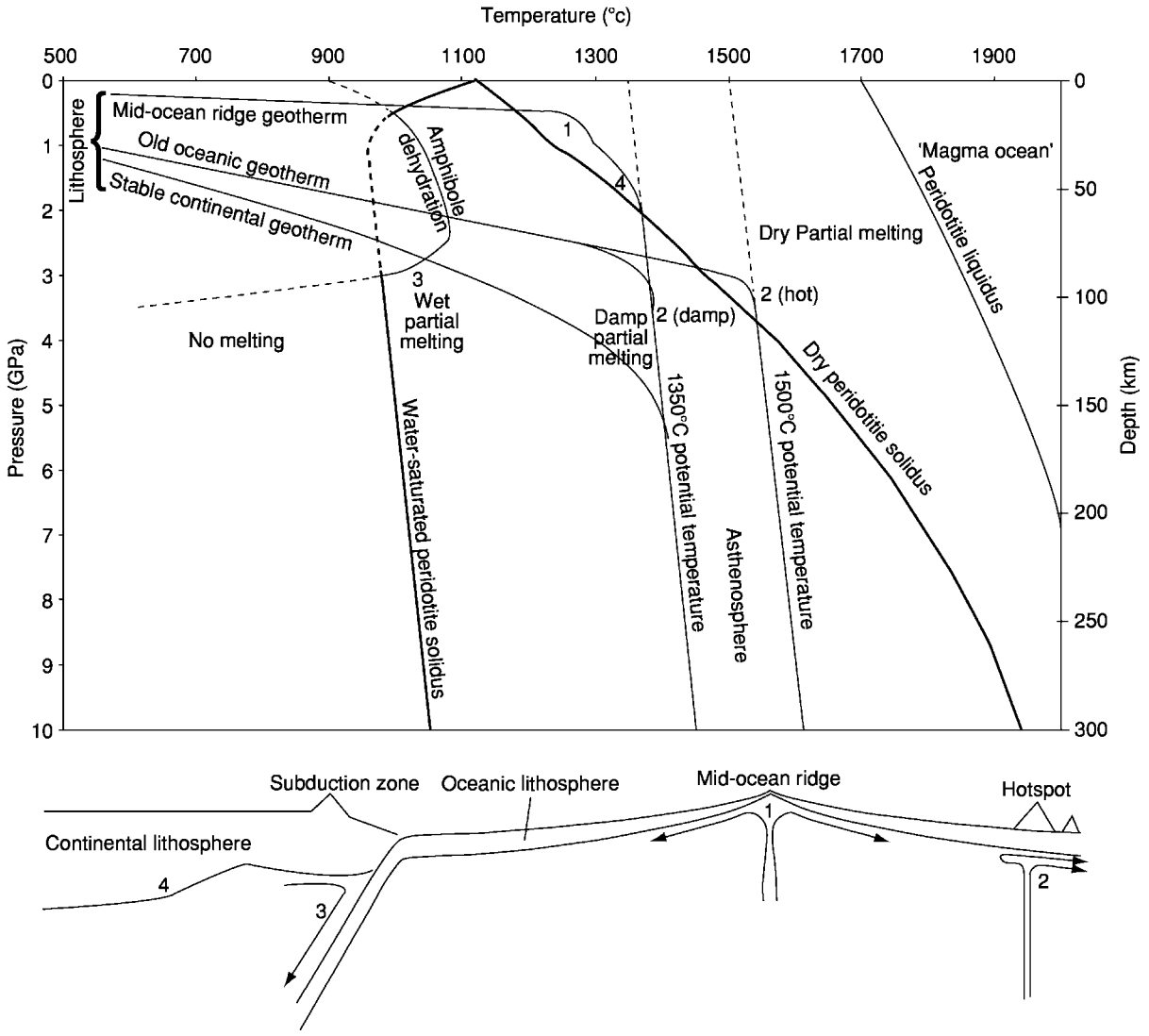


Figure 1 The pressure temperature structure and physical locations of major igneous processes on Earth. Curves shown in pressure temperature space include the liquidus of peridotite, the major rock composing the Earth's upper mantle; the solidus of peridotite under dry conditions; the solidus of peridotite in the presence of excess water; and the breakdown of the hydrous mineral amphibole, which can store water and prevent wet melting as long as it is stable. Compared to these underlying phase relations, the temperature structures of lithosphere and asthenosphere underlying mid ocean ridges, hotspots, and continents are shown. This shows that melting of peridotite is expected at low pressure and dry conditions under mid ocean ridges (1), at high pressure and either hot or damp conditions at intraplate hotspots (2), and at wet and cold conditions above the water saturated solidus and amphibole stability limit within subduction zones (3).

flow upwards with minimal cooling and fill the resulting gap. This is the case at mid-ocean ridges (region 1 in Figure 1) and in well-developed continental rifts (region 4 in Figure 1). At higher potential temperatures or higher mantle water contents, the asthenospheric temperature profile will cross the peridotite solidus at higher pressure, so that upwelling flow can drive melting even beneath a thick lithosphere. This is the case at hotspot localities such as Hawaii (region 2 in Figure 1) and in large-scale igneous provinces such as the Ontong-Java plateau or the Columbia River flood basalt (*see* **Large Igneous Provinces**). Whether decompression melting occurs in any particular environment depends on the potential temperature, the water content of the source rocks (see later), and the extent to which the lithosphere is thinned or removed to allow the asthenospheric temperature gradient to extend close to the surface.

It is important to realize that decompression melting is always a partial melting process; there is always a solid residue left behind in the mantle. Although a completely liquid magma can separate from the residue, this is the result of magma transport

(see later), not of complete melting. With continued decompression above the point of intersection with the solidus, partial melting continues and reaches higher extents of melting for longer melting columns. Thus, the volume or flux of magma generated depends on the same variables that determine whether melting occurs in the first place (lithosphere thickness, potential temperature, and water content) as well on as the flow rate of mantle source rocks through the melting region.

Flux Melting

At subduction zones, a cold slab of old oceanic lithosphere flows down into the mantle and induces downwards corner flow of the mantle wedge above the slab and below the thick lithosphere of the overlying plate. The shape of the dry solidus of peridotite and the energetics of decompression melting show that melting is to be expected under thin lithosphere, where the mantle is hot, and where the flow direction is upwards. Subduction zones presumably show none of these characteristics, yet most of Earth's most recognizable volcanoes and nearly all of the hazardous ones form above subduction zones. Clearly, a great deal of melting occurs in such settings. This primarily results from the introduction of water into the mantle wedge via subduction of oceanic sediments and hydrothermally altered oceanic crust. At high pressures, when water can readily dissolve in magma, it acts as a flux; Figure 1 shows that the water-saturated solidus of peridotite is several hundred degrees below the dry solidus. At intermediate water contents, melting begins on a damp solidus curve between the two limits shown. The process of melting by addition of water to material below its dry solidus, but at or above the wet solidus, is called flux melting, which is the second most efficient source of magmatism on Earth.

Subduction zone magmatism is complex, but generally, there is consensus on the essential elements. The subducting slab remains relatively cold, and only the sedimentary component is thought to melt directly, creating a mobile liquid that ascends into the shallow part of the overlying mantle wedge and refreezes, modifying the composition of the wedge. At somewhat greater depth, the basaltic component of the slab undergoes a series of dehydration reactions that create a water-rich fluid. The stability limits of hydrous minerals in peridotite are different from those in basalt, so that as this fluid ascends out of the slab and into the immediately overlying cold part of the mantle wedge, it may also freeze in place and generate an enriched, hydrated mantle source. However, with further downflow along the slab, or if the fluid can migrate far enough into the hot interior of the mantle wedge, this material crosses the hydrous solidus (region 3 in Figure 1) and partially melts to create primary arc basalt.

Heating by conduction At locations remote from plate boundaries and hotspots (for example, in continental interiors), volcanism and magmatic intrusion do occur, requiring a mechanism other than decompression or flux melting. These locations include large rhyolitic caldera-forming systems such as Long Valley, in California. Their source materials are embedded in the lithosphere (which is both too cold and too rigid for decompression melting to be effective) and, on a stable long-term geotherm, sit below even the water-saturated solidus curve. Direct heating by conduction is therefore the most likely mechanism for bringing these sources into their melting range. Why should conduction deliver more heat in one place than in another? The answer generally goes back to decompression melting in the underlying mantle. Geochemical and geophysical evidence typically shows that, although the principal source of intraplate volcanism may be within the crust, there is often a mantle component. Ponding of basaltic magma and subsequent underplating of basaltic rocks at the base of the crust are the most efficient ways to focus a large heat flow into a specific region of the crust. Because basalts crystallize at temperatures above 1000°C and the rocks of the continental crust can begin melting (in the presence of water) near 700°C, it is clear that crustal melting is a likely consequence of the arrival of a large mass of basalt at the base of the crust. If the basalt actually intrudes the crust and assimilates or mixes with the resulting melts of surrounding crustal rocks, the process may be described as a differentiation process (see later) affecting the basalt, as well as a melting process affecting the wallrock.

Magma Transport

The transport of magma from regions of melting to regions of emplacement or eruption is a fundamental aspect of igneous phenomena; indeed, the outpouring of lava or explosive eruption of volcanic ash is the most obvious and hazardous manifestation of igneous activity. Most of the melting processes that occur inside Earth are partial melting processes, producing a mixture of liquid and residual minerals. Somehow the liquid component of this mixture is physically separated from the residue and is transported, generally to shallower depths. Thus, for example, a mid-ocean ridge basalt created by 10% melting of its mantle source may erupt as a 100% liquid, because the liquid and residue have been separated by melt

migration. Indeed, in the absence of separation between melts and residues, igneous processes would be unable to drive differentiation of Earth into mantle and crust, and would be unable to produce many of the eruptive phenomena associated with volcanism.

There are several mechanisms of magma transport to discuss (Figure 2), but all of them are driven essentially by the same force: gravity acting on the buoyancy of melts or melt–gas mixtures relative to residual minerals or wallrocks. That is, when a rock partially melts, at least at the pressures at which most melting takes place inside Earth today, the liquid is less dense than its surroundings are. Gravity will therefore cause the melt to rise, if a pathway is available that allows it to do so. Such a pathway can be established by intergranular porous flow, if the liquid

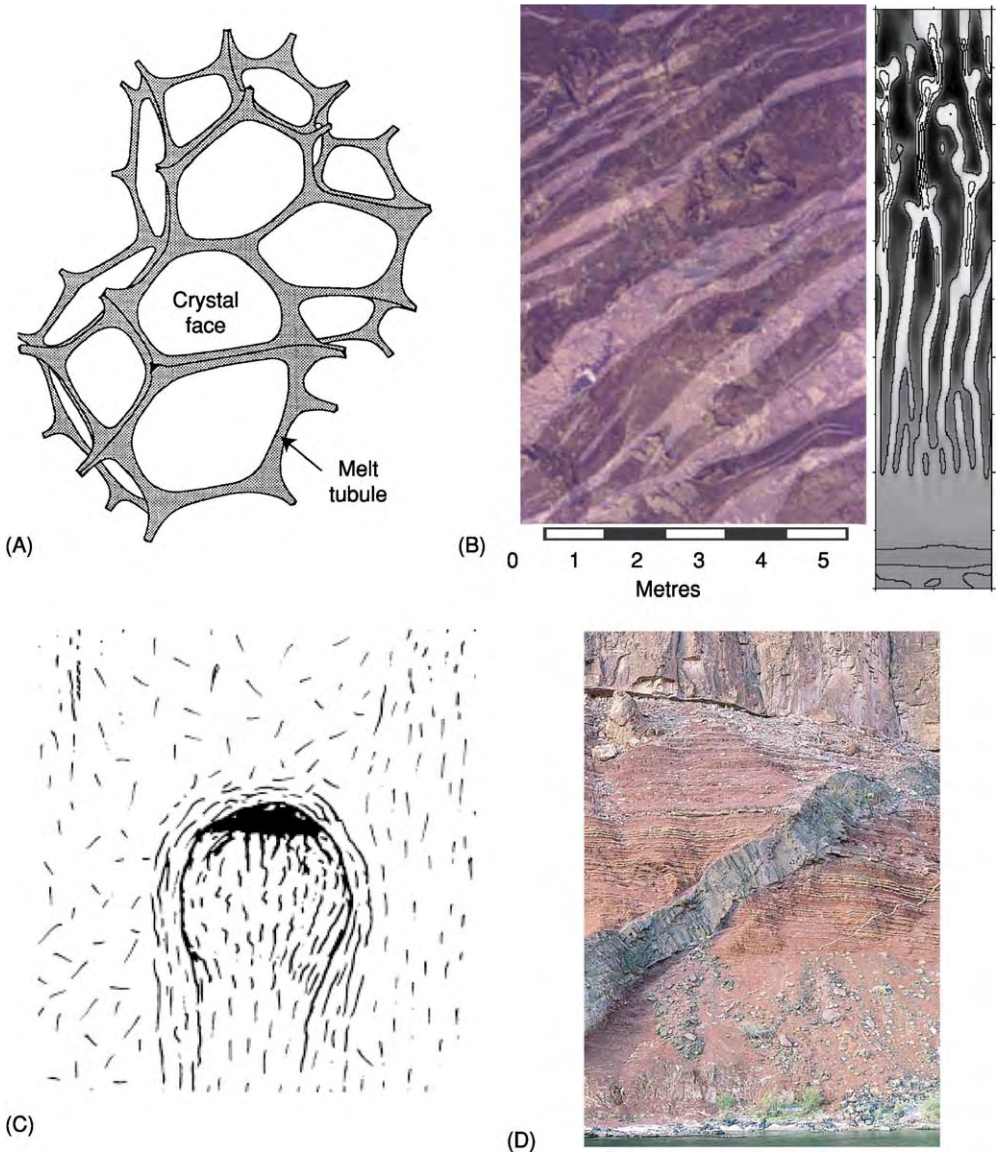


Figure 2 Illustrations of magma transport mechanisms. (A) During partial melting of mantle rock, the melt phase organizes into an interconnected network of tubules along the intersections where three or four solid crystals of the residual rock meet. This allows the melt to move by porous flow. (B) Bands of dunite (light colour, olivine only) and harzburgite (dark colour, olivine plus orthopyroxene) in a section of former oceanic upper mantle exposed in the Sultanate of Oman; the adjacent diagram is a numerical model of the development of such structures by porous flow, in which melt is able to develop preferred flow pathways by dissolving orthopyroxene. (C) Sketch of a buoyant diapir of partially molten rock rising through denser asthenosphere containing less melt. (D) A dyke of basalt intruded into sedimentary rocks in the Grand Canyon, Arizona, USA. (C) Reprinted with permission from Basaltic Volcanism Study Group (1981) *Basaltic Volcanism on the Terrestrial Planets*, p. 499. (B) Reprinted with permission from Braun MG and Kelemen PB (1992) Dunite distribution in the Oman ophiolite: Implications for melt flux through porous dunite conduits, with permission from American Geophysical Union. (D) Reprinted with permission from Kelemen PB, Hirth G, Shimizu N, Spiegelman M, and Dick HJB (1997) *Philosophical Transactions of the Royal Society of London, Series A* 355, p. 296.

forms an interconnected network along the grain boundaries of the partially molten rock, or by opening a crack or dyke in the overlying rock to create a conduit for melt flow. Also, the partially molten assemblage is less dense than is the unmelted rock of the same bulk composition, and therefore melts can rise by diapirism, or upward flow of the entire partially molten region, so long as the surroundings are weak enough to allow such flow. Finally, as melts containing dissolved volatiles reach low pressures in the crust, the dissolved gas components form bubbles. This leads to a runaway drop in density, increase in buoyancy, faster uprise, and continued bubble growth; the end result may be fragmentation of the magma as bubbles begin to touch each other, and the explosive eruption of tiny shards of volcanic glass called ash.

Porous Flow

The initial stage of the melt separation process is presumably porous flow. Experiments on the texture of partially molten mantle-like rocks shows that, at mantle temperatures and small degrees of partial melting, the melt phase organizes into a network of tubules along the boundaries of crystals that make up the solid residue (Figure 2A). The melt thereby establishes an interconnected network through which it can migrate relative to the solid. Such porous flow is driven by pressure gradients, principally due to gravity but also due to shearing forces as the host rock deforms, and resisted by the viscosity of the magma and the permeability of the host rock. Although the exact relationship is unknown, permeability is an increasing function of the fraction of melt present, such that more melt can flow through a region where more melt is present. Porous flow is a rather slow melt migration process and is thought to allow continuous chemical equilibration between melts and solids, under mantle conditions. Such chemical equilibration can, in some cases, lead to extra melting and an increase in local melt fraction. Because this increases the permeability, increased flow can lead to yet more extra melting and an instability can develop whereby a porous flow system evolves into a set of high-porosity, high-flux conduits embedded in a low-porosity, low-flux matrix. This process is thought to be important in the rapid extraction of basaltic magma from the mantle, in determining the chemical characteristics of mid-ocean ridge basalt, and in explaining the rock types and distributions seen in outcrops of former oceanic mantle rocks (Figure 2B).

Diapirism

In addition to relative flow between melt and solids in a partial melt, the bulk partially molten assemblage can migrate relative to surrounding regions (Figure 2C). When a mass of bulk partial melt flows upwards because of its thermal or melt-induced buoyancy, this is called a diapir. The significance of diapirism in many magmatic settings is unclear. There may be a component of active, buoyancy-driven flow beneath mid-ocean ridges, but, in general, this is not necessary to explain the existence of ridges or the eruption of magmas on-axis. There may be buoyant upwellings of hydrated and/or partially molten material from near the slab in subduction zones, and this may help to explain the large degrees of melting and large volumes of melt emplaced in such settings, but, again, this is controversial. The most likely setting in which diapirism is essential to explaining observed field relations is in the emplacement of granitic plutons within continental crust.

Dyke Injection

In the lithosphere, porous flow becomes an inefficient means of moving melts. The low temperatures and conductive heat flow imply that melts migrating slowly through the lithosphere will begin to freeze. This reverses the permeability feedback that occurs in the asthenosphere, and chokes off melt conduits where melt is able to react with wallrocks. Also, the plastic strength of minerals increases with decreasing temperature to the point at which the host minerals are unable to compact or expand to accommodate changes in melt fraction. On the other hand, this allows differential stresses to accumulate rather than relax. As pressure decreases, the differential stress needed to cause brittle failure and crack propagation decreases as well. Hence, at shallow depths, melt migration becomes dominated by the formation of cracks and the flow of melt through the resulting conduits. This process is called dyke injection, and the tabular bodies of igneous rock that end up frozen in such crack-related conduits are called dykes (Figure 2D). By its nature, dyke injection is an episodic process; it requires stress to build to the point of failure and it requires a large pool of melt to flow suddenly into the dyke and maintain stress at the crack tip. On the other hand, porous flow is a continuous process. The transition from porous flow to dyke injection, which is associated in some way with the transition from asthenosphere to lithosphere, is therefore a likely location for melt to pond and accumulate in some temporary storage reservoir. This is one mechanism for developing a magma chamber, the primary site where differentiation takes place (see later).

Eruption

When the products of igneous activity within Earth, in the form of lava or ash, flow or explode onto the surface, the site of eruption is called a volcano.

This is a type of igneous process, clearly involving melt transport. (*see* **Pyroclastics, Volcanoes**).

Differentiation Processes

Much of the diversity of igneous rock compositions found on Earth results from the evolution of magmas once they have migrated away from the melting region of their source and have arrived at a shallow enough depth to begin crystallizing. The sets of processes that act on a primary magma after separation from its source are collectively called differentiation, and they include partial crystallization, assimilation of foreign material, and mixing of various magmas.

Just as the melt phase composition is distinct from the residue composition during melting, allowing separation of chemical components by partial melting, so the crystals that form during differentiation differ in composition from the liquid, and so by mass balance cause the liquid to evolve in composition as it cools and crystallizes. During crystallization, the liquidus is the temperature at which the first crystals appear and the solidus is the temperature at which the last bit of liquid freezes.

Fractional Crystallization

The dominant process relating a suite of igneous rocks within a given locality is often fractional

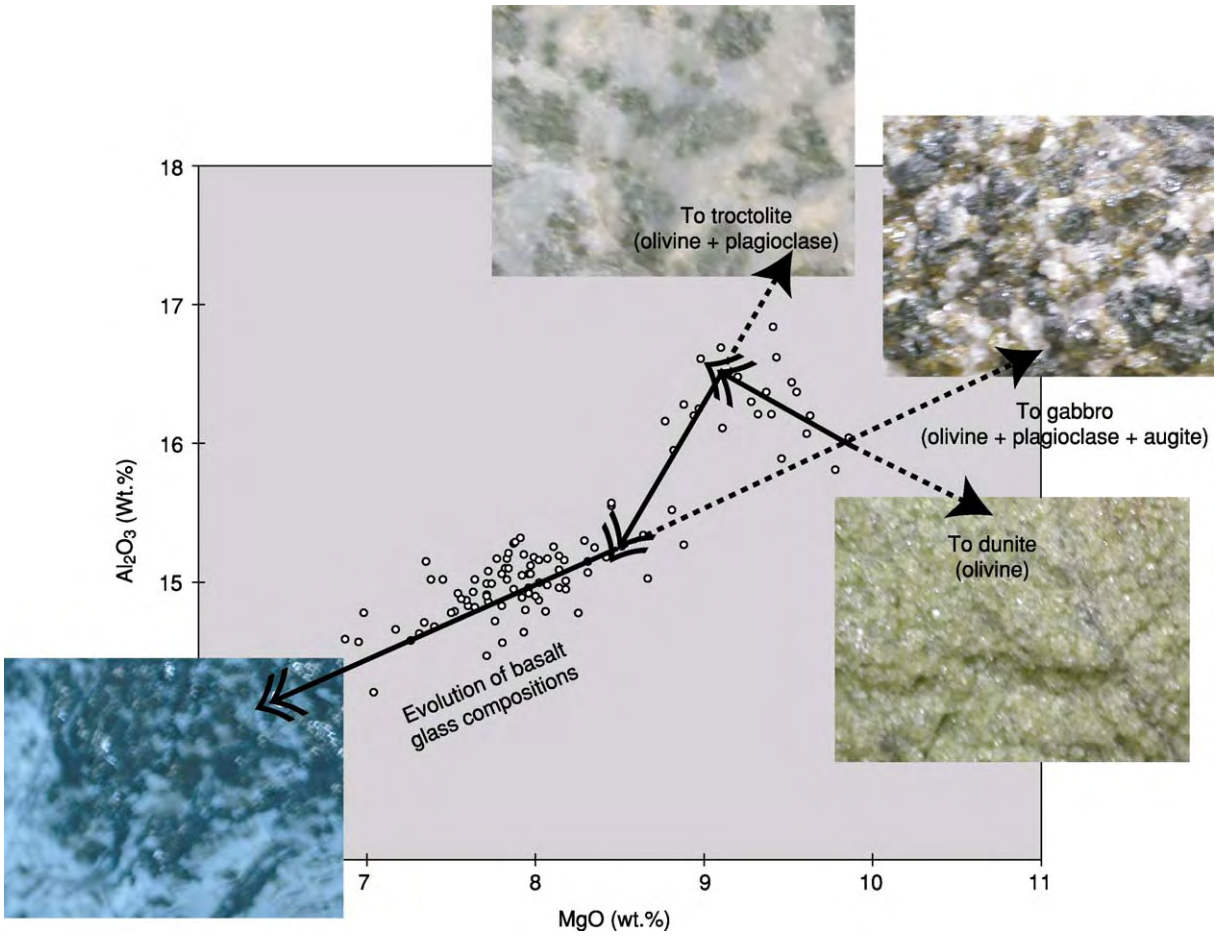


Figure 3 Illustration of complementary formation of plutonic and volcanic rocks by fractional crystallization. The graph shows MgO and Al_2O_3 contents measured in a suite of samples of volcanic glass recovered from a small area of the Mid Atlantic Ridge. The solid double headed arrows show the evolution of magma composition that results from the withdrawal of particular combinations of minerals, and the dashed arrows point towards the rocks that form by accumulation of the fractionated minerals. The first mineral to crystallize from the parental liquid is olivine (pale green). Olivine is high in MgO and low in Al_2O_3, and its removal causes the liquid to evolve towards lower MgO and higher Al_2O_3. A rock consisting entirely of olivine is called dunite. The second mineral to join olivine in the fractionating assemblage is plagioclase (white), which is high in Al_2O_3; removal of olivine and plagioclase causes the rock to decrease in MgO and Al_2O_3. A rock consisting of olivine and plagioclase is called a troctolite. Finally, a composition of clinopyroxene, called augite (black), joins the fractionating assemblage, driving the liquid along a shallower trend of decreasing MgO and Al_2O_3. A rock containing olivine, plagioclase, and augite is a gabbro. Eruptions of liquid occur at different stages in this process, whereupon rapid cooling by contact with seawater forms basaltic glass (shiny black) that can be collected and measured.

crystallization. This is the process whereby, on cooling, crystals grow from a liquid and become chemically and physically separated from the evolving liquid. Accumulating the resulting crystals results in plutonic igneous rocks called cumulates, such as the gabbro of the lower oceanic crust or the layered basic intrusions. Sampling the remaining liquid by eruption at various stages of fractionation results in a series or suite of liquids with compositions that can be related by subtraction of the crystal phases that are being fractionated. An example is shown in Figure 3; the compositions of a suite of basalt glasses dredged from a small region of the Mid-Atlantic Ridge show evidence of removal first of olivine, followed by olivine and plagioclase, and finally by olivine, plagioclase, and augite. The lower crust and upper mantle beneath mid-ocean ridges contain the rocks that result from removal of these combinations of minerals: dunite, troctolite, and gabbro, respectively. Thus, the fractional crystallization process in this case explains the mineralogy both of plutonic rocks formed from the fractionated crystals and of erupted volcanic rocks formed from the complementary evolving liquids. From the same primary liquid, a diversity of evolved rocks can result, whether by continued evolution along a particular liquid line of descent or along different liquid lines of descent determined by the conditions of fractionation. The important variables determining which fractionation trend a given liquid follows include the pressure, water activity, and oxidation state in the magma chamber.

Assimilation and Magma Mixing

Some suites of igneous rocks display compositional relationships, textures, or other field relations showing that they did not evolve by simple removal of fractionating crystals from a single primary liquid. Many other processes can affect the evolution of igneous rocks in the lithosphere. These include assimilation (the melting and dissolution of surrounding rocks with lower melting points) and magma mixing (the combination of various batches of primary liquid that may have been created by melting of different sources) to different extents and/or at different conditions. Generally, magmas do not arrive in the crust with very much superheat; i.e., they are on their liquidus and will begin crystallizing if energy is taken out of the magma. On the other hand, assimilation of solid wallrocks requires transfer of energy from the magma to the assimilant in order to melt it. Hence, assimilation and fractional crystallization are generally coupled, occurring together and with the balance between mass of wallrock melted and mass of magma crystallized determined by the energy balance of the overall system. Assimilation can readily change the chemical, isotopic, and volatile composition of an evolving magma, because the wallrocks encountered are likely to be much older, more chemically more evolved, and more hydrothermally altered than the primary intrusion is. Evidence that processes such as assimilation and magma mixing have occurred may be drawn from chemical analysis of igneous rocks, but may also be seen directly through the occurrence of incompletely melted enclaves or stoped blocks within a magmatic body.

Glossary

adiabatic gradient The pressure–temperature path followed by rocks moving upwards or downwards without exchanging heat with their environment; given in K GPa $^{-1}$.

assimilation The melting and dissolution into a magma of surrounding rocks, encountered during melt transport or differentiation.

asthenosphere That part of Earth's mantle where temperatures are high enough and the rocks weak enough to allow heat transport by convection.

diapir A buoyant upwelling, especially a discrete mass or episodic upwelling, rather than a continuous flow.

dyke A (near-vertical) crack filled with crystallized magma; a former conduit for the transport of magma through the lithosphere.

fractional crystallization The formation of crystals from a magmatic liquid under conditions in which the crystals are physically or chemically isolated from further interaction with the remaining liquid.

igneous rock A rock formed by cooling from the molten state.

lava Molten or partially molten rock as it erupts or flows on the surface of a planetary body.

liquidus, liquidus curve The maximum temperature (as a function of pressure) at which a rock of a particular composition will be completely molten; also the temperature at which a magma or lava of a given composition will begin crystallizing.

lithosphere The outer shell of Earth, where heat transport is dominated by conduction.

magma Molten or partially molten rock in the interior of a planetary body.

porous flow The motion of a fluid driven by a pressure gradient through interstitial space in a solid, such as grain boundaries or pore space.

potential temperature The temperature that a parcel of asthenosphere would reach if expanded to 1 atmosphere pressure without melting or exchanging heat with its environment; given in °C.

solidus, solidus curve The minimum temperature (as a function of pressure) at which a rock of a

particular composition will begin to melt; also the temperature at which a magma or lava of given composition will be completely crystallized.

See Also

Earth: Mantle. **Igneous Rocks:** Granite; Komatiite. **Large Igneous Provinces**. **Mantle Plumes and Hot Spots**. **Mining Geology:** Magmatic Ores. **Pyroclastics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Asimow PD, Hirschmann MM, and Stolper EM (1997) An analysis of variations in isentropic melt productivity. *Philosophical Transactions of the Royal Society of London, Series A* 355: 255 281.

Basaltic Volcanism Study Project (1981) *Basaltic Volcanism on the Terrestrial Planets*. New York: Pergamon.

DePaolo DJ (1981) Trace element and isotopic effects of combined wallrock assimilation and fractional crystal lization. *Earth and Planetary Science Letters* 53(2): 189 202.

Grove TL, Kinzler RJ, and Bryan WB (1992) Fractionation of mid ocean ridge basalt (MORB). In: Morgan JP, Blackman DK, and Sinton JM (eds.) *Mantle Flow and Melt Generation at Mid Ocean Ridges*, pp. 281 310. American Geophysical Monograph 71. Washington DC: American Geophysical Union.

Hirschmann MM (2000) Mantle solidus: experimental constraints and the effects of peridotite composition. *Geochemistry Geophysics Geosystems* 1: 70.

Kelemen PB, Hirth G, Shimizu N, Spiegelman M, and Dick HJB (1997) A review of melt migration processes in the adiabatically upwelling mantle beneath oceanic spread ing ridges. *Philosophical Transactions of the Royal Soci ety of London, Series A* 355: 283 318.

Langmuir CH, Klein EM, and Plank T (1992) Petrological systematics of mid ocean ridge basalts: constraints on melt generation beneath ocean ridges. In: Morgan JP, Blackman DK, and Sinton JM (eds.) *Mantle Flow and Melt Generation at Mid Ocean Ridges*, pp. 183 280. American Geophysical Monograph 71. Washington DC: American Geophysical Union.

Sigurdsson H, Houghton B, McNutt SR, Rymer H, and Stix J (eds.) (2000) *Encyclopedia of Volcanoes*. San Diego, CA: Academic Press.

Wickham SM (1987) The segregation and emplacement of granitic magmas. *Journal of the Geological Society* 144: 281 297.

Yoder HS (1979) The evolution of the igneous rocks: fif tieth anniversary perspectives. Princeton, NJ: Princeton University Press.

IGNEOUS ROCKS

Contents

Carbonatites

K Bell, Carleton University, Ottawa, ON, Canada

Introduction

Few other igneous rocks are as intriguing and as fascinating as carbonatites. Made up of at least 50% carbonate minerals, carbonatites have distinctive trace element chemistries, many contain unusual, accessory minerals and some are even associated with economic mineral deposits. Because carbonated melts can be produced by very low degrees of partial melting (<0.01%) of a volatile-rich source, they may be sensitive indicators of thermal instabilities in the mantle (*see* **Earth:** Mantle) and may mark asthenospheric upwellings, deep mantle plumes, and crustal delamination.

The main physical property that separates carbonatitic melts from most others is their viscosity. Because these melts are ionic liquids, there is little or no polymerization and hence these magmas have very low viscosities. Carbonatitic melts can therefore sample

large mantle volumes, and infiltration experiments demonstrate that they can percolate rapidly in polycrystalline olivine by chemical exchange between melt and matrix. Estimates of ascent rates based on fluid-flow calculations suggest that carbonatitic melts can migrate to the surface at speeds of 20 to 65 ms ^{1}.

Once considered rare, carbonatites are now fairly commonplace, with more than 500 occurrences reported to date. Although carbonatites are found on all continents, including Antarctica, almost half are found in Africa where they are intimately associated with the East African Rift Valley system. Brazil, Canada, and north-western Russia are other regions where carbonatites are abundant. So far only two oceanic island occurrences have been found: the Cape Verde and the Canary Islands. About 50% of carbonatites are associated with extensional environments, such as rift valley systems, while others are associated with major faults and large-scale domal swells. Carbonatites are rarely, if ever, associated with subduction-related environments. In constructing any model for the origin of carbonatites, the restriction of most carbonatites to continental areas suggest that thickened lithosphere plays an important role in the production of CO_2-rich melts. Figure 1 shows the general distribution of carbonatites on a worldwide basis. Also shown for reference, are areas underlain by Archaean cratons.

Even though carbonatites are volumetrically insignificant, they provide insights into terrestrial differentiation unrivalled by few other igneous rocks. First of all, they provide information about the geochemistry of the Earth's mantle, secondly they can be used to monitor the evolution of the mantle during the last 3 Ga of Earth history, and lastly they tell us something about the migration of low viscosity melts at both mantle and crustal levels.

Field Relationships

Just like any other igneous rock, carbonatites can take the form of plutons, dykes, sills (although somewhat rare), cone sheets, lava flows, and pyroclastic deposits. Table 1 gives localities for some of these

Table 1 Forms that carbonatites can take

Form	*Example*
Lava flows	Oldoinyo Lengai, Tanzania
Lava flows	Fort Portal, Uganda
Lava flows	Kontozero, Russia
Tuffs	Kerimasi, Tanzania
Tuffs	Kaiserstuhl, Germany
Tuffs	Cape Verde Islands
Linear dykes	Gross Brukkaros, Namibia
Linear dykes	Alno, Sweden
Linear dykes	Turiy, Russia
Sills	Kaluwe, Zambia
Cone sheets and ring dykes	Oka, Canada
Cone sheets and ring dykes	Dicker Willem, Namibia
Plutons	Jacupiranga, Brazil
Plutons	Sokli, Finland
Plutons	Firesand, Canada

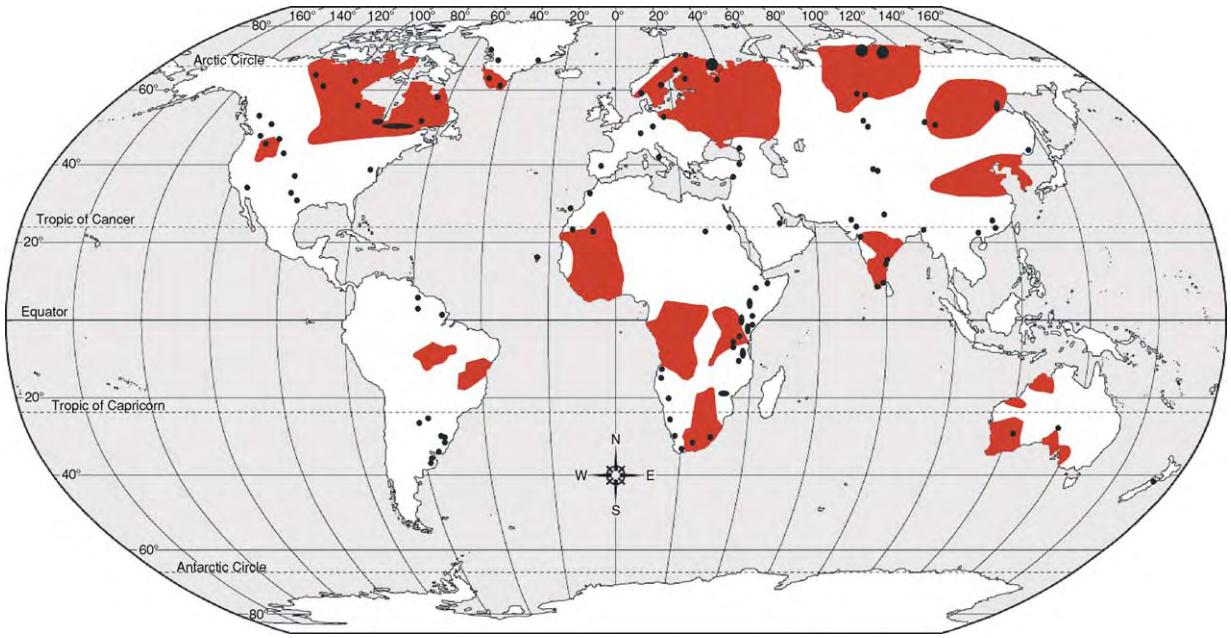


Figure 1 Distribution of carbonatites and presently exposed Archaean cratons. Note the widespread distribution of carbonatites, and that only two carbonatite occurrences are found in oceanic islands. © Hougthon Mifflin Company.

different forms. Although it is difficult to make generalizations about carbonatitic complexes, they usually consist of annular or cylindrical bodies, many less than 5 km across, of silicate rocks and/or carbonatite material cut by younger carbonatite. Invariably, the carbonatite is much younger than the associated silicate rocks, and commonly much less in volume. Many carbonatites occur in parallel dyke swarms, and many are associated with small diatreme breccias and sub-volcanic pipes and plugs. Figure 2A and B show idealised sections through a typical carbonatite complex.

When first proposed, the magmatic origin for carbonatites was vigorously challenged, but the evidence is now so overwhelming that even the most critical of sceptics accept carbonatites as true igneous

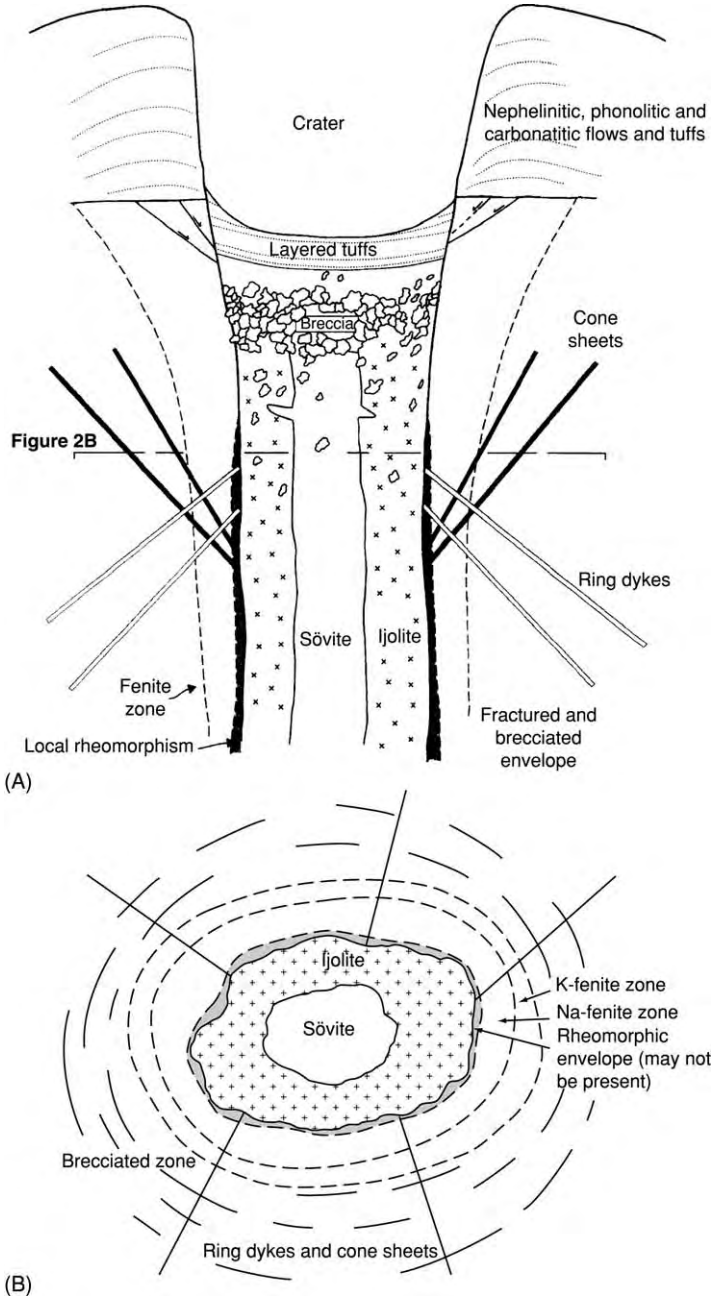


Figure 2 Idealized sections through a carbonatite complex. The early silicate rocks, such as ijolite and urtite, are cut by carbonatite. Sovite is a carbonatite rock type. Ferite is a contact metasomatic rock type. Dykes, cone sheets, and ring dykes, as well as veins, are among the last events to occur. In many complexes, the youngest events are represented by veins of dolomite carbonatites, ferrocarbonatites, and REE rich carbonatites. Adapted from Sage and Watkinson (1991).

Table 2 Rock associations

	Age	*Rock type*
Carbonatite melilitolite		
Oka complex, Quebec, Canada	120 Ma	Melilitolite (*okaite*), pyroxenite, ijolite, urtite, alnoite
Turiy complex, Russia	380 Ma	Melilitolite, melteigite, ijolite, urtite, phoscorite
Carbonatite phonolite nephelinite		
Oldoinyo Lengai	Active	Ijolite, nephelinite, phonolite
Cape Verdes	<12 Ma	Alkali basalt, basanite, nephelinite, phonolite, tephrite, syenite
Napak, Uganda	<40 Ma	Ijolite, nephelinite, phonolite
Carbonatite pyroxenite		
Vuoriyavi, Russia	380 Ma	Pyroxenite, ijolite, melteigite, syenite
Phalaborwa, South Africa	ca 2050 Ma	Pyroxenite, syenite
Carbonatite syenite		
Salitre, Brazil	80 Ma	Syenite, pyroxenite
Khibina, Russia	380 Ma	Syenite, ijolite, jacupirangite, foyaite, trachyte
Carbonatite lamprophyre		
Kandalaksha dykes, Russia	380 Ma	Ultramafic lamprophyres
Gardar, Greenland	ca 1200 Ma	Ultramafic lamprophyres
Carbonatite kamafugite		
Western Uganda	Holocene	Kalsilitite, melilitite, leucitite
Italy	Holocene	Melilitite

rocks. Evidence includes the presence of chilled margins, Peleé's tears, vesicles, lava flows, dykes, veins, and flow textures. Although most carbonatites occur as plutonic rocks, there are well over forty localities with extrusive carbonatites, at nine of which there are lavas. It should also be noted that carbonatites can take the form of hydrothermal and replacement bodies.

Carbonatites rarely occur on their own. Most are spatially associated with a variety of silicate rocks of similar age. Of the classic carbonatite complexes scattered throughout East Africa, most are associated with nephelinites and phonolites, or their plutonic equivalents. Silicate rocks, when present, form a much greater volume than the carbonatites and are usually characterized by great structural and textural complexity. The close proximity between silicate rocks and carbonatites within the same complex means that the origin of one cannot be divorced from the origin of the other. Table 2 groups possible carbonatite-silicate rock associations.

Ages

Carbonatitic magmatism extends back to at least 3 Ga, although carbonatites of Archaean age are comparatively rare. The oldest known carbonatite to date occurs in Greenland (Tupertalik, 3.0 Ga). Other Archaean carbonatites include Silliinjarvi in Finland, and Dolodau and Lac Shortt in Canada. The youngest carbonatite is represented by the lava flows from Oldoinyo Lengai in Tanzania, which is the only known active carbonatite volcano on Earth strategically situated in the floor of the East African Rift Valley system. Although carbonatitic magmatism on a worldwide scale has been continuous throughout geological time, distinct groupings occur at about 2.8, 2.1, 1.8, 1.0, .60, and .35 Ga, ages that roughly correspond to periods of major orogenic activity. As a generalization, however, most carbonatites occur in extensional, rather than compressional, environments.

In terms of age, carbonatites appear to become more abundant towards more recent times. This might simply reflect the difficulties encountered in preserving carbonatites or alternatively it might simply reflect the much higher geothermal gradient that was in existence during the Archaean.

Mineralogy

Calcite and dolomite are the major rock-forming minerals in carbonatites. Other carbonates include the Fe-rich varieties and some of the much rarer REE (rare-earth element) carbonates. Table 3 lists some of the minerals commonly found in carbonatites. The division of most carbonatites into three groups is based on whether calcite, dolomite, or Fe-rich carbonates (*see* **Minerals:** Carbonates) are the predominant minerals. The plutonic variety of calcite carbonatite is sometimes called sövite and when medium to fine grained is referred to as an alvikite. Both of these names have been carried over from the very early days of carbonatite research. One unusual carbonatite, found only at Oldoinyo Lengai volcano in Tanzania, consists of sodium, potassium,

Table 3 Some minerals found in carbonatites

Carbonates	*Phosphates*
Calcite	Apatite
Dolomite	Monazite
Ankerite	*Halides*
Siderite	Fluorite
Bastnasite $(Ce,La)FCO_3$	Halite
Burbankite	Sylvite
Parisite	*Oxides*
Nyerereite $(Na,K)_2Ca(CO3)_2$	Baddeleyite
Gregoryite $(Na,K)_2CO_3$	Hematite
Silicates	Ilmenite
Amphibole	Magnetite
Aegerine augite	Perovskite
Diopside	Pyrochlore
Olivine	
Phlogopite	
Sulphides	
Pyrite	
Pyrrhotite	
Galena	
Sphalerite	

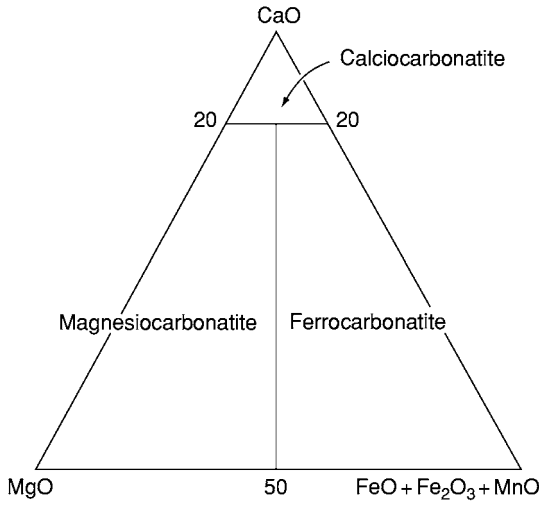


Figure 3 Classification scheme for carbonatites suggested by the IUGS based on chemical composition. Oxides in wt percent.

and calcium carbonates, a variety known as natrocarbonatite. Other minerals that are commonly found in carbonatites include apatite, amphibole, magnetite, mica, olivine, and pyroxene.

As an alternative to using a classification scheme based on mineralogy, carbonatites can also be classified on the basis of their chemical compositions (*see* **Rocks and Their Classification**). Figure 3 shows one such classification scheme that divides carbonatites into calciocarbonatites, magnesiocarbonatites, and ferrocarbonatites. The term silicocarbonatite is sometimes used for those rare carbonatites that contain greater than 20 wt% SiO_2.

Economic Deposits

Carbonatites are a major source of Nb, phosphates (*see* **Sedimentary Rocks:** Phosphates), F, and the REEs. Minerals occur in magmatic, hydrothermal, and supergene deposits. A list of some of the more important deposits is given in Table 4. Carbonatites still constitute the world's greatest source of Nb. Pyrochlore, the most important Nb-bearing mineral, is principally mined in Brazil (e.g., Araxá, Tapira) and some is mined in Canada (St Honoré). However, not all carbonatites contain pyrochlore.

Because carbonate minerals are easily dissolved during chemical weathering, the more resistant minerals, such as pyrochlore and monazite, can be concentrated and easily mined as residual material. Rare earth elements in carbonatites are common and can be found as magmatic minerals or in minerals directly precipitated from hydrothermal solutions. Among such minerals are included monazite, bastnaesite, britholite, burbankite, parisite, and synchesite.

Although most of the world's phosphate production comes from marine phosphorites, the presence of primary apatite in all carbonatites makes them a potential source for fertilisers. More money is generated from this commodity than any other

Table 4 Some examples of mineral deposits associated with carbonatites

Complex	*Commodity*	*Details*
Araxá, Brazil	Nb	Pyrochlore in laterite
St Honoré, Canada	Nb	Pyrochlore in carbonatite
Jacupiranga, Brazil	Phosphate	Primary apatite in carbonatite
Sukulu, Uganda	Phosphate	Residual apatite in weathered carbonatite
Mountain Pass, USA	REE	Bastnaesite in carbonatite
Gakara Karonge, Burundi	REE	Hydrothermal veins of bastnaesite
Tapira, Brazil	Ti	Anatase in laterite
Powderhorn, USA	Ti	Perovskite in pyroxenite
Phalaborwa, South Africa	Cu + Vermiculite	Vermiculite in pyroxene pegmatoid
Ipanema, Brazil	Vermiculite	Vermiculite in glimmerite
Amba Dongar, India	Fluorite	Hydrothermal overprint on fenite
Okurusu, Namibia	Fluorite	Hydrothermal overprint on fenite

extracted from carbonatites. Apatite, for example, is mined at Khibina in the Kola region of Russia (*see* **Fluid Inclusions**).

The presence of accessory fluorite in many carbonatites shows the importance of F in carbonatitic liquids. Most fluorite deposits, however, are formed by late-stage hydrothermal activity. Among the most important deposits are those at Amba Dongar, India, and Okorusu, Namibia. Fluid inclusion studies from the fluorite at Amba Dongar reveal low crystallization temperatures between 100 and 150°C.

Th and U are also associated with many carbonatite complexes, and are contained within minerals such as thorite and monazite.

Another intriguing economic aspect of carbonatites is the possibility that they may contain diamonds. Diamond-bearing carbonatites have been reported from Uzbekistan and although abundant, the diamonds are very small. This discovery supports any model that proposes that diamonds are formed from strongly compressed carbonatitic melts.

Geochemistry

Carbonatites can be easily separated from other carbonate-rich rocks, especially marbles, by their unusual and distinctive trace element contents. Representative carbonatite analyses are given in Table 5. Among the most characteristic features are high REE (>500 ppm; all are light-REE enriched), high Sr (>700 ppm), Ba (>250 ppm) and V (>20 ppm) contents. Table 6 gives some trace element ratios. Figure 4A, a diagram of Ba + Sr (ppm) versus REE + Y (ppm), clearly discriminates between carbonatites and other rock types. Enrichment in the light REE can be seen from Figure 4B. Normalized trace element data from Oldoinyo Lengai, along with values for average oceanic island basalt, are given in Figure 5.

The Nd, Pb, and Sr isotopic systematics from young carbonatites (<200 Ma) tell a fascinating story, and show that most carbonatites share many isotopic features seen in oceanic island basalts, in spite of the fact that carbonatites are mainly found on continents. Clearly shown is the mantle origin for carbonatites. The two diagrams shown in Figures 6 and 7 illustrate these isotopic similarities, as well as the isotopic heterogeneity shown by carbonatites. Isotopic data from East African carbonatites (Figure 8) provide an interesting case, where the near-linear array reflects the mixing of two end-members, both of which occur in oceanic island basalts. Even within individual centres the spread of isotopic data can be large, particularly if the silicate rocks are included. The idea of a closed system, magma chamber undergoing differentiation cannot be reconciled with the isotopic data. Melts can only be generated by a series of discrete partial melting events that take place within an isotopically heterogeneous mantle.

Because carbonatites date back to at least 3 Ga, they can be used to monitor the chemical evolution of sub-continental upper mantle over a considerable part of the Earth's history. Data from Canada, Greenland, and the Fennoscandian Shield, are shown in Figure 9. Although there is considerable scatter of the data, the low data points define a source for the carbonatites that was clearly established early in the Earth's history, and which has remained isotopically undisturbed since at least 3 Ga. This is also supported by Hf isotope data.

The stable isotopic systematics tell another story about magma differentiation, and magma sources. Unfortunately, carbonatites are particularly prone to isotopic exchange processes involving weathering, alteration under atmospheric conditions, and hydration, so care must be taken to ensure that only fresh samples are analysed.

C and O isotopic data from freshly erupted samples from the Oldoinyo Lengai natrocarbonatite overlap the mantle field based on data from oceanic basalts (Figure 10) and support the conclusion that carbonatites are ultimately derived from a mantle-derived parental melt. For the most part, carbonatites have much lighter C and O than limestones. A comparison is made in Figure 10 among the stable isotope composition of carbonatites from several different continents and limestones of both Phanerozoic and Precambrian age. A convergence towards mantle values is shown by much of the data from carbonatites from Greenland, Europe, North America, and South America.

Few S isotopic data exist for carbonatites. The range of $\delta^{34}S$ values from sulphides from carbonatites are quite variable and show marked differences to mantle values. A single sulphide sample from the Oldoinyo Lengai natrocarbonatite has a δ value of +2.8‰, thus overlapping with the range of sulphide values found in high temperature carbonatites. Sulphides from several Proterozoic calciocarbonatites from Canada, as well as carbonatites from Russia, indicate a large range of values (−7 to +3‰) which coincide with most of the values obtained from other carbonatites. There is some indication that each complex may have its own distinctive average S isotopic composition. In view of these observations, some of the differences in S isotope compositions between carbonatite complexes have been attributed to isotope heterogeneity within the mantle. In contrast to the sulphides, the sulphates are invariably higher in their S isotope compositions (range +4 to +14) reflecting the oxidation state of the S.

Table 5 Chemical compositions of carbonatites

wt%	*Cacite carbonatite*	*Dolomite carbonatite*	*Ferrocarbonatite*	*Natrocarbonatite*
SiO_2	2.72	3.63	4.7	0.16
TiO_2	0.15	0.33	0.42	0.02
Al_2O_3	1.06	0.99	1.46	0.01
Fe_2O_3	2.25	2.41	7.44	0.05
FeO	1.01	3.93	5.28	0.23
MnO	0.52	0.96	1.65	0.38
MgO	1.80	15.06	6.05	0.38
CaO	49.1	30.1	32.8	14.0
Na_2O	0.29	0.29	0.39	32.2
K_2O	0.26	0.28	0.39	8.38
P_2O_5	2.10	1.90	1.97	0.85
H_2O^+	0.76	1.20	1.25	0.56
CO_2	36.6	36.8	30.7	31.6
BaO	0.34	0.64	3.25	1.66
SrO	0.86	0.69	0.88	1.42
F	0.29	0.31	0.45	2.50
Cl	0.08	0.07	0.02	3.40
S	0.41	0.35	0.96	nd
SO_3	0.88	1.08	4.14	3.72
ppm				
Li	0.1	nd	10	270
Be	2	<5	12	nd
Sc	7	14	10	nd
V	80	89	191	116
Cr	13	55	62	<3.0
Co	11	17	26	1.8
Ni	18	33	26	1
Cu	24	27	16	nd
Zn	188	251	606	88
Ga	<5	5	12	<20
Rb	14	31	nd	178
Y	119	61	204	7
Zr	189	165	127	2
Nb	1204	569	1292	28
Mo	nd	12	71	125
Ag	nd	3	3	bd
Cs	20	1	1	6
Hf	nd	3	nd	<0.4
Ta	5	21	1	0.3
W	nd	10	20	49
Au	nd	nd	12	18
Pb	56	89	217	22
Th	52	93	276	4
U	9	13	7	11
La	608	764	2666	545
Ce	1687	2183	5125	645
Pr	219	560	550	19
Nd	883	634	1618	102
Sm	130	45	128	8
Eu	39	12	34	2
Gd	105	nd	130	2
Tb	9	5	16	0.1
Dy	34	nd	52	2
Ho	6	nd	6	0.1
Er	4	nd	17	0.3
Tm	1	nd	2	0.3
Yb	5	10	16	bd
Lu	1	0.1	nd	0.01

bd below detection; nd not determined.

Data from Woolley and Kempe (1989), Dawson *et al.* (1995), Keller and Spettel (1995) and Simonetti *et al.* (1997).

Table 6 Trace element ratios

	Calcite carbonatite	*Dolomite carbonatite*	*Ferrocarbonatite*	*Natrocarbonatite*[a]
Ba/Sr	0.41	0.97	3.85	1.44
Ba/La	4.93	7.46	10.8	50.3
Th/U	5.8	7.2	39	1.0
Ce/Pb	30	25	24	12.6
Zr/Nb	0.16	0.29	0.1	0.1
Ba/Nb	2.49	10	22.4	65.6
La/Nb	0.50	1.3	2.1	12
Th/La	0.10	0.75	9.7	0.1
Nb/U	133	44	185	14.4

[a]Data from natrocarbonatites of the June, 1993 eruption of Oldoinyo Lengai with <0.3% silica. From Simonetti *et al.* (1997).

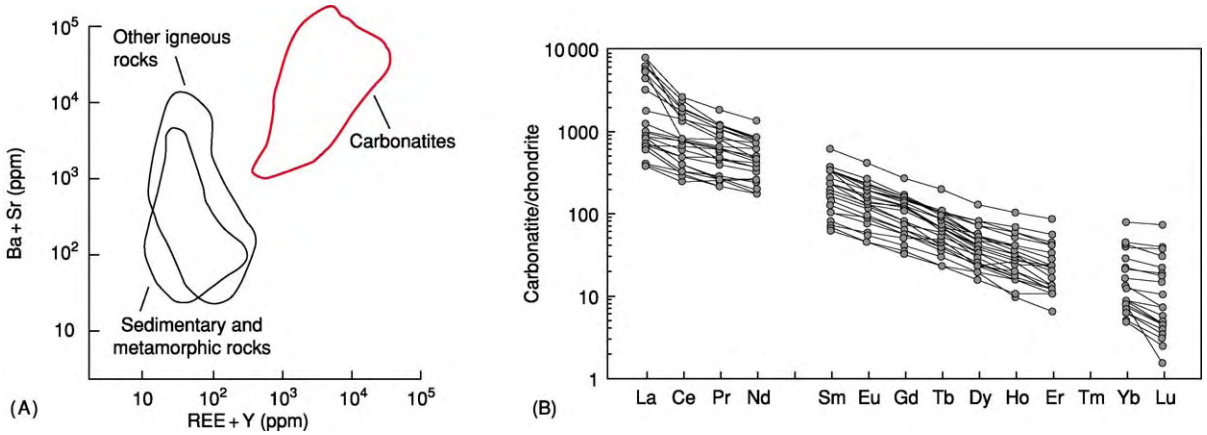


Figure 4 (A) Ba + Sr (ppm) vs REE + Y (ppm). Note the separation of the carbonatites from all other rocks types. From Samoilov (1991). (B) Chondrite normalised plot for some whole rock sovite from Europe and North America. Note the similarity in patterns and especially the enrichment in the LREEs.

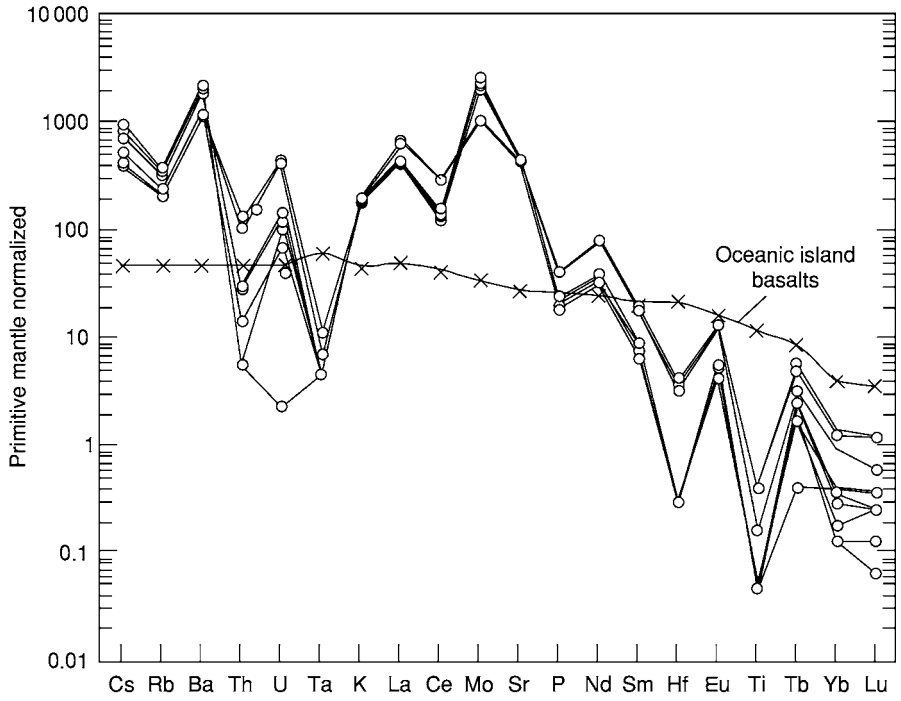


Figure 5 Normalized trace element diagram of the June 1999 lavas from Oldoinyo Lengai, normalized to primitive mantle. Note the peaks for Ba and Sr, and the depletions for Ta, Hf, and Ti. Also an average value is given for oceanic island basalts. From Simonetti *et al.* (1997).

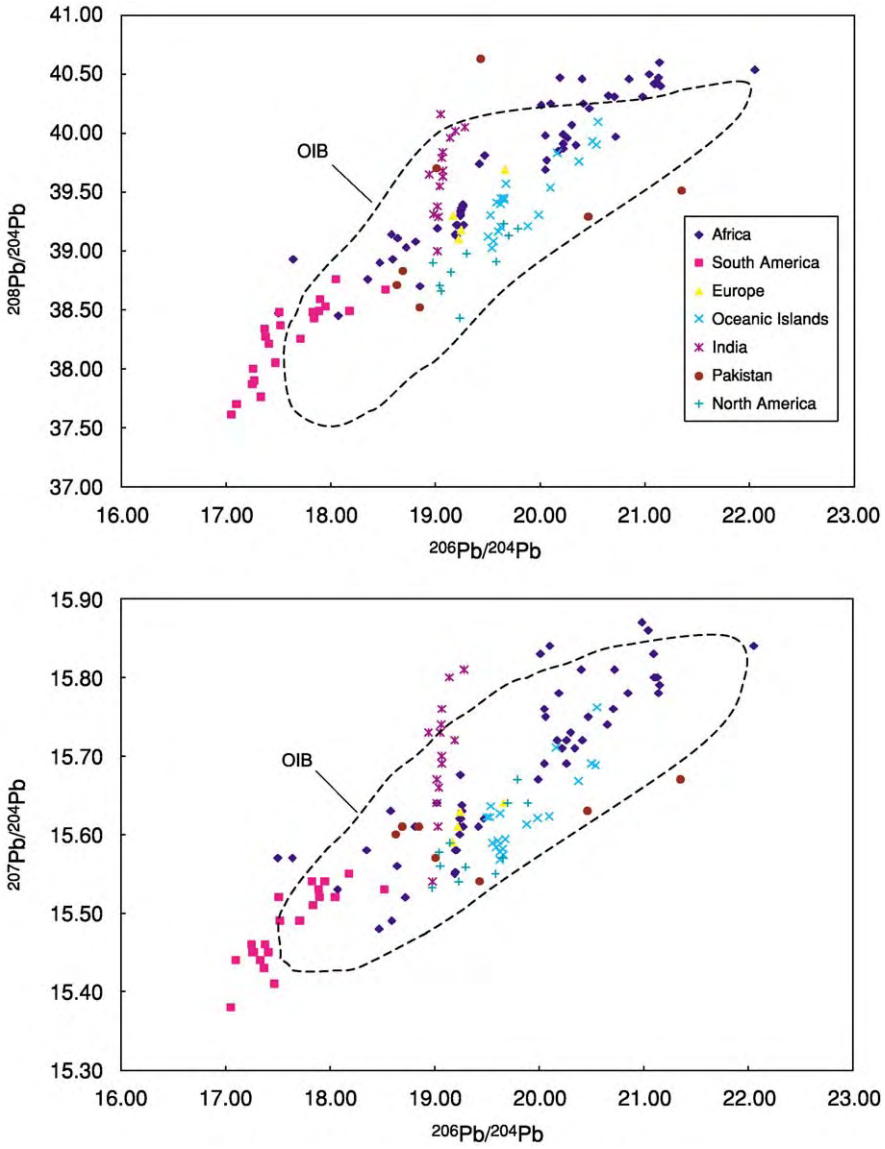


Figure Pb Pb isotope ratio diagrams of young (<200 Ma) carbonatites. OIB field for oceanic island basalts. Because these samples are young, the ratios plotted are assumed to be those that existed in the magmas at the time of their formation. Note the similarity of many of the carbonatite data points to those from oceanic island basalts.

Chlorine isotopes have also been measured in carbonatites. Unaltered carbonatites from Africa and Germany have δ^{37}Cl of between −0.8 and +0.1‰. Because the whole mantle δ^{37}Cl is approximately +4.7, the anomalously low chlorine values in carbonatites suggests anomalous mantle involving a crustal component, or processes that fractionate the isotopes, perhaps during magma degassing.

Oldoinyo Lengai

Because Oldoinyo Lengai is the only active carbonatite volcano it deserves special mention. Rising to nearly 3000 m above Tanzania's Eastern Rift Valley, the stratovolcano has a basal diameter of about 12 km and an approximate volume of 60 km^3. During the last few hundred thousand years, the volcano has erupted a series of carbonatitic, phonolitic, and nephelinitic lavas as well as pyroclastic material. Major ash eruptions documented during historic time occurred in 1917, 1926, 1940, and 1966–1967.

The natrocarbonatite flows from Oldoinyo Lengai (Figures 11 and 12) are unusual, both in terms of their mineralogy and their chemical composition. They are made up mainly of two alkali-rich minerals, nyrereite and gregoryite, both of which are

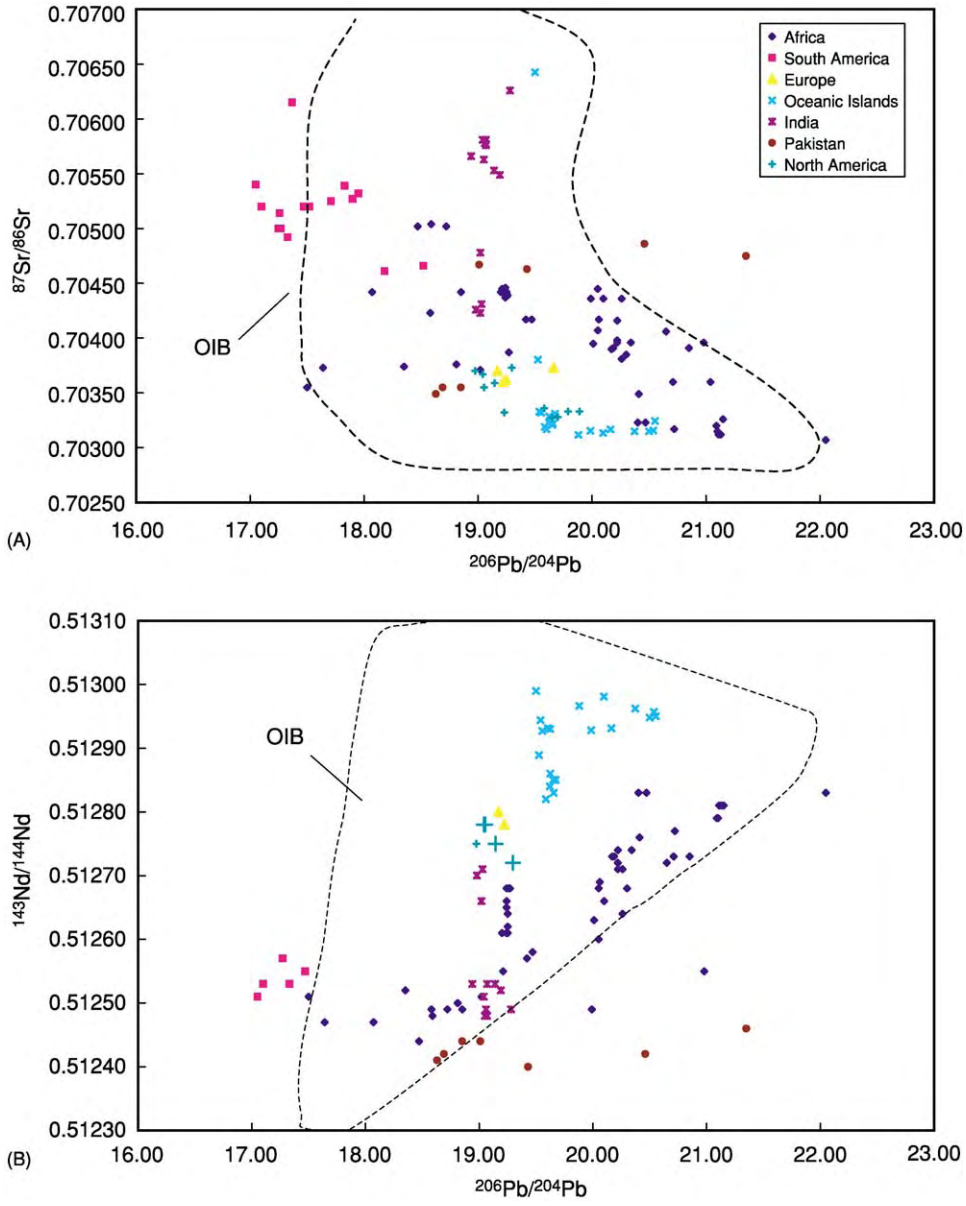


Figure 7 (A) $^{87}Sr/^{86}Sr$ vs $^{206}Pb/^{204}Pb$ for young carbonatites. (B) $^{143}Nd/^{144}Nd$ vs $^{206}Pb/^{204}Pb$ for young carbonatites (<200 Ma). Note the similarities of the carbonatite data with those from oceanic island basalts. A handful of data points for carbonatites with values >0.710 have been omitted.

alkali carbonates with Na well in excess of K. These two carbonates form phenocryst phases, and are present in the groundmass along with fluorite, sylvite, and other minerals. The lavas are black on eruption but because of their hygroscopic nature quickly turn white as the lava reacts with water from the atmosphere.

Natrocarbonatites also have the lowest measured eruption temperatures of any terrestrial magma (490 to 590°C), and they also have a viscosity, thermal diffusivity, specific heat capacity, and latent heat of fusion considerably lower than basaltic lavas. Important features of the trace element geochemistry of natrocarbonatites are enrichment in Ba, Ba/Sr >1, low Zr (<5 ppm), Th/U ~1, LREE normalized values >1000 × chondrites, and $(La/Sm)_N$ >40.

The existence at Oldoinyo Lengai of highly-alkaline carbonatitic liquids (Na-rich) and the spatial association of sodic-rich silicate rocks led to a model involving liquid immiscibility in which a carbonated alkali-rich, silicate parent magma exsolved to form two immiscible liquids. Laboratory studies have shown that immiscibility can indeed take place with liquids of appropriate compositions.

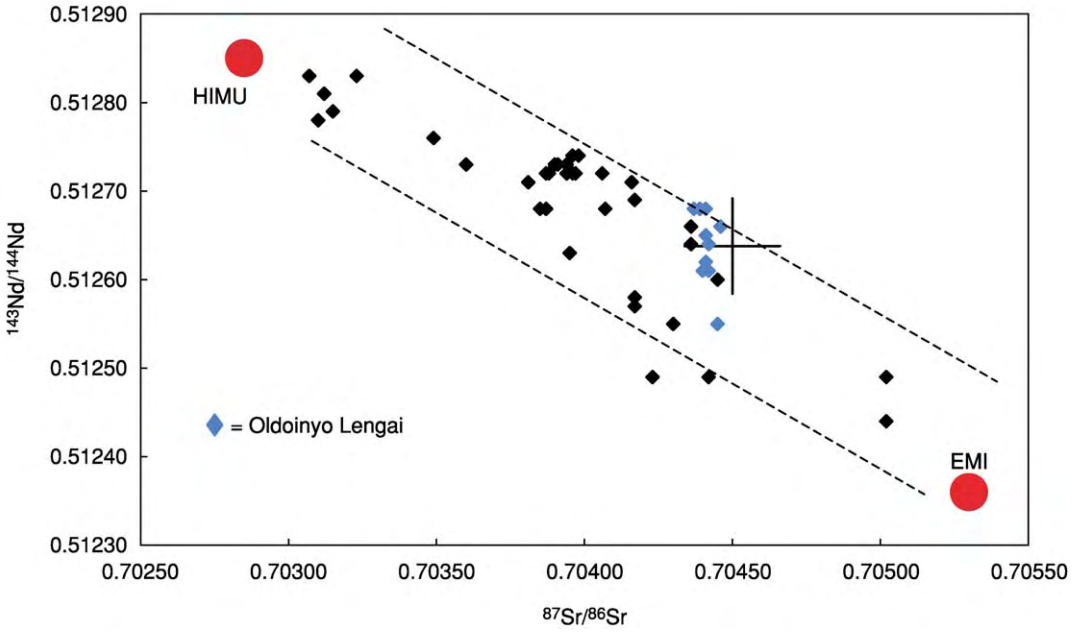


Figure 8 Nd and Sr isotopic data from young East African carbonatites. Note the near linear array, that suggests mixing between two distinct mantle sources, shown by the red dots, similar to those found in oceanic island basalts (HIMU and EMI). The data from Oldoinyo Lengai cluster close to the cross that marks the position of values for the undifferentiated silicate part of the Earth. Although the natrocarbonatites from Oldoinyo Lengai are quite different in chemical composition to other carbonatites, their isotopic data follow the same trend as other carbonatites from East Africa.

The absence of the high alkali contents in all other types of carbonatite raises doubts about the seminal role that Oldoinyo Lengai plays in understanding carbonatite genesis. It was once thought that all carbonatitic melts were alkali rich, and that alkalies were lost during fluid migration from the melt during fenitisation. Fascinating as the findings from Oldoinyo Lengai may be, in an introduction to a collection of papers dedicated solely to Oldoinyo Lengai, the consensus was that "natrocarbonatites should be considered the result of an extreme process rarely encountered in nature".

Mode of Origin

There has been a great deal of debate about the origin of carbonatites, and this is still a matter of ongoing concern. On the basis of evidence from phase equilibrium studies, field relationships, and geochemistry, there are three possible ways of producing carbonatitic melts. Firstly, they can be direct partial melts of a mantle containing carbonate phases. Secondly, they can be produced by liquid immiscibility from a carbonated alkali-rich melt. And finally, they can be generated by extreme crystal fractionation of a silicate magma. Studies involving geochemistry, mineralogy, petrography, and field relationships show that even within one igneous province, such as the Kola Alkaline Province, or even within an individual complex, it is likely that carbonatites are produced in more than one way.

As to how individual carbonatites form is a difficult question to answer, since we still lack good criteria to separate one mechanism from another. If carbonatitic melts are primary liquids, they are probably dolomitic in composition, have a Ca/Ca + Mg ratio of between 0.5 and 0.7 and a silica content of 4–5 wt%. However, many dolomite carbonatites can be found as late stage veins, dykes, and stringers cutting older calcite carbonatites.

The Source(s)

One of the more important questions concerning carbonatites is the site of generation of their parental liquids. At first it was thought that the carbonated melts were generated by melting metasomatized lithosphere, but this is not supported by the radiogenic isotopic data, which would require the involvement of more than one mantle source similar to those reflected in oceanic island basalts (Figure 8).

The mantle source for carbonatites must contain carbonate phases. The stability fields of Mg-rich calcite, dolomite ($CaMg(CO_3)_2$), and magnesite

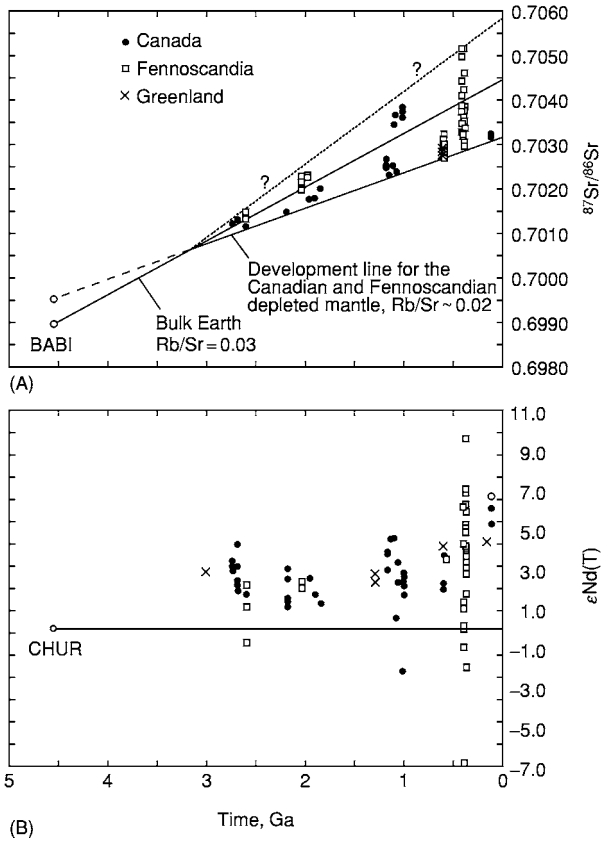


Figure 9 Development lines for Sr and Nd. (A) $^{87}Sr/^{86}Sr$ ratio versus time is shown for Canadian and Fennoscandian carbonatites. The development line indicates an ancient mantle reservoir at least 3.0 Ga old. The Earth is considered to have started with a $^{87}Sr/^{86}Sr$ ratio similar to that found in basaltic achondritic meteorites. The initial value is shown by the point marked BABI. (B) εNd(T) versus time where εNd(T) $(^{143}Nd/^{144}Nd_{sample}$ $^{143}Nd/^{144}Nd_{CHUR})/^{143}Nd/^{144}Nd_{CHUR}$ where $^{143}Nd/^{144}Nd_{sample}$ is the initial ratio in the sample and $^{143}Nd/^{144}Nd_{CHUR}$ is the initial ratio of CHUR at the same point in time. Any of the data points above the horizontal line reflects an ancient reservoir that has undergone depletion relative to achondritic meterorites. Note that the depletion event must have occurred before 3.0 Ga. From Bell and Tilton (2002).

($MgCO_3$) make these minerals ideal candidates for trapping CO_2 in the mantle. Whether the carbonate is primary, in other words present in the mantle since the Earth formed, or whether it is recycled (*see* **Carbon Cycle**) during subduction into the mantle in the form of limestone or carbonated sea floor material, is an interesting question. From the results of recent experiments, however, it now seems that carbonates from a carbonated eclogite (the metamorphosed equivalent of basalt plus calcite) will probably be removed from the downgoing slab before reaching a depth of 300 km. Thus, the possibility of carbonates being recycled into the transition zone and into the deep mantle seems unlikely.

Figure 13 shows an example of a volatile-rich mantle containing H_2O and CO_2. At depths of about 75 km the solidus forms a shoulder or ledge under conditions of near constant pressure, which coincides with solubility of the $MgCO_3$ component in the magma when dolomite becomes stable under mantle conditions. Although there is still some question about the exact depth of this ledge and its shape, there is general agreement that it does exist and that it plays a key role in suppressing volatile-rich melts as they attempt to reach the surface. At pressures below this ledge, very low degrees of partial melting will generate a dolomitic carbonatite magma. As the temperature of melting progressively increases, and hence the degree of partial melting, the melt compositions become progressively more silica-rich, yielding melts of lamprophyric, melilitic, and nephelinitic compositions.

As we have seen, most carbonatites have mantle isotopic signatures similar to those found in oceanic island basalts. Far from being constant in their isotopic composition, the heterogeneities found in oceanic island basalts have been generally attributed to material subducted into the deeper parts of the mantle, that have been allowed to age over geologic time. The isotopic data restrict the mantle source well below the source region of mid-ocean ridge basalts.

The task of evaluating the depths of the mantle sources on the basis of Sr, Nd, and Pb is difficult, if not impossible. Fortunately, noble gas data have come to the rescue and provide some indications about the relative depths of the source material. Noble gases (e.g., Ar, Ne, Kr, Xe) trapped in minerals, such as apatite and magnetite, point to a relatively primitive mantle, which is less degassed than the source that generated mid-ocean ridge basalts. Of particular interest are the ^{129}Xe anomalies, low $^{40}Ar/^{36}Ar$ and $^{4}He/^{3}He$ ratios, and $^{20}Ne/^{22}Ne$ vs $^{21}Ne/^{22}Ne$ regressions similar to those from Loihi, Hawaii, that contain a known plume component. Wherever the source is in the mantle, it appears that it has not had the opportunity to mix with gases from our present atmosphere, nor did it let the ^{129}Xe escape.

Plumes and Carbonatites

Attention has recently focused on the relationship between carbonatites and mantle plumes, and although perhaps not all are plume-related, there is some evidence to support this idea. Evidence for plume-related magmatism includes: (i) the similarity of isotope signatures between oceanic island basalts and carbonatites; (ii) the association of carbonatites

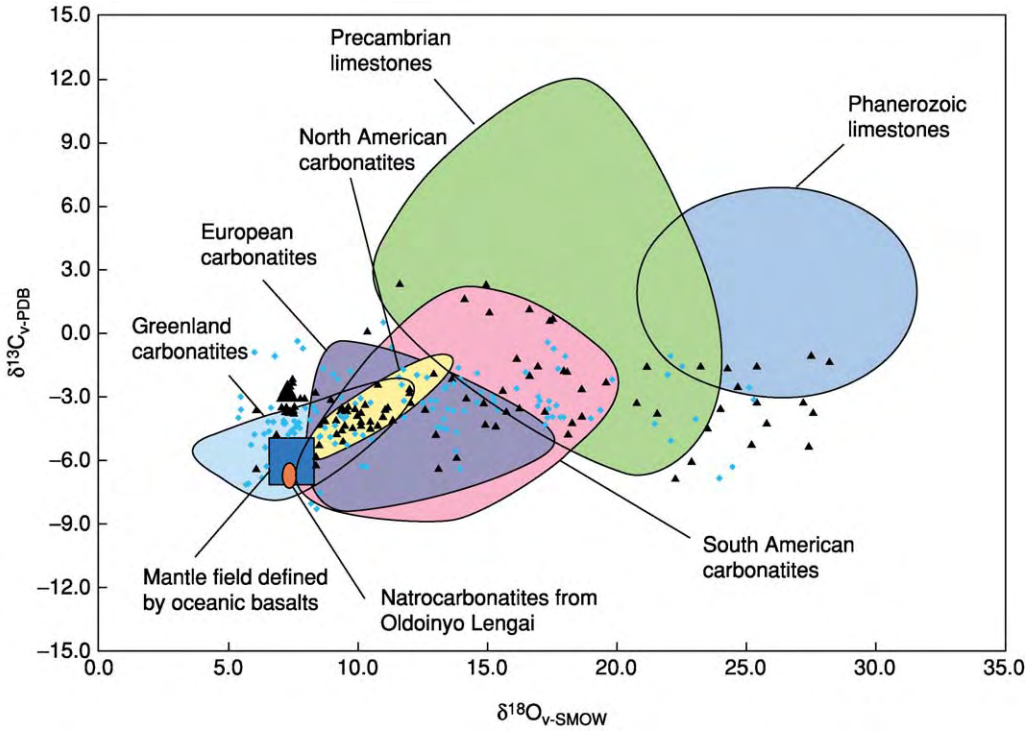


Figure 10 Stable isotope data. Comparison between limestone and carbonatite isotopic signatures. Note that the isotopic data from carbonatites end at the mantle box based on oceanic basalts. Each of the fields enclose >90% of the data points for each particular group. The data from Africa carbonatites are marked by diamonds, those from India by triangles.

with two known oceanic plumes, the Canary and the Cape Verde; (iii) the temporal and spatial relationships of some carbonatites to flood basalt provinces (for example the Deccan, Siberian, and Keeweenawan events); and (iv) the primitive natures of some of the noble gas data. Plume-related magmatism has also been proposed for the young carbonatites in East Africa, including Oldoinyo Lengai.

One issue in understanding carbonatite magmatism is the role that the continental lithosphere plays in their origin. Some plume models involve mixing between material from the plume and the overlying lithosphere, whereas others regard the sole role of the lithosphere as being one of mechanically constraining volatiles within a plume head. Plume heads that become attached to the lithosphere might even form sites for the generation of carbonated melts.

Based on models that have been proposed for plume-related magmatism, carbonatites could well be generated from the cooler, volatile-rich parts of a plume head. With the presence of the shoulder on the volatile-rich, mantle solidus at depths of about 75 km, the release of volatiles may induce crack propagation and explosive activity and promote the movement of low viscosity silicate and carbonatitic melts. A schematic representation of this is shown in Figure 14.

Metasomatism

Migration of volatiles and fluids (probably aqueous), at the margins of many carbonatites, infiltrate and metasomatise the surrounding country rocks. These rocks become highly fractured, and new minerals are developed (such as feldspar, sodic pyroxenes, and alkali amphiboles) as the volatiles and fluids chemically interact with the country rocks. The process of metasomatism at crustal levels is called fenitisation and the resulting rocks fenites, named after the Fen complex in Norway, where they were first recognized. Potassic metasomatism is considered to characterize high level, lower temperature metasomatism, whereas the sodic type is characteristic of unspecified deeper levels and higher temperatures. It has even been suggested that some syenitic rocks may represent the ultimate products of fenitisation.

The low viscosity that characterizes carbonatitic melts along with their low dihedral wetting angles make them ideal agents for metasomatism within the mantle. With their high reactivity, such melts can

Figure 11 (A) Oldoinyo Lengai, Tanzania during one of its phases of explosive activity (Photo: JB Dawson, taken August, 1966). Ash clouds were reported by aircraft at flight levels of 10 700 m. Oldoinyo Legai rises about 2886 m above the floor of the East African Rift Valley System, and was first shown on a map compiled by two missionaries in 1855. (B) Aerial view from the north west of ash cone that formed in the crater in 1966 (Photo: JB Dawson, taken August, 1966).

Figure 12 (A) Active north crater of Oldoinyo Lengai looking north from the summit (Photo: J Keller, taken June, 1989). Lake Natron can be seen in the background. The crater is about 250 m across. On far side of crater are two active vents erupting black, very fluid natrocarbonatite lava flows. Inactive parasitic vents with lava flows occur on the rim of the crater. Note the white colour of the weathered carbonatite flows. (B) Looking into the north crater from the summit (Photo: J Keller, taken October, 2003). Upper left the rift escarpment, upper right Lake Natron (2300 m below!) The crater is full, and compared with **Figure 12A** hardly any of the crater wall is visible. The active vents have formed spatter cones. Note the very fluid lava flows in the foreground.

change the mineralogy of the mantle, and melts may be changed themselves, undergoing 'chemical death' without reaching the surface. Reactions that can take place within the mantle are:

$$3\text{dolomite} + \text{diopside} = 2\text{forsterite} + 4\text{calcite} + 2CO_2$$

$$\text{dolomite} + 4\text{orthopyroxene} = \text{clinopyroxene} + \text{olivine} + CO_2$$

Incursions of carbonatitic melts can have a profound effect, not only on the mineralogy of the mantle, but also on the chemical composition of the mantle. If carbonatitic melts reach the lithosphere, minerals can precipitate to generate a veined mantle with unusual mineralogy. Melting of these veins, along with the wall rocks that surround them, can generate unusual magmas that may lead to

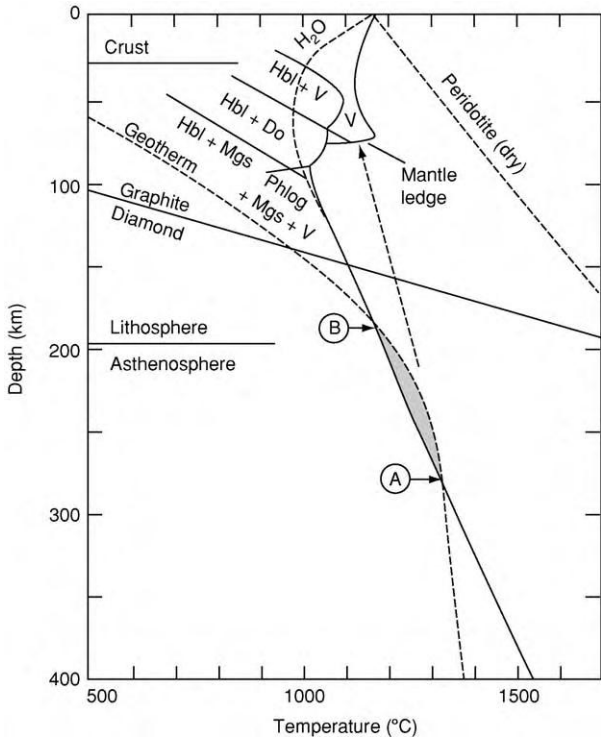


Figure 13 Generalized diagram of the system peridotite +C+H+O. Note the position of the mantle ledge at a depth of about 75 km. Melts are generated between A and B where the geotherm intersects the mantle solidus. Melts at relatively low temperatures will hit the mantle ledge producing metasomatism and possibly explosive activity. System is one in which CO_2 is well in excess of H_2O. Phlog phlogopite, Do dolomite, Hbl hornblende, Mgs magnesite and V vapour. Adapted from Wyllie *et al.*(1990).

melts of kimberlitic, lamproitic, and lamprophyric affinity (*see* **Igneous Rocks:** Kimberlite).

Phase Equilibrium Studies

Numerous experiments have been carried out in various laboratories using both natural and synthetic mixtures to help understand the melting conditions needed to generate carbonatitic melts, and to monitor the evolution of carbonated melts within a magma chamber. Using the system CO_2-H_2O-CaO-($\pm Al_2O_3$)-MgO-SiO_2 as a reasonable approximation to a volatile-rich mantle peridotite, we can place some constraints on the origin and evolution of carbonatitic melts. Such studies show that it is indeed possible to produce carbonatitic liquids by: (i) small degrees of partial melting of a carbonated peridotite, by liquid immiscibility; and (ii) by fractional crystallization of a carbonated silicate melt.

Primary carbonatitic melts can be generated at depths of >200 km and these have compositions that lie close to dolomite. As the melt migrates to the surface, at a depth of about 75 km, the magma 'freezes' and volatiles are given off that can generate explosive diatremes. If the melt reacts to produce wehrlite then the melt can become increasingly enriched in $CaCO_3$ and SiO_2 and hence siliceous carbonatite might escape to the surface.

If the parental melt is a carbonated silicate melt, such as an olivine melanephelinite, such a liquid can take various paths. One path can lead to the silicate-carbonatite liquidus field boundary and the liquid can precipitate out cumulate carbonatites. Another path at relatively low pressures can lead to a miscibility gap, and can exsolve an immiscible carbonate-rich melt. These phase relations are shown in Figure 15. Many systems indicate that a carbonatite produced in this way can only contain a maximum of 80% $CaCO_3$. With CO_2 saturation, immiscible separation can take place at much deeper levels. If the high temperature, immiscibly-derived liquid differentiates then it must trend towards enrichment in $(Na,K)_2CO_3$. Magnesiocarbonatite magmas are also capable of precipitating calciocarbonatite rocks. In addition, it should not be overlooked that many calciocarbonatites may represent crystal cumulates.

Concluding Remarks

Carbonatites must originate from a volatile-rich mantle (*see* **Earth:** Mantle), and the obvious choice of minerals containing the CO_2 necessary for their origin are calcite, dolomite, and magnesite. Thinking in much broader terms about carbonatites, one key question involves the carbon cycle (*see* **Carbon Cycle**). Does carbonatitic magmatism reflect recycling of C into the mantle by the subduction of carbonate-bearing material such as limestone or carbonated eclogite, or does it simply reflect an attempt by the Earth to purge itself of primitive volatiles? On the basis of isotopic data, especially the noble gas data, crustal recycling seems to be unlikely and we are more inclined to turn to those models that involve outgassing of primitive material.

It is difficult to imagine a primitive Earth without a CO_2-rich atmosphere. It seems reasonable to assume that some CO_2 was trapped during the Earth's accretionary history and remnants of this are still retained within the lower parts of the mantle. Our two neighbouring planets, Venus and Mars, have CO_2 rich atmospheres, and there is no reason to believe that the Earth was any different from them in its early history (*see* **Solar System:** Mars; Venus). Some of the volcanic features observed from Magellan imagery (Figure 16), especially the channels (canali), have been considered to be the result of carbonatitic magmatism.

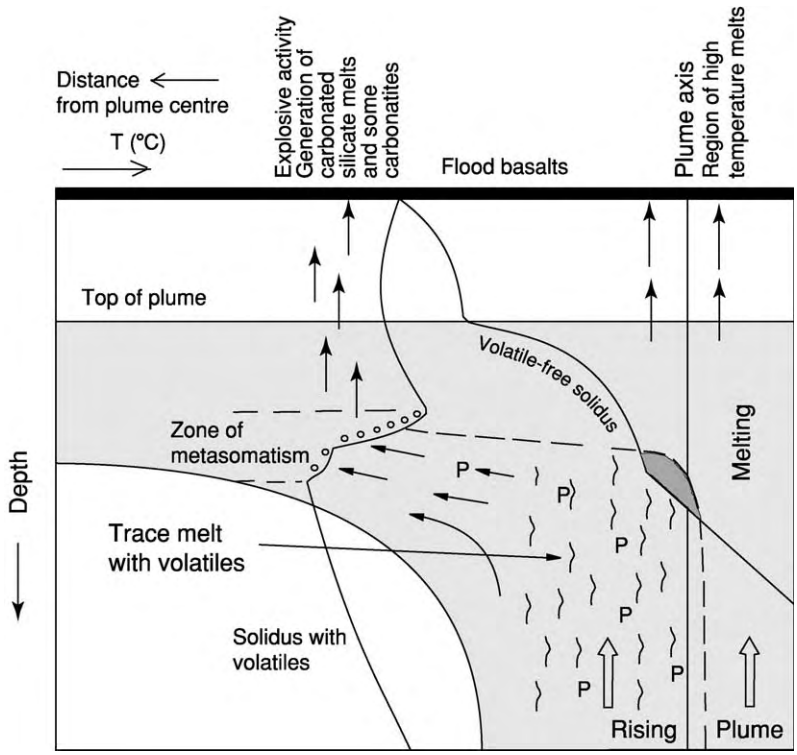


Figure 14 Plume head showing the position of mantle solidii, one vapour free and the other vapour rich. Migration of volatile rich melts from the cooler, outer parts of the plume hit the mantle ledge producing metasomatism, explosive activity, and the possible release of low viscosity melts. P represents material from the deeper parts of the mantle. Adapted from Wyllie (1988).

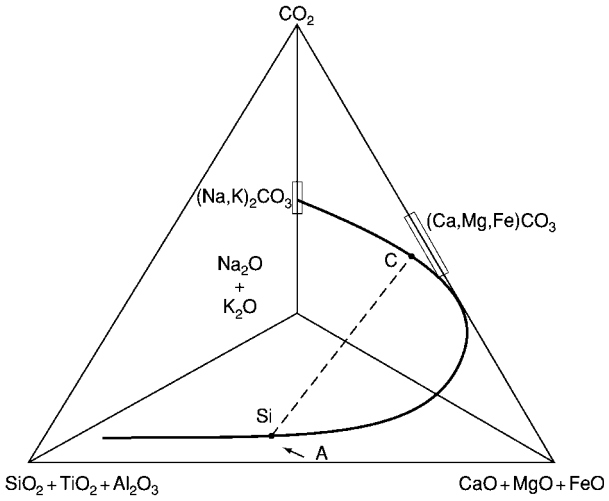


Figure 15 Quaternary diagram showing liquid trends in a silicate carbonate system. Solid curved line is a hypothetical solvus. Point A represents a carbonated silicate melt which on cooling and fractionation moves into the two liquid field where carbonatite of composition C is exsolved. This coexists with a silicate liquid of composition Si. The dashed line connects the conjugate silicate carbonate pair. Adapted from Kjarsgaard and Hamilton (1989).

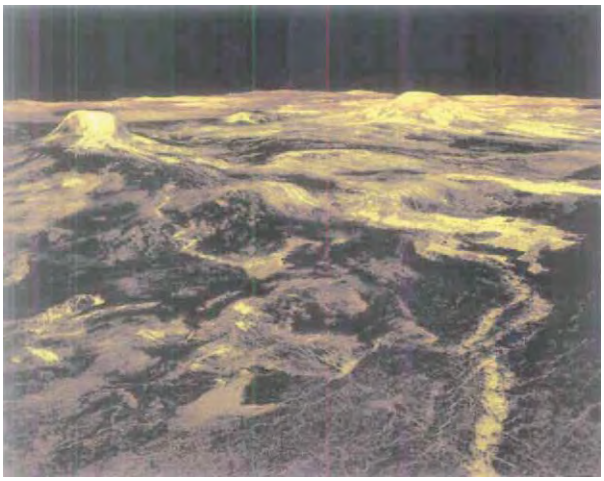


Figure 16 Lavas on Venus. This 3D, false colour, computer generated, radar image shows Venusian lava flows (yellow in colour) extending for at least 700 km from Gula Mons (upper left), a volcano about 3 km high. The other volcano is Sif Mons that lies about 730 km from Gula Mons. Note the intense fracturing in the right foreground. (JPL image PIA 00102; Magellan mission to Venus.)

See Also

Carbon Cycle. **Earth:** Mantle. **Igneous Rocks:** Kimberlite. **Lava**. **Minerals:** Carbonates. **Rocks and Their Classification**. **Sedimentary Rocks:** Phosphates. **Solar System:** Venus; Mars. **Volcanoes**.

Further Reading

Bailey DK (1993) Carbonatite magmas. *Journal of the Geological Society of London* 150: 637 651.

Bell K (ed.) (1989) *Carbonatites: Genesis and Evolution*. Unwin Hyman.

Bell K and Keller J (eds.) (1995) *Carbonatite Volcanism*. Springer Verlag.

Bell K and Tilton GR (2002) Probing the mantle: the story from carbonatites. *Eos* 83.

Dawson JB, Pinkerton H, Norton GE, Pyle DM, Browning P, Jackson D, and Fallick AE (1995) Petrology and geochemistry of Oldoinyo Lengai lavas extruded in November 1988: magma source, ascent and crystallization. In: Bell K and Keller J (eds.) *Carbonatite Volcanism,* pp. 47 69. Springer Verlag.

Harmer RE and Gittins J (1998) The case for primary, mantle derived carbonatite magma. *Journal of Petrology* 39: 1895 1903.

Keller J and Spettel B (1995) The trace element composition and petrogenesis of natrocarbonatites. In: Bell K and Keller J (eds.) *Carbonatite Volcanism,* pp. 70 86. Springer Verlag.

Kjarsgaard BA and Hamilton DL (1989) The genesis of carbonatites by immiscibility. In: Bell K (ed.) *Carbonatites: Genesis and Evolution,* pp. 388 404. Unwin Hyman.

Kogarko LN, Kononova VA, Orlova MP, and Woolley AR (1995) *Alkaline Rocks and Carbonatites of the World. Part 2: Former USSR*. Chapman and Hall.

Le Bas MJ (1977) *Carbonatite Nephelinite Volcanism*. John Wiley and Sons.

Le Bas MJ (1984) Oceanic carbonatites. In: Kornprost J (ed.) *Kimberlites 1: Kimberlites and Related Rocks, Proceedings of the Third International Kimberlite Conference*, pp. 169 178. New York: Elsevier.

Sage RP and Watkinson DH (1991) Alkalic rock carbonatite complexes of the Superior Structural Province, northern Ontario, Canada. *Chronique de la Recherche Minière* 504: 5 19.

Samoilov VS (1991) The main geochemical features of carbonatites. *Journal of Geochemical Exploration* 40: 251 262.

Simonetti A, Bell K, and Shrady C (1997) Trace and rare earth element geochemistry of the June 1993 natrocarbonatite lavas, Oldoinyo Lengai (Tanzania): implications for the origin of carbonatite magmas. *Journal of Volcanology and Geothermal Research* 75: 89 106.

Wall F and Zaitsev AN (eds.) (2004) Phoscorites and carbonatites from mantle to mine: the key example of the Kola Alkaline Province. *The Mineralogical Society Series,* 10.

Woolley AR (1987) *Alkaline Rocks and Carbonatites of the World. Part 1: North and South America*. University of Texas Press.

Woolley AR (2001) *Alkaline Rocks and Carbonatites of the World. Part 3: Africa*. London: The Geological Society.

Woolley AR and Kempe DRC (1989) Carbonatites: nomenclature, average chemical compositions, and element distribution. In: Bell K (ed.) *Carbonatites: Genesis and Evolution,* pp. 1 14. Unwin Hyman.

Wyllie PH (1988) Solidus curves, mantle plumes, and magma generation beneath Hawaii. *Journal of Geophysical Research* 93: 4171 4181.

Wyllie PJ, Baker MB, and White BS (1990) Experimental boundaries for the origin and evolution of carbonatites. *Lithos* 26: 3 19.

Granite

A I S Kemp, University of Bristol, Bristol, UK

Introduction

The transformation of the Earth's earliest mafic to ultra-mafic crust into the stable high-standing continental landmasses was achieved by granitic plutonism. This involves the generation and ascent of incompatible element-rich silicic magmas, which leave behind dense dehydrated residues that either accumulate in the deep crust or sink into the mantle (**Figure 1**). Dating of the highly resistant mineral zircon, which crystallizes from silicic magmas, hints that this differentiation process was underway by approximately 4.4 Ga, shortly after the Earth's formation, and it continues to the present day. Unlike their mafic counterparts basalts and gabbros, granitic rocks are composed of light silica-rich minerals and are difficult to destroy by subduction or tectonic delamination. Granitic rocks therefore preserve a continuous record of the changing thermal and possibly atmospheric conditions throughout Earth's history.

In broad terms, granite is a crystalline plutonic igneous rock consisting essentially of quartz and feldspar, in which the relative proportion of the former is between 20% and 60%. Yet, this simple definition belies the remarkable variety of colour, grain size, texture, mineralogy, and composition that is intrinsic to the granitic family. Furthermore, granite bodies are fossil magma chambers and as such provide a 'snapshot' of part of a dynamic magmatic system; however,

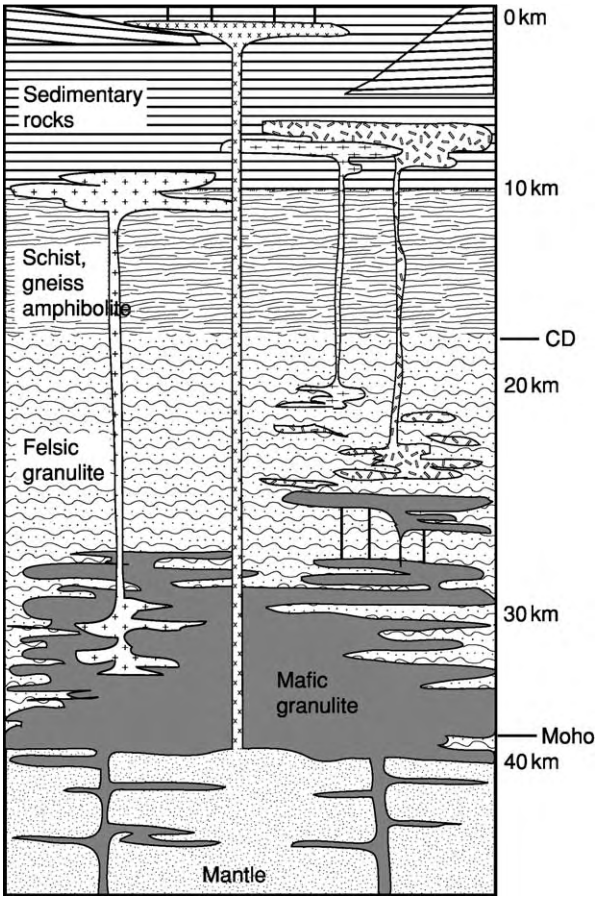


Figure 1 Simplified view of crustal differentiation during partial melting and granite formation. Granitic rocks have crossed patterns. CD, Conrad Discontinuity. (Adapted from Johannes W and Holtz F (1996) *Petrogenesis and Experimental Petrology of Granitic Rocks*. Berlin: Springer Verlag.)

as with animal fossils, the picture is often incomplete and potentially unrepresentative. Valuable information lies encoded in textures and in mineral and bulk-rock compositions, though its interpretation is rarely straightforward.

For these reasons, models explaining the nature and origin of granitic rocks have always been contentious. Debate culminated in the high-profile and sometimes acrimonious 'granite controversy' of the early twentieth century, waged between the transformists and the magmatists. This focused on the perceived 'room problem' for the emplacement of granitic magmas, the role as fluids as 'granitizing' agents, the link between granitic and volcanic rocks, and whether the large amounts of granite observed could be differentiated from basaltic liquid. Concerning this last aspect, a prominent disputant, H H Read, remarked that it 'degrades granitic magma to the status of the dregs of the primary basaltic', epitomizing the polarity of views expressed at this time.

Nonetheless, the debate stimulated conceptual advances in granite geology and in the Earth sciences in general. For example, the experimental studies of N L Bowen and co-workers that firmly resolved the debate in the magmatists favour introduced some fundamental concepts, such as fractional crystallization and partial melting, that remain the cornerstone of modern igneous geochemistry and the key to understanding crustal differentiation. With the advent of plate tectonics, it became possible to place granitic magmatism into a wider geodynamic framework, whereby the physical processes driving the extraction, transport, and emplacement of granitic magmas became quantifiable. This framework has been coupled with thermal models, to constrain the nature of the heat sources required to generate granitic magmas and the most favourable tectonic scenarios for this to occur. Modern studies seek to integrate these physical and geochemical parameters and ultimately to build a picture of the source rocks of the granite plutons, the way in which these melted to generate the silicic magma, and how this traversed the crust and acquired its present mineralogy and chemical composition.

Classification Schemes

Rocks of the granite family exhibit a continuum of compositions, reflecting derivation from diverse and multiple crustal and mantle sources, all of which are typically a matter of deduction. It is therefore prudent to classify granitic rocks according to parameters that can be directly observed and measured, such as modal mineralogy and chemical composition.

The simplest subdivision is made on the basis of the ratio of alkali feldspar to plagioclase, giving rise to the alkali feldspar granite, granite (*sensu stricto*), granodiorite, and tonalite categories. The abundance of mafic minerals typically increases with the proportion and anorthite content of the plagioclase.

A host of chemical classification schemes also exist. Of these, the alumina saturation index (ASI), based on the molecular ratio of Al_2O_3 to $(CaO + Na_2O + K_2O)$, is particularly useful, since it can be tied readily to mineralogy (Table 1). Further subdivision can be made on the basis of alkali–lime relationships, which has the advantage of highlighting the differentiation trends shown by different granitic suites and the links to potential parental magmas (Figure 2).

Finally, granites may be classified according to their oxidation state, which is manifested in the ratio of ferrous to ferric iron (Fe^{2+}/Fe^{3+}) or more simply in the assemblage of iron oxides and mafic silicates present relative to a reference reaction or 'buffer'. The most useful oxygen buffer for granitic rocks is the fayalite–magnetite–quartz (FMQ) reaction (Figure 3).

Table 1 Classification of granitic rocks according to the alumina saturation index (ASI), listing the diagnostic mineralogy and common accessory phases occurring in rocks of each category (accessory zircon and apatite are ubiquitous). Note that metaluminous magmas will become weakly peraluminous by fractionation of calcic phases (e.g. amphibole) or assimilation of aluminous sediments. In applying this classification, care must be taken to ensure that the ASI is determined from pristine samples, as subsolidus alteration and weathering can leach soluble CaO and Na_2O, thereby enriching the sample in immobile Al_2O_3 and elevating the ASI beyond its primary magmatic value

Category	*ASI*	*Diagnostic mineralogy*	*Biotite colour*	*Accessory phases*
Strongly peraluminous	>1.1	Cordierite, garnet, muscovite, andalusite sillimanite, topaz	Rusty red brown	Monazite, xenotime, uraninite
Weakly peraluminous	1.0 1.1	Biotite, aluminous amphibole (cummingtonite or gedrite), calcic amphibole	Red brown to dark brown	Monazite, titanite, allanite
Metaluminous	<1.0	Clinopyroxene, hornblende, epidote, calcic plagioclase (to An_{90})	Dark chocolate brown to greenish brown	Titanite, allanite
Peralkaline	<1.0	Alkali amphibole, aegirine, albitic plagioclase	Very dark brown to black	Fluorite, titanite, allanite, thorite

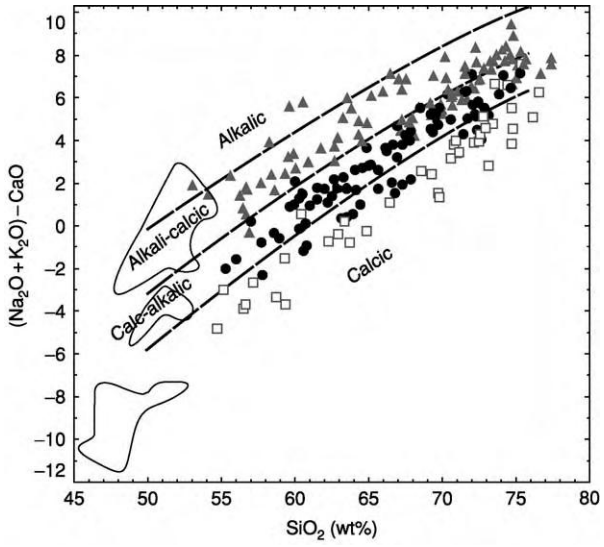


Figure 2 Subdivision of granitic rocks according to the modified alkali lime index of Frost BR, Barnes CG, Collins WJ, *et al.* (2001), as given by a plot of $(Na_2O + K_2O)$ CaO versus SiO_2. Note that the trends defined by the three granitic suites the subduction related Peninsula Ranges (mostly calcic; open squares) and Sierra Nevada (mostly calc alkalic; filled circles) batholiths (western USA) and the syn to post collisional Caledonian plutons of the UK (mostly alkali calcic; grey triangles) project at low silica towards coeval mantle derived rocks in the three areas (circled fields).

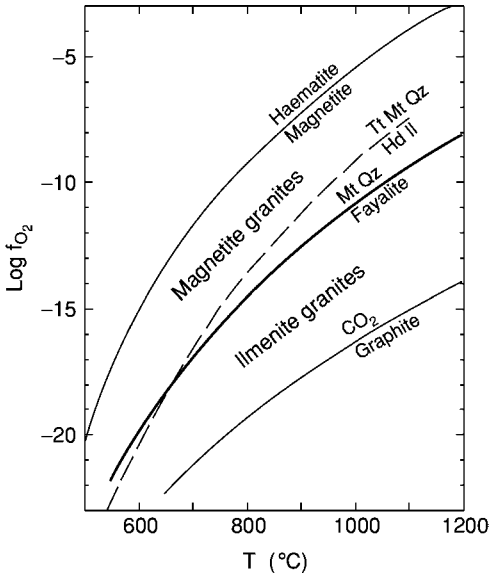


Figure 3 Plot of log f_{O_2} (oxygen fugacity) against temperature, showing the locations of commonly used buffer curves and the crystallization fields of ilmenite and magnetite bearing granitic rocks. Hd, hedenbergite; Il, ilmenite; Mt, magnetite; Qz, quartz, Tt, titanite.

Granitic magmas located below FMQ are considered to be reduced and crystallize ilmenite ($Fe^{2+}TiO_3$), whereas magnetite ($Fe^{2+}Fe_2^{3+}O_4$) appears in those relatively oxidized magmas that have a higher oxygen fugacity (f_{O_2}) than FMQ. Ilmenite is destabilized at even higher f_{O_2}, whereupon magnetite is accompanied by titanite as the major titanium-bearing phase. Mafic silicates (e.g. hornblende, pyroxene) tend to be enriched in magnesium in strongly oxidized granites, since the iron is sequestered by magnetite.

Determining the oxidation state of granitic rocks is important for three reasons. First, it has been suggested that the oxygen fugacity of a granitic magma reflects its particular source rocks, and so this parameter can provide valuable petrogenic information. Second, the differing magnetic susceptibilities of oxidized and reduced granites result in contrasting aeromagnetic signatures, and this can be used as a mapping and exploration tool in areas of limited outcrop. Last, the redox state of granitic magmas has a bearing on the types of ore deposits associated with granitic plutons (*see* **Mining Geology:** Magmatic Ores). For example, copper, gold, and molybdenum mineralization tend to be associated with magnetite-bearing granites,

whereas economic deposits of tin, tungsten, and tantalum are related to ilmenite-bearing intrusives.

Where are Granites Found?

Granitic rocks, their extrusive equivalents (rhyolites and dacites), and sedimentary and metamorphic rocks of granitic composition are easily the most abundant constituents of the upper continental crust, and they are widespread on all continents. An early observation was that large composite granitic bodies (batholiths) were commonly, but not exclusively, found as linear belts defining mountain ranges and, therefore, that their emplacement was linked in some way to orogenesis or at least to large-scale interactions between lithospheric plates (*see* **Tectonics:** Mountain Building and Orogeny). The paucity of granitic bodies in oceanic environments is testimony to the importance of continental materials, and possibly lithospheric thickness, in the genesis of these magmas, implying that granitic generation is associated with the remobilization of older crust.

However, granitic bodies are encountered in all tectonic settings, although their volumes, compositions, and the nature of the associated igneous rocks vary accordingly (Table 2). The relationship between granitic composition and tectonic setting can be used to reconstruct the tectonic evolution of ancient orogenic belts.

Ascent and Emplacement Mechanisms

Granitic bodies range in volume from less than 1 km^3, for single intrusions, to in excess of 10^6 km^3, for example in batholithic structures such as the Coast Plutonic Complex in British Columbia. They are classically depicted as circular-to-elliptical plutons on geological maps, though the shapes, field relationships, and internal structures of these bodies may be far more complex. The outcrop pattern of a granite pluton depends on the level and topography of exposure and on the three-dimensional shape of the body. To a large extent, the latter is dictated by the emplacement mechanism, which is in turn related to the depth of emplacement – and thus to the rheology of the host rocks – and to the geodynamic regime into which the pluton was intruded. Emplacement occurs at a neutral buoyancy level or mechanical discontinuity, such as the brittle–ductile transition, whereupon magma flow switches from dominantly vertical to dominantly horizontal.

The way in which large plutonic masses are emplaced in the rigid upper crust has long been contentious. The so-called 'room problem' has even formed the basis of arguments that granite has a non-magmatic origin. However, our concepts of how large volumes of granitic magma are extracted from the source region, transported tens of kilometres through the crust, and intruded at higher structural levels have recently been revolutionized in two ways. First, the traditional view that partial melting in the deep crust proceeds until a rheological threshold is reached, whereupon the source region slowly upwells *en masse* as a gravity-driven hot-Stokes diapir has been largely supplanted by a more dynamic scenario, in which the entire magma segregation, extraction, and transport process is controlled by deviatoric stress attending tectonic activity, perhaps augmented by compaction at high melt percentages. The exact way in which this occurs is probably specific to the source rock type and the nature and rate of deformation, but valuable insight is provided by deep-crustal migmatite terranes. These areas are characterized by deformation-enhanced segregation of partial melts from their residues and subsequent migration along planar anisotropies (foliations or lithological layering) to dilational structural sites, such as boudin-necks or fold axial surfaces (Figure 4A). An additional driving force is the volume increase associated with some partial-melting reactions, which can increase permeability by promoting microfracturing. In areas of local melt accumulation, high magma pressures may develop in response to compressional stress and/or fluctuation in magma supply. Subsequent magma (and thus heat) transfer to higher crustal levels occurs through interconnected networks of narrow structurally controlled channels or active shear zones during transpression, and as a result partial melts can be efficiently expelled from the source region. Excellent examples of these relationships occur in the Arunta Inlier of central Australia and in the northern Appalachians of eastern North America (*see* **North America:** Northern Appalachians), where plutons are localized in regional shear-zone systems and rooted in migmatites.

In addition to its pervasive flow in distributed pathways, granitic magma may ascend from source regions in a more focused fashion by exploiting pre-existing faults, or as self-propagating dykes, essentially driven by buoyancy. Thermal models have established that such magmatic conduits penetrating cold crust must maintain a critical thickness (typically 3–12 m) and magma-flow rate to avoid freezing and so are most viable where the generation and extraction of melt from the source is rapid. Magmatic diapirism operating over short length scales may be of limited importance in the ductile deep crust, where thermal and rheological contrasts with the host rocks are minimal, and where the viscosity of granitic

Table 2 Empirical relationships between the common granitic rock types, their typical compositions, and the geodynamic environments into which they are emplaced. The various granite types may be juxtaposed in plutonic belts whose tectonic setting has evolved during an orogenic cycle

Tectonic setting	*Occurrence*	*Dominant granite type*	*Composition*	*Associated igneous rocks*	*Other features*	*Examples*
Mid ocean ridges, ophiolites	Dykelets and net vein complexes to small plutons	Tonalite	Metaluminous; calcic with very low K, Rb	Pillow basalt, sheeted dykes, gabbro	May be associated with amphibolites in shear zones	Western Ophiolite Belt, Albania; Lizard Complex, south west England; Oman Ophiolite
Oceanic arc	Strongly zoned plutons	Tonalite, granodiorite	Metaluminous; calcic to calc alkaline	Gabbro, quartz diorite, diorite, monzodiorite	Low LILE contents, isotopically primitive Cu Au deposits	New Britain; Papua New Guinea; Aleutian Islands; Alaska
Continental arc	Vast linear batholiths elongate parallel to the trench; zoned plutons are common	Tonalite, granodiorite, granite	Mostly metaluminous to weakly peraluminous; vary from calcic, calc alkaline to alkali calcic with time and distance from the trench	Gabbro, quartz diorite; andesite to dacite volcanics	High Na/K, Sr/Y; primitive to mixed isotopic signatures; Cu Mo deposits	Peninsula Ranges Batholith, USA; Sierra Nevada Batholith, USA; Coastal Batholith, Peru; Patagonian Batholith, Chile
Syn collisional	Batholiths orientated parallel to the structural grain, sheets in shear zones	Granodiorite, granite	Metaluminous to strongly peraluminous; calc alkaline and alkali calcic	Gabbro, diorite (uncommon)	Moderate Na/K, mixed to crust like isotopic signatures; high T, low P metamorphism; Sn W deposits	Lachlan Fold Belt, Australia; Massif Central, France; Appalachians, USA; Himalayas
Post collisional	Individual plutons or groups of plutons oblique to structural trends	Granodiorite, granite	Metaluminous to strongly peraluminous; alkali calcic	Appinite, shoshonite, lamprophyre	Low Na/K; mixed to crust like isotopic signatures	Caledonian Fold Belt, UK; western Lachlan Fold Belt; Hercynian Belt, western Europe.
Continental rift or intraplate	Discrete plutons, ring complexes	Granite	Metaluminous to peralkaline; alkalic	Gabbro, basalt, silicic volcanics	Isotopically primitive highly differentiated, high Nb, Zr, Zn, Ga, Nb, Th, U, REE deposits	East African Rift; Eastern Red Sea Margin; Corsica; Tertiary Igneous Province, UK

LILE, large ion lithophile elements.
REE, rare earth elements.

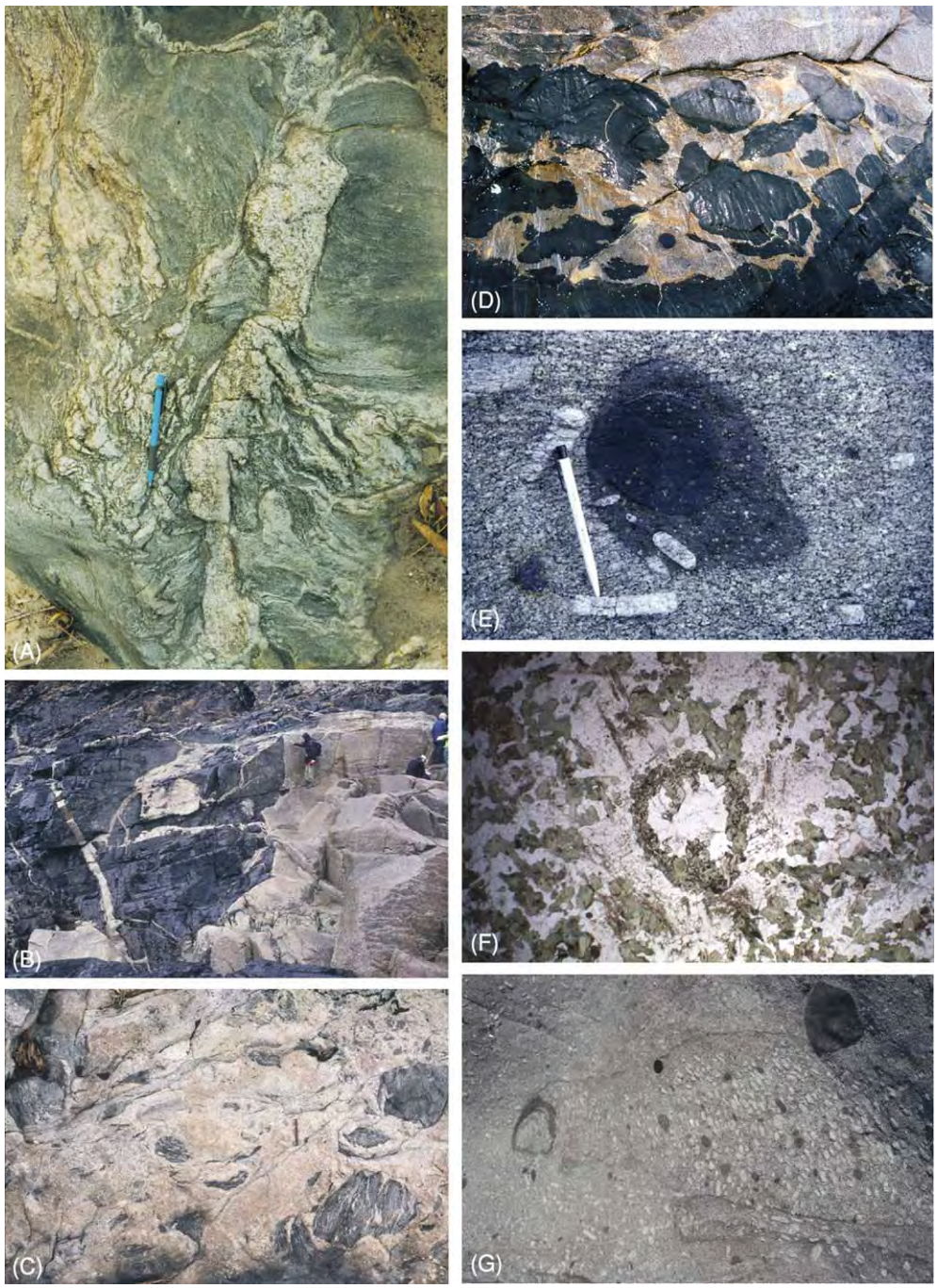


Figure 4 Granitic rocks in the field. (A) Migmatite outcrop from south eastern Australia, showing the deformation enhanced segregation and accumulation of locally derived partial melts into a dilatant shear zone. In this case, the leucogranite in the shear zone is compositionally equivalent to the kilometre scale leucogranitic plutons of the area. (B) Strongly peraluminous leucogranite invading and stoping its metasedimentary roof rocks along bedding planes and cracks (Cornwall, UK). (C) Enclaves of gneiss and migmatite in a deep crustal (*ca.* 20 km) granitic pluton in south eastern Australia that has been retained close to the generative region. The enclaves correspond to less fertile horizons of the granite's source rocks. (D) Mingling between a syn plutonic basaltic dyke and a metaluminous tonalite, Moruya Batholith, eastern Lachlan Fold Belt (Australia). (E) Composite enclave, in which a mafic igneous enclave (basaltic andesite) is enclosed by a microgranular enclave. Note the physical incorporation of potassium feldspar pheno crysts from the enclosing granite (Elba Island, Italy). (F) Thin section of a microgranular enclave, showing a large quartz grain rimmed by pyroxene, which is overgrown in turn by hornblende (3 mm across). The quartz grain is interpreted as a xenocryst incorporated from the host granite, which reacted with, and was partially resorbed by, the hotter enclave magma. (G) Unusual vortex like structure in granitic rock, which is thought to represent a cross section through a magma flow conduit that has subsequently been 'infilled' by crystal rich magma during crystallization. Features such as this hint at the dynamic nature of upper crustal magma chambers and demonstrate how considerable compositional heterogeneity can be caused by simple magmatic flow processes in granitic crystal mushes (Elba Island, Italy).

magmas is elevated by the carriage of unmelted source material (restite).

Furthermore, it is becoming increasingly apparent from geophysical imaging (seismic, gravity), analysis of magnetic directional fabrics, and detailed field mapping that the three-dimensional geometries of many granitic plutons are tabular or funnel-shaped, with inclined feeder zones, and that plutons have an internally sheeted structure on a metre to kilometre scale. The implication is that, rather than being intruded in a single episode, granitic plutons (and, on a larger scale, batholiths) were constructed sequentially and *in situ* by the coalescence of multiple sill-like magma pulses. The most compelling demonstrations of this are in the Himalayas and Greenland Caledonides (*see* **Europe:** Scandinavian Caledonides (with Greenland)), where great topographical relief and superb exposure reveal that leucogranitic bodies have laccolithic geometry, where the constituent lenticular intrusions are separated by thin screens of country rock. The same situation applies in some magmatic arcs. Gravity surveys show that the vast Coastal Batholith of Peru is floored at surprisingly shallow depths (less than 7 km) and comprises a collage of numerous sheet-like plutons.

Tabular plutons in the upper crust are fed by an underlying magma conduit, and, as highlighted above, this can occur in several ways. Networks of feeder dykes are clearly exposed beneath the sheeted leucogranites of the Himalayas. The importance of felsic-magma transport by dyking in the brittle uppermost crust receives further support from the geometry of sub-volcanic ring complexes, in which magma emplacement occurs as 'ring dykes' in conical fractures. Some of the best examples of these are the alkaline ring complexes of the British Tertiary Igneous Province.

One important consequence of incremental and dynamic models of pluton assembly is that the 'room' problem is largely circumvented. The incoming granitic magmas are progressively accommodated by a combination of tectonism – involving movement on pre-existing tensional faults and fractures and ponding of magma in dilatant jogs or pull-apart structures – and roof lifting during pluton inflation or ballooning, with subsidence of the floor of the magma chamber, possibly in response to magma extraction at depth. The proximity of many plutons to major fault systems and their parallelism to regional strain patterns, as is especially evident in the Circum-Pacific subduction-related batholiths, support the notion of structurally controlled emplacement. However, it is worth noting that plutons emplaced late in the orogenic history overprint deformational fabrics and are randomly aligned relative to the structural grain. There is sometimes evidence for the localized stoping and assimilation of the roof rocks (Figure 4B), though this is probably a late-stage phenomenon rather than a major ascent mechanism.

Enclaves

Pieces of rock enclosed by granitic bodies that exhibit a colour or textural contrast with the surrounding granite are termed enclaves. Although comprising a small volumetric portion of most plutons, enclaves are eye-catching and exhibit a bewildering array of colours, shapes, sizes, textures, and compositions. Enclaves are a source of information about the sorts of processes that may have operated during the evolution of granitic magmas, and so their interpretation is an important aspect of granitic studies.

The main enclave types encountered in granitic rocks are listed in Table 3. Some are xenoliths spalled from the country rock, but the origin and significance of other enclave varieties remain debated. High-grade metamorphic enclaves have been entrained from depth, though whether they are restite from the granitic source or simply mid-crustal xenoliths is difficult to determine (Figure 4C). A restitic origin seems reasonable for those enclaves with melt-depleted compositions, such as the cordierite–spinel lumps found in some peraluminous granites. Mafic igneous enclaves have basaltic to dioritic compositions and are derived from the injection and fountaining of syn-plutonic mafic magma into the granitic pluton whilst it was partially crystalline (Figure 4D). As a result of the thermal and rheological contrasts between the hot fluid basaltic magma and the cool viscous granite, the mafic enclave commonly exhibits quenched margins and complex involute shapes resulting from contraction.

In appearance, mafic igneous enclaves are transitional to microgranular enclaves. These may also have a finer-grained margin, though, rather than chilling, in some instances this is related to concentration of biotite resulting from a reaction between the iron- and magnesium-rich enclave and the potassium- and water-rich granite. Microgranular enclaves have unusual igneous-looking textures, and their mineralogy and composition commonly, though not invariably, correlate with those of the host. They are variously considered to represent an intermingled globule of relatively mafic magma that has undergone hybridization with the host granite, a reincorporated chilled margin derived from a less-evolved facies of the host, or partially melted restite entrained from the source. Each of these origins may be correct in specific instances. A mingling origin seems apposite where the enclave exhibits features that confirm its incorporation in the molten state, such as magmatic

Table 3 Summary of the various enclave types that are encountered in granitic rocks. Note that, for deep seated granitic bodies enclosed by their migmatitic source rocks, country rock enclaves may correspond to fragments of the protolith or to relatively refractory source horizons

Enclave type	*Occurrence*	*Morphology*	*Description*	*Possible origin*
Country rock	All granites; may be concentrated towards the periphery	Tabular, angular	Matches exposed country rock; may have a reaction rim (especially calc silicates) or be in the arrested stages of assimilation	Accidental entrainment
High grade metamorphic	Most common in strongly peraluminous granites; virtually absent from peralkaline granites	Tabular, lenticular	Higher metamorphic grade and typically more complexly deformed than host rocks; micaceous objects are common	Mid crustal xenolith; restite
Mafic igneous	All granites; occur in swarms or associated with syn plutonic mafic dykes	Variable globular, elliptical, lenticular, pillow like, or amoeboid; cuspate or crenulate margins	Fine grained, commonly with chilled margins; may show evidence of magmatic deformation, usually stretching or contortion	Intermingled globules of relatively mafic magma
Microgranular	All granites, though rare in peralkaline; dispersed more or less evenly throughout the pluton	Rounded, elliptical to ovoid, rarely amoeboid	Generally darker and finer grained than the host, though the mineralogy may be the same; typically biotite rich; may entrain crystals of the host granite	Hybridized mafic globule; fragment of less evolved granite; restite
Coarse igneous	All granites, but uncommon	Rounded or angular	Fragments of relatively coarse grained plutonic rock; may be mafic or felsic	Cumulate; xenolith of unexposed pluton

flow textures or the engulfing of phenocrysts from the surrounding granite (Figure 4E). The occurrence of disequilibrium or hybridization features also attests to an intermingled origin (Figure 4F).

Mineralogy and Textures of Granitic Rocks

Besides quartz and feldspar, a variety of minor and accessory minerals may occur in granitic rocks, as dictated by the bulk composition of the magma (including its oxidation state and water content) and the pressure and temperature at which igneous crystallization took place (Table 1). An additional complication is that, instead of precipitating directly from the granitic magma, some of the minerals may be of extraneous origin (xenocrysts derived by assimilation of the country rocks) or be residual or restitic crystals entrained from the crustal source of the granitic magma. In some cases this may be resolved by textural examination, but this is by no means diagnostic; plutonic rocks cool slowly, and thus there is ample opportunity for conversion of early (restitic or magmatic) phases to minerals stable at the pressure and temperature conditions of emplacement. Primary high-temperature textures can be obscured, or even obliterated, by these reactions.

Strongly peraluminous granites contain an aluminium-saturating phase, sometimes accompanied by orthopyroxene or pseudomorphs after this mineral. Orthopyroxene is especially common in deep-seated plutons of granulite terranes, such as the tonalites of the Hidaka Metamorpic Belt, Japan. Of the peraluminous minerals, cordierite is texturally the most complex (Figure 5). It can occur as (a) euhedral crystals, (b) large poikilitic grains intergrown with quartz and alkali feldspar, (c) anhedral masses sieved with orientated inclusions of biotite, fibrous sillimanite, and/or spinel, commonly enveloped by euhedral inclusion-free cordierite, and (d) overgrowths on garnet. These different textural varieties may even be encountered in the same sample, and they highlight the type of information available from understanding mineral textures in granitic rocks. The first and possibly the second of these textures are magmatic, whereas the third is either restitic or xenocrystic. The cordierite rims on garnet reflect decompression, which could be related to magma ascent. The garnets themselves (mostly almandine-rich) are also potentially either magmatic or restitic/xenocrystic in origin, though the small manganese-rich garnets found in leucogranites are probably magmatic. In contrast to the felty aggregates of sillimanite, andalusite may form euhedral melt-precipitated crystals in peraluminous

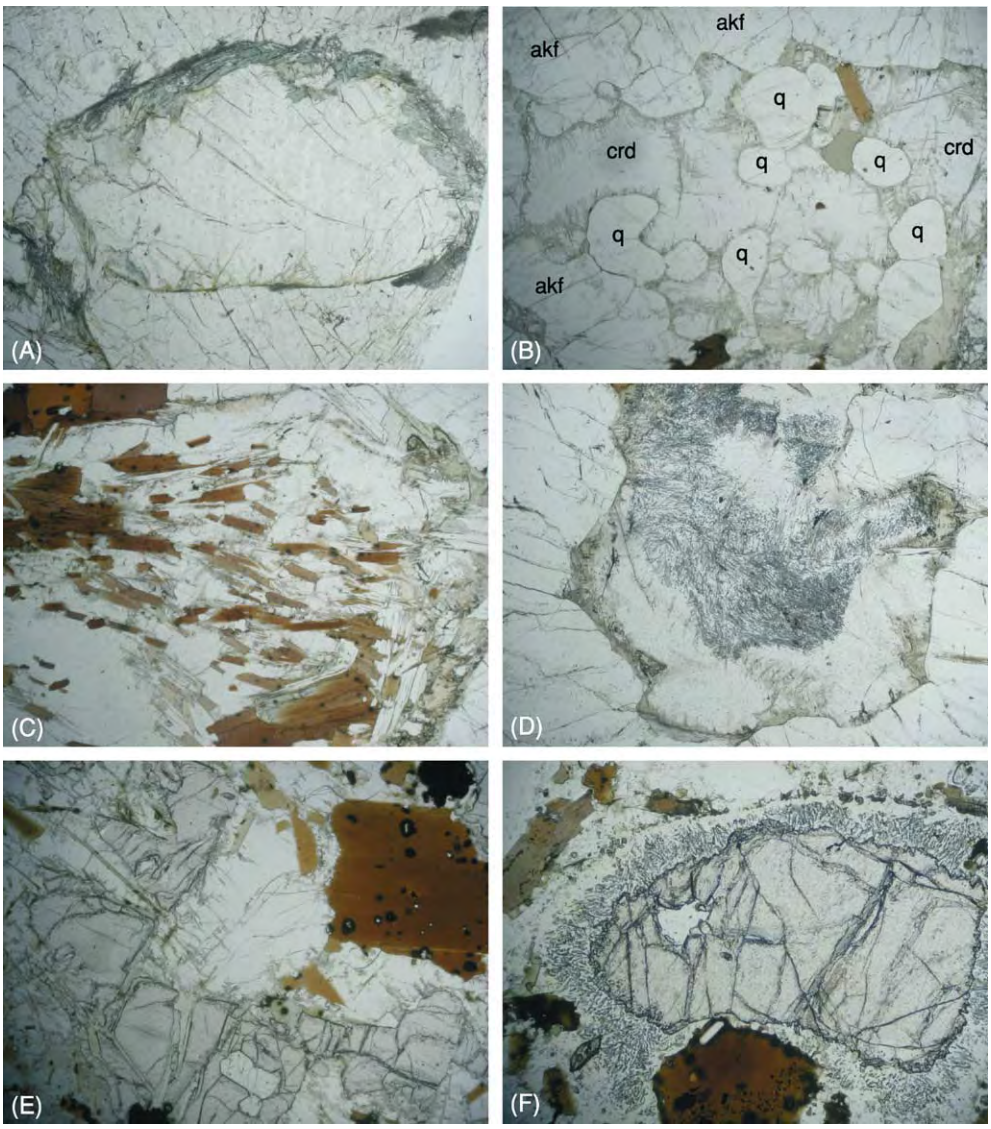


Figure 5 The different cordierite morphologies encountered in strongly peraluminous granitic rocks of the Lachlan Fold Belt, south eastern Australia. All photographs were taken in plane polarized light and are 3.5 mm across. (A) Well shaped magmatic cordierite crystals with a thin rim of greenish biotite and fine andalusite. (B) Irregularly shaped cordierite grain intergrown with quartz (q) and alkali feldspar (akf), crd cordierite. Note the fine micaceous alteration along cracks and cleavages. (C) Large cordierite mass enclosing numerous biotite inclusions, which preserve an earlier directional fabric. The granite is undeformed, so the cordierite is either restitic or (more likely) xenocrystic. (D) Fibrous aggregates of sillimanite and fine spinel grains in the core region of a cordierite grain. The clear inclusion free rim is euhedral and melt precipitated. (E) Reaction of garnet (high relief, brownish) to cordierite (centre, clear). (F) Garnet with a reaction corona of cordierite and wormy intergrowths of orthopyroxene and plagioclase. The small circular yellow patches inside the clear cordierite are haloes around minute zircon inclusions and reflect radiation damage caused by the decay of uranium and thorium in these inclusions.

granites. These are most commonly observed in highly differentiated plutons (co-existing with topaz), where the solidus temperature has been depressed into the andalusite stability field by high concentrations of fluorine and/or boron. Magmatic andalusite of this type occurs in some granites of the South Mountain Batholith, Nova Scotia.

Muscovite in strongly peraluminous granites also has two potential modes of occurrence. Large euhedral plates, sometimes with oscillatory zoning and 'dovetail' intergrowths with biotite, are thought to be magmatically precipitated. However, poorly shaped muscovite flakes enclosing and/or replacing other aluminous phases are probably of subsolidus origin, especially where the rock has been deformed or hydrothermally altered. Secondary muscovite is a common replacement product of feldspar and thus may occur in altered metaluminous rocks.

Metaluminous granites have a simpler mineralogy. The diagnostic minerals are hornblende and clinopyroxene (augite), which may be accompanied by orthopyroxene, accessory titanite and/or allanite. Generally, clinopyroxene is enveloped and partially replaced by calcic amphibole, and, as this process releases silica, amphiboles formed by this reaction may be perforated with quartz. Epidote is a common alteration product, though euhedral prisms in metaluminous granites emplaced at moderate pressures are magmatically precipitated.

The mineralogy and texture of peralkaline granites are most distinctive, and it is clear that these rocks are produced by the crystallization of high-temperature completely liquid magmas at shallow crustal levels. Many are brick-red and have unusually turbid feldspars containing minute haematite inclusions, reflecting pervasive hydrothermal alteration. This typically occurs by circulation of meteoric waters contained in near-surface aquifers. Alkali feldspars show irregular perthitic features ('patch' perthite) and exhibit striking granophyric intergrowth with quartz, particularly near miarolitic cavities, which are formed by exsolution of the magmatic vapour phase at low confining pressures. The plagioclase is sodic, and so most peralkaline plutons are true granites. An alkali amphibole is usually present, sometimes accompanied by sodium-pyroxene and rarely fayalite. Biotite is iron-rich and interstitial to other phases, consistent with late crystallization and moderate-to-low magmatic water contents (2% or less).

Petrogenic Studies

Silicic magmas can be produced in two fundamental end-member ways: first, by differentiation of high-temperature mantle-derived mafic magmas, and, second, by partial melting of pre-existing crustal rocks. The end-point of the first of these and the incipient stages of the second both produce hydrous melts that cluster around the thermal minimum/eutectic in the quartz–albite–orthoclase–water system.

However, the geochemical diversity of granitic rocks is controlled by a complex interplay between the composition of the source rocks – which may be a function of the tectonic setting – the conditions of partial melting, and the operation of processes such as magma mixing, restite unmixing, assimilation, fractional crystallization, and other crystal–liquid processes (e.g. convective crystallization, filter-pressing), which may operate in tandem or sequentially from the source region to the emplacement site (Figure 6). Deconvolving these effects is a major challenge facing granite petrologists (e.g. Figure 7).

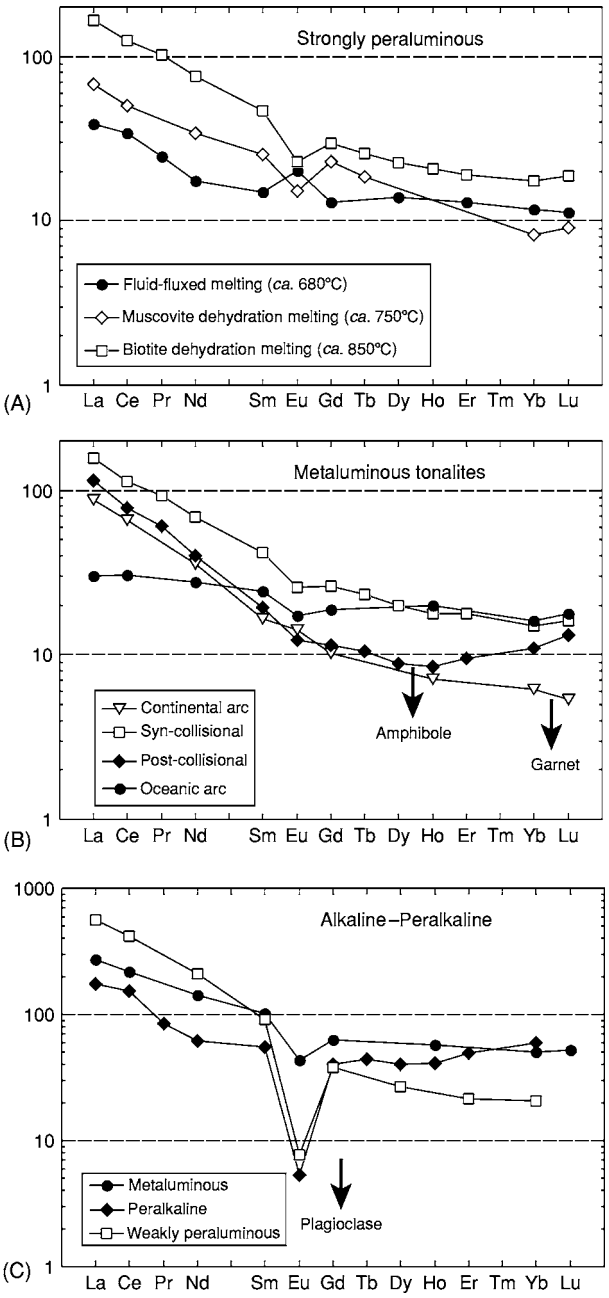


Figure 6 Chondrite normalized rare earth element (REE) plots comparing the different granite compositions produced by melting (A) metasedimentary sources and (B) meta igneous sources, under different conditions, with (C) alkaline to peralka line plutons. (A) The increasing REE content with temperature in strongly peraluminous granites reflects greater dissolution of REE rich accessory phases during melting. (B, C) Arrows indi cate compositional control by garnet (heavy REE), amphibole (middle REE), and plagioclase (europium), either by fractional crystallization or by retention of these phases in the source region during partial melting. (B) Specifically, the continental arc tonalite is thought to have been derived by high pressure melting of garnet amphibolite at the base of the crust or in the mantle, whereas (C) the alkali granites have experienced plagio clase fractionation at shallow depths.

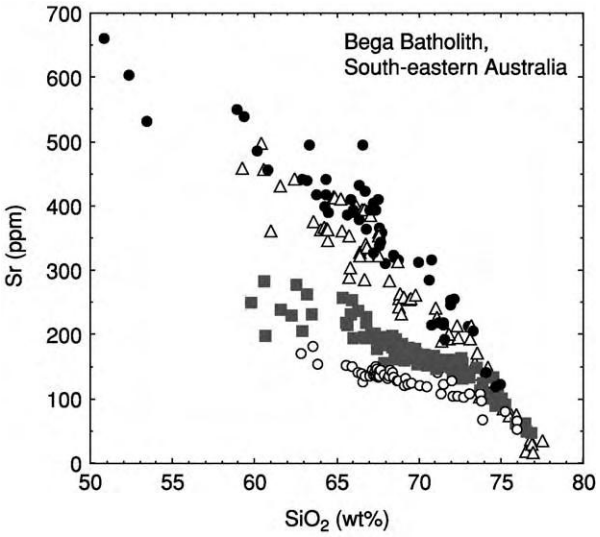


Figure 7 Plot of strontium versus silica contents for metaluminous granites of the Bega batholith, south eastern Australia. The broadly linear variation defined by the different groups of granites has been variously ascribed to restite unmixing, two component magma mixing, and fractional crystallization. In this context, the compositional convergence at high silica could reflect evolution towards a 'minimum melt' composition or that a similar felsic end member magma was involved in mixing. Examples such as this highlight the potential problems of interpreting the whole rock geochemistry of plutonic rocks in general. The various symbols represent granite suites. These are groups of rocks with similar textural and geochemical features and are thought to be related through a common source or process.

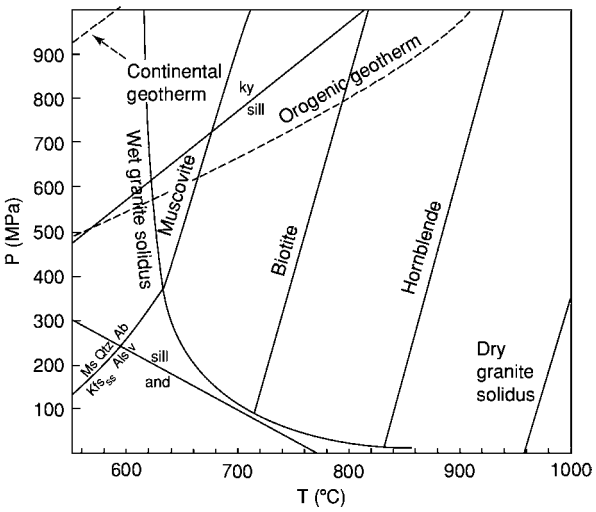


Figure 8 Pressure temperature diagram showing the beginning of melting curves for granitic compositions and the locations of the muscovite, biotite, and hornblende dehydration melting reactions. The important role of water in crustal melting is borne out the large temperature difference between the water saturated ('wet') and water absent ('dry') granite solidi. Note that melting reactions in crustal P T space (except hornblende dehydration) are intersected only by unusually hot geotherms, such as that of young orogenic crust, and not by the geotherm of stable continental areas. Abbreviations: ky, kyanite; sill, sillimanite; and, andalusite; ms, muscovite; qz, quartz; ab, albite; Kfs_{ss}, solid solution of potassium feldspar; Als, aluminosilicate. (Modified from Johannes W and Holtz F (1996) *Petrogenesis and Experimental Petrology of Granitic Rocks*. Berlin: Springer Verlag.)

Furthermore, there has been increasing recognition that, rather than representing liquid compositions, many granites were emplaced as mushes of precipitated or entrained restitic crystals, or even represent crystal cumulates in which the proportion of trapped melt is below 30%.

The origin of granitic rocks is clearest in oceanic areas remote from continental materials. Here, silicic magmas are generated either by extreme fractional crystallization or by the remelting of mantle-derived materials. Granitic rocks are exposed as the small-volume plagiogranites of mid-ocean ridges (now part of ophiolite complexes) and the alkali granites of some oceanic plateaux. The quartz diorite to tonalite plutons of intraoceanic arcs have a more complex genesis, though primitive isotopic ratios suggest that the source materials are essentially the same.

In contrast, the potential protoliths of the voluminous granitic plutons of continental areas include a diverse range of common crustal rock types of various ages, which may melt at a variety of pressures and temperatures. This is borne out by the spectrum of strontium, neodymium, and oxygen isotopic compositions shown by continental granitic rocks, which range from depleted mantle-like to highly evolved and similar to the compositions in the average continental crust. To some extent the source-rock composition is governed by tectonic setting. For example, the immature volcanogenic greywackes of convergent plate margins differ from the more weathered flysch of passive margins, and the compositions of the derivative granitic magma differ accordingly. For this reason, chemical features of silicic magmas may be diagnostic of the tectonic setting of their sources, rather than that under which magmatism occurred.

Water is a crucial ingredient for partial melting of pre-existing crust (Figure 8). It may be externally derived, for example by dewatering of underthrust sediments or crystallization of hydrous magmas, or it may be liberated during melting by decomposition of hydrous minerals ('dehydration' melting). The role of fluids in the lowermost crust, especially above subduction zones, requires further investigation. However, melting promoted by fluid influx into mid-crustal amphibolite-facies terranes tends to generate near-solidus peraluminous leucogranitic magmas that have limited vertical mobility. For this reason, most granitic magmas are thought to have been generated by higher temperature dehydration-melting reactions involving micas and amphiboles. These reactions have

negative dP/dT slopes, permitting melts generated in this fashion to ascend to shallow crustal levels before crystallizing at the water-saturated solidus (see Figure 8).

During dehydration melting, the nature and amount of the resultant granitic magma are controlled by the specific hydrous mineral participating in melting and its proportion in the rock. At the lowest temperatures (less than 750°C), muscovite dehydration melting in metapelites yields strongly peraluminous felsic liquids; this mechanism is thought to be responsible for the formation of the Himalayan leucogranites and the leucogranites of the Appalachians, Cornubian batholith (south-west England), and Iberian massif (Spain). The evolved isotopic ratios of these rocks (i.e. high $^{87}Sr/^{86}Sr$ and low $^{143}Nd/^{144}Nd$ at the time of formation) are consistent with derivation from metasedimentary protoliths. Modelling demonstrates that temperatures sufficient for partial melting can be achieved by internal radioactive heating and thermal relaxation associated with crustal thickening. Melting may be triggered or enhanced by decompression during orogenic collapse and exhumation, where additional heat is supplied by frictional heating along active fault systems. However, the melt fraction is limited by the modal abundance of muscovite in the rock and is thus generally less than 20%.

The large relatively biotite-rich granitic bodies that dominate most magmatic belts require higher-temperature (more than 850°C) biotite and hornblende dehydration-melting reactions from protoliths such as metagreywackes, intermediate igneous rocks (i.e. meta-andesites, metatonalites), or even amphibolites. Modelling and experimental studies predict that these reactions are capable of generating large proportions (up to 50 vol.%) of water-undersaturated granitic melt from protoliths with the appropriate (fertile) composition.

In nature the situation is more complex. In the absence of fluids, the elevated temperatures required for biotite and hornblende dehydration melting can be achieved only by significant heat input into the crust from the mantle, for example by the direct incursion of hot (more than 1000°C) basaltic liquids. Under these circumstances, it seems inevitable that the basaltic liquids will also physically interact with the crustal melts and therefore make some material contribution to the resultant granitic magmas. Several lines of evidence support this. First, the large-scale mingling and in some cases mixing of granitic and mafic (basaltic) magmas is commonly observed in the field, both in upper-crustal magma chambers and composite dykes and (more rarely) at deep crustal levels. The widespread occurrence of mafic or microgranular enclaves and mafic mineral clots derived therefrom demonstrates that dispersal of the mafic ingredient is pervasive throughout such plutons. Second, experimental melting studies of metasedimentary protoliths have consistently failed to reproduce the composition of the cordierite-rich granites found in thickened convergent orogens. However, these compositions are satisfactorily reproduced when a basaltic component is added to the experimental mixture. Third, some granitic rocks, especially metaluminous varieties, share compositional features with spatially associated mantle-derived magmas, strongly suggesting a genetic link. Fourth, the isotope ratios of many granitic bodies are systematically shifted to more primitive values than their inferred crustal sources. In some areas, granitic rocks form covariant plausibly mixing arrays on strontium–neodymium–oxygen isotope diagrams that converge towards depleted mantle-like values (Figure 9).

The way in which the mantle components are incorporated is not well understood. Simple mixing models are challenged by the physical and thermal difficulties of efficiently blending hot highly fluid basaltic magma with cooler viscous granitic magma,

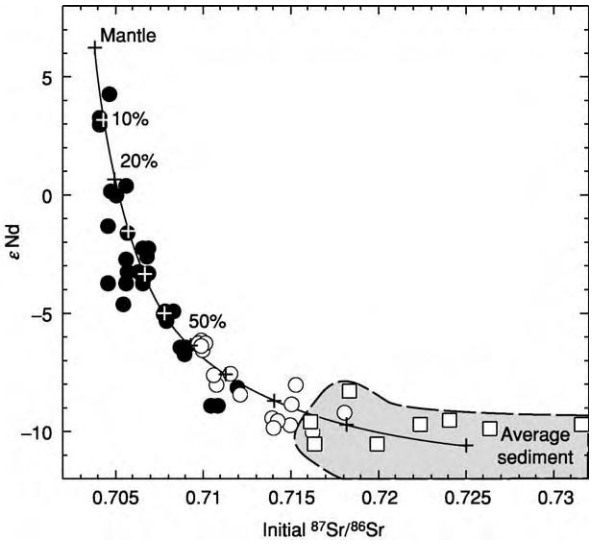


Figure 9 Diagram of epsilon neodymium versus $^{87}Sr/^{86}Sr$ for granitic rocks of the Lachlan Fold Belt (south eastern Australia) calculated at their time of formation (the epsilon notation signifies deviation from a reference chondritic reservoir in parts per 10 000). The field defined by the country rocks Ordovician turbi dites (open squares) is also shaded. The single array defined by the metaluminous (hornblende granite; filled octagons) and strongly peraluminous (cordierite granite; open octagons) gran ites has been taken to imply that the two groups formed by differ ent degrees of mixing between mantle derived magma and a continental crust like end member, approximated by the host sediments. Mixing proportions are indicated as crosses on the calculated mixing curve. However, the mismatch between the measured bulk rock composition of the granite and that predicted by the isotopic mixing proportions suggests the operation of other processes, which are difficult to resolve with bulk rock data.

although the viscosity contrast may be less than originally thought (10^4–10^6 Pa s for hydrous crystal-poor granite versus 10^2 Pa s for a typical basalt). The thermal constraints are satisfied by models in which granitic magmas are generated by the wholesale assimilation of crustal materials by thick convecting basaltic sills in thermally perturbed lower crust, or by the combination of felsic liquids derived from the crystallization of such sills with partial melts of the overlying crust. Compelling field evidence for such scenarios can be directly observed in deep-crustal exposures upthrust in continental-collisional zones, such as the Ivrea Zone of the Italian Alps and in the Norwegian Caledonides.

A promising new way of unravelling the various source components of granitic magmas involves decoding the isotopic information preserved in the fine-scale growth zoning of certain minerals (Figure 10). This is done by *in situ* microanalysis, using laser-ablation sampling or secondary ion mass spectrometry, both of which are capable of measuring precise isotope ratios and trace-element contents at sub-ppm levels in very small areas of the crystal (20 μm or less). Zircon is an ideal mineral for this approach since the growth zones can be directly dated by uranium–lead isotopes. Zircon is also robust and may survive as a refractory residue in granitic magmas (*see* **Minerals:** Zircons), thereby preserving age and compositional information about the granite's source rocks and the otherwise inaccessible deep crust. Hafnium is a particularly abundant trace element in zircon, and fluctuations in hafnium isotope composition within zircon crystals record the progress of processes such as magma mixing and crustal assimilation. In this way, the crystallization history of silicic magmas can be reconstructed as they ascend from their deep crustal source to their final emplacement site.

The Time-Scales of Granitic Magmatism

The time-scales of geological processes provide important clues about the formation of granitic magmas and the various factors that influence their compositional evolution. For silicic magmas, this area of study is still in its relative infancy. However, there is growing evidence that the segregation, ascent, and emplacement of granitic magmas occur on time-scales of hundreds or thousands, rather than millions, of years (Figure 11). When coupled with the crystallization rates, this places constraints on the sites and nature of magmatic differentiation and on partial-melting processes in the source region.

Information on the time-scales of melt segregation and extraction in the granite's source region comes from the accessory minerals zircon and monazite,

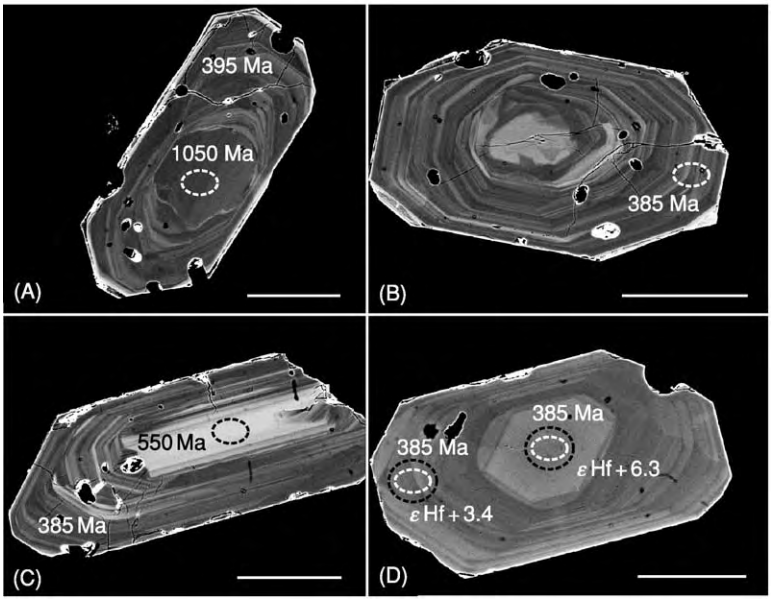


Figure 10 Cross sections of zircon crystals imaged by back scattered electron microscopy (scale bar is 100 μm) to show internal growth structures. Circles in (A) (C) represent ion probe pits, and the circles in (D) represent laser ablation pits. (A) Zircon with a 390 Ma igneous crystallization age showing corroded, partially resorbed older core dated at 1050 Ma. (B) Small, irregularly shaped zircon nucleus (too small to analyse) enclosed by strongly zoned magmatic zircon. (C) Elongate 550 Ma core surrounded by 385 Ma igneous zircon. Note the cracks radiating from the older zircon core. (D) Magmatic zircon analysed by laser ablation, showing strong core to rim zoning in hafnium isotopic composition, expressed in the standard epsilon notation (see caption to Figure 9). This reflects the operation of an open system process, such as crustal assimilation, during crystallization.

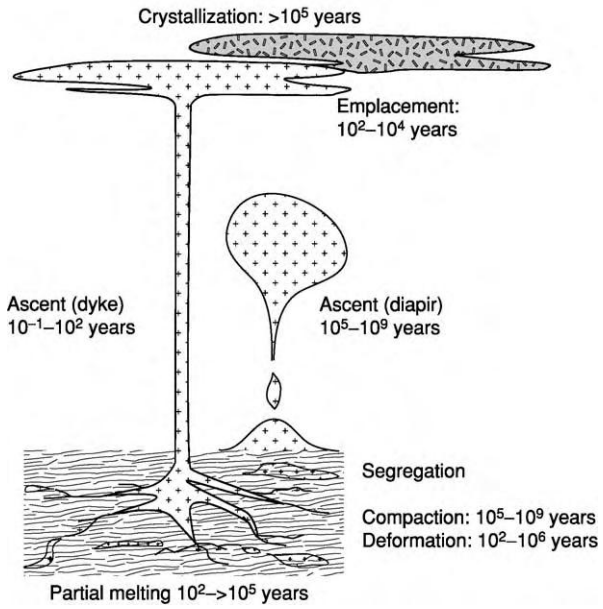


Figure 11 Summary of the time scales estimated for the various stages in the formation of a granitic pluton.

which (together with some lithophile elements) dominate the trace-element budget of peraluminous granites. Many peraluminous leucogranites contain lower concentrations of zirconium and rare earth elements than are predicted by equations describing the solubility of the host minerals (i.e. zircon and monazite) at the temperature of partial melting. Assuming that the granitic magmas are unmodified by fractional crystallization, this is attributed to the rate of melt segregation exceeding that at which equilibrium can be maintained with dissolving zircon and/or monazite crystals in the source. As zircon dissolution kinetics can be numerically modelled, it is possible to calculate the residence time of the partial melt in the source region from the zirconium content of the melt or the degree of zirconium undersaturation. Remarkably, such calculations reveal that the Himalayan leucogranites, some of which exceed 3000 km^3, were extracted from their sources in less than 150 years. It remains to be established whether these extraction rates are applicable to granitic magmas formed from different source rocks in other tectonic settings.

Evidence concerning the rates of ascent and emplacement of granitic magmas stems from the commonly observed transport of granitic magma in dykes. As indicated above, to avoid freezing, dykes must maintain a critical width and flow rate, both of which are determined by the temperature and viscosity of the magma. In view of this, fast melt transport rates are predicted, leading to rapid filling of large high-level granitic plutons. For example, in the Himalayas, leucogranitic magmas are thought to have been transported vertically through about 10 km of crust. Using an experimentally determined melt viscosity, a critical dyke width of 3 m permits an average melt ascent velocity of 9.5 cm s^{-1}, a factor of up to 10^8 faster than magma transport modelled by diapiric ascent. At these magma delivery rates the largest laccoliths in the Himalayas could have been emplaced in about 10 years, though this is an extreme case. Rapid rates of melt transport and emplacement are also inferred for granitic plutons in subduction-related settings. For the large Cordillera Blanca batholith of the Andes, a melt ascent rate of 1 cm s^{-1} and a dyke width of 6 m lead to an emplacement time of 350 years. Peak pluton-filling times of 1000 years are modelled for the plutons of the Sierra Nevada batholith.

The time-scales of crystallization of felsic magmas are constrained by the study of radiogenic isotopes in recent high-silica volcanic rocks. These are the eruptive analogues of granitic plutons, and, in some cases, such as the voluminous (*ca.* 5000 km^3) Fish Canyon Tuff (Colorado), are thought to form by the rejuvenation of a partially crystalline granitic magma chamber. For the peralkaline rhyolites of the East African Rift, precise rubidium–strontium and uranium–thorium dating reveals that the crystals and their host fractionated liquids formed about 15 000 years before eruption. Similarly, rubidium–strontium mineral dating establishes that most feldspar in the Long Valley rhyolites (California) crystallized in less than 15 000 years. In contrast, a much longer crystallization time (more than 200 000 years) for zircons in the same rocks was inferred using uranium–thorium isotope disequilibria. Zircons in pumices from the Taupo Volcanic Zone (New Zealand) also crystallized over a period of 300 000 years prior to eruption, suggesting a substantial residence time in the magma chamber.

Although it is unclear to what extent these studies can be extrapolated to plutonic rocks, they serve to highlight the fact that the rates of crystallization of silicic magmas are potentially far out-paced by the rates of magma transport and emplacement. The implication for granitic magmas is that crystal–liquid differentiation, and thus crustal assimilation, is likely to operate in high-level magma chambers rather than in the deep crust.

A final point is that, if dyke transport is a valid description of silicic magma ascent in general, the rate-determining step in the construction of granitic batholiths is probably the melt generation in the source. The sheeted structure and internal isotopic heterogeneity characteristic of many plutons suggest an episodic or pulsed, rather than continuous, transport of magma to the emplacement site. If the

magma supply is produced by crustal fusion, then this in turn is limited by the prograde heating rate. Where felsic melt is produced from mantle-derived precursors, the rate of melt production depends on the intrusion rate of these into the crust and on the timescales of differentiation. Constraining the rate of silicic magma generation by both of these processes in nature has thus far proved elusive and is likely to be a focus of future studies.

See Also

Earth: Crust. **Europe:** Scandinavian Caledonides (with Greenland). **Igneous Processes**. **Minerals:** Feldspars; Quartz; Zircons. **Mining Geology:** Magmatic Ores. **North America:** Northern Appalachians. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Bouchez JL, Hutton DHW, and Stephens WE (eds.) (1997) *Granite: From Segregation of Melt to Emplacement Fabrics*. Dordrecht: Kluwer.

Brown M (1994) The generation, segregation, ascent and emplacement of granite magma: the migmatite to crustally derived granite connection in thickened orogens. *Earth Science Reviews* 36: 83 130.

Frost BR, Barnes CG, Collins WJ, *et al.* (2001) A geochemical classification for granitic rocks. *Journal of Petrology* 42: 2003 2048.

Johannes W and Holtz F (1996) *Petrogenesis and Experimental Petrology of Granitic Rocks*. Berlin: Springer Verlag.

Kemp AIS and Hawkesworth CJ (2003) Granitic perspectives on the generation and secular evolution of the continental crust. In: Rudnick R (ed.) *Treatise of Geochemistry, Volume 12. The Crust,* pp. 349 410. Oxford: Elsevier.

Patiño Douce AE (1999) What do experiments tell us about the relative contribution of crust and mantle to the origin of granite magmas? In: Castro A, Fernandez C, and Vigneresse JL (eds.) *Understanding Granites: Integrating New and Classical Techniques,* pp. 55 76. Special Publication 168. London: Geological Society.

Petford N, Cruden AR, McCaffrey KJW, and Vigneresse J L (2002) Granite magma formation, transport and emplacement in the Earth's crust. *Nature* 408: 669 673.

Pitcher WS (1998) *The Nature and Origin of Granite,* 2nd edn. Glasgow: Blackie.

Thompson AB (1999) Some time space relationships for crustal melting and granitic intrusion at various depths. In: Castro A, Fernandez C, and Vigneresse JL (eds.) *Understanding Granites: Integrating New and Classical Techniques,* pp. 7 25. Special Publication 168. London: Geological Society.

Kimberlite

G J H McCall, Cirencester, Gloucester, UK

Introduction

The name kimberlite is derived from the name of the town in South Africa where diamonds were first mined in quantity after their discovery there in 1870. Coincidentally, the region of Western Australia where economic deposits of diamonds were discovered in the 1970s was named the 'Kimberleys' after the same Colonial Secretary in England.

Kimberlites are rare and volumetrically insignificant rocks, which are nonetheless studied exhaustively because they are the main source of diamonds (directly or from placer deposits derived from them).

Definition

The early definitions of kimberlite were so broad that they permitted extension to a wide variety of brecciated igneous rocks found in diatremes (explosion pipes) and carrying mantle-derived xenoliths and high-pressure macrocrysts, including mafic minettes (lamprophyres), olivine lamproites, nephelinites, and melilitites. Stricter definitions applied more recently have placed greater emphasis on groundmass composition and rock chemistry.

Kimberlites are petrographically complex in that they contain not only phases crystallized from the liquid magma but also polycrystalline fragments or crystals derived from various types of fragmental xenolith. The texture is inequigranular because of the xenoliths and macrocrysts, and the latter include both xenocrysts and high-pressure phenocrysts.

There have been innumerable definitions advanced in the literature. The AGI *Glossary of Geology* offers the following.

> Kimberlite: an ultramafic igneous rock containing at least 35% olivine with one or more of the following in the groundmass: monticellite, phlogopite, carbonate, serpentine, diopside. No leucite is allowed. Two types have been distinguished:
>
> Basaltic or type 1 kimberlite: the classic diatreme filling diamond bearing rocks of South Africa, much more widespread than type 2 micaceous kimberlite.

Micaceous or type 2 kimberlite: phlogopite is the main component and the smaller micas are phlogopite or tet raferriphlogopite: other groundmass minerals are diop side (mantled by titanian aegirine), spinels (chromite to magnetite), apatite (Sr rich), perovskite, carbonates and serpentine. There is no monticellite and olivine is com monly subordinate to phlogopite in abundance. These type 2 kimberlites were called orangeites by Wagner in 1928 and the name has recently been revived, though whether it is necessary is debatable. It is restricted to South Africa.

JB Dawson has emphasized that the presence of diamond in kimberlites is not essential, and indeed kimberlites that contain diamonds are in the minority.

In 1995 RH Mitchell used a definition that improved the textural and mineralogical description. He defined kimberlites as volatile-rich (dominantly carbon dioxide) potassic ultrabasic rocks, commonly exhibiting a distinctive inequigranular texture resulting from macrocrysts (or even megacrysts) set in a fine matrix and consisting of anhedral crystals of olivine, magnesian ilmenite, chromium-poor titaniam diopside, pyrope garnet (commonly subcalcic), phlogopite, enstatite, and titanium-poor chromite. Olivine macrocrysts are a characteristic constituent of all but fractionated kimberlites. The matrix contains a second generation of primary euhedral-to-subhedral olivine, which occurs together with one or more of the following primary minerals: monticellite, phlogopite, perovskite, spinel (magnesian-ulvo-/magnesiochromite-ulvo-, and magnetite solid solutions), apatite, and serpentine. Many kimberlites contain late-stage poikilitic micas belonging to the barian phlogopite–kinoshitalite series. Nickeliferous sulphides and rutile are common accessory minerals.

Early-formed olivine, phlogopite, monticellite, and apatite are commonly replaced by deuteric serpentine and carbonate.

Chemistry

Table 1 gives some typical major-element analyses of kimberlites. Rare earth elements are enriched in kimberlites relative to other ultrabasic rocks.

Types of Intrusion

Kimberlites occur as diatremes (pipes), dykes, and sills, and one ring-dyke occurrence is known in Sierra Leone. Small surface extrusions do occur, for example in Kasma, Mali (pipe surrounded by laterized tuffs 4 m thick) and Igwise, Tanzania (three small tuff rings and a lava flow). Thermal metamorphism of the country rocks may be present, as at the Premier Mine near Pretoria, South Africa, where shale and dolomite have been metamorphosed.

Diatremes

Diatremes tend to occur as clusters or scatters along elongate zones. They have small surface areas, varying from a few square metres to several hectares (some surface areas of diatremes are listed in Table 2). The original pipe at Kimberley covered only 4 ha, whereas the Mwadui occurrence in Tanzania covers 141 ha. Double, paired, or multiple intrusions are common in Yakutia (Figure 1), and these may coalesce at depth or represent quite different eruptive phases. The pipes have the form of simple cones, tapering downwards, but they vary in surface plan

Table 1 Some chemical analyses (major elements) of kimberlites. (Reproduced from Nixon PH (1980) The morphology and mineralogy of diamond pipes. In: Glover JE and Groves DI (eds.) *Kimberlites and Diamonds*, pp. 32 47. Nedlands: University of Western Australia.)

	Lesotho kimberlite	*South African kimberlite*	*Yakutian kimberlite*	*Indian kimberlite*	*Riley Co. USA*	*Ultrabasic (peridotite)*
SiO_2	33.21	35.2	31.1	36.09	24.15	40.6
TiO_2	1.97	2.32	2.03	5.10	1.50	0.05
Al_2O_3	4.45	4.4	4.9	3.81	2.03	0.85
Fe_2O_3	6.78	9.8	10.5	4.77	6.51	12.6
FeO	3.43	9.8	10.5	4.45	1.85	12.6
MnO	0.17	0.11	0.10	0.15	0 (not detected)	0.19
MgO	22.78	27.9	23.9	24.71	24.45	42.9
CaO	9.36	7.6	10.6	3.50	15.87	1.0
Na_2O	0.19	0.32	0.31	0.20	0.30	0.77
K_2O	0.79	0.98	2.1	1.14	0.30	0.04
P_2O_5	0.65	0.72	0.66	2.31	0.62	0.04
CO_2	4.58	3.3	7.1	0.72	12.04	0.04
H_2O^+	8.04	7.4	5.9	13.05	8.50	
H_2O	2.66					0.84
No. of samples	25	80	623	12		

Table 2 Surface areas of some larger kimberlite diatremes. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths*. Berlin, Heidelberg, New York: Springer Verlag.)

	Hectares
Southern Africa	
Premier	31.5
Kimberley	4.1
Bultfontein	9.5
Dutoitspan	12.3
AK/1 (Orapa), Botswana	112.3
Kao (Lesotho)	17.4
Zaire	
Taiala	50.0
Kambeli	26.0
M'Bo	23.5
Tanzania	
Mwadui	141.4
Russia	
Mir	13
Zarnitsa	28
Udachinaya	52

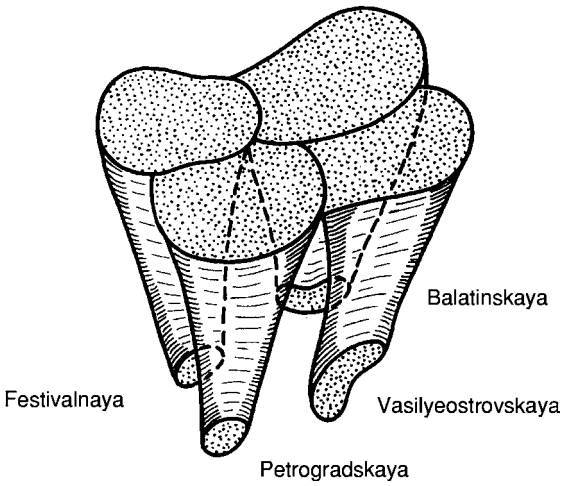


Figure 1 Structure of a group of diatremes in the Ukukinskoya Field, Yakutia (Reproduced from Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks*. New York: Plenum Press.)

(Figure 2). The generalized model of a kimberlite shown in Figure 3 takes the form of a pipe open to the surface and represented there by a tuff ring, but some pipes are blind. This model shows how different diatremes may be preserved eroded to different levels.

Dykes

Dykes occur both singly and in swarms of parallel dykes and may be multiple, representing several phases of injection of magma. They are characteristically narrow (not more than 1–2 m), but may be persistent – one over a metre wide south-west of Kimberley persists for 30 km. They may show enlargements or

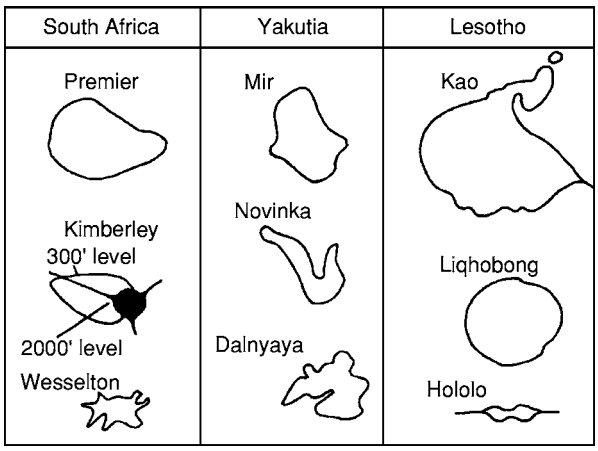


Figure 2 Some plans of kimberlite diatremes. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths*. Berlin, Heidelberg, New York: Springer Verlag.)

pods that connect with the surface, but these are not diatremes, being lenticular in section and plan: they do not contain brecciated material or fragments of overlying formations such as those typically found to have descended into a diatreme.

Sills

Sills of kimberlite are much rarer than diatremes or dykes, probably because the present erosion level must coincide closely with the injection level of these bodies, which is restricted. However, they are known to occur in South Africa, Zimbabwe, Tanzania, and Greenland.

Diatreme–Dyke–Sill Relationships

Some diatremes have been excavated sufficiently to show that they contract downwards into dykes or groups of dykes. The zones of diatremes appear to be a surface expression. The dykes occur near the roots of diatremes, which rise to considerable heights above this zone. These dykes have been referred to as antecedent dykes, but the relationship is complicated, because the diatremes flare well above the dykes, intrude the dykes, and contain different types of kimberlite. Dykes may branch into sills.

The complexity of the relationships between the various bodies is shown in Figures 4 and 5, which show a diatreme in Yakutia, and Figure 6, which depicts an idealized system.

Geotectonic Settings

Most kimberlites are found on stable cratons that have not been deformed since the Precambrian (Figure 7). Such areas have experienced epeirogenic

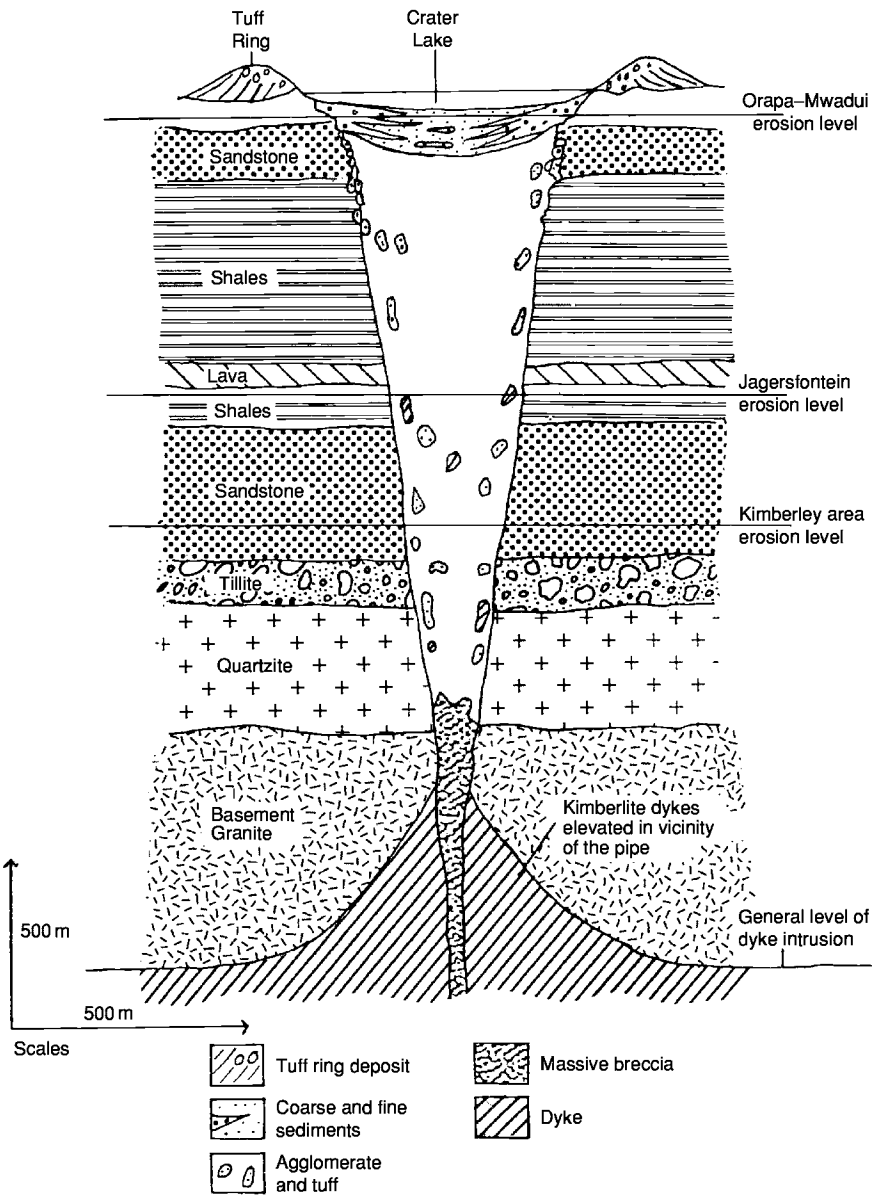


Figure 3 Generalized model of a kimberlite diatreme and its sub diatreme dykes. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.)

flexuring prior to the intrusion. They may be associated with local dome or swell structures. They occur within fundamental fractures that transect the deepest known basement terrains. Although they also occur in circumcratonic fold belts, they were not intruded there during orogenic cycles. They typically occur both on cratons and in the marginal fold belts where magmatic activity was otherwise very limited. Where the intrusions are underlain by circumcratonic orogenic belts with limited contemporaneous magmatic activity, they are characteristically non-diamondiferous.

Age Relationships

Kimberlite developments in the various shield areas or provinces show various ages. For example, the Premier diatreme near Pretoria is transected by a basic sill that is confidently dated at 1115 Ma, which means that the kimberlite is Precambrian and older than this intrusion into it. A date in the Permian has been obtained for the Dokolwayo occurrence in Swaziland on the eastern edge of the Kaapvaal Craton, although the majority of the South African occurrences can be dated radiometrically as

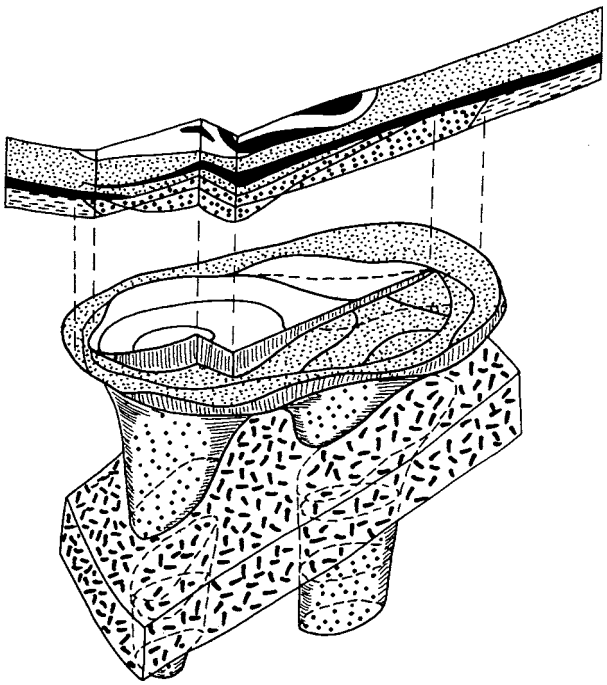


Figure 4 Isometric diagram of the double Palaeozoic age Krasnopresneskaya pipe, Yakutia; the pipe has been intruded by a basaltic sill (dashed), and, at the top, bowl shaped depressions are filled with younger volcaniclastic sedimentary rocks and covered by basalt flows. (Reproduced from Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks.* New York: Plenum Press.)

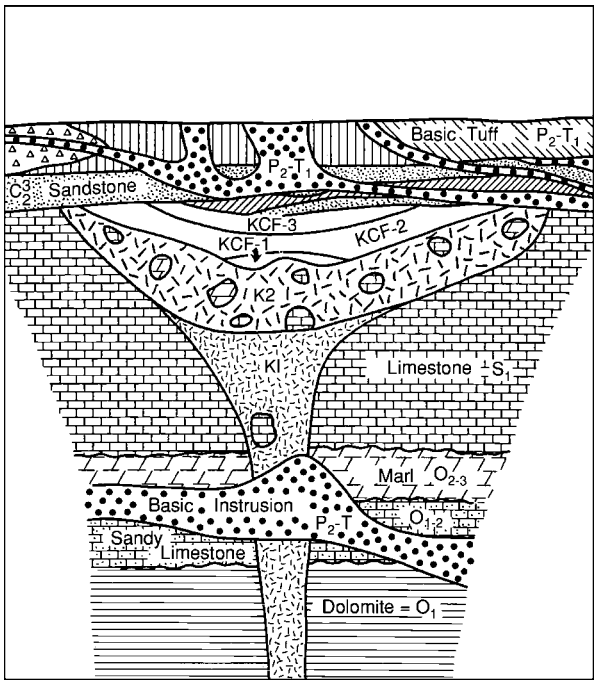


Figure 5 Cross section of the diatreme shown in Figure 4. K1 and K2, kimberlite breccias; O, Ordovician country rocks; S, Silurian country rocks; C, later Carboniferous rocks; P T, later Permo Triassic rocks. (Reproduced from Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks.* New York: Plenum Press.)

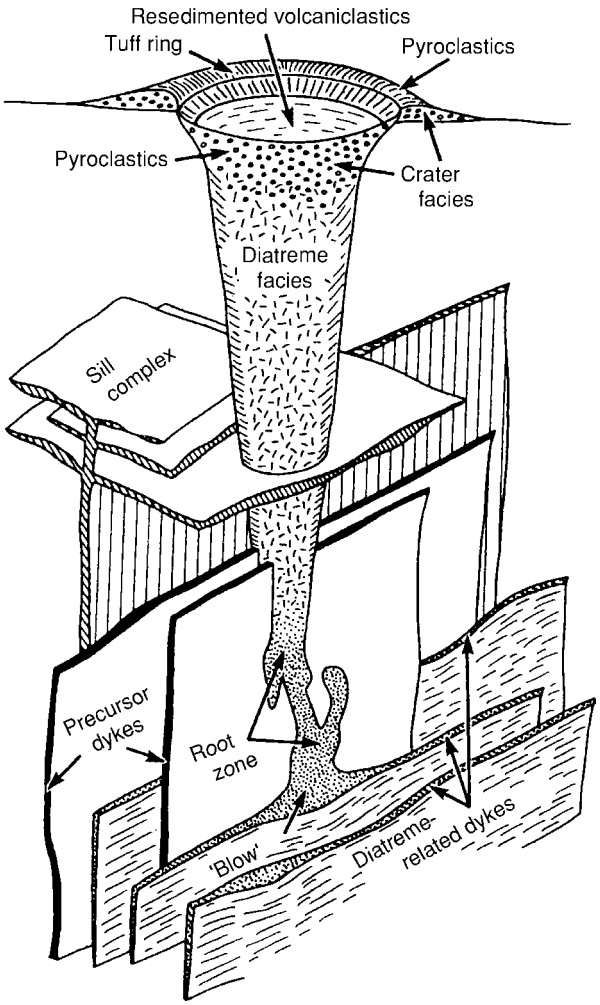


Figure 6 An idealized kimberlite magmatic system (not to scale). Hypabyssal rocks include sills, dykes, root zone, and a 'blow' (Reproduced from Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks.* New York: Plenum Press.)

post-Jurassic (Figure 8). In West Africa there are four ages of intrusions into the Eburnian Craton: 2100–2300 Ma, 1150 Ma, 700 Ma, and 60–100 Ma. The Mwadui occurrence, the largest kimberlite diatreme known, is post-Precambrian and has not been substantially eroded: the original crater-like sediments at the top are preserved. The Igwise occurrences are recent. Two distinct ages of 1140 Ma and 840 Ma have been obtained for diatremes in India; the Siberian kimberlites are of at least four ages: Precambian, Palaeozoic, and both Early and Late Mesozoic (Figure 9). Whereas single diatremes or diatreme complexes represent intrusions of a single suite over a short period, repeated intrusions far apart in time within a single shield may use the same fracture systems.

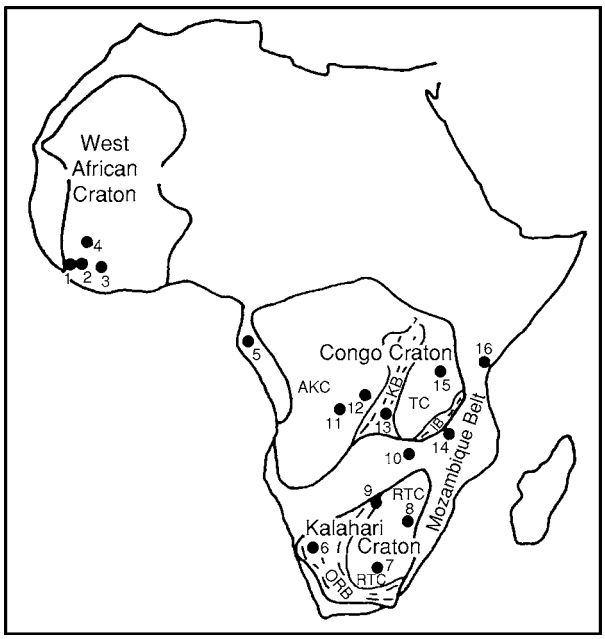


Figure 7 The relation of kimberlite occurrences in Africa to ancient cratons and their boundaries. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.)

Kimberlite magmatism has occurred at various times in the Earth's history, one major phase being in the Cretaceous. South and West Africa have seen three or four periods of kimberlite magmatism, and there have been several epochs of such magmatism in the central USA and Yakutia. Certain geographical areas coinciding with old cratons are 'kimberlite prone'.

Xenoliths

Within kimberlites a wide variety of rock types occur as fragments. Minerals derived from these rocks also become incorporated within the host kimberlite. Thereby, the mineralogy becomes extremely complex. There are five sources of xenoliths:

- fragments of the immediate wall rocks at the level of preservation;
- fragments derived from earlier formations that existed at the time of intrusion but have since been removed by erosion;
- blocks of still-buried formations through which the intrusion passed;

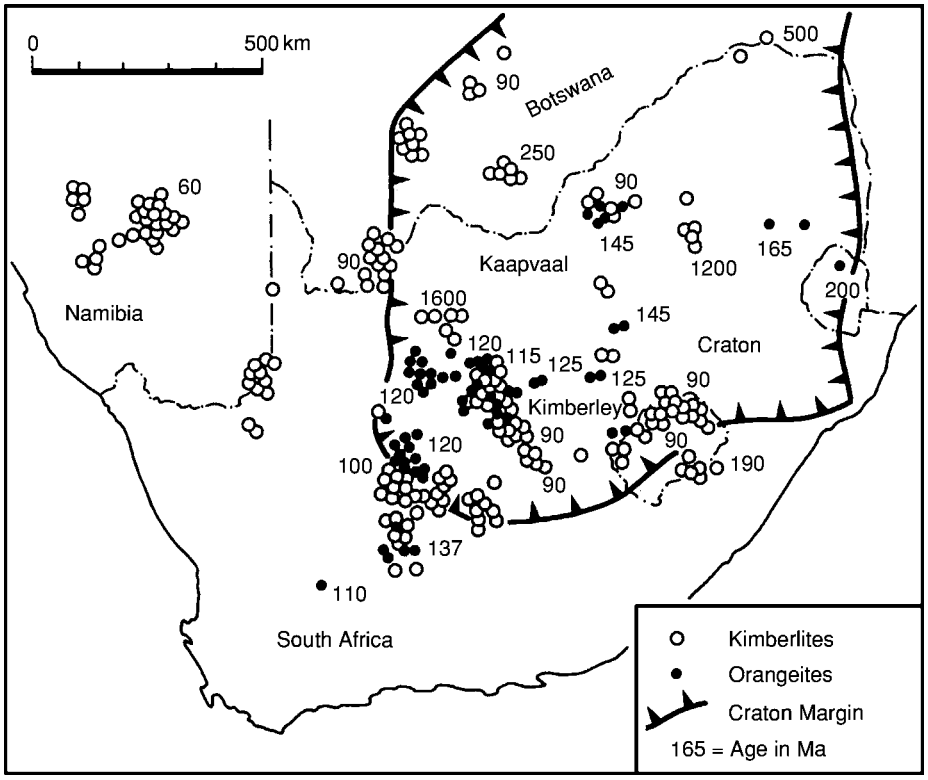


Figure 8 Kimberlite and 'orangeite' occurrences in southern Africa, with their radiometrically derived ages. Note the Precambrian age (1200 Ma) at the site of the Premier Mine and the single Permian (200 Ma) age in Swaziland. (Reproduced from Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks.* New York: Plenum Press.)

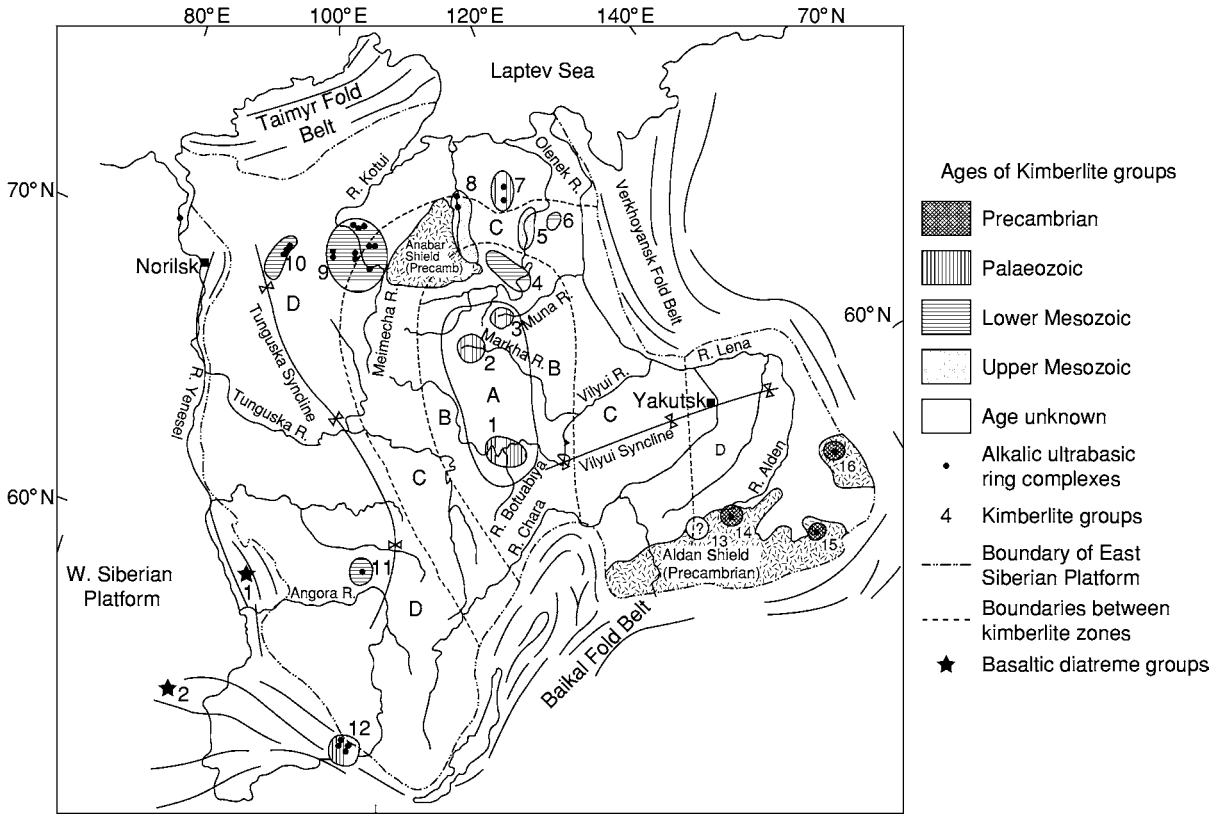


Figure 9 Distribution of kimberlites in the East Siberian Platform, showing the different ages. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.)

- granulites derived from deep metamorphic terrains (in the crust); and
- fragments believed to come from the Earth's upper mantle (containing high-pressure mineral phases).

Examples of xenoliths derived from earlier formations are the Karoo basalt xenoliths found at Kimberley, 350 km from their nearest outcrop: they are estimated to have descended 1500 m within the eroded-off top of the intact diatreme before it was mined. In Yakutia, blocks derived from still-buried formations may include fragments of the ancient Precambrian shield (the crystalline Siberian Platform).

Xenoliths of rock types with a mineralogy and density indicating higher pressures and temperatures than those normally pertaining within the Earth's crust are believed, by inference, to emanate from the mantle. They not only provide the most important evidence of the components of the mantle but also provide an insight into the processes that generated the kimberlite diatreme, which originated at great depths. The have been exhaustively studied, and state-of-the-art methods of investigating them are continually being developed, in particular those based on isotopes. Such methods are being extended to single mineral grains and inclusions within diamonds.

The deep-seated xenoliths may be divided into five main groups: peridotite–pyroxene suite; eclogites and grospydites; metamorphosed peridotites (rich in amphibole or mica); glimmerites and the MARID suite; and miscellaneous. Together, these xenoliths usually do not amount to more than 2% of the bulk rock, but in Lesotho they may form 20–30% (**Figure 10**). There is much variation from diatreme to diatreme: in some the peridotite suite predominates, whereas in others eclogite is dominant.

The peridotite–pyroxene suite, of ovoid to discoidal xenoliths up to 30 cm in maximum dimension, includes various types distinguished on the basis of modal mineral composition – dunite, lherzolite, harzburgite, websterite, and pyroxenite. The Russians in Yakutia have noted that the spinel in these rocks is magnetite, not the chromite that is characteristic of peridotites not occurring as xenoliths in kimberlite. The origin of the rounding is obscure. Most xenoliths are structureless, but a few are banded. Globules of

Figure 10 Kimberlite choked with mantle derived xenoliths; Matsoko pipe, Lesotho. (Reproduced from Nixon PH (ed.) (1973) *Lesotho Kimberlites.* Lesotho National Development Corporation.)

sulphide and oxides of ore minerals may be present. Some garnet lherzolites contain high concentrations of rare earth elements with the lighter ones being enriched, and there is some evidence that they were originally cumulates.

The eclogites are coarse-grained rocks composed of pyroxene and garnet, with many accessory minerals (Figure 11), and they are commonly banded. Again, rounded xenoliths are common. The peraluminous grospydites are eclogites containing grossular garnet, kyanite and/or corundum, and rarely sillimanite; there are also quartz eclogites, amphibole eclogites, and orthopyroxene eclogites. Diamond-bearing eclogites are known from at least nine intrusions in South Africa, the AK/1 pipe in Botswana, and the Mir pipe, Yakutia; the diamonds are mostly high-temperature octahedra (though some are rhombic dodecahedra), and they may be accompanied by graphite.

Temperature ranges of equilibration of eclogite xenoliths, peridotite and eclogite inclusions in diamonds, and the megacryst suite are given in Figure 12.

Glimmerites are xenoliths dominated by phlogopite, which are found in Yakutia. The MARID-suite xenoliths, found in South Africa, are also dominated by phlogopite, with various amounts of amphibole, clinopyroxene, ilmenite, rutile, apatite, and olivine. These are thought to have crystallized high in the mantle under oxidizing conditions, but alternatively it has been suggested that they are of wall-rock metasomatic origin.

The miscellaneous xenoliths include alkremites (found in the Udachnaya pipe in the former Soviet

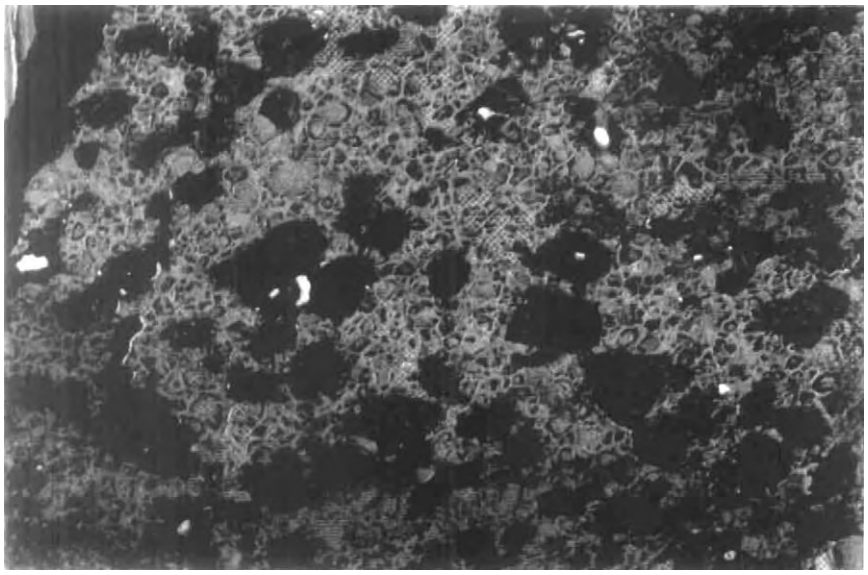


Figure 11 Eclogite nodule; Kao pipe, Lesotho. (Reproduced from Nixon PH (ed.) (1973) *Lesotho Kimberlites.* Lesotho National Development Corporation.)

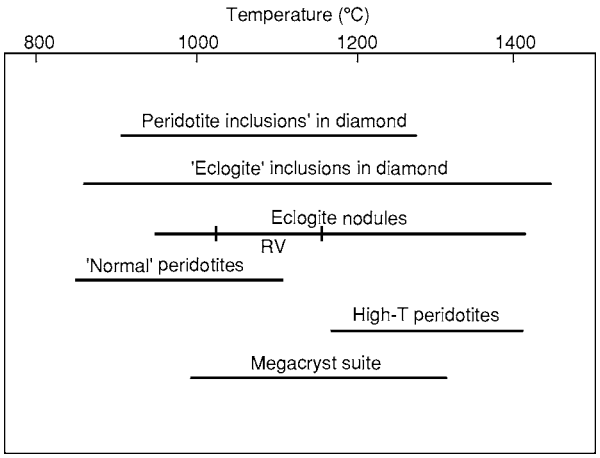


Figure 12 Temperature ranges of various inclusions and 'nodules', normal peridotites, and the megacrysts suite (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths*. Berlin, Heidelberg, New York: Springer Verlag.)

Union), which are coarse assemblages of garnet and dark green spinel. They are interpreted as early dense cumulates of an aluminous mantle under pressures equivalent to a depth of 70 km.

Megacrysts

Single large crystals (more than 2 cm) are common in kimberlites; they are alternatively referred to as discrete nodules. Two series are recognized: chrome-rich and chrome-poor. Such megacrysts may contain other minerals. There are also diopside, mica, and zircon megacrysts. It has been suggested that some megacrysts are genetically linked to the kimberlite, but such a relationship has been rejected by some authors, who consider the megacrysts to be accidental mantle-derived inclusions.

The minerals found in kimberlites and their origins are summarized in Table 3.

Studies of Diamonds and Their Inclusions

Nowadays many studies are carried out on diamonds in kimberlites and the inclusions within them, in an attempt to determine the pressure and temperature (i.e. mantle depth) at which they were formed. In a study of the diamonds from Pipe DO-27, Lac de Gras, Slave Craton, North-west Territories, Canada, a high proportion were found to be cubo-octahedral stones with resorption characteristics. Syngenetic inclusions are of three suites: 25% are peridotitic (garnet, clinopyroxene, sanidine, sulphide); 50% are eclogitic (garnet, olivine, orthopyroxene, sulphide); and 25% are super-deep (ferropericlase, magnesium-perovskite, nickel). The super-deep suite of minerals represents the mantle at depths greater than 670 km.

Upper Mantle Configuration

The Earth's upper mantle is petrologically complex. Xenoliths and xenocrysts from kimberlites provide a wealth of high-temperature and high-pressure rock types therefrom, but the problem is to delineate where in the mantle they originated and how they relate to each other genetically. Two main zones – the harzburgite zone and the garnet lherzolite zone – can be recognized, but these exhibit large- and small-scale heterogeneities, represented by pyroxenites, eclogites, and the MARID suite of rocks. These differences probably represent both lateral and vertical heterogeneities in the mantle. Shearing and metasomatism are also evident; brittle and plastic deformation accompany magma rise and open up channels. Convective processes may also operate. A tentative model for the upper mantle and crust derived from studies of kimberlite xenoliths and xenocrysts is shown in Figure 13. The super-deep inclusions in diamonds, mentioned above, represent a further zone below 670 km.

This field of research is on going, and each kimberlite conference produces new evidence and models, without necessarily clarifying the picture, such are the contradictions and differing interpretations of the evidence.

Diamond Provenance

There are several models for the origin of the diamonds in kimberlites. Since they are found within eclogites of established mantle provenance, it is reasonable to suppose that they come from the mantle, where high pressures and temperatures occur. Current models of diamond formation differ mainly in the source of the carbon. Juvenile methane or other hydrocarbons, oxidized during ascent through the upper mantle or at the lower boundary of the lithosphere, is favoured by one school. This is supported by the presence of peridotite inclusions in some diamonds. Other models invoke crystallization from kimberlite liquids or from ultrabasic melts during the formation of cratonic roots. Another set of models introduces the carbon to the mantle by subduction processes, the carbon being not juvenile and possibly biogenic in origin. This model is supported by the existence of diamonds containing inclusions of the eclogite suite. It seems likely, however, that several diamond-forming processes operate. Regardless of the model favoured, it is generally accepted that most diamonds are xenocrysts in the transporting kimberlite magma. Disruption and disaggregation of diamond-bearing horizons by the passage of magma

Table 3 A summary of the disaggregated minerals found in kimberlites, and their origins x, present; xx, fairly abundant; xxx, abundant. (Reproduced from Nixon PH (1980) The morphology and mineralogy of diamond pipes. In: Glover JE and Groves DI (eds.) *Kimberlites and Diamonds*, pp. 32 47. Nedlands: University of Western Australia.)

Earth zone (Host rock)	*Disaggregated resistant mineral suite*	*Composition (approximate values)*	*Relative abundance*	*Comment*
Upper Crust	Widely variable;		xxx	Scapolite, hornblende, apatite, etc. also present
Lower Crust	Garnet,	Py alm,	x	
(granulites,	Clinopyroxene,	Omphacite,	x	
etc.)	Orthopyroxene,	Hypersthene	x	
	Rutile		x	
Lithosphere	Olivine	Fo_{93}	xxx	Depleted mineralogy high Mg/Fe, Cr/Al and
(lherzolites	Orthopyroxene	En_{94}	xxx	low Ca and Na, although not as extreme as
etc.)	Clinopyroxene	Cr diopside	xx	in the diamond inclusion suite. Differentiation
	Garnet	Cr pyrope	xx	within the mantle locally extends chemical
	Spinel	Chromite	x	ranges. Metasomatism produces additional
	Mica	Phlogopite	xx	minerals similar to kimberlite suite below.
				Diamond and graphite recorded
Mantle	Garnet	Py alm gross	x	Rare suite but is locally abundant. Generally
differentiates	Clinopyroxene	Omphacite	x	much higher Fe/Mg ratios than in lherzolites
(eclogite and	Rutile, ilmenite		x	
grospydite)	Kyanite		x	
	Corundum		x	
	Diamond		x	
Kimberlite suite	Olivine	$Fo_{85–95}$	xxx ?	Many secondary minerals not listed but note
(including	Mica	Phlogopite	xxx	wide variety of fine grained spinels and
MARID	Amphibole		x	perovskite. Zircon present
'precipitates')	Rutile		x	
	Ilmenite	Mg bearing	xx	
	Clinopyroxene	Diopside	xx	
	Garnet	Py alm	x	
	Diamond		x	
Mantle residue	Olivine	$Fo_{93–94}$	x	Rare suite in concentrates. 'Ultra depleted'
at kimberlite	Orthopyroxene	$En_{92–96}$	x	extremely high Mg/Fe, Cr/Al and low Ca, Ti,
foci (diamond	Clinopyroxene	Cr diopside	x	and Na
inclusion	Garnet	Knorringite pyrope	x	
suite)	Spinel	Chromite	x	
Asthenosphere	Olivine	Fo_{90}	x	'Fertile' composition, i.e., relatively low Mg/Fe
(lherzolites,	Orthopyroxene	Calcic bronzite (En_{87})	x	and Cr/Al but higher Ti compared with
etc.)	Clinopyroxene	Subcalcic diopside	x	typical lithosphere
	Garnet	Pyrope	x	
Discrete	Orthopyroxene	Calcic bronzite (En_{87})	x(x)	Silicates have similar composition to those in
nodules	Clinopyroxene	Subcalcic diopside	xx	fertile nodules with a tendency to higher
(megacrysts)	Garnet	Pyrope (little alm)	xxx	Fe/Mg ratios. Ilmenite pyroxene lamellar
	Ilmenite	Mg bearing	xxx	intergrowths are part of this suite
	Zircon		x	

from great depths incorporates the diamonds as xenocrysts in the magma. Very hot magmas may completely resorb the diamonds. Only magmas rising from great depths will carry diamonds.

Major Provinces

The global distribution of major kimberlite provinces is shown in Figure 14. The recently discovered province of diamondiferous kimberlites in the Slave Craton of Canada, is shown. It should be noted that province number 9, in the north of Western Australia, is a lamproite province not a kimberlite province, although it is a rich source of diamonds. Southern Africa is overwhelmingly the site of the most numerous kimberlite developments, as shown in Figure 8.

Prospecting Methods

Because kimberlite pipes rarely have a surface expression, prospecting for diamonds is arduous and requires the application of specialized methods. This in turn requires specialized training. The methods include:

- separation of heavy minerals (coarse and fine fractions): red-brown pyrope (discrete nodule

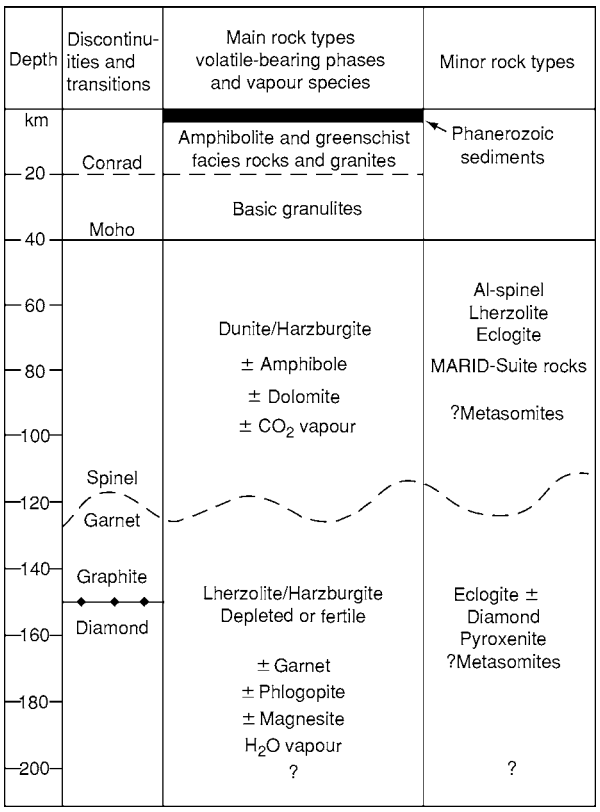


Figure 13 Tentative model for the depth distributions of the various rock types and the volatile bearing and vapour phases in the subcratonic lithosphere, derived from kimberlite occurrences. The depth of the Conrad discontinuity is based on seismic work on the South African craton. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.)

suite); purple-red chromium-pyrope (depleted lherzolite); ilmenite (coarse, rounded, with perovskite alteration or cream skin, distinct from the tabular variety); chromium-diopside (depleted lherzolite); coarse chromite; enstatite, and olivine grains;

- geochemistry (trace element suite);
- scintillometry;
- satellite imagery;
- air-borne photography; and
- geophysical surveys (magnetic, electromagnetic, induced polarization, resistivity, seismic, gravity).

Very few pipes carry diamonds in the kimberlite rock (Figure 15), and those that do may not contain economic deposits.

Mining

Mining of kimberlites may be opencast, as in the case of the Kimberley occurrence (Figure 16). Here, before it was consolidated, there were innumerable operators working small claims in the pit and transporting the ore up to the surface on a profusion of cable ways. Mining may also involve deep mining methods, as in the case of the Premier Mine, where a thick basic sill prevented opencast working below a certain depth. Here, the operation is large scale and standard-gauge electric trains move about underground, powered by overhead lines. This major operation in a huge pipe aims to recover diamonds that constitute a minute fraction of the rock, something like 1/100 000.

Lamproites

Lamproites are much less abundant than kimberlites, but like them most are the products of diatremes emanating from great depths. They were not considered important until the 1960s, when J B Jeppe, a South African, followed up a long-standing report of a single diamond find in the north of Western Australia and looked at the lamproites of the Fitzroy Region, earlier mapped by A Wade and petrologically described under a collection of rather unnecessary rock designations (cedricite, wolgidite, etc). Jeppe found a single diamond in one diatreme outcrop, but it was not until the early 1970s that a team from Tanganyika Holdings and London Tin, in a syndicate with four other companies, carried out extensive prospecting using well-tried methods and located the Argyle occurrences of the AK1 pipe and also alluvial deposits, both of which contained diamonds (Figure 17). Approximately one-quarter of the world's diamond production is now concentrated in this area (Figure 18). The pipes are of Miocene age (20 Ma), and there are more than 100 separate pipes, plugs, and sills, with some rare dykes.

Lamproites are potassic leucite-bearing rocks and include olivine lamproite, olivine–diopside lamproite, and leucite lamproites containing phlogopite, diopside, and postassic richterite. They range from ultrabasic (20–29% MgO) to basic (5% MgO). The lamproites are derived from depleted (garnet–clinopyroxene-poor) peridotite mantle beneath the Kimberleys region. This means that kimberlite indicator minerals are very scarce compared with regions where minerals are being shed from kimberlite sources, making exploration more difficult. Lamproites seem to occur on the margins of Archaean cratons. They have been found in Spain, Corsica, Zambia, West Africa, India, the USA (Figure 19), Gaussberg Island off Antarctica, and, recently, Brazil, where there are also kimberlites containing diamonds. There are two or three occurrences in South Africa, which have an atypical mineralogy and are possibly transitional to kimberlites. Diamonds were discovered in the Prairie Creek, Arkansas, occurrence in 1906, but it was misidentified as a kimberlite;

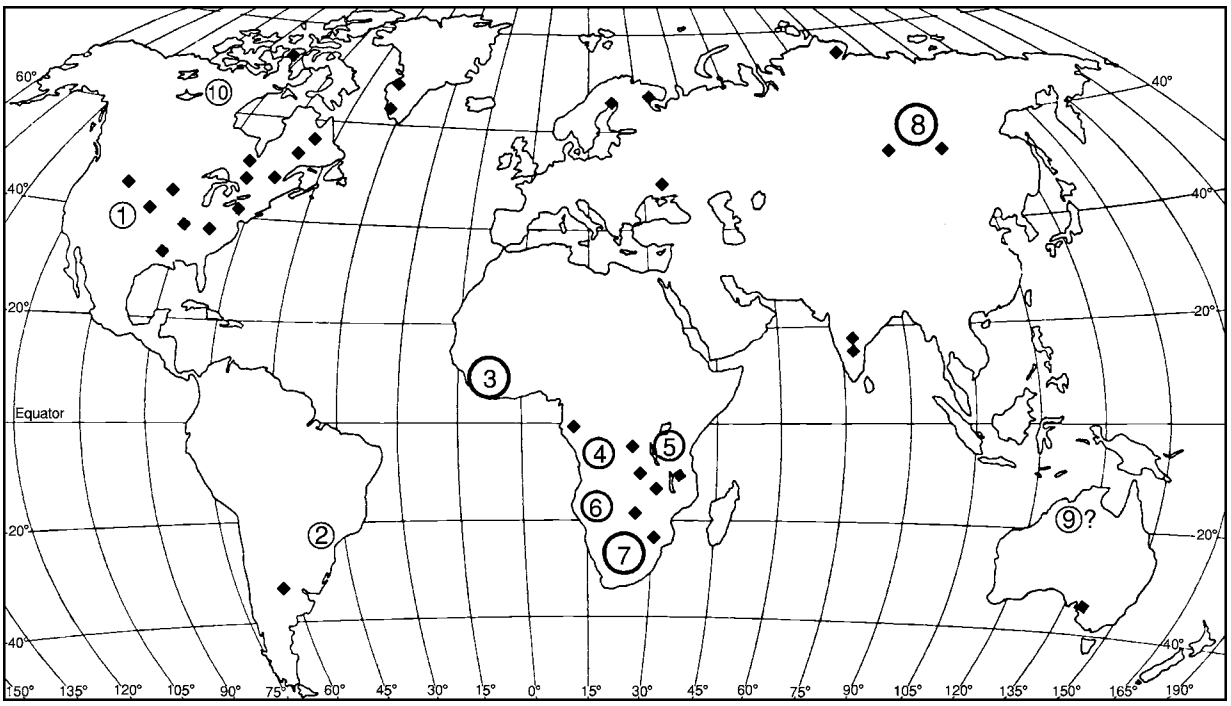


Figure 14 Global distribution of major kimberlite and diamond provinces (circles with numbers) and more isolated occurrences (black diamond symbols). Number 9 is a lamproite province, not a kimberlite province, and number 10 is a recent major development.

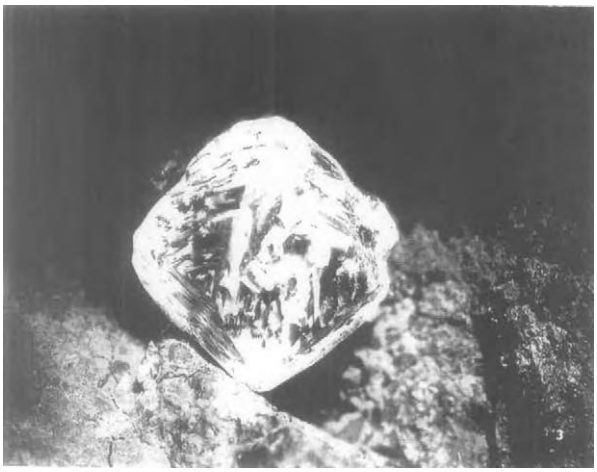


Figure 15 An uncut diamond in kimberlite, South Africa, 10×10 mm. (Reproduced from Bauer J and Bouska V (1983) *A Guide in Colour to Precious and Semi Precious Stones*, pp. 80 83. London: Octopus.)

the Mahjgawan, India, occurrence yielded diamonds even earlier, in 1827, but was also misidentified as a kimberlite.

Lamproites and the diamonds within them are derived from the mantle, but they do not appear to have much petrogenetic relationship to kimberlites, for

Figure 16 The Kimberley diatreme worked out. (Reproduced from Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.)

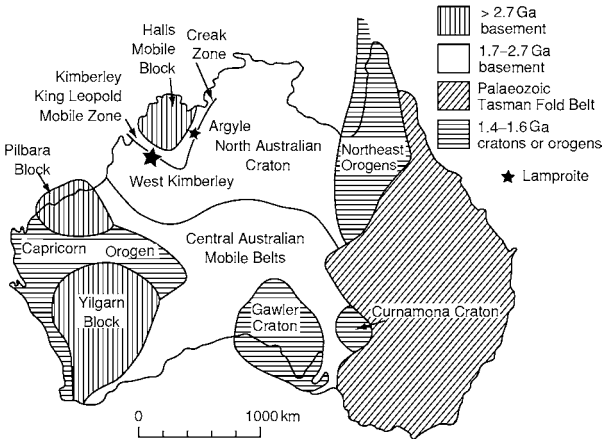


Figure 17 The Precambrian cratonic areas of Australia, showing the location of the Leopold Zone lamproite occurrences and the Argyle diamondiferous pipe. (Reproduced from Mitchell RH and Bergman SC (1991) *Petrology of Lamproites.* New York: Plenum Press.)

Figure 18 A 58 carat pink cut diamond worth $A38 000 (≃£18 000 sterling) from the Argyle Mine in the Kimberleys lamproite province, Western Australia (photograph supplied by the Western Australian Museum).

which they have been mistaken, although the South African occurrences do suggest the possibility of a connection.

Kimberlites and Carbonatites

It has been suggested, particularly in Russia, that there is a relationship between kimberlites and carbonatite intrusions and volcanic eruptives, but, although both are of deep-seated origin, carbonatites occur in and close to rifted geotectonic zones (for example East Africa), characteristically form ring complexes in the underworks of volcanoes (as in the huge eroded Kisingiri volcano in Kenya), and are associated with sodic rather than potassic alkaline igneous suites (intrusive ijolites, melteigites, and jacupirangites carrying nepheline and aegirine, also intrusive uncomopahgrites and turjaites carrying melilite and perovskite, and surface volcanics, nephelinites, and melilitites).

Figure 19 Boars Tusk, a lamproite neck in the Leucite Hills, Wyoming, USA. Some Western Australian lamproites form similar but less dramatic features. (Reproduced from Mitchell RH and Bergman SC (1991) *Petrology of Lamproites.* New York: Plenum Press.)

See Also

Earth: Mantle. **Gemstones**. **Geochemical Exploration**. **Igneous Processes**. **Igneous Rocks:** Carbonatites. **Shields**. **Volcanoes**.

Further Reading

Bauer J and Bouska V (1983) *A Guide in Colour to Precious and Semi Precious Stones,* pp. 80 83. London: Octopus.

Davies RH, Griffin WL, Pearson NL, *et al.* (1999) Diamonds from deep pipe DO 27, Slave Craton, Canada. In: Gurney JJ, *et al.* (eds.) *The J B Dawson Volume: the P H Nixon Volume,* Proceedings of the 7th International Kimberlite Conference April 1998, Cape Town, pp. 148 155. Red Roof Design.

Dawson JB (1980) *Kimberlites and Their Xenoliths.* Berlin, Heidelberg, New York: Springer Verlag.

Glover JE and Groves DI (eds.) (1980) *Kimberlites and Diamonds.* Nedlands: University of Western Australia, Geology Department and Extension Service.

Glover JE and Harris PG (eds.) (1985) *Kimberlite Occurrence and Origin: A Basis for Conceptual Models in*

Exploration. Nedlands: University of Western Australia, Geology Department and Extension Service.
Gurney JJ, *et al.* (eds.) (1999) *The J B Dawson Volume: the P H Nixon Volume*. Proceedings of the 7th International Kimberlite Conference April 1998, Cape Town. Red Roof Design.
Janse AJA (1995) A history of diamond sources in Africa: part I. *Gems and Gemmology* 31: 228 255.
Janse AJA (1998) A history of diamond sources in Africa: part II. *Gems and Gemmology* 32: 2 30.
Jaques AL, Lewis JD, and Smith CB (1986) The kimberlites and lamproites from Western Australia. *Geological Survey of Western Australia Bulletin* 132: 1 268.
Kornprobst T (ed.) (1984) *Kimberlites and Related Rocks*. Proceedings of the 3rd International Kimberlite Conference. Amsterdam, Oxford, New York, Tokyo: Elsevier.
Mitchell RH (1995) *Kimberlites, Orangeites and Related Rocks*. New York: Plenum Press.
Mitchell RH and Bergman SC (1991) *Petrology of Lamproites*. New York: Plenum Press.
Nixon PH (ed.) (1973) *Lesotho Kimberlites*. Lesotho National Development Corporation.
Nixon PH (1980) The morphology and mineralogy of diamond pipes. In: Glover JE and Groves DI (eds.) *Kimberlites and Diamonds*, pp. 32 47. Nedlands: University of Western Australia, Geology Department and Extension Service.
Nixon PH (1980) Regional diamond exploration: theory and practice. In: Glover JE and Groves DI (eds.) *Kimberlites and Diamonds*, pp. 64 75. Nedlands: University of Western Australia, Geology Department and Extension Service.
Sobolev NV (1977) *Deep Seated Inclusions in Kimberlite and the Problem of the Upper Mantle*. Washington: American Geophysical Union.

Komatiite

N T Arndt, LCEA, Grenoble, France
C M Lesher, Laurentian University, ON, Canada

Introduction

It is easy to explain roughly what a komatiite is but difficult to give a rigorous definition. The simple definition, as given by Arndt and Nisbet in 1982; is that komatiite is an ultramafic volcanic rock. A limit of 18% MgO separates komatiites from less magnesian volcanic rocks such as picrites, ankaramites, or magnesian basalts. The term komatiitic basalt is applied to volcanic rocks containing less than 18% MgO that can be linked, using petrological, textural, or geochemical arguments, to komatiites. The definition seems to include intrusive rocks; this was probably not intended.

Implicit to the definition of komatiite is the notion – difficult to prove – that komatiites crystallize from liquids that contained more than about 18% MgO. Complications arise from the existence of other MgO-rich volcanic rocks that either formed through the accumulation of olivine from less magnesian liquids, or crystallized from magmas with geochemical characteristics quite unlike those of most komatiites. An example of the first type is a phenocryst-charged basaltic liquid (a picrite according to some definitions); an example of the second is meimechite, a rare alkaline ultramafic lava (*see* **Lava**) with unusual major and trace element composition.

To distinguish komatiite from other types of highly magnesian volcanic rock, it is useful to include spinifex texture in the definition (**Figure 1A, B**). Spinifex, a texture characterized by the presence of large skeletal or dendritic crystals of olivine or pyroxene, is present in many, but not all komatiite flows. A workable definition of komatiite should include the phrase "komatiite is an ultramafic volcanic rock containing spinifex or related to lavas containing this texture". With the last part of the definition, we can make allowance for the manner in which texture varies within komatiitic units. For example, many komatiite flows have an upper spinifex-textured layer and a lower olivine-cumulate layer (**Figure 1C**); and other flows grade along strike from layered spinifex-textured portions to massive olivine-phyric units. With the inclusion of the phrase about spinifex, the lower olivine-cumulate portions of layered flows or the olivine-phyric units can also be described as komatiite. On the other hand, meimechites, picrites, and other rock types that contain no spinifex are excluded. For further discussion, see the **Further Reading** section at the end of this article.

Because of the facility with which olivine fractionates or accumulates in low-viscosity ultrabasic liquids, the compositions of komatiite lavas vary considerably. MgO contents range from 18% (the limit between komatiite and komatiitic basalt), to as high as 50% in the lower cumulate portions of layered flows (**Figure 2**). Elements that are immobile during the metamorphism and hydrothermal alteration, which affects all komatiites to a greater or lesser extent,

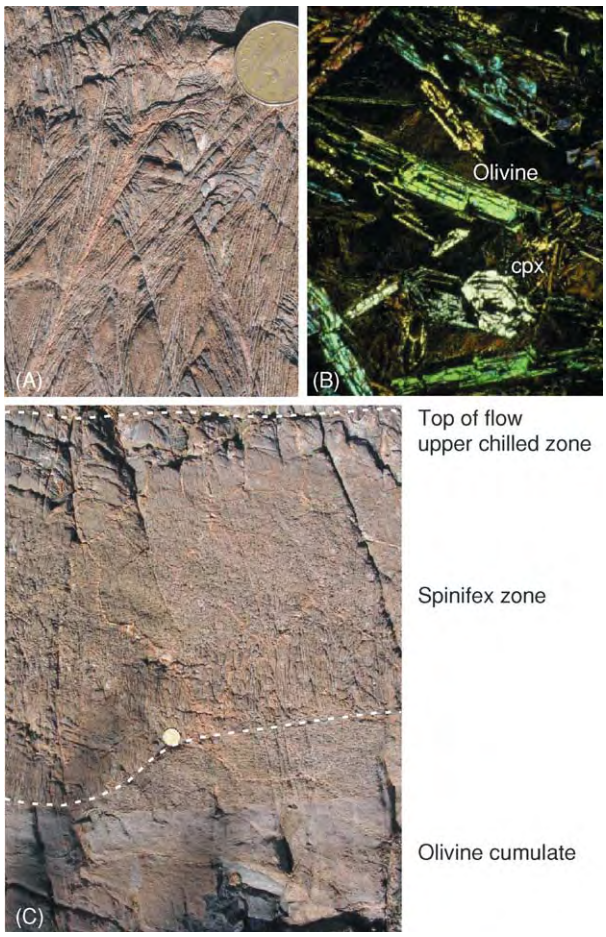


Figure 1 (A) Spinifex textured komatiite in outcrop (sample from Munro Township, Canada). The bladed habit of large paral lel olivine crystals is clearly visible. (B) Spinifex textured koma tiite in thin section. Sample from the Belingwe Belt, Zimbabwe. Skeletal olivine crystals lie in a matrix of clinopyroxene and altered glass. (C) The upper spinifex textured portion and the upper part of the olivine cumulate of a layered komatiite flow from Munro Township. The curved contact between the spinifex and cumulate zones is unusual; in most flows this contact is horizontal.

plot on olivine-control lines in variation diagrams. The least magnesian compositions are found in the lower parts of spinifex zones, the most magnesian in the olivine cumulates. Chilled flow tops and porphyritic lavas have intermediate compositions.

The maximum MgO contents of komatiite liquids, estimated using the compositions of chilled flow margins and the forsterite contents of olivine, are between 28 and 30%. The dry 1-atmosphere liquidus temperature of these liquids, calculated from experimental data, is between 1560 and about 1600°C.

Several different geochemical types can be distinguished using Al_2O_3/TiO_2 and rare earth elements (REE). The komatiites first discovered by Viljoen and Viljoen in 1969 are of the 'Al-depleted' or Barberton type. As shown in Figure 3, these rocks have relatively low Al_2O_3/TiO_2 and Gd/Yb, the latter ratio being a measure of relative depletion of the heavy rare Earth elements (HREE). The second group of komatiites, called 'Al-undepleted' or Munro-type, has near chondritic ratios of Al_2O_3/TiO_2 and Gd/Yb. A third type, Al-enriched komatiite, is common in komatiites from Gorgona Island, Colombia, described by Echeverria in 1980, and a fourth type, Ti-rich komatiite occurs in the Baltic shield and in other parts of Ontario.

The typical habitat of komatiite is an Archaean greenstone belt. Ultramafic lavas comprise between 0 and about 20% of well-preserved volcanic successions and appear to have similar abundances in both Middle- and Late-Archaean belts. True komatiites are rare or absent in Proterozoic sequences – the spinifex-textured lavas of the Cape Smith belt have komatiitic basaltic compositions – but reappear in one notable example in the Cretaceous. The ~90 Ma ultramafic lavas of Gorgona Island are true komatiites which crystallised as spinifex-textured flows from liquids containing at least 20% MgO.

Formation of High-MgO Liquids

As shown in Figure 4A, magmas with highly magnesian, ultrabasic compositions form either through melting at high pressures, or by high percentages of mantle melting. The effect of increasing the pressure or depth of melting is to increase the stability of orthopyroxene relative to olivine, and that of garnet relative to the more magnesian mantle minerals (*see* **Earth:** Mantle). The consequence is the formation of ever more magnesian magma as the pressure increases. It has been shown that at high pressures, above about 8 GPa, near-solidus melts (liquids produced by melting of mantle peridotite at temperatures only slightly above the solidus) contain more than 30% MgO and have ultrabasic compositions. Increasing the percentage of melting has a similar effect. At shallower levels in the upper mantle, at pressures of ~0.5 to 5 GPa, the minerals that melt at low temperatures (plagioclase, spinel, garnet, and clinopyroxene) have relatively low MgO contents. As the degree of melting increases, the more magnesian minerals, olivine and orthopyroxene, progressively enter the liquid, increasing its MgO content. Figure 4A shows schematically how magmas with 30% MgO form either through deep melting near the solidus, or by higher degrees of melting at shallower levels. However, komatiites contain low concentrations of incompatible elements, which indicates that they formed through relatively high degrees of partial melting. Near-solidus melting therefore is an improbable

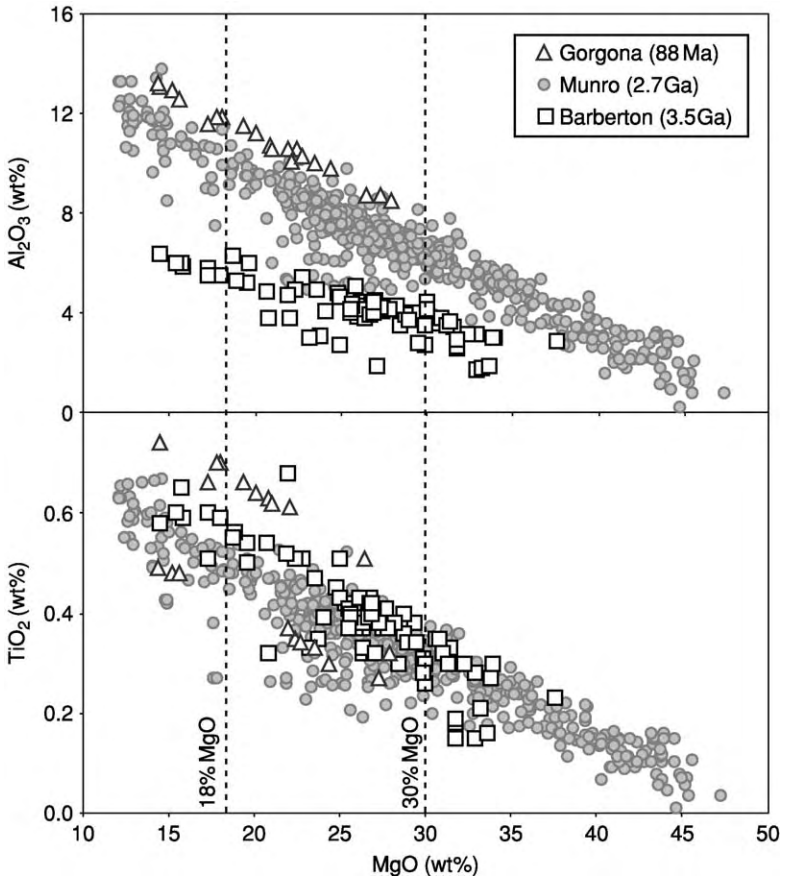


Figure 2 Variation diagrams illustrating the range of compositions of the main types of komatiite. The limit at 18% MgO separates komatiitic basalt from komatiite; the limit at 30% MgO indicates the probable maximum MgO content of komatiitic liquids. Rocks with more than 30% MgO are olivine cumulates.

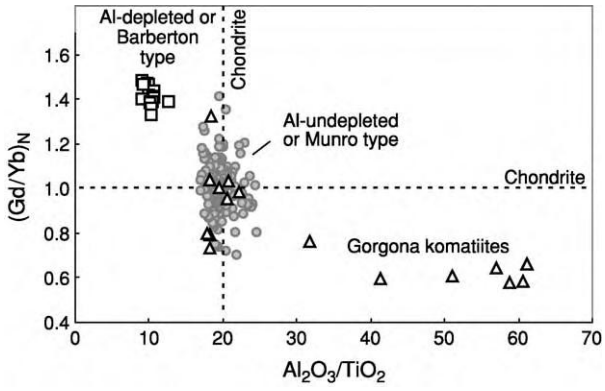


Figure 3 Al_2O_3/TiO_2 versus Gd/Yb showing the main types of komatiite. The high Gd/Yb and low Al_2O_3/TiO_2 of Barberton type komatiite indicates that garnet was retained in the residue of melting and that the magmas formed at extreme mantle depths.

explanation for most komatiites: from Figure 5, we see that on the basis of Ce and Zr contents, they would have formed by 30 to 60% melting of mantle peridotite.

Eruption and Solidification

The viscosity of komatiite magma is very low, one to two orders of magnitude less than that of basalt. A viscosity of 1–2 Pascal-seconds was estimated for dry komatiite containing 28% MgO, which compares with 500–1000 Pascal-seconds for typical basalts. The low viscosity influences the way komatiite segregates from its source, rises through the lithosphere, and erupts on the surface.

Komatiite is a fragile magma. It is far hotter than surrounding rocks, especially when it passes through the lithosphere, and it has a strong capacity to interact with them. When it flows rapidly past crustal rocks, it is capable of thermally eroding and assimilating them; if it ponds in a crustal magma chamber, it will fractionally crystallise and lose its ultramafic character. For komatiite to reach the surface it must flow rapidly and continuously, without pausing on its way. Komatiite is also a relatively dense magma. Anhydrous komatiite, containing 28% MgO, has a density of about 2800 kg m ^{3}, significantly higher than

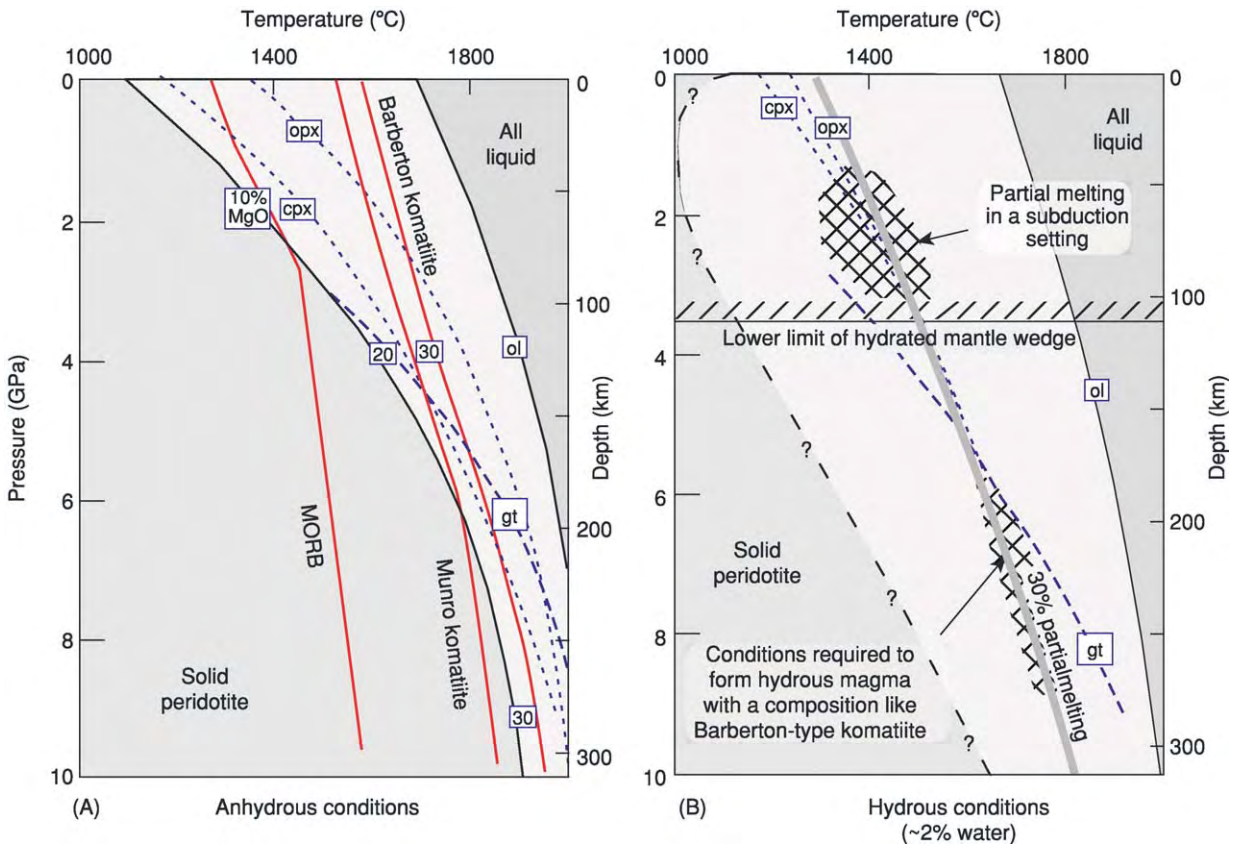


Figure 4 Melting behaviour of mantle peridotite and the formation of ultramafic magmas. (A) Anhydrous conditions. (Phase relations from Herzberg (1999)). The grey lines show the paths taken by mantle that undergoes partial melting. MORB forms from ambient, low temperature mantle. Much high source temperatures, as in mantle plumes, are required to produce komatiites. The plumes intersect the solidus at great depths. The source of Munro type komatiite starts to melt at about 200 km depth and it undergoes fractional melting. The relatively low pressure and the fractional melting process eliminates garnet from the source of these magmas. Barberton type komatiites start to melt at greater depths their source may have been molten as it transited the transition zone. These komatiites are formed by about 30% batch melting at depths greater than 200 km. (B) Hydrous conditions. (Phases relations from Asahara *et al.* (1998)). The position of the solidus is very uncertain but its location is less important than that of the 30% melting curve, which corresponds to the conditions under which Barberton type komatiites form. Pressures greater than about 6GPa are needed to stabilize garnet in the residue of a 30% partial melt. In contrast, the subduction zone komatiite of Grove *et al.* (1999) (shown as a star) forms at much shallower depths. Under these conditions garnet is not stable in the residue of fusion and the melt does not have the geochemical characteristics of Barberton komatiite.

that of many rocks in the upper crust. For komatiite to reach the surface and erupt, it must fill a continuous liquid column within rocks whose average density is greater than that of the komatiite itself. This would be the case when komatiite erupts in an oceanic setting where solidified basalt near the surface has a density similar to that of the komatiite liquid, and the cumulates or other intrusive rocks at deeper levels have higher densities. However, when komatiite traverses or erupts onto a granitic substrate, as in the Kambalda area in Western Australia, the high density of mantle rocks lower in the liquid column must counterbalance the low density of the granites.

We have very little idea how a komatiite behaves during eruption. The best analogue is probably the sheet flows of continental flood basalt sequences. On this basis, we can predict that komatiites probably erupted initially along fissures, as a series of lava fountains. The violence of this fountaining is difficult to judge. It will be enhanced by the low viscosity of the silicate liquid but mitigated by the high density. The primary control, however, is the volatile content in the komatiite magma, which probably is low in most komatiites.

We know from work in areas of good outcrop that once komatiites escape the vent they form highly mobile flows. The maximum length of a komatiite flow is unknown, our knowledge being limited by the quality and continuity of outcrop. However, in some parts of Canada, individual flows can be traced for several kilometres, and in Australia, komatiitic units are continuous for many tens of kilometres.

Thick, massive, olivine-rich units are present in most regions where komatiite lavas are abundant. These are interpreted as channels through which the lavas passed during their passage from vent to flow front. Flowage through these channels may have been very rapid and turbulent. Most of the komatiite flows preserved in greenstone belts represent lateral facies – small sheets or lobes of lava that spread out from a central feeder. These small pulses of lava may never have moved far from the feeder channel before ponding and crystallising beneath a thin elastic crust.

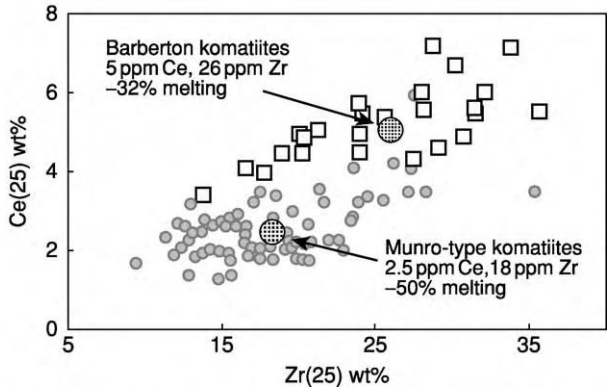


Figure 5 Concentrations of Ce and Zr in komatiites, normalized to 25% MgO by subtracting or adding olivine. The average compositions of the two groups are shown, together with calculated percentages of mantle melting. These were calculated assuming a primitive source for Barberton komatiites (1.6 ppm Ce and 9.7 ppm Zr) and a depleted source for Munro (1.3 ppm Ce and 9.7 ppm Zr). Residual mineralogy: ol 50; opx 30; cpx 10; gt 10. (Partition coefficients are from Green (1974).) (Principal sources of komatiite data are from Sun and Nesbitt (1978); Bickle *et al.* (1993); Jahn *et al.* (1983).)

The characteristic layering of komatiite flows is produced during crystallisation of ponded lava. Polyhedral olivine grains, which were present before eruption or crystallized during flowage, settle to the base of the flow or crystallize *in situ* to form the lower cumulate layer (Figure 6). At the same time, the spinifex-textured upper part of the flow crystallizes through downward growth of crystals from the crust of the flow. Faure *et al.*, showed in 2002 that the presence of a thermal gradient, such as exists at the margin of every flow and high-level intrusion, is instrumental in the formation of spinifex texture. The texture forms as a result of constrained, *in-situ* crystallization of olivine or pyroxene during moderately rapid cooling of low-viscosity ultramafic liquid.

Melting and Segregation of Komatiite Liquids

During mantle melting, if certain conditions are met, the silicate liquid segregates efficiently from its

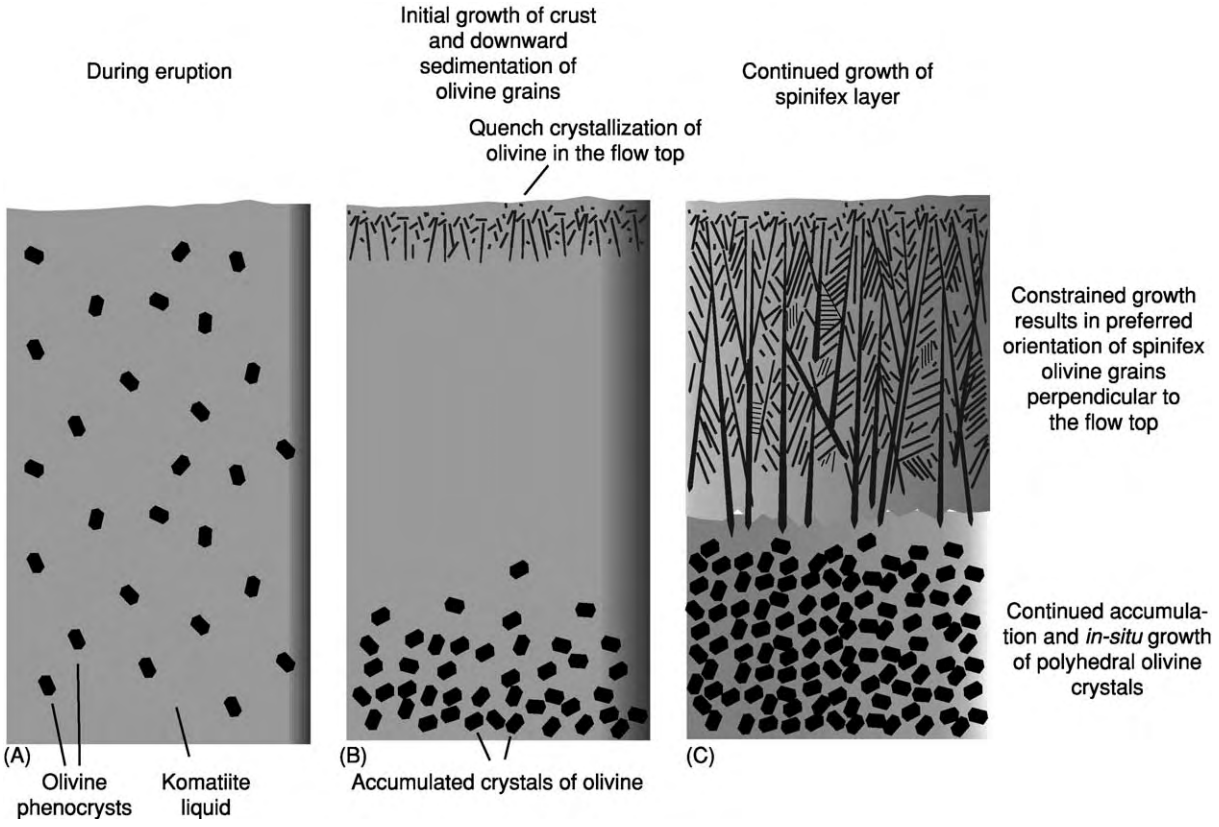


Figure 6 Diagram illustrating the crystallization of a layered spinifex textured komatiite flow.

source, even at low degrees of melting. The conditions are: (i) the viscosity of the liquid is low; (ii) the solid matrix is deformable; and (iii) the density of the liquid is less than that of the unmelted residue. Although this is the normal situation for melting at low pressures where basaltic magmas are produced, it may not always be the case for komatiites. It has been shown through experimental studies that because silicate liquid is more compressible than solid silicate minerals, the density contrast between solid and liquid decreases as pressure increases. At pressures greater than about 8 GPa, which correspond to a mantle depth of about 250 km, the density of an ultramafic komatiite liquid exceeds that of olivine. It remains less, however, than that of garnet, the densest upper-mantle mineral. It has also been shown that at 8 GPa, komatiite magma forms through 30–50% melting, leaving a residue of olivine and majorite garnet. The density of the liquid is slightly less than that of the olivine-garnet residue. Although under static conditions there would be little impetus for the liquid to segregate from its source, mantle melting is normally due to adiabatic decompression in an ascending source, as in a mantle plume. As the source rises, the pressure decreases and the density difference between melt and solid increases. Eventually the density contrast becomes sufficiently large that komatiite of the Al-depleted Barberton type escapes from its source, probably in a single batch of high-degree mantle melt. This type of komatiite is probably one of the rare types of magma that forms through batch melting of the mantle Figure 7.

Al-undepleted or Munro-type komatiites lack the geochemical signature that signals melting in equilibrium with garnet. This does not necessarily require that the source was garnet-free; only that when the komatiite magma separated from its source, garnet was absent in the solid residue. Three processes can contribute to the elimination of garnet: low pressure, which destabilizes garnet; a high degree of melting, which eliminates low-temperature phases; and fractional melting, which preferentially removes the first-melting garnet-rich component from the source. In a rising mantle plume, all three processes may operate together.

Al-undepleted Munro-type komatiites are characterized by relative depletion of the more incompatible trace elements, such as the light rare-earth elements (LREE), Nb and Th. In some cases, notably for the Cretaceous Gorgona komatiites, the extent of LREE depletion varies widely within a suite of rocks of constant Nd isotopic composition. This pattern is a

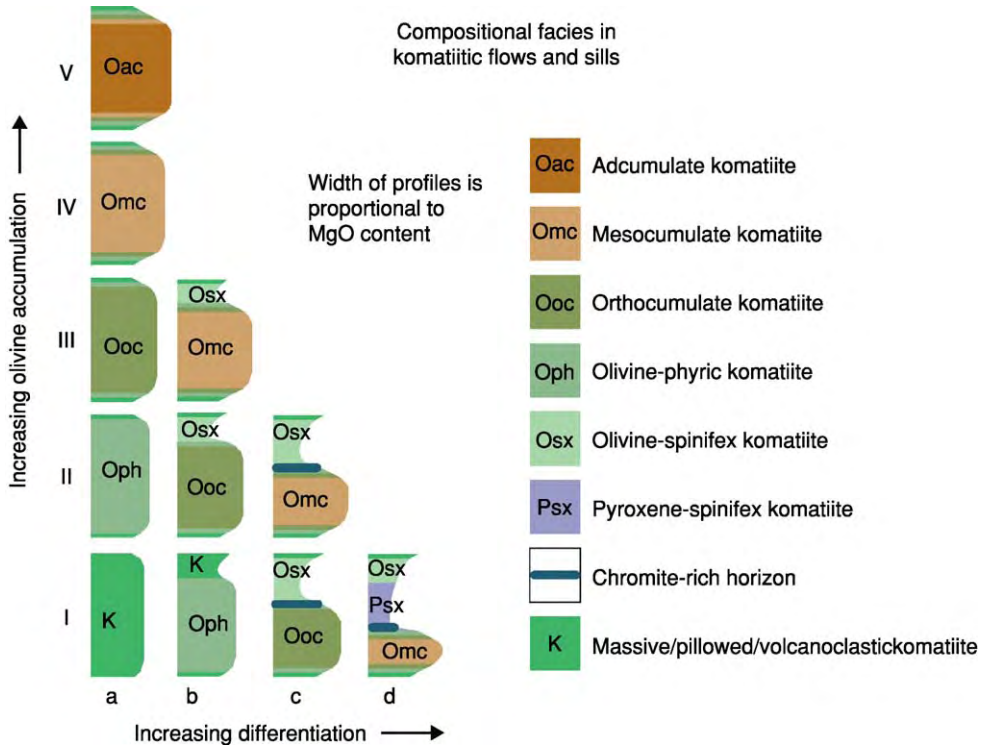


Figure 7 Classification of komatiitic flows and sills on the basis of the relative degrees of olivine enrichment and differentiation *in situ*. (Modified from Lesher *et al.* (1984).) UN, undifferentiated non cumulate (massive, pillowed, or volcaniclastic); DN, differentiated non cumulate; UC, undifferentiated cumulate; DC, differentiated cumulate; illustrating the crystallization of a layered spinifex textured komatiite flow.

clear signature that the rocks formed through fractional melting. The source apparently started to melt at pressures lower than those that produced Barberton-type komatiites. Under these conditions, the komatiite liquid is less dense than the residual solid–perhaps the source initially contained a small water content, which entered the melt and lowered its density. This melt then escaped from the source, taking with it a high proportion of the incompatible elements or low-temperature components, including garnet, and the komatiites then formed through more advanced melting of the now-depleted source.

Wet Komatiites?

Grove and others and Parman and others have proposed that Barberton komatiite is a hydrous magma that forms by partial melting of metasomatized mantle above a subduction zone. In their model, the komatiite magma did not erupt as lava flows but crystallized in high-level sills. Arguments in support of the model are as follows.

i. The presence of water in the mantle source reduces the melting temperatures to 'reasonable' levels.
ii. The morphologies of olivine crystals in spinifex texture are best explained by crystallization in magmas that either contained a significant water content, or had recently exsolved water.
iii. The compositions of augite in spinifex-textured komatiites from the Barberton region resemble those of pyroxene in moderate-pressure hydrous experiments.
iv. Analyses of melt inclusions in chromite in komatiites from Zimbabwe suggest that the parental komatiite contained about 0.9% water.
v. Komatiites from the Boston Creek in Canada and in some other areas contain vesicles and minor amounts of magmatic amphibole.
vi. The trace-element compositions of some komatiitic basalts resemble those of modern boninites.

Many other geologists and geochemists believe that most komatiites are essentially anhydrous and that only some rare examples (excluding the Barberton komatiites) contain moderate water contents. The principal arguments are:

i. The Archaean mantle probably was hotter than the modern mantle because of higher heat production from more abundant radioactive elements, release of accretion energy and core formation. In this context, the ~1600°C liquidus of anhydrous komatiite is not 'unreasonably' hot.
ii. Spinifex texture has been reproduced in anhydrous experiments.
iii. Recent mapping has shown that Barberton komatiite, like komatiite in other regions, erupted on the surface as highly mobile, long-lived, non-explosive lava flows. This corresponds to the eruption behaviour of anhydrous magma.
iv. The extrusive setting of Barberton komatiite casts doubt on the interpretation of pyroxene compositions.
v. Melt inclusions in komatiitic olivine have low water contents.
vi. The trace-element compositions of most komatiites are very different from those of boninites.
vii. The major and trace-element contents of komatiites indicate that they formed at depths far below those of subduction zones (**Figure 4B**).

Ni-Cu-(PGE) Mineralization

Because they have relatively high Ni, Cu, and platinum-group element (PGE) contents and are capable of eroding S-rich crustal rocks, komatiites are capable of forming Ni-Cu-(PGE) sulphide deposits. The best-known examples, of Archaean age, are in the Kambalda region of Western Australia. These deposits are localized in the lower parts of thick dunitic units that are interpreted as lava channels. Komatiite lava flowing turbulently through the channels thermally eroded and melted S-rich floor rocks, leading to the segregation of Ni-Cu-PGE-rich immiscible sulphide liquids that accumulated at the base of the units to form the ore deposits (*see* **Mining Geology:** Magmatic Ores). Proterozoic deposits, such as those of the Cape Smith Belt in Canada, formed by a similar process within invasive lava channels.

See Also

Earth: Mantle. **Igneous Processes**. **Lava**. **Mining Geology:** Magmatic Ores.

Further Reading

Arndt NT (1982) Proterozoic spinifex textured basalts of Gilmour Island, Hudson Bay. *Geological Survey of Canada* Paper 83 1A: 137 142.

Arndt NT (1994) Archean komatiites. In: Condie KC (ed.) *Archean Crustal Evolution,* pp. 11 44. Amsterdam: Elsevier.

Arndt NT and Nisbet EG (1982) What is a komatiite? In: Arndt NT and Nisbet EG (eds.) *Komatiites,* pp. 19 28. London: George Allen and Unwin.

Arndt NT and Nisbet EG (1982) *Komatiites*. London: George Allen and Unwin.

Asahara Y, Ohtani E, and Suzuki A (1998) Melting rela tions of hydrous and dry mantle compositions and the genesis of komatiites. *Geophysical Research Letters* 25: 2201 2204.

Bickle MJ, Ford CE, and Nisbet EG (1977) The petro genesis of peridotitic komatiites; evidence from high pressure melting experiments. *Earth Planetary Sci entific Letters* 37: 97 106.

Echeverria LM (1980) Tertiary or Mesozoic komatiites from Gorgona Island, Colombia; field relations and geochem istry. *Contributions to Mineral Petrology* 73: 253 266.

Faure F, Arndt N, and Libourel G (2002) Crystallisation of plate spinifex texture at 1atm. pressure in a thermal gradient. *Geochim. Cosmochim. Acta* v. Goldschmidt Conference Abstracts, p. A225.

Green DH (1974) Genesis of Archaean peridotitic magmas and constraints on Archaean geothermal gradients and tectonics. *Geology* 3: 15 18.

Herzberg C (1999) Phase equilibrium constraints on the formation of cratonic mantle. In: Fei Y, Bertka CM, and Mysen BO (eds.) *Mantle Petrology: Field Obser vations and High Pressure Experimentation,* pp. 13 46. Houston: The Geochemical Society.

Jahn BM, Gruau G, and Glickson AY (1982) Komatiites of the Onverwacht Group, South Africa: REE chemistry, Sm Nd age and mantle evolution: *Contributions to Min eral Petrology* 80: 25 40.

Le Maitre RW, Bateman P, Dudek A, *et al.* (1989) *A Classification of Igneous Rocks and Glossary of Terms.* Oxford: Blackwell.

Lesher CM, Arndt NT, and Groves DI (1984) Genesis of komatiite associated nickel sulphide deposits at Kambalda, Western Australia: A distal volcanic model. In: Buchanan DL and Jones MJ (eds.) *Sulphide Deposits in Mafic and Ultramafic Rocks,* pp. 70 80. London: Institution of Mining and Metallurgy.

Nesbitt RW, Jahn BM, and Purvis AC (1982) Komatiites: an early Precambrian phenomenon. *Journal of Volcanic Geothermal Research* 14: 31 45.

Nisbet EG (1982) The tectonic setting and petrogenesis of komatiites. In: Arndt NT and Nisbet EG (eds.) *Komatiites,* pp. 501 520. London: George Allen and Unwin.

Nisbet EG, Cheadle MJ, Arndt NT, and Bickle MJ (1993) Constraining the potential temperature of the Archaean mantle: a review of the evidence from komatiites. *Lithos* 30: 291 307.

Parman S, Dann J, Grove TL, and de Wit MJ (1997) Em placement conditions of komatiite magmas from the 3.49 Ga Komati Formation, Barberton Greenstone Belt, South Africa. *Earth Planetary Scientific Letters* 150: 303 323.

Sun SS and Nesbitt RW (1978) Petrogenesis of Archean ultrabasic and basic volcanics: evidence from rare earth elements. *Contributions to Mineral Petrology* 65: 301 325.

Obsidian

G J H McCall, Cirencester, Gloucester, UK

Introduction

Obsidian is an extremely siliceous volcanic rock, found in lava flows and volcanic plugs, domes, and necks. It is closely related to rhyolite but has solidified from an extremely viscous magma as a glass with only minute crystallites within it. Like granite, it has a composition close to the eutectic, which explains the lack of crystallization (i.e., the glass solidified before crystals could form). Obsidian is of rare occurrence in volcanic suites, in which case it is mostly in calc-alkaline island-arc-type suites: this rarity is probably because of the fact that obsidian is the final residual product of differentiated magmatic suites, and the fact that the stickiness of the magma restricts it to protrusions such as domes and flows of small extent, which are squeezed out with difficulty. It has immense archaeological significance, having been used widely by ancient people for ornaments, arrow-heads, knives, and scrapers. The Romans and Greeks quarried obsidian extensively for gemstones and the Aztecs mined it in the fourteenth through sixteenth centuries, presumably for ornaments. It is still listed as semiprecious gemstone, and the famous jeweller Peter Carl Fabergé (1846–1920) used it to create animal-shaped ornaments. Colour (black, brown, green, yellow, and red), translucency and transparency, reflectance, relative hardness, and sharp edges when fractured are all qualities that have made obsidian desirable through the ages.

Historical

The term 'obsidian' is a very ancient word for natural glasses. In the first century AD, Pliny the Younger wrote: "Among the various kinds of glass we may also reckon obsidian, a substance very similar to the stone which Obsidius discovered in Ethiopia. The stone is a very dark colour and somewhat transparent, but it is dull to the sight, and reflects, when attached as mirror to walls, the shadow of the object rather than the image". Thus the origin of the term clearly goes back to the Romans, and the first record was in Ethiopia. That the name of the rock comes from that of a person is, however, doubtful. John Hill in 1740 wrote, based on Theophrastus, a Greek philosopher who wrote extensively on plants, stones, and climatic

topics in about 300 BC, that "the antients had two or three of these dark marbles, of fine texture, of great use amongst them. They took a polish, were transparent to some degree when cut into thin plates, and reflected the image as our looking glasses do. The first kind was called Οψιανος απο τηϛ οψεοϛ, which expressed its property of reflectivity and was afterwards written in the Latin as opsidianus or obsidianus".

This early derivation of the name does seem to agree better with the known fact that Greeks and Romans used obsidian as a gemstone and obtained it from the Island of Melos in the Aegean, where quarries have yielded it for 12 000 years. In 1773, the German mineralogist UFB Bruckmann wrote that obsidian was probably a black lava and geologist Leopold von Buch in 1809 noted that it flowed out, and was not cast out, from volcanoes. In 1822, the American geologist Parker Cleaveland wrote: "This variety has a strong resemblance to glass. Its fracture is distinctly conchoidal, with large cavities and strongly shining with a lustre more or less vitreous. The surface of the fracture often exhibits a striated or wavy appearance, and its appearance is a little unctuous. It scratches glass, gives fire with steel, but is brittle, and falls into sharp-edged fragments. Most commonly it is translucent at the edges, or opaque, but some varieties are translucent or in thin scales transparent. Its colour is black, either deep or pure, or tinged with brown, green, blue or grey, and sometimes passes to blue, green, brown or gray, even yellow or red. The darkest colours often discover a tinge of green by translucent light".

Composition

The Norwegian geologist and petrologist JHL Vogt in 1923 wrote that "compositions of eutectic or nearly eutectic proportions promote the formation of glass, since the eutectic has the lowest melting point; consequently, at that temperature the melt is more viscous than elsewhere on the curve, and points near the eutectic tend to reach solidifying point before reaching the crystallizing point. With relatively quick cooling the crystallization will be entirely or nearly restrained. Thus it is no accidental circumstance that by far the most obsidians have nearly the chemical composition of the granitic eutectite". As now used, the term 'obsidian' is applied to massive, usually dense, but often slaggy glasses of deep brown or black, grey, red, or mottled red and black colour. The viscosity of obsidian as a flow stems from branching and tangled chains of tetrahedral silicon and aluminium combined with oxygen. When solidified, obsidian is quite hard and its conchoidal fracture results in sharp, even cutting edges to the brown fragments. In many cases, the rock is spotted or banded. Spherulites and lithophysae occur in some obsidians, and may be abundant, also concentrated in certain layers. Normally obsidians are natural glasses of rhyolites, but any acid (siliceous) volcanic rock may solidify as similar glass by rapid cooling, and thus the terms 'trachyte' and 'dacite-obsidian' in common use, though strictly obsidians are of rhyolite composition.

The specific gravity of obsidian ranges from 2.30 to 2.58. The refractive index ranges from $n = 1.48$ to 1.53. The hardness on Moh's scale ranges from 5.6 to 7. The chemical compositions of various obsidians are given in Table 1; also shown in the lower part of the table are the CIPW norms (named after the petrologists Cross, Iddings, Pirsson, and Washington, in 1931). A norm is a means of converting a chemical composition of an igneous rock to an ideal mineral composition. In this way, similarities in rocks with contrasting mineral assemblages can be noted. Some of the factors considered are temperature, pressure, and mineral content; in the CIPW norm calculation, the magma is considered to be anhydrous and at low pressure.

Chemically, obsidian has a low water content, but even so, this is an order or more greater than is the case for tektites, which resemble obsidian and were once referred to as obsidianites. For example, water content of moldavites from Central Europe ranges from 0.006 to 0.010. Tektites also contain lechatelierite, an amorphous form of quartz that is never found in volcanic glasses (*see* **Tektites**).

Occurrences Worldwide

Obsidian Cliff, Yellowstone National Park

Obsidian Cliff in Yellowstone National Park (Wyoming, USA) is considered a typical occurrence. The chemical compositions of red and black obsidian samples from the site are given in Table 1. The composition is rhyolitic. The cliff forms a giant flow 120–160 m thick. The rock is locally columnar, and at the lower part is traversed by bands or layers of small grey spherulites, but cavities or lithophysae are almost absent. Higher up, the obsidian is less massive and contains large lithophysae (concentric shells of flattened fine material with a central cavity) parallel to the plane of flow.

Eolian Islands

Three recent obsidian flows from the island of Lipari have been described, being the youngest (from the sixth to eighth centuries AD) eruptives on the island (Figure 1). The Rocche Rosse flow is of obsidian

Table 1 Chemical analyses of obsidians

	Sample[a]						
Mineral/norms	*1*	*2*	*3*	*4*	*5*	*6*	*7*
SiO_2	75.52	74.70	76.20	75.23	74.37	74.05	73.84
TiO_2							0.14
Al_2O_3	14.11	13.72	13.17	12.36	12.65	14.67	13.00
Fe_2O_3	1.74	1.01	0.34	0.96	2.58	0.89	1.82
FeO	0.08	0.62	0.73	1.24	n.d	n.d	0.79
MnO			0.10				0.07
MgO	0.10	0.14	0.19	0	0.20	0.26	0.49
CaO	0.78	0.78	0.42	1.00	1.22	0.97	1.52
Na_2O	3.92	3.90	4.31	4.00	3.87	3.99	3.82
K_2O	3.63	4.02	4.42	4.62	4.57	5.11	3.92
H_2O	0.29	0.62	0.33	0.73	0.24	0.91	0.53
P_2O_5				0.27			0.01
FeS_2	0.11	0.40					0.02
Total	100.38	99.91	100.25	100.42	99.79	99.85	99.97
CIPW norms[b]							
Q	36.72	34.86	32.40	32.76	29.88	26.94	
Or	21.68	23.35	26.69	27.24	27.24	30.98	
Ab	33.01	33.01	36.16	33.54	33.01	32.54	
An	3.80	3.89	1.95	2.22	3.34	5.00	
C	2.24	1.53	0.51			0.71	
Di				0.75	2.20		
Hy	0.30	0.66	1.69	1.06	3.57	2.28	
Rut	0.23	1.39	0.46	1.39			
Hm	1.00						
Ap				0.67			

[a]n.d., Not determined. Samples: 1, red obsidian, Obsidian Cliff, Yellowstone National Park; 2, black obsidian, Obsidian Cliff, Yellowstone National Park; 3, obsidian, Obsidian Hill, Tewan Mountains, New Mexico; 4, obsidian, Cerro de Los Navajos, Mexico; 5, obsidian, Forgia Vecchia, Lipari; 6, obsidian, Tenerife; 7, average of 44 obsidians.
[b]See text for discussion of CIPW norms.

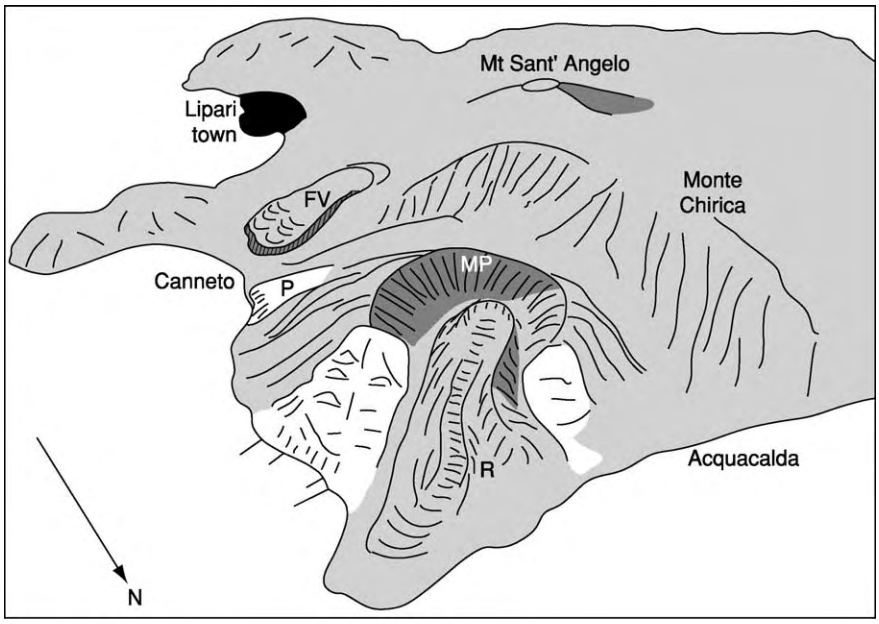


Figure 1 The island of Lipari, north of Sicily, showing the obsidian flows dating to the sixth to eighth centuries AD (R, Rocche Rosse; FV, Forgia Vecchia; MP, Monte Pilato; P, Pomiciazzo), the most recent eruptives on the island. Reproduced with permission from Guest JE, Cole PD, Duncan AM, and Chester DK (2003) *Volcanoes of Southern Italy.* Bath: Geological Society Publishing House.

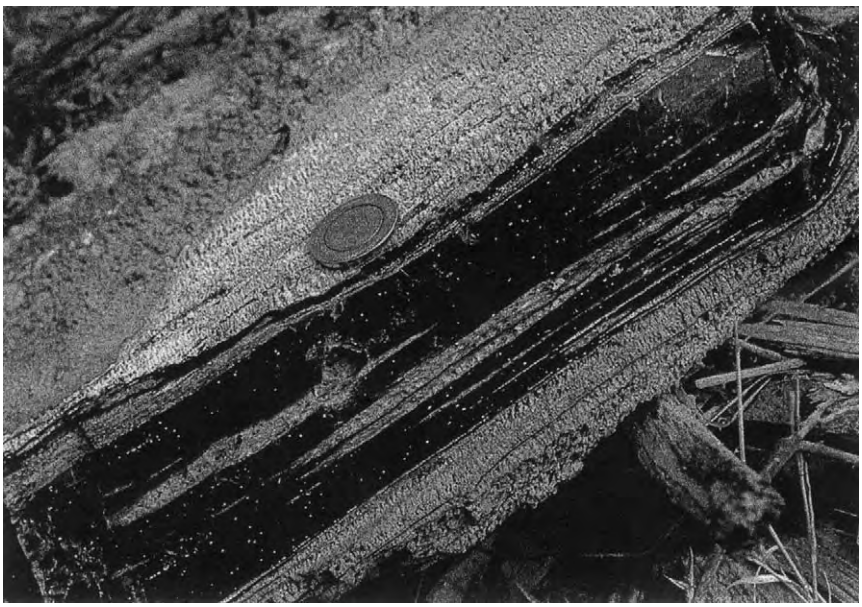


Figure 2 Close up of the Rocche Rosse obsidian flow, Lipari, showing alternations of fresh and devitrified glass layers. Reproduced with permission from Guest JE, Cole PD, Duncan AM, and Chester DK (2003) *Volcanoes of Southern Italy*. Bath: Geological Society Publishing House.

block lava: the surface has arcuate surface ridges (ogives) and talus margins. The flow is several tens of metres thick and is typical of high-viscosity silicic lavas. There are devitrified layers and lenses alternating with the black obsidian (Figure 2). There are also obsidian flows on the nearby island of Vulcano.

Armenia

The Quaternary Gutansar volcano in Armenia shows a progression from basalts and andesites, through liparites, transitional liparite obsidians, to massive obsidian flows. There is a fissure extrusion of black obsidian at the base, changing gradually to brownish and greyish slightly swelled glass, then to perlite and pumice. The obsidian extrusion has a fanlike structure (Figure 3). The distribution of the obsidian bodies indicates very shallow magma chambers. Such viscous magmas increase in volume and heave up continuously to the surface with regular accelerations, the process being one of tectovolcanic displacements.

Newberry Caldera, Oregon

Newberry Caldera in the Oregon Cascades (USA), 8 km in diameter, lies atop of a volcano initiated in the Pliocene, but its main activity occurred during the Pleistocene and the last eruptions occurred just over 2000 years ago. The obsidian flows lie within the caldera and are among the most recent eruptives. They present expanses of barren, shiny black obsidian, almost devoid of vegetation, and surfaces sharp as razors can cut the legs of the unwary geologist. There are four separate steep-sided flows of glistening black obsidian. The Big Obsidian Flow (Figure 4) covers 1.5 km^2 and has its own dome of pumiceous obsidian plugging the vent. Two smaller flows straddle a north-east-trending fissure from which they were erupted. Another small flow was erupted from a vent on the caldera wall. This development at Newberry Caldera is undoubtedly the most spectacular development of obsidian flows known, but the terrain is extremely difficult inside the caldera.

California

Obsidian in the central eastern Sierra Nevada mountains of California, USA, at Mono Craters, occurs in a flow of viscous lava, as lenses and bands interleaved with rhyolite (Figure 5). The obsidian is very regularly banded (Figure 6). Two other spectacular obsidian flows occur further north, at Glass Mountain, California (Figure 7), in the Medicine Lake region of the Cascade Range, overlooking the Nevada and Oregon borders. These flows are less than 1000 years old and are associated with small obsidian domes. They display remarkable flow structures, of curved pressure ridges akin to those on the surface of glaciers.

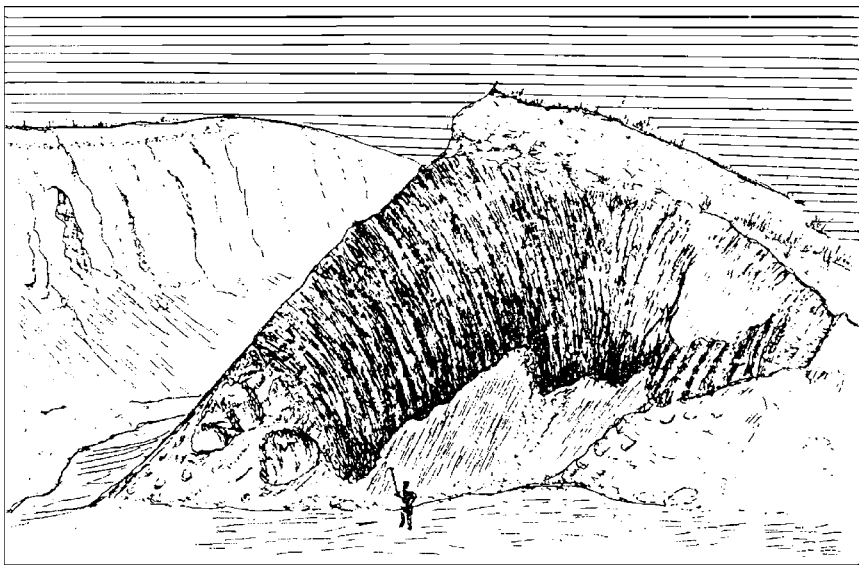


Figure 3 Fan like fluidal structure in an obsidian extrusion, Gutansar volcano, Armenia. Reproduced from Mkrtchian SR, Paffengoltz KN, Shirinian KG, Karepetian KI, and Karepetian KG (1971) *Late Orogenic Acid Volcanism of Armenian SSR*. Yerevan: Publishing House of the Academy of Sciences of the Armenian SSR.

Figure 4 'Big Obsidian Flow', the Newberry Caldera, Oregon, showing also the obsidian dome from which it emanates. The flow is about 5 km long from source to front. Note the corrugations, or 'corda', which are pressure ridges as on a glacier, each of an amplitude of ~5 m. Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, and Berlin: Springer Verlag.

Obsidian as a Semiprecious Stone

The physical attributes of obsidian might seem to make it an unlikely material for ornamental use, but it is in fact classified as a semiprecious stone (Figure 8). The small carved rhinoceros in Figure 8 is an example of work by Fabergé. Obsidian has been described and illustrated by Bauer and Bouska, who noted that it is a product of rapid solidification of molten magma and is amorphous and without cleavage. These authors mentioned the large obsidian formations in the Eolian

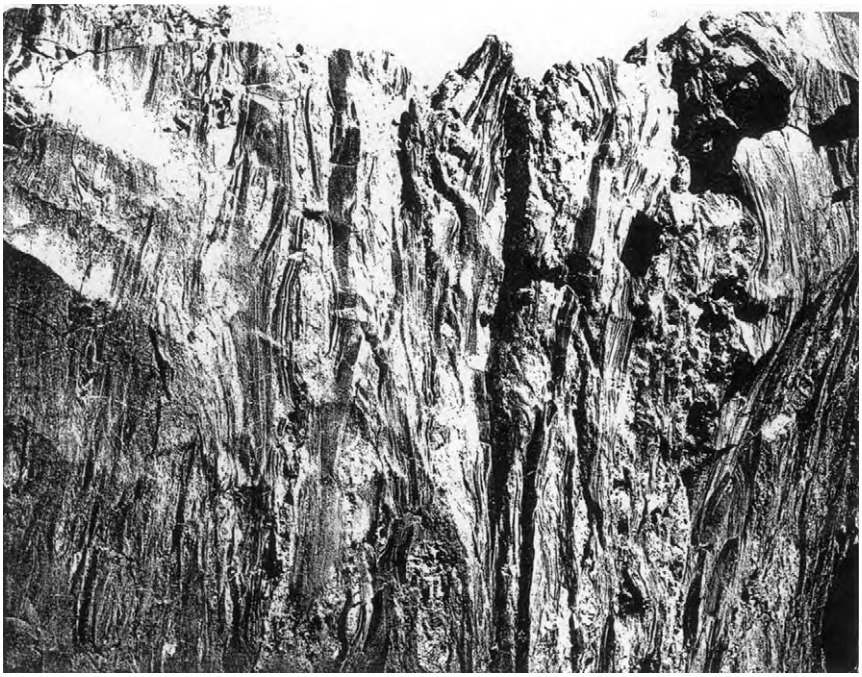


Figure 5 Banded rhyolite and obsidian (black), Panum Crater, Mono Craters, California (height of face is 1.5 m). Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, and Berlin: Springer Verlag.

islands, especially Lipari, and those in Oregon, Arizona, Utah, Hungary, Slovakia, the Caucasus, the Urals, and New Zealand. Most obsidians reveal abundant minute crystallites under the microscope. Obsidians from Mexico, the United States, and the Lake Sevan area of Armenia display an outstanding sheen and silky lustre, due to the presence of such crystallites. The pre-Colombian Indians of Central America, besides fashioning knives, scrapers, and spearpoints from it, used it for ornamentation. The Mexican Indians have long been famous for fashioning decorative objects and amulets, necklaces, and bracelets of obsidian (jet is used in a similar way around Whitby, Yorkshire; it is not obsidian, but is lignitic and is believed to be derived from water-logged driftwood in shales).

As a gemstone, obsidian is cut in various ways. Perfectly transparent varieties are faceted, mainly into steep cuts, but brilliants are quite common. Materials of lower transparency are faceted into cabochons, or table-cuts. Large irregular cabochons are commonly cut from Mexican or North American obsidian and contain tiny crystallites of the tetragonal polymorph of silica, cristobalite. A very attractive obsidian with grey, eye-like spots in the red, haematite-coloured groundmass comes from Gyumishkoe in Armenia. Pitchstone is more siliceous than obsidian, and lacks its lustre, and tachylite also resembles it, but is basaltic in composition.

Obsidian Artefacts

Obsidian is one of the most useful materials known to archaeology. Towards the end of the Stone Age, humans increasingly sought obsidian for making cutting tools. Skilled artisans made knives and daggers and ornamental bowls and polished mirrors; even the making of less sophisticated tools such as scrapers and arrowheads required expertise in chipping off flakes. Such tools appear abundantly at archaeological sites in Europe and Asia, in South, Central, and North America, and in the Pacific Islands. Obsidian is rare in Australia, though obsidian artefacts have been found in Queensland. It is mainly found in quite recent island-arc volcanic belts, and these do not traverse Africa, though the original obsidian find in Roman times was reportedly in Ethiopia. Because obsidian hydrates rapidly, most fresh obsidian is less than 10 million years old.

Archaeological Tracing Methods

Obsidian served early humans well for making cutting tools, due to its relative hardness and the manner in which it fractures into pieces of sharp-edged

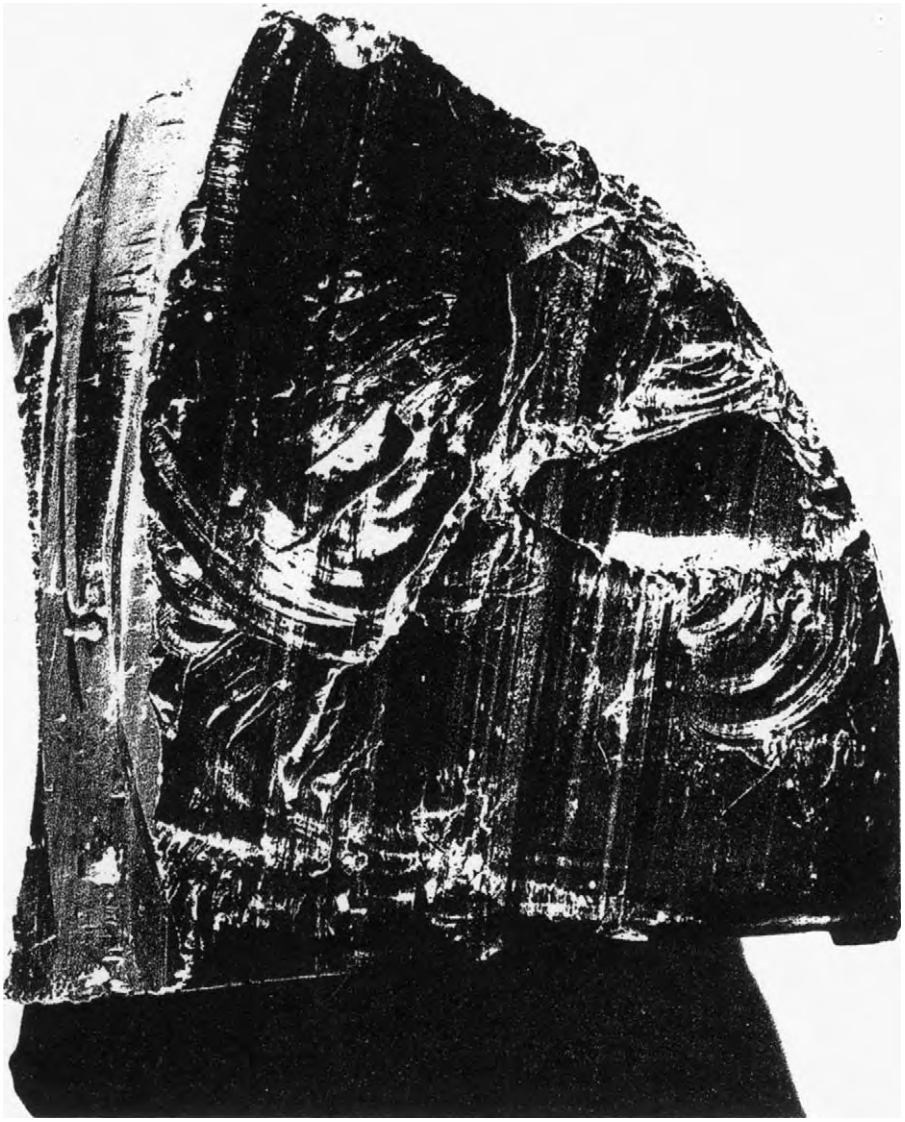


Figure 6 Banded obsidian, Mono Craters, California (maximum dimension of specimen, 12 cm). Note the conchoidal fracture. Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary.* New York, Heidelberg, and Berlin: Springer Verlag.

wedges. Because it is isotropic, there is no preferred direction for striking off flakes. The age of the artefacts, as distinct from the glass of which they are made, can be determined by radiocarbon dating or other means such as stratigraphic position or associated potsherds. The hydration state of the obsidian can be used for dating the glass, because the rock hydrates at a steady rate with the passage of time, though this would seem to be a method liable to error. By any means, dating of obsidian artefacts has helped to delineate ancient trade routes extending for thousands of kilometres.

Newly developed geochemical techniques have made it possible to trace the source of the glass utilized in artefacts. The aim of these tracing techniques is to trace the obsidian from the archaeological sites back to the parent lava flow, by geochemical matching. Minor and trace elements that have a wide compositional range variation in obsidians are utilized, especially manganese and sodium. Such variations are due to the trace element composition differences in the parent rocks, which were partially melted in the mantle or crust to produce the magma. Variations also derive from changes due to crystals forming on the way up to the surface (e.g., in the magma chamber); partition of minor or trace elements is imposed between crystals, which separate off and melt. Successive flows from the same magma chamber are erupted

Figure 7 Two obsidian flows and several obsidian domes, Glass Mountain, California. The northern of the two flows emanates from an obsidian summit dome with a minute axial orifice for gas escape, and there is a succession of very small pancake like effusions from a chain of centres in the upper left hand corner of the photograph. The pressure structures on the flows, their glassy nature, and their tongue like form, with steep sides and fronts on account of their viscosity, are features clearly illustrated. Reproduced with permission from Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary*. New York, Heidelberg, and Berlin: Springer Verlag.

at different times in the magma chamber crystallization, and so they too will show differences in minor or trace element ratios. Each obsidian flow thus has a unique minor and trace element 'fingerprint'. Geochemical matching begins with obtaining a number of different minor and trace element values for each source flow, as a baseline. Having identified the minor trace element values of the likely source flow, the geochemist can with confidence match it with obsidian artefacts from archaeological sites and delineate the trade routes. In J Glover's *Geological Journeys*, three cases were described in which this has been done successfully.

In various northern and central locations of what is now the United States are found the Hopewell sites, which are remarkable burial sites built by the ancestors of Native Americans between 100 BC and AD 500. Obsidian artefacts are found there, but there are no developments of the volcanic rock nearby. Possible sources on the Pacific Coast or in Alaska, Yellowstone National Park, New Mexico, and Mexico have been investigated by JB Griffin and colleagues. Neutron activation analyses confirmed two groups with Na/Mn ratios of 150 and 90. These results fingerprinted two flows at Yellowstone National Park in Colorado. Figure 9 shows the path over which the obsidian had to travel, over 2400 km eastward, to the Hopewell sites. The process may have involved intermediate stages of bartering across this route.

Artefacts in the south-west Pacific have also been analysed. Emission spectroscopy analysis results on flakes from the Santa Cruz Islands, part of the nation of the Solomon Islands (one of the most northern group of the chain previously called the New Hebrides, now Vanuatu), show that the chemistry matches obsidians cropping out on New Britain, not the basaltic volcanic rocks of Santa Cruz. Radioacarbon dating shows that the flakes were deposited in Santa Cruz about 1000 BC, so the population of the islands 3000 years ago must have carried obsidian in small craft by sea east–southeastwards over at least 2000 km (Figure 10). Work in 1996 by R Service has shown, further, that obsidian from New Britain and the Admiralty Islands (a group close to New Britain) was transported earlier to a

Figure 8 Gem quality obsidian: (1) black, massive obsidian from the Caucasus (60 × 45 mm); (2) obsidian with a silvery, silky sheen, Yerevan, Armenia (120 × 80 mm); (3) Neolithic tool made of obsidian, Cejkov, Slovakia (140 × 100 mm); (4) rhinoceros figurine by Carl Fabergé, Russia (length, 60 mm). Reproduced with permission from Bauer J and Bouska V (1983) *Precious and Semi Precious Stones.* London: Octopus.

6000-year-old site in northern Borneo, in the reverse direction (Figure 10).

An obsidian trade in the Middle East and Near East has been described by O Williams-Thorpe. About 8000 BC, humans in the Fertile Crescent of the Middle East, extending from the Lebanon to Iraq, for the first time left the hunter–gatherer mode of life, instead domesticating animals and growing cereals and vines – this has been called the Neolithic revolution. Obsidian was being traded extensively through this region during this period, presumably mostly for cutting and scraping tools, and here again the movements have been geochemically traced.

One further source of artefacts has been quarries and mines. Quarries were used as a source of obsidian by the Romans on Melos, and in Mexico, on a volcanic peak near Veracruz, the Aztecs mined obsidian from AD 1350 to 1520. The obsidian was removed

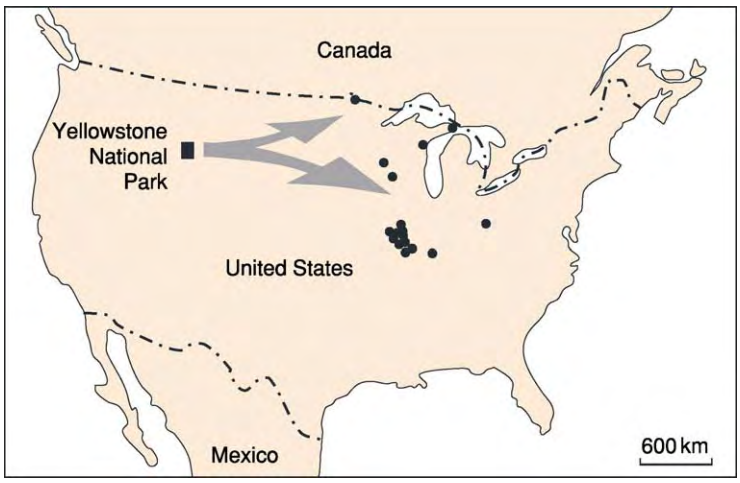


Figure 9 The path of transport of obsidian from Yellowstone, Colorado, to the Hopewell Indian sites. Reproduced with permission from Glover J (2003) *Geological Journeys: From Artifacts to Zircon.* Perth, Western Australia Division: Geological Society of Australia.

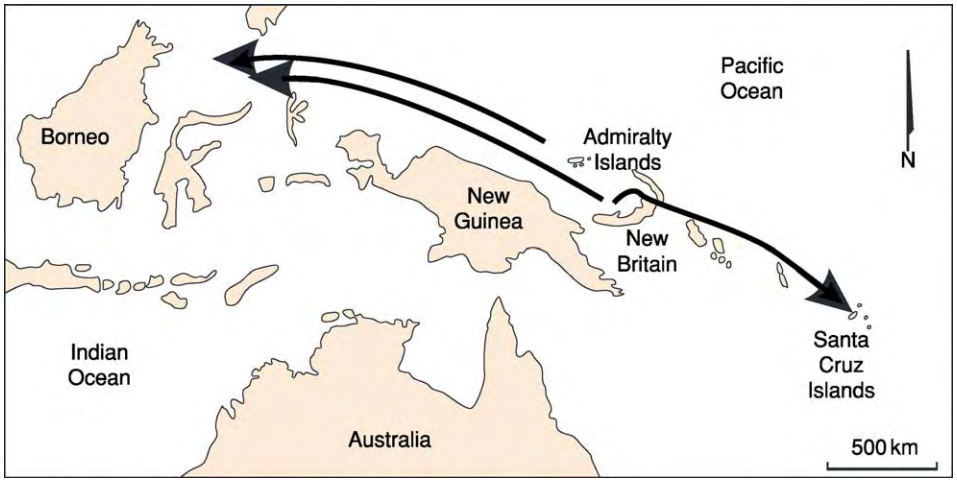


Figure 10 The trade routes for obsidian from New Britain and the Admiralty Islands to Santa Cruz Islands and Borneo, respectively. Reproduced with permission from Glover J (2003) *Geological Journeys: From Artifacts to Zircon.* Perth, Western Australia Division: Geological Society of Australia.

by driving passageways 70 m into the rock. Using levers of sharpened pine poles, the Aztecs exploited weaknesses and fractures in the obsidian rock, even using hammer stones of obsidian.

See Also

Earth: Mantle; Crust. **Geoarchaeology**. **Lava**. **Plate Tectonics**. **Volcanoes**.

Further Reading

Ambrose WR and Green RC (1972) First millennium BC transport of obsidian from New Britain to the Solomon Islands. *Nature* 237: 31.

Bauer J and Bouska V (1983) *Precious and Semi Precious Stones*. London: Octopus.

David B, Bird R, Fullagar R, and Little L (1993) Glassy obsidian artifacts from North Queensland: the Nolan's Creek source and some archaeological occurrences. *The Artifact* 15: 25 30.

Dixon J, Cann JR, and Renfrew C (1968) Obsidian and the origins of trade. *Scientific American* 218(3): 38 46.

Glover J (2003) *Geological Journeys: From Artifacts to Zircon.* Perth, Western Australia Division: Geological Society of Australia.

Green J and Short NM (1971) *Volcanic Landforms and Surface Features: A Photographic Atlas and Glossary.* New York, Heidelberg, and Berlin: Springer Verlag.

Griffin JB, Gordus AA, and Wright GA (1969) The identifi cation of sources of Hopewellian deposits in the Middle West. *American Antiquity* 34: 1 14.
Guest JE, Cole PD, Duncan AM, and Chester DK (2003) *Volcanoes of Southern Italy*. Bath: Geological Society Publishing House.
Johannsen A (1952) *A Descriptive Petrology of Igneous Rocks: Volume 2, The Quartz Bearing Rocks*. Chicago: University of Chicago Press.
Mkrtchian SR, Paffengoltz KN, Shirinian KG, Karepetian KI, and Karepetian KG (1971) *Late Orogenic Acid Vol canism of Armenian SSR*. Yerevan: Publishing House of the Academy of Sciences of the Armenian SSR.
Peterson NV and Grob EA (1965) *State of Oregon Lunar Geological Field Conference Guide Book*. Corvallis: Uni versity of Oregon Department of Geology and New York Academy of Sciences.
Service RF (1996) Rock chemistry traces ancient traders. *Science* 274: 212 213.
Torrence R (1986) *Production and Exchange of Stone Tools: Prehistoric Obsidian in the Aegean. New Stu dies in Archaeology*. Cambridge: Cambridge University Press.
Williams Thorpe O (1995) Review article: obsidian in the Mediterranean and Near East: a provenancing access story. *Archaeometry* 37: 215 238.

IMPACT STRUCTURES

R A F Grieve, Natural Resources Canada, Ottawa, ON, Canada

Introduction

The systematic study of the terrestrial impact record is a recent endeavour in geology. The first terrestrial impact structure was recognized in 1906. By the late 1930s, the number of terrestrial impact structures had risen to about 20 but there was considerable controversy over their origin. The recognition, however, of diagnostic mineralogical indicators of, so-called, shock metamorphism established reliable criteria for the occurrence of extreme transient and dynamic pressures in the geologic environment and, hence, the occurrence of an impact event. These criteria have been developed further by observation and experiment over the intervening years and have stood the test of time and new observations. Terrestrial impact structures are essentially the sole source of ground truth data on large-scale natural impact events, with respect to the character and spatial distribution of impact-related lithologies and structure, particularly in the third dimension. Although the terrestrial impact record serves as an analogue for impact cratering processes on the other terrestrial planets, impact was not generally regarded, until recently, as a process of importance to the evolution of the Earth. This changed dramatically with the evidence for the occurrence of a major impact at the Cretaceous–Tertiary (K–T) boundary 65 My ago. Although originally contentious to many in the larger geoscience community, the involvement of a major impact at the K–T boundary and the related mass-extinction event in the biosphere is now generally accepted.

General Character of the Record

The terrestrial impact record is incomplete. It is characterized by inherent biases, largely related to the high level of endogenic geological activity of the Earth. At the time of writing, the known impact record consists of approximately 170 known impact structures or crater fields, and some dozen impact events, which are recorded in the stratigraphic record, some of which are related to known impact structures: for example, the modavite tektites with the Ries impact structure, Germany and the Ivory Coast microtektites with the Bosumtwi impact structure. Some of these tektite–microtektite strewn field cover a considerable area of the Earth's surface, e.g., the Australasian strewn field covers an area in excess of 50×10^6 km^2.

Due to the effects of erosion, the terrestrial impact record contains a mixture of topographic forms, and it is more appropriate to use the generic term impact structures than impact craters, which by definition require a negative topographic form. The discovery rate of new impact structures is about five per year, with the rate of discovery having increased substantially in the past three decades (**Figure 1**). A current listing of known terrestrial impact structures can be found at http://www.unb.ca/passc/ImpactDatabase.

The spatial distribution of known terrestrial impact structures is biased towards the stable interior or cratonic areas of the North American, Australian, and European continents (**Figure 2**). These are areas with low rates of erosional and tectonic activity, over extended periods of geological time. They are, thus, the best available surfaces for the preservation of

impact structures in the terrestrial geological environment. They are also areas where there have been active programmes to search for and study impact structures.

Erosion and sedimentation result in characteristics of the terrestrial impact record that are not as dominant on other planetary bodies. For example, approximately one-third of known impact structures are buried by post-impact sediments. Most buried impact structures were detected initially as geophysical anomalies and later drilled, for economic or scientific purposes, which confirmed their impact origin. Although a few submerged impact structures are known, they occur on the relatively shallow, continental shelves. No impact structures are known from the true ocean floors. This reflects the relatively young age and the generally poor resolution of geological knowledge of the ocean floors.

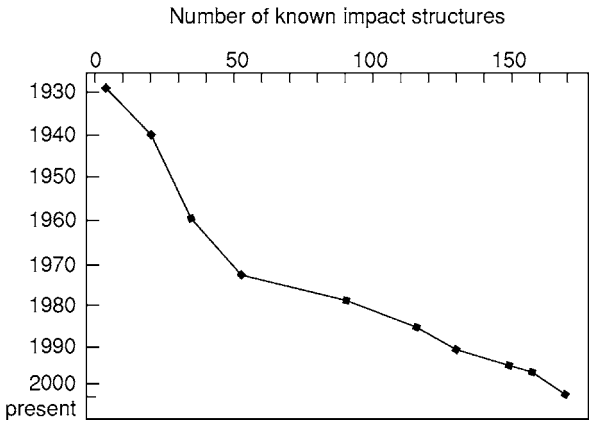


Figure 1 Number of known terrestrial impact structures recognized with time. Note the increase in the rate of discovery after the 1970s, due to the establishment of shock metamorphism (see text for details) as a reliable criterion for identifying the impact origin of specific structures.

There are also biases in the ages and sizes of known terrestrial impact structures. The majority are <200 My old. This reflects problems of preservation and, to a lesser extent, recognition in the highly active geological environment of the Earth. At larger diameters, the cumulative size–frequency distribution is similar to the production distribution observed on the other terrestrial planets. In the terrestrial case, however, this distribution more likely represents a steady-state condition between the formation and removal of impact structures from the record. At diameters below approximately 20 km, the cumulative size–frequency falls off, with an increasing deficit of structures at smaller diameters. This drop-off is an inherent property of the terrestrial record, as it has remained even with the addition of new structures to the known record over the past 30 years. The deficit of small craters is due to atmospheric crushing of smaller impacting bodies, the relative ease with which smaller structures can be buried or eroded, and the intrinsic difficulty in recognizing smaller structures.

Morphology

Small impact structures have the form of a bowl-shaped depression, with an upraised rim, and are known as simple craters (Figure 3). The rim, walls, and floor define the so-called apparent crater. At the

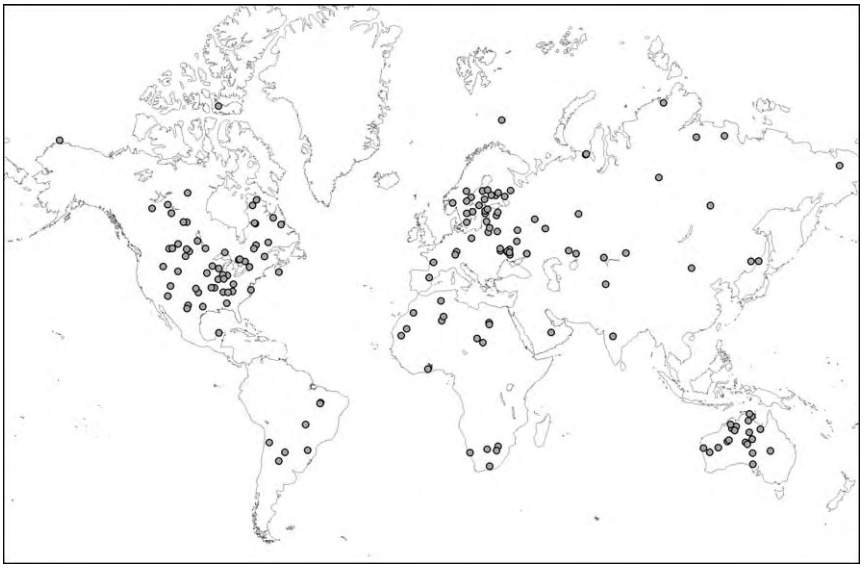


Figure 2 Spatial distribution of known terrestrial impact structures. Note current concentrations in the stable interiors of North American, Australian, and European continents.

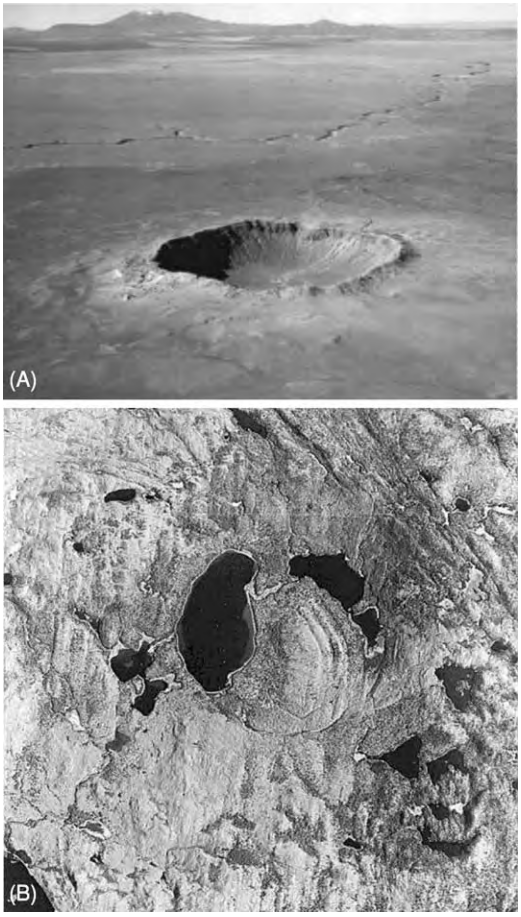


Figure 3 (A) Oblique aerial view of 1.2 km diameter, 50 000 year old simple crater, Meteor or Barringer Crater, Arizona, USA. (B) Vertical aerial view of 3.8 km diameter, 450 ± 30 My old, Brent Crater, Ontario, Canada. Note how this ancient crater has no rim, has been filled by sediments and lakes, and is a generally subtle topographic feature.

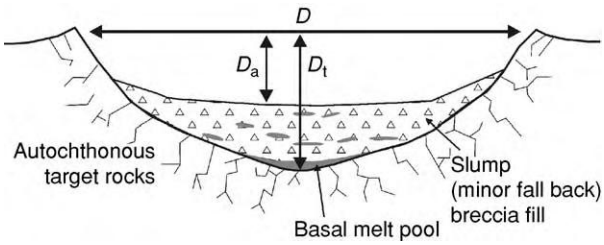


Figure 4 Schematic cross section of a simple crater. D is the diameter and D_a and D_t are the depths of the apparent and true crater, respectively.

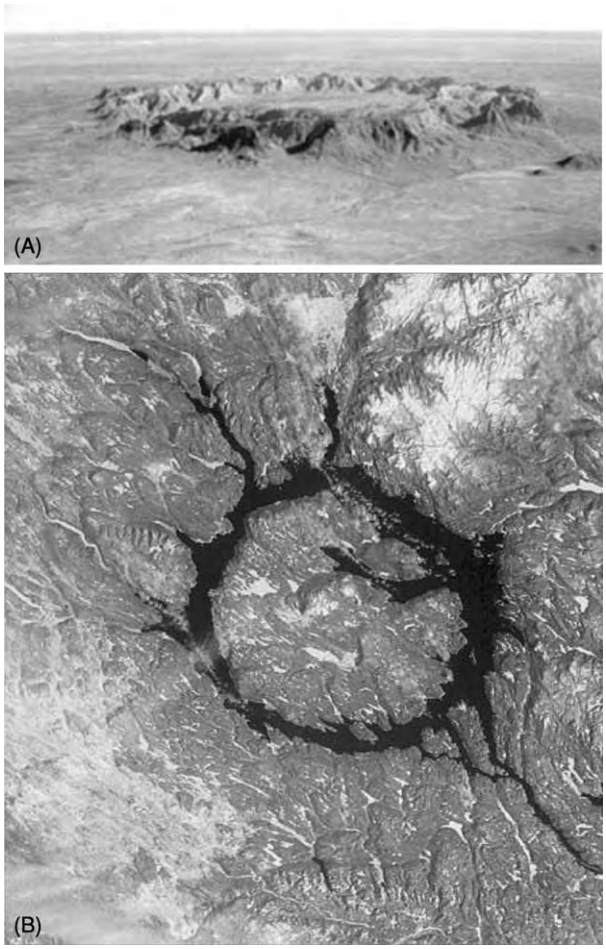


Figure 5 (A) Oblique aerial photograph of the Gosses Bluff impact structure, Australia. Note that all that is visible of the originally 22 km, 142.5 ± 0.8 My old structure is a 5 km annulus of hills, representing the eroded remains of a central uplift. (B) Shuttle photograph of the Manicouagan impact structure, Canada, 100 km in diameter and 214 ± 1 My old. Note that the annular trough (with a diameter of ~65 km) is filled by water.

rim, there is an overturned flap of ejected target materials. Beneath the apparent floor is a lens of brecciated target material that is roughly parabolic in cross-section (Figure 4). This breccia lens is allochthonous and polymict. In places, this breccia lens contains highly shocked and melted target materials. Beneath this breccia lens, parautochthonous, fractured target rocks define the walls and floor of what is known as the true crater. The depth to the base of this breccia lens is roughly one-third of the rim diameter and the depth to the top of the breccia lens is about one-sixth. Shocked rocks in the parautochthonous materials of the true crater floor are confined to a small central volume at the base of the true crater.

With increasing diameter, simple craters show increasing evidence of wall and rim collapse and evolve into complex craters (Figure 5). Complex craters on Earth first occur at diameters greater than 2 km in sedimentary target rocks but not until diameters of 4 km or greater in stronger, more coherent, igneous or metamorphic, crystalline target rocks. The rim of complex craters is a structural feature, corresponding

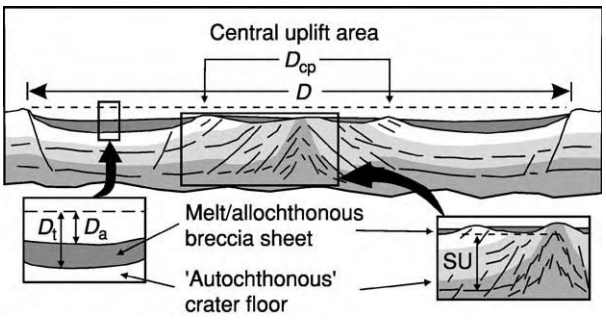


Figure 6 Schematic cross section of complex impact structure. Notation as described in the legend to Figure 4 with SU corresponding to structural uplift and D_{cp} to the diameter of the central uplift. Note preservation of beds in outer annular trough of the structure, with excavation limited to the central area.

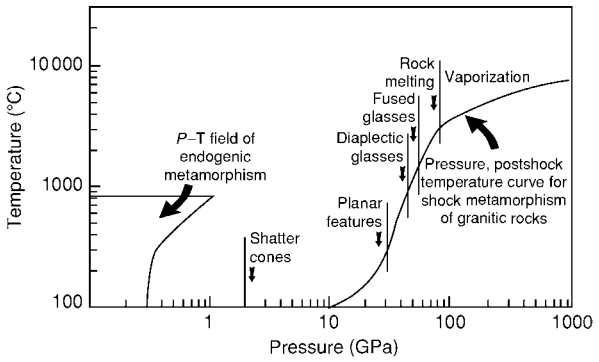


Figure 7 Temperature and pressure range of shock metamorphic effects compared to that of endogenic terrestrial metamorphism. Planar features include planar deformation features and planar fractures. Note scale is logarithmic.

to a series of fault terraces. Interior to the rim lies a down-faulted annular trough, which is partially filled by a sheet of impact-melt rock and/or polymict allochthonous breccia (Figure 6). Only in the central area of the crater is there evidence of substantial excavation of target materials. This central region is structurally complex and can be occupied by a central peak (Figure 5), which is the topographic manifestation of a much broader and extensive area of structurally uplifted target rocks that occurs beneath the centre of complex craters.

With increasing diameter, a fragmentary ring of interior peaks appears, marking the transition from complex craters to impact basins. There have been claims that the largest known terrestrial impact structures, e.g., Chicxulub, Mexico; Sudbury, Canada; and Vredefort, South Africa, have multiring basin forms. Although certain of their geological and geophysical attributes form annuli, it is not clear that these correspond, or are related in origin, to the obvious topographical rings observed, for example, in lunar multiring basins.

A small number of relatively young, and, therefore, only slightly eroded, complex impact structures (e.g., Haughton, Canada; Ries, Germany; Zhamanshin, Kazakhstan) do not have an emergent central peak or other interior topographical expression of a central uplift. These structures are in mixed targets of platform sediments overlying crystalline basement. This difference in form is probably a target rock effect but it has not been studied in detail.

Geology of Impact Structures

Although an anomalous circular topographic, structural, or geological feature may indicate the presence of an impact structure, other terrestrial geological processes can produce similar features. The burden of proof for an impact origin for a specific structure generally lies with the occurrence of shock metamorphic effects. Only small, young, simple structures, where the impacting body has been slowed by atmospheric deceleration, preserve physical evidence of the impacting body. These are generally fragments of iron or stony-iron meteorites. Stony meteorites are weaker and small stones are crushed as a result of atmospheric interaction. Larger impacting bodies (>100–150 m in diameter) survive atmospheric passage with undiminished impact velocity.

On impact, a portion of the impacting body's kinetic energy is partitioned into kinetic energy in the target rocks, which results in the formation of a craterform. The remainder of the impact kinetic energy is partitioned into increasing the internal energy of the target rocks. The target rocks are compressed by the passage of a shock wave and experience extremely high transient pressures and temperatures. For example, the peak pressure experienced in crystalline target rocks by the impact of a stony body, such as a chondritic asteroid, at 25 km s ^{1} is approximately 100 GPa. During shock compression, considerable pressure–volume work is done. On decompression, not all this mechanical work is recovered. This excess work is manifest as waste heat, leading to the heating, melting, and even vapourization of part of the target rocks and the impacting body, destroying it as a physical entity.

The combined effect of this compression and heating is the production of a series of irreversible changes that occur in individual minerals and rocks, which are known collectively as shock metamorphic effects. Shock metamorphic effects are produced at pressures and temperatures well beyond those in endogenic terrestrial metamorphism (Figure 7). The physical conditions upon impact are a function of initial impact parameters. Projectile and target rock type and impact velocity determine peak pressures on

impact, and projectile size determines the absolute radial distance at which particular shock metamorphic effects occur. Shock metamorphic effects are also produced on vastly different time-scales from endogenic metamorphic effects, and disequilibrium is the rule, not the exception.

Impact Melting

Impact-melted lithologies occur as glass bombs in crater ejecta, as dykes within the crater floor and walls, as glassy to crystalline lenses within the breccia lenses of simple craters, or as coherent annular sheets (Figure 8) lining the floor of complex craters. When crystallized, impact-melt sheets have igneous textures, and may, therefore, resemble endogenic igneous rocks. An important textural characteristic, however, of impact-melt rocks is the presence of mineral and rock fragments, which exhibit shock metamorphism to different degrees. The size of such fragments ranges from millimetres to several hundreds of metres and gradational changes in fragment content are observed, with highest concentrations towards the lower and upper contacts of coherent impact-melt sheets.

The composition of impact-melt rocks reflects the wholesale melting of a mix of target rocks, as opposed to partial melting and/or fractional crystallization relationships for endogenous igneous rocks. The composition of impact-melt rocks can be reproduced by a mixture of the various target rock types, in their appropriate geological proportions. Such parameters as $^{87}Sr/^{86}Sr$ and $^{143}Nd/^{144}Nd$ ratios also reflect the preexisting target rocks, although other isotopic systems, e.g., Ar^{39}/Ar^{40}, reflect remelting at the time of impact. In general, even relatively thick impact-melt sheets are chemically homogeneous over distances up to tens of kilometres. Differentiation is not a characteristic of impact-melt sheets (with the exception of the extremely thick, $\geq$2.5 km, Sudbury igneous complex, at the Sudbury Structure, Canada).

Enrichments above target rock levels in siderophile elements and Cr have been identified in some impact-melt rocks. These represent an admixture of up to a few percent of meteoritic material from the impacting body. In some melt rocks, the relative abundances of the various siderophiles have constrained the composition of the impacting body to the level of meteorite class. In other melt rocks, no siderophile anomaly has been identified. The latter may be due to the inhomogeneous distribution of meteoritic material or to differentiated and, therefore, non-siderophile-enriched impacting bodies, such as basaltic achondrites. High-precision chromium and osmium-isotopic analyses have also been used to detect a meteoritic signature at terrestrial impact structures.

Fused and Diaplectic Glasses

Shock-fused minerals are characterized by flow structures and vesiculation (Figure 9). Peak pressures required for shock melting of single minerals are 40 to 60 GPa, for which postshock temperatures exceed the melting points of typical rock-forming minerals. Under these conditions, the minerals in the rock will melt immediately and independently, after the passage of the shock wave. Melting is mineral selective, producing unusual textures in which one or more minerals show typical melting features; whereas others, even juxtaposed ones, do not. One of the most common fused glasses observed at terrestrial impact structures is that of quartz, i.e., lechatelierite.

Conversion to an isotropic, dense, glassy phase is a shock metamorphic effect unique to framework silicates. These phases are called diaplectic (from the Greek 'destroyed by striking') glasses, and are produced by breakdown of long-range order of the crystal lattice without fusion. Based on shock recovery experiments, the formation of diaplectic glass occurs between 30 and 45 GPa for feldspar and 35 to 50 GPa for quartz. The morphology of the diaplectic glass is

Figure 8 Approximately 80 m high outcrop of coherent impact melt rocks at the Mistastin complex impact structure, Canada. These rocks resulted from the melting of the target rocks by shock pressures in excess of approximately 60 GPa or 600 kbars (Figure 7).

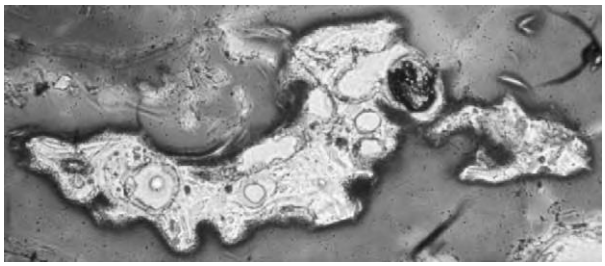


Figure 9 Photomicrograph of fused glass (lechatelierite) of the mineral quartz, from the Ries impact structure, Germany. Field of view 2.5 mm.

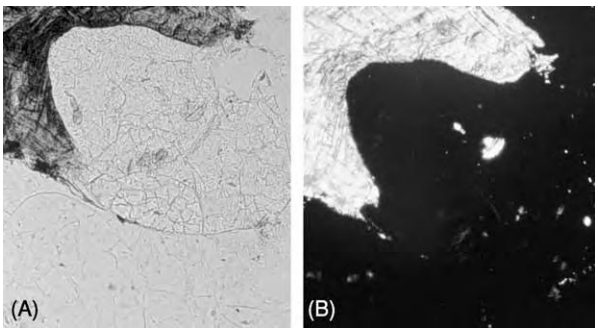


Figure 10 Diaplectic mineral glasses. Photomicrograph of granite from the Mistastin complex impact structure with biotite (upper left), quartz (centre), and plagioclase (bottom and far right). Field of view 2.5 mm. (A) Plane light. (B) Crossed polars. Quartz and plagioclase (black) are isotropic, while retaining original mineral shape; i.e., they are diaplectic glasses. This requires approximately 40 GPa shock pressure (Figure 7).

the same as the original mineral crystal and shows no evidence of fluid textures. Maskelynite, the diaplectic form of plagioclase, is the most common example from terrestrial rocks; diaplectic glasses of quartz and of alkali feldspar also occur (Figure 10).

High-Pressure Polymorphs

Shock can result in the formation of metastable polymorphs, such as stishovite and coesite from quartz and diamond and lonsdaleite from graphite. Coesite and diamond are also products of endogenic terrestrial geological processes, including high-grade metamorphism, but the paragenesis and the geological setting are completely different from that in impact structures. In terrestrial impact structures, stishovite and coesite polymorphs occur in small or trace amounts, as very fine-grained aggregates, and are formed by partial transformation of the host quartz.

Planar Microstructures

The most common documented shock-metamorphic effect is the occurrence of planar microstructures in tectosilicates, particularly quartz (Figure 11). The utility of planar microstructures in quartz is a function of the ubiquitous nature of the mineral and its stability, in the terrestrial environment, and the relative ease with which they can be documented optically. Planar deformation features in quartz have various orientations and are produced under pressures of ~10 to ~35 GPa (Figure 7), with their crystallographic orientation providing a measure of the recorded shock pressures.

Shatter Cones

Shatter cones (Figure 12) are the only known diagnostic shock effect that is megascopic in scale. They

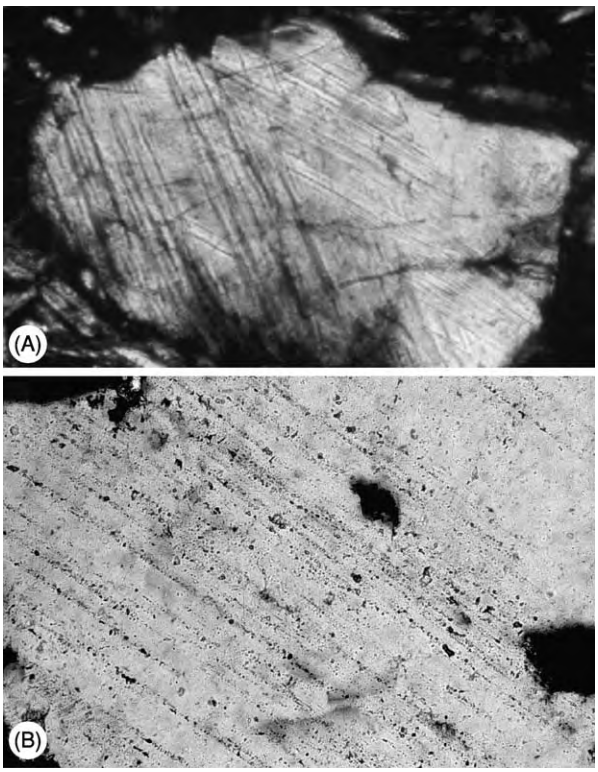


Figure 11 Photomicrographs of planar deformation features in quartz crystals: (A) nondecorated features, New Quebec impact structure, Canada; and (B) decorated features, Charlevoix impact structure, Canada. Field of view 1 mm crossed polars.

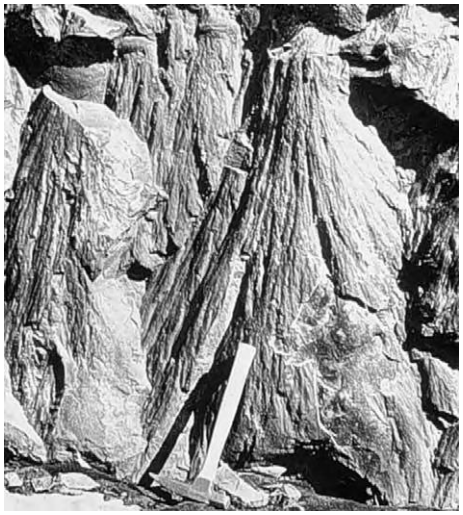


Figure 12 Shatter cones in quartzites at the Sudbury impact structure, Canada.

are unusual, striated, and horse-tailed conical fractures ranging from millimetres to tens of metres in length. Shatter cones are initiated most frequently in rocks that experienced moderately low shock pressures, 2–6 GPa, but have been observed in rocks

that experienced ~25 GPa. Shatter cones are best developed in fine-grained, structurally isotropic lithologies, such as carbonates and quartzites. They do occur in coarse-grained crystalline rocks but are less common and more poorly developed.

Impacts and Earth Evolution

The basic working hypotheses for the formation of the terrestrial planets is the accretion of small bodies by collision and the subsequent growth to larger planetary embryos. This early stage of planetary accretion is believed to be followed by a stage involving truly giant impacts between the embryos. In the context of planetary formation, therefore, impact is the most fundamental process (*see* **Solar System:** Moon).

In the case of the Earth, the last embryonic planetary collision may have resulted in the formation of the Earth's Moon. Numerical simulations of a glancing impact of a Mars-sized body with the proto-Earth result in the formation of an Earth-orbiting accretionary disk from which a Moon-like body can form. The key is that the bulk of the material that forms the accretionary disc is originally terrestrial or impacting body mantle material. Condensation and reaccretion result in volatile and siderophile-depleted material, which goes to form the Moon. Although the impact hypothesis for the origin of the Moon is consistent with the constraints of the dynamics of the Earth–Moon system and the geochemical nature of the Moon, it remains a model. The consequences of such a giant impact for the proto-Earth, however, would have been severe. They would have included massive remelting and the likely loss of the original atmosphere.

After planetary formation, the subsequent high rate of bombardment by the remaining tail of accretionary debris is recorded on the Moon and other terrestrial planets that have preserved portions of their earliest crust. This is not preserved on Earth, because of the high level of endogenic geological activity and the resultant relatively young surface of the Earth. In the case of the Moon, a minimum of 6000 craters with diameters >20 km are known to have been formed during the early period of approximately 4.5–3.8 Ga. There are also approximately 45 known impact basins ranging in diameter from Bailey at 300 km to the south pole Aitken basin at 2600 km. Throughout geological time, the Earth has received more impacts than the Moon. The Earth is a physically larger target for incoming bodies and has a much larger gravitational cross-section to attract incoming bodies. There has been considerable speculation as to the potential effects of these basin-sized impacts on the Earth. By analogy with the lunar case, it is likely that few terrestrial surface rocks would have survived intact through this period of heavy bombardment. Basin-sized impacts on the early Earth may have also affected the existing atmosphere, hydrosphere, and the potential development of the biosphere.

The oldest known terrestrial impact structures are Sudbury and Vredefort. They have reconstructed original diameters of approximately 250–300 km, and ages of 1.85 and 2.0 Ga, respectively. Although it is expected that such large events would have had a deleterious effect on the climate and biosphere, no direct evidence is known at present. These structures, however, also affected the local geology in a manner that is to human benefit. Both impact structures are the sites of world-class ore deposits.

An extensive literature concerning the evidence for a major impact event at the K–T boundary and its association with a mass extinction in the terrestrial biosphere 65 Ma exists. The unequivocal physical evidence for impact contained in K–T boundary deposits consists of planar microstructures in quartz, feldspar, and zircon, and the occurrence of stishovite, impact diamonds, high-temperature magnesioferrite spinels (believed to be vapour condensates), and various melt spherules, generally altered, but including the tektite-like glass spherules in Haiti and other Caribbean sites. The chemical evidence consists of a global siderophile anomaly in K–T boundary deposits, indicative of an admixture of meteoritic material.

Although originally contentious, there is now little doubt that the Chicxulub structure in the Yucatan peninsula, Mexico is the K–T impact structure. Chicxulub is buried by some 1 km of younger sediments but evidence for impact, in the form of planar microstructure quartz and feldspar in deposits interior and exterior to the structure, as well as impact-melt rocks, has been documented. In addition, the geochemistry of K–T tektite-like glasses from Haiti matches the mixture of lithologies found at the Chicxulub site. Isotopic ages for the impact-melt rocks at Chicxulub of 64.98 ± 0.05 Ma are indistinguishable from the K–T tektites at 65.07 ± 1.00 Ma (*see* **Mesozoic:** End Cretaceous Extinctions).

The original hypothesis for the killing mechanism for the K–T mass extinction suggested global darkening and cessation of photosynthesis due to ejecta in the atmosphere. Soot has also been identified in K–T deposits and ascribed to global wildfires. A considerable thickness of anhydrite ($CaSO_4$) occurs in the target rocks at Chicxulub. Impact heating of anhydrite would produce sulphur aerosols in the atmosphere. Modelling suggests that these sulphur aerosols would reduce light levels below those needed for photosynthesis for 6–9 months. In addition, if most of the aerosols were in the form of SO_2, solar

transmission would drop to 10–20% of normal (a cloudy day) for a few decades. It may be, therefore, that the devastating effects of the K–T impact are due to the character of the target rocks at Chicxulub. Whatever the case, the temporal association of an extremely large impact crater with a worldwide ejecta layer and a global mass-extinction event is well established. The cause–effect relationship is less well established and is the subject of current research.

The frequency of K–T-sized events on Earth is on the order of one every 100 ± 50 My. Smaller, but still significant, impact events occur on shorter time-scales. Dust loadings from the formation of impact craters as small as 20 km could produce light reductions and temperature disruptions on relatively short time-scales. Such impacts occur on Earth with a frequency of a few every million years or so and are not likely to have an effect upon the biosphere. The most fragile component of the present environment, however, is human civilization. There is little doubt that if it lasts long enough it will suffer severely or may even be destroyed by an impact event. The greatest threat (the product of probability and the expected death and destruction) is likely from ocean-wide impact-induced tsunamis, which occur on time-scales of tens of thousands of years. For example, model calculations indicate that the impact of even a relatively small body (approximately 200 m in diameter) would result in a wave in the open ocean that would still be 10 m high 1000 km from the point of impact. Potential impact events do, in fact, occur on human time-scales. For example, the Tunguska event on 30 June 1908 was due to the atmospheric explosion of a relatively small, <100 m, body at an altitude <10 km.

Terrestrial impact structures are the manifestation of extremely high-energy, transient geological events and, as such, have resulted in unusual local geological environments. Some of these have produced significant economic deposits. About 25% of known terrestrial impact structures have some form of economic deposit, and about half of these are currently exploited or have been exploited in the recent past. They range from local to world-class (e.g., reserves of 1.6 billion tonnes Ni–Cu ores at Sudbury). They also include hydrocarbon deposits (e.g., reserves of 50 million barrels of oil and 60 billion ft^3 of gas at Ames, Oklahoma, USA). The Ames impact structure produced the structural trap and also provided the source rocks, which are locally developed postimpact oil shales within the crater. In addition hydrocarbon production at Ames includes nontraditional reservoirs, such as the fractured crystalline rocks of the central uplift. The world-class Campeche Bank oilfield, Gulf of Mexico, produces most of its hydrocarbons from K–T breccias, with 10% porosity, resulting from seismic activity related to the nearby Chicxulub impact. Proven reserves are 30 billion barrels of oil and 15 trillion ft^3 of gas and exceed the entire onshore and offshore US reserves.

Summary Remarks

The detailed study of impact events on Earth is a relatively recent addition to the spectrum of studies engaged in by the geological sciences. More than anything, it was preparations for and, ultimately, the results of the lunar and planetary exploration programme that provided the impetus and rationale for their study. The terrestrial record of impact has made important contributions to our understanding of impact processes. Answers to many questions are known to, at least, the first order. Many details require clarification, however, and some problems, for example, the characteristics of superficial impact deposits, such as ejecta, are difficult to address in the active terrestrial geological environment.

It is apparent that impact can no longer be considered a process of interest only to the planetary community. It is a process that has fundamentally affected terrestrial evolution. For example, without the K–T impact, the present-day biosphere may have been quite different. Similarly, without a Mars-sized impact forming the Moon, and the resulting tidal forces on the Earth, one can only speculate how the littoral zone, the most important area in terrestrial ecosystem, would have evolved and been populated. Impact is the most catastrophic geological process known and is fundamental to the nature of the Solar System. Impact events have happened on Earth throughout geological time and will happen again. Although the occurrence of a large impact event has a low probability on the time-scale of human civilization, the consequences of its occurrence could be globally disastrous.

See Also

Earth Structure and Origins. **Engineering Geology:** Natural and Anthropogenic Geohazards. **Mesozoic:** End Cretaceous Extinctions. **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Mercury; Venus; Moon; Mars. **Tektites**

Further Reading

Australian impact structures (1996) *AGSO Journal of Australian Geology and Geophysics* 16(4): 371 625.

Frankel C (1999) *The End of the Dinosaurs*. Cambridge: Cambridge University Press.

French BM (1998) *Traces of Catastrophe: A Handbook of Shock Metamorphic Effects in Terrestrial Meteorite*

Impact Structures. Lunar and Planetary Institute Contribution 954. Houston: Lunar and Planetary Institute.
Grady MM, Hutchison R, McCall GJH, and Rothery RA (eds.) (1998) *Meteorites: Flux with Time and Impact Effects*. Geological Society of London Special Publication 140. London: Geological Society of London.
Gehrels T (ed.) (1994) *Hazards due to Comets and Asteroids*. Tucson: University of Arizona Press.
Hodge PW (1994) *Meteorite Craters and Impact Structures*. Cambridge: Cambridge University Press.
Hoyt WG (1987) *The Coon Mountain Controversies*. Tucson: University of Arizona Press.
Melosh HJ (1989) *Impact Cratering: A Geologic Process*. Oxford: Oxford University Press.
Plado J and Pesonen LJ (eds.) (2002) *Impacts in Precambrian Shields*. New York: Springer Verlag.
Plucker Ehrenbrink BE and Schnitz B (eds.) (2001) *Accretion of Extraterrestrial Matter throughout Geologic Time*. Norwell, MA: Kluwer Academic Press.
Spudis P (1993) *The Geology of Multi ring Impact Basins*. Cambridge: Cambridge University Press.

INDIAN SUBCONTINENT

A B Roy, Presidency College, Kolkata, India

Introduction

The Indian Sub-Continent constitutes a distinctive geographic entity: the countries included in the Sub-Continent, including Bangladesh, India, Nepal, and Pakistan, virtually cut off from the rest of Asia by lofty mountain chains. Almost half of the Sub-Continent's boundary in the north is bordered by the Himalayas and its associated branches: the Sulaiman and Kirthar Ranges to the Hindukush in the northwest, and the Naga Hills and the Arakan Yoma constituting the Indo-Myanmar Arc (**Figure 1**). The Himalayas extend for over 2500 km, from the Pamir in the west to the Mismi Hills in the east. The landscape changes sharply on the southern side of the great mountain ranges, where the high mountain terrain descends down to a huge plain land of unconsolidated sediments, known as the Indo-Gangetic Alluvial Plain (IGAP). The IGAP, which constitutes the enormous flood plains of the Indus and the Ganga River systems, also includes the narrow basin of the Brahmaputra River in the east and the Thar Desert in the west (including the North Gujarat Plain). South of the IGAP lies the rocky landmass of Peninsular India.

Except for the Himalayas and associated mountain ranges, the Indian Sub-Continent is traditionally considered as a shield of Precambrian rocks with a younger cover. The geophysical data, especially the seismic, Bouguer gravity anomaly, and heat flow patterns, suggest that much of the shield area attained considerable 'mobility' during post-Precambrian time. In this respect at least, the Indian Shield appears distinctly different from the better-known shield areas like the Canadian or the Ukrainian Shields.

The history of geological evolution of the Indian Sub-Continent is quite long and complex and, broadly speaking, took place in two stages. The first stage covered the entire Precambrian, which was the period of growth and final cratonisation. During the second stage of its evolution, the cratonised Indian Shield underwent considerable reconstitution that ultimately produced the present-day geomorphology as well as the tectonic character of the region.

The following description discusses the regional geology of the Sub-Continent mainly in terms of the history of evolution of the crust, starting from the evolution of the Archaean basement and the formation of Proterozoic basins to the different stages of the Phanerozoic reconstitution. Evidence for all of these come from smaller regions, which preserved these stages of crustal development. The description of regional geology of the Indian Sub-continent will be divided into the following heads:

i. Precambrian Indian Crust;
ii. Geology of the Gondwana basins;
iii. Between the Jurassic break-up and Himalayan collision.
iv. Geology of the Himalayas;
v. Quaternary sedimentation and neotectonics.

Precambrian Indian Crust

The Peninsular India lying south of the Indo-Gangetic Alluvial Plain is an old landscape, a considerable part of which is covered by Phanerozoic rocks, the Gondwanas, Deccan Traps, marine Mesozoic-Tertiary formations, and Recent alluvium. Tectonically speaking, the peninsula is a shield area that has remained free of any orogenic deformation since the Cambrian. The entire region was considered a 'terra incognita', even up to about three decades ago. Detailed studies in several interdisciplinary fields helped to characterise

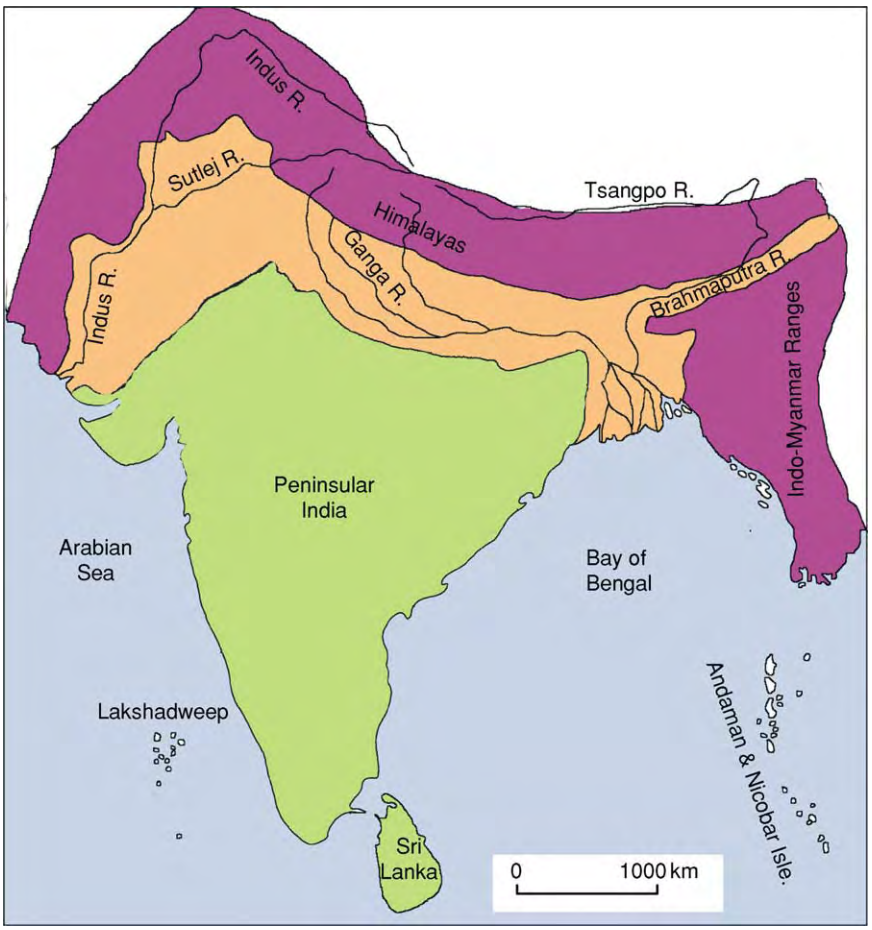


Figure 1 Physical map of the Indian Sub Continent.

the Indian Shield as a coherent unit, comprising a number of Precambrian crustal blocks sutured together along joins marked by prominent lineaments. The Precambrian rocks occur as an isolated fault-bounded rectangular block at the Shillong Plateau. Tectonic affinity of this small block east of the Bengal Basin remains unclear. The crustal blocks include four cratons (Dharwar, Bastar, Singhbhum, and the Aravalli-Bundelkhand) and two granulite terrains (Eastern Ghats and Southern Granulite Terrains) and are characterised by distinctive evolutionary history and metallogenic patterns (Figure 2). The most ancient rocks constituting the basement in each craton are between 3.3 Ga and 3.5 Ga old. Before becoming a craton, all these crustal blocks were mobile belts. Typical cratonic platformal sediments cover large areas in the different cratons, apart from the Singhbhum and the granulite belts.

Dharwar Craton

Divided into the Eastern and the Western belts, this craton is marked by the evolution of a number of greenstone belts, called the schist belts, over sialic basements. No universally acceptable stratigraphical succession is known from the craton, which dates from the Archaean (3.5 Ga to 2.5 Ga). The principal structural trend of the Dharwar Craton is north-south, showing an eastward convexity (Figure 3). A change of metamorphic grade is indicated by the presence of very low grade rocks in the north to a transitional granulite facies in the south. A number of mineral deposits occur in the different schist belts. Important ones are iron-ores (as BIFs), chromite, nickel (as sulphides), copper, and gold.

The eastern Dharwar Craton by comparison shows more extensive Late Archaean reconstitution and granitic activities. The chronology of growth of the Dharwar Craton is given in Table 1.

Bastar Craton

The Bastar Craton also known as the Bhandara Craton or the Bastar-Bhandara Craton, occupies a quadrangular area with well-defined boundaries. Lithologically, the craton comprises patches of Archaean supracrustal rocks engulfed in a vast expanse of gneiss-granite of various ages, well-defined

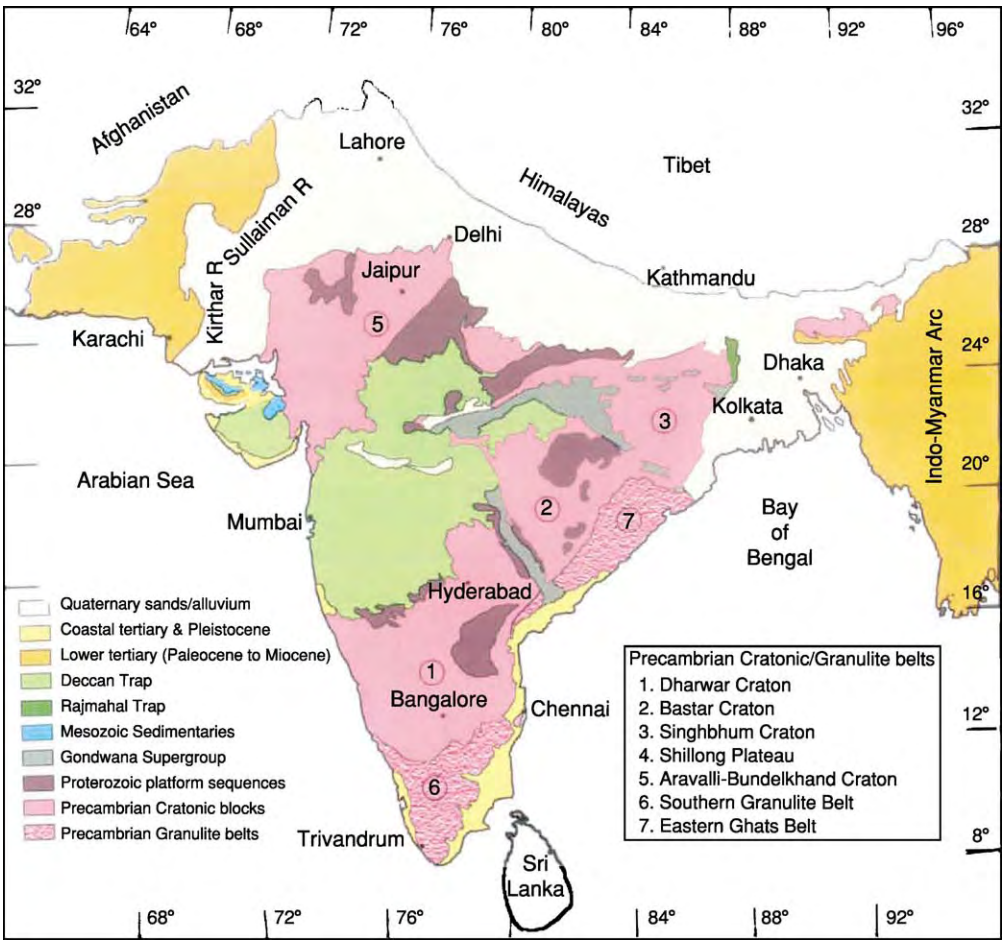


Figure 2 Generalised geological map of the Indian Sub Continent.

Proterozoic fold belts (comprising the Sausar, Sakoli, and Dongargarh Groups), and cratonic platformal cover rocks of the Chattisgarh Group and its equivalents. The Deccan Trap covers it in the north (Figure 4).

The antiquity of the Bastar Craton is indicated by the occurrence of 3.5 Ga old trondhjemite gneiss. The oldest supracrustal rocks host large deposits of iron-ore (of BIF type). Barring the general high-grade metamorphic character of rocks, the supracrustals show features of dismembered greenstone sequences. There are occurrences of granulite-facies rocks in the Bastar Craton. The geochronological data, although poor suggests a number of episodes of reconstitution of the basement rocks during both the Archaean and the Proterozoic.

Important mineral deposits in the Bastar Craton include iron-ore in the Bailadela Group, manganese ores in the Sausar Group, and copper ores in the Malanchkhand Granite.

Singhbhum Craton

The most significant aspect of the recent works carried out during the last four decades is the recognition of an old Archaean nucleus, which was enlarged through the accretion of younger bodies around it. The components making up the Archaean nucleus include: (i) the oldest Archaean supracrustals, called the Older Metamorphic Group (OMG), and a variety of tonalitic gneisses that intruded it; (ii) massifs of the Archaean granitoids; and (iii) the supracrustals called the Iron Ore Group (IOG). The pre-IOG events are shown by the addition of juvenile materials during 3.3 and 3.0 Ga. The presence of a pre-OMG basement of still older age is inferred, based on the geochemical character of the OMG ortho-amphibolites. The granitic activities at around 2.6 Ga marked the closure of Archaean history in the craton. The stratigraphic status of the very lowly metamorphosed platformal sequence, the Kolhan Group, which occurrs in close association with the IOG, remains a matter of debate.

North of the Archaean nucleus lies arcuate Proterozoic fold belts (also known as the North Singhbhum mobile belts) which show a prominent northward convexity (Figure 5). Close to the southern boundary lies the prominent intracrustal shear zone that

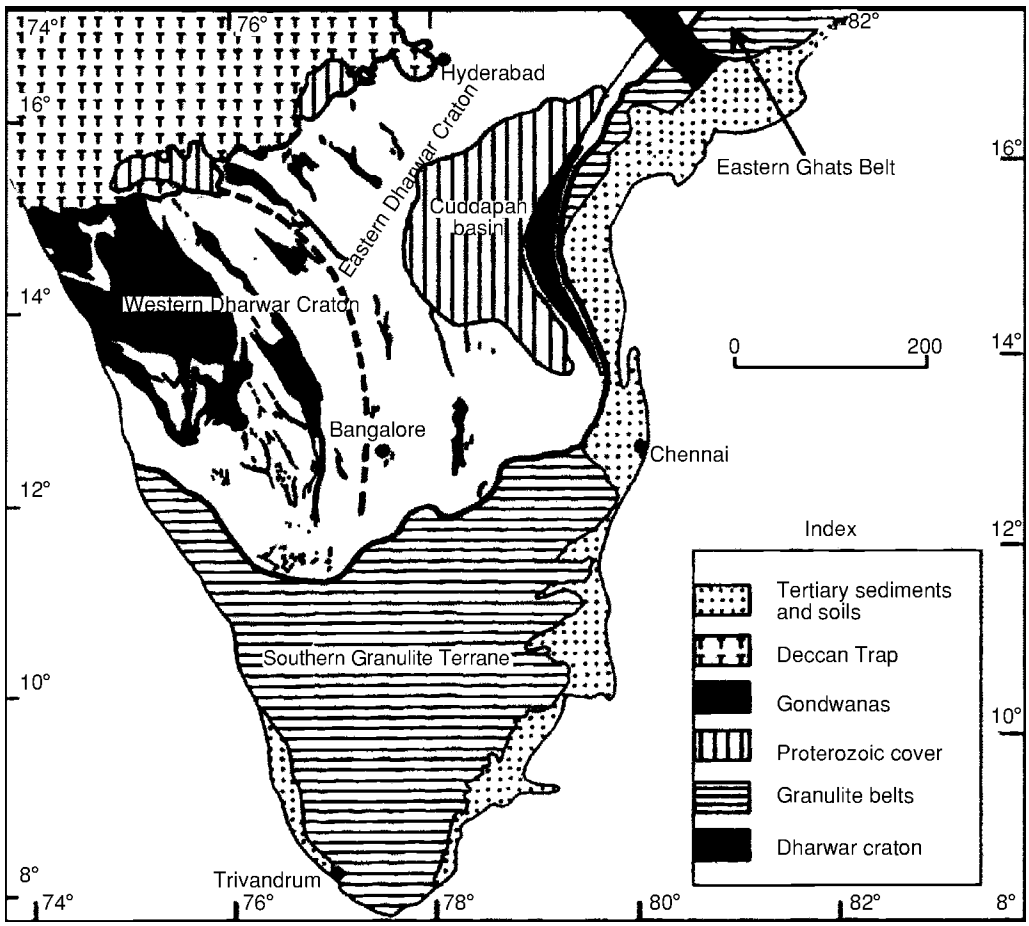


Figure 3 Generalised geological map of the Dharwar Craton and the Southern Granulite Terrain (adapted with permission from Radhakrishna 1983; © Geological Society of India, Bangalore).

Table 1 Chronology of evolution of crust of the Dharwar Craton (adapted from Radhakrishna and Vaididhyanadhan, 1997)

2.5 Ga	Closepet granite and granulites
	Dharwar type schist belts (2.9 2.6 Ga) (Holenarasipur, Javanhalli, Bababudan, Chitradurga, Shimoga, Sandur)
3.0 Ga	Kolar type auriferous schist belt (Kolar and Hutti)
>3.0 Ga	Sargur Group
~3.5 Ga	Formation of earliest basement

virtually wraps the northern half of the oval outcrop of the Singhbhum nucleus. There is hardly any unanimity on the stratigraphic succession of the Proterozoic fold belts. A simplified succession of the Singhbhum Craton is given in Table 2.

Aravalli-Bundelkhand Craton

Separated by the Son-Narmada Lineament from the three major cratonic blocks, the Dharwar. Bastar, and the Singhbhum, this large cratonic block comprises two distinctive crustal blocks, the Aravalli and the Bundelkhand, divided by an important tectonic boundary, the Great Boundary Fault (Figure 6). The Bundelkhand block comprises a gneissic basement formed between 3.3 and 2.5 Ga. The Proterozoic supracrustals include the Gwalior, the Bijawar, and the Mahakoshal Groups. The block did not undergo any tectothermal event later than 1850 Ma. By contrast, the Aravalli part of the craton shows continuous geological orogenic histories that ended at around 850 Ma (Table 3).

This craton contains rich deposits of phosphorite, in addition to lead-zinc and copper ore bodies.

Southern Granulite Terrane

The popular view on the evolution of this terrane assumes a model of amalgamation of microterranes. The network of dominantly dextral ductile shear zones emanating from the Cauvery Shear zones is considered central to the problem of assembly of Precambrian crust of the SGT (Figure 3). Geochronological data suggest the formation of granulite crust mostly during the Archaean, followed by crustal

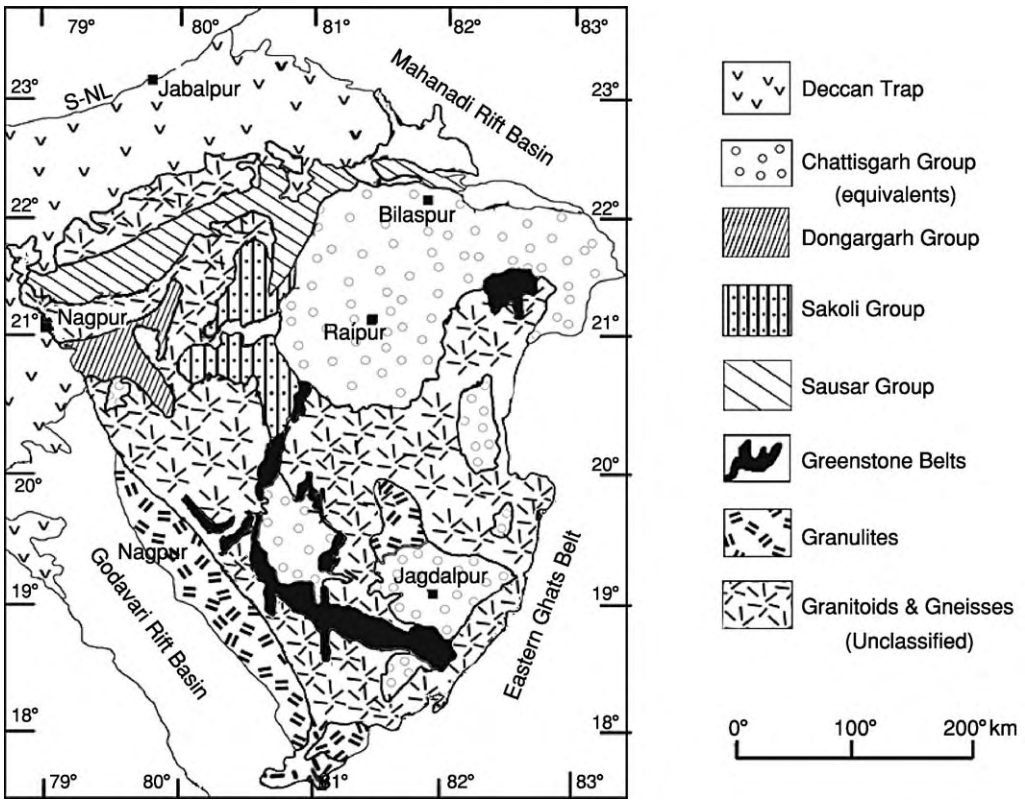


Figure 4 Geological map of the Bastar Craton. S NL Son Narmada Lineament (adapted with permission from Roy *et al.*, 2000; © Geological Survey of India).

reworking during the Palaeoproterozoic and (in some only) during the Neoproterozoic.

Eastern Ghats Granulite Belt

This linear belt of deep sections is subdivided into four crustal provinces with widely different geological evolutionary histories. The northernmost part of the belt bordering, the Bastar Craton, formed in the late Archaean. The other domains provided indication of reconstitution by tectonothermal events in Mesoproterozoic, Grenville, and Pan-African times. Considerable similarity is reported between the Eastern Ghats Belt and the Rayner Province of East Antarctica in terms of their geological evolution.

Cratonic Basins

Several cratonic basins have developed in different parts of the Peninsular Shield area, of which the two most important ones are the Cuddapah Basin and the Main Vindhyan Basin. Both basins evolved quite early, in all probability during the late Palaeoproterozoic, signifying cratonisation of the Archaean crust prior to the basin opening. A break in sedimentation is indicated by the intrusion of kimberlite pipes in the older sequence and the deposition of diamond-bearing conglomerate at the base of the upper upper formations. The youngest cratonic basin developed in the west of the Aravalli-Bundelkhand Craton over a basement of the late Neoproterozoic plume-related volcanic to plutonic Malani Group.

Geology of Gondwana Supergroup

Definition, Classification, and Distribution

After a break in sedimentation for over 200 million years between the Ordovician and the Early Permian, the deposition of sediments in the Indian Sub-Continent started with the formation of tillites and glacial boulder beds in close association with Permian marine beds. This was accompanied by the deposition of fluvial and fluvio-lacustrine sediments in linear intracontinental rift basins. These sediments, along with intercalated plant remains that ultimately turned into coal seams, constitute the Gondwana Supergroup.

Precise definition of the Gondwana formations as formal stratigraphic units has suffered because of the inclusion of rocks deposited in diverse geological conditions into its ambit, and also because of the overemphasis given to the floral evidence. The Gondwana Supergroup has been redefined to include

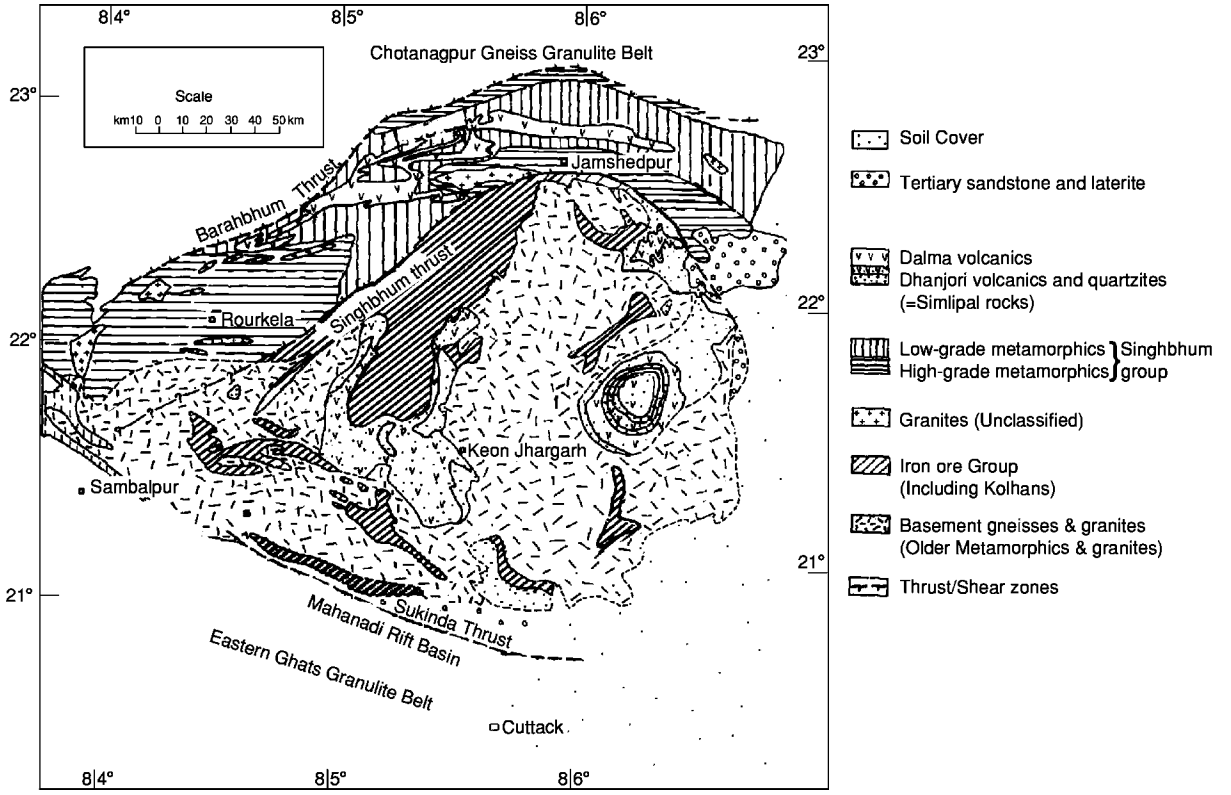


Figure 5 Geological map of the Singhbhum Craton (adapted with permission from Mahadevan TM, 1994; © Geological Society of India, Bangalore).

dominantly continental rift-basin deposits formed during the Early Permian and Middle Jurassic with minor marine inputs. Adoption of such a definition helps in the finer tuning of the stratigraphic succession by removing the contemporaneous and other Jurassic-Creataceous formations, which are correlatable when the Jurassic Gondwana broke up and the Cretaceous-Eocene plume outburst occurred. Recent studies favour adoption of smaller subdivisions of the Gondwana Supergroup instead of the two- and three-fold classifications based on floral assemblages.

Redefinition of the Gondwana Supergroup helps to identify three belts of the Gondwana basins. These are:

i. E–W trending Damodar Valley basins. Included in this belt is the narrow sub-parallel belt of subsidiary basins occurring north of the main belt.
ii. NW–SE trending Son-Mahanadi Valley basins. The belt widens and thickens in a northwesterly direction, meeting the extension of the Damodar Valley basins.
iii. NW–SE trending Pranhita-Gondwana Valley basins.

Stratigraphic classification of the Gondwana formations developed in the three belts is given in Table 4. The basal Talchir unit includes glacial tillites with shale and shale-siltstone rhythmites. All the other units comprise sandstone with shale as the dominant lithology. A number of coal seams (along with carbonaceous shale) occur in the Karharbari, Barakar, and Ranigani formations. The Barren Measure, which occurs between the Barakar and the Raniganj is (as the name implies) devoid of any coal seam. The Panchet and the Supra-Panchet are also coal-free sequences. The latter is termed the Mahadeva Formation in the Satpura Range.

Sedimentation, Basin Morphology and Evolution

Evidence from fossil flora and palynological studies indicate repeated changes in climate during the Gondwana sedimentation, which is also reflected in the nature of repetitive cycles of lithofacies. After the initial cold glacial condition, the climate warmed up along with increased humidity, which varied between medium and very high. This highly humid condition favoured enormous growth of plants. The en masse deposition of plant remains ultimately led to the formation of coal seams. The depositional condition changed to shallow freshwater with increased evidence of desiccation and salinity during later Panchet time.

Palaeoslope studies in the Damodar Valley provide inconsistent results, but a general north-northwesterly to northerly palaeoslope is suggested for these

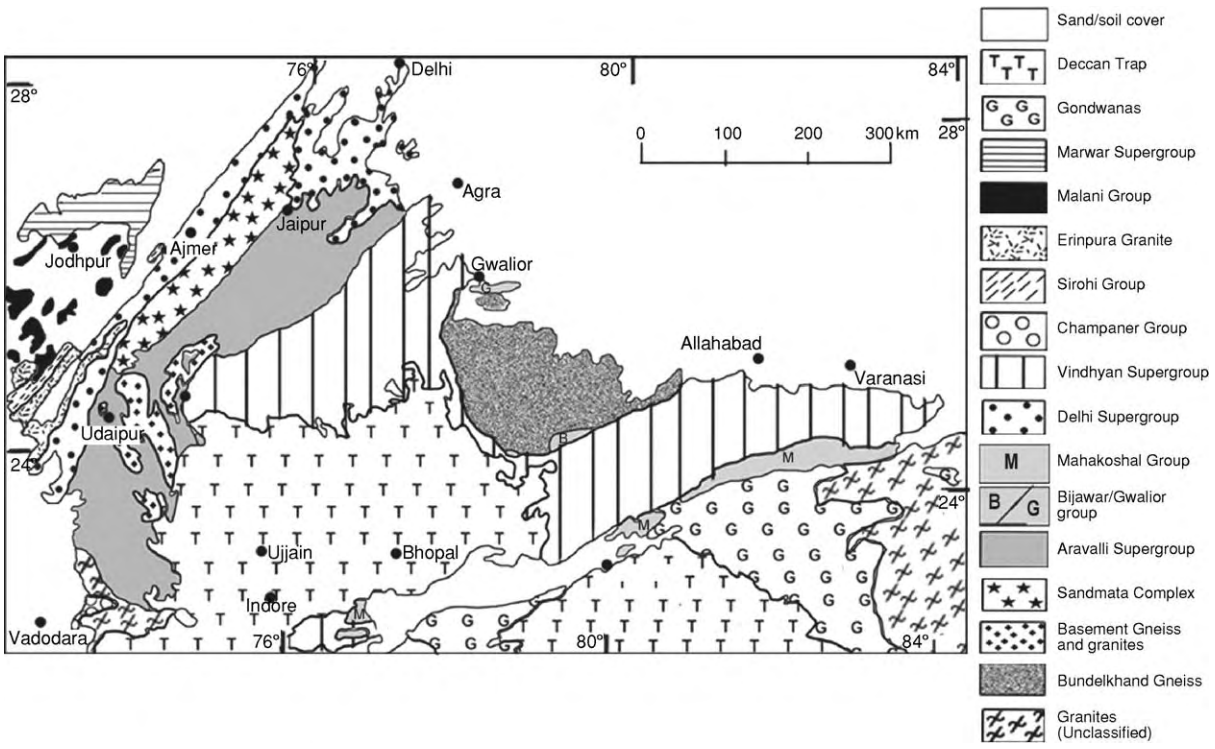


Figure 6 Geological map of the Aravalli Bundelkhand Craton (adapted from Soni *et al.*, 1987; © Geological Society of India, Bangalore and Roy and Jakhar, 2002; © AB Roy, 2002).

Table 2 Generalised stratigraphic succession of Precambrian rocks in the Singhbhum Craton (adapted from Mahadevan, 1993 and Mukhopadhyay, 2001)

Phanerozoic covers (soils and laterites and Gondwana formations)	
Dolerite dikes and sills	
Gabbro Anorthosite suites	
Kolhan Group[a] (?)	
Dalma lavas, Dhanjori lavas with quartzite[b] (Simplipal lavas and quartzites)	
Singhbhum Group	Pelitic and arenaceous metasediments with mafic sills
Iron Ore Group	Upper Shales with sandstones and volcanics Banded hematite jasper with iron ore Mafic tuffs and lavas Tuffs, felsic volcanics and tuffaceous shales Sandstone and conglomerate (local)
Granitoids	(3.0 2.6 Ga)
Older metamorphic tonalitic gneisses	(3.3 3.4 Ga)
Older Metamorphic Group	Pelitic schist, quartzite, para and ortho amphibolite
Sialic basement of unknown age	

[a]Some authors consider the Kolhan Group as a component of the Archaean Iron Ore Group.
[b]Dhanjori Group is placed by some below the Singhbhum Group.

Table 3 Precambrian tectonostratigraphic framework of the Aravalli cratonic block (adapted from Roy and Jakhar, 2002)

Marwar Supergroup (cratonic sediments)	Youngest Proterozoic
Malani Group (volcanics, granites, sediments)	780 680 Ma
Erinpura Granites	~850 Ma
Sirohi Group (metasediments)	
Intrusion of Gabbro diorites	~1000 Ma
Synorogenic granites in the Delhi Fold Belts	~1450 Ma
Delhi Supergroup (metasediments and metavolcanics)	
Granulite exhumation in Sandmata Complex	1725 1625 Ma
Synorogenic granites (Darwal Granite)	~1850 Ma
Aravalli Supergroup (metasediments and minor metavolcanics)	
Basement Gneiss (Mewar Gneiss Complex) and granites	~2500 ~3300 Ma

basins. In contrast, the sedimentation both in the Pranhita-Godavari and Son-Mahanadi Valley basins possibly took place along linear fluvial basins of river systems draining southeast to northwest in the direction of the palaeoslope.

Gondwana sedimentation during the initial stage (primarily during the Talchir sedimentation) took place in eroded topographic depressions. Based on

Table 4 Generalised correlation of the Indian Gondwana formations (adapted from Shastri *et al.*, 1977)

Age	*Damodar Valley*	*Mahanadi Son Valley*	*Pranhita Godavari*
JURASSIC			
Lower			Kota
TRIASSIC			
Upper	Supra Panchet	Parsora Mahadeva	Dharmavaram
			Maleri
Middle		Pali	
			Bhimavaram
Lower	Panchet		Yerapalli
			Mangali
PERMIAN			
Upper	Raniganj Barren Measure	Kamthi	Kamthi
			Motur
Lower	Barakar	Barakar	Barakar
	Karharbari	Karharbari	
	Talchir	Umaria Talchir	Talchir

the similarities in sedimentary succession in several of the now-isolated basins, the presence of a master basin for the deposition of the Gondwana sediments has been suggested. Faulting, as well as topographic relief during the post depositional phase, presumably caused the varying morphology of the individual basins. Popular opinion, however, insists on a rift origin of the Gondwana basins, which at a later stage developed into half or full grabens. The progressive, as well as repetitive movements along block faulted basement slabs underlying the basin, controlled the sedimentation in different Gondwana basins. The present-day basin geometry is a combined effect of faulting in three stages, at the initiation, during, and post-dating sedimentation.

Between Jurassic Breakup and the Himalayan Collision

The Gondwana sedimentation, which began in the Permian, continued until the Lower Jurassic. The next major global event that grossly affected Indian continental block was the breakup of Gondwana at around 165 Ma. The initial separation resulted in marine incursions and deposition of sediments in western Rajasthan and in the Kachch region along WNW–ESE trending rift basins. The deposition of continental sediments, which had earlier stopped in different Gondwana basins before the Lower Jurassic was also resumed, at least in certain cases. The Gondwana breakup event is also responsible for the development of arrays of fracture systems in the Indian Continental block (Figure 7). Geomorphologically expressed as lineaments, these fracture systems, which developed either as new sets of fractures or as reactivated old tectonic grains, helped to significantly change the geophysical character of the Indian crust in a variety of ways.

As the Indian continental block (along with Madagascar and Seychelles, and Antarctica remaining attached to it) moved northward following the dismemberment of Gondwana, it was affected successively by the outbursts of four plume heads centred at Marion, Reunion, Crozet, and Kerguelen Islands. The manifestations of the Crozet Plume outbursts are virtually unknown. The Marion Plume outbursts resulted in the separation of Madagascar from the Indian continental block during 80 to 90 Ma (Figure 8). Evidence for this comes from the occurrence of 88–90 Ma old acid as well as mafic rocks in different parts of central and North Kerala, St Mary's Island off the Karnataka coast and also from Madagascar. The Rajmahal Traps and the Sylhet Traps are the manifestations of Kerguelen Plume activities, which lasted from 130 Ma to 110 Ma. The plume caused the separation of Antarctica from the Indian continental block and induced Cretaceous marine ingress both in the south-eastern part of the peninsula as well as in the south-east of the Bengal basin.

The timing and the passage of impingement of the Reunion Plume are well recorded in the form of different features that developed between Kohistan in the north to the Lakhsadweep-Maldive Islands in the south. The important features that developed during between 70 Ma and 64 Ma include:

i. the formation of narrow sedimentary basins, locally having volcanic inputs (many of which turned into oil-bearing formations) in western Rajasthan and northern Gujarat;
ii. extensive mafic (continental tholeiite basalt) volcanicity in western and central India;
iii. intrusions of plutonic alkaline masses of diverse composition which were also responsible for the development of isolated patches and linear belts

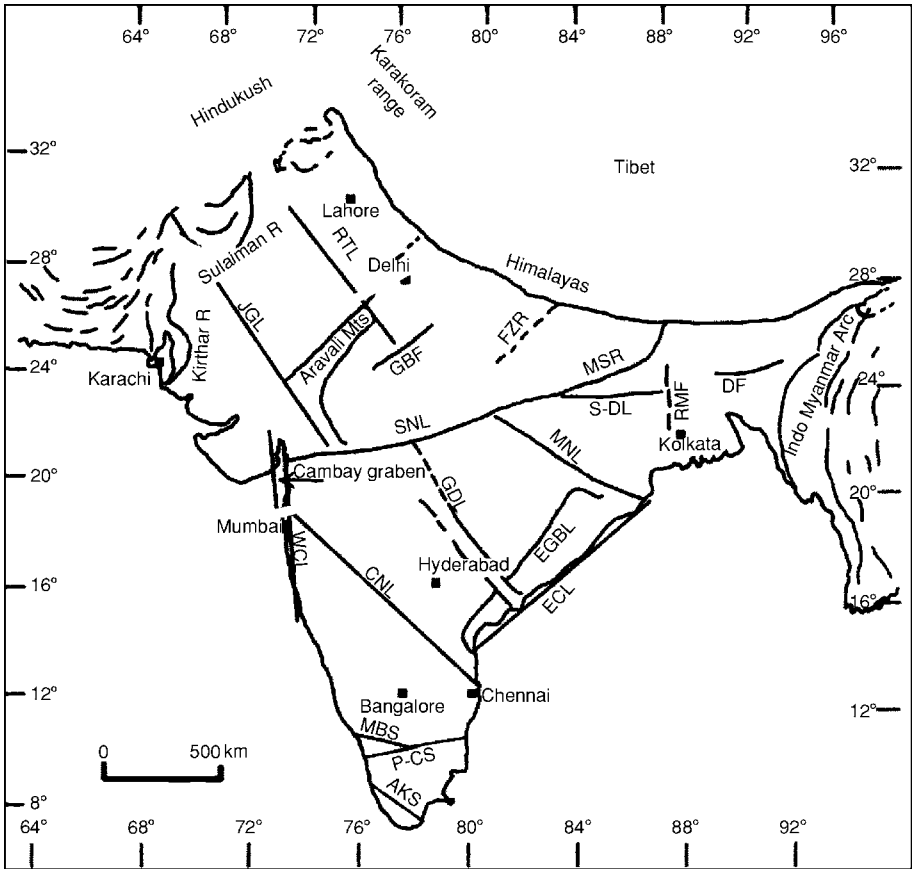


Figure 7 Generalised map showing the distribution of major lineament fractures in the Indian Sub Continent (adapted with permission from Roy, 2004; © Gondwana Research, International Association for Gondwana Research, Japan). Legend: AKS Achankovil Shear Zone: CNL Chennai Nasik Lineament: DF Deuki Fault: ECL East Coast Lineament; GDL Godavari Lineament; GBF Great Boundary Fault; FZR Faizabad Ridge: JGL Jaisalmer Barwani Lineament; MBS Moyar Bovani Shear Zone; MNL Maha nadi Lineament; MSR Munger Saharsa Lineament; P CS Palghat Cauvery Shear Zone; RMF Rajmahal Fault; SNL Son Narmada Lineament; S DL Son Damodar Lineament; RTL Raisinghgarh Tonk Lineament; WCL West Coast Lineament.

of positive Bouguer gravity anomaly and zones of high heat flow; and

iv. the development of fractures/lineaments. The palaeomagnetic data from the Deccan flood basalt suggests their development in two stages during C30N and C29R.

Geology and Evolution of the Himalayas

The Himalayas, the grand crescent-shaped mountain ranges with a prominent southward convexity, fringes the entire northern margin of the Indian Sub-continent. The world's loftiest and youngest mountain ranges extend for over 2500 km from south of the Indus Valley beyond Nanga Parbat (height 8114 m) in the west to Namcha Barwa (height 7755 m) in the east. Topographically, the Himalayas is bent sharply at the western end to join with the Sulaiman and Kirthar Ranges. With a similar sharp

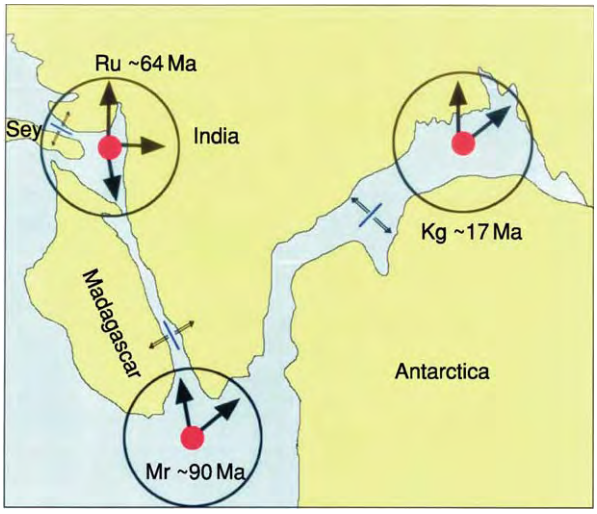


Figure 8 Simplified map of India, Antarctica and Madagascar, showing timing of the three different plume outbursts around the Indian Sub Continent. (Adapted with permission from from Raval and Veeraswamy, 2003; © Gondwana Research, International Association for Gondwana Research, Japan.)

bending at the eastern end, the mountain ranges join the north-trending Indo-Myanmar Arc represented by the Naga Hills and the Arakan Yoma (Figure 9).

The Himalayas are divided axially into the following five units, each showing distinctive litho-tectonic character and evolutionary history:

i. The Sub-Himalayas. 10 to 50 km wide belt of Late Tertiary molasse sediments age constituting the Siwalik Group. The belt also includes the older Murree formations and their equivalent, the Dharamshalas.
ii. The Lesser Himalayas. 60 to 80 km wide belt predominantly comprising Proterozoic low-grade metamorphosed rocks overlain by thrust sheets of granites and metamorphic rocks.
iii. The Higher (or Great) Himalayas. 10 to 15 km thick belt of dominantly Precambrian metamorphites and young granites of Cenozoic age. This is also the zone of highest uplift.
iv. Trans-(or Tethyan) Himalayas, a belt of dominantly shelf (usually fossil-bearing) sediments of Late Proterozoic to Cretaceous age, bounded by the Indus-Tsangpo Suture Zone (ITSZ), a relatively narrow belt of ophiolites and associated sediments. Though not a thrust contact, the ITSZ is an important tectonic contact welding the Indian continental block with the Tibetan block Immediately north of the ITSZ is a belt of 40 Ma to 100 Ma old granitoids, known as the Trans-Himalayan batholith granites.

The tectonic architecture of the Himalayas is built on three prominent intracrustal thrusts. From north to south these thrusts are:

i. The Main Central Thrust (MCT) which separates the crystalline rocks of the Higher Himalayas from the low-grade metamorphites of the Lesser Himalayas.
ii. The Main Boundary Thrust (MBT) which regionally separates the Lesser Himalayas from the Sub-Himalayas.
iii. The Himalayan (Main) Frontal Thrust (HFT or MFT) which demarcates the tectonic and physiographic boundary between the Siwaliks and the Indo-Gangetic Alluvial Plain.

A unique feature of the Himalayas is its crustal thickness, which rises from about 35 km in the IGAP to a thickness of between 65 and 80 km over

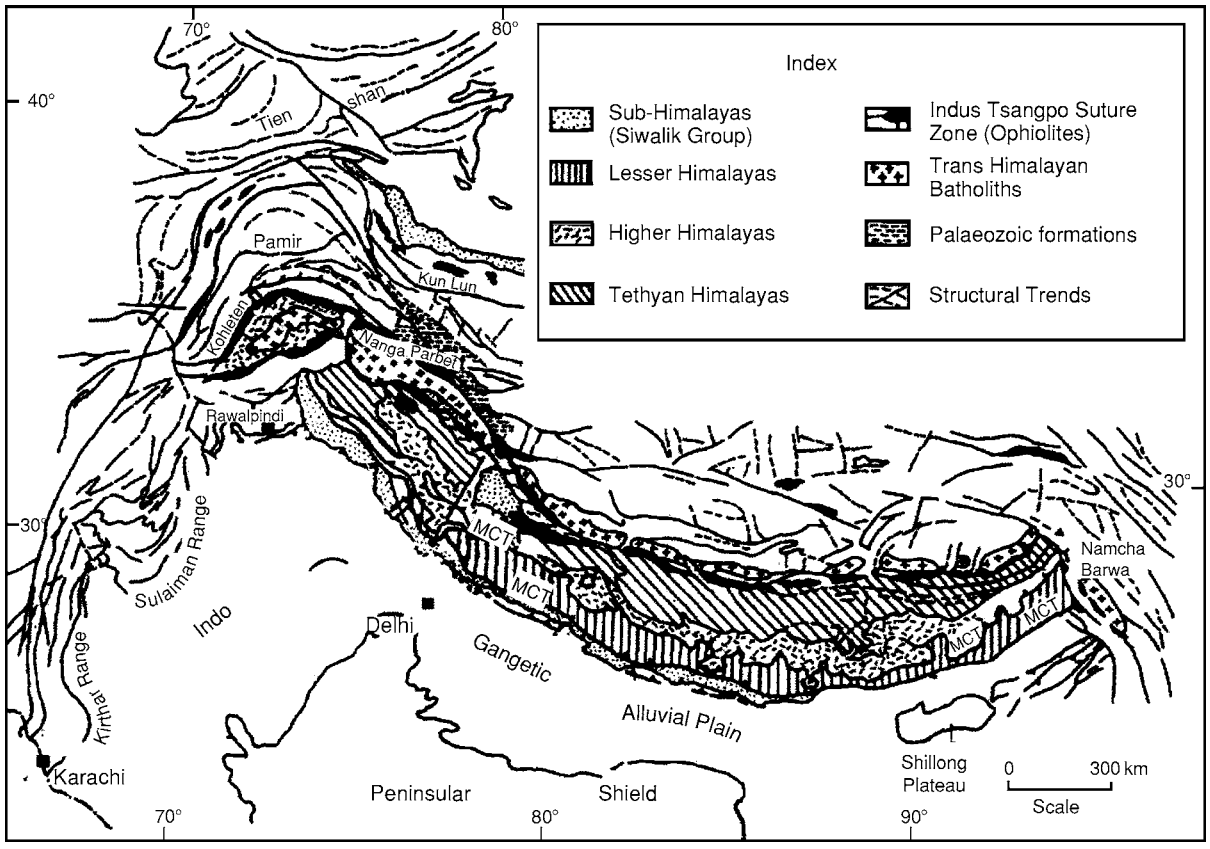


Figure 9 Geological map of the Himalayas showing important tectonic elements. MBT Main Boundary Thrust; MCT Main Central Thrust. (Adapted with permission from Mahadevan, 1994; © Geological Society of India, Bangalore.)

the Higher Himalayas. The increasing crustal thickness is reflected in the MOHO dip, which has been estimated to be 7°–8° N under the Sub-Himalayas up to over 15° further north.

Lithotectonic Units and Sedimentation

A considerable part of the Higher Himalayan mass is the Precambrian rocks which also constitute Peninsular India. An important component of the Lesser Himalayas is the Lower Palaeozoic (*ca*. 500 ± 50 Ma) dominantly per-alkaline anorogenic granites, which are popularly termed the Pan-African. The sedimentary rocks suggest continuity of sedimentation in some parts in the northern shelf sea across the Precambrian-Cambrian boundary into the Lower Palaeozoic, especially during the Cambrian. Records of continuity of sedimentation in these basins during the Ordovician and Silurian are, however, quite equivocal, although both periods are represented by fossils. Deposition of shelf facies sediments was resumed during Carboniferous in zones of marine incursions along intracratonic rift basins that had developed in the Salt Range and in the Kashmir region. The deposition in some basins continued until the Triassic. There are records of contemporary volcanicity in the Pir Panjal Ranges and in other places in the east. Both marine and continental sediments are correlatable with those of the Gondwana deposits which occur in parts of Nepal and the Sikkim Himalayas.

Continuous sedimentation from the Cambrian to the Eocene, with a number of breaks, is recorded in the Tethys belt. Fossil records from different parts of this belt indicate that the extent of these breaks were not of uniform duration.

The closure of the Tethys Ocean by the Eocene caused a brief pause in sedimentation, which was resumed around mid-Miocene times in two important basins. One in the north opened as a major intermontane (back arc) basin in the suture zone, leading to the deposition of Indus Group. The Siwalik Group was deposited in the southern foreland basin that developed in front of the rising Himalayas from around 18 Ma. The Siwalik Group bears rich fossil records of plants, molluscs, fishes, reptiles, and mammals. The last Himalayan upheaval at around 1.7 Ma caused shifting of the depocentres to the south to build up the flood plains of the Indo-Gangetic Alluvial Plain.

While the closure of Tethys Sea marked the end of sedimentation in the north, the marine shelf sedimentation continued both along the eastern and western margins of the Indian continental block, in the Naga Hills and Arakan Yoma in the east and the Sulaiman and the Kirthar Ranges in the west. Sedimentation in these basins, which began in the Eocene, continued at least until the Oligocene.

Himalayan Tectonics

The earliest Himalayan deformation coincided with the final closure of the Tethys Sea at around 50 Ma, affecting the rocks on either side of the suture zone. There was a distinct southward polarity of the deformation across the Tethys to the Higher Himalayan Crystalline Complex. A series of south-directed recumbent folds and thrusts were produced in the Higher Himalayas, resulting in thickening of the crust, with attendant Barrovian metamorphism, anatexis and the generation of leucogranites. The southward transmission of thrust nappes by the MCT, continued till around 22 Ma. This was also the time when the Barrovian metamorphic isograds underwent inversion.

Almost simultaneously with the piling of the fold-thrust nappes in the Higher Himalayas, the Indus molasse basin in the north and the Siwalik molasse basin in the south developed as rapidly subsiding troughs.

Like the piggy-back thrusting model, the southward transmission of the fold-thrust nappes, which was initially along the MCT, was later carried on by MBT in the south. The HFT (or MFT) which overrides the Recent sediments was the last thrust to form in the Himalayan tectonics.

The Himalayas represent a classic example of continent-continent collision. The very similar tectonic pattern observed over the entire length of the Himalayas is primarily an expression of the impact of two continental blocks. Complexities noted in the western end of the Himalayas arise because of the development of an island arc complex (Kohistan-Dras Island Complex) prior to its collision with the Karakoram microplate (possibly during the mid-Crataceous).

Palaeomagnetic data indicate an initiation of the continental collision at equatorial latitudes, resulting in the progressive suturing from the Paleocene in the north-western Himalayas until the Eocene in the eastern Himalayas. Continued convergence and indentation of the Indian continental block with southern Asia (or Tibet) up to the Early Miocene, resulted in the doubling of the crustal thickness over a large region of the Himalayas, the Pamir-Hindukush and Tibet. The total area of the thickened crust may account for about 2000 km of crustal shortening in the entire orogen. As to the origin of the Himalayan arc, palaeomagnetic observations seem to favour a steady-state model of formation of the arcuate bending of the mountain ranges due to Late Tertiary

counterclockwise rotational underthrusting of the Indian continental block beneath the Tibetan Plateau after the latest Miocene.

Quaternary Sedimentation and Neotectonics

The Quaternary geology, which began with the waning phase of the Siwalik sedimentation, came to an end with the most recent upheaval of the Himalayas. The depocentres had by then shifted to their subsiding southern parts which ultimately evolved as the Indo-Gangetic Alluvial Plains (IGAP). Geographically, apart from the Indus-Ganga flood plains, the IGAP also include the narrow basin of the Brahmaputra River in the east and the Thar Desert (along with the North Gujarat Plain) in the west. The alluvial sediments over the entire IAGP belt range between 400 m and 800 m, with a maximum thickness of about 6 km along the edge of the Himalayas. The belt is divided into a number of sub-basins by several submerged ridges (basement highs) lying across it.

Quaternary sediments outside the IGAP occur in the Narmada and Tapti Basins in Peninsular India, and along the eastern coastlines. Thick laterite formations (some of which contain rich bauxite deposits) were produced at this time in parts of central India, eastern Ghats and the Konkan coasts in the western Ghats.

The formation of the Thar Desert in the east of the Indus Basin, which had a fluvial pre-history, is linked with the establishment of the monsoon system over the Sub-Continent by the mid-Pleistocene, with the high rising Aravalli Mountains producing the rain-shadow zone to its west. The series of saline lakes that dot the entire desert-land were formed by the segmentation and blocking of river channels due to neotectonic movements.

The Quaternary neotectonic movements caused spectacular geomorphic changes in the entire Sub-Continent, primarily through movements of fault-bounded blocks. The Rann of Kachch in northern Gujarat is a classic example of regional uplift during historical times. The development of the Ganga-Brahmaputra-Megna Delta Complex (also known as the Sundarban Delta) is a very important geological landform feature which evolved in three stages of tectonically-influenced delta sedimentation processes during the Late Pleistocene. A point of interest in the delta sediments is the arsenic toxicity in the groundwater aquifer within the Holocene sediments.

See Also

Gondwanaland and Gondwana.

Further Reading

Dasgupta S and Sengupta P (2003) Indo Antactica correl ation: a perspective from the Eastern Ghats Granulite Belt, India. In: Yoshida M, Windley BF, and Dasgupta S (eds.) *Proterozoic East Gondwana: Supercontinent As sembly and Breakup*. Geological Society of London, Special Publications 206: 131 143.

Mahadevan TM (1994) *Deep Continental Structure of India: A Review.* Geological Society of India Memoir 28.

Mahadevan TM (2002) *Geology of Bihar & Jharkhand.* Bangalore: Geological Society of India.

Mukhopadhyay D (2001) The Archaean Nucleus of Singhbhum: the present state of knowledge. *Gondwana Research* 4: 307 318.

Naqvi SM and Rogers JJW (1987) *Precambrian Geology of India.* New York: Clarendon Press. Oxford: Oxford University Press.

Radhakrishna BP (1983) Archaean granite greenstone terrain of the South Indian Shield, 1 46. In: Naqvi SM and Rogers JJW (eds.) *Precambrian of South India.* Geological Society of India Memoir 4.

Radhakrishna BP and Vaidyanadhan R (1997) *Geology of Karnataka.* Bangalore: Geological Society of India.

Raval U and Veerasswamy K (2003) India Madagascar separation: Break up along a pre existing mobile belt and chipping of the craton. *Gondwana Research* 6: 467 485.

Roy, Abhinaba, Raamachandra HM, and Bandopadhyay BK (2000) Supracrustal belts and their significance in the crustal evolution of Central India. *Geological Survey of India, Special Publication* 55: 387 406.

Roy AB (2004) The Phanerozoic reconstitution of Indian Shield as the aftermath of break up of the Gondwana land. *Gondwana Research* 7(2): 387 406.

Roy AB and Jakhar SR (2002) *Geology of Rajasthan: Northwest India: Precambrian to Recent.* Jodhpur: Scientific Publishers (India).

Sastry MVA, Acharyya SK, Shaw SC, *et al.* (1977) Strati graphic lexicon of Gondwana Formations of India. *Geo logical Survey of India Miscellaneous Publication* 36, 170.

Searle MP, Windley BF, Coward MP, *et al.* (1987) The closing of Tethys and tectonics of the Himalaya. *Geo logical Society of America Bulletin* 98: 678 701.

Soni MK, Chakraborty S, and Jain VK (1987) Vindhyan Supergroup A review. *Geological Society of India Memoir* 6: 87 138.

Subbarao KV (1999) Deccan Volcanic Province. *Geological Society of India Memoir* 43(1) and 43(2), 947.

Valdiya KS (1998) Dymanic Himalaya. Hyderabad: Universities Press.

JAPAN

J Tazawa, Niigata University, Niigata, Japan

Introduction

Japan, lying off the eastern coast of Eurasia, has an area of 377 819 km^2; slightly larger than the United Kingdom, and comprises four major islands, Hokkaido, Honshu, Shikoku, Kyushu, and some thousands of smaller islands. Honshu accounts for over 60% of the total area. The Japanese Islands are about 3800 km long, ranging from Kunashiri (45·33·N, 148·45·E) to Yonakuni (24·27·N, 122·56·E) (Figure 1).

The Japanese Islands consist of two arc-trench systems, the North-east Japan and South-west Japan Systems. The NE Japan System, related to the subduction of the Pacific Plate, extends from the Chishima (Kuril) arc-trench to the Izu-Ogasawara (Bonin) arc-trench, through the NE Japan Arc-Japan Trench. The SW Japan System, related to the subduction of the Philippine Sea Plate, extends from the SW Japan Arc-Nankai Trough to the Ryukyu arc-trench. Two back-arc basins, the Okhotsk Sea and the Japan Sea, are placed behind the NE Japan System. Moreover, another new, weakly-developed convergent plate boundary (*see* **Tectonics:** Mountain Building and Orogeny) extends along the eastern margin of the Japan Sea from Sakhalin to central Honshu. The Itoigawa-Shizuoka Tectonic Line (west border of the 'Fossa Magna') is considered to be the boundary between the North American Plate and the Eurasian Plate. Thus Japan exhibits the interaction of four plates, the Pacific, North American, Philippine Sea, and Eurasian Plates. Japan's current tectonic setting has existed only since mid-Tertiary (Early Neogene) time.

Japan was born at the site of a trench, namely, a subduction zone bordering the eastern margin of North China (Sino-Korea) in the Early Ordovician (about 500 Ma). Since then, the Proto-Japan has experienced successive accretions, with episodic high-P/T metamorphism, thrusting, and large scale strike-slip faulting and therefore the geology of Japan is very complicated.

Geological exploration of Japan dates from the pioneer work of E. Naumann in 1875–1885. He first described the geological features of the Japanese Islands, and proposed the 'Fossa Magna' and the Median Tectonic Line. Between 1940 and 1960, T. Kobayashi and M. Minato put forward syntheses on the tectonic development of the Japanese Islands based on the Geosyncline concept. In contrast to the classical syntheses, K. Kanmera, A. Taira, and T. Kimura expressed their new progressive opinions in the 1970 to the 1980s, after the establishment of the Plate Tectonic theory. In the early 1990s, three models of the tectonic framework of the Japanese Islands and its development were proposed: (i) the microcontinent model by K. Ichikawa and others; (ii) the nappe model by Y. Isozaki and S. Maruyama; and (iii) the strike-slip model by J. Tazawa. Since then, an enduring controversy has continued between the three models.

In this article, the tectonic division and evolution of the Japanese Islands are based on the strike-slip model. The 'Economic Geology' of the area is omitted, since most of the mines in Japan are no longer working.

Topography

Japan is very mountainous, with mountains occupying about 71% of the total land area (Figure 2). Altogether, 532 of these mountains are over 2000 m high; Mt Fuji, the highest, rises to 3776 m. The high mountains are concentrated in central Honshu, just west of the 'Fossa Magna', and form three echelon ranges, the Northern, Central, and Southern Japan Alps from north to south. Other ranges coincide with the Palaeozoic to Mesozoic tectonic and regional metamorphic belts or the Quaternary volcanic belts. Trenches are particularly deep along the Pacific Plate boundary, reaching about 7000 m in the Chishima and Japan Trenches, and 8000 m in the Izu-Ogasawara Trench. The Nankai Trough along the Philippine Sea Plate is not as deep as a trench, being less than 6000 m. The Japan Sea, one of the back-arc basins, is roughly divided into northern and southern parts: the Northern Japan Sea has a flat floor, the Japan Basin, of about 3500 m: the Southern Japan Sea has an uneven floor, with the Yamato Bank, Yamato Basin, Tsushima Basin, and Korean Plateau.

Volcanism

About 200 Quaternary volcanoes, of which more than 60 are still active, occur in the Japanese Islands (Figure 3). These volcanoes are distributed in two belts, the East Japan and West Japan Volcanic Belts, and are related to the subduction of the Pacific Plate and the Philippine Sea Plate, respectively. The East Japan Volcanic Belt extends parallel to the Chishima-Japan-Izu-Ogasawara trenches, from the Chishima Islands to the Izu-Ogasawara Islands, through Hokkaido and northern Honshu. Mt Fuji, a typical

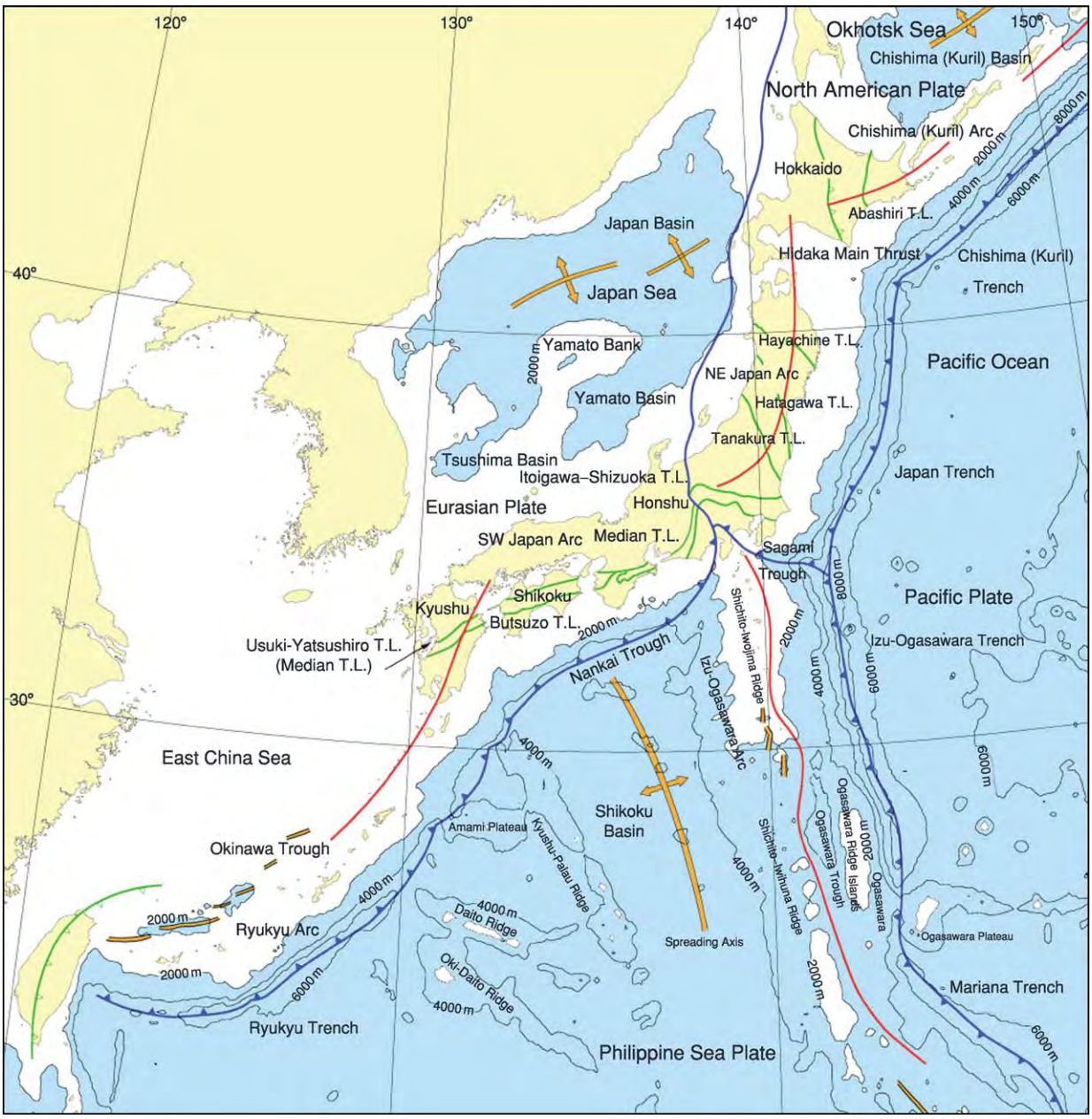


Figure 1 Location map of Japan and the surrounding regions. (After Editorial Committee for the Geology of the Japanese Islands (2002).)

stratovolcano, belongs to this belt. The West Japan Volcanic Belt runs parallel to the Nankai Trough and the volcanoes assigned to this belt occur in the Japan Sea side of western Honshu and Kyushu. Some volcanoes, such as Mt Usu in Hokkaido and Mt Aso in Kyushu, formed large calderas after the eruption of vast amounts of pyroclastic material. In general, the volcanic rock types range from basalt, andesite, dacite, and rhyolite of the tholeiitic series on the east, near the volcanic front, to andesite and basalt of more alkalic or high-alumina basalt types on the west, far from the volcanic front.

Earthquakes

Figure 4 shows the hypocentre distribution of deep-focus earthquakes, over M 2 and deeper than 30 km in Japan. Most of the deep-focus earthquakes occur along the seismic plane (Wadati-Benioff zone) of both arc-trench systems, the NE Japan and SW

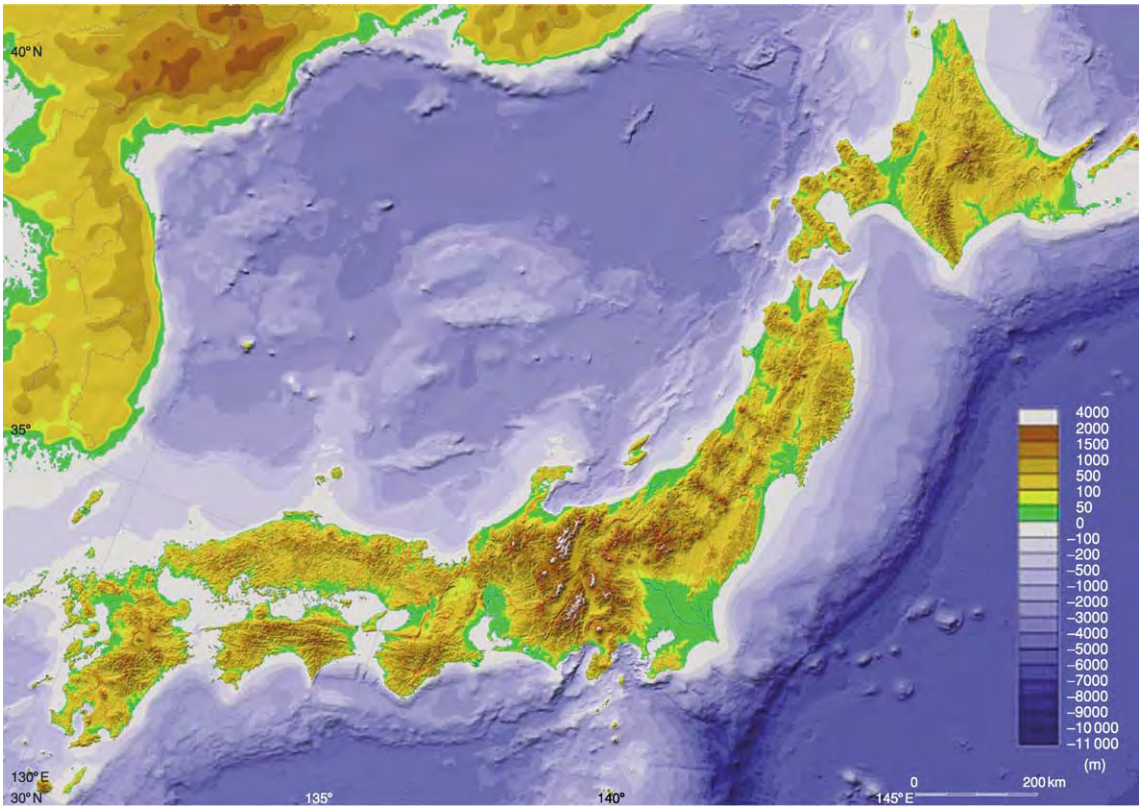


Figure 2 Topographical map of Japan. (After Editorial Committee for the Geology of the Japanese Islands (2002).)

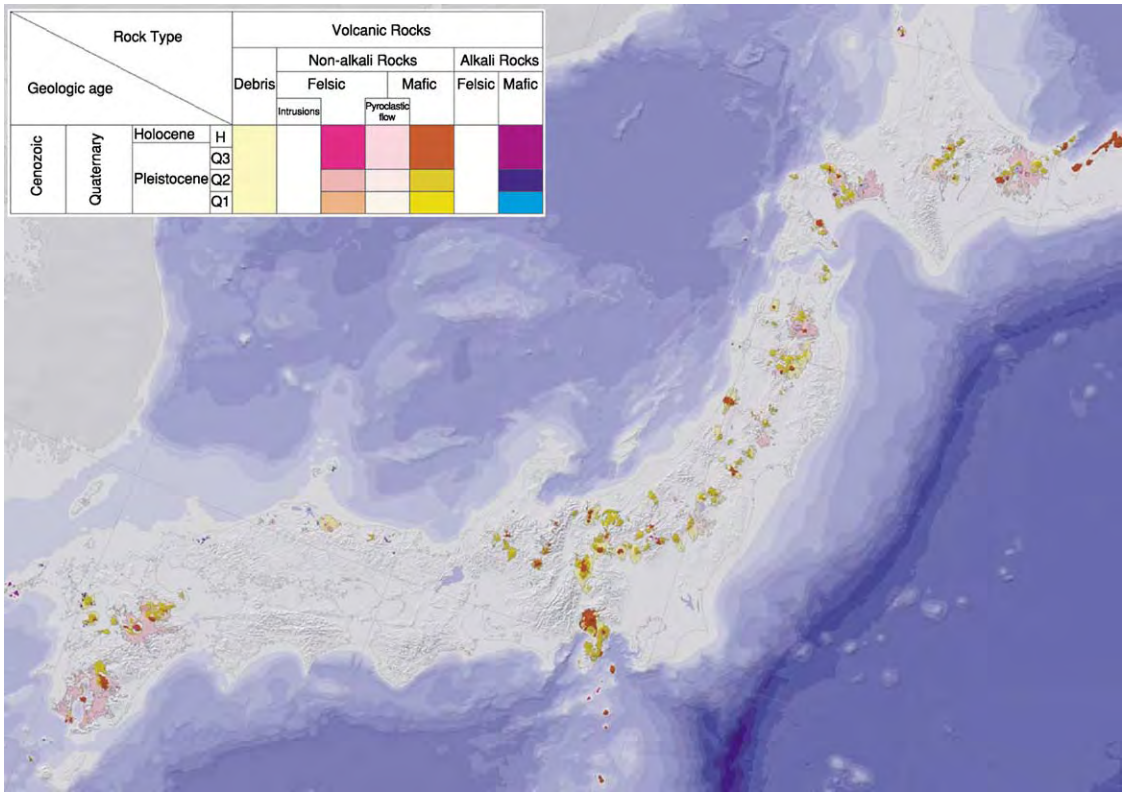


Figure 3 Distribution of Quaternary volcanoes and volcanic rocks in Japan. (After Editorial Committee for the Geology of the Japanese Islands (2002).)

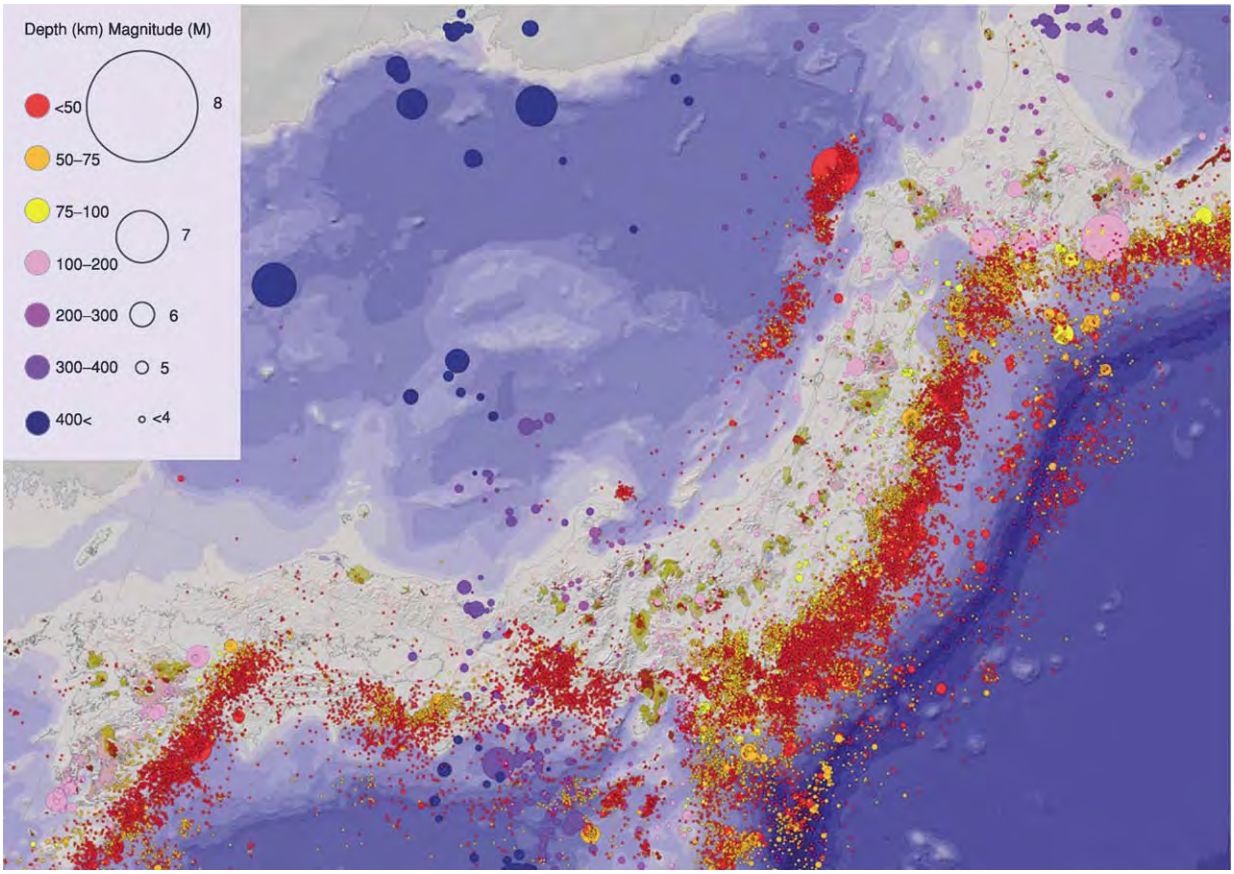


Figure 4 Map showing the distribution of deep focus earthquakes and Quaternary volcanoes, overprinted on **Figure 3**. (After Editorial Committee for the Geology of the Japanese Islands (2002).)

Japan Systems. Occasionally, shallow submarine earthquakes generate large tsunamis along both the Pacific and Japan Sea sides, particularly in northern Honshu and the Ryukyus. The Great Kanto Earthquake (*M* 8.2; 1 September 1923) killed over 140 000 people and destroyed a third of Tokyo and most of Yokohama. The Meiji Sanrikuoki Earthquake (*M* 8.6; 15 June, 1896) caused a great tsunami of 38 m high, and killed about 22 000 people in the Sanriku coast of Iwate Prefecture, Pacific side of northern Honshu.

In the NE Japan System, the distribution of hypocentres along the seismic plane is smooth and continuous. However, the dipping angle of the seismic plane, i.e., the boundary between the Eurasian Plate and the subducting Pacific Plate varies from place to place. The average angle is about 30° under northern Honshu, but it is steeper under Hokkaido and the Izu-Ogasawara Islands. The maximum depth of deep-focus earthquakes is estimated to be about 700 km. The distribution of hypocentres in the SW Japan System is irregular and they are intermittently distributed beneath central Honshu and Shikoku. The maximum depth in Kyushu is estimated to be about 200 km.

Geological Outline

The geology of Japan (**Figure 5**) has been complicated by long-ranging subduction-related tectonic movement since the Early Ordovician. The tectonic framework of the Japanese Islands is difficult to recognize owing to the thick covering of the Neogene and Quaternary sediments, which include volcanic rocks.

The Pre-Neogene rocks of the Japanese Islands are divided into the following four terranes: (i) the South Kitakami Terrane (Early Ordovician to Late Devonian accretionary terrane, comprising the Hida Gaien, South Kitakami, and Kurosegawa Belts); (ii) the Akiyoshi Terrane (Middle to Late Permian accretionary terrane, comprising the Akiyoshi, Maizuru, Suo, Ultra Tanba, and Joetsu Belts); (iii) the Mino Terrane (Early Jurassic to Early Cretaceous accretionary terrane, comprising the

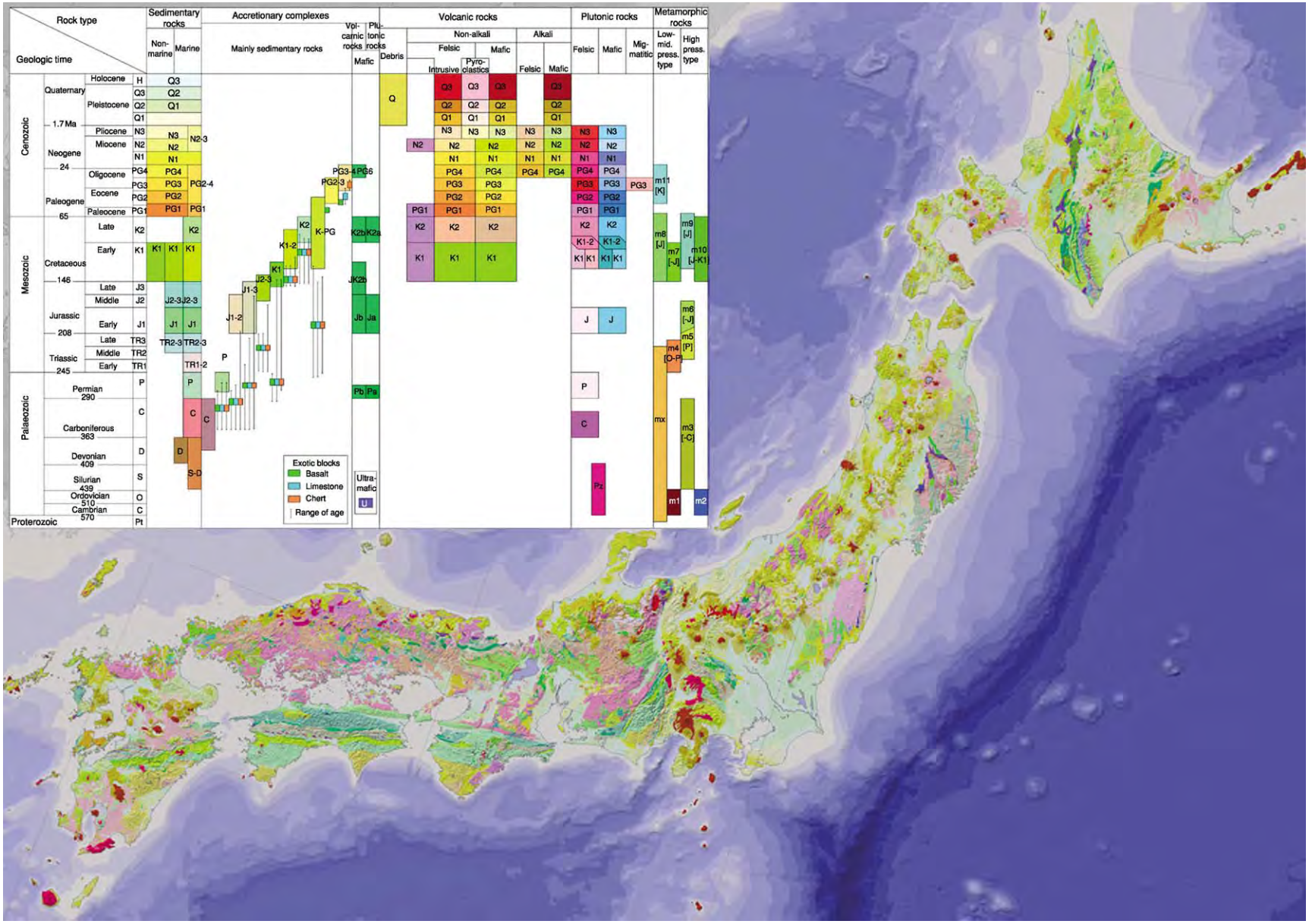


Figure 5 Geological map of Japan. (After Editorial Committee for the Geology of the Japanese Islands (2002).)

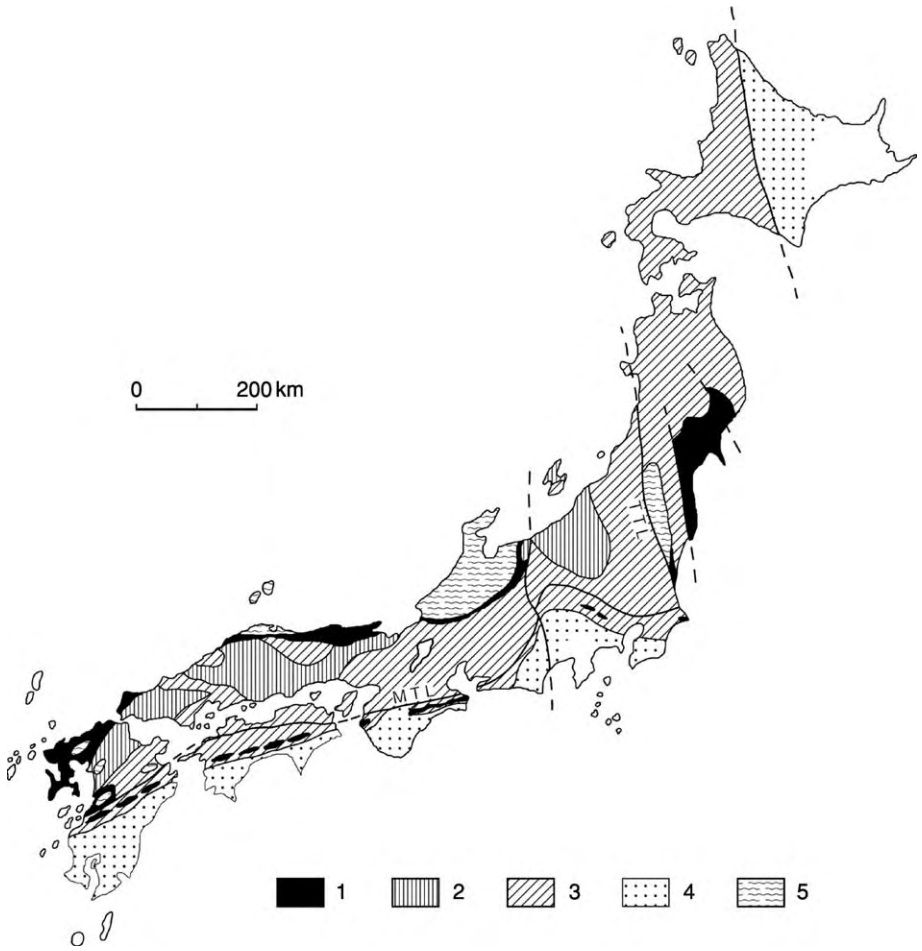


Figure 6 Tectonic map of the Japanese Islands, showing the distribution of pre Neogene terranes and nappes. MTL: Median Tectonic Line, TTL: Tanakura Tectonic Line, Legend 1: South Kitakami Terrane, 2: Akiyoshi Terrane, 3: Mino Terrane, 4: Shimanto Terrane, 5: Hida Abukuma Nappe. (After Tazawa (2004).)

Mino, Kamuikotan, North Kitakami, Chizu, Ryoke, Sanbagawa, Northern Chichibu, and Southern Chichibu Belts and the Gozaisho metamorphic rocks); and (iv) the Shimanto Terrane (Late Cretaceous to Early Neogene accretionary terrane, comprising the Shimanto, Hidaka and Tokoro Belts and part of the Idonnappu Belt) (Figure 6). These terranes are arranged in a NE–SW direction, subparallel with the extension of the Japanese Islands, and younging towards the south-east, from the Japan Sea side to the Pacific side. Besides the four terranes, several nappes (Hida Nappe, Abukuma Nappe, Higo Nappe, etc.) in the Hida-Abukuma Nappe, consisting of Palaeozoic and Mesozoic gneisses, schists, amphibolites, crystalline limestones, and granites, all of which experienced medium-pressure metamorphism in the Triassic (250–220 Ma), thrust over the South Kitakami, Akiyoshi, and Mino Terranes towards the east to south-east. These metamorphic rocks were probably derived from the Qinling-Dabie Suture Zone, a collision zone between the North China and South China blocks.

Tectonic Evolution

The tectonic framework of the Japanese Islands was formed through long-term subduction onto mostly the eastern margin of North China, since the Early Ordovician. The Silurian to Permian marine faunas (brachiopods, cephalopods, and trilobites) and the Permian floras indicate affinities to those of central Asia (Tien Shan, north Xinjiang), Inner Mongolia, North-east China, and Far East Russia (Primorye), except for the Permian Tethyan-Panthalassan fusulinoid and brachiopod faunas from exotic limestone blocks in the Jurassic melange

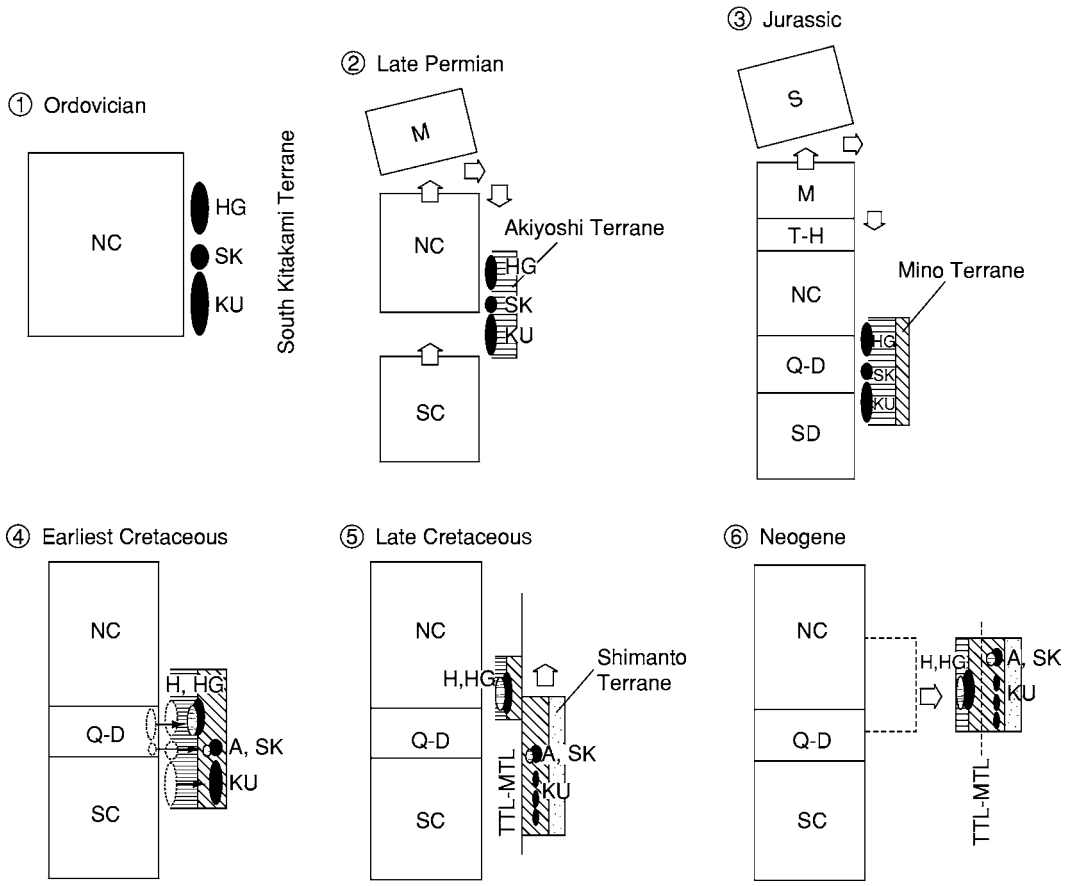


Figure 7 Generalized kinematic model for the tectonic development of the Japanese Islands in Ordovician to Neogene times. HG: Hida Gaien, KU: Kurosegawa, M: Mongolia, NC: North China, Q D: Qinling Dabie, S: Siberia, SC: South China, SK: South Kitakami, T H: Tienshan Hinggan, TTL MTL: Tanakura Median Tectonic Line. (After Tazawa (2004).)

of the Mino Terrane. For example, the Permian brachiopod faunas of the Hida Gaien and South Kitakami Belts in the South Kitakami Terrane are characterized by a mixture of Boreal and Tethyan elements, and are similar to those of Inner Mongolia, North-east China, and Primorye. Figure 7 shows the tectonic evolution of the Japanese Islands together with some continental blocks in east Asia.

The oldest trench, related to the accretion of Ordovician to Late Devonian sediments in the Hida Gaien, South Kitakami, and Kurosegawa Belts, all of which belong to the South Kitakami Terrane (Figure 7(1)), changed to a continental shelf in the Early Carboniferous. Subsequently, a new subduction zone developed in front of the continental shelf in the Middle Permian. The Middle to Late Permian sediments of the Akiyoshi Terrane accumulated in this trench (Figure 7 (2)). After that, the South Kitakami Terrane (together with the Akiyoshi and Mino Terranes) moved towards the south along the dextral strike-slip faults on the eastern margin of North China in the Late Permian to the Late Jurassic. This dextral strike-slip movement may have been caused by the collision of the Mongolian and North China blocks in the Late Permian to Triassic (Figure 7(2)) and the Siberian and Mongolian blocks in the Jurassic (Figure 7(3)). Next to the dextral strike-slip faulting, a thrusting towards the east took place in this region in the latest Jurassic to earliest Cretaceous. This thrusting resulted in the emplacement of several nappes, such as the Hida Nappe, South Kitakami Nappe, and Kurosegawa Nappe (Figure 7(4)). The single tectonic unit, therefore, was then separated into three belts, the Hida Gaien Belt in the Inner Zone (continental side), and the South Kitakami and Kurosegawa Belts in the Outer Zone (oceanic side). Then a rapid northward motion of the Izanagi Plate (*ca.* 20–30 cm year) caused a large-scale sinistral displacement (1500–2000 km) of the Outer Zone, including the South Kitakami and

Figure 8 Reconstruction of Early Cretaceous Japan. Solid arrow shows the direction of strike slip motion in that time. Shading as for **Figure 6**. (After Tazawa (2004).)

Kurosegawa Belts along the Tanakura Tectonic Line, Median Tectonic Line and some associated faults in the Early Cretaceous to Palaeogene, mainly in the Early Cretaceous (Figures 7(5) and 8). The pre-Cretaceous arrangement of the Joetsu, Hida Gaien, South Kitakami, and Kurosegawa regions in this order (from north to south) is supported by the biogeographical evidence from the Middle Permian brachiopod faunas, as well as the Late Jurassic to Early Cretaceous floras. This spatial arrangement of the pre-Cretaceous Japan was greatly changed by the sinistral strike-slip faulting in the Early Cretaceous to Palaeogene. From the Early Jurassic to Palaeogene, the Mino and Shimanto Terranes accreted on to the subduction zone at the front of the Proto-Japan. Finally, the opening of the Japan

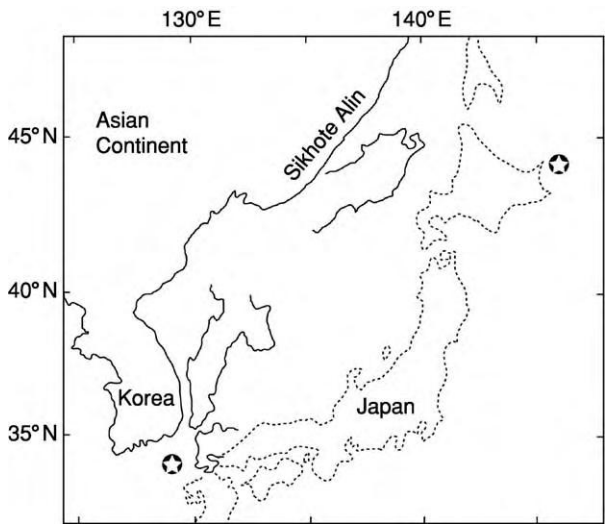


Figure 9 Reconstruction of Early Neogene Japan, showing the 'Double door opening model' of NE Japan and SW Japan. Stars show the rotation pivots for NE Japan and SW Japan, respectively. (Modified from Otofuji *et al.* (1985).)

Sea caused a clockwise rotation of South-west Japan and a counterclockwise rotation of North-east Japan in the Early Neogene (Miocene) (Figure 7(6) and 9).

See Also

China and Mongolia. **Russia**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mountain Building and Orogeny.

Further Reading

Editorial Committee for the Geology of the Japanese Islands (ed.) (2002) *Computer Graphics: Geology of the Japanese Islands, CD ROM Version*, Tokyo: Maruzen.

Hashimoto M (ed.) (1991) *Geology of Japan*. Tokyo: Terra Scientific Publishing Company.

Ichikawa K, Mizutani S, Hara I, Hada S, and Yao A (eds.) (1990) *Pre Cretaceous Terranes of Japan*, Publication of IGCP Project No. 224. Osaka: Nippon Insatsu Shuppan.

Isozaki Y (1996) Anatomy and genesis of a subduction related orogen: A new view of geotectonic subdivision and evolution of the Japanese Islands. *The Island Arc* 5: 289 320.

Isozaki Y and Maruyama S (1991) Studies on orogeny based on plate tectonics in Japan and new geotectonic subdivision of the Japanese Islands. *Journal of Geography* 100: 697 761. (In Japanese).

Kanmera K, Hashimoto M, and Matsuda T (eds.) (1980) *Geology of Japan, Earth Sciences* 15. Tokyo: Iwanami Shoten. (In Japanese).

Kimura T, Hayami I, and Yoshida S (1991) *Geology of Japan*. Tokyo: University of Tokyo Press.

Otofuji Y, Matsuda T, and Nohda S (1985) Palaeomagnetic evidence for the Miocene counter clockwise rotation of Northeast Japan rifting process of the Japan Arc. *Earth and Planetary Science Letters* 75: 265 277.

Taira A and Tashiro M (eds.) (1987) *Historical Biogeography and Plate Tectonic Evolution of Japan and Eastern Asia*. Tokyo: Terra Scientific Publishing Company.

Tazawa J (1993) Pre Neogene tectonics of the Japanese Islands from the viewpoint of palaeobiogeography. *Journal of the Geological Society of Japan* 99: 525 543. (In Japanese).

Tazawa J (2002) Late Paleozoic brachiopod faunas of the South Kitakami Belt, northeast Japan, and their paleo biogeographic and tectonic implications. *The Island Arc* 11: 287 301.

Tazawa J (2004) The strike slip model: A concept on the tectonic evolution of the Japanese Islands. *Journal of the Geological Society of Japan* 110: 503 517. (In Japanese).

JUPITER

See **SOLAR SYSTEM: Jupiter, Saturn and Their Moons**

LAGERSTÄTTEN

S E Gabbott, University of Leicester, Leicester, UK

Introduction

This article will describe in detail Fossil Lagerstätten, with a special emphasis on conservation Lagerstätten – deposits containing exceptionally preserved fossils. Concentration Lagerstätten are deposits where fossils are especially abundant and the processes that operate to produce these types of fossil horizons will be described. Fully articulated skeletons and the preservation of the soft parts of organisms, which are usually rapidly lost to decay after death, both constitute examples of exceptional preservation. In order to create these extraordinary fossil deposits, scavengers must be prohibited from disarticulating and consuming the carcass and bacterial putrefaction of soft tissues must be circumvented. Authigenic mineralization of soft tissues produces fossils with the highest degree of fidelity known, where even subcellular features can be recognised. Mineralization processes, including replacement by apatite, pyrite, clays, and silica are documented using examples from conservation Lagerstätten. Even exceptionally preserved fossils show some degree of decay-induced morphological change, so that decay experiments have proved crucial in their accurate interpretation. In addition, decay experiments have determined some of the physical and chemical factors, such as degree of oxygenation, pH, temperature, and soft tissue composition which all affect preservation. Exceptionally preserved faunas not only produce some of the world's most spectacular fossils, but they are also critical in the reconstruction and understanding of ancient life and evolutionary processes. Such deposits may be temporally constrained because some geological periods are replete in examples compared with others; the putative explanations for such temporal trends are explored.

Lagerstätten, a German word, was originally used to describe mineral and ore deposits of economic worth. Adolf Seilacher first coined the term Fossil Lagerstätten (singular Lagerstätte) to describe a body of rock that is unusually rich in palaeontological information because, either the fossils are so well preserved and/or the fossils are so abundant that they warrant exploitation and scientific attention. The distinction between exceptionally preserved fossils and exceptionally numerous fossils has led to the terms conservation Lagerstätten and concentration Lagerstätten being applied, respectively. It is the conservation Lagerstätten that have received the most attention because they contain fossils that are extraordinary in the quality of their preservation. They may contain completely articulated skeletons or soft-bodied animals composed of nonmineralized, so-called soft-parts and soft-tissues. Fossil conservation Lagerstätten include some of the world's most celebrated deposits, such as the Cambrian Burgess Shale of Canada and the Jurassic Solnhofen Lithographic Limestone of Germany. Exceptionally preserved faunas are confined to particular environmental settings, where a number of processes must occur in order to fossilise the nonmineralised tissues. In recent years, some of these processes have become well documented and will be described below; however, there are still many unanswered questions relating to exceptional preservation.

It should be appreciated that there is no absolute Criterion to discriminate a Lagerstätte from a fossiliferous horizon. However, this is consistent with the notion that Lagerstätten are not distinct rock units, but are end-members of a range of sedimentary facies in which fossils are preserved.

Concentration Deposits

In these deposits it is the sheer abundance of fossils, although not often very well preserved, that is important and it is vitally important to reveal how such profusion occurred. Concentration may be a real reflection of community ecology or behaviour, or concentration may be a sedimentological artefact produced by time-averaging and physical processes, such as winnowing, which act to distillate skeletal remains. Two types of concentration deposits are considered in more detail below.

Stratiform Concentration Deposits

Where shelly fossils are locally abundant and form dense concentrations they may be referred to as 'shell beds' or 'coquinas'. They are formed in a variety of ways, reflecting a range of processes and time-scales. For example, they may form from a sudden, rapid influx of shells from a mass mortality, or from the slow accumulation of shells over many years during times of low deposition rates. An example to illustrate concentration by multiple, physical processes is the famous ammonite coquina of the Middle Jurassic (Normandy Coast, France) where beds contain different ammonites representing about 2 million years. The ammonites clearly do not reflect a single mass

mortality event, rather sedimentation to preserve the shells (so that they were not corroded) and probably several mixing events during storms that concentrated and yet did not destroy the shells.

Bone bed coquinas comprise an unusually rich concentration of phosphatic vertebrate remains, including bones, teeth, scales, and coprolites. Remains are not articulated and commonly include terrestrial and marine components. There are several models proposed to explain bone bed genesis. Winnowing involves episodes of erosion of the seafloor to rework and concentrate phosphatic material, which was previously more widely dispersed. Condensation involves extremely low sedimentation rates so that the input of phosphatic material is not diluted by sediment, but builds up on the seafloor over long time periods. Another model suggests that other biogenic material, composed of carbonate, may be dissolved, resulting in the apparent concentration of phosphatic material. Finally, in the transgressive lag model, a transgressive sea picks up phosphatic clasts from previously formed sediments and concentrates them into a basal bone bed. Examples of bone beds include the Westbury Formation of the Rhaetian Penarth Group, Wales and the Silurian of Ludlow, England.

Concentration Traps

In terrestrial and marine environments, fossils may be concentrated in protected environments such as cavities. Terrestrial cave and fissures constitute the most spectacular of these deposits where animals are concentrated either because they used the cavity as a dwelling, or because the cavity formed a death trap, successively killing and preserving animals over time. The Early Holocene Shield Trap Cave (Montana, USA) records many disarticulated bison bones and was formed in a bell-shaped limestone cave after roof collapse.

Conservation Deposits

The Importance of Exceptional Preservation

Entirely soft-bodied organisms or the soft parts of organisms usually rapidly disappear after death, due to decay processes and, consequently, most fossils comprise the remaining hard, mineralized parts of animals. This has two consequences: firstly, entirely soft-bodied animals are usually not preserved (except in conservation Lagerstätten); and secondly, if the animal has hard parts, these become disarticulated as the soft tissues no longer support them. Disarticulation may pose significant problems when trying to understand ancient life. Plants in particular undergo disarticulation when transported and because their component parts behave differently hydrodynamically their leaves, pollen, seed, fruit, bark, stems, roots, etc. often become widely separated, making it difficult to reconstruct the whole plant. Animals also suffer disarticulation with some, such as echinoderms, being more susceptible than others.

Conservation Lagerstätten (exceptionally preserved faunas) provide palaeontologists with a unique and unrivalled window through which to view the biology, ecology, and evolution of life. Up to two-thirds of modern shallow water marine communities are composed of entirely soft-bodied animals (*see* **Palaeoecology**). So conservation Lagerstätten are vital because they preserve a more complete picture of community diversity. However, it is important to determine how much of the original biota is represented in a fossil assemblage; in other words, is a particular taxon absent because it was not part of the community, or was it just not preserved? Understanding the taphonomy (processes of fossilization) of different conservation Lagerstätte, individual fossil specimens, and a range of different tissues types, is crucial in gauging preservational bias.

The preservation of soft tissues is also extremely helpful when trying to interpret animal affinities. Many animals have hard parts that are uninformative as to the underlying soft part anatomy. For example, conodonts (*see* **Microfossils:** Conodonts) (microscopic, phosphatic, tooth-like elements) were known for over 100 years before their affinities were resolved. This is because, although conodont elements are made of apatite and therefore have a good fossil record, the rest of the animal, to which these conodonts belonged, had never been found. Fortunately, in 1982, a set of conodont elements was reported with associated soft tissues from the Carboniferous Granton Shrimp Bed near Edinburgh. The analyses of these fossils led to the interpretation that conodonts were in fact vertebrates with an eel-like form, a notochord, V-shaped muscle blocks, a caudal fin, and that the conodont elements were located in the animal's oral cavity and functioned as a feeding apparatus.

Finally, the preservation of soft-bodied organisms provides palaeontologists with the evidence to explore biological phenomenon such as the Cambrian explosion, the nature of early metazoans, extinctions, and radiations.

Death, Decay, and Destruction

After death, several processes usually occur which would normally inhibit the potential of an animal or plant from becoming a fossil, and this is particularly the case for nonmineralised tissues. Necrolysis is the organic modification to carcasses, and occurs through autolysis, which is the breakdown of cells by their

own enzymes, scavenging which is the consumption of carcasses by macroscavengers, decomposition which is degradation by fungi, and putrefaction which is the destruction by bacteria. Biostratinomy describes the physical alterations that affect a carcass prior to burial, and after burial, diagenesis includes the chemical and mechanical changes that take place to fossils. Taphonomy is the study of all of these processes of fossilization, from death of the animal to its discovery as a fossil. In order for exceptional preservation to occur, the taphonomy must be exceptional to bypass decay and ultimately destruction – the normal fate of nonmineralized tissues.

Decay Experiments

Decay experiments have proved to be an extremely valuable tool to document and understand decay-induced morphological changes and the processes of fossilisation. Decay affects morphology and it is useful to compare the morphological features of a fossil with similar morphological features that have undergone decay, rather than the same features from an extant, living animal. Decay experiments have been undertaken on a range of taxa including vertebrates, from entirely soft-bodied hagfish (Figure 1) through to deer carcasses, invertebrates such as cnidarians, various worms, various echinoderms, cephalopods, crustaceans, insects, cephalochordates, and prokaryotes. In addition to assessing decay-induced morphological changes, these experiments have investigated the role of different factors (pH, oxygen concentration, temperature, salinity, and even microbe type) in the process of decay and fossilization. Decay experiments have provided important insights in to why some organic tissues survive decay and how other tissues become replaced very rapidly by authigenic minerals. In the future, understanding decay and fossilization processes may be used to predict the environmental conditions, and thus the type and location of sediments that may favour exceptional preservation.

Conservation Traps

There are two broad groups of conservation deposits: conservation traps and stratiform conservation deposits.

Conservation traps are deposits that are generally very geographically localised and are usually temporally restricted. They occur where animals and plants are trapped in a setting where macroscavengers are unable to access the remains, and the destruction by bacteria and enzymes is inhibited. Often the processes responsible for preservation in conservation traps are akin to those we employ to preserve our food, such as freezing, drying, and pickling. Examples include deep frozen mammoths from the Pleistocene found in clefts in the glacial permafrost in Siberia, mummified, desiccated ground sloths in caves in South America, and insects trapped in amber resin. Perhaps the most famous conservation trap is the Pleistocene Rancho La Brea tar pits of California, USA, which have preserved three million fossils including mammals, birds, fish, reptiles, invertebrates, and plants. In this respect, Rancho La Brea tar pits are a concentration, conservation trap. Conservation traps are temporally restricted because if conditions change, for example, if a carcass is re-hydrated or thawed, bacteria will resume their relentless destruction of the organic remains.

Stratiform Conservation Deposits (Table 1)

Stratiform conservation Lagerstätten are deposits where incomplete necrolysis occurs throughout the

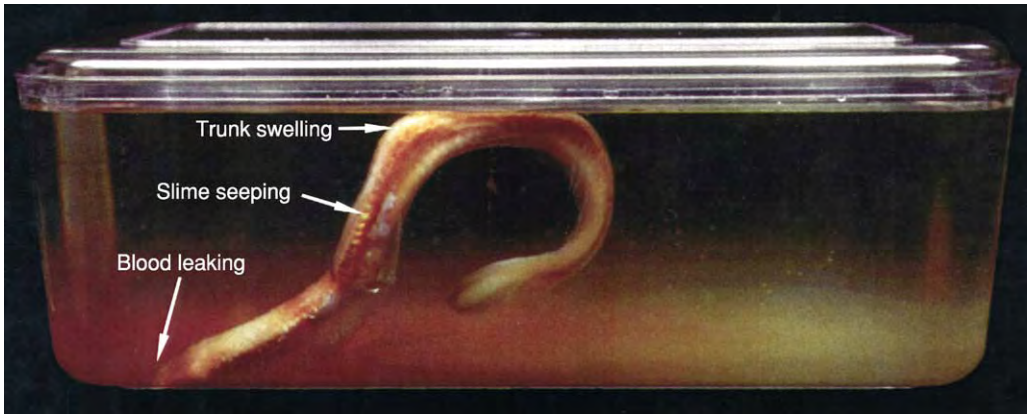


Figure 1 Decay experiment showing a hagfish in normal marine water just 6 days after death. Decay gases in the body cavity have resulted in the carcass floating. On day 7 of this experiment, the skin tore away and began to sink; disintegration of the carcass occurred shortly afterwards. This experiment demonstrates that exceptional preservation requires that a carcass be rapidly buried, or kept on the substrate by other means, such as by microbial mats, and that exceptional preservation must occur very rapidly after death. (Experiment and photographs by Dr K Freedman.)

Table 1 A small selection of conservation Lagerstatten with age, location, and brief comments on the environmental conditions and preservation of soft parts (sp.)

Lagerstätten	*Age*	*Location*	*Brief comments*
Doushantuo Formation	Latest Proterozoic	South China	One deposit in carbonaceous shales, another in phosphorites phosphatisation of sp.
Chengjiang	Early Cambrian	Yunnan Province, China	Burgess Shale type fauna, microturbidites various mineral replacement of sp. and possible organic remains
Burgess Shale	Middle Cambrian	British Columbia, Canada	Burgess Shale type fauna, obrution and anoxia, kerogenised organic films and clay mineral replacement of sp.
Orsten	Late Cambrian	Southern Sweden, Germany and Poland	Anthraconite limestone nodules in bituminous shale, apatite (phosphatic) coating and replacement of sp.
Soom Shale	Late Ordovician	Cape Province, South Africa	Euxinic and anoxic bottom and porewaters, clay mineral replacement of sp.
Herefordshire	Silurian	Herefordshire, UK	Calcite nodules in volcanic ash deposit after decay three dimensional animal void infilled by calcite
Hunsruck Slate	Early Devonian	Western Germany	Obrution by microturbidites pyrite replacement of sp.
Rhynie Chert	Early Devonian	Eastern Scotland	Plants and arthropods engulfed in Si rich hotspring water silicification of sp.
Mazon Creek	Pennsylvanian Carboniferous	Illinois, USA	Rapid burial in oxygen depleted sediment and siderite nodules sp. represented as highly compressed, light on dark impressions
Monte San Giorgio	Middle Triassic	Southern Switzerland	Bottom waters anoxic, but surface waters normal marine rare phosphatised sp., mostly articulated skeletons
Posidonia Shale	Lower Jurassic	Holzmaden region, Germany	Anoxic bottom waters, soupy sediment and occasional sediment blanketing, sp. of belemnites, cephalopods
Solnhofen Limestone	Upper Jurassic	Southern Germany	Hypersaline, oxygen depleted bottom waters, soupy sediment obrution and microbial mats. Soft parts mostly impressions rare phosphatisation of sp. and organic residues
Santana Formation	Lower Cretaceous	NE Brazil	Hypersaline waters, soupy sediment, and mass mortalities many carbonate concretions and sp. phosphatised
Grube Messel	Mid Eocene	West Germany	Quiet, anoxic lacustrine waters silhouettes of sp. made of autolithified, sideritic bacterial films
Baltic Amber	Late Eocene Early Oligocene	Baltic, NW Europe	Amber is fossilized resin from trees, which traps arthropods; resin stops bacterial and fungal decay and acts as desiccant and antibiotic

sediment layer(s) rather than being constrained to special, localized settings, as is the case for conservation traps.

Scavenging by macro-organisms is a ubiquitous event in most environments and leads to disarticulation, and sometimes wide dispersal of the animal hard parts, and the removal of soft parts. Terrestrial and aquatic environments abound with macroscavengers and they may devour soft tissues in surprisingly little time. For example, a fish carcass introduced to a cold (4°C), dysaerobic seabed off the coast of California was stripped of soft parts and disarticulated in 2–3 days by brittle stars. Clearly, one of the most important prerequisites for exceptional preservation is the precludement of macroscavengers from carcasses and plants. This may be accomplished in a variety of ways, some of which are illustrated below, using examples from fossil conservation Lagerstätten.

Obrution Catastrophic, rapid burial, by turbidites or tempestites, may help to preserve nonmineralised tissues because it protects the animal (which may be buried dead or alive) from macroscavengers and bioturbators, (*see* **Trace Fossils**) and it tends to induce anoxia (see below). However, whilst there are fewer bacteria with depth, and anaerobic bacteria are less efficient at breaking down organic carbon, a rapidly buried carcass still requires other processes to occur to inhibit or circumvent the decay of organic remains. Sediment smothering often produces an ecological bias because it will affect bottom-living organisms more than nektonic ones. In the Phyllopod Bed of the Middle Cambrian Burgess Shale (British Columbia, Canada) there is evidence to suggest that much of the biota was catastrophically buried, such as graded beds and variable orientations of specimens relative to bedding (Figure 2). Furthermore, Phyllopod Bed sediments are parallel laminated (i.e., undisturbed) suggesting that, owing to anoxia, conditions in the sediment porewaters were inimical to bioturbators and scavengers.

Soupy substrates Burial owing to high rates of sedimentation is not the only way of rapidly covering a

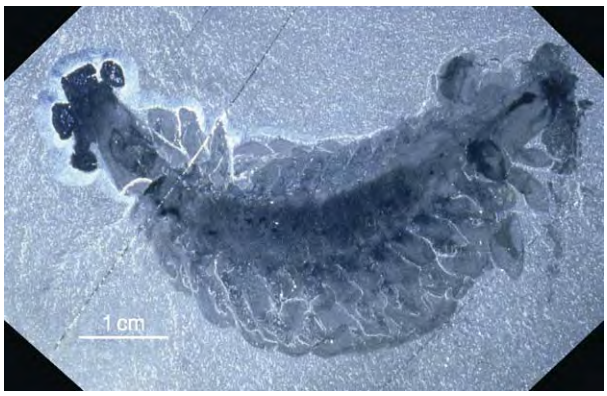


Figure 2 *Opabinia* from the Middle Cambrian Stephan Formation, Burgess Shale, Canada. Only three of the five eyes are evident and the flexible proboscis is still buried, and may be curved back beneath the head. The trunk consists of 15 segments, each bearing a pair of lateral lobes.

carcass and thus prohibiting many macroscavengers. If the sediment is soft, or even soupy, carcasses can sink into the sediment and thus be buried. Articulated vertebrate skeletons and the preservation of the soft parts of ichthyosaurs of the Jurassic Oxford Clay (UK) are suggested to have occurred because carcasses sunk into a soupy seafloor. This example is fascinating because areas of vertebrate carcasses that protruded above the soupy sediment were often encrusted and bored, suggesting that vertebrate skeletons provided 'benthic islands' of hard substrate, amid the otherwise inhospitable, soupy seafloor.

Stagnation, anoxia Anoxia is the absence of dissolved oxygen and results if the rate of carbon deposition exceeds that of oxygen supply. Euxinic conditions describe waters that are anoxic and sulphidic. Benthic organisms are prohibited from living in such environments as are bioturbators, resulting in sediments with parallel (i.e., undisturbed) laminations. Carcasses falling into such bottom water conditions have an enhanced potential for exceptional preservation for several reasons. Firstly, macroscavengers would not be able to devour and disarticulate the carcass. Furthermore, anaerobic bacteria putrefy soft tissues more slowly and so increase the time for rapid mineralization (see below) to replace the tissues. In addition, anaerobic bacteria may not be able to utilise very recalcitrant organic matter, which then has the potential to survive into the fossil record as an organic residue. A classic example of a deposit where anoxia and stagnation led to exceptional preservation is the Jurassic Posidonia Shales (Holzmaden, Germany). Here black, bituminous shales containing extremely rare benthonic organisms were overlain by an upper water body rich in oxygen and life. The fauna includes plants and many invertebrates, but the shale is most famous for its complete vertebrate skeletons such as sharks, ichthyosaurs, crocodiles, and pterosaurs. However, anoxia cannot account for the occurrence of the articulated crinoids, because if they lay unprotected on the seafloor, albeit an anoxic one, they would disarticulate in a matter of days. It has been suggested, therefore, that rapid sediment smothering also occurred in the Posidonia Shales.

It should be appreciated that exceptional preservation in several conservation Lagerstätten can be attributable to more than one environmental circumstance. The Jurassic Solnhofen Limestone (Bavaria, Germany) was deposited in a basin where the stratified bottom waters were inhospitable to life owing to high salinity and low oxygen content. Periodic storm events washed animals into the basin, which were then rapidly buried by carbonate ooze. So rapid burial, hypersalinity, and anoxic stratified waters may all have contributed to exceptional preservation. Moreover, on the basis of the anatomy of the commonly preserved *Saccocoma*, a crinoid, it has been suggested that periodically the bottom waters were not saline, but that the sediment was soupy, allowing carcasses to sink into the limey muds.

Bacteria

Even where organisms are protected from macroscavengers, their soft parts are still vunerable, owing to the action of autolytic enzymes and bacteria. In the presence of oxygen, bacteria decompose soft tissues rapidly. Once oxygen has been used up, decay proceeds anaerobically. In freshwater environments, anaerobic decay is dominated by nitrate reduction and methanogenesis, whereas in the marine environment sulphate reduction and methanogenesis dominate. These reactions all lead to the decomposition of soft tissues, although there is evidence to suggest that anaerobic decay proceeds more slowly and is less efficient than aerobic decay. Decay may be inhibited by microbial mats (*see* **Biosediments and Biofilms**) because these mats effectively limit the diffusion of toxic bacterial metabolic by-products away from the site of decomposition, and the influx of oxygen and sulphate (marine environments) needed to fuel the bacteria. Mats may also aid soft part preservation by trapping organisms on the seafloor and creating a 'closed' microenvironment that may well favour rapid mineralization processes (see below). Under certain conditions, bacterial decomposition may also be retarded in fine-grained mudrocks by the prevalence of clay minerals. It has been suggested that the digestive enzymes used by bacteria in decomposition may become adsorbed onto specific clay minerals and thus rendered inert. Such a model has been proposed

to explain the preservation of organic films in the Burgess Shale (Figure 2). However, some of the highest fidelity of soft-part preservation comes about when the soft tissues are mineralized before they are decomposed.

Soft Tissue Mineralization

Nonmineralized tissues display a spectrum of resistance to decay depending on their physical and chemical makeup. Decay-resistant (sometimes called recalcitrant) tissues, such as chitin of arthropod cuticles and woody plant tissues, may retain their organic composition as fossils. Decay-susceptible (sometimes called labile) tissues cannot survive unless authigenic minerals rapidly replace them, a process sometimes referred to as permineralisation. Mineralization can be extremely rapid, occurring within days, weeks, and months of death. Authigenic mineralization is controlled by a number of factors, including the organic makeup of the soft tissues, the geochemistry of the sediment, the pH and Eh of the porewaters and the concentrations of mineral forming ions. In addition, bacteria may mediate mineralization by breaking bonds, speeding up reactions, and concentrating authigenic mineral ions on to the tissue surfaces – the exact nature of their participation is not yet understood. In some deposits it is the bacteria themselves that autolithify, and they can pseudomorph details of the underlying animal morphology. For example, in the Eocene oil shale of Grube Messel (Germany) exquisite details, such as single hairs or the barbules of feathers, have been preserved as layers of lithified bacteria.

Only a few minerals are known to replace soft tissues, the most important of which are apatite, pyrite, clay minerals, and silica.

Apatite

Apatite is the mineral most often involved in replacing soft tissues and it does so with the highest degree of fidelity. Details, down to the subcellular level, such as Cretaceous fish cell nuclei sitting on muscle fibres, have been recorded and such fine detail is possible because the apatite crystallites can be very small (commonly <30 nm). Apatite mineralization requires a sufficient concentration of phosphate, which may be released from the decaying animal, or concentrated in the surrounding sediment. Decay experiments have also shown that slightly acidic pH favours apatite authigenesis over carbonate authigenesis, which does not have the same potential to replace soft tissues. Experimental phosphatisation (the replacement of soft tissues by apatite) and the analyses of experimentally produced and fossil apatite

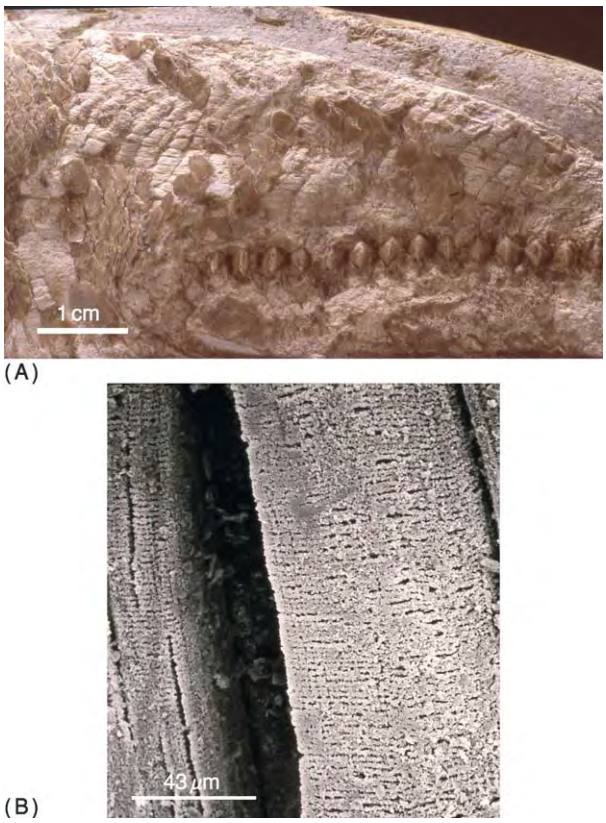


Figure 3 Colour photograph shows the elopomorph fish *Notelops* sp. with phosphatised striated muscle tissue (white fibrous material). Black and white photograph shows phosphatised muscle fibres from another *Notelops* sp, obtained using a Scanning Electron Microscope. Both specimens are from the Early Cretaceous Romualdo Member, Santana Formation, NE Brazil. (Photographs by Dr D Martill.)

textures has greatly improved our understanding of apatite authigenesis. Some of the more spectacular examples of soft tissues replaced by apatite include: delicately preserved tiny arthropods in nodules from the Upper Cambrian 'Orsten' (northern and central Europe) and fish muscles (Figure 3), gill filaments, and eggs from the Cretaceous Santana Formation (Brazil).

Pyrite

Pyrite is commonly formed in the sediment of anoxic black shales but it is relatively rarely involved in mineralizing animal soft tissues. Although it is quite commonly associated with the preservation of plant remains where it may preserve cellular details (e.g., Eocene London Clay of England). Pyrite crystals are generally larger than apatite and clay minerals, and this constrains the fidelity of pyritised soft tissues. Pyrite commonly fills voids or coats the gross morphology or outline of tissues and so delicate, rapidly decayed tissues, such as muscle are not known to be

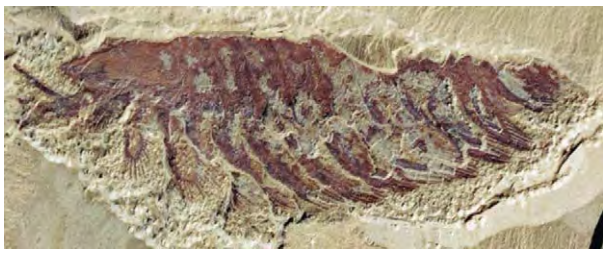


Figure 4 A specimen of the arthropod *Leancholia* from the Early Cambrian Chengjiang Formation, China, 2.3 cm in length. (Photo graph by Dr Derek Siveter.)

pyritised. Replacing animal tissues by pyrite requires that pyrite must grow on the carcass rather than throughout the sediment. In order to do this the sediment must have high concentrations of iron in porewater solutions and low total organic matter. This allows sulphide generated from seawater sulphate by the decay of soft tissues to be trapped instantly by formation of iron sulphide minerals within, or very close to the carcass. An example of soft part pyritisation in seen in the Early Devonian Hunsrück Slate (Germany) where several taxa (including echinoderms, molluscs, arthropods, vertebrates, and cnidarians) are beautifully preserved with some relief owing to early pyritisation. In addition, the soft parts of several taxa from the Early Cambrian Chengjiang biota of China were preserved by pyrite, which later oxidised to create the characteristic pink and orange fossils (Figure 4).

Clay Minerals

Clay minerals (*see* **Clay Minerals**) are capable of replacing soft tissues with fidelity slightly less than that of apatite but considerably better than that of pyrite. Clays have an affinity for organic matter when certain geochemical conditions are met. It has been demonstrated that a particularly low pH may be important in this respect, and may explain why, although clay minerals are ubiquitous in fine-grained sediments, clay mineral replacement of soft tissues is not. It is possible that detrital, colloidal clays which are extremely small (1 nm to 1 μm across) may become attracted towards, and template on to decaying soft tissues. Or clay authigenesis may be responsible for soft tissue preservation. In either case, it has been suggested that bacteria play an important role in somehow controlling and mediating the process. Several taxa from the Upper Ordovician Soom Shale of South Africa show soft tissue preservation by clay minerals, including the muscle fibres and fibrils of a conodont animal. Whilst it has been shown that the Middle Cambrian Burgess Shale fossils of Canada are composed of kerogen (or another graphite-like structure), clay minerals are also involved in their preservation and are seen to replace and coat organic surfaces.

Other Minerals

Silica and calcium carbonate are also involved in mineralising soft tissues. The Earth's oldest Lagerstätte, the Apex Cherts (*see* **Precambrian:** Prokaryote Fossils) (Warrawoona Group, Western Australia) contain silicified microfossils 3450–3470 Ma in age. The Early Devonian Rhynie Chert of Scotland preserves terrestrial arthropods and plants in superb detail. Here chert beds contain beautiful, three-dimensional plant material with preserved cell structures which became silicified when plants were engulfed in Si-rich hotspring waters (*see* **Sedimentary Rocks:** Chert).

The Role of Nodules

In a number of conservation Lagerstätten, early diagnetic concretions (nodules) formed around the fossil and protected it from later compaction. The Silurian Herefordshire fauna (UK) has produced spectacular three-dimensionally preserved fossils, owing to early stiffening of decaying animals by volcanic ash, followed by calcite infilling of the void left by the decayed tissues, and precipitation of calcite nodules around the fossil. A novel technique was developed to extract the maximum amount of information from these three-dimensional fossils. The nodules containing the fossils are serially ground at 30 μm intervals, and digitally photographed so that a set of 'slice images' of the whole fossil is created. These images are then used to create impressive 3D computerised reconstructions of the fossils (Figure 5). Other deposits where 3D preservation occurs owing to concretions include a diverse biota from the Carboniferous Mazon Creek (Illinois, USA) in siderite concretions, and vertebrates and invertebrates from the Lower Cretaceous Santana Formation (NE Brazil) in carbonate concretions (Figure 6).

Temporal Trends in Exceptional Preservation?

It has been suggested that the conditions required to create fossil Lagerstätten have not been uniformly present through the time of metazoans on Earth. In fact, taphonomic windows have appeared, leading to unique exceptionally preserved faunas, and then been 'closed'. The preservation potential of Cambrian, Ordovician, and Silurian organisms is not greatly different, and yet the Cambrian is relatively replete in fossil Lagerstätten compared with the other two periods. This may reflect a temporal difference in the processes of preservation. For example, it has been suggested that during the Cambrian, deep bioturbators had yet to invade deeper water settings,

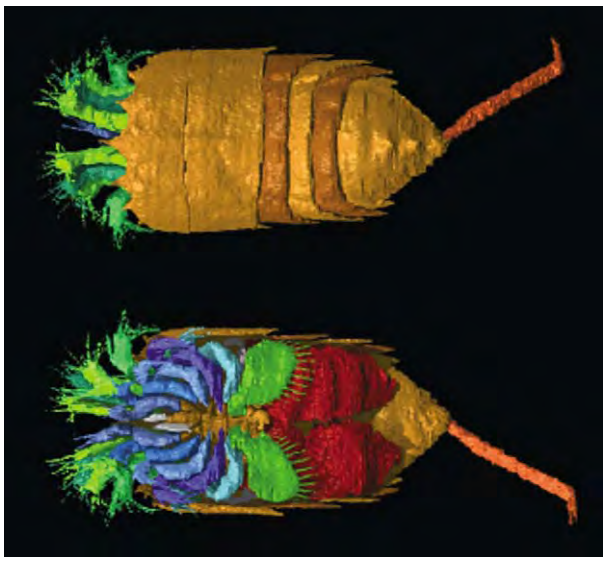


Figure 5 A computer based reconstruction of the three dimensionally preserved arthropod *Offacolus kingi* from the Silurian Herefordshire (UK) Lagerstatte. Top is dorsal view, bottom is ventral view. Carapace *ca* 2 mm in width. (Image by Dr Mark Sutton.)

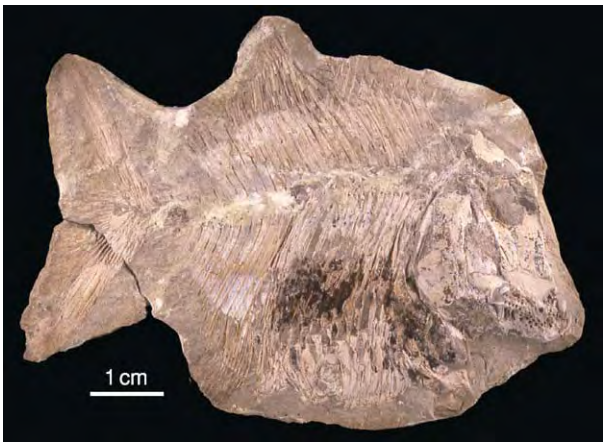


Figure 6 A pycnodont fish preserved in an early diagenetic carbonate nodule showing fully articulated skeleton. Specimen from the Early Cretaceous Romualdo Member, Santana Formation, NE Brazil. (Photograph by Dr Dave Martill.)

but after the Cambrian they radiated from onshore to offshore, decreasing the likelihood of organic preservation after the Cambrian. There are several other examples where exceptional preservation appears to be favoured during particular time periods.

See Also

Biosediments and Biofilms. **Clay Minerals**. **Fossil Invertebrates:** Arthropods. **Microfossils:** Conodonts. **Palaeoecology**. **Precambrian:** Prokaryote Fossils. **Sedimentary Rocks:** Chert; Phosphates. **Trace Fossils**.

Further Reading

Aldridge RJ, Gabbott SE, and Theron JN (2002) The Soom Shale. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology II*, pp. 340–342. Oxford: Blackwell Science.

Allison PA and Briggs DEG (1991) Taphonomy of non-mineralized tissues. In: Allison PA and Briggs DEG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record*, pp. 26–71. New York and London: Plenum Press.

Bartels C, Briggs DEG, and Brassel G (1998) *The Fossils of the Hunsrück Slate: Marine Life in the Devonian*. Cambridge: Cambridge University Press.

Barthel KW, Swinburne NHM, and Conway Morris S (1990) *Solnhofen: A Study in Mesozoic Palaeontology*. Cambridge: Cambridge University Press.

Bottjer DJ, Etter W, Hagadorn JW, and Tang C (2002) *Exceptional Fossil Preservation – A Unique View on the Evolution of Marine Life*. New York: Columbia University Press.

Briggs DEG (2002) Exceptionally preserved fossils. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology II*, pp. 328–332. Oxford: Blackwell Science.

Briggs DEG (2003) Annual Review of Earth and Planetary Science. *Earth and Planetary Science* 31: 275–301.

Briggs DEG, Erwin DH, and Collier FJ (1994) *Fossils of the Burgess Shale*. Washington and London: Smithsonian Institution Press.

Butterfield NJ (1990) Organic preservation of non-mineralizing organisms and the taphonomy of the Burges Schale: *Palaeobiology*, v. 16, pp. 272–286.

Donovan SK (1991) *Processes of fossilization*. London: Belhaven Press.

Hou X G, Aldridge RJ, Bergström J, Siveter DJ, and Feng X H (2004) *The Cambrian Fossils of Chengjiang, China; The Flowering of Early Animal Life*. Oxford: Blackwell Science Ltd.

Kidwell SM (1991) The stratigraphy of shell concentrations. In: Allison PA and Briggs DEG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record*, pp. 212–279. New York and London: Plenum Press.

Knoll AH and Shuhai X (2002) Precambrian Lagerstätten. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology II*, pp. 332–337. Oxford: Blackwell Science.

Martill DM (1993) *Fossils of the Santana and Crato Formations, Brazil*. Dorcherter: Henry Ling Ltd.

Müller KJ (1990) Taphonomy of Fossil Lagerstätten, Upper Cambrian 'Orsten.' In: Briggs DEG and Crowther PR (eds.) *Palaeobiology I*, pp. 274–277. Oxford: Blackwell Scientific Publications.

Selden PA and Nudds JR (2004) *Evolution of Fossil Ecosystems*. London: Manson Publishing Ltd.

Seilacher A (1990) Taphonomy of Fossil Lagerstätten, Overview. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology I*, pp. 266–270. Oxford: Blackwell Scientific Publications.

Spicer RA (1991) Plant taphonomic processes. In: Allison PA and Briggs DEG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record*, pp 72 115. New York and London: Plenum Press.

Sutton MD, Briggs DG, Siveter D, and Siveter D (2001) Methodologies for the visualization and reconstruction of three dimensional fossils from the Silurian Hereford shire Lagerstätte. *Palaeontologica Electronica* 4(1). http://laeoelectronica.org/2001 1/ s2/issue1 01.htm

Sutton MD, Briggs DEG, Siveter DJ, Siveter DJ, and Orr PJ (2002) The arthropod *Offacolus kingi* (Chelicerata) from the Silurian of Herefordshire, England: computer based morphological reconstructions and phylogenetic affinities. *Proceedings of the Royal Society, London B 269*: 1195 1203.

Trewin RH (2002) The Rhynie Chert. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology II*, pp. 342 346. Oxford: Blackwell Science Ltd.

LARGE IGNEOUS PROVINCES

M F Coffin, University of Tokyo, Tokyo, Japan
O Eldholm, University of Bergen, Bergen, Norway

Introduction

Large igneous provinces (LIPs) are massive crustal emplacements of predominantly iron- and magnesium-rich (mafic) rock that form by processes other than normal seafloor spreading; they are the dominant form of near-surface magmatism on the terrestrial planets and moons of our solar system. On the Earth's surface, LIP rocks are readily distinguishable from the products of the two other major types of magmatism – mid-ocean ridge magmatism and arc magmatism – on the basis of petrologic, geochemical, geochronological, geophysical, and physical volcanological data. LIPs occur both on the continents and in the oceans, and include continental flood basalts, volcanic passive margins, oceanic plateaus, submarine ridges, seamount chains, and ocean-basin flood basalts (Figure 1 and Table 1). LIPs and their contemporary small-scale analogues, hotspot volcanoes, are commonly attributed to decompression melting of hot low-density mantle material ascending from the Earth's interior in mantle plumes, and thus provide a window onto mantle processes. This type of magmatism currently accounts for about 10% of the mass and energy flux from the Earth's deep interior to its crust. The flux may have been higher in the past, but is episodic over geological time, in contrast to the relatively steady-state activity at seafloor spreading centres. Such episodicity reveals dynamic non-steady-state circulation within the Earth's mantle, perhaps

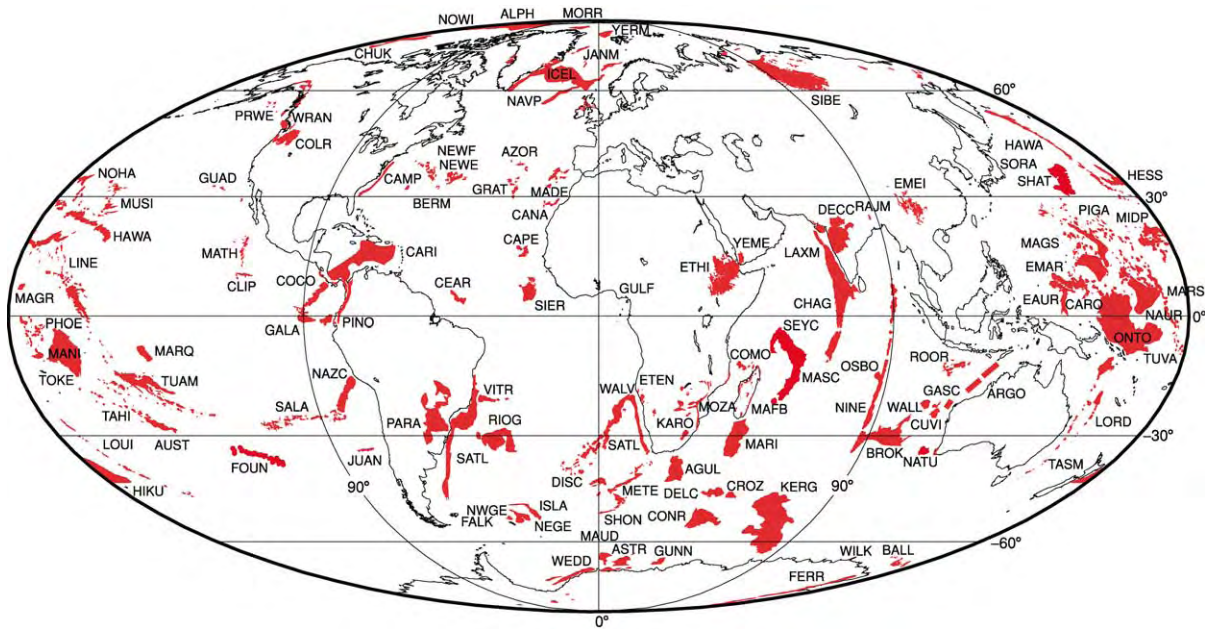


Figure 1 Phanerozoic global LIP distribution (red), with LIPs labelled (*see* Table 1).

Table 1 Large igneous provinces

Large igneous province	*Abbreviation (see Figure 1)*	*Type*
Agulhas Ridge	AGUL	SR
Alpha Mendeleyev Ridge	ALPH	SR/OP
Argo Basin	ARGO	VM
Astrid Ridge	ASTR	VM
Austral Seamounts	AUST	SMT
Azores	AZOR	SMT
Balleny Islands	BALL	SMT
Bermuda Rise	BERM	OP
Broken Ridge	BROK	OP
Canary Islands	CANA	SMT
Cape Verde Rise	CAPE	OP
Caribbean Flood Basalt	CARI	OBFB (partly accreted)
Caroline Seamounts	CARO	SMT
Ceara Rise	CEAR	OP
Central Atlantic Magmatic Province (VM only)	CAMP	CFB/VM
Chagos Laccadive Ridge	CHAG	SR
Chukchi Plateau	CHUK	VM
Clipperton Seamounts	CLIP	SMT
Cocos Ridge	COCO	SR
Columbia River Basalt	COLR	CFB
Comores Archipelago	COMO	SMT
Conrad Rise	CONR	OP
Crozet Plateau	CROZ	OP
Cuvier (Wallaby) Plateau	CUVI	VM
Deccan Traps	DECC	CFB/VM
Del Caño Rise	DELC	OP
Discovery Seamounts	DISC	SMT
East Mariana Basin	EMAR	OBFB
Eauripik Rise	EAUR	OP
Emeishan Basalts	EMEI	CFB
Etendeka	ETEN	CFB
Ethiopian Flood Basalt	ETHI	CFB
Falkland Plateau	FALK	VM
Ferrar Basalts	FERR	CFB
Foundation Seamounts	FOUN	SMT
Galapagos Carnegie Ridge	GALA	SMT/SR
Gascoyne Margin	GASC	VM
Great Meteor Atlantis Seamounts	GRAT	SMT
Guadelupe Seamount Chain	GUAD	SMT
Gulf of Guinea	GULF	VM
Gunnerus Ridge	GUNN	VM
Hawaiian Emperor Seamounts	HAWA	SMT
Hess Rise	HESS	OP
Hikurangi Plateau	HIKU	OP
Iceland Greenland Scotland Ridge	ICEL	OP/SR
Islas Orcadas Rise	ISLA	SR
Jan Mayen Ridge	JANM	VM
Juan Fernandez Archipelago	JUAN	SMT
Karoo	KARO	CFB
Kerguelen Plateau	KERG	OP/VM
Laxmi Ridge	LAXM	VM
Line Islands	LINE	SMT
Lord Howe Rise Seamounts	LORD	SMT
Louisville Ridge	LOUI	SMT
Madagascar Flood Basalts	MAFB	CFB
Madagascar Ridge	MARI	SR/VM?
Madeira Rise	MADE	OP
Magellan Rise	MAGR	OP
Magellan Seamounts	MAGS	SMT
Manihiki Plateau	MANI	OP
Marquesas Islands	MARQ	SMT
Marshall Gilbert Seamounts	MARS	SMT
Mascarene Plateau	MASC	OP
Mathematicians Seamounts	MATH	SMT
Maud Rise	MAUD	OP
Meteor Rise	METE	SR
Mid Pacific Mountains	MIDP	SMT
Morris Jesup Rise	MORR	VM
Mozambique Basin	MOZA	VM
Musicians Seamounts	MUSI	SMT
Naturaliste Plateau	NATU	VM
Nauru Basin	NAUR	OBFB
Nazca Ridge	NAZC	SR
New England Seamounts	NEWE	SMT
Newfoundland Ridge	NEWF	VM
Ninetyeast Ridge	NINE	SR
North Atlantic Volcanic Province	NAVP	CFB
North east Georgia Rise	NEGE	OP
North west Georgia Rise	NWGE	OP
North west Hawaiian Ridge	NOHA	SR/SMT
Northwind Ridge	NOWI	SR
Ontong Java Plateau	ONTO	OP (partly accreted)
Osborn Knoll	OSBO	OP
Paraná	PARA	CFB
Phoenix Seamounts	PHOE	SMT
Pigafetta Basin	PIGA	OBFB
Piñón Formation (Ecuador)	PINO	OP (accreted)
Pratt Welker Seamounts	PRWE	SMT
Rajmahal Traps	RAJM	CFB
Rio Grande Rise	RIOG	OP
Roo Rise	ROOR	OP
Sala y Gomez Ridge	SALA	SR
Seychelles Bank	SEYC	VM
Shatsky Rise	SHAT	OP
Shona Ridge	SHON	SR
Siberian Traps	SIBE	CFB
Sierra Leone Rise	SIER	OP
Sorachi Plateau (Japan)	SORA	OP (accreted)
South Atlantic Margins	SATL	VM
Tahiti	TAHI	SMT
Tasmantid Seamounts	TASM	SMT
Tokelau Seamounts	TOKE	SMT
Tuamotu Archipelago	TUAM	SMT
Tuvalu Seamounts	TUVA	SMT
Vitória Trindade Ridge	VITR	SR/SMT
Wallaby Plateau (Zenith Seamount)	WALL	OP
Walvis Ridge	WALV	SR
Weddell Sea	WEDD	VM
Wilkes Land Margin	WILK	VM
Wrangellia	WRAN	OP (accreted)
Yemen Plateau Basalts	YEME	CFB
Yermak Plateau	YERM	VM

CFB, continental flood basalt; OBFB, ocean basin flood basalt; OP, oceanic plateau; SMT, seamount; SR, submarine ridge; VM, volcanic margin.

extending far back into Earth history, and suggests a strong potential for LIP emplacements to contribute to, if not instigate, major environmental changes.

Composition, Physical Volcanology, Crustal Structure, and Mantle Roots

LIPs are defined by the characteristics of their dominantly iron- and magnesium-rich (mafic) extrusive rocks; these typically consist of subhorizontal subaerial basalt flows. Individual flows can extend for hundreds of kilometres, be tens to hundreds of metres thick, and have volumes as great as 10^4–10^5 km^3. Silica-rich rocks also occur as lavas and intrusive rocks and are usually associated with the initial and late stages of LIP magmatic activity. Relative to mid-ocean-ridge basalts, LIPs include higher MgO lavas, basalts with more diverse major-element compositions, rocks with more common fractionated components, both alkalic and tholeiitic differentiates, basalts with predominantly flat light-rare-earth-element patterns, and lavas erupted in both subaerial and submarine settings.

As the extrusive component of LIPs is the most accessible for study, nearly all of our knowledge of LIPs is derived from the lavas forming their uppermost crusts. The extrusive layer may exceed 10 km in thickness. On the basis of geophysical, predominantly seismic, data from LIPs and from comparisons with normal oceanic crust, LIP crust beneath the extrusive layer is believed to consist of an intrusive layer and a lower crustal body, characterized by P-wave velocities of 7.0–7.6 km s^{-1}, at the base of the crust (Figure 2). Beneath continental crust this body may be considered as a magmatically underplated layer. Seismic-wave velocities suggest an intrusive layer that is probably gabbroic and a lower crust that is ultramafic. If the LIP forms on pre-existing continental or oceanic crust or along a divergent plate boundary, dikes and sills are probably common in the middle and upper crust. The maximum crustal thickness, including extrusive and intrusive layers and the lower crustal body, of an oceanic LIP is about 35 km, as determined from seismic and gravity studies of the Ontong Java Plateau (Figure 1 and Table 1).

Low-velocity zones have recently been observed in the mantle beneath the oceanic Ontong Java Plateau and the continental Deccan Traps and Paraná flood basalts (Figure 1 and Table 1). Interpreted as lithospheric roots or keels, these zones can extend at least 500–600 km into the mantle. In contrast to the high-velocity roots beneath most continental areas and the absence of lithospheric keels in most oceanic areas, the low-velocity zones beneath LIPs apparently reflect primarily residual chemical, and perhaps some thermal, effects of mantle-plume activity. High-buoyancy roots extending well into the mantle beneath oceanic LIPs would suggest a significant role for LIPs in continental growth via accretion of oceanic LIPs to the edges of continents.

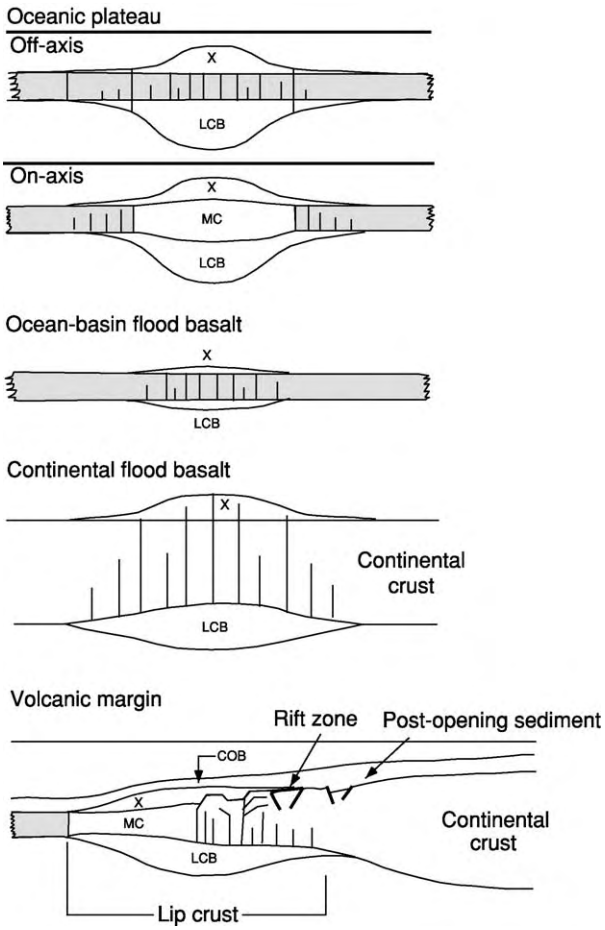


Figure 2 Schematic LIP plate tectonic settings and gross crustal structure. The LIP crustal components are extrusive cover (X), middle crust (MC), and lower crustal body (LCB). Normal oceanic crust is shown in grey, and intrusives are denoted by vertical lines. The continent ocean boundary (COB) is indicated for the volcanic margin.

Distribution, Tectonic Setting, and Types

LIPs occur worldwide, in both continental and oceanic crust, in purely intraplate settings and along present and former plate boundaries (Figure 1 and Table 1), although the tectonic setting at the time of formation is unknown for many features. If a LIP forms at a plate boundary, the entire crustal section is LIP crust (Figure 2). Conversely, if one forms in an intraplate setting, the pre-existing crust must be intruded and sandwiched by LIP magmas, albeit to an

extent that is not resolvable by current geological or geophysical techniques.

Continental flood basalts, which are the most intensively studied LIPs owing to their exposure, are erupted from fissures in continental crust (Figure 1 and Table 1). Most continental flood basalts overlie sedimentary basins that formed via extension, but it is not clear what happened first, the magmatism or the extension. Volcanic passive margins form as a result of excessive magmatism during continental breakup along the trailing rifted edges of continents. In the deep ocean basins, four types of LIPs are found. Oceanic plateaus, commonly isolated from the major continents, are broad typically flat-topped features generally lying 2000 m or more above the surrounding seafloor. They can form at triple junctions (e.g. the Shatsky Rise), at mid-ocean ridges (e.g. Iceland), or in intraplate settings (e.g. the northern Kerguelen Plateau). Submarine ridges are elongated steep-sided elevations of the seafloor. Some form along transform plate boundaries (e.g. the Ninetyeast Ridge). In the oceanic realm, oceanic plateaus and submarine ridges are the most enigmatic LIPs with respect to the tectonic setting in which they formed. Seamounts, which are closely related to submarine ridges, are local elevations of the seafloor; they may be discrete, form linear or random groups, or be connected along their bases and aligned along a ridge or rise (*see* **Seamounts**). They commonly form in intraplate regions (e.g. Hawaii). Ocean-basin flood basalts (e.g. the Nauru Basin and the Caribbean province) are the least-studied type of LIP and consist of extensive submarine flows and sills lying above and postdating the normal oceanic crust.

Ages

Age control for all LIPs apart from continental flood basalts is poor owing to their relative inaccessibility, but the $^{40}Ar/^{39}Ar$ dating technique is having a particularly strong impact on studies of LIP volcanism. Geochronological studies of continental flood basalts (e.g. the Siberian Traps, the Karoo, the Ferrar Basalts, the Deccan Traps, and the Columbia River Basalt; Figure 1) suggest that most LIPs result from mantle plumes, which initially transfer huge volumes (*ca.* 10^5–10^7 km^3) of mafic rock into localized regions of the crust over short intervals (*ca.* 10^5–10^6 years) but which subsequently transfer mass at a far lesser rate, albeit over significantly longer intervals (10^7–10^8 years). Transient magmatism during LIP formation is commonly attributed to mantle-plume 'heads' reaching the crust following transit through all or part of the Earth's mantle, whereas persistent magmatism is considered to result from steady-state mantle-plume 'tails' penetrating the lithosphere, which is moving relative to the plume (Figure 3). However, not all LIPs have obvious connections with mantle plumes or even hotspot tracks, suggesting that more than one source model may be required to explain all LIPs.

LIPs are not distributed uniformly over time. For example, many LIPs formed between 50 Ma and 150 Ma, whereas few have formed during the past 50 Myr (Figure 4). Such episodicity probably reflects variations in rates of mantle circulation, and this is supported by high rates of seafloor spreading during a portion of the 50–150 Ma interval, specifically during the long Cretaceous Normal Superchron

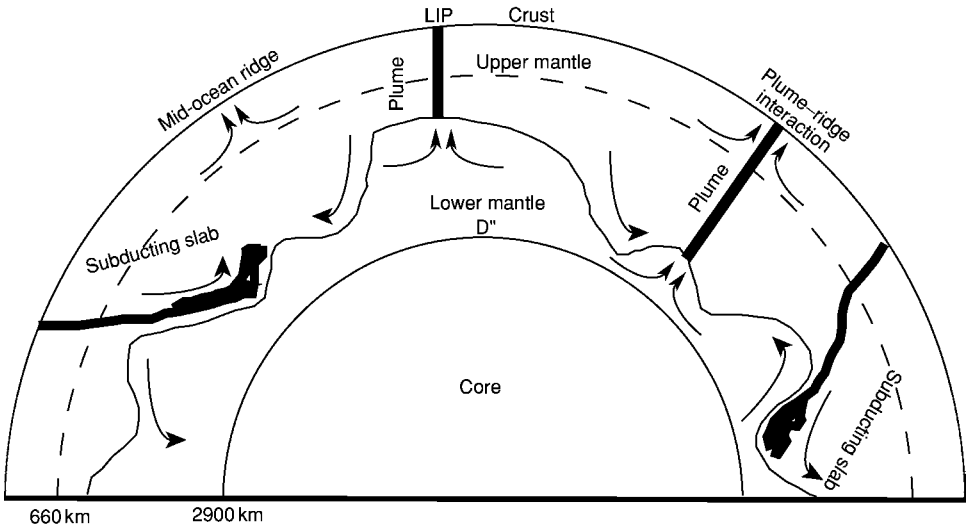


Figure 3 Diagram of the Earth's interior, showing plumes (tails), subducting slabs, and two mantle layers that move in complex patterns but never mix.

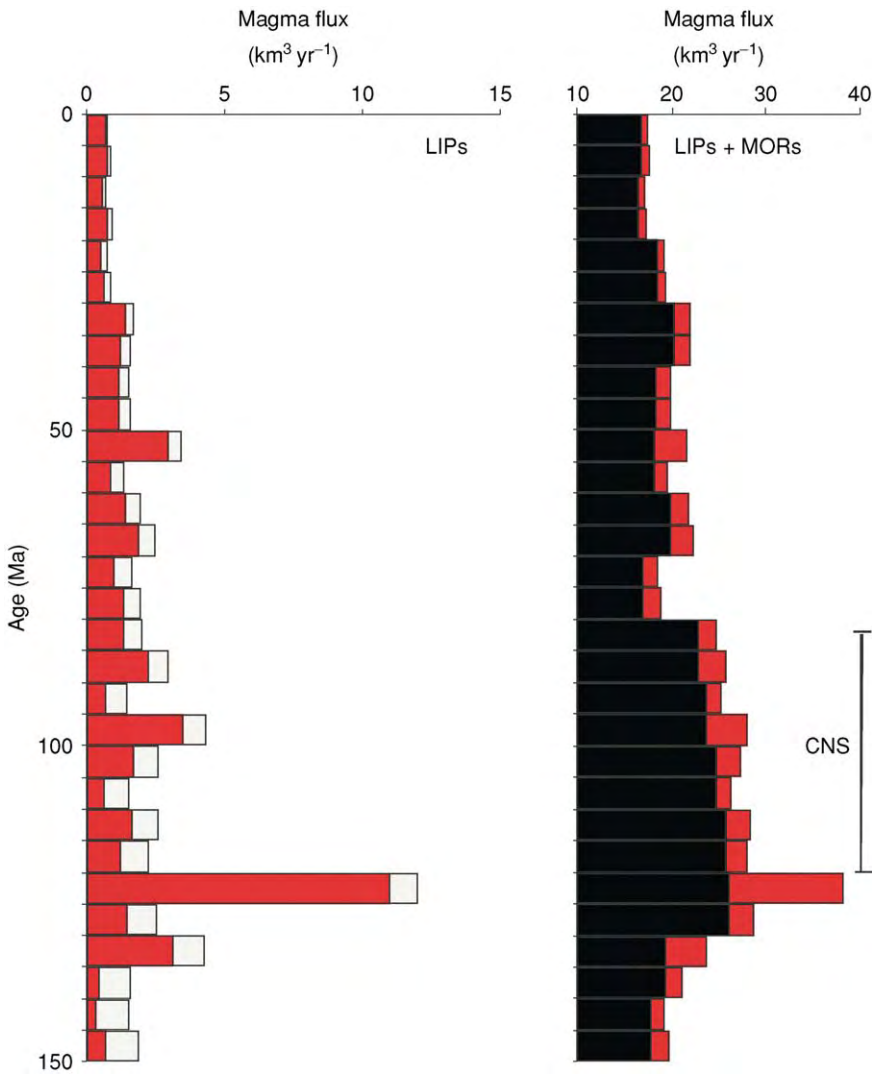


Figure 4 LIP, corrected for subduction (left; preserved in red and subduction correction in white), and summed LIP (red) and mid ocean ridge (MOR; black) (right) magma production since 150 Ma. Note the difference in the *x* axis scales. CNS, Cretaceous Normal Superchron.

(*ca*. 120–80 Ma), a time during which the Earth's magnetic field was of normal polarity. Thus, although LIPs manifest types of mantle processes distinct from those resulting in seafloor spreading, the waxing and waning rates of overall mantle circulation probably affect both sets of processes. A major question that emerges from studies of the global LIP production rate is whether the mantle is circulating less vigorously as the Earth ages.

LIPs and Mantle Dynamics

The formation of various sizes of LIP in a variety of tectonic settings on both continental and oceanic lithosphere suggests that a variety of thermal anomalies in the mantle give rise to LIPs and that the lithosphere strongly controls their formation. Equivalent thermal anomalies beneath continental and oceanic lithosphere should produce more magmatism in the latter setting, as oceanic lithosphere is thinner, allowing more decompression melting. Similarly, equivalent thermal anomalies beneath an intraplate region (e.g. Hawaii) and a divergent plate boundary (e.g. Iceland) (Figure 1 and Table 1) will produce more magmatism in the latter setting, again because decompression melting is enhanced. Recent seismic tomographic images of mantle-velocity (a proxy for temperature) structure beneath Iceland and Hawaii show significant differences between the two.

Only recently, seismic tomography has revealed that slabs of subducting lithosphere can penetrate the entire mantle to the D″ layer just above the

boundary between the mantle and the core, at a depth of approximately 2900 km (Figure 3). If we assume that the volume of the Earth's mantle has remained roughly constant throughout geological time, then the mass of crustal material fluxing into the mantle must be balanced by an equivalent mass of material fluxing from the mantle to the crust. Most, if not all, of the magmatism associated with the plate-tectonic processes of seafloor spreading and subduction is believed, on the basis of geochemistry and seismic tomography, to be derived from the upper mantle (above *ca*. 660 km depth). It is reasonable to assume that the lithospheric material that enters the lower mantle is eventually recycled, in some part contributing to the emplacement of LIPs at the Earth's surface.

Although LIPs are commonly believed to have originated from mantle plumes generated solely by solid Earth processes, alternative mechanisms have also been proposed. The spatial, if not temporal, association of flood basalts and impact craters on the Moon, as well as limited evidence on Earth, suggests that massive decompression melting of the mantle or at least significant crustal thinning and fracturing forming conduits for mantle material to reach the surface of a terrestrial planet could account for the emplacement of some LIPs. Such a mechanism has been proposed as an alternative to the plume hypothesis for the Siberian Traps, the Ontong Java Plateau, and the Deccan Traps. Other LIPs may originate as a result of a combination of plate divergence or fracturing and co-located underlying thermally anomalous mantle. Thus, multiple mechanisms may be required to explain all LIPs, both on Earth and elsewhere in our solar system.

LIPs and the Environment

The formation of LIPs has had documented environmental effects both locally and regionally. The global effects are less well understood, but the formation of some LIPs may have affected the global environment, particularly when conditions were at or near a threshold state. Investigations of volcanic passive margins and oceanic plateaus have demonstrated widespread and voluminous subaerial basaltic eruptions. The eruption of enormous volumes of basaltic magma during LIP formation releases volatiles such as carbon dioxide, sulphur, chlorine, and fluorine (Figure 5). A key factor affecting the magnitude of volatile release is whether the eruptions are subaerial or submarine; hydrostatic pressure inhibits vesiculation and degassing of relatively soluble volatile components (water, sulphur, chlorine, and fluorine)

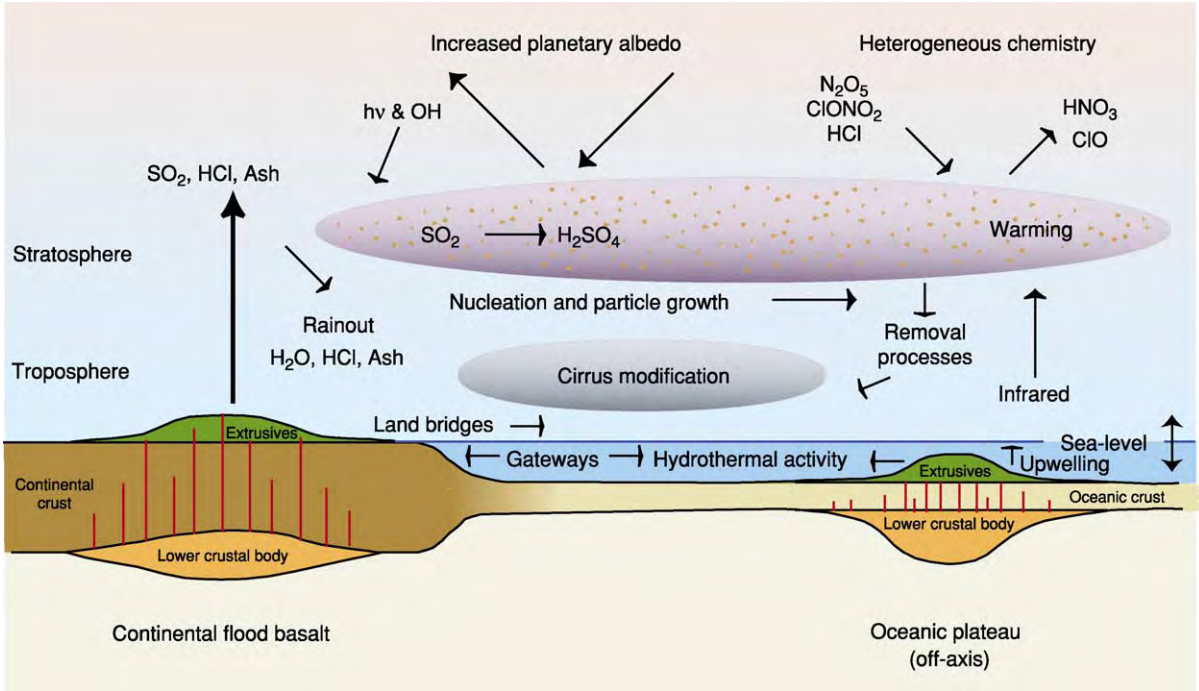


Figure 5 Environmental effects of LIP formation. LIP eruptions can perturb the Earth ocean atmosphere system significantly. Note that many oceanic plateaus form, at least in part, subaerially. Energy from solar radiation is hv, where h Planck's constant, and v frequency of electromagnetic wave of solar radiation.

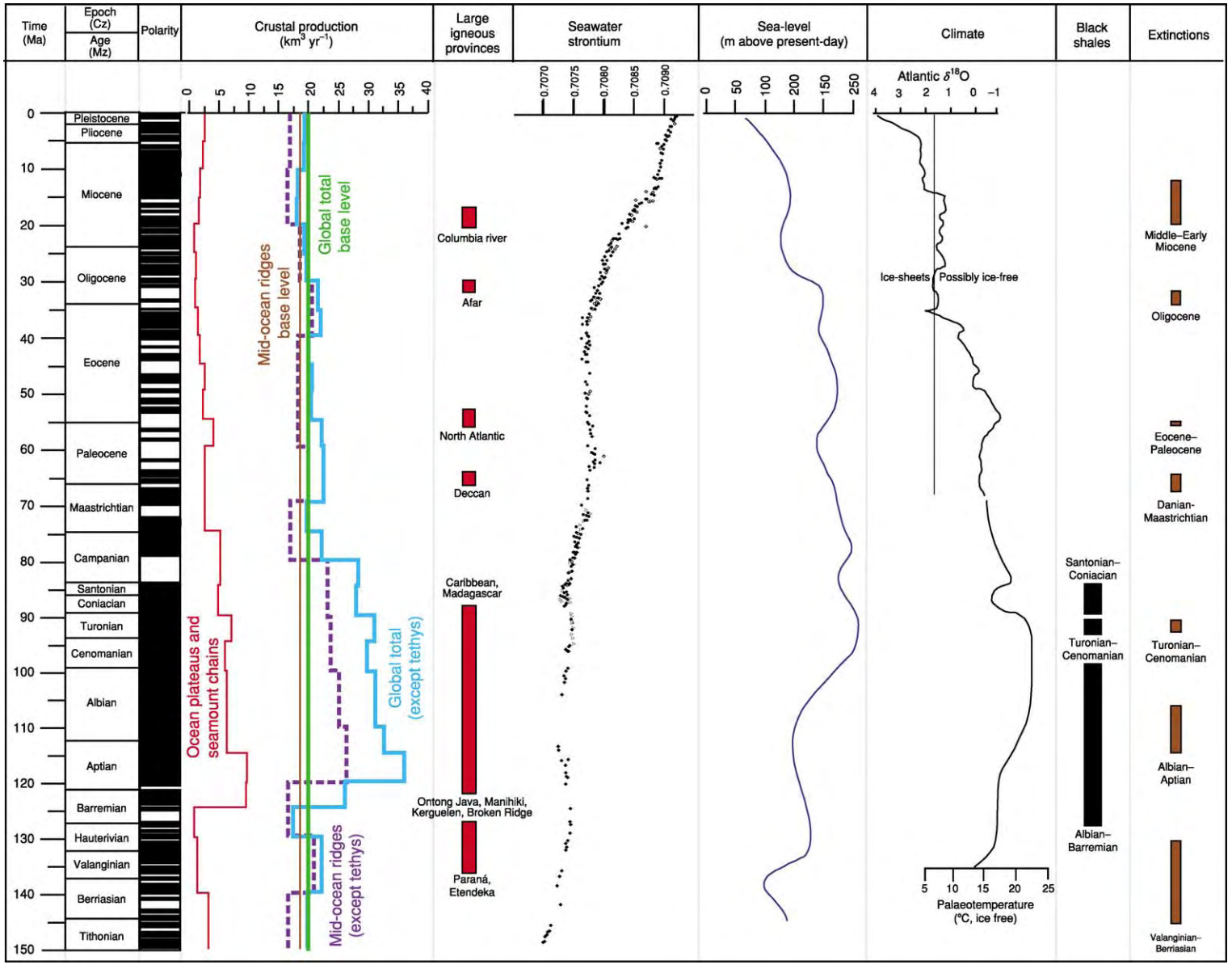


Figure 6 Temporal correlations among geomagnetic polarity, crustal production rates, LIPs, seawater strontium, sea level, climate, black shales, and extinctions.

during deep-water submarine eruptions, although low-solubility components (carbon dioxide and noble gases) are mostly degassed even at abyssal depths.

Another important factor affecting the environmental impact of LIP volcanism is the latitude at which the LIP forms. In most basaltic eruptions, released volatiles remain in the troposphere. However, at high latitudes, the tropopause is relatively low, allowing large-mass-flux basaltic fissure eruption plumes to transport SO_2 and other volatiles into the stratosphere. Sulphuric acid aerosol particles that form in the stratosphere after such eruptions have a longer residence time and greater global dispersal than if the SO_2 remains in the troposphere; therefore, the effects on climate and atmospheric chemistry are greater. The large volume of volatiles released, over relatively brief geological intervals, by the subaerial flood basalts of high-latitude LIPs would contribute to potential global environmental effects.

Highly explosive felsic eruptions, such as those documented from volcanic passive margins, an oceanic plateau (Kerguelen; Figure 1 and Table 1), and continental flood basalt provinces, can also inject both particulate material and volatiles (SO_2 and CO_2) directly into the stratosphere. The total volume of felsic volcanic rocks in LIPs is poorly constrained, but they may account for a small, but not negligible, fraction of the volcanic deposits in LIPs. Significant volumes of explosive felsic volcanism would further contribute to the effects of predominantly mafic LIP volcanism on the global environment.

Between about 145 Ma and 50 Ma, the global oceans were characterized by variations in chemistry, relatively high temperatures, high relative sea-level, episodic deposition of black shales, high production of hydrocarbons, mass extinctions of marine organisms, and radiations of marine flora and fauna (Figure 6). Temporal correlations between the intense pulses of igneous activity associated with LIP formation and environmental changes suggest a causal relationship. Perhaps the most dramatic example is the eruption of the Siberian Traps (Figure 1 and Table 1) at approximately 250 Ma, coinciding with the largest extinction of plants and animals in the geological record. It is estimated that 90% of all species became extinct at that time (*see* **Palaeozoic:** End Permian Extinctions). On Iceland, the 1783–1784 eruption of Laki provides the only human experience of the type of volcanism that constructs LIPs. Although Laki produced a basaltic lava flow representing approximately 1% of the volume of a typical (10^3 km^3) LIP flow, the eruption's environmental impact resulted in the deaths of 75% of Iceland's livestock and 25% of its population from starvation.

Conclusions

Oceanic plateaus, volcanic passive margins, submarine ridges, seamount chains, ocean-basin flood basalts, and continental flood basalts share geological and geophysical characteristics that indicate an origin distinct from that of igneous rocks formed at mid-ocean ridges and arcs. These characteristics include

- a broad areal extent (in excess of 10^4 km^2) of iron- and magnesium-rich lavas;
- massive transient basaltic volcanism occurring over 10^5–10^6 years;
- persistent basaltic volcanism from the same source lasting 10^7–10^8 years;
- lower crustal bodies characterized by P-wave velocities of 7.0–7.6 km s^{-1};
- a component of more silica-rich volcanic rocks;
- higher MgO lavas, basalts with more diverse major-element compositions, rocks with more common fractionated components, both alkalic and tholeiitic differentiates, and basalts with predominantly flat light-rare-earth-element patterns, relative to mid-ocean-ridge basalts;
- thick (tens to hundreds of metres) individual basalt flows;
- long (up to 750 km) single basalt flows; and
- lavas erupted in both subaerial and submarine settings.

There is strong evidence that many LIPs both manifest a fundamental mode of mantle circulation, commonly distinct from that which characterizes plate tectonics, and contribute episodically, at times catastrophically, to global environmental change. Nevertheless, it is important to bear in mind that we have literally only scratched the surface of oceanic and continental LIPs, and that LIPs on other terrestrial planets await investigation.

See Also

Earth: Mantle. **Igneous Processes**. **Lava**. **Mantle Plumes and Hot Spots**. **Palaeozoic:** End Permian Extinctions. **Plate Tectonics**. **Seamounts**. **Tectonics:** Mid-Ocean Ridges.

Further Reading

Campbell IH and Griffiths RW (1990) Implications of mantle plume structure for the evolution of flood basalts. *Earth and Planetary Science Letters* 99: 79 93.

Coffin MF and Eldholm O (1993) Large igneous provinces. *Scientific American* 269: 42 49.

Coffin MF and Eldholm O (1994) Large igneous provinces: crustal structure, dimensions, and external consequences. *Reviews of Geophysics* 32: 1 36.

Condie KC (2001) *Mantle Plumes and Their Record in Earth History*. Cambridge: Cambridge University Press.
Davies GF (2000) *Dynamic Earth: Plates, Plumes and Mantle Convection*. New York: Cambridge University Press.
Duncan RA and Richards MA (1991) Hotspots, mantle plumes, flood basalts, and true polar wander. *Reviews of Geophysics* 29: 31 50.
Eldholm O and Coffin MF (2001) Large igneous provinces and plate tectonics. In: Richards MA, Gordon RG, and van der Hilst RD (eds.) *The History and Dynamics of Global Plate Motions*. Geophysical Monograph 121, pp. 309 326. Washington: American Geophysical Union.
Ernst RE and Buchan KL (2001) *Mantle Plumes: Their Identification through Time*. Special Paper 352. Boulder: Geological Society of America.
Ernst RE and Buchan KL (2003) Recognizing mantle plumes in the geological record. *Annual Review of Earth and Planetary Sciences* 31: 469 523.
Macdougall JD (ed.) (1989) *Continental Flood Basalts*. Dordrecht: Kluwer Academic Publishers.
Mahoney JJ and Coffin MF (eds.) (1997) *Large Igneous Provinces: Continental, Oceanic, and Planetary Flood Volcanism*. Geophysical Monograph 100. Washington: American Geophysical Union.
Morgan WJ (1981) Hotspot tracks and the opening of the Atlantic and Indian oceans. In: Emiliani C (ed.) *The Oceanic Lithosphere, The Sea, Volume 7*, pp. 443 487. New York: John Wiley & Sons.
Richards MA, Duncan RA, and Courtillot VE (1989) Flood basalts and hot spot tracks: plume heads and tails. *Science* 246: 103 107.
Sleep NH (1992) Hotspot volcanism and mantle plumes. *Annual Review of Earth and Planetary Sciences* 20: 19 43.
White RS and McKenzie D (1995) Mantle plumes and flood basalts. *Journal of Geophysical Research* 100: 17 543 17 585.

LAVA

N Geshi, Geological Survey of Japan, Ibaraki, Japan

Introduction

Red-hot lava is characteristic of an active volcano, and the spectacle of flowing lava gives us an impression that the Earth is really living. Lava flows are the most common volcanic feature on Earth, and large volcanic structures, such as shield volcanoes and stratovolcanoes, are formed mainly from a pile of lava flows. Furthermore, the upper part of the ocean floor consists of submarine lava flows produced at the mid-ocean ridges, which form the largest volcanic zone on Earth. The structural characters of a lava flow, such as its size, composition, and viscosity, tell us much about its nature and emplacement mechanism. This article describes the primary features of lava flows and cites some examples of disasters caused by them.

Eruption of Lava

Lava is the term for molten magma erupted onto the Earth's surface as a continuous melt and also for the rock that solidifies from it. A flow of hot molten rock over the ground surface is called a lava flow, as is the solidified rock that forms from it.

Lava filling a broad depression or a vent is often called a 'lava lake'. A jet of fluid lava sprayed 10–100 m into the air is called a 'lava fountain'. Lava fountains are typical of extensive basaltic eruptions. Lava fountains may occasionally reach more than 500 m into the air (Figure 1).

Lava is generally a product of non-explosive moderate eruption. By contrast, fragments of molten lava produced in an explosive eruption, such as volcanic ash, pumice, scoria, and volcanic bombs, are called 'pyroclastics' or 'tephra' (*see* **Pyroclastics**). Some basaltic and andesitic lava flows are generated during explosive eruptions with high effusion rates by refusion and reactivation of an accumulation of still-molten pyroclastics in the vicinity of the vent. This type of lava is often called 'clastogenic lava' or 'rootless lava'.

The explosivity of a magma is controlled by the behaviour of bubbles that are present in the magma while it is ascending. When magma containing bubbles ascends rapidly, decompression causes expansion and rupturing of the bubbles and an explosion will occur. In contrast, effective separation of the bubbles from the magma will result in non-explosive eruption of lava flows. Bubble separation is controlled by many factors, such as viscosity, decompression speed, and gas content, and its mechanism is one of the current topics in volcanology.

Temperature and Viscosity of Lava

The highest measured temperature of the lava tends to correspond to the liquidus of a rock of the same composition. The liquidus is the temperature at which magma begins to crystallize, and the liquidus of dry magma at a pressure of one atmosphere ranges

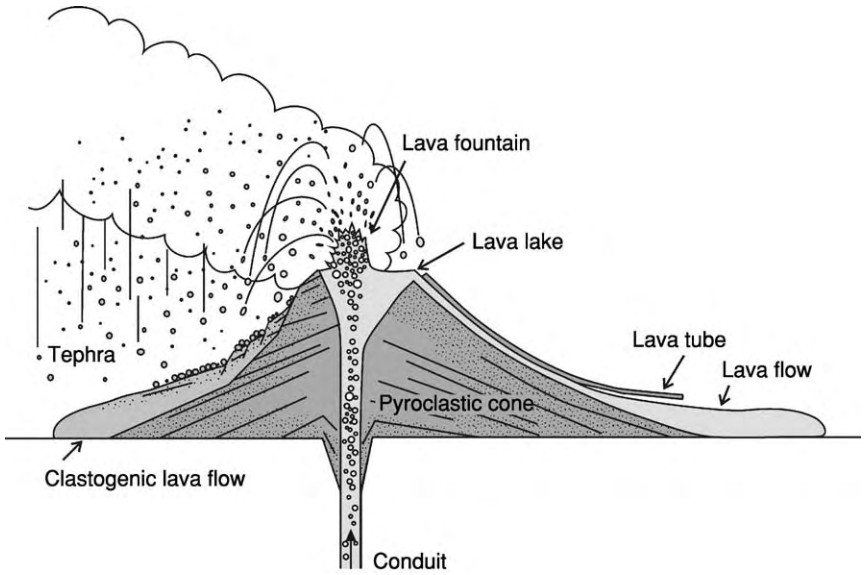


Figure 1 The eruption of quite fluid magma. Bubbles rise from conduit ruptures in the vent and produce pyroclastic particles. Degassed magma flows down the side of the cone and forms a lava flow. Lava travels through a lava tube crusted with solidified lava. Rapid sedimentation of still molten pyroclastic materials causes welding and remobilization of the pyroclastics, forming a clastogenic lava flow.

from 1150°C to 1250°C for basalt and from 900°C to 1000°C for dacite and rhyolite. However, the temperature of flowing lava is usually less than the liquidus temperature because it is cooled quickly by radiation from its surface and conduction to the atmosphere. Flowing basaltic lavas from the Hawaiian volcanoes, Etna, and Izu-Oshima have temperatures of between 1000°C and 1100°C, and andesitic lavas from the 1946 eruption of Sakurajima had temperatures of 850–1000°C.

Viscosity is one of the important parameters controlling the movement and emplacement of lava flows. The viscosity of the melt increases as its silica content increases because silica particles form chain structures in silicate-rich melts. Elements such as magnesium can sever these chain structures, so the viscosities of magmas rich in these elements are low. Water also disrupts the silica chains, and the viscosities of magmas containing water are lowered dramatically. Viscosity is also strongly controlled by the temperature of the lava. Basaltic lava is usually of low viscosity (less than 10^4 Pa), and dacite and rhyolite lavas display very high viscosities (more than 10^{10} Pa).

Natural magma usually contains crystals that have crystallized from the cooling melt. The abundance of these crystals also controls the viscosity of the magma because of the way that the crystal grains interact. Collision of the crystals may check the flow of melt and increase the bulk viscosity of the magma, so the magma becomes more viscous with increasing crystal content. When the quantity of crystals exceeds 30%, this effect becomes noticeable, and magmas that are more than 60–70% crystals are unable to flow. This effect also depends strongly on the shape and size distribution of the crystals. Interaction among bubbles in the magma has a similar effect on bulk viscosity.

Volume, Effusion Rate, and Speed

The volume of a lava flow depends on the scale of the eruption, viscosity, and various other factors. The largest lava flow in the historic record was issued during the 1783 eruption in Iceland and is called the Laki flow. The lava was erupted from a 25 km long fissure over a period of 6 months and the total volume reached 12 km^3. The longest branch of the lava flow extended for more than 60 km and covered an area of more than 500 km^2. Flows of lava more than 10 km long generally have a basaltic composition (although the Yatta Plateau in Kenya, which is an intermediate flow of phonolite, extended for 300 km), and the maximum volume of lava flows decreases as their silica content increases. The effusion rate of lava is usually less than several cubic metres per second for basaltic lava flows and decreases as silica content increases because its high viscosity prevents flow through the conduit. This tendency may be controlled by many factors such as eruption style, the size of the magma batch, and the viscosity of the magma. However, this does not mean that a huge silicic eruption

Figure 2 (A) An advancing basaltic pahoehoe lava; Kilauea, Hawaii (photograph supplied by K Niihori). (B) The surface of an a'a' lava of basaltic andesite composition that was issued during the 1983 eruption of Miyakejima, Japan. Clinkers several centi metres across cover the surface of the flow. (C) The flow front of an andesitic block lava issued in the 1783 eruption of Asama, Japan. The surface of the lava flow is covered with angular blocks several metres in diameter. (D) Dacite lava dome of Unzen, Japan, formed during the 1991 1995 eruption.

cannot occur: silicic magmas sometimes cause Plinian eruptions – the most violent eruption style – and issue ignimbrites covering several square kilometres within a few hours or days.

Many scientists have tried to model the eruption mechanism that forms lava flows. Recent studies using numerical simulation have shown that eruption rate loosely controls the width of a single lobe of lava. In contrast, these studies have also shown that the length of a lava flow is affected by many complex parameters, such as slope angle, viscosity, and effusion time. In fact, a real lava flow is more complex. A large lava field usually consists of a complex of many small lava flows fed for a long period by tube systems, and the final shape is controlled by many other factors, such as the topography of the surface.

The maximum speed of a lava flow is usually less than several tens of metres per second. Basaltic lava, with its low viscosity, sometimes flows very fast. Hawaiian basaltic lava often flows at more than 15 m s ^{1} through lava tubes. During the 1950 eruption on Mauna Loa, Hawaii, lava flows travelled from the vent to the ocean, 10 km away, in as little as 3 h. In contrast, silicic lava flows very slowly. During the 1991–1995 eruption on Unzen, Japan, the maximum flow speed of the dacitic lava was less than 50 m day ^{1}.

Structure of Lava Flow

Subaerial Lava

A subaerial flow of lava is largely controlled by its viscosity. Lavas with a mafic composition – basalt and andesite – are normally fluid enough to flow away from the vent under gravity and form relatively thin flow lobes. In contrast, silicic magma cannot flow far because of its high viscosity and forms short thick flows (Figure 2).

Basaltic lava with a smooth or ropy surface is called 'pahoehoe lava'. Pahoehoe is a native Hawaiian term for lava with a smooth surface. During the advance of pahoehoe lava, small lobes and toes continuously break out from a cooled crust. When a red-hot molten lobe breaks out of the front of a previous lobe, a thin glassy crust is quickly formed on its surface by rapid cooling. Injection of molten lava into the lobe stretches the crust at its tip. As cooling progresses, the crust becomes rigid, and, rather than stretching further, it breaks to allow issue of the next new lobe. Sometimes flexible cooled crust is wrinkled by the motion of the molten lava below, forming a remarkable ropey structure on its surface. If the effusion rate

is high enough, the lava will form a large flat sheet-like lobe with a smooth surface. Pahoehoe lava is typical of basaltic lava with low viscosity and is rarely found in andesitic or rhyolitic lava. The thickness of a pahoehoe flow is usually less than 1 m, although lobes may often inflate to more than several metres thick as a result of the additional injection of molten lava.

Another typical type of basaltic lava is called 'a'a' lava'. The term a'a' also originates from the native Hawaiian term for lava with a rough surface. The surface of a'a' lava is completely covered by well-vesiculated rugged blocks called 'clinker'. Flow lobes of a'a' lava are, in many cases, several metres thick, much thicker than pahoehoe lava. During the advance of a'a' lava, flow cooling and vesiculation form a brittle crust at the top of the flow. Motion of the inner molten part breaks the crust and forms clinker. Since the velocity of the flow is at its maximum at the surface, clinkers on the top of the flow move forwards and tumble down the steep front of the flow. The advancing a'a' flow buries the fallen clinker, which forms a basal clinker layer. Some pahoehoe lavas change to a'a' lavas during their travel, but the reverse case is rare. During the transition from pahoehoe to a'a', the less-flexible cooled crust at the surface of the pahoehoe flow is fragmented by the motion of the inner molten part and forms a'a' clinker. This process indicates that the yield strength of the cooling crust controls the transition from pahoehoe to a'a' (Figure 3).

Viscous andesitic lava forms 'block lava'. The surface of block lava is covered by angular dense fragments with smooth faces. The angular blocks were formed by the break-up of the partly or wholly congealed upper part of the flow as the still-mobile magma moved beneath the crust. Block lava is usually thicker than a'a' lava: sometimes more than 100 m of fragmented material may constitute the entire thickness, making up a greater proportion of the flow than in a'a' lava.

Highly viscous lava, such as rhyolite and trachyte, cannot flow any distance and will form a thick short lobe of block lava around the extrusive point. Such a rise of lava is called a 'lava dome'. Usually the surface of a lava dome is covered with a block of lava produced by the internal motion of the viscous core and so resembles blocky lava.

Underwater Eruption of Lava

When magma is extruded at the bottom of a sea, lake, or glacier, rapid cooling by the surrounding water

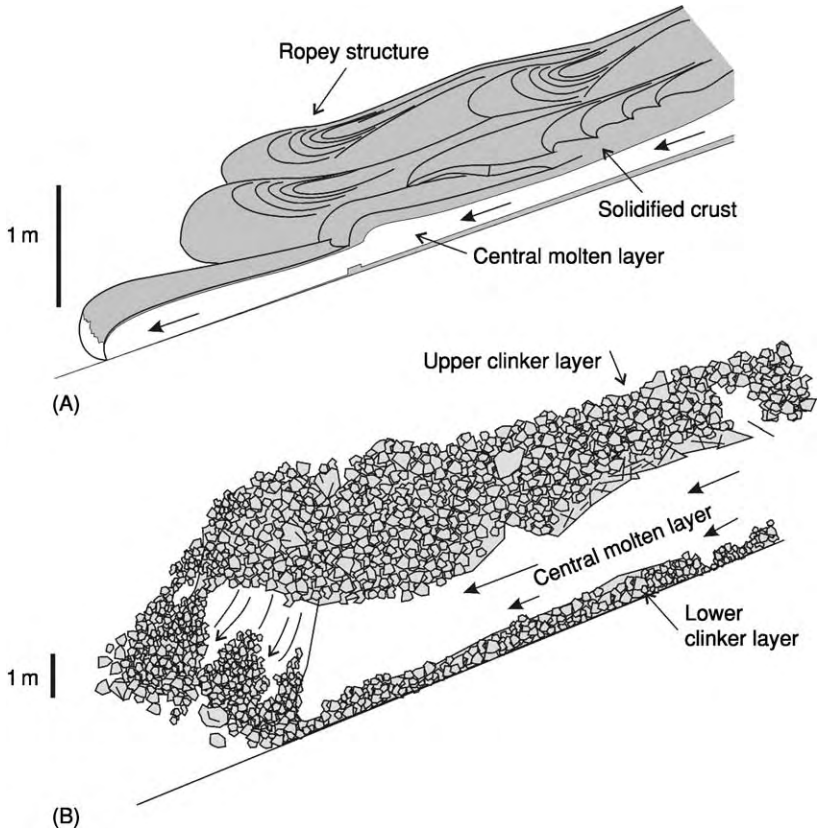


Figure 3 The structure of (A) pahoehoe lava and (B) a'a' lava. Note that the scales of the two diagrams are different.

forms a unique lava structure. Because of the large specific heat, thermal diffusivity, and evaporation heat of water, lava will be cooled more effectively in water than in air. When lava with this unique structure is exposed on land, a geologist can tell that the rock was formed underwater because of its structure.

One of the typical structures of lava so erupted is 'pillow lava', which forms mounds of elongated 'sacks' with quenched glassy rims. Repeated oozing and quenching of hot basaltic magma produces the pillow structure. A newly extruded lobe of lava is quickly surrounded by a flexible glassy skin due to rapid cooling by the surrounding water. Continuous injection of lava into the lobe expands it and forms a mass of basalt with a pillow shape. Finally, the skin breaks and new basalt extrudes into another lobe. Repetition of this sequence forms a thick deposit of pillow-like lobes of basalt (Figure 4). Pillow lava is a characteristic structure of fluid basaltic lava, corresponding to pahoehoe lava on land. When the effusion rate is high, basaltic lava spreads out on the seafloor to form 'sheet lava'. Rare pillows are known from more siliceous lavas.

More viscous lava such as andesite forms a volcanic breccia (hyaloclastite) by brittle fracturing due to quenching on contact with water. The surface of the flow lobe is chilled and forms a brittle crust while the viscous lava is still travelling. Motion of the molten interior fractures the crust and, because the surrounding water invades the inside of the flow, brittle fracturing advances deeply into the flow. As a result, a highly brecciated lobe consisting of angular glassy and massive blocks is formed.

Figure 4 A pile of pillow lavas produced during the Miocene by the Ogi, Japan, submarine volcano. Each pillow lobe has a glassy skin and a massive interior with radial cooling joints. (Photograph supplied by T Oikawa.)

Lava Tubes

Lava tubes are natural tunnels through which lava travels beneath the surface of a lava flow. Flowing lava is cooled from its surface by radiation and thermal convection of air, and a rigid crust is formed. Once the rigid crust is formed, it provides insulation because of its small thermal conductivity, and the inner molten lava can flow without cooling. In a broad lava-flow field, lava tube systems with a main tube from the vent and a series of smaller branches develop, which can supply lava to the front of the flow without it cooling. When the supply of lava ceases at the end of an eruption or the lava is diverted elsewhere, lava in the tube flows out and leaves a partially empty tunnel beneath the ground. Lava can also erode downwards, deepening the tube and leaving empty space above the flowing lava, because the walls and floor of the tube consist of lava of the same composition as the hot lava flowing in the tube, so the flowing lava can melt the wall rock. Lava tubes often develop in basaltic lava flows with low viscosity and are rare in highly viscous felsic lava.

Cooling Joints

One of the remarkable internal structures of lava is the systematic cooling joint. As the temperature of lava drops, its volume decreases and strain within the lava causes it to fracture (Figure 5). A typical cooling-joint system is 'columnar jointing', which forms prismatic columns of rock with polygonal cross sections. Columnar jointing is formed as follows: the spread out lava cools from its upper and lower surfaces, and the volume of lava decreases as its temperature falls. Since the surface area of the lava is fixed, tension from the contraction of the main part of the lava will form a polygonal – in many cases pentagonal or hexagonal – fracture system. Shrinkage fractures parallel to the surface rarely develop because tension in the vertical dimension is accommodated by a decrease in the thickness of the flow. A similar phenomenon, called 'sun cracks', is often observed in mud when a puddle dries up. In this case, the mud shrinks as water in the mud evaporates. As in the progress of cooling inside the lava, the fractures propagate inwards and polygonal pillars surrounded by platy fractures are formed. The axis of the pillar is normal to the isothermal plane. Columnar jointing often develops not only in lava but also in tabular intrusions such as dykes and sills.

Another type of cooling joint is the 'platy jointing'. This joint system consists of subparallel fractures forming thin plates. Tabular joints are typically observed at the bottom of a lava flow and at the wall of a dyke, where the shear strain acts during

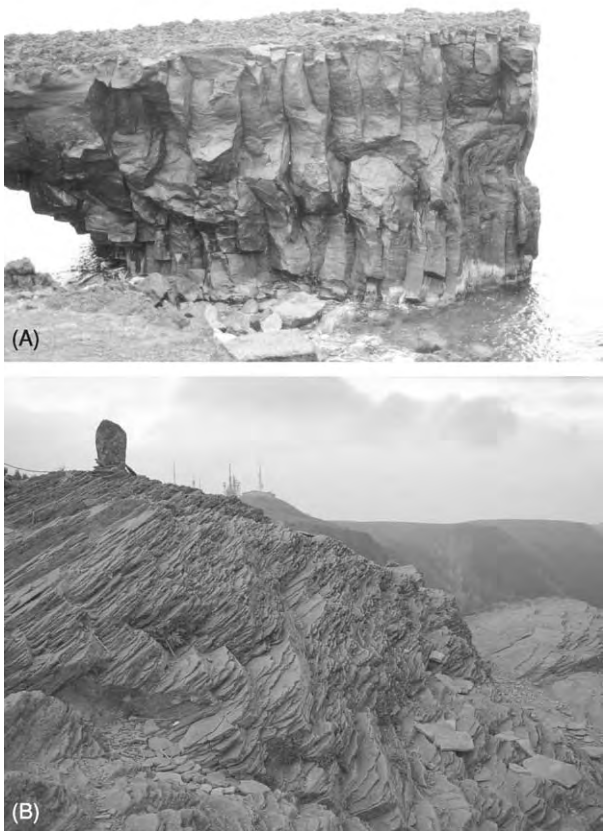


Figure 5 (A) Columnar jointing in thick basaltic lava from Miyakejima, Japan. (B) Tabular jointing in the middle Pleistocene andesitic lava of Utsukushigahara, Japan (Photographs supplied by T Oikawa).

flow. The brittle cooling margin at the base of the lava flow is dragged by the still-mobile magma inside the flow, and fractures parallel to the flow are formed. Shear deformation occurs at the bottom of the flow. Platy jointing is commonly observed in silicic lava flows such as those composed of andesite or dacite.

Flood Basalt

At certain times in the geological record, prolonged volcanic activity of immense scale resulted in thick piles of numerous basaltic lava flows that covered thousands of square kilometres (similar flows of the intermediate, more siliceous, lava type phonolite are also known in East Africa). The large-volume basaltic lava flows are called 'flood basalts' or 'plateau basalts'. The total volume of a flood-basalt eruption may sometimes exceed 100 000 km^3, and this huge volume is erupted over a short time interval, usually less than 1 Ma. Flood-basalt sequences consist of a pile of thousands of flows, and the individual flows may be more than several tens of metres thick. The individual flows may extend for hundreds of kilometres. How can such huge lava flows be emplaced? The old idea for the emplacement of flood basalts was that these flows were emitted at incredible velocities. However, based on the detailed survey of the structure of flood basalts, geologists arrived at an alternative idea that these flows are emplaced by slow movement, with most of the great thickness being achieved by injecting lava into the interior of an initially thin flow (*see* **Large Igneous Provinces**).

The most famous example of a flood basalt is the Deccan Traps of central India, where about 1 000 000 km^3 of basaltic lavas were erupted within a half million years at the end of the Cretaceous. The Siberian Traps, which erupted in Permian–Triassic times, are much larger than the Deccan Traps but less well understood. Many scientists consider that the volcanism that produces flood basalts is related to the rise of hot mantle plumes. The origin of the flood basalts is closely associated with entire-mantle dynamics, and their magma genesis is one of the currently popular topics in Earth science.

Lava Flow Hazards

Lava flows destroy or ignite all things in their path. When lava invades developed areas, it can be disastrous. In the cases of many volcanoes, the people living nearby make great efforts to stop the advance of lava flows, but much land and many buildings are destroyed and buried by lava flows. Fortunately, the speed of advance of a lava flow is, usually, very slow, and in many cases people can escape from it.

Kilauea (Hawaii, USA)

Kilauea is one of the most active volcanoes in the world. More than 90% of the land surface of the Kilauea volcano has been re-covered by new lava within the last 1500 years. In particular, the southeastern flank of the east rift zone of Kilauea has been threatened frequently by lava flows, and more than 30% of the land surface between the east rift zone and the coast has been covered by lavas since 1955, less than half a century ago (Figure 6).

The Puu Oo–Kupaianaha eruption of Kilauea from 1983 to the present (2004) is the longest duration of a single eruption in the historical record. Lava flows from the Puu Oo and Kupaianaha craters spread widely on the southern flank of the east rift zone and poured down to the ocean. Lava frequently changed direction, and one of the flows invaded Kalapana village in 1990. By the end of October that year, lava had destroyed all of Kalapana and

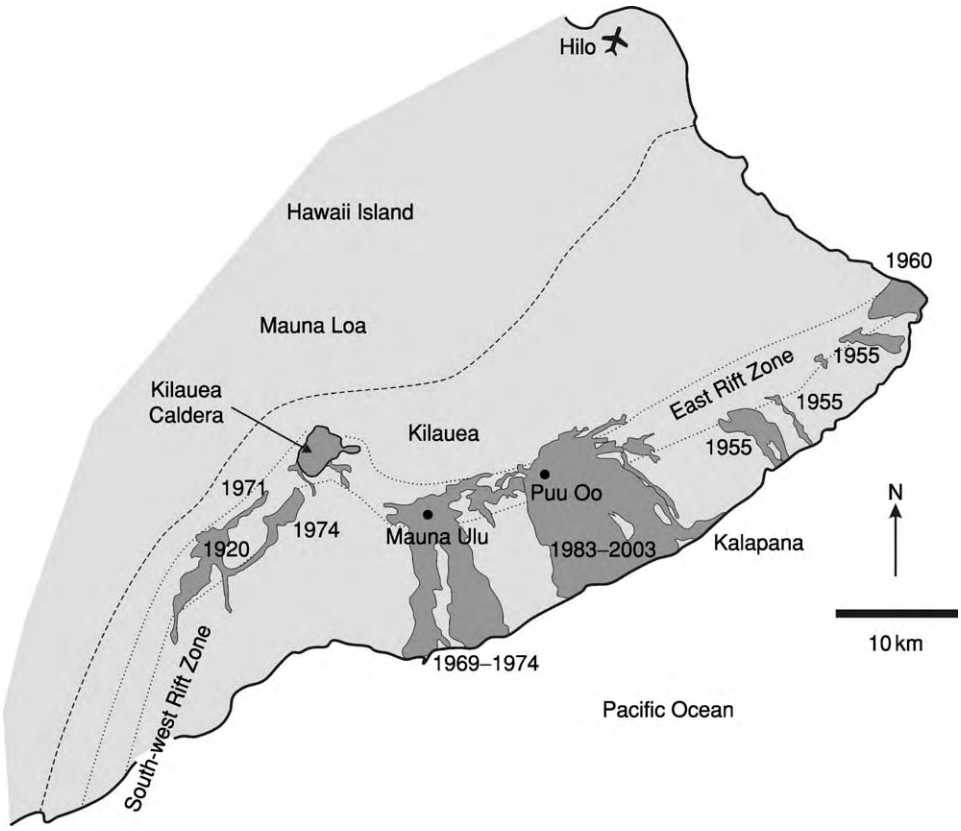


Figure 6 Distribution of the lava flows of Kilauea in the last 100 years. Presently, eruption is occurring at the Puu Oo vent in the east rift zone. From 1983 to 2003, about 2.3 km^3 of lava erupted and covered an area of more than 110 km^2.

filled up Kamimu bay. The lava spread slowly and people could evacuate safety, but more than 170 structures were destroyed. The lavas have added more than a square kilometre of new land to Hawaii Island, the largest of the Hawaiian island chain.

Nyiragongo (Congo)

Lava flowing from the Nyiragongo volcano in eastern Congo has also been disastrous. Nyiragongo is a stratovolcano that erupts predominantly nephelinitic lavas, with very low viscosity, and it is one of most active volcanoes in the East African rift zone. This volcano is famous for the active lava lake in the summit crater. A major fissure eruption occurred in 1977 on the south flank, which drained the lava lake in the summit caldera: very fluid lava rapidly covered several square kilometres and destroyed some villages. Because the lava was of very low viscosity and the eruption rate was very high, lava quickly ran down the steep slope of the volcano. The lava travelled about 10 km within 20 min and reached Goma, one of the major cities of Congo. The maximum speed of the lava was more than 50 $km\,h^{-1}$ in the vicinity of the fissure, which is extraordinarily fast for a lava flow. Over 70 people were killed by the lava flow. In January 2002, another major flank eruption occurred, and streams of lava spread over the centre of the city of Goma, dividing it in two. More than 40 people were killed by this eruption (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards).

Etna (Sicily)

People living on the flanks of Etna are forever fighting against lava. Mount Etna is the largest stratovolcano in Europe and has erupted more than 150 times in the last 2700 years. The flank areas of Etna have frequently been invaded by lava flows during its long history. In 1966, a major fissure eruption occurred on the south-eastern flank. Lava erupted from a 12 km long fissure towards the city of Catania. Some people dug into the levee at the side of the lava flow to try to change the flow direction, but their effort ended in failure. In March 1983, a fissure on the southern flank began to produce a'a' lava, which destroyed several buildings. The lava flow advanced more than 6 km, and three towns were threatened by the lava by the end of April. To change the direction of the lava flow, people tried to breach the side levee of lava and

lead the flow into a diversion channel. The levee was reduced to only a few metres thick, and 400 kg of dynamite were set in the hot wall of the levee. Because of numerous technical difficulties, the breach was smaller than originally planned, and only 20% of the flow was diverted out of the main channel. Next, people tried to construct a rubble barrier about 10 m high, 30 m wide, and 400 m long along the western margin of the flow. Even though lava overtopped the first barrier, more barriers were formed and finally they succeeded in preventing lateral spreading of the flow field into developed areas.

Heimaey (Iceland)

The 1973 eruption on the island of Heimaey is a famous example of fighting a lava flow by cooling it with water. A fissure eruption began in January from a 2 km long fissure across the island in the vicinity of the centre of the town of Vestmannaeyjar, one of Iceland's major fishing ports. Lava flow from the vent began to threaten the town and the port. As the flow advanced to the north and east, the mouth of the port began to be buried by lava. A second large lava flow moved north-west on the west side of the main flow and had covered many houses by the end of March. By early May, some 300 buildings had been engulfed by lava flows or gutted by fire. To prevent the advance of the lava and save the town and port, people sprayed seawater onto the moving lava during the eruption. More than 30 km of water pipes and 43 pumps were used to deliver seawater at up to $1\ m^3\ s^{-1}$. A total of 6 000 000 m^3 of water was poured onto the lava. The front of the lava flow was solidified by the cooling effect of the water and it stopped moving. The eruption ended in July that year and the port was saved. After the eruption, people made a great effort to remove the lava and tephra from the centre of the city.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards. **Igneous Processes**. **Large Igneous Provinces**. **Mantle Plumes and Hot Spots**. **Plate Tectonics**. **Pyroclastics**. **Tectonics:** Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Bardintzeff JM and McBirney AR (2000) *Volcanology,* 2nd edn. Sudbury, MA: Jones & Bartlett Publishers.

Cas RAF and Wright JV (1987) *Volcanic Successions: Modern and Ancient*. London: Chapman & Hall.

Decker R and Decker B (1989) *Volcanoes: Revised and Updated Edition*. New York: WH Freeman and Company.

Decker RW, Wright TL, and Stauffer PH (eds.) (1987) *Volcanism in Hawaii*. Professional Paper 1350. US Geological Survey.

Fink JH (ed.) (1990) *Lava Flows and Domes: Emplacement Mechanisms and Hazard Implications*. IAVCEI Proceedings in Volcanology 2. New York: Springer Verlag.

Green J and Short NM (1971) *Volcanic Landforms and Surface Features A Photographic Atlas and Glossary*. New York: Springer Verlag.

Hall A (1996) *Igneous Petrology,* 2nd edn. London: Longman Group Limited.

Macdonald GA, Abbot AT, and Peterson FL (1990) *Volcanoes in the Sea: The Geology of Hawaii,* 2nd edn. Honolulu: University of Hawaii Press.

Schminke HU (2003) *Volcanism*. New York: Springer Verlag.

Sigurdsson H, Houghton B, McNutt ST, Rymer H, and Stix J (eds.) (2000) *Encyclopedia of Volcanology*. San Diego: Academic Press.

Wright T, Takahashi TJ, and Griggs JD (1992) *Hawaii Volcano Watch: A Pictorial History, 1779 1991*. Honolulu: University of Hawaii Press and Hawaii National History Association.

MAGNETOSTRATIGRAPHY

S G Lucas, New Mexico Museum of Natural History, Albuquerque, NM, USA

Introduction

Most of the dense core of the Earth is iron. The outer portion of the core is liquid, and the motion of this liquid produces a magnetic field, so that the Earth behaves like a giant bar magnet. This dynamo, however, changes easily, and for unknown reasons the magnetic field periodically reverses itself – the north and south magnetic poles switch positions.

On average, the magnetic field reverses itself about every 500 000 years, though the pattern of reversals is erratic. Flip-flops of the Earth's magnetic field, when recorded in a stratigraphical succession of rocks, are the basis of magnetostratigraphy (a contraction of 'magnetic-polarity stratigraphy').

Magnetostratigraphy correlates rocks on the basis of similarities in their magnetic-reversal patterns and is generally used to correlate surface exposures of rocks, though it can also be applied to subsurface cores. As explained below, magnetostratigraphy is not an independent method of correlating rocks. Nevertheless, it is a powerful tool because magnetostratigraphical correlation is based on matching magnetic reversals, which are geologically simultaneous events worldwide.

The Geomagnetic Polarity Time-Scale

Magnetic reversals have occurred frequently but irregularly during Earth history. The process of reversal seems to take about 4000–5000 years. The current state of the magnetic field (in which a compass needle points towards the north magnetic pole) has persisted for the last 700 000 years and is referred to as an interval of normal polarity. Geologists refer to periods when the poles had switched positions (so that a compass needle would have pointed to the south magnetic pole) as intervals of reversed polarity.

The first attempts at magnetostratigraphy were made in the 1950s, especially by the Russian scientist A N Khramov. Since the 1960s, geologists have made a concentrated effort to decipher the history of the Earth's magnetic field, and this research is ongoing.

During much of Earth history, the magnetic field reversed frequently (Figure 1). This has been the case throughout most of the Mesozoic and Cenozoic, but during the Late Carboniferous and most of the Permian, an interval of about 70 Ma, the magnetic field was stable (reversed). The pre-Carboniferous nature of the magnetic field is still not as well understood as its later history.

Because of plate tectonics and the subduction of oceanic crust, the oldest seafloor preserved on Earth dates from the beginning of the Late Jurassic, about 160 Ma ago. Geologists have determined the magnetic polarities of rocks from the seafloor, which are lavas for which some numerical ages have been calculated. Bands of cooled lava on the seafloor adjacent to spreading ridges preserve magnetic stripes that are symmetrical about the ridge. This seafloor magnetization provides a template that geologists have used

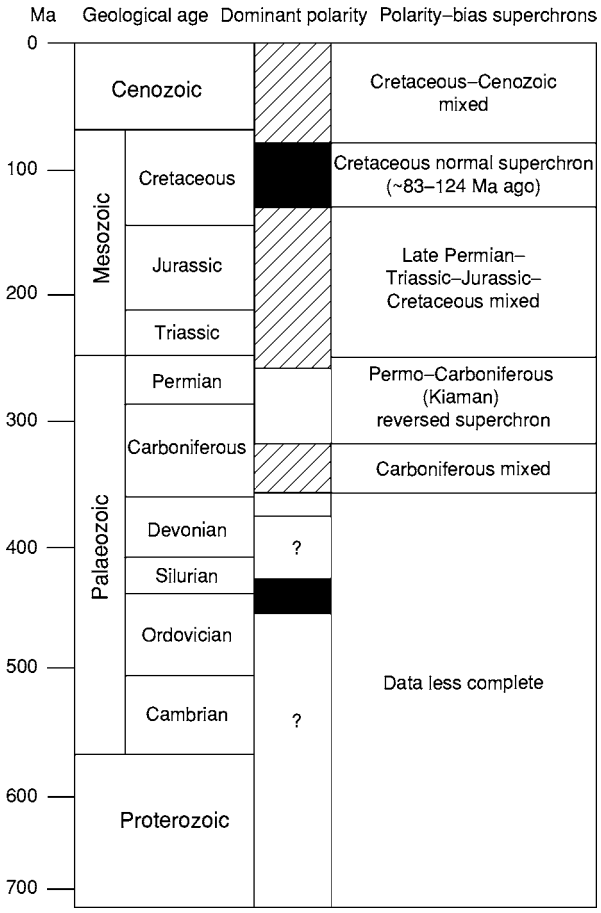


Figure 1 The polarity bias superchrons during the last 700 Ma. The polarity history is not well understood before about 350 Ma, and since then it has been mixed (many magnetic reversals) except for two long intervals of polarity stability: the Permo Carboniferous reversed and the Cretaceous normal superchrons.

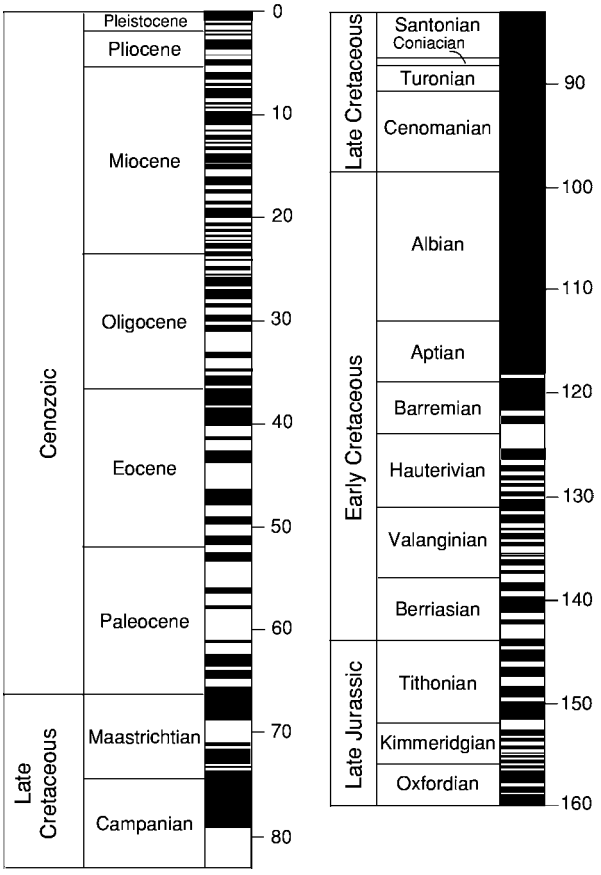


Figure 2 The global polarity time scale for the last 160 Ma is a well established and widely accepted record of magnetic polarity reversals since the beginning of the Late Jurassic.

to plot the magnetic-polarity history of the Earth back to the beginning of the Late Jurassic (Figure 2). This polarity history is referred to as the geomagnetic-polarity time-scale (GPTS). It provides a globally consistent pattern of normal- and reversed-polarity intervals that can be used to estimate the ages of rocks and the events that they record during the last 160 Ma of Earth history.

In the history of the Earth's magnetic field, a superchron is an interval of tens of millions of years during which the polarity remains constant. There are two well-established superchrons: the Cretaceous normal superchron, from about 118 Ma to 83 Ma ago, and the Permo-Carboniferous (also called Kiaman, after a place in Australia) reversed superchron, from about 316 Ma to 262 Ma ago (Figure 1).

Early studies of magnetostratigraphy named the magnetic-polarity intervals (then called 'epochs') after scientists and mathematicians (Brunhes, Matuyama, Gauss, Gilbert) and the polarity events after the places where they were first identified (Jaramillo, Mammoth, Olduvai). However, it was subsequently realized that there are so many intervals (now called 'chrons') and events (now called 'subchrons') that numbering them is simpler than assigning them names. Chron numbers are followed by an 'n' or an 'r' to indicate whether they are normal or reversed.

Remnant Magnetization

Most rocks contain minerals that are naturally magnetic, such as the iron oxide minerals haematite (Fe_2O_3) and magnetite (Fe_3O_4). In the crystals in which they are bound, minute grains of magnetic minerals act like tiny bar magnets, and these mineral grains can record the direction of the Earth's magnetic field. Thus, when lava flows cool, these magnetic minerals align themselves with the Earth's magnetic field. Magnetic mineral grains also align themselves with the magnetic field when they are deposited in sediments. These processes provide a record of the state of the magnetic field (normal or reversed polarity) when a rock is formed. This record is called the remnant magnetization.

Heat destroys the magnetization of a rock. Indeed, magnetic minerals lose their magnetization at a certain temperature (usually above 500°C), called the Curie point (after the chemist Pierre Curie). The natural remnant magnetization is locked into a rock when it cools below its Curie point, which is approximately 650°C for haematite and 580°C for pure magnetite.

There are three kinds of remnant magnetization. Thermal remnant magnetization is the result of a molten rock cooling below its Curie point, at which magnetic minerals align with the current magnetic field and become locked into the crystal with that alignment.

Detrital remnant magnetization occurs when an igneous rock erodes, and its magnetic minerals become loose sedimentary particles. These tiny magnetic grains (which are only a few micrometres in diameter) act as bar magnets and align with the magnetic field as they settle through the water column and are deposited as sedimentary particles (detritus).

Chemical remnant magnetization takes place when iron weathers out of a rock, moves through groundwater, and precipitates elsewhere. Usually it precipitates as some form of haematite. During the precipitation, the magnetic minerals, which were initially aligned when the original rock was formed, realign themselves with the magnetic field at the time of precipitation. Thus, the new alignment is a younger magnetization than the original magnetization that was acquired when the rock formed.

Field Sampling and Laboratory Analysis

To determine the magnetic polarity a rock, sampling begins in the field, either by drilling small cores out of well-indurated rocks or by cutting small blocks (cubes) out of less resistant rocks. The orientations of these samples with respect to the current magnetic field and their precise geographical locations are recorded. Orientated samples are taken to the laboratory to have their magnetization analysed.

The intensity of the magnetic field is measured using a unit called a gauss (named after the German

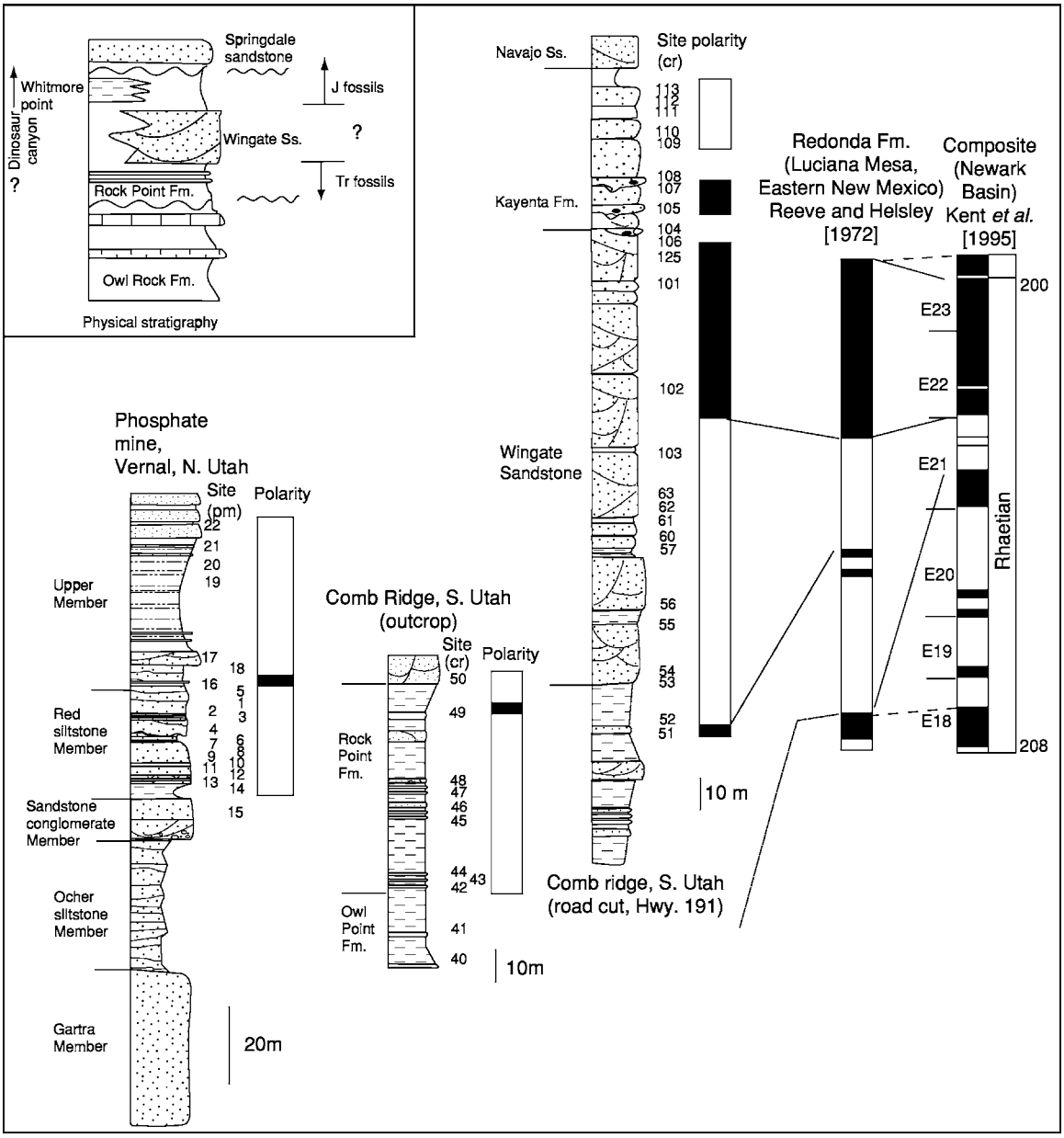


Figure 3 An example of how rock sections can be correlated with each other based on their magnetic polarity zonation. The three columns on the left show the magnetostratigraphy of three stratigraphical sections in Utah, with black indicating normal polarity magnetozones and white indicating reversed polarity magnetozones. The short interval of normal polarity between two long intervals of reversed polarity allows the three Utah columns to be correlated with each other. On the right, an attempt is made to correlate the Utah magnetostratigraphies with magnetozones in New Mexico and New Jersey (Newark). The match between the Utah and New Mexico magnetostratigraphies is fairly straightforward, but matching these with New Jersey data is more complex, because the Newark column has many more magnetic reversals. Molin Garza RS, Geissman JW and Lucas SG (2003) Paleomagnetism and magnetostratigraphy of the lower Glen Canyon and upper Chinle groups, Jurassic Triassic of northern Arizona and northeast Utah. *Journal of Geophysical Research*, v. 108, no. B4, 2181.

mathematician Karl Gauss). In the laboratory, a magnetometer is used to measure the intensity and direction of the magnetic vector of the rock sample. At this point, all the magnetizations acquired by the rock since it was formed are present. The goal is to identify the original magnetization at the time of rock formation (i.e. the natural remnant magnetization). Subsequent (younger) magnetizations are referred to as overprints. These younger magnetizations need to be removed; in a sense we can think of the process as cleaning up the magnetization history of the rock so that only the natural remnant magnetization remains (or at least can be identified). This is done in one of two ways, depending on which kind of magnetometer is used. Alternating-field demagnetization subjects the rock samples to a strong alternating magnetic field that destroys the weaker magnetizations in the rock, so that only the stronger natural remnant magnetization remains. Thermal demagnetization heats the rock; during this process weaker magnetizations tend to disappear first, at lower temperatures than the stronger natural remnant magnetization. Neither method is usually strong enough to realign the primary magnetic mineral grains in the rock, and thus remagnetize it.

Magnetostratigraphical Correlation

Once the magnetic polarity of a succession of rocks has been determined, a magnetostratigraphy can be established (Figure 3). The basic unit of such a magnetostratigraphy is the magnetostratigraphical polarity zone (magnetozone for short), a body of rock with normal or reversed polarity. Now, the problem is to establish whether the observed succession of reversed and normal intervals has a pattern that can be correlated with that of another succession and/or with one or more segments of the global polarity time-scale. In other words, this piece (preserved by a local rock succession) needs to be matched with another piece or with a piece of the geomagnetic-polarity time-scale. This matching is termed 'correlation'. In correlation, a signature – a distinctive pattern of magnetic-polarity reversals – is looked for in order to establish a match (Figure 3).

The global pattern of magnetic reversals is irregular and nonperiodic, so that distinctive long intervals of magnetic reversals can be recognized. Reversals took place wordwide, independently of rock types and environments, and were geologically instantaneous (in rocks millions of years old, the 5000 or so years it takes for the reversal to happen is insignificant). However, polarity events are not unique, so that only a long succession of polarity intervals of distinctive lengths can be correlated. Furthermore, sedimentary hiatuses (unconformities) and changes in sedimentation rate can confuse the picture. Because of these problems, an independent method of correlation is often needed to provide a tie point (datum) against which to correlate the magnetostratigraphy. In other words, some other idea of the general age of the local rock succession – either derived from an index fossil or a numerical age – is usually needed to help narrow the possible correlation of magnetic-reversal histories. This means that magnetic-polarity-based correlations are typically not an independent means of correlating strata, although, once an index fossil or numerical age places the local slice of magnetic-polarity history 'in the ballpark', the matching of magnetic signatures often provides a more exact correlation than can be obtained from fossils or numerical ages alone.

Secular Variation

The Earth's magnetic north pole is close to, but not the same as, the geographical north pole. This means that, in most places, there is a small east-west difference between true north and magnetic north. The angle of this east-west deviation, measured from anywhere on Earth, is called the declination. For example, in California the declination is about 20° to the east, whereas in New York it is about 10° to the west of true north. The angle that the magnetic field makes with the Earth's surface is called the inclination. At the equator, the inclination is nearly horizontal, whereas at the magnetic pole it is vertical.

The magnetic field varies globally on geologically short time-scales of a few hundred years. These variations in declination, inclination, and field intensity are called secular variation. Secular variations are not magnetic reversals, but they are well documented over at least the last 10 000 years, and such palaeosecular variation can provide a succession of magnetic events that may be useful in correlation, particularly in archaeological research.

See Also

Analytical Methods: Geochronological Techniques. **Lava**. **Palaeomagnetism**. **Sedimentary Rocks:** Ironstones. **Stratigraphical Principles**. **Tectonics:** Mid-Ocean Ridges.

Further Reading

Butler RF (1992) *Paleomagnetism: Magnetic Domains to Geologic Terranes*. Boston: Blackwell.

Kennett JP (ed.) (1966) *Magnetic Stratigraphy of Sedi ments*. Stroudsburg, PA: Dowden, Hutchinson and Ross.

Khramov AN (1958) *Paleomagnetism and Stratigraphic Correlation*. Canberra: Australian National University [English translation, published 1960].

Khramov AN (1987) *Paleomagnetology*. Berlin: Springer Verlag.
McElhinny MW (1973) *Paleomagnetization and Plate Tectonics*. Cambridge: Cambridge University Press.
McElhinny MW and McFadden PL (2000) *Paleomagnetization: Continents and Oceans*. San Diego: Academic Press.
Opdyke ND and Channell JET (1996) *Magnetic Stratigraphy*. San Diego: Academic Press.
Tarling DH (1983) *Paleomagnetization*. London: Geological Society.
Tauxe L (1998) *Paleomagnetic Principles and Practice*. Dordrecht: Kluwer Academic Publishers.

MANTLE PLUMES AND HOT SPOTS

D Suetsugu, B Steinberger, and T Kogiso, Japan Marine Science and Technology Center, Yokosuka, Japan

Introduction

Hotspots are defined as anomalous volcanism that cannot be attributed to plate tectonics, unlike that associated with island arcs and spreading ridges. Mantle plumes, which are upwelling instabilities from deep in Earth's mantle, are thought to be responsible for hotspots that are relatively stationary, resulting in chains of islands and seamounts on moving oceanic plates. The volcanic rocks associated with hotspots have signatures in trace elements and isotopes distinct from those observed at mid-oceanic ridges and island arcs. Seismic imaging has revealed low-velocity anomalies associated with some deep-rooted hot mantle plumes, but images of their full-depth extent are of limited resolution, thus evidence for plumes and hotspots is primarily circumstantial. Commonly, it is not even clear which areas of intraplate volcanism are underlain by a mantle plume and should be counted as a hotspot.

Surface Expression of Hotspots

The primary surface expression of mantle plumes consists of hotspot tracks. These are particularly evident in the oceans as narrow (≈100 km) chains of islands and seamounts, such as the Hawaiian–Emperor chain, or as continuous aseismic ridges, such as the Walvis Ridge, up to several kilometres high. These tracks are thought to form as lithospheric plates move over plumes. The active hotspot is at one end of the chain; radiometric dating has determined that the ages of the volcanics along the chain tend to increase with distance from the active hotspot. Interpretation of age data is complicated, because volcanics do not necessarily erupt directly above a plume. Late-stage volcanism may occur several million years (My) after passage over a plume. Many hotspot tracks begin with a flood basalt or large igneous province. Volcanic volumes and age data indicate that these form during short time-spans with much higher eruption rates than are found at present-day hotspots. Examples of continental flood basalts (CFBs) include the Deccan Traps (associated with the Reunion hotspot) and the Parana basalts (associated with the Tristan hotspot). The Deccan Traps have erupted a volume of $\approx 1.5 \times 10^6\,km^3$ within less than 1 My, whereas the present-day eruption volume at the Reunion hotspot is $\approx 0.02\,km^3\,year^{-1}$. For other tracks, older parts have been subducted, and yet others, particularly shorter ones, begin with no apparent flood basalt. The length of tracks shows that hotspots may remain active for more than 100 My. For example, the Tristan hotspot track indicates continuous eruption for 120 My. Numerous shorter tracks exist as well, particularly in the south central Pacific, commonly without clear age progression. This may indicate either that the region is underlain by a broad upwelling or that widespread flow from a plume is occurring beneath the lithosphere, with locations of volcanism controlled by lithospheric stresses. Geometry and radiometric age data of hotspot tracks indicate that the relative motion of hotspots is typically slow compared to plate motions. However, for the Hawaiian hotspot between 80 and 47 million years ago (Ma), inclination of the magnetization of volcanics indicates formation at a palaeolatitude further north than Hawaii, with hotspot motion southward of several centimetres per year. The Hawaiian–Emperor bend may therefore represent more than a change in Pacific plate motion. In most other cases in which palaeolatitude data are available, inferred hotspot motion is slow or below detection limit. Associated with many tracks is a hotspot swell (≈1000 km wide, with up to 3 km anomalous elevation). Swells are associated with a geoid anomaly. Swell height slowly decreases along the track away from the active hotspot, and the swell also extends a few 100 km 'upstream' from the hotspot. The geoid-to-topography ratio remains approximately constant along swells, and this value indicates isostatic compensation at depths ≈100 km. From the

product of swell cross-sectional area and plate speed relative to the hotspot, plume anomalous mass flux (volume flux times density anomaly of plume relative to ambient mantle - this is the quantity that is directly estimated from observations), volume flux (volume flowing through plume conduit per time unit), and heat flux, (volume flux density times heat capacity times temperature anomaly of plume relative to ambient mantle), can be estimated. For Hawaii, the hotspot track and swell are both evident (Figure 1). The anomalous mass flux determined for Hawaii is the largest of any plume. Its estimated heat flux corresponds to ≈16% of global hotspot heat transport; global hotspot heat transport in turn is estimated to be ≈5% of total global heat flux.

Direct measurements of heat flow above hotspot swells yield only small anomalies (≈5–10 m W m $^{-2}$) or no anomalies at all. This may indicate that the swell is caused by buoyant uplift due to hot plume material spreading beneath the lithosphere, rather than by thermal erosion and heating of the lithosphere. The amount of volcanics produced at hotspots is estimated to be $<1\,km^3\,year^{-1}$, ≲5% of volcanics produced at mid-ocean ridges. Magma production at mid-plate hotspots thus appears less efficient than at mid-ocean ridges. However, during formation of large igneous provinces, magma production rate and heat flow were higher than at present.

Hotspots cluster in two antipodal regions around the Pacific and Africa. They are mostly absent from regions where subduction has occurred in the past 100 My, and tend be in highs of the 'residual' geoid (actual geoid minus contribution of subducted slabs). Probably the hotspot plumes are not the primary cause of these geoid highs. Rather, both may be due to the same cause: a less dense and presumably hotter lower mantle in these regions. Plate boundaries, in particular mid ocean ridges, may move across hotspots; hotspots may hence, successively or simultaneously, leave tracks on different plates. Examples are the Reunion hotspot, with tracks on the Indian and African plates, and the Tristan hotspot, with tracks on the South American and African plates. If a plume is located close to a spreading ridge, eruption of plume material may occur not only directly above the plume, but also at the section of the ridge

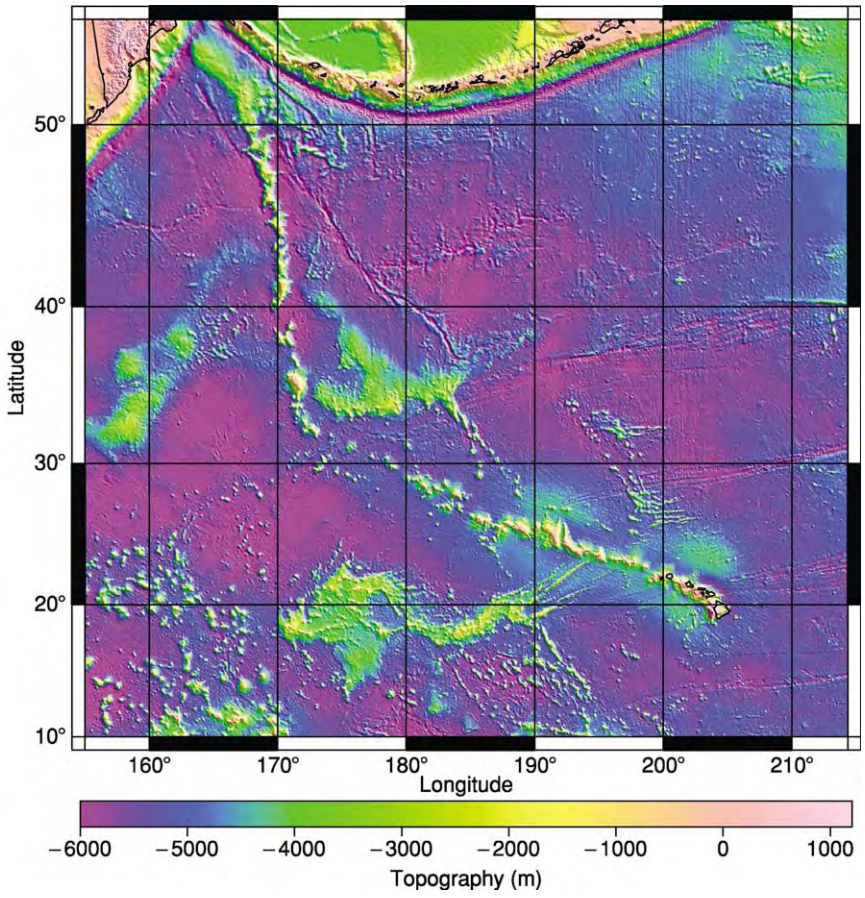


Figure 1 Topographic map of the North Pacific, showing both the swell and the track associated with the Hawaiian plume.

close to the hotspot. Elongated aseismic ridges may be formed by volcanics above channels along which plume material flowed to the spreading ridge (for example, in the Musicians Seamounts near 200° E, 25° N) (Figure 1). Hotspots and their tracks are less obvious on continents. The Yellowstone National Park area in the United States is frequently regarded as a hotspot plume, with the Snake River Plain being its hotspot track.

Seismic Images of the Mantle Plumes

Column-like anomalies of low seismic velocity associated with high temperatures are expected under hotspots, if hotspots are the surface expression of mantle plumes. Seismic imaging of hotspots has advanced in the past decade, and seismic images beneath some hotspots have been obtained. Commonly, these are imaged only at specific depths. For some hotspots, no low-velocity anomalies have been found.

Upper Mantle

Global mapping of the upper mantle by long-period (50–300 s) surface waves has revealed low seismic velocities associated with hotspots and spreading ridges. There is a distinct difference between hotspots and ridges, concerning the depth extent of these low-velocity regions : low seismic velocities beneath hotspots extend to a depth of 200 km, whereas low seismic velocities beneath ridges are confined to the upper 100 km. This suggests that hotspots are caused by active upwellings (mantle plumes) with deeper sources, compared to ridges, which may be caused by passive upwelling. The 200-km depths of the slow velocities under hotspots do not necessarily correspond to the actual source depths of mantle plumes, but rather to the depths to which surface waves can resolve. Seismic array observations have been carried out in hotspot regions to resolve fine structures such as plume conduits. A recent example of an S-velocity model beneath the Icelandic hotspot was obtained from body and surface wave data recorded by a temporary seismic array; a low-velocity plume can be seen beneath the hotspot (Figure 2). A 200-km-thick low-velocity zone extends laterally beneath Iceland; a vertical column of low velocities under central Iceland extends to a depth of at least 400 km. Similar array observations carried out at other hotspots (e.g., Hawaii, Yellowstone, and Massif Central) detected low-velocity anomalies extending to sublithospheric depths in the upper mantle.

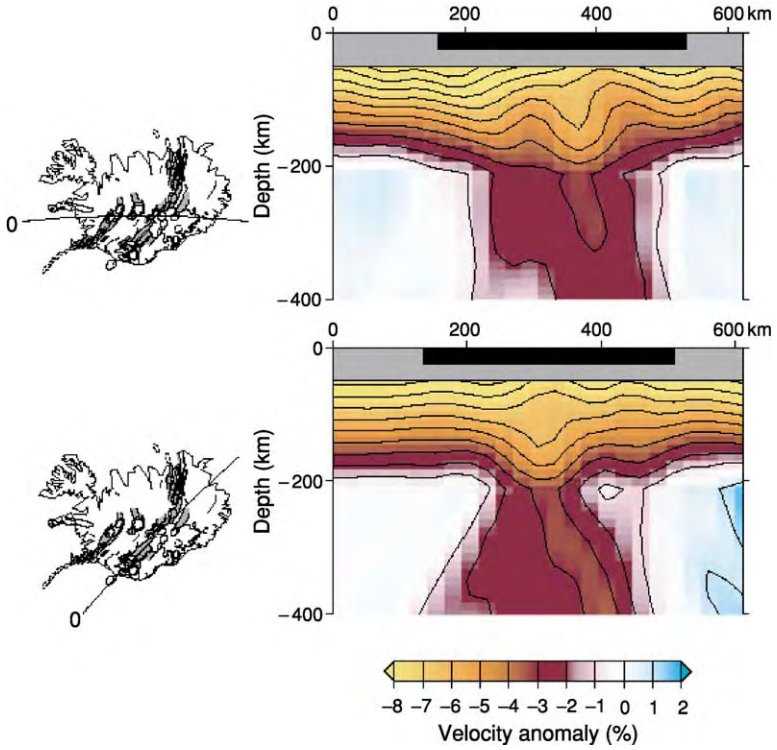


Figure 2 The S velocity profile beneath the Icelandic hotspot. Reproduced with permission from Allen RM, Nolet G, Morgan WJ, *et al.* (2002) Imaging the mantle beneath Iceland using integrated seismological techniques. *Journal of Geophysical Research Solid Earth* 107(B12): 2325, doi:10.1019/2001 JB000595.

The Transition Zone

The mantle transition zone is bounded by seismic discontinuities at depths around 410 and 660 km (referred to hereafter simply as '410' and '660'); the discontinuities at these two depths are interpreted to be due to olivine–β-olivine and postspinel phase changes, respectively. Topography of the 410 and 660 provides important data for understanding thermal structure and the ascending process of hot plumes; the 410 and 660 are expected to be depressed and elevated, respectively, in and near a hot plume, because the olivine–β-olivine and the postspinel phase changes have positive and negative Clapeyron slopes, respectively. The thickness of the transition zone (defined as the interval between the 410 and 660) is also used as a measure for thermal anomalies, because estimates are less affected by upper mantle velocity structure and therefore can be more reliably determined. Discontinuity depths are determined by detecting waves reflected or converted at the discontinuities, and measuring their timings. Using P- to S-converted waves recorded by regional seismograph arrays, the transition zone beneath Iceland and Hawaii has been estimated to be thinner than in the surrounding mantle by ≈20 and ≈40 km, respectively, for areas 400 km in diameter. Amounts of thinning can be converted to temperature anomalies, with the Clapeyron slope obtained experimentally, suggesting temperatures of ≈150 K and ≈300 K beneath Iceland and Hawaii, respectively. The transition zone beneath the Society hotspot in the South Pacific Superswell was estimated, using S waves reflected beneath the discontinuities, to be 20–30 km thinner than it is in the surrounding area (an area 500 km in diameter), suggesting temperatures that are 150–250 K hotter (Figure 3). These results suggest that mantle plumes beneath these hotspots originate at least from transition zone depths, and possibly from the lower mantle.

Lower Mantle and D″ Layer

To explore the deeper structure beneath hotspots requires analysis of body waves, which penetrate much deeper in the mantle. Global tomographic studies using body waves in the lower mantle have identified possible 'plume conduit signatures' beneath some hotspots, but not all. At present, different studies using different data and methods give somewhat inconsistent results for the same hotspot, making it difficult to derive definitive conclusions about the geometry of plume conduits and the source depths of hotspots. Different depth extents may indicate various source depths for hotspots. Two robust features – broad, low-velocity anomalies – have been identified beneath the South Pacific and Africa (Figure 4); often referred to as 'superplumes', these features have diameters of 1000–2000 km. The latter is tilted from south-west to the north-east, suggesting that the plume conduit may be advected by large-scale mantle flow. These superplumes may be related to the hotspots in the South Pacific Superswell and Africa, although robust seismic images of the plume conduits to each of the hotspots have not been obtained.

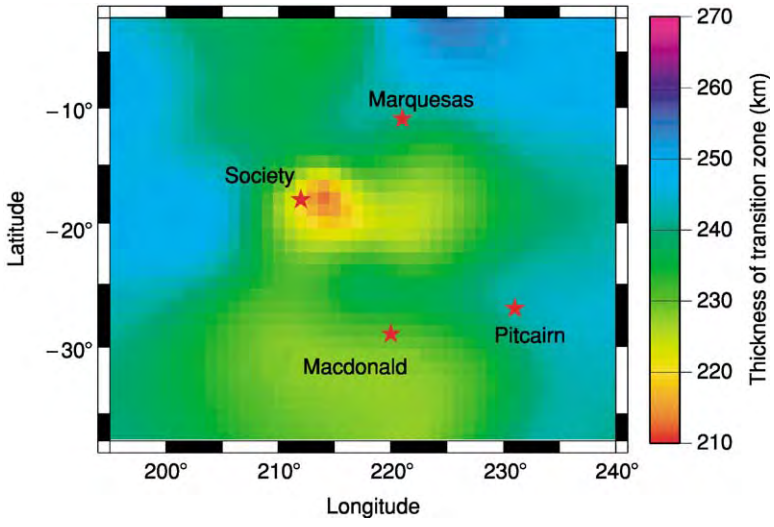


Figure 3 Thickness of the transition zone beneath the South Pacific Superswell. Deviation from the global average is shown (colour bar, kilometres). Stars represent hotspots. Reproduced with permission from Niu F, Solomon SC, Silver PG, Suetsugu D, and Inoue H (2002) Mantle transition zone structure beneath the South Pacific Superswell and evidence for a mantle plume underlying the Society hotspot. *Earth and Planetary Science Letters* 198: 371 380.

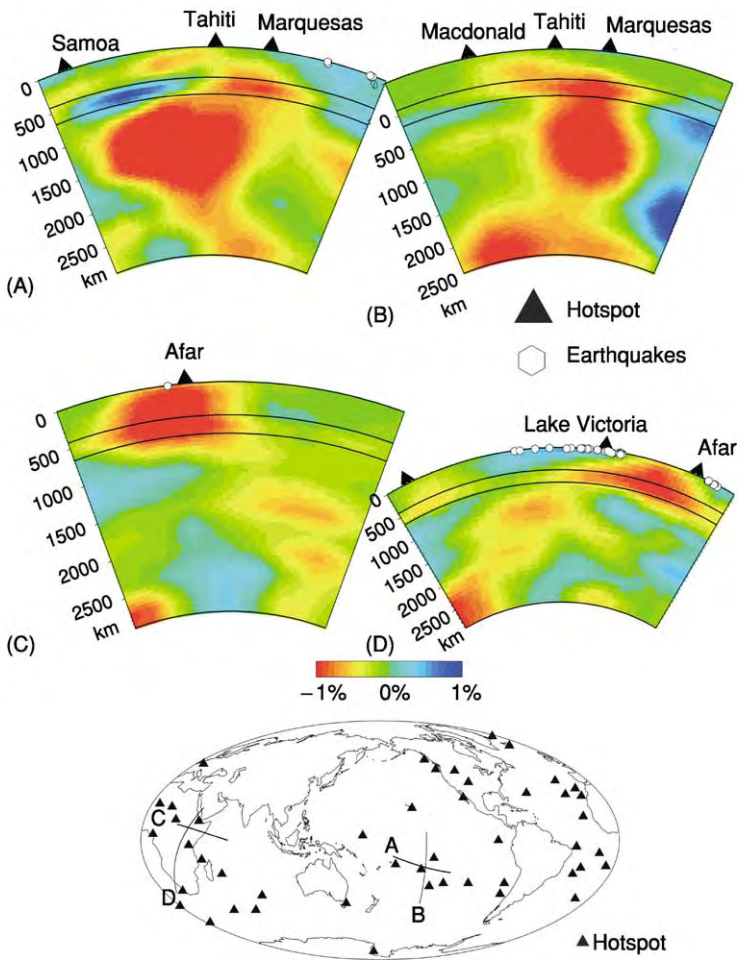


Figure 4 The P velocity profile of the whole mantle beneath the South Pacific Superswell (top) and Africa (middle). (Bottom) Locations of profiled areas. Reproduced with permission from Zhao D (2001) Seismic structure and origin of hotspots and mantle plumes. *Earth and Planetary Science Letters* 192: 251 265.

Over most of the lower mantle depth range, lateral variation of seismic velocities is less than it is in the upper mantle, but pronounced lateral variation does occur at depths below 2500 km. Broad, low-velocity anomalies located beneath the South Pacific Superswell and Africa seem to be linked to the superplumes; high-velocity anomalies beneath circum-Pacific regions may be associated with subducted slabs. At the bottom of the D″ layer (the lowermost part of the mantle), above the core and mantle boundary (CMB), 5- to 50-km-thick zones of very low velocities have been found; these ultra-low-velocity zones (ULVZs) have low P and S velocity anomalies (10% or greater) and may represent a layer of partial melting and/or chemical heterogeneities. There is a spatial correlation between the ULVZs and hotspots, suggesting that the ULVZs may be a possible source of hotspots. Many plume conduits may be too narrow to be detected by present seismic data and techniques; dense seismic arrays on the land and seafloor and new analysis techniques may be required to resolve mantle plumes at full depth.

Petrological and Geochemical Signatures of Hotspot Rocks

Volcanic rocks associated with plume-related (hotspot) volcanism are quite variable in rock type and chemical composition and show significant correlation with the tectonic environment in which they erupt. Within the oceanic environment, hotspot volcanism produces mainly basaltic lavas, often called ocean island basalts (OIBs); these have chemical characteristics distinct from those of mid-ocean ridge basalts (MORBs), of which most of the ocean floor is composed. Hotspot volcanism in the oceanic environment sometimes produces voluminous oceanic plateaus or rises (e.g., the Ontong Java Plateau). In

continental environments, mantle plume activity is commonly manifested by eruption of huge floods of basaltic lavas, commonly called continental flood basalts (CFBs). Less voluminous magmatic activities also occur in continents with extensional tectonics such as intracontinental rifts and grabens. Volcanics in continental rift systems are thought to be products of mantle plume activity, although some continental rift volcanism is not clearly related to mantle plumes.

OIBs are commonly olivine-bearing lavas with subalkalic (tholeiitic) and alkalic compositions. Some large ocean islands, such as Hawaii and Iceland, are dominantly composed of tholeiitic basalts and their derivatives, with small amounts of alkalic rocks at later stages of eruptive sequences. A vast majority of ocean islands are composed entirely of alkalic basalts and their derivatives. The major element composition of tholeiitic OIBs is generally similar to that of MORBs, although lavas with higher MgO content are more abundant than at mid-oceanic ridges. Alkali olivine basalt is the most abundant type in alkalic OIB suites, but more alkalic lavas, such as basanite and nephelinite, also commonly erupt in many ocean islands. Compared with MORBs, both tholeiitic and alkalic OIBs have higher concentrations of incompatible trace elements, and their chondrite-normalized rare earth element (REE) patterns show variable but strong enrichment in light REEs relative to heavy REEs. Alkalic OIBs are typically more enriched in incompatible elements, including light REEs, compared to tholeiitic OIBs. Sr–Nd–Pb–Os isotopic ratios of OIBs show quite large variations relative to MORBs, which extend from nearly MORB-like values towards one component with highly radiogenic Pb isotopes (designated by the term 'HIMU', referring to the high U/Pb source ratios) and other components with nonradiogenic Pb isotopes (designated as 'enriched mantle 1', or EM1) and with high $^{87}Sr/^{86}Sr$ ratios (EM2). Variation in helium isotopic ratios of OIBs extends considerably towards the primordial value with high $^{3}He/^{4}He$ ratios, although some OIB suites have slightly lower $^{3}He/^{4}He$ ratios than MORBs have. Volcanic rocks of oceanic plateaus and rises are commonly aphyric lavas with tholeiitic composition. Their major element composition is relatively homogeneous compared to tholeiitic OIBs and more akin to MORBs. Oceanic plateau basalts have trace element characteristics similar to those of MORBs, but they are slightly more enriched in incompatible elements. Chondrite-normalized REE patterns are typically flat or slightly enriched in light REEs. Radiogenic isotope ratios of oceanic plateau basalts are quite variable, but their entire range is smaller than that of OIBs.

Petrological features of CFBs are similar to those of oceanic plateau basalts. Most CFB suites are composed of voluminous aphyric lavas with tholeiitic composition, and their major element compositions cover similar ranges as those covered by MORBs. However, compared to MORBs, CFBs generally have higher FeO/MgO ratios, and SiO_2-rich (andesitic) lavas are more dominant in CFBs. Most CFB suites are associated with minor amounts of alkalic rocks. In contrast to major elements, trace element features of CFBs are more like those in OIBs than in oceanic plateau basalts and MORBs. CFB suites are highly enriched in incompatible elements, sometimes even more than OIBs are, and have REE patterns with strong enrichment in light REEs. Sr–Nd–Pb–Os isotopic ratios of CFBs are much more variable than those of OIBs are, extending beyond the highest $^{87}Sr/^{86}Sr$ and the least radiogenic Pb isotopes of OIBs. Volcanic rocks associated with continental rift systems show considerable diversity in rock type, from transitional olivine basalt-rhyolite to strongly alkalic nephelinite–phonolite series, melilitites, and even carbonatite lavas (Ol Doinyo Lengai, Tanzania). A noticeable feature of continental rift volcanics is that acidic rocks, such as rhyolite, trachyte, and phonolite, are more abundant than in other environments. Enrichment of incompatible elements in continental rift volcanics is as great as or sometimes much greater than in alkalic OIBs. Variations of radiogenic isotope ratios are as wide as those in CFBs.

Geochemical diversity in plume-related volcanic rocks is generally attributed to chemical heterogeneity of mantle plume material. In particular, isotopic variations of OIBs, which are least affected by crustal material, clearly indicate that mantle plumes contain several geochemical components. The origin of such geochemical components in the mantle is still in dispute, but the most commonly accepted model is that subducted crustal materials (oceanic basaltic crust plus silicic sediments) contribute to formation of distinct geochemical reservoirs in the mantle, such as HIMU, EM1, and EM2. Chemical interaction between core and mantle has also been proposed to explain isotopic diversity of OIBs, especially for Pb and Os. High $^{3}He/^{4}He$ ratios in some OIB suites require the existence of some primordial volatile component in mantle plumes. Extremely radiogenic Sr isotopic ratios observed in some CFBs and continental rift volcanics have been explained by contamination or assimilation of continental crust. Major element diversity of plume-related volcanic rocks has been interpreted in terms of potential temperature of mantle plumes. A conventional view is that tholeiitic basalts in relatively large ocean islands and CFB suites are produced at higher degrees of melting in hot

mantle plumes, and alkalic rocks are produced at lower degrees of melting in colder plumes. On the other hand, the major element diversity of plume rocks can be explained by heterogeneity in major element composition of mantle plumes, because subducted crustal material, if contained in mantle plumes, may produce both tholeiitic and alkalic magmas under mantle conditions. In this case, no significant excess temperature is required for the plume, because the melting temperature of crustal materials is much lower than that of mantle peridotite.

Dynamics of Mantle Plumes

In Earth's mantle, the Rayleigh number, a measure of the vigour of convection, is estimated to be 5×10^6 to 5×10^7. For this range, time-dependent convection, with instabilities originating at thermal boundary layers, is expected. Plume formation may, however, be suppressed by large-scale flow related to plate motions, which, if sufficiently fast, advects growing instabilities towards large-scale upwellings before individual plumes can form. For somewhat slower large-scale flow, individual plumes may exist, but are advected towards, and hence cluster in, regions of large-scale upwellings. A thermal boundary layer exists above the core–mantle boundary, which is therefore a likely plume source depth. Another thermal boundary and possible source depth is at the 660. Plumes may also originate at chemical boundaries within the mantle. No conclusive evidence exists for those, but anticorrelation of bulk sound and shear wave anomalies in the lowermost mantle indicates that the D″ layer at the base of the mantle may be chemically distinct from the overlying mantle. Topography of the chemical boundary could be up to a few 100 km, and plumes may rise from the high points (cusps) on the boundary. Plumes may entrain material from the underlying layer; the amount of entrained material depends on the density difference between the two layers, due to chemical and temperature differences. Both laboratory and numerical experiments show that a plume that is less viscous than the surrounding mantle tends to have a large, roughly spherical or mushroom-shaped head, followed by a narrow tail (conduit) connecting the head and the source region (Figure 5). Scaled to Earth dimensions, a head diameter of several 100 to 1000 km is expected, if it rises from the lowermost mantle. Such a size is also required to explain the volume of many large igneous provinces, thus adding support to the notion that these originate from upwellings from the lowermost mantle, whereas hotspots not associated with flood basalts may be caused by plumes from shallower depth. Conduits should be only ≈100 km wide. Thinner conduits may occur for strongly temperature-dependent viscosity. If mantle viscosity increases strongly with depth, conduit diameter could be up to a few hundred kilometres in the lower mantle. Ascent times of plume heads through the whole

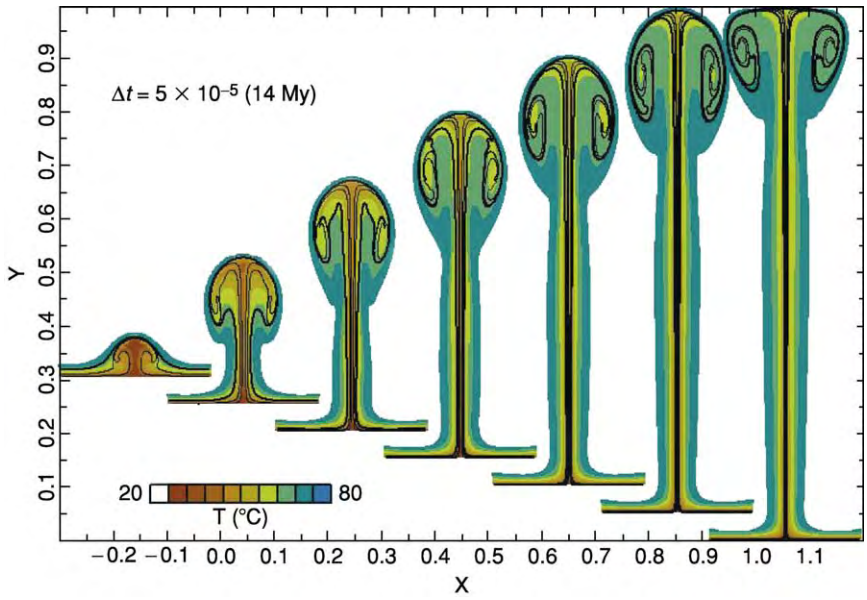


Figure 5 Numerical model of an axisymmetric mantle plume with strongly temperature dependent viscosity. Colours indicate temperature (red = hottest, blue = ambient mantle); black lines indicate marker chains. Plots are at 14 My time intervals. Reproduced with permission from van Keken P (1997) Evolution of starting mantle plumes: a comparison between numerical and laboratory models. *Earth and Planetary Science Letters* 148: 1 11.

mantle are estimated to be ≈10 My to several tens of millions of years. For nonlinear mantle rheology, decrease of effective viscosity due to larger stresses around plume heads may cause rise times shorter than for Newtonian viscosity. As a plume head rises, the surrounding mantle heats up and becomes buoyant and less viscous. Mantle material is entrained into the rising plume, which therefore contains a mixture of materials from the source region and ambient mantle. Formation of a flood basalt is thought to occur when a plume head reaches the base of the lithosphere. Subsequently, the conduit may remain in existence as long as hot material is flowing in at its base – for 100 My or longer. The conduit consists of a narrow core, where most of the material transport occurs, and a thermal halo. Due to thermal diffusion, heat is lost from plume conduits as they traverse the mantle. Though strong plumes, such as the Hawaii plume, are not significantly affected, heat loss significantly reduces the temperature anomaly expected for weaker plumes. For weak plumes from the core–mantle boundary, with anomalous mass flux of $\lesssim$500–1000 $m^3 s^{-1}$, the sublithospheric temperature anomaly is low and no melting is expected. Weak hotspots may therefore have shallower origins. On the other hand, the temperature anomaly of strong plumes, inferred from observations, is much less than expected from the temperature drop across a thermal boundary layer between core and mantle. This may indicate a chemically distinct layer at the base of the mantle, with plumes rising from its top. Mantle plumes may coexist with superplumes, and conduits are expected to be tilted and distorted in large-scale mantle flow. The rising of a tilted conduit may cause further entrainment of ambient mantle material. If the tilt exceeds ≈60°, the conduit may break into separate diapirs, which may lead to extinction of the plume. As a consequence of conduit distortion, overlying hotspots are expected to move. Hence, mantle plumes probably do not provide a fixed reference frame. However, if plumes arise from a high-viscosity lower mantle, hotspots should move much more slowly than lithospheric plates move. Conduits are likely to be time variable, with disturbances traveling along them; these may be wave-like or may take the shape of secondary plume heads. Waves are associated with increased conduit flux, which may explain flux variations in mantle plumes. Ascending plumes interact with mantle phase transitions. The 660 somewhat inhibits flow across but is unlikely to block penetration of plumes. Experiments involving high pressure suggest phase relations of a pyrolite mantle such that, at the high temperatures of mantle plumes, this phase boundary does not hinder flow across. Beneath the lithosphere, buoyant plume material flows out of the conduit, spreads horizontally in a low-viscosity asthenosphere, and is dragged along with moving plates. Plume material buoyantly lifts up the lithosphere and causes a hotspot swell. Partial melt extraction at the hotspot may leave behind a buoyant residue that also contributes to swell formation. Plume material does not necessarily erupt directly above the conduit. It may also flow upward along the sloping base of the lithosphere, and enhanced melting may occur at steep gradients. A sloping base exists near spreading ridges. If ridge and hotspot are less than a few hundred kilometres apart, eruption of volcanics may occur at the ridge rather than, or in addition to, directly above the plume. Also, in other cases, such as in Africa, the spatial distribution of plume-related melting and magmatism may be controlled by the lithosphere rather than by the plume position. Formation of vertical fractures and ascent of magma through the lithosphere preferably occur for tensile lithospheric stresses. Loading of the lithosphere by hotspot islands causes stresses that may influence formation of fractures and therefore determine the spacing of hotspot islands along tracks. Feeding of plume material to a nearby ridge may put the lithosphere above the plume under compression and shut off eruption directly above the plume. If a hotspot (e.g., Iceland) is located close to the ridge, the viscosity contrast between plume and ambient mantle may become a factor 1000 or more beneath thin lithosphere. Such large viscosity variations facilitate ridge-parallel flow of plume material and help to explain geochemical anomalies south of Iceland. Propagation of pulses in plume flux explains the V-shaped topography and gravity anomalies at the Reykjanes Ridge. Probably not all intraplate volcanism is caused by plumes as described. In many cases, the origin of intraplate volcanism may be shallow, due to cracks in the lithosphere caused by tensional stresses, or due to edge-driven convection at locations where lithospheric thickness varies laterally.

See Also

Igneous Processes. **Large Igneous Provinces**. **Lava**. **Plate Tectonics**. **Rocks and Their Classification**. **Seamounts**. **Tectonics:** Propagating Rifts and Microplates At Mid-Ocean Ridges; Seismic Structure At Mid-Ocean Ridges.

Further Reading

Allen RM, Nolet G, and Morgan WJ, *et al.* (2002) Imaging the mantle beneath Iceland using integrated seismological techniques. *Journal of Geophysical Research Solid Earth* 107(B12): 2325, doi:10.1019/2001 JB000595.

Courtillot V, Davaille A, Besse J, and Stock J (2003) Three distinct types of hotspots in the Earth's mantle. *Earth and Planetary Science Letters* 205: 295 308.

Dickin AP (1995) *Radiogenic Isotope Geology*. Cambridge: Cambridge University Press.

Hofmann AW (1997) Mantle geochemistry: the message from oceanic volcanism. *Nature* 385: 219 229.

Jackson I (ed.) (1998) *The Earth's Mantle: Composition, Structure, and Evolution*. Cambridge: Cambridge University Press.

Mahoney JJ and Coffin MF (eds.) (1997) *Large Igneous Provinces: Continental, Oceanic, and Planetary Flood Volcanism*. American Geophysical Union Geophysical Monograph 100. Washington, DC: American Geophys ical Union.

Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl RE, and Hung S H (2004) Finite frequency tomography reveals a variety of plumes in the mantle. *Science* 303: 338 343.

Nataf H C (2000) Seismic imaging of mantle plumes. *Annual Review of Earth and Planetary Sciences* 28: 391 417.

Niu F, Solomon SC, Silver PG, Suetsugu D, and Inoue H (2002) Mantle transition zone structure beneath the South Pacific Superswell and evidence for a mantle plume underlying the Society hotspot. *Earth and Planet ary Science Letters* 198: 371 380.

Ritsema J and Allen RM (2003) The elusive mantle plume. *Earth and Planetary Science Letters* 207: 1 12.

Schubert G, Turcotte DL, and Olson P (2001) *Mantle Convection in the Earth and Planets*. Cambridge: Cambridge University Press.

van Keken P (1997) Evolution of starting mantle plumes: a comparison between numerical and laboratory models. *Earth and Planetary Science Letters* 148: 1 11.

Wilson M (1989) *Igneous Petrogenesis*. London: Unwin Hyman.

Zhao D (2001) Seismic structure and origin of hotspots and mantle plumes. *Earth and Planetary Science Letters* 192: 251 265.

Zindler A and Hart S (1986) Chemical geodynamics. *Annual Reviews for Earth and Planetary Sciences* 14: 493 571.

MARS

See **SOLAR SYSTEM: Mars**

MERCURY

See **SOLAR SYSTEM: Mercury**

MESOZOIC

Contents

Triassic

S G Lucas, New Mexico Museum of Natural History, Albuquerque, NM, USA
M J Orchard, Geological Survey of Canada, Vancouver, BC, Canada

Introduction

In 1834, the German geologist Frederich August von Alberti coined the term 'Triassic' for rocks originally recognized in Germany as the Bunter, Muschelkalk, and Keuper formations. Today, the rocks of Triassic age (~200 to 251 million years ago) are recognized on all continents (Figure 1). Most of these are sedimentary rocks consisting of dominantly shallow-water carbonates of marine origin and siliciclastic red beds of non-marine origin. These rocks represent a record of sedimentation on and around the vast Pangaean supercontinent and tell the tale of its final union and the initiation of its subsequent fragmentation. In this brief overview of the global Triassic, the rock record, time-scale, palaeogeography, tectonics and sedimentation, sea-levels, climate, and biota of the time period are considered.

Triassic Rocks

Triassic rocks are exposed on all the world's continents (Figure 1). Estimates of their maximum thickness have been given as 9 km, and their total volume is ~45 million km^3. These estimates are slightly more than the values estimated for the Permian, but substantially less than estimates for the Jurassic or Cretaceous. Triassic rocks are mostly sedimentary in origin, and volcanic rocks do occur in relatively minor amounts: they have been estimated as constituting ~1–2% overall of the exposed Triassic rocks in the Americas. Triassic volcanic rocks can be substantial in some regions, however, as exemplified in the Pacific north-west of North America.

The Triassic was a time of great continental emergence due to a combination of widespread epeirogenic uplift and relatively low sea-level. Marine deposition was mostly confined to the Tethys, the circum-Pacific, and the circum-Arctic (Figure 1). A worldwide survey identifies 15 significant Triassic outcrop belts: (1) the Cordillera of the western United States and western Canada, which exposes significant accumulations of both marine and non-marine strata, as well as a substantial record of Triassic rocks in accreted terranes; (2) the Newark Supergroup non-marine rift basins of eastern North America; (3) extensive marine and non-marine deposits of eastern Greenland, Franz Josef Land, and Svalbard; (4) western Europe, from the dominantly non-marine deposits of the Germanic basin system to the dominantly marine strata of the northern Mediterranean; (5) the extensive, dominantly marine deposits of north-eastern Siberia; (6) shallow marine deposits in Israel; (7) marine deposits in the Transcaucasian region of Iran and Azerbaijan, which include some very fossiliferous sections of the Permian-Triassic boundary (PTB); (8) dominantly marine strata of the Caspian Basin and Mangyshlak Peninsula of western Kazakhstan; (9) the Himalayan belt from Afghanistan and Pakistan through Kashmir into Tibet, also the location of some very fossiliferous PTB sections; (10) extensive non-marine deposits of the Junggur and Ordos basins of northern China; (11) marine deposits of southern China, South-east Asia, and Indonesia, including the most well-studied PTB section at Meishan in China and the phenomenal ammonite-bearing beds of Timor; (12) mixed marine-non-marine deposits on the western and eastern coasts of Australia; (13) extensive marine deposits in New Zealand; (14) deep marine deposits in Japan; and (15) non-marine strata exposed in the Transantarctic Mountains of Antarctica.

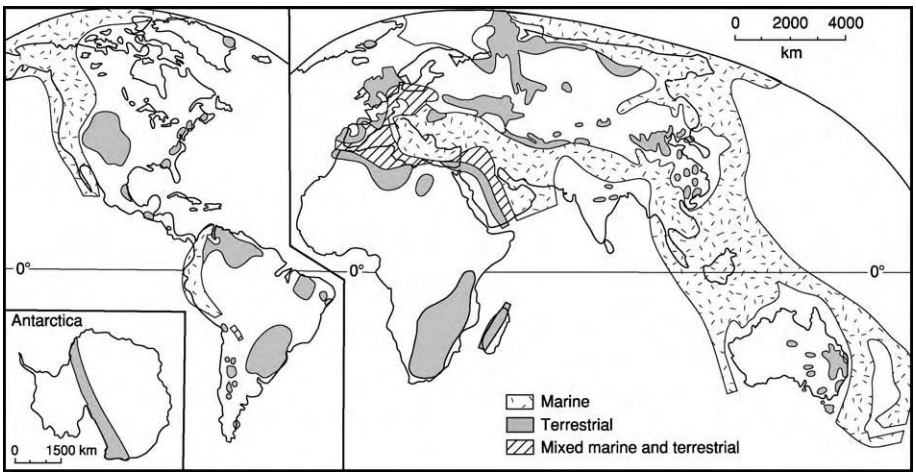


Figure 1 Map of the present world, showing the distribution of Triassic rocks. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

Time-Scale

The standard global chronostratigraphic scale for the Triassic is divided into seven stages. In ascending order, these are the Induan, Olenekian, Anisian, Ladinian, Carnian, Norian, and Rhaetian (Figure 2). Many other stage names of various scope and utility are employed regionally. For example, the Lower Triassic Griesbachian, Dienerian, Smithian, and Spathian stages are based on Arctic successions and have widespread currency in North America. Similarly, Tethyan substage names remain widely used in Europe. Definition and subdivision of the Triassic stages have been based largely on ammonoid biostratigraphy. However, only a single Global Stratotype Section and Point (GSSP) has been defined for the Triassic time-scale (for the Induan), and this is based on a conodont first occurrence. This datum defines the base of the Triassic system within the stratotype section at Meishan in southern China. International agreement and definition of the bases of the remaining six stages are under consideration.

The relatively low level of Triassic volcanism results in a dearth of numerical ages to provide geochronological control of the time-scale. Nevertheless, some important advances have been made in the past decade. The most intensively studied is the Permian–Triassic boundary at Meishan, South China, where U/Pb ages measured from zircons in ash beds both above and below the defined boundary provide a range of ages around 250 to 252 Ma. An age of 251 Ma for the PTB seems most consistent with ages derived from other PTB sections (Figure 2). The age of the Lower–Middle Triassic (Olenekian–Anisian) boundary is now determined from the Nanpanjiang Basin in Guizhou and Guangxi Provinces, South China. Several ash bands bracketing this interval

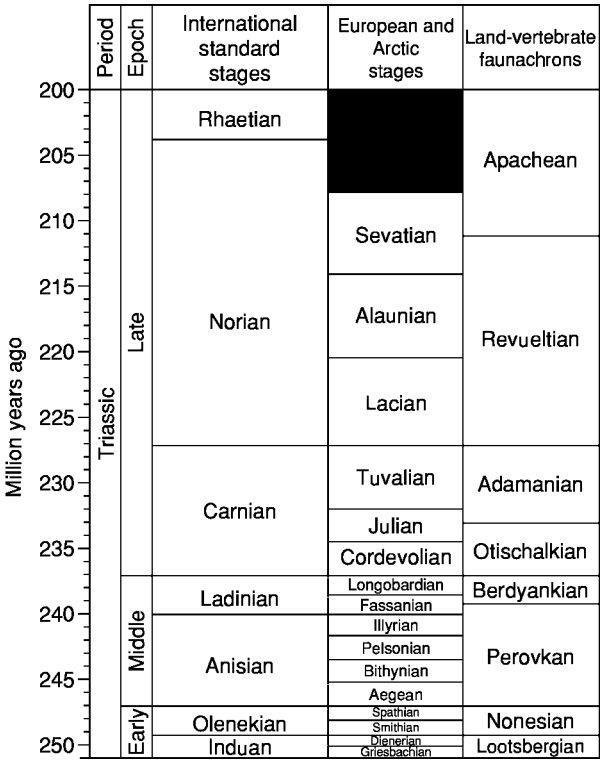


Figure 2 A simplified Triassic time scale. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

have been dated recently, and the boundary age seems confidently set at ~247 Ma. One staggering implication of this is that the Early Triassic would appear to be on the order of only 4 million years long!

Several radioisotopic ages are available from around the Anisian–Ladinian boundary. One potential position, at the base of the *Nevadites secedensis* ammonoid zone in Italy, is dated with U/Pb at

~240 Ma, whereas tuffs within the underlying *Reitziites reitzi* ammonite zone in Hungary give maximum ages closer to 241 Ma. The Upper Triassic stages are poorly constrained radioisotopically. The Ladinian–Carnian boundary approximates 237 Ma based on U/Pb ages from Italy and Hungary. Hence, the Upper Triassic is over half the length of the duration of the Triassic Period. Estimates for the Carnian–Norian boundary have been based on extrapolation from lithological cyclicity recognized in the non-marine Newark Supergroup of eastern North America. These cycles are interpreted as a response to Milankovitch orbital periodicities and can thus be measured in absolute terms. This astronomically calibrated numerical scale is set against the magnetostratigraphic polarity profile, which, in turn, can be correlated with successions elsewhere. However, there is no current agreement either on a position for a marine-based boundary or on precise correlations of the fossil scale with the Newark succession. Independent data come from an Ar/Ar age of a tuff that underlies Adamanian (=Late Carnian) tetrapods in Argentina. This gives an age of ~228 Ma. A Triassic–Jurassic boundary age of 202 Ma has been based on the ages of Newark Supergroup basalts that overlie the palynologically determined boundary. However, the palynological correlation of the boundary is debated and a more recent U/Pb age from British Columbia gives a Triassic–Jurassic boundary age very close to 200 Ma.

There is essentially no preserved Triassic seafloor, so there is no agreed geomagnetic polarity time-scale for the Triassic. However, a composite polarity time-scale is now becoming available, based on successions being cobbled together from non-marine and marine sections in North America, Europe, and Asia. With continued refinement, this scale will be an important supplement to that provided by biostratigraphy in both the marine and non-marine realms.

Palaeogeography

At the onset of the Triassic, the world's continents were assembled into a single supercontinent called Pangaea (*see* **Pangaea**) (Figure 3). The rest of the globe comprised a single vast ocean called Panthalassa, with a westward-extending arm called Tethys. This followed the Late Palaeozoic assembly of the continents when Laurentia, Asia, and Gondwana collided along the Alleghanian–Variscan–Ural mountain chains. The nearly hemispheric Pangaean supercontinent was encircled by subduction zones that dipped beneath the continents while the Panthalassan and Tethyan plates carried island arcs and oceanic plateaus that were destined to become accreted to the continental margins.

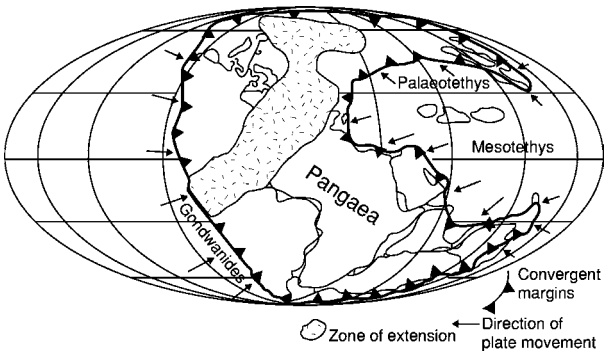


Figure 3 Triassic Pangaea, showing major tectonic elements. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

The supercontinent drifted northward and rotated clockwise throughout the Triassic, so there was considerable latitudinal spread to the landmass, which was nearly symmetrical about the equator (Figure 3). However, no sooner had the supercontinent been assembled than significant fragmentation began. Thus, Gondwana and Laurasia began to separate in Late Triassic time with the onset of rifting in the Gulf of Mexico basin. It was not until the Early Jurassic, though, that significant marine sedimentation took place in the nascent Atlantic Ocean basin.

Tectonics and Sedimentation

At the broadest level, the tectonics of Triassic Pangaea were simple. The accreted supercontinent was simply surrounded by convergent margins (Figure 3). However, these margins were actually complex belts of magmatic arcs and terranes moving in various directions to produce compressional and transpressional tectonics. Late Triassic Pangaea was the site of widespread extensional tectonism, especially the initial opening of the Atlantic Ocean basin by rifting of the North American and African plates. During the Late Triassic, in the Tethys, North Atlantic, and Arctic, multidirectional rift systems developed (Figure 3). Rifting also took place along a zone of transforms that extended well into the Gulf of Mexico basin and, punctuated by volcanism, dominated the northern border of western Tethys. This rifting in the North Atlantic and Tethyan regions subjected western and central Europe to progressive regional extension, culminating in the development there of complex multidirectional systems of troughs and grabens. During the Early–Middle Triassic, terminal thrusting took place along the entire Gondwanan margin of Pangaea, which was followed in the Carnian by extension in southern South America and eastern Australia.

Most Triassic sedimentation took place in one of three types of basins: foreland, fore-arc, or extensional. Perhaps the best example of a Triassic foreland basin is the Karoo Basin of South Africa, a retroarc foreland basin originally formed by the collision of the palaeo-Pacific and Gondwana plates during the Late Carboniferous. In the Karoo Basin, 12 km of Carboniferous-Jurassic red beds accumulated. Most of the Pangaean marginal basins were part of an array of arc-trench systems that surrounded much of the supercontinent. A good example is the complex Cordilleran basin of western North America, in which deposition took place between an offshore island arc and the continental margin. In the western United States portion of this basin, 1.2-km siliciclastic red beds were shed to the north-west and interfinger with marine carbonates deposited in the arc-trench system.

Of the (mostly Late Triassic) extensional basins, perhaps the best studied is the Newark basin in the eastern United States. This was a dip–slip-dominated half graben in which ~7 km of mostly lacrustine Upper Triassic–Lower Jurassic sediments accumulated. There were also other types of Triassic extensional basins more complex than the Newark half grabens, such as those of the Germanic basin system of north-western Europe.

Sea-Levels

Early Mesozoic plate reorganization was apparently associated with the development of new seafloor-spreading axes, which caused a general reduction of ocean basin volume during the Triassic. Pangaea was very emergent and, because of its high freeboard, the Triassic was a time of relatively low sea-level, which may be termed a first-order Pangaean global lowstand. After the major sea-level fall of the latest Permian, sea-level apparently rose through much of the Triassic, to peak during the Norian and then fall near the end of the period (Figure 4). There were, however, short significant falls in sea-level, especially during the Ladinian and Carnian. There are generally five second-order transgression–regression cycles recognized in the Triassic: these encompass the Lower Triassic and Middle Triassic and the Carnian, Norian, and Rhaetian.

Glacio-eustasy could not have driven Triassic sea-level change, so its underlying cause must be tectonism. Indeed, a large amount of regional Triassic tectonism has been invoked to explain sea-level changes in the western Tethys and the Arctic Sverdrup Basin. Triassic sequence boundaries caused by local tectonism or global eustasy show a remarkable degree of synchrony across Pangaea, and 12 high-order boundaries of global extent have been identified and attributed to episodic, major plate tectonic reorganizations.

Climate

Triassic climates marked the transition from the Late Palaeozoic ice house to the Mid-Late Mesozoic greenhouse (Figure 5). During the Triassic, there were no glacial ages, and there is no evidence of pack ice in the boreal or austral realms. The Triassic was thus a time of increased warmth with relatively wide subtropical dry (desert) belts at 10° to 30° latitude, as attested to by the broad latitudinal distribution of

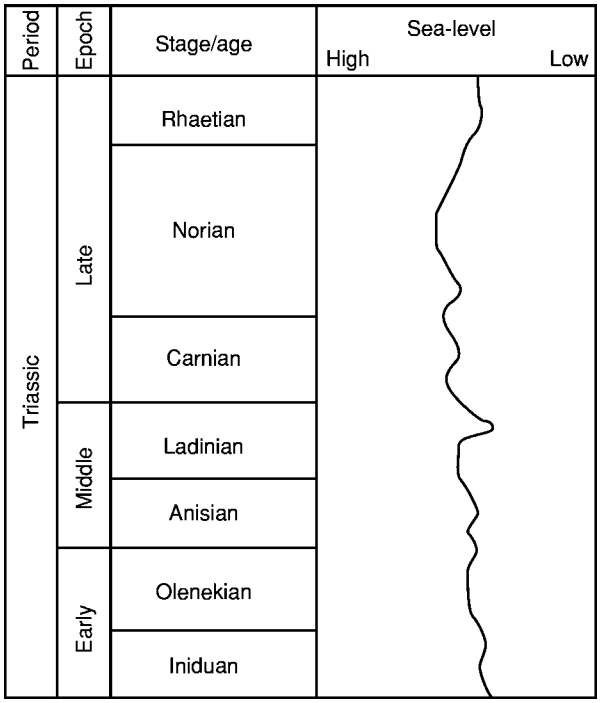


Figure 4 Triassic sea level curve. After Lucas (2000) The epi continental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

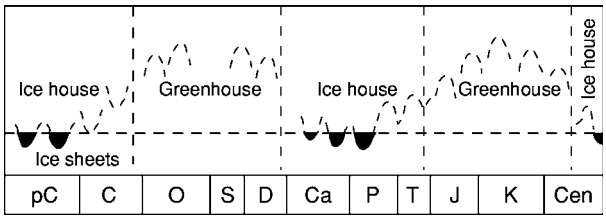


Figure 5 Triassic climate was a transition between the Late Palaeozoic ice house and the Late Mesozoic greenhouse. pC, Precambrian; C, Cambrian; O, Ordovician; S, Silurian; D, Devonian; Ca, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Cen, Cenozoic. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

Triassic evaporites. There was also strong east-west climatic asymmetry across Pangaea, with eastern Pangaea (at least between latitudes 40° S and 40° N) being relatively warmer and wetter because of the presence of Tethys and the absence of an Atlantic Ocean to facilitate oceanic heat exchange.

With the Pangaean landmass centered near the equator during the Triassic, and a prominent Tethyan bight, climate models suggest that seasonality was monsoonal. Hence, there were only two seasons, wet and dry. The abundant rainfall was concentrated in the summer months, and there was little annual temperature fluctuation. During the northern hemisphere summer, the northern landmass would have been relatively hot, whereas the southern land mass would have been relatively cool. Moisture from Tethys would have been pulled into the northern hemisphere low-pressure cell, producing extensive rains, whereas the southern hemisphere high-pressure cell would have remained relatively dry. During the southern hemisphere summer, this process would have occurred in reverse. Thus, seasonality across Triassic Pangaea would have been alternating hemisphere-wide wet and dry seasons. The warm and highly seasonal climates (wet–dry) of Triassic Pangaea are reflected in its biota. The Triassic saw an increase in the diversity of gymnosperms, particularly of xeromorphic scale-leaved conifers and seed ferns and cycadophytes with thick cuticles. Similarly, during the Triassic, in the evolution of reptiles, more water-efficient (putative uric-acid-excreting) diapsids diversified at the expense of less water-efficient (probably urea-excreting) synapsids.

Extinctions

The Permian ended with the greatest biotic extinction of Phanerozoic history (here termed the PTB biotic crisis) (*see* **Palaeozoic:** End Permian Extinctions). This extinction is best documented in the marine realm (Figure 6), where it is estimated that ~90% of the species, and more than half of the families of shelled marine invertebrates, became extinct. The magnitude and synchrony of the terrestrial extinction are much less clear. The Triassic records the recovery of the global biota from this massive extinction. The period also bore witness to further marine extinctions within the Late Triassic, and was terminated by a series of Late Triassic marine and non-marine extinctions.

The cause of the PTB biotic crisis remains uncertain. Some workers have identified a complex and interrelated group of terrestrial events as a possible cause: (1) major marine regression that reduced marine shelfal habitat areas and increased climatic variability, (2) eruption of Siberian flood basalts, (3) release of gas hydrates and erosion/oxidation of marine carbon due to the regression, and (4) elevated atmospheric CO_2 due to all of these phenomena resulting in ocean anoxia and global warming. During the Carnian, there was a substantial marine extinction of many kinds of conodonts, ammonoids, bivalves, echinoids, and reefal organisms, although the impact on land was less obvious, with evolutionary turnover occurring throughout the Late Carnian and Early Norian. A further extinction has been identified at the Norian–Rhaetian boundary with the disappearance of the ubiquitous flat clam *Monotis*. This, in fact, was part of a series of Late Triassic extinctions that included the disappearance of the conodonts, near extinction of the ammonites, decimation of about half of the marine bivalves, and collapse of the reef ecosystem. On land, there were also profound extinctions of tetrapods between the end of the Triassic and sometime in the Middle Early Jurassic (Sinemurian), but it has been difficult to establish the exact timing. A major carbon isotope anomaly has been identified in both marine and terrestrial environments at the end of the Triassic. This major perturbation in the global carbon cycle has been variously linked to a significant fall in sea-level, extraterrestrial impact, flood-basalt volcanism, and/or methane release.

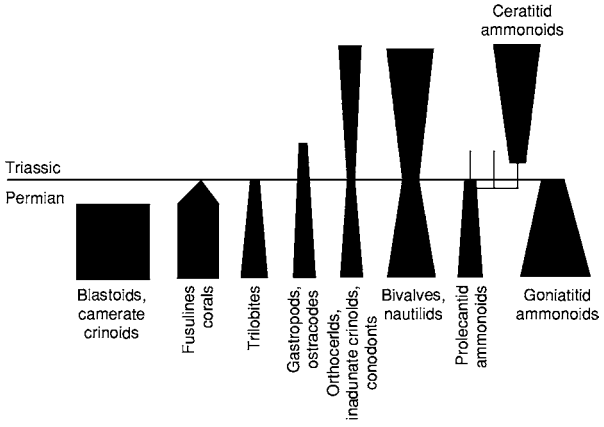


Figure 6 Characteristic extinction/diversity patterns of marine invertebrates across the Permian Triassic boundary. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

Flora

During the Permian and Triassic, there was a complex and prolonged replacement of the palaeophytic flora by the mesophytic flora. This was the global change from pteriodophyte-dominated floras of the Palaeozoic to the gymnosperm-dominated floras that characterized much of the Mesozoic. Thus, the

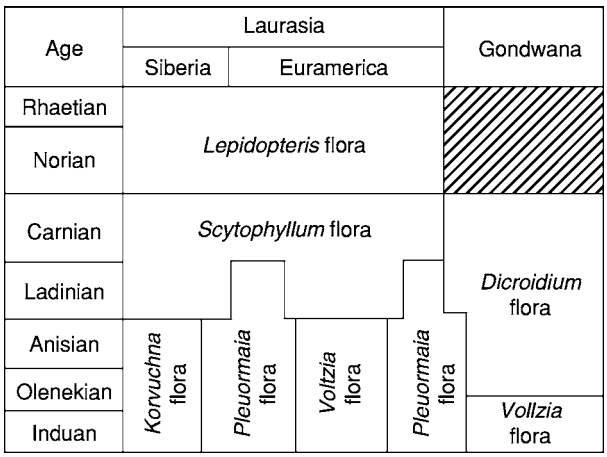


Figure 7 Triassic floral provinces and floras. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

arborescent lycopods and sphenopsids gave way to Triassic floras dominated by seed ferns, ginkgophytes, cycads, cycadeoids, and conifers. Distinct Gondwanan and Laurasian floras can be recognized, and within Laurasia two or three provinces are recognized – more boreal Siberian and more equatorial Euramerican provinces (Figure 7). However, the endemism of these floral provinces was not great. Triassic Laurasian floras were dominated by primitive conifers, ferns, cycads, bennettitaleans, and sphenopsids. Conifers were the dominant large trees, whereas the other plant types formed the understory. In coastal settings, stands of the lycopsid *Pleuromeia* were dominant.

Gondwanan floras of the Triassic were dominated by a wide range of seed ferns, especially the genus *Dicroidium*. These floras were generally composed of only a few (no more than 10) genera. *Dicroidium* was dominant in a variety of vegetation types, from heath to broad-leaved forest to dry woodland. Other important elements of Gondwanan floras were conifers and some Laurasian groups of cycadaleans and ginkgoes. Near the end of the Triassic, the *Dicroidium* flora declined and was replaced by a cosmopolitan conifer–benettitalean flora. From the end of the Permian until the Middle Triassic, there is a global coal discontinuity – there are no Early Triassic coal beds. This has been attributed to either an extinction of peat-forming plants at the PTB or to unfavourable tectonic conditions for coal preservation, though unfavorable climatic conditions for coal formation and preservation may also have been a factor.

Shelled Marine Invertebrates

Late Palaeozoic seas were dominated by pelmatozoans, brachiopods, and bryozoans, but molluscs

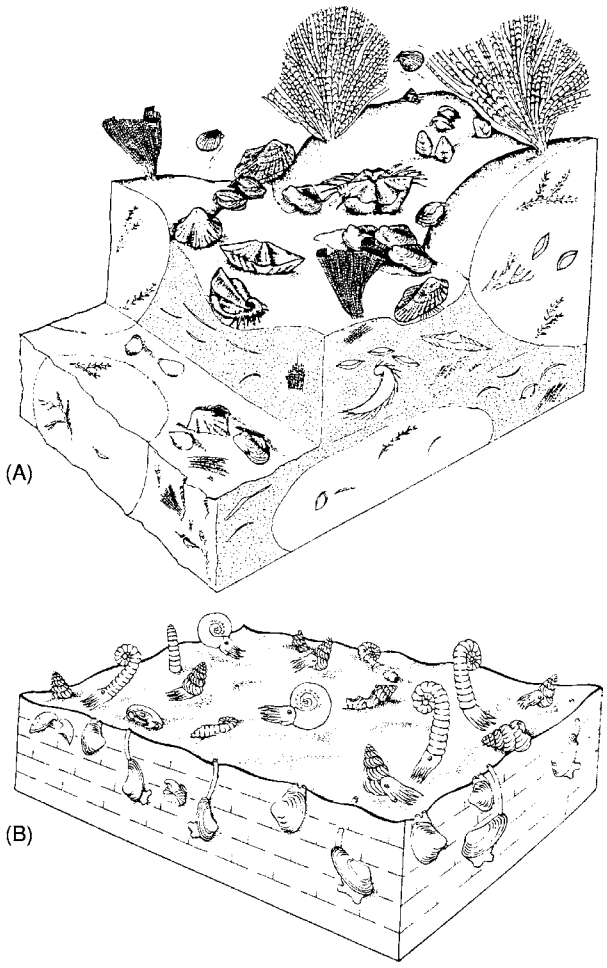


Figure 8 Reconstruction of (A) a characteristic Permian sea floor dominated by brachiopods, bryozoans, and crinoids and (B) a Triassic seafloor dominated by molluscs. After Lucas (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

dominated the Triassic seas (Figure 8). Most prominent of the Triassic molluscs were ammonoid cephalopods their rapid diversification during the Triassic provides a fossil record by which Triassic time has long been measured. Most Triassic ammonoids were ceratitidans with relatively simple suture lines. These were descended from only two ammonoid stocks that survived the PTB crisis: the otoceratids and the xenodiscids. Triassic ammonoid genera define three broad marine palaeobiogeographic provinces – Tethyan, Boreal, and Notal, the last of which is not well differentiated. The ammonoid palaeobiogeography of Triassic Panthalassa was complex and remains little understood. Triassic nautiloid cephalopods appear to have undergone relatively little change at the PTB, but reached great diversity in the Triassic, only to suffer an extensive (but not complete) extinction near the end of the period.

Bivalves were common Triassic molluscs and underwent a substantial diversification. Earliest Triassic assemblages are dominated by epifaunal pteriomorphs and detritus-feeding nuculoids, and they are very abundant as fossils. The Middle–Late Triassic saw a diversification of arcoid, mytiloid, trigonioid, and veneroid genera. The thin-shelled bivalves (so-called flat clams) *Claraia, Daonella, Halobia,* and *Monotis* are characteristic Triassic forms widely used in biostratigraphy. In contrast to ammonoids and bivalves, Triassic gastropods are relatively uncommon and not particularly diverse. A well-described Early Triassic (Smithian) gastropod assemblage from the western United States contains many genera that are also known from the Permian. Younger Triassic gastropod faunas are more diverse, but still contain numerous Permian holdover genera. The major Mesozoic change in gastropods took place after the end of the Triassic.

Brachiopods, bryozoans, and crinoids did not suffer total extinction at the PTB, although their numbers were greatly reduced (**Figure 6**). They were relatively minor, but persistent, components of Triassic marine faunas. More interesting is the distribution of corals and other reef-building organisms, which are virtually unknown in the Early Triassic (an exception are basal Triassic *Renalcis* biostromes in south China). In Middle Triassic time, Permian-type reef communities were re-established by *Tubiphytes*, bryozoans, calcisponges, and calcareous algae. The Carnian–Norian marine extinction was followed by a rapid turnover of the reef-building organisms, so that Norian reefs were characterized by abundant scleractinian corals, probably a result of the evolution of coral–zooxanthellae symbiosis. This presaged the extensive radiation of, and reef building by, corals that typified the Jurassic.

Insects

There was a major turnover in insect orders during the PTB biotic crisis, followed by a Triassic adaptive radiation, especially of beetles and cockroaches. A review of the Gondwanan Triassic record of plants and insects supports the concept of a co-evolution that led to the establishment of most modern insect orders by the end of the period.

Fishes

Fishes underwent a significant diversification during the Triassic. This is particularly evident in the appearance of new kinds of primitive actinopterygians, lungfishes, hybodontid sharks, and coelacanths. The extent of the extinction of fishes at the PTB biotic crisis is uncertain, but it appears to have been more significant within the marine realm. Conodonts, regarded by many now as a primitive fish group, were relatively unaffected by the end-Permian extinction. For them, the major biotic turnover occurred at the end of the Griesbachian, and this was followed by an explosive radiation at the start of the Olenekian. By the end of the Early Triassic, stocks had dwindled and there was a paucity of genera for the remainder of the period. The conodonts, nevertheless, continued to evolve rapidly, and their tooth-like elements now provide very useful biostratigraphic markers. After a long record, stretching throughout the Palaeozoic, conodonts became extinct at the end of the Triassic.

Tetrapods

Tetrapod vertebrates dominated Triassic landscapes and underwent at least two successive evolutionary radiations and extinctions during the period. Early Triassic tetrapod faunas were very similar to those of the Late Permian in being dominated by a relatively low diversity of dicynodont therapsids and capitosauroid/trematosaurid temnospondyls. Most notable is the dicynodont *Lystrosaurus,* the broad geographic distribution of which has provided classic evidence of the integrity of Triassic Pangaea (*Lystrosaurus* fossils have been found in Antarctica, South Africa, India, China, and Russia).

Middle Triassic tetrapod faunas remained dicynodont and temnospondyl dominated. However, by this time, the shift toward archosaur domination of the terrestrial tetrapod fauna had begun. By Late Triassic time, dicynodonts were rare, and the temnospondyls were greatly reduced in diversity. Instead, archosaurs were the most abundant terrestrial tetrapods. It was at this time that new groups of tetrapods appeared – turtles, dinosaurs (**Figure 9**), crocodiles, pterosaurs, and mammals – making the Late Triassic one of the most significant junctures in the history of vertebrate life. Indeed, it is fair to say that during the Late Triassic, the Mesozoic tetrapod fauna was born.

Marine reptiles also had their highest diversity during the Triassic. These reptiles were huhpehsuchians, nothosaurs, thallatosaurs, and placodonts, groups for which the entire diversification period was confined to the Triassic, and ichthyosaurs and plesiosaurs, groups that became prominent marine predators throughout much of the Jurassic and Cretaceous. There appears to have been a substantial extinction of marine reptiles (loss of 64% of families) at about the Middle–Late Triassic boundary. Apparently, the Reptilia successfully and explosively invaded the marine realm after the PTB crisis, but their diversity diminished rapidly, possibly due to

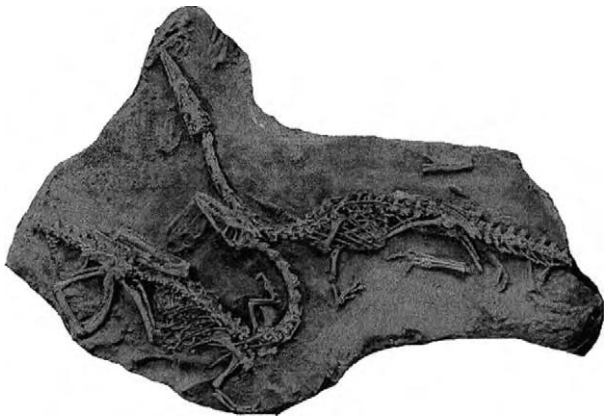


Figure 9 Dinosaurs appeared during the Triassic. These 3 m long skeletons of the Late Triassic theropod *Coelophysis* are from Ghost Ranch in northern New Mexico.

the overall Late Triassic lowering of the sea, which reduced the epicontinental seaways in which most of the marine reptiles lived.

Glossary

compressional tectonics Deformation of the crust through compression (pushing together).

cycadophytes A group of gymnosperm plants with compound leaves, including the cycads.

diapsids The lizard-like and ruling reptiles, including lizards, snakes, crocodiles, and dinosaurs.

extensional tectonism Deformation of the crust through extension (pulling apart).

fore-arc The region between an island arc and an oceanic trench.

foreland The region in front of a deformed area of the crust.

freeboard The difference between mean sea-level and mean continental altitude.

graben A block of crust dropped down along faults relative to blocks on either side.

gymnosperm A vascular plant with seeds that are not covered by an ovary, such as conifers.

mesophytic The time of intermediate land plants, approximately equivalent to the Middle Triassic, Jurassic, and Cretaceous.

palaeophytic The time of ancient land plants, approximately equivalent to the Palaeozoic and Early Triassic.

pteridophyte A fern-like division of vascular plants that reproduce by spores.

retroarc foreland basin A zone of thickened sediment (basin) and extensional tectonism behind an island arc, floored by continental crust.

synapsids Mammal-like reptiles, including the ancestors of mammals.

transpressional tectonics Deformation of the crust by a combination of strike-slip (horizontal) motion and oblique compression.

trough An elongate depression in the crust with gently sloping borders.

xeromorphic scale-leaved conifers A group of conifers with thick, scaly leaves that retain water and thus allow the plants to live in relatively dry climates.

See Also

Atmosphere Evolution. **Fossil Invertebrates:** Molluscs Overview. **Mesozoic:** Jurassic. **Microfossils:** Conodonts. **Palaeoclimates**. **Palaeozoic:** Permian; End Permian Extinctions. **Pangaea**. **Time Scale**.

Further Reading

Alberti F von (1834) *Beitrag zu einer Monographie des Bunten Sandsteins, Muschelkalks und Keupers, und die Verbindung dieser Gebilde zu einer Formation*. Stuttgart: Cotta.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Callaway JM and Nicholls EL (eds.) (1997) *Ancient Marine Reptiles*. San Diego: Academic Press.

Dobruskina IA (1994) Triassic floras of Eurasia. *Österreichische Akademie Wissenschaften Schriftenreihe Erdwissen Kommission* 10: 1–422.

Embry AF (1988) Triassic sea level changes: Evidence from the Canadian Arctic archipelago. *SEPM Special Publication* 42: 249–259.

Embry AF (1997) Global sequence boundaries of the Triassic and their identification in the western Canada sedimentary basin. *Bulletin Canadian Petroleum Geology* 45: 415–433.

Erwin DH (1993) *The Great Paleozoic Crisis: Life and Death in the Permian*. New York: Columbia University Press.

Kummel B (1979) Triassic. In: *Treatise on Invertebrate Paleontology, Part A, Introduction. Fossilization (Taphonomy) Biogeography and Biostratigraphy*, pp. 351–389. Lawrence, KS: Geological Society of America and University of Kansas Press.

Lucas SG (1998) Global Triassic tetrapod biostratigraphy and biochronology. *Palaeogeography, Palaeoclimatology, Palaeoecology* 143: 347–384.

Lucas SG (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7–8: 475–496.

Sherlock RL (1948) *The Permo-Triassic Formations*. London: Hutchinson's Scientific and Technical Publ.

Tozer ET (1984) The Trias and its ammonoids: the evolution of a time scale. *Geological Survey Canada, Miscellaneous Report 35*. Ottawa: Geological Survey Canada.

Yin H (ed.) (1996) *The Palaeozoic–Mesozoic Boundary. Candidates of the Global Stratotype Section and Point of the Permian–Triassic Boundary*. Wuhan, China: University of Geosciences Press.

Ziegler PA (1989) *Evolution of Laurussia*. Dordrecht: Kluwer Academic Publ.

Jurassic

K N Page, University of Plymouth, Plymouth, UK

Introduction

The Jurassic System is the second of the three systems comprising the Mesozoic Era. It takes its name from the Jura Mountains of eastern France and Switzerland, where, as long ago as 1795, its distinctive character was recognized, although it was not until 1829 that Alexander Brongniart first used the term 'Jurassique'. The stratigraphical and palaeontological meaning of the system was first clearly defined, however, by Alcide d'Orbigny's publication of his classic *Palaeontologie Française, terrains Jurassique* from 1842 to 1849.This work established a system of remarkably modern-looking stages, most of which are still in use today.

D'Orbigny's stages or '*étages*' were designed to be of worldwide use and were based on the assumption that periodic mass extinctions followed by the rapid re-establishment of new faunas characterized stage boundaries. The existence of a particular 'fauna' therefore correlated rocks belonging to a specific stage. Later research showed that these patterns are little more than an artefact of environmental changes in north-western Europe at the levels taken to represent stage boundaries by d'Orbigny, with faunal changes being more ecological than catastrophic.

Each of d'Orbigny's stages included a sequence of fossil 'zones', representing the general stratigraphic ranges of specific taxa. This term was further refined by Albert Oppel who, from 1856 to 1858 in his *Die Juraformation Englands, Frankreichs und des Südwestlichen Deutschlands*, developed a sequence of such divisions for the entire Jurassic System, crucially using the units in the sense of time divisions. Oppel used a very similar sequence of '*grupen*' or '*etagen*' to d'Orbigny, and these form the basis of today's Jurassic System.

Jurassic Stratigraphy and Chronology

Chronostratigraphy: Stages, Standard Zones, Subzones, and Horizons

Chronostratigraphy aims to produce rigorous definitions for named subdivisions of geological time, using actual rock units as standards for reference. This approach includes the designation of a Global Stratotype Sections and Points (GSSP), through international agreement, for the base of each geological time period (or 'System') and each of their component stages. The sequence of stages recognized in the Jurassic and their actual or candidate GSSPs are shown in Table 1. Below the level of the stage, subdivisions at the level of chronozones and ultimately zonules can be used. In the Jurassic, the frequent abundance of ammonoid cephalopod molluscs in marine sequences and their wide geographical distribution have led to their use for correlating sequences of 'standard zones'. Such units have, from the days of Oppel, been used in the same sense as the modern concept of chronozones, although they are still frequently confused with biozones, where the use of fossils in correlation has no

Table 1 Jurassic stages and their definitions

Stage	*Historical type locality*	*GSSP or candidate GSSPs*
Tithonian (152 145 Ma)	South Eastern France, after the Goddess 'Tithon'	Mont Crussol, Ardèche/Canjuers, Provence, France
Kimmeridgian (155 152 Ma)	Kimmeridge Bay, Dorset, England	Staffin Bay, Isle of Skye, Scotland; Mont Crussol, Ardèche, France
Oxfordian (157 155 Ma)	Oxford, Southern Central England	Savournon/Thoux, Provence, France; Weymouth, Dorset, England
Callovian (160 157 Ma)	Kellaways, Wiltshire, England	Albstadt Pfeffingen, Wurttemberg, southern Germany
Bathonian (166 160 Ma)	Bath, South West England	Cabo Mondego, Portugual; Bas Auran, Provence, France
Bajocian (174 166 Ma)	Bayeux, Normandy, France	Cabo Mondego, Portugual (GSSP)
Aalenian (178 174 Ma)	Aalen, Wurttemberg, Germany	Fuentalsaz, Iberian Cordillera, Spain (GSSP)
Toarcian (183 178 Ma)	Thouars, Deux Sèvres, France	Peniche, Portugual
Pliensbachian (192 183 Ma)	Pliensbach, Wurttemberg, Germany	Robin Hood's Bay, North Yorkshire, England
Sinemurian (197 192 Ma)	Semur En Auxois, Burgundy, France	East Quantoxhead, Somerset, England (GSSP)
Hettangian (200 197 Ma)	Hettange, Lorraine, France	St Audries Bay, Somerset, England; Muller Canyon, Nevada, USA; Chilingote, northern Peru; Kunga Island, BC, Canada

explicit meaning for geological time. As such, ammonite chronozones are always treated as simple subdivisions of chronostratigraphical stages, a usage that would be impossible if they were biozones.

Most Jurassic ammonoid chronozones are divided into subchronozones, largely for historical reasons and to achieve a degree of nomenclatural stability at the chronozone level. Classically, ammonoid zones have offered a very precise and detailed relative time-scale with which to date Jurassic geological sequences and events. Over the last 30 years or so, however, even smaller subdivisions or units have been established, in particular in Europe, which are known collectively as 'horizons'. There are conceptually two types of sub-subzonal 'horizon' in use in Jurassic ammonoid stratigraphy – first, zonules, which are the smallest component of a chronostratigraphical hierarchy, and, second, biohorizons, which can be used in the sense of faunal 'events'.

Both classes of intra-subzonal unit provide a high-resolution time-scale for the Jurassic, which is probably unique for pre-Quaternary sequences, with subdivisions averaging less than 200 000 years, and even potentially as little as 120 000 years in the Lower Jurassic. Their practical value is potentially immense and only just beginning to be realized: one of the most dramatic demonstrations of this is the tracing of the basal Callovian *Kepplerites keppleri* Biohorizon circumglobally in the northern hemisphere from Europe to Russia to Japan to Alaska to Arctic Canada to Greenland and back to Europe again, thereby precisely correlating widely spaced regions with a single faunal 'event' that potentially lasted less than 200 000 years.

Other fossil groups, especially microfossils, have been used to construct true biozonal schemes for the marine Jurassic in particular, but the resolution of these schemes is usually inferior to that of the ammonite scale, with a single microfossil biozone spanning several ammonite chronozones. Indeed, the latter scale is typically used as a 'standard' against which biozonal schemes are correlated. Nevertheless, where ammonites are very rare or absent (e.g. in non-marine or restricted-marine facies or in boreholes), they can be very useful. Continued research can only improve these correlations, and detailed sampling of sequences of dinoflagellate floras, for instance, has begun to demonstrate that they may even ultimately provide a correlative scale at least as detailed as that provided by ammonite chronozones.

Other Stratigraphical Methods in the Jurassic

Considerable advances have been made in recent years in establishing chemostratigraphy as a potential tool for correlating Jurassic sequences, in particular for sequences where high-resolution biostratigraphical tools are rare or lacking (e.g. continental sequences), for linking terrestrial with marine systems, and for linking different faunal provinces where detailed biochronostratigraphical correlations have yet to be established. A variety of isotope systems have been assessed, in particular carbon, oxygen, and strontium, but also osmium, sulphur, and nitrogen. Trace elements, such as magnesium and manganese, have also been shown to have some potential.

Initial results are promising but still need to be adequately integrated with established high-resolution biostratigraphical schemes. Some attempts to infer correlations solely on the basis of geochemical data are less persuasive when independently assessed using, for instance, ammonite zonal and biohorizonal schemes.

Similarly, magnetostratigraphy has the potential to provide a globally useful framework for correlating existing regional biostratigraphical and chronostratigraphical schemes, but it is essential that an accurate calibration is achieved between the two different 'time-scales' if the results are to be meaningful and applicable.

The economic significance of Jurassic sequences, in particular as sources of oil in Europe, has led to a massive development of sequence stratigraphical analysis, and, for specific depositional basins, the tool has a high value for the correlation of geophysical logs in particular.

Jurassic Geochronology

Absolute time-scales for the Jurassic have been somewhat elusive, owing to the rarity of reliably dated and correlated calibration points, with as few as five such being used until relatively recently. As a result various methods have been used to estimate, for instance, stage duration, in particular the persistent fantasy that ammonite zones average 1 Ma in duration. That such zones can contain as few as one or more than 19 correlatable subunits (as biohorizons or zonules) emphasises that different zones are likely to be of widely differing durations. Such assumptions are no longer tenable.

A contemporary and more reliable time-scale has been developed in Canada, however, which uses at least 50 calibration points, mainly in biostratigraphically well-constrained volcanic and volcaniclastic rocks. Nevertheless, there remain problems with the dating of the later Jurassic, from the Callovian to the Tithonian, as reliable data remain sparse. In particular, analysis of the number of biostratigraphical and chronostratigraphical units recognizable for each of the stages in this interval in Europe suggests that the Callovian and Oxfordian stages are too short and

the Tithonian is too long. This dating problem may be compounded by problems of interprovincial correlation in the Late Jurassic, when ammonoid provincialism was strong (Figure 1).

The Jurassic World

Continents

The Early Jurassic inherited the pole-to-pole massive supercontinent of Pangaea (*see* **Pangaea**). The opening of the North Atlantic began the process of dismantling this huge landmass, initially with South America moving away from North America to form the marine 'Hispanic Corridor'. The latter linked the previously separated East Pacific regions and the west Tethyan regions of Europe. There is circumstantial evidence from faunal migrations that this passage may have been intermittently open during the Lower Jurassic, providing a migration passage for shallow-marine faunas, but it was not until the later Middle Jurassic that it became a more open passage.

During the Middle Jurassic, oceanic crust began to form in the Atlantic, as rifting began to separate Europe from Greenland and North America. Despite this gradual opening up, however, the general layout of the continents maintained strong east–west and north–south physical barriers to marine faunal migrations throughout the period, leading to well developed bioprovincialism (Figure 1) that did not finally break down until the mid-Cretaceous.

Climate

Greenhouse effects dominate Jurassic climates worldwide. Lithological indicators, such as coals, evaporites, and aeolian sands, show a general symmetry of climate zones about the palaeoequator. Floral data

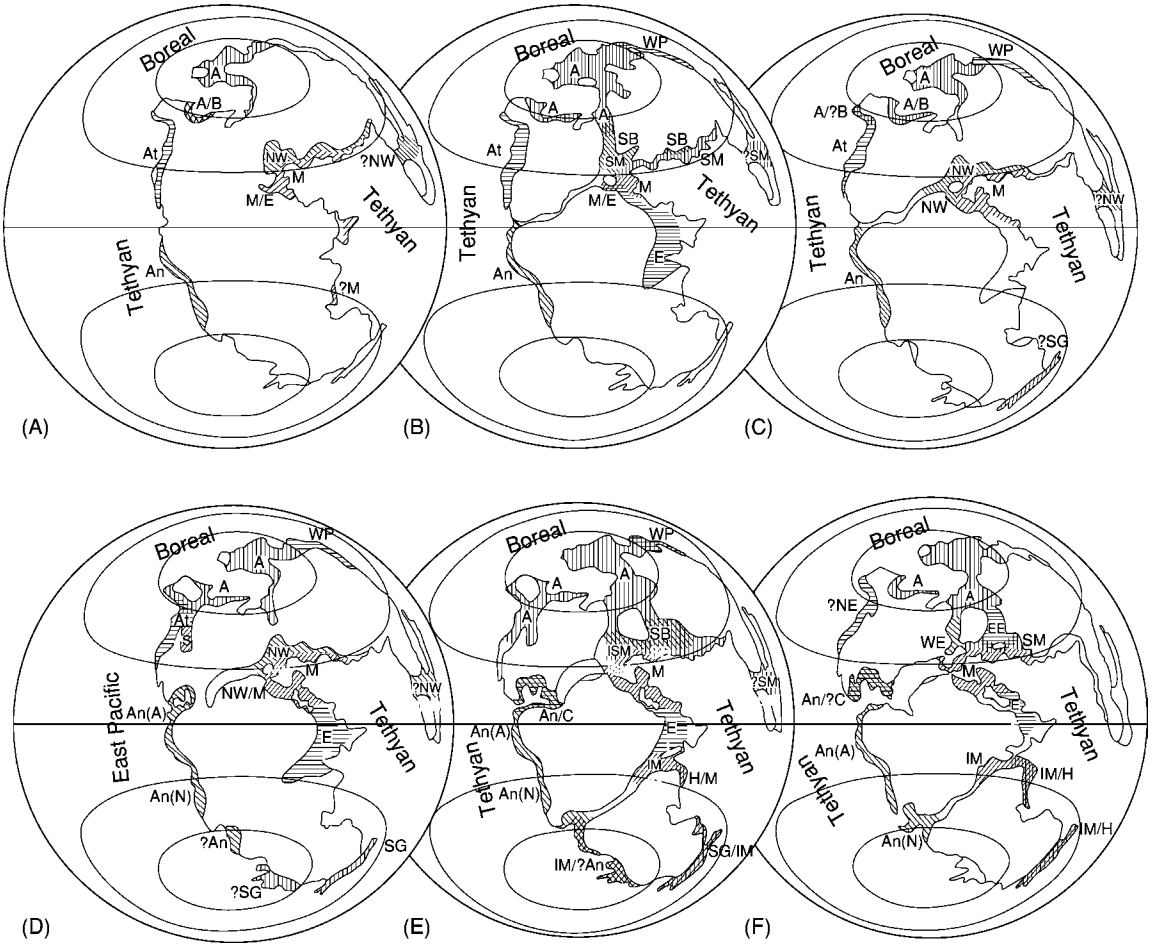


Figure 1 The evolution of the Jurassic world. (A) Hettangian Lower Pliensbachian, (B) Upper Pliensbachian Toarcian, (C) Aalenian Lower Bajocian, (D) Upper Bajocian early Lower Callovian, (E) Late Lower Callovian Kimmeridgian, and (F) Tithonian Valanginian. Ammonoid provinces as follows: Tethyan Realm M, Mediterranean; NW, North western European; SM, Sub Mediterranean; H, Himalayan; E, Ethiopian; IM, Indo Malgach; SG, Sula New Guinean; C, Cuban; WP, West Pacific. East Pacific (Late Bajocian Bathonian) or Tethyan Realm At, Athabascan; S, Shoshonean; An, Andean. Boreal Realm A, Arctic; SB, Sub Boreal; WE, West European; EE, East European; B, Bering. Cross hatching indicates where provincial barriers fluctuated.

suggest that the Jurassic world was one in which low latitudes were seasonally dry and were succeeded polewards in both hemispheres by desert, seasonally dry, warm temperate, and cool temperate biomes. Modelling studies have suggested that the tropical regions reached temperatures as high as 40°C during parts of the summer months, whilst winters were around 0–20°C. High-latitude oceans during the Jurassic were certainly significantly warmer than they are today and were generally ice-free.

Ocean palaeotemperature estimates for the European area range from 12°C to 30°C, again considerably warmer than today, with the Lower and Upper Jurassic appearing to be the warmest. The possible existence of Boreal and Austral realms, delineated by temperature, during parts of the Mesozoic, however, certainly lends support to the existence of cold conditions at the poles at certain times. General Circulation Model simulations of the Jurassic climate also reveal that high-latitude regions may have developed cold temperatures, providing adequate leeway to argue for either the presence or the absence of high-latitude ice. Only equivocal Jurassic tillites in Siberia have, however, been described. The Late Jurassic–early Cretaceous was characterized in many areas of the world by a sea-level lowstand, resulting in large semi-restricted epicontinental seas and widespread deposition of non-marine sediments. The latest Jurassic Purbeck group of the UK, which contains evidence of evaporite deposition, may be the local expression of the spread of an equatorial arid zone across Europe at this time.

A characteristic of Jurassic shelf sequences is a strong cyclicity, observed as sedimentary microrhythms, typically with an alternation of limestone and marl, although in clastic-dominated sequences this may manifest itself as an alternation between carbonate-rich and carbonate-poor mudrocks or even as coarser silty or sandy bands alternating with finer-grained claystones. The periodicity of these microrhythms is consistent with orbital forcing due to Milankovitch cyclicity (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)), with, for instance, the more carbonate-rich levels representing improved climates and consequently higher production by calcareous plankton. There has been much discussion about the potential of using microrhythms assignable to specific Milankovitch frequencies to improve or even calibrate Jurassic chronostratigraphy, although few attempts have been made to correlate results with high-resolution biostratigraphical or chronostratigraphical schemes, to ensure that local effects such as gaps in the record due to local processes (e.g. tectonic) do not compromise conclusions.

Key Events

The Early Jurassic saw a recovery from the devastating Late Triassic mass extinction, which was linked to the destruction of shelf seas and to climactic changes associated with the final assembly of the supercontinent of Pangaea. Radiations of many taxa followed, and extinction events that have been reported during the later Jurassic, for instance across the Pliensbachian–Toarcian and Callovian–Oxfordian boundaries, are probably consequences of no more than local or regional facies changes.

The Lower Toarcian in north-western Europe is characterized by a well-known anoxic event, whose causes and extent have been much discussed. Much of the latter depends on the dating of the event, which has been claimed to be global in extent. Nevertheless, precise correlations have not always been established and, even in Europe, there are possibly either two 'events' or a single diachronous anoxic event. For example, the ammonite chronology suggests that levels with the highest organic carbon contents are slightly older in Mediterranean/reas than in north-western Europe. It has also been suggested that the $\delta^{13}C$ minimum at this general level is related to a massive release of methane from gas-hydrate deposits, although a lack of synchroneity across Europe would make this scenario less likely. Perhaps more compelling, however, is the relatively simple explanation that anoxia developed in the stratified waters of the relatively enclosed and circulation-restricted marine basins of north-western Europe, perhaps as a result of a regional rise in sea-level, as suggested by facies changes and faunal migrations. The $\delta^{13}C$ minimum can be explained by the recycling of ^{12}C-enriched carbon derived from the remineralization of organic matter. Hence the 'global' significance of this event beyond Europe remains to be proved conclusively.

Jurassic Faunas and Floras, Including Key Groups with Stratigraphical Value

Benthic Algae (Including Chlorophyta (Dasycladales), Charophyta, and Rhodophyta)

Dasycladean algae that produce calcareous structures are locally sufficiently common in peri-Tethyan shallow-water carbonate-platform deposits to be useful locally as biostratigraphical markers, especially in the virtual absence of other stratigraphically useful groups. Coralline rhodophytes are sometimes important in Jurassic reefs and bioherms, contributing to their bound structure and to associated bioclastic sediments.

In freshwater carbonate sequences, the calcified stems and reproductive bodies ('oogonia') of charophytes lend themselves to preservation as fossils, and hence they may be locally abundant and even stratigraphically useful from the mid- to the Late Jurassic.

Planktonic Algae (Including Prynesisophyta, Coccosphaeales (Coccoliths) and Pyrrhophyta (Dinoflagellates))

Jurassic calcareous nannofloras, including coccoliths, are still relatively poorly known, especially in the Early Jurassic, owing to their great sensitivity to diagenesis. Nevertheless, they can be locally so abundant as to be rock-forming, especially in some fine-grained mud-rich carbonate sequences, such as pelagic deposits, where diagenesis has not destroyed traces of their structure – indeed they may well be the main contributor of calcareous material to many such fine-grained Jurassic marine carbonates. As knowledge improves, their stratigraphical value will improve, especially following the resolution of problems related to provincialism, including an apparently marked diachroneity of first occurrence across different regions.

The cysts of dinoflagellates, representing the dormant phase of the life cycle of a marine planktonic alga, are frequently abundant in Jurassic marine and quasi-marine sequences, especially those where the sediments are relatively organic rich. As such they can be excellent stratigraphical guide fossils in borehole studies, in particular in petroleum exploration, and, although correlative units are currently stratigraphically relatively long, the cysts' great diversity from the Middle Jurassic onwards hints at the potential, at least locally, to achieve a resolution virtually equivalent to that of ammonoid chronozones.

Protists

The calcareous or agglutinated tests of benthic foraminifera are frequently common in Jurassic marine sediments. In basins and platforms dominated by argillaceous sediments these associations are often dominated by Nodosariidae, although on some carbonate platforms Textularina may be more significant. Lower Jurassic assemblages are typically of relatively low diversity, following the end-Triassic mass extinctions, but diversity increases markedly through the Middle Jurassic. Planktonic foraminifera (Globigerinina) first appear in the Early–Middle Jurassic and by the early Upper Jurassic may be abundant in some sequences – although understanding of their distribution and biostratigraphy is often reduced by lumping them together under the term 'protoglobigerinids'. The stratigraphical use of Jurassic foraminifera is relatively limited, however, owing to the relatively long ranges of many taxa, although the group can be locally important for recognizing divisions on the scale of stages.

Jurassic radiolaria are most characteristic of relatively deep-water siliceous rocks, such as cherts, formed below the carbonate compensation depth. In such areas they can be valuable stratigraphical tools.

Porifera (Sponges)

Sponges (*see* **Fossil Invertebrates:** Porifera) can be locally common in Jurassic carbonate rocks, especially from the Middle Jurassic. In the Upper Jurassic, in particular, Hyalospongea (siliceous sponges) can locally be very important as rock formers in some fine-grained carbonate-platform sequences bordering Tethys, including building spectacular biohermal structures.

Cnidaria, Scleractinia (Corals)

The origin of hermatypic (reef-building) corals with symbiotic algae, in the Triassic, paved the way for the development of true coral reefs in the Jurassic, with a bound framework formed by the corals themselves – although typically in association with calcareous algae (including rhodophytes) and sometimes bryozoans and sponges. Coral patch-reefs are most typical where suitable warm shallow-water environments are developed, mainly in the Middle and Upper Jurassic.

Brachiopoda and Bryozoa

Although the last spiriferid brachiopods persist into the Lower Jurassic, the articulate orders Terebratulida and Rhynconellida dominate normal-marine Jurassic brachiopod faunas. Locally, in shallow-marine carbonate deposits these groups can be a major component of shelly faunas, even outnumbering bivalves. Despite strong provincialism, this local abundance gives the group biostratigraphical use, and resolutions more or less equivalent to that of ammonoid zones can be achieved across some basins or sub-basins. The remarkable inarticulate survivor *Lingula* persists in more restricted environments, especially in shallow quasi-marine environments but also occasionally in deeper-water situations where organic-rich mudrocks exclude other brachiopod groups.

Encrusting and coralline bryozoans with calcified skeletons are locally common in shallow-marine Jurassic carbonate sediments; the order Cyclostomata is particularly important and tends to dominate.

Mollusca

Jurassic bivalves are often extremely abundant in a great range of water-lain sediments, from fully

marine, through brackish, to freshwater – each environment has a characteristic range of taxa. As their morphology is tightly constrained by their ecology, they have been extensively used for palaeoecological studies – a process aided by the fact that virtually all Jurassic bivalve orders survive today. Some groups, in particular Trigonoida and some Pteroida (including some pectenids and the well-known oyster *Gryphaea*), can also have limited, albeit somewhat crude, biostratigraphical value. More useful, however, especially in some extreme Boreal areas in the Late Jurassic, are thin-shelled pseudo-planktonic buchids, and these form the basis of a biozonation, for instance in Arctic Canada and Russia. Certain Jurassic bioclastic limestones may be composed largely of bivalves in varying degrees of fragmentation, and occasionally epifaunal groups such as oysters and lithiotids may form small bioherm-like structures.

Jurassic gastropods are dominated by epifaunal archaeogastropods and mesogastropods and are considerably less diverse than the neogastropod assemblages of the later Cretaceous and Tertiary. Very occasionally, however, groups such as nereneiids may be very abundant in certain restricted-marine environments and may consequently even have local biostratigraphical value.

At the end of the Triassic the Ammonoidea faced a massive crisis, and perhaps only two genera survived into the Jurassic, the early phylloceratid *Rhacophyllites* and its direct descendant, the first ammonitine ammonite, *Psiloceras*. A massive evolutionary explosion followed, and by the Middle Jurassic five ammonoid suborders can be recognized, with a sixth appearing right at the end of the period (*see* **Fossil Invertebrates:** Ammonites). Phylloceratine and lytoceratine ammonites remain morphologically very conservative throughout the Jurassic and are most abundant in areas influenced by open ocean – studies of shell strength indicate that they could live at greater water depths than the other Jurassic suborders.

Other groups, such as the Ammonitina, Haploceratina, and Perisphinctina, are much more morphologically varied and thrived in epicontinental seas. They frequently show distinctive geographical distribution patterns, reflecting ecological and physical controls on individuals and populations (**Figure 1**). Such patterns are characterized as biogeographical provinces, and the inevitable consequence of using ammonites for correlation purposes is that every province, almost by definition, will have a different scheme of standard zones. These differences inevitably make interprovincial correlations at zonal level, and especially at subzonal and horizon levels, difficult. Nevertheless, faunal links often exist, frequently facilitating remarkably good interprovincial correlations, and ammonites remain the preeminantly high-resolution stratigraphical tools in the Jurassic.

Early Jurassic distributions continue the basic pattern of the Late Triassic, with a clearly distinguishable Arctic Boreal Realm and a much broader Tethyan Realm to the south. By the Middle Jurassic, however, communication with the western margin of the Americas was so restricted that a separate East Pacific Realm developed. The establishment of direct Boreal–Tethys connections in Europe in the late Middle and Upper Jurassic provided more possibilities to exchange faunas between the northern Arctic Sea and the trans-equatorial Tethys Ocean. Latitudinal controls on faunal distribution are nevertheless still evident, and a complex series of faunal belts developed, ranging from high-latitude restricted Boreal faunas, through sub-Boreal to sub-Mediterranean and eventually low-latitude diverse Mediterranean Province faunas. Analogous transitions from Boreal to more Tethyan faunas are also present in East Pacific regions at this time, although latitudinal faunal belts are only occasionally discernible southwards from the Jurassic equator towards the Antarctic.

Belemnoidea, in contrast, are never as diverse as Ammonoidea, although they can still show analogous bioprovincial distributions, with distinct Boreal and Tethyan assemblages. Their robustness in a sedimentological context gives them some value as stratigraphical indicators, occasionally at a resolution equivalent to around two ammonite zones. Much more important, however, is their use in geochemical studies, especially for assessing changes in oceanic chemistry (particularly related to temperature and climatic change), using oxygen, carbon, and strontium isotopes.

Crustacea, Including Ostracoda

Jurassic decapod crustaceans, although generally seen as body fossils only under exceptional conditions (such as rapid burial in a phosphate-rich environment), were apparently major contributors to the bioturbation of shallow-marine aerated sediments. Generally only burrow traces are seen, however, including the branching form *Thalassinoides*, which was made by a callianassid shrimp.

In contrast, the calcareous carapaces of microscopic Ostracoda can be very important environmental indicators, as different genera and species can be characteristic of marine or brackish-water deposits. They can also have value as stratigraphical indicator fossils. Although resolution tends to be at the level of substage and interbasin provincialism is potentially

well developed, when combined with results from the Foraminifera obtained in the same sieved residues, the group can be very useful for correlation purposes.

Uniramia, Insecta

Beetles and dragonflies are typical components of Jurassic insect faunas although their fossil record is very patchy in the Jurassic, prior to the more abundant record from Cretaceous amber. Although obviously representing terrestrial faunas, Jurassic insects not infrequently turn up in fine-grained organic-rich marine sediments where land was not too distant.

Echinodermata

By the Jurassic, Crinoidea had become much less important as producers of bioclastic sediment in shallow-marine environments than earlier, for instance in the Late Palaeozoic. They remain locally common, however, and planktonic and pseudo-planktonic forms can be locally significant, the former including the Upper Jurassic *Saccocoma* and the latter attached to drift wood.

Mobile benthic echinoderms, such as ophiuroids and holothurians, may have been more abundant in some marine Jurassic environments than their macrofossil remains may suggest, as their spicules or fragmented remains can be common in microfossil residues. Echinoids, including both burrowing 'irregular' and epifaunal 'regular' forms, are, however, more conspicuous in some marine deposits where seabed oxygenation was relatively good, and the former, in particular, can even have some local stratigraphical value.

Chordata (Vertebrates)

Jurassic fish faunas are characterized by a dominance of bony fishes (Osteicthyes), in particular advanced actinopterygians ('ray-finned fishes'). Teleosts, in particular, were on the rise. Chondrosteans (which include today's sturgeons) were in decline, as were coelocanths and lungfishes, which are both generally rare in the Jurassic. Whilst showing a reduced diversity compared with the Late Palaeozoic, chondrichthyans (cartilaginous fishes, including sharks and rays) remain relatively common, with neoselachian sharks becoming dominant (*see* **Fossil Vertebrates:** Fish). Otoliths (fish ear-bones) are well known from some Jurassic sediments.

Marine large-reptile faunas are very characteristic of the Jurassic, with dolphin-like ichthyosaurs, long-necked plesiosaurs, and shorter-necked pliosaurs being particularly significant – the latter including the world's largest known carnivore, *Liopleurodon*,

Figure 2 A typical marine scene from the Lower Jurassic, showing nektonic plesiosaurs, actinopterygian fishes, ammonites, and belemnites, with a benthos dominated by bivalves with gastropods and crinoids. Reconstruction based on a locality in southern England. (Reproduced with permission from © University of Bristol, UK.)

at over 15 m in length (Figure 2). Marine crocodiles include the highly adapted geosaurs, with limbs modified as paddles, but their terrestrial cousins also radiated extensively. Other terrestrial reptiles included true lizards (squamates). Pterosaurs become increasingly important through the Jurassic, although they do not achieve the great diversity of form typical of their Cretaceous descendants.

Of all terrestrial animals in the Jurassic, the dinosaurs will always remain the most emotive (*see* **Fossil Vertebrates:** Dinosaurs). From humble beginnings in the Triassic, the group rapidly diversified to dominate all terrestrial vertebrate faunas worldwide. Although Early Jurassic assemblages are still dominated by Triassic groups, during the Middle Jurassic, many well-known groups appeared, including the giant herbivorous sauropods, large carnivorous theropods, and armoured stegosaurs and ankylosaurs (Figure 3).

By the Upper Jurassic, some small dinosaurs had developed feathers, and by the Kimmeridgian the transition to birds was well underway with the famous *Archaeopteryx*. Although mammals had arisen in the latest Triassic, they remained small and

Figure 3 A typical Middle Jurassic terrestrial scene, showing dinosaurs (including a ceratosaur theropod, a stegosaur, a fabro saurid ornithopod, and the sauropod *Cetiosaurus*), rhamphorynchid pterosaurs, a tritylodont mammal like reptile (*Stereognathus*), a docodont mammal, and, associated with the pond, a goniopholid crocodile, a salamander, a disoglossid frog, and an actinopterygian fish. The vegetation includes horsetails, ferns, cycads, seed ferns, and conifers. Reconstruction based on faunal remains from a single locality in southern England. (Reproduced with permission from © University of Bristol, UK.)

relatively insignificant members of terrestrial faunas throughout the Jurassic.

Jurassic Terrestrial Floras

Jurassic floras are sporadically recorded where terrestrial sediments, especially those formed by rivers and in lakes, are well developed. Assemblages are typically dominated by ferns and various gymnosperm groups such as cycads, bennettites, caytonias, ginkgos, and conifers. Where conditions allowed the development of marsh vegetation, horsetails can also be abundant. Macrofloras are also well known, however, in certain shallow-marine sequences, although they are often somewhat different in composition from those of terrestrial (fluviolacustrine) facies, including more conifers and *Ptilophyllum* and *Sphenozamites* bennettites, with fewer ferns, ginkgos, cycads, and caytonias. Crucially, many families of extant ferns and conifers first appear in the fossil record in the Jurassic.

A characteristic feature of Jurassic floras is the relative lack of latitudinal variation, with northern 'temperate' floras in north-western Canada and Siberia being broadly similar to 'tropical' floras in the Early and Middle Jurassic of Europe, North America, Central Asia, and China. The main difference between these floras is a slight reduction in diversity in northern areas, with a dominance of ginkgos and lepidostrobaleans and with very few cheirolepidiacean conifers. Southern 'temperate' latitude floras, in contrast, were poor in ginkgos and dominated by the cheirolepidiacean conifers, with lepidostrobaleans being absent. These southern 'temperate' floras extended into very high near-polar latitudes in Antarctica, indicating the absence of polar ice.

The situation changed in the Late Jurassic owing to the development of arid conditions over much of the palaeotropics and parts of the southern mid-latitudes. The floras of these more arid regions became very low in diversity and were dominated by cheirolepidiaceans and a few ferns that were adapted to drier conditions (e.g., *Weichselia*). The floras of the northern high- and mid-latitudes and that part of the southern mid- and high-latitudes that escaped the aridification became much more diverse and more comparable with the floras of the palaeotropics in the Early and Middle Jurassic.

See Also

Earth: Orbital Variation (Including Milankovitch Cycles). **Fossil Invertebrates:** Ammonites; Porifera. **Fossil Vertebrates:** Fish; Dinosaurs. **Mesozoic:** Triassic; Cretaceous. **Pangaea**.

Further Reading

Arkell WJ (1956) *Jurassic Geology of the World*. London: Oliver & Boyd.

Behrensmeyer AK, Damuth JD, DiMichele WA, *et al.* (eds.) (1992) *Terrestrial Ecosystems Through Time.* Chicago: Chicago University Press.

Benton MJ (ed.) (1993) *The Fossil Record 2.* London: Chapman & Hall.

Cariou E and Hantzpergue P (eds.) (1994) *3ème Symposium International de Stratigraphie du Jurassi que (Poitiers France, 22 29 Septembre 1991.* Geobios Mémoire Spécial 17. Lyon: Université Claude Bernard.

Cariou E and Hantzpergue P (eds.) (1997) Biostratigraphie du Jurassique Ouest Européen et Méditerranéen. *Bulletin du Centre Recherche Elf Exploration et Production, Mémoire* 17: 15 24.

Hall RL and Smith PL (eds.) (2000) *Advances in Jurassic Research 2000: Proceedings of the Fifth International Symposium on the Jurassic System (Vancouver, Canada, August 12 25, 1998).* GeoResearch Forum 6. Zurich: Trans Tech Publications.

Ineson JR and Surlyk F (eds.) (2003) *The Jurassic of Denmark and Greenland.* Geological Survey of Denmark and Greenland, Bulletin 1. Copenhagen: Danmark og Grønlands Geologiske undersøgelse (GEUS).

Jenkyns HC, Jones CE, Gröcke DR, Hesselbo S, and Parkinson DN (2002) Chemostratigraphy of the Jurassic System: applications, limitations and implications for palaeoceanography. *Journal of the Geological Society* 159: 351 378.

Page KN (1995) Biohorizons and zonules: intra subzonal units in Jurassic ammonite stratigraphy. *Palaeontology* 38: 801 811.

Page KN (1996) Mesozoic ammonoids in space and time. In: Landman NH, Tanabe K, and Davis RA (eds.) *Ammo noid Paleobiology,* pp. 755 794. Topics in Geobiology 13. London: Plenum Press.

Parisi G (guest editor) (2004) The 6th International Sympo sium on the Jurassic System, 16 19 September 2002. *Rivista Italiana di Paleontolgia e stratigrafa* 110(1): 428. Milano.

Riccardi A (ed.) (1996) *Advances in Jurassic Research, Proceedings of the Fourth International Symposium on the Jurassic System (Mendoza, Argentina, 1995).* Geo Research Forum 1, 2. Zurich: Trans Tech Publications.

Westermann GEG (ed.) (1992) *The Jurassic of the Circum Pacific.* World and regional geology 3. Cambridge: Cambridge University Press.

Cretaceous

N MacLeod, The Natural History Museum, London, UK

Introduction

The Cretaceous System/Period is the last major subdivision of Mesozoic time. It was established by JJ d'Omalius d'Halloy in 1822 and was divided into Upper and Lower series/epochs by WD Conybeare and William Phillips that same year. Cretaceous rocks are currently assigned to 12 stage/age-level subdivisions, the combination of which represent an interval estimated to lie between 145.5 and 65.5 Ma. This is the longest single system/period of the Phanerozoic, and Cretaceous rocks are found on all continents. The Cretaceous is also the oldest system/period to be entirely represented in the ocean basins. Palaeogeographically, the Cretaceous represents the time during which Pangaea continued to fragment and continental plates began to move into their current positions. Prominent tectonic features of Cretaceous time include the rifting of Laurasia to form the northern and southern embayments of the proto-Atlantic Ocean, the joining of these embayments into a single north–south-trending Atlantic Ocean Basin, and the rifting of the southern supercontinent of Gondwana into South America, Antarctica, India, and Australia. During the Cretaceous, the sea-level stood high, flooding the interiors of most continental platforms and resulting in the establishment of marine chemical conditions that favoured the deposition of calcite. This, in turn, led to the widespread deposition, and subsequent preservation, of the rock type uniquely associated with the Cretaceous: chalk. As a result of this high sea-level, the Cretaceous was characterized by warm, equable, greenhouse-type climates over most areas, with temperatures that both exceeded and were more stable than those at present are. There was very little or no ice at the Cretaceous poles, and reefs, swamps, crocodiles, and even dinosaurs reached latitudes in the vicinity of – and, at times even greater than – 60°. Cretaceous atmospheric composition also differed from that of today, with higher levels of oxygen and carbon dioxide. The Cretaceous was a quiet time for magnetic reversals, but also a time of widespread volcanism. Biotically, the Cretaceous represents the culmination of many evolutionary–ecological trends begun in the Jurassic, including the diversification of many plant and animal lineages (e.g., diatoms, coccoliths, gymnosperms, angiosperms, foraminifera, ammonoids, molluscs, insects, dinosaurs, ichthyosaurs, plesiosaurs, mosasaurs, pterosaurs, and mammals). Finally, the Cretaceous includes two extinction events: an 'Aptian' event, which may be an analytic artefact, and the Cenomanian–Turonian event, in

which 27% of all marine genera disappeared as a result of the coincidental juxtaposition of sea-level rise, the (possibly volcanically accentuated) upwelling of oxygen-poor waters into the shallow chalk seas, and the effects of global cooling induced by more efficient marine circulation patterns. This period was also ended by a very large extinction event. (*see* **Mesozoic:** End Cretaceous Extinctions.)

Stratigraphy

Terrain Cretacé was the name originally given, in 1822, by the French geologist JJ d'Omalius d'Halloy, to a sequence of chalk beds, underlain by tufas, sands, and marls, that crop out in the structural basis of southern England, northern France, and Belgium. d'Omalius d'Halloy's usage followed that of William Smith's (*see* **Famous Geologists:** Smith) map, which had identified four sequences of strata between the Tertiary 'lower clay' and the (Jurassic) Portland stone. These were (from oldest to youngest) micaceous clay, also known as 'Brickearth', Greensand, brown or grey chalk, and white chalk. Later that same year, the English geologists WD Conybeare and William Phillips gathered these four units into two groups, the upper chalk facies and the lower, predominately clastic facies, and first used the term 'Cretaceous' to describe the entire stratigraphic package. Conybeare and Phillips' subdivision is reflected today in the fact that most time-scales recognize only two Cretaceous epochs (Upper and Lower), instead of the more typical three-fold subdivision of Upper, Middle, and Lower.

From 1840 to 1871, each Cretaceous epoch was further subdivided into six stages (Figure 1), based on rocks cropping out in France, Switzerland, and southern Holland. Cretaceous chronostratigraphy was originally based on molluscan biostratigraphy, especially the biostratigraphy of ammonites. These fossils were used to subdivide each stage into either threefold or twofold substage intervals. Beginning in the 1950s, however, Cretaceous biostratigraphic zonations were further refined through the use of microfossils, chiefly planktonic foraminifera (28 biozones) and calcareous nannoplankton (26 biozones). At least seven key radioisotopic tie points have been identified in Cretaceous sediments, providing good geochronometric control, especially for Upper Cretaceous stage boundaries.

The current base of the Cretaceous is undefined by a boundary stratotype, but is taken as being near the first occurrence of the ammonite *Berriasella jacobi*. The base of the Cenozoic Danian stage defines the top of the Cretaceous and was established in 1991 at the base of the 'boundary clay layer' associated

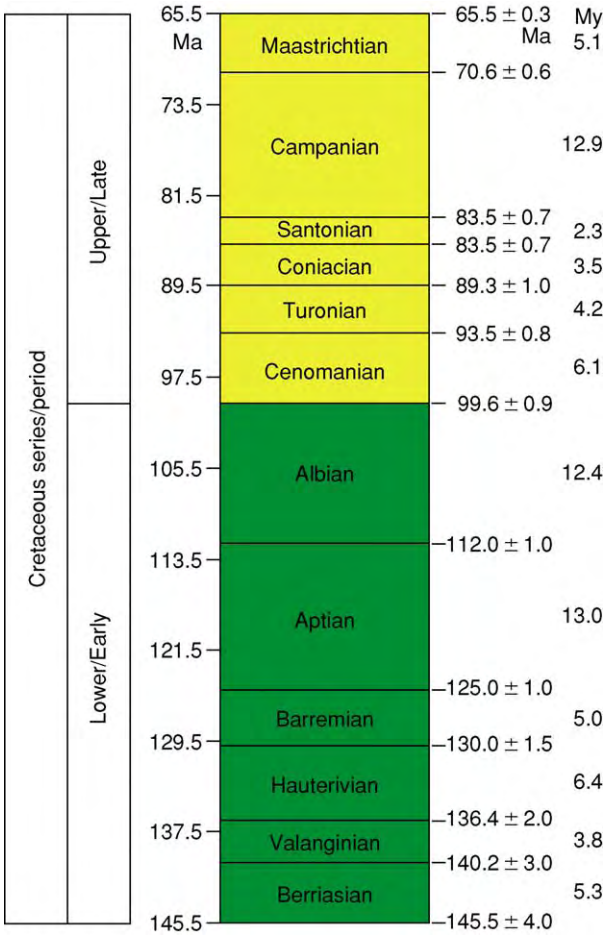


Figure 1 Chronostratigraphy and geochronometry of Cretaceous system/stages. Time scale based on that of Gradstein and Ogg (*see* **Time Scale**). Stage thicknesses scaled to reflect relative durations. Interpolated duration estimates given in millions of years (far right column).

with a geochemical iridium anomaly and a major extinction of planktonic foraminifera and calcareous nannoplankton in the stratotype section outside the town of El Kef, Tunisia. This stratotype definition is somewhat unusual in that it is not taken at a biostratigraphic datum and coincides with a local extinction horizon rather with than a first occurrence horizon. It is also unfortunate that the El Kef stratotype was destroyed after 1991 due to local farming practices. Studies are currently under way to re-establish the stratotype in the Tunisian type area at an outcrop less susceptible to such damage.

The Cretaceous system has the greatest duration of any Phanerozoic stratigraphic system (76.5 My). Its distinctive sediments, especially the characteristic Upper Cretaceous chalk facies, are present on all large continental platforms. The Cretaceous is also the first stratigraphic system to be well represented in the deep sea, owing to the fact that most pre-Cretaceous

sediments deposited in the deep-ocean basins have been subducted.

Palaeogeography and Tectonics

At the beginning of the Cretaceous (Figure 2A), the post-Permian breakup of Pangaea had progressed to the point at which the northern (Laurasian) continents of North America, North China, Siberia, and Eurasia had rifted away from Gondwana (Africa, South America, India, Antarctica, Australia), though the latter remained coherent. Gondwana was, in turn, separated from North America by a narrow Atlantic Seaway, and from North China by a broad Tethys Ocean. Surrounding this region of intense tectonic activity, the Pacific Ocean occupied fully half of the surface of the planet.

In terms of continental landmasses, the primary Cretaceous event was the rifting of Gondwana. Near the Jurassic–Cretaceous boundary (~140 Ma), a large rift between Africa and India formed and propagated south until, by 150 Ma, Gondwana has

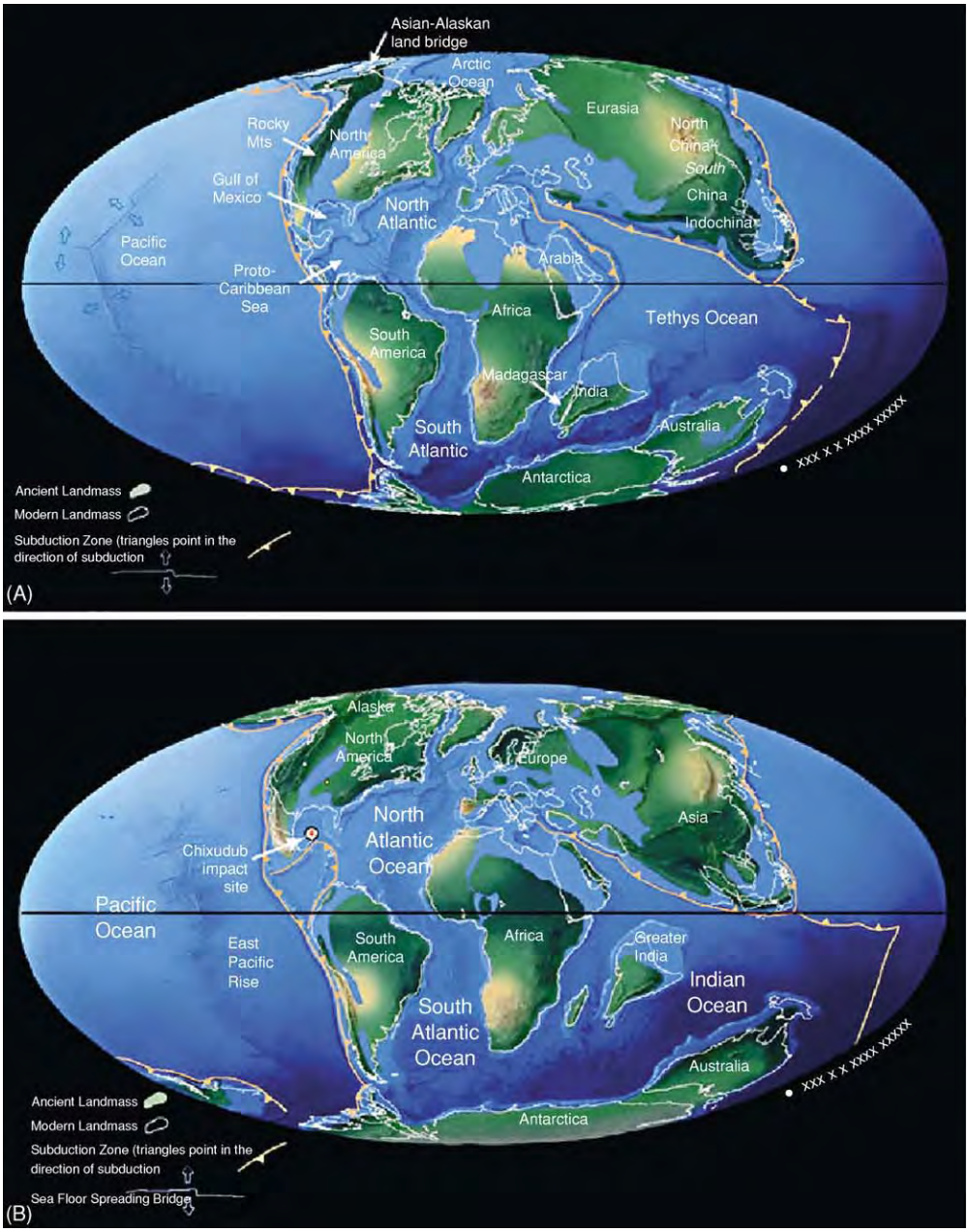


Figure 2 Reconstructions of Cretaceous palaeogeography at (A) 94 Ma (Cenomanian) and (B) 66 Ma (Maastrichtian). See text for discussion. Palaeogeographical maps generated as part of the GeoMap Project. Reprinted with permission.

Table 1 Cretaceous large igneous provice eruptions (from Courtillot *et al.*, 1996; Courtillot, 1999)

Name	*Location*	*Assoc. mantle plume*	*Age (Ma)*	*Stage*
Deccan Traps	India	Reunion	65	Maastrichtian Danian
Madagascar Traps	Madagascar	Crozet(?)	88	Turonian
Ontong Java Plateau	SE Asia	Ontong Java	122	Aptian
Rajmahal Tasman Traps	India	Kerguelen	115	Aptian
Etendeka Parana Traps	Brazil Namibia	Tristan da Cunha	133	Valanginian

been effectively split into two parts: South America–Africa and India–Antarctica–Australia. Late in the Early Cretaceous (~120 Ma), a rift opened up along the southern coast of South America–Africa and propagated north to form a proto-South Atlantic Seaway. Early in the Late Cretaceous (~110 Ma), southward-propagating rifts had opened up between India and Antarctica as well as along the north-eastern coast of Australia. Seafloor spreading took place along all of these rifts throughout the Late Cretaceous (Figure 2B), causing (1) the South Atlantic Seaway to open northward where, by 100 Ma, it had joined with the Central Atlantic Seaway, (2) India to drift northward along the coast of Africa, and (3) Antarctica–Australia to drift southward. In order to compensate for the north-eastern drift of the Tethyan seafloor and the westward drift of the Atlantic seafloor, oceanic subduction systems developed along the north-eastern margin of the Tethys and along virtually the entire circumference of the Pacific Plate.

Patterns of marine circulation were also effected substantially by these tectonic reconfigurations. During the Early Cretaceous, tradewinds blowing westward across the Tethys Ocean would have set up a strong westward equatorial current that circled the globe owing to its passage between North America and the northern part of Gondwana down the Central Atlantic Seaway. Direct evidence for the existence of this current comes from the similarity of shallow marine faunas from submerged parts of the Europe and southern Asia through to (now) submerged seamounts (island systems in the Cretaceous) over 1000 miles west of Hawaii. Aside from this circumequatorial current, paired cyclonic gyres would have been present in the Pacific Ocean, leading to the evolution of distinctive northern and southern hemisphere marine invertebrate faunas. With continued continental fragmentation and northern drift through the Cretaceous, this very simple Early Cretaceous marine circulation pattern would have grown more complex, especially once the South Atlantic Seaway had opened up between South America and Africa.

In terms of physiochemical characteristics, the Cretaceous is noteworthy for the number of large igneous province eruptions (Table 1) and the number of large bolide impact craters (Table 2) that occurred during its span, as well as the large number of major marine-anoxic events (Aptian, Albian, Cenomanian, Turonian, Coniacian, Santonian, and Campanian) and very small number of magnetic polarity reversal events (the Long Cretaceous Normal interval stretches from the Aptian to the Santonian, some 40 million years). Cretaceous rates of weathering are inferred to have been relatively low, partly as the result of relatively high sea-levels and partly because of a relative lack of mountain building, especially during the Early Cretaceous. These, along with the proliferation of phytoplankton (see later), are thought to have been responsible for elevated levels of both carbon dioxide and free oxygen in the Cretaceous atmosphere, relative to modern concentrations.

Table 2 Large Cretaceous bolide impact crater

Name	*Location*	*Diameter (km)*	*Age (Ma)*	*Stage*
Gosses Bluff	Australia	22	122.5	Beriasian
Tookoonooka	Australia	55	128.0	Hauterevian
Carswell	Canada	39	115.0	Aptian
Steen River	Canada	25	91.0	Cenomanian
Kara	Russia	60	70.3	Campanian
Mansan	United States	35	73.8	Campanian
Lappajarvi	Finland	23	73.3	Campanian
Chicxulub	Mexico	170	65.5	Maastrichtian

Sea-Level and Sedimentation Patterns

As with other intervals of Phanerozoic history, the determination of Cretaceous sea-level history is necessarily tied to analyses of sedimentation patterns and is complicated by the fact that regional tectonic factors can modify, or in some cases even obscure, the global or eustatic signal. The first-order trend towards rising sea-level in the Jurassic culminated in the Tithonian and was followed by a eustatic sea-level fall of ~50 m into the Berriasian (Figure 3) The nadir of Late Hauterivian sea-level represents the lowest sea-level documented for the entire system. Major sea-level rises took place in the Early Barremian, Albian, and Early

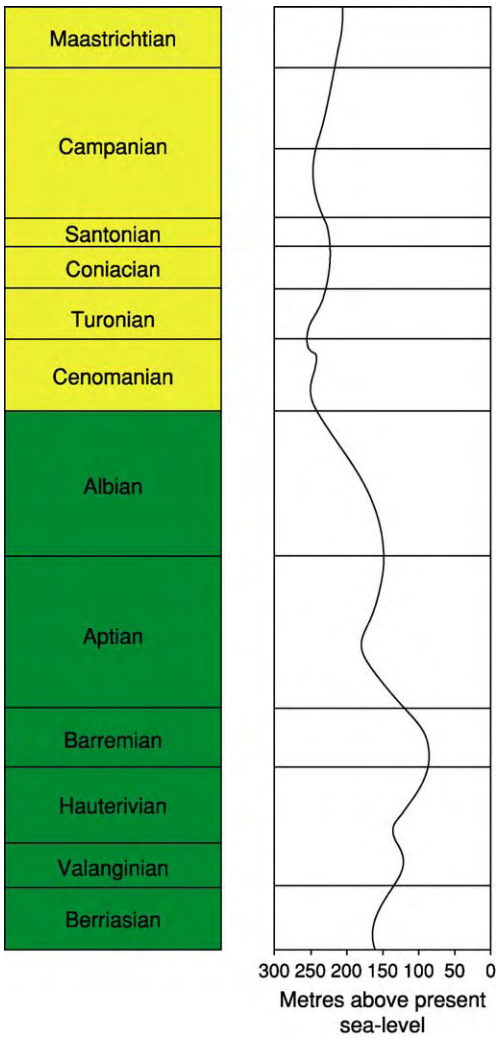


Figure 3 Global variation in sea level throughout the Cretaceous. Sea level curve partly based on data from Haq.

Campanian, with the Turonian transgression achieving the highest global sea-level stand, some 250 m above present sea-level, not only for the Cretaceous System, but also for the entire Mesozoic–Cenozoic interval. Global sea-level remained high throughout the Upper Cretaceous, but suffered quite a sudden (tens of thousands of years) and deep (50–100 m) sea-level fall more or less coincident with the Cretaceous–Tertiary (K–T) boundary. Due to the lack of evidence for widespread glaciation in the Cretaceous rocks of Gondwana, it is presumed that these sea-level fluctuations were largely driven by changes in the heat flow surrounding mid-ocean ridge systems, causing the ridge systems to swell and contract, with consequent effects on the volume of the deep-ocean basins.

The very high sea-levels achieved throughout the Late Cretaceous meant that large portions of the continental platforms were flooded to form very broad, but shallow, epicontinental seas (Figure 2). The substantial sediments that accumulated in these seas are largely responsible for the excellent Cretaceous geological and biological record. In North America and South America, collisions between the western margins of those plates and the eastern Pacific subduction centres resulted in the fold–thrust uplift of the Cordilleran mountain ranges (e.g., Sevier Orogeny) (*see* **North America:** Northern Cordillera; Southern Cordillera), along with associated volcanic and plutonic activity. Sediments from these mountains were shed to the east and west. In North America, this erosion led to deposition of the predominately clastic Great Valley sequences in California, which were subsequently deformed and uplifted during the Cretaceous accretion of microcontinental fragments (e.g., Wrangalia).

To the east of the Sevier–Laramide mountains, a large epicontinental sea (the Mowry Sea) encroached from the north and south as a result of the late Early Cretaceous sea-level rise. During its initial transgression (Albian), this sea was characterized by dysaerobic to anoxic conditions as evidenced by the abundant oil shales and black shales of this age. Clastic deposition characterized the northern part of the Mowry Sea during the Early Cretaceous, whereas carbonate-evaporite deposition characterized its southern arm. These two arms coalesced in the early Late Cretaceous (during the sea-level maximum), and a single interior seaway occupied the central portion of North America through to the Maastrichtian, during which time a more typical basinal carbonate-clastic depositional pattern become dominant. This same pattern of Early Cretaceous drowning of continental platforms also took place in South America, Europe, southern Asia, and Australia.

During the Upper Cretaceous, the characteristic chalk lithofacies developed in most large, epicontinental, marine ocean basins. These enormous chalk seas represented a singular environment that had no equivalent prior to the Late Cretaceous nor in all but the earliest part of the subsequent Cenozoic. Chalk is predominately an epipelagic sedimentary deposit composed of astronomical numbers of calcareous microfossil skeletons, chiefly nannoplankton and planktonic foraminifera. These organisms are present in the world's oceans today, but large areas of modern chalk deposits are not being created because the steady rain of calcareous from the water column is diluted by clastic sediments and by the dissolution of calcareous materials in deeper water. The shallow Late Cretaceous epicontinental seas, however, combined shallow depths with high productivity (because of their chemistry; see later) and low clastic input (because of their size) to produce near-ideal conditions for the development and preservation of plankton tests.

Moreover, these planktonic groups served as the basis for a highly productive and stable Late Cretaceous marine food chain, fostering the diversification of both marine invertebrate and vertebrate groups (see later). With the Late Maastrichtian sea-level regression, though, these chalk seas retreated from the continental interiors, although a few regional centres of chalk deposition continued into the Paleocene.

Climate

Continuing on from the Jurassic greenhouse, world climates in the Cretaceous were, if anything, even more stable, uniform, and equable, even at very high latitudes (Figure 4A). During the Early Cretaceous, climate bands with essentially modern latitude limits were found in both the northern and the southern hemispheres. Exceptions include a tropical zone largely confined to the western Tethys (as evidenced by widespread reef facies), a paratropical embayment reaching across southern Europe and into Eurasia (as evidenced by high-latitude bauxite deposits), and a very large arid region (as evidenced by widespread evaporite and calcrete deposits) in southern North America and northern South America. The overall equability of the Early Cretaceous climate is supported by the observation of coal deposits throughout Pangaea and Laurasia – even in areas reconstructed as being near the Early Cretaceous poles – along with crocodile fossils above 30° N latitude and, amazingly, above 60° S latitude!

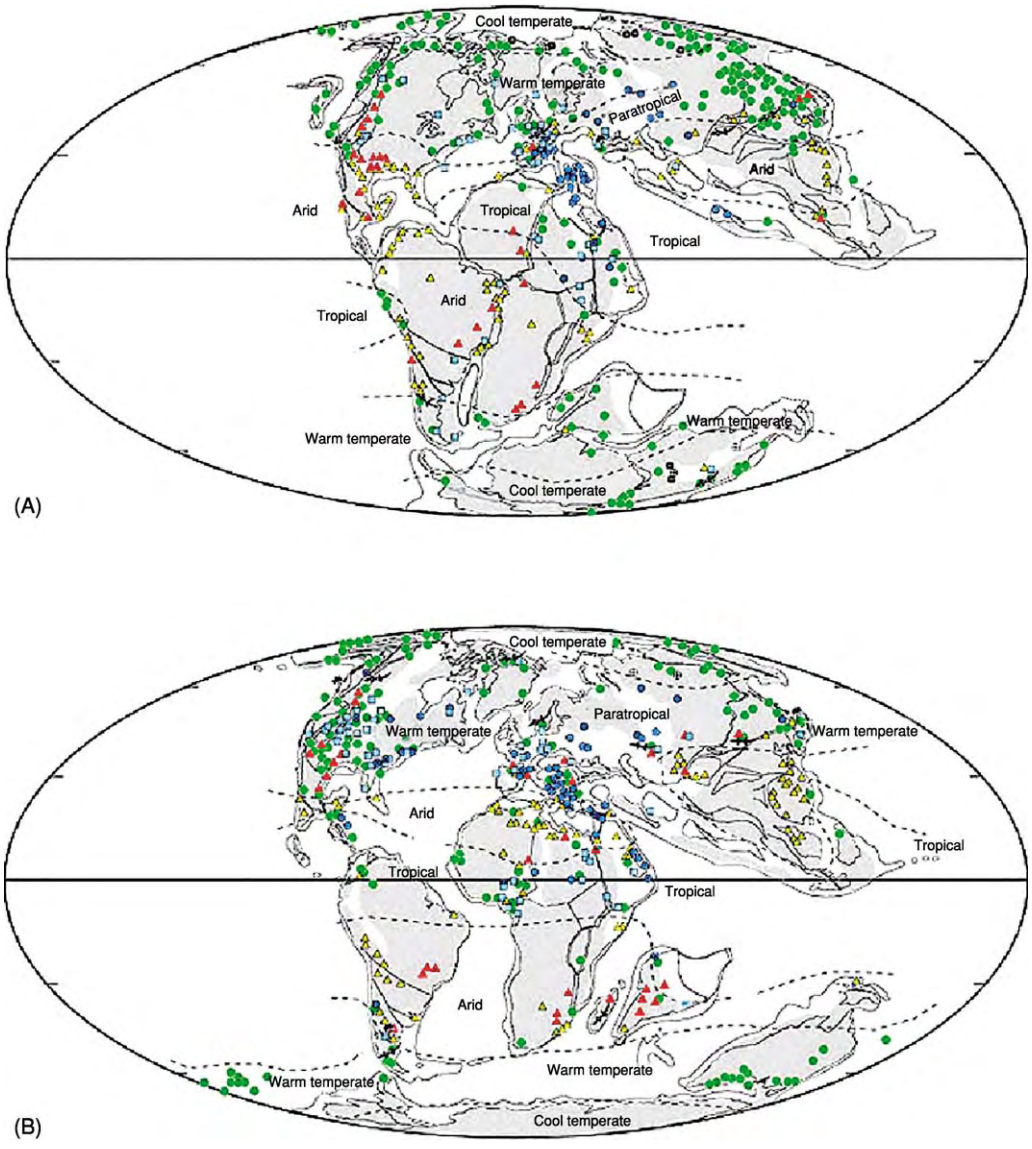


Figure 4 Environmental zones and climate variation (as inferred from sediments and biota) at (A) 120 Ma (Aptian) and (B) 80 Ma (Campanian). Palaeogeographical maps and environmental data generated as part of the GeoMap Project. Reprinted with permission.

The Late Cretaceous sea-level rise further intensified these climatic gradients (Figure 4B) by creating arid belts both north and south of the tropical equatorial zone and pushing the paratropical embayment well into Asia. However, an asymmetry developed between the north and south polar areas in the Late Cretaceous: the northern polar region contains coal deposits, crocodile fossils, and even dinosaur fossils well above the Arctic Circle, whereas these deposits stop at or near 60° S latitude. No doubt the fact that Antarctica had drifted to occupy a position at the Late Cretaceous south pole was a significant factor in the development of this climatic contrast between the hemispheres.

In the sea, $\delta^{18}O$ analyses of planktonic foraminifera indicate that both surface- and bottom-water temperatures rose steadily through the Lower Cretaceous, with a single, strong, high-temperature anomaly in the Upper Berriasian–Lower Aptian. Cretaceous marine temperatures peaked in the Turonian (average surface temperatures of ~18°C) and then went into a rapid decline to the Campanian. The Campanian–Maastrichtian interval was characterized by strong marine surface-water temperature instabilities, with several strong reversals between warm and cool phases that varied by as much as 4°C continuing to the K–T boundary. In contrast, marine bottom-water temperatures do not appear to have been subject to these strong variations.

Biota

Fossil Protists

The Cretaceous is, in many ways, the acme of the microfossil record. Two of the three modern phytoplankton groups, calcareous nannoplankton (coccoliths) and diatoms (Figure 5), underwent major family-level diversifications during this period, with diatoms being a particularly rapid family-level diversification in the

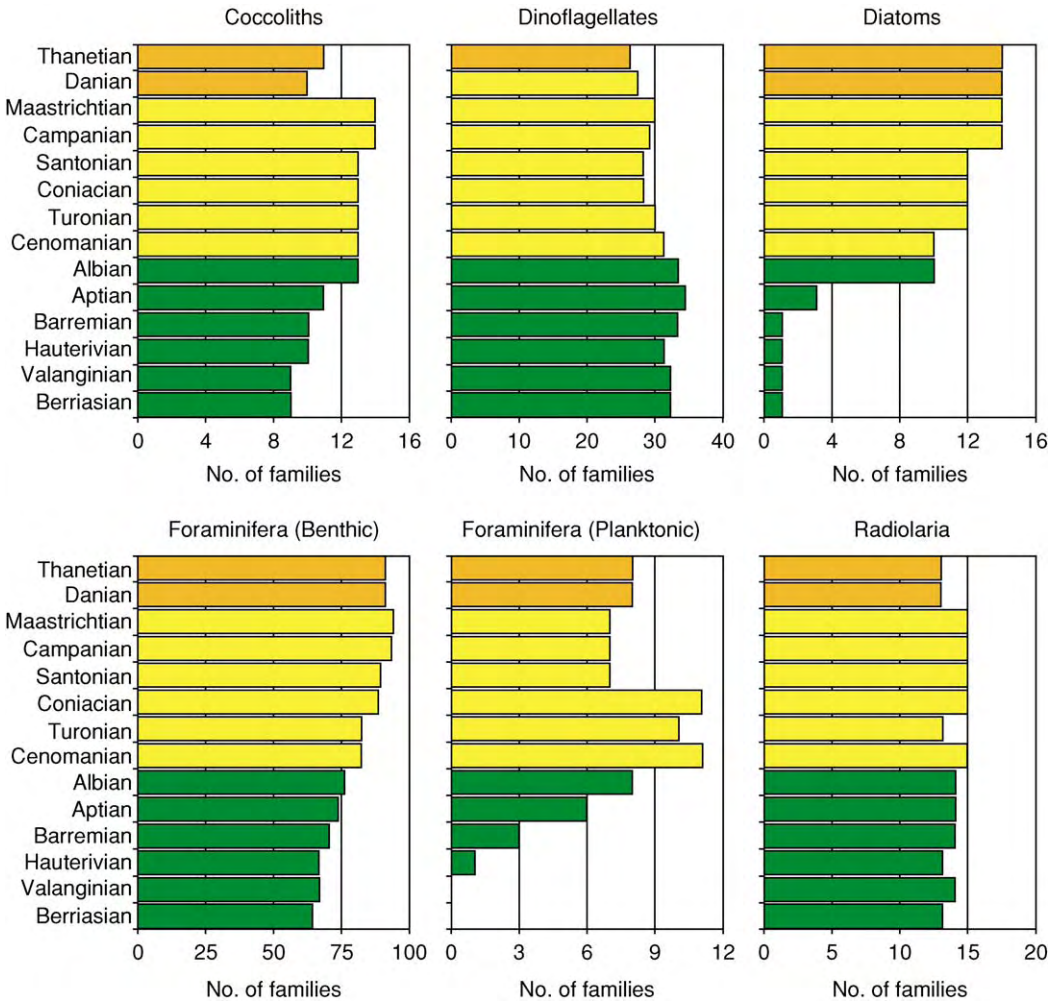


Figure 5 Fossil protist family richness patterns through the Cretaceous and the first two stages of the Paleocene. Note different scales on each graph. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

late Lower Cretaceous. Dinoflagellates maintained their very high diversities throughout the interval, albeit with a long-term drift to slightly lower family-richness values through the Late Cretaceous, and this continued into the Paleocene. Foraminifera responded to this change in the marine environment by maintaining a steady diversification of benthic forms through the late Cretaceous (probably a response to enhanced carbon deposition to the seafloor by planktivorous zooplankton and nekton). Planktonic foraminifera first appeared in the Hauterivian and, after an initial radiation that extended into the early part of the Upper Cretaceous (e.g., appearance of the first hedbergellids, heterohelicids, and guembelitrids), appear to have settled back to a steady-state family-level diversity in the upper part of the Late Cretaceous. Radiolaria appear to have maintained a more or less steady-state family-level diversity throughout the interval.

As a result of these radiations, the Cretaceous seabeds were, for the first time, blanketed with calcareous oozes. Of course, the most dramatic example of this was the extensive and economically important chalk deposits of North America and Europe. Massive Cretaceous chalk production appears to have occurred because of the very low Mg/Ca ratio of Cretaceous seawater. This chemical environment favours the production of calcite, which is then preserved because the small size of the nannoplankton-produced grains makes the (later uplifted) chalk deposit nearly impenetrable to groundwater. Widespread chalk deposition ended in the Middle Paleocene when the seawater Mg/Ca ratio began to rise from its all-Phanerozoic Cretaceous low.

Marine Invertebrates

Like most marine protist groups, corals, marine molluscs (chiefly gastropods, cephalopods, and bivalves), and marine arthropods (chiefly crustaceans and ostracods) underwent long-term family-level radiations through the Cretaceous and extending into the overlying Paleocene (Figure 6). Corals diversified strongly throughout the interval, but declined in abundance in the Late Cretaceous, presumably because the low Mg/Ca ratio made it more difficult to secrete their aragonite skeletons. The total mollusc patterns shown in Figure 6 mask strongly differing patterns among the

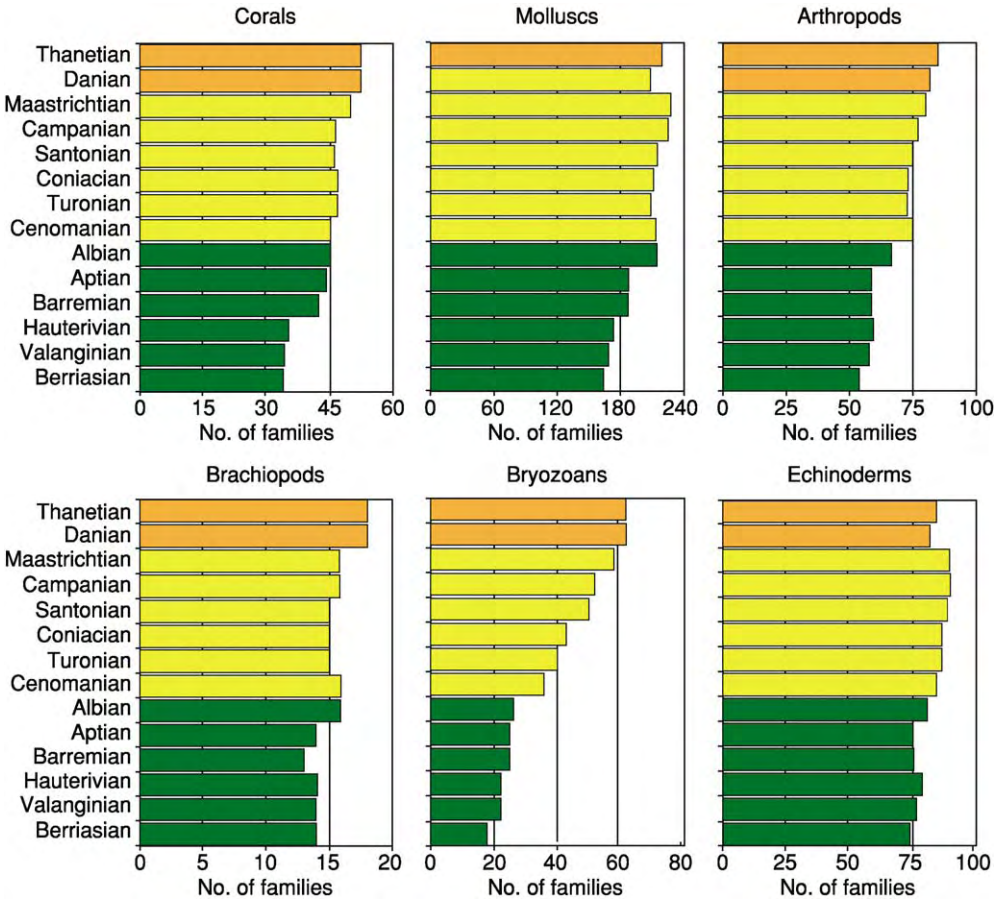


Figure 6 Fossil marine (invertebrate) family richness patterns through the Cretaceous and the first two stages of the Paleocene. Note different scales on each graph. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

constituent groups. For example, cephalopods exhibited a particularly impressive diversification in the latter half of the Early Cretaceous and across into the earlier part of the Late Cretaceous, whereas gastropods and bivalves had the greater part of their family-level diversification histories in the Late Cretaceous. Burrowing bivalves and gastropods of modern aspect (neogastropods) diversified strongly throughout the period. The Cretaceous was also a time of gigantism among several surface-dwelling bivalve lineages; the genus *Inoceramus*, for example, could be as much as a metre across (natural inoceramid casts are often mistaken by amateur geologists for 'dinosaur footprints'). One oddly shaped bivalve group, the rudists, even managed to replace corals as the principal reef builders of the shallow Late Cretaceous seas, owing to their superior chemical control of skeleton-secretion processes. The Cretaceous arthropod radiation was driven largely by crustaceans, presumably in response to the diversification of neogastropods, their principal prey item. During this time, gastropods and crabs continued their 'arms race', with gastropods developing ever more elaborate predator deterrent mechanisms (e.g., reinforced apertures, shell ribbing, and spines) and crabs responding through the development of improved claw designs (e.g., strength and shape). This crustacean diversification was particularly pronounced in the Albian and Campanian. Brachiopod family-richness values fell slightly in the Barremian, but recovered quickly and were maintained throughout the Late Cretaceous. Bryozoans exhibited perhaps the most striking richness-change pattern, with a strong and sustained diversification throughout the entire Late Cretaceous. This trend was driven by cheilostome bryozoans, which almost trebled their number of families through the course of the Late Cretaceous, though ctenostomes and cyclostomes also underwent modest diversifications through this interval as well. Echinoderms appear to have had the most conservative diversification history of any major marine invertebrate clade, with the slight drift to higher Late Cretaceous values being driven primarily by an asteroid (starfish) radiation.

Marine Vertebrates

The Cretaceous marine vertebrate faunas exhibit surprising similarities (and differences) between the richness histories of 'fish' (including sharks, rays, and bony fish) and marine 'reptiles' (including marine turtles, ichthyosaurs, plesiosaurs, and mosasaurs). Both groups (Figure 7) exhibit low, steady-state Early Cretaceous values and both exhibit much higher Late Cretaceous values. Moreover, both groups underwent the transition from low Early Cretaceous values to higher Late Cretaceous values, within a stage of one another. This similarity is all the more remarkable when the difference between the sizes of these faunas is taken into consideration. The Late Cretaceous richness increase was primarily driven by diversification within fish clades, whereas the driver of the marine 'reptile' pattern was chiefly a turtle–mosasaur diversification.

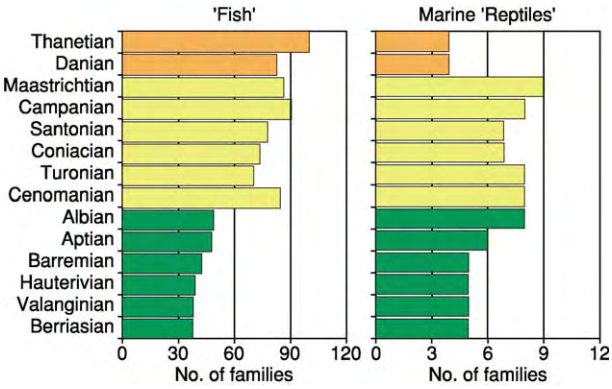


Figure 7 Fossil marine (vertebrate) family richness patterns through the Cretaceous and the first two stages of the Paleocene. Note different scales on each graph. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

Terrestrial Invertebrates

There were several groups of terrestrial and freshwater molluscs (e.g., pulmonate gastropods and unionid bivalves) as well as freshwater arthropods (e.g., crustacea and ostracodes), in addition to the ubiquitous Cretaceous insects. Among these, only the molluscs exhibit a sustained diversification trend (Figure 8). This pattern was driven primarily by the Late Cretaceous proliferation of both terrestrial and freshwater gastropods. Both non-insect arthropods and insects exhibited more or less steady-state diversification histories throughout this interval, with perhaps some suggestion of a Late Cretaceous increase in the former, driven primarily by a proliferation of Late Cretaceous chelicerate families (e.g., scorpions). It is interesting to note the very high richness values for Cretaceous insects, a group many would assume to have a 'poor' fossil record.

Terrestrial Vertebrates

Though dinosaurs understandably get most of the attention when it comes to Cretaceous terrestrial vertebrate faunas, it is important to note that essentially all modern terrestrial vertebrate groups had representatives in the Cretaceous. Although not as well known to the popular audience, salamanders, frogs, lizards, snakes, turtles, crocodiles, birds,

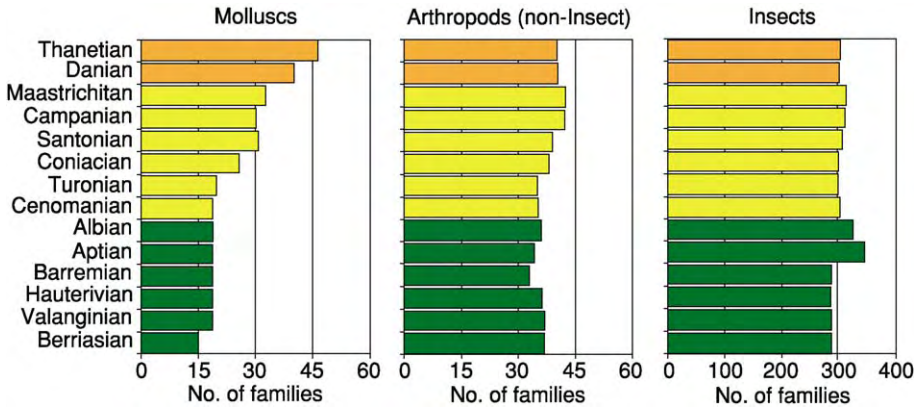


Figure 8 Fossil terrestrial (invertebrate) family richness patterns through the Cretaceous and the first two stages of the Paleocene. Note different scales on each graph. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

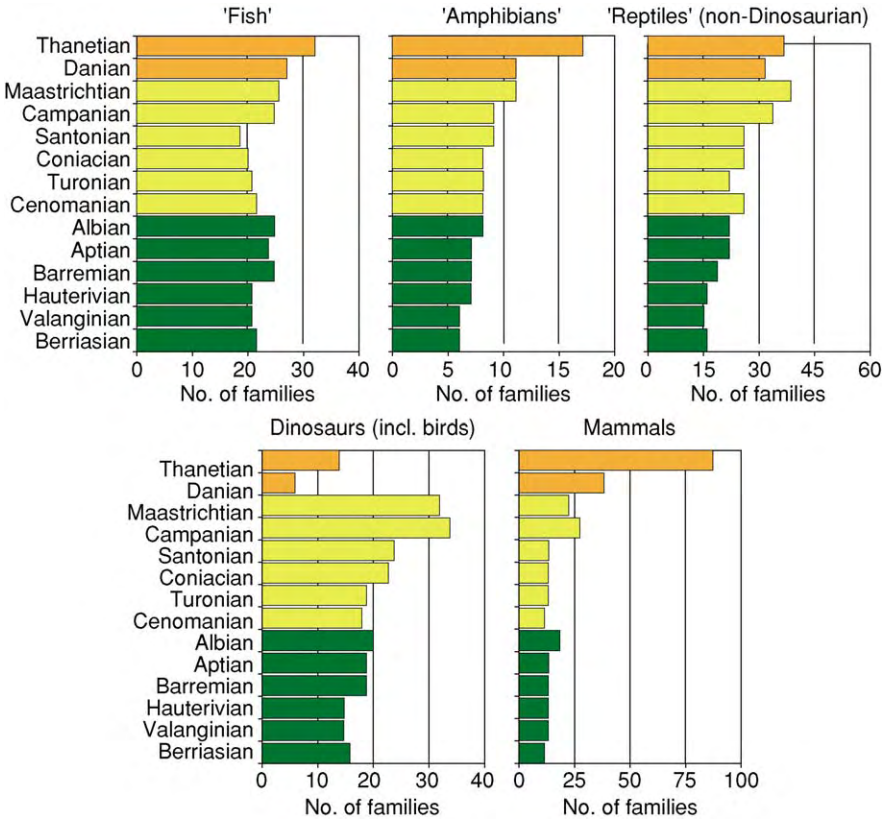


Figure 9 Fossil terrestrial (vertebrate) family richness patterns through the Cretaceous and the first two stages of the Paleocene. Note different scales on each graph. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

and mammals all shared the Cretaceous landscapes along with the non-avian dinosaurs, non-dinosaur archosaurs (e.g., crocodiles and champosaurs), and pterosaurs (Figure 9). Perhaps the most unusual Cretaceous diversification history is that of freshwater 'fish' (including sharks and rays as well as bony fish). This group began the Cretaceous with a healthy 20 or so families and underwent a modest diversification event in the latter part of the Early Cretaceous. The subsequent early part of the Late Cretaceous, when many other groups were diversifying, was a time of progressive diversity reduction in the freshwater fish fauna. This pattern was driven largely by reductions in the number of teleost and sarcopterygian clades. However, this declining pattern reversed dramatically in the Campanian (driven by a Late Cretaceous

radiation in teleost clades) and continued unbroken across the K–T boundary and into the Paleocene.

Both amphibians and non-dinosaurian reptiles (e.g., turtles, snakes, and lizards) exhibited similar progressive patterns throughout the Cretaceous that also extended into the Paleocene. Dinosaurs had strong diversification throughout the Late Cretaceous up to their Campanian peak, from which there was a marked retreat (involving both theropod and sauropodomorph forms) in the Maastrichtian. All non-avian dinosaurs became extinct by the end of the Maastrichtian, and the Danian values in Figure 9 all represent bird families. This said, there are persistent reports of dinosaur remains and signs (e.g., nests and trackways) in Danian strata and it would not be surprising if these reports were confirmed and accepted by most vertebrate palaeontologists in the near future.

Last, but by no means least, mammals exhibit very low family-richness values throughout most of the Early Cretaceous and the lower part of the Late Cretaceous. There is a hint, however, that mammalian diversification started in the Campanian, at about the same time as the final dinosaur diversification. Mammals also underwent a richness reduction in the Maastrichtian coincident with a similar reduction in dinosaur family-richness values. Indeed, mammals lost more families in the Maastrichtian than did dinosaurs! The fates of these two clades were decoupled by other events in the Maastrichtian, but it is false to regard their diversity histories as being mirror images of one another, as has been the case in many popular accounts and more than a few scientific treatises. Rather than ships passing in the night, the Cretaceous history of mammals and dinosaurs suggests several striking similarities, the interpretation of which would benefit from further investigation.

Terrestrial Plants

The Cretaceous bore witness to a fundamental transformation of the terrestrial flora, with the initial diversification of seed-bearing (angiosperm) plants. Though angiosperms first appeared in the Jurassic, 'naked seed' gymnosperms such as conifers, cycads, and ginkgoes dominated the earliest Cretaceous landscapes. Nevertheless, the strong plant diversification event documented in Figure 10 is almost entirely the result of family production within angiosperms. During the Cretaceous, over 30 new angiosperm families appeared, including beech, birch, fig, holly, magnolia, oak, palm, sycamore, and walnut. Angiosperm seed development is considered superior to that of gymnosperms because the seed is protected from damage inside the plant's ovary. This means that the plant can produce a much larger number of smaller seeds from the same energy investment required to make the less efficient gymnosperm cone. Flowers were another Cretaceous angiosperm development, and these literally opened up a host of new possibilities for the attraction of pollinator organisms, effectively co-opting other creatures into the plants' reproductive strategy. Like the first plants, the first angiosperms appear in Early Cretaceous lake and stream habitats, where quick-developing, opportunistic, weedy forms have an advantage. However, there were no grasses, and so no grassy plain or savanna habitats, in Cretaceous times.

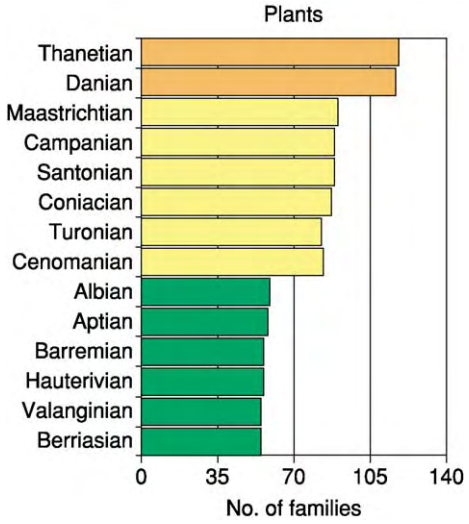


Figure 10 Fossil terrestrial (plant) family richness patterns through the Cretaceous and the first two stages of the Paleocene. Based on data from Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

Extinctions

Two extinction events are widely held to have occurred during the Cretaceous, an Aptian event and a Cenomanian event. There is considerable debate on whether the Aptian event is important, or even recognizable in most palaeontological datasets. In the late JJ (Jack) Sepkoski's family-level and genus-level compendia, an Aptian extinction-intensity peak stands slightly above what he and David Raup regarded as background extinction levels. The extent of this peak's distinctness has changed over successive versions of the Sepkoski dataset and is, to some extent, dependent on how time is 'binned' across the relatively long Aptian and Albian stages. In addition, an extinction peak in the vicinity of 119 Ma was predicted on the basis of Raup and Sepkoski's extinction-periodicity hypothesis.

Irrespective of these analytic results, specialist biostratigraphers and systematists have been hard

pressed to identify which Aptian organismal groups bore the brunt of this putative extinction event. Modest losses among planktonic foraminifera, benthic foraminifera, calcareous algae, and rudistid bivalves have been noted, but none of these appears to have had a dominant effect on the diversity histories of any of these groups, not to mention the biosphere as a whole. Possible extinction mechanisms also abound in the Aptian, including a local sea-level lowstand, an ocean-anoxic event, and a large volcanic eruption that emplaced the submarine Ontong-Java Plateau. Nevertheless, the Aptian extinction record appears to have a relatively weak claim to be a major turning point in Earth's biodiversity history.

The case for a major extinction event at the Cenomanian–Turonian boundary is much stronger. Beginning in the Upper Cenomanian, and continuing into the Lower Turonian, marine taxa suffered the lost of about 7% of all families with fossilizable hard parts and 26% of all constituent genera. This event occurred over an interval of from 3 to 5 Ma and was complexly structured both temporally and geographically. Ecologically, the extinction appears to exhibit a distinctly 'bottom-up' character, with deep-dwelling taxa (e.g., gastropods, bivalves; and benthic foraminifera, especially agglutinated-walled forms; deep-dwelling planktonic foraminifera; and bottom-feeding ammonoids) being differentially affected. Over a few metres of section, species-level extinction rates ranging from 20% to 50% (and occasionally as high as 80%) are commonly reported. At least locally, however, taxa in shallower water (including coccoliths, pelagic ammonites, and rudistid bivalves, as well as sharks, ichthyosaurs, and plesiosaurs) fell victim. Geographically, the extinction appears to have been centred on tropical and temperate biota, with virtually no extinctions being recorded above 60° north or south of the Cretaceous palaeoequator.

Geochemically, the extinction event is correlated with the widespread interruption of chalk formation in epicontinental seaways in favour of clastic deposition, especially in the form of the black shales and limestones that signal low-oxygen or dysaerobic conditions. In addition, this event is associated with pronounced $\delta^{13}C$ and $\delta^{18}O$ anomalies, a sea-level highstand, and an interval of widespread submarine and subareal volcanism. No convincing or consistent evidence for bolide impact has been recovered from Cenomanian–Turonian boundary sediments to date. Two Ir anomalies are associated with the Cenomanian–Turonian interval, but, owing to similarities between the trace-element signature of these anomalies and the trace-element composition of mid-Atlantic Ridge basalts, these anomalies have been interpreted as resulting from volcanic, rather than extra-terrestrial, inputs.

Virtually all of the popular extinction mechanisms, including bolide impact, have been advanced to explain the Cenomanian–Turonian extinction. The current geological and palaeontological consensus suggests that this is a multicausal event that occurred because a unique juxtaposition of independent factors affecting marine habitats coincided during this interval of Earth history. The most important proximal cause was probably anoxic deep-ocean waters invading middle and, in some cases, shallow shelf habitats as a result of eustatic sea-level rise. The Cenomanian–Turonian interval represents the highest stand of sea-level for the entire Phanerozoic as well as the highest Cretaceous sea-level stand. This primary mechanism was likely intensified by submarine volcanism (which can alter the buoyancy of dysoxic and anoxic deep-marine waters, causing them to rise further up the continental shelves than would have otherwise been the case) and global cooling resulting from improved marine circulation patterns that were probably a by-product of the sea-level highstand. This improved circulation may have increased the efficiency of heat transfer from tropical to polar regions, thus cooling the topics and (perhaps) exceeding the tolerance of many warm-adapted tropical species, including reef-building rudists and corals. As is the case with most modern extinction events, once a sufficient number of key species had been eliminated, other species would become increasingly susceptible to extinction because of complex ecological dependencies rather than physiochemical tolerances per se.

It has been suggested that the magnitude of the Cenomanian–Turonian extinction event has been overestimated as a result of the geographic migration of certain types of marine environments in the stratigraphic record, and because any single section or core presents a picture of changing local conditions superimposed over changing global conditions. Though all fossil records must be examined by biases of this sort, the widespread and, on the whole, consistent nature of the biotic patterns described from the Cenomanian–Turonian interval argues that such biases – though undoubtedly present – can be invoked to discount the entire event. Additional research will be needed in order to test these hypotheses and further sharpen our understanding of this fascinating interval of Earth history.

See Also

Atmosphere Evolution. **Biozones**. **Fossil Invertebrates:** Arthropods; Brachiopods; Bryozoans; Corals and Other Cnidaria; Echinoderms (Other Than Echinoids); Molluscs Overview. **Fossil Plants:** Calcareous

Algae. **Fossil Vertebrates:** Fish; Palaeozoic Non-Amniote Tetrapods. **Mesozoic:** End Cretaceous Extinctions. **Microfossils:** Foraminifera. **Plate Tectonics**. **Sequence Stratigraphy**. **Stratigraphical Principles**. **Time Scale**.

Further Reading

Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

Gould SJ (ed.) (1993) *The Book of Life*. New York: WW Norton.

Hallam A (1992) *Phanerozoic Sea Level Changes*. New York: Columbia University Press.

Hallam A and Wignall PB (1997) *Mass Extinctions and Their Aftermath*. Oxford: Oxford Science Publications.

Haq B, Hardenbol J, and Vail PR (1987) Chronology and fluctuating sea levels since the Triassic. *Science* 235: 1156 1166.

Stanley S (1989) *Earth and Life through Time*, 2nd edn. San Francisco: WH Freeman.

Stanley SM (1999) *Earth System History*. New York: WH Freeman.

End Cretaceous Extinctions

N MacLeod, The Natural History Museum, London, UK

Introduction

Although the end-Cretaceous extinction event goes by a variety of names in both the technical and popular literature (e.g., Cretaceous–Tertiary (K–T) mass extinction, K–T boundary extinction, Cretaceous–Paleocene (K–P) mass extinction, Cretaceous–Palaeogene (K–P) extinction), it is most closely associated with the uppermost Cretaceous period/stage—the Maastrichtian—and, to a lesser extent, the lowermost Palaeogene epoch/age—the Danian (**Figure 1**). As such, any review of effects and causes that may (or may not) have occurred during this episode of Earth history must begin with a review of the stratigraphy of these two intervals, not only because they provide the temporal, environmental, geographic, and tectonic context within which such phenomena must be understood, but also because the study cause–effect associations in Earth history is, by definition, a largely stratigraphic exercise.

Dumont first defined the Maastrichtian in 1850 as a distinct stratigraphic subdivision of the Upper Cretaceous well expressed in the sediments around the town of Maastricht in southern Holland. The current boundary stratotype was established in 1911 by the Comité d'étude du Maastrichtian as the Tuffeau section, exposed in the ENCI quarry at St. Pietersburg on the Maastricht outskirts. Unfortunately, only the upper Maastrichtian is exposed in this quarry and its upper boundary has traditionally been regarded as incomplete (though a 1996 restudy of the quarry concluded that a complete K–T boundary succession was present in certain man-made cave sections within the quarry). In terms of boundary stratotypes the Campanian–Maastrichtian boundary was established by the International Commission on Stratigraphy (ICS) in 2001 at the 115.2-m level in Grande Carrière Quarry, Tercis-les-Bains, Landes Province, in south-western France. A total of 12 criteria (all of equal weight) were defined as useful in recognizing the Campanian–Maastrichtian boundary, which falls just above the first appearance of the ammonite *Pachydiscus neubergicus* in this section.

The overlying Danian was established by Desor in 1847 for the stratigraphic successions present at Stevens Klint and Faxse in Denmark. Desor originally regarded the Danian as a Cretaceous stage because it is characterized in these two localities by chalk lithofacies. Later, it was shown that these sections were equivalent temporally to the Montian stage of Belgium that had long been regarded by continental European stratigraphers as the lowermost stage of the Tertiary. After a short debate concerning whether to place the new Danian–Montian stage in the Cretaceous or the Tertiary, the latter was accepted, largely on the basis of similarities between this fauna and that of the Midway Formation of the US Gulf Coast, which had also long been regarded as being basal Tertiary. These correlations notwithstanding, controversy regarding the correct placement of the Danian continues, largely on the basis that a number of characteristically 'Cretaceous' taxa (including bryozoans, brachiopods, echinoids, gastropods, bivalves, and, perhaps, planktonic foraminifera) did not become extinct until the end of the Danian where chalk deposition also effectively ceased worldwide. Some have regarded the Danian as the time interval between the (Maastrichtian) 'white chalk' exposed at Stvens Klint and the basal Selandian conglomerate exposed at Hvalløse in Jutland.

The Danian boundary stratotype was established by the GSSP in 1996 at the base of the boundary clay in the El Haria section, near the town of El Kef,

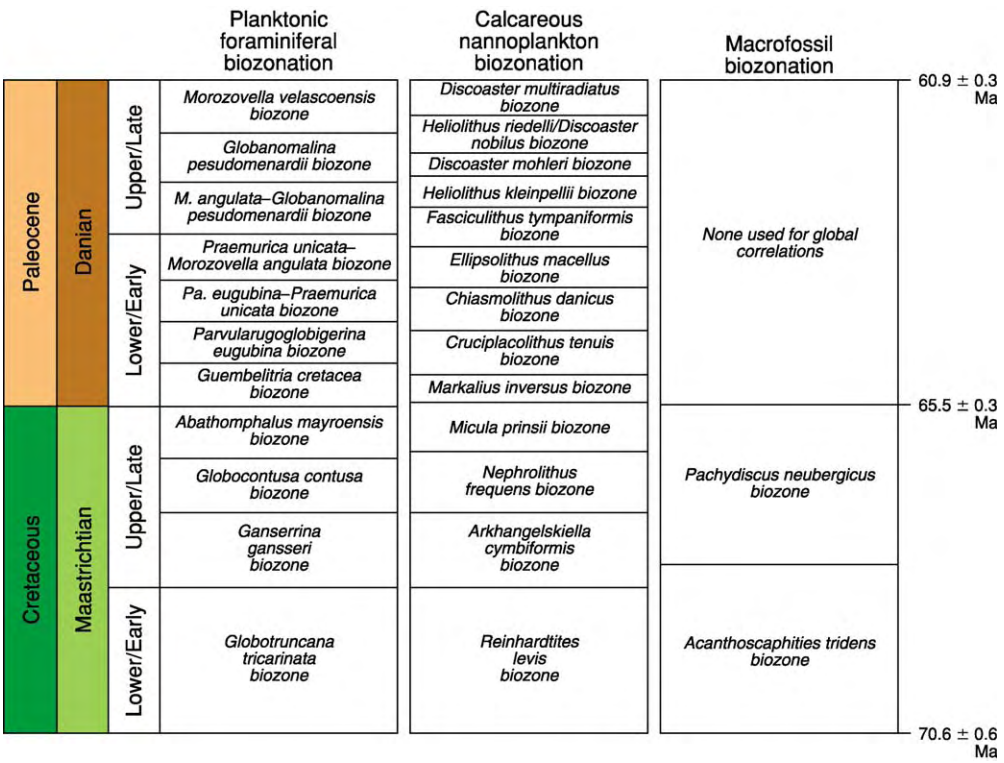


Figure 1 Chronostratigraphy, biostratigraphy, and geochronology of the (Cretaceous) Maastrichtian through (Palaeogene, Paleocene) Danian interval. Boundaries between Early/Lower and Late/Upper Maastrichtian and Danian represent undated, informal units based on biostratigraphic zones.

Tunisia. Like the Campanian–Maastrichtian boundary, multiple criteria are used to recognize the Maastrichtian–Danian boundary (which is also the Cretaceous–Tertiary, Cretaceous–Palaeogene, and Mesozoic–Cenozoic boundary), including an iridium (Ir) anomaly and a major extinction horizon for foraminifera and calcareous nanoplankton. The boundary stratotype for the base of the Selandian Epoch (= top of the Danian) has not been established to date.

Like almost all chronostratigraphic boundaries, the Maastrichtian–Danian interval is difficult to correlate because:

1. most so-called boundary successions are profoundly incomplete, often juxtaposing sediments of markedly different ages (e.g., Campanian beneath Eocene), and
2. biozones (see Figure 1) contained within the interval are demonstrably diachronous and often facies limited.

For example, the traditional uppermost Maastrichtian planktonic foraminiferal zone—the *Abathomphalus mayroensis* taxon range zone—is based on a open ocean, deep-dwelling species not present in shallower marine environments, including those represented by the El Kef boundary stratotype. The result has been confusion regarding what constitutes a complete record of the uppermost Maastrichtian–lowermost Danian sequences of events, especially between different ecological realms and locations remote from one another.

The Nature of the Maastrichtian–Danian Turnover Patterns

Historical Concept

Despite the problems associated with attempts to achieve a high-resolution, temporal correlation of end-Cretaceous and Early Danian sediments, it is abundantly clear that something extraordinary happened to the Earth's biosphere in the Late Cretaceous. In many lineages characteristic elements of the Late Cretaceous marine and terrestrial biota are simply not present in the Early Paleocene fossil record. The standard list of Late Cretaceous victims includes coccoliths and planktonic foraminifera (both of which survive into the Danian but with an almost complete species-level taxonomic turnover), rudistid and inoceramid bivalves (both of which were common constituents of the Late Cretaceous chalk seas and both of which became entirely extinct),

ammonite cephalopods (the traditional index fossil group of Cretaceous biostratigraphy, all of whose representatives became entirely extinct), marine reptiles (e.g., ichthyosaurs, plesiosaurs, mosasaurs; all entirely extinct), flying reptiles (=pterosaurs, extinct), and all nonavian dinosaurs (extinct). (Modern birds are nested phylogenetically within the dinosaur clade and so are considered dinosaurs in terms of their ancestry. Recent discoveries of feathered Cretaceous dinosaurs in China confirms this phylogenetic link and puts the 'obvious morphological distinctions' between extinct nonavian dinosaurs and modern birds—which are often cited by those who would try to preserve phylogenetically unwarranted taxonomic distinctions between these two groups—in a somewhat different light.) On one hand, this might not seem a very long list, especially given the fact that these groups comprise a relatively small collection of suborders, superfamilies, families, and genera in a much larger Late Cretaceous biota. Nevertheless, several victim groups were among the set of dominant marine and terrestrial players on the Late Cretaceous ecological stage whose demise affected many other groups in both direct and indirect ways. Also, since many evolutionary advantages flow from ecological incumbency, the removal of these groups from the ecological scene (by whatever means) opened up opportunities for evolutionary–ecological transformation and diversification among Late Cretaceous survivor lineages that simply would not have been possible otherwise. (This dependency on the unpredictable elimination of competition between lineages is referred to as contingency in the evolutionary literature, and it is widely reported that, were it not for the contingent elimination of nonavian dinosaurs in the Late Cretaceous, the explosive radiation of mammals—in which our own lineage, primates, took part—would not have taken place. Although this is true in a general sense, it is not the case that we owe our existence solely to the end-Cretaceous extinction since many other antecedent as well as subsequent contingencies also played important roles in mammal-primate evolution.)

In order to understand the character of this extinction event, the obvious questions to ask are as follows:

1. When did these extinctions occur?
2. Did they affect only these traditional victim groups?
3. Did any groups diversify or prove unusually extinction-resistant over this interval?
4. Were these Late Cretaceous extinctions associated with any physical event(s) that could have been responsible for the extinctions?

Before answers to these questions can be attempted, however, some further aspects of the Late Cretaceous–Early Tertiary fossil record must be understood.

Sources of Bias in the Fossil Record

Although all information we possess about fossils is traceable ultimately to observations made on actual specimens, many factors conspire to constrain the scope and depth of this knowledge for different fossil groups. For example, the contrast between the Late Cretaceous planktonic foraminiferal and dinosaur records could hardly be more striking even though both have played roughly equivalent roles in the understanding of end-Cretaceous extinction patterns and processes. Planktonic foraminifera are very common constituents of Maastrichtian and Danian marine sediments deposited over a broad range of marine depth habitats. Species can be recovered from these sediments worldwide, sampled in great temporal detail (owing, not least, to the small size of foraminiferal shells), and placed within a very highly resolved biostratigraphy–taxonomy, albeit the latter of whose phylogenetic dimensions are not especially well known. Dinosaurs, on the other hand, are comparatively rare constituents of Maastrichtian terrestrial sediments and occur in only a narrow range of terrestrial habitats. Uppermost Maastrichtian dinosaurs are known to occur in only one area of the world (western USA) and cannot be sampled in great temporal detail, because of the spotty occurrence of complete or near-complete skeletons. Dinosaur biostratigraphy and taxonomy are not as well known as planktonic foraminiferal taxonomy (e.g., it is difficult to assess the shifting patterns of erosion-prone terrestrial sections for chronostratigraphic relations to stratotype successions, since new dinosaur species turn up each year). Nevertheless, owing to the complex nature of vertebrate skeletons, more is known about dinosaur phylogeny than planktonic foraminiferal phylogeny. The Late Maastrichtian–Danian fossil records of other groups can be thought of as ranging through a spectrum of temporal, geographic, taxonomic, and phylogenetic resolutions, the boundaries of which are set by dinosaurs and planktonic foraminifera.

Not only are there intrinsic differences between the quality of the fossil records of different groups, there are also several fundamental sources of uncertainty about the details of all fossil records, which constrain interpretations based on palaeontological observations. First, lack of a detailed, global chronostratigraphy for Maastrichtian–Danian sediments means that the vast majority of fossils can only be located as occurring—or not occurring—within these relatively coarse temporal intervals. Thus, although it is

possible to say that the extinction of plagioptychid bivalves and azhdarchid pterosaurs both occurred in the Maastrichtian, it is not possible to place these extinctions accurately within the span of the Maastrichtian, or to say whether they occurred simultaneously with respect to one another, without going back and restudying the original material. In particular, it is not appropriate to assume that simply because an extinction event is listed as Maastrichtian, all species composing the group in question ranged through the entire Maastrichtian and then simultaneously became extinct at a horizon coincident with the Maastrichtian–Danian boundary.

It is standard practice for biostratigraphic range charts to represent the chronostratigraphic ranges of fossil groups as solid lines joining individual occurrence horizons that denote intervals of time along a temporal axis (Figure 2). Because we experience time as a continuum it is tempting to regard these axes as representing time as a continuous variable. In fact, these charts represent time as a discontinuous variable with the vertical range line always being drawn through the entire interval, irrespective of whether the actual time of extinction is known. This graphing convention often gives the (erroneous) impression that all extinction events occur at stage/age boundaries and that there is some pronounced tendency for extinctions to occur together in time.

Similarly, use of higher taxonomic categorizations (e.g., families, genera) as proxies for species in extinction studies often leads to unappreciated distortions of the fossil record by inexperienced interpreters. The taxonomic categories of family and genus are used most often in extinction studies because these are regarded as more stable and comparable than species-level data across the broad scope of life's diversity. Because these are composite categories, however, the presence of a family at one point in time may represent a rather large number of species, whereas at another point in time (especially if the latter is close to the group's extinction event) the actual number of species represented may be much smaller (Figure 2).

Pseudoextinction is another problem. In an evolving lineage two types of morphological transformations can occur. The first (anagenesis) results in the progressive transformation of the entire species from

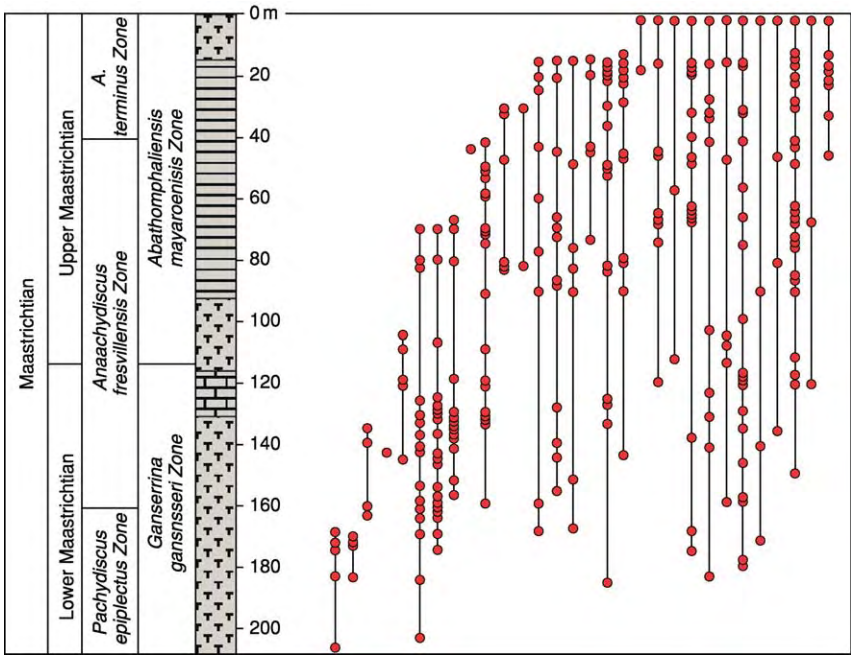


Figure 2 Maastrichtian ammonite biostratigraphy at Zumaya, Spain. Note that the majority of species (and, by implication, the genera and families they represent) disappear from this section during the course of the Maastrichtian, not at the Maastrich tian Danian boundary. In fact, the last appearance datum of no ammonite species has been found to coincide with the Maastrichtian Danian boundary in this section. Moreover, of the 12 species whose last appearance is within 1.5 m of the boundary, approx.half have exceedingly rare or patchy biostratigraphic distributions, which suggests patchy, uneven, extinction susceptible population struc tures. In this section, the mass extinction of ammonites is equivalent to the near coincident disappearance of 6 7 previously abundant species, a phenomenon that would hardly be noticed in other parts of the Cretaceous ammonite record. Note also that, because of the patchy occurrence pattern characteristic of all ammonite species in this section, the idea that some ammonite species may have survived into the lowermost Danian interval cannot be rejected statistically. Redrawn from Marshall and Ward (1996) Sudden and gradual molluscan extinctions in the latest Cretaceous of western European Tethys. *Science* 274: 1360 1363.

one morphological condition or state into another. The other (cladogenesis) results when a single species is split into two or more daughter species, one of which may continue to exhibit the morphological condition of the ancestral population(s). Whereas the extinction of a lineage by the physical death of the last individual within the population must be regarded as a true extinction event, the physical death of the last individual within an extant population that simply happens to exhibit an ancient or atavistic morphological condition is not the same sort of event. This latter situation is termed pseudoextinction (Figure 3). Although the phenomenon of pseudoextinction need not be distinguished from true extinction in most routine biostratigraphic studies (after all, the pseudoextinction event does take place at a particular time), the differences between pseudoextinction and true extinction are profound in the context of extinction studies. Mammal systematist/palaeobiologist David Archibald has estimated that as many as 25% of the extinctions recorded in three different Early Puercan (=lowermost Danian) mammal lineages were pseudoextinctions.

Raw observations can also be deceiving in terms of the simple observation of fossil groups' last occurrence distributions. Since the last observed occurrence of each species is only an estimate of its true extinction coordinate in time and space (*see* **Biozones** for additional discussion), the vagaries of preservation and sampling will conspire to distort the observed extinction record for each species (and, by

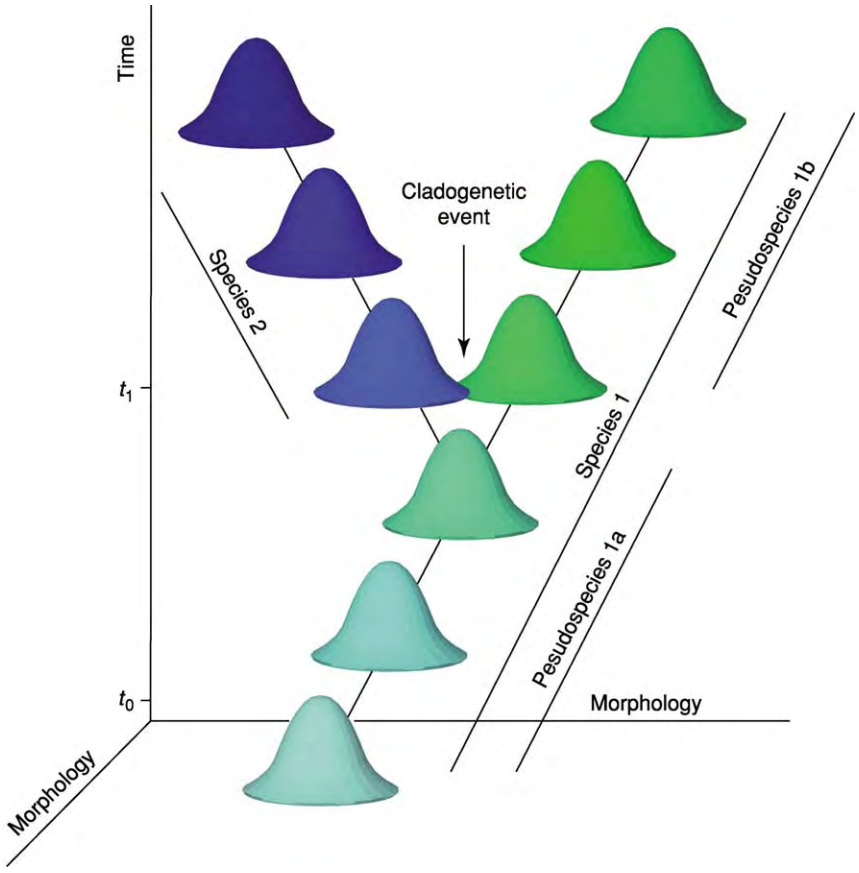


Figure 3 An example of the distinction between true extinction and pseudoextinction. Morphological change often occurs during the course of a lineage's evolutionary history. Morphological changes that result in a daughter species diverging from its ancestor are said to have arisen via cladogenesis (=lineage splitting). In the diagram cladogenesis has taken place approximately halfway up the time axis and resulted in the production of Species 2 from Species 1. Cladogenetically produced species are true species with a definite beginning (the cladogenetic event), history, and end (the extinction event). However, after a lineage has undergone so much within lineage (=anagenetic) morphological change that no individuals at time t_1 exhibit the same morphological states as individuals in ancestral populations (time t_0), biostratigraphers often find it useful to differentiate between the t_0 and t_1 populations by giving a different species name to the latter. This convention, in effect, results in the 'extinction' of the t_0 species, not because of the death of the last individual belonging to the lineage, but merely as the result of a nomenclatural change. Such nomenclatural extinctions are termed pseudoextinctions. Without a detailed knowledge of the phylogenetic and nomenclatural history of a lineage, it is often impossible to distinguish true extinction from pseudoextinction.

extrapolation, for each genus, family, etc.) to a greater or lesser extent (Figure 4). This phenomenon was termed the Signor-Lipps Effect by the nominal authors in 1982, although Alan Shaw offered an essentially identical description of the same phenomenon as early as 1964. There is no way to correct for the Signor-Lipps Effect. Using the distribution of gaps between known occurrences within a species' or higher taxon's stratigraphic range, and making use of several simplifying assumptions, it is possible to estimate a confidence interval above which it is appropriate to regard the species or group as being truly extinct. Nevertheless, there is no method whereby an investigator may retrospectively pinpoint a taxon's 'true' extinction level. For groups of taxa various scenario-based extinction geometries can be evaluated using extensions of the confidence-interval method, but this approach will always identify a spectrum of geometries, ranging from randomized patterns to a strictly simultaneous geometry, as being equiprobable.

Finally, simple taxonomic uncertainty conspires to complicate data reported in the palaeontological literature. Certain species or higher groups that appear to become extinct at particular stratigraphic levels appear again, later in time at the same locality or in a different region altogether. Sometimes the morphological condition of the specimens at the second appearance is similar to that of predisappearance specimens and sometimes it is not. If the second appearance is regarded as a continuation of the predisappearance taxonomic entity it is referred to as a Lazarus taxon and should not be regarded as having undergone anything other than a local extinction at

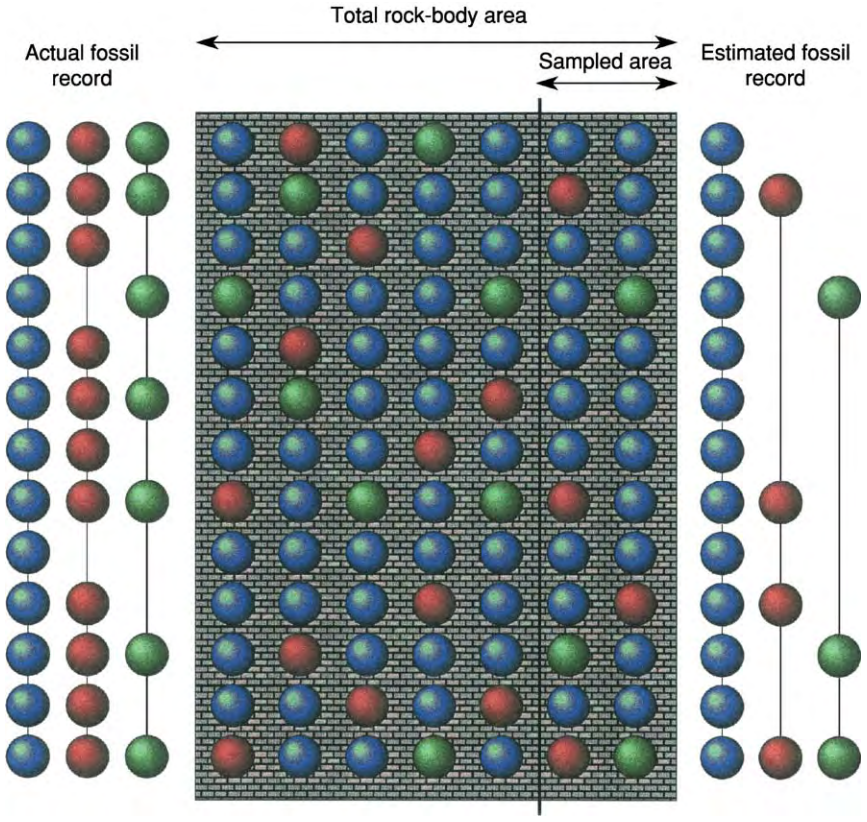


Figure 4 A hypothetical example of the Signor Lipps Effect. The total fossil record of any species is provided by the sum total of all occurrences of the fossil within a body of rock. However, because a relatively small portion of that total rock body is ever available for sampling, the known fossil record represents a subset of the total fossil record. Because of this distinction between the total and known fossil records, and because of additional biases that might arise as a result of changes in the relative abundance of fossil species over time, changes in facies within single stratigraphic successions, and sampling patterns, estimates of the pattern of last appearance data (e.g., coincident vs progressive) are usually biased towards the recovery of progressive patterns. In this example, note that despite the fact that all three species range through the entire section, ranges based on the subset of occurrences in the two right hand columns yield a sequential extinction pattern. This phenomenon has been used to argue that progressive extinction patterns imply an abrupt extinction mechanism. Such arguments are invalid as was pointed out by Signor and Lipps (1982) because the progressive bias applies to genuinely progressive as well as genuinely abrupt true extinction patterns. The best one can do is not to reply on the apparent pattern of biostratigraphic last appearance data as reliable estimates of true extinction rates.

the lower horizon. However, if the second appearance specimens differ in some way from their characteristic predisappearance forms such that their placement within the previously existing lineage is questionable, the second appearance may regarded as a suspect or 'Elvis' taxon, in which case the status of the group's extinction is uncertain.

Taking all of these potential biasing factors into consideration, the following sections outline our current understanding of Maastrichtian–Danian biotic turnover patterns for six different ecologically subdivided biotas. In all cases these data are based on the family-level compilations provided by *The Fossil Record 2* (Benton, 1993) as the most up-to-date single source of stratigraphic information across the broad spectrum of fossil groups. Although not as detailed taxonomically as JJ Sepkoski's *Compendium of Fossil Marine Animals* (2002), the former has the advantage of encompassing terrestrial as well as marine taxa and being compiled, reviewed, and adjudicated by specialists in each group.

Marine Microfossils

Marine microfossils, including protistan autotrophs and heterotrophs, have traditionally been thought to be one of the broad taxonomic groups most affected by the end-Cretaceous extinction. Review of the family-level fossil record largely bears this out. Among the major marine microfossil groups (Figure 5), only diatoms fail to exhibit a Maastrichtian extinction-intensity peak. For coccoliths, benthic foraminifera, and radiolaria this peak is more-or-less isolated, suggesting operation of a causal process or processes that were confined to the Maastrichtian. For dinoflagellates and planktonic foraminifera though, the Maastrichtian peak appears to be part

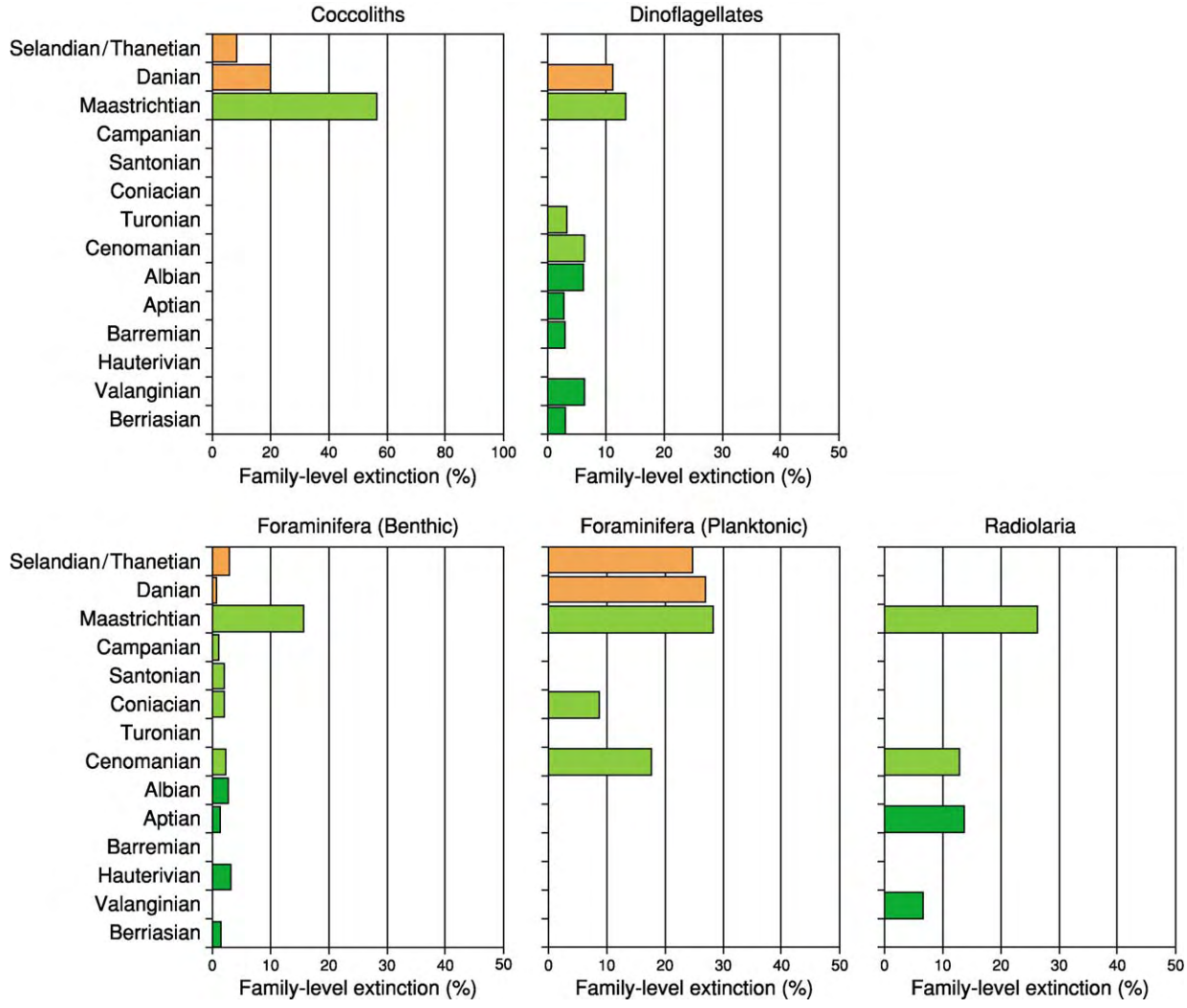


Figure 5 Cretaceous Paleocene, stage level extinction patterns for marine microfossil families. Note change in scale in the coccolith diagram. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

of a larger pattern that encompasses both the Maastrichtian and Paleocene intervals. Overall Maastrichtian extinction intensities across all six marine microfossil groups are less than 20%, suggesting an overall species loss similar to that of the total Maastrichtian estimate (see above).

Interpretation of this Maastrichtian–Danian microfossil record is complicated by several factors. A considerable controversy regarding the correct interpretation of Cretaceous planktonic foraminiferal species routinely found in lowermost Danian sediments continues. Some specialists regard Cretaceous isotopic values obtained from the analysis of particular species' skeletons, along with the widespread chaotic disruption of bedding patterns in lowermost Danian sediments, as indicative of widespread shelf failure at the Maastrichtian–Danian boundary with consequent reworking of Cretaceous species into Danian sediments. This shelf failure was presumably caused by the physical shock of bolide impact along with subsequent earthquakes. Others regard the recovery of Danian isotopic results from other Cretaceous species, the pristine preservation of many millions of Cretaceous microfossil skeletons in Danian sediments (fully comparable with those of undoubted Danian species and distinct from obviously reworked Cretaceous species), and the fact that Danian occurrences of Cretaceous species exhibit a clear biogeographic signal of greater penetration into the Danian in higher latitudes (where the effect of boundary disturbances is known to be reduced) as evidence for the survivorship of some species into Danian times. (There are also specialists who cite phylogenetic criteria as bearing on the survivorship question despite the fact that these arguments can be rejected on both logical and analytic grounds.) Two families are involved in this controversy, Globotruncanidae and Rugoglobigerinidae.

Relatively low Maastrichtian extinction intensities for dinoflagellates and diatoms have been accounted for by noting that the biology of these groups includes resting cyst stages and that these may have enhanced their overall survivorship potential. Species-level data for both groups from the K–T section on Seymour Island, Antarctica do not support this interpretation. Seymour Island cyst-forming dinoflagellates and diatoms exhibit progressive turnover patterns across the K–T boundary, suggesting that extinction-inducing environmental changes were not confined to any single horizon. There is an increase in diatom resting spores in the Upper Maastrichtian interval of this high-latitude section, but this occurs throughout the succession and is not confined to any single stratigraphic horizon. Moreover, since no modern diatom resting spore has been successfully revived after more than two years' dormancy, this sets an inferred maximum duration of environmental disruption that could be tolerated before wholesale extinction of the indigenous diatom flora—for which there is no evidence—would occur.

Marine Invertebrates

Extinctions among marine invertebrate groups exhibit a range of patterns (Figure 6) different from that of marine microfossils. Here, family-level data suggest that poriferans were among the most strongly affected groups, whereas corals exhibit scarcely any effect at all. Between these extremes molluscs, brachiopods, bryozoans, and echinoderms exhibit broadly progressive extinction patterns that rise throughout the Late Cretaceous to a Maastrichtian peak and then fall off into the Paleocene, whereas marine arthropods exhibit a coral-like indifference to whatever factors were driving these long-term changes within other groups.

Given its iconic status within the pantheon of end-Cretaceous victims, the mollusc record is particularly interesting. Throughout the Cretaceous mollusc extinction intensities are comparatively low. The traditional mollusc victim families of Inoceramidae, and the rudistid families Radiolitidae and Hipporitidae all record last appearances in the Maastrichtian, but these three losses alone represent over 20% of all marine bivalve family-level extinctions. Overall, Maastrichtian mollusc losses are subequally split between bivalves (14 families lost), gastropods (8 families lost), and cephalopods (12 families lost). Note that, in contrast to the popular perception of ammonites being iconic K–T victims, at the family level bivalves suffered more than ammonites during the Maastrichtian. It should also be noted that all of these groups suffered the bulk of their losses within the Maastrichtian and not at the K–T boundary itself.

Bryozoans are another group with an especially intriguing Cretaceous extinction record. Numbers of extinct bryozoan families appear to rise in a more-or-less uniform stage-level pattern to a Maastrichtian peak that represents just over 12% of the total Maastrichtian assemblage. Although all three bryozoan orders are affected by this protracted extinction, the ecological hammer appears to have fallen more forcefully on cheilostomes than on ctenostome or cyclostome forms. Interestingly, none of these extinctions are thought to have effected long-term patterns of bryozoan diversification substantially. The relatively comparable level of Maastrichtian extinction intensity exhibited by the bryozoans' sister group, the brachiopods, is a bit misleading in that the data presented in Figure 6 have been expressed as percentages.

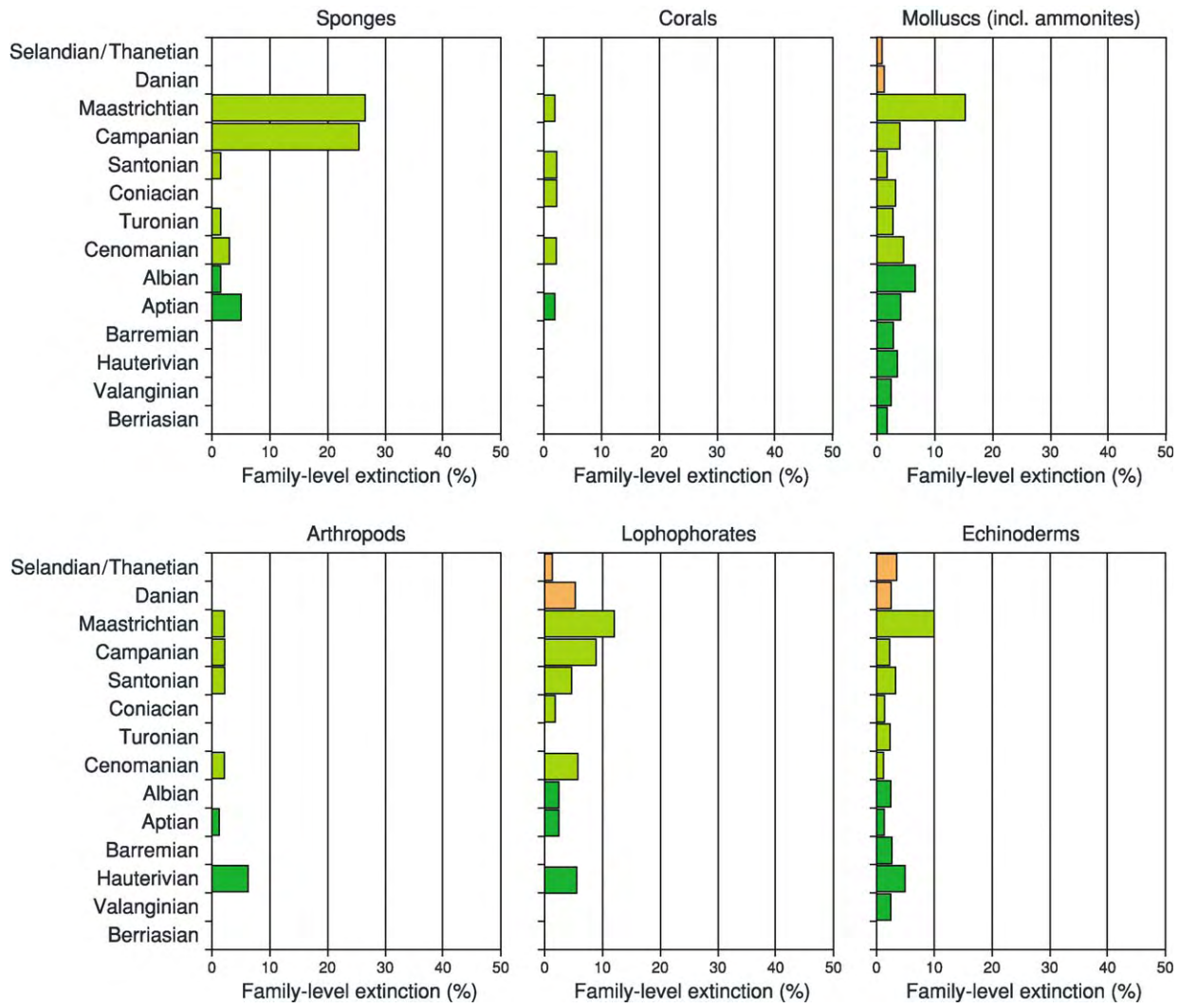


Figure 6 Cretaceous Paleocene, stage level extinction patterns for marine invertebrate families. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

In fact, overall Maastrichtian brachiopod family richness is less than a third that of bryozoans, and the actual number of brachiopod families that became extinct during this stage was comparably reduced. Among echinoderms, irregular echinoid families bore the brunt of the Maastrichtian extinction pulse.

Taken as a whole, it is difficult to avoid the impression that marine invertebrates were being subjected to much longer-term extinction pressures than their protistan counterparts with the Maastrichtian representing the apotheosis of trends that first become evident in the early part of the Late Cretaceous. Moreover, once the Maastrichtian catharsis was reached, it had relatively less overall effect on marine invertebrate family-level diversity than did the protistan reductions. As for the patterns of species-level extinction distributions, despite two decades of intensive study the fine-scale extinction pattern is known only for ammonites in two Late Maastrichtian sections (Seymour Island, Antarctica and Zumaya, Spain). In both cases extinctions observed extinction horizons are not concentrated at the K–T boundary, but rather are spread throughout the entire interval (see Figure 2).

Marine Vertebrates

This group, largely comprising fish, ichthyosaurs, mosasaurs, sauropterygians, and marine turtles, has rarely received much attention in Maastrichtian extinction studies despite the iconic status of mosasaurs and plesiosaurs as stereotypic K–T victims. One sometimes also sees ichthyosaurs mentioned as participants in the end-Cretaceous extinction event, but this is erroneous. The last ichthyosaur became extinct in the Albian. Overall (Figure 7), marine vertebrates

exhibit a distinct and rather abrupt Cenomanian peak, followed by a protracted Late Cretaceous buildup to a subordinate Maastrichtian peak, after which extinction intensities decline throughout the Paleocene. Although it should be noted that the pattern shown in Figure 7 is dominated by fish data, and that marine reptiles do show a much more distinct Maastrichtian extinction peak, it is also true that comparatively few Maastrichtian marine reptile families were present and that these were formed from relatively few genera. Moreover, no marine reptile extinction horizon is known to coincide with the K–T boundary.

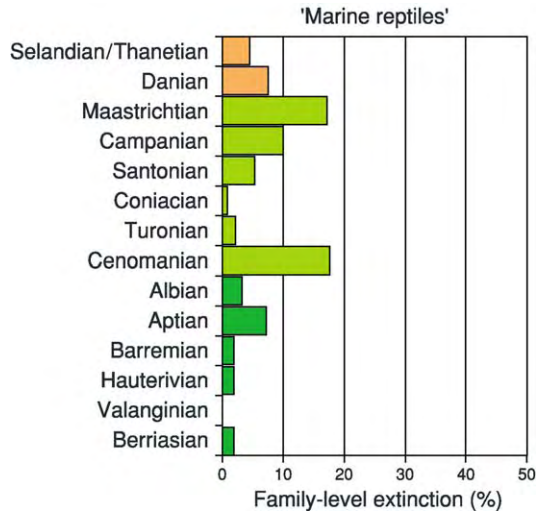


Figure 7 Cretaceous Paleocene, stage level extinction patterns for marine vertebrate families. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

Terrestrial Invertebrates

Relative to comparable marine taxa, freshwater and terrestrial invertebrates (gastropods, bivalves, chelicerates, crustaceans, ostracods, and insects) exhibit very modest levels and episodic patterns of extinction intensity throughout the Cretaceous with the Maastrichtian stage representing a decidedly subordinate intensity peak (Figure 8). Combined data from this ecological realm are inevitably biased by the very large number of insect families that, in all Cretaceous stages, are represented by values well over an order of magnitude larger than those for any other group. Despite this faunal size discrepancy though, Maastrichtian levels of percent extinction intensity for both freshwater and terrestrial mollusc and noninsect arthropod families are both less than those for insect families. The general character of Maastrichtian family-level extinctions for these former groups appears fully consistent with the qualitative notion of background, rather than mass, extinction. It should also be noted in passing that insects suffer their greatest Cretaceous extinction in the Albian rather than in the Maastrichtian.

Terrestrial Vertebrates

The Maastrichtian terrestrial vertebrate extinction record (freshwater fish, amphibians, reptiles, dinosaurs, and mammals) is by far the most diverse of any ecological realm (Figure 9). Even more than their marine counterparts, terrestrial fish faunas provide no evidence for heightened Maastrichtian extinction intensities. Indeed Maastrichtian fish families exhibit the lowest family-level extinction intensity of any Cretaceous stage. This remarkably quiescent

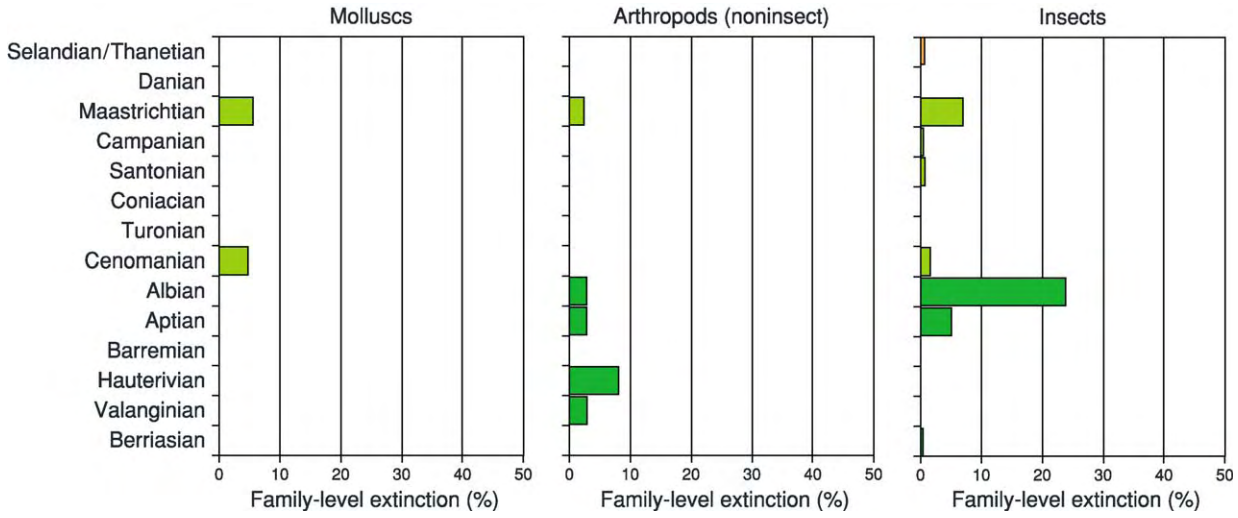


Figure 8 Cretaceous Paleocene, stage level extinction patterns for terrestrial freshwater invertebrate families. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

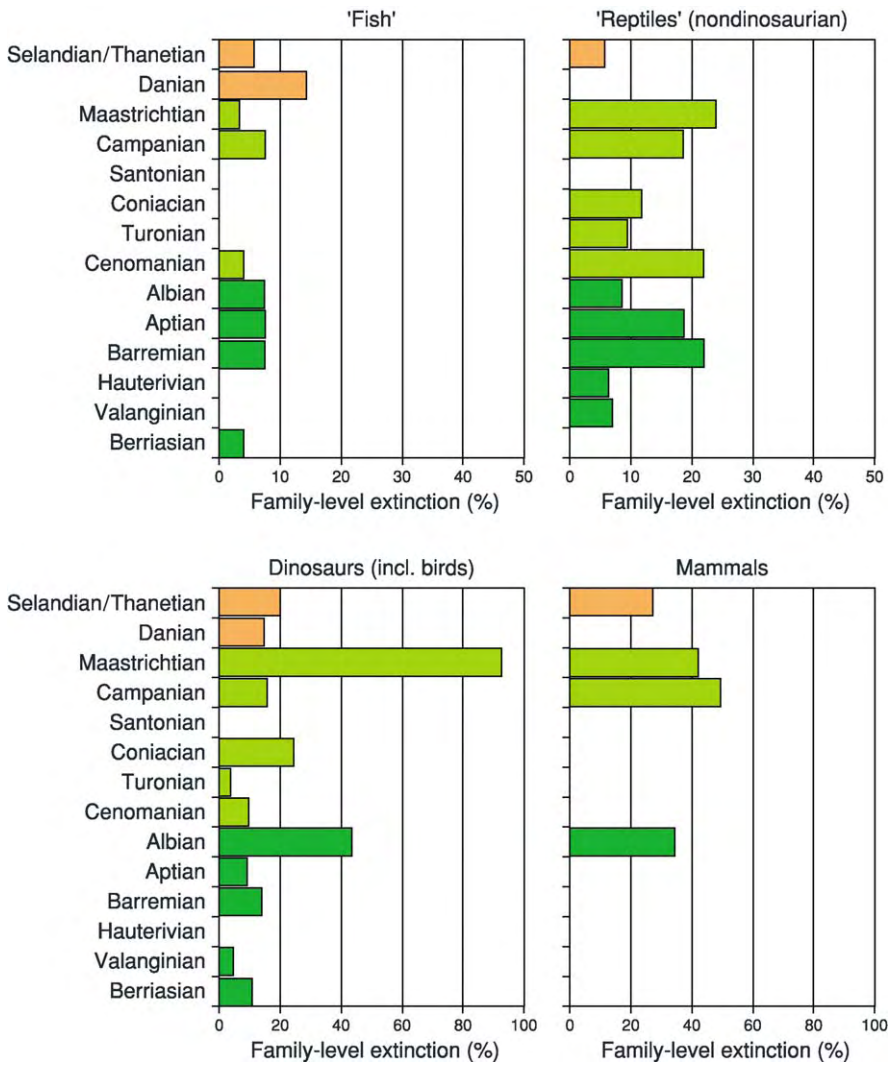


Figure 9 Cretaceous Paleocene, stage level extinction patterns for terrestrial freshwater vertebrate families. Note change in scale in the dinosaur and mammal diagrams. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

record is exceeded only by that of amphibians, which fail to record any family-level extinctions in any Cretaceous stage. Cretaceous extinction levels for non-dinosaurian reptiles are much higher, but neglect to exhibit a pronounced Maastrichtian peak despite the loss of five crocodylomorph families. This pattern is reversed, however, in the Dinosauria with its impressive 21 dinosaur families lost over the course of the Maastrichtian. Supporters of catastrophic dinosaur extinction scenarios hold that most, if not all, of these extinction occurred at the K–T boundary despite the fact that no 'in-place' dinosaur bone has been found within more than a metre of that horizon to date. Finally, as noted for Cretaceous family richness patterns, the Cretaceous mammalian extinction record bears an intriguing similarity to that of the Dinosauria, with coincident Albian and Maastrichtian peaks along with a unique Campanian peak that exceeds that of the Maastrichtian in terms of overall extinction intensity. Although dinosaurs (including birds) and mammals represent the two most severely affected clades among all terrestrial groups, and two of the most severely affected across all ecological realms, it should nevertheless be kept in mind that, in terms of overall extinction intensity, dinosaurs and mammals represent only 10% of all Maastrichtian land-based families and less than 5% of the overall Maastrichtian biota.

Plants

As has been noted elsewhere in the extinction literature, Cretaceous plants appear to have been relatively extinction-resistant (Figure 10). There is a hint of a slight increase in Maastrichtian plant extinction intensities, but nothing that could be spoken of as being extraordinary relative to other Late Cretaceous

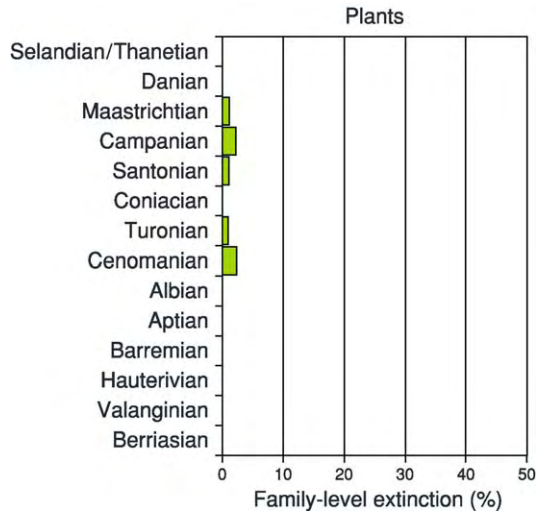


Figure 10 Cretaceous Paleocene, stage level extinction patterns for terrestrial plant families. Data tabulated from Benton (1993) *The Fossil Record 2*. London: Chapman & Hall.

stages. In addition, the pattern of extinctions suggested heightened extinction susceptibility beginning in the Campanian and continuing into the Danian. Over this time interval the Maastrichtian extinctions that do occur are focussed in gymnosperm lineages.

Extinction Causes

There are three single-cause models and one composite cause model currently used to account for the end-Cretaceous extinctions. The former are based on changes in global sea-level, increases in large-province volcanism, and bolide impact.

Sea-Level

Three distinct, short-term, sea-level drops occurred during the Maastrichtian. These were superimposed on a long-term declining trend from the all-Cretaceous sea-level highstand at or near the Cenomanian–Turonian boundary. The oldest of these is an ~100-m drop across the Campanian–Maastrichtian boundary, followed by a slightly larger and more acute drop across the Early-Late Maastrichtian, followed by a significantly larger (~200-m drop) and still more abrupt drop that took place at or just before the end of the Maastrichtian. These sea-level drops are reflected in $\partial^{18}O$ isotopic values that suggest a long-term cooling of global climates throughout the Turonian–Maastrichtian interval (presumably due to decreased albedo and reduced atmospheric CO_2) coupled with extreme sea-surface temperature instabilities in the Maastrichtian. Aside from further reducing the already-reduced area of submerged continental shelf available for colonization and foraging by marine taxa, these sea-level drops would be expected to have resulted in intensified global climate gradients (due to reduced equator-to-pole heat transfer efficiency), increased seasonality, and the development of local centres of bottom water anoxia. Sea-level has long been a suspected cause of many 'mass extinction' events in the geological past. A total of seven major sea-level lowstands have occurred over the past 250 My, and of these three exhibit a stage-level association with peaks in the extinction–intensity curve.

Large Igneous Province (LIP) Volcanism

The Maastrichtian witnessed a particularly large flood-basalt volcanic event when the (then) island continent of India moved over the Reunion mantle plume or 'hotspot' currently located in the southern Indian Ocean. Radioisotopic dates indicate that these Deccan volcanics were erupted in a series of quasi discrete events that range in age from 65.7 ± 1.3 Ma to 65.4 ± 1.3 Ma. These dates suggest that over 1×10^6 km^3 of extrusive volcanics were emplaced over an ~300 000- to 500 000-year interval that straddled the Maastrichtian–Danian boundary (65.5 Ma). The primary climatic effect of these eruptions would have been a brief global cooling (due to increased albedo brought about by injection of SO_2 into the stratosphere), longer-term global warming (due to release of volcanogenic CO_2, possibly along with the melting of frozen methane hydrates on the continental shelves), and acid rain (due to the SO_2). Secondary effects would include enhanced marine anoxia (due to global warming and sluggish marine circulation), disruption of weather patterns (due to the thermal anomaly created by the cooling lava), enhanced local-regional earthquake activity, and increase rates of subareal weathering. The emplacement of large igneous-volcanic bodies such as the Deccan Traps has long been suspected as a cause of many mass extinction events in the geological past. There have been six associations between large LIP volcanic eruptions and extinction-intensity peaks during the past 250 My.

Bolide Impact

Few people interested in science can be unaware of the dramatic discovery that an ~10-km (diam.) comet or asteroid collided with the Earth at the precise time of the Maastrichtian–Danian boundary. Indeed, the iridium (Ir) anomaly in marine and terrestrial sediments that marks this event has been accepted as the primary defining feature of this boundary in temporally complete K–T successions worldwide. Subsequent subsurface geological

investigations, first in cores from boreholes drilled near the town of Chixculub on Mexico's Yucatan Peninsula, and later from regional Bouger gravity surveys, revealed the presence of a 250-km-wide, circular, multiringed basin that corresponds to predictions of the crater size such an impact would be expected to leave. Radioisotopic dates of melted basement rock from the original Chixculub cores confirmed the age of this crater as 65.5 ± 0.1 Ma. In 1996 this date was accepted as the official age of the Maastrichtian–Danian (=K–T) boundary.

The physical and climatic effects of a Chixculub-size impact would be expected to be similar to those of an LIP volcanic event. Unlike a volcanic event, a thermal blast would precede the arrival of a massive shock wave at localities thousands of miles away from the impact site itself. At the same time, a variety of materials would be injected into the stratosphere, resulting in a sharp-but-short global cooling phase, followed by a longer-term global warming. Swarms of major earthquakes would also have been produced locally as the Earth's crust adjusted to the deformation. Some have predicted global darkness brought about by this stratospheric dust injection, plunging the Earth into a months- to years-long night, but recent reviews of this proposed effect have reduced the duration of impact-related global darkening considerably. Acid rain would be produced as a result of both the bolide's passage through the atmosphere and its ejection of a large volume of vaporized evaporitic material from the impact site, though, once again, the severity of this acid-rain phase has been downgraded recently from initial estimates. The initiation of global wildfires as a result of melted ejecta is another part of the original bolide impact scenario that, once again, has been downgraded as a result of subsequent research findings. The remaining, accepted, primary and secondary phenomena (see above) are all remarkably similar to the predicted effects of an LIP volcanic event, with the bolide impact effects capable of being distinguished from the latter only in terms of:

1. different trace element signatures in impact vs volcanogenic glass spherules,
2. the production of different mineralogic artefacts (e.g., multiple sets of shock lamellae in the crystal lattices of mineral grains), and
3. the shorter time period over which the primary and secondary effects would have operated.

With respect to the latter, it should be noted that, although the intensity of the postimpact environmental effects would be greater than those generated as a result of LIP volcanic eruptions, bolide impact effects would also be expected have attenuated at a much more rapid rate over time since—unlike LIP eruption event scenarios—disruptive event renewal would not have been possible. In effect, the important ecological distinction that can be drawn between bolide impact and volcanic disruptions is that of a single, massive, disruptive shock to global ecosystems after which normal conditions are reestablished versus a continuous series of less intense shocks spread over a much longer time interval. Bolide impacts had long been discounted as a major extinction cause in the past but interest in this mechanism has undergone a dramatic turn-around since 1980. Still, of the six major bolide impact events (crater diam. $\geq$50 km) that are known to have occurred over the past 250 My, only three exhibit stage-level associations with peaks in the extinction-intensity curve.

Multiple Causes

Although it has not been defined precisely—especially in terms of positive or negative feedbacks between different disruption sources—a commonly stated body of opinion suggests that the most reasonable stance is to admit that all of the previous mechanisms, as well as others, took place in the Maastrichtian and may have contributed to precipitating the end-Cretaceous extinctions, each in their own way. The interesting feature of this model is not the existence of causal mechanisms but their apparently coincident timing. This model gains credibility not from the detail with which its adherents can currently construct scenarios of predicted environmental effects, but from:

1. the variety of extinction patterns present among the various groups affected by the overall extinction event (see above),
2. the lack of a consistent and/or simple ecological signal among organisms that inhabit different ecological realms,
3. the fact that all local peaks in extinction—intensity curves, regardless of size, are associated with the time series of at least one of these component mechanisms, and
4. the fact that, over the past 250 My, all three large extinction events occur during a time of confluence between two or more of these component mechanisms.

Opinion surveys among research palaeontologists with regard to the most likely cause of the end-Cretaceous extinctions were conducted in 1984 and 1996. On both occasions 'geologists' (=researchers with expertise outside palaeontology) were overwhelmingly in favour of the single-cause, bolide impact explanation for end-Cretaceous extinctions—especially the dinosaur extinctions—whereas palaeontologists (including experts in dinosaur and nondinosaur groups) were overwhelmingly in favour

of the multiple-cause model. If this survey were to be repeated this author is quite sure the results would be similar.

See Also

Atmosphere Evolution. **Biozones**. **Fossil Invertebrates:** Arthropods; Brachiopods; Bryozoans; Corals and Other Cnidaria; Echinoderms (Other Than Echinoids); Molluscs Overview. **Fossil Vertebrates:** Fish; Palaeozoic Non-Amniote Tetrapods. **Microfossils:** Foraminifera. **Plate Tectonics**. **Sequence Stratigraphy**. **Stratigraphical Principles**. **Tertiary To Present:** Paleocene. **Time Scale**. **Volcanoes**.

Further Reading

Archibald JD (1996) *Dinosaur Extinction and the End of an Era: What the Fossils Say. Critical Moments in Paleobiology and Earth History.* New York: Columbia University Press.

Benton MJ (1993) *The Fossil Record 2*. London: Chapman & Hall.

Galvin C (1998) The great dinosaur extinction controversy and the K T research program in the late 20th Century. *Earth Sciences History* 17: 41 55.

Glasby GP and Kunzendorf H (1996) Multiple factors in the origin of the Cretaceous Tertiary boundary: The role of environmental stress and Deccan Trap volcanism. *Geologische Rundschau* 85: 191 210.

Hallam A and Wignall PB (1997) *Mass Extinctions and Their Aftermath*. Oxford: Oxford Science.

Hoffman A and Nitecki MA (1985) Reception of the asteroid hypothesis of terminal Cretaceous extinctions. *Geology* 13: 884 887.

MacLeod N, *et al.* (1997) The Cretaceous Tertiary biotic transition. *Journal of the Geological Society of London* 154: 265 292.

Marshall CR and Ward PD (1996) Sudden and gradual molluscan extinctions in the latest Cretaceous of western European Tethys. *Science* 274: 1360 1363.

Ryder G, Fastovsky D, and Gartner S (1996) *The Cretaceous Tertiary Event and Other Catastrophes in Earth History*, Special Paper 307. Boulder, CO: Geological Society of America.

Sarjeant WAS (1999) Dinosaur extinction: sudden or slow, cataclysmic or climatic? *Geoscience Canada* 23(3): 161 164.

Sepkoski JJ Jr (2002) *A Compendium of Fossil Marine Animals*. Bulletins of American Paleontology. 363. Ithaca, NY: Paleontological Research Institute.

Sharpton VL and Ward PD (1990) *Global Catastrophes in Earth History: An Interdisciplinary Conference on Impacts, Volcanism, and Mass Mortality*, Special Paper 247. Boulder, CO: Geological Society of America.

Shaw A (1964) *Time in Stratigraphy.* New York: McGraw Hill.

Signor PW III and Lipps JH (1982) Sampling bias, gradual extinction patterns and catastrophes in the fossil record. In: Silver LT and Schultz PH (eds.) *Geological Implications of Impacts of Large Asteroids and Comets on the Earth*, Special Paper 190, pp. 291 296. Boulder, CO: Geological Society of America.

Silver LT and Schultz PH (1982) *Geological Implications of Impacts of Large Asteroids and Comets on the Earth*, Special Paper 190. Boulder, CO: Geological Society of America.

Williams ME (1994) Catastrophic versus noncatastrophic extinction of the dinosaurs: Testing, falsifiability, and the burden of proof. *Journal of Paleontology* 68: 183 190.

METAMORPHIC ROCKS

Contents

Classification, Nomenclature and Formation

G Hoinkes and C A Hauzenberger, University of Graz, Graz, Austria
R Schmid, ETH-centre, Zurich, Switzerland

Introduction

Many names for metamorphic rocks are inconsistently used. A recommended classification of these rocks, which is acceptable worldwide, does not yet exist but will, however, be presented in near future by the IUGS Subcommission on the Systematics of Metamorphic Rocks (SCMR). In this chapter, a classification principle and definitions for metamorphic rocks are presented which have been extracted from provisional recommendations of SCMR, presented on their homepage at 'http://bgs.ac.uk/SCMR/' (see at end of this article **Further Reading** list) and from our own considerations.

How to Classify and Name a Metamorphic Rock

Metamorphic rocks originate, if not already of metamorphic origin, from sedimentary and igneous rocks. Therefore, their bulk chemical composition is extremely variable, and because they can have formed under pressures (*see* **Ultra High Pressure Metamorphism**) and temperatures (*see* **Thermal Metamorphism**) ranging between those existing some kilometres beneath the Earth's surface and those existing when rocks start to melt, under the influence of varying fluid compositions or metasomatic processes, the mineralogical composition (*see* **Minerals:** Definition and Classification) is far more variable than that of igneous or sedimentary rocks. This contrasts with the small number of well-known and most frequently used names for metamorphic rocks.

The definitions of these names unfortunately are not based on the same features or properties. Some definitions use purely mineralogical, others structural or a combination of structural and mineralogical properties, others chemical parameters or genetic considerations, or otherwise (Tables 1a, b, and 2). This means that only some of the terms can be used to derive a coherent classification system encompassing all the possible varieties of metamorphic rock types, because such a system requires that each subdivision in that system use only one criterion for each subdivision. If it is not built up in this manner, it will contain many gaps, and will not be able to offer names for rocks falling within such gaps.

The Principal Solution of This Problem

The available rock names are split into three groups, and four major rules are set up on how to use the terms in these groups

Main Specific Rock Names

Eight terms (Table 1a) are applied to frequently occurring metamorphic rock types and in their definition describe the essential mineral mode of these rock types. They are valuable because the nearest macroscopic characterization of a metamorphic rock is achieved by indicating its mineral mode. Therefore, the first rule of the classification system proposed here is that when a rock sample fulfils the requirements of the definition of a main specific rock term listed in Table 1a the pertinent rock name is the correct name for the sample in question. The rule excludes in this case one of the structural root names (see below) being applied. It would give rise to ambiguities, if two different names for one and the same rock type were in use. The rule should, therefore, be strictly followed.

Minor Specific Rock Names

Terms define the mineral mode less well or define other properties of the rock (Table 1b). Here, a second rule similar to the first units is applied, but less strictly.

Table 1a Definitions of the main specific rock names which have to be given preference over the equivalent structural root names schist, gneiss, or granofels

Amphibolite: *Metamorphic rock mainly (to more than 50% vol.) consisting of green, brown, or black amphibole and plagioclase. The modal content of pyroxene is larger than 30% vol. and larger than the percentage of any one of the other mafic minerals. Plagioclase is a major mineral (>5% vol.).*

Other common minerals in amphibolite are quartz, chlorite, epidote, zoisite, biotite, garnet, titanite, scapolite, and calcite. Their presence should be indicated by prefixing them, e.g., *garnet bearing clinopyroxene amphibolite*, where garnet is a minor and clinopyroxene a major constituent. Only in the case of the special type of amphibole or plagioclase being present, these constituents should be prefixed as well (e.g., *bytownite amphibolite*).

Calc-silicate rock: *Metamorphic rock mainly composed of Ca rich silicate minerals such as wollastonite, vesuvianite, diopside hedenbergite, titanite, grossular and andradite rich garnet, prehnite, meionite rich scapolite, zoisite, clinozoisite, epidote, and pumpellyite.*

Less Ca rich silicates such as plagioclase, tremolite, etc., are common additional minerals as well as opaque minerals. Also carbonate (calcite or aragonite ± dolomite) may be present in an amount of up to 50% by vol.

Rocks of similar mineral composition, formed by metasomatism or contact metamorphism, should be classified as skarns or calc silicate hornfelses, respectively.

Eclogite: *Plagioclase free metamorphic rock composed of more than 75% omphacite and garnet, both of which are present as major minerals, the amount of neither of them being higher than 75% vol.*

Granulite: *High grade metamorphic rock, in which anhydrous Fe Mg silicates are dominantly anhydrous. Cordierite may be present and is not counted as either a hydrous or anhydrous mineral. The term should not be applied to ultramafic rocks, calc silicate rocks, marbles, and ironstones. This rule implies that granulites contain more than 10% vol. of feldspar and/or quartz.*

The mineral composition of the rock is to be indicated by prefixing all of the major constituents present, e.g.:

- Biotite sillimanite garnet plagioclase granulite: rock sample composed of the major minerals plagioclase, garnet, sillimanite, and biotite, where plagioclase?garnet?sillimanite?biotite.
- Pyroxene plagioclase granulites ± garnet, amphibole: rock series in which each individual sample contains as major minerals pyroxene (cpx and/or opx) and plagioclase, and sometimes also garnet and/or amphibole.

The following collective terms for certain groups of granulites are proposed:

- Granulites containing the phase assemblage orthopyroxene, quartz, K feldspar or mesoperthite and ±plagioclase may be called *charnockitic granulites*.
- The term *mafic granulites* embraces all granulites containing pyroxene±amphibole±garnet as the dominant ferromagnesian phase assemblage. They contain plagioclase, rarely K feldspar and quartz, and very rarely aluminosilicates (only at high P).
- In contrast, those granulites containing garnet±biotite±cordierite as the dominant ferromagnesian phase assemblage may be called, due to the lack of a better term, *felsic granulites*, even if they are frequently restitic and may contain a large amount of garnet. They contain plagioclase and/or K feldspar or mesoperthite, ± quartz and ±aluminosilicates.
- The collective term *quartzo feldspathic granulites* or *quartz feldspar granulites* should only be applied, if all members of the rock group contain both of the minerals quartz and feldspar.

Marble: *Metamorphic rock containing more than 50% vol. of calcite and/or dolomite and/or aragonite.*

Pure marble contains more than 95% vol. of these carbonate minerals, whereas the remainder are classified as *impure marble*.

Phyllite: *Metamorphic rock, in which the individual grains are large enough to be seen by the unaided eye (>0.1 mm) and which is characterized by a lustreous sheen and a well developed schistosity resulting from the parallel arrangement of phyllosilicates.*

Phyllite is usually of low metamorphic grade.

Pyriclasite: *Metamorphic rock mainly (to more than 50% vol.) consisting of pyroxene (cpx and/or opx) and plagioclase. The modal content of pyroxene is larger than 30% vol. and larger than the percentage of any one of the other mafic minerals (>5% vol.).*

Other common minerals in pyriclasite are garnet and/or amphibole and/or biotite. The presence of them may be indicated by prefixing them, e.g., *biotite bearing garnet pyriclasite*. Only in the case of the special type of pyroxene or plagioclase being present, these constitutents should be prefixed as well, e.g., *cpx pyriclasite*, if only clinpyroxene is present, or *cpx opx pyriclasite*, if both pyroxenes are present.

The term was created and first defined by Berthelsen in 1960.

Slate: *Metamorphic rock, in which the individual grains are too small to be seen by the unaided eye (<0.1 mm) and in which the schistosity is developed on the grain scale.*

Slate is usually of very low metamorphic grade and rich in phyllosilicates.

A minor specific rock name may be used in cases where it is appropriate and makes sense to do so.

Structural Root Names

The three terms 'schist', 'gneiss', and 'granofels' are defined in (Table 2). The third rule simulates that for a rock sample, for which no major or minor specific rock term can or should be applied, one of the three structural root names schist, gneiss, and granofels can to be applied. These names are delineated from each other by one single criterion, the quality of schistosity developed within them (Table 2). If the schistosity is well developed, it is a schist, if poorly developed, it is a gneiss, and if there is no schistosity present, it is a granofels. Thus, any metamorphic rock, for which no specific name is available, can be described by one of these three terms. In order to indicate the mineral mode of the sample, all minerals present in the rock in a volume of more than 5% (= major minerals) have to be arranged in front of the structural root name

Table 1b List of minor specific rock names names which need not be given preference over the equivalent structural root names schist, gneiss, or granofels, because they only poorly define the modal composition of a rock and should therefore only be applied when it is appropriate to do so. The list only presents some examples of minor specific rock names and is not complete

1.Rock names defined by mode and rock colour:
Blueschist, Greenschist, Greenstone, Whiteschist.
2.Rock name refering to the protolith and its alteration:
Spilite.
3.Genetic to semi genetic rock names:
Restite, Rodingite, Skarn.
4a.Names for cohesive fault rocks:
Mylonite, Protomylonite, Mesomylonite, Ultramylonite, Augen mylonite, Blastomylonite, Phyllonite. Pseudotachylite.
4b.Names for cohesive or incohesive fault rocks:
Cataclasite, Protocataclasite, Mesocataclasite, Ultracataclasite, Fault breccia
4c.Names for incohesive fault rocks:
Fault gouge
5.Names for composite rocks:
Migmatite
6.Names for contact metamorphic rocks:
Hornfels
7.Names for impact rocks:
Impact breccia, Impact melt rock, Impactite

Table 2 Definitions of the three structural root names, and of the structural terms used in these definitions

Schist: *A metamorphic rock displaying a schistose structure.*
The term schist may also be applied to a rock displaying a linear rather than a planar fabric, in which case the expression 'lineated schist' is applied
For phyllosilicate rich rocks, the term schist is reserved for medium to coarse grained varieties, wheras finer grained rocks are termed phyllites or slates.
The mineral composition of the rock is to be indicated by prefixing all of the major constituents present.
Gneiss: *A metamorphic rock displaying a gneissose structure.*
The term gneiss may be also applied to a rock displaying a linear rather than a planar fabric, in which case the expression 'lineated gneiss' is applied.
The mineral composition of the rock is to be indicated by prefixing all of the major constituents present
Granofels: *A metamorphic rock displaying a granofelsic structure.*
The mineral composition of the rock is indicated by prefixing all of the major constituents present
For granofels containing layers of different composition the expression 'layered (or banded) granofels' may be used.
Note: In English there was no term for a massive rock devoid of a planar or linear fabric, until R. Goldsmith proposed, in 1959, the name granofels for such a rock type.
Schistosity: *A type of foliation produced by metamorphic processe, characterised by the preferred orientation of inequant mineral grains or grain aggregates.*
A schistosity is said to be well developed if inequant mineral grains or grain aggregates are present in a large amount and show a high degree of preferred orientation.
If the degree of preferred orientation is low or if the inequant grains or grain aggregates are only present in small amounts, the schistosity is said to be poorly developed.
Schistose structure: *A type of structure characterized by a schistosity which is well developed, either uniformly throughout the rock or in narrowly spaced repetitive zones, such that the rock will split on a scale of one cm or less.*
Gneissose structure: *A type of structure characterized by a schistosity which is either poorly developed throughout the rock or, if well developed, occurs in broadly spaced zones, such that the rock will split on a scale of more than one centimetre.*
Granofelsic structure: *A type of structure resulting from the absence of schistosity, such that the mineral grains and aggregates of mineral grains are equant, or if inequant, have a random orientation.*
Mineralogical or lithological layering may be present.

chosen, in order of increasing modal abundance, and hyphenated amongst themselves, but not with the rock name. Optionally, minor minerals, present in a volume of $\geq 5\%$, may be mentioned by prefixing them with '…. bearing' (see Flow Chart (Figure 1)). This comprises the fourth rule of the classification system.

A flow chart How to name a Metamorphic Rock is presented in Figure 1, which when followed, will enable each Earth scientist to find the correct name for each metamorphic rock. The flow chart consists of subsequent questions, and on each sides of the questions, comments which help in gaining the right

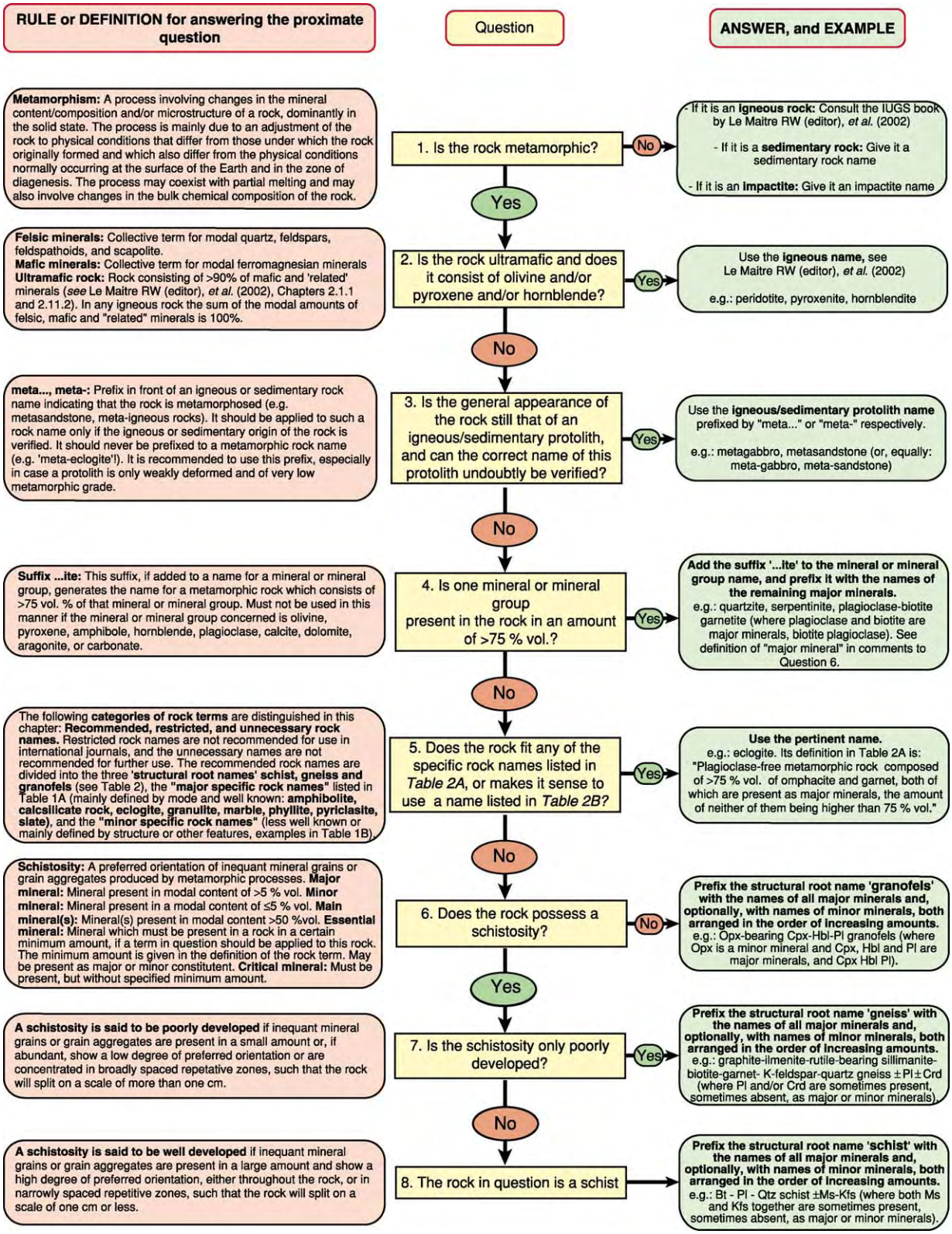


Figure 1 Flow Chart 'How to name a Metamorphic Rock'.

Table 3 Terms for structures partially used in the rock definitions of this chapter

Structure: *The arrangement of the parts of a rock mass irrespective of scale, including spatial relationships between the parts, their relative size and shape and the internal features of the parts.*

Dependent on the scale of the feature, the terms micro (thin section or smaller scale), meso (hand specimen scale), and mega (outcrop or larger scale) can be used as a prefix.

Fabric: *The relative orientation of the parts of a rock mass.*

This is commonly used to refer to the crystallographic and/or shape orientation of mineral grains or groups of grains, but can also be used on a larger scale. Preferred linear orientation of the parts is termed linear fabric, preferred planar orientation planar fabric and the lack of a preferred orientation is referred to as random fabric.

Augen: *Distinctive large lenticular mineral grains or grain aggregates in a finer grained matrix of a rock, having the shape of an eye in cross section, around which the foliation of the matrix is wrapped, commonly forming symmetric or asymmetric trails.*

Cleavage: *The property of a rock to split along a regular set of parallel or sub parallel closely spaced surfaces.*

More than one cleavage may be present in a rock.

Fault: *A dislocation surface along which there has been displacement of the rock mass on one side relative to the one on the other side.*

Foliation: *Any repetitively occurring or penetrative planar feature in a rock body. Examples include:*

- *layering*
- *preferred planar orientation of inequant mineral grains (planar fabric)*
- *preferred planar orientation of lenticular or elongate grain aggregates.*

More than one kind of foliation with more than one orientation may be present in a rock. Foliations may become curved or distorted. The surfaces to which they are parallel are called s surfaces. Where possible, the type of foliation should be indicated.

Fracture: *A general term for any surface of break in a rock mass, whether or not it causes displacement, due to mechanical failure.*

Fracture includes cracks, joints, and faults.

Joint: *A single fracture in a rock with or without a small amount (<1 cm) of either dilational or shear displacement.*

Joints may be infilled during or after their formation.

Layer: *One of a sequence of near parallel, tabular shaped rock bodies.*

Layering: *The feature displayed by layered rocks.*

Band: *Used as a synonym for layer.*

Banding: *Used as a synonym for layering.*

Lineation: *Any repetitively occurring or penetrative visible linear feature in a rock body. This may be defined by:*

- *alignment of the long axes of elongate mineral grains (= mineral lineation)*
- *alignment of elongate mineral aggregates*
- *alignment of platy mineral grains parallel to and around an axis*
- *parallelism of hinge lines of small scale folds (= crenulation lineation)*
- *intersection of two foliations (= intersection lineation)*
- *slickenside striations or fibres.*

More than one kind of lineation, with more than one orientation, may be present in a rock. Lineations may become curved or distorted. The lines to which they are parallel are called l lines. Where possible the type of lineation should be indicated.

answers. The comments comprise definitions or rules to be taken into consideration, and a practical example to each of the questions. Complementary information to the individual Questions 1 to 8 of the flow chart is given below. Terms for structures used in some of the rock definitions of this chapter are presented and defined in Table 3.

Additional Information to Questions 1 to 8 of the Flow Chart

Question 1: Is the Rock Metamorphic?

There are two possible cases in which it may be difficult to answer this question, especially when this has to be done after macroscopic inspection of a rock, and before thin section or analytical data are available:

i. The rock looks like an unmetamorphosed sedimentary or igneous rock but may, in fact, be weakly metamorphosed. When in doubt as to whether it is metamorphosed or not, choose the name of the protolith. When, during later laboratory work, it can be established that the rock is in fact metamorphosed, prefix the protolith name with 'meta…' or 'meta-', respectively (see comment in the flow chart to Question 3).

ii. The rock is mafic or ultramafic, the mafic minerals are olivine and/or pyroxene and/or amphibole, the microstructure is granoblastic, and a schistosity is not or only weakly developed (see definition of schistosity in the comment to Question 6 in the flow chart). It might be an igneous rock which was emplaced at deep crustal levels, where it crystallized and/or re-equilibrated at ambient pressures and temperatures. At these pressures and temperatures, igneous and metamorphic microstructures and phase assemblages converge, and thus the following is proposed:

- When the rock is ultramafic, use the pertinent ultramafic rock name in the IUGS-publication by Le Maitre (ed.) and others, published in 2002. For

the definition of 'ultramafic rock', see comment to Question 2 in the flow chart.
- When the rock is mafic; answer Question 1 with 'yes' and proceed down the flow chart. You will then have the choice to either use an igneous name prefixed by 'meta-' (Question 3) or call your rock a 'granofels' (Question 6).

iii. The rock is an impactite. Even if it is metamorphic, first have a look at the paper by Stoeffler and Grieve about impactites, on the homepage of SCMR at http://bgs.ac.uk/SCMR.

Question 5: Does the Rock Fit Any of the Specific Rock Names Listed in Table 1a, or Does it Make Sense to Use a Name Listed in Table 1b?

i. A name for a metamorphic rock should at best characterize its modal mineral contents as well as its structure. Whereas the structure can easily be described by prefixing specific rock names with schistose, gneissose, or granofelsic, indication of the minerals which are present in a metamorphic rock and of their relative amounts is less easy to perform. Therefore, the specific rock names listed in Table 1a, which are defined by mineral content and which are well known, are important and have, therefore, to be given preference over the the equivalent structural root names, schist, gneiss, or granofels. As an example: a rock containing the essential minerals plagioclase and amphibole in the amounts defined in the definition of amphibolite (Table 1a) should always be called amphibolite, and never an amphibole-plagioclase gneiss. The names in Table 1a should be prefixed by those major minerals, which are not essential minerals (see comment to Question 6 in the flow chart).

ii. The specific rock names in Table 1b are defined using structural or other features or properties. The list is by far from complete. If such special features or properties should be highlighted in the name rather than mineral content, such rock names may be used, giving them preference over the the equivalent structural root names, schist, gneiss or granofels. As an example: during field observations the term mylonite may be applied to a fault rock which is derived from a biotite-quartz-plagioclase gneiss. In a hand specimen collection of a museum, the same rock may be called a very fine grained biotite-quartz-plagioclase schist, with or without mentioning that it originates from a fault zone.

Questions 6–8:

i. *(i) Schistosity as a parameter for the subdivision of remaining rocks into schist, gneiss, and granofels:* It is obvious that the parameter 'perfection of schistosity' used for the subdivision into schist, gneiss, and granofels is only a qualitative one. Mainly rocks at the transition 'well developed' and 'poorly developed schistosity' will be qualified by some colleagues as schists, and by others as gneisses. Fortunately, no severe complications are expected to arise from this.

ii. *Mentioning the mineral contents of metamorphic-rocks when applying structural root names:* The main information about a metamorphic rock comes from its mode which is semiquantitatively fixed. This is done by arranging the names of the major minerals in order of increasing modal abundance and hyphenated amongst themselves in front of the structural root name or, if not present in each individual sample of a group of rocks, behind the root name, preceeded by '±' and separated by a comma, or if not co-existing, by a hyphen, respectively. It is very important that all major minerals are prefixed with reference to a single rock sample or all those major minerals consistently occurring in each specimen of a group of rocks. Ambiguities would arise, if this rule was not obeyed.

The presence of minor minerals (present in a volume of $\leq 5\%$ by volume) may be accounted for by prefixing them, hyphenated with the suffix '-bearing', either in front of the major minerals or behind the structural root name. When not present in each sample of a set of rocks, a '±' is set in front of the respective mineral or the paragenesis of minerals.

SCMR is preparing a list of abbreviations of mineral names containing the abbreviations already published, as well as additional ones. These abbreviations may preferentially be used when prefixing minerals in front of rock names.

Formation of Metamorphic Rocks

Introduction

In petrology, the term metamorphism refers to a change in a rock's mineralogy, structure and/or composition during geological processes that occur predominantly in a solid state under conditions between diagenesis and large-scale melting. Most metamorphic rocks retain some of their parental heritage such as chemical composition and to some extent primary structural features such as bedding. The process of metamorphism is driven by changes in physical and/or chemical conditions, usually as a result of large-scale geological processes (plate tectonics).

Temperature Increasing temperature promotes recrystallization of fine-grained rocks as well as

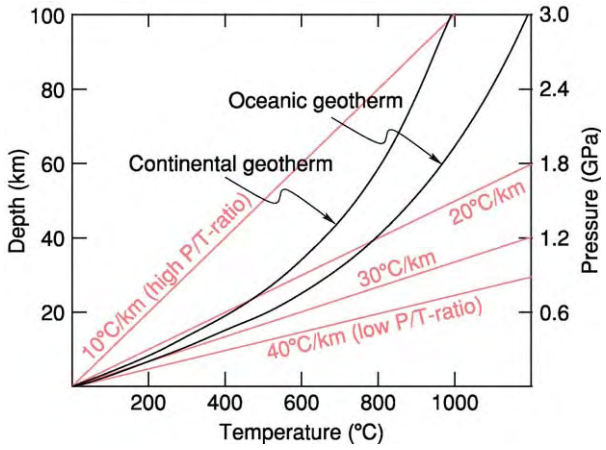


Figure 2 Typical geothermal gradients observed in the Earth's crust. Geotherms in oceanic crust are higher than in continental crust.

the crystallization of new minerals. The temperature experienced by a rock is controlled by the production, conduction, and advection of heat during geological processes (*see* **Thermal Metamorphism**). The main source of heat production in the lithosphere is radioactive decay, which is typically some microwatts per cubic metre. Minor contributions come from mechanical (e.g., friction) and chemical (latent heat of reaction) processes. The transport of heat within the lithosphere occurs by conduction and/or advection. The first describes the diffusive heat transport, the latter the active transport of heat, for example, by magmatic intrusions or fluids. Geothermal gradients describe the temperature increase with depth (Figure 2). Typical geothermal gradients are in the range of 15–30°C km.

Pressure Pressure increases with depth due to the weight of overlying rocks and is called lithostatic pressure (*see* **Ultra High Pressure Metamorphism**). The relationship between depth and pressure ($P = N/m^2$) is given by the density (ρ) of the rock ($\rho = m/V$). The density of crustal rocks is ~2500–2800 kg/m^3; that of mantle rocks ~3100–3300 kg/m^3. Thus, pressure increases by ~0.027 GPa (= ~270 bar) per kilometre in the crust and by ~0.032 GPa (= ~320 bar) per kilometre in the mantle.

Fluid During most metamorphic processes, an intergranular fluid of varying chemical composition is interacting with the solid phases of a metamorphic rock. The pressure of the fluid is generally considered to be close to the lithostatic pressure except for porous rocks in shallow crustal levels where the fluid forms a continuous network extending to the Earth's surface. In that case, the fluid pressure is equivalent to the overlying fluid weight, which is the hydrostatic pressure. Intergranular metamorphic fluids are usually dominated by H_2O. Additionally, CO_2 and minor amounts of CH_4 and N_2 as well as NaCl may be present in the fluid phase. Fluids can come from meteoric sources, juvenile magmatic sources, rocks dehydrating during prograde metamorphism, and trapped sedimentary brines. Fluids play an important role in enhancing the speed of metamorphic reactions, crystallization, and crystal growth, and in material transport within the crust.

Time The duration of metamorphism has an important influence on the formation of metamorphic rocks. The shortest period of time is realised by impact metamorphism (*see* **Impact Structures**), which is completed within seconds. Contact metamorphism can last from tens of thousands of years to almost a million years. Typical time spans for regional metamorphism are in millions to tens of millions of years. During this time, the rocks are subject to changes in temperature, pressure, and possibly fluid composition. Usually the structure and minerals formed at the highest temperature and pressure are preserved. However, during exhumation some late-stage mineral forming reactions may lead to a retrograde overprint of the peak assemblages.

Classification of Metamorphism

Regional metamorphism (*see* **Regional Metamorphism**) occurs over wide areas, affects large volumes of rocks, and is associated with tectonic processes such as plate collision and crustal thickening (orogenic metamorphism) and ocean-floor spreading (ocean-floor metamorphism). During the subsidence of large basins, sedimentary piles can be affected by PT conditions of 200–400°C and 0.05–0.2 GPa. These conditions mark the transition from diagenesis to metamorphism and are named burial metamorphism.

i. ***Orogenic Metamorphism*** is associated with various phases in the course of an orogenic cycle and involves compressional and extensional regimes. The pressure–temperature conditions cover a wide range (300–1000°C, 0.3–3 GPa), depending on the specific mountain building processes.
ii. ***Ocean-floor metamorphism*** is caused by circulating hot seawater, resulting in hydration of the formerly water-free magmatic mineral assemblages of the newly formed oceanic crust at mid-ocean ridges. The temperature range may vary considerably (200–700°C), depending on the proximity of the spreading centre. Pressure is typically low (<0.5 GPa).

Local metamorphism occurs in relatively small areas around magmatic intrusions (contact metamorphism), meteorite impacts (impact metamorphism), or certain fault zones (dislocation metamorphism).

i. ***Contact metamorphism*** is caused by igneous intrusions as a result of the thermal effects of hot magma on the surrounding cooler country rock. Temperatures may be up to 1000°C at relatively low pressures (0.05–0.5 GPa). Pyrometamorphism is a special form of contact metamorphism at very high temperatures (>1000°C) and very low pressures (<0.1 GPa), affecting small fragments of country rock in a volcanic or subvolcanic environment.
ii. ***Hydrothermal metamorphism*** is caused by hot H_2O-rich fluids. Metasomatism is commonly associated with this type of metamorphism.
iii. ***Impact metamorphism*** is due to an enormous amount of kinetic energy which is released to surface rocks by a large-scale meteorite (asteroid, or comet) impact. Extremely high temperatures partly vaporise and melt the rocks. The impact also produces some ultra-high pressure minerals (*see* **Ultra High Pressure Metamorphism**), such as stishovite, usually not found in the Earth's crust and also lead to amorphous (isotropic) and other anomalous behaviour of minerals.

Metamorphic Processes and Reactions

Mineral assemblages observed in metamorphic rocks are directly related to their chemical composition (X) and to a specific temperature (T) and pressure (P). Rocks can be considered as mechanical mixtures of phases (minerals, melts, fluids, gases), which can be mixtures of phase components. As an example, we will look at a rock which consists of the minerals garnet (grossular; $Ca_3Al_2Si_3O_{12}$), kyanite (Al_2SiO_5), and quartz (SiO_2). At a given temperature and pressure, these three mineral phases have a unique molar Gibbs free energy (G), a quantity that describes the energy state of the rock. The Gibbs free energy of the rock is the sum of the Gibbs free energies of its phases, in this case garnet, kyanite, and quartz. However, the same chemical system can be described by a single mineral, the plagioclase end-member anorthite ($CaAl_2Si_2O_8$), if the following relationship is applicable:

$$1\,\text{Grossular} + 2\,\text{Kyanite} + 1\,\text{Quartz} = 3\,\text{Anorthite} \quad [1]$$

According to the thermodynamic laws, the mixture with the lowest (most negative) total Gibbs free energy is the stable mixture, while the other mixture is metastable. Within a P-T diagram (*see* **Metamorphic Rocks: PTt - Paths**) we find exactly one line (= univariant reaction curve), where the Gibbs free energies of the assemblage garnet, kyanite, and quartz, and of anorthite are equal (eqn [2]; **Figure 3**). Below this reaction curve anorthite has a lower Gibbs free energy than 1 garnet + 2 kyanite + 1 quartz and hence is stable (eqn [3]) above the reaction curve 1 garnet + 2 kyanite, + 1 quartz are stable.

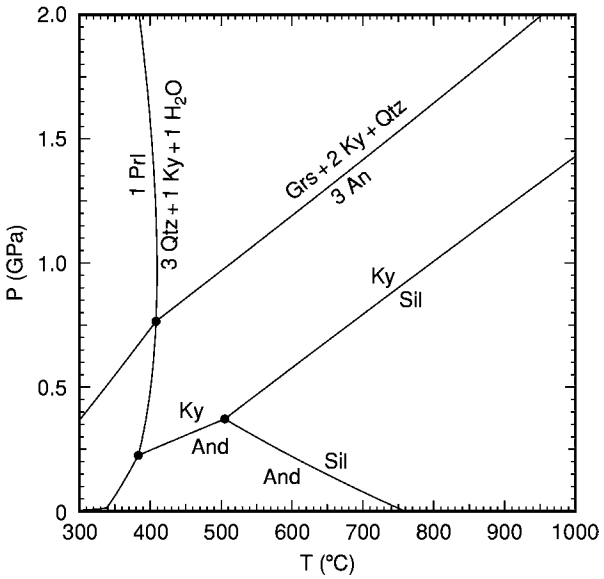


Figure 3 PT diagram showing the reactions Grossular (Grs) + 2 Kyanite (Ky) + Quartz (Qtz) 3 Anorthite (An) and Pyrophyllite (Prl) 3 Qtz + Ky + H_2O. In addition, the stability fields of the Al_2SiO_5 phases Ky, Andalusite (And), and Sillimanite (Sil) are shown.

$$\Delta G_r = 3G_{An} - G_{Grs} - 2G_{Ky} - G_{Qtz} = 0 \quad [2]$$

$$\Delta G_r < 0 \quad [3]$$

The slope of a reaction in a P-T diagram can be approximated by a simple relationship of change of reaction volume (ΔV) and reaction entropy (ΔS):

$$dP/dT = \Delta S_r/\Delta V_r\,(\text{Clausius} - \text{Clapeyron equation}) \quad [4]$$

Reactions involving only solid phases (minerals) phases tend to have a relatively large change in volume (ΔV_r) compared to their change in entropy (ΔS_r) and show relatively flat-lying reaction curves in a P-T diagram (eqn. [1], **Figure 3**). Reactions involving a fluid phase (e.g., H_2O) are different. As an example, take the pyrophyllite (Prl) dehydration reaction (**Figure 3**):

$$1\,\text{Prl} = 3\,\text{Qtz} + 1\,\text{Ky} + 1\,H_2O \quad [5]$$

The volume of H_2O decreases significantly with increasing pressure so the slope of the reaction curve can

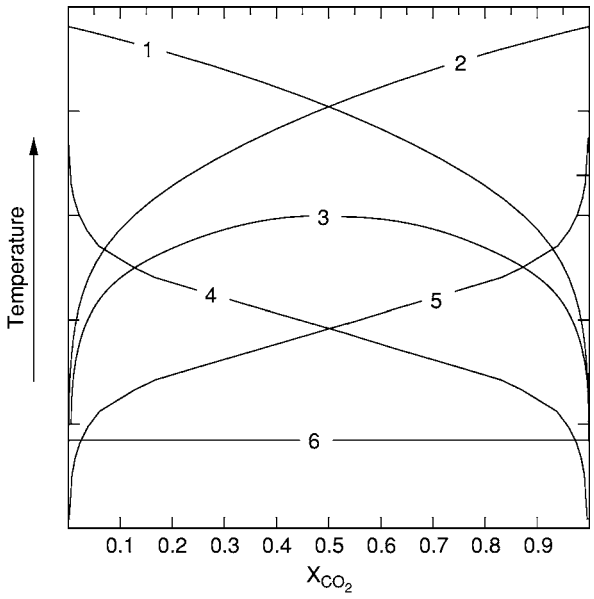


Figure 4 Different reaction types in T X_{CO_2} space.

become negative. Reactions involving only an exchange of cations (e.g., the exchange of Fe and Mg between garnet and biotite) typically have a small ΔV_r and a large change in entropy (ΔS_r), and hence a steep slope in *P-T* space.

The stability of carbonates and hydrous silicates is not only influenced by P and T, but also by the fluid composition in terms of CO_2 and H_2O (X_{CO_2}) which can be shown in a T-X_{CO_2} diagram. The reactions that take place in marbles and carbonate-bearing rocks can be grouped into six types, each of which has a distinctive shape to the equilibrium curve on a *T-X* diagram (Figure 4): (i) dehydration reactions; (ii) decarbonation reactions; (iii) dehydration-decarbonation reactions; (iv) carbonation-dehydration reactions; (v) hydration-decarbonation reactions; and (vi) fluid-absent reactions).

Metamorphic Structures

Metamorphic rocks can form in many different stress environments and at different heating rates. They contain structures ranging from massive to highly foliated and show extremely fine to very coarse grain sizes. Recrystallization and/or mineral growth during metamorphism are the main mechanisms, producing structures that are diagnostically important in recognizing metamorphic rocks. Minerals formed by metamorphic processes are called blasts. Porphyroblasts occur as coarse grained crystals in a finer grained matrix (Figure 5A). Poikiloblasts contain numerous inclusions of other minerals (Figure 5B).

Structural criteria for determining the relative ages for deformation and metamorphic mineral growth are characterized by the relationship between the internal schistosity (S_i) within the mineral blasts and the external schistosity (S_e) of the dominant matrix of the rocks. Within polydeformed areas, it is necessary to explain the relation between the mineral growth and the respective deformational phase. Pre-deformational growth is defined, when S_i is discordant to S_e, or when a break between S_i and S_e develops. During the younger deformational overprint (S_2), the foliation in the matrix often wraps around the blasts and pressure shadows develop (Figure 5C). Syn-deformational growth is defined by a blast that consist of sigmoidal, spiral shaped inclusion patterns, that develope during rotational growth by shearing along the schistosity planes (Figure 5D). The shape of the inclusion patterns are, therefore, best indicators for establishing the shear direction during metamorphic mineral growth. Figure 5D shows no pressure shadows around the blasts and a continuously developed inclusion pattern from S_i (S_1) to S_e (S_2). Post-deformational growth shows blasts that contain inclusion trails that are parallel or continuous to the external schistosity ($S_i = S_e$) (Figure 5E). Corona structures are observed when a mineral breaks down to a new phase or phase assemblage which prevents the progress of reaction. This is commonly observed in granulite-facies rocks. Symplectites develop when a mineral becomes unstable over a very short period of time, which is the case in the exhumation of eclogites (Figure 5G).

Metamorphism of Different Protoliths

Protoliths (other than already metamorphosed rocks) may be of either magmatic or sedimentary origin. For simplicity, magmatic protoliths may be divided into three different groups: ultramafic, mafic, and quartzo-feldspathic rocks. The most important sedimentary protoliths are pelitic (clay-stones), quartzo-feldspathic, carbonate rocks and mixtures of the three groups, such as marls (carbonate-rich pelitic sediments to clay-rich carbonates). Figure 6 shows metamorphic rock names in relation to protolith and grade of metamorphism.

Ultramafic protoliths are mainly peridotites and related rocks from the Earth's mantle and reflect a rather simple MgO-rich silicate composition. The most abundant metamorphic products are serpentinites.

Mafic protoliths are mainly mafic igneous rocks (basic to intermediate intrusive and extrusive rocks) and to a much lesser extent marls of unusual composition similar to basalts. With increasing metamorphic grade they result in very significant metamorphic rock types, depending on the *P/T*-ratio (Figure 2). Low to

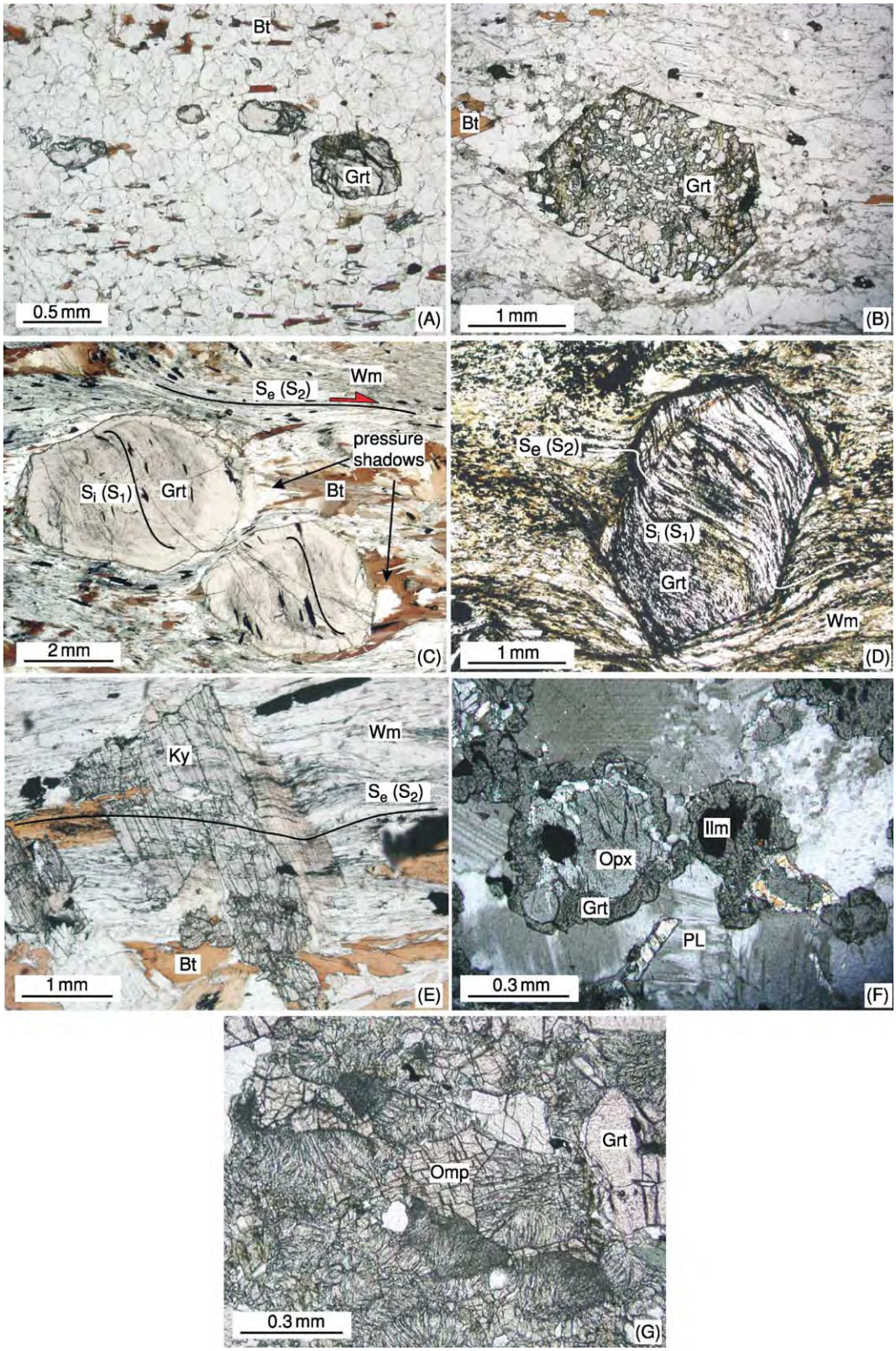


Figure 5 Typical mineral structures observed in metamorphic rocks; (A) A garnet porphyroblast occurs in a finer grained granoblastic matrix; (B) Poikiloblastic garnet containing numerous inclusions of other minerals; (C) Pre deformational garnet porphyroblasts is indicated by a discordant internal foliation S_i (S_1) in garnet. During the younger deformational overprint (S_2), the foliation in the matrix S_e commonly wraps around the blasts and pressure shadows develop; (S) Syn deformational growth is defined

Continued

medium P/T-gradients lead to greenschists, amphibolites, mafic granulites, and/or eclogites, whereas high P/T-gradients are responsible for blueschists and eclogites.

Quartzo-feldspathic protoliths are either represented by intermediate to acid magmatic rocks (intrusive and extrusive) or by clastic sediments derived from these magmatites. If derived from sediments, they form metamorphic schists, paragneisses, migmatites and felsic granulites with increasing grade. If derived from granitoids, the metamorphic products are meta-granitoids, orthogneisses, migmatites, and felsic granulites with increasing grade.

Pelitic protoliths are mainly composed of clay minerals derived from weathered and eroded continental crust. With increasing metamorphic grade, they are transformed to slates, phyllites, mica-schists, and granulites.

Carbonate rocks mainly originate from biogenic sediments. Pure carbonate sediments are SiO_2-free. The metamorphic products are marbles, composed of calcite and/or dolomite. Impurities by SiO_2-rich clastic sediments result in silicate marbles during metamorphism.

Marls may resemble mafic protoliths in terms of chemical composition but usually differ by significant amounts of $CaCO_3$. Low-grade metamorphic marls are calcareous micaschists, whereas the formation of Ca-rich silicate minerals during higher-grade metamorphism results in calcsilicates.

Diagnostic Mineral Assemblages with Prograde Metamorphism

Ultramafic rocks

The primary mineral assemblage of peridotite is represented by forsterite–enstatite±diopside. In order to enable prograde metamorphism, these anhydrous assemblages must first be hydrated to serpentinites in the course of ocean floor metamorphism. The starting assemblage for prograde metamorphism of diopside-free peridotite is the low temperature form of serpentine (chrysotile), brucite, and/or talc, depending on the amount of enstatite and forsterite in the protolith. At about 200°C, antigorite starts replacing chrysotile and at 300°C antigorite is the only serpentine phase stable up to more than 500°C, where it breaks down to forsterite and talc. This assemblage may be replaced by anthophyllite at slightly more than 600°C and low pressures. From 670°C, the peridotite assemblage forsterite and enstatite is also the stable metamorphic assemblage. In peridotites containing considerable amounts of CaO, metamorphic diopside and tremolite are present. Diopside is also a common stable phase in serpentinites at greenschist-facies conditions but is replaced by tremolite and forsterite at amphibolite-facies conditions (*see* **Metamorphic Rocks:** Facies and Zones). It comes in again with enstatite at the amphibolite–granulite-facies transition replacing tremolite and forsterite. Additional Al_2O_3 stabilizes

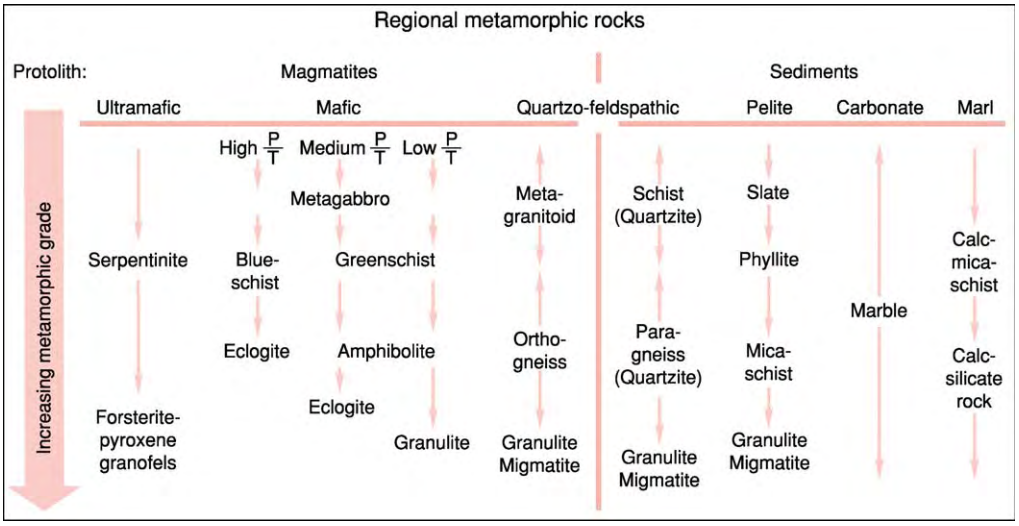


Figure 6 Chart showing the metamorphic rock name in relation with protolith and metamorphic grade.

by a blast that consists of sigmoidal, spiral shaped inclusion patterns, that developed during rotational growth by shearing along the schistosity planes; (E) Post deformational growth is seen in discordant growth of kyanite and biotite which may contain inclusion trails parallel to the external schistosity (S_i S_e); (F) Orthopyroxene forms a corona structure around garnet and ilmenite; and (G) Symplectite formation of diopside and plagioclase after omphacite.

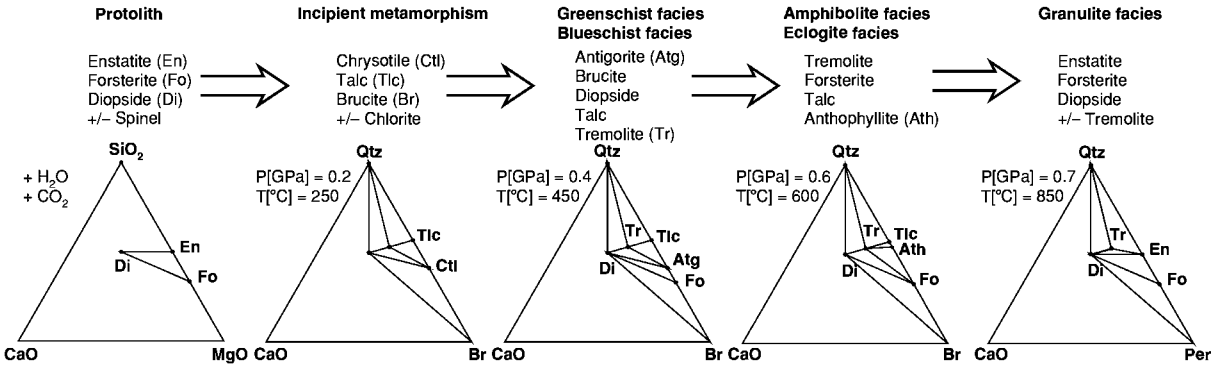


Figure 7 For graphical presentation of mineral assemblages and chemical compositions of ultramafic rocks, the ternary system SiO_2 M_gO CaO with H_2O and CO_2 in excess is used. The significant mineral assemblages for ultramafic rocks at different metamorphic grades are shown here.

chlorite in ultramafic rocks up to the amphibolite–granulite-facies transition where it is replaced by spinel (Figure 7).

Mafic Rocks

Basaltic or gabbroic protoliths are usually anhydrous and must first be hydrated and carbonated at low temperatures (by ocean floor metamorphism) to allow prograde metamorphic reactions to proceed. A typical low-T assemblage is albite + chlorite + carbonates. Compared to other protoliths, mafic lithologies are particularly sensitive to very low-grade metamorphic conditions. Incipient metamorphism is recognized by the occurrence of various zeolite minerals (i.e., laumontite). The subsequent prograde assemblages depend strongly on the geothermal gradient. Index minerals along a path with a high P/T-ratio (Figure 2) are pumpellyite and lawsonite (stable up to ca. 0.8 Gpa at 300°C), glaucophane (stable up to ca. 1.6 GPa at 500°C), and omphacite. Subgreenschist-facies metamorphism along a medium P/T-gradient is recognized by prehnite (at pressures below 0.3 GPa) or pumpellyite at higher pressures. The transition to greenschist-facies assemblages occurs at about 280°C, due to the breakdown of prehnite and pumpellyite to epidote and actinolite. Together with the matrix minerals albite+chlorite+carbonates, they form the characteristic assemblage of a greenschist. The amphibolite-facies assemblage plagioclase+hornblende grows at ~550°C from chlorite+epidote. In course of this reaction garnet may be formed. White micas, particularly paragonite, as well as kyanite coexisting with hornblende, are indicative of pressure-dominated amphibolite-facies conditions (epidote–amphibolite-facies). Additionally to the change in the ACF-topology, the transition from greenschist- to amphibolite-assemblages is also marked by a significant change in the chemical composition of feldspars from albite to oligoclase–andesine and amphiboles from actinolite to Tschermakitic hornblende. Eclogite-facies conditions are reached when plagioclase is replaced by omphacite and quartz. The lower pressure limit of the amphibolite/eclogite transition is *ca.* 1.2 GPa at 600°C. In metamafics metamorphosed along low P/T-gradients (Figure 2), the transition to granulite-facies conditions occurs between 700–800°C. Index mineral phases of mafic granulites are clinopyroxene and orthopyroxene, which form at the expense of hornblende (Figure 8).

Quartzo-Feldspathic Rocks

For the chemographic presentation of mineral assemblages, the AKF-system is applied (Figure 9). The chemical composition of quartzo-feldspathic rocks plot in the Al-poor part of the triangle and therefore these compositions contain few and not easily diagnosed assemblages. Contrary to clastic sediments, magmatites need to be hydrated to make metamorphic reactions possible. However, granitoids often remain in a dry state preventing the rock from completely metamorphosing. Otherwise, incipient metamorphism is demonstrated by phengitic muscovite, chlorite±stilpnomelane. Stilpnomelane is only stable in lower greenschist- and blueschist-facies conditions whereas muscovite remains stable up to about 650–700°C, then breaking down to K-feldspar, quartz, and an Al_2SiO_5-modification. However, the chemical composition of muscovite changes significantly with T and particularly with *P* and may be used as a geobarometer. In the assemblage K-feldspar–biotite–muscovite–quartz, the phengite content of muscovite expressed as Si per formula unit decreases with temperature but increases with pressure due to substitution of the Tschermak-molecule $Mg_{+1}Si_{+1}Al_{-1}Al_{-1}$. Metamorphism along a medium P/T-gradient leads to only a few metamorphic minerals, such as almandine-rich garnet at lower amphibolite facies

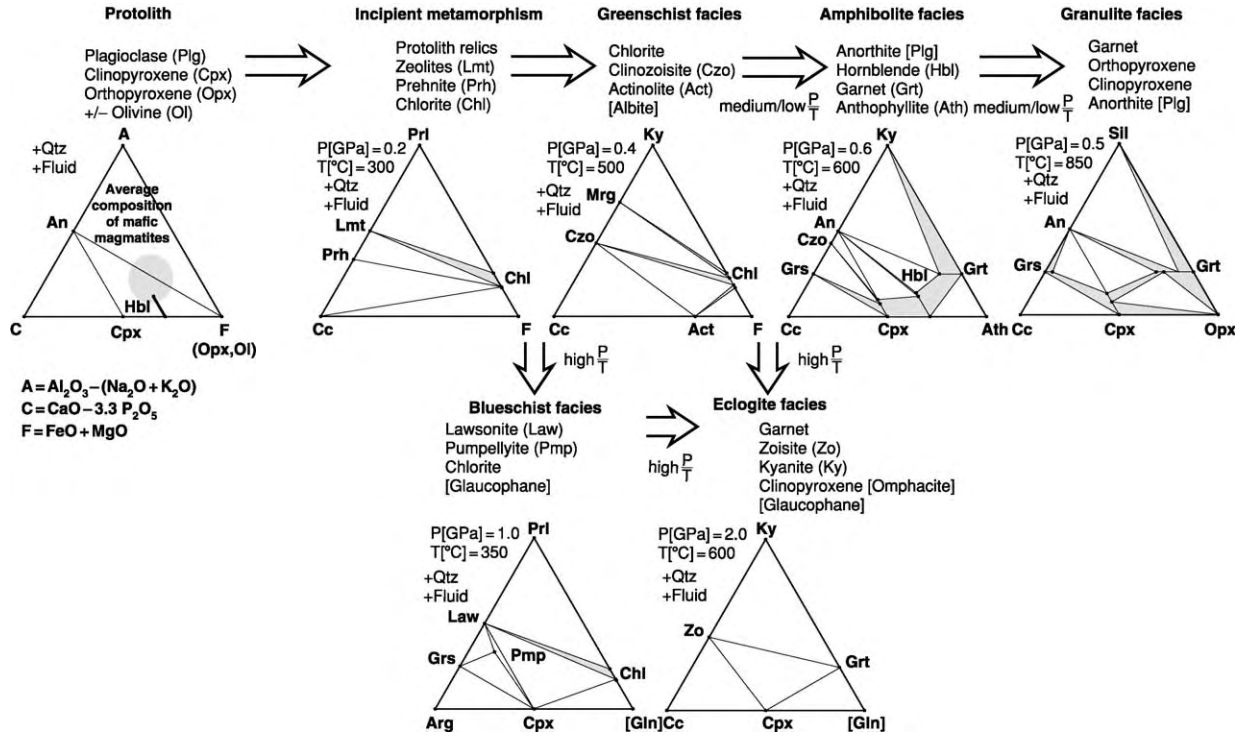


Figure 8 Al_2O_3 (A) is an important component in mafic rocks besides CaO (C) and FeO + MgO (F), hence the ACF projection is commonly used for the graphical representation of mineral assemblages and chemical compositions of metamafics. In order to include silicates, hydrates, and carbonates in the ACF projection SiO_2, H_2O and CO_2 are excess components. The chemographic projections show the significant mineral assemblages for mafic rocks at different metamorphic grades.

conditions – and partial melting takes place at the temperatures of the granite minimum in H_2O-saturated conditions (*ca.* 650°C), resulting in the formation of migmatites. In water-feel conditions, orthopyroxene may form at the expense of biotite at granulite-facies conditions (*ca.* 750°C). In quartzo-feldspathic rocks, which have undergone eclogite-facies metamorphism, plagioclase is usually replaced by jadeite, zoisite, and quartz and biotite by garnet and phengite.

Pelitic Rocks

Pelites are rich in sedimentary water ocurring as H_2O-molecules and as lattice-bound OH-groups. This enables mineral reactions to take place at very low temperatures. The mineral assemblage prior to metamorphism comprises mainly detrital clay minerals in addition to relict quartz, feldspars, and micas. Additionally authigenic Fe-(hydr)oxides, sulphates, and organic carbonaceous material are common. Clay minerals show significant structural changes already in diagenetic conditions, leading mainly to an improvement of lattice ordering. At the transition to metamorphism, white mica and chlorite are formed from clay minerals at about 200°C. These rocks typically have a slaty cleavage due to compaction and parallel orientation of the sheet silicates. Incipient metamorphism is clearly demonstrated by the production of pyrophyllite from kaolinite at about 300°C. In chlorite-rich assemblages, pyrophyllite may in turn be replaced by chloritoid and the terminal breakdown of pyrophyllite to kyanite occurs at *ca.* 400°C. Approximately at this temperature, biotite appears for the first time from K-feldspar and chlorite. In the upper greenschist facies, first garnet appears in an assemblage with biotite at the expense of chlorite and muscovite. At *ca.* 500°C, the terminal reaction of chloritoid leads to the formation of staurolite and garnet. This marks the transition to the amphibolite facies. Staurolite and biotite, formed at the expense of chlorite, garnet, and muscovite, constitute an important coexisting assemblage in the medium grade amphibolite facies at *ca.* 600°C. At the same temperature, paragonite is replaced in quartz-bearing assemblages by albite and an Al_2SiO_5-modification (typically kyanite). At about 700°C, staurolite disappears in quartz-bearing assemblages with the formation of the assemblage sillimanite, kyanite, and garnet, whereas muscovite and quartz are replaced by K-feldspar, sillimanite, or kyanite and melt. Hence metapelites of the upper amphibolite facies are poor in phyllosilicates (the surviving mica

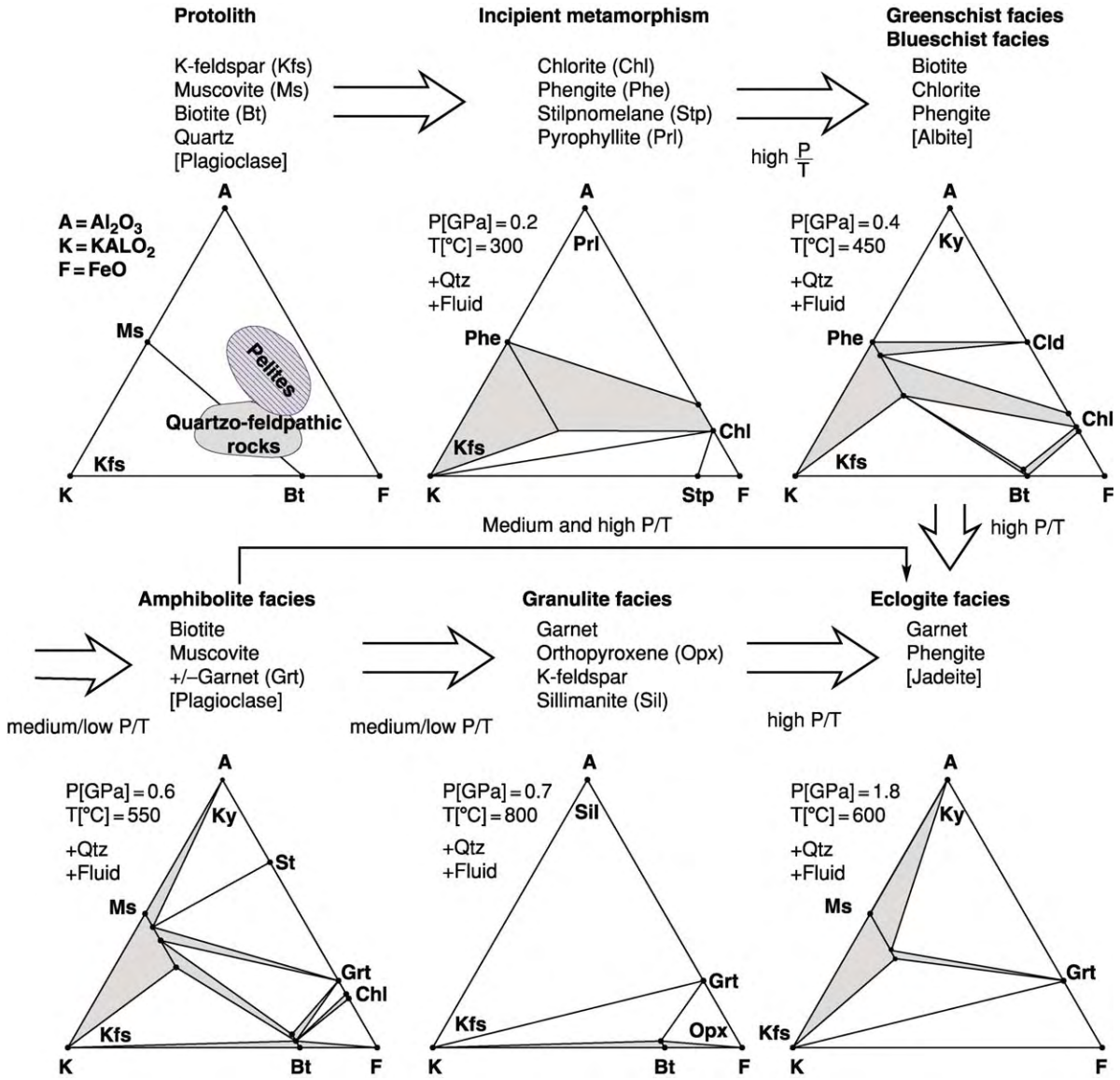


Figure 9 For the chemographic presentation of quartzo feldspathic rocks, the AKF system is applied. The difference to the ACF projection is the consideration of K instead of Ca. Again feldspar, quartz, and H_2O are used as projection phases. K feldspar, muscovite, biotite, and stilpnomelane are the only four important potassium bearing minerals, besides the K free Fe Mg Al phases plotting at the AF line. The chemical composition of quartzo feldspathic rocks plot in the lower Al poor part of the triangle and therefore these lithologies contain only few and not diagnostic assemblages. The AKF chemographic projections show the significant mineral assemblages at different metamorphic grades.

being biotite) but rich in feldspars, garnet, and sillimanite or kyanite. At low pressures, garnet and sillimanite may be replaced by cordierite in the amphibolite facies. In the upper amphibolite-facies, also an orthoamphibole (anthophyllite, gedrite) may form together with cordierite. At *ca*. 750–800°C, the formation of orthopyroxene from orthoamphibole, or from remaining biotite in quartz-bearing rocks, marks the transition to the granulite facies. At still higher temperatures, between 850 and 950°C, spinel, sapphirine, and ossumilite are diagnostic in quartz-bearing metapelites. Eclogite-facies metamorphism of metapelites can best be recognized in MgO-rich bulk chemical compositions, where talc and kyanite form the diagnostic 'white schist'-assemblage. High *P/T* gradients are also indicated by carpholite at lower greenschist facies and talc–phengite assemblages at higher temperatures (Figure 10).

Carbonate Rocks

Pure sedimentary carbonate rocks consist of either calcite, calcite and dolomite, or dolomite and have a

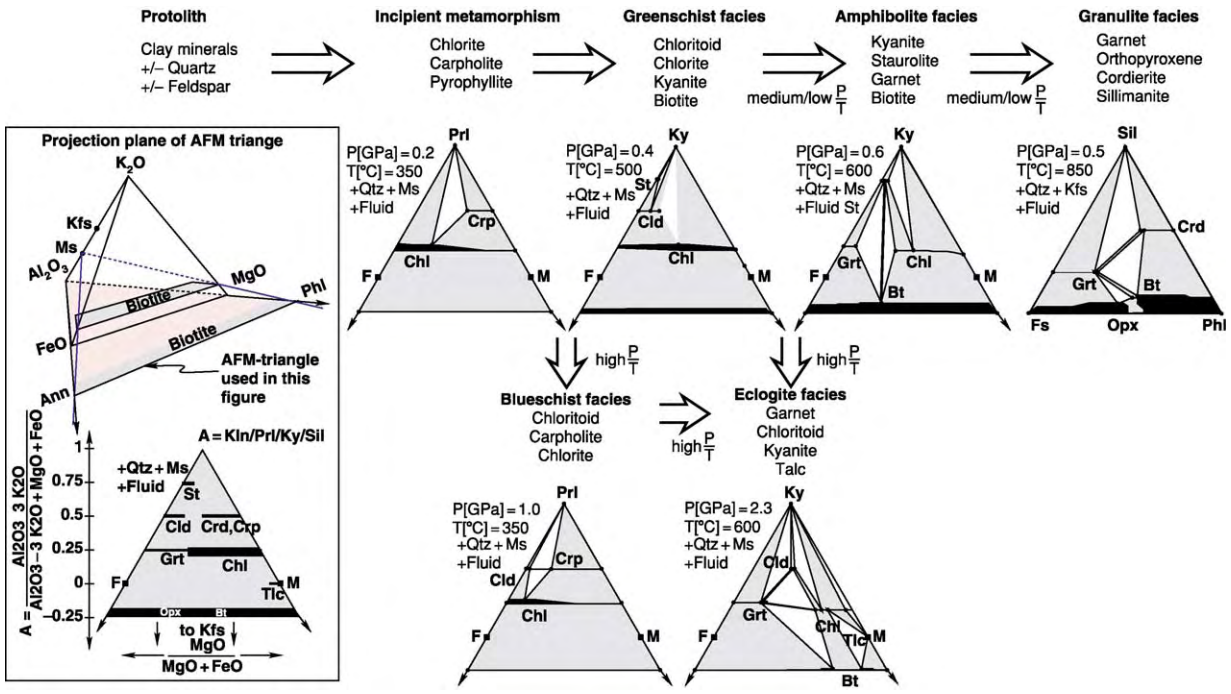


Figure 10 The chemical system applied to metapelites is identical with the AKF system. However, an essential difference has to be taken into account: pelites plot in the Al rich part of the AKF triangle (Figure 9) and are mainly capable of producing Fe Mg Al silicates plotted at the AF line. This rock composition probably contains the most key (or index) minerals with changing metamorphic grade. In order to show the reaction sequence in terms of the Fe Mg Al silicates, it is necessary to consider FeO and MgO as separate components. Thus these four components build up an AKFM tetrahedron and for practical use the AFM plane (projected from muscovite or K feldspar) is used for graphical presentation of the mineral phases.

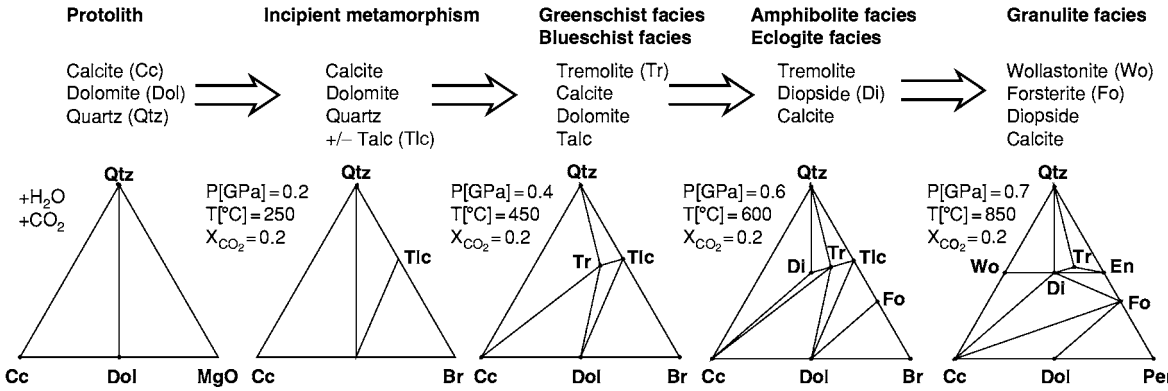


Figure 11 Possible mineral assemblages in metacarbonates with increasing metamorphic grade.

H_2O-rich pore fluid. Quartz has to be added to enable mineral reactions of the carbonates. Hence the relevant chemical system comprises CaO–MgO–SiO_2–CO_2–H_2O. (Figure 11). The temperature of incoming or breakdown of diagnostic assemblages in the carbonate system is dependent on the fluid composition (X_{CO2}) and a *T*-X_{CO2} plot is usually applied to show this effect (Figure 4). The prograde sequence of metamorphic index minerals in calcite–dolomite–quartz marbles affected by an intermediate *P*/*T*-gradient is generally: talc, tremolite, diopside, forsterite, and wollastonite. The maximum transition-temperature from talc+calcite to tremolite is about 500°C. The maximum stability of tremolite+calcite is *ca*. 700°C. The first appearance of diopside is at *ca*. 600°C when it is replacing the mineral assemblage tremolite + calcite + quartz. At *ca*. 800°C, diopside + dolomite breaks down to forsterite. At low P/T-gradients,

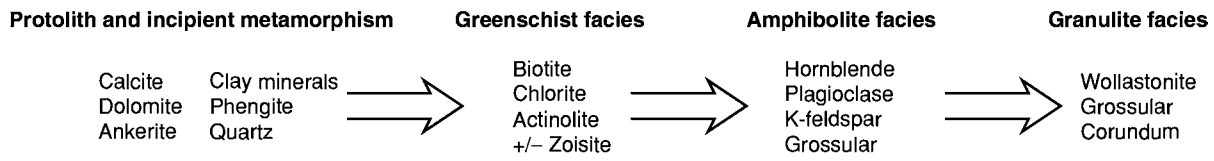


Figure 12 Possible mineral assemblages in metamarls with increasing metamorphic grade.

typical of contact metamorphism, these temperatures are reduced by about 100°C. The regular sequence observed all over the world of index minerals in metacarbonates is indicative of an internal buffering of the fluid composition during prograde metamorphism. An exception is probably the formation of wollastonite from calcite + quartz, or periclase from dolomite at high grade and low X_{CO_2}.

Marls

These rocks represent lithological and chemical mixtures between carbonate rocks and pelites, resulting in a rather complicated chemical system resembling mafic lithologies. Hence, in meta-marls, most minerals occurring in metapelites, metacarbonates, and metamafics may be present. For simplicity, meta-marls are considered here as contaminated metacarbonates where additional K + Na, Al, and Fe lead to several additional mineral phases such as plagioclase, zoisite/epidote, garnet, margarite, paragonite, muscovite, biotite, chlorite, and K-feldspar. Since carbonates and OH-bearing silicates occur, the diagnostic mineral assemblages depend on X_{CO_2}. Starting with an assemblage of ankeritic carbonates, white mica, and quartz, the mineralogy changes to biotite and chlorite at the greenschist–amphibolite facies transition (500°C). At lower amphibolite facies conditions, hornblende, plagioclase, and K-feldspar are diagnostic minerals and occur together with calcite and quartz. Hornblende, calcite, and quartz react to diopside at higher amphibolite facies conditions. Zoisite replaces plagioclase and calcite at very low X_{CO_2} at upper greenschist-to lower amphibolite-facies conditions. At slightly higher temperatures but the same low X_{CO_2}-values, plagioclase, calcite, and quartz are replaced by grossular-rich garnet (Figure 12).

See Also

Impact Structures. **Metamorphic Rocks:** Facies and Zones; PTt - Paths. **Minerals:** Definition and Classification. **Regional Metamorphism**. **Shock Metamorphism**. **Thermal Metamorphism**. **Ultra High Pressure Metamorphism**.

Further Reading

Árkai P, Sassi FP, and Desmons J (2002) *Towards a unified nomenclature in metamorphic petrology: Very low grade to low grade metamorphic rocks*. A proposal on behalf of the IUGS Sucommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Blatt H and Tracy RJ (1996) *Petrology: Igneous, Sedimentary, and Metamorphic*, 2nd edn. WH Freeman and Company.

Brodie KH, Fettes D, Harte B, and Schmid R (2002) *Towards a unified nomenclature in metamorphic petrology: Structural terms, including fault rocks*. A proposal on behalf of the IUGS Sucommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Bucher K and Frey M (2002) *Petrogenesis of Metamorphic Rocks*. Berlin, Heidelberg, New York: Springer Verlag.

Desmons J, Smulikowski W, and Schmid R (2002) *Towards a unified nomenclature in metamorphic petrology: High P/T rock terms*. A proposal on behalf of the IUGS Subcommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Dudek A, Coutinh JMV, Desmons J, *et al.* (2002) *Towards a unified nomenclature in metamorphic petrology: Amphibolite and granulite*. A proposal on behalf of the IUGS Sucommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html .

Le Maitre RW (ed.) (2002) *Igneous Rocks. A Classification and Glossary of Terms*. Cambridge: Cambridge University Press.

Philpotts AR (1990) *Principles of Igneous and Metamorphic Petrology*, 1st edn. London: Prentice Hall.

Rosen OM, Desmons J, and Fettes D (2003) *Towards a unified nomenclature of metamorphism: Carbonate and related rocks*. A proposal on behalf of the IUGS Subcommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Schmid R, Fettes D, Harte B, Davis E, Desmons J, and Siivola J (2001) *Towards a unified nomenclature in metamorphic petrology: How to name a metamorphic rock*. A proposal on behalf of the IUGS Sucommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Smulikowski W, Desmons J, Harte B, Sassi FP, and Schmid R (2003) *Towards a unified nomenclature in metamorphic petrology: Types, grade and facies*. A proposal on behalf of the IUGS Subcommission on the Systematics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/scmr products.html.

Stöffler D and Grieve RAF (2003) *Towards a unified no menclature of metamorphism: Impactites*. A proposal on behalf of the IUGS Subcommission on the Systemat ics of Metamorphic Rocks. http://www.bgs.ac.uk/SCMR/ scmr products.html.

Wimmenauer W and Bryhni I (2002) *Towards a unified nomenclature of metamorphism: Migmatites and related rocks*. A proposal on behalf of the IUGS Subcommis sion on the Systematics of Metamorphic Rocks. http:// www.bgs.ac.uk/SCMR/scmr products.html.

Winter JD (2001) *An Introduction to Igneous and Meta morphic Petrology,* 1st edn. London: Prentice Hall.

Yardley BWD (1996) *An Introduction to Metamorphic Petrology*. Addison Wesley Publication Company.

Facies and Zones

K Bucher, University of Freiburg, Freiburg, Germany

Intensity of Metamorphism

The intensity of metamorphism is mainly related to the pressure and temperature conditions prevailing during rock transformation. A rock recrystallized at 800°C in the lower crust contains characteristic minerals or mineral assemblages that reflect the high PT conditions under which the rock formed. In contrast, a shale in the shallow upper crust equilibrates perhaps at 300°C and is said to be weakly metamorphosed. It is of prime interest in geology to characterize the intensity of metamorphism and to use a reliable classification system to compare it within and between metamorphic terrains. The metamorphic facies system is such a tool for characterizing the intensity or grade of metamorphism (*see* **Metamorphic Rocks:** Classification, Nomenclature and Formation).

The distinctive grade of metamorphism of rocks from an outcrop can be evaluated and described by a large variety of techniques and methods. Because pressure and temperature are the main variables that control metamorphism, one could simple determine P and T for each collected sample by using methods of geological thermobarometry. This is, however, a tedious and costly affair and not necessary, except for special research studies. The metamorphic facies system is a much simpler system that can be used in the field and normally does not require laboratory work or even thin section observation.

The Metamorphic Facies Concept

The concept of metamorphic facies, as it is used today, is simple in principle. It makes use of the fact, that at a given pressure and temperature the chemical components of a rock are distributed among a unique set of minerals with fixed compositions. This is an inevitable consequence of chemical thermodynamics, provided the rocks reach a state of equilibrium. The unique sets of minerals that form in a reference material at one P and T can be compared with assembages of minerals that formed in other rock compositions under the same P and T. The collective sets of mineral assemblages that form under the same conditions in many kinds of rocks are called a metamorphic facies.

The reference material that is used in the metamorphic facies scheme is tholeiitic basalt (mid-ocean ridge basalt or MORB) because of its well-defined composition and global abundance. Mafic rocks of basaltic composition are extremely widespread and metamorphosed basalts (metabasalt) constitute a major rock type in orogenic belts, the oceanic lithosphere, and continental crust.

For example, at an outcrop one finds a mafic rock consisting of hornblende (Hbl) and plagioclase (Pl, e.g., andesine), that is termed an amphibolite. A metapelitic rock exposed at the same outcrop consists of kyanite, staurolite, biotite, muscovite, and quartz, and an associated band of dolomite marble contains dolomite, calcite, tremolite, and diopside. All three kinds of rocks with very diverse compositions have been metamorphosed under the same pressure and temperature conditions. They are of the same metamorphic grade.

At this outcrop, the reference material consists of Hbl + Pl and hence the set of rocks with their characteristic mineral assemblages each belong to the same metamorphic facies. The facies name could be the hornblende-plagioclase facies. For simplicity and because Hbl + Pl rocks are amphibolites, the name amphibolite facies is used for conditions under which Hbl and Pl are formed in the reference material, namely MOR basalt.

The kyanite-staurolite micaschist at the same outcrop has also been metamorphosed under amphibolite facies conditions, i.e., where basalt is transformed into amphibolite, and the micaschist is also an amphibolite facies rock. However, it contains no hornblende and no andesine. Similarly, the

tremolite-diopside marble belongs to the same amphibolite facies.

At another outcrop some 20 km from the first, one finds micaschist but no amphibolite and no marble. The micaschist contains kyanite, staurolite, biotite, muscovite, and quartz, i.e., the same assemblage as at the first outcrop. Consequently, the micaschist at this second outcrop is also an amphibolite facies rock, metamorphosed under conditions where hornblende and andesine would have formed in metabasalt if this rock were present.

Metamorphic Facies Systems

The example above illustrates the practical use and the concept of metamorphic facies. The amphibolite facies is the total of all pressure and temperature (P-T) conditions under which metabasalt would consist predominantly of hornblende and plagioclase. These PT conditions can be broadly outlined on a pressure versus temperature diagram (Figure 1).

In a next step, a series of characteristic metabasalt mineral assemblages can be used to define additional facies to the amphibolite facies described above. In fact, one can select typical assemblages in metabasalts that form at any P-T conditions rocks can be subjected to on Earth.

A decision has to be made, however, how detailed the assemblages should be for use in the metamorphic facies system. First we use, as in the example above, just two characteristic minerals, a mineral pair, occurring in metabasalt to define a series of additional facies. Other versions and additional complexities will be introduced later.

Amphibolite Facies

Mafic rocks contain hornblende and plagioclase. Hornblende is an aluminous amphibole with a composition that contains predominantly a pargasite component. Hornblende is dark in hand specimen and green to brown under the microscope. Plagioclase contains significant amounts of an anorthite component. Oligoclase is the minimum requirement, typical is andesine or labradorite.

Greenschist Facies

In the greenschist facies mafic rocks contain actinolite and albite. Actinolite is an Al-poor amphibole; it contains mainly tremolite. Actinolite is green in hand specimen and colourless to pale green under the microscope. Plagioclase is pure albite and contains no Ca. The two minerals are the major constituents of greenschists. The greenschist facies represents lower P-T conditions than the amphibolite facies, i.e., a lower metamorphic grade. Greenschist and amphibolite facies rocks form in typical regional metamorphism related to a geodynamic setting of continental collision tectonics (Figure 2). In many orogenic belts worldwide, metamorphic rocks of the greenschist and amphibolite facies predominate over all other metamorphic facies rocks.

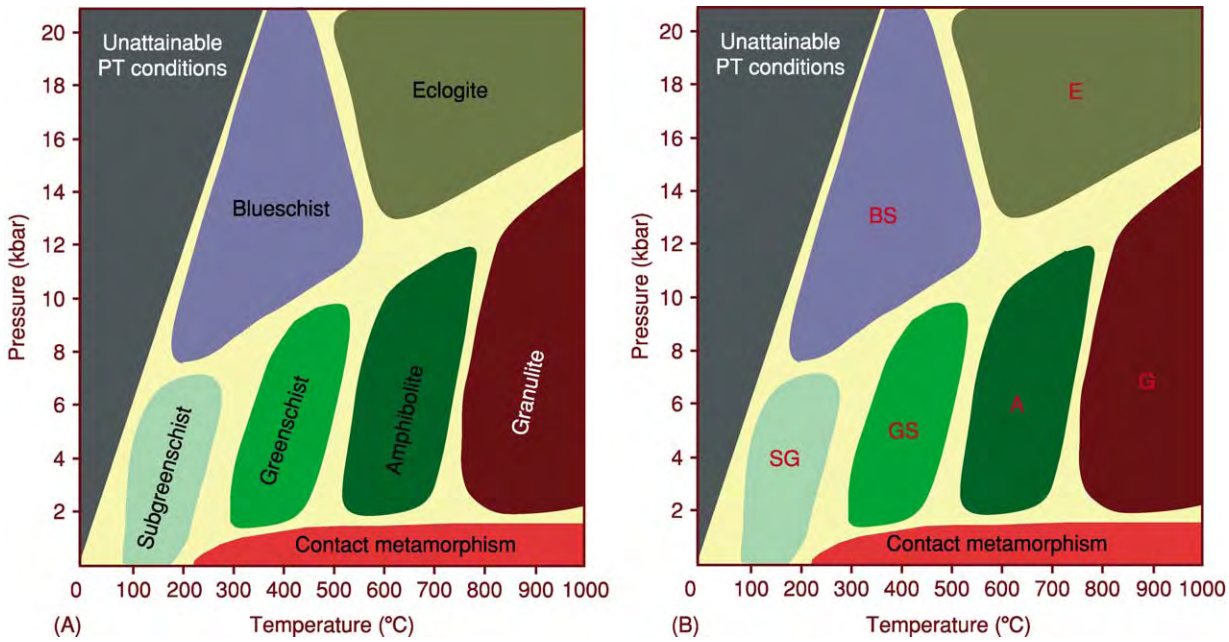


Figure 1 (A) Metamorphic facies scheme, general version (full names). (B) Metamorphic facies scheme, general version (abbreviation). SG sub greenschist, GS greenschist, A amphibolite, G granulite, BS blueschist, E eclogite.

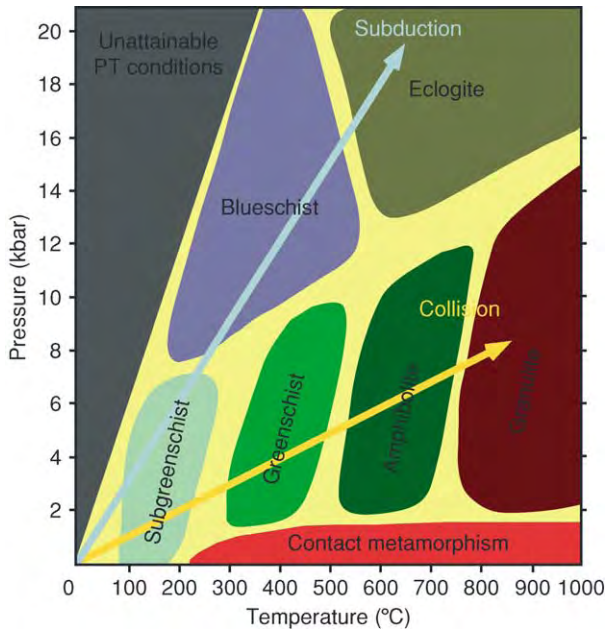


Figure 2 Metamorphic facies scheme, general version (full names). ''subduction'' path of prograde metamorphism during subduction of oceanic lithosphere; ''collision'' path of pro grade metamorphism during continental collision.

Blueschist Facies

Mafic rocks contain glaucophane and epidote in the bluschist facies. Glaucophane is an alkali-amphibole that is dark or blue in hand specimen and colourless to blue under the microscope. Plagioclase is not normally present, but can occur as albite. Epidote or clinozoisite occurs in abundance and carries much of the calcium of the basaltic precursor rock. The two minerals are major constituents of blueschists, hence blueschist facies represent high-P low-T conditions typical of subduction zone metamorphism (Figure 2). These conditions are normally not compared to the greenschist or the amphibolite facies, in the sense of lower or higher grade rocks, because blueschists reflect metamorphism in a different geodynamic context.

Eclogite Facies

Some mafic rocks contain garnet and omphacite and are termed eclogites, hence eclogite facies. Eclogite garnet is a ternary mixture of almandine, pyrope, and grossular. Omphacite is an calcalkali pyroxene made of jadeite, diopside, and hedenbergite components. Omphacite is green in hand specimen and colourless to green under the microscope. Plagioclase is absent. Eclogite facies rocks form at high pressure in diverse geodynamic settings, ranging from subduction of oceanic lithosphere, to the formation of thickened continental crust during continent collision. Eclogites also are present at depth in stable anorogenic settings. If eclogites are formed by continuous prograde metamorphism, they are always the highest grade rocks. Although eclogite facies rocks are widespread in some orogenic belts, they typically occur in small isolated outcrops in a matrix of surrounding rocks that reflect conditions of lower grade facies.

Granulite Facies

Metamorphic mafic rocks that contain orthopyroxene (Opx) and plagioclase belong to the granulite facies. Orthopyroxene is a mixture of enstatite and ferrosilite components and is also called hypersthene. It is dark with a brownish tint in hand specimen and colourless to reddish under the microscope. Plagioclase is calcic and a labradorite composition is typical. The two minerals are present in mafic granulites. However, the Opx + Pl pair is present in granulite facies mafic rocks only at relatively low pressure. The major mineral pair in all mafic granulites is clinopyroxene (Cpx) and plagioclase. Note that this assemblage is identical to that of the unmetamorphosed basalt, however, the composition of the minerals and the texture of the rocks are different. Hence, the Cpx + Pl pair that makes up the bulk of granulite facies mafic rocks is unfortunately not exclusive to that facies.

As mentioned, the high-pressure portion of the granulite facies P-T field (Figure 1) lacks the diagnostic assemblage Opx + Pl. How does one know then, that a suite of rocks reflects high-pressure granulite facies conditions? The defining assemblage in this case is Cpx + Pl + garnet. The Cpx in mafic granulites is an augite with a low jadeite content. This is a black mineral in hand specimen and greenish under the microscope. Garnet is a ternary mixture of almandine, pyrope, and grossular components much like those in eclogite.

The granulite facies represents PT conditions higher than that of the amphibolite facies, so compared with the amphibolite facies, it represents higher grade rocks. Granulite facies rocks occur in extensive terrains, particularly in Precambrian shield areas. They are rare in Phanerozoic orogenic belts and very rare in Alpine chains. The boundary of granulite terrains to surrounding amphibolite facies rocks is usually of a retrograde nature.

Zeolite Facies

If metamorphism is weaker than the conditions of the greenschist facies, mafic rocks may contain an abundance of minerals that could be used to define metamorphic facies. Typical greenschist facies rocks require about 300°C to form from basalt. Greenschist facies rocks are typically well equilibrated and relics

from the original igneous assemblage are rare. At conditions of very low-grade metamorphism this is typically not the case. At, for example, 150°C, a number of characteristic zeolite minerals may form in vesicles and fractures of basaltic lava but the transformation is incomplete and igneous minerals are still present, although often severely altered. The alteration of igneous plagioclase and clinopyroxene is a process of hydration and water must have access to the reaction sites. Alteration by hot aqueous solutions is known as hydrothermal alteration (*see* **Thermal Metamorphism**). It may take place at the sea-floor and then be termed oceanic metamorphism or it may represent the initial stages of metamorphism in continental settings. It is rarely pervasive. Hydrothermal alteration is always accompanied by chemical transport, deposition of new minerals in the pore space of rocks and thus is not isochemical. The zeolite assemblages formed in basaltic rocks do not completely replace the original rock and are characteristic of PT conditions at the Earth's surface to about 150–200°C (Figure 3).

Prehnite–Pumpellyite Facies

At higher temperatures, but still not hot enough for greenschists to form, two characteristic minerals may be found in mafic rocks: prehnite and pumpellyite. The two minerals rarely occur together in the same rock. However, mafic rocks with prehnite or pumpellyite are typical of metamorphic conditions between the zeolite facies and the greenschist facies (Figure 3). There is a considerable overlap of mineral assemblages of these two very low-grade metamorphic facies. Pumpellyite often occurs in metabasalts together with a number of different zeolites. Prehnite and zeolites are also not exclusive. In greenschist facies rocks, however, zeolites. prehnite, and pumpellyite are absent (line 1 in Figure 5).

Sub-Greenschist Facies

Some typical features of the zeolite and prehnite-pumpellyite facies include the persistence of igneous minerals, overlap of assemblages of the two facies, widespread occurrence of metastable assemblages, and non-equilibrium textures and assemblages. Because of these features it is better to characterise the rocks at grades below that of the greenschist facies as sub-greenschist facies in many metamorphic terrains. This would comprise all rocks and all processes that transform rocks below about 300°C. Sediment diagenesis is normally excluded from the facies scheme but all hydrothermally altered igneous and metamorphic rocks can be described by either terms, sub-greenschist facies or very-low-grade metamorphism (Figure 1).

Ultra-High-Pressure Metamorphism

At very high pressures, mafic rocks contain omphacite and garnet. This is the mineral pair that makes up eclogite and defines the eclogite facies (see above). Rocks from as deep as 250 km below the surface have been found at various localities. The corresponding pressure is on the order of 50–60 kbar. Such ultra-high-pressure (UHP) conditions are normally especially pointed out in rock descriptions and separated from normal eclogite facies rocks (Figure 4).

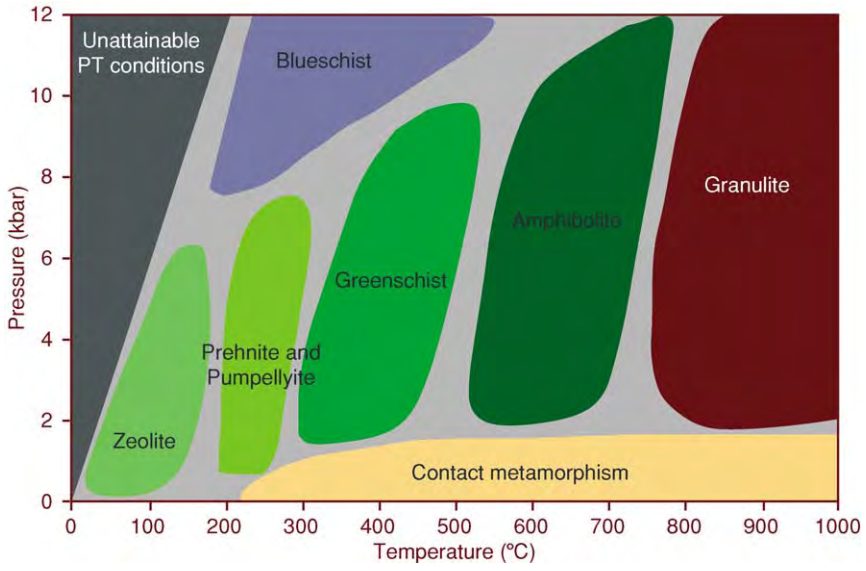


Figure 3 Metamorphic facies scheme, showing the PT fields of the zeolite and prehnite pumpellyite facies.

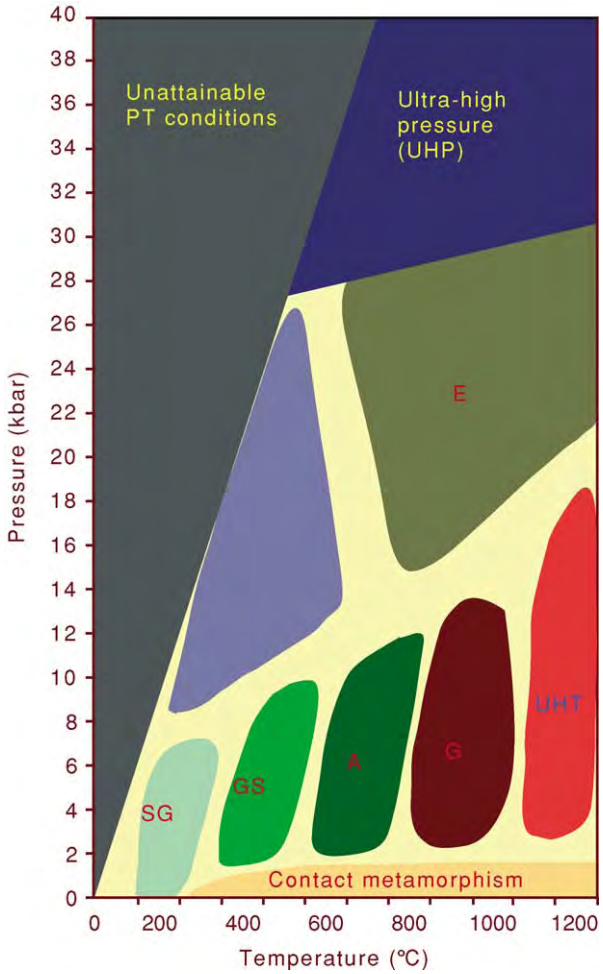


Figure 4 Metamorphic facies scheme, general version (abbreviations as in Fig. 1B). Extended P T range showing conditions of UHP (ultra high pressure) and UHT (ultra high temperature) metamorphism.

The most diagnostic mineral of the UHP facies is coesite, indicating that UHP conditions are above the quartz-coesite transition. Coesite-eclogites are typical rocks of the UHP facies. One could also refer to a coesite facies. However, many UHP eclogites do not contain coesite. Among a number of other assemblages diagnostic of UHP facies conditions is talc + chloritoid.

Ultra-High-Temperature Metamorphism

The granulite facies includes all rocks that have been metamorphosed at high-grade conditions in the stability field of plagioclase (*see* **Ultra High Pressure Metamorphism**). Normally, granulites are associated with metamorphic temperatures in the range of 700–900°C. Increasing evidence shows that some rocks may have experienced metamorphic temperatures up to 1100°C. These ultra-high-temperature rocks may be classified as a separate facies, the ultra-high-temperature facies (UHT), analogous to the UHP facies (Figure 4).

Contact Metamorphism

Heating of rocks at low pressure in the shallow crust by intrusion of igneous rocks causes metamorphism that is restricted in areal extent, close to the igneous contact which is normally a pipe or subsurface magma chamber. The process is termed contact metamorphism and typical rocks are termed hornfels which are very fine grained metamorphic rocks resulting from rapid heating. It should be noted however, that contact metamorphic marbles can be coarse grained. The maximum temperature reached by the country rock aureole surrounding the igneous intrusion can be high and depends on the amount, nature, and composition of the magma. Basaltic magmas can reach temperatures in excess of 1200°C and contact rocks may be heated to 900°C or higher by gabbro intrusions.

In principle, contact metamorphic rocks can also be assigned to a metamorphic facies such as albite-epidote hornfels, hornblende hornfels, pyroxene hornfels, and sanidinite facies. Rocks of each of these facies would typically occur as successive zones of several tens of metres in thickness surrounding the igneous body. In the study of such contact aureoles, most researchers have used isograd mapping as the preferred tool to characterise of the metamorphic (= thermal) structure.

Volatile Components and the Role of Water

The metamorphic facies scheme relies on the premise that the stable mineral assemblage of a rock is entirely determined by pressure, temperature, and the composition of the rock. This is only true in a strict sense if the rock in question has had no chemical exchange with the surroundings, i.e., the metamorphism is isochemical. This is normally not important for the major components of a rock. However, it is a matter that has to be evaluated for volatile components in each case. The prime volatile component of rocks is H_2O. In prograde metamorphism, dehydration reactions typically create an H_2O pressure in rocks that equals the total pressure, that is H_2O is present as a free fluid phase. The presence of fluid is evidenced by the presence of ubiquitous fluid inclusions in minerals of metamorphic rocks. The rock controls its fluid phase at a given P and T.

In some geologic environments this is not the case. If rocks undergo prograde dehydration in regional metamorphism (*see* **Regional Metamorphism**), they

gradually loose their fluid phase. At a certain stage, the minerals are not in contact with a free aqueous fluid and under such circumstances hydrous minerals such as micas and amphiboles decompose at significantly lower temperatures compared with the case where free water would have been present. Consequently, high-grade mineral assemblages appear in rocks at lower grade. Lack of a free aqueous fluid is particularly typical of granulite facies rocks. Thus, the amphibolite–granulite facies transition (line 3, in Figure 5) is not only a PT boundary but is also related to decreasing H_2O pressure in the rocks. Also, in very low-grade rocks, H_2O may not pervasively wet all mineral grain boundaries in primary igneous rocks as outlined above (*see* **Metamorphic Rocks:** PTt - Paths). As a consequence, very high-grade assemblages may be found at ambient conditions, a fortunate circumstance that permits survival of the high-grade rocks at conditions of the earth surface.

Another aspect of volatile components and their effect on the facies scheme is the frequent occurrence of CO_2 and other gaseous components in metamorphic fluids. In sedimentary sequences that undergo metamorphism, carbonate-bearing strata are very common. When heated, these rocks produce CO_2 in addition to H_2O. No problems arise if the fluid produced remains in the rock that generates it. However, if the CO_2 producer exports its fluid to carbonate-free rocks, hydrous minerals may decompose at lower grade than if free pure H_2O would have been present.

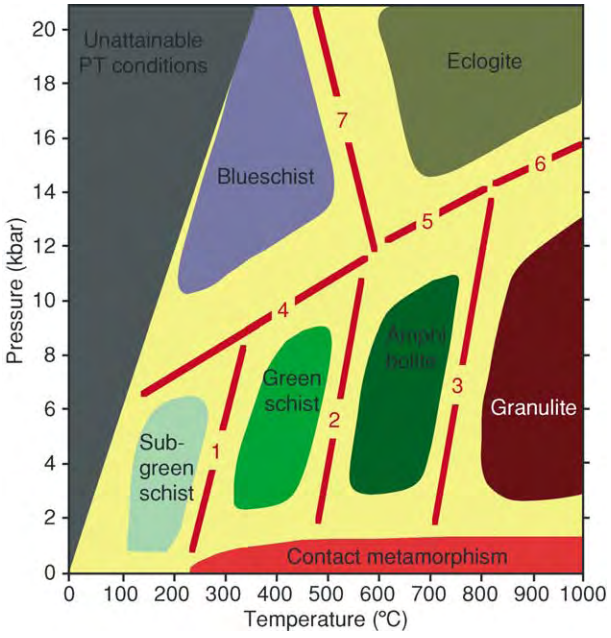


Figure 5 Metamorphic facies scheme, general version (full names) showing numbered facies boundaries as discussed in the text.

The effect on facies boundaries is similar to those cases lacking a fluid phase and it typically affects the amphibolite–granulite facies boundary (line 3, in Figure 5). The consequence of this fluid effect is that at the same temperature and pressure, granulite and amphibolie facies rocks may occur side by side depending on the H_2O pressure (or the chemical potential of H_2O in the rock).

Partial Melting and Igneous Processes

The metamorphic facies scheme (Figures 1 and 2) requires the presence of a free aqueous fluid. The facies figures have been prepared with this premise. Above, some exceptions from this requirement have been given with respect to the granulite and subgreenschist facies. An additional complication arises from the fact that in the presence of free H_2O-fluid, mafic rocks begin to melt at higher temperatures. The minimum melt temperature for metabasalts can be as low as 700°C in the middle crust. The onset of partial melting and the presence of a silicate melt at higher metamorphic grades has several consequences for the facies concept and the characterization of metamorphic grade.

H_2O and other volatiles dissolve in the silicate melt or may be expelled from the melt. The fine details of the melt-producing process has, therefore, consequences for the requirement of a free H_2O fluid in rocks. At conditions of the granulite facies, the amount of melt present depends on the H_2O pressure. The absence of migmatites (partially melted rocks) in many granulite facies terrains is strong evidence of low water pressure during metamorphism.

At conditions of the UHT facies, melt is produced even in the absence of H_2O. The absolute upper end of crustal metamorphism is given by 'dry' melting of granites (about 1150°C).

Melt produced in metamorphic rocks has the important implication that a once homogeneous precursor rock separates into a melt phase and a residuum (or restite) of different composition. Thus, isochemical conditions, the requisite basis for comparing metamorphic rocks from different grades, is no longer maintained in rocks that undergo partial melting. One has to be aware of these complications when working in migmatite areas.

Isofacial and Allofacial Conditions

It is commonly observed in metamorphic terrains that certain outcrops are of rather different metamorphic grade compared with all other outcrops in the area. Rocks belonging to very different facies may occur side by side along a road cut and the variation in

grade cannot be related to a continuous process of metamorphism that has transformed all rocks of the area. A good example of this is found along the west coast of Norway. There are a large number of 100 m-sized lenses and blocks of eclogite or of UHP rocks (coesite eclogites) in an amphibolite facies gneiss terrain. No relic assemblages of the eclogite facies are present in the gneiss. The eclogite blocks have tectonic contacts with the gneiss and are characterised as allofacial relative to them. Other examples are outcrops of garnet peridotite in the Central Swiss Alps that occur in amphibolite facies rocks. These UHP rocks are also clearly allofacial.

An the other hand, serpentinites of the Zermatt-Saas ophiolite complex of the Central Swiss Alps occur together with metabasalts and metagabbros of the eclogite and blueschist facies. The serpentinites represent hydrated mantle of the oceanic lithosphere and their assemblages antigorite + forsterite + diopside is isofacial with the eclogite facies metabasites.

The Metamorphic Facies Scheme, Additional Aspects and Facies Boundaries

Features of the amphibolite–granulite facies boundary (line 3, in Figure 5) and the lower limit of the greenschist facies (line 1, in Figure 5) have been briefly discussed.

The greenschist–amphibolite facies boundary (line 2, in Figure 5) is characterized in mafic rocks by the relatively abrupt increase of the calcium (anorthite)-content of plagioclase and by a simultaneous increase of the tschermak- and edenite-content of amphibole. In metapelites, staurolite-bearing assemblages gradually replace chloritoid-bearing assemblages.

The boundaries 4, 5, and 6 in Figure 5 all mark the disappearance of plagioclase in mafic rocks towards higher pressures. At low temperature, albite-bearing mafic rocks are replaced by assemblages with sodic amphibole (line 4, in Figure 5). Sodium is transferred from feldspar to amphibole and mica (paragonite). At higher temperature, amphibolite and granulite are directly replaced by eclogite (lines 5 and 6, in Figure 5). The transition from blueschist to eclogite facies (line 7, in Figure 5) is a very gradual boundary. In typical low-temperature eclogite, omphacite and garnet often coexist with glaucophane and clinozoisite over a fairly wide range of conditions. This means that assemblages of both facies occur in the same metabasalt. This is mostly due to the strong dependence of the stable assemblages on the redox conditions prevailing during high-pressure metamorphism.

Mineral Zones

Characterizing the metamorphic grade can also be achieved by mapping in the field the occurrence of so called index minerals in a homogeneous rock unit with constant composition over a large outcrop area. Index minerals have a PT sensitive distribution in the rock type studied. In going up-grade, the first occurrence of an index mineral can be placed as a line on a map. The line connecting all outcrops of the first occurrence of an index mineral is called a zone boundary of a mineral zone that is characterized by that index mineral. The upper grade limit of the mineral zone is defined by the mineral zone boundary of the next index mineral with increasing grade.

The classic example for the use of this simple and straightforward technique is the map of mineral zones in metapelites (micaschists) in the Scottish Highlands (see Further Reading). At the lowest grade outcrops, the micaschists contain chlorite as a diagnostic mineral. The first occurrence of biotite in the micaschists can be mapped as a line in the field. It defines the zone boundary of the biotite zone. The next diagnostic mineral that appears in the rocks is garnet. Its first appearance defines the beginning of the garnet zone. At still higher grade staurolite appears in the schist and the staurolite zone boundary marks also the upper end of the garnet zone. Note that both biotite and garnet may still be present in the rocks. Continuing upgrade, kyanite can be found in staurolite-garnet-biotite micaschists, defining a kyanite mineral zone. Finally at the highest metamorphic grade but still within the amphibolite facies, sillimanite can be found in the rocks. The complete zonal sequence of index minerals is known as the Barrovian sequence of metamorphism. It is a typical pattern of metamorphism and it has been described from many orogenic belts.

See Also

Metamorphic Rocks: Classification, Nomenclature and Formation; PTt - Paths. **Regional Metamorphism**. **Thermal Metamorphism**. **Ultra High Pressure Metamorphism**.

Further Reading

Austrheim H (1990) The granulite eclogite facies transition: A comparison of experimental work and a natural occurrence in the Bergen Arcs, western Norway. *Lithos* 25.

Barrow G (1912) On the geology of lower Deeside and the southern Highland border. *Proceedings of the Geologists Association* 23: 268 284.

Bucher K and Frey M (2002) *Petrogenesis of Metamorphic Rocks*. Berlin, Heidelberg: Spring Verlag.

Coombs DS, Ellis AJ, Fyfe WS, and Taylor AM (1959) The zeolite facies, with comments on the interpretation of hydrothermal syntheses. *Geochimica et Cosmochimica Acta* 17: 53 107.

Eskola P (1921) On the eclogites of Norway. *Skrifter Viden sabelig Selskab. Christiania, Mat. nat. Kl. I* 8: 1 118.

Evans BW and Brown EH (1986) *Blueschists and Eclogites*. Geological Society of America Memoir, 423 pp. Boulder, Colorado: The Geological Society of America.

Frost BR and Frost CD (1987) CO_2, melts, and granulite metamorphism. *Nature* 327: 503 506.

Newton RC (1985) Temperature, pressure and meta morphic fluid regimes in the amphibolite facies to gran ulite facies transition zones. In: Tobi AC and Touret JCR (eds.) *The Deep Proterozoic Crust in the North Atlantic Provinces*, pp. 75 104. Dordrecht: Reidel.

Pattison DRM (1991) Infiltration driven dehydration and anatexis in granulite facies metagabbro, Grenville Province, Ontario, Canada. *Journal of Metamorphic Geology* 9: 315 332.

Pattison DRM (2003) Petrogenetic significance of orthopyr oxene free garnet + clinopyroxene + plagioclase ± quartz bearing metabasites with respect to the am phibolite and granulite facies. *Journal of Metamorphic Geology* 21: 21 34.

Winter JD (2001) *An introduction to igneous and meta morphic petrology*. New Jersey: Prentice Hall.

PTt-Paths

P J O'Brien, Universität Potsdam, Potsdam, Germany

Introduction

A pressure–temperature–time (usually abbreviated as PTt) path is, very simply put, a record of the ups and downs in temperature and pressure experienced by a metamorphic rock during its lifetime. This information is cryptically recorded in the form of different minerals present in a rock, in their chemical and isotopic compositions and degree of chemical homogeneity, and in the fabrics (size, shape, distribution, and orientation) of the different minerals. In order to utilize PTt paths of rocks for interpreting Earth history it is necessary to understand the different processes causing temperature and pressure changes, how the rates of change of these processes differ due to different geological processes, and how this is then reflected in mineralogical and textural changes in rocks. Once these processes are understood and quantified, they can be used to model and predict the changes expected in metamorphic rocks as a result of a chosen tectonometamorphic scenario. The complementary situation is the reconstruction of the magnitude and duration of metamorphic processes based on PTt paths derived from natural rock samples that were 'eye-witnesses' to one or more tectonometamorphic events, i.e., rocks that underwent heating, cooling, burial, and exhumation as a result of tectonometamorphism. In contrast to theoretical PTt-path modelling, the reconstruction of PTt paths from actual samples requires determination of the equilibration pressure and temperature conditions for mineral assemblages, the age of these assemblages, and also how mineral compositions and assemblages have changed over time as a function of evolving pressure and temperature changes: factors all fraught with uncertainty and error. For this reason the information from natural samples is far from complete but it is information from natural samples that provide key constraints and thus allow fine-tuning of predictive tectonometamorphic models.

PTt Paths: The Basics

If a metamorphic rock preserves a perfect equilibrium mineral assemblage then its history can be reflected as a single point in a PT diagram (**Figure 1A**). The fact that it was most likely collected at the Earth's surface means that the mineral assemblage is metastable and also that the route from the depths corresponding to the PT point back to the surface is not recorded. If the age of the mineral assemblage can be determined then average cooling and exhumation rates relative to today's surface conditions are available. However, if a rock does not show equilibrium, then more than one PT point can be deduced and, assuming that the sequence of reactions or changes can be determined, a PT path can be constructed (**Figure 1B**). Disequilibrium is typically evidenced by frozen, incomplete reactions where both old and new phases are present; as compositional differences between matrix minerals and the same phases present as armoured inclusions in large porphyroblasts; as compositional zoning of minerals; or as multiple growth events for the same mineral. If the age of the individual metamorphic stages in the history can be determined then all three requirements for a PTt path are met, i.e., pressure, temperature, and time changes (**Figure 1C**). However, the very fact that different stages in the PT evolution

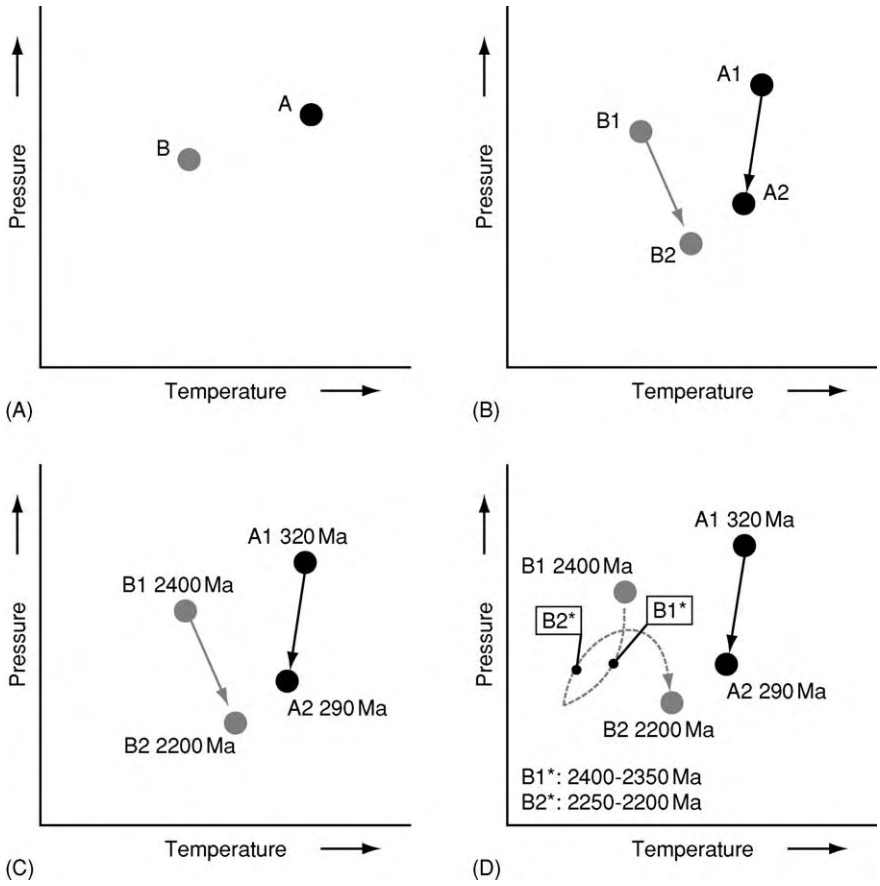


Figure 1 Schematic diagrams showing the relationship between PT point (A), PT path (B), and PTt path (C, D).

are determinable means that the rock is no longer in equilibrium, and thus the use of geothermobarometers based on equilibrium thermodynamics requires a significant degree of interpretation of textural and compositional variations. For samples that underwent several reaction stages at high temperature it is highly likely that compositional information has been modified or even lost from these earlier stages in the reaction history. Alternatively, it is by no means certain that all temperature and pressure changes of a rock's environment are actually recorded in textural and compositional information (Figure 1D). Thus the simple linking of different PT points to form a PTt path, although common practice, should not be accepted as the only solution.

The quantitative study of metamorphic rocks is essentially subdivided into three different but interrelated fields, namely formation, preservation, and exhumation. The PTt path is a historical record of these factors. The most simple factors controlling the formation of metamorphic rocks are changes in temperature and pressure. The interrelationship between rate of temperature change, rate of pressure change, and tectonic environment is the fundamental feature of a PTt path. The absolute pressure and temperature controls the thermodynamic driving force for reaction, i.e., which minerals should form and with which compositions. The actual formation and preservation of minerals is related to kinetic factors that control the rates of material transport processes, mineral breakdown or growth, and compositional homogenization. Exhumation of metamorphic rocks, leading to cooling and decompression, is a necessity if they are to be sampled at the Earth's surface. Processes such as surface erosion, tectonic extension, buoyancy, or a combination of these may all play a role and may also yield characteristic PTt-path segments. All three factors can be modelled for given tectonometamorphic scenarios so the input parameters for models, the shape of resultant PTt paths, and their relative importance will be outlined in turn.

Pressure–Temperature Controls

Pressure, as a geological parameter, is easy to understand. This is purely the average of the stresses applied to the rock. The main stress usually corresponds to

the force of the mass of rocks overlying the point of interest (lithostatic pressure) and can be determined as $P = \rho g h$ (where P = pressure in pascals, ρ = density of the overlying rock in kilograms per cubic metre, g = acceleration due to gravity = 9.81 m s^{-2}, and h = depth of the point of interest in metres). Pressure in the geological literature is usually given in kilobars (1 bar = 10^5 Pa, 1 kbar = 10^8 Pa), megapascals (1 Mpa = 10^6 Pa) or gigapascals (1 Gpa = 10^9 Pa). For typical rock densities (granite: 2700 kg m^{-3}; basalt: 3000 kg m^{-3}; peridotite: 3300 kg m^{-3}), it takes a thickness of 3.8 km of granite, 3.4 km of basalt, or 3.1 km of peridotite to produce a pressure of 100 MPa. Alternatively, pressure increases at a rate of between 26.5 MPa (for granite), 29.4 MPa (for basalt), and 32.4 MPa (for peridotite) per kilometre depth. Generally, pressure deep in the Earth is said to be hydrostatic, i.e., equal in all directions, but in some situations pressures in the crust may be higher than that calculated for the depth alone due to the accumulation of directed horizontal stresses (nonhydrostatic or deviatoric stress) during collision. These stresses cannot build up to significant levels, especially at higher temperatures, because rocks will break or flow to reduce stress build-up, but they are very important for deformation and the production of oriented fabrics in rocks. This means that, within error, pressure can be correlated with depth.

In contrast, the temperature distribution in the crust is not so easy to estimate because it is necessary to know about heat sources, heat sinks, and heat transport as well as how these factors are affected during tectonometamorphism. Heat sources in the crust are either internal, external, or transported (advected). Within the crust the internal heat sources are the heat derived from the decay of radioactive elements, especially K, U, and Th, exothermic mineral reactions (typically retrograde reactions), crystallizing magmas (latent heat of crystallization), or deformation (frictional or shear heating). The radioactive elements do not fit well into the structure of mantle minerals—they are thus termed incompatible—and so over geological time they have been preferentially removed from the Earth's mantle, during melting episodes, and enriched in the continental crust, especially in its upper part. It is the distribution of these radioactive, heat-producing elements in the crust that plays an important role in the possible temperature–depth pattern in the crust before, during, and after a tectonometamorphic event. External, with respect to the crust, is heat added from the underlying mantle, whereas transported or advected heat is that introduced by intruded (magmas) or tectonically emplaced (crustal slices or nappes) hotter bodies. Heat sinks, the consumers of heat energy, are endothermic mineral reactions, the minerals themselves via their heat capacities (i.e., the amount of heat needed to be added in order to raise their temperature), and the atmosphere (whereby heat is ultimately lost from the solid Earth at the surface and can be measured as the surface heat flow). Heat transport in the crust is primarily by either conduction (relatively slow) or advection (relatively fast) with radiation being generally insignificant.

Simple Models: The Stable Geotherm

Before we can start to consider the thermal consequences of tectonometamorphic events we must firstly consider the stable temperature distribution in the crust—in its simplest form this is the stable (or steady-state) geotherm. From measurement of the actual heat flow from the Earth at the present-day surface, realistic models for the internal heat production potential (i.e., higher amounts in the upper crustal than in lower crustal rocks), and typical values for heat conduction (measured on rocks in the laboratory), it is possible to estimate the mantle heat input. Then, under the assumption that all rocks are at equilibrium and therefore that the only heat sink is the surface, it is relatively easy to calculate the temperature distribution in the crust, the so-called steady-state geotherm, in a one-dimensional model. This method works well in stable continental regions (Figure 2, right-hand side). Here the measured flow of heat at the Earth's surface, surface heat flow (Q), is around 50–70 mW m^{-2}. From the heat flow alone and application of Fourier's Law

$$Q = -k\mathrm{d}T/\mathrm{d}z$$

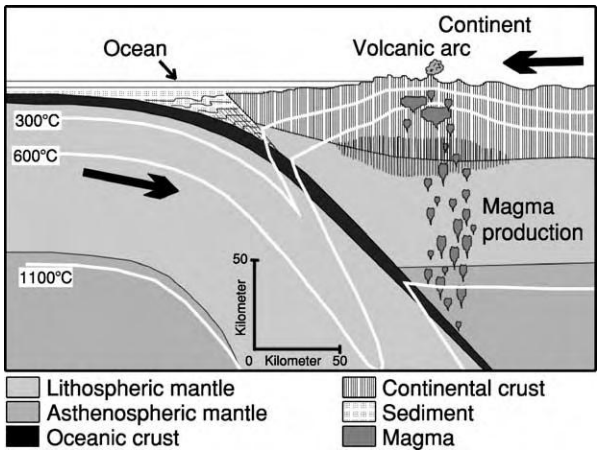


Figure 2 Simplified cross section of an active continental margin showing the different thermal gradients in the normal crust, magmatic arc, and subduction zone domains.

where Q = heat flow (W m $^{-2}$ or J m $^{-2}$ s $^{-1}$), k = thermal conductivity (W m $^{-1}$ K $^{-1}$ or J s $^{-1}$ m $^{-1}$ K $^{-1}$), T = temperature (K), and z = depth (m). This heat flow value corresponds to a thermal gradient (rate of change of temperature with depth, i.e., dT/dz) of 22–31°C km $^{-1}$ taking an average thermal conductivity of 2.25 W m $^{-1}$ K $^{-1}$ (Figure 3A). Consider the environments represented by an active continental margin (Figure 2). Here the subduction of cold oceanic crust causes transfer of heat downwards to the subducting slab and so surface heat flow is lower than in the craton, $Q < 40$ m W m $^{-2}$, and the thermal gradient is also lower (<18°C km $^{-1}$). In contrast, the addition of magma in the arc regime results in a much higher heat flow (Q = 80–200 m W m $^{-2}$) and a correspondingly higher thermal gradient (36–90°C km $^{-1}$). These three different thermal regimes define distinctly different trends within the framework of the metamorphic facies scheme (Figure 3B). The subduction zone trend is one of relatively low temperatures for high pressures, the magmatic arc regime is of relatively high temperatures at low pressures, whereas the cratonic trend lies in-between. If it were possible to drill down into the crust in each of these environments then the formation conditions for the mineral assemblages in the extracted rocks, when plotted in a metamorphic facies diagram, would document these distinctly different trends. The defined trends would not be linear, however, as the irregular distribution of heat-producing elements in the crust defines thermal gradients in the crust that are curves (geotherms) and not simple linear extrapolations of these thermal gradients to depth (Figure 3C).

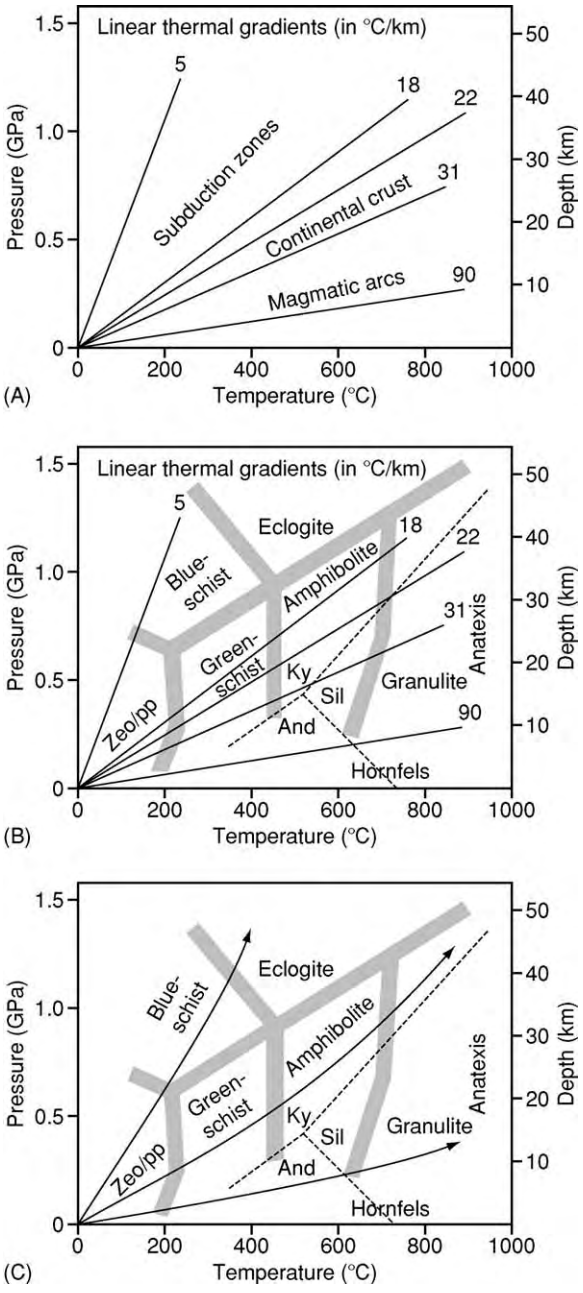


Figure 3 Simple linear temperature depth trends for subduction, continental crust, and magmatic arc environments (A), the same gradients superimposed on the metamorphic facies grid (B), and more realistic curved geotherms (C) for the same environments.

Simple Models: Perturbation of the Stable Geotherm and Regional Metamorphism

Many plate tectonic processes result in a disturbance or perturbation of these simple temperature–depth patterns. This is because the three-dimensional distribution of heat sources and sinks is also disturbed as a result of crustal stacking, crustal extension, or magmatic intrusion. Consider modern examples of continental collision zones such as the European Alps or Himalayas. Here the present-day crust, deduced from geophysical methods, reaches a thickness twice that normally expected. Assuming a normal crustal geotherm existed before collision, the result of the thickening event has significantly changed the distribution of heat-producing elements such that the potential stable geotherm for the thickened crust is very different and the perturbed geotherm will tend to 'relax' towards this new stable state. In addition, the thickened crust is likely to be mechanically unstable and so will, by a combination of erosion and extension processes, gradually return to a stable thickness. The consequence of crustal thickening, thermal relaxation, and erosion/extension for a discrete rock 'parcel' in the crust is a change in depth and temperature over time, i.e., a PTt path.

A typical PTt evolution for a rock buried and exhumed during a collisional orogeny is depicted in Figure 4. Such PTt paths are calculated by solving the heat flow equation

$$dT/dt = \kappa(\delta^2 T/\delta z^2) + A/(\rho c) - U_z(dT/dz)$$

where T = temperature (K); t = time (s); z = depth below the surface (m); A = radioactive heat production (μW m^{-3}); ρ = density (kg m^{-3}); c = heat capacity (J kg^{-1} K^{-1}); κ = thermal diffusivity (= k (thermal conductivity)/$c\rho$) (W m^{-1} K^{-1} or J s^{-1} m^{-1} K^{-1}); U_z = vertical (exhumation) rate (ms^{-1}). Solving the equation involves heat conduction, heat production, and heat convection components, in this case in one-dimension, for a specified range of parameters. The PTt path follows a single point in the crust, 'A', lying on a prethickening geotherm (dashed line), that firstly shows a marked pressure increase (the crustal thickening stage) to point 'B' on the geotherm that exists immediately after thickening. If this thickened crustal state persisted, then the resulting thermal state would be the T_∞ geotherm (Figure 4). Thermal relaxation of the post-thickening geotherm towards the T_∞ geotherm results in heating of the observed rock unit. If, as is usual, thermal relaxation is accompanied by erosion or tectonic thinning, then any rock unit would show heating to a thermal maximum ('C' in Figure 4), reached at a pressure below that of the pressure maximum, before cooling. In this case the T_∞ geotherm is never reached because the factors controlling the shape of the geotherm, such as the distribution of heat-producing elements, have been continually modified. Instead, a new geotherm will emerge once erosion and tectonic thinning has ceased. The PTt path resulting from this tectonometamorphic event has the form of a clockwise loop in a pressure–temperature diagram drawn with the pressure-axis upwards. (Note that the same path would be anticlockwise if the pressure axis is downwards.) The segment from A to C is designated the prograde path and that after C is the retrograde part of the PTt path.

It should be noted that many published PTt paths have been modelled by a one-dimensional solution to the heat production–conduction–advection equation, which only allows instantaneous thickening. A consequence for models of crustal thickening by overthrusting is the generation of a saw-tooth-shaped geotherm because the initial post-thickening geotherm is simply a segment of a normal stable geotherm stacked on top of itself (Figure 5). This perturbed geotherm then relaxes over time. In reality, and as realized in two-dimensional modelling, thermal conduction between over- and underthrusting units works to overcome the advected heat component transported in the hanging wall such that the saw-tooth geotherm never occurs. Only in a large-scale underthrusting situation, as in a subduction zone, can an overturn or inversion in the thermal gradient occur. Thus, the oft-shown instantaneous isothermal (i.e., at constant temperature) pressure increase (A to B in Figure 5), followed by an isobaric (i.e., at constant pressure) heating stage for tens of millions of years, is an artefact of one-dimensional thermal models that should be disregarded.

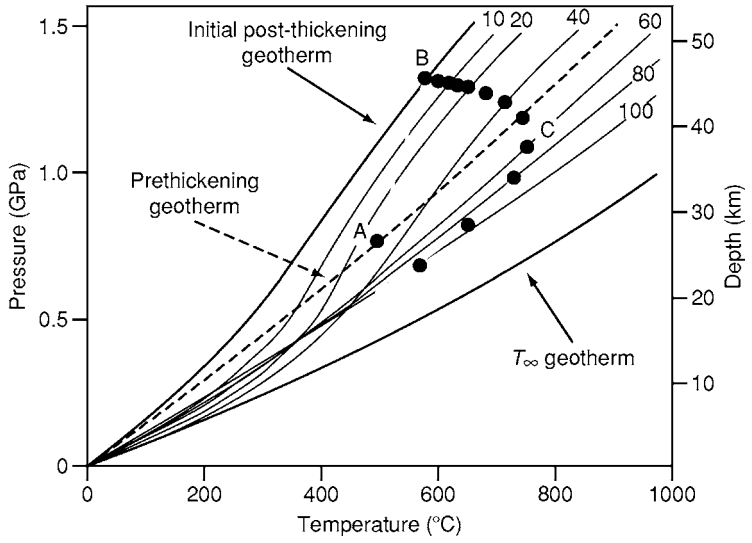


Figure 4 Crustal geotherms before and after a thickening event. The numbers indicate time, after the end of the thickening event, for development of that particular geotherm. A, B, and C mark the prethickening starting point, the pressure peak, and the temperature peak, respectively, for the observed PTt path.

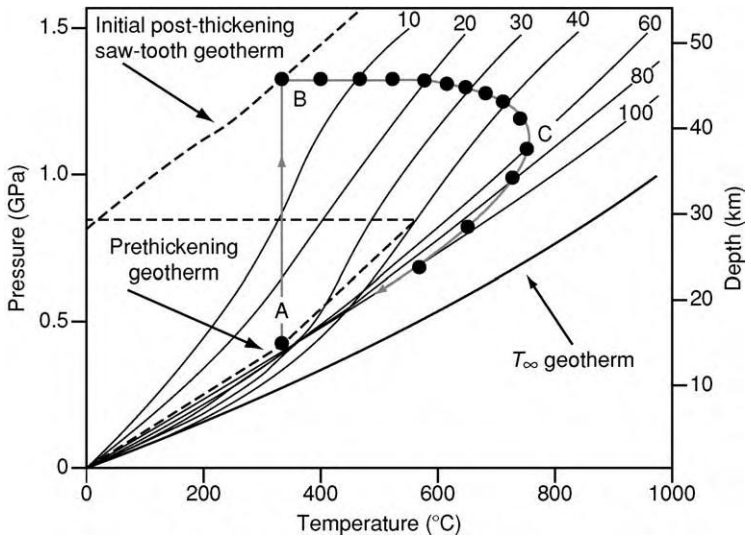


Figure 5 Schematic PTt path for crustal thickening by instantaneous overthrusting as typical for one dimensional modelling. Point A on the prethrusting geotherm undergoes an isothermal pressure increase to B followed by isobaric heating for 20 Ma before erosion is initiated.

How can the clockwise PTt loop of Figure 4 be used as an aid to understanding the PTt evolution of natural metamorphic rocks? The PTt loop shows a wide range in PT conditions and, if superimposed on a facies diagram, would show an evolution passing through several different facies. Mineral reactions and diffusive transport are, however, thermally activated with kinetics following an Arrhenius relationship. For this reason, reaction rates increase exponentially with temperature and therefore the peak temperature is likely to be the point where most reaction occurs. Cooling after the thermal peak will see a slowing of reaction and material transport such that the peak temperature mineral assemblage is preferentially preserved. In addition, the general shape of dehydration reactions means that a higher degree of fluid release, useful for material transport during prograde reaction, will occur for a path of increasing temperature with minor pressure change. Once fluids have left the system at the thermal peak, any retrogression to hydrous assemblages during cooling will be hindered by the absence of a free fluid phase thus again favouring preservation of the peak temperature assemblage. Combining this information it is possible to predict the most likely determinable PT point for a rock that followed a standard clockwise PTt path, based on preserved mineral assemblage for the temperature peak.

If, for a single segment of crust, several rocks formerly at different depths are traced, it is possible to plot the loci of their peak temperature points (Figure 6A). These are the most probable PT points that would be determinable for rocks that had followed these particular paths and represents the sort of information available from field geology for a tilted and peneplained regional metamorphic terrane: hence the name metamorphic field gradient (sometimes also piezothermic array). Several important points can be deduced from this plot especially when additional temperature–age (Figure 6B) and depth–age plots (Figure 6C) for the same model are presented. Firstly, the metamorphic field gradient defines a PT trend that lies between that of the initial and final geotherms. However, the points defining this trend represent different ages, visible from Figures 6B and 6C, and so the curve does not represent any actual, temporary geotherm that occurred during the thermal relaxation. Thus, the sequence of preserved metamorphic rocks is not the same as would be predicted, for example, along one of the Figure 3 thermal trends. The age difference between the different preserved points is also important. If a major deformation event occurred at 25 Ma, this would be syntectonic with respect to rock B but would be pretectonic for rock A and post-tectonic for rock C. In addition, the difference in depth of the peak-temperature points (see Figure 6C) is less than the true depth difference, thus leading to an apparent thinning of the crustal section. Such thermal models for crustal thickening followed by exhumation, although relatively simple, illustrate some of the fundamental problems inherent in trying to interpret natural metamorphic sequences.

Contact Metamorphism

So far the models presented have considered only the effects of tectonically induced crustal-scale

disturbance of geotherms. These have a time-scale measured in tens of millions of years. A further possibility to disturb a geotherm is by the introduction of heat transported (advected) by magmas. This is a discrete quantity of heat, quickly emplaced as a result of intrusion, and dissipated due to conduction, mineral reaction (i.e., contact metamorphism), and perhaps also hydrothermal convection. In contrast to regional metamorphism, contact metamorphism is more restricted in range and has a duration of only a few million years. In many cases, contact metamorphism is a simple thermal process not related to erosion or tectonic thinning. In such cases, PT paths are essentially isobaric heating trends starting from an initial geotherm, quickly reaching a peak, and then cooling back to the initial geotherm. This is illustrated for the case of a 2-km wide, vertical, mafic intrusion initially at 1175°C, intruded into country rocks with an initial temperature of 300°C, solved by a conduction equation:

$$T_{(a,t)} = T_R + 0.5(T_I - T_R)\ (\mathrm{erf}[(a - x)/(2\sqrt{\kappa t})] + \mathrm{erf}[(a + x)/(2\sqrt{\kappa t})])$$

where $T_{(a,t)}$ = temperature (°C) at point x at time t; T_I = temperature of the intrusion; T_R = preintrusion temperature of the host rocks; a = half-width of the intrusion (m); x = distance from the centre of the intrusion (m); κ = thermal diffusivity of the rocks ($m^2\,s^{-1}$), in this example taken to be 1×10^{-6} $m^2\,s^{-1}$; t = time (s); and erf is the mathematical error function.

The equation is solved for points located at different distances from the intrusion and the resultant T–t-curves plotted in Figure 7. The following features should be recognized. The highest contact metamorphic temperature is reached close to the intrusion, this peak temperature is reached quickly, and the time period where the rock is close to T-max is short. Further away from the contact, the peak temperature is lower, it is reached later, and the duration of the period close to T-max is much longer. The duration of the thermal disturbance is only a few million years,

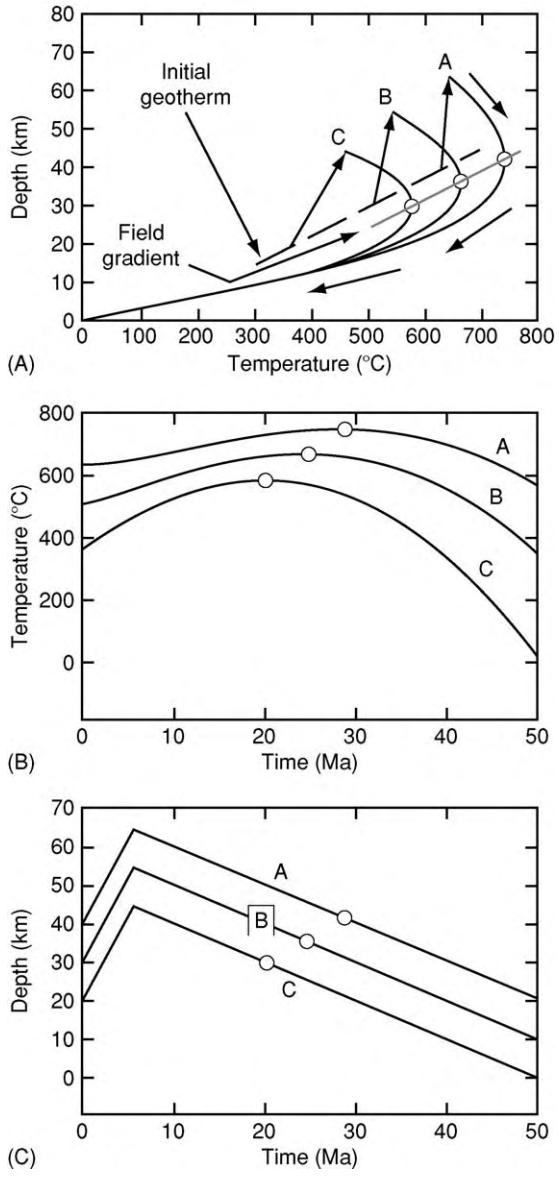


Figure 6 Schematic pressure temperature, temperature time, and pressure time paths for rocks originally sitting on the stable geotherm at depths of 40 (path A), 30 (path B), and 20 (path C) km that were overthrust by a 30 km crust unit in the first 5 Ma and then underwent erosion at 1 mm a^{-1}. Note age difference for peak temperature and depth at peak temperature.

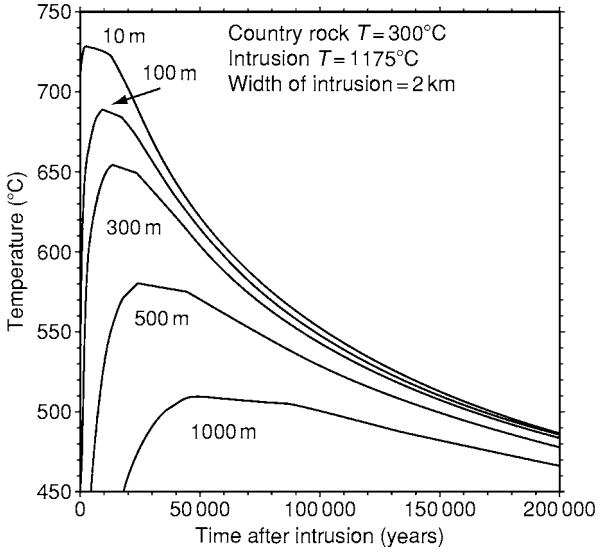


Figure 7 Temperature time paths for positions at different distances from a 2 km wide intrusion emplaced at a temperature of 1175°C into country rocks at 300°C (conduction model).

much shorter than the tens of millions of years seen for regional metamorphism, and the variation in *T–t* paths over a small distance is much greater. For these reasons there is a considerable variation, for a given bulk composition, in mineral assemblages as well as textures in a contact aureole.

Age Determination: The Small 't' in PTt

Qualitatively, the reaction sequence in natural rocks is often visible in the form of growth features, compositional zoning, or breakdown textures. Using simple 'rule-of-thumb' techniques these changing assemblages can be attributed to a sequential path in the metamorphic facies diagram. The sequence tells us the relative age but not the absolute age of the different stages. Did the rock experience a monophase metamorphism, a multistage reactivation during a single orogeny, or a complex polyphase evolution? This is only possible by determining the ages of the different metamorphic stages. Weakly metamorphosed rocks may still preserve fossils, thus putting a maximum age limit, or rocks may have fossiliferous deposits overlying them, thus giving a minimum age limit. More commonly, only isotopic dating of minerals within the studied rock, or from related intrusive of extrusive rocks, can yield absolute ages. The time-dependent decays of radioactive isotopes in the systems Rb–Sr, K–Ar, U–Th–Pb, Nd–Sm, or Lu–Hf are commonly used in geochronology. Important to recognize here is that minerals used for isotopic dating show different resistances to loss of radiogenic isotopes important for geochronology. Some minerals, like zircon, can survive even the highest temperatures unscathed, whereas other minerals, like micas and feldspar, only start their use as geochronometers at lower temperatures. The last few kilometres of exhumation may be recorded in the number of unhealed fission-induced traces in minerals such as apatite, titanite, or zircon. Although radioactive decay is time-dependent, the preservation of daughter isotopes or fission tracks is temperature–time-dependent because of the diffusion processes acting to homogenize the phases in question. Generally, in order to quantitatively determine a PTt path, it will be necessary to extract age information using a number of different methods on different minerals.

PTt Paths as Interpretative Tools

The shapes of certain PTt paths or path segments are often characteristic. A rock that remains in the subduction channel for both its prograde and retrograde history will generally stay cool and thus show a 'hairpin-shaped' PTt path (1 in Figure 8). The clockwise path described already is typical for many crustal collision orogens, although there has been considerable argument as to whether the shape of the retrograde path reflects erosion or extension (2 in Figure 8). For a given crustal segment an extensional path should cool quicker but the poor resolution of PTt paths derived for natural rocks rarely allows such a distinction. If exhumation occurs much faster than conductive loss to the surface, the shallow-level, early-exhumed rocks may receive a short-lived thermal pulse. In contrast, some rocks exhibit anticlockwise (for pressure upwards) PTt paths—paths requiring heat addition during burial followed by cooling accompanied by little erosion or extension (3 in Figure 8). This is a significantly different tectonometamorphic scenario only seldom reliably demonstrated. Important PTt path segments that are characteristic are isobaric heating, isobaric cooling, or isothermal decompression paths (6, 5, and 4 in Figure 8). The latter two are very commonly deduced, especially in granulites, where they are segments of incompletely-defined overall paths, and reflect distinct differences in exhumation rates. An isothermal decompression path segment may reflect a fast exhumation in response to buoyancy or extensional processes, whereas an isobaric cooling path segment may reflect the final cooling back to a stable geotherm after the crust has regained isostatic equilibrium. Some rocks may show both these characteristic paths at different times of their history. An isobaric heating path is most probably a reflection of magmatic heat addition or of mantle processes and is most commonly a minor segment of a more usual clockwise path (the spike in path 6 of Figure 8).

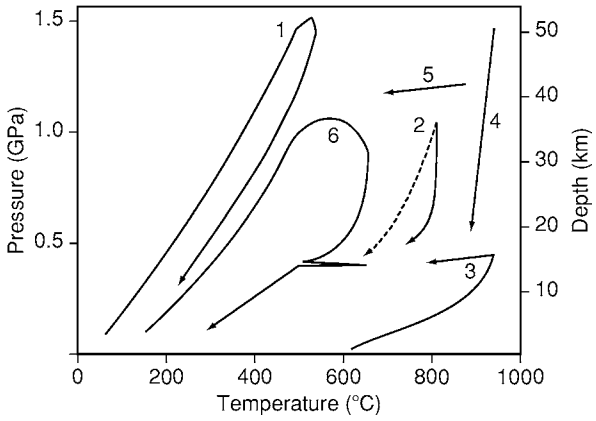


Figure 8 Schematic PT paths and path segments. (1) Subduction zone 'hairpin'; (2) erosional (solid line) or extensional (dotted line) exhumation from the same PT point; (3) anticlockwise path; (4) isothermal decompression; (5) isobaric cooling; (6) clockwise path with short lived contact metamorphic isobaric heating cooling segment superimposed.

Conclusions

The PTt path is the curriculum vitae of a rock. Regional or local variations in characteristic PTt paths may allow the boundaries between large-scale rock units to be defined or allow major tectonic boundaries to be identified. The difficulty is that the rocks deliver only a few 'snapshot' positions on a PTt path and do not actually fully define the path. The quantitative estimation of PT conditions, and ages, for the formation of the identified metamorphic minerals or mineral assemblages, requires both luck and judgement: luck in finding an 'informative' sample and judgement in deciding reaction sequence and which phases, or parts of phases, should be analysed chemically and/or isotopically. However, the PTt data from natural rocks deliver important boundary conditions for the theoretical mathematical modelling of orogenic processes. The actual numerical values of the parameters controlling production, transport, and consumption of heat energy, and mechanical stability driving tectonic thickening and thinning of the lithosphere, are so varied in their possible individual ranges that a definitive model for any actual orogenic event does not exist. Regardless, such models must predict: (1) the PT conditions recorded by the mineral assemblages and (2) the heating/cooling, burial/exhumation rates (if the ages or relative ages of mineral assemblages are available). In other words, the true test of the model is how it relates to reality—in this case the actual rocks that have been formed.

See Also

Analytical Methods: Geochronological Techniques. **Metamorphic Rocks:** Facies and Zones. **Plate Tectonics**. **Regional Metamorphism**. **Tectonics:** Mountain Building and Orogeny. **Thermal Metamorphism**.

Further Reading

Bohlen SR (1987) Pressure temperature time paths and a tectonic model for the evolution of granulites. *Journal of Geology* 95: 617 632.

England PC and Molnar P (1990) Surface uplift, uplift of rocks and exhumation of rocks. *Geology* 18: 1173 1177.

England PC and Richardson SW (1977) The influence of erosion upon the mineral facies of rocks from different metamorphic environments. *Journal of the Geological Society, London* 134: 201 213.

England PC and Thompson AB (1984) Pressure temperature time paths of regional metamorphism I. Heat transfer during the evolution of regions of thickened continental crust. *Journal of Petrology* 25: 894 928.

Ernst WG (1973) Blueschist metamorphism and P T regimes in active subduction zones. *Tectonophysics* 17: 255 272.

Harley SL (1989) The origin of granulites: A metamorphic perspective. *Geological Magazine* 126: 215 247.

Spear FS (1993) *Metamorphic Phase Equilibria and Pres sure Temperature Time Paths*. MS America Monograph. Washington, DC: Mineralogical Society of America.

Spear FS and Peacock SM (1989) *Metamorphic Pressure Temperature Time Paths*. Washington, DC: American Geophysical Union Short Course in Geology.

Stüwe K (2002) *Geodynamics of the Lithosphere: An intro duction*. Berlin: Springer Verlag.

Thompson AB and England PC (1984) Pressure tempera ture time paths of regional metamorphism II. Some petrological constraints from mineral assemblages in metamorphic rocks. *Journal of Petrology* 25: 929 955.

Treloar PJ and O'Brien PJ (eds.) (1998) *What Drives Meta morphism and Metamorphic Reactions?* Geological So ciety, London: Special Publications, 138.

METEORITES

See **SOLAR SYSTEM: Meteorites**

MICROFOSSILS

Contents

Acritarchs

K J Dorning, University of Sheffield, Sheffield, UK

Introduction

The acritarchs are a very large group of organic-walled microfossils of unknown or uncertain affinity. They are mostly unicellular, or apparently unicellular, although they may be found in clusters. The acritarchs are an informal, polyphyletic, organic-walled microfossil group, conceived as a holding category for the very large numbers of *incertae sedis* recorded in palynological assemblages that have no clear affinities. Many acritarchs are likely to be the cysts, temporary resting stages, or phycoma produced as part of the life cycle of marine planktonic algae. A few forms are recorded in apparently lacustrine, fluvial, and terrestrial environments. They were formerly regarded as part of the hystrichospheres (Hystrichophyta) until the type material and many others from the Mesozoic and Tertiary were recognized to be dinoflagellate cysts. Some others are now considered to be representatives of the Prasinophyta or Chlorophyta, although, in the Palaeozoic, they are generally studied as 'acritarchs and associated forms', because many described taxa show little or no evidence of their affinities. A few forms have morphological similarities with copepod eggs and masuelloids. The acritarchs form the main fossil record of global photosynthetic production during the Late Precambrian and Early Palaeozoic, prior to the colonization of the land by plants and consequent terrestrial photosynthesis. There are over 7000 species described, which, given the number of undescribed forms, would suggest that there are well over 10 000 acritarch species preserved in the fossil record. Acritarchs are routinely recorded in abundance in Late Precambrian, Palaeozoic, and Early Mesozoic marine sediments, and are regularly recorded, together with dinoflagellate cysts, in later Mesozoic and Cenozoic marine sedimentary sequences. They are generally studied by examining palynological preparations produced by dissolving rock samples in hydrochloric and hydrofluoric acids in specialist laboratories, although they can also be examined in rock thin sections. Most routine studies are undertaken with strew mounts using transmitted light microscopes, with ultrastructural studies using the scanning electron microscope and transmission electron microscope. Because of their small size, abundance, and diversity, they are of considerable importance in geological interpretation, including biostratigraphy, palaeoenvironmental interpretation, and geothermal alteration within depositional basin analysis studies.

Acritarch Occurrence, Preservation, and Geothermal Alteration

Acritarchs are regularly recorded in abundance from siliciclastic and carbonate marine sediments, including mudstones, silty mudstones, sandy mudstones, shales, siltstones, sandstones, and limestones. In many sequences deposited in shelf sea areas, there is a continuous record of the preserved phytoplankton productivity. Generally, they are most abundant in fine-grained sediments, because of the slower rate of sedimentation. Specimens preserved in grey silty mudstones and muddy siltstones are often of excellent preservation, reflecting the dysoxic to anoxic conditions at the sediment–water interface. Specimens recorded from sandstones and bioclastic limestones are of very varied preservation, depending on the oxygen availability in the sediment. Those recorded in highly oxic conditions are often badly degraded,

whilst some sandstones and early diagenetic limestones can preserve specimens in three dimensions, uncrushed between the mineral grains. Forms recovered from laminated dark shales deposited with slow rates of sedimentation are normally flattened, and are often associated with the growth of minute pyrite framboids that can impact on the internal and external walls of palynomorphs, including acritarchs. Recrystallization in limestones, particularly dolomites, may also distort the acritarch wall. In areas of deep burial or high heat flow, the temperature increase has the effect of changing the colour of acritarchs from transparent to pale greenish-yellow, through increasingly darker shades of brown, to grey or black, depending on the wall thickness (Figure 1).

These series of colour and transparency changes with different temperatures are affected by the wall thickness, as well as by the composition of the different complex organic compounds, comparable with the changes that can be observed with complex carbohydrates when preparing toast. Abundant long-ranging forms, including *Leiosphaeridia* and *Veryhachium*, have been used to estimate maximum rock palaeotemperatures. This colour change has little impact on the preservation of the acritarchs until temperatures associated with the transition of mudstones to slate are reached, when the cleavage that penetrates the sediments and the associated shrinkage of the organic matter tend to fragment the acritarchs. In these situations, thin sections made parallel to the bedding may be useful for identifying the acritarchs present. In the Ordovician, Silurian, and Devonian, acritarchs are often recorded in palynological assemblages together with lower numbers of chitinozoans, graptolite fragments, and scolecodonts, the jaw apparatus of annelid worms. From the Middle Triassic, forms recognizable as dinoflagellate cysts may also be present in the marine microflora; from the Middle Jurassic, dinoflagellate cysts generally dominate the organic-walled marine microflora in marine shelf sediments.

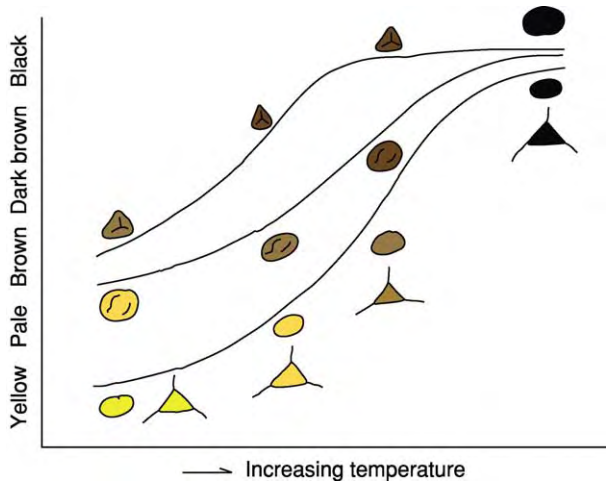


Figure 1 Diagram illustrating the change in acritarch colour with increasing temperature for *Veryhachium*, *Leiosphaeridia* and a thick walled sphaeromorph, in comparison with a trilete spore.

Acritarch Morphology

The general morphology of most acritarchs consists of a hollow, organic-walled, enclosed cell of varied shape that may extend into flanges or processes with a wide variety of morphology. In almost all forms, the wall is very resistant to decay, and appears to be as resistant to degradation as sporopollenin. Forms that show a reflected paratabulation, an angular opening or archaeopyle, or evidence of a sulcus are generally considered to be dinoflagellate cysts. Three main morphologies can be recognized: acritarchs without processes or flanges, acritarchs with flanges but no processes, and acritarchs with processes with or without flanges (Figure 2).

For all acritarchs, the main body is known as a vesicle, with wing-like projections or alae known as flanges and large projections referred to as processes, which in many forms are spinose. Some specimens show evidence of an excystment opening, which

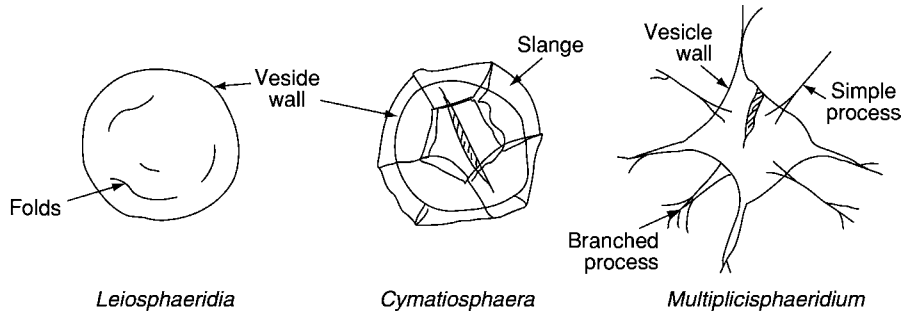


Figure 2 Morphology of *Leiosphaeridia*, a sphaeromorph acritarch, *Cymatiosphaera*, an acritarch with flanges, and *Multiplicisphaeridium*, an acritarch with processes.

enabled the cell contents of the next stage in the life cycle to be released.

The vesicle shape is highly varied, and has been used as a simple method of separating different groupings of acritarchs. The shapes observed include spherical to subspherical, ovoidal to elongate, flattened triangular to rectangular to polygonal (Figure 3).

The vesicle size varies from about 2 to 2000 μm, although many acritarchs fall into the size range of 10–50 μm.

Flanges and Processes

Some acritarchs have thin vertical projections, wings, or alae, known as flanges, projecting from the surface, which divide the vesicle surface into two or more fields. Many of these forms are considered to be prasinophytes, although this is less certain in forms in which processes support the flanges. Projections that arise from the surface, whether spinose or bulbous, are termed processes. These are distinguished from ornament at an arbitrary size of 2 μm. The process shape, number, and distribution on the vesicle surface are of particular importance in the speciation of many acritarchs. The processes may be solid or tubular, tapering or inflated, and may have a sharp to wide base at the contact with the vesicle (Figure 4). In forms with hollow processes, the interior of the process may communicate with the vesicle or be plugged by the inner vesicle wall. The processes may be simple and unbranched, or branched in a regular or irregular manner. The unbranched portion of the process is sometimes called the trunk. The first order of branching forms two or more branches, or pinnae, that are of similar or different length, and at various branching angles. The branches may continue to branch at the second, third, fourth, or more orders in a regular or irregular manner. In addition, the processes may be totally or partly ornamented or, in some forms, interconnected with a thin veil.

Vesicle and Process Wall

There are both single and double vesicle walls, with a wall thickness in the range 0.5–3.0 μm. Several different types of wall structure can be recognized. In the micrhystridian wall type, the single wall is thin, apparently homogeneous in nature, and generally appears to be of similar thickness within a species. The micrhystridian wall type is known in *Micrhystridium* and *Acanthodiacrodium*. In the tasmanitid wall type, the wall is uniform, but laminated with narrow radial pores. The wall is often thick. The tasmanitid wall type is found in *Tasmanites*. Other forms, including *Baltisphaeridium*, have walls with very narrow radial pores. The visbysphaerid wall type includes forms with a double wall, with the inner wall generally thicker and, in transmitted light, visibly darker than the outer wall. The visbysphaerid wall type is found in *Visbysphaera*. Each of the wall types may be smooth (laevigate) or ornamented with granulate, microgranulate, foveolate, costate, echinate, or striate ornament, or a combination of these.

Excystment Opening

Some acritarchs show excystment openings that allowed the cell contents to escape the cyst stage to form the next stage of the life cycle (Figure 5). In some acritarchs, this is a circular pylome, or cyclopyle, with apparently the same function as the archaeopyle in dinoflagellate cysts. Some forms have a particularly large pylome, or macropyle. In others, the vesicle splits into two equal halves by means of a median split. Others show large splits in the vesicle wall to form a wide gape, while a few forms almost split into four quarters, similar to an opening flower. Some specimens show a short split with a flap, also known as an epityche, or an irregular split in the vesicle wall, although these can sometimes be difficult to distinguish from accidental damage as a result of compaction in the sediment. Within each species, the proportion of specimens showing an excystment mechanism can be used in palaeoenvironmental interpretation.

Acritarch Clusters

Most of the organic-walled marine microflora, including the acritarchs and dinoflagellate cysts, occur as isolated individuals in palynological preparations and thin sections. Most are the benthonic cysts or resting stages of the planktonic marine microflora, although some of the cysts or phycoma were probably planktonic. If care is taken during palynological processing, some forms can be recovered as monospecific clusters. In the Late Precambrian, clusters of sphaeromorph acritarchs, including *Leiosphaeridia*, are particularly common. In the Cambrian and Ordovician, clusters of *Acanthodiacrodium*, *Cymatiogalea*, *Leiosphaeridia*, *Micrhystridium*, and *Polygonium* are not infrequently recorded, whilst, in the Silurian, clusters of *Cymbosphaeridium*, *Dilatisphaera*, *Leiofusa*, *Leiosphaeridia*, *Micrhystridium*, and *Veryhachium* are sometimes recorded. Clusters have also been observed in a few Mesozoic acritarchs and dinoflagellate cysts. Typically, all of the individuals in a cluster are identical in size. Clusters are particularly abundant in laminated sediments that lack bioturbation, which would account for their frequency in the Precambrian and their widespread distribution in the Early Palaeozoic, prior to the shelly

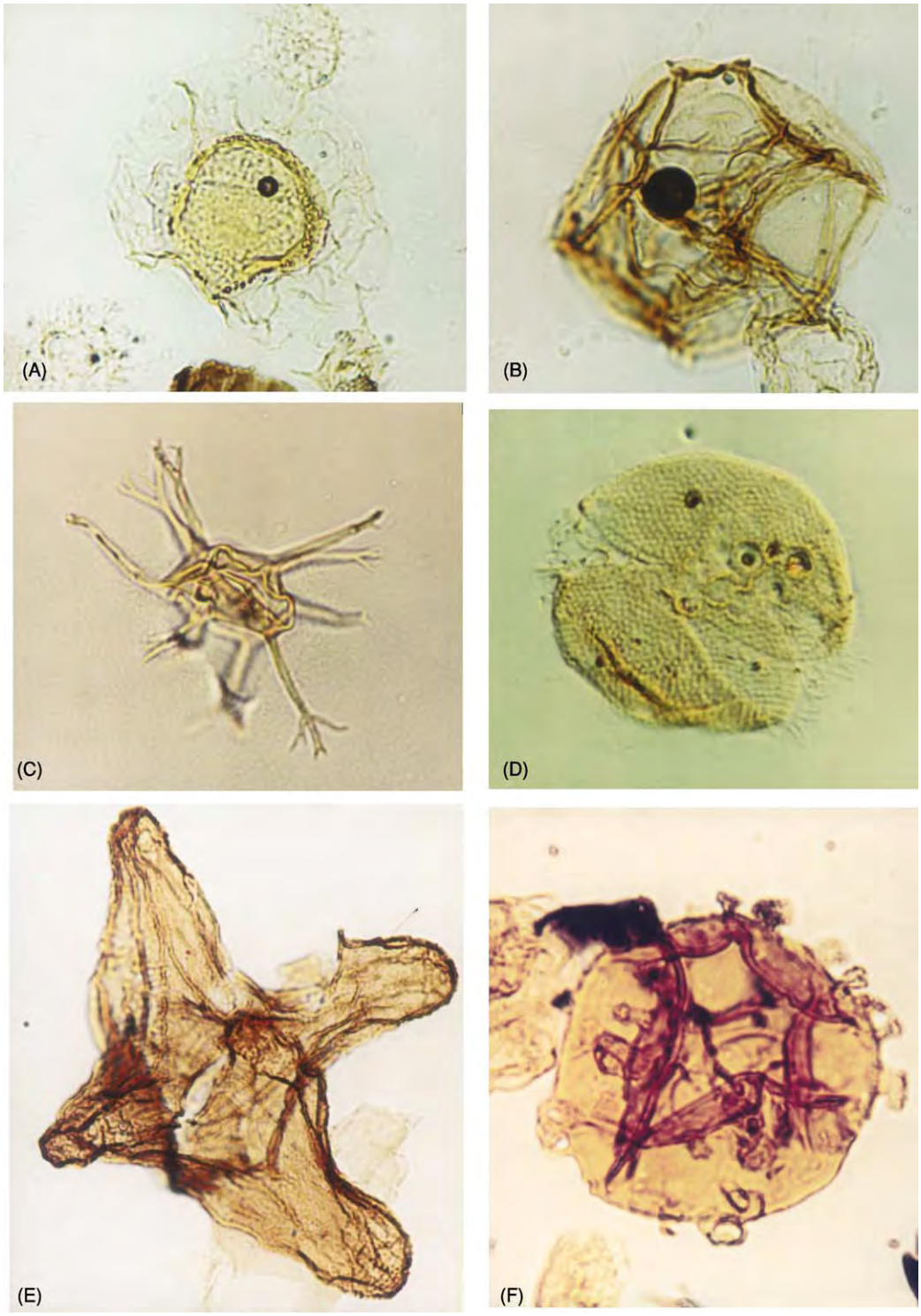


Figure 3 Colour photomicrographs of acritarchs: (A) *Pterospermella* (×1000); (B) *Cymatiosphaera* (×1000); (C) *Multiplicisphaeridium* (×1000); (D) *Helosphaeridium* (×1000); (E) *Pulvinosphaeridium* (×500); (F) *Visbysphaera* (×1000).

macrofossil migration from shallower into deeper shelf areas which caused bioturbation. Although it is possible that some monospecific clusters were derived from production in a sporangia, it is probable that others formed aggregates to make up part of the plankton snow, as a mechanism against predation by zooplankton by increasing the speed of sinking from the surface water to the sediment-water interface.

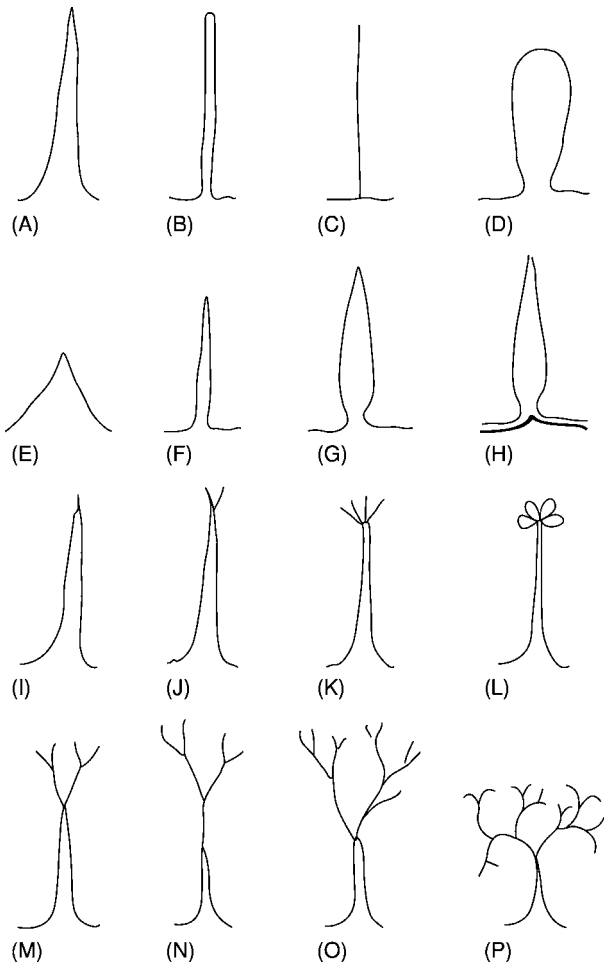


Figure 4 Process morphology: (A) tapering, distally sharp; (B) tubular, distally blunt; (C) solid, wire like; (D) expanded; (E) wide base; (F) narrow base; (G) constricted base; (H) plugged base; (I) simple, unbranched; (J) bifurcate; (K) multifurcate; (L) distal loops; (M) regularly branched in two orders; (N) irregularly branched in two orders; (O) irregularly branched in three to four orders; (P) irregularly branched at a high angle.

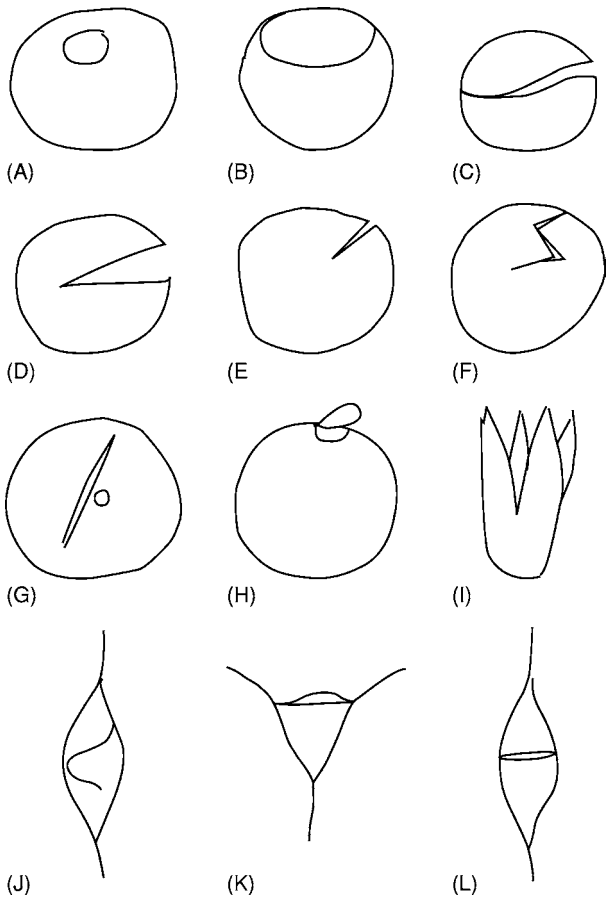


Figure 5 Excystment openings: (A) pylome; (B) macropyle; (C) split into two equal hemispheres; (D) gape split; (E) straight split; (F) irregular split; (G) straight split and small pylome; (H) hinged pylome; (I) split into four; (J) sinusoidal split; (K) flap style epityche; (L) split into two halves.

Acritarch Classification

In general, the acritarchs are considered as form genera, and, as such, it is inappropriate to use a formal classification system. For this reason, the acritarchs are treated as *incertae sedis* microfossils, organic-walled microphytoplankton, or group Acritarcha Evitt, 1963. They have been divided, mostly on the basis of vesicle shape, into a number of informal subgroups. This system of subdivision is less than satisfactory, as there are a number of gradational forms between subgroups, particularly in forms with processes between spherical and polygonal vesicles. None of the subdivisional schemes for the acritarchs is entirely satisfactory, and many research workers list all the acritarchs, algal cysts, and colonial algae together alphabetically when describing palynological assemblages. The following classification refers to the subgroups, but also documents other groupings that may provide progress towards a more natural classification.

Acritarchs without Processes or Flanges (Figure 6)

This category includes the subgroup Sphaeromorphitae, the sphaeromorph acritarchs, together with the Schizomorphitae, Scutellomorphitae, and *Navifusa* group. The grouping is almost certainly polyphyletic, and includes forms attributable to the green algal groups Chlorophyta (Chlorophyceae) and Prasinophyta (Prasinophyceae). Some forms with an indistinct wall may be related to the Cyanophyta (cyanobacteria; blue–green algae). The first organisms that could be seen as acritarchs date from 3800 Ma, but these should probably be considered to be Cyanophyta, with the first sphaeromorph acritarchs of 20–200 μm in diameter known from 1800 Ma, following the rise in oxygen levels in the

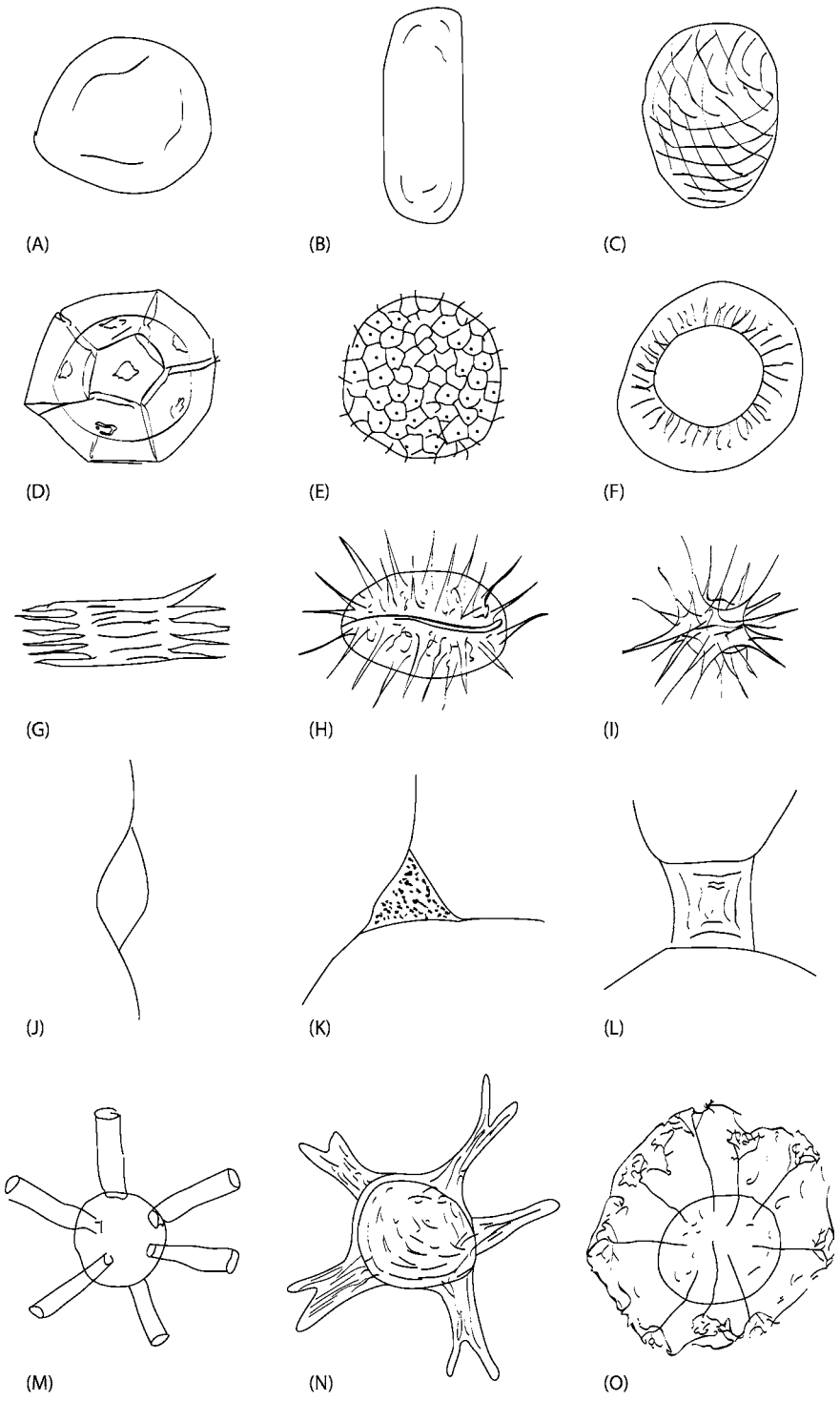


Figure 6 Acritarchs without processes or flanges: (A) *Leiosphaeridia*; (B) *Navifusa*; (C) *Moyeria*. Acritarchs with flanges: (D) *Cymatiosphaera*; (E) *Dictyotidium*; (F) *Pterospermella*. Acritarchs with processes: (G) *Acanthodiacrodium*; (H) *Salopidium*; (I) *Micrhystridium*; (J) *Leiofusa*; (K) *Villosacapsula*; (L) *Neoveryhachium*; (M) *Dilatisphaera*; (N) *Leptobrachion*; (O) *Tunisphaeridium*.

atmosphere. The sphaeromorph acritarchs have a long continuous record through the Late Precambrian, Palaeozoic, Mesozoic, and Cenozoic, and continue through the Holocene to the present day. Some sphaeromorph acritarchs, including forms referable to *Leiosphaeridia* and *Lophosphaeridium*, occur in abundance in apparently freshwater palaeoenvironments, as well as shallow shelf to open ocean

environments. The inclusion of the sphaeromorph acritarchs within the informal group Cryptarcha has not been generally adopted. The clear distinction between some thick-walled sphaeromorph acritarchs, cryptospores, and thin alete plant spores is often difficult in Palaeozoic palynological assemblages. The *Navifusa* group of elongate acritarchs includes marine forms in the Ordovician, and *Quisquilites*, a possible non-marine form, in the Devonian. The *Moyeria* group includes the discoidal acritarchs *Moyeria*, *Chomotriletes*, and *Concentricystes*, acritarchs of apparent terrestrial or fluvial affinity that may possibly have affinities with the Euglenophyta or Prasinophyta.

Acritarchs with Flanges, but without Processes (Figure 6)

This category includes many forms in the subgroups Herkomorphitae (herkomorph acritarchs) and Pteromorphitae (pteromorph acritarchs). Some species attributed to *Cymatiosphaera*, *Dictyotidium*, and *Pterospermella* are clearly related to the Prasinophyta, whilst others are of uncertain affinity. The cysts or phycoma of many prasinophytes are photosynthetic and therefore planktonic, which may explain their wide distribution in the marine realm from shelf seas to the open ocean. The size of individuals within many species of *Cymatiosphaera* and *Dictyotidium* is often rather variable, making consistent identification difficult.

Acritarchs with Processes, with or without Flanges (Figure 6)

This category includes many forms in the subgroups Acanthomorphitae (acanthomorph acritarchs), Coryphomorphitae, Diacromorphitae (diacrodian acritarchs), Netromorphitae (netromorph acritarchs), Oomorphitae (oomorph acritarchs), Polygonomorphitae (polygonomorph acritarchs), and Prismatomorphitae (prismatomorph acritarchs). Some very large acritarchs with rigid processes may have affinities with the masuelloids (Muellerisphaerida), planktonic zooplankton with an organic and phosphate wall that are mostly found in deep-water oceanic sediments.

Several groupings can be recognized in acritarchs with processes, based on wall structure and ornament, process style, and excystment method. These groupings are particularly useful in palaeoenvironmental interpretations.

- *Acanthodiacrodium* group. This group is characterized by a simple wall structure and elongate body shape. The group is particularly abundant in the Late Cambrian to Early Ordovician, where it grades into forms with four corners, attributable to *Coryphidium*.
- *Ammonidium* group. This group has an excystment structure that divides the vesicle into two equal halves. *Salopidium*, *Ammonidium*, and *Gracilisphaeridium* show increasing process complexity. The Silurian acritarchs, *Helosphaeridium* and *Percultisphaera*, have a similar excystment mechanism.
- *Cymatiogalea* group (galeate acritarchs). This group is characterized by a large cyclopyle as found in *Cymatiogalea*, *Priscogalea*, and *Stelliferidium*, and is characteristic of the Early Ordovician.
- *Diexallophasis* group. This group, which includes *Evittia* and *Tylotopalla*, is characterized by an irregular echinate ornament. A few species are found in the Late Ordovician, but it is a common component of the Silurian and Devonian assemblages.
- *Domasia* group. This group of small acritarchs with a simple wall and low number of processes of variable length includes *Deunffia*. It is abundant in the Late Llandovery to Early Wenlock.
- *Diliatisphaera* group. *Dilatisphaera* is unusual in having thin tubular processes that open distally. The wall composition is different from most acritarchs in that it is reluctant to take safranin stain.
- *Duvernaysphaera* group. This group, which includes *Quadraditum*, has an equatorial flange supported by processes. It is possible that the equatorial flange may suggest that it could be included in the Prasinophyta.
- *Eisenackidium* group. This group with a double vesicle wall and a thin flexible process wall includes *Eisenackidium*, with simple processes, and *Leptobrachion*, with branched processes.
- *Eupoikilofusa* group. This group of large elongate acritarchs includes large forms of *Leiofusa*. It is commonly recorded in the Silurian and Devonian.
- *Estiastra* group. These are very large acritarchs with a very thin wall, sometimes with dark blotches, and include *Hogklintia* and *Pulvinosphaeridium*. They are only found in Late Ordovician to Silurian tropical carbonate environments. Their large size, together with an apparently different wall composition, suggests that they might not be cysts of planktonic algae.
- *Micrhystridium–Veryhachium* group. This is a large group of small acritarchs with a simple wall and simple processes. The group, which includes some small forms of *Leiofusa*, is long ranging from the latest Precambrian to Tertiary. Representatives of this group are the main acritarchs with processes in the Permian, Triassic, Jurassic, Cretaceous, and Cenozoic.

- *Onondagella* group. This group has three different processes and a wall that is reluctant to take stain. It is possibly related to the chlorophyta, given that four specimens may be arranged to form a tetrahedral colony similar to *Deflandrastrum*, although this has not been observed.
- *Oppilatala* group. This grouping has a large gape excystment opening and includes *Oppilatala* and some species attributed to *Dateriocradus*.
- *Orthosphaeridium* group. This group with plugged bases to the processes and a straight split as an excystment opening includes *Bacispheridium*.
- *Tunisphaeridium* group. This group with solid, wire-like processes includes *Carminella*, *Geron*, and *Elektoriskos*.
- *Visbysphaera* group. This group normally has a double-walled vesicle and numerous heteromorphic processes. Specimens without the inner wall are often observed.

Biostratigraphy

The Precambrian record of the acritarchs is dominated by the sphaeromorph acritarchs. In the Late Palaeoproterozoic and Mesoproterozoic, the low-diversity assemblages are composed of laevigate sphaeromorph acritarchs. Striate vesicles appear in the Late Mesoproterozoic, with large acanthomorph acritarchs and ornamented sphaeromorphs in the Neoproterozoic. Many acritarchs appear to become extinct at the Varanger ice age, but, after this event, diverse assemblages of ornamented sphaeromorph acritarchs are seen until the latest Proterozoic, where low-diversity laevigate sphaeromorphs are dominant. Acritarchs diversified rapidly in the Early Cambrian, with the first appearance of *Micrhystridium* close to the Ediacaran–Cambrian boundary, and diverse assemblages including *Skiagia* and *Archaeodiscina* soon after. Throughout the Cambrian, Ordovician, Silurian, and Devonian, there are successions of diverse acritarch microfloral assemblages that are of particular value in the biostratigraphy of Palaeozoic successions. Formal and informal biostratigraphical zonation schemes have been proposed for sections in the Cambrian, Ordovician, Silurian, and Devonian, based on assemblage characteristics and the first appearance of distinctive acritarch taxa. Major changes of note include the occurrence of *Eliasum* in the Middle Cambrian, and the proliferation of the *Acanthodiacrodium* and *Cymatiogalea* groups, together with *Vulcanisphaera*, in the Late Cambrian and Early Ordovician in western Europe and eastern North America. Acritarch abundance was also exceptionally high in this area, with values of over 100 000 per gram of rock regularly recorded. Middle Ordovician sediments contain the distinctive acritarchs *Coryphidium* and *Frankea*. Late Ordovician assemblages are characterized by an abundance of acritarchs with plugged process bases, including *Baltisphaeridium*, *Orthosphaeridium*, and *Ordovicidium*, together with *Peteinosphaeridium*. Late Ashgill assemblages associated with the Early Hirnantian glacial events contain many small distinctive acritarchs, including *Villosacapsula* and *Pheoclosterium*, together with cryptospores. The acritarchs show a remarkably low diversity in the Early Llandovery, perhaps associated with continuing cool conditions. In Arabia and North Africa, sediments with a high total organic carbon are associated with acritarch assemblages with large numbers of *Leiosphaeridia*. Cryptospores and recycled acritarchs are regularly recorded in this interval, suggesting that relative sea-levels were still fairly low. By the Late Llandovery, acritarch diversity had recovered, and high-diversity assemblages are routinely recorded from this interval. The Late Llandovery to Early Wenlock is characterized by diverse acritarchs, including species of *Deunffia* and *Domasia* in tropical areas. Diversity continued to be high during the rest of the Silurian, with some evidence of shifts in assemblage related to cyclic alternations from warm and dry to cooler and wetter climatic episodes. Assemblages during the Devonian continued to be diverse, with species of *Poledryxium* being conspicuous, and species of *Cymatiosphaera* becoming an important component. The latest Devonian to Early Carboniferous interval shows a remarkable decline in the number of acritarchs recorded, although this is partly because there is a significant drop in the numbers of the acritarchs of medium to large size, together with many palynological assemblages being dominated by terrestrial spores. Carboniferous acritarchs are of low diversity, with small forms of *Leiosphaeridia*, *Micrhystridium*, *Unellium*, and *Veryhachium* recorded from marine sediments. There are very few records of acritarchs from Late Carboniferous sequences, but, during the Permian, the numbers recovered to some extent, with new species of *Unellium* conspicuous. The Triassic is notable for the abundance of small species of *Micrhystridium* and *Veryhachium*, which continue into the Early Jurassic, together with *Metaleiofusa* and *Multiplicisphaeridium* and dinoflagellate cysts of low diversity. Rapid diversification of the dinoflagellate cysts occurred in open marine shelf areas during the Middle Jurassic, although species of *Micrhystridium* and *Veryhachium* are often found in palynological assemblages through the Mesozoic and Cenozoic. Oceanic sediments continue to contain

acritarchs of small size, and they are notable in the Late Cretaceous pelagic sediments, including the Chalk. Acritarchs and dinoflagellate cysts were first studied in detail within flints in the Chalk, and continue to be of value in the sourcing of archaeological flint and chert. Assemblages in the Cenozoic are also dominated by dinoflagellate cysts, although the marine phytoplankton record also includes the calcareous nanoplankton (Haptophyta, Coccosphaerales) and diatoms (Bacillariophyta). Acritarchs are recorded together with dinoflagellate cysts in Holocene marine sediments, although some may include copepod eggs.

Palaeoenvironmental Distribution

In common with present-day marine phytoplankton, the acritarchs show distributions related to water masses, so that different nearshore, offshore shelf, and basinal assemblages can be recognized (**Figure 7**). Forms that are interpreted as benthonic cysts are most common in shelf areas, whereas forms with planktonic cysts, including photosynthetic phycoma, are widespread and common in deep-water oceanic sediments. On a global basis, some forms are apparently restricted to broad latitudinal belts, comparable to modern phytoplankton. The palaeoenvironmental interpretation of Late Precambrian acritarchs is uncertain, although it is probable that different sizes of sphaeromorph acritarch were influenced by nutrient availability. In the Cambrian, many acritarchs are widespread, although *Micrhystridium* favoured inshore areas, whereas more complex forms, including *Skiagia*, are found in offshore areas. During the Ordovician and Silurian, clear palaeoenvironmental distribution patterns become established, with distinct assemblages associated with nearshore shelf, offshore shelf, and deep-water areas. In the Late Ordovician, nearshore areas contain many *Leiosphaeridia*, with *Peteinosphaeridium* and *Dicommopalla* characteristic of shoal areas, and diverse acritarchs with *Baltisphaeridium*, *Veryhachium*, and *Polygonium* in open sea shelf areas. In the Silurian, nearshore assemblages are of low diversity and low to moderate abundance, with many small, thin-walled *Leiosphaeridia*, together with *Veryhachium* with three short processes, *Micrhystridium*, and *Diexallophasis*, together with land plant spores. The offshore shelf assemblage is of high diversity and moderate abundance, with no one taxa dominating the assemblage. Species of *Veryhachium* often have three or four medium to long processes. This pattern of increasing length of process in an offshore direction is also documented from the Early Jurassic. The assemblages associated with deep water are of low diversity and low to moderate abundance, with *Leiosphaeridia* dominant, particularly thick-walled forms. Other genera present in low numbers include *Cymatiosphaera* and *Pterospermella* that have planktonic phycoma, together with occasional *Diexallophasis*, *Micrhystridium*, and *Diexallophasis*, which are the most common acritarchs on the shelf. There is some evidence from the Silurian to suggest that acritarch assemblages may be used to predict oceanic phytoplankton productivity and palaeoclimate, as certain forms are associated with argillaceous sediments formed during cool, wet, primo episodes, with others associated with more calcareous sediments deposited during drier, warmer, secundo episodes.

During the Silurian and Devonian, distribution patterns are also recognized within carbonate environments, with distinct assemblages associated with patch reef, inter-reef and deeper non-reef areas

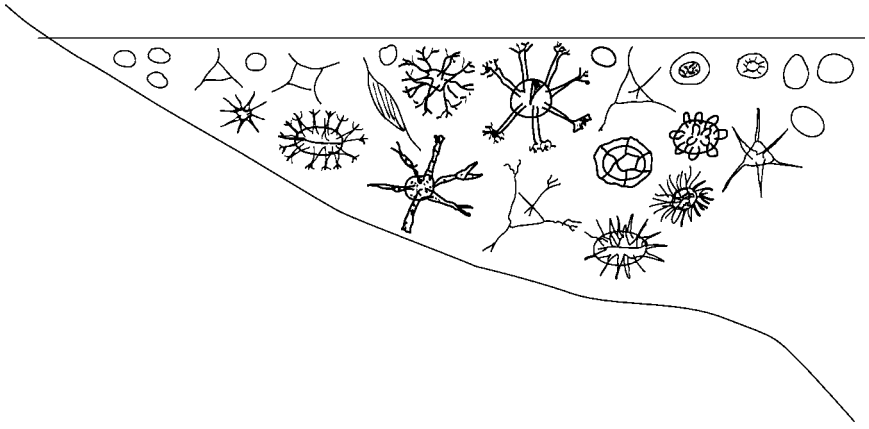


Figure 7 Diagrammatic section from a shallow to deep shelf tropical sea during the mid Silurian, illustrating the distribution of acritarchs. Data in part from Dorning KJ (1981) Silurian acritarch distribution in the Ludlovian shelf sea of South Wales and the Welsh Borderland. In: Neale JW and Brasier MD (eds.) *Microfossils from Recent and Fossil Shelf Seas*, pp. 31 36. Chichester: Ellis Horwood Ltd.

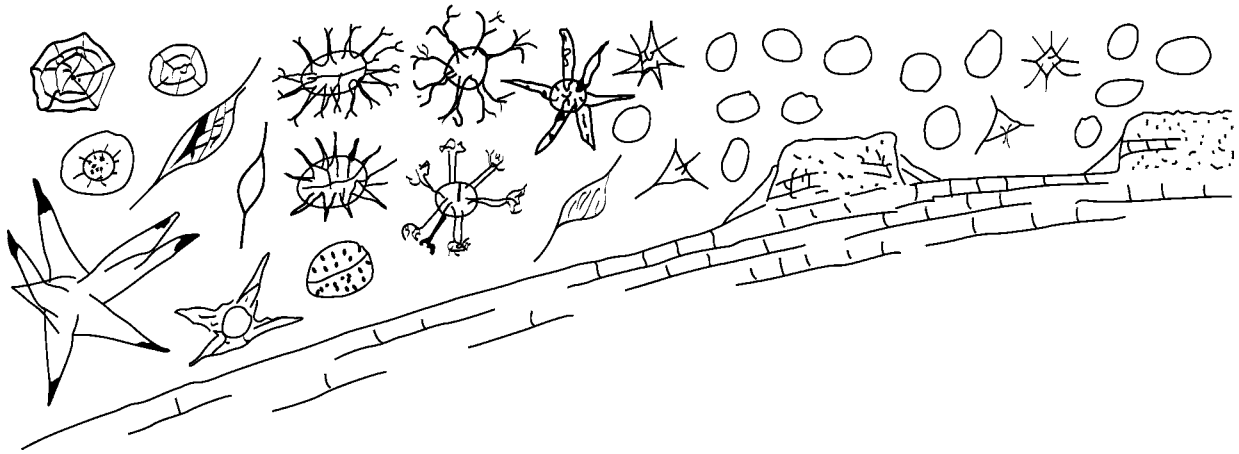


Figure 8 Diagrammatic section from inter patch reef to off reef shelf areas during the Late Wenlock, showing the distribution of acritarchs. Data in part derived from Dorning KJ and Bell DG (1987) The Silurian carbonate shelf microflora: acritarch distribution in the Much Wenlock Limestone Formation. In: Hart MB (ed.) *Micropalaeontology of Carbonate Environments*, pp. 266 287. Chichester: Ellis Horwood Ltd., and Dorning KJ and Harvey C (1999) Wenlock cyclicity, palynology, and stratigraphy in the Buildwas, Coalbrookdale, and Much Wenlock Limestone formations, Shropshire, England. *Bolletino della Societa Paleontologica Italiana* 38: 155 166.

(Figure 8). Reef and reef-associated sediments contain a high percentage but very low numbers of *Leiosphaeridia*. Inter-reef sediments contain much higher numbers of acritarchs, including many species with simple processes, such as *Micrhystridium*. Deeper water areas contain higher diversity shelf assemblages, with many species with branched processes, including *Oppilatala* and *Multiplicisphaeridium*. On a global scale, in the Silurian, cold-water high latitudes are characterized by assemblages with *Neoveryhachium*, large *Leiofusa*, and *Eupoikilofusa* in North Africa and Arabia, whereas warmer, low-latitude areas contain *Deunffia*, *Domasia*, and the large acritarchs *Estiastra*, *Hogklintia*, and *Pulvinosphaeridium*.

Applications

The organic-walled microplankton are the most abundant and diverse microfossil groups from the Late Precambrian through the Palaeozoic and Mesozoic, and are still important in the Cenozoic. Because the acritarchs are the principal phytoplankton in the Late Precambrian, Palaeozoic, and Early Mesozoic, they are the main biostratigraphical tool used in oil and gas exploration and production in marine sediments. Many forms have short stratigraphical ranges, which enable acritarch biohorizons and biozones to be established for the correlation of surface and borehole sections by analysing small samples of core, sidewall cores, or drill cuttings. Some acritarchs are associated with intervals of high total organic carbon, which are the main oil and gas source rock horizons. The colour of acritarchs may be used to estimate maximum palaeotemperatures, which can be employed to provide an indication of the maximum burial depth of rock sequences, in order to predict the timing of oil and gas generation. In regional studies, the distribution of acritarchs can be used to map areas of shallow to deep water as part of basin analysis. A combination of acritarch biostratigraphy, palaeotemperatures, and recycled acritarchs can be used to unravel the geological history of complex tectonic provinces.

See Also

Biosediments and Biofilms. **Microfossils:** Chitinozoa; Conodonts; Foraminifera; Ostracoda; Palynology. **Palaeoclimates**. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils.

Further Reading

Aldridge RJ, Jeppsson L, and Dorning KJ (1993) Early Silurian oceanic episodes and events. *Journal of the Geological Society, London* 150: 501 513.

Commission Internationale de Microflore du Palaeozoique (CIMP) Acritarch Subcommission (from 1986) *Acritarch Newsletter*. www.shef.ac.uk/~cidmdp/ac16.doc

Dorning KJ (1981) Silurian acritarch distribution in the Ludlovian shelf sea of South Wales and the Welsh Border land. In: Neale JW and Brasier MD (eds.) *Microfossils from Recent and Fossil Shelf Seas*, pp. 31 36. Chichester: Ellis Horwood Ltd.

Dorning KJ (1993) Group ('Phylum') Acritarcha Evitt, 1963. In: Benton MJ (ed.) *The Fossil Record 2*, pp. 33 35. London: Chapman and Hall.

Dorning KJ and Bell DG (1987) The Silurian carbonate shelf microflora: acritarch distribution in the Much Wenlock Limestone Formation. In: Hart MB (ed.)

Micropalaeontology of Carbonate Environments, pp. 266 287. Chichester: Ellis Horwood Ltd.
Dorning KJ and Harvey C (1999) Wenlock cyclicity, paly nology, and stratigraphy in the Buildwas, Coalbrookdale, and Much Wenlock Limestone formations, Shropshire, England. *Bolletino della Societa Paleontologica Italiana* 38: 155 166.
Downie C (1973) Observations on the nature of the acritarchs. *Palaeontology* 16: 239 259.
Downie C (1984) Acritarchs in British Stratigraphy. *Geological Society of London, Special Report* 17: 1 26.
Downie C, Evitt WR, and Sarjeant WAS (1963) Dinoflagel lates, hystrichospheres and the classification of the acri tarchs. *Stanford University Publications (Geological Sciences)* 7: 1 26.
Fensome RA, Williams GL, Barss MS, *et al.* (1990) Acri tarchs and fossil prasinophytes: an index to genera, species and infraspecific taxa. *AASP Contributions Series* 25: 1 771.
Jeppsson L, Aldridge RJ, and Dorning KJ (1995) Wenlock (Silurian) oceanic episodes and events. *Journal of the Geological Society, London* 152: 487 498.
Molyneux SG, Le Hérissé A, and Wicander R (1996) Palaeozoic phytoplankton. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications,* vol. 2, pp. 492 529. Salt Lake City, Utah: AASP Foundation.
Tappan H (1980) *Paleobiology of Plant Protists.* San Francisco: W. H. Freeman & Co.
Traverse A (1988) *Paleopalynology.* London: Unwin Hyman Ltd.

Chitinozoa

F Paris, University of Rennes 1, Rennes, France
J Verniers, University of Ghent, Ghent, Belgium

Introduction

Chitinozoans are strange and fascinating organic-walled microfossils. They appeared in the Early Ordovician (Tremadocian) and proliferated worldwide for 120 million years, until their extinction in the topmost Devonian (latest Famennian). The chitinozoans are exclusively recovered from marine deposits. They are usually abundant in sediments deposited under a low to moderate hydrodynamic regime. The biological significance of these microfossils and their systematic position have long been a matter of debate. The hypothesis that chitinozoans were eggs of soft-bodied metazoans is considered as the most likely.

Biological Aspects

Morphology and Structure

A chitinozoan vesicle corresponds to a small purse-, vase-, or flask-like structure formed by an organic wall, called a tegument or test. It is organized around a radial symmetry axis (Figure 1). The vesicle length ranges most frequently from 100 to a few hundred micrometres. Depending on the thickness of the wall, the weight of a single vesicle is between 10^{-3} and 10^{-5} mg. The chitinozoans are frequently linked along their longitudinal axis. In such chain-like structures, however, each vesicle does not communicate with its neighbours.

Basically, a chitinozoan comprises a chamber closed by a plug. The chamber, i.e., the bulging part of the vesicle, may be spherical, lenticular, ovoid, conical, cylindrical, bell-shaped, or claviform. A circular opening (aperture), frequently surrounded by a collarette, is located directly on the chamber or at the end of an apertural tube (neck) developing from the chamber. The plug is called an operculum (disc-like shape) when it seals the chamber directly (Figure 1B). It is named a prosome when it is situated within the neck, and corresponds to a cylinder segmented by horizontal septa (Figure 1A). The operculum and the prosome are extended anti-aperturally by the rica. This variously developed membranous expansion lines the upper part of the chamber. The vesicle wall and the plug act as a temporary protection for the inner contents.

At the junction between the flanks and the bottom of the chamber (i.e., the anti-apertural end) is the margin, whose shape and ornamentation (i.e., processes, spines, carina) are of primary importance in generic assignment. The ornaments of the vesicle display a large variety of size, structure, and shape. They may be either randomly or geometrically distributed (e.g., rows, rims, ridges) on part or on the whole vesicle. An anti-apertural mark occurs frequently at the apex of the chamber. This structure may be discrete (e.g., concentric rims, central scar, or even central pit) or more developed, forming the mucron (short membranous tube), the copula (long hollow tube), the siphon (bulb-like membranous expansion), or the peduncle (solid structure).

Investigations by transmission electron microscopy (TEM) and scanning electron microscopy (SEM) have demonstrated that the chitinozoan wall is composed

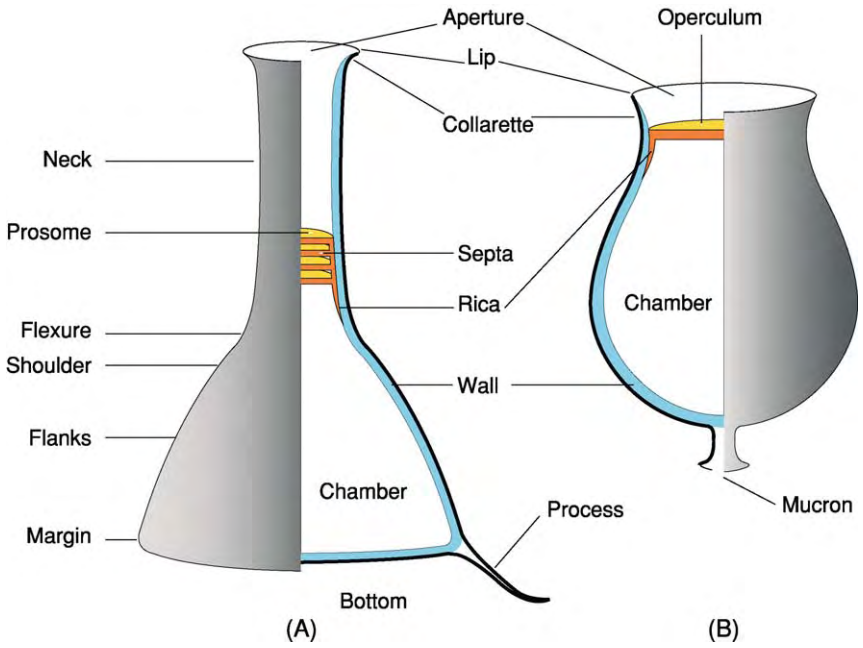


Figure 1 Diagram of a chitinozoan vesicle with morphological terminology: (A) Lagenochitinidae; (B) Desmochitinidae. Black, outer layer; blue, inner layer; orange and yellow, operculum and prosome. Adapted from Paris F (1981) Les Chitinozoaires dans le Paléozoıque du sud ouest de l'Europe (cadre géologique étude systématique biostratigraphie). *Mémoire de la Société Géologique et Minéralogique de Bretagne* 26: 1 496.

of an inner and an outer layer. The actual chemical composition and molecular structure of this wall, however, are still unknown. The inner layer constitutes the frame of the vesicle (Figure 1). TEM observations have revealed neither ultrastructures, nor tiny pores. The inner layer is usually thicker than the outer layer and extends even under the central pit. Moreover, it survives longer than the outer layer when exposed to various physical (abrasion) and biological (e.g., fungi and/or bacterial attacks) degradations. No structures are observed within the chamber or on the internal surface of the wall.

The outer layer gives rise to the ornamentation through its outgrowth (e.g., verrucae, cones, granules, spines), evagination (e.g., hollow spines or processes), or folding (e.g., carina, crests, wrinkles). Its ultrastructure is highly variable (e.g., spongy, microlamellar, or perforated with microcanals). Thin-walled taxa (e.g., Ancyrochitininae) have a very thin outer layer (less than 1 μm), whose existence is mainly documented from scars on the hollow spines or processes. In thick-walled taxa (e.g., Desmochitininae), the outer layer itself can be rather thick (up to 10 μm), but fragile and easily removed. The aspect of the external surface of the vesicle is controlled by the texture of the outer layer (e.g., smooth, felt-like, spongy). The shape, size, density, and location and organization of the various ornaments of the vesicle are used as taxonomic criteria.

Intervesicle Adjustments

Probably all the chitinozoans had linked vesicles, at least at one time in their development. The less resistant connections were, however, destroyed during fossilization and/or processing. The most common vesicle adjustments are linear or chain-like structures. Planar aggregates are also known.

In linear catenary structures, the linkage between two successive vesicles occurs along the longitudinal axis. The simplest linkage, but also the most fragile and therefore less commonly observed (e.g., *Angochitina*, *Cyathochitina*; vesicles are usually free in palynological residues), resulted from close contact of the flaring lip of one individual with the convex anti-apertural part of the preceding vesicle (Figure 2A). Such an apparatus, in all probability, required a covering with mucilage. When the mucron, or the copula, is also in full contact with the operculum of the succeeding vesicle, the linkage is much more robust and the operculum frequently remains stuck to the preceding vesicle (e.g., *Cingulochitina*, *Urnochitina*) (Figure 2B). A more sophisticated stage is reinforced linkage, where important modifications of the vesicle are observed. The apex of the chamber extends towards the anti-apertural pole and covers the operculum of the succeeding vesicle. This is well illustrated in the genus *Margachitina*, where the operculum is fully part of the preceding vesicle (Figure 2C).

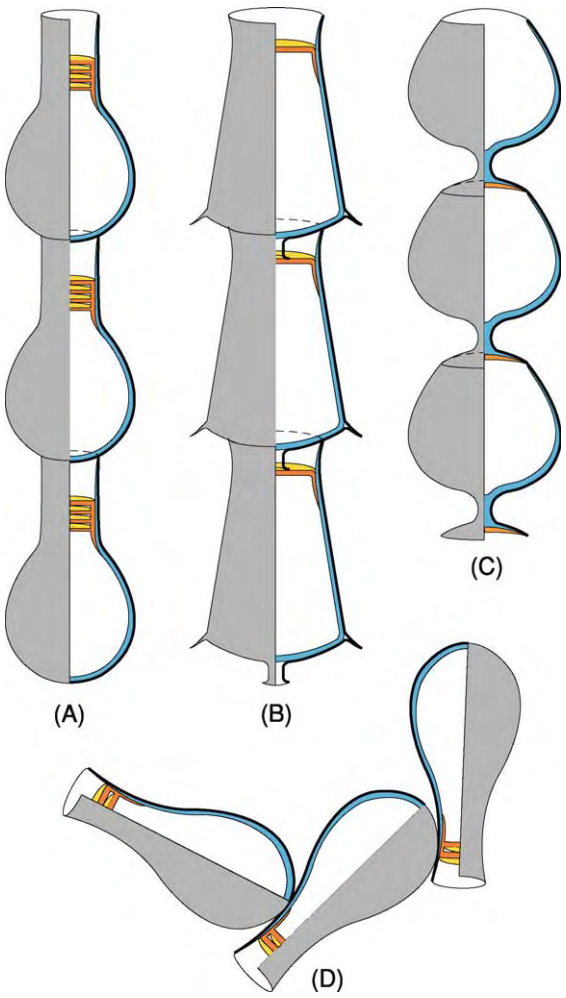


Figure 2 Schematic longitudinal cross section illustrating the main linkage types in chitinozoans: (A) simple adherence; (B) reinforced linkage; (C) double linkage; (D) helicoidal catenary structure. Black, outer layer; blue, inner layer; orange and yellow, operculum and prosome. Adapted from Paris F (1981) Les Chit inozoaires dans le Paléozoıque du sud ouest de l'Europe (cadre géologique étude systématique biostratigraphie). *Mémoire de la Société Géologique et Minéralogique de Bretagne* 26: 1 496.

Helicoidal catenary structures are less frequently observed. These chains appear normally as a coiled structure. However, due to compaction of the surrounding sediments, the original organization (several ten to several hundred linked vesicles) is frequently disturbed (folded or broken chains). Such structures have been reported in several genera (e.g., *Desmochitina*, *Bursachitina*, *Urnochitina*, *Margachitina*, *Pterochitina*). They are used for interpreting the actual biological meaning of the chitinozoans. Additional helicoidal catenary structures exist in which the linkage occurs by the adherence of an external part of the neck to part of the chamber of the following individual (e.g., *Lagenochitina navicula*). In this case, the aperture remains entirely free (Figure 2D).

Non-linear structures have been reported. This section regroups planar structures and aggregates. In planar structures, each vesicle is linked to a neighbour with an organic film. Geometrically arranged aggregates are known in *Desmochitina* species, in which vesicles are placed side-by-side, evoking an egg-laying, a cocoon, or the content of an egg capsule. In these structures, the aperture of every individual is completely free and its operculum is frequently absent. Chaotic aggregates (e.g., *Calpichitina*, *Cyathochitina)* usually correspond to individuals caught by organic matter. This arrangement, however, is not casual because it is always monospecific. When chaotic aggregates contain exclusively broken vesicles, they probably represent faecal pellets.

Teratological chitinozoans have been reported in chitinozoan assemblages, but they are extremely rare. They do not represent true intervesicle arrangements, but suggest aborted separations between two or several vesicles. For a time, they were interpreted as growing or rejuvenating stages. Biometric studies on large populations, however, support the existence of 'juvenile' chitinozoans.

Taxonomy and Classification

Progressively, a suprageneric classification, including orders, families, and subfamilies, and introducing a hierarchy amongst the morphological criteria available, has been adopted. Chitinozoans diversified rapidly throughout the Ordovician. Of the 56 usually accepted chitinozoan genera, 28% appeared in the Early Ordovician, 24% in the Middle Ordovician, and 11% in the Late Ordovician. Only seven genera became extinct at the end of the Middle Ordovician. The Ordovician biodiversification was accompanied by major morphological innovations (e.g., the appearance of the copula, carina, siphon, spiny ornaments, ornament in rows, processes). More than half of these innovations appeared before the Late Ordovician. During the Silurian, additional innovations appeared (e.g., the *Margachitina* linkage-type in the latest Llandovery; the reticulate chamber wall of *Pseudoclathrochitina*). Only a few changes occurred during the Devonian (e.g., the development of the solid and complex peduncle of *Urochitina* in the Lochkovian).

Suprageneric classification The order is the highest taxonomic subdivision in use for Chitinozoa. It is based upon the occurrence of an operculum in the Operculatifera and of the prosome (internal structure, not functional in the intervesicle linkage) in the Prosomatifera. The Operculatifera includes the unique Desmochitinidae family, whose representatives are

devoid of a neck but may have a collarette. Two families, i.e., the Lagenochitinidae (well-individualized neck) and Conochitinidae (without a shoulder, but with a deep prosome), are differentiated within the Prosomatifera. The third-order taxonomic subdivision is the subfamily, whose diagnostic criteria are regarded as reflecting evolutionary trends (e.g., differentiation of the outer layer). The Desmochitinidae include six subfamilies, the Conochitinidae seven, and the Lagenochitinidae six (Figure 3A–C).

Generic classification and specific criteria The generic assignments are based on the arrangement, location, and development of morphological features used as subfamily criteria. The outline of the vesicle, which is one of the most easily discernible elements by transmitted-light microscope, is also used. After emendations, restriction, and revision of the original diagnoses, only 56 genera from about 150 nominal ones are presently used.

Specific criteria for chitinozoans correspond to slight variations of the morphological parameters considered as generic criteria (e.g., length, density, location, shape of the ornamentation). From about 1120 species already described, only two-thirds seem to be workable taxa.

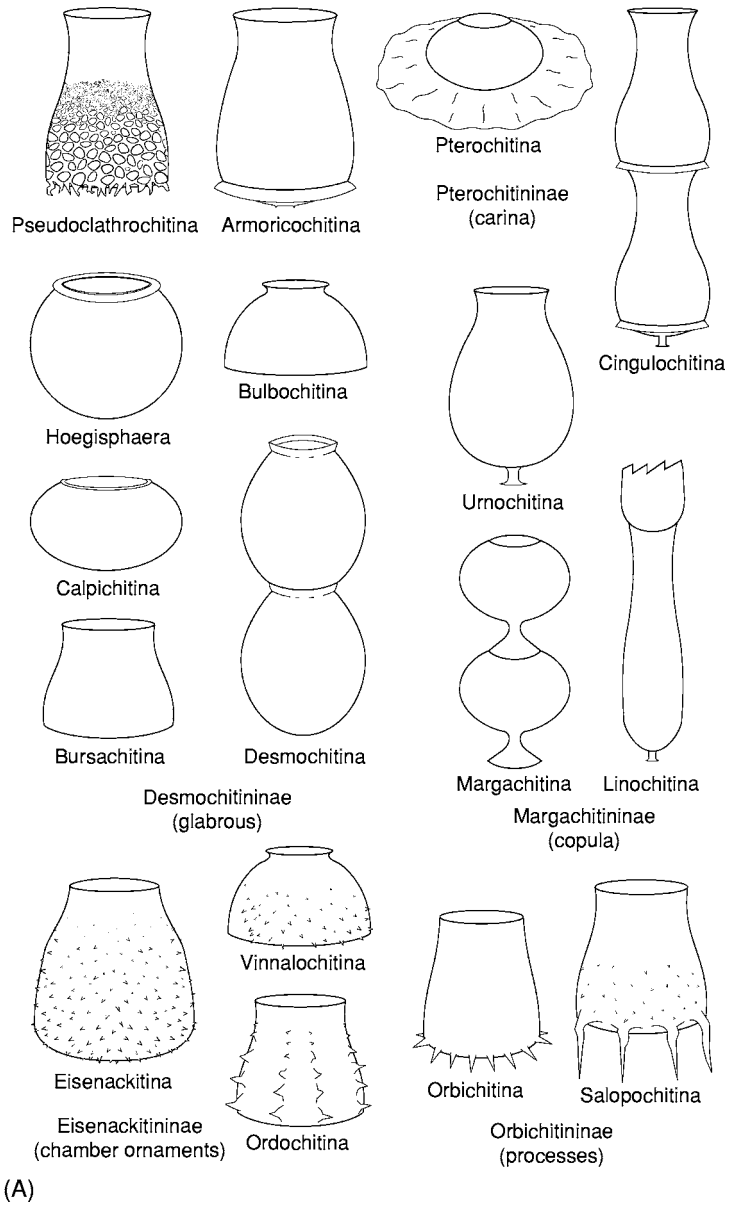


Figure 3 Outline of the most common chitinozoan genera: (A) Desmochitinidae; (B) Conochitinidae; (C) Lagenochitinidae.

Biological Interpretation and Systematic Position

The irritating question of the biological affinities of the chitinozoans has long been controversial. In general, successive proposals have rested on morphological features occurring in a small number of species belonging to the simplest shaped forms (i.e., *Desmochitina*, *Conochitina*, *Lagenochitina*). Chitinozoans have been successively assigned to protozoans (e.g., testaceans, rhizopods, tintinnids, flagellate protozoans), protists (e.g., dinoflagellates), and even to fungi ('Chitinomycetes'). However, all of these present-day groups show fundamental differences in structure, chemical composition, and environmental control from the chitinozoans. Moreover, a gap exceeding 100 million years separates the last chitinozoan occurrence from the first record of their possible, more modern analogues.

A more convincing hypothesis calls attention to the similarities between chitinozoans and eggs or egg capsules of diverse metazoans (e.g., annelids, gastropods (*see* **Fossil Invertebrates:** Gastropods), rotifers, and nematodes). The discovery of a complete coiled catenary structure preserved in three dimensions played the role of a 'Rosetta Stone' in providing the missing elements in this 'eggs theory'. This structure and the much more common fragments of chitinozoan chains are interpreted as an intraoviduct stage, and all of these catenary structures represent

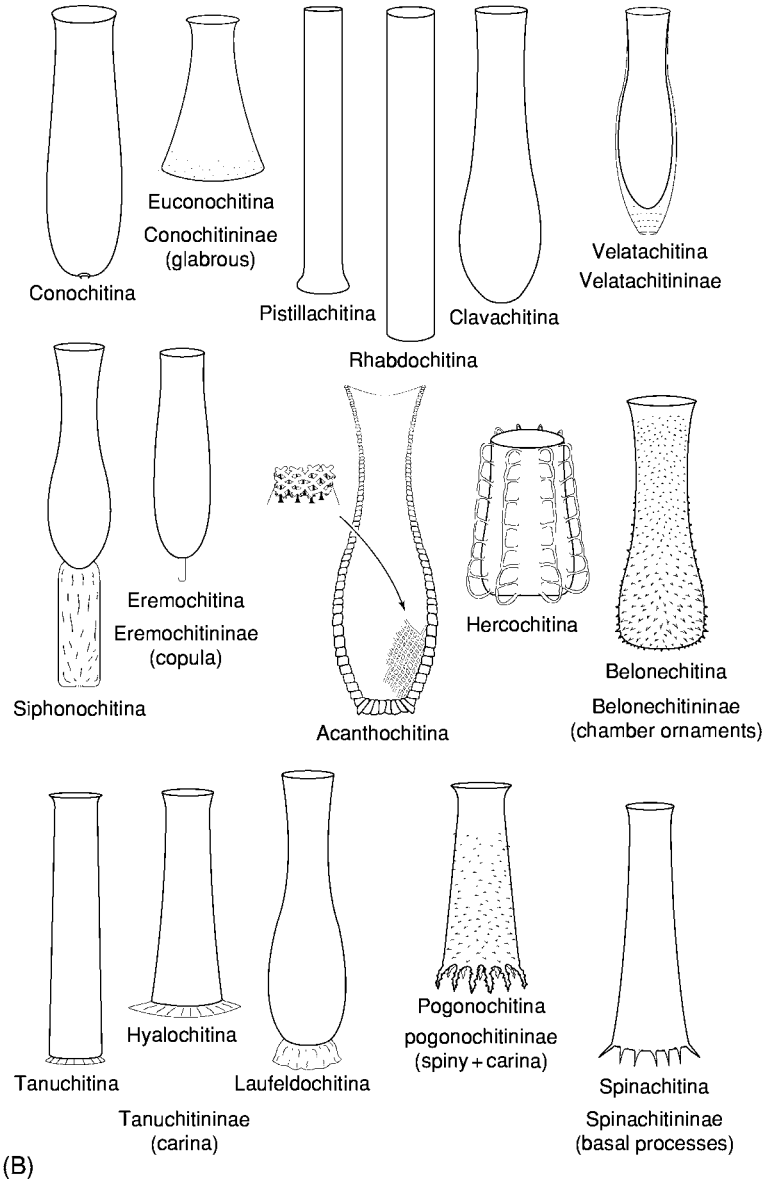


Figure 3 Continued

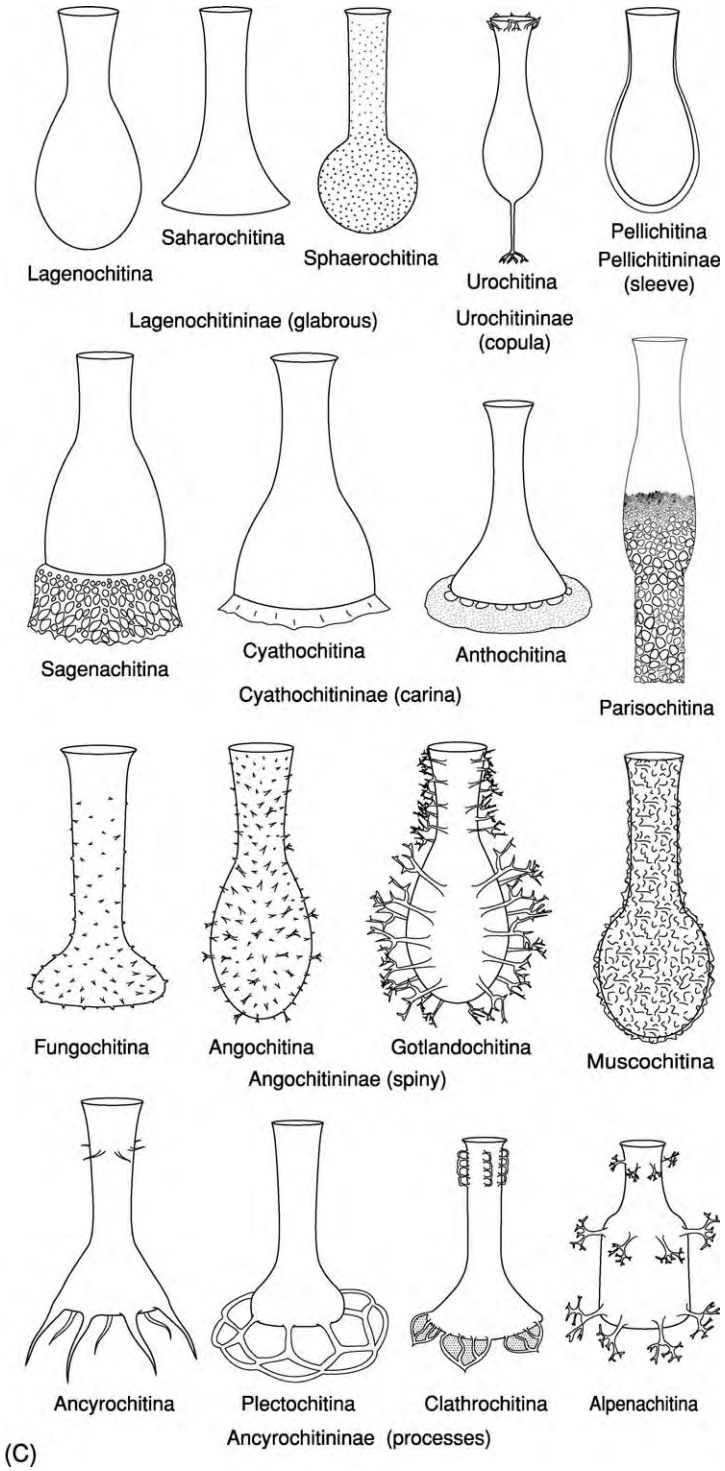


Figure 3 Continued

the release of immature eggs after the death and decay of the producing organism. The normal development of the chitinozoans included a maturation phase in the oviduct, prior to their laying in seawater, either as free isolated vesicles (e.g., *Pterochitina*, *Urnochitina*, Conochitinidae, Lagenochitinidae), as egg capsules, or as egg masses deposited through an ovipositor (e.g., *Desmochitina nodosa*). In all of these cases, each chitinozoan aperture was free. This is a prerequisite for the removal of the plug and the subsequent liberation of the chamber contents, i.e., a free larva, or a juvenile 'chitinozoan animal'.

With regard to the systematic affinities of the organisms producing the chitinozoans (i.e., the 'chitinozoan animals'), none of the body fossil groups usually recorded in Palaeozoic rocks display the same range as the chitinozoans. The tentaculitids, for instance, disappeared in the lowermost Famennian before the extinction of the chitinozoans in the topmost Famennian. The graptolites (*see* **Fossil Invertebrates:** Graptolites) appeared in the Cambrian, whereas the first chitinozoans are recorded in the Tremadocian. Moreover, the extinction of the monograptids in the Emsian had no consequences in the chitinozoan record. An extinct soft-bodied metazoan group is thus the most likely. According to the dimensions of the chitinozoans, these animals should be a few millimetres to 1 cm in length. These 'chitinozoans animals', with a worm-like soft body and a planktic or epiplanktic mode of life, were not preserved in normal conditions of fossilization. Some organic remains or discrete casts, however, probably survived fossilization in very quiet anoxic environments (e.g., 'konservat-Lagerstätten', black shale). Then, the 'chitinozoan animal' might be preserved as a carbonaceous film outlining a metazoan body with a flattened coiled chitinozoan chain inside. Attention should be paid, however, to avoid possible confusion with stomach and/or intestine contents, or with faecal pellets including chitinozoan remains.

Evolutionary Trends

If it is accepted that chitinozoan vesicles are the eggs of metazoans, their morphology is genetically controlled. This is critical for their biostratigraphical application. Amongst the examples of evolutionary trends, one of the most convincing is the progressive migration of the carina in the genus *Pterochitina*. This carina moved from an anti-apertural position in the Darriwilian (*P. retracta*) to a location below the equator of the chamber during the Llandovery (*P. deichaii*), then above this equator in the Pridoli (*P. perivelata*) and, finally, around the aperture in the Lochkovian (*P. megavelata*). Simultaneously, this carina developed significantly from the latest Ludlow to the mid-Lochkovian. Another obvious evolutionary trend is provided by a *Margachitina* lineage, from the latest Llandovery (*M. margaritana*) to the Pragian (*M. tenuipes*). Several morphological elements in this lineage present concomitant modifications affecting the chamber (from an ovoid to a lenticular shape), wall (development of thickened rims), peduncle (diameter progressively reduced), and operculum aperture adjustment (strengthening of the seal through a mortice-and-tenon assemblage system).

The widely held idea of a general shortening of the chitinozoan vesicles through time is not supported by the facts. The average smaller size of the vesicles recorded in Devonian samples is, in reality, due to an early extinction of the larger sized Conochitinidae in the Late Silurian (Figure 4). The Desmochitinidae and Lagenchitinidae themselves showed no general vesicle shortening from the Ordovician to the latest Devonian.

Chitinozoan Applications

Because of their rapid morphological changes through time (Figure 5), their records in a large variety of marine sedimentary rocks, their wide palaeogeographical distribution, and a fairly simple technical preparation, chitinozoans are amongst the most useful biostratigraphical fossil groups for Early Palaeozoic time. Chitinozoans, however, have a wider use in geology as they also provide information on palaeoenvironments, palaeoclimates, palaeobiogeography, and on the post-depositional history of the bearing rocks (e.g., diagenetic evolution, palaeothermometry, reworking events, and tectonic deformation).

Biostratigraphy

The biostratigraphical value of the chitinozoans has been progressively improved thanks to: (1) closely spaced sampling, with even bed-by-bed studies around the Global Stratotype Sections and Points (GSSPs); (2) better investigation techniques (e.g., sorting of the chitinozoans); (3) routine SEM observations; (4) more complete stratigraphical and geographical coverage of the sampled sequences and areas; (5) independent palaeontological controls provided by graptolites and/or conodonts; and (6) an improved taxonomy.

Global biozonations are now available for the Silurian and Devonian. For the Ordovician, due to palaeobiogeographical constraints, three regional biozonations are used.

The proposed chitinozoan biozones include interval range biozones and a few total range biozones. Most of the chitinozoan biozones benefit from direct chronostratigraphical calibration in the type sections of the GSSPs. Locally, and depending on the sedimentation, the chitinozoans allow high-resolution biostratigraphy, with a power of resolution approaching 0.5 million years for the most accurate biozones.

Ordovician regional biozones Because of the dispersion of the main palaeoplates and their highly contrasted latitudinal positions, three different biozonations have been defined for the Ordovician chitinozoans (Figure 6). The first represents the high-latitude assemblages from northern Gondwana regions. The second represents the assemblages from Baltica, which was located at moderate latitude

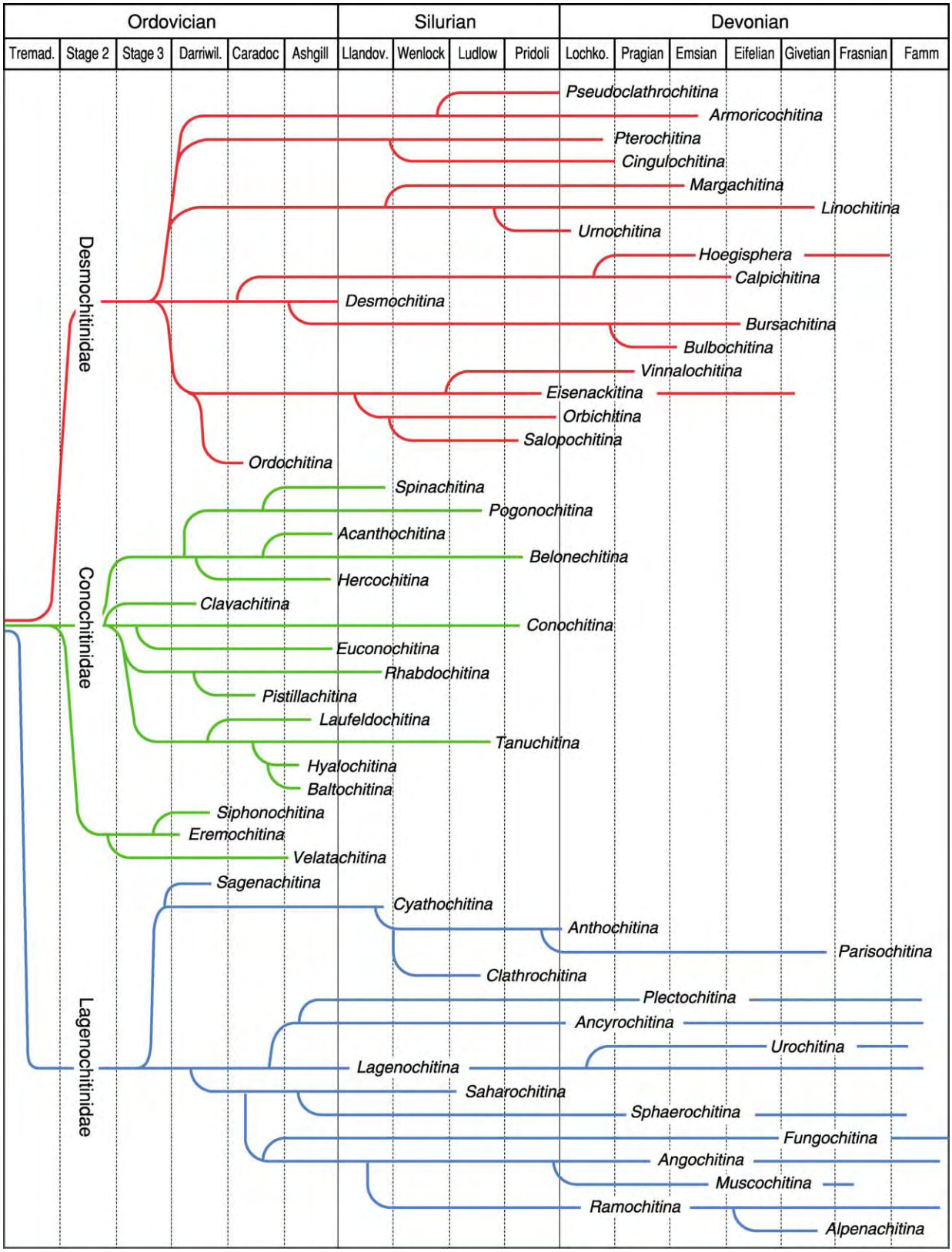


Figure 4 Morphological and temporal relationships between the chitinozoan genera.

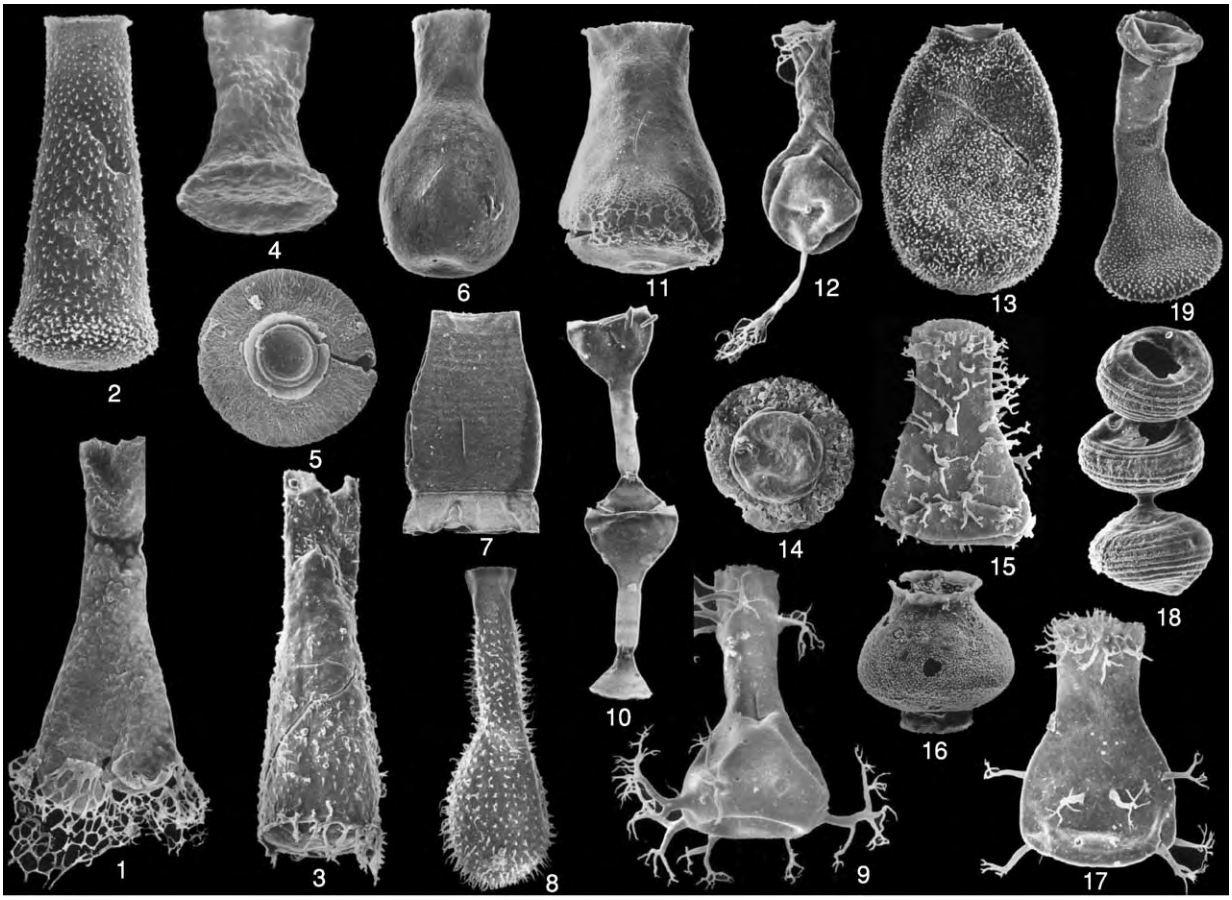


Figure 5 Scanning electron microscopy views of selected chitinozoan species. Ordovician species: 1, *Sagenachitina oblonga*; 2, *Belonechitina robusta*; 3, *Pogonochitina spinifera*; 4, *Euconochitina vulgaris*; 5, *Calpichitina lenticularis*; 6, *Lagenochitina baltica*; 7, *Armoricochitina nigerica*. Silurian species: 8, *Angochitina longicollis*; 9, *Ancyrochitina desmea*; 10, *Margachitina elegans*; 11, *Pseudoclathrochitina carmenchui*. Devonian species: 12, *Urochitina simplex*; 13, *Eisenackitina bohemica*; 14, *Pterochitina megavelata*; 15, *Ramochitina ramosi*; 16, *Bursachitina bursa*; 17, *Alpenachitina eisenacki*; 18, *Margachitina catenaria*; 19, *Fungochitina* sp.

during most of the Ordovician, and the third represents the low-latitude assemblages from Laurentia. A direct chronostratigraphical tie exists with the base of the Darriwilian in south China, and indirect calibration (through graptolites and conodonts) is available for most of the regional chitinozoan biozones.

The 25 Ordovician chitinozoan biozones presently identified in northern Gondwana are fairly regularly distributed, with five, nine, and 11 biozones in the Early, Middle and Late Ordovician, respectively. The average duration of one Ordovician chitinozoan biozone is about 2 million years. This duration is shorter (close to 1 million years) for the Middle Ordovician biozones. Eastern Avalonia, which left northern Gondwana in the Late Cambrian to earliest Ordovician, shared many chitinozoan species (including biozone index species) up to the Middle Ordovician.

A total of 22 chitinozoan biozones and subzones have been defined in Baltica, in spite of some hiatuses in the Ordovician chitinozoan record. The zonation is more accurate in the Late Ordovician, with an average duration of 1.2 million years per biozone. In contrast, only two chitinozoan biozones are so far available in the Early Ordovician of Baltica.

The Laurentian regional biozonation includes 19 biozones. Through successive improvements, the average duration of a Late Ordovician biozone in Laurentia is about 1.2 million years. The time resolution is much less precise in the Early and Middle Ordovician (average duration of about 3 million years per biozone).

Silurian global biozones Silurian chitinozoans are known from all the present-day continents, except Antarctica. The adopted Silurian global zonation includes 17 chitinozoans biozones (Figure 7). Most of these biozones benefit from direct chronostratigraphical calibration through the GSSPs of the Silurian Series in England and in the Czech Republic. Depending on the radiometric calibration accepted

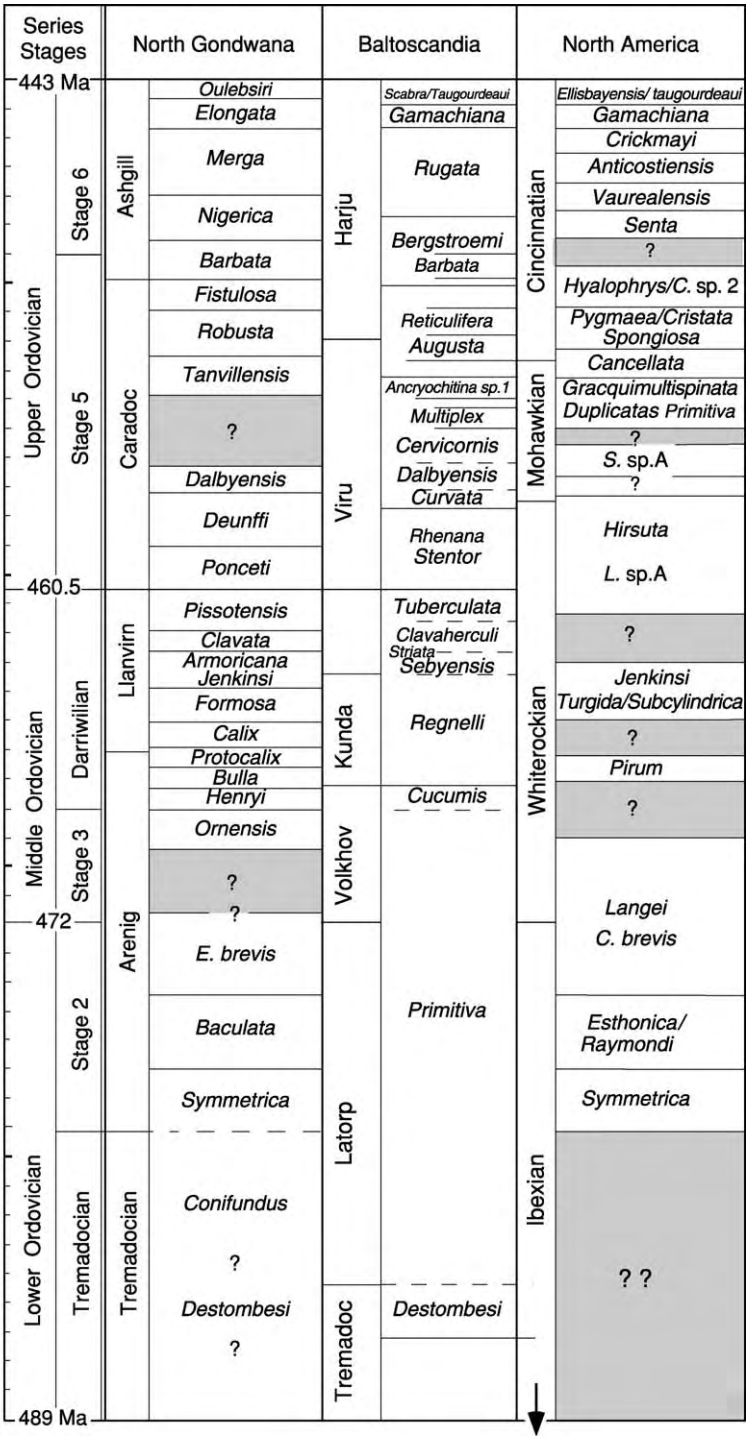


Figure 6 Ordovician chitinozoan biozones for northern Gondwana. Adapted from Webby, *et al.* (2004) Stratigraphic framework and time slices. In: Webby BD, *et al.* (eds.) *The Great Ordovician Biodiversification Event*, pp. 41 47. New York: Columbia University Press.

for the Silurian Series, the mean duration of a Silurian chitinozoan biozone is about 1.5–2 million years. A few time intervals are not yet characterized by a global chitinozoan biozone. In some cases, indirect age control of the biozone is available by means of graptolites. The present stage of the zonation will soon be improved as high-resolution biozonations become available at a regional scale (e.g., Llandovery and Wenlock of Estonia; Llandovery of Saudi Arabia).

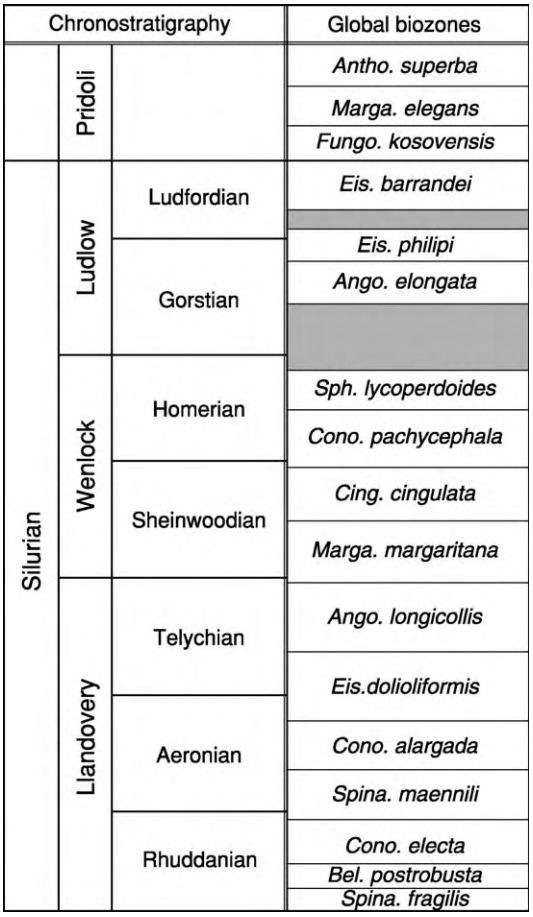


Figure 7 Silurian global chitinozoan biozones.

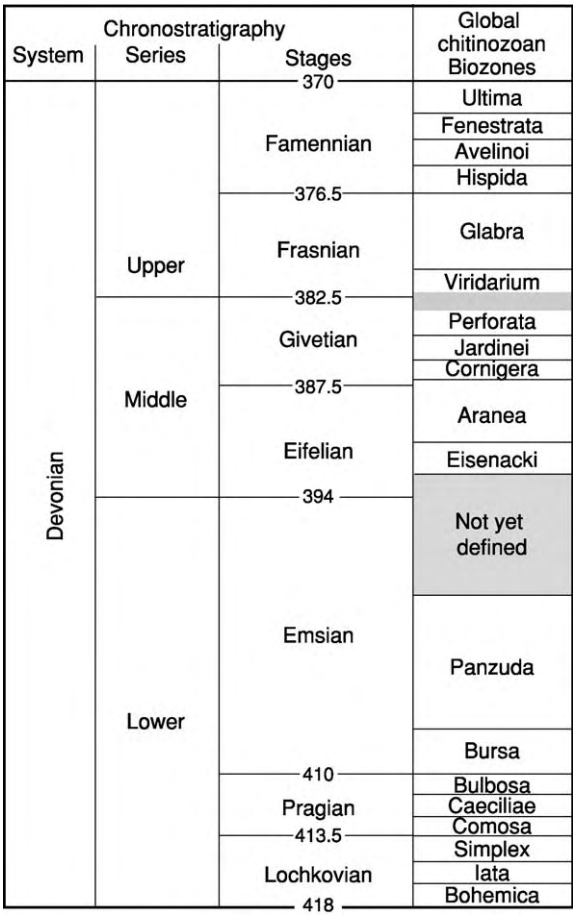


Figure 8 Devonian global chitinozoan biozones.

Devonian global biozones Most of the chitinozoan data available for the Devonian are from Gondwana (i.e., North Africa, southern and central Europe, South America, Australia) and from Laurussia. The present zonation, which benefits from a direct chronostratigraphical calibration with most of the GSSPs of the Devonian Stages (Czech Republic for the Lochkovian and the Pragian; southern France for the Frasnian and the Famennian), can be regarded as global. However, some endemic chitinozoan species have been recorded from Middle and Late Devonian strata in South America (Brazil and Bolivia). The Devonian chitinozoan zonation comprises 19 biozones, and two time intervals with no biozones yet defined (Figure 8). The average duration of one biozone is about 2.5 million years. Some high-resolution chitinozoan regional zonations are available (e.g., central Sahara, Brazil). They demonstrate that a more precise time slicing can be expected with the Devonian chitinozoans.

No undisputed chitinozoans have been recorded in the Carboniferous. The group apparently became extinct at the end of the Famennian, when *Fungochitina ultima* disappeared (Figure 4). However, this extinction may be an artefact of record as the 'chitinozoan animals' might have adopted a different mode of life (e.g., parasitism or non-marine). In both cases, dramatic changes might have occurred in the composition and structure of the wall of the chitinozoan vesicles, making fossilization no longer possible.

Palaeoenvironments

Chitinozoans disappeared at about 360 Ma. The assessment of the influence of environmental parameters on these microfossils therefore rests exclusively on indirect sedimentological and palaeoclimatological evidence, and on the ecology of associated organisms. Very little has been published on the environmental control of chitinozoans. Some authors have postulated a benthic life mode. However, many observations favour a planktic or epiplanktic mode of distribution for the chitinozoans as: (1) chitinozoans, which are exclusively marine microfossils, occur in

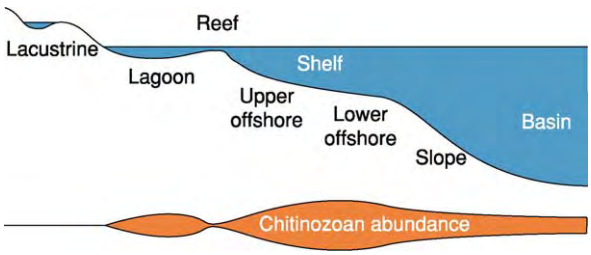


Figure 9 Diagram of the abundance of chitinozoans along a continent basin transect.

shallow-water deposits as well as in shelf and slope sediments; (2) chitinozoans are present and often very abundant (up to several thousand specimens per gram of rock) in marine anoxic or epoxic deposits devoid of benthic or endofaunas (e.g., Silurian black shales); (3) distal deposits (outer shelf and slope) yield abundant chitinozoans (several hundred to several thousand specimens per gram of rock; Figure 9), whereas acritarchs and/or miospores are rare or virtually lacking; (4) the geographical distribution of many chitinozoan species is frequently much wider than that of benthic or neritic organisms; and (5) some chitinozoan species extended throughout all the palaeoclimatic belts. Thus, the environmental control prevailing for pelagic organisms should also affect the chitinozoan vesicles, as well as the 'chitinozoan animals', if the 'egg theory' is accepted. This control is assessed through a variety of parameters, such as: (1) chitinozoan abundance (expressed as the number of specimens per gram of rock); (2) chitinozoan diversity (number of species and genera recorded per sample); (3) the relative frequency of the represented taxa; and (4) the chitinozoan ratio in palynological residues containing marine phytoplankton (acritarchs) and to land-derived elements (miospores, tracheid fragments). All of these numerical data are closely related to the production of the chitinozoans (nutrients, temperature), predation pressure, oxidation of organic matter, hydrodynamic constraints, and the fossilization processes.

Palaeobiogeography

Chitinozoans display a rather discrete provincialism compared with benthic Palaeozoic faunas. Most of the chitinozoan genera and many species have a worldwide distribution, comparable with the palaeogeographical distribution of graptolites or conodonts. However, the peculiar palaeogeographical context prevailing during the Ordovician (i.e., wide dispersion and contrasting latitudinal positions of the main palaeoplates; a very large ocean occupying most of the northern hemisphere) favoured a weak provincialism of the chitinozoans at a generic level, as illustrated by the restricted distributions of *Eremochitina*, *Velatachitina*, *Pogonochitina*, *Siphonochitina*, and *Sagenachitina*. These genera have been considered as typical northern Gondwana taxa but, probably due to the southern hemisphere gyre, some of these genera are now also reported in south China. *Hercochitina* was long believed to be an endemic genus from Laurentia, but recent investigations have revealed its occurrence in the Middle Ordovician of the northern Gondwana regions. As the record is being completed, the number of endemic genera amongst the Ordovician chitinozoans is decreasing. At species level, the geographical control is more effective. This has led to three different chitinozoan biozonations for low (Laurentia), medium (Baltica), and high latitudes (northern Gondwana), respectively. A few species only, usually corresponding to long-ranging and poorly discriminated taxa, display a more or less global distribution (e.g., *Cyathochitina campanulaeformis*, *Desmochitina minor*, *Belonechitina micracantha*, and *Lagenochitina baltica*). It is worth noting that the specific diversity of the Ordovician chitinozoan is slightly higher in the subequatorial assemblages (about ten species per sample) than in the high-latitude ones.

The Silurian and Devonian were times of convergence of the main palaeoplates. This is reflected by the distribution of the chitinozoan species. The weaker geographical control during these epochs allowed the appearance of chitinozoan biozones with global value. However, some regions displayed sporadic endemism at species level. The most obvious case concerns the assemblages recorded in the very large marine gulf which extended temporarily over westernmost Africa, Brazil, and Bolivia. These assemblages include a number of endemic species in the Early Silurian (e.g., *Pogonochitina djalmai*), Middle Devonian (e.g., *Ancyrochitina biconstricta*), and Late Devonian (e.g., *Sommerochitina langei*). These endemic species, however, coexisted with more widespread taxa, indicating communication with the global ocean.

Miscellaneous

Chitinozoans registered certain chemical parameters of the environment. The $\delta^{13}C$ values measured on carefully sorted chitinozoan vesicles show that the carbon of their wall is in equilibrium with the carbon reservoir of the ocean. Therefore, chitinozoans can be used for an accurate record of the $\delta^{13}C$ variations through time. At high latitudes, where carbonates are missing, they represent an alternative carbon source as they have an almost continuous record.

Like other organic-walled microfossils, chitinozoans registered the thermal events affecting the host sediment. A temperature increase led to a progressive darkening of their wall. This corresponds to a new molecular equilibrium (i.e., maturation of the organic matter by a progressive increase in the carbon ratio). Such changes can be measured through the reflectance of the vesicle wall, which provides a precise geothermometer for the early diagenetic to very low-grade metamorphic history of the host sediment.

Because the organic wall of the chitinozoan vesicles remained elastic for a long time, they were susceptible to collapse during compaction of the sediments, especially when the interstitial water escaped from argillaceous deposits. Vesicles preserved in full relief indicate an early lithification prior to compaction in, for example, cherts, limestones, phosphates, and ironstones. This phenomenon can be used to detect reworking in argillaceous deposits.

See Also

Biozones. **Fossil Invertebrates:** Graptolites; Gastropods. **Lagerstätten**. **Palaeoecology**. **Palaeozoic:** Ordovician; Silurian; Devonian.

Further Reading

Achab A (1988) Mise en évidence d'un provincialisme chez les chitinozoaires ordoviciens. *Canadian Journal of Earth Sciences* 25: 635 638.

Achab A (1989) Ordovician chitinozoan zonation of Québec and western Newfoundland. *Journal of Paleontology* 63: 14 24.

Combaz A, Calandra F, Jansonius J, *et al.* (1967) *Microfossiles Organiques du Paléozoïque. Les Chitinozoaires (2): Morphographie*. Paris: Centre National de la Recherche Scientifique.

Eisenack A (1972) Beitrage zur chitinozoen forschung. *Palaeontographica A* 140: 117 130.

Jansonius J (1970) Classification and stratigraphic application of Chitinozoa. In: *Proceedings of the North American Paleontologists' Convention,* 1969, vol. G, pp. 786 808. Lawrence.

Jenkins WAM (1970) Chitinozoa. *Geoscience and Man* 1: 1 21.

Kozlowski R (1963) Sur la nature des Chitinozoaires. *Acta Palaeontologica Polonica* 8: 425 449.

Laufeld S (1974) Silurian Chitinozoa from Gotland. *Fossils and Strata* 51: 1 130.

Miller MA (1996) Chitinozoa. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications,* vol. 1, pp. 307 336. Dallas, TX: American Association of Stratigraphic Palynologists' Foundation.

Paris F (1981) Les Chitinozoaires dans le Paléozoïque du sud ouest de l'Europe (cadre géologique étude systématique biostratigraphie). *Mémoire de la Société Géologique et Minéralogique de Bretagne* 26: 1 496.

Paris F (1990) The Ordovician chitinozoan biozones of the northern Gondwana Domain. *Review of Palaeobotany and Palynology* 66: 181 209.

Paris F (1996) Chitinozoan biostratigraphy and palaeoecology. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications,* vol. 2, pp. 531 552. Dallas, TX: American Association of Stratigraphic Palynologists' Foundation.

Paris F, Grahn Y, Nestor V, and Lakova I (1999) A revised chitinozoan classification. *Journal of Paleontology* 73: 547 568.

Paris F and Nõlvak J (1999) Biological interpretation and paleobiodiversity of a cryptic fossil group: the 'chitinozoan animal'. *Geobios* 32: 315 324.

Paris F, Winchester Seeto T, Boumendjel K, and Grahn Y (2000) Toward a global biozonation of Devonian chitinozoans. *Courier Forschung Institute Senckenberg* 220: 39 55.

Verniers J, Nestor V, Paris F, *et al.* (1995) A global Chitinozoa biozonation for the Silurian. *Geological Magazine* 132: 651 666.

Webby, *et al.* (2004) Stratigraphic framework and time slices. In: Webby BD, *et al.* (eds.) *The Great Ordovician Biodiversification Event,* pp. 41 47. New York: Columbia University Press.

Conodonts

R J Aldridge, University of Leicester, Leicester, UK

Introduction

Conodonts are a group of extinct marine animals that were entirely soft-bodied, except for an apparatus of tooth-like elements, composed of calcium phosphate and situated in the mouth and/or the pharynx. These elements normally became scattered in the sediment on the seafloor after the death and decomposition of the animals that bore them. The elements are usually microscopic, 0.2–2 mm in size, but rare larger specimens reach up to 25 mm in length. They display a range of morphologies, from coniform shapes through denticulated bars and blades to highly ornamented plates (Figure 1). The true conodonts have a stratigraphical range of Late Cambrian to Late Triassic and are important biostratigraphical indices for

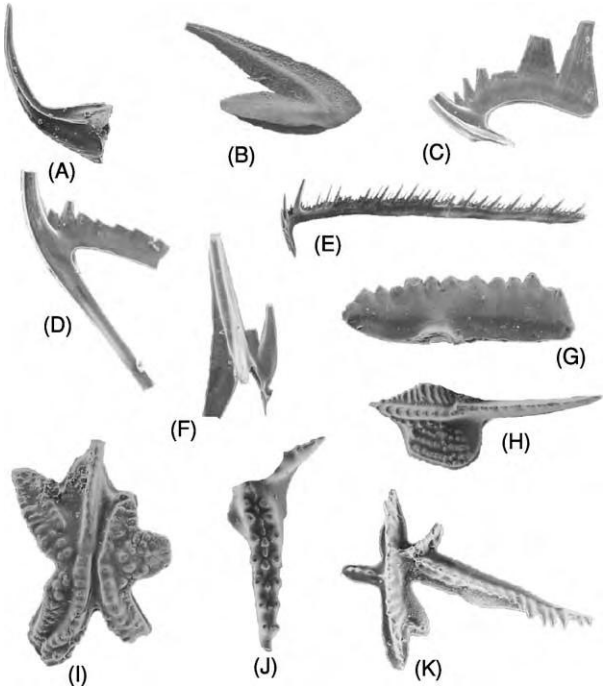


Figure 1 A vignette of conodont elements to illustrate the variety of morphologies. (A) Coniform element of the genus *Dapsilodus*, ×25, Silurian. (B) Coniform element of the genus *Drepanoistodus*, ×40, Ordovician. (C) Ramiform (alate S) element of *Gamachignathus*?, ×45, Silurian. (D) Ramiform (dola brate S) element of *Gamachignathus*?, ×45, Silurian. (E) Ramiform (bipennate S) element of *Idiognathodus*, ×30, Carboniferous. (F) Ramiform (tertiopedate S) element of *Gamachignathus*?, ×40, Silurian. (G) Pectiniform P element of *Ozarkodina*, ×20, Silurian. (H) Pectiniform P element of *Gnathodus*, ×25, Carboniferous. (I) Pectiniform P element of *Aulacognathus*, ×20, Silurian. (J) Pectiniform P element of *Icriodus*, ×25, Devonian. (K) Pectini form P element of *Eoplacognathus*, ×25, Ordovician.

marine strata of basal Ordovician to end-Triassic age. In addition, their colour is related to the temperatures that have been subsequently experienced by the rocks in which they are found and can be used in the investigation of the thermal history of sedimentary basins.

Fossils that preserve evidence of the soft tissues of conodonts are extremely rare. Research on these specimens, and on the internal microstructure of the phosphatic elements, has provided evidence that conodonts were chordates, although hypotheses regarding their precise relationships are the subject of continuing debate. The conodont apparatus itself was used to capture and process food, and most evidence suggests that the conodonts were active, macrophagous vertebrates.

Anatomy of the Animals

Conodont fossils preserving features of the soft tissues of the head and trunk are known from only two places in the world. The Upper Ordovician Soom Shale of Cape Province, South Africa, has yielded a single specimen that displays structures of the head and of the anterior portion of the trunk (Figure 2), as well as several tens of specimens that preserve paired head structures. The Lower Carboniferous Granton Shrimp Bed of Edinburgh, Scotland, has provided ten specimens, one of which is more or less complete from head to tail; the others all show a portion of the trunk, together with features of the head or of the tail (Figure 3).

The soft tissue features of the Scottish animals have been preserved through replacement by calcium

Figure 2 Conodont specimen with preserved soft tissues, genus *Promissum*, Upper Ordovician Soom Shale Member, Sandfontein, Cape Province, South Africa. Specimen C721b, Geological Survey of South Africa, Pretoria; scale bar, 10 mm. The conodont apparatus is at the lower right of the specimen with the position of the eyes preserved in white above and just to the right of the apparatus.

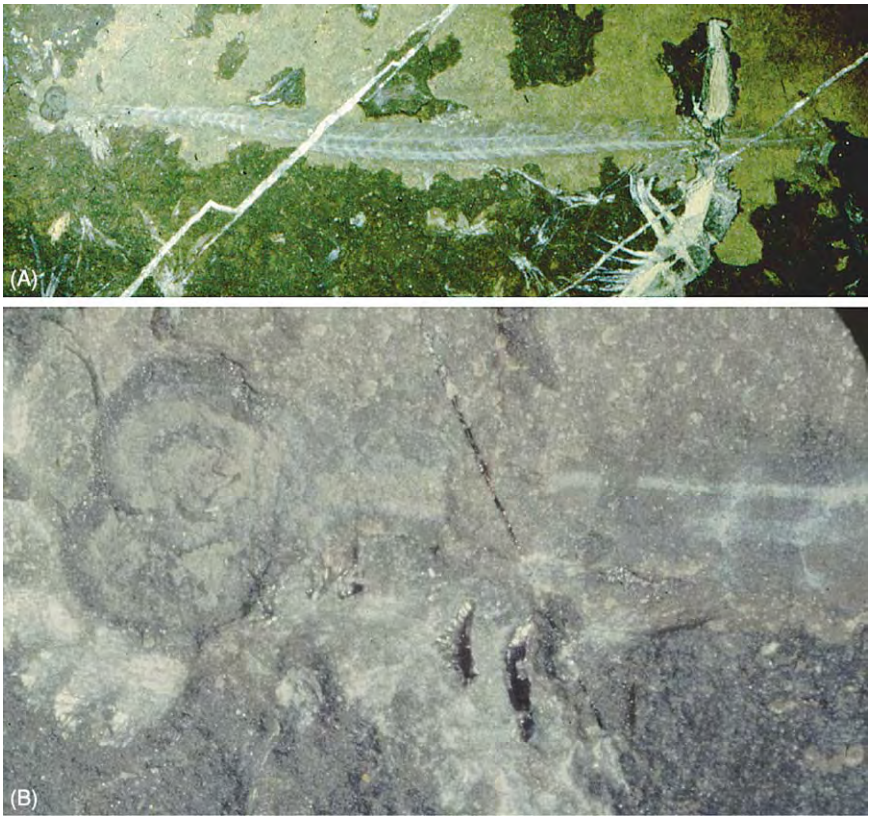


Figure 3 Conodont specimen with preserved soft tissues, genus *Clydagnathus*, Lower Carboniferous Granton Shrimp Bed, Edinburgh, Scotland. Specimen RMS GY 1992.41.1, Royal Museums of Scotland, Edinburgh, UK. (A) Complete specimen, anterior to left, ×2.3. (B) Anterior portion, showing the feeding apparatus (partly buried in the rock matrix), the eye capsules, and the anterior portion of the trunk, ×15.

phosphate, which forms a film on the bedding surface. The replacement was probably selective, and not all tissues and structures have been preserved. However, the animals can be seen to be elongate, 40–55 mm in length and less than 2 mm in height. They have a short head with a pair of ring-like structures that have been interpreted as representing the eyes. The skeletonized feeding apparatus, comprising 15 elements of various morphologies, is situated below and just behind the eyes; there is little evidence preserved of the soft tissue that housed the apparatus. The body traces show a distinct pair of subparallel longitudinal lines, which appear to represent a collapsed notochord, a precursor of the backbone. V-shaped muscle blocks (myomeres), with the apex anterior, are evident along the entire length of the trunk; it is not impossible that the myomeres were originally W-shaped, but that only the central V is preserved. Two of the specimens have an asymmetrical bilobed caudal fin with supporting fin rays. The fin rays are unbranched, set closely together, and there is no evidence for supporting musculature; the bilobed portion appears to occur on the dorsal margin. Two specimens show a possible dorsal nerve cord that stops short of the eyes, and the most complete individual shows faint traces of possible auditory capsules and gill pouches behind the eyes.

The specimens from South Africa are unusually large, and belong to a different order of conodonts, the Prioniodontida, from that represented by the fossils from Scotland (the Ozarkodinida). The most complete shows the anterior portion of an animal, with eyes, a feeding apparatus, and 100 mm of preserved trunk; the soft tissues are preserved through replacement by clay minerals. As in the Scottish specimens, the apparatus occupies a position ventral to and slightly posterior to the eyes; in this case, it comprises 19 individual elements up to 10 mm in length. The trunk shows a sequence of V-shaped myomeres, with the line of the notochord probably represented by a longitudinal gap in clay mineral replacement at the apices of the muscle chevrons. Under high magnification, high-fidelity preservation of original muscle fibres and fibrils can be seen. There is also a black organic patch within the mid-trunk region that may be the remains of a visceral organ. The body is nearly 20 mm in height; if the linear proportions are

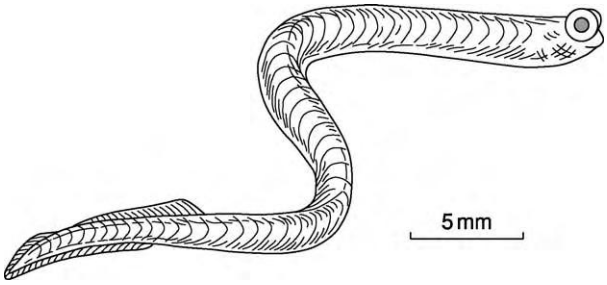


Figure 4 Reconstruction of a conodont animal, based on the specimens from Scotland and South Africa.

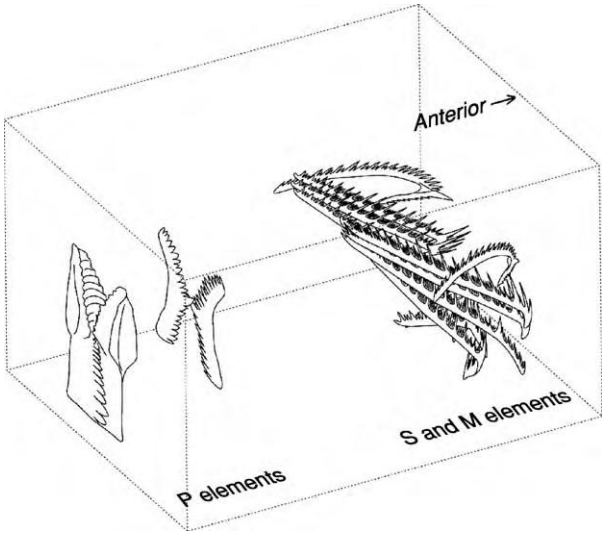


Figure 5 Architecture of an ozarkodinid conodont apparatus. Reproduced with permission from Aldridge RJ and Purnell MA (1996) The conodont controversies. *Trends in Ecology and Evolution* 11: 463–468.

comparable with those of the Scottish animals, then the entire length of this individual may have been as much as 400 mm.

Several tens of specimens from the Soom Shale preserve just the apparatus and the eyes. From these it can be determined that the eye capsules formed outwardly expanding cups, not much larger than those of the Scottish specimens. In a few specimens, including the most complete, the position of the eyes is occupied by fibrous tissue, which probably represents extrinsic eye musculature.

The evidence from these rare fossils permits a broad reconstruction of the morphology of a living conodont (Figure 4). A single, very poorly preserved, specimen with soft tissue is also known from the Silurian strata of Wisconsin, USA. This individual is from an order of conodonts, the Panderodontida, that possessed an apparatus of coniform elements, and the indistinct remains hint at a rather broader, flatter morphology for this type of animal.

Morphology and Internal Structure of the Elements

The more primitive conodonts possessed oro/pharyngeal skeletal apparatuses that consisted of an array of coniform elements. These elements vary within and between apparatuses in characteristics such as the degree of curvature, presence or absence of surface striations, and the development of costae. More derived conodont groups possessed elements of more complex morphology, with greater morphological distinction within each apparatus. Animals belonging to the most derived conodont order, the Ozarkodinida, bore a bilateral apparatus of 15 elements, which have been differentiated into two sets: a rostral array of nine denticulate, ramiform 'S' elements flanked by a pair of pick-shaped 'M' elements; behind these are two pairs of robust 'P' elements (Figure 5). The prioniodontid apparatuses from the Soom Shale differ in having four pairs of P elements that are situated above the S elements and behind the M elements, but the ozarkodinid 15-element pattern seems to apply to the vast majority of complex conodont apparatuses.

Conodont elements are composed of two structures: the crown and the basal body. The crown is heavily mineralized and relatively coarsely crystalline, and its upper surface may be elaborated into complex denticulated processes or ornate platforms. The basal body is finely crystalline with a higher content of organic material. The two components grew by the external apposition of layers of calcium phosphate. In many derived conodont elements, the basal body is either loosely attached or unmineralized and the crown has an open basal cavity. The tip of the basal cavity marks the origin of growth of the element.

In categorizing the morphology of individual conodont elements, the curvature of the cusp is taken conventionally as pointing towards the 'posterior', although this does not always correspond to the true biological orientation of the element. The cusp is defined as the denticle immediately above the tip of the basal cavity, and it is often the most prominent denticle of an element. Ramiform, or bar-like, elements are categorized by the disposition of processes (Figure 6), which radiate from the cusp and are commonly denticulate, but may be adenticulate: alate elements are bilaterally symmetrical, with a posterior process and a lateral process on each side; tertiopedate elements are asymmetrical with a long posterior process and a lateral process on each side; digyrate elements are asymmetrical, with a lateral process on each side, but normally with a very weakly developed

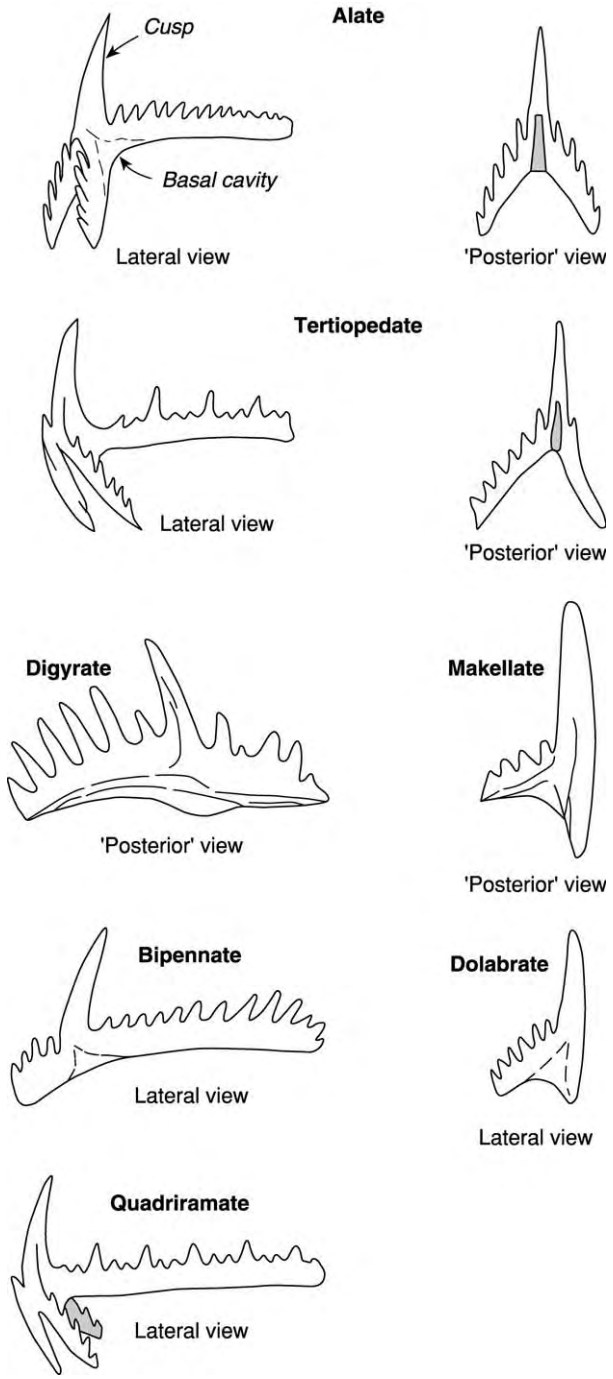


Figure 6 Morphological categories of ramiform conodont elements.

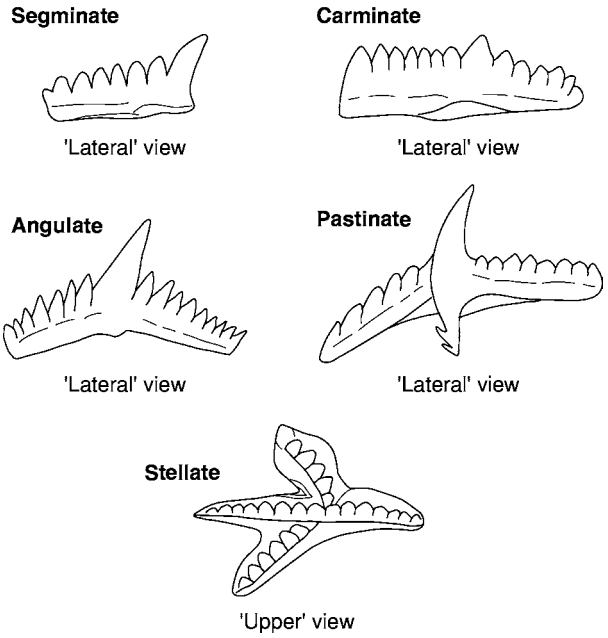


Figure 7 Morphological categories of pectiniform conodont elements.

posterior process; makellate elements have one lateral process and a downwardly directed anticusp on the other side; bipennate elements have anterior and posterior processes; dolabrate elements have only a posterior process, but there may be an extended anticusp below the cusp; quadriramate elements have four processes. Apart from the alate elements, each element is asymmetrical in itself, but left and right forms can be recognized that represent mirror-image pairs within the apparatus.

The P positions in complex apparatuses are often occupied by pectiniform elements, a term that encompasses blade-like morphologies and more complex platform or blade-plus-platform shapes. Pectiniform elements are accommodated in five major categories (Figure 7): segminate elements have only one process, conventionally termed anterior; carminate elements have anterior and posterior processes and are essentially straight in lateral view; angulate elements have anterior and posterior processes and are arched in lateral view; pastinate elements have three primary processes, anterior, posterior, and lateral; stellate elements have at least four primary processes.

The crown of conodont elements is composed of lamellar tissue, which may contain areas of opaque tissue known as white matter. The lamellar crown tissue is composed of crystallites that are typically a few micrometres in length, forming layers that are bounded by incremental growth lines. White matter is more finely crystalline, and is rendered opaque by the inclusion of small pores; it was deposited synchronously with the lamellar tissue. The basal body is formed of basal tissue, which normally shows incremental growth lines, but may be internally globular and/or tubular or may lack both globules and tubules. Most post-Devonian taxa lack biomineralized basal tissue.

Some authors have compared the lamellar crown tissue of conodont elements to the enamel of vertebrate teeth and scales, and the basal tissue to dentine. Other commentators have questioned whether these conodont and vertebrate tissues are truly homologous. White matter is certainly a type of tissue unique to conodonts.

Biological Affinity of Conodonts

The presence of chevron-shaped myomeres and a notochord indicates that conodonts belong to the chordates. Gill slits have not been recognized to date, but they would lie in the antero-ventral region where soft tissues have not been preserved in the fossils. Within the chordates, the presence of paired sensory organs (eyes), extrinsic eye musculature, a ray-supported caudal fin, a bilateral feeding apparatus, phosphatic skeletal elements, enamel, and dentine all suggest an assignment to the vertebrates, rather than to the more primitive cephalochordates, exemplified by the living lancelet *Branchiostoma*. The termination of the notochord anteriorly before the feeding apparatus is also a vertebrate feature. Some authors, however, prefer to place conodonts in the non-vertebrate chordates, pointing out the simplicity of the V-shaped myomeres and sometimes disputing the presence of eye musculature, enamel, and dentine.

Cladistic analyses, carried out using characteristics of conodont soft and hard tissues (Figure 8), have placed conodonts either as a sister group to the lampreys or, most recently, as more derived than both hagfishes and lampreys, the most primitive of living vertebrates. It has also been shown that this phylogenetic position does not rely on the interpretation of conodont hard tissues as homologues of enamel and dentine; if enamel, dentine, and extrinsic eye musculature are scored as absent, cladistic analysis still places the conodonts as more derived than the hagfishes. The weight of current evidence, therefore, indicates that the best hypothesis for conodont affinity is that they are crown group vertebrates (i.e., nested within the clade delimited by extant vertebrates). Their position as more derived than lampreys also means that they are stem group gnathostomes (jawed vertebrates), albeit more primitive than the array of extinct armoured agnathans (e.g., heterostracans, thelodonts, osteostracans) known from the Cambrian to Devonian.

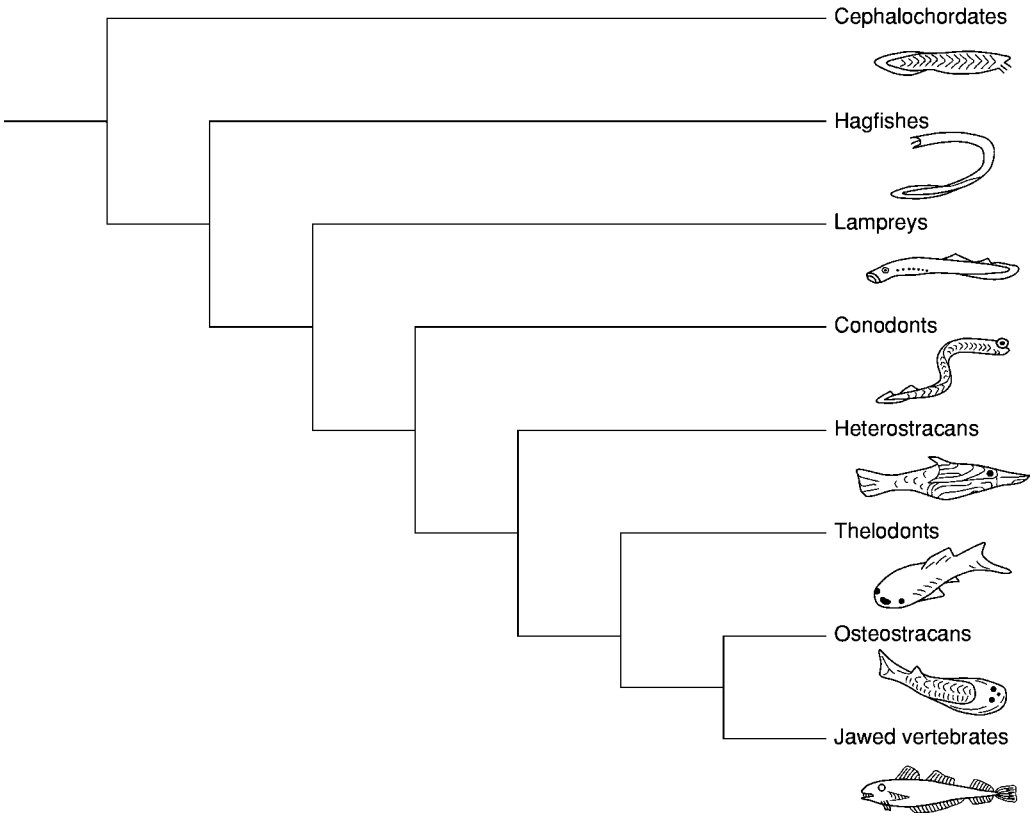


Figure 8 Cladogram of relationships between representative chordate groups, showing the most parsimonious position of the conodonts. Simplified from Donoghue PCJ, Forey PL, and Aldridge RJ (2000) Conodont affinity and chordate phylogeny. *Biological Reviews* 75: 191 251.

Classification

The classification of conodonts is in a phase of evolution. As the apparatus structures of various taxa and the homologies between elements become better understood, relationships are being increasingly studied using cladistic and other methods, but these have yet to be reflected in a firm classificatory scheme. The most widely used classification currently has conodonts assigned to seven orders, although it is realized that some of these are paraphyletic.

1. Order Proconodontida – conodonts with apparatuses mostly comprising coniform elements with deep basal cavities and smooth surfaces.
2. Order Belodellida – conodonts with apparatuses typically of coniform elements with thin walls, very deep cavities, and smooth surfaces; the elements commonly bear distinct costae, which may develop fine serration or denticulation.
3. Order Protopanderodontida – conodonts with apparatuses of longitudinally striated coniform elements that lack lateral furrows.
4. Order Panderodontida – conodonts with apparatuses of longitudinally striated elements with lateral furrows; the elements are coniform or bear denticles on the concave ('posterior') margin.
5. Order Prioniodontida – conodonts with P positions of the apparatus occupied by pastinate coniform or pastinate pectiniform elements; the more derived taxa have complex apparatuses, with strong differentiation between the morphologies of pectiniform P and ramiform M + S elements, and sometimes with the development of platform morphologies, especially in the P positions.
6. Order Prioniodinida – conodonts with apparatuses typically of ramiform elements, with those in the P positions characteristically digyrate; the denticles on all elements are commonly discrete and peg-like.
7. Order Ozarkodinida – conodonts with complex apparatuses with strong differentiation between the P and M + S elements; the P elements are pectiniform, consisting of a carminate pair and an angulate pair; platform development is common, especially on the carminate elements; the M and S elements are ramiform.

Architecture and Function of the Conodont Apparatus

Reconstruction of the architecture of conodont apparatuses relies on exceptionally preserved fossils, where evidence of the original geometrical relationship between the component elements is preserved. Such fossils include those with associated soft tissues, plus a larger number of, still very uncommon, undisturbed apparatuses preserved on bedding planes. When the conodont animal died and the soft tissues decayed, the elements were generally released to become scattered on the seabed. Scavengers and burrowers will have further dispersed them, and they were commonly transported by currents and waves before finally becoming incorporated into the sediment. In normal circumstances, therefore, it is impossible to recognize the various elements that came from an individual animal. However, in some deposits, in which decomposition of the conodont carcass took place in very quiet conditions, the skeletal apparatus has simply collapsed onto the bedding surface. These 'natural apparatuses' are normally found in shales or fine-grained limestones, and they preserve a two-dimensional representation of the original three-dimensional architecture. Skeletons that have collapsed onto bedding planes in different orientations can be used to reconstruct the original configuration, in much the same way that a building can be envisaged from drawings of the elevation and plan. Some additional information comes from clusters of elements that have become fused together diagenetically within the sediment. However, direct evidence of apparatus architecture is still only available for a small number of conodont genera. The best known apparatuses architecturally are those of a number of ozarkodinid taxa and that of the giant prioniodontid, *Promissum*, from the Soom Shale. There is also direct evidence of the three-dimensional structure of a few prioniodinid and panderodontid genera.

In the ozarkodinid apparatus (**Figure 5**), the pairs of P elements are orientated with their long axes perpendicular to that of the animal; the carminate pair is behind the scissor-like angulate pair. The rostral array of S elements is positioned with the long denticulate processes rising caudally in relation to the body axis and tilted towards the centre. The pair of M elements is above the S array, with the cusps directed downwards and inwards.

The apparatus is generally considered to be a feeding device, and functional interpretations have focused on two hypotheses. One model suggests that conodonts were microphagous suspension feeders, in which the rostral S and M elements supported a ciliated structure that captured food particles that were passed back to the tissue-covered P elements to be gently crushed and ingested. The alternative is that conodonts were macrophagous, with the S and M elements actively grasping food which was then sliced and crushed by the P elements. For this mode, the elements would have been exposed during function, and there is a need to reconcile this with the appositional accretion

of the lamellar tissue; this may be effected by periodic soft tissue cover.

The recognition of microwear features on a variety of conodont elements (Figure 9) provides strong evidence that they performed an active tooth-like function. This is supported by evidence from undisturbed ozarkodinid apparatuses in which the platform elements are constructed to occlude so precisely that there would have been no available space for an intervening cover of soft tissue. Internal discontinuities within the lamellar crown tissue can be interpreted as resulting from episodes of function alternating with episodes of growth, with worn surfaces being covered by newly secreted phosphate during the phases of dormancy.

There is less evidence for the function of conodont apparatuses consisting entirely of coniform elements. They could plausibly have been used to grasp prey, but direct evidence of surface wear is currently sparse.

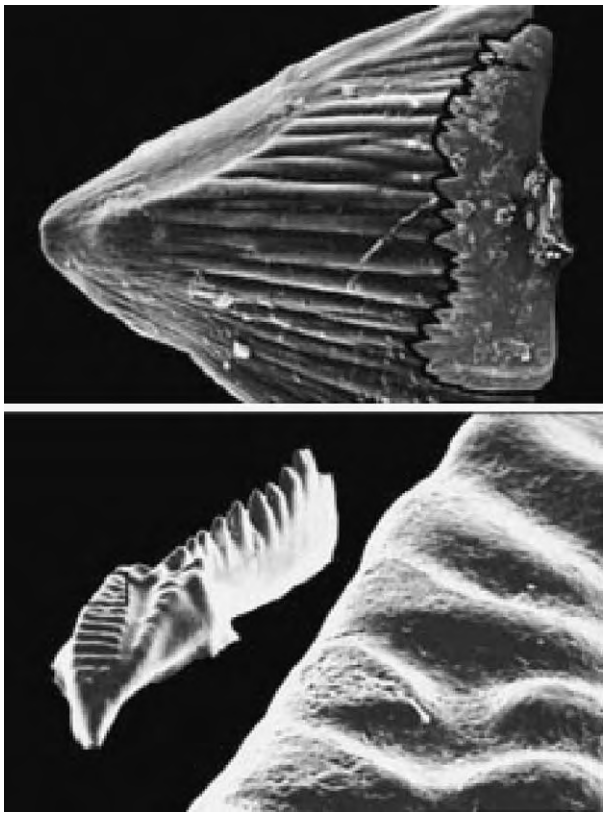


Figure 9 Microwear on the surfaces of conodont elements. Top photographs: P element of *Ozarkodina*; complete element to right, ×33; to left, close up of highlighted single denticle showing wear facet, ×1150. Lower photographs: P element of *Idiognathodus*, ×25; close up of pitted wear facets on oral surface, ×200. Re printed with permission from *Nature* (Purnell MA (1995) Microwear on conodont elements and macrophagy in the first vertebrates. *Nature* 374: 798 800). © 1995 Macmillan Magazines Limited.

If conodonts were the most primitive of stem group gnathostomes, as suggested above, the functional interpretation of their skeletal apparatus assumes major significance in the assessment of the origin of the vertebrate skeleton. More primitive chordates than the conodonts all lacked a biomineralized skeleton. The evidence that conodont elements fulfilled a tooth function, albeit without jaws, therefore implies that the vertebrate skeleton first arose, in the conodonts or their ancestors, as an oro/pharyngeal raptorial feeding system.

Many conodont species are globally widespread and broadly independent of facies in their distribution, suggesting that they were free-swimming pelagic animals. Other taxa show a more restricted distribution and a closer relationship to facies belts, and these were more probably nektobenthic or even benthic in habit.

Conodont Evolution and Biostratigraphy

The ancestry of the true conodonts ('euconodonts') discussed above appears to lie within a group of taxa represented by scattered phosphatic sclerites referred to the Paraconodonta. Paraconodonts lack an enamel crown, but are formed of a tissue that has been compared with dentine. No complete apparatuses of paraconodonts have been discovered, but fused clusters of small numbers of elements have been found that indicate that some taxa, at least, had multielement apparatuses with some morphological differentiation between the component elements. Paraconodonts range from the Middle Cambrian into the Ordovician.

The earliest euconodonts appeared in the Late Cambrian, where they are predominantly represented by coniform proconodontid elements. There was a major radiation of conodonts in the Ordovician, with the appearance and diversification of protopanderodontid and prioniodontid apparatuses in the earliest Ordovician, followed by belodellid, panderodontid, prioniodinid, and, finally, ozarkodinid taxa. No new orders appeared after the Ordovician, and the diversity of genera seen during this period is not matched again in later conodont history. However, conodonts are important biostratigraphical indices from the base of the Ordovician to their disappearance at the top of the Triassic.

The most important group biostratigraphically for Ordovician strata is the prioniodontids, which were highly diverse throughout the period and evolved rapidly. However, the prioniodontids were severely depleted by the end-Ordovician extinction event and were reduced again at the end of the Llandovery Epoch (Early Silurian); the last representatives of the order disappeared at the end of the Devonian. During

the Silurian, ozarkodinids flourished and there was a major diversification of this order in the Devonian. Conodont biozonal schemes from the Silurian to the Early Triassic rely heavily on ozarkodinid taxa. Prioniodinids remained conservative for much of their history, but diversified in the Triassic, with several lineages developing platform P elements; this order outlasted all the others and prioniodinids provide the basis for Middle and Upper Triassic conodont biozonation. In general, taxa with coniform elements had declined by the end of the Silurian, with only a very few coniform genera recorded above the Devonian. Conodont diversity was low in the Permian, but conodonts were relatively unaffected by the end-Permian mass extinction; however, they declined again in the latest Triassic and succumbed at the very end of that period.

Conodont Elements as Indices of Thermal Maturation

If conodont elements are heated, organic material within the lamellar crown tissue breaks down, and carbon fixing produces a continuous colour change from pale amber through brown to black; at even higher temperatures, elements become grey, then opaque white, and then clear. This colour alteration is dependent on temperature and time, but, where temperatures have been maintained for the order of millions of years, differences in the length of time make only a small difference to the colour achieved. A colour alteration index (CAI) has been established for such long-term heating, calibrated to temperature ranges, with CAI 1–5 covering the range from pale amber to black (<50–350°C) and CAI 6–8 covering metamorphic temperatures. This index has been widely applied in investigations of the thermal history of sedimentary basins containing strata of Cambrian to Triassic age, and provides information on thermal cut-offs in oil and gas exploration.

See Also

Biozones. **Evolution**. **Mesozoic:** Triassic. **Palaeozoic:** Ordovician. **Petroleum Geology:** Exploration. **Stratigraphical Principles**.

Further Reading

Aldridge RJ and Purnell MA (1996) The conodont contro versies. *Trends in Ecology and Evolution* 11: 463 468.

Donoghue PCJ, Forey PL, and Aldridge RJ (2000) Cono dont affinity and chordate phylogeny. *Biological Reviews* 75: 191 251.

Epstein AG, Epstein JB, and Harris LD (1977) Conodont color alteration an index to organic metamorphism. *United States Geological Survey, Professional Paper* 995: 1 27.

Purnell MA (1995) Microwear on conodont elements and macrophagy in the first vertebrates. *Nature* 374: 798 800.

Sweet WC (1988) *The Conodonta: Morphology, Tax onomy, Paleoecology and Evolutionary History of a Long Extinct Animal Phylum. Oxford Monographs on Geology and Geophysics 10*. New York and Oxford: Clarendon Press.

Foraminifera

M A Kaminski, University College London, London, UK

Introduction

The term 'Foraminifera' is derived from the Latin *foramen*, little hole and *ferre*, to carry or bear. They are a phylum of predominantly marine heterotrophic testate protozoans with tubular mitochondrial cristae and granuloreticular pseudopodia emanating from one or more openings in their tests. With over 50 000 known fossil and living species, the foraminifera constitute a diverse and geologically long-ranging group of organisms found in virtually all marine habitats. Morphologically, they form a heterogeneous and perhaps polyphyletic group with a fossil record that begins in the latest Precambrian, though forms with organic or unmineralised tests probably existed earlier. The group as a whole is characterised by the presence of an organic, agglutinated, or secreted biomineralised test partially enclosing the amoeboid body. The test may be single-chambered, pseudocolonial, pseudochambered, or multichambered with interconnected chambers added as the cell grows. Openings between the chambers [foramens] allow the cytoplasm to flow freely, while one or more apertures enable the cell to communicate with its external environment. A pseudopodial network may arise from a single apertural opening forming a distinct pseudopodial trunk or radiate in all directions from numerous openings or pores in the walls of some calcareous taxa. Pseudopodia exhibit bidirectional streaming, and are used for locomotion in free-living taxa, to anchor attached forms to the

substrate, to capture and ingest food items, or to build new chambers and growth or reproductive cysts. Reproduction is remarkably complex, and typically involves an alternation of generations between a multinucleate diploid [agamontic] asexually reproducing stage and a mononucleate haploid [gamontic] sexually reproducing stage. The latter gamontic individuals may also reproduce asexually by budding or fission, and sexually by the release of flagellated gametes. The two generations may produce morphologically distinct tests, with the gamont possessing a relatively large [megalospheric] first chamber or 'proloculum', and the agamont possessing a much smaller [microspheric] proloculum, but often a more complex and larger adult test. This dimorphism is most evident in the more advanced groups. Foraminifera are heterotrophic, and food items include other protozoans, algae, ciliates, small crustaceans, or other foraminifers. Some planktonic and larger benthic species inhabiting the photic zone harbour endosymbionts; some benthic forms burrow, ingesting sediment to crop bacteria, while others can make use of dissolved organic matter or phytodetritus.

The Rank of the Foraminifera

The discovery by Dujardin (1835) that Foraminifera are protozoans rather than cephalopods led d'Orbigny in 1852 to classify the group as a class with six orders based on chamber arrangement, and a seventh for single-chambered forms. Subsequent to d'Orbigny's original classification, later workers variously regarded the group to be of lower taxonomic rank. The widely accepted classification of Loeblich and Tappan regarded the group as a class with 12 major groups treated as orders. The class rank is maintained in the second edition of the *Illustrated Guide to the Protozoa*. However, over the last three decades, Protozoologists in both Russia and North America have assigned the group to a higher rank. Margulis first elevated the Foraminifera to the rank of a phylum, a rank that is maintained in her popular textbook *Five Kingdoms*. In his expanded classification of the Kingdom Protozoa, Cavalier-Smith first regarded the Foraminifera as a subphylum of the phylum Reticulosa (= Granuloreticulosa of earlier authors), but in his latest revision Cavalier-Smith (1998) quoted cytological evidence that removes the naked athalamids from that phylum. As a result, Cavalier-Smith removes the Granuloreticulosa/Reticulosa from his classification and elevates the foraminifera to the status of a phylum. In Russia, foraminiferal workers were quick to embrace the idea of a higher rank for the Foraminifera, with Mikhalevich regarding the group as a subphylum, later elevated to the status of phylum. This is the rank used in the monumental volume *Protista: Handbook on Zoology* published by the Russian Academy of Sciences (Alimov, 2000), which adopts the classification of Mikhalevich. The *International Working Group on Foraminiferal Systematics* [IWGFS] of the Grzybowski Foundation now accepts the phylum rank adopted by the protozoological community.

Classification of the Foraminifera

The two most widely accepted classifications of the Foraminifera are currently based on the morphology of the foraminiferal test. Important features include the composition and layering of the wall, the presence or absence of pores, the overall test shape, the mode of coiling, the shape of chambers and the presence of any internal structures such as pillars, the number and position of the apertures or openings, and any apertural modifications such as lips or teeth. Two schools of thought have recently emerged, both of which utilise morphological criteria with different weightings.

Since the mid-nineteenth century, wall ultrastructure has been regarded as a prime criterion for classification at a higher level, when Carpenter first subdivided the Foraminifera into two suborders (Perforata and Imperforata) based on the presence or absence of perforations in the test wall. In his classification, Carpenter also took into account the composition of the wall and remarked "The imperforate sub-order may be divided into three very natural groups, according as the nature of the envelope is membranous, porcellanous, or arenaceous; and thus we have the families Gromida, Miliolida, and Lituolida". In 1876, TR Jones raised the status of the 'arenaceous' and 'porcellanous' forms to that of groups with equal rank to the perforate forms. Jones' idea of grouping the agglutinated, imperforate porcellanous, and perforate 'hyaline' forms into separate higher-order categories was used in many later classifications. In their first comprehensive classification in 1964, Loeblich and Tappan used wall composition and microstructure as the defining character for the hierarchy of the foraminiferal groups, recognizing five suborders: (i) the organic-walled Allogromiina; (ii) the agglutinated-walled Textulariina; (iii) the calcareous microgranular Fusulinina; (iv) the porcelaneous Miliolina; and (v) the calcareous perforate Rotaliina. The most recent update by Loeblich and Tappan continued to use wall composition (including features such as mineralogy, layering, or nature of the cement in the case of agglutinated forms), to define 14 foraminiferal orders, which is here extended to 16 groups based on new research. If the class rank for the foraminifera adopted by North American workers is

Figure 1 Summary of compositionally based groups of the Foraminifera, (modified after Loeblich and Tappan (1992), Sen Gupta (1999) and Kaminski (2004).) The 'orders' of Loeblich and Tappan (1992) are here regarded as informal groupings pending revision of their status. The agglutinated Trochamminida and the Carterinida were included within the Lituolida by Kaminski (2004).

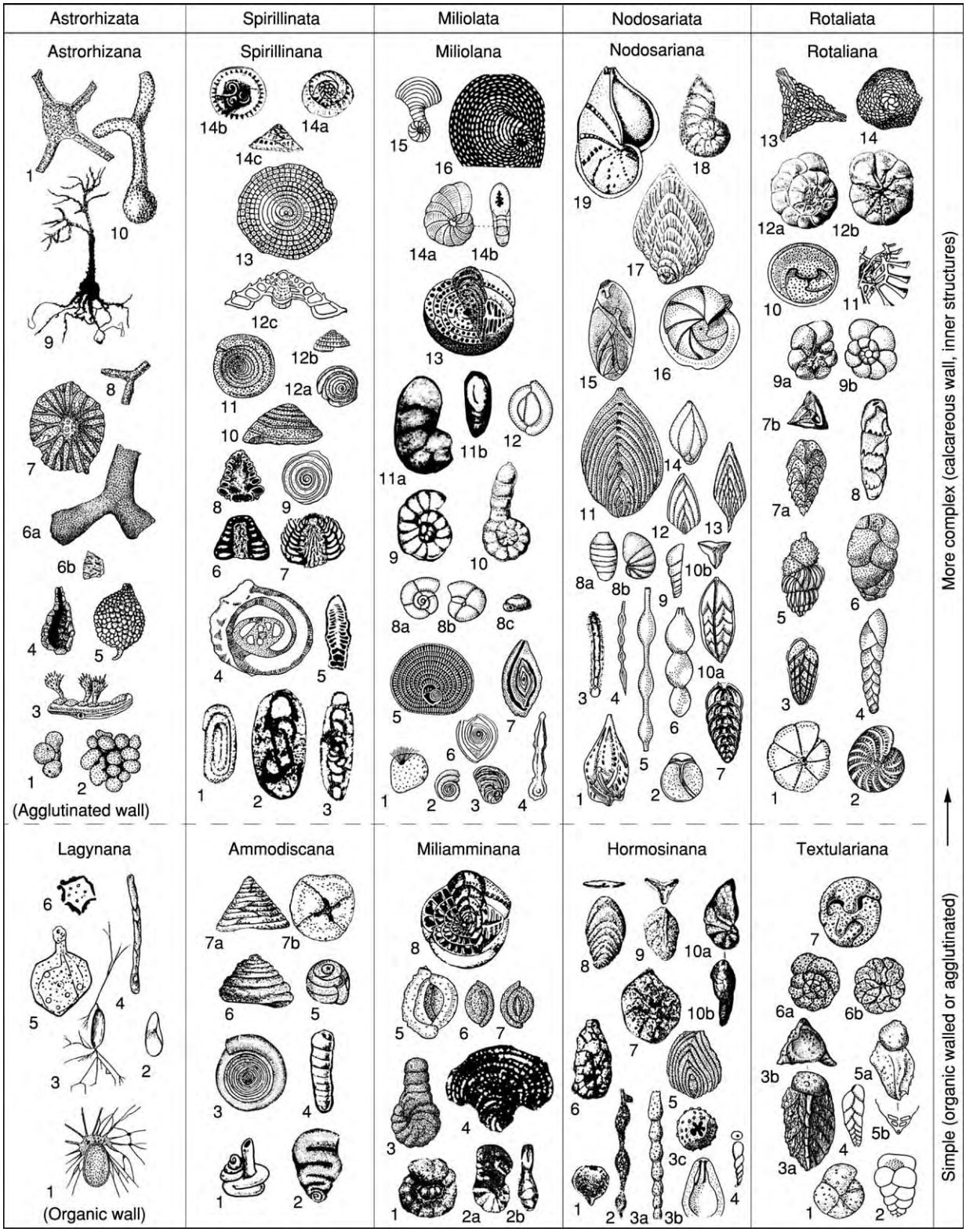


Figure 2 The main phyletic lines of benthic foraminifera (classes); with primitive and advanced subclasses and representative genera (modified after Mikhalevich (2004)).

Class Astrorhizata. Subclass Lagynana: 1. *Diplogromi*a, 2. *Lagynis*, 3. *Phainogulmia*, 4. *Xenothekella*, 5. *Heterogromia*, 6. *Microcometes*. Subclass Astrorhizana: 1. *Sorostomasphaera*, 2. *Thuramminopsis*, 3. *Halyphysema*, 4. *Lagenammina*, 5. *Astrammina*, 6a,b. *Schizammina*, 7. *Masonella*, 8. *Rhabdammina*, 9. *Notodendrodes*, 10. *Astrorhiza*, 11. *Saccorhiza*.

Continued

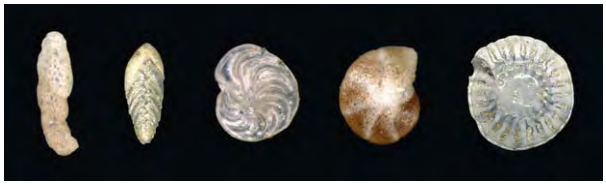


Figure 3 Fossil foraminifera, showing diversity of form of the chambered test (or shell). Left to right, they comprise the genera Ammotium, Bolivina, Planulina, Cyclammina, and Pseudorotalia, respectively. The first and fourth are agglutinating foraminifera, making their test from mineral grains 'glued' together with an organic cement. The other three are calcareous, secreted by the foraminifer itself. Magnifications between ×25 and ×150.

retained and the foraminiferal wall structure and composition is used at the highest taxonomic level, the eight main systematic groupings within the Foraminifera may be defined at the rank of a subclass. However, in view of the elevation of the Foraminifera to the status of a phylum, the rank of these groups will probably change in the near future. Pending a decision on their rank by the IWGFS, the main wall-structure groups are treated here as informal categories. The main groups are the allogromiids, agglutinates, fusulinids, miliolids, silicoloculinids, robertinids, and other minor perforate aragonitic groups, the spirillinids which display a unique crystal stucture, and the perforate calcitic groups consisting of rotaliids, buliminids, and globigerinids (Figure 1). This subdivision, based on wall structure, is not strictly a phylogenetic scheme, and even separates some groups based on habitat (i.e., planktonic and benthic groups) in spite of their identical wall structure.

Alongside the work of Loeblich and Tappan, a conceptually different version of the higher systematics of the foraminifera ('the Russian school') was pioneered by Fursenko and Rauser-Chernousova, and developed in detail by Mikhalevich and others. According to these authors, the starting point for the systematic hierarchy is the overall test morphology rather than wall structure. The general 'bauplan' of the foraminiferal test is regarded to be associated with the function and habitat of the whole organism. The mode of coiling and apertural characteristics of the test are considered to be the most conservative feature of the test, and the overall test shape is given priority in those cases where the test composition and/or microstructure has undergone modification within evolutionary lineages. Therefore, the composition and ultrastructure of the foraminiferal wall has subordinate significance compared with test morphology. Mikhalevich uses the following criteria to define the main foraminiferal classes: 1. the number of chambers, 2. the chamber shape, 3. pathways of chamber formation resulting in the nature of chamber conjunction, 4. the shape of the test and the predominant mode of coiling, 5. the shape and position of the main aperture, 6. the development of any inner apertural structures, 7. the development of integrative apertural systems, 8. the presence or absence of additional apertures, 9. the presence of additional skeletal plates, and 10. the presence or absence of canal systems. According to the Mikhalevich classification, the foraminifera are composed of a fewer number of larger groups of equal rank, and consist of the classes Astrorhizata, Spirillinata, Miliolata, Nodosariata, and Rotaliata (the latter also includes the planktonic Globigerinana as a subclass). The more advanced multichambered classes have both secreted calcareous and agglutinated subclasses (Figures 2 and 3). In each case, the agglutinated subclass (or in the case of the Astrorhizata, the organic-walled Lagynana) is regarded to be the ancestral group. Within each subclass, there are numerous examples of strikingly isomorphic pairs of

Class Spirillinata. Subclass Ammodiscana: 1. *Ammovertellina*, 2. *Ammovertella*, 3. *Ammodiscus*, 4. *Turritelella*, 5. *Repmanina*, 6. *Arenotur rispirillina,* 7a,b. Tetrataxis.

Subclass Spirillinana: 1. *Miliospirella*, 2. *Glomodiscus*, 3. *Archediscus dubitabilis*, 4. *A. karreri*, 5. *Cylindrotrocholina*, 6. *Howchinia*, 7. *Lasio trochus*, 8. *Babelispirrillina*, 9. *Coronipora*, 10. *Trocholina*, 11. *Spirillina*, 12. *Spirotrocholina* (a,b. views of the test, c. view of the canal in axial section), 13. *Annulopatellina*, 14. *Paleopatellina*.

Class Miliolata. Subclass Miliamminana: 1. *Recurvoides*, 2a,b. *Charentia*, 3. *Lituola*, 4. *Alzonella*, 5. *Dentostomina*, 6. *Sigmoilopsis*, 7. *Ammomassilina,* 8. *Reticulinella*. Subclass Miliolana: 1. *Squamulina*, 2. *Cornuspira*, 3. *Cornuspiroides*, 4. *Gheorgianina*, 5. *Discospirina*, 6. *Cornuloculina*, 7. *Spirophtalmidium*, 8. *Fisherinella*, 9. *Zoella*, 10. *Spirolina*, 11a,b. *Danubiella*, 12. *Quinqueloculina*, 13. *Neoalveolina*, 14a,b. *Dendritina*, 15. *Laevipeneroplis*, 16. *Parasorites*.

Class Nodosariata. Subclass Hormosinana: 1. *Saccammina*, 2. *Hormosinella*, 3a,b,c. *Nodosinum,* 4. *Adelungia*, 5. *Pseudopalmula*, 6. *Nouria*, 7. *Agardhella*, 8. *Flabellammina*, 9. *Triplasia,* 10a,b. *Ammomarginulina*. Subclass Nodosariana: 1. *Lagena*, 2. *Parafissurina*, 3. *Syzrania*, 4,5. *Grigelis*, 6. *Nodosaria*, 7. *Multiseptida,* 8a,b. *Lingulina* (a. microspheric, b. megalospheric forms), 9. *Marginulina*, 10a,b. *Tristix*, 11. *Kyphopixa*, 12. *Dyofrondicularia*, 13. *Flabellina*, 14. *Polymorphina*, 15. *Laryngosigma*, 16. *Lenticulina,* 17. *Planularia*, 18. *Hemicristellaria*, 19. S*aracenaria*.

Class Rotaliata (the third subclass, Globigerinana, is not shown). Subclass Textulariana: 1. *Haplophragmoides*, 2. *Minouxia*, 3a,b. *Gaudryina*, 4. *Pseudobolivina*, 5a,b. *Clavulina*, 6a,b. *Asterotrochammina,* 7. *Tiphotrocha*. Subclass Rotaliana: 1. *Bermudezinella*, 2. *Elphidium*, 3. *Brizalina*, 4. *Bolivinellina*, 5. *Euuvigerina*, 6. *Sporobulimina*, 7a,b. *Reussella*, 8. *Pseudobuliminella*, 9a,b. *Discorbis*, 10. *Neoconorbina*, 11. *Rotalia* (detail of canal system), 12a,b. *Ammonia*, 13. *Baculogypsinoides* (horizontal section of megalospheric individual showing canals), 14. *Eulinderina*.

agglutinated and calcareous genera. The Mikhalevich classification also makes fuller use of the Linnean hierarchy of classes, subclasses, orders, suborders, superfamilies, etc.

Further Reading

Cavalier Smith T (1998) A revised six kingdom system of life. *Biological Reviews* 73: 203 266.

Kaminski MA (2004) The year 2000 classification of the agglutinated foraminifera. In: Bubik M and Kaminski MA (eds.) *Proceedings of the Sixth International Workshop on Agglutinated Foraminifera. Grzybowski Foundation Special Publication.* 8: 237 255.

Lee JJ (1990) Phylum Granuloreticulosa (Foraminifera). In: Margulis L, Corliss JO, Melkonian M, and Chapman DJ (eds.) (1990) *Handbook of Protoctista.* Boston: Jones and Bartlett.

Lee JJ, Pawlowski J, Debenay JP, *et al.* (2000) Class For aminifera. In: Lee JJ, Leedale GF, and Bradbury P (eds.) *An Illustrated Guide to the Protozoa,* second edition, pp. 877 951. Society of Protozoologists. Lawrence Kansas: Allen Press.

Loeblich AR and Tappan H (1964) Part C. Protista 2. Chiefly 'Thecamoebians' and Foraminiferida. In: Moore RC (ed.) *Treatise on Invertebrate Paleontology,* p. 900. Lawrence Kansas: The Geological Society of America and the University of Kansas.

Loeblich AR and Tappan H (1987) *Foraminiferal Genera and their Classification.* New York: Van Nostrand Reinhold.

Loeblich AR and Tappan H (1992) Present status of For aminiferal Classification. In: Takayanagi Y and Saito T (eds.) *Studies in Benthic Foraminifera.* Tokyo: Tokai Uni versity Press.

Loeblich AR and Tappan H (1994) Foraminifera of the Sahul Shelf and Timor Sea. *Cushman Foundation for Foraminiferal Research Special Publication* 31: 661.

Margulis L (1974) Five kingdom classification and the origin of evolution in cells. *Evolutionary Biology* 7: 45 78.

Margulis L and Schwartz K (1988) *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth,* 2nd edn. New York: W.H. Freeman and Co.

Mikhalevich VI (1980) Sistematika i evolyutsiya foramini fer v svete novykh dannykh po ikh tsitologii i ul'trastruk ture. *Trudy Zoologicheskogo Instituta Akademii Nauk SSSR* 94: 42 61.

Mikhalevich VI (1998) Makrosystema Foraminifer. *Izves tiya Akademii Nauk, Seriya Biologicheskaya* 1998(2): 266 271.

Mikhalevich VI (2000) Typ Foraminifera d'Orbigny, 1826. In: Alimov AF (ed.) *Protisty: Rukovodstvo po Zoologii,* pt. 1, pp. 533 623. St. Petersburg: Nauka Publishers.

Mikhalevich VI (2004) On the heterogeneity of the former Textulariina (Foraminifera). In: Bubik M and Kaminski MA (eds.) *Proceedings of the Sixth International Work shop on Agglutinated Foraminifera,* 8, pp. 317 349. Grzybowski Foundation Special Publication.

Sen Gupta BK (1999) Systematics of modern Foraminifera. In: Sen Gupta BK (ed.) *Modern Foraminifera,* pp. 7 36. Dordrecht: Kluwer Academic Publishers.

Ostracoda

D J Horne, University of London, London, UK

Introduction

Ostracods, small crustacean arthropods (*see* **Fossil Invertebrates:** Arthropods) characterized by a bivalved carapace that can totally enclose the body and appendages, have an excellent fossil record by virtue of their small size and calcite shells (valves). Their bodies have reduced trunk segmentation and up to eight pairs of specialised limbs that are protruded from the gaping valves for locomotion, feeding, and reproductive activity. Adult ostracods are typically 0.5–2.0 mm long; some interstitial forms, however, are as small as 0.2 mm, some freshwater species grow up to 8.0 mm, and the pelagic marine myodocopan *Gigantocypris* reaches 32 mm. They are one of the most diverse crustacean groups (there are estimated to be more than 20 000 living species, of which only about 8000 have been described) and have a 500 million-year fossil record, from the Ordovician onwards, with more than 65 000 fossil species described. They are all essentially aquatic, inhabiting both marine and non-marine environments, although some taxa are adapted to a semi-terrestrial life. They have many applications in palaeoenvironmental analysis, palaeoclimatology and biostratigraphy.

Classification

The Class Ostracoda is divided into two subclasses, the Myodocopa and the Podocopa, each comprising three orders (Table 1). Of the Myodocopa only the Cladocopina and Thaumatocypridoidea (both Halocyprida) have strongly calcified valves and good fossil records, while those of most Myodocopida and Halocypridoidea (Halocyprida) are weakly calcified and hence less common as fossils. The three podocopan

Table 1 A summary classification of the Ostracoda

Class ***Ostracoda***
Subclass **Podocopa**
Order **Palaeocopida**
Suborder **Beyrichicopina**
Including superfamilies Beyrichioidea, Tetradelloidea, Eurychilinoidea, Aparchitoidea, Primitiopsoidea
Suborder **Binodicopina**
Including superfamilies Bollioidea, Aechminoidea, Drepanelloidea, Nodelloidea, Limbatuloidea
Suborder **Kirkbyocopina**
Including superfamilies Kirkbyoidea, Puncioidea
? Suborder **Eridostracina** (may not be ostracods)
Order **Platycopida**
Suborder **Platycopina**
Including superfamilies Leperditelloidea, Kloedenelloidea, Cytherelloidea
Suborder **Metacopina**
Including superfamilies Healdioidea, Thlipsuroidea
Order **Podocopida**
Suborder **Bairdiocopina**
Superfamily Bairdioidea
Suborder **Sigilliocopina**
Including superfamilies Sigillioidea, Bairdiocypridoidea
Suborder **Cypridocopina**
Including superfamilies Macrocypridoidea, Pontocypridoidea, Cypridoidea
Suborder **Darwinulocopina**
Including superfamilies Carbonitoidea, Darwinuloidea
Suborder **Cytherocopina**
Including superfamilies Quasillitoidea, Cytheroidea, Terrestricytheroidea
Subclass **Myodocopa**
Order **Myodocopida**
Including superfamilies Cypridinoidea, Cylindroleberidoidea, Sarsielloidea, Cyprelloidea, Bolbozooidea
Order **Halocyprida**
Suborder **Entomozocopina**
Superfamily Entomozooidea
Suborder **Halocypridina**
Including superfamilies Thaumatocypridoidea, Halocypridoidea
Suborder Cladocopina
Superfamily Cladocopoidea
? Order Leperditicopida (may not be ostracods)

Figure 1 Some living ostracods. A C: marine brackish cytheroidean podocopids; D: freshwater cypridoidean podocopids. Scale bars approx. 1 mm.

orders are the Platycopida, the ubiquitous Podocopida (the most diverse group of Ostracoda at the present day; Figure 1) and the Palaeocopida (diverse and widespread in the Palaeozoic, but now extremely rare). The Cambrian Bradoriida and Phosphatocopida, once included within the Ostracoda, are now considered to be separate bivalved arthropod groups. The ostracod affinities of some other groups, such as the Eridostracina and Leperditicopida, are currently being questioned.

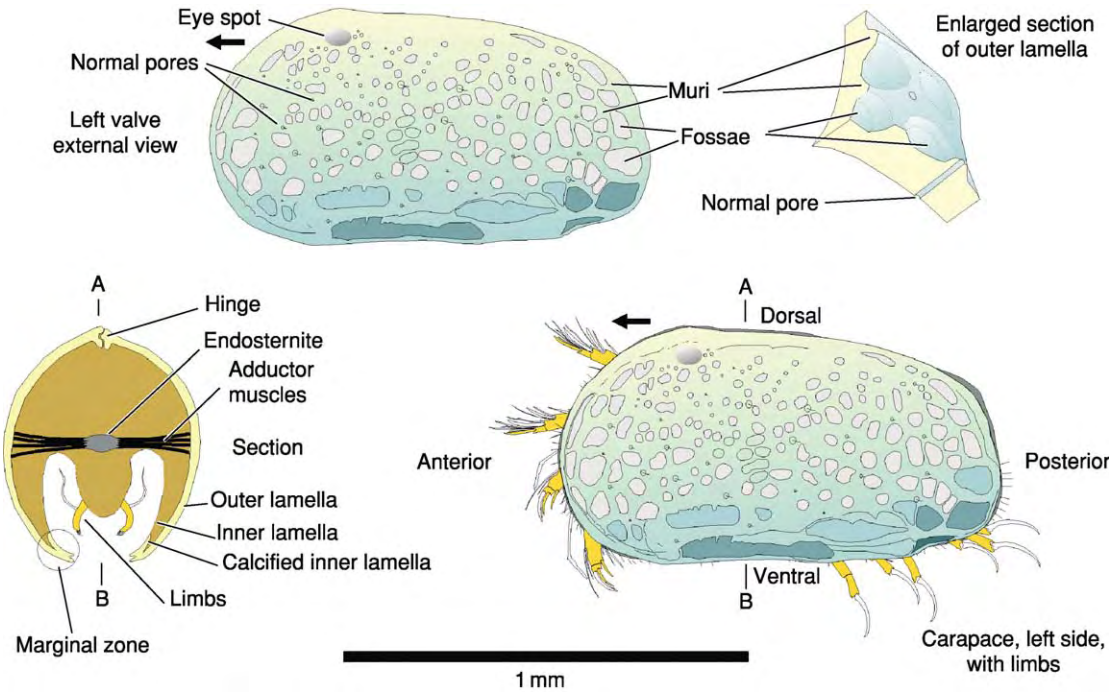


Figure 2 Morphology of a typical ostracod: external carapace morphology and cross section of a cytheroidean podocopid. Arrows indicate anterior direction in this and subsequent figures.

Morphology and Biology

The ostracod carapace consists of two valves formed by lateral folds (duplicatures) of the epidermis, completely enclosing the body and limbs (Figures 2 and 3). In podocopans the valves, comprising outer and inner lamellae, are mineralized with low-magnesium calcite; in fossil material, only the calcified inner lamella indicates the duplicature, since the uncalcified part is unlikely to be preserved. Many myodocopans are weakly or not at all calcified, and their ultrastructure differs from that of podocopans; furthermore, fossil valves of myodocopans are often secondarily calcified and calcareous nodules can form as postmortem artefacts in the valves of specimens caught alive; the probable equivalent of the podocopan calcified inner lamella is termed the infold. In the Podocopa one valve is usually larger and overlaps the smaller valve along part or all of its margin, while in the Myodocopa the valves are usually more symmetrical but may show posterior overlap. A few taxa show a narrow gape even when the valves are fully closed.

Although typically 'bean-shaped' or 'mussel-shaped', ostracod carapaces are extremely varied in shape and ornament (Figures 4–6). Sexual dimorphism is common; males may be larger or smaller than females and sometimes inflated posteriorly to accommodate the relatively large copulatory appendages, while in podocopid taxa with brood care the female is larger and more inflated. The valves are closed by means of adductor muscles attached to the inner surface of the calcified outer lamellae, where distinctive scars are formed that are a useful taxonomic character, especially at superfamily level; other muscle scars on the inner surfaces of valves are associated with various appendages. Dorsally the flexible cuticle connecting the two valves has been referred to as a ligament, but it probably plays no part in the opening of the valves, which is achieved by relaxation of the adductor muscles, hydrostatic pressure, and/or appendage movements. The valves of some taxa (notably cytheroidean podocopids) have a dorsal hinge structure of interlocking grooves and bars or teeth and sockets, also a useful taxonomic character. The outer lamella is pierced by normal pore canals which may terminate as simple pores through which sensilla protrude externally, sieve pores (also bearing sensilla) or exocrine pores (without sensilla; associated with moulting). The marginal zone, formed by the coincidence of the calcified inner lamella with the outer lamella, contains many taxonomically useful characters; the former may be fused throughout its width to the inner surface of the outer lamella, or the two lamellae may separate and diverge inwards, forming a space known as a vestibulum. Sensilla are often densely spaced along the free margins of the valves,

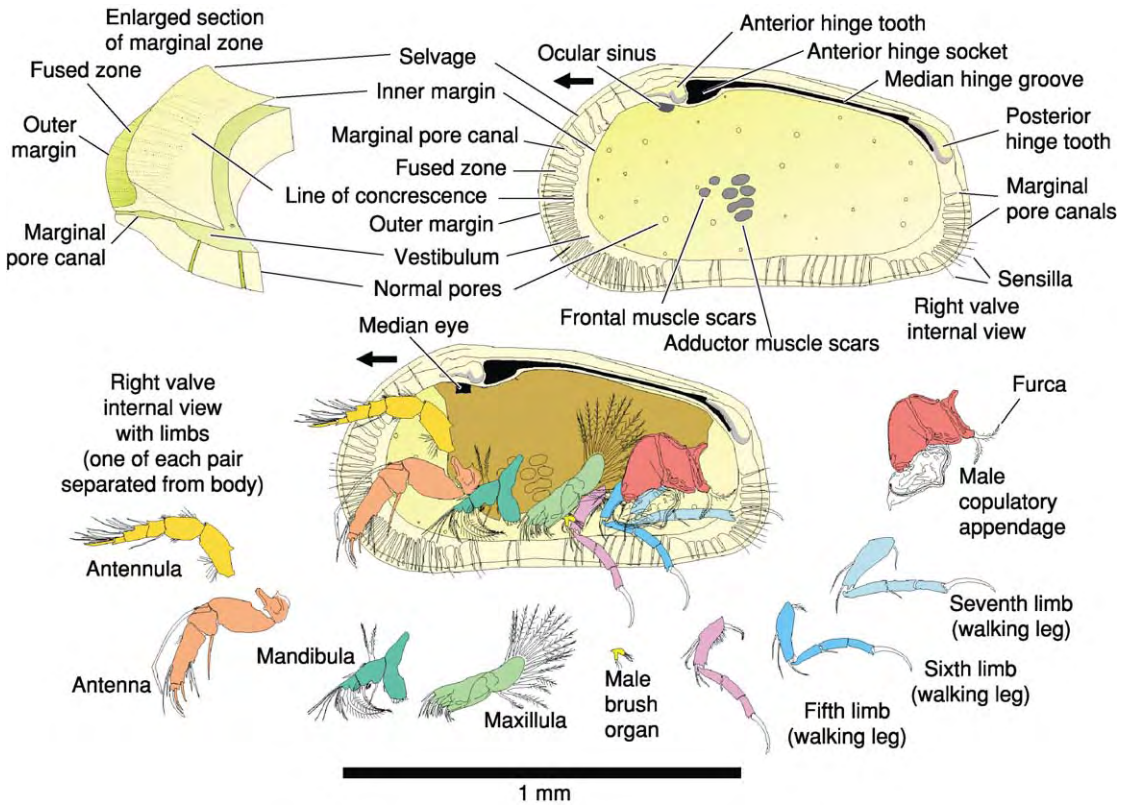


Figure 3 Morphology of a ostracod: internal valve morphology and limbs of a cytheroidean podocopid.

where the nerve-bearing canals leading to the pores pass through the fused zone, forming marginal pore canals. The calcified inner lamella may bear a ridge called a selvage, running sub-parallel to the outer margin; this forms an effective seal when the valves are closed, and there may be other, less prominent ridges called lists.

Many podocopan and myodocopan ostracods have a median or naupliar eye consisting of three optic cups which may be fused or separated, the left and right cups being inserted into ocular sinuses in the corresponding valves. Externally the calcite valves may show a clear eye-spot, a lens, or even a prominent eye tubercle. Most Myodocopida additionally have paired compound lateral eyes. The Halocyprida and some podocopan taxa are blind.

The paired limbs of adult ostracods are, from front to back, the antennules, antennae, mandibles, maxillulae, fifth limbs, sixth limbs, and seventh limbs; in some taxa (e.g., Cladocopina) posterior limbs may be absent. The eighth limbs are only present in the rare Puncioidea (the only living palaeocopids) but the copulatory appendages of other groups may have originated partly or wholly as modifications of these limbs. Additionally, a furca (a pair of caudal rami) is situated near the posterior termination of the body; in males of some taxa, there is also a pair of brush organs which may be vestigial limbs. The limbs are highly adapted to a variety of special functions, including locomotion, feeding and reproduction, and many of them bear sensory setae. The four or five head limbs are attached to the cephalon, comprising forehead, upper and lower lips, and hypostome. Most ostracods show little trace of post-cephalic segmentation other than the presence of paired appendages, but the evidence in some taxa suggests 10 or 11 trunk (thorax plus abdomen) segments (plus a posterior telson) in the Podocopa and 4 to 7 in the Myodocopa. Ostracod limbs are only fossilised under exceptional circumstances, but such occurrences are extremely valuable for resolving problems of classification and phylogeny.

Like all arthropods, ostracods grow by moulting. Podocopid ontogeny usually consists of 9 instars (moult stages): 8 juvenile and 1 adult, designated A, A-1, A-2, etc., in descending order of size. The first juvenile instar is already enclosed by two (non- or very weakly calcified) valves and has 3 pairs of limbs: antennules, antennae, and rudimentary mandibles. The genus *Manawa* (Palaeocopida: Puncioidea) is unusual in that the earliest instars have a single shield-like carapace, and only the later instars are bivalved.

Limbs are added progressively through ontogeny, accompanied by changes in carapace shape. In many taxa, the features of the marginal zone and hinge are only fully developed in adults, while in juvenile valves they are weakly developed or absent. The ornament of juveniles is usually a subdued version of that of corresponding adults, but some taxa show a marked change in ornament at the final moult. Myodocopan ontogeny consists of 4 to 7 juvenile instars and a single adult instar, the first juveniles having 5 or 6 limbs (or limb rudiments).

Most marine ostracods (myodocopans and podocopans) reproduce sexually (a few may be parthenogenetic). Males of some myodocopids produce bioluminescent courtship displays. Three reproductive modes are recognized in non-marine ostracods: fully sexual, exclusively parthenogenetic (e.g., darwinuloideans, termed 'ancient asexuals' since they are believed to have reproduced without sex for 200 million years) and mixed, for example cypridoideans with widespread parthenogenetic populations but geographically restricted sexual populations. Males of some Cypridoidea have the longest spermatozoa, relative to body size, of any animal (up to ten times the length of the adult carapace), which are kept coiled inside the duplicature and during copulation are passed through a muscular pump called a Zenker's Organ.

Most Podocopida deposit their eggs singly or in clusters, but brood care of the eggs and early instars in the posterior brood space of the adult females is known in several cytheroidean families and all darwinuloideans. Female platycopids brood eggs (usually only 4–8) but not juvenile instars. All myodocopids have brood care, but it is rare in halocyprids, most of which release eggs directly into the sea. In the Palaeozoic, palaeocopid beyrichicopine females had distinctive anteroventral brood pouches known as cruminae (Figure 9), in which fossilized juveniles have sometimes been found.

In Podocopa, life cycles vary from a few months to as long as 4 years. Shallow marine cytheroideans typically have a single generation per year, development taking place mainly in spring and summer, with delayed development of eggs or instars during the winter months; some species, however, manage 4 or 5 generations during the warmer part of the year. Many non-marine cypridoideans, especially those living in temporary ponds, have short life cycles of only a few weeks, but their desiccation-resistant eggs can remain viable for years or even decades. Total life span in Myodocopa can be as little as 1 month, or up to 4 years.

Ecology

The great majority of marine ostracods are benthonic or nektobenthonic. The only pelagic ostracods belong to the almost exclusively marine Myodocopa (Figure 4); they include active predators (e.g., on copepods) and scavengers. Pelagic Cypridinoidea

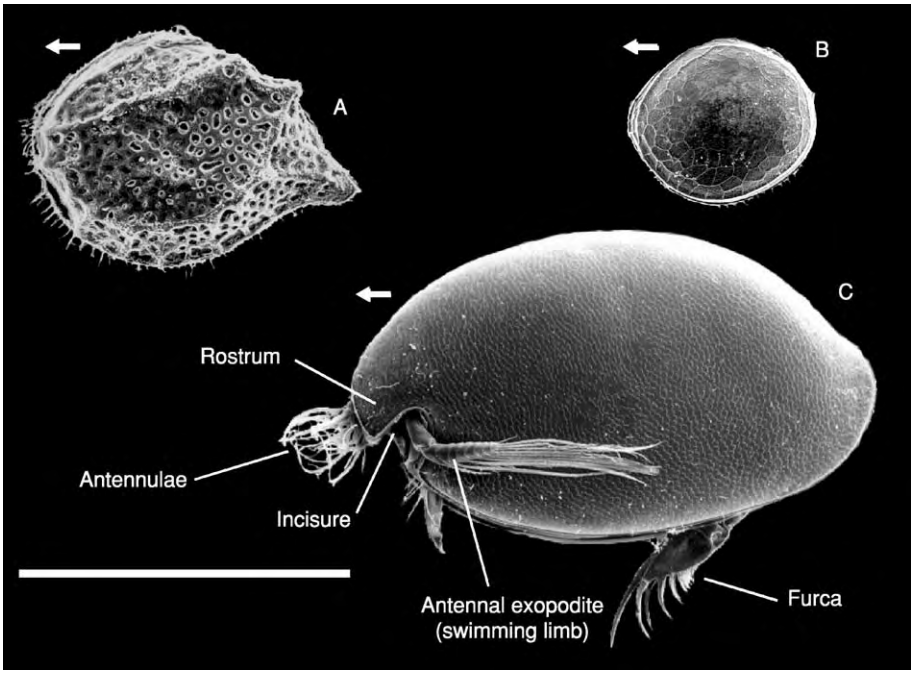


Figure 4 Shells of some Recent marine myodocopan ostracods. A: Myodocopida, Sarsielloidea; B: Halocyprida, Cladocopina; C: Myodocopida, Cypridinoidea (with limbs). Scale bar 1 mm.

(Myodocopida) live mainly in mesopelagic and abyssal zones but are also found in shallower waters. Of the Halocyprida, all of the Halocypridoidea and a few Thaumatocypridoidea, also predators and scavengers, are truly pelagic. Of the non-pelagic Myodocopa, the nektobenthonic halocyprid Thaumatocypridoidea inhabit not only abyssal and bathyal depths but also shallow marine caves which, like deep water environments, are dark and oligotrophic, with little temperature variation. Another halocyprid group, the Cladocopina, are benthonic/nektobenthonic; they live on fine sediment bottoms in deep water, but at least some shallow water species live interstitially in coarse sediments. The Cylindroleberidoidea and Sarsielloidea (both Myodocopida) are essentially benthonic/nektobenthonic, sometimes burrowing in fine sediments; they have a wide bathymetric range, being more prevalent in shallow coastal (even intertidal) waters. Cylindroleberidoids are filter-feeders, while many sarsielloids are voracious predators. Only a few myodocopans tolerate reduced salinities and are found in estuaries (e.g., some Sarsielloidea).

Of the Podocopa, the Platycopida (Figure 5A) are all benthonic filter-feeders found predominantly in marine environments (although a few are known from brackish waters), being most diverse in warm, shallow, carbonate environments, but also living in deep waters. The Palaeocopida, highly diverse in the Palaeozoic, are today extremely rare (the Puncioidea), living interstitially in shallow marine, high-energy biogenic sands off New Zealand.

The predominantly benthonic Podocopida are the most diverse marine ostracods (Figure 5B–M), comprising representatives of six extant superfamilies. The Bairdioidea are most abundant and diverse in warm, shallow carbonate environments; benthonic detritus-feeders, unable to swim, they are mostly epifaunal crawlers but some live interstitially. Benthonic Macrocypridoidea, benthonic/nektobenthonic Pontocypridoidea and a few benthonic/nektobenthonic Cypridoidea (the candonid Paracypridinae) are found in marine and brackish waters. These three cypridocopine groups include herbivores, detritivores and carnivores; some pontocypridoids are commensal on other invertebrates such as starfish. Cytheroidea are benthonic crawlers, climbers on algae, burrowers in fine sediment, or live interstitially in coarse sand, and some are commensal, such as the Entocytheridae on freshwater and marine crustaceans. Cytheroideans include herbivores, detritivores, and carnivores. Extant Sigillioidea are rare, living in warm, shallow marine environments (probably interstitially), submarine caves, and (in at least one case) the deep sea. At least one species of the predominantly non-marine Darwinuloidea lives in brackish as well as freshwater conditions.

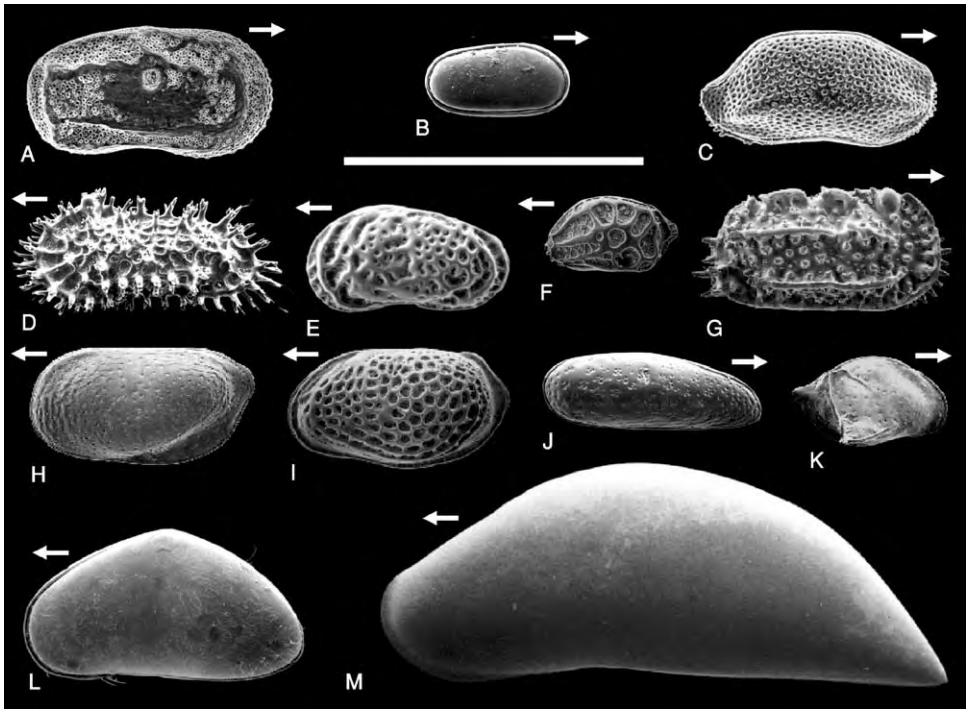


Figure 5 Shells of some Quaternary Recent marine podocopan ostracods: A: Platycopida; B M: Podocopida (B: Sigillioidea; C: Bairdioidea; D K: Cytheroidea; L: Pontocypridoidea; M: Macrocypridoidea). Scale bar 1 mm.

No myodocopans, and only three podocopan superfamilies, are found today in non-marine waters (which can include saline as well as fresh, flowing (lotic) and standing (lentic), temporary and permanent waterbodies) (Figure 6). The Cypridoidea are the most successful, with desiccation-resistant eggs allowing many taxa to inhabit temporary ponds, survive periods of drought, and achieve wide dispersal (eggs being wind-blown or carried by birds and other animals); they are benthonic/nektobenthonic, and herbivores, detritivores, predators and scavengers. Several families of the Cytheroidea have non-marine representatives, most notably the Limnocytheridae, of which many have female brood care and live in permanent lakes (but a few have desiccation-resistant eggs and live in temporary ponds). Most Darwinuloidea live in freshwater lakes, rivers, and springs; benthonic crawlers, they may filter-feed as well as grazing on algae and detritus, and the exclusively parthenogenetic females brood eggs and early instars. Some non-marine ostracods have highly specialized habitats, such as the limnocytherid genus *Elpidium* which lives in the water held by bromeliad plants in South America. Some freshwater ostracods are hypogean (predominantly cypridoidean Candonidae), living in deep, well-oxygenated ground water; others are found where ground water reaches the surface in springs and seepages, including hot springs up to 54°C.

Some representatives of essentially freshwater ostracod groups inhabit merely damp environments such as fen soils, mosses, and leaf-litter, including species of Cypridoidea, Cytheroidea, and Darwinuloidea. The Terrestricytheroidea are thought to have entered such habitats from marine littoral environments, however, and are exclusively associated with high intertidal and supratidal vegetation.

Geological History and Evolution

Thanks to an excellent fossil record (the best of any arthropod group) the geological history of the Ostracoda is well-documented and understood in most marine and non-marine environments (Figure 7). Some ostracod habitats have poor preservation potential, however: commensal ostracods (marine and non-marine) and semi-terrestrial taxa have little or no fossil record. The fossil record of Ostracoda begins in the Ordovician and their origins and relationships to other crustacean groups are obscure.

Ordovician ostracods were marine, probably meiobenthonic (possibly nektobenthonic) animals; faunas dominated by Palaeocopida (beyrichicopines and binodicopines) but including Podocopida (e.g., sigillioideans), Platycopida and Leperditicopida, were essentially confined to shelf areas and show lithofacies-related assemblages that can be related to depth; marginal marine tidal flat assemblages dominated by leperditicopids and palaeocopids are recognized, but there is no evidence for deep marine (bathyal–abyssal) faunas. The earliest myodocopans, marine nektobenthonic forms, appear in the Late Ordovician (Figure 7).

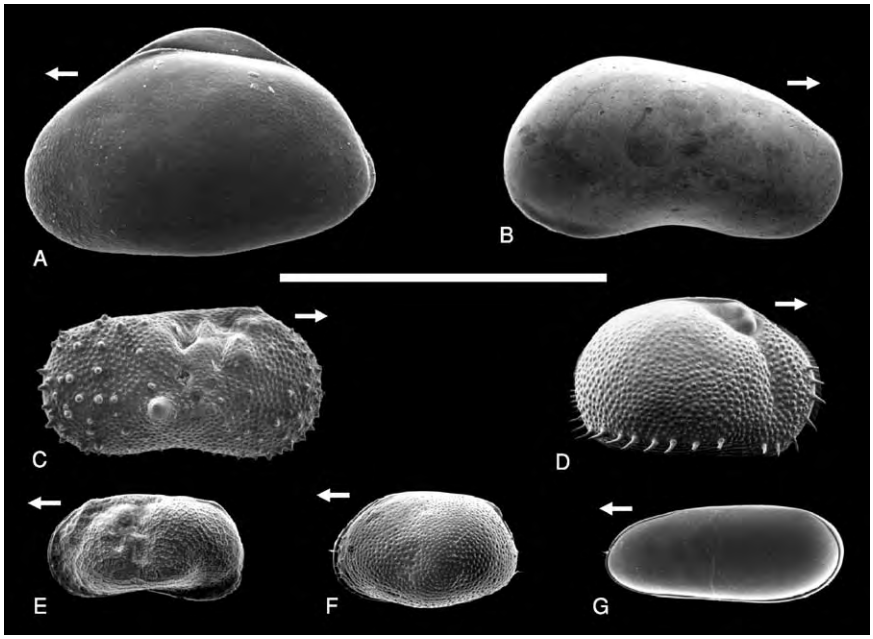


Figure 6 Shells of some Recent freshwater podocopid ostracods. A D: Cypridoidea; E F: Cytheroidea (Limnocytheridae); G: Darwinuloidea. Scale bar 1 mm.

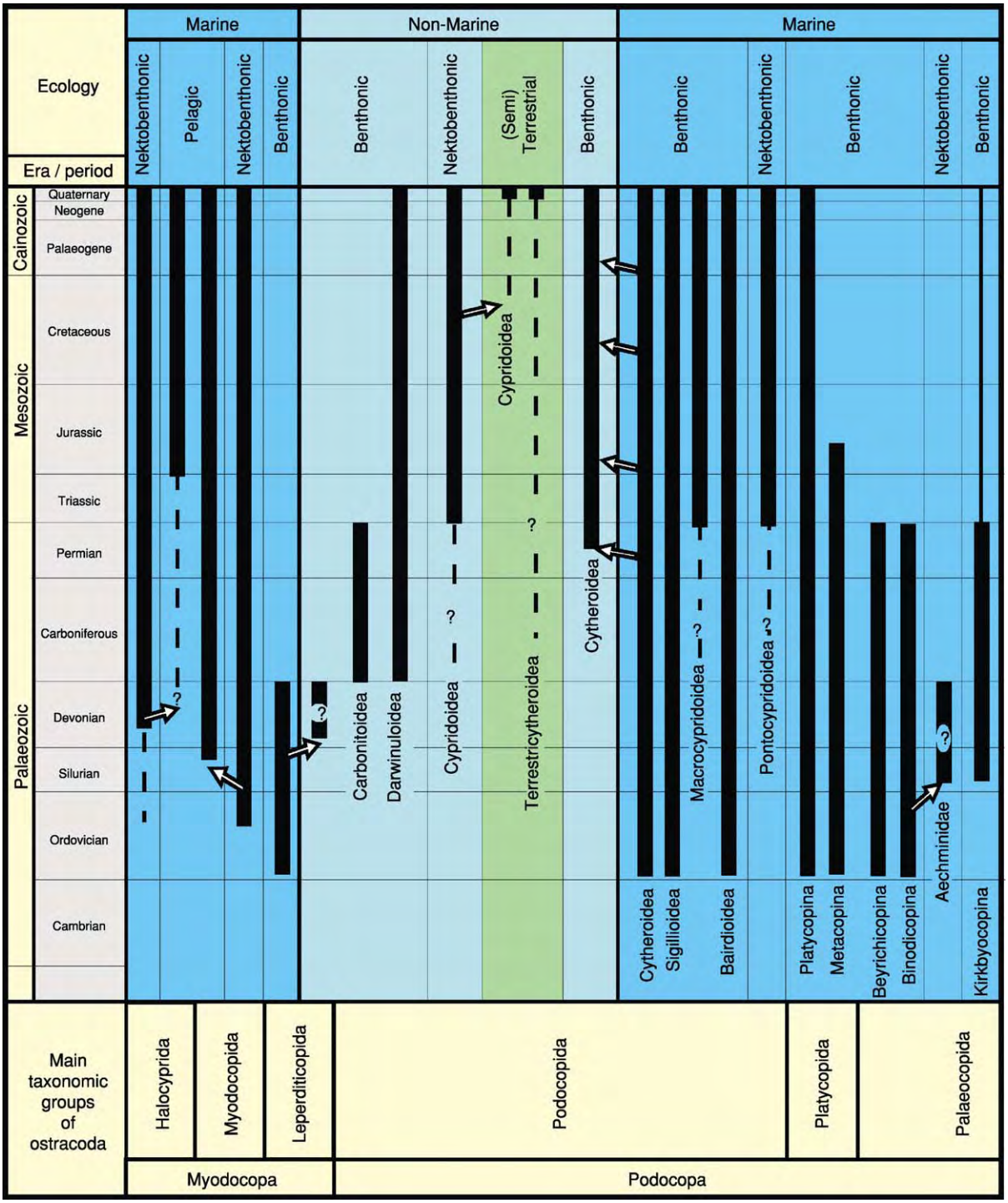


Figure 7 Stratigraphical time lines of the major ecological radiations of the Ostracoda.

In the Silurian, some palaeocopids (Figure 8), platycopids, and leperditicopids may have begun to adapt to marginal marine, brackish water (and perhaps hypersaline) conditions; diverse palaeocopids dominated shallow marine shelf waters, giving way to 'non-palaeocope' faunas (including metacopines) in deeper water. The Ordovician–Devonian leperditicopids include the largest known ostracods (up to 50 mm long) and inhabited marginal marine environments, including tidal flats and estuaries, possibly entering freshwater environments in the Devonian; it has been speculated that they occupy an ancestral

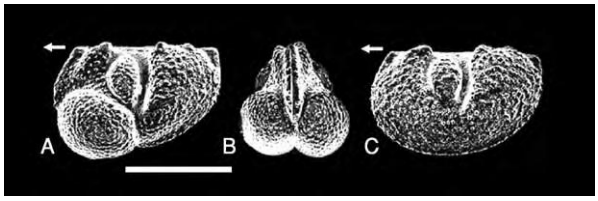


Figure 8 A Silurian palaeocopid ostracod (Beyrichicopina). A B: female, lateral and anterior views (note large anteroventral brood pouches); C: male. Scale bar 1 mm.

position in myodocopan evolution, but their affinities are controversial and there is some doubt that they are true ostracods.

The exclusively marine bairdioideans, one of the oldest extant podocopid lineages, are known from the mid-Ordovician onwards; they show a considerable variety of both deep- and shallow-water taxa by the Devonian, although post-Palaeozoic forms are most diverse in shallow carbonate environments. The podocopid cytheroidean family Bythocytheridae originated in the Ordovician, possibly sharing a common ancestor with the Bairdioidea, and underwent major radiations in the Devonian and again in the Jurassic–Cretaceous; while many cytheroidean families include at least some taxa that have adapted to brackish waters, bythocytherids never seem to have achieved this and are found only in fully marine salinities.

Distinctive Palaeozoic marine assemblages from the Devonian onwards have been related to shallow benthonic (high energy; dominated by thick-shelled ornamented podocopans), deep benthonic (low energy; thin-shelled spinose podocopans), and pelagic environments (myodocopans). The low-energy benthonic assemblage may represent a 'palaeopsychrospheric' fauna inhabiting deep cold waters below the thermocline in a stratified ocean, but this interpretation remains controversial.

The pioneer colonization of pelagic marine environments by myodocopan ostracods took place in mid-Silurian (Figure 7), after an ecological shift from a nektobenthonic mode of life, probably in response to bottom water oxygen deficiency and the opportunities offered by the plankton-rich, well-oxygenated upper waters. Another myodocopan group, the halocypridoidean halocyprids, may have undergone a similar shift, but they have an extremely poor fossil record and evidence for the existence of pelagic halocypridans since the Devonian, as shown on Figure 7, is limited and controversial; in the Silurian – Carboniferous they are represented by entomozooideans, a group of questionable affinity. A Silurian–Devonian palaeocopid binodocopine family, the aechminids, with a large, hollow dorsolateral spine on each valve, are also thought to have been pelagic, since it is difficult to imagine such a carapace morphology as functionally suitable for a benthonic organism; since the end of the Palaeozoic, however, there have been no pelagic podocopans.

The first undoubted non-marine ostracods, the podocopid superfamilies Carbonitoidea and Darwinuloidea, entered freshwater environments in the Early Carboniferous and radiated to high diversity in the Late Carboniferous–Permian; at the end of the Permian the former became extinct and the latter reduced in diversity (Figure 7). Post-Palaeozoic non-marine faunas are dominated by Cypridoidea and limnocytherid Cytheroidea; both lineages may have Late Palaeozoic origins but their early history is controversial, as is the possibility of a Carboniferous brackish-freshwater radiation of some platycopids, which today are an exclusively marine group. Ostracods, which may represent the first limnocytherids, proliferated in Late Carboniferous and Permian lakes and coal swamps. It is notable that darwinuloideans, which in Palaeozoic assemblages show distinctive sexual dimorphism (the females having expanded posterior brood chambers), have apparently been exclusively parthenogenetic since the Early Mesozoic.

Almost all of the Palaeocopida, so characteristic of Palaeozoic marine faunas, became extinct at the end of the Permian; only the Puncioidea survived and are represented today by a single living species. Post-Palaeozoic marine faunas are dominated by cytheroidean Podocopida. Major radiations of cytheroidean families took place in the Mesozoic, for example the Cytheruridae (Triassic onwards), the Progonocytheridae and Schulerideidae (mid-Jurassic–Early Cretaceous), the Trachyleberididae and Brachycytheridae (Late Cretaceous onwards), and the Cytherettidae, Hemicytheridae, Loxoconchidae, Leptocytheridae, and Xestoleberididae (Tertiary). The origins of the diverse modern deep-sea (psychrospheric) benthonic ostracod fauna (predominantly cytheroideans) were in Mesozoic faunas which evolved in a thermospheric ocean and were forced to adapt to cooling conditions in the Tertiary. Metacopine platycopids radiated in the Triassic and Early Jurassic and then became extinct. Platycopina were often the dominant group (in terms of abundance rather than diversity) in the chalk seas of the Late Cretaceous; their highest diversity today is in sub-tropical or tropical shallow carbonate environments.

In contrast to podocopans, the fossil record of myodocopans is poor, since in many the valves are weakly or not at all mineralized. The nektobenthonic cladocopine halocyprids are usually better calcified, however; they originated in the Palaeozoic and became relatively diverse in the Mesozoic, achieving their greatest diversity in the deep (bathyal) waters of

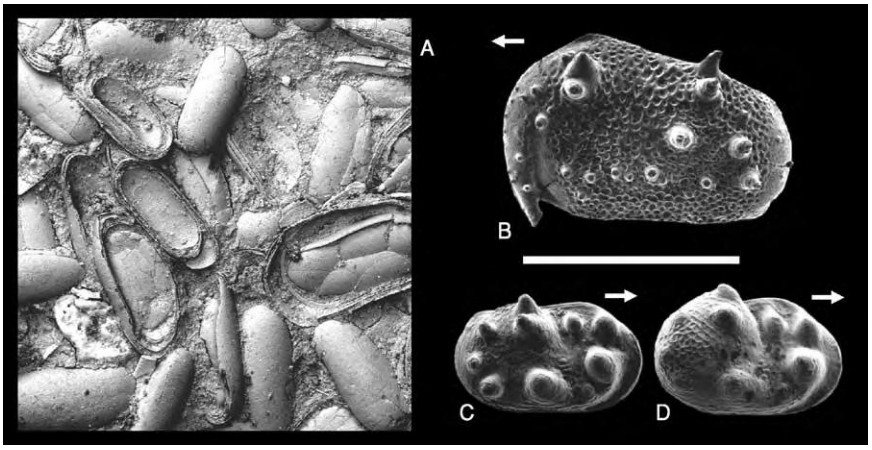


Figure 9 Some Early Cretaceous nonmarine ostracods. A: bedding plane assemblage, mainly Darwinuloidea; B: *Cypridea* (Cypridoidea); C D: Cytheroidea, Limnocytheridae (C: male; D: female, with expanded posterior brood chamber). Scale bar 1 mm.

the modern Arctic Ocean, where competing pandemic deep-sea cytheroideans are excluded by physical barriers such as the shallow Bering Strait.

Mesozoic to modern non-marine faunas are dominated by Cypridoidea, but include significant cytheroidean (particularly the Limnocytheridae) and darwinuloidean components (Figure 9). The marine cytheroidean family Cytherideidae also gave rise to significant post-Palaeozoic non-marine radiations in brackish, hypersaline, and freshwaters. A major Mesozoic radiation of cypridoidean podocopids took place in non-marine environments from the latest Jurassic onwards. A key factor may have been their possession of desiccation-resistant eggs, enabling them to inhabit temporary water bodies and achieve wide dispersal. However, much of the Early Cretaceous cypridoidean diversity is attributable to one family, the Cyprideidae (*Cypridea* and related genera), which subsequently became extinct in the Palaeogene, although they co-existed with representatives of extant families; the reasons for their differential success and subsequent demise are obscure.

Applications

Benthonic marine ostracods, being facies-controlled, are of limited value for wide-scale biostratigraphy, but their small size, morphological diversity, and excellent fossil record have nevertheless rendered them extremely useful on regional or intrabasinal scales from the Ordovician to the Quaternary. Ostracods have also been used with great success in establishing zonation schemes and correlations in non-marine depositional settings, such as the Mesozoic 'Wealden' basins of England.

Quaternary deep-sea benthonic ostracod assemblages characteristic of different water masses have considerable potential for palaeoceanographic and palaeoclimatic studies, such as the identification of mode switches in thermohaline circulation. Ostracods were used to great effect in documenting the development of a stratified ocean (with a shallow, warm thermosphere overlying a deep, cold psychrosphere) during the Cenozoic, as well as other palaeoceanographic events, such as the opening of the Drake Passage and the closure of the Panama Isthmus.

Shallow and marginal-marine ostracods have proved valuable as indicators of sea-level change, both by the recognition of depth-related shelf assemblages and by the identification of intertidal taxa as sea-level proxies.

Two groups of marine ostracods have been used as fossil proxies for dissolved oxygen levels. Platycopids, as filter-feeders and brooders, may be better able than podocopids to survive episodes of low oxygen availability in the benthic marine environment, by virtue of more efficient mechanisms (branchial plates) for circulating water (for respiration) through the domiciliar and brood spaces in the carapace. The 'platycopid signal' indicative of low-oxygen conditions, first recognized in differential success of platyocopids compared to podocopids during the mid-Cretaceous Cenomanian–Turonian Oceanic Anoxic Event, has been widely applied, not only in the Mesozoic but also in the Palaeozoic; in the latter case palaeocopids, metacopines, and platycopines have all been counted as filter-feeders. Variation in the size and shape of the vestibulum in the podocopid cytheroidean genus *Krithe* and its close relatives has also been used (controversially) as an indicator of mid-Cretaceous to Quaternary marine oxygen levels. However, it must be noted that neither of these attractive applications is so far adequately supported by evidence of the biology, ecology and distribution of modern taxa.

Ostracods are particularly effective for palaeoenvironmental analysis in marginal marine settings, bearing in mind the need to distinguish between autochthonous and allochthonous components of assemblages (the former representing the local enviroment, the latter being mixed by post-mortem transport, perhaps by tidal currents, from a wider range of environments). The transition from marine to brackish conditions is usually reflected in changes in species composition and a reduction in the diversity of assemblages. Intraspecific variation, for example in carapace ornament, can often be related to different environmental conditions, although caution is advised since different taxa undoubtedly respond in different ways. In the living euryhaline cytherideid *Cyprideis torosa* (a cytheroidean podocopid) variations in the development of nodes and the shape of sieve pores are influenced by salinity, and this knowledge has been effectively applied to the interpretation of Neogene and Quaternary fossil examples; the relationships are complex, but noded forms with round sieve pores may generally be taken to indicate low-brackish salinities, while smooth forms with irregular pores record hypersaline conditions.

Differences in non-marine ostracod assemblages reflect different water chemistries and habitat types as well as overall salinity. That Mesozoic Purbeck-Wealden faunas of England are essentially non-marine is as much evident from the absence of platycopids and bairdioideans as from the abundance of cypridoideans (especially *Cypridea* and its allies), cytheroideans (limnocytherids and cytherideids), and darwinuloideans (**Figure 9**); some cytheroidean taxa with marine affinities may indicate periodic marine incursions in coastal settings, or alternatively the existence of saline inland lakes. Darwinuloideans and some limnocytherids had brood care and must have lived in permanent water bodies, while many *Cypridea* species may represent temporary ponds, perhaps in ephemeral river beds. Limnocytherids have also been related to high alkalinity waters, *Cypridea* to low alkalinities.

Quaternary non-marine ostracods can be interpreted in similar ways and it has been shown that different species inhabit different lakes with the same salinity (ionic concentration) but different water chemistry (ionic composition). They are increasingly being used as sources of biogenic calcite for the analysis of trace metals, stable isotopes, and even amino acids, with applications in palaeoclimatology (e.g., as palaeotemperature and palaeosalinity proxies) and age determination. Temperature is an important influence on the geographical distribution of both marine and non-marine ostracods, and many taxa are effective indicators of cold (glacial) and warm (interglacial) conditions. Ostracods are frequently included in multi-proxy studies of Quaternary archaeological sites.

See Also

Fossil Invertebrates: Arthropods.

Further Reading

Athersuch J, Horne DJ, and Whittaker JE (1989) *Marine and brackish water ostracods*. Synopses of the British fauna (New Series), No. 43. EJ Brill: Leiden.

Benson RH (1990) Ostracoda and the discovery of global Cainozoic palaeoceanographical events. In: Whatley R and Maybury C (eds.) *Ostracoda and global events*, pp. 41 58. London: Chapman and Hall.

Carbonel P, Colin J P, Danielopol DL, Löffler H, and Neustrueva I (1988) Palaeoecology of limnic ostracodes: a review of some major topics. *Palaeogeography, Palaeo climatology, Palaeoecology* 62: 413 461.

De Deckker P, Colin J P, and Peypouquet J P (eds.) (1988) *Ostracoda in the Earth Sciences*. Amsterdam: Elsevier.

Holmes JA and Chivas AR (eds.) (2002) *The Ostracoda: applications in Quaternary research*. AGU Geophysical Monograph **131**. Washington DC: American Geophys ical Union.

Griffiths HI and Holmes JA (2000) *Non marine ostracods and Quaternary palaeoenvironments*. Quaternary Re search Association Technical Guide No. 8. London: Quaternary Research Association.

Horne DJ (2002) Ostracod biostratigraphy and palaeoecol ogy of the Purbeck Limestone Group in southern England. *Special Papers in Palaeontology* 68: 53 70.

Martens K (ed.) (1998) *Sex and parthenogenesis. Evolu tionary ecology of reproductive modes in non marine ostracods*. Leiden: Backhuys.

Meisch C (2000) *Freshwater Ostracoda of western and central Europe*. Susswasserfauna von Mitteleuropa **8/3**. Heidelberg: Spektrum Akadamischer Verlag.

Park LE and Smith AJ (eds.) (2003) *Bridging the gap: trends in the ostracode biological and geological sciences*. Palaeontological Society Papers **9**.

Siveter DJ (1984) Ecology of Silurian ostracods. *Special Papers in Palaeontology* 32: 71 85.

Siveter DJ, Vannier JMC, and Palmer D (1991) Silurian Myodocopes: pioneer pelagic ostracods and the chron ology of an ecological shift. *Journal of Micropalaeontol ogy* 10(2): 151 173.

Siveter DJ, Sutton MD, Briggs DEG, and Siveter DJ (2003) An ostracode crustacean with soft parts from the Lower Silurian. *Science* 302: 1749 1751.

Whatley R (1995) Ostracoda and oceanic palaeoxygen levels. *Mitteilungen aus dem Hamburgischen Zoolo gischen Museum und Institut, Hamburg* 92: 337 353.

Williams M, Floyd JD, Miller CG, and Siveter DJ (2001) Scottish Ordovician ostracodes: a review of their pa laeoenvironmental, biostratigraphical and palaeobiogeo graphical significance. *Transactions of the Royal Society of Edinburgh: Earth Sciences* 91: 499 508.

Palynology

P Coxon and G Clayton, Trinity College, Dublin, Ireland

Introduction

'Palynology' is a term that was introduced in 1944 to cover a broader range of studies than was covered by the earlier term, 'pollen analysis', which had distinct Quaternary connotations. Initially, palynological research dealt mainly with spores and pollen, but many other types of organic-walled microfossils were subsequently investigated. The useful term 'palynomorph', thus includes a broad spectrum of microfossils (such as the miospore shown in Figure 1) that are resistant to the mineral acids and other reagents commonly used for fossil extraction from rocks and sediments. As a consequence of the inclusive definition, based on chemical composition of the microfossil wall, palynology studies encompass microfossils of diverse biological affinities, both plant and animal. Although many different techniques for the study of palynomorphs have been employed, by far the most commonly used are transmitted light microscopy for routine examination and scanning electron microscopy for the elucidation of surface structure and ornamentation, at much higher magnifications.

Major differences separate Quaternary and pre-Tertiary palynology. In the former, almost all taxa recorded are from extant plants, whereas in the latter, the taxa are almost invariably extinct. This is strongly reflected in the palynomorph classifications utilized, with pre-Tertiary palynomorphs typically being assigned to 'form' (i.e. morphology-based) rather than natural genera and species. Equally striking differences are apparent in terms of biostratigraphy, with Quaternary dating and correlation strongly tied to the recognition of changing climate but with pre-Tertiary biostratigraphy being normally based on the first appearances and extinctions of taxa in the stratigraphic record. The Tertiary represents a complex 'overlap' interval, with uncertainties concerning both the appropriate scheme for the classification of palynomorphs and the biostratigraphic approach adopted.

Figure 1 *Knoxisporites stephanephorus* Love 1960. A transmitted light photomicrograph of a typical Mississippian miospore.

Quaternary Palynology

Palynologists working on Quaternary material have, for the most part, concentrated primarily on the pollen of gymnosperms and angiosperms as well as on the spores of bryophytes and pteridophytes (this work being colloquially referred to as 'pollen analysis'). However, a range of Quaternary fossil material is considered as palynological and could be referred to here (including fungal spores). The problems of reconstructing complex Quaternary environments from a wealth of detailed evidence have led palynologists to use a wide range of fossil groups in order to provide multiproxy evidence that can allow more accurate models of past changes. There is also a wide literature on other microscopic fossil material, e.g., charcoal, cuticle fragments, phytoliths, and testate amoebae.

Quaternary palynology has the distinct advantage that taxa are, by and large, still extant, allowing the use of modern reference material collected from living plants in addition to keys, pollen and spore atlases, and World Wide Web-based identification pages offering digital images of pollen and spores. Such a wealth of information allows identification of Quaternary (and some Tertiary) fossil palynological material to the level of family, often to genus, and sometimes to species, giving the Quaternary palynologist a wide range of options with regard to palaeoenvironmental reconstruction and other applications relevant to palynological data.

Early palynological research in the Quaternary concentrated on vegetation history and produced (albeit initially unrefined) palaeoecological reconstructions. This research quickly became more sophisticated,

leading to the gathering of much detailed information on former plant distributions and to a tradition of palaeobotanical studies that continues, with considerable enhancement of the technique, to this day. The palynological elucidation of detailed vegetation histories of north-western Europe was followed by attempts to produce biostratigraphic correlations, which, in some instances, merged into climatostratigraphies. Holocene palynological research was inevitably linked to the human influence on both landscape and vegetation and early work began to show the importance of the anthropogenic impact on natural ecosystems. Palynological research into such influence on the environment is still of primary importance today, and a detailed knowledge of past environment has allowed present and future management and protection of sensitive ecosystems to be better planned.

Problems are inherent in the time-transgressive nature of temporally and spatially detailed Holocene palynological records; this has caused obvious problems with correlation and has led to a movement away from relying on biostratigraphy deduced from Holocene sequences. For Late Pleistocene and Holocene work, radiocarbon dating has permitted detailed correlation and synthesis of palynological data from sequences younger than, in theory, *ca.* 40 000 years, although the most certain dates come from within the limits of reliable radiocarbon calibration methodologies (<13 000 years). Even within these limits, the complexities of atmospheric radiocarbon production severely hamper attempts to date important environmental/geological transitions (e.g., the radiocarbon 'plateau' of the critical last glacial–interglacial transition some 13 000–10 000 radiocarbon years ago).

Despite calibration problems with radiocarbon dating, the ability to date palynological assemblages from the Holocene has enabled palynologists to map the timing of migration and vegetational succession in great detail, allowing isopoll and isochron maps to be constructed and giving palynologists a great deal of evidence regarding plant migration, vegetational succession, and the timing of climatic (and other environmental) impacts on vegetation. **Figure 2** shows reconstructions of European vegetation zones at four different time periods. The use of radiocarbon-dated palynological data allows such detailed palaeoecological maps to be presented. However, biostratigraphy remains important in many older (Pleistocene) palynological records, for which 'pollen analysis' has long been used to characterize the oscillating nature of the climate at a coarse level (i.e., warm stage/cold stage) and at a high resolution (e.g., the analysis of vegetation history and biostratigraphy of interglacial or temperate-stage sequences). **Figure 3** shows the palynological record from four successive interglacials from southern France. Pollen diagrams such as these from interglacial sequences have an important role in Pleistocene biostratigraphy, palaeoecological studies, and biogeographical analyses, and in understanding regional and continental vegetational histories.

The distinctive nature of successive temperate stages is emphasized in Europe by the geographical 'trap' that is formed during cold stages, when climatic conditions deteriorate and thermophilous taxa become locally extinct. During cold stages, plant taxa inhabit refugia to the south and east, but there are severe barriers to migration (e.g., the Pyrenees, the Alps, and the continentality of eastern Europe and the Mediterranean). Each successive cold stage saw extinction of taxa from Europe, and so throughout the Pleistocene, the number of taxa recorded in each interglacial generally declines. This decline in represented taxa provides unique assemblages in temperate stages and the differing rates of migration of taxa, and their relative success provides distinctive vegetational successions in each interglacial (see **Figure 3**).

More recently, problems with using biostratigraphy on short or incomplete sequences have been highlighted, and the use of advanced dating methodologies and mammalian biostratigraphy (e.g., in the UK and north-western Europe) has suggested that there has been a compression of the number of temperate stages that actually exist into a smaller number of named stages, because of a lack of palynological resolution. The 'missing stages' are glaringly apparent from the far more complete marine oxygen isotope records. However, the apparent resolution in the biostratigraphy of the interglacial sequences in south central France (**Figure 3**) shows the value of palynological work on complete temperate-stage deposits, and palynology still holds great value for analysing temperate-stage palaeoenvironments.

Inferences regarding climate change are of paramount importance in research stemming from modern palynology, and the need to understand climate variability has led to palynological data becoming particularly important as a proxy data source for climatic parameters. In this regard, the use of taxa to act as a proxy for climate variables was recognized early on; this approach has continued to be used as models of the relationships between individual taxa, and even whole fossil assemblages, and climate parameters are clarified. This research has proved particularly important in large-scale modelling of global climate change using fossil data from widespread regional sources.

The most recent geological period, the Quaternary, is characterized by dramatic shifts in climate forced

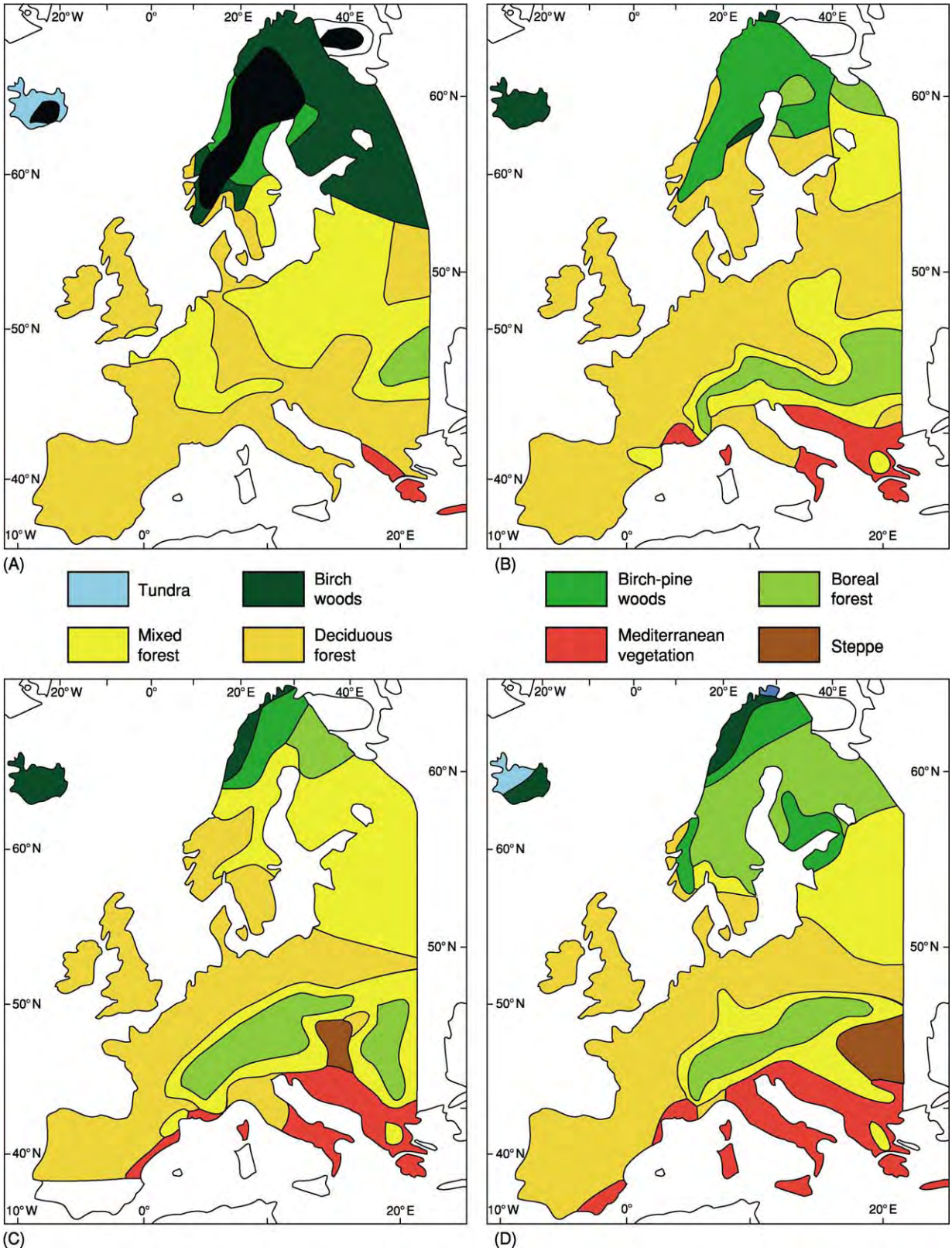


Figure 2 Vegetational regions of Europe reconstructed from pollen data for (A) 9000, (B) 6000, and (C) 3000 years BP and (D) the present. After Huntley B and Prentice IP (1993) Holocene Vegetation and Climates of Europe. In: Wright HE Jr, *et al.* (eds.) *Global Climates Since the Last Glacial Maximum*. University of Minnesota Press.

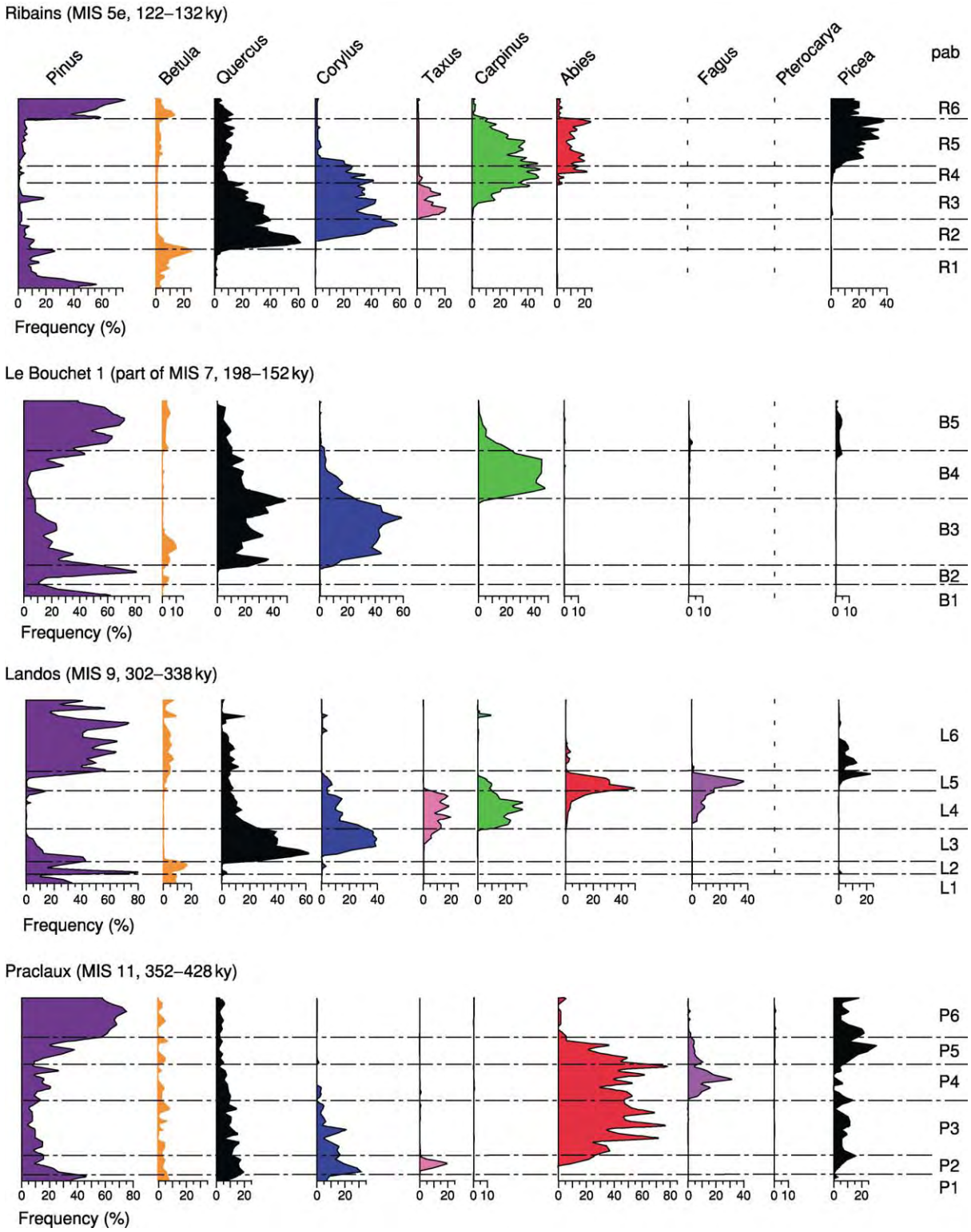


Figure 3 Relative percentage pollen diagrams (selected taxa) of four successive Pleistocene temperate stages (interglacials) from the Velay region of France. The correlation to marine isotope stages (MIS) and tentative ages of the deposits are shown. The behaviour of some of the individual taxa (coloured curves) during each temperate stage and the unique succession of pollen assemblages are apparent. Possible local pollen assemblage biozones (pab) are indicated; these can be subdivided as necessary to produce more detailed descriptions of assemblages. The regional disappearance of taxa e.g., *Pterocarya*, after the Middle Pleistocene (MIS 11), and *Fagus*, by the last interglacial (MIS 5e) is also apparent. After De Beaulieu JL and Reille M (1995) Pollen records from the Velay craters: A review and correlation of the Holsteinian Interglacial with isotopic stage 11. *Mededelingen van de Geologische Dienst* 52: 59 70.

by the astronomical position of Earth relative to the sun. The problems of reconstructing complex Quaternary environments have led palynologists to use a wide range of fossil groups in order to provide multiproxy evidence that can allow more accurate models of past changes.

Pre-Quaternary Palynology

Hydrocarbon exploration provided the main impetus for palynological biostratigraphy (or 'palynostratigraphy') in the second half of the twentieth century, due largely to the realization that small fragmentary rock samples, such as 'ditch cuttings' from drilling, could yield thousands of intact palynomorphs. The erection of zonal schemes for non-marine successions based on spores and pollen in the Carboniferous (Figure 4) was followed rapidly by the publication of similar schemes for the Devonian and later systems. The stratigraphic resolution of these palynological zonations compares very favourably with many schemes based on invertebrates; the average duration of Upper Palaeozoic spore and pollen biozones/sub-biozones is less than 3 Ma.

Detailed zonations of marine Palaeozoic rocks have been erected, based mainly on acritarchs (*see* **Microfossils:** Acritarchs) and chitinozoa (*see* **Microfossils:** Chitinozoa), whereas palynostratigraphic work on marine Mesozoic and Tertiary successions has relied principally on dinoflagellate studies. In relatively recent times, palynological investigations of Precambrian rocks have produced spectacular results in terms of both enabling much needed biostratigraphic correlation and shedding light on the nature of early life in Earth's oceans.

The limited palaeogeographical distribution of many pre-Quaternary spore and pollen species restricts the applicability of most palynological zonations to specific regions, and in most cases rules out any possibility of 'global' correlation. However, the converse of this obvious disadvantage is that floral provinces can be recognized, which can complement palaeomagnetic and other evidence in making large-scale palaeogeographical reconstructions.

Other Geological Applications

Not all Phanerozoic palynological research has been biostratigraphic. Inspired by palaeoecological studies of Quaternary peats, pioneering work on spore and pollen assemblages from Palaeozoic coal seams led to convincing interpretations of the gradually changing environmental conditions responsible for the accumulation of the peat. More comprehensive

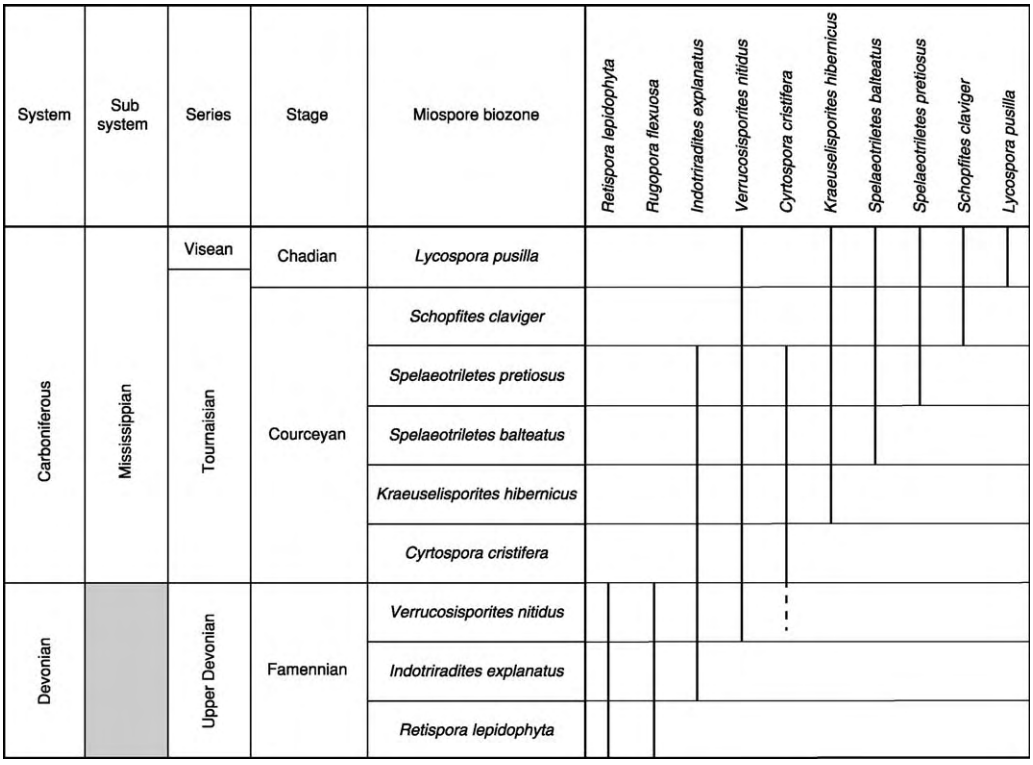


Figure 4 Part of the miospore zonation for the Late Devonian and Early Carboniferous of western Europe, showing the stratigraphic ranges of selected miospore taxa.

Figure 5 Specimens of the acritarch *Veryhachium* from four Devonian samples of different maturity. With increasing temperature, the colour changes from almost colourless, through yellow, to brown, and eventually to black.

palaeoenvironmental interpretations based on palynological investigations of lithologies other than coal are often referred to as 'palynofacies analyses'. These studies extend beyond the identification of just the palynomorphs present, to consideration of all of the 'palynodebris' – the microscopic organic matter present, including particles such as woody debris, charcoal, and cuticle. Interpretations based on palynofacies analysis typically enable distinction to be made between sediments deposited in marine and non-marine environments, and, in the former case, permit some estimate of proximity to shoreline and energy level of the depositional environment.

A further example of the non-stratigraphic applications of palynology concerns the gradual carbonization of organic matter resulting from increased temperature. The 'thermal maturity' of sediments is critical in terms of hydrocarbon generation but is often difficult to assess accurately. The irreversible darkening of palynomorphs that results from carbonization (Figure 5) can be measured and may be used in place of traditional methods such as vitrinite reflectance determination to estimate maturity in rocks deficient in vitrinite.

See Also

Biozones. **Microfossils:** Acritarchs; Chitinozoa. **Palaeoclimates**. **Palaeoecology**. **Petroleum Geology:** The Petroleum System. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Batten DJ (1999) Palynofacies analysis. In: Jones TP and Rowe NP (eds.) *Fossil Plants and Spores: Modern Techniques*, pp. 194 198. London: Geological Society.

Clayton G and Coxon P (1999) Spore and pollen biostratigraphy. In: Jones TP and Rowe NP (eds.) *Fossil Plants and Spores: Modern Techniques*, pp. 225 229. London: Geological Society.

Donald G (1996) Non Aquatic Quaternary. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 879 910. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Fensome RA, Riding JB, and Taylor FJR (1996) Dinoflagellates. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 107 169. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Jarzen DM and Nichols DJ (1996) Pollen. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 261 291. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

MacDonald GM (1996) Non Aquatic Quaternary. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 879 910. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Miller MA (1996) Chitinozoa. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 307 336. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Moore PD, Webb JA, and Collinson ME (1991) *Pollen Analysis*, 2nd edn. London: Blackwell Scientific Publications.

Playford G and Dettmann ME (1996) Spores. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 227 260. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Reille M (1992) *Pollen et Spores d 'Europe et d' Afrique du Nord*. Marseille: Laboratoire de Botanique Historique et Palynologie.

Strother PK (1996) Acritarchs. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 81 106. Salt Lake City: American Association of Stratigraphic Palynologists Foundation.

Wright HE Jr, Kutzbach JE, Webb T III, Ruddiman WF, Street Perrott FA, and Bartlein PJ (eds.) (1993) *Global Climates Since the Last Glacial Maximum*. Minneapolis: University of Minnesota Press.

MICROPALAEONTOLOGICAL TECHNIQUES

I J Slipper, University of Greenwich, Chatham Maritime, UK

Introduction

The techniques used in the study of micropalaeontology tend not to be discussed in any great detail in most published accounts; many reports dismiss the subject of methods used with 'standard techniques were employed' or some such statement. This is often a mistake, because each group of microfossils will benefit from different treatments and each rock type will present challenges to the technician. To judge results of any scientific work, it is necessary to know which methods were used (e.g., during sample preparation). The focus of this article is the many ways in which microfossils may be extracted from their matrix and then prepared for study. Topics covered include field sampling/procedures and laboratory techniques, including extraction of microfossils of different compositions, such as calcareous, organic walled, phosphatic, and siliceous, from rocks of varying lithology. General methods of washing and sieving and specialist methods of separation, such as flotation and magnetic separation, are also covered.

Field Procedures

The approach in the field varies according to the type of work undertaken and the fossil group being studied. The sampling interval should be commensurate with the objectives: a high-resolution study may require taking samples every 10 cm whereas a reconnaissance survey can be carried out with samples taken every metre. The amount taken in the field is also dependent on the fossil group, on the fossil sizes, and the estimated abundance of specimens. Sample quantities taken in the field are usually larger than is needed for laboratory study. This allows some of the raw sample to be held in reserve for use either as reference material or as replacement in case of mishaps, or, more likely, when the results prove that further study of a different nature is required. The most important aspect of field procedure is awareness and avoidance of possible contamination from sources other than that of interest. The route to avoiding contamination is routine cleaning of the sampling equipment between taking successive samples. Of equal importance is good record keeping and labelling. Accurate logs should be drawn up in the field and the horizons sampled must be measured in relation to key marker horizons. Microfossils may also be found in abundance on bedding surfaces, such as ostracods in the Cretaceous Wealden Clay. In these cases, laboratory processing will destroy many specimens, and it is far better to recover large oriented blocks and to examine the surfaces directly under a binocular microscope. Where spot sampling is undertaken, the procedure is to remove all weathered material from the surface of the outcrop and to excavate a clean, unweathered block of the sediment, which is then placed in a clean labelled bag for transport and storage.

Extraction Methods

The techniques that have been devised to extract microfossils from their matrix are as varied as the number of different fossil groups and the types of rock in which they are found. There is no single overall method for extracting microfossils from rock, though some general-purpose methods may enhance achieving this. To be successful, a preparator selecting a method needs to be aware of the composition not only of the fossils concerned, which may be calcareous, siliceous, phosphatic, organic, or, in some cases, combinations of these, but also of the composition of the matrix, which could be any of the manifold types of sedimentary rocks. In cases in which the compositions of the microfossil and matrix differ, such as ostracods in a clay, simple methods may suffice, but in cases in which the two are similar, such as foraminifera in chalk or radiolaria in chert, the processes involved become complex. In these cases, other factors, such as differences in grain size, crystallinity, structure, and porosity of the matrix, must become the focus of attention for the method.

The methods surveyed here have been devised by many workers in laboratories around the world looking for efficient ways to separate the precious microfossils from the rock, while retaining a representative sample of the original assemblage in terms of abundance and diversity, in addition to minimizing any damage to the fragile specimens; which are easily damaged during processing. Some of the processes are dangerous and should be undertaken only in a properly equipped laboratory with fume cupboards and protective clothing, whereas others can easily be carried out with minimal equipment, such as an oven, a sieve, and some water. Though the techniques may be associated with certain groups, many

methods are suitable for use across different groups. The selection of the method, therefore, will depend as much on the facilities and methods of the laboratory and the experience of the technician as on the 'correct' laboratory manual method.

Calcareous Microfossils

Foraminifera (*see* **Microfossils:** Foraminifera) and ostracods (*see* **Microfossils:** Ostracoda) both secrete calcitic material that is preserved in the fossil record, and are generally of a similar size and are often found together in the same environments. There are exceptions to this, but techniques suitable for one group may be applied to the other. Calcareous nannofossils, however, by virtue of their much smaller size, require slightly more specialized treatment.

Foraminifera and ostracods The sample size required is generally between 250 and 500 g; once cleaned, coarsely broken up, and oven dried, the sample is weighed prior to the start of processing. Weighing enables the specimen abundance to be estimated on completion of the laboratory work. During the drying stage, there is a possibility that any clay in the sample could be baked hard, rendering any further processing more difficult, so drying is carried out at low temperatures for many hours. The extraction of foraminifera and ostracods from argillaceous sediments is relatively straightforward and comes close to being a standard technique; this is because both the composition and the grain size differ between fossil and matrix.

If the sample is poorly consolidated, a simple boiling in distilled water may be sufficient to break down the rock; this can be done in a microwave oven or on a hotplate. Stirring the sample is avoided because this can easily damage specimens. Various agents can be added to the water if the initial boiling is ineffective. A detergent such as Teepol may help, but this has been shown to cause some dissolution of delicate calcitic material. One of the problems with clay-rich sediments is that they tend to be sticky and difficult to disperse. In these circumstances, a deflocculating agent will help, such as sodium carbonate (washing soda), sodium hydroxide, or sodium hexametaphosphate (Calgon). The ideal result is to turn the rock into a fine soup that may then be sieved to concentrate the residue.

Better lithified sediments, such as shales, require treatment harsher than simple boiling in a water softener. In these cases, a weak solution of hydrogen peroxide (15%) is used. On contact with the rock, the effect of the generation of small bubbles of oxygen within the pore spaces breaks the rock apart. However, hydrogen peroxide can cause dissolution of the microfossils, so the time spent immersed has to be kept to minimum. This is particularly important if the rock contains any pyrite, because reaction with hydrogen peroxide will create sulphuric acid, which further dissolves the microfossils. Other similar methods employ sodium hypochlorite (domestic bleach) or sodium hydroxide; these are less destructive but take a longer time to be effective.

For marly sediments, a very effective method is to use an organic solvent, such as petroleum spirit, paraffin, or white spirit (turpentine substitute). After the rock has been dried and all pore spaces opened, it is immersed in spirit for several hours. The soaking time is dependent on the porosity of the rock; for very fine-grained sediments, a longer time is required for the spirit to enter the rock. The spirit is then decanted off, filtered, and saved for reuse. The vessel containing the rock is then topped up with water, which will cause the sample to disintegrate and form a thick slurry. The slurry is then boiled in excess water and washed with a detergent before sieving.

If the rock type is carbonate rich, the difference between the composition of the fossil and the rock is minimal and chemical methods may not be effective. In these cases, physical breakdown of the rock is achieved by forces of crystallization within the pore spaces; chalks will respond well to this approach. The most commonly used material for this procedure is sodium sulphate decahydrate (glauber salt) in a supersaturated solution. After the rock chips have been dried in the oven, the still-warm sample is immersed in the solution and returned to the oven and allowed to soak thoroughly. When saturated, the liquid is decanted and the sample is placed in a freezer overnight. Rapid freezing has the effect of causing many nucleation sites, allowing fast growth of very small crystals in the pore spaces. This is sufficient to break the rock apart and has less destructive power than does slow crystal growth, and therefore gives better fossil recovery.

Calcareous nannofossils Calcareous nannofossils (*see* **Fossil Plants:** Calcareous Algae) as a group are composed of several forms, such as coccoliths, nannoconids, and discoasters. They range in size between 0.25 and 30 μm, and, due to this very small size, the number of specimens in fine-grained pelagic sediments can often exceed 1 million cm ^{3}. Therefore, when collecting in the field, only a very small sample size (approximately 1 cm^3) is required. Contamination must be very carefully avoided in all stages of nannofossil preparation: the surface of the rock must be scraped clean, cleaned tools must be used to extract the sample, and the sample must be placed in a clean, labelled bag. In the laboratory, any glassware

must be cleaned with a 10% solution of hydrochloric acid and rinsed with distilled water.

Several methods, each of increasing complexity, can be used for extracting nannofossils from the surrounding matrix. The simplest may be done rapidly by making a smear slide. This is a small scraping of sediment onto a glass slide; this specimen is then diluted and mixed with a little distilled water. The slide is dried on a hotplate and finished by mounting a cover slip with a medium that has a refractive index between the two extremes of calcite, 1.484 and 1.658. Caedax or Cellosize and Elvacite have been used successfully. These smear slides are useful for reconnaissance work, but for more detailed analysis, other methods are available. The most commonly used technique is separation by centrifuge, which has replaced methods such as wet sieving, gravity settling, and elutriation. The object is to prepare slides that contain a high proportion of nannofossil to sediment while retaining the diversity of the original sample. To achieve this, the sample is broken down by crushing or shaking in distilled water buffered with sodium carbonate to a pH of around 8 to 8.5. This should then pass through a 45μm sieve. Often a water softener such as sodium hexametaphosphate is added to deflocculate the clays, but some authorities avoid this because it has a slightly corrosive effect on the fossils. Similarly, ultrasonic vibration is used by some to disaggregate the sample further, but this has been reported to damage some specimens. Many variations on the centrifuge speeds have been suggested, but a simple two-stage method involves initially centrifuging with a low speed to settle out particles coarser than 30 μm, retaining the finer material in suspension, then a second stage of rapidly centrifuging the liquid to concentrate the particles down to 3 μm. Speeds and times will vary according to the equipment, so the instrument manufacturer's charts should be consulted for calculating the correct variables.

Nannofossils may be examined either by light or electron microscopy. For light microscopy, a drop of the suspension is placed onto a slide and prepared as for the smear slide. A high-powered research-grade optical microscope with magnifications of 800× or 1200× (oil immersion lenses) is required. Observations are made either by using polarizing attachments and viewing in cross-polarized light or by using phase-contrast methods. Though the results obtained in this way are rapid, and often used in industrial applications, more detailed images may be obtained by employing electron microscopy. The transmission electron microscope (TEM) has been used to see the structural relationships of the elements of the nannofossils, but the preparation for the TEM requires making carbon replicas of the specimens, which is a painstaking and skilled operation and is rarely used now. A faster and more convenient method is to coat the specimens in a thin layer of gold and observe using scanning electron microscopy (SEM). Figure 1 shows examples of specimens using various observation methods.

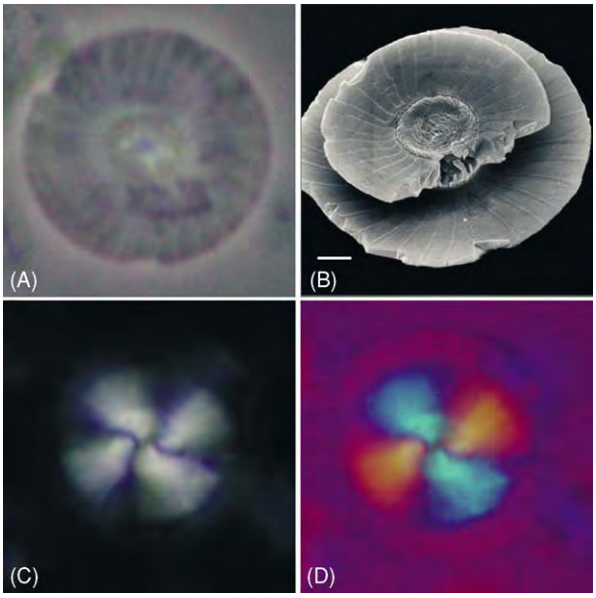


Figure 1 Calcareous nannofossils (*Calcidiscus*) visualized using different methods. (A) Phase contrast; (B) scanning electron microscopy; scale bar 1 micron; (C) crossed polarized light; (D) gypsum tint plate. Photographs by Dr J Young.

Acid-Insoluble Microfossils

Conodonts Conodonts (*see* **Microfossils:** Conodonts) are found in marine rock types as diverse as limestones, shales, marls, ironstones, and cherts, and the methods of extracting them from their matrix are just as varied. However, unlike the calcareous microfossil groups, the phosphatic nature of conodonts renders them suitable for some acid preparations, and a variety of acids have been used for this purpose. Although conodonts may be visible on bedding surfaces in the field, it is far more usual to employ spot sampling at intervals that have been accurately logged. A bulk sample size of 1–2 kg should be sufficient to yield many conodonts. Because conodonts are durable and quite resistant to weathering and erosion, care needs to be taken to avoid contamination of samples from material washed down slopes. Again, only fresh rock should be sampled. For all lithologies, samples are pretreated by cleaning the exterior of the bulk sample, oven drying, and then coarsely crushing to centimetre-sized pieces.

Carbonate-cemented rocks will respond to diluted acids and many methods have been tried. The use

of monochloroacetic acid, although giving a rapid reaction, has been shown to damage conodont elements. Similarly, the use of unbuffered acids, such as formic and acetic acids, can also cause dissolution. However, when 10% formic acid (HCOOH) is buffered with calcium carbonate and tricalcium carbonate to a pH of less than 3.6, it becomes a suitable digesting medium. Similarly, 10% glacial acetic acid (CH_3COOH), when buffered with sodium acetate or calcium acetate to a pH of 3.5, can be used quite effectively. The technique is similar with both acids and involves immersing the crushed rock into the acid in a fume cupboard and waiting for the effervescence to cease. Formic acid may require up to 1 day before the residue is washed with water and sieved, whereas acetic acid will require many regular changes of acid to dissolve the matrix, and this process can continue for several weeks. For other rock types, such as shales, methods similar to those used for calcareous microfossils are generally used. Once the residue has been sieved, conodonts may be further concentrated by either electromagnetic or heavy liquid separation.

Siliceous microfossils Several groups of microfossils have siliceous skeletons, including radiolaria, diatoms, silicoflagellates, and ebridians. These may occur in both clastic and carbonate rock types, and by virtue of their chemistry are resistant to some acids that may be used successfully to extract them. However, this is not always necessary, because diatoms may be found in unconsolidated sediments and can even form the majority of the rock, which is then termed a 'diatomite'. Radiolarians may also be numerous enough to form radiolarian earth. Various methods are used for loosely consolidated sediments, involving several stages: (1) removal of the carbonate component by using dilute acids, such as hydrochloric, formic, nitric, or acetic acids, (2) removal of organic matter using oxidizing agents, such as hydrogen peroxide or stronger nitric acid, and (3) dispersal of the clay component by using water softeners. With more indurated material, the strength of the acid and the times of digestion are increased. For extraction of radiolarians from mudstones, shales, and marls, a combination of 15% hydrogen peroxide and a degreasing agent such as MP 10 is effective. Very often radiolarians occur in chert beds, siliceous shales, siltstones, or mudstones, which contain silica of the same composition as the fossils. In these cases, 10% hydrofluoric acid (HF) has to be used with extreme caution to etch the radiolarians out of the rock.

The observation of siliceous fossils such as radiolaria is made more difficult because the refractive index of the shells is similar to that of the glass on which the specimens are mounted, rendering them invisible. Techniques to overcome this involve selecting a mounting medium with a high refractive index to contrast with that of the glass, and observing in transmitted light with dark-field illumination, or by coating the specimens with silver nitrate, thus allowing oblique incident light to be used. Larger specimens may be observed directly using a stereozoom microscope. The electron microscope is also used for studies requiring observation of the finest detail and is invaluable in taxonomic work.

Organic microfossils As a group, organic microfossils include pollen and spores, dinoflagellates, acritarchs (*see* **Microfossils:** Acritarchs), tasmanitids, Chitinozoa (*see* **Microfossils:** Chitinozoa), and scolecodonts. With any palynological methods, a clean working environment is essential to reduce the risks of sample contamination, because the specimens are only micrometres in size and are easily transferred between tools and vessels. Although methods vary widely according to the fossil group and the lithology being treated, a broad working plan would follow several stages, including cleaning, disaggregation and dispersal, chemical extraction, and concentration. Palynological preparations differ from other techniques in the chemical extraction stage, because the fossils are composed of sporopollenin or pseudochitin, which allows the use of strong acids; these methods should only be carried out in properly equipped laboratories. Initial disaggregation is achieved by cleaning, crushing, and soaking or boiling in strong detergents such as 'Quaternary O'. Other dispersal agents are used depending on the lithology, such as KOH, NaOH, or Na_3PO for peat; KOH for coals; or NaOH for shales.

Chemical extraction takes several stages to complete and may consist of demineralization and maceration, depending on the rock type. Demineralization is needed to remove the inorganic component of the rock, and different acids are used to achieve this. Carbonates are removed by treatment with HCl; silicates are removed using HF. If sulphates or sulphides are present, they are removed by treatment with HNO_3. Other phases, such as heavy minerals, are removed using heavy liquids. The second stage involves the removal of unwanted organic material; plant disintegration products in the residue may consist of cellulose, hemicellulose, and lignin. Cellulose is broken down by the process of acetylation using a reagent that is a mixture of concentrated sulphuric acid (H_2SO_4) and glacial acetic acid. The broken down cellulose is removed by centrifuging and washing in glacial acetic acid, and finally washing in distilled water. Acetylation also has the beneficial effect of darkening spores and dinoflagellate cysts, which

enhances their appearance for microscopical study. The lignin found in peats, though resistant to acetylation, is susceptible to oxidation and becomes soluble in a hot alkali solution. However, pollen and spores are also easily destroyed by the process of oxidation, so the effect on these grains has to be monitored and the process stopped before damage occurs. Many methods of oxidation may be used, depending on the lithology of the specimen. For example, silicates and argillaceous rocks will respond to 10% nitric acid to which is added 10% sodium hydroxide or Schulze's solution (a mixture of potassium chlorate and concentrated nitric acid); carbonates require fuming nitric acid (95%); peats can be treated with glacial acetic acid, sodium chlorate, and sulphuric acid; and oxidation of coals can be achieved by using Schulze's solution, fuming nitric acid, or hydrogen peroxide.

After these various chemical techniques have been completed, the final stages involve washing by centrifuge. The resulting palynomorphs in the residue are finally washed in 95% alcohol before staining with compounds such as Safranine-O, Malachite green, or Bismarck brown in 0.1% solutions, and are then transferred to a storage phial in distilled water with a mould inhibitor. To prepare slides for study, a small amount of the residue is pipetted onto a cover slip, and after the water has evaporated, a mounting medium, such as Elvacite 2044, Canada Balsam, or epoxy resin, is applied to the glass slide and the cover slip is placed on the slide and allowed to cure.

The usual method of observation is with a conventional biological microscope. However, modern methods of illustration and study involve the use of confocal laser scanning microscopy (CLSM), which gives high-resolution, blur-free images with a greater depth of focus than is possible with conventional microscopy. These images may also be reconstructed to give three-dimensional information and can be viewed as rotating computer animations.

Physical Separation and Concentration Methods

Additional techniques are sometimes required after the initial processing has been completed and the sample is thoroughly disaggregated. The purpose is to concentrate the sample to increase the ratio of microfossil specimens to rock particles such that time spent examining the residue is kept to a minimum.

Sieving

The principal aim of sieving the sample is to remove particles of finer or larger sizes, compared to the fossils of interest. Usually the finer particles are the deflocculated and dispersed clays; fragments larger than the microfossils may be stones, shelly material, or rock that has not broken down completely. The choice of sizes of sieves is often dependent on the fossil group being studied. Usual practice for the larger calcareous microfossils is to use a nest of sieves, of sizes ranging from (bottom to top) finest (75 μm), to medium (250 μm), to coarse (1 mm). If studying planktonic foraminifera, it is essential that the finest sieve is either 75 or 63 μm, because heterohelicids and similar-sized foraminifera will easily fall through coarser sizes. It is worth noting that sieves are designed to separate sedimentary particles that are equant in shape; however, many microfossils are elongated in one direction and may pass through a sieve lengthways. Therefore, the smallest microfossil dimension should be considered when choosing sieves to retain a certain element of the fauna. Occasionally microfossils can become stuck in the mesh, and are then liable to be transferred to another sample as a contaminant. In order to identify such errant specimens, the sieve is stained with a solution of methylene blue; once the surplus stain is washed off, any contaminating material takes the blue colour and is readily spotted in the residue.

Magnetic Separation

Rock particles in residues often contain a proportion of iron, and calcareous or phosphatic microfossils do not, thus it is possible to use differences in magnetic properties to separate the two fractions. For example, this technique is particularly successful in concentrating calcareous microfossils in a residue containing glauconite. This is achieved by using a Frantz Isodynamic Magnetic Separator, which consists of a very large electromagnet that can be tilted in two planes, a vessel into which the sample is placed, and a vibrating motor that moves the particles down a chute. The chute divides into two sections so that separate collecting buckets can receive the magnetic and non-magnetic grains.

Flotation

A common method for concentrating specimens in a residue is to use differences in the specific gravities of the fossil and the matrix. A suitable heavy liquid can then be used to separate the two components, such that one floats and one sinks. Each fraction can then be collected separately. Toxic liquids such as carbon tetrachloride, bromoform, and tetrabromoethane, which have been used widely in the past, have now been superseded by safer compounds, such as sodium

polytungstate, the density of which is altered by the addition of water. To separate conodonts, for example, a liquid with a specific gravity of about 2.75 is used; this allows calcite to float and the conodonts to sink. The specific gravity of pollen ranges from just over 1 to about 1.5, whereas the minerals encountered in pollen analyses have specific gravities ranging from 2.0 to over 3.0 (gypsum, 2.3; feldspar, 2.55; quartz, 2.6; calcite, 2.7; dolomite, 2.8). Therefore, a solution specific gravity between 1.56 and 2.0 should be used for pollen separation.

Picking

The final stage in any processing of the groups of large microfossil is that of picking the specimens. This is usually carried out by thinly spreading out the residue on a flat plate such that each grain can be easily seen, and then using a low-power binocular microscope at 10–20× magnification, scanning each grain to identify the microfossils. When a specimen is spotted, it is removed by using a fine, sable-hair brush of gauge 000 that has been wetted in distilled water. The specimen is transferred onto a numbered gridded slide that has been prepared with some adhesive. Gum tragacanth with a mould inhibitor has been widely used, but water-soluble glues such as Pritt Stick work just as well. Alignment and careful arrangement of the specimens on the slide aids in identification and counting.

See Also

Conservation of Geological Specimens. **Fossil Plants:** Calcareous Algae. **Microfossils:** Acritarchs; Chitinozoa; Conodonts; Foraminifera; Ostracoda; Palynology. **Sedimentary Environments:** Depositional Systems and Facies.

Further Reading

Aldridge RJ (1990) Extraction of microfossils. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology A Synthesis*, pp. 502 504. Oxford: Blackwell Scientific Publications.

Allman M and Lawrence DF (1972) *Geological Laboratory Techniques*. London: Blandford.

Austin RL (1987) *Conodonts: Investigative Techniques and Applications*. Chichester: Ellis Horwood Limited.

Brasier MD (1980) *Microfossils*. London: Allen and Unwin.

Feldman RM, Chapman RE, and Hannibal JT (1989) *Paleo techniques. The Paleontological Society Special Publication 4*. New Haven, CT: The Paleontological Society.

Green OR (2001) *A Manual of Practical Laboratory and Field Techniques in Palaeobiology*. Dordrecht: Kluwer Academic Publishers.

Hodgkinson RL (1991) Microfossil processing: a damage report. *Micropaleontology* 37(no. 3): 320 326.

Kummel B and Raup D (1965) *Handbook of Palaeontological Techniques*. San Francisco: WH Freeman and Co.

Taylor RJ and Hamilton GB (1982) Techniques. In: Lord AR (ed.) *A Stratigraphical Index of Calcareous Nannofossils*, pp. 11 15. Chichester: Ellis Horwood Limited.

MILANKOVITCH CYCLES

See **EARTH: Orbital Variation (Including Milankovitch Cycles)**

MILITARY GEOLOGY

E P F Rose, Royal Holloway, University of London, Egham, UK

Introduction

Throughout the nineteenth century, geology was widely perceived as a science with potential military applications. However, geologists were not operationally deployed upon the battlefield until the two twentieth-century world wars. Then, German forces made far greater use of military geologists than their British and American opponents, although the roles on both sides were similar. Military geologists guided the quarrying of aggregates and other mineral resources for tactical or strategic use; the development of secure water supplies adequate to support troop concentrations; aspects of military engineering, particularly with regard to fortification and the construction of underground facilities; and terrain analysis, notably

to facilitate or impede amphibious assault and cross-country military 'going'. By the dawn of the twenty-first century, small numbers of military geologists were established for routine use by armed forces, both European (notably British and German) and American.

Historical Background

Military applications of geology first became apparent in Napoleonic times. Indeed, the first general to take geologists as such on campaign was Bonaparte himself. The French army he led into Egypt in 1798 was accompanied by a civilian Commission of Sciences and Arts that included Déodat de Dolomieu (after whom the mineral dolomite was later named) (Figure 1) and several of Dolomieu's former geology students, recent graduates of the School of Mines in Paris – notably Louis Cordier (later immortalized by the mineral name cordierite), François-Michel de Rozière, and Victor Dupuy. However, their role was that of 'mineralogists' to support the army by exploring the geological resources of the country, rather than by contributing tactical or strategic advice. Karl von Raumer (1783–1865), from 1811 Professor of 'Mineralogy' at the University of Breslau (now Wroclaw in Poland), served as a staff officer in Prussia's 1813–14 war of independence from Napoleonic domination, but was used to communicate dispatches rather than appraise geology. The distinguished Swiss mining geologist, Johann Samuel Gruner (1766–1824) (also known as von Grouner), having moved to Bavaria in 1803, became a captain in a volunteer rifle battalion when Bavaria joined the alliance against Napoleon late in 1813. Postwar in 1820, seemingly at the instigation of the Bavarian Bureau of Military Topography, he combined his military and geological experience to write a memorandum on the relationship between geology and military science. Published posthumously in 1826, this is the earliest known work in its field.

Figure 1 Déodat de Dolomieu (1750 1801), in 1798, by André Dutertre. A former cavalry officer and Knight of Malta, Dolomieu taught geology at the School of Mines, Paris, before serving as a senior scientist with Napoleon's expeditionary forces in Egypt. From portrait supplement to Saintine, X B, Marcel JJ and Reybaud L (eds.) (1830 1836) *Histoire Scientifique et Militaire de l'Expédition Française en Égypte*. Paris: Dénain & Delamare. By permission of the British Library, London (Shelfmark1311.h.2). Also from Rose EPF (2004) Napoleon Bonaparte's Egyptian campaign of 1798: the first military operation assisted by geologists? *Geology Today* 20: 24 29, by permission of Blackwell Publishing Ltd.

In the UK, geological mapping was initially perceived as a military skill. From 1809 to 1814, military (Board of Ordnance) objectives and funding generated geological fieldwork by J. MacCulloch, and from 1814 to 1826 his geological mapping in Scotland. From 1826 to 1846, Royal Engineer officers (successively captains J.W. Pringle, J.E. Portlock, and H. James) pioneered government geological surveys in Ireland. The 'British' Geological Survey was founded in 1835 and sustained until 1845 under military (Board of Ordnance) auspices, and its earliest directors (H.T. De la Beche until 1855, R.I. Murchison until 1871 (*see* **Famous Geologists:** Murchison)) were both men who had received a military rather than a university education. The world's oldest geological society, the Geological Society of London founded in 1807, included amongst its earliest influential members those also active in the reserve army (Lieutenant G.B. Greenough) or the regular army (T.F. Colby, J.W. Pringle, and J.E. Portlock), and a veteran of the Napoleonic wars in receipt of military half-pay throughout his geological career (Lieutenant W. Lonsdale).

Because of its evident practical applications, geology was soon introduced into the curriculum at many military colleges. It was taught at Addiscombe in Surrey to officer cadets of the East India Company's army, by J. MacCulloch from 1819 to 1835 and by D.T. Ansted from 1845 until college closure in 1861. For the British army, J. Tennant lectured on geology to Royal Engineer and Royal Artillery cadets at the Royal Military Academy, Woolwich, from 1848 to 1868; T. Rupert Jones to cadets of the infantry and cavalry at the Royal Military College, Sandhurst, from 1858 to 1870; Jones also to young officers of all arms at the Staff College, Camberley,

until 1882; and A.H. Green at the School of Military Engineering, Chatham, from about 1888 to 1896. Jones (Figure 2) was employed full time as a military professor of geology, but the others were essentially visiting professors who concurrently held better known appointments elsewhere – Ansted and Tennant for a while at King's College London, Green at Oxford. The United States Military Academy at West Point began some geological teaching in about 1823. Publications by R.B. Smith in 1849, F.W. Hutton in 1862, and articles within a massive three-volume, *Aide-Mémoire to the Military Sciences*, by G.G. Lewis and others (first edition, 1846–52; second edition, 1853–62), published primarily for use by the Royal and the East India Company's Engineers, demonstrated quite specifically the importance of geology to the British military profession; however, military teaching of the subject declined sharply towards the end of the nineteenth century as perception of its practical value waned.

Professional geologists were not used operationally as such until World War I, and then primarily as a response to the near-static battlefield conditions on the Western Front in Belgium and northern France. The German army deployed the engineering geologist W. Kranz in 1914, and H. Philipp plus many more geologists in 1915. A German military geological service was constituted in 1916, and made use of some 250 geologists by the end of the war in 1918 when it was disbanded. In the British army, far fewer geologists were used: Lieutenant (later Captain) W.B.R. King from 1915 to guide the development of water supplies; Major (later Lieutenant-Colonel) T.W. Edgeworth David from 1916 to guide siting of mine tunnels and dugouts; and a few others, primarily in the Tunnelling Companies of the Engineer Corps. All returned to their former civilian life at the end of the war: King to the 'British' Geological Survey and then the academic staff of the University of Cambridge and later London, and David as Professor of Geology and Physical Geography to the University of Sydney in Australia. Led by Lieutenant-Colonel A.H. Brooks, nine (of potentially 18) geologists were assigned for service with the American Expeditionary Force in 1918, primarily from the United States Geological Survey, but the war ended before their full operational deployment.

The German army was quick to reinstate a military geological organization prior to World War II, from 1937. Military geology textbooks were generated by J. Wilser in 1921, E. Wasmund in 1937, K. von Bülow and others, C. Mordziol, and W. Kranz, all in

Figure 2 Professors of the Staff College, Camberley, in 1874. The geologist, T. Rupert Jones, is seated at the left rear. Reproduced with permission from Rose EPF, Hausler H, and Willig D (2000). In: Rose EPF and Nathanail CP (eds.) Comparison of British and German applications of geology in world war, pp. 107 140. *Geology and Warfare: Examples of the Influence of Terrain and Geologists on Military Operations*. London: Geological Society. Courtesy of the Joint Services Command and Staff College, Swindon, and the Geological Society, London.

1938. Military geologists were deployed with troops invading Poland in 1939, and France and the Low Countries in 1940. By late 1941, there were 32 teams of military geologists providing support to the German army throughout its area of occupation, the number of terms increasing to 40 by November 1943. Making use eventually of some 400 geologists, largely recruited or conscripted from university staffs, the German army thus developed the largest organization ever to provide military applications of geoscience in wartime. Additionally, smaller but still significant numbers of geologists served with the German air force, navy, Waffen-SS, or the paramilitary construction agency 'Organisation Todt'.

In contrast, the British army again used very few geologists as such: Major (later Lieutenant-Colonel) W.B.R. King from 1939 to 1943 and his Cambridge protégé Captain (later Major) F.W. Shotton from 1940 to 1945; this group was increased to perhaps a dozen in various roles by 1944, largely in preparation for the Allied invasion of Normandy, and many of them formerly academic geologists from British universities. America entered the war in December 1941, its forces soon supported by a Military Geology Unit at the United States Geological Survey. With a wartime roster of 88 geologists, 11 soil scientists, 15 other specialists, and 43 support staff, it produced 313 studies – including 140 major terrain folios, 42 other major reports, and 131 minor investigations – in total containing about 5000 maps, 4000 photographs and figures, 2500 large tables, and 140 terrain diagrams.

After the war, the British army maintained continuity in geological expertise through a small group of officers in the reserve army (the Territorial Army or, from 1953 to 1967, the Army Emergency Reserve) for peace-time engineering projects and operational planning, and active service in times of crisis, whether occasioned by military conflict or humanitarian need. German capability in military geology was abolished at the end of the war, but later re-established within the Bundeswehr. By the end of the Cold War in 1990, the German army employed more than 20 full-time geologists – as civilians but with commitment as reserve army officers. In the USA, geological support for the armed forces continued to be based on the United States Geological Survey which, between 1945 and 1972, briefly used about 150 scientists to compile terrain intelligence on a global scale.

Quarrying, Aggregates, and Mineral Resources

Quarrying became important to the German army in World War I. From 1915 especially, in regions such as Alsace, German military geologists produced maps showing sites for quarrying of construction materials – building stone, sand, and aggregates – maps that were used to guide the construction of fortification systems by the Engineer Corps. Other maps plotted reserves of economic significance, such as coal and metalliferous ores. A significant proportion of German military geologist tasks (Figure 3) were related to the winning of raw materials in occupied land. The Allies, fighting to defend their own terrain, made use of local civilian rather than military geological expertise for such tasks.

Raw materials also formed a significant proportion (Figure 3) of German military geologist tasks in World War II. Fortification systems, such as the West Wall bordering Germany and France, and the Atlantic Wall that marked the western limit of German-occupied Europe, consumed vast quantities of aggregates for concrete in 1938–40 and 1940–44, respectively. The Atlantic Wall alone included some 15 000 bunker and casemate complexes with outer walls of reinforced concrete at least 2 m thick, eventually absorbing about 17.6 million tonnes of concrete.

Knowledge of potential sources of raw materials was also required for operational planning. For Operation Sealion, the German invasion of England scheduled for September 1940, planning included the preparation of maps by military geologists at scales of 1 : 100 000 and 1 : 50 000 to show the positions of all working and disused quarries in south-east England, the nature of material produced (e.g., limestone, sand, and clay), and its potential engineering use. Such maps, their data derived mainly from published topographical maps plus geological maps and memoirs, were widely used by the German army in preparation for attack.

In defence, the Allies relied largely on civilian agencies, such as the Geological Survey of Great Britain, for geological advice. When counter-attacking, however, military geologists were used. To support the Allied liberation of Normandy in June 1944, the (British) Geographical Section of the General Staff reprinted all the French 1 : 80 000-scale geological maps of the region. These, plus military geologist advice, guided the operation of units, such as the Quarry Group, Royal Engineers. During the 11 month campaign in north-west Europe, this generated in total some 1 376 000 m^3 of crushed rock from 49 different quarry sites (Figure 4), successively in France (Figure 5), Belgium, and Germany – aggregates used for road repair, widening, and construction; hard standings for ammunition and stores depots; vehicle and gun parks; and airfield runways. The mobile, mechanized conflict of World War II gave the maintenance and construction of roads and airfields a high priority.

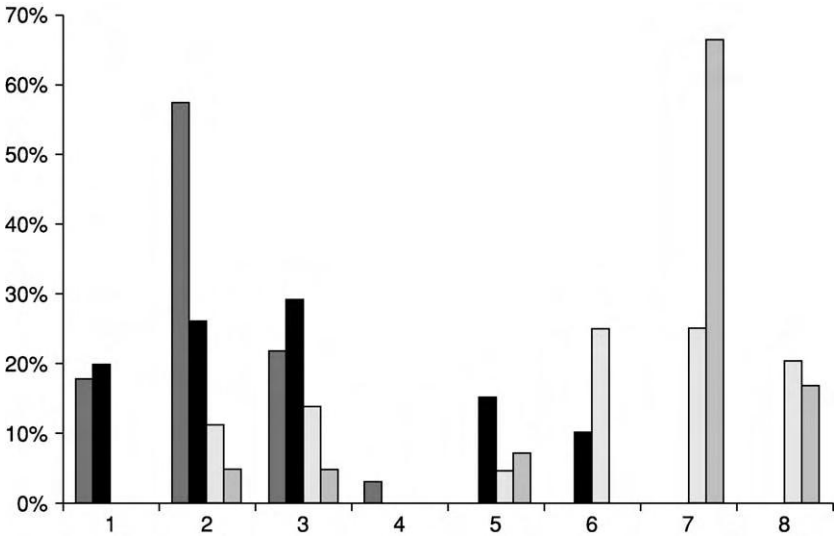


Figure 3 Categories of major tasks undertaken by German military geologists, showing percentage of tasks calculated for each of four time periods to demonstrate a change in focus from World War I to the end of the twentieth century. ■, World War I (1914 18), based on 215 tasks undertaken by the military geology group directed by C. Mordziol between February and November 1917; ■, World War II (1939 45), based on 640 tasks; □, Cold War (1947 89), based on tasks for III (German) Corps between April 1971 and March 1972; ▨, late twentieth century, based on 42 main tasks undertaken by III (German) Corps in 1990. 1, Raw materials; 2, water supply/drainage; 3, infrastructure; 4, general and regional geology; 5, trafficability/exercises; 6, obstacles/barriers; 7, environmental protection; 8, other tasks. Environmental protection features more highly in post 1970 work for III Corps than for other German corps. Data courtesy of D. Willig, derived from Rose EPF, Hausler H, and Willig D (2000). Comparison of British and German applications of geology in world war, pp. 107 140. In: Rose EPF and Nathanail CP (eds.) *Geology and Warfare: Examples of the Influence of Terrain and Geologists on Military Operations*. London: Geological Society.

In the postwar years, military quarrying has waned considerably in importance (Figure 3), but the expertise has been maintained for occasional operational use. Thus, following the Falkland Islands' conflict between the UK and Argentina in 1982, aggregate for repair and extension of the Islands' principal airport was provided from quarries operated by Royal Engineers, supported by advice from a military geologist.

Water Supply

Potable water has long been recognized as a vital military resource. Without a secure water supply, no defended locality, however well fortified and garrisoned, can withstand siege for long. Moreover, troops massed for campaign need, in total, copious supplies of drinking water.

Problems of water supply and drainage related to the army's field positions were, from late 1914, amongst the earliest German military geologist tasks assigned on the Western Front in World War I. Overall, they represented by far the greatest number of tasks (Figure 3). From 1915, German military geologists were preparing a variety of water supply maps to provide information in a manner readily understood by combat troops, especially to guide the siting of new boreholes and dams for ponding small reservoirs.

Also in 1915, the British army deployed its first operational military geologist, 'Bill' King, as a hydrogeologist. Most of his early work consisted of compiling water supply maps. Later work involved the selection of sites for boreholes. Over 470 borings for water were emplaced behind the Western Front by the British army during the war, and King supervised and interpreted many of these. Improved techniques for recording and mapping hydrogeological data were developed as a military expedient at this time.

In World War II, geological appraisals relating to water supply were still important to the German army, if less so than in World War I (Figure 3) because conflict was relatively more mobile. Military water supply maps ranged from the very detailed, large-scale maps prepared for some areas under long-term occupation, to the simpler, smaller scale maps required for operational use. Thus, in the 2 month planning period for Operation Sealion, German military geologists gave a high priority to preparing water supply maps at scales of 1 : 50 000, 1 : 100 000, and 1 : 250 000. These depicted the major lithologies which cropped out in south-east England, typically grouped into three categories according to

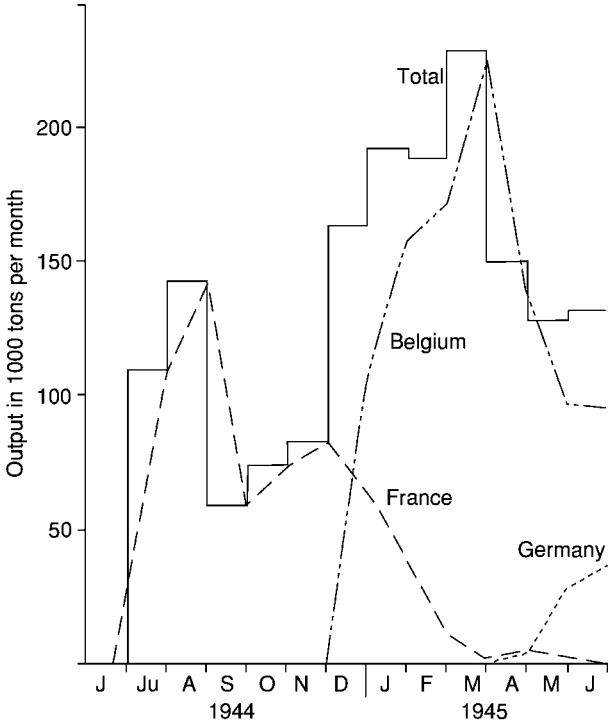


Figure 4 Monthly 1000 ton (1020 tonne) quantities of stone produced or distributed by the Quarry Group, Royal Engineers, from June 1944 to June 1945, in total (full line) and progressively from sources in France, Belgium, and finally Germany. Redrawn, with permission from Blackwell Publishing Ltd., from Rose EPF and Pareyn C (1995) Geology and the liberation of Normandy. *Geology Today* 11: 58 63.

suitability for the emplacement of shallow wells: suitable (overprinted with crossed diagonal lines), unsuitable (no overprint), and only partly suitable (with unidirectional diagonal line overprint) (Figure 6).

In the British army, 'Bill' King was deployed once more to France as a hydrogeologist to serve with the British Expeditionary Force, from 1939 until its evacuation in 1940. Dowsers rather than geologists were initially used to guide well drilling in North Africa and the Middle East, but with little success. There was a marked improvement in 1941 when 'Fred' Shotton was deployed as a military hydrogeologist, with support from 42nd Geological Section of the South African Engineer Corps, whose members were experienced in the use of electrical resistivity and other geophysical techniques employed in the exploration for groundwater in arid zones. From 1943, first King and then Shotton were tasked with geological preparations for the liberation of Normandy, which included the complication of water supply maps, amongst others (principally for beach landings, and airfield construction). Following the D-Day landings, well drilling in Normandy to develop groundwater, largely by borings through Middle Jurassic limestones, was carried out primarily under military geological supervision.

Postwar, the British army has retained a well drilling capability on a near-permanent basis. Part-time reserve army geologists have been used to guide its operation

Figure 5 Military quarrying of Middle Jurassic limestone near Caen, Normandy, August 1944. © The Imperial War Museum, London: photograph CL811.

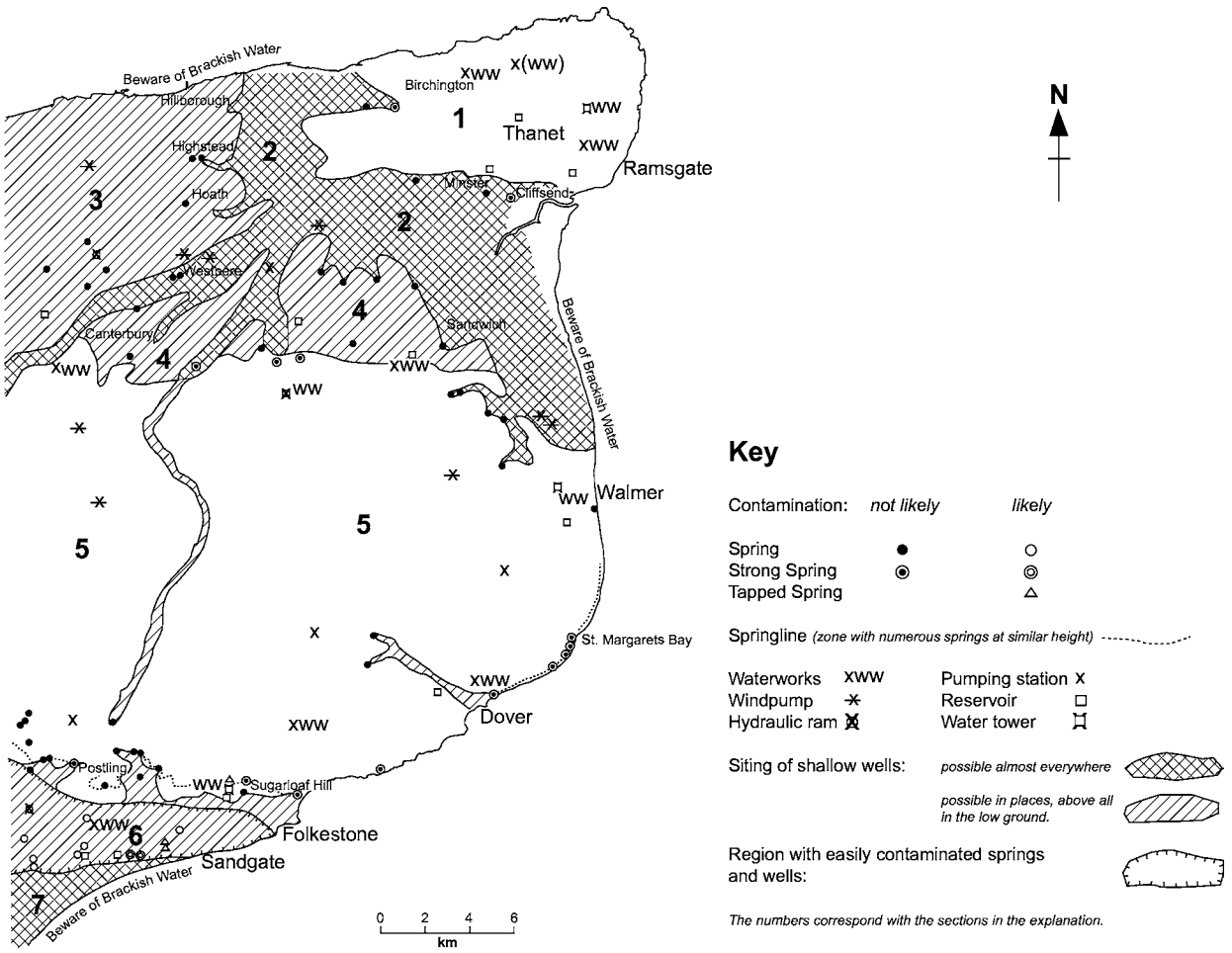


Figure 6 Map of east Kent, showing the seven water supply regions depicted on the 1 : 50 000 scale water supply map sheet 117 prepared by German military geologists to facilitate planning for the invasion of England in September 1940. Key translated from German, but retaining the symbols adopted for the original map. Redrawn with permission from Rose EPF, Mather JD, and Willig D (2002) German hydrogeological maps prepared for Operation 'Sealion': the proposed invasion of England in 1940. *Proceedings of the Geologists' Association* 113: 363 379, Fig. 4. © 2002 The Geologists' Association, London.

both in the UK and overseas (Figure 7). The German army has also used its military geologists in a water supply role (Figure 3), notably under United Nations or North Atlantic Treaty Organization auspices, in countries such as Somalia, Bosnia-Herzegovina, Kosovo, and Afghanistan.

Engineering Geology

Although tunnelling to emplace mines is one of the oldest applications of engineering to the science of war, during World War I it was employed on a vast and unprecedented scale on the Western Front. Beginning in 1915, it reached a peak in 1916 with some 25 000 Allied troops actively engaged in tunnelling, faced by a comparable number of Germans, and nearly 1500 mines fired on both sides in the process of more or less continuous underground warfare. Through test boring and collation of geological data, including the compilation of some of the first water table maps to be prepared in Europe, Edgeworth David provided detailed information on the extent and depth of strata most suitable for mining by the British Expeditionary Force (Figure 8).

Later, as from 1917 the front line was held by firepower rather than manpower, troops were deployed in depth and protected from artillery and aerial bombardment by dugouts (underground shelters), subways (connecting tunnels) and other subsurface excavations. To guide their siting, David prepared specialist geotechnical maps (Figure 9), illustrating the relative suitability of ground for dugout construction – credited to be amongst the first environmental/engineering geology maps ever published. German military geologists were used on similar 'infrastructure' tasks (Figure 3) related to the construction of field positions,

Figure 7 Royal Engineers boring for water in the South Arabian Federation in 1964. Military geologists from the British reserve army were used to guide such deployments during a time of intermittent conflict. © Crown Copyright/MOD. Reproduced with the permission of the Controller of Her Majesty's Stationery Office, and courtesy of the Royal Engineers Library, Chatham.

especially trenches, mines, and dugouts (to minimize time, labour, and materials, and the risk of collapse or water inflow), and site selection for standing camps, munition dumps, heavy gun positions, and airfields.

On demobilization, there were many more German than Allied geologists who returned to civilian life with practical experience in engineering geology. From them developed the expertise more fully deployed in World War II (Figure 3). The German army and navy were noted for using geological guidance when planning and constructing major fieldworks, whether fortifications, underground installations and factories, airfields, military bases, or command HQ facilities. In addition to the massive fortifications of the West Wall and Atlantic Wall, examples include the successful construction of bomb-proof submarine pens on the coast of Norway at sites with considerable geotechnical problems, such as the instability of beach sands and weak 'quick' clays, and active residual stresses and associated rebound-relief structures in the rock formed as a consequence of the region's glacial and tectonic history. Moreover, the Germans successfully located many military installations and critical manufacturing plants underground throughout Europe during the war.

The British seldom used military geologists for such tasks, except on Gibraltar, where a military geologist was briefly available during 1943 to guide some of the extensive tunnelling work undertaken to enhance the rocky peninsula as a fortress.

When America entered the war, its military engineers had to operate in terrain which varied from the

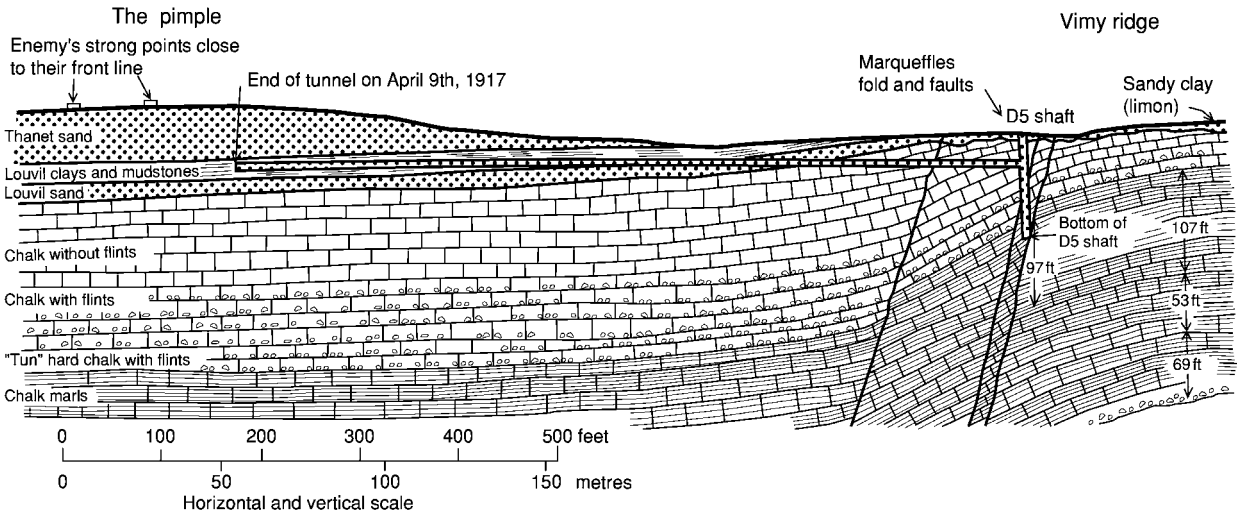


Figure 8 Geological cross section east west through Vimy Ridge, France, showing a mine tunnel excavated forward from the British lines, mostly in Louvil clays, which provided the quiet, dry, and stable conditions ideal for tunnelling. Redrawn, with permission from Blackwell Publishing Ltd., from Rosenbaum MS and Rose EPF (1992) Geology and military tunnels. *Geology Today* 8: 92 98.

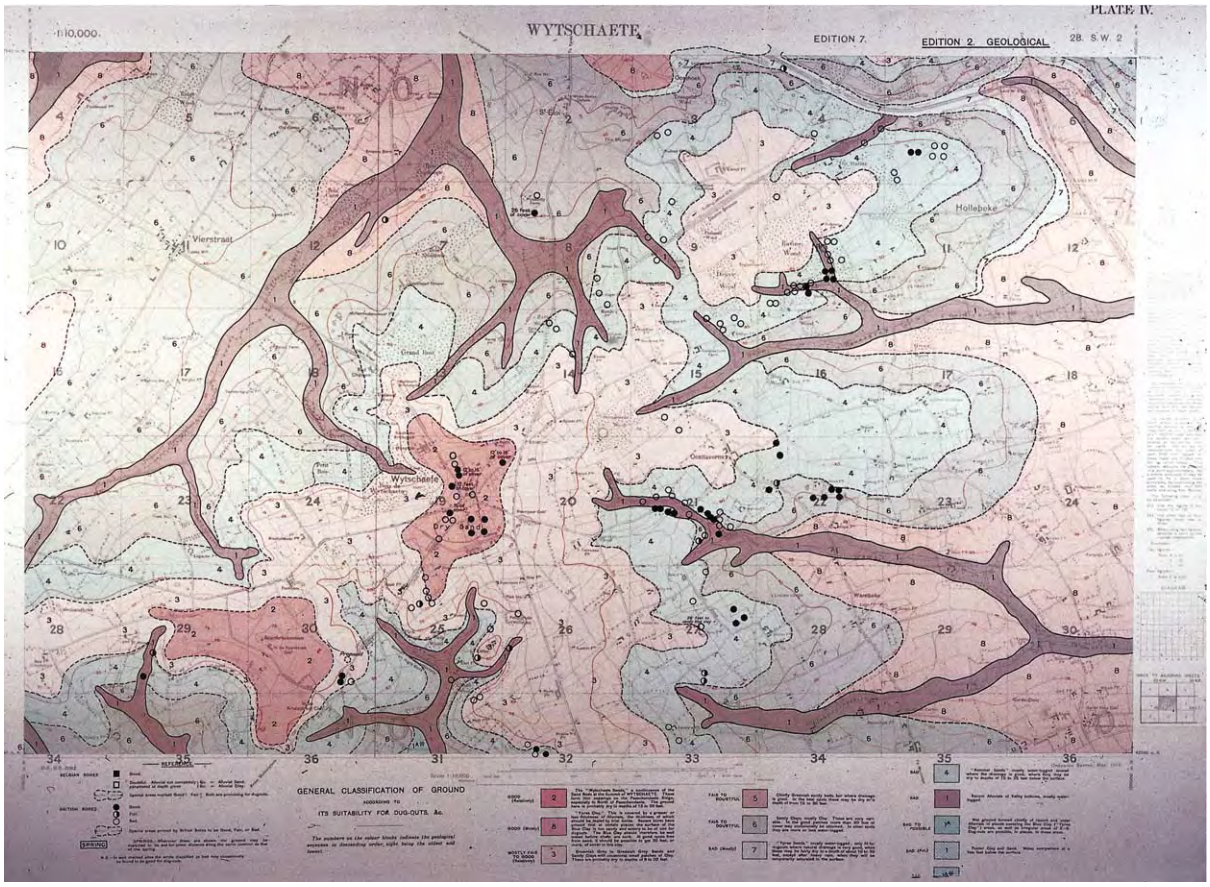


Figure 9 'Geological' map of Wytschaete, Belgium, originally at a scale of 1 : 10 000, showing the relative suitability of the ground for the construction of dugouts. The redder the formation shown on the map, the drier and more suitable for dugout construction; the bluer the formation, the wetter and less suitable. Reproduced with permission from the Ordnance Survey map of May 1918.

frozen Arctic to the jungles of the tropics. In northern regions, permafrost was a problem. In the island region of the South Pacific, terraced coral reefs and associated soft to weathered clay-mud and sand provided a very different challenge, because of physical variability and rapid variations in rock strength. Postwar, military engineering geology continued to be important to America as its role as the global superpower developed; however, its importance in European priorities declined sharply (Figure 3) as during peacetime major engineering works for military clients were largely undertaken by civilian agencies or contractors rather than by uniformed military engineers.

Terrain Analysis

In the nineteenth century, German language authors showed early appreciation of the relationships between geology, terrain, and warfare, but terrain analysis for military operations proved to be of relatively minor importance in World War I, at least in the near-static battlefield conditions of the Western Front. Studies of regional geology formed a very minor proportion of German military geological tasks (Figure 3), and the same was true of tasks carried out by the much smaller number of Allied military geologists.

In the more mobile conflict of World War II, the situation was different. A significant proportion of tasks undertaken by German military geologists related to the effect of terrain on the off-road trafficability of ground for military vehicles of different types, tracked or wheeled, light or heavy, and on the avoidance in attack or enhancement in defence of natural obstacles to troop movements (Figure 3). For the invasion of south-east England scheduled for September 1940, German military geologists prepared coastal maps at a scale of 1 : 50 000, which categorized beaches and cliffs in terms of suitability for amphibious assault, based on features related to both geomorphology and geology. A start was made on construction/engineering maps of inland areas, showing the distribution of major rock types, with assessment of the workability (by hand tools

to explosives), stability of excavations (with or without gabions), and permeability to groundwater (Figure 10). Another series was started to show variations in off-road trafficability, i.e., maps from which the rate of military 'going' could be predicted.

Later in the war, the German army developed map-making roles for all major combat areas, and the German navy developed specialist geotechnical maps for coastal regions to facilitate defence from Allied assault. Around 1942, a marine geographical unit was established by the Naval High Command with the primary objective of preparing specialist maps of selected coastal areas to facilitate defensive planning and fortification. The unit included a nucleus of professional geographers and at least one geologist. Also in 1942, a research detachment for special duties was created to provide terrain mapping and evaluation for the Armed Forces High Command. Initially, this was tasked with investigating the passability of the central Libyan Desert for troops moving from the south. Later, the scope of its work widened, and specialist maps were generated for many areas, at scales ranging from 1 : 50 000 to 1 : 500 000, by teams which included one or more specialists in geography, geology, plant ecology, meteorology, soil science, cartography, and photogrammetry, working in cooperation. This multidisciplinary work generated some 36 terrain evaluation maps between September 1943 and February 1945, covering areas from northern Finland to the mouth of the River Dnieper, and from north-west Germany to Greece. These were complemented by specialist trafficability maps of the Caucasus, the Pyrenees, and the eastern Alps, generated largely in 1942–43 by the German army's military geographical service. 'Military geology' maps were prepared by teams of German army geologists for some areas, e.g., to guide fortification of the Channel Islands (Figure 11).

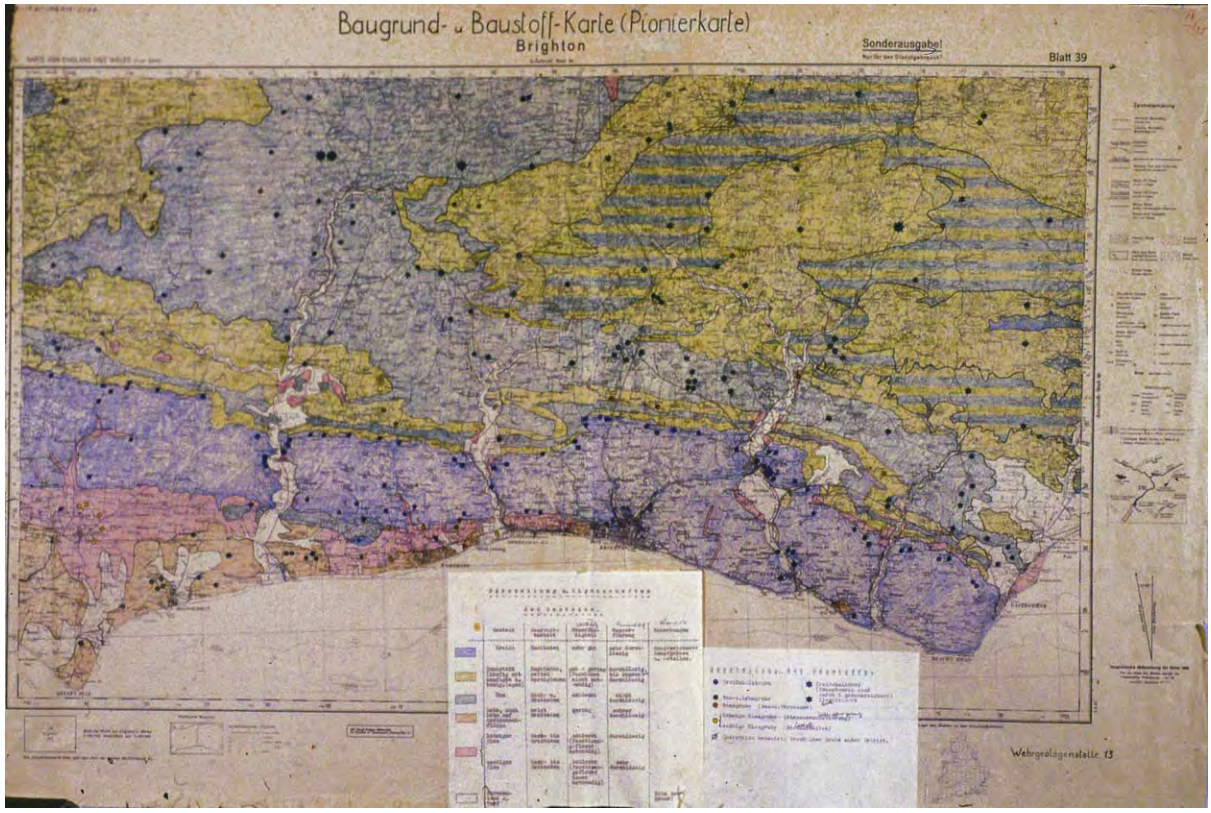


Figure 10 Building terrain and construction materials (engineering) map of the Brighton area of southern England, prepared at a scale of 1 : 100 000 by German military geologists to assist planning for the invasion scheduled for September 1940. Quarry sites are plotted on to a simplified geological map, and the key classifies the major rock types in terms of engineering properties. Reproduced with permission from Baugrund und Baustoff Karte, Brighton [Cartographic Record]; #23 2; German Language Geologic Maps, Record of the Geological Survey, Record Group 57; National Archives at College Park, College Park, MD, USA. Also by permission from Rose EPF and Willig D (2004). Specialist maps prepared by German military geologists for Operation Sealion: the invasion of England scheduled for September 1940. *The Cartographic Journal* 41: 13 35.

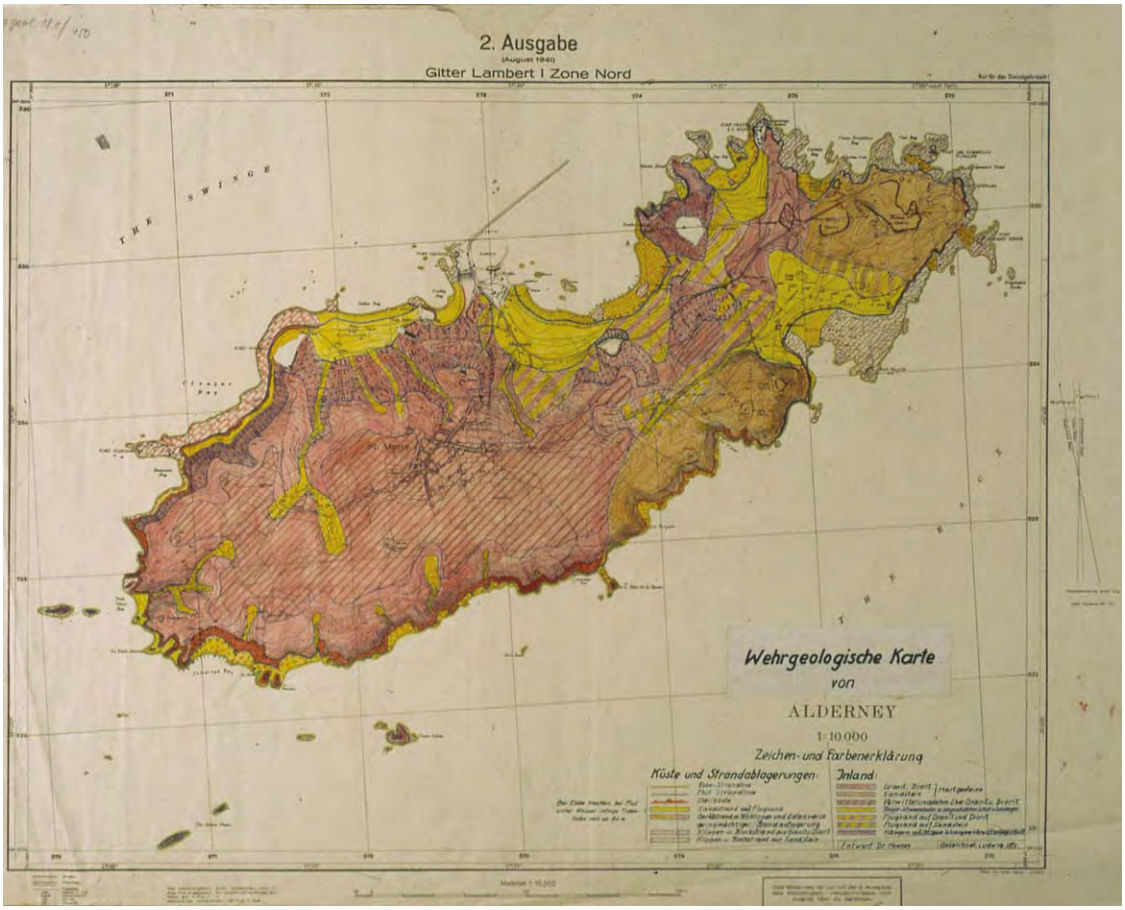


Figure 11 Military geology map of Alderney, one of the British Channel Islands occupied by German troops during World War II, compiled by German military geologists in 1942. The key provides an appraisal of rock type and geomorphology for both coastal and inland areas in terms that would facilitate fortification and troop deployments against British counterattack. Reproduced from Wehrgeologische Karte von Alderney [Cartographic Record]; #20 Great Britain Channel Is.; German Language Geologic Maps, Records of the Geological Survey, Record Group 57; National Archives at College Park, College Park, MD, USA.

Allied geologists also helped to generate specialist maps as the war progressed, the British especially to facilitate operations in Normandy in 1944 (Figure 12). Additionally, late in the war, cross-country movement (CCM) maps were developed by the Allies, especially the Americans, by multidisciplinary effort.

The preparation of such specialist maps, including those for areas without ground access, became a focus of interdisciplinary military activity during the Cold War of 1947–89. During this time, the Military Geology Branch of the United States Geological Survey focused research on strategic studies, areas of the Pacific region formerly occupied by the Japanese, Alaska and other areas of permafrost, military engineering geology and CCM maps for Germany, suitability of arid lands both inside and outside the USA for airfield construction, special intelligence, and studies to assist interpretation of global seismic signals for the Nuclear-Test Detection Program.

Conclusion

During the nineteenth century, there was a closer relationship between the military and some of the founding fathers of the science of geology than is commonly realized. In Britain, in particular, military objectives and funding generated initiatives both in the teaching of geology and in government-sponsored geological mapping. However, such initiatives waned through the second half of the century.

During the twentieth century, rather than teach the basic elements of geology to army or at least engineer officers in general, Britain, Germany, and the USA all incorporated professional geologists into their armed forces. During World Wars I and II, the role of geologists was partly to gather, collate, and interpret geological information for particular military use, and partly to guide site investigation and well drilling. Much military geology was merely

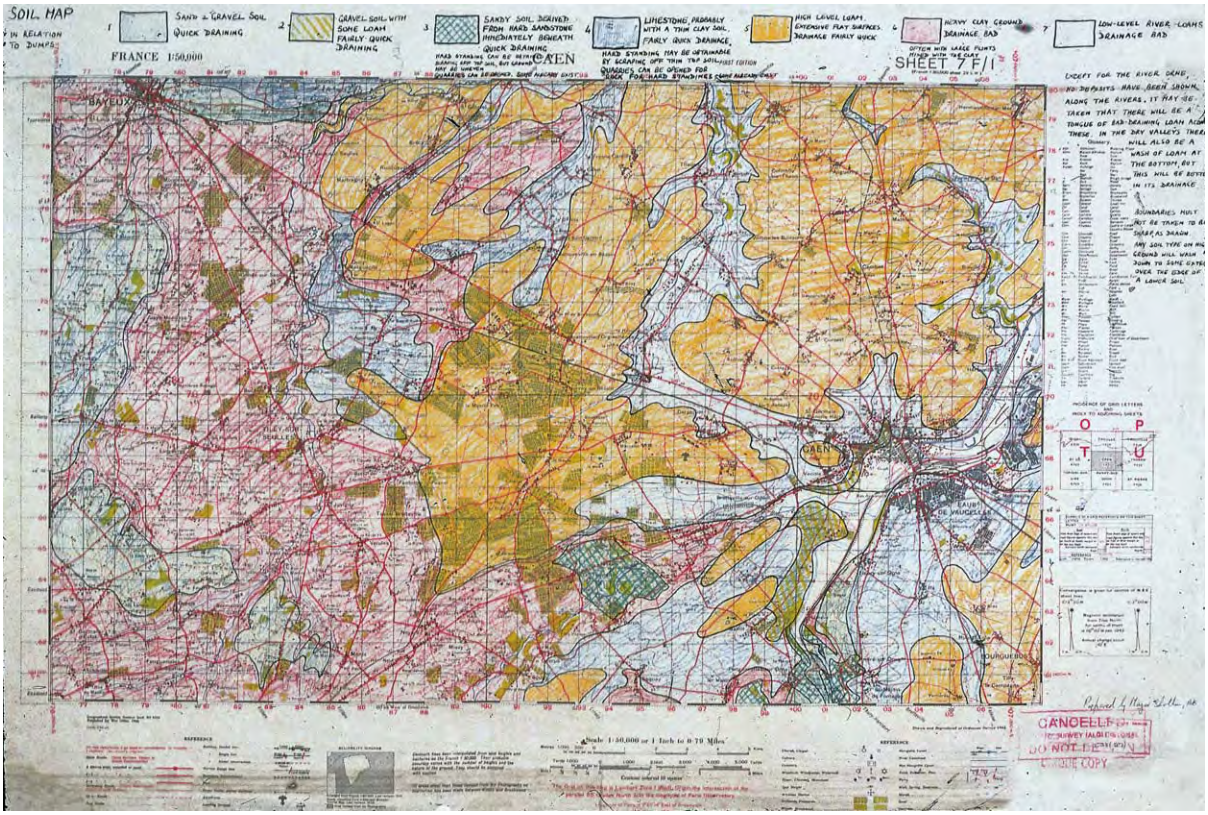


Figure 12 Soil map in relation to dumps, Caen: one of the specialist maps prepared by British military geologists in support of the 1944 campaign in Normandy. This 1 : 50 000 scale topographical map was hand coloured by Major F. W. Shotton to distinguish areas according to seven categories, given (from left to right) in the key at the top: 1, sand and gravel soil, quick draining; 2, gravel soil with some loam, fairly quick draining; 3, sandy soil derived from hard sandstone immediately beneath, quick draining; hard standing can be obtained by scraping off top soil but ground may be uneven; quarries can be opened, some already exist; 4, limestone, probably with a thin clay soil, fairly quick drainage; hard standing may be obtained by scraping off thin top soil; quarries can be opened for rock for hard standings, some already exist; 5, high level loam, extensive flat surfaces, drainage fairly quick; 6, heavy clay ground, drainage bad; often with large flints mixed with the clay; 7, low level river loams, drainage bad; peat marsh along coast at Meuvaines. Donated by the Ministry of Defence to British Geological Survey. © NERC. See also Rose EPF and Pareyn C (2003) *Geology of the D Day Landings in Normandy, 1944, Geologists' Association Guide 64*. London: Geologists' Association.

hydrogeology or engineering geology constrained by military objectives and urgency. Additionally, the need to communicate complex data to non-geologists under operational conditions led to the development of a wide range of specialist engineering geological maps to depict ground conditions in order to guide specific military activities. Examples include: water supply maps to guide the siting of boreholes and the abstraction of potable groundwater; geotechnical maps to guide the excavation of mines and dugouts; resource maps to guide the quarrying of materials for tactical or strategic use; 'diggability' maps to guide the siting of trench systems and anti-tank ditches; and CCM (trafficability or 'going') maps to indicate the potential effect of terrain and varying climatic conditions on the off-road movement of troops and vehicles. Some but not all of these map categories were derived from, or subsequently gave rise to, similar non-military specialist maps – for use in planning environmental or civil engineering works. Some depicted very complex data, but the maps most popular with military commanders, irrespective of nationality, were commonly those that depicted ground in terms of only three categories for a specific military activity: 'go' (suitable); 'no go' (unsuitable); and 'slow go' (intermediate or needing further qualification) (e.g., **Figures 6** and **9**).

By the start of the twenty-first century, in addition to these traditional roles, there was a trend for military geologists in peacetime to be used to protect the environment rather than the state, particularly in Germany (**Figure 3**) and the USA. In the Western world, large areas of land are used for military

training, and there is a continuing need to monitor and control the use of this land for a number of reasons: to protect soil and aquifers from significant contamination by explosive residues, propellant combustion products, and fuel spillage; to minimize scarring or erosion of soil pavements by off-road use of vehicles and foot traffic; and to assess the effects of vehicular compaction of soils in these areas, with its implications for permeability, rainfall runoff, and consequent patterns of erosion. Remote sensing, coupled with geological and geographical expertise, has facilitated the preparation of a wide range of specialist maps. It has also been used to predict sites of clandestine tunnels and underground facilities; cave systems potentially adapted for terrorist use; and areas in which bombing might artificially induce landslides and so deny road communication to an enemy. Geophysical techniques have been used to supplement geological appraisals in the detection and location of underground facilities potentially of use for espionage or as a means to avoid arms control inspections. Geologists and the military thus continue to share an interest in a variety of ground properties.

See Also

Engineering Geology: Aspects of Earthquakes; Geological Maps; Geophysics. **Environmental Geology**. **Famous Geologists:** Murchison. **Geological Engineering**. **Geological Field Mapping**. **Geological Surveys**. **Geology, The Profession**. **History of Geology From 1780 To 1835**. **History of Geology From 1835 To 1900**. **History of Geology From 1900 To 1962**. **Remote Sensing:** Active Sensors; GIS; Passive Sensors. **Sedimentary Rocks:** Dolomites.

Further Reading

Caldwell DR, Ehlen J, and Harmon RS (eds.) (2004) *Studies in Military Geology and Geography*. Dordrecht: Kluwer Academic Publishers.

Doyle P (1998) *Geology of the Western Front, 1914 1918, Geologists' Association Guide 61*. London: Geologists' Association.

Doyle P and Bennett MR (eds.) (2002) *Fields of Battle: Terrain in Military History*. Dordrecht: Kluwer Academic Publishers.

Ehlen J and Harmon RS (eds.) (2001) *The Environmental Legacy of Military Operations, Reviews in Engineering Geology 14*. Boulder, CO: Geological Society of America.

Häusler H (1995) Die Wehrgeologie im Rahmen der Deutschen Wehrmacht und Kriegswirtschaft. *Informationen des Militärischen Geo Dienstes, Vienna* 47: 1 155; 48:1 119.

Rose EPF (1996) Geologists and the army in nineteenth century Britain: a scientific and educational symbiosis? *Proceedings of the Geologists' Association* 107: 129 141.

Rose EPF (2004) The contribution of geologists to the development of emergency groundwater supplies by the British army. pp. 159 182. In: Mather JD (ed.) *200 Years of British Hydrogeology, Geological Society Special Publication 225*. London: Geological Society.

Rose EPF and Hughes NF (1993) Sapper geology. *Royal Engineers Journal* 107: 27 33, 173 181, 306 316.

Rose EPF and Nathanail CP (eds.) (2000) *Geology and Warfare: Examples of the Influence of Terrain and Geologists on Military Operations*. London: Geological Society.

Rose EPF and Pareyn C (2003) *Geology of the D Day Landings in Normandy, 1944, Geologists' Association Guide 64*. London: Geologists' Association.

Rose EPF and Rosenbaum MS (1993) British military geologists. *Proceedings of the Geologists' Association* 104: 41 49, 95 108.

Rose EPF, Mather JD, and Willig D (2002) German hydrogeological maps prepared for Operation 'Sealion': the proposed invasion of England in 1940. *Proceedings of the Geologists' Association* 113: 363 379.

Underwood JR Jr and Guth PL (eds.) (1998) *Military Geology in War and Peace, Reviews in Engineering Geology 13*. Boulder, CO: Geological Society of America.

von Bülow K, Kranz W, and Sonne E (1938) *Wehrgeologie*. Leipzig: Quelle and Meyer.

Willig D (1999 2003) Entwicklung der Wehrgeologie: Aufgabenspektrum und Beispiele. *Fachliche Mitteilungen des Amtes für Wehrgeophysik, Traben Trarbach* 225: 1 116; 226: 1 82; 227: 1 77.

MINERAL DEPOSITS AND THEIR GENESIS

G R Davis, Imperial College London, London, UK

The Nature of Mineral Deposits

The modern industrialized world makes use of an astonishing range of rocks and minerals, either for direct use or for processing to other useful forms such as metals. Mineral deposits are formed in places where natural processes have concentrated certain minerals, and some of these deposits are rich enough to be exploited as economic deposits. New rocks (including mineral deposits) are generated through weathering and sedimentation, and through volcanic activity and igneous intrusion, and existing rocks (including mineral deposits) are metamorphosed by tectonic activity and changes in their pressure/temperature environment within the Earth's crust. The formation of mineral deposits should be expected to be part of this general scheme of petrogenesis, bodies of rock where valuable minerals have become concentrated by processes active in their crustal environment over long periods of time. Many mineral deposits are simply expressions of peak geological conditions within more widely dispersed and less intense mineralization. For instance, the thin black Kupferschiefer shale of the large central European Permian basin is typically mineralized with copper sulphides. In only a few places, however, does mineralization reach sufficient concentration and extent to be called a deposit, and in fewer still, such as Lubin, is it an economic mineral deposit. Once seen as geological accidents, mineral deposits are now widely regarded as a normal and revealing part of the overall geological picture. This transformation was illustrated in 1972 by the appearance of a major textbook entitled Ore Petrology (Stanton) alongside classical texts on Igneous Petrology and Sedimentary Petrology. The processes by which anomalously high concentrations are achieved in mineral deposits, and particularly metalliferous deposits, are far less well understood than those operating to form anomalously high, but more abundant and familiar concentrations of common minerals such as SiO_2 as quartzite and $CaCO_3$ as limestone or chalk. Fortunately, our understanding of many ore-forming processes is constantly being broadened through detailed observation of both known deposits in existing mines, and new kinds of deposits revealed by mineral exploration. Field observations are supplemented by the wide range of laboratory studies including fluid inclusions, stable isotopes of sulphur and oxygen, and age dating, in addition to petrological methods of microscopy on transparent and opaque minerals. Knowledge about mineral deposits, the geological controls that influence their characteristics, the environments in which they develop and the genetic processes that produce them, is the basic science that is applied by economic geologists in their professional functions as mining geologists and exploration geologists (*see* **Economic Geology**).

The following introduction is a brief and necessarily incomplete outline of some aspects of the great variety of mineral deposits. The scheme of arrangement is based on the association of mineral deposits with the major geological rock groups – igneous, sedimentary and metamorphic. The order of presentation is pragmatic, moving from deposits and their genetic processes that are readily understood, to those that are enigmatic and controversial. We start, therefore, with those deposits and processes that we can observe at the present time in our own familiar surface environments, and progress through associations with basic igneous rocks, felsic igneous rocks, ancient sedimentary rocks, and, finally, metamorphism.

Present Day Associations and Processes

The easiest processes to understand are those we can actually observe in operation at or near the Earth's surface, namely the weathering of rocks, the transportation of the products to form secondary deposits, and the precipitation of minerals from solution. Prolonged weathering of suitable bedrock, by no means a simple process in detail, produces some major resources that can be easily mined (**Figure 1**). The world's dominant Al resource is bauxite (139 million tonnes per annum) derived either from limestones as 'terra rossa', mainly in Jamaica, or from laterite type profiles, with the largest deposits in Australia, Brazil and Guinea. Even gold may be concentrated in, and extracted from, some lateritic and clay profiles developed over auriferous bedrock, for example in Western Australia. New Caledonia in the late nineteenth century supplied most of the world's nickel and continues to supply about 10% of annual production, mainly as green garnierite (hydrated Ni-Mg silicate) concentrated from ultrabasic host rocks into weathered bedrock and the overlying laterite.

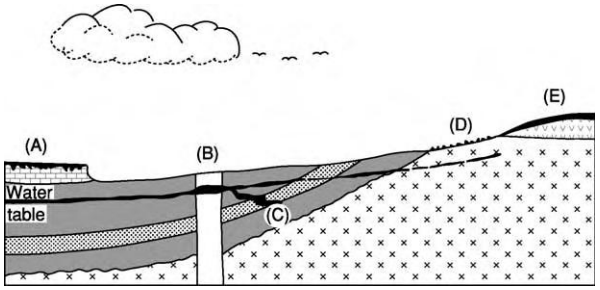


Figure 1 Diagrammatic crustal profile showing the habitats of economic deposits related to weathering processes. (A) Bauxite formed from the weathered residue of limestone. (B) Supergene enrichment of sulphide ore. (C) Uranium 'roll front' ore body. (D) Residual placer formed on weathered basement. (E) Lateritic iron, nickel, and manganese deposits formed on weathered ferromagnesian rich volcanics. Reproduced from Selley RC (2000) *Applied Sedimentology*, 2nd edn. p. 36. San Diego: Academic Press.

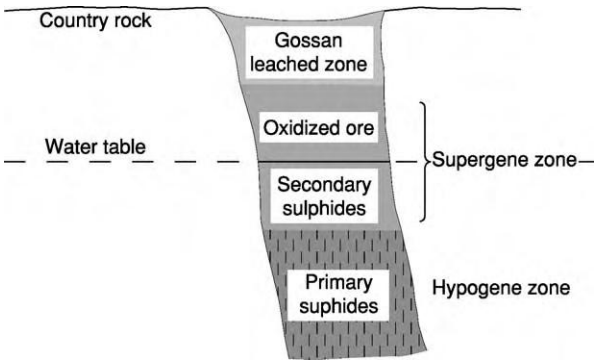


Figure 2 A simplified sketch section showing typical near surface zonation produced by weathering of a metalliferous sulphide deposit. Reproduced from Selley RC (2000) *Applied Sedimentology*, 2nd edn. p. 37. San Diego: Academic Press.

Nickeliferous laterite with associated Co and Mn is also being mined or tested elsewhere as appropriate extractive metallurgy is developed. It is probable that residual deposits of laterite type constitute a major resource bank for the future.

Metalliferous sulphide minerals are reactive and consequently metal sulphide-rich mineral deposits weather readily in most climates to a secondary 'oxide' mineral assemblage capped by a leached iron-rich' gossan' and underlain around water table depth by an enriched supergene sulphide mineralization (Figure 2). These weathered profiles, especially for copper ores, have been economically important since early man found that the brightly coloured minerals, such as malachite, were easy to discover and to smelt to metal. In large ore bodies, such as porphyry copper sulphide deposits, the oxidized ores and supergene enriched chalcocite blankets may form an important

Figure 3 Colluvial boulders of kyanite being mined at Lapsa Buru, India. Kyanite ($A1_2$ SiO_5) is an industrial mineral used in mullite refractories such as sparking plugs. Photo: GR Davis.

separate resource. These require different techniques of mining and treatment, but in many cases they provide an early return on capital because of the shallow working depth.

A most striking and economically important effect of prolonged weathering cycles is the conversion of banded iron formation (BIF) into high-grade iron ore by leaching of silica, thus raising the iron content and breaking down hard rock to softer fabrics that are cheaper to mine.

Minerals resistant to weathering may be liberated from their host rocks and concentrated nearby as eluvial and colluvial deposits, such as kyanite (Figure 3).

Resistate minerals may also be carried away in water by erosion, to be deposited elsewhere as placer deposits of heavy minerals such as gold, wolfram and cassiterite. Gold, the eternal metal, has provided wealth and romance through the ages. Jason and his Argonauts sought the Golden Fleece, the sheep's coat in which fine gold was trapped by washing alluvium. Two millennia later came the famous rush to the placer deposits of the Yukon, and today's prospecting by metal detector in Australia for eluvial nuggets grown in the weathering environment. Modern placers no longer provide a major source of the noble metals, but platinum placer mining is important in Russia.

Mechanically concentrated by the action of gravity in moving water, placer deposits, both fluvial and marine, are of major economic importance (Figure 4).

Cassiterite placers in Indonesia, Malaysia and Nigeria are fluviatile, and the coastal environment includes the titanium sands of Florida, and a growing number of black beach sand deposits. Mining of ilmenite, rutile and zircon sands dominates the market for these industrial minerals, which occur as very

low grade (3–6% total heavy minerals), but cheaply mined deposits at places along the wave-swept sandy shores of southern Africa, Australia, and India. Special interest attaches to placer deposits of the 'forever' mineral diamond, in South Africa. Average gem quality and value is progressively enhanced as stones are swept from their parent kimberlite pipes down the Orange River via river terrace alluvial deposits to the even more vigorous environment of the marine gravel terraces and wave cut platforms of the south-western coast of Africa. Only the best survive the journey (Figure 5).

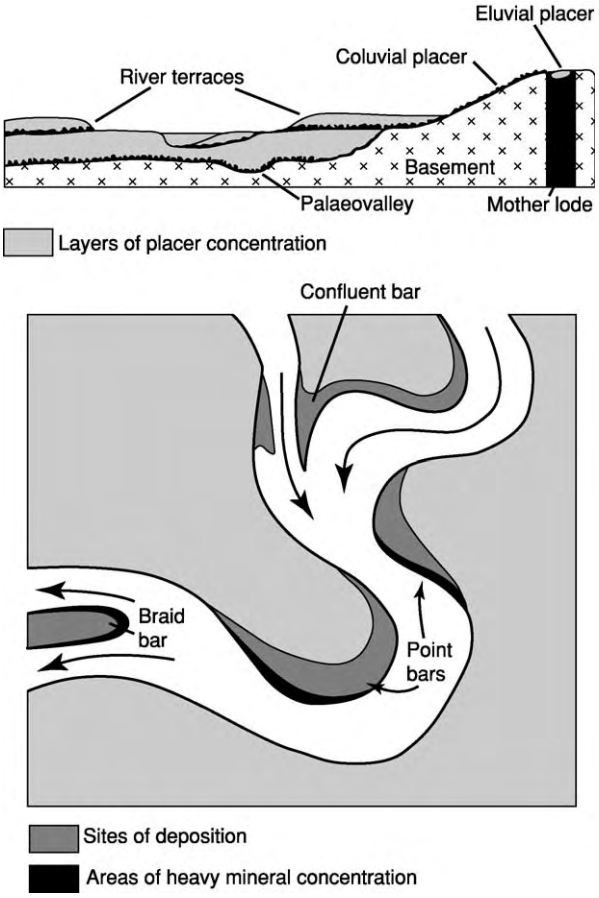


Figure 4 Cross section and plan showing the loci of alluvial and related placer mineral deposits. Reproduced from Selley RC (2000) *Applied Sedimentology*, 2nd edn. p. 205. San Diego: Academic Press.

Precipitation of salts from solution has been known to mankind since time immemorial. Salinas, or salt pans, are widely used to evaporate sea water and recover a succession of salts, notably halite (common salt), and so mimic the natural process of salt accumulation is closed or restricted marine basins. A somewhat different process concentration of gypsum and other salts can be observed in the sabkha flats fringing hot desert seas. Both processes contribute to the thick salt strata found in many sedimentary basins. Such salt deposits not only form the basis of large scale salt mining, but play an important role in petroleum geology as source rocks, traps and seals.

One observable, but little understood process that deserves much more attention concerns the metal-fixing powers of bacteria, as seen, for example, in some iron and manganese deposits. Suggestions concerning the possible role of bacteria have been made for the Late Precambrian disseminated Cu deposits of the Zambian Copperbelt, where stromatolites are developed. Pyritic Au-U ores of the Late Archaean (*ca.* 2.8 Ga) Witwatersrand conglomerates in South Africa also contain carbonaceous material. This has been attributed to primitive organic matter or remnants of algal colonies that grew at the time of sedimentation, but another theory holds that it was introduced as oily material much later during fracturing of the lithified sediments. Progress in this field is likely to come via research in mineral processing

Figure 5 Diamond deposits on the south western coast of Africa. (A) Diamonds are concentrated in deeply gullied wave cut rock platforms exposed after stripping off later wind blown sand cover. (B) Well rounded hard pebbly gravel exposed on a raised beach. The heavy mineral content includes diamonds in hydraulic balance with larger pebbles. (Fountain pen lower left shows scale). Photo: GR Davis.

technology, where increasing use is being made of bacterially assisted leaching of refractory metallic ores. Increased understanding of the role of bacteria in removing and precipitating metals from dilute solutions may have a profound influence upon theories of ore genesis.

The clearly important role of meteoric water in residual and depositional deposits in the present-day superficial environment reflects 'the fundamental and powerful part played by fluids' in nearly every kind of mineral deposit. Subsurface fluids may be molten magmas, circulating meteoric waters, connate waters squeezed out of sedimentary basins, magmatic waters resulting from the solidification of igneous magmas, and waters generated by metamorphism. All may contain dissolved solids and/or gases, and if heated they are known as hydrothermal solutions. Because most processes of mineral concentration cannot be directly observed, various genetic theories have been advanced on the basis of available field observations and laboratory evidence. Such theories seek to explain the origin of both the mineralizing fluid and its dissolved ore-forming constituents, and to relate the time of mineral concentration to the time of formation of the host rocks. Mineralization considered to have been formed at the same time as the host rock is termed syngenetic, whereas epigenetic mineralization is considered to be introduced into the host rock at a later time.

Conceptual theories are vital to scientific understanding and advancement, but invite debate and controversy. They tend also to trap dogmatic adherents into forcing good, but inconvenient new evidence into existing pigeonholes, or brushing it aside, unless a coherent new genetic mechanism is revealed. There is no better example of this than the denial by many geologists of the evidence for continental drift, until the discovery of sea-floor spreading provided a credible mechanism for embracing that same evidence. Development of the new theory of plate tectonics during the 1960s transformed the scientific framework of geology, including views on the nature and origin of mineral deposits. During that same period some other discoveries of observable natural mineral concentration systems, currently active at or near the surface of the Earth, had a major impact on theories of ore genesis. Modern sea-floor exploration revealed the vast extent of the manganese nodules described by the Challenger expedition of 1873. This demonstrated not only an enormous potential resource of Cu, Co, Ni and other associated metals, but also the potential of cold sea water as a dilute mineralizing fluid. Also at the Earth–Ocean interface, the hot brine pools and underlying soft ferruginous muds rich in Zn, Cu and Ag discovered in 1965 in the Red Sea deeps demonstrated an exhalative deposit still forming at the present day in a continental rift system. On the mid-Atlantic ridge the discovery of active 'black smoker' hydrothermal vents and massive sulphide deposits made a dramatic impact. A 1600 m deep geothermal well in the Salton Sea in California (1962) tapped hot metalliferous brines that precipitated dark siliceous deposits containing about 20% Cu, 1% Ag and other metals in the discharge pipe at surface temperatures and pressures. Contemporaneously with these revelations, fluid inclusion and chemical studies were highlighting the potency of brines and chloride complexes as solvents and transporters of metals.

Despite current knowledge of present-day processes, the interpretation of many features of mineral deposits in ancient sediments remains controversial, and will be discussed later.

Associations with Basic and Ultrabasic Rocks

These deep-seated rock types with low silica, high Fe-Mg content are the source of leading world supplies of minerals as diverse in use as chromite, platinum group minerals, vanadium, nickel, and diamonds. In all these rocks, the problems of ore genesis are very much part of the problems of rock genesis. The valuable elements or minerals are an integral part of the geochemistry of the parent magma, which is itself the ore forming fluid.

Layered basic intrusions display magmatic segregation during undisturbed cooling and differential crystallization, a process readily understood through clear field relationships supported by laboratory studies on ore textures. Heavy oxide minerals such as chromite crystallize and settle into discrete layers as syngenetic magmatic sediments. Sulphide minerals in contrast are deposited late (in many cases at the base of the intrusion or injecting wall rocks) from an immiscible sulphide melt that persists to the last stages of magma crystallization. The oxide assemblage in strongly layered basic rocks is exemplified by the repetitive and regionally extensive chromite and vanadiferous magnetite layering of the giant early Proterozoic Bushveld Igneous Complex in South Africa. The closely associated and famous platiniferous Merensky reef is coarse grained (pegmatoidal), with minor sulphides of Ni and Cu. The sulphide assemblage is best known from the nickel mines of Sudbury, Canada, a differentiated, but not layered basic complex, now considered to have been triggered by meteoric impact, and the Norilsk deposit in flood basalts in Siberia. Ultrabasic rocks of Archaean age (which include komatiite lavas) are also host to important nickel-copper-iron sulphide segregations in

the Yilgarn block of Western Australia (another discovery of the 1960s) and to very large masses of metallurgical grade chromite at Selukwe (now named Shurugwi) in Zimbabwe. The Selukwe occurrences contrast with the genetically similar, but much smaller podiform chromite segregations in ophiolites of the Tertiary Alpine belt in eastern Europe.

A source of ilmenite alternative to beach sands is found in alkaline igneous complexes such as the anorthosites at Egersund in Norway, Allard Lake in Canada and Stillwater in the USA. The distinctive pipe-like intrusive carbonatites are closely associated with peralkaline intrusive complexes and are mined for an unusual range of minerals. The highly mineralized Palabora complex, South Africa (Figure 6), produces copper (average ore grade 0, 54% Cu), apatite, vermiculite, magnetite, baddeleyite (a zirconium oxide), uranium and by-product sulphuric acid. Kimberlite pipes, diatremes and some dykes are famous as the primary source of diamonds. Economic grades are typically only a few carats to the tonne (1 carat = 0.2 g), but most kimberlites do not contain diamonds in sufficient quantity and/or quality to meet economic criteria. These deep-seated intrusives bring xenoliths of mantle material to surface, and diamond mining has spurred geological research. This has resulted in genetic understanding that has in turn lent impetus to the current exploration boom and new discoveries – a sequence of events typical of progress in economic geology.

Figure 6 Palabora mine open pit, averages 1750 M across the rim and has reached its economic working depth of 822 M in 38 years. Beneath the open pit underground mass mining by block caving methods is now producing 30 000 tonnes per day of copper ore from the core of this very productive carbonatite pipe. Photo graph courtesy of Rio Tinto plc.

Associations with Felsic Igneous Activity

Granitic magmas consolidate in a dynamic and high-energy environment dominated by superheated fluids at high temperatures and pressures. As rising magma slowly consolidates the portion still melted becomes progressively more enriched in volatiles. Release of pressure, by volcanic eruption, for example, may result in the mechanical release of gases capable of extensive fracture damage to both the intrusive and also the intruded country rocks. At high crustal levels the heat around the magma may create convection cells of surrounding groundwaters and augment the fluid content, so that episodic igneous and hydrothermal activity may persist over tens of millions of years. The fluids incorporated into the magmatic system may thus include deep connate waters and meteoric waters, both of which could be mineralized through leaching of the rocks that they traverse. These mineralized fluids may then transport and precipitate their dissolved load in rocks at higher levels, at lower temperatures and pressure, or escape to the surface. The terms hypothermal, mesothermal and epithermal are used to describe mineral assemblages attributed to hydrothermal fluids at relatively high, moderate and low temperatures and pressures (Figure 7).

The vast range of mineral deposits associated with felsic intrusions include pyrometasomatic replacement such as skarns (tactites) forming scheelite at contacts with limestones, and corundum at contacts with basic rocks. Complex coarse grained pegmatites are a product of late-stage fluids, and are mined for a variety of minerals including cassiterite, wolframite, beryl, mica and lithium minerals. Extensive vein systems may form in and around intrusive granite cupolas, well displayed by the Cornubian batholith of southwest England, mined since pre-Roman times, but no longer economic. Rich metallic mineralization of Sn-Cu-Zn-Pb displays a marked zonal pattern, and hydrothermal alteration of some granite areas (supplemented by later surface weathering) formed extensive kaolin deposits. Zonal patterns of wall-rock alteration and of metallic-sulphide mineralization also characterize the immensely productive, but low-grade porphyry copper and molybdenum deposits that contribute so much to current world production of base metals and by-product gold. These huge ore bodies are formed as impregnations at depth within and around hydrofractured calc-alkaline intrusives above subduction zones at destructive plate margins. Many gold- and silver-rich epithermal ore deposits formed nearer the surface are attributed to late-stage magmatic or mixed magmatic-meteoric fluid systems linked to the porphyry bodies. One interesting

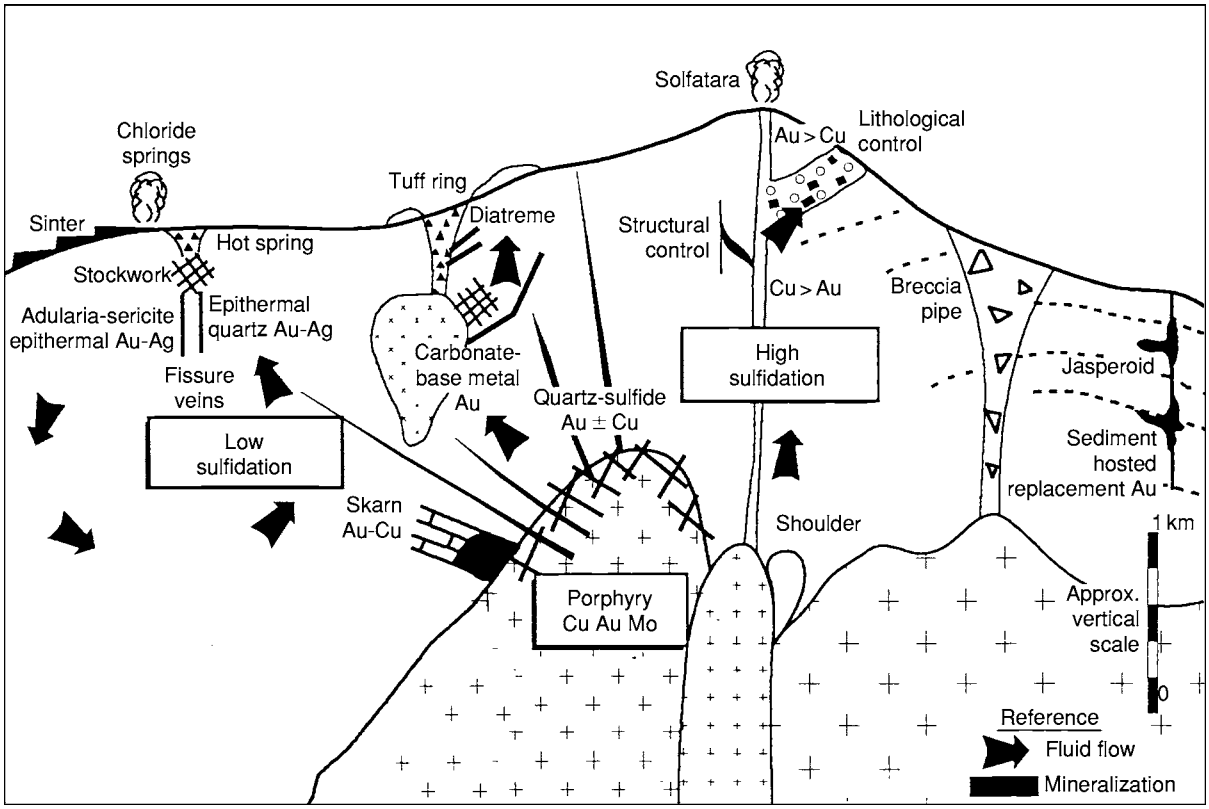


Figure 7 Pacific Rim gold copper mineralization models. This cross section sketch is an example of the types of mineralization associated with felsic intrusive rocks. Reproduced with permission from the Society of Economic Geologists Inc., *Southwest Pacific Rim Gold Copper Systems: Structures, Alteration, and Mineralization*, p. 6, Fig. 1.1, Corbett GJ and Leach TT (1998).

example is at Lihir island within the Pacific 'ring of fire', where large-scale mining recently commenced on very fine-grained gold in pyritic mineralization within an inactive volcanic caldera. The Carlin-type deposits of Nevada, on the other hand, are situated along regional structural trends. The gold mineralization is commonly so finely disseminated within the dark impure calcareous host rocks that mining limits are determinable only by assay, and the discovery in the 1970s was hailed as 'virus gold'.

Rhyolites and dacites, the extrusive forms of granites and granodiorites, are host rocks to economically and geologically important types of base-metal sulphide mineralization that are transitional between sub-surface and surface. The Iberian pyrite belt of southern Spain and Portugal, and the Kuroko deposits of Japan are leading examples. Subsurface mineralization of transgressive stockworks in explosive rhyolite domes underlies massive stratiform sulphide bodies that are interpreted as deposited at surface in a submarine environment (Figure 8). These two closely associated deposits, of distinctly different style and appearance, are attributed to the same mineralizing ore fluids at work in the subsurface and, after exhalation, above the rock-water interface.

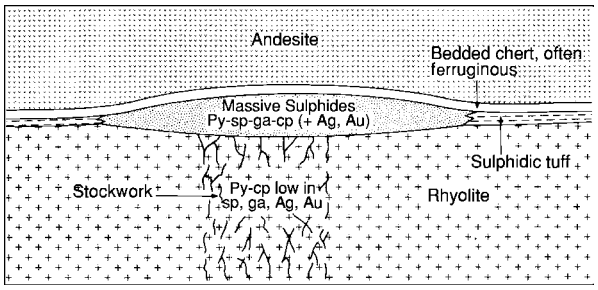


Figure 8 Schematic section through an idealized volcanic associated massive sulphide deposit showing the underlying feeder stockwork and typical mineralogy. Py pyrite, sp + spha sphalerite, ga galena, cp chalcopyrite. Reproduced from Evans AM (ed.) (1995) *Introduction to Mineral Exploration*. Oxford: Blackwell Science.

Associations with Ancient Sedimentary Rocks

Knowledge of present-day surface processes assists in the understanding of ancient sedimentary deposits such as coal, and salt deposits formed by evaporation of brines from closed basins or sabkhas. Geological details vary, but the genetic picture is clear enough to guide exploration for further resources. However,

controversy and lack of understanding surrounds many deposits found in sediments, especially metalliferous ores. Dispute tends to centre on the nature and origin of the ore forming fluid, its movement through the crust, the origin of the metals, and the mechanism of their precipitation.

The controversy is well illustrated by the continuing debate concerning the origin of the giant Witwatersrand gold deposits of South Africa, the world's greatest gold producer for over a century. A well-preserved Late Archaean sedimentary basin, about 6 km thick and 280 km across, contains pyrite-uraninite-gold mineralization in conglomerates developed at several unconformities from the bottom to the top of the sequence (**Figure 9**). Economic interest is focused on a series of conglomerates towards the top of the basin and situated near its edge. Despite compelling evidence for a paleoplacer origin, undisputed signs of some hydrothermal activity in the sediments are still being used to argue the minority view that epigenetic hydrothermal processes played a major role in the introduction of the gold. Another syn-sedimentary hypothesis holds that gold was precipitated from solution, on the evidence of carbon ('thucholite') that may be derived from primitive organisms. The uncertainties continue, as some others hold that the carbonaceous matter was introduced from an external source long after consolidation of the sediments.

Ores in sediments may be discordant irregular masses of clearly epigenetic origin; or concordant and possibly, therefore, either syngenetic with the enclosing sediment, or diagenetic (that is, due to processes of diagenesis). The term 'stratabound' is used to indicate confinement of the mineral deposit between or within sedimentary strata, and 'stratiform' to indicate that the deposit shows internal layering or stratification.

Most Phanerozoic ironstones are sedimentary, and the intriguing Banded Iron Formation, widespread in Proterozoic and Archaean basins, is well documented as the major sedimentary protore of high-grade iron ore production. Many stratiform massive base-metal sulphide deposits, once considered by many to be epigenetic replacements due to hydrothermal fluids from hidden igneous bodies, have been shown to be the products of hydrothermal exhalations and a volcano-sedimentary environment, as for the Kuroko and Iberian deposits mentioned above. At the other pole are the Late Proterozoic stratabound Cu-Co disseminated sulphide ores of the Copperbelt in northern Zambia and the adjacent Congo. Mineralization of just a few percent sulphides is confined to strata only tens of metres thick, but extending (in large regional folds) over thousands of square kilometres. The only igneous rocks known nearby are basement granites that form the hilly topography upon which the cupriferous sediments were deposited, and much later basic dykes. The detailed evidence, including a marked facies zoning upwards and shoreward of pyrite-chalcopyrite-bornite-chalcocite (**Figure 10**), convinced most geologists that the ores are not simple hydrothermal replacements as favoured after mine development in the 1920s. An origin due to syngenetic precipitation of pyrite and copper sulphides by bacterial action has long been advocated, together with effects due to diagenesis and metamorphism, but not convincingly demonstrated. Current research based on increased regional knowledge (over 1000 million

Figure 9 A typical specimen of Witwatersrand gold ore from South Africa. Pyrite and minor gold in the groundmass supporting well rounded, unmineralized quartz pebbles, 25 cm across. Photo: GR Davis.

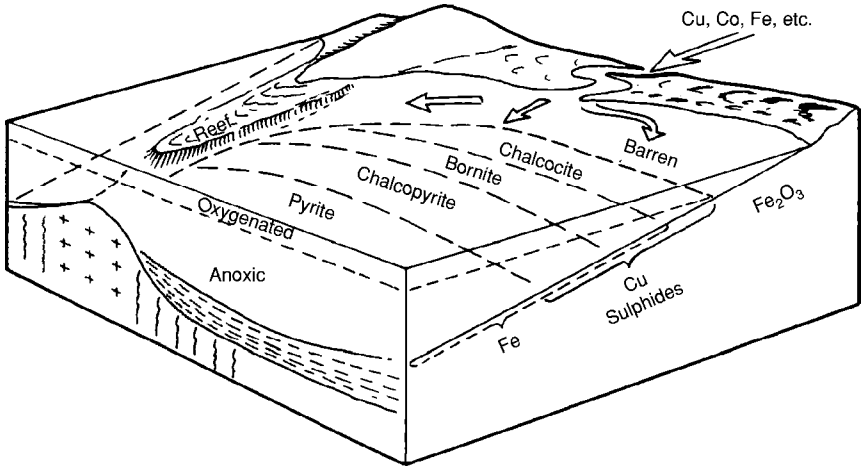


Figure 10 A diagram to illustrate the environment of sedimentation of Zambian Copperbelt host rocks. A marked zonation of copper and iron sulphides exists near the shoreline formed by the hilly basement rocks. The diagram (1976) illustrates a syngenetic theory of origin for mineralization, which is not universally accepted. Reproduced from Fleischer VD, Garlick WD, and Haldane R (1976) Geology of the Zambian copperbelt. In: Wolf KH (ed.) *Handbook of Stratabound and Stratiform Ore Deposits,* vol. 6, pp. 223 350. New York: Elsevier.

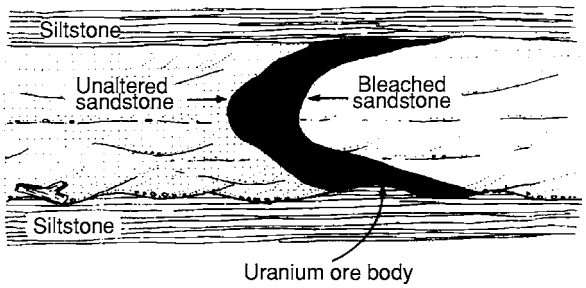


Figure 11 Sketch illustrating the occurrence of carnotite roll front ore bodies in fluvial channel sandstones. The direction of front migration in this case was from right to left. Reproduced from Selley RC (2000) *Applied Sedimentology*, 2nd edn. p. 38. San Diego: Academic Press.

tonnes averaging about 2.7% Cu has been mined) and laboratory studies suggests that metal-bearing sulphate brines were introduced through tectonic thrust faulting and that copper sulphides were epigenetically precipitated by organic matter in the host rocks. Neither theory has demonstrated the source of the metals in the seawater (for syngenesis) or in the deep-seated fluids (for epigenesis), or fully explained all the observed field relationships and laboratory data.

Knowledge of diagenetic processes is advancing and is illustrated by roll-front uranium deposits in Colorado, where carnotite has been precipitated from migrating pore waters in confined aquifers (Figure 11). The mottled lead sulphide sandstone ore bodies of the extensive Laisvall and related deposits of the Swedish Caledonides may also be due to migrating mineralized fluids.

The discordant masses of zinc and lead sulphide mineralization that are widespread in sedimentary carbonate sequences, such as the Mississippi Valley deposits of the USA, have clearly been deposited later than their host rocks. The ore forming fluids were for many years attributed to igneous sources, for lack of a better option. It has now been shown in some well-documented cases that oilfield brines migrating up-dip from deep sources are the ore forming fluids from which sulphides were precipitated. From years of meticulous field observations backed up by laboratory studies (such as fluid inclusions, stable isotope and geochemical analysis) there has emerged an elegant and logical understanding that both oil and ore deposits are linked through such common controls as source, migration path, trap and seal. The effect of this advance in knowledge about mineral deposits has been dramatic. The philosophy of prospecting for new mineral fields of this type has switched, to put the case over-simply, from searching carbonate sequences near intrusive granites to studying fluid migration paths in sedimentary basins containing hydrocarbon source rocks and evaporites. An understanding is developing of the thermochemical reactions (at elevated pressures and temperatures) between hydrocarbons and sulphate brines that can release aggressive rock-altering fluids and hydrogen sulphide, a powerful precipitant of base metals. This example, and that of the Zambian Copperbelt, demonstrates that ores in ancient sediments are best studied, like the hydrocarbon deposits to which they may be linked, in the context of the entire geological history of the containing sedimentary basin.

Metamorphic Associations

Metamorphic processes may either generate new mineral deposits, or modify the characteristics of existing deposits. A number of industrial minerals are formed by metamorphism, including kyanite for refractories, and garnet for abrasives. Talc forms in metamorphosed ultrabasic rocks, which are also the major source of chrysotile asbestos. When olivine rich rocks such as dunite are serpentinized, the fibrous form of serpentine, chrysotile, is formed in cross-fibre veins that seem to develop under conditions of tensile stress and hydrofracturing. A dense enough network of veins (a few per cent by volume) of good-quality fibre converts large masses and lenses of serpentinized rock into a mineral deposit of economic potential, usually suitable for mass-mining methods (Figure 12). Despite the health hazards associated with asbestos, world production is still about 2 million tonnes per annum. There are large deposits in the Urals, the Palaeozoic of the eastern Townships in Quebec, and in the Archaean at Zvishavani (Shabani) in Zimbabwe. Tectonic deformation is also an important precursor to epigenetic mineralization ('ground preparation') as it forms shear zones or fractures that provide channelways for mineralizing fluids. Dynamic metamorphism and the repetitive action of a process known as seismic pumping may drive mineralizing fluids from depth into higher level vein systems. Activated by earthquake movements, fluids accumulating at depth in large fault systems are expelled upwards through channelways where dissolved substances may be precipitated as successive generations of vein fillings. Large numbers of vein systems, commonly containing gold, are associated with the shear and fracture zones typical of greenstone schist belts in Archaean cratons and Precambrian Shield areas. Geological and mining conditions may thus be very similar in such gold fields areas as far apart as Australia, Canada and West Africa.

A second and geologically important aspect of metamorphism relates to the fact that pre-existing mineral deposits, like all other rocks, must suffer the effects of regional metamorphism. Deposits not only become faulted, folded and sheared into new geometric shapes and attitudes that affect the cost of mining, but may respond by recrystallization in ways that materially affect their amenability to mineral extraction and beneficiation. Sulphide minerals are particularly prone to recrystallize and anneal, or become mobilized into new open spaces such as rock cleavage. Veins of gangue and ore minerals may also be formed by lateral secretion driven by metamorphic temperatures and stresses. The overall effect is to overprint and mask the original ore fabrics, especially at the micro and meso scales. Such difficulties lead to additional debate about controls on ore formation and ore genesis, which in their turn affect decisions on how and where to explore for similar mineralization. Some examples include the famous massive sulphide Pb-Zn ores of Broken Hill (Australia) in strongly folded sillimanite grade gneisses, the folded disseminated Cu deposits of the Zambian copperbelt, the great conglomerate hosted Au-U deposits of the Witwatersrand (South Africa) and the twice-folded and metamorphosed stratabound Cu-Co deposits at Kilembe (Uganda).

Figure 12 A network of cross fibre chrysotile asbestos veins in serpentinite, Havelock mine, Swaziland. This length of fibre and intensity of mineralization constitutes superior grade ore. Photo: GR Davis.

Discovery of Further Mineral Deposits

The AGI definition of economic geology includes the 'application of geological knowledge and theory to the search for.... mineral deposits', recognizing the fact that the extractive industries must meet future demands for new materials and replace mined out deposits with new discoveries. Over the centuries the world's need for mineral supplies has always been met by some combination of chance finds at outcrop, and indirect indications from prospecting methods such as panning, geophysical and geochemical surveys, aerial photography and earth satellite imagery. Whether the search is for an extension to ore at a working mine or a new grassroots discovery, the choice of where to look for success is based on someone's idea that a given part of the Earth's crust will actually contain economic mineralization. The best ideas are based upon two factors – superior knowledge of the observed and detailed characteristics and field associations of the type of deposit targeted, and a valid theory of its genesis. These two factors have long been used empirically in prospecting, but over the past 20 years the available information has been systematically gathered and codified into a large number of 'deposit models'. These models are continually being tested, improved, and added to as discoveries of new and unexpected types of mineralization are made. For example, the huge Olympic Dam copper-uranium deposit, discovered under 350 m of barren cover rocks in South Australia in 1975, now represents a broad and complex group referred to as iron oxide-copper-gold deposits.

The search becomes more difficult as the number of undiscovered deposits decreases and the emphasis moves further towards deposits deeply buried and hidden under overlying rock cover. This growing challenge is being met with an integrated approach of new and broader ideas backed by an ever-improving tool-kit of geophysical and geochemical field exploration technologies and elegant laboratory techniques. The tool-kit also includes readily available space satellite imagery that now provides the broad perspective on which to study major crustal features and lineaments that may link the surface to the depths below the crust. Economic mineral concentration is seldom an isolated event, but rather one result from a dynamic system linking tectonics, magmatism, volcanicity, and sedimentation in complex ways. To guide exploration to best advantage the entire system should be investigated and understood as far as possible. Twenty-first century economic geologists, striving to understand the ore forming systems in our complex Earth's crust, may be poised to take the biggest step forward in the subject since 1556, when Georgius Agricola (Bauer) famously classified ore deposits and recorded the state of the art of mining in his classic work, *De Re Metallica*.

See Also

Economic Geology. **Mining Geology:** Exploration; Mineral Reserves; Hydrothermal Ores; Magmatic Ores. **Sedimentary Rocks:** Banded Iron Formations; Ironstones.

Further Reading

Cooke DR and Pongratz J (eds.) (2002) *Giant Ore Deposits: Characteristics, Genesis and Exploration. CODES Special Publication 4*. Hobart: University of Tasmania.

Craig JR, Vaughan DJ, and Skinner BJ (2001) *Resources of the Earth: Origin, Use and Environmental Impact*, 3rd edn. New York: Prentice Hall.

Davis GR (1988) Is Metallogeny a Practical Exploration Tool? *Episodes* 11(2): 105 110.

Derry DR (1980) *World Atlas of Geology and Mineral Deposits*. London: Mining Journal Books.

Evans AM (1993) *Ore Geology and Industrial Minerals, an introduction*, 3rd edn. Oxford: Blackwell.

Kirkham RV, Sinclair WD, Thorpe RI, and Duke JM (eds.) (1993) *Mineral Deposit Modeling. Geological Association of Canada Special Paper 40*.

Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

Stanton RL (1972) *Ore Petrology*. New York: McGraw Hill.

MINERALS

Contents

Definition and Classification

E H Nickel, CSIRO Exploration and Mining, Wembley, WA, Australia

Introduction

Minerals, the individual components comprising rocks, are generally defined in terms of chemical composition and crystal structure, and most classification systems are based on these properties. Minerals are formed by geological processes and include both terrestrial and extraterrestrial materials. The total number of minerals generally accepted as valid by the mineralogical community is about 4000. Mineraloids are mineral-like substances such as synthetic materials, human-influenced substances, and some biological materials that do not satisfy all the criteria for the definition of a mineral species.

Definition of a Mineral Species

Minerals are substances formed by geological processes that occur on Earth or in extraterrestrial bodies. A mineral is defined on the basis of its chemical composition and crystal structure, and to qualify as a mineral species, a substance must have a unique combination of these properties. Some minerals have a well-defined composition (quartz, for example, is SiO_2), but many have a variable composition whereby some chemical constituents are replaced by others. For example, the mineral olivine can be regarded as a solid-solution series between forsterite (Mg_2SiO_4) and fayalite (Fe_2SiO_4) in which Mg^{2+} and Fe^{2+} substitute for each other in the crystal structure. In such

a case, the two end-members are regarded as species, with a compositional range extending from the end-member composition to the midpoint of the solid-solution series, namely 50 mol%. 'Olivine' is therefore regarded as a series or group name. Other solid-solution series are more complex, as in the case of ternary or higher order solid solutions, as exemplified by the mutual replacement of F, OH, and Cl in the apatite series, giving rise to the end-member species fluorapatite, hydroxylapatite, and chlorapatite, respectively. Further complications may involve coupled substitutions, such as in britholite-(Ce), in which the replacement of Ca^{2+} in the apatite structure by Ce^{3+} is balanced by the substitution of P^{5+} by Si^{4+}. In all such cases, the species is defined on the basis of the predominant ion in a specified structural site.

Some minerals with identical compositions have different crystal structures; examples include quartz, tridymite, and cristobalite, all of which have the chemical formula SiO_2 but which crystallize in the hexagonal, orthorhombic, and tetragonal crystal systems, respectively. Such substances are called polymorphs and qualify as separate mineral species because they have different structures. On the other hand, there are minerals with the same crystal structure but with different compositions, such as galena (PbS), periclase (MgO), and halite (NaCl). These, too, are regarded as separate species.

A structural variant that is sometimes encountered, notably in micaceous minerals, is created by the rotation of one structural unit with respect to its structurally equivalent unit. When this perturbation is distributed throughout the structure as a regular stacking sequence, the structural variant is called a polytype. Polytypes are not regarded as separate species and are indicated by the addition of a hyphenated suffix to the root mineral name. The suffix takes the form of a numeric symbol that represents the periodicity with respect to the basic structural element, an alphabetic symbol that represents the crystal system, and sometimes a numeric subscript that represents the type of stacking. Some examples of mica polytypes are muscovite-1M (monoclinic, with a periodicity of 1), muscovite-$2M_2$ (monoclinic, with a periodicity of 2 and a particular stacking sequence), muscovite-3T (trigonal, with a periodicity of 3), and muscovite-2A (anorthic, i.e., triclinic, with a periodicity of 2).

Although the crystal structure of a mineral is one of the important properties defining a mineral species, some mineral substances do not have a long-range structural arrangement of the atoms comprising the mineral. Minerals lacking this long-range order are termed amorphous (if they were created in this form) or metamict (if they originally possessed long-range structural ordering, but the structural order was later destroyed, usually as a result of ionizing radiation). An amorphous mineral may be accepted as a valid species if there is evidence that it is chemically homogeneous and if spectroscopic characterization shows that it is unique. A metamict mineral may be accepted as a valid species if it can be established with reasonable certainty that the original substance was a crystalline mineral, generally by returning it to its original crystalline state by heat treatment.

Many minerals were formed under conditions of high temperature or pressure (or both) and are metastable under ambient conditions; others may tend to hydrate or dehydrate when removed from their place of origin. Such minerals may require special procedures to prevent their decomposition before an investigation is complete. The use of special procedures in the investigation does not preclude the acceptance of a metastable or unstable substance as a mineral, if it can be adequately characterized and if it meets the other criteria for a mineral.

The Validation of Mineral Species

Minerals have been given names since before the dawn of history, and some of these names have survived as valid species names to the present day. However, the vast majority have not. Following the establishment of the Commission on New Minerals and Mineral Names (CNMMN) of the International Mineralogical Association (IMA) in 1959, the introduction of new mineral names has been strictly controlled, and publication of new mineral names and descriptions requires prior approval by the national representatives of the CNMMN. Proposals for the creation and naming of a new species require a complete description of the mineral, including chemical composition; crystallographic, physical, and optical properties; crystal structure (if possible); and specification of the type locality where the mineral was discovered. If approved, the specimen providing this information becomes the type specimen for the species and is the specimen to which all subsequent descriptions of the mineral from other localities must be compared.

Minerals in the literature prior to the formation of the CNMMN have been progressively winnowed down to a generally accepted list of valid mineral species by the mineralogical community in general, and by activities of the CNMMN in particular. Subcommittees have been established by the CNMMN for the review of mineral groups, including the amphibole, mica, platinum-group alloys, pyrochlore, pyroxene, and zeolite groups, and reports of these subcommittees have been published in the open literature. Discreditation or redefinition of individual species has also been done under the auspices of the CNMMN.

Mineral Names

The author(s) of the original description of a valid new mineral have the prerogative of naming the mineral. However, the name must be approved by the CNMMN prior to publication. Mineral names have various derivations, the principal ones being the geographical locality of the discovery, a particular characteristic of the mineral, and the name of a person. Such names are sometimes called "trivial" to distinguish them from systematic ones. In an effort to reduce the proliferation of trivial names, the CNMMN has approved the use of root names with suffixes. Suffixed mineral names were introduced by AA Levinson in 1966 for rare-earth minerals, and involves the addition of a hyphenated chemical symbol in brackets after the root name, e.g., synchysite-(Ce). Such suffixes are generally referred to as Levinson modifiers, and the nomenclature of all minerals with one or more rare-earth elements predominating in a structural site must conform to this usage. The system of Levinson modifiers has subsequently been extended to some other mineral groups, notably the zeolites.

Mineral Varieties and Varietal Names

In addition to valid mineral names, which apply to mineral species or groups, other names are commonly used for particular varieties of minerals, generally those with distinctive coloring. This practice is especially common in gemology. Deeply coloured varieties of corundum, for example, have been given varietal names such as ruby (red) and sapphire (blue). Such names have no validity as species names, and their use is not controlled by the CNMMN.

Mineraloids

Mineraloids are substances that have some of the properties of minerals, but are not regarded as valid minerals, usually because they have not been formed exclusively by geological processes. One class of mineraloids is those produced by biological processes, such as mineral-like calculi in animals or organic crystals in plants. A pearl is therefore classified as a mineraloid rather than as a mineral. Some biogenic substances are subsequently found to be formed by geological processes as well, such as the urinary calculi whewellite and weddellite, and these then qualify as mineral species.

Another class of mineraloids includes synthetic substances. Even though they may have a definite chemical composition and a known crystal structure, they do not qualify as minerals because geological processes have not been involved in their creation. Such substances are called anthropogenic. Substances that are formed by a combination of anthropogenic and geological processes are also classified as mineraloids. These include substances formed as a result of mine or waste-dump fires and by the action of water on man-made substances. In the past, such substances were accepted as mineral species, as, for example, the submerged ancient Laurium slag 'minerals', but in recent years the CNMMN has enunciated a policy of not accepting occurrences of such substances as minerals.

Mineral Classification

Historical Background

The ancient classification of minerals was based mainly on their practical uses, minerals being classified as gemstones, pigments, ores, etc. Probably the earliest classification based on external characteristics and on some physical properties, such as colour, fusibility, malleability, and fracture, was that of Geber (Jabir Ibn Hayyaan, 721–803), later extended by Avicenna (Ibn Sina, 980–1037), Agricola (1494–1555), and AG Werner (1749–1817). This system was substantially refined by F Mohs (1773–1839) in his *Natural-History System of Mineralogy* (1820). With Werner, physical classification attained its maturity, and was in general use, by the end of the eighteenth century. Linnaeus (1707–78) attempted to classify minerals primarily by their external morphology, with a hierarchical system involving subdivision into genus, order, and class. A purely chemical classification was proposed by T Bergmann (1735–84), but this approach was premature, because many chemical elements had not been discovered at that time and analytical procedures were in their infancy. AF Cronstedt (1722–65) seems to have been the first to devise a classification scheme involving both chemical and physical properties, with chemistry predominating. Systematic crystallography was initiated by JBLR de l'Isle (1736–90), and this concept was applied by RJ Haüy (1743–1822) in *Traité de Minéralogie* (1801), in which he presented a mineral classification scheme based on the 'nature of metals', or, as it would be expressed now, the nature of cations.

With advances in chemistry, chemical properties became increasingly important, and a chemical classification of minerals was proposed in 1819 by JJ Berzelius (1779–1848). He recognized that minerals with the same non-metal (anion or anionic group) have similar chemical properties and resemble each other far more than do minerals with a common metal. He considered minerals as salts of anions and anionic complexes, namely, as chlorides, sulphates,

silicates, etc., rather than as minerals of Zn, Cu, etc. At this time, CS Weiss (1780–1856) introduced the seven crystal systems (1815) and Mitscherlich discovered isomorphy (1819) and polymorphy (1824). Gustav Rose (1798–1873) combined chemistry, isomorphy, and morphology to produce a chemical–morphological mineral system, and this was further developed by P von Groth (1843–1927) in his five editions of *Tabellarische Übersicht der Mineralien nach ihrer Kristallographisch-chemischen Beziehungen*. JD Dana (1813–95), in his *System of Mineralogy* (1837), based his classification system primarily on chemistry, and this emphasis has been maintained throughout subsequent editions of the *System*.

After 1912, following the discovery of the phenomenon of X-ray diffraction by crystals by M von Laue (1879–1960) and WH Bragg (1862–1942), and the elucidation of the first crystal structure (the mineral halite) in 1914 by WL Bragg (1890–1971), the crystal structures of minerals began to be taken into account in mineral classification schemes. The first classification of this type, which took into account the distribution of interatomic bonds in a structure, involved the structures of silicates, determined by Machatschki in 1928. This new field was rapidly expanded by WL Bragg in *The Crystalline State* (1933) and in the first edition of *Atomic Structures of Minerals* (1937). This combination of chemistry and structure in mineral classification was subsequently applied to many other categories of minerals, such as fluoraluminates (by Pabst; 1950), aluminosilicates (by Liebau; 1956), silicates and other minerals with tetrahedral complexes (by Zoltai; 1960), phosphates (by Liebau, in 1966, and by Corbridge, in 1971), sulphosalts (by Makovicky; 1981 and 1993), and borates (by Heller, in 1970, and by Strunz, in 1997).

The classification of silicates on the basis of polymerization of corner-sharing SiO_4 tetrahedra from insular groups to dimers, chains, rings, sheets, and frameworks proved to be a particularly useful scheme but, with the exception of the borates, this concept could not be comprehensively extended to other categories of minerals. The polymerization of cation-centred polyhedra by sharing corners, edges, and faces to form various configurations has been applied in the classification of some minerals, such as sulphates (by Sabelli and Trosti-Ferroni; 1985), copper oxysalts (by Hawthorne; 1993), and phosphates (by Hawthorne; 1998), but such schemes have only limited applicability to minerals as a whole. Other classification schemes, such as those stressing genetic aspects of mineral formation (by Kostov; 1975) or interatomic bonding (by Godovikov; 1997) have not been generally adopted by the mineralogical community.

Current Comprehensive Classification Systems

Chemical composition and crystal structure are the two properties that define a mineral species, thus it is no surprise that systems based on one or both of these properties are the most widely used. Russian mineralogists have been particularly prolific in devising crystallochemical classification schemes, e.g., Betechtin (in 1961), Povarennykh (in 1966), Lazarenko (in 1971), Godovikov (in 1983), Semenov (in 1991), and Bulakh (in 1995). Among other European mineralogists, such classification schemes appear to be less profuse, and the following few examples have been selected to illustrate the diverse approaches to mineral classification.

A classification based entirely on chemical composition is that of Hey (Table 1), whose *Chemical Index*

Table 1 Principal categories in Hey's chemical classification system

Category	*Description*
1	Elements and Alloys (including the arsenides, antimonides, and bismuthides of Cu, Ag, and Au)
2	Carbides, nitrides, silicides, and phosphides
3	Sulphides, selenides, tellurides, arsenides, and bismuthides (except the arsenides, antimonides, and bismuthides of Cu, Ag, and Au)
4	Oxysulphides
5	Sulphosalts sulpharsenites, sulphantimonites, and sulphobismuthites
6	Sulphosalts sulphostannates, sulphogermanates, sulpharsenates, sulphantimonates, sulphovanadates, and sulphohalides
7	Oxides and hydroxides
8	Halides
9	Borates
10	Borates with other anions
11	Carbonates
12	Carbonates with other anions
13	Nitrates
14	Silicates not containing aluminium
15	Silicates of aluminium
16	Silicates containing aluminium and other metals
17	Silicates containing other anions
18	Niobates and tantalates
19	Phosphates
20	Arsenates
21	Vanadates
22	Phosphates, arsenates, or vanadates with other anions
23	Arsenites
24	Antimonates and antimonites
25	Sulphates
26	Sulphates with halide
27	Sulphites, chromates, molybdates, and tungstates
28	Selenites, selenates, tellurites, and tellurates
29	Iodates
30	Thiocyanates
31	Oxalates, citrates, mellitates, and acetates
32	Hydrocarbons, resins, and other organic compounds

of Minerals (1950) has gone through three editions. In Hey's classification, minerals are divided into 32 main categories that are subdivided into smaller groupings on the basis of predominant cations. Most of the other classification systems combine the criteria of chemical composition and crystal structure. The system used by the International Centre for Diffraction Data (1993) categorizes the minerals into 177 groups based on crystallographic criteria; the groups are divided into subgroups on the basis of chemistry and/or crystallography. The principal groupings in *Fleischer's Glossary of Mineral Species* (1999) are similar, but lack the subgroupings.

In the system advocated in 1983 by Lima-de-Faria (Table 2), minerals are divided into aqueous and nonaqueous categories; these are then subdivided on the basis of element ratios and then on the basis of crystal-structure type. In the Dana system (Table 3), minerals are divided into 22 categories, based mainly on anion composition, except for the silicates, which comprise six categories based on polymerization of the SiO_4 tetrahedra. Further subdivision of the nonsilicates is based primarily on composition, whereas that of the silicates is based on the configuration of the SiO_4 units.

The Strunz system (Table 4), probably the most widely used classification system, divides minerals into 10 classes, based on anion composition. Further subdivisions are based on chemical and structural criteria, which are different in each of the classes. In Class 1, the primary subdivision is based on composition and the secondary one is based on structure. In Class 2, the primary subdivision is based on composition (chiefly the cation/ion ratio), the secondary one is based on further compositional criteria, and the tertiary one is based on structure. In Class 3, the primary subdivision is based on composition (principally on the compositional complexity and the presence or absence of combined H_2O), the secondary one is based on cation:anion ratio, and the tertiary one is based on structure. In Class 4, the primary subdivisions are based on cation:anion ratio and the presence or absence of H_2O, with separate subdivisions for the uranyl hydroxides, vanadates with 5- or 6-coordinated V atoms, and arsenites, antimonites, bismuthites, sulphites, selenites, tellurites, and iodates. Further subdivisions are based on cation size and structure type. In Class 5, the carbonates are subdivided first on the presence or absence of additional anions and H_2O, secondly on cation size, and ultimately on crystal structure. The nitrates are subdivided on the basis of the presence or absence of OH and H_2O. In Class 6, the primary subdivision is based on the number of borate units in the chemical formula and the secondary one is based on structure. In Class 7, the primary subdivision is focussed on the presence or absence of additional anions and H_2O,

Table 2 Principal categories in the classification system of Lima de Faria

Category	*Type*
Minerals without water molecules in their structure	A
	A_mB_n
	$A_pB_qC_r$
	$A_pB_qC_rD_s$
	$A_pB_qC_rD_sE_x$
	$A_pB_qC_rD_sE_xF_y$
	$A_pB_qC_rD_sE_xF_yG_z$ ...
	$A_pB_qC_rD_sE_xF_yG_z$
Minerals with water molecules in their structure[a]	A_pB_q ... E_xF_y ... *n* (aq.)

[a]Work on this category is in progress.

Table 3 Principal categories in the Dana classification system

Native elements and alloys	Phosphates, arsenates, and vanadates
Sulfides and related compounds	Antimonates, antimonites, and arsenites
Oxides	Vanadium oxysalts
Halogenides	Molybdates and tungstates
Carbonates	Organic compounds
Nitrates	Nesosilicates: insular SiO_4
Iodates	Sorosilicates: isolated tetrahedral noncyclic groups, $N > 1$
Borates	Cyclosilicates
Sulphates	Inosilicates: two dimensionally infinite silicate units
Selenates and tellurates; selenites and tellurites	Phyllosilicates
Chromates	Tektosilicates

Table 4 Principal categories in the Strunz classification system

Class	*Description*
1	Elements (metals and intermetallic alloys; metalloids and nonmetals; carbides, silicides, nitrides, phosphides)
2	Sulphides and sulphosalts (sulphides, selenides, tellurides; arsenides, antimonides, bismuthides; sulpharsenites, sulphantimonites, sulphbismuthites, etc.)
3	Halides
4	Oxides (hydroxides, $V^{[5,6]}$ vanadates, arsenites, antimonites, bismuthites, sulphites, selenites, tellurites, iodates)
5	Carbonates (+ nitrates)
6	Borates
7	Sulphates (selenates, tellurates; chromates, molybdates, wolframates)
8	Phospates, arsenates, vanadates
9	Silicates (germanates)
10	Organic compounds

the secondary one is based on cation size, and the tertiary one is based on structure; uranyl sulphates, chromates, molybdates/wolframates, and thiosulphates are in separate subdivisions. In Class 8, the primary subdivision is based on the presence or absence of additional anions and H_2O, the secondary one is based on cation size and OH:H_2O ratio, and the tertiary one is based on crystal structure. Separate subdivisions are reserved for uranyl phosphates and arsenates and for polyphosphates, polyarsenates, and polyvanadates (with V in fourfold coordination). In Class 9, the primary subdivision is based on the degree of polymerization of the SiO_4 tetrahedra in the structure; the secondary one is based on the presence or absence of additional anions, the coordination number of the cations, and, in certain cases, on the periodicity of the polymerized units. Tertiary subdivision is based on crystal structure. Zeolites are in a separate division and are further subdivided on the basis of structure. The minerals of Class 10 are divided into acetates, oxalates, benzine salts, hydrocarbons, and miscellaneous organic minerals.

See Also

Analytical Methods: Mineral Analysis.

Further Reading

Bayliss P (2000) *Glossary of Obsolete Mineral Names*. Tucson: The Mineralogical Record Inc.

Blackburn WH and Dennen WH (1997) *Encyclopedia of Mineral Names. The Canadian Mineralogist, Special Publication 1*. Ottawa: Mineralogical Association of Canada.

Clark AM (1993) *Hey's Mineral Index*, 3rd edn. London: Chapman & Hall.

Gaines RV, Skinner HCW, Foord EE, *et al.* (1997) *Dana's New Mineralogy*, 8th edn. New York: Wiley and Sons.

Johnsen O (2002) *Minerals of the World*. Princeton, NJ: Princeton University Press.

Kampf A and Gerhold G (eds.) (1998) *The Photo Atlas of Minerals*. CD format. Los Angeles: Gem and Mineral Council, Los Angeles County Museum of Natural History.

Klein C (2002) *The Manual of Mineral Science*, 22nd edn. New York: John Wiley & Sons.

Lima de Faria J (2001) *Structural Classification of Minerals, vol. 1: Minerals with A, A_mB_n and $A_pB_qC_r$ General Formulas*. Dordrecht: Kluwer Academic Publishers.

Lima de Faria J (2003) *Structural Classification of Minerals, vol. 2: Minerals with $A_pB_qC_rD_s$ to $A_pB_qC_rD_sE_xF_yG_z$ . . . General Formulas*. Dordrecht: Kluwer Academic Publishers.

Lima de Faria J (2004) *Structural Classification of Minerals, vol. 3: Minerals with A_pB_q. . .E_xF_y. . .nA_q. General Chemical Formulas and Organic Minerals*. Dordrecht: Kluwer Academic Publishers.

Mandarino JA and Back ME (2004) *Fleischer's Glossary of Mineral Species 2004*. Tucson: Mineralogical Record Inc.

Martin RF (ed.) (1998) *The Nomenclature of Minerals: A Compilation of IMA Reports*. Ottawa: Mineralogical Association of Canada.

Nickel EH and Grice JD (1998) The IMA Commission on New Minerals and Mineral Names: procedures and guidelines on mineral nomenclature, 1998. *Canadian Mineralogist* 36: 1 14.

Strunz H and Nickel EH (2001) *Strunz Mineralogical Tables*, 9th edn. E. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung (Nägele u. Obermiller).

Amphiboles

R A Howie, Royal Holloway, University of London, London, UK

The minerals of the amphibole group differ from the pyroxenes (*see* **Minerals:** Pyroxenes) in having a double-chain silicate structure and in having hydroxyl ions as an essential constituent. Their name is from the Greek *amphibolos* (ambiguous) in allusion to the great variety of compositions and appearances shown by this mineral group. All have a perfect {210} cleavage, and typically range from white to yellow, green and dark green, or blue, the coloured varieties having variable to strong pleochroism.

As in the pyroxene group, both orthorhombic and monoclinic amphiboles occur; their flexibility of ionic replacement is due to their structure. The composition of the simplest calcium-rich amphibole, tremolite, may be expressed by the formula $Ca_2Mg_5[Si_8O_{22}](OH)_2$, the Ca atoms occupying the largest positions of between six- and eight-fold coordination. The general chemical formula for an amphibole can be given as $A_{0\ 1}B_2C_5[T_8O_{22}](OH,F)_2$ where A represents the larger cations, B the cations in *M*4, C the cations in *M*1, *M*2, and *M*3, and T those in the T1 and T2 sites (**Figures 1** and **2**).

The four main substitutions seen in the amphibole structures are:

i. Monovalent ion (Na or K) for vacancy (□) at A.
ii. Monovalent (Na,K) for divalent Ca at B (*M*4 sites).
iii. Trivalent (Al) for tetravalent (Si) at T sites.
iv. Trivalent (Al,Fe^{3+}) for divalent (Mg, Fe^{2+}, etc) at C (*M*1, *M*2, *M*3 sites).

To maintain charge balance, it is necessary for more than one of these substitutions to take place, e.g., in pargasite, $NaCa_2(Mg, Fe^{2+})_4Al[Si_6Al_2O_{22}](OH, F)_2$, where the Na ion is in the A position (type 1 substitution), Al is in the *M* sites (type 4 substitution) and the excess charge is balanced by the substitution of 2Al in the T site (type 3 substitution), the maximum substitution allowed for Si. At the same time, there may be complete substitution between Mg and Fe^{2+} to give ferro-pargasite. A complete range of possible amphibole substitutions was published by Leake and others, and also Deer and others in 1997.

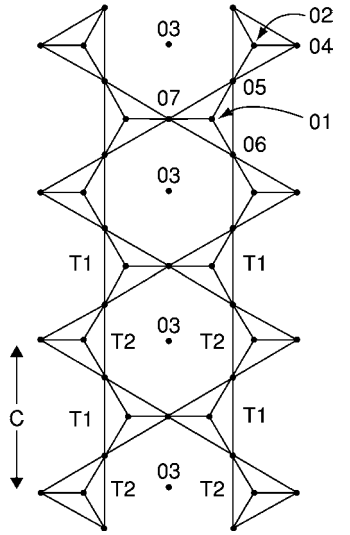


Figure 1 Double chain of linked $(Si,Al)O_4$ tetrahedra characteristic of all amphiboles.

There are, thus, four main groups with boundaries as follows:

i. Iron-magnesium-manganese-lithium amphiboles: $(Ca + Na)_B < 1.00$
ii. Calcic amphiboles: $(Ca + Na)_B \geq 1.00$ and $Na_B < 0.50$
iii. Sodic-calcic amphiboles: $(Ca + Na)_B \geq 1.$ and $0.50 \leq Na_B \leq 1.50$
iv. Sodic amphiboles: $(Na)_B \geq 1.50$

Because of excellent argon retention properties and the common incorporation of potassium in their structures, the amphiboles are particularly useful for K–Ar dating. Hornblende K–Ar ages have been used to date a wide variety of metamorphic and igneous rocks.

Leaving aside the lithium-bearing holmquistite, the two orthorhombic amphiboles are anthophyllite, $(Mg, Fe^{2+})_7[Si_8O_{22}](OH, F)_2$, and gedrite (Mg, Fe^{2+})

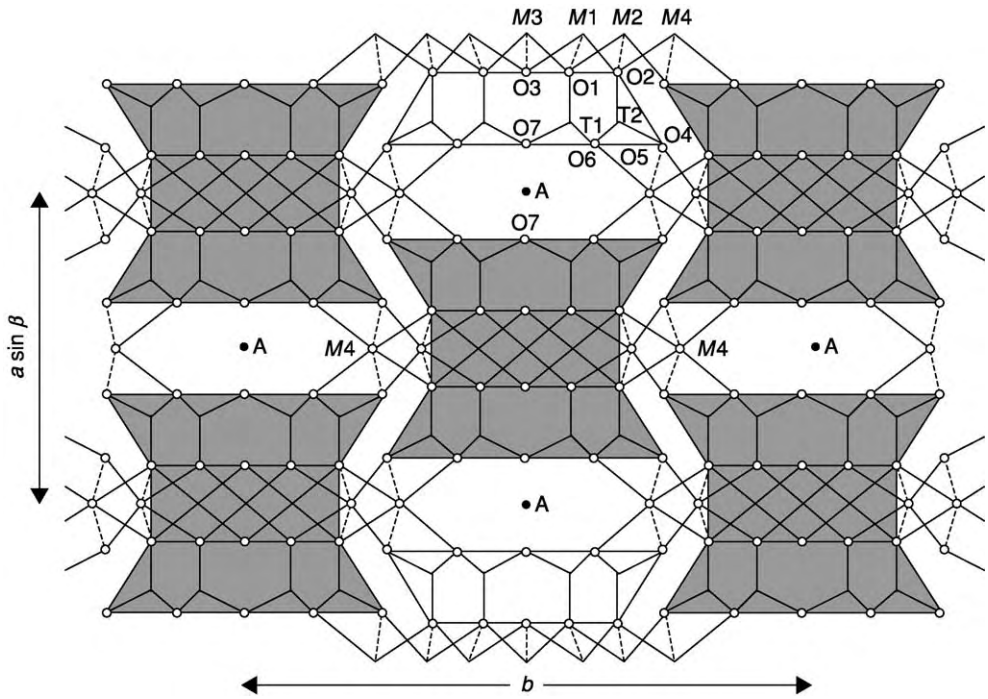


Figure 2 The crystal structure of an amphibole as viewed along *z*. Pairs of chains are linked by cations *M*1, *M*2, *M*3 and *M*4, and in some cases A.

$Al_2[Si_6Al_2O_{22}](OH,F)_2$. The replacement of Mg by Fe raises the refractive indices as does the substitution of Mg,Si by Al,Al; Mg-rich anthophyllites are optically negative, whereas the more Fe-rich anthophyllites and the gedrites are optically positive. In general, these orthorhombic amphiboles are virtually unknown in igneous rocks, but occur in a great variety of metamorphic and metasomatic rocks. The cummingtonite–grunerite series, $(Mg,Fe,Mn)[Si_8O_{22}](OH)_2$, are monoclinic and show characteristic multiple twinning on (100); cummingtonite is optically positive, but the more iron-rich grunerite is optically negative. Cummingtonite occurs in amphibolites derived by the regional metamorphism of basic igneous rocks and in hybrid rocks of intermediate composition. The more iron-rich (and sometimes manganese-rich) grunerites are typical of the banded iron formations of regional metamorphism, where they form a characteristic magnetite–grunerite–quartz association.

In the calcic amphiboles, the tremolite–ferro-actinolite series, $Ca_2(Mg,Fe^{2+})_5[Si_8O_{22}](OH,F)_2$, are colourless to yellow and green, and are essentially metamorphic minerals occurring in both contact and regionally metamorphosed rocks. In the metamorphism of siliceous dolomites, tremolite forms early on by reaction between dolomite and quartz:

$$5CaMg(CO_3)_2 + 8SiO_2 + H_2O \rightarrow$$
dolomite quartz
$$Ca_2Mg_5Si_8Oc\ \text{rocks}\ {}_{22}(OH)_2 + 3CaCO_3 + 7CO_2$$
tremolite calcite

Both tremolite and actinolite are characteristic minerals in low-grade regionally metamorphosed ultrabasic rocks, and actinolite is a common mineral in the greenschist facies.

Hornblende is used to describe a specific calcic amphibole, $Ca_2(Mg,Fe^{2+})_4(Al,Fe^{3+})[Si_7AlO_{22}](OH)_2$, but the continuous chemical variations towards pargasite (see above) and tschermakite, $Ca_2(Mg,Fe^{2+})_3(Fe^{3+}Al)_2[Si_6Al_2O_{22}](OH)_2$, as well as to Al-poor tremolite–ferro-actinolite tend to be called hornblendes in petrographic descriptions, when more appropriate names would pargasitic-, tschermakitic- or ferro-actinolitic-hornblende. This is illustrated in Figure 3, showing the typical chemical variations in analysed calcic amphiboles. The 'hornblendes' are typically pleochroic from pale green to yellow-brown or brown-green and have extinction angles in the range 12–34° (as compared with $\leq 45°$ for the clinopyroxenes). The hornblendes are typical minerals of intermediate plutonic rocks, and occur as products of primary crystallization of igneous rocks

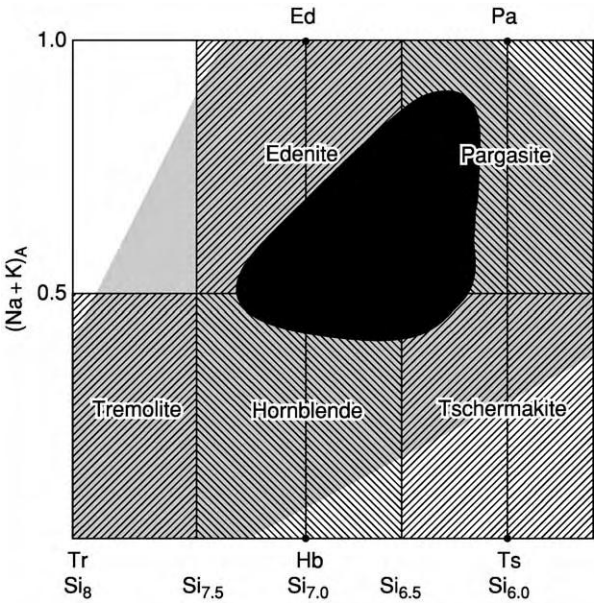


Figure 3 The chemical variations of calcic amphiboles expressed as the numbers of (Na + K) atoms in *A* sites and Si atoms per formula unit. Tr tremolite, Hb hornblende, Ed edenite, Pa pargasite, Ts tschermakite. The more densely stippled areas show the more commonly occurring compositions.

ranging from granites (*see* **Igneous Rocks:** Granite) and syenites to gabbros and ultrabasic rocks. They are neither as common or as abundant in volcanic rocks (due to the loss of volatiles from typical basalts and trachytes); they do, however, occur in a variety of andesites and dacites. Hornblende is also a dominant constituent in many regionally metamorphosed rocks, from the greenschist to the lower parts of the granulite facies.

Kaersutite, $NaCa_2(Mg,Fe^{2+})_4Ti[Si_6Al_2O_{22}](O,OH,F)_2$, is brown to reddish brown under the microscope and is characterized chemically by its high Ti content (TiO_2 5–10 wt.%, equivalent to 0.5–1.0 atom pfu). It is a typical constituent of alkali volcanic rocks, occurring as phenocrysts in trachybasalts, trachyandesites, trachytes, and alkali rhyolites.

Among the sodic amphiboles, glaucophane, in the glaucophane–riebeckite series, $Na_2(Mg,Fe^{2+})_3(Al,Fe^{3+})_2[Si_8O_{22}](OH,F)_2$, is the iron-poor end-member and characteristically shows blue–violet pleochroism in thin section, and may give a blue shade to its rocks, forming typical blueschists as the result of regional metamorphism of basaltic rocks. Riebeckite, the iron-rich equivalent, is an even deeper blue to almost black, and occurs in peralkaline granites, syenites, microgranites and in acid volcanic rocks. The eckermannite-arfvedsonite series, $NaNa_2(Mg,Fe^{2+},Mn)_4(Al,Fe^{3+})[Si_8O_{22}](OH,F)_2$, is even richer in sodium and ranges compositionally between both the Mg and Fe^{2+} and the Al and Fe^{3+}

end-members; they are generally associated with aegirine in peralkaline igneous rocks and have a bluish green to yellow-brown pleochroism.

See Also

Igneous Rocks: Granite. **Minerals:** Definition and Classification; Other Silicates; Pyroxenes. **Rocks and Their Classification**.

Further Reading

Deer WA, Howie RA, and Zussman J (1997) *Rock Forming Minerals:* Vol. 2B. *Double Chain Silicates*, 2nd edn. xii + 764 pp. London: Geological Society.

Leake BE, Woolley AR, Arps CES, *et al.* (1997) Nomenclature of Amphiboles: Report of the Subcommittee on Amphiboles of the International Mineralogical Association Commission on New Minerals and Mineral Names. *Mineral Magazine* 61: 295 321.

Arsenates

K Hudson-Edwards, University of London, London, UK

Crystal Structure

The arsenates are a subclass of the phosphate mineral class, which has a basic chemical unit of $[AO_4]$ with a negative three charge (-3) and a tetrahedral symmetry. In the case of the arsenates, the 'A' in the tetrahedron is the element arsenic (As). These tetrahedra are generally linked to one or more metal–oxygen, –hydroxide, and/or –water octahedra, which in turn are linked by edge- and corner-sharing arrangements. Other weaker bonds exist in the arsenate structures, such as H bonds between the octahedra and arsenate tetrahedra atoms. Some of the arsenates are sheet-like, with intersheet spaces occupied with, and bonded by, cations, halogens, or H_2O molecules. A good example of an arsenate structure is that of the mineral scorodite $[FeAsO_4 \cdot 2H_2O]$, which is composed of Fe(III)–O octahedra that share the oxygen atoms of arsenate tetrahedra. The arsenate tetrahedra are also weakly linked to the Fe(III)–O octahedra through H bonding.

Several dimorphous (same chemical formula, different structure) arsenates exist (e.g., roselite, monoclinic–beta-roselite, triclinic; symplesite, triclinic–parasymplesite, monoclinic).

Chemistry and Nomenclature

There are well over 180 arsenates. They form a diverse subclass of minerals, exhibiting widely varying chemistries and structures, and comprising a number of groups and solid-solution series. Table 1 lists examples of the most common arsenate groups, and their crystal system(s) and point group(s).

A common feature of the arsenates is their incorporation of water or hydroxide into their structure. The amount of water incorporated depends on temperature, vapor pressure, and crystal structure. Some of the more common hydrous and complex arsenates are summarized in Table 2. Those arsenates that have a phyllosilicate-like, layered structure, with interlayer water molecules, are often incorrectly termed 'micas' (e.g., the autunite arsenates are often referred to as 'uranyl micas').

Many of the arsenates form solid-solution series, with the As in the arsenate tetrahedron being replaced by P or V. In fact, many of the groups listed in Table 1 also include phosphates and vanadates (only the arsenate members of the groups are shown in Table 1). For example, the apatite group contains arsenates (e.g., mimetite $Pb_5(AsO_4)_3Cl$), phosphates (e.g., pyromorphite $Pb_5(PO_4)_3Cl$), and vanadates (e.g. vanadinite $Pb_5(VO_4)_3Cl$), and the mixite group contains arsenates (e.g., mixite, $BiCu_6(AsO_4)_3(OH)_6 \cdot 3H_2O$) and phosphate minerals (e.g., petersite $(Ca,Fe,Y,Ce)Cu_6(PO_4)_3(OH)_6 \cdot 3H_2O$). Solid solutions also exist between members of different arsenate groups, where metal cations substitute for one another (the 'X' positions in Table 1). In some cases, the solid solutions are complete, with no miscibility gaps (e.g., annabergite–erythrite), but in others the series are incomplete (e.g., annabergite-köttigite). At least 75 solid solution series involving arsenates have been described.

Physical Properties and Stability

Probably because of their wide-ranging chemistries, the arsenates exhibit a wide range of physical properties. Many of the arsenates are green or a variety of green (bluish, bright, bright apple, emerald, grass, grey, olive, pale, yellowish), but a wide number of colours have been reported, including blue, brown, grey, pink, orange, purple, red, yellow, white, and shades in-between. The colour is often dictated by the incorporation of transition metals such as Ni, Cu, U, and Co into the mineral structure. Similarly,

Table 1 Examples of common arsenate groups

Group	*General formula*	*Examples*	*Crystal system(s)*	*Point group(s)*
Adamite	$X(AsO_4)(OH)$	Adamite	Triclinic	bar 1
	X Cu,Zn,Mn,Zn	Olivenite	Orthorhombic	2/m 2/m
		Paradamite		2/m
Adelite	$XYAsO_4OH$	Austenite	Orthorhombic	2/m 2/m
	X Ca,Pb	Conichalcite		2/m
	Y Cu,Zn	Duftite		2 2 2
Apatite pyromorphite	$X_5(AsO_4)_3(F,Cl,OH)$	Clinomimetite	Hexagonal	6/m
	X Ca,(Ba,Ca,Pb),(Ca,Pb),(Ca,Sr),Pb	Mimetite		
		Hedyphane		
Arthurite	$XFe_2(AsO_4)_2(O,OH)_2\ 2H_2O$	Arthurite	Monoclinic	2/m
	X Cu,Fe,Mn,Zn	Cobaltarthurite		
		Ojuelaite		
Autunite	$X(UO_2)_2(AsO_4)_2\ 8\ 12H_2O$	Heinrichite	Monoclinic	2/m 2/m
	X H,Mg,Fe,Cu,Ca,Ba,Mn,1/2(HAl)	Kahlerite	tetragonal	4/m 2/m
		Uranospinite		2/m
		Zuenerite		bar 4
Beudantite	$XY_3(AsO_4PO_4)SO_4(OH)_6$	Beudantite	Trigonal	
	X Ca,Ba,Ce,Pb,Sr,H_3O	Gallobeudantite		
	Y Fe,Al,Ga	Hidalgoite		
Crandallite	$XAl_3(AsO_4)_2(OH,F)_{1-6}\ 1\ 15H_2O$	Arsenocrandallite	Trigonal	3m 3m
	X (Ca,Sr),(Sr,Ca,Ba),Ba,Pb,(Ce,La),Fe	Philipsbornite		3m 3
				2/m
Meta autunite	$X(UO_2)_{1-2}(AsO_4)_2\ 4\ 8H_2O$	Abernathyite	Tetragonal	4/m 2/m
	X Mg,Fe,Co,Cu,Ca,Ba,K,Na,Zn	Metalodevite		2/m
		Metanovacekite		
Mixite	$XCu_6(AsO_4)_3(OH)_6\ 3H_2O$	Agardite	Hexagonal	
	X Bi,Al,Ca,Ce,La,Y,Nd,Th	Goudeyite		
		Mixite		
Pseuomalachite	$X_{3-5}(AsO_4)_2(OH)_{3-4}$	Cornwallite	Triclinic, monoclinic	1
	X Cu,Mn	Arsenoclasite		2/m
				2 2 2
Roselite	$Ca_2X(AsO_4)2\ 2H_2O$	Brandtite	Monoclinic	2/m
	X Co,Mn,Mg,Zn	Roselite		
		Zincroselite		
Variscite	$XAsO4\ 2H_2O$	Mansfieldite	Orthorhombic	2/m 2/m
	X Al,Fe,In	Scorodite		2/m
		Yanomamite		
Vivianite	$X_3(AsO_4)_2\ 8H_2O$	Annabergite	Monoclinic	2/m
	X Co,Fe,Mg,Ni,Zn	Erythrite		
		Kottingite		
		Hornesite		
		Parasymplesite		

mineral streak is often green or a shade of green (bluish, olive, pale), but grey, pale blue, brown red and yellow, orange yellow, and white are known.

Arsenate crystals tend to range from transparent to translucent, and some opaque varieties have also been documented. Luster is adamantine, dull (in massive or powdery varieties), greasy, pearly, resinous, silky, sub-adamantine, or vitreous. Hardness is generally low, ranging from 1.5 up to 4.5, and specific gravity tends to be slightly below average to heavy for translucent minerals (2.4 to 6.7).

A huge variety of crystal habits have been described for the arsenates. Examples include acicular, botryoidal, fibrous, foliated, micaceous, prismatic, radiating, striated blade and tabular single crystals, aggregates, clusters, or masses (Figure 1). Crystal faces exhibit diverse termination surfaces: domes with bipyramids, triangles, double triangles, pinacoids, and wedges with complex facets. Several varieties are twinned. Most commonly, however, the minerals occur as crusts, globular aggregates, nodules, or powdery masses with little crystal shape. Cleavage ranges from absent to poor or perfect in one or two directions. Fracture is described as conchoidal, fibrous, flaky, or lamellar.

The arsenate adamite ($Zn_2AsO_4(OH)$) is well known for its fluorescent properties. Arsenates belonging to the autinite group (e.g., meta-zeunerite, walpurgite, zeunerite) are radioactive due to the uranium in their structures.

Table 2 Examples of hydrous and complex arsenates

Mineral	*Formula*	*Crystal system*	*Point group*
Adamite	$Zn_2AsO_4(OH)$	Orthorhombic	2/m 2/m 2/m
Arsentsumebite	$Pb_2Cu(AsO_4)(SO_4)OH$	Monoclinic	
Bayldonite	$Cu_3Pb(AsO_4)_2\ H_2O$	Monoclinic	
Carminite	$PbFe_2(AsO_4)_2(OH)_2$	Orthorhombic	2/m 2/m 2/m
Clinoclase	$Cu_3AsO_4(OH)_3$	Monoclinic	2/m
Cuproadamite	$(Cu,Zn)_2(AsO_4)(OH)$	Orthorhombic	2/m 2/m 2/m
Kankite	$Fe(AsO_4)_3\ 5H_2O$	Monoclinic	
Legrandite	$Zn_2AsO_4(OH)\ H_2O$	Monoclinic	2/m
Liroconite	$Cu_2Al(AsO_4)(OH)_4\ 4H_2O$	Monoclinic	2/m
Olivenite	$Cu_2AsO_4(OH)$	Orthorhombic	2/m 2/m 2/m
Paradamite	$Zn_2AsO_4(OH)$	Triclinic	bar 1
Pharmacolite	$Ca(AsO_3OH)\ 2H_2O$	Monoclinic	
Pharmacosiderite	$KFe_4(AsO_4)_3(OH)_4\ 6\ 7H_2O$	Cubic	bar 4 3m
Picropharmacolite	$H_2Ca_4Mg(AsO_4)_4\ 11H_2O$	Triclinic	bar 1
Symplesite	$Fe_3(AsO_4)_2\ 8H_2O$	Triclinic	
Talmessite	$Ca_2Mg(AsO_4)\ 2H_2O$	Triclinic	bar 1
Tilasite	$CaMg(AsO_4)F$	Monoclinic	2/m
Tsumcorite	$Pb(Zn,Fe)_2(AsO_4)_2\ H_2O$	Monoclinic	
Walpurgite	$(BiO)_4UO_2(AsO_4)_2\ H_2O$	Trigonal	

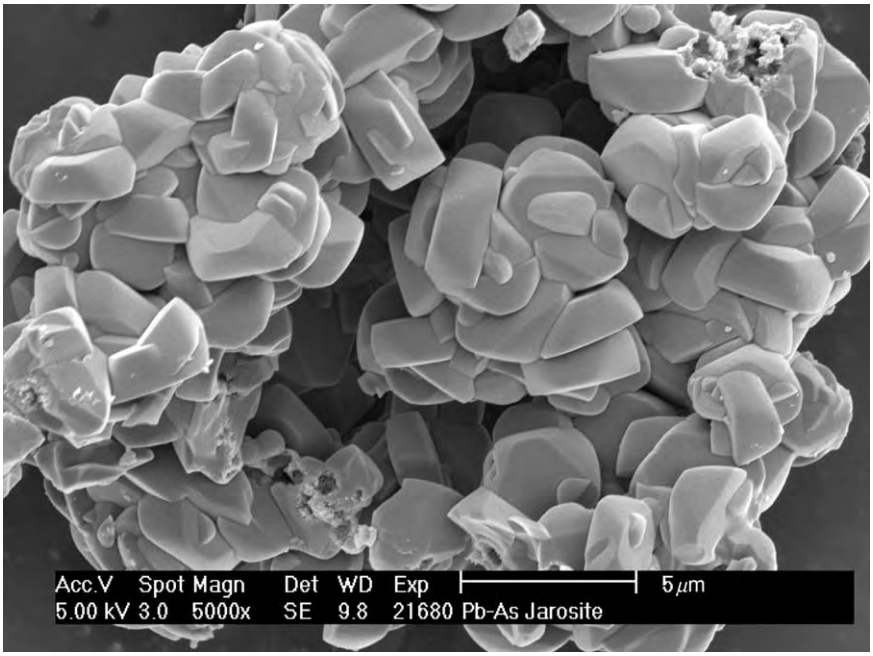


Figure 1 Photomicrograph of synthetic beudantite ($PbFe_3AsO_4SO_4(OH)_6$) crystals, exhibiting tabular cluster crystal shape. Image courtesy of AML Smith.

Solubility and Alteration

The solubilities of the arsenates are dependent on their degree of crystallinity and water content. Many have high log K_{sp} values under standard conditions, and are thus quite soluble (e.g., ferrarisite, guerinite, Table 3). Others, such as annabergite and scorodite, are less soluble and, as a result, tend to control the distribution of arsenic in soils and other environmental media. Scorodite is one of the most common arsenates in soils, but is only stable under acid (pH 4) oxidizing conditions. Under more neutral pH or reducing conditions, scorodite dissolves incongruently to form Fe oxides and soluble As species; amorphous scorodite is thought to be more soluble than crystalline forms (Table 3). Similarly, the arsenate beudantite ($PbFe_3AsO_4SO_4(OH)_6$) dissolves incongruently in acid and

Table 3 Solubilities of selected arsenates at 25°C, 1 bar

Mineral	*Dissociation reaction*	*Log K_{sp}*
Annabergite	$Ni_3(AsO_4)_2\ 8H_2O \quad 3Ni^{2+} + 2AsO_4^{3-} + 8H_2O$	28.38
TriCa arsenate	$Ca_3(AsO_4)_2\ 4H_2O \quad 3Ca^{2+} + 2AsO_4^{3-} + 4H_2O$	21.257
Calcium arsenate hydrate	$Ca_4(OH)_2(AsO_4)_2\ 4H_2O \quad 4Ca^{2+} + 2H_3AsO_4 + 6H_2O \quad 8H^+$	40.2
Ferrarisite	$Ca_5(HAsO_4)_2(AsO_4)2\ 9H_2O \quad 5Ca^{2+} + 4H_3AsO_4 + 9H_2O \quad 10\ H^+$	27.75
Guerinite	$Ca_5(HAsO_4)_2(AsO_4)2\ 9H_2O \quad 5Ca^{2+} + 4H_3AsO_4 + 9H_2O \quad 10\ H^+$	28.55
Scorodite, amorphous	$FeAsO_4\ 2H_2O \quad Fe^{3+} + AsO_4^{3-} + 2H_2O$	22.89
Scorodite, crystalline	$FeAsO_4\ 2H_2O \quad Fe^{3+} + AsO_4^{3-} + 2H_2O$	25.89

alkali solutions, yielding As(V) ions, α-FeOOH, $Fe(OH)_3$, and $PbSO_4$.

One of the arsenates, meta-zeunerite, forms by dehydration and pseudomorphing of the parent arsenate zeunerite.

Occurrence

Most of the arsenates are rare to very rare. They are prized by mineral collectors because of their range of spectacular colours, lusters, and crystal habits. The arsenate balydonite, found in Cornwall, England, has even been used to craft cabochons (convex gems and beads) and cabinets.

Many arsenates are found in the oxidation zone of sulphide ore deposits or in oxidized mine waste, where they occur as discrete precipitates or coatings on other mineral grains. They are associated with sulphide minerals such as arsenopyrite, chalcopyrite, sphalerite, and galena, and with other arsenates and secondary ore minerals such as iron oxides, azurite, and malachite. The most common of these arsenates include annabergite, beudantite, conichalcite, hörnesite, parasymplesite, rauenthalite, and, particularly, scorodite. Although not common, two of the most famous of the oxidation zone arsenates are erythrite and annabergite, known by miners as 'cobalt bloom' and 'nickel bloom', respectively. These are used as indicator minerals of Co and Ni sulphide ores, and as ores of Co and Ni themselves. Several other arsenates are also minor ores of metals; examples are agardite (REE), beudantite (Pb), chalcopyllite (Cu), conichalcite (Cu), köttigite (Zn), mimetite (Pb), and walpurgite (U and Bi). These mine-related arsenates have been found at well-known mines in Devon and Cornwall, England; Mexico (Mapimi); Cobalt, Ontario, Canada; Sweden (Långban, Warmland); Nambia (Tsumeb); Alsace, France; Arizona; Romania; Russia; and Zaire. The arsenate nealite ($Pb_4Fe(AsO_4)_2Cl_4$) is also mine-related, but has been found in Greek and Roman slags in Lavrio, Greece, rather than outcrops, leading to some controversy over its classification as a mineral, since by definition, minerals should be natural, rather than man-made.

The arsenates scorodite and beudantite have been reported frequently in sulphide tailings. In uranium tailings, Ca–arsenates occur as coprecipitates with gypsum and as amorphous precipitates. $Ca_3(AsO_4)$ is the arsenates generally thought to control As solubility in aqueous systems where Ca^{2+} is present, but $Ca_4(OH)_2(AsO_4)_2 \cdot 4H_2O$ and $CaHAsO_4 \cdot H_2O$ have also been suggested to occur in tailings with high Ca/As ratios and alkaline pH (10) conditions, and low Ca/As ratios (0.88–1.00) and slightly acid pH (5.76–6.22) conditions, respectively.

Arsenic occurs most commonly in soil waters as the complexes H_2AsO_4 (low pH) and $HAsO_4^2$ (higher pH). The species $H_3AsO_4^0$ and AsO_4^3 occur in extremely acidic and alkaline conditions, respectively. Sources of these complexes are many and varied, and include pesticides, mining, munitions waste, wood preservatives, and tannery wastes. The arsenate complex is sorbed onto common soil minerals, such as Fe oxides and clays, or forms discrete arsenates, many of which are listed in Tables 1 and 2. The type of arsenate formed in soils depends on the availability of other cations (e.g., Ca, K, Mg, H, Pb) in the soils. Many of the hydrous arsenates (Tables 1, 2) are found in soils (e.g., carminite, kankite, pharmacosiderite, talmessite, tilasite, and members of the vivianite group).

Arsenates such as scorodite have also been reported to occur as crusts in hot spring deposits. Due to its insolubility, scorodite is also used in metallurgy for the disposal of arsenic wastes from metallurgical effluents and flue dusts.

See Also

Economic Geology. **Environmental Geochemistry**. **Mineral Deposits and Their Genesis**. **Minerals:** Vanadates.

Further Reading

Bothe JV and Brown PW (1999) As immobilization by calcium arsenate formation. *Environmental Science Technology* 33: 3806 3811.

Bothe JV and Brown PW (1999) The stabilities of calcium arsenates. *Journal of Hazardous Materials* 69: 197 207.

Donahue R and Hendry MJ (2003) Geochemistry of arsenic in uranium mill tailings, Saskatchewan, Canada. *Applied Geochemistry* 18: 1733 1750.

Dove PS and Rimstidt JD (1985) The solubility and stability of scorodite, $FeAsO_4 \cdot 2H_2O$. *American Mineralogist* 70: 838 844.

Frost RL (2004) An infrared and Raman spectroscopic study of the uranyl micas. *Spectrochimica Acta A* 60: 1469 1480.

Frost RL, Kloprogge R, Weier ML, *et al.* (2003) Raman spectroscopy of selected arsenates implications for soil remediation. *Spectrochimica Acta Part A* 59: 2241 2246.

Frost RL, Weier ML, Martents W, Kloprogge JT, and Ding Z (2003) Thermal decomposition of the vivianite arsenates implications for soil remediation. *Thermochimica Acta* 403: 237 249.

Kitahama K, Kiriyawa R, and Baba Y (1975) Refinement of the crystal structure of scorodite. *Acta Crystallographica Section B* 31: 322 332.

Krause E and Ettel VA (1988) Solubility and stability of scorodite, $FeAsO_4 \cdot 2H_2O$: New data and further discussion. *American Mineralogist* 73: 850 854.

Krause E and Ettel VA (1989) Solubilities and stabilities of ferric arsenate compounds. *Hydrometallurgy* 22: 311 337.

Langmuir D and Mahoney J (1998) Appendix 3 McClean Lake Project, JEB Tailings Management Facility, Construction License Additional Information, Geochemistry, Sub Appendix F Thermodynamic Data for Selected Species and Slids of Arsenic and Nickel. Cogema Resources Inc.

Mahoney J (2002) *The Corrected Solubility Product of Scorodite and Its Application to Arsenic Behaviour in Buried Mine Tailings*. GSA Program with Abstracts, Denver Annual Meeting, October 27 30, 2002, Paper 84 17.

Parkhurst DL and Appelo CAJ (1999) User's guide to PHREEQC A computer program for speciation, batch reaction, 1D transport and inverse geochemical calculations. US Geological Survey Water Resources Investigations Report 99 4259.

Rimstidt JD, Chermak JA, and Gagen PA (1994) Rates of reaction of galena, sphalerite, chalcopyrite, and arsenopyrite with Fe^{3+} in acidic solutions. In: Alpers CN and Blowes DW (eds.) *Environmental Geochemistry of Sulfide Oxidation*. p. 2 13. Washington, DC: American Chemical Society.

Robins RG (1981) The solubility of metal arsenates. *Metallurgical Transactions B* 12: 103 109.

Robins RG (1987) Solubility and stability of scorodite, $FeAsSO_4 \cdot 2H_2O$. *American Mineralogist* 72: 842 844.

Smith AML (2004) Mechanisms and products of the breakdown of contaminant element bearing jarosites. PhD thesis, Birkbeck, University of London.

Borates

C Helvaci, Dokuz Eylül Üniversitesi, İzmir, Turkey

Introduction

Borate has a very long history. Derived from the Persian burah (boorak), borax was known to the Babylonians, who brought it from the Himalayas some 4000 years ago for use in the manufacture of rings, amulets, and bracelets. The Egyptians used borax in mummifying, and around AD 300 the Chinese were familiar with borax glazes, as were the Arabs three centuries later. Borax was first brought to Europe in the thirteenth century, presumably by Marco Polo, and it has been supplied since by traders from Tibet and Kashmir.

By the 1770s the French had developed a source of tincal, the old name for crude borax, in Purbet Province, India, and at about the same time natural boric acid (sassolite) was discovered in the hot springs in the Maremma region of Tuscany, Italy. The middle of the nineteenth century was a particularly active time for the discovery and commercial development of borate deposits. In particular, Chile started to mine the borate resources of the Salar de Ascotan in 1852, and within a few years output accounted for a quarter of the world's annual supply of 16 000 tonnes. In 1856 John Veatch discovered borax in Clear Lake, Lake County, California, and this led eventually to the formation of the California Borax Company in 1864 and to California's dominance of the borate industry. In Turkey modern borate mining began in 1865 when the Compagnie Industrialle des Mazures mined borates from the Aziziye Mine near Susurluk and shipped the ore to France for processing. Demand encouraged the exploitation of large-scale deposits in Turkey and the USA and overwhelmed more modest producers.

Borates are among the most interesting of the world's industrial minerals; they were used first in precious-metal working and later in ceramics. They are an unusually large group of minerals, but the number of commercially important borates is limited, and their chemistry and crystal structure are both unusual and complex. There are only a few large deposits, although there are many non-commercial occurrences in other rocks and brines. The accounts of the early exploration, mining, and processing of

borates are fascinating, because the remote locations of the deposits often led to unusual difficulties and hardships in recovering the desired products. These varied from workers wading into Himalayan lakes to harvest the 'floor' and then transporting the borax in saddlebags on sheep across the Himalayas to the markets, to the 'Dante's Inferno' of the Larderello boric-acid fumaroles, and the colourful 20 mule teams of the western USA. Such operations transformed borates from expensive minor minerals into the large-volume industrial commodities they are today. Boron's chemistry and reactivity are also fascinating because it forms oxygen compounds in an essentially unending variety of simple to exceedingly complex molecules. Determining the crystal structures of these compounds has given rise to a separate subfield of crystallography. The boron isotopes ^{10}B and ^{11}B have very different reactivities during both physical and chemical changes, which has allowed them to be used to understand many geological and other events, again forming a specialized field in geology.

A borate is defined as any compound that contains or supplies boric oxide (B_2O_3). A large number of minerals contain boric oxide, but the three that are most important from a worldwide commercial standpoint are borax, ulexite, and colemanite. These are produced in a limited number of countries (Figure 1), and production is dominated by the USA and Turkey, which together furnish about 90% of the world's borate supplies. Production in the USA is concentrated in the Mojave Desert of California: borax and kernite are mined from the large deposit at Boron. Borate-containing brines are pumped from Searles Lake, and a limited amount of colemanite is mined in Death Valley. Turkish production is controlled by Eti Maden (Eti mine), the national mining enterprise, which supplies most of the commercially traded ulexite and colemanite from mines in the Bigadiç and Emet districts, together with borax from the huge deposit at Kyrka.

Borate minerals have been used in many ways since at least the eighth century, when they were used primarily as a flux for assaying and refining gold and silver. Their valuable properties and relative rarity soon stimulated an international trade in borates. Borates were traded at relatively high prices for highly specialized applications until the late nineteenth century. At that time they were being used in medicines, food preservatives, ceramic glazes, and as metal fluxes. Borates are often defined and sold according to their boric oxide or B_2O_3 content, and most statistical data are listed in tonnes of B_2O_3. Borax pentahydrate and boric acid are the most commonly traded commodities. Boric acid plants are operated by all the major borate producers. Glass-fibre insulation is the major end use in the USA, followed by textile glass fibre and borosilicate glass, detergents, and ceramics. Detergents continue to be a major end use in Europe.

Geology and Mineralogy

Over 150 boron-bearing minerals have been identified, the most common being sodium, calcium, or magnesium salts. Table 1 gives a more complete list of the common borates. However, just four minerals – borax, ulexite, colemanite, and datolite – are commercially significant today. Borax or tincal, a natural sodium borate decahydrate, is the major commercial source of boron, with major supplies coming from the USA, Argentina, and Turkey. The principal commercial mixed sodium–calcium borate, ulexite, is produced in Turkey and several countries in South America, whereas large-scale production of the main calcium borate, colemanite, is restricted to Turkey.

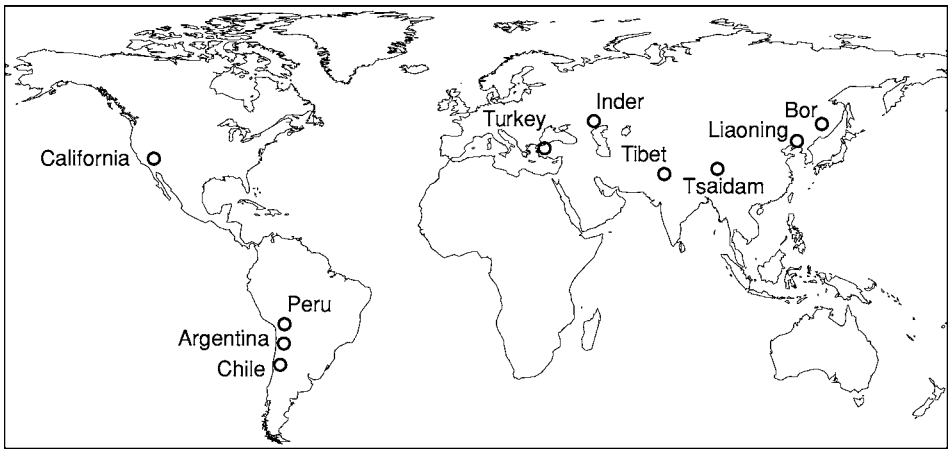


Figure 1 Current world resources of borates.

Table 1 Composition of principal boron minerals

Mineral	*Empirical formula*	*B_2O_3 Content (Weight %)*
Sassolite	$B(OH)_3$ or B_2O_3 $3H_2O$	56.4
Borax (tincal)	$Na_2B_4O_7$ $10H_2O$	36.5
Tincalconite	$Na_2B_4O_7$ $5H_2O$	48.8
Kernite	$Na_2B_4O_7$ $4H_2O$	51.0
Ulexite	$NaCaB_5O_9$ $8H_2O$	43.0
Probertite	$NaCaB_5O_9$ $5H_2O$	49.6
Priceite (Pandermite)	$Ca_4B_{10}O_{19}$ $7H_2O$	49.8
Inyoite	$Ca_2B_6O_{11}$ $13H_2O$	37.6
Meyerhofferite	$Ca_2B_6O_{11}$ $7H_2O$	46.7
Colemanite	$Ca_2B_6O_{11}$ $5H_2O$	50.8
Hydroboracite	$CaMgB_6O_{11}$ $6H_2O$	50.5
Inderborite	$CaMgB_6O_{11}$ $11H_2O$	41.5
Kurnakovite	$Mg_2B_6O_{11}$ $15H_2O$	37.3
Inderite	$Mg_2B_6O_{11}$ $15H_2O$	37.3
Szaibelyite (ascharite)	$Mg_2B_2O_5$ H_2O	41.4
Suanite	$Mg_2B_2O_5$	46.3
Kotoite	$Mg_3B_2O_6$	36.5
Pinnoite	MgB_2O_4 $3H_2O$	42.5
Boracite (strassfurite)	$Mg_3B_7O_{13}Cl$	62.2
Datolite	$Ca_2B_2Si_2O_9$ H_2O	21.8
Cahnite	Ca_2AsBO_6 $2H_2O$	11.7
Danburite	$CaB_2Si_2A_8$	28.3
Howlite	$Ca_4Si_2B_{10}O_{23}$ $5H_2O$	44.5
Vonsenite (paigeite)	$(Fe, Mg)_2FeBO_5$	10.3
Ludwigite	$(FeMg)_4Fe_2B_2O_7$	17.8
Tunnellite	SrB_6O_{10} $4H_2O$	52.9

Production of datolite, a silicate mineral, is confined to Russia.

Boron minerals may be divided for convenience into three broad groups according to their origin and geological environments: first, skarn minerals related to intrusives (mainly silicates and iron oxides); second, magnesium oxides related to marine sediments; and, third, hydrated sodium and calcium borates related to continental sediments and volcanic activity (Figure 2).

In the first and second groups are the Russian and some of the Chinese sources, the major minerals being datolite and szaibelyite. In the third group are borax, kernite, colemanite, and ulexite, which provide the sources for most of the production from Turkey, and Argentina (Table 2). Minerals occurring in Turkey and the USA are shown in Table 3.

Table 2 Commercial borate minerals

Minerals	*Emprical formula*	*B_2O_3 content (weight %)*
Colemanite	$Ca_2B_6O_{11}$ $5H_2O$	50.8
Ulexite	$NaCaB_5O_9$ $8H_2O$	43.0
Borax	$Na_2B_4O_7$ $10H_2O$	35.5
Kernite	$Na_2B_{14}O_7$ $4H_2O$	51.0
Pandermite	$Ca_4B_{10}O_9$ $7H_2O$	49.8
Hydroboracite	$CaMgB_6O_{11}$ $6H_2O$	50.5

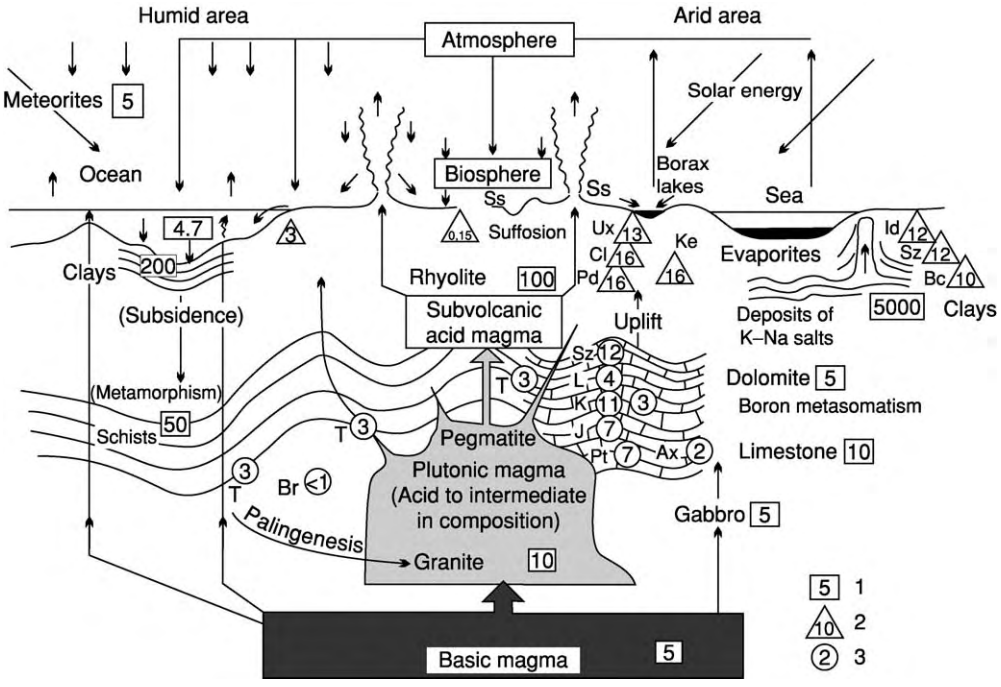


Figure 2 Scheme for the cycle and concentration of boron. The geochemical abundance of boron in rocks is cited according to Goldschmidt, Landergren and Harder; 1, average content of boron in rocks (g tonne ^{1}); 2, exogenic concentration of boron (%); 3, endogenic concentration of boron (%); Ss, sassolite; Ux, ulexite; Cl, colemanite; Pd, pandermite; Ke, kernite; In, inderite; Bc, boracite; Sz, szaibelyite; L, ludwigite; K, kotoite; J, jimboite ($Mn_3(BO_3)_2$); Dt, datolite; Ax, axinite; T, tourmaline; Br, braunite.

Table 3 Borate minerals occurring in Turkish and Argentine Miocene borate deposits

			Turkish borate deposits					*Argentine borate deposits*		
	Mineral	*Formula*	*Bigadiç*	*Sultançayir*	*Kestelek*	*Emet*	*Kirka*	*Tincalayu*	*Sijes*	*Loma Blanca*
Ca borates										
	Inyoite	$Ca_2B_6O_{11}\ 13H_2O$	+				+	+	+	+
	Meyerhofferite	$Ca_2B_6O_{11}\ 7H_2O$	+			+	+		+	
	Colemanite	$Ca_2B_6O_{11}\ 5H_2O$	+	+	+	+	+		+	+
	Tertschite	$Ca_4B_{10}O_{19}\ 20H_2O$	+							
	Pandermite (priceite)	$Ca_4B_{10}O_{19}\ 7H_2O$	+	+			+			
	Nobleite	$CaB_6O_{10}\ 4H_2O$							+	
	Gowerite	$CaB_6O_{10}\ 5H_2O$							+	
	Ginorite	$Ca_2B_{14}O_{23}\ 8H_2O$						+		
Ca/Na borates										
	Ulexite	$NaCaB_5O_9\ 8H_2O$	+		+	+	+	+	+	+
	Probertite	$NaCaB_5O_9\ 5H_2O$	+		+			+	+	
Na borates										
	Borax	$Na_2B_4O_7\ 10H_2O$					+	+		+
	Tincalconite	$Na_2B_4O_7\ 5H_2O$					+	+		+
	Kernite	$Na_2B_4O_7\ 4H_2O$					+	+		
	Ezcurrite	$Na_4B_{10}O_{17}\ 7H_2O$						+		
	Ameghinite	$NaB_3O_5\ 2H_2O$						+		
Other borates										
(Mg)	Hydroboracite	$CaMgB_6O_{11}\ 6H_2O$	+		+	+	+		+	
(Mg)	Inderborite	$CaMgB_6O_{11}\ 11H_2O$					+		+	
(Mg)	Inderite	$Mg_2B_6O_{11}\ 15H_2O$					+	+		
(Mg)	Kurnakovite	$Mg_2B_6O_{11}\ 15H_2O$					+	+		
(Mg)	Rivadavite	$Na_6MgB_{24}O_{40}\ 22H_2O$	+					+		
(Mg)	Mcallisterite	$Mg_2B_{12}O_{20}\ 15H_2O$						+		
(Sr)	Tunellite	$SrB_6O_{10}\ 4H_2O$	+			+	+			
(Sr)	Veatchite A	$Sr_4B_{22}O_{37}\ 7H_2O$				+				
(As)	Teruggite	$Ca_4MgAs_2B_{12}O_{28}\ 20H_2O$				+				+
(As)	Cahnite	$Ca_2BAsO_6\ 2H_2O$				+				
(Si)	Howlite	$Ca_4Si_2B_{10}O_{23}\ 5H_2O$	+	+						
(Si)	Bakerite	$Ca_8B_{10}Si_6O_{35}\ 5H_2O$		+						
(Si)	Searlesite	$NaBSi_2O_6\ H_2O$						+		
(Na Mg)	Aristarainite	$Na_2MgB_{12}O_{20}\ 8H_2O$						+		
(Ca Sr)	Estroncioginorite	$(Sr,Ca)_2\ B_{10}O_{17}\ 7H_2O$						+		

+, present; , not present.

Borax is by far the most important mineral for the borate industry. It crushes freely and dissolves readily in water; its solubility and rate of solution increase with water temperature. Borax in large tonnages is present in the deposits at Boron (California), Kyrka (Turkey), and Tincalayu (Argentina). Kernite is present in minor amounts at Kyrka and Tincalayu, but it makes up about a third of the total reserve at Boron and has a higher B_2O_3 content than borax.

Colemanite is the calcium-bearing borate preferred by the non-sodium fibreglass industry. It has low solubility in water, although it dissolves readily in acid. Some colemanite is used in European chemical plants to produce boric acid because the supply from Turkey provides B_2O_3 at the lowest cost. Turkey is the world's major source of high-grade colemanite. The USA has important reserves in the Death Valley area, but only limited amounts are produced there at this time. Colemanite is not known to occur in major deposits outside Turkey and North America, although the higher hydrate, inyoite, is mined on a limited scale in Argentina.

Ulexite is the usual borate found on or near the surface, in playa lakes and marshes of Holocene to Quaternary age throughout the world, where it occurs as soft, often damp, masses of fibrous crystals. These 'cotton balls' or 'papas' are collected in large salars in South America and China. Ulexite of Neogene age is mined in Turkey and occurs at Boron and Death Valley in the USA. It is well lithified and, therefore, is hard, dense, and commonly well-bedded.

Szaibelyite (ascharite) is a major source of both Chinese and Russian borate. It is a magnesium borate and, like colemanite, has low solubility in water. Although it is less satisfactory, owing to its magnesium content, for most uses than either borax or colemanite, substantial tonnages are used in Eastern Europe, Russia, and Asia; it is not traded internationally as a mineral concentrate on a major scale.

The Russians also produce substantial amounts of borate from skarn borosilicates, mainly datolite, with some reports of minor amounts of danburite, ludwigite, and tourmaline. These minerals must be liberated, concentrated, and then dissolved in acid to make a usable product because their natural melting points exceed those of the other minerals used in common glass furnaces.

Pandermite (priceite) was mined in Turkey, and hydroboracite was mined in Russia and Argentina. Other minerals such as inyoite, howlite, meyerhofferite, and kurnakovite are found intimately associated with the major ores. Boracite was used in Germany prior to 1945, where it and minor magnesium borates were recovered as a by-product of potash mining.

A large number of silicates contain boron in their lattices in varying amounts. Axinite, suanite, kotoite, and others are listed in the literature as occurring with the Russian borosilicate ores. Sassolite has only a mineralogical interest at most occurrences, as the quantity found is generally very small. In the Lardarello region of Italy, however, natural steam carries boric acid recoverable as sassolite, and for a long period prior to 1965 several thousand tonnes per year were produced.

Chemistry

Boron is the fifth element of the periodic table and is the only electron-deficient non-metallic element. Thus, boron has a high affinity for oxygen, forming strong covalent boron–oxygen bonds in compounds known as borates. Boron is also the only light element with two abundant isotopes, B^{10} and B^{11}; the former has a large capture cross-section, which makes it an excellent neutron absorber. Within the borate group of minerals, BO_3 units can polymerize (in a similar way to the SiO_4 tetrahedral groups in the silicates) to form chains, sheets, and isolated multiple groups. This is possible because the small B^{3+} ion, which generally coordinates three oxygens in a triangular group, has bond strengths to each oxygen ion with an e.v. of 1; this is exactly half the bonding energy of the oxygen ion. This permits a single oxygen to be shared between two boron ions, linking the BO_3 triangles into expanded structural units (double triangles, triple rings, sheets, and chains). Because of the triangular coordination, boron is also found in tetrahedral groups. In addition to BO_3 and BO_4 groups, natural borates may contain complex ionic groups such as $[B_3O_3(OH)_5]^2$, which consists of one triangle and two tetrahedra. In the structure of colemanite, $CaB_3O_4(OH)_3 \cdot H_2O$, complex finite chains of tetrahedra and triangles occur, and in borax, $Na_2B_4O_5(OH)_4 \cdot 8H_2O$, a complex ion, $[B_4O_5(OH)_4]^2$, consisting of two tetrahedra and two triangles is found. Borates can be classified on the basis of the structural anionic linking into insular (independent single or double BO_3 or BO_4 groups), chain, sheet, and framework structures.

Although it is possible to prepare a three-dimensional framework made up of BO_3 triangles only and having the composition B_2O_3, such a configuration has a very low stability and disorders readily, yielding a glass. Because of its tendency to form somewhat disordered networks of BO_3 triangles, boron is regarded as a 'network former' in glass manufacture and is used in the preparation of special glasses with light weight and high transparency to energetic radiation.

The chemistry of borates is somewhat analogous to, but more complex than, that of silicates. The boron atom in borates can be either three- or four-coordinate bonded to oxygen, forming either planar trigonal BO_3 units or negatively charged tetrahedral BO_4 units; crystalline sassolite or boric acid, $B(OH)_3$, has only BO_3 units. Other borate minerals are typically salts containing a mixture of BO_3 and BO_4 units. The nomenclature for borates can be confusing because three different formula systems are in common use. For example, ulexite can be described by an empirical formula, $NaCaB_5O_9 \cdot 8H_2O$, an oxide formula, $Na_2O \cdot 2CaO \cdot 5B_2O_3 \cdot 16H_2O$, and a crystal-structure formula, $NaCa[B_5O_6(OH)] \cdot 5H_2O$. Only the crystal-structure formula conveys information about how the atoms are organized in the crystal.

Many borate minerals contain waters of hydration and, with the exception of borax, are stable at ambient conditions. Borax dehydrates to tincalconite under ambient conditions unless the relative humidity is high. Upon heating, borates lose their attached water molecules before fusing to form a melt. The common borate salts melt at low temperatures; the melts are excellent fluxes or solvents for dissolving other more refractory oxides.

Only alkali-metal borates, ammonium borates, and boric acid have appreciable solubility in water. In dilute aqueous solutions, borates exist as an equilibrium mixture of two species: non-ionized molecular boric acid molecules, $B(OH)_3$, and metaborate ions, $B(OH)_4$. For example, borax dissolves in water to give a 1:1 mixture of $B(OH)_3$ and $B(OH)_4$; it is a good buffer system at pH 9.4.

$$B(OH)_3 + 2H_2O \ \{\text{Revers React}\} \ B(OH)_4 + H_3O^+$$

Most borate minerals dissolve in strong mineral acids to liberate the weaker boric acid; for example, boric acid can be produced by mixing ulexite with sulphuric acid. Boric acid is a weak Lewis acid and electron pair acceptor, with a solution pH of about 5.

Crystalline borate minerals are made up of both trigonal and tetrahedral borate units. These borate units can link through boron–oxygen–boron bonds to form chain polyborates, or they may join into rings, which can contain three or four boron atoms (i.e. triborates and tetraborates). These rings can in turn link together to form pentaborates, hexaborates, and higher borates.

Non-borate boron-containing compounds (those containing no boron–oxygen bond) are not found in nature. Such compounds include elemental boron, borides, boron halides, boranes, and organoboron compounds. In general, these manmade materials may be stable inert solids or may hydrolyze in the presence of moisture or oxidize in the presence of oxygen to form the more stable borates.

Borates are widely distributed in nature in low concentrations as alkali-metal and alkaline-earth borate and borosilicate minerals and less commonly as boric acid. They are typically found in soil and rock in concentrations of up to about 450 ppm total boron, distributed in over 150 minerals, primarily as salts of sodium, calcium, and magnesium. The average concentration of boron dissolved in land-surface water is about 0.1 ppm, and in seawater the concentration is about 4.6 ppm.

At low levels, water-soluble boron is an essential micronutrient for the growth and viability of plants; the range between insufficient boron and excess boron is narrow (0.25–15.0 ppm boron), but most soils fall within this range. Borate is found in animal tissues at about 1 ppm as a result of the ingestion of fruits and vegetables; it is not known to have an essential biochemical function, although it may play a role in the body's ability to use calcium. Borate transported in plants and animals is usually complexed with polyalcohols in the aqueous phase.

Depositional Setting and Formation of Borate Deposits

Boron is extremely dispersed in nature, averaging 0.1 ppm in land-surface water, 3 ppm in the Earth's crust, and 4.6 ppm in seawater. There are relatively few occurrences where the element is sufficiently concentrated to be economic. Where a degree of concentration does occur, it is usually a result of local volcanic activity (as a source of boron), a body of water such as a lake (to dissolve boron compounds), evaporative conditions (to concentrate the solution to the point of precipitation), and the deposition of a protective layer of sediment (to preserve the highly soluble borate minerals).

Major borate deposits throughout the world are found in tectonically active extensional regions associated with plate boundaries. Most of the commercial borate deposits in the USA, South America, and Turkey are thought to be associated with continental sediments and volcanism of Neogene age. Many of the older skarn deposits also appear to be related to continental volcanic sources. Marine borate deposits are apparently the product of evaporation of seawater in a restricted basin, probably associated with a seafloor borate source, and/or progressive decanting that preferentially concentrated the borates to many times natural seawater concentrations. Borates associated with igneous and some metamorphic rocks are thought to be an end phase of specialized magmatic segregation or leached from the intruded rocks by

associated hydrothermal fluids (*see* **Mining Geology:** Hydrothermal Ores).

Most of the South American deposits are associated with calcareous tuff, which occurs as a late-stage capping over the borates, and in some cases with halite and gypsum. Recent volcanic activity is indicated by basaltic to rhyolitic flows in adjacent areas, and a volcanic source for the borates is presumed. The Salar deposits of South America consist of beds and nodules of ulexite, with some borax or inyoite, associated with Holocene playa sediments, primarily mud, silt, halite, and gypsum.

Borates can be concentrated in different ways:

- by chemical precipitation in the neighbourhood of boron-bearing springs in playa-type basins (e.g. the Boron, Searles Lake, and Billie deposits of California, and the Kirka, Sultançayir, Bigadiç, and Emet deposits in Turkey);
- by precipitation from seawater in the closing phase of a salt-forming evaporate cycle (e.g. at Stassfurt in Germany);
- by contact metasomatism with dolomite or magnesite, forming magnesium borates such as ludwigite, kotoite, and ascharite (e.g. the Teazhnoe deposits at Yakutia in Russia, and the Hol-Kol deposit in North Korea);
- by contact metasomatism with limestone, forming boron silicates such as datolite and danburite (e.g. at Ak Akhdar, Pamir, and Dalnegorskoye, Primorsky, Russia); and
- by volcanic exhalations of boric acid, i.e. sublimates (e.g. at Clear Lake, Lake County, California, at Salar de Surire, Chile, and in the Maremma area, Tuscany, Italy).

The formation of borate deposits can be tentatively summarized as follows.

- Although boron is a rare element (average content in the Earth's crust is 10 ppm), extraordinary concentrations can be found in certain places.
- Over 150 minerals are known to contain boron, and they are found in various geological environments. These can be subdivided into three groups:
  1. a skarn group associated with intrusives and consisting of silicates and magnesium-iron oxides;
  2. a magnesium oxide group hosted by marine evaporitic sediments; and
  3. a sodium– and calcium–borate hydrates group associated with lacustrine (playa lake) sediments and explosive volcanic activity.
- The following conditions are essential for the formation of economically viable borate deposits in playa-lake volcanosedimentary rocks (Figure 3):

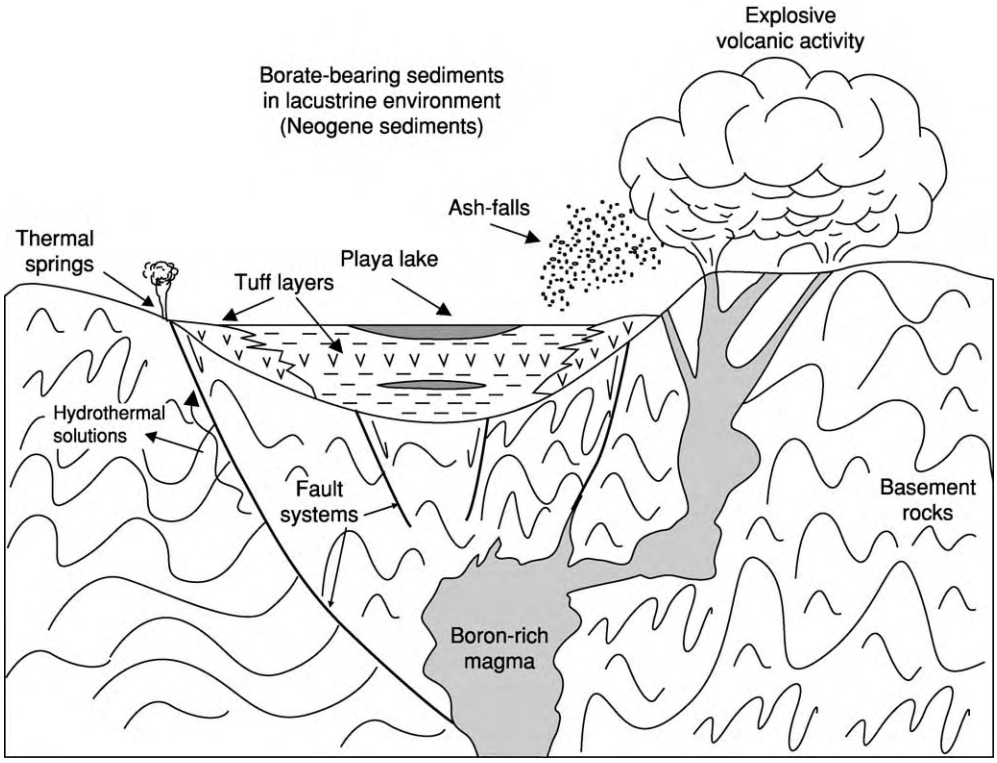


Figure 3 Generalized playa lake depositional model, showing the formation of borate deposits in the Neogene basins of western Anatolia, Turkey.

1. formation of playa-lake environment;
2. concentration of boron in the playa lake, sourced from andesitic to rhyolitic volcanics, direct ash fall into the basin, or hydrothermal solutions along graben faults;
3. thermal springs near the area of volcanism;
4. arid to semi-arid climatic conditions; and
5. lake water with a pH of between 8.5 and 11.

Borate Deposits in Non-Marine Basins

The largest known borate deposits originated as chemical precipitates and are found interbedded with clays, mudstones, tuffs, limestones, and similar lacustrine sediments. There is evidence that most of these deposits were closely related in time to active volcanism. Thermal springs and hydrothermal solutions associated with this volcanic activity are therefore regarded as the most likely source of the boron.

Several South American springs in volcanically active areas are depositing borates, and the first borax discovered in the USA was found in the muds associated with warm springs at Clear Leak, California, a volcanically active area. The Italian (Tuscany) steam vents from which sassolite was recovered also represent an active volcanic source.

In addition to the concentrated source of the borates and a 'basin' in which they can collect, an arid to semiarid climate also seems to be essential during the deposition and concentration of economic amounts of soluble borates. These soluble borates can, in the long run, be preserved only by burial; however, the lack of deposits of soluble borates older than mid-Tertiary may indicate that even burial is not able to protect borates over long periods of geological time.

Hydrated borates may accumulate in several ways within a non-marine basin. They may be deposited in layers in a spring apron around a borate spring, with ulexite, borax, or inyoite as the primary borate mineral. Borates may also form in a pool dominantly fed by a borate spring, with borax crystals formed in bottom muds or at the intermittently dried margins (as at Clear Lake, and at Salar de Surire, Chile). If the spring flow is low or intermittent, evaporation develops a surface efflorescence or precipitate or an accumulation of crystals just below the surface (examples are the marsh or playa deposits that have been mined in California and Nevada and some of the salar deposits of South America). Finally, there are lake deposits, whose occurrence requires much more than seasonal flooding (examples are the borax deposits at Boron and Kirka, which were formed by chemical precipitation in a closed basin). There is some disagreement as to whether the thicker ulexite deposits, such as those in Turkey and Death Valley, are spring-apron or lake deposits. In fact, there is probably a gradation between spring-apron and small-lake deposits. These borate lakes are essentially monomineralic, in the sense that no other salts occur in major quantities. There is, however, another type of borate lake deposit, consisting of mixed salts and/or brine containing borates in sufficient quantity of justify recovery. Searles Lake in California has been cited as a type example of a multicomponent deposit formed by the evaporation of lake waters. Numerous studies of Searles Lake have concluded that boron and the other dissolved constituents originated along the eastern front of the Sierra Nevada, were concentrated and decanted in a series of up-drainage lakes, and were finally precipitated and preserved in Searles Lake itself. This scenario would also fit several of the Chinese and Tibetan lakes. The borates found in the large South American salars, such as Uyuni and Atacama, may also have formed by leaching of the surrounding rocks and subsequent evaporation, although the role of local mineralized spring waters containing boron has not been fully evaluated.

Marine Evaporites

Borates of marine origin have been found in commercial quantities only in Europe. These are magnesium borates associated with Permian salt deposits. They were produced in Germany, as a by-product of potash mining, and in the Inder region of Kazakhstan. The Inder deposits, where the borates occur as veins in the cap of a very large salt dome complex, are thought to have been remobilized and concentrated from the salt during the intrusion of the salt dome itself. Some of the Chinese deposits of the Liaoning Peninsula may be of similar origin, although they occur as veins in Precambrian metamorphosed limestone and magnesite.

The Inder Lake brines, which are also a source of Kazakhstan borate, appear to be simply a sump accumulation of borates leached from the huge Inder salt dome complex. The Kara-Bagaz-Gol Lagoon borates on the east shore of the Caspian Sea appear to have leached from marine brines.

Magmatic Sources

Pegmatites and contact-metamorphic rocks contain assemblages of various boron-containing minerals, such as datolite, ludwigite, paigeite, and tourmaline. These represent concentrations of boron that relate more or less directly to the crystallization of intrusive granitic magma. Analyses show that granites average about 10 ppm boron, with a few exceptions ranging up to 300 ppm. However, boron does not readily enter into the crystal structures of the common

rock-forming minerals; hence, when magma crystallizes, the boron is thought to leave with the released water. The combination of high temperatures (300–400°C) and fluids under high pressure at an intrusive contact could also extract some boron from the adjacent country rocks. These borate skarn deposits, some of which are associated with iron ores and magnesium deposits of commercial grade, are mined in both eastern Russia and China.

Exploration

Borate exploration consists of detailed prospecting of favourable areas followed by drilling, and uses all the tools available to the exploration geologist. The recognition of trends of favourable host rocks and structures is an important guide to areas that are of possible interest. Satellite imagery, both real and false-colour, and standard photographic interpretation can be successfully used under certain conditions.

In most parts of the world, the identification of a Cenozoic suite of non-marine fine-grained sediments and tuffs is the usual starting point for the field geologist, because most commercial borates are associated with these rocks. Any accumulation of salts in these sediments is examined closely. Because many borates are associated with volcanic rocks, volcanic centres, flows, ash deposits, and tuffs, particularly if they are zeolite-bearing, may also be favourable guides to borate prospecting.

Borate exploration begins with visits to Neogene basins that appear to be favourable for borate deposition. It should be remembered that the mobility of boron in the Earth's crust is very high, like chlorine, bromium, sulphur, etc. Because of its rapid alteration by water and carbon dioxide, borate can be altered to calcite and boric acid washed away as follows:

$$2CaO \cdot 3B_2O_3 \cdot 5H_2O(\text{colemanite}) + 2H_2CO_3(\text{carbonic acid}) + 2H_2O \rightarrow 2CaCO_3(\text{calcite}) + 6H_3BO_3(\text{boric acid})$$

The following criteria, based on the criteria outlined above in the section on the formation of borate deposits, are used during exploration in Turkey, the USA, and Argentina:

- contribution of Miocene volcanics and volcanoclastics to Neogene playa-lake basins;
- formation of capping limestone;
- claystone–limestone–marl intercalations;
- evidence of hot springs or hydrothermal solutions carrying boron into the playa lake; and
- evaporitic horizons or an indication of leaching of them.

The skarn borates of Eastern Europe and Asia were found by careful prospecting in geologically well-preserved fold belts where limy sediments are in contact with potassic to alkaline volcanics. Datolite and danburite occur in skarns where the limestone was originally rich in calcium and silica; axinite, kotoite, and ludwigite skarns are hosted by dolomitic limestones. The magnesium borates and borates associated with iron ores are generally of low grade. Marine borates are sought in tectonically stable areas with shallow to outcropping salt structures capped by gypsum, where near-surface borate deposits are identified by whitish soil cover and scanty vegetation.

In arid regions, ulexite often accumulates at, or just beneath, the current surface of salt flats and playas, indicating that boron is moving within the system. These recent crusts may also indicate brine deposits containing boron concentrations of interest. In either case, additional prospecting is usually justified; playas are usually prospected by pitting on a 100–500 m grid. Springs and recent spring deposits containing anomalous borates may also be used as a guide to the presence of ore in certain areas.

Geochemical surveys are useful for narrowing down prospect areas to a drill target. Both soil and rock-chip sampling techniques are used in exploration programmes, with boron, strontium, arsenic, and lithium as a common suite of elements. Beryllium is also used in the search for skarn borates in Russia, as are complex boron-magnesium-calcium-chlorine ratios. Sampling of both surface and well water may be useful. Certain plants are boron sensitive, and vegetation surveys may prove interesting.

Geophysical surveys, particularly gravity and magnetic, are used to outline target basins or structures beneath sedimentary basin fill. Resistivity and seismic surveys have been used to define basin structures and formations that may be associated with the borates in that area. Various down-hole well-logging techniques, including natural gamma and neutron probes, can indicate the approximate percentage of borates and clay in zones of special interest.

Geological mapping followed by drilling is still the definitive test in most areas of the world. While rotary drill methods may be used, cores are generally taken of the most prospective zones. Assays of B_2O_3 and other associated elements (arsenic, lithium, and strontium) are then run on the horizons that appear favourable for borates. Because the saline borates are water soluble, short core runs are used, but the common borates generally core well, with recoveries of more than 90%.

In areas of doubt, two easy field tests can determine the presence of borates. The original flame test, where the mineral is soaked with sulphuric acid and alcohol

and ignited ('she burns green, Rosie. . .') is still one of the most diagnostic. The turmeric test, where hydrochloric acid and turmeric solution turn the specimen red-brown, is also used. As chemical tests performed in the laboratory are more diagnostic and accurate, most field samples are sent to the laboratory for analysis.

Under current economic conditions, bedded deposits of borax, colemanite, and ulexite are not generally sought at depths greater than 500 m. Brines with a high borate content, particularly those associated with other salts of value, might be extracted from greater depths in certain circumstances. The skarn and magnesium borates are currently economical only from surface and near-surface excavations.

Mining and Mineral Processing

Most of the world's commercial borate deposits are mined by open pit methods. The world's major borate operations – the Boron mine of US Borax at Boron (Kramer) in California and the Kyrka mine of Eti Holding in Turkey – are huge open pit mines using large trucks and shovels and front-end-loader methods for ore mining and overburden removal. Ores and overburden are drilled and blasted for easier handling. The Boron operation uses a belt conveyor to move ore from the in-pit crusher to a coarse ore stockpile from which it is reclaimed by a bucket wheel that blends the ore before it is fed to the refinery. Kirka uses trucks to haul material to a crusher near the refinery, which is about 0.5 km from the current ore faces.

Smaller operations in Argentina, Chile, China, Turkey, and Russia use similar methods, but on a smaller scale appropriate to the scale of the operation. Some of the South American and Chinese salar operations use hand labour to mine the thin salar borates, generally after stripping the overburden with a small bulldozer or front-end loader.

Borates are mined by underground methods in the Liaoning area of north-eastern China and at the Billie and Gerstley mines in Death Valley, California. Borate brines are recovered at Searles Lake in California and in the Qinghai Basin of China; brines are also used in the Inder region of Kazakhstan. Borate-containing brines from several salars in South America are being considered for production. Also under study is the recovery of low-grade borates by in situ acid leaching, producing a crude boric acid for subsequent refining.

Processing techniques depend on both the scale of the operation and the ore type, with either the upgraded refined mineral (borax, colemanite, ulexite) or boric acid as the final product for most operations.

Borax–kernite ores (Boron, Kyrka, Tincalayu) are crushed to less than 2.5 cm and then dissolved in hot water or recycled borate liquor. The resultant strong liquor is clarified and concentrated in large countercurrent thickeners, filtered, fed to vacuum crystallizers, centrifuged, and dried. The final product is refined borax decahydrate or pentahydrate or fused anhydrous borax, or is used as feed for boric acid production.

The ulexite from most of the South American salars is air dried, screened, and bagged. It is then combined with locally available sulphuric acid to produce a relatively low-grade boric acid or exported as feed for boric acid plants elsewhere.

Colemanite concentrates are used directly in specific glass melts or as a feed for boric acid plants. The magnesium borates are generally concentrated, dissolved in acid to remove the magnesium, and converted to boric acid or sodium borates. The borosilicates of the Bor deposit in eastern Russia, with their relatively low B_2O_3 grades, are crushed and then run through a complex plant, which includes magnetic separators, heavy-media separators, and flotation cells. The concentrates are then dried, leached, and calcined before being converted to boric acid or a sodium borate.

Brines from Searles Lake, and presumably the Chinese sources, are recovered by either controlled evaporation or carbonation. In the latter process, carbon dioxide produced from lime kilns or flue gas is bubbled through the brine to crystallize sodium bicarbonate; borax is then crystallized in vacuum crystallizers. In the 'evaporation' process, a rapid controlled cooling electively crystallizes the various salts. The remaining borate liquor is fed to tanks containing borax seed crystals, which aid the recovery of borates from the liquor. The resultant slurry is filtered, washed, redissolved, and fed to vacuum crystallizers, which produce dehydrate borax products or boric acid.

Boric acid is one of the final products produced from most of the processes (Table 4). The world's largest boric acid facility is located adjacent to the Boron pit. It uses either refined borates or hydrated kernite as feed, with sulphuric acid as the reacting

Table 4 Commercial refined borate productions

Product	*Formula*	*%B_2O_3*
Borax decahydrate	$Na_2B_4O_7 \cdot 10H_2O$	30.5
Borax pentahydrate	$Na_2B_4O_7 \cdot 5H_2O$	47.8
Boric acid	H_3BO_3	56.3
Borax anhydrous	B_2O_3	100.0
Sodium perborate	$NaBO_3 \cdot 4H_2O$	22.0
Raw borax anhydrous	$Na_2B_2O_3$	69.2

agent. Other smaller facilities around the world use smelter acid or other locally available acid feed-stock to produce products acceptable to their local markets.

Uses

The principal uses of borates have not changed much in the past decade. Major uses include fibreglass, insulation, textile or continuous-filament glass fibres, glass, enamels and frits, and fertilizers (Figure 4). Bleaches and detergents are also a major end use; however, demand for borates for glass and glass fibres, including fibreglass, is increasing. Boron fibre-reinforced plastics are used in quantity for aerospace frame sheathing, where they combine flexibility and light weight with strength and ease of fabrication. Relatively minor uses that are expected to increase in the near future include fertilizers, wood preservatives, alloys and amorphous metals, fire and flame retardants, and insecticides. However, the promising field of boron–iron–silicon electrical transformers has not developed as rapidly as predicted owing to various cost factors.

Miscellaneous uses include pharmaceuticals, cosmetics, anticorrosion compounds, adhesives, abrasives, insecticides, metallurgical processes, and nuclear shielding. Research is still continuing in many areas. One of the more publicized applications is in super-magnets, where borates, combined with rare earths, nickel, and iron, produce an alloy that can be used to make electromagnets for computer drives, high-fidelity speakers, automobile starter motors, and various household appliances. Another new field is biological growth-control agents.

Borates can be used to protect the environment by aiding the conversion of heavy metals in industrial waste streams into recoverable free metals and by removing impurities from polymers used to bleach wood pulp for paper production. Borates also help to control the refractive index in optical fibres for medical research, where precise control is needed. Other medical applications include treatments for cancer, in which the ^{10}B isotope reacts with low-energy neutrons to give off short-range alpha particles that can be used for microsurgery in previously inoperable areas of the brain. Current tests on boron analogues indicate that they may be effective in reducing serum cholesterol and other disease-causing proteins.

Borates have become a relatively modestly priced industrial mineral commodity in recent years, following the development of the large deposits at Boron and, more recently, Kyrka. Prices are directly related to the cost of production, of which the major part is the cost of fuel for drying, dehydrating, and melting the refined ore into the products desired by industry. Industry prices for most products have increased only in line with inflation.

Both Western Europe and Japan, neither of which have local borate sources, are major markets for US and Turkish production. South America is largely self sufficient, and is producing an increasing excess, which is exported mainly to Europe and Japan. Russia and China both appear to be self sufficient in borates

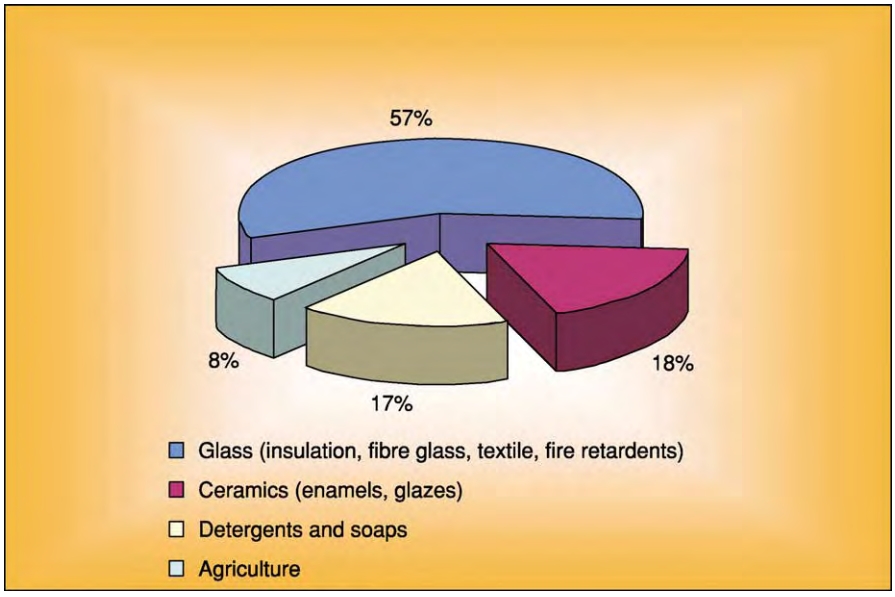


Figure 4 World borate end uses.

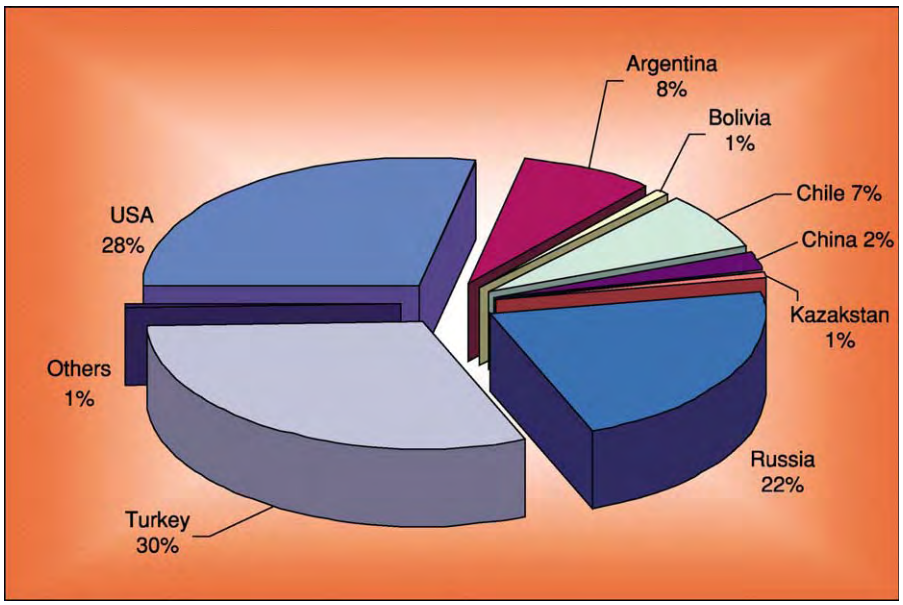


Figure 5 World borate production.

at this time, although their costs of production per tonne of B_2O_3 are thought to be relatively high.

Borates are a lightweight commodity and are generally sold in bulk by carload lots, in intermediate bulk containers or 'super-sacks', and by palletized bags. Overseas shipments are mostly made in bulk from special terminals at Los Angeles, California and from Bandirma on the Marmara Sea in Turkey to similar terminals in the Netherlands, Belgium, and the UK, from where they are moved by barge, as well as by rail and truck. Other imports to Europe arrive in Italy and Spain. Imports to the Far East are generally sold in small bag lots. Bulk imports to the USA (mainly colemanite) usually land in Charleston, South Carolina, where there are grinding facilities; this colemanite is then shipped to eastern fibreglass manufacturers. There is no import duty on borates brought into the USA.

Boron consumption is directly related to the use of glass, glass fibres, and ceramics. These materials, along with certain plastics that contain borate products, are seen as having a steady consumer demand in the construction and houseware markets well into the next century. Other major uses, such as detergents, plant foods, and wood preservation, are expected to show a slowly rising demand. Total world borate demand is expected to grow at about 3% per annum in the near future, based on industry forecasts.

Known reserves of borate minerals are large, particularly in Turkey, South America, and the USA, and production from Turkey and the USA will continue to dominate the world market. However, borates from other areas will probably take up an increasing share of the world market (Figure 5). This trend is already evident, with boric acid from Chile reaching the Far East and Europe, and both Russia and China beginning to export (Table 5).

Table 5 The reserves and life estimates of the world's borate deposits

Country	*Known economic reserve (million tonnes of B_2O_3)*	*Total reserve (million tonnes of B_2O_3)*	*Estimated life of known reserve (years)*	*Estimated life of total reserve (years)*
Turkey	224 000	563 000	155	389
USA	40 000	80 000	28	55
Russia	40 000	60 000	28	69
China	27 000	36 000	19	25
Chile	8000	41 000	6	28
Bolivia	4000	19 000	3	13
Peru	4000	22 000	3	15
Argentina	2000	9000	1	6
Kazakstan	14 000	15 000	10	10
Total	363 000	885 000	253	610

There are few substitutes for borates. In most applications, they provide unique chemical properties at a reasonable price; this is particularly true for glass fibres and in the field of heat- and impact-resistant glass. Borates are an essential part of certain plant foods. Their use in nuclear-reactor shielding and control is well documented. Future markets are difficult to predict. Based upon recent history, the

major world consumers of borates will continue to be the developed countries of North America, Europe, and Japan.

See Also

Geysers and Hot Springs. **Minerals:** Definition and Classification. **Mining Geology:** Exploration; Mineral Reserves; Hydrothermal Ores. **Sedimentary Environments:** Lake Processes and Deposits. **Sedimentary Rocks:** Evaporites.

Further Reading

Aristarain LF and Hurlbut CS Jr (1972) Boron minerals and deposits. *Mineralogical Record* 3: 165 172, 213 220.

Floyd PA, Helvaci C, and Mittwede SK (1997) Geochemical discrimination of volcanic rocks, associated with borate deposits: an exploration tool? *Journal of Geochemical Exploration* 60: 185 205.

Foshag W (1921) The origin of the colemanite deposits of California. *Economic Geology* 16: 194 214.

Garrett DE (1998) *Borates. Handbook of Deposits, Processing, Properties, and Use*. London: Academic Press.

Grew ES and Anovita LM (eds.) (1996) *Boron. Mineralogy, Petrology and Geochemistry*. Reviews in Mineralogy 33. Washington DC: Mineralogical Society of America.

Helvaci C (1978) A review of the mineralogy of the Turkish borate deposits. *Mercian Geology* 6: 257 270.

Helvaci C (1986) Geochemistry and origin of the Emet borate deposits, western Turkey. *Faculty of Engineering Bulletin, Cumhuriyet University, Series A. Earth Sciences* 3: 49 73.

Helvaci C (1984) Occurence of rare borate minerals: veatchite A, tunelite, terggite and cahnite in the Emet borate deposits, Turkey. *Mineralium Deposita* 19: 217 226.

Helvaci C (1989) A mineralogical approach to the mining, storing and marketing problems of the Turkish borate production. *Geological Engineering* 34/35: 5 17.

Helvaci C (1995) Stratigraphy, mineralogy, and genesis of the Bigadiç borate deposits, western Turkey. *Economic Geology* 90: 1237 1260.

Helvaci C and Orti F (1998) Sedimentology and diagnesis of Miocene colemanite ulexite deposits (western Anatolia, Turkey). *Journal of Sedimentary Research* 68: 1021 1033.

Helvaci C and Alonso RN (2000) Borate deposits of Turkey and Argentina: a summary and geological comparison. *Turkish Journal of Earth Sciences* 24: 1 27.

Helvaci C and Firman RJ (1976) Geological setting and mineralogy of Emet borate deposit, Turkey. *Transactions/Section B, Institute of Mining and Metallurgy* 85: 142 152.

Kistler RB and Helvaci C (1994) Boron and borates. In: Carr DD (ed.) *Industrial Minerals and Rocks*, 6th edn, pp. 171 186. Littleton, CO: Society for Mining, Metallurgy and Exploration Inc.

Muessig S (1966) Recent South American borate deposits. In: Rau JL (ed.) *Proceedings of 2nd Symposium on Salt, Northern Ohio Geological Society, Cleveland*, vol. 1, pp. 151 159.

Palmer MR and Helvaci C (1995) The boron isotope geochemistry of the Kirka borate deposit, western Turkey. *Geochimica et Cosmochimica Acta* 59: 3599 3605.

Palmer MR and Helvaci C (1997) The boron isotope geochemistry of the Neogene borate deposits of wetsern Turkey. *Geochemica et Cosmochimica Acta* 61: 3161 3169.

Travis NJ and Cocks EJ (1984) *The Tincal Trail. A History of Borax*. London: Harrap.

Watanebe T (1964) Geochemical cycle and concentration of boron in the Earth's crust. *Verdenskii Institute Geochemistry and Analytical Chemistry USSR* 2: 167 177.

Carbonates

B Jones, University of Alberta, Edmonton, AB, Canada

Introduction

The term 'carbonates' refers to any sediments or sedimentary rocks that are formed primarily of $CaCO_3$ and/or $CaMg(CO_3)_2$. Carbonate sedimentary rocks, known from strata of all ages, house up to 50% of the world's oil and gas reservoirs and are commonly used for building stone, building aggregates, and cement production. Their chemical simplicity belies the bewildering array of textures evident in these rocks and the complexity of the processes that govern their formation. Indeed, the origins of many of these rocks, their constituents, and their fabrics are still matters of ongoing debate.

Carbonate sediments form and accumulate in a broad spectrum of depositional environments (Figure 1). Although most commonly associated with shallow tropical seas (Figure 2), carbonate sediments and precipitates can also form in oceans, lakes, streams, caves, and cold or hot springs. Animals and plants are multifaceted actors in the formation of these sediments – their calcareous skeletons contribute to the sediment, they promote precipitation through modification of the local environment, and they trap and bind sediments.

Figure 1 Carbonate depositional settings. (A) Marine environment with a marginal reef separating the forereef from a lagoon; oblique aerial view of a reef system on the coast of Caicos Island. (B) Hot spring environment with carbonate deposits formed of dendritic calcite crystals; spring KL6 at Loburu, Lake Bogoria, in the Kenyan Rift Valley. Photo courtesy of Robin Renaut. (C) Lake environment, showing giant oncoids (20 30 cm) in the early Mid Holocene Galana Boi Formation, Lake Turkana, Kenya; wind blown sand covered the oncoids after the lake regressed. Photo courtesy of Robin Renaut. (D) Lake environment with large towers of tufa that formed around springs vents on a lake floor; the towers formed when the lake level was considerably higher than it is today (highest tower is ~5 m high); Mona Lake, California. (E) Cave environment, showing stalactites formed of calcite hanging from the roof of Crystal Cave, Bermuda.

Constituent Minerals

Solid $CaCO_3$ exists in three different crystal structures (polymorphs), namely, calcite (trigonal system), aragonite (orthorhombic system), and vaterite (hexagonal system). Vaterite is very rare and needs no further consideration. Aragonite is a common component of many carbonate sediments despite the fact that it is metastable at the low temperatures

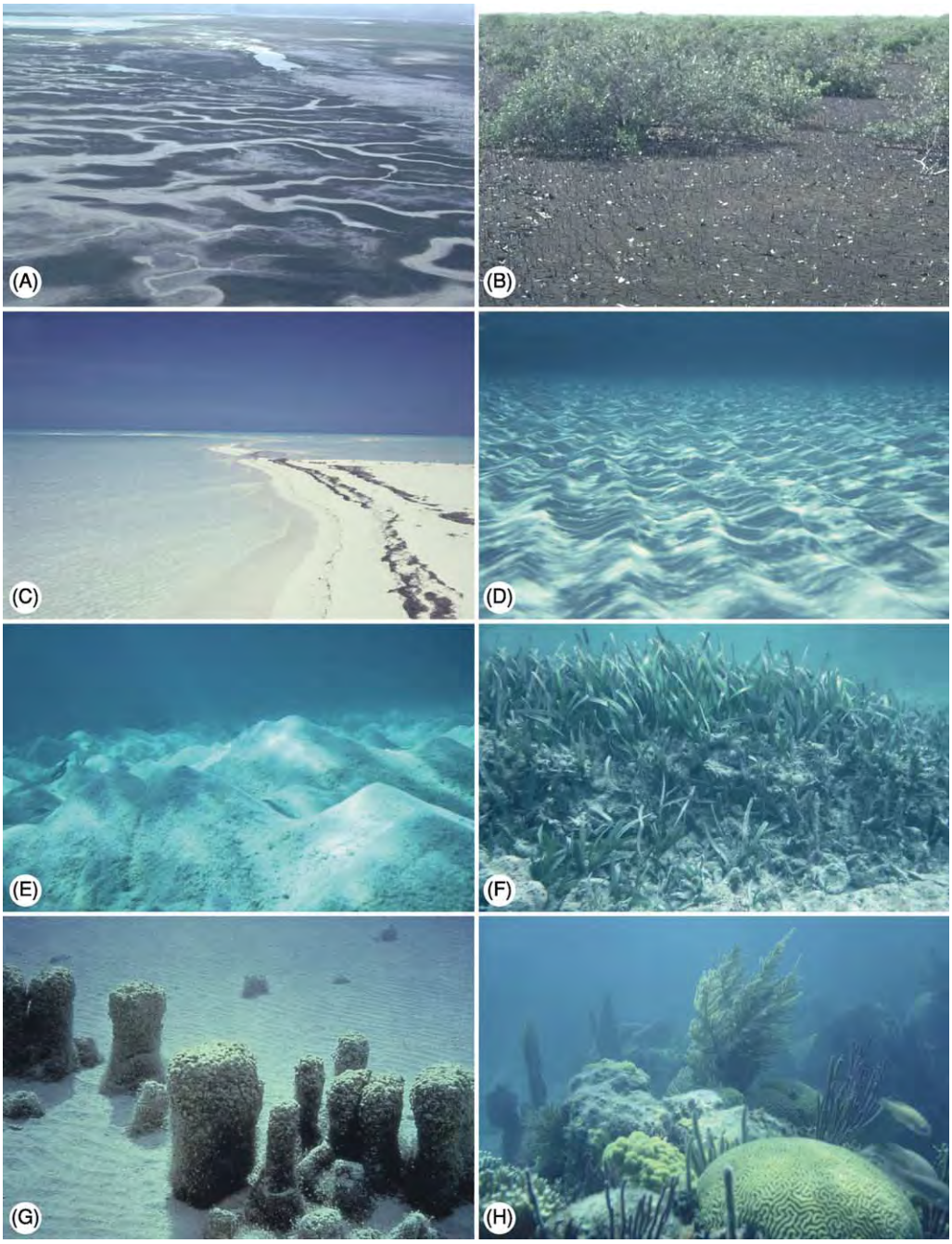


Figure 2 Marine carbonate depositional environments. (A) Intertidal and supratidal flats cut by numerous tidal creeks; aerial view, Caicos Island. (B) Intertidal flat covered by thick algae mat and mangroves; intertidal flat, Caicos Island. (C) Beach and shallow subtidal setting covered with ooid sands; Joulters Key, Bahamas. (D) Shallow subtidal seafloor covered with rippled ooid sands; Joulters Key, Bahamas. (E) Lagoonal sands with numerous mounds created by the burrowing ghost shrimp, *Callianassa*; water depth ~6 m, East Sound, Grand Cayman. (F) Shallow lagoon with *Thalassia* (seagrass), which has an extensive root system that binds sediment; water depth ~0.6 m, Pease Bay Lagoon, Grand Cayman. (G) Lagoon with stromatolites growing on the seafloor; columns ~1 m high, Shark's Bay, Australia. Photo courtesy of Ian Hunter. (H) Small patch reef formed of corals and other organisms; water depth ~5 m, near North Rock, Bermuda.

and pressures found at Earth's surface. With time, aragonite will invert to calcite, the more stable polymorph, by processes that are still poorly understood.

Dolomite, $CaMg(CO_3)_2$, a common component of many carbonate rocks, typically forms through the replacement of a precursor limestone. The origin of dolomite has been hotly debated ever since the term

was introduced over 200 years ago. The fact that it cannot be produced in the laboratory under low-temperature and low-pressure conditions has seriously impeded the understanding of its genesis. Calcite and dolomite are both prone to isomorphic substitutions. Magnesium commonly substitutes for calcium in calcite to produce magnesium calcite. Calcite, with <4 mol% $MgCO_3$, is referred to as low-magnesium calcite (LMC), whereas calcite, with >8 mol% $MgCO_3$, is referred to as high-magnesium calcite (HMC). Similarly, Fe commonly substitutes for Ca or Mg in calcite and dolomite to produce ferroan calcite and ferroan dolomite. Such substitutions are important because they affect crystal lattice structures, which can have a profound effect on the solubility and recrystallization of these minerals during diagenesis.

Components

Carbonate sediments are typically formed of various chemical or biochemical precipitates that are held in a matrix or cement (Figure 3). These precipitates, termed allochems, form within a basin and are organized into discrete aggregated bodies that have usually undergone some type of transport. A matrix refers to any mechanically deposited sediment (e.g., mud) that is deposited between the allochems, whereas cement refers to crystals that have been passively precipitated from pore fluids between the allochems. Non-marine carbonates typically, but not always, contain fewer allochems than do their marine cousins. Most cave deposits and hot-spring deposits, for example, are formed of crystalline calcite and aragonite that have been precipitated directly from water. Bizarre and unusual crystal morphologies commonly characterize these deposits.

Allochems may be coated grains (including ooids; Figure 2C and D), pellets/peloids, bioclasts, and lithoclasts. Although most coated grains are <2 mm in diameter (e.g., ooids), some of those found in lacustrine and spring settings are up to 0.5 m in diameter (Figure 1C). Depending on the environment in which the coated grain forms, the nucleus can be anything from a small sand grain to a large coral fragment to an animal bone. The cortex may be formed of radial crystals and/or laminae that are symmetrically or asymmetrically arranged around the nucleus. The cortical laminae may originate through organic processes and/or inorganic processes. The plethora of names applied to coated grains commonly reflects their size, internal structures, and/or the organism that mediated their formation. Common coated grains include ooids (<2 mm diameter), pisoids (>2 mm diameter), rhodolites (growth mediated by red algae), and oncolites (growth mediated by algae).

Faecal pellets, typically ovate and formed of micrite, are produced by animals (e.g., sea cucumbers, gastropods, worms) that extract organic material from the vast quantities of sediment that they ingest and process through their digestive tract. Although typically <1 mm long, some of the largest sea cucumbers produce faecal pellets up to 1 cm in diameter. Faecal pellets, which are coated with mucus when excreted by the animal, are delicate structures prone to breakdown and disintegration. Compaction can, for example, transform a faecal pellet grainstone into a featureless micrite. Not all ovate grains formed of micrite, however, are faecal pellets. Such grains, produced by micritization of skeletal calcite grains, are known as peloids.

Bioclasts are grains of any size that are derived from the calcareous skeletons of animals and plants (Figure 3A, B, and D). The composition and size of bioclasts depend on the skeletal architecture of the original animal/plant and/or the amount of postmortem physical and biological degradation that it underwent. Compositional variance arises because different animals and plants have skeletons formed of aragonite (e.g., corals, bivalves, calcareous algae), LMC (e.g., brachiopods, bryozoans, trilobites), or HMC (e.g., crinoids, echinoids, red algae). Some animals even have the ability to produce skeletons formed of alternating layers of aragonite and calcite. Tropical sediments, dominated by aragonite, are highly reactive, with inversion of the metastable aragonite to calcite being the dominant process. By comparison, temperate carbonates, which are dominated by LMC and HMC, are less prone to change because they contain only minor amounts of metastable aragonite. Bioclastic sediments typically have a polymodal grain size distribution because the constituent grains are derived from the skeletons of many different animals and plants.

Lithoclasts, derived from lithified or semi-lithified substrates, are important components of some carbonates. Tabular carbonate lithoclasts are commonly produced by storms and hurricanes that generate currents with power sufficient to rip up and transport lithified sediments from the seafloor.

The matrix in many carbonate rocks is microcrystalline calcite, or micrite, which commonly refers to any calcareous sediment that is formed of grains that are <4 μm long. Although originally defined as mechanically deposited sediment, micrite can also be formed as cement. Accordingly, the term 'micrite' is commonly used without any genetic connotation. Many carbonates, irrespective of their place of formation, are formed of allochems that are held in calcite and/or aragonite cements that were

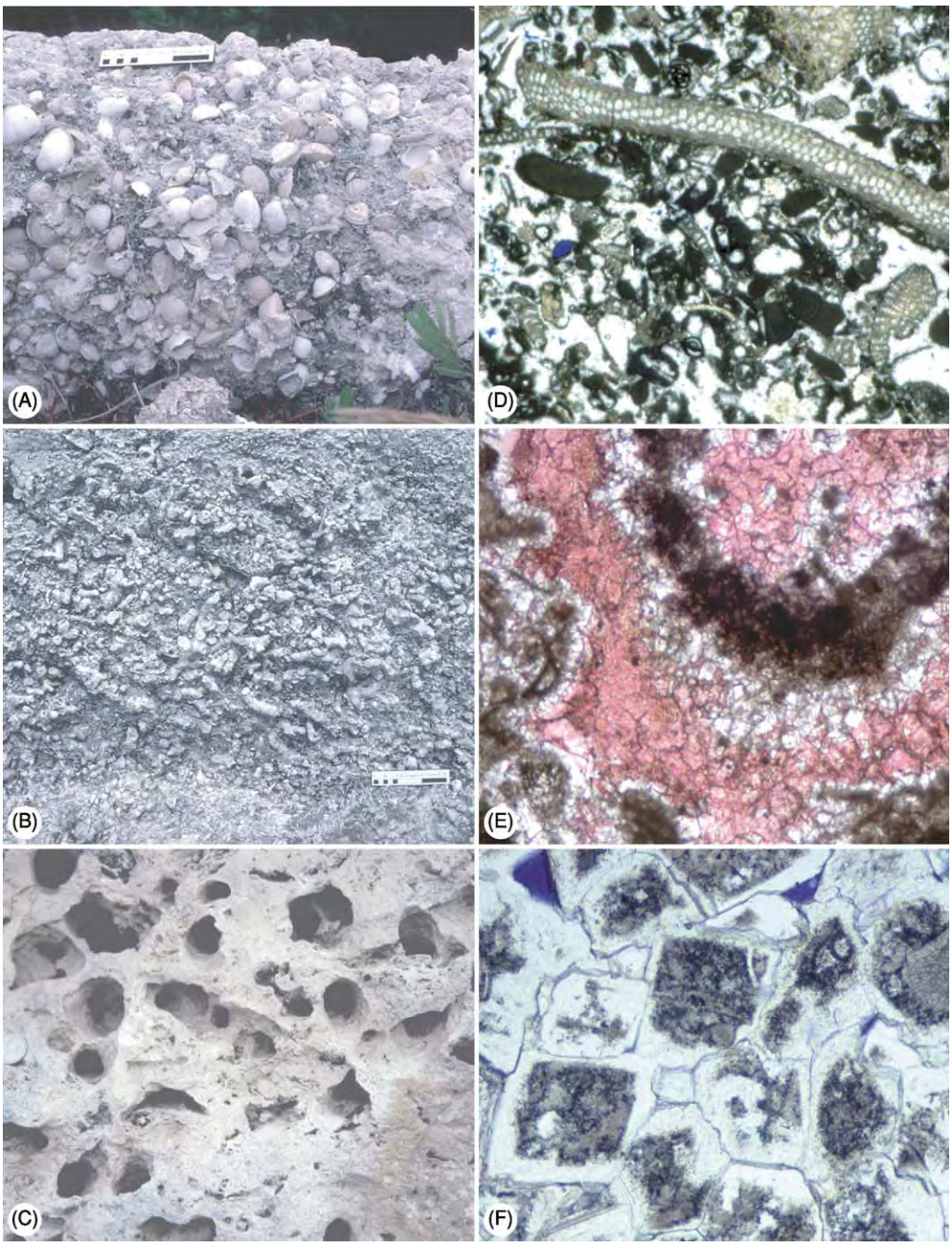


Figure 3 Lithified carbonates. (A) Articulated bivalves embedded in a packstone matrix; Ironshore Formation, Pleistocene, Grand Cayman. (B) Well preserved patch reef formed of *Porites*; Ironshore Formation, Pleistocene, Grand Cayman. (C) Limestone with fossil mouldic porosity (after corals); Pleistocene, Barbados. (D) Thin section photomicrograph, showing grainstone formed of bioclasts embedded in spar calcite cement; Brac Formation, Oligocene, Cayman Brac. Field of view covers ~6 × 6 mm. (E) Finely crystalline dolostone with cavities lined with dolomite cement (white) and filled with calcite cement (stained red by Alizarin Red S solution); Cayman Formation, Oligocene, Cayman Brac. Field of view covers ~1 × 1 mm. (F) Sucrosic dolomite formed of euhedral dolomite crystals that have a dirty core surrounded by a clear rim; note ghost structures of skeletal fragments in cores, inherited from precursor limestone; Brac Formation, Oligocene, Cayman Brac. Field of view ~1.5 × 1 mm.

precipitated following their deposition (Figure 3D). Cementation may take place while the sediments are still in their original depositional setting (e.g., hardgrounds, reefs) or millions of years later, after the sediments have been buried.

Classification

Numerous classification schemes have been proposed for carbonate sediments and sedimentary rocks. The Folk and Dunham classification schemes, which

reflect different philosophical approaches to the naming and genesis of carbonate rocks, are the most commonly used schemes. Usage of one or the other of these schemes is a matter of personal preference and the type of study that is being undertaken. Fundamentally, names derived from Folk's scheme reflect the dominant allochem (e.g., ooid, pellet, bioclast, or lithoclast) and the micrite matrix or spar calcite cement found between the allochems (Figure 4). For example, an oomicrite refers to a rock formed of ooids embedded in micrite matrix, whereas an oosparite refers to a rock formed of ooids embedded in a spar calcite cement.

Dunham's scheme is based on the depositional texture of the rock and makes no specific reference to the allochems, the matrix, or the cement (Figure 5). The progression from mudstone to wackestones to packstone to grainstone reflects, for example, a progressive decrease in the amount of mud (i.e., micrite) and hence ever-increasing energy levels. This scheme was

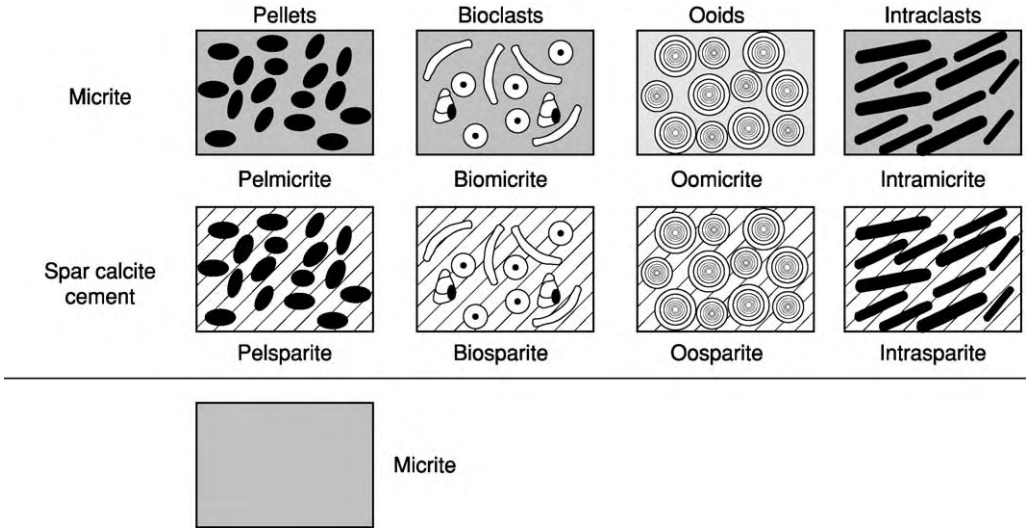


Figure 4 Folk's classification system for carbonate rocks.

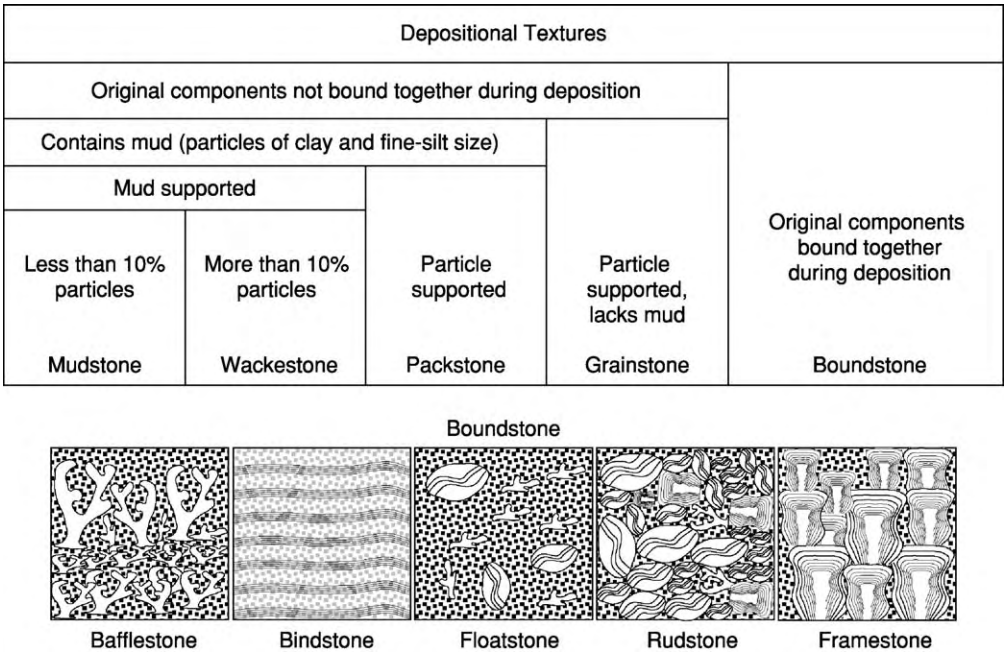


Figure 5 Dunham's classification system for carbonate rocks. The diagrams illustrating the various structures are labelled using the terminology subsequently introduced for different types of boundstones.

subsequently modified with the introduction of terms such as boundstone, bafflestone, and bindstone, which are typically applied to reefal limestones. In some cases, prefixes (e.g., skeletal grainstone) are used to signify the dominant allochem.

Marine Carbonate Deposits

Distribution and Composition

Depending on local conditions, many different minerals can be precipitated from marine waters. Seawater contains many salts, but under conditions of increasing temperature, $CaCO_3$ is one of the first compounds to be precipitated, either as calcite or aragonite. In modern oceans, carbonate sediments have the potential of forming anywhere and accumulating wherever the water depth is shallower than the carbonate compensation depth (CCD) and where there is little or no influx of siliciclastic sediments. In areas where large volumes of siliciclastic sediments are being introduced (e.g., where the Mississippi River empties into the Gulf of Mexico), the carbonate sediments, which form at a much slower rate, will be masked to the extent that they will appear to be non-existent. Extensive areas of carbonate sediment accumulation are found on shallow-water shelves (typically <30 m deep) that receive little or no influx of siliciclastic sediment, or on isolated banks that are surrounded by deep oceanic waters. In modern oceans, the mineralogy of the carbonate sediment is largely a function of the water temperature and pressure. Thus, aragonite dominates sediments forming in shallow tropical seas whereas calcite dominates temperate waters and the deeper, cooler waters in the tropics.

Depositional Systems

Marine carbonate sediments typically accumulate on a platform that develops as sediments accumulate in an area of subsidence. These depositional systems, which commonly cover hundreds of square kilometres, are referred to as a carbonate 'shelf' or a 'bank'. A carbonate shelf has a continental landmass on one side and deep oceanic waters on the other side, whereas a bank is completely surrounded by deep oceanic water (Figure 6). Carbonate shelves are classified as a ramp, an unrimmed shelf, or a rimmed shelf according to the slope of their depositional surface and the nature of their margins (Figure 6). A ramp slopes gradually (typically $<1°$) from the shoreline to the deep ocean with no appreciable change in slope. In contrast, rimmed and unrimmed shelves have a distinct break in slope on their oceanic margin (Figure 6). A rimmed shelf has islands, sand shoals, or actively growing reefs along their oceanic margins whereas an unrimmed shelf is devoid of such structures (Figure 6).

The margins of many modern tropical platforms are commonly characterized by large, robust coral reefs, especially on the windward coasts (Figure 1A). Spur and groove structures develop on the oceanward side of the reef crest. Unrimmed shelves or ramps are usually found on the leeward side of tropical islands and in temperate zones. The presence of islands, sand shoals, and/or reefs along the margin of a rimmed shelf significantly dampens the strength of any waves that encroach from the open ocean. As a result, low-energy environments and little current movement generally characterize the shelf. Only major storms and hurricanes will interfere with these tranquil settings, by generating high waves and strong currents for short periods of time. In stark contrast, ramps and unrimmed shelves have open communication with the ocean and therefore receive the full impact of onshore waves and currents. The fact that ocean waves move freely across the shelves with little or no impediment means that ramps and open shelves are characterized by high-energy conditions, complex arrays of nearshore facies, and physical transportation of sediment across the shelf.

The 'carbonate factory', where most carbonate sediment is generated, is typically located on the shallow, illuminated parts of the shelves. Conditions are ideal for the precipitation of carbonate sediment directly from seawater and for the growth of animals and plants that are the sources of aragonite and/or calcite skeletons (Figure 2F, G, and H). Most of the sediment stays in or close to the place where it formed. Nevertheless, storm-generated waves and currents can move sediment shoreward onto the

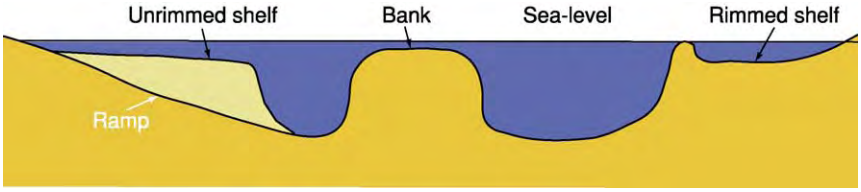


Figure 6 Schematic diagram showing differences between rimmed and unrimmed shelves, a ramp, and a bank.

intertidal and supratidal flats or seaward into the open ocean. Viewed in this context, it is apparent that the sediment production capacity of the carbonate factory carries important ramifications for sediment deposition throughout the carbonate depositional realm.

Factors critical to the health of the carbonate factory include water depth, water temperature, water circulation, oxygenation, salinity, and light penetration. In particular, these parameters control and strongly influence the growth and development of the plants and animals that contain the calcareous skeletons that will eventually become part of the sediment. Shallow, well-oxygenated, warm waters of normal salinity are, for example, ideal for the growth of many plants and animals, including calcareous algae, corals, bivalves, and gastropods. Conversely, many plants and animals cannot survive low or high salinities, water that is too cool or too hot, or poor circulation. Adverse environmental conditions commonly result in a biota formed of numerous individuals of a few taxa. Similarly, light penetration plays a major role in controlling the distribution of plants (e.g., calcareous algae and *Thalassia*) or animals (e.g., corals) that have a symbiotic relationship with plants (e.g., algae).

In many parts of the world, extensive carbonate deposits are accumulating in the intertidal (between low- and high-tide levels) and supratidal (above high-tide level) environments (**Figure 2A** and **B**). Conditions in these areas are highly variable because they are inundated by marine waters during high tide and storms, but are exposed to the atmosphere and rainwater during low tide. Such exposure means that climatic conditions play a critical role in the formation of sediments in these areas, especially in the supratidal area that is commonly exposed for extended periods of time. In hot, dry climates, evaporation dominates to the extent that evaporitic minerals, such as gypsum and halite, are produced. In humid climates, evaporates are rare and any that do form will dissolve during subsequent rainstorms. Under these conditions, the tidal flat commonly becomes covered with thick, leathery microbial mats and stromatolites that are formed of various consortiums of bacteria, fungi, and cyanobacteria (**Figure 2B**). The intertidal flats are complex depositional systems because meandering tidal creeks commonly cut them and isolated ponds are common (**Figure 2A**). Small-scale arrays of subtidal, intertidal, and supratidal microenvironments are associated with each creek and pond despite the fact that they are located on the intertidal flats. Such complex environmental diversity means that these depositional systems are usually characterized by complex and rapid vertical and lateral facies changes and a biota that is usually characterized by large numbers of only a few species. Little sediment (apart from evaporites) actually forms on the intertidal and supratidal flats. Instead, most of the sediment forms in the nearby lagoons and is carried onshore during severe storms and hurricanes.

Sediment that is transported in an oceanward direction will settle in the forereef area, where it may be transported downslope under the effects of gravity. As a result, carbonate sediments originally formed in shallow water may accumulate at considerable depth on the forereef slope or, in some cases, on the basin floor. Collapse of the reef, whether due to instability caused by overgrowth of the reef or to earthquake activity, will produce large blocks of reefal limestones (commonly the size of houses) that will eventually come to rest in deep water on the basin floor.

Sedimentation in marine systems is a dynamic process that is ultimately controlled by the balance between seafloor subsidence, sedimentation, and sea-level changes. The carbonate factory will continue to function while conditions conducive to plant and animal growth are maintained. A sudden increase in water depth, irrespective of its cause, will shut down the carbonate factory and in extreme cases will 'drown' the platform. Similarly, a decrease in water depth can have profound affects on the depositional system because it may lead to the exposure of some reefs and to the seaward extension of the tidal flats, making the subtidal conditions less conducive to the growth of plants and animals.

Reefs and Mounds

A reef, a positive seafloor structure constructed by animals and plants, is capable of withstanding strong currents and waves. In modern oceans, reefs are constructed mostly by coral-dominated communities and are typically found in shallow, well-illuminated tropical waters that are conducive to rapid coral growth. A mound, in contrast, is typically smaller and usually develops in more tranquil settings, with growth being mediated by fewer and more delicate creatures.

Modern reefs vary in size from small patch reefs (**Figure 2H** and **3B**) that may be only a few metres in diameter and include few coral heads, to immense structures, such as the Great Barrier Reef off the north-east coast of Australia, that are thousands of kilometres long and are formed of numerous corals along with various other animals and plants. Such large reefs are important because they separate the open ocean from the landward shelf. Atolls, which are circular to ovate structures that enclose a central lagoon, are formed where coral reefs have colonized and grown upwards from the rim of a subsiding

volcano. The construction and rigidity of a reef, irrespective of its size and morphology, are controlled by the growth of the animals/plants and the deposition of sediment and/or precipitation of cements in cavities between those formative organisms. Destructive processes, which offset the constructive processes, include bioerosion and physical damage caused by high-energy events such as storms and hurricanes. Not all reefs have been formed by corals. Archaeocyathids (Cambrian), sponges (Ordovician), stromatoporoids (Silurian, Devonian), tabulate corals (Ordovician, Silurian), phylloid algae (Pennsylvanian), and rudist bivalves (Cretaceous), for example, have all been responsible for the construction of large reefal structures at different times throughout the Phanerozoic.

Mounds, which vary from low-relief lenses to mounds with slopes up to 40°, are generally categorized as microbial mounds, skeletal mounds, or mud mounds. Their genesis is poorly understood because the origin of the constituent mud is poorly constrained, the organisms that mediated their formation are typically poorly preserved or absent, and there are no modern examples of mud mounds. Growth and development of microbial mounds were typically mediated by cyanobacteria, algae, diatoms, and other micro-organisms that can become calcified, trap sediment, and/or induce $CaCO_3$ precipitation. Skeletal mounds formed where organisms such as bryozoa, corals, stromatoporoids, sponges, and rudist bivalves induced local sedimentation through current baffling, sediment trapping and binding, and sediment stabilization.

Marine Sedimentary Processes

The formation of carbonate deposits through deposition and/or precipitation depends on the depositional environment and, in particular, whether animals and/or plants are present. Plants play many different roles in the formation of marine carbonate sediment. Calcareous algae, seagrasses, and mangroves, for example, play important roles in carbonate sedimentation. Carbonate sedimentation is influenced by the binding action of their roots, the baffling action of their leaves, and the sediment production by the epibionts that live on their leaves and/or roots. The roles played by each plant vary. Calcareous algae, with their calcareous skeletons, are major sediment producers but play a minimal role in terms of current baffling or sediment binding. Conversely, *Thalassia* ('turtle grass') has long-bladed leaves that reduce current strengths and thereby cause deposition of any sediment suspended in that current, as well as long complex rhizomes that bind the sediment in place and prevent erosion except by the strongest of storm-driven currents. *Thalassia* is also a major sediment source because the calcareous skeletons of the small animals and plants that live on their leaves become part of the sediment load following the death and decay of the leaves (**Figure 2F**). Mangroves (**Figure 2B**), found in coastal regions, are characterized by extensive prop-root systems that bind sediment in place and are extremely effective at reducing currents, thereby causing sediment deposition. Sediment production, however, is generally limited to those plants and animals that attach themselves to the roots of trees.

Micro-organisms commonly play a critical role in carbonate sedimentation. Various consortia of microbes can form microbial mats that cover vast areas of the seafloor and the intertidal flats. In many areas, these microbial communities mediate the growth and construction of stromatolites (**Figure 2G**) by trapping and binding sediment to the seafloor or by acting as nucleation sites for aragonite and/or calcite precipitation. Removal of the mats commonly leads to erosion and transportation of the underlying sediment. In many areas, the microbial communities mediate the growth and construction of stromatolites, which are highly variable in terms of their size and morphology.

Non-Marine Carbonates

In non-marine settings, carbonate deposits will form from any water that is supersaturated with respect to $CaCO_3$. Rapid degassing of CO_2 and/or evaporation commonly trigger such precipitation (**Figure 1B** and **E**). In most situations, abiotic precipitation, in contrast to biologically influenced precipitation, is more important. Nevertheless, coated grains (**Figure 1C**), bioclasts, peloids, and lithoclasts can form in many of these settings. Indeed, it is commonly difficult to distinguish between allochems that form in marine and non-marine settings if their depositional context is unknown.

The precipitation of aragonite, as opposed to calcite, in non-marine settings has commonly been attributed to water temperature, with calcite being precipitated from water that is cooler than 40°C and aragonite being precipitated from water that is warmer than 40°C. This relationship, however, is not universally true. Exceptions are found when extreme CO_2 degassing leads to aragonite precipitation or when slow ion delivery, caused by microbial biofilms or high viscosity, leads to calcite precipitation. In the vents of some hot springs ($T > 90$°C), alternating aragonite and calcite precipitation can take place in response to variations in the rate of CO_2 degassing, even though the water chemistry remains essentially

constant. Precipitation under conditions of rapid CO_2 degassing and/or evaporation commonly produces bizarre and unusual crystals. Trigonal, dendritic, skeletal (i.e., hollow), or platy calcite crystals, for example, are found in many deposits that have formed around the vents of hot springs and geysers. Many of these crystals are composite crystals, with the constituent subcrystals being readily apparent in high-magnification, high-resolution scanning electron microscope images.

Compared to the role of plants, animals, and other eukaryotes in forming marine carbonates, their role in forming non-marine carbonates is relatively minor. The distribution of the biota in these settings is controlled primarily by water temperature, water acidity, and, in some spring systems, the presence of elements that may be toxic. One of the main controls over the distribution of the biota is the maximum water temperature that each taxon of can tolerate (e.g., 45–50°C for fungi). In the simplest sense, however, the diversity of plants and other organisms tends to decrease with increasing temperature. In some cold-water streams, lakes, and spring systems, however, calcite is commonly precipitated around charophytes (freshwater green algae) and bryophytes (mosses, liverworts, and hornworts). On the sides of valleys or tufa dams in some cold-water spring systems, calcite precipitation around these species can lead to the construction of large fan-shaped deposits that stretch from one side of a valley to another and cause ponding on the upstream side. Microbial communities formed of cyanobacteria, bacteria, and/or fungi are common in many freshwater lakes, streams, and spring systems. As in the marine environments, the microbes can play a major role in $CaCO_3$ accumulation by providing nucleation sites for crystal growth, by inducing precipitation through modification of the physiochemical conditions in their surrounding microenvironment, and/or by trapping and binding sediment to the substrate. Such activity commonly leads to the formation of microbial mats and/or stromatolites that are akin to those found in marine environments.

Diagenesis

Carbonate diagenesis is primarily driven by the fact that carbonate minerals are highly reactive to changes in temperature and pressure conditions, especially in the presence of vast quantities of water. Dissolution, cementation, inversion, recrystallization, and replacement, which are commonly triggered by pressure and temperature changes, may lead to the complete transformation of carbonate sediments and limestones. In many cases, the original components of the limestones will be obliterated, and porosity and permeability are either created or destroyed. Dolomitization, which is still poorly understood, is commonly pervasive, with thick successions of limestone being replaced by dolomite. Carbonate diagenesis will start on the seafloor and continue until the rocks have been buried to depths at which metamorphism takes place. In general, the factors that control diagenesis are poorly understood because they operate on a microscale in settings that cannot be directly observed or monitored. Much more is known about diagenesis on the seafloor and in near-surface settings than is known about diagenesis that takes place at depth, simply because the former settings are much more amenable to observation and monitoring. Diagenesis on the seafloor (e.g., hardground formation, reef lithification) and along the shoreline (e.g., beachrock) typically involves the precipitation of aragonite and/or calcite cements directly from seawater. Once exposed in the vadose zone (i.e., above the water table), carbonate sequences are prone to significant changes. Climate plays a critical role because maximum carbonate diagenesis takes place where vast quantities of water flow through the rocks under high-temperature conditions. Thus, carbonate rocks located in hot, humid climates tend to undergo more rapid diagenesis than do those in cool, dry climates. The surface and subsurface landforms associated with karst terrains provide clear evidence of the effects that surface and near-surface diagenesis has on carbonates. Dissolution is mediated by the weak carbonic acid that is formed as rainwater absorbs CO_2 from the atmosphere and from decaying vegetation on the ground. Aragonitic components are either dissolved, producing fossil-mouldic porosity (Figure 3C), or are transformed to calcite. Dissolution of aragonite and calcite means that the groundwaters commonly become supersaturated with respect to $CaCO_3$, allowing precipitation of aragonite and/or calcite cements to take place in cavities in which suitable physiochemical conditions exist. Diagenesis in the freshwater phreatic zone, which depends on water chemistry, may involve dissolution or cementation. Under conditions of deep burial, the rocks are subjected to considerable overburden pressures that may lead to dissolution and the formation of stylolites.

Many ancient carbonate successions have been pervasively dolomitized (Figure 3F) by processes that are still not well understood. All dolomite models, irrespective of their mechanics, require a source of Mg, a mechanism for transporting the Mg to the dolomitization site, and a dolomitization site that is physiochemically conducive to dolomite formation. Some models, such as the reflux model, involve near-surface

dolomitization that is mediated by brines with a high Mg:Ca ratio. Other models operate on the premise that dolomitization took place in the mixing zone, wherein mixtures of freshwater and seawater were responsible for dolomite formation. At the other extreme are models that have dolomitization taking place after deep burial. Tertiary 'island dolomites' are commonly attributed to dolomitization by slightly modified seawater in near-surface settings. It seems probable, however, that pervasive dolomitization can be achieved under many different conditions and that it is impossible to explain all dolomitization by means of a single model.

See Also

Biosediments and Biofilms. **Carbon Cycle**. **Diagenesis, Overview**. **Fossil Invertebrates:** Bryozoans; Corals and Other Cnidaria. **Fossil Plants:** Calcareous Algae. **Microfossils:** Foraminifera. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Rocks:** Mineralogy and Classification; Dolomites; Limestones.

Further Reading

Adams AE, MacKenzie WS, and Guilford C (1984) *Atlas of Sedimentary Rocks under the Microscope*. England: Longman.

Bathurst RGC (1975) *Carbonate Sediments and Their Diagenesis*. Amsterdam: Elsevier.

Budd DA, Saller AH, and Harris PM (1995) *Unconformities and Porosity in Carbonate Strata. American Association of Petroleum Geologists, Memoir 63*. Tulsa, OK: American Association of Petroleum Geologists.

Loucks G and Sarg JF (1993) *Carbonate Sequence Stratigraphy. American Association of Petroleum Geologists, Memoir 57*. Tulsa, OK: American Association of Petroleum Geologists.

Purser B, Tucker M, and Zenger D (1994) *Dolomites. A Volume in Honour of Dolomieu. International Association of Sedimentologists, Special Publication Number 21*. Oxford: Blackwell.

Scholle PA, Bebout DG, and Moore CH (1983) *Carbonate Depositional Environments. American Association of Petroleum Geologists, Memoir 33*. Tulsa, OK: American Association of Petroleum Geologists.

Tucker ME and Bathurst RGC (1990) *Carbonate Diagenesis. International Association of Sedimentologists Reprint Series*, vol. 1. Oxford: Blackwell.

Tucker ME and Wright VP (1990) *Carbonate Sedimentology*. Oxford: Blackwell.

Vacher HL and Quinn TM (eds.) (1997) *Geology and Hydrogeology of Carbonate Islands. Developments in Sedimentology 54*. Amsterdam: Elsevier.

Walker RG and James NP (1992) *Facies Models. Response to Sea Level Change*. Ottawa: Geological Association of Canada.

Wilson JE (1975) *Carbonate Facies in Geologic History*. New York: Springer Verlag.

Chromates

P A Williams, University of Western Sydney, Penrith South DC, NSW, Australia

Introduction

Naturally occurring chromate minerals are confined to the oxidized zones of base metal orebodies and to extremely oxidizing desert environments. Under such conditions, chromium reaches its highest oxidation state (+6). The only commercially significant ore mineral of chromium is the primary spinel chromite, $FeCr_2O_4$, which contains Cr(III). Chromium (VI) can form under very oxidizing conditions, and nearly all of its minerals contain the simple tetraoxochromate(VI) or chromate ion, CrO_4^{2-} (see Table 1). Two exceptions are chrombismite, $Bi_{16}CrO_{27}$, a complex lattice oxide, and lopezite, naturally occurring potassium dichromate. The anion in the latter is generated reversibly by the acid-assisted dimerization of chromate, as shown in eqn [1]:

$$2CrO_4^{2-}(aq) + 2H^+(aq) \leftrightarrow Cr_2O_7^{2-}(aq) + 2H_2O \quad [1]$$

Geographical Distribution of Chromates

The chromates listed in Table 1 occur in two quite disparate settings: deserts and oxidized zones of a few base metals. Desert environments include those in Jordan (the only known locality for hashemite) and, more spectacularly, the Atacama Desert of Chile. In the latter case, chromates are associated with the caliche nitrate deposits, which besides nitrate contain other highly oxidized species, such as bromate, iodate, periodate, perchlorate, and native iodine. The persistence of highly oxidizing species in this environment is due to the extreme paucity of organic material with

Table 1 Chromate(VI) minerals

Type	*Name*	*Composition*
Simple chromates and dichromates		
	Chromatite	$CaCrO_4$
	Crocoite	$PbCrO_4$
	Hashemite	$BaCrO_4$
	Lopezite	$K_2Cr_2O_7$
	Tarapacaite	K_2CrO_4
Complex chromates		
	Cassedanneite	$Pb_5(CrO_4)(VO_4)_2\ H_2O$
	Deanesmithite	$Hg_5S_2O_2(CrO_4)$
	Edoylerite	$Hg_3S_2(CrO_4)$
	Embreyite	$Pb_5(CrO_4)_2(PO_4)_2\ H_2O$
	Dietzeite	$Ca_2(CrO_4)(IO_4)_2\ H_2O$
	Fornacite	$CuPb_2[(Cr,As,P)O_4]_2OH$
	Hemihedrite	$Pb_{10}Zn(CrO_4)_6(SiO_4)_2F_2$
	Iranite	$Pb_{10}Cu(CrO_4)_6(SiO_4)_2(F,OH)_2$
	Macquartite	$Pb_7Cu_2(CrO_4)_4(SiO_4)_2(OH)_2$
	Phoenicochroite	Pb_2OCrO_4
	Santanaite[a]	$Pb_{11}O_{12}CrO_4$
	Vauquelinite	$Pb_2Cu(CrO_4)(PO_4)(OH)$
	Wattersite	$Hg_5O_2CrO_4$
	Yedlinite	$Pb_6Cl_6O_2CrO_4\ 2H_2O$

[a]Contains both Pb(II) and Pb(IV).

which they would ordinarily react. Dietzeite, chromatite, lopezite, and tarapacaite have all been found in certain of the caliche deposits. These minerals are either freely soluble or decompose in water and they owe their preservation as well to the extreme aridity of the Atacama region.

Chromate minerals that form sparingly in the oxidized zones of base metal deposits form the bulk of the minerals listed in Table 1, and many of these chromate species are involved in complex solid solutions with both constituent cations and anions. Molybdate substitutes for chromate in the fornacite lattice and the alternate end-member molybdofornacite is recognized as a separate species. Nevertheless, all of the chromates are rare and none is a used as a source of chromium. One exception to this is the mineral crocoite, normal lead(II) chromate, which has been used as a paint pigment. The element chromium was first extracted and identified in 1796 by the French chemist Louis Vauquelin in samples of crocoite from Beresov, Russia. Fine specimens of crocoite together with the rarer phase, phoenicochroite, were recovered from Beresov, and the mineral has been found at a number of other localities worldwide. By far the most spectacular specimens have come from the Magnet mine and other deposits in the Dundas area of north-western Tasmania, Australia. Single crystals weighing several hundred grams each were recovered from the Red Lead mine, Dundas, and the neighbouring Adelaide mine has been a prolific producer of fine specimens for several decades. Crocoite in these deposits is formed by the interaction of chromium-bearing solutions derived from weathering serpentinites with oxidized, fault-hosted argentiferous galena ores. Associated with the crocoite is yellow cerussite, $PbCO_3$; the yellow colour is thought to arise from the substitution of minute amounts of chromate in the lattice.

See Also

Minerals: Definition and Classification; Molybdates; Nitrates; Tungstates.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2003) *Handbook of Mineralogy. Volume 5. Borates, Carbonates, Sulfates, Chromates, Germanates, Iodates, Molybdates, Tungstates, etc., and Organic Materials*. Tucson, AZ: Mineral Data Publishing.

Baes CF, Jr and Mesmer RE (1986) *The Hydrolysis of Cations*. Malabar, FL: Krieger Publishing Company.

Bard AJ, Parsons R, and Jordan J (1985) *Standard Potentials in Aqueous Solution*. New York: Marcel Dekker.

Ericksen GE (1981) *Origin of the Chilean Nitrate Deposits. United States Geological Survey Professional Paper, 1118*. Washington, DC: US Geological Survey.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, 8th edn. London: Wiley Europe.

Haupt J (1988) Minerals of western Tasmania. In: *The Mineralogical Record*, vol. 19, pp. 381 388. Tucson, AZ: Mineralogical Record Inc.

Mandarino JA (1999) *Fleischer's Glossary of Mineral Species 1999*, 8th edn. Tucson, AZ: Mineralogical Record Inc.

Roberts WL, Campbell TJ, and Rapp GR, Jr (1990) *Encyclopedia of Minerals*, 2nd edn. New York: Van Nostrand Reinhold.

Williams PA (1990) *Oxide Zone Geochemistry*. Chichester: Ellis Horwood.

Williams SA (1974) The naturally occurring chromates of lead. *Bulletin of the British Museum (Natural History)* 8: 379 418.

Feldspars

R A Howie, Royal Holloway, University of London, London, UK

The minerals of the feldspar group are the most abundant minerals in the Earth's crust and are also found on the Moon and in meteorites. Their ubiquity and their wide range of chemical composition has led to their use as a primary tool in the classification of the igneous rocks (*see* **Rocks and Their Classification**). In the great majority of these rocks, whether acid, alkaline, intermediate, or basic, the feldspars are the major constituents; they are absent only from some ultrabasic and rare alkaline rocks. They are important constituents of simple pegmatites and are common in mineral veins. Feldspars are major constituents of most gneisses and schists, and occur also in any thermally as well as regionally metamorphosed rocks. Although susceptible to alteration and weathering, the feldspars are second in abundance to quartz in the arenaceous sediments (*see* **Sedimentary Rocks:** Sandstones, Diagenesis and Porosity Evolution), in which they occur as detrital grains. It is only in the pelitic, and to a greater degree in the carbonate rocks that they are of relatively minor importance.

Nomenclature

Most feldspars may be classified chemically as members of the ternary system $NaAlSi_3O_8$ (albite, Ab)–$KAlSiO_8$ (orthoclase, Or)–$CaAl_2Si_2O_8$ (anorthite, An) (*see* **Sedimentary Rocks:** Mineralogy and Classification). Compositions between $NaAlSi_3O_8$ and $KAlSi_3O_8$ are referred to as alkali feldspars and those between $NaAlSi_3O_8$ and $CaAl_2Si_2O_8$ as plagioclase feldspars. These two series each generally contain less than 5–10% of the other. The distinction between alkali and plagioclase feldspars near the $NaAlSi_3O_8$ composition is an arbitrary one.

To achieve a proper understanding of feldspar relationships it is necessary to characterise them not only by chemical composition, but also according to their structural state, the latter depending on the temperature of crystallisation and on subsequent thermal history (*see* **Igneous Processes**). Feldspars that retain a crystal structure appropriate to their high-temperature formation are called high-temperature feldspars; most feldspars in volcanic rocks are of this type. Low-temperature feldspars are those with structures appropriate either to crystallization at lower temperatures, or to slow cooling from elevated temperatures as, for example, in plutonic igneous rocks.

The difference between high- and low-structural state in a feldspar can involve the degree of ordering of Al and Si atoms between distinct tetrahedral sites. Feldspars can thus be classified as:

i. homogeneous high-temperature disordered feldspars; or
ii. ordered feldspars, in most of which phase separation (i.e., phases with appreciably different Na, K, or Ca content) occurs on a macro-, micro- or submicroscopic scale.

The nomenclature of the high-temperature feldspars is based on crystal symmetry. The sanidines (Figure 1A), mostly Or_{50}–Or_{80}, have monoclinic symmetry; the anorthoclase field embraces those feldspars that are triclinic at room temperature but

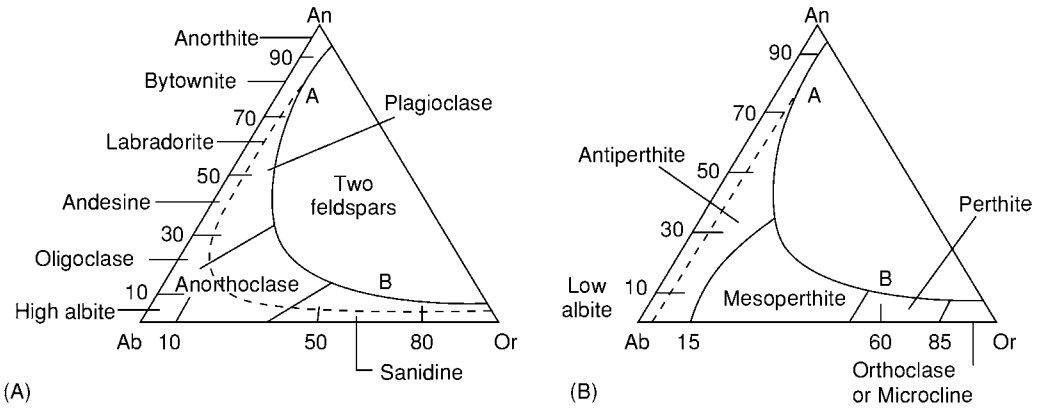


Figure 1 Feldspar nomenclature for (A) disordered ternary feldspars; (B) ordered ternary feldspars in which phase separation is at or below the resolution of the optical microscope. Composition in mol.%. Curve A B limit of ternary solid solution.

which may have possessed monoclinic symmetry at crystallization temperatures. For pure $NaAlSi_3O_8$, both low- and high-temperature forms are triclinic, but the remainder of the plagioclases have triclinic symmetry both at room temperature and at their temperature of crystallization.

In the alkali feldspars Al/Si and Na/K ordering lead to marked changes in symmetry, optics, morphology, and exsolution textures. In the compositional range Or_{15}–Or_{85}, exsolution gives rise to cryptoperthite (in which exsolution is on a submicroscopic scale), or to perthite or microperthite (Figure 1B). Microcline is triclinic, and may show varying degrees of order. The name orthoclase is used to describe intermediately ordered K-rich feldspars. Adularia is a low-temperature ordered potassium feldspar with a characteristic habit (large prism faces) and paragenesis ('alpine' veins).

A purely chemical definition of a plagioclase can be given in terms of Ab–An 'molecular' percentages, but specific names are used to denote the six compositional ranges into which the series has been divided. Thus, albite, oligoclase, andesine, labradorite, bytownite, and anorthite refer to the An percentages 0–10, 10–30, 30–50, 50–70, 70–90 and 90–100, respectively. Position within a chemical range may be further indicated by such terms as sodium-rich (or sodic) oligioclase or calcic labradorite, etc. Many plagioclases consist of a fine lamellar intergrowth of two phases, one more Ca-rich than the other; those within the An_5–An_{20} range are called peristerites and their surfaces may exhibit iridescence. Other ranges for such intergrowths may be found at An_{45}–An_{70} (Bøggild) which shows labradorescence and at An_{75}–An_{85} (Huttenlocher), also with associated iridescence effects.

Most perthites (named from Perth, Quebec, an early locality) are exsolution intergrowths of sodium-rich feldspar in a potassium-rich feldspar host; an intergrowth of K-rich feldspar in a plagioclase host is called antiperthite. The name feldspar as originally given was feldtspat and is believed to be in reference to the presence of this spar (spath) in tilled fields (Swedish: feldt or fält) overlying granite (*see* **Igneous Rocks:** Granite), rather than to the German Fels, meaning rock.

Structure

The feldspars are framework silicates with tetrahedra of $(Si,Al)O_4$ linked to one another (by shared oxygens) in all directions rather than in chains or sheets. Although discrete chains of tetrahedra do not occur in the structure, its nature may be more readily understood by considering the atomic arrangement as the linking of chains in two directions perpendicular to their length. The chains themselves are formed by the linking of horizontal rings of four tetrahedra, as shown in Figure 2. When viewed in the direction of the chain axis, a horizontal ring appears approximately, as shown in Figure 2B, and this can be further simplified by its representation, as in Figure 2C. The configuration shown in Figure 2 is often described as a 'double crankshaft'; the upwards- and downwards-pointing tetrahedra in a horizontal ring are labelled U and D, respectively. In the actual structure, the rings are considerably distorted; successive horizontal rings of a chain are related by vertical glide planes passing through their centres; the view down the chain axis may be idealized, as in Figure 2D, the first, third, fifth, etc. rings being represented by thick and the even number rings by thin lines. At the level of the first ring, a network of oxygen linkages is formed, producing a plane of four-membered and eight-membered rings of tetrahedra. The eight-membered rings are of two kinds, UUUUDDDD and DUUDUDDU, characteristic of feldspars, as distinct from other framework silicates, (*see* **Minerals:** Other Silicates) such as zeolites.

Twinning in feldspars is common and can occur in three different ways: (i) as a primary phenomena during crystal growth; (ii) as glide twinning induced by deformation; or (iii) on thermal transformation to lower symmetry. There are several twin laws shown

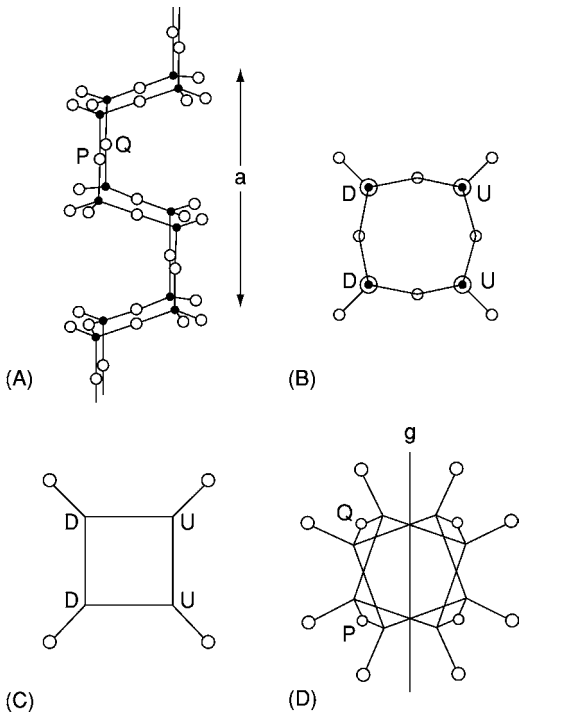


Figure 2 Idealized illustrations of the feldspar 'chain'. In (B) and (C) U and D indicate tetrahedra pointing up and down, respectively.

by feldspars; basically these may appear as single twins or as repeated or polysynthetic twins, an example of the latter being the very frequent twinning on the albite law in triclinic feldspars, giving the characteristic bands of alternating birefringence with an (010) composition plane. Measurement of the positions of optical extinction of these twins in plagioclase can be used to give an indication of the composition of the crystals.

Optical and Physical Properties

All the common members of the feldspars have D 2.55–2.5, with H 6–61/2. They are generally colourless, white or grey, but may be yellow, red, or green; colourless in thin section. Refractive indices are low, alkali feldspars having values <1.54 and plagioclase in the range 1.53–1.59 (the values increasing with increase in An content); the birefringence ranges 0.006–0.013. Single or multiple twins are common.

Alkali Feldspars

(K,Na)[($AlSi_3O_8$]. Monoclinic or triclinic. K-feldspar melts incongruently at ~1150° to leucite ($KAlSi_2O_6$) under anhydrous condition at 1 atm (Figure 3). The eutectic temperature (E) between K-feldspar and tridymite is ~990°C. From a melt of composition A, leucite (B) is the first phase to crystallize, and continues to do so as the liquid composition moves along the liquidus curve to the reaction point (R). At this temperature, the liquid reacts with the leucite transforming it to K-feldspar. As the temperature continues to fall, the composition of the melt moves along the curve RE with K-feldspar continuing to crystallize; at the eutectic (E) crystallization is complete and the whole mass solidifies as a mixture of K-feldspar and tridymite. Later experimental work has shown that at 4 kbar the eutectic point is lowered to ~735°C.

The refractive indices for the alkali feldspar series increase steadily with increasing albite content (Figure 4), values for α and γ in the high-temperature series sanidine–anothoclase–high-albite being marginally lower. Microcline is distinguishable from most other alkali feldspars by its 'tartan' twinning in which two sets of twin lamellae are approximately at right angles to each other.

The K–Na feldspars are essential constituents of alkali and acid igneous rocks and are particularly abundant in syenites, granites, granodiorites, and their volcanic equivalents. In plutonic rocks, the alkali feldspar is usually orthoclase, microcline, microcline microperthite, or microcline perthite, whereas in volcanic rocks it is sanidine, anorthoclase cryptoperthite, or anorthoclase. Granophyric intergrowths, the bulk composition of which is close to the melting minimum in the orthoclase–albite–quartz–H_2O system, consisting of roughly equal amounts of quartz, K-feldspar, and Na-feldspar components, occur in the mesostasis of acid rocks of high-level origin.

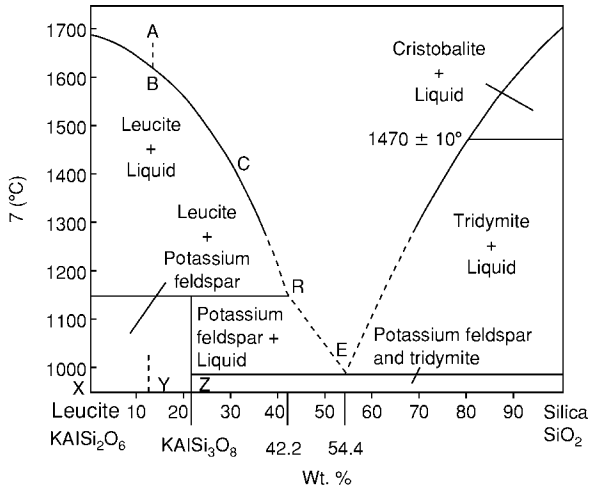


Figure 3 Equilibrium diagram of the binary system $KAlSi_2O_6$ (leucite) SiO_2. Reaction point R represented approximately by $5\,KAlSi_2O_6(\text{leucite}) + 3\,KAlSi_3O_8 + 5\,SiO_2(\text{melt}) \rightarrow 8\,KAlSi_3O_8$.

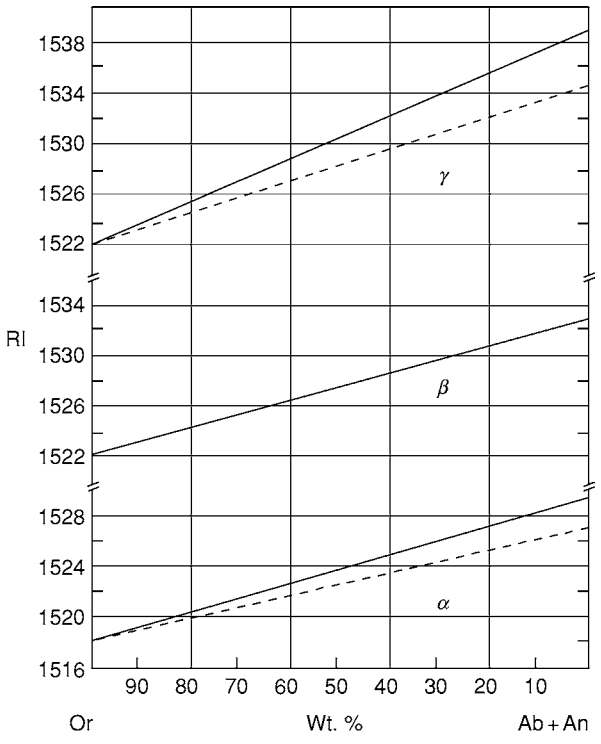


Figure 4 Variation of refractive indices with composition for the alkali feldspars; α and γ for the sanidine anorthoclase high albite series are shown as dashed lines, and those for the ortho clase (microcline) low albite series as full lines. The β index is essentially the same in both series.

K-feldspar is a stable product of both high-grade thermal and regional metamorphism. It is a typical mineral of the sillimanite zone of metamorphism and, in rocks of argillaceous composition, does not appear at lower grades. Where K-feldspars released by weathering of igneous or metamorphic rocks have undergone rapid transportation and burial, they occur as relatively abundant detrital constituents in arkosic sediments. They may also develop as authigenic minerals occurring as rims on detrital feldspar, often in optical continuity.

Plagioclase

$Na[AlSi_3O_8]$—$Ca[Al_2Si_2O_8]$. Triclinic. The phase diagram for the plagioclase series was one of the first to be determined experimentally (Figure 5). It shows a solid solution without maximum or minimum, with the melting point of anorthite at 1550°C and that of albite at ~1100°C at $P = 1$ bar. A liquid of composition $Ab_{50}An_{50}$ (A) begins to crystallize at ~1450°C, the first crystal having the composition of $\sim Ab_{82}An_{18}$(B). With further cooling under equilibrium conditions, both liquid and crystals change their composition along the liquidus and solidus, respectively, until at 1285°C the crystals reach a composition of $Ab_{50}An_{50}$(D) as the last of the liquid now of composition (C) is used up. This continuous change in composition of the plagioclase crystals with falling temperature occurs only if there is sufficient time for the earlier-formed crystals to react with the liquid: if there is insufficient time for this interchange of material, the crystals will be zoned. The resultant product will have an average composition of $Ab_{50}An_{50}$, but the inner core will be more calcic and the outer zones more sodic. With the addition of H_2O to the system, at $P_{H_2O} = 5$ kbar, the liquidus and solidus temperatures are over 300° lower than for the anhydrous system (Figure 5, curve 2). An even greater effect is shown on the addition of SiO_2 to the system at 2 kbar (curve 3).

The optical properties of the plagioclases are directly related to their anorthite content. Both the relief and birefringence are low and similar to quartz. A determinative chart is given in Figure 6; it is important to note that measurement of the refractive indices must be accurate to ±0.001 to obtain an accuracy of ± 2% An. The composition may also be determined by melting a small amount of material, quenching it to a glass and determining the refractive index of the glass, the refractive index for an isotropic glass being more easily determined than for a triclinic mineral and, for the plagioclases, having twice as great a rate of change with composition (Figure 6). This method can give the average composition for plagioclase which is strongly zoned or has exsolution intergrowths; also, the original structural state of the plagioclase is of no consequence.

In the majority of twinned plagioclase crystals, the composition plane is parallel to the crystal length.

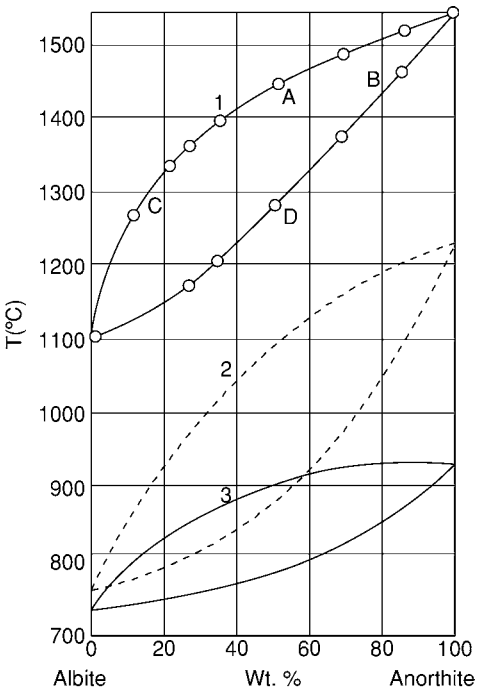


Figure 5 Equilibrium diagram of the plagioclase feldspars: (1) in anhydrous conditions, (2) at 5 kbar water pressure, (3) eutectic liquidus and solvus in the albite anorthite quartz H_2O system at 2 kbar.

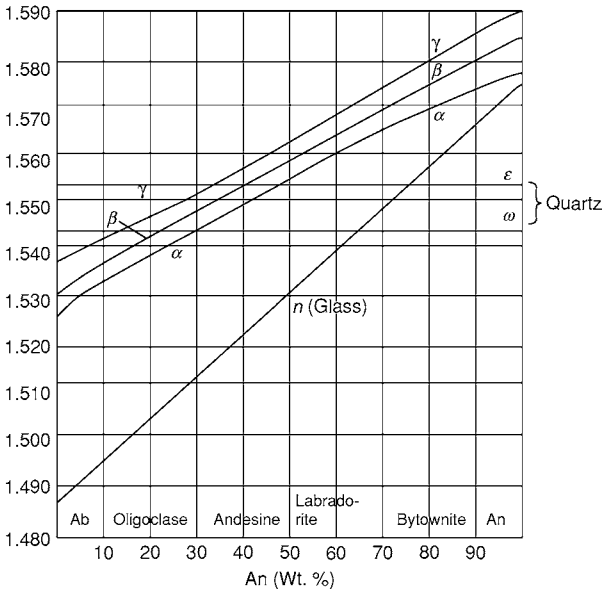


Figure 6 Refractive indices of the plagioclases. Values for glasses of plagioclase composition are also shown.

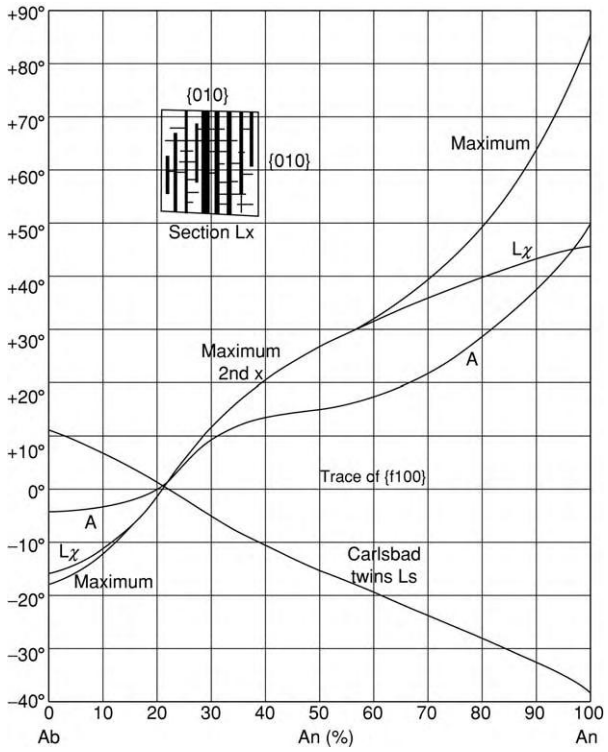


Figure 7 Extinction angles of plagioclase (with respect to α') in the 'symmetrical zone' and in sections normal to *x*.

The maximum extinction angles normal to {010} or in the so-called symmetrical extinction zone are diagnostic (Figure 7). In thin section, the alternate twin lamellae give symmetrical extinction angles on either side of the twin plane. Such sections may be recognised by the sharpness of the composition plane between albite twin lamellae, by the equal interference colours of the twin lamellae when the twin plane is parallel to the vibration direction of the polarizers, and by adjacent twin lamellae giving equal extinction angles on either side of the twin plane. Values which show more than a 5% divergence in the extinction angles for adjacent twin lamellae should be discarded, but lesser variations may be averaged; it is essential to take measurements from several (6–12) suitable grains, and the highest extinction angle must then be used.

Plagioclase is the most abundant mineral in the great majority of basic and intermediate lavas, in which it occurs both as phenocrysts and as a groundmass constituent. In basalts, the plagioclase phenocrysts usually have a wide homogeneous core of bytownite composition surrounded by a narrow zone of more sodic plagioclase. The broad cores of uniform composition indicate slow crystallization and it is clear that the growth of these crystals occurred before extrusion and final consolidation of the magma. Under plutonic conditions the first plagioclase to crystallize from most basic magmas, like that of basalts, has a bytownite composition. In basic igneous intrusions, plagioclase sometimes occurs in feldspar-rich bands formed by the accumulation of the primary precipitate plagioclase (the cumulus phase). In contrast, primary precipitate plagioclase may be absent from the ferromagnesian-rich bands in which the feldspar occurs as a product of the crystallization of intercumulus liquid. In the differentiated basic rocks of layered intrusions, the compositional range of the plagioclase is normally restricted to between An_{85} and An_{30}. In addition to the plagioclase-rich bands of layered intrusions, plagioclase also occurs as the only essential constituent of large masses of anorthosite; in these rocks the plagioclase may be bytownite, labradorite, or andesine in composition. Plagioclase feldspars are the main constituent of dolerites and many other hypabyssal rocks.

In metamorphic rocks, the composition of the plagioclase is generally related to the grade of the host rock. Thus, albite is the stable plagioclase in the chlorite and biotite zones of regional metamorphism, occurring in such rocks as chlorite-biotite-epidote-albite amphibolites and chlorite-albite schists; the anorthite component is not present in notable quantity until the garnet zone. Plagioclase between An_5 and An_{20} in composition is absent in low- to medium-grade schists; this compositional break (the peristerite gap) corresponds with the change in grade between the greenschist and the almandine amphibolite facies. In calc-silicate rocks of the amphibolite facies, clear compositional minima are seen corresponding also with the Bøggild and Huttenlocher gaps (Figure 8). The plagioclase feldspars of the intermediate and acid rocks of the granulite facies commonly are sodic andesines; thus in the rocks of the charnockite series, the majority of the plagioclase is between An_{30} and An_{35} in composition, although labradorite is present in some of the more basic charnockitic rocks. Plagioclase is not stable in the P–T environment of the eclogite facies; under these conditions the albite and anorthite components enter the compositions of omphacite and garnet, respectively. In the thermal metamorphism of calcareous sediments plagioclase again becomes increasingly calcic as the grade increases, with the production of anorthite marking a high grade of metamorphism.

In sedimentary rocks, plagioclase feldspars may occur via a variety of sources, including detrital minerals, authigenic minerals, and the products of volcaniclastic activity. In arkosic sediments, K-feldspar is usually dominant over plagioclase, but this varies according to the source rock and the weathering

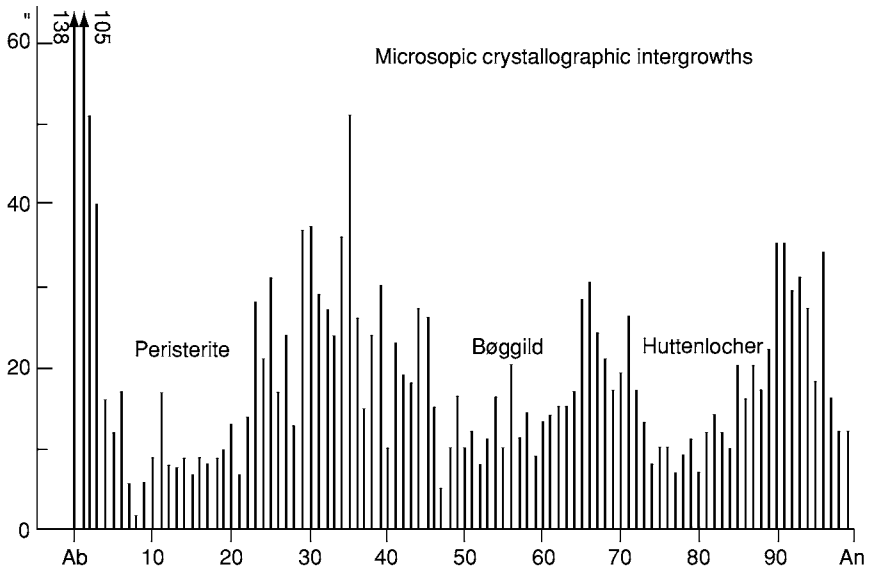


Figure 8 Frequency distribution of plagioclase compositions from 1883 microprobe results from 147 metamorphic calc silicate rocks in the central Alps (after Wenk *et al.* (1991)).

conditions to which they have been subjected. In meteorites, a whole range of plagioclase compositions occur, but there are two main clusters at around An_{90} and An_{15}. In lunar rocks the great majority of feldspars are calcic plagioclase An_{90} to An_{95} representing the lunar anorthosites.

See Also

Igneous Processes. **Igneous Rocks:** Granite. **Minerals:** Definition and Classification; Other Silicates. **Rocks and Their Classification**. **Sedimentary Rocks:** Mineralogy and Classification; Sandstones, Diagenesis and Porosity Evolution. **Solar System:** Meteorites; Moon.

Further reading

Deer WA, Howie RA, and Zussman J (1992) *An Introduction to the Rock Forming Minerals,* 2nd edn. London: Longman.

Wenk E, Schwander H, and Wenk H R (1991) Microprobe analyses of plagioclases from metamorphic carbonate rocks of the central Alps. *European Journal of Minerals* 3: 181 191.

Feldspathoids

M D Welch, The Natural History Museum, London, UK

Structure

Feldspathoid structures consist of rings of AlO_4 and SiO_4 tetrahedra that are connected to form a fully polymerized three-dimensional framework enclosing cavities occupied by large interstitial monovalent or divalent cations (sodium, potassium, calcium, and caesium) and, in some cases, anions (chloride, sulphide, and sulphate). These cavities are aligned to form channels that pass through the structure. Compared with feldspars (*see* **Minerals:** Feldspars), the feldspathoids are silica poor.

Three main tetrahedral-framework topologies are found in the feldspathoid group, and these are represented by the minerals nepheline, leucite, and sodalite. They are distinguished primarily by their different ring types and by how these rings are connected to each other. The nepheline structure consists of rings of six tetrahedra, whereas the structures of leucite and sodalite consist of both four-membered and six-membered rings.

Nepheline Subgroup

Minerals in the nepheline subgroup have the chemical composition ABO_4, where A is Na, K, Ca, or vacant,

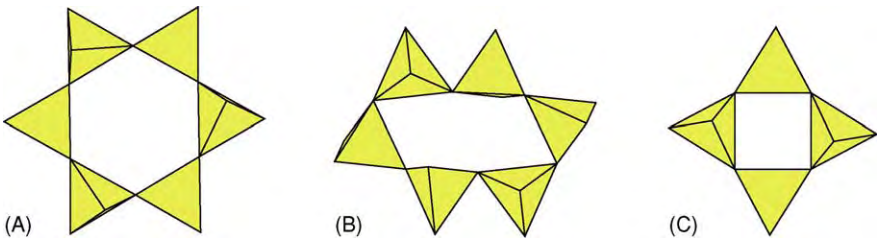


Figure 1 The different tetrahedral rings in feldspathoids. (A) The undistorted six membered rings exhibited by the high temperature form of tridymite. (B) The highly distorted six membered rings exhibited by nepheline and the low temperature form of tridymite. (C) The four membered rings that occur in leucite and sodalite.

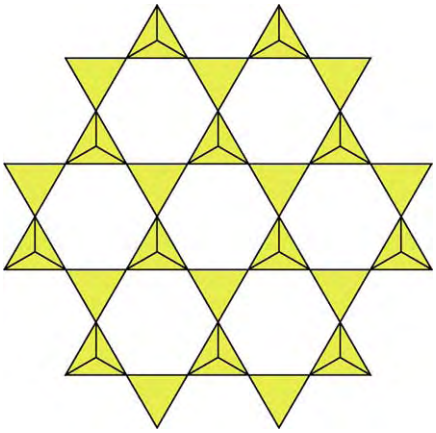


Figure 2 A sheet of undistorted hexagonal rings, as occur in the high temperature form of tridymite.

and B is Al or Si. The tetrahedral framework of nepheline-type structures is related to that of the mineral tridymite (*see* **Minerals:** Quartz), which has two polymorphs: a high-temperature largely undistorted hexagonal framework collapses upon cooling to a low-temperature structure in which all the six-membered rings are highly distorted (Figure 1B). A sheet of undistorted hexagonal rings in the high-temperature form of tridymite is shown in Figure 2. Half the tetrahedra point up; the other half point down. In the nepheline structure these sheets are connected to each other by apical oxygen ions to form cavities that are aligned into channels running parallel to the crystallographic *c*-axis. Insertion of cations into the channels causes distortions of the framework that are similar to those associated with the thermally induced collapse of high-temperature tridymite to low-temperature tridymite. For this reason, minerals of the nepheline group are often referred to as having 'stuffed tridymite' structures. The common stoichiometry of natural nephelines is $Na_3KAl_4Si_4O_{16}$, which arises from the ordering of potassium and sodium in the two different channels (Figure 3).

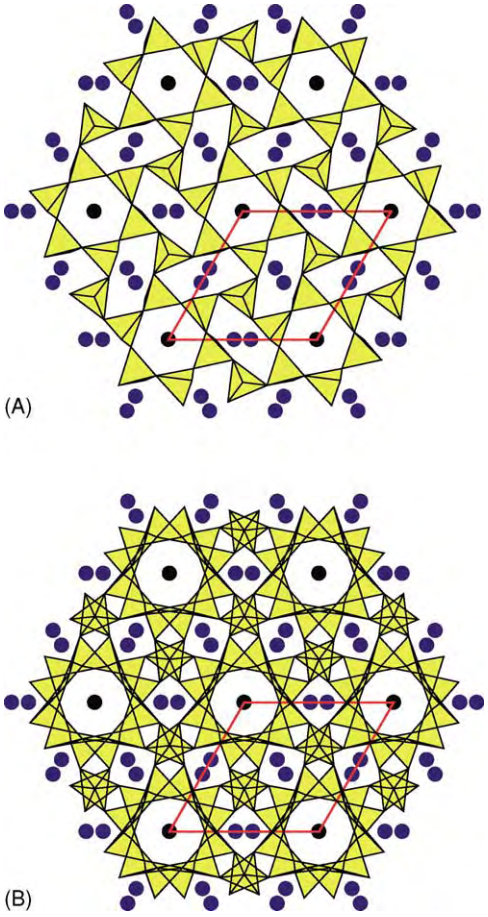


Figure 3 (A) The characteristic sheet of distorted and undistorted six membered rings of AlO_4 and SiO_4 tetrahedra in the nepheline type structure. Blue circles are sodium ions occupying smaller channel sites above and below the sheet, and black circles are potassium ions occupying the larger regular channels. The projection of the unit cell is shown in red. (B) A complete double sheet of the nepheline type structure, showing the two kinds of channel.

Leucite and Sodalite Groups

The leucite and sodalite structures differ from the nepheline structure in that they also contain squares of four tetrahedra (Figure 1C). However, the ways in

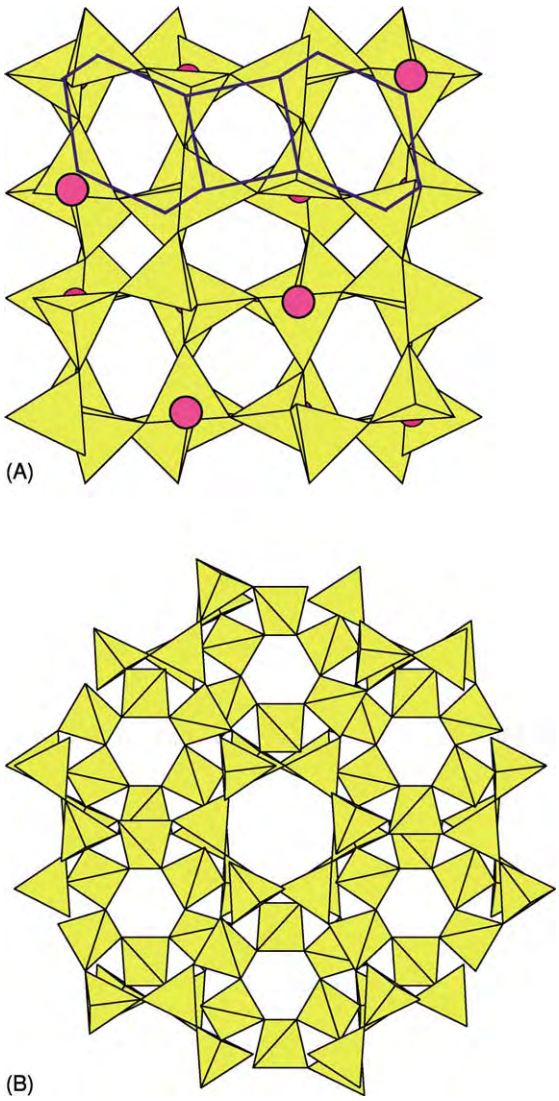


Figure 4 (A) The leucite structure is composed of connected four membered and six membered rings of tetrahedra. Two six membered rings and one four membered ring are shown in blue. Some potassium cations in the channels are shown in pink. (B) A view of the high temperature cubic leucite structure looking down the <111> channels. Channel cations are omitted for clarity. In leucite these channels do not intersect, but in sodalite they do.

which the six-membered and four-membered rings are connected to each other differ in the leucite and sodalite structures. In leucite, the hexagonal rings are aligned to form a set of non-intersecting channels in four orientations corresponding to the triad axes of the cubic high-temperature phase. In sodalite, these channels intersect to form very large cavities. Figure 4 shows the structure of leucite, in which the four-membered and six-membered rings are connected and potassium ions occupy the channels. Natural leucites are usually extensively twinned as a result of a reduction in symmetry upon cooling from the high-temperature cubic polymorph to a tetragonal structure. In end-member leucite, $KAlSi_2O_6$, this transition occurs at about 625°C.

The leucite group has the general chemical formula $AB_3O_6 \cdot xH_2O$, where A is K, Na, or Cs, B is Al or Si, and x is between zero and one. The three main minerals of the group are leucite ($KAlSi_2O_6$), analcite ($NaAlSi_2O_6 \cdot H_2O$), and pollucite ($CsAlSi_2O_6 \cdot xH_2O$, where $x = 0$ to 1). Water molecules occur in the larger channels.

The structure of the sodalite group differs from the leucite structure primarily in that the channels intersect to form large cavities, which are occupied by chloride, sulphide, and sulphate anions. The hexagonal channels are occupied by sodium or calcium ions. The sodalite group has three main mineral types: sodalite (*sensu stricto*) ($Na_8Al_6Si_6O_{24}Cl_2$), nosean ($Na_8Al_6Si_6O_{24}SO_4$), and häuyne ($(Na,Ca)_{4-8}Al_6Si_6O_{24}(SO_4,S)_{1-2}$).

Natural Occurrences of Feldspathoids

Nepheline-group feldspathoids occur predominantly in silica-undersaturated igneous rocks such as syenites and phonolites, of which they are major diagnostic minerals. More rarely, nepheline can occur as a metasomatic alteration product in metamorphosed sedimentary rocks. Nepheline is used in the manufacture of ceramics and glasses.

Leucite-group and sodalite-group minerals occur only in alkaline volcanic rocks that are more siliceous than syenites and phonolites. Leucite itself is unstable and, upon slow cooling, can transform to pseudomorphs of potassium feldspar and nepheline.

See Also

Igneous Processes. **Minerals:** Definition and Classification; Feldspars; Quartz; Zeolites.

Further Reading

Deer WA, Howie RA, and Zussman J (2004) *Rock Forming Minerals*, vol. 4B, *Silica Minerals, Feldspathoids, and Zeolites*. London: Geological Society.

Palmer DC (1994) Stuffed derivatives of the silica polymorphs. *Reviews in Mineralogy* 29: 83 118.

Glauconites

J M Huggett, Petroclays, Ashtead, UK and The Natural History Museum, London, UK

Green clay can be Fe smectite, glauconitic-smectite, smectitic-glauconite, berthierine, odinite (formerly phyllite v), a ferric chlorite (phyllite c), chamosite, ferric illite, or celadonite. Although the environment of deposition can be a guide to identification, careful chemical and mineralogical analysis is required to unequivocally identify green clay minerals. The green granule-forming clays are glauconitic, odinite, berthierine, or rarely chamosite, all forming where the rate of deposition is slow. Glauconite is a green, ferric iron-rich micaceous mineral that forms in marine sediments. Odinite and ferric chlorite are the ferric iron-rich precursors of the ferrous iron-rich clays berthierine and chamosite. Together odinite and ferric chlorite comprise the verdine facies in recent sediments. Odinite is only known from marine sediments, but berthierine and chamosite can be non-marine. An assemblage of mainly glauconitic granules is referred to as the glaucony facies, whereas an assemblage of mainly iron-rich 7-Å granules is called the verdine facies (although the latter term is used for recent assemblages rather than ancient ones).

Composition

Glauconite (*sensu strictu*) is an Fe-rich, K-rich dioctahedral mica with tetrahedral Al (or Fe^{3+}) usually >0.2 atom per formula unit and octahedral R^{3+} correspondingly >1.2 atoms. Typically 5–12% of the total iron is ferrous (Table 1). The structural formula may be written as

$$K_{(x+y)}(Si_{4-x}, Al_x)_{\sim 4}(Fe^{3+}, Al, Mg, Fe^{2+})_{\sim 2} O_{10}(OH)_2$$

where x is 0.2–0.6 and y (the sum of the divalent octahedral cations) is 0.4–0.6.

It is distinguished chemically from ferric illite in having more total iron, and from celadonite in having greater substitution of aluminium for silicon in the tetrahedral layer and a higher octahedral charge (Table 1). Glauconitic granules that have replaced faecal pellets are almost always inhomogeneous, with inclusions of common detrital minerals: clays, quartz, feldspars, carbonates, or apatite. Glauconitic-smectite is a mixed layer clay that is less potassic than glauconite and which characterizes immature glauconitic granules. Fe smectite is a ferrous iron-bearing dioctahedral smectite, half way between montmorillonite and nontronite (iron-rich smectite). Odinite is a poorly ordered, ferric iron-rich, 7-Å 1:1 green clay with ~0.8 $Fe^{3+}/Fe^{2+} + Fe^3$ and low tetrahedra substitution of Al_2O_3 for SiO_2. It is believed to be the precursor of the poorly – ordered, ferrous iron-rich, 7-Å green clay berthierine, which has ≤0.2 $Fe^{3+}/Fe^{2+} + Fe^3$ and less Mg than odinite. Phyllite C has

Table 1 Chemical composition of various glauconites

	1	*2*	*3*	*4*	*5*	*6*
SiO_2	48.17	50.87	49.73	50.18	35.41	27.20
Al_2O_3	4.25	3.71	9.78	6.74	15.73	20.33
TiO2	0.07	0.07	0.05	0.00	0.00	0.00
Fe_2O_3	23.44	23.00	21.05	23.43	26.99	?
FeO[a]	2.34	2.30	2.10	2.34	1.30	37.96
MnO	0.04	0.04	0.03	0.01	0.14	0.00
MgO	3.59	3.93	4.34	2.77	4.95	2.97
CaO	0.09	0.29	0.22	1.10	0.00	0.00
Na_2O	0.41	0.52	0.15	0.56	0.00	0.00
K_2O	8.55	7.19	6.56	6.89	0.00	0.00
Total	90.95	91.92	94.01	94.02	84.52	88.46
Si	7.49	7.70	7.24	7.43	3.79	2.72
Al	0.51	0.30	0.76	0.57	0.21	1.28
Total tetrahedral	8.00	8.00	8.00	8.00	4.00	4.00
Al	0.27	0.36	0.92	0.60	1.76	1.12
Ti	0.01	0.01	0.01	0.00	0.00	0.00
Fe^{3+}	2.51	2.39	2.09	2.39	2.40	3.01
Fe^{2+} [a]	0.30	0.29	0.26	0.28	0.12	0.00
Mn	0.01	0.01	0.00	0.00	0.00	0.00
Mg	0.83	0.89	0.94	0.61	0.79	0.44
Total octahedral	3.93	3.95	4.21	3.89	5.07	4.57
Ca	0.02	0.05	0.03	0.17		
Na	0.12	0.15	0.04	0.16		
K	1.70	1.39	1.22	1.30		
Total interlayer	1.84	1.59	1.29	1.63		

1 Shallow water glauconite (Pliocene) from ODP leg 174 A,
2 deep water glauconite (Oligocene) from ODP leg 174 A,
3 smectitic glacuonite from the Weches Formation (Eocene, Texas),
4 smectitic glauconite from the Karai Formation (Cretaceous, India).
5 odinite from Weches Formation (Eocene, Texas),
6 berthierine. Interlayercations in berthierine are due to contamination by micaceous minerals. Number of cations are on the basis of O20(OH)4 for glauconitic clays and O10(OH)8 for odinite and berthierine.

[a]FeO is assumed to be 10% of the total iron for glauconite, the odinite FeO is <5% of total Fe (analysed by EELS).

chemistry similar to that of odinite but a 14-Å 2:1 lattice structure similar to that of chlorite; it is only found in recent verdine granules. Chamosite is chemically similar to berthierine, but has a 14-Å 2:1 clay structure; some chamosite forms by the replacement of berthierine.

Formation

Granules formed from faecal pellets are often referred to as pellets; where the origin is not clear, the term peloid may be used. Faecal pellets are the main substrate for glaucony formation in both ancient and recent sediments. For iron-rich, 7-Å clay they are the main substrate only in recent sediment; most ancient berthierine (and some chamosite) occurs as ooids or pore-lining or pore-filling clay. Other modes of occurrence are replacement of mica, quartz, chert, feldspar, calcite, dolomite, phosphate, and volcanic rock, cementation of microfossil cavities, and formation of thin films on limestone hardgrounds, quartz grains, and flint pebbles. The current consensus on the origin of iron-rich clays is that they form by direct precipitation, close to the sediment–water interface. The crucial component of the faecal pellets and microfossil cavities that predisposed them to replacement by iron-rich clay is the presence of labile organic matter, the composition of the clay-rich clastic component apparently having little if any affect upon the mineralization process. It is likely that microbial oxidation of the organic matter may be critical in creating redox conditions favourable to iron fixation and detrital clay dissolution. The formation of glaucony and iron-rich, 7-Å clays (verdine) is favoured by a semiconfined oxic microenvironment. This is indicated by the following observations:

1. Green clays form in microfossil tests.
2. Grains $<100\,\mu$m are commonly only slightly glauconitized, whereas grains of 100–500 μm are typically homogeneously glauconitized. Grains >1 mm have a zone of most intense glauconitization just inside the margin.
3. There is an absence of diffuse glaucony or iron-rich, 7-Å clay layers from sediment.
4. There is a concentration of green granules in invertebrate burrows.

Glauconitization

The iron in glauconite and its precursors is mainly ferric (unusual for a sedimentary silicate), but mobile iron is ferrous; hence, it is believed to form close to the redox boundary. The first stage in glauconitization of faecal pellets is the dissolution of carbonate, detrital clay and formation of Fe(III) smectite, while the organic matter content decreases. The overall Fe, and to a lesser extent K_2O (2–4%), content of the faecal pellets increases while the Al content decreases. At this stage the granules have the particulate or vermiform fabric of the detrital faecal pellets are ochre to light green in colour and soft; i.e., reworking will disperse them. This is the 'nascent' stage in the classification of Odin (see Further Reading). The 'slightly evolved stage' is characterized by the almost complete disappearance of detrital minerals (nonclay, noncarbonate minerals are the most resistant to replacement). Fe and K uptake continues through this phase, which is characterized by 4–6% K_2O, 20–28% Fe oxides, and a mixed layer expandable/10-Å clay composition. At this stage the pellets are olive green, and in backscatter SEM images still have a particulate or vermiform appearance (**Figure 1**). The 'evolved stage' is the result of recrystallization where dense, indurated dark green pellets, frequently cracked at the margins, are formed. The original fabric is destroyed and X-ray diffraction spectra show a strong reflection at 10 Å (diagnostic of potassium-bearing micas, including glauconite). At this stage there are few expandable layers remaining, K_2O content is 6–8%, and the iron content is no higher than that for the slightly evolved stage. In-filling of fractures with glauconite represents the highly evolved stage. By this stage the granule is made of end-member, *sensu strictu* glauconite with $>8\%$ K_2O. Note that iron is fixed in the octahedral layer before potassium fixation is complete. It is also potassium content that increases as the layer charge increases (due to loss of swelling layers and

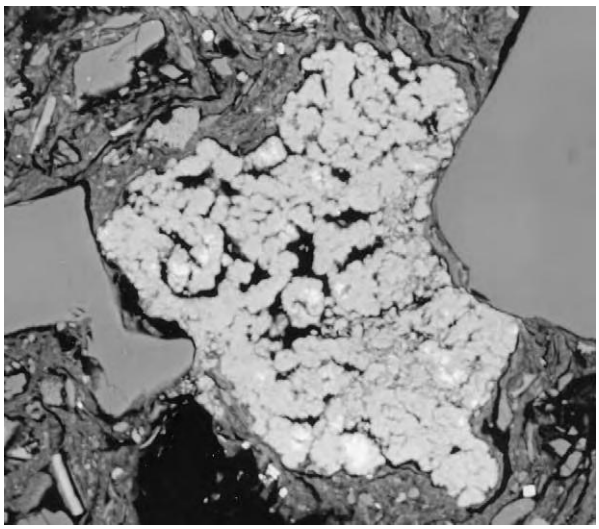


Figure 1 Thin section of vermiform (white in this picture) glauconite pellets in the ''slightly evolved'' stage of glauconization.

reduction of octahedral iron), and is consequently the only reliable chemical indicator of glaucony maturity. If glauconitic clay is not buried, it will remain stable so long as there is not a major sea-level fall. In a very shallow marine, fluvial, or palaeosol environment, glauconite will oxidize to kaolinite and goethite. Recently it has been found that tunnelling into glauconitic sand can result in a sudden (and harmful to life) drop in oxygen level, due to oxidation of the ferrous iron in glauconitic clay to ferric oxides or hydroxides. This instability only affects incompletely evolved glaucony; the use of sandstone containing highly evolved glauconite as building stone that has remained green over several centuries demonstrates this. In a reducing environment glauconite may be replaced by phosphate. Burial to more than a few decimetres halts the glauconitization process. Hence, burial prior to the glaucony attaining the fully mature state will preserve the immature state. Burial diagenesis results in increasing substitution of Al for Si, a decrease in swelling (smectite) layers, and an increase in crystallinity; hence, the chemistry of ancient glauconite differs slightly from that of Recent and Quaternary glauconite.

Glauconite forms coatings on particles and as laterally extensive (many squared kilometres) surfaces on hardground. Coatings of glauconite on particles are most frequent on flint pebbles in chalk and on bioclasts. Thin section study of these films shows that the clay mainly develops along fissures, biogenic borings, and any other void of suitable size to provide the semiconfined microenvironment necessary to glauconite formation. The thickness of the film depends upon the substrate and extent of alteration: on silicates it is $\leq$3 mm for silicates and $\leq$1 cm for carbonates. Film glaucony is commonly present in Jurassic limestone and Late Cretaceous chalk and provides useful stratigraphic markers.

The origin of the elements necessary for the formation of glauconite has been the subject of much discussion. Experimental work has indicated that a high silica concentration is necessary for glauconite and iron-rich smectite formation, as at low silica concentration iron-rich, 7-Å clays will form instead. This may link glaucony formation to the presence of biogenic silica.

The concentrations of Fe, Al, and K in normal seawater are also too low for direct precipitation of glaucony, although in river water the concentrations of Fe and Si are sufficient. Reducing conditions develop as soon as all dissolved oxygen in the pore fluid is used up in microbial oxidation of organic matter. Subsequently iron-reducing bacteria become the major consumers of organic matter until the sediment is buried sufficiently deep for sulphate-reducing bacteria to become active. In this narrow depth zone, just below the sediment–water interface in oxic sediments, the concentration of dissolved (ferrous) iron increases. Where there is a low sedimentation rate, intraformational reworking, or winnowing, much of the dissolved iron will diffuse towards the oxic zone and be precipitated as ferric iron hydroxides. Such precipitates can sorb silica, Fe^{2+}, Al, K, Mg, B, and other ions from solution. Experimental work has shown that these poorly crystalline materials can reprecipitate as iron-rich smectite. Glauconite cannot be directly precipitated at the low K concentrations in natural pore water, but evolves gradually through increased interlayer charge and K fixation.

Studies of recent glaucony have provided information on the rate of glauconitization. On the Congolese continental shelf glaucony from depths $<$110 m is younger than 20 000 years and has an average K_2O content of 5%, indicating that it is slightly evolved, whereas the age of recent glaucony off Vancouver Island is estimated to be $<$3000 years. Exceptionally, very small mica grains in nonevolved faecal pellets may undergo K and Fe enrichment in less than 10 years. In general it is thought that a slightly evolved granule typically takes the order of 10^4 years to form, and a mature pellet 10^5–10^6 years. However, laser-argon probe dating suggests that the period of genesis may be as long as $\sim$5 million years. It is therefore possible to use glaucony maturity as an indication as to the length of a hiatus. However, the extent and intensity of glauconitization appears to depend upon the size of the host material.

Verdinization

Verdine granules are mostly odinite, with minor ferric chlorite and detrital minerals; they occur only in recent sediments and have never been buried. Also referred to as the process of *verdissement*, the replacement of faecal pellets by iron-rich, 7- and 14-Å minerals follows a maturation process similar to that for glauconite, although without the interlayer K it is less easy to monitor the degree of chemical evolution. Verdine formation requires oxic pore waters because both odinite and the less frequent ferric chlorite are rich in ferric iron. The formation of verdine is less well known than that of glauconite. It follows a similar morphological transformation from nascent to evolved, though detrital minerals may be more persistent (Figure 2). It has been suggested that the first part of the process is odinite formation, followed by formation of ferric chlorite, and finally replacement by ferrous iron-rich clays. However, as very few intermediate stages between recent odinite and diagenetically modified chlorite and berthierine deposits have

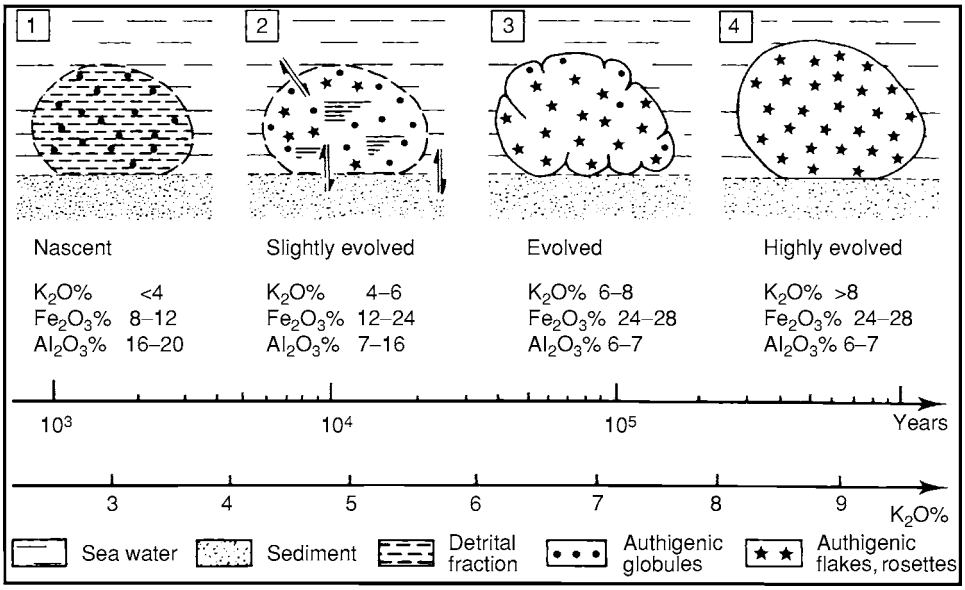


Figure 2 Morphological transformation from nascent to highly evolved glauconite.

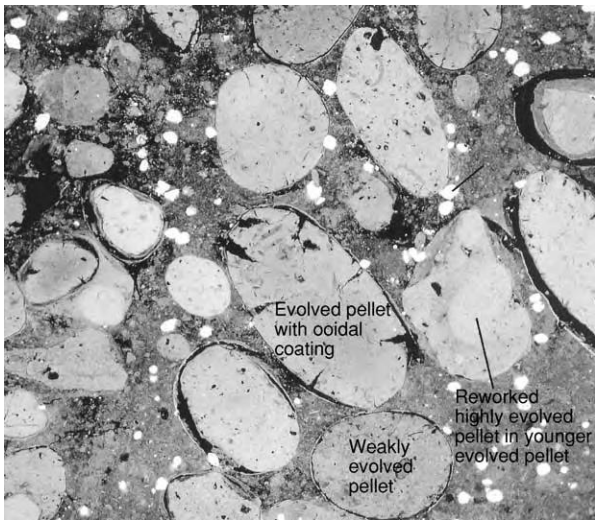


Figure 3 Thin section of faecal pellets with different stages of maturation towards full glauconization.

yet to be identified, the reaction pathway or pathways is not known with certainty. In some of the older literature it is suggested that berthierine forms by reaction of goethite and kaolinite. Although this may be true in some instances, it seems more likely that, like glaucony, formation of 7-Å, iron-rich clays is substrate independent. Verdinization may, like glaucony formation, follow a general pattern of iron-rich smectite formation followed by iron-rich, 7- and 14-Å clay formation (an example is the Claibourne Group of Texas, rich in pellets of an odinite-like mineral). Before it was realized that ferric iron-rich clays can be precursors of chlorite and berthierine it was commonly argued that these clays are indicative of a reducing environment at the time of sedimentation, rather than during diagenesis. It is possible to find many examples of this often-erroneous view in the literature.

Verdine is known only as granules and microfossil test in-fills. In contrast chamosite chlorite and berthierine frequently occur as ooidal coatings on grains, as pore-linings, replacement of detrital chlorite or discrete 'cryptocrystalline' clay, and only rarely as granules. This discrepancy has not been explained, although it might in part be due a widespread presumption that green granules are glauconitic. The odinite-like mineral found in the Claibourne Group of Texas does occur as ooidal coatings, but only rarely (Figure 3).

X-Ray Diffraction Identification of Green Clays

Optically the two various green granules are indistinguishable; hence, X-ray diffraction can assist in identifying the composition and purity of such granules. Due to their much higher Fe content, glauconite and ferric illite have a 002 (~5 Å) reflection much weaker than that of illite. Additionally, the iron-rich clays have a diagnostic 060 reflection at 1.51 Å. However, where both Fe-rich and-poor, 10-Å clays are present, this will not be so pronounced as that for the pure clays. All except the most evolved glauconitic clay will typically have some swelling clay layers, which can be identified by the broad area of reflection intensity between 10 and 14 Å and its shift to $\leq$17 Å

upon treatment with glycol and collapse to 10 Å upon heating at >400°C. Verdine granules yield a broad 001 reflection at ~7.2 Å, which is the strongest reflection of odinite, and also of berthierine, but the latter has a slightly smaller 002 reflection. Chlorites (both ferrous and ferric) are distinguished from other green clays by their 001 reflection at ~14 Å, which is unaffected by glycol treatment but shifts to a slightly smaller spacing upon heating at >500°C.

Distribution

In Time

Glaucony is absent from Precambrian sediments, but two major episodes of glauconite (and chamosite/berthierine) formation occurred during Phanerozoic time, the first in the Early Palaeozoic, and the second in the Middle-Late Mesozoic. Both episodes were characterized by temperate to subtropical climates, dispersed cratonic blocks, global sea-level rise (transgressive sediments), and periods of low shelf sedimentation. After the break-up of the Precambrian supercontinent about 600 My ago, glauconite became increasingly common through the Cambrian as cratonic blocks were rifted and dispersed from the Late Proterozoic supercontinent, the sea-level rose, and invertebrate faecal pellet makers became abundant. Glauconite formation in the Palaeozoic peaked in the Late Cambrian–Early Ordovician (*see* **Palaeozoic:** Ordovician). From then until the Cretaceous glauconite is relatively uncommon. The greatest period of glauconite formation occurred in the Cretaceous, notably along the continental margins of the widening North Atlantic Ocean, when global sea-level was again high (*see* **Mesozoic:** Cretaceous). Abundant glaucony continued until the Middle Cainozoic.

The earliest chamositic granules formed in the Early Proterozoic in association with greenalite. Chamosite ooids are not known until the Phanerozoic, and berthierine is only known from a few banded iron formations prior to the Mesozoic, whereas ancient odinite-like minerals are recorded only from the Eocene and younger rocks. The formation of ferrous clays in the Proterozoic was probably favoured by the lack of atmospheric oxygen. The major periods of chamosite/berthierine formation occurred when cratonic blocks were dispersed during the Ordovician, Devonian, and Jurassic–Cretaceous eras. Also global temperature may have effected the variation in abundance of chamosite and berthierine. In the Eocene a decrease in maturity of odinite-like mineral granules and an increase in Fe smectite may be linked to global cooling.

In Space

Recent glaucony apparently forms principally on the outer shelf at a water depth of 100–300 m, at a temperature of <20°C (possibly <13°C) and between 50° N and 50°S, although the true depth and latitude range may be masked by Late Pleistocene and Holocene sea-level fluctuations. Glaucony occurs where there are bottom currents that cause constant remobilization of sediment and winnowing. This association occurs because the intraformational reworking results in grains being exposed at the seafloor for sufficient time to permit glauconitization. Most recent glaucony occurs around the margins of the North and South Atlantic, and the Pacific west coast of the Americas, with smaller occurrences around the Mediterranean Sea and the Gulf of Oman. Glaucony is particularly abundant (up to 90% of the sediment) at the shelf-slope transition; however, apparently *in situ*, unweathered glauconite has been reported from ancient marine sediments deposited beneath 10–30 m water (e.g., in marginal marine to estuarine Tertiary sediments of the Hampshire basin, UK) and from clinoform toe sediments that must have been deposited beneath at least 600 m water. Very deep water glaucony (2000–3000 m) occurs in some recent sediments, as on the Chatham Rise, on the Scotia Ridge, and on both sides of Japan. The glauconite in the Chatham Rise and the Scotia Ridge sediments has not been investigated in detail, whereas that around Japan is associated with a fauna indicative of water shallower than the present depth. This could be due to either redeposition or rapid subsidence in this tectonically active area. Examples of ancient deep-water glauconitic sediments are the Miocene turbidites south-west of the Rockall Plateau (North Atlantic) and in Cretaceous turbidites in south-east France, e.g., the 'Petit Verolles'. The Petit Verolles is concentrated in coarser laminae, which suggests it was redeposited by turbidity currents along with the rest of the sediment with which it occurs. Glaucony in turbidites typically comprises fragments of mature pellets, because immature pellets are relatively soft and so become disaggregated during transport. It is likely that the precipitation temperature (which is reflected in the latitudinal constraints) and the availability of iron and potassium are ultimately more important than water depth. The importance of temperature is illustrated by the decrease in neoformed glaucony throughout the period of shallow marine sedimentation in the Eocene of the Hampshire basin (UK). This decrease coincides with a period of global

temperature decrease, whereas all other factors for glauconitization remained favourable. Indeed, the presence of the appropriate bacteria (which may be temperature-controlled) for organic matter degradation and iron reduction, plus faecal pellets (which are most abundant on the continental shelf in water <100 m deep), may prove to be the ultimate controls on glauconite formation.

Verdine in recent sediments is much more restricted than is glaucony. Its formation requires water temperatures >20°C; hence it is only known from the tropics. Verdine preferentially forms where the water depth is 20–60 m (but <5 m in the Casamance Estuary, Senegal, and up to 200 m elsewhere), where large amounts of land-derived iron compounds have been deposited in shallow marine sediments. Major recent examples are seaward of the Congo River, the Niger Delta, and between the mouths of the Amazon and Orinoco rivers. Recent verdine is also associated with iron derived from volcanic rocks, as in the case of offshore from New Caledonia. Berthierine, which in some instances has probably replaced odinite as the principal component of verdine, apparently precipitates over a much wider temperature range. For example, it is known from Arctic palaeosols and estuarine sediments of western Scotland. However, like glaucony, berthierine can be an indicator of hiatus, and in the case of the latter mineral this applies to nonmarine as well as to marine sediments. Hence berthierine and chamosite may be of use for onshore sequence stratigraphic interpretation.

The different types of substrate for replacement by green clays in a marine environment are typically found at different water depths. For example, on the Western African margin there is a succession of substrates from planktonic foraminiferans on the outer shelf margin, to benthic bioclasts, faecal pellets, mineral grains, and shell bioclasts in water <5 m deep. Of these the deeper water substrates are glauconitized, and the shallower ones are replaced by verdine. Such a succession is only possible in a period of stable sea-level. During a transgression, successive substrates will be glauconitized. Note that this produces a glauconitic sediment that may be used for facies correlation, although the age of the sediment will decrease up-dip.

Because green clay granules are slow to form, they are favoured by a low sedimentation rate, which may partially explain the depth constraint. The slow rate of formation also explains the common association of abundant glaucony with transgressive sediments. The frequent association of glaucony with transgressions and condensed sections predisposes it to reworking. Indeed, any glaucony found outside the range suggested above should be checked carefully for evidence of reworking. There have been several reports of nonmarine glauconite but detailed investigation has shown that it was reworked from underlying marine deposits.

The principal criteria for evidence of *in situ* granules are:

1. Concentration of glaucony in burrows rather than the host sediment; and
2. The presence of immature granules, which are soft and unlikely to withstand transport over more than a few metres; and
3. Deeply fractured mature granules (e.g., Plate), which would not survive transport without breaking up.

However, for reworked granules one or more of the following criteria may apply:

1. Fragments of granules;
2. Sand-size green granules in silt- or clay-dominated sediment;
3. Cross-bedding (which implies a high-energy environment), especially with laminae of green granules, whereas green clays are associated with slow sedimentation; and
4. Inclusions of material not representative of the host sediment; at the immature stage glaucony or verdine has a mineralogy similar to that of the enclosing sediment.

Age Dating

Potassium-rich glaucony provides ~40% of the absolute age dates for the geological time-scale of the past 250 million years. However, there are potential problems with this method of K/Ar radiogenic isotopic age dating:

1. Argon loss.
2. Residual detrital illite or K feldspar will result in incorrect (too old) dates. Both K/Ar and Rb–Sr apparent ages progressively decrease as glauconitization proceeds, but even the most evolved grains may retain around 10% of the initial substrate, which may affect the apparent age, depending upon its mineralogy. This is true for both ancient and recent glaucony.
3. Potassium-rich glaucony, being the most resistant to weathering, is the most likely to survive reworking; hence, ages greater than time of sedimentation may be measured.

4. A single glaucony assemblage may represent a cluster of glauconitization events; hence, the bulk age will be an average of these.

To get a meaningful, if not completely accurate date, the glauconite should be *in situ*, then sieved, mixed with distilled water, and subjected to one hour in an ultrasonic bath to break up the less-evolved granules. The remaining intact granules can then be separated off by washing, and if necessary any feldspar can be separated by magnetic separation. *In situ* glaucony with >7% K_2O gives the most reliable dates. Argon loss can be avoided by the microencapsulation technique. And clustered glauconitization events can be individually identified by single-grain age dating.

Ferric Illite

Ferric illite (sometimes misleadingly called nonmarine glauconite) most often forms green illite-rich claystones rather than granules. The iron content is much less, and the aluminium content is higher than that in marine glauconitic minerals. It forms in palaeosols and hypersaline lacustrine environments, but is not known to originate from marine environments. In lacustrine environments it is often associated with limestone and evaporite minerals, where alkaline, potassium-rich pore water favours illite precipitation. In palaeosols illite forms by the wetting and drying of smectite. Ferric iron in the smectite is reduced while the clay is wet, resulting in increased layer charge, loss of swelling capacity, and K fixation in the interlayer sites. Note that green clay in palaeosols can also be berthierine, but the latter is not associated with wetting and drying.

See Also

Analytical Methods: Geochronological Techniques. **Clay Minerals**. **Mesozoic:** Cretaceous. **Palaeozoic:** Ordovician. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Ironstones; Sandstones, Diagenesis and Porosity Evolution.

Further Reading

Amorosi A (1995) Glaucony and sequence stratigraphy: A conceptual framework of distribution in siliciclastic se quences. *Journal of Sedimentary Research B* 65: 419 425.

Bailey SW (1988) Odinite, a new dioctahedral trioctahedral Fe^{3+} rich 1:1 clay mineral. *Clay Minerals* 23: 237 248.

Brindley GW (1982) Chemical compositions of berthier ines a review. *Clays and Clay Minerals* 30: 153 155.

Chamley H (1989) *Clay Sedimentology*. Berlin: Springer Verlag.

Clauer N, Keppens E, and Stille P (1992) Sr isotopic constraints on the process of glauconitization. *Geology* 20: 133 136.

Hesselbo PS and Huggett JM (2001) Glaucony in ocean margin sequence stratigraphy (mid Cenozoic, offshore New Jersey, USA, ODP, Leg 174A). *Journal of Sediment ary Petrology* 74: 599 607.

Huggett JM and Gale AS (1997) Petrology and palaeoenvir onmental significance of glaucony in the Eocene succes sion at Whitecliff Bay, Hampshire Basin, UK. *Journal of the Geological Society, London* 154: 897 912.

Ireland BJ, Curtis CD, and Whiteman JA (1983) Compos itional variation within some glauconites and illites and implications for their stability and origins. *Sedimentol ogy* 30: 769 786.

Odin GS (1988) *Green Marine Clays,* Developments in Sedimentology, 45. Elsevier: Amsterdam.

Rao VP, Thamban M, and Lamboy M (1995) Verdine and glaucony facies from surficial sediments of the eastern continental margin of India. *Marine Geology* 127: 105 113.

Stille P and Clauer N (1994) The process of glauconitiza tion: chemical and isotopic evidence. *Contributions to Mineralogy and Petrology* 117: 253 262.

Toth TA and Fritz SJ (1997) An Fe berthierine from a Cretaceous Laterite: Part 1. Characterization. *Clays and Clay Minerals* 45: 564 579.

Van Houten FB and Purucker ME (1984) Glauconitic peloids, and chamositic ooids Favorable factors, constraints, and problems. *Earth Science Reviews* 20: 211 243.

Micas

R A Howie, Royal Holloway, University of London, London, UK

Introduction

The minerals of the mica group show considerable variation in their chemical and physical properties, but all are characterized by a platy morphology and perfect basal cleavage which is a consequence of their layered atomic structure. They all have a negative optic sign and have the optic direction α approximately perpendicular to their perfect cleavage. The most common micas are muscovite (or white mica), the phlogopite-biotite series (brown or black mica), and lepidolite (lithium mica), of which muscovite, phlogopite, and lepidolite are of economic importance. The classification of

micas has been reviewed recently by an international committee.

Crystal structure

The basic structural feature of mica is a composite sheet in which a layer of octahedrally coordinated *Y* cations is sandwiched between two identical layers of linked $(Si,Al)O_4$ tetrahedra. Two of these tetrahedral sheets, with a composition $(Si,Al)_2O_5$, are illustrated in Figure 1. On the left is a sheet in which all tetrahedra are pointing upwards (see the elevation view), and on the right is a sheet of tetrahedra which all point downwards. These two sheets are superimposed and are linked by a plane of octahedrally coordinated cations (Figure 2). Additional hydroxyl ions, together with the apical oxygens of the inward-pointing tetrahedra, complete the octahedral coordination of the sandwiched cations. The central *Y* ions determine the positions of the two tetrahedral sheets so that they are displaced relative to one another by *a*/3 in the [001] direction. Thus the micas, although superficially hexagonal, have a monoclinic unit cell (outlined in Figure 2). The perfect cleavage is along the weakly bound sheet of *X* ions (K, Na, etc.) lying between the tetrahedral layers.

The tetrahedral hexagons may be superimposed in six different ways. Thus one hexagon may be related to the next by rotation through 0° or by multiples of 60°; this, in combination with the stagger of *a*/3 introduced by the *Y* layer (Figure 2), determines the location of corresponding atoms in successive cells. Various sequences of layer rotations are possible, and when repeated regularly these build up unit cells with one, two, three, or more layers. The commonest stacking sequences lead to either one- or two-layered monoclinic polytypes (symbols 1M, $2M_1$), a different two-layered monoclinic ($2M_2$) or a three-layered trigonal (3T) polytype.

Chemistry

The general formula, which may be used to describe the chemical composition of micas, is $X_2Y_{4-6}Z_8O_{20}(OH,F)_4$, where:

X is mainly K, Na or Ca, but also Ba, Rb, Cs, etc.
Y is mainly Al, Mg or Fe, but also Mn, Cr, Ti, Li, etc.
Z is mainly Si or Al, but perhaps also Fe^{3+} and Ti

The micas can be subdivided into di-octahedral and tri-octahedral classes in which the number of *Y* ions is 4 and 6, respectively. Thus muscovite $K_2Al_4[Si_6Al_2O_{20}(OH,F)_4$ is di-octahedral whereas phlogopite $K_2(Mg,Fe^{2+})_6[Si_6Al_2O_{20}](OH,F)_4$ is tri-octahedral.

Di octahedral			
	X	*Y*	*Z*
Muscovite	K_2	Al_4	Si_6Al_2
Paragonite	Na_2	Al_4	Si_6Al_2
Glauconite	$(K,Na)_{1\ 2-2\ 0}$	$(Fe,Mg,Al)_4$	$Si_{7-7\ 6}Al_{1\ 0-0\ 4}$
Tri octahedral			
	X	*Y*	*Z*
Phlogopite	K_2	$(Mg,Fe^{2+})_6$	Si_6Al_2
Biotite	K_2	$(Mg,Fe^{2+},Al)_6$	$Si_{6-8}Al_{2-3}$
Zinwaldite	K_2	$(Fe, Li,Al)_6$	$Si_{6-7}Al_{2-1}$
Lepidolite	K_2	$(Li,Al)_6$	$Si_{6-5}Al_{2-3}$

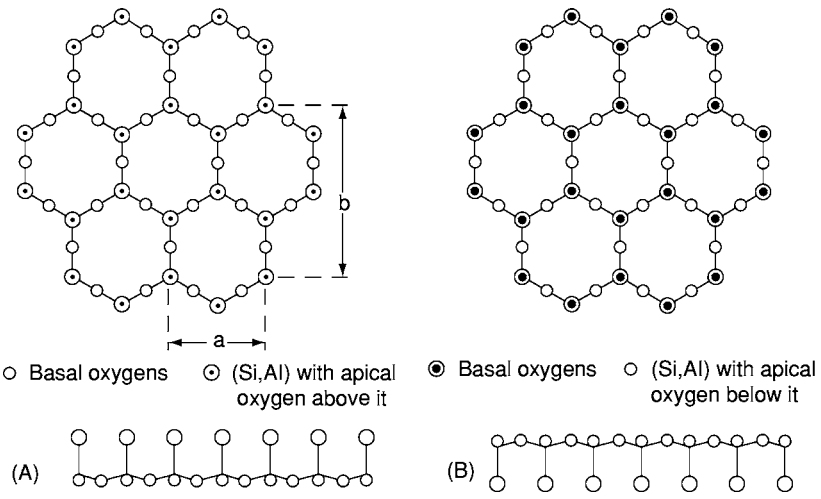


Figure 1 Tetrahedral layer $[(Si,Al)_4O_{10}]$ in an ideal mica structure viewed in *z* and *y* axis projection with tetrahedra pointing (A) upwards and (B) downwards. (After Deer *et al.* (1992).)

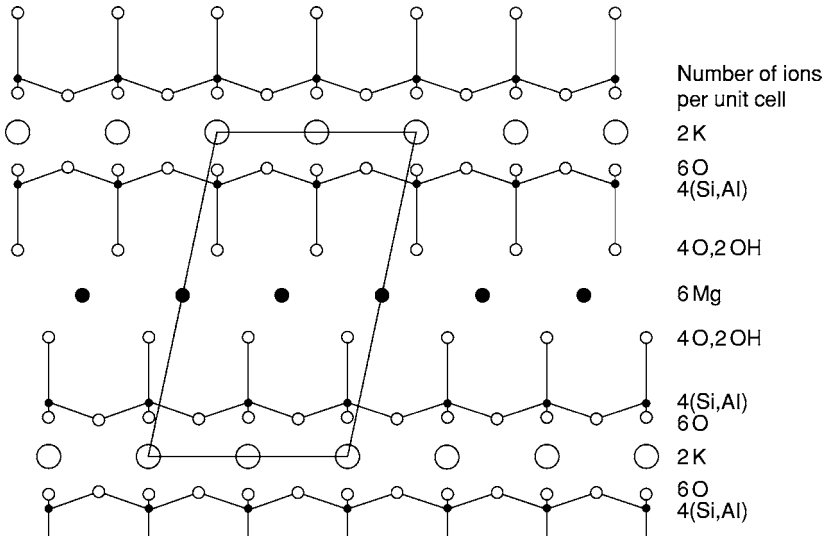


Figure 2 The crystal structure of phlogopite [$KMg_3AlSi_3O_{10}(OH)_2$] viewed in *y* axis projection, highlighting the three layer (tetrahe dral octahedral octahedral [*t o t*] sandwich unit, and the layer of interlayer cations (K^+). (After Deer *et al.* (1992).)

Some 40 mica species are recognized, but the more common micas are be listed as below. For muscovite and paragonite, a Si/Al ratio greater than 6:2 can be balanced by equivalent substitution of divalent ions for Al in the *Y* sites.

A chemical feature in most micas is their water content, which is generally around 4–5 wt% H_2O^+, except for those with a high fluorine content. The F/(F + OH) ratio rises in acid rocks, particularly in late-stage granites (*see* **Igneous Rocks:** Granite) and in the micas of pegmatites.

Optical and physical properties

The optical properties of micas cover a wide range but all have negative optic sign, low 2V, and have the α optic direction approximately perpendicular to their perfect cleavage. Birefringence is generally very weak in the plane of cleavage flakes but strong in transverse sections. Pleochroism is strong in coloured micas such as biotite, where it may range from yellow in the α direction to dark reddish-brown perpendicular to this direction. The perfect {001} cleavage of all micas is a useful characteristic. The sheet-like nature of the micas makes muscovite and phlogopite of economic interest in electrical insulation applications

Paragenesis

Minerals of the mica group occur in igneous, metamorphic, and sedimentary rocks. The principal occurrences are as follows:

i. *Igneous rocks*
   - Muscovite: granites, granitic pegmatites, and aplites.
   - Phlogopite: peridotites and kimberlites.
   - Biotite: gabbros, norites, diorites, syenites, granites, and pegmatites.
   - Lepidolite and zinnwaldite: pegmatites and high-temperature veins.

ii. *Metamorphic rocks*
   - Muscovite, paragonite, and biotite: phyllites, schists, and gneisses.
   - Phlogopite: metamorphosed limestones and dolomites.

iii. *Sedimentary rocks*
   - Muscovite and paragonite: detrital and authigenic sediments.
   - Glauconite: greensands.

See Also

Clay Minerals. **Igneous Rocks:** Granite. **Minerals:** Definition and Classification; Amphiboles; Other Silicates; Pyroxenes.

Further Reading

Deer WA, Howie RA, and Zussman J (1992) *An Introduc tion to the Rock Forming Minerals*, 2nd Ed. London: Longman.

Fleet M (2003) *The Micas. Rock Forming Minerals*, 2nd edn. Vol. 3A. London: Geological Society.

Rieder M, Cavazzini G, D'Yakonov YS, *et al.* (1999) Nomenclature of the micas. *Mineralogical Magazine* 63: 267 279. *Clays and Clay.*

Smith JV and Yoder HS, Jr (1956) Experimental and theor etical studies of the mica polymorphs. *Mineralogical Magazine* 31: 209 234.

Molybdates

P A Williams, University of Western Sydney, Penrith South DC, NSW, Australia

Inroduction

Molybdates are compounds containing negatively charged oxymolybdenum ions. For the most part, mineralogically, the simple tetraoxomolybdate(VI), or molybdate ion, MoO_4^{2} , is present. Under acid conditions, molybdate ions polymerize and this process can incorporate other chemical entities. Resulting heteropolymolybdates are represented in the mineral kingdom, but are rare. Indeed, only a few minerals containing essential molybdate are known, and overwhelmingly, molybdate minerals contain the simple molybdate ion. In this connection, the chemistry of molybdenum displays some differences compared to that of its congeners, chromium and tungsten. The former does not polymerize to any significant extent in its highest oxidation state, aside from the formation of dichromate(VI), $Cr_2O_7^{2}$, and tungsten and molybdenum minerals in the primary zones of deposits are markedly differentiated by virtue of redox differences. Standard potentials at 25°C are –0.78 and –1.26 V for eqns [1] and [2], respectively. This accounts for the fact that tungstates are favoured even in the primary environment; tungstenite, WS_2, is an extremely rare species.

$$MoO_4^{2} + 2H_2O + 2e \rightarrow MoO_2 + 4OH \quad [1]$$

$$WO_4^{2} + 2H_2O + 2e \rightarrow WO_2 + 4OH \quad [2]$$

Primary Molybdates

The primary mineralogy of molybdenum is dominated by molybdenite, MoS_2. Aside from the substitution of molybdate in primary scheelite, $CaWO_4$, molybdates are generally confined to the supergene environment. A remarkable exception to this pattern of paragenesis concerns the rare occurrence of primary powellite, $CaMoO_4$, in centimetre-sized crystals from basalt-hosted deposits of the Deccan Trap in India. In primary or other reducing environments, lower molybdenum valencies are stabilized in various complex oxides. Some molybdenum oxides are termed 'fumarolic condensates', and the more complex mineral, kamiokite, $Fe_2Mo_3O_8$, found in such settings contains Mo(IV). Sedovite is thought to be a U(IV) molybdate, and mourite, $UMo_5O_{12}(OH)_{10}$, may also contain U(IV). Alternatively, it may be a mixed-valency Mo(V,VI) molybdate or polymolybdate of the uranyl ion, but further studies are required to characterize this species fully.

Secondary Molybdates

Apart from the preceding examples, molybdates are mostly confined to the oxidized zones of base metal orebodies. Here molybdate is derived from the oxidation of molybdenite. In such environments, a number of simple salts are found, together with rarer double salts and polymolybdates, as listed in Table 1. For simple salts such as powellite, substitution by tungstate is common and solid solution over the full range of compositions to scheelite is known. The end-member composition is shown in Table 1, as is the case with other examples (aspects of solid solution are dealt with later, where they are significant). Where solid solution is more complex, stoichiometries reflect compositions of known materials.

In similar fashion to the case with the powellite–scheelite pair, complete solid solution between wulfenite, $PbMoO_4$, and stolzite, $PbWO_4$, can occur, although most naturally occurring material contains minor tungstate. Other ions substitute in the lattice as well, including chromate, arsenate, and vanadate, the latter involving some other charge compensation mechanism. Molybdofornacite, $CuPb_2[(Mo,Cr)O_4][(As,P)O_4]OH$, is structurally related to the brackebuschite group of monoclinic arsenates, phosphates, sulphates, and vanadates of general formula $AB_2(XO_4)_2(OH,H_2O)$; A = Al, Cu, Fe, Mn, Zn; B = Ba, Ca, Pb, Sr; X = As, P, S, V. Chromate substitution is pronounced in this species. However, not all of the species in Table 1, even those of comparatively simple stoichiometry, contain isolated molybdate ions. Koechlinite is an example of a layered-structure compound; this type of chemical and structural feature is more pronounced in related minerals of W(VI). For example, there is no Mo(VI) analogue of the dimorph of stolzite, raspite, that contains chains of edge-linked WO_6 octahedra (*see* **Minerals:** Tungstates). The heteropolymolybdates betpakdalite, sodium betpakdalite, and mendozavilite are all rare minerals, although betpakdalite is commonly found in small amounts in the oxidized zones of molybdenite deposits. It is possible that the very rare minerals

Table 1 Molybdate(VI) minerals

Mineral	*Chemical composition*
Simple molybdates	
Powellite	$CaMoO_4$
Wulfenite	$PbMoO_4$
Umohoite	$UO_2MoO_4\ 4H_2O$
Sedovite	$U(MoO_4)_2$
Ferrimolybdite	$Fe_2(MoO_4)_3\ 8H_2O$
Basic double salts	
Koechlinite	Bi_2MoO_6
Lindgrenite	$Cu_3(MoO_4)_2(OH)_2$
Szenicsite	$Cu_3MoO_4(OH)_4$
Parkinsonite	$Pb_6O_4(MoO_4)Cl_2$
Molybdofornacite	$CuPb_2[(Mo,Cr)O_4][(As,P)O_4]OH$
Complex uranium salts	
Deloryite	$Cu_4UO_2(MoO_4)_2(OH)_6$
Calcurmolite	$Ca(UO_2)_3(MoO_4)_3(OH)_2\ 11H_2O$
Tengchongite	$Ca(UO_2)_6(MoO_4)_2(OH)_{10}\ 7H_2O$
Moluranite	$H_4U(UO_2)_3(MoO_4)_7\ 18H_2O(?)$
Cousinite	$Mg(UO_2)_2(MoO_4)_2(OH)_2\ 5H_2O(?)$
Mourite	$UMo_5O_{12}(OH)_{10}(?)$
Polymolybdates	
Iriginite	$UO_2(Mo_2O_7)\ 3H_2O$
Betpakdalite	$MgCa_2[Mo_8As_2Fe_3O_{36}(OH)](H_2O)_{23}$
Sodium betpakdalite	$MgCa_2[Mo_8As_2Fe_3O_{36}(OH)](H_2O)_{23}$
Mendozavilite	$Na(Ca,Mg)_2Fe_6(PO_4)_2(PMo_{11}O_{39})(OH,Cl)_{10}\ 33H_2O$
Other complex species	
Obradovicite	$H_4(K,Na)CuFe_2AsO_4(MoO_4)_5\ 12H_2O$
Melkovite	$CaFeH_6PO_4(MoO_4)_4\ 6H_2O$
Chiluite	$Bi_6(TeO_4)_2(MoO_4)_2O_5$

obradovicite, melkovite, and chiluite also belong to this class of heteropolymolybdate.

Molybdate minerals have little commercial significance, although wulfenite assumed an important role in this regard between the two World Wars. The Mammoth–St Anthony mine at Tiger, Arizona, USA, was the principal producer of molybdenum during this period, all from oxidized ores carrying wulfenite. Today, however, the mineral is of interest only to collectors. Nevertheless, it is a common mineral in many deposits, the most noteworthy of which are found in the south-western states of the United States and in Mexico.

See Also

Minerals: Definition and Classification; Chromates; Tungstates.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2003) *Handbook of Mineralogy. Volume 5. Borates, Carbonates, Sulfates, Chromates, Germanates, Iodates, Molybdates, Tungstates, etc., and Organic Materials.* Tucson, AZ: Mineral Data Publishing.

Anthony JA, Williams SA, Bideaux RA, and Grant RW (1995) *Mineralogy of Arizona,* 3rd edn. Tucson, AZ: University of Arizona Press.

Baes CF Jr and Mesmer RE (1986) *The Hydrolysis of Cations.* Malabar, FL: Krieger Publishing Company.

Bard AJ, Parsons R, and Jordan J (1985) *Standard Poten tials in Aqueous Solution.* New York: Marcel Dekker.

Bideaux RA (1980) Famous mineral localities: Tiger, Arizona. In: *The Mineralogical Record,* vol. 11, pp. 155 180. Tucson, AZ: Mineralogical Record Inc.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana,* 8th edn. London: Wiley Europe.

Mandarino JA (1999) *Fleischer's Glossary of Mineral Species 1999,* 8th edn. Tucson, AZ: Mineralogical Record Inc.

Ottens B (2003) Minerals of the Deccan Traps, India. In: *The Mineralogical Record,* vol. 34, pp. 1 82. Tucson, AZ: Mineralogical Record Inc.

Roberts WL, Campbell TJ, and Rapp GR Jr (1990) *Encyc lopedia of Minerals,* 2nd edn. New York: Van Nostrand Reinhold.

Williams PA (1990) *Oxide Zone Geochemistry.* Chichester: Ellis Horwood.

Native Elements

P A Williams, University of Western Sydney, Parramata, Australia

Introduction

A surprising number of the elements of the Periodic Table occur naturally, in the so-called native state. These range from the familiar coinage metals – copper, silver, and gold – to quite reactive metals, such as iron, zinc, and perhaps aluminium. Non-metallic elements such as sulphur and carbon are also well represented in the mineral kingdom (Table 1). The amounts of native elements present in rocks and ores vary widely, from minute grains to bodies weighing millions of tonnes. A number of these occurrences play extremely important economic roles, yet others are no more than mineralogical curiosities. The elements and their associations in widely disparate environments are reviewed, but ordered intermetallic species are not included.

Gases and Liquids

The majority of the atmosphere is made up of two gaseous elements, nitrogen and oxygen. Traces of hydrogen are known to be present. All of the noble gases – helium, neon, argon, krypton, xenon, and radon – are present in the atmosphere, the latter in small amounts, particularly in soil gas, where it is produced as the result of the radioactive decay of certain elements. No doubt other elements, including astatine, not considered in detail below, exist transiently in the free state as members of various radioactive decay series, but this aspect of the existence of the elements is rather arcane and is beyond the scope of this article. The third most abundant gas in the atmosphere is argon. Its apparently abnormal distribution is due to ^{40}Ar being a stable decay product of ^{40}K. While these gaseous elements are ubiquitous on Earth, albeit in highly variable amounts, they are not considered to be minerals. Another element present in volcanic gases is chlorine.

Under ambient conditions, only one element is found on Earth as a liquid. Native mercury is widespread, usually in small amounts, in low-temperature hydrothermal deposits and in hot springs. One exceptional occurrence is in the celebrated mercury deposits of Almadén, Spain, where it occurs in sufficient amounts to make it an economic mineral.

Solid Metals

The frequent occurrence of the coinage metals – copper, silver, and gold – gave rise to their use from early times. Native gold (*see* **Gold**) is overwhelmingly the commonest mineral form of the element and

Table 1 The naturally occurring native elements

Gases	
Argon	Ar
Astatine	At
Chlorine	Cl_2
Helium	He
Hydrogen	H_2
Iodine	I_2
Krypton	Kr
Neon	Ne
Nitrogen	N_2
Oxygen	O_2
Radon	Rn
Xenon	Xe
Liquids	
Mercury	Hg
Solids	
Allargentum	$Ag_{1-x}Sb_x$ (x 0.09 0.16)
Aluminium	Al
Antimony	Sb
Arsenic	As
Cadmium	Cd
Copper	Cu
Chromium	Cr
Diamond	C
Gold	Au
Graphite	C
Indium	In
Iodine	I_2
Iridium	Ir
Iron	Fe
Kamacite	(Ni, Fe)
Lead	Pb
Osmium	Os
Palladium	Pd
Phosphorus	P_n
Platinum	Pt
Rhodium	Rh
Rosickýite	S_8; γ sulphur
Ruthenium	Ru
Selenium	Se
Silver	Ag
Stibarsen	SbAs
Sulphur	S_8; α sulphur
Taenite	(Ni, Fe)
Tellurium	Te
Tin	Sn
Unnamed	(Ni, Fe); unnamed tetragonal phase
Unnamed	S_8; β sulphur
Unnamed	S_n; amorphous sulphur
Zinc	Zn

remains its economically most important ore mineral. It is frequently alloyed with silver (electrum) or mercury (an amalgam) and usually contains small amounts of other alloyed elements. While silver and gold form a continuous solid solution, the mutual solubilities of copper and silver are quite restricted. Gold of very high fineness is known from supergene deposits. Silver is a comparatively common native element, forming a range of solid solutions with other elements. It also is known to occur as ordered alloys and intermetallic compounds, which are not considered here. Allargentum is a hexagonal mineral that may be considered to be the hexagonal form of silver, at least in part. Native silver is present in both primary and secondary ores, but it is now chiefly produced by the refining of other metals or concentrates, especially those of copper, lead, and zinc. Native copper is a very common mineral in the oxidized zone of base metal ore bodies. It is also found in the form of masses amounting to many tonnes in weight in the famous Keweenaw Peninsula of northern Michigan, USA. Native copper remains a significant ore mineral, especially in the oxidized zones of base metal deposits.

The other precious metals – iridium, osmium, palladium, platinum, rhodium, and ruthenium – are well known minerals in primary and placer deposits. They also form alloys with each other and with base and noble metals. They remain economically important species in layered deposits of platinum-group elements such as those in the Bushveldt (South Africa), Stillwater (Montana, USA), and related igneous complexes. Platinum was originally extracted from extensive placer and alluvial deposits in South America, the Urals, and elsewhere, but these are not now important sources of the element.

Many other metallic elements are known to exist in the natural environment in amounts varying from isolated grains to large masses. Native iron is a frequent accessory mineral of basic rocks, and masses weighing many tonnes have been recovered from weathering basalts on Disko Island, Greenland. It is also found in meteoritic materials, alloyed with nickel and small amounts of cobalt and other elements. Native nickel is a minor component of nickel-rich ultramafic rocks. Taenite (Ni, Fe) and kamacite (Ni Fe) are ordered and disordered cubic minerals that have been recovered from Ni–Fe meteorites, and a third tetragonal phase is known. Elemental chromium has been reported from unspecified localities in Russia and China. A number of high- and low-temperature hydrothermal deposits are known to host native antimony, arsenic, and bismuth. The last is a significant ore mineral in some deposits. The intermetallic compound stibarsen (SbAs) is known to occur naturally, and both native selenium and tellurium, included here with the metals, are frequent accessories in certain epithermal deposits. Native indium has been reported to occur in several unspecified Russian deposits.

Very rarely, metals whose natural existence would be unexpected because of their redox potentials have been found in small amounts. Aluminium, zinc, and cadmium have been found in highly reduced skarn and intrusive igneous deposits. Native zinc is the most common of these and has been found in other settings. Elemental tin has been found in a number of igneous deposits. Grains of tin have also been recovered from placer deposits in northern New South Wales, Australia. It is thought that these may be the result of the reduction of cassiterite, SnO_2, by carbon during bush fires. Some doubt has attended many reports of native lead because a number of cases may be due to the recovery of shot or other human artefacts. However, it has been reliably reported from a dozen or so deposits, and fine crystal groups have been found at Långbarn and the Harstigen mine, Sweden.

Solid Non-Metals

Only four solid non-metals, other than perhaps selenium, occur naturally as elements. These are carbon, sulphur, phosphorus, and iodine. Two allotropes of carbon – diamond and graphite – are recognized as minerals. No doubt the buckminsterfullerines and related compounds occur in nature as well, but they have not yet been formally described. Deposits of millions of tonnes of α-sulphur (the mineral sulphur) occur in so-called 'salt dome' deposits, and these are the most important commercial source of the element, which is recovered by the Frasch process. Sulphur is found in other sedimentary rocks, in the oxidized zones of many sulphide ore bodies, as a product of various micro-organisms, and very frequently as a sublimate associated with volcanic fumaroles, where it is sometimes associated with selenium. Two further modifications of sulphur have been reported to occur as minerals. Rosickýite is the natural counterpart of the well-known allotrope γ-sulphur. It occurs as a volcanic sublimate at Vulcano in the Mediterranean Sea and in sediments in Czechoslovakia and the USA. Naturally occurring β-sulphur has been reported from volcanic fumaroles on Vulcano and Mount Vesuvius, Italy. Amorphous or 'plastic' sulphur occurs in the Kobui sulphur mine, Japan, but it crystallizes over time to α-sulphur. Native phosphorus has been found in the Saline Township meteorite from Kansas, USA. Finally, the occurrence of elemental iodine is worthy of note. Iodine-rich caliche from certain

deposits of the Atacama Desert, Chile, is coloured blue-purple by elemental iodine. In addition, specimens of marshite (CuI) frequently possess the characteristic odour of iodine. This is perhaps not surprising given the mode of formation of marshite by the interaction of aqueous solutions carrying Cu^{2+} and I^- ions:

$$Cu^{2+}(aq) + 2I^-(aq) \rightarrow CuI(s) + 0.5I_2(s)$$

See Also

Gold. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Solar System:** Meteorites. **Tectonics:** Hydrothermal Activity.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (1990) *Handbook of Mineralogy. Volume 1. Elements, Sulfides, Sulfosalts*. Tucson: Mineral Data Publishing.

Bard AJ, Parsons R, and Jordan J (1985) *Standard Potentials in Aqueous Solution*. New York: Marcel Dekker.

Boyle RW (1968) *The Geochemistry of Silver and its Deposits*. Quebec: Geological Survey of Canada.

Boyle RW (1987) *Gold: History and Genesis of Deposits*. New York: Van Nostrand Reinhold.

Cabri LJ (ed.) (1981) *Platinum Group Elements Mineralogy, Geology, Recovery*. Quebec: Canadian Institute of Mining and Metallurgy.

Cox PA (1995) *The Elements on Earth*. New York: Oxford University Press.

Emsley J (1991) *The Elements*, 2nd edn. Oxford: Clarendon Press.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, 8th edn. London: Wiley Europe.

Gasparrini C (1993) *Gold and Other Precious Metals*. Berlin: Springer Verlag.

Guilbert JM and Park CF Jr (1986) *The Geology of Ore Deposits*. New York: W H Freeman.

Mandarino JA (1999) *Fleischer's Glossary of Mineral Species 1999*, 8th edn. Tucson: Mineralogical Record Inc.

Williams PA (1990) *Oxide Zone Geochemistry*. Chichester: Ellis Horwood.

Nitrates

P A Williams, University of Western Sydney, Penrith South DC, NSW, Australia

Introduction

Naturally occurring nitrate mineral deposits are found in a number of different environments, but most are found in arid climates or in areas in which evaporation rates are high. This is because nitrate salts are generally highly soluble in water; exceptions are the metastable nitrate-bearing members of the hydrotalcite groups and certain basic copper salts.

Geographical Distribution of Nitrate Minerals

Despite the high degree of solubility of most nitrates in water, aside from basic salts, they are quite widely distributed in the natural environment, given appropriate conditions for their crystallization. Simple nitrates (Table 1), including gwihabaite, nitratine, niter, nitrobarite, nitrocalcite, and nitromagnesite, are found in quantities ranging up to thousands of tonnes in dry caves, where rates of evaporation are significant. The nitrate in this case (together with the ammonium ion in the case of gwihabaite) is usually derived from the decomposition of bat guano or is derived from the leaching of nitrate-bearing volcanic or sedimentary rocks. Some of the simple anhydrous species are also formed as ephemeral efflorescences on rock faces and in playa lakes in arid climates. Aside from this, both simple and complex nitrates (again, excluding the basic salts) are present in quantities amounting to millions of tonnes in the celebrated Chilean nitrate deposits of the Atacama Desert. These represent a significant source of nitrate fertilizer and have been mined as such for centuries. The rare mineral ungemachite was first found in during mining operations in the oxidized zone of the Chuquicamata copper deposit together with clinoungemachite; the mine is located in the Atacama region and ungemachite has been identified elsewhere only in a single specimen from the oxidized zone of the New Cobar copper–gold deposit near Cobar, New South Wales, Australia. The caliche-hosted nitrate deposits of the Atacama Desert contain other highly oxidized species (besides nitrate, with nitrogen in its highest oxidation state), including bromate, iodate, periodate, perchlorate, and native iodine.

Three nitrate-bearing members of the hydrotalcite group containing nickel are confined to a single locality, the Mbobo Mkulu cave in South Africa. Nitrate is formed from the leaching of nitrate-bearing volcanic or sedimentary rocks, and nickel and copper are

Table 1 Nitrate(v) minerals

Type	*Name*	*Composition*
Simple nitrates	Gwihabaite	$(NH_4,K)NO_3$
	Nitratine	$NaNO_3$
	Niter	KNO_3
	Nitrobarite	$Ba(NO_3)_2$
	Nitrocalcite	$Ca(NO_3)_2\ 4H_2O$
	Nitromagnesite	$Mg(NO_3)_2\ 6H_2O$
Hydrotalcite related phases	Hydrombobomkulite	$(Ni,Cu)Al_4[(NO_3)_2,(SO_4)](OH)_{12}\ 12\ 14H_2O$
	Mbobomkulite	$(Ni,Cu)Al_4[(NO_3)_2,(SO_4)](OH)_{12}\ 3H_2O$
	Nickelalumite	$(Ni,Cu)Al_4[(SO_4),(NO_3)_2](OH)_{12}\ 3H_2O$
	Sveite	$KAl_7(NO_3)_4Cl_2(OH)_{16}\ 8H_2O$
Compound nitrate sulphates	Darapskite	$Na_3(NO_3)(SO_4)\ H_2O$
	Humberstonite	$K_3Na_7Mg_2(NO_3)_2(SO_4)_6\ 6H_2O$
	Ungemachite	$K_3Na_8Fe(NO_3)_2(SO_4)_6\ 6H_2O$
Basic copper nitrates	Gerhardtite	$Cu_2NO_3(OH)_3$
	Rouaite	$Cu_2NO_3(OH)_3$
	Likasite	$Cu_3NO_3(OH)_5\ 2H_2O$
	Buttgenbachite	ca $Cu_{36}(NO_3)_2Cl_8(OH)_{62}\ 4\ 10H_2O$

derived from oxidizing sulphide minerals in associated shale horizons. Sveite is also a cave mineral; it has been found in the Autana cave, Venezuela, and has been reported as a coating on rock faces at an unspecified California locality. All of these minerals are exceptionally rare and are based on the hydrotalcite structure, $Mg_6Al_2CO_3(OH)_{16} \cdot 4H_2O$, consisting of an $M(OH)_2$ layer that sandwiches layers containing anions, metal ions, and water molecules. The minerals are metastable phases with respect to Al(III) oxyhydroxides and other soluble components.

The basic copper(II) nitrates of Table 1 are an exception to the pattern of nitrate mineral distribution in arid regions in that they are found sparingly in the oxidized zones of a number of copper orebodies. Perhaps the most common of the minerals is buttgenbachite. The composition of this mineral is complicated by the fact that it is involved in a complex solid-solution series involving the sulphate-chloride analogue, connellite. Various amounts of chloride and hydroxide ions are found in the lattice, depending on the composition of the solution from which the mineral crystallizes and with respect to requirements for charge compensation. Connellite, the sulphate-dominant analogue, almost invariably carries some nitrate, although the nitrate-free end-member can be synthesized.

The origin and persistence of nitrate in the natural environment are both complex phenomena. Biological activity serves both to produce and to consume nitrate, depending on the micro-organisms involved. The accumulation of vast amounts of nitrate in the Chilean deposits is due to the extreme paucity of living matter that would ordinarily consume it. Large amounts of nitrate minerals can accumulate only in the absence of significant biological activity or in a steady-state environment when losses due to metabolic processes are balanced by biological nitrate production or inputs from other sources. In this connection, aside from contributions from leaching of rocks, nitrate is formed in significant amounts by electrical discharge in the atmosphere during thunderstorm activity.

Despite the large quantities of some of the world's nitrate deposits, including those found in caves, nitrates generally are rare minerals and are confined to special environments. Aside from the few simple minerals that dominate the exploitable resources, nitrate minerals are of interest only in the academic sense or to collectors.

See Also

Minerals: Definition and Classification; Chromates. **Sedimentary Rocks:** Evaporites.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2003) *Handbook of Mineralogy. Volume 5. Borates, Carbonates, Sulfates, Chromates, Germanates, Iodates, Molybdates, Tungstates, etc., and Organic Materials*. Tucson, AZ: Mineral Data Publishing.

Blackburn TH (1983) The microbial nitrogen cycle. In: Krumbein WE (ed.) *Microbial Geochemistry*, pp. 63 90. Oxford: Blackwell Scientific Publications.

Böhlke JK, Ericksen GE, and Revesz K (1997) Stable iso tope evidence for an atmospheric origin of desert nitrate deposits in northern Chile and southern California, USA. *Chemical Geology* 136: 135 152.

Doner HE and Lynn WC (1989) Carbonate, halide, sulfate, and sulfide minerals. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, 2nd edn. pp. 331 378. Madison, WI: Soil Science Society of America.

Ericksen GE (1981) *Origin of the Chilean Nitrate Deposits. United States Geological Survey Professional Paper, 1118*. Washington, DC: US Geological Survey.

Fenchel T, King GM, and Blackburn TH (1998) *Bacterial Biogeochemistry The Ecophysiology of Mineral Cycling*, 2nd edn. New York: Academic Press.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, 8th edn. London: Wiley Europe.

Hill C and Forti P (eds.) (1997) *Cave Minerals of the World*, 2nd edn. Huntsville, AL: National Speleological Society.

Mandarino JA (1999) *Fleischer's Glossary of Min eral Species 1999*, 8th edn. Tucson, AZ: Mineralogical Record Inc.

Roberts WL, Campbell TJ, and Rapp GR, Jr (1990) *Encyc lopedia of Minerals*, 2nd edn. New York: Van Nostrand Reinhold.

Williams PA (1990) *Oxide Zone Geochemistry*. Chichester: Ellis Horwood.

Olivines

G Cressey, The Natural History Museum, London, UK
R A Howie, Royal Holloway, University of London, London, UK

Introduction

Olivine is an important rock-forming mineral and occurs in mafic igneous rocks such as peridotite, gabbro, and basalt, along with pyroxene and plagioclase. The ultramafic rock dunite consists almost entirely of olivine, and is the result of the fractional crystallisation of peridotite-composition melts to form cumulate olivine. Magnesium-rich olivine is abundant in the majority of mantle-derived ultramafic nodules in kimberlite pipes. Olivine is common in the Earth, Moon, (*see* **Solar System:** Moon) and in the stony meteorites, and much research is conducted into the physical, chemical, and mechanical properties of olivine in order to understand the physical state of the interior of the Earth, the origin of meteorites, and the history and processes on planetary bodies (*see* **Solar System:** Mars).

Crystal Chemistry of the Olivine Group

Crystals of olivine possess orthorhombic symmetry and have a structure consisting of an array of individual (SiO_4) tetrahedral units that do not share linking oxygen atoms (unlike the pyroxenes, amphiboles, or feldspars which form linked SiO_4 polymers). The oxygen atoms that form the individual tetrahedral units of the olivine structure are arranged in planes perpendicular to the x-axis; in each plane the oxygen atoms are in approximate hexagonal array, but they are not touching. The tetrahedra point alternately up and down in the x and y directions of the unit cell, and there are two non-equivalent sets of cation sites linking the tetrahedra. Half of the cation sites lie on a centre of symmetry between the tetrahedral bases (M1 sites), and half lie on mirror planes between edges and bases of tetrahedra (M2 sites). Although in each of these sites the cations are in six-coordination with oxygen atoms of neighbouring SiO_4 tetrahedra, the cation sites are geometrically distinct (**Figure 1**). At elevated temperatures, the cation–oxygen bond lengths of both M1 and M2 sites increase, but the SiO_4 tetrahedra remain the same, thus reflecting the greater bond strength of Si–O compared with that between other cations and oxygen. In most natural olivines, Mg^{2+} or Fe^{2+} occupy the cation sites. Because of their similar ionic size, Mg^{2+} (ionic radius 0.72 Å in 6-coordination with oxygen) or Fe^{2+} (ionic radius 0.78 Å) can be accommodated in either cation site, and there exists a complete solid solution series between these two end-members having compositions Mg_2SiO_4 (forsterite) and Fe_2SiO_4 (fayalite). Almost all natural olivines have compositions intermediate between these two extremes, the exact composition reflecting the availability of Mg and Fe during crystallisation. Mg^{2+} and Fe^{2+} can occupy M1 and M2 with almost equal preference, so Mg^{2+} and Fe^{2+} remain disordered over these two sites down to low temperatures, and the solid solution is close to ideal with no tendency for ordering or exsolution to occur. However, it has been established that there is a slight tendency for Fe^{2+} to occupy the M1 site rather than the M2 site (this is paradoxical since M1 is the smaller site and Fe^{2+} the larger cation), therefore the Mg–Fe distribution is not completely random, but this effect is only slight.

Larnite, Ca_2SiO_4, does not have an olivine-type structure, but Ca^{2+} ions can be accommodated in the olivine structure on half the cation sites, i.e.,

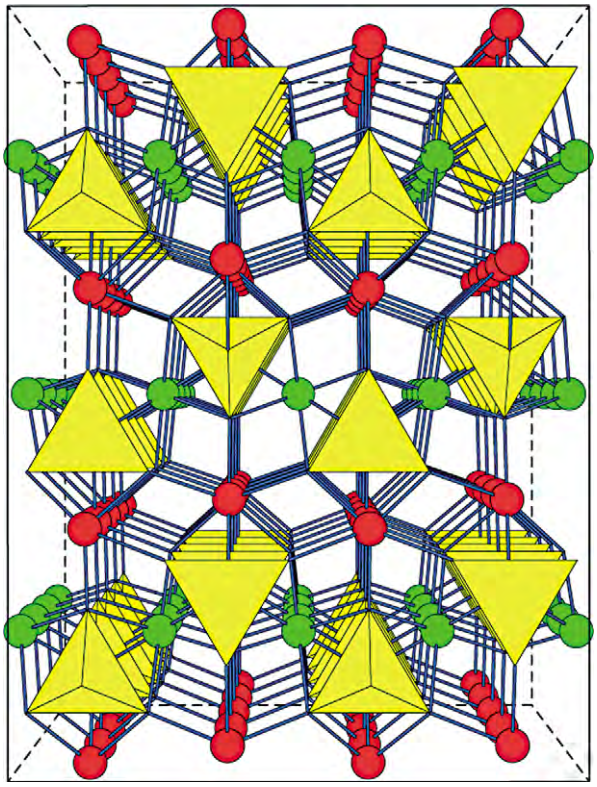


Figure 1 The olivine structure (view looking down the *x* axis of the unit cell): SiO_4 tetrahedral units (yellow) point alternately either way in both the *x* and *y* directions, and there are two different cation sites, M1 (green) and M2 (red).

$CaMgSiO_4$, forming the mineral monticellite. The Fe analogue, $CaFeSiO_4$ (kirschsteinite) is known only in synthetic systems, but complete solid solution is possible between montcellite and kirschsteinite as Fe replaces Mg in the structure. The most iron-rich naturally occurring examples in the $CaMgSiO_4$–$CaFeSiO_4$ series contain about 70% $CaFeSiO_4$ component. However, there is very limited solid solution between the montcellite-kirschsteinite series and the forsterite-fayalite series, because the size of the Ca^{2+} ion (ionic radius 1.00 Å) is substantially larger than Mg^{2+} and Fe^{2+} ions. Substitution of the large Ca^{2+} ion into the olivine structure causes strain that distorts the SiO_4 tetrahedra.

The olivine structure cannot tolerate the degree of distortion that would be produced by a random occupation of the cation sites by ions of such different sizes, so very little solid solution can exist between the Ca-olivines and the Mg, Fe-olivines, even at high temperatures. In the Ca-bearing olivines, the Ca and Mg, Fe ions are completely ordered in the structure; Ca^{2+} occupies the (larger) M2 site, while Mg^{2+} and Fe^{2+} are randomly distributed on the M1 sites. In many natural Mg-Fe olivines, particularly those rich in Fe, there is often a small proportion of Ca and Mn present in the structure. The substitution of Mn^{2+} (ionic radius 0.83 in 6-coordination) for Fe^{2+} in fayalite also occurs when Mn is present during crystallization, and a complete solid solution series exists between Fe_2SiO_4 (fayalite) and Mn_2SiO_4 (tephroite). In natural occurrences, the low Mg-content of fayalite-tephroite phases is more likely to be related to the limited amounts of Mg available in such crystallization environments, rather than the inability of the structure to tolerate differences between these cations sizes. Zn-bearing tephroites have been reported, and $CaMnSiO_4$ (glaucochroite), although rare in nature, has an olivine structure.

Other compositions that also possess the olivine structure, such as Ni_2SiO_4 and Mg_2GeO_4, have been prepared by laboratory synthesis. Small amounts of Ni^{2+} (ionic radius 0.69 Å in 6-coordination) are commonly present in the structure of natural Mg-rich olivines. However, the presence of Cr^{3+} and/or Fe^{3+} (which would produce unfavourable charge imbalances in the olivine structure) is often found to be associated with exsolved sub-microscopic crystallites of chromite or magnetite inside the olivine. Fe^{3+} is also likely to be present in the oxidation products commonly formed during the hydrothermal alteration of olivine.

Nomenclature

Compositions within the forsterite-fayalite isomorphous solid solution series are the most abundant of the naturally occurring members of the olivine group, and the term olivine has come to signify compositions between these two end members, with general formula $(Mg,Fe)_2SiO_4$. In the past, different names have been assigned to specific ranges of Mg:Fe ratio (e.g., chrysolite, hyalosiderite, hortonolite, ferrohortonolite), but it is preferable to indicate the composition by giving the mole proportion of either the forsterite (Fo) or the fayalite (Fa) component in the solid solution. This can be expressed in different ways; for example, an olivine with specific composition $Mg_{1.8}Fe_{0.2}$ SiO_4 can be expressed as Fo_{90} (or Fa_{10}), meaning that the olivine contains 90% of forsterite component (and 10% of fayalite component). Alternatively, the cation site occupancy in this olivine structure can be expressed as $X_{Mg} = 0.90$, $X_{Fe} = 0.10$, where $X_{Mg} = Mg/(Mg + Fe)$, $X_{Fe} = Fe/(Mg + Fe)$, and $X_{Fe} + X_{Mg} = 1$. This nomenclature can be extended to any number

of components, if small amounts of Mn, Ca, Ni, Zn, etc. are present.

Physical Properties and Stability

The densities of olivine solid solutions vary smoothly and linearly from forsterite (3.22 gm cm ^{3}) to fayalite (4.39 gm cm ^{3}). Similarly, their unit cell dimensions and optical properties vary systematically with composition. At atmospheric pressure the melting point of pure forsterite occurs at 1890°C; under anhydrous conditions the melting point increases with pressure, and under water-saturated conditions the melting point decreases with pressure. Pure fayalite melts at a much lower temperature (1205°C at 1 atm) than forsterite, because the Fe^{2+}-oxygen bonds are weaker than the shorter Mg^{2+}-oxygen bonds. For cations of the same charge, bonds to oxygen are weaker for the larger cation, so as more of the larger cations enter the structure there is a progressive reduction in the melting points of intermediate compositions. This is reflected in the crystallisation behaviour of olivine solid solutions; the first olivine crystals to separate from a liquid of given composition are richer in Mg than those of later in the crystallisation sequence, and consequently the larger Fe^{2+} ions are concentrated in the residual liquid.

Igneous olivines formed during a slow, undisturbed cooling regime will have sufficient opportunity to re-equilibrate their Mg-Fe composition with the coexisting liquid as the temperature gradually falls, so that the final crystals have a single composition. However, rapid cooling can often result in zoned olivine crystals that have a Mg-rich core and progressively more Fe towards their rims; the core represents the initial composition to form which was subsequently prevented from equilibrating with the remaining liquid by a protective overgrowth. In very rapidly cooled magmas, olivines sometimes form dendritic crystals. The morphology of olivine crystals may be equant, tabular, acicular, or dendritic; these can, in general, be related to the rate of cooling of the magma during crystallisation. In a natural silicate melt, the melting point of Mg-rich olivine is much depressed, such that crystals of composition Fo_{88} will melt in basalt magma at about 1250°C.

At high pressures in the deep earth, the olivine structure transforms to a denser spinel-type structure. For forsterite-rich compositions, the transformation to a distorted spinel phase occurs at about 140 kbar pressure at a temperature of 1000°C. This structural transition is of geophysical significance, as it is believed to be the cause of the abrupt change in seismic velocity in the Earth's mantle at about 400–500 km depth.

Alteration

Olivine is very susceptible to hydrothermal alteration, during low-grade metamorphism and weathering, and readily reacts to form serpentine and a range of iron oxides, chlorites, and smectite clays. Serpentinisation is the most common and widespread alteration process to affect olivine-rich rocks such as dunite and peridotite. The serpentine polymorphs chrysotile, lizardite, and antigorite, along with brucite, talc, and carbonates (if CO_2 is also present in the fluid) are commonly the products of this retrograde metamorphism. The following chemical re-arrangement represents the serpentinisation process:

$$\underset{\text{forsterite}}{2Mg_2SiO_4} + 3H_2O = \underset{\text{serpentine}}{Mg_3Si_2O_5(OH)_4} + \underset{\text{brucite}}{Mg(OH)_2}$$

This hydrothermal alteration of forsterite takes place at about 400°C, and for an olivine of composition Fo_{90} this reaction takes place nearer 340°C. Any iron present in the original olivine can only be accommodated in the serpentine structure to a very limited extent and is, therefore, usually converted to iron oxide or oxyhydroxide (magnetite, hematite, or goethite) inclusions that give the otherwise green serpentine a dark brownish or reddish appearance.

Occurrence

Pure forsterite is rare, but is found in certain metamorphosed siliceous dolomitic limestones (marbles). Olivine with compositions between Fo_{96} and Fo_{82} are major constituents in ultramafic rocks such as dunite and peridotite, and in prograde metamorphosed serpentintes. Mantle-derived ultramafic nodules in basalts and kimberlites usually contain olivines in the range Fo_{91} to Fo_{86}. Olivines with compositions Fo_{80} to Fo_{50} are common in gabbros and basalts (Figure 2); the actual forsterite-fayalite composition provides a useful index of fractionation for these basic igneous rocks. A wide range of olivine compositions from Fo_{98} to Fo_{40} (often as zoned crystals) and occasional almost pure fayalite grains are important constituents in some carbonaceous chondrite meteorites. Fe-rich olivines occur in alkaline and acid igneous rocks; fayalite is present in small amounts in many volcanic rocks such as

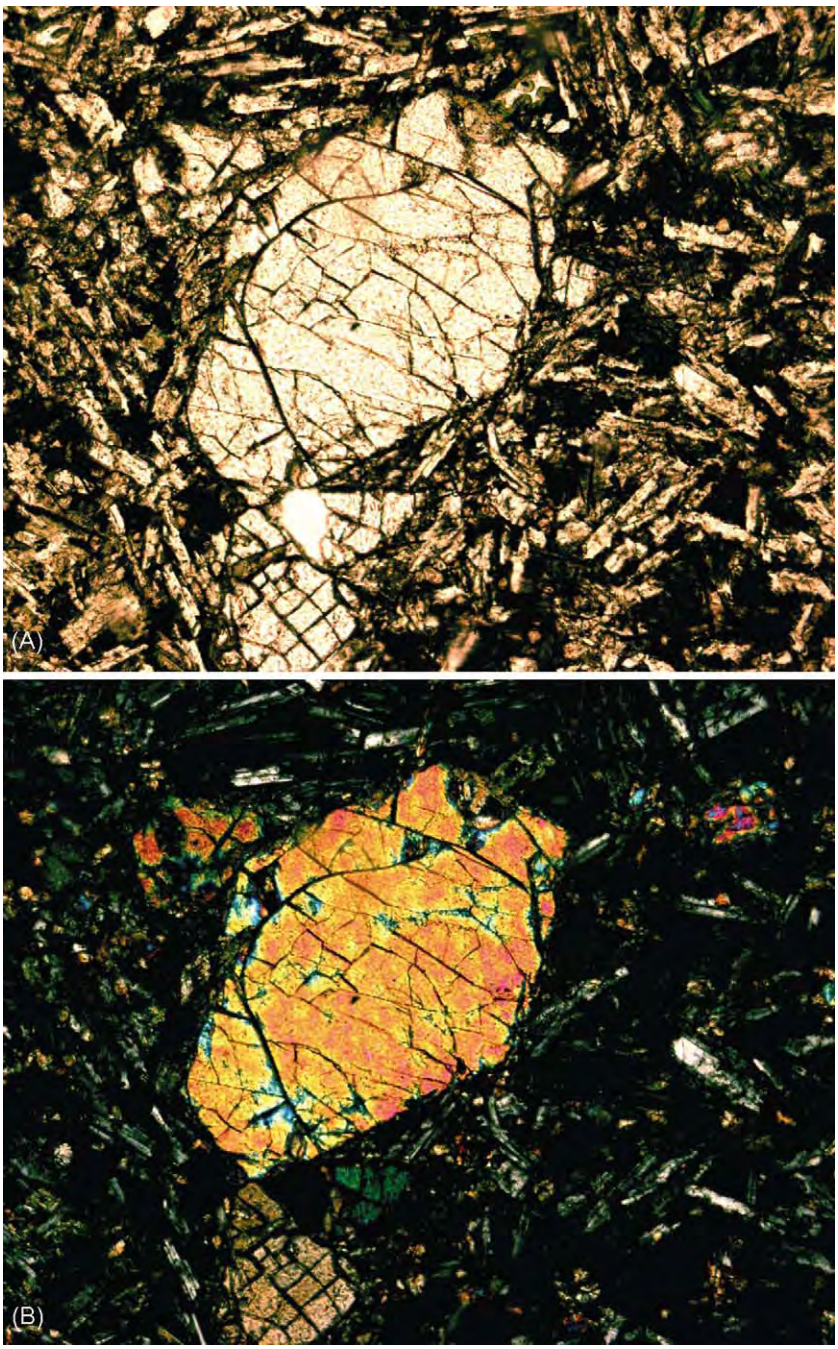


Figure 2 Basalt (from Melbourne, Victoria) seen in thin section through a polarizing microscope showing a euhedral phenocryst of olivine in plane polarized light (A), and with crossed polars (B).

obsidian, rhyolite, trachyte, and phonolite. Fayalitic and tephroitic olivines occur in rocks of metamorphic origin, such as thermally and regionally metamorphosed Fe-rich and Mn-rich sediments, respectively. Monticellite occurs in metamorphosed siliceous and magnesian limestones and in skarn zones developed by metasomatic alteration of limestones at intrusive igneous contacts. Monticellite also occurs in kimberlite and alnöite.

Meteoritic Olivine

Olivine is a major component of stony meteorites and some stony-irons (pallasites). In the former, it

occurs mainly in rounded chondrites. It is also found in some meteorites, including the almost monomineralic Chassigny stone and some basaltic shergottites. These are believed to come from Mars. Likewise, it is found in lunar-sourced meteorites and lunar surface breccias.

See Also

Earth: Mantle. **Gemstones**. **Igneous Rocks:** Komatiite. **Minerals:** Definition and Classification; Other Silicates. **Rocks and Their Classification**. **Solar System:** Moon; Mars.

Further Reading

Boyd FR and Nixon PH (1978) Ultramafic nodules from the Kimberly pipes, South Africa. *Geochemica et Cosmochemica Acta* 42: 1367 1382.

Deer WA, Howie RA, and Zussman J (1992) *An introduction to the Rock Forming Minerals*. London: Longman.

Deer WA, Howie RA, and Zussman J (1997) Rock Forming Minerals, vol. 1A, p. 219, Orthosilicates. London: The Geological Society.

Donaldson CH (1976) An experimental investigation of olivine morphology. *Contributions to Mineral Petrology* 57: 187 213.

Ghose S, Wan C, and McCallum IS (1976) Fe^{2+} Mg^{2+} order in an olivine from the lunar anorthosite 67075 and the significance of cation order in lunar and terrestrial olivines. *Industrial Journal of Earth Science* 3: 1 8.

Green DH (1964) The petrogenesis of the high temperature peridotite intrusion in the Lizard area, Cornwall. *Journal of Petrology* 5: 134 188.

MacKenzie WS, Donaldson CH, and Guilford C (1982) *Atlas of Igneous Rocks and Their Textures*. Longman.

MacKenzie WS and Guilford C (1980) *Atlas of Rock Forming Minerals in Thin Section*. London: Longman.

Wenk HR and Raymond KN (1973) Four new refinements of olivine. *Zeit. Krist.* 137: 86 105.

Other Silicates

R A Howie, Royal Holloway, University of London, London, UK

Introduction

Silicates not grouped in the major series separately dealt with, but which are of geological importance, include the garnet group, the aluminosilicates (the Al_2SiO_5 polymorphs), staurolite, the epidote group, cordierite, tourmaline, the chlorite group, and the serpentines.

The Garnet Group

The minerals (*see* **Minerals:** Definition and Classification, **Gemstones**) of the garnet group are particularly characteristic of metamorphic rocks but are also found in some igneous rocks, and as detrital grains in sediments. The group is subdivided into six principal species, which represent the end-members of an isomorphous series:

Pyrope	$Mg_3Al_2Si_3O_{12}$
Almandine	$Fe_3^{2+}Al_2Si_3O_{12}$
Spessartine	$Mn_3Al_2Si_3O_{12}$
Grossular	$Ca_3Al_2Si_3O_{12}$
Andradite	$Ca_3(Fe^{3+},Ti)_2Si_3O_{12}$
Uvarovite	$Ca_3Cr_2Si_3O_{12}$

A garnet corresponding in composition with any one end-member is rare, however, and the name is assigned according to the dominant 'molecular' type present. The garnets may be considered as two series: pyralspite (*py*rope, *al*mandine, *sp*essartine) and ugrandite (*u*varovite, *gr*ossular, *and*radite). Fairly complete and continuous variation in composition occurs within these two series, but there appears to be no continuous variation between them.

Garnets are cubic and are normally isotropic, but some calcic varieties show complex and sector twinning, which may be visible in birefringent varieties. Their colour ranges from red, orange, brown, and black for the pyralspites, to green, yellow, brown, and black for the ugrandites; they are colourless to pale pink, yellow, or brown in thin section. They are notable in having a strong power of crystallization, typically forming rhombic dodedcahedra or icositetrahedra.

The pyrope–almandine series occur in metamorphic rocks of a variety of types, of which the garnet-mica schists of regional metamorphism (*see* **Regional Metamorphism**) are typical; they are also found in certain ultrabasic rocks such as the mica peridotites and kimberlites. The pyralspites are also

known in peraluminous volcanic rocks, granites, and pegmatites. The distribution of Fe and Mg between coexisting equilibrium assemblages of garnet and biotite or garnet and clinopyroxene can be used in geothermometry. The equilibrium constant $K_{D\ Mg/Fe}^{Gt/Bi}$ for different metamorphic grades is shown in Figure 1. For the garnet–clinopyroxene pair, the pressure–temperature dependence of the Fe–Mg partition coefficient has been determined experimentally; the use of thermodynamic data for the garnet group is illustrated in Figure 2. Purplish chrome-pyrope is typical of diamondiferous kimberlites and peridotitic xenoliths.

Garnets with appreciable spessartine are usually found in granitic pegmatites and skarns.

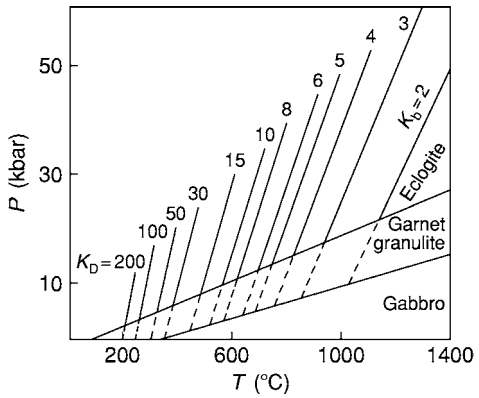


Figure 1 $K_{D\ Fe/Mg}^{Gt/Cpx}$ as a function of pressure and temperature (after Råheim and Green (1974)).

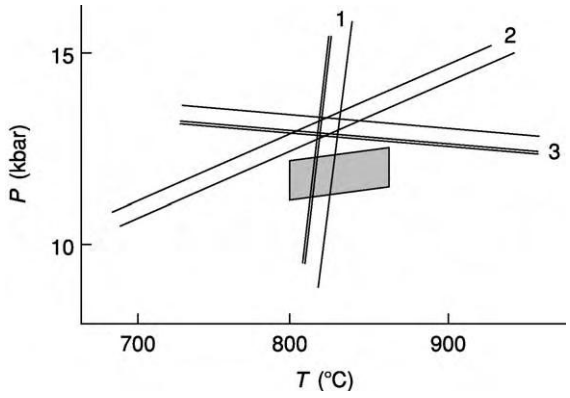


Figure 2 Thermodynamic data for natural equilibrium assem blages from pyroxene granulites of South Harris, Scotland, derived from: (1) Fe Mg distribution between garnet and clinopyroxene; (2) the activity coefficient for the $Ca_3Al_2Si_3O_{12}$ component in garnet using the equilibrium 3 $CaAl_2Si_2O_8$ (plagioclase) $Ca_3Al_2Si_3O_{12}$ + 2 Al_2SiO_5 + SiO_2; and (3) the activity coefficient for the equilibrium $CaAl_2Si_2O_8$ $CaAl_2SiO_6$ (clinopyroxene) + SiO_2. The shaded area is the apparent region of 'bracketing' of metamorphic pressure and temperature derived from other equilibria (after Wood 1975 and Wood (1997)).

Grossular is characteristically found in regionally or thermally metamorphosed impure calcareous rocks; andradite also occurs in such rocks but is more typically found in metasomatic skarn deposits.

The black garnet melanite is a Ti-bearing andradite occurring in alkaline igneous rocks and some skarns. Uvarovite is rare and any appreciable Cr content in metasomatic garnets can be detected from their green colour.

The aluminosilicates (Al_2SiO_5)

This family comprises three polymorphs, all typically found in metamorphosed pelitic rocks formed in different metamorphic conditions. Their occurrence is, thus, of particular interest in metamorphic petrology.

Sillimanite is orthorhombic, typically colourless or white, and found as optically length-slow prismatic crystals or needles.

Andalusite is also orthorhombic, typically pink in hand specimen, and found as stout, length-fast prismatic crystals. In low-grades of metamorphism the anhedral grains frequently encloses a host of inclusions of foreign matter, but rapidly acquire a prismatic outline pushing aside the impurities which then form a characteristic chiastolite pattern.

Kyanite is triclinic, typically giving inclined exctinction and having a blue colour in hand specimen.

The experimentally determined phase diagram for the Al_2SiO_5 composition is given in Figure 3, indicating the position of the triple point between sillimanite, andalusite, and kyanite.

The application of these results to metamorphic petrology, thus, clearly shows that sillimanite is the high-temperature phase, to be found in the highest

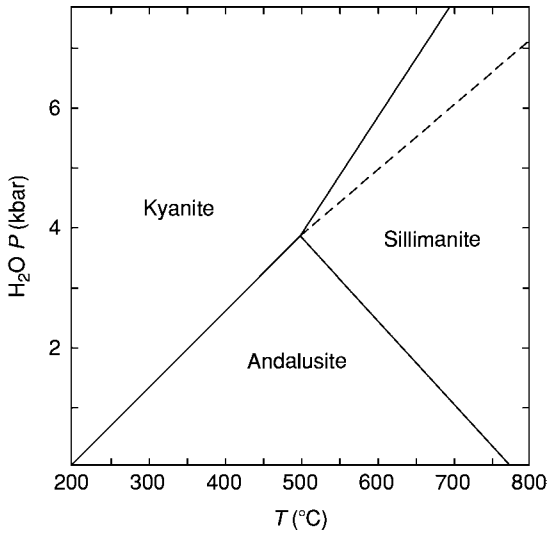


Figure 3 Phase diagram for kyanite andalusite and kyanite sillimanite equilibria.

grades of thermal metamorphism of pelitic rocks or in the highest grade of regional metamorphism of similar rocks. Andalusite is normally confined to thermally metamorphosed pelitic rocks, but may be found in some granites, probably as a result of contamination. Kyanite is typically a mineral of high-grade metamorphism of pelitic rocks, but has also been found in thermal aureoles in situations where appreciable shear-stress arose during emplacement of the igneous intrusion.

Staurolite

This yellow to brown ferromagnesian aluminous silicate characteristically occurs in medium grade metamorphosed pelitic schists and gives its name to a regional metamorphic zone. The staurolite–garnet–kyanite assemblage is typical of Barrovian-type metamorphic terrain, whereas the staurolite–cordierite–andalusite/sillimanite assemblage results from the lower pressure Buchan-type metamorphism. During the prograde metamorphism of pelitic rocks, staurolite develops earlier than kyanite, and then together with kyanite, before being replaced by kyanite and garnet:

$$\text{staurolite} + \text{muscovite} + \text{quartz} = \text{almandine} + Al_2SiO_5 + \text{biotite} + H_2O$$

Its variable Fe/Mg ratio allow it to be used in conjunction with garnet and other ferromagnesian silicates as a geoithermometer. Staurolite is affected by retrograde metamorphism, resulting in it being mantled and partially replaced by chlorite. Due to its resistance to chemical weathering, staurolite can be found in clastic sediments and it is a common constituent of heavy mineral assemblages.

The Epidote Group

All members of the epidote group contain both single [SiO_4] tetrahedral and double [Si_2O_7] tetrahedral groups, with large cavities within this framework typically occupied by large cations, typically Ca. The general formula can be given as $X_2Y_3Z_3(O,OH,F)_{13}$, in which:

X = Ca, Ce, La, Y, Th, Fe^{2+}, Mn^{2+}, Mn^{3+}

Y = Al, Fe^{3+}, Mn^{3+}, Mn^{2+}, Fe^{2+}, Ti,

Z = Si, Be

The major compositions within the group are:

Zoisite (clinozoisite)	$Ca_2(Al,Fe^{3+})Al_2O \cdot OH \cdot Si_2O_7 \cdot SiO_4$
Epidote	$Ca_2(Fe^{3+},Al)Al_2O \cdot OH \cdot Si_2O_7 \cdot SiO_4$
Piemontite	$Ca_2(Mn^{3+},Fe^{3+},Al)_3O \cdot OH \cdot Si_2O_7 \cdot SiO_4$
Allanite	$(Ca,Mn,Ce,La,Y)_2(Fe^{2+},Fe^{3+},Al)_3O \cdot OH \cdot Si_2O_7 \cdot SiO_4$

Epidote minerals occur in a wide variety of parageneses. They are typical products of greenschist and epidote–amphibolite facies of regional metamorphism, but also form under conditions of thermal metamorphism and metasomatism, during late-stage crystallisation of acid igneous rocks, and in peraluminous granites and volcanic rocks. Epidote also occurs in the hydrothermal alteration of plagioclase. Epidote itself often has a distinctive pistachio green colour and relatively high birefringence. Piemontite is reddish brown or black, but in thin section displays striking pleochroism in yellow–amethyst–crimson; it occurs as a product of low-grade regional metamorphism and as a hydrothermal product associated with manganese deposits. The rare-earth-bearing allanite is typically dark brown and occurs mainly as an accessory mineral in many granitic rocks.

Cordierite

Cordierite is a ring silicate, typically with low refractive indices and birefringence, and tends to be confused with quartz or feldspars (*see* **Minerals:** Feldspathoids; Feldspars). Twinning is known on {110} and {310}, forming simple, lamellar, or cyclic twins. It has the composition $(Mg,Fe)_2[Si_5Al_4O_{18}] \cdot nH_2O$, and is mainly found in a metamorphic environment. The Fe/Mg ratios in coexisting cordierite and garnet in divariant equilibrium are functions of pressure and temperature and may be of use in geothermometry or geobarometry. It also occurs in contaminated igneous rocks, for example, cordierite norites.

Tourmaline

Tourmaline is a boron-bearing ring silicate with a complex chemistry. Structural determinations have shown that its general formula can be taken as $XY_3Z_6B_3Si_6(O,OH)_{30}(OH,F)$. The principal recognised varieties can be distinguished in terms of their *X*, *Y* and *Z* cations;

	X	*Y*	*Z*
Elbaite	Na	Al,Li	Al
Olenite	Na	Al	Al
Dravite	Na	Mg	Al
Schorl	Na	Fe^{2+}	Al

Tsilaisite	Na	Al,Mn	Al
Buergerite	Na	Fe^{3+}	Al
Liddicoatite	Ca	Li,Mg	Al
Uvite	Ca	Mg	Al,Mg
Feruvite	Ca	Fe^{2+}	Al,Fe^{3+},Mg
Ferridravite	Na	Mg	Fe^{3+}
Chromdravite	Na	Mg	Cr
Foitite	□	Fe^{2+}	Al,Fe^{3+}

In the *X* position, Na may be partially replaced by K or Ca if valency conditions are satisfied; thus, in uvite the substitution NaAl = CaMg. For the commonest tourmaline varieties, *Y* can be predominantly Fe^{2+} as in schorl, Mg as in dravite, or (Al + Li) as in the elbaite series.

Optically, pleochroism is variable in intensity but is particularly strong for the iron-bearing tourmalines. The absorption is always $\omega > \varepsilon$, with the result that maximum absorption occurs when the *z*-axis is lying perpendicular to the vibration direction of the polariser. Thus schorl is black, but in thin section shows strong pleochroism from blue to yellow. Dravite is brown, while in the elbaites show a range of colours in hand specimen from red (rubellite), green (verdelite), blue (indicolite), to bi-coloured or colourless (achroite). Tourmaline crystals typically show striations along the length of prismatic faces.

Tourmaline is typically a mineral of granite pegmatites, pneumatolytic veins, and some granites; it is also commonly found in metamorphic rocks, as a product of boron metasomatism, or as a result of recrystallisation of dertrital grains from the original sediment. In granitic rocks, the tourmalines belong to the schorl–elbaite series and are generally iron-rich; typical tourmaline-bearing granites have black prismatic crystals visible in hand specimen and showing yellow or bluish yellow pleochroism in thin section. Both black and coloured tourmalines occur in the original pegmatitic phase, while rubellite and the zoned tourmalines are mainly restricted to the sodium replacement unit in the so-called pneumatogenic stage.

The Chlorite Group

The term chlorite is a portmanteau name for a mineral with a layered structure resembling the micas, but generally with a green colour. The chlorites principally occur in low-grade regionally metamorphosed rocks, as hydrothermal alteration products of ferromagnesian minerals in igneous rocks, and together with clay minerals in argillaceous sediments.Their general formula can be given as (Mg, $Fe^{2+}Fe^{3+}$,Mn,Al)$_{12}$[(Si,Al)$_8$O$_{20}$](OH)$_{16}$. The chlorites have a perfect basal cleavage and occur commonly as fine-grained scaly or massive aggregates.

The occurrence of very substantial, varied, and often continuous cation substitution in the chlorites has led to a proliferation of varietal names based mainly on chemical composition; commonly these cannot be specifically characterized optically.

The crystal structure (**Figure 4**) consists of regularly alternating negatively charged tetrahedral–octahedral–tetrahedral, 2:1, layers: talc-like layers with composition $Y_6Z_8O_{20}(OH)_4$ and positively charged interlayer brucite-like layers, $Y_6(OH)_{12}$, where *Y* and *Z* represent octahedral and tetrahedral sites, respectively. The chlorites are monoclinic

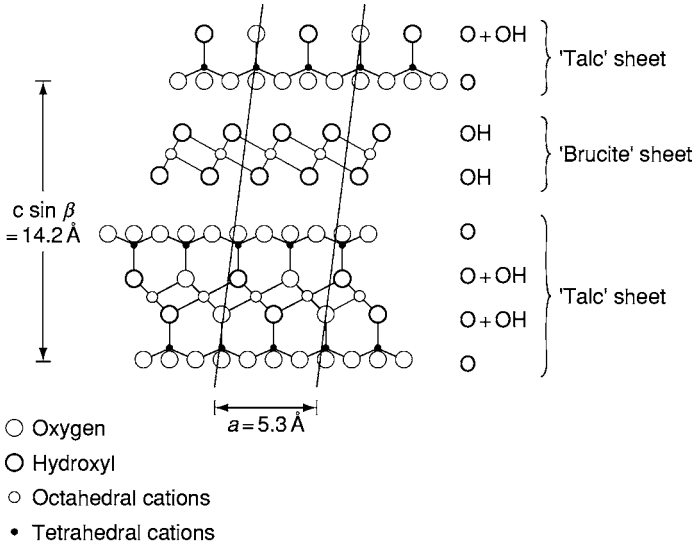


Figure 4 Idealized crystal structure of chlorite; projection on (010) (after Brown and Bailey (1962)).

(pseudohexagonal), like the micas, because of the offset of the layers.

In most common chlorites there are 12 octahedral cations per $O_{20}(OH)_{16}$ and approximately equivalent amounts of Al in tetrahedral and octahedral sites; such minerals are referred to as tri-octahedral chlorites. In a small number of chlorites, however, the number of octahedral cations is <10; these varieties are described as dioctahedal.

The principal factors influencing the optical properties of the chlorites are the iron/magnesium ratio (Figure 5). Most chlorites are optically positive, but the more iron-rich varieties are optically negative. Near the crossover point, where the birefringence passes through zero, abnormal birefringence colours (i.e., those not in the normal spectral range) may be seen.

In metamorphic rocks, chlorite is an early product of lithification and diagenesis, forming during the development of slaty cleavage in pelitic sediments. They are developed from mixed-layer clay minerals in such sediments, resulting from the expulsion of interlayer alkali ions, the fixation of (Fe,Mg) and increased tetrahedral substitution that occurs concurrently with increasing grain size. Chlorite is also a common constituent of rocks of the zeolite facies; in the lower part of the facies it is often associated with laumontite, stilbite, or heulandite, whereas in the upper part it occurs in assemblages with prehnite and pumpellyite. In the greenschist facies, chlorite–actinolite–epidote–albite assemblages are common and with increasing grade chlorite is involved in the production of amphiboles. In rocks of higher metamorphic grade, chlorite is a typical breakdown product associated with the retrogression of such ferromagnesian phases as garnet, staurolite, and biotite.

In igneous rocks, chlorite is a common product of the hydrothermal alteration of pyroxenes, amphibole, or biotite. Partial or complete chloritisation of biotite is particularly common in granites, and in most cases the transformation is pseudomorphous. The association of chlorite with albite and quartz is an essential characteristic of adinoles. Low-temperature hydrothermal Alpine-type veins in lowgrade metamorphosed sediments carry chlorite in addition to adularia and quartz.

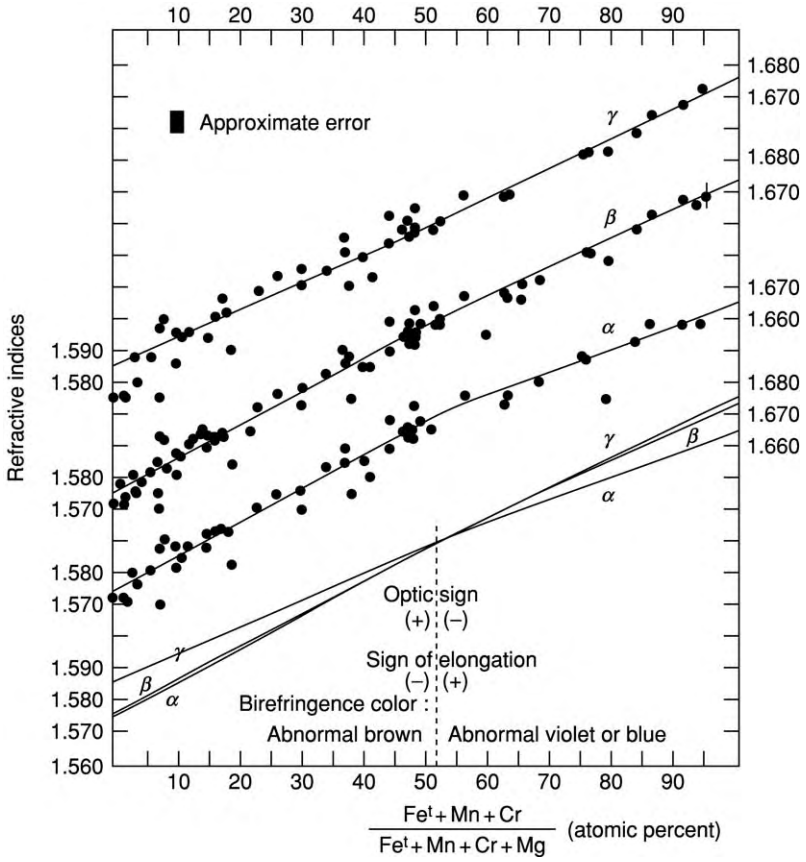


Figure 5 The relationship between optical properties and $(Fe^t + Mn + Cr)/(Fe^t + Mn + Cr + Mg)$ ratio (atomic %) for the principal chlorites (after Albee (1962)).

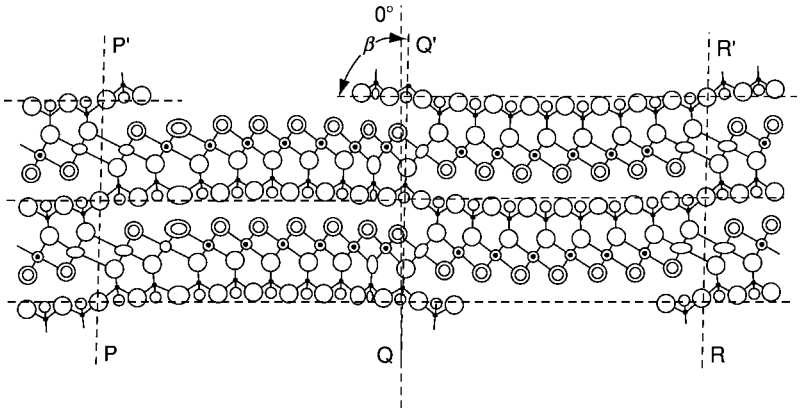


Figure 6 The crystal structure of antigorite as viewed along the *y* axis. The curved layers (radius of curvature 75 Å) reverse polarity at PP′, RR′, and near QQ′ (after Kunze (1956)).

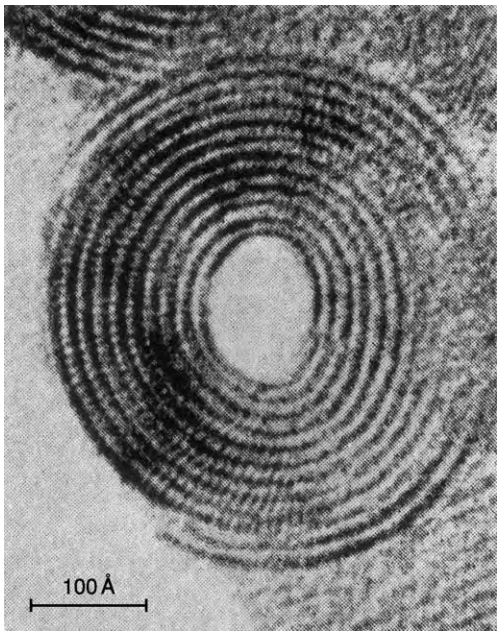


Figure 7 Transmission electron micrograph of the cross section of a fibril of chrysotile. The structural layers can be seen to be in concentric cylindrical fashion.

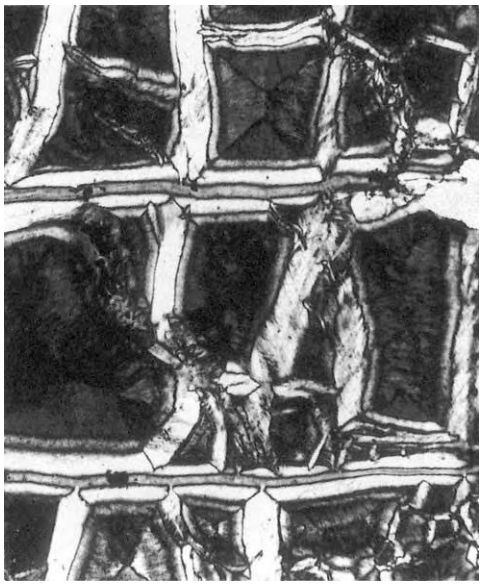


Figure 8 'Window' and 'hour glass' textures of chrysotile and lizardite in serpentinite. Crossed polars ×120.

Serpentine

There are three principal forms of serpentine – lizardite, antigorite, and chrysotile – all with approximate composition $Mg_3Si_2O_5(OH)_4$. The most abundant is lizardite and the least common is chrysotile, though the latter is perhaps best known as it often occurs in veins of silky fibres, and was an important source of commercial asbestos. All three species have a layered structure: one part consists of a pseudo-hexagonal network of linked SiO_4 terahedra all pointing one way; joined to this sheet is a brucite layer in which, on one side only, two out of every three hydroxyls are replaced by apical oxygens of the SiO_4 tetrahedra. Although much of the matrix surrounding chrysotile veins is extremely fine-grained lizardite, electron microscopy shows that it has platy morphology. The fibrous nature of chrysotile is explained by it consisting of layers curved cylindrically or spirally, usually about the *x*-axis. The fundamental fibrils typically have average outer and inner diameters of about 250 and 75 Å, respectively (Figure 6). Antigorite is structurally distinct in having a large *a* cell parameter of ~33–51 Å, which can be related to the curvature of serpentine layers about *y* and their inversion to form a corrugated sheet structure with periodicity *a* (Figure 7).

The chemistry of most serpentines is relatively simple, with minor substitution of Al for Si and of Fe for Mg. The fine-grained nature of most serpentines makes a complete optical description impossible, but the refractive indices (1.54–1.55) and birefringes are low. They show a variety of textures, many of which ('mesh' and 'hour-glass') are pseudomorphous after olivine (**Figure 8**) or 'bastite' after pyroxene or amphibole.

The serpentine minerals typically form by the retrograde hydrothermal alteration of ultrabasic igneous rocks such as dunite and peridotite. The main mineral altered is olivine which frequently contains too much iron to be accommodated in the serpentine structure, resulting in its exsolution as iron oxide; this may occur as magnetite which alters to hematite to give the reddish mottled appearance reputedly resembling the skin of a serpent. Serpentines may also form, associated with thermally metamorphosed siliceous dolomitic limestones.

See Also

Gemstones. **Minerals:** Definition and Classification; Feldspars; Feldspathoids. **Regional Metamorphism**.

Further Reading

Albee AL (1962) Relationship between the mineral association, chemical composition and physical properties of the chlorite series. *American Mineralogy* 47: 850 870.

Bailey SW and Brown BE (1962) Chlorite polytypism: I. Regular and semi random one layer structures. *American Mineralogist* 47: 819 850.

Deer WA, Howie RA, and Zussman J (1992) *An Introduction to the Rock Forming Minerals*, 2nd ed. London: Longman.

Holdaway MJ (1971) Stability of andalusite and the aluminum silicate phase diagram. *American Journal of Science* 271: 97 131.

Kunze G (1956) Die gewellte Struktur des Antigorits. *Z. Krist.* 108: 82 107.

Råheim A and Green DH (1974) Talc garnet kyanite quartz schist from an eclogite bearing terrane, western Tasmania. *Contributions to Mineralogy and Petrology* 43: 223 231.

Wood BJ (1975) The influence of pressure, temperature and bulk composition on the appearance of garnet in orthogneiss an example from South Harris, Scotland. *Earth and Planetary Science Letter* 26: 299 331.

Wood BJ (1977) The activities of components in clinopyroxene and garnet solid solutions and their applications to rocks. *Phil. Trans. Royal Society of London* 286: 331 342.

PHOSPHATES

See **SEDIMENTARY ROCKS: Phosphates**

Pyroxenes

R A Howie, Royal Holloway, University of London, London, UK

Pyroxenes are the most important group of rock-forming ferromagnesian silicates, and occur as stable phases in almost every type of igneous rock. They are also found in many rocks of widely different compositions formed under conditions of both regional and contact/thermal metamorphism.

The pyroxenes include both orthorhombic and monoclinic species: the orthopyroxenes and the clinopyroxenes. The orthopyroxenes consist essentially of a simple chemical series of (Mg, Fe)SiO_3 minerals (*see* **Minerals:** Definition and Classification) in contrast to the larger group of clinopyroxenes which have a very wide range of chemical compositions. There is a virtually complete series with a complete range of Fe/Mg ratios.

The clinopyroxenes can be considered, as a first approximation, as members of the four-component system $CaMgSi_2O_6$–$CaFeSi_2O_6$–$Mg_2Si_2O_6$–$Fe_2Si_2O_6$. The nomenclature used to describe these pyroxenes is illustrated in Figure 1 (*see* **Rocks and Their Classification**). The $Ca_2Si_2O_6$ end-member (wollastonite) is not a pyroxene (see below).

The crystal structure of the pyroxenes is based on infinite single chains of linked silicon—oxygen tetrahedra, the [Si_2O_6] groups, these chains being linked laterally by octahedral layers of (Ca, Mg, Fe, Na, etc.) cations (Figures 2 and 3). The formulae of the commoner pyroxenes are as follows:

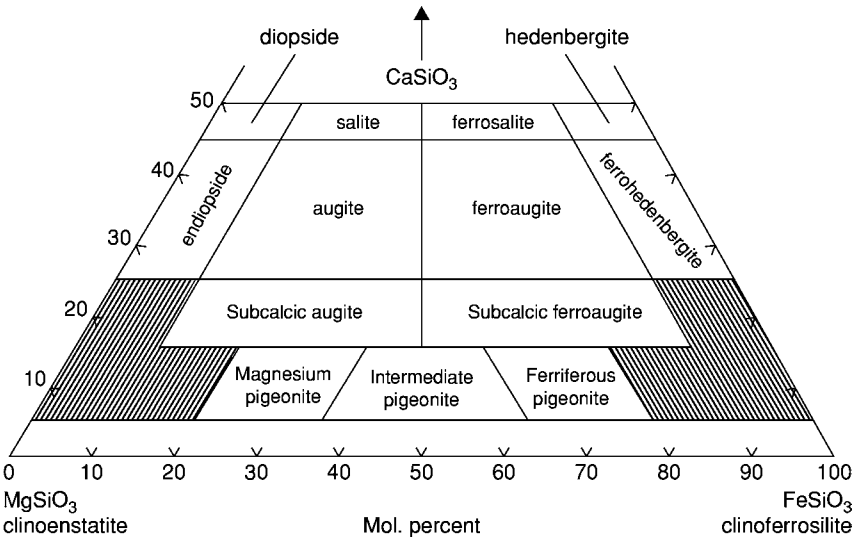


Figure 1 Nomenclature of clinopyroxenes in the system $CaMgSi_2O_6$ $CaFeSi_2O_6$ $Mg_2Si_2O_6$ $Fe_2Si_2O_6$.

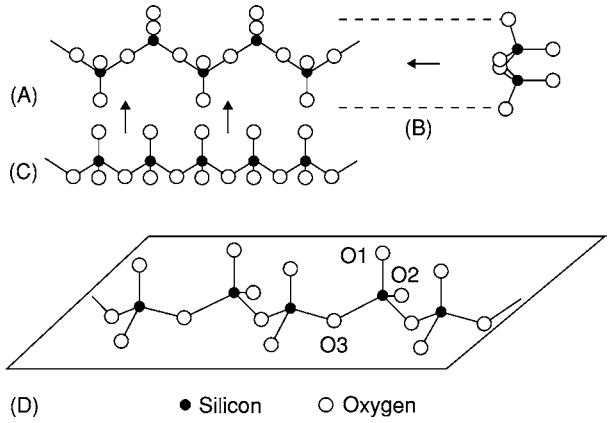


Figure 2 Idealized illustration of a single pyroxene chain as seen in projections on (A) (100), (B) along *z*, and (C) along *y*.

Orthopyr-	$Mg_2[Si_2O_6]$	enstatite
oxenes	$Fe_2^{2+}[Si_2O_6]$	ferrosilite
Clinopyr-	$CaMgSi_2O_6$	diopside
oxenes	$CaFe^{2+}[Si_2O_6]$	hedenbergite
	$(Mg, Fe^{2+}Ca)$	pigeonite
	$(Mg, Fe^{2+})[Si_2O_6]$	
	$(Ca, Mg, Fe^{2+}, Fe^{3+}, Ti, Al)_2$	augite
	$[(Si, Al)O_6]$	ferroaugite
	$(Ca, Na)(Mg, Fe^{2+}, Fe^{3+}, Al)$	
	$[Si_2O_6]$	omphacite
	$NaAl[Si_2O_6]$	jadeite
	$NaFe^{3+}[Si_2O_6]$	aegirine
	$(Na, Ca)(Fe^{3+}, Fe^{2+}, Mg)$	aegirine
	$[Si_2O_6]$	augite
	$LiAl[Si_2O_6]$	spodumene

Optically, the orthopyroxenes are colourless to greenish under the microscope and bronze to black in hand specimen. They have a good {210} cleavage and straight extinction. Their other properties vary with the Fe/Mg ratio.

The clinopyroxenes also are colourless to greenish under the microscope, have a good {210} cleavage, but with variable extinction angles from 0 to 45°. Within the $CaMgSi_2O_6$–$CaFeSi_2O_6$–$Fe_2Si_2O_6$–$Mg_2Si_2O_6$ quadrilateral, which covers 90% of the common rock-forming clinopyroxenes, their optical properties vary systematically, allowing their composition to be estimated.

The occurrence of members of the pyroxene group in general reflects the compositions of their parent rocks. Mg-rich enstatites are important constituents associated with olivine, diopside, and spinel in ultrabasic and ultramafic rocks. Other orthopyroxenes of intermediate composition are essential constituents of the granulite facies of metamorphic rocks. Diopside is formed by the metamorphism of siliceous dolomites, but is not uncommon as an early of basic igneous rocks. Pigeonite is a constituent of many andesites and dacites and can be distinguished by its small optic axial angle. In layered igneous intrusions, it may invert to lamellae of augite ± orthopyroxene. The most common clinopyroxenes are augites; these are essential constituents of basic igneous rocks, and also occur in ultramafic rocks, lherzolites and websterites, and in ultrabasic nodules and xenocrysts in basic rocks. They also occur in dolerites and in basic lavas.

The sodium-bearing clinopyroxene Jadeite is a diagnostic mineral of rocks generally considered to have formed in a high-pressure environment; it is particularly characteristic of eclogites and related

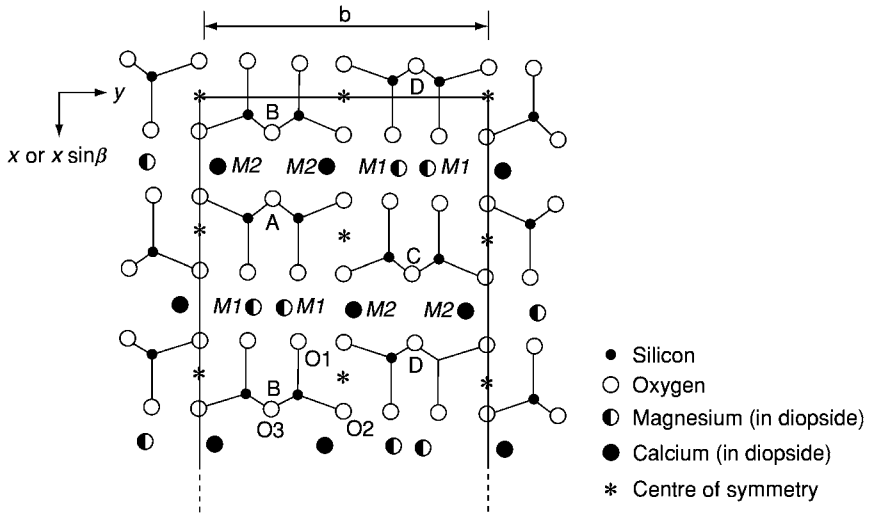


Figure 3 Projection along *z* of an idealized pyroxene structure, showing the pyroxene chains and the two distinct cation sites *M*1 and *M*2. The latter are in oxygen polyhedra which also form chains or bands parallel to *z*.

rocks. When near its end-member composition of $NaAlSi_2O_6$ Jadeite, is colourless or pale green (one variety of the gem material jade); it is typically found in metagreywackes and related rocks of regional metamorphism, particularly in the blueschist facies. Aegirine and aegirine-augite illustrate the replacement of ferric iron in the substitution $NaFe^{3+} = Ca(Mg, Fe^{2+})$; they are mainly found in alkaline or peralkaline igneous rocks. The lithium pyroxene spodumene is generally white and is restricted to Li-rich granitic pegmatites.

Wollastonite ($CaSiO_3$) and the three manganese-rich silicates rhodonite, bustamite, and pyroxmangite are all triclinic. Sometimes referred to as 'pyroxenoids', they differ from the pyroxenes in having, respectively, three, five, five, and seven silica tetrahedra in their structural chains. Wollastonite is a common mineral of metamorphosed siliceous limestones, whereas the three manganese-rich species typically occur in Mn-orebodies, usually associated with metasomatic activity.

See Also

Minerals: Definition and Classification; Amphiboles; Other Silicates. **Rocks and Their Classification**.

Further Reading

Deer WA, Howie RA, and Zussman J (1992) *An Introduction to the Rock Forming Minerals,* 2nd edn. London: Longman.

Quartz

R A Howie, Royal Holloway, University of London, London, UK

Quartz is the second most abundant mineral in the Earth's crust being hard (7 on Mohs scale) and resistant to many forms of chemical weathering, with the result that it becomes concentrated in sedimentary rocks while the coexisting feldspars are rapidly weathered away. It has a compact framework structure and has trigonal symmetry. It is an essential mineral in many acid plutonic igneous rocks such as the granites (*see* **Igneous Rocks:** Granite) and granodiorites, and also in hypabyssal and volcanic rocks of equivalent composition. Metamorphism of such igneous or sedimentary rocks gives rise to quartzites and quartz-rich veins.

Although the chemical composition is comparatively simple, most quartz consisting of 99.9% SiO_2, it has α (low-temperature) and β (high-temperature) forms, with a transition at around 573°C, and several polymorphic forms, mostly at higher temperatures and pressures. This leads to the phase diagram for SiO_2 being extremely complex. There are the high-temperature polymorphs tridymite and cristobalite, both typical of volcanic rocks, and the more recently discovered high-pressure phases coesite and stishovite (Figure 1). Coesite occurs at a static pressure of

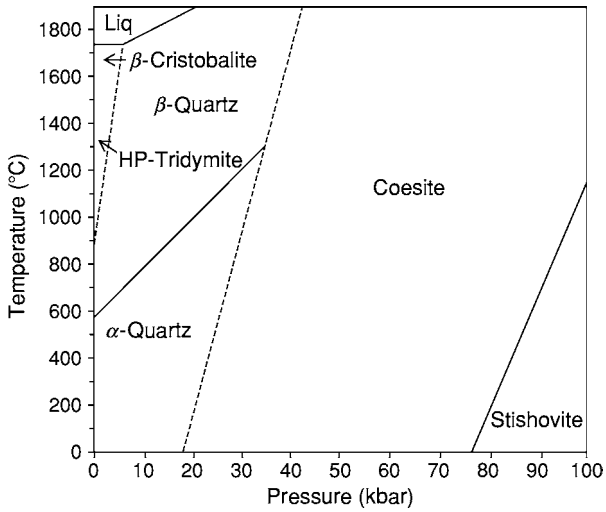


Figure 1 Phase diagram for the silica system (after Klein and Hurlbut, (1973) *Manual of Mineralogy*, 21st edn, p. 527. © *John Wiley and Sons* (Reprinted with permission).

$\geq$29 kbar; it is found as inclusions, generally partially replaced by quartz, in zircon, garnet, monazite, and other phases formed in low-temperature–high-pressure conditions and is used as an indicator of such ultrahigh-grade metamorphic conditions in rocks which have generally been exhumed relatively rapidly in orogenic belts (as in the Himalaya, the Alps and in SE China). Stishovite requires even higher pressures (of around 130 kbar) for its formation and is typical found in meteorites or meteorite impact craters, also in tektites.

Microscopic fluid inclusions often occur in quartz, and may contain a gas phase as well as a liquid and one or more microscopic crystals of such minerals as halite, sylvite, calcite, etc., believed to have been precipitated from the same source as the quartz. The development of sophisticated microanalytical techniques has allowed the composition of these liquids and solids to be determined and to the determination of the temperature at which the original precipitation and crystallization occurred. This has important applications in unravelling the petrogenetic history of both igneous and metamorphic rocks as well as in the geochemistry and conditions of deposition of ore deposits and the provenance of sedimentary rocks (*see* **Sedimentary Rocks:** Mineralogy and Classification).

Chalcedony is a group name for the compact varieties of silica composed of minute crystals of quartz with submicroscopic pores. The colour and texture vary considerably according to the impurities present, but in general such materials may be subdivided into chalcedony in which the colour is fairly uniform, and agate in which the colour may be arranged in bands or concentric zones. The terms 'chert (*see* **Sedimentary Rocks:** Chert)' and 'flint' are used for opaque dull-coloured or black chalcedony, and in common usage chert is taken as the name for this material when it occurs in stratified or massive form in rocks, while the term flint is generally restricted to dark chalcedony occurring in nodular form in a rock matrix, particularly in the chalk. Jasper is a red opaque massive form of chalcedony.

Natural hydrous silica may be divided into three well-defined structural groups (*see* **Minerals:** Other Silicates): opal-C (well-ordered γ-cristobalite), opal-CT (disordered α-cristobalite, α-tridymite), and opal-A (highly disordered, nearly amorphous). Opal may be colourless, milky white, yellow, red, green, blue, or black. In precious opal a play of delicate colours is seen, due not to chemical impurities but to optical diffraction from a randomly faulted close-packed structure of minute transparent silica spheres (around 0.25 μm). Colourless opal is a major constituent in hot-spring deposits and deep-sea precipitates.

Quartz is an extremely important in many industrial processes. At the simplest level it is raw material for the production of glasses, ceramics, and foundry moulds. Most quartz sands are a golden yellow due to the quartz particles being encased by a thin coating of iron oxides and/or hydroxides, and the iron would cause any glass to be dark green. In certain geological situations, however, the beach sands may be rolled to and fro by the tides for several million years, leading to the erosion of the iron coating and forming silver sands, which are much sought after for glass manufacture. For the more technical applications requiring an even higher degree of purity, colourless vein quartz is used, or nowadays reliance is placed on the manufacture of synthetic single crystals of quartz in an autoclave. One of the main technical aspects of quartz lies in its piezoelectric properties, whereby the crystal develops an electric polarization when mechanically stressed in an appropriate direction. In the construction of quartz oscillators, thin plates of quartz are cut in a particular direction specified to generate a desired resonant vibrational frequency.

Although most quartz is colourless or white, well-known coloured varieties include amethyst (violet), citrine (yellow to brown), and smoky quartz. Each of these contains either a substitutional or interstitial component other than SiO_2. For amethyst and citrine, iron is an important constituent (Fe 25–280 ppm). In smoky quartz the brownish or black colour is due to natural radiation acting upon Al^{3+}-bearing quartz.

Quartz is an essential constituent in many igneous, sedimentary, and metamorphic rocks, and also occurs as secondary material, often forming a cementing

medium in sediments. Quartz is also a common constituent of hydrothermal veins.

In granites, microgranites, tonalities, etc., quartz typically forms anhedral grains due to it being one of the final minerals to crystallize. Quartz phenocrysts in fine-grained rapidly cooled rhyolites, pitchstones, and quartz porphyries may show euhedral outlines, although in some of these rocks the quartz phenocrysts may suffer magmatic corrosion. In intermediate igneous rocks, the amounts of quartz are less than in those of granitic composition, and in basic rocks it usually amounts to less than 5% (though it is more abundant in some quartz dolerites).

Because of its chemical and physical resistance to erosion, quartz is an abundant mineral and becomes concentrated during sedimentary processes to give rise to sands and sandstones of various types. Secondary quartz is often deposited around pre-existing grains (of quartz or other minerals) and is a common cementing material in sediments.

Quartz is a common mineral in many metamorphic rocks, occurring in the metamorphosed equivalent of quartz-bearing sediments and igneous rocks. Although in low-grades of metamorphism quartz may survive unchanged, in the higher grades it undergoes recrystallization with concomitant increase in grain size. In addition, much is developed by the release of SiO_2 in reactions taking place during metamorphism. Intergrowths of quartz and other minerals are fairly common, particularly those of quartz and K-feldspar as in graphic or micrographic granite, and quartz and plagioclase as in myrmekite in association with K-feldspar.

The typical occurrence of tridymite is in acid volcanic rocks such as rhyolite, obsidian, trachyte, andesite, and dacite. In such rocks it is often found in cavities and may be associated with such minerals as sanidine, and less frequently augite or fayalite. Cristobalite is also a mineral of volcanic rocks, and may be found in (metastable) association with tridymite. It is often a late product of crystallization, sometimes replacing tridymite. The ability of cristobalite to occur as a metastable form, outside its equilibrium field, means that no definite conclusions can normally be drawn as to conditions at the time of its crystallization.

The SiO_2 polymorph coesite was first identified with quartz and fused silica glass in sheared porous sandstones at Meteor Crater, Arizona, and in granite and pumiceous tuff near the rim of the Ries crater, Bavaria, developed by the shock wave generated by the meteoritic impact. It is also being increasingly recognised as microscopic inclusions in such minerals as zircon, pyrope garnet, and other high-grade metamorphic minerals where it has often been partially inverted to quartz, resulting in radial cracking of the garnet around the inclusions. This coesite is considered to have crystallized in nearly static pressure conditions at $\geq$28 kbar, followed by fairly rapid exhumation.

See Also

Igneous Rocks: Granite. **Minerals:** Definition and Classification; Other Silicates. **Sedimentary Rocks:** Mineralogy and Classification; Chert; Sandstones, Diagenesis and Porosity Evolution. **Tektites**. **Ultra High Pressure Metamorphism**.

Further Reading

Audetát A and Gunther D (1999) Mobility and H_2O loss from fluid inclusions in natural quartz crystals. *Contributions to Mineral Petrology* 137: 1 14.

Chopin C (1984) Coesite and pure pyrope in high grade blueschists of the Western Alps: a first record and some consequences. *Contributions to Mineral Petrology* 86: 107 118.

Heaney PJ, Prewitt CT, and Gibbs GV (eds.) (1994) Silica: Physical Behavior, Geochemistry and Materials Applications Reviewal Mineralogy 29: 606.

Klein C and Hurlbut CS (1993) *Manual of Mineralogy,* 21st edn. New York: John Wiley.

Liou JG and Zhang RY (1996) Occurrence of intergranular coesite in ultrahigh *P* rocks from the Sulu region, eastern China: implications for lack of fluid during exhumation. *American Mineralogy* 81: 1217 1221.

O'Brien PJ, Zotov N, Law R, Khan MA, and Jan MQ (2001) Coesite in Himalayan eclogite and implications for models of India Asia collision. *Geology* 29: 435 438.

Sulphates

G Cressey, The Natural History Museum, London, UK

Introduction

Gypsum ($CaSO_4 \cdot 2H_2O$), the most common of the sulphate minerals, is also known as alabaster (a fine-grained massive form), satin spar (a fibrous variety of gypsum), or selenite (colourless transparent gypsum crystals). Gypsum is often found in considerable thicknesses within evaporite sequences and in association with limestones and shales. Large deposits of alabaster gypsum are commonly observed at the sole of many large-scale tectonic dislocations and thrust faults. Gypsum is of great importance as a raw material in the production of cement and plaster for the building industry.

Crystal Structure of Gypsum

In the crystal structure of gypsum, calcium ions are coordinated by six oxygens from sulphate (SO_4) tetrahedral groups, and by two oxygens from water molecules (H_2O). Two layers of SO_4 tetrahedra are bound together by calcium ions forming a double-layer sheet. Water molecules are located at each side of these double-layer sheets. The water molecules are oriented such that one of the two hydrogens of each water molecule points almost perpendicular to the sheets; this facilitates weak hydrogen bonding to the next layer in the structure. These hydrogen bonds point approximately along the *y*-axis of the crystal, forming a layer of weak hydrogen bonding perpendicular to the *y*-axis, and explain the easy (010) cleavage of gypsum. The proton (H^+) positions in gypsum have been determined in deuterated gypsum by neutron diffraction, as weakly scattering proton positions cannot be determined accurately by X-ray diffraction methods (Figure 1).

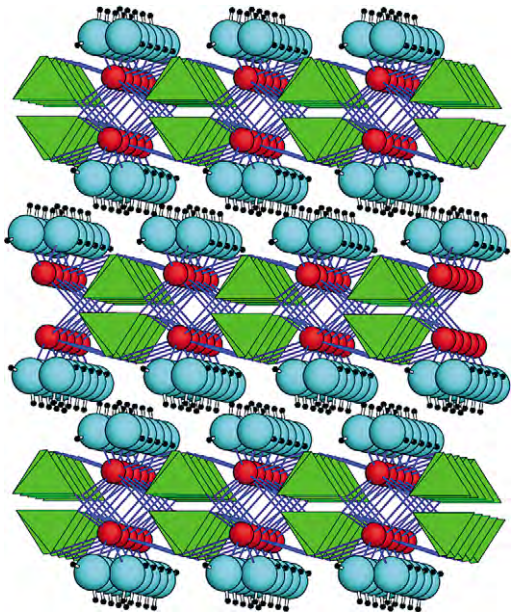


Figure 1 The structure of gypsum (view looking almost down the *z* axis): SO_4 tetrahedral units (green) and calcium ions (red) form double layer sheets perpendicular to the *y* axis. At each side of these sheets, water molecules (with oxygen in blue and hydrogen shown as small black dots) form weak hydrogen bonds to the next layer in the structure.

Physical Properties of Gypsum

The weak hydrogen bonding in the structure of gypsum, occurring in layers perpendicular to the *y*-axis, accounts for the perfect (010) cleavage and is also responsible for the anisotropic thermal expansion behaviour of gypsum. The *b* dimension of the unit cell increases the most with temperature, and this is principally because the hydrogen bonds oriented across the sheets easily lengthen with increasing temperature. The static compressibility of gypsum is also highly anisotropic and, as is the case with the thermal expansion, is related to the hydrogen-bonding orientation across the sheet structure; these hydrogen bonds more easily shorten with pressure and therefore the *b* lattice parameter has the greatest compressibility.

Upon heating to about 65°C, gypsum dehydrates to form the hemihydrate phase, bassanite ($CaSO_4 \cdot 0.5H_2O$), with its *z*-axis parallel to that of the gypsum; this is known as plaster of Paris. Further heating dehydrates the sulphate completely, forming γ-$CaSO_4$ (a polymorph of anhydrite, $CaSO_4$) at about 95°C. γ-$CaSO_4$ is also known as 'soluble anhydrite', as unlike anhydrite it can be re-hydrated. Anhydrite has a different structure and cannot be easily hydrated. Commercial plaster of Paris consists mainly of calcium sulphate hemihydrate (bassanite) and is made by heating gypsum to about 200°C. Heating gypsum at higher temperatures produces anhydrite and this change is irreversible. At the stage where bassanite and/or γ-$CaSO_4$ are produced, re-hydration to gypsum ($CaSO_4 \cdot 2H_2O$) easily occurs; this is the process of the setting of plaster of Paris to form an interlocking mass of gypsum crystals.

Occurrence of Gypsum

Gypsum occurs mainly as sedimentary deposits associated with limestones and shales and in evaporite deposits. Gypsum can form by the direct evaporation of brine, by hydration of anhydrite and by oxidation of sulphides. Gypsum is also one of the products of the reaction of acid sulphate solutions with carbonate rocks. Seawater contains about 3.5% dissolved material, of which 80% is sodium chloride and about 4% is dissolved calcium sulphate. The experimental evaporation of seawater results in the crystal deposition sequence: calcium carbonate – calcium sulphate – sodium chloride – sulphates/chlorides of magnesium – sodium bromide/potassium chloride.

Gypsum also occurs in soils either as disseminated crystals or in horizons beneath calcrete layers. Percolating waters, which in dry seasons are drawn to the surface by capillary action, can evaporate and crystallize gypsum, sometimes in the form described as 'desert roses'. The calcium sulphate formed in evaporates is sometimes gypsum, sometimes anhydrite, and often both minerals occur together. Some geological evidence suggests that the original material of certain gypsum beds was anhydrite; this is supported by the fact that some gypsum beds grade into anhydrite at depth. Along with jarosite and goethite, gypsum is one of the products formed by the oxidation of iron sulphide (pyrite). Gypsum is also found in volcanic regions along with native sulphur where it forms by the reaction of sulphur-bearing fluids and gases with calcium-bearing rocks. Natural bassanite is rare, and nearly always forms as a product of alteration of gypsum. It has been found along with gypsum in volcanic craters, and in some soils and peat deposits.

Other Sulphates

The most abundant barium mineral in the Earth's crust is barite, $BaSO_4$, and this is also the least soluble sulphate mineral. It commonly occurs in hydrothermal metalliferous veins, but also occurs in cementations and nodules in sedimentary rocks. Because of its high density (4.5 g cm ^{3}) barite is widely used as a component of drilling muds in the petroleum industry, or in other dense fluid media, and also as a filler or extender in the manufacture of paper, plaster, rubber and plastics. There is a complete solid solution series between $BaSO_4$ (barite) and $SrSO_4$ (celestine), although natural compositions are usually near to end-member compositions, reflecting either the Ba-rich or Sr-rich geochemical environments. $PbSO_4$ (anglesite) is isostructural with $BaSO_4$ (barite), so a complete solid solution between these two end members would also be expected, and some natural samples occur that are intermediate in composition. At room temperature, about 6% Ca^{2+} can be accommodated in the $BaSO_4$ structure. About 8% Ba^{2+} replacement of Ca^{2+} is the limit of solid solution tolerated by the $CaSO_4$ (anhydrite) structure.

The oxidation of pyrite (FeS_2), especially when mediated by the action of bacteria, frequently results in the formation of powdery assemblages of iron sulphates. This alteration commonly occurs during the storage of pyrite-bearing materials under humid conditions. Sulphates produced in this way include melanterite ($FeSO_4 \cdot 7H_2O$), rozenite ($FeSO_4 \cdot 4H_2O$), szomolnokite ($FeSO_4 \cdot H_2O$) and rhomboclase ($Fe^{3+}(SO_4)_2\,(H_3O^+) \cdot 3H_2O$).

Several hundred compounds that contain essential sulphate (SO_4) groups are known to occur in nature; these include sulphates, sulphate hydrates, sulphate hydroxy hydrates, sulphate/chlorides, and sulphate/arsenate/phosphate/vanadates of alkali- and transition-metals.

See Also

Minerals: Definition and Classification; Sulphides. **Sedimentary Rocks:** Evaporites.

Further Reading

Alpers CN, Jambor JL, and Nordstrom DK (eds.) (2000) *Sulfate Minerals Crystallography, Geochemistry and Environmental Significance*. Reviews in Mineralogy and Geochemistry, vol. 40. Mineralogical Society of America.

Bragg WL (1937) *The Atomic Structure of Minerals*. New York: Cornell University Press, and London: Oxford University Press.

Chang LLY, Howie RA, and Zussman J (1996) *Rock Forming Minerals,* vol. 5B, *Sulphates, Carbonates, Phos phates and Halides*. London: Longman.

Jerz JK and Rimstidt JD (2003) Efflorescent iron sulfate minerals: paragenesis, relative stability, and environmen tal impact. *American Mineralogist* 88: 1919 1932.

Pedersen BF and Semmingsen D (1982) Neutron diffraction refinement of the structure of gypsum, $CaSO_4 \cdot 2H_2O$. *Acta Crystallographica* B38: 1074 1077.

Wooster WA (1936) On the crystal structure of gypsum $CaSO_4 \cdot 2H_2O$. *Zeitschrift fur Kristallographie* 94: 375 396.

Sulphides

D J Vaughan, University of Manchester, Manchester, UK

Introduction

The sulphide minerals are the major source of world supplies of a very wide range of metals and are the most important group of ore minerals. In addition to their concentrated occurrence in ore deposits and areas of mineralization, a limited number of sulphide minerals are found as accessory minerals in rocks. However, only pyrite, pyrrhotite, galena, sphalerite, chalcopyrite, and possibly the binary copper sulphides of the chalcocite family could be classed as rock-forming minerals. Pyrite (FeS_2) is by far the most abundant sulphide mineral. The very fine particle iron sulphides found in reducing environments beneath the surfaces of Recent sediments and soils, although transient species, are also volumetrically important. Formerly referred to as amorphous iron sulphide, these phases are now known to be largely composed of fine-particle mackinawite (tetragonal FeS) and, in some cases, the sulphospinel mineral greigite (Fe_3S_4). Both phases are metastable in relation to pyrite and pyrrhotite. As well as being important sources of metals, sulphide minerals are potential sources of pollution. In particular, the release of sulphur through the weathering of sulphides in natural rocks or in mine wastes generates sulphuric acid, resulting in acid rock drainage or acid mine drainage, and a significant proportion of the sulphurous fumes generated by combustion of coal (or by the smelting of sulphide ores) originates from sulphide minerals; these fumes react to form acid rain. A more positive role may have been played by sulphide minerals in the emergence of life on Earth; although speculative, such theories have been developed following the discoveries of ocean-floor hydrothermal systems generating large volumes of sulphide minerals and associated with novel life forms and ecosystems.

The mineralogy, mineral chemistry, and geological occurrences of sulphide minerals have been well researched, particularly because of their economic and environmental importance. The crystal structures, chemical compositions, physical properties, thermochemistry, and phase relations of the major sulphides were well established by the latter part of the twentieth century, as were the macroscopic and microscopic characteristics of sulphide minerals in natural occurrences. Recent decades have seen major advances in the understanding of the chemical bonding in sulphide minerals and in the surface chemistry of sulphides. An understanding of the genesis of sulphide-ore deposits has developed through systematic geological, petrographic, compositional, and isotopic studies of ores and host rocks.

In this article, the compositions, crystal structures, physical properties, and certain aspects of mineral chemistry and geological occurrence of the major sulphide minerals will be reviewed, with suggestions given for further reading.

Compositions and Structures

Several of the common sulphide minerals were among the first minerals to be studied by X-ray crystallography, and since that time the structures of nearly all the mineralogically significant sulphides have been determined. It is possible to categorize the mineral sulphides into a series of groups based on major structure types, or having key structural features in common, as shown in Table 1. In many cases, these structures are the classic structures of crystalline solids, such as the rock-salt structure of the galena group (Figure 1A), the sphalerite and wurtzite forms of zinc sulphide (Figures 1B and 1C), or the nickel arsenide structure (Figure 1E). The disulphides are characterized by the presence of dianion (S–S, S–As, As–As, etc.) units in the structure: as well as the pyrite structure, in which FeS_6 octahedral units share corners along the *c*-axis direction, there is the marcasite form of FeS_2, in which octahedra share edges to form chains along the *c*-axis (Figure 1D). The structures of loellingite ($FeAs_2$) and arsenopyrite (FeAsS) are variants of the marcasite structure that have shorter and alternate long and short, respectively, metal–metal distances across the shared octahedral edge (see Figure 1D). A few sulphides such as molybdenite, covellite (Figure 1F), and mackinawite have layer structures, and a small number exhibit structures that are best characterized as containing rings or chains of linked atoms (e.g. realgar, AsS). A diverse group of sulphides, the metal-excess group, is composed of an unusual set of structures that is well illustrated by the important example of the mineral pentlandite ($(Ni,Fe)_9S_8$; Figure 1G).

As can be seen from Table 1, in many of these groups a number of minerals share the actual structure type, but there are also, commonly, other minerals that have structures based on these 'parent' structures and that can be thought of as being derived from them. The relationship between a derivative structure and the parent structure may involve

Table 1 Sulphide structural groups

1. The disulphide group					
Pyrite structure	*Marcasite structure*	*Arsenopyrite structure*	*Loellingite structure*	*Derived by As/S ordered substitution*	
FeS_2 pyrite	FeS_2 marcasite	FeAsS arsenopyrite	$CoAs_2$ safflorite	(Co,Fe)AsS cobaltite	
CoS_2 cattierite		FeSbS gudmundite	$FeAs_2$ loellingite	(Ni,Co,Fe)AsS gersdorffite (I)	
			$NiAs_2$ rammelsbergite		
2. The galena group					
PbS galena					
α MnS alabandite					
3. The sphalerite group					
Sphalerite structure	*Derived by ordered substitution*	*Stuffed derivatives*			
β ZnS sphalerite	$CuFeS_2$ chalcopyrite	$Cu_9Fe_8S_{16}$ talnakhite			
CdS hawleyite	Cu_2FeSnS_4 stannite	$Cu_9Fe_9S_{16}$ mooihoekite			
Hg(S,Se) metacinnabar	Cu_2ZnSnS_4 kesterite	$Cu_4Fe_5S_8$ haycockite			
4. The wurtzite group					
Wurtzite structure	*Composite structure derivatives*	?*Further derivatives*	*Derived by ordered substitution*		
α ZnS wurtzite	$CuFe_2S_3$ cubanite	$Cu_2Fe_2SnS_6$ hexastannite	Cu_3AsS_4 enargite		
CdS greenockite	?$AgFe_2S_3$ argentopyrite				
5. The nickel arsenide group					
NiAs structure	*Distorted derivatives*	*Ordered omission derivatives*			
NiAs niccolite	FeS troilite	Fe_7S_8 monoclinic pyrrhotite			
NiSb breithauptite	CoAs modderite	Fe_9S_{10}, $Fe_{11}S_{12}$ hexagonal pyrrhotite, etc?			
6. The thiospinel group					
Co_3S_4 linnaeite					
$FeNi_2S_4$ violarite					
$CuCo_2S_4$ carrollite					
7. The layer sulphides group					
Molybdenite structure	*Tetragonal PbO structure*	*Covellite structure*			
MoS_2 molybdenite	$(Fe,Co,Ni,Cr,Cu)_{1+x}S$	CuS covellite			
WS_2 tungstenite	mackinawite	$\sim Cu_3FeS_4$ idaite			
8. Metal excess group					
Pentlandite structure	*Argentite structure*	*Chalcocite structure*	*Digenite structure*	*Derived by ordered substitution*	*Nickel sulphide structures*
$(Ni,Fe)_9S_8$ pentlandite	Ag_2S argentite	Cu_2S chalcocite	Cu_9S_5 digenite	Cu_7S_4 anilite	NiS millerite
		?derivative			Ni_3S_2 heazlewoodite
		$Cu_{1.96}S$ djurleite			
9. Ring or chain structure group					
Stibnite structure	*Realgar structure*	*Cinnabar structure*			
Sb_2S_3 stibnite	As_4S_4 realgar	HgS cinnabar			
Bi_2S_3 bismuthinite					

(I) one polymorph.

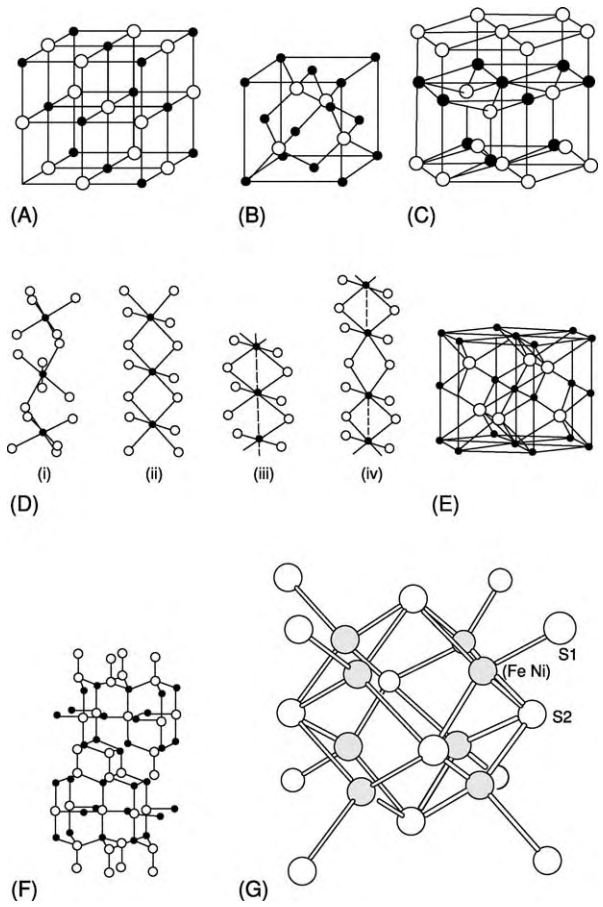


Figure 1 Crystal structures of the major sulphides: (A) galena (PbS) structure; (B) sphalerite (ZnS) structure; (C) wurtzite (ZnS) structure; (D) pyrite structure and the linkage of metal sulphur octahedra along the *c* axis direction in (i) pyrite (FeS_2), (ii) mar casite (FeS_2), (iii) loellingite ($FeAs_2$), and (iv) arsenopyrite (FeAsS); (E) niccolite (NiAs) structure; (F) covellite (CuS) struc ture; and (G) cube cluster of tetrahedrally coordinated metals in the pentlandite ($(Ni,Fe)_9S_8$) structure. In each case the metals are shown as smaller or shaded spheres. Adapted from Institute of Materials, Minerals and Mining, Sulphite Deposits their origin and Processing, ed. P. Gray, 1990.

- ordered substitution (the structure of chalcopyrite ($CuFeS_2$) is derived from that of sphalerite (ZnS) by alternate replacement of zinc atoms with copper and iron, resulting in an enlarged (tetragonal) unit cell; the structure of stannite (Cu_2FeSnS_4) results from further ordered substitution of half of the iron atoms in chalcopyrite by tin (Figure 2A));
- a stuffed derivative (talnakhite ($Cu_9Fe_8S_{16}$) is derived from chalcopyrite by the occupation of additional normally empty cavities in the structure (Figure 2B));
- ordered omission (monoclinic pyrrhotite (Fe_7S_8) is derived from the nickel arsenide structured troilite by removal of iron atoms, leaving holes (vacancies) that are ordered (Figure 2C)); or
- distortion (the troilite form of FeS is simply a distortion of the parent nickel arsenide structure form (Figure 2C)).

In some cases, the relationships involved are more complex, as, for example, in certain of the sulphosalts (minerals with a general formula $A_mT_nX_p$ in which common elements are *A*: Ag, Cu, and Pb; *T*: As, Sb, and Bi; *X*: S, and which contain pyramidal TS_3 groups in the structure). Here, the resulting structure is composite and made up of slabs or units of a parent structure (or structures) arranged in an ordered fashion.

The compositional variations of the major sulphide minerals are reasonably well characterized, both as a result of numerous analyses of natural samples from a wide variety of deposits and as a result of systematic laboratory investigations of the phase equilibria. Many metal sulphides show evidence that the elements comprising them are not combined in simple whole-number ratios, i.e. they exhibit nonstoichiometry. In certain cases, the deviation from a simple ratio is considerable. For example, the pyrrhotites are sometimes given the general formula $Fe_{1-x}S$ where $0 < x < 0.125$ and the varying compositions correspond to varying concentrations of vacancies in sites that would otherwise be occupied by iron atoms. However, in systems such as these, ordering of the vacancies occurs at low temperatures, and the result may be a series of stoichiometric phases of slightly different compositions. Although Fe_7S_8 has a (monoclinic) superstructure resulting from vacancy ordering, the situation in the intermediate or hexagonal pyrrhotites is more complex. Some of these pyrrhotites may represent ordered phases with clearly defined compositions (Fe_9S_{10}, $Fe_{11}S_{12}$, etc.), but more complex and partial ordering in these systems may occur. One problem is certainly that free-energy differences between a series of phases resulting from vacancy ordering are very small, making a successful investigation of the relationships between synthetic products very difficult.

Other examples of nonstoichiometry involve relatively small deviations from the simple ratio. For example, galena (PbS) exhibits a range of nonstoichiometry of 0.1 atomic %. Galena is apparently stable over a wide range of values of aS_2 (activity of sulphur), and at high aS_2 it has lead vacancies, whereas at low aS_2 there are sulphur vacancies.

Electrical, Magnetic, and Optical Properties

The metal sulphides also show a tremendous range of electrical and magnetic behaviours. As Table 2

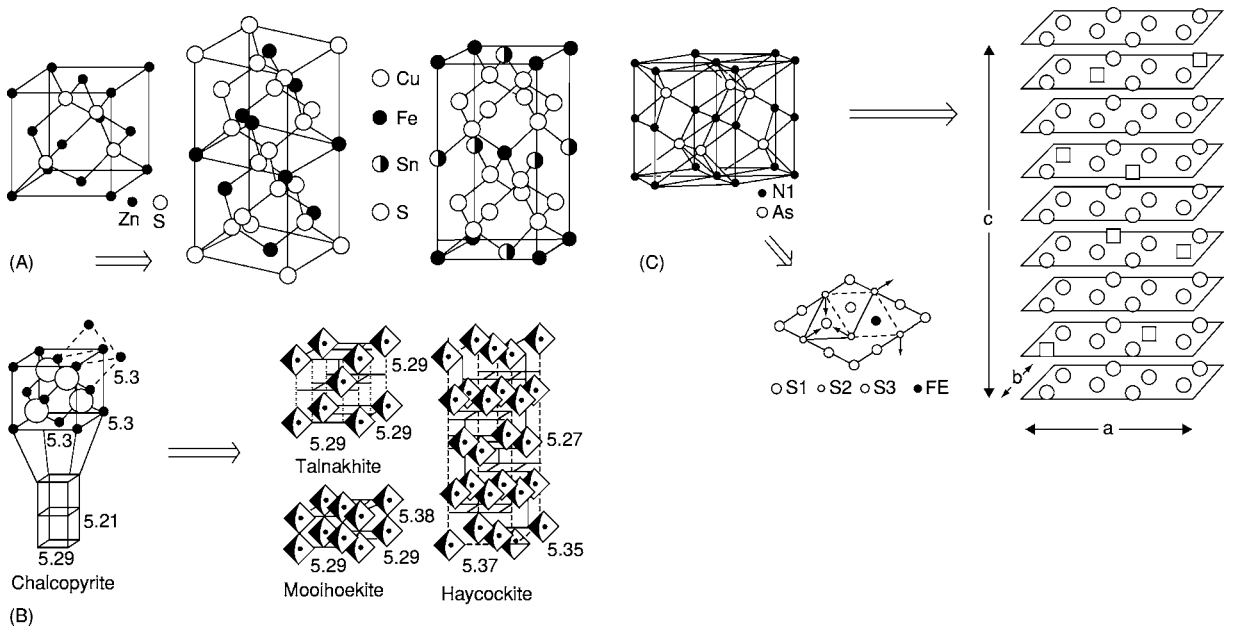


Figure 2 Parent and derivative crystal structures in the sulphide minerals. (A) The sphalerite (ZnS) structure with the chalcopyrite ($CuFeS_2$) and stannite (Cu_2FeSnS_4) structures. (B) The sphalerite and chalcopyrite unit cells with octahedra of metals outlined, within which may be an additional metal ion in the minerals talnakhite ($Cu_9Fe_8S_{16}$), mooihoekite ($Cu_9Fe_9S_{16}$), and haycockite ($Cu_4Fe_5S_8$); the arrangement of additional occupied metal sites is as shown (also shown are the dimensions of the parent sphalerite cell in (A)). (C) The niccolite unit cell of high temperature FeS, which has vacancies in place of iron atoms in monoclinic pyrrhotite (Fe_7S_8) that are ordered as shown, with vacancies represented by squares (and only iron atom layers shown); also shown in a projection onto the basal plane are the distortions that occur in the troilite modification of FeS. Adapted from Institute of Materials, Minerals and Mining, Sulphite Deposits their origin and Processing, ed. P. Gray, 1990.

Table 2 Magnetic and electrical properties of some major sulphide minerals

Mineral species	*Magnetic properties*	*Electrical properties*
Sphalerite (ZnS)	diamagnetic	insulator (E_g~3.7 eV)
'Iron sphalerite' (Zn,Fe)S	paramagnetic	semiconductor (E_g~0.5 eV for *ca.* 12 at.% Fe)
Galena (PbS)	diamagnetic	semiconductor (n and p type, E_g~0.41 eV)
Pyrite (FeS_2)	diamagnetic	semiconductor (n and p type, E_g~0.9 eV)
Cattierite (CoS_2)	ferromagnetic (T_c 110 K)	metallic
Chalcopyrite ($CuFeS_2$)	antiferromagnetic (T_N 823 K)	semiconductor (n type, E_g~0.5 eV)
Covellite (CuS)	diamagnetic(?)	metallic
(monoclinic) Pyrrhotite (Fe_7S_8)	ferrimagnetic (T_c 573 K)	metallic
Carrollite ($CuCo_2S_4$)	Pauli paramagnetic	metallic
Pentlandite $(Ni,Fe)_9S_8$	Pauli paramagnetic	metallic

E_g, band or 'energy' gap; T_c, Curie temperature; T_N, Nèel temperature.

indicates, whereas such non-transition-metal sulphides as sphalerite and galena are diamagnetic; diamagnetism is also exhibited by pyrite and marcasite (the other FeS_2 polymorph). Substitution of iron for zinc in sphalerite leads to paramagnetic behaviour, and many transition-metal sulphides show various forms of magnetic ordering at lower temperatures, including antiferromagnetism (e.g. chalcopyrite), ferromagnetism (e.g. cattierite), and ferrimagnetism (e.g. monoclinic pyrrhotite, although hexagonal pyrrhotites are antiferromagnetic). Other transition-metal sulphides that are metallic conductors exhibit the weak temperature-independent paramagnetism known as Pauli paramagnetism. Metallic conductivity occurs in many sulphides of diverse magnetic character, and, in addition, numerous sulphides are semiconductors and a smaller number (e.g. pure sphalerite) are insulators. Among semiconducting sulphides, both intrinsic and impurity (or extrinsic) conduction mechanisms occur, and conduction both via electrons (n-type) and via holes (p-type) is common. Such variations are often a consequence of the presence of very minor impurities or of

slight nonstoichiometry (e.g. galena (PbS) shows p-type conductivity in lead-deficient samples and n-type conductivity in sulphur-deficient samples).

Given that the great majority of sulphide minerals are opaque, the optical properties of interest are those observed in reflected light from polished sections. The qualitative optical properties commonly serve for routine identification, but definitive characterization can be greatly aided by measuring reflectance at set wavelengths, or a whole series of wavelengths, across the visible-light region.

Sulphide Mineral Stability

Much work has been done to establish the stability relations of sulphide minerals in terms of variables that include temperature, pressure, composition, and the activities of various components.

The Fe–S system is the most important binary sulphide system because the iron sulphides are, by far, the dominant sulphide minerals in terms of geological abundance and variety of geological occurrences. Furthermore, the Fe–S system is an integral part of other important more complex systems such as the Fe–Zn–S, Cu–Fe–S, and Fe–Ni–S systems. The phase relations in the condensed Fe–S system above 400°C (**Figure 3A**) have been the subject of numerous studies and, in contrast to the low-temperature relationships, are well understood. The central part of the system is dominated by the large high-temperature pyrrhotite field of solid solution from stoichiometric FeS towards more sulphur-rich compositions. This high-temperature form with a hexagonal NiAs-type structure accommodates solid solution by random vacancies at the iron sites within the lattice. Hence, the compositions of high-temperature pyrrhotites, except for pure FeS, are best given as $Fe_{1-x}S$, where x is 0 to ~0.14. The maximum thermal stability of the pyrrhotite solid solution at pressures below about 1 atmosphere is 1192°C, above which it melts congruently. Pyrite has, at low pressures, a maximum thermal stability of 742 ± 1°C. The upper thermal stability of pyrite rises by approximately 14°C per kilobar of confining pressure. In spite of numerous studies, the phase relations in the Fe–S system at temperatures below 350°C are incompletely understood. A temperature–composition diagram for the system between FeS and FeS_2 from 0°C to 350°C is shown in **Figure 3B**. Phase relations in the compositional regions Fe–FeS and FeS_2–S remain straightforward, essentially as shown in **Figure 3A**, but the central part of the system is exceedingly complex. This complexity is caused by the crystal chemistry of the pyrrhotites, as discussed above, where iron atoms can be omitted, leaving holes or vacancies. At low temperatures, ordering of vacancies occurs, resulting in the development of various superstructures. The best-known of these superstructures is that of monoclinic pyrrhotite (Fe_7S_8). In this case, the vacancies occur in alternate layers of iron atoms parallel to the basal plane and in alternate rows in those layers (**Figure 2C**). This is known as the 4C structure, because the superstructure has a unit cell that is four times the c dimension of the parent nickel arsenide-type cell.

Although the structure of monoclinic pyrrhotite is well established, the structures of those compositions lying between FeS and Fe_7S_8 remain uncertain. Numerous superstructure types have been reported for these intermediate pyrrhotites, including examples with nonintegral multiples of the parent-cell c and a dimensions (so-called NC and NA types). Many reported superstructures have not been observed in natural samples, where there is evidence of compositions clustering around $Fe_{11}S_{12}$, $Fe_{10}S_{11}$, and Fe_9S_{10}, as well as Fe_7S_8. These compositions also correspond to superstructures with high degrees of order, the 5C (Fe_9S_{10}) and 6C ($Fe_{11}S_{12}$) superstructures. The thermodynamic stability of the various pyrrhotites, particularly the intermediate pyrrhotites, is problematic, with kinetic controls exerting the major influence over the compositions and structure types observed. The highly ordered monoclinic pyrrhotite (Fe_7S_8) and troilite (FeS, a 2C superstructure produced by distortion of the nickel arsenide high-temperature form) may be stable phases in the system, but reported or measured free-energy minima suggest that only pyrite and troilite are truly stable phases.

Much sulphide is formed in sedimentary or hydrothermal environments, so phase equilibria in aqueous systems are of particular interest. The stability of pyrite, for example, has been considered as a function of many variables. Perhaps the best-known relationships are those at 25°C, for which the stability of pyrite in aqueous solution has been examined as a function of the changing activities of two or more components in the system when the others are held constant. Thus, for the system Fe–S–O–H, the activity of hydrogen ions (pH) has been plotted against the 'activity of electrons' (Eh), and either one or both of these variables have been plotted against the activity of sulphur (expressed as pS^{2-} or pS_2). **Figure 4** shows Eh–pH plots illustrating the stability of pyrite and other sulphides and oxides of iron in water at 25°C and 1 atmosphere total pressure. **Figure 4A** shows relationships at a total dissolved sulphur activity of 10^{-1} and illustrates the large stability field of pyrite under reducing conditions over a wide pH range. By contrast, the stability field of pyrrhotite is extremely small. Under more oxidizing conditions, haematite is the dominant iron mineral, with magnetite having

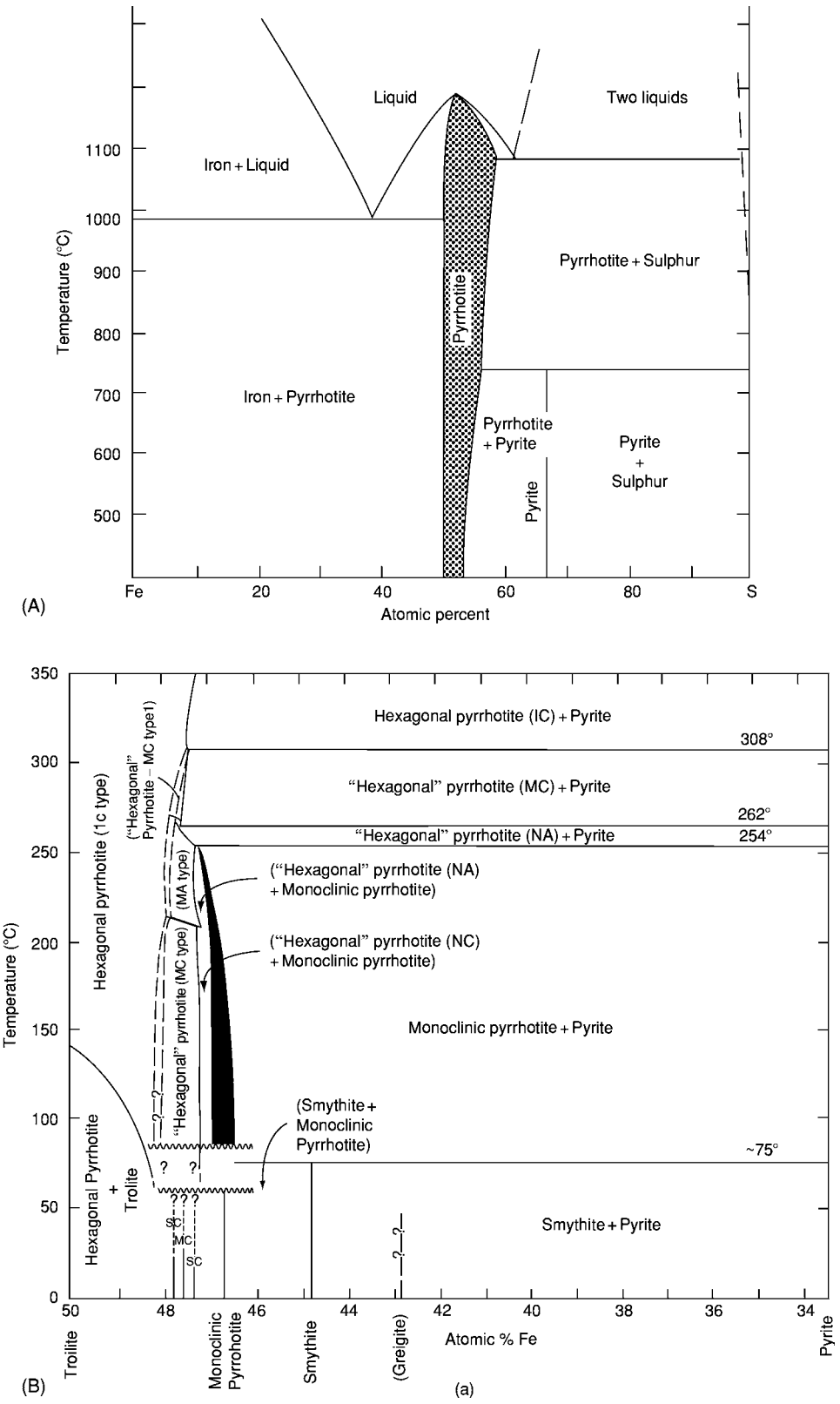


Figure 3 The iron sulphur system: (A) phase relations in the Fe S system above 400°C at low pressure; and (B) phase relations in the central portion of the Fe S system below 350°C. Reproduced with permission from Geochemistry of Hydrothermal Ore Deposits, ed. H. L. Barnes, 1997, John Wiley and Sons.

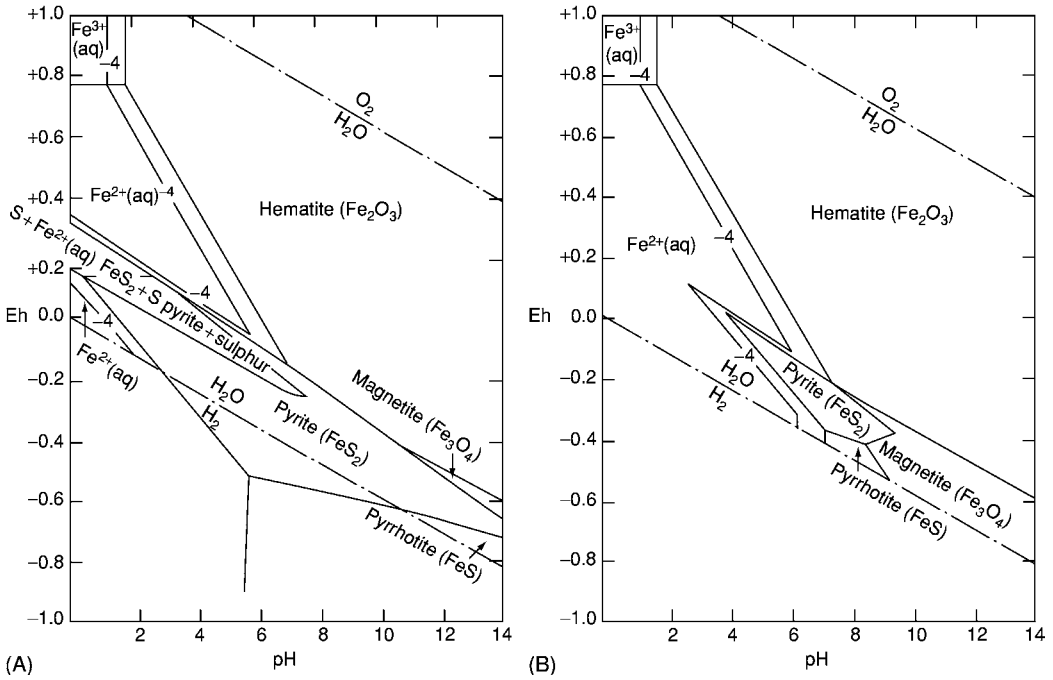


Figure 4 Stability relations of iron sulphides and oxides in water at 25°C and 1 atmosphere total pressure (A) at a total activity of dissolved sulphur of 10^{-1} and (B) at a total activity of dissolved sulphur of 10^{-6}.

only a very small stability field. The effect of lowering the total sulphur activity to 10^{-6} is shown in Figure 4B: the pyrite field is now much smaller and the magnetite field has increased in size. Notable from examination of Figure 4 is the very large stability field of pyrite relative to pyrrhotite in sediments. It can also be seen that quite low sulphide activities can stabilize pyrite over a considerable range of Eh conditions, and that iron oxides or carbonate will form only when the sulphide activity of the environment is very small.

The problem of metastability is implicit in the study of iron sulphides. For example, certain iron sulphides that are common in nature, such as marcasite (FeS_2) and mackinawite ($Fe_{1+x}S$), are not included in the phase diagrams of Figure 3. These sulphides cannot be synthesized by a direct reaction between iron and sulphur in dry systems but only by precipitation from solution. In the natural aqueous systems where such iron sulphides form, reduced sulphur is readily available from the bacterial reduction of sulphate. Here mackinawite is the dominant precursor phase, with the ultimate stable product being pyrite. Details of the reaction pathways involving mackinawite and other less common precursors are still being established, but Figure 5 makes an attempt to summarize known and more speculative phase transformations.

When interpreting the mineral assemblages and textures in order to understand the genesis of sulphide

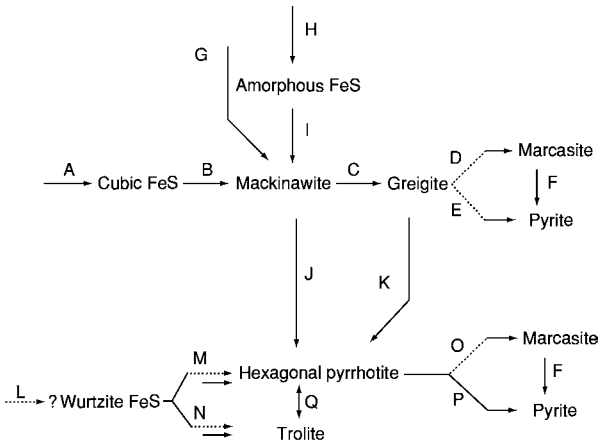


Figure 5 Summary of known and speculative phase transformations and reactions in the low temperature iron sulphur system (redrawn after Lennie and Vaughan, *Geochem. soc. spec pubh. No. 5*, p.128, 1996). The letters on the diagram are as follows: A, formation from Fe^{2+} or Fe and HS or H_2S at a pH of 4 5 ($T < 523$ K); E, inferred sulphidation of greigite at a pH of more than 5 ($T < 523$ K); F, solid state transformation; G, sulphidation of Fe by H_2S at a pH of 3 10 ($T < 523$ K); I, ageing of amorphous precipitate; J, solid state transformation, kinetics not established; L, precipitation from Fe^{2+} by HS at a pH of 4 6 (wide range of temperatures to above 573 K); M,N, reaction of Fe^{2+} and HS or H_2S at a pH of 4 5 (or speculative transformation of proposed wurtzite FeS structure) to form either (M) hexagonal pyrrhotite ($T > 413$ K) or (N) troilite ($T < 413$ K); O, marcasite after pyrrhotite; P, pyrite formed by sulphidation of pyrrhotite (this reaction is rapid above 573 K); Q, troilite hexagonal pyrrhotite reversible phase transition. Reproduced with permission from Geochemis try of Hydrothermal Ore Deposits, ed. H. L. Barnes, 1997, John Wiley and Sons.

ores, temperature–composition relations in the subsolidus region are particularly useful. This is because many sulphide minerals undergo changes in the solid state down to relatively low temperatures. These changes give rise to a wealth of intergrowth textures produced by exsolution or unmixing, as exemplified by the Ni–Fe–S system (Figure 6). Such changes occur because, at elevated temperatures in many sulphide systems, there are extensive fields of solid solution that shrink with falling temperature. In the case of the Ni–Fe–S system, it is the breakdown of the so-called monosulphide solid solution and the segregation of pentlandite according to the reaction

$$(Fe, Ni)S = Fe_{1-x}S + (Fe, Ni)_9S_8$$

monosulphide solid solution = pyrrhotite + pentlandite

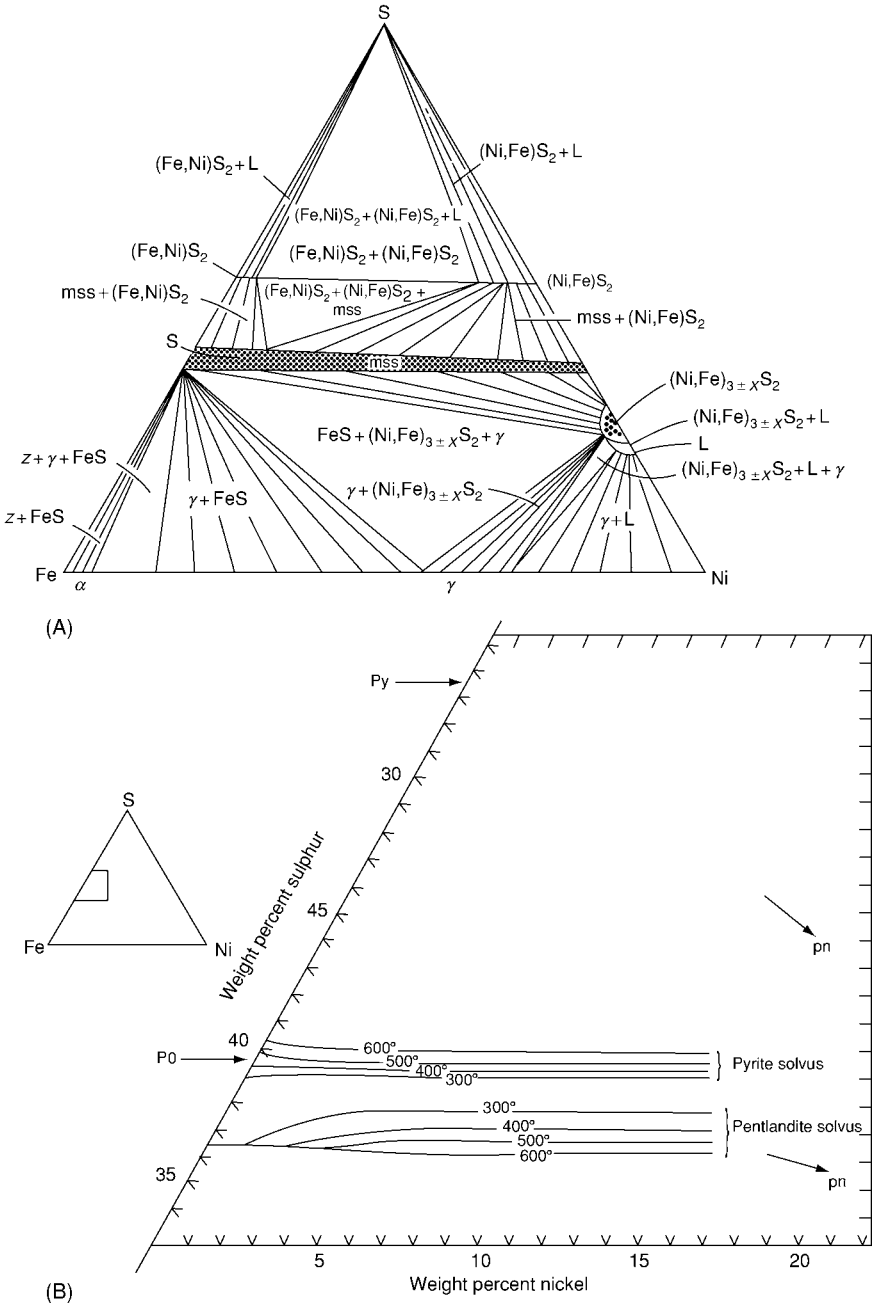


Figure 6 (A) The condensed phase relations in the Fe Ni S system at 650°C. Note the absence of pentlandite and the area of monosulphide solid solution (*mss*) that spans the system from $Fe_{1-x}S$ to $Ni_{1-x}S$. (B) A portion of the Fe Ni S system, showing the compositional limits of the monosulphide solid solution in (A) at 600°C, 500°C, 400°C, and 300°C. Pentlandite flames exsolve from the *mss* as the sulphur poor boundary retreats during cooling. Reproduced with permission from Geochemistry of Hydrothermal Ore Deposits, ed. H. L. Barnes, 1997, John Wiley and Sons.

that is central to understanding the assemblages and textures in the sulphide nickel ores. Such processes may lead to completely separate grains of the two phases being formed on exsolution, or the two phases may be intergrown as laths, blebs, etc., often with a clearly defined crystallographic relationship. The crystal structures of the two phases will partly dictate the kind of texture that forms, but other crucial factors are the monosulphide solid solution bulk starting composition and the cooling history. The majority of binary and ternary and many quaternary sulphide systems have now been studied as regards temperature–composition relations, yielding evidence of the maximum stability temperatures of phases, solid solution limits, and coexistence of phases under the equilibrium conditions pertaining in such experiments.

The stable phase relations may also be defined in terms of the activities of components and, in regard to the sulphide minerals, it is the activity of sulphur (aS_2) that is most important. Thus, a 'petrogenetic grid' commonly used to present sulphide mineral relationships plots aS_2 against temperature (actually log aS_2 versus 1/temperature because sulphidation boundaries then plot as straight lines in many cases) as shown in Figure 7. Such diagrams show the progressive development of assemblages through sulphidation from metal to monosulphide to (where appropriate) disulphide.

A limited number of sulphide systems have proved valuable in geothermometry and geobarometry. Notably useful in geothermometry is the mineral arsenopyrite (FeAsS). Phase relations in the Fe–As–S system are shown schematically over the temperature range 688°C to 363°C in Figure 8. In the higher temperature range, arsenopyrite coexists with pyrrhotite, loellingite, arsenic, or an (As,S) liquid (represented on cooling by an arsenic sulphide) in the assemblages labelled 1 to 4. From 491°C to 363°C, arsenopyrite can coexist with pyrite, pyrrhotite, loellingite, arsenic, or (As,S) liquid in the assemblages labelled 5 to 8. Because arsenopyrite composition is a function of both temperature and the equilibrium with the coexisting phases in the system, it can be used in geothermometry only in buffered assemblages, i.e. those labelled 1 to 8 in Figure 8. For these assemblages, the composition of arsenopyrite has been calibrated as a function of temperature to provide a geothermometer. A method has also been devised of using the d_{131} X-ray spacing curve to determine arsenopyrite composition

$$(\text{As, at.}\%) = 866.67d_{131} - 1381.12.$$

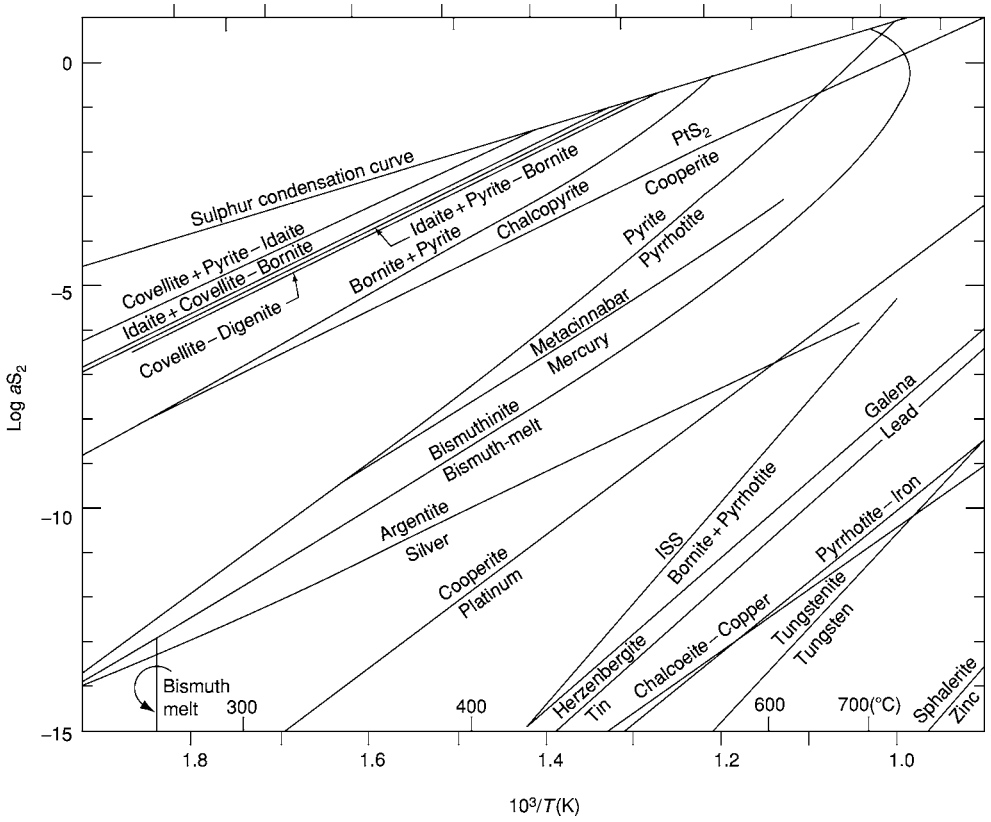


Figure 7 Sulphidation curves for various sulphides as a function of temperature. ISS intermediate solid solution.

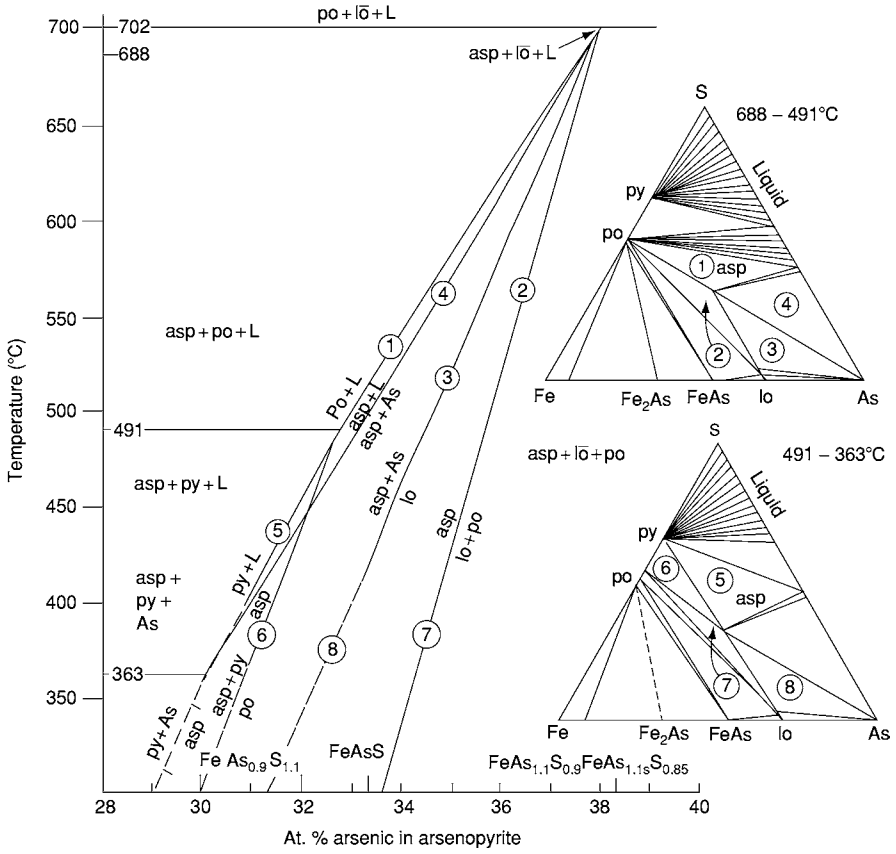


Figure 8 Pseudobinary temperature composition plot, showing arsenopyrite composition as a function of temperature and equilibrium mineral assemblage. The assemblages numbered 1 8 in the Fe As S phase diagrams on the right correspond to the labelled curves in the left hand diagram. Abbreviations: asp, arsenopyrite; py, pyrite; po, pyrrhotite; lo, lollingite; L, liquid.

Critical evaluation of the arsenopyrite geothermometer, using volume calculations, has revealed that pressure will shift the sulphidation buffer curves in Figure 8 for the system Fe–As–S to higher sulphur activities (e.g. by up to 10 times for arsenopyrite coexisting with pyrite). The effect of pressure on the composition of arsenopyrite buffered by two other phases is not large and can be ignored for low-pressure hydrothermal deposits but needs to be taken into account in metamorphosed deposits.

The sphalerite geobarometer is based on the observation that increasing pressure reduces the FeS content of a sphalerite equilibrated with pyrite and hexagonal pyrrhotite in a manner that is independent of temperature (Figure 9). Sphalerite is sufficiently refractory to preserve the original composition unless it remains in contact with pyrrhotite or chalcopyrite: low-temperature re-equilibration of sphalerite generally reduces the original FeS content and results in pressure estimations that are too high.

A very important aspect of sulphide geochemistry and mineral chemistry is concerned with the mechanisms and kinetics of sulphide oxidation, particularly in regard to pyrite, the overwhelmingly dominant sulphide in most natural systems. This is because of the applications of such studies in both mineral-processing technology and environmental chemistry (particularly in relation to the generation of acid waters). Despite a large amount of work, reflected by a substantial literature, many aspects of pyrite oxidation remain unresolved. This is because, in addition to the large number of variables (whether oxidation is in air, water, or other aqueous fluid; pH; Eh; temperature; presence of various oxidants; role of bacteria), the oxidation of pyrite to form sulphate requires the transfer of eight electrons per sulphur atom and must involve several steps, with sulphur species of intermediate oxidation states. The literature almost universally acknowledges that any oxidation reaction mechanism is surface-reaction controlled. It is suggested that, in aqueous solution, pyrite oxidizes by a combination of the half-reactions

$$FeS_2 + 8H_2O = Fe^{3+} + 2SO_4^{2} \quad + 16H^{+} + 15e$$

and

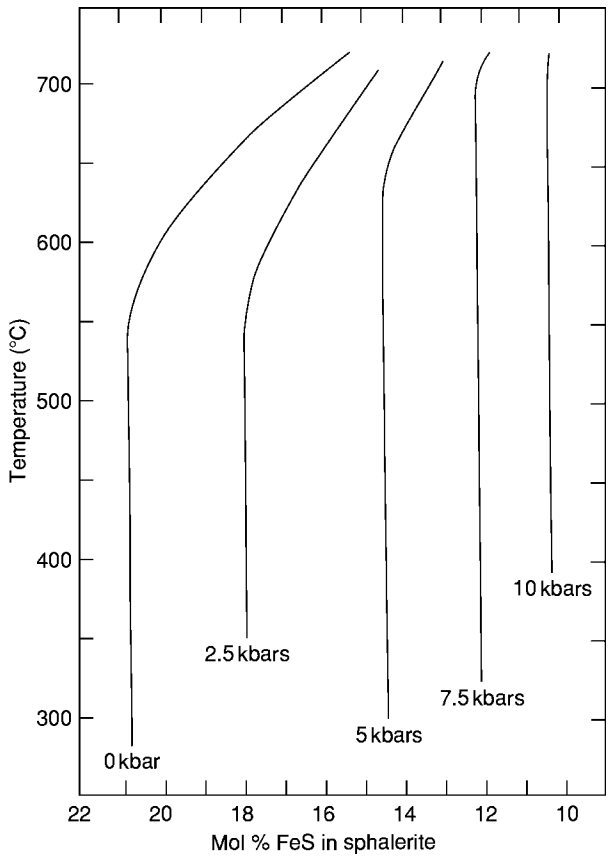


Figure 9 Plots of the FeS contents of sphalerite coexisting with pyrite and hexagonal pyrrhotite at 0, 2.5, 5, 7.5, and 10 kbar at temperatures from 300°C to 700°C. Reproduced with permission from Geochemistry of Hydrothermal Ore Deposits, ed. H. L. Barnes, 1997, John Wiley and Sons.

$$FeS_2 = Fe^{3+} + 2S + 3e \quad ,$$

for which the Eh and pH determine the dominant pathway: the sulphate route dominates under ambient conditions. At low pH, pyrite may be oxidized by O_2 and Fe^{3+}_{aq}:

$$FeS_2 + 7/2O_2 + H_2O = Fe^{2+} + 2SO_4^{2} + 2H^+$$

and

$$FeS_2 + 14Fe^{3+} + 8H_2O = 15Fe^{2+} + 2SO_4^{2} + 16H^+$$

The rates of pyrite oxidation are dramatically affected by the presence of bacteria; for example, at low pH, *Thiobacillus ferrooxidans* can accelerate the oxidation of Fe^{2+} by a factor of more than 10^6. Recent work on the mechanism of pyrite oxidation emphasizes the importance of the fact that pyrite is a semiconductor. Electrons can be transferred from sulphur atoms at an anodic site, where oxygen atoms from water molecules attach to sulphur atoms to form sulphoxy species, through the crystal to cathodic Fe(II) sites, where they are acquired by the oxidant species (usually O_2 or Fe^{3+}). Bulk measurements of oxidation rate show that the concentration of O_2 or Fe^{3+} is critical, so it appears that the 'cathodic' reaction is the rate-determining step.

The study of the surface structures and reactivities of sulphide minerals is being revolutionized by the application of the scanning probe microscopy techniques of scanning tunnelling microscopy and atomic force microscopy. These are leading to models of surface structures and reactions that have been developed from direct observations of the systems at the atomic scale.

Sulphide Paragenesis: Rocks, Sediments, and Ore Deposits

The major ore deposit types that contain substantial amounts of sulphide minerals are shown in Table 3 along with the major minerals they contain, the metals extracted from them, and examples of specific deposits.

Pyrite is exceptionally widespread in its occurrence, being known as an accessory mineral in many of the major rock types that make up the crust of the Earth and as a major or minor phase in many types of ore deposit. However, in ultramafic and mafic intrusive rocks, especially in the so-called 'sulphide nickel' deposits, it is pyrrhotite that dominates. The sulphides in these deposits (pyrrhotite, pentlandite, chalcopyrite, and, in some cases, pyrite) are regarded as having formed by the crystallization of an immiscible sulphide melt that separated from the main silicate melt. Studies of the sulphides in the best-known of all layered basic intrusions, the Bushveld Complex, also show a dominance of pyrrhotite, pentlandite, and chalcopyrite along with minor but economically vital platinum group minerals. In the classic Skaergaard intrusion (Greenland), pyrite is present chiefly in small amounts in the latest rock fraction, the acid granophyre. In the Stillwater Complex, the principal sulphides include pyrite, chalcopyrite, and pyrrhotite, which are fairly conspicuous near the floor of the complex and at one or two higher horizons.

In felsic igneous rocks, the predominant sulphide is pyrite. In addition to its occurrence as an accessory phase in a wide range of such rocks, pyrite is the dominant sulphide mineral in porphyry copper deposits, in which chalcopyrite is the most important ore mineral, along with various binary copper sulphides. Anhedral to euhedral grains of sulphides occur in these ores, where the mineralization is as disseminated grains or veinlets in host intrusions, which range in composition from quartz diorite to

Table 3 The major types of sulphide ore deposit

Type	*Major minerals*	*Metals extracted*	*Examples*
Ores related to mafic and ultramafic intrusions			
Sudbury nickel copper	po, pn, py, cpy, v	Ni, Cu, Co, PGM	Sudbury, Ontario
Merensky reef platinum	po, pn, cpy	Ni, Cu, PGM	Merensky Reef, South Africa
Ores related to felsic intrusive rocks			
Tin and tungsten skarns	py, cass, sph, cpy, wolf	Sn, W	Pine Creek, California, USA
Zinc lead skarns	py, sph, gn	Zn, Pb	Ban Ban, Australia
Copper skarns	py, cpy	Cu, Au	Carr Fork, Utah, USA
Porphyry copper/molybdenum	py, cpy, bn, mbd	Cu, Mo, Au	Bingham Canyon, Utah, USA Climax, Colorado, USA
Polymetallic veins	py, cpy, gn, sph, ttd	Cu, Pb, Zn, Ag	Carnsell River, NWT
Ores related to marine mafic extrusive rocks			
Cyprus type massive sulphides	py, cpy	Cu	Cyprus
Besshi type massive sulphides	py, cpy, sph, gn	Cu, Pb, Zn	Japan
Ores related to subaerial felsic to mafic extrusive rocks			
Creede type epithermal veins	py, sph, gn, cpy, ttd, asp	Cu, Pb, Zn, Ag, Au	Creede, Colorado, USA
Almaden mercury type	py, cinn	Hg	Almaden, Spain
Ores related to marine felsic to mafic extrusive rocks			
Kuroko type	py, cpy, gn, sph, ttd, asp,	Cu, Pb, Zn, Ag, Au	Japan
Ores in clastic sedimentary rocks			
Quartz pebble conglomerate gold uranium	py, uran, Au	Au, U	Witwatersrand, South Africa
Sandstone hosted lead zinc	py, sph, gn	Zn, Pb, Cd	Laisvall, Sweden
Sedimentary exhalative	py, sph, gn, cpy, asp	Cu, Pb, Zn, Au, Ag	Sullivan, British Columbia, Canada
lead zinc (Sedex)	ttd, po		Tynagh, Ireland
Ores in carbonate rocks			
Mississippi Valley type	py, gn, sph	Zn, Pb, Cd, Ga, Ge	South eastern Missouri

Abbreviation: po, pyrrhotite; pn, pentlandite; py, pyrite; cpy, chalcopyrite; v, violarite; cass, cassiterite; sph, sphalerite; wolf, wolframite; gn, galena; bn, bornite; mbd, molybdenite; ttd, tetrahedrite; asp, arsenopyrite; cinn, cinnabar; uran, uraninite; PGM, platinum group metals.

quartz monzonite. Large masses of pyrite, sphalerite, galena, or chalcopyrite are also found in ore deposits formed by contact metamorphism (skarn deposits). Pyrite occurs, often as a major phase, in the majority of hydrothermal vein deposits and dominates the ore mineralogy of those deposits that might broadly be described as 'volcanogenic'. These include the ores that occur in thick volcanic sequences, such as the Kuroko deposits of Japan. In the so-called 'black ore' of Kuroko-type deposits, irregular masses of galena are intergrown on a fine scale with sphalerite, pyrite, and chalcopyrite. In the Besshi-type deposits of Japan, the sulphides occur in dominantly sedimentary sequences and are again dominated by pyrite, chalcopyrite, sphalerite, and galena.

In sulphide ores in volcanosedimentary sequences that are associated with ophiolite complexes, as in the Troodos Complex deposits (Cyprus), pyrite and chalcopyrite dominate. An understanding of the genesis of these deposits has rapidly developed in recent years through the observation of present-day volcanic and hydrothermal activity on the seafloor. The disseminated to massive stratiform sulphide ores that occur, often conformably, within sedimentary sequences grade into the volcanogenic deposits discussed above. Again, pyrite is the dominant ore mineral in these deposits, which include the Kupferschiefer-Marl Slate of northern Europe and the Copperbelt of Zambia and Zaire. The latter contains not only a range of copper and copper–iron sulphides (chalcopyrite, bornite, chalcocite, covellite) but also significant amounts of cobalt in pyrite and the sulphospinel mineral carrollite.

In the context of sedimentary rocks, sulphides (galena and sphalerite with some pyrite) are major phases in many of the lead–zinc–barite–fluorite ores, which occur chiefly in limestones, and pyrite, along with the copper sulphides, occurs in the uranium–vanadium–copper ores that are associated with sandstones in areas such as the Colorado Plateau, USA. In the gold–uranium ores that occur in conglomerates in the Witwatersrand, South Africa, and Elliot Lake, Ontario, Canada, pyrite is the major opaque phase. As well as occurring as common accessory minerals in black shales, sulphides, especially pyrite and marcasite, are the major opaque mineral in coals. The formation of sulphides in all of these environments is probably linked to the reaction of metals, particularly iron, released by the dissolution of detrital oxides and silicates, with sulphur produced by the bacterial reduction of sulphate present in interstitial waters. This is observed in Holocene sediments, with pyrite (preceded by mackinawite and possibly greigite) forming through a diagenetic process occurring in the reduced

zone beneath the sediment surface. The occurrence of marcasite is much more restricted than that of pyrite. It is never found as a primary phase in any magmatic association and is indicative of environments that have probably been low temperature and where an aqueous phase was present at the time of formation.

Many studies have been undertaken of sulphur isotope ratios ($^{32}S : {}^{34}S$) in sulphide minerals from rocks and ore deposits. Initially these studies were aimed at developing criteria to distinguish ore deposits of igneous–hydrothermal origin from those of sedimentary syngenetic origin. Since bacterial reduction of sulphate substantially enriches the involved sulphur in ^{32}S (by 50% or more in the case of H_2S), such enrichment is taken as an indication that the sulphur has passed through the sedimentary cycle. This key reaction and others that involve isotope exchange and occur during ore formation have been the basis for the application of sulphur isotope studies to genetic problems. However, studies of both synthetic systems and natural ores have shown that the isotopic composition of sulphur is also controlled by changes in pH and oxygen fugacity in the reacting system. When these can be determined, for example by using carbon isotope ratios, valuable information about the formation of a deposit may be obtained. Coexisting pairs of sulphide minerals, such as pyrite co-existing with galena, sphalerite, chalcopyrite, or pyrrhotite, have also been used for sulphur isotope geothermometry. The fractionation of sulphur isotopes between co-existing mineral pairs has been calibrated as a function of temperature for use in such geothermometric studies.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray); Mineral Analysis. **Mining Geology:** Hydrothermal Ores. **Sedimentary Environments:** Anoxic Environments. **Tectonics:** Hydrothermal Activity; Hydrothermal Vents At Mid-Ocean Ridges. **Thermal Metamorphism**.

Further Reading

Barnes HL (ed.) (1979) *Geochemistry of Hydrothermal Ore Deposits*, 2nd edn. New York: Wiley Interscience.

Barnes HL (ed.) (1997) *Geochemistry of Hydrothermal Ore Deposits*, 3rd edn. New York: Wiley Interscience.

Cabri LJ and Vaughan DJ (1998) *Modern Approaches to Ore and Environmental Mineralogy*. Short Course 27. Ottawa: Mineralogical Association of Canada.

Cotter Howells J, Campbell L, Batchelder M, and Valsami Jones E (eds.) (2000) *Environmental Mineralogy: Microbial Interactions, Anthropogenic Influences, Contaminated Land and Waste Management*. Monograph 9. London: Mineralogical Society of Great Britain and Ireland.

Craig JR and Vaughan DJ (1994) *Ore Microscopy and Ore Petrography*, 2nd edn. New York: Wiley Interscience.

Garrels RM and Christ CL (1965) *Solutions, Minerals and Equilibria*. New York: Harper and Row.

Jambor JL and Blowes DW (eds.) (1994) *The Environmental Geochemistry of Sulfide Minewastes*. Short Course Handbook 22. Ottawa: Mineralogical Association of Canada.

Kostov I and Micheeva Stefanova J (1981) *Sulphide Minerals: Crystal Chemistry, Parageneses and Systematics*. Sofia: Bulgarian Academy of Sciences.

Mills KC (1974) *Thermodynamic Data for Inorganic Sulphides, Selenides and Tellurides*. London: Butterworth.

Ramdohr P (1980) *The Ore Minerals and their Intergrowths*, 2nd English edn. Oxford: Pergamon.

Ribbe PH (ed.) (1974) *Sulfide Mineralogy*. Short Course Notes Volume 1. Washington, DC: Mineralogical Society of America.

Russell MJ and Hall A (1997) The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *Journal of the Geological Society of London* 154: 377 402.

Shuey RT (1975) *Semiconducting Ore Minerals*. Developments in Economic Geology 4. Amsterdam: Elsevier.

Stanton RL (1972) *Ore Petrology*. New York: McGraw Hill.

Vaughan DJ and Craig JR (1978) *Mineral Chemistry of Metal Sulfides*. Cambridge: Cambridge University Press.

Vaughan DJ and Lennie AR (1991) The iron sulphide minerals: their chemistry and role in nature. *Science Progress* 75: 371 388.

Tungstates

P A Williams, University of Western Sydney, Penrith South DC, NSW, Australia

Introduction

Tungstates are compounds containing negatively charged oxytungstate ions. In the mineral kingdom, the simple tetraoxotungstate(VI), or tungstate ion, WO_4^{2-}, is present. As with the case of molybdate(VI), tungstate ions can polymerize under acid conditions to produce polytungstate and heteropolytungstate species, but minerals containing such complex anions are exceedingly rare and remain poorly understood and characterized. Only a few minerals containing essential tungstate are known, though some of these have great commercial significance. The standard potential at 25°C is −1.26 V for

eqn [1], and this explains the fact that tungstates with W in its highest oxidation state (+6) dominate the mineralogy of the element in both the primary and the secondary environments.

$$WO_4^{2} + 2H_2O + 2e \rightarrow WO_2 + 4OH \qquad [1]$$

Tungstenite, WS_2, is an extremely rare species, isomorphous with its common congener molybdenite, MoS_2. Table 1 provides a list of known naturally occurring tungstates. The so-called tungstic acids of stoichiometry $WO_3 \cdot nH_2O$ are not listed because they are simple lattice compounds. However, due to possible confusion as to the attribution of certain species to the tungstate class, Table 1 does include certain species that are in fact related. For example, alumotungstite and ferritungstite are derivatives of WO_3 with the pyrochlore structure. Substitution of W by Al or Fe gives a positive charge discrepancy that is compensated by incorporation of other cations in a vacant lattice site. In related fashion, raspite, $PbWO_4$, the dimorph of stolzite, contains chains of edge-linked WO_6 octahedra. Pinalite, russellite, and tungstibite are complex layer structure oxides; cerotungstite-(Ce) and yttrotungstite-(Y) are complex secondary oxyhydroxide species containing chains of WO_6 octahedra. The same situation may obtain for anthoinite and mpororoite, two poorly characterized minerals that require further investigation.

Table 1 Tungstate(VI) minerals

Mineral	*Chemical composition*
Simple tungstates	
Ferberite	$MnWO_4$
Hubnerite	$FeWO_4$
Paraniite (Y)	$Ca_2Y(AsO_4)(WO_4)_2$
Raspite	$PbWO_4$
Sanmartinite	$ZnWO_4$
Scheelite	$CaWO_4$
Stolzite	$PbWO_4$
Basic double salts	
Anthoinite	$WAl(O,OH)_3(?)$
Cuprotungstite	$Cu_3(WO_4)_2(OH)_2$
Mpororoite	$WAlO_3(OH)_3\ 2H_2O(?)$
Pinalite	$Pb_3WO_5Cl_2$
Russellite	Bi_2WO_6
Tungstibite	Sb_2WO_6
Complex uranium salts	
Uranotungstite	$(Fe,Ba,Pb)(UO_2)_2(WO_4)(OH)_4\ 12H_2O$
Polytungstates	
Phyllotungstite	$(Ca,Pb)Fe_3H(WO_4)_6\ 10H_2O$
Rankachite	$CaFeV_4W_8O_{36}\ 12H_2O$
Other complex species	
Cerotungstite (Ce)	$CeW_2O_6(OH)_3$
Ferritungstite	$(W,Fe)(O,OH)_3$[a]
Alumotungstite	$(W,Al)(O,OH)_3$[a]
Yttrotungstite (Y)	$YW_2O_6(OH)_3$

[a]Other cations, including Ca, Na, and Pb, may substitute in a site vacancy to compensate for charge imbalance (see text).

Primary Tungstates

All of the primary tungstates, including ferberite, hübnerite, paraniite-(Y), sanmartinite, and scheelite, contain the simple WO_4^{2} ion. Paraniite-(Y) and sanmartinite are both extremely rare minerals, but other members of the group have great economic importance, constituting the only minerals of tungsten to have value commercially. Scheelite takes up molybdenum in the lattice and a complete solid solution extends to the isomorphous mineral powellite, $CaMoO_4$. Primary scheelite, however, does not usually contain much molybdenum. The term 'wolframite' was formerly applied to members of the ferberite–hübnerite series and the nomenclature is still commonly in use. Solid solution in the series is in fact limited to about 20 mol% in each end-member and the separate end-member terms are now applied to individual specimens, depending on the exact composition.

Economically valuable deposits of the Alpine Cleft or sedimentary types of scheelite are well documented, but the main settings of scheelite, ferberite, and hübnerite mineralization are either with acid-intrusive rocks or associated skarns. Frequent associates of the tungstates are cassiterite, SnO_2, molybdenite, MoS_2, base metals such as copper, and minor amounts of gold. Important mines were once located in all continents, but at present deposits in China, Indochina, and Brazil are the main sources of tungsten.

Secondary Tungstates

Scheelite is also well known as a secondary mineral, especially associated with the oxidation of ferberite and hübnerite. In this setting, it also may incorporate molybdate in its lattice. Other secondary tungstates are rare, with perhaps the most frequently encountered species being stolzite. Notable locations of stolzite include the Clara mine and other mines in the Black Forest in Germany, and at Broken Hill and the Cordillera mine in New South Wales, Australia. The Clara, Broken Hill, and Cordillera mines have produced raspite as a mineralogical curiosity and the latter is renowned for the association of the dimorphs with cuprotungstite.

As with molybdate (*see* **Rocks and Their Classification**), polymerization of tungstate in acid solutions yields polymeric species. These are represented in the mineral kingdom by phyllotungstite and rankachite. Both are much rarer than are their molybdenum-bearing congeners and are of academic interest

alone. None of the secondary tungstates represents an economically important source of tungsten.

See Also

Minerals: Definition and Classification; Chromates; Molybdates. **Rocks and Their Classification**.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2003) *Handbook of Mineralogy. Volume 5. Borates, Carbonates, Sulfates, Chromates, Germanates, Iodates, Molybdates, Tungstates, etc., and Organic Materials*. Tucson, AZ: Mineral Data Publishing.

Baes CF Jr and Mesmer RE (1986) *The Hydrolysis of Cations*. Malabar, FL: Krieger Publishing Company.

Bard AJ, Parsons R, and Jordan J (1985) *Standard Potentials in Aqueous Solution*. New York: Marcel Dekker.

Birch WD (ed.) (1999) *Minerals of Broken Hill*. Broken Hill, NSW: Broken Hill City Council.

Cox DP and Singer DA (1986) *Mineral Deposit Models. United States Geological Survey Bulletin 1693*. Washington, DC: United States Geological Survey.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, 8th edn. London: Wiley Europe.

Guilbert JM and Park CF Jr (1986) *The Geology of Ore Deposits*. New York: WH Freeman.

Mandarino JA (1999) *Fleischer's Glossary of Mineral Species 1999*, 8th edn. Tucson, AZ: Mineralogical Record Inc.

Roberts WL, Campbell TJ, and Rapp GR Jr (1990) *Encyclopedia of Minerals*, 2nd edn. New York: Van Nostrand Reinhold.

Sahama TG (1981) *The Secondary Tungsten Minerals, a Review. The Mineralogical Record*, vol. 11. pp. 81 87. Tucson, AZ: Mineralogical Record Inc.

Williams PA (1990) *Oxide Zone Geochemistry*. Chichester: Ellis Horwood.

Vanadates

P A Williams, University of Western Sydney, Sydney, Australia

Naturally occurring vanadate(V) minerals are described here, together with the environments in which they are found. Attention is drawn to the existence of polyvanadates as well as to normal vanadates. In the latter, extensive solid solution involving arsenate and phosphate is evident. The chemistry of polymerisation of vanadate in acidic solution is well understood and is reflected in many examples in the mineral kingdom. Depending upon the prevailing redox potential in the solutions from which such species crystallise, highly coloured, mixed V(IV, V) species are formed, the so-called vanadium bronzes. Significant amounts of vanadium are recovered from secondary vanadate minerals, but much is derived from the treatment of vanadium-rich fly ash and the mining of primary oxide minerals that carry minor amounts of the element.

Vanadates are compounds containing negatively charged oxyvanadium ions and are extremely widespread in the natural environment in small amounts. A list of them is given in Table 1. Nearly all of the vanadates are found in the oxidised zones of base metal ore bodies or in other supergene environments, such as drainage channels and oxidised sedimentary rocks. A few, including wakefieldite-(Y), clinobisvanite, and pucherite occur as accessory minerals in granite pegmatites, while others such as leningradite, averierite, and stoiberite are volcanic sublimates. Many vanadate minerals contain the simple tetraoxovanadate(V) or vanadate ion, VO_4^3 . The arrangement of such minerals under the first heading in Table 1, however, obscures the fact that in many cases complex solid solution phenomena attend naturally occurring examples. Some of the formulae indicate solid solution series where these are found to be significant for all known examples of particular phases. Nevertheless, all of the simple vanadates incorporate other ions in their lattices. With cations, this phenomenon is common and several series are known, including the descloizite-mottramite and mounanite-krettnichite pairs, for example. Solid solution involving anionic substitution is ubiquitous. Phosphate and arsenate frequently substitute for vanadate in these minerals, albeit to different extents. The vanadate ion, VO_4^3 , is somewhat larger than phosphate and arsenate ions, whose thermochemical radii are 230 and 237 picometres, respectively. Other

Table 1 Vanadate minerals

Normal vanadate, oxyvanadate and chlorovanadate(V) minerals	
Averievite	$Cu_5O_2(VO_4)_2$ n(Cu,Cs,K)Cl ($n \approx 1.2$)
Brackebuschite	$Pb_2(Mn,Fe,Zn)(VO_4)_2(OH,H_2O)$
Čechite	$Pb(Fe,Mn)VO_4(OH)$
Cheremnykhite	$Pb_3Zn_3O_2(TeO_4)(VO_4)_2$
Clinobisvanite	$BiVO_4$
Descloizite	$Pb(Zn,Cu)VO_4(OH)$
Dreyerite	$BiVO_4$
Duhamelite	$Pb_2Cu_4Bi(VO_4)_4(OH)_3$ $8H_2O$
Fingerite	$Cu_{11}O_2(VO_4)_6$
Gamagarite	$Ba_2(Fe,Mn)(VO_4)_2(OH)$
Hechtsbergite	$Bi_2(VO_4)(OH)$
Heyite	$Pb_5Fe_2O_4(VO_4)_2$
Howardevansite	$NaCu(Fe,Al,Mn)_2(VO_4)_3$
Kolovratite	$Ni_xZn_y(VO_4)_z$ nH_2O
Kombatite	$Pb_{14}O_9(VO_4)_2Cl_4$
Krettnichite	$PbMn_2(VO_4)_2(OH)_2$
Leningradite	$PbCu_3(VO_4)_2Cl_2$
Lyonsite	$Cu_3Fe_4(VO_4)_6$
Mcbirneyite	$Cu_3(VO_4)_2$
Mottramite	$Pb(Cu,Zn)VO_4(OH)$
Mounanaite	$PbFe_2(VO_4)_2(OH)_2$
Namibite	$Cu(BiO)_2(VO_4)(OH)$
Palenzonaite	$NaCa_2Mn_2[(V,As,Si)O_4]_3$
Pottsite	$HPbBi(VO_4)_2$ $2H_2O$
Pucherite	$BiVO_4$
Pyrobelonite	$PbMnVO_4(OH)$
Reppiaite	$Mn_5(VO_4)_2(OH)_4$
Rusakovite	$(Fe,Al)_5[(V,P)O_4]_2(OH)_9$ $3H_2O$
Santafeite	$(Na,Ca,Sr)_3(Mn,Fe)_2Mn_2$ $(VO_4)_4(OH,O)_5$ $2H_2O$
Schumacherite	$Bi_3[(V,As,P)O_4]_2(OH)$
Stoiberite	$Cu_5O_2(VO_4)_2$
Tangeite	$CaCuVO_4(OH)$
Turanite	$Cu_5(VO_4)_2(OH)_4$
Vanadinite	$Pb_5(VO_4)_3Cl$
Vésigniéite	$BaCu_3(VO_4)_2(OH)_2$
Wakefieldite (Ce)	$(Ce,La,Nd,Y,Pr,Sm)[(V,As)O_4]$
Wakefieldite (Y)	YVO_4
Layered uranyl vanadate(V) minerals	
Carnotite	$K_2(UO_2)_2(VO_4)_2$ $3H_2O$
Curienite	$Pb(UO_2)_2(VO_4)_2$ $5H_2O$
Francevillite	$(Ba,Pb)(UO_2)(VO_4)_2$ $5H_2O$
Margaritasite	$(Cs,K,H)_2(UO_2)_2(VO_4)_2$ H_2O
Metatyuyamunite	$Ca(UO_2)_2(VO_4)_2$ $4H_2O$
Metavanuralite	$Al(UO_2)_2(VO_4)_2(OH)$ $8H_2O$
Sengierite	$Cu(UO_2)(VO_4)(OH)$ $3H_2O$
Strelkinite	$Na_2(UO_2)_2(VO_4)_2$ $6H_2O$
Tyuyamunite	$Ca(UO_2)_2(VO_4)_2$ $8H_2O$
Vanuralite	$Al(UO_2)_2(VO_4)_2(OH)$ $11H_2O$
Vanuranylite	$(H_3O,Ba,Ca\ K)_2(UO_2)_2(VO_4)_2$ $4H_2O(?)$
Pyrovanadate(V) minerals	
Blossite	$Cu_2(V_2O_7)$
Chervetite	$Pb_2(V_2O_7)$
Fianelite	$Mn_2(V_2O_7)$ $2H_2O$
Volborthite	$Cu_3(V_2O_7)(OH)_2$ $2H_2O$
Ziesite	$Cu_2(V_2O_7)$
Minerals containing the decavanadate(V) isopolyanion	
Hummerite	$K_2Mg_2(V_{10}O_{28})$ $16H_2O$
Huemulite	$Na_4Mg(V_{10}O_{28})$ $24H_2O$
Pascoite	$Ca_3(V_{10}O_{28})$ $17H_2O$
Rauvite	$Ca(UO_2)_2(V_{10}O_{28})$ $16H_2O$
Metavanadate(V) minerals[a]	
Alvanite	$(Zn,Ni)Al_4(VO_3)_2(OH)_{12}$ $2H_2O$
Delrioite	$CaSr(V_2O_6)(OH)_2$ $3H_2O$
Metadelrioite	$CaSrV_2O_6(OH)_2$
Metamunirite	$NaVO_3$
Metarossite	$Ca(V_2O_6)$ $2H_2O$
Munirite	$NaVO_3$ $(2\ x)H_2O$, chains
Rossite	$Ca(V_2O_6)$ $4H_2O$
Layered $V_6O_{16}^{n-}$ vanadate(IV,V) minerals of the hewittite group[a]	
Barnesite	$(Na,Ca)(V_6O_{16})$ $3H_2O$
Grantsite	$(Na,Ca)(V_6O_{16})$ $4H_2O$
Hewettite	$Ca(V_6O_{16})$ $9H_2O$
Hendersonite	$Ca_{1\ 3}(V_6O_{16})$ $6H_2O$
Metahewettite	$Ca(V_6O_{16})$ $3H_2O$
Layered $V_8O_{20}^{n-}$ vanadate(IV,V) minerals of the straczekite group[a]	
Bariandite	$Al_{0\ 6}(V_8O_{20})$ $18H_2O$
Bokite	$(Al,Fe)_{1\ 4}[(V,Fe)_8O_{20}$ 7 $4H_2O$
Corvusite	$(Na,KCa,Mg)_2(V_8O_{20})$ 6 $104H_2O$
Fernandinite	$Ca_{0\ 6}(V_8O_{20})$ $10H_2O$
Straczekite	$(Ca,K,Ba)_2(V_8O_{20})$ $6H_2O$
Other vanadium bronze type minerals[a]	
Bannermanite	$Na_{0\ 7}(V_6O_{15})$
Fervanite	$Fe_4(V_4O_{16})$ $5H_2O$
Kazakhstanite[b]	$Fe_5(V_{15}O_{39})(OH)_9$ $9H_2O$
Melanovanadite	$Ca(V_4O_{10})$ $5H_2O$
Schubnelite	$Fe_{2-x}(V_2O_4)(OH)_8$ (x is small)
Miscellaneous vanadate minerals	
Phosphovanadylite[c]	$(Ba,Ca,K,Na)_x[(V,Al)_4P_2(O,OH)_{16}]$ $12H_2O$
Rankachite[d]	$CaFeV_4W_8O_{36}$ $12H_2O$
Satpaevite[e]	$Al_{12}V_8O_{37}$ $30H_2O$
Sherwoodite[f]	$Ca_9[AlV_{14}O_{40}]_2$ $56H_2O$
Simplotite[g]	CaV_4O_9 $5H_2O$
Uvanite	$(UO_2)_2(V_6O_{17})$ $15H_2O(?)$
Vanalite[h]	*ca* $NaAl_9(V_{12}O_{44})(OH)_4$ $33H_2O$

[a] Vanadium bronzes that contain variable amounts of V(IV); stoichiometries are indicative.
[b] Related to minerals of the straczekite group.
[c] Contains V(IV) in a V_4O_{16} cluster linked by phosphate groups.
[d] A heteropolytungstate containing V(V) and W(VI).
[e] A hydrated, mixed Al(III) V(IV,V) oxide cluster species.
[f] Sherwoodite is the calcium salt of the 14 vanadoaluminate heteropolyanion $(AlV_{14}O_{40})^n$, in which a portion of the vanadium is in the V(IV) state.
[g] Contains a V(IV) oxyanion cluster.
[h] Contains a mixed V(IV,V) isopoly or hereropolyanion.

comparable ions, such as silicate, may also substitute with charge balance compensation. Thus, many of the formulae conceal more complex compositions for natural materials, although substitution for vanadate is more limited in many cases than is found for mutually soluble phosphate and arsenate analogues and related minerals.

Separate mention is made in Table 1 of the layered uranyl vanadates, which represent important ore minerals both of uranium and vanadium. Carnotite and tyuyamunite are perhaps the most significant in this respect. All of these species are members of the group of minerals known as the uranium micas. Their structures are based on infinite sheets of linked uranyl

(UO_2^{2+}) and vanadate ions (phosphate and arsenate in the cases of the autunite, meta-autunite, and related groups), with other cations, hydroxide ions, and water molecules lying between the sheets. Limited substitution of phosphate and arsenate is reported for the vanadate minerals, which are extremely insoluble species.

In aqueous V(V) solutions at low concentrations, only mononuclear species are present, these comprising VO_2^+, $VO(OH)_3$, $VO_2(OH)_2$, VO_3OH^2, and VO_4^3. In contrast to this, at higher total V(V) concentrations under acid conditions, extensive polymerisation of vanadate occurs according to the following equations:

$$2VO_4^3 + 2H^+ \leftrightarrow V_2O_7^4 + H_2O \quad \text{(pyrovanadate)}$$

$$3VO_4^3 + 6H^+ \leftrightarrow V_3O_9^3 + 3H_2O \quad \text{(metavanadate)}$$

$$2V_2O_7^4 + 4H^+ \leftrightarrow V_4O_{12}^4 + 2H_2O \quad \text{(metavanadate)}$$

$$5V_4O_{12}^4 + 8H^+ \leftrightarrow 2V_{10}O_{28}^6 + 4H_2O \quad \text{(decavanadate)}$$

Protonated analogues are also present in solution. The pyrovanadates, metavanadates, and decavanadates are all expressed in a number of mineral structures, as listed in Table 1. Metavanadates, in the solid state, consist of infinite chains of VO_3 composition, $(VO_3)_n$.

Divanadate and tetravanadate chains are the basic structural units of many of the vanadium bronzes, also listed in Table 1. These minerals assume a variety of colours, owing to the fact that a variable though usually small proportion of the V(V) present is reduced to V(IV). Consequent variations of stoichiometry are noted as a result of charge compensation and this is an additional feature of other members of the mineral groups containing both metavanadate and decavanadate, and their protonated derivatives. Other vanadium bronzes are based on different structural motifs, but all share the above characteristics in terms of the presence of V(IV) and attendant compositional variation. The natural expression of the vanadium bronzes, exhaustively studied as artificial compounds, is dependent upon subtle fluctuations of redox potential in the solutions from which they form. They, and other vanadium species, are important constituents of sandstone vanadium deposits and their geochemical relationships have been thoroughly explored, based on pioneering studies dating back nearly five decades. The mineral phosphvanadylite contains V(IV) in a V_4O_{16} cluster linked by phosphate groups, but its chemistry of formation remains to be elucidated.

Like the tetaoxomolybdate(VI) and tetaoxotungstate(VI) ions, vanadate forms heteropolyanions in acid solutions. Rankachite and sherwoodite are two examples, the former carrying W(VI) in the cluster and the latter being a naturally occurring salt of the well-known synthetic 14-vanadoaluminate heteropolyanion $(AlV_{14}O_{40}]^n$. No doubt, other congeners exist in nature in line with observations concerning the geochemistry of molybdenum and tungsten.

As mentioned above, vanadate minerals are an important source of the element vanadium. In addition to species already singled out in this connection, vanadinite, mottramite, and descloizite have been exploited as vanadium ores. However, the bulk of present vanadium production comes from vanadium-bearing oxides (principally magnetite) and by-product refining.

See Also

Minerals: Arsenates. **Sedimentary Rocks:** Phosphates.

Further Reading

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (1997) *Handbook of Mineralogy. Volume 3. Halides, Hydroxides, Oxides*. Tucson: Mineral Data Publishing.

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2000) *Handbook of Mineralogy. Volume 4. Arsenates, Phosphates, Vanadates*. Tucson: Mineral Data Publishing.

Anthony JA, Bideaux RA, Bladh KW, and Nichols MC (2003) *Handbook of Mineralogy. Volume 5. Borates, Carbonates, Sulfates, Chromates, Germanates, Iodates, Molybdates, Tungstates, etc., and Organic Materials*. Tucson: Mineral Data Publishing.

Baes CF, Jr and Mesmer RE (1986) *The Hydrolysis of Cations*. Reprint Edition. Malabar, Florida: Krieger Publishing Company.

Bard AJ, Parsons R, and Jordan J (1985) *Standard Potentials in Aqueous Solution*. New York: Marcel Dekker.

Evans HT, Jr and Hughes JM (1990) Crystal chemistry of the natural vanadium bronzes. *The American Mineralogist* 75: 508 521.

Evans HT, Jr and Garrels RM (1958) Thermodynamic equilibria of vanadium in aqueous systems as applied to the interpretation of the Colorado Plateau ore deposits. *Geochimica et Cosmochimica Acta* 15: 131 149.

Gaines RV, Skinner HCW, Foord EE, Mason B, and Rosenzweig A (eds.) (1997) *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, 8th edn. London: Wiley Europe.

Mandarino JA (1999) *Fleischer's Glossary of Mineral Species 1999*, 8th edn. Tucson: Mineralogical Record Inc.

Roberts WL, Campbell TJ, and Rapp GR, Jr (1990) *Encyclopedia of Minerals*, Second Edition. New York: Van Nostrand Reinhold.

Schindler M, Hawthorne FC, and Baur WH (2000) A crystal chemical approach to the composition and occurrence of vanadium minerals. *The Canadian Mineralogist* 38: 1443 1456.

Williams PA (1990) *Oxide Zone Geochemistry*. Chichester: Ellis Horwood.

Zeolites

W S Wise, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

Zeolites are aluminosilicate minerals that occur as low-temperature (generally less than 200°C) alteration products of volcanic and feldspathic rocks. They are well known in cavities of basalt, having crystallized as a result of diagenetic or hydrothermal alteration (Figure 1). Some zeolites completely replace rhyolitic tuff in saline alkaline lacustrine environments or through groundwater percolation. Thick sequences of sediment from arc-source terranes contain several different kinds of zeolite, which formed through diagenetic alteration and very low-grade metamorphism. These occurrences of zeolites have economic significance, because they either produce useful rock or affect the porosity of reservoir rocks.

A zeolite is a hydrated aluminosilicate mineral with a structure characterized by a framework of linked tetrahedra, each consisting of four oxygen atoms surrounding a silicon or aluminium cation. This three-dimensional network has open cavities in the forms of channels and cages, which are occupied by water molecules and non-framework cations. The channels are large enough to allow the passage of guest ionic and molecular species. This feature allows certain zeolites to exchange non-framework cations with the surrounding solution or to act as molecular sieves. For example, clinoptilolite (Table 1) can selectively remove radiogenic strontium from nuclear wastewater, and chabazite can remove carbon dioxide from methane in the gas generated in landfill sites.

Figure 1 Crystals of laumontite and calcite in a cavity in an altered basalt near Kailua, northeast Oahu, Hawaii. This area is within the former conduit system of the 2.6 Ma Koolau volcano. It is likely that the pervasive rock alteration and zeolite formation is a result of hydrothermal activity when the volcano was active. The long dimension of the cavity is 8 cm.

Each zeolite mineral or, in many cases, zeolite series has a different framework arrangement. Table 1 lists the currently recognized zeolite mineral species and series, their generalized formulae, and structure-type codes. Many more zeolites have been synthesized in the laboratory. The Structure Commission of the International Zeolite Association assigns a three-letter code to each different arrangement, including the synthetic ones. At the time of writing there are 145 structure types, of which 42 occur as minerals. For updates see the International Zeolite Association website http://www.iza-structure.org/.

Structure of Zeolites

All silicates have silicon (or other elements such as aluminium) located in the small space between four oxygen anions. This site is called a tetrahedral site because the cation is strongly bonded to the four oxygen anions. In framework silicates each oxygen anion is shared between two tetrahedral sites. Therefore, the resulting framework has two oxygen anions for each silicon cation, giving the composition $Si^{4+}O_2^2$, as in the mineral quartz (*see* **Minerals:** Quartz). Most other framework silicate minerals have aluminium in some of the tetrahedral sites, which results in a charge imbalance. For example, in some feldspars (*see* **Minerals:** Feldspars) the framework has the composition $[AlSi_3O_8]$, and the charge is balanced by the addition of alkali cations in sites near the oxygen anions bonded to the aluminium tetrahedral site.

The zeolite minerals all have framework arrangements with channels or cages (large pockets). In order to illustrate some of the variety among the zeolites, we will examine three different framework structures: natrolite (NAT), chabazite (CHA), and heulandite (HEU).

Natrolite forms long prismatic crystals, in which the basic unit of the structure is the $[Al_2Si_3O_{10}]^2$ chain (Figure 2). Continuous repetition of this unit forms long chains parallel to the prism length. Cross-linking of these chains forms channels parallel to the chains. To balance the charge in the framework, caused by the inclusion of aluminium in two of the

Table 1 Names, general formulae, and framework type of accepted zeolite species

Name	*Generalized formula*	*Z (Z = number of formula units per unit cell.)*	*Structure type code*
Amicite	$K_4Na_4[Al_8Si_8O_{32}].10H_2O$	1	GIS
Ammonioleucite	$(NH_4)[AlSi_2O_6]$	16	ANA
Analcime	$Na[AlSi_2O_6].H_2O$	16	ANA
Barrerite	$Na_8[Al_8Si_{28}O_{72}].26H_2O$	2	STI
Bellbergite	$(K,Ba,Sr)_2Sr_2Ca_2(Ca,Na)_4[Al_{18}Si_{18}O_{72}].30H_2O$	1	EAB
Bikitaite	$Li[AlSi_2O_6].H_2O$	2	BIK
Boggsite	$(Ca,Na_{0\ 5},K_{0\ 5})_9[Al_{18}Si_{78}O_{192}].70H_2O$	1	BOG
Brewsterite series	$(Sr,Ba)_2[Al_4Si_{12}O_{32}].10H_2O$	1	BRE
Brewsterite Ba	$(Ba,Sr)_2[Al_4Si_{12}O_{32}].10H_2O$		
Brewsterite Sr	$(Sr,Ba)_2[Al_4Si_{12}O_{32}].10H_2O$		
Chabazite series	$(Ca_{0\ 5},Na,K)_x[Al_xSi_{12-x}O_{24}].12H_2O$, x 2.4 5.0	1	CHA
Chabazite Ca	$(Ca_{0\ 5},K,Na)_x[Al_xSi_{12-x}O_{24}].12H_2O$, x 2.4 5.0		
Chabazite K	$(K,Na,Ca_{0\ 5})_x[Al_xSi_{12-x}O_{24}].12H_2O$, x 3.0 4.5		
Chabazite Na	$(Na,K,Ca_{0\ 5})_x[Al_xSi_{12-x}O_{24}].12H_2O$, x 2.5 4.8		
Chabazite Sr	$(Sr_{0\ 5},Ca_{0\ 5},K)_4[Al_4Si_8O_{24}].12H_2O$		
Chiavennite	$CaMn[Be_2Si_5O_{13}(OH)_2].2H_2O$	4	CHI
Clinoptilolite series	$(Na,K,Ca_{0\ 5},Mg_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$	1	HEU
Clinoptilolite Ca	$(Ca_{0\ 5},Na,K,Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Clinoptilolite K	$(K,Na,Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Clinoptilolite Na	$(Na,K,Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Cowlesite	$Ca[Al_2Si_3O_{10}].5.3H_2O$	52 unknown	
Dachiardite series	$(Ca_{0\ 5},Na,K)_{4-5}[Al_{4-5}Si_{20-19}O_{48}].13H_2O$	1	DAC
Dachiardite Ca	$(Ca_{0\ 5},Na,K)_5[Al_5Si_{19}O_{48}].13H_2O$		
Dachiardite Na	$(Na,K,Ca_{0\ 5})_4[Al_4Si_{20}O_{48}].13H_2O$		
Edingtonite	$Ba[Al_2Si_3O_{10}].4H_2O$	2	EDI
Epistilbite	$(Ca,Na_2)_3[Al_6Si_{18}O_{48}].16H_2O$	1	EPI
Erionite series	$K_2(Na,Ca_{0\ 5})_8[Al_{10}Si_{26}O_{72}].30H_2O$	1	ERI
Erionite Ca	$K_2(Ca_{0\ 5},Na)_8[Al_{10}Si_{26}O_{72}].30H_2O$		
Erionite K	$K_2(K,Na,Ca_{0\ 5})_7[Al_9Si_{27}O_{72}].30H_2O$		
Erionite Na	$K_2(Na,Ca_{0\ 5})_7[Al_9Si_{27}O_{72}].30H_2O$		
Faujasite series	$(Na,Ca_{0\ 5},Mg_{0\ 5},K)_x[Al_xSi_{12-x}O_{24}].16H_2O$, x 3.2 3.8	16	FAU
Faujasite Ca	$(Ca_{0\ 5},Na,Mg_{0\ 5},K)_x[Al_xSi_{12-x}O_{24}].16H_2O$, x 3.3 3.9		
Faujasite Mg	$(Mg_{0\ 5},Ca_{0\ 5},Na,K)_{3\ 5}[Al_{3\ 5}Si_{8\ 5}O_{24}].16H_2O$		
Faujasite Na	$(Na,Ca_{0\ 5},Mg_{0\ 5},K)_x[Al_xSi_{12-x}O_{24}].16H_2O$, x 3.2 4.3		
Ferrierite series	$(K,Na,Mg_{0\ 5},Ca_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$	1	FER
Ferrierite K	$(K,Na,Mg_{0\ 5},Ca_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Ferrierite Mg	$(Mg_{0\ 5},K,Na,Ca_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Ferrierite Na	$(Na,K,Mg_{0\ 5},Ca_{0\ 5})_6[Al_6Si_{30}O_{72}].20H_2O$		
Garronite	$(Ca_{0\ 5},Na)_6[Al_6Si_{10}O_{32}].14H_2O$	1	GIS
Gaultite	$Na_4[Zn_2Si_7O_{18}].5H_2O$	8	VSV
Gismondine	$Ca_4[Al_8Si_8O_{32}].18H_2O$	1	GIS
Gmelinite series	$(Na,Ca_{0\ 5},K)_8[Al_8Si_{16}O_{48}].22H_2O$	1	GME
Gmelinite Ca	$(Ca_{0\ 5},Sr,K,Na)_8[Al_8Si_{16}O_{48}].22H_2O$		
Gmelinite K	$(K,Ca_{0\ 5},Na)_8[Al_8Si_{16}O_{48}].22H_2O$		
Gmelinite Na	$(Na,K,Ca_{0\ 5})_8[Al_8Si_{16}O_{48}].22H_2O$		
Gobbinsite	$Na_5[Al_5Si_{11}O_{32}].12H_2O$	1	GIS
Gonnardite	$(Na,Ca_{0\ 5})_{8-10}[Al_{8+x}Si_{12\ \ x}O_{40}].12H_2O$, x 0 2	1	NAT
Goosecreekite	$Ca[Al_2Si_6O_{16}].5H_2O$	2	GOO
Gottardiite	$(Na,K)_3Mg_3Ca_5[Al_{19}Si_{117}O_{272}].93H_2O$	1	NES
Harmotome	$(Ba_{0\ 5},Ca_{0\ 5},K,Na)_5[Al_5Si_{11}O_{32}].12H_2O$	1	PHI
Heulandite series	$(Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5},Na,K)_9[Al_9Si_{27}O_{72}].24H_2O$	1	HEU
Heulandite Ca	$(Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5},Na,K)_9[Al_9Si_{27}O_{72}].24H_2O$		
Heulandite Na	$(Na,Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5},K)_9[Al_9Si_{27}O_{72}].24H_2O$		
Heulandite K	$(K,Ca_{0\ 5},Sr_{0\ 5},Ba_{0\ 5},Mg_{0\ 5},Na)_9[Al_9Si_{27}O_{72}].24H_2O$		
Heulandite Sr	$(Sr_{0\ 5},Ca_{0\ 5},Ba_{0\ 5},Mg_{0\ 5},Na,K)_9[Al_9Si_{27}O_{72}].24H_2O$		
Hsianghualite	$Li_2Ca_3[Be_3Si_3O_{12}]F_2$	8	ANA
Kalborsite	$K_6[Al_4Si_6O_{20}]B(OH)_4Cl$	2	?EDI
Laumontite	$Ca_4[Al_8Si_{16}O_{48}].18H_2O$	1	LAU
Leucite	$K[AlSi_2O_6]$	16	ANA

Continued

Table 1 Continued

Name	*Generalized formula*	*Z (Z = number of formula units per unit cell.)*	*Structure type code*
Levyne series	$(Ca_{0.5},Na,K)_6[Al_6Si_{12}O_{36}].17H_2O$	1	LEV
Levyne Ca	$(Ca_{0.5},Sr,K,Na)_6[Al_6Si_{12}O_{36}].17H_2O$		
Levyne Na	$(Na,K,Ca_{0.5})_6[Al_6Si_{12}O_{36}].17H_2O$		
Lovdarite	$K_4Na_{12}[Be_8Si_{28}O_{72}].18H_2O$	1	LOV
Maricopaite	$(Pb_7Ca_2)[Al_{12}Si_{36}(O,OH)_{100}].nH_2O$	1	?MOR
Mazzite	$(Mg_{2.5}K_2Ca_{1.5})[Al_{10}Si_{26}O_{72}].30H_2O$	1	MAZ
Merlinoite	$(K,Ca_{0.5},Ba_{0.5},Na)_{10}[Al_{10}Si_{22}O_{64}].22H_2O$	1	MER
Mesolite	$Na_2Ca_2[Al_6Si_9O_{30}].8H_2O$	8	NAT
Montesommaite	$K_9[Al_9Si_{23}O_{64}].10H_2O$	1	MON
Mordenite	$(Na_2,Ca,K_2)[Al_8Si_{40}O_{96}].28H_2O$	1	MOR
Mutinaite	$Na_3Ca_4[Al_{11}Si_{85}O_{192}].60H_2O$	1	MFI
Nabesite	$Na_2BeSi_4O_{10}.4H_2O$	4	NAB
Natrolite	$Na_2[Al_2Si_3O_{10}].2H_2O$	8	NAT
Offretite	$CaKMg[Al_5Si_{13}O_{36}].16H_2O$	1	OFF
Pahasapaite	$(Ca_{5.5}Li_{3.6}K_{1.2}Na_{0.2}\square_{13.5})Li_8[Be_{24}P_{24}O_{96}].38H_2O$	1	RHO
Parthéite	$Ca_2[Al_4Si_4O_{15}(OH)_2].4H_2O$	4	PAR
Paulingite series	$(K,Ca_{0.5},Na,Ba_{0.5})_{10}[Al_{10}Si_{32}O_{84}].27\ 44H_2O$	16	PAU
Paulingite K	$(K,Ca_{0.5},Na)_{10}[Al_{10}Si_{32}O_{84}].44H_2O$		
Paulingite Ca	$(Ca_{0.5},K,Na,Ba_{0.5})_{10}[Al_{10}Si_{32}O_{84}].27\ 34H_2O$		
Perlialite	$K_9Na(Ca,Sr)[Al_{12}Si_{24}O_{72}].15H_2O$	1	LTL
Phillipsite series	$(K,Na,Ca_{0.5})_x[Al_xSi_{16-x}O_{32}].12H_2O,\ x\ \ 3.8\ \ 6.4$	1	PHI
Phillipsite Na	$(Na,K,Ca_{0.5})_x[Al_xSi_{16-x}O_{32}].12H_2O,\ x\ \ 3.7\ \ 6.7$		
Phillipsite K	$(K,Na,Ca_{0.5})_x[Al_xSi_{16-x}O_{32}].12H_2O,\ x\ \ 3.8\ \ 6.4$		
Phillipsite Ca	$(Ca_{0.5},K,Na)_x[Al_xSi_{16-x}O_{32}],12H_2O,\ x\ \ 4.1\ \ 6.8$		
Pollucite	$(Cs,Na)[AlSi_2O_6].nH_2O\ (Cs + n\ \ 1)$	16	ANA
Roggianite	$Ca_2[BeAl_2Si_4O_{13}(OH)_2].2\ \ 2.5H_2O$	8	ROG
Scolecite	$Ca[Al_2Si_3O_{10}].3H_2O$	8	NAT
Stellerite	$Ca_4[Al_8Si_{28}O_{72}].28H_2O$	2	STI
Stilbite series	$(Ca_{0.5},Na,K)_9[Al_9Si_{27}O_{72}].30H_2O$	1	STI
Stilbite Ca	$(Ca_{0.5},Na,K)_9[Al_9Si_{27}O_{72}].30H_2O$		
Stilbite Na	$(Na,K,Ca_{0.5})_9[Al_9Si_{27}O_{48}].28H_2O$		
Terranovaite	$NaCa[Al_3Si_{17}O_{40}].13H_2O$	1	TER
Thomsonite series	$(Ca,Sr)_2Na[Al_5Si_5O_{20}].6\ \ 7H_2O$	4	THO
Thomsonite Ca	$Ca_2Na[Al_5Si_5O_{20}].6H_2O$		
Thomsonite Sr	$(Sr,Ca)_2Na[Al_5Si_5O_{20}].7H_2O$		
Tschernichite	$(Ca,Mg,Na_{0.5})[Al_2Si_6O_{16}].8H_2O$	8	BEA
Tschortnerite	$Ca_4(K_2,Ca,Sr,Ba)_3Cu_3(OH)_8[Al_{12}Si_{12}O_{48}].nH_2O$	16	TSC
Wairakite	$Ca[Al_2Si_4O_{12}].2H_2O$	8	ANA
Weinebeneite	$Ca[Be_3PO_8(OH)_2].4H_2O$	4	WEI
Willhendersonite	$Ca_{3-x}K_x[Al_6Si_6O_{24}].10H_2O,\ x\ \ 0.0\ \ 2.0$	1	CHA
Yugawaralite	$Ca[Al_2Si_6O_{16}].4H_2O$	2	YUG

Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

five tetrahedral sites, sodium cations are bonded within the channels (Figure 3). The bonding of these sodium cations to opposite sides of the channel partially collapses the channel from a possible square cross-section to an elliptical one. Each sodium cation is bonded by two water molecules, which are in turn each bonded to two sodium cations.

The crystal form of chabazite is a rhombohedron that is almost cubic. The framework consists of six-member rings of tetrahedra, which are arranged in the sequence AABBCC... (Figure 4). Each different letter represents a different position of the six-member ring around the stacking direction parallel to the *c*-axis. The distribution of aluminium in the tetrahedral sites is nearly random. Non-framework cations partially fill four different sites: (i) outside the double six-member ring and bonded with three oxygen atoms of the ring; (ii) near the eight-member ring port of the chabazite cage and bonded with three framework oxygen atoms; (iii) near the centre of the cage and not bonded with the framework; and (iv) at the centre of the eight-member ring. Positions (i) and (iii) generally contain most of the cations.

Crystals of heulandite and clinoptilolite are tabular with a perfect cleavage. The framework consists of sheets of tetrahedra with no openings large enough

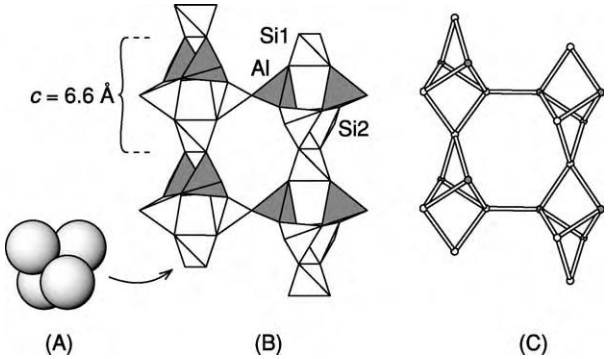


Figure 2 The framework arrangement in natrolite. (A) A representation of four oxygen anions surrounding a silicon cation (unseen). This site can be represented by a tetrahedron by reducing the oxygen spheres to points and connecting the four points with lines, as in (B). (B) Two chains of the natrolite framework. A unit of the natrolite chain consists of a Si1 tetrahedron, two Si2 tetrahedra (all three shown as white), and two Al tetrahedra (grey). The length of this unit, which is repeated continuously along the length of the chain, is 6.6 Å. Si2 tetrahedra connect with Al tetrahedra to link chains together. Because an Al O Al bond sequence is energetically less favourable than an Al O Si sequence, they tend not to occur. (C) Another way to represent the framework arrangement (topology) is to eliminate the oxygen positions and connect the tetrahedral sites with double lines. The method makes visualization of more complex arrangements easier.

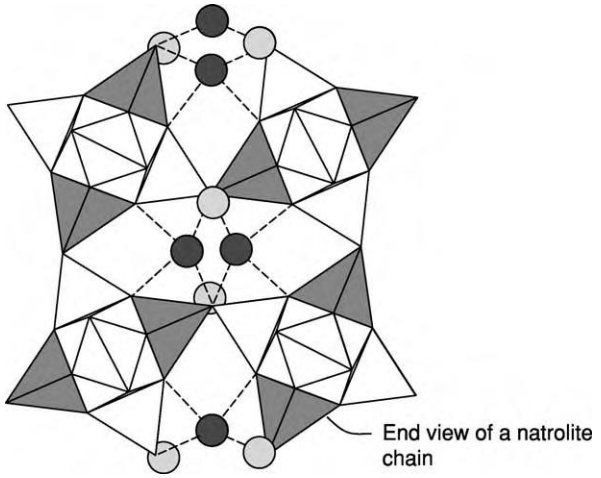


Figure 3 The bonding of sodium cations (dark grey circles) and water molecules (light grey circles) in the channels of natrolite. Sodium cations are bonded to four of the framework oxygen anions between Al and Si tetrahedra (dashed lines show bonds) and to two water molecules, which are located in spaces between linked chains (see **Figure 2**).

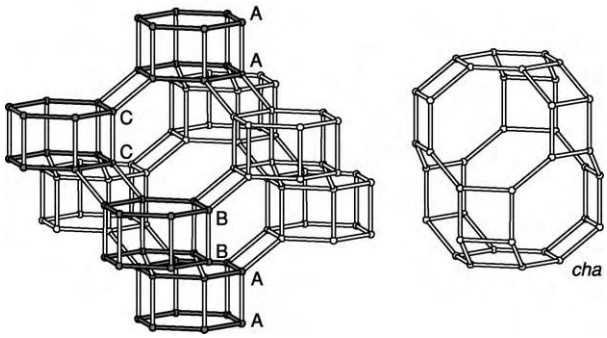


Figure 4 The arrangement of the framework of chabazite, in which only tetrahedral sites and linkages are plotted. A set of six member rings that defines the framework is darkened, to illustrate the stacking sequence AABBCC.... Within the spiral stair stepped linked double six member rings is a large space, called a *cha* cage. For clarity this *cha* cage has also been drawn outside the structure. Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

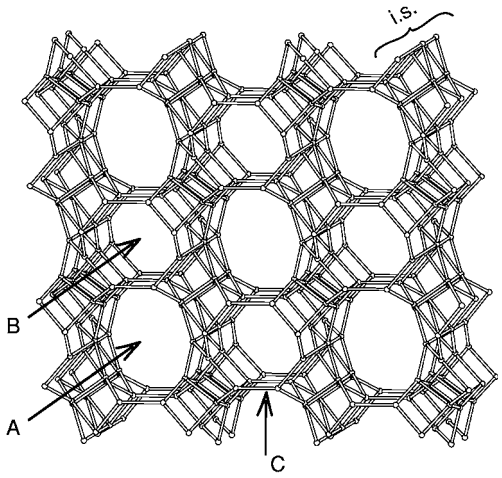


Figure 5 The arrangement of the framework of heulandite and clinoptilolite, in which only the tetrahedral sites are plotted and connected without the bridging oxygen ions. The framework consists of impermeable sheets (i.s.) that are cross linked, forming three different types of channel, A, B, and C; A and B are almost perpendicular to the page, while C is parallel to it. Non framework cations and water molecules occupy various sites within these channels. Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

to allow the passage of ions or molecules. In effect the sheets are impermeable, as shown in Figure 5. The sheets are bonded together with widely spaced oxygen bridges, between which are three different channels. There are relatively few aluminium cations in this structure, leaving the framework silicon-rich, and the distribution of the aluminium is random. Non-framework cations are located along the centres of the channels, weakly bonded to the sides and to surrounding water molecules.

Chemical Composition of Zeolites

A perusal of Table 1 shows that zeolite minerals are for the most part aluminosilicates of sodium, potassium, or calcium, with magnesium, strontium, and barium being less common. There is significant variation in composition between mineral species, but also within several of the species. Compositional variation may occur within the framework, i.e. the amount of aluminium in tetrahedral sites, and in the charge-balancing non-framework cations. One way to illustrate the range of zeolite compositions is with a plot like that shown in Figure 6. This diagram shows the wide range of silicon (or aluminium) content observed among zeolite species, as well as the variability in the non-framework cations.

Some zeolite mineral series exhibit a considerable variation in the framework composition. For example, within the chabazite structure group the silicon content ranges from high, Si/Al = 3.0, giving a framework composition of $[Al_{2.4}Si_{9.6}O_{24}]$, to low, Si/Al = 1.0, yielding $[Al_6Si_6O_{24}]$ in willhendersonite (Table 1). Other mineral series that have wide ranges of framework composition are heulandite, clinoptilolite, and phillipsite. The ratio Si/Al = 1.0 appears to be the lower limit for the amount of aluminium in any framework. More aluminium would require some Al–O–Al bond sequences, which are energetically less stable than Al–O–Si bond sequences.

Detailed structural studies of heulandite, clinoptilolite, and phillipsite indicate that the Si/Al distribution throughout the framework is random. The framework arrangements of other zeolites seem to require a fixed Si/Al ratio, with the aluminium and silicon in specific sites, i.e. an ordered arrangement. A good example is the arrangement and composition of natrolite (Figure 2). Interestingly, this is true only for well-formed crystals. Crystals that are forced to grow rapidly form small crystals with disordered structures, in which the distribution of aluminium and silicon tends to be random and the ratio of Si : Al varies from the 2 : 3 of ordered natrolite.

The ordered frameworks of some species are commonly preferential to one of these cations. For example, the sodium in analcime tightens the framework such that inclusion of calcium in the sodium sites is strongly restricted. Because only one calcium cation is required to balance two aluminium cations in the framework, they tend to distort the structure.

An important chemical property of any zeolite is its ion-exchange capacity, i.e. the extent to which external cations can be exchanged with those naturally present in the structure. High-silica zeolites (Figure 6), for example clinoptilolite, have few aluminium cations in tetrahedral sites, and therefore have few non-framework cations available for exchange. Conversely, the low-silica zeolites, such as phillipsite, have high ion-exchange capacities.

Included in the original definition of a zeolite by Cronstedt in 1756 was the release of bubbles of water vapour upon heating a crystal in a blowpipe flame. All presently recognized zeolite minerals (Table 1) contain water molecules in the channels and cages. In some, such as natrolite, the molecules are tightly bonded, whereas in others, such as clinoptilolite, the water is loosely held. Heating of the zeolite generally releases the water at temperatures of less than about 500°C. Most structures will rehydrate with exposure to a humid environment. The number of molecules listed for each zeolite is one of the least-restrictive features of their composition, because the amount held in the structure may vary with humidity.

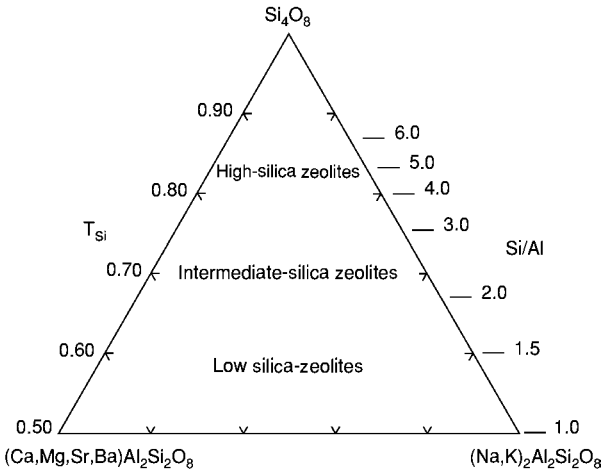


Figure 6 Compositional diagram illustrating major variations of zeolite composition. Components for the plot are: Si_4O_8 (molecular proportion [Si 2(Ca + Mg + Sr + Ba) 2(Na + K)]/4); $(Ca,Mg,Sr,Ba)Al_2Si_2O_8$ (molecular proportion Ca + Mg + Sr + Ba); and $(Na,K)_2Al_2Si_2O_8$ (molecular proportion (Na + K)/2). T_{Si} = Si/(Si + Al). Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

Occurrences of Zeolites

Zeolite minerals occur in sedimentary, metamorphic, and igneous rocks, almost exclusively as alteration products. Most form at temperatures less than 250°C and at depths less than 10 km in water-saturated environments (Figure 7). The alteration process must produce an abundance of silicon and aluminium along with alkali or alkaline earth elements. This commonly occurs where volcanic materials – lava or tuff – react with surface water or groundwater, or where plagioclase-bearing rock undergoes a reaction with pore fluids.

Zeolites in Sedimentary Rocks

The most widespread and economically important of all zeolite occurrences are in sediments and sedimentary rocks composed partially or wholly of volcanic clasts. During diagenesis (*see* **Diagenesis, Overview**) the volcanic component of these rocks reacts with pore fluids to form clay and zeolite minerals. This alteration can take place in various environments, including shallow alkaline lakes, the deep seafloor, thick terrestrial pyroclastic deposits, and sandstone beds in shallow to deep marine basins. Diagenetic reactions merge with those of burial metamorphism (Figure 7), and the distinction is a matter of the terminology used by researchers in this field. In this review, diagenesis refers to all changes of the initial sediment between deposition and the completion of lithification or cementation, and metamorphism refers to changes occurring after diagenesis, generally involving the breakdown of the first diagenetic minerals in response to increases in temperature and pressure.

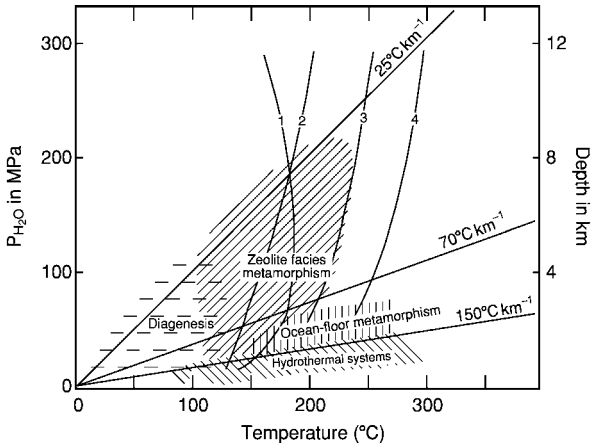


Figure 7 Pressure temperature ranges of environments of zeolite formation. The boundary between diagenesis and meta morphism overlaps considerably more than shown here. Numbered curves are experimentally determined or calculated stability limits of selected zeolites: 1, analcime + quartz albite + water; 2, heulandite laumontite + quartz + water; 3, laumontite + prehnite epidote + quartz + water at low pressure end and laumontite + pumpellyite epidote + chlorite + quartz + water at high pressure end; and 4, laumontite wairakite + water. The depth scale is based on a lithostatic gradient of 25 MPa km ^{1}. Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

Shallow alkaline lakes Beds of zeolite-rich rock a few centimetres to several metres thick occur in several areas of the world, such as the western USA, Serbia, and western Turkey. Clinoptilolite is the most abundant zeolite in these beds, but it may be associated with chabazite, phillipsite, erionite, and analcime. The zeolitic beds are interlayered with silt and were deposited in shallow playa lakes in closed basins. With no outlets, these lakes become saline and alkaline through extended periods of evaporation. Nearby explosive rhyolitic volcanism produced vitric ash that reacted with the alkaline water of the lake to produce the beds of zeolite-rich rock. Figure 8 shows the relationships between mineral zoning and hydrology in this type of environment.

Glass persists unaltered where pore solutions have a pH of less than about 8.5. The clinoptilolite zone may contain other zeolites, such as erionite, chabazite, and phillipsite, all of which pseudomorph glass shards. Analcime, which is abundant in the next zone inwards, replaces both pre-existing zeolites of the clinoptilolite zone and, less commonly, glass shards. Potassium feldspar, dominant in the innermost zone, replaces earlier-formed zeolites. This horizontal zonation of the zeolites and potassium feldspar along a single tuff bed is characteristic of this type of alteration, and puzzling to understand.

A kinetic model has been proposed to explain the formation of zeolites in tuffaceous rocks. The model

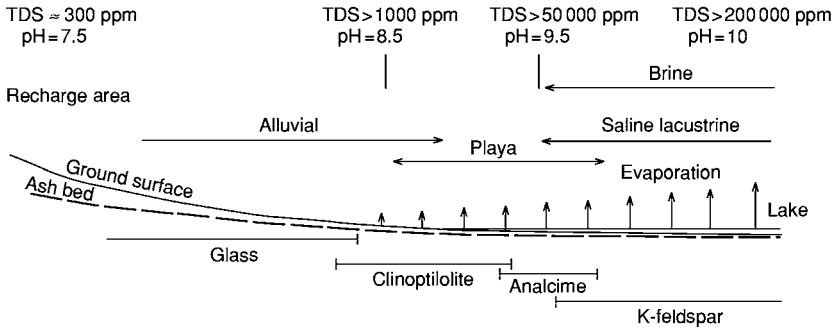


Figure 8 Relationships among hydrology, geology, and mineralogy in a playa saline lake. TDS, total dissolved solids. Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

incorporates the idea that glass dissolution proceeds at much greater rates than stable authigenic minerals, such as analcime and feldspar, can crystallize. The less-stable disordered alkali zeolites (chabazite, clinoptilolite, phillipsite, erionite) nucleate and crystallize readily, lowering the free energy of the system. With time, these metastable phases are replaced by stable ones, but changes in intensive parameters, such as temperature or pH, can greatly accelerate the reaction.

Deep sea sediment Zeolites, predominantly phillipsite and clinoptilolite, occur as a moderate component (5–15%) of many deep-marine sediments that include calcareous ooze and chalk, radiolarian ooze, hemipelagic mud, hyaloclastite, ash layers, or volcaniclastic sand. Zeolite-bearing sediment has been recovered from all parts of all oceans. Phillipsite tends to be common at shallow core depths, while clinoptilolite is more abundant in deeper (and older) sediment. Because phillipsite has been recovered from very near the sediment–water interface, it is believed that initial reactions of seawater with clay and/or volcanic debris occur at or just below the seafloor.

Thick terrestrial pyroclastic deposits Most terrestrial accumulations of volcanic material – lavas and pyroclastic deposits (*see* **Pyroclastics; Lava**) – are subject to alteration by meteoric water and groundwater. Initial alteration occurs as meteoric water percolates through permeable pyroclastic deposits and reacts with very-fine-grained glass, forming smectitic clay and alkaline vadose water (Figure 9).

Above the water table, fresh glass shards persist, even though the outer surface has been altered to smectite, and the very fine particles may be replaced by smectite and opal. Because the water-to-rock ratio is high (between 5 and 10), through-flowing vadose water carries away alkali and alkaline earth cations as they are released from the altering glass surface. The remaining alumina, iron, and silica readily form clay, but zeolite growth is minor. Below the water table, where fluid flow rates are much slower, the water-to-rock ratio is less than one, and zeolite crystallization occurs. At this depth clinoptilolite is an abundant replacement of glass shards and may be associated with chabazite, phillipsite, erionite, and mordenite, along with smectite and a silica phase. At somewhat greater depths, analcime generally replaces shards and pumice lapilli, and fills voids, but in some areas textural evidence suggests that analcime replaces clinoptilolite. At the deepest levels, authigenic albite and potassium feldspar replace pre-existing zeolites.

Compositions of the pyroclastic rock other than rhyolite lead to different zeolitic alteration products.

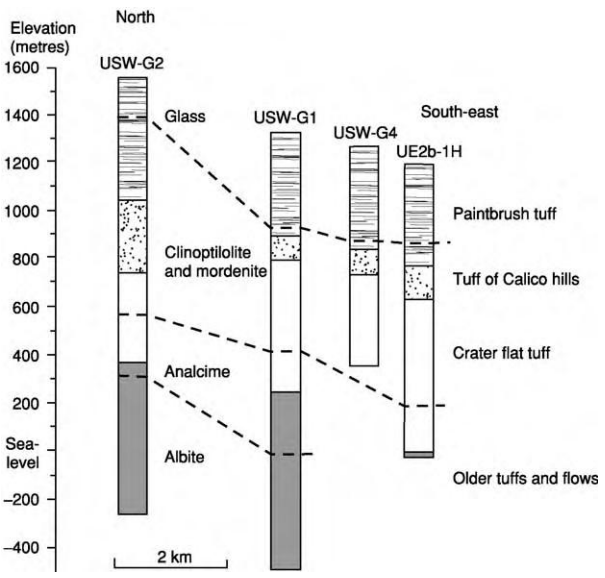


Figure 9 The relationships between diagenetic mineral facies (heavy dashed lines) and ash flow tuff stratigraphy (patterns in columns) at the Nevada Test Site, Yucca Mountain, Nevada, USA. The listed mineral is the dominant diagenetic mineral for that facies. The boundaries are higher to the north (left side of dia gram) because that is the direction of the source of eruptions and presumably the temperatures were higher. Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

For example, phonolitic and tephritic tuffs in Italy alter mainly to phillipsite.

Sandstone beds in shallow to deep basins The largest volumes of zeolite-bearing sedimentary rock are adjacent to volcanic arcs, where volcaniclastic debris is shed into fore-arc or back-arc basins. Depending somewhat on the composition and amount of the volcanic detritus, common diagenetic minerals are clinoptilolite, mordenite, heulandite, analcime, and laumontite. The amount of sediment accumulated in some basins approaches 10 km, such as that now exposed in the Southland Syncline, New Zealand.

In Japan, diagenetic minerals in Tertiary pyroclastic rocks are distributed in the following zones: zone I, unaltered glass; zone II, alkali-clinoptilolite and mordenite; zone III, analcime and heulandite; zone IV, laumontite; and zone V, albite. The boundaries between zones are approximately parallel to stratification. Drilling through these zones in oilfields shows that the zones are temperature-dependent, with the boundary between zones I and II at about 45°C, that between zones II and III at about 85°C, and that between zones III and IV at about 123°C. A similar distribution of zeolite phases is observed in the Southland Syncline, New Zealand, in a general sense,

but is complicated in detail, suggesting that fluid composition, as well as temperature, may play a role.

Because these sequences are so thick, the alteration of the original volcanic material extends beyond diagenesis into burial metamorphism, especially where early-formed heulandite and analcime are replaced by laumontite and albite. In fact, the zeolite and feldspar minerals of zones III, IV, and V in Japan show replacement textures, indicating early metamorphism.

Metamorphosed Rocks

In volcanic rock sequences, laumontite and analcime are common products following diagenesis and are developed through burial, contact, or ocean-floor metamorphism. Burial of sediment derived from arc-terrains, containing abundant volcanic detritus and plagioclase sand grains, produces such zeolite minerals as heulandite during diagenesis. Early minerals are replaced by higher temperature (or more stable) zeolites with deeper burial. The type example of this style of metamorphism, called burial metamorphism, is the Southland Syncline, New Zealand. At higher temperature and pressure, the zeolite phases ultimately give way to prehnite, pumpellyite, and feldspar.

The lowest-temperature zones of thermal metamorphic aureoles in volcanic terrains, lavas, and pyroclastic rocks may contain zeolite assemblages. Laumontite-bearing assemblages are common in the zeolite zone of the thermally metamorphosed tholeiitic pillow lavas and pyroclastic rocks of the Karmutsen Group, British Columbia. Here, too, the zeolite facies assemblages are replaced by the prehnite–pumpellyite facies. Similar sequences of mineral assemblages have been described in the Tanzawa Mountains, central Japan, although the geological history suggests a combination of thermal and burial metamorphism.

Metamorphism of pillow lavas and dykes near mid-ocean spreading ridges, now observed in cores from deep-sea drilling sites and in ophiolite sequences, commonly produces zeolite-bearing assemblages. This style of metamorphism is known as ocean-floor metamorphism; analcime and natrolite occur at the shallowest depths, and laumontite, scolecite, and heulandite occur at somewhat deeper levels.

Zeolites in Lava Flows

Many zeolites occur in cavities in basaltic rocks. Some notable localities are: the Tertiary lavas of eastern Iceland; County Antrim, Northern Ireland; the Faeroe Islands; the Deccan Plateau, India; the Jurassic basalt units of New South Wales, Australia; and the Triassic basalt belts in and near Paterson, New Jersey, and in Nova Scotia, Canada. This type of zeolite occurrence has been widely referred to as 'hydrothermal' in the literature. However, the zeolites in these occurrences may have originated in any of a number of different ways, including: burial diagenesis; burial metamorphism; hydrothermal alteration and metamorphism; reaction with groundwater in areas of high heat and fluid flow (high geothermal gradient); and reaction with surface water during the initial cooling of a lava flow or plug.

Studies of thick accumulations of basalt and andesite in extensional basins have documented regional low-grade metamorphism. In basins along the Andes of Peru and Chile the rocks are little deformed and the metamorphic grade increases with depth from zeolite to greenschist facies.

The effect of depth on zeolite crystallization in subaerial flood-basalt flows is demonstrated by the distribution of amygdale minerals in the basalt lavas of eastern Iceland. Zeolites or assemblages of zeolites define nearly flat-lying zones that cut across the lava stratigraphy (Figure 10). Similar regionally zoned occurrences of zeolite have been described in sequences in the Faeroe Islands, eastern Greenland, western Greenland, the Deccan Traps in western India, the North Shore Volcanic Group of Minnesota, the Parana lavas of Brazil, and Northern Ireland.

In Iceland and Greenland zeolite crystallization did not occur until after the accumulation and weak

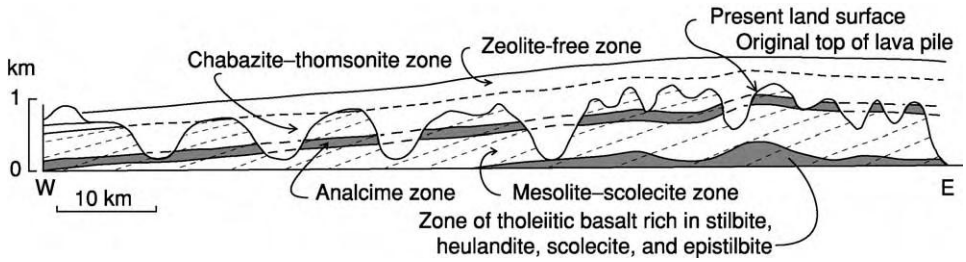


Figure 10 Diagrammatic section through the Tertiary lava pile of eastern Iceland, showing the inferred relationship between the lava stratigraphy (sloping dashed lines) and the distribution of amygdale minerals (subhorizontal bands of grey and white). Reproduced with permission from Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

deformation of the lava pile, because the zones are parallel to the inferred original surface. The reaction of groundmass glass with interstitial groundwater results in the replacement of the glass by smectite and the crystallization of smectite, celadonite, and zeolites in open spaces. Because the water-to-rock ratio is small, the chemical composition of the glass in the host rock exerts important controls on the Si/Al ratio of the framework and on the non-framework cation content of the resulting zeolites. In the Icelandic sequences, tholeiitic basalt vesicles contain a somewhat different assemblage of zeolites from those in olivine basalt. For example, analcime occurs in olivine basalt, whereas in the same depth zone mordenite is found in tholeiitic basalt. Along with evolving zeolite assemblages, the associated phyllosilicate minerals also change with depth. In the shallowest affected rocks smectite, commonly ferroan saponite, is abundant, and this gives way to mixed-layer smectite–chlorite and ultimately to chlorite with increasing temperature. However, the range of permeability within a single flow can cause variations in the degree of crystallization and in mineral assemblages.

Hydrothermal Alteration

Hydrothermal alteration is a chemical replacement of the original minerals (or glass) in a rock by new minerals where hot water delivers reactants and removes the aqueous reaction products. Hydrothermal systems may develop alteration aureoles within sequences of volcanic rocks (*see* **Tectonics:** Hydrothermal Activity). The best-known examples are in active geothermal areas, where the alteration process is presently occurring. However, the term hydrothermal is also used to refer to the alteration caused by a fluid phase following magmatic crystallization of pegmatitic bodies.

Active geothermal systems Zeolites have been found in drill cores from several active geothermal areas, for example in Japan, Iceland, the Wairakei geothermal field (New Zealand), and Yellowstone National Park (Wyoming, USA). These systems involve warm to hot (70–150°C) water circulating through relatively large volumes of rock. Common zeolites found in this environment are mordenite, analcime, heulandite, clinoptilolite, laumontite, yugawaralite, and wairakite.

Late-stage crystallization and alteration of pegmatitic bodies Zeolites are common phases in the late stages of pegmatite crystallization. Early in the formation of a pegmatite dyke, magmatic crystallization produces the bulk of the minerals, but as the magma is depleted and the temperature decreases excess water forms a fluid phase. This fluid produces minerals at lower temperatures, many as alteration products of earlier phases. Excellent examples of hydrothermal zeolite formation in pegmatite bodies are found in the Lovozero Massif, Kola Peninsula, Russia. Here, all of the pegmatite pods, lenses, pipes, and dykes are nepheline syenite. Some analcime may be magmatic in origin but much of it, along with the very abundant natrolite, is of hydrothermal origin.

Zeolites also occur in vapour pockets late in the crystallization of granitic pegmatite dykes, mostly as a result of the alteration of feldspar either by residual magmatic fluid or by introduced water. Laumontite and stilbite are commonly precipitated along fractures or in miarolitic cavities. A well-known example is the granodiorite pegmatite dykes near San Piero in Campo on the island of Elba, Italy, where mordenite, dachiardite, stilbite, epistilbite, heulandite, and chabazite occur with pegmatite pocket minerals quartz, tourmaline, beryl, pollucite, and lepidolite.

Fractures and cavities in granitic gneiss Gash fractures formed by the semi-brittle deformation of granitic gneiss may fill with water before the rock has completely cooled, and new minerals may crystallize in the open space. These are the well-known alpine clefts, which contain chlorite, adularia, and quartz crystals that grow at the expense of wall-rock minerals. In some cases reactions continue, producing zeolites such as stilbite, heulandite, scolecite, and chabazite.

Alteration along faults An early discovery in the drill hole into the San Andreas fault zone at Cajon Pass, southern California, was the pervasive occurrence of laumontite and stilbite in shattered granodiorite and gneiss. Both minerals replace plagioclase and fill microfractures throughout the fault zone. This mineralization is attributed to groundwater circulation in fractured and frictionally heated rocks within the fault zone.

Uses of Zeolites

In those areas of the world where beds of vitric tuff or other volcanic-rich sedimentary rock have been altered to nearly pure zeolite, the deposits are suitable for economic exploitation. Over the past 40 years considerable effort has been directed towards finding uses for those zeolites that occur in relative abundance. Applications have been developed in agriculture, in wastewater treatment, in gas separation or enrichment, in building construction, as catalysts, and in energy storage from solar collectors. The most common zeolite

occurring in large quantities is clinoptilolite, so most applications involve this mineral.

The cation-exchange capacity, cation selectivity, and molecular-sieve property of bulk zeolites are being applied to various aspects of agriculture. Some of the beneficial effects of using clinoptilolite are improved growth and feed utilization, improved eggshell quality in chickens, reduced free ammonia in the rumens of sheep and cattle, and reduced exposures to ammonia and odours in feed lots. Zeolites are used in soil conditioning to improve plant productivity, increase fertilizer efficiency, and aid remediation of soils. For example, clinoptilolite-bearing tuff is used to increase the yield of wheat in Azerbaijan. Zeoponics is a new method of plant growth, using artificial soil with zeolite as a major component. As in other agricultural applications, the zeolite helps to control soil solutions.

The selectivity of clinoptilolite for certain heavy metals is the basis for many applications in water and wastewater treatment. Examples are the removal of ^{137}Cs from contaminated waters resulting from the Chernobyl accident and the removal of copper from electroplating effluents.

The molecular-sieve property of many zeolites is used to separate gases. Phillipsite has been shown to have adsorption selectivity of ammonia over methane in gases from coal gasification. In Japan activated mordenite is used to increase the oxygen in air for steel making, and in the USA chabazite is used to separate carbon dioxide from methane in gas from a landfill site in southern California.

Where zeolitic tuff is abundant it has been used as a dimension stone at least since Roman times to construct castles and cathedrals in Italy and other countries of Europe. These tuffs are also used as a constituent of lightweight aggregate, as in the production of pozzolanic cement.

Although most zeolites used as catalysts are synthetic and made for specific applications, a few natural zeolites have also been tried. Natural ferrierite has been shown to be an effective material as a Claus tail-gas catalyst, to remove elemental sulphur from gas streams, and natural mordenite is effective in the hydroisomerization of n-hexane, with results similar to those of synthetic mordenite.

Glossary

Hydrothermal alteration Alteration of a host rock by hot water percolating through fractures, generally where the water-to-rock ratio is very large.

Metastable zeolites Those zeolites that crystallize quickly but are later replaced by more stable phases (zeolites or feldspars).

Oxygen bridge The oxygen anion between two tetrahedral sites.

Tetrahedral sites The space between four tightly held oxygen anions.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray); Mineral Analysis. **Diagenesis, Overview**. **Lava**. **Metamorphic Rocks:** Facies and Zones. **Minerals:** Definition and Classification; Feldspars; Quartz. **Pyroclastics**. **Tectonics:** Hydrothermal Activity.

Further Reading

Bargar KE and Beeson MH (1985) Hydrothermal alteration in Research Drill Hole Y 3, Lower Geyser Basin, Yellow stone National Park, Wyoming. Professional Paper 1054 C. Washington, DC: US Geological Survey.

Bish DL and Ming DW (2001) *Natural Zeolites: Occur rence, Properties, Applications*. Reviews in Mineralogy and Geochemistry 45. Washington, DC: Mineralogical Society of America.

Coombs DS, Alberti A, Armbuster T, *et al.* (1997) Recom mended nomenclature for zeolite minerals: report of the subcommittee on zeolites of the International Mineral ogical Association, Commission on New Minerals and Mineral Names. *Canadian Mineralogist* 35: 1571 1606.

Deer WA, Howie RA, Wise WS, and Zussman J (2004) *Rock Forming Minerals, Volume 4B Framework Sili cates: Silica Minerals, Feldspathoids and the Zeolites*. London: The Geological Society.

Dibble WE Jr and Tiller WA (1981) Kinetic model of zeolite paragenesis in tuffaceous sediments. *Clays and Clay Mineralogy* 29: 323 330.

Gottardi G (1989) The genesis of zeolites. *European Jour nal of Mineralogy* 1: 479 487.

Gottardi G and Galli E (1985) *Natural Zeolites*. Berlin: Springer Verlag.

Hay RL (1966) *Zeolites and Zeolitic Reactions in Sedi mentary Rocks*. Special Paper 85. Boulder: Geological Society of America.

Neuhoff PS, Fridriksson T, and Bird DK (2000) Zeolite para geneses in the North Atlantic Igneous Province: implica tions for geotectonics and groundwater quality of basaltic crust. *International Geological Review* 42: 15 44.

Sand LB and Mumpton FA (1978) *Natural Zeolites: Occurrence, Properties, Use*. Elmsford, New York State: Pergamon Press.

Sheppard RA and Gude AJ III (1973) *Zeolites and Associ ated Authigenic Silicate Minerals in the Tuffaceous Rocks of the Big Sandy Formation, Mohave County, Arizona*. Professional Paper 830. Washington, DC: US Geological Survey.

Sheppard RA, Gude AJ III, and Fitzpatrick JJ (1988) *Distri bution, Characterization, and Genesis of Mordenite in Miocene Silicic Tuffs at Yucca Mountain, Nye County, Nevada*. Bulletin 1777. Washington, DC: US Geological Survey.

Zircons

G J H McCall, Cirencester, Gloucester, UK

Introduction

The word 'zircon' is very old; the word in Arabic, *zarqun*, is said to derive from the original form of the word in Persian *zar* ('gold') and *gun* ('colour'). Zircon is an orthosilicate mineral with the formula $ZrSiO_4$. Its hardness on Moh's scale is 7.5 and it is slowly attacked by hot concentrated sulphuric acid. Its density is 4.6–4.7. Its fracture is uneven and may be conchoidal in metamict varieties. Its streak is white. Some of the properties of zircon are given in Table 1.

Structure

Zircon is tetragonal (Figure 1), with a unit-cell structure where $a = 6.61$ Å and $c = 5.99$ Å. X-Ray diffraction analysis shows that each zircon atom is surrounded by a tetrahedral group, with four oxygen atoms surrounding a central silicon atom (Figure 2).

Chemistry

Zircon always contains an amount of hafnium, with the HfO_2/ZrO_2 ratio varying. The amount is usually minute, but certain zircons from Norway contain up to 24% of HfO_2. Hafnian zircon occurs mainly in pegmatites. Zoned crystals of members of the zircon–hafnon series from Mozambique were described by Correia Neves and colleagues in 1974; these crystals ranged from hafnian zircon to zirconian hafnon to hafnon ($HfSiO_4$). Zircon is thus a source of hafnium and well as zirconium. Phosphorous may also replace silicon. Zircons often contain rare earth elements and may be metamict (amorphous). Yttria, which is commonly present, is related to the isostructural nature of xenotime (YPO_4) to zircon. Some metamict zircons are mixtures of silica, cubic ZrO_2, and baddeleyite (monoclinic ZrO_2). Uranium and thorium are present in zircons from granites, a property that makes zircons valuable for radiometric age determinations. Chemical analyses of zircons from two geographical locations are given in Table 2.

Table 1 Properties of zircon

Property	*Description*
Refractive index	Optically positive ω 1.923 1.960 ε 1.961 2.015
Birefringence	δ 0.042 0.065
Dispersion	Very strong D
Cleavage	{110} imperfect, {111} poor
Twinning	Rare, on {111}; some zoning may occur
Colour	Reddish brown, yellow, grey, green, colourless; in thin section, colourless to pale brown; transparent to translucent, vitreous to adamantine
Pleochroism	Very weak; in thick sections, may show absorption, $\omega > \varepsilon$
Unit cell	*a* 6.61, *c* 5.99; A type; *Z* 4; space group $I4_1/amd$

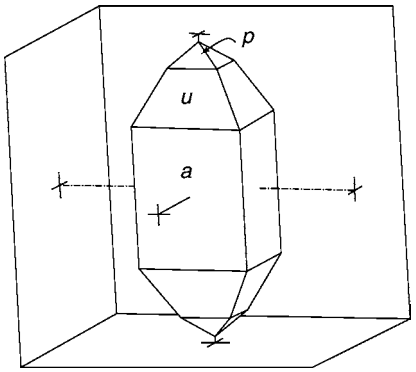


Figure 1 Sketch of a typical zircon crystal, showing crystal faces *a* (100), *p* (101), and *u* (301). Reproduced from Berry LG and Mason B (1959) *Mineralogy Concepts, Descriptions and Determinations*. San Francisco: WH Freeman.

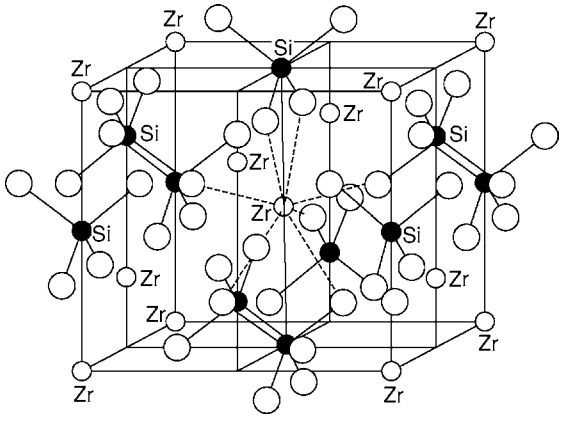


Figure 2 The structure of zircon. The eight bonds from each zircon atom to neighbouring oxygen atoms are shown only at the centre of the figure. Reproduced from Deer WA, Howie RA, and Zussman J (1962) *Rock Forming Minerals*, 2nd edn, vol. 1A. London: Longmans.

Occurrence

Zircon, a commonly occurring accessory mineral in igneous rocks, is fairly common as a detrital mineral in sediments and is also found in metamorphic rocks. It is always present in ancient shoreline concentrations of heavy minerals ('black sands'); on the coasts of Queensland–New South Wales and Western Australia (Capel, Yoganup, and Eneabba), it is typically subordinate to ilmenite and rutile.

Gemstone Zircon

Mined for a very long time for gemstone purposes in Ceylon (Sri Lanka), zircon did not really become popular until the 1920s. It is perhaps the most variable of the gemstone minerals, and its legendary mystical qualities are now known to be due to the contained radioactive elements. Grains in alluvia and heavy mineral sand deposits are commonly rounded by erosion and do not possess the full crystalline shape, short to long columnar prisms with bipyramidal terminations. Zircon has the second highest index of refraction after diamonds; the dispersion gives it an intensive lustre and cut stones have a vivid fire. Though zircon is quite hard, it is brittle and it is susceptible to damage by scratch marks. The radioactive elements in zircon do with time destroy the internal structure, producing the metamict forms, and such crystals are of lower than gemstone quality. Zircons of this type are classified as 'low quality'. Those not metamict are classified as 'high quality', and there may be varieties classified as 'intermediate'. Table 3 shows how the optical properties vary systematically with the amount of radioactivity. Heating up to 800–1000°C transforms minerals of intermediate and low quality to high quality, with colour changes, and the colour varies according to whether an oxidizing or reducing atmosphere is present during the heating. Pale blue 'starlite' and straw-yellow 'jargoons' are popular. Colourless, lucid zircon gemstones have been incorrectly termed 'Matura diamonds'. Artificially produced zircons may change colour over time, either growing pale or returning to their original hue. Gem-quality zircons never occur as very large crystals and are mined from placer deposits in Sri Lanka, Cambodia, and Thailand. Some natural gem-quality zircons are shown in Figure 3. Zirconia is an artificial imitation of diamond, a double oxide of zirconium and yttrium.

Table 2 Two Chemical analyses of zircon

Mineral	*Sample 1 (%)*[a]	*Sample 2 (%)*[b]
SiO_2	32.51	31.52
ZrO_2	67.02	64.69
HfO_2		1.31
TiO_2		0.27
Al_2O_3	0.21	0.27
Fe_2O_3	0.08	0.43
RE_2O_3[c]	0.04	0.20
MgO	0.01	0.57[d]
CaO	0.22	0.04
ThO_2		0/04
Nb_2O_5		
UO_3		0.05[e]
P_2O_5		0.19
H_2O+	0.03	0.03
H_2O		0.11
Total	100.12	100.23

[a]From North Burgess, Ontario, Canada.
[b]From South Africa.
[c]RE_2, rare earth elements.
[d]Y_2O_3.
[e]U_3O_8.

Zircon as an Economic Heavy Mineral

Zircon occurs typically in placer deposits (Table 4). The placers in which zircon occurs are commonly known as 'black sands'.

Beach Placers

The most important minerals in beach placers are cassiterite, diamond, gold, ilmenite, magnetite, monazite, rutile, xenotime, and zircon. Black sand deposits commonly contain ilmenite, monazite, rutile, xenotime, and zircon, all valuable minerals. Ilmenite may be dominant, as in Travencore, India, and Capel/Yoganup, Western Australia, but elsewhere rutile is dominant, as in the deposits of Queensland and Eneabba, Western Australia. Recent marine placers occur at different levels, owing to Pleistocene sea-level changes, and they may be many kilometres inland, as at Eneabba. The tidal zone and wave-cut terraces are optimum situations for such deposition.

Table 3 Variation in optical properties with amount of radioactivity

Classification	*D*	*ω*	*ε*	*δ*	*Radioactivity*
Normal zircon	4 6 4 71	1 924 1 934	1 970 1 977	0 036 0 053	Low
Intermediate	4 2 4 6	1 903 1 927	1 921 1 970	0 017 0 043	Medium
Metamict zircon	3 9 4 2	1 782 1 864	1 827 1 872	0 0 008	High

Figure 3 Gem quality zircons. (A) Zircon crystal in host rock. (B) Hyacinth zircon from Sri Lanka. (C) Cut zircons from Thailand. (D) Cut zirconia. Reproduced with permission from Bauer J and Bouska V (1983). *A Guide in Colour to Precious and Semi precious Stones*, pp. 108 109. London: Octopus Books.

Table 4 Classification of placers

Mode of origin	*Class*	*Alternative classification*
Accumulation *in situ* during weathering	Residual placers	Eluvial
Concentration in a moving solid medium	Eluvial placers	Colluvial
Concentration in a moving liquid medium (water)	Stream or alluvial placers, beach placers, offshore placers	Fluvial, strandline, marine placers
Concentration in a moving gaseous medium (air)	Aeolian placers	Desert aeolian, coastal aeolian

In Eastern Australia, these deposits of rutile and zircon occur over a stretch of 900 km in a strip 13 km wide, the thickness of the deposits averaging 30-40 m. Typical sites of beach placer deposition are shown in Figure 4. Deposition is by wave and shore current action. Preservation of beach deposits of heavy minerals would seem to require either progradation or sea-level fall, to shield the deposits from further erosion. The two shorelines in south-west Western Australia are shown in Figure 5, and cross-sections of the deposits there are shown in Figure 6 (Eneabba and Yoganup).

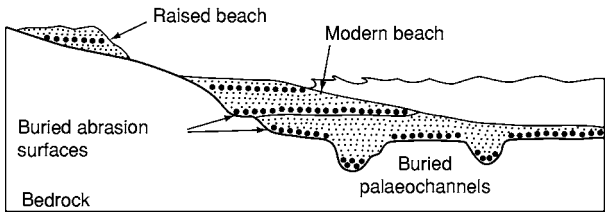


Figure 4 Typical sites of deposition of beach placer deposits. Reproduced with permission from Evans AM (1993) *Ore Geology and Industrial Minerals*, 3rd edn. Oxford: Blackwells Science Publications.

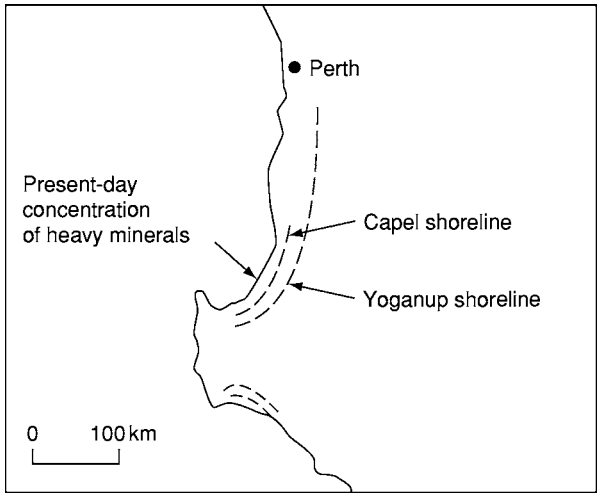


Figure 5 Sketch of the coastline in south west Western Australia, showing the two shorelines of the younger (outer) Capel deposit and older (outer) Yoganup deposit of heavy minerals. The Eneabba deposits are about 300 km north of Perth and include an older beach deposit, about 25 km inland from the present coast, as in the case of the Yoganup deposit, and aeolian dune concentrations in the Jennings part of the Eneabba deposits just to the east of the beach placers. The Eneabba deposits are dominated by rutile whereas ilmenite is dominant in the south. Reproduced with permission from Evans AM (1993) *Ore Geology and Industrial Minerals*, 3rd edn. Oxford: Blackwells Science Publications.

Aeolian Placers

Aeolian placers are formed by reworking of the beach placers by wind. The coastal dunes that are created in this way may have a great volume and may be an important economic source for extraction of heavy minerals, even if of low grade. Three types of dunes may contain mineral deposits: foredunes, transgressive dunes that have moved inland from the beach area, and cliff-top dunes. Aeolian placers may be minor and subsidiary to large beach placer deposits, but they may be large and contain deposits in their own right. The Jennings deposit at Eneabba appears to have formed as a cliff-top dune, whereas the deposits just to the west of it are beach placer deposits. High-grade deposits of rutile and zircon occur in dunes at Richards Bay, Natal.

World Production

Australia is by far the biggest producer of zircon, with 70% of the world's supply; zircon deposits are found in other regions of the world (Figure 7), primarily in South Africa. World production of zirconium mineral concentrates is shown in Table 5.

Mining and Processing

Heavy mineral sands are extracted by dredge (Figure 8), bucket-wheel excavators or draglines, and bulldozers. The raw material is then processed, first by gravity concentrators, to remove light minerals; the concentrate is scrubbed and dried, ilmenite, leucoxene, and rutile being removed by magnetic and high-tension techniques (Figure 9). The remaining minerals are then separated by a system of spirals, gravity tables, high-tension separators, magnets, and air tables to produce zircon. The complex process requires large plants on site: the different minerals of economic interest are separated at the plant location because transport of the raw mined material would be prohibitively costly. Mining these deposits is usually done under a requirement to restore the land, and at Capel in Western Australia, the land is restored to agricultural use.

Uses

The production of zirconia (ZrO_2) for ceramic refractory purposes is the main usage of zircon, but about 10% has a variety of other uses, in ferro alloys, paints, pharmaceuticals, and abrasives. Zircon is also used in leather tanning, chemical manufacture, and in nuclear industries, the latter requiring high-purity zirconium metal and extraction of the ~2% of hafnium contained in the zircon. There are foundry uses and alloy uses, including use in welding rods and sandblasting. The hafnium contained in the zircon is used in unalloyed form in nuclear reactors. Hafnium carbide is also one of the most refractory materials known.

Zircon Analyses

Zircon has a unique status as a key mineral for placing the evolution of Earth within a clear framework of precisely measured time.

Radiometric Geochronology

Before 1896, when Henri Becquerel discovered radioactivity, estimates of the ages of minerals, rocks, and Earth were dependent on theological dogma, outright

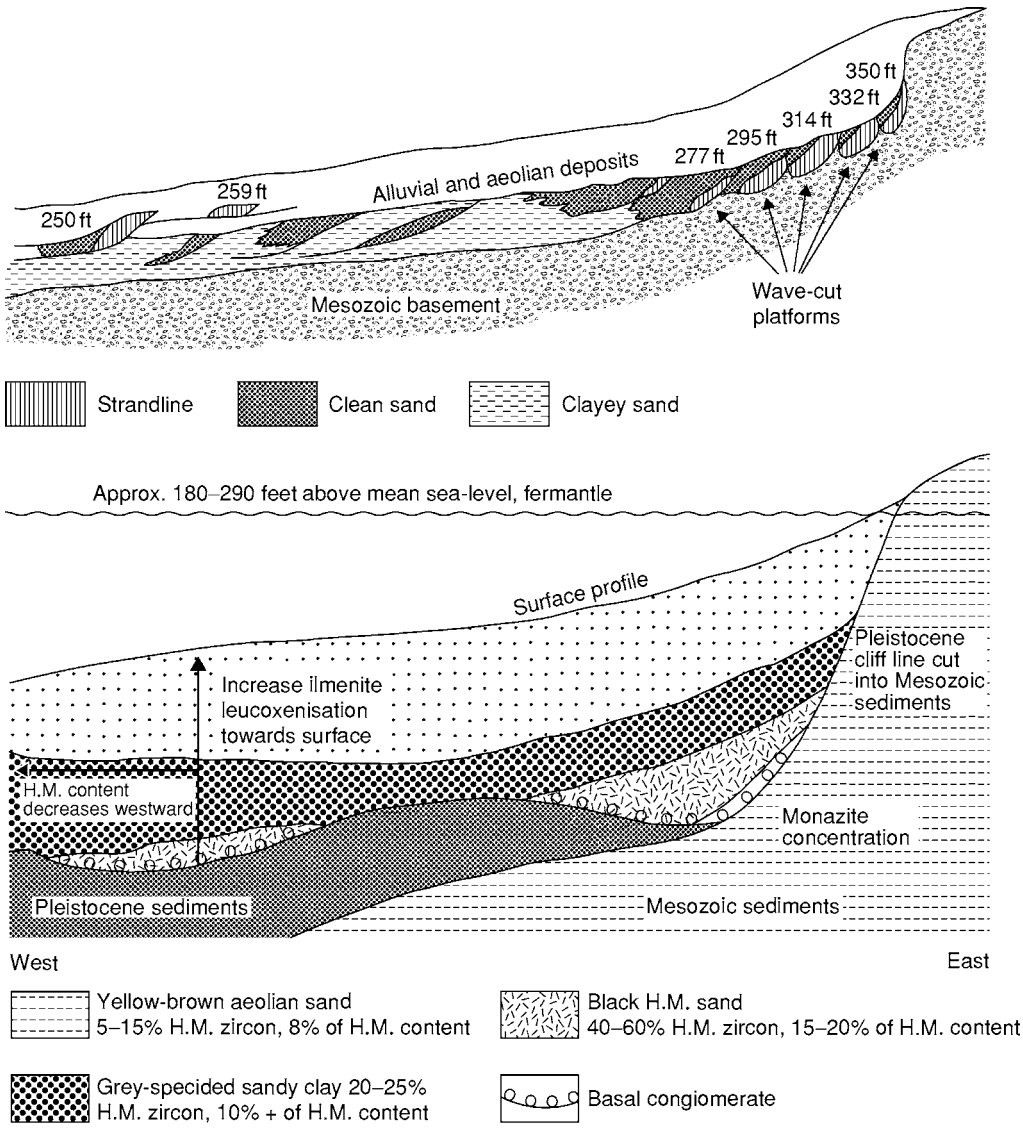


Figure 6 Cross section of the Eneabba (top) and Yoganup (bottom) beach placer deposits. H.M., Heavy mineral. Horizontal scale, 1:120; vertical scale, 1:24. Reproduced with permission from Evans AM (1993) *Ore Geology and Industrial Minerals*, 3rd edn. Oxford: Blackwells Science Publications.

speculation, or crude physical measures (e.g., estimates based on ocean salinity). By 1903, Marie and Pierre Curie had been awarded the Nobel Prize in physics for showing that a radioactive parent element decays to various daughter elements at a constant rate, and this laid the foundations for precise physical measurement of mineral and rock ages. Despite Wilhelm Rutherford realizing in 1905 that the decay of U to He and Pb provided a powerful technique for this purpose, precise dating methods were slow to develop, awaiting the discovery of isotopes by JJ Thomson and the development of mass spectrometric techniques in the 1920s by Francis Aston and Alfred Nier. It was not until the 1940s that that a wide variety of decay schemes were tested in succession (U–Pb, K–Ar, K–Ca, Rb–Sr, Lu–Hf, Re–Os, and Sm–Nd). The basic requirements are (1) that the radioactive parent element is reasonably abundant geologically, (2) that, in the sample, the daughter element remains bound with the mineral in which it formed over time, and (3) that the latter can be distinguished reliably from any concentration of the same element that may have been originally present in the mineral or rock.

Zircon Samples

Zircon forms an ideal sample material with respect to the U–Pb decay scheme, because it is an abundant

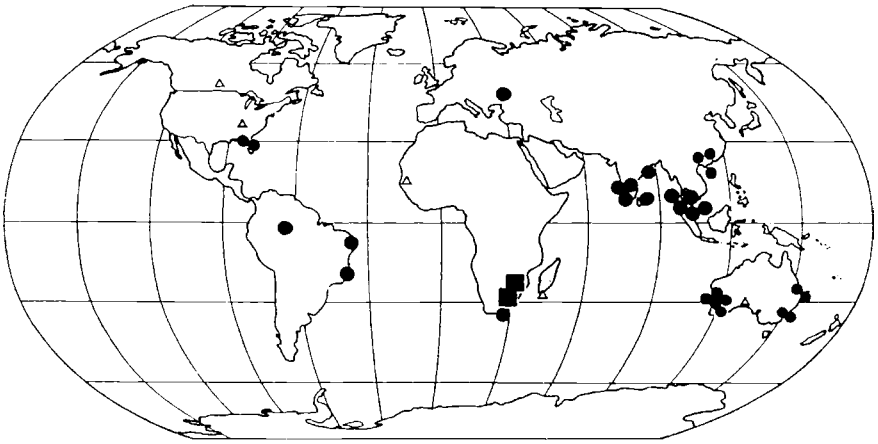


Figure 7 Sites of zirconium deposits throughout the world: (●) active zircon mines; (■) active baddeleyite mines; (△) deposits not yet worked. Reproduced with permission from Cary DD (ed.) (1994) *Industrial Minerals and Rocks*, 6th edn. Littleton, CO: Society for Mining, Metallurgy and Exploration.

Table 5 World production of zirconium mineral concentrates by year[a]

Country	*1986*	*1987*	*1988*	*1989*	*1990*
Australia	451 824	456 590	480 049	511 000	442 000
Brazil	15 116	18 131	28 029	32 970	33 000
China	15 000	15 000	15 000	15 000	15 000
India	10 000	15 000	17 000	18 000	18 000
Malaysia	12 633	17 828	25 671	18 704	4 279
South Africa, Republic of	140 000	140 000	150 000	180 000	180 000
Sri Lanka	4 000	4 000	3 000	3 000	3 000
Thailand	1 705	1 532	5 098	1 496	2 000
Former USSR	85 000	90 000	90 000	90 000	90 000
United States	W	W	117 606	118 388	102 073
Total	741 278	753 081	929 453	987 758	889 352

[a] W, Withheld; *Source: Minerals Yearbooks*, US Bureau of Mines.

accessory in a wide range of igneous, metamorphic, and epiclastic/pyroclastic sedimentary rocks. Its structure is such that it normally does, at the time of crystallization, accommodate trace quantities of U at levels sufficient for precise analysis, and the daughter Pb tends to be retained tightly. The double isotopic decay of ^{238}U to ^{236}Pb and ^{237}Pb provides a check on any loss of radiogenic Pb. Early on, in the 1950s, the method required milligram quantities (thousands of grains) of zircon, which were separated into fractions of varying magnetic susceptibility, dissolved in acid, and isotopically analysed by conventional thermal ionization mass spectrometry (TIMS) in ion-exchange columns. Improved techniques later allowed individual zircon crystals or parts of them to be analysed.

Sensitive High-Resolution Ion Microprobe Technique

In the 1980s, the sensitive high-resolution ion microprobe (SHRIMP) technique was successfully developed at the Australian National University, Canberra, based on secondary ion mass spectrometry (SIMS). In SHRIMP analysis, a focused beam of oxygen ions sputters the surface material of a mounted and polished zircon crystal into a high-voltage field, which accelerates a complex mix of ions through an electrostatic separator and then through a conventional magnet, as in the TIMS instrument. Experimentally difficult and time-consuming stages of dissolution and ion-exchange are eliminated. SHRIMP analysis has the advantages of being faster than TIMS analysis, is essentially non-destructive, and allows analysis of very small spots ($\sim$20 μm in diameter). A disadvantage is that SHRIMP analysis requires analysis by TIMS of a standard natural zircon crystal with perfect internal homogeneity and large enough for the crushed fragments to be supplied to and used by different laboratories. In practice, finding such a standard has proved difficult. In addition, the physics of the sputtering process is poorly understood and there may be an in-built bias in isotope ratios derived.

The internal complexities of zircons have been studied in great detail using SHRIMP, particularly using electron backscatter images. Age can vary significantly within a single zircon crystal. Euhedral prismatic zircons from granites commonly contain rounded cores inherited from source rocks, and high-grade orthogneisses contain patchy areas that yield

Figure 8 A dredge working on a deposit of rutile and zircon on Stradbroke Island off the Queensland coast. Such large and complex installations require capitalization running into millions of dollars. Photograph by GJH McCall.

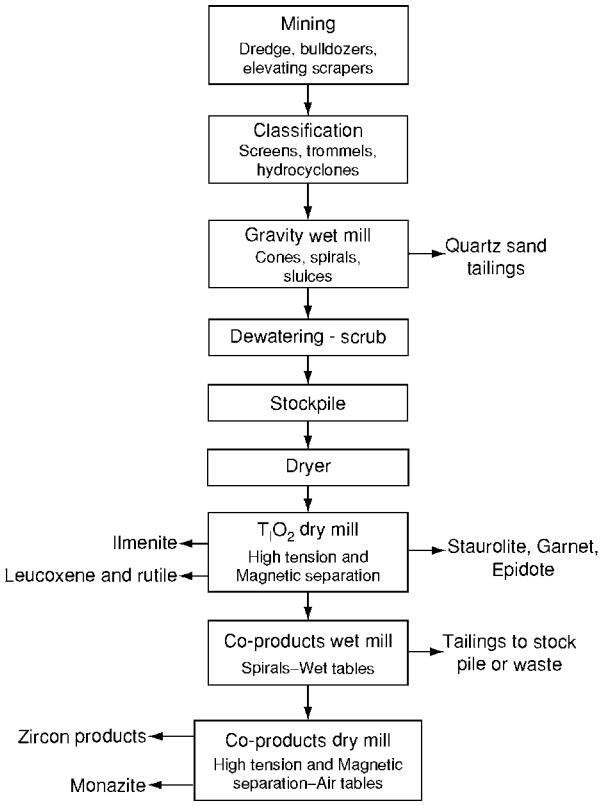


Figure 9 Simplified flow diagram showing the stages in process ing a zircon bearing placer deposit. Reproduced with permission from Cary DD (ed.) (1994) *Industrial Minerals and Rocks*, 6th edn. Littleton, CO: Society for Mining, Metallurgy and Exploration.

reset metamorphic ages, rather than the original igneous age. Such ages may have no geological meaning, but development of a combination of TIMS and SIMS methods, supported by field studies, goes beyond the

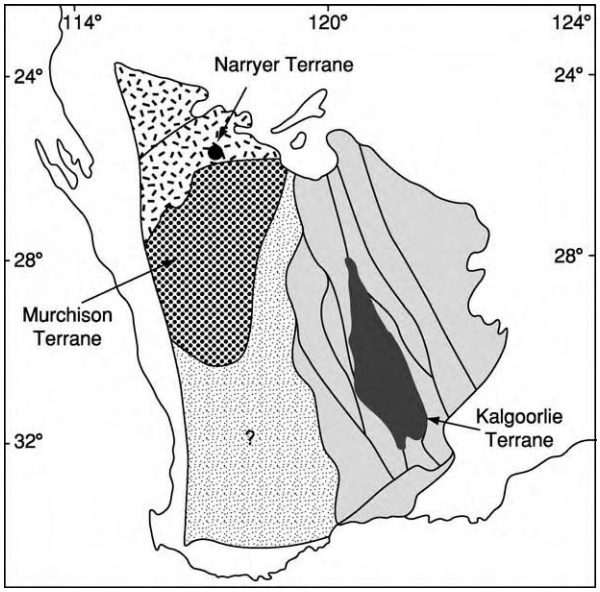


Figure 10 Location of the Narryer Gneiss Complex and Jack Hills, Western Australia.

limits of zircon geochronology based on igneous rocks, allowing complex metamorphic histories and the source rocks of sedimentary basins to be deciphered. This is a field of research that is really only just opening up.

Amelin and co-workers reported in 2001 on detrital zircons aged between 3.90 and 4.25 million years; these zircons came from the Jack Hills within the Narryer Gneiss Complex, Western Australia (Figure 10). The host rock is a metaconglomerate. Separation for SHRIMP analysis was possible for a large number of ancient grains. The experiments were carried out on SHRIMP-II at the Geological Survey of Canada, Ottawa, on 605 detrital grains from sample W74 of the metaconglomerate. Ages between 4000 and 4100 million years were found for 86 grains and 17 gave ages between 4100 and 4250 million years. Wilde and colleagues have found two zircon grains with an age of 4404 ± 8 million years using the same sample. Some grains are uniform but others show oscillatory zoning. Analysis on some spots within crystals gave no significantly different values. Some grains have visible uniform rims overgrown on zoned zircon, but, again, there is no discordance in age. A few had younger rims (3580–3650 My) and several 4100- to 4150-My-old grains have 4100- to 4150-My-old rims. These ancient zircons are not only the world's oldest dated minerals, they preserve a record of the terrestrial crust formed 100–400 My earlier than the oldest known rocks. The earliest known evidence for liquid water and a terrestrial hydrosphere comes from the 3.85-Gy-old sediments (including

banded ironstones) of south-west Greenland. The evidence derived from the Mount Narryer zircons (including oxygen isotope ratios) strongly implies that water existed on Earth at 4.4 Ga. Life may even have developed by this time, within the Hadean.

Overgrowths on Zircons

Dating of sedimentary rocks by radiometric methods has long been a difficult matter and has been achieved only to broad limits by dating igneous intrusions into them or concordant lava flows or tuffs close to the sedimentary stratum concerned. Authigenic overgrowths on zircon, a mineral that is formed at high temperature and is insoluble in groundwater, have long attracted attention. They are composed of xenotime, essentially yttrium phosphate (Figure 11). Such overgrowths occur in the Millstone Grit (Carboniferous, Namurian) of northern England. Though diagenetic glauconite has been radiometrically dated, it suffers from the defect that it is prone to postdepositional alteration. The SHRIMP technique has recently been applied to such xenotime outgrowths (xenotime contains traces of radioactive elements). The outgrowths have been shown to precede all other authigenic minerals such as quartz, and are thus thought to be very early diagenetic deposits. The xenotime is precipitated from seawater with other biogenic phosphates shortly after burial, in a zone of high bacterial activity and sulphate reduction. Xenotime outgrowths thus provide a method of obtaining what is virtually the age of deposition of a sedimentary rock, by radiometric methods. The method can be applied to rocks of any geological age, to date strata directly. This new method is especially important in the case of sedimentary rocks devoid of fossils.

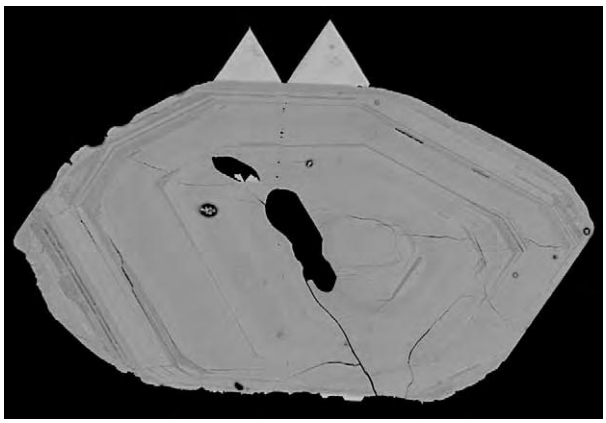


Figure 11 Two xenotime outgrowths on a zircon crystal. Reproduced from Glover J (2003) *Geological Journeys from Artifacts to Zircon*, pp. 105 106. Perth: Geological Society of Australia, Western Australian Division.

See Also

Analytical Methods: Geochronological Techniques. **Mineral Deposits and Their Genesis**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings; Shoreline and Shoreface Deposits.

Further Reading

Amelin Y, Stern R, Wiechert U, Davis D, Lee D, Halliday A, and Pidgeon R (2001) Combined U Pb, trace element, oxygen, Zr and Lu Hf isotopic systematics of 3.90 4.25 Ga detrital zircons from Jack Hills, Western Australia. In: Cassidy KF, Dunphy JM, and Van Kranendonk MJ (eds.) *4th International Archaean Symposium Extended Abstracts AGSO Geoscience Australia Record 2001/37*, pp. 33 34. Perth, WA: Geoscience Australia.

Bauer J and Bouska V (1983) *A Guide in Colour to Precious and Semi precious Stones*, pp. 108 109. London: Octopus Books.

Berry LG and Mason B (1959) *Mineralogy Concepts, Descriptions and Determinations*. San Francisco: WH Freeman.

Cary, DD (ed.) (1994) Industrial Minerals and Rocks, 6th edn. Littleton, CO: Society for Mining, Metallurgy and Exploration.

Correia Neves JM, Lopez Nunes JE, and Sahama ThG (1974) High hafnium members of the zircon hafnon series from the granite pegmatites of Zambezia, Mozambique. *Contributions to Mineralogy and Petrology* 48: 73 80.

Deer WA, Howie RA, and Zussman J (1982) *Rock Forming Minerals*, 2nd edn, vol. 1A, pp. 418 442. London: Longmans.

Evans AM (1993) *Ore Geology and Industrial Minerals*, 3rd edn. Oxford: Blackwells Science Publications.

Glover J (2003) *Geological Journeys from Artifacts to Zircon*, pp. 105 106. Perth: Geological Society of Australia, Western Australian Division.

Harben PW and Bates BC (1984) *Geology of the Non metallics*. New York: Metal Bulletin Inc.

Mitchell RS (1973) Metamict minerals: a review. *Mineral Record* 4: 177 182.

Wilde SA, Valley JW, Peck WH, and Graham CM (2001) Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Ga ago. *Nature* 409: 175 178.

Wilde SA, Valley JW, Peck WH, and Graham CM (2001) Evidence for the early growth of continents and oceans from >4 Ga detrital zircons. In: Cassidy KF, Dunphy JM, and Van Kranendonk MJ (eds.) *4th International Archaean Symposium Extended Abstracts* AGSO *Geoscience Australia Record 2001/37*, pp. 106 108. Perth, WA: Geoscience Australia.

MINING GEOLOGY

Contents

Exploration Boreholes

M Vaněček, Charles University Prague, Prague, Czech Republic

Introduction

Exploration boreholes are the primary tools used in the preliminary investigation of coal and industrial mineral deposits. Complex ore deposits are studied by means of boreholes in the initial stages of prospecting and exploration. The most frequently used technique used is core drilling. The reliability of the results of exploration drilling depends on core recovery, which is influenced by choice of method and equipment (e.g., type of flushing, use of special core barrels, bit design). Correct interpretation depends also on correct assessment of the borehole inclination.

Borehole Drilling

The drilling of boreholes for mineral exploration is practised with the aim to locate mineral deposits, to define their extent and reserves, to determine hydrogeological and engineering geological conditions, and to take samples for laboratory investigations (chemical assays, dressing characteristics, etc.). Drilling intercepts yield satisfactory data in the case of mineral deposits that are not very complex, such as sedimentary deposits and other deposits forming more or less homogeneous mineral bodies. Exploration of homogeneous mineral bodies by means of drilling is usually made in square or other regularly shaped grids. If any feature of the mineral body has an anisotropic distribution, the exploration grid should be rectangular, with shorter distances between boreholes orientated in the direction of the changing feature. Where ore deposits dip, the exploratory boreholes are arranged in profiles perpendicular to the strike of the mineral body. The entire ore deposit, however, cannot be explored in detail without complete mining operations.

Rotary core drilling is the most commonly used drilling technique for creating exploration boreholes. Drilling machines rotate an annular bit to cut a narrow kerf around a central core and thus obtain unaltered samples of the drilled rocks. The core is cut by a tungsten carbide or diamond-studded drill bit and is collected in the core barrel (**Figures 1** and **2**). When the depth of the ore body is relatively shallow, the core barrel containing the cut core is retrieved by pulling the entire length of the drill stem. If the target interval is deep and a considerable number of drill

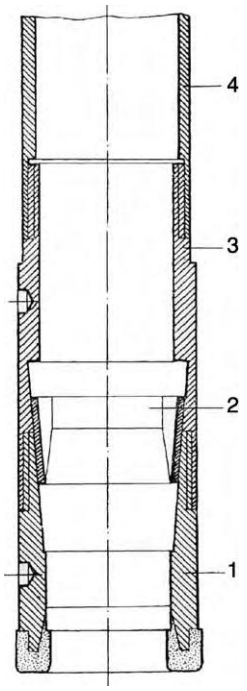


Figure 1 Conventional core drilling string. 1, Diamond bit; 2, core breaker; 3, core breaker tube; 4, core barrel.

Figure 2 Examples of diamond drill bits for exploration drilling. Photograph by Atlas Copco.

Figure 3 Underground exploration drilling (Simba H1354 for Echo Bay, Canada). Photograph by Atlas Copco.

stem rods are involved, then a wire–line system is more feasible. In this case, double-tube core barrels are used; the inner barrel is removed from the bottom of the borehole by means of wire and is thus retrieved without having to bring the drill stem rods to the surface. Use of the wire–line system is limited, depending on the hardness of the rocks being drilled. If the rock is very hard, drill bits may be destroyed in a short time and the entire drill stem will have to be retrieved frequently.

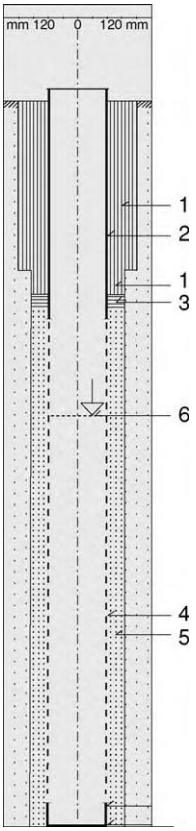


Figure 4 Detail of the hydrogeological borehole casing. 1, Cementation; 2, full casing; 3, sand; 4, perforated casing; 5, quartz pebbles (graining, 4 8 mm); 6, underground water level detected.

Special drilling machines have been designed for drilling exploration boreholes from locations below the surface (Figure 3). In comparison with surface-drilled boreholes, which may extend for some hundreds of metres, the length of subsurface-drilled boreholes usually does not exceed 50–100 m. These boreholes are drilled, often from one location, with different inclinations, from vertical to horizontal. Geological cross-sections are then based on results of fans of boreholes.

Assessment of mining conditions for mineral deposit development or extraction sometimes requires special hydrogeological approaches to exploration drilling. In these situations, the borehole casing is perforated at the underground water horizons (see Figure 4). The boreholes can be drilled using any of non-coring techniques (rotary drilling, percussion drilling, rotary–percussion drilling, etc.).

Core Recovery

One of the main problems encountered in exploration drilling is core recovery. There is a general expansion

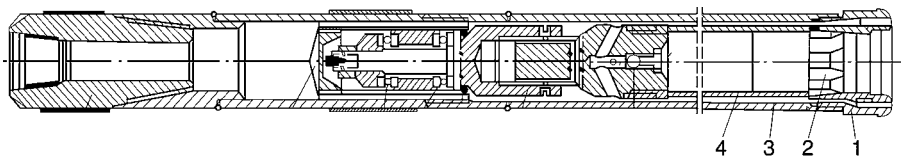


Figure 5 Double tube core barrel for exploration of coal deposits. 1, Diamond bit; 2, core breaker; 3, outer tube of the core barrel; 4, inner tube of the core barrel.

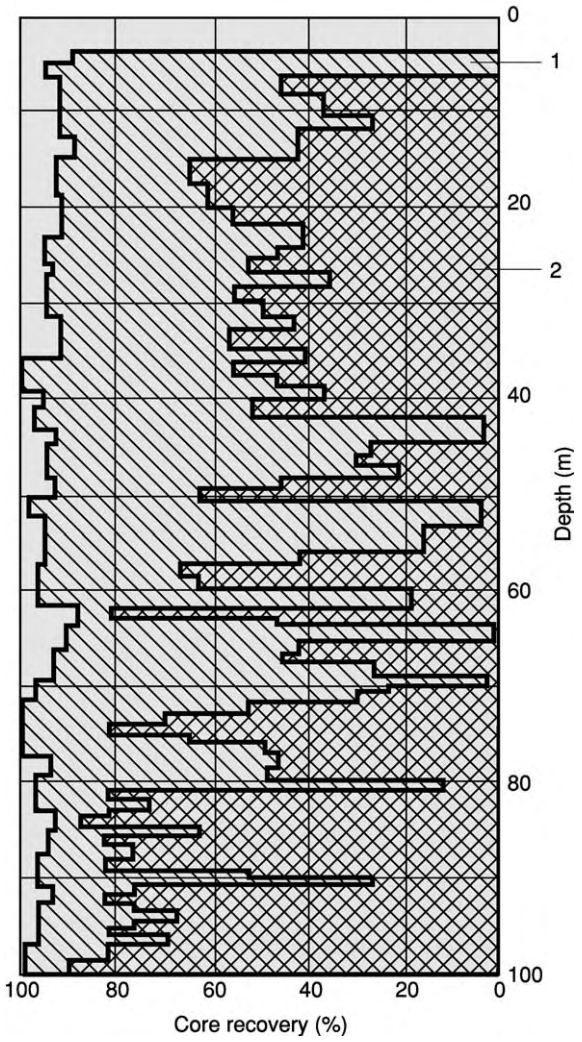


Figure 6 Comparison of the core recovery from two boreholes drilled in identical geological conditions. 1, Drilled using benton ite mud flushing; 2, drilled using water flushing. Modified with permission from Tkaný Z, Juránek O, and Brych J (1966) *Deep Diamond Drilling* (in Czech). Prague: Technical Literature Publ. House (SNTL).

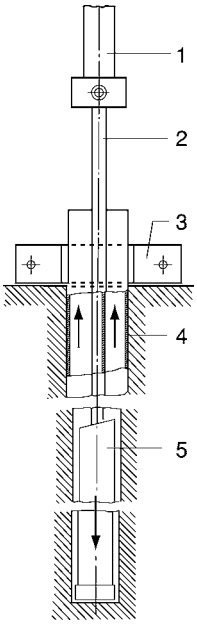


Figure 7 Detail of direct flushing tubular drill rod. 1, Drill head; 2, drill stem; 3, casing clamp; 4, casing; 5, core drilling string.

of core sample length on removal of the sample from the core barrel, related to fracturing of rocks during the drilling process. If soft rock intercalations exist in the borehole location, the core sample is likely to be lost. Because it may not be possible to determine the presence or position of soft rock prior to drilling, use of double-tube (Figure 5) or even triple-tube core barrels may increase the chances of obtaining 100% core sample recovery.

Core recovery may also depend on the type of flushing medium used. The effect of using water versus bentonite mud as the circulation fluid is shown in Figure 6. Circulation media (water, clay mud, or air) may be pumped into drill holes through the inner part of a tubular drill rod (Figure 7); in counterflush drilling, media are injected into drill holes through the annulus between the wall casing and the drilling pipe or, more often, between the inner and outer wall pipes (see Figure 8). Counterflush drilling, also called reverse-circulation drilling, has the distinct advantage over normal rotary drilling of providing relatively complete and uncontaminated samples. Though the method is considerably slower and more expensive than normal rotary drilling is, it is used increasingly in contemporary explorations of mineral deposits. Costs incurred in reverse-circulation drilling depend on geological formation hardness.

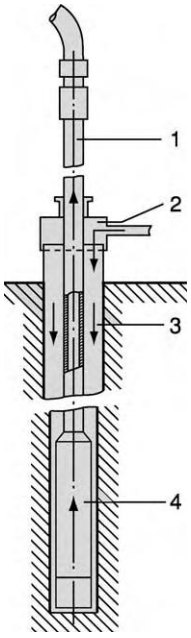


Figure 8 Detail of reverse circulation flushing drill. 1, Drill stem; 2, flush head; 3, casing; 4, core drilling string.

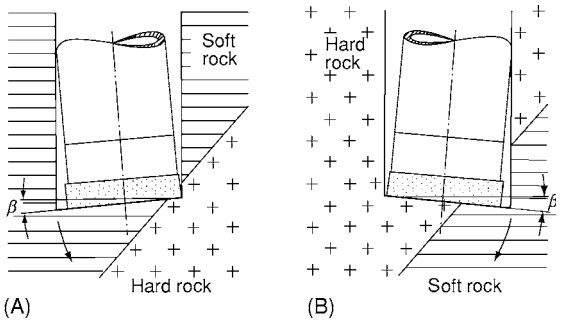


Figure 9 Change in the borehole inclination at the boundary between two layers. (A) Transition from soft to hard rocks. (B) Transition from hard to soft rocks.

Correct interpretation of exploration drilling results requires knowledge of the course of the drilled boreholes, which also relates to rock hardness. For example, if a drill bit reaches a boundary between two layers with different levels of hardness, the inclination of the borehole may change (see Figure 9) and the hole may become crooked and off-course.

Drill Bits

Drill bits of different geometry and design are used for core drilling. Surface-set bits contain a single layer of big diamonds, layered bits are made of diamonds of medium size and small size, and impregnated bits have diamond fragments distributed irregularly throughout the copper matrix of the crown. Core bits with tungsten carbide are made either with tungsten carbide inserts or with cutting edges hardened by welded-on grains of tungsten carbide. Non-coring diamond bits may be used when a core is not desired. The lack of core may be compensated by geophysical logging (recording changes in lithology, resistivity, temperature, porosity, hole diameter, etc., using special probes).

See Also

Engineering Geology: Site and Ground Investigation; Ground Water Monitoring at Solid Waste Landfills. **Geological Maps and Their Interpretation**. **Mining Geology:** Exploration; Mineral Reserves; Hydrothermal Ores; Magmatic Ores.

Further Reading

Anonymous (1963) *Symposium on Exploration Drilling*. Golden: Colorado School of Mines.

Bender F (ed.) (1986) *Angewandte Geowissenschaften*, vol. IV. Stuttgart: Enke Verlag.

Clark GB (1982) *Principles of Rock Drilling and Bit Wear. Colorado School of Mines Quarterly*, vol. 77(1 and 2). Golden: Colorado School of Mines.

Compton RR (1962) *Manual of Field Geology*. New York: Wiley.

Davenport B (1992) *Horizontal and Vertical Drilling*. New York: Osborne McGraw Hill.

Finkl CW Jr. (ed.) (1988) *Encyclopedia of Field and General Geology*, vol. XXV. New York: Van Nostrand Reinhold.

Kennedy JL (1983) *Fundamentals of Drilling: Technology and Economics*. Tulsa: PenWell Books.

Kužvart M and Boehmer M (1986) *Prospecting and Explor ation of Mineral Deposits*, 2nd edn. Prague: Academia.

Lahee FH (1961) *Field Geology*, 6th edn. New York: McGraw Hill.

Tkaný Z, Juránek O, and Brych J (1966) *Deep Diamond Core Drilling* (in Czech). Prague: Technical Literature Publ. House (SNTL).

Exploration

N C White, Brisbane, QLD, Australia

Introduction

Mineral exploration aims to discover deposits of minerals and rocks that can be used to meet the resource needs of society. It encompasses the search for industrial raw materials (e.g., clay, limestone, sulphur, salts, and fertilizer minerals and rocks), ores from which metals are extracted (e.g., iron, copper, and zinc ores), and gemstones (diamonds, sapphires, and opals), and includes the search for solid fuels (coal, oil shale, and uranium) but not liquid or gaseous fuels (petroleum and natural gas). Mineral exploration can be as basic as prospecting, using elementary techniques such as panning for gold, or it can be very sophisticated, involving the use of complex technology for data gathering and interpretation. This article considers our need for mineral resources, why different groups explore for minerals and the strategies they adopt and the tools they employ, what success and failure mean in the industry, the role of governments, and the future of exploration.

Our Need for Minerals

Human beings have always been dependent on natural materials to satisfy the needs of everday life. All these materials must come from Earth, either grown as plants or animals or extracted as rocks, minerals, or fossil fuels. Some of the rocks and minerals that are important in everyday life are readily available, but many are not, and we must search for them: that search is what exploration is about.

The importance of rocks and minerals to humankind is conveyed by terms such as 'stone age', 'bronze age', 'iron age'; the emphasis on stages in the evolution of technology and civilization is related to our mastery of different rocks and our understanding of the metals that can be extracted from them. In modern civilization, there is amazing diversity in the rocks and minerals we use and in the variety of ways in which we use them. Some mineral resources are used directly, with little processing apart from shaping, polishing, crushing, or sizing (e.g., rocks used for building, aggregate used for road-making, talc used in talcum powder, or gemstones used in jewellery). Many mineral resources are consumed in industrial processes that transform them into new products (e.g., clays used to make bricks or ceramics, silica sand used for glass-making, limestone used to make lime or cement, or phosphate rock used for superphosphate fertilizer). The large group of rocks we call ores (e.g., iron, copper, and gold ores) are important because of the metals they contain. Other rocks and minerals are used in industrial processes because they aid the processing of other materials, not because anything will be extracted from them (e.g., many rocks and minerals are used as flux in smelting).

Who Explores for Minerals, and Why?

The motivation for mineral exploration is the same as for many other human activities: either we need the materials we seek or, by satisfying other people's needs, we can earn money. Though exploration is necessary to maintain the supply of essential raw materials, individuals or corporations undertake exploration for a variety of reasons, but underlying them all is the belief that there will be a good commercial return from what is discovered. The commercial return may come from selling the deposits that are discovered, from setting up a company based on the discovery and trading the company's shares, from starting a mine and selling the products, or from using the products as feed in some downstream business based on the commodity, and profiting from sales of those new products. Companies of different sizes have different objectives and reasons for exploring, and target different commodities and styles of mineralization.

Prospectors, Small Syndicates, and Small Companies with No Mines

If the small-scale exploration group plan to mine what they discover, commodities such as gold, gemstones, and tin are favoured, because processing of the ore is simple, capital costs are low, and the unit value of the commodity produced is high, so small-scale operations may be viable. Work occurs in prospective areas that are known, typically close to where the group is based. If the small-scale group plans to sell the prospects they discover, whatever the market prefers at the time is favoured, because those targets are readily marketable. Work is undertaken in prospective areas that are familiar to them, or where their financial resources allow operations.

Small Exploration Companies with Mines or Mineral-Based Businesses

If small groups that own mines or businesses plan to use the product in their business, they will explore

for the raw materials they need for their current or planned operations. This will typically take place near current or planned operations to minimize transport costs. If the plan is to produce minerals for sale, small-scale owners will explore for commodities for which they believe they have the technical skills (mining, processing, marketing) to handle, and that can be sold profitably. Typically, work is undertaken in known regions, but may expand to other regions.

Large Mining and Exploration Companies

Some large companies focus on specific commodities or metals for which they perceive they have a technical and operational advantage, or for which wish to appeal to investors wanting commodity focus (e.g., Barrick and Newmont for gold, De Beers for diamonds, Cominco for zinc, Alcoa for aluminium, and INCO for nickel). They explore where they believe they have the greatest chance of success, based on geology and legal, commercial, and security criteria. Other large companies focus on any commodities meeting their commercial criteria: typically these will be commodities with very large global markets, no barriers to entry, low production costs, and major cash flows. These companies explore where they believe they have the greatest chance of economic success, based on geology and legal, commercial, and security criteria.

Mega-Mining Houses

This small group of very large mining houses has emerged only recently as a result of mergers and acquisitions (e.g., Anglo American, BHP Billiton, and Rio Tinto). For these corporations, exploration alone can no longer deliver their growth objectives because very few individual discoveries can meet their commercial criteria, and because explorers cannot reliably predict when they will make their next economically viable discovery. Depending on the level of competition, these companies may reduce their exploration expenditure and aim instead to acquire major discoveries made by successful smaller companies. If they acquire a small company that has grown by exploration success, in many instances they discard the successful exploration group.

Strategies for Exploration

The strategies and tools applied to exploration vary with the size of the group's exploration budget and its technical capabilities. Prospectors generally work as individuals or small syndicates. Corporations employ specialist explorers, mostly geologists, but include in-house specialist geochemists and geophysicists where there is sufficient demand for their services. Large exploration groups may comprise several hundred people, with technical specialists including geologists, geochemists, and geophysicists, and a wide array of technical support staff.

The sequence typically followed by explorers is summarized in Figure 1. The first major strategic issue they face is what to explore for and where to explore for it. There are many possible targets and commodities for which to search, and the choice can be as broad as 'anything we can make money from' or as narrow as one specialist commodity. For companies planning to use minerals in their own operations, the choice of commodities is already decided by what they need. For other explorers, the choice of targets for exploration can be driven by opportunity or by strategy. In either case, the main criteria are the perceived likelihood of success and the reward for success.

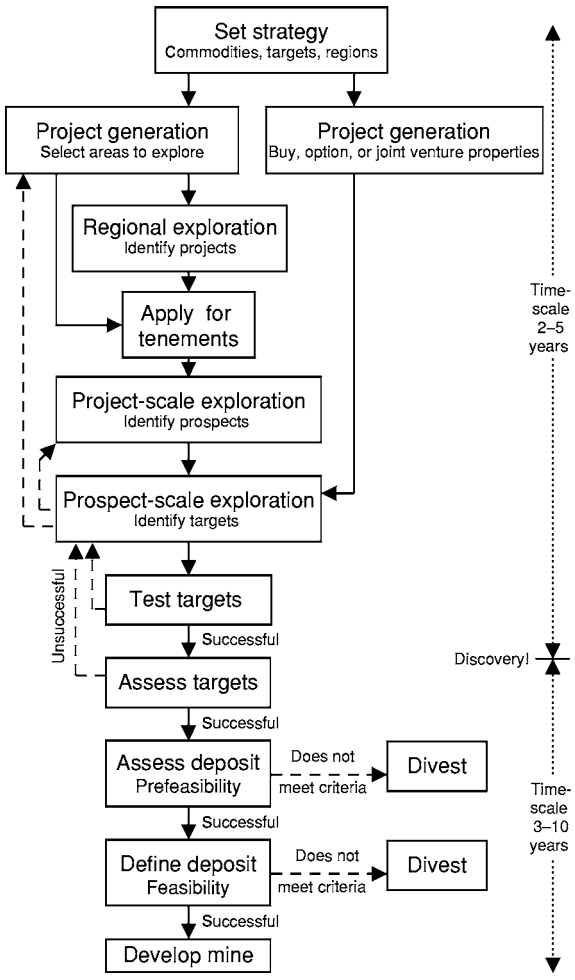


Figure 1 The sequence of stages from mineral exploration to mining.

Selecting commodities and targets is probably most crucial for large companies allocating substantial resources to exploration. They have the opportunity to explore for whatever they choose, and few restrictions on where they explore. However, because they spend a lot of money on exploration, their shareholders expect to see substantial returns for the investment. For a very large company, a small discovery that might seem to be a major success for a small company has no discernible impact on overall profitability or net capital worth. To justify large expenditures, large corporations need to focus on finding deposits that are very large and that can have a significant impact on profitability.

Once the exploration targets have been selected, the next decision is where to explore: locally, regionally, nationally, or globally. For small players, that decision is largely dictated by where they can afford to operate. It is easy to spend a lot of money on travel before any work is done, so small exploration groups typically choose a limited area of operations. To increase their chances of success, every exploration group, regardless of size, must balance between trying to be in all the best places and having as many good prospects as possible against the risk of spreading their effort too thinly and spending too much of their budget on access and support costs. The response of different explorers to these issues varies widely. At the peak of the exploration boom of the late 1990s, many exploration groups spread their activities thinly around the world, while others focused all their resources into a few priority regions. There is no right way, and either strategy can lead to success.

Even when decisions such as what to explore for and where to explore have been decided, there remain many important issues. Prospective regions may be very large, so the exploration group needs to select a region of manageable size on which to focus attention. The most important factor in making a discovery is favourable geology, but because the aim of exploration is economic success, not merely technical success, many other factors come into play. These factors include having secure title over the land to be explored, having access to the land, and having safe operating conditions, favourable legal and taxation laws, certainty of permission to progress from exploration to mining, availability of infrastructure and skilled workers, and many other issues. Previous exploration history must also be considered, because there is no point in exploring an area where previous work has been thorough, and new explorers can only repeat what has been done before. Availability of infrastructure has greatest impact for bulk commodities or industrial minerals, for which transport costs may be a substantial proportion of the sale price obtained, and the least impact for small-bulk, high-value commodities such as gemstones or precious metals, for which the product can easily be flown to market in a small plane.

Once a region has been selected for exploration, there may be opportunities to gain access to advanced projects by entering agreements with explorers already active in the area. Possible arrangements include outright purchase of an established project, option agreements (these involve progressive payments during exploration, with a major payment or share in profits or production if exploration is successful), and joint ventures involving working together to share costs and benefits. If arrangements of this kind are not possible, or the best opportunities are no better than can be obtained alone, then the best option is to explore alone.

This then raises further questions. Should the company focus its attention on known deposits and prospects, or get as close as possible to them (known as brownfields exploration)? Or should it search in other areas with apparently similar geology, but lacking known deposits (known as greenfields exploration)? The relative merits of the two approaches are a contentious issue. In brownfields exploration, it is easier to justify work and expenditures on the basis of being close to a known deposit, but how likely is it that a new deposit will be found close to an old one? With greenfields exploration, the apparent risk is higher because it involves stepping out beyond what is known, in the hope that there are new, unknown deposits to be found. Success is possible using either approach. One thing that is certain, however, is that brownfields opportunities ultimately are limited: new deposits cannot continue to be found by endlessly revisiting the same areas.

A major goal of exploration is to get to the point of testing a potential orebody as quickly and as cheaply as possible. Identifying a volume of rock that potentially hosts an orebody (the target) is an important step along that path. Identification of the target may result from the evaluation of previous work or application of a new exploration technique on a brownfields site, or from a systematic exploration program on a greenfields site. Most exploration starts with identification of a broad target area that has the potential to contain an orebody: the challenge is then to narrow the focus to a specific target that can be tested, hopefully leading to discovery. In systematic exploration, the starting region might be as large as hundreds of thousands of square kilometres, whereas the target usually has an area of much less than $1\,km^2$ (commonly $0.003\,km^2$ or less). Explorers can apply a wide array of tools and exploration methods to narrow down the area of search and define

targets for drilling. The skill in exploration lies in how those tools are used to discover an orebody.

Exploration Tools

The wide variety of exploration techniques available for exploration can be broadly grouped into three types: geological, geochemical, and geophysical. The selection of the technique or combination of techniques used varies with the target sought, the area to be explored, the geological conditions, the stage of exploration, the weathering regime, and factors such as location, topography, vegetation cover, climate, and social and cultural issues. Some techniques can be applied over large regions, whereas others can be applied only over small areas. Some require detailed ground access, whereas others can be flown by aircraft and require minimal ground access. The following discussion focuses mostly on exploration for metallic ore deposits, because for them, discovery of an economic tonnage and grade is typically the crucial step. For many bulk commodities (especially most industrial minerals), discovery is important, but the ultimate value of the find relates less to grade and more to marketability, which involves many different characteristics, such as physical properties, amenability for the ultimate application, and presence or absence of deleterious impurities. Thus, the relative importance of the early search and discovery stages, and the later testing and evaluation stages, changes with the commodity sought.

Geological Techniques

If the geological situation in which mineralization is likely to occur can be recognized by observation, then geological mapping of rock types, stratigraphy, or structure can be used (Table 1; Figure 2). Work on very prospective areas may involve more detailed studies, including petrography, fluid inclusion studies, or alteration mineralogy. Geological techniques are applicable only where geological observations can be made (i.e., where there is sufficient outcrop or where appropriate samples can be collected) or where geological inferences can be made indirectly (from aerial geophysical or spectral surveys, aerial photographs, satellite images, etc.).

Geochemical Techniques

Geochemical techniques involve the chemical analysis of geological materials such as stream sediment, soil,

Table 1 Main geological tools in mineral exploration

Tool	*Basis*	*Coverage*	*Applicability*	*Target*
Geological mapping	Systematic mapping of surface rock types, structures, hydrothermal alteration, etc.	Regional scale available globally; detailed coverage very limited	Requires reasonable distribution of exposure of bedrock, though other tools can be used to fill in missing information	An important tool in exploration for all deposit types; provides the basis for interpretation of all other datasets
Aerial photograph interpretation	A stereoscopic view of Earth's surface is provided by overlapping photographs; lineaments, colour, structures, weathering patterns, and vegetation can be used to generate a geological interpretation	Global coverage available with variable resolution and quality	Applicable if the surface is visible (no cloud cover); best in areas lacking thick vegetation cover	An important aid to geological mapping; can be used to locate orebodies that are exposed at the surface and show distinctive features
Satellite image interpretation	Satellite based scanners generate detailed images of Earth's surface	Global	Available everywhere, but application varies with density of buildings and agriculture, vegetation, rainfall, weathering, etc. Most useful in arid, well exposed regions	An important aid to geological mapping; can seldom be used to locate orebodies directly, but in favourable conditions in remote areas, can highlight alteration zones
Spectral data interpretation	A wide variety of spectral information is now available from satellites, aircraft, and ground surveys	Global for satellite systems, large areas from aircraft, small areas with ground instruments	Applicability varies with density of buildings and agriculture, vegetation, rainfall, weathering, etc. Most useful in arid, well exposed regions	A useful aid to geological mapping; in favourable conditions, can locate hydrothermal alteration related to mineralization, or, rarely, large mineralized areas

Figure 2 Geological mapping in the Farallon Negro district, Argentina. All rocks show intense hydrothermal alteration related to a major porphyry copper molybdenum gold deposit called Agua Rica.

rock, and, in some cases, water, vegetation, or air (Table 2; Figure 3). Analytical results ('anomalies') may suggest prospective areas. Geochemical techniques are applicable only where the rock that hosts the ore, or material derived directly from it, is accessible. Sampling of overlying transported material (e.g., glacial debris) is not effective unless the source of the transported material is known, or unless chemical elements related to mineralization have migrated up through the transported cover. In areas of shallow cover, drilling may be used to collect samples from beneath the cover. Proper orientation surveys are essential for effective geochemical surveys. These involve testing the response of known mineralization in a similar nearby area to verify the effectiveness of the method under the prevailing conditions, before using it routinely to look for unknown mineralization.

Geophysical Techniques

Geophysical techniques rely on variations in the physical properties of the mineralized rocks or the rocks that surround them to indicate the location of mineralization directly or indirectly (Table 3; Figure 4). The physical response may be a direct property of the rock (e.g., its density, natural radioactivity, magnetic properties, or resistivity), or an induced response produced by exposing the rock to a physical stimulus such as an electrical or magnetic field. Geophysical techniques may indicate mineralization directly, or may allow the search area to be reduced to a favourable host rock or a favourable structural environment. With the increasing availability of large-scale surveys, often involving more than one technique, the data from geophysical surveys can be interpreted to produce geological maps and to identify favourable settings for mineralization, exactly as they would be from conventional geological maps. A major advantage of geophysical techniques is that they can detect responses from mineralization buried several hundred metres below the ground surface, at depths too great to be reflected in the surface geology or detected by geochemical surveys.

Recognizing Targets

Sometimes orebodies are exposed and their surface expression can be recognized by explorers, but as exploration continues, new surface targets become fewer and fewer, and increasingly there is a need to look for more subtle targets. That typically involves the use of multiple exploration tools. The data from all the techniques used are compiled onto maps, and the strength and location of anomalies in each technique are compared. Areas that stand out with coincident anomalies from more than one technique are usually priority targets for drill testing.

One additional geological tool universally used by modern explorationists is the ore deposit model. There are models available for a wide range of ore deposit types, and new ones are regularly developed. These models summarize all the important geological characteristics (including spatial relationships) of a particular deposit type. Models include large-scale aspects such as the tectonic and structural settings,

Table 2 Main geochemical tools in mineral exploration

Tool	*Basis*	*Coverage*	*Applicability*	*Target*
Soil geochemistry	Systematic sampling of soil, usually of a selected size fraction	Large to small areas, depending on sampling density	Areas with residual soils; a very commonly used method	Most metallic deposit styles
Rock chip geochemistry	Systematic sampling of exposed bedrock as composites of small chips	Small areas only	Requires extensive exposure of bedrock	Most metallic deposit styles
Bedrock geochemistry	Systematic sampling of the upper parts of bedrock, usually by drilling through shallow cover	Large to small areas, depending on sampling density and depth of cover	Areas of strong near surface leaching, or covered by transported overburden	Most deposit styles
Stream sediment geochemistry	Systematic sampling of stream sediments, usually of a selected size fraction	Mostly large areas; application to small areas may be limited by the stream pattern density	Widely applicable in regional exploration; requires bedrock outcrop or shallow subcrop	Most metallic deposit styles
Water geochemistry	Sampling of water in streams or wells and bore holes	May be applied over large areas; sample density typically low	Depends on access to water samples in streams or bores	Limited application for base metal and uranium deposits
Soil gas geochemistry	Systematic sampling and analysis of gases trapped in soil	Mostly small areas; sampling density is high	Suitable for areas with deep soil and transported cover	Locates structures and may help find some sulphide rich deposit styles
Plant geochemistry	Sampling of one plant species or part of one plant species, followed by chemical analyses (mostly after ashing)	Mostly over restricted areas	Best in areas with high plant density but low plant diversity (usually temperate to cold areas)	Most base metal deposit styles

Figure 3 Geochemical sampling in eastern Taiwan. Sediments are dug from the creek bed, sieved to remove the largest pieces, and the finer fraction is bagged and sent to a laboratory for chemical analysis.

Table 3 Main geophysical tools in mineral exploration

Tool	*Basis*	*Coverage*	*Applicability*	*Target*
Magnetics	Measures spatial variations in the intensity of Earth's natural magnetic field	Very large areas covered by aerial surveys; smaller areas by ground surveys	Widely used as an aid to geological mapping; also a direct detection method for some deposits: applicable wherever rock magnetic properties vary	All deposits with a well defined geological setting in magnetically variable rocks; direct detection of deposits containing magnetic minerals (e.g., iron ore, chromite, diamonds, some volcanic massive sulphide deposits, some porphyry copper deposits)
Radiometrics	Measures spatial variations in the intensity of natural radiation from potassium, thorium, and uranium	Very large areas covered by aerial surveys; smaller areas by ground surveys	Widely used as an aid to geological mapping; also aids direct detection of some deposits: applicable wherever there are variations in abundance of natural radioactive elements	All deposits with a well defined geological setting; direct detection of deposits containing uranium or thorium minerals, or deposits with strong potassium enrichment in alteration zones (e.g., some volcanic massive sulphide deposits, some porphyry copper deposits, some epithermal gold deposits)
Gravity	Measures spatial variations in the intensity of Earth's natural gravitational attraction	Very large areas covered by aerial gravity gradiometer surveys; smaller areas by ground surveys	Huge areas covered by very low density ground surveys; limited coverage with high resolution aerial surveys; detailed ground surveys only in areas of low topographic relief: applicable wherever rock densities vary	Direct detection of deposits with strong density contrast to surrounding rocks (e.g., iron ore, volcanic massive sulphide deposits, kimberlites)
Electromagnetics	Measures the induced response in the Earth from an applied electromagnetic field	Very large areas covered by aerial electromagnetic surveys; smaller areas by ground surveys	Can be used as an aid to geological mapping; also aids direct detection of some deposits: applicable wherever there are variations in natural rock conductivities	Direct detection of deposits with high concentrations of conductive minerals (e.g., volcanic massive sulphide deposits, manganese deposits) or where there are associated anomalous conductivity zones (e.g., some porphyry copper deposits, kimberlites)
Induced polarization	Measures the induced electrical polarization response in minerals from an applied electrical field	Ground technique mostly applied over small areas, but large surveys also possible	Detects the presence of polarizable minerals (mainly pyrite, pyrrhotite, galena, chalcopyrite, etc.)	Direct detection of deposits with high concentrations of polarizable minerals (e.g., volcanic massive sulphide deposits, porphyry copper deposits)
Resistivity	Measures spatial variations in the resistivity of the rock mass being surveyed	Mostly applied over small areas	Detects variations in resistivity	Sulphide rich mineralization has low resistivity compared to the surrounding rock; applied to direct detection of sulphide rich masses (e.g., massive sulphide deposits)
Seismic	Records underground surface reflections produced by explosions or mechanical thumpers at the surface	Applied over huge areas for petroleum exploration; seldom used in mineral exploration	Indicates surfaces in the bedrock (e.g., faults, beds, igneous layers)	In mineral exploration, mainly used for layered complexes (e.g., Bushveld) or for engineering purposes

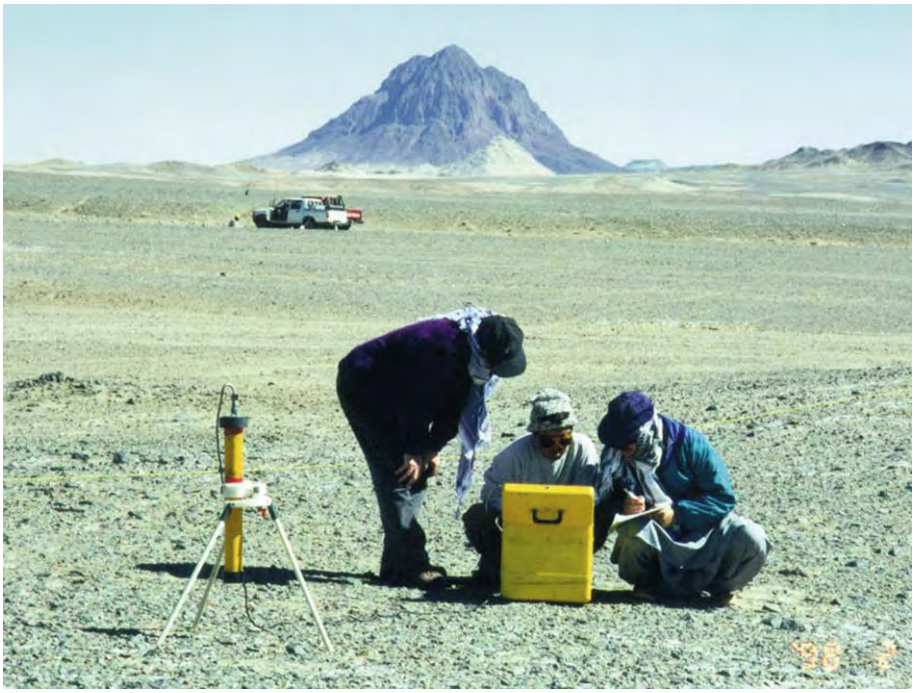


Figure 4 A geophysical survey being conducted in western Pakistan. This is an electromagnetic survey, in which electrical pulses are used to induce a response from conductive rocks in the ground. Photograph by R Irvine, reproduced with permission.

down to small-scale features such as the mineralogy of the ore, and details such as the chemistry of fluid inclusions. In most but not all cases they include genetic models summarizing how geologists believe the deposits form. They also reflect the kind of geological, geochemical, and geophysical expressions an orebody of this type should have, and so they help in the interpretation of exploration data.

Testing Targets

Once possible mineralization has been located or a target identified, it is then necessary to test it to confirm that it is what is sought, and how commercially significant it is (Figure 5). The cheapest approach in near-surface situations is to dig trenches across the mineralization to create complete exposure and to allow detailed sampling. If the target is more than a few metres below the surface, drilling is needed (Figure 6). Drilling is the most reliable means to investigate a target and is used to define the grade and extent of the mineralization. If testing confirms that potentially economic mineralization has been found, then a discovery has been made (Figure 7). Further work is then required to quantify the full significance of the discovery.

The assessment stage that follows a discovery involves gathering huge the amounts of data needed to assess whether the mineralization can be mined economically. Typically, widely spaced drilling is used first to assess the overall size of the mineralized body, and then more detailed infill is used to determine the shape, distribution, tonnage, grade, and metallurgical properties. If it is found that the indicated tonnage and grade could potentially justify starting a mine, a new stage of data-gathering follows; typically projects pass through several stages, at each of which it is reassessed to see if another more detailed stage of expensive assessment work is warranted. These stages typically include prefeasibility, full feasibility, and, if the outcomes are positive, finally a decision to go ahead and develop a mine. Depending on the commodity, the planned scale of operation, and the location of the deposit, developing a mine might cost anything from a few million dollars up to several billions of dollars. (*see* **Mining Geology: Exploration Boreholes**).

Success and Failure in Exploration

Exploration is a technical activity with a commercial objective. It is common to see these two aspects, the technical and commercial, confused in claims of success. In the case of junior companies, it is important that any success (not necessarily a commercial success) is publicly highlighted, because it is important for these groups to be seen as successful, to increase their ability to raise further funds for ongoing

Figure 5 Exploration success! The geologists are standing around the drill hole that discovered the high grade Cannington silver lead zinc orebody in western Queensland, Australia. A drill rig can be seen on the skyline, where drilling is being done further along the same orebody. The deposit was brought into production and quickly became the world's largest producer of silver.

Figure 6 A helicopter being used to move a diamond drilling rig in the Kimberley region of Western Australia. The helicopter allows the drill to be positioned with minimal disturbance and environmental damage.

exploration. By contrast, it is sometimes in the interests of major mining companies to play down expectations from their exploration, and their successes often go unmentioned, or are announced only in a low-key way. Too much publicity can increase the level of competition for ground near the discovery, and overly enthusiastic expectations may mean shareholders are ultimately disappointed if a technical discovery does not turn into a commercial success.

The most important justification for exploration is that it supplies society with the mineral products it needs. This benefit, however, is realized only if the value of the products is greater than the cost of discovery and production. Real success in exploration and mining must therefore be judged in commercial, not technical, terms. In this respect, the role of the large exploration and mining companies has particular importance. They have financial and technical

Figure 7 Examining trays of drill core from the Eloise copper gold deposit in western Queensland, Australia. Diamond drilling techniques are used, which allows deep holes to be drilled, and the core recovered allows detailed geological examination.

resources that small companies lack, and they are not so dependent on the day-to-day demands of the share market. This means they can pursue longer term goals and are more likely to discover and develop long-life mines that contribute stability to the long-term supply of major mineral commodities. Some junior companies are fast and economically efficient explorers and play an important role by supplying new exploration opportunities to larger companies and by bringing into production deposits that would be too small for major companies.

The Role of Governments in Exploration

In most parts of the world, governments regulate exploration and mining. Governments control the granting of titles, the conditions that are applied, and regulations governing safety and environmental protection. Governments also determine the availability of regional data and the cost of accessing it, and they set the levels of taxation and other imposts and concessions that are applied to the industry. Direct government involvement in exploration and mining is not regarded favourably these days, because having a government acting as both regulator and competitor involves them in a conflict of interest that may undermine trust in the regulatory process.

Although favourable geology is by far the most important determinant of how prospective a region is for exploration in a purely technical sense, the potential for commercial success from exploration is strongly influenced by the regulatory and financial environment in which it operates, as laid down by governments. Countries with relatively poor geology may still be attractive for exploration if the legal, economic, and regulatory environment is attractive; conversely, even the best geology does not compensate for a bad environment, because confidence in the stability of an operating environment is very important for explorers and miners.

The Future of Exploration

Mineral exploration is an essential part of the process of supplying the resource needs of societies: while that demand continues, so will mining and the exploration that is needed to sustain it. But exploration faces many challenges, and is currently undergoing major changes.

With the rise in environmental awareness in the second half of the twentieth century, mining was widely criticized for its environmental impact. Much of that criticism was justified, and there have been major changes, such that the modern mining industry can now operate with minimal environmental impact in the areas surrounding mines. However, it is inescapable that any process that takes material from one place to another must have an environmental impact. Mineral exploration has also come under attack by environmentalists, but most parts of the exploration process have very little to no environmental impact. Most opposition to exploration is either misinformed about what is actually done, or more commonly is based on fear of what might happen in the unlikely event that a discovery is

made, leading to mining. In fact, most exploration efforts leave no long-term trace and make no discoveries. Even when targets are identified, on further testing less than 1% prove to be of economic interest.

The mining industry as a whole, including exploration, has been criticized for not returning a satisfactory profit on money invested, but such analyses can only be sensibly done on a case-by-case basis. Over the past decade, major changes have occurred in financial markets and in how businesses operate, with emphasis on short-term financial performance. Because exploration cannot guarantee very quick returns, it lost favour with investors, with the result that in the late 1990s, exploration expenditure dropped drastically, and investment in mining developments were also sharply reduced in response to low commodity prices. If this situation were to continue, there would eventually be a shortage of some commodities, which should cause prices to rise, making exploration and mining of those commodities attractive again.

The last few years of the twentieth century represented a period of dramatic change for the mining and exploration industry. These changes are still being played out, but a new industry structure and way of operating have to be found. Throughout the world, all people seek a better life, and to fulfil that desire requires more and better houses, vehicles, roads, and everyday things. All those things have to be made from some material commodity, and most commodities can be obtained only by mining. Mining will ultimately die without exploration, so, as long as populations continue to increase and people strive for a better standard of living, requiring the use of mineral commodities, the future of mineral exploration is assured.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Economic Geology**. **Geochemical Exploration**. **Geological Maps and Their Interpretation**. **Minerals:** Definition and Classification. **Mining Geology:** Exploration Boreholes; Hydrothermal Ores; Magmatic Ores. **Quarrying**.

Further Reading

Boyle RW (ed.) (1971) *Geochemical Exploration, CIMM Special Volume 11*. Quebec: Canadian Institute of Mining, Metallurgy, and Petroleum.

Dobrin MB and Savit CH (1988) *Introduction to Geophysical Prospecting*, 4th edn. New York: McGraw Hill.

Evans AM (1995) *Introduction to Mineral Exploration*. London: Blackwell Science Ltd.

Harben PW and Kuzvart M (1997) *Industrial Minerals Global Geology*. Surrey, UK: Industrial Minerals Information Ltd.

Levinson AA (1980) *Introduction to Exploration Geochemistry*, 2nd edn. Wilmette, Illinois: Applied Publishing.

Peters WC (1987) *Exploration and Mining Geology*, 2nd edn. New York: John Wiley & Sons.

Rose AE, Hawkes HE, and Webb JS (1979) *Geochemistry in Mineral Exploration*, 2nd edn. New York: Harper and Row.

Sillitoe RH (1995) *Exploration and Discovery of Base and Precious Metal Deposits in the Circum Pacific Region during the Last 25 Years, Resource Geology Special Issue 19*. Tokyo: The Society of Resource Geology.

Sillitoe RH (2000) *Exploration and Discovery of Base and Precious Metal Deposits in the Circum Pacific Region A Late 1990s Update, Resource Geology Special Issue 21*. Tokyo: The Society of Resource Geology.

Telford EM, Geldart LP, Sheriff RE, and Keys DD (1976) *Applied Geophysics*. Cambridge, UK: Cambridge University Press.

Van Blaricom R (1992) *Practical Geophysics II for the Exploration Geologist*. Spokane: Northwest Mining Association.

Vanecek M (ed.) (1994) *Mineral Deposits of the World; Ores, Industrial Minerals, and Rocks, Developments in Economic Geology*. vol. 28. Amsterdam: Elsevier.

Mineral Reserves

M Vaněček, Charles University, Prague, Czech Republic

Introduction

Mineral reserves and mineral resources are terms that cannot be separated from each other. Their definitions are derived from the grade of geological certainty of the basic information about the mineral accumulation and from expected profitability of extracting and marketing the mineral commodity. In all main classifications (i.e., the United Nations, the USA, and the Russian classifications) the definitions of mineral resources and mineral reserves are slightly different, but in all cases the reserves form only a better known part of the resources. Procedures for mineral reserves calculation can be divided into the scalar-geometrical methods and into geostatistical methods. For the most part, in mineral statistics,

world reserves and reserve base of different minerals are introduced.

The terms mineral reserves and mineral resources are inseparably bound to each other. One of the possible definitions for resource is: "something in reserve or ready if needed". This can be made precise for mineral resources as follows: "valuable mineral deposits of an area that are presently recoverable and may be so in the future, including known ore bodies and potential ore". Mineral reserves usually represent a part of mineral resources that could be defined with respect to the quantity and quality of mineral bodies.

The evaluation of reserves/resources of a mineral deposit is based on reserve calculation procedures that would include description and analysis of main geological characteristics of the deposit, such as tonnage, grade, shape and thickness of ore bodies, and depth of mineralization etc., and on a feasibility study that analyzes expected profitability of extracting and marketing the mineral commodity in a given economy and at a given time.

Methods of reserves calculation and of economic analyzes of mining activities have developed over a long period of time. The various connoisseurs in prospecting, mining and metallurgy were the very first to be able to evaluate the worth of a mineral accumulation. In 1910, during the XI, International Geological Congress in Stockholm, the iron ore resources of the world were classified into different groups of exactitude and the symbols A, B, and C were used for the first time. At present time a special branch of economic geology, called mineral economics, deals with the reserves/resources evaluation and classification.

Due to the needs of industrialization, different systems of reserves classification and calculation have arisen. So far as the classification, i.e., group segregation, of mineral reserve tonnage is concerned, we find two schools of thought in existence. Governmental agencies, for example different geological surveys, are mostly concerned with the determination of the future potential of mineral resources. On the other hand, a reserve analysis made for or by a private enterprise usually is designed to show the various mineral tracts classified on the basis of their currently mineable nature. The classification terminology used by the governmental agencies is as follows: measured, indicated, and inferred resources. On the other hand, in mining geology there are used more readily the terms developed ore, proved ore, probable ore, and possible and/or extension ore. Single members of both classifications are not direct equivalents. For example, the developed ore may be only a part of measured or proved reserves that are prepared for immediate withdrawal by mining. Some probable ore may be the essential counterpart of some measured reserve or/and of some indicated resource. Possible ore, which is sometimes called future ore, and extension ore, are to all intents and purposes the equivalent of inferred reserve.

At present, three different reserves/resources classifications are mainly used: the United Nations International Framework Classification, the Classification System of the U.S. Bureau of Mines and U.S. Geological Survey, and the Classification of the State Commission for Reserves of the Russian Federation.

The UN Framework Classification provides information about the stage of Geological Assessment, the stage of Feasibility Assessment, and the degree of Economic Viability. Geological Study is subdivided into four consecutive stages of geological assessment, which are: Reconnaissance, Prospecting, General Exploration, and Detailed Exploration. Feasibility Assessment is subdivided into three consecutive stages which are: Geological Study, Prefeasibility Study, and Feasibility Study/Mining Report. There are two categories of Economic Viability: economic, and potentially economic, which are only quoted in the stages of Mining Report/Feasibility Study and Prefeasibility Study (**Table 1**).

With respect to a variety of meanings in the use of the terms 'reserve' and 'resource' they are redefined as follows: the Total Mineral Resource is defined as the naturally occurring concentration of mineral raw material of economic interest and with specified geological certainty. A Mineral Reserve is the economically mineable part of the Total Mineral Resource as demonstrated by Feasibility Assessment. The Remaining Mineral Resource is the balance of the Total Mineral Resource that has not been identified as a Mineral Reserve. Only the groups 'proved' and 'probable' are reserves, whereas the group 'inferred' belongs to resources. The UN Framework Classification is mostly applied in Western Europe and in some European countries with 'economies in transition' (**Figure 1**).

The U.S. Resource/Reserve Classification for Minerals defines Resource as a concentration of naturally occurring solid, liquid, or gaseous material in or on the Earth's crust in such form and amount that economic extraction of a commodity from the concentration is currently or potentially feasible. Identified Resources are resources whose location, grade, quality, and quantity are known or estimated from specific geological evidence. Identified resources include economic, marginally economic, and subeconomic components. To reflect varying degrees of geologic certainty, these economic divisions can be subdivided into 'measured', 'indicated', and 'inferred'. For the sum of measured plus indicated, the term 'demonstrated' is used. The part of an identified resource

Table 1 United Nations International Framework Classification for Reserves/Resources. Solid fuels and mineral commodities

UN International Frame work				
	National system			
	Detailed exploration	*General exploration*	*Prospecting*	*Reconnaissance*
Feasibility study and/or mining report	1 Proved Mineral Reserve (111)	Usually not relevant	Usually not relevant	Usually not relevant
	2 Feasibility Mineral Resource (211)	Usually not relevant	Usually not relevant	Usually not relevant
Prefeasibility study	1 Probable Mineral Reserve (121) + (122)	1 Probable Mineral Reserve (121) + (122)	Usually not relevant	Usually not relevant
	Prefeasibility Mineral Resource (221) + (222)	Prefeasibility Mineral Resource (221) + (222)	Usually not relevant	Usually not relevant
Geological study	1 2 Measured Mineral Resource (331)	1 2 Indicated Mineral Resource (332)	1 2 Inferred Mineral Resource (333)	? Reconnaissance Mineral Resource (334)

Economic viability categories: 1 economic; 2 potentially economic; 1 2 economic to potentially economic (intrinsically economic); ? undetermined.

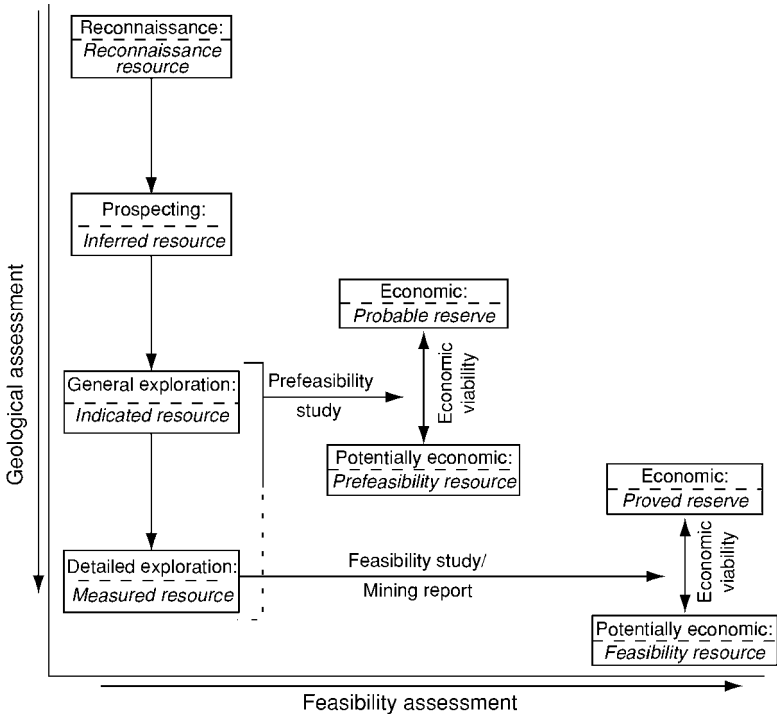


Figure 1 Schematic illustration of the UN Framework Classification.

that meets specified minimum physical and chemical criteria related to current mining and production practices, including those of grade, quality, thickness, and depth, is called Reserve Base. Reserves are that part of the reserve base which could be economically extracted or produced at the time of determination. The whole U.S. system of resource/reserve classification is illustrated by Tables 2 and 3. The U.S. Resource/Reserve Classification for Minerals seems to be the most widespread, not only in the USA, but in most of world.

In the classification of the State Commission for Reserves (GKZ) of the Russian Federation reserves and resources are distinguished. According to the grade of geological certainty the reserves are divided into the explored ones (categories A, B, C 1), and the estimated ones (category C 2). From the point of view of the economic importance, the reserves are

Table 2 Major elements of mineral resource classification, excluding *reserve base* and *inferred reserve base*

	Identified resources			*Undiscovered resources*	
	Demonstrated			*Probability range*	
Cumulative production	*Measured*	*Indicated*	*Inferred*	*Hypothetical*	*Speculative*
Economic	Reserves	Reserves	Inferred reserves		
Marginally economic	Marginal Reserves	Marginal Reserves	Inferred marginal reserves		
Subeconomic	Demonstrated subeconomic resources	Demonstrated subeconomic resources	Inferred subeconomic resources		
Other occurrences	Includes non conventional and low grade materials	Includes non conventional and low grade materials	Includes non conventional and low grade materials	Includes non conventional and low grade materials	

Table 3 Reserve base and inferred reserve base classification categories

	Identified resources			*Undiscovered resources*	
	Demonstrated			*Probability range*	
Cumulative production	*Measured*	*Indicated*	*Inferred*	*Hypothetical*	*Speculative*
Economic	Reserve base	Reserve base	Inferred reserve base		
Marginally economic	Reserve base	Reserve base	Inferred reserve base		
Subeconomic	Reserve base	Reserve base	Inferred reserve base		
Other occurrences	Includes non conventional and low grade materials	Includes non conventional and low grade materials	Includes non conventional and low grade materials	Includes non conventional and low grade materials	

Table 4 Russian classification of reserves

	Grade of geological certainty		*Group of economic importance*	
Reserves	Explored	A	Economic	Subeconomic
		B	Economic	Subeconomic
		C_1	Economic	Subeconomic
	Estimated	C_2	Economic	Subeconomic
Resources				
	Prognostic	P_1		
		P_2		
		P_3		

subdivided into economic and sub economic. In the Russian mining industry a classification of the explored reserves according to their mineable nature is used as follows: developed, prepared for extraction, and ready to extraction. Resources are divided into three groups of prognostic resources (P 1, P 2, P 3) depending on the grade of geological certainty. Main patterns of the Russian reserves/resources classification are shown on Table 4. The Russian reserves classification is used mostly in the states of the former USSR and also in some Eastern European countries.

The general procedure of a mineral reserve calculation is the same regardless of the concept that is used in making the ultimate classification, since it follows that no classification of any mineral reserve can be visualized completely until all necessary working data have been evaluated. Two main groups of traditional resource estimations can be distinguished:

1. The scalar-geometric methods or methods of geometrization
2. The spatial ('geostatistical') methods.

In the first case a complex natural shape of mineral bodies are geometrized into defined geometric bodies, the volume of which can be calculated. The following methods are distinguished:

1. Methods of blocks
2. Methods of geological sections
3. Methods of isolines.

Methods of blocks can be divided depending on a geometrical arrangement of the exploration points

(i.e., boreholes, shafts, etc.) into triangular, regular squares, rectangular, polygonal, and irregular or geological blocks. These methods differ slightly in ways of determination of volume and grade of mineral bodies. The average thickness is calculated as an arithmetical mean in all cases with an exception of the polygonal method where thickness and grade are taken from the borehole lying at the geometrical centre of each polygon. In all other kinds of blocks an average grade of the mineral body is calculated as an arithmetical mean weighted by thicknesses. Polygonal blocks and geological blocks are applied if the arrangement of the exploration points is irregular. Contours of geological blocks are derived from the geological situation (tectonic elements, facial changes, changes of thickness, etc.).

Methods of sections can be divided as follows:

1. Longitudinal sections (along the axis of the body)
2. Cross-sections.

Sections may be vertical, horizontal, and inclined, each other being parallel and non-parallel. Owing to the application of geological sections the natural shape of the body is nearly preserved. Volumes of blocks can be calculated either between two sections or using area of only one section. The first way is applied in the case of similar in shape and near in dimension geological sections, the second one is used if the shape of the mineral body is very irregular and/or if due to geological reasons a connection between two sections is not assumed. The average grade of blocks is calculated as an arithmetical mean weighted by the areas of geological sections of the body. Areas of the sections are measured by means of instruments (planimeter, scanning, etc.) or average thickness may be used.

For calculation by means of isolines there are used isopachs (i.e., isolines of thicknesses), isohypses (isolines of altitude), grade, density and their combination can be used. Volume of mineral bodies is given by contours of isolines and can be calculated using the Simpson formula. If a combination of isopachs, of isolines of grades, and of isolines of density is used the computed volume is equal to the reserve.

Geostatistics, in their most generally accepted use, are concerned with the study of the distribution in space of useful values for mining engineers and geologists, such as grade, thickness of accumulation, including a most important practical application to the problems arising in mineral deposit evaluation. Spatial or geostatistical methods are working with statistical processing of primary data and with description of a statistical character of the distribution of main properties of the deposit in space by means of variograms. The discovered statistical characteristics that allow the prediction of the distribution of studied properties in the mineral body are used for the computing of reserves. One of the most known and used geostatistical methods is so called 'kriging', a method developed for gold reserves computing in South Africa.

In the international statistical overviews of the world reserves and of the regional resources of a mineral commodity, the terms 'resources' and 'reserves' base are usually given. For the metallic ores the reserve are given mostly in terms of metal content. Reserves of naturally occurring non-metals and fossil fuels are expressed in the tonnage of raw mineral material. Quantities, qualities, and grades may be expressed in different terms and units to suit different purposes, but usage must be clearly stated and defined.

Mineral reserves can be defined also as: "a quantity of mineral raw materials in the Earth's crust which is estimated as a result of geological exploration works or during their extraction". They could be calculated for a single deposit, for an ore field or district, for a basin, for a region, for states, for continents, and for the whole world.

World mineral reserves of each raw mineral commodity are immense and their quantity is subject of changes due to the development of geological knowledge, in conjunction with considerable mining activity. For example, world resources of iron ores are estimated to exceed 800 billion metric tonnes of crude ore containing more than 230 billion tonnes of iron. World reserves in iron content are estimated to be 71 000 million metric tonnes, the reserve base being 160 000 million metric tonnes. Annual world output of iron ore mines varies by about 1 billion metric tonnes per year. World reserves of phosphate rocks are recorded as high as 12 billion metric tonnes, reserve base being 37 billion metric tonnes. World resources of coal are estimated to be about 15×10^{12} metric tonnes, of which about 1067 billion metric tonnes are recoverable reserves. In 2001 the annual world mining output of coal was 3630 metric tonnes. These few examples may give an idea as to the order of huge quantities of world mineral resources and reserves.

See Also

Military Geology. **Mining Geology:** Exploration. **Petroleum Geology:** Reserves.

Further Reading

Agterberg F P (1974) Geomathematics: mathematical back ground and geoscience applications. In: *Developments in Geomathematics,* vol.1, p. 596. Amsterdam, London, New York: Elsevier.

Bárdossy G, Szabó I R, and Varga G (2003) *A New Method of Resource Estimation for Bauxite and Other Solid Mineral Deposits,* pp. 57 64. New York: Springer Wien.

Carr D D and Herz N (eds.) (1989) *Concise Encyclopedia of Mineral Resources,* p. 426. Oxford New York: Pergamon Press.

Clark I (2000) *Practical Geostatistics,* pp. 68 85. Columbus, Ohio: Greyden Press.

Goovaerts P (1997) *Geostatistics for Natural Resources Estimation,* p. 483. New York: Oxford University Press.

Jewbali A and Mousset Jones P (2002) A survey of sampling and resource/reserve estimation practices in the surface gold mining industry. *Mining Resources of England* 11(2): 183 195.

Mineral Commodity Summaries 2003 (2003) *U.S. Geological Survey,* p. 199. Washington: U.S. Department of the Interior.

Rudawski O (1986) *Mineral Economics,* p. 192. Amsterdam: Elsevier.

U.N. Economic and Social Council (1999) *United Nations International Framework Classification for Reserves/Resources.* New York: U.N. Economic and Social Council.

Wellmer F W (1983) *Klassifikation von Lagerstättenvorräten mit Hilfe der Geostatistik,* pp. 9 43. Schriftenreihe der GDBM, H.39, Weinheim: GDBM (GDBM = Society of German Miners and Metallurgists).

Working Group of IMM in Conjunction with EFG and IGI (2000) *European Code for Reporting of Mineral Exploration Results, Mineral Resources and Mineral Reserves.* (The European Code). Consultation draft.

Hydrothermal Ores

M A McKibben, University of California, CA, USA

Introduction

Hydrothermal ores are natural, economically valuable concentrations of metallic minerals whose origins can be directly tied to the physical and chemical actions of hot aqueous fluids within the Earth's crust. Hydrothermal ores are ideally distinguished from those that form by other geological processes, including magmatic ores (formed by segregation from a melt), sedimentary ores (formed by sedimentation or precipitation from ambient surface waters), and residual ores (formed by surficial weathering). Such strict genetic distinctions are not always possible, however; for example, magmatic-hydrothermal ores (such as skarns and pegmatites) may arise from a combined process whereby an aqueous phase exsolves from a crystallizing magma. The origins of some diagenetic ores likewise may fall within a realm between purely sedimentary and hydrothermal processes. So while the defining criteria for classifying minerals and ores as hydrothermal is evidence of transport and precipitation from a largely aqueous fluid phase at elevated temperatures, there is always the caveat that transitional processes can take place.

For millennia, humans have observed connections between areas of active subaerial volcanism and the mineralizing actions of hydrothermal fluids, manifested in surficial features such as fumaroles, hot springs, geysers, mud pots, and sinter and travertine terraces. More recently, marine scientists have discovered similar active hydrothermal features on the ocean floors near mid-ocean ridge volcanic centres, particularly 'black smokers' and their associated mounds of accumulated hydrothermal minerals. In all of these instances, it is obvious that minerals containing elements of economic significance are periodically concentrated by such volcanic-related hydrothermal phenomena.

Chemical and physical clues gathered from older ore deposits, as well as deep commercial and scientific drilling into active hydrothermal systems over the last few decades, have provided additional details of hydrothermal processes that are only hinted at in surficial manifestations. These field-based data on modern and ancient systems, coupled with essential theoretical and experimental laboratory studies of hydrothermal processes, have added to our knowledge of the full spectrum of hydrothermal processes that have taken place within the Earth's crust through geological time. Thus, although hydrothermal ores are often associated with volcanism, geologists now recognize that some of these ores can be generated from processes that are peripheral to or entirely unrelated to magmatic processes.

Hydrothermal Fluids

Because of their high temperatures and pressures, hydrothermal fluids in the Earth's crust can contain relatively elevated concentrations of dissolved salts (such as NaCl) and gases (such as H_2S) (Table 1). Dissolved salts and gases can arise from a variety of sources, including evaporative concentration of

Table 1 Representative chemical compositions (mg kg ^{1}) of geothermal fluids and brines in selected geothermal fields of the Salton Trough

Field					
Well	*Salton Sea*	*Imperial*	*Cerro Prieto*	*East Mesa*	*Heber*
Temperature (°C)	*52 14*	*L2 28*	*M 5*	*6 1P*	*5*
Depth (m)	*330*	*275*	*300*	*190*	*195*
Constituent[a]	*2500 3220*	*3290 4270*	*1200*	*2164*	*1800*
Na	54 800	50 466	5004	6362	4019
Ca	28 500	18 140	284	759	750
K	17 700	9555	1203	1124	333
Fe	1710	3219	<1	NA	NA
Mn	1500	985	1	NA	NA
SiO_2	>588	465	569	257	237
Zn	507	1155	NA	NA	NA
Sr	421	1500	NA	NA	41
B	271	217	11	NA	4
Ba	210	2031	NA	NA	4
Li	209	252	13	NA	7
Mg	49	299	<1	9	2
Pb	102	>262	NA	NA	NA
Cu	7	>1	NA	NA	NA
Cd	2	4	NA	NA	NA
NH_4	330	NA	NA	NA	6
Cl	157 500	131 000	9370	11 668	7758
Br	111	NA	31	NA	NA
CO_2	1580	30 000	2400	NA	186
HCO_3	NA	NA	NA	221	NA
H_2S	10	>47	180	NA	1
SO_4	53	NA	4	51	66
TDS	26.5%	25.0%	1.6%	2.2%	1.3%

[a]The compositions given have been corrected for loss of gases and steam during wellbore phase separation and therefore represent the *in situ* reservoir compositions.

NA, not analysed.

TDS, Total Dissolved Solids in weight percent.

Modified from McKibben MA and Hardie LA (1997) Ore forming brines in active continental rifts. In: Barnes HL (ed.) *Geochemistry of Hydrothermal Ore Deposits*, 3rd edn., pp. 875 933. New York: Wiley Interscience.

surface waters, dissolution of evaporites, water–rock interaction, and exsolution from magmas.

The solvent (dissolving) properties of hydrothermal fluids impart their ability to create ore deposits. The high dielectric constant of water and the ability of metal cations to bond with aqueous anions (such as chloride and sulphide) give natural aqueous fluids in the Earth's crust the potential to dissolve, transport and re-deposit the elements that make up hydrothermal ore minerals.

Water vapour is far less effective than liquid water in transporting most valuable elements. The temperature and pressure conditions under which aqueous liquids can effectively transport metals depend not only on the physical properties of water but also on the amounts of dissolved minerals and gases in the water. Salts such as NaCl dramatically increase the critical temperature and pressure of hydrothermal fluids, allowing brines to exist up to magmatic conditions as very effective liquid transporters of metals.

Metals are often solubilized and carried in solution as metal–anion complexes, such as $ZnCl^{+}$ and $Au(HS)_2$. Precipitation of these metals as ore minerals occurs when physico-chemical conditions destabilize the complexes in favour of the minerals.

Cooling, boiling, depressurization, oxidation, fluid mixing and fluid–rock reaction are examples of common processes which can trigger ore mineral precipitation from fluids. For example, native gold can be deposited during shallow effervescence and oxidation of hot spring waters:

$$2Au(HS)_2 + 7.5O_{2(g)} + 2HCO_3 = 2Au_{(s)} + CO_{2(g)} + 4HSO_4$$

Likewise, Zn sulphide can precipitate when chloride brines react with carbonate wall rocks such as limestone:

$$ZnCl^{+} + H_2S_{(aq)} + CaCO_{3(s)} = ZnS_{(s)} + Ca^{2+} + H_2CO_{3(aq)} + Cl$$

In addition to reacting chemically, hydrothermal fluids will flow in response to thermal gradients, density gradients, and gradients in hydrological head or pressure. Flow rates will be highest in zones of maximum permeability, such as fracture and fault zones in brittle rocks, as well as in more porous lithologies such as sandstones and carbonates. Concentrations of ore minerals are thus formed in loci where a number of chemical and physical environmental factors have acted in concert to promote both sustained fluid flow rates and effective precipitation rates of ore minerals.

Hydrothermal Minerals

Economic geologists divide hydrothermal minerals into ore minerals (economically valuable) and gangue minerals (economically worthless). While some hydrothermal ore minerals occur as native metals or alloys, most consist of a metal combined with another non-metallic element or anion group – examples are shown in Table 2. Gangue minerals usually consist of common silicate, oxide, carbonate and sulphate minerals that do not contain sufficient concentrations of valuable metals. Nonetheless, gangue minerals can be just as important as ore minerals in providing scientific clues to the origins of hydrothermal ore deposits, particularly in regard to their age and the origins of the water in the ore-forming fluid.

Table 2 Examples of common hydrothermal ore minerals

Elements	Copper	Cu
	Gold	Au
	Mercury	Hg
	Silver	Ag
Alloys	Electrum	(Au,Ag)
Oxides	Cassiterite	SnO_2
	Cuprite	CuO_2
	Magnetite	Fe_3O_4
	Pyrolusite	MnO_2
	Uraninite	UO_2
Sulphides	Argentite	Ag_2S
	Arsenopyrite	FeAsS
	Bornite	Cu_5FeS_4
	Chalcocite	Cu_2S
	Chalcopyrite	$CuFeS_2$
	Cinnabar	HgS
	Cobaltite	(Co,Fe)AsS
	Covellite	CuS
	Galena	PbS
	Molybdenite	MoS_2
	Realgar	AsS
	Sphalerite	ZnS
	Stannite	Cu_2FeSnS_4
	Stibnite	Sb_2S_3
Sulphosalts	Enargite	Cu_3AsS_4
	Pyrargyrite	Ag_3SbS_3
	Tennantite	$(Cu,Zn)_{12}(As,Sb)_4S_{13}$
Tellurides	Calaverite	$AuTe_2$
	Sylvanite	$(Au,Ag)Te_2$
Tungstates	Scheelite	$CaWO_4$
	Wolframite	$(Fe,Mn)WO_4$
Carbonates	Cerussite	$PbCO_3$
	Smithsonite	$ZnCO_3$
Sulphates	Anglesite	$PbSO_4$

The crystal habits and textures of hydrothermal ore and gangue minerals are controlled by many factors, including fluid saturation states, nucleation/precipitation kinetics and rock surface textures. Fluids can move through brittle rocks most easily by virtue of interconnected fracture and pore permeability, and precipitation of ore minerals in such open spaces coupled with sustained flow rates (nutrient supply) often promote the development of abundant crystal faces. A straightforward textural sequence of precipitation may be recorded by bands or crusts of crystals deposited one after the other along a fracture or pore surface (Figure 1). Vugs or cavities lined with pristine euhedral crystals may remain as a consequence of the eventual cessation of precipitation or fluid flow; such mineral textures can look quite spectacular and are often a reason that people initially become interested in collecting and studying hydrothermal minerals.

Less efficient chemical flow through rocks can occur by the process of diffusion, where fluid components migrate in response to concentration gradients rather than fluid flow (advection). For example, many hydrothermal veins exhibit the development of alteration envelopes, areas of immediately adjacent wall rock in which new minerals have formed via diffusion and infiltration of chemical components from the hydrothermal fluid as it flows through the fracture (Figure 2). New hydrothermal phases such as silicates, carbonates and sulphides often form in these envelopes at the expense of primary phases in the wall rock. Some phases in alteration envelopes, such as potassium feldspars, can prove quite useful in making

Figure 1 Banded, crustiform sphalerite (ZnS, yellow to brown) and marcasite (FeS_2, green) with later pore filling, cross cutting galena (PbS, grey) in polished slab of hydrothermal ore from Olkusz mine, Silesian Cracovian lead zinc district, Poland. US 10 cent coin for scale.

Figure 2 Vein of quartz molybdenite (MoS_2, grey) ore surrounded by alteration envelope of pink alkali feldspar in polished slab of altered porphyry granitoid, Nevada Moly (Hall) mine, near Tonopah, Nevada. US 10 cent coin for scale.

Figure 3 Porphyry Cu Mo ore showing cross cutting silica veins, from Gross Peak pit, Ithaca Peak (Mineral Park) mine, Arizona. US 10 cent coin for scale.

radiogenic age determinations of the age of formation of the veins.

At elevated temperatures rocks become ductile rather than brittle and open fractures and pores cannot be sustained. Under these conditions, hydrothermal mineral deposition occurs more commonly by replacement rather than by open-space filling. Extreme instances of wholesale hydrothermal replacement of wall rock by ore and gangue minerals can accompany metasomatism, a metamorphic process which substantially changes the original wall-rock composition. Both fluid flow and diffusion occur during metasomatism. Skarns are probably the best example of metasomatic hydrothermal ore deposits, where wall-rock carbonate is replaced by coarse calc–silicate minerals. Such deposits typically form near the contact zones of hydrous silicate magmas, with heat and silica provided by the cooling melt.

The textures of hydrothermal ores, especially their cross-cutting relationships, are an important aspect of the study of hydrothermal ore deposits (Figure 1, Figure 3). The term paragenesis is used for the inferred sequence of mineral precipitation based on textural studies of ores. Because most hydrothermal ores consist of a mixture of opaque and non-opaque minerals, the use of reflected-light microscopy in addition to transmitted-light microscopy is essential in their study.

After raw ore is mined, it is ground up and processed in such a manner as to separate the ore and gangue minerals, often by taking advantage of the chemical, density, magnetic, or surface adsorption differences among the particular minerals. For example, cyanide solutions can be used to preferentially dissolve gold from ore, and many base metal sulphide minerals can be efficiently separated from silicate minerals by their different wetability characteristics in oil and water. Once separated, the ore minerals are then smelted and refined to produce the pure metal.

Often there are practical economic constraints in processing ores that may only be revealed by very careful microscopic study of the paragenesis. For example, some hydrothermal gold ores contain native gold or electrum encapsulated in later silica, making grinding and processing to release the gold much more difficult and expensive. In the case of one famous California gold mine, the mine plant and smelter were designed before this practical paragenetic constraint was realized, forcing expensive last-minute changes to the processing circuit!

Hydrothermal Alteration

Because the total volume of rock subject to fluid flow and diffusion is typically much greater than that subject to more localized ore precipitation, hydrothermal alteration accompanying ore mineralization is often a larger and more obvious target for exploration than the ore deposit itself. Economic geologists therefore

look for characteristic alteration mineral assemblages that tend to develop in wall rocks around hydrothermal ore mineralization. Common alteration types in aluminosilicate wall rocks are listed below. For each type, the set of alteration minerals that appears in a rock is a function of its bulk composition.

Advanced argillic:
- alunite, sericite, kaolinite, dickite, pyrophyllite, pyrite

Phyllic:
- sericite, pyrite

Argillic:
- sericite, kaolinite, halloysite, montmorillonite, calcite, biotite, chlorite, pyrite

Propylitic:
- sericite, montmorillonite, zeolite, epidote, calcite, biotite, albite, K-feldspar, ankerite, chlorite, Fe sulphides, Fe oxides, siderite

Potassic:
- sericite, calcite, anhydrite, ankerite, chlorite, biotite, K-feldspar, pyrite, Fe oxides, siderite

Other more monomineralic alteration types include silica and carbonate alteration.

Because the formation of both ores and alteration envelopes are a function of permeability and fluid flow patterns, the geometry of the alteration assemblages can sometimes give clues to the likely location of additional ore during exploration and drilling, as well as aid in reconstructing the palaeohydrology of the ore-forming system. Potassic, phyllic and advanced argillic alteration often mark the highest-temperature, most intense part of a hydrothermal system close to the ore, whereas propylitic alteration often marks the peripheral regions. Silica alteration forms mainly during cooling and tends to occur in the hydrological discharge regions of the system, whereas carbonate alteration typically occurs during heating in the recharge regions.

On the basis of empirical data and proposed genetic models for the origins of different ore types, many national geological surveys have published exploration models that aim to assist in the discovery of new deposits. These idealized models emphasize characteristic alteration assemblages, geochemical fingerprints, geological structures, rock associations and geological ages of major ore deposit types. The models should be viewed as works in progress, subject to revision as new data and theories materialize.

Origins of Ore-Forming Hydrothermal Fluids

The origins of all of the constituents in an ore-forming hydrothermal fluid can be complex and may not be the same, but usually such a fluid can be classified in terms of the origin of its main component, water. The water can be (1) derived directly from magma, (2) derived from ground or connate water heated by magmatic processes, or (3) derived from water heated by non-magmatic processes such as subsidence, tectonism and metamorphism. The following is a list of examples of hydrothermal ore types grouped on this basis.

Magma-hydrothermal fluids:
- pegmatite
- skarn
- porphyry base metal (early stage)

Magma-heated surficial and subsurface waters (including seawater):
- volcanogenic massive sulphide (black smoker)
- porphyry base metal (late stage)
- epithermal vein

Waters heated mainly by tectonism, metamorphism and subsidence:
- Mississippi Valley type (Pb–Zn carbonate)
- stratiform/stratabound base metal
- greenstone belt (orogenic) gold

The origins of water in a hydrothermal ore deposit can sometimes be inferred from the geological setting, but often it also must be based on stable and radiogenic isotope ratios, elemental geochemical signatures, and fluid inclusion compositions in the ore and gangue minerals.

Magma-Hydrothermal Fluids

Magma-hydrothermal fluids arise from the crystallization of magmas that contain water. Magmas can contain significant amounts of dissolved water, sometimes exceeding 10% by weight. They can acquire water initially from the partial melting of 'wet' source rocks containing hydrous mineral phases (amphiboles, micas), and they can also acquire water subsequently via assimilation of hydrous wall rocks during their ascent. In either case, such waters will often exhibit H and O isotopic values that reflect equilibrium exchange with magma at high temperature. Silicic magmas tend to hold more dissolved water than mafic magmas, because of the ability of water molecules to interact with highly polymerized silicate melts.

As a hydrous magma cools and crystallizes, the early precipitation of largely anhydrous silicate liquidus phases (orthosilicates, feldspars) forces the dissolved water to be contained in a progressively shrinking mass of melt. Eventually, the diminishing melt may become saturated in water, forcing exsolution of a separate aqueous fluid (often violently) upon further crystallization. Aqueous fluid exsolution also can be

triggered by a sudden drop in confining pressure on the magma in response to rapid ascent, stoping, seismic events, or collapse of an overlying volcanic edifice.

This aqueous exsolution process, traditionally termed 'resurgent boiling', can result in injection of the magma-hydrothermal fluid into the crystallized outer carapace of the magma as well as the surrounding wall rock. The aqueous fluid often carries incompatible large ion lithophile elements as well as volatile components such as CO_2, HCl, HF, $B(OH)_3$, SO_2 and H_2S. Metal ions may be carried along with the halides and sulphur, typically as metal chloride or sulphide complexes.

As this exsolved fluid reacts with the extraneous rocks and pore fluids it encounters, it cools and evolves chemically; it may also undergo further phase separations (boiling and gas exsolution). Because the stabilities of aqueous metal complexes can be strongly affected by changes in temperature, pressure, pH and other chemical parameters, the cooling, reacting fluid can precipitate hydrothermal ore minerals in open fractures and pores, as well as deposit ore minerals by replacement in suitably reactive wall rock.

Economic geologists look for a number of lines of evidence to link a specific ore deposit to a magma-hydrothermal origin. Field evidence may show spatial and temporal proximity of the ores to a source pluton; for example, veins, breccias or pegmatites which begin in a pluton and then branch out into and cross-cut surrounding wall rock. Hydrothermal alteration developed in the surrounding wall rock may show concentric mineralogical and geochemical zonation around a likely source pluton (**Figure 4**). Evidence can also include remnants of the magma-hydrothermal fluid itself, preserved as aqueous fluid inclusions trapped in the ore and gangue minerals. Such inclusion fluids will typically have very high salinities, high homogenization temperatures and daughter minerals indicative of their origin. Stable isotope ratios of H and O in the minerals may also yield values and calculated equilibration temperatures consistent with a magmatic-hydrothermal origin.

Many magma-hydrothermal ore deposits (such as the numerous skarn and porphyry base metal deposits found in the south-western United States, Chile, Papua New Guinea, and Indonesia) exhibit evidence of early exsolution of a magma-hydrothermal fluid followed by late-stage heating and entrainment of surrounding groundwaters as the pluton

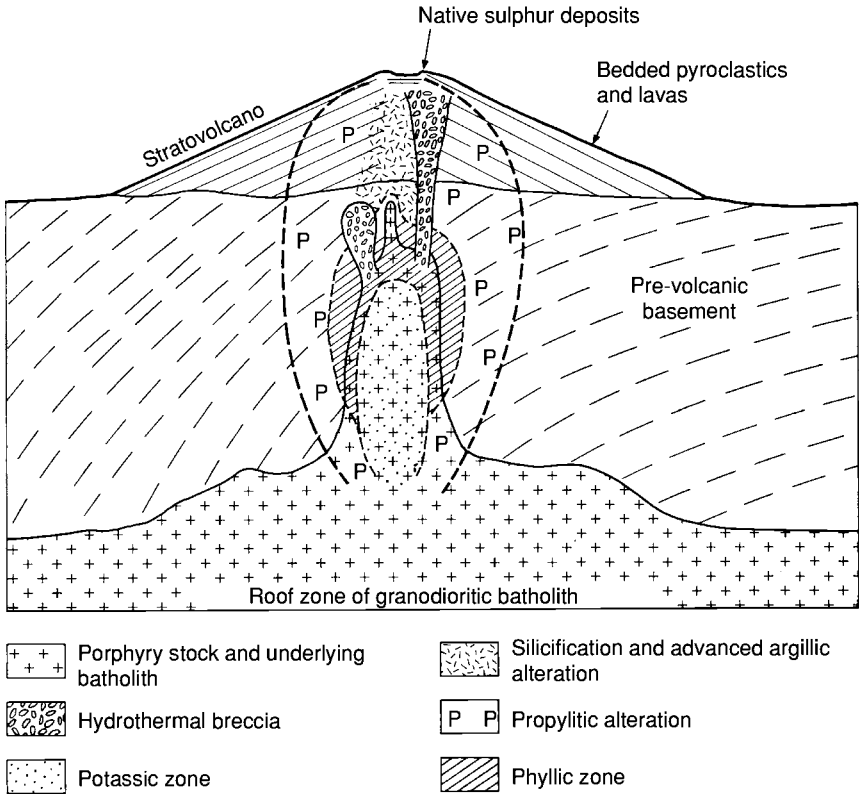


Figure 4 Diagramatic representation of a simple porphyry copper system. (Reproduced with permission from Evans AJ (1993) *Ore Geology and Industrial Minerals: An Introduction*, 3rd edn. Oxford: Blackwell.)

cools. This may result in redistribution of the early-formed ore metals and development of a larger-scale hydrothermal convective meteoric system more akin to the magma-heated systems discussed in the next section.

Examples of some major magma-hydrothermal ore deposits include Grasberg (porphyry and skarn Cu in Indonesia), Bingham (porphyry and skarn Cu in Utah), King Island (skarn W in Tasmania), and Chiquicamata (porphyry Cu in Chile).

Magma-Heated Surface, Ground and Ocean Waters

Some hydrothermal ores owe their origins to waters that were mainly heated by, but not directly exsolved from, magmas. In many cases any aqueous geochemical signature from magma degassing or exsolution is simply swamped out by the availability of shallow water from other sources. Such non-magmatic water sources include meteoric waters, groundwaters, lake waters and ocean waters. There are specific types of plate tectonic settings where such waters often become heated by proximity to magmatic processes.

In continental settings, shallow (epizonal) magmas often ascend into close proximity to surface and subsurface waters in volcanic arcs. Being topographically elevated, active continental volcanic arcs can receive abundant meteoric precipitation and are sufficiently porous and permeable (due to recurring tectonism and volcanism) that rapid recharge of their subsurface waters occurs. Crater lakes, caldera lakes, and tectonic lakes (formed from landslide blockage of drainages) provide additional reservoirs of surface water for recharge in arcs. At higher elevations and latitudes, snow and glaciers on the flanks of volcanoes can provide recent meteoric water upon melting.

This combination of shallow heat sources plus readily available meteoric water in permeable rocks creates ideal conditions for shallow convective hydrothermal systems, in which cold meteoric recharge waters descend, become heated by magmas, and rise buoyantly to discharge at or near the surface. Fumaroles, hot springs, geysers and mud pots at the surface attest to such fluid circulation cells at depth. As the heated meteoric waters circulate through the wall rocks and even the outer crystallized portions of a magma, they can acquire dissolved salts and metals and become ore-forming fluids, re-precipitating the metals as ores when they cool, boil or react with shallower groundwaters and rocks. Such shallow vein and disseminated hydrothermal ores are often termed epithermal on account of their shallow depth of formation. Examples include Creede, Colorado and Hishikari, Japan.

Arc magmas do not passively contribute only heat to overlying epithermal systems. Exsolving magmatic gases such as SO_2 and H_2S can become entrained in epithermal circulation systems, oxidizing to sulphuric acid near the surface. This results in a common association between some epithermal ores and acidic alteration (micas, clays, sulphates) of shallow wall rocks. Metals from magmas may also be exsolved into shallow epithermal systems, so they really should be viewed as hybrid systems with variable degrees of chemical input from magmatic and non-magmatic sources. Epithermal ores with evidence for a high contribution of components (especially acidic sulphur gases) from shallow degassing magmas are called high sulphidation state ores; they typically form in settings where a crater lake sits right above a shallow volcanic magma. Examples include Summitville, Colorado and El Indio, Chile.

Figure 5 schematically shows examples of many types of settings for hydrothermal systems in which epithermal and mesothermal gold deposits are formed.

It should not be surprising that magma-hydrothermal and epithermal hydrothermal processes can overlap in time and space, or that transitions between these processes should be common. Economic geologists use tools such as fluid inclusions and stable isotopes (O, H, S) to evaluate the relative roles of magmatic and meteoric components in the genesis of specific ore deposits. It is not uncommon to find deep, largely magma-hydrothermal ore deposits (such as porphyry base metal or skarn deposits) beneath temporally equivalent epithermal deposits (such as gold and silver vein and disseminated deposits) in shallower wall rock and volcanics.

Continental rifts are another plate tectonic setting where magmas and non-magmatic waters can interact to form hydrothermal ores. In such rifts, extension, subsidence and rapid sedimentation take place in conjunction with bimodal volcanism, providing opportunities for shallow basaltic and rhyolitic magmas to interact with lakes and groundwaters. Moreover, evaporation of closed-basin surface waters in rifts can lead to the development of very saline brines, which when heated by magmatism have tremendous capacity to dissolve, carry and re-precipitate ore metals derived from sedimentary and volcanic sources. Subsurface vein and disseminated ores can form in rift sediments and volcanics, along with overlying stratiform lenses and layers of ores formed when upwelling brines are injected or expelled parallel to shallow sediment layers. The Salton Trough of western North America (Table 1) and the Red Sea are modern examples of such settings.

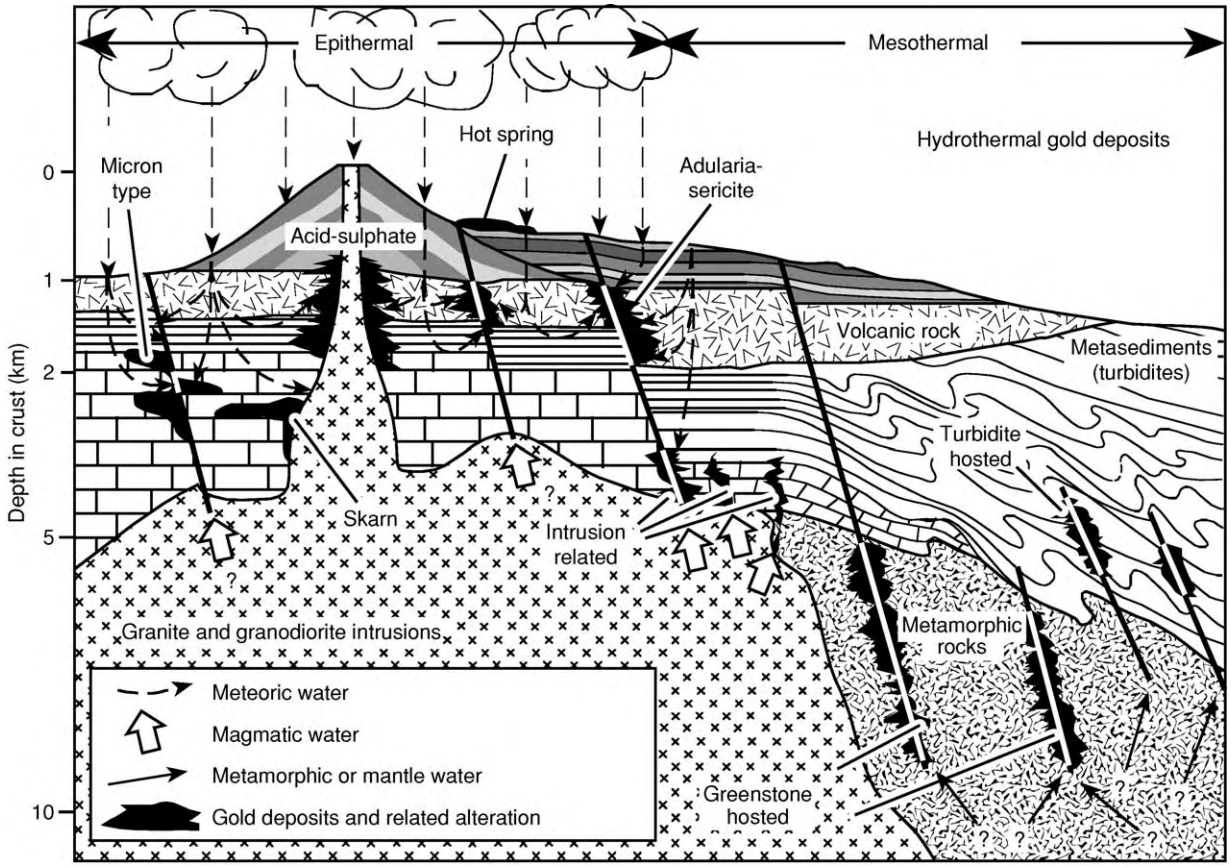


Figure 5 Schematic illustration of the wide variety of occurrences of hydrothermal gold deposits in continental crust. The terms epithermal and mesothermal refer to shallow crustal and mid crustal level deposits, respectively. (Reproduced with permission from Kesler SE (1994) *Mineral Resources, Economics and the Environment*. New York: MacMillan.)

In oceanic settings, volcanic arcs and rifts can lead to the formation of hydrothermal ores similar in principle to those found on continents. The main differences are the pervasiveness of seawater as a source of water and the dominance of oceanic crust and basaltic magmatism. Mid-ocean ridges, where roughly two-thirds of the Earth's heat flow occurs, develop large-scale hydrothermal convection systems that flux tremendous amounts of seawater through hot oceanic crust. The cold, oxidized metal-deficient and sulphate-rich seawater reacts with the hot basaltic crust and becomes a hot, reduced metal- and H_2S-rich ore-forming fluid. The metals that have been mobilized from the oceanic crust are re-deposited as vein and disseminated sulphide ores in the subsurface of the ridge structure, as well as expelled from submarine hot springs (black smokers) to form mounds and precipitates of metal sulphide on the seafloor. Only within the last few decades have earth scientists become aware of the spectrum of oceanic hydrothermal deposits that form at oceanic ridges, seamounts and back-arc basins. Many volcanogenic massive sulphide (VMS) deposits, long known to have spatial associations with altered mafic rocks, are now recognized as remnants of ancient oceanic ridge and back-arc hydrothermal systems that have since been accreted to or obducted onto the continents. The Cu–Pb–Zn districts of Kuroko, Japan and Jerome, Arizona are ancient examples; modern settings include massive sulphide deposits forming along the Juan de Fuca ridge in the north-eastern Pacific.

Hydrothermal Fluids Not Directly Affiliated with Magmatic Processes

Hydrothermal fluids and ores can be formed via heating mechanisms that have no direct association with magmatic processes. Basin subsidence through the normal geothermal gradient can heat sediment pore fluids and induce free convection, enhancing their ability to carry and redistribute metals. Such processes may be responsible for unconformity-related basinal uranium deposits such as those found in Saskatchewan (Canada) and Northern Territory (Australia). Rapid subsidence may also induce

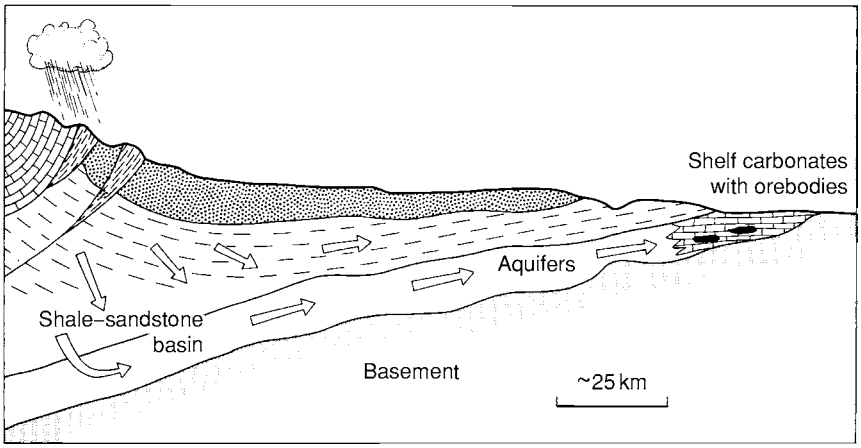


Figure 6 Schematic model for formation of carbonate hosted Mississisppi Valley type Pb Zn ores, via lateral gravity driven flow of basinal brines from an upland region. (Reproduced with permission from Evans AJ (1993) *Ore Geology and Industrial Minerals: An Introduction*, 3rd edn. Oxford: Blackwell.)

compaction and overpressuring of basin sediments, promoting the flow of warm pore fluids.

Pore fluids in deforming sedimentary basins adjacent to uplifting forelands can be forced to flow laterally in response to gravity, a scenario that is best developed during continental plate collisions. A widely accepted theory for the origin of many carbonate-hosted Mississippi Valley-type (MVT) Pb–Zn deposits in central and eastern North America is westward gravity-driven flow of warm basinal brines, caused by Late Palaeozoic collisions between the ancient North American and European continents (Figure 6). Compositional and colour banding ('sphalerite stratigraphy'), fluid inclusions, and other geochemical features of the ores are remarkably consistent over long distances between the Pb–Zn deposits, supporting the occurrence of regional fluid-flow events. Gravity-driven lateral flow of basinal brines has emerged as the most effective mechanism that can account for the thermal and mass balances in many MVT systems.

Overthrusting and consequent sediment compression in fold-mountain belts may also drive the flow of deeper, warmer fluids into shallower or more peripheral environments. Shale-hosted stratiform copper deposits such as the Kupferschiefer (Europe) and White Pine (North America) may have formed via this type of mechanism.

Dehydration reactions associated with regional metamorphism can also provide sources of hot fluids, although they are initially low in salinity and high in CO_2. Such fluids may have been responsible for hydrothermal gold deposits associated with CO_2 metasomatism, found in shear zones and faults in Archaean and younger greenstone belts. Earthquake-induced seismic pumping could provide a mechanism of fluid flow along fault zones.

See Also

Economic Geology. **Gold**. **Minerals:** Native Elements; Sulphides. **Mining Geology:** Hydrothermal Ores. **Tectonics:** Hydrothermal Activity.

Further Reading

Barnes HL (1979) *Geochemistry of Hydrothermal Ore Deposits*, 2nd edn. New York: Wiley Interscience.

Barnes HL (1997) *Geochemistry of Hydrothermal Ore Deposits*, 3rd edn. New York: John Wiley.

Barton PB, Rye RO, and Bethke PM (2000) Evolution of the Creede Caldera and its relation to mineralization in the Creede mining district, Colorado. In: Bethke PM and Hay RL (eds.) *Ancient Lake Creede; Its Volcano Tectonic Setting, History of Sedimentation, and Relation to Mineralization in the Creede Mining District*, vol. 346, *Special Paper Geological Society of America*, pp. 301 326.

Berger BR and Bethke PM (1985) *Reviews in Economic Geology, vol. 2: Geology and Geochemistry of Epithermal Systems*. El Paso, Texas: Economic Geology Publishing Company.

Cox DP and Singer DA (1986) *Ore Deposit Models*. US Geological Survey Bulletin 1693.

Craig JR and Vaughan DJ (1994) *Ore Microscopy and Ore Petrography*, 2nd edn. New York: Interscience.

Evans AJ (1993) *Ore Geology and Industrial Minerals: An Introduction*, 3rd edn. Oxford: Blackwell.

Giggenbach WF (1992) Magma degassing and mineral deposition in hydrothermal systems along convergent plate boundaries. *Economic Geology* 87: 1927 1944.

Guilbert JM and Park CF Jr (1986) *The Geology of Ore Deposits*. New York: WH Freeman.

Hedenquist JW and Lowenstern JB (1994) The role of magmas in the formation of hydrothermal ore deposits. *Nature* 370: 519 527.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (1995) *Seafloor Hydrothermal Systems: Physical, Chemical, Biological and Geochemical Interactions*. AGU Geophysical Monograph 91, Washington, DC: American Geophysical Union.

McKibben MA and Hardie LA (1997) Ore forming brines in active continental rifts. In: Barnes HL (ed.) *Geochemistry of Hydrothermal Ore Deposits*, 3rd edn., pp. 875 933. New York: Wiley Interscience.

Pirajno F (1992) *Hydrothermal Mineral Deposits: Principles and Fundamental Concepts for the Exploration Geologist*. Heidelberg: Springer Verlag.

Rye RO (1993) The evolution of magmatic fluids in the epithermal environment: the stable isotope perspective. *Economic Geology* 88: 733 753.

Thompson JFH (1995) *Magmas, Fluids and Ore Deposits*. Mineralaogical Association of Canada, Short Course Series 23.

White DE (1981) *Active Geothermal Systems and Hydrothermal Ore Deposits*, pp. 392 423. Economic Geology Seventy Fifth Anniversary Volume. El Paso, Texas: Economic Geology Publishing Company.

White NC and Herrington RJ (2000) Mineral deposits associated with volcanism. In: Sigurdsson H (ed.) *Encyclopedia of Volcanoes*, pp. 897 912. San Diego: Academic Press.

Magmatic Ores

J E Mungall, University of Toronto, Toronto, ON, Canada

Introduction

Magmatic ore deposits may be defined as rocks of igneous origin (*see* **Igneous Processes**), which can profitably be mined for their constituent chemical elements. For example, the worth of annual global production from magmatic ore deposits exceeded $10 billion in 2001, and dominates or contributes significantly to world supplies of elements representing more than a third of the periodic table. This article summarises the petrogenetic controls on the genesis of magmatic ore deposits, with an emphasis on the application of geochemical models. Readers are encouraged to consult the Further Reading Section at the end of this article for detailed descriptions of the multitude of types of magmatic ore deposits (Figure 1).

Fundamental Controls

The generation of a magmatic ore deposit depends upon the successful operation of four fundamental processes. Elements that are normally widely distributed at low concentrations must be extracted from a large volume of rock by a natural melting event. The resulting magma must be collected into or channelled through a relatively small volume of the Earth's crust, such as a conduit or magma chamber. While within the conduit or magma chamber the magma must become saturated with a phase within which the element of interest is highly concentrated. The phase containing high concentrations of the target element must then be mechanically sorted from the remainder of the magma, and collected in sufficient quantity and at sufficient grade to constitute an economically attractive ore deposit (*see* **Economic Geology**). The study of magmatic ore deposits is, therefore, inextricably tied to studies of the partitioning of elements between coexisting phases in magmas, and of the processes of intrusion and differentiation of magmatic systems (Figure 2).

Element Partitioning

The key to a quantitative approach to the description of the evolution of a magmatic system lies in the distribution of chemical components among the coexisting phases. The distribution of an element between two phases at equilibrium may be described through the use of a partition coefficient D, which is defined as the ratio of the concentration of the element in one phase (e.g., mineral or sulphide melt) to its concentration in another (e.g., silicate melt) (*see* **Minerals:** Sulphides); for example, the partition coefficient for Cu between sulphide and silicate melts can be expressed as

$$D_{Cu}^{sulphide/silicate} = \frac{C_{Cu}^{sulphide}}{C_{Cu}^{silicate}} \quad [1]$$

where C denotes concentration of the subscripted element in the superscripted phase. Partition coefficients can be used in quantitative thermodynamic calculations if the element in question is a trace element whose concentration varies through ranges too small to affect D (Henry's Law is obeyed). If the element is a major or stoichiometric constituent of one of the phases, then D loses its utility as a constant. Values of

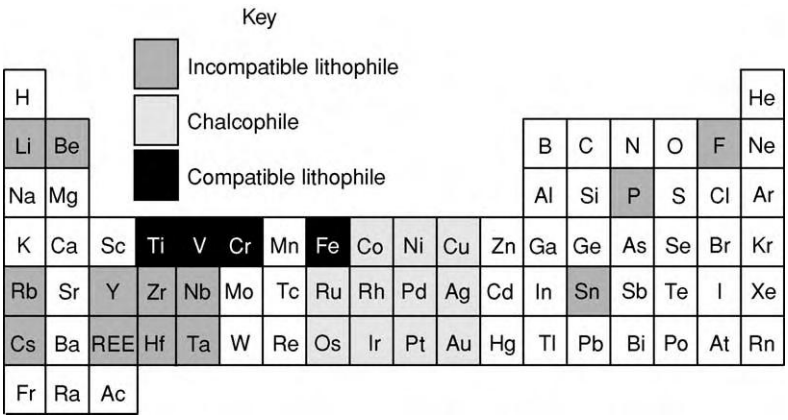


Figure 1 Elements extracted in significant quantities from magmatic ore deposits.

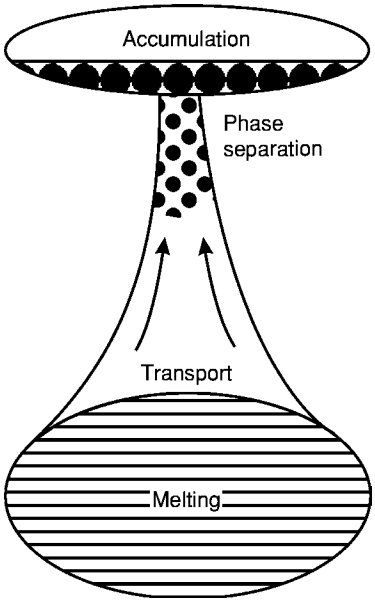


Figure 2 Processes required to generate a magmatic ore deposit.

D quoted in this contribution are intended to be used not as thermodynamic constants but rather as approximate indicators of the affinity of elements for various magmatic phases.

If two phases coexist at equilibrium in a closed system, eqn [1] can be combined with mass balance constraints to give concentration of an element in the melt as a function of the degree of partial melting:

$$\frac{C_i^{melt}}{C_i^{bulk}} = \frac{1}{D_i^*(1-F)+F} \qquad [2]$$

where C_i^{bulk} is the concentration of element i in the entire system, F is the fraction of melt present and D_i^* is a bulk partition coefficient for element i between solids and melt, obtained by summing over all solids the products of the individual phase/melt D and the weight fraction of the phases in the total solid assemblage.

Two processes that may occur during the differentiation of a magma are of critical importance to the generation of magmatic ore deposits. The first is the fractional removal of crystalline phases as the magma loses heat by assimilation of wall rock or by conduction. The second is the separation of an immiscible Fe-S-O liquid if the magma exceeds its carrying capacity for sulphide. Both processes govern the concentration and eventual segregation of ore metals from magmas after melting has occurred.

The effect of fractional crystallization on the concentration of a trace element with constant solid/liquid partition coefficient D_i^* is described by the Rayleigh distillation equation:

$$C_i^{melt} = C_i^o F^{(D_i^* \ 1)} \qquad [3]$$

where C_i^o is the initial concentration of an element and F is the proportion of liquid remaining. Application of eqn [3] to a highly incompatible element in the limiting case of complete solidification would cause the singular occurrence of infinite concentration of a trace element in the last vestiges of melt. What occurs instead is that, at some point, the concentration of a trace element rises to the point at which a mineral that concentrates that element begins to crystallize, effectively increasing D and avoiding the singularity. Depending on the solubility of the concentrating phase, and the value of the element in question, an ore deposit may be generated simply by the intrusion and solidification of a magma that contains sufficient concentration of the element solely through the action of fractional crystallization from a magma which originated by small degrees of partial melting of a

fertile source. Metals which begin to crystallise in a concentrating phase, when the bulk concentration of the element in the magma itself is has not yet attained ore grade, must undergo a final concentrating stage in which the mineral phase carrying the element is physically separated from the magma and collected in one place to form an ore body.

The effect of separation of sulphide melt from silicate melt is usually described using the assumption of batch equilibration of the two phases in a closed system:

$$C_i^{sulphide} = \frac{C_i^{bulk}(R+1)}{R + D_i^{sulphide/silicate}} \qquad [4]$$

where R is the ratio of the mass of silicate melt divided by that of sulphide melt, and C and D are defined as above. Equation [4] can be derived from eqn [2] by redefining some of the terms, and is effectively identical to the equilibrium crystallization equation.

Based upon their relative affinity for the three principal types of phase present in magmatic systems, the elements shown in Figure 1 may be classed into three main groups (*see* Table 1). The chalcophile elements are partitioned strongly into sulphide melt at equilibrium with oxides or silicates, whereas the lithophile elements are those which reside in silicate or oxide phases in preference to metal or sulphide liquids. The lithophile elements may be subdivided into those which are partitioned into oxide or silicate minerals (the compatible lithophile elements) and those which are preferentially retained in silicate melt (the incompatible lithophile elements).

There is considerable overlap between the three groups of elements listed above; for example, V shows variable compatibility because it is an incompatible lithophile element during melting of the mantle, but becomes highly compatible when a magma is saturated with ilmenite or magnetite. Pd is strongly chalcophile, but in the absence of sulphide melt it behaves as an incompatible element.

Incompatible Lithophile Elements

Some combination of very small degrees of partial melting and fractional crystallization to very small residual melt fraction is required to elevate the concentrations of highly incompatible lithophile elements from their trace abundances in the bulk silicate Earth to the weight percent levels necessary for economic extraction. The principal host silicate magmas are granitic pegmatites or highly evolved alkaline or peralkaline felsic plutons. Carbonatites are a related class of rocks derived from carbonate magmas, either through liquid immiscibility with felsic alkaline rocks, or directly by extremely small degrees of partial melting of carbonate-rich mantle.

Incompatible lithophile element deposits may be subdivided into the Li-Cs-Ta (LCT) and the Nb-Y-F (NYF) classes. Deposits of the LCT class are almost exclusively granitic pegmatites associated with highly fractionated intrusions of reduced ilmenite-series (S-type) granites, whereas the NYF class tend to be derived from alkaline or peralkaline felsic intrusions.

LCT pegmatites are thought to originate by fluid-absent melting of pelitic or semipelitic metasedimentary schists in Abukuma-type regional metamorphic belts. Pelitic sediments are generally rich in alkali elements including Li and Cs, and also contain large amounts of F and B. Incongruent melting of white micas and complete melting of tourmaline produces water-poor peraluminous magmas rich in the fluxing components P, F, and B. The compositional undercooling caused by degassing of H_2O as water-rich magmas rise through the crust, causes them to

Table 1 Summary of partitioning behaviour of metals

Element group	*Examples*	D_i^*	*Concentrating phases*	D_i	*Magmatic environments of ore deposition*
Incompatible lithophile	Li, Be, F, Rb, Y, Zr, Nb, Cs, REE, Hf, Ta	<0.1	Melt, rare element minerals	na	Pegmatite fields alkaline intrusions carbonatites
Variable compatibility	P, Ti, V, Cr, Fe	<0.1	Oxide minerals, apatite	10 100	Stratiform intrusions, podiform bodies immiscible oxide melts
Incompatible chalcophile	Rh, Pt, Pd, Au, Cu	<0.1	Sulphide melt	2000 50 000	Magmatic sulphides
Compatible chalcophile	Os, Ir, Ru, Ni	1 10	Olivine, oxide minerals	500 50 000	Chromitites magmatic sulphides

D_i^* is the bulk partition coefficient between mantle minerals and silicate melt during conditions typical of the generation of basaltic magmas. D_i is the partition coefficient for the phase which ultimately concentrates the element (e.g., sulphide melt for chalcophile elements, oxides for Ti, V, Cr). Elements with variable compatibility are those which are generally incompatible during melting but which are concentrated in minerals that crystallize from basaltic to intermediate magmas.

freeze at their eutectic before they can undergo extensive fractional crystallization. In contrast, the water-poor magmas parental to LCT pegmatites do not exsolve a free vapour phase, and do not evolve toward the water-saturated eutectic in the haplogranite system. Instead, LCT magmas are high-variance assemblages that can continue to evolve down to temperatures as low as 450°C, while continously becoming enriched in the fluxing elements F, B, and P as well as the trace elements including Li, Cs and Ta. The residual liquid may be enriched in the lithophile elements to such an extent that its bulk composition may be of ore grade. This magma can then be injected into the country rocks surrounding the cooling pluton and crystallized as swarms of rare-element pegmatite dikes (Figure 3).

The NYF suite of pegmatites and related plutons are probably derived by small degrees of partial melting of upper mantle similar to the sources of ocean island basalts, to produce alkaline magmas or magmas transitional from alkaline to tholeiitic affinity. Such small degree melts are enriched in incompatible elements like Nb, Zr, and Y, and can evolve through a prolonged process of fractional crystallization to peralkaline silica-saturated (pantelleritic) or silica-undersaturated (phonolitic) products. Deposits of the NYF suite of elements may form either directly by crystallization of the rare-element-rich magma. Alternately, the incompatible lithophile elements (e.g., Be, Nb, Y, Zr) may be transported by peralkaline fluids expelled from the cooling pluton, to be deposited where they mix with meteoric or formational waters surrounding the pluton.

There is an immense variety of rare-element pegmatites and related rocks, and there are undoubtedly a number of deposits whose genesis differs in significant ways from the models described here. However, the key requirements for any rare-element magmatic ore deposit must be the essential combination of small degrees of partial melting of fertile source rocks, and extreme fractional crystallization, both processes acting to concentrate these elements in progressively smaller volumes of silicate magma.

Carbonatites (*see* **Igneous Rocks:** Carbonatites) are igneous rocks that contain more than 50% carbonate minerals (*see* **Mining Geology:** Mineral Reserves). They host a great diversity of ore deposits, but the primary commodities extracted from these enigmatic rocks are rare earth elements (REE) and Nb (for which carbonatites are the principal source) as well as Fe, P, F, V, Zr, and Cu. Carbonatite may occur as small, usually concentrically-zoned plutons, or as lavas or pyroclastic deposits of regional extent.

Since carbonate is immiscible with alkaline CO_2-rich magmas, almost any alkaline or peralkaline magma has the potential to form an immiscible carbonatite liquid, given sufficient excess CO_2. Associated carbonatite and alkaline silicate magmas are found in tectonic settings including hotspots and convergent margins. The primary carbonatite magma is thought usually to be calcitic, and concentrates elements with high ionic field strength such as Nb, P, Zr, and F relative to coexisting silicate melt. Carbonatites evolve by fractional crystallization to dolomitic compositions which may become saturated with ore minerals such as pyrochlore, magnetite, and apatite. Late-stage carbonatite magmas may evolve to extremely sodic or ferroan compositions, and can crystallize important volumes of REE minerals such as monazite and bastnaesite within core zones or dykes cutting across carbonatite-alkaline rock complexes.

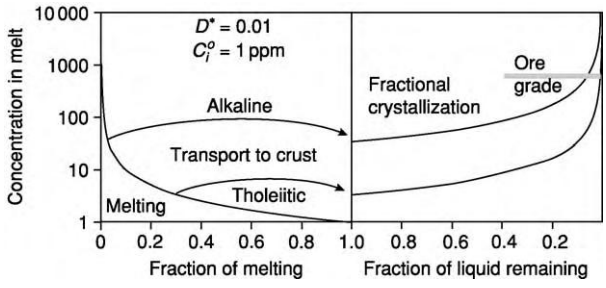


Figure 3 Formation of incompatible element deposits of the NYF class. Melting curve was calculated with eqn [2]; fractional crystallization with eqn [3], for a hypothetical element similar to Nb, Y, or Zr. Arrows represent the extraction of magmas from the mantle to the crust. Only alkaline magmas derived by small degrees of partial melting (left panel) can evolve through fractional crystallisation to very high concentrations of rare lithophile elements (right panel).

Oxides

Cumulates

Oxide minerals hosted by igneous rocks constitute important or principal sources of V, Ti, Cr and, to a lesser degree, Fe. There is considerable controversy surrounding the origins of many magmatic oxide deposits, between proponents of models for their deposition as cumulates from silicate magmas, or as oxide magmas derived by immiscibility with silicate magmas.

There is little doubt that the extremely refractory mineral chromite is deposited as a cumulus phase from silicate melts. Chromite deposits occur either as stratiform horizons in layered intrusions or as podiform accumulations in dunitic rocks. Mixing of two magmas cosaturated with olivine and chromite

will lead to oversaturation of the hybrid magma with chromite alone. Since primitive magmas commonly have only olivine and chromite on their liquidus, each magma recharge event in a large mafic magma chamber will lead to the temporary cessation of olivine accumulation, and the formation of a crystal layer composed solely of chromite. The thickness of the layer that forms will depend upon the temperatures and proportions of the magmas involved as well as the height of the magma column above the cumulate pile. Large layered intrusions, such as the Bushveld intrusion of South Africa, commonly contain numerous chromitite layers in their ultramafic basal series, the largest of which can be mined.

The mode of formation of podiform chromitites has been a more difficult conundrum than that of stratiform bodies. Podiform chromitites are found as wispy, nodular, or layered masses entirely enclosed within pipe- or dyke-like bodies of dunite in the harzburgitic mantle section of ophiolite complexes. Arai recently proposed that the same magma mixing mechanism used to explain the formation of stratiform chromitites can be applied to podiform chromitites. In his model, the more primitive magma is carried by porous flow through the dunite pipe. Melts that have migrated into successively lower-temperature environments farther from the core of the pipe become more evolved. Hybrid, chromite-oversaturated magma forms in places where the flow regime allows some of the cooler, more evolved melt to be drawn back into the core of the system. If the system reaches a steady state, chromitite pods can form and grow to considerable size in the zone of mixing between the two magmas.

Titanomagnetite can be deposited as stratiform horizons in large layered intrusions in much the same manner as chromite. Several processes have been proposed to explain how magnetite can arrive on the liquidus of a basaltic magma, to form monomineralic oxide layers in gabbroic cumulate rocks. The simplest involves magma mixing, analogous to the chromite saturation mechanism. Others include sudden changes in pressure or oxygen fugacity. Whatever the cause, once magnetite begins to crystallize from a basaltic magma, the transition metals Ti and V experience a sudden change from incompatible to strongly compatible behaviour. The first layer of magnetitite to form can constitute an ore of Ti and V, with Fe as an important byproduct. Because magmas form with a range of V and Ti contents, and because the partitioning behaviour of V is extremely sensitive to oxygen fugacity, there is a wide possible range of V abundance in magnetitites, and only a few are of economic grade. Stratiform accumulations of magnetitite are the world's primary repositories of V ore.

Immiscible Oxide Liquids

There are a number of deposits of Fe-Ti oxide (ilmenite and titanomagnetite with or without accessory apatite) worldwide that show textural and structural features, suggesting that they were emplaced as oxide magmas or lavas. There appears to be little doubt that the world's largest Ti deposits (Lac Tio, Canada; Tellnes, Norway) consist of intrusive sills and dykes composed of ilmenite, magnetite, and apatite. More controversial are the magnetite deposits of Kiruna, Sweden, and the El Laco deposit in Chile. In oxide rocks that have mesoscale textures and structures that perfectly replicate shallow subvolcanic and extrusive rocks, including pyroclastic flows, bombs, and lava flows, both deposits present textures closely resembling dendritic quench textures seen in slags. On the other hand, no experimental study has yet demonstrated that iron oxides can exist as liquids at reasonable magmatic temperatures. Furthermore, there is abundant evidence for marginal metasomatic replacement of common lavas and pyroclastic rocks by the same mineral assemblage as is found in the allegedly magmatic oxide rocks, raising the possibility that the oxide deposits may have formed by hydrothermal replacement of pre-existing silicate extrusive rocks. Controversy will likely continue for many years to come.

Chalcophile Elements

The upper mantle is thought to contain about 250 ppm of sulphur in the form of sulphide minerals. Under the conditions of formation of basaltic magmas, the sulphide will be entirely molten, so that the chalcophile elements will be sequestered within the sulphide melt. Under these conditions, the highly chalcophile elements will not enter the basaltic magma in significant quantities, and even the more moderately chalcophile elements such as Cu and Ni will be impoverished in the basalt. Since sulphide is sparingly soluble in basaltic magmas, progressively higher degrees of partial melting will dissolve progressively more of the sulphide. Upon approximately 25% partial melting of the mantle, all sulphide originally present will have been dissolved by the silicate melt, whereupon the chalcophile elements will now reside in a system containing only silicate minerals and silicate melt. Most chalcophile elements behave as incompatible lithophile elements in the absence of a free sulphide phase, so that at high degrees of partial melting of the mantle the chalcophile elements can be extracted into basaltic magmas and transported toward the surface. Such high degrees of melting are generally thought to be related to the melting of

mantle plumes (*see* **Mantle Plumes and Hot Spots**), and result in the extraction of picritic or komatiitic magmas (*see* **Igneous Rocks:** Komatiite).

Suprasubduction zone magmas typically remain sulphide-saturated during melting due to the introduction of sulphide to their source regions by the same fluids that induce partial melting (Figure 4). At best, arc magmas initially saturated in sulphide can generate low-grade Ni-Cu deposits with exceedingly low precious-metal contents.

Once a magma separates from its source and begins to rise toward the surface, it will fall into a state of sulphide-undersaturation due to the inverse dependence of sulphide solubility on pressure in silicate melts. When deeply-sourced mafic or ultramafic magmas reach the Earth's surface, they typically contain less than half the amount of sulphide necessary to induce sulphide saturation.

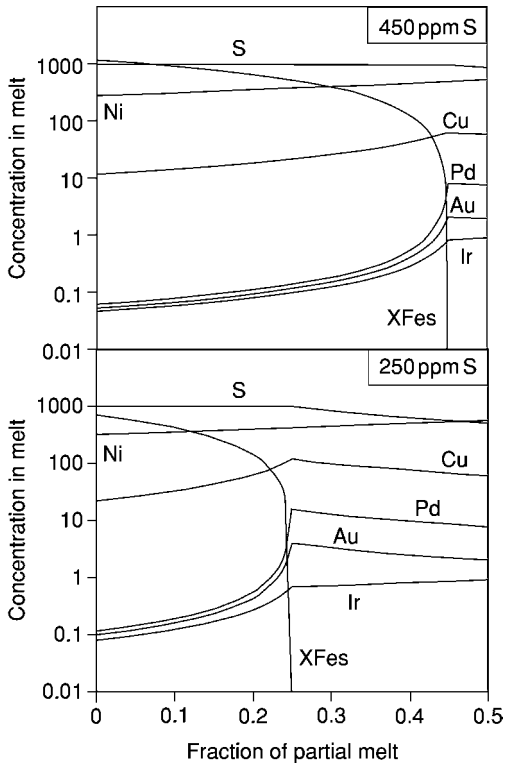


Figure 4 Concentration of chalcophile elements in basaltic magma during partial melting of mantle containing either 450 ppm S (metasomatized sub arc mantle; upper panel) or 250 ppm S (plume head melting; left panel). Chalcophile element concentrations increase as the amount of sulphide remaining in the source (XFes) diminishes. At the moment when sulphide is completely consumed, incompatible chalcophile elements reach their maximum possible concentrations, whereas compatible chalcophile elements (Ni, Ir) continue to increase with further melting.

Magmatic Sulphide Deposits

Magmatic sulphides represent accumulations of immiscible sulphide liquid separated from silicate magmas, and account for the majority of the value of metal produced from magmatic ore deposits worldwide. Although all magmatic sulphide deposits are intrinsically polymetallic concentrations of the base metals Fe, Ni, Cu, and the precious metals Au, Ag, and the platinum group elements (PGE; especially Pt and Pd), it is convenient to categorise them as either precious metal or base metal deposits, each of which yields the elements of the other group as important byproducts.

Precious metal sulphide deposits generally contain less than 1 modal % sulphide, whereas base metal sulphide deposits contain 5 to 100 modal % sulphide. The distinction is an arbitrary one, but it is a consequence of the extreme sensitivity of precious metal concentrations to the mass ratio of silicate to sulphide melt that equilibrate during sulphide segregation, as illustrated in Figure 5. Precious metal sulphide deposits typically form where very small amounts of sulphide have separated from silicate melt and collected in stratiform bodies within large mafic intrusions. Base metal-dominated magmatic sulphide deposits form by the collection of immiscible sulphide liquid

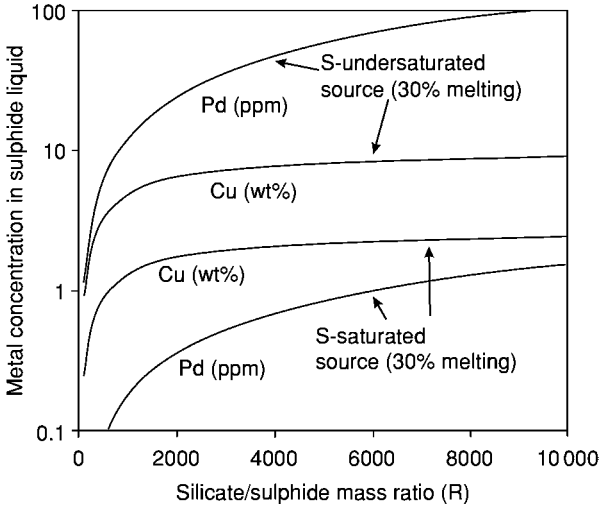


Figure 5 Concentrations of Cu and Pd in sulphide liquid equilibrated with basaltic liquids as functions of the silicate/sulphide mass ratio (R in eqn [3]). The curves are calculated assuming that the initial basalt composition resulted from 30% partial melting of the upper mantle, using the compositions for the sulphide saturated and sulphide undersaturated cases shown in Figure 4. Sulphide rich systems ($100 < R < 1000$) can generate large quantities of base metal rich sulphide which contains relatively low PGE contents. At very high values of R the resulting sulphide is PGE rich, but correspondingly smaller in volume. Magmas derived from sulphide saturated mantle sources can only generate low grade base metal deposits without significant PGE.

from mantle-derived mafic, or ultramafic silicate magmas in lava channels, magma conduits, or places where conduits feed into larger bodies. The Sudbury mining camp, Canada, the world's largest collection of Ni-Cu-PGE deposits, is a unique exception. The Sudbury Igneous Complex and its base-metal sulphide ore deposits formed by the batch equilibration and gravitational segregation of sulphide liquid with several thousand km^3 of shock-melted continental crust following a large meteorite impact into a shallow marine basin filled with sulphidic sediments.

Base Metal Sulphide Deposits

A more common path to sulphide saturation in fertile, previously sulphide-undersaturated magmas is the assimilation of crustal rocks rich in sulphide. This has the combined effect of lowering the temperature of the magma while sharply increasing the amount of sulphur it contains. Addition of large quantities of sulphide can drive a magma into a state of extreme sulphide-oversaturation, so that the sulphide equilibrates at a relatively low silicate/sulphide ratio ($100 < R < 5000$; eqn [3]). The result can be the generation of sulphide in sufficient quantity to separate and form economically attractive tonnages of the base metal sulphides, however, at these relatively lower silicate/sulphide ratios, the concentrations of the precious metals in the resulting sulphide tend to contribute a relatively small part of the total value of the ores.

Emplacement of magmatic sulphides The dominant mode of transport of magmas in the crust is by the emplacement of dykes and sills. Figure 6 provides schematic illustrations of the most important mechanisms for sulphide segregation. Two phenomena related to dyke and sill emplacement are of particular importance in the generation of magmatic sulphide deposits. These are the formation of long-lived conduits, and the formation of channelized lava flows. Because thin, tabular sheets of magma cool rapidly, flow along dykes and sills tends to become focused along their widest portions, which then enlarge themselves by heating and assimilating their wall rocks. The result is the generation of long-lived blade-shaped or tubular magma conduits in which flowing magma cools relatively slowly. Similarly, large fissure-fed lava flow fields are thought to develop distributary systems resembling deltas, in which deeply incised lava channels proximal to the vent feed smaller branching channels and distal sheet or pillowed flow lobes. Within either a conduit or a lava channel, it is possible for a steady state to be reached, in which the temperature of magma passing a given point remains constant for extended periods of time. If the rate of supply of sulphide by degassing or erosion of the channel floor or conduit walls, and the rate of cooling and crystallization also approach a steady state, then sulphide liquid first reaches saturation in the magma always as it passes a single point within the conduit or channel. Sedimentation of sulphide droplets from the flowing lava, perhaps also with a co-saturated phenocryst phase (olivine or pyroxene), can thus lead to highly efficient sorting of the phase and the elements it concentrates from a very large volume of magma into a small, localized volume of rock.

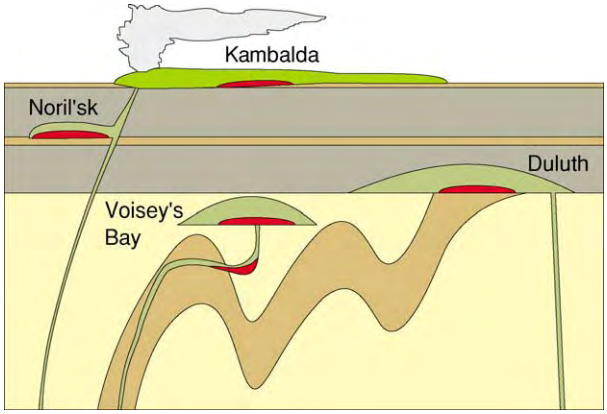


Figure 6 Cartoon of principal mechanisms for the generation and segregation of magmatic base metal sulphides. The image shows a cross section of the upper crust, with brown areas representing sulphur rich crustal rocks which act as sources of sulphur for dykes and plutons of mantle derived magma (darker green) or lava flows (lighter green). Sulphide liquid collects where it formed (Duluth), in traps within conduits or where con duits enter magma chambers (Voisey's Bay), in subvolcanic sills (Noril'sk), or in lava channels cutting through sulphidic sediments (Kambalda).

A related possibility with a similar result is that a magma which has become saturated with sulphide liquid may carry the exsolved sulphide suspended as small droplets in the rapid and turbulent flow within the conduit. A transition within the dyke to laminar flow or a sudden drop in flow velocity due to its passage through a magmatic breccia or a widening of the conduit will then cause the entrained sulphides to settle together in one place within the conduit.

In practice it may be impossible to distinguish between the two sulphide collection mechanisms described above. Classic examples of valuable sulphide deposits accumulated from flowing magma in conduits or lava channels are given in Table 2, with the Sudbury Camp for comparison.

If sulphides are deposited in the place at which sulphide saturation occurred, without being transported and collected from a large volume into a small concentration, then a low-grade disseminated

Table 2 Characteristics of major base metal sulphide deposits*

Deposit or camp	Structural setting	Host magma type	Grades		Production and reserves (Mt)
			Ni %	Cu %	
Sudbury, Canada	Meteorite impact crater	Quartz diorite	1.2	1.0	>1600
Thompson, Canada	Magma conduit	Komatiite	2.5	<0.5	90
Voisey's Bay, Canada	Magma conduit	Picrite	2.8	1.7	32
Kambalda, Australia	Lava channel	Komatiite	3.5	<0.5	50
Perseverance, Australia	Lava channel	Komatiite	0.6	<0.5	270
Duluth, USA	Base of large intrusion	Continental tholeiite	0.2	0.7	>4000
Noril'sk, Russia	Magma conduit	Continental tholeiite	2.7	2.1	>500
Pechenga, Russia	Magma conduit	Ferropicrite	1.0	<0.5	30
Jinchuan, China	Magma conduit	Continental tholeiite	1.1	0.7	>500

*Data sources: Production and reserves from www.minecost.com; and Eckstrand, O.R., Magmatic nickel copper platinum group elements (Eckstrand OR, Sinclair WD, and Thorpe RI (eds.) Geology of Canadian Mineral Deposit Types, Geological Survey of Canada, *Geology of Canada* 8: 583 605).

deposit like those of the Duluth Complex (USA) will result.

Precious metal magmatic sulphide deposits An initially sulphur-undersaturated magma may become sulphur-saturated because crystallization of silicate minerals causes an enrichment of sulphur in residual melt while cooling simultaneously lowers the solubility of sulphide. Highly evolved basaltic magmas may thus passively evolve into a state of sulphide saturation in large magma chambers, causing the separation of extremely small fractions of sulphide melt, as may have occurred in the Great Dyke of Zimbabwe and the Munni Munni intrusion of Australia. Recharge of a large magma chamber by a fresh pulse of magma is another popular explanation for the sudden onset of sulphide saturation in large magma chambers (Figure 7). However, the magma body does become saturated with sulphide, removal of sulphide melt at high R value (eqn [4]) from a silicate magma barely saturated in sulphide may efficiently scavenge precious metals from the entire magma chamber into a single thin stratabound layer. The sulphides thus generated will also be highly enriched in the base metals, but at total sulphide abundances on the order of 1 modal %, the deposits are only valuable for their precious metals.

An alternate possibility for the generation of stratiform sulphide horizons in a layered intrusion involves migration of deuteric fluids in the cooling cumulate pile (Figure 6). As the initially vapour-undersaturated intercumulus melt cools and approaches its solidus it will eventually exsolve an aqueous fluid. Orthomagmatic fluids are capable of dissolving considerable amounts of precious metals and sulphur as they rise through the cumulate pile. If such fluids remain sulphide-saturated and equilibrate with the small quantities of sulphide that exist within the intercumulus liquid, then they will carry

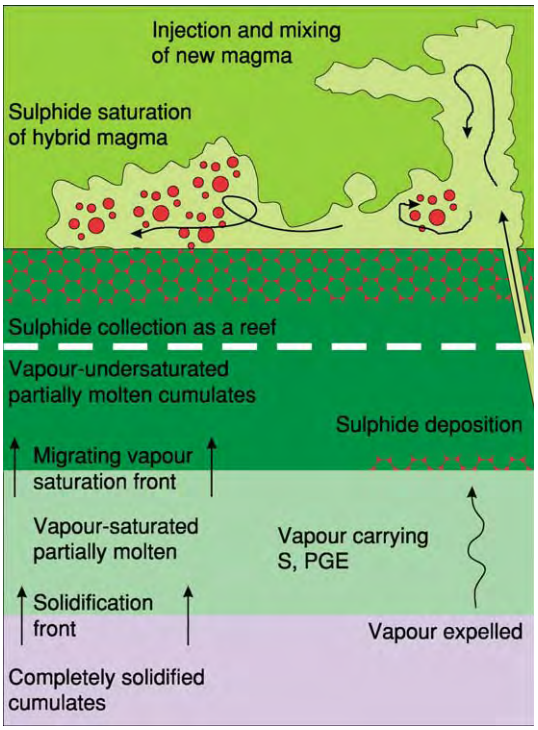


Figure 7 Schematic view of two popular models for the generation of reef type PGE sulphide mineralization in layered intrusions. The orthomagmatic model proposed by Campbell and co workers is illustrated above the white dashed line. The deuteric fluid model proposed by Boudreau and coworkers is illustrated below the line.

considerable amounts of dissolved sulphide and precious metals upward through the pile of cumulates. When such fluids reach the top of the vapour-saturated zone and encounter vapour-undersaturated intercumulus liquid, they will redissolve within it. Dissolution of the vapour phase will cause reprecipitation of the entrained sulphide and precious metals at the level of the transition to vapour-undersaturated conditions. As a vapour-saturation surface migrates

upward through a cumulus pile it will bring with it a sulphide- and precious metal-rich zone akin to a chromatographic front. Such fronts may be frozen in place by the eventual solidification of the cumulate pile, to form stratiform PGE-rich sulphide horizons.

Although there is a potential for the generation of a precious metal sulphide deposit wherever small volumes of sulphide liquid reach equilibrium with large volumes of formerly sulphur-undersaturated silicate magma, nearly all production from precious metal sulphide deposits comes from just two large layered intrusions; the Bushveld complex of South Africa and the Stillwater intrusion of the USA. About half of world production of Pd is generated as a by-product of base metal sulphide mining at Noril'sk in Russia.

See Also

Economic Geology. **Igneous Processes**. **Igneous Rocks:** Carbonatites; Komatiite. **Mantle Plumes and Hot Spots**. **Minerals:** Sulphides. **Mining Geology:** Mineral Reserves.

Further Reading

Boudreau AE, Mathez EA, and McCallum IS (1986) Halogen geochemistry of the Stillwater and Bushveld complexes: evidence for transport of platinum group elements by Cl rich fluids. *Journal of Petrology* 27: 967 986.

Campbell IH and Naldrett AJ (1979) The influence of silicate:sulphide ratios on the geochemistry of magmatic sulphides. *Economic Geology* 74: 1503 1505.

Campbell IH, Naldrett AJ, and Barnes SJ (1983) A model for the origin of the platinum group rich sulphide horizons in the Bushveld and Stillwater complexes. *Journal of Petrology* 24: 133 165.

Crowson P (2001) *Minerals Handbook 2000 2001*. Mining Journal Books.

Drew LJ, Qingrun M, and Wiejun S (1990) The Bayan Obo iron rare earth niobium deposits, Inner Mongolia, China. *Lithos* 26(1/2): 43 65.

Eckstrand OR, Sinclair WD, and Thorpe RI (eds.) (1995) *Geology of Canadian Ore Deposit Types*. Geological Survey of Canada.

Keays RR, Lesher CM, Lightfoot PC, and Farrow CEG (eds.) (1999) *Dynamic Processes in Magmatic Ore Deposits and their Application in Mineral Exploration*. Short Course Volume 13. Geological Association of Canada.

Kolker A (1982) Mineralogy and geochemistry of Fe Ti oxide and apatite (nelsonite) deposits and evaluation of the liquid immiscibility hypothesis. *Economic Geology* 77: 1146 1158.

London D (1992) The application of experimental petrology to the genesis and crystallization of granitic pegmatites. *The Canadian Mineralogist* 30: 499 540.

Naldrett AJ (1989) *Magmatic Sulphide Deposits*. New York, Oxford: Clarendon Press Oxford University Press.

Naslund HR (1983) The effect of oxygen fugacity on liquid immiscibility in iron bearing silicate melts. *American Journal of Science* 283: 1034 1059.

Reynolds IM (1985) The nature and origin of titaniferous magnetite rich layers in the Upper Zone of the Bushveld Complex: a review and synthesis. *Economic Geology* 80: 1089 1108.

Roberts RG and Sheahan P (eds.) (1988) *Ore Deposit Models*. Geoscience Reprint Series 3. Geological Association of Canada.

Whitney JA and Naldrett AJ (eds.) (1989) *Ore Deposition Associated with Magmas*. Society of Economic Geologists. *Reviews in Economic Geology* Vol. 4.

MOHO DISCONTINUITY

P Giese, Freie Universität Berlin, Berlin, Germany

Introduction

A century ago, instrumental seismology began to reveal the internal structure of Earth. Analysing seismic data in 1909, following an earthquake 39 km south-east of Zagreb, the Austro-Hungarian meteorologist Adrija Mohorovičić discovered the boundary between the crust and mantle of south-eastern Europe. During the following decades, it turned out that this discontinuity, now known as the 'Mohorovičić discontinuity' (or simply 'Moho-discontinuity' or 'M-discontinuity' or Moho), is present worldwide in continental as well as in oceanic regions. The following discussion focuses on the morphology, structure, and age of the European Moho-discontinuity and the bearing that these features have on the structure and evolution of the crust. The data are largely based on the results of national and international projects carried out during the past 50 years. The European Geotraverse (EGT) project, from 1983 to 1990, involved studies of the European lithosphere along a north- and south-trending corridor from the North Cape of Scandinavia down to the northern coast of Africa (Tunisia), an ambitious

enterprise, in that it attempted to resolve a geodynamic evolution from ~3.5 billion years ago to the present. The EGT project was followed by the EUROPROBE project, which aimed at analysing the structure, properties, and evolution of 'deep Europe' and exemplary cases of crustal accretion, thickening, and extension. The summary of data from these projects covers the geophysical tools used, the petrological significance of the data, and the main structural features of the European Moho-discontinuity.

Geophysical Tools

Geophysical methods are indispensable for the recognition and interpretation of structures in Earth's crust and mantle. Among the numerous geophysical methods suitable to explore structure and composition of the lithosphere, seismic techniques provide the best tools for detailed crustal and upper mantle investigations (*see* **Seismic Surveys**). In explosion seismology, the use of deeply, nearly vertical, reflected waves (P waves) in the subcritical range provides detailed information about the internal structure of the crust and the uppermost mantle (Figure 1A). Another method is based on recording P waves termed 'PmP' wide-angle arrivals; these are recorded as overcritical reflected waves from the Moho-discontinuity and/or as diving waves generated by a strong velocity gradient zone in the model of the crust/mantle transition (Figure 1B). Tomographic studies using travel times of waves generated by earthquakes or by big explosions allow determination of the velocity field of the lithosphere (Figure 1C). During the past decade, a new 'receiver function method' has been developed (Figure 1D). It is based on earthquake records of broadband stations and makes use of P-to-S converted phases that are created at the Moho-discontinuity or other prominent boundaries in the lithosphere. For the study of the lithosphere/asthenosphere system, analysis of surface waves is another powerful seismic tool. The data discussed here are mainly based on recordings of P waves (compressional waves).

Nature and Stability of the Moho-Discontinuity

From a seismological point of view, the Moho-discontinuity is defined as the depth at which the velocity v_P of compressional waves increases from normal crustal velocity values of $<7.2\,km\,s^{-1}$ to typical upper mantle values of $>7.6\,km\,s^{-1}$. Modern studies see the Moho-discontinuity as a sandwich-like mix of crust and mantle materials (i.e., low- and high-velocity materials). In addition, the Moho-discontinuity is not only a compositional boundary,

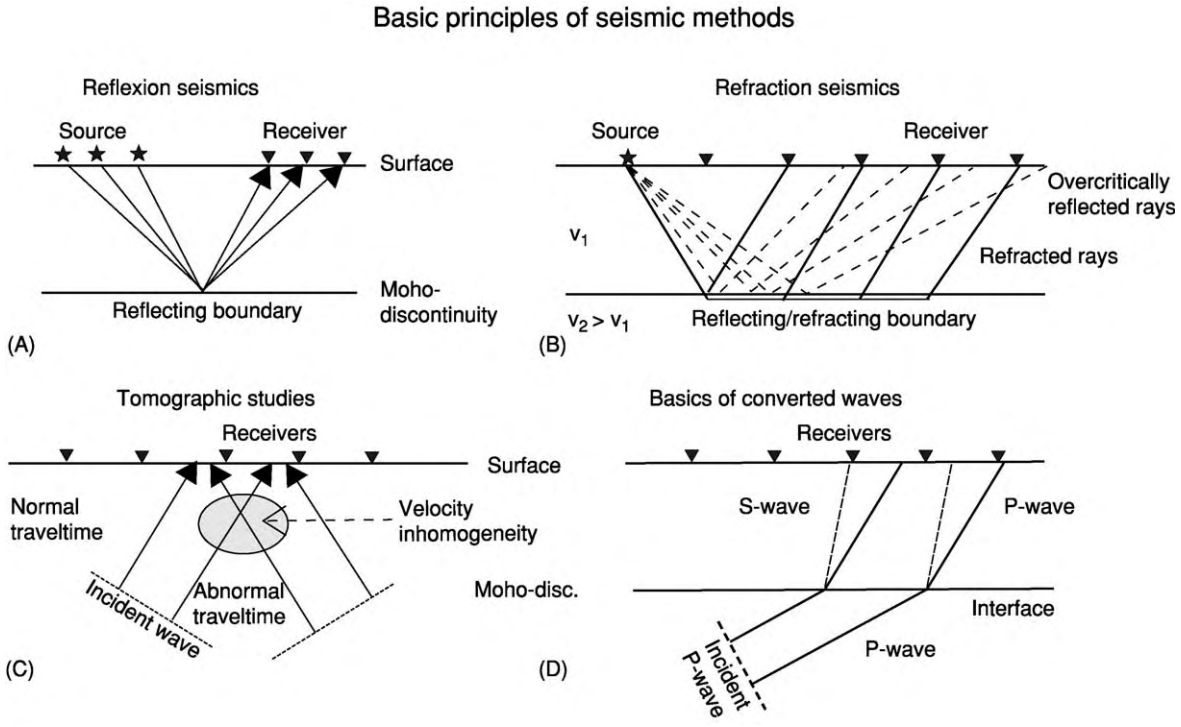


Figure 1 Basic principles of deep seismic sounding methods: (A) reflection seismology, (B) refraction seismology, (C) tomographic seismology, and (D) receiver function analysis.

but is additionally characterized by an abrupt change in the length scale of heterogeneities between the lower crust and upper mantle, and therefore also images various geodynamic processes (Figure 2). In many regions, the Moho depth derived from reflected waves generated at the base of the lower crust coincides with the depth of the Moho-discontinuity as determined from seismic refracted and wide-angle reflected waves. The difference between both results does not exceed a few kilometres. On the other hand, petrologists define the Moho-discontinuity between the crust and the uppermost mantle as a petrological boundary. The lower crust consists of high-temperature/high-pressure metamorphic rocks (granulite facies) with an average mafic composition approaching that of a primitive basalt (*see* **Ultra High Pressure Metamorphism**). As evidenced by mantle fragments brought to the surface by volcanism (xenoliths), the lower crust shows an interlayering of felsic and intermediate lithologies, causing a high degree of seismic reflectivity in the lower crust, as observed in many regions. The layer beneath the Moho-discontinuity, the uppermost mantle, is built up of rocks rich in olivine (*see* **Minerals:** Olivines).

The ambiguity between the geophysical and petrological definitions of the Moho-discontinuity causes problems of interpretation. If lower crustal rocks such as pyroxene granulite or gabbro are subducted to depths below 50–60 km, they are metamorphosed into the high-pressure rock type, eclogite, which shows the same density (3300 kg m ^{3}) and the same P-wave velocity (>7.6 km s ^{1}) as normal mantle rocks (peridotites) show. From this geophysical perspective, eclogites would be interpreted as upper mantle rocks underlying the Moho-discontinuity. Geological, petrological, and geophysical studies reveal that the Moho-discontinuity is not a permanently stable boundary and that it can be destabilized during all stages of the orogenic cycle by tectonic, magmatic, and metamorphic processes:

- Collision of continental blocks causes crustal thickening, associated with the formation of a surface topography and the generation of crustal and lithospheric roots. Intrusion of magmas may also cause crustal thickening.
- When the crust of a thickened mountain belt is reduced by erosion, the crustal root will be buoyantly uplifted until the crust has attained the same thickness as in the surrounding areas.
- When deeper levels of the crust are heated by the underlying mantle, crustal rocks may become mechanically soft and the crustal root may spread laterally over the Moho-discontinuity, much like oil on water.
- A pseudo-thinning of the crust may be caused by metamorphic processes in the lower crust, such as the transformation of gabbroic rocks into eclogite. Such 'pseudo-mantle' rocks might represent portions of subducted and metamorphosed lower crust, so that the true (petrological) crustal thickness would be higher than expected from the study of the recent seismic velocity distribution. Furthermore, such high-density layers may delaminate and sink downward into the deeper lithosphere and asthenosphere due to their negative buoyancy.

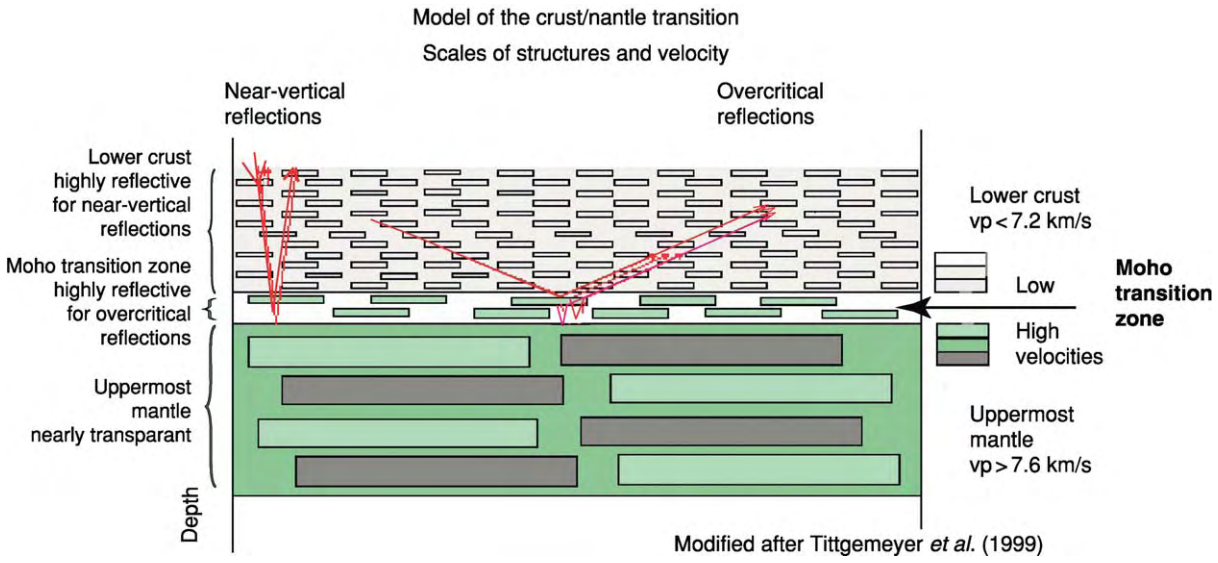


Figure 2 Material properties around the continental Moho discontinuity and their response to seismic signals. The thickness of the transition zone between the lower crust and the upper mantle may vary between ~100 m and a few kilometres. Modified from Tittgemeyer *et al.* (1999).

- The thickened orogenic crust may be thinned by syn- to post-orogenic extension driven by mantle convection, which creates new continental rifts and oceanic basins.

This brief overview demonstrates that the Moho-discontinuity is a sensitive indicator for the detection of past and recent geodynamic processes.

Main Tectonic Units of the European Continent

Present-day Europe forms the western part of the large continental Eurasian Plate. It is bordered in the east by the mountain belt of the Urals and in the north and west by the Atlantic Ocean, which opened in the Triassic. The southern border of Europe is formed by the Mediterranean Sea, the Black Sea, and the Caspian Sea (Figure 3). The growth of the European continent started in the northern region of the Baltic Shield, and the European crust becomes younger and younger from north to south and from east to west. Five main tectonic domains can be distinguished: (1) The East European (Russian) Platform and the Baltic Shield, the north-western part, were formed during Precambrian times, 3.6 Ga to 660 Ma (*see* **Europe:** East European Craton). The western margin of the East European Platform is formed by the complexly structured Trans-European Suture Zone (TESZ), which separates the Variscan mobile belts of Central Europe from the stable platform of Eastern and Northern Europe. This suture zone extends over 2000 km from the North Sea down to the north-western Black Sea. (2) The north-western margin of the Baltic Shield is marked by the Caledonide Orogen, which was formed during the Late Silurian to Early Devonian (*see* **Europe:** Scandinavian Caledonides (with Greenland)). The Caledonides continue into northern parts of England and Ireland (*see* **Europe:** Caledonides of Britain and Ireland). (3) On the eastern side of the East European Platform, the Uralian Orogen (380–250 Ma) marks the tectonic boundary

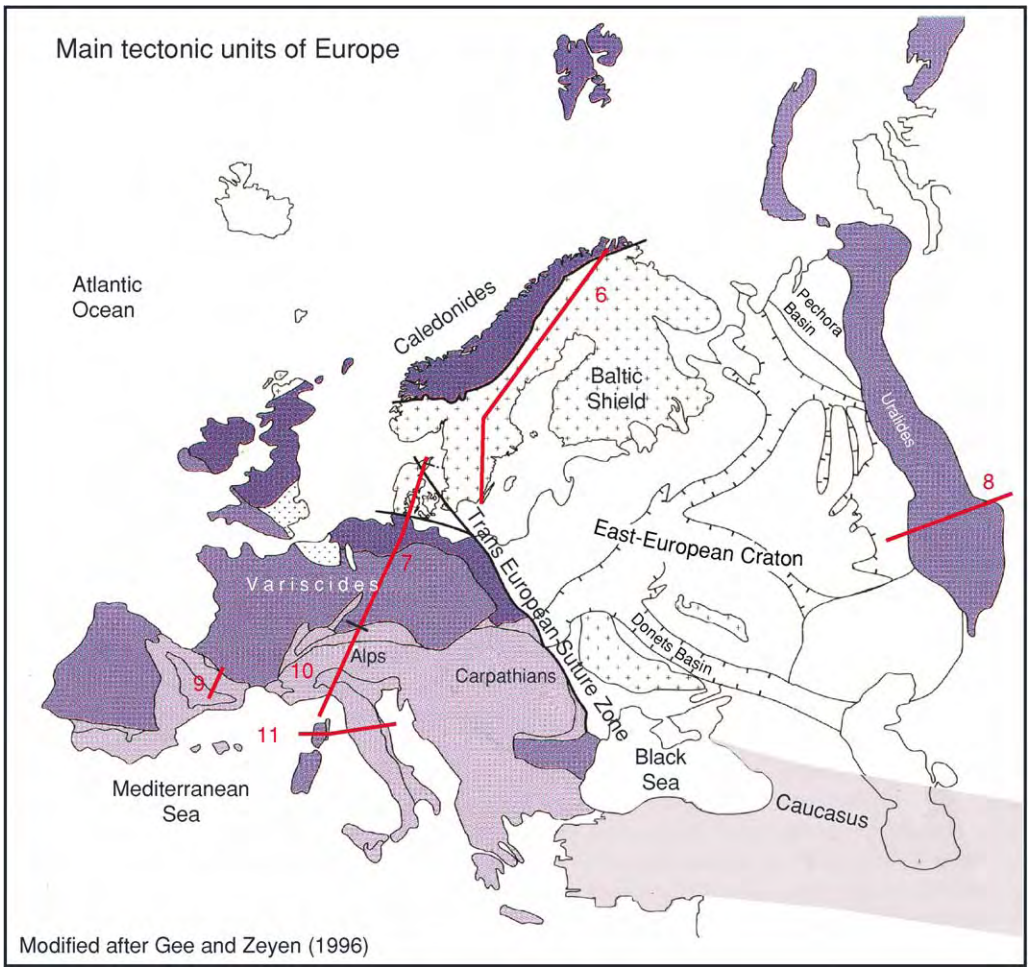


Figure 3 Main tectonic units of Europe. The position of the crustal cross sections described in this paper is marked by red lines. Modified from Gee and Zeyen (1996).

between the East European and the Siberian Plates (*see* **Europe:** The Urals). (4) Central and Western Europe were formed during the Cadomian (~620–540 Ma) Orogeny and severely were overprinted by the Variscan (or Hercynian) Orogeny (400–295 Ma) (*see* **Europe:** Variscan Orogeny). (5) In Southern Europe, the Cenozoic Alpine–Mediterranean mountain belts resulted from the convergence between Africa and Europe during Mesozoic to Cenozoic times (220 Ma). The Mediterranean Sea contains relicts of the oceanic crust of the Tethys and young back-arc basins that are partly built up by young oceanic crust (*see* **Europe:** Mediterranean Tectonics). Besides these compressional structures, there are also some prominent extensional regions in Europe. The formation of the North German and Polish basins and the evolution of the Central European Rift System were generated by post-Variscan extensional processes (*see* **Europe:** Permian to Recent Evolution).

Main Features of the Moho-Discontinuity in Europe

During the past five decades, a large amount of seismic data (active and passive seismology) has provided detailed information on regional and local crustal structures and resulting geophysical anomalies. Figure 4 is a contour map of the depth of the Moho-discontinuity of the Eurasian plate, extending from the Atlantic Ocean in the west to the Siberian Platform in the east. A more detailed Moho contour map of Central, Western, and Southern Europe is presented in Figure 5, which is based on seismic data measured in Europe from 1958–1988. This map comprises large parts of Pre-Caledonian, Caledonian, Variscan, and Alpine domains in Europe. There are very pronounced variations of crustal thickness within and between the different tectonic units. The most prominent structural features are as follows:

- The crust of the Baltic Shield and the East European Platform shows a rather large thickness, with an average of 40–50 km.
- The Trans-European Suture Zone is one of the most important geological boundaries in Europe north of the Alpine Orogen. This suture zone forms a graben-like structure with a thickened crust and a depth of the Moho-discontinuity between 45 and 55 km.
- In contrast to the Baltic Shield and the East European Platform, the depth of the Caledonian and Variscan Moho-discontinuity in Central and

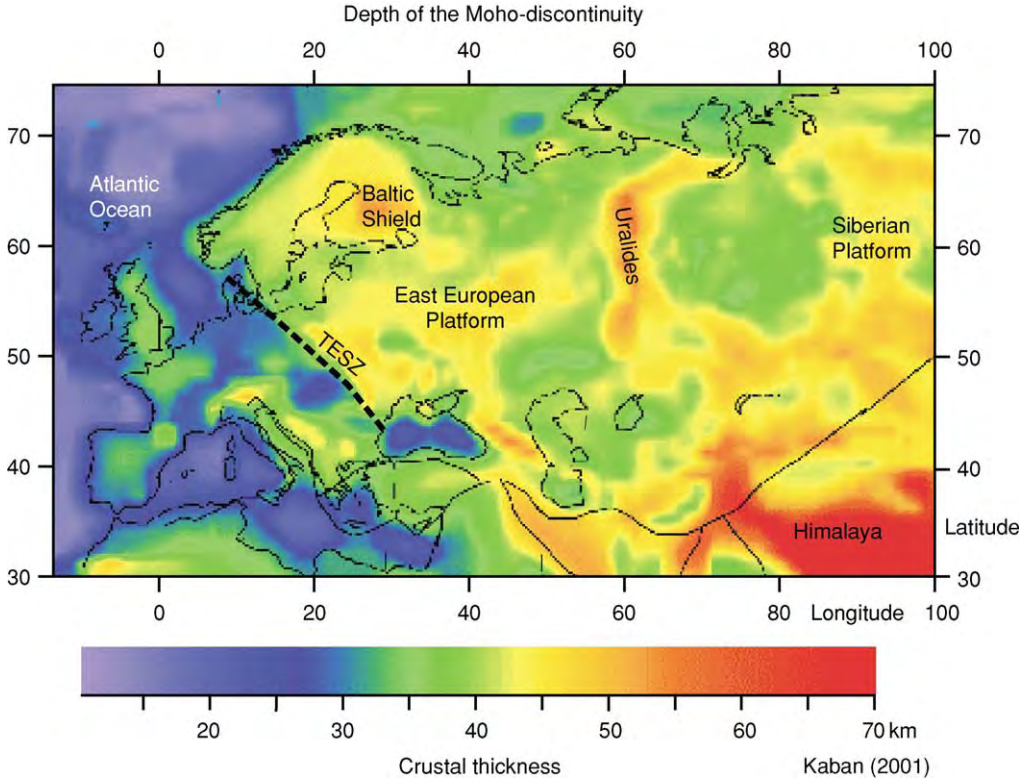


Figure 4 Crustal thickness of the European continent and the West Siberian platform, derived from numerous seismic profiles (mainly from deep seismic sounding studies) and gravity data. Modified and reproduced from Kaban (2001).

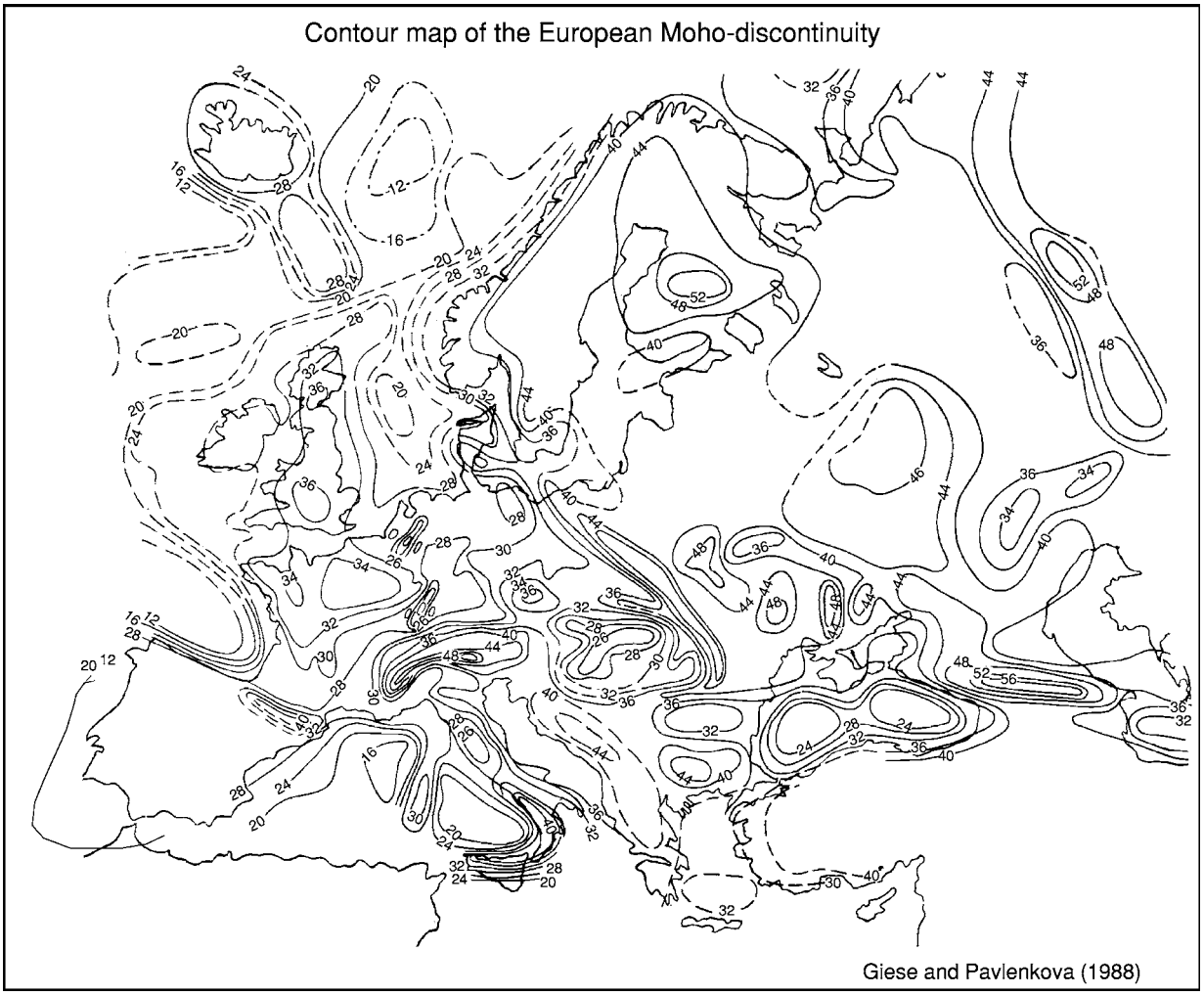


Figure 5 Morphology of the Moho discontinuity in Central, Western, and Southern Europe. The contours (spacing, 2 to 4 km resp) reveal the main geological structures of young orogens such as the Alps, the Apennines, and the Pyrenees. Giese and Pavlenkova (1988).

Western Europe measures only 30–35 km. No crustal roots are recognizable.

- The picture is distinctly different in the Variscan Uralides, where the crustal root or parts of the root have been preserved since the Late Palaeozoic. In the root zone, the Moho-discontinuity has been detected at a depth of 50–60 km.
- The southern (Mediterranean) region of Europe presents a very heterogeneous set of crustal structures. This region is built up of a system of several Cenozoic orogenic belts with thick mountain roots up to 60 km deep.

Other regions were formed by extensional processes and are characterized by a shallow Moho-discontinuity. Extension in continental crust is documented in the Permo-Mesozoic basins of Northern Germany and Northern Poland, in the Cenozoic European rift system. The Cenozoic Pannonian Basin with a crustal thickness of only 24 km was formed by back-arc spreading. Oceanization of continental crust is also observed in back-arc regions of the western Mediterranean, where the Moho-discontinuity is found at a depth of only 10 km. Atlantic Ocean crust borders the western and north-western margin of the European continent. Oceanic crust is also present in the central Black Sea and in the Bay of Biscay, which was formed by counterclockwise rotation of the Iberian peninsula with respect to mainland Europe.

Main Structural Features of Pre-Variscan and Variscan Europe

Northern and Eastern Europe

Precambrian Europe is built of the Baltic Shield and the East European Platform (*see* **Europe:** East European Craton; Scandinavian Caledonides (with

Greenland)). The Baltic Shield is a collage of cratonic blocks and island arcs. The orogenic belts resulting from multiple collision have been levelled out by erosion at the surface and partial re-equilibration of the Moho-discontinuity. Seismic studies reveal variations of crustal thickness ranging from 40 km in Sweden to more than 55 km in southern Finland. This anomalously large crustal thickness is mainly caused by a relatively thick lower crust (15–20 km). Figure 6 shows a north to south crustal cross-section through the Baltic Shield between the North Cape and the southern tip of Sweden at the Baltic Sea. Along this section, the average crustal thickness is about 40 km. Lateral inhomogeneities with several distinct offsets of the Moho-discontinuity may be related to fault zones in the upper crust. These probably reflect amalgamation of microplates that were derived from the fragmentation of Late Archaean megacontinents and that were welded together along Proterozoic collisional belts. The large thickness of the Baltic Shield and the Eastern European Craton requires an average crustal density of 2900–3000 kg m ^{3} By comparison, the average crustal density in Central Europe amounts only to about 2800 kg m ^{3}. Thinned crust is present only in the Dniepr–Donets Basin, which is a product of intracratonic rifting in the Devonian and Early Carboniferous times. Parts of the Caledonian belt are exposed along the coast of Norway and in Scotland and central England. Another branch of the Caledonides is suspected beneath the Northern German and Polish basins. The recent average crustal thickness is about 28–34 km and no orogenic roots are visible (Figure 5). As in the Variscides, the Caledonian crust is distinctly thinner than that of the Baltic Shield, which again suggests crustal thinning by delamination of lower crust.

The Trans-European Suture Zone

The Trans-European Suture Zone is one of the most prominent geological boundaries in Europe. It separates the East European Platform from the mobile Variscan Domain in Central Europe. This zone, with its graben-like structure and a crustal thickness of 40–50 km, extends from the western coast of the Black Sea up to the western Baltic Sea. Here this system splits into two branches, the northern Sorgenfrei–Tornquist Zone (STZ) and the southern Trans-European Fault Zone (Figure 3). This zone is well recognizable in the contour map of the Moho-discontinuity in Figure 5.

The Variscides of Central and Western Europe

The main part of Central and Western Europe has a Late Proterozoic (Cadomian) fundament, but the present-day structure is controlled by a strong Variscan overprint (*see* **Europe:** Variscan Orogeny). This region extends between the North Sea and Baltic Sea to the north and the Alpine mountain belts to the south and, in the west to east direction, between the Atlantic Ocean and the Trans-European Suture Zone (Figure 3). The upper section of Figure 7 shows the crustal structure of the transition from the Baltic Shield to the Variscan Domain in Central Europe (*see* **Europe:** Variscan Orogeny). The lower section of shows the crustal structure of the area between the

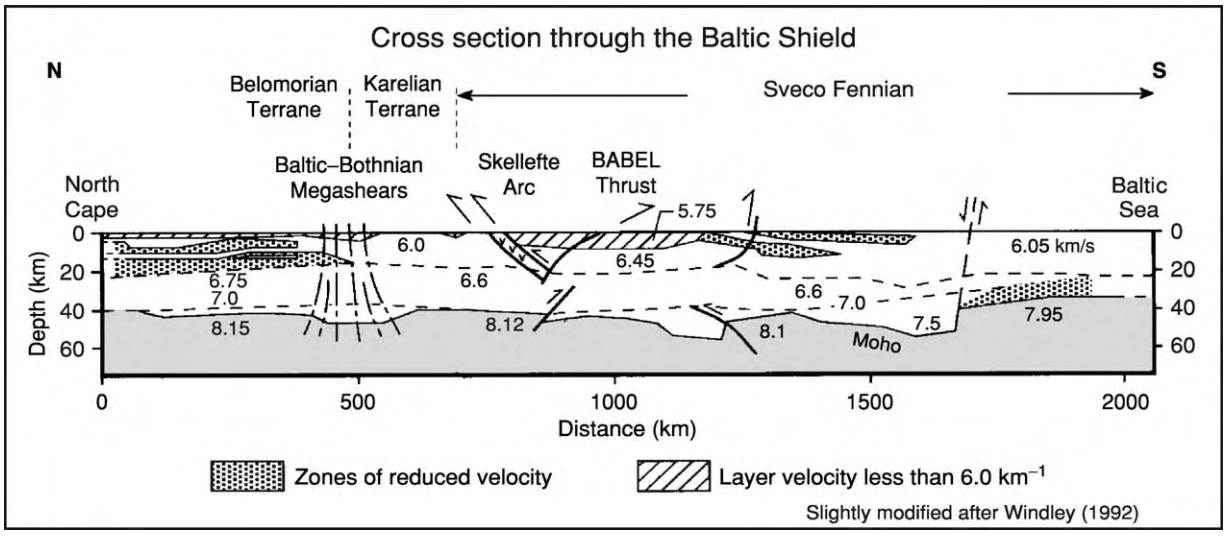


Figure 6 N S cross section through the Baltic Shield with the main crustal structures, displayed as an isoline velocity model (vertical exaggeration: 5:1). Modified after Windley *et al.* (1991) in Blundell *et al.* (1992).

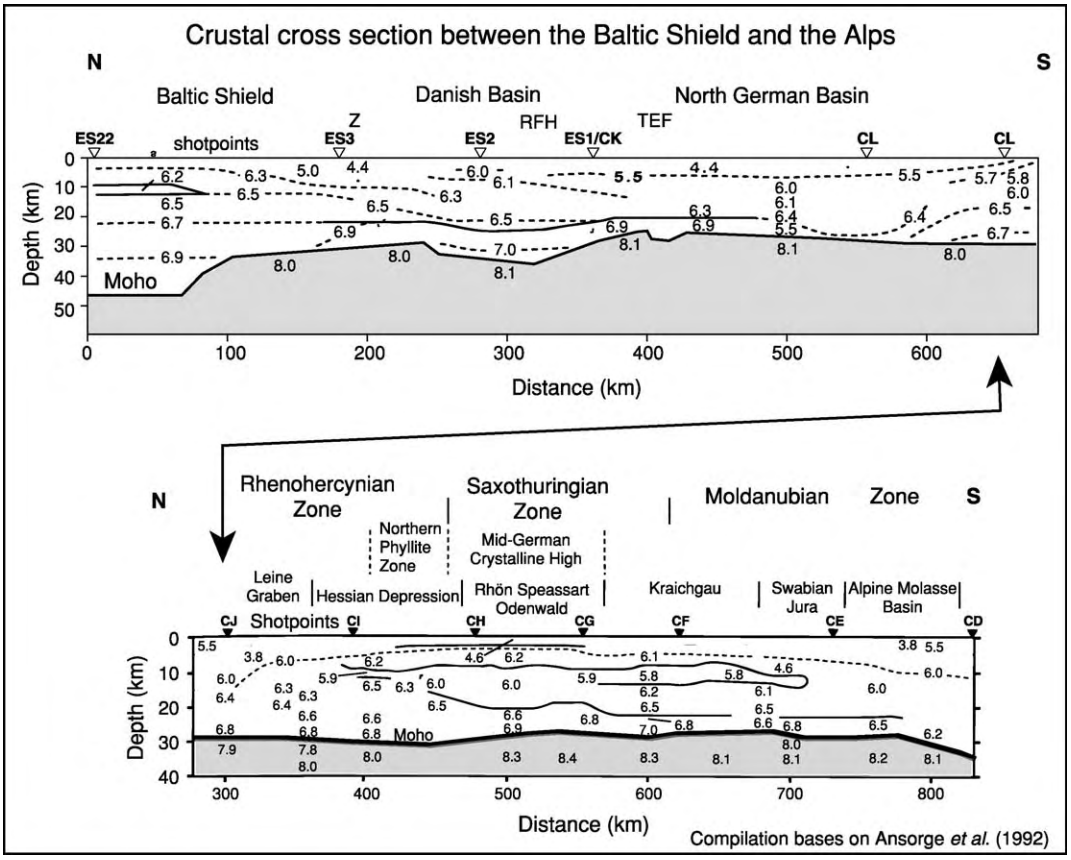


Figure 7 (Upper part) Structure of the transitional zone between the Baltic Shield and the Variscan crust adjacent to the south, which has undergone Permo Mesozoic extension. STZ, Sorgenfrei Tornquist zone; RFH, Ringkbing Fyn High; TEF, Trans European Fault. (Lower part) North south section from the northern margin of the Rhenohercynian zone to the northern margin of the Alps. Although the internal crustal structure is very heterogeneous and the section cuts prominent suture zones, the crustal thickness is rather constant (28 30 km). Modified from Ansorge *et al.* (1992).

North German Basin and the northern margin of the Alps. Along this profile, the recent Moho-discontinuity lies at a depth between 28 and 30 km The depth increases to 34 km at the northern margin of the Alps. The present-day morphology of the Moho-discontinuity does not show any significant relicts of Variscan roots. Velocities in the uppermost mantle change significantly from a minimum value of 7.8 km s ^{1} in the north to as much as 8.4 km s ^{1} under the southern margin of the Mid-German Crystalline High; from there it decreases again to a more normal value of 8.1 km s ^{1}. The Variscan orogenic systems were formed during Late Palaeozoic times by the collision of numerous continental fragments or terranes of different size with the Baltic Shield and the East European Platform. Because seismic reflection profiles provide evidence of important crustal stacking, compressional deformation must have formed thick and deep crustal roots, comparable with the mountain roots of the recent Mediterranean orogens. These roots have not survived. Destruction of the Variscan roots is usually referred to as 'collapse' of the orogen. This comprehensive term summarizes a variety of processes, including erosion and isostatic uplift; gravitational spreading of hot, low-viscosity materials; extension by active rifting; and delamination of dense lower crust and mantle lithosphere.

Spectacular Crustal Root of the Uralides

The Ural Mountains are the product of collision of the East European Craton with the Siberian Shield (*see* **Europe:** The Urals). Figure 8 shows the results of a seismic profile across the southern Uralides. The Moho-discontinuity, well defined by seismic reflection data, dips from the west and east under the central part of the belt, down to a depth of 50 km. Wide-angle reflections in the central part of this belt reveal a still existing crustal root, with the Moho-discontinuity situated at a depth of 50–60 km. This unusual preservation of an orogenic root may be explained by the small density contrast (only 100 kg m ^{3}) between the root and the surrounding

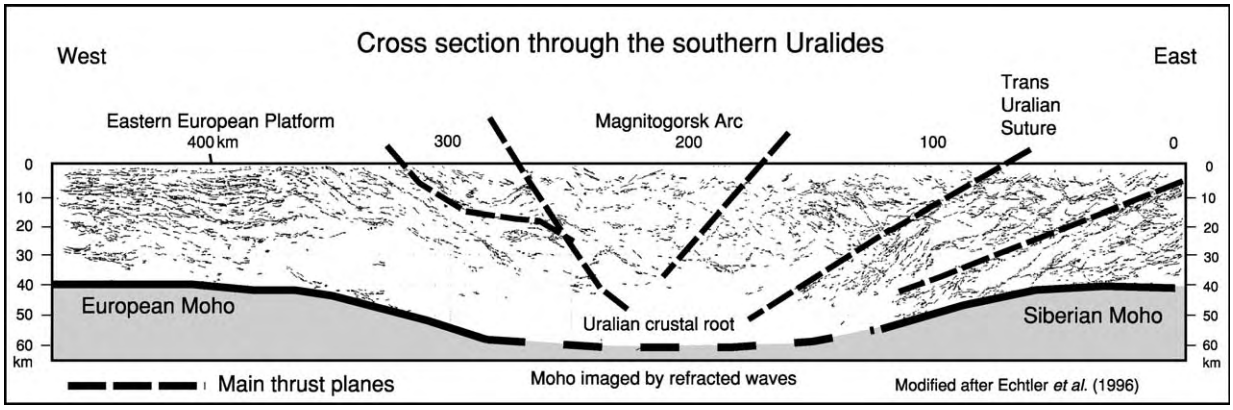


Figure 8 (Upper part) West to east directed section across the southern Uralides, based on seismic reflection and refraction measurements. Note the bilateral symmetry of reflections dipping from either side under the thick root. Modified from Echtler *et al.* (1996).

mantle, which was not sufficient to drive buoyant lateral spreading. In addition, there is no indication of post-orogenic active rifting in the Uralides.

The Post-Variscan Basins of Central and Western Europe

The post-Variscan regions of Central and Western Europe are characterized by an average crustal thickness of only 28–32 km (Figures 5 and 7). In these regions, the post-Variscan evolution is dominated by extensional tectonics, with formation of a new Moho-discontinuity at depths around 30 km. The main extensional elements are the Permo-Carboniferous basins in northern Germany and Poland and the Cenozoic European Rift System (*see* **Europe:** Permian to Recent Evolution). In the North German and Polish basins, the thickness of the post-Variscan sediments may attain 10 km, and the base of the crust has been detected at depths between 23 and 30 km. This implies crustal thinning, decreasing from a thickness of ~30 km to 13–23 km. Basin development started in Early Permian times and was probably caused by a combination of strike–slip and extension along the south-western margin of the East European Platform, and thermal subsidence subsequent to an Early Permian episode of mantle-derived magmatism. A second magmatic event occurred during the Cretaceous.

The Cenozoic European Rift System traverses Western and Central Europe up to the southern border of the Baltic Shield. Over a distance of about 1800 km, this north–south-oriented rift system is composed of a sequence of separated segments (Figure 5). This rift system starts in southern France in the Rhône Graben. Northwards, the Limagne Graben, west of the Rhône Valley, is a very significant segment of the rift that traverses the Massif Central and is flanked by volcanic chains. Further north, the Bresse Graben system continues from the Rhône Valley. Both the Limagne Graben and the Bresse Graben systems terminate again in the Burgund transform fault system, which links the grabens of south-eastern France with the Rhine Graben near Basel. The Rhine Graben (Figure 5) between Basel and Frankfurt/Main forms the central segment of the European Cenozoic Rift System. At the northern end of the Rhine Graben, at a triple junction near Frankfurt, the rift system splits into two branches. To the north, it follows the Hessen depression and the Leine Graben near Göttingen and disappears beneath the North German Basin. The Oslo Graben forms the northernmost segment of this rift system. The western branch starts at the northern end of the Rhine Graben near Frankfurt and continues in a north-west direction through the Rhenish Massif into the Lower Rhine Basin and disappears beneath the North Sea.

As shown in Figure 5, the total crustal thickness along the Central European Rift System is distinctly reduced, from 30 to 32 km outside the rift to about 28 km on top of the rift. Even the splitting of the rift near Frankfurt is recognizable in the contour map of the Moho-discontinuity. (Figure 5). The creation of the layering in the lower crust has been widely discussed in the literature. Deep seismic reflection profiles suggest that the layering of the lower crust resulted in an extensional regime, and that magmatic intrusions caused partial crustal melting and the transformation of deep crustal mafic rocks into eclogite. The unusual large amplitudes of the Moho reflections and the presence of anomalously high seismic velocities in the lowermost crust beneath Tertiary rifts (e.g., the Rhine and Rhône Grabens) imply that the Moho-discontinuity and part of the layering

of the lower crust are comparatively young features, related to interactions between the lower crust and the mantle.

Crustal Structure of the Mediterranean Region

Towards the south, Europe is flanked by Mesozoic–Cenozoic orogenic systems created by the convergence and collision of the European continent with the African/Arabian Plate. During their evolution, since the early Mesozoic, the Alpine–Mediterranean orogenic belts underwent a more or less complete cycle of extension, subsidence, subduction, collision, topographic buildup, extensional collapse, and back-arc spreading associated with the creation of new oceanic basins. The morphology of the Moho-discontinuity shows that this discontinuity was strongly involved in these compressional and extensional deformation processes.

All Mediterranean mountain belts show the typical subdivision into three main tectonic zones, each with its own specific crustal structures: the foreland (fore-arc), the central zone, and the hinterland (back-arc). In all the studied cases (Alps, Apennines, Betic Cordillera, Carpathians, Caucasus, Hellenides, and Pyrenees), this basic picture is clearly expressed in the asymmetric geometry of the Moho-discontinuity. The Moho-discontinuity of the foreland dips under the central zone and in some cases even beyond, under the hinterland, down to a depth of 60–70 km. Still deeper crustal rocks cannot be excluded, but have been metamorphosed into eclogites, which owe their thickness to tectonic stacking of sedimentary and crystalline units, including mantle slices, as well as to magmatic additions to the crust. The hinterland, with its thin crust, overlies the underthrusted continental lithosphere of the foreland and oceanic crustal fragments of the central zone. In this case, there is no continuity between the Moho boundary of the foreland and the Moho boundary of the hinterland. The vertical offset between the two Moho-discontinuities may attain 10–30 km. In the closely spaced array of Mediterranean microplates (*see* **Europe:** Mediterranean Tectonics), the tectonic role of a crustal block may change across a short distance: the Adriatic Microplate (*see* **Europe:** The Alps) is the upper plate in the Alpine system, whereas it represents the lower plate in the Apennines. Three typical crustal cross-sections (Pyrenees, Western Alps, and Northern Apennines) highlight the main characteristics of the Moho topography in the Mediterranean belts (Figures 3, 9, 10, and 11).

Pyrenees

The Pyrenees are situated between the Iberian continental block and the European Plate. Extensional deformation was active from the Triassic through the Cretaceous. This extensional phase was followed by convergence between Iberia and Europe that occurred from the Late Cretaceous through the Middle Miocene, creating the Pyrenees. The present-day crustal structure across the Pyrenees displays a well-preserved crustal root created by the collisional process (Figure 9). The seismic reflection profile reveals how the Iberian crust was subducted below the European crust. A Moho offset of about 30 km has been detected beneath the North Pyrenean fault zone, and the reflecting lower crust of the Iberian crust can be traced down to a depth of 60 km.

Central Alps and Northern Apennines

The Alps are considered as a classical example of an orogen formed by the convergence between the continents of Europe and Africa, including the intervening Adriatic Microplate (*see* **Stratigraphical**

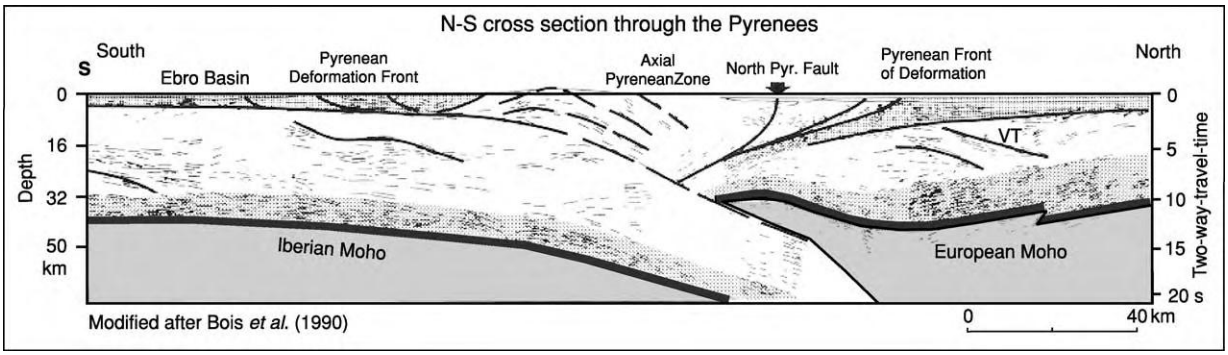


Figure 9 Line drawing and tectonic interpretation of a deep seismic profile across the Pyrenees. Beneath the central zone the offset of the Moho discontinuity is clearly recognizable. Modified from Bois and ECORS (1994).

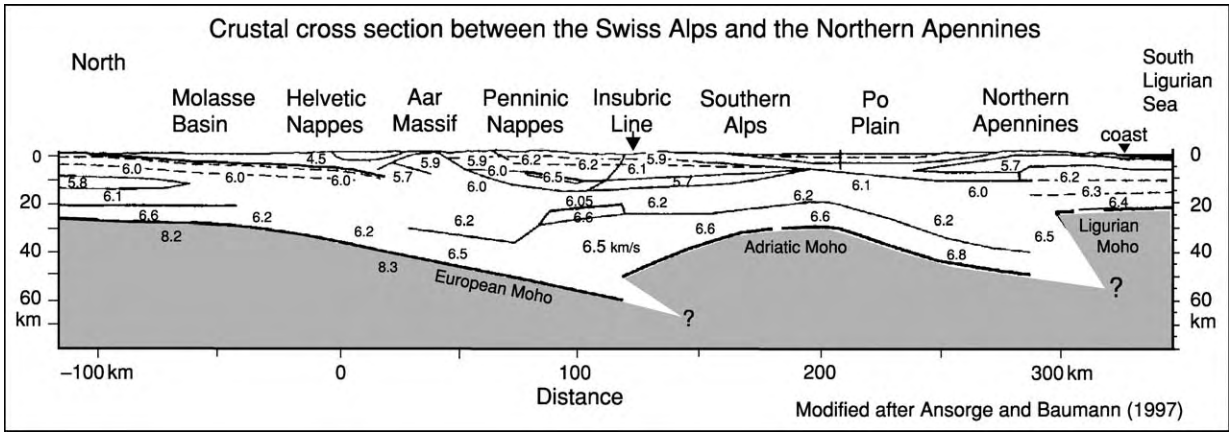


Figure 10 North to south directed crustal section between the Central (Swiss) Alps and the Ligurian Sea. This section shows two clearly expressed offsets of the Moho discontinuity. Thus three separated Moho discontinuities can be defined from north to south (the European, the Adriatic, and the Ligurian sea Moho). Modified with permission from Kissling (1997) in Pfiffner *et al.* (1997).

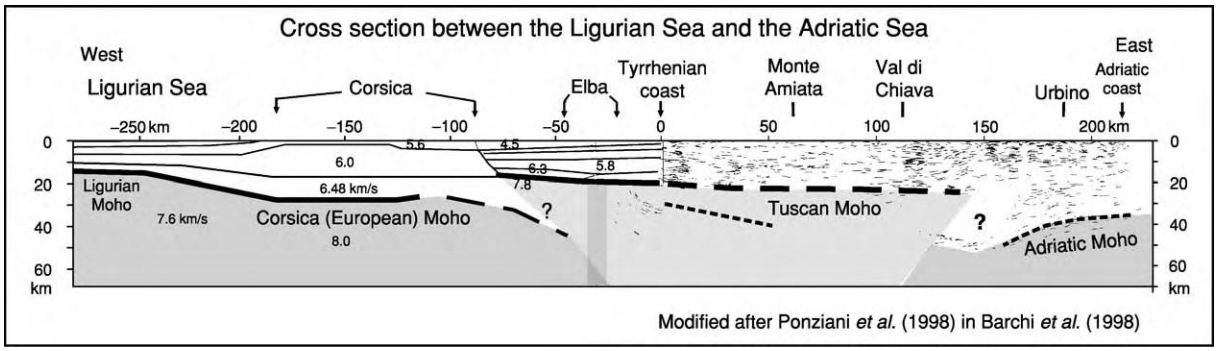


Figure 11 West to east directed crustal section across the Northern Apennines between the Ligurian Sea and the Adriatic Sea. This section is based on seismic refraction data in the western part and on seismic reflection observations in the eastern part. Note the different Moho discontinuities from west to east: the thin oceanic crust of the Ligurian Sea, the European crust of Corsica (and Sardinia) with a normal Variscan crustal thickness of 30 km, the thinned crust in the Tuscany region, and the subducting crust of the Adriatic foreland. Reproduced from Ponziani *et al.* (1998).

Principles). After the Variscan Orogeny, a vast embayment in the eastern margin of Pangaea propagated westwards during the Mesozoic to produce the Tethys Ocean. In the Alps, convergence started in the mid-Cretaceous with the consumption of ocean crust and was followed by overthrusting of the marginal units of the Adriatic Microplate onto the European continental margin. In the Southern Alps (the Alpine hinterland), thrusting of the sedimentary cover was directed southwards towards the Po Plain. South of the Po Plain, along the southern and western margins of the Adriatic Microplate, the mountain belt of the Northern Apennines was formed in the Cenozoic. These repeated collisions have strongly affected the geometry and depth of the Moho-discontinuity. Figure 10 shows a north–south crustal cross-section between the Swiss Alps in the north and the northern Ligurian Sea in the south. This section traverses the Central and Southern Alps, the western Po Plain, and the Northern Apennines. The complex shape and geometry of the Moho-discontinuity are clearly imaged in the records of the seismic reflected and refracted waves. The seismic section shows the well-developed structural asymmetry of both mountain belts. From the Alpine foreland (the lower plate), the Moho-discontinuity dips southwards, from 30 km at the northern Alpine front to about 60 km beneath the Central Alps. Under the Southern Alps, the Adriatic Moho-discontinuity situated at a depth of about 40 km is clearly separated from the European Moho-discontinuity overlap of both Moho-discontinuities. The southern part of the section (Figure 10) traverses the Southern Alps and the Po Plain, the Northern Apennines, and the northern part of the Ligurian Sea. The continental crust of the Southern Alps and of the Po Plain must be related to the margin of the Adriatic Microplate. Beneath the Po Plain, the Adriatic Moho-discontinuity shows a minimum

depth of 35 km. Towards the north and south, the Adriatic Moho-discontinuity dips down to 40–50 km beneath the Southern Alps (as the upper plate) and also beneath the Northern Apennines (as the lower plate). Beneath the Northern Apennines, there is a clear offset of 20–30 km between the deep northern and the shallow southern Moho-discontinuity. The thinned crust of the Northern Apennines (about 20–25 km) must be regarded as an extended marginal part of the Adriatic Microplate. Southwards, the thin crust of the Northern Apennines passes into the thin oceanic crust of the Ligurian Sea.

Ligurian Sea, Tuscany, and Adriatic Sea

Figure 11 shows a WSW–ENE-trending cross-section between the Ligurian Sea in the west and the Adriatic Sea in the east, which passes by northern Corsica, Elba, Tuscany, and the Umbrian Apennines. This profile demonstrates the complex structure of this divergent orogenic belt, which resulted from repeated deformation with changing transport directions along the margin of the Adriatic Microplate. The counterclockwise rotation of the Corsica–Sardinia block has strongly affected the geometry and nature of the Moho-discontinuity west and east of Corsica. Along a south-west and north-east section, four different Moho-discontinuities can be identified (Figure 11): (1) a young oceanic Moho-discontinuity beneath the Ligurian Sea and Provençal Basin is situated at a depth of 14–18 km; (2) beneath the Variscan continental crust of Corsica, the European Moho-discontinuity has been detected at a depth of 25–30 km; (3) the young Moho-discontinuity beneath Elba and Tuscany is situated at a depth of only 20–25 km (the structural relations between the Corsican and the Tuscany Moho are a matter of debate: there is probably no direct link between the discontinuities); and (4) a well-expressed offset of 20–30 km separates the shallow Tuscan Moho-discontinuity from the deep Moho-discontinuity beneath the Northern Apennic Arc (this offset defines the crustal root of the Apennines).

Along the section (Figure 11), there are three different zones in which the crustal thickness changes. In the westernmost part, the transition from the thin and young oceanic crust of the Ligurian Sea to the Variscan continental crust of Corsica can be interpreted as a passive continental margin. The situation is completely different in the contact zone between Corsica and Tuscany. Due to the counterclockwise rotation of Corsica, an eastward-dipping subduction zone must be postulated beneath Elba and western Tuscany. In the reflection profile CROP03 beneath Elba and western Tuscany, weak, eastward-dipping reflections indicate the existence of an interface that may be interpreted as the top of an eastward-dipping subduction zone. Very probably there is no direct link between the Moho-discontinuity beneath Corsica and the Moho-discontinuity beneath Tuscany. The Tuscany crust in the hinterland of the Northern Apennines is characterized by back-arc extension that has reduced the crust to a thickness of only 20 km. The third change in crustal thickness occurs beneath the Apenninic Arc. In this area, the Tuscany Moho-discontinuity terminates and another deeper Moho has been detected at a depth of 40–50 km. There is no link between the shallow Tuscany Moho and the deep Moho-discontinuity of the Adriatic Microplate. The offset between the two Moho-discontinuities amounts to about 20 km. This break can be interpreted as a continent–continent collision zone. A normal crustal thickness of 30–35 km has been measured beneath the Adriatic coast, the foreland of the Apenninic Arc.

The Subcrustal Lithosphere in Europe

The lithosphere is defined as the rigid outermost shell of Earth, comprising continental and oceanic crust and the uppermost mantle. The Moho-discontinuity is an intralithospheric interface. The lithosphere is a strong, elastic layer overlying the weak, ductile asthenosphere with a more or less gradual transition. The ductile asthenosphere is believed to represent the zone in the mantle where the melting point is most closely approached by the actual temperature curve. Because it transmits shear waves, this layer cannot be completely molten, but small amounts of melt may be present. The thickness of the lithosphere, like that of the crust, exhibits important variations. The thickness of the European lithosphere is depicted by the contour map in Figure 12. The cross-section reveals four main features. First, beneath the Baltic Shield and the Eastern European Platform, the thickness of the lithosphere is about 200 km. This feature correlates with the thick continental crust and low heat-flow values (40–50 mW m ^{2}). Second, a thin lithosphere of about only 50 km exists beneath the western region of Central Europe, especially along the Rhine and Rhône rivers. Here, also, the crust is relatively thin ($\leq$30 km) and the heat flow is distinctly higher than in Northern and Eastern Europe (80–90 mW m ^{2}). Third, in the Mediterranean region, the thickness of the lithosphere is very heterogeneous. A thick lithospheric root has been detected beneath the Alps and the Northern Apennines. This feature is at least partly explained by sinking delaminated lithospheric slabs, subsequent to continental collision. Under the young oceanic basins such as the

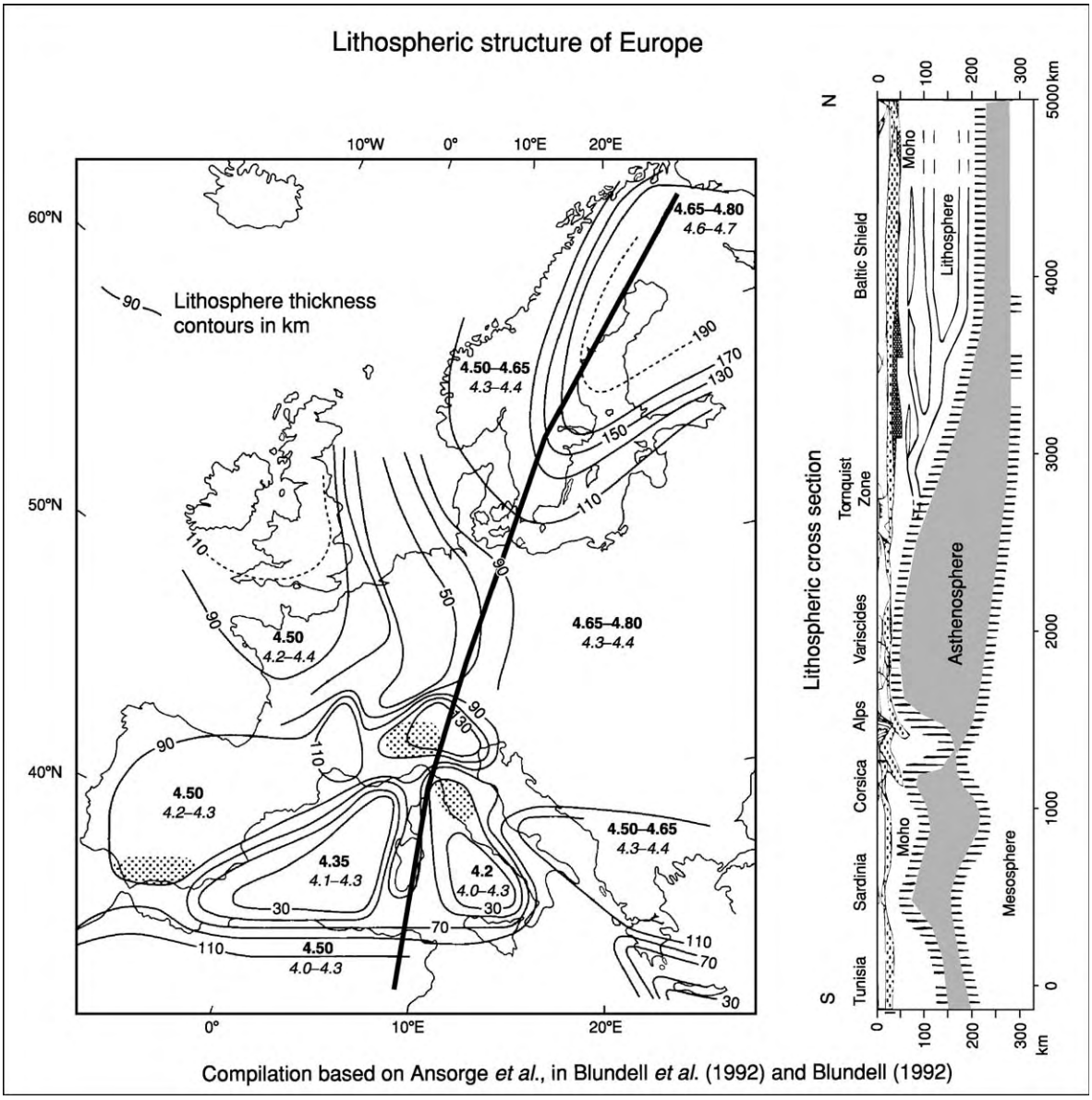


Figure 12 (Left) Contour map of the depth to the base of the lithosphere of Europe, derived from an interpretation of surface wave data. Numbers refer to S wave velocities above and below the base of the lithosphere. Shaded areas indicate thickened lithosphere. Reproduced with permission from Panza (1992). (Right) North to south cross section of Europe, demonstrating very clearly the basic differences of lithospheric thickness between the Baltic Shield in the north and the Variscan and Alpine part in Central and Southern Europe. Note the lithospheric root beneath the Alps. Modified from Blundell *et al.* (1992).

Tyrrhenian Sea, the asthenosphere is thin (10–20 km). Due to the very different tectonic situation, the heat flow varies strongly between 50 and 150 mW m $^{-2}$. Fourth, there is relatively thin lithosphere (about 50–80 km) in the Baltic Shield, but a thick one (about 200 km) exists beneath Central Europe.

Although recent tectonic and magmatic events in Europe are mainly concentrated in the Mediterranean region, there has been some recent activity in Central and Western Europe. Late Cretaceous to Holocene volcanism is found in an east- to west-trending belt that extends from western Poland to western Germany, continuing southwards across the French Massif Central to the Mediterranean. Recent seismological studies in the Eifel region and in the northern Massif Central have revealed low-velocity anomalies in the lithosphere and down to at least depths of 300–400 km. These findings support the concept that deep-seated mantle plumes can produce intraplate mantle volcanism.

Analysis

Studies of the European Moho-discontinuity offer the possibility to compare crustal structures of very different ages and evolutionary stages. The Moho-discontinuity is not a static feature in the lithosphere, but rather evolves geodynamically. In Europe, the Moho-discontinuity in each of the four main regions is distinguished by a specific nature and geometry. The Baltic Shield, including the East European Platform, is characterized by a 40- to 50-km-thick continental crust. In the shield regions, the average crustal velocity and density are relatively high due to a thick lower crust with mafic components. These shield regions have remained rather stable and undeformed in the Phanerozoic Eon, during which the crust of Central, Western, and Southern Europe has been repeatedly reworked. A relatively thin crust, only 30 km thick, characterizes the Variscan belts in Central and Western Europe. The Variscan Orogen was formed in Devonian and Carboniferous times by the sequential collision of continental fragments, producing a thickened crust such as that in the recently formed Alps. The Variscides subsequently collapsed and then re-equilibrated in Permian times with extensional overprinting. This re-equilibration was associated with widespread crustal melting, extensive volcanism, and the formation of late- to post-orogenic sedimentary basins. Consequently, the post-Variscan Moho image is mainly related to post-orogenic pervasive subhorizontal shearing associated with magmatic underplating. This layered structure of the lower crust is imaged by the highly reflective crust under sedimentary basins. An exception to this general evolution exists in the Palaeozoic orogen of the Urals, which have preserved their collisional structure with a crustal root. Apparently, this root was never destabilized by major fluid flow and heating.

The third main European region was formed by the young orogens in the Mediterranean regions of Southern Europe, which are characterized by a very heterogeneous crustal and Moho-discontinuity structure, morphology, and composition. Finally, spreading along the Mid-Atlantic Ridge since the Early Jurassic produced oceanic crust and an oceanic Moho-discontinuity.

The differences between the Baltic Shield and west-Central and Western Europe do not require different plate tectonic mechanisms, but can be explained by higher temperatures and rheological softening of the Variscan crust. This may ultimately be due to the repeated emplacement of melts rich in heat-producing radioactive elements. The Variscides have passed the three stages typical of an orogenic cycle: compression and crustal thickening, metamorphism and extensional collapse of the orogen. By comparison with the Variscides, the evolution of most of the young Mediterranean mountain belts has advanced only to the second stage or to an early part of the third stage. The mountain roots of these young orogenic belts are still preserved and are clearly recognizable in the seismic records. In all orogenic belts, it cannot be excluded that lower crustal rocks are present in eclogite facies and therefore escape detection by geophysical methods.

See Also

Earth: Crust. **Europe:** East European Craton; Caledonides of Britain and Ireland; Scandinavian Caledonides (with Greenland); Variscan Orogeny; The Urals; Permian to Recent Evolution; The Alps; Mediterranean Tectonics. **Minerals:** Olivines. **North America:** Continental Interior. **Plate Tectonics**. **Russia**. **Seismic Surveys**. **Shields**. **Tectonics:** Mountain Building and Orogeny. **Ultra High Pressure Metamorphism**.

Further Reading

Blundell D, Freeman R, and Mueller S (eds.) (1992) *A Continent Revealed: The European Geotraverse,* pp. 1 275. Cambridge: Cambridge University Press.

Bois C, Cazes J, Choukroune P, Gariel O, Hirn A, La Gall B, Lefort J P, Matte P, and Pinet B (1994) Seismic reflection image of the Pre Meseozoic crust in France and adjacent areas. In: Keppi JD (ed.) *Pre Mesozoic Geology in France and Related Areas,* pp. 3 47. Berlin Heidelberg: Springer Verlag.

Echtler HP, Stiller M, Steinhoff F, Krawczyk CM, Suleimanov A, Spiridonov V, Knapp JH, Menshikov Y, Alvarez Marron J, and Yunusov N (1996) Preserved col lisional crustal architecture of the southern Urals revealed by Vibroseis CMP profiling. *Science* 274: 224 226.

Gee DG and Zeyen HJ (eds.) (1996) *Europrobe Litho sphere Dynamics: Origin and Evolution of Continents,* pp. 1 138. Uppsala, Sweden: EUROPROBE Secretariat.

Giese P and Pavlenkova N (1988) *The Structural Maps of the Earth's Crust of Europe* (In Russian). Moscow: Nauka.

Kaban MK (2001) Gravity model of the North Eurasia crust and upper mantle: 1. Mantle and static residual gravity anomalies *Russian Journal of Earth Sciences* 01.3. No.2.

Meissner R, Wever T, and Flüh ER (1987) The Moho in Europe: Implications for crustal development. *Ann. Geo phys.* 3: 357 364.

Morelli C (1998) Lithospheric structure and geodynamics of the Italian Peninsula derived from geophysical data: a review. *Memorie della Societä Geologica Italiana* 52: 113 122.

Pfiffner OA, Lehner P, Heitzmann P, Mueller S, and Steck A (eds.) (1997) *Deep Structure of the Alps, results from NRP 20,* pp. 1–380. Basel Switzerland: Birkhäuser.

Ponziani F, de Franco R, and Biella G (1998) Geophysical reinterpretation of 1974 and 1978 DSS experiments along CROP03 profile. *Memorie della Società Geologica Italiana* Vol. LII: 215–224.

Tittgemeyer M, Wenzel F, Ryberg, T, and Fuchs K (1999) Scales of heterogeneities in the continental crust and upper mantle. *Pure Appl. Geophys.* 156: 29–52.

MOON

See **SOLAR SYSTEM: Moon**

ENCYCLOPEDIA OF
GEOLOGY

ENCYCLOPEDIA OF GEOLOGY

EDITED BY

RICHARD C. SELLEY
L. ROBIN M. COCKS
IAN R. PLIMER

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

NEW ZEALAND

N Mortimer, Institute of Geological and Nuclear Sciences, Dunedin, New Zealand

Introduction

The south-west Pacific Ocean is a region of isolated islands and submerged plateaus and ridges (Figure 1). The three main islands of New Zealand (North, South, and Stewart) make up the largest landmass group in the region. Schists, greywackes, and granitoids are exposed in islands on the Challenger Plateau and Chatham Rise and have been sampled in dredges on the Campbell Plateau and Norfolk Ridge, thus demonstrating their continental geological character. Abyssal Pacific oceanic-crustal floor typically lies at water depths of about 5000 m, and the boundary between continental crust and oceanic crust is marked by a generally pronounced slope break at about 2500 m water depth. The wider area of continental crust in the New Zealand region (Figure 1) is about one-third the area of on-land Australia and is commonly referred to as Zealandia.

On-land New Zealand contains a wide variety of Phanerozoic rocks (Figure 2), which preserve a detailed record of the Cambrian to early Early Cretaceous convergent margin of southern Gondwana, late Early Cretaceous rifting, a Late Cretaceous–Palaeogene passive margin, and the Neogene–Holocene active convergent and strike-slip margin. So much of continental Zealandia is submerged because of the widespread Cretaceous extension and rifting. It was only with the development of the Neogene–Holocene convergent plate boundary that about 10% of Zealandia emerged above sea-level.

A distinction is generally drawn in New Zealand between pre-late Early Cretaceous (more than 105 Ma) 'basement' rocks, which are commonly metamorphosed and generally highly deformed, and 'cover' rocks, which are younger than 105 Ma, poorly indurated, well stratified, and less deformed.

Palaeogeographical Reconstructions

The shape of the continental crust of Zealandia has changed throughout the Phanerozoic. From the Cambrian to the Early Cretaceous, the New Zealand part of the Gondwanan margin grew by the magmatic and tectonic addition of batholiths and terranes. In the last 100 Ma this continental crust has been thinned, rotated, and translated in response to multiple tectonic events. Reconstructions (Figure 3) involve subtracting the 480 km Neogene dextral strike-slip movement on the Alpine Fault, the 45° Oligocene–Miocene rotation between the Pacific and Australian plates, and 4000 km of northwards drift. Small crustal blocks within 100 km of the Alpine Fault (i.e. most of on-land New Zealand) have undergone strong Cenozoic deformation. In the pre-rift (100 Ma) palaeogeography (Figure 3A) Zealandia is in a near-polar position and contiguous with Tasmania and Antarctica. By 10 Ma, some movement on the Alpine Fault had taken place and modern-day New Zealand had been isolated by seafloor spreading.

Geological Basement

At a regional scale, the volcanic, sedimentary, plutonic, and metamorphic basement rocks of New Zealand can be described in terms of a number of western and eastern tectonostratigraphic terranes, composite regional batholiths intruding these terranes, and schist, gneiss, and mélange overprints on

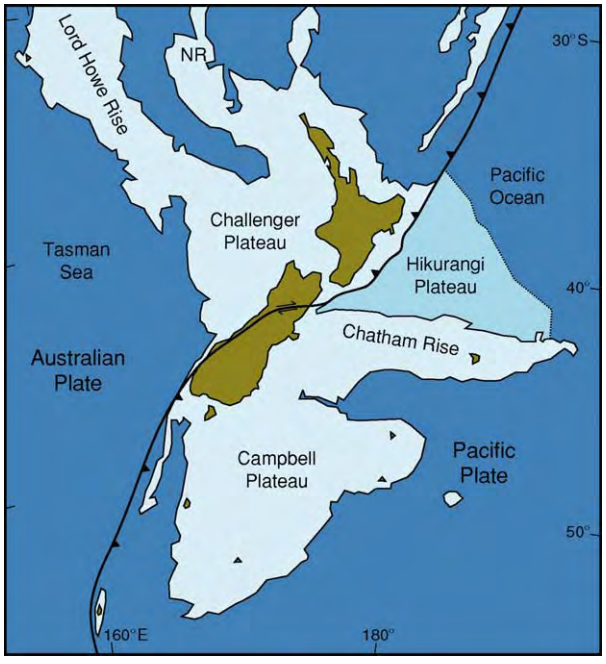


Figure 1 The outline of the area of continental crust in the New Zealand region (Zealandia). Land and major islands are pale brown; water less than 2500 m deep is pale blue. Deep ocean floor is dark blue and Hikurangi plateau large igneous province is intermediate blue. The present day Pacific Australian plate boundary is shown by the thick black line, with teeth on the over riding plate. Only about 10% of Zealandia is emergent above sea level as the North and South Islands. NR, Norfolk Ridge.

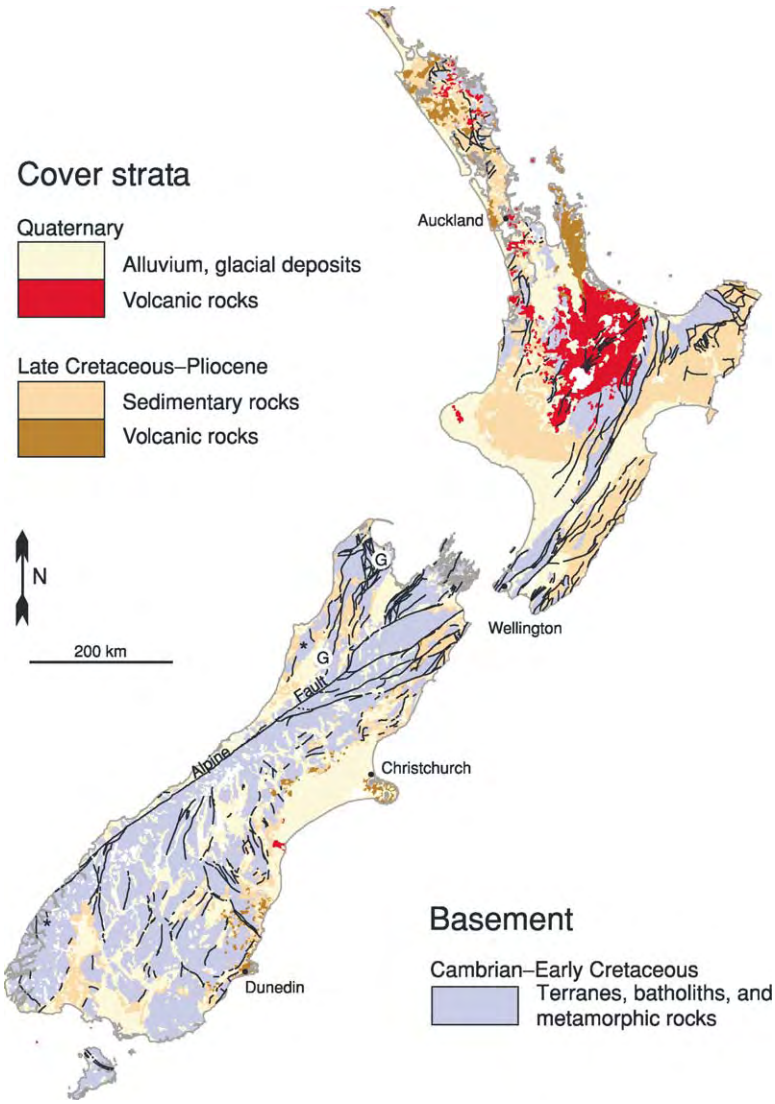


Figure 2 Simplified present day on land geology of New Zealand. The black lines are faults. Many of the lakes and valleys of the South Island are the result of Quaternary glacial erosion. G, locations of Devonian to Jurassic Gondwanan sequences; ∗, locations of Cretaceous metamorphic core complexes.

the terranes and batholiths (Figure 4). Figure 5 shows the names, age ranges, and mutual geometric relationships of the constituent basement units on an Early Cretaceous reconstruction (Figure 3A). No Precambrian rocks are exposed; New Zealand has been near a continent–ocean margin throughout the Phanerozoic.

Western Province Terranes

The Western Province terranes lie west of the Median Batholith and comprise the Early Palaeozoic Buller and Takaka terranes. The Buller Terrane consists of variably metamorphosed siliciclastic sandstones and mudstones, of continental Gondwanan provenance, and is the westernmost recognized terrane in New Zealand (i.e. the terrane closest to the Gondwanan cratonic core). Rare fossils are of Ordovician age, but a Buller Terrane paragneiss contains detrital zircons as old as 3400 Ma (Archean; New Zealand's oldest known geological material). Intercalated volcanics are absent. The Takaka Terrane consists of siliciclastic, carbonate, and volcanic rocks. Middle Cambrian trilobites in the Takaka Terrane are New Zealand's oldest known fossils. The Takaka Terrane is generally well stratified and lithologically diverse, and includes Cambrian ultramafics and boninites, Ordovician limestones, and Silurian orthoquartzites. The Buller and Takaka Terranes were accreted to Gondwana by the Devonian.

Eastern Province Terranes

The Eastern Province terranes lie east of the Median Batholith and comprise the Brook Street, Murihiku,

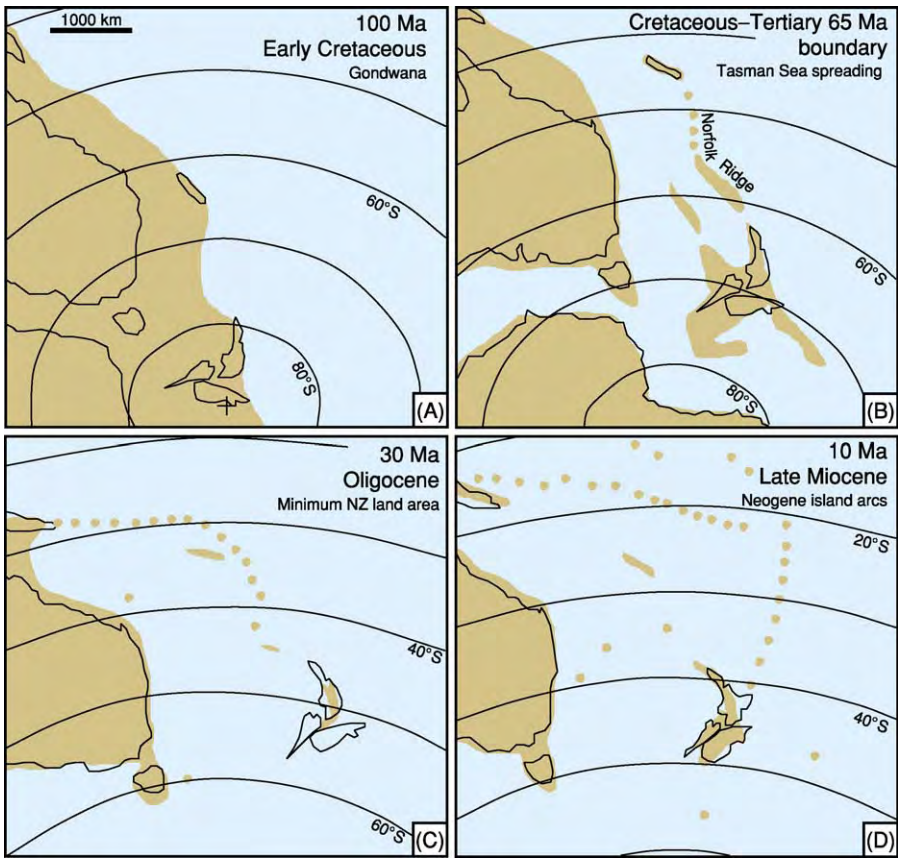


Figure 3 Palaeogeographical reconstructions of New Zealand at approximately (A) 100 Ma, (B) 65 Ma, (C) 30 Ma, and (D) 10 Ma. Brown, inferred land including schematic island arc chains; black lines, present day coastlines of North and South Islands and the Alpine Fault. On these reconstructions, the Alpine fault divides the South Island into eastern and western parts. (Reproduced with permission from Lee DE, Lee WG, and Mortimer N (2001) Where and why have all the flowers gone? Depletion and turnover in the New Zealand Cenozoic angiosperm flora in relation to palaeogeography and climate. *Australian Journal of Botany* 49: 341 356.)

Maitai, Caples, Rakaia, and Pahau terranes. Carboniferous conodonts are known from a limestone in the Rakaia Terrane, but the age range of clastic rocks in Eastern Province terranes is from Permian to Early Cretaceous (Figure 5). The Eastern Province terranes are thus entirely younger than the Western Province terranes and represent accretion of material to Gondwana in the Mesozoic.

The Brook Street Terrane is a Permian subduction-related isotopically primitive pyroxene-rich basalt-dominated volcanic pile and volcaniclastic apron, in places up to 14 km thick, which is intruded by Permian layered gabbros and trondhjemite plutons that are now part of the Median Batholith. New Zealand's only known *Glossopteris*, a Gondwanan leaf fossil, occurs in the Brook Street Terrane. The Murihiku Terrane comprises a 9–13 km Late Permian to Late Jurassic volcaniclastic marine succession of sandstone with lesser conglomerates, mudstones, and tuffs. It has the simplest internal structure of all the Mesozoic New Zealand terranes, a broad synclinorium that is traceable for 450 km through the North and South Islands.

The Maitai Terrane consists of the eastern Early Permian (285–275 Ma, according to uranium–lead dating of zircon) Dun Mountain Ophiolite Belt, which is unconformably overlain by 6 km of well-stratified Late Permian to Middle Triassic volcaniclastic sedimentary rocks. The ophiolite originated in a near-arc setting. The Brook Street, Murihiku and Maitai terranes are adjacent to each other as a Permian–Triassic arc, fore-arc, and exhumed near-arc ophiolite, respectively.

The Caples, Bay of Islands, and Rakaia terranes contrast with the aforementioned Eastern Province terranes in that their Permian–Jurassic clastic sequences are tectonically imbricated with ocean-floor basalt, chert, and limestone associations; all three terranes grade into the pumpellyite–actinolite to amphibolite facies Haast Schist. Deposition occurred as submarine-fan deposits in lower trench-slope basins, before juxtaposition in an accretionary

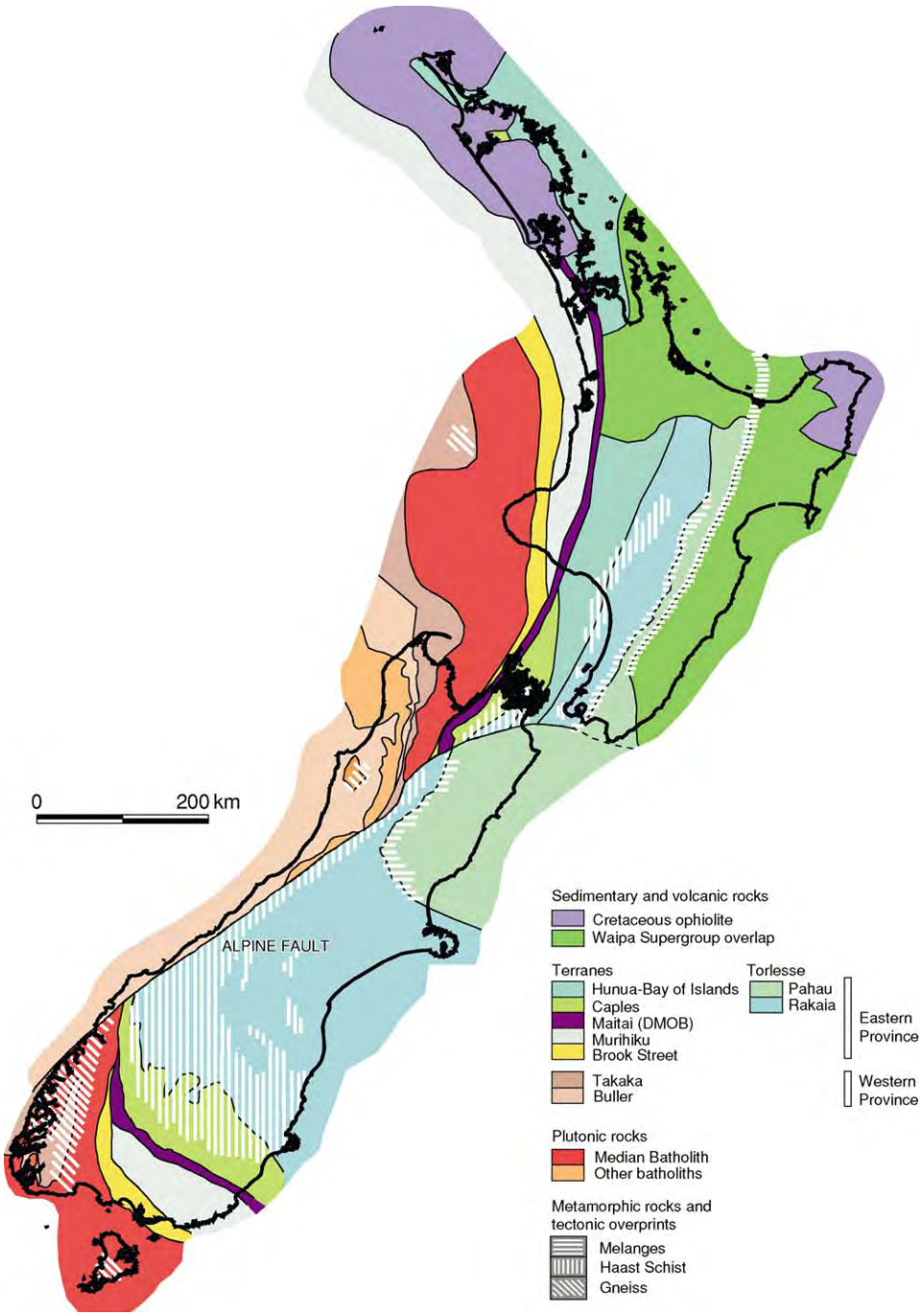


Figure 4 Basement geological subdivisions of New Zealand. Minor outliers of Permian Gondwanan sequences rest on Takaka Terrane, and Devonian, Triassic and Jurassic Gondwanan sequences on Buller Terrane. DMOB, Dun Mountain Ophiolite Belt. (Reproduced with permission from Mortimer N (2004) New Zealand's geological foundations. *Gondwana Research* 7: 261 272. © International Association for Gondwana Research.)

prism. Compositional and provenance differences are used to discriminate the three terranes: Rakaia sandstones are quartz rich, plutoniclastic, and of average rhyodacitic composition and are thus compositionally distinct from the more dacitic to andesitic volcaniclastic-dominated sandstones in Caples and the Bay of Islands. The Pahau Terrane has a similar lithology and structure to the Rakaia Terrane, but its depositional ages extend into the Late Jurassic and Early Cretaceous and it contains tuffs. Much of the Pahau clastic detritus is probably recycled from Rakaia rocks, but a volcanic input, probably from the Median Batholith, is also required. The Pahau Terrane probably represents trench deposits that

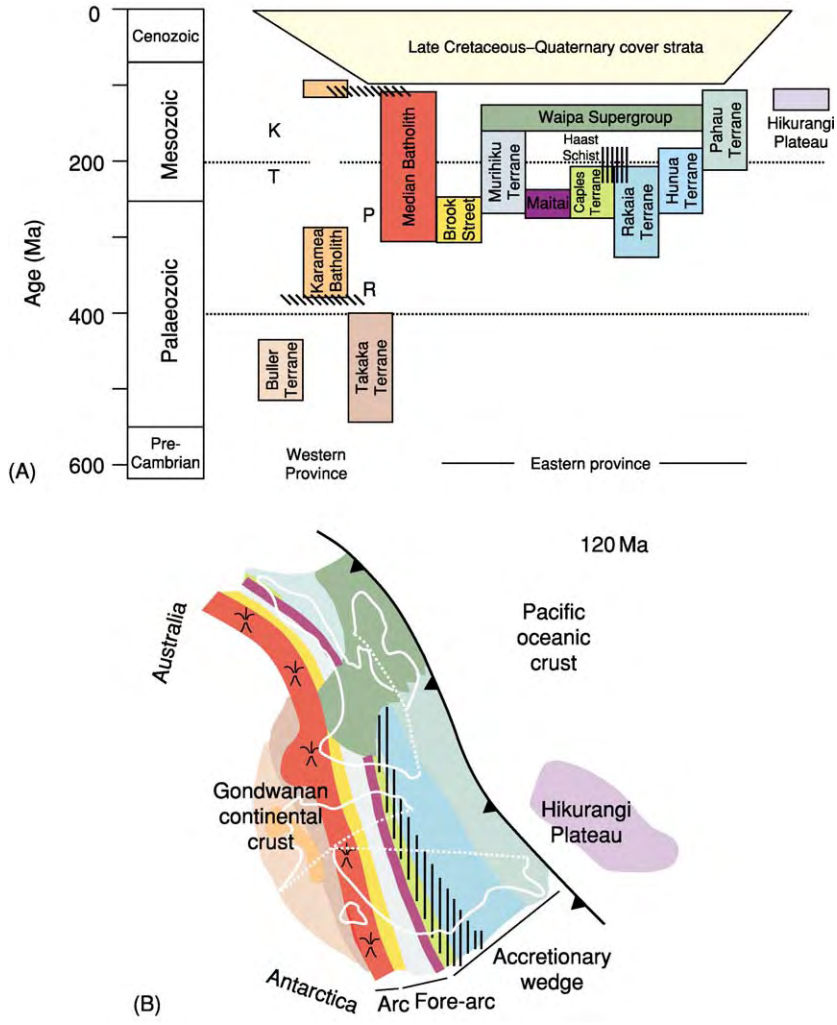


Figure 5 (A) Summary of the age ranges of New Zealand's basement terranes (green, blue, brown, yellow), batholiths (red, orange), and metamorphic rocks (overprint stripes). Gondwana Sequence rocks are shown by letters: K, Kirwans Dolerite; T, Topfer Formation; P, Parapara Group; R, Reefton Group. The terranes can be grouped into Eastern and Western provinces. (B) One possible palaeogeographical reconstruction of south eastern Gondwana at about 120 Ma (the end of the convergent margin phase). The present day New Zealand coastlines are shown as white lines; the Alpine Fault and other faults are shown as white dotted lines. (Reproduced with permission from Mortimer N (2004) New Zealand's geological foundations. *Gondwana Research* 7: 261 272. © International Association for Gondwana Research.)

were laid down and deformed towards the end of Cretaceous subduction.

These nine terranes make up the bulk of the New Zealand volcanosedimentary basement. Smaller tectonostratigraphic units can be regarded as components of the larger terranes. The Median Tectonic Zone is still used by some New Zealand geologists to describe a zone of terrane shards and igneous complexes of uncertain status and correlation that lies between the Brook Street and Takaka terranes (see Plutonic Rocks, below).

Overlap Sequences

Overlap sequences of varying ages can be recognized by their lesser deformation and metamorphism and distinctive petrofacies, as compared with older immediately underlying rocks. Lateral correlatives are used to constrain models of terrane amalgamation and accretion. The only New Zealand rocks that have been correlated with autochthonous Gondwanan sequences occur in two small outliers in northern South Island (marked G in Figure 2). Four units – the Reefton Group (marine Devonian), Parapara Group (marine Permian–Triassic), Topfer Formation (nonmarine Triassic), and Kirwans Dolerite (a Middle Jurassic low-titanium tholeiite sill intrusion) – indicate that the Buller and Takaka terranes had been accreted to Gondwana by the end of the Palaeozoic. In North Island, a postulated Late Jurassic overlap sequence – the Waipa Supergroup, possibly sourced

from the Median Batholith – may indicate that most of the basement terranes had accreted by this time. The post-105 Ma cover sediments provide a firm minimum age for the mutual juxtaposition and accretion of the basement terranes.

Plutonic Rocks

There are three composite regional batholith-sized belts (more than 100 km^2) of plutons in New Zealand, and numerous smaller isolated plutons. The Median Batholith is a composite Cordilleran batholith with intrusive contacts against the Brook Street and Takaka terranes. It comprises dozens of 1–10 km Carboniferous to Early Cretaceous gabbroic to granitic subalkaline I-type plutons. The eastern half of the Median Batholith has also been called the Median Tectonic Zone. The ages and average compositions of the rocks change across the batholith axis: Permian gabbroids dominate the eastern edge, Triassic to early Early Cretaceous dioritoids dominate the central part, and late Early Cretaceous adakitic granitoids are found on the western margin. Roof pendants of petrologically related volcanosedimentary rocks occupy about 5% of the batholith area. How much of the Median Batholith is allochthonous is debatable.

The Karamea–Paparoa Batholith lies entirely within the Buller Terrane. Its constituent plutons are dominated by Devonian–Carboniferous I-type and S-type granites. The Hohonu Batholith, also within the Buller Terrane, represents 105–82 Ma plutonism associated with the change from convergence to rifting.

Metamorphic Overprints

Regionally extensive metamorphic overprints include Devonian and Cretaceous polymetamorphic amphibolite–granulite facies gneisses (formed from Buller, Takaka, Median, and Karamea–Paparoa protoliths). Some of the gneisses are confined to the lower plates of two Cretaceous metamorphic core complexes in the South Island (stars in Figure 2). The Haast Schist of Jurassic–Cretaceous pumpellyite–actinolite to amphibolite facies overprints the Caples, Bay of Islands, and Rakaia terranes. Metamorphism probably took place in the deep parts of a Jurassic–Early Cretaceous accretionary wedge. Exhumation of the Haast Schist belt was episodic, with most of the schist being at the surface by 105 Ma and deeper levels being exhumed along the Alpine Fault from 20 Ma to the present day. The third major kind of regional tectonic overprint is the Early Cretaceous mélange that is present as belts in and between the Rakaia and Pahau terranes. In part this mélange is probably coeval with the Haast Schist mineral growth, but it formed in the shallower parts of the accretionary wedge. The steep mélanges may represent zones of strike-slip deformation.

Episodes of tectonic activity were previously described as orogenies. In the early 1980s it was realized that the Mesozoic Rangitata Orogeny was probably a composite of an older subduction-related event and a younger extension-related event. With the recognition that much of the intrusion, metamorphism, and deformation in the Buller and Takaka terranes ('Tuhua Orogen') occurred during the Cretaceous (of 'Rangitata' age), the orogen terminology became obsolete.

Cover Strata

Late Early Cretaceous to Holocene rocks rest unconformably on all the older basement units. The last 100 Ma of New Zealand's geological history can conveniently be divided into four periods, which are discussed in the following sections.

Late Early Cretaceous Intracontinental Rifting

Estimates of the timing of the end of Palaeozoic–Mesozoic subduction range from 125 Ma to 85 Ma. The youngest clearly subduction-related plutonic suites are 125 Ma old, and the oldest rift-related alkaline volcanics are about 100 Ma old. Detrital zircon dates of about 100 Ma have been obtained from sandstones imbricated in the accretionary wedge of the Pahau Terrane, but ignimbrites that fill extensional half-grabens are 105–100 Ma old. The two metamorphic core complexes, Paparoa and Fiordland, also attest to extreme local continental extension at around 105 Ma. Hypotheses about the reason for the cessation of subduction in the Cretaceous include migration of a spreading ridge along the trench, stalling of spreading outboard of the trench, and collision of the Hikurangi Plateau large igneous province (Figure 1) with the Gondwanan margin. Late Early Cretaceous sedimentary rocks typically comprise synrift non-marine deposits succeeded by passive-margin marine transgressive strata.

Late Cretaceous–Palaeogene Passive Margin

The oldest oceanic crust adjacent to the Challenger and Campbell Plateaus is about 85 Ma old. Spreading in the Tasman Sea ceased at about 55 Ma. New Zealand moved north (Figure 3) due to continuing spreading on the Pacific–Antarctic Ridge. Marine basins developed across Zealandia as a result of post-rift thermal subsidence. The maximum marine inundation of Zealandia, with widespread limestone deposition, occurred in the Oligocene, but local fluvial and coal deposits dating from throughout the

Late Cretaceous–Cenozoic (albeit in different parts of North and South islands) indicate that Zealandia was never entirely submerged.

Miocene–Pliocene Active Margin Development

By earliest Miocene times, a new plate boundary (in broadly the same place as the present-day boundary; Figure 1) had propagated through Zealandia and cut across the trend of the basement terranes. The development of the Pacific–Australian plate boundary had profound geological consequences. The Northland and East Coast allochthons, consisting of Late Cretaceous–Palaeogene ophiolitic and sedimentary rocks, were thrust onto North Island from the north-east at the end of the Oligocene (*ca*. 25 Ma). Subduction-related arc volcanism started at about 25 Ma and became widespread in Northland and the Coromandel between 15 Ma and 5 Ma. On South Island, intraplate stratovolcanoes developed along the east coast. The Neogene succession is generally thick, clastic-dominated, and regressive.

Quaternary

Although they are part of the continuing Neogene volcanotectonic phase, Quaternary rocks are shown separately in Figure 2 because of their wide areal extent and strong association with present-day landforms. The subduction-related Quaternary volcanoes of North Island have erupted extensive ignimbrite sheets, which blanket the older rocks. In South Island large fluvioglacial outwash plains issue from glacially eroded valleys.

Conclusions

The Phanerozoic geological history of New Zealand can be interpreted in terms of the progressive Pacificwards growth of Gondwana by terrane accretion and batholith intrusion at an obliquely convergent margin. Continental growth was terminated by widespread extension in southern Gondwana from about 105 Ma and was followed by seafloor spreading from about 85 Ma in the Tasman Sea and Southern Ocean. The entire basement and its cover of passive-margin Late Cretaceous to Palaeogene sediments were subjected to renewed deformation in the Neogene with the inception of the modern Australia–Pacific plate margin.

See Also

Antarctic. **Australia:** Tasman Orogenic Belt. **Gondwanaland and Gondwana**. **Large Igneous Provinces**. **Oceania (Including Fiji, PNG and Solomons)**. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges.

Further Reading

Adams CJ, Campbell HJ, Graham IJ, and Mortimer N (1998) Torlesse, Waipapa and Caples suspect terranes of New Zealand: Integrated studies of their geological history in relation to neighbouring terranes. *Episodes* 21: 235 240.

Ballance PF (ed.) (1993) *South Pacific Sedimentary Basins*. Sedimentary Basins of the World 2. Amsterdam: Elsevier.

Bradshaw JD (1989) Cretaceous geotectonic patterns in the New Zealand region. *Tectonics* 8: 803 820.

Cole JW (1986) Distribution and tectonic setting of Late Cenozoic volcanism in New Zealand. *Bulletin of the Royal Society of New Zealand* 23: 7 20.

King PR (2000) Tectonic reconstructions of New Zealand: 40 Ma to the present. *New Zealand Journal of Geology and Geophysics* 43: 611 638.

Korsch RJ and Wellman HW (1988) The geological evolution of New Zealand and the New Zealand region. In: Nairn AEM, Stehli FG, and Uyeda S (eds.) *The Ocean Basins and Their Margins*, vol. 7B, pp. 411 482. New York: Plenum Press.

Lee DE, Lee WG, and Mortimer N (2001) Where and why have all the flowers gone? Depletion and turnover in the New Zealand Cenozoic angiosperm flora in relation to palaeogeography and climate. *Australian Journal of Botany* 49: 341 356.

Mortimer N (2004) New Zealand's geological foundations. *Gondwana Research* 7: 261 272.

Sutherland R (1999) Basement geology and tectonic development of the greater New Zealand region: an interpretation from regional magnetic data. *Tectonophysics* 308: 341 362.

NORTH AMERICA

Contents

Precambrian Continental Nucleus

W Bleeker, Geological Survey of Canada, Ottawa, ON, Canada

Introduction

North America is a large continent and much of it is ancient (Figure 1). Fringed along several of its coastlines by younger mountain ranges, the broad interior of the continent is underlain by crust that ranges in age from >4.0 Ga to <1 Ga. This ancient crust (Figure 2) formed and progressively amalgamated during several broad 'orogenic' or mountain-building cycles. Although recording net crustal growth, this protracted history of orogenic cycles was far from linear – it involved numerous crust-forming (e.g., growth of volcanic islands) and crust-destroying (e.g., partial subduction) events, punctuated in time. It also involved a complex choreography of tectonic collisions between small lithospheric plates, subsequent rifting and dispersal events, re-aggregation, associated deformation, partial melting and metamorphism, and large-scale rearrangements along faults. Ultimately, at about 1 Ga, this protracted crustal evolution culminated in the growth of a large supercontinent known as Rodinia. In essence, the ancient core of continental North America, which is generally referred to as Laurentia, is a large fragment of this vast late Proterozoic supercontinent (*see* **Precambrian:** Overview).

Between 800 and 600 Ma, supercontinent Rodinia started to break up, resulting in the plate tectonic dispersal of about a dozen large continental fragments. This event liberated the large North American fragment (Laurentia) out of the parental landmass of Rodinia. Since this breakup, the west coast of Laurentia (present coordinates) has been a long-lived active margin facing oceanic plates and colliding island arcs, whereas its eastern margin was modified by yet another major collision and rifting cycle, first forming the Appalachian-Caledonian Mountain (*see* **North America:** Southern and Central Appalachians) Belt, and finally rifting, starting at *ca*. 200 Ma, to form the present-day Atlantic Ocean basin. (*see* **North America:** Atlantic Margin) These events have modified the margins of Laurentia and added marginal terranes of mostly younger crust, such as the Coast Mountains of British Columbia, large parts of Alaska (*see* **North America:** Northern Cordillera), and Gondwana-derived terranes along the eastern seaboard. Nevertheless, North America remains dominated by Precambrian crust, with a mean isotopic age >2 Ga.

The ancient crustal core of the continent, parts of which are exposed in the Canadian Shield (Figure 2), is underlain by subcontinental mantle lithosphere of above average thickness, which locally reaches down to a depth of 200–300 km into the hotter, convective mantle. This 'mantle keel' developed during or shortly after the amalgamation events in the crust, either by thermal growth (i.e., cooling from the top downwards) or by lateral accretion of buoyant slabs of depleted peridotite. The lithospheric keel is currently a reservoir of diamonds, a high-pressure polymorph of carbon. Transported to the surface by exotic volcanic rocks known as kimberlites, these valuable gemstones are now being produced from several mines across northern Canada (Figure 3). The mantle keel is mechanically coupled with the overlying crust, forming a thick and somewhat cooler lithosphere that enhances the strength and

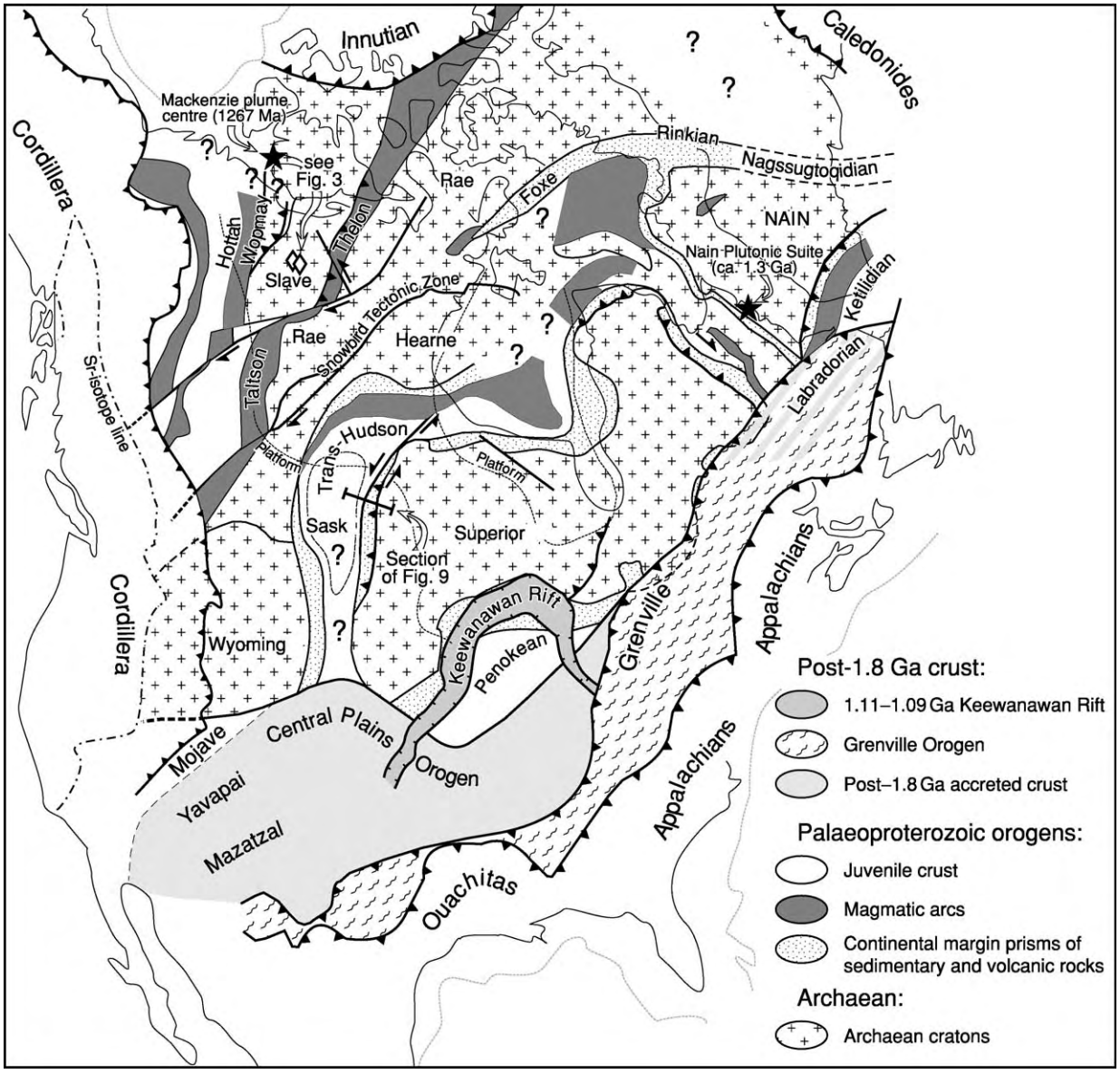


Figure 1 Simplified tectonic map of the Precambrian basement of North America, with Greenland in predrift position (modified after Hoffman, 1988; and Ross *et al.*, 1991).

long-term stability of the ancient continental core of Laurentia.

Modern Plate Tectonic Context and Extent of Laurentia

The continent of North America is an integral part of the North American plate, one of about 12 large lithospheric plates that presently form the segmented outer mechanical and thermal boundary layer of the Earth. These plates float on a solid but convecting mantle. Due to forces acting on their bounding surfaces (e.g., 'slab pull', but also 'ridge push', 'basal drag', and frictional forces), plates drift and interact laterally. Typical absolute plate velocities on the modern Earth are on the order of 1–10 cm year, with the North American plate moving in a west-south-westerly direction at *ca*. 2 cm year (centre of the continent), around a rotation pole located in north-western Equador.

Along its leading edge, the North American plate overrides subducting oceanic plates (Juan de Fuca, Cocos), which fuel arc volcanism along the Cascade and Central American margins. Elsewhere, relative plate motion, with respect to the main Pacific plate, is oblique or sideways, resulting in a transform fault boundary with right-lateral motion (the San Andreas and related faults). Collectively, these processes cause

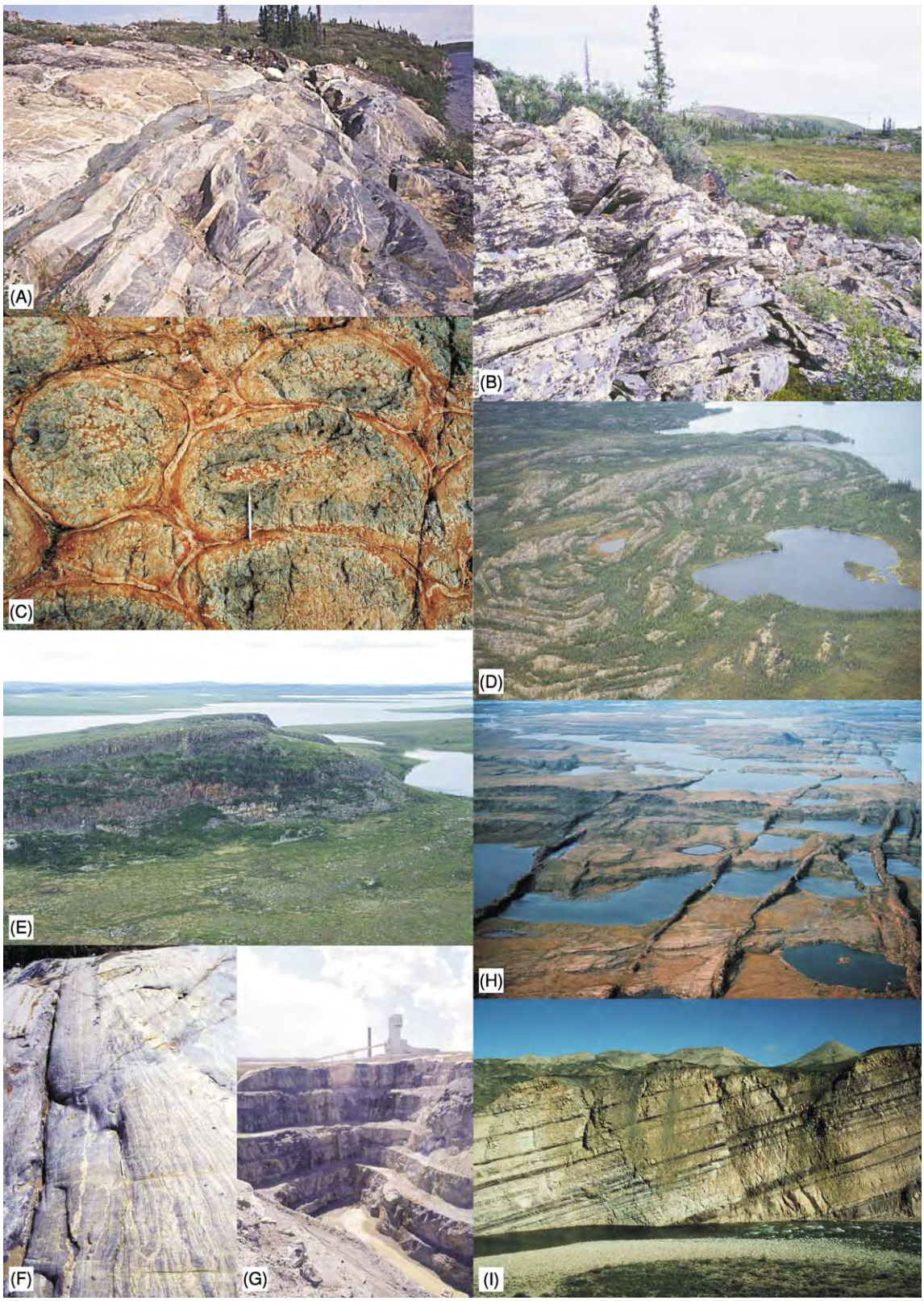


Figure 2 Field photos of representative elements of the Precambrian geology of North America. (A) Earth's oldest intact rocks, the Acasta gneisses, along the western edge of the Slave Craton. *Ca*. 4.03 Ga gabbros and tonalitic gneisses cut by 3.6 Ga isoclinally folded granite sheets and younger mafic dykes. (B) *Ca*. 2.8 Ga fuchsitic quartzites overlying ancient basement gneisses at the base of the Neoarchaean Yellowknife Supergroup, Slave craton. (C) *Ca*. 2710 Ma variolitic pillow basalts of Fe tholeiitic composition, a typical component of Archaean greenstone belts; Yellowknife greenstone belt, Slave Craton. (D) Multiply deformed, *ca*. 2660 Ma metaturbidites

Figure 3 North America's first diamond mine, the Ekati Mine of the central Slave Craton. Main picture shows the flat barren lands of the central Slave Craton, with several pipe like kimberlite bodies being excavated. Inset (upper left) shows a close up of one of the partially excavated pipes. Several mm size gem quality diamonds are shown on upper right (photos provided by David Snyder and Grant Lockhart).

deformation and modification of the Precambrian crust and lithosphere, which locally extends up to 1000 km inboard of the margin.

The eastern margin of the North American plate is the slow-spreading mid-Atlantic ridge, which runs north through Iceland and to the east of Greenland into the Arctic basin. Hence, Greenland, almost entirely underlain by Precambrian crust, is geologically part of North America and, more specifically, part of its Precambrian core, Laurentia. With progressive northward opening of the North Atlantic rift, Greenland started to rift apart from the main part of Laurentia at about 80 Ma, opening up the Labrador Sea. This rifting was subsequently aborted in favour of propagation of the mid-Atlantic ridge to east of Greenland, along the present trace of the ridge. Here, ridge propagation split off a fragment of Precambrian crust from Laurentia, which remained attached to the Eurasian plate and now underlies parts of Scotland, Ireland, and the adjacent continental shelf. This Precambrian sliver, known as the Lewisian Gneiss Complex, is equally part of Laurentia. Other Precambrian crustal elements may have rifted and drifted off Laurentia with the opening of a proto-Arctic basin (i.e., prior to the Innutian Orogeny, see **Figure 1**), but here relationships are presently less well understood.

The stable Precambrian lithosphere underlying much of North America responds to far-field boundary forces by gentle flexing and arching. As a consequence, Precambrian crust in the north-central parts of the continent is gently arched up, giving rise to the physiographic province of the Canadian Shield. The latter is characterized by scattered rocky outcrops of generally low relief (**Figure 2**), although, locally, rift

of the Burwash Formation, Slave Craton. Turbidites such as these are typical for the uppermost units of Archaean greenstone belts. Here they have been folded into regional mushroom interference patterns by two upright fold generations. (E) Shallow dipping strata of the *ca.* 2.0 Ga Coronation Supergroup onlapping onto peneplaned Archaean basement (flat foreground), western margin of Slave Craton. These strata represent some of the first true passive margin sequences in the geological record. (F) Archaean Palaeoproterozoic unconformity, metamorphosed to upper amphibolite facies, along the highly deformed western margin of the Superior Craton. Archaean gneissic layering (right) truncated by primary layering and foliation in basal quartzitic schists of the *ca.* 2.0 Ga Ospwagan Group (left). (G) Thompson Mine, exploiting one of the largest komatiite associated nickel deposits in the world, hosted by the highly deformed Palaeoproterozoic Ospwagan Group, along the western margin of the Superior Craton. (H) Proterozoic mafic dykes of the 1267 Ma giant radiating Mackenzie swarm, northern Canadian Shield. (I) Tilted strata of Proterozoic fluvial sandstones overlying the *ca.* 1.8 Ga crustal collage of Laurentia; 100 m high cliffs along the Brock River, North west Territories. Deposits such as these form widespread overlap assemblages as early as 1.75 Ga.

flank uplift has created significant new relief, exposing up to 1000 m-high sections through Precambrian rocks, particularly along the coasts of Labrador, Baffin Island, and Greenland. Minor rejuvenation of relief is currently taking place along arctic coastlines in response to post-glacial rebound. Elsewhere, peneplaned Precambrian basement dips gently beneath a thin Phanerozoic sedimentary cover, giving rise to interior platforms (*see* **North America:** Continental Interior). Presence or absence, and thickness, of this sedimentary cover was, of course, further modulated by changes in global sea-levels, either expanding or shrinking the extent of interior seaways.

In a number of localities, subsidence of the Precambrian lithosphere has been more pronounced, resulting in downwarped intra-continental basins filled with thicker sedimentary successions. One example is the Hudson Bay Basin in the middle of the Canadian Shield, which has been subsiding intermittently during the Phanerozoic. The processes responsible for this intermittent subsidence are poorly understood. Much better understood is the Western Canada Basin, a broad foreland basin forming in response to downflexing of Precambrian lithosphere in advance of the marginal load of the North American Cordillera.

In addition to the broad exposure of the Canadian Shield, crust of Precambrian age is also exposed in more restricted tectonic or plateau uplifts within the Cordilleran and Appalachian-Alleghanian mountain chains that fringe Laurentia. Important examples are the Laramide uplifts in the western United States (e.g., the Wind River Ranges in Wyoming) (*see* **North America:** Southern Cordillera), thrust sheets in the Rocky Mountains, and domal 'core complexes' throughout the internal Cordillera. Deep incision of the Colorado River in the uplifted Colorado Plateau also exposes an important window of Precambrian crust. The Blue Ridge Mountains of the south-eastern United States and the Long Range inlier of Newfoundland are examples of major thrust-related uplifts of Precambrian basement of Grenvillian age in the Appalachians (*see* **North America:** Northern Appalachians). In general, the subsurface edge of Precambrian basement can be traced with isotopic ratios of younger granitoids. An example is the $^{87}Sr/^{86}Sr$-isotopic line of the Cordillera (Figure 1).

Precambrian Nucleus of North America: General Structure

The basic architecture of the Precambrian nucleus of North America is that of a 'collage' comprising crustal elements of different ages (Figure 1). Many of the crustal elements involved in this collage were exotic with respect to each other, having originated in different places around the globe before being amalgamated. Other elements represent crustal growth along the margins of previously assembled continental crust. In places, the crustal collage is truncated by younger rifted margins, providing evidence that, originally, the collage must have extended well beyond its currently preserved margins.

Although the internal structure of the collage is complex, the complexity is far from random. In some ways, the overall architecture can be compared with the 'nested' structure of 'Russian dolls' (Figure 4):

- North America, in its present form, is a fragment of the latest supercontinent in Earth history, Pangaea (Figure 5) (*see* **Pangaea**), which existed from about 300 to 200 Ma and is currently still in the process of progressive breakup and dispersal, although a number of plates have already started their collisional history (e.g., India with Eurasia).
- The North American fragment of Pangaea hosts within itself a large fragment of the earlier supercontinent Rodinia (Figure 4), which was amalgamated from 1200 to 980 Ma. As discussed earlier, this fragment of Precambrian crust, several thousand kilometres in diameter, is commonly referred to as Laurentia.
- On a smaller scale, Laurentia hosts within itself 6–7 large fragments of Archaean (>2.5 Ga) crust, the Archaean 'cratons' (*see* **Precambrian:** Overview). Among these, the Superior, Wyoming, and Slave cratons are some of the better-known examples. These crustal fragments are called cratons because they show long-term stability, having been affected by younger deformation only around their edges. The typical length scale of these Archaean cratons is *ca.* 1000 km. The Superior Craton is the largest preserved Archaean craton on Earth.
- Furthermore, the larger Archaean cratons are typically composite, consisting of a number of domains with disparate crust formation ages. Included among these domains are ancient crustal fragments that are dominated by gneissic granitoids with ages of crystallization or inheritance of >2.9–3.0 Ga. Such ancient gneiss domains (e.g., the Central Slave Basement Complex of the Slave Craton; or the North Caribou Terrane of the Superior Craton; see Figure 4) have typical length scales of 100–300 km.
- Finally, these ancient gneiss domains are themselves heterogeneous. Embedded within them are found Earth's oldest rocks, including, for instance, the 4.03 Ga Acasta gneisses of the Slave Craton (Figures 2A and 4). Individual examples of these

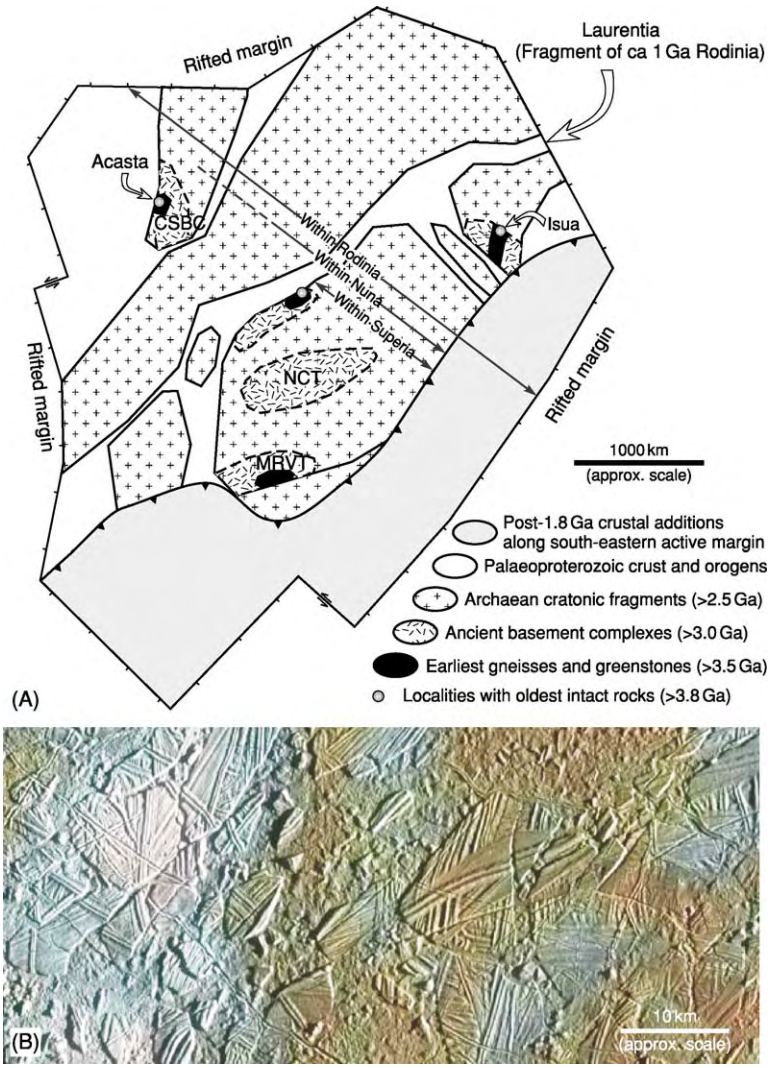


Figure 4 (A) Cartoon illustrating the systematic fractal architecture of the Precambrian crust. The Precambrian record consists of a collage of 'nested' fragments across a variety of scales. At the largest scale, Laurentia itself represents a rifted fragment of Rodinia, a *ca* 1 Ga supercontinent. Its margins formed between 780 600 Ma, when Rodinia started to break up. At the smallest scale, we find individual gneiss complexes consisting of some the oldest intact rocks on Earth (e.g., Acasta, Isua), nested within larger Archaean cratons. The overall fractal structure is one of preservation and reflects repeated cycles of fragmentation and re aggregation, with smaller older fragments preserved in larger younger fragments. Abbreviations: CSBC, Central Slave Basement Complex, Slave Craton; NCT, North Caribou Terrane, Superior Craton; MRVT, Minnesota River Valley Terrane. (B) An identical fractal pattern of nested fragments, but at smaller scales, is seen in this picture of the icy crust of Europa, one of Jupiter's large moons (source: NASA, Galileo Orbiter PIA01127). Linear features are ridges formed by upwelling and subsequent freezing of water along cracks.

Early Archaean gneiss complexes are typically preserved at the 1–10 km scale, in some cases ranging up to *ca*. 100 km.

Hence, the structure of the Precambrian nucleus of North America, and that of the geological record in general, is self-similar (fractal) in nature, repeating a basic motif across a variety of scales, ranging from the size of modern plates to that of individual ancient gneiss complexes or the greenstones belts embedded within them. As plate tectonics is the dominant process in shaping this pattern on the modern Earth, it is tempting to conclude that plate tectonics must have been equally dominant since at least *ca*. 4.0 Ga, the time of the Earth's oldest preserved crust (Figure 6). One might also conclude that plates have grown in size over time. These conclusions seem reasonable but there are a number of important caveats.

The similarity in basic pattern strongly suggests repeated fragmentation and dispersal events of pre-existing crust, with fragments becoming incorporated in new crustal collages through time. However, similarity in pattern and implied kinematic processes

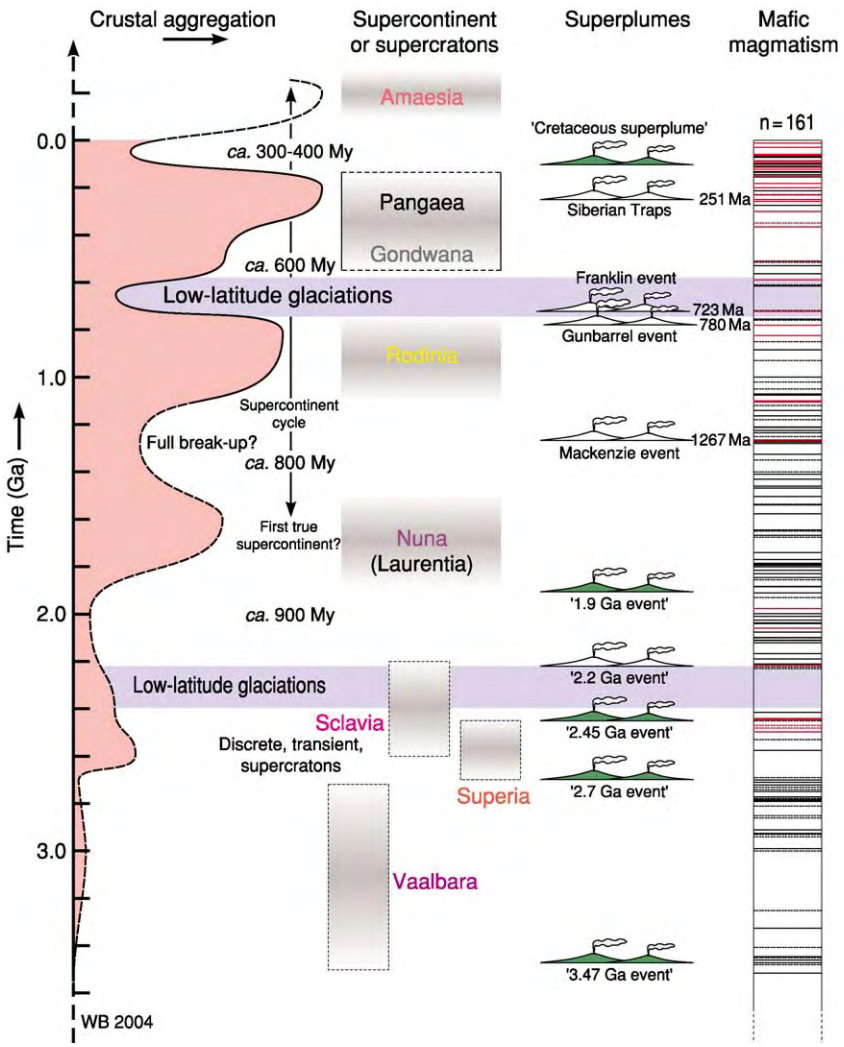


Figure 5 The history of crustal aggregation states (supercratons, supercontinents) through time (vertical axis; modified after Bleeker, 2003). Mid Proterozoic Nuna, including the large 1.8 Ga core of Laurentia, was probably the first true supercontinent in Earth history. The Late Archaean may have been characterized by several discrete, transient, aggregations referred to as super cratons: Vaalbara, Superia, Sclavia and possibly others. The diachronous break up of these supercratons, in the Palaeoproterozoic, spawned the present ensemble of *ca.* 35 Archaean cratons, which now are variably incorporated into younger crustal collages. Since the assembly of Nuna, the time gaps between successive crustal aggregation maxima appear to have become shorter. Note the correlation of intervals of global glaciation with two periods of continental breakup and dispersal, and with possible minima in the frequency of mafic magmatic events in the continental record (legend for the latter: red line, well established mantle plume event; black line, other mafic magmatic event; dashed line, poorly dated event; modified after Ernst and Buchan (2001).

does not necessitate similarity at a dynamic process level. Modern lithospheric plates tend to be large and relatively rigid, and negative buoyancy of subducting, old and cold oceanic plates ('slab pull') is a major component in the overall force balance driving plate motions. At present, it is controversial whether in a hotter Archaean Earth, with 2–4 times the present heat production and a more substantial fraction of the primordial heat budget still preserved inside the Earth, thinner and smaller lithospheric 'plates' interacted the same way as their modern counterparts. A key question is whether the return flow of oceanic mafic-ultramafic material was by rigid slab subduction (i.e., 'plate tectonics') or whether residues, after tonalitic melt extraction, descended back into the convecting mantle by more disorganized gravitational sinking ('drip tectonics')?

Attempts to answer this question from the geological record, as preserved in Archaean cratons, leaves little doubt that the cratons preserve a record of interaction between different crustal fragments. Nevertheless, Archaean cratons lack a fully diagnostic set of criteria confirming modern plate behaviour (e.g., passive margins, flexural basins, accretionary prisms, uncontested ophiolites, high pressure-low temperature metamorphic rocks and

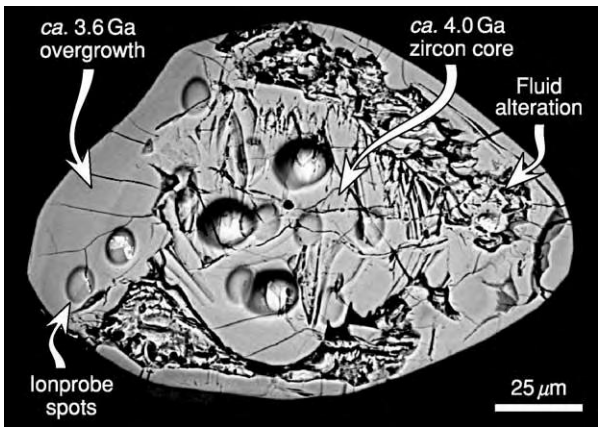


Figure 6 Backscatter electron image of a *ca.* 4.0 Ga zircon crystal from the Acasta gneisses, Earth's oldest intact rocks (see also **Figure 2A**). Metamorphic rims formed at *ca.* 3.6 Ga. Oval spots are scars from the ion probe beam (from Stern and Bleeker, 1998).

paired metamorphic belts, etc.). Hence, it is useful to consider alternative explanations. In doing so, it is also instructive to consider other planetary bodies and their unique pathways of lithospheric evolution. For a comparison with the fractal pattern of repeated fragmentation and amalgamation in the Earth's crust, the icy crust of Europa, one of Jupiter's large moons (**Figure 4B**), is of particular interest (*see* **Solar System: Jupiter, Saturn and Their Moons**). Europa's crust of water-ice, overlying an interior ocean, shows a fractal pattern that is essentially similar to that of Earth's crust, requiring repeated fragmentation events, upwelling and freezing of water along cracks, and lateral mobility of fragments. Yet it is not the product of plate tectonics driven by rigid slab subduction of dense silicate rocks.

Both on Europa and Earth, the fractal pattern of smaller, older fragments embedded in younger crustal collages, is an artefact of the process by which fragments are created and preserved. With each repeated breakup event, crustal fragments can only become smaller, while their number will increase. On the other hand, erosion, tectonic slivering, and partial subduction will, over time, reduce the number of fragments or hide them in the lower crust. Furthermore, a fragment of a certain size and age can only be preserved in a younger fragment that is either the same size or larger. Hence, the negative correlation between fragment size and age, as illustrated in **Figure 4**, is primarily a function of how continental crust is preserved, and may not provide information on whether continental plates have grown in average size over geological time.

Nevertheless, it is intriguing that few of the Archaean cratons come close in size to even the smallest of the major continental plates today. The Superior Craton of North America, the largest extant Archaean Craton, with a surface area of *ca.* $1.57 \times 10^6\,km^2$, is about half the size of the Arabian Plate. Indeed, several independent lines of evidence suggest that Archaean continental aggregations may have been smaller and more transient than their modern counterparts:

- A hotter Archaean Earth, with a substantially higher radiogenic element budget, favours smaller and faster plates to dissipate the increased heat production and avoid catastrophic heating of the planetary interior.
- The record of $^{87}Sr/^{86}Sr$ isotopic compositions of Precambrian seawater, as measured from marine carbonates, shows that prior to 2.6–2.5 Ga, Sr ratios were buffered by dominant interaction with oceanic crust and young mantle-derived mafic rocks. Only after *ca.* 2.5 Ga do we see a rapid rise in the $^{87}Sr/^{86}Sr$ ratios, reflecting more significant input from weathering of aged continental crust.
- Much of the modern detrital sediment load to sedimentary basins and continental shelf-slope systems is provided by large river systems draining large continental hinterlands. The fact that there are few examples of large sedimentary basins of a shelf-slope affinity preserved in the Archaean record, suggests that river systems and their continental hinterlands were smaller, and that large, emerged continental plates were rare or absent.
- As attested by Laurentia, once large cratonic landmasses have aggregated and become underlain by a lithospheric keel, they tend to resist breakup. So, as noted above, the observation that none of the Archaean fragments, with their lithospheric keels, approaches modern continents in size may be significant after all.
- Characteristics of the supercontinent cycle (**Figure 5**) suggest that the successive time lags between breakup and dispersal of one supercontinent (e.g., Rodinia) and the re-aggregation of a subsequent supercontinent (e.g., Pangaea) have become shorter over the last 2.5 billion years. The easiest explanation for this observation is that, following supercontinent breakup and dispersal, modern continental plates, due to their large average size, quickly run out of room and start colliding. More substantial time lags in the past suggest smaller continental plates.
- And finally, Earth's present rate of heat loss is about twice the rate of internal heat production, while at the same time heat production (mainly from K, Th, U) is undergoing slow exponential decay. Thermal arguments predict therefore that,

over time, the lithosphere will slowly thicken and stiffen, evolving ultimately to a relatively immobile 'one-plate' state. This state has already been reached by our closest sister planet, Venus, which is only marginally smaller than Earth (Venus' radius is 95% that of Earth).

Precambrian Nucleus of North America: A Systematic Overview

From 2.0 Ga to 1.8 Ga, over a time-span of about 200 million years, the core of Laurentia formed by progressive amalgamation of 6–7 large fragments of Archaean crust and intervening island arcs into a broad orogenic collage (Figure 1). Crustal shortening and thickening accompanying the various accretion and collision events led to uplift and exhumation and, ultimately, stabilization of the newly formed crust, as indicated by the initiation of widespread intracontinental sedimentary basins that overstep the boundaries between different crustal blocks as early as *ca* 1.75 Ga. Interestingly, the time-frame of relatively rapid crustal aggregation and growth implied for Laurentia is of a similar order to that for the progressive development of the most recent supercontinent, Pangaea, which started with assembly of the large southern continent Gondwana in the latest Proterozoic and terminated with final collision of Gondwana and Laurasia in the Carboniferous to Permian (i.e., an overall time-frame of *ca*. 250 Ma).

The widespread 2.0–1.8 Ga orogenic events in Laurentia, known in North America as the Hudsonian Orogeny, have counterparts in most other continents (e.g., the Amazonian Orogeny in Brazil; the Svecofennian Orogeny of the Baltic Shield; the Capricorn Orogeny of Western Australia). Through the collective amalgamation of a large number of disparate Archaean cratons or microcontinents into a global collage, this first-order orogenic event may have led to the first true supercontinent in Earth history at about 1.75 Ga named Nuna (Figure 5). Through further crustal growth along its external margins, and final continent–continent collision to form the Grenville Orogen along its southeastern margin (Figure 1), Nuna evolved into *ca*. 1 Ga Rodinia.

Clearly, the Archaean cratons are the dominant building blocks in the Palaeoproterozoic assembly of Laurentia, and the core of Nuna. Hence, they form a logical starting point for a brief systematic overview of the Precambrian crust.

The Archaean Cratons

Individual Archaean cratons that constitute the core of Laurentia are: the Superior, Slave, and Wyoming cratons; the highly reworked Rae and Hearne cratons, collectively known as the Churchill structural province; the enigmatic Sask Craton, which underlies much of the Palaeoproterozoic Trans-Hudson Orogen in central Canada; and the Nain or North Atlantic Craton underlying southern Greenland (Figure 1). These cratons represent a subset out of a total ensemble of about 35 large Archaean crustal fragments preserved around the world.

The independent nature of some of the component cratons of Laurentia is still a matter of debate. Clearly, the well-studied Superior and Slave cratons are two pieces of crust that are exotic relative to each other and most likely originated from unrelated ancestral landmasses – the Late Archaean supercratons Superia and Sclavia, respectively (Figure 5). The ancestry of some of the other cratons is less clear.

A putative suture between the Hearne and Rae cratons, the Snowbird Tectonic Zone (Figure 1), remains controversial. Hence, it is possible that the Hearne and Rae cratons represent a contiguous fragment of Archaean crust. The Hearne Craton is locally overlain by an Early Proterozoic cover sequence, the *ca*. 2.4–2.1 Ga Hurwitz Group, which shows similarities to the Huronian Supergroup overlying the southern Superior Craton. Both sequences contain evidence for *ca*. 2.3 Ga low-latitude glaciations. Futhermore, both cratons are cross-cut by *ca*. 2450 Ma mafic dykes. Palaeomagnetic data from these dykes suggest that the Hearne Craton may have originated, at 2450 Ma, from just south of the 'Huronian margin' of the Superior Craton (Figure 7). Thus, both cratons likely originated from within supercraton Superia, possibly along with others such as the Karelia, Nain and Yilgarn cratons.

Whether the Wyoming Craton, underlying much of the north-western United States, is truly distinct or just a southern continuation of the Hearne Craton also remains unclear. The proposed suture between the Hearne and Wyoming cratons is buried underneath thick platform strata. The Sask Craton may be an exotic piece of crust or, alternatively, a partially detached piece of either the neighbouring Hearne or Superior Craton, incorporated in the Trans-Hudson Orogen. Interestingly, many of these fundamental questions can be resolved, in principle, with detailed palaeomagnetic studies and high-precision age dating of the Palaeoproterozoic mafic dyke swarms that cut the cratons, in conjunction with stratigraphic comparisons of the Archaean cratons and the Palaeoproterozoic cover sequences that overlie them.

The Slave Craton

As an example of the Archaean components of Laurentia, the Slave Craton is described in more detail.

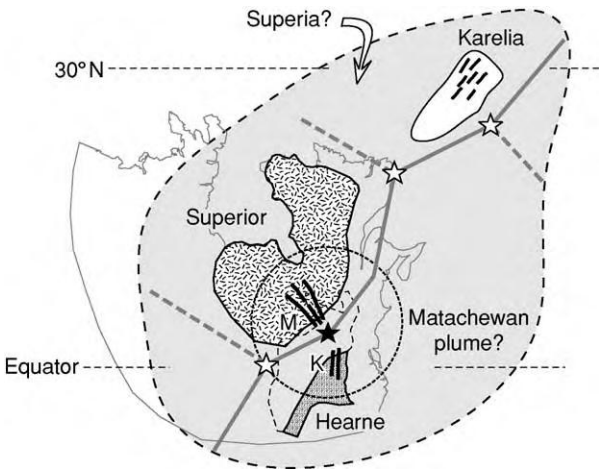


Figure 7 Possible reconstruction of the relative position of the Superior and Hearne crustal fragments at *ca.* 2450 Ma, prior to the inferred break up of their ancestral supercraton, Superia. Both cratons are cut by *ca.* 2450 Ma mafic dykes that are petrographi cally similar, the Kaminak dykes (K) of the northern Hearne and the Matachewan dykes (M) of the southern Superior, respect ively. Palaeomagnetic data from both dyke swarms allow the reconstruction shown here. Both dyke swarms and the inferred break up of Superia may have been consequences of the *ca* 2450 Ma Matachewan plume (black star). Karelia, another Ar chaean craton now embedded in the Baltic Shield, also hosts *ca* 2450 Ma mafic dykes and, thus, may have been part of Superia. Also shown is a possible topology of intracontinental rift zones along which supercraton Superia broke up, liberating individual cratons. Plumes or hotspots along this rift zone (open stars) may have had ages spreading out over a 100 200 million year interval, similar to those along the mid Atlantic rift during progressive break up of Pangaea.

This well-exposed fragment of Archaean crust measures about 500 (E–W) by 700 (N–S) km and is surrounded by orogens of Palaeoproterozoic or younger age (Figures 1 and 8).

The west-central part of the craton is underlain by an ancient gneiss complex that hosts some of the oldest intact rocks in the world, the *ca.* 4.03 Ga Acasta Gneisses (Figure 2A). This basement complex is overlain by a thin, *ca.* 2.8 Ga quartzite and banded iron formation sequence (Figures 2B and 8), which in turn is overlain by the main cycle of *ca.* 2.73–2.62 Ga volcanic and sedimentary rocks (Figures 2C,D and 8). Very similar stratigraphic relationships are seen in other cratons of Laurentia and around the world.

The Neoarchaean supracrustal rocks of the Slave Craton, known as the Yellowknife Supergroup, were intruded by voluminous granitoid rocks that range in age and affinity from *ca.* 2.71–2.67 Ga, synvolcanic, tonalite-granodiorite plutons to *ca.* 2.58 Ga, post-tectonic, granite plutons. The supracrustal and pre-tectonic granitoid rocks are deformed by several discrete folding events, resulting in complex fold interference structures (Figure 2D) and high-amplitude basement-cored domes. The overall result is an archetypical Archaean 'granite-greenstone terrane'.

The younger granites are part of a craton-wide, *ca.* 2.59–2.58 Ga 'granite bloom', which transported heat, partial melt fractions, and aqueous fluids, as well as most of the incompatible heat producing elements (K, U, Th), from the hot lower crust to the upper crust. This critical and irreversible step allowed the lower crust to cool and to mechanically couple with the mantle lithosphere. Shortly following this terminal granite bloom, Slave crust stabilized and became 'cratonic'.

Following cratonization, from 2.58 Ga to 2.2–2.0 Ga, Slave crust was gradually exhumed by *ca.* 10–15 km (on average) through slow uplift and erosion. This is indicated by the onlap of unconformably overlying Palaeoproterozoic cover sequences onto exhumed Archaean basement (Figure 2E). Starting at about 2230 Ma, the Slave Craton, still within the larger context of its host supercraton Sclavia, was intruded and weakened by numerous mafic dyke swarms and anorogenic alkaline complexes. These events eventually led to rifting and breakup of Sclavia and independent drift of the Slave Craton as a small microcontinent. In general, these Early Proterozoic rifting and dispersal events resulted in some of the first true sediment-rich passive margin sequences in the geological record (e.g., the Coronation Supergroup overlying the western margin of the Slave Craton, Figure 2E).

Proterozoic Orogens: the Glue of the Laurentian Collage

The oldest Palaeoproterozoic orogen within Laurentia is the Taltson-Thelon Orogen along the western margin of the Rae Craton. It is marked by a *ca.* 2.02–1.95 Ga magmatic arc, with a conspicuous aeromagnetic signature that allows it to be traced for several thousand kilometres (Figure 1). This magmatic arc suggests significant consumption of oceanic lithosphere below the western edge of the Rae Craton, leading to collision and eventual suturing of the Slave Craton to the Rae. Partial subduction of Slave crust below the Rae led to flexural subsidence of the Slave and its overlying platform sequence, resulting in one of the best examples of a Palaeoproterozoic foreland trough—the Kilihigok Basin. The short flexural wavelength of this basin suggests a low flexural rigidity of Slave lithosphere at 1.9 Ga, and hence argues against a thick lithospheric keel at that time.

During the Slave–Rae collision, subduction is also active on the western side of the Slave Craton, underneath an outboard Hottah Arc. This arc accreted to

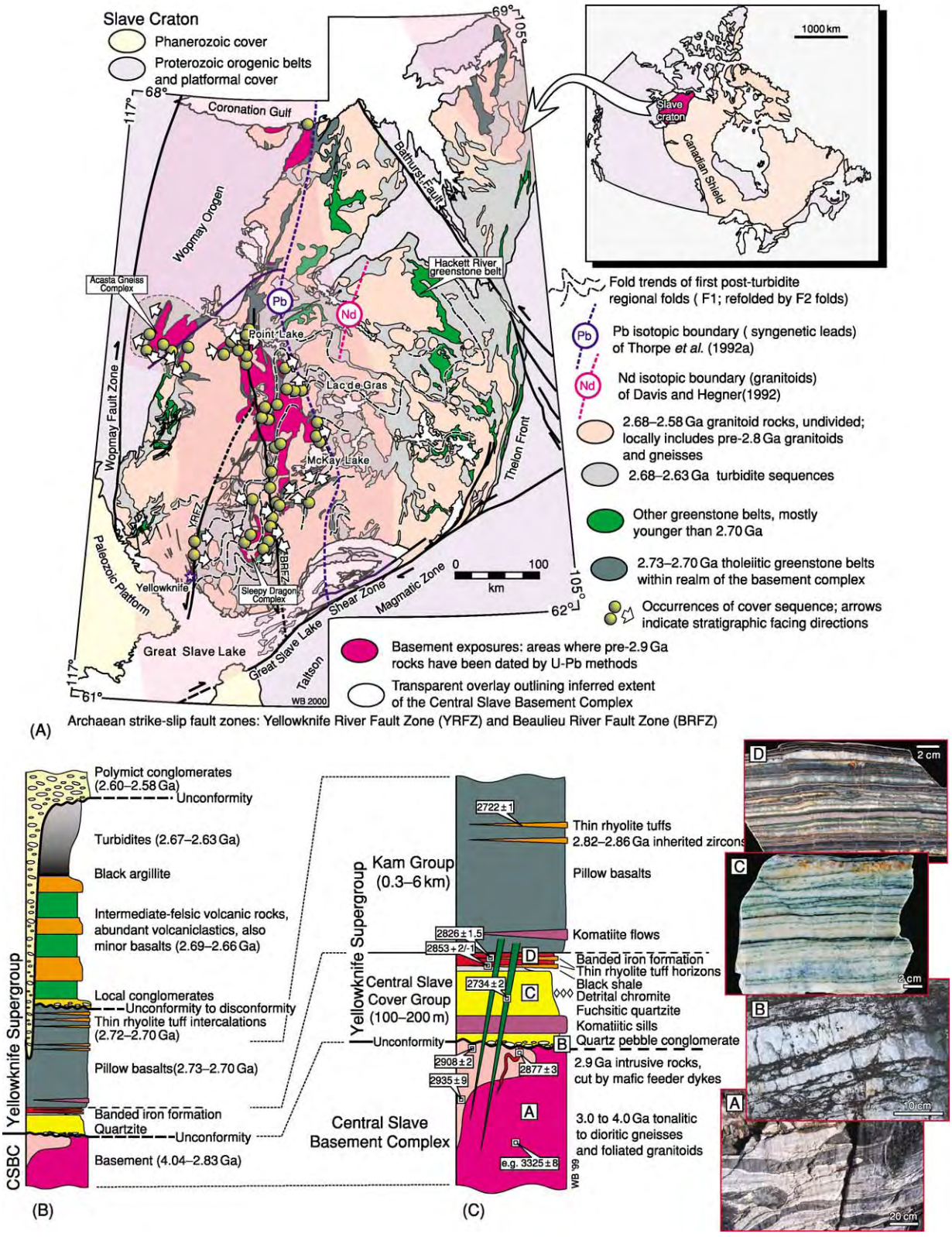


Figure 8 Geology and generalized stratigraphy of the Slave Craton. (A) Geological map; transparent overlay (light purple) outlines minimum extent of the Mesoarchaean to Hadean basement complex of the Slave Craton. (B) Generalized stratigraphy for large parts of the craton. (C) Typical stratigraphy immediately overlying the Mesoarchaean basement complex; photos A D illustrate different units in the section.

the Slave Craton, followed by a subduction polarity flip and building of a new arc on the Hottah–Slave Suture at *ca*. 1.88 to 1.86 Ga.

Elsewhere, the eastern margin of the Rae-Hearne block (already assembled or always contiguous?) started to interact with outboard island arcs, representing the early stages of the Trans-Hudson Orogen (Figure 1). Similar processes were occurring around the margins of the other cratons, some still at remote locations (e.g., the onset of arc accretion along the southern margin of the Superior Craton, initiating the Penokean Orogen). Then, in a spectacular climax of orogenic activity, between about 1850 Ma and 1800 Ma, the remaining cratons gradually aggregated, collided, and amalgamated. This final stage involved the putative closure of the wide Manikewan Ocean between the Rae-Hearne and Superior cratons to form the >3000 km-long Trans-Hudson Orogen through the centre of the Canadian Shield. Terminal collision in Trans-Hudson Orogen (Figure 9) involved northward indentation of the large Superior Craton into a northern hinterland of the Rae-Nain-Hearne-Slave Collage. This northward indentation caused oroclinal bending of the hinterland and tectonic escape of crustal blocks along large strike-slip faults. The northward movement also caused late-stage transcurrent motion on the lateral margins of the Superior indentor (Figure 1).

Post-1.8 Ga Growth and Modification of Laurentia

Following the climactic growth of Laurentia at *ca*. 1.8 Ga, orogenic activity transferred to the external margins of Laurentia, in the west and the south-east. A complete assessment of growth along the western margin needs to involve a western landmass (most likely a combined Australia-Antarctica continent and its amalgamation history), which has since rifted off.

In the south-east, 1.8 Ga orogenic activity must have left major tracts of oceanic lithosphere along the 'trailing edges' of the plates that carried the Archaean cratons into the Laurentian collage. This oceanic lithosphere was subsequently consumed. A number of successive arcs formed, either as outboard

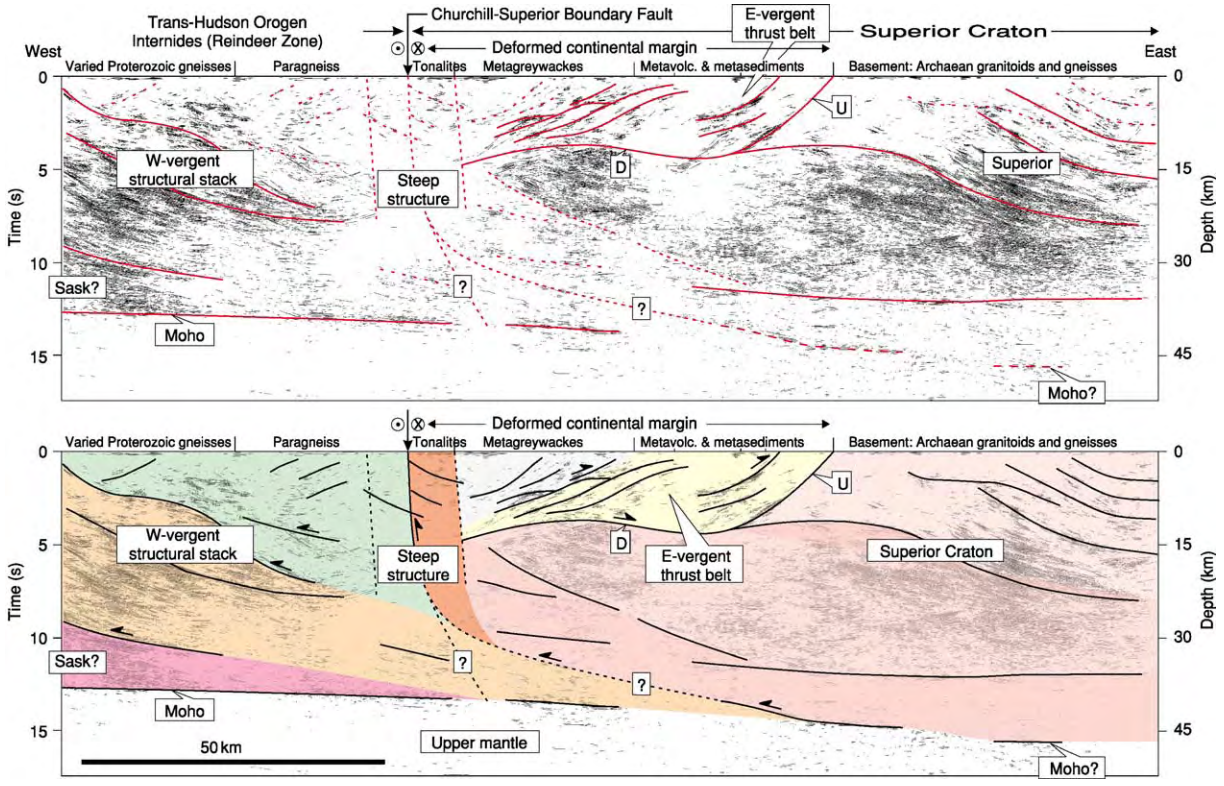


Figure 9 Representative seismic reflection profile (upper diagram) and interpreted cross section (lower diagram) through one of the major *ca*. 1.8 Ga sutures of Laurentia, the 'Churchill Superior Boundary Zone', along the western margin of the Superior Craton (see Figure 1 for location of profile). Early deformation imbricated the continental margin of the Superior Craton, including the *ca*. 2.0 Ga Ospwagan Group (Figure 2F and G) into an E vergent fold thrust belt. During terminal collision, vergence changed and the buoyant Superior margin started to override juvenile Palaeoproterozoic rocks of the internal zone of the Trans Hudson Orogen. Oblique convergence degenerated into late stage sinistral strike slip deformation along the steep Churchill Superior Boundary Fault (modified after White *et al.*, 2002; see also Bleeker, 1990).

island arcs or, at other times, as magmatic arcs on the growing margin of Laurentia. These long-lived active margin processes led to progressive outward growth of and accretion to Laurentia, culminating with the final collisional history of the *ca*. 1.19–0.98 Ga, high-grade, Grenville Orogen (Figure 1).

Significant pulses of outward growth that can be recognized along various parts of the margin are the Yavapai (1.79–1.71 Ga) and Mazatzal (*ca*. 1.70–1.62 Ga) belts of the south-western United States (Figure 1). The latter is broadly coeval with the Labradorian event of north-eastern Laurentia (Figure 1), which involved accetion of a 1.70–1.66 Ga outboard arc to the margin of Laurentia at about 1655 Ma. The Labradorian Arc and a number of younger arcs (e.g., the Pinwarian Arc) were severely reworked during terminal Grenville collisions. The latter involved final suturing of Laurentia to a southeastern landmass (Amazonia?) to form Rodinia.

During its long-lived Proterozoic active margin evolution, lasting from about 1.8 Ga to 1.0 Ga, Laurentia likely faced a 'Pacific-type' ocean to the south-east (present coordinates). Inboard of this margin, several major magmatic events suggest the upwelling of hot mantle underneath the growing continent. Similar processes ultimately led to the rifting and breakup of Laurentia (as part of Rodinia), starting at about 780 Ma. Important examples of mid-Proterozoic igneous events are several pulses of granite-rhyolite magmatism concentrated in south-eastern Laurentia; the intrusion of the 1.3 Ga Nain plutonic suite in Labrador; the 1267 Ma giant Mackenzie dyke swarm; several more localised mafic dyke swarms; and the basalt-rhyolite volcanism of the Keewanawan event in the mid-continent rift (Figure 1).

Glossary

Nuna Earth's first true supercontinent that formed by progressive amalgamation of numerous microplates and intervening island arc terranes towards the end of the Palaeoproterozoic era, from about 1.9 Ga to 1.7 Ga.

Sclavia Late Archaean ancestral landmass of the Slave Craton and allied cratons. This landmass or supercraton amalgamated during Late Archaean orogenic activity, cratonized shortly after 2.6 Ga, and experienced rifting and break-up between 2.2 Ga and 2.0 Ga, spawning several microplates, including its type craton, the Slave Craton of the Canadian Shield. Existence of this late Archaean supercraton is indicated by the Slave Craton being a cratonic fragment surrounded by rifted margins.

Superia Late Archaean ancestral landmass of the Superior Craton and allied cratons. This landmass or supercraton amalgamated during Late Archaean orogenic activity, cratonized shortly between 2.68 Ga and 2.62 Ga, and experienced rifting and break-up starting at 2.48 Ga, spawning numerous microplates, including its type craton, the Superior Craton of the Canadian Shield. Existence of this Late Archaean supercraton is indicated by the Superior craton being a cratonic fragment surrounded by rifted margins.

Supercraton One of several Late Archaean cratonic landmasses, not necessarily a single supercontinent, that on Palaeoproterozoic break-up spawned the present ensemble of *ca*. 35 Archaean cratons with their rifted margins.

See Also

Earth Structure and Origins. **Gondwanaland and Gondwana**. **North America:** Atlantic Margin; Continental Interior; Northern Appalachians; Northern Cordillera; Southern and Central Appalachians; Southern Cordillera. **Pangaea**. **Plate Tectonics**. **Precambrian:** Overview. **Solar System:** Jupiter, Saturn and Their Moons.

Further Reading

Anderson DL (2002) How many plates? *Geology* 30: 411 414.

Bleeker W (1990) New structural metamorphic constraints on Early Proterozoic oblique collision along the Thompson nickel belt, northern Manitoba. In: Lewry JF and Stauffer MR (eds.) The Early Proterozoic Trans Hudson orogen of North America. Geological Association of Canada Special Paper 37, pp. 57 74.

Bleeker W (2002) Archaean tectonics: a review, with illustrations from the Slave craton. In: Fowler CMR, Ebinger CJ, and Hawkesworth CJ (eds.) *The early Earth: physical, chemical and biological development,* Geological Society of London Special Publication 199, pp. 151 181. UK: Bath.

Bleeker W (2003) The late Archean record: a puzzle in ca. 35 pieces. *Lithos* 71(2 4): 99 134.

Buchan KL and Ernst RE (2004) *Diabase dyke swarms and related units in Canada and adjacent regions.* Geological Survey of Canada Map No. 2022A, scale 1:5 000 000, Ottawa, Canada.

Davies GF (1999) *Dynamic Earth: plates, plumes and mantle convection.* Cambridge: Cambridge University Press.

de Wit MJ and Ashwal LD (1997) *Greenstone Belts.* Oxford Monographs on Geology and Geophysics 35, Oxford: Oxford University Press.

Ernst RE and Buchan KL (2001) Large mafic magmatic events through time and links to mantle plume heads. In: Ernst RE and Buchan KL (eds.) *Mantle plumes: their identification through time,* Geological Society of America, Special Paper 352, pp. 483 575. Colorado, Boulder.

Gower CF and Krogh TE (2002) A U Pb geochronological review of the Proterozoic history of the eastern Grenville Province. *Canadian Journal of Earth Sciences* 39: 795 829.
Grotzinger JP and Royden L (1990) Elastic strength of the Slave Craton at 1.9 Gyr and implications for the thermal evolution of the continents. *Nature* 347(6288): 64 66.
Heaman LM (1997) Global mafic magmatism at 2.45 Ga: remnants of an ancient large igneous province? *Geology* 25(4): 299 302.
Hoffman PF (1988) United plates of America, the birth of a craton: Early Proterozoic assembly and growth of Laurentia. *Annual Reviews in Earth and Planetary Science* 16: 543 603.
Hoffman PF (1989) Precambrian geology and tectonic his tory of North America. In: Bally AW and Palmer AR (eds.) *The geology of North America An overview*, The Geology of North America, vol. A, pp. 447 512. Doubler, Colorado: Geological Society of America.
Hoffman PF (1991) Did the breakout of Laurentia turn Gondwanaland inside out? *Science* 252(5011): 1409 1412.
Hoffman PF (1992) Supercontinents. In: Encyclopedia of Earth System Science 4, pp. 323 328. London: Academic Press Inc.
Ross GM, Parrish RR, Villeneuve ME, and Bowring SA (1991) Geophysics and geochronology of the crystalline basement of the Alberta basin, western Canada. *Canadian Journal of Earth Sciences* 28: 512 522.
Stern RA and Bleeker W (1998) Age of the world's oldest rocks refined using Canada's SHRIMP: the Acasta gneiss complex. *Geoscience Canada* 25: 27 31.
Wheeler JO, Hoffman PF, Card KD, *et al.* (1996) *Geological map of Canada.* Geological Survey of Canada 'A' Series Map No. 1860A, scale 1:5 000 000. Ottawa, Canada.
White D, Lucas SB, Bleeker W, Hajnal Z, Lewry JF, and Zwanzig HV (2002) Suture zone geometry along an ir regular Paleoproterozoic margin: the Superior boundary zone, Manitoba, Canada. *Geology* 30(8): 735 738.

Continental Interior

D F Merriam, University of Kansas, Lawrence, KS, USA

> PB King wrote "The [Continental] Interior Lowlands are differentiated from the Laurentian [Canadian] Shield mainly by their retention of a sedimentary cover, so that the tectonic boundary between the two divisions is arbitrary and the differences are not fundamental."
>
> (*The Tectonics of Middle North America*, 1951)

Introduction

The vast country of essentially flat-lying sediments rimmed on three sides by the old Appalachian, Ouachita, and Rocky Mountain mountain chains and to the north the poorly exposed, heavy glaciated, low-lying Canadian Shield (nucleus) composed of Precambrian crystalline igneous and metamorphic rocks, comprises the North American Continental Interior (Figure 1). This immense area, also known as the Central Stable Region, is drained by the mighty Mississippi River and tributaries – and contains the large, relatively shallow, freshwater Great Lakes. This physiography accurately reflects the tectonic provinces (Figure 2). Occasionally, bits of the basement poke through the veneer of this essentially gently-dipping sedimentary cover in the Black Hills of South Dakota, Wichita Mountains of Oklahoma, Ozarks of Missouri, and Baraboo Range of Wisconsin. The shape and form of the sedimentary rocks over the area conform well to the shape and form of the Precambrian basement foundation on which they lie.

Although the topography is subdued and the spectacular scenery is formed for the most part, by erosional remnants or incised valleys, it reflects geological conditions of the region and thus the landforms are a key to understanding recent developments. This part of North America forms the physiographic provinces of the Interior Lowlands and the Great Plains – the bread basket of the continent.

The nucleus of the North American continent is the Precambrian Canadian Shield – the craton, a stable feature around and on which subsequent events took place (*see* **North America:** Precambrian Continental Nucleus). The craton itself had a long and complex history but because of poor exposure and structural complexities it has been difficult to decipher. With a few notable exceptions, the craton is made up of ancient igneous and metamorphic crystalline rocks. The notable exception is the Midcontinent Rift System (MRS), extending from Minnesota to Oklahoma, which is an elongated, younger, faulted, and down-dropped trough containing extrusive igneous rocks interspersed with thick, locally derived, immature sediments (Figure 3). Near the end of the Precambrian, a prolonged period of erosion beveled and smoothed the old fractured/faulted surface prior to it being covered by the younger sedimentary rocks.

Figure 1 Index map of North American physiographic provinces showing location of Interior Lowlands and Great Plains of Continental Interior wedged in between the Appalachian Mountains on the east and the Rocky Mountains on the west. (From Kay and Colbert (1965), John Wiley & Sons.)

This part of the Earth was relatively stable during the next several hundred million years, and as the seas advanced and retreated over the low relief surface, they deposited fairly uniform marine sediments, layer upon layer, over extensive areas. The type and distribution of the sediments responded to the global change in sea-level as well as to changes in local conditions. All of these events were recorded in the sedimentary section, which, as the history unfolded, was sprinkled liberally with long intervals of no record. Because of the sensitivity to conditions in the accumulation of sediments in this type of environment, much of the history is not chronicled; in fact, it has been estimated that as much as 85% of time in any one place is not represented by the rock record. This situation is aptly described as layer-cake geology.

The Foundation

The foundation is the craton, which was formed by adding on (accretion of) material to the original

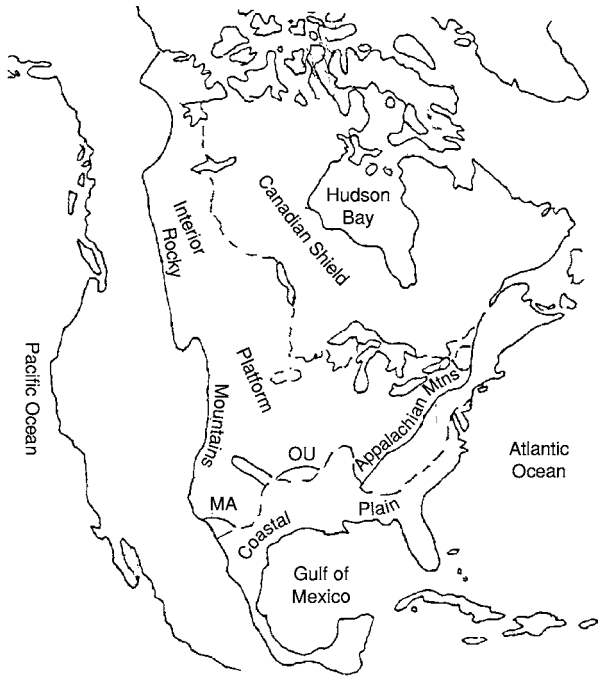


Figure 2 Simplified tectonic map of the North American Continent (MA Marathon Uplift; OU Ouachita Mountains). (Modified from Bally, Scotese and Ross (1989) GSA, The Geology of North America, v. A.) Note physiographic provinces mimic tectonic ones.

primaeval nucleus (Archaean age; 3600 to 2600 Ma) so that the younger Precambrian crystallines (Proterozoic age; 2500 to 900 Ma) rim the nucleus to the south (Figure 3). The Midcontinent Rift System (MRS; 1200 to 1000 Ma) cuts across the older Precambrian terrain. This foundation forms part of the continental crust which ranges in thickness from 40 to 50 km (Figure 4).

In the Continental Interior south of the Canadian Shield, the Precambrian basement complex is exposed in only a few places. Thus, the knowledge of this time period in the Earth's history is known from the thousands of wells drilled in the search for petroleum, water, and mineral resources, plus, of course, geophysical data.

The Sedimentary Veneer

Overlying the Precambrian foundation is the sedimentary veneer, the composition and configuration of these younger sedimentary deposits giving clues to the conditions under which they were deposited and accumulated. Thus, information is contained in the rocks about the location of sea margins, climate, and life at that time.

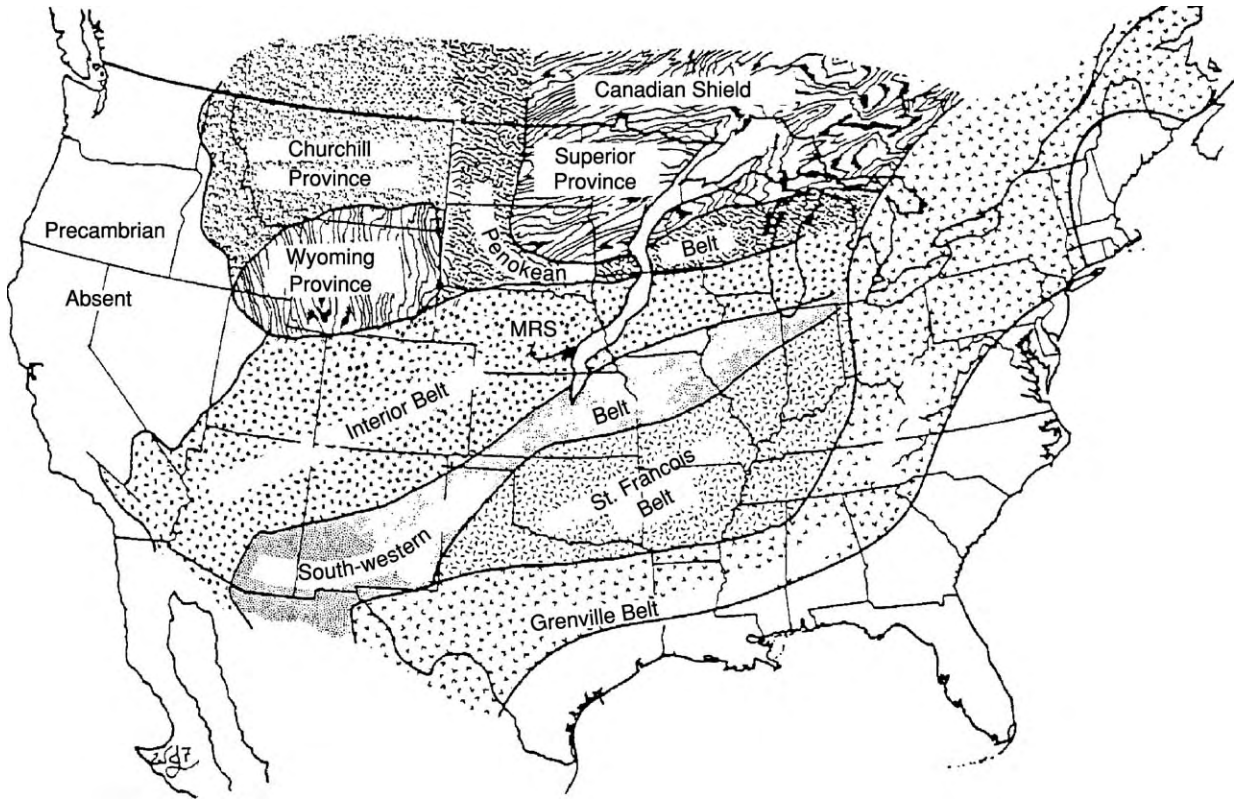


Figure 3 Proterozoic crustal provinces in North America. Archaean: Canadian Shield (>2500 Ma); Proterozoic: Penokean Belt (1900 1820 Ma), Interior Belt (1780 1690 Ma), South western Belt (1680 1610 Ma), St Francois Belt (1480 1380 Ma), Greenville Belt (1000 550 Ma). Midcontinent Rift System (MRS). (Compiled by Van Schmus and Bickford (eds.) (1993) GSA, The Geology of North America, v. C 2.)

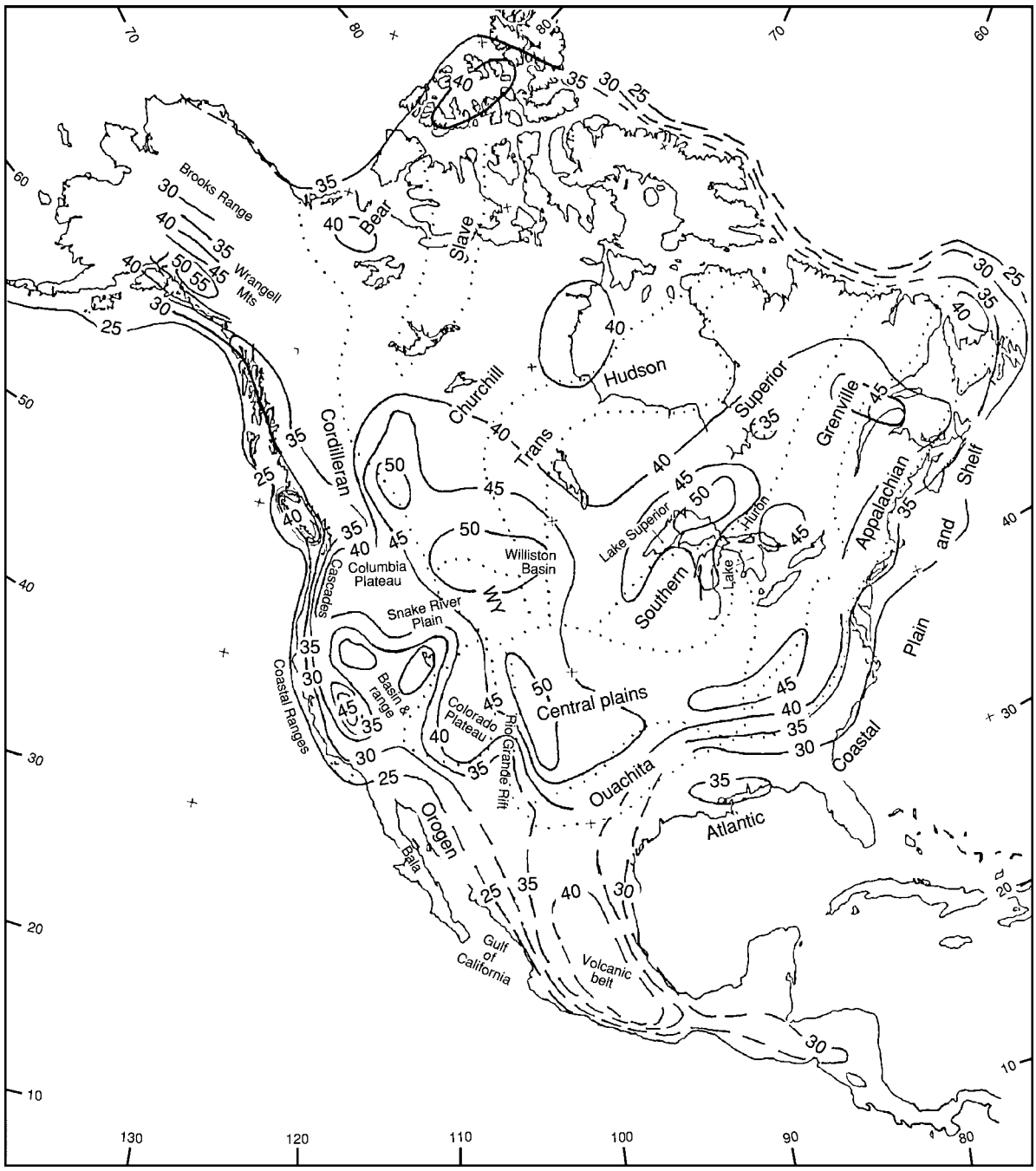


Figure 4 Crustal thickness of North America superposed on simplified geologic province map. Thickness in Continental Interior ranges from 40 to 50 km; Contour Interval = 5 km. (From Houston (1993) GSA, The Geology of North America, v. C 2.)

The original major geological time divisions of the rock sequence were established in Europe and equivalents correlated to North America in the nineteenth Century (Figure 5). It was soon learned, however, that the classic classification of rock units did not always coincide with the major divisions on the craton; these divisions, separated by major gaps in the record, became known as sequences. Sequences (*see* **Sequence Stratigraphy**) are major cycles in the Earth's history recording events on and adjacent to the craton. In a simplistic view, they are records of a significant advance and retreat of the sea lapping on the edge of the craton, ideally starting with a sequence of nonmarine clastics, followed by marine sediments,

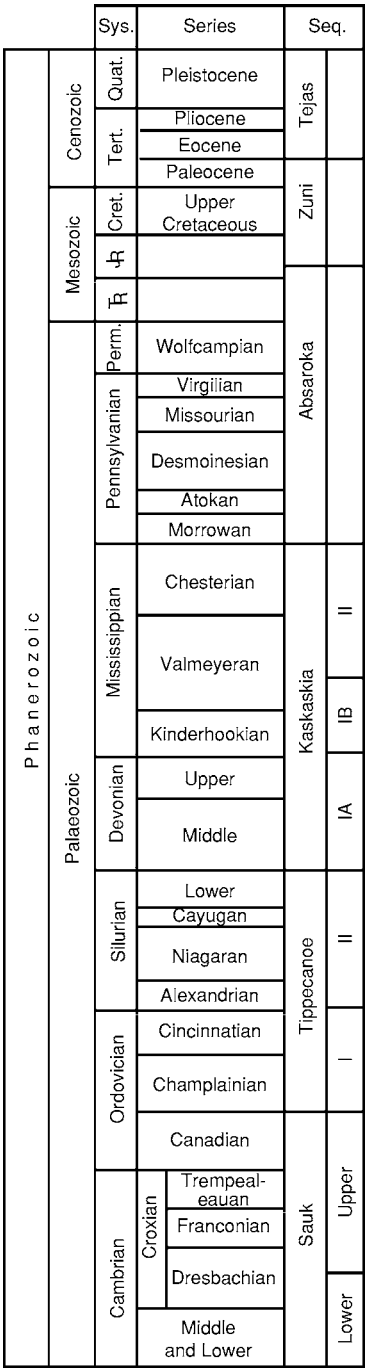


Figure 5 Correlation chart showing relation of sequences and subsequences to standard chronostratigraphic scheme for Phanerozoic. (From Sloss (1988) GSA, The Geology of North America, v. D 2.)

culminating with evaporites and clastics, and ending with a hiatus (gap), then followed by another cycle (Figure 6). All or parts of the sequence may or may not be present at any one location, depending on local conditions.

The craton was never very high above sea-level, a few tens or a couple of hundred metres at most, so that slight changes in sea-level brought inundations or withdrawals of the sea over wide expanses. The seas that spread over the craton, thus, were never very deep, and as the craton gradually foundered, the sediments accumulated. In concert with the rise and fall of sea-level, tectonic forces were busy forming and shaping mountain belts around the edges of the craton. On the craton, the tectonic forces were transmitted through the rigid basement rocks, causing a differential vertical movement of the fractured basement blocks as they readjusted to the stress. A series of uplifts and basins formed as a result of these structural movements and the sediments responded accordingly, filling in the low areas and either not being deposited on the high areas, or deposited and then stripped off. Most of the basins are shallow but some on the margins, such as the Anadarko Basin of Oklahoma and Denver Basin of Colorado, are deeper structural troughs with corresponding thicker sections of sedimentary rocks.

The Phanerozoic Geological Record

As the seas advanced over the low-lying craton, sediments accumulated and thus preserved a record of the history of that episode. The first Phanerozoic sediments on the craton were detrital material derived from erosion of the craton itself. In fact, the exposed part of the craton contributed clastics to the sequence during each episode. The Sauk sequence (latest Proterozoic to Early Ordovician) saw a major inundation as the seas advanced over the craton (Figure 7). The Tippecanoe sequence (Middle Ordovician to Early Devonian) was less widespread and confined mostly to the eastern part of the craton and distribution of the Kaskaskia sequence was similar. However, the early part of the Absaroka sequence (latest Mississippian to Early Permian) was more extensive, and represents a calumination of events, as all but the most northern part of the craton was affected.

Near the end of the Palaeozoic, major changes in the sedimentation and structural history (*see* **North America:** Southern Cordillera) of the central part of North America took place as the result of the Ancestral Rocky Mountain, Ouachita, and Alleghany/Appalachian orogenies (*see* **North America:** Southern and Central Appalachians). The next major inundation of the continent took place in the Zuni sequence, when Cretaceous seas covered much of the craton. Again, major changes in the structural history near the end of the Mesozoic, with the rise of the Rocky Mountains to the west, mark the end to

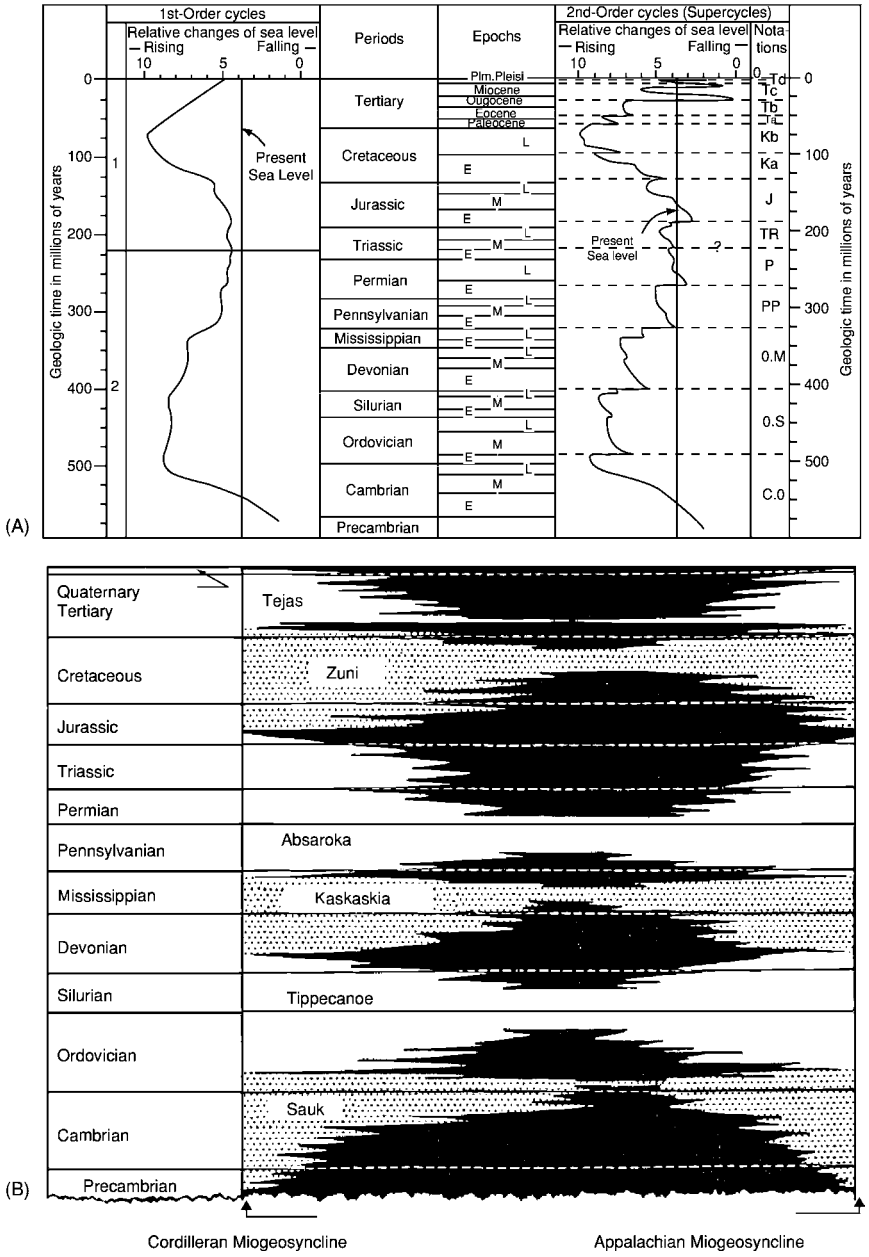


Figure 6 (A) First and second order global cycles of relative change of sea level during Phanerozoic time. (Adapted from (1989) GSA, The Geology of North America, v. A.) (B) Sloss' (1963) diagram of relation of sequences to craton. Black areas are gaps (hiatuses) and white and stippled areas represent deposition of sediments during successive depositional episodes. Note the stratigraphic section is more complete off the craton.

the sea incursions on the continent, so that the Tejas sequence is represented only by terrigenous sediments – material derived from the erosion of the newly-formed Rocky Mountains. This is now the situation remains today – the interior of the continent is high and dry. The end of each episode was marked by withdrawal of the seas and prolific erosion eradicating parts or all of the recorded history of the immediate and preceding chapters.

Several episodes of igneous intrusions into the sediments in the southern part of the Continental Interior took place during the later Cretaceous and Early Tertiary. These small igneous bodies were intruded from the mantle at great depths rapidly along zones of weakness in the continental crust into the sedimentary section. They are of interest, not only for their composition, time of emplacement, and implication for the tectonics, but because some of them, on

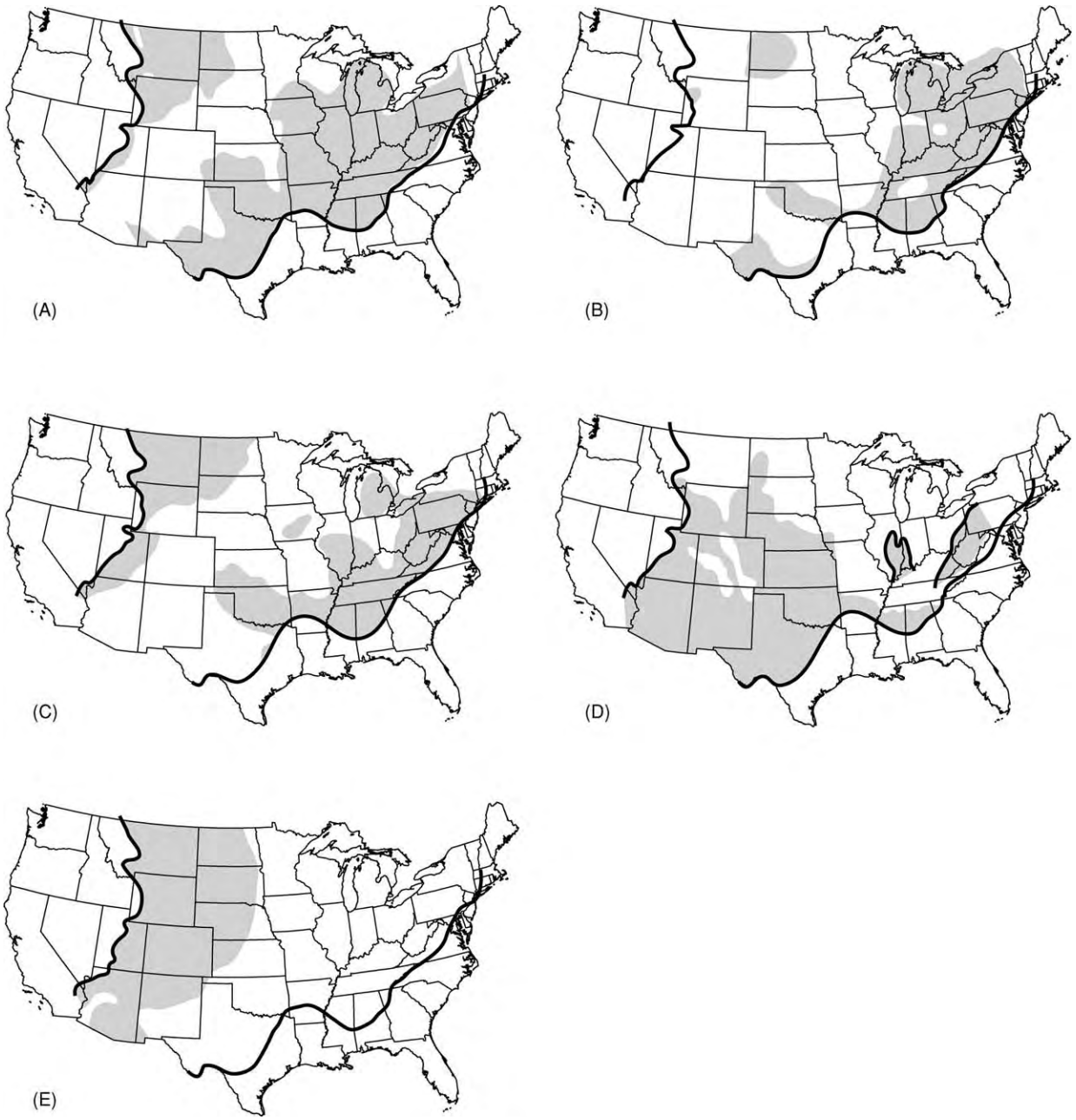


Figure 7 Location of sedimentary sequences preserved on North American craton by sequence. (Modified from Sloss (1988) GSA, The Geology of North America, v. D 2.) (A) Sauk sequence; (B) Tippecanoe sequence; (C) Kaskaskia sequence; (D) Absaroka sequence; and (E) Zuni sequence.

their way up from the depths, incorporated bits and pieces of mantle, continental crust, and Phanerozoic sediments.

The beginning of the last chapter of the story is concerned with the large continental glaciers that covered much of the north-eastern and east-central part of North America (*see* **Tertiary To Present:** Pleistocene and The Ice Age). Numerous advances and retreats of the glaciers are recorded in the remains they left behind – glacial debris (outwash and till; Figure 8A and B) and large erratics (out-of-place foreign boulders; Figure 8C) with thick deposits of loess (wind-blown glacial dust; Figure 8D). In places this forms a cloak over the area, the icing on the layer cake, so to speak.

The glaciers had disrupted the inplace drainage from the end of the Tertiary so that with their retreat, the drainage reoriented to its present

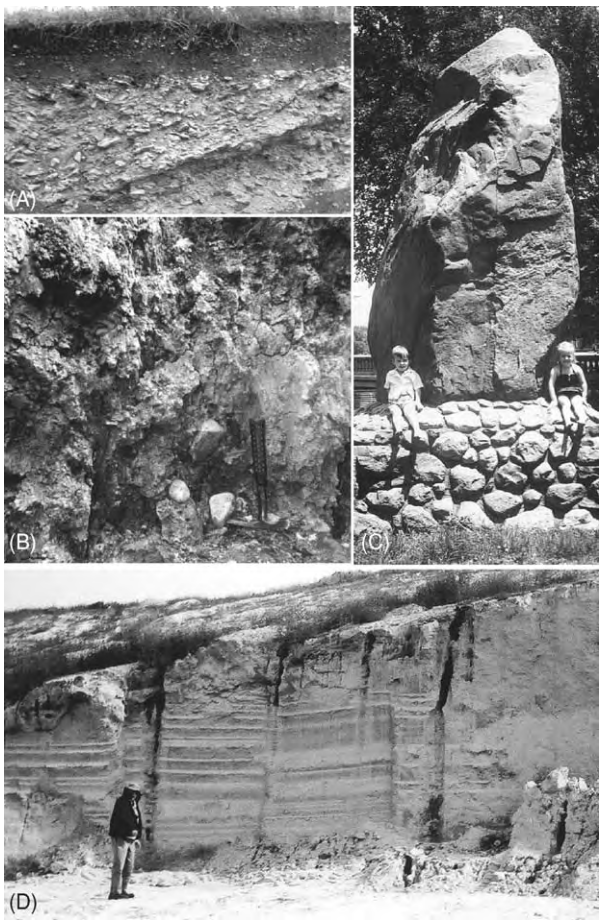


Figure 8 Glacial debris. (A) Glacial outwash deposit, which covers extensive areas of the Continental Interior; (B) glacial gumbo till deposited directly from the glacier; (C) large glacial erratic in Lawrence, Kansas; (D) thick loess deposit (glacial wind blown dust) in central Nebraska.

position. Drainage south of the southern extent of the glaciers was profoundly affected with overloads of water and material, and this latest episode of history is evidenced by the underfit streams, stream piracy, and extensive stream and river terrace development. The final adjustments are taking place today as the Continental Interior returns to 'normal'.

The Sediments

The succession of stratigraphic units on the craton could be and have been described as monotonous, but they are anything but boring. Special features of these unique sedimentary units are their; (i) lateral persistence; (ii) extreme thinness, the tremendous ratio of width to thickness; (iii) sharpness of boundaries; and (iv) cyclic nature.

The contained micro- and macro-fauna and flora (or lack thereof) in these sediments not only gives information on conditions under which the sediments formed, but provides information on relative age for correlation from one area to another, where the sediments are not continuous. Understanding life as it occurred in ancient times is based mainly on understanding the occurrence and distribution of similar forms today.

In addition to the usual sedimentary types, such as sandstone, siltstone, shale, limestone, dolomite, anhydrite, and gypsum, that occur in the Phanerozic sequence, there are four rock types that are formed under special circumstances: chalk, salt, black shale, and coal. Chalk is made up of tiny micro-organisms that accumulated in the open sea in warm water (**Figure 9A and B**). Major chalk deposits throughout the Earth are prominent in the Cretaceous (Zuni) and the Kansas chalk is famous for its marine vertebrate fauna. Salt is formed when and where marine waters are restricted and evaporate (**Figure 9C**). Salt is used here as an all-inclusive term for minerals precipated from the evaporation of sea-water. The Silurian (Tippecanoe) and Permian (Absaroka) are two of those times in Earth history when vast areas of salt were formed on the craton under these conditions.

The black shale accumulates under confined conditions at the bottom of the sea, where because of lack of water circulation and reducing conditions, all forms of carbonate and other soluble material are dissolved (**Figure 9D**). Radioactive material is concentrated in the black shales, making them distinctive subsurface marker beds on gamma-ray wireline logs and easy to correlate. On the North American craton, black shales are prevalent in the Late Devonian to Early Mississippian (Kaskaskia) and numerous thin but persistent ones in the Middle and Upper Pennsylvanian and Lower Permian (Absaroka). Lignite and coal form near sea-level and are thus a good palaeoenvironment indicator (**Figure 10A and B**). In the Interior Province, several basins contain bituminous coal, mostly Pennsylvanian in age, whereas in the Northern Great Plains it is mostly sub-bituminous coal or lignite, and Cretaceous or Tertiary in age. The plant material grew in the shallow water along the shore, dying in place, accumulated, and was buried, eventially forming peat. As the peat was buried, the material compacted to coal and in doing so created a handy indicator to determine the depth of burial. However, it never has been buried deep enough in the Continental Interior, or subjected to tectonic forces, to create the highest grade of coal, anthracite.

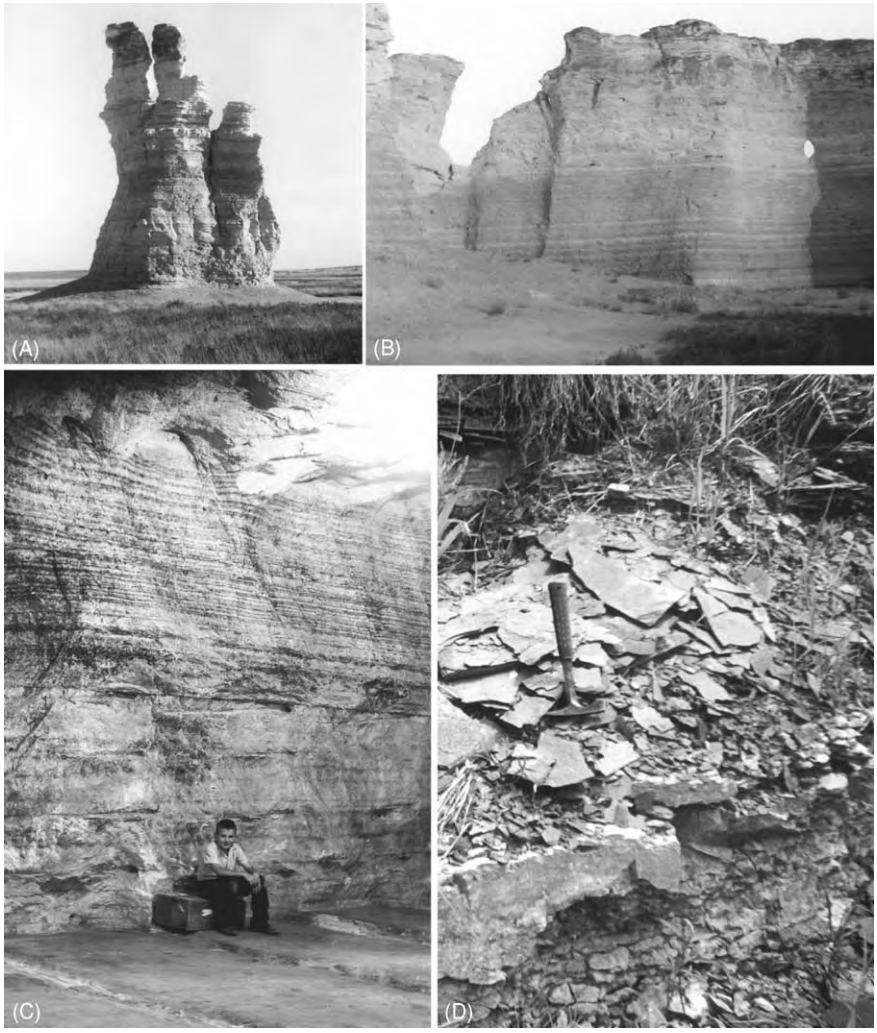


Figure 9 Examples of special rock types in the Continental Interior. (A) Cretaceous chalk in western Kansas; (B same exhibiting rhythmic layering of chalk and bentonite; (C) Permian salt in mine in central Kansas (note: rhythmic layering); and (D) Pennsylvanian thin bedded, fissile black shale.

Several other less widely distributed rock types are distinctive and informative: volcanic ash (Figure 10C) and its altered form bentonite, fossil soils (palaeosols), and extensive thin-layered chert beds (Figure 10D). Volcanic ash is recognised easily in the Tertiary and its altered form of bentonite in the Cretaceous.

The bentonites occur in distinctive thin bands over large areas and form instant time-lines. Multiple fossil soils are widespread in the Upper Permian and indicators of subareal exposure and climatic conditions at the time. In the Lower Permian, thin but persistent nodular chert beds occur and individual beds can be traced for many miles, indicating uniform conditions in the sea over a wide area. And lastly, the mantle of Pleistocene glacial material (either deposited from the glaciers or derived from their action) is an indicator of climatic conditions. These masses of thick, heavy ice deformed the crust and modified the landscape extensively.

Accumulation time of sediments can be estimated by applying known sediment accumulation rates of today to their thickness or using the relation of total sediment thickness to the known time period. Because of the limiting factors in both methods, the results are only an indication.

Erosion, usually by water and wind, and the rates at which it takes place, is of interest; generally the rates are higher in young terrains and less in the older ones, such as the Continental Interior. The obvious exception to the erosion rate, of course, is in those areas that are farmed and thus subject to the plough and other forms of tillage that break the surface and hasten the destruction of the land. Erosion rates generally are steady, but accelerated in times of crisis such

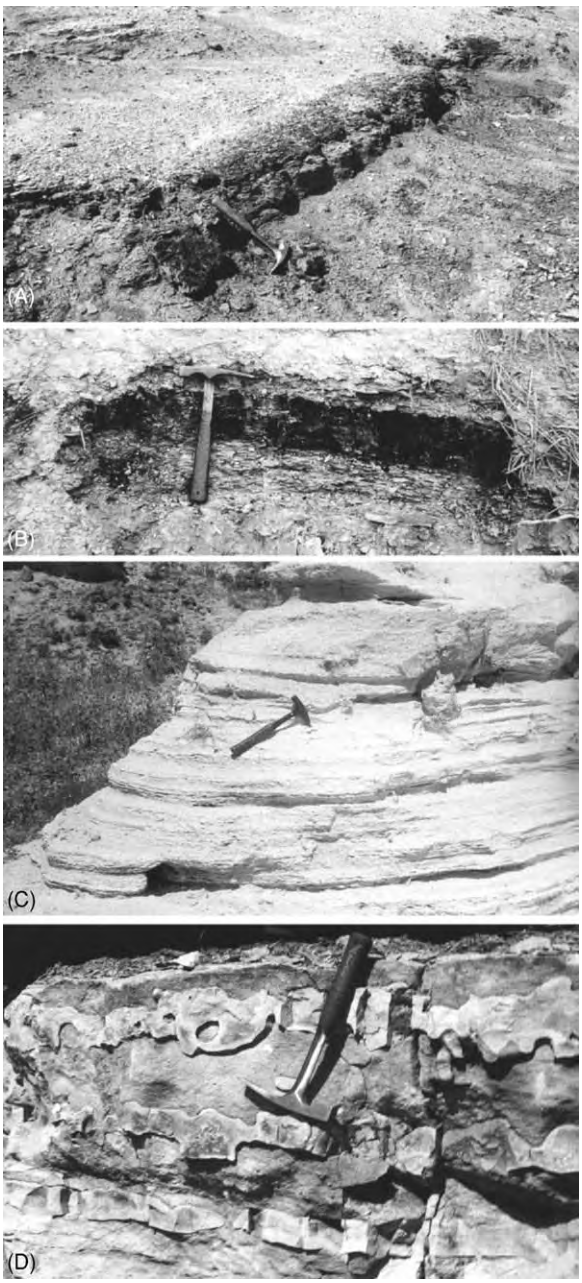


Figure 10 Other distinctive rock types that occur in the Continental Interior. (A) Lignite from Cretaceous of central Kansas; (B) thin bituminous coal (Upper Pennsylvanian); (C) Pliocene volcanic ash, a distinctive geochemical signature makes it an excellent marker bed for correlation; (D) layered chert in limestone in the Flint Hills of central Kansas.

as floods or devastating wind storms. More 'normal' rates generally prevail in between catastrophic events.

Rates of denudation can be determined by measuring the amount of material carried by streams and rivers in a unit of time. This usually is reported as so many centimetres reduction in the land surface per millennium. Because of the relatively high elevation of the North American Continent today, these rates are high and probably are not representative of the past. Another approach to denudation rates is to determine how much material was present and now evoded away and to divide that amount by the time interval. Neither approach is totally satisfactory, but both give a fair estimate.

The Sedimentary Sequence

With special conditions, unique rock sequences occur. The North American Craton is one place where these special sequences – rhythmites and cyclothems – are preserved. Rhythmites are alternations of two sediment types, such as the chalk and bentonite that developed in the Cretaceous (Zuni), or siltstone and the shales that are present in the Pennsylvanian (Absaroka). Cyclothems are repetitious sequences that occur repeatedly in the Permo-Pennsylvanian (Absaroka). The classic cyclothem records a single advance and retreat of the sea in an area, giving rise to a symmetrical rock cycle of non-marine to marine to non-marine sediments; however, the ideal seldom occurs. Because the transgressive part of the cycle usually is better represented (or better preserved) than the regressive part, most cycles are asymmetrical and many are incomplete or have components repeated.

Bundles of cyclothems have been termed megacyclothems and an example is given in Figure 11. Origin of cyclothems has been proposed as being the result of: (i) change of sea-level; (ii) tectonic movements; or (iii) climatic factors; the truth is that all three are probably responsible to some degree. In Kansas alone, in just the Middle and Upper Pennsylvanian (Absaroka) section, there are parts or all of about 40 cyclothems. Their formation probably was the result of changes in sea-level as a result of the waxing and waning of southern-hemisphere glaciers. Sediments were accommodated spatially by the craton gradually sinking in relation to sea-level, perhaps partially as a result of the weight of the added sediments and the glaciation was caused by changes in the climate.

Structural Development

The development of the structural elements on the craton to their present disposition has not been spectacular, but has been incremental and persistent. The movement is recorded, not only by the sediments, but by the gaps in the record – the unconformities. Much of the regional structural development on the craton has been by epeirogenesis; that is, up-and-down movement and tilting of the area in response to orogenic (tectonic) activity elsewhere. Four major unconformities indicate four major changes in the structural regimen – the end

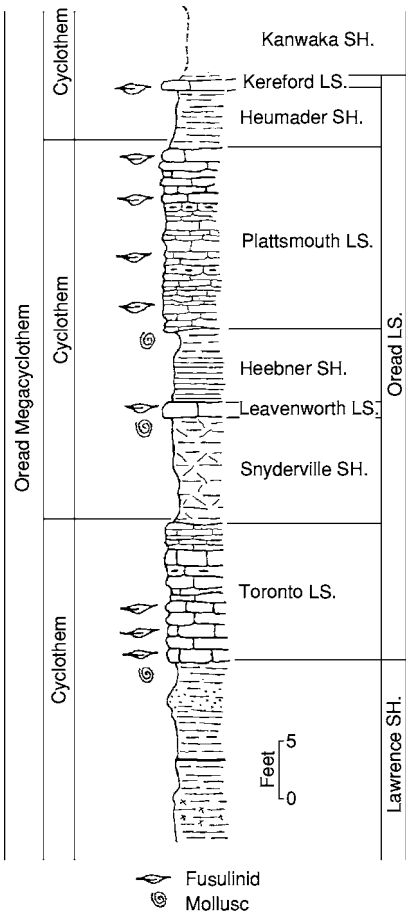


Figure 11 Example of Pennsylvanian megacyclothem from Kansas, showing individual cyclothems in the Oread Limestone Formation (Virgilian, Pennsylvanian). (Adapted from Merriam (1963) Kansas Geological Survey Bulletin 162.)

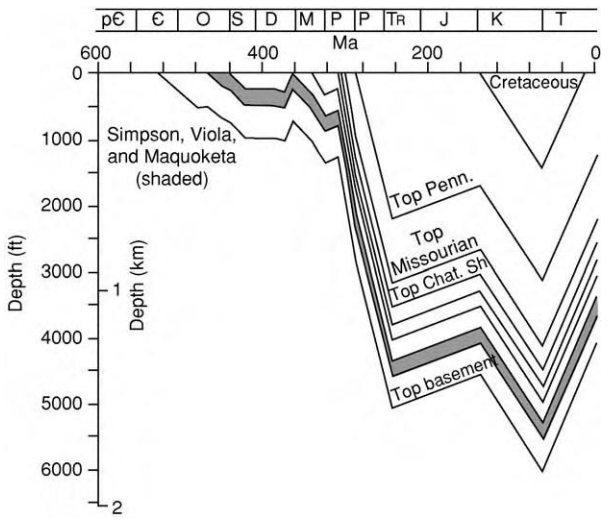


Figure 12 Burial history diagram of the Salina Basin (central Kansas; for location see Figure 13. Diagram shows amount of sediment deposited and amount removed giving depth of burial of sediments. (From Newell and Hatch (1997) Kansas Geological Survey Bulletin 240.) The diagram illustrates that the basin sub sided and received sediments from Late Cambrian to end of the Devonian when it was uplifted. Major subsidence commenced in the Mississippian, with minor uplift near end of the Mississippian and continued from then, with minor ups and downs until the present.

Table 1 Different procedures used to determine overburden removed. (Adapted from Merriam and Forster (2001) Kansas Academy of Science Program and Abstracts.)

• compaction depth relations
• vitrinite reflectance
• oil field overpressure
• maturation models
• thermal modeling
• mineralogy and rock textures of intrusives
• fluid inclusions
• geologic restoration

of the Precambrian, the Ouachita Orogeny (Mississippian), the Allegheny or Appalachian Orogeny (Late Palaeozoic), and the Laramide Orogeny (Cretaceous).

An important part of the story on the structural development of the Continental Interior is knowing how deep and when the sediments were buried. The ups and downs of the craton at any location is profiled by constructing a two-dimensional burial-history diagram (Figure 12). Another question is how much material was removed at an unconformity. There are numerous ways in which this can be done, but a combination of several approaches for any area is the most satisfactory (Table 1). This information then can be used together with the sediment distribution to understand the history.

Cratonic Structures

The sediments are preserved more completely in the basins than on the adjoining uplifts, as would be expected. The major basins, uplifts, and structural discontinuities in the Continental Interior are shown in Figure 13. A series of cross-sections of the Continental Interior are shown in Figure 14. These cross-sections give the best impression of the subsurface layer-cake nature and the relation of stratigraphic units to one another. Most impressive is the thinness of the sedimentary cover (when viewed in perspective (Figure 15); indeed, the sediments are but a facade over the Precambrian foundation. The structures shown in Figure 13 are those that formed near the end of the Mississippian and in the Early Pennsylvanian during the Alleghany/Appalachian orogenies and since have been slightly modified.

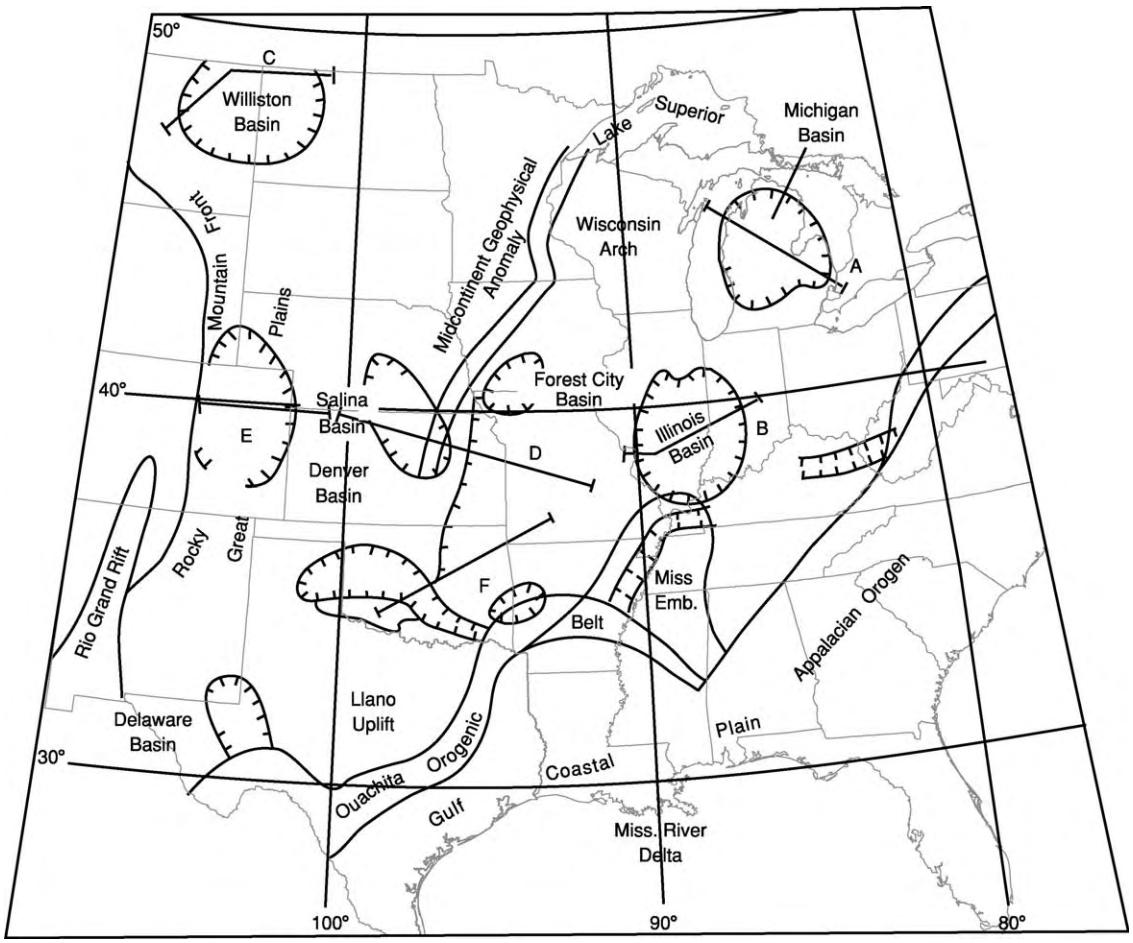


Figure 13 Location of major structural features in the Continental Interior. A, New York Alabama Lineament; B, Ouachita Foldbelt; C, Wichita Uplift; D, Anadarko Basin; E, Ardmore Basin; F, Aricoma Basin; G, Nemaha Uplift. (From Hinze and Braile (1988).)

There is no apparent spatial pattern of the basins; they seemingly formed wherever there was a weakness in the foundation. Through time, several of these regional structures were inverted, (that is, basins were uplifted to become positive areas, or vice versa), and some positive features foundered to form depressions.

The Precambrian basement also is the key to the location of the minor structures of the region. These secondary features show a preferential orientation: one set north-east–south-west and the other north-west–south-east, with a ancillary set north–south and east–west. These structures were formed by the vertical displacement of basement fault blocks in response to outside tectonic forces in which the overlying sediments were draped over the offset blocks by differential compaction of the unconsolidated, soft sediments. These anticlines (plains-type folds) are present in the sedimentary cover on cratons worldwide (Table 2; Figure 16). After initial formation of the structural regimen in Late Mississippian to Early Pennsylvanian time (Absaroka), continued movement of the basement blocks accentuated the folds. This incremental movement has continued from the Late Palaeozoic until today.

The incremental movement of the fault blocks is recorded in the Permo-Pennsylvanian (Absaroka) sedimentary section by thinning of beds over the structure and by convolute features known as seismites (Figure 17). The seismites, formed by dewatering of soft sediments, were triggered by palaeoearthquakes. By noting the geographic and stratigraphic occurrence of the seismites, it is possible to determine the relative intensity and size of the area affected by the palaeoquake. Earthquakes are the result of the adjustment of the basement to stress and are recorded on the craton today. In historic times, several large earthquakes occurred – the New Madrid quake (1811–1812) in Missouri and Manhattan quake (1867) in Kansas are two examples; fortunately there was little to damage at those early dates.

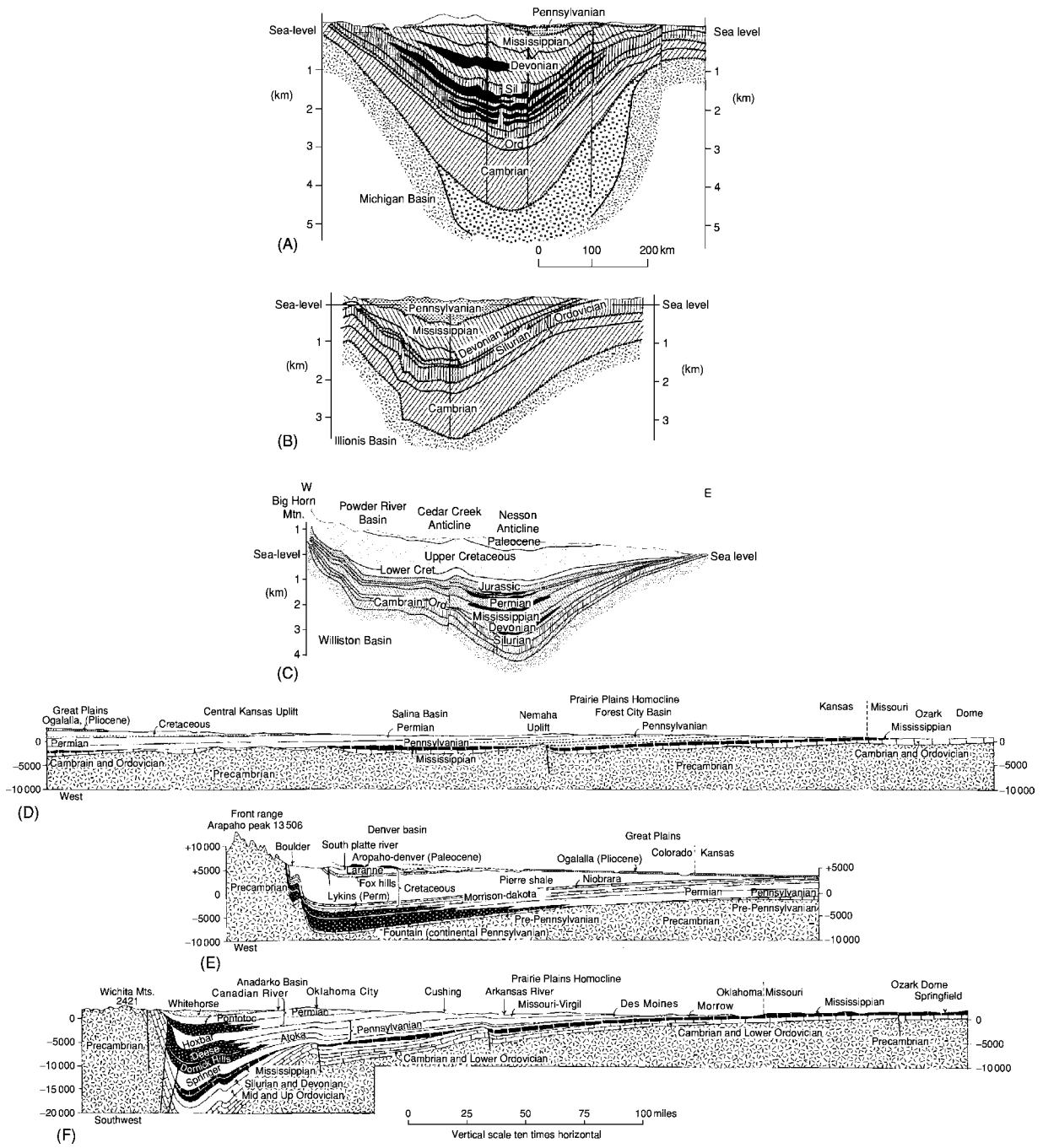


Figure 14 Cross sections of selected areas in the Continental Interior. A Michigan Basin; B Illinois Basin; C Williston Basin; D Ozark Dome west to Great Plains; E Great Plains through Denver Basin to Rocky Mountains; F Ozark Dome south west through the Anadarko Basin to Wichita Mountains. Denver and Anadarko basins are two of the deepest troughs in the Continental Interior. (A, B, and C from Bally (1989) Phanerozoic basins of North America, *in* The Geology of North America, v. A; E, F, and G from King (1951) The tectonics of middle North America: Princeton University Press.)

Mineral Resources

The Continental Interior has long been known as a rich source of metallic minerals, mostly produced from Precambrian rocks. Gold has been produced from the Precambrian at Lead, South Dakota and the Porcupine and Kirkland Lake districts in Ontario, Canada. The extensive Lake Superior (Michigan) copper deposits occur in Precambrian rocks and have been known for

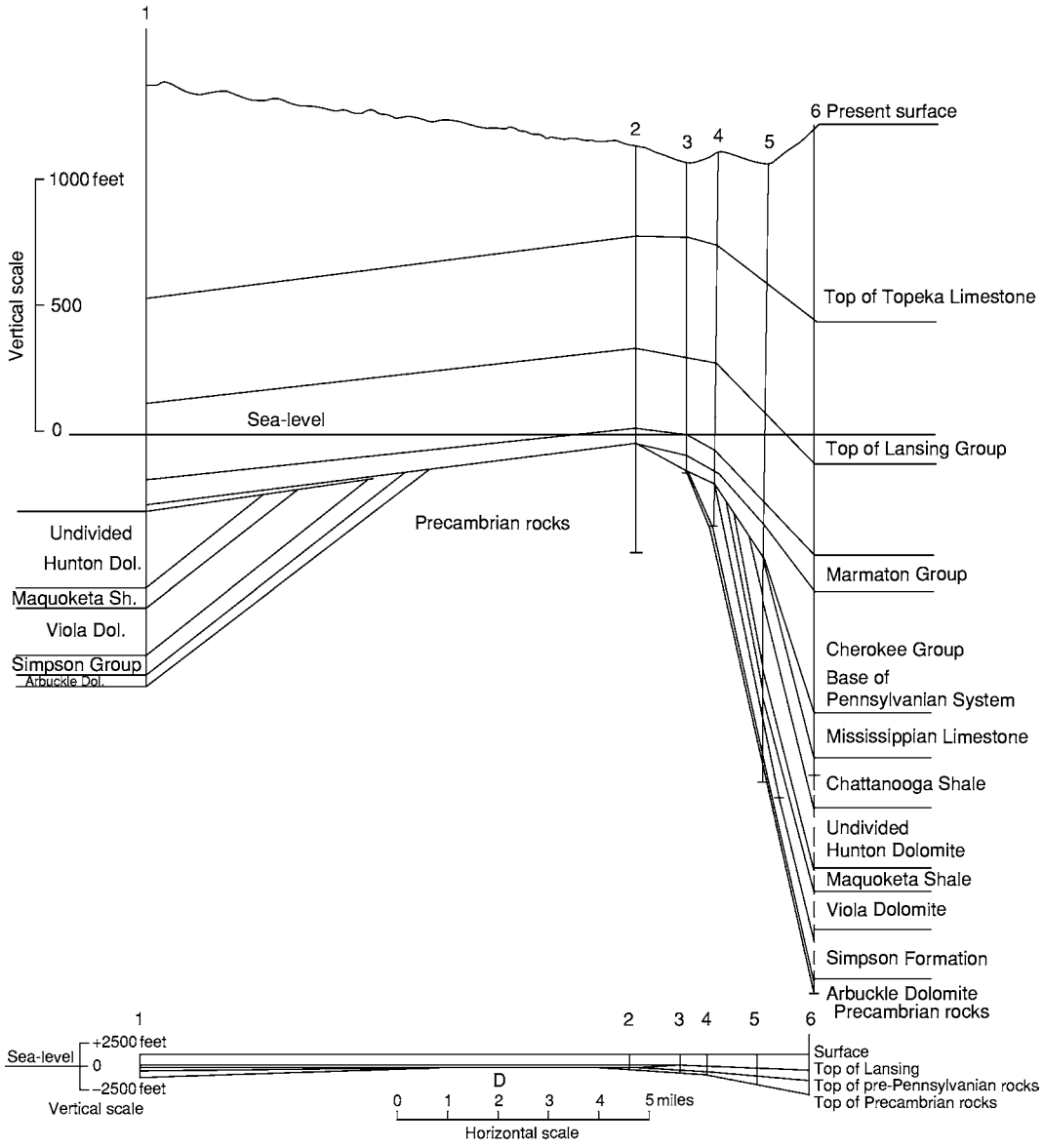


Figure 15 Cross section of the Nemaha Anticline, a major structural feature showing the result of exaggeration (upper) to at scale (lower). Exaggeration is used by geologists to show relationships better in flat lying stratigraphic units (from Lee, 1956, Kansas Geological Survey Bulletin 121).

Table 2 Feature of plains type folds. (Modified from Powers (1925) Geological Society of America Bulletin, v. 36.)

- Structural noses are dominant structure at shallow depth (ie in surface units)
- Most folds are small in areal extent and amount of structural closure
- Closure increases downward and structure becomes sharply defined
- Amount of closure is proportional to size in many structures

centuries and were utilized by the Indians. Copper was also produced at Ducktown, Tennessee from Cambrian metamophosed sediments. The occurrences of lead and zinc deposits in south-eastern Missouri were known from the eighteenth century. The Tri-State District centred on Joplin, Missouri, described as the greatest zinc district on Earth, occurs in Mississippian carbonates. The rich Precambrian iron-ore deposits of the Mesabi Range in Minnesota and in the adjacent states of Wisconsin, and Michigan have been utilised for a century. Sudbury, Ontario, Canada is a major world source of nickel that occurs in a Precambrian complex.

Extensive sub-bituminous and bituminous coal deposits occur in the Midcontinent area and is mined both underground and in strip pits where shallow. Petroleum seeps in Miami County, Kansas were used by the Indians for medicinal purposes, and pioneers

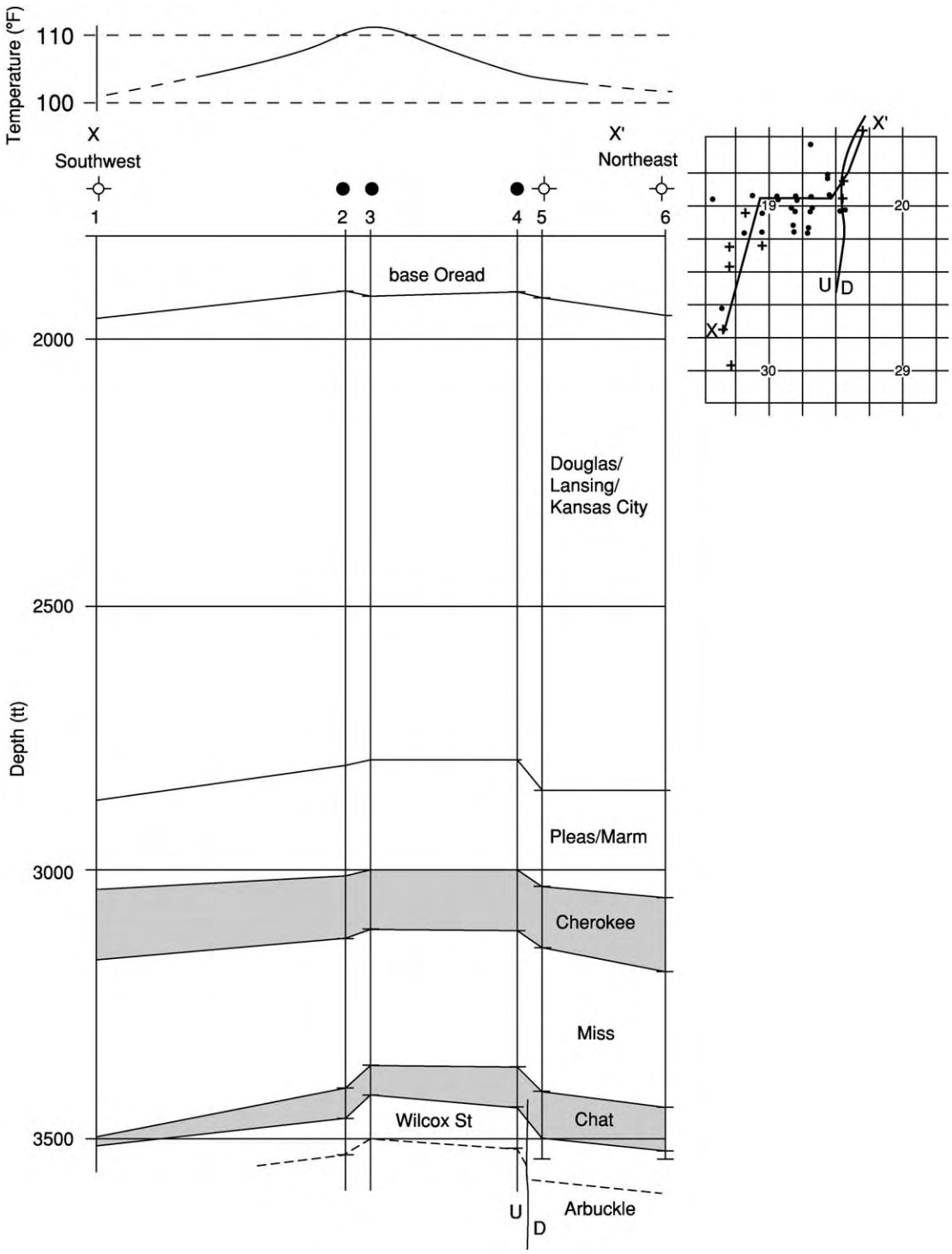


Figure 16 Cross section of typical 'plains type fold' (Slick Carson oil field, T32S, R3E, Cowley County, Kansas). Note the anticline closure increases with depth, faulted on east flank, and temperature high over crest of anticline; all characteristics of oil producing plains type folds.

on the Santa Fe and Oregon trails on their way west used it for lubricating their wagon wheels. Petroleum was discovered in a well drilled in 1860 near Paola, Kansas, opening the large Midcontinent Oil and Gas Province shortly after the Drake well was drilled in 1859 in Pennsylvania. Natural gas was exploited for town lighting, starting in about 1870.

Gold discoveries in the Black Hills of South Dakota and in the Rocky Mountains in Colorado spurred the gold rushes, with thousands of get-rich

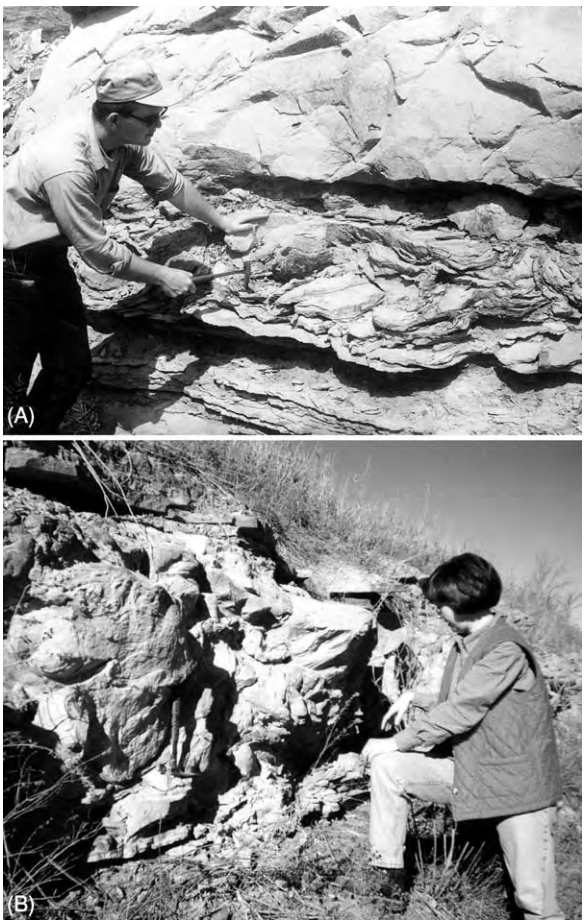


Figure 17 Sedimentary soft sediment deformation indicating incremental movement of Precambrian fault blocks. (A) Soft sediment deformation (seismite) in sandstone in Douglas Group (Upper Pennsylvanian) in south eastern Kansas; (B) seismite in Upper Pennsylvanian Douglas Group clastic section in Montgomery County, south eastern Kansas.

gold seekers making their way across the plains to the 'promised land'. As the area was populated, building material and water were needed and many other resources besides metals were exploited, including gypsum, salt, fluorite, clay, lime, glass sand, sand and gravel, building and crushed stone, and fertilizer minerals.

See Also

North America: Precambrian Continental Nucleus; Southern and Central Appalachians; Southern Cordillera. **Sequence Stratigraphy**. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Frazier WJ and Schwimmer DR (1987) *Regional stratig raphy of North America*. New York: Plenum Press.

Houston RS (ed.) (1993) The Wyoming Province. In: Reed JC Jr (ed.) *Precambrian conterminous US: The Geology of North America*. Geology Society of America, vol. C 2, pp. 121 170.

Merriam DF (1963) The geologic history of Kansas. *Kansas Geological Survey Bulletin* 162: 317.

Rascoe B Jr and Hyne NJ (eds.) (1989) Petroleum geology of the Midcontinent. *Tulsa Geology Society Special Pub lication* 3: 162.

Sloss LL (ed.) (1988) *Sedimentary cover North American craton: US: The Geology of North America*. Geology Society of America, v. D 2, p. 506.

Van Schmus WR and Bickford ME (eds.) (1993) Transcon tinental Proterozoic provinces, In: Reed JC Jr (ed.) *Precambrian conterminous US: The Geology of North America*. Geology Society of America, v. C 2, pp. 171 334.

Northern Cordillera

J W H Monger, Geological Survey of Canada, Vancouver, BC, Canada and Simon Fraser University Burnaby, BC, Canada
R A Price, Queens University, Kingston, ON, Canada
W J Nokleberg, United States Geological Survey, Menlo Park, CA, USA

Introduction

The northern Cordillera extends some 3500 km north-westward from latitudes 47–48° N in Washington–Montana, USA, through Canada, and westwards across Alaska to terminate physiographically, but not geologically, at the Bering and Chukchi Seas between Alaska and the Russian Far East. Its southern boundary is delineated by the southern limits of the Olympic Mountains and North Cascade Ranges in north-western Washington, the northern Columbia Plateau in central Washington, and a system of west–north-west-trending faults in northern Idaho and western Montana called the 'Lewis and Clark line'.

The northern Cordillera follows an ocean–continent boundary created by about 540 Ma, emerged as a mountain belt between 185 and 60 Ma, and is still growing. A noteworthy feature is that about three-quarters of its surface, and all of Alaska, is underlain

by an orogenic collage composed of allochthonous rocks accreted to the ancient continental margin of western North America.

Physiography

The northern Cordillera contains three major physiographical divisions, the Rocky Mountains System, the Intermontane Plateau System, and the Pacific Mountains System, each comprising subsidiary ranges or other physiographical features (Figure 1; Appendix A). It is flanked on the east by the Interior Plains System, and on the north by a continental shelf and slope less than 200 km wide that terminates in water depths of about 2000 m in the Arctic Ocean. The Rocky Mountains System contains many summits that exceed 2000 m in elevation, with Mount Robson near latitude 53° N the highest peak at 3955 m. Between latitudes 48° and 60° N, the great valley called the Rocky Mountains Trench divides the system longitudinally, but its continuation north of this, called the Tintina Trench, extends into the Intermontane Plateau System in east-central Alaska. The Intermontane Plateau System has lower relief than flanking systems, and is mostly between 1000 and 2000 m in height, but in Alaska includes extensive lowlands under 300 m. Summit elevations in the Pacific Mountains System are mostly between 2000 and 3000 m, except for the Alaska Range in southern Alaska and Saint Elias Mountains in south-western Yukon where, respectively, Mount McKinley (or Denali) at 6193 m and Mount Logan

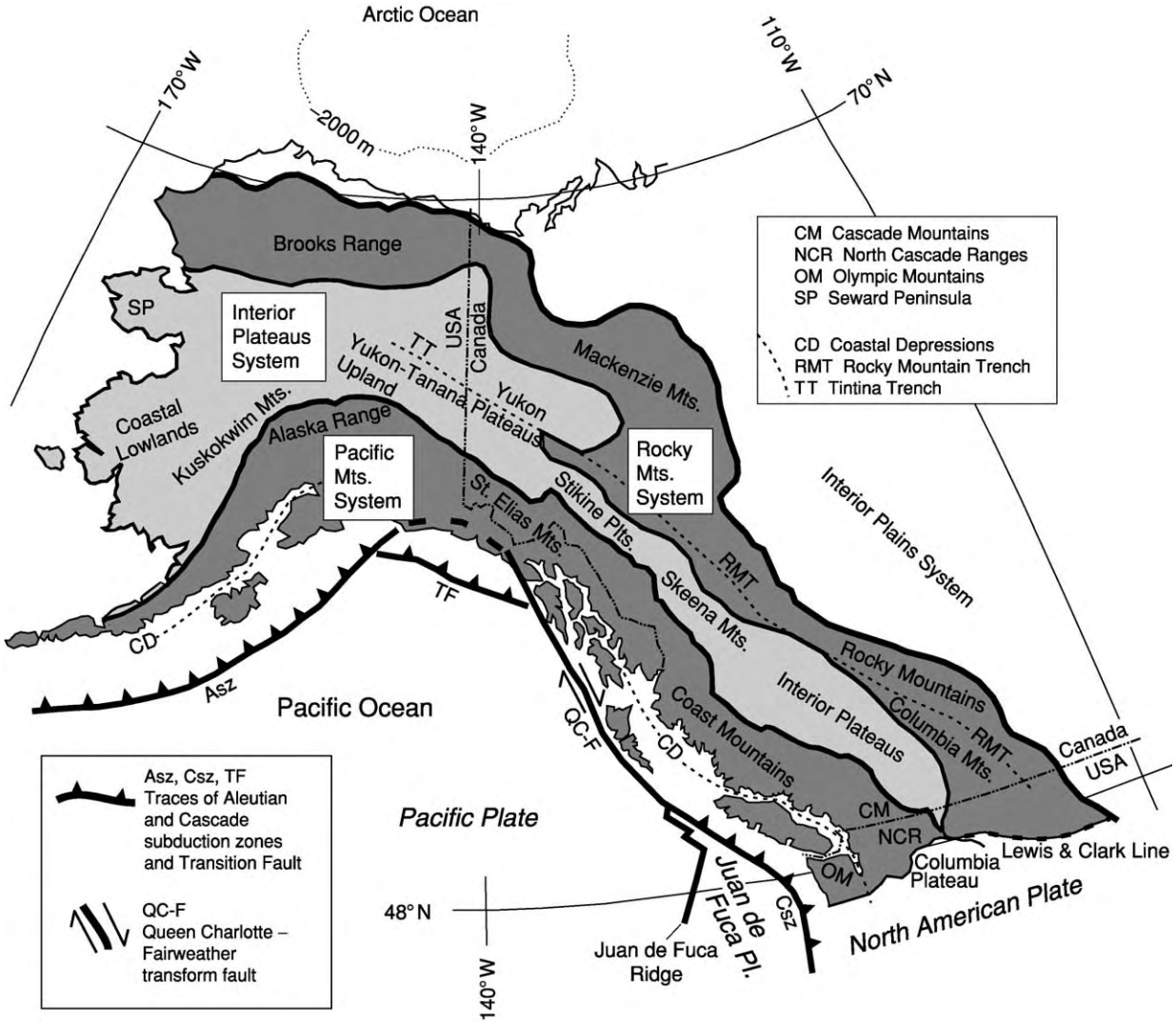


Figure 1 Major physiographical systems of the northern Cordillera, some subsidiary physiographical features, and plate boundaries.

at 6050 m are the highest peaks on the continent. The Pacific Mountains System contains a longitudinal but discontinuous coastal depression, oceanward of which are many islands and the Olympic Mountains in the south and Saint Elias Mountains in the north. The continental shelf and slope along the Pacific Ocean are about 200 km wide; the base of the continental slope is about 2000 m deep south of latitude 60° N, but westwards across the northern Gulf of Alaska its depth increases to 7000 m in the Aleutian Trench off the western Aleutian Islands. The Bering and Chukchi Seas are mostly less than 200 m deep, except for a 3000 m deep basin north of the western Aleutians.

The Cordilleran ice-sheet covered most of the region several times between 2 Ma and 12 Ka, was up to 2.5 km thick in places, and shaped local landforms. Parts of central Alaska, the eastward extension of this region into Yukon, and the Arctic coastal plain remained largely ice-free. Today, ice cover is restricted to mostly retreating glacier complexes mainly at higher elevations, but which descend to sea-level in parts of southern and south-eastern Alaska.

Neotectonics

The northern Cordillera occupies a 'soft' plate margin up to 1200 km wide landward of an offshore convergent–transform–convergent plate boundary (Figure 1). Earthquakes are concentrated near most plate boundaries, but are scattered across the entire region. South of latitude 51° N, the small, young (≤9 Ma) Juan de Fuca and Explorer oceanic plates converge with the North American Plate at rates of 2–4.6 cm year ^{1} along the currently aseismic, locked Cascadia subduction zone, 200–300 km landward of which is the active Cascade magmatic arc. Between latitudes 51° and 60° N, the plate boundary is the dextral transform Queen Charlotte–Fairweather Fault, west of which the enormous Pacific Plate moves northwards to converge with the North American Plate at rates ranging from 5.7 cm year ^{1} near the transform to 8 cm year ^{1} south of the western Aleutian Islands. The Pacific Plate descends beneath Alaska along the Aleutian subduction zone, and above it the Aleutian–Wrangell magmatic arc extends eastwards some 3800 km from the Komandorsky Islands near Kamchatka, along the Aleutian Island chain, and into the mountains of southern Alaska and south-westernmost Yukon. The North American Plate continues for some 2000 km west of the Bering Strait to terminate against the Eurasian Plate near longitude 140° E.

There is widespread evidence for neotectonic activity in the northern Cordillera in addition to earthquakes and volcanoes. Seismically imaged folds and thrust faults in Late Tertiary and Quaternary strata are in the accretionary wedges above the Cascadia and Aleutian subduction zones, and faults and folds occur in the Arctic shelf-slope region off northern Alaska and Canada. On-land surface evidence for latest Tertiary to Holocene deformation is sparse, possibly because erosion by ice eliminated the surface expression of many small-scale tectonic features. Segments of the Denali Fault in southern Alaska and south-western Yukon are active dextral strike-slip faults, as are other fault segments in central Alaska. Glaciated valleys crossing the Fairweather Fault in south-eastern Alaska are dextrally offset, and neotectonic faults occur on Vancouver Island in south-western British Columbia. In western Washington, east–west-trending Holocene folds and reverse faults in the coastal depression are related to north–south compression, as are structures involving Late Miocene lava flows in the Columbia Plateau, east of the Cascade arc.

On a larger scale, regional differential uplift of up to several kilometres in the last 5–10 million years in the Alaska Range, Saint Elias Mountains, Coast Mountains, and Olympic Mountains is documented by elevated Late Tertiary erosion surfaces, changes from wet to dry floras in regions now in rain-shadows inland of the coastal mountains, palaeodrainage patterns, and fission track studies. Differential regional uplift is not well documented elsewhere, but major reversals of drainage in upper reaches of the Fraser and Columbia Rivers in central and south-eastern British Columbia, respectively, possibly reflect Neogene differential uplift of parts of the Rocky Mountains System, in addition to documented blockage by temporary ice dams and landslides during waning stages of Cordilleran glaciation. The late differential uplift means that physiographical system boundaries only approximately coincide with major bedrock boundaries, many of which formed between the Early Jurassic and the Neogene.

Crustal Thickness

Seismic refraction and reflection profiles across the northern Cordillera show the crust to be nearly 50 km thick under the south-eastern Canadian Cordillera and parts of northern Alaska (Figure 2). Below the Intermontane Plateau System in Alaska and northern Canada, it is about 35 km thick, and under the southern Canadian Cordillera only 30 km thick. Above subducting lithosphere in southern Alaska and south-western Canada, the crust is apparently underplated by material detached from the lower plate and is over 50 km thick, whereas landward of the Queen Charlotte–Fairweather transform fault it is only about 25 km thick.

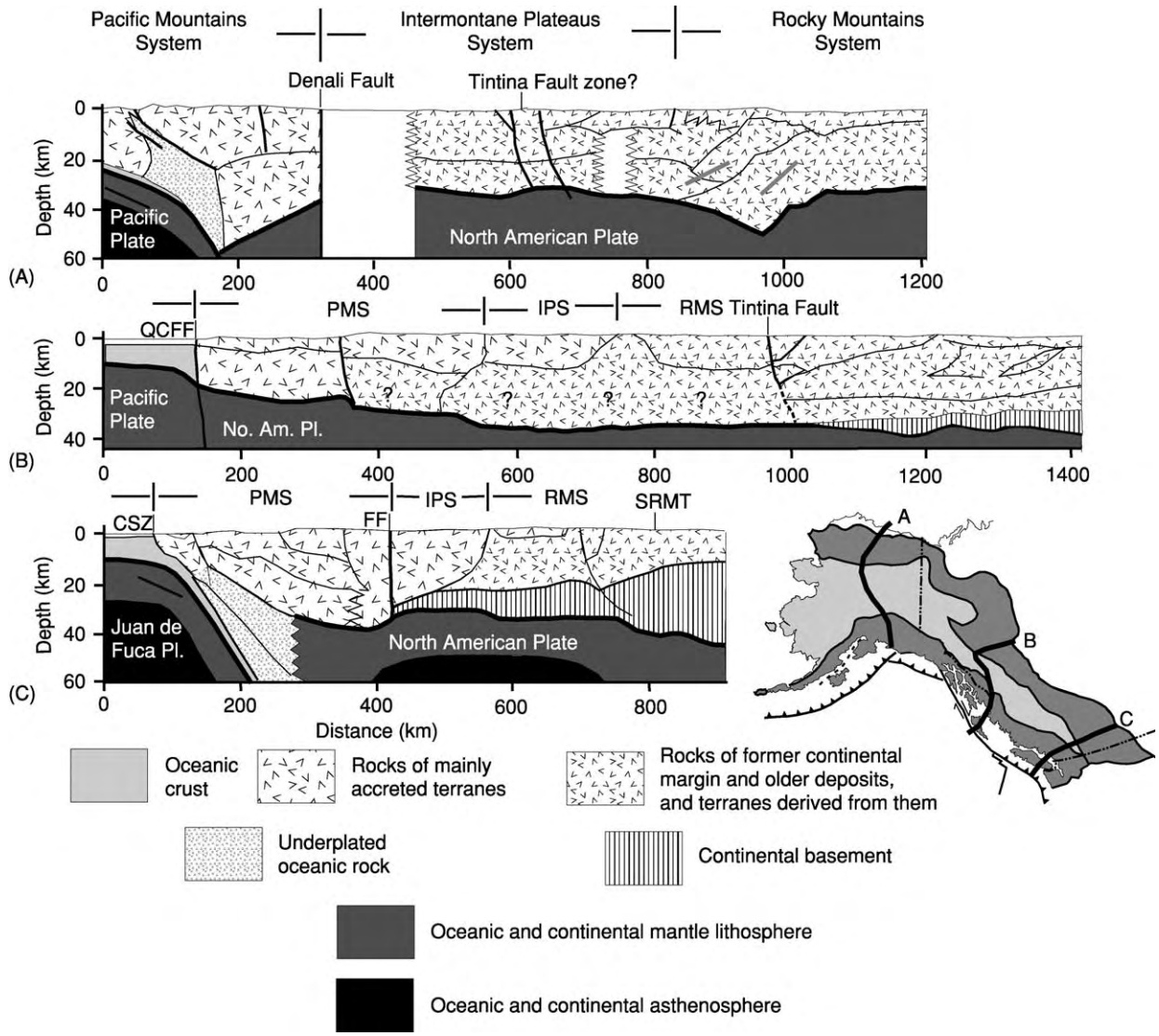


Figure 2 Crustal thickness, crustal origins, and lithospheric structures interpreted from seismic reflection and refraction profiles and surface geology across the northern Cordillera. Locations on map: (A) Trans Alaska Crustal Transect; (B) Lithoprobe Northern Canadian Cordillera Transect; (C) Lithoprobe Southern Canadian Cordillera Transect. CSZ, Cascadia subduction zone; QCFF, Queen Charlotte Fairweather Fault; SRMT southern Rocky Mountains Trench. Modified from Clowes RM and Hammer PTC (2000) (*see* Further Reading).

Bedrock Features

Two fundamentally different divisions of older rocks in the northern Cordillera are separated by a boundary that snakes back and forth across the physiographical boundary between the Rocky Mountains and Intermontane Plateau Systems. In most of the Rocky Mountains System, autochthonous or parautochthonous rocks were deposited on or near continental basement. Much of the remainder is an orogenic collage of allochthonous terranes, many of which show little geological or isotopic evidence of continental influence. Superimposed on the older rocks are features formed during the evolving Cordilleran mountain belt in Mesozoic and Tertiary time (Figures 3 and 4).

Autochthonous and Parautochthonous Rocks

Rocks of the ancient continental margin, and deposited on or near it, form five divisions based on age and the tectonic setting in which they formed.

1. The continental basement that extends under the eastern Cordillera from Montana–Idaho–eastern Washington to north-western Yukon is a Palaeoproterozoic orogenic collage composed of fragments of Archaean (4.0–2.6 Ga) continental crust embedded in metamorphosed remnants of

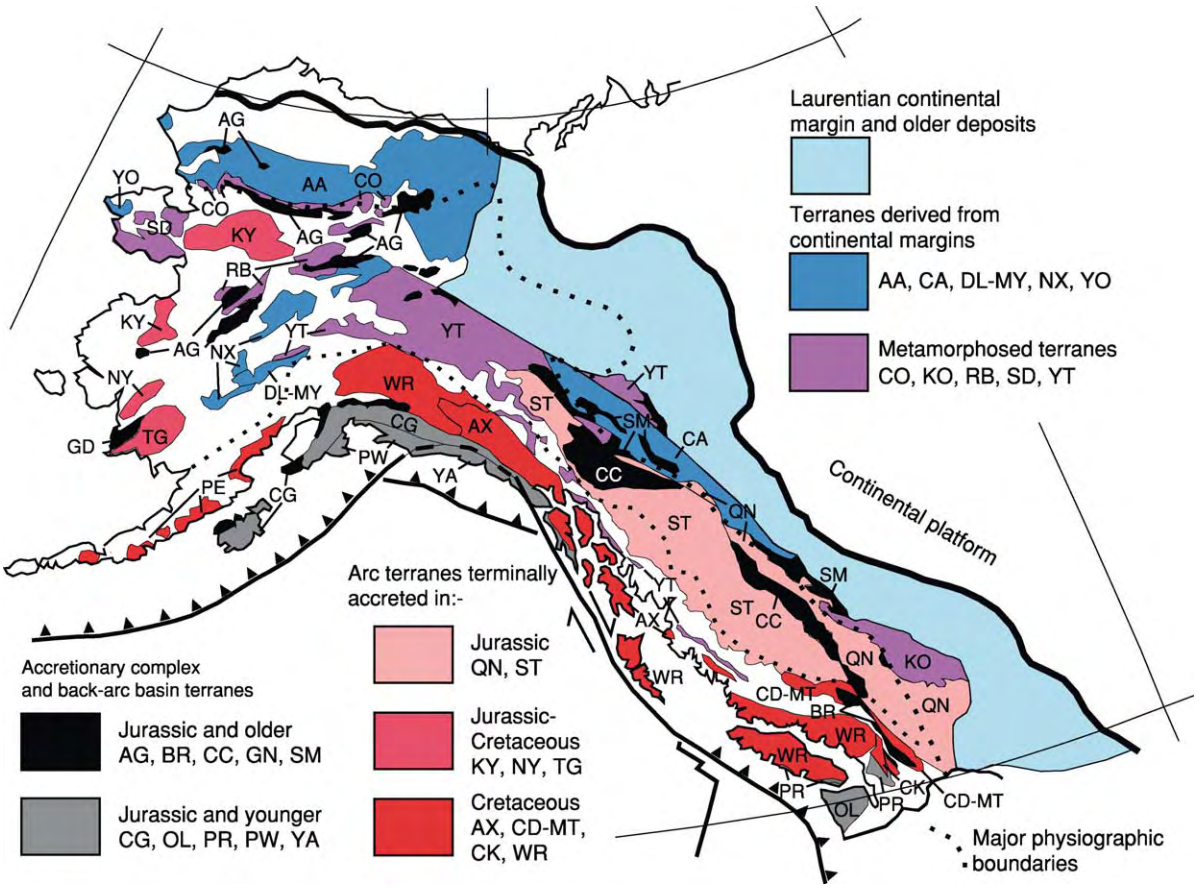


Figure 3 Distribution, origins, and times of accretion of terranes. Terrane names, abbreviated herein as AG, BR, etc., and brief terrane descriptions are in Appendix B.

Palaeoproterozoic (2.3–1.8 Ga) magmatic arcs and accretionary wedges. Although mostly buried, it is exposed locally in south-eastern British Columbia (Monashee Complex) and northern Idaho (Priest River Complex). The continental basement slopes beneath the eastern Rocky Mountains System as a layer about 35 km thick, but thins abruptly and appears to be no more than a few kilometres thick beneath the southern Intermontane Plateau System, and is absent below the Pacific Mountains System.

2. Palaeoproterozoic (1.8–1.7 Ga) basinal to platformal deposits (the Wernecke Supergroup) in northern Canada and east-central Alaska (Tindir Group) are rift-related and as much as 13 km thick. However, it is uncertain whether they formed in an intercontinental basin or during continental break-up. An angular unconformity between folded and slightly metamorphosed rocks of the Wernecke Supergroup and younger (1.4–1.0 Ga) Mesoproterozoic basinal to platformal deposits brackets the age of the Racklan Orogeny. Mesoproterozoic (1.5–1.4 Ga) strata (Belt-Purcell Supergroup) in Montana, Idaho, and adjoining parts of British Columbia and Alberta form a north-west-trending prism up to 20 km thick. The lower part consists of deep-water turbiditic quartzite, siltstone and argillite, and voluminous mafic sills; the upper part comprises shallow-water siliciclastics and stromatolitic carbonates deposited in intertidal and floodplain settings. The rocks apparently accumulated within and near an intracontinental rift basin.
3. Neoproterozoic (780–550 Ma) strata are the oldest rocks to occur more or less continuously along the entire northern Cordillera. Glaciogenic deposits (Sturtian?) at the base of the Windermere Supergroup contain mafic volcanic rocks with ages of 780–760 Ma. These date the initiation of rifting that led eventually to dispersal of components of the Neoproterozoic Rodinian supercontinent and creation of what became the margin of Laurentia. The overlying succession, up to 10 km thick, consists of shale with intercalated quartz-and-feldspar 'grit' turbidites, and grades upwards into slope deposits, carbonate, and evaporite shelf deposits.

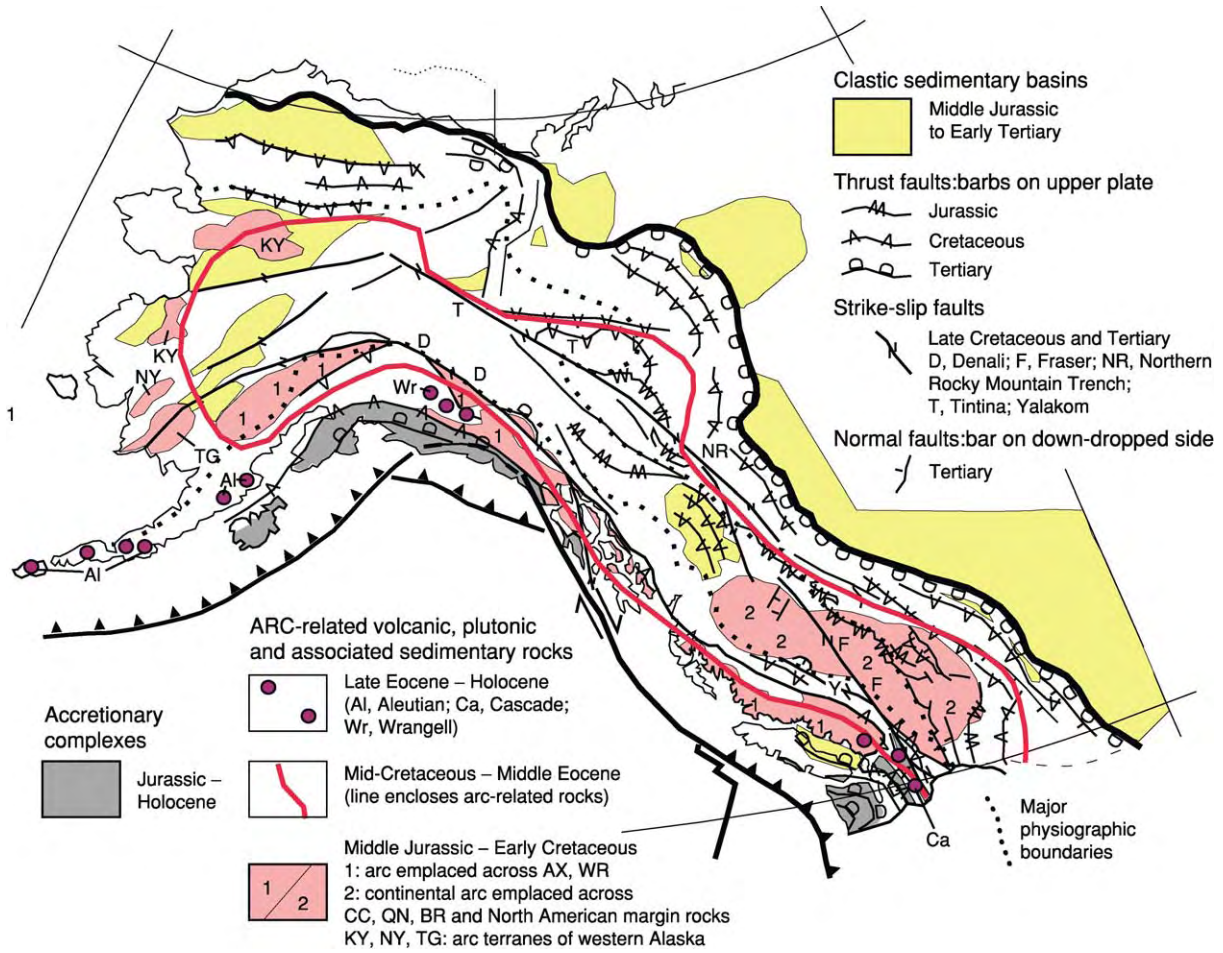


Figure 4 Features accompanying Cordilleran orogenesis: magmatic arcs and associated accretionary complexes; basins resulting from contraction, crustal thickening, uplift, and erosion; and syn and post accretionary structures. Terrane names, abbreviated herein as AX, WR etc; are in Appendix B.

Younger glaciogenic sediments (Varanger equivalent?) occur locally in the upper Windermere. A second major rifting episode at the top of the succession is recorded by an angular unconformity beneath Cambrian rocks and by coarse siliciclastic deposits and associated magmatic rocks dated at 550–600 Ma. Rifting culminated in complete supercontinental separation by about 540 Ma at latitudes 48–55° N, but took place earlier further north.

4. Early Cambrian to Late Jurassic (540–150 Ma) miogeoclinal strata were deposited along the newly formed continental margin of Laurentia. In the south, the continental shelf appears to have been relatively narrow and contains abundant shallow-water carbonates, but in the north it is broader and deeper water sediments are more abundant, characteristics that have been ascribed to asymmetric rifting. Cambrian to Late Devonian carbonates with locally interbedded shales exhibit lateral facies changes from incomplete platform successions about 2 km thick below the plains and in the eastern mountains, to shallow-water accumulations up to 10 km thick that were deposited on the former continental shelf. Beyond the shelf, the strata are replaced by calcareous shale and shale, and further west by deep-water carbonaceous shale with local intercalated mafic volcanic rocks and coarse-grained siliciclastic deposits. Devonian–Mississippian and Permian–Triassic siliciclastic rocks, widespread along the Laurentian continental margin from northern Alaska to north-eastern Washington, have distinctive detrital zircon and geochemical signatures indicating their derivation from northern source areas. The predominance of latest Palaeozoic and Early Mesozoic siltstone and shale over carbonate may reflect relocation of this margin to higher latitudes as a result of the northward drift and rotation of Laurentia.
5. Latest Jurassic to Palaeogene (150–40 Ma) clastic deposits were eroded from older rocks that

were uplifted and exposed above sea-level during folding and thrusting towards, and over, the edge of the continental platform, and accumulated in foreland basins along the margins of the advancing fold and thrust belt. The sediments, up to 4 km thick, consist of deltaic and alluvial sandstone, shale, conglomerate, and coal, as well as interfingering marine sediments deposited during three transgressive–regressive cycles of latest Jurassic and Early Cretaceous, mid-Cretaceous, and Late Cretaceous and Palaeogene ages.

Rocks of the Orogenic Collage

Major components of the orogenic collage have been called terranes. These are fault-bounded regions of upper crustal rocks, detached from their lower crust and mantle lithospheres, and with different geological histories from one another. Many terranes contain palaeomagnetic, palaeobiogeographical, and/or detrital zircon signatures very different from those of neighbouring terranes and from the north-western (in present co-ordinates) Laurentian margin. Most terranes in the Pacific Mountains and Intermontane Plateau Systems south of latitude 60° N are remnants of arcs and accretionary complexes whose generally low Sr_i and positive ε_{Nd} values indicate the mantle origins of the magmatic rocks within them. Others, mainly in the western Rocky Mountains System and Intermontane Plateau System north of latitude 60° N, are displaced former continental margin deposits (**Figure 3**; Appendix B). Evidence of terrane provenance and amounts of displacement can be gathered by comparing these signatures with those in coeval rocks of the ancient continental margin, which has been oriented approximately longitudinally since the Late Palaeozoic, and thereby provides latitudinal control.

Continental margin terranes mostly formed along the Laurentian margin. The Arctic Alaska Terrane in the northernmost Cordillera consists of Neoproterozoic to Jurassic continental margin strata. The terrane has affinities with parts of the Innuitian Orogen of Arctic Canada in that it contains rocks deformed, metamorphosed, and intruded during the Devonian Ellesmerian Orogeny. However, there is disagreement about whether or not it underwent large amounts of counterclockwise rotation or translation relative to the rest of North America during the Early Cretaceous rifting and seafloor spreading that created adjacent parts of the Arctic Ocean basin. The Coldfoot, Kootenay, Seward, and Yukon–Tanana terranes contain metamorphosed Neoproterozoic and Palaeozoic strata formed in distal continental margin settings, and some contain Middle Devonian to Early Carboniferous arc magmatic rocks and underwent latest Palaeozoic metamorphism. Possible exceptions to a Laurentian provenance are the Dillinger, Mystic, and Nixon Fork terranes in Intermontane Alaska, whose Early and Middle Palaeozoic stratigraphy and faunas are similar to those of the Siberian Craton, and possibly parts of the Kootenay Terrane, for which a peri-Gondwanan provenance has been suggested.

A key question, whose answer constrains the potential amount of displacement of magmatic arc terranes relative to the former continental margin, is whether subduction was beneath the continent or away from it. Mid-Mesozoic and younger magmatic arcs evidently faced away from the continent, because contemporary accretionary complexes are located oceanward of them, with the exception of the Koyukuk–Nyak–Togiak Arc terranes in western Alaska, which are outboard of the oceanic Angayucham terrane. The facing directions of Early Mesozoic arc rocks in the Quesnel Terrane and of the Middle Jurassic arc in the Stikine Terrane were also away from the continent, but those of Early Mesozoic and older arcs of the Stikine, Alexander, and Wrangellia terranes are unknown. It has been suggested that the facing direction of the Late Palaeozoic arc in the Quesnel Terrane in southern British Columbia was away from the continent, but, conversely, that of a Late Palaeozoic arc in the Yukon–Tanana Terrane, where Carboniferous and Permian subduction evidently involved the Slide Mountain Terrane, is thought to have been towards it. Middle and Late Devonian arc rocks extend into continental margin deposits in the southeastern Canadian Cordillera, suggesting that at that time subduction was beneath the continent.

Terranes mainly in the Intermontane Plateau System south of latitude 60° N have counterparts as far south as central California and north-western Nevada. Permian to earliest Jurassic faunas in the Quesnel, Slide Mountain, and Stikine terranes differ from those in rocks deposited on the ancient continental margin in Canada, but are comparable to coeval faunas in the western conterminous USA and Mexico. Middle Permian to Middle Triassic faunas in the accretionary complex called the Cache Creek Terrane, located between the Stikine and Quesnel terranes, are similar to faunas found in eastern Asia and in rocks of Permian Tethys.

The Wrangellia and Alexander terranes in the Pacific Mountains System of Canada and Alaska contain Palaeozoic faunas most similar to those in the Canadian Arctic and Siberia. Their Early Mesozoic faunas are more comparable to those further south in the western Cordillera, although Wrangellia also contains an Early Jurassic ammonite of a characteristic Russian species unknown elsewhere from North America. The Alexander Terrane features a

distinctive record of Neoproterozoic and Early Palaeozoic arc magmatism and orogeny; suggestions have been made for its origin in regions as far apart today as eastern Australia and the Barents Sea, with the latter currently favoured.

Northern Cordilleran Mountain Building

The final accretion of terranes to the former continental margin started in the Early Jurassic and was associated with the subduction of oceanic lithosphere beneath the western North American Plate, accompanying arc magmatism, and the collapse of back-arc basins. By the Late Cretaceous, almost the entire region was above sea-level.

For much of its length, the northern Cordillera features two tracts of Mesozoic and Palaeogene high-grade metamorphic rocks that are loci of the greatest amounts of burial, crustal thickening, differential uplift, and erosion and/or tectonic exhumation. The metamorphic rocks are located in internal parts of the Rocky Mountains and Pacific Mountains Systems and are flanked on both sides by far less metamorphosed rocks in much of the remainder of the Cordillera (Appendix A). Some metamorphic rocks in Intermontane Alaska and Yukon (in the Ruby, Seward, and Yukon–Tanana terranes) are older. Starting in the latest Early Jurassic ($\leq$185 Ma), terranes (Cache Creek, Quesnel, Stikine, and Slide Mountain) in the Intermontane Plateau System in Canada were thrust towards the continent and deeply buried former continental margin rocks that are exposed today in internal parts of the Rocky Mountains System. Metamorphism along the southern Arctic Alaska Terrane and in parts of Intermontane Alaska similarly appears to be related to Middle Jurassic–Early Cretaceous thrusting of oceanic rocks and arc terranes (Angayucham, Koyukuk, Nyac, and Togiak) on to Arctic Alaska. Beginning in the mid-Cretaceous ($\sim$95 Ma), the terranes (Alexander, Wrangellia) near the present coast tectonically overlapped the previously accreted terranes, mainly preserved in the Intermontane Plateau System, along a boundary located in the internal part of the Pacific Mountains System.

Compressional structures in the remainder of the northern Cordillera can be directly related to those in the two tracts of high-grade metamorphic rocks (Figure 4). External parts of the Rocky Mountains System contain Cretaceous and Paleocene thrust faults that carried miogeoclinal and older deposits for distances of up to 200 km over the continental platform. In the Pacific Mountains System, Late Cretaceous to Holocene folds and thrust faults mostly verge oceanward. In the Intermontane Plateau System in Canada, oceanward-verging folds and thrust faults are Jurassic, whereas those directed towards the continental interior are mainly Cretaceous. Convergent deformation is contemporaneous with uplift, erosion, and deposition in marine to non-marine basins that are both internal and external to the Cordillera (Figure 4).

In western Alaska, Cretaceous and Palaeogene convergence between the North American and Eurasian Plates may have driven the 40–60° counterclockwise rotation of south-western Alaska. Rotation, combined with dextral displacement on major faults in Intermontane Alaska, created the curvilinear grain of the Alaskan Cordillera and its expanded western termination.

Late Cretaceous, Tertiary, and Holocene dextral strike-slip faults lace the northern Cordillera and apparently reflect oblique subduction and partial coupling of the western North American Plate margin to the northward-moving Pacific Plate and to its precursor, the Kula Plate, which had been entirely subducted by $\sim$37 Ma (Figure 4). The total amount of displacement on these faults, as determined from geological features offset across them, approaches 1000 km. In the north, there are displacement amounts of 425 km on the Tintina Fault and up to 400 km on the Denali Fault, and, in the south, about 300 km on combined Fraser and Yalakom Faults. A major unresolved discrepancy exists between the above numbers and those arising from palaeomagnetic studies on Late Cretaceous rocks ($\leq$85 Ma), which suggest dextral displacements of as much as 3500 km for parts of the western Cordillera.

In addition to the above, a sparse record of Early Cretaceous sinistral, Early to Middle Jurassic dextral, and Triassic–Jurassic sinistral strike-slip faulting was possibly responsible for the fragmentation, displacement, and duplication of arc terranes in the collage prior to their final accretion and further disruption by dextral faulting. In the Pacific Mountains System, a Middle Jurassic–Early Cretaceous island arc founded on Wrangellia and Alexander terranes was apparently displaced southward in Early Cretaceous time. There is about 800 km of overlap between it and the northern end of the contemporary continental arc in the southern Intermontane Plateau and internal Rocky Mountains Systems (Figure 4). In addition, Early Mesozoic oroclinal bending, or left-lateral strike-slip faulting, has been invoked to account for apparent duplication of a Quesnel–Stikine Arc and enclosure of the associated Cache Creek Accretionary Complex (Figure 3).

Palaeogene (58–40 Ma) normal faults are widespread south of latitude 55° N in the Intermontane System and in parts of the Pacific and Rocky Mountains Systems. Many faults are oriented north–north-easterly or north-easterly, at angles to the prevailing

general grain of this part of the Cordillera, and some clearly are associated with movement on dextral strike-slip faults. Normal faults bound graben and half-graben structures containing coeval clastic and volcanic rocks, and many, especially in south-eastern British Columbia and adjoining parts of Idaho and Montana, contributed to the tectonic exhumation of formerly deeply buried rocks in metamorphic core complexes, in some of which Palaeoproterozoic continental basement is exposed.

Economic Deposits

Hydrocarbon deposits occur mainly in the Rocky Mountains System and include oil and gas originating from Palaeozoic and Early Mesozoic strata, and coal in foreland basins. Coal was formerly exploited in Late Mesozoic and Early Tertiary basins in the western Cordillera, and oil and gas are known in rocks of the Pacific continental shelf.

Metallic mineral deposits occur widely. Interior parts of the Rocky Mountains System contain widespread lead, zinc, and silver deposits, two of which, the former Sullivan Mine in south-eastern British Columbia and the active Red Dog mine in north-western Alaska, are some of the largest known. The Intermontane Plateau and Pacific Mountains Systems contain several large copper and molybdenum deposits, with associated golds in porphyry systems in arc-related plutons. Placer gold deposits were of historical importance, but currently are less so.

Evolution of the Northern Cordillera

The origins of the northern Cordillera extend back some 770 Ma to rifting events that marked initial stages in the break-up and dispersal of components of the Rodinian supercontinent. By Early Cambrian time (~540 Ma), rifting had culminated in the seafloor spreading that opened the distant ancestor of the Pacific Ocean basin and created the Laurentian continent–ocean boundary that now is embedded in the Rocky Mountains System.

The intraplate continent–ocean boundary persisted until the Middle Devonian (~390 Ma), when it was succeeded by a convergent plate boundary that initially generated arc magmatic rocks within distal continental margin strata in a belt that extends from California to the Canadian Arctic. Arc magmatism has continued until the present, but has varied greatly in character and volume. From the Early Carboniferous until the Early Mesozoic (355–185 Ma), the period spanning amalgamation and eventual break-up of the supercontinent called Pangaea, the convergent plate boundary lay well offshore and featured island arc chains (Quesnel, Stikine) separated from the continental margin by back-arc basins (Slide Mountain). The region offshore western Pangaea probably was similar to the present western Pacific Ocean basin.

The accretion of arc, back-arc basin, and accretionary prism rocks and former continental margin fragments to the old continental margin started in the Jurassic at about 185 Ma. The accretion records the initiation of Cordilleran mountain building and coincided with the break-up of Pangaea and the advance of the North American continent towards the trench on its western margin. It resulted in the relocation of the continental margin some 500 km oceanward of its original position by the Late Cretaceous (~90 Ma), when the Cordillera probably resembled parts of the modern Andes.

During and following accretion, the collage and parts of the former continental margin were sliced and shuffled by strike-slip faults associated with oblique plate convergence, stretched and dismembered by extension and, in Alaska, apparently rotated counterclockwise. Most of the major physiographical features of at least the Pacific Mountains System, and probably much of the remainder of the northern Cordillera, probably emerged in the later Neogene ($\leq$12 Ma).

Appendix A: Northern Cordilleran Physiographical Systems and Underlying Bedrock Geology

The names Rocky Mountains, Intermontane Plateau, and Pacific Mountains Systems, used for the major physiographical divisions in Alaska, are approximately, but not completely, equivalent to the Eastern, Interior, and Western Systems used in British Columbia.

Features east and north of the Northern Cordillera: Interior Plains System, Arctic Coastal Plain, and Arctic Ocean

Rocky Mountains System
External part

Some major physiographical units: Montana and Canadian Rocky Mountains, Mackenzie Mountains, northern Brooks Range.

Bedrock: (1) Archaean and Palaeoproterozoic continental basement; (2) Mesoproterozoic (locally Palaeoproterozoic) clastic, carbonate, and minor magmatic rock associated with rift-related basins; (3) Neoproterozoic clastic, minor carbonate, and mafic magmatic rocks formed during rifting of the Rodinian supercontinent; (4) Cambrian to Jurassic carbonate and shale deposited on platform, shelf, and slope of Laurentian continental margin; and

(5) latest Jurassic to Palaeogene marine to non-marine clastics deposited in foreland basins concomitantly with folding and thrusting towards and over the continental platform.

Rocky Mountains and (in part) Tintina Trenches
Internal part

Some major physiographical units: Columbia, Cassiar, and southern Brooks Range.

Bedrock: (1) local Palaeoproterozoic continental basement; (2) Neoproterozoic rift-related clastic and volcanic rocks; (3) Palaeozoic, mainly deep-water, clastic rock, rift- and arc-related volcanic and plutonic rocks; (4) local Late Palaeozoic to Early Jurassic magmatic and sedimentary rocks formed in island arcs and marginal basins; (5) Middle Jurassic to Paleocene plutonic rocks; and (6) local Palaeogene magmatic and sedimentary rocks; rocks typically metamorphosed up to high grades and complexly deformed (mainly) during Middle Jurassic to earliest Tertiary compression, and also (in the south) by Palaeogene extension and (in the north) dextral strike-slip faulting.

Intermontane Plateau System

Some major physiographical units: northernmost Columbia Plateau, Interior Plateaus, Skeena Mountains, Yukon Plateaus, Yukon–Tanana Upland, Kuskowim Mountains, Coastal Lowlands, Seward Peninsula.

Bedrock: (1) variably metamorphosed Neoproterozoic and Palaeozoic continental margin deposits in Alaska and Yukon; elsewhere (2) Devonian to Early Jurassic sedimentary and magmatic rocks formed in island arcs, accretionary complexes, and back-arc basins; (3a) Middle Jurassic to Palaeogene continental magmatic arc rocks, and (3b) marine and non-marine clastic deposits eroded from tectonically uplifted regions; (4) local Neogene flood basalt; rocks deformed by compression mainly in the Mesozoic and extension–transtension in the Early Tertiary.

Pacific Mountains System
Internal part

Some major physiographical units: North Cascade Ranges (in Washington) and Cascade Mountains (in British Columbia), Coast Mountains, northern Alaska Range.

Bedrock: (1) dominant (~80%) Middle Jurassic to Palaeogene plutonic rocks (called the Coast Plutonic Complex); (2) Palaeozoic to Holocene volcanic and sedimentary rocks formed mainly in magmatic arcs, but locally in accretionary complexes; most rocks metamorphosed up to high grades and complexly deformed mainly in mid-Cretaceous to Palaeogene time.

Coastal depressions
External part

Some major physiographical units: Olympic Mountains, Insular Mountains, Saint Elias Mountains, Alaska Ranges, Aleutian Range, Aleutian Islands, continental shelf and slope.

Bedrock: (1) latest Proterozoic to mid-Cretaceous volcanic and sedimentary rock formed mainly in island arc and (in Triassic) oceanic plateau settings; (2) mid-Cretaceous and younger clastics eroded mainly from Coast and Cascade Mountains region; (3) Early Palaeozoic to Palaeogene plutons; (4) Early Jurassic to Holocene clastic-rich accretionary complexes in part underlying continental shelf and slope; most rocks folded and thrust oceanwards in Cretaceous to Holocene time.

Features west and south of the Northern Cordillera: Pacific Ocean here underlain by Juan de Fuca and Pacific oceanic plates

Appendix B: Summary Descriptions of Northern Cordilleran Terranes

The abbreviations in parentheses are used in Figure 3; most palaeogeographical references are made with respect to the North American continental interior.

Continental Margin Terranes

Displaced Neoproterozoic, Palaeozoic, and, locally, Early Mesozoic sedimentary rocks.

Arctic Alaska (AA): former North American continental margin deposits, locally with Devonian plutons, possibly rifted away by counterclockwise rotation of ~70° from its former position near the Canadian Arctic Islands in Early Cretaceous time (≤130 Ma).

Cassiar (CA): former North American continental margin deposits, sliced off and translated northwards by 425–2000 km on right-lateral strike-slip faults.

Coldfoot (CO), Kootenay (KO), Ruby (RB), Seward (SD), and Yukon-Tanana (YT): derived mainly from Neoproterozoic and Early Palaeozoic rift and lower slope deposits, and Middle Devonian to Early Carboniferous continental margin arcs; metamorphosed

and multiply deformed in Palaeozoic (?), Mesozoic, and Palaeogene times.

Dillinger (DL), Mystic (MY), and Nixon Fork (NX): small terranes in Intermontane Alaska whose stratigraphy and Early and Middle Palaeozoic faunas resemble those of the Siberian (or North Asian) craton margin.

Accreted Terranes

Arc terranes terminally accreted in the Jurassic

Quesnel (QN) and Stikine (ST): Devonian to Permian clastic rock and carbonate, arc-related volcanic, and plutonic rocks; local Palaeozoic ultramafic rocks, basalt, chert, pelite, and carbonate formed in marginal basins or ocean floor (=SM?); overlying Late Triassic and Early Jurassic arc-related magmatic and clastic rock and minor carbonate. Permian, Triassic, and Early Jurassic faunas most similar to those further south in the western conterminous USA and north-western Mexico; contrarily, palaeomagnetic data suggest little or no northward displacements. ST apparently duplicated QN in the Early Mesozoic by oroclinal bending or left-lateral strike-slip faulting and enclosed the associated accretionary complex (CC).

Arc terranes terminally accreted in the Cretaceous

Alexander Terrane (AX): latest Proterozoic and Early Palaeozoic magmatic arc and associated clastic rock and carbonate, and Late Palaeozoic and Early Mesozoic rift-related clastic and volcanic rock and carbonate. AX and WR apparently linked by Carboniferous plutons and overlapped by Middle Jurassic–Early Cretaceous arc rocks. Palaeomagnetic data suggest an Early Palaeozoic palaeogeographical location either near eastern Australia or the Barents Sea; Silurian and Devonian faunas favour the latter location; Triassic palaeolatitudes of AX comparable with those of Alaskan WR.

Cadwallader (CD) and Methow (MT): Upper Triassic magmatic arc rocks and associated sedimentary rock overlying Permian ocean floor rocks; unconformably overlain by Early Jurassic to Early Cretaceous marine arc-derived clastic and local volcanic rocks; overlying Late Cretaceous continental arc volcanics carry palaeomagnetic signatures indicating $\sim 35^\circ$ of northward displacement.

Chilliwack (CK): Devonian to Early Jurassic arc-related volcanic and clastic rocks and carbonate; in general, stratigraphy and faunas resemble ST and QN; possibly a fragment of ST, sliced off and translated southwards by sinistral strike-slip faults in the Early Cretaceous.

Koyukuk (KY), Nyac (NY), and Togiak (TG): Late Jurassic–Early Cretaceous arc rocks in western Alaska associated with the subduction of AG and GN oceanic terranes.

Wrangellia (WR): Devonian to Middle Jurassic magmatic arc rocks, with associated clastic rock and carbonate, and conspicuous Upper Triassic plume-related (?) tholeiitic basalt overlain by carbonate; Permian faunas comparable with some in Canadian Arctic and Urals; Early Jurassic faunas comparable with those elsewhere in the western Cordillera; palaeomagnetic data from Triassic strata on Vancouver Island indicate little or no latitudinal displacement, whereas those from southern Alaska indicate about 20° northward translation; interpreted as due to Tertiary separation of the two regions; WR linked to AX by Middle Jurassic and possibly Carboniferous.

Accretionary complex and back-arc basin terranes; ordered from continental interior to ocean Unless noted otherwise, all are characteristically highly disrupted and most contain melanges; pillow basalts, local alpine-type ultramafic rocks, and rare blueschists, in places associated with eclogite. Pre-Late Jurassic accretionary complexes in the Cordillera typically contain abundant radiolarian chert, whereas younger ones contain abundant clastic rocks including sandstone.

Angayucham (AG), Goodnews (GN), and Slide Mountain (SM): remnants of oceanic and marginal basins. SM is Carbonifereous to Triassic; AG and GN range in age from Devonian to Late Jurassic and appear to be accretionary complexes related to KY, NY, and TG arc terranes in western Alaska; SM in Yukon is associated with a Late Palaeozoic arc in YT and contains blueschist; further south, SM appears to be a back-arc basin, lacking melanges, eclogites, and blueschists, and separating QN from the former continental margin. The late Palaeozoic palaeomagnetic record of SM suggests northward displacement of $\sim 20^\circ$ with respect to the craton, in accord with displacements indicated by its Middle Permian fauna, which is similar to that in the south-western USA and north-western Mexico.

Cache Creek (CC) and Bridge River (BR): CC is Early Carboniferous to Early Jurassic and contains distinctive large masses of Late Carboniferous to Late Triassic shallow-water carbonate that probably formed as atolls capping seamounts, as do several similar complexes of this age in the circum-Pacific region. Middle Permian to Middle Triassic faunas

are similar to those in eastern Asia and the region occupied by Permian Tethys. CC contains Triassic blueschist and is interpreted as the accretionary complex accompanying QN and ST magmatic arcs. BR is similar to CC, but is of Early Carboniferous to Late Jurassic age, lacks distinctive large carbonate bodies, and was finally accreted in the Cretaceous.

Chugach (CG) and Pacific Rim (PR): CG is a long-lived Mesozoic accretionary complex. The inboard part of CG is a chert-rich assemblage that, in one locality, contains a limestone block with Permian Tethyan fauna and Early Jurassic blueschist; it is the accretionary complex associated with an Early Mesozoic arc on WR. The outboard part of CG is a Late Jurassic and Cretaceous clastic-rich assemblage with local Cretaceous blueschist; PR is similar but local; it is the accretionary complex accompanying the Middle Jurassic–Early Cretaceous arc founded on WR and AX.

Olympic (OL) and Prince of Wales (PW): Eocene to Oligocene; OL contains the youngest Cordilleran accretionary complex exposed on land; PW lies outboard of CG.

Yakutat (YA): Late Mesozoic accretionary complex overlain by Tertiary marine and continental clastic rocks; probably displaced northwards along the continental margin on the Queen Charlotte–Fairweather transform fault.

See Also

North America: Southern Cordillera. **Palaeomagnetism**. **Plate Tectonics**. **Tectonics:** Mountain Building and Orogeny. **Terranes, Overview**.

Further Reading

Armstrong RL (1988) Mesozoic and early Cenozoic magmatic evolution of the Canadian Cordillera. In: Clark SP Jr., Burchfield BC, and Suppe JE (eds.) *Processes in Continental Lithospheric Deformation, Geological Society of America, Special Paper 218*, pp. 55 91. Denver, Colorado: Geological Society of America.

Clowes RM and Hammer PTC (2000) Comparison of lithospheric structures across the Alaskan and Canadian Cordillera. In: *GeoCanada 2000, GAC MAC CGU CSEG CSPG Conference, Calgary, Alberta, Canada, Conference CD* Abstract 529, Calgary: Alberta.

Coney PJ, Jones DL, and Monger JWH (1980) Cordilleran suspect terranes. *Nature* 228: 329 333.

Engebretson DC, Kelley KP, Cashman HP, and Richards MA (1992) 180 million years of subduction. *GSA Today* 2: 93 100.

Enkin RJ, Mahoney JB, Baker J, Reisterer J, and Haskin ML (2003) Deciphering shallow magnetic inclinations: 2. Implications from Late Cretaceous strata overlapping the Insular/Intermontane superterrane boundary. *Journal of Geophysical Research* 108(B4): 2186.

Gabrielse H and Yorath CJ (eds.) (1991) Geology of the Cordilleran orogen in Canada. In: *Geological Survey of Canada, Geology of Canada 4,* and *Geological Society of America, The Geology of North America G 2,* Ottawa, Ontario: Geological Survey of Canada and Denver, Colorado: Geological Society of America.

Gehrels GE and Ross GM (1998) Detrital zircon geochronology of Neoproterozoic to Permian miogeoclinal strata in Alberta and British Columbia. *Canadian Journal of Earth Sciences* 35: 1380 1402.

Ghosh DK (1995) Nd Sr isotopic constraints on the interactions of the Intermontane Superterrane with the western edge of North America in the southern Canadian Cordillera. *Canadian Journal of Earth Sciences* 32: 1740 1758.

Nolkelberg WJ, Parfenov LM, Monger JWH, *et al.* (2000) *Phanerozoic Tectonic Evolution of the Circum North Pacific. United States Geological Survey Professional Paper 1626.* Menlo Park, CA: United States Geological Survey.

Oldow JS, Bally AW, Ave Lallement HG, and Leeman WP (1989) Phanerozoic evolution of the North American Cordillera: United States and Canada. In: Bally AW and Palmer AR (eds.) *The Geology of North America An Overview, Geological Society of America, The Geology of North America A,* pp. 139 232. Denver, Colorado: Geological Society of America.

Plafker G and Berg HC (eds.) (1994) *The Geology of Alaska. Geological Society of America, The Geology of North America G 1.* Denver, Colorado: Geological Society of America.

Samson SD and Patchett PJ (1991) The Canadian Cordillera as a modern analogue of Proterozoic crustal growth. *Australian Journal of Earth Sciences* 38: 595 611.

Scotese CR, Nokleberg WJ, Monger JWH, *et al.* (2001) *Dynamic Computer Model for the Metallogenesis and Tectonics of the Circum North Pacific. United States Geological Survey Open File* 01 26, version 1.0 (CD). Menlo Park, CA: United States Geological Survey.

Smith PL, Tipper HW, and Ham DM (2001) Lower Jurassic Amaltheidae (Ammonitina) in North America: paleobiogeography and tectonic implications. *Canadian Journal of Earth Sciences* 38: 1439 1449.

Southern Cordillera

A W Snoke, University of Wyoming, Laramie, WY, USA

Introduction

The Spanish term 'cordillera' refers to a series of parallel ranges or chains of mountains. The word was first applied in the western hemisphere to the mountain ranges of western South America, i.e., Las Cordilleras de los Andes. In the western United States, the Rocky Mountains and Coast Ranges and mountains between are collectively called the Cordillera. In Mexico, the cordillera includes the Sierra Madre Occidental, Del Sur, and Oriental systems, as well as the Sonoran Basin and Range Province and the Baja California peninsula (Figure 1). The focus in this article is on the southern half of the western North American Cordillera, including the mountain belt from ~47° N to ~16° N, an area that extends >4500 km. Thus this region is referred to as the 'Southern Cordillera'.

Diverse physiographic provinces are included in the Southern Cordillera. Where crossed by the Fortieth Parallel, it includes, from east to west, the Rocky Mountains, the Basin and Range Province, the Sierra Nevada, the Great Valley, and the Coast Ranges (Figure 1). Farther south, the Colorado Plateau is a prominent physiographic province between the Rocky Mountains and the Basin and Range Province, whereas in the north-western United States the Cenozoic volcanic provinces (Snake River Plain, Columbia Plateau, and Cascade Range) are prominent. An important topographic feature in the Southern Cordillera is the tract of high elevation that constitutes the Rocky Mountains and adjacent Colorado Plateau and Great Plains. Most of the 14 000-foot or greater mountain peaks in the conterminous United States are within the Southern Rocky Mountains, with the exception of Mt Whitney (14 494 ft; ~4421 m) in the Sierra Nevada and Mt Shasta (14 162 ft; ~4319 m) and Mt Rainier (14 410 ft; ~4395 m) in the Cascade Range.

Crustal thickness in the Southern Cordillera is strikingly variable. The Colorado Rocky Mountains and Sierra Nevada both exceed 50 km in thickness. The thinnest continental crust is in the north-western and north-eastern parts of the Basin and Range Province (~25–20 km) and around the northern part of the Gulf of California (Salton Trough), where the crust of south-eastern California and south-western Arizona has been rifted.

The fundamental nature of the Southern Cordillera is a Mesozoic and Cenozoic history as a convergent margin. The subduction of oceanic lithosphere along the Pacific margin and partial subduction of continental lithosphere in the interior of the orogen led to the development of accretionary complexes, magmatic belts, a mobile igneous and metamorphic core zone, and a foreland belt of thin- and thick-skinned contractional deformation. In contrast, the Late Cenozoic history of this circum-Pacific orogen is characterized by partitioning of deformation and magmatism into discrete magmatotectonic zones related to ongoing subduction of oceanic lithosphere, evolution of the San Andreas transform-fault system, north-westward migration of the Mendocino triple junction, and intraplate continental lithosphere extension.

Precambrian Framework

The nucleus of North America is the Precambrian craton, commonly referred to as 'Laurentia'. This Precambrian craton consists of seven Archaean microcontinents welded together along Palaeoproterozoic collisional orogenic belts. The Archaean Wyoming province is exposed in basement-involved, Laramide (Late Cretaceous and Early Palaeogene) uplifts of the Rocky Mountain foreland. The Wyoming province is bordered by Proterozoic orogenic belts on its northern, eastern, and southern margins, whereas its western margin, which extends into north-eastern Nevada and under the Snake River Plain, is a rifted margin formed during Neoproterozoic and Early Cambrian breakup of the Late Mesoproterozoic supercontinent 'Rodinia'. Along the southern margin of the Wyoming province, a Palaeoproterozoic geosuture (Cheyenne belt) is well exposed in the Medicine Bow Mountains of south-eastern Wyoming, and is interpreted as a Palaeoproterozoic collisional zone between a rifted continental margin (Archaean Wyoming province and overlying Palaeoproterozoic rocks) and a Palaeoproterozoic oceanic supra-subduction complex. This collisional orogenic event, the Medicine Bow Orogeny, developed during the interval 1.78–1.74 billion years ago (Ga). It was apparently the harbinger of a series of Palaeoproterozoic accretion events, which continued until ~1.65 Ga. This long history of lateral accretion of

Figure 1 Major physiographic provinces of the United States and Mexican Cordillera. Base map © 1992, Raven Maps & Images.

chiefly juvenile Palaeoproterozoic crust yielded a zone of accretion at least 1200 km wide, now extensively exposed in parts of Colorado, New Mexico, and Arizona. In the south-western United States, these accretion events have been traditionally referred to as the Yavapai and Mazatzal orogenies.

By ~1.6 Ga, a supercontinent (i.e., 'Nuna', after P F Hoffman) had been assembled that included Laurentia and other ancient elements that are now part of the Greenland and Baltic shields. This supercontinent experienced multiple magmatic and tectonic events in the next 500 million years. In the first event, a broad belt of Mesoproterozoic magmatism and associated tectonism developed across much of the supercontinent. In the Rocky Mountains, this magmatism is well represented by ~1.4-billion-year-old plutonic complexes that are widespread south of the Cheyenne Belt. Plutons of similar age are common in the Basin and Range Province of Arizona, in the Mojave Desert of south-eastern California, and in the subsurface of West Texas and eastern New Mexico. Near the end of the Mesoproterozoic, the North American Craton experienced both rifting (manifested by the Mid-Continent Rift System) and severe contraction, a product of the collisional Grenville Orogeny (~1.1–0.9 Ga). That orogeny culminated in assembly of the Late Mesoproterozoic supercontinent Rodinia, the reconstruction of which remains controversial.

Breakup of Rodinia, and Neoproterozoic to Mid-Palaeozoic Sedimentation

The breakup of Rodinia was followed by development of a thick wedge of Neoproterozoic through mid-Palaeozoic continental terrace deposits along the Cordilleran and Appalachian margins. In what is now the Canadian Cordillera, rifting began ~750 Ma and evolved into a Neoproterozoic passive margin where the Windermere Supergroup was deposited. In the Southern Cordillera, initial rifting was incomplete, and only during a second rifting event was the rift-to-drift transition achieved. Tectonic subsidence studies indicate that the drift phase in the US Cordillera did not occur until the latest Neoproterozoic or earliest Cambrian (~600–555 Ma). Nevertheless, Neoproterozoic clastic deposits, including glacial diamictite, can be traced from the Mackenzie Mountains in the Canadian Cordillera to the Death Valley region in southern California. The Neoproterozoic rocks in the Southern Cordillera are considered rift deposits that accumulated in isolated basins, but by the Early Cambrian, a continuous, Atlantic-type, passive margin existed virtually the length of the western North American Cordillera.

The Neoproterozoic deposits were the initial stratigraphic units of an enormous wedge of chiefly clastic and carbonate rocks that accumulated along the rifted, western margin of the Southern Cordillera through Late Devonian time. This wedge of continental margin deposits is often referred to as the 'Cordilleran Miogeocline', and it reached ~10 000 m in total thickness. The miogeocline is separated from a partially equivalent, but considerably thinner, cratonic sequence by the 'Wasatch Line', interpreted as a hinge line in the depositional framework of Southern Cordillera. This fundamental boundary is also interpreted as the eastern limit of Neoproterozoic, syndepositional faulting related to initial rifting of Rodinia in the Southern Cordillera. The Phanerozoic cratonic sedimentary sequence is characterized by disconformities and in some cases complete periods are unrepresented (e.g., the Silurian over most of Wyoming). Subsequently, the Wasatch Line represents an important boundary during Pennsylvanian–Permian basin development (e.g., Oquirrh Basin), coincides with a regional ramp during foreland fold-and-thrust belt development, and is the approximate eastern margin of the Cenozoic Basin and Range extensional province. Furthermore, a segment of the Intermountain seismic belt follows the Wasatch Line.

Another fundamental boundary in the Southern Cordillera is the $Sr_i = 0.706$ line, a boundary based on the initial $^{87}Sr/^{86}Sr$ ratio in Mesozoic and Cenozoic igneous rocks. This isotopic boundary has commonly been interpreted as the western extent of Precambrian basement rocks (Figure 2) and therefore also roughly correlates with the western extent of miogeoclinal sedimentary rocks.

Palaeozoic Orogenies

Passive margin sedimentation ended in Late Devonian time when Cambrian through Devonian oceanic rocks of the Roberts Mountains allochthon were thrust onto the continental shelf during the Antler Orogeny, best documented in north–central Nevada. Slope-and-rise sedimentary rocks and seafloor mafic volcanic rocks were thrust upon coeval, shallow-water shelf strata along a regional thrust fault (Roberts Mountains Thrust). Detailed palaeontological and stratigraphic studies have shown that the Roberts Mountains allochthon consists of fault-bounded packets of oceanic-facies rocks imbricated into a tectonic wedge. A great clastic wedge (Antler Flysch) was shed eastward from the resulting highlands into a broad foredeep that included much of eastern Nevada and extended into Utah. Upper Pennsylvanian and Permian limestone and clastic rocks unconformably overlie the Roberts Mountains

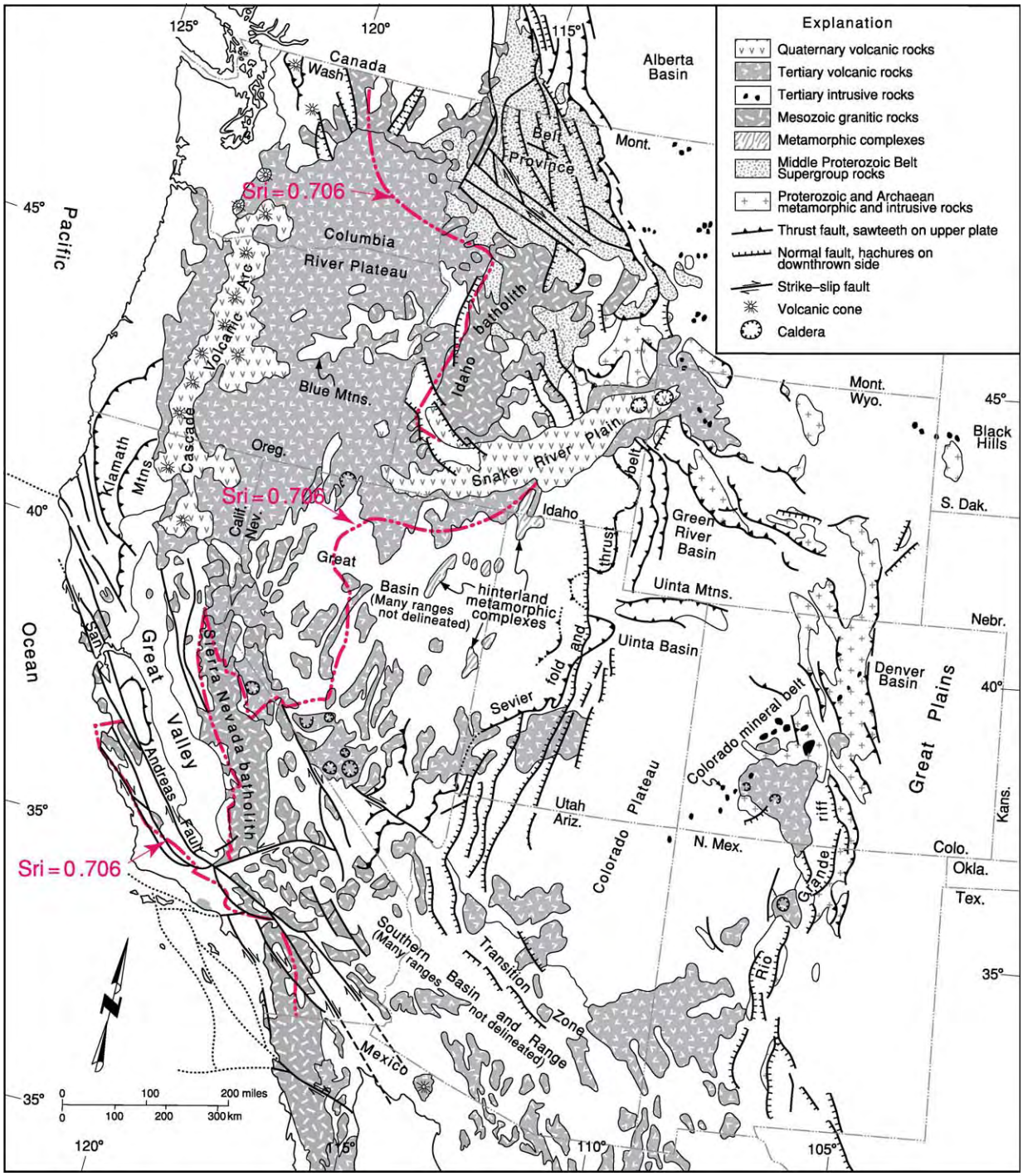


Figure 2 A tectonic map of the United States Cordillera, showing selected geological/tectonic features.

allochthon. Similar relationships occur to the north-east in the Pioneer Mountains of central Idaho and to the south-west in roof pendants in the Sierra Nevada batholith. Various plate tectonic models have been suggested for the Antler Orogeny and considerable debate has focused on the Antler Magmatic Arc that plays a role in virtually all of them. Some have argued that the Antler Arc subsided after collision with the western North American continental margin and was subsequently buried by younger rocks of later orogenic cycles. Others have stated that Devonian oceanic-arc rocks in the eastern Klamath Mountains and northern Sierra Nevada are the magmatic arc elements involved in the Antler Orogeny; however,

there is debate about the facing direction of the arc. One hypothesis is that the arc faced and migrated south-eastward during progressive rollback of the subducted slab. Still another model suggests that the arc faced westward, and the emplacement of the Roberts Mountain allochthon was related to the collapse of a back-arc basin.

Near the Permian–Triassic transition, a tectonic event similar to the Late Devonian Antler Orogeny occurred in the Southern Cordillera. The field relationships for this event are known best in north-western Nevada. The deep-water chert–argillite-limestone–greenstone of the Havallah sequence was thrust eastward onto autochthonous, shallow-water Upper Palaeozoic strata along the Golconda Thrust. This Golconda allochthon is composed of numerous fault-bounded slices of rock, such as the Roberts Mountains allochthon. This contractional orogeny is referred to as the 'Sonoma Orogeny', and again there is debate over the facing direction of the oceanic arc involved in this orogenic event. One school of thought interprets the tectonic setting for the Sonoma Orogeny as an oceanic arc–continental margin collision involving incipient subduction of continental crust beneath an east-facing oceanic arc (i.e., subduction directed westward). Others view the Sonoma Orogeny as another example of the collapse of a back-arc basin developed behind a west-facing oceanic arc (i.e., subduction directed eastward), remnants of which are preserved in the northern Sierra Nevada and eastern Klamath Mountains. An anomaly of the Sonoma Orogeny, unexplained by any model, is the fact that it did not create an extensive foreland basin.

Far to the east of the Antler and Sonoma orogenies, and approximately between them in time of origin, are structural and sedimentological effects related to the amagmatic, Pennsylvanian–Permian, intraplate Ancestral Rocky Mountains Orogeny. Manifestations of the Late Palaeozoic Ancestral Rockies are best demonstrated in the present-day Colorado Rocky Mountains and environs, but structural and sedimentological effects related to the Ancestral Rockies Orogeny can be traced from southern Oklahoma to northern Nevada. The uplifts supplied salmon pink to red arkosic sandstones, which grade into marine strata; these deposits are important components of the Late Palaeozoic stratigraphic section of the central Rocky Mountain region. The Ancestral Rockies Orogeny is commonly interpreted as an intraplate orogeny related to the Late Palaeozoic collision of the South American–African plates of Gondwana with the southern margin of Laurentia during the development of the supercontinent Pangaea. Orogenic effects of this continent–continent collision are manifested in the Marathon–Ouachita Orogeny in the south–central United States and are part of an extensive Late Palaeozoic orogenic system that can be traced from West Texas to central Europe.

Truncation of the Cordilleran Miogeocline and Pre-Cenozoic Strike–Slip Faulting along the South-western Margin of the Cordillera

In a 1969 synthesis of the plate tectonic evolution of California and environs, Warren Hamilton noted the apparent truncation of the south-west-striking Cordilleran miogeocline in southern California and north-western Mexico. Although Hamilton favoured Late Palaeozoic to Triassic truncation, subsequent studies have demonstrated that in Early to Middle Pennsylvanian time, the depositional framework in the Death Valley region experienced a fundamental change in orientation from north-east/south-west to north-west/south-east. This important change in orientation of Late Palaeozoic depositional facies in south-eastern California is interpreted as the result of a sinistral transform fault zone that was initiated during the Early or Middle Pennsylvanian and continued to be active into the Early Mesozoic. In Late Triassic time (Norian), this north-west/south-east strike was maintained during the initiation of the Cordilleran continental-margin magmatic arc that can be traced from southern Arizona into the eastern Sierra Nevada. Late Palaeozoic magmatism is recognized in eastern Mexico where it invades Gondwanan crust; whereas scarce Late Permian plutonic rocks occur in the western Mojave Desert and environs. How these igneous rocks relate to the development of the continental-margin magmatic arc is uncertain, but the occurrences in the Mojave Desert area suggest magmatism shortly after the establishment of a newly formed continental margin bounded by a transform-fault system.

Another widely cited tectonic hypothesis concerning major strike–slip displacement along the south-western margin of the Southern Cordillera postulates a major Late Jurassic, sinistral, transform-fault boundary designated the 'Mojave–Sonora Megashear'. This regional fault zone has been considered to be significant in the translation of part of northern Mexico (e.g., 'Caborca block') into its present position in the Southern Cordillera after the initial rifting and breakup of Pangaea and during the development of the Gulf of Mexico. However, the recognition that Palaeozoic depositional and structural trends were initially truncated in the Southern Cordillera in Late Palaeozoic time has dramatically reduced the potential significance of the Mojave–Sonora Megashear in reshaping the south-western

margin of the Southern Cordillera. Furthermore, detailed field studies in southern California have failed to locate the trace of the Mojave–Sonora Megashear in that region, although Late Jurassic (?) faults possibly related to it have been identified in Sonora, Mexico.

Possibly significant Early Cretaceous, dextral strike–slip faulting has been suggested in the wall rocks of the Sierra Nevada batholith; and younger, right-slip, crystal-plastic shear zones are known from the Sierra Nevada batholith. In one speculative model, a large-scale, Early Cretaceous dextral strike–slip fault system is hypothesized to have extended from the Mojave Desert region to the western margin of the Idaho batholith, suggesting significant northward translation of a large tract of the accreted terranes of the Western Cordillera. The region inferred to have been displaced northward following its accretion to western North America includes the Blue Mountains (north-eastern Oregon), Klamath Mountains (south-western Oregon and north-western California), and northern Sierra Nevada (Figure 1).

Early Mesozoic Continental to Oceanic Magmatic Arc

The Middle Pennsylvanian–Middle Early Triassic oblique truncation of the continental margin of the Southern Cordillera was subsequently overprinted by the development of a north-west/south-east-trending magmatic arc. It can be traced from southern Arizona, where it is built on continental (sialic) crust, to the eastern Klamath Mountains, where it is built on oceanic (ultramafic to mafic) crust. The transition from continental to oceanic arc is inferred to occur at about 39° N latitude in the Sierra Nevada. This Early Mesozoic magmatic arc is thought to have been west-facing with an eastward-dipping subduction zone. Late Triassic (~220 Ma) blueschist-facies metamorphic rocks in the northern Sierra Nevada, Klamath Mountains, and near Mitchell, Oregon (inlier of the Blue Mountains Province) are interpreted as the innermost subducted rocks of an accretionary complex that developed seaward of the magmatic arc. Its development initiated a long-lived convergent plate boundary zone along the Southern Cordillera. This west-facing magmatic arc served as the 'backstop' for the accretion of numerous tectonostratigraphic terranes that were added to the western North American continental margin from mid-Jurassic through Early Tertiary time.

Accreted Terranes

The 'terrane concept' was conceived and initially applied by W Porter Irwin (United States Geological Survey) to explain complex geologic relationships in the south-eastern Klamath Mountains of California. Subsequently, the concept has been utilized throughout the western North American Cordillera as well as in other orogens (e.g., the Appalachians and the Caribbean region). A tectonostratigraphic (or lithotectonic) terrane is an allochthonous, fault-bounded assemblage of rocks with a different geological history than that of adjacent rock units. Tectonostratigraphic terranes can consist of continental-margin features such as a displaced continental margin or part of an original fringing island arc and/or its subduction complex. Oceanic features such as plateaus, seamounts, or even back-arc basins may occur as discrete terranes in accretionary orogens. The tectonostratigraphic terranes in the western North American Cordillera are tectonic slices of such crustal elements and are not lithospheric or even crustal sections. The juxtaposition and amalgamation of such slices produced the Cordilleran 'collage' of terranes that characterizes the western part of the orogen.

In the Southern Cordillera, Phanerozoic tectonostratigraphic terranes are the basic 'building-blocks' of continental accretion west of the $Sr_i = 0.706$ line (i.e., the inferred western margin of Laurentia in the Southern Cordillera). Beginning in the late 1970s and extending to the present, many geological studies have focused on determining the affinity of various tectonostratigraphic terranes that are part of the Cordilleran Orogen. Clearly some of these terranes are 'pericratonic', having originated near the western margin of Laurentia, whereas other terranes are truly 'exotic' to the Cordillera. These far-traveled terranes represent major additions to western North America.

Ophiolites, sometimes dismembered, are an important component of some of the accreted, tectonostratigraphic terranes of the Southern Cordillera. In the Klamath Mountains of north-western California and south-western Oregon, well-preserved ophiolitic sequences range in age from Early Palaeozoic through Late Jurassic. The oldest are in the eastern parts of the Klamath Mountains and younger sequences are to the west. Some examples are the Trinity Subterrane, North Fork, Rattlesnake Creek, and Western Klamath terranes. The Josephine Ophiolite of the Smith River Subterrane (part of the Western Klamath Terrane) is the best-studied ophiolite sequence in western North America. To the north-east in the Blue Mountains of north-east Oregon, the Baker Terrane includes important ophiolitic sequences. In east–central California, ophiolitic sequences commonly occur as tectonic slices along the Foothills fault system of the western Sierra Nevada metamorphic belt. Many,

if not all, of these Cordilleran ophiolites developed in supra-subduction-zone settings, which indicate origins by rifting and spreading within oceanic arcs. This intra-arc extension may have been a result of oblique subduction and broad-scale transtension within the arc, similar to the present rifting and spreading in the Andaman Sea north of Sumatra. In contrast, the rifting and spreading may have been more orthogonal, such as the ongoing propagating rift and spreading centre related to the opening of the Lau Basin behind the Tofua (Tonga) Arc in the south-western Pacific. This supra-subduction interpretation of Cordilleran ophiolites implies proximity to, and temporal overlap with, oceanic-arc deposits, and has been demonstrated throughout the Cordillera by detailed geological mapping coupled with geochemical and geochronological studies. A particularly well-documented example of this relationship is the deposition of the Upper Jurassic Rogue and Galice formations and development of the Late Jurassic Josephine Ophiolite—all within the western Klamath Terrane. The Galice Formation, consisting of slaty metashales and metagreywackes with subordinate metaconglomerate and metavolcanic rocks, lies depositionally on both the Rogue Formation and Josephine Ophiolite. This situation suggests a close proximity in space and time between an oceanic arc (Rogue Formation), its adjacent sedimentary apron (Galice Formation), and the development of a complete ophiolite sequence (i.e., Josephine Ophiolite).

Another type of tectonostratigraphic terrane common in the western North American Cordillera is the accretionary complex terrane. Some of these terranes include ophiolitic components (e.g., the Baker Terrane in north-eastern Oregon), but most are either chert rich or clastic rich and are characterized by tectonic melange, broken formation, and/or olistostromal deposits. Fossiliferous rocks in some accretionary complex terranes provide some of the most reliable palaeogeographic data obtainable from tectonostratigraphic terranes. In the Cordillera, limestone blocks (of tectonic or olistostromal origin) yield fossils long recognized as 'exotic' to the western North American Cordillera. For example, Permian Tethyan fossils have been discovered in the Klamath and Blue Mountain provinces (North Fork, Rattlesnake Creek, and Baker terranes) and are especially characteristic of the Cache Creek Terrane of British Columbia, Canada. The significance of the Tethyan fauna in the Cordillera is still debated by palaeontologists. However, the association of these exotic limestones with blueschist-facies blocks strongly suggests that these terranes represent accretionary complexes that incorporated various rock types during the subduction of Pacific Ocean crust along the margin of western North America. The polarity of the original magmatic arcs related to these accretionary complexes is commonly controversial. Some accretionary complexes in the Southern Cordillera may be composites produced by collision of oppositely facing arcs. A possible example is the Baker Terrane (north-east Oregon) that apparently developed during Late Palaeozoic through Early Mesozoic time between the partly coeval Olds Ferry and Wallowa arcs. Such a situation is analogous to the arc–arc collision presently ongoing in the Molucca Sea between the oppositely facing Sangihe and Halmahera island-arc systems.

The largest accreted tectonostratigraphic terrane in the Southern Cordillera is the Guerrero Superterrane of western Mexico; its full geographic extent is still poorly known. The Guerrero Superterrane includes various volcanogenic terranes ranging in age from Late Jurassic to Early Cretaceous. They yield geochemical and isotopic data that suggest an intraoceanic arc setting. However, presence of a possible basement unit (the Arteage Complex with negative initial ε_{Nd} values) unconformably(?) overlain by the volcanogenic units suggests a composite history for the superterrane. In contrast, some tectonicists view the Arteage Complex as an underthrusted subduction complex overridden by the Guerrero Superterrane Arc. In many recent tectonic models, the Guerrero Superterrane is interpreted as an exotic, east-facing, intraoceanic arc accreted to mainland Mexico in mid-Cretaceous time. After collision with the Mexican continental margin, a polarity reversal occurred, with the development of a west-facing, continental-margin magmatic arc. In effect, it is the southern continuation of the Cretaceous arc of the Peninsular Ranges of southern California and northern Baja California. Some workers have argued that the Guerrero Superterrane arc is only part of an enormous, east-facing intraoceanic arc that included the Great Arc of the Caribbean. The latter is considered to be the source of the numerous allochthonous, arc-related terranes that rim the circum-Caribbean region from Cuba to Tobago and along the northern margin of South America.

Jurassic Magmatic and Tectonic Events

The Jurassic history of the western North American Cordillera is particularly complex and includes various events along the continental margin as well as orogenic effects within the interior of the continent. How these Jurassic orogenic processes interrelate is a major unresolved problem in Cordilleran tectonics. A classic Late Jurassic orogeny of the continental

margin is the Nevadan Orogeny originally defined by Eliot Blackwelder in 1914. Recent studies have suggested that the age of this orogeny can be tightly bracketed in the Klamath Mountains, at ~150 Ma. However, in light of the plate tectonic paradigm, coupled with recognition that terrane accretion is commonly progressive along a continental margin, the regional significance of a tightly defined orogenic event has been questioned. The classic definition of the Nevadan Orogeny is based on rock relationships in the western Sierra Nevada of California and the Klamath Mountains of Oregon, in comparison to the California Coast Ranges. These include folds and associated cleavage in Kimmeridgian–Oxfordian (Upper Jurassic) metasedimentary rocks of the Mariposa and Galice formations and the nearby association of relatively undeformed Tithonian (uppermost Jurassic stage) sedimentary rocks of the Knoxville Formation in the Coast Ranges. This classic definition was reinforced with the recognition that Early Cretaceous granitic plutons cut these structural features in the Sierra Nevada.

The plate tectonic setting of the Late Jurassic Nevadan Orogeny is especially controversial with respect to the inferred geological evolution of the western Sierra Nevada as compared to that of the Klamath Mountains. Many tectonicists have favoured a collision between a west-facing Early Mesozoic arc (i.e., an eastern arc) and an exotic, east-facing oceanic arc (i.e., a western arc) to explain the geological relations manifested in the western Sierra Nevada metamorphic belt. The accreted western arc would be separated from the eastern arc by a composite accretionary complex, in part manifested by the Palaeozoic and Triassic Calaveras Complex east of the Sonora fault. In the Klamath Mountains, structural relationships involving west-directed regional thrusting preclude the accretion of an exotic east-facing oceanic arc. In this light, the most widely accepted tectonic model for the Nevadan Orogeny in the Klamath Mountains is the collapse of a back-Arc basin behind a west-facing oceanic arc (Rogue–Chetco Arc). In this model, collapse could be related to increased coupling across the subduction system (i.e., the arc became contractional after an earlier extensional history). Resolution of these two contrasting tectonic scenarios has not been completely achieved, although some favour a polarity reversal of the western arc from east facing in the western Sierra Nevada to west facing in the Klamath Mountains. Such polarity reversals do occur along the strike of modern intra-oceanic arcs. However, this compromise requires drastically different tectonic settings for the deposition of the Galice and Mariposa formations, which have been recognized as probable stratigraphic and temporal equivalent units for over 100 years.

Cretaceous Palaeogeographic Belts and Transition to an Andean-Type Continental Margin

Late Cretaceous palaeogeography of western North America was dominated by three tectonic belts that extended virtually the length of the Cordillera: (1) an Andean-type volcanic–plutonic magmatic arc on the west (i.e., Cordilleran Magmatic Arc), (2) a thin-skinned fold-and-thrust belt on the east (initiated in mid-Cretaceous time), and (3) a hinterland (zone of metamorphism, magmatism, and deformation) (Figure 3). Another extensive feature of western North America during the Cretaceous was the Western Interior Basin. At its maximum development, the Western Interior Basin extended from the Arctic Ocean to the Gulf of Mexico.

Magmatism related to the development of the Cordilleran Magmatic Arc began in the Middle Triassic, but was only developed in the continental crust of Mexico, Arizona, and south-eastern California. From east–central California north, Late Triassic through Middle Jurassic magmatism is recorded in

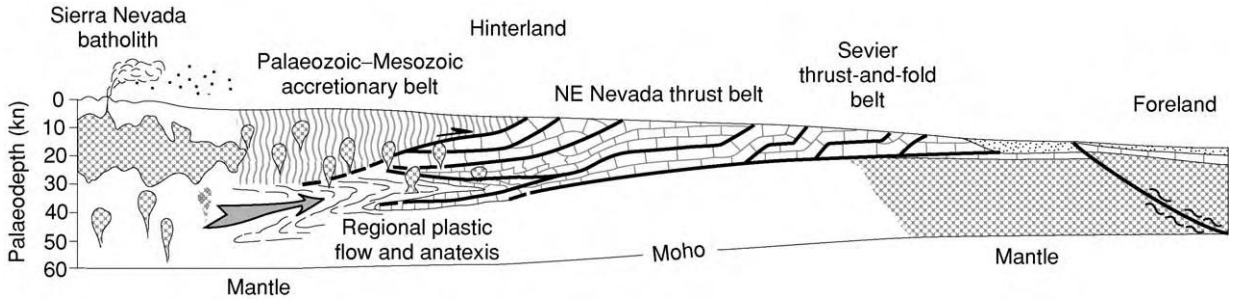


Figure 3 Schematic Late Cretaceous cross section of the western United States Cordillera, showing the relationship between the Sevier fold and thrust belt and fractured Rocky Mountain foreland on the east, metamorphic and igneous rocks of the hinterland, and Sierra Nevada batholith on the west. Reproduced (but modified) with permission from Miller EL and Gans PB (1989) Cretaceous crustal structure and metamorphism in the hinterland of the Sevier thrust belt, western U.S. Cordillera. *Geology* 17: 59 62.

tectonostratigraphic terranes not yet completely attached to the continental margin. During the Middle and Late Jurassic, magmatism spread eastward to include areas well within Laurentian crust (e.g., the Ruby Mountains of north-eastern Nevada). Also, during this time the Sundance Sea occupied an incipient foredeep in the Western Interior Basin and important terrane accretion took place in the western Sierra Nevada, Klamath Mountains, and Blue Mountains.

The geological history of the Southern Cordillera during the early part of the Cretaceous Period (i.e., Neocomian) is poorly understood. Exposures of Neocomian magmatic and sedimentary rocks are rare in the western North American Cordillera. The time interval ~140–125 Ma is commonly referred to as 'the Early Cretaceous magmatic lull'. Global sea-level was low during this time, and in the Rocky Mountain region and environs a regional unconformity exists between the chiefly fluvial deposits of the Upper Jurassic Morrison Formation and overlying mid-Cretaceous strata of the Western Interior Basin. In contrast, widespread Late Cretaceous magmatism in the Sierra Nevada and shortening in the foreland fold-and-thrust belt (i.e., Sevier Orogenic Belt of south-eastern Idaho, western Wyoming, and northern Utah) were synchronous with the culmination of seaway advancement. The fold-and-thrust belt is characterized by sled-runner thrust faults that are directed eastward and merge into a décollement near the top of the crystalline Precambrian basement. Major thrust systems are younger to the east, and the shortening history extends from ~119 to ~52 Ma. The Late Cretaceous culmination of magmatism in the Cordilleran Magmatic Arc (~100–85 Ma) coincides with intense shortening in the fold-and-thrust belt. One tectonic model argues for a close tie between Cretaceous magmatism in the California arc and continental lithospheric underthrusting beneath the arc (so-called A-type subduction). Although Sevier belt thrusting was initiated in mid-Cretaceous time, earlier (Jurassic) back-arc shortening is well documented in central Nevada (e.g., the Luning-Fencemaker fold-and-thrust belt).

While Late Cretaceous magmatism occurred in the Cordilleran Arc and shortening took place in the Sevier Orogenic Belt, regional metamorphism, large-scale folding, and the generation of peraluminous granitic rocks (commonly containing two micas, garnet, ±sillimanite) characterized the deep crust of the hinterland (Figure 3). The regional thrust faults of the Sevier Belt root into this igneous–metamorphic infrastructure, and remobilized Precambrian basement rocks were involved in both the plastic deformation and deep-crustal anatexis. Juvenile magmatic additions from the mantle did not play a significant role in the petrogenesis of the granitic rocks. Tectonic thickening coupled with dehydration melting during late decompression were the key processes responsible for metamorphism and anatexis in the Sevier hinterland. The deep-crustal rocks of the Sevier hinterland are exposed in windows developed during Tertiary crustal extension.

Laramide Orogeny

Distribution of igneous activity in the Cordillera of the western United States changed dramatically near the end of the Cretaceous Period. At ~75 Ma, magmatism in the Sierra Nevada batholith ended, and magmatism migrated eastward into the Rocky Mountains and beyond. This change in the geographic distribution of igneous activity in the Southern Cordillera is roughly synchronous with the Laramide orogeny of the Rocky Mountains. In the Rocky Mountain foreland, the Laramide Orogeny is characterized by the development of basement-involved uplifts and adjacent deep basins during Late Cretaceous through Early Eocene time (Figures 3 and 4). This 'thick-skinned' tectonic style is dramatically different from the contemporaneous 'thin-skinned' style of the Sevier fold-and-thrust belt. Although both deformations may be broadly related to plate tectonic processes near the continental margin, the local decoupling levels for these foreland deformation belts are different. The decoupling zone for the Sevier fold-and-thrust belt lies above the Precambrian basement except in the more interior (western) parts of the belt, whereas seismic-reflection studies show the decoupling zone for the basement-involved Laramide uplifts to be rooted into the deep crust.

The typical plate tectonic explanation for both the Late Cretaceous change in geographic distribution of Cordilleran magmatism and development of a thick-skinned tectonic style in the eastern Rocky Mountains involves a change to a shallow-dipping subduction mode. In this interpretation, subduction occurs at a shallow depth beneath the overriding plate in a wide segment of the orogen that extended from south-western Montana to central New Mexico (Figure 4). Why such segmentation of the subduction system would occur along the strike of the orogen is a major unanswered question in Cordilleran tectonics. The Andean orogenic system of South America is commonly proposed as a present-day analogue for the Late Cretaceous–Early Tertiary orogen of western North America. In the present Andean system, segmentation of the subduction system may be related to the character of the subducting plate; i.e., the presence of an oceanic plateau or aseismic ridge. The age of the subducted oceanic lithosphere may also be important

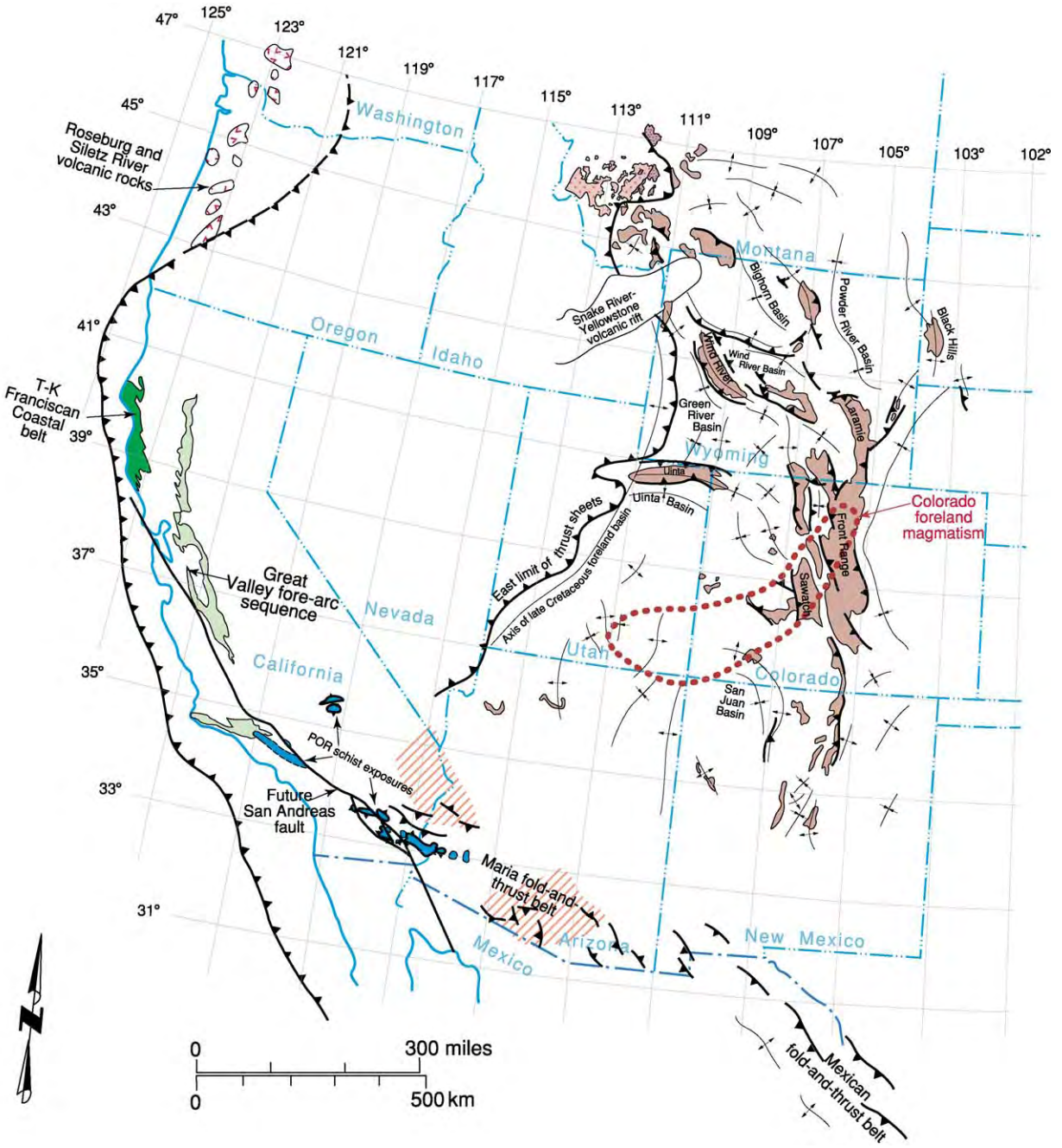


Figure 4 Distribution of the effects of the Late Cretaceous Early Palaeogene Laramide Orogeny in the south western part of the western North American Cordillera. T K, Tertiary Cretaceous; POR, Pelona Orocopia Rand; oblique red lines denote areas of late Cretaceous, high grade, regional metamorphism.

in the segmentation process, because young lithosphere is more buoyant than old, dense lithosphere. In the southern Andes, the absence of Quaternary volcanic centres in the western part of the orogen and development of basement-involved uplifts (Sierras Pampeanas) in the eastern part of the orogen coincide with the subduction of the Juan Fernandez Ridge and an eastward shift in subduction-related seismicity.

Additional evidence for a shallow-dipping subduction during Late Cretaceous–Early Palaeogene time in the Southern Cordillera is provided by a series of windows in southern California that expose oceanic rocks (Pelona, Orocopia, and Rand schists) beneath an upper plate of continental-affinity rocks. These windows are the product of Cenozoic deformation (both extensional and contractional) and provide

exceptional exposures of underthrusted, Late Cretaceous–Early Tertiary oceanic rocks that apparently underlie much of southern California and adjacent south-western Arizona.

Post-Laramide, Early Cenozoic Magmatic and Tectonic History

The post-Laramide Cenozoic history of the Southern Cordillera is characterized by new patterns of magmatism and tectonic strain. The throughgoing tectonic belts, which characterized Late Mesozoic time, were replaced by domains of extension, contraction, and strike–slip deformation. During mid-Eocene time, a broad belt of magmatism extended from southern British Columbia into central Idaho and north-western Wyoming, and a roughly contemporaneous zone of magmatism existed in southernmost Arizona and New Mexico and extended farther south into Mexico (Sierra Madre Magmatic Zone). These zones of Eocene magmatism were separated by a broad amagmatic corridor in the west–central United States that became the site of a large, Middle Eocene lake system. Accompanying the mid-Eocene magmatism in the Pacific north-west, metamorphic core complexes developed in areas of large-magnitude crustal extension. Typical complexes are characterized by a hanging wall of upper crustal rocks, sometimes including syntectonic volcanic and sedimentary deposits, separated from a footwall of mid-crustal igneous and metamorphic rocks by a plastic-to-brittle, normal-sense shear zone. Younger rocks are commonly structurally emplaced on older rocks, and brittle deformation features, including low-angle detachment faults, are superposed on the crystal-plastic deformation of the normal-sense, mylonitic shear zone. This northern belt of magmatism and accompanying localized, large-magnitude crustal extension migrated southward in Late Eocene and Early Oligocene time. Initiation of core-complex development in the eastern Great Basin was later than in the Pacific north-west, and large-magnitude crustal extension continued into the Early Miocene in the Ruby–East Humboldt and Snake Range core complexes (eastern Nevada). During the Late Eocene through Early Miocene, enormous amounts of volcanic ejecta erupted as ash-flow tuff sheets in the Great Basin (Nevada and western Utah). This 'ignimbrite flare-up' has significant implications for the crustal composition of the Great Basin, including substantial mafic magmatic intra- or underplating of the extended crust of the region. Volcanic ash from these enormous eruptions spread eastward in the upper atmosphere and formed a conspicuous air-fall component in post-Laramide, Late Eocene to Miocene strata of the Rocky Mountains (especially in Wyoming and environs). By Early Miocene time, the northern and southern magmatic zones had merged, and a continuous Neogene magmatic arc could be traced from the early Western Cascades arc into the Mojave–Sonoran Volcanic Zone. Numerous examples of Miocene core-complex development are well documented from south-eastern California across southern Arizona and into Sonora, Mexico. Still younger examples of core-complex development (Late Miocene to Pleistocene) are present in other areas of large-magnitude extension in Southern Cordillera, such as the ongoing rifting of continental crust in the northern Gulf of California, Mexico, and in the Salton Trough, California.

Elevations on the Colorado Plateau range from approximately 1.5 to 3.5 km, with the highest elevations typically associated with igneous centres such as the San Francisco volcanic field in northern Arizona. In deep canyons (e.g., the Grand Canyon), the elevation is considerably less than 1.5 km. The average elevation of the plateau is ~2 km, and the crustal thickness is ~45 km. Stratified rocks exposed on the plateau indicate that the area was near sea-level for much of the Phanerozoic and that uplift occurred after the deposition of Upper Cretaceous marine sedimentary rocks. The western and southern margins of the plateau are delineated by normal-fault systems related to the Basin and Range Province, whereas its northern and eastern margins merge into the eastern Rocky Mountains. The south-eastern margin of the Colorado Plateau in central New Mexico is delineated by normal faults related to the Rio Grande rift. The processes that facilitated uplift of the Colorado Plateau remain controversial, as well as the age or ages of uplift. One tectonic model relates the uplift of the plateau to eastward, intracrustal flow toward the Colorado Plateau from the overthickened, Sevier hinterland (now part of the Basin and Range Province). Another model argues for uplift related to lithospheric attenuation as a by-product of shallow-dipping subduction associated with the Laramide Orogeny. Still other tectonic models favour a polyphase uplift history: initially during the Laramide Orogeny and subsequently in the Late Cenozoic as part of a regional uplift, including the Southern Rocky Mountains and Great Plains. Clearly, the cause of the uplift of the Colorado Plateau remains a major unresolved problem in Southern Cordilleran tectonics.

As the San Andreas Fault (transform) system developed off the west coast of Mexico, and the triple junction between the North American, Pacific, and Juan de Fuca plates (Mendocino triple junction) migrated north-westward, the Neogene magmatic arc was shut off at its south end. In a broad area east of the late Western Cascade Arc, the Columbia River

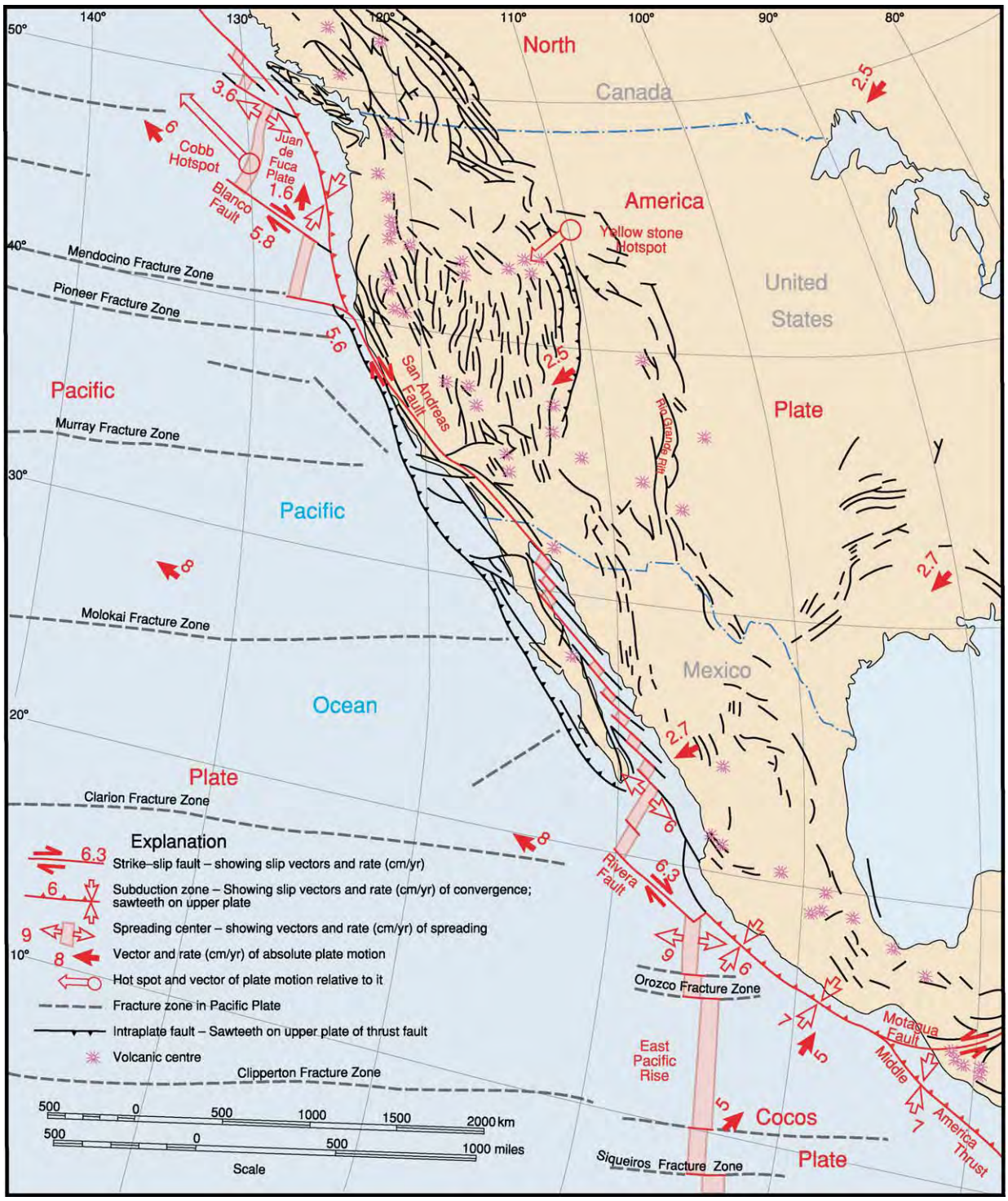


Figure 5 Present day plate tectonic setting of the southern segment of the Cordilleran Orogen of western North America and adjacent north eastern Pacific Basin. Reproduced (but modified) with permission from Drummond KJ, *et al.* (1982) *Pacific Basin sheet of plate tectonic map of the circum Pacific region (scale 1:20 000 000)*. Tulsa, OK: American Association of Petroleum Geologists.

flood basalt province developed between ~17 and 14 Ma. Basaltic dike swarms of this age are present in north–central Nevada and indicate the initiation of rifting in that area that eventually produced the Late Cenozoic Basin and Range Province.

Late Cenozoic Tectonic/Volcanic Systems and Seismicity

The Late Cenozoic tectonic evolution of the Southern Cordillera is dominated by four large-scale, tectonic/volcanic systems: (1) the San Andreas transform Fault and north-westward migration of the Mendocino triple junction, (2) the crustal extension in the Basin and Range Province and Rio Grande rift, (3) the north-east propagation of the Yellowstone hotspot track and concurrent north-west-propagation of the Newberry volcanic trend, and (4) the Cascadia subduction zone and volcanic arc. The development of these features was in part concurrent, and elements of all four systems are presently active. A relationship between the San Andreas transform-fault system and crustal extension in the Basin and Range Province has been suggested in several analyses of the Late Cenozoic tectonic history of the Southern Cordillera. However, effects associated with plate boundary slip are significant only in the south-western Basin and Range Province (e.g., eastern California shear zone). Late Cenozoic, regional crustal extension characteristic of the Basin and Range Province reflects the removal of the Farallon slab and subsequent asthenospheric upwelling coupled with north-west retreat of the Pacific Plate with respect to the interior of the North American Plate. At ~5 Ma, the Baja California peninsula was rifted from mainland Mexico when the San Andreas fault system shifted inland from a former position on the Pacific side of the peninsula. The north-western separation of Baja California from the mainland is an example of the interplate transfer of continental lithosphere.

The present-day seismicity of the Southern Cordillera is concentrated in belts at least partly related to fundamental plate tectonic boundaries (Figure 5); these boundaries are either strike–slip or convergent. Off the western coast of Oregon, seismicity is related to the Cascadia subduction zone and the Blanco Fracture Zone. Farther south, in western California and offshore, the Mendocino triple junction and San Andreas Fault system are the loci of present-day seismicity. Still farther south along the western coast of Mexico, seismicity is related to convergence associated with the Middle America Thrust.

Other prominent belts of seismicity are within the North American Plate and include the eastern California shear zone and its extensions into western Nevada (e.g., Walker Lane). The seismicity associated with this intraplate tectonic zone is interpreted as distributed strain associated with the plate boundary zone. Still another prominent zone of intraplate seismicity is the Intermountain seismic belt. It can be traced from southern Utah into western Montana and is particularly prominent in the Yellowstone National Park area. This belt of seismicity reflects a combination active normal faulting along the eastern margin of the Basin and Range Province (e.g., Wasatch front) as well as magmatism related to the Yellowstone hotspot.

See Also

North America: Northern Cordillera. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges.

Further Reading

Armstrong RL and Ward PL (1993) Late Triassic to earliest Eocene magmatism in the North American Cordillera: implications for the western interior. In: Caldwell WGE and Kauffman EG (eds.) *Evolution of the Western Interior Basin, Geological Association of Canada Special Paper 39,* pp. 49 72. St John's, NL: Geological Association of Canada.

Burchfiel BC, Lipman PW, and Zoback ML (eds.) (1992) *The Cordilleran Orogen: Conterminous U.S. The Geol ogy of North America*, vol. G 3. Boulder, CO: Geological Society of America.

Coney PJ, Jones DL, and Monger JWH (1980) Cordilleran suspect terranes. *Nature* 288: 329 333.

DeCelles PG (2004) Late Jurassic to Eocene evolution of the Cordilleran thrust belt and foreland basin system, west ern U.S.A. *American Journal of Science* 304: 105 168.

Dickinson WR (2000) Geodynamic interpretation of Paleo zoic tectonic trends oriented oblique to the Mesozoic Klamath Sierran continental margin in California. In: Soreghan MJ and Gehrels GE (eds.) *Paleozoic and Triassic Paleogeography and Tectonics of Western Nevada and Northern California, Geological Society of America Special Paper 347,* pp. 209 245. Boulder, CO: Geological Society of America.

Dickinson WR (2002) The Basin and Range province as a composite extensional domain. *International Geology Review* 44: 1 38.

Dickinson WR and Lawton TF (2001) Carboniferous to Cretaceous assembly and fragmentation of Mexico. *Geological Society of America Bulletin* 113: 1142 1160.

Drummond KJ, *et al.* (1982) *Pacific Basin sheet of plate tectonic map of the circum Pacific region (scale 1:20 000 000)*. Tulsa, OK: American Association of Petrol eum Geologists.

Hamilton W (1969) Mesozoic California and the underflow of Pacific mantle: *Geological Society of America Bulletin* 80: 2409 2430.

Hamilton W (1978) Mesozoic tectonics of the western United States. In: Howell DG and McDougall KA (eds.) *Mesozoic*

Paleogeography of the Western United States, Pacific Coast Paleogeography Symposium 2, pp. 33 70. Los Angeles, CA: The Pacific Section, Society of Economic Paleontologists and Mineralogists.
Miller EL and Gans PB (1989) Cretaceous crustal structure and metamorphism in the hinterland of the Sevier thrust belt, western U.S. Cordillera. *Geology* 17: 59 62.
Moores EM (1998) Ophiolites, the Sierra Nevada, "Cordil leria," and orogeny along the Pacific and Caribbean margins of North and South America. *International Geology Review* 40: 40 54.
Oldow JS, Bally AW, Avé Lallemant HG, and Leeman WP (1989) Phanerozoic evolution of North American Cordil lera: United States and Canada. In: Bally AW and Palmer AR (eds.) *The Geology of North America An Overview, The Geology of North America*, vol. A, pp. 139 232. Boulder, CO: Geological Society of America.
Sedlock RL, Ortega Gutiérrez F, and Speed RC (1993) *Tectonostratigraphic Terranes and Tectonic Evolution of Mexico, Geological Society of America Special Paper 278*. Boulder, CO: Geological Society of America.
Severinghaus J and Atwater T (1990) Cenozoic geometry and thermal state of the subducting slabs beneath western North America. In: Wernicke BP (ed.) *Basin and Range Extensional Tectonics near the Latitude of Las Vegas, Nevada, Geological Society of America Memoir 176*, pp. 1 22. Boulder, CO: Geological Society of America.
Wernicke BP, Christiansen RL, England PC, and Sonder LJ (1987) Tectonomagmatic evolution of Cenozoic exten sion in the North American Cordillera. In: Coward MP, Dewey JF, and Hancock PL (eds.) *Continental Ex tensional Tectonics, Geological Society of London Special Publication 28*, pp. 203 221. London: Geological Society of London.

Ouachitas

K C Nielsen, The University of Texas at Dallas, Richardson, TX, USA

Introduction

The Ouachita orogenic belt extends approximately 2100 km along the southern margin of North America (Figure 1). These Palaeozoic sedimentary rocks are associated with the rifted cratonic margin developed during Late Proterozoic to Early Cambrian time. Two promontories, Alabama and central Texas, and two re-entrants, Arkansas/Oklahoma and West Texas, characterized this margin. The Ouachita Mountains of Arkansas/Oklahoma and the Marathon uplift of West Texas developed in these re-entrants. While stratigraphic evidence suggests some regional adjustments coeval with the earlier stages of Appalachian tectonics, the Early Palaeozoic succession of the Ouachita orogen primarily records a long period of stable carbonate deposition along the continental shelf and slow deep-water clastic sedimentation off the shelf. This pre-orogenic phase ended abruptly in the Lower Carboniferous. At that time, the Ouachita basin began subsiding rapidly in response to closure of the Iapetus Ocean and the assembly of Pangaea (*see* **Pangaea**). A thick succession of mud-rich turbidites filled the basin, beginning at various times during the Early Carboniferous (Mississippian). The oldest deformation of this sequence is inferred from the Mississippian clastic wedge of the Black Warrior basin. Additionally, some authors have suggested that early folding documented in the metasediments of the uplifted Ouachita facies may correlate to early subduction in the Ouachita basin. The Late Carboniferous (Pennsylvanian) synorogenic sediments continued filling the basin and were translated inboard over the coeval shelf sequence, forming a generally north-verging fold and thrust belt. Subsidence within the associated foreland basins suggests a diachronous character to this advancing thrust system, beginning in the Late Mississippian in the east and in the Middle Pennsylvanian in the foreland of Oklahoma. Stratigraphic constraints indicate that thrusting ended in the Late Pennsylvanian in Oklahoma while thrusting continued in West Texas until the Lower Permian. The classic Permian carbonate succession unconformably overlies the deformed Marathon sequence. During the later stages of the deformation, basement cored uplifts developed in the Ouachita Mountains of Arkansas and Oklahoma, creating large north-east-trending uplifts. Timing of these uplifts is debateable, but suggests post-thrust deformation and a significant shift in regional shortening, possibly coincident with the later stages of the Alleghanian orogen of the southern Appalachian Mountains.

Unconformable Mesozoic units overlie approximately 80% of the Ouachita Orogenic Belt. Two large exposures, the Ouachita Mountains of Oklahoma and Arkansas and the Marathon uplift of West Texas, provide the bulk of our knowledge about the belt (Figure 1). Geophysical and well data provide the remainder of the database. Two important references serve as the foundation for discussion about the Ouachita orogenic belt and the

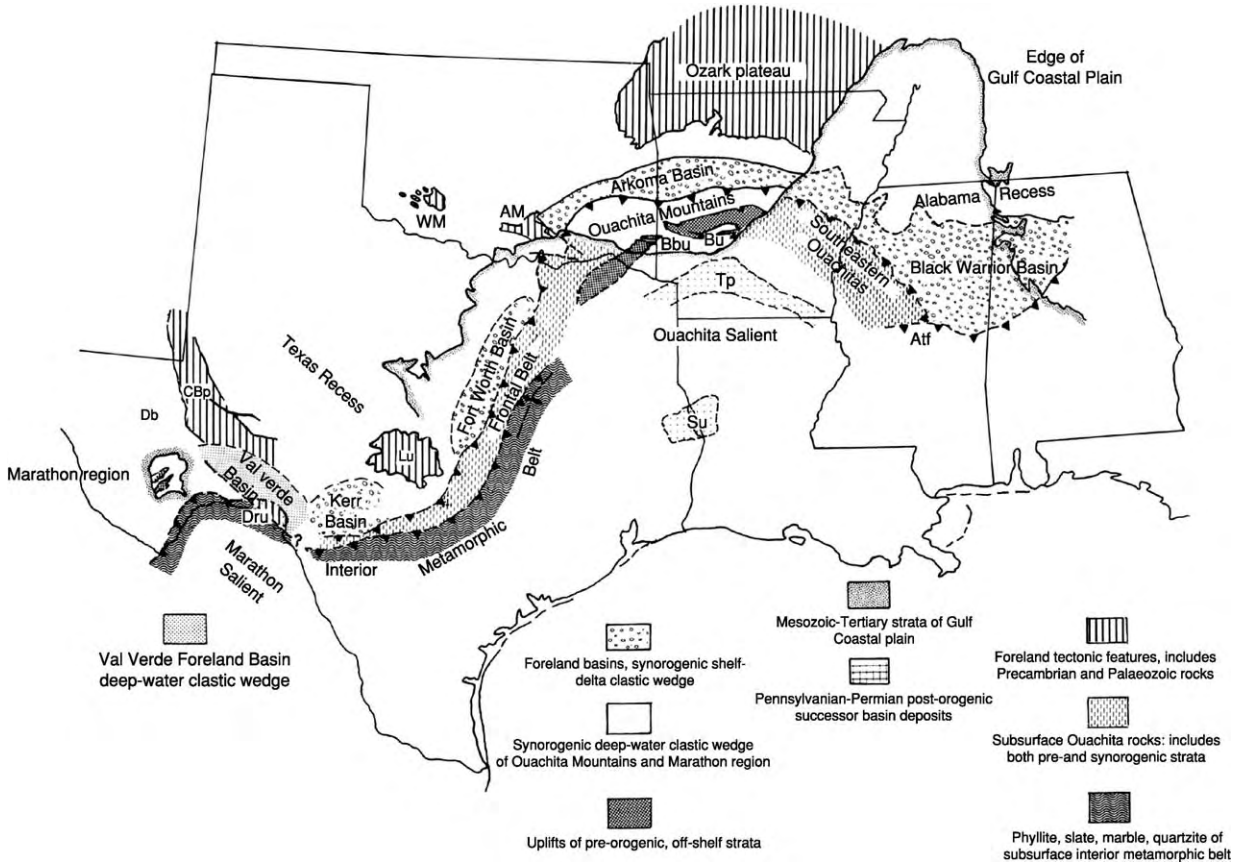


Figure 1 Tectonic elements of the Ouachita Orogen. Abbreviations (AM Arbuckle Mountains, Atf Appalachian Tectonic Front, Bbu Broken Bow Uplift, Bu Benton Uplift, CBp Central Basin Platform, Db Delaware Basin, Dru Devils River Uplift, Tp Texarkana Platform, Su Sabine Uplift, WM Wichita Mountains). (After Viele and Thomas, 1989.)

interested reader should consult these. The seminal work of Flawn and his colleagues in 1961 provided a complete review of the mountain belt and has been a cornerstone for all subsequent work. 28 years later, the Geological Society of America published a thorough update of the Ouachita Orogen. This discussion is synthesized from these sources.

Cambrian Continental Margin

The trace of the Ouachita Orogen is irregular (Figure 1) and this spatial distribution is directly related to the Late Proterozoic–Early Cambrian rifting of the North American Craton. From central Alabama and the junction with the Appalachian Mountains, the fold and thrust belt extends WNW across the Mississippi embayment into Arkansas and Oklahoma. There the trend shifts to the SSW, extending from Oklahoma into central Texas. At that point the trend shifts westerly continuing into West Texas, only to turn south again near the Marathon Uplift. The mountain belt extends a few hundred kilometers into Mexico before it is obscured by a Mesozoic tectonic overprint. This orthogonal pattern of recesses and salients reflects the rifting geometry along the southern margin of North America during the Late Precambrian and Early Cambrian. The spreading ridge for this rifting is inferred to be subparallel to the east coast of North America and to the south-trending margin in Texas (Figure 1). The connection between Alabama and Oklahoma and between central and West Texas are considered to have been transform boundaries. The Southern Oklahoma aulacogen is coeval with this rifting event and continues the WNW trend of the transform boundary into the craton. Additional Late Precambrian intracratonic rifting has been documented in the Mississippi Valley; the grabens trend NNE from the Ouachita Front through eastern Arkansas and western Tennessee. Igneous rocks exposed in the southern Oklahoma aulacogen reveal a 570–525 Ma bimodal volcanic package consistent with the rift model. Approximately 1 km of Cambrian clastic units in the Mississippi Valley graben indicate additional regional extension. However, outside these rift zones, including the Ouachita Orogen, there is little evidence of this rift sequence.

Stratigraphy

A fairly consistent lithostratigraphic succession characterizes the Palaeozoic sequence of the Ouachita Orogen. These sediments were deposited in two distinct phases. A pre-orogenic package consists of carbonate shelf sediments north and west of the continental margin and a deep-water shale, sandstone, and chert sequence on the continental slope or rise south and east. The second phase is dominated by syn-orogenic turbidites filling the basin and fluvial-deltaic sediments covering the shelf.

Pre-Orogenic Sequence

A classic carbonate shelf sequence accumulated on the craton over the Precambrian basement during the Early to Middle Palaeozoic (Figure 2). A thin basal sandstone unit is in contact with the Precambrian basement. This sandstone is overlain by a transgressive carbonate facies. In the Southern Oklahoma aulacogen, this carbonate succession is approximately 3.5 km thick and coeval sequences on the craton is approximately 1 km thick (Figure 2). Within the Arkoma Basin, this sequence is 1.2 km thick and about 2 km thick in the Black Warrior Basin. In many places, such as the Arkoma Basin, this pre-orogenic package records nearly continuous deposition and consistent thickness. The section in the Black Warrior Basin records influences of the Taconic Orogeny (Upper Ordovician) along the south-east margin of the basin and reveals more rapid deposition of the Devonian chert sequences along the south-west margin.

Within the Ouachita Basin, south and east of the rift margin, approximately 2–4 km of 'Ouachita facies' accumulated in generally deep marine settings (Figure 2). These shale, sandstone, and chert units currently form most of the exposures in the central uplifts of the Ouachita Mountains. Basinal pre-orogenic units have been subdivided into a lower shale-dominated section with two sandstone units followed by an upper black siliceous shale and chert succession with one interbedded sandstone (Figure 2). The change in lithology is interpreted as a function of progressively slower depositional rate and a similar subsidence history can be seen off the east coast of North America on the Blake Plateau. Current directions for the turbidites in the pre-orogenic sequence indicate that much of the deposition was coming from the north, but that the Silurian Blaylock Sandstone was sourced from the south-east: therefore the basin may have been two-sided during the pre-orogenic phase. The transition between shelf and deep water deposits is missing in the Oklahoma salient, due to subsequent shortening of the sequences. In the Marathon Region, there is a thinner pre-orogenic sequence that can also be described in two phases (Figure 2). The early units, in contrast to the Ouachita Mountains, include a lower turbidite, an overlying deep-water carbonate sequence (Marathon Formation) with overlying deep-water shale, calcarenite, conglomerates, and boulder beds. The upper phase is a chert and limestone sequence, similar in character to the chert sequence in the Ouachita Mountains. There has been significant debate about whether these cherts were deposited in shallow or deep-water settings. There is little argument about the fact that they record a long period of very slow sedimentation.

Syn-Orogenic Sequence

The tectonic picture changed dramatically in the Lower Carboniferous and this is reflected in the stratigraphic sequence. The subsidence rate and sediment input increased. The result was a rapid change up-section to deep-water shales and sandstones with interbedded debris flows. Within the Ouachita Basin, this contact appears to have been gradational, although some chert breccias have been found near the base of the section. These turbidite sequences filled the basin from east to west throughout the Carboniferous. Thickness estimates vary, but generally range from 12–14 km of section in the Ouachita Mountains where four major units are identified: Stanley and Jackfork Groups, Johns Valley, and Atoka Formations. The Stanley Group is the oldest of these flysch units, extending from the Middle to Late Mississippian. The characteristic lithologies include shale, sandstone, tuff, impure chert, and siliceous shale. Olistostromal deposits are frequent and several thin volcaniclastic units are found near the base of the Stanley Group around the Broken Bow Uplift. Gradationally overlying the Stanley Group, the Jackfork Group is characterized by prominent ridge-forming sandstone units. The Johns Valley Formation is distinguished from the underlying, gradational Jackfork Group by the presence of abundant exotic boulders and associated contorted shale. These boulders are correlated to the shelf sequence along the North American margin as well as the pre-orogenic Ouachita facies. These boulder beds are concentrated along the north-west margin of the thrust belt. The overlying unit is the widely spread Atoka Formation. Within the Ouachita Basin, the Atoka Formation is a typical turbidite, with nearly equal proportions of sandstone and shale. The maximum thickness of 8.5 km is found in the frontal Ouachita Mountains, but 2.5–4.9 km of coeval deltaic sandstones are found across the Arkoma Basin. The Atoka Formation thickens across regional, down-to-the-south normal faults, generated by tectonic loading of the continental

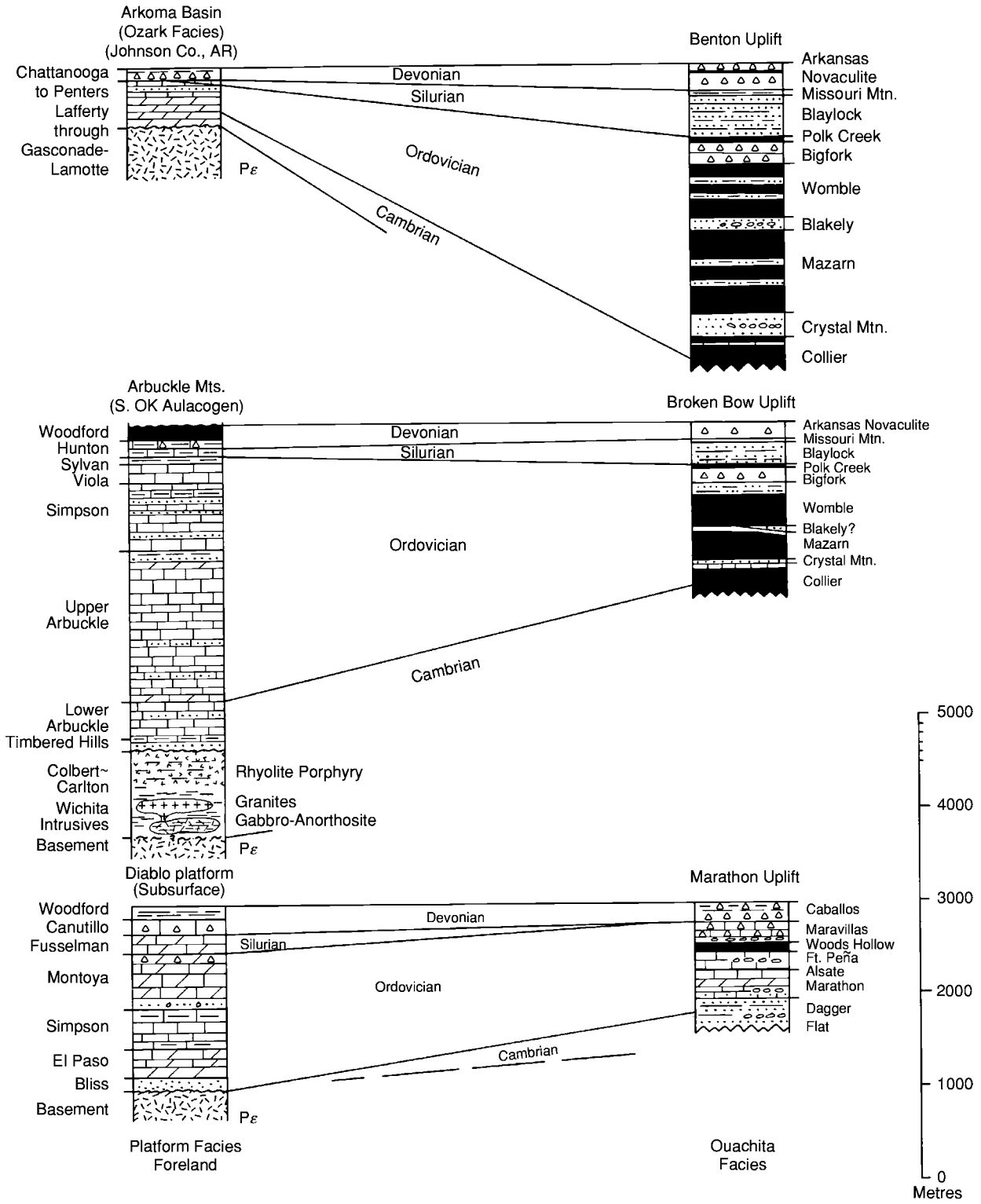


Figure 2 Representative lithological columns of the pre orogenic sequences at various settings along the Ouachita Orogen. (After Arbenz, 1989.)

margin. A northward migrating depositional centre characterizes the syn-orogenic sequence. The Stanley Group is 3.3 km thick just north of the Broken Bow Uplift. The Jackfork Group reaches a maximum thickness of 2.1 km along the axis of the Ouachita Mountains, and the Atoka Formation is thickest in the frontal Ouachita Mountains. While the entire section is allochthonous, these displaced depocentres are believed to be in relative order. Within the thrust belt, the Atoka Formation is the youngest deformed unit, but in the Arkoma Basin the Atoka Formation is overlain by additional deltaic deposits of Late Pennsylvanian Carboniferous age. The syn-orogenic sequence of the Marathon Uplift is also dominated by turbidite deposition but the sequence is thinner (Figure 3). The lower turbidite is 2.0 km thick in the eastern portion of the basin and thins to approximately 100 metres of shale in the west. There are abundant soft sediment deformational features in the section and these indicate extension parallel to the north-west trending palaeocurrents. The Tesnus Formation extends through the Mississippian into the Early Pennsylvanian and is overlain by the Dimple Limestone. This thin carbonate turbidite contains a clast assemblage of cratonic shelf rocks. The palaeocurrent data confirms a north-west source. The Haymond Formation is a repetitive siliciclasitc turbidite package that maintains a consistent 1.2 km thickness across the Marathon Basin. In this case, the palaeocurrents are trending WSW. The Haymond Formation includes important boulder beds that yield Early Palaeozoic rocks representing the pre-orogenic sequence and a full range of exotic rock types including various igneous clasts, some of which are Siluro-Devonian. These exotic clasts are commonly interpreted to have come from a southern source. The youngest unit in the Marathon Basin is Latest Pennsylvanian molasse type of sediments of the Gaptank Formation.

Regional Subdivisions

The Ouachita Orogen can be subdivided into four parts. The eastern section extends from central Mississippi north-westwards into Arkansas. This section is entirely in the subsurface. North of the tectonic front, a passive margin carbonate shelf of Cambrian to Early Mississippian age is overlain by Late Mississippian to Middle Pennsylvanian shallow marine to deltaic sediments. This sequence dips beneath the tectonic front which translates deep-water undifferentiated mudstones, cherts, and sandstones of the Ouachita facies.

The Ouachita Mountains represent the largest outcrop area along this orogenic belt (Figure 1). Similar to the pattern documented to the east, the carbonate shelf and overlying Carboniferous clastic sequences are structural below the allocthonous Ouachita facies. A minimal estimate for this translation is in the order of 100 km and maybe much greater. The Ouachita Mountains are commonly subdivided into the frontal imbrication zone between the Choctaw/Y-City and Windingstair faults, the central zone extending southward to the Broken Bow-Benton Uplifts, and then the Athens Plateau south of the uplifts. The imbrication zone is characterized by closely spaced thrust sheets made up of the younger Carboniferous units. The central zone, on the other hand, contains older Carboniferous units folded into a few long wavelength (8–16 km), north vergent synclines each truncated by a large thrust fault. The uplift areas trend ENE obliquely across the general east–west structural fabric of the Ouachita Mountains. These uplifts expose intensely deformed Ouachita facies and the Lower Stanley Group. The deformational history is more complicated than seen in the overlying central zone with at least three folding events and associated cleavage development recognized. Two distinctive aspects of these older rocks include a significant development of southerly verging folds and a disharmonic relationship to the large-scale folds of the central zone. The uplifts are also the location of low-grade metamorphism as reflected in recrystallized cherts (novaculite), chlorite growth, vitrinite reflectance, and ubiquitous quartz veins. The maximum temperature for these metasediments approach 300°C in the southern Broken Bow and eastern Benton Uplifts (Figure 4). The Athens Plateau has a structural style similar to the central zone to the north-west, northerly verging faults-truncating the Carboniferous sequence.

Within Texas, the Ouachita Orogen has been traced almost 1000 km in the subsurface. The structural subdivisions of the Ouachita Mountains have been extended (Figure 1). In this case, the 'frontal zone' includes all of the unmetamorphosed northerly verging structures inboard of the uplifts. The 'interior' zone is correlated to the uplifts of the Ouachita Mountains and includes deformed metasediments presumed to be equivalent to the pre-orogenic sequence. The Athens Plateau sequence is defined as the 'southern Carboniferous province' in the subsurface, but is only documented immediately south of the exposed Ouachita Mountains. The frontal zone varies in width and can be followed south to central Texas. The most distinctive structural element in the frontal zone is the leading thrust as this boundary is clearly proven by drilling. The interior zone extends from the Broken Bow Uplift, but is not continuous across north-east Texas (Figure 1). In general, the

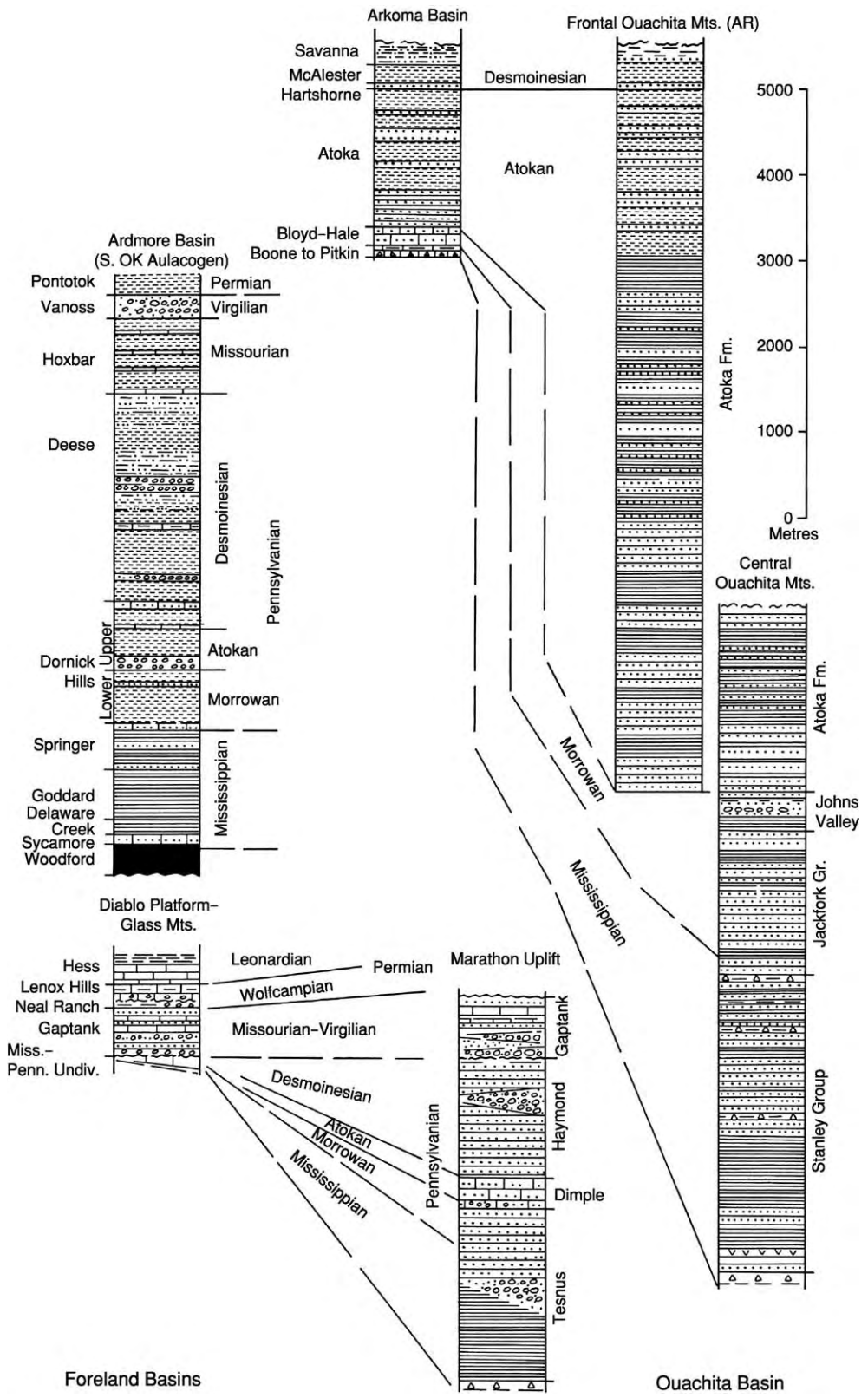


Figure 3 Representative lithological columns of syn orogenic sequences showing the transition from the craton to the Ouachita Basin. (After Arbenz, 1989.)

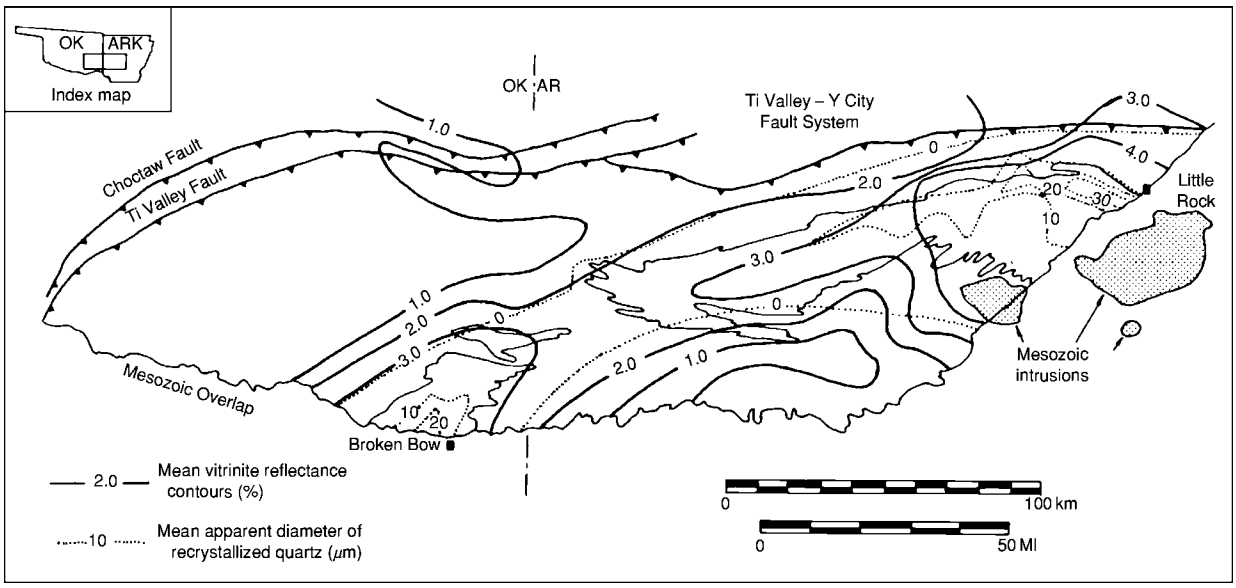


Figure 4 Thermal maturity of exposed strata of the Ouachita Mountains. (After Viele and Thomas, 1989.)

interior zone follows the frontal zone and together they form an 80 km wide belt. The frontal zone appears to be cut out in south-central Texas only to reappear east of the Marathon Uplift. Two notable uplifts are documented along this subsurface trend: the Waco Uplift along the south-trending section and the Devils River Uplift along the westerly-trending section. The first is interpreted to be a segment of the North American Craton thrust inboard carrying an overlying allocthonous section of Ouachita facies. The second uplift is an approximately 120 km by 50 km northwest-trending Precambrian block juxtaposed to the Carboniferous foreland basin.

The last segment of the Ouachita Orogen is located in West Texas and includes the exposures of the Marathon Uplift and two smaller outcrops to the south-west (Figure 1). Similar to the Ouachita Mountains, the entire outcrop belt of Marathon Uplift is allochthonous. This area is correlated to subsurface frontal zone. The stratigraphic sequence is similar to that of the Ouachita Mountains. The folding style is directly related to the nature of the lithologic package, long wavelength folds in the thick flysch sequence and complex, tight folding and faulting in the pre-orogenic sequence. The measured shortening has been utilised to restore this belt to a palaeogeographic position as much as 200 km to the south-east.

Five foreland basins are an integral part of the Ouachita Orogen (Figure 1). From east to west, these include the Black Warrior, Arkoma, Fort Worth, Kerr, and Val Verde basins. The Black Warrior Basin is critical in that it is bordered on two sides by convergent zones; the Ouachita thrust faults cut the southwest side of the basin and the Appalachian thrust belt on the south-east intersects these faults. The Black Warrior Basin is a homocline dipping gently towards the tectonic front and extending beneath the front. This homocline rises up to a large domal structure, the Nashville Dome, to the north. The pre-orogenic sequence rests on Precambrian basement and is dominated by carbonate shelf deposition. The Upper Mississippian-Pennsylvanian syn-orogenic clastic wedge progrades north-eastward. This section thickens towards the tectonic front and is consistent with tectonic loading at a convergent margin. The other foreland basins have similar geometries and pre-orogenic successions. They differ primarily in the nature of the syn-orogenic sequence, both in timing and depositional environments.

Regional Geophysics

The Bouguer anomaly map represents the most extensive geophysical data set for the region (Figure 5). A positive anomaly associated with the interior zone can be followed from central Arkansas through Texas into the Marathon region. The extension of this belt from Arkansas into Mississippi is more obscure. While uplifted basement blocks, mafic intrusions, and metamorphism contribute to the anomaly, the anomaly requires a major transition in crustal structure

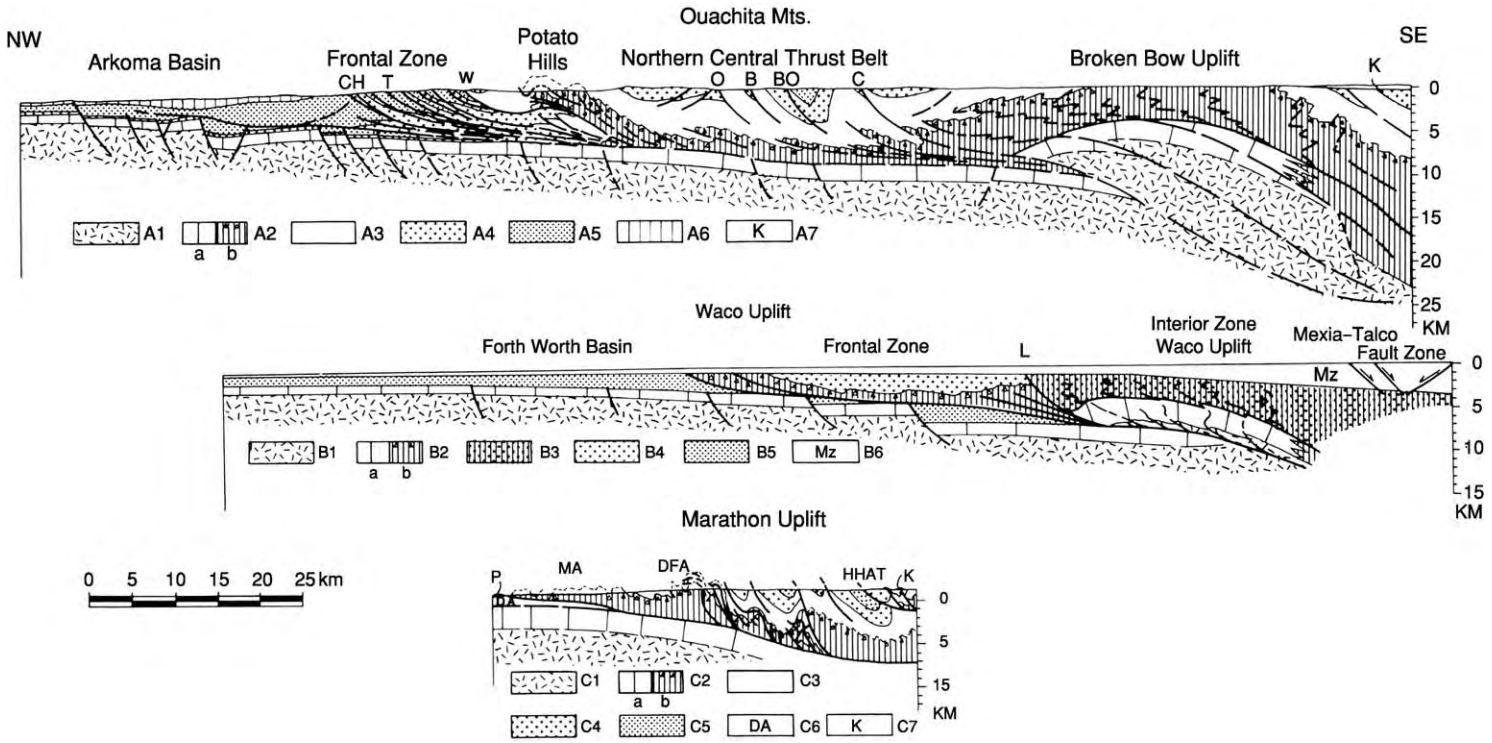


Figure 5 Generalized cross sections across the Ouachita Orogen in Oklahoma and Texas. (From Hatcher *et al.*, 1989) (Modified from Arbenz, 1989.)

across this trend. In other words, it is proposed that this anomaly represents the Early Palaeozoic continental margin. Of similar amplitude is the northwest-trending positive anomaly associated with the Southern Oklahoma aulacogen (Figure 6). Allowing for the uplifted Precrambrian blocks and mafic intrusions, this implies significant crustal modification during rifting. The aulacogen trend is clearly developed from the Panhandle area of Texas southeastward to the Ouachita Orogen where the two trends form a distinctive X-pattern. Closed gravity minima are observed for three of the foreland basins: Arkoma, Fort Worth, and Kerr basins. Gravity lows are noted in the Black Warrior and Val Verde basins, but the pattern is more subdued. Of these Basins, the Arkoma Basin is the most dramatic (~80 mgal). Interestingly, the lowest values of this minimum are located just north of the Broken Bow Uplift where a steep gradient is also apparent. This pattern argues for a thick sedimentary section in the central Ouachita Mountains, some of it possibly below the allocthonous Ouachita facies. The Benton, Broken Bow, and Waco uplifts are actually located on the gradient inboard of the interior zone anomaly. Thus, these uplifts have a similar density to the Ouachita facies; that they are not deeply rooted in the basement and have been translated significant distances. In addition, the uplifts probably reflect late stage compressional deformation along the transitional crust of the continental margin. In contrast, the Devils River Uplift does reveal a distinct positive anomaly related to the large Precambrian block.

Regional seismic data is widely scattered. Early refraction and surface wave studies document a normal continental crust north and west of the margin and an ambiguous noncratonal crust model south of the margin. The COCORP line in southwestern Arkansas is the longest line across the fold and thrust belt. Crustal models based on the COCORP and additional data to the south show an increase in velocities, which have been interpreted to be interleaved granitic and basaltic crust. In this area, the Early Palaeozoic transitional crust is thought to be preserved; and the deformed Palaeozoic section is detached and transported northward over the continental margin. Seismic reflection data support the allocthonous, northerly verging geometry of this fold and thrust belt. Additional features identified in

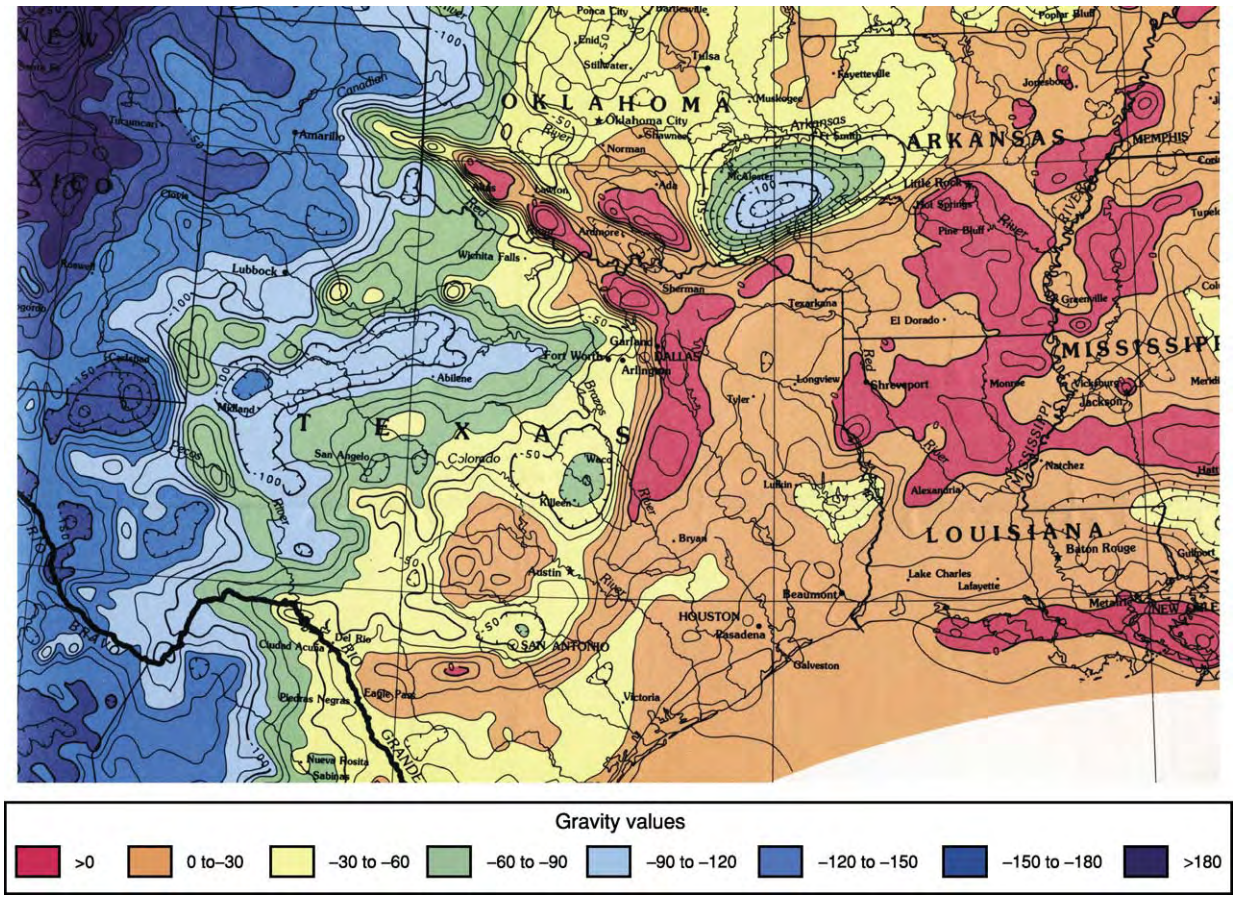


Figure 6 Bouguer Anomaly Map of south central North America. Contour Interval is 10 Milligals. (After Keller *et al.*, 1989.)

the COCORP line are the basement cored uplift, Early Palaeozoic sections overlying a presumed Precambrian basement, allocthonous Ouachita facies structurally overlying the Early Palaeozoic section, and a tectonically thickened section of Carboniferous units south of the uplift. A final feature identified in the subsurface is a successor basin of Late Pennsylvanian sediments in unconformable contact with the deformed Carboniferous units south of the Ouachita Mountains (Figure 1).

Tectonic Synthesis

Representative cross-sections provide a clear picture of the structural style of the Ouachita Orogenic Belt (Figure 6). Common to these sections is the dominant thin skin style of deformation. The amount of translation is minimally 50 to 100 km; the greatest of these displacements appear to be in the salients. The foreland basin reveals the tilt of the basement and the pre-orogenic sequence towards the tectonic zone. Also, the strongly asymmetric nature of the foreland basin is outlined. In the Ouachita Mountains, the extremely thick sedimentary succession north of the Broken Bow Uplift is illustrated along with the large folds of the central zone. The basal detachment is above the shelf sequence, possibly with a Carboniferous section also in the footwall. Commonly, the detachment is interpreted to extend over the basement uplift. The pattern is similar in the Waco Uplift, although the basement block is much thinner. The Marathon cross-section is limited to the exposed uplift and reflects a style similar to the frontal zone of the Ouachita Mountains, but well data has extended the thrust further to the south where some metamorphic rocks have been associated with the 'interior zone'.

The Ouachita Orogen represents the imprint of the collision of North America, South America, and possibly several small continental blocks and/or island arcs. The specific nature of these blocks is debatable; however, there is agreement about the general timing. Initial subsidence is recorded by the shift from the Devonian chert and siliceous shale deposition to the Carboniferous turbidite sequence. In the Ouachita Mountains, this transition is in the Lower Mississippian. Similar timing is envisioned for the Caballos Novaculite of the Marathon region. Coincident with the change in the basin, the shelf areas of the Black Warrior Basin experienced a change from the Lower Mississippian shallow marine chert and limestone to an Upper Mississippian clastic wedge. This Mississippian clastic wedge prograded northeastward, reflecting a source to the south-west. In the Arkoma Basin, however, the shelf was stable with Mississippian shallow marine limestone grading southward into mudstones. A second clastic wedge in the Black Warrior Basin of Pennsylvanian age is more extensive, and merges with a another wedge prograding from the Appalachian tectonic front. Within the Arkoma Basin, deltaic sedimentation on a stable shelf continued until Middle Pennsylvanian (Lower Atokan). This was followed by the development of down-to-south normal faults and syn-orogenic deposition. Thrust faults truncate the Pennsylvanian sediments in the foreland basins. The Arkoma Basin is divided into a southern compressional zone and a northern extensional zone related to plate loading and normal fault development. The youngest units of the Arkoma Basin involved with the Ouachita Orogenesis are mid-Desmoinesian (Late Pennsylvanian). Coincident with this compression is the development of a successor basin south of the exposed Ouachita Mountains. Desmoinesian shallow marine deposits rest unconformably on deformed Ouachita Carboniferous units. In the Marathon area the compressional deformation is subdivided into two phases. The first phase corresponded with the folding of the Dimple Limestone, formation of the Val Verde foreland Basin, and the transition syn- to post-orogenic sediments (Gaptank Formation). The timing of this phase is Middle to Upper Pennsylvanian. The younger phase deformed these folded and faulted Pennsylvanian units as part of the development of the Dugout Creek Thrust. This younger phase extended into the Early Permian (Wolfcampian).

While initial subsidence appears to be nearly coincident along the length of the Ouachita Basin, the compressional events began earlier in the east (Middle Mississippian). It has been suggested that this early phase was related to a collision with an island arc along the south-east margin of North America. The 'two tier' fold style developed in the Ouachita Mountains is interpreted as reflecting deformation in two distinctly different settings. The earliest deformation involved Cambrian through to Early Mississippian basinal sediments. Tight folds and faults developed in a growing accretionary wedge above the oceanic and transitional continental crust. The remainder of the Carboniferous flysch sequence was then deposited over these deformed units and the entire accretionary prism was thrust onto the southern margin of North America in the Late Pennsylvanian. As mentioned above, the final stages of thrust faulting were Desmoinesian in the Ouachita Mountains. The principal thrusting began about this time in West Texas and continued into the Early Permian. Therefore a diachronous model is commonly applied to the Ouachita Mountains, closing from east to west. Thrust faults of the Appalachian Alleghenian

Orogeny truncate the Ouachita trends in the Black Warrior Basin. The final stages of deformation in the Ouachita Mountains are also documented in the rotated folds and cleavages of the Broken Bow and Benton uplifts. These basement cored structures trend ENE and reflect a rotation of the principal shorten direction. It has been proposed that these uplifts represent the far-field effects of the Alleghenian collision to the south-east. Various microplates have been proposed to account for the translation of the Ouachita facies inboard over the coeval shelf sequences. Collectively these have been identified as the Sabine Plate and represent elements trapped between South and North America during the assembly of Pangaea. Subsequent Mesozoic rifting associated with the development of the Gulf of Mexico have unfortunately obscured these relationships.

See Also

North America: Northern Cordillera; Southern Cordillera; Southern and Central Appalachians. **Pangaea**.

Further Reading

Arbenz JK (1989) The Ouachita system. In: Bally AW and Palmer AR (eds.) *The Geology of North America: An Overview,* v. A, pp. 371 396. Boulder, Colorado: Geological Society of America, The Geology of North America.

Ethington RL, Finney SC, and Repetski JE (1989) Biostrati graphy of the Paleozoic rocks of the Ouachita orogen, Arkansas, Oklahoma, west Texas. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States,* v. F 2, pp. 563 574. Boulder, Colorado: Geological Society of America, The Geology of North America.

Flawn PT, Goldstein A Jr., King PB, and Weaver CE (1961) *The Ouachita system*. Austin, University of Texas, Bureau of Economic Geology Publication number 6120, p. 401.

Hatcher RD Jr., Thomas WA, and Viele GW (1989) *The Appalachian Ouachita Orogen in the United States,* v. F 2, p. 767. Boulder, Colorado: Geological Society of America, The Geology of North America.

Keller GR, Kruger JM, Smith KJ, and Voight WM (1989) The Ouachita system: A geophysical overview. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States,* v. F 2, pp. 689 694. Boulder, Color ado: Geological Society of America, The Geology of North America.

King PB (1937) *Geology of the Marathon region, Texas,* U.S. Geological Survey Professional Paper 187, p. 148.

Lillie RJ, Nelson KD, de Voogd B, *et al.* (1983) Crustal structure of Ouachita Mountains, Arkansas: A model based on integration of COCORP reflection profiles and regional geophysical data. *American Association of Petroleum Geologists Bulletin* 67: 907 931.

Lowe DR (1989) Stratigraphy, sedimentology, and depos itional setting of pre orogenic rocks of the Ouachita Mountains, Arkansas and Oklahoma. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States;* v. F 2, pp. 575 590. Boulder, Colorado: Geological Society of America, The Geology of North America.

McBride EF (1989) Stratigraphy and sedimentary history of Pre Premian Paleozoic rocks of the Marathon uplift. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States:* v. F 2, pp. 603 620. Boulder, Color ado: Geological Society of America, The Geology of North America.

Morris RC (1989) Stratigraphy and sedimentary history of post Arkansas Novaculite Carboniferous rocks of the Ouachita Mountains. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States:* v. F 2, pp. 591 602. Boulder, Colorado: Geological Society of America, The Geology of North America.

Muehlberger WR and Tauvers PR (1989) Marathon fold thrust belt, west Texas. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States,* v. F 2, pp. 673 680. Boulder, Colorado: Geological Society of America, The Geology of North America.

Nicholas RL and Waddell DE (1989) The Ouachita system in the subsurface of Texas, Arkansas, and Louisiana. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States;* v. F 2, pp. 661 672. Boulder, Color ado: Geological Society of America, The Geology of North America.

Nielsen KC, Viele GW, and Zimmerman J (1989) Structural setting of the Benton Broken Bow uplifts. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States:* v. F 2, pp. 635 660. Boulder, Colorado: Geo logical Society of America, The Geology of North America.

Thomas WA (1989) The Appalachia Ouachita orogen be neath the Gulf Coastal Plain between the outcrops in the Appalachian and Ouachita Mountains. In: Hatcher RD (ed.) *The Appalachian Ouachita Orogen in the United States:* v. F 2, pp. 537 553. Boulder, Colorado: Geo logical Society of America, The Geology of North America.

Viele GW and Thomas WA (1989) Tectonic synthesis of the Ouachita orogenic belt. In: Hatcher RD (ed.) *The Appa lachian Orogen in the United States:* v. F 2, pp. 695 728. Boulder, Colorado: Geological Society of America, The Geology of North America.

Southern and Central Appalachians

R D Hatcher, Jr, University of Tennessee, Knoxville, TN, USA

> The Appalachian Chain is the most elegant on Earth, so regularly arranged that its belts of formations and structures persist virtually from one end to the other. . . . But the apparent simplicity is deceiving; actu ally it is full of guile, and its geology has aroused contro versies as acrimonious as any of those in our science.
>
> (Philip B King, 1970)

Introduction

The exposed Appalachians extend from Alabama to Newfoundland, with buried components extending eastward and southward beneath the Gulf and Atlantic coastal plains, and the Atlantic continental shelf (Figure 1). The southern and central Appalachians reach their greatest width in Tennessee, the Carolinas, and Georgia. To the north, the chain reaches its narrowest point at the latitude of New York City, and from there it widens northward again into New England. Physiographical provinces from north-west to south-east include the Cumberland–Allegheny Plateau, Valley, and Ridge from Pennsylvania southward, the Blue Ridge from South Mountain in southern Pennsylvania to northern Georgia, and the Piedmont from New Jersey to Alabama. Basins filled with Late Triassic–Early Jurassic sedimentary (and some volcanic) rocks that formed prior to the opening of the present Atlantic Ocean are superposed on the Piedmont from North Carolina northward. Physiographical provinces to the west correspond roughly to geological provinces, but there is little correspondence with geological provinces in the interior of the chain. The southern Appalachians contain the highest mountains in the entire chain. Though there is one mountain (Mt Washington, New Hampshire) in New England with an elevation greater than 6000 ft (~2000 m), and none farther north, there are at least 39 named peaks in the southern Appalachian Blue Ridge of western North Carolina and eastern Tennessee with elevations exceeding 6000 ft. Topographic relief here can exceed 5000 ft (1600 m).

The Appalachians were created following the rifting of the Rodinia supercontinent during three or more Palaeozoic orogenic events: the Ordovician–Silurian Taconian, the Devonian Acadian and Devonian–Mississippian Neoacadian, and the Pennsylvanian–Permian Alleghanian orogenies (Figure 2). All orogenies were diachronous. The character of the orogen changes roughly at the latitude of New York City, from one constructed largely from west to east to the north, to an orogen built from the inside out to the south, with the exception of the Alleghanian Thrust Belt that extends up the Hudson River Valley to just south of Albany, New York (Figure 3).

Philip King doubtlessly had the southern and central Appalachians in mind when he made his observation about the Appalachians. These mountains have been studied for over 150 years, and several fundamental concepts in geological science were developed in the process. The continentward segment of the chain consists of a foreland fold–thrust belt that changes style within; it is fold-dominated from Pennsylvania to south-west Virginia, thrust-dominated from south-west Virginia to Alabama, and fold-dominated again in Alabama (Figure 3). This along-strike change in character is related to changes in mechanical stratigraphy, to relationships of the shape of the original continental margin, and to the shape of the Gondwanan indenter that collided with Laurentia at the end of the Palaeozoic. The internal parts of the chain to the east initially consist of deformed rifted margin sedimentary rocks and recycled Grenville basement. These give way eastward to accreted terranes of distal Laurentian metasedimentary oceanic and arc-affinity material (Tugaloo, Chopawamsic, Potomac, and Baltimore terranes), then a sequence of mid-Palaeozoic metasedimentary rocks (Cat Square Terrane), and a truly exotic peri-Gondwanan (Pan-African) Carolina–Avalon Superterrane of Neoproterozoic to Ordovician arc volcanic and volcaniclastic rocks, including the Smith River allochthon (Figure 3). The subsurface Suwannee Terrane lies south of an east- and west-trending suture zone in southern Georgia and Alabama (Figure 4), and contains Gondwanan-affinity basement and cover sedimentary rocks to the south. The Appalachian collage records a complete Wilson Cycle of the opening and closing of a series of Palaeozoic oceans, starting with rifting of the supercontinent Rodinia and ending with the assembly of the supercontinent Pangaea.

The reference frame used throughout this article is the present geography. However, the equator trended north–south (present geographic frame), bisecting North America during the Early and Middle Palaeozoic. Laurentia rotated counterclockwise during the Middle to Late Palaeozoic, bringing the continent into its present orientation, but still at equatorial

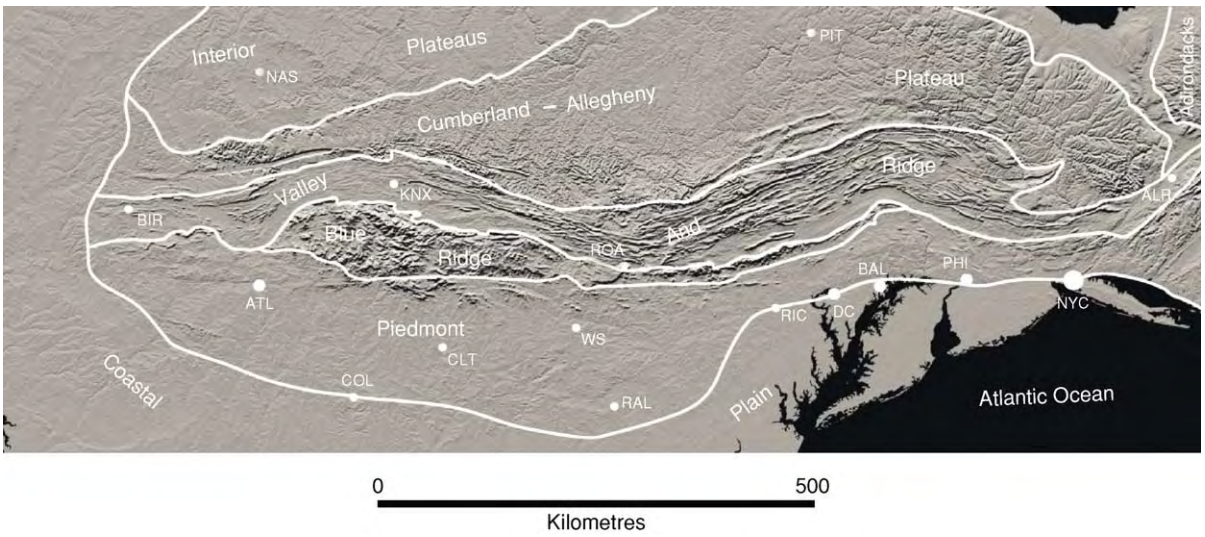


Figure 1 Shaded relief map showing the physiographical provinces of the southern and central Appalachians. ALB, Albany, New York; ATL, Atlanta, Georgia; BAL, Baltimore, Maryland; BIR, Birmingham, Alabama; CLT, Charlotte, North Carolina; COL, Columbia, South Carolina; DC, Washington, DC; KNX, Knoxville, Tennessee; NAS, Nashville, Tennessee; NYC, New York City; PHI, Philadelphia, Pennsylvania; PIT, Pittsburgh, Pennsylvania; RAL, Raleigh, North Carolina; RIC, Richmond, Virginia; ROA, Roanoke, Virginia; WS, Winston Salem, North Carolina. Image from the United States Geological Survey.

latitude. Most subsequent movement of Laurentia during the Mesozoic and Cenozoic has been northward to its present location north of the equator.

Neoproterozoic to Ordovician Iapetan Margin Development and Destruction

Rifting of Rodinia

Supercontinent Rodinia formed 1.2–1 billion years ago (Ga) as a product of the worldwide Grenvillian Orogeny. The continent remained intact for ~300 million years (My) and then rifted apart at ~750 Ma in western and north-eastern North America, but failed rifting occurred at ~735 Ma in the southern and central Appalachians. Failed rifting, however, produced the alkalic A-type Crossnore plutonic–volcanic suite; the Grandfather Mountain and Mount Rogers Formation marine, non-marine, and glaciogenic sedimentary and bimodal volcanic rocks in North Carolina, Tennessee, and Virginia; and the Robertson River igneous complex in central Virginia. Successful rifting of Rodinia along the entire southern-central Appalachian margin occurred at ~565 Ma, with basins that deposited the Ocoee Supergroup in Tennessee, North Carolina, and Georgia, and the Catoctin volcanic–sedimentary assemblage in Virginia. Neither of these assemblages is physically connected, because initial rifting created an irregular margin in south-eastern Laurentia, leaving basement highs that separated rift basins.

Deposition of the Ocoee Supergroup locally reached thicknesses exceeding 15 km, with most of the thickness contained in the Great Smoky Group. The Ocoee Supergroup contains three groups: (1) the Snowbird Group, which rests unconformably on Grenvillian basement, (2) the Great Smoky Group, which nowhere has been observed in direct contact with basement rocks and is faulted onto the Snowbird Group rocks, and (3) the Walden Creek Group, which conformably overlies both the Great Smoky Group and the Snowbird Group, revealing an extraordinarily complex relationship between the components of the Ocoee Supergroup. The Snowbird Group consists of immature basal sandstone overlain by much cleaner quartz arenite to arkosic sandstone, then a pelitic unit. The Great Smoky Group consists of a rapidly deposited deep-water sequence of immature greywacke and pelite (some of which was deposited in an anoxic environment), minor amounts of cleaner sedimentary rocks, and polymictic vein quartz-dominated conglomerate that also contains quartzite, sandstone, granitoid, black shale rip-ups, and variably textured limestone and dolostone clasts. The Walden Creek Group consists of banded chlorite–sericite slate that contains minor amounts of limestone overlain by shale, clean sandstone, and limestone. The latter unit has yielded soft-bodied wormlike fossils, and possible ostracodes and inarticulate brachiopods, indicating a Cambrian (probably Tommotian) age. Coarse polymictic vein quartz-dominated conglomerate facies also occur within this unit, and again the clasts are

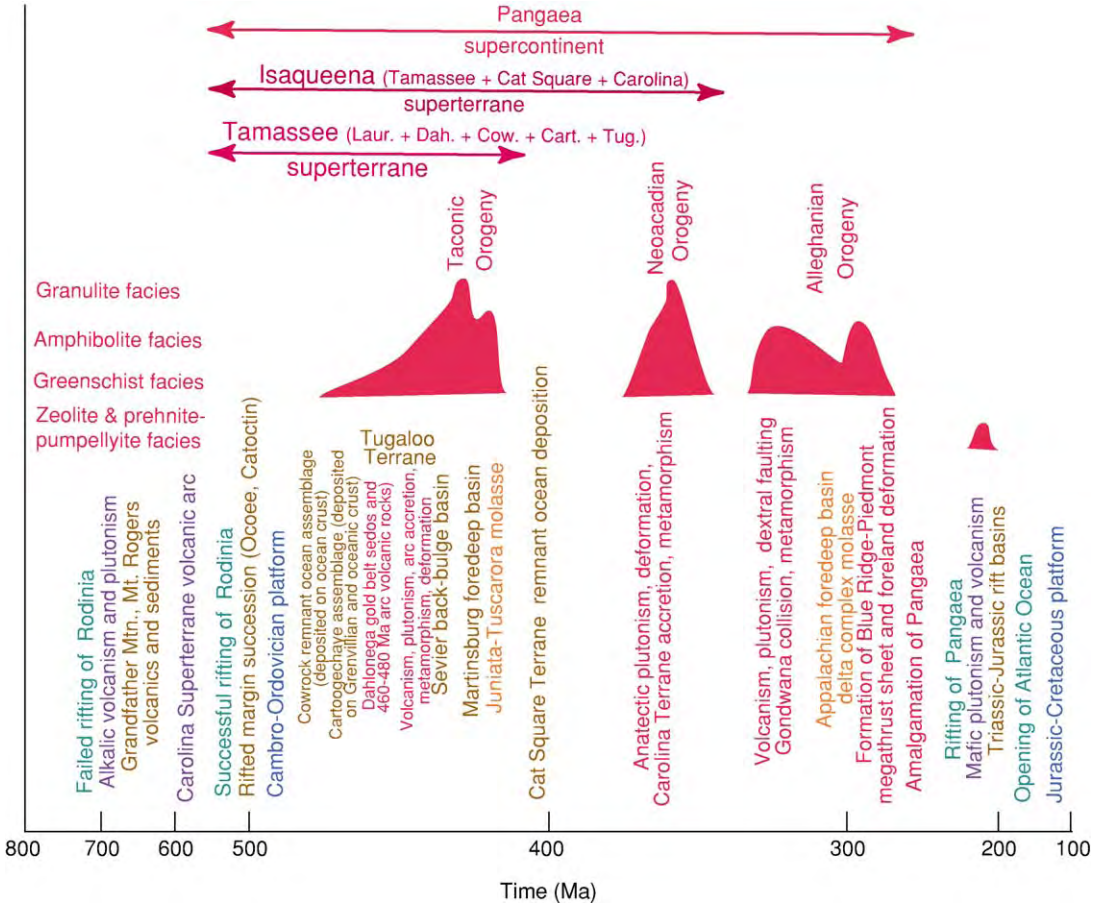


Figure 2 Events distributed through time in the Appalachian Wilson Cycle. High points on the curves denoting orogenies indicate locations of maxima in time and maximum pressure and temperature conditions. Compressional events and related activities are red; arc and rifted margin volcanism is purple; clastic wedges are brown and orange; carbonate platforms are blue; rifting events are green. Note that the horizontal scale is expanded between 500 and 200 Ma.

dominated by vein quartz; several textural and compositional variations of limestone and dolomite; and quartzite, sandstone, and black shale rip-ups. Many of the conglomerates clearly were deposited in channels; armoured mudballs have been observed in at least one palaeochannel. Though the Ocoee Supergroup thins gradually southward to near Cartersville, Georgia, its thickness rapidly decreases northward from its maximum of 15 km in the Great Smoky Mountains National Park, to zero south-east of Johnson City, Tennessee.

In central Virginia, the Ocoee-equivalent ~565-My-old Catoctin Formation consists of basalt and rhyolite, with some interlayered clastic sedimentary rocks that are locally underlain by an older suite of immature rift-related sedimentary rocks (Snowbird Group equivalent?). East of the Blue Ridge anticlinorium in northern North Carolina and Virginia is a sequence of distal, deeper water clastic sedimentary and rift-related volcanic rocks, the Ashe–Alligator Back-Lynchburg sequence. More distal eastward and south-eastward equivalents of these metasedimentary and metavolcanic rocks (Ashe–Tallulah Falls–Sandy Springs) have been recognized in the eastern Blue Ridge and Inner Piedmont from the Carolinas to Alabama. They are overlain by Cambrian–Ordovician(?) metasedimentary rocks (Candler, Chauga River), and by Middle Ordovician(?) mid-oceanic ridge basalt-arc volcanic rocks (Slippery Creek–Poor Mountain–Ropes Creek). Farther east in the Virginia Piedmont is a sequence of felsic–mafic volcanic rocks from ~470 Ma (Chopawamsic) that were intruded by plutons at ~455 Ma, then overlain by fossiliferous Late Ordovician (Ashgill) fine-grained clastic rocks comprising the Chopawamsic Terrane.

The Lower Cambrian Chilhowee Group was deposited on the Laurentian margin during the rift-to-drift transition, and consists of upward-maturing alternating sandstone and shale units. The lowest unit consists of immature sandstone and greywacke with some basalt. The upper part of the Chilhowee Group

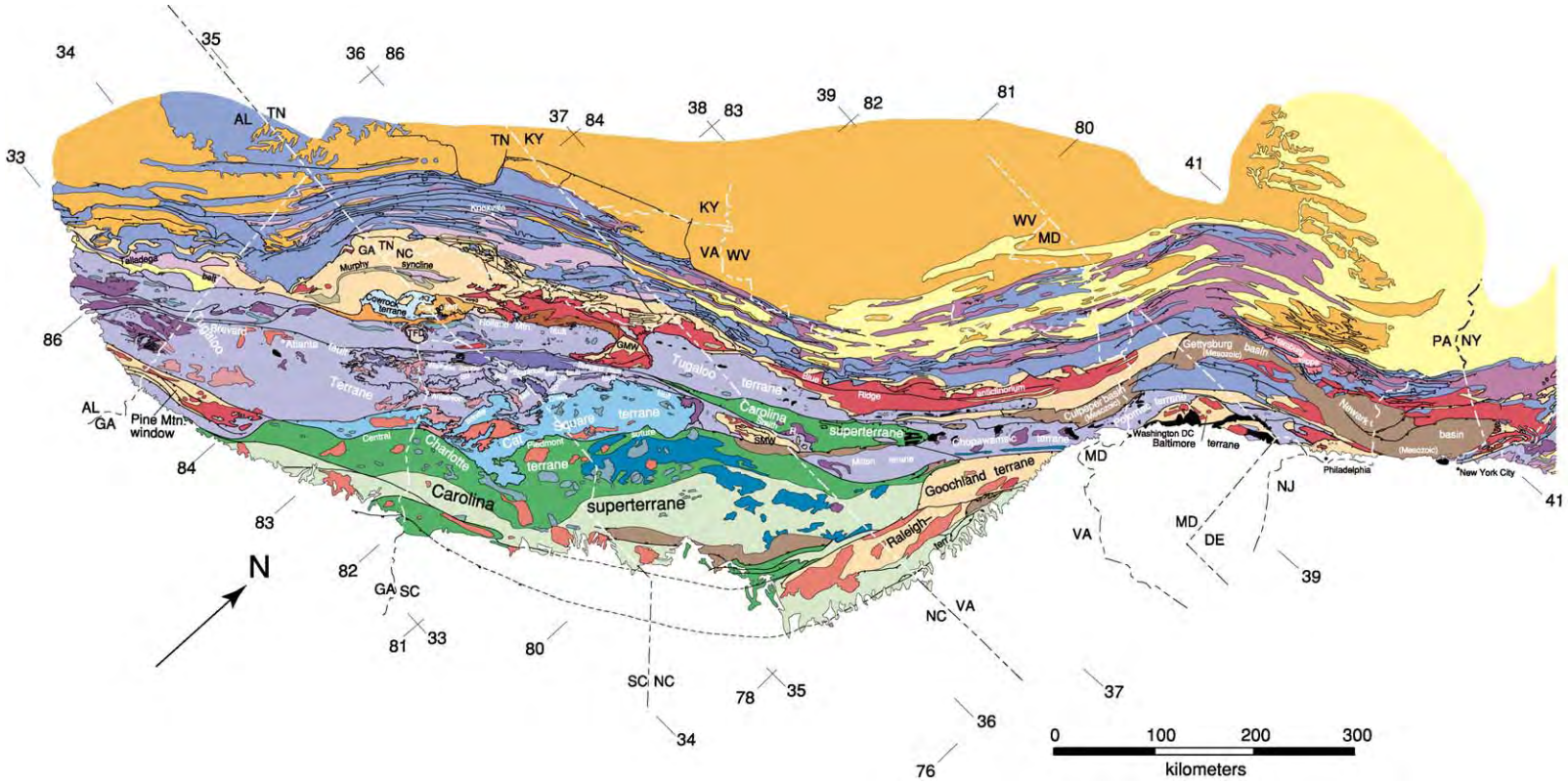


Figure 3 Tectonic map of the southern and central Appalachians. GMW, Grandfather Mountain window; SMW, Sauratown Mountains window; TFD, Tallulah Falls Dome; Cart. Terr., Cartoogechaye Terrane. Colour key: ultramafic rocks are black; Alleghanian clastic wedge is orange (to north west); Acadian clastic wedge is light yellow (north eastern and south western parts of map); Middle Ordovician (Sevier, Blountian) clastic wedge is lavender; Upper Ordovician Silurian Martinsburg Tuscarora (type Taconian) clastic wedge is wine; Cambro Ordovician platform rocks are dark blue; rifted Laurentian margin Neoproterozoic to Cambrian clastic and volcanic rocks are tan; Murphy syncline is olive green with lavender in centre (imbedded in tan in Georgia and North Carolina); 1.1 Ga (Grenvillian) rocks are red; pre Grenvillian 1.8 Ga Mars Hill Terrane is brown orange (next to large red area in the Blue Ridge); Cowrock Terrane is light blue (in Georgia and North Carolina); Cartoogechaye Terrane is light orange (in Georgia and North Carolina); Dahlonega gold belt is burnt orange (Georgia and North Carolina); Hamburg Klippe is pink (north eastern part of map); high grade parts of Carolina Superterrane are dark green; low grade parts of Carolina Superterrane are light green; Neoproterozoic pluton is light grey (north west of GMW); Neoproterozoic to Early Cambrian plutons (Carolina Superterrane) are dark blue green; Ordovician plutons are dark purple; likely Ordovician and Silurian plutons are medium purple; Devonian plutons dark gray green; Carboniferous to Permian plutons are red orange; Late Triassic Jurassic sedimentary and volcanic rocks are brown. Scale 1:2 500 000.

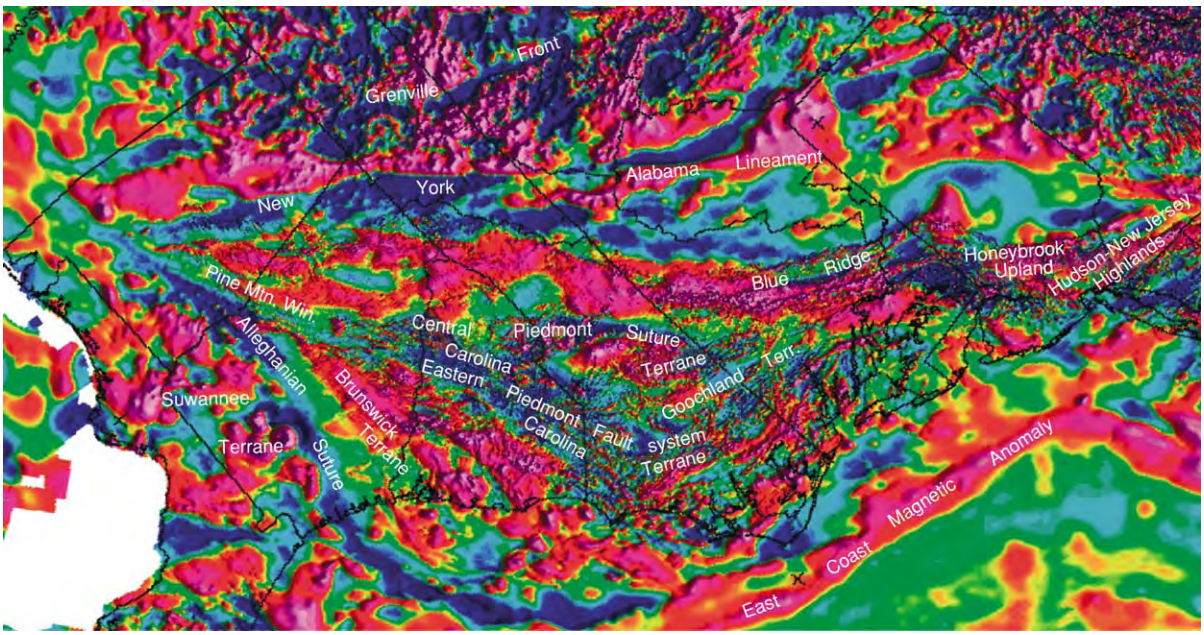


Figure 4 Magnetic map of major crustal features in southern and central Appalachia. Magnetic anomalies are shown in spectral colours, with red indicating highs and blue indicating magnetic lows. Image from the United States Geological Survey.

was probably deposited along the open-ocean margin, because it grades upward into the first successful carbonate bank assemblage. Both the Chilhowee Group and the overlying Shady (–Tomstown–Dunham) Dolomite are continuous throughout the Appalachian margin.

Passive Margin Evolution

Once the first carbonate bank was established, the south-eastern Laurentian margin continued to develop, with pulses of clastic sediment derived from the eroding Grenvillian mountains of progressively lower relief. From Late Cambrian to earliest Middle Ordovician time, carbonate deposition transgressed across the entire margin into the central interior of North America, extending westward as far as present-day westernmost Texas, and northward into southern and eastern Canada. This formed the first extensive carbonate platform across much of interior Laurentia. Following this great incursion of shallow-water marine carbonate deposition into interior North America, the entire carbonate platform was uplifted and a karst surface formed over all, except for an area in northern Virginia, Maryland, and Pennsylvania. Carbonate deposition resumed during the Middle Ordovician (Llanvirn). Shortly afterwards, the eastern margin began rapidly to subside as Ordovician volcanic arcs were obducted onto the eastern Laurentian margin (Figure 5). A foredeep formed in front of the arriving thrust sheets, to the west of that a forebulge, and a back-bulge basin formed ~200 km to the west into the platform. This

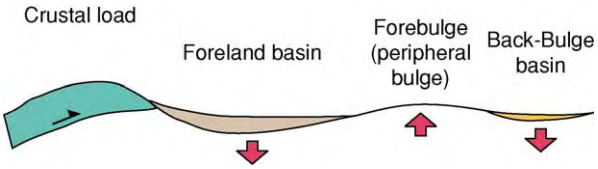


Figure 5 Ordovician loading of the Laurentian margin by obducted volcanic arcs, producing a foredeep basin, forebulge, and a back bulge basin farther into the continent.

back-bulge basin accumulated >3 km of clastic sediment in the southern Appalachians, forming the Sevier clastic wedge that graded westward into the thin platform limestone sequence. Potassium (K)-bentonite beds (Late Llanvirn) from continuing volcanic activity to the east were deposited throughout the southern and central Appalachians and on the adjacent platform to the west.

The younger (Caradocian) Martinsburg clastic wedge formed a thin successor basin atop the Sevier wedge in the southern Appalachians, but accumulated ~5 km of deep-water turbidic clastics in Pennsylvania, probably in a true foredeep. The Martinsburg assemblage graded upward into the shallower water Ashgill Juniata red-bed succession, and then into the Llandoverian Clinch–Tuscarora quartz arenite and Shawangunk Conglomerate (to the north-east). Though the source of the clastic sedimentary assemblage in the Sevier back-bulge and Martinsburg foredeep basins was clearly to the east, detrital zircon suites contain

no Palaeozoic components. All are dominated by zircons derived from Grenvillian and older Laurentian sources, suggesting provenance from exhumed and cannibalized platform, and rifted margin sedimentary and volcanic rocks. Clasts in Llanvirn polymictic conglomerates in the Sevier Basin sample platform carbonate and older lithologies, but none are derived from sources in the Taconian Orogen. Detrital zircon suites from the rifted margin succession (Ocoee Supergroup, Ashe Formation, and equivalents) across the southern Appalachian Blue Ridge and western Inner Piedmont (Tugaloo Terrane) are all dominated by zircons from ~1.1 Ga, with smaller components of zircons from 1.3–1.5 Ga. Dahlonega gold belt detrital zircons, though also dominated by 1.1- to 1.4-Gy-old components, contain components from 2.7–2.9 Ga, 1.9–2.2 Ga, and 650–700 Ma. These rocks were traditionally considered only a more distal facies of the Ocoee Supergroup, but detrital zircons suggest they may be more exotic. In addition, the Dahlonega gold belt contains an arc complex from 460 and 480 Ma.

Taconic Events

The Taconic Orogeny is clearly diachronous in the southern and central Appalachians, with earlier pulses beginning in the Late Arenig in the southern Appalachians and the younger event—the type Taconic Orogeny—beginning sometime in the Llandeilo or Caradoc in the central Appalachians. Corresponding suites of plutons, penetrative deformation, and metamorphism to conditions as high as granulite facies were a product of subduction, obduction, and arc accretion. The 455- to 460-My-old plutonic–volcanic suite (Hillabee, Ropes Creek, Poor Mountain, Chopawamsic, Baltimore) in the southern and central Appalachians has a mid-oceanic ridge basalt (MORB) to arc affinity. At that time, the Neoproterozoic to Cambrian rifted margin assemblage was A-subducted beneath the arc, penetratively deformed, and metamorphosed to mid-crustal metamorphic assemblages.

During the Taconic events, the Laurentian margin underwent A-subduction and several oceanic and arc assemblages were obducted onto the margin, with accompanying emplacement of the Cowrock, Cartoogechaye, and Tugaloo Terranes in the southern Appalachians, and the Chopawamsic, Potomac, and Baltimore Terranes in the central Appalachians (Figure 3). In New Jersey and eastern Pennsylvania, Grenvillian basement was remobilized in the core of the recumbent Musconnectong Nappe. Basement was also remobilized to the north-east (Hudson–New Jersey Highlands) and south-west, forming external massifs in the core of the Blue Ridge from Virginia to Georgia. Internal massifs in the eastern Virginia Goochland Terrane, the Tugaloo Terrane west of the Brevard fault zone in the Carolinas and north-eastern Georgia, and possibly in the Pine Mountain window in Georgia and Alabama were remobilized, although remobilization of basement in the latter may be younger (see later).

A suite of thrust sheets carrying deep-water sedimentary rocks was emplaced onto the New England and Canadian Maritimes margin at this time, locally carrying oceanic crust and mantle in ophiolite sheets (Quebec and Newfoundland). These extend into the central Appalachians as the Hamburg Klippe in New Jersey and the south-eastern Pennsylvania Piedmont. None has been identified farther south. Other premetamorphic thrusts formed at this time from the Pennsylvania Piedmont southward. Amalgamation of old and new crust generated during the Taconian events formed a superterrane along the south-eastern Laurentian margin.

Cat Square Terrane

Deposition continued until post-430 Ma in the still-open remnant of the Theic Ocean that lay between south-eastern Laurentia and the approaching Carolina Superterrane. Here a sequence of immature deep-water sandstones (subgreywacke to greywacke) and pelites that today consist of aluminous schist and biotite gneiss were probably deposited on oceanic(?) crust. Detrital zircons include the usual prominent 1.1- to 1.4-Ga suite, but with the addition of major components of zircons from 600, 500, and 430 Ma, along with small populations of zircons from 1.9–2.2 and 2.7–2.9 Ga. This diverse population suggests sediment was derived from both Laurentia and Carolina. Presence of younger zircons precludes docking of the Carolina Superterrane before the Late Devonian or Early Carboniferous.

Pine Mountain Terrane

The Pine Mountain Terrane in west-central Georgia and eastern Alabama is a complex window exposing 1.1-Gy-old Grenvillian basement and a thin cover sequence of Early Palaeozoic(?) schist, quartzite–dolomite, and a higher schist. The window is framed by the dextral Towaliga and Goat Rock and Dean Creek faults to the north and south, respectively, but is closed by the Box Ankle Thrust at the east end. All of these faults yield Alleghanian ages, but the Box Ankle Fault was emplaced at sillimanite-grade pressure and temperature (P–T) conditions, whereas the Towaliga and Goat Rock(?) faults moved during garnet-grade conditions, and the Dean Creek Fault moved under chlorite-grade conditions (the same as the Modoc, farther east). Detrital zircons from the

cover-sequence quartzite inside the window yield the usual dominance of ages from 1.1 Ga, but a significant component of zircons from 1.9–2.2 Ga occur there, suggesting a Gondwanan provenance for this sequence and basement. If this is a rifted block of Gondwanan basement, the emplacement kinematics would be very complex, because of the Late Palaeozoic ages of all of the faults framing the window. This basement block and its cover have traditionally been considered Laurentian and the cover sequence is thought to be metamorphosed platform rocks. It remains a suspect terrane.

Peri-Gondwanan Carolina Superterrane

Coeval with the rifting of Rodinia was the development of a number of Peri-Gondwanan terranes in the intervening oceans separating Laurentia from Gondwana. All of these terranes developed proximal to Gondwana, formed composite superterranes, and were accreted to the Laurentian assemblage during the mid-Palaeozoic. Peri-Gondwanan components located in the southern and central Appalachians consist of the Carolina Superterrane and outlier Smith River allochthon (Figure 3). The Neoproterozoic component of the Carolina Terrane consists of a mafic to felsic volcanic arc assemblage and associated volcaniclastic sedimentary rocks. A number of plutons were generated in the range of 550–600 Ma and intruded into the underpinning of the arc complex. This assemblage was metamorphosed during the Cambrian because plutons from 530 Ma cut upper amphibolite to greenschist facies metamorphic rocks. Neoproterozoic Ediacaran metazoan fossils have been identified on bedding planes in low-grade felsic volcanic rocks near Durham, North Carolina.

Overlying the arc complex is a sequence of Cambrian and possible Ordovician clastic sedimentary rocks. A Middle Cambrian Acado-Baltic (*Paradoxides*) fauna occurs in these sedimentary rocks in South Carolina, and Ordovician conodonts have also been reported in another part of the sequence in North Carolina. The high-grade western Carolina Superterrane contains a 350- to 360-My-old metamorphic overprint and numerous younger (Devonian to Carboniferous) granitoids probably related to the mid-Palaeozoic to Late Palaeozoic docking of the Carolina Terrane. The greenschist and lower grade central and eastern Carolina Terrane consists mostly of volcanic and volcaniclastic rocks interrupted by the Alleghanian Kiokee Belt metamorphic core. South-east of this metamorphic core are more low-grade volcanic and volcaniclastic rocks.

Mid-Palaeozoic Sedimentation and Neoacadian Docking of the Carolina Superterrane

Evidence for the Acadian Orogeny from 410–380 Ma in the southern and central Appalachian internides is mostly lacking. A suite of granitoid plutons from 374–382 Ma (Rabun, Pink Beds, Looking Glass, Mt Airy, Stone Mountain) intruded the eastern Blue Ridge of the Carolinas and Georgia, and a suite of anatectic granitoids (Toluca) from 380 Ma exists in the Cat Square Terrane. The Catskill delta spread Middle and Upper Devonian sediment from New England into New York and Pennsylvania, but pinches to the south. Late Silurian–Early Devonian sedimentary rocks comprise the Talladega Group in Alabama. Deposition of the Upper Devonian–earliest Carboniferous Chattanooga–Brallier (and equivalents) black shale unconformably on folded and faulted Silurian and older rocks on the platform may be an indicator of the Neoacadian Orogeny in the internal parts of the southern and central Appalachians.

Detrital zircon ages require that the Cat Square Terrane lying immediately east of the Tugaloo Terrane and west of Carolina (Figure 3) be accreted during the mid-Palaeozoic to Late Palaeozoic. Anatectic granites formed at ~380 and 366 Ma. Metamorphic rims on zircons, not only in the Cat Square Terrane, but also throughout the Tugaloo Terrane, along with deformed 366-My-old plutons in the Cat Square Terrane, indicate peak metamorphism occurred from 360–350 Ma. U/Pb thermal ionization mass spectroscopy (TIMS) ages on monazite yield similar 360 (and 320) Ma ages. Metamorphism to upper amphibolite and granulite facies conditions probably corresponds to time of docking of the Carolina Superterrane, whereby the previously assembled superterrane plus the Cat Square Terrane were A-subducted beneath Carolina, achieving the burial depths necessary to produce wholesale migmatization in the Inner Piedmont and eastern Blue Ridge.

The Inner Piedmont, which now consists of the Cat Square and the eastern Tugaloo Terrane, records a unique pattern of crustal flow that probably is the key to the docking kinematics of the Carolina Terrane. This flow pattern, from south-east to north-west, consists of a north-west-directed suite of thrust sheets that to the west became south-west directed as they were buttressed against the eastern Blue Ridge rocks along the Neoacadian Brevard fault zone. The docking process was probably thus oblique transpressional, beginning in the north and propagating south-westward. As the Carolina Superterrane subducted the components previously joined to Laurentia, the

Inner Piedmont may have decoupled to form an orogenic channel through which the partially melted plastic mass escaped and extruded south-westward.

The western boundary of the Tugaloo Terrane is the Chattahoochee–Holland Mountain fault in the eastern Blue Ridge, which truncates the 374-My-old Rabun Granodiorite. This fault transported migmatitic kyanite to sillimanite-grade Tugaloo Terrane rocks over staurolite–kyanite–sillimanite-grade rocks of the Dahlonega gold belt in southern North Carolina and northern Georgia, and was tightly folded after emplacement. The Chattahoochee–Holland Mountain must therefore be either a Neoacadian or an Alleghanian fault. Though the detrital zircon suites in the Tugaloo Terrane have quite uniform ages and probably reflect a Laurentian source, both Ordovician and Devonian plutons in the western Tugaloo Terrane contain a wide range of inherited components, from dominant 1.1-Gy-old to Archaean zircons. In contrast, almost none of the Ordovician (~466 Ma) and Devonian (~380 Ma) plutons in the eastern Tugaloo and Cat Square Terranes contain any inheritance. The magmas in the western Tugaloo Terrane must have passed through an appreciable thickness of old crust, whereas the magmas crystallizing the plutons in the Inner Piedmont east of the Brevard fault zone, which were generated in more oceanic to arc environments, did not pass through older continental crust on their way to crystallization. Inner Piedmont and eastern Blue Ridge rocks cooled sufficiently to block the metamorphic zircon rims and monazite from 360–350 Ma, and the western Carolina Superterrane recorded $^{40}Ar/^{39}Ar$ cooling from 350 to 360 Ma. Both terranes were reheated at 325 Ma, probably marking the initial collision of Gondwana with Laurentia. The 350- to 360-Ma event actually records the first thermal connection between the Inner Piedmont and the Carolina Superterrane.

Late Mississippian to Permian Alleghanian Zippered Collision with Gondwana and the Amalgamation of Pangaea

Actual collision of Gondwana with crust amalgamated during the Early to Middle Palaeozoic orogenies began the final chapter in the formation of the Appalachian Mountain chain. This collision was oblique, transpressive, rotational, and involved north-to-south collision zipper-like closing of Gondwana with the previously assembled Laurentian–Peri-Gondwanan collage. Initial collision produced a suite of largely S-type granites. Additionally, the Brevard Fault zone was reactivated dextrally under chlorite-grade conditions, and additional dextral faults—the Hylas and other components of the eastern Piedmont Fault system (Roanoke Rapids, Nutbush Creek, Modoc, Augusta, Towaliga, Goat Rock, Bartlett's Ferry)—formed as crustal blocks attempted to escape from the collision zone by moving south-westward from the vicinity of present-day northern Virginia–Maryland–Pennsylvania–New Jersey. In addition to the 325-My-old plutonic suite, a younger suite of plutons from 300–265 Ma is largely confined to the Carolina Terrane. These plutons were likewise generated during various stages of collision. A fault-bounded metamorphic belt, the Kiokee–Raleigh–Goochland Terrane extending from Georgia to the eastern Virginia Piedmont (Figure 3), was also generated during the early and middle stages of collision. Here low-grade Carolina Terrane rocks were metamorphosed to middle to upper amphibolite facies assemblages. In South Carolina and Georgia, the north-west flank of this metamorphic belt consists of the dextral Modoc Fault, whereas the down-to-the-south Augusta fault comprises the south-eastern flank. The Modoc Fault has been traced south-westward to the south-eastern flank of the Pine Mountain window in Georgia and Alabama, where it joins the Dean Creek Fault.

The collision process with its accompanying faulting and other indications of compression caused the interior of the mountain chain to be uplifted beginning in the Late Mississippian (about 325–320 Ma). The uplift process was complex and is recorded by current indicators, variable sediment thickness, and decreased maturity of sediments that demonstrate the orogen was unroofed piecemeal from Alabama to Pennsylvania. At that time, a near-sea level delta formed on the western flank of the chain on the Laurentian margin and the first sediments that contain Palaeozoic detrital zircons were deposited from Alabama to Pennsylvania. The transition from marine to non-marine depositional conditions took place from the Late Mississippian into the Early Pennsylvanian, and abundant land plants formed the coal and the associated sedimentary rocks of the Appalachian Basin.

The final stage of the Alleghanian scenario involved head-on collision of south-eastern Laurentia with Gondwana (Figure 6). The major product of head-on collision is the Blue Ridge–Piedmont megathrust sheet that transported crust amalgamated during all previous Palaeozoic events at least 350 km onto the North American platform. The Brevard Fault remained a suitably oriented weak fault and was again reactivated, this time as an out-of-sequence thrust in the interior of the megathrust sheet The Late Mississippian to Permian clastic wedge was

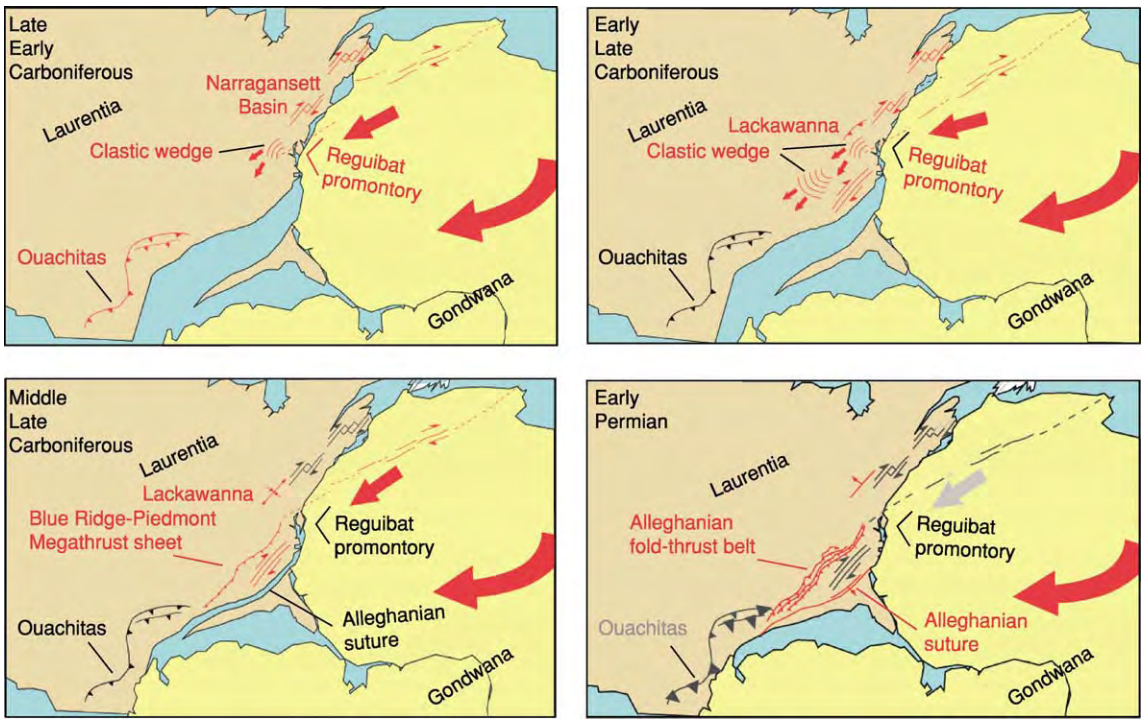


Figure 6 Rotational transpressive collision of Gondwana with Laurentia to form the Supercontinent of Pangaea at the end of the Palaeozoic. Features in red were active at the time indicated. Reproduced with permission from Catalán JRM, Hatcher RD Jr, Arenas R, and García FD (eds.) (2002) *Geological Society of America Special Paper 364*. Boulder, CO: Geological Society of America.

ultimately deformed as the Blue Ridge Piedmont megathrust sheet pushed into the continent, deforming the platform into a foreland fold–thrust belt. In addition, the delta complex that had formed to the west of the rising mountain chain was ultimately deformed. Though final collision was head-on and the megathrust sheet moved westward, uplift of the interior of the chain was non-uniform. The oldest parts of the delta complex are in the southern part of the orogen, possibly related to uplift and erosion of the earlier Ouachita Orogen. This part of the foredeep was superseded to the north by a younger delta complex that fed south-westward along the trough of the foredeep.

The collision was finished about 265 Ma, forming Supercontinent Pangaea and completing the Palaeozoic Wilson Cycle of opening and closing of oceans. The subsurface Suwannee Terrane south of the almost east- and west-trending suture in southern Georgia and Alabama consists of Gondwanan basement and cover Ordovician to Devonian sedimentary rocks of West African affinity. These rocks were left on the Laurentian side in the wake of the breakup of Pangaea. The southern and central Appalachians constitute a microcosm in the assembly of all existing continents along the huge suture zone that extends from south-eastern Laurentia through North-west Africa, western Europe, and the Ural Mountains in Russia.

Breakup of the supercontinent began during the Late Triassic, at ~50 My following the amalgamation of Pangaea, beginning the initial stages of opening of the present Atlantic Ocean. The assembly of Pangaea took place over a span of at least 100 My, but ironically, its breakup occurred over a much shorter timespan. This contrasts with the ≥200-My time-frame for assembling Rodinia and the 300 My that it took for it to break apart.

See Also

Grenvillian Orogeny. **North America:** Precambrian Continental Nucleus; Northern Appalachians. **Pangaea**.

Further Reading

Aleinikoff JN, Zartman RE, Rankin DW, and Burton WC (1995) U Pb ages of metarhyolites of the Catoctin and Mt. Rogers Formations, central and southern Appalachians: evidence of two pulses of Iapetan rifting. *American Journal of Science* 295: 428 454.

Bream BR, Hatcher RD Jr, Miller CF, and Fullagar PD (2004) Detrital zircon ages and Nd isotopic data from the southern Appalachian crystalline core, GA SC NC TN: new provenance constraints for part of the Lauren tian margin. In: Tollo RP, Corriveau L, McLelland J, and

Bartholomew MJ (eds.) *Geological Society of America Memoir 197,* pp. 459 475. Boulder, CO: Geological Society of America.

Faill RT (1998) A geologic history of the north central Ap palachians. Part 1: Orogenesis from the Mesoproterozoic through the Taconic orogeny; Part 2: The Appalachian basin from the Silurian through the Carboniferous. *American Journal of Science* 297: 551 569, 729 761.

Faill RT (1998) A geologic history of the north central Appalachians. Part 3: The Alleghany orogeny. *American Journal of Science* 298: 131 179.

Hatcher RD Jr (2001) Rheological partitioning during mul tiple reactivation of the Paleozoic Brevard Fault Zone, Southern Appalachians, USA. In: Holdsworth RE, Strachan RA, Macloughlin JF, and Knipe RJ (eds.) *The Nature and Significance of Fault Zone Weakening: Geological Society of London Special Publication 186,* pp. 255 269. London: Geological Society of London.

Hatcher RD Jr (2002) The Alleghanian (Appalachian) orogeny, a product of zipper tectonics: rotational transpressive continent continent collision and closing of ancient oceans along irregular margins. In: Catalán JRM, Hatcher RD Jr, Arenas R, and García FD (eds.) *Variscan Appalachian dynamics: The Building of the Late Paleozoic Basement: Geological Society of America Special Paper 394,* pp. 199 208. Boulder, CO: Geological Society of America.

Hatcher RD Jr, Thomas WA, and Viele GW (eds.) (1989) *The Appalachian Ouachita orogen in the United States: The Geology of North America,* vol. F 2. Boulder, CO: Geological Society of America.

Hibbard JP, Stoddard EF, Secor DT Jr, and Dennis AJ (2002) The Carolina zone: overview of Neoproterozoic to early Paleozoic peri Gondwanan terranes along the eastern flank of the southern Appalachians. *Earth Science Reviews* 57: 299 339.

Thomas WA, Becker TP, Samson SD, and Hamilton MA (2004) Detrital zircon evidence of a recycled orogenic foreland provenance for Alleghanian clastic wedge sand stones. *Journal of Geology* 112: 23 37.

Northern Appalachians

C R van Staal, Geological Survey of Canada, Ottawa, ON, Canada

Introduction

The Appalachian Orogen is a remarkably linear, north-east-trending, Palaeozoic mountain belt that generally follows the eastern seaboard of North America, with the segment between Long Island Sound and Newfoundland (Figure 1) being referred to as the Northern Appalachians. Prior to the Mesozoic opening of the Atlantic Ocean, the Appalachians continued into the Caledonides of the British Isles and Scandinavia, forming a long linear mountain chain that was created by the closing of the Iapetus (Cambrian–Early Devonian) and Rheic (Devonian–Carboniferous) Oceans.

Iapetus started to open at the end of the Neoproterozoic (about 570 Ma) as a result of the final breakup of the supercontinent Rodinia, and had achieved a width of about 5000 km by the end of the Cambrian (500–490 Ma). The south-facing margin of ancient North America (Laurentia), which was then situated near the equator in a roughly east–west orientation, defined the northern limit of Iapetus. The northern margin of Gondwana, a large continent centred on the south pole and comprising present-day Africa, Australia, Antarctica, India, and large parts of South America, represented the southern limit of Iapetus (Figure 2). The large Gondwanan landmass was assembled as a result of late Neoproterozoic to Early Cambrian collisions; thus the opening of Iapetus overlapped in time with the final assembly of Gondwana. The Rheic Ocean opened when the Avalonian microcontinent rifted from Gondwana during the Early to Middle Ordovician. The Rheic Ocean was host to several small continental terranes, including Meguma, which were accreted to Laurentia before the final Carboniferous arrival of Gondwana and the formation of the Pangaean supercontinent.

Tectonostratigraphical Divisions

The tectonic architecture and the evolution of the distinctive rock assemblages in the Northern Appalachians are generally described within a framework of tectonostratigraphical zones and subzones. The concept of zonal divisions, based on sharp contrasts in lithology, stratigraphy, fauna, structure, geophysics, plutonism, and metallogeny of Early Palaeozoic and older rocks, was first introduced more than 30 years ago by Harold Williams, following detailed studies of the well-exposed coastal sections in Newfoundland. The identification of these zonal divisions throughout the Northern Appalachians implies a common geological evolution within each zone and highlights the usefulness of each in tectonic analysis.

From west to east, the Northern Appalachians have been divided into the Humber, Dunnage, Gander, Avalon, and Meguma zones (Figure 1). The Humber

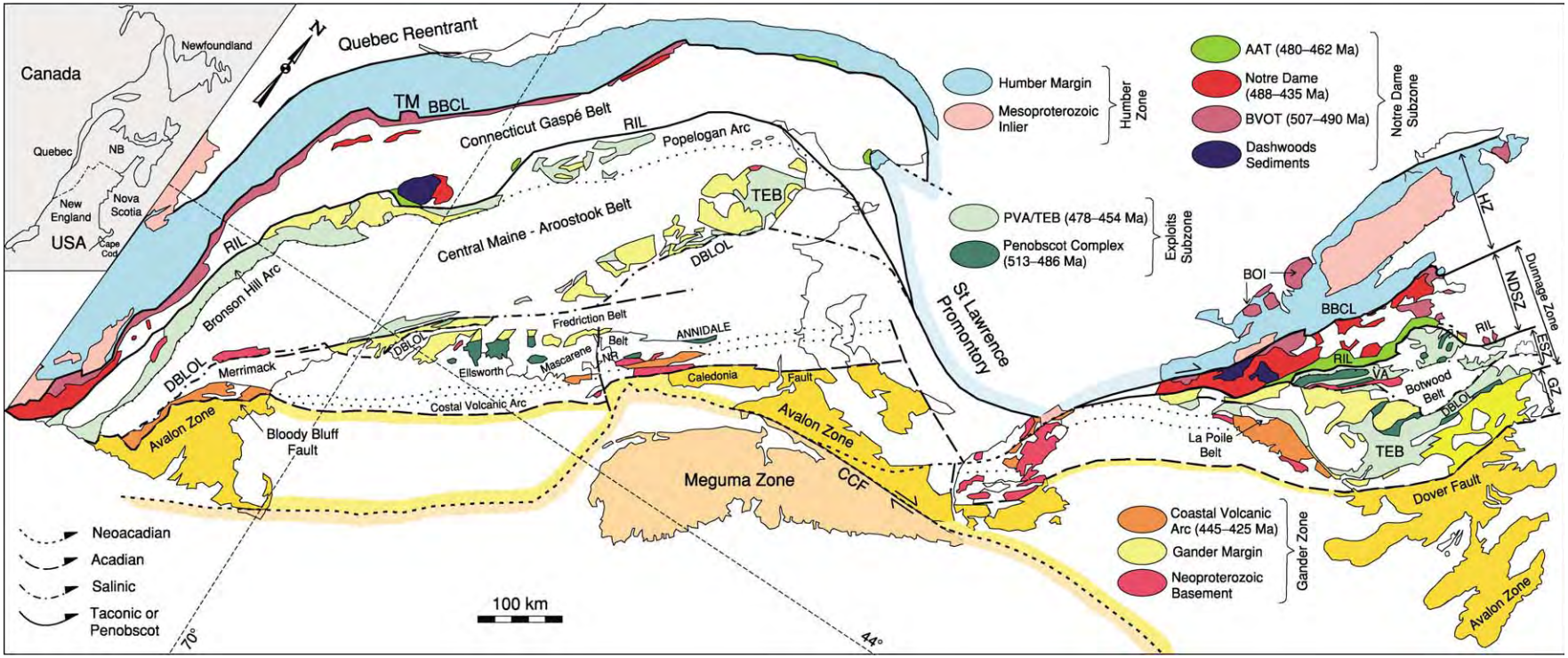


Figure 1 Tectonostratigraphic map of the Northern Appalachians. Zonal divisions are principally based on sharp contrasts in Lower Palaeozoic geology; the main belts of Siluro Devonian sedimentary and volcanic rocks are outlined in white. The major structures and principal orogenic events responsible for their formation are indicated by line patterns. The extension of structures off the land is based on geophysical interpretation and well data. AAT, Annieopsquotch accretionary tract; BBCL, Baie Verte Brompton Cameron Line; BOI, Bay of Islands Ophiolite; BVOT, Baie Verte Oceanic Tract; CCF, Cobequid Chedabucto fault; DBLOL, Dog Bay Liberty Orrington Line; ESZ, Exploits Subzone; GZ, Gander Zone; HZ, Humber Zone; NDSZ, Notre Dame Subzone; NR, New River Belt; RIL, Red Indian Line; PVA, Popelogan Victoria Arc; TEB, Tetagouche Exploits back arc Basin; TM, Thetford Mines Ophiolite; VA, Victoria Arc.

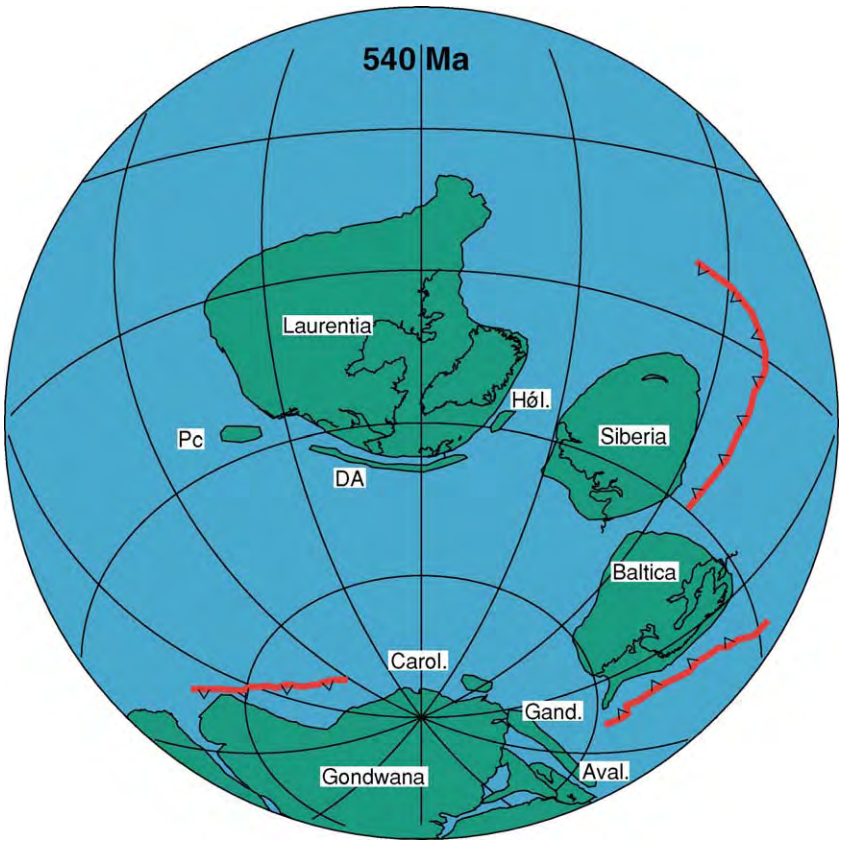


Figure 2 Early Cambrian (around 540 Ma) palaeogeography of the Iapetus Ocean. Seafloor spreading that led to the opening of Iapetus started at around 570 Ma. Aval., Avalonia; Carol., Carolina; DA, Dashwoods; Gand., Ganderia; Hól., Hólanda; Pc, Precordillera Terrane. Meguma was probably situated further east, just outside the field of view, along the north African margin of Gondwana near elements of Armorica. Hólanda and Precordillera are peri Laurentian microcontinents like Dashwoods but ended up in Baltica and the South American part of Gondwana, respectively, after the closure of Iapetus. Made with the help of Conall MacNiocaill.

Zone represents the peripheral part of the Laurentian Craton, which was involved in Appalachian orogenesis. The Gander, Avalon, and Meguma zones represent peri-Gondwanan microcontinents that accreted to Laurentia during the Early to Middle Palaeozoic (450–380 Ma). The Dunnage Zone contains the remnants of oceanic terranes that formed within the realm of the Iapetus Ocean, and is subdivided into the Notre Dame and Exploits subzones.

Orogenesis (the combined effects of spatially and temporally associated deformation, metamorphism, magmatism and sedimentation) was mainly confined to the central portion of the Northern Appalachians, often referred to as the Central Mobile Belt.

Humber Zone

The Humber Zone is mostly underlain by a Late Neoproterozoic–Ordovician rifted passive margin sequence deposited on Laurentian basement that had experienced multiple orogenic events during the Mesoproterozoic (around 1.5–0.95 Ga), culminating in the Grenville orogeny (around 1.2–0.95 Ga). These basement rocks are exposed in several windows from Newfoundland to southern New York state. Its Late Neoproterozoic–Palaeozoic cover preserves evidence of rifting, development of a warm-water carbonate shelf and slope, and conversion into a convergent margin. Rift-related magmatism took place intermittently during an approximately 70 Ma time-span during the latest Neoproterozoic (620–550 Ma) and was coeval with the deposition of consanguineous clastic sedimentary rocks. Exactly when during this time interval the opening of the Iapetus Ocean began is at present unclear, but palaeomagnetic and geological arguments suggest that it occurred at around 570 Ma. Several lines of evidence suggest that the final stage of rift magmatism (555–550 Ma) and the subsequent Early Cambrian (around 540 Ma) transition from rift to drift sedimentation preserved in Newfoundland, Quebec, and New England (Figure 3), relates to the departure of a small microcontinent, referred to as Dashwoods. This event thus postdates the seafloor spreading associated with the earlier opening of Iapetus. Dashwoods equates

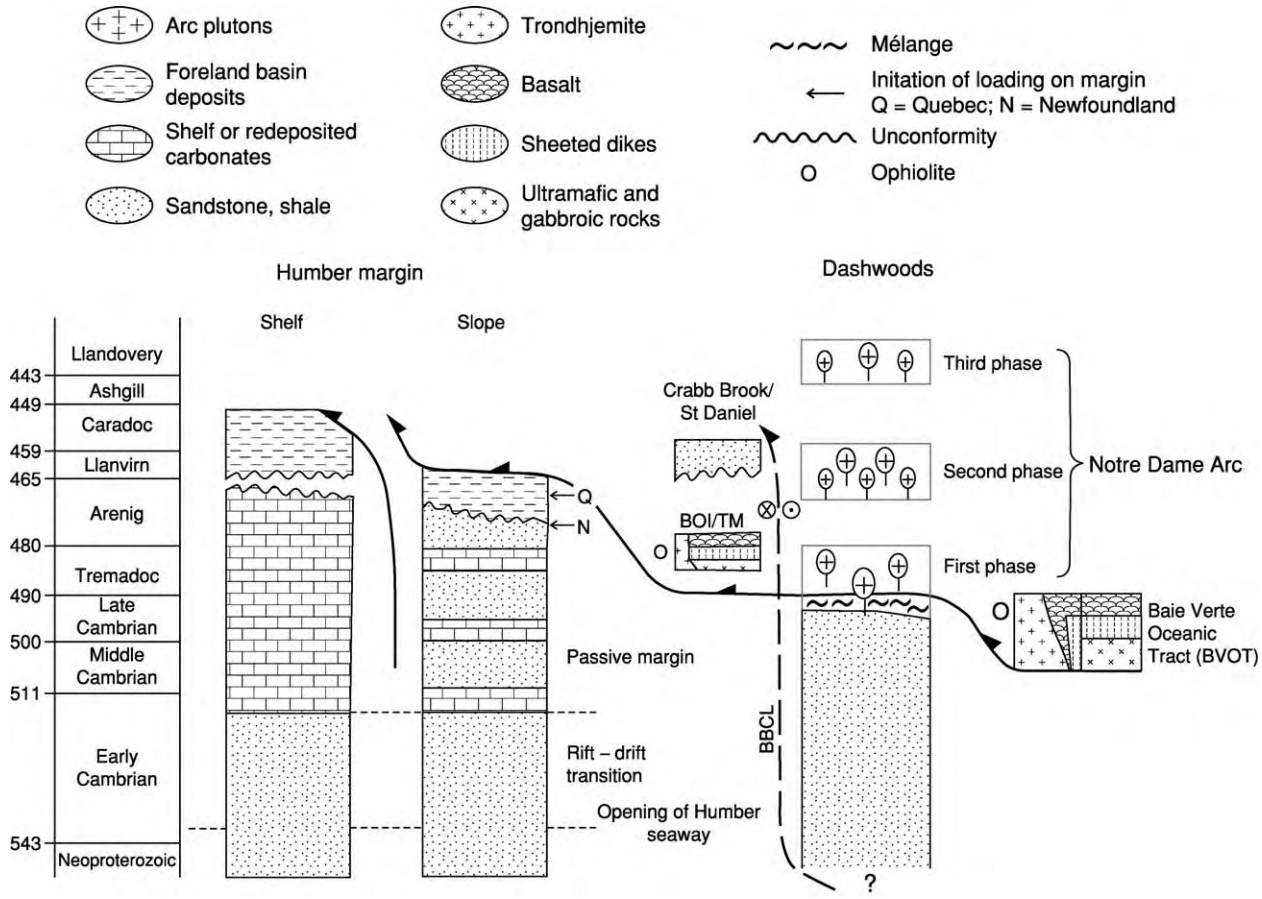


Figure 3 Tectonostratigraphic relationships pertinent to the convergence between the Baie Verte Oceanic Tract and Dashwoods, and between Dashwoods and the Humber margin. East directed subduction started in the Middle Cambrian immediately outboard of the Dashwoods microcontinent and led to underthrusting of the Dashwoods microcontinent beneath the Baie Verte Oceanic Tract (BVOT). The Bay of Islands (BOI) and Thetford Mines (TM) ophiolites formed during local rifting and concurrent seafloor spreading of the allochthonous BVOT after the Dashwoods microcontinent was subducted beneath the BVOT and the east directed subduction zone stepped back into the Humber seaway. The Crabb Brook and St Daniel were deposited in piggyback basins formed during obduction of the ophiolites onto the Humber margin. BBCL, Baie Verte Brompton Camerons Line.

to the Boundary Mountain Terrane in New England. Spreading between Laurentia and the Dashwoods microcontinent was either very slow or aborted soon after formation, because faunal and palaeomagnetic data suggest that the Dashwoods was separated from Laurentia by only a narrow oceanic seaway (the Humber Seaway). The along-strike extent of the Dashwoods is uncertain. Isotopic evidence for the involvement of Laurentian crust in the generation of Early to Middle Ordovician magmatic rocks of the Notre Dame Arc and Laurentian rift-related clastic rocks positioned in an upper plate setting during the Taconic Orogeny (Figures 1 and 3) suggests that this microcontinent extended southward into New England.

The conversion of the Humber passive margin into a convergent margin is indicated by cratonward migration of a foreland basin formed in response to tectonic loading of the margin by an overriding ophiolite terrane (Figure 3). Foreland-basin development was immediately preceded by uplift and local karst erosion due to the passage of a peripheral bulge across the margin. This process started in the Middle to Late Arenig (around 474 Ma) and lasted at least until the Late Caradoc (around 450 Ma) in Newfoundland. Loading of the margin started slightly later in Quebec and New England (Figure 3) probably owing to the presence of the Quebec Reentrant–St Lawrence Promontory pair (Figure 1).

Dunnage Zone

Palaeomagnetic, fossil, and other geological evidence indicates that the oceanic terranes of the Dunnage zone (Figure 1) can be separated into peri-Laurentian (Notre Dame Subzone) and peri-Gondwanan (Exploits Subzone). Detailed geochemical studies have revealed that virtually all the preserved oceanic rocks have compositions typical of suprasubduction zone

settings (i.e. fore-arc, arc, and back-arc). True oceanic lithosphere formed at mid-oceanic spreading centres, far removed from continental margins and subduction zones, is either very rare or has been lost during subduction.

Notre Dame Subzone The Notre Dame Subzone (Figure 1) lies immediately east of the Humber Zone, from which it is separated by a complex and long-lived fault zone, the Baie Verte–Brompton–Camerons Line. The Notre Dame Subzone comprises several unrelated Middle Cambrian to Middle Ordovician (507–462 Ma) oceanic terranes and an important continental magmatic arc (the Notre Dame Arc), which was intermittently active between 488 and 433 Ma (Figure 3). Rocks of the Notre Dame Subzone can be traced from south-western Connecticut through western New England and Quebec into Newfoundland (Figure 1), where they are best exposed and studied. Rocks belonging to the Notre Dame Arc are commonly referred to as the Shelburne Falls Arc in New England. The eastern boundary of the Notre Dame Subzone with the Exploits Subzone is marked by a major suture (the Red Indian Line) that juxtaposes rocks formed on opposite sides of the Iapetus Ocean (Figure 4).

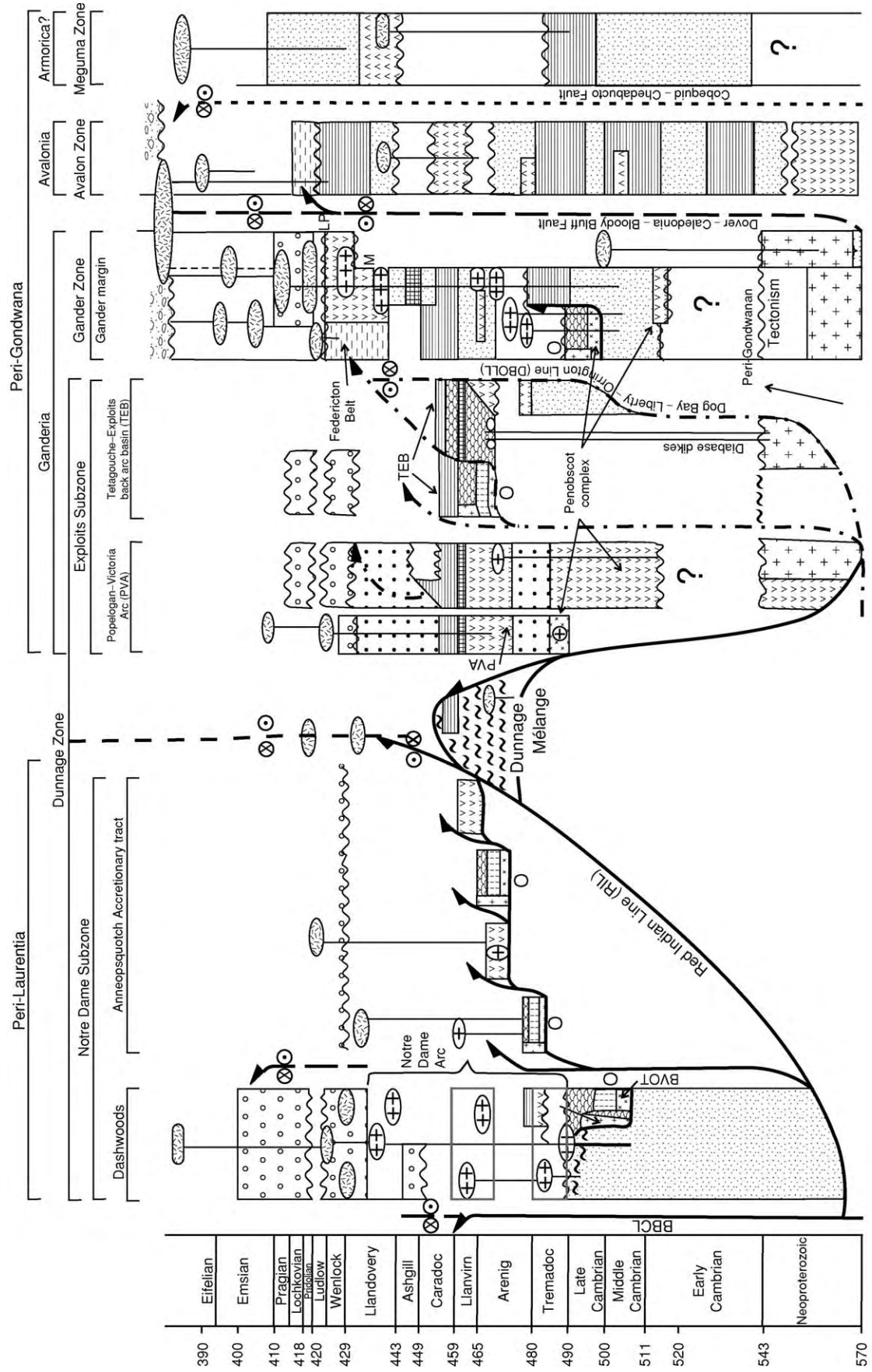
The oldest oceanic rocks in the Notre Dame Subzone occur in the Middle to Upper Cambrian (around 507–490 Ma) Baie Verte Oceanic Tract, which mainly represents an infant arc terrane (wide zone of supra-subduction-zone oceanic lithosphere created during and shortly after the initiation of subduction, before the establishment of a well-defined linear narrow arc axis). The Baie Verte Oceanic Tract was obducted on to the Dashwoods shortly after formation owing to eastward-directed (present coordinates) subduction initiated close to Laurentia. Stitching plutons of the first phase of the Notre Dame Arc (around 488 Ma) indicate that obduction in Newfoundland took place during the latest Cambrian. Rare evidence of Late Cambrian tectonometamorphic events and the age range of Early Ordovician arc plutonism in New England and Quebec suggest that there was little diachroneity in this process along the length of the Northern Appalachians. The first phase of Notre Dame Arc magmatism resulted from the stepping back of subduction into the Humber seaway, transferring the Dashwoods microcontinent from a lower-plate to an upper-plate setting (Figure 3). Closure of the Humber Seaway brought Dashwoods back to Laurentia in the late Arenig (475–470 Ma) and started the Taconic orogeny. The Taconic Orogeny was accompanied by the second phase of Notre Dame Arc magmatism (469–458 Ma) and intense deformation and Barrovian metamorphism of the Dashwoods, including the rocks generated during the first phase of the Notre Dame Arc development.

A younger, unrelated, oceanic terrane, the Annieopsquotch accretionary tract, is situated along the eastern margin of the Notre Dame Subzone, immediately west of the Red Indian Line (Figure 4). The Annieopsquotch accretionary tract comprises a tectonic collage of Arenig–Llanvirn (480–462 Ma) infant arc, arc, and back-arc terranes that formed as a result of west-directed subduction outboard of the Dashwoods. Ultramafic rocks are rarely preserved in the Annieopsquotch accretionary tract, in contrast to the Baie Verte Oceanic Tract, probably because their accretion to the Dashwoods involved underthrusting and structural underplating rather than obduction. Accretion of the Annieopsquotch accretionary tract to Laurentia started at around 470 Ma, suggesting a causal relationship to the Taconic, Laurentia–Dashwoods, collision further west, and ended before 455–450 Ma, when the leading edge of the Exploits subzone (Popelogan–Victoria Arc) started to accrete to Laurentia. The Annieopsquotch accretionary tract is poorly preserved in New England, mainly because the Red Indian Line is here largely buried and/or obscured by Mid-Palaeozoic cover sequences and tectonism. The Red Indian Line approximately follows the eastern boundary of the Connecticut–Gaspé Belt (Figure 1) and is locally marked in Maine by mélanges and ophiolitic rocks.

The Notre Dame Subzone experienced both uplift and subsidence during the Late Ordovician and Silurian, and is locally overlain by a marine to terrestrial Silurian volcanosedimentary sequence. Terrestrial rocks are dominant in Newfoundland, while marine deposits characterize New England and Quebec.

Notre Dame Arc magmatism had ended by the early Silurian and was replaced by a phase of Early to Late Silurian (433–424 Ma) bimodal within-plate volcanism and plutonism; the latter has been ascribed to breakoff of the west-dipping, down-going slab responsible for accretion of the Annieopsquotch accretionary tract, the Exploits Subzone, and the Gander Zone. Early Devonian (Acadian) orogenesis, very penetrative in the adjacent Exploits Subzone and Gander Zone, is weak or absent, showing that the locus of orogenesis had shifted to the south-east by Late Silurian times. Mild orogenesis returned by the Middle to Late Devonian (Neoacadian, 395–350 Ma).

Exploits Subzone The oldest known subduction-related rocks preserved in the Exploits Subzone have Lower Cambrian to Tremadoc (513–486 Ma) ages and compositions typical of an arc or back-arc setting. Some of the rocks have compositions and stratigraphies typical of ophiolite complexes, while


Peri-Laurentia
Peri-Gondwana
Dunnage Zone
Ganderia
Notre Dame Subzone
Exploits Subzone
Gander Zone
Avalon Zone
Meguma Zone
Dashwoods
Anneopsquotch Accretionary tract
Tetagouche–Exploits back arc basin (TEB)
Popelogan–Victoria Arc (PVA)
Gander margin
Avalonia
Armorica?
Cobequid – Chedabucto Fault
Dover – Caledonia – Bloody Bluff Fault
Federicton Belt
Dog Bay – Liberty – Orangeton Line (DBOLL)
Diabase dikes
Penobscot complex
Peri-Gondwanan Tectonism
TEB
PVA
Dunnage Mélange
Red Indian Line (RIL)
Notre Dame Arc
BVOT
BBCL
LP
M
O
?
Eifelian
Emsian
Pragian
Lochkovian
Pridolian
Ludlow
Wenlock
Llandovery
Ashgill
Caradoc
Llanvirn
Arenig
Tremadoc
Late Cambrian
Middle Cambrian
Early Cambrian
Neoproterozoic
390
400
410
418
420
429
443
449
459
465
480
490
500
511
520
543
570

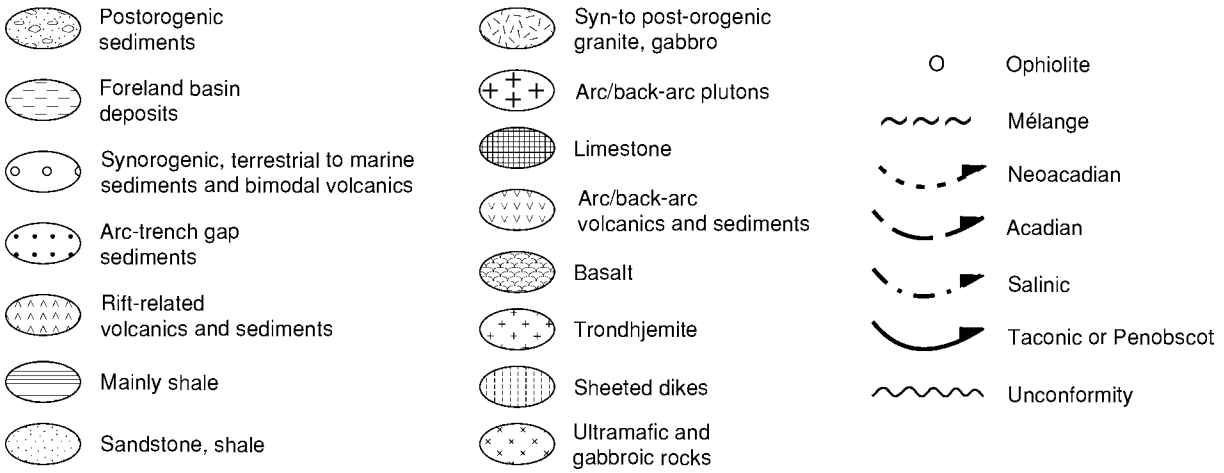


Figure 4 Tectonostratigraphic relationships pertinent to the closure of the main tract of the Iapetus Ocean, the Tetagouche Exploits back arc basin, and the oceanic crust that separated Ganderia from Avalonia and Avalonia from Meguma. The Red Indian Line marks the site of the collision between the Notre Dame and Popelogan Victoria arcs. The line patterns of the major structures correspond to those in **Figure 1**. The arrows on the faults indicate the polarity of subduction and/or thrusting. Obliquity of collision and subsequent strike slip fault reactivation is indicated by the symbols on each side of the major faults: a point represents moving towards and a cross moving away from the viewer. BBCL, Baie Verte Brompton Camerons Line. BVOT, Baie Verte Oceanic Tract. LP, La Poile Belt volcanic rocks; M, Mascarene belt; PVA, Popelogan Victoria Arc; TEB, Tetagouche Exploits back arc basin.

others have close spatial associations with felsic-arc plutonic rocks that are at least as old as 565 Ma (Figure 4). The latter probably represent ensialic basement and suggest that at least some of these rocks are not truly oceanic. Coeval and lithologically similar Cambrian arc or back-arc volcanic rocks and associated Neoproterozoic plutonic rocks also occur in the Gander Zone (see below), particularly in southern New Brunswick (e.g. Annidale and New River Belts) and adjacent Maine (Ellsworth Belt) (Figure 1). Collectively, these rocks have been interpreted as an arc–back-arc complex, the Penobscot Complex, formed near the leading edge of the Gondwanan margin on which the arenites and shales of the Gander Zone (see below) were deposited. Stitching plutons and isotopic evidence show that the Cambrian oceanic rocks of the Penobscot Complex were obducted onto the continental-margin rocks of the Gander Zone before 474 Ma, during the Penobscot Orogeny (Figure 4).

The composite crust of the Penobscot Complex and the Gander Zone sedimentary rocks are disconformably overlain by younger rocks related to the Arenig–Llanvirn ensialic Popelogan–Victoria Arc/Tetagouche–Exploits back-arc system, which was active between 478 and 454 Ma. The Popelogan–Victoria Arc/Tetagouche–Exploits back-arc system has been traced from Newfoundland through central New Brunswick into Maine, mainly on the basis of a remarkably consistent and distinct lithological association of coeval plutonic, volcanic, and sedimentary rocks. Correlations further to the south into Massachusetts become progressively more difficult, owing to a lack of age control and the presence of substantial cover sequences. The Popelogan–Victoria Arc is probably continuous with the Bronson Hill Arc, whereas most of the products of the Tetagouche–Exploits back-arc have been buried beneath the Siluro-Devonian cover sequences of the Central Maine and Merrimack Belts (Figure 1). The Popelogan–Victoria Arc had migrated to a mid-Iapetus position by at least the late Llanvirn (around 462 Ma), according to faunal and sparse palaeomagnetic evidence, while the passive-margin side of the Tetagouche–Exploits back-arc, which was deposited on Gander-Zone rocks, was positioned further south owing to subduction-zone retreat and concurrent back-arc spreading. Rifting and seafloor spreading in the Tetagouche–Exploits back-arc is necessary to account for the formation of Llanvirn back-arc oceanic crust (464–459 Ma), which is locally preserved as incomplete ophiolite complexes and large structural slices of highly tectonized back-arc oceanic basalt and gabbro in mélange belts in New Brunswick and Newfoundland. Fossil and palaeomagnetic evidence suggest that the Tetagouche–Exploits back-arc nowhere achieved a width of more than about 1000 km (largely based on the assumption that the pelagic larvae of animals such as brachiopods could traverse oceans up to this width, but no further).

Gander Zone

The Gander Zone (Figures 1 and 4) is defined by a distinct monotonous sequence of Lower Cambrian to

Tremadoc (520–480 Ma) arenites and shales, capped by Tremadoc black shale. This sequence can be traced from north-east Newfoundland into New England and is generally accepted to represent the outboard part (outer shelf to slope) of a passive margin built on Gondwanan continental crust. This inference is based on an extensive set of isotope, geochemical, detrital mineral, palaeomagnetic, and fossil data. The sedimentary source area has to be cratonic and contain Archean, Palaeo-, Meso-, and Neoproterozoic rocks. Based on a process of elimination, this craton is generally inferred to be Amazonia, although other Gondwanan sources cannot be ruled out.

With the possible exception of a few outcrops in southern New Brunswick, the basement on which this sedimentary sequence was deposited is not exposed. However, spatially associated Late Neoproterozoic to Early Cambrian arc rocks sporadically exposed in the Gander Zone are, on the basis of distinct isotope, geochemical, and age data, inferred to represent basement. The basement rocks show many similarities in their Neoproterozoic history to coeval rocks of the adjacent Avalon Zone, from which they are separated by a major long-lived ductile–brittle fault zone (Dover–Caledonia–Bloody Bluff Fault), but have experienced a very different Cambrian to Early Devonian evolution. Avalon developed a shale-rich platformal sedimentary succession during the Early Palaeozoic and did not experience orogenesis until the Latest Silurian–Early Devonian Acadian Orogeny. Gander, on the other hand, experienced multiple phases of arc magmatism and orogenesis from the Cambrian to the Silurian (Figure 4). Furthermore, Nd-isotope data indicate that their Precambrian basements are strikingly different. Hence, the Gander and Avalon zones are considered to represent two distinct Gondwanan microcontinents during the Palaeozoic (Ganderia and Avalonia, respectively). Ganderia was accreted to Laurentia in two stages. The leading edge of Ganderia, with the Popelogan–Victoria Arc suprastructure (Figure 4), accreted to Laurentia in the Late Ordovician. Convergence between Ganderia and Laurentia continued, leading to final closure of the Tetagouche–Exploits back-arc by the late Early Silurian (430–425 Ma). During this collision with Laurentia, a Late Ordovician to late Early Silurian (445–425 Ma) extensional arc–back-arc system (coastal volcanic arc/Mascarene–La Poile back-arc basin) was constructed on its southern margin (Figures 1 and 4) owing to convergence between Avalonia and Ganderia. This arc–back-arc system forms a narrow volcanic–plutonic belt that roughly follows the coastline of New England and maritime Canada to southern Newfoundland.

Avalon Zone

The exposed rocks of the Avalon Zone mainly represent the suprastructure of the microcontinent Avalonia. Its distinctive rocks have been traced from the British Isles into Massachusetts, and mainly comprise Neoproterozoic dominantly juvenile arc-related volcanosedimentary successions and associated plutonic rocks, which experienced a complicated and long-lived Neoproterozoic tectonic history that largely predates the opening of Iapetus. Characteristic of the Avalon Zone is the deposition of a Cambrian–Lower Ordovician shale-rich platformal sedimentary sequence that locally includes rift-related volcanic rocks (Figure 4). Fossil and palaeomagnetic data show a strong connection with Gondwana during this period. However, starting in the late Early Ordovician, the data indicate that Avalon started to drift north independently across Iapetus. Early to Middle Ordovician rift-related volcanic rocks (479–460 Ma), preserved in the Avalonian successions of Nova Scotia and New Brunswick, support this assertion. The accretion of Avalonia to Ganderia, and by implication Laurentia, is poorly constrained, because the oldest preserved sedimentary and magmatic (stitching plutons) linkages have Middle to Late Devonian ages. Palaeomagnetic and structural evidence, particularly relating to the Late Silurian (422–418 Ma) start of the inversion of the Mascarene–La Poile back-arc basin on the southern edge of Ganderia and the ages of syntectonic plutons, suggest however that collision started during the latest Silurian and culminated in the mainly Early Devonian (418–395 Ma) Acadian Orogeny.

Meguma Zone

The Meguma Zone represents the most outboard terrane in the northern Appalachians. It is exposed on land only in southern Nova Scotia, but its regional extent is much larger. It has been traced offshore by an impressive set of geophysical and well data from the southernmost part of the Grand Banks south-east of Newfoundland, across the Scotian Shelf and the Gulf of Maine, to southernmost Cape Cod. The basal part of the exposed Meguma zone comprises a thick latest Neoproterozoic to Early Ordovician turbiditic sandstone–shale sequence of the Meguma Supergroup, which was deposited on the continental rise and/or slope to outer shelf of a Gondwanan passive margin. A combination of detrital-zircon, sedimentological, and sparse fossil data suggest a north-west African provenance, but the dataset at present is small and other parts of Gondwana cannot be ruled out as source areas. The Meguma Supergroup is disconformably overlain by Late Ordovician

to Early Devonian shallow-marine dominantly siliciclastic sedimentary rocks, which are locally interlayered with rift-related bimodal Early Silurian (442–438 Ma) volcanic rocks. The Meguma Zone experienced intense orogenesis during the late Early Devonian to Early Carboniferous (395–350 Ma; Neoacadian orogenesis). Fossil evidence (fish and crinoid) suggests that, during the Late Silurian, Meguma was close to Avalonia and/or Baltica, and separated from Gondwana. When combined with the magmatic evidence for Late Ordovician rifting and Neoacadian orogenesis, this suggests that Meguma was a microcontinent or part thereof (perhaps Armorica) during at least the Silurian and Devonian, as Gondwana was not accreted to Meguma until the Carboniferous–Permian Alleghanian Orogeny.

Overview and Summary of the Tectonic Evolution of the Northern Appalachians

The oldest orogenic events associated with the closure of Iapetus are the Late Cambrian–Early Ordovician (495–484 Ma) accretion of the Baie Verte Oceanic Tract to the Dashwoods and the Penobscot complex to Ganderia, with the Baie Verte Oceanic Tract and the Penobscot complex both in upper-plate settings (Figures 3 and 4). Exposed, inferred Ganderian, basement locally contains evidence of older Early Cambrian orogenesis, but this is probably related to peripheral subduction around the Gondwanan continent outside the realm of Iapetus. The accretions of the Baie Verte Oceanic Tract and the Penobscot complex were nearly coeval, but took place on opposite sides of the Iapetus Ocean, when the ocean was at its widest. Because these events took place shortly after the initiation of subduction on either side of the ocean, they signal a major plate reorganization that changed Iapetus from an Atlantic-type ocean into a Pacific-type ocean. Time constraints provided by stitching plutons indicate that both accretionary events were short-lived, forming mélanges and ductile–brittle deformation, and did not lead to any significant regional metamorphism.

The first major orogenic event recorded in the Northern Appalachians was the Early to Middle Ordovician (474–455 Ma) Taconic Orogeny, which principally resulted from closure of the Humber seaway and accretion of the Dashwoods microcontinent to Laurentia. Closure of the Humber Seaway initiated loading of the Humber margin, initially by emplacement of oceanic terranes, some of which are younger than the Baie Verte Oceanic Tract (484–479 Ma). These younger oceanic terranes, which include the Bay of Islands and Thetford Mines ophiolites in Newfoundland and Quebec, respectively, formed either during the initiation of subduction in the Humber Seaway or as a result of localized pericratonic seafloor spreading in pull-aparts or above embayments in the downgoing plate. The Taconic Orogeny culminated in imbrication, folding, and locally intense Barrovian metamorphism, mainly in upper-plate rocks associated with the Dashwoods. Age constraints on Taconic metamorphism (470–455 Ma) are remarkably consistent along the length of the orogen.

Docking of Dashwoods was rapidly followed by accretion of the Annieopsquotch accretionary tract between 470 and 460 Ma and the Popelogan–Victoria Arc at 455–450 Ma to composite Laurentia along its ocean-facing eastern margin. The effects of these accretionary events, which are centred on the Red Indian Line, are difficult to separate from those of the orogenesis accompanying the Dashwoods–Laurentia collision to the west. Since separation of these events on the basis of kinematic arguments is often impossible outside Newfoundland, these events are included in the Taconic orogeny. Structural relationships and the presence of young supra-subduction-zone oceanic rocks (480–464 Ma) in the Annieopsquotch accretionary tract indicate that a new west-dipping subduction zone had formed in the Arenig to the east of Dashwoods. Accretion involved south-east-verging duplex-style thrust complexes, folding, mélange development, and a sinistral oblique reverse shear zones.

The Popelogan–Victoria Arc was formed in the Early Arenig (at about 478 Ma) on the leading edge of Ganderia after the Penobscot orogeny. It was active until the Late Caradoc and formed above an east-dipping subduction zone. Upper-plate extension and rifting as a result of slab rollback dispersed the Popelogan–Victoria Arc into Iapetus and formed a wide (up to 1000 km) Japan Sea-style back-arc basin (Tetagouche–Exploits back-arc) that was partly underlain by Llanvirn oceanic crust.

The opposite subduction polarities associated with the Annieopsquotch accretionary tract (west-dipping) and the Popelogan–Victoria Arc (east-dipping) require that the Red Indian Line marks a collision between two arc terranes (Figure 4). Structural and seismic-reflection data in central Newfoundland indicate that the Popelogan–Victoria Arc was partly subducted beneath the Annieopsquotch accretionary tract. Palaeomagnetic evidence suggests that between 480 and 450 Ma these two subduction zones consumed approximately 3000 km of ocean, implying a minimum closure rate of around 10 cm yr^{-1} in the main Iapetus tract. The remaining tracts of Iapetus (Humber Seaway, Tetagouche–Exploits back-arc, and the ocean

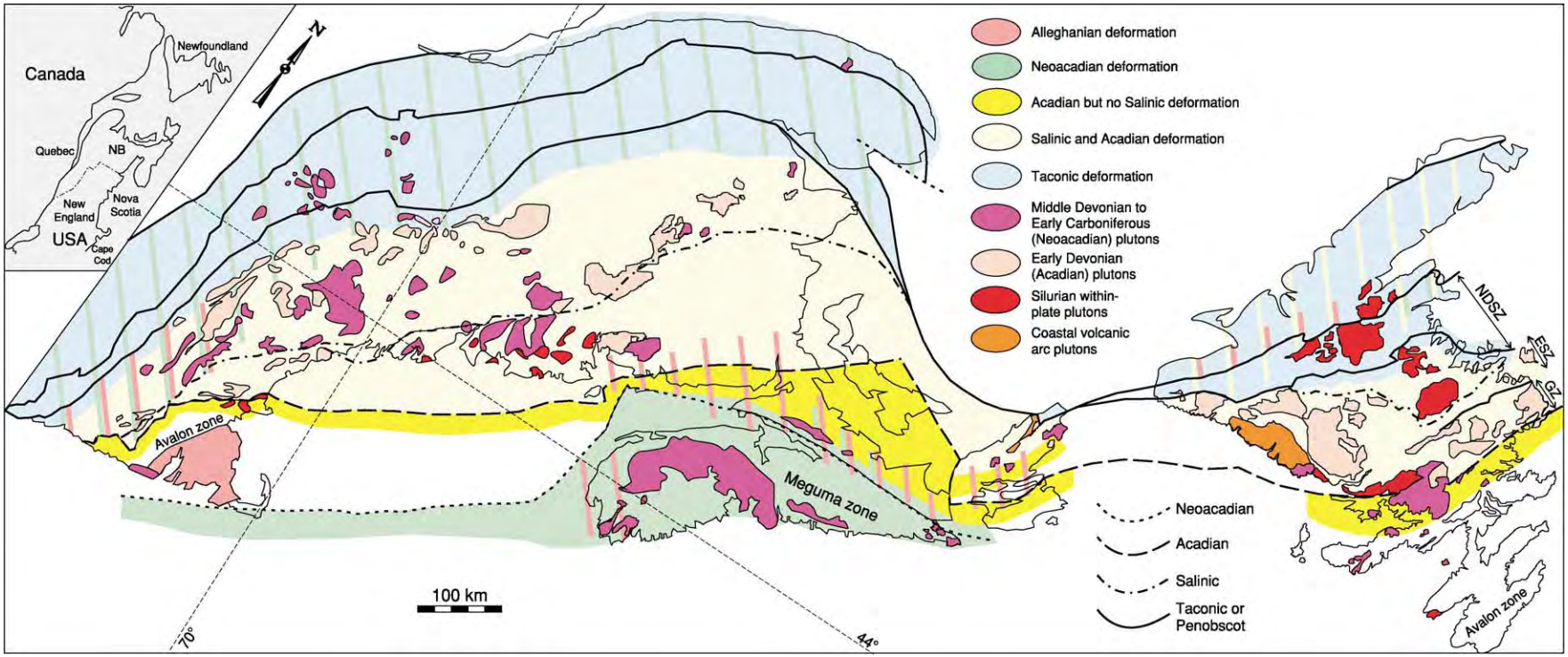


Figure 5 Map of the Northern Appalachians showing the spatial and temporal distributions of deformation, associated metamorphism, and Siluro Early Carboniferous plutons. For abbreviations see **Figure 1**.

basin separating Avalonia and Ganderia) had a combined width in the order of 2000 km.

After accretion of the Popelogan–Victoria Arc to Laurentia, the west-dipping subduction zone stepped back into the Tetagouche–Exploits back-arc. Late Ordovician–Early Silurian mélanges, blueschists, and foredeep deposits leave little doubt that subduction continued. Closure of the Tetagouche–Exploits back-arc in the Early Silurian (around 430 Ma) accreted the bulk of Ganderia firmly to Laurentia along the Dog Bay–Liberty–Orrington Line. This event, the Salinic Orogeny, was contemporaneous in Newfoundland, New Brunswick, and Maine. Although mainly restricted to rocks east of the Red Indian Line, localized Salinic-age deformation west of this suture is probably related (Figure 5).

Sinistral oblique accretion of Avalonia to Laurentia started at around 421 Ma, shortly after the docking of Ganderia. This caused inversion of the Mascarene–La Poile back-arc basin in southern New Brunswick, Maine, and Newfoundland and led to penetrative orogenesis of Ganderian, including its Siluro-Devonian cover, and, to a much lesser extent, immediately adjacent Avalonian rocks. This collision is the Acadian Orogeny as defined in its type area along the south coast of New Brunswick and neighbouring Maine. It produced polyphase folding including an early phase of large south-east-verging fold nappes and thrusts, a north-west-verging retrowedge also involving fold nappes and ductile thrusts, high-temperature metamorphism, and extensive syntectonic granitic plutonism with bimodal magmatism developing later on. Acadian orogenesis finished before the end of the Early Devonian and did not extend west of the Red Indian Line, other than as localized faulting and development of clastic wedges in foreland basins. Subsequent late Early Devonian to Early Carboniferous Neoacadian orogenesis is present throughout most of the Northern Appalachians (Figure 5) and is at least in part related to the docking of Meguma to the Laurentian margin. The Neoacadian Orogeny was accompanied by a westerly progradation of clastic wedges and granitic magmatism in western New England and maritime Canada and widespread dextral transpression.

The Carboniferous–Permian Alleghanian Orogeny was due to tectonic processes related to the oblique convergence and collision of Gondwana with Laurentia, which is in an upper-plate setting in the Northern Appalachians. Dextral strike-slip faulting and the development of intermontane transtensional basins prevailed during the Early Carboniferous. Transpressional inversion of these basins involved several stages during the late Early Carboniferous to Middle Permian and was accompanied locally by intense deformation and metamorphism near the boundary between the Avalon and Meguma zones. Intense late Alleghanian orogenesis accompanying renewed north-directed underthrusting of Avalonian and Ganderian rocks beneath Laurentia occurred in southern New England, but is absent elsewhere.

The temporal and spatial relationships of the tectonic events discussed above indicate that the Northern Appalachians are an example of a Palaeozoic accretionary orogen, punctuated by the accretion of small suprasubduction-zone terranes and five collisional events related to the arrival of four microcontinents (Dashwoods, Ganderia, Avalonia, and Meguma) and finally Gondwana at the Laurentian margin. Laurentia was thus progressively expanding oceanwards concurrent with an eastward shift of the locus of orogenesis (Figure 5).

Glossary

Boundary Mountain Terrane Equivalent of Dashwoods microcontinent in New England-Southern Quebec. Is inferred to underlie most of the Connecticut-Gaspè Belt

Ganderia Gondwanan-derived microcontinent. Characterized by island arc and back-arc magmatism during the Early Palaeozoic.

See Also

Europe: Caledonides of Britain and Ireland. **North America:** Southern and Central Appalachians. **Plate Tectonics**. **Sedimentary Environments:** Depositional Systems and Facies. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Faults; Folding; Mountain Building and Orogeny; Ocean Trenches.

Further Reading

Barr SM, White CE, and Miller BV (2002) The Kingston Terrane, southern New Brunswick, Canada: evidence for an Early Silurian volcanic arc. *Geological Society of America Bulletin* 114: 964 982.

Dunning G, O'Brien SJ, Colman Sadd SP, *et al.* (1990) Silurian orogeny in the Newfoundland Appalachians. *Journal of Geology* 98: 895 913.

Fortey RA and Cocks LRM (2003) Palaeontological evi dence on global Ordovician Silurian continental recon structions. *Earth Science Reviews* 61: 245 307.

Karabinos P, Samson SD, Hepburn JC, and Stoll HM (1998) Taconian orogeny in the New England Appalach ians: collision between Laurentia and the Shelburne Falls arc. *Geology* 26: 215 218.

Kerr A (1997) Space time relationships among Appalachian cycle plutonic suites in Newfoundland. In: Sinha AK, Whalen JB, and Hogan JP (eds.) *The Nature of Magmatism in the Appalachian Orogen,* pp. 193 220.

Geological Society of America Memoir 191, Boulder, Colorado: Geological Society of America.

Kim J and Jacobi RD (2002) Boninites: characteristics and Appalachian constraints, north eastern Appalachians. *Physics and Chemistry of the Earth* 27: 109 147.

Robinson P, Tucker RD, Bradley D, Berry HN IV, and Osberg PH (1998) Paleozoic orogens in New England, USA. *Geologiska Föreningens Stockholm Forhandlingar* 120: 119 148.

Swinden HS, Jenner GA, and Szybinski ZA (1997) Magmatic and tectonic evolution of the Cambrian Ordovician Laurentian margin of Iapetus: Geochemical and isotropic constraints from the Notre Dame Subzone, Newfoundland. In: Sinha K, Whalen JB, and Hogan JP (eds.) *The nature of magmatism in the Appalachian Orogen,* pp. 337 365. Geological Society of America Memoir 191, Boulder, Colorado: Geological Society of America.

Van der Pluijm BA, Johnson RJE, and van der Voo R (1993) Paleogeography, accretionary history, and tectonic scen ario: a working hypothesis for the Ordovician and Silur ian evolution of the northern Appalachians. In: Roy DC and Skehan JW (eds.) *The Acadian Orogeny,* pp. 27 40. Geological Society of America Special Paper 275, Boulder, Colorado: Geological Society of America.

Van Staal CR, Sullivan RW, and Whalen JB (1996) Proven ance and tectonic history of the Gander Margin in the Caledonian/Appalachian Orogen: implications for the origin and assembly of Avalonia. In: Nance RD and Thompson MD (eds.) *Avalonian and Related Peri Gondwanan Terranes of the Circum North Atlantic,* pp. 347 367. Geological Society of America Special Paper 304, Boulder, Colorado: Geological Society of America.

Van Staal CR, Dewey JF, Mac Niocaill C, and McKerrow WS (1998) The Cambrian Silurian tectonic evolution of the northern Appalachians: history of a complex, south west Pacific type segment of Iapetus. In: Blundell DJ and Scott AC (eds.) *Lyell: the Past is the Key to the Present,* pp. 199 242. Geological Society Special Publication 143, London: Geological Society.

Waldron JF and van Staal CR (2001) Taconian orogeny and the accretion of the Dashwoods block: a Peri Laurentian microcontinent in the Iapetus Ocean. *Geology* 29: 811 814.

Whalen JB, Jenner GA, Longstaffe FJ, Gariépy C, and Fryer BJ (1997) Implications of granitoid geochemical and iso topic (Nd, O, Pb) data from the Cambrian Ordovician Notre Dame arc for the evolution of the Central Mobile belt, Newfoundland Appalachians. In: Sinha K, Whalen JB, and Hogan JP (eds.) *The Nature of Magmatism in the Appalachian Orogen,* pp. 367 395. Geological Society of America Memoir 191, Boulder, Colorado: Geological Society of America.

Williams H (ed.) (1995) *Geology of the Appalachian Caledonian Orogen in Canada and Greenland*. Geo logical Survey of Canada, Geology of Canada, no. 6 (also Geological Society of America, The Geology of North America, v. F1).

Williams H, Currie KL, and Piasecki MAJ (1993) The Dog Bay Line: A major Silurian tectonic boundary in north east Newfoundland. *Canadian Journal of Earth Sciences* 30: 2481 2494.

Atlantic Margin

D R Hutchinson, US Geological Survey, Woods Hole, MA, USA

Published by Elsevier Ltd.

Introduction

The North American Atlantic Continental Margin stretches between Florida and Newfoundland and extends from on land, near the exposed Mesozoic Newark basins, to offshore, near the western limit of the abyssal plain. It is a classic example of a passive or trailing-edge continental margin, in which continental rifting followed by seafloor spreading, sedimentation, and subsidence are the primary controls on margin morphology, tectonics, and sedimentary style. At almost 4500 km long, the margin considered here spans climatic zones from subtropical off the coast of Florida to subarctic off the coast of Newfoundland. During Mesozoic continental separation, the rifting style changed from volcanic, during the North America–Africa separation, to non-volcanic, during the North America–Europe separation. Sediments in the five major offshore basins contain varying volumes and ages of evaporite, carbonate, coarse terrigenous, and fine siliciclastic deposits.

This margin is documented by one of the most thorough high-quality publicly available geological and geophysical datasets of any continental margin in the world. Exploration began in the 1960s, when the new theory of plate tectonics ignited an excitement about Earth science and stimulated a new generation of enthusiastic scientists to rethink how the Earth was formed. The Arab Oil embargo in the 1970s led to the initiation and funding of a sustained period of continental-margin exploration aimed at understanding energy resources in US domestic frontier environments, such as the unexplored continental margins. Finally, the adoption of the United Nations Convention on the Law of the Sea in 1982 provided the legal basis for nations to claim up to 200 nautical miles (or more) for their exclusive economic use. Together, these three events made exploration of the vast

submerged portion of the North American Atlantic Continental Margin possible. This article summarizes the diverse geology of this passive margin discovered during the last 40 years of scientific research.

Morphology of the Margin

The bathymetry of the North American Atlantic Continental Margin (Figure 1) shows a continental shelf that dips gently seawards, an abrupt shelf–slope break at a depth of about 100–200 m, a steep and sometimes dissected continental slope, and, finally, a gently sloping rise that merges with the abyssal plains of the deep ocean. The Blake Plateau is an exception to this morphology. It is a submerged platform at a depth of about 800–1200 m, which ends abruptly in a near-vertical escarpment that descends to more than 5000 m. The morphology at the north and south ends of the margin reflect modification by climatic zonation. For example, between New England and Newfoundland, the margin was shaped by Pleistocene glaciation (*see* **Sedimentary Processes:** Glaciers): banks, channels, ridges, knolls, rocky outcrops, and troughs are observed. In contrast, off the coast of Florida, subtropical reef growth, carbonate deposition, submarine carbonate lithification (*see* **Minerals:** Carbonates), and carbonate solution in deep water have led to oversteepened slopes.

Superimposed on the broad climatic effects are features resulting from oceanographic, geological, and tectonic modifications. The oceanographic processes that modify the seafloor are primarily tides, waves, and currents (for example, creating mobile sand bodies atop Georges Bank and around Sable Island). In deeper water, the Blake Outer Ridge (Figure 1) was formed by a huge sediment drift deposited at the confluence of southwards-flowing and northwards-flowing deep currents.

Perhaps the most striking features of the North American Atlantic Continental Margin are the many submarine canyons that incise the continental slope and rise between Cape Hatteras and Newfoundland.

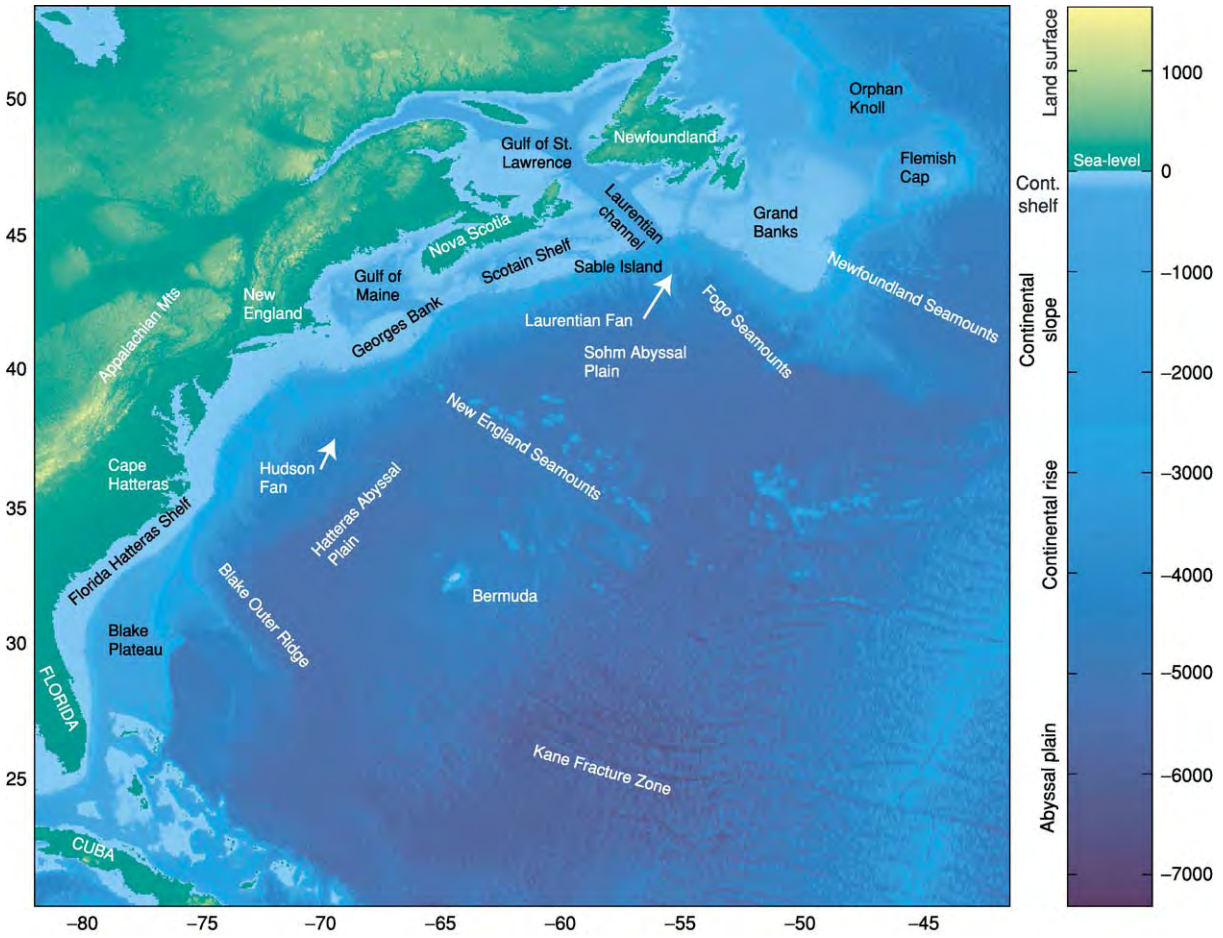


Figure 1 Bathymetric map of the North American Atlantic Continental Margin. Bathymetry was generated using ETOPO2 values with shaded relief. (ETOPO2 provides $2' \times 2'$ depth and elevation values, from NOAA/National Geophysical Data Center.)

More than 200 canyons can be identified on the upper slope off the US margin. Many of these canyons also indent the shelf edge. The largest canyon system, the Hudson Shelf Valley and Canyon, can be traced back to the mouth of the Hudson River (Figure 2). Abundant slumps, debris-flow deposits, and other evidence of mass movement are seen on the sides of many canyons. Fewer canyons exist on the lower slope because the canyons coalesce there. The lower slope and rise also contain large submarine fan deposits (e.g. the Laurentian and Hudson Fans off Newfoundland and New York, respectively).

Deeper-water geological processes are reflected in the presence of three seamount (*see* **Seamounts**) chains that impinge upon the margin: the New England Seamounts off Georges Bank, the Fogo Seamounts south of the Grand Banks, and the Newfoundland Seamounts off Newfoundland. A rifted continental fragment forms the submerged circular platform of Flemish Cap.

Several large landslides (*see* **Sedimentary Processes: Landslides**) also modify the margin morphology. The undated Cape Fear slide (Figure 3), south of Cape Hatteras, extends downslope for more than

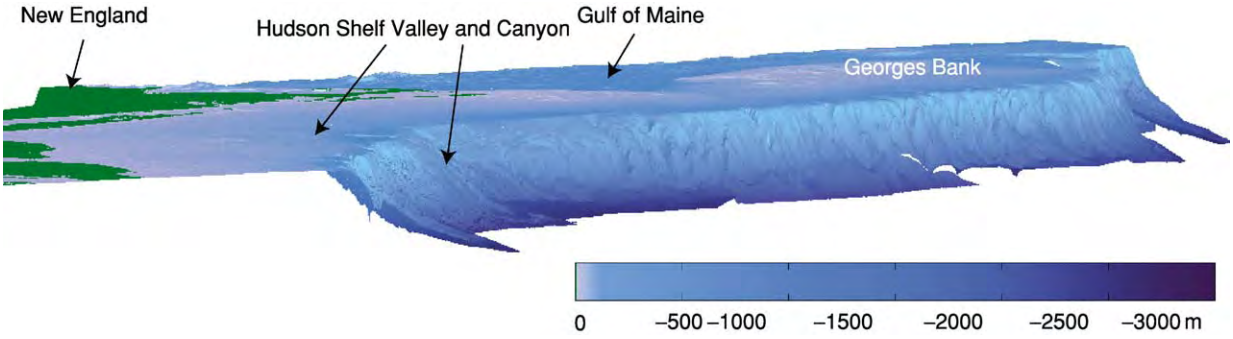


Figure 2 Perspective view of the canyons on the continental margin using bathymetric values from the NOAA coastal relief model (90 m × 90 m grid).

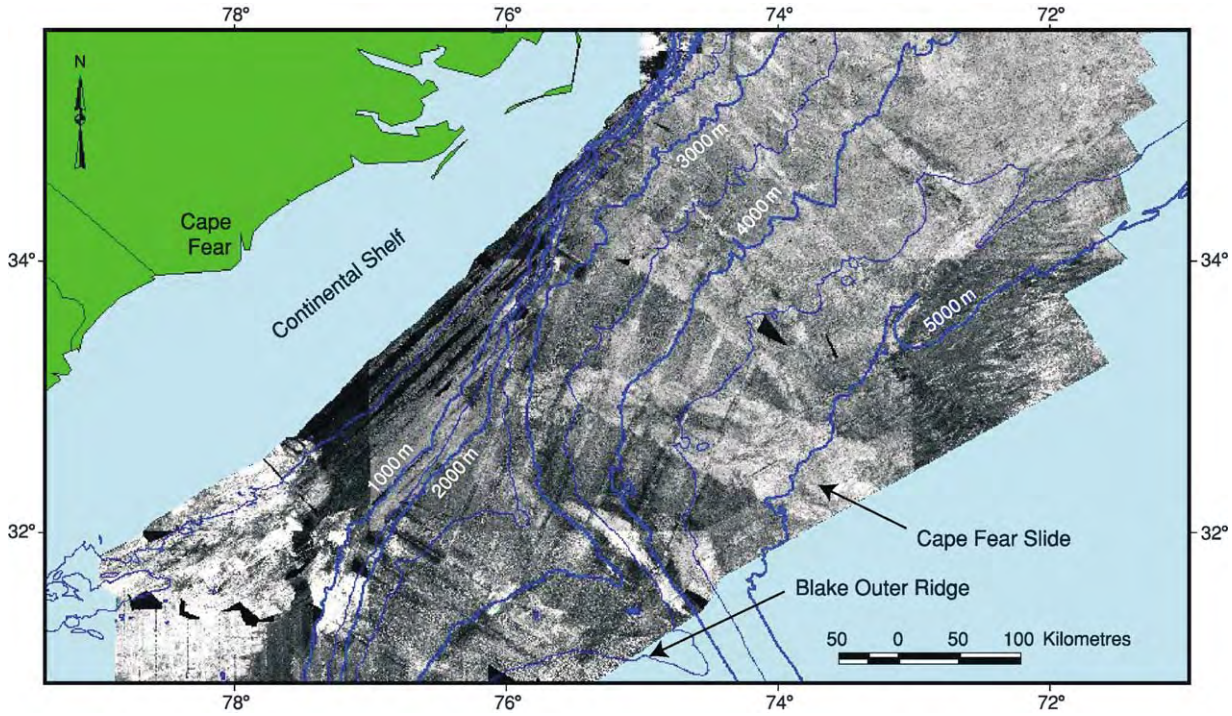


Figure 3 GLORIA side scan sonar image of the continental slope along a segment of the south eastern US margin, showing the Cape Fear slide, an undated massive submarine landslide south east of Cape Hatteras, NC. Dark grey areas are regions of low backscatter signal, commonly from fine grained deposits; light grey areas are regions of high backscatter signal, commonly from coarse grained deposits. The Cape Fear slide appears as a light grey region.

300 km between about 1200 m and 5500 m water depth. Off Newfoundland, the slides, turbidites, and other mass-flow bedforms resulting from the 1929 earthquake on the south-eastern Grand Banks can be identified in water depths from about 500 m to 3000 m.

Tectonic Evolution

The North American Atlantic Continental Margin is the product of continental rifting and seafloor spreading (*see* **Plate Tectonics**), which began in the Mesozoic between North America, Africa, and, eventually, Europe. Rifting is the phase of stretching and fracturing of continental crust that culminates in seafloor spreading (drifting), which creates new oceanic crust as the continental blocks move apart. These tectonic processes are directly responsible for the modern configuration of basins, platforms, magnetic anomalies, and oceanic fracture zones along the margin (Figure 4).

Early Triassic rifting began in an extensive interior mountain range in the supercontinent of Pangaea (*see* **Pangaea**) (Figure 5A). A series of rift basins formed between what are now the Gulf of Mexico and the Bay of Biscay. Rifting intensified in the Late Triassic (Norian), and true seafloor spreading progressed from the south at about 200 Ma to the north at about 185 Ma in what is now maritime Canada. Spreading near Newfoundland initially did not continue north of the Gibraltar transform zone at the northern edge of the African block. The Grand Banks were located north of the Gibraltar transform and, although extended during this time, behaved as a unit with Gibraltar and the European Plate.

Rifting and drifting along the entire margin are inferred to have overlapped during the Early Jurassic. A major basaltic extrusion event is recorded in the rift basins at about 200 Ma (Early Jurassic). Basalt flows are exposed throughout the Newark Supergroup. Numerous feeder dykes can be mapped in the exposed rocks surrounding the rifts (e.g. Shelburne Dyke in Nova Scotia, the Great Dyke in the Carolinas). These dykes probably also exist beneath the submerged continental margin, but are difficult to map beneath large thicknesses of postrift sediments. Rifting ceased during the Middle Jurassic.

The East Coast Magnetic Anomaly (ECMA; Figure 4), marks the western edge of the oceanic crust. The oldest oceanic crust near the ECMA was created during the Jurassic magnetic quiet zone and, therefore, is does not contain datable seafloor magnetic anomalies. The oldest seafloor-spreading isochron that can be identified on both sides of the Atlantic is M25 (about 156 Ma, Late Jurassic).

During the Middle and Late Jurassic, the continents dispersed and readjusted to tectonic events that were occurring in other parts of the fragmenting remnants of the Pangaean supercontinent. The initially shallow and restricted ocean basins deepened and widened as North America and Africa separated. In the Late Jurassic, a major plate (*see* **Plate Tectonics**) reorganization between North America and Europe rejuvenated synrift basins that initially formed in the Triassic on the Grand Banks of Canada (e.g. Jeanne d'Arc Basin). This rifting ended when seafloor spreading began between the Grand Banks and Iberia during the early Barremian (126 Ma, Early Cretaceous), initiating the final separation of Europe from North America (Figure 5B). Another plate reorganization in the Early Cretaceous (middle Aptian, about 110 Ma) completed the separation of North America from Europe when the northern part of Flemish Cap (and Orphan Knoll) separated from the Rockall region of the north-west European margin as seafloor spreading began in the Labrador Sea.

Alkaline basaltic magmatism, not obviously associated with breakup, created the New England Seamounts, (*see* **Seamounts**) which extend from the continental slope (Bear Seamount) to the foothills of the Mid-Atlantic Ridge. The age of the seamount chain has been difficult to establish, but it is now generally felt to have formed episodically between the Early Jurassic (125 Ma) and the middle Cretaceous (90–82 Ma). A younger, middle Cretaceous age for some of the White Mountain magma series in New Hampshire indicates that this younger magmatic pulse may have extended inland.

The Cenozoic has been a period of stable plate configurations, in which the Atlantic Ocean has widened and deepened. There are two exceptions to this otherwise stable tectonic regime. First, the Labrador spreading centre was abandoned and the Norwegian and Arctic Oceans began opening in the Late Paleocene. This gave rise to the mid-ocean ridge configuration that persists today between Newfoundland and Florida (Figure 5C). Second, at least two large bolides hit the margin (*see* **Impact Structures**). One excavated a crater 85 km wide beneath Chesapeake Bay in the late Eocene (35 Ma; Figure 6). A second smaller impact structure, 45 km wide, called the Montagnais structure, was formed at about 51 Ma near the shelf–slope break off Nova Scotia.

Passive Margin Structure

During the last 200 Ma, three major tectonic processes have affected the structure of the North American Atlantic Continental Margin: rifting, which broke apart Pangaea, thinned the continental crust by

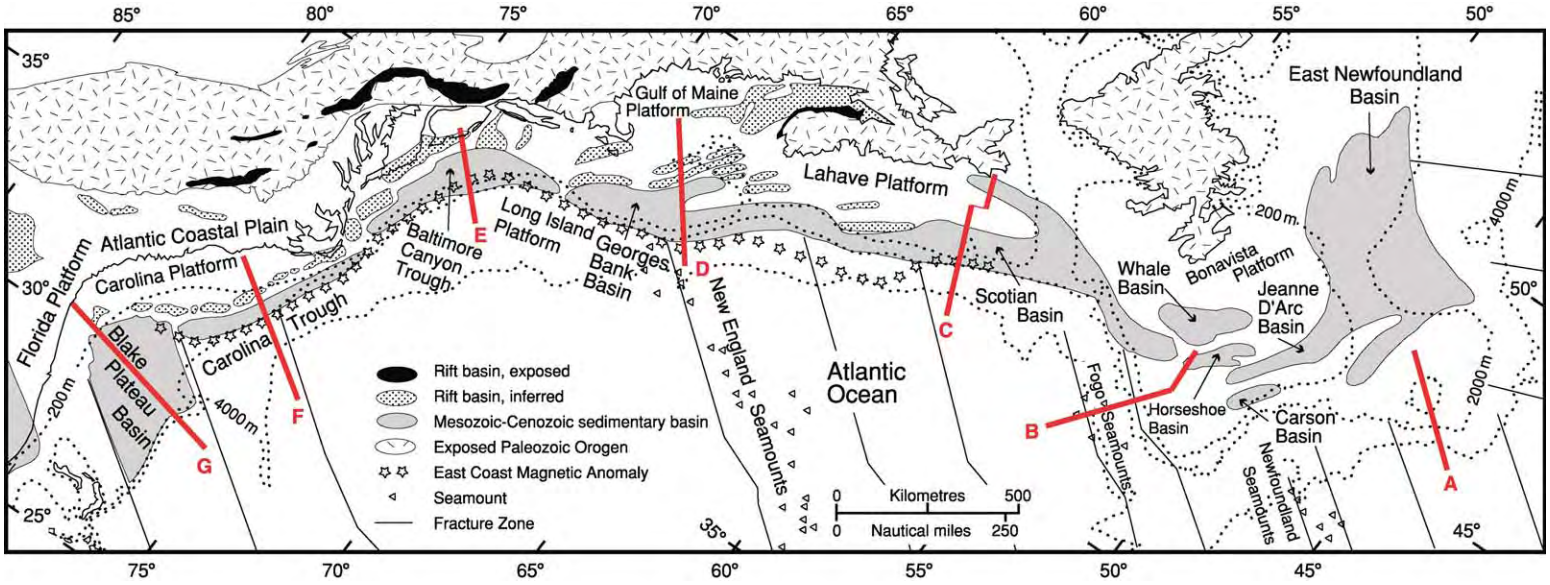


Figure 4 Tectonic map of the North American Atlantic Continental Margin showing the major postrift basins, platforms, and rift basins associated with Mesozoic breakup. The Newark Supergroup consists of the basins labelled as 'Rift basin, exposed.' The red lines show the locations of the crustal cross sections in **Figure 8**. The East Coast Magnetic Anomaly marks the landwards extent of the oceanic crust.

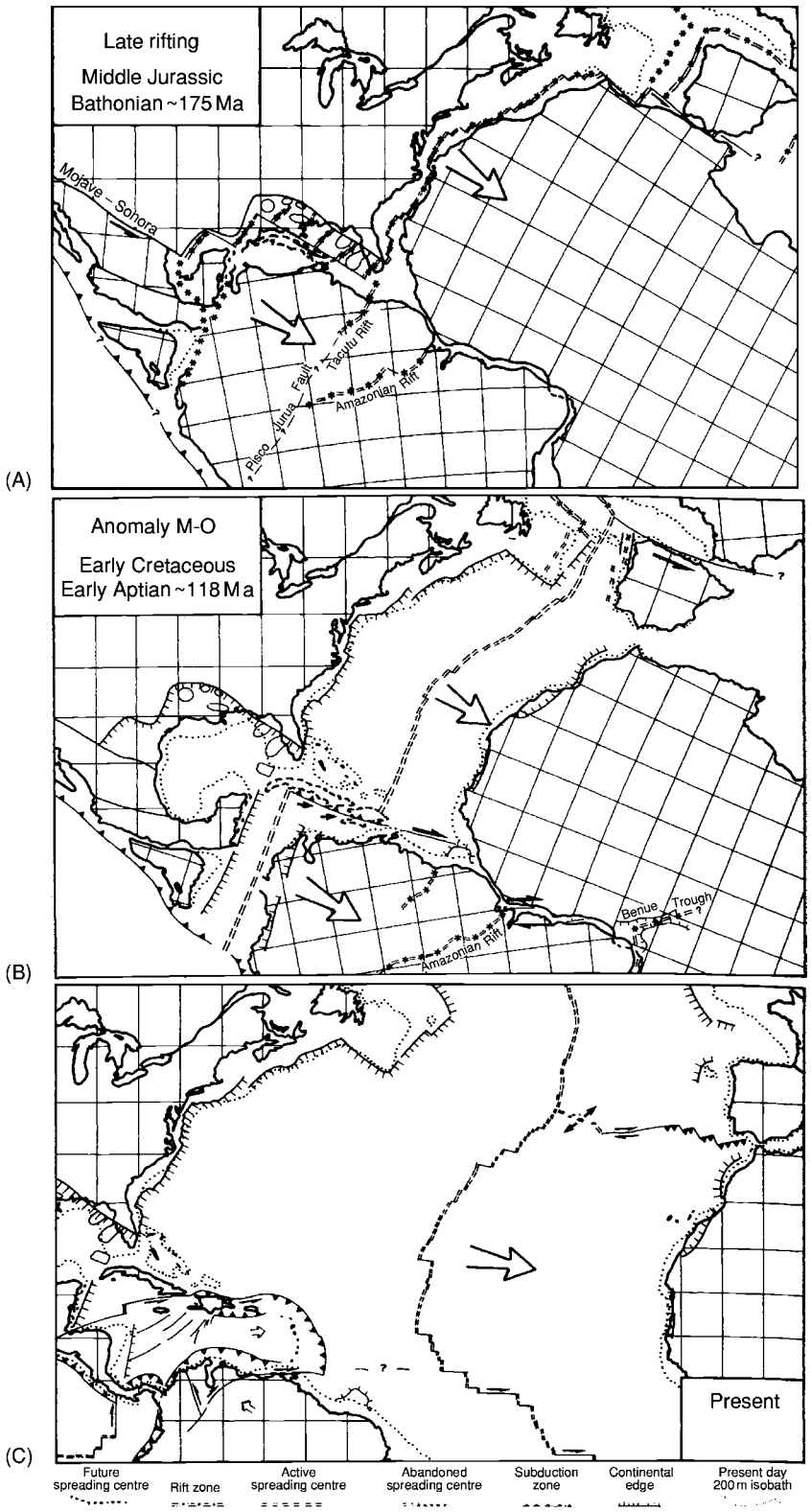


Figure 5 Three plate tectonic configurations relative to North America. (A) Rifting (Middle Jurassic) and initiation of seafloor spreading when Iberia and Newfoundland were part of a single plate. (B) Midway through the creation of the Atlantic Ocean (Early Cretaceous), shortly after Iberia and Newfoundland separated. (C) The current plate separations. Large arrows indicate the plate separation directions relative to North America. (Modified from Klitgord KD and Schouten H (1986) Plate kinematics of the central Atlantic. In: Vogt PR and Tucholke BE (eds.) (1986) *The Geology of North America, Volume M: The Western North Atlantic Region*, pp. 351 378. Boulder, Colorado: Geological Society of America.)

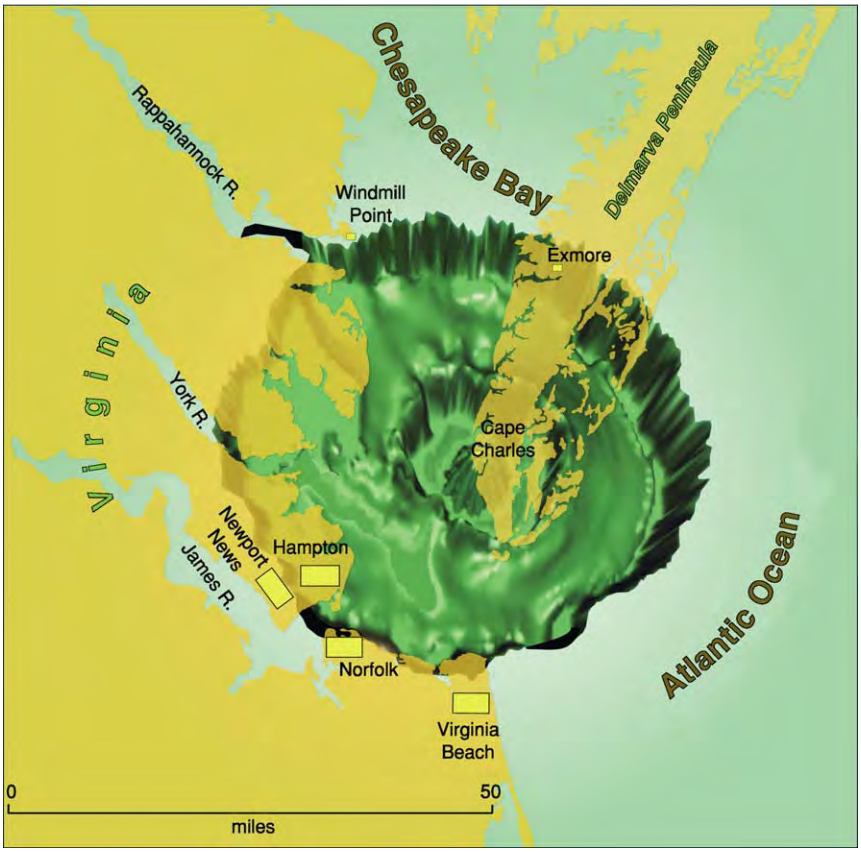


Figure 6 Three dimensional perspective of the Chesapeake Bay impact crater, which was formed at about 35 Ma on the North American Atlantic Continental Margin. The inner and outer crater walls are shown superimposed on a map of the Virginia part of Chesapeake Bay. (Reproduced from Poag CW, Koeberl C, and Reimold WU (2004) *The Chesapeake Bay Crater: Geology and Geophysics of a Late Eocene Submarine Impact Structure.* New York: Springer Verlag.)

variable amounts, and created numerous synrift basins; drifting, which created hot new oceanic crust initially adjacent to the rifted continent and now far away along the Mid-Atlantic Ridge; and passive subsidence, caused by cooling of the lithosphere and loading of the crust by sediments, which resulted in the development of large offshore sedimentary basins. Superimposed on these first-order lithospheric events are various oceanographic, magmatic, fluvial, eustatic, sedimentary, diagenetic, and biological processes that have also affected the final shape of the margin. The resulting structure of the margin is an alternating sequence of platforms, which have relatively little sedimentary cover, and basins, which contain enormous thicknesses of sediment (Figure 4).

Mesozoic rifting created the exposed basins of the Newark Supergroup (Figure 4). These fault-bounded basins and/or half-grabens form a linear chain along the exposed part of the margin. The basins contain deposits up to several kilometres thick. Similar synrift basins have been detected seismically and drilled near the coast and beneath the continental shelf. The best-known of these offshore basins is the petroliferous Jeanne d'Arc basin on the Grand Banks, in which the thickness of sediment exceeds 14 km. Synrift deposits are inferred to underlie the deepest portions of the margin, but have not yet been sampled. Synrift basins are also known on the conjugate north-west African continental margin.

A prominent margin-wide unconformity separates the synrift basins from the overlying postrift sedimentary units and is called the postrift or the breakup unconformity (Figure 7). Consequently, sedimentary-thickness maps of the margin often include only those units above the postrift unconformity (i.e. postrift deposits). The postrift unconformity may become a conformable surface within the deepest sediments in the centres of the postrift basins, but evidence for this is ambiguous, because seismic imaging in these deeper regions is often of poor quality and drilling data are lacking. Beneath the Grand Banks, the postrift unconformity is termed the Avalon unconformity.

Above the postrift unconformity, five major sedimentary basins have been documented. From north to south, they are the Scotian basin, Georges Bank basin,

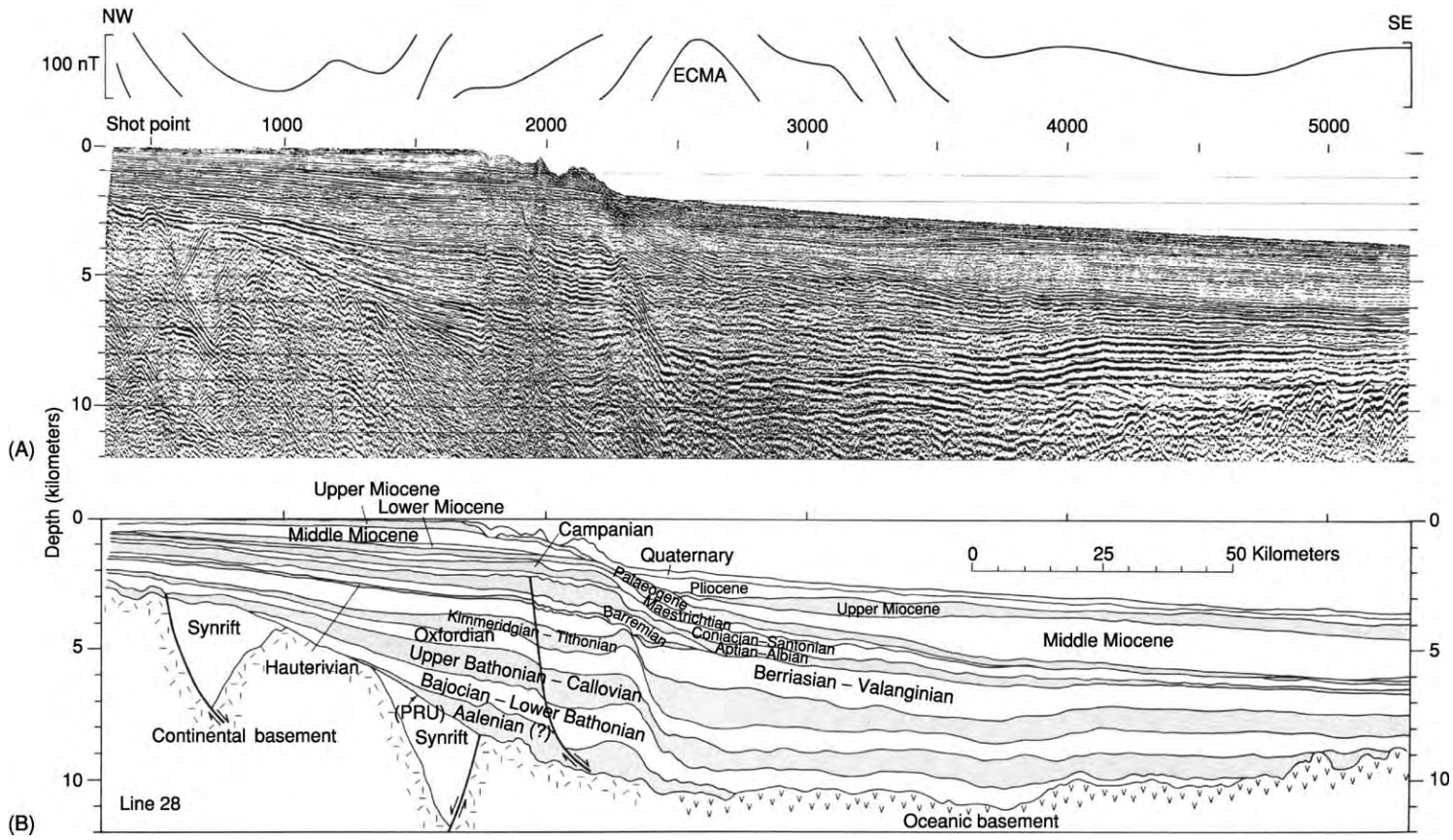


Figure 7 Stratigraphic cross section across the southern Baltimore Canyon trough. (A) Uninterpreted multichannel seismic reflection data. The magnetic anomaly is plotted across the top of the section and shows the location of the East Coast Magnetic Anomaly (ECMA). (B) Interpretation, giving ages and geometries of units. PRU, postrift unconformity. (Reproduced from Poag CW (1992) US Atlantic continental rise: provenance, dispersal, and deposition of Jurassic to Quaternary sediments. In: Poag CW and de Graciansky PC (eds.) *Geologic Evolution of Atlantic Continental Rises*, pp. 100 156. New York: Van Nostrand Reinhold.)

Baltimore Canyon trough, Carolina trough, and Blake Plateau basin (Figure 4). In plan view, they range in shape from nearly circular (Blake Plateau basin) to extremely elongated (Carolina trough). In cross section (Figure 7), the basin fill is roughly lens shaped, thinning towards both the continent and the ocean. Maximum sediment thickness also varies, from more than 15 km in the Baltimore Canyon trough and Scotian basin to less than 9 km in the Georges Bank basin. The composition of the basin fill is variable, and the sediment is often distributed among several sub-depocentres that together make up a basin complex.

The basins are bounded on their sides by structural highs, which often coincide with major faults. These faults both segmented the margin during rifting and coincided with (and perhaps localized) fracture zones in the newly forming oceanic crust. Most of the fracture zones can be mapped continuously from magnetic-anomaly offsets to the present Mid-Atlantic Ridge and are used as flow-lines of plate movement in plate-tectonic and palaeoenvironmental reconstructions of the ocean basin. Fracture-zone faults near the margin are mostly deeply buried, but they can be identified in the seafloor morphology near the mid-ocean ridge (e.g. Kane Fracture Zone; Figure 1).

Landwards of the large postrift basins are structural platforms. These are, from north to south, the Bonavista, LeHave, Long Island, Carolina, and Florida platforms. These regions may contain locally thick synrift deposits (e.g. the Jeanne d'Arc basin) but generally consist of prerift crystalline basement at shallow depths. Postrift deposits are thin (less than 4 km) and represent the onlapping edges of much thicker offshore units. Coastal-plain deposits are the exposed portions of these thinner younger Mesozoic and Cenozoic units. A hinge zone frequently marks the transition from a platform to a basin, i.e. where sediment thickness changes abruptly. These platforms are also affected by eustatic changes; hence, many parts of the sedimentary section are missing. The platforms are underlain by crystalline rocks from older Palaeozoic orogenies that built the Appalachian, Caledonian–Variscan, and Mauritanian mountains of North America, Europe, and Africa, respectively. The basin-bounding faults of the synrift basins frequently reactivate deeper basement faults and structures.

Variations in the amount of crustal stretching can be seen in cross sections through different parts of the North American Atlantic Continental Margin (Figure 8). Three crustal types are shown: oceanic crust, which is 5–6 km thick and similar to normal oceanic crust; transitional crust, which is generally less than 20–25 km thick, changes thickness rapidly, and has seismic velocities intermediate between those of oceanic and continental crust; and thinned continental crust, which is 30–35 km thick and may contain synrift basins but otherwise appears to be continental in seismic-velocity character (i.e. the platforms). ('Seismic' is used throughout this article to include coincident reflection and refraction techniques, unless otherwise noted.) A fourth type of crust, which is not shown on these cross sections, is normal continental crust, which would be 40–45 km thick, as found in continental interiors. The change from one crustal type to another may be abrupt or gradational and depends on the geological, tectonic, magmatic, and thermal history of the margin prior to and during breakup.

The transform margin of the southern Grand Banks basin (Figure 8) changes abruptly from continental crust to oceanic crust across a 40 km wide zone, whereas most of the basins have transitional crust that is 100–200 km wide (Scotian basin, Georges Bank basin, Baltimore Canyon trough, Carolina trough). The Blake Plateau basin is the most unusual, because the transition from continental crust to oceanic crust extends over nearly 400 km. Deep seismic data do not exist for the Blake Plateau basin, so the interpretation of crustal type is based primarily on poorly constrained gravity models. The basement and crustal structure in this southernmost basin are the least understood of any part of the margin.

The nature of the transitional crust also varies along the margin. For example, the southern part of the margin, towards Florida, is considered to be strongly volcanic, whereas the northern part, near Flemish Cap, is considered to be non-volcanic. The EDGE seismic experiment showed that a thick wedge of high-velocity presumably mafic material exists and coincides with rifting beneath the Carolina trough. Thus, this part of the margin is classified as a volcanic margin. Similar velocities have been interpreted from beneath the Baltimore Canyon trough. In contrast, a recent seismic experiment across Flemish Cap revealed that both serpentinized mantle and unusually thin oceanic crust lie adjacent to thinned continental crust. These anomalous features are interpreted as evidence of ultraslow seafloor spreading associated with a non-volcanic margin. The conjugate Iberian margin also has serpentinized rocks, which are thought to indicate non-volcanic processes. These have been drilled in the Ocean Drilling Program. The deeper structures of the Scotian and Georges Bank basins have not been well imaged using modern seismic techniques. Therefore, the nature of the deeper crust in these regions is uncertain, hindering our understanding of both the transition from continent to ocean and the transition from volcanic to non-volcanic parts of this margin.

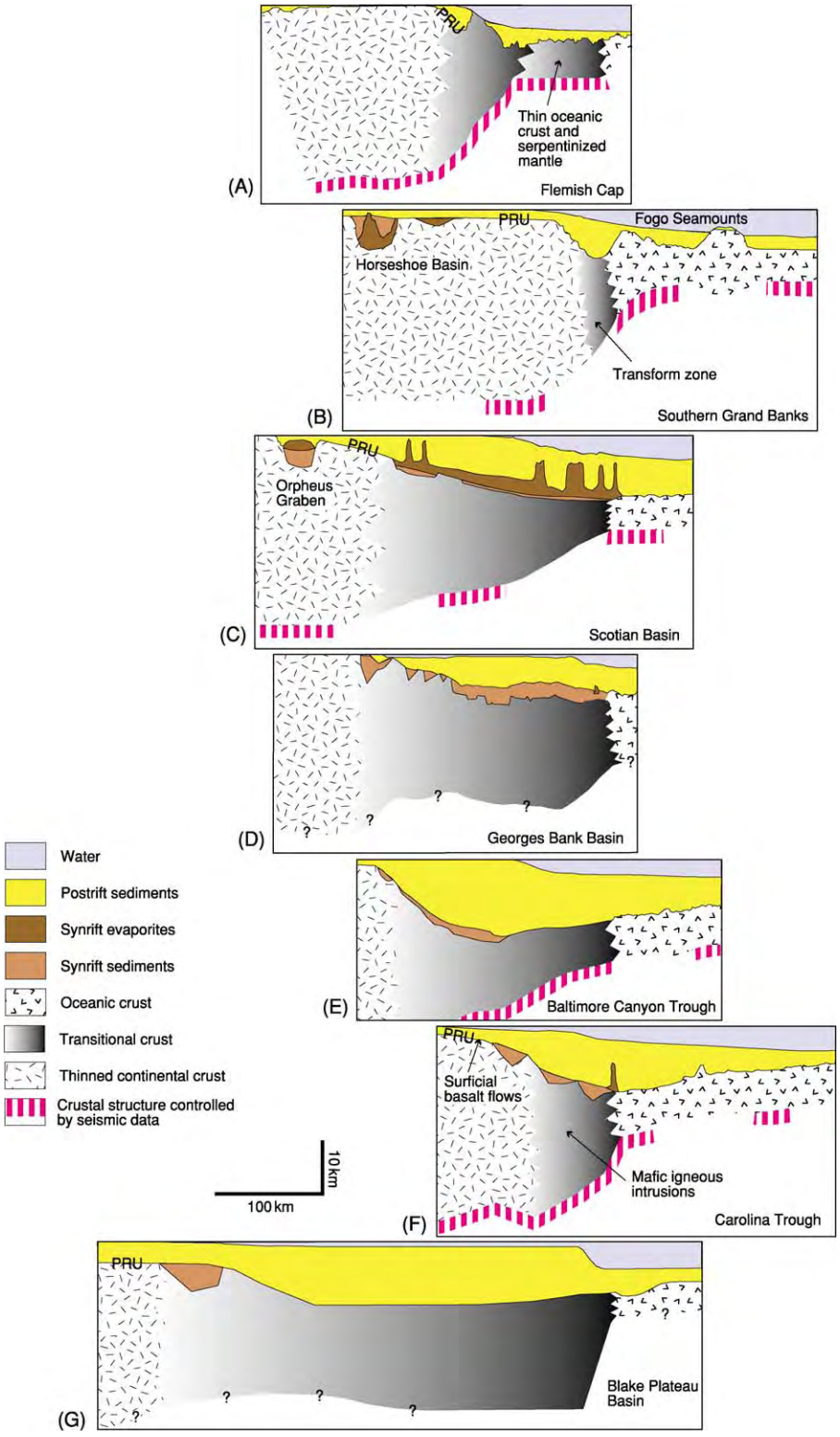


Figure 8 Simplified crustal cross sections through the major postrift basins of the North American Atlantic Continental Margin, aligned on the western edge of the oceanic crust. The transitional zone between continental crust and oceanic crust varies in thickness and width. PRU, postrift unconformity. (A) Flemish Cap, interpreted as a non volcanic transition (the two transitional crust patterns differentiate highly thinned continental crust on the landwards side from anomalous oceanic or serpentinized crust adjacent to the oceanic crust). (B) Southern Grand Banks, showing a narrow transition at the position of the fossil transform fault zone. (C) Scotian Basin. (D) Georges Bank Basin. (E) Baltimore Canyon Trough. (F) Carolina Trough, interpreted as a volcanic transition (the transition region is thought to contain magmatically emplaced crust). (G) Blake Plateau Basin.

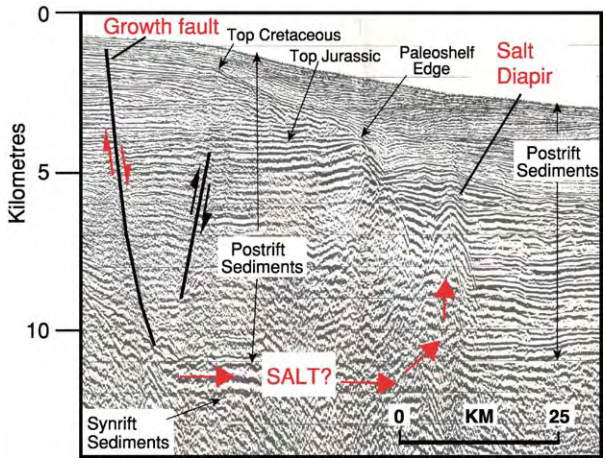


Figure 9 Seismic reflection profile and interpretation across the Carolina trough, showing the relationship between salt diapirism and faulting. Salt movement, shown by the arrows, is inferred to have caused growth faulting. PRU, postrift unconformity.

Numerous faults that offset postrift sediments are secondary structures on the continental margin. Beneath the Scotian basin and Carolina trough, faults are attributed to salt tectonism (Figure 9). Numerous small Cretaceous and younger north-east-trending reverse faults (e.g. the Block Island and New York Bight faults on the continental margin south of New England) are part of growing evidence for regional compression of the margin from Early Cretaceous through to at least Pleistocene times. The stress field east of the Rocky Mountains remains compressional today. Most of the faulting associated with the Chesapeake Bay bolide impact appears to be localized to the region surrounding the crater.

Sedimentary History and Palaeoenvironments

Interpretations of sedimentary history and palaeoenvironments are based on a combination of seismostratigraphic interpretation and direct samples from dredges, cores, drill cuttings, and submersible observations.

Triassic rifting began in central Pangaea (*see* **Pangaea**), where average elevations were probably more than 1 km above Mesozoic sea-level. Continental clastic redbed sequences filled large lakes that occupied subsiding rift basins, not unlike the large lakes of the modern East African Rift valley. The region from Florida to the Grand Banks spanned a latitudinal range from about 5° S to 20° N and encompassed equatorial rainforests to tropical savannahs. Evaporites (*see* **Sedimentary Rocks:** Evaporites) accumulated in shallower settings. Rift topography, and especially uplifted rift shoulders, exerted a local control on climatic processes (*see* **Tectonics:** Rift Valleys).

By the Late Triassic, shallow-marine conditions had developed around the site of the future continental margin. The initial marine incursion began in the north, when hypersaline waters transgressed the Grand Banks region, probably from across the Newfoundland–Gibraltar fracture zone. A 2000 m thick section of halite in the Osprey evaporites was deposited above redbeds of the Carson sub-basin. More extensive flooding proceeded southwards, and additional evaporites accumulated along the Nova Scotian margin (in a region now called the slope diapiric province). The youngest salt was deposited in the Carolina trough at the southern end of the future continental margin. The marine flooding marked a time when the sedimentary basins along the future margin became interconnected. Isolated rift basins still existed, however, on the distal edges of the zone of extension (i.e. exposed basins of the Newark Supergroup).

Evaporite deposition ended in the Early Jurassic when normal marine conditions became established. The ocean basin may have been as much as 350 km wide, 900 km long, and less than 1 km deep, although seafloor spreading did not become fully established until the Middle Jurassic. Carbonate platforms developed around the edges of the subsiding seaway, and calcareous mudstones were deposited in the deeper central portions. The developing North Atlantic Ocean was connected to the open ocean north of the Grand Banks (via Greenland and Europe) and to the developing Gulf of Mexico to the south. Uplift and basaltic volcanism interrupted and contributed to the Middle Mesozoic sedimentary record.

Between the Late Triassic and Middle Jurassic, the North American continent drifted northwards by about 10–15°, which created more arid conditions in the basins of the Grand Banks, while maintaining humid conditions in the south. The transition to more grey siliciclastic sequences during this interval indicates more humid palaeoenvironments.

The Middle Jurassic was a time of relative sealevel rise, which flooded the continental margins and started the transition from semi-closed to partially open circulation. Final plate separation occurred during the Middle Jurassic, and the ocean basin grew significantly wider (600 km) and deeper (2000 m, perhaps as much as 3000 m). All the large offshore sedimentary basins had formed by this time (Figure 10) and were accumulating enormous volumes of siliciclastics. This influx of siliceous material may have contributed to extensive radiolarian blooms. By the middle Middle Jurassic, local carbonate tracts were developing along the margins; these prograded rapidly seawards

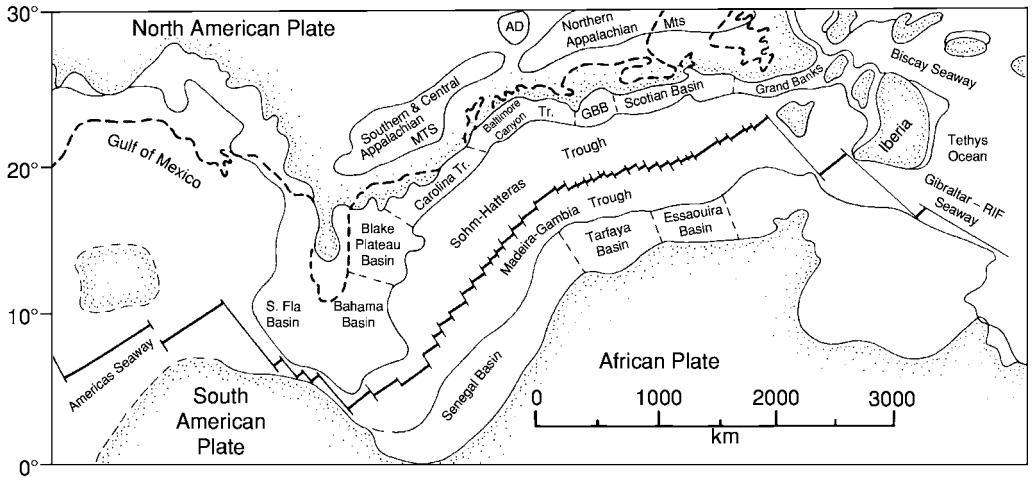


Figure 10 Structural, depositional, and physiographic features of the proto Atlantic seaway, reconstructed at the end of the Jurassic. Landmass boundaries and palaeolatitudes are approximate. GBB, Georges Bank Basin, AD, Adirondack Mountains. (Reproduced from Poag CW (1991) Rise and demise of the Bahama Grand Banks gigaplatform, northern margin of the Jurassic proto Atlantic seaway. *Marine Geology* 102: 63 130.)

in the latest Middle Jurassic and established carbonates as the dominant sediment type.

By the Late Jurassic, a giant carbonate platform extended for more than 5000 km from south of Florida to north of the Grand Banks (Figure 11). The Atlantic Ocean had opened to about 1400 km wide, and maximum seafloor depths reached nearly 4 km. The margin stretched from the equator to about 25° N. A very large carbonate bank formed across the entire Blake Plateau shelf. A more linear semicontinuous carbonate bank–reef system developed in the Baltimore Canyon trough. Farther north, in the Georges Bank and Scotian basins, the reef was more discontinuous. The morphology of the North American Atlantic Continental Margin at this time resembled that of the modern Great Barrier Reef of Australia, in which steep reef fronts mark the edge of the continental shelf and carbonate-rich aprons form on the continental rise seawards of the reef front. Gaps in the reef system channelled siliciclastics into large submarine fan complexes, which carried terrigenous material as far as 300–400 km from the shelf edge. A series of coalescing siliciclastic deltas also deposited terrigenous material shorewards of the carbonate platforms.

Except around the Blake Plateau, the giant carbonate platform ceased to exist in the Early Cretaceous because of rising sea-level, which stressed the carbonate-producing communities, and burial by siliciclastic sediments, which finally filled the back-reef basins and overflowed the bank–reef system. Overall, siliciclastic sedimentation slowed in the major basins, and the margins began to assume shapes similar to their present morphology. Globally, the Cretaceous was a period of poorly oxygenated marine waters,

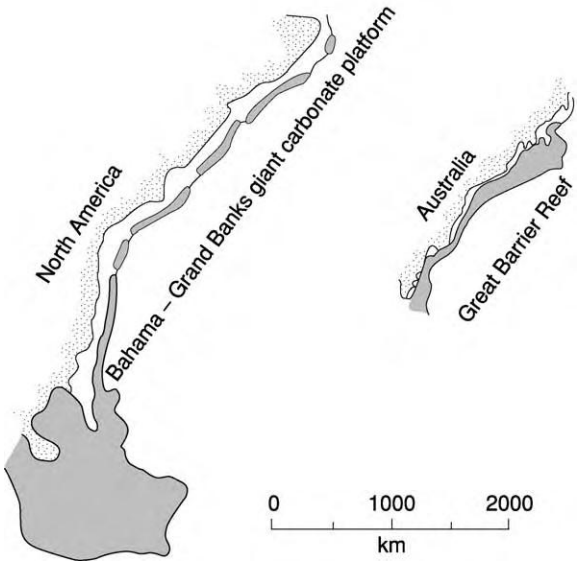


Figure 11 Carbonate tracts (stippled areas) associated with the Late Jurassic Bahama Grand Banks giant carbonate plat form along the North American Atlantic Continental Margin. The modern Great Barrier Reef system of Australia (rotated through 180°) is shown for comparison. (Reproduced from Poag CW (1991) Rise and demise of the Bahama Grand Banks gigaplat form, northern margin of the Jurassic proto Atlantic seaway. *Marine Geology* 102: 63 130.)

and studies of the dark carbon-rich Cretaceous deposits of the Hatteras Formation, on the rise and abyssal plain off the coast of North America, have contributed substantially to understanding black-shale sequences. The Cretaceous was also a time when the carbonate compensation depth cycled

between shallow (2000 m) and deep (more than 4000 m).

During the Cenozoic, continental-margin sedimentation was increasingly affected by currents, especially contour-parallel currents (*see* **Sedimentary Environments:** Contourites). Depocentres shifted seawards, causing the continental rise to be built up (Figure 12). The Paleocene and Eocene probably saw the onset of a current-dominated regime (e.g. the ancestral Gulf Stream) and the initiation of small canyons, but much of this early record is thin or missing because of subsequent erosion during the Oligocene. Global temperatures began falling as the Earth completed its transition from a 'greenhouse' in the Cretaceous and Paleocene to an 'icehouse' in the Oligocene. During the Middle Oligocene sea-level lowstand, the Baltimore Canyon shelf retreated by about 30 km. Cold water in the North Atlantic contributed to deep thermohaline circulation, and contour currents became more vigorous in the Early Miocene when cold Norwegian bottom waters overflowed the Norwegian–Greenland ridge and entered the Atlantic, initiating the formation of both the elongate Chesapeake and the Blake Outer Ridge drift deposits. Glaciation of North America in the Late Tertiary increased siliciclastic deposition, intensified down-slope sedimentary processes (e.g. turbidity currents, slumps, slides, and debris flows), and

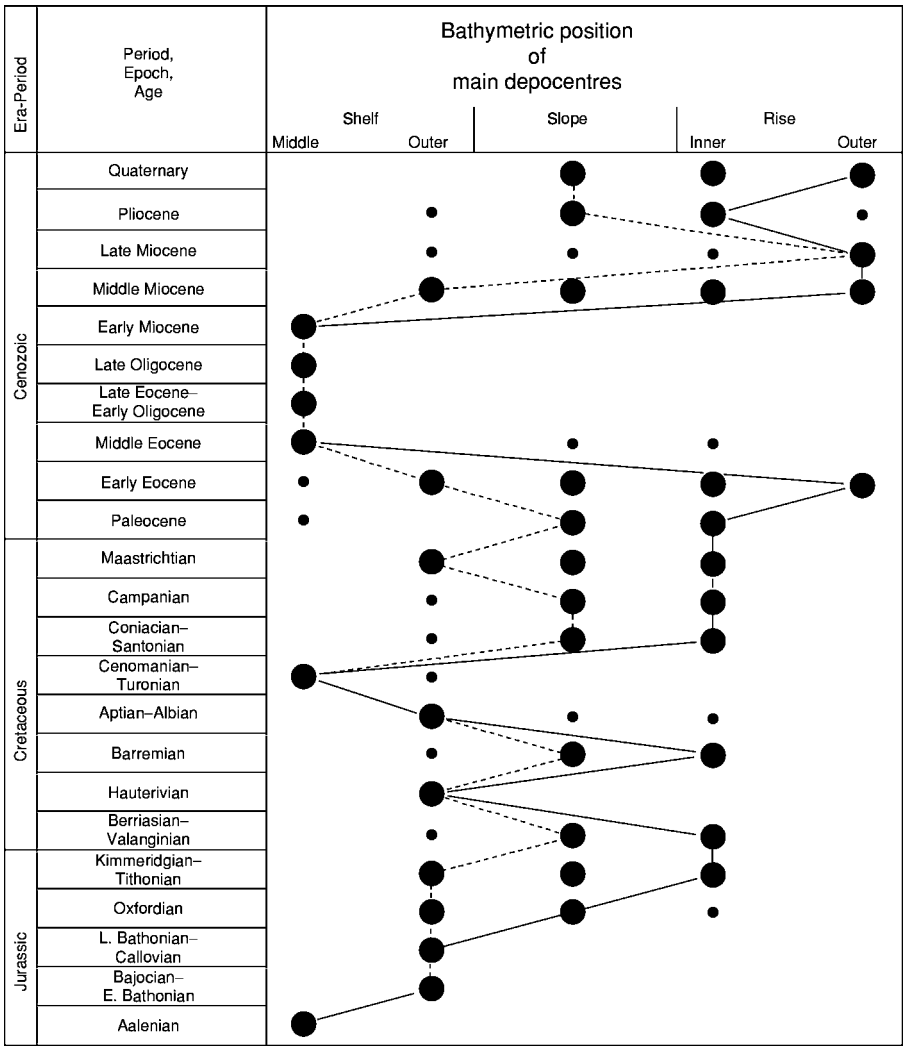


Figure 12 The bathymetric positions of the main depocentres in the Baltimore Canyon trough through the last 187 Ma. Large dots indicate major depocentres (i.e. thickest sediments). Small dots indicate secondary depocentres. Solid line shows the shift in the primary deep water depocentres through time. Dashed line shows the shift in the primary shallow water depocentres. This chart shows the migration of the trough seawards and landwards during the history of the margin. Note the marked period of limited shelf and slope deposition in the Late Miocene. (Reproduced from Poag CW and Sevon WD (1989) A record of Appalachian denudation in postrift Mesozoic and Cenozoic sedimentary deposits of the US Middle Atlantic Continental Margin. *Geomorphology* 2: 119 157.)

shifted sedimentary depocentres to deeper water on the rise and abyssal plain. Many new canyons also developed.

Energy, Mineral, and Water Resources

The knowledge base about energy resources on the North American continental margin varies by country: drilling off the US East Coast in the late 1970s and early 1980s produced disappointing shows for both oil and gas in the Baltimore Canyon trough and Georges Bank basin. Neither the Carolina trough nor the Blake Plateau basin have been drilled or tested for petroleum maturation, although models suggest that maturation in the Carolina trough should have reached the oil window only in its deepest Early and Middle Jurassic rocks. The petroleum potential in the rift basins buried beneath the margin remains unknown. Currently the USA has a moratorium on leasing and drilling on its offshore Atlantic margin.

In contrast, Canada has had an active exploration and production programme on the Scotian Shelf and Grand Banks since the late 1970s. The Mesozoic Jeanne d'Arc rift basin contains the most significant discoveries on the Grand Banks (Hibernia, Terra Nova, Hebron-Ben Nevis, and White Rose fields). Both oil and gas have also been produced from Upper Jurassic and Lower Cretaceous sandstones on the Scotian Shelf in the vicinity of Sable Island. In the 25 years since exploration began, a robust offshore technology and infrastructure has grown to support the Canadian petroleum industry.

One of the best-known deposits of an unconventional hydrocarbon, gas hydrate, exists on the Blake Ridge off the coast of the south-eastern USA (Figure 13). Gas hydrate is an ice-like solid that holds high concentrations of methane in a crystalline water matrix, which forms under pressures, temperatures, and geochemical conditions that are often found in sediments near the seafloor deeper than about 500 m water depth (*see* **Petroleum Geology:** Gas Hydrates). The first dedicated gas-hydrate drilling leg of the Ocean Drilling Program occurred on the Blake Ridge in 1995 (Leg 164). Estimates of the total volume of methane in both hydrate and associated free-gas deposits on the Blake Ridge vary from 60 to 100 trillion cubic metres.

Occurrences of offshore minerals are known along the North American continental margin, but they have not been as extensively developed or exploited as their onshore equivalents, primarily because recovery and transportation are more expensive in an offshore environment. Phosphates, used in agricultural fertilizers, are successfully mined in coastal-plain (i.e. postrift) deposits in Florida and North Carolina, but phosphate deposits on the continental shelf between Georgia and North Carolina are largely unexploited. Manganese nodules and ferromanganese crusts, potential sources of manganese, nickel, cobalt, copper, and platinum, have been sampled on the Newfoundland and New England Seamounts, but their distributions are largely unmapped. More detailed studies of similar nodules and crusts have been carried out on samples from the Blake Plateau.

Placer deposits also exist offshore, the most famous being gold-bearing sands and alluvium along the Nova Scotian inner shelf, which are presumed to have a source in pre-Mesozoic basement rocks. Mining for these intriguing finds has been largely uneconomical. Placer deposits containing titanium and iron have been mapped off Florida and are inferred to exist on other parts of the margin. The most plentiful mineral deposits are quartz sands, which are dredged mainly from the inner shelf for beach nourishment (for example, off the coasts of most of the states between New Jersey and Florida). Other resources that have been identified but not developed offshore include high-purity silica sand in the Gulf of St Lawrence and on the Grand Banks (which could be used for glass making), calcium carbonate sand (which is a key ingredient of Portland cement), clay deposits (which supply material for bricks, sewer pipes, and other construction materials), and coarse sand and gravel (which could be used in road building and maintenance). Quaternary glaciation (*see* **Sedimentary Processes:** Glaciers) and modern depositional processes have controlled much of the distribution of unconsolidated sands and gravels on the margin from New England to Newfoundland.

One offshore resource that has received little attention is fresh groundwater within sediments of the continental margin. Freshwater springs are known along the Florida shelf. One of the most spectacular is the Crescent Beach spring, about 3 km off the coast of north-east Florida, in about 20 m water depth. This spring produces a boil with a hydrogen sulphide odour on the sea surface during times of high discharge. Within the fractured and porous limestone (*see* **Sedimentary Rocks:** Limestones) aquifer system of the south-eastern USA, this offshore fresh groundwater is derived from rainfall on land, and its pressure is considered to be in equilibrium with sea-level. In contrast, shallow wells drilled on the continental shelf off New Jersey show a thin wedge of almost-fresh water that extends more than 100 km across the margin at relatively shallow depths (less than 100 m beneath the seafloor, Figure 14). This fossil freshwater accumulated during the Pleistocene falls in sea-level and has been protected from saltwater contamination

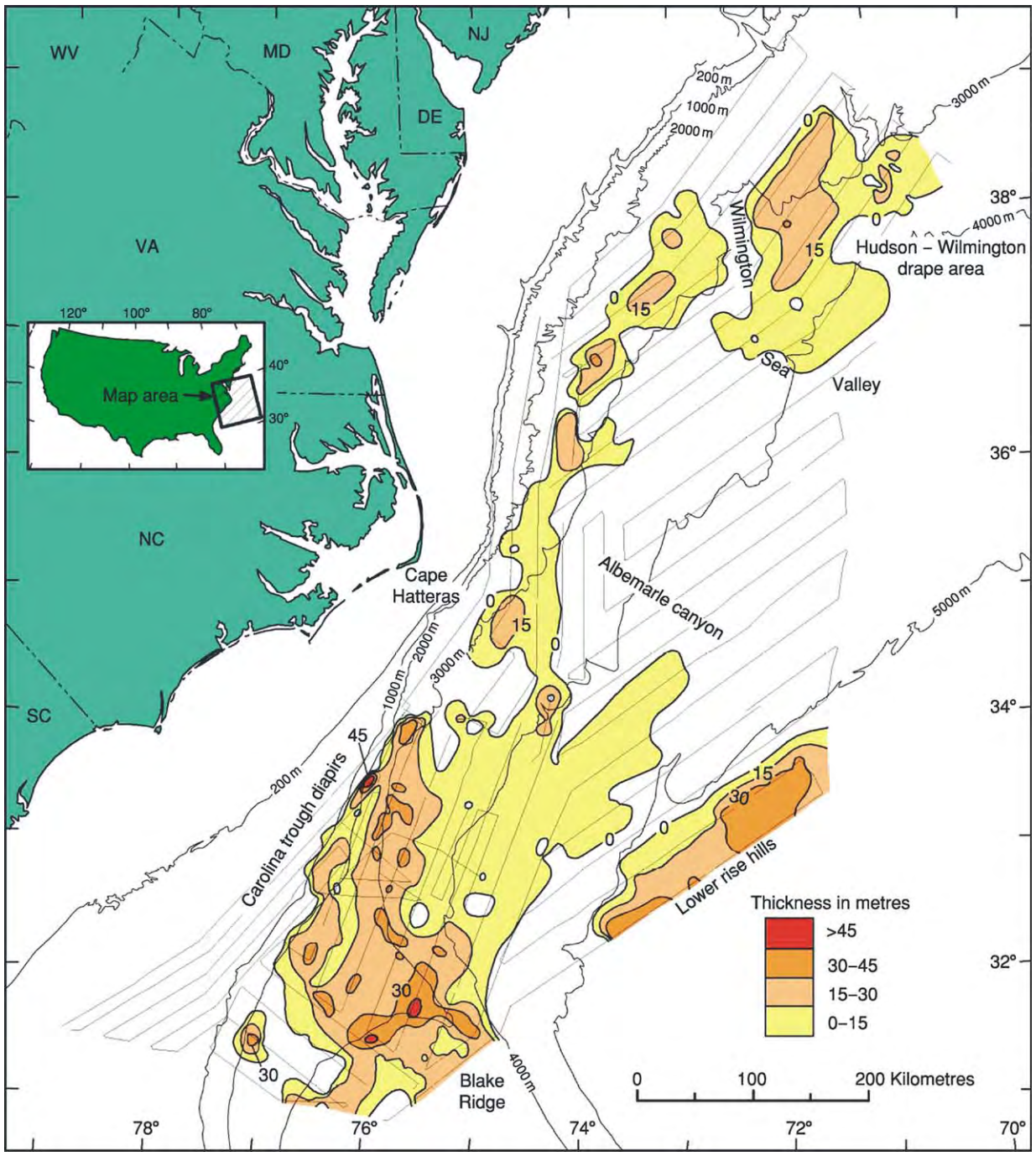


Figure 13 Distribution of gas hydrate inferred from seismic reflection data on the US Atlantic margin. Contours show the volume of pure gas hydrate in terms of thickness in metres. Volumes were estimated by mapping the extent of blanking in seismic profiles and then using a model to transform blanking to proportion of gas hydrate in the sediment. (Reproduced from Dillon WP, Fehlhaber K, Coleman DF, Lee MW, and Hutchinson DR (1995) *Maps Showing Gas Hydrate Distribution off the East Coast of the United States*. Miscellaneous Field Studies Map MF 2268, 1 : 1 000 000, 2 plates. US Geological Survey.)

by a low-permeability confining clay unit, which has prevented it from reaching equilibrium with the present saltwater hydrologic system. Freshwater is unknown in the continental-margin sediments off the coast of Canada.

Current and Future Societal Issues

As population growth has intensified in coastal regions, interactions between humans and nature have created an urgent need to understand surficial

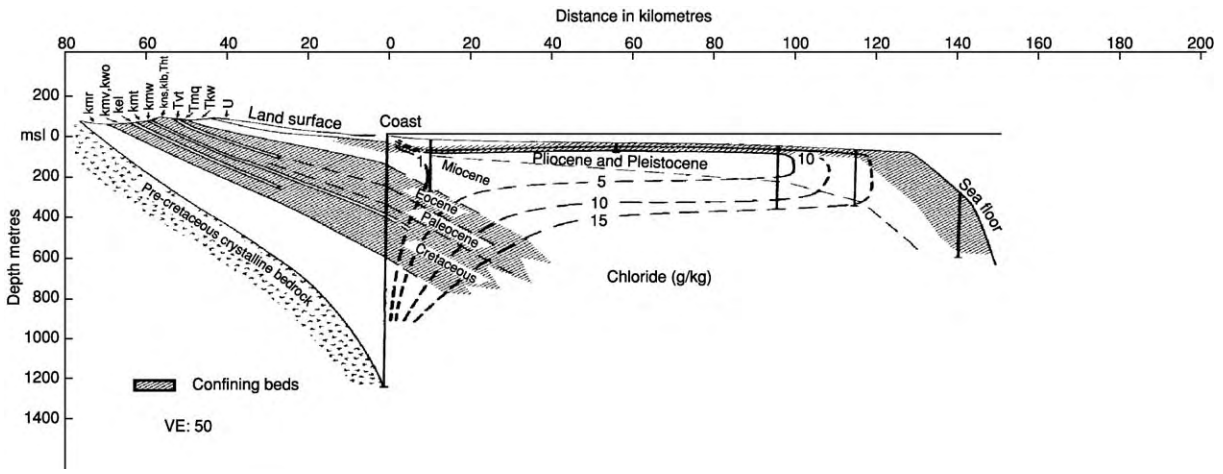


Figure 14 Schematic hydrological section through postrift sediments of the central Baltimore Canyon trough. Dashed contours show the chlorinity of the pore waters. The chlorinity of typical ocean water is about 19‰. (Modified from Kohout FA, Meisler H, Meyer FW, *et al.* (1988) Hydrogeology of the Atlantic continental margin. In: Sheridan RE and Grow JA (eds.) *The Geology of North America, Volume I 2: The Atlantic Continental Margin: U.S.*, pp. 463 480. Boulder, Colorado: Geological Society of America.)

geological processes that affect, for example, coastal erosion, sediment contamination, and habitat occurrence. Erosion from large storms results in loss of and damage to property and sometimes life. Beaches of the USA are among the most heavily artificially replenished of any country in the world: sand is often mined from the inner shelf. Three-dimensional bathymetric and geological mapping offshore has re-emphasized the importance of the geological framework of the shallow continental margin in understanding sediment transport, defining fragile habitats, and predicting the ultimate fate of industrial waste products in the oceans. Mapping techniques that integrate bathymetry, shallow subsurface profiles, video, and sampling have helped to determine which fishing grounds to open on Georges Bank and to manage other offshore fisheries. Many of these societal issues will increase in importance as population growth continues in coastal regions.

New technologies and high-speed computers are yielding new ways to view the continental margin and revealing new opportunities for fundamental scientific research. High-resolution bathymetric maps of the Hudson Shelf Valley reveal major bedforms and debris-flow deposits on the outer shelf, which appear to have been formed by a previously unrecognized enormous flood event. Sampling, dating, and interpretation of these deposits have not yet been achieved but are likely to modify interpretations of the late glacial and postglacial Quaternary history of the northern Atlantic region. Likewise, three-dimensional mapping of gas hydrates continues to challenge researchers to understand the importance of geology in controlling the occurrence of these highly dynamic potential resource reservoirs.

Recent ratification of the United Nations Convention on the Law of the Sea by Canada and the imminent ratification expected by the USA will create the need to identify geodetically accurate positions of the foot of the continental slope and to map accurately the sediment thickness in deep water. Such data can be used by nations to extend the legal boundary of their continental margin. These circumstances provide opportunities in which fundamental science can be integrated with practical needs.

Conclusions

The North American Atlantic Continental Margin is the type example of an Atlantic passive continental margin. For more than 200 million years, this margin has evolved from its initial equatorial rift configuration, accompanied by shallow oceanic flooding, to its present mature mid-latitude position bounding a deep ocean. The margin contains a rich record of diverse geological and tectonic processes and features, which include evaporite basins, extensive carbonate banks and reefs, and the modern sand-dominated partly glaciated sediment system. This margin provides sources of energy and minerals for human use and may hold a potential reserve of freshwater. The twentieth century was a time of great exploration, discovery, and characterization of the margin. At the start of the twenty-first century, new challenges confront researchers, in particular the need to understand how to balance human demands (e.g. for resources, stable coastlines, and waste disposal) with natural

processes (e.g. coastal erosion, habitat maintenance, and sediment and contaminant transport). Much of the continental margin remains an unexplored frontier area with regard to these new challenges.

See Also

Earth: Crust. **Impact Structures**. **Minerals:** Carbonates. **Pangaea**. **Petroleum Geology:** Gas Hydrates. **Plate Tectonics**. **Seamounts**. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Contourites. **Sedimentary Processes:** Glaciers; Landslides; Particle-Driven Subaqueous Gravity Processes. **Sedimentary Rocks:** Evaporites; Limestones. **Tectonics:** Rift Valleys.

Further Reading

Dennison JM and Ettensohn FR (1994) *Tectonic and Eustatic Controls on Sedimentary Cycles*. Concepts in Sedimentology and Paleontology volume 4. Society of Economic Paleontologists and Mineralogists. Tulsa: Oklahoma.

Ford D and Golonka J (2003) Phanerozoic paleogeo graphy, paleoenvironment, and lithofacies maps of the circum Atlantic margins. *Marine and Petroleum Geology* 20: 249 285.

Gardner JV, Field ME, and Twichell DC (eds.) (1996) *Geology of the United States' Seafloor: The View from GLORIA*. New York: Cambridge University Press.

Keen MJ and Williams GL (eds.) (1990) *The Geology of North America, Volume I 1: Geology of the Continental Margin of Eastern Canada*. Geological Society of America. Ottawa, Ontario, Canada.

Manspeizer W (ed.) (1988) *Triassic Jurassic Rifting: Con tinental Breakup and the Origin of the Atlantic Ocean and Passive Margins*. New York: Elsevier.

Poag CW (ed.) (1985) *Geologic Evolution of the United States Atlantic Margin*. New York: Van Nostrand Reinhold.

Poag CW (1991) Rise and demise of the Bahama Grand Banks gigaplatform, northern margin of the Jurassic proto Atlantic seaway. *Marine Geology* 102: 63 130.

Poag CW and de Graciansky PC (eds.) (1992) *Geologic Evolution of Atlantic Continental Rises*. New York: Van Nostrand Reinhold.

Scrutton RA, Stoker MS, Shimmield GB, and Tudhope AW (eds.) (1995) *The Tectonics, Sedimentation, and Palaeo ceanography of the North Atlantic Region*. Special Publication 90. London: Geological Society.

Sheridan RE and Grow JA (eds.) (1988) *The Geology of North America, Volume I 2: The Atlantic Continental Margin: US*. Boulder, Colorado: Geological Society of America.

Tankard AJ and Balkwill HR (eds.) (1989) *Extensional Tectonics and Stratigraphy of the North Atlantic Margins*. Memoir 46. Tulsa: American Association of Petroleum Geologists.

Vogt PR and Tucholke BE (eds.) (1986) *The Geology of North America, Volume M: The Western North Atlantic Region*. Boulder, Colorado: Geological Society of America.

Withjack MO, Schlische RW, and Olsen PO (1998) Di achronous rifting, drifting, and inversion on the passive margin of Central Eastern North America: an analog for other passive margins. *American Association of Petroleum Geologists Bulletin* 82: 817 835.

OCEANIA (INCLUDING FIJI, PNG AND SOLOMONS)

H Davies, University of Papua New Guinea, Port Moresby, Papua New Guinea
P Bani, Institut de la Recherche pour le Développement, Nouméa, New Caledonia
P Black and I Smith, Auckland University, Auckland, New Zealand
E Garaebiti, Department of Geology and Mines, Port Vila, Vanuatu
P Rodda, Mineral Resources Department, Suva, Fiji

Introduction

The islands of the south-west Pacific extend for 5000 km from New Guinea in the west to Fiji, Tonga, and Samoa in the east, and include the Solomon Islands, Vanuatu, and New Caledonia (Figure 1). The early European explorers called the area Melanesia because of the generally dark skin colour of the people. Subsequently, the parallel island chains have sometimes been referred to as the inner and outer Melanesian arcs. New Guinea was colonized from the west more than 40 000 years ago; New Britain, at least 35 000 years ago; New Ireland, at least 20 000 years ago; and the northern Solomon Islands, 28 000 years ago. The other Melanesian islands, New Caledonia, Vanuatu, and Fiji, were settled only 3000 years ago.

The islands have many similarities in their geology, in that, for the most part, they have been constructed on oceanic crust by one or more cycles of volcanic activity. Exceptions are the island of New Guinea, which is part of the Australian continent; the older rocks of New Caledonia, which first formed at the margin of the Australian continent; and the outermost Solomon Islands, which are elevated oceanic crust. There are deep-sea trenches on either side of the island chains. On the outer (north-eastern) side are the Manus, Kilinailau, Solomons, and Vitiaz trenches, which are only weakly active or are inactive, and the Tonga–Kermadec Trench east of Tonga (Figure 1), which is active. On the inner (south-western) side is an active system of trenches, the New Britain–Makira–Vanuatu–Hunter–Kadavu trenches. Other seafloor features include the massive Ontong Java submarine plateau of thickened oceanic crust (Figure 1), smaller oceanic plateaus of rifted continental lithosphere in the Coral Sea, and a number of small ocean basins, some of which are actively spreading.

The islands occupy the boundary zone between the north-moving Australian plate and the west–north-west-moving Pacific Plate, and for this reason are tectonically and volcanically active (Figure 2). The plate boundary is not a simple one but, rather comprises a number of smaller plates that exhibit a variety of plate interactions, from seafloor spreading to subduction, transform faulting, and rotation. The lines of earthquakes on the map (Figure 2) coincide with the plate boundaries, and give an idea of the high level of seismic activity in the region. The plot of depth-coded hypocentres shows the trace of a subducted slab dipping westward from the Tonga–Kermadec Trench, in the east, and of a subducted slab dipping northward from the New Britain Trench, in the north-west. The area has evolved since the Cretaceous by repeated episodes of crustal extension and crustal shortening. Extension produced small ocean basins, and shortening resulted in subduction and thrust faulting. The volcanic islands have grown through this time, initially (in the Eocene to Early Miocene) by volcanism related to subduction of the Pacific Plate on the outer trench system, and more recently (the past 10 million years) by volcanism related to subduction of the Australian Plate on the inner trench system.

Papua New Guinea

Papua New Guinea (PNG) comprises the eastern half of the island of New Guinea, the adjacent islands of the Bismarck Archipelago, and the northernmost of the Solomon Islands (Bougainville and Buka). The geology of PNG can be described in terms of four geological provinces (Figure 3): a stable platform in the south-west, where the Australian craton extends beneath the island of New Guinea; a collisional zone in the centre and north of the island, where terranes have been sutured to the Australian craton; a province built by volcanic arc activity in the north-east; and the Ontong Java submarine plateau.

Platform and Foldbelt

The stable platform comprises a continental basement of Palaeozoic and partly Precambrian metamorphic and igneous rocks overlain by a 4-km-thick paraconformable sequence of shelf facies Mesozoic siliciclastic, and Cenozoic carbonate and siliciclastic, sedimentary rocks (Figure 3). Rocks along the northern margin of the platform are buckled and faulted to form a basement-involved foreland fold-and-thrust belt. Northward of the fold-and-thrust belt, the Mesozoic sediments are transitional into a kilometres-thick sequence of black mudstones. These

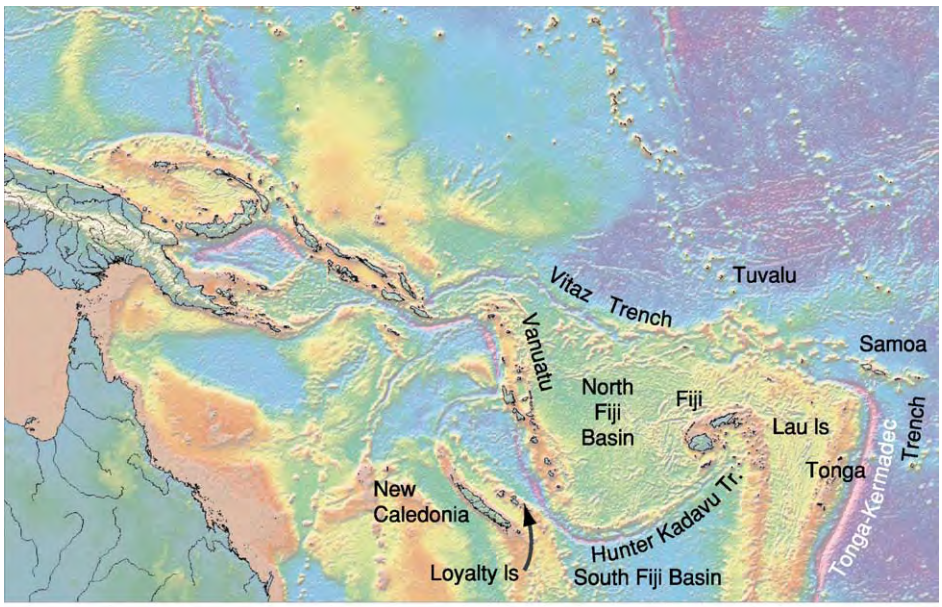


Figure 1 Physiographic map of the south west Pacific. Reproduced with permission from Smith WHF and Sandwell DT (1997) Global seafloor topography from satellite altimetry and ship depth soundings. *Science* 277: 1957 1962. Available on the Internet at http://topex.ucsd.edu/marine_topo/mar_topo.html.

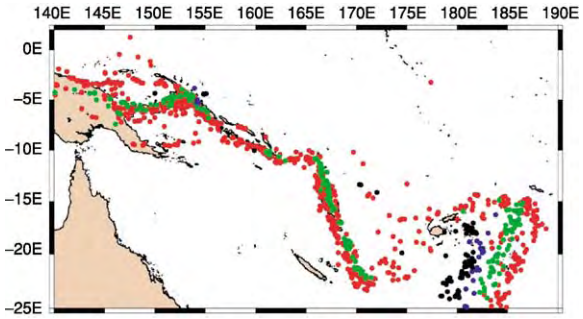


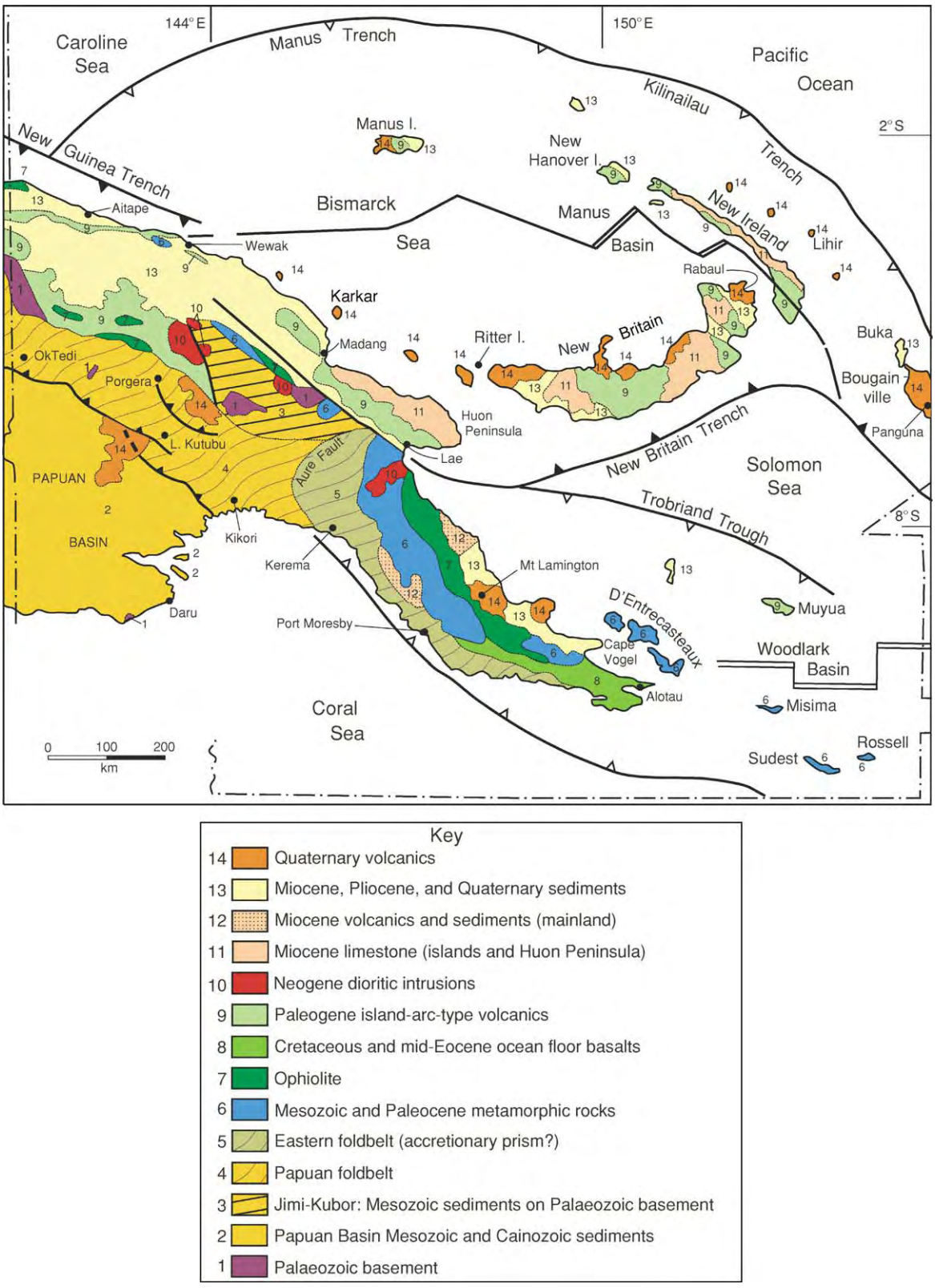
Figure 2 The map shows earthquakes with magnitude 6 or greater, recorded during 1963 2002. Red indicates hypocentre depth of 0 70 km, green 70 300 km, blue 300 500 km and black >500 km. Features of the map are the west dipping seismic zone beneath Tonga in the east, the north dipping seismic zone be neath New Britain in the north west, and the steeply dipping seismic zone beneath Vanuatu. Shallow seismic zones (red dots) mark the spreading ridges and transforms in the Bismarck and Solomon seas and the North Fiji Basin. Map prepared by Emile Okal.

in turn are transitional northwards into greenschist facies graphitic schists. The sediments of the stable platform and foldbelt comprise the Papuan Basin.

The foldbelt changes character east of the Aure Fault (Figure 3). Beyond this point, the foldbelt comprises fault-bounded segments of Late Cretaceous, Paleocene, and Eocene fine siliceous siliciclastic and calcareous sedimentary rocks, associated with faulted slivers of ultramafic rock; Oligocene and younger coarser, partly volcanogenic, dominantly clastic sediments; and locally extensive Oligocene gabbro. There is no indication of a continental or cratonic basement, though such may exist. The Paleocene and much of the Eocene sediments were deposited in a deep marine environment and probably have been added to the mainland by accretion above a north-dipping subduction system.

Collisional Zone

The western part of the central collisional zone comprises fault-bounded terranes of Jurassic to Eocene sedimentary, igneous, metamorphic, and ultramafic rock, and younger (Eocene to Miocene) dioritic intrusive bodies. The association of ultramafic rocks with eclogite, blueschists, submarine volcanic rocks, and volcanogenic sediments indicates a former volcanic arc environment. Sediments associated with the volcanic rocks are of two ages, mid-Eocene and Oligocene. This suggests a history of successive development of volcanic arcs and successive collisions; the Oligocene volcanic rocks are found only on the north side of the northern ranges. Also included in the western part of the central collisional zone are two fault-bounded blocks of Palaeozoic basement. One of these lies on the border with Indonesian Papua and the other extends northward from the Kubor Range, to underlie the Jimi valley (Figure 3). These Palaeozoic terranes lie outward (northward) of younger accreted terranes and thus appear to be 'out of sequence'.



Key

14 Quaternary volcanics
13 Miocene, Pliocene, and Quaternary sediments
12 Miocene volcanics and sediments (mainland)
11 Miocene limestone (islands and Huon Peninsula)
10 Neogene dioritic intrusions
9 Paleogene island-arc-type volcanics
8 Cretaceous and mid-Eocene ocean floor basalts
7 Ophiolite
6 Mesozoic and Paleocene metamorphic rocks
5 Eastern foldbelt (accretionary prism?)
4 Papuan foldbelt
3 Jimi-Kubor: Mesozoic sediments on Palaeozoic basement
2 Papuan Basin Mesozoic and Cainozoic sediments
1 Palaeozoic basement

Figure 3 Geological map of Papua New Guinea, showing active (solid triangles) and inactive or slow acting trenches (open triangles), transforms, and spreading ridges; major thrust faults on land (solid triangles) are also indicated. There are four geologic provinces: Platform in the south west, collisional in the centre, volcanic arcs in the north east, and Ontong Java Plateau. Map drawn by L Cotterall, Geology Department, Auckland University.

Possibly they are indications that the Palaeozoic craton extends northward as a substrate beneath the terranes of the collisional zone. Mesozoic amphibolite facies metamorphics near the north coast (Prince Alexander Range) and Oligocene metamorphics associated with ultramafic rocks just beyond the border on the north coast (near Jayapura) also are 'out of sequence' and may represent material ramped up from a subjacent older stratum.

The geology of the eastern part of the collisional zone, from Lae south-eastward, is much simpler and essentially comprises two rock units: an inner core of metamorphic rocks, flanked on the north-east side by a major ophiolite complex. The protolith of the metamorphics was felsic and psammitic sediments and some submarine basalts, partly of Cretaceous age, and these were metamorphosed in the Paleocene. The ophiolite comprises a 4- to 8-km thickness of tectonite harzburgite, overlain by lenses of cumulate-textured ultramafic rocks, which are in turn overlain by 4 km of gabbro and 4 km of basalt. Paleocene and Eocene volcanic rocks that overlie the ophiolite may be arc related. Ocean floor basalts that form the eastern end of the island are Late Cretaceous and Middle Eocene and may be related to the ophiolite. Miocene and younger, partly volcanogenic, dominantly siliciclastic sediments of the North New Guinea Basin unconformably overlie basement in the western part of the collision zone, and Miocene and younger volcanics and sediments of the Cape Vogel Basin unconformably overlie basement in the eastern part of the collision zone.

North-Eastern Province

The north-eastern province, constructed by volcanic arc activity, embraces the islands of the Bismarck Archipelago and the mainland mountain ranges of the Huon Peninsula. The oldest rocks are Eocene and Oligocene volcanics with some dioritic intrusions. These are partly overlain by Miocene limestone, which is in turn partly overlain by volcanic rocks of the present-day (Plio-Quaternary) volcanic cycle. Bougainville, in the Solomon Islands chain (see later), has similar geology. The land areas are separated by several small ocean basins: the Bismarck Sea, Solomon Sea, and Coral Sea (Figure 3). There are active spreading ridges and transforms in the eastern Bismarck Sea (Manus Basin) and southern Solomon Sea (Woodlark Basin) and active trenches on the inner side of New Britain and the Solomon Islands, on the outer side of the Solomon Islands and extending westward to Manus, and off the New Guinea coast west of Wewak. A trench on the south side of the Solomon Sea, at the Trobriand Trough, is masked by thick sediment. Volcanic activity is associated with the New Britain Trench but not with the trench west of Wewak (New Guinea Trench). The Coral Sea opened by seafloor spreading in the Paleocene (65–55 Ma), the Solomon Sea in the Oligocene (40–30 Ma), and the eastern Bismarck Sea and Woodlark Basin in the Plio-Quaternary (since 6 Ma).

Economic Minerals

The main mineral deposits of Papua New Guinea are porphyry copper–gold deposits, disseminated and vein gold deposits, and alluvial gold. Other less common deposits include lateritic nickel–cobalt ore over ultramafic rocks, minor manganese ore, and the small copper–gold massive sulphide orebodies that have been mined near Port Moresby. The Panguna mine (now closed) on Bougainville Island and the Ok Tedi mine in Western Province are classic porphyry copper–gold deposits. In both locations, copper and gold occur in Plio-Quaternary dioritic to granitic rocks and in the adjacent country rocks. Similar mineralization is known at Frieda River (West Sepik), Yandera (north-west of Goroka), and Wafi, 50 km south-west of Lae.

Disseminated and vein gold deposits are mined at Porgera; Ladolam on Lihir Island; Tolukuma, north of Port Moresby; and, until recently, Misima Island. Production is planned from Irumafimpa near Kainantu and from Hidden Valley, near Wau. Production from these mines, together with Ok Tedi, is sufficient to place Papua New Guinea in the top five of gold-producing nations, worldwide. The Panguna mine also was a major producer of gold. At Porgera, gold mineralization occurs in intrusive dioritic rocks and in the adjacent, baked sedimentary rocks. At Ladolam, the gold is in the floor of the caldera of an eroded volcano, with hot springs nearby. At Tolukuma, Irumafimpa, and Hidden Valley, the gold is in quartz veins. Alluvial gold is mined on a small scale in many parts of Papua New Guinea, notably near Amanab and Maprik in the Sepik, in the Western and Eastern Highlands, around Wau, and in Milne Bay Province.

From 1930 to 1965, with a break for the Second World War, alluvial gold was mined by dredges in the Bulolo valley. A total of about 65 tonnes of gold was produced. Within the last few years, a rich alluvial field was found at Mount Kare, south-west of Porgera. Where earlier gold rushes had been almost exclusively managed by overseas miners, Mount Kare was the first entirely Papua New Guinean gold rush. As many as 8000 miners were on the field at any one time and an estimated 15 tonnes of gold was taken out in 2 years. The only other fields to have produced more than a tonne of gold are the Lakekamu and Sepik alluvial fields, and Misima and Muyua (Woodlark), both of which were mixed alluvial and lode mining. Lateritic soils over ultramafic rocks south of

Madang contain economic concentrations of nickel and cobalt and await development.

Energy Resources

Oil currently is produced from several thrust-bounded anticlines containing the Kutubu, Gobe, and Moran oil fields, all of which are in the vicinity of Lake Kutubu in the Southern Highlands Province. Production from the South-east Mananda oil field within the giant Mananda structure between the Kutubu oil fields and the Hides gas field is planned. The oil was derived from Jurassic black shales and has accumulated in clean, shallow marine quartz sandstones of Late Jurassic and Early Cretaceous age. At the time of the initial discovery, in the Iagifu anticline, recoverable reserves were estimated at around 200 million barrels of oil. By the beginning of 2004, 354 million barrels had been produced from the various fields and remaining recoverable oil reserves stood at 222 million barrels. Production, which began in June 1992, has declined from a peak of almost 150 000 barrels day^{-1} in 1993 to around 50 000 barrels day^{-1} in 2004 as the fields have been depleted.

A small part of the large volume of gas in the Hides anticline, south-west of Tari, provides electrical power to the Porgera mine. The gas at Hides and in other structures in the foldbelt is sufficient to be developed commercially, provided that a market can be found. A pipeline to Queensland, Australia is planned, but the project awaits confirmed markets before detailed engineering and design work and construction can proceed. Large volumes of gas also have been discovered offshore in the Gulf of Papua at Pandora on the north-western corner of the Eastern Plateau, about 150 km south-west of Kerema. This gas occurs in a deeply buried cavernous Miocene limestone reef. Oil and gas seeps are known in the North New Guinea Basin and there are indications of petroleum in the Cape Vogel Basin. A pilot program to utilize geothermal energy at the Ladolam Mine on Lihir Island has proved successful and is being expanded. This is the first use of geothermal energy in PNG.

Natural Hazards

Papua New Guinea has 14 active and 22 dormant volcanoes and these pose a potential danger to 204 000 people. Ten of the fourteen active volcanoes and 12 of the 22 dormant volcanoes are in the Bismarck volcanic arc, from Bam, Manam, Karkar, and Long Island in the west to Langila, Garbuna, Makalia, Pago, Ulawun, and Rabaul in the east. Other groupings of active volcanoes are in the Admiralty Islands, on Bougainville Island, and in eastern Papua. The agency charged with monitoring the PNG volcanoes is the Rabaul Volcanological Observatory.

Rabaul volcano erupted in 1937, killing more than 500 people. When it erupted again in 1994, only five people were killed, but much of Rabaul town was devastated, leading to a decision to relocate government and most business activity to Kokopo. The worst volcanic disaster in historical time was the explosive eruption of Lamington volcano in Oro Province in 1951; 3000 people were killed. A major tsunami generated by the collapse of the west flank of Ritter Island volcano in 1888 caused untold loss of life on New Britain and adjacent islands. More recently, in 1998, the Aitape tsunami destroyed two villages, damaged other villages, and caused more than 1650 deaths.

Solomon Islands

The central and western Solomon Islands form a double chain on either side of a central trough and are bounded on both flanks by deep-sea trenches (Figure 4). The islands comprise three geological provinces: a north-eastern province of Cretaceous–Paleocene oceanic basalts and pelagic sediments, a central province of more complex geology that includes metamorphic and ultramafic rocks and Oligocene–Miocene arc-related volcanics, and a south-western province of Late Miocene to Quaternary volcanic activity. The islands of the north-eastern province are Malaita, the small island of Ulawa, and the north-eastern side of Santa Isabel. On these islands, Cretaceous and Paleocene submarine basalts with intercalated pelagic limestone and mudstone are overlain by younger Cenozoic marine sediments that contain terrigenous material. The Cretaceous and Paleocene rocks originated as part of the Ontong Java submarine volcanic plateau. On Malaita, an intrusion of alnoite, a silica-poor mafic rock, dating to 34 Ma, has attracted interest in the past because of the occurrence of unusual xenoliths, including garnet peridotite.

The islands of the central province are Choiseul, the south-western side of Santa Isabel, the Florida islands, Guadalcanal, and Makira (San Cristobal). Basement on these islands comprises Cretaceous basalt, greenschist and amphibolite facies, mafic schists, and ultramafic rocks; the metamorphic rocks have radiometric ages around 50 Ma, Early to mid-Eocene. Volcanic rocks of Late Eocene(?), Oligocene, and Early Miocene age unconformably overlie the basement rocks, and volcano-related dioritic stocks intrude basement. A thick sequence of Miocene to Pliocene clastic sediments covers much of Guadalcanal Island but is not strongly developed on the other islands.

The south-west province comprises Late Miocene to Holocene arc volcanic rocks and associated

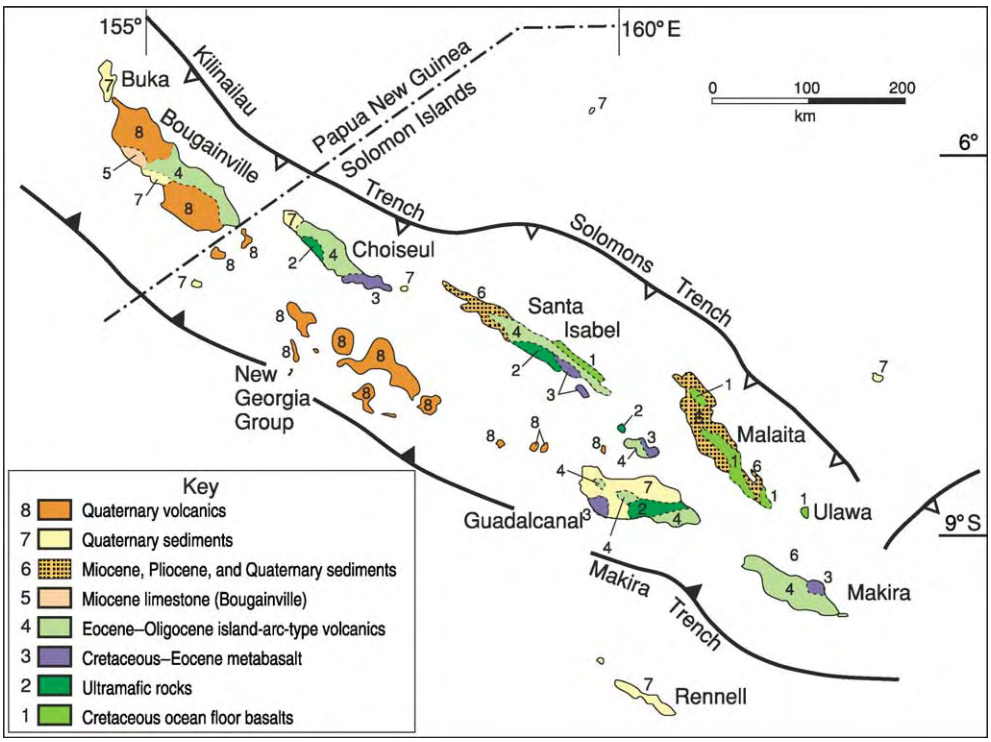


Figure 4 Geological map of the Solomon Islands, including Bougainville and Buka Islands (Papua New Guinea), showing the Makira (San Cristobal) active trench, the Kilinailau inactive trench, and the North Solomons slow acting trench. The dashed line marks the international border. Map drawn by L Cotterall, Geology Department, Auckland University.

sediments and intrusive rocks. These extend from Bougainville (PNG) and the Shortland Islands in the north-west, to the near end of Guadalcanal in the south-east. There is an outlier of Pliocene volcanics and intrusive rocks in central Guadalcanal, at Gold Ridge. Bougainville Island, politically a part of Papua New Guinea, is the north-westernmost of the Solomon Island chain and combines characteristics of both the central and south-western geological provinces. A basement of Eocene and Oligocene arc-related volcanic rocks is partly overlain by Miocene limestone and Plio-Quaternary volcanoes and is intruded by Pliocene dioritic stocks. The Solomon outer eastern islands are far removed from the central and western islands and are shown on the Vanuatu geological map (Figure 5). The islands of the Santa Cruz Group comprise Miocene to Holocene volcanic rocks and sediments. The Duff Islands, and the islands of Anuta, Fatutaka, and Tikopia (not shown on Figure 5) comprise Quaternary volcanic rocks and sediments. Tinakula in the Santa Cruz Group is the only active volcano.

Geological Evolution

The islands of the north-east province developed as a result of continued convergence across the line of the Kilinailau Trench (Figure 4), and resulting delamination and eastward overthrusting of an upper part of the Ontong Java volcanic plateau. The rocks of the lower part of the Ontong Java plateau continue to be subducted steeply beneath the Solomon Islands. The central province association of ultramafic rocks, Cretaceous basalt, and Eocene metamorphic rocks is thought to have developed as a result of subduction of Pacific plate at the Kilinailau Trench in the Eocene, but the apparently rapid exhumation of these rocks (exhumed before the Oligocene) is not readily explained. The same cycle of subduction yielded the Oligocene to Early Miocene arc-type volcanic rocks that overlie basement on the islands of the central province. The Late Miocene to Holocene volcanic rocks of the south-west province were generated by subduction of the Australian plate at the Makira (San Cristobal) Trench.

Economic Geology

Gold associated with Pliocene intrusives has been mined at Gold Ridge on Guadalcanal Island, and soils on ultramafic rocks on Santa Isabel Island have been tested for lateritic concentrations of nickel and cobalt.

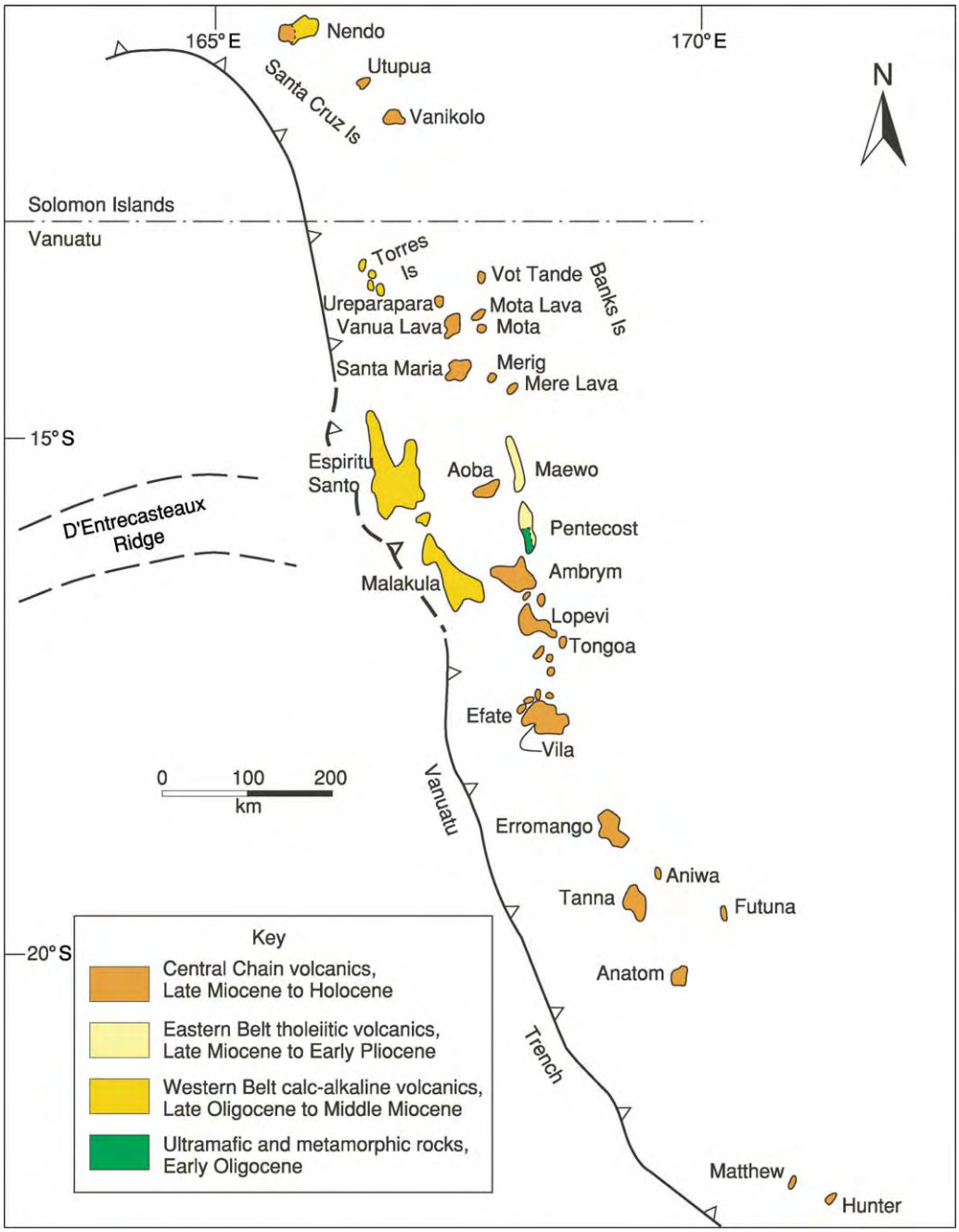


Figure 5 Geological map of Vanuatu, including the Santa Cruz Group in the Solomon Islands (the dashed line marks the international border). The map shows the active Vanuatu trench, the western belt Islands of Late Oligocene to Middle Miocene volcanics, the eastern belt (Maewo and Pentecost Islands) of Late Miocene to Early Miocene volcanic rocks, and the central chain of Late Miocene to Holocene volcanoes. Ultramafic and metamorphic rocks (dating to 35 Ma) are exposed on southern Pentecost Island. Map drawn by L Cotterall, Geology Department, Auckland University.

Vanuatu

The islands of the Vanuatu archipelago (Figure 5) form a narrow chain that extends for more than 1200 km, from the Torres Islands in the north to Matthew and Hunter islands in the south. The island chain is bounded eastward by the North Fiji Basin and westward by the Vanuatu Trench. The islands are, with one exception, entirely volcanic in origin, with associated sediments and intrusive rocks. The exception is on southern Pentecost Island, where ultramafic rocks associated with blocks of schistose amphibolite, metamorphosed pillow lava, and metagabbro are in

fault contact with Pliocene sediments. The metamorphic minerals have a radiometric age of around 35 Ma, and thus are older than any of the volcanic rocks.

The geology of Vanuatu has been described in terms of three volcanic provinces: a western province of Late Oligocene to Middle Miocene age (27–14 Ma; calc-alkaline volcanics), an eastern province of Late Miocene to Pliocene age (7–4 Ma; tholeiitic volcanics), and a central province of Miocene and Pliocene to Holocene volcanic rocks (6 Ma to the present day; Figure 5). The volcanic rocks of the western province developed in response to westward subduction of the Pacific plate at the Vitiaz Trench. The Vitiaz Trench, at that time, lay immediately east of Vanuatu. The trench moved away from Vanuatu (see Figure 1) with the opening of the North Fiji Basin in the past 6 million years.

The volcanic rocks of the eastern province developed in response to eastward subduction of Australian plate at the Vanuatu Trench. Magmatism in the central chain began at 6 Ma and was initially focused on Erromango, Tanna, and Anatom in the south. Volcanism subsequently developed along the entire length of the arc and shifted closer to the trench because of a steepening of the subduction zone to its present inclination of 70°.

The D'Entrecasteaux aseismic ridge, an Eocene–Oligocene island-arc complex, collided with central Vanuatu at 3–1.5 Ma, near Epi, and has migrated northwards as a result of the obliquity of the ridge relative to the trench. The collision has caused a slowing of subduction that may in turn have caused the development of back-arc thrust faults east of Maewo and Pentecost Islands, and the elevation of the ultramafic and metamorphic rocks on Pentecost. Five of the volcanoes of the central province are potentially dangerous. These are Ambae, Gaua, Ambrym, Tanna, and Lopevi. Only Yasur volcano on Tanna Island is monitored.

Economic Geology

Manganese oxide has been mined on Efate, where it typically is developed at contacts between tuff and overlying limestone. Occurrences of copper, zinc, molybdenum, and gold are known.

New Caledonia

The island of New Caledonia and the adjacent islands (Belep Islands to the north and Isle of Pines to the south) lie on the Norfolk Ridge. To the west, beyond the New Caledonia Basin, is Lord Howe Rise, and to the east are the Loyalty Ridge and Loyalty Islands (Figure 1). The rock units of New Caledonia (Figure 6) do not form a simple stratified sequence but, rather, comprise a series of terranes, juxtaposed during a complex history of convergent and extensional tectonic activity, and successor basin sediments. Four terranes make up a pre-Cretaceous basement. This is unconformably and tectonically overlain by Late Cretaceous and younger successor basin sediments and terranes.

Basement Terranes

The pre-Cretaceous basement is a complex amalgam of ophiolite, volcanics, and sediments that developed along the East Gondwana margin during the Permian to Jurassic. Four distinct terranes are recognized. The oldest, the Late Carboniferous Koh ophiolite, is a dismembered fore-arc ophiolite that is probably the basement on which two of the other three terranes, the Teremba and Central Chain terranes, were deposited. The Teremba Terrane, of mid-Permian to Late Jurassic age, has arc-related volcanics and volcaniclastics at its base, but is dominantly composed of a simply deformed thick sequence of diagenetically altered, fossiliferous shallow water sediments. The Central Chain Terrane, considered to be the distal equivalent of the Teremba Terrane, is composed of volcaniclastic sediments and minor arc-tholeiitic basalts that all have been weakly metamorphosed. The Boghen Terrane is a thick sequence of metagreywackes and schists, with sedimentary compositions ranging from pelitic to mafic, on a basal sequence of tholeiitic pillow lavas and red and green radiolarian cherts; detrital zircons in the sedimentary rocks indicate the sediments are of Jurassic age and derived from a south-east-Gondwana arc system and the Antarctic continent. The Boghen Terrane is considered to be a Jurassic accretionary complex, metamorphosed in the Late Jurassic.

Successor Basin Sediments

A transgressive sequence of Late Cretaceous coal measures and conglomerates rests unconformably on all of the older terranes, and forms thick sequences in the Nouméa–Bourail region and in the northern part of New Caledonia. Coarse carbonaceous sediments at the base of this sequence become finer upsection and are overlain by Paleocene to mid-Eocene siliceous shales and a pelagic chert and micrite sequence (pthanites). The coarser grained lower parts of the Late Cretaceous–Paleocene sequence locally contain shoshonitic basic volcanics, and there are also felsic horizons of rhyolitic flows and pyroclastics that appear to be younger than the basic volcanics. Later Eocene sequences, unconformably overlying the pelagic sediments, are carbonates. In the Bourail region, these are overlain by an alternating sandstone and siltstone flysch sequence. These are topped by an

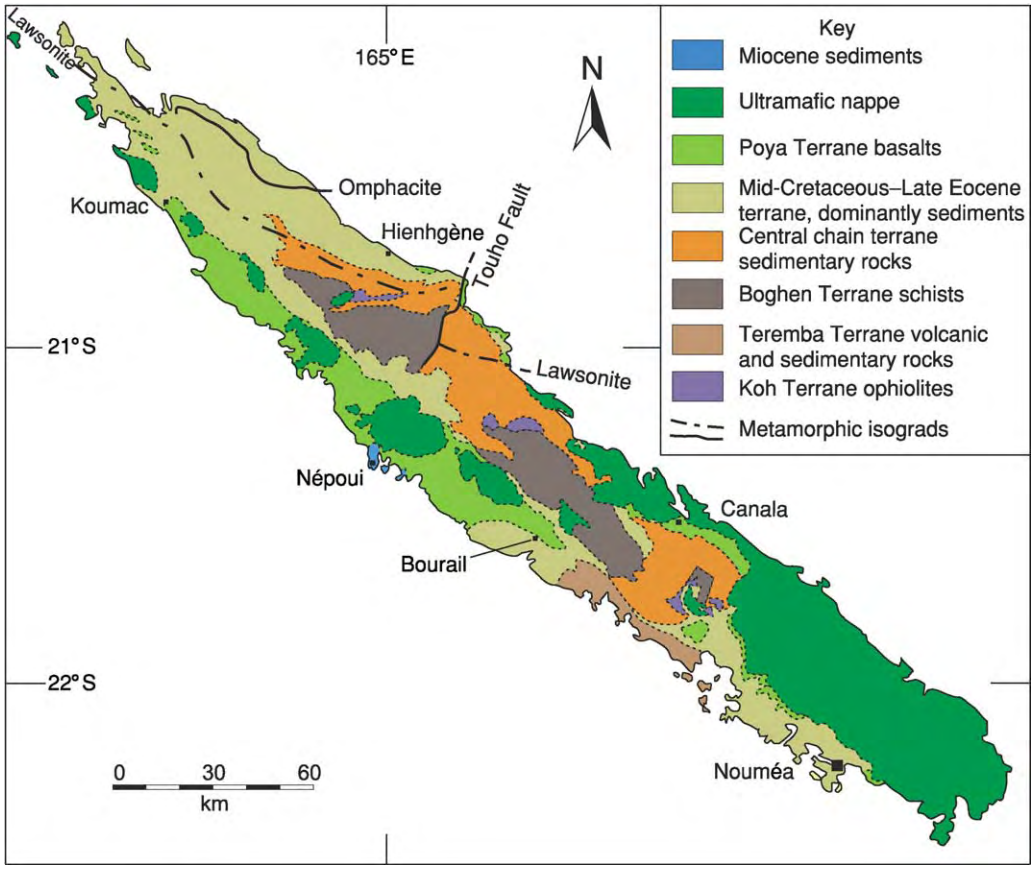


Figure 6 Geological map of New Caledonia (excluding the Loyalty Islands, Belep Islands, and Isle de Pines). The pre Cretaceous basement comprises the Koh ophiolite and the Teremba, Boghen, and Central Chain terranes. These are unconformably overlain by mid Cretaceous sediments forming a mid Cretaceous to Late Eocene terrane or successor basin sequence. Great thrust sheets of Poya Terrane basalts and ultramafic rocks (the ultramafic nappe), emplaced in the Late Eocene, overlie the older terranes. Miocene marine sediments are preserved on the south west coast. Map drawn by L Cotterall, Geology Department, Auckland University.

olistostrome that contains reworked components of the successor basin sediments, including upper Eocene bioclastic limestone and the flysch.

The Younger Terranes

The New Caledonian ophiolite, emplaced in the Late Eocene, overlies all the older rocks on the island. It consists of two units that appear to be geochemically unrelated: an overlying mainly harzburgitic ultramafic sheet (the ultramafic nappe) and the dominantly basic volcanic Poya terrane. The Poya terrane is composed of slices of geochemically variable submarine mafic volcanic rocks, mainly pillow lavas, together with dolerites and minor gabbros, and associated abyssal sediments. The ophiolite was emplaced as a result of Eocene convergence along the eastern margin of New Caledonia. Convergence was driven by the opening of the North Loyalty and South Fiji basins. Eastward-dipping subduction associated with the convergence consumed most of the South Loyalty Basin and caused high-pressure metamorphism of the Cretaceous–Eocene sediments of northern New Caledonia. The Eocene Loyalty Island volcanic arc formed above the subduction zone.

Miocene

Early Miocene coral reef and shallow-water marine sediments containing material derived from the ultramafic sheet are exposed at Nepoui. Small intrusions of granodiorite of Miocene age intrude the ultramafic sheet east of Nouméa, and north-east of Canala. Buoyancy-driven uplift in the Miocene has exhumed and exposed the high-pressure metamorphosed Cretaceous–Eocene sediments on the north-eastern coast of New Caledonia.

Economic Minerals

Since the Miocene, New Caledonia has been emergent and subjected to intensive tropical weathering, with the result that thick lateritic soils developed on the ultramafic rocks. The laterites have been mined for nickel, cobalt, and iron since the late nineteenth

century and still comprise the largest single reserve of nickel worldwide. Chromite has been mined directly from the ultramafic rock. Manganese deposits are associated with the Poya terrane basalts. Copper, lead, and zinc deposits have been mined from volcano-sedimentary horizons in the metamorphic rocks of the Diahot valley in northern New Caledonia. Gold-bearing quartz veins have been mined in the Diahot valley and gold and molybdenum deposits are associated with the Miocene granodiorite intrusions.

Fiji

The Fiji archipelago (Figure 7) comprises 800 islands, most of which lie between 16° and 19° south, 177° east, and 178° west. The Lau Group of islands in the east extends to 21° south (Figure 1). The largest islands are Viti Levu and Vanua Levu. The island of Rotuma, 650 km north of Viti Levu, is also a part of the Republic of Fiji. The islands are mainly volcanic and of island-arc origin, of basaltic, basaltic andesite, andesitic, and dacitic composition, and tholeiitic, calc-alkaline, or shoshonitic affinities. The younger volcanic rocks are mostly more alkaline and mainly belong to the alkali–basalt suite, and are intraplate rather than subduction related.

Viti Levu

The oldest rocks in the Fiji archipelago are exposed on Viti Levu. These are the volcanic and plutonic rocks of the Late Eocene to Early Oligocene Yavuna arc, and of the Late Oligocene to Middle Miocene (and possibly Late Miocene) Wainimala arc. The two volcanic suites are separated by an unconformity. The Yavuna arc rocks are exposed in western Viti Levu and include basaltic pillow lavas and volcaniclastic rocks, dacite, tonalite, and limestones of both Eocene and Oligocene ages. The basalts are geochemically primitive to arc like.

The Wainimala arc rocks are exposed throughout southern and central Viti Levu and on the islands of Malolo and Waya, immediately to the west. The rocks include mainly dacites in the centre of the island and basalts overlain by pelagic limestone and dacitic and andesitic lavas and volcaniclastics in the south-west. On Malolo and Waya, there are basaltic pillow lavas and breccia, and swarms of basalt dykes, indicating the local arc axis. Gabbro and tonalite of the Colo Plutonic Suite are thought to be part of the Wainimala arc activity and have mid- to Late Miocene radiometric ages (12–7 Ma). They were exposed to erosion in the Late Miocene. The Colo, Yavuna and Wainimala rocks show the effects of some deformation and weak metamorphism. They are unconformably overlain by latest Miocene to Early Pliocene basaltic and andesitic volcanic rocks and associated sediments (radiometric ages, 5.6–4 Ma). These cover much of the northern half of the island. The centre of the island was elevated at about 4 Ma. In the south-east, sediments were deposited in relatively deep water during the latest Miocene and through the Early Pliocene, the Messinian sea-level low being represented by a subaerially eroded coral reef. Shallow-water sediments were laid down in the latest Pliocene. In the south-west, sandstone and siltstone, overlain by limestone, were deposited in the Early Pliocene.

Vanua Levu

Vanua Levu comprises Late Miocene to Pliocene volcanic rocks (7–2.8 Ma). Volcanic activity appears to have progressed with time from east–north-east to west–south-west. The volcanic rocks are generally tholeiitic to calc-alkaline and include basalt, andesite, and some dacite, including a strong development of dacite and rhyolite in the north-east. Boninitic basalts were erupted on Cikobia to the north-east. Vanua Levu was elevated at about 3.5 Ma and Seatura Volcano began to form in western Vanua Levu not later than 3.3 Ma. Clastic sediments were deposited along the south-east coast.

Yasawa Group

The Yasawa and other islands north-west of Viti Levu, other than those of the Wainimala arc, were formed by volcanic arc activity (Yasawa arc) in the Late Miocene, from around 8 Ma. Early submarine dacites and basalts were succeeded by subaerial basalts in the central Yasawa islands, and by varied andesites. Strata generally dip towards Viti Levu, but at the north-eastern and southern ends of the chain have been strongly deformed with the development of overturned folds and thrust faults.

Islands Immediately South of Viti Levu

Vatulele, Beqa, and Yanuca (Serua) are Early Pliocene volcanoes south of Viti Levu. On Vatulele, shoshonitic rocks were erupted through limestone.

Islands of the Koro Sea

The islands of the Koro Sea (Lomaiviti and the Moala Group) mostly formed by volcanic activity in the Early Pliocene. Exceptions are the island of Koro, which formed in the Late Pliocene, and a Late Pleistocene cone on Nairai. The rock types range from tholeiitic through calc-alkaline to alkali basaltic and shoshonitic, and are mostly basaltic, with some hornblende andesite, monzonite, olivine monzonite, and tonalite. Olivine nodules occur in basalts on southern Moala and the western Koro islands.

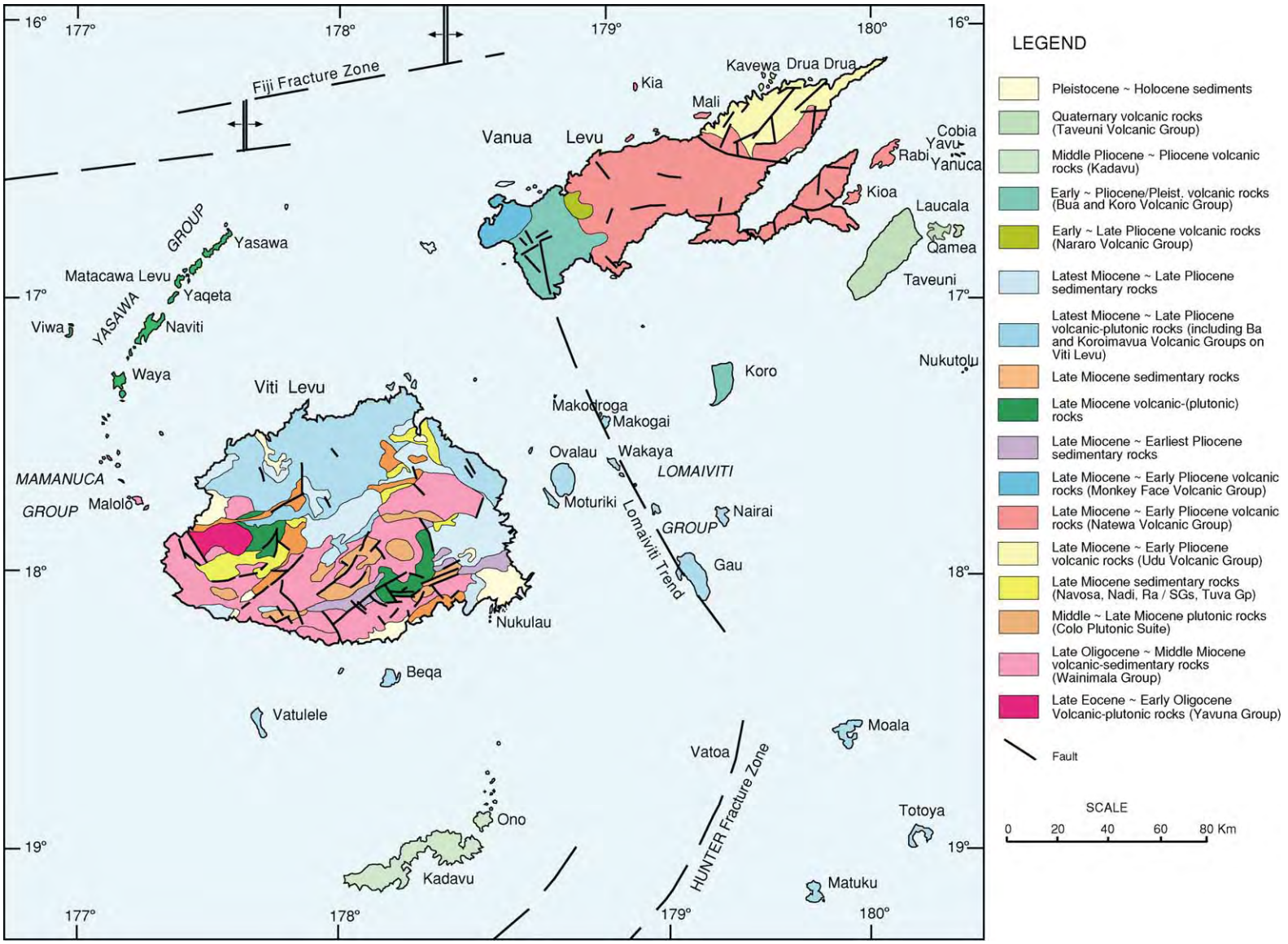


Figure 7 Geological map of Fiji, excluding Lau Islands and smaller Islands. On Viti Levu, the main rock units are Late Eocene Early Oligocene Yavuna volcanic and plutonic rocks, Late Oligocene Middle Miocene Wainimala volcanic rocks, Middle to Late Miocene Colo plutonic rocks, Late Miocene sediments, and Late Miocene to Pliocene volcanics and sediments. On Vanua Levu, the main rock units are Late Miocene Early Pliocene Udu volcanics, Late Miocene Early Pliocene Natewa volcanics, Early to Late Pliocene Nararo volcanic rocks, and Early Pliocene Pleistocene volcanics. On Koro Island are Early Pliocene Pleistocene volcanics; and on the Kadavu Islands are Middle Pliocene volcanic rocks. Map reproduced with permission from Mineral Resources Department, Fiji.

Kadavu Islands

The Kadavu islands, south of Viti Levu, are a volcanic island arc related to subduction at the Kadavu Trench. Ages range from 3.2 Ma to Holocene. Compositions range from basaltic to dacitic and are generally potassic.

Lau Islands

The islands on the Lau Ridge (Figure 1) formed in several periods of volcanic activity, probably beginning in the Eocene. The oldest exposed rocks, on Yacata island, are Middle Miocene (14 Ma). Further south, Late Oligocene rocks have been recovered by dredging. The exposed volcanic rocks are of three ages: Middle and Late Miocene, Middle Pliocene, and Late Pliocene to Pleistocene. Eruptive rocks in the two earlier episodes were tholeiitic to calc-alkaline, and in the third episode were alkali basalt. Limestone was deposited in the Middle to Late Miocene, Late Miocene to Early Pliocene, and Late Pliocene to Quaternary.

Quaternary Volcanic Activity

Quaternary volcanic activity of mainly alkali basalt composition formed the island of Taveuni immediately south-east of Vanua Levu (where the most recent dated eruption was at 340 ± 70 years), and the island of Rotuma, 650 km north of Viti Levu. Many small centres of Quaternary alkali-basalt volcanism are scattered around the archipelago in the north-east, the north-west, in Lomaiviti, adjoining Kadavu and in the Lau Group, but none is known on the Viti Levu platform. High-K andesite to dacite erupted at the western end of Kadavu island.

Economic Geology

Gold has been mined at Vatukoula on Viti Levu since 1933 and has been mined at other locations on Viti Levu and at Mt Kasi on Vanua Levu at different times. Manganese ore was mined from 1950 to 1971, most production being from Viti Levu in the 1950s. Magnetite was mined at Tuveriki in south-west Viti Levu from 1957 to 1963. Deposits of gold, polymetallic sulphides, and disseminated deposits are being explored, notably the Waisoi and other porphyry–copper deposits at Namosi on Viti Levu. Bauxite deposits occur in western Vanua Levu, mainly around Wainunu, and some islands of the Lau Group have deposits of phosphatic clay.

Plate Tectonic Setting

The early volcanic arcs that now form the basement of the Fiji platform developed above subducted Pacific plate in the Eocene, Oligocene, and Miocene. Initially the Yavuna arc and Lau–Tonga Ridge probably were parallel to, and close to, New Caledonia, but were moved to the north-east by the opening of the South Fiji Basin (36–25 Ma). By the Middle Miocene, the Lau–Tonga arc, Fiji, and Vanuatu arcs formed a single east-facing arc above subducted Pacific plate. This simple arc, sometimes referred to as the Vitiaz Arc, was disrupted in the Late Miocene with the onset of back-arc extension in the North Fiji Basin. This extension caused the clockwise rotation of Vanuatu and the anticlockwise rotation of the Fiji platform. In the Pliocene, spreading extended east and then south, causing separation of the Lau and Tonga ridges and the oblique subduction of South Fiji Basin lithosphere at the Kadavu Trench. Rotation of the Fiji platform caused local compression and extension. As Fiji moved away from sites of active subduction, volcanism became mainly alkali basalt in composition.

Tonga

The Tongan archipelago is a north–north-east-trending double-island chain that straddles the northern end of the active convergent plate boundary that separates the Pacific and Australian plates. A well-defined Benioff zone dips westward at 43–45° beneath Tonga and extends to a depth of 700 km. The larger islands of the archipelago form the eastern chain, 130–150 km west of the Tonga–Kermadec Trench. These are low-lying islands made up mainly of coraline limestone. On Eua, the southernmost of this eastern chain, the stratigraphic sequence has been described as consisting of coraline limestone, Miocene tuffs and tuffaceous limestones, Late Eocene tuffaceous marlstone (thickness, 60–80 m), Late Eocene foraminiferal and algal limestone, and Middle Eocene (46 Ma) basaltic and intermediate lava, volcanic breccia, and tuffs cut by dykes. Although the Eocene and Miocene rocks are not exposed elsewhere in Tonga, seismic profiles indicate that they form the basement of the Tonga Ridge. The crust beneath the ridge is 12–18 km thick. Development of the Tonga Trench and uplift of the frontal arc occurred in Middle to Late Oligocene. The Miocene tuffs are thought to represent the initial stages of arc volcanism. The islands of the western chain, which lie 150–200 km west of the trench, comprise the Tonga or Tofua volcanic arc. The volcanoes have a history of intense modern activity, with more than 35 eruptions occurring during the past 200 years, including both subaerial and submarine events. The volcanic products are mainly basalt and basaltic andesite with subordinate dacite, as is to be expected from an intra-oceanic arc system. West of the Tonga or Tofua arc is

the Lau basin, an area of attenuated crust and high heat flow that is interpreted as an extensional back-arc basin.

Samoa

The Samoan archipelago lies between latitudes 13° and 15° S and longitudes 169° and 173° W and extends for 500 km on azimuth 288°, from Rose Island in the east–south-east to Savaii Island in the west–north-west. The archipelago comprises a chain of hotspot-related intraplate volcanoes made up of the alkaline and tholeiitic basalts that are typical of Pacific island chains. The two largest islands are Savaii (1700 km^2) and Upolu (1115 km^2). These, together with the small islands of Manono and Apolima, comprise Western Samoa.

Savaii is the summit of a large basaltic volcano. Many small parasitic cones occur about its flanks and the lavas of the historical eruptions of 1760 and 1905–11 are widely distributed across the northern part of the island. On Upolo Island, young cones and lava flows overlie older, deeply eroded volcanic sequences that form spectacular topography on the eastern end of the island. Similar geology is seen on Tutuila, the third largest island, where post-erosional lavas overlie an older group of volcanics. Further east, the volcano chain is reduced to subaerial remnants in the Manu'a islands and the easternmost Rose Atoll.

Pacific island chains typically increase in age westward. Dating of Samoan lavas and dredged tholeiitic samples from seamounts west of Savaii confirm that this is the case in the older volcanic rocks of the Samoan chain. For example, shield volcanics on Upolu range in age from 2.8 Ma to 1.7 Ma and on Tutuila, to the east, range from 1.54 to 1.28 Ma. The post-erosional volcanic activity on Savaii, Upolu, and Tutuila, from 700 000 years ago to the present, does not fit the 'older westward' pattern. Possibly this is a side effect of the sharp change in the orientation of the Tonga–Kermadec Trench which, at a point about 100 km south of the Samoan islands, swings from a near north–south azimuth to near east–west (Figure 1). Flexural bending of the Pacific plate at this point presumably could cause rifting in the Samoan archipelago and thus provide a conduit for renewed volcanic activity.

See Also

Gondwanaland and Gondwana. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Hydrothermal Vents At Mid-Ocean Ridges.

Further Reading

Coleman PJ (ed.) (1973) *The Western Pacific: Island Arcs, Marginal Basins and Geochemistry*. Perth: University of Western Australia Press.

Hillis RR and Muller RD (2003) *Evolution and Dynamics of the Australian Plate. Geological Society of Australia Special Publication 22*. Sydney: Geological Society of Australia.

Kroenke LW (1984) *Cenozoic Development of the South west Pacific. SOPAC Technical Bulletin 6*. Suva: South Pacific Applied Geoscience Commission.

Veevers JJ (ed.) (2000) *Billion Year Earth History of Australia and Neighbours in Gondwanaland*. Sydney: GEMOC Press, Macquarie University.

General

The following Earth Science Series, publications on *The Geology and Offshore Resources of Pacific Island Arcs*, are published by the Circum Pacific Council for Energy and Mineral Resources (Menlo Park, CA) and are available from the AAPG Bookshop, PO Box 979, Tulsa, OK, USA.

Series 1: Howell DG (ed.) (1985) *Tectonostratigraphic Terranes of the Circum Pacific Region*.

Series 2: Scholl DW and Vallier TL (eds.) (1985) *Tonga Region*.

Series 3: Brocher TM (ed.) (1985) *Geological Investigations of the Northern Melanesian Borderland*.

Series 4: Vedder JG, Pound KS, and Boundy SQ (eds.) (1986) *Central and Western Solomon Islands*.

Series 7: Taylor B and Exon NF (eds.) (1987) *Marine Geology, Geophysics and Geochemistry of the Woodlark Basin Solomon Islands*.

Series 8: Greene HG and Wong FL (eds.) (1988) *Vanuatu Region*.

Series 9: Marlow MS, Dadisman SV, and Exon NF (eds.) (1988) *New Ireland and Manus Region, Papua New Guinea*.

Series 12: Vedder JG and Bruns TR (eds.) (1989) *Solomon Islands and Bougainville*.

Fiji

Colley H and Flint DJ (1995) *Metallic Mineral Deposits of Fiji. Fiji Mineral Resources Department Memoir 4*. Suva: Fiji Mineral Resources Department.

Hathway B and Colley H (1994) Eocene to Miocene geology of southwest Viti Levu, Fiji. In: Stevenson AJ, Herzer RH, and Ballance PF (eds.) *Geology and Submarine Resources of the Tonga Lau Fiji Region. SOPAC Technical Bulletin 8*, pp. 153 169. Suva: South Pacific Applied Geoscience Commission.

Rodda P (1994) Geology of Fiji. In: Stevenson AJ, Herzer RH, and Ballance PF (eds.) *Geology and Submarine Resources of the Tonga Lau Fiji Region. SOPAC Technical*

Bulletin 8, pp. 131 151. Suva: South Pacific Applied Geoscience Commission.

New Caledonia

Aitchison JC, Ireland TR, Clarke GL, Cluzel D, Davis AM, and Meffre S (1998) Regional implications of U/Pb SHRIMP age constraints on the tectonic evolution of New Caledonia. *Tectonophysics* 299: 333 343.

Black PM (1995) High Si rhyolites and shoshonitic volcan ics; a late Cretaceous bimodal association, Nouméa basin, New Caledonia. In: Mauk JL and St George JD (eds.) *Proceedings of the Pacrim Congress, September 1995, Auckland, New Zealand,* pp. 55 58.

Cluzel D and Meffre S (2002) L'unité de la Boghen (Nouvelle Calédonie, Pacifique sud ouest): un complexe d'accrétion jurassique. Données radiochronologiques preliminaries U Pb sur les zircons détritiques. *CR Geoscience* 334: 867 874.

Cluzel D, Aitchison JC, and Picard C (2001) Tectonic ac cretion and underplating of mafic terranes in the Late Eocene intraoceanic fore arc of New Caledonia (South west Pacific): geodynamic implications. *Tectonophysics* 340: 23 59.

Paris JP (1981) *Géologie de la Nouvelle Calédonie. Un Essai de Synthèse. Memoirs du BRGM, No. 113.* Orleans: BRGM.

Yokoyama K, Brothers RN, and Black PM (1986) Re gional facies in the high pressure metamorphic belt of New Caledonia. In: Veevers JJ and Powell CM (eds.) *Permian Triassic Pangean Basins and Foldbelts along the Panthalassan Margin of Gondwanaland. Geological Society of America Memoir 184,* pp. 407 571. Boulder, CO: Geological Society of America.

Papua New Guinea

Corbett GJ and Leach TM (1998) *Southwest Pacific Rim Gold Copper Systems: Structure, Alteration and Miner alisation. Economic Geology Special Publication 6.* Littleton, CO: Society of Economic Geologists.

Davies HL, Perembo RCB, Winn RD, and KenGemar P (1997) Terranes of the New Guinea orogen. In: Hancock G (ed.) *Proceedings of the Geology Exploration and Mining Conference, Madang,* pp. 61 66. Mel bourne: Australasian Institute of Mining & Metallurgy.

Dow DB (1977) *A Geological Synthesis of Papua New Guinea, Bulletin 201.* Australia: Bureau of Mineral Resources.

Pigram CJ and Davies HL (1987) Terranes and accretionary history of the New Guinea orogen. *BMR Journal of Australian Geology and Geophysics* 10: 193 211.

Solomon Islands

Coulson FI and Vedder JG (1986) Geology of the Central and Western Solomon Islands. In: Vedder JG, Pound KS, and Boundy SQ (eds.) *Geology and Offshore Resources of Pacific Island Arcs Central and Western Solomon Islands, Earth Science Series 4,* pp. 59 87. Menlo Park, CA: Circum Pacific Council for Energy and Mineral Resources.

Hughes GW, Craig PM, and Dennis RA (1981) *Geology of the Outer Eastern Islands. Geological Survey Division Solomon Islands Bulletin 4.* Honiara, Solomon Islands: Mines and Minerals Division.

Vanuatu

Auzende JM, Pelletier B, and Eissen J P (1995) The North Fiji Basin: Geology, structure and geodynamic evolution. In: Taylor B (ed.) *Backarc Basins: Tectonics and Magma tism,* pp. 139 175. New York: Plenum.

Calmant S, Pelletier B, Lebellegard P, Bevis M, Taylor FW, and Phillips DA (2003) New insights on the tectonics along the New Hebrides subduction zone based on GPS results. *Journal of Geophysical Research* 108(B6): 2319.

Greene HG, Collot J Y, Fisher MA, and Crawford A (1994) Neogene tectonic evolution of the New Hebrides island arc: a review incorporating ODP drilling results. In: Greene HG, Collot J Y, Stokking LB, *et al.* (eds.) *Pro ceeding of the Ocean Drilling Program, Scientific Results 134,* pp. 19 46. College Station, TX: Ocean Drilling Program.

Macfarlane A and Carney JN (1987) *Geology and Mineral isation of North and Central Malekula. General Report.* Republic of Vanuatu: Department of Geology, Mines & Rural Water Supplies.

Macfarlane A, Carney JN, Crawford AJ, and Greene HG (1988) Vanuatu a review of the onshore geol ogy. In: Greene HG and Wong FL (eds.) *Geology and Offshore Resources of Pacific Island Arcs Vanuatu Region, Earth Science Series 8,* pp. 45 91. Menlo Park, CA: Circum Pacific Council for Energy and Mineral Resources.

Maillet P, Ruellan E, Gérard M, Person A, Bellon H, Cotten J, Joron JL, Nakata S, and Price RC (1995) Tectonics, magmatism, and evolution of the New Hebrides back arc (south west Pacific). In: Taylor B (ed.) *Backarc Basins: Tectonics and Magmatism,* pp. 177 235. New York: Plenum.

Mitchell AHG and Warden AJ (1971) Geological evolution of the New Hebrides island arc. *Journal of the Geological Society of London* 129: 501 542.

Warden AJ (1971) *Manganese Mineralisation in the New Hebrides. New Hebrides Condominium Geological Survey.* New Hebrides: British Service.

ORIGIN OF LIFE

J Bailey, Anglo-Australian Observatory and Australian Centre for Astrobiology, Sydney, Australia

Introduction

Establishing the nature of the processes by which life originated is one of the most fundamental unsolved scientific problems. Fifty years of study have enabled us to assemble many pieces of the puzzle but have also left a number of critical gaps in our understanding. Even the issues of when and where life originated are still poorly constrained, but life probably appeared on Earth between about 4.2 Ga and 3.5 Ga ago. There is good evidence that the current DNA–RNA–protein basis for life developed from a stage known as the 'RNA world' in which RNA performed the information-storage and catalytic roles currently performed by DNA and proteins. Direct formation of the RNA world from prebiotic chemistry seems unlikely, so much current research focuses on possible precursors to the RNA world. Various suggestions for pre-RNA worlds have been made, but none has yet proved entirely satisfactory. A number of terrestrial and extraterrestrial processes can produce prebiotic organic molecules, but there are still problems in obtaining the right molecules in sufficient quantities.

Development of Ideas on the Origin of Life

The idea that all life on Earth has a common origin became well established only in the twentieth century. Early ideas on the evolution of life were most clearly expressed by Lamarck, who described the process as one of progression from simpler to more complex and advanced forms (*see* **Evolution**). The observation that both simple and complex organisms are currently present therefore required that simple organisms are appearing even today from inorganic matter by a process of 'spontaneous generation' (a concept that has its origins in antiquity). Thus, the ideas of transformism (evolution) and spontaneous generation were closely linked in the early nineteenth century and put Lamarck and his followers in conflict with the religious and scientific establishment, who argued for the immutability of species, which were divinely created in their current forms. Lamarck's ideas were criticized by scientists such as Cuvier (*see* **Famous Geologists:** Cuvier) (in a celebrated debate with Geoffroy Saint-Hilaire in 1830) and later by Pasteur, who carried out his famous 'swan-necked flask' experiments to discredit the idea of spontaneous generation.

Darwin's theory of evolution by natural selection opened the way for the modern view of the origin of life (*see* **Famous Geologists:** Darwin). In Darwin's theory there is not necessarily a ladder of progress from simple to more complex forms. Simple organisms can be as evolutionarily successful as complex ones. This allowed the idea that all life, simple and complex, had a single origin in the distant past. While the subject of the origin of life is hardly mentioned in Darwin's published writings, the following quote (from a letter to his botanist friend Joseph Hooker) gives us an inkling of his thoughts on the subject at that time:

> If (and oh! what a big if!) we could conceive in some warm little pond, with all sorts of ammonia and phos phoric salts, light, heat, electricity present, that a protein compound was chemically formed, ready to undergo still more complex changes... .

Oparin and Haldane in the early twentieth century developed the idea that chemical reactions on the early Earth could have led to the production of a range of organic compounds, forming a 'primordial soup' in which the required building blocks for life would have been present.

But it was not until 1953 that these ideas received experimental support. Stanley Miller, then a graduate student working in the laboratory of Harold Urey, set up his famous experiment in which electrical discharges were passed through a mixture of gases (methane, ammonia, hydrogen, and water vapour) simulating a thunderstorm on the primitive Earth. The experiment produced a mixture of several amino acids, the building blocks of proteins. Miller speculated that this was how organic compounds had been made on the early Earth. In the same year, Crick and Watson published their structure for DNA, the first step in elucidating the fundamental molecular basis of life. These two discoveries meant that we had both a plausible way of generating simple organic building blocks and an understanding of the macromolecules on which

life depends. A detailed experimental and theoretical study of the origin of life was now possible.

The Earliest Life

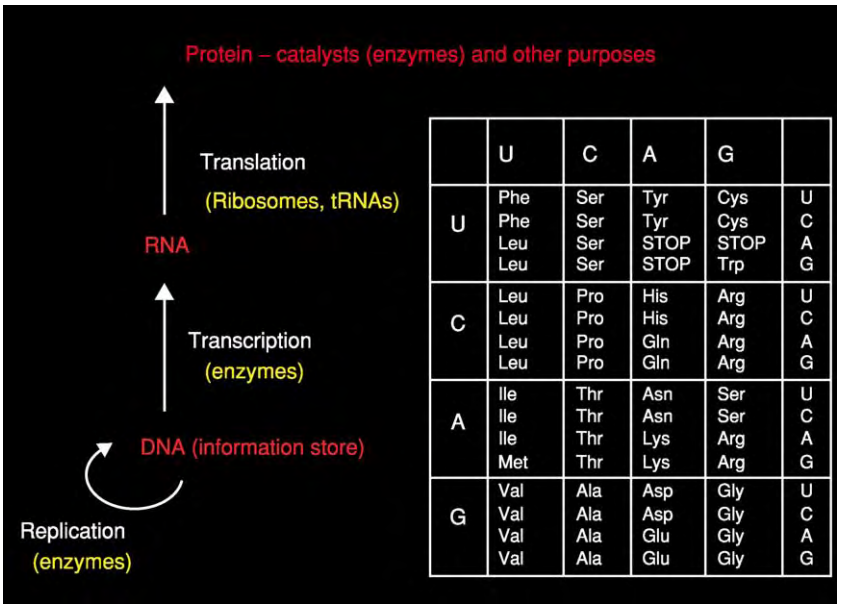
Sedimentary rocks of about 3.5 Ga old are found in the Pilbara region of Western Australia and in the Barberton Mountain Land in South Africa; these are the oldest such rocks that are sufficiently well preserved to contain fossils. These rocks contain structures that are interpreted as fossil stromatolites (layered structures built by colonies of micro-organisms) as well as filamentous microstructures interpreted to be microfossils of bacteria (*see* **Biosediments and Biofilms**). The biogenicity of many of these early fossils is controversial, but, if accepted, their presence indicates that life was well established on Earth at 3.5 Ga. It is even more difficult to say much about the nature of this early life. Some of the Pilbara microfossils have been suggested to resemble present-day cyanobacteria, but recent reinterpretation of the stratigraphy of the region suggests that the rocks originated in a hydrothermal environment.

Evidence for life has also been claimed in rocks of about 3.8 Ga from Greenland. These rocks are too metamorphosed to preserve physical fossils, but they contain carbon depleted in ^{13}C, which is consistent with the occurrence of biochemical reactions. These claims are also controversial.

The Tree of Life

The early history of life can also be studied through the techniques of molecular biology. It has been found that all life on Earth shares the same fundamental chemical processes. All life uses DNA as its basic genetic material. The DNA message is read by copying it onto RNA (transcription), and this RNA message is used to synthesize proteins (translation) by means of an RNA–protein complex called the ribosome (Figure 1). The resulting proteins are most frequently used as catalysts (enzymes) to drive the many chemical processes in the cell. The genetic code – which translates three-'letter' sequences of the four DNA bases (A, C, G, and T) into the corresponding amino acid to be added to a protein – is almost universal (with a few minor variations).

The common chemical basis for life strongly supports the idea that all life on Earth is related and descended from a single universal ancestor, sometimes known as LUCA (last universal common ancestor). Moreover, by comparing the sequences of genes involved in these fundamental processes common to all life, we can construct a 'family tree' showing the relatedness of all groups of life. Such molecular phylogenetic analysis has led to the 'three domain' taxonomy, in which life is divided into the domains Bacteria, Archaea, and Eucarya (Figure 2).



	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Figure 1 Summary of the key molecular processes common to all life. Genetic information is carried by DNA and can be copied (replication) onto new DNA molecules. The genetic message is read by first copying onto RNA (transcription) and then using the RNA message to synthesize a protein (translation). Three letter sequences of the RNA bases (uracil (U), cytosine (C), adenine (A), and guanine (G)) are translated into a corresponding amino acid to include in a protein according to the genetic code shown on the right.

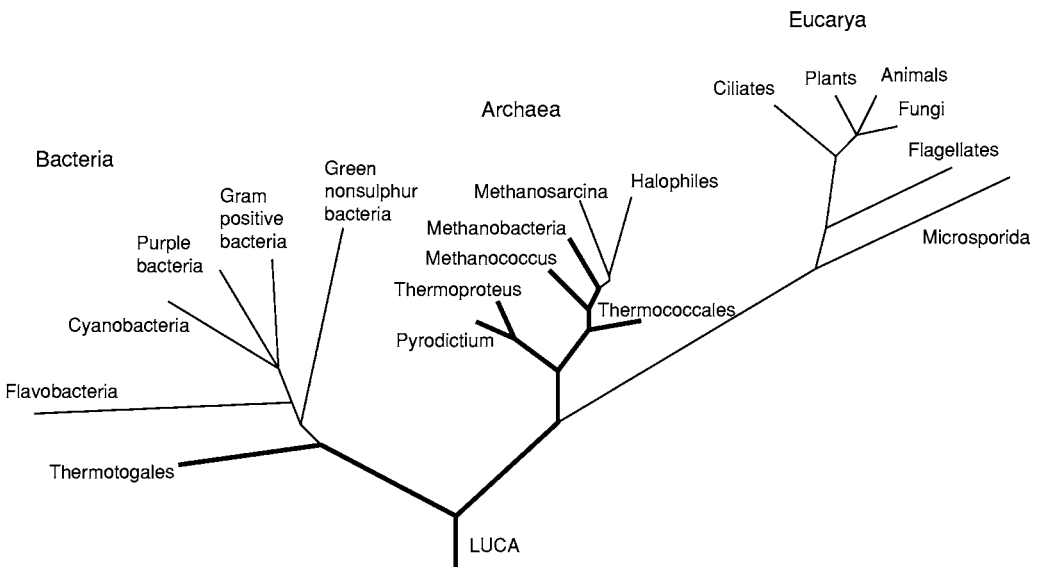


Figure 2 The tree of life derived from analysis of the small subunit ribosomal RNA gene, showing the division of life into the three domains Bacteria, Archaea, and Eucarya. Heavy lines indicate the branches containing hyperthermophilic organisms and illustrate the hypothesis that LUCA (the last universal common ancestor) was a hyperthermophile.

One feature of this phylogenetic tree is that many of the deepest branches closest to the 'root' of the tree are found to contain the hyperthermophiles (organisms adapted to very high temperatures). This has led to the suggestions that LUCA itself was probably a hyperthermophile and that perhaps life originated in a high-temperature environment. However, the interpretation of the tree is complicated by the likely occurrence of horizontal gene transfer (transfer of genes between different species), and there are possible interpretations of the tree that do not require LUCA to be hyperthermophilic.

Even if LUCA was a hyperthermophile, this does not necessarily tell us much about the environment in which life originated, since we do not know how much evolution occurred before LUCA. For example, one scenario is that life had already adapted to a wide range of environments before LUCA, but a catastrophic event such as a major impact wiped out everything except the hyperthermophiles.

LUCA is likely to have been similar to present-day bacteria. While these are usually regarded as simple organisms, even the simplest bacteria still contain substantial complexity, having genome sequences of several hundred thousand base pairs and using several hundred enzymes to catalyse a wide range of metabolic processes. The basic problem of the origin of life is to understand how such complexity could arise from a system simple enough to have assembled by chance in a prebiotic environment.

The key process is likely to be evolution by Darwinian natural selection. This process is capable of developing increased complexity. We therefore need to identify simpler systems that preceded the current DNA–RNA–protein basis for life but retain the basic requirements for natural selection to operate. While we have no fossil record to guide us in studying this stage of life's history, there are clues in the molecular mechanisms of modern organisms.

The RNA World

The DNA–RNA–protein basis for life presents a chicken-and-egg problem. DNA requires protein enzymes to catalyse its replication and transcription. However, these proteins can be made only by using the DNA genes that code for them. Hence, DNA is dependent on proteins and vice versa. A possible solution to the problem of how such a system originated was provided by the discovery by Cech and Altman that RNA can exhibit catalytic activity. This means that RNA can perform the roles of both a replicating information store (like DNA) and a catalyst (like protein enzymes).

This discovery led to the concept known as the RNA world – the idea that there was a stage in the development of life in which RNA acted as both the genome and the catalyst. These roles were subsequently taken over by DNA and proteins, leaving RNA in its modern role as an intermediary between DNA and proteins. The RNA-world concept is supported by a number of features of the molecular structure and chemistry of modern cells.

First, the structure of the ribosome, which catalyses protein synthesis, has recently been determined in detail. Although the ribosome contains both RNA and protein components, it has been found that the important catalytic activity comes from the RNA components, with the protein playing a peripheral role. This shows that protein synthesis (translation) is carried out almost exclusively by RNA and therefore probably originated at a time when only RNA catalysts were available. By contrast, the replication and transcription of DNA is carried out mostly by enzymes. Second, the synthesis of the nucleotides of DNA is carried out by building the corresponding RNA nucleotides first and then converting them to DNA nucleotides. Third, there are a number of important biochemical processes that use RNA (but not DNA) nucleotides or closely related molecules. These include the use of ATP (adenosine triphosphate) as the main energy carrier, and the cofactors used to assist a number of enzymes.

All these lines of evidence point to RNA being more ancient than either DNA or proteins and suggest that a pure RNA world 'invented' protein synthesis, giving rise to an RNA–protein world; subsequently, DNA was adopted as the primary genetic material.

For a pure RNA world (i.e. one in which proteins play no role) there must have been an RNA catalyst (or ribozyme) capable of catalysing the replication of RNA. No such molecule exists in modern organisms (although there are protein enzymes, known as RNA-dependent RNA polymerases, that do this). However, attempts have been made to synthesize such an RNA polymerase ribozyme. This is done by a process of 'in-vitro evolution', in which many copies of an RNA strand are made with random variations and then selected for appropriate catalytic activity. After many cycles of selection, ribozymes have been made that are capable of copying short sections of RNA (up to 14 bases) but are too inefficient to copy larger molecules.

The Origin of the RNA World

An RNA world is now widely accepted as being a stage in the development of life. The RNA world was either the first stage of life or preceded by an earlier stage of life, the pre-RNA world.

If the RNA world was the first stage of life then there must be a way of making a viable RNA world through the random processes of prebiotic chemistry. However, there are a number of difficulties with such a process. RNA consists of a sugar–phosphate backbone carrying four possible bases – guanine, adenine, cytosine, and uracil. Prebiotic syntheses have been suggested for the bases, but some require contrived conditions. The ribose sugar can be produced by the formose reaction but only in small quantities as part of a complex mixture of other sugars. The assembly of a complete ribonucleotide from these components presents further difficulties.

The nucleotides would then need to be assembled into RNA strands of sufficient length to show useful catalytic activity. Experiments using montmorillonite clay as a catalyst have led to RNA molecules of up to about 40 nucleotides. However, for this to be possible the nucleotides must be provided in their activated, or energy-rich, triphosphate form.

A further difficulty with an RNA world is imposed by the chirality of RNA. RNA replication (like many processes in biological chemistry) depends on the RNA nucleotides being based on the right-handed version of the two mirror-image structures for the sugar (i.e. D-ribose rather than L-ribose). However, any prebiotic process would be expected to make the sugar as a racemic mixture (i.e. equal quantities of D-ribose and L-ribose).

Pre-RNA Worlds

The difficulties with the prebiotic synthesis of RNA lead many researchers to conclude that direct production of an RNA world is highly improbable, and there must have been a preceding stage, or pre-RNA world, that used some other genetic material. Many different suggestions have been made.

The most obvious precursor to RNA is some other organic polymer that carries information in the same way as DNA and RNA. One suggestion is the peptide nucleic acids (PNA) (**Figure 3**). These molecules use the same set of bases as DNA and RNA but have a much simpler backbone linked by peptide bonds. The prebiotic synthesis of PNA seems to be more straightforward than that of RNA. Furthermore, PNA is achiral, thereby avoiding the chirality problems associated with RNA.

Another possible RNA precursor is TNA. This is a nucleic acid in which the four-carbon sugar threose replaces the five-carbon sugar ribose used in RNA. TNA can base pair with RNA or DNA and is the structurally simplest member of the RNA family. The four-carbon sugar should be easier to make than ribose in prebiotic conditions.

Some other suggestions for pre-RNA worlds involve very different ideas. For example, it has been suggested that inorganic clay crystals, rather than organic polymers, provided the first genetic material. Another suggestion is the 'compositional genome', in which the genetic information is carried by a set of molecules

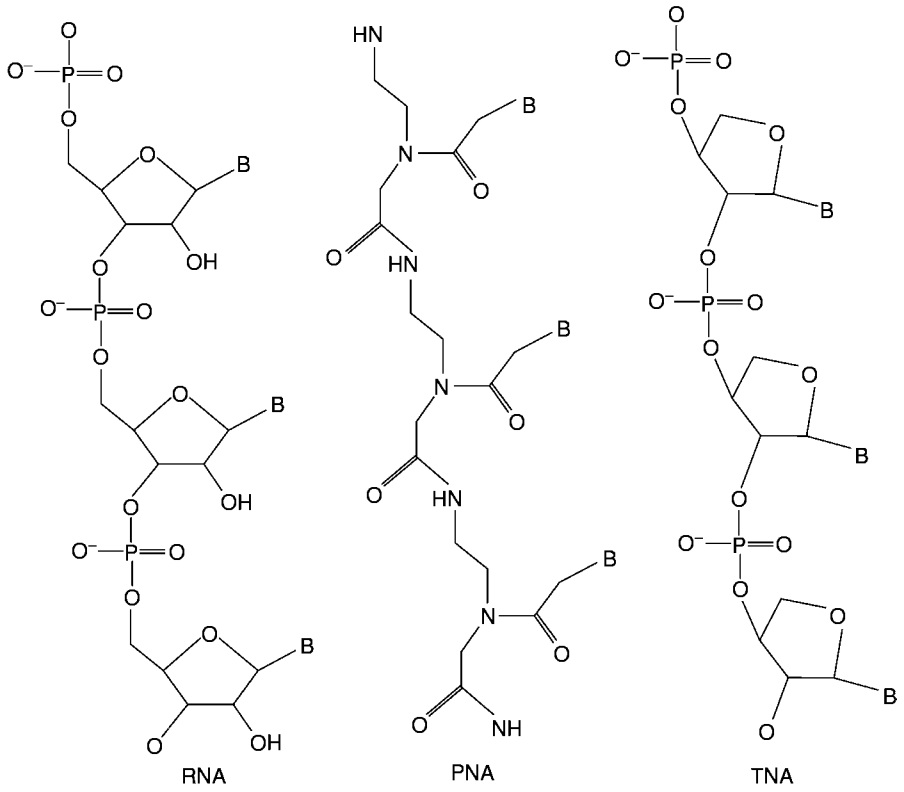


Figure 3 The molecular structures of RNA and two molecules that have been suggested as possible precursors of RNA. PNA (peptide nucleic acid) uses a simpler backbone based on peptide bonds, and TNA replaces the five carbon sugar ribose with the four carbon sugar threose. All three molecules can use the same set of four bases (attached at the positions marked B) and can base pair with RNA or DNA.

enclosed by a membrane. Other models propose that networks of reactions that are the forerunners of modern metabolic processes preceded the development of genes. All such models present a number of problems, not least in explaining how the transition to an RNA world could occur (Figure 4).

Suffice it to say that no model of a pre-RNA world has yet achieved any degree of consensus, and the problem of how to build an RNA world out of the available prebiotic material remains unsolved.

Sources of Prebiotic Organic Molecules

The Miller experiment, described above, provided the first demonstration of how organic molecules of biological relevance might be produced in the atmosphere of the primitive Earth. Similar experiments have produced a wide range of organic molecules in this way. However, to achieve a substantial yield, a strongly reducing atmosphere (i.e. one containing gases such as hydrogen, methane, and ammonia) is required. Current ideas suggest that the atmosphere of the early Earth was instead composed mostly of carbon dioxide and nitrogen (*see* **Atmosphere Evolu-**

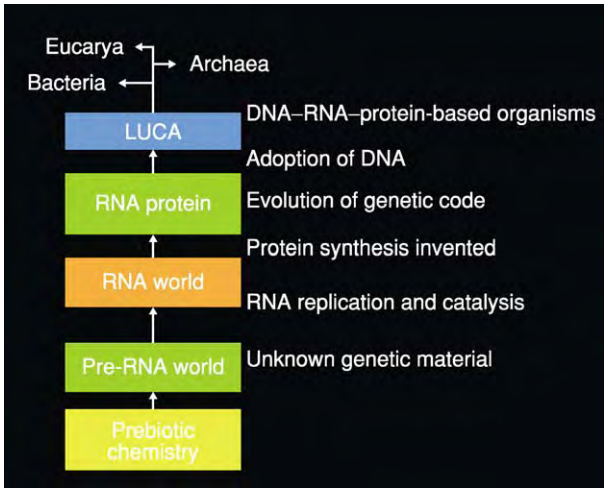


Figure 4 Outline of a possible series of steps in the development of life from prebiotic molecules up to the modern DNA RNA protein basis for biological chemistry.

tion). There may have been a small amount of methane (up to about 100 ppm), making the atmosphere, at best, weakly reducing and allowing the production of small amounts of organic material in this way.

A possible alternative source of organic molecules is extraterrestrial delivery. We know from astronomical observations that interstellar clouds contain at least 100 species of simple organic molecule, while experiments simulating interstellar conditions have shown that more complex molecules, including amino acids, might be produced by photochemistry in ice-coated dust grains. The Murchison meteorite, which fell in Murchison, Australia in 1969, has been found to contain a huge range of organic compounds including more than 70 different amino acids. During the phase of heavy bombardment in the first few hundred million years after the solar system was formed, the Earth would have been subject to frequent impacts of comets and asteroids (*see* **Earth Structure and Origins**). While the heat produced by the impacts of these large bodies may destroy much of the organic material, recent experiments and simulations suggest that at least some organic material may be able to survive these major impacts. Organic material can also be delivered to the Earth's surface by interplanetary dust particles. However, current estimates of the amount of organic material that might have been delivered to the early Earth vary widely.

One intriguing discovery is that at least some of the amino acids in the Murchison meteorite show a small excess of the L or left-handed form. This excess is found in amino acids not known to be present in the biosphere, so it cannot be due to contamination, and suggests that a chiral imbalance existed in the organic material of the early Solar System. This could be relevant to the origin of biological chirality. A possible cause of the original excess of L-amino acids is photolysis by circularly polarized light, which could have occurred in the star-formation region in which the Sun originated.

Another source that has been suggested for organic compounds is reactions occurring in the superheated water around deep-sea hydrothermal vents (*see* **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges). Reactions involving mineral catalysts could build organic molecules from simple molecules such as carbon dioxide. However, some workers argue that the high temperatures in hydrothermal vents are more likely to destroy organic molecules than create them. It has been calculated that virtually all the water in the oceans circulates through hydrothermal vents on a timescale of about 10^7 years, and the temperatures would be high enough to destroy molecules such as amino acids. This limits the amount of organic material that can build up in the oceans from any of the above sources.

Thus, while there are a number of possible sources of prebiotic organic compounds, we are still left with two significant problems. First, the total amount of organic material produced by any of these mechanisms seems to be enough to lead to only very weak concentrations in the ocean. Second, the molecules that are most easily made do not necessarily appear to be the most relevant. For example, amino acids are readily produced and are certainly important in modern biology, but the adoption of proteins seems to be a relatively late step in the development of life. The building blocks needed for RNA (such as ribose) are more difficult to make, although RNA appears to be essential for early life.

Where did Life Originate?

Given that we do not yet understand the chemical processes required for the early stages of life, it is hard to say much about the likely location. However, there are some general considerations that can be applied. The early Earth would have been subject to bombardment by asteroids at a high but rapidly reducing rate, and some of the largest impacts would have been 'sterilizing impacts' capable of vaporizing the oceans and destroying any life. Clearly life must have appeared after the last sterilizing impact, which might have been around 4–4.2 Ga ago. (There could, of course, have been one or more earlier origins of life that were wiped out in the impacts.) However, somewhat smaller impacts might have destroyed life near the surface while allowing it to be preserved deep in the oceans or within the Earth's crust. This suggests that life might have originated in a deep-ocean location or even deep within the crust. Another important consideration is that life requires an energy source (for modern life this is normally solar energy via photosynthesis). For a deep location the likely energy source is geothermal energy. A location on or near the surface would make available other energy sources such as ultraviolet radiation and lightning. A surface origin would, however, be much more exposed to impacts, and life would therefore have to have arisen somewhat later, towards the end of the heavy bombardment. This leaves a rather small time window between the end of the bombardment and the first evidence of life on Earth. However, the timings of both these events are somewhat uncertain.

Finally, the possibility must be considered that life did not originate on Earth at all. More than 20 meteorites have been identified as coming from Mars. The most famous is ALH84001, which has been sug-

gested to contain fossil evidence of Martian life. While many are unconvinced by these claims, the meteorite does provide evidence that material can be transferred from Mars to Earth with relatively good preservation, and suggests that something like a bacterial spore could seed life from one planet to another. There is evidence that early Mars had liquid water and may have been at least as favourable as Earth as a site for the origin of life. During the heavy-bombardment phase transfer of material from Mars to Earth would have been much more frequent than it is today.

Glossary

ALH84001 Martian meteorite discovered in the Allan Hills ice-field in Antarctica in 1984. It has been suggested that this meteorite contains evidence of Martian microbial life.

Amino acid Simple organic molecule containing an amino group (NH_2) and a carboxyl group (COOH). Amino acids link together in chains to form proteins.

Archaea One of three domains of life, the other two being the Bacteria and the Eucarya. The archaea are prokaryotic organisms, which were shown to be distinct from the bacteria by molecular phylogeny using 16S ribosomal RNA. The archaeal domain includes many extremophilic organisms.

Bacteria One of three domains of life, the other two being the Archaea and the Eucarya. The bacterial domain includes all prokaryotic organisms not classified as archaea.

Base pairing (Watson–Crick base pairing) The pairing of the DNA and RNA bases through chemical bonds, which link the double-helix structure of DNA and allow the copying of information. Adenine always pairs with thymine (or uracil), and cytosine pairs with guanine.

Bases The components of DNA and RNA that carry the genetic information in their sequence. DNA uses adenine, guanine, cytosine, and thymine. In RNA, thymine is replaced by uracil.

Chirality A chiral molecule is a molecule with an asymmetric structure that can exist in two mirror-image forms (or enantiomers). In living organisms such molecules are usually found in only one of the two possible enantiomers (homochirality). Thus, amino acids are normally in the left-handed or L-enantiomer, while sugars are in the right-handed or D-enantiomer.

Cyanobacteria A class of bacteria that make use of oxygen-producing photosynthesis; commonly referred to as blue-green algae.

DNA Deoxyribonucleic acid; the molecule that constitutes the genome of living cells. DNA molecules have a double-stranded helical structure built from a sugar–phosphate backbone and a set of four bases (adenine, guanine, cytosine, and thymine). The sequence of bases specifies the genetic information.

Domain The highest taxonomic division in the classification of living organisms. The three domains are the Archaea, the Bacteria, and the Eucarya. Domains are subdivided into kingdoms. While the three-domain model is widely used in astrobiology, some biologists prefer other schemes, such as the five-kingdom system.

Enzyme A protein that acts as a catalyst. Most chemical processes in living cells are enzyme catalysed.

Genetic code The set of rules by which three-letter 'words' in a DNA or RNA sequence describe an amino acid to be incorporated in a protein.

Genome The complete set of genetic information for a particular organism.

Heavy bombardment During the first few hundred million years after the formation of the solar system, the Earth and the other planets were subject to an intense bombardment by the debris left over from the formation of the solar system. It is during this phase of heavy bombardment that most of the craters on the Moon were formed. The emergence of life on Earth appears to coincide roughly with the end of the heavy bombardment.

Hyperthermophile An organism adapted to life at very high temperatures. Hyperthermophiles have optimum growth temperatures above 80°C, and a number can grow at temperatures above 100°C.

Interstellar molecule Molecules in interstellar space are most commonly detected by means of radio-frequency emission lines coming from the gas in molecular clouds. Molecules can also be detected from the infrared spectra emitted from dust. Well over a hundred molecular species have been detected by these methods. Some of the more complex molecules found include acetic acid, acetone, and ethanol.

Lateral gene transfer The transfer of genes between different species. Lateral gene transfer may have been widespread in the early stages of life on Earth, and this complicates the interpretation of the tree of life.

LUCA Last universal common ancestor; the last common ancestor of all living organisms.

Martian meteorites Meteorites that originate from Mars, also known as SNC meteorites. Their Martian origin is demonstrated by bubbles of gas trapped within them, which have identical composition to the atmosphere of Mars.

Miller–Urey experiment (or Miller experiment) An experiment carried out by Stanley Miller and Harold Urey in 1952, which demonstrated the synthesis of amino acids in conditions simulating a thunderstorm on the early Earth. The experiment can be seen as marking the beginning of the experimental study of the chemistry of life's origins.

Nucleic acid The molecules that carry genetic information – DNA and RNA.

Phylogeny The evolutionary relationships between different species of organism, represented in the form of a phylogenetic tree. In molecular phylogeny these relationships are determined by analysing the differences in the sequences of genes common to the various species.

Pre-RNA world A hypothetical early stage in the development of life, which preceded the RNA world and used some other genetic material in place of RNA or DNA.

Ribosomal RNA The RNA components of a ribosome. One of these components, the small subunit ribosomal RNA (also known as 16S ribosomal RNA in prokaryotes or 18S ribosomal RNA in eukaryotes), has been widely used to determine the tree of life.

Ribosome A structure composed of protein and RNA molecules that reads genetic information from messenger RNA and synthesizes the corresponding protein.

Ribozyme An RNA molecule that acts as a catalyst. The discovery of RNA catalysis led to a Nobel prize for Sidney Altman and Thomas Cech, and to the concept of the RNA world.

RNA Ribonucleic acid; a molecule that can carry genetic information in a similar way to DNA. In modern cells, RNA molecules are important in the process of protein synthesis, in the form of messenger RNA, ribosomal RNA, and transfer RNA.

RNA world A hypothetical early stage in the development of life, in which RNA molecules provided both the genome and the catalysts, roles that were subsequently taken over by DNA and proteins.

Stromatolites Layered structures built by colonies of micro-organisms that are commonly found in the Archaean and Proterozoic fossil records. Modern examples can be found in sites such as Shark Bay in Western Australia.

Tree of life A phylogenetic tree covering all groups of life on Earth. The term is commonly used for the tree derived by molecular phylogeny using small subunit ribosomal RNA, as pioneered by Carl Woese in the 1970s.

See Also

Atmosphere Evolution. **Biosediments and Biofilms**. **Earth Structure and Origins**. **Evolution**. **Famous Geologists:** Cuvier; Darwin. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils. **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges.

Further Reading

Bailey J (2001) Astronomical sources of circularly polarized light and the origin of homochirality. *Origins of Life and Evolution of the Biosphere* 31: 167 183.

Davies P (1998) *The Fifth Miracle*. London: Penguin Books.

De Duve C (1995) *Vital Dust*. New York: Basic Books.

Ferris JP (1998) Catalyzed RNA synthesis for the RNA world. In: Brack A (ed.) *The Molecular Origins of Life*, pp. 255 268. Cambridge: Cambridge University Press.

Kasting JF and Brown LL (1998) The early atmosphere as a source of biogenic compounds. In: Brack A (ed.) *The Molecular Origins of Life*, pp. 35 56. Cambridge: Cambridge University Press.

Miller S (1953) A production of amino acids under possible primitive Earth conditions. *Science* 117: 528 529.

Miller SL and Lazcano A (2002) Formation of the building blocks of life. In: Schopf JW (ed.) *Life's Origin*, pp. 78 112. Berkeley: University of California Press.

Orgel LE (2002) The origin of biological information. In: Schopf JW (ed.) *Life's Origin*, pp. 140 157. Berkeley: University of California Press.

Stetter K (1998) Hyperthermophiles and their possible role as ancestors of modern life. In: Brack A (ed.) *The Molecular Origins of Life*, pp. 315 335. Cambridge: Cambridge University Press.

Wills C and Bada J (2000) *The Spark of Life*. New York: Perseus Publishing.

Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. *Proceedings of the National Academy of Sciences USA* 87:4576 4579.

Yusupov MM, Yusupova GZ, and Baucom A (2001) Crys tal structure of the ribosome at 5.5Å resolution. *Science* 292: 883 896.

PALAEOCLIMATES

B W Sellwood, University of Reading, Reading, UK
P J Valdes, University of Bristol, Bristol, UK

Introduction

An evaluation of the climatic regime under which sedimentary successions accumulated has often been integral to their overall interpretation. It is an essentially uniformitarian approach. Palaeoclimatology, the understanding of past climatic regimes (palaeoclimates), has only relatively recently become a subject in its own right and is currently undergoing a major phase of expansion. One reason for this is the light thrown by palaeoclimatic studies on the natural variability that exists in the Earth's climate in the context of future, anthropogenically influenced, climate changes. Another is the realization that many of the Earth's sedimentary resources (e.g. coal, hydrocarbons, and other extractable minerals) formed under particular types of climatic regime.

Through time, the climate of the Earth has undergone many changes – the most recent ice age ended a mere 500 human generations ago. Astronomical data suggest that the Sun is of a type that steadily increases its heat output as it ages, increasing by about 1% per 100 Ma (*see* **Solar System:** The Sun). This presents a problem (the 'faint young sun paradox'). Simply stated, in its early days the Sun would have emitted about 45% less than its current heat output, so Earth should have been frozen, but there is ample evidence from Earth's most ancient rocks (e.g. in Greenland, South Africa, and Australia) to suggest that aqueous processes were operating. This paradox can be overcome if it is accepted that Earth's atmosphere was much denser and much richer in greenhouse gases (such as carbon dioxide and methane) during its early history. The rock record makes it clear that the Earth was periodically affected by major freezes during its early history, and controversy exists over the possibility that the Earth periodically froze from the poles to the equator (the 'snowball Earth hypothesis') during Proterozoic and older times (*see* **Atmosphere Evolution**).

Despite the possibility that the Sun was significantly cooler 500 Ma ago, abundant and widespread geological evidence suggests that through much of the subsequent time the Earth has been significantly warmer than at present. Since the Cambrian there appear to have been only three episodes that were cooler than now. One was in the Ordovician, when glaciation was centred on the present Sahara (*see* **Palaeozoic:** Ordovician), and one was in the Late Carboniferous (*see* **Palaeozoic:** Carboniferous) and Early Permian, with glaciation centred on southern Gondwana (*see* **Gondwanaland and Gondwana**). The latest 'ice ages' took place during the Quaternary (*see* **Tertiary To Present:** Pleistocene and The Ice Age), from about 1.67 Ma, but a growing body of evidence suggests that this latest phase had its origins at around the Eocene–Oligocene boundary (36 Ma), when ice began to accumulate over eastern Antarctica. For around 250 Ma, since the Late Permian, there appears to have been a period of extreme warmth, often referred to as the 'greenhouse world'. The Earth's past climate regimes can now be simulated using general circulation models (GCMs) similar to those used in long-term weather forecasting.

Orbital Forcing of Climate

In 1875, James Croll was the first to describe the cyclic way in which the Earth's orbit behaves. The eccentricity varies from near-circular to elliptical over a cycle of about 110 Ka and affects the total amount of solar radiation reaching the Earth. Changes in the tilt of the Earth's axis of rotation (obliquity; currently about 23.4° but varying between 21.8° and 24.4°), which affect the latitudinal distribution of received solar radiation, take around 40 Ka. In addition, a rotational wobble in the Earth's axis of rotation (as if the poles describe a circle against the background of the stars) causes the precession of the equinoxes and takes around 22 Ka to complete a cycle. In the 1930s and 1940s, Milutin Milankovitch, improving on Croll's work, realized that each of these cycles interacts with the others to cause changes in the amount of insolation received by the Earth that are sufficient to promote climate changes (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). These cycles are now named after Milankovitch and have been used to predict the time-frame of glacial cycles in the Quaternary. Changes in palaeoclimate have been recognized in Neogene oceanic sediments, which provide a metronomic time signal, synchronized with the Milankovitch time-frame, and hence a means of short-term correlation that was unavailable until recently.

Geological Proxies of Palaeoclimate (Figures 1A and 1B)

Sedimentary facies and palaeontological data generally provide qualitative evidence of palaeoclimates

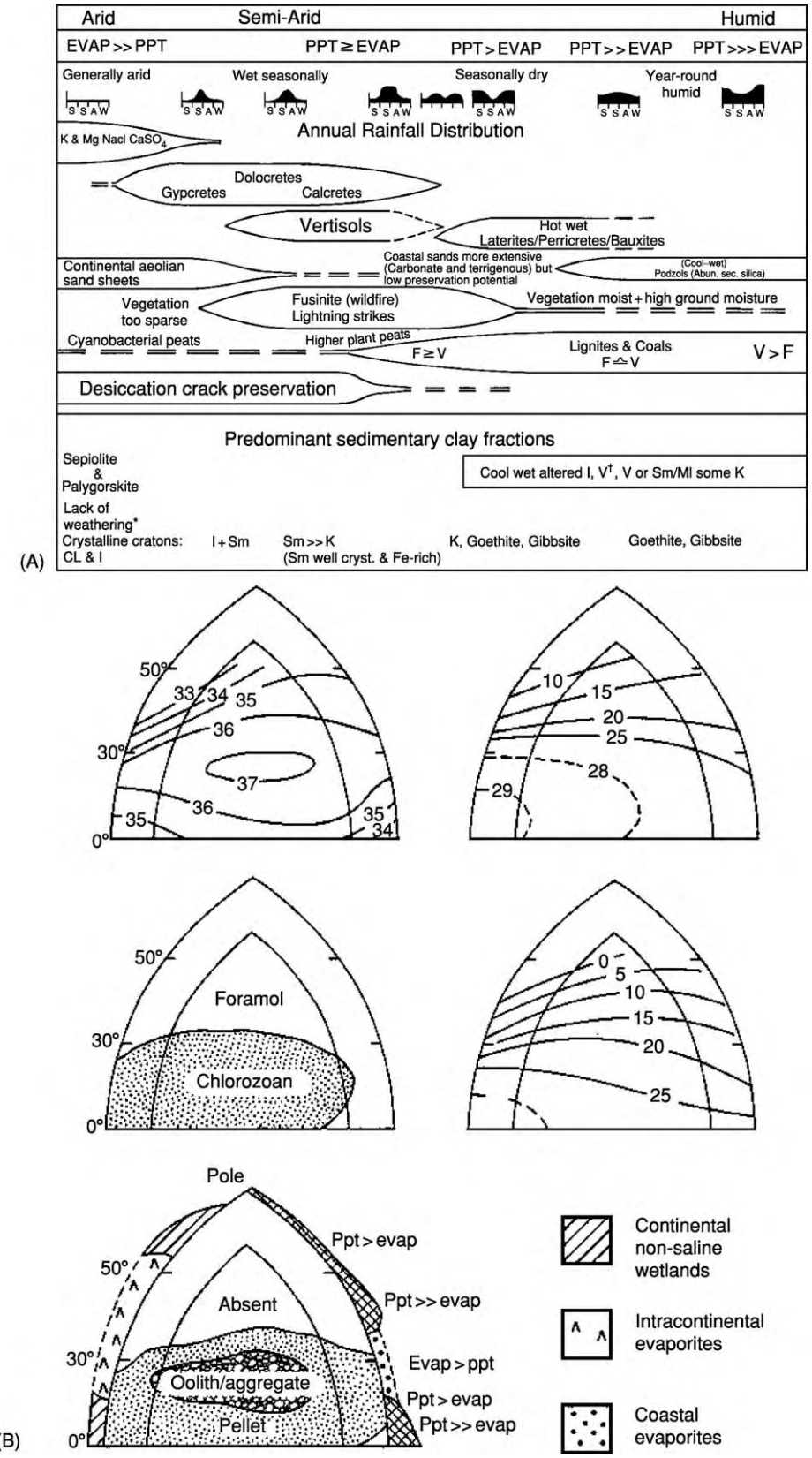


Figure 1 Geological palaeoclimate proxies for (A) terrestrial environments and (B) marine carbonates and evaporites (top left, salinity (‰); top right, maximum temperature (°C); middle left, skeletal associations; middle right, minimum temperature (°C) Lower left, predicted modern carbonate and evaporite distributions. Cross hatching shows maritime latitudes where precipitation exceeds

and climate changes. These are palaeoclimate proxies. The most climatically informative structures are glaciogenic features (striated pavements, tills, and tillites), distinctive soil types (e.g. laterites and bauxites), and indicators of aridity (e.g. evaporites and aeolianites). Supplementary evidence of climate is provided by a range of other features, such as other types of palaeosol (e.g. calcretes, gypcretes, and vertisols), clay mineralogy (reflecting weathering regime), storm deposits, reefs and oolites, glendonites, and fusain (from wildfires). Coals, often considered to require moist tropical climates, are now known to accumulate equally well in temperate mires. Critical information on past climates (particularly quantitative information on palaeotemperature) comes from oxygen isotopic ($\delta^{18}O$) data derived from carbonate fossils, but such data have to be treated with care because of problems associated with diagenetic resetting. Palaeobotanical data, particularly regarding plant appearance (physiognomy), are generally qualitative, and have also been claimed to provide quantitative information. The density of leaf stomata may be related to the concentration of atmospheric carbon dioxide, and there is, today, a striking relationship between mean annual temperature and leaf shape (percentage of entire-margined, as opposed to toothed-margined, species) in a biome. Entire-margined leaves dominate warm biomes.

Fossil reptiles such as crocodilians, like their modern descendants, have global distributions that are principally controlled by temperature. A coldest-month mean temperature of 5.5°C marks the minimum thermal limit for the group (corresponding to a modern mean annual temperature of *ca*. 14.2°C). Such evaluations of temperature have been successfully applied globally for the Mesozoic and Cenozoic and compared with GCM output on climate for those times.

Marine Carbonates

Modern shelf seas are dominated by carbonate facies in areas where organic productivity is high, generally in low latitudes (between about 30°N and 30°S) with strong insolation and warm normal-salinity seawaters (>20°C and 35‰ salinity). Carbonate shelves also occur in temperate waters, particularly where rates of terrigenous run-off are low. However, rates of productivity in these areas are much lower than in warmer waters, and the biota is dominated by different species. Nonetheless, dominance of carbonates in a particular rock succession may not necessarily indicate warm waters, particularly for more ancient rocks in which the biota have only tenuous links with modern forms (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves).

Present-day warm-water shelves unpolluted by fine-grained terrigenous run-off have diverse communities with hermatypic (reef-building) corals and codiacian calcareous algae and may include ooids, aggregates, and pellets; these latter grain types reflect carbonate precipitation and early cementation. This is known as the chlorozoan skeletal grain association and can be recognized as far back as the Triassic. Modern temperate-water carbonate factories are dominated by benthic foraminiferans, molluscs, bryozoans, barnacles, and calcareous red algae. This is termed the foramol skeletal grain association and can be recognized, albeit equivocally, into the Mesozoic Era. The polewards expansion of reef and carbonate facies (by about 10° latitude) in the Mesozoic and Early Cenozoic, relative to today (30° today, *ca*. 40° in the Mesozoic and Palaeogene), suggests that the Earth was more equable during these times.

The principles of palaeotemperature determination from carbonates using oxygen isotope ratios are well established, and for many years oxygen isotopic ratios have been used to quantify sea-surface temperatures and bottom-water temperatures. The technique has most successfully been applied using the fossil shells of organisms considered to be robust to diagenetic changes (e.g. planktonic forams, coccolithic oozes). Such data have shown that ocean waters have changed their isotopic ratios in concert with the waxing and waning of the ice-caps during Quaternary glacial cycles, the oceans becoming relatively depleted in ^{16}O during glacial episodes (shells therefore showing ^{18}O enrichment) by comparison with interglacials. In the longer term the $\delta^{18}O$ trend for ocean waters suggests that seawater has become progressively depleted in ^{16}O (leading to positive $\delta^{18}O$ values) since about 55 Ma. This has occurred in a series of steps, probably reflecting, first, the cooling of deep waters, followed by the onset of permanent ice on Antarctica at around the Eocene–Oligocene boundary. Then followed a series of significant more positive excursions, through the Oligocene and Miocene, before a major and permanent swing in the Miocene (15 Ma), which is taken to reflect the formation of the Antarctic ice-sheet. The growth of the northern hemisphere ice-sheets is

evaporation). Under lignites and coals V, vitrinite; F, fusinite; I, illite; V, vermiculite; Sm, smectite; K, kaolinite; CL, chlorite; ppt, precipitation; evap, evaporation; *, unaltered parent rock with fines under both arid and frigid terrains; †, vermiculite can form under a wide variety of conditions.

marked by a further more positive excursion in the Quaternary.

Evaporites

Evaporite salts may form anywhere on Earth where evaporation exceeds rainfall (*see* **Sedimentary Rocks:** Evaporites). Their preservation is controlled by the nature of the post-depositional climate, the ingress of groundwaters, and subsequent burial history. Modern coastal evaporites accumulate between 15° and 35° latitude, but intracontinental evaporites may occur beyond 50°, sometimes being frozen for several months in the year. So salts do not necessarily imply permanent warmth. The formation and preservation of the more soluble salts of potassium require low atmospheric humidity (less than 50%). Salts may form very rapidly under ideal conditions (100 m in 1 Ka), so short-lived episodes of evaporation may lead to great thickness of salts.

Glaciogenic and Cold Water Sediments

About 10% of the Earth is covered by ice today, but this increased to 30% during the Quaternary glacials. Ice and permafrost provide a series of erosional and depositional fingerprints (e.g. striated pavements and boulders, dropstones, ice wedges, and glacial tills (diamicton)) (*see* **Sedimentary Processes:** Glaciers) as well as a range of geomorphological features, such as eskers, kames, etc., some of which have been recognized in association with ancient glacial systems (e.g. in the Lower Palaeozoic of the Sahara and the Late Palaeozoic of Gondwana). However, boulders may be dropped into deep water through mechanisms other than ice rafting (e.g. from the roots of floating trees). More cryptic evidence of cold-water precipitation comes from glendonites – carbonate nodules pseudomorphing the mineral ikaite, the low-temperature polymorph of calcium carbonate ($CaCO_3 \cdot 6H_2O$). It is today associated with clathrates in cold deep waters and in shallow polar waters. Glendonites are reported from Cretaceous localities in Australia and Canada and from the Jurassic of Siberia.

Coals and Lignites

Coals and lignites are indicative of an excess of precipitation over evaporation: vegetation accumulates in wetland mires, which, if rain occurs on most days of the year, will grow upwards above the general level of the regional water table. Preservation of peat requires rapid burial and/or anoxic conditions. Frigid conditions favour preservation, as do environments in which rainfall is evenly spread throughout the year. Today such conditions occur in equatorial zones and in high mid-latitudes. In the Mesozoic equatorial zones appear to have been more arid than they are today.

Wildfires and Palaeosols

Fusain, almost pure carbon in sedimentary rocks, is a product of wildfires. Natural fires, mostly triggered by lightning strikes, are favoured in strongly seasonal climates where vigorous plant growth is promoted by a distinct wet season, followed by a dry season creating tinder that becomes fire-prone during thunder storms at the onset of the next wet season. Fusain-rich sediments are frequently associated with other evidence of strong seasonality (distinct wet–dry seasons) such as vertisols (and associated smectite-rich clay mineralogies) and calcretes.

Laterites (or ferricretes) reflect the relative accumulation of haematite, goethite, aluminium hydroxides (e.g. gibbsite), and kaolinite within soil profiles (*see* **Soils:** Palaeosols). Pedogenic and groundwater laterites form best in humid tropical climates with a distinct dry season. Bauxites (aluminium-rich laterites) often occur in association with laterites. Some, like those of the Cretaceous of the Tethyan region, occur on karstic surfaces cut into carbonate platforms (*see* **Sedimentary Processes:** Karst and Palaeokarst). In such cases the detrital materials are likely to have been brought into the system, probably by aeolian activity.

Clay minerals in oceanic sediments reflect the weathering processes in the adjacent continental areas and the climatically controlled distribution of soils. Differential settling causes some depositional segregation of clay species (e.g. near-shore accumulation of kaolinite); nonetheless, recognizable patterns in the ancient can be preserved (*see* **Clay Minerals**).

Aeolianites

Although coastal dune complexes may form across a range of latitudes, major sand seas occur predominantly in arid environments. Such aeolianites provide both local and regional palaeowind data, which may be used to test palaeoclimate models generated by GCMs (see below). (*see* **Sedimentary Processes:** Aeolian Processes)

Palaeoclimate Models

A recent approach to understanding past climate regimes on Earth has been through the application of complex computer models, specifically atmospheric general circulation models (AGCMs), ocean general circulation models (OGCMs), and recently even more complex coupled ocean–atmosphere general circulation models (OAGCMs). There are now many contributions in this field, the palaeoclimatic

approach having been pioneered by John Kutzbach and Eric Barron.

General Circulation Models

GCMs use the laws of physics and an understanding of past geography to simulate climatic responses. They are objective in character. They require powerful computers to handle vast numbers of calculations. Nevertheless, it is now possible to compare the results of different GCMs for a range of times and over a wide range of parameterizations for the past, present, and future (e.g. in terms of predictions of surface air temperature, surface moisture, precipitation, etc.). GCMs are currently producing simulated climate predictions for the Mesozoic and Cenozoic that compare favourably with the distributions of the geological climate proxies discussed above. They can be used effectively to predict sites of oceanic upwelling and the distribution of petroleum source rocks and phosphorites. Models also produce evaluations of parameters that do not leave a geological record (e.g. cloud cover, snow cover) and quasi-parametric phenomena such as storminess. Parameterization is the main weakness of GCMs (e.g. palaeogeography, palaeobathymetry, sea-surface temperature, orography, cloud behaviour), and model output for continental interiors is still colder in winter than indicated by palaeontological data. The sedimentary and palaeontological record provides an important way of evaluating GCMs, and this is important because the same GCMs are currently being used to predict possible changes in future climate.

The outputs discussed below were generated by an AGCM (HadAM3) and an OAGCM (HadCM3L), which is currently state-of-the-art. The model was developed at the Hadley Centre for Climate Prediction and Research, which is part of the UK Meteorological Office. The GCM consists of a linked atmospheric model, ocean model, and sea-ice model. The horizontal resolution of the atmospheric model is 2.5° latitude and 3.75° longitude. This provides a grid spacing at the equator of 278 km north–south and 417 km east–west. The atmospheric model consists of 19 layers. It also includes a radiation scheme that can represent the effects of minor trace gases. Its land-surface scheme includes a representation of the freezing and melting of soil moisture. The representation of evaporation includes the dependence of stomatal resistance on temperature, vapour pressure, and carbon dioxide concentration. There is an adiabatic diffusion scheme, to simulate the horizontal mixing of tracers.

The ocean model has the same spatial resolution as the atmosphere model and 20 vertical layers, with a time step of 30 min. This contrasts with HadCM3 (the standard version of the Hadley centre OAGCM), which uses a horizontal resolution of $1.25° \times 1.25°$.

Palaeoclimate of the Mesozoic – Model Output and Geological Data

The Mesozoic Earth was an alien world, as illustrated here by reference to a Triassic GCM simulation and geological data. Throughout the Mesozoic dense forests grew close to both poles, experiencing months of continual daylight in warm summers and months of continual darkness in cold snowy winters. Neither Triassic nor Jurassic oceanic sediments provide good evidence, but in the Late Cretaceous, from ODP (Ocean Drilling Program)/DSDP (Deep Sea Drilling Program) data, the ocean depths appear to have been warm (8°C or more at the ocean floor), and reefs grew 10° further north and south than at the present time. During the era the whole Earth was warmer than now by at least 6 °C, giving more atmospheric humidity and a greatly enhanced hydrological cycle. However, from modelling studies, it seems that much of the rainfall could have been predominantly convective in character, often focused over the oceans. The model output might help to explain geological data suggesting that major desert expanses extended across the continents in low latitudes. From the model, polar ice sheets are unlikely to have been present because of the high summer temperatures. The model suggests the possibility of extensive sea ice in the nearly enclosed Arctic seaway through parts of the year, but there is as yet no proxy data against which such predictions may be tested. The Triassic world was a predominantly warm world; the model outputs for evaporation and precipitation conform well to the known distributions of evaporites, calcretes, and other climatically sensitive facies.

Triassic: Comparison of Model and Proxy Data – A Case Study

Modelled temperatures (Figure 2) A significant feature of the Triassic Earth is that the landmasses were almost symmetrically distributed in a broad arc about the equator (*see* **Mesozoic:** Triassic). A major aspect of the modelled Earth is its overall warmth. Despite temperatures plunging to −20°C and below over Siberia in the northern-hemisphere winter and to similarly low values over southernmost Gondwana in the southern-hemisphere winter, the annual average temperature is subdued in these high-latitude areas because of the high summer values achieved there (*ca.* 24°C). These high summer values preclude the possibility of year-round ice and snow.

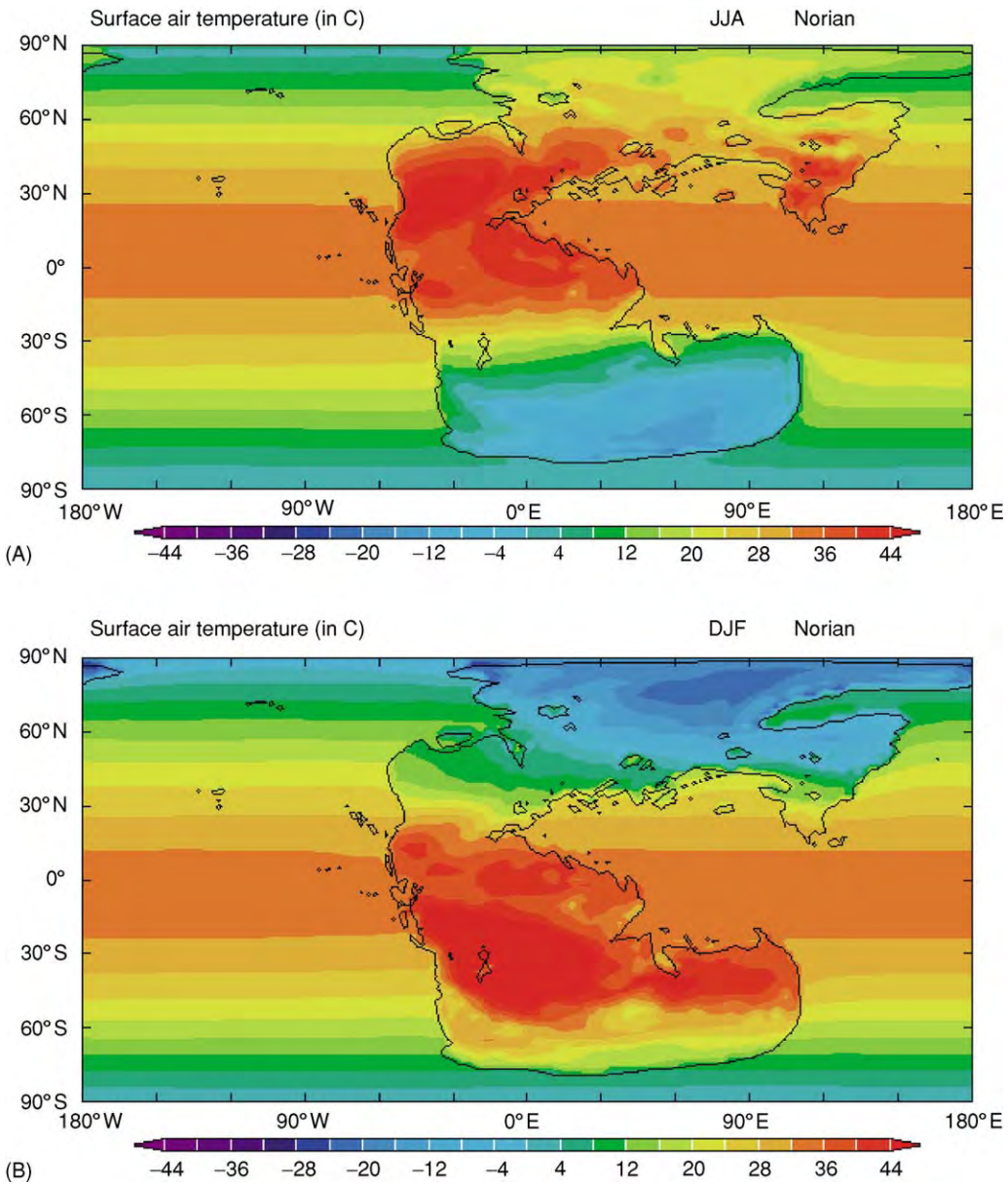


Figure 2 Model simulated mean seasonal temperatures for the Norian for (A) December February and (B) June August. Units are °C and the contour interval is every 4°C. The ocean sea surface temperatures were prescribed. JJA, June, July, August; DJF, December, January, February.

Temperature-limited facies According to the model, low temperatures would not have been a significant barrier to coral reef development around the Tethys. Reef systems developed and diversified through the Triassic, extending their range from between 2°S and 25°N in the mid-Triassic to a maximum range of 35°S to 33°N in the Late Triassic. Range limits are very close to the 20°C isotherm in the model.

Precipitation (Figure 3) Large tracts of Pangaea between about 40°N and 40°S are modelled to receive very little rainfall. Much of the rainfall is over the oceans, being convective in character, with the main zone of rainfall migrating north and south through the year with the movement of Intertropical convergence zone (ITCZ). The equatorial lands surrounding eastern and southern Tethys are modelled to receive relatively little rainfall during December–April. North-eastern Gondwana is modelled to have a monsoonal climate (winter wet). Northern and eastern Siberia have some rainfall throughout the year, the wettest period generally being through June and the

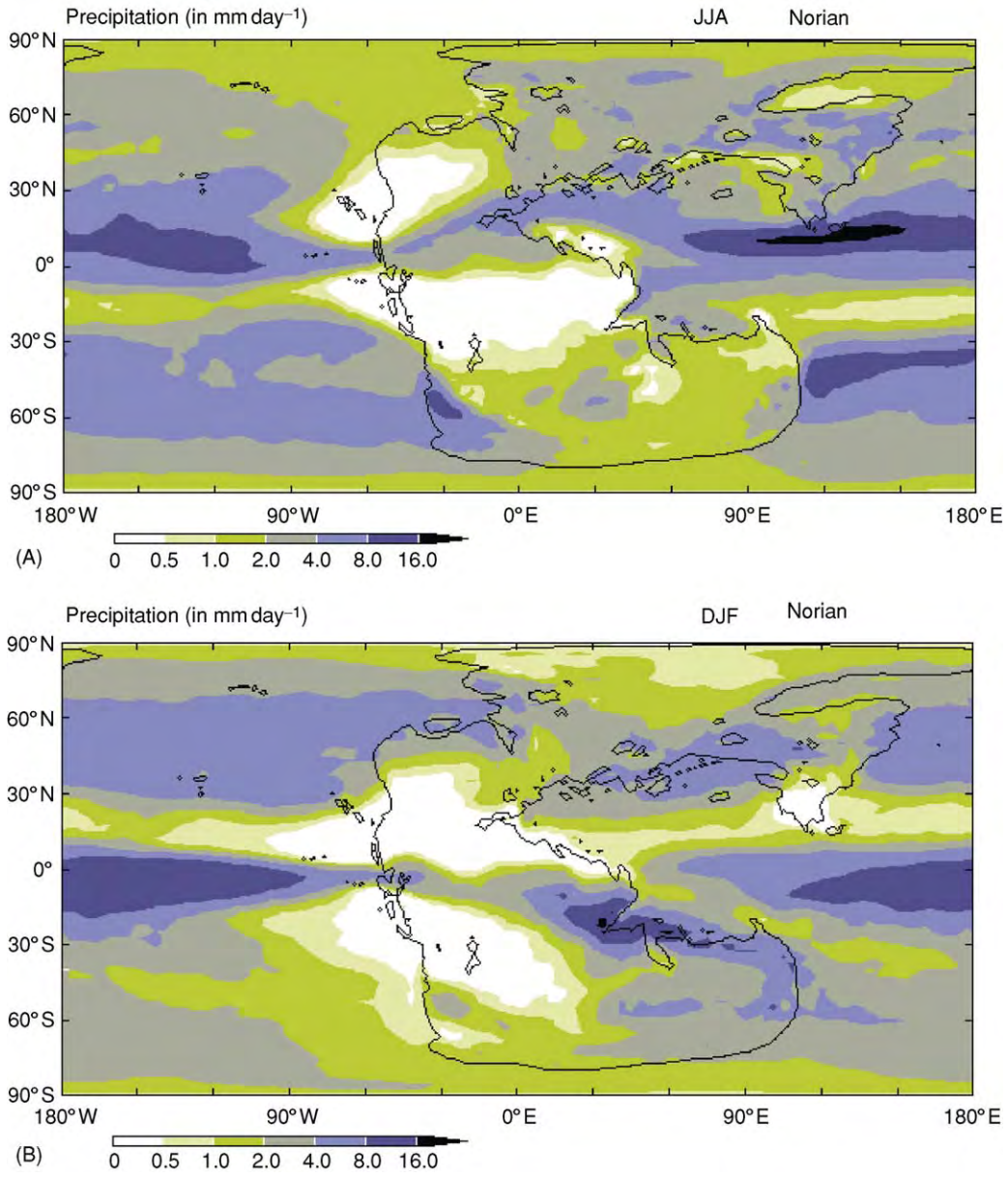


Figure 3 Model simulated mean seasonal precipitation for the Norian for (A) December February and (B) June August. Units are in mm day ^{1} and the contour interval is not regular. JJA, June, July, August; DJF, December, January, February.

driest during the summer months of July and August (also equivalent to a winter-wet biome). South-western Gondwana is also modelled to be winter wet; the eastern parts of Gondwana are moister than the western parts. Nonetheless, large tracts of Pangaea, particularly in the tropics but extending to nearly 50°N and 50°S, have a large excess of evaporation throughout much of the year, as does a large part of western Tethys. Southernmost Gondwana is also winter wet, including the southern polar area during its months of darkness.

In the Triassic (Figure 4), the following climate zones have yet to be recognized from geological data: ice; polar; cold temperate; and cold temperate arid (steppe/desert). Tillites have not been reported. An absence of both ice caps and several of these biomes would be expected from the model simulations, but with some caveats as the biomes are modelled by reference to modern equivalents (e.g. cold temperate).

In the Triassic the highest-latitude floras contain a lycopod that has a worldwide distribution, reflecting a global climate largely devoid of frost. From

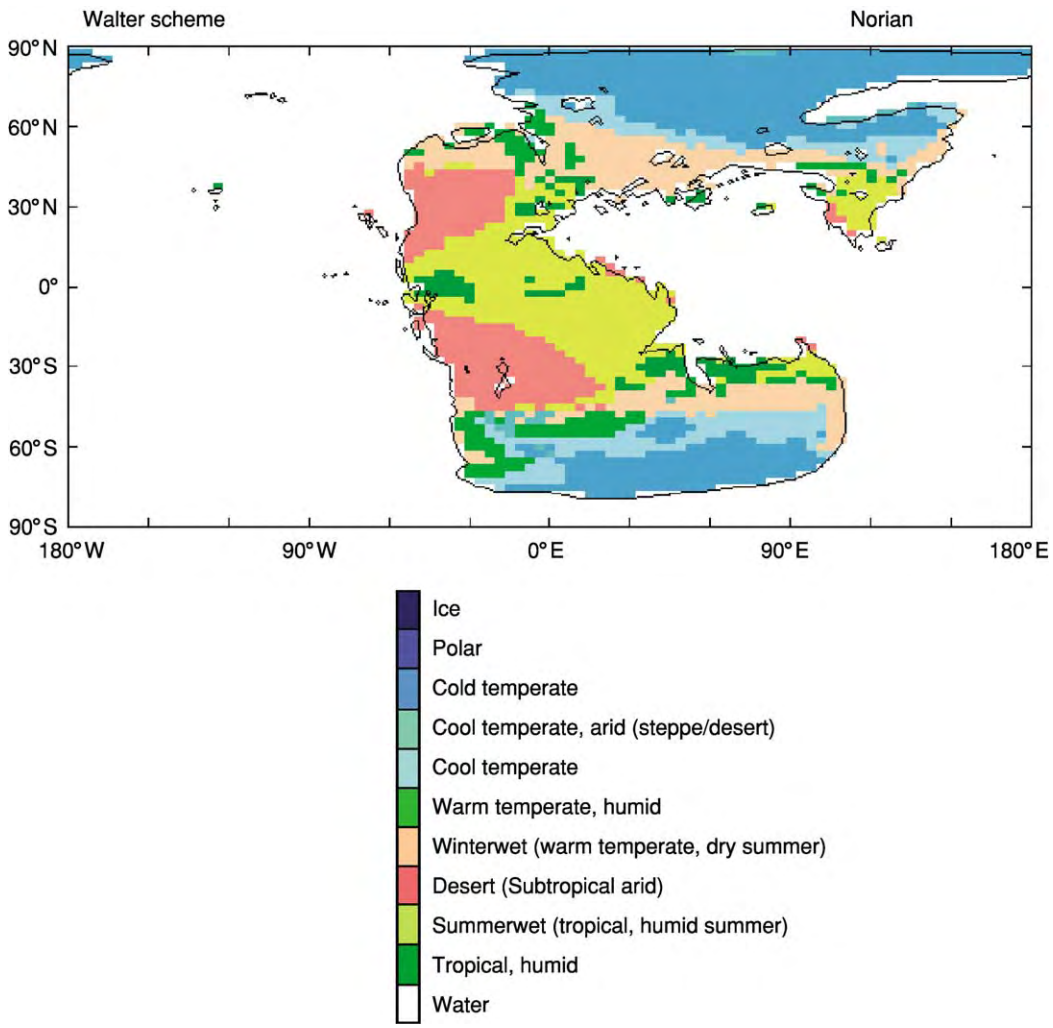


Figure 4 The Walter biome zones for the Triassic based on the model predicted temperatures and precipitation.

geological proxy data the interior of Pangaea was hot and dry, with warm temperate climates extending to the poles. In Siberia the flora was deciduous and displays well-defined annual growth bands; the climate is interpreted to be cool temperate, with cool winters and warm summers and a growing season of four to eight months. For this biome to extend to 75° in both hemispheres, as is observed, the poles must have been far warmer than at present. This is consistent with modelled output. The distribution of extensive calcretes throughout Europe and the western Tethyan margin is in agreement with the modelled climate: hot and predominantly arid but with a short wet season.

Relatively low-diversity communities of conifers were confined to lower latitudes and are often found in the proximity of evaporites (usually found in both hemispheres between about 10° and 40° latitude). They thus represent the subtropical dry environment, and their distributions are closely in line with model predictions in terms of the ratio between precipitation and evaporation.

On the basis of facies and palaeobotanical data, warm temperate climates existed between 5° and the equator, with high-diversity floras of ferns, cycads, and seed ferns, as well as conifers and ginkgoes. The cool temperate zone extended to central Siberia and is associated with the postulated uplands of the Central Siberian Volcanic Traps.

During the Middle Triassic, abundant evaporites and calcretes were deposited in Europe, whereas in northern and southern China warm temperate and cool temperate floras predominated. These observations are fully in agreement with model output. Warm temperate floras extended to about 70° N in north-eastern Siberia, and the first northern-hemisphere coal swamps since the end-Permian appear in this area. Southern-hemisphere coal swamps are found in

South America and Antarctica, again compatible with modelled zones where precipitation exceeded evaporation for much of the year. Biomes and facies suggest that the equatorial zone is a predominantly arid belt, again lending support to the validity of the model. Exquisitely preserved fossil wood from the Middle Triassic of Antarctica (70–75° S in the Triassic) indicates forests with growth rates that were one to two orders of magnitude greater than those seen at such latitudes today.

In the Late Triassic the warm temperate biome was dominant between 30° and 50° of latitude in both hemispheres. The model generally supports this scenario, but lack of precipitation typifies many regions between these latitudes, especially away from the coastal zones.

Thus, the model output shows a good correlation with globally derived data, particularly for terrestrial plant biomes, terrigenous proxy facies, and the distribution of reefs. Terrestrial reptile distributions are difficult to relate directly to palaeoclimate, but can be interpreted to be broadly in line with model output. This illustrates the sort of model–proxy correlation that is currently possible.

Conclusions

Model simulations for the Mesozoic throw important light on climatic processes and the behaviour of Earth systems under greenhouse conditions. Model output generally compares favourably with climate proxy data, but discrepancies between data and models reveal some serious modelling problems (or possibly incorrect interpretation of the proxies). In continental interiors, model simulations generate winter conditions that are considered to be too cold (by 15°C or more) according to the interpretation of some palaeontological data.

Ocean temperatures could have been much warmer than today, not only at the surface but also at great depths. Such temperatures would have greatly reduced the oceanic storage capacity of both carbon dioxide (in ocean waters) and methane hydrates (in ocean-floor sediments) relative to the present day. Model output is in close agreement with proxy climate data in coastal and open-sea areas.

In its 'greenhouse' mode the Earth's palaeoclimate has greatly enhanced evaporation and precipitation, by comparison with the present, but much of the rainfall is convective in character, falling over the oceans. To replicate climatic conditions similar to those indicated by proxy facies throughout the Mesozoic a very large increase in atmospheric carbon dioxide is required (at least four times the present-day values).

Through the Phanerozoic, the Earth's climate has changed significantly, both on a variety of time-scales and over a range of climatic states, usually baldly referred to as 'greenhouse' and 'icehouse', although these terms disguise more subtle states between these extremes.

See Also

Atmosphere Evolution. **Clay Minerals**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Gondwanaland and Gondwana**. **Mesozoic:** Triassic. **Palaeozoic:** Ordovician; Carboniferous. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Processes:** Aeolian Processes; Glaciers; Karst and Palaeokarst. **Sedimentary Rocks:** Evaporites. **Soils:** Palaeosols. **Solar System:** The Sun. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Allen JRL, Hoskins BJ, Sellwood BW, Spicer RS, and Valdes PJ (eds.) (1994) *Palaeoclimates and Their Modelling: With Special Reference to the Mesozoic Era*. London: Chapman and Hall.

Barron EJ (1987) Eocene equator to pole surface ocean temperatures: a significant climate problem. *Palaeocli matology, Palaeoecology, Palaeogeography* 2: 729 740.

Billups K and Schrag DP (2002) Paleotemperatures and ice volume of the past 27 Myr revisited with paired Mg/Ca and $^{18}O/^{16}O$ measurements from benthic foraminifera. *Paleoceanography* 17:10.1029/2000PA000567.

Crowley TJ and North GR (1991) *Palaeoclimatology*. Oxford: Oxford University Press.

D'Hondt S and Arthur MA (2002) Deep water in the late Maastrichtian ocean. *Paleoceanography* 17:10.1029/ 1999PA000486.

Frakes LA, Francis JE, and Syktus JI (1992) *Climate Modes of the Phanerozoic*. Cambridge: Cambridge University Press.

Haywood AM, Valdes PJ, Sellwood BW, and Kaplan JO (2002) Antarctic climate during the middle Pliocene: model sensitivity to ice sheet variation. *Palaeogeography, Palaeoclimatology, Palaeoecology* 182: 93 115.

Huber BT, MacLeod KG, and Wing ST (eds.) (2000) *Warm Climates in Earth History*. Cambridge: Cambridge Uni versity Press.

Kiessling W, Flügel E, and Golonka J (eds.) (2002) *Phaner ozoic Reef Patterns*. SEPM Special Publication 40072. Tulsa: Society for Sedimentary Geology.

Kutzbach JE and Gallimore RG (1989) Pangean climates megamonsoons of the megacontinent. *Journal of Geo physical Research Atmospheres* 94: 3341 3357.

Markwick PJ (1998) Fossil crocodilians as indicators of Late Cretaceous and Cenozoic climates: implications for using palaeontological data in reconstructing palaeo climate. *Palaeogeography, Palaeoclimatology, Palaeoe cology* 137: 205 271.

Price GD, Sellwood BW, and Valdes PJ (1995) Sedimento logical evaluation of general circulation model simula tions for the 'greenhouse' Earth: Cretaceous and Jurassic case studies. *Sedimentary Geology* 100: 159 180.
Ruddiman WF (2000) *Earth's Climate: Past and Future*. New York: WH Freeman and Co.
Scotese CR (2000) Paleomap Project. http://www.scotese.com.
Valdes PJ, Spicer RA, Sellwood BW, and Palmer DC (1999) *Understanding Past Climates: Modelling Ancient Weather*. CD ROM. OPA (Overseas Publishers Association), Amsterdam.

PALAEOECOLOGY

E M Harper, University of Cambridge, Cambridge, UK
S Rigby, University of Edinburgh, Edinburgh, UK

Introduction

Ecology is the scientific study of the way in which modern organisms interact with their environment. In practice, ecologists study a very wide range of questions, but there are two main branches of study, autecology and synecology. Autecology is concerned with the ecology of an individual taxon, such as a cheetah, blue mussel, or vestimentiferan worm. Questions tackled might concern mode of life, feeding and breeding habits, and physical and chemical factors that limit distribution. Synecology examines the complex interactions between populations of a single species or between communities of organisms; examples of synecological studies might concern the relationships between a variety of big cats and their prey on the plains of Africa, the organisms associated with a mussel patch on a rocky shore, or the chemosymbiotic relationships within deep-sea vent communities, examining topics such as trophic structure and diversity. Palaeoecologists seek to gain similar understandings about fossilized organisms and ancient communities.

Palaeoecological studies may have relevance outside the sphere of palaeobiology. Organisms are not randomly distributed and are adapted to particular sets of environmental conditions (e.g., salinity, light levels, substrate consistency, and temperature). Once it has been established that a particular fossil taxon had a particular set of habitat requirements, the presence or absence of that taxon may be used in a broader geological context as a palaeoenvironmental indicator. For example, stenohaline organisms such as echinoderms and brachiopods, which have no ability to excrete water gained by osmosis, are limited to normal marine salinity, thus the presence of the body fossils of these higher taxa has been widely used as an indicator of prevailing normal salinity. For other groups, a detailed knowledge of the palaeoecologies of their component genera or species has been invaluable in finely subdividing palaeoenvironments. For example, the relative proportions of milioline, rotaliine, and textularine foraminifera in sediments give important information concerning the salinity of a depositional environment from brackish to hypersalinity. Brackish waters are dominated by rotaliinids and textularinids, whereas miliolinids and rotaliinids make up the dominant species in areas of normal or elevated salinity.

In terrestrial sediments, details of the palaeoecology of plants, such as the relationship between leaf shape and temperature and between concentration of stomata and atmospheric carbon dioxide levels, may be used in palaeoenvironmental reconstruction. Most recently, there has been an added interest in palaeoecology because the study of how the ecologies of different organismal groups and communities have altered during periods of dramatic environmental change (for example, during glacial periods or after mass extinctions) may give useful clues to how life on Earth may react to future climatic perturbations linked to human activities.

Palaeoautecology

The palaeoautecology of an extinct organism may be taken broadly to equate to its mode of life. Of course, there are aspects of fossil palaeoautecologies for which there is very little hope of being able to reconstruct; for example, it is difficult to imagine that anything will ever be known of mating dances in extinct flies, but a remarkable amount can be gleaned. Arguments made from homology with living relatives and analogy from organisms with apparently similar adaptations, from trace fossils or from models of fossils, can be used. It is almost invariably preferable to use a combination of approaches in order to allow cross-referencing.

When fossil organisms belong to taxonomic groups that still have extant representatives, much can be learned by applying knowledge about living taxa, using principles of homology. Clearly this is likely to

be most successful in groups which are most closely related, and as this relationship is stretched, any conclusions that extant and fossil groups may have lived in the same way are tenuous. Interpretation of the modes of life of most fossil bivalve molluscs is relatively straightforward. Several features of bivalves make these organisms ideal for palaeoecological reconstruction. First, modern bivalves are more diverse, both in terms of taxonomic units and in life habits, than ever before in their 500-million-year evolutionary history, meaning that interpretations can be based on a wide range of possible models. Second, a wide variety (for example, oysters, scallops, mussels, and clams) are of great commercial interest and consequently their biologies have been studied intensively. Third, it is well demonstrated that bivalve shell morphology and life habit are intimately related (*see* **Fossil Invertebrates:** Bivalves).

An example of this methodology is illustrated in Figure 1, which shows the right valve of *Eopecten*, a large bivalve known from Middle and Upper Jurassic hardground faunas of Normandy (France) and Portugal. Several aspects of its morphology can be used to interpret the palaeoecology of this animal, based on comparison with a living bivalve of the same family. The oldest part of the shell, close to the umbo, is a regular scallop shape and shows the morphological features associated with attachment to a hard substrate by byssal threads spun by the foot. The surrounding section (later growth) of the shell is highly irregular and its form varies among individuals, because at this size the bivalve began to cement to the seafloor and modified its shape to suit the substrate. Experimental evidence has shown that cementation in bivalves is a good defence against predators.

Clearly some caution is required in using modern organisms to interpret the palaeoecology of extinct fossils, even when they are closely related. Some organisms have undergone a major shift in aspects of their life habits over their evolutionary history. For

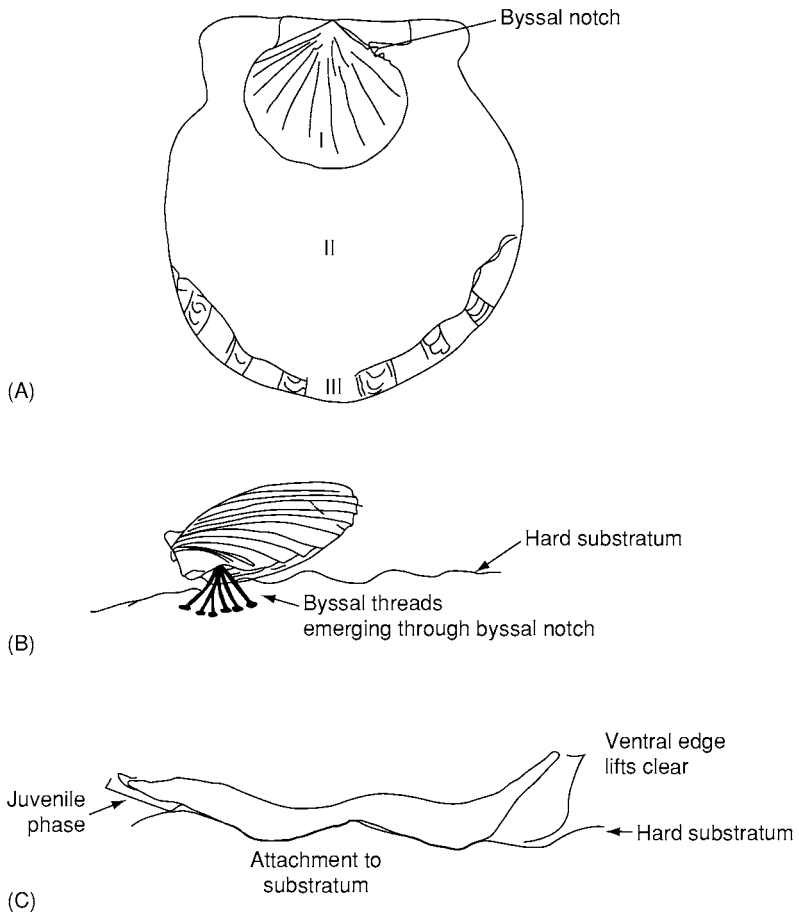


Figure 1 Reconstruction of life habits of Jurassic *Eopecten*. (A) External view of the right valve. I, Juvenile phase (the byssal notch is later filled with shell material); II, later growth (attachment scar, no ribbed ornament); III, valve lifts clear of substrate, coarse ribbed ornament. (B) Juvenile stage byssally attached to hard substrate. (C) Cross section through a mature animal, showing extensive cementation scar and raising of ventral margins above substrate.

example, monoplacophorans, a primitive caplike molluscan class (*see* **Fossil Invertebrates:** Molluscs Overview), are known from a variety of shallow marine sediments from the Lower Palaeozoic, apparently living in communities alongside brachiopods and trilobites. They then apparently disappear from the fossil record during the Devonian and were presumed extinct until the celebrated rediscovery of *Neopilina* during dredging of deep water (>3500 m) off the coast of Mexico during the Danish-led Galathea expedition in 1952. However, it would clearly be inappropriate to reconstruct the palaeoecology of the Palaeozoic forms based solely on our understanding of the biology of *Neopilina*. Presumably during the Upper Palaeozoic, monoplacophorans were forced, perhaps by competition or predation pressure, into deep-water refuges that are not represented in the sedimentary record. Other Palaeozoic taxa that were important and widespread in shelf communities, such as brachiopods and stalked crinoids, are postulated by some also to have been driven into refugia. These examples clearly demonstrate the need to use several lines of evidence when reconstructing palaeoecology and, most importantly, not to divorce the study of fossils from their lithological context.

What happens when there are no convenient living homologues? In such cases, the palaeoecologists must use more imaginative approaches. Trilobites (*see* **Fossil Invertebrates:** Trilobites) became extinct by the close of Palaeozoic, although much may be gleaned of their palaeobiology by homology and by analogy with other arthropod groups (such as understanding their moulting and likely internal anatomy). It is believed that trilobites exploited a wide range of life habits; although most were benthic, others were pelagic or nektonic. Some (e.g., *Calymene*) are interpreted as having been voracious predators, whereas others were thought to have been filter feeders or deposit feeders. The lines of evidence used to elucidate these different palaeoecological details are varied. Much is reliant on detailed studies of functional morphology, but evidence from trace fossils (including some that show evidence of trilobites tracking worms) and experimental modelling have also played a role. An interesting example is the interpretation of the Lower Palaeozoic olenellids as possible chemosymbionts in an elegant study by Richard Fortey. Using a combination of evidence based on the functional morphology of the family, principally the large number and wide nature of the thoracic segments and degenerate hypostome, the lithological evidence that olenellids tended to inhabit sulphur-rich unbioturbated black shales, and by analogy with totally unrelated modern chemosymbiotic organisms such as lucinid bivalves, Fortey suggested that these trilobites might well have harboured sulphate-reducing bacteria on their gills, which allowed them to live in 'difficult' oxygen-poor environments.

When living relatives and even good analogues are absent, more unconventional approaches can be taken in the study of the fossil mode of life. Graptoloids (*see* **Fossil Invertebrates:** Graptolites) were a significant group of macrozooplankton living in the Ordovician to mid-Devonian open seas. They have very remote living relatives, the pterobranchs, but their benthic, encrusting mode of life makes direct comparisons unrealistic. Instead, work has concentrated on the response of graptoloid shapes to seawater, which had the same properties in the Lower Palaeozoic as it does now. Physical and computer-derived models show that both the overall colony shape and its detailed morphology would have had important consequences for the behaviour of the species in water. Colonies were designed to rotate during movement, thereby increasing the volume of water that could be sampled. Flow of water was directed into the thecal apertures where the zooids lived by an arrangement of spines, hoods, or hooks in many species. A standard orientation relative to the dominant direction of water flow was maintained by structures such as the virgella and virgula, arrayed at the proximal and distal ends of the colony (**Figure 2**). Understanding the physical effects of adaptations shown by fossils allows inferences about life habits to be made (for example, concerning the likely feeding position of the zooids and the sampling requirements that conditioned their fitness and hence evolution).

Palaeosynecology

Ecologists define a community as a recurring group of taxa (in fixed proportions) that live together in a particular habitat. The focus of interest is in the way in which populations of single species or different taxa interact and the manner in which resources within that habitat are shared between the different groups. Ecological studies can be well founded on quantitative analyses based on direct observations. Palaeoecologists, however, are faced with a number of problems and must use less direct methods. First, a high proportion of any community may be soft-bodied or weakly skeletonized such that they are unlikely to be preserved in the fossil record. It is estimated that more 70% of modern taxa fall into this category and the Middle Cambrian Burgess Shale lagerstätte (*see* **Lagerstätten**) provides a salutary lesson, showing the wealth of poorly or unskeletonized organisms (such as worms and arthropods) not usually preserved in contemporary deposits. Second, relatively few organisms are preserved in life position and many may have been

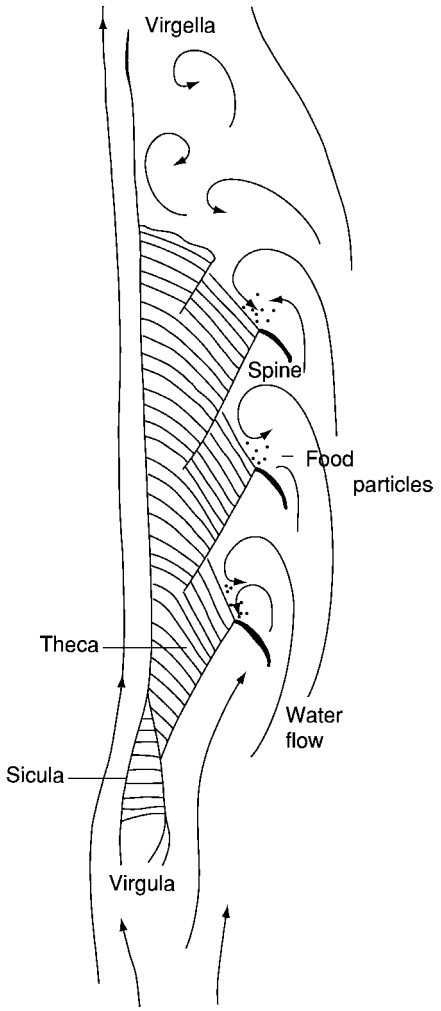


Figure 2 Sketch of flow patterns of water over a monograptid graptoloid model. The virgula and virgella maintain the colony in a stable orientation relative to the main current. Spines at the thecal openings produce local turbulence and direct flow over the apertures, concentrating food particles conveniently for animals living in the thecae.

transported far from their original habitats. As a result, the palaeoecologist must consider whether the group of fossils being studied is autochthonous (i.e., preserved in life position), parautochthonous (i.e., elements are disturbed and disarticulated, but the fossils are not necessarily far removed from their origins), or allochthonous (i.e., organisms from different habitats have been swept together) before drawing any conclusions. Third, even if a group of fossils is considered autochthonous, it has to be recognized that any bed, or even bedding plane assemblage, is going to be time averaged, and it will thus contain a sample of individuals that lived on that site over many centuries or even thousands of years. In this way, individuals of the same species but of very different generations may be found next to one another, or taxa that lived in successive communities on that site may appear to have co-existed. As a result of these complications, palaeoecologists prefer to use the term 'assemblage' rather than 'community'. In addition, the soft-bodied components of the original communities are seldom found fossilized. If similar assemblages are found to occur widely, either geographically or temporally, this provides a good indication that they are a real ecological phenomenon, and these are given the name 'associations'.

Investigating Fossil Populations

Populations of organisms are sensitive to characteristics of the environment and can be studied in two main ways, by the examination of life strategy and through survivorship analysis (Figure 3). Two extreme life strategies, commonly termed *r* and *K*, are derived from the logistic equation that states that the number of individuals (N) in a population will be a function of reproductive rate (r) modified by closeness of the number of individuals to the total number that can be supported by an environment (the carrying capacity, or K):

$$dN/dt = rN[(K - N)/K]$$

The *r*-strategist species depend heavily on high rates of reproduction, often breed early (at small size), and may die after breeding. They generally invest little in the care of their offspring, but have many offspring at once. The strategy is characteristic of unstable environments in which the carrying capacity varies significantly over short time-scales. The *K*-strategist species breed late, often bearing single young over whom care is expended. Individuals often live for a considerable time, even after breeding has ceased. This strategy is typical of stable environments in which the carrying capacity is highly predictable. Although these two descriptions are extremes, and most organisms show a mixture of *r* and *K* features, the generalizations have great use in making first-level interpretations of population dynamics.

Survivorship analysis is a complementary tool for use in the examination of fossil populations. The number of survivors of each age group of a population is plotted on a logarithmic scale as a function of age. In fossil populations, the number of survivors can be inferred if the bedding plane is time averaged, or can be observed if mass mortality of a single group produced the association. Most survivorship curves fall into three distinct categories, though variations exist, and organisms with very different life habits at different stages of their lives may show composite patterns. In cases in which high infant mortality occurs, as might be seen, for example, in fish stocks,

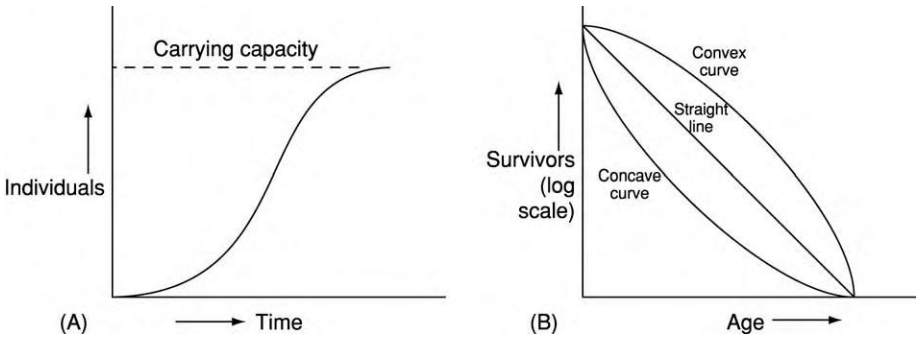


Figure 3 Key graphs for the interpretation of fossil populations. (A) A plot of the logistic equation, showing the rate of population increase slowing as the population size approaches the carrying capacity of the local environment. (B) Survivorship curves for populations with a high juvenile mortality (concave curve), an equal chance of mortality at any age (straight line), and a high gerontic mortality (convex curve).

the survivorship curve will be concave. In environments in which there is an equal risk of mortality at any age, the survivorship curve will be straight. The classic example of this type of curve is that of garden birds. Palaeontological examples are found, especially in extreme environments in which mortality is generated by frequent, unpredictable, and unselective events. In more stable environments, mortality rate is low initially but increases with age after a certain point, producing a convex survivorship curve. In these cases, organisms can be fit enough to survive the variation seen in their environment, but this characteristic is reduced later in life through illness or gerontic effects.

Investigating Interactions between Organisms

Organisms may interact in a number of different ways. Most important among these are competition for resources (e.g., for nutrients, space, light, or mates), either intra- or interspecifically; predator–prey interactions; and symbiosis, which runs through a spectrum, from being beneficial to both partners (commensalism) to parasitism, in which one partner's advantage is deleterious to the other. Ecologists have found that these interactions play an important role in controlling the fitness of modern organisms, and competition and predation in particular are thought to be extremely important driving forces in evolution. At first sight it might seem that the fossil record is likely to be virtually mute on the interactions between organisms; however, this is not the case. Clearly, individual cases must be carefully assessed, but it is possible to generate quantitative data on interactions between organisms from the fossil record.

Competition

The competitive exclusion principle suggests that two species that have exactly the same requirements cannot exist together within a habitat. Where two species do appear to have the same requirements, one will be prevented from maintaining a viable population by the success of the other, or one will be forced to adapt in order to use slightly different resources. The latter strategy may involve character displacement (e.g., modification of filter-feeding apparatus to cope with particles of a different size).

Competition has often been invoked as a cause of macroevolutionary change, although it is difficult to demonstrate. For example, some have suggested that the post-Palaeozoic decline in articulated brachiopods was at least in part influenced by the superior competitive advantages of bivalves that may occupy the same niche, and a similar argument has been made to link the success of dinosaurs with the decline of synapsids towards the end of the Triassic. One situation in which it has been possible to collect high-quality data concerns interactions between borers and skeletonized encrusters (e.g., bryozoans, oyster-like bivalves, and serpulid worms). These organisms have the advantage of having been preserved *in situ* and thus the fossils maintain their spatial relationships to one another and to the substratum. It must be remembered that soft-bodied encrusters (e.g., anemones, sponges) that may also have co-existed will be unrepresented, although evidence of their existence may be revealed by bioimmuration (preservation by the overgrowth of a skeletonized organism). In these cases, it is also possible to examine competitive overgrowth. It is known that space on hard substrates is at a premium in modern marine habitats, from which it may be inferred that taxa that are able rapidly to colonize space will be at an advantage over those that do so less aggressively. Much of this type of research has been centred on two clades of extant bryozoans, the cyclostomes and cheilostomes (*see* **Fossil Invertebrates:** Bryozoans). Both evolved independently from soft-bodied ctenostome ancestors, the cyclostomes in

the Ordovician and the cheilostomes in the Cretaceous. The cheilostomes replaced the cyclostomes as the dominant bryozoan clade and it has been argued that the former were more successful competitors. An analysis of competitive overgrowths between the two groups from a range of localities (Cretaceous to Recent) by McKinney has revealed that cheilostomes consistently 'won' around 66% of all interactions between the two clades, although interestingly this figure appears to have remained constant over geological time.

Predation

All organisms die, and a great many deaths are the result of being eaten by predators, becoming part of the food chain. The need to avoid predation is a very important selection pressure, hence the battery of defensive morphologies (e.g., spines and thick shells), behaviours (such as leaping by cockles or swimming by scallops), and life habits (such as deep burrowing) that may be demonstrated in modern organisms. Very rarely, predators may be fossilized in the act of predation, with prey remains preserved in their guts, for example. But predator–prey interactions may be interpreted from the fossil record in a number of other ways, such that palaeoecologists are not limited to exceptional preservation to chart the evolution of different predatory groups and methods. Various aspects of an organism's functional morphology, such as possession of teeth or claws, may suggest that it was a predator. For example, the large, toothed chelicerae of Palaeozoic eurypterids (large aquatic scorpions) have been used to suggest that they may have been powerful predators of the early fish. Most theropod dinosaurs (*see* **Fossil Vertebrates:** Dinosaurs), including the well-known *Tyrannosaurus rex*, have narrow, sharp-pointed curved teeth, many with serrated edges, which suggests that they were carnivores. The most useful evidence, however, has come from the study of prey remains. Although some predatory groups eat their prey without leaving any evidence of damage (for example, starfish that prise apart bivalve shells, or fish that simply swallow their prey whole), there are two important and widely used predatory methods on shelled prey, crushing and drilling, that can be recognized in the fossil record. A whole variety of arthropod and vertebrate predatory groups feed by crushing open the shells of their victims, many leaving very diagnostic breakage patterns. Similarly, a number of worms, gastropods, and octopods drill neat holes in their victims, for either the extraction of tissue or the injection of toxins (Figure 4). Where such damage is evident, it is possible to undertake quantitative surveys of the frequency of attack by a particular predatory group and the frequency of prey repair (which provides a measure of the effectiveness of their defences), and also some aspects of behaviour, such as prey preference or stereotypic (predictable and so diagnostic of the predator) attack strategy.

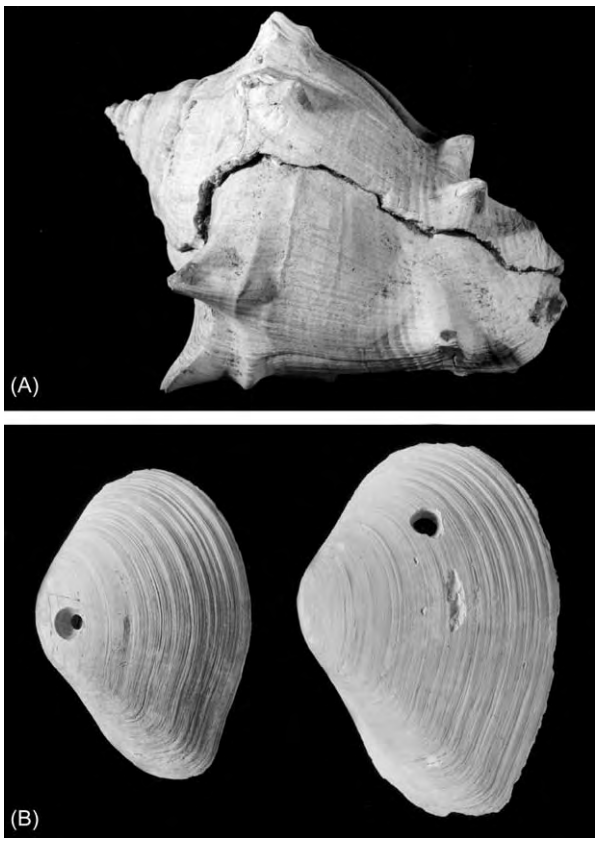


Figure 4 (A) Repaired predator induced breakage to gastropod prey. Miocene, France (×1). (B) Predatory drill holes produced by naticid gastropods in bivalve prey. Pliocene, United States (×1.5).

There is now a strong body of evidence to suggest that levels of predation pressure have increased during the Phanerozoic. In particular, it is evident that there have been marked increases in the past 270 million years, especially during the Cretaceous and Early Tertiary, during which time the numbers of predator taxa increased and more specialized feeding techniques evolved. This is part of the so-called Mesozoic Marine Revolution, a major restructuring of marine communities following the Permian–Triassic extinctions and leading to ecosystems taking on their 'modern' aspect. It can be argued that this increase in predation pressure had a profound effect on the evolution of individual taxa and changed the ecological properties of communities. It has been effectively shown that post-Palaeozoic gastropods have become better armoured (thickening their shells,

more taxa with slit-like apertures, and fewer taxa with loose coiling) and therefore better able to resist crushing predators, such as crabs. Many bivalve taxa adopted 'new' life habits such as cementation and deep burrowing, which can be shown to have decreased the threat from predators, and stalked crinoids moved away from shallow water communities into deeper habitats where there would have been fewer predators.

Symbioses

If two different fossil taxa are frequently found together, this may be evidence of a symbiotic relationship, but it may not be easy to demonstrate where the benefits and the costs of the relationships lay. Upper Palaeozoic platyceratid gastropods are frequently preserved in close association with the calyces of a variety of crinoids, in particular in a position close to or over the anus. The margins of these snails mirror closely the morphology of the calyx, implying that the animals were in close proximity long-term and that the gastropod literally grew to fit its location. The relationship appears to have been detrimental to the crinoid (i.e., a case of parasitism), because those with platyceratids are usually smaller than those without.

Hermatypic scleractinian corals (*see* **Fossil Invertebrates:** Corals and Other Cnidaria) are involved in an intimate symbiotic relationship with zooxanthellae that live within their soft tissue. Theirs is a commensal relationship of mutual benefit to both partners: the photosynthetic zooxanthellae gain a living site free from sediment cover, held aloft in the water and protected from the stinging nematocysts of the host, and the coral uses carbon fixed by the zooxanthellae to augment dramatically their own calcification processes. It is a strong relationship and many corals cannot survive without their zooxanthellae, which explains the very stringent requirements of scleractinian corals to inhabit clear water within the photic zone.

Examples of Palaeosynecological Studies

The Middle Cambrian Burgess Shale is one of the most important lagerstätten to have been discovered. The vast collection of material from the site has allowed a detailed examination of the palaeosynecology of the deposit. In a survey of more than 65 000 specimens collected from the Phyllopod Bed of the Burgess Shale, Conway Morris was able to examine the life habits of each of the taxonomic groups represented and to make calculations of biomass. It was also possible to undertake a trophic analysis (i.e., examining the feeding methodologies of different elements of the fauna). Largely using evidence from functional morphology, as well as some instances of preserved gut contents, Morris demonstrated the presence of deposit and suspension feeders, predators, and scavengers and allowed the tentative reconstruction of a trophic web.

In another study, Fürsich, Palmer, and Goodyear examined the palaeosynecology of bivalve-dominated patch reefs from the Upper Jurassic of the southern UK. Their studies recognized a number of 'guilds', or species that exploit a particular way of life or resource. Primary framebuilders (mostly bivalves, solenoporacean algae, and bryozoans) accounted for 55–70% of the volume of the reef, but there were also accessory encrusters (other bivalves, bryozoans, serpulids, and forams); a crevice-dwelling fauna (including byssate bivalves, small terebratulid brachiopods, and worms) that occupied cavities within the reef and borings made by endoliths; borers on a variety of scales, from large borings of bivalves to microscopic borings made by a variety of taxa (including phoronids,

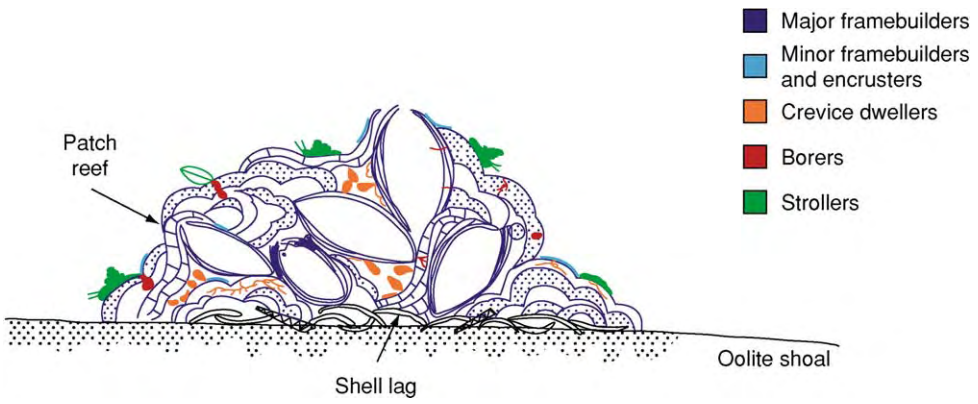


Figure 5 A reconstruction of a patch reef from the Upper Jurassic Portland Beds of South England. The key shows the different guilds of organisms, linking taxa by function within the reef, rather than by family tree.

crustaceans, and bryozoans); and a small number of vagile taxa, or strollers, largely gastropods, that roamed the surfaces of the reefs. It is likely that this latter guild is under-represented. These studies suggest that the reefs built up after initially unstable oolitic shoals had been stabilized by both infaunal and epifaunal bivalves, the shells of which then provided the firm seafloor onto which the reef-forming organisms were recruited (**Figure 5**).

See Also

Fossil Invertebrates: Trilobites; Bryozoans; Corals and Other Cnidaria; Graptolites; Molluscs Overview; Bivalves; Gastropods. **Fossil Vertebrates:** Dinosaurs. **Lagerstätten**. **Microfossils:** Foraminifera. **Palaeopathology**.

Further Reading

Brenchley PJ and Harper DAT (1998) *Palaeoecology: Ecosystems, Environments and Evolution*. London: Chapman and Hall.

Briggs DEG and Crowther PR (1990) *Palaeobiology: A Synthesis*. Oxford: Blackwell.

Briggs DEG and Crowther PR (2001) *Palaeobiology II*. Oxford: Blackwell.

Conway Morris S (1986) The community structure of the Middle Cambrian Phyllopod Bed (Burgess Shale). *Palaeontology* 29: 423 467.

Fortey RA (2000) Olenid trilobites: the oldest known che mautotrophic symbionts? *Proceedings of the National Academy of Sciences* 97: 6574 6578.

Fürsich FT, Palmer TJ, and Goodyear KL (1994) Growth and disintegration of bivalve dominated patch reefs in the Upper Jurassic of southern England. *Palaeontology* 37: 131 171.

Gould SJ and Calloway CB (1980) Clams and brachiopods ships that pass in the night. *Paleobiology* 6: 383 396.

Harper EM and Palmer TJ (1993) Middle Jurassic cemented pectinids and the missing right valves of *Eopec ten*. *Journal of Molluscan Studies* 59: 63 72.

Kelley PH, Kowalewski M, and Hansen TA (2002) *Preda tor Prey Interactions in the Fossil Record. Topics in Geobiology 20*. New York: Kluwer Academic/Plenum Publishers.

McKinney FK (1995) One hundred million years of com petitive interactions between bryozoan clades: asymmet rical but not escalating. *Biological Journal of the Linnean Society* 56: 465 481.

Rollins HB and Brezinski DK (1988) Reinterpretation of crinoid platyceratid interaction. *Lethaia* 21: 207 217.

Taylor PD and Wilson MA (2003) Palaeoecology and evolution of marine hard substrate communities. *Earth Science Reviews* 62: 1 103.

PALAEOMAGNETISM

T H Torsvik, Geological Survey of Norway, Trondheim, Norway

Introduction

Palaeomagnetism is the study of the Earth's magnetic field preserved in rocks. The discovery that some minerals, at the time of their formation, can become magnetized parallel to the Earth's magnetic field was made in the nineteenth century. Early in the twentieth century, Bernard Brunhes made the startling discovery that some rocks are magnetized in the opposite orientation to the Earth's present-day magnetic field. This led him to propose that the Earth's magnetic field had reversed its polarity in the past. These reversals have subsequently been shown to be non-periodic and the Earth's magnetic field reversal history is now well known for the past 175 million years and more sketchingly understood to the beginning of the Palaeozoic (*ca*. 545 Ma).

Palaeomagnetism has a range of application potential (*see* **Magnetostratigraphy, Analytical Methods:** Geochronological Techniques), and the focus here is on understanding the importance of palaeomagnetism as an investigative tool in assembling palaeogeographical reconstructions.

Fundamentals

The Earth's magnetic field is believed to originate from the outer fluid core and, at the surface, the field is described by its inclination (angle with respect to the local horizontal plane), declination (angle with respect to the Greenwich meridian), and field strength (**Figure 1**). The inclination of the Earth's field varies systematically with latitude, which is of prime importance for palaeomagnetic reconstructions. At the north magnetic pole, the inclination of the field is $+90^\circ$ (straight down), at the equator the field inclination is zero (horizontal) pointing north, and at the south magnetic pole the inclination is -90° (straight up; **Figure 2**). The magnetic north and south poles currently differ from the geographical north and south poles by 11.5° because the magnetic axis is inclined from the geographical (=rotation) axis. The

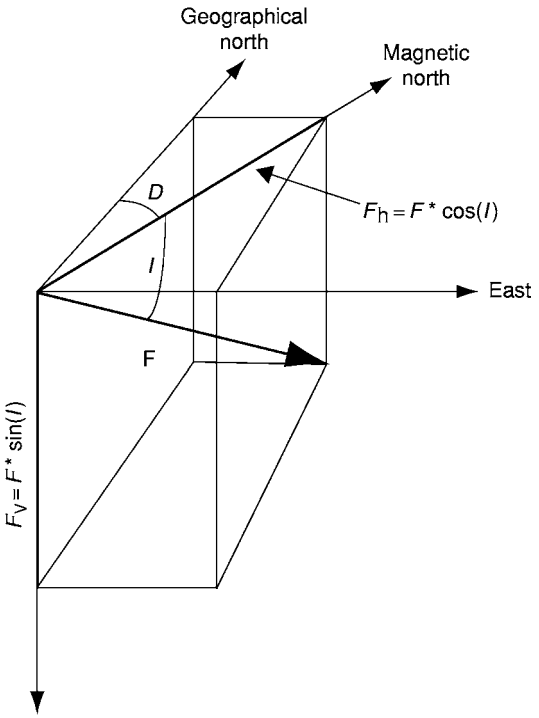


Figure 1 The direction and intensity of the total field vector (**F**) decomposed into declination from geographical north (D) and inclination from the horizontal (I). The equations relate the horizontal (F_h) and vertical (F_v) components of the total field (**F**) with I and D.

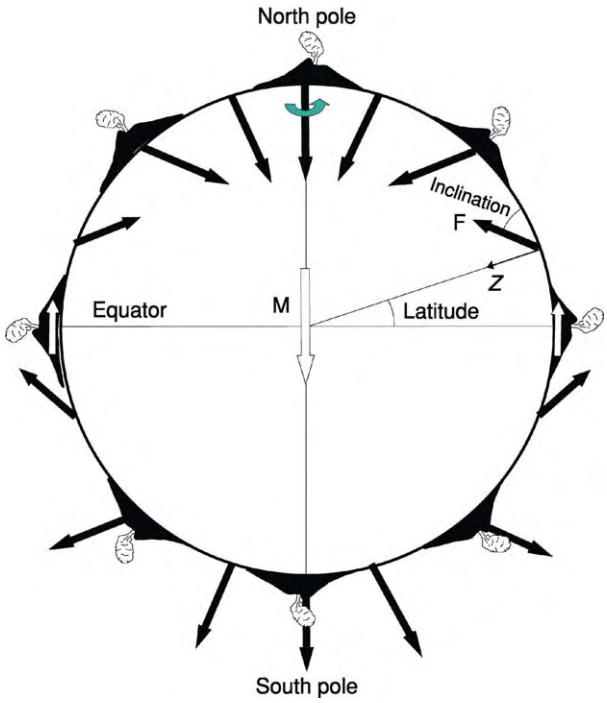


Figure 2 Field lines at the Earth's surface for a geocentric axial dipole. At the equator, the inclination is flat (zero) and, at the north and south poles, the inclination is vertical (+90° and −90°, respectively). The inclination recorded in volcanoes formed on the Earth's surface is dependent on the latitude (see also **Figure 3**). Declinations in a normal polarity field, such as today's, should point to the north.

magnetic axis, however, is slowly rotating/precessing around the geographical axis (known as secular variation) and, over a period of a few thousand years, it is hypothesized that the averaged magnetic poles correspond reasonably well with the geographical poles. This is known as the geocentric axial dipole (GAD) hypothesis. We can therefore imagine that a magnetic dipole is placed at the centre of the Earth and aligned with the Earth's rotation axis (**Figure 2**). In palaeomagnetic studies, it is therefore important to sample rocks whose ages range over more than a few thousand years; a study of a single dyke or basalt flow, for example, represents an instantaneous reading of the Earth's magnetic field (cools within a scale of days to weeks) and will not accurately record the position of the Earth's rotation axis.

When rocks form on an ideal planet, they acquire a remanent (permanent) magnetization parallel to the Earth's magnetic field at that location (**Figures 2 and 3**). There are a number of ways in which a rock can acquire a remanent magnetization, but most rocks are magnetized by one of the following processes: (1) as magma solidifies and cools below the Curie temperature (T_C, i.e., temperature above which a magnetic material loses its magnetism because of thermal agitation), magnetic minerals acquire a thermoremanent magnetization aligned with the Earth's magnetic field at the time of cooling; (2) during the deposition of sediments, magnetic mineral grains settle statistically in the direction of the Earth's magnetic field and a detrital remanent magnetization is acquired; and (3) when magnetic minerals are formed during chemical processes (diagenesis or metamorphism), the magnetic minerals grow as magnetized crystals with their magnetization in the direction of the external magnetic field; this creates a chemical remanent magnetization. Most magnetic minerals are iron–titanium oxides that belong to two solid solution series (**Figure 4**): the titanomagnetites (e.g., the end-member magnetite) and the titanohaematites (e.g., haematite). In the titanomagnetite series ($Fe_{3-x}Ti_xO_4$), there is an approximately linear variation of spontaneous magnetization ($\mathbf{M}_s$) and T_C with composition (x). Other common magnetic minerals include goethite and pyrrhotite (**Table 1**).

As an example, we can consider the remanence acquisition of a basaltic lava flow during cooling. The most important magnetic mineral in basaltic rocks is titanomagnetite. Low-titanium phases (e.g., pure

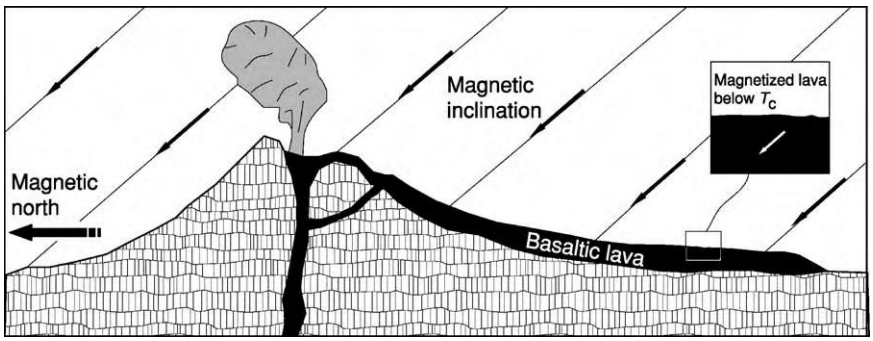


Figure 3 Example of acquisition of a thermoremanent magnetization (TRM) at intermediate northerly latitudes (acquired in a normal polarity field similar to today's). A lava will acquire a TRM upon cooling below the Curie temperature (see text), and the inclination will parallel the inclination of the external field and have declinations due north.

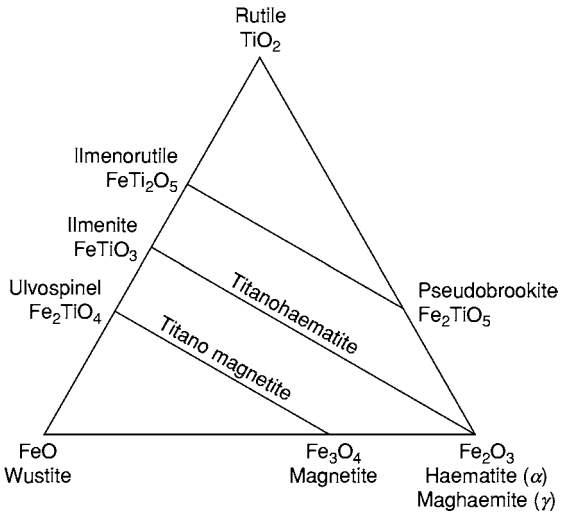


Figure 4 Ternary diagram showing the magnetically important iron oxide minerals in the titanomagnetite and titanohaematite solid solution series.

Table 1 Magnetic properties of some common minerals in rocks

Mineral (composition)	M_s(kA m ^{1})	T_C (°C)
Titanomagnetite ($Fe_{2\,4}Ti_{0\,6}O_4$)	125	150
Magnetite (Fe_3O_4)	480	580
Maghaemite (γ Fe_2O_3)	380	590 675
Haematite (α Fe_2O_3)	~2.5	675
Goethite (α FeOOH)	~2	120
Pyrrhotite (Fe_7S_8)	~80	320

magnetite; Table 1) have T_C values close to 580°C, whereas the presence of titanium lowers the T_C value. During a basaltic volcanic eruption, the temperature of a lava is approximately 1200°C; when a lava flow cools below T_C, the Earth's magnetic field is recorded within the lava flow (Figure 3). The declination, inclination, and magnetization intensity, which are proportional to the strength of the field, can today be measured in the laboratory. The magnetization intensity can vary by several orders of magnitude between different rock types, and thus different laboratory instruments are required to measure the magnetization precisely. Volcanic rocks normally have high intensities and the magnetization can be measured on standard spinner magnetometers. Conversely, sedimentary rocks can be extremely weakly magnetized, and highly sensitive superconducting magnetometers (superconducting quantum interference devices, SQUIDs) are required to measure and unravel their magnetization history.

Palaeomagnetic Analysis

In the early days of palaeomagnetic studies, it was common to measure the magnetization in a rock and assume that this magnetization, referred to as the natural remanent magnetization (NRM), represented a primary magnetization that had survived magnetic resetting from subsequent thermal or chemical activity. However, during the 1970s, it became more and more evident that rocks can undergo magnetic resetting, and it is therefore now standard procedure to test the stability of the NRM by thermal, alternating field, or chemical (rare) demagnetization. With the former method, a sample is measured following heating to higher and higher temperatures in a 'zero' field oven. From the 1980s, it became standard procedure to display the demagnetization data in orthogonal vector plots, also referred to as Zijderveld diagrams (Figure 5). These diagrams portray directional and intensity changes on a single diagram – magnetization components are identified as linear segments in both the horizontal ('declination') and vertical ('inclination') planes. Components and the degree of linearity can be computed using least-squares algorithms. A single-component

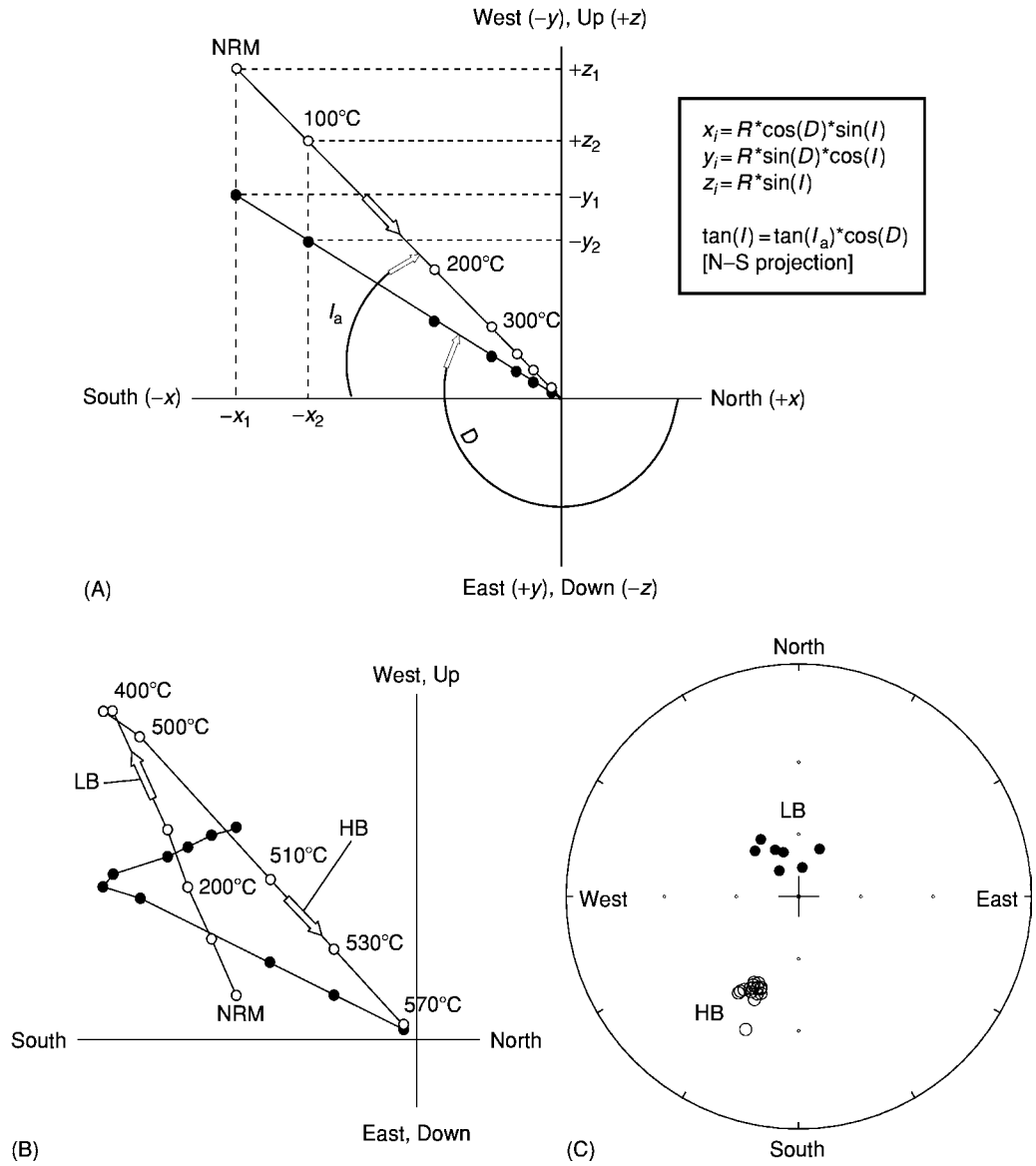


Figure 5 A Zijderveld plot illustrating vector components of magnetization during progressive thermal demagnetization, projected onto orthogonal horizontal (declination; filled symbols) and vertical (inclination; open symbols) planes. For each demagnetization step, the measured declination (D), inclination (I), and magnetization intensity (R) are decomposed to cartesian coordinates, x_i, y_i, and z_i, using the formulae in the box in (a). N S (+x, x) is selected as the projection plane as the declination is closer to N S than E W. The horizontal component is plotted as x_i,y_i, whereas the vertical plane is plotted as x_i,z_i. This procedure is repeated for each demagnetization step (in the example, natural remanent magnetization (NRM): 100°C, 200°C, etc.). Using this procedure, a magnetic component is recognized as a linear vector segment. The declination for a component can be read directly from the diagram, normally numerically computed by least squares analysis, whereas the inclination is apparent (I_a) and always larger than the real I; the 'distortion' of I_a depends on the projection plane. (A) Single component magnetization decaying towards the centre of the diagram. (B) Two component magnetization with a low unblocking component (LB) identified below 400°C and a high unblocking component (HB) decaying towards the centre of the diagram; the stability up to 570°C suggests pure magnetite as the remanence carrier. (C) Mean site compilation of HB and LB components shown in a stereoplot. LB is a recent overprint, whereas HB should be considered to be a Late Permian magnetization from the Oslo area (Baltica). The mean declination/inclination in this hypothetical study is 205°/ 41° (95% confidence circle around the mean, α_{95}, is 2.3°). The calculated palaeomagnetic pole for the site (60° N, 10° E) is 49.3° N and 152.3° E. In the stereoplot, open (filled) symbols denote negative (positive) inclinations.

magnetization is identified by single vector decay towards the origin of the diagram as the sample is progressively demagnetized (Figure 5A). Multi-component magnetizations, in which a primary component has been partly overprinted by younger components, can be recognized by the presence of two or more linear segments (Figure 5B). In the latter example, a hypothetical Permian dyke magnetization

from the Oslo region (Norway), with SSW declinations and negative (upward-pointing) inclinations (high unblocking component), is partially overprinted by a younger magnetization of Holocene origin, with NNW declinations and positive (downward-pointing) inclinations (low unblocking component). When the blocking temperatures overlap, curved segments are observed.

In a typical palaeomagnetic study, 5–10 samples are analysed from each site. Magnetization components are computed by least-squares analysis, a mean direction is computed for each site using Fisher statistics, and, finally, a mean direction is calculated from all the sites. In our example, we can imagine that the stereoplot in Figure 5C represents the mean directions from numerous individual dykes in the Oslo region.

Palaeomagnetic Stability Tests

Magnetic overprinting or resetting presents a problem. However, there are four fundamental tests used to check the stability and the potentially primary character of magnetizations: (1) the fold test; (2) the reversal test; (3) the conglomerate test; and (4) the contact test (Figure 6). The fold test determines whether a magnetization was formed prior to or after folding of a rock, and is therefore a relative test that does not prove a primary origin of the remanence. If a remanence is pre-fold, the site vectors should be dispersed after folding (Figure 6A); conversely, if a magnetization is post-fold, the site vectors should be similar throughout the fold structure (Figure 6B). The presence of antipodal stratigraphically linked reversals in a sedimentary or basaltic sequence is the best evidence for primary remanence (Figure 6A). The conglomerate test is also a powerful test. If magnetizations are random between individual boulders (Figure 6A), this is an indication for a primary magnetization in the host rock. Conversely, if boulder magnetizations concur with those in the surrounding rocks, magnetic overprinting is indicated (Figure 6B). A contact test is employed to check whether an intrusion or a dyke carries a primary magnetization. A dyke and its baked margin should coincide, whilst non-baked samples should differ if significantly older

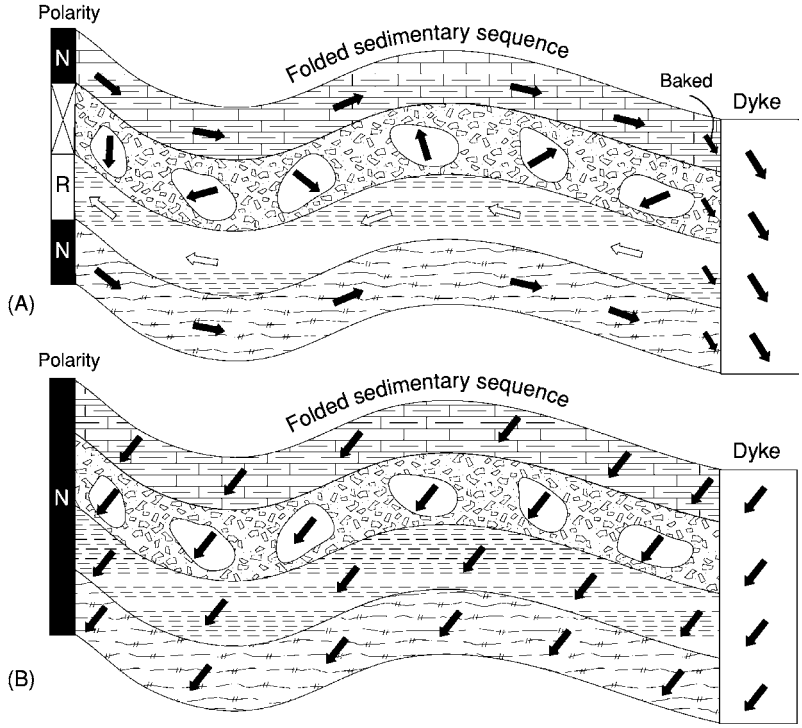


Figure 6 (A) Positive field tests. We assume that a sedimentary sequence was deposited at the equator (flat inclinations) and later folded. Polarity is indicated by the arrows. As the inclinations follow the bedding (except in the conglomerate), this is a pre fold magnetization. Evidence for a primary magnetization is witnessed by layers with antipodal polarity, and we can therefore establish a magnetostratigraphy (alternating normal and reversed magnetic fields). Further evidence for a primary magnetization is observed from a layer with boulders that show random magnetization vectors. A dyke intruded the folded sequence later, and steeper inclinations (indicating that the continent must have moved to higher latitudes) from the dyke and its baked/chilled margin indicate a primary dyke magnetization. (B) Negative field tests. All layers, conglomerate boulders, and the dyke have similar magnetizations and no reversals are observed. This indicates a regional secondary overprint after the folding and dyke intrusion.

than the dyke (Figure 6A). Conversely, if the dyke and all the surrounding rocks have a similar magnetization, the dyke records a younger overprint (Figure 6B).

Palaeomagnetic Poles and Reconstruction of a Continent

Based on the measurement of the remanent inclination, we can calculate the ancient latitude for a continent when the rock formed from the formula: $\tan(I) = 2 \times \tan(\text{latitude})$. In addition, the remanent declination, which deviates from 0° or 180° (depending on the polarity of the Earth's magnetic field), provides information about the rotation of a continent.

The inclination and declination change with the position of the sampled rock on the globe (Figure 2), but the position of the magnetic pole of a geocentric axial dipole is independent of the locality at which the rock acquired its magnetization. Thus, it is practical to calculate pole positions in order to compare results from various sites or to perform plate tectonic reconstructions.

Ideally, as a time average, a palaeomagnetic pole (calculated from the declination, inclination, and the geographical site location) for a newly formed rock will correspond with the geographical north or south pole. If a continent moves later, the palaeomagnetic pole must move with the continent. To perform a reconstruction with palaeomagnetic poles, we therefore have to calculate the rotation (Euler) pole and angle which will bring the palaeomagnetic pole back to the geographical north or south pole, and then rotate the continent by the same amount (Figure 7A). In our example, a palaeomagnetic pole (latitude, 49.3° N; longitude, 152.3° E), calculated from the situation depicted in Figure 5C, will position the Baltica continent (most of northern Europe eastward to the Urals) at latitudes between 15 and 50° N, causing the city of Oslo to have been located

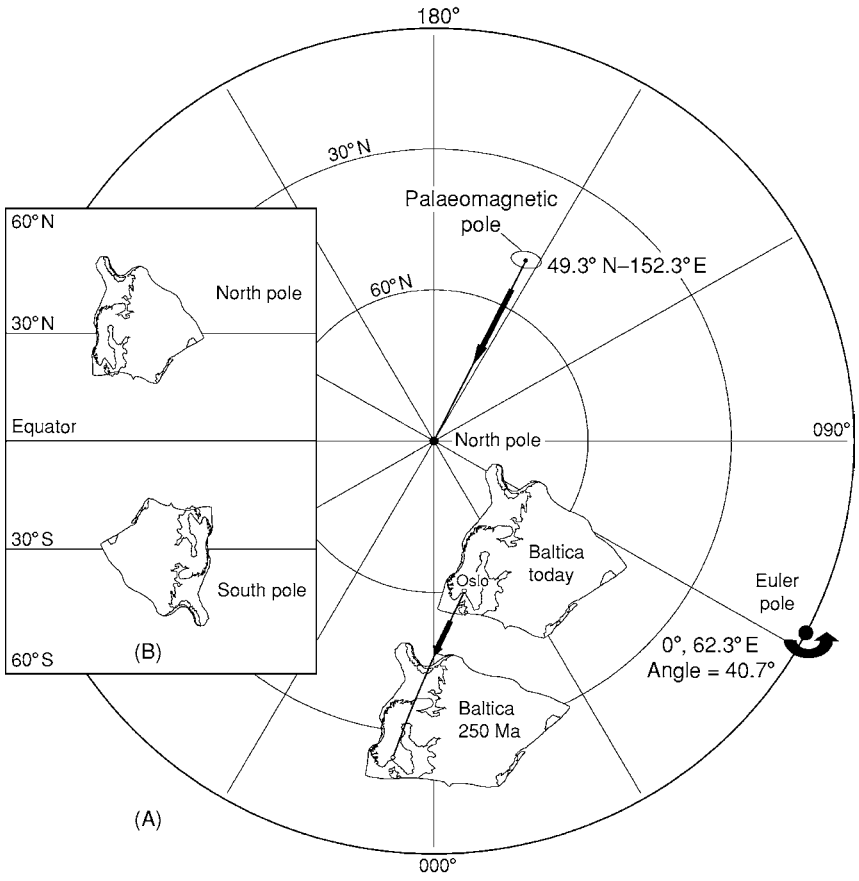


Figure 7 (A) The reconstruction of a continent, for example Baltica, is performed as follows. Determine the Euler pole needed to rotate a palaeomagnetic pole (in our case 49.3° N, 152.3° E) to the geographical north pole (we calculate 0°, 62.3° E and a rotation of 40.7°). This Euler pole is then used to rotate the continent by the same amount. Thus, Baltica today is rotated back about this pole to the position it occupied in Permian times. (B) In (A), we assumed that the palaeomagnetic pole was a north pole. If we assume a south pole, then the continent will be placed in the opposite hemisphere and geographically inverted.

at 24° N in the Late Permian (Figure 7A). Because the current latitude of Oslo is 60° N, Baltica must have drifted northwards since the Permian.

Palaeomagnetic data can only constrain latitude (based on inclination) and the amount of angular rotation (based on declination). Because the palaeolongitude is unknown, we can position Baltica at any longitude we wish, subject to other geological constraints. In addition to that uncertainty, we cannot tell in old rocks whether a palaeomagnetic pole is a south or north pole. In Figure 7A, we assumed that the pole was a north pole, but if we used a south pole, Baltica would plot in the southern hemisphere but in a geographically inverted orientation (Figure 7B). Hence, there is freedom to select north or south poles when producing reconstructions, placing the continent in an opposite hemisphere and rotated by 180°.

Apparent Polar Wander Paths

Apparent polar wander (APW) paths represent a convenient way of summarizing palaeomagnetic data for a continent or terrane, instead of producing palaeogeographical maps at each geological period. APW paths represent the apparent motion of the rotation axis relative to the continent, depending on whether one plots the movement of the north or south pole. APW paths can therefore be constructed as north or south paths. To construct an APW path, a set of palaeomagnetic poles of varying geological age

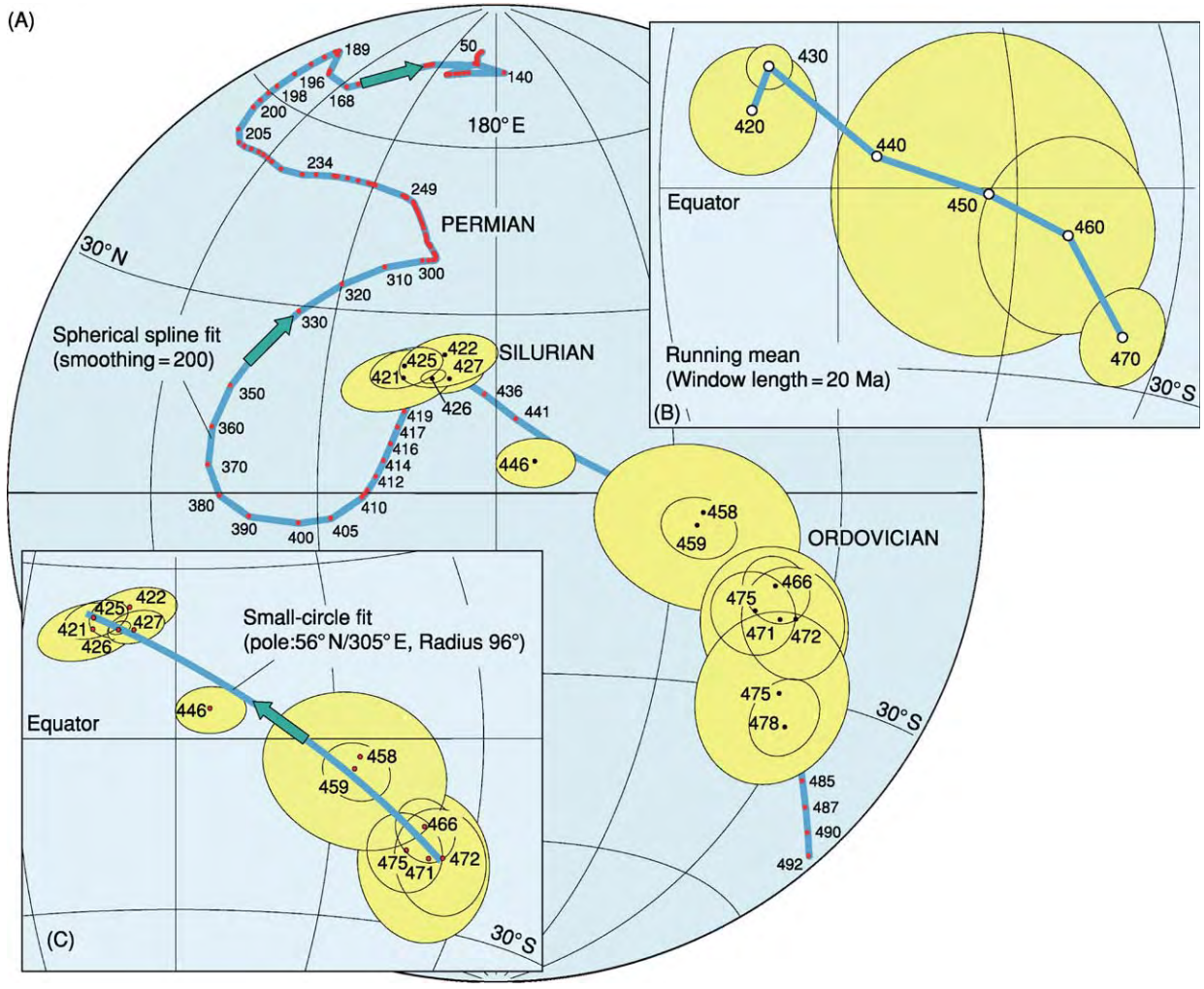


Figure 8 Examples of apparent polar wander (APW) paths for Baltica (include stable European data from Permo Carboniferous times). (A) Moderately smoothed spherical spline APW path from Early Ordovician to Early Tertiary times. Only Ordovician through Silurian input poles are shown. (B) Running mean path using the same input poles as in (A) and a 20 million year window. Mean poles are shown with 95% confidence ellipses (A95), except for the 440 Ma mean pole for which there was only one pole entry. The running mean path is only shown for the Ordovician Silurian section of the APW path. (C) Small circle path fitted to Ordovician Silurian poles as in (A). Input poles in (A) and (C) are shown with 95% confidence ovals (known as dp/dm).

are presented in a single diagram, and a synthetic path is fitted to the incrementing poles (Figure 8A). There are three common methods for generating APW paths: (1) spherical splines; (2) running mean (sliding time window); and (3) the small circle method.

The spherical spline method of modelling APW paths has been employed since the late 1980s. In brief, a spline constrained to lie on the surface of a sphere is fitted to the palaeomagnetic poles (Figure 8A), themselves weighted according to the precisions of the input palaeopoles. In the running mean method, palaeomagnetic poles from a continent are assigned absolute ages, a time window is selected (e.g., 20 million years), and all palaeomagnetic poles with ages falling within the time window are averaged. Using Fisher statistics, 95% confidence ellipses (known as A95 when averaging poles) can be calculated for each mean pole (Figure 8B). Both the spline method and the running mean technique are effective in averaging out random noise and allowing the basic pattern of APW paths to be determined.

The small circle method is based on the fact that movements of continents, APW paths, hotspot trails, ocean fracture zones, etc., must describe small circular paths if the Euler pole is kept constant. It is reasonable to assume that continents may drift around Euler poles that are kept constant for, say, some tens of millions of years. One can therefore fit APW segments along an APW path. This is demonstrated in Figure 8C where we can fit a small circle to Baltica poles from 475 to 421 Ma. However, after 421 Ma, the path changed direction markedly and this resulted from the collision of Baltica with Laurentia (North America, Greenland, and the British Isles north of the Iapetus Suture), which radically changed the plate tectonic boundary conditions and the APW path for Baltica.

Palaeolatitudes and Drift Rates – Links to Facies

Based on APW paths, we can calculate palaeolatitudes and plate velocities for a specific geographical location. Plate velocities are minimum velocities as the longitude is unconstrained; we only calculate latitudinal velocities. Figure 9 shows an example of such calculations based on the APW path in Figure 8A. In this diagram, we have also separated the northward and southward drift of Baltica. Drift velocities are typically below 8 cm per year, but peak velocities of around 14 cm per year are seen after collision of Baltica with Laurentia in Late Silurian–Early Devonian times.

The calculation of latitudinal velocities is important in order to check whether drift rates are

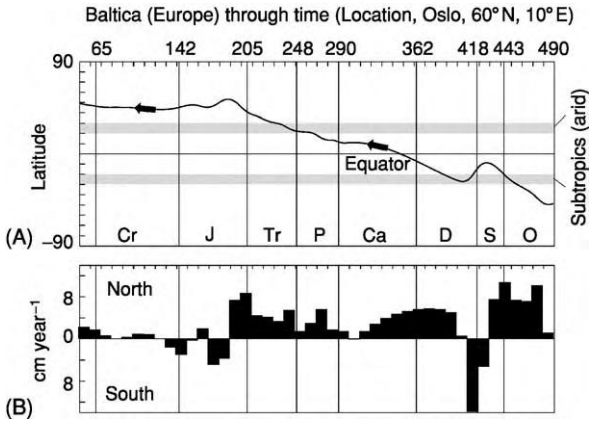


Figure 9 Latitude motion (A) and velocities (B) for Baltica (city of Oslo) from Ordovician to present times based on palaeomagnetic data (Figure 8A). Baltica was in the southern hemisphere during Ordovician through Devonian times, crossed into the northern hemisphere in the Early Carboniferous, and continued a general northward drift throughout the Mesozoic and Cenozoic. Northward and southward latitudinal translations throughout the Early Palaeozoic were accompanied by velocity peaks in the Early Silurian (northward) and the earliest Devonian (southward). Cr, Cretaceous; J, Jurassic; Tr, Triassic; P, Permian; Ca, Carboniferous; D, Devonian; S, Silurian; O, Ordovician.

compatible with 'modern' plate tectonic velocities. A rate of 18 cm per year (India) is the highest reliable value reported for the last 65 million years. When values appear unrealistically high (e.g., more than 20–30 cm per year), some authors have appealed to true polar wander (TPW) as a plausible explanation. TPW is a highly controversial subject that implies rapid tilting of the Earth's rotation axis, and is not generally accepted.

The distribution of climatically sensitive sediments, such as glacial deposits, coal, carbonates, and evaporites, is useful to check the palaeolatitudes derived from palaeomagnetic data. Glacial deposits are usually confined to polar latitudes and, except during the recent ice ages, there is no evidence for such deposits in Southern Baltica, as predicted by the palaeomagnetic data (maximum 60° S in the Early Ordovician). Carbonates, particularly in massive build-ups, such as reefs, are more common in lower latitudes. During Ordovician and Silurian times, Baltica drifted to subtropical and tropical latitudes, as witnessed by the presence of Bahamian-type reefs in Southern Baltica. Evaporites typically record dry climates within the subtropics (20–30°). During the Late Permian, Baltica was located at subtropical northerly latitudes, and the Late Permian coincides with large evaporite deposits in the North Sea area that subsequently became important in hydrocarbon trap development.

Palaeomagnetism and Palaeogeography: the Big Picture

Quantitative reconstructions are most commonly derived from hotspot traces (Cretaceous–Tertiary times) and ocean-floor magnetic anomalies, but prior to the earliest *in situ* ocean floor preserved today (approximately 175 Ma), the positioning of continents can only be quantitatively recognized by palaeomagnetism. As longitude is unknown from palaeomagnetic data, the identification and discrimination of faunas and floras can indicate that continents with similar faunas were in proximity to one another; conversely, different faunas of the same age can indicate the separation of the continents.

In order to construct a global palaeogeographical map, palaeomagnetic data from individual continents or terranes must be compiled and evaluated in terms of reliability. At any given time, palaeomagnetic data may not exist for all continents and additional criteria, such as fauna, flora, facies, and tectonic and

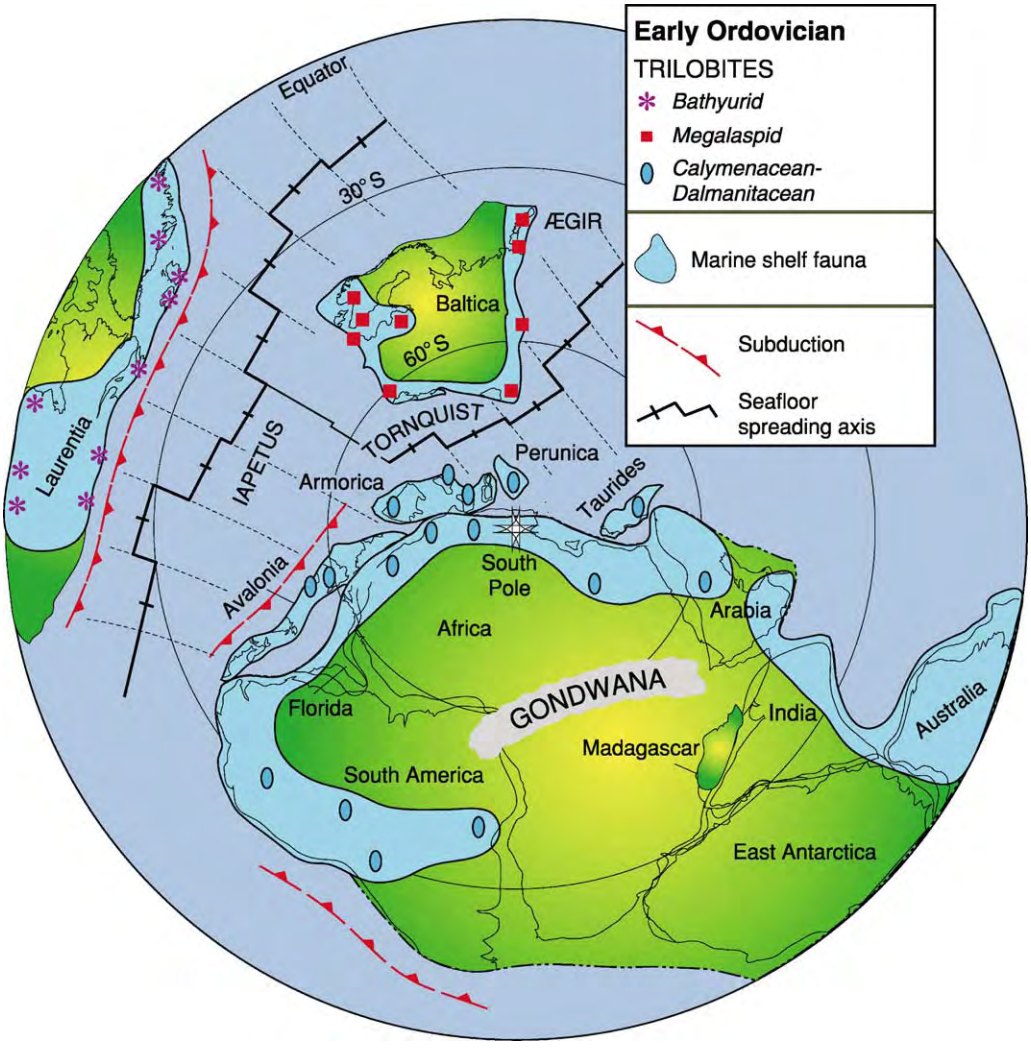


Figure 10 Reconstruction for Early Ordovician times (490 470 Ma) with the major terranes and some key Arenig Llanvirn trilobite faunas. Laurentia (located at the equator) includes North America, Greenland, and the British Isles north of the Iapetus Suture. Baltica (intermediate southerly latitudes) includes most of Scandinavia and Russia eastwards to the Urals. The core of Gondwana consists of Africa, Arabia, Madagascar, Greater India, most of Antarctica, most of Australia, Florida, and most of South America. Gondwana formed at around 550 Ma and covered more than 90° of latitude in the Early Ordovician. Gondwanan dispersal history commenced with the rifting off of Avalonia at *ca.* 465 Ma. Avalonia includes the British Isles and north west Europe south of the Iapetus Suture, eastern Newfoundland, most of the Maritime Provinces of Canada, and parts of the eastern USA. Armorica includes the Armorican Massif of Normandy and Brittany, the Massif Central, and the Montagne Noire areas of France, together with parts of the Iberian Peninsula. Perunica comprises the area north of the Barrandian basin of Bohemia. The Taurides comprises most of central and southern Turkey.

magmatic history, must be incorporated in order to construct rational maps. **Figure 10** illustrates an Early Ordovician (490–470 Ma) reconstruction; it does not show all the continents or terranes that existed at this time, but some major players, such as Gondwana, Baltica, Laurentia, and some selected peri-Gondwanan terranes with palaeomagnetic and/or faunal data (Avalonia, Armorica, Perunica, and Taurides). Not all the Gondwanan continents have reliable palaeomagnetic data for this time. However, Gondwana was amalgamated at around 550 Ma and continental elements, such as South America and India (no palaeomagnetic data), remained attached to Africa until the breakup of the Pangea supercontinent during the Mesozoic.

An integrated approach of palaeomagnetic and faunal analysis is applicable for the entire Phanerozoic, but works best for the Early Palaeozoic and, notably, the Early Ordovician (**Figure 10**). At this time, Gondwana stretched from the south pole (Africa) to the equator (Australia and East Antarctica), Baltica occupied intermediate southerly latitudes, separated by the Tornquist Sea, whereas Laurentia straddled the equator. Laurentia was separated from both Baltica and Gondwana by the Iapetus Ocean that had opened in the Late Precambrian. The Iapetus Ocean (approximately 5000 km across the British sector) and the Tornquist Sea (approximately 1100 km between southern Baltica and Armorica–Perunica) were at their widest. This is probably why benthic trilobites from Laurentia (bathyurid) and north-west Gondwana (calymenacean–dalmanitacean) are so markedly different from those of Baltica (megalaspid).

In summary, palaeomagnetism can be seen to be the best and only quantitative method to establish the positions of old terranes and continents as they drifted across the globe over geological time.

See Also

Analytical Methods: Geochronological Techniques. **Gondwanaland and Gondwana**. **Magnetostratigraphy**. **Mantle Plumes and Hot Spots**. **Palaeozoic:** Ordovician. **Plate Tectonics**.

Further Reading

Butler RF (1992) *Paleomagnetism Magnetic Domains to Geologic Terranes*. Oxford: Blackwell Scientific Publications.

Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. *Journal of the Geological Society, London* 159: 631 644.

Cox A and Hart RB (1986) *Plate Tectonics How it Works*. Oxford: Blackwell Scientific Publications.

Dunlop DJ (1979) On the use of Zijderveld vector diagrams in multicomponent palaeomagnetic studies. *Physics of the Earth and Planetary Interiors* 20: 12 24.

Fisher NI, Lewis T, and Embleton BJJ (1987) *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.

Gordon RG, Cox A, and O'Hare S (1984) Paleomagnetic Euler poles and the apparent polar wander and absolute motion of North America since the Carboniferous. *Tectonics* 3: 499 537.

Kirschvink JL (1980) The least square line and plane and the analysis of palaeomagnetic data. *Geophysical Journal of the Royal Astronomical Society* 62: 699 718.

McElhinny MW and McFadden PL (2000) *Paleomagnetism, Continents and Oceans*. Cambridge: Academic Press.

Torsvik TH, Smethurst MA, Meert JG, *et al.* (1996) Continental break up and collision in the Neoproterozoic and Palaeozoic: A tale of Baltica and Laurentia. *Earth Science Reviews* 40: 229 258.

Van der Voo R (1993) *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*. Cambridge: Cambridge University Press.

PALAEONTOLOGY

L R M Cocks, The Natural History Museum, London, UK

Introduction

The word 'fossil' has no exact definition, but it is loosely taken to mean any organism whose remains or traces of remains are preserved in some kind of sediment. The term is derived from the Latin *fossare*, meaning 'to dig'. The definition implies nothing about age: fossils range from those that are about three billion years old to those that are preserved in lime-saturated water deposited only a few days or even hours ago. The study of fossils is termed palaeontology, which is a word derived from Greek that literally means 'knowledge of ancient things'; however, for more than two centuries the word 'palaeontology' has been restricted to the study of formerly living (organic) not inorganic remains or traces. The word palaeobiology is sometimes used as an alternative. In this encyclopaedia there follow numerous

articles on the various groups of fossils, and this article is designed simply so that the enquiring reader may find his or her own way through the ensuing contributions.

The Preservation and Condition of Fossils

Very few fossils are found that show no changes from the original organism. There is a very approximate correlation in many cases between the age of the fossil and the degree of change, but age is in itself no guide. For example, a fossil may be distorted by tectonic activity within a few hundred thousand years of its deposition, whilst other rocks and their contained fossils undergo little tectonic change over long periods. For example, the Lower Cambrian shale in the St Petersburg area of Russia, which is some 430 Ma old, is so soft that it can be dug with a spade, and the marine shells preserved within it can be washed out and look not dissimilar in preservation from those found on a beach today. The term diagenesis (*see* **Diagenesis, Overview**) describes the changes within both rocks and fossils as the sediments become dewatered and chemical transformations occur. Voids, from microscopic size upwards, present both in the original organism and in the surrounding sediment, are usually filled during diagenesis as a result of mineral-charged fluids circulating through the rock. This may occur at any time from soon after the fossil was deposited to many millions of years later, when the entombing rock is subjected to sedimentary pressure or tectonic events. Replacement of the chemicals in the fossils themselves is widespread; for example, the calcium carbonate ($CaCO_3$) of a brachiopod shell is often replaced by silica (SiO_2) during diagenesis.

Fossils may also be hugely distorted from the original shape of the organisms by tectonic processes. In addition the fossils may dissolve, and their chemical contents may form other substances; for example all 'fossil fuels', such as crude oil (*see* **Petroleum Geology:** The Petroleum System) and natural gas, are made up of the concentrated remains of fossil plankton, and coal is formed from the now usually unrecognizable remains of fossil plants.

Another process is termed disarticulation. Obviously, if an animal or plant has only a single hard part, such as the shell of a snail, then that may become weathered or broken, but there is still only the one shell. However, if an organism has more than one hard part (ranging from the two shells of a bivalved mollusc or brachiopod to the several hundred bones found in a mammal skeleton), then the process of transport from its place of death to its final burial place, where it will eventually form a fossil, will often cause the organism to break up or disarticulate. The degree of disarticulation often depends on how strong or weak the original articulation was; for example, some brachiopods have weak hinge structures, and others, such as the common Silurian genus *Atrypa*, have complex interlocking mechanisms between the two valves and are thus very commonly found fossilized with their two valves still closed together.

Classification of Organisms

Fossils and organisms living today are all included within the same system of classification: whether an organism is living or dead, or the species extinct, is not relevant to its position in the classification. Of course, all the various systems of classification, whilst attempting to reflect truly natural relationships and groupings, are actually man made and in the last analysis subjective, although objective measurements and methods of analysis have in many cases assisted the systematists in their analyses and helped them to reach their taxonomic conclusions. The term 'systematics' is used in a specialized sense by biologists and palaeontologists to describe the study and arrangement of organisms into classifications and hierarchies.

All organisms are classified within the binomial method; in other words, each organism or taxon is defined as a species within a genus. The binomial system was invented by the eighteenth-century Swedish naturalist Carl Linnaeus (or Linné). Linnaeus published many classificatory books and papers, and subsequent systematists have agreed that the 12th edition of his book *Systema Naturae*, published in 1758, should be the technical starting date for animal classification; however, there are several different starting dates for plant names. All generic names start with capital letters, and all species names are written in lower case: for example, our own genus is named *Homo* and our species name is *sapiens*. Although the generic and specific names are the formal minimum for all distinctively named organisms, some genera are divided into subgenera (which are therefore above the species level), and some species are divided into subspecies. For example, the English song thrush is the species *Turdus philomelos*, and those birds in my garden in London are the main subspecies *Turdus philomelos philomelos*, whilst the subspecies occurring only in northwest Scotland is *Turdus philomelos hebridensis*. Generic, subgeneric, specific, and subspecific names are normally printed in italic. Above the genus level, organisms are grouped within families, which may again be divided into subfamilies or grouped together as superfamilies. Above the family are the order, the class, the

phylum, and the kingdom. Plants are one kingdom, animals are another. Within the animal kingdom, the phylum Mollusca (*see* **Fossil Invertebrates:** Molluscs Overview), for example, is divided into the class Bivalvia (*see* **Fossil Invertebrates:** Bivalves) (including oysters), the class Cephalopoda (*see* **Fossil Invertebrates:** Ammonites; Cephalopods (Other Than Ammonites)) (which includes ammonites, nautiloids, and squids), the class Gastropoda (*see* **Fossil Invertebrates:** Gastropods) (snails), and various smaller classes with few included orders and families. Family names and above are not normally italicized in print. Families, genera, and species are subject to the Law of Priority, but orders and the higher taxa are not. This means that if a genus is named but is subsequently recognized as being within the same generic concept as another genus named earlier (but after 1758), then the later-named genus is placed within the synonymy of the earlier-named one.

Organisms are also classified, but less formally, according to their mode of life. They are obviously divided into marine (living in the sea), freshwater, and non-marine. Those that drift in the water without substantial self-propulsion are termed plankton, those that swim are termed nekton, and those that are restricted to the ocean floor are termed benthos. Benthic animals may be anchored to the seafloor (like most corals), move around on it (like most gastropods and trilobites), or burrow within the sediment (like most bivalves).

Trace Fossils

As stated in the introduction, the term 'fossil' includes not only the remains of an organism but also its traces preserved in the rock. These traces fall into two main groups. The first is when, for example, the shell of a bivalve is preserved in a sediment that is subsequently changed through diagenesis into a lithified rock, after which the shell vanishes, usually through the action of circulating fluids, leaving only the mould of the original creature's hard parts (in that case the shell) in the rock. That mould will, in favourable circumstances, faithfully reproduce the image of the shell (Figure 1). The second group of trace fossils are the burrows, tracks, feeding scrapes, and other biological traces preserved in soft sediment that has eventually hardened into rock. These are often found even though all the remains of the organism that caused the traces have entirely disappeared.

Human Understanding of Fossils

Fossils have been noticed by humans from time immemorial. Their use as grave goods dates back at least to the Neolithic, and superstitions abounded; for example, the incurved fossils of the bivalve *Gryphaea* were identified as toenails of the Devil, and the Chinese thought that dinosaur bones were the remains of dragons. The naturalist Gilbert White described the ammonites from the Cretaceous chalk of Selbourne in

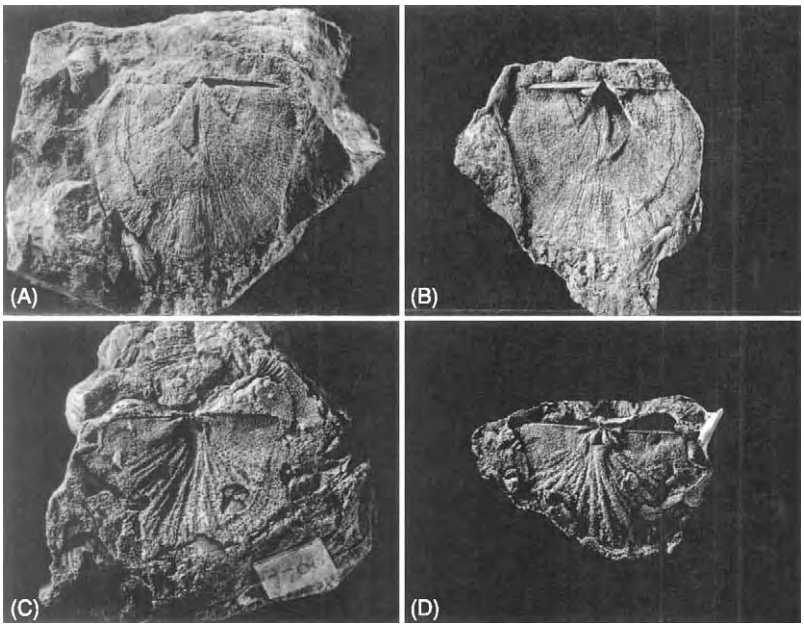


Figure 1 The (A,B) ventral (×1.5) and (C,D) dorsal (×2.0) valves of the Late Silurian brachiopod *Castelloarina fascifer* from Argentina. The views in (A) and (C) are natural internal moulds, as found in the geological outcrop, with the original carbonate shell entirely dissolved by the action of circulating groundwater. The views in (B) and (D) are artificial latex casts of the moulds, faithfully reproducing the topology of the original shell.

Hampshire, England, and retold the legend of St Keyna, of Keynsham near Bristol, who lived alone in a wood full of snakes, which she turned into stone (ammonites) by the power of prayer. It was usually thought that fossils represented dead specimens of living species of animals and plants, but it was not until the eighteenth century that it was realized that in many cases those animals and plants were now extinct. It was William Smith (*see* **Famous Geologists:** Smith) who realized that rocks could be dated in a relative way because different suites of fossils were found in rocks of successive ages. In the past 200 years all Phanerozoic (Cambrian to Holocene) rocks have been dated using fossils, and it is only since about 1960 that relatively reliable radioisotopic ages have been assigned to those zones. Even today, isotopic dates are not precise to within a million or so years over most of the Phanerozoic, whilst by interpolation many fossil zones are understood to have lasted for as little as 100 000 years. Thus, fossils have been extensively used to correlate rocks over considerable distances, and this knowledge has been of vast economic benefit in the exploration and discovery of coal and hydrocarbons, as well as being used to understand past Earth geographies.

Fossils in this Encyclopaedia

Animals are divided into those without backbones (invertebrates) and those with backbones (vertebrates). The chief Encyclopedia entries on fossils are arranged under the major headings of "Fossil Invertebrates", "Fossil Vertebrates", "Fossil Plants" and "Microfossils". Invertebrates are the most varied and commonest fossils. They are arranged here in order of increasing biological complexity. For about 50 years, each genus of fossil invertebrate has been illustrated and described in the 40 or so volumes of *The Treatise on Invertebrate Paleontology* (published jointly by the Geological Society of America and Kansas University Press between 1953 and 2004, and which is continuing), which is to be found in most geological libraries and in which much more detail on each group may be found; however, publication intervals between successive editions of the various parts of the *Treatise* are usually substantial. Despite the fact that fossils with original hard parts only occur in Cambrian and later rocks, a very substantial amount is now known about Precambrian fossils. In this Encyclopedia those are treated under **Origin of Life, Biosediments and Biofilms**, and also in the four articles within the Precambrian section, namely **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran.

There are also, at appropriate and separate places, articles on more general palaeontological and biological topics, namely **Biodiversity, Biological Radiations and Speciation, Creationism, Evolution, Fake Fossils, Lagerstätten, Palaeoecology, Palaeopathology, Pseudofossils** and **Trace Fossils**. In addition, the treatment of fossil specimens is considered in **Conservation of Geological Specimens** and **Micropalaeontological Techniques**. Of course the names of many fossil biozones are given in every one of the main stratigraphical articles on systems, which are grouped into Palaeozoic, Mesozoic, and Tertiary to Present, and also in other articles, particularly those on regional geology.

See Also

Biodiversity. **Biological Radiations and Speciation**. **Biosediments and Biofilms**. **Conservation of Geological Specimens**. **Creationism**. **Diagenesis, Overview**. **Evolution**. **Fake Fossils**. **Famous Geologists:** Smith. **Fossil Invertebrates:** Molluscs Overview; Bivalves; Cephalopods (Other Than Ammonites); Gastropods; Ammonites. **Lagerstätten**. **Micropalaeontological Techniques**. **Origin of Life**. **Palaeoecology**. **Palaeopathology**. **Petroleum Geology:** The Petroleum System. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran. **Pseudofossils**.

Further Reading

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman and Hall.

Benton MJ and Harper DAT (1997) *Basic Palaeontology*. Harlow: Addison Wesley Longman.

Briggs DEG and Crowther PR (eds.) (1990) *Palaeobiology: A Synthesis*. Cambridge: Blackwell Science.

Clarkson ENK (1998) *Invertebrate Palaeontology and Evolution*, 4th edn. Cambridge: Blackwell Science.

Fortey RA (2002) *Fossils: The Key to the Past*, 2nd edn. London: The Natural History Museum.

McKerrow WS (ed.) (1978) *The Ecology of Fossils: An Illustrated Guide*. London: Duckworth. MIT: USA.

Murray JW (ed.) (1985) *Atlas of Invertebrate Macrofossils*. London and New York: Longman and John Wiley.

Rudwick MJS (1992) *Scenes from Deep Time*. Chicago: University of Chicago Press.

Stewart WN and Rothwell AC (1993) *Paleobotany and the Evolution of Plants*, 2nd edn. Cambridge: Cambridge University Press.

Willis KJ and McElwain JC (2002) *The Evolution of Plants*. Oxford: Oxford University Press.

PALAEOPATHOLOGY

S G Lucas, New Mexico Museum of Natural History, Albuquerque, NM, USA

Introduction

In 1913, the British scientist, Sir Armand Ruffer, originally defined palaeopathology as "the science of diseases which can be demonstrated in human and animal remains of ancient times". Palaeopathology thus encompasses two distinct areas, one palaeoanthropological, in which ancient human remains are examined, and the other more strictly palaeontological, in which non-human animal fossils are examined. Furthermore, not only disease, but also evidence of injuries, is normally studied by palaeopathologists. Moreover, pathological conditions are not only known in fossil animal remains, but have also been reported in fossil plants. Here, discussion focuses on the strictly palaeontological aspects of palaeopathology. Palaeopathology is also used here as a noun to refer to any evidence of disease or injury in a fossil.

Assumptions and Methods

Palaeopathology relies on a straightforward uniformitarian assumption, namely that the manifestations of diseases have been relatively stable through time. Thus, for example, a disease that produces a particular skeletal abnormality at present would have done just the same millions of years ago. If this is the case, then it will be possible to diagnose medically (determine the underlying cause of) a palaeopathology.

Also critical to palaeopathology is the need to distinguish an abnormal structure in a fossil, caused by disease or injury, from the effects of post-mortem damage, the processes of rock formation and fossilization, and of erosion and weathering. Abnormal structures in fossils are sometimes caused by these effects, and have been mistakenly identified as manifestations of pathology.

Once a pathological condition is identified in a fossil, diagnosis is strictly by analogy. The palaeopathology is compared with similar pathologies in extant plants and animals, and a match in pathologies suggests identical underlying causes in the living organism (for which we know the cause) and the fossil (for which we infer the cause). Bacteria and viruses that cause disease are almost never preserved in fossils, and so the 'diagnosis by analogy' method is the principal way to determine the underlying causes of palaeopathologies.

Palaeopathologies in Fossil Plants

Pathologies caused by diseases have been identified in fossil leaves and wood. These include evidence of bacterial and fungal infections of ancient plants (*see* **Fossil Plants:** Fungi and Lichens). Most important, however, in the study of fossil plant pathologies is the fossil record of predation and infestation of plants by insects.

This record of predation and infestation is seen in damage of the fossil plant, and specific types of damage include:

- leaf mining – the extraction of living leaf tissue by larvae or mites;
- galling – when an insect or mite inserts an egg into or burrows into plant tissue, so that an abnormal, three-dimensional expansion of the plant tissue (a gall) is formed;
- wood boring – when insects burrow into wood;
- piercing and sucking – when insects penetrate plant tissue to remove fluid;
- sporivory, pollinivory, and nectarivory – consumption by insects of spores, pollen, and nectar, respectively.

The most significant aspect of these palaeopathologies is that they provide direct evidence of plant–insect interaction in the fossil record. This enables palaeontologists to determine the history and intensity of insect herbivory in the fossil record, and allows inferences to be made about the evolution of plant defences against such herbivory.

Palaeopathologies in Fossil Invertebrates

Evidence of deformities and other abnormalities in fossil shells is attributed to disease or genetic defects. Evidence of parasitism on the shells of stationary Palaeozoic animals, such as crinoids (sea lilies), is also well documented. Amongst the fossil invertebrates, there has been extensive study of healed injuries due to predation, especially in fossil clams (bivalves) and snails (gastropods). This evidence of predation (Figure 1) is particularly significant as it allows palaeontologists to understand the history and evolution of attack and defence behaviours in fossil invertebrates.

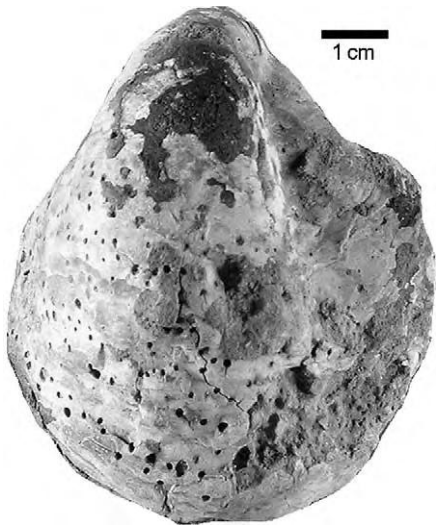


Figure 1 Shell of a 100 million year old clam (*Texigryphaea*) from eastern New Mexico covered with tiny holes, which are the drill marks left by snails (gastropods) that attempted to prey on the clam.

Palaeopathologies in Fossil Vertebrates

The majority of palaeopathologies (and the vast majority of the literature on them) are found in the teeth and bones of extinct vertebrates, especially reptiles and mammals. The most commonly identified palaeopathologies are: (1) so-called arthritic conditions, although many are not demonstrably arthritis; (2) mechanical injuries (Figure 2) and related infections, and the healing of these injuries if the animal survived them; (3) wounds due to predation or combat between animals; (4) dental anomalies, including caries (tooth decay); and (5) bone lesions caused by infections and/or parasites.

A list of the most common skeletal pathologies of fossil vertebrates includes:

- arthritis – joint diseases that can be caused by infections or mechanical injuries;
- dental anomalies – rotated, misplaced, malformed, or extra teeth (Figure 3);
- dental caries – cavities (decay) in the teeth;
- exostosis – swellings on bone surfaces;
- fracture – broken bones which, when naturally healed, leave a bone callus (a patch of healed bone) (Figure 2);
- hypertrophy – increase in size of the bone or portions of the bone;
- necrosis – pitting and disintegration of dead bone;
- osteomyelitis – bone infections;
- osteoma – bone tumours;
- osteoporosis – loss of bone density;

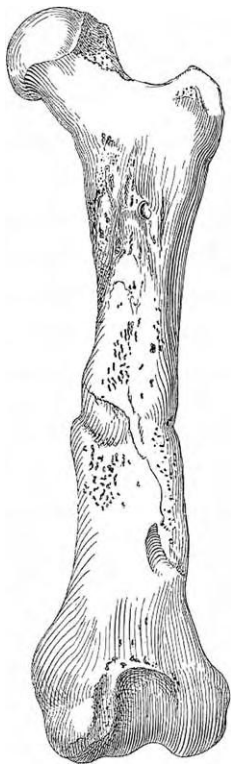


Figure 2 Anterior view of the left femur (thigh bone) of a Pleistocene cave bear from Europe, showing a healed fracture of the bone shaft. The pits above and below the fracture may also indicate necrosis. From Moodie RL (1930), Fig. 5.

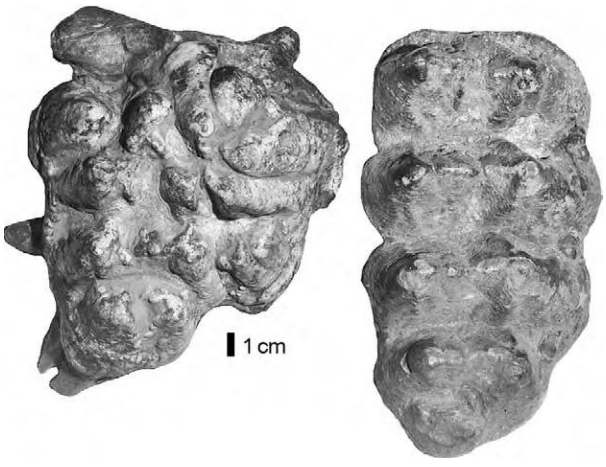


Figure 3 Left: crown view of a pathological molar of an American mastodon (*Mammut americanum*) from the Pleistocene of Oklahoma. Right: normal mastodon molar for comparison. On the pathological molar, there is a normal shaped molar (left side) covered by a second (right side), highly deformed molar. This secondary (or supernumerary) tooth probably resulted from the fusion of two tooth buds where, normally, only one tooth bud would have been present.

- synostosis – co-ossification (fusion) of normally separate bones.

Dinosaur Palaeopathology

The study of palaeopathology in fossil vertebrates is very active, with many new cases documented in the technical literature each year. This is especially true of dinosaur palaeopathology, which, during the last two decades, has become the largest field of palaeopathological research on non-human fossils. Prior to this interest, reports of palaeopathologies in dinosaurs go back to the 1800s, but they were mostly of isolated curiosities that often led to unfounded speculation. A good example of this is the unsupported suggestion that the frills (head shields) of horned dinosaurs (ceratopsians) grew so large because the dinosaur had too much growth hormone.

More recent studies of dinosaur palaeopathology have examined large samples (populations) using stringent standards for disease recognition. They have been accompanied by attempts to identify blood cells, DNA, and disease pathogens in dinosaur bone. These studies have identified a variety of common palaeopathologies in dinosaur skeletons. Most obvious are broken and healed bones (Figure 4), fused bones, and certain types of bone tumour. These palaeopathologies are well known in horned dinosaurs (ceratopsians), duck-billed dinosaurs (hadrosaurs), and tyrannosaurs. Indeed, the famous skeleton of *Tyrannosaurus rex* nicknamed 'Sue', has many healed injuries in its bones. The new studies of dinosaur palaeopathology demonstrate that dinosaur skeletons show much evidence of injury and trauma, but little evidence of disease.

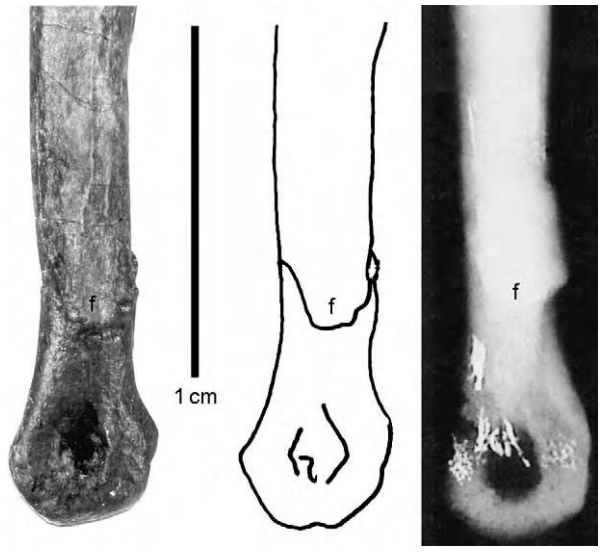


Figure 4 Toe bone (metatarsal) of a small, meat eating dinosaur from the Upper Cretaceous of New Mexico (about 73 million years old) with a healed fracture (f) visible in the photograph (left) and the X ray (right). Courtesy of R. M. Sullivan.

Application of Palaeopathology

As the above paragraphs make clear, the study of palaeopathology provides important information about more than just the history of disease. It also provides direct evidence of interaction between plants and animals, and between animals, mostly between predators and prey. It is thus important to the understanding of the evolution of behaviour. Palaeopathology can also contribute to an understanding of palaeoecology. For example, leaf mining by insects is extremely intense in today's tropics, and the intensity of fossil leaf mining has been used to infer the presence of ancient tropical conditions. Palaeopathology is thus an important subject in its own right, and a valuable tool for inferring ancient behaviour and ecology from the fossil record.

Although observations of palaeopathologies have been made for nearly two centuries, it is fair to say that the study of palaeopathology remains a relatively understudied field. Disease-causing agents generally do not fossilize, and so their history, which would be gleaned directly from the fossils of bacteria and viruses, is still very poorly understood. New technologies promise the ability, at least in some cases, to identify fossil bacteria and viruses associated with palaeopathologies.

Most diseases do not affect the fossilizable parts of an animal (its skeleton), and so much about ancient disease may never be extracted from the fossil record. Most published records of palaeopathologies are of isolated occurrences, often questionably diagnosed. Surveys that provide an idea of the frequency of palaeopathologies are almost unknown. Moreover, sweeping claims about the role of disease in the past – such as the idea that disease caused dinosaur extinction – lack supporting evidence. What is needed, and what is being pursued today, are more extensive efforts to identify palaeopathologies in all kinds of fossils, coupled with more rigorous attempts at diagnosis.

See Also

Fossil Plants: Fungi and Lichens. **Fossil Vertebrates:** Dinosaurs.

Further Reading

Baker J and Brothwell D (1980) *Animal Diseases in Archaeology*. London: Academic Press.

Brothwell DR (1972) *Digging up Bones*. British Museum (Natural History), London.

Jarcho S (ed.) (1966) *Human Palaeopathology*. New Haven: Yale University Press.
Labandeira C (1998) The role of insects in Late Jurassic to Middle Cretaceous ecosystems. *New Mexico Museum of Natural History and Science Bulletin* 14: 105 124.
Moodie RL (1923) *Paleopathology: An Introduction to the Study of Ancient Evidences of Disease*. Urbana: University of Illinois Press.
Pindborg JJ (1970) *Pathology of the Dental Hard Tissues*. Philadelphia: W. B. Saunders.
Rothschild BM and Martin LD (1993) *Paleopathology: Disease in the Fossil Record*. Boca Raton, FL: CRC Press.
Tanke DH and Rothschild BM (2002) Dinosaurs: an anno tated bibliography of dinosaur paleopathology and related topics 1838 2001. *New Mexico Museum of Natural History and Science Bulletin* 20: 1 96.
Vermeij GJ (1987) *Evolution and Escalation: An Eco logical History of Life*. Princeton: Princeton University Press.
Wilf P, Labandeira CC, Johnson KR, Coley PD, and Cutter AD (2001) Insect herbivory, plant defense, and early Cenozoic climate change. *Proceedings of the National Academy of Sciences of the USA* 98: 6221 6226.

PALAEOZOIC

Contents

Cambrian

N C Hughes, University of California, Riverside, CA, USA
N A Heim, University of Georgia, Athens, GA, USA

Introduction

The Cambrian period, the oldest division of the Phanerozoic time-scale, encompasses several important global-scale environmental and evolutionary transitions. Recent advances in the understanding of Cambrian environments, faunas, and chronostratigraphy have all provided an enriched view of this critical interval in the history of the Earth – the time during which life's direct impact upon the rock record first became explicit. The term 'Cambrian' was first applied by Sedgwick (*see* **Famous Geologists:** Sedgwick) in 1835 to a series of sedimentary rocks well exposed in north Wales. The Cambrian system, as defined today, equates to only a portion of what Sedgwick originally considered to be 'Lower Cambrian'. Cambrian deposits are distributed widely, and a basic stratigraphical and palaeogeographical framework has been established for almost all regions. Nevertheless, a great deal of additional research will be necessary to secure our understanding of the geological framework in which the events of the Cambrian were structured. This article presents an overview of the defining temporal, geographical, and environmental attributes of the Cambrian, and then considers these in relation to the most striking aspect of the Cambrian system – the diversification of abundant metazoan life.

Chronostratigraphical Framework

The ratification of formal boundaries for the Cambrian system and improvements in the precision of isotope geochronology have recently resulted in substantial adjustments to views concerning the absolute age and duration of the Cambrian period. These results have fundamentally changed perspectives on the historical events encompassed within the Cambrian.

The Bounds of the Cambrian System

Formal definitions for both the base and the top of the Cambrian system have been ratified within Global Standard Stratotype Sections and Points (GSSPs) in recent years, both with type sections in Newfoundland, Canada. The base of the Cambrian system is currently placed at the occurrence of the trace fossil, *Treptichnus pedum*, 2.4 m above the base of Member 2 of the Chapel Island Formation at Fortune Head, Newfoundland. The base of the subsequent geological system, the Ordovician system, is placed at the first occurrence of the conodont, *Iapetognathus fluctivagus*, at a level of 101.8 m, within Bed 23, in the measured section of the Lower Broom Point Member, Green Point Formation at Green Point, western Newfoundland.

Stratotype points are located at the occurrence of diagnostic markers within particular stratotype sections, and both the top and the base of the Cambrian system rely on the occurrence of fossils. Such boundaries commonly relate to the stratigraphically lowest occurrence of the diagnostic marker fossil in the stratotype section. In the case of the base of the Cambrian, this is not so, because unequivocal specimens of *Treptichnus pedum* have recently been described some 4 m below the GSSP at the stratotype section. The occurrence of other *Treptichnus* specimens still lower in this faulted section could compromise the utility of this section as a stratotype. Nevertheless, the diversification of trace fossils during the Precambrian–Cambrian boundary interval was a significant evolutionary event, and offers the possibility of global correlations based on the impressions of homologous morphological characteristics.

The absolute ages of both the base and the top of the Cambrian system rely on recent U–Pb zircon dating of volcanic deposits intercalated within sedimentary successions. Suitable volcanic deposits are not known from the critical intervals of either of the stratotype sections. Hence, all absolute age estimates of the boundaries of the Cambrian rely on radioisotopic dates obtained from other sections that have been correlated with the stratotypes using alternative criteria, such as fossils or geochemical signatures. The base of the Cambrian system is commonly considered to be at about 542 Ma. This view is based on a minimum date for the first appearance of *Treptichnus pedum* of 539.4 ± 1.0 Ma in the Swartkloofberg section of Namibia, and a maximum age of 543.3 ± 1.0 Ma based on the occurrence of an assemblage of Ediacaran fossils in the same section. An age of 542 Ma also corresponds with a sharp negative excursion in carbon isotope ratios found in several parts of the world. Nevertheless, Ediacaran fossils have recently been recovered from rocks bearing *Treptichnus pedum* in South Australia, and the exact correlation of the Namibia section with the Newfoundland stratotype awaits resolution. The maximum age of the Cambrian–Ordovician boundary has recently been estimated from latest Cambrian volcaniclastic sandstone beds in north Wales to be 489 ± 0.6 Ma. The minimum age is 483 ± 1.0 Ma based on a Tremadocian ash from Nova Scotia.

Stratigraphical Subdivision of the Cambrian System

Formal stratigraphical stages have been established for Cambrian sequences around much of the globe. These sequences are correlated using cosmopolitan fossils, isotopic signatures, and correlative stratigraphical surfaces. Certain key horizons, characterized by distinctive and widespread markers, can be correlated with some accuracy around the globe. Other portions of the Cambrian are less securely correlated due to such factors as faunal endemism (partly real, partly the result of provincial biases in taxonomic descriptions) or questionable alignments of isotopic profiles. All global correlation schemes for Cambrian stratigraphical sequences (e.g., **Figures 1** and **2**) should be viewed as provisional.

The Cambrian period is commonly divided into three portions (**Figure 3**), although the global utility and bounds of these divisions have yet to be formalized. The perception of the nature of the Cambrian period has changed dramatically recently. The decision to place the base of the system near the first occurrence of *Treptichnus pedum* has meant that the first appearance of trilobites (and some other skeletonized macrofossils) occurs after about one-third of Cambrian time had passed. Furthermore, the discovery that the oldest trilobites date from around 520 Ma has dramatically reduced the apparent duration of the trilobite-bearing Cambrian to an interval of approximately 35 million years in duration. This suggests that the biostratigraphical zones of trilobite-bearing portions of the Cambrian are of a time resolution comparable to those of Mesozoic ammonites, on average as short as 1 million years.

Cambrian Palaeogeography

Palaeogeographical reconstructions for the Cambrian, constructed by different earth scientists, generally agree in terms of the overall distribution of the major landmasses, although doubts are commonly expressed about the quality of the palaeomagnetic database used to underpin such reconstructions. The assembly of the core Gondwana components is

generally considered to have been completed prior to the onset of the Cambrian. The positions and movements of many of the 'outboard' terranes that jostled about the margins of Gondwana, however, remain poorly constrained. This is particularly true of the equatorial sector of Gondwana, where faunal provinciality – commonly useful for regional demarcation – was low. Overall, Cambrian palaeogeography was characterized by open water in the northern polar region, and dominated by the major landmass, Gondwana, that stretched from the lowest latitudes in the south, across the equator, and well into the northern hemisphere. Other major landmasses, such as Laurentia, Siberia, and Baltica, were equatorial or lay within the southern hemisphere (Figure 4).

With the exception of the movements of minor terranes, global palaeogeography remained broadly consistent throughout the Cambrian period, with most of the changes apparently related to the rotation and dispersion of Laurentia, Siberia, and perhaps Baltica relative to a stable Gondwana. These movements, coupled with the patterns of sedimentation, are consistent with the breakup of a late Precambrian supercontinent that has been named Pannotia. Details of late Neoproterozoic palaeogeography remain hotly debated and these debates have important implications for interpreting secular changes during Cambrian time.

Cambrian Environments and Climates

The deposits of most major palaeocontinents record a substantial transgression during much of the Cambrian. Such a pattern is clearly displayed in Laurentia, where deposits encroached progressively further into the interior of the continent. Cambrian sedimentation patterns are consistent with relatively high rates of thermal subsidence, perhaps related to the fragmentation of Pannotia. In equatorial settings, particularly within Gondwana, latest Neoproterozoic carbonate platforms were drowned by relatively rapid flooding events that were followed by the deposition of phosphorites and black shales. The shoreward spread of oxygen-depleted and nutrient-enriched waters onto the shelf was associated with the replacement of fossils by secondary phosphates and within cherts. Substantial evaporite deposits developed within interior basins. In the absence of evidence of persistent glaciation, the Cambrian transgression is inferred to have been related to high rates of seafloor spreading.

The analysis of secular changes in the stable isotopes of strontium and carbon provides supplementary tools for correlation amongst sections, together with insights into the evolution of Cambrian environmental conditions (Figure 5). The striking rise in Cambrian stable isotopic ratios is thought to reflect enriched ^{87}Sr values due to the uplift and erosion of cratonic material. This rise in $^{87}Sr/^{86}Sr$ ratios is comparable with that seen in the Neogene, associated with the uplift of the Himalaya. Events associated with the latest stages of the prolonged 'Pan-African' orogeny (*see* **Africa:** Pan-African Orogeny) and the final consolidation of Gondwana may have been responsible, but the sequence and timing of these events remain incompletely resolved.

Investigations of Cambrian seawater chemistry, both directly from fluid inclusions and indirectly from inorganic and organic precipitates, suggest that Cambrian seawaters were generally characterized by relatively low Mg/Ca ratios. This is consistent with the deposition of calcitic, as opposed to aragonitic, biomineralized metazoan skeletons. The late Proterozoic to Cambrian interval witnessed the transition from an 'aragonite sea' to a 'calcite sea', corresponding to a transition from 'icehouse' to 'greenhouse' climatic conditions, and a change from $MgSO_4$ to KCl late-stage salts in evaporites. When viewed in the context of trends seen throughout the Phanerozoic, such transitions correspond with long-term changes in plate tectonic regime. The rapid seafloor spreading associated with the breakup of Pannotia probably drove changes in mid-ocean ridge brine fluxes, caused a global sea-level rise, and increased global volcanism.

Marked fluctuations in Cambrian carbon isotopic ratios over short time-scales have been noted in association with the base of the Cambrian and the Early–Middle Cambrian boundary. In these cases, the excursions were towards lighter carbon isotopes. A more prolonged, positive excursion towards the heavier isotope, ^{13}C, has been defined within the Late Cambrian Steptoean stage within Laurentia, Kazakhstan, and portions of core and outboard Gondwana. The determinants of such fluctuations are complex because carbon isotopic ratios are influenced by multiple factors. Explanations commonly invoked include alternations between periods of well-oxygenated, warmer waters on continental shelves and periods in which the shelves were flooded by cooler, oxygen-poor waters. These explanations receive additional support from marked periodic fluctuations in macrofaunal diversity that roughly parallel the excursions, at least in some cases. Cambrian oceans were similar to oceans throughout the Phanerozoic in that oxygen penetrated the sediment–water interface even within deep ocean basins. Recent models of atmospheric oxygen levels throughout the Phanerozoic suggest that O_2 constituted between 10% and 20% of the composition of the atmosphere. Although these estimates are subject to considerable uncertainty, they suggest that Cambrian

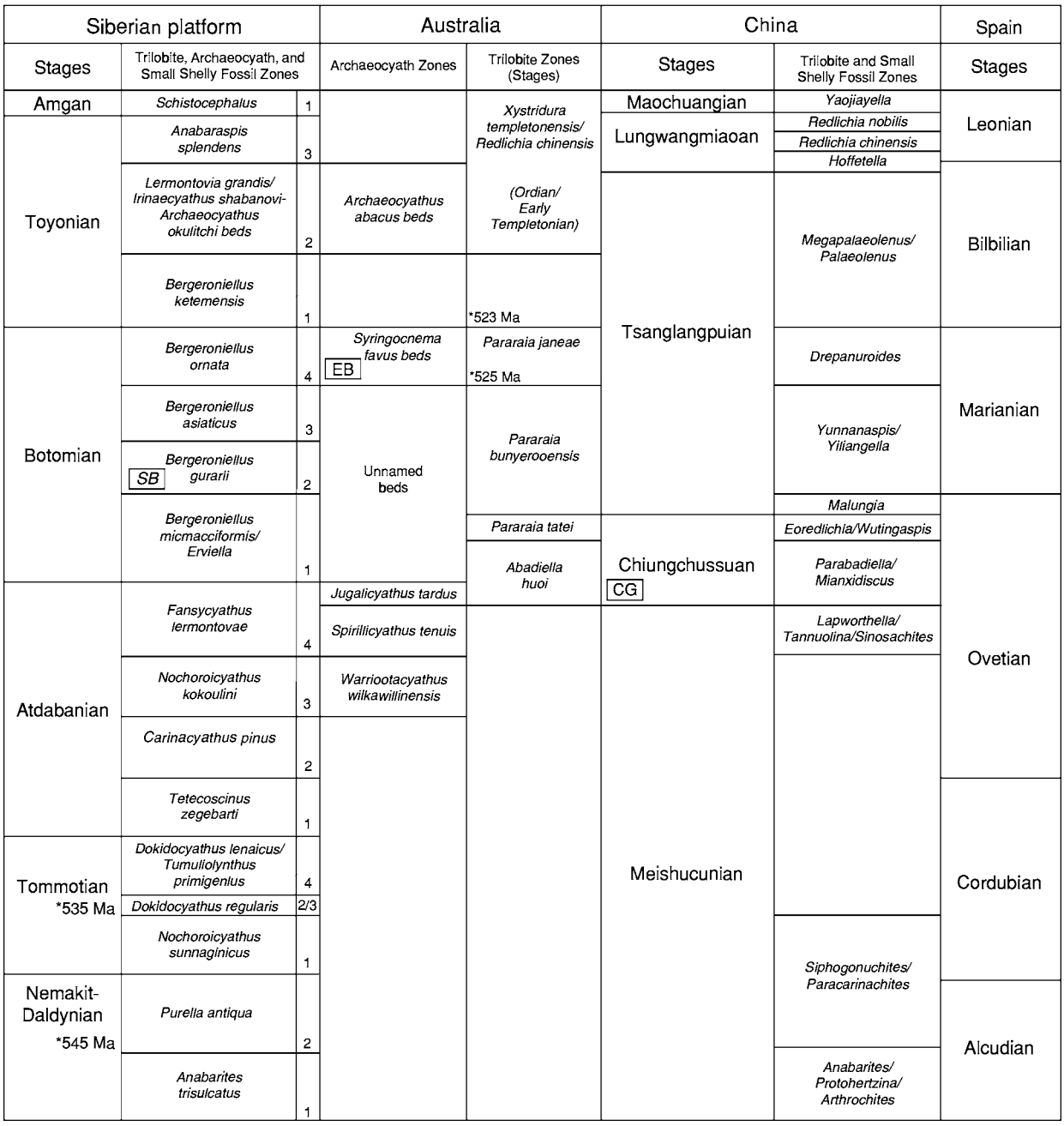


Figure 1 Correlation chart for the major regions of Lower Cambrian rocks. CG, Chengjiang fauna; EB, Emu Bay Shale; MC, Mount Cap Formation; SB, Sinsk fauna; SP, Sirius Passet fauna; RW, Ruin Wash fauna. Modified with permission from Zhuravlev AY and Riding R (2000) *The Ecology of the Cambrian Radiation*. New York: Columbia University Press. © Columbia University Press, 2000.

oxygen profiles and levels in both ocean and atmosphere were essentially modern in aspect.

As noted above, the Cambrian marked a transition interval between 'icehouse' and 'greenhouse' climatic conditions. The details of this transition are poorly resolved, however. Direct evidence of warm conditions in the later Cambrian, derived from oxygen stable isotope ratios, is questioned by some authorities. Yet, the presence of widespread evaporites suggests warm conditions, at least in the equatorial Gondwana region. Broad, but indirect, climatic indicators, such as the distributions of evaporites, ironstones, and carbonates, are consistent with climatic gradients similar to those that characterized much of the Palaeozoic. Recent suggestions of Early Cambrian glacial deposits in West Africa are yet to be accepted generally, but it is likely that the overall climatic conditions warmed through the Cambrian.

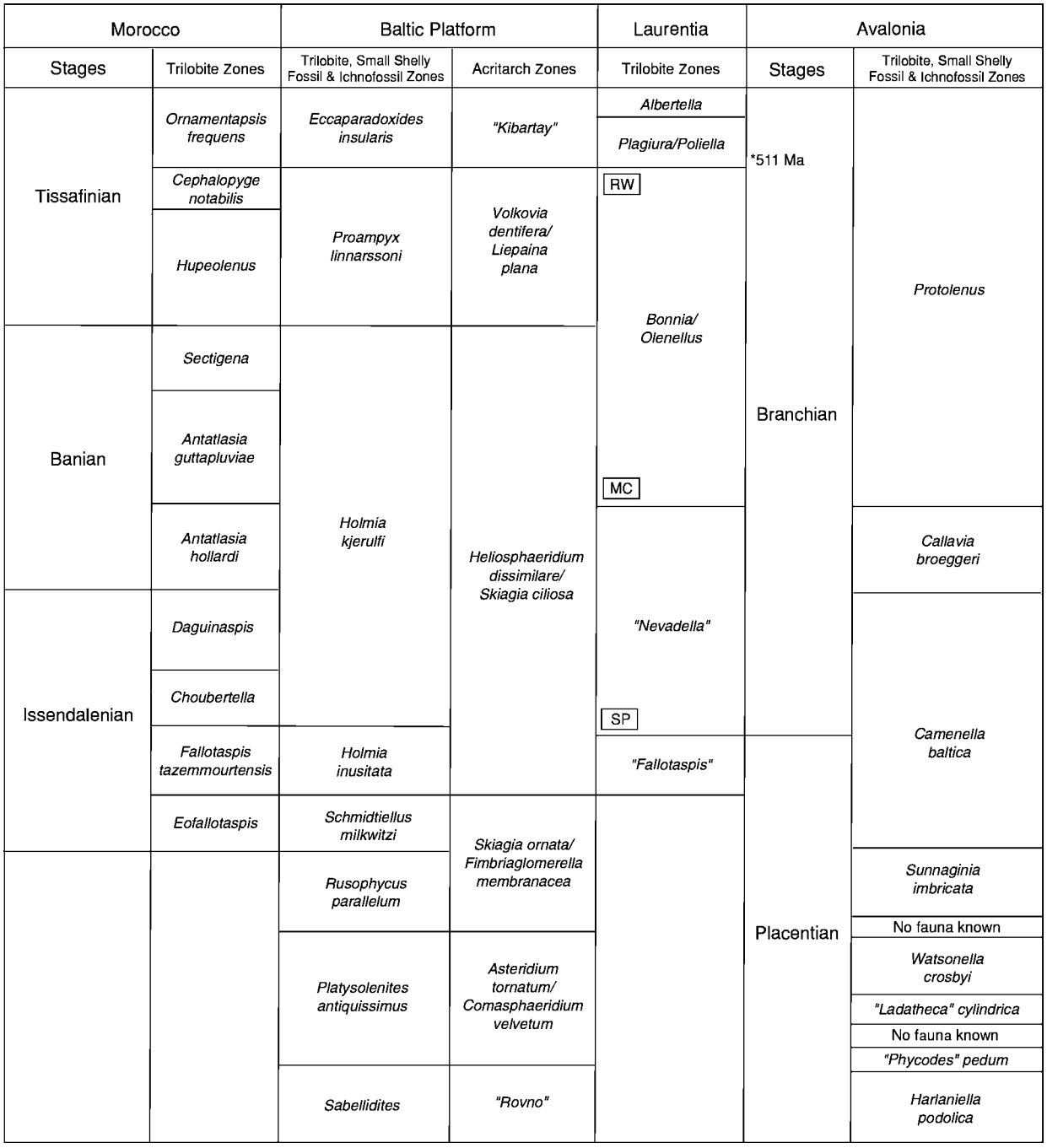


Figure 1 Continued

Information on Cambrian terrestrial conditions is sparse, and the transition between marine and terrestrial sedimentary rocks is difficult to determine in many Cambrian sequences, partly because fine-grained sedimentary particles were easily winnowed in the absence of binding by macroflora. Definitive non-marine sedimentary structures are relatively rare within Cambrian sandstones. Nevertheless, the common tacit assumption of an abiotic terrestrial realm during the Cambrian is probably unrealistic. A diverse microbiota may have existed in some settings. Furthermore, there is abundant trace fossil and sparse body fossil evidence of both diploblastic and triploblastic metazoans active in marginal marine environments, together with spores that are apparently non-marine in origin.

Due to the decreasing rotation speed of the Earth, the Cambrian year would have contained more days

Kazakhstan & Siberia			Australia		China
Ungurian	*Dilelokephalina*	1	Warendian	*Cordylodus lindstromi*	Xingchangian
Batyrbayan	*Euloma limitaris/ Batyraspis*	1	Datsonian	*Cordylodus prolindstromi*	
				Hirsutodontus simplex	
				Cordylodus proavus	
	Lotagnostus hedini	3	Payntonian	*Mictosaukia perplexa*	Fengshanian
	Harpidoides/Troedsonia				
	Lophosaukia			*Neoagnostus quasibilobus/ Shergoldia nomas*	
	Trisclcagnostus trisulcus			*Sinosaukia impages*	
Aksayan	*Eolotagnostus scrobicularis*	2	Iverian	*Lophosukia*	Changshanian
				Rhaptagnostus clarki prolatus/ Cazbnaia sectarix	
	Neoagnostus quadratiformis	1		*Rhaptagnostus clarki patulus/ Caznaia squamosa/ Hapsidocare lilyensis*	
	Oncagnostus ovaliformis			*Peichiashania tertia/ Peichiashania quarta*	
	Oncagnostus kazachstanicus			*Peichiashania secunda/ Peichiashania glabella*	
	Pseudagnostus pseudangustilobus			*Wentsua iota/ Rhaptagnostus apsis*	
Sakian	*Ivshinagnostus ivshini*	3		*Irvingella tropica*	
	Pseudagnostus "curtare"	2	Idamean	*Stigmatoa diloma*	
	Oncagnostus longifrons	1		*Erixanium sentum*	
				Proceratopyge cryptica	
	Glyptagnostus reticulatus			*Glyptagnostus reticulatus*	
Aysokkanian	*Glyptagnostus stolidotus*	6	Mindyallan	*Glyptagnostus stolidotus*	Kushanian
	Agnostus pisiformis	5		*Acmarhachis quasivespa*	
Mayan	*Leiopyge laevigata/ Aldanaspis truncata*	4	Boomerangian *495 Ma	*Erediaspis eretis*	Changhian
				Damesella torosa/ Ascionepea jantrix	
	Anomocarioides limbataeformis	3		*Holteria arepo*	
				Proampyx agra	
				Ptychagnostus cassis	
	Anopolenus henrici/ Corynexochus perforatus	2	Undillan	*Goniagnostus nathorsti*	
				Doryagnostus notalibrae	
				Ptychagnostus punctuosus	
			Late Templetonian/ Floran	*Euagnostus opimus*	
Amgan	*Pseudanomocarina*	3		*Acidusus atavus*	Hsuchuangian
				Triplagnostus gibbus	
	Kounamkites	2		*Xystridura templetonensis/ Redlichia chinensis*	Maochuangian KF
	Schistocephalus	1			

Figure 2 Correlation chart for the major regions of Middle and Upper Cambrian rocks. Global correlations for the top of the Cambrian System are not finalized, but the top of the Sunwaptan Stage of North America approximates the top of the Cambrian in these figures. Rocks above this belong to the Lower Ordovician. BS, Burgess Shale; KF, Kali fauna; MF, Marjum Formation; OR, *orsten*; WF, Wheeler Formation. Modified with permission from Zhuravlev AY and Riding R (2000) *The Ecology of the Cambrian Radiation*. New York: Columbia University Press. © Columbia University Press, 2000.

China *(cont.)*	Scandinavia		North America (Laurentia)	
Yosimuraspis	*Rhabdinopora*	*Rhabdinopora flabelliforme*	*Symphysurina*	Ibexian
Richardsonella/ Platypeltoides	*Acerocare*	*Acerocare ecorne*		
		Westergaardia		
		Peltura costata		
		Peltura traniens		
Missisquoia perpetis	*Peltura*	*Peltura scarabaeoides*	*Missisquoia*	
			Eurekia apopsis	
Mictosaukia cf. M. orientalis			*Saukiella serotina*	Sunwaptan
Tsinania/Ptychaspis			*Saukiella junia*	
Kaolishania pustulosa		*Peltura minor*	*Saukiella serotina/ Rasettia magna*	
	OR *492 Ma	*Protopeltura praecursor*		
	Leptoplastus	*Leptoplastus stenotus*	*Ellipsocephaloides*	
		Leptoplastus angustatus		
		Leptoplastus ovatus		
		Leptoplastus crassicorne		
		Leptoplastus raphidophorus	*Idahoia*	
		Leptoplastus paucisegmentatus		
	Parabolina	*Parabolina spinulosa*		
Maladioidella		*Parabolina brevispina*	*Taenicephalus*	
Changshania conica	*Olenus*	*Olenus scanicus*	*Elvinia*	Steptoan
		Olenus dentatus		
		Olenus attenuatus	*Dundenbergia*	
		Olenus wahlenbergi	*Aphelaspis*	
Chuangia batia	OR	*Olenus truncatus*		
		Olenus gibbosus		
Drepanura	*Agnostus pisiformis*	*Agnostus pisiformis*	*Crepicephalus*	Marjuman
Blackwelderia	OR		*Cedaria*	
Damesella/Yabeia	*Paradoxides forchhammeri*	*Lejopyge laevigata*		
Leiopeishania			*Bolaspidella*	
Taitzuia/Poshania		*Jinsella brachymetopa*		
Amphoton		*Goniagnostus nathorsti*		
	Paradoxides paradoxissimus	*Ptychagnostus punctuosus*		
Crepicephalina		*Hypagnostus parvifrons*		
Bailiella/Lioparia		*Tomagnostus fissus/ Acidiscus atavus*		MF / WF
Poriagraulos/ Hsuchuangia/Ruichengella		*Triplagnostus gibbus*	*Ehmaniella*	
Shantungaspis	*Eccaparadoxides oelandicus*	*Eccaparadoxides pinus*	*Glossopleura*	
Yaojiayella			*Albertella*	BS

Figure 2 Continued

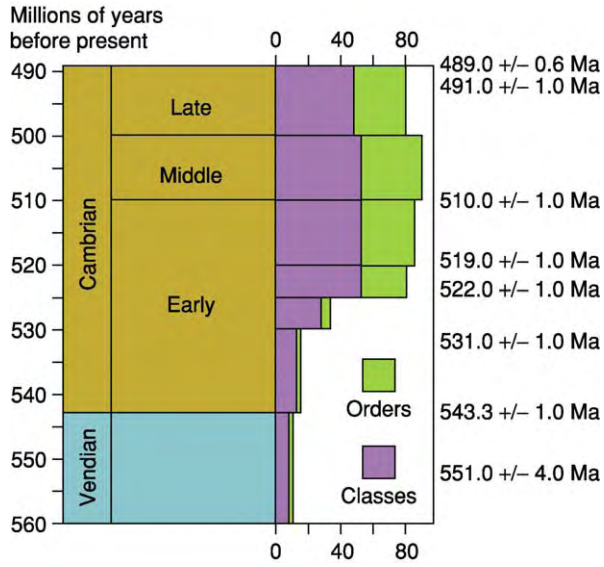


Figure 3 A broad chronology of the Cambrian, and part of the underlying Precambrian Vendian period, showing the major stratigraphical divisions of the system that are commonly used, an approximation of the diversity of metazoan order and classes, and some radiometric dates. Modified with permission from Bowring SA and Erwin DH (1998) A new look at evolutionary rates in deep time: uniting paleontology and high precision geo chronology. *GSA Today* 8: 1 7.

than at present, and each day would have been shorter than 24 h. The Cambrian year was likely to have been about 420 days long. If the shorter distance between the Earth and its moon resulted in a stronger tidal action during the Cambrian, as might also be expected, Cambrian epeiric seas may have been dominated by tides. However, there is no observational support for increased tidal dominance in the Cambrian relative to comparable recent settings. Overall, Cambrian environmental conditions were apparently primarily controlled by the global tectonic setting. The dispersion of the continents, rates of thermal subsidence, and seafloor spreading strongly influenced both the accumulation of sedimentary rocks and their preservation potential. Cambrian environmental conditions were similar to those seen at other parts of the Phanerozoic and essentially modern in aspect. In marked contrast, several aspects of ocean chemistry and stratification during the Neoproterozoic apparently differed significantly from the Phanerozoic conditions. Hence, the Neoproterozoic–Cambrian interval remains a fascinating and important time of environmental transition, in addition to encompassing critical biotic transitions.

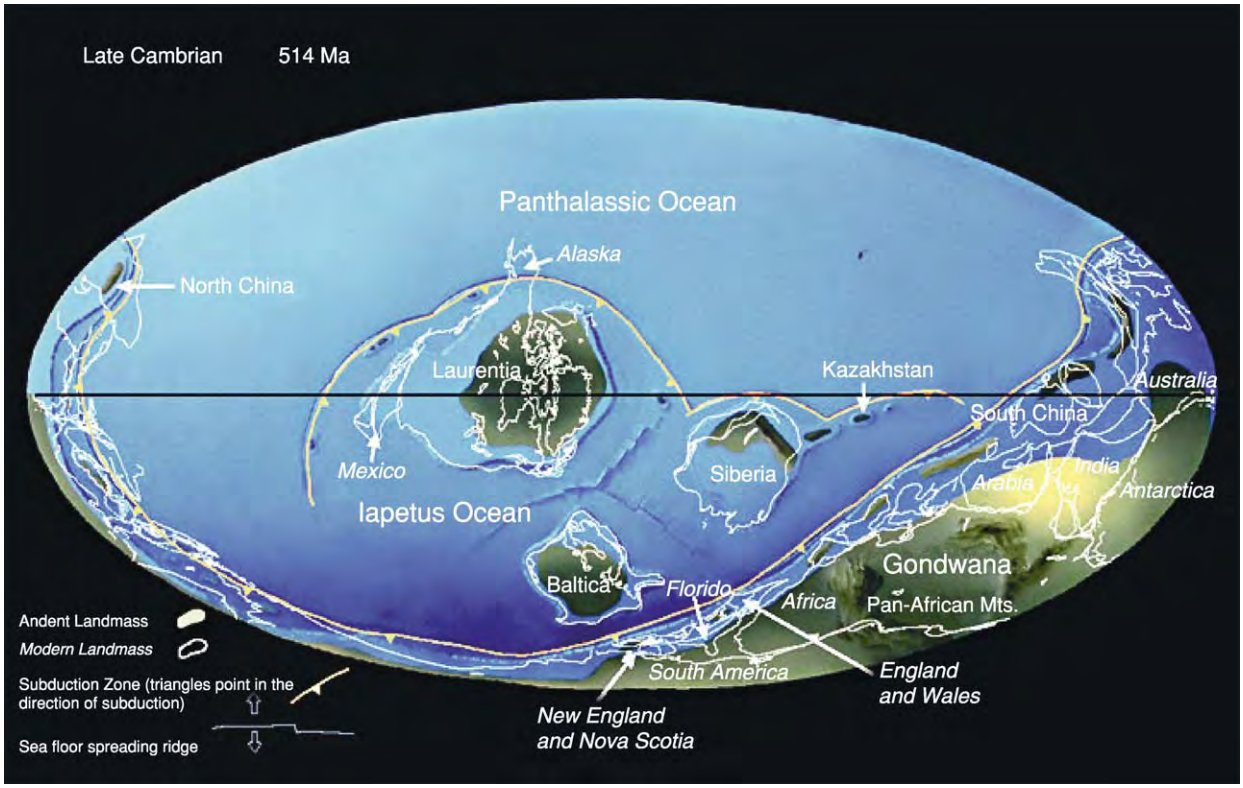


Figure 4 A reconstruction of Cambrian palaeogeography at approximately 514 Ma. Reproduced with permission from Chris Scotese and the Paleomap project.

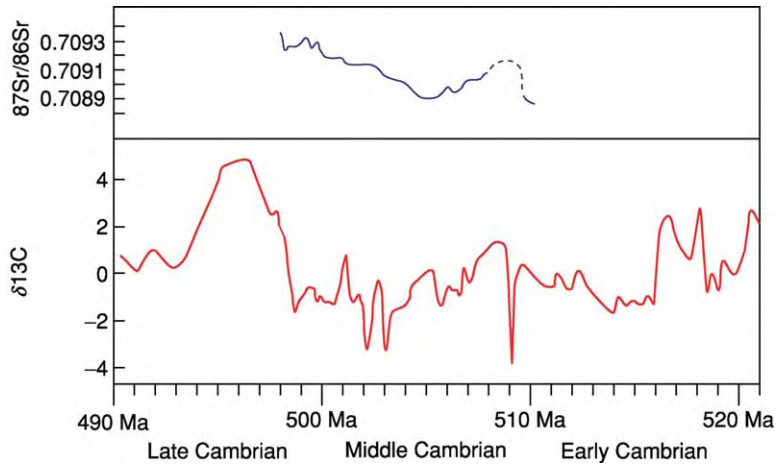


Figure 5 Secular changes in carbon and strontium isotope values during the Cambrian. Modified with permission from Montañez IP, Oselger DA, Banner JL, Mack LE, and Musgrove ML (2000) Evolution of the Sr and C isotope composition of Cambrian oceans. *GSA Today* 10: 1 7.

The Neoproterozoic–Cambrian Evolutionary Radiation

Cambrian Life

The morphological and taxic diversity of trace fossils expanded markedly in association with the basal Cambrian boundary. Surficial trails of simple morphology present in Late Neoproterozoic Ediacaran assemblages were joined by varied morphologies, suggesting motile, triploblastic organisms that mined the sediment infaunally and possessed complex anterior–posterior body patterning. The first traces that can be attributed to the scleritized appendages of arthropods also occurred at about 530 Ma, significantly before the appearance of arthropod body fossils, but after the base of the Cambrian system.

Biomineralized skeletons were present in the latest Neoproterozoic, but were uncommon there relative to their abundance in the Early Cambrian. The pre-trilobitic Cambrian contains a myriad of relatively small biomineralized fossils, collectively known as 'small shelly fossils'. These were commonly composed of calcium carbonate or calcium phosphate and displayed shapes that ranged from simple conical tubes to highly sculptured plates. The diversity of these forms increased through the lowermost Cambrian. Although the phylogenetic relationships of many small shelly fossils remain obscure, in some instances considerable progress has been made in assessing their affinities. Whilst some small shelly fossils represented protective skeletons surrounding the bodies of a variety of diploblastic organisms, others are known to be components of a sclerite meshwork containing hundreds or thousands of separate units that covered portions of the external surface of triploblastic organisms (Figure 6). The synchronous biomineralization of numerous separate metazoan and protist lineages, together with the appearance of Foraminifera that agglutinated inorganic minerals to their external surfaces, suggests that the possession of a reinforced external surface conferred a selective advantage in the Early Cambrian. This pattern, together with evidence from body fossil size and mouthpart anatomy of some Early Cambrian arthropods, suggests that Cambrian marine ecologies were structured with significant feedback between higher and lower food chain levels.

The appearance of trilobites marked the start of the traditional Cambrian fauna, as it has come to be known following Jack Sepkoski's static modelling of Phanerozoic marine diversity. This assemblage of fossils, based on the record of biomineralized fauna, was dominated by certain basal trilobite, linguliform brachiopod, echinoderm, and poriferan clades, and persisted into the lower Ordovician. Although the designations 'Cambrian fauna' and 'Palaeozoic fauna' aptly describe the marked contrast in the composition of biomineralized faunas, dynamic modelling of diversity does not suggest that these two faunas behaved as distinct ecological entities. Moreover, it is clear that basal representatives of many of the clades that rose to dominance in the Ordovician were also present in the Cambrian.

Cambrian benthic macrofossils were also characterized by a periodic series of major evolutionary radiations and extinctions. Several of these extinctions (e.g., those associated with the Early to Middle Cambrian transition) had a cosmopolitan effect,

Figure 6 The Early Cambrian triploblastic metazoan *Halkieria evangelista*. Note the external surface consisting of multiple skel etal elements. Each element individually constitutes a 'small shelly fossil', many of which were combined on a single individ ual. Photograph: John S. Peel.

during which important groups, such as the archaeocyathid sponges and olenelloid trilobites, were exterminated. Marked episodes of diversity expansion were typically followed by stratigraphically sharp extinctions. These striking cycles of diversity fluctuation occurred repeatedly during the Cambrian, and have been documented most clearly in the Late Cambrian 'biomeres' of Laurentia. Although the mechanisms responsible for the biomere pattern are still debated, it appears that large-scale eustatic changes of sea-level, probably coupled with fluctuations in seawater temperature and oxygen levels, were driving factors.

Cambrian Faunal Provinces

The distributions of Cambrian taxa can be used as tools for both biostratigraphical correlation and palaeogeographical reconstruction. Geologically short-lived, abundant, and cosmopolitan species (e.g., agnostoid trilobites) have great utility for global correlations. Other forms had more restricted distributions, with their geographical ranges limited by barriers (e.g., temperature tolerances, the inability to cross deep ocean basins; Figure 7). Distributions of fossils can be used as independent criteria to constrain palaeogeographical reconstructions based on other data. Laurentian shelf faunas were apparently the most distinctive, which is consistent with the notion that Laurentia was geographically isolated during Cambrian times. A wide-ranging shelf fauna also occurred about the peri-Gondwanan margin, although restriction of some elements to specific regions suggests some palaeolatitudinal limits on faunal distribution. In general, species adapted to cooler waters had more widespread occurrences than those restricted to equatorial shelf environments. Additionally, the occurrence of Cambrian faunas diagnostic of one province surrounded by rocks of another province has proved key to the recognition of Cambrian microcontinents.

The Neoproterozoic–Cambrian Biotic Transition

The traditional criterion for recognizing the base of the Cambrian is the first appearance of macrofossils in the stratigraphical record – the change from 'hidden' to 'obvious' life. Although it is certain that Cambrian global tectonics favoured the accumulation and preservation of fossils, and may even have fuelled evolutionary radiation, the extent to which Cambrian tectonic settings triggered the metazoan radiation is less clear. Clarification of this depends on a knowledge of the nature of biotic diversity and structure prior to the base of the Cambrian. If the clades that first appeared in the Cambrian had phylogenetic origins significantly prior to the base of the Cambrian and had also assumed their 'modern' ecological roles within Neoproterozoic ecosystems, the appearance of abundant fossils in the Cambrian may be related to conditions that favoured both biomineralization and the preservation of shelfal sediments. Alternatively, the Cambrian might chronicle the fundamental steps in metazoan phylogenetic and ecological diversification, in which case clues for the trigger might be sought within Cambrian deposits.

It is now apparent that the base of the Cambrian did not coincide with the phylogenetic appearance of most of the clades, both metazoan and protist, that became common in Cambrian rocks. Although metazoan divergence patterns and times based on molecular data are fuzzy, they are consistent in several major aspects, placing the origins of diploblastic organisms before the triploblasts at least tens of millions of years before the base of the Cambrian. Furthermore,

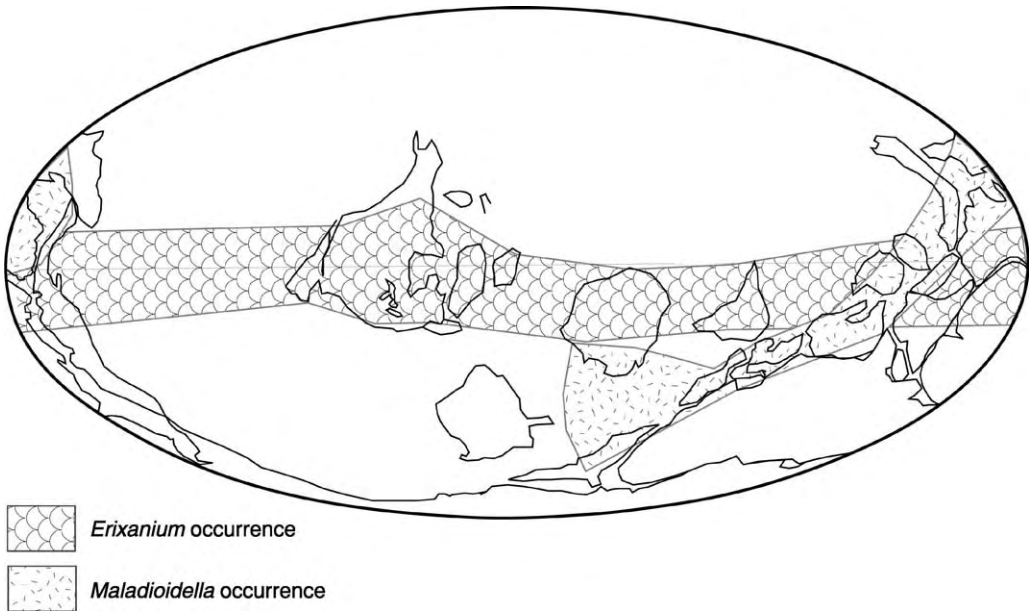


Figure 7 Contrasting biogeographical distributions of two widespread Late Cambrian trilobite genera. The *Erixanium* distribution was apparently constrained by factors related to latitude. In contrast, the *Maladioidella* distribution was apparently limited by open ocean basins. Modified with permission from Zhuravlev AY and Riding R (2000) *The Ecology of the Cambrian Radiation*. New York: Columbia University Press. © Columbia University Press, 2000.

direct fossil evidence of Neoproterozoic metazoans is provided by phosphatized embryos aged about 575 Ma from southern China, in addition to other various late Neoproterozoic fossils such as definitive sponges and possible cnidarians. Evidence for Neoproterozoic triploblasts is still being debated. The phylogenetic affinities of the famous Ediacaran fossil assemblages (*see* **Precambrian:** Vendian and Ediacaran), consisting of large bodies commonly constructed metamerically, also remain hotly contested. Even for some of the best known forms, such as *Dickinsonia*, recent interpretations range from it being a member of an extinct kingdom, a multicellular prokaryote, a lichen, a cnidarian, and a basal triploblastic metazoan capable of movement. Unequivocal Precambrian triploblasts pursuing ecologies closely comparable to their roles in the Phanerozoic have yet to be described, although the Ediacaran *Kimberella* – which may be a mollusc with an associated grazing trail – provides the most plausible candidate.

The presence of likely Late Precambrian metazoans that differed in gross morphology and ecology from their Cambrian counterparts suggests that the transition into 'obvious life' was not merely an artefact related to the advent of biomineralized skeletons with high preservation potential (although biomineralization clearly did appear synchronously and independently during Early Cambrian time in various metazoan and protist clades). Such a suspicion is confirmed by two other data sets that are arguably free of strong preservational biases across the Neoproterozoic–Cambrian boundary. The first of these is the record of trace fossils. Ediacaran fossils are commonly found preserved in sandstones with evidence of mat-like bedding surfaces bound by algae. Finer grained substrates, such as mudstones, were apparently quite firm throughout the Neoproterozoic to Cambrian transition interval. Such substrates would have provided a suitable medium for the preservation of surficial movement traces and shallow, infaunal burrowing. Examination of the trace fossil record shows that the diversity, complexity, depth, and extent of bioturbation increased progressively across the Neoproterozoic–Cambrian boundary, with trends begun in the Neoproterozoic continued into the Early Cambrian. These observations suggest that this interval witnessed important evolutionary innovations in metazoan size, morphology, and behaviour coincident with the establishment of benthic ecologies of essentially modern aspect. The second critical data set is the burst in morphological diversity of pelagic microflora at the onset of the Cambrian following marked conservatism during the late Neoproterozoic. New morphologies found in the Cambrian are consistent with the presence of active pelagic herbivores (mesozooplankton) of which there is direct evidence from Lower

Cambrian rocks. If mesozooplankton were essential to the establishment and maintenance of the Cambrian benthic communities, as they are in modern marine communities, it could be that innovations in the zooplankton drove diversification of the Cambrian macrobenthos.

It seems that the phylogenetic origins of the clades that first appeared in the Cambrian probably predated the base of the Cambrian by at least 50 Ma (and possibly far longer), as did most of the genetic regulatory architecture required to organize large, complex, body plans. This is indicated by the remarkable extent to which body patterning mechanisms are shared amongst disparate metazoan groups. For this reason, it seems unlikely that the Cambrian radiation was directly coupled to basic biological innovations that made metazoan diversity mechanistically possible. Nonetheless, the Cambrian radiation was more than a simple expansion of biotic diversity and ecological structure present in the terminal Neoproterozoic triggered by favourable tectonics. Cambrian ecologies, both pelagic and benthic, were different from those of the Neoproterozoic in ways that suggest enhanced trophic interactions and ecospace utilization. Biotic innovations coincident with the Neoproterozoic–Cambrian transition, both morphological and ecological, appear to have been critical to the adaptive radiation that clearly took place during Cambrian time.

Although there were marked differences between late Neoproterozoic and Cambrian biotas, this does not imply that the transition between the two was sharp, even on a geological time-scale. Both the sequence and apparent spacing of events are consistent with evolutionary processes and mechanisms operative today, although the rates of anatomical and behavioural innovations may have been elevated in some lineages relative to later times. This pattern, and the persistent compatibility of new data with the expectations of the evolutionary model, belies any basis for the idea that the Cambrian radiation requires alternative, non-evolutionary explanations. Furthermore, it can be argued that the types of evolutionary transformation that took place during the Precambrian–Cambrian boundary interval persisted into the Cambrian.

Many Early Cambrian body fossils arguably belong to stem group lineages of clades whose derived crown group representatives only appeared later in the Cambrian or thereafter. According to this view, many varieties of early arthropods present in Cambrian Lagerstätten (*see* **Lagerstätten**) (e.g., Chengjiang in Yunnan, China; the Burgess Shale in British Columbia, Canada) represent steps *en route* to derived euarthropods, and scenarios for the evolution of basic features of arthropod morphology, such as rigid, metameric sclerites, can be postulated on the basis of character state transformations seen amongst Cambrian fossils. Although the specifics of such scenarios are subject to debate, the fact that they can be reasonably postulated affirms that major aspects of arthropod evolution were probably accomplished subsequent to the appearance of the earliest arthropod body fossils. Furthermore, the opportunity to explore the connection between the ontogenetic and phylogenetic development of body patterning in early Palaeozoic arthropods (e.g., Figure 8) suggests that aspects of the basic control of body plan (e.g., the numbers and varieties of segments) evolved during the Early Palaeozoic. Although such opportunities may offer a fascinating glimpse of the evolution of developmental processes, they apparently concern relatively minor adjustment of developmental systems established prior to the base of the Cambrian.

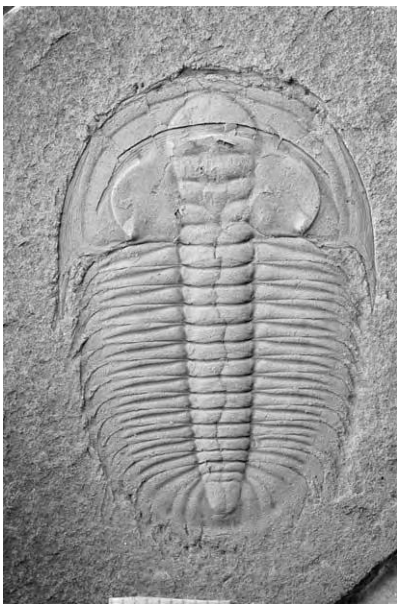


Figure 8 The Middle Cambrian trilobite, *Xystridura templetonensis*, from Australia, a typical representative of the 'Cambrian fauna'. Photograph: Nigel C. Hughes.

See Also

Africa: Pan-African Orogeny. **Analytical Methods:** Geochronological Techniques. **Evolution**. **Famous Geologists:** Sedgwick. **Fossil Invertebrates:** Trilobites. **Lagerstätten**. **Palaeozoic:** Ordovician. **Precambrian:** Vendian and Ediacaran. **Trace Fossils**.

Further Reading

Berner RA (2001) Modeling atmospheric O_2 over Phaner ozoic time. *Geochimica et Cosmochimica Acta* 65: 685 694.

Bowring SA and Erwin DH (1998) A new look at evolution ary rates in deep time: uniting paleontology and high precision geochronology. *GSA Today* 8: 1 7.

Budd GE and Jensen S (2000) A critical reappraisal of the fossil record of the bilaterian phyla. *Biological Reviews* 75: 253 295.

Conway Morris S (1998) *Crucible of Creation.* Oxford: Oxford University Press.

Cooper RA, Nowlan GS, and Williams SH (2001) Global Stratotype Section and Point for base of the Ordovician System. *Episodes* 24: 19 28.

Davidek K, Landing E, Bowring SA, *et al.* (1998) New uppermost Cambrian U Pb date for Avalonian Wales and age of the Cambrian Ordovician boundary. *Geo logical Magazine* 135: 305 309.

Dickson JAD (2002) Fossil echinoderms as monitor of the Mg/Ca ratio of Phanerozoic oceans. *Science* 298: 1222 1223.

Droser ML, Jensen S, and Gehling JG (2002) Trace fossils and substrates of the terminal Proterozoic Cambrian transition: implications for the record of early bilaterians and sediment mixing. *Proceedings of the National Academy of Sciences, USA* 99: 12 572 12 576.

Gehling JG, Jensen S, Droser ML, Myrow PM, and Narbonne GM (2001) Burrowing below the basal Cambrian GSSP, Fortune Head, Newfoundland. *Geo logical Magazine* 138: 213 218.

Grotzinger JP, Bowring SA, Saylor BZ, and Kaufman AJ (1995) Biostratigraphic and geochronologic constraints of early animal evolution. *Science* 270: 598 604.

Hughes NC (2003) Trilobite body patterning and the evo lution of arthropod tagmosis. *BioEssays* 25: 386 395.

Johnson WJ and Goldstein RH (1993) Cambrian sea water preserved as inclusions in marine low magnesium calcite cement. *Nature* 362: 335 337.

Landing E, Bowring SA, Davidek KL, *et al.* (2000) Cam brian Ordovician boundary age and duration of the lowest Ordovician Tremadoc Series based on U Pb zircon dates from Avalonian Wales. *Geological Magazine* 137: 485 494.

Landing E, Bowring SA, Fortey RA, and Davidek KL (1997) U Pb zircon date from Avalonian Cape Breton Island and geochronologic calibration of the Early Ordovician. *Canadian Journal of Earth Sciences* 35: 724 730.

Montañez IP, Oselger DA, Banner JL, Mack LE, and Musgrove ML (2000) Evolution of the Sr and C isotope composition of Cambrian oceans. *GSA Today* 10: 1 7.

Narbonne GM, Myrow PM, Landing E, and Anderson MM (1987) A candidate stratotype for the Precambrian Cambrian boundary, Fortune Head, Bruin Peninsula, southeast Newfoundland. *Canadian Journal of Earth Sciences* 24: 1277 1293.

Sepkoski JJ Jr (1979) A kinetic model of Phanerozoic taxo nomic diversity. II. Early Phanerozoic families and multiple equilibria. *Paleobiology* 5: 222 251.

Zhuravlev AY and Riding R (2000) *The Ecology of the Cambrian Radiation.* New York: Columbia University Press.

Ordovician

R A Fortey, The Natural History Museum, London, UK

Introduction

The Ordovician System, which is radioisotopically dated to have lasted from 495 Ma to 443 Ma, follows the Cambrian in the geological time-scale and is overlain by the Silurian. The Ordovician comprises the middle part of the Lower Palaeozoic, and has been traditionally divided into six series, of which two (Llanvirn and Llandeilo) have been combined to give a five-fold subdivision. These series are currently in the process of revision, largely to bring them into line with the threefold division (Lower, Middle, and Upper) used in other geological systems. It is still, however, convenient to describe Ordovician rocks in terms of five classical subdivisions, which are in ascending order: Tremadocian (489–483 Ma), Arenigian (483–463 Ma), Llanvirnian (463–*ca*. 455 Ma), Caradocian (*ca*. 455–*ca*. 448 Ma), and Ashgillian (*ca*. 448–443 Ma). Of these, the Tremadocian was previously included in the topmost part of the Cambrian by some authorities, but has now been assigned to the Ordovician by international agreement. The top of the Ordovician is marked by a major ice age and an important mass extinction of marine organisms, which forms a natural boundary with the base of the Silurian. Ordovician rocks are found on every continent except the Indian Peninsula. During the Ordovician the continental masses were widely dispersed, other than the Gondwana supercontinent. Ordovician geology, therefore, varies richly from one area to another. The Ordovician saw the inception of several marine ecosystems that are still with us today.

Several important animal groups (e.g. corals and nautiloid cephalopods) also underwent evolutionary radiations during the Ordovician.

History

The Ordovician System was a comparative latecomer to the geological column. It was born out of a historical compromise between competing claims of what strata should be included in the underlying Cambrian and overlying Silurian systems. The seminal studies of this part of the stratigraphical succession were carried out in the British Isles. Cambrian rocks had been studied in Wales by the Woodwardian Professor at the University of Cambridge, the Reverend Adam Sedgwick (*see* **Famous Geologists:** Sedgwick) – a pioneering field geologist whose classes had been attended by the young Charles Darwin. Rocks comprising the Silurian System were described in considerable detail by Sir Roderick Murchison (*see* **Famous Geologists:** Murchison) in one of the classics of geology, *The Silurian System* (first published in 1839). Murchison was a gentleman of private means and a dominant personality, who had studied rock successions in Shropshire and across southern Wales, where inevitably he overlapped stratigraphically with Sedgwick's ground. Murchison's attempts to 'appropriate' some of Sedgwick's Cambrian rocks into the Silurian met with stern opposition from the Cambridge cleric, and a vigorous controversy ensued, which was still not resolved on the death of both protagonists. A compromise solution to the problem was devised by Charles Lapworth, a Professor in the Department of Geology at the new University of Birmingham. This compromise involved hiving off the disputed middle ground into a separate System, the Ordovician. The name was derived – like that of the Silurian – from the name of an ancient British tribe. Although Lapworth's proposal was made in 1879, it was a long time before it was universally accepted. Indeed, one can find Murchison's 'Silurian' still in use in Russia more than 50 years later. However, as with many of Lapworth's innovations, it came to be seen as a wise solution, and the Ordovician has since been accepted as one of the more natural divisions of geological time.

An outstanding confusion remained, however, about whether to include the rocks around the north Welsh town of Tremadoc in the Cambrian or in the Ordovician. Lapworth had originally stated that the base of the Ordovician should be drawn at the base of the Arenigian rocks, thereby assigning the underlying Tremadocian rocks to the Cambrian. To many observers at the close of the nineteenth century, the Tremadocian seemed to show 'intermediate' features between the two stratigraphical contenders. In Scandinavia, the Tremadocian was incorporated in the Ordovician. The situation was further complicated because in much of North America – where fossil faunas are very different from those in Europe – Tremadocian rocks were not recognized as such, and a different set of names (e.g. 'Ozarkian' and 'Canadian') were applied to rocks of Late Cambrian and Early Ordovician age. Only when correlation between these separate continents improved did it become practicable to apply a standard series of names to the subdivisions of Ordovician time. As late as 1972, the Tremadoc 'series' continued to be allocated to the Cambrian by British scientists, although by then a majority of other nations had settled on a horizon for the base of the Ordovician that approximated the base of the Tremadocian. Such problems lend themselves to arbitration by the International Union of Geological Sciences, and the Tremadocian was eventually deemed to be the basal Ordovician subdivision, by general consensus, in 1988. The horizon chosen to mark the base of the Ordovician also marks the appearance of planktonic graptolites of *Rhabdinopora* (formerly *Dictyonema*) *flabelliformis* type, and thus respects the spirit of Lapworth, who recognized the stratigraphical utility of these ubiquitous fossils well over a century earlier.

Type Areas and Sections

Since Ordovician rocks are so varied, it has not proved possible to define a single 'type area' for subdivision of the System, although the Anglo-Welsh area remains historically important. The base of the Ordovician System has recently been defined by international agreement within the Cow Head Group at Green Point, on the western coast of Newfoundland, Canada. Here, graptolite and conodont fossils occur together in a series of limestones and shales, permitting ready international correlation. In some areas (e.g. the western USA) conodonts are found in the absence of graptolites, whereas the situation is reversed in Spain and northern Africa. The index fossil is the first appearance of the conodont *Iapetognathus fluctivagus*. The base of the Silurian is, by definition, the top of the Ordovician System and has been similarly defined by international agreement in one of Lapworth's classical sections at Dob's Linn, near Moffat, in Scotland. This horizon lies at the base of the *Akidograptus acuminatus* biozone, which is characterized by the eponymous graptolite.

Attempts to define Lower, Middle, and Upper Ordovician Series are in progress. The base of the Middle Ordovician has been mooted, on the basis of conodonts, to lie at a horizon within the Arenig 'Series' of traditional usage. Correlation remains a problem,

however, and to date a type section has yet to be chosen. A standard time division close to the Middle Ordovician, the Darriwilian, has been recognized on the basis of graptolites. The base of the Upper Ordovician has been agreed to be the base of the widespread *Nemagraptus gracilis* biozone. This horizon coincides with the base of the Caradoc 'Series' as now defined. A type section in the Fågelsang section, Sweden, has been selected for its definition. Because this threefold series subdivision has not yet achieved wide currency, it is expedient to continue to use the fivefold subdivision in this chapter. Strictly speaking, these should no longer be referred to formally as 'Series', as they are effectively regional stages. The suffix '-ian' is therefore employed below. Because of the high degree of regional differences in Ordovician strata, several widely separated areas have become important in defining Ordovician temporal subdivisions on a regional to global scale. Biozonal standards have been based on graptolites, conodonts, and trilobites, but these are found together in only a minority of rock sections and they, too, show provinciality. Much effort has accordingly been expended in tying together graptolitic and conodont-based zonal schemes.

Tremadocian

The small town of Tremadoc lies in North Wales at the northern end of Cardigan Bay. A series of dark shales and mudstones of the Dol cyn Afon Formation – often cleaved and weakly metamorphosed – overlie rocks of the Upper Cambrian. There is often a stratigraphical break between the two, but in at least one section a complete and conformable transition between the two systems has been demonstrated. The Tremadocian is recognized by the incoming of net-like rhabdosomes of the widespread graptolite *Rhabdinopora flabelliformis* subspecies, the oldest planktonic graptoloid species. These rocks were deposited in comparatively open shelf environmental conditions, and the graptolites are accompanied by a variety of trilobites, of which members of the Family Olenidae are a prominent component, some species of which are also very widespread. The genus *Jujuyaspis*, for example, is known from North and South America, Scandinavia, and China in earliest Ordovician strata. In North America, the stage name Ibexian (approximately the same as 'Canadian' used by earlier authors) is employed for Early Ordovician strata, based on well-exposed sections in Utah in the Great Basin. The Ibexian embraces all of the Tremadocian and part of the Arenigian. Both the Tremadocian and Ibexian have been further subdivided into chronostratigraphical subdivisions of regional application. The graptolite and conodont succession is particularly well-known in the Baltoscandian area and indicates the presence there of an upper part of the Tremadocian (Hunnebergian) that is not well-developed in Britain.

Arenigian

Arenigian rocks overlie the Tremadocian rocks in North Wales, and were named after a mountain, Arennig Fawr, where they are well exposed. However, in North Wales there is an unconformity at the base of the Arenigian, which is marked by a transgressive sandstone, and so it is inadequate as a type area (Figure 1). The Arenigian succession is better developed in South Wales and Shropshire, where thick sequences of mudstones, shales, sandstones, and turbidites have yielded diverse trilobite faunas of shallow-to-deep shelf aspect with, more locally, graptolites and brachiopods. Because of the disparate biogeography of the Early Ordovician, most of the species are of relatively local occurrence, and correlation of these strata internationally can be difficult. For some years a *de facto* base of the Arenigian has been recognized at the base of the widespread graptolite *Tetragraptus* (*Etagraptus*) *approximatus* biozone, which can be identified in North and South America,

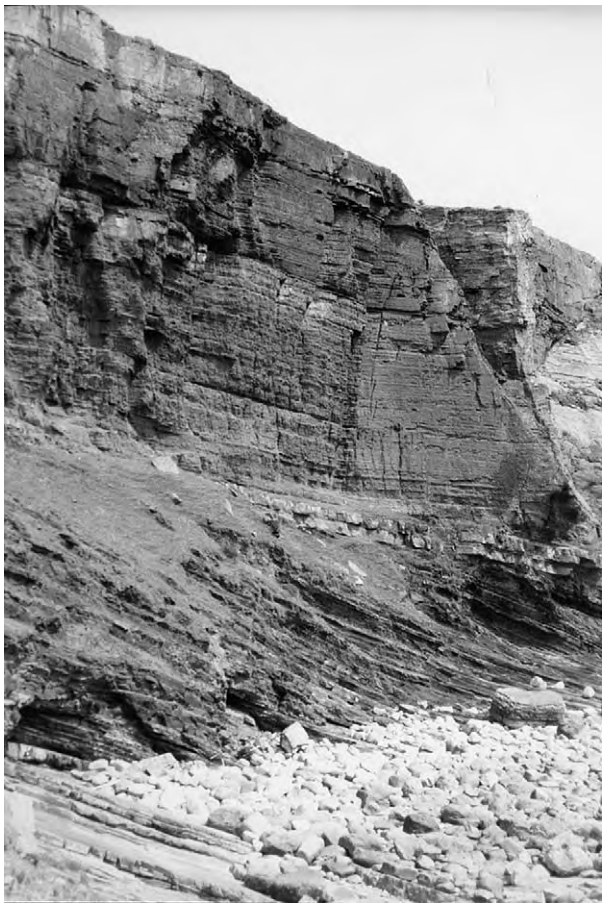


Figure 1 Ordovician (Arenigian) strata lying unconformably over Cambrian rocks on the Llyn Peninsula, North Wales.

Scandinavia, Australia, and New Zealand. The most varied succession of graptolite faunas is found in the shales of the region of Bendigo, Victoria, Australia; this has achieved the status of a global standard, and a fine-scale division into biozones (and stages) has been established, which has been widely applied elsewhere. In south-western China sequences span facies from inshore to deep-water, as do sequences in Scandinavia and the Russian Platform. In North America, strata equivalent to the Arenigian embrace the upper part of the Ibexian and the lower part of the Whiterockian, and have been documented in largely calcareous successions in the Great Basin. In many limestone sequences, conodonts have become the biostratigraphical standard. This disparity of stratigraphical criteria together with regional differences have meant that a plethora of stratigraphical subdivisions of Arenigian strata has grown up over the years, and different regional names are used in, among others, China, Scandinavia, Britain, Australia/New Zealand, and North America. Correlation between these schemes is difficult. However, certain horizons have proved to have international utility. Near the top of the Arenigian, for example, the appearance of biserial graptolites marking the *austrodentatus* biozone has been recognized on most palaeocontinents, and this will form one basis of future standardization.

Llanvirnian

The Llanvirn 'Series' was recognized in the nineteenth century in black slates exposed on the coast in south-western Wales, where it takes its name from an insignificant farmhouse. It is typified by a multitude of 'tuning fork' graptolites (e.g. *Didymograptus artus*, *D. murchisoni*, and related forms), appearing in the Aber Mawr Formation, which have been recognized widely in continental Europe and Scandinavia. The type area is complicated by volcanics and difficult structure, as is the area around Fishguard, and the succession of strata in Britain is better inspected in Shropshire and central South Wales. In Scandinavia, the Upper *Didymograptus* Shales and their equivalents in limestone strata have enabled comparisons to be made with a rich conodont fauna, which provides an international basis for correlation. The Llandeilan (the lower part of the original Llandeilo 'Series') has recently been incorporated into the upper part of the Llanvirnian. The flaggy limestones around the town of Llandilo in central South Wales were already well known by Murchison's time and are among the most fossiliferous rocks of the British Ordovician. Similar problems to those of the Arenigian apply to the international correlation of subdivisions of the Llanvirnian. The Whiterockian of the North American standard includes the Llanvirnian but extends downwards into the Arenigian. Across Europe and Asia the base of the Llanvirnian is marked by a transgressive event, indicated by a deepening in the biofacies and the appearance of graptolites.

Caradocian

The type area of the Caradocian is in Shropshire, England, and it takes its name from a prominent hill, Caer Caradoc, in the vicinity of Church Stretton. In that area the Caradoc is transgressive and its base is defined by an unconformity at the base of the Hoar Edge Grit. As the sequence deepens upwards into a varied succession of mudstones, sandstones, and siltstones (from which several formations are mapped), a great variety of trilobites and brachiopods appear, which were originally studied in detail by B. B. Bancroft. He divided the Caradocian into fine subdivisions ('stages') based on faunal turnover, which, while useful at a local level, are of limited service internationally. In the Shelve Inlier, in South Wales, and near Builth the facies represent deeper water and the sections are without unconformities. However, the effect of the Caradocian transgression is to bring in many graptolite species of stratigraphical utility, including *Dicellograptus* and *Nemagraptus*, replacing the restricted fauna of the Llandeilan *teretiusculus* biozone beneath. This event is of worldwide significance, and the graptolite species concerned can be widely employed for correlation purposes.

The base of the Caradocian (and of the Upper Ordovician) is defined as the base of the *Nemagraptus gracilis* biozone, which can be traced into many areas including the Laurentian platform edge, Bohemia, continental Europe, and the standard graptolitic sequences of Australia and New Zealand. *Nemagraptus gracilis* itself is probably associated with relatively 'oceanic' conditions, and its appearance in basinal successions is probably not entirely synchronous, but the co-occurrence of some of its attendant species of *Dicellograptus* usually places correlation on a sound footing. This, together with the beginnings of a breakdown of the faunal provinciality typifying the earlier Ordovician, means that international correlation of Caradocian strata is less of a problem than is the case with Arenigian and Llanvirnian strata. Nonetheless, there are still distinctive fossil faunas and concomitant separate stratigraphical schemes in the platform limestone successions (Trentonian and Blackriverian) of North America and the Baltic areas. In China, a distinctive diachronous red often nautiloid-rich deep-water formation, the Pagoda Limestone, appears in the Caradocian and continues

into the Ashgillian. North African Gondwanan Caradocian successions include endemic genera adapted to cool-water conditions close to the pole. In deep-water sites, graptolitic successions are well-studied, particularly in the Glenkiln and Hartfell Shales of the Southern Uplands of Scotland, where Lapworth was a pioneer in establishing the succession of faunas. In Wales, the Caradocian was a time of intense volcanic activity in the area around Snowdon, with several kilometres of varied pyroclastic and eruptive rocks, which are classics in the study of 'fossil' volcanoes.

Ashgillian

The youngest, and shortest, subdivision of the Ordovician System (Figure 2) has a type section in the English Lake District, at Ash Gill. A correlative section at Foggy Gill shows continuous passage downwards into the Caradocian. The base of the Ashgillian at the base of the Pusgillian local 'stage' is recognized entirely on a shelly fauna of brachiopods and trilobites. Conodonts and, until recently, graptolites were unrecognized from this locality, and this has made it difficult to locate the base of the Ashgillian in the graptolitic (and conodont) schemes, although a horizon within the *Pleurograptus linearis* biozone has often been accepted. Fossil faunas are not strikingly different from those of the underlying beds and are diverse, but higher in the succession there is a decline in the Rawtheyan local 'stage', presaging a much more drastic reduction as the climate deteriorated at the end of the Ordovician. Shallow-water tropical strata in North America correlative with the Ashgillian, the Cincinnatian limestones, are prodigiously fossiliferous. The stratigraphy of Ashgillian strata in Bohemia, Scandinavia, and China is known in equal detail. There is, in general, a more cosmopolitan fauna in this part of the Ordovician, but cool-water high-latitude faunas still endured close to the pole in northern Africa at the core of the Gondwana supercontinent. Graptolitic deposition continued in deep-water sites distant from continental influences, as in southern Scotland and Taimyr. The youngest subdivision of the Ashgillian, the Hirnantian, marked the onset of the Ordovician glaciation, which is accompanied by drastic lithological and faunal changes in most stratigraphical sections caused by a worldwide regression. In many places there is an unconformity at this horizon, while in all sections spanning the interval there is a reduction in faunal diversity, often accompanied by shallow-water clastic sedimentation.

Figure 2 Late Ordovician (Ashgillian) strata underlying Conway Castle, North Wales.

Life in the Ordovician

The Ordovician is arguably the period of greatest marine diversification in the fossil record. Groups of animals that appeared in the Cambrian (e.g. bivalved molluscs and gastropods) became much more diverse and widespread. Other groups appeared and diversified for the first time. These included corals and bryozoans, which, after a rare occurrence in the Tremadocian, formed diverse bioherms and even reefs before the Caradocian. Cephalopod molluscs – the nautiloids – achieved great size and variety of form, and presumably were among the most effective predators that served to structure the ecosystem. The free-floating graptolites (*see* **Fossil Invertebrates:** Graptolites) became widespread in the world's oceans and are often abundant macrofossils in off-shelf sites where few Cambrian animals can be found. Trilobites (*see* **Fossil Invertebrates:** Trilobites) continued to diversify from the Cambrian and included free-swimming as well as benthic forms, which did not survive into the Silurian. Filter-feeding trilobites were highly characteristic of soft substrates; for example, trinucleid trilobites are abundant in, and confined to, Ordovician strata. Among predatory trilobites, the largest example known from the whole group (*Isotelus*) was discovered in Ordovician strata in Canada.

Articulated brachiopods became numerous enough to form shell beds and evolved into the orders that would dominate the Palaeozoic shallower seas. Echinoderms saw radiations in the starfish, sea lilies, and the earliest sea urchins. By the mid-Ordovician vertebrates with phosphatic external skeletons had become quite varied. Conodonts (now considered to be related to vertebrates) (*see* **Microfossils:** Conodonts) became abundant, and showed a variety of feeding apparatuses. It is likely that the first land plants, allied to liverworts, had appeared by Llanvirnian times, although their fossil record is dominated by spores alone. Moreover, because of the diverse palaeogeography of the time, various groups of animals evolved differently on different palaeocontinents, adding to the total biodiversity. Climatic zonation aided this differentiation. If there was a watershed between Cambrian-style faunas and those characterizing later strata, it occurred at the base of the Whiterockian, when taxa typical of younger strata increased at the expense of those with an archaic history extending back into the Cambrian. It has also been shown that there was an increase in the complexity of trace fossils and bioturbation through the period, with younger Ordovician animals burrowing deeper in the sediment and making 'galleries'. The kind of tiering characteristic of later infaunal assemblages had its origins in the Ordovician. It is possible that increased predation drove these innovations forwards. Allowing for fluctuations in the fossil record, there was an increase in global biodiversity through the Ordovician until the Caradocian (**Figure 3**). The enrichment process was interrupted by the extinction event at the end of the Ordovician, but communities were able to re-establish comparable complexity in the Silurian.

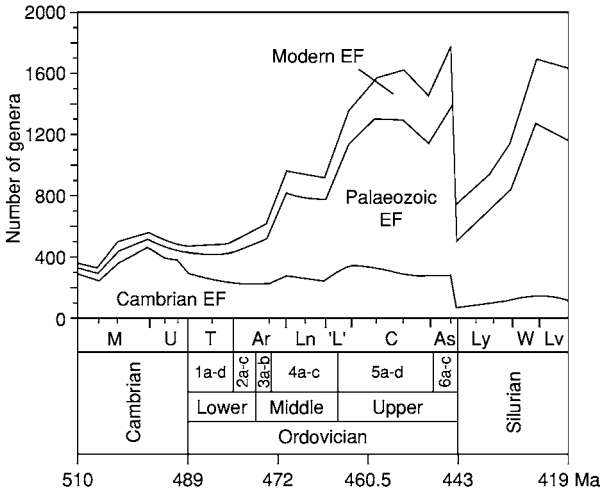


Figure 3 Changes in biodiversity through the Ordovician, showing changes in the diversity of taxa belonging to the different 'evolutionary faunas' (EF) identified by JJ Sepkoski. Maximum diversity is reached late in the Ordovician. The numbers 1a d, 2a c etc., refer to finer scale temporal subdivisions used in cali brating this diagram. Reproduced from Webby BDE, Paris F, Droser ML, and Percival IG (eds.) (2004) *The Great Ordovician Biodiversification Event*. New York: Columbia University Press.

The End-Ordovician Extinction and Glaciation

The mass extinction at the end of the Ordovician is one of the major events in the fossil record, being exceeded only by those at the end of the Palaeozoic and at the Cretaceous–Tertiary boundary. While the ultimate controls on the latter events are still a subject of dispute, the end-Ordovician extinction coincided closely with a short-lived but very extensive glaciation, and it is natural to link the biological event with the physical event. Exactly how the secondary effects associated with the glaciation caused the extinction is still the subject of research.

The ice age climaxed in the final subdivision of the Ordovician, the Hirnantian 'stage'. It was evidently short-lived, and current evidence indicates that the Hirnantian may have been as brief as 0.5 Ma. There is good reason to suppose that the climate was deteriorating before the glacial climax. Evidence for glacial deposits is widespread near the Ordovician south pole in North Africa, and they have also been discovered in South America and the Arabian Peninsula. All these areas were part of the Gondwanan continent in the Ordovician, and the scenario of an extensive ice-sheet expanding outwards from a Gondwanan core is supported by evidence on the ground. In Morocco, there are spectacular glacial pavements, which are associated with tillites and finer-grained diamictites. Evidence of ice movement away from the polar centre has been obtained from striae, roches moutonées, and the fining directions of glacial deposits (*see* **Sedimentary Processes:** Glaciers). The glacial event does not seem to have involved multiple advances and retreats, as in the Pleistocene, although a double ice-advance pulse has been claimed. Correlation within non-marine sediments is difficult at this early time, however.

The effects of the glaciation extended well beyond Gondwana. Oceanic water became tied up in the ice-sheet, resulting in a general marine regression. This event was accompanied by climatic cooling, such that the tropical carbonate belt became extremely restricted. This obviously affected the many organisms adapted to warm carbonate environments, which either died out or were pushed into small refugia. Clastic deposits are typical of the Hirnantian, but in many localities the interval was one of general scouring and erosion. A striking faunal change occurs

in the same interval. A low-diversity set of fossil genera known as the *Hirnantia* Fauna is found almost ubiquitously in strata deposited during the regression. *Hirnantia* itself is a brachiopod, and brachiopods are probably most characteristic, but a dalmanitoid trilobite, *Mucronaspis*, is almost invariably found alongside the other shelly fossils. Nearer the palaeoequator a somewhat more diverse fossil fauna can be collected, but there is little sign of the extraordinarily diverse life of the Early Ashgillian. The fact that most members of the *Hirnantia* fauna had a previous Gondwanan pedigree supports the idea that it represents an invasion of cool-water taxa into lower latitudes – a 'crisis' fauna. It even appears in the midst of graptolitic sequences in China, where the sea-level must have dropped sufficiently to allow the colonization of the seafloor by benthos.

This extinction event was much more profound than one just affecting the shelf benthos. The diversity of those animals inhabiting the open ocean was severely reduced. The fossil record is probably better for the planktonic graptolites than for anything else, and they were extinguished to the extent that only two or three species survived through the crisis. These are thought to have given rise to all the Silurian faunas. Oceanic trilobites such as agnostids and olenids, which had endured for tens of millions of years, disappeared completely, along with many of the typical Ordovician families of shelf trilobites (asaphids and trinucleids among them). It has been suggested that these trilobites had planktonic larvae, in which case an oceanic event affecting graptolites and larvae alike is implied. Such an event could have been extensive anoxia. Some, at least, of the shelf faunas were able to lurk in refugia through the crisis, and in some cases they reappeared as Lazarus taxa as late as the Wenlock (Silurian).

Ordovician Geography and Tectonics

The Ordovician was a period when major continents were separated by wide oceans. At the edges of these continents, microcontinents and islands arcs – many of which were subsequently accreted to one continent or another – produced a complex local geography, which was an additional factor in the rich biotic composition of the Ordovician.

The largest continent was Gondwana, which comprised Africa and the southern part of Europe, South America, the Indian and Arabian Peninsulas, Antarctica, and Australia (Figure 4). It extended from the south pole in Africa to north of the palaeoequator in Australia. Much of it, especially central Africa and India, was land throughout the period and largely lacks Ordovician strata. But shallow seas immersed northern Africa, Arabia, and large parts of Australia, and these areas preserve strata rich in fossils. South China was close, if not actually attached, to Gondwana through much of the Ordovician. The kinds of rocks that accumulated reflect the Ordovician palaeolatitudes: mostly clastic rocks near the pole, and predominantly carbonate rocks as the palaeoequator is approached. Spanning the palaeoequator, but well removed from Gondwana, the Laurentian palaeocontinent included what is now North America and Greenland, plus some fragments that lie today on the opposite side of the Atlantic Ocean. Throughout most of the Ordovician, thick deposits of limestones accumulated over platform Laurentia (Figure 5). The Laurentian continent appears to have been stable in

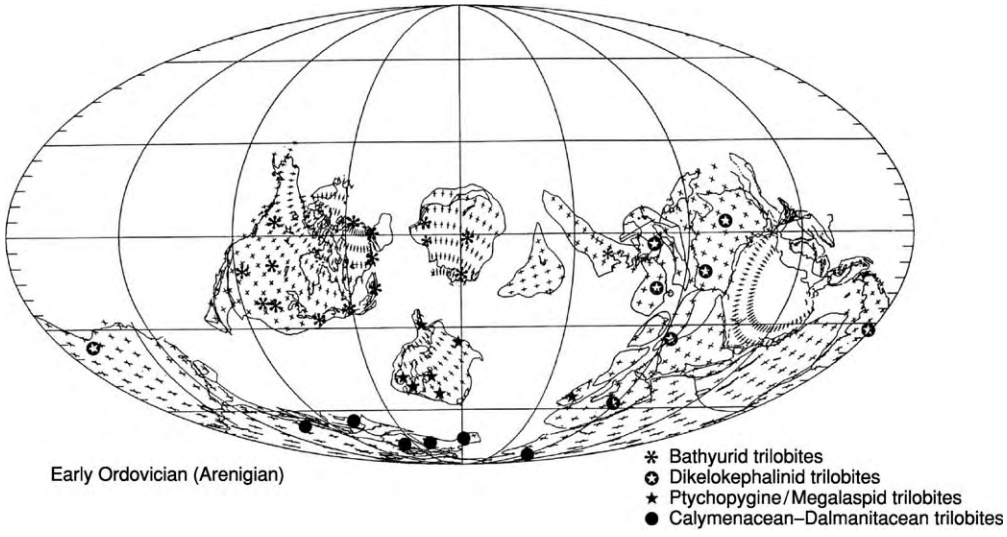


Figure 4 World geography in the Early Ordovician. The distributions of some of the characteristic trilobites are shown.

Figure 5 Limestone strata typical of Ordovician tropical palaeolatitudes: the white limestones of the Meiklejohn Bioherm (Middle Ordovician) near Beatty, Nevada, USA.

Figure 6 A Caradocian world reconstruction showing the migration of the Avalonian Terrane across Iapetus in the middle of the map.

its palaeoequatorial position throughout the Ordovician. The Siberian continent lay between Laurentia and Gondwana. Lying at an intermediate latitude through much of the Ordovician, the continent of Baltica comprised most of present-day Scandinavia and the Russian platform as far east as the Urals. Palaeomagnetic evidence suggests that it rotated between the Middle Cambrian and the Middle Ordovician to attain its current orientation. The palaeocontinents were surrounded by continental shelf and slope deposits or marginal basins related to offshore island arcs. In these, graptolite-bearing shales and mudstones accumulated, and benthic animals different from those inhabiting the shallow shelves were able to prosper. Further seawards again there were oceans, direct evidence for which is provided by obducted ophiolites preserved in mountain belts. These vanished oceans have been given several names: Iapetus separated Laurentia from Baltica and Gondwana; the Tornquist Sea separated Baltica from Gondwana. The oceans separated the major continents sufficiently for each to evolve its own endemic faunas in the earlier Ordovician.

Geography changed continually during the Ordovician, as oceans waned. For example, the Tornquist Sea closed, and Iapetus became much narrower. As this happened, former ocean islands became added to the edges of active continental margins along the subduction zones. The changing geography also facilitated the migration of faunas from one continent to another. Microcontinents migrated across the oceans: Avalonia – comprising the Anglo-Welsh area plus part of what is now eastern North America – moved from Gondwana towards Laurentia, and a new ocean (the Rheic Ocean) opened up behind it (Figure 6). Simultaneously, part of Laurentia drifted away to become accreted to what is now South America before the end of the period. This crustal mobility makes the Ordovician one of the more difficult geological periods for which to reconstruct the geography in detail.

Mountain belts grew along the lines of the subduction zones. The Ordovician saw the early phases of the growth of the Appalachian–Caledonian chain along the western margin of Iapetus. Another mountain belt (now partly buried beneath younger strata) marks the suture between Baltica and continental Europe. Another belt runs along eastern Australia. Volcanic activity along island arcs and in connection with subduction produced varied local rock sequences in different marginal terrains, in which volcanogenic and sedimentary rocks are interbedded. Palaeogeographical (latitudinal) coordinates for these contentious fragments can be determined from palaeomagnetic data, which, combined with evidence from lithology and faunas, contribute to a kinematic description of the shifting palaeogeography. Some mineralization of economic significance is associated with the Ordovician mountain belts. Extensive lead and silver deposits were mined in Wales and elsewhere along the Caledonian chain. Gold is known from many sites close to the main Iapetus oceanic suture in, for example, Ireland and Newfoundland.

Vulcanicity and Geochronology

The high level of tectonic activity during Ordovician times means that there is much evidence of explosive volcanic eruptions that spread ash deposits widely over the seafloor. These are recognized today as bentonites, which usually appear as very soft pale clay beds contrasting with the stratified deposits with which they are associated. In some successions, such as that of the eastern USA, bentonites occur more or

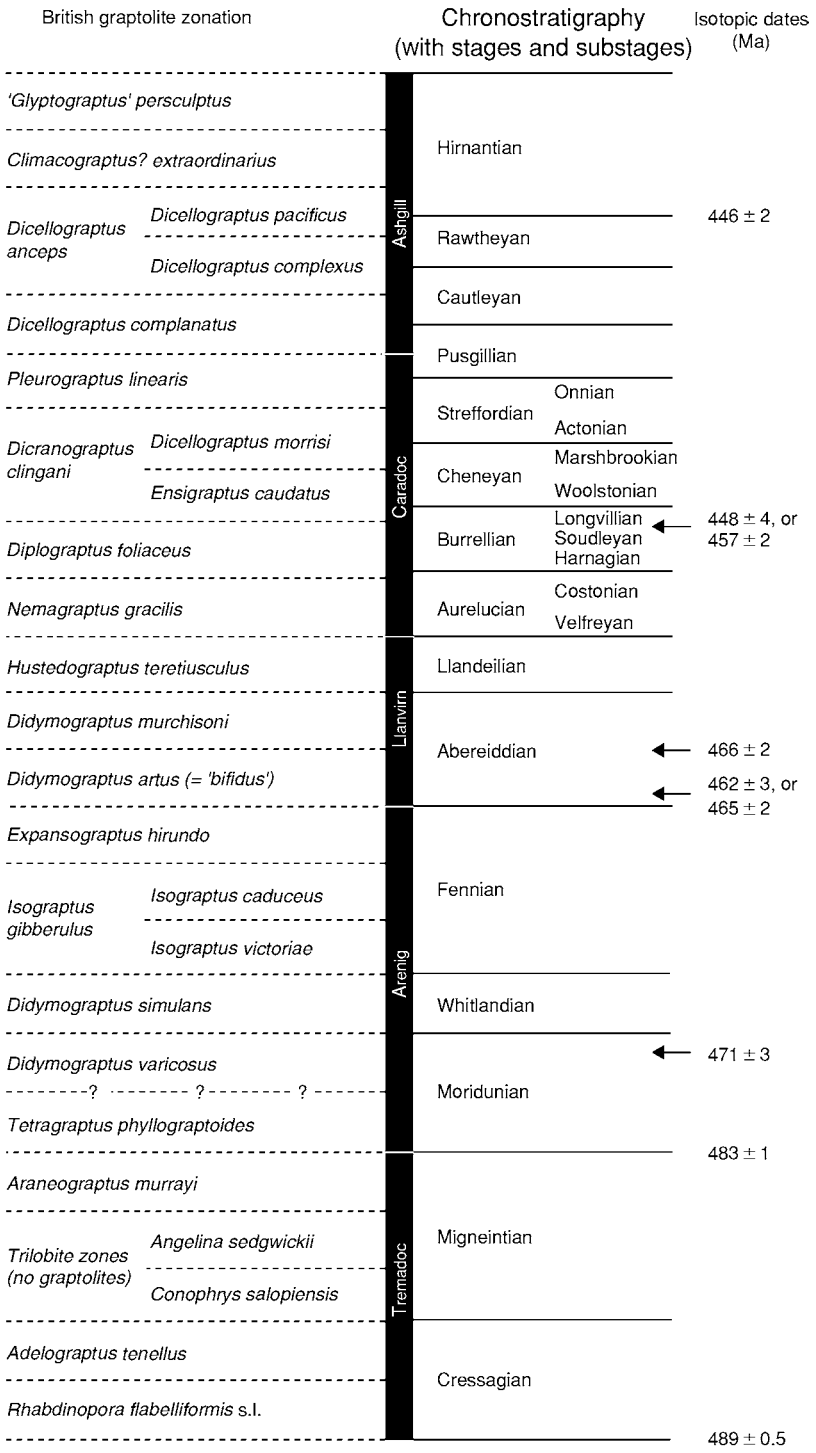


Figure 7 Chronostratigraphy of the Ordovician, showing some of the radioisotopic dates on the right. The traditional series names are employed without the ' ian' suffix in this diagram. Taken from Fortey RA, Harper PAT, Ingham JK, Owen AW, Parks MA, Rushton AWA and Woodcock NH (2000). A revised correlation of Ordovician Rocks in the British Isles. Geological Society. Special Report No. 24, pp. 38.

less regularly through hundreds of metres of strata. The Caradocian 'Big Bentonite' has been claimed to be the product of the explosion of an enormous 'super-volcano', an eruption greater than any that has been recorded in historical times. Bentonites useful in geochronology because they yield datable zircons and are often interbedded with graptolitic deposits, enabling a precise biostratigraphical tie-in (Figure 7). Other

sources of zircons include ash beds and tuffs. Subdivisions within the Ordovician are well-dated with zircon ages, often to an accuracy of ±1.0 Ma.

Biostratigraphy

Historically, biostratigraphical zonation using graptolites was the first scheme to be applied to Ordovician rocks for the purposes of international correlation of strata. Graptolites underwent a successive series of easily recognizable evolutionary changes through the Ordovician, which, even today, give a rapid method of estimating geological age. The 'stipe reduction series' – in which the graptolite fauna as a whole changes from species with many (eight or more) stipes to species with just two – still affords a 'ready reckoner' to distinguish Tremadocian to Llanvirnian strata in the field. Caradocian strata often include numerous Y-shaped colonies of *Dicranograptus*. However, finer zonal subdivision requires the accurate determination of species. Currently, about 25 biozones have been recognized. The solution of limestones to yield conodont fossils provided an additional and refined biostratigraphical system. This was especially useful because it could be applied to the bulk sampling of otherwise unpromising carbonate rocks. Since the 1950s conodont stratigraphy has complemented the graptolite-based schemes, with only slightly lesser refinement. However, such work requires dedicated laboratories. Recently, the stratigraphic utility of organic-walled microfossils – acritarchs and chitinozoa – has been appreciated, especially for dating clastic rocks from which conodonts are difficult to extract. Laboratory facilities capable of handling hydrofluoric acid are essential for this kind of work. These microfossils have proved exceptionally useful in dating rocks from high palaeolatitudes in the Ordovician. Trilobites and brachiopods have been used in local stratigraphy, but individual species are not usually of wide intercontinental geographical distribution, the notable exception being the *Hirnantia* Fauna. For international correlation, sections that include more than one index group are considered ideal: a combination of graptolites and conodonts is especially useful.

See Also

Biozones. **Famous Geologists:** Murchison; Sedgwick. **Fossil Invertebrates:** Trilobites; Graptolites. **Microfossils:** Conodonts. **Palaeomagnetism**. **Palaeozoic:** Cambrian; Silurian. **Sedimentary Processes:** Glaciers.

Further Reading

Bassett MG (ed.) (1976) *The Ordovician System*. Cardiff: University of Wales Press and National Museum of Wales.

Brenchley PJ (1984) Late Ordovician extinctions and their relationship to the Gondwana glaciation. In: Brenchley PJ (ed.) *Fossils and Climate*, pp. 291 316. New York: John Wiley and Sons.

Bruton DL (ed.) (1984) *Aspects of the Ordovician System*. Oslo: Universitetsforlaget.

Cocks LRM and Fortey RA (1982) Faunal evidence for oceanic separations in the Palaeozoic of Britain. *Journal of the Geological Society of London* 139: 467 478.

Fortey RA (1995) Trilobite life habits and Lower Palaeozoic continents. *Geoscientist* 5: 16 17.

Ross RJ (1982) The Ordovician System in the United States: correlation chart and explanatory notes. *Mono graph 12*. International Union of Geological Sciences. Denver, Colorado.

Webby BDE, Paris F, Droser ML, and Percival IG (eds.) (2004) *The Great Ordovician Biodiversification Event*. New York: Columbia University Press.

Webby BDE and Laurie J (eds.) (1992) *Global Perspectives on Ordovician Geology*. Rotterdam: Balkema.

Silurian

L R M Cocks, The Natural History Museum, London, UK

Introduction

The Silurian System, which is radioisotopically dated to have extended from 443 Ma to 418 Ma (*see* **Time Scale**), follows the Ordovician (*see* **Palaeozoic:** Ordovician) in the geological time-scale and is immediately overlain by the Devonian (*see* **Palaeozoic:** Devonian). The Silurian forms the latest part of the Lower Palaeozoic and is divided into four standard series: the Llandovery (from 443 Ma to 429 Ma), the Wenlock (from 429 Ma to 424 Ma), the Ludlow (from 424 Ma to 420 Ma), and the Pridoli (from 420 Ma to 418 Ma). The type sections for the first three are in Great Britain and the last in the Czech Republic. The base of the Silurian is marked by an unconformity in many parts of the world, since the glaciation that occurred near

the Ordovician–Silurian boundary was responsible for a global lowering of sea-level. The system is well represented in most regions today, apart from Antarctica and peninsular India, which both formed interior parts of the immense supercontinent of Gondwana.

This article describes the history, type areas, other key areas, palaeontology, tectonic activity, climate, and palaeogeography of the Silurian.

History

The Silures were an ancient British tribe who occupied southern Wales before the Roman occupation in the first century AD. The name 'Silurian' was coined by Roderick Impey Murchison (1792–1871) (*see* **Famous Geologists:** Murchison) to denote a geological system in a short paper in 1835. That paper was followed in 1839 by *The Silurian System*, a substantial book that not only described the geology of the Silurian, particularly of Wales and the Welsh Borderland of England, but also contained many illustrations and descriptions of characteristic Silurian strata and fossils. Successive editions, with the shortened title *Siluria*, followed until just after Murchison's death in 1872, by which time the scope of both the book and the system had been extended to cover the whole world (Figures 1 and 2).

The Silurian System was originally envisaged as encompassing the rocks between the Cambrian below and the Devonian above. However, it became increasingly clear that the 'Lower Silurian' of Murchison included the same rocks and geological time as the upper part of Adam Sedgwick's underlying Cambrian System. Sedgwick (*see* **Famous Geologists:** Sedgwick) was originally a friend of Murchison's (they named the Devonian together), but as time progressed fierce arguments broke out between the two men over the scope and definition of both the Cambrian and the Silurian. The matter was not resolved until after both Sedgwick and Murchison had died, when, in 1879, Charles Lapworth erected the Ordovician System to encompass both the upper part of Sedgwick's Cambrian and the lower part of Murchison's Silurian, which had at its base the Llandeilo Flags.

For many years the restricted Silurian was divided into three parts: the lowest was termed the Llandovery (or sometimes the Valentian from the graptolite-dominated faunas in the southern Uplands of Scotland), and this was followed by the Wenlock and the Ludlow. At the type area in the Ludlow area of Shropshire, marine Silurian rocks are conformably overlain by the Old Red Sandstone. At the boundary between the Ludlow and the non-marine Old Red Sandstone is a remanié deposit rich in disarticulated fish remains, known as the Ludlow Bone Bed. This thin bed was, for many years, regarded by British geologists as defining the top of the Silurian. However, in continental Europe, beds stratigraphically higher than the Ludlow Bone Bed equivalent were regarded as Silurian in age, largely because they contained graptolites. It was not until the 1960s that the views of British and continental geologists were reconciled by erecting the Pridoli Series (type area in the Bohemian part of the Czech Republic), which is now internationally accepted as a separate series within the Silurian, above the Ludlow Series and below the Devonian System.

Type Areas and Sections

The base of the Silurian System, which is also the top of the Ordovician System, the base of the Llandovery Series, and the base of the Rhuddanian Stage, has been set by international agreement within the mudstones of the Birkhill Shale Formation at the base of the *Parakidograptus acuminatus* graptolite Biozone (*sensu lato*) at Dob's Linn, near Moffat, Scotland. Some biostratigraphers, however, now divide that zone into a lower *P. ascensus* Biozone and an upper *P. acuminatus* Biozone (*sensu stricto*). The Birkhill Shales, and most of the southern Scottish Uplands in which they occur, were deposited as deep-water ocean-floor deposits and thus represent fairly continuous deposition of sediments, and faunas occurred without substantial break over the Ordovician–Silurian boundary time interval. At Dob's Linn, the underlying latest Ordovician *complexus, extraordinarius*, and *persculptus* graptolite biozones are all represented by diagnostic graptolite faunas, as are the Early Llandovery *atavus, acinaces*, and *cyphus* graptolite biozones above the *acuminatus* Biozone. Fossils other than the abundant graptolites are less common in the Dob's Linn section, but include some inarticulated brachiopods, the blind dalmanitid trilobite *Songxites*, and conodont, acritarch, and chitinozoan microfossils, the last of which are helpful in international correlation.

The Llandovery Series

Llandovery is a town standing on late Ordovician rocks in the Twyi Valley, central Wales, but the low hills to the east of the town are formed from rocks of Early Silurian age. These form the type area of the Llandovery Series (Figure 3). This series is formally divided into three stages, which are, in ascending order, the Rhuddanian, Aeronian, and Telychian. The base of the Rhuddanian, which is synonymous with the base of the Silurian, is formally defined at Dob's Linn, Scotland, some 300 km from Llandovery. Nevertheless, Rhuddanian rocks are well exposed in the type Llandovery area, consisting of up to 250 m of

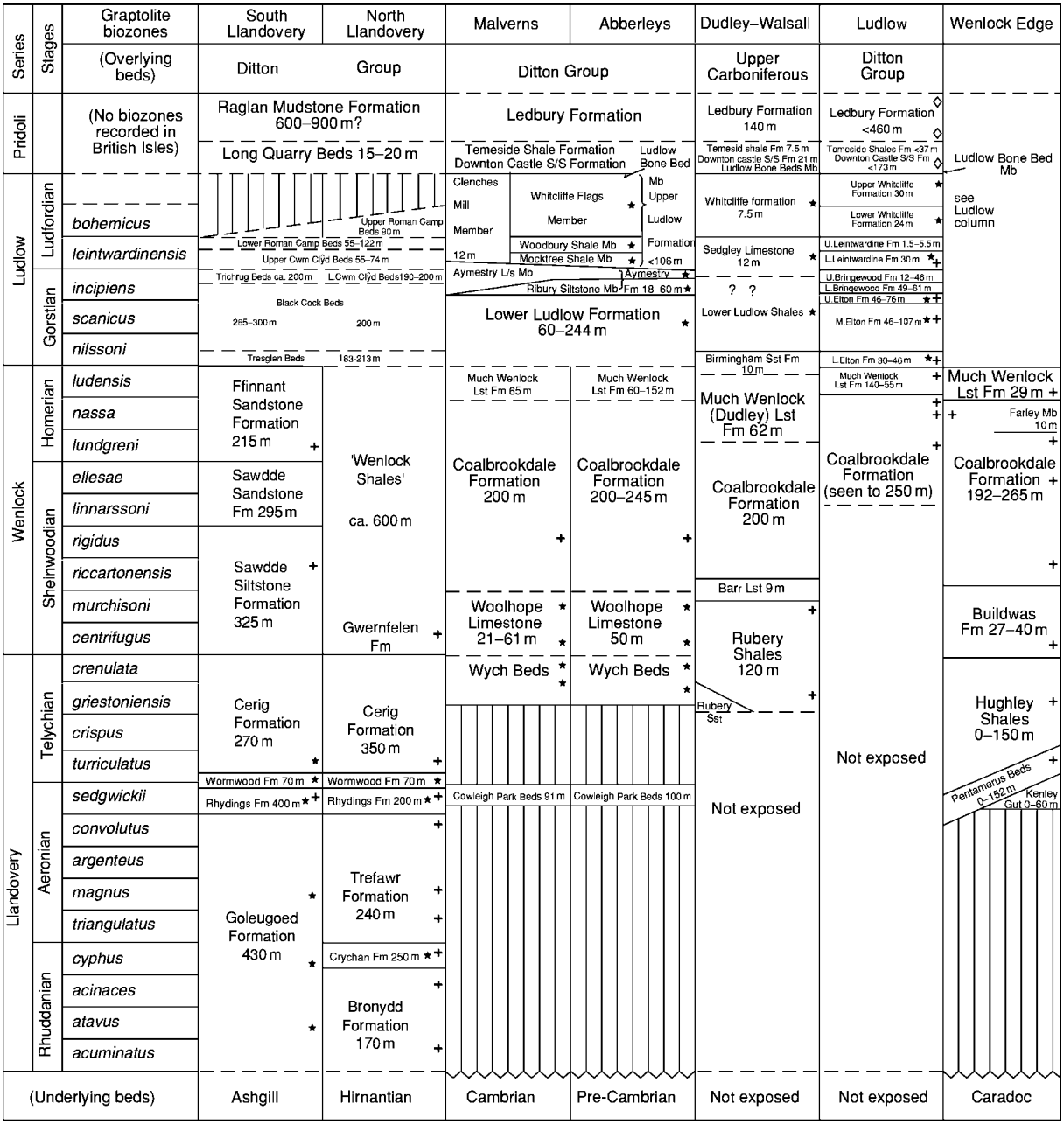


Figure 1 Correlation of key Silurian sections in the British Isles. (Reproduced from Cocks LRM, Holland CH, and Rickards RB (1992) *A Revised Correlation of Silurian Rocks in the British Isles*. Special Report 21. London: Geological Society.)

clastic rocks varying from sandstone through siltstone to shale. These sediments contain a wide variety of fossils. Macrofossils are dominated by brachiopods, including the early members of the *Stricklandia* lineage, whose species are so important for correlating Llandovery and Early Wenlock rocks internationally. Over 35 brachiopod genera are recorded from the Rhuddanian Bronydd, Crychan, and Lower Trefawr formations and their lateral equivalents at Llandovery, together with numerous trilobites, molluscs, and echinoderms and rare corals, bryozoans, and dasyclad algae. Microfossils are represented by conodonts of the *Icriodella discreta–I. deflecta* Assemblage Biozone, as well as by various acritarchs, chitinozoa, and scolecodonts.

The Aeronian Stage has its 'golden spike' base in the Forestry Commission Trefawr track near Cwm Coed Aeron Farm, Llandovery, and its base lies within the blocky mudstones of the Trefawr Formation at a horizon corresponding to the base of the *triangulatus*

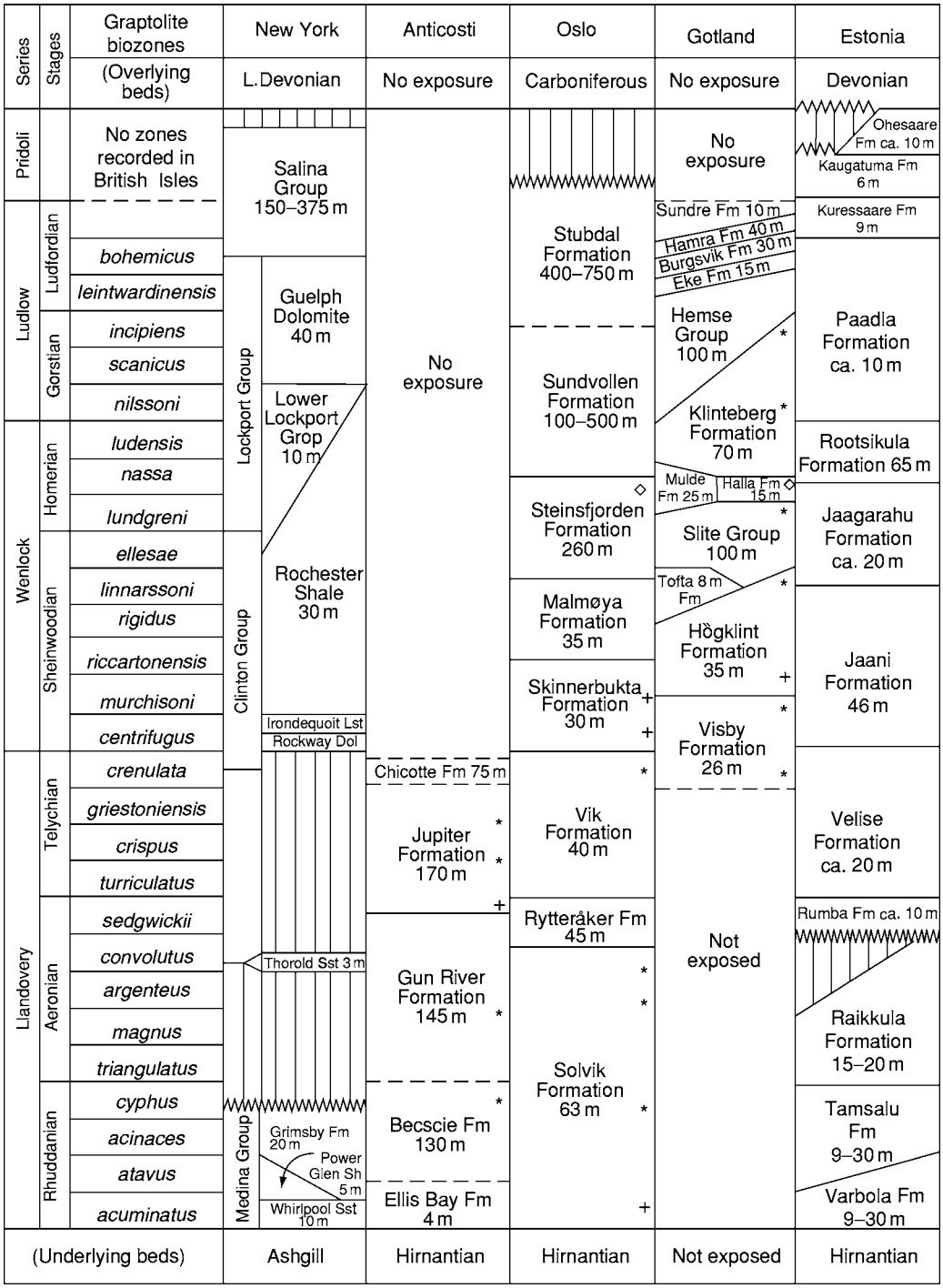


Figure 2 Correlation of key Silurian sections in New York State, Anticosti Island (Canada), the Oslo Region of Norway, the island of Gotland (Sweden), and the northern area of Estonia. (Reproduced from Cocks LRM, Holland CH, and Rickards RB (1992) *A Revised Correlation of Silurian Rocks in the British Isles*. Special Report 21. London: Geological Society.)

graptolite Biozone. The Aeronian is represented at Llandovery by over 450 m of rocks carrying, once again, an abundant macrofauna, microfauna, and microflora, including two more *Stricklandia* brachiopod species. Also present for the first time are two species of the rhynchonellide *Eocoelia*, a brachiopod genus that, in the upper half of the Llandovery, evolved even more quickly than *Stricklandi* and whose species are therefore invaluable for detailed correlation. The Aeronian formations at

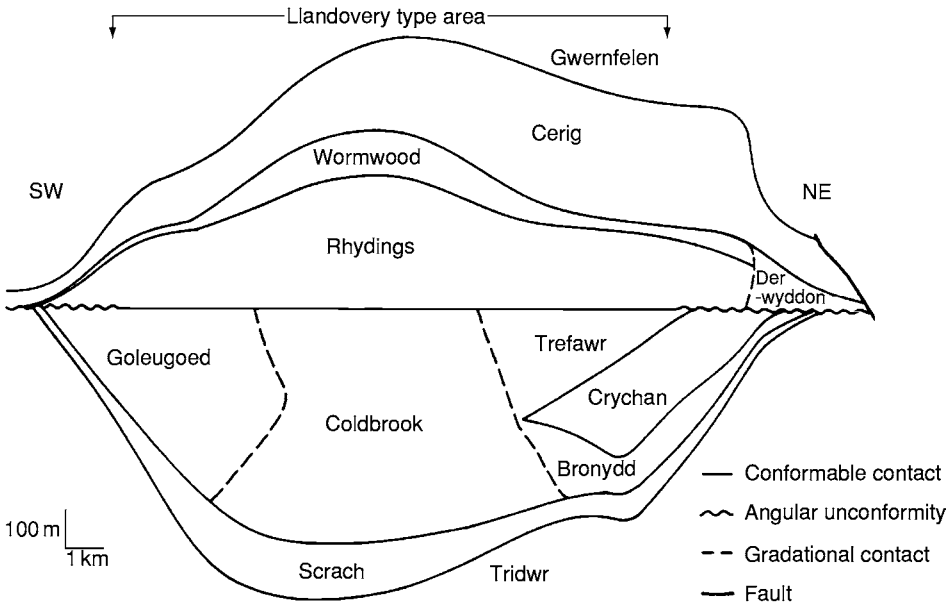


Figure 3 Schematic cross section of the formations and their thicknesses in the type Llandovery area, Wales. (Reproduced from Cocks LRM, Woodcock NH, Rickards RB, Temple JT, and Lane PD (1984) The Llandovery series of the type area. *Bulletin of the British Museum* (*Natural History*) *Geology* 38: 131 182.)

Llandovery are the Upper Trefawr, Rhydings, and Lower Wormwood Formations and their lateral equivalents.

The Telychian Stage has its formal base within a 30 cm sandstone in a small quarry near Pen-lan-Telych Farm, east of Llandovery. This sandstone occurs just below the top of the Wormwood Formation at a horizon that correlates with the base of the *turriculatus* graptolite Biozone. It is also just above the last occurrence of the brachiopod *Eocoelia intermedia* and just below the earliest record of *Eocoelia curtisi*. The higher levels of the Telychian in the type Llandovery area, the Cerig Formation, consist of over 350 m of monotonous and rather poorly fossiliferous purple mudstones. The thickness of this sequence, along with the fact that the overlying Gwernfelen Formation carries the basal Wenlock *centrifugus* Biozone graptolites, suggests that the great majority of Llandovery time is represented by rocks in the type Llandovery area.

The Wenlock Series

The Wenlock has its type area near the town of Much Wenlock in Shropshire, England, after which the series is named. The Wenlock is divided into two stages, a lower Sheinwoodian Stage and an upper Homerian Stage. The base of the Wenlock Series and the Sheinwoodian Stage is defined in Hughley Brook, Shropshire, where the Llandovery Purple (or Hughley) Shale Formation grades upwards into the overlying Buildwas Formation. This horizon, which is in an area of apparently continuous sedimentation, correlates with the base of the *centrifugus* graptolite Biozone and is accompanied by abundant and superbly preserved microfossils as well as many macrofossils, chiefly brachiopods and molluscs. The Buildwas Formation, which contains sporadic large calcareous nodules, was deposited fairly deep on the Silurian continental shelf and is about 35 m thick. This is followed conformably by the 192–265 m thick Coalbrookdale Mudstone Formation. The base of the type Homerian Stage lies within the Coalbrookdale Formation in a stream section at Whitwell Coppice, Shropshire, at a horizon corresponding to the base of the *lundgreni* graptolite Biozone. The Coalbrookdale Formation becomes more limey near its top, formally termed the Farley Member, and is overlain by the famous Much Wenlock Limestone Formation, which caps the escarpment of Wenlock Edge for more than 30 km. The Wenlock Limestone (as it is more commonly called) is only 29 m thick, but contains massive bioherms exposed in the many quarries along Wenlock Edge. The Wenlock Limestone is very extensive laterally both to the south and east of Shropshire and crops out substantially near Birmingham, where it is known locally as the Dudley Limestone, in particular at a locality named Wren's Nest. The Coalbrookdale Formation carries an abundant macrofauna, especially cephalopods and brachiopods, and varied microfossils, but the Wenlock Limestone fossils are even more varied and are exquisitely

preserved: the trilobites and corals are the best known. The bioherms consist mostly of stromatoporoids and calcareous algae, with subsidiary tabulate corals. Above the Wenlock Limestone on Wenlock Edge the strata pass up without substantial break into the overlying Elton Formation, which is of Ludlow age and extends laterally into the type Ludlow area, some 25 km away.

The Ludlow Series

The mediaeval town of Ludlow lies south-west of Much Wenlock in Shropshire, England, and the low cliffs to the south and west of the town form a shallowly plunging anticline, which includes the type rocks of the Ludlow Series and its two stages, the lower Gorstian and the upper Ludfordian. The formal base of the Ludlow Series and Gorstian Stage is located in Pitch Coppice quarry, 4.5 km south-west of Ludlow, where the Much Wenlock Limestone Formation is succeeded conformably by the Elton Formation. This boundary correlates with the base of the *Neodiversograptus nilssoni* graptolite Biozone. The 110–225 m thick Elton Formation consists largely of blocky fine siltstones, which were deposited in the middle and deeper parts of the shelf and contain many macrofossils, particularly brachiopods, trilobites, and cephalopods, and a large variety of microfossils, particularly conodonts, acritarchs, chitinozoa, and ostracods. Above the Elton Formation is the 62–105 m thick Bringewood Formation, which includes the Aymestrey Limestone of Murchison, famous for its banks of the large pentameride brachiopod *Kirkidium knightii*. The formal base of the Ludfordian Stage is at Sunnyhill Quarry, 2.5 km south-west of Ludlow, where the Bringewood Formation is overlain by the Leintwardine Formation. The base of this unit is defined in beds correlating with the base of the *Saetograptus leintwardinensis* graptolite Biozone. The 35 m thick Leintwardine Formation and the overlying 55 m thick Whitcliffe Formation both consist of siltstones and mudstones with, again, common brachiopods and cephalopods and rich microfaunas and microfloras. At the top of the Whitcliffe Formation is the Ludlow Bone Bed (see above), which is overlain by the Downton Castle Sandstone. The latter can be correlated indirectly with beds of the Pridoli Series in its type section in the Czech Republic.

The Pridoli Series

The Pridoli (technically spelt Přídolí, but the Czech accents are often omitted) Series was of much shorter duration than the other three series of the Silurian and as a result is not subdivided into stages. Its type section is in the disused Požáry Quarry at Řeporyje, in the western suburbs of Prague, Czech Republic. There the Požáry Formation is a uniform facies of dark platy limestones with intercalations of calcareous shale. Brachiopods occur commonly throughout, with abundant cephalopods at certain horizons and rich microfossil floras and faunas. The base of the Pridoli Series is defined at the first appearance of the graptolite *Monograptus parultimus*, a key constituent of the *Monograptus ultimus* graptolite Biozone. This horizon is close to the gradational boundary between the underlying Kopanina Formation and the overlying Požáry Formation. The Pridoli Series lies entirely within one conodont biozone, that of *Ozarkodina eosteinhornensis*, whose first and last appearances fall below and above the Pridoli rocks. The upper limit of the series is about 8 m above the base of the Lochkov Formation, which overlies the Požáry Formation. The top of the Pridoli Series is defined by the internationally agreed base of the Devonian System and its basal Lochkovian Stage, which is taken at the base of the *Monograptus uniformis* graptolite Biozone, whose 'golden spike' is at Klonk, also in the Czech Republic and about 23 km south-west of Požáry.

Other Key Silurian Areas

New York

Niagara Falls, in New York State, USA, and bordering Ontario, Canada, runs over a Silurian escarpment of relatively flat-lying rocks. There the Medina Group, consisting of the Whirlpool Sandstone, the Power Glen Shale, the Grimsby Formation, and the Thorold Sandstone, is of Llandovery age and is overlain paraconformably by the Clinton Group, which is of Wenlock age. The latter unit includes the very fossiliferous Rochester Shale, famous for its trilobites and other invertebrate fossils, and other formations (Figure 4). Above this lie the Lockport Group, of late Wenlock and Ludlow age, and the Salina Group, which includes many anhydrites of latest Ludlow and Pridoli age. Much of the sequence is relatively unfossiliferous, since the limestones have been largely replaced by dolomites.

Norway and Sweden

The graben running north–south through the Oslo region of Norway includes a rich succession of Silurian rocks. The Solvik, Rytteråker, and Vik formations form a Llandovery sequence that conformably overlies the latest Ordovician rocks and underlies the Wenlock rocks. The latter include the Skinnerbukta, Malmøya, and Steinsfjorden formations. The Solvik

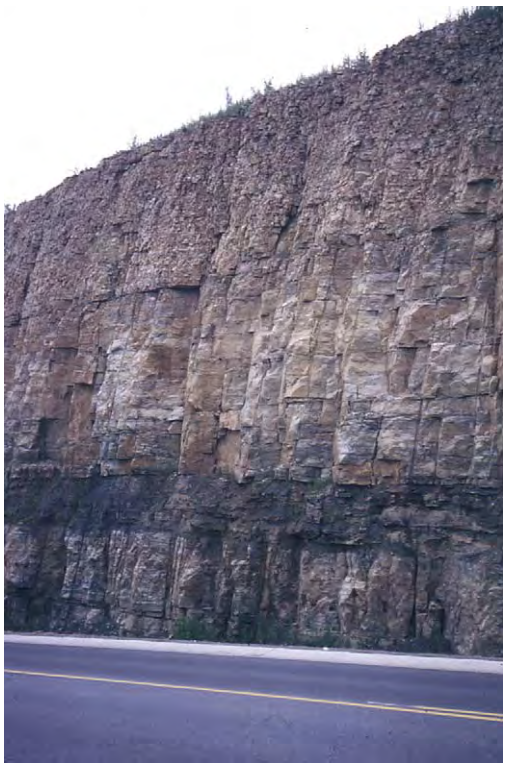


Figure 4 Section through most of the Silurian in southern Ontario, Canada, near Niagara Falls. The lower limestones are Late Llandovery and Early Wenlock; they are overlain by the Late Wenlock Rochester Shale Formation, which is in turn overlain by Ludlow, Pridoli and earliest Devonian limestones.

to the Steinsfjorden formations were all deposited in shallow-marine environments on the continental shelf and include abundant brachiopods and other macrofossils and microfossils. Above these, the Late Wenlock and Ludlow Sundvollen and Stubdal formations were deposited under freshwater or continental non-marine conditions. The Pridoli is not known in this region. On the Baltic island of Gotland, Sweden, there are flat-lying outcrops of Late Llandovery to Ludlow rocks, in many cases forming imposing sea cliffs and containing notable bioherms. The visible sequence starts with the Visby Marl, which spans the Llandovery–Wenlock boundary, and this is succeeded by the bioherms of the Högklint Formation. Ten further Wenlock and Ludlow formations follow, including the Klinteberg Formation, which spans the Wenlock–Ludlow boundary and is famous for its reefs of the pentameride brachiopod *Conchidium*, originally described by Linnaeus in the early eighteenth century.

Anticosti

Anticosti Island lies near the mouth of the St Lawrence River, Canada, and is nearly 100 km long, with a relatively flat-lying Late Ordovician and Early Silurian succession. The Silurian follows the latest Ordovician with only a minor break within the Ellis Bay Formation, which is followed by the Becscie, Gun River, Jupiter, and Chicotte formations, all of which are of Llandovery age. These rocks consist of alternating limestones and shales, all deposited on a shallow-marine continental shelf and yielding excellently preserved macrofossils (particularly brachiopods, corals, and bryozoans) and microfossils (particularly condonts, ostracods, acritarchs, and chitinozoans). Since graptolites are poorly represented, the principal correlations have been achieved by using conodont biozones.

Podolia

The area bordering the Dnestr River has been part of both Poland and the USSR in the past and is now in the west of Ukraine. A very thin band of Middle Llandovery rocks lies unconformably on Middle Ordovician rocks and is followed paraconformably by the Wenlock Kitaigorod Formation, the Bagovitsa Formation, which spans the Wenlock–Ludlow boundary, the Ludlow Malmovtsy Formation, and the Pridoli Rashkov and Dzwinogorod formations, which are themselves followed conformably by earliest Devonian strata. All are marine, with many limestones but much clastic fine-grained input, and have yielded very abundant macrofossils and microfossils. The latest Silurian rocks, locally termed the Skala Stage, were a strong candidate for the stratotype of the latest formal series of the Silurian, but the vote went to the Pridoli area of the Czech Republic.

Methods of Dating and Correlation

The system is well-constrained by radioisotopic ages down to the level of the series (see above for the dates in Ma), but within the series the dating and correlation are achieved by using quickly evolving fossils. The best of these are the graptolites, upon which more than 30 biozones have been based, covering the entire Silurian. Because of their planktonic mode of life, these quickly evolving lineages dispersed very rapidly, enabling correlation across all continents and terranes. However, graptolites are often uncommon or even absent in rocks deposited under very shallow-water conditions, and, in these rocks, microfossils, particularly conodonts, are used. However, there are often four or five successive graptolite biozones during the period of a single Silurian conodont biozone. Conodonts, acritarchs, chitinozoans, and spores can often be recovered from small samples, for example the rock chips produced through boring by

the hydrocarbon industry. In the Early Silurian (Llandovery and Early Wenlock), quickly evolving brachiopods, such as *Stricklandia* and *Eocoelia* (mentioned above), are also used for local and international correlation.

Life in the Silurian

Within the marine realm, invertebrates were abundant and diverse throughout the Silurian. The benthos was dominated by brachiopods, bivalves, gastropods, corals, stromatoporoids, bryozoans, various echinoderms (particularly crinoids), calcareous algae, and trilobites. The plankton included cephalopods (most of which swam), graptolites, and a variety of microplankton, particularly acritarchs, chitinozoans, and spores. Ostracods and conodonts were certainly mobile and in some cases dispersed quickly, but their detailed ecologies and modes of life are less certain. Fishes became progressively more abundant during the period. There is debate as to whether all Silurian fishes were marine or whether some were entirely freshwater: some fish fossils are to be found in apparently non-marine rocks. Fossils of land-living creatures are very rare, and the land was clearly not yet much colonized. Various groups of arthropods probably lived onshore at the time, as did some algae whose remains are equivocally preserved. Judging by the occurrence of fossil spores in marine rocks, some land plants probably lived as early as the Ordovician, but the oldest land plants yet represented by convincing body fossils are from rocks of Middle Silurian (Wenlock) age.

Tectonic Activity

The Silurian was considerably shorter than either the Ordovician or the Devonian, and thus there was a smaller variety of tectonic activity during the duration of the System. The docking between Avalonia and Baltica at the start of the period was relatively mild tectonically, since the collision was oblique and therefore relatively 'soft' in tectonic terms. Subsequently, during most if not all of Silurian time, there

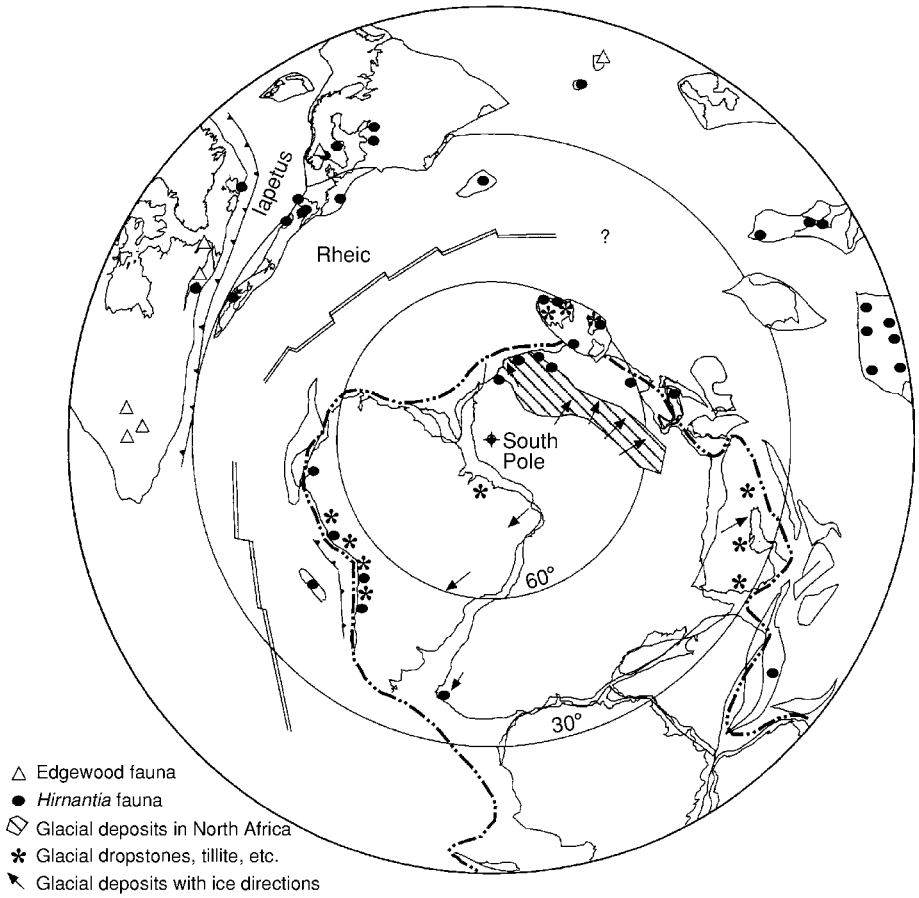


Figure 5 Palaeogeography at the Ordovician Silurian boundary (approximately 440 Ma), showing the distribution of the latest Ordovician glacial deposits and the *Hirnantia* brachiopod fauna. (Reproduced from Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. *Journal of the Geological Society, London* 159: 631 644.)

was a major collision between Avalonia–Baltica and Laurentia, resulting in a major tectonic event termed the Caledonian Orogeny (*see* **Europe:** Caledonides of Britain and Ireland; Scandinavian Caledonides (with Greenland)), which led to mountain building and metamorphism along a belt that today embraces all of the North Atlantic rim from the eastern USA (Carolina) to Norway, including eastern North America, Ireland and Britain, and the Scandinavian Caledonides of Norway and Sweden. Away from this area, there appears to have been less tectonism, apart from in today's central Asia, where a number of small terranes and their marginal island arcs were breaking up and regrouping, particularly in Kazakhstan and adjacent countries. Around the vast Gondwanan margin there were also island arcs and tectonic activity in south-east Australia (Victoria and New South Wales) and New Zealand.

Palaeogeography and Climate

Silurian geography was dominated by the vast supercontinent of Gondwana (*see* **Gondwanaland and Gondwana**), which included today's South America, Africa, peninsular India, Antarctica, and Australasia, as well as adjacent terranes in southern Europe and the Middle and Far East, which may or may not have been physically attached to the supercontinent (Figure 5). Other substantial terranes were Laurentia (which included most of North America, Greenland, and the easternmost part of Siberia), Siberia (which was only part of today's political entity), and Baltica–Avalonia. Baltica, which today comprises most of northern Europe eastwards to the Ural Mountains, had been an independent terrane until about the Ordovician–Silurian boundary (443 Ma), when it collided with Avalonia, whose long narrow

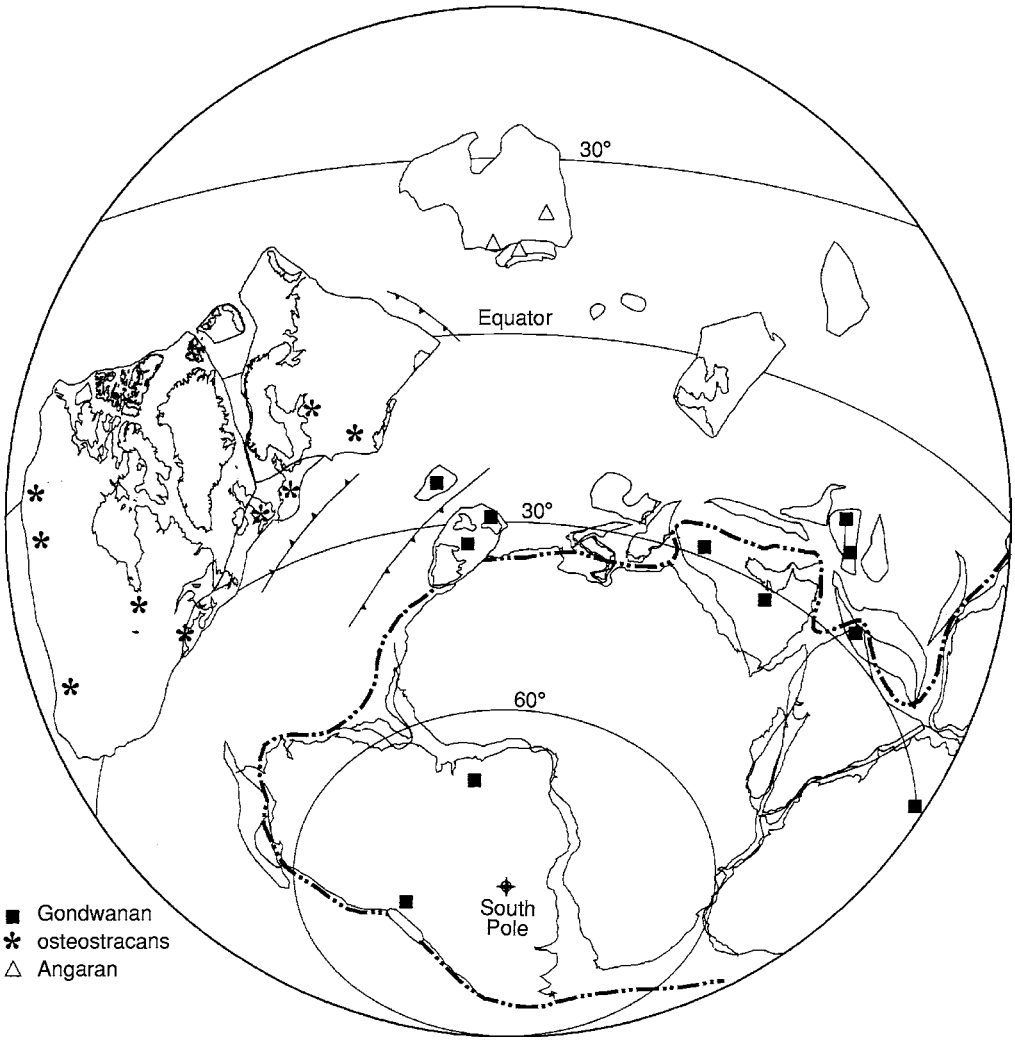


Figure 6 Palaeogeography at the Silurian Devonian boundary (approximately 420 Ma), showing the distribution of the earliest Devonian fishes. (Reproduced from Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. *Journal of the Geological Society, London* 159: 631 644.)

area today includes the Maritime States of Canada and the USA as far south as Cape Cod, Massachusetts, eastern Newfoundland, southern Ireland, England, Wales, Belgium, Holland, and parts of northwestern Germany. North and South China were also independent terranes, as was Sibumasu, which stretched from Burma through Thailand and western Malaysia to Sumatra in Indonesia. There were also many smaller terranes, including Perunica (Bohemia, Czech Republic).

The oceans were dominated by the vast Panthalassic Ocean, which was comparable in size to today's Pacific. The Iapetus Ocean between Laurentia and Avalonia–Baltica, which had been very substantial in Cambrian and Ordovician times, was steadily closing. During the Late Ordovician some of the island arcs within Iapetus collided with and accreted to the eastern margin of Laurentia or the western margin of Avalonia–Baltica. The Silurian saw the major continent–continent collision between Laurentia and Avalonia–Baltica, which resulted in the closure of Iapetus and the Caledonian Orogeny (see above). This orogeny caused considerable uplift and formed the land supporting the Old Red Sandstone Continent in the Late Silurian and Devonian (**Figure 6**). The combined terrane that resulted from the Laurentia–Avalonia–Baltica collision is termed Laurussia. To the south of the Avalonian part of Laurussia the widening ocean between it and Gondwana is termed the Rheic Ocean.

Over a vast area of the continental shelf of Gondwana, today including France, Bohemia, the Iberian Peninsula, Italy, and much of northern Africa, anoxic conditions prevailed, precluding the colonization by benthos. The fossils preserved in the black shales deposited there consist largely of plankton, including graptolites and cephalopods. It was not until Late Silurian times in some of these areas, for example Bohemia, that the seafloor became oxygenated, enabling its colonization by brachiopods, crinoids, and other benthos.

The climate of the Silurian started cold. The latest Ordovician (Hirnantian) ice-cap was very widespread, but gradually receded in Llandovery times; the latest known Silurian glacial sediments are in the early Wenlock of Brazil, which was near the then south pole. It is not until the Late Llandovery (Telychian) that warmer-water carbonate deposits become common, but from then on into the Wenlock, Ludlow, and Pridoli the global temperature was obviously warm to hot. This can be deduced from the carbonate build-ups and bioherms present in the Wenlock type area of England and in Gotland and elsewhere and also from the Late Silurian evaporite deposits, which are widespread but best known from New York State, USA.

See Also

Europe: Caledonides of Britain and Ireland; Scandinavian Caledonides (with Greenland). **Famous Geologists:** Murchison; Sedgwick. **Fossil Invertebrates:** Brachiopods; Graptolites. **Gondwanaland and Gondwana**. **Palaeozoic:** Ordovician; Devonian. **Time Scale**.

Further Reading

Bassett MG, Cocks LRM, Holland CH, Rickards RB, and Warren PT (1975) The type Wenlock Series. *Institute of Geological Sciences Report* 77(13): 1 33.

Cocks LRM and Rickards RB (eds.) (1988) *A global analysis of the Ordovician Silurian boundary. Bulletin of the British Museum (Natural History) Geology* 43: 1 394.

Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomag netic review. *Journal of the Geological Society, London* 159: 631 644.

Cocks LRM, Woodcock NH, Rickards RB, Temple JT, and Lane PD (1984) The Llandovery series of the type area. *Bulletin of the British Museum (Natural History) Geology* 38: 131 182.

Cocks LRM, Holland CH, and Rickards RB (1992) *A Revised Correlation of Silurian Rocks in the British Isles*. Special Report 21. London: Geological Society.

Holland CH and Bassett MG (eds.) (1989) *A global stand ard for the Silurian System National Museum of Wales Geological Series* 9: 1 325.

Holland CH, Lawson JD, and Walmsley VG (1963) The Silurian rocks of the Ludlow District, Shropshire. *Bulletin of the British Museum (Natural History) Geology* 8: 93 171.

Landing E and Johnson ME (eds.) (2003) Silurian lands and seas: paleogeography outside of Laurentia. *New York State Museum Bulletin* 493: 1 400.

Martinsson A (ed.) (1977) *The Silurian Devonian Bound ary*. Stuttgart: Schweizerbart'sche.

Sengor AMC and Natalin BA (1996) Paleotectonics of Asia: Fragments of a synthesis. In: Yin A and Harrison M (eds.) *The Tectonic Evolution of Asia,* pp. 486 640. Cambridge: Cambridge University Press.

Devonian

G R McGhee, Rutgers University, New Brunswick, NJ, USA

Introduction

The Devonian Period of the Palaeozoic Era was only slightly shorter than the entire Cenozoic Era (63 million years as opposed to 65) and was a time of many 'firsts' in Earth history. The Devonian saw the explosive diversification of vertebrate life (e.g., new fish species that proliferated in both marine and freshwater environments) including the evolution of the first terrestrial vertebrates, the amphibians. In the oceans, the Devonian saw the evolution of the largest reef ecosystems in Earth history. Both plant and animal groups were rapidly evolving and invading the terrestrial realm, and in the Devonian the Earth's first forests evolved with trees that towered some 30 metres high. Yet it is also in the Devonian that the Earth suffered one of its greatest biodiversity crises, one of the 'Big Five' mass extinctions of life, and it is in the Devonian that the entire climate of the planet switched from a hot greenhouse state to a cold icehouse phase.

The Devonian System is named after fossiliferous marine strata exposed in Devonshire, southwest England. The English palaeontologists Adam Sedgwick (*see* **Famous Geologists:** Sedgwick) and Roderick Murchinson (*see* **Famous Geologists:** Murchison) first proposed that the fossils in Devon were younger than Silurian fossils, yet older than Carboniferous, in 1839. Subsequent stratigraphic work demonstrated that the Old Red Sandstone of Scotland and Wales, long thought to lie in the basal Carboniferous, was the non-marine facies equivalent of Devonian marine strata. The Devonian Period is divided into seven Ages, from oldest to youngest: the Lochkovian, Pragian, and Emsian (constituting the Early Devonian), the Eifelian and Givetian (Middle Devonian), and the Frasnian and Famennian (Late Devonian). Lochkovian and Pragian are in the Czech Republic, Emsian and Eifelian in Germany, and Givetian, Frasnian and Famennian in Belgium.

Life in the Sea

The Devonian was a time of major evolutionary innovation in the marine realm. The Devonian has been called the 'Age of Fishes', as it is in this period that the evolution of the major fish groups of today's oceans occurred (*see* **Fossil Vertebrates:** Fish). The explosive diversification of bony, ray-finned fishes (the Actinopterygii) took place in the Devonian, as did the origination of the first true sharks. Of particular note is the origination of the first lobe-finned fishes (the Sarcopterygii) as this group later gave rise to the amphibians, also in the Devonian. The ammonoids – externally shelled, swimming cephalopods distantly related to the modern-day chambered nautilus – first evolved in Devonian seas (*see* **Fossil Invertebrates:** Cephalopods (Other Than Ammonites)). These animals were active swimmers with widespread geographic ranges and possessed a characteristic high rate of speciation, two traits that made them valuable for biostratigraphic dating of sedimentary rocks for over 300 million years before their lineage perished in the Cretaceous–Tertiary mass extinction.

The largest development of reefal ecosystems in Earth history occurred in the Devonian. The great Devonian reefs are estimated to have covered over 5 000 000 square kilometres of seafloor at their maximum development, almost ten times the areal extent of reefal ecosystems seen in modern oceans. The Devonian was also the 'Golden Age of Biconvex Brachiopods' (*see* **Fossil Invertebrates:** Brachiopods), the period of maximum diversification of these 'shellfish of the Palaeozoic', an ecological role that is dominated by the bivalve molluscs today.

We are terrestrial animals, and we tend to think of life in terms of life on land, rather than in the seas. It was during the Devonian that life staged its dramatic invasion of the land, and that the origination of the first complex terrestrial ecosystems occurred.

The Invasion of the Land

The Plant Invasion

Life existed in the Earth's oceans for over three billion years before finally venturing out onto the harsh environments of dry land. The biological invasion of the land has deep roots, beginning with terrestrial autotrophic microbes in the Cambrian and progressing to simple mosses and liverworts in the Ordovician and Silurian. It was only in the Late Silurian to Early Devonian, however, that life's invasion of the land dramatically accelerated with the evolution of vascular land plants.

In the Lochkovian vascular land plants were confined mostly to lowland floodplains and stream margins. These plants were simple in structure, many with short vertical stems sometimes covered

with scale-like leaves (resembling asparagus spears), others with short vertical stems that branched into progressively smaller forks towards the top of the plant. The tallest of these plants was less than half a metre high, and they would remain the tallest terrestrial organisms on Earth for the first 17 million years of the Devonian (Figure 1).

By the time of the Emsian, competition for sunlight and space resulted in the development of woody tissues by several plant groups. Such support structures enabled plants that possessed them to grow taller than their neighbours. These included very primitive ferns and small tree-like lycopods (relatives of modern club mosses) that could reach two to three metres in height, and they were the tallest plants on Earth for the next 20 million years (Figure 1).

Vegetative height increased dramatically again in the Late Eifelian to Early Givetian with the evolution of arborescent plants that are characterized by secondary wood. The Earth's first forests evolved in the Givetian, and rapidly spread across the continents in the hot climates of the Frasnian, from tropical to almost subpolar regions. In the Early Frasnian the first true woody trees appeared, the progymnosperms of the genus *Archaeopteris* (ancient relatives of modern nonflowering evergreens). *Archaeopteris* trees towered from 10 to 30 metres high in Late Devonian forest landscapes (*see* **Fossil Plants:** Gymnosperms). These were accompanied by strange sphenopsid trees (which looked like gigantic reeds) and lycopsid trees 10 to 20 metres tall.

The last major Devonian advance in plant evolution was the evolution of seed plants in the Late Famennian. Reproduction by seeds (rather than spores) freed land plants from their confinement to moist lowland areas and allowed them to spread into much drier inland areas and into the harsher climates of highlands (Figure 1).

The evolution of increasing vertical height in land plants during the Devonian is mirrored in the evolution of deeper and deeper root penetration into the ground below, and the progressive formation and spread of new soil types. In the Late Pragian some of the herbaceous lycopods penetrated just a single centimetre into the ground. Root penetration increased to 20 centimetres with the evolution of arborescence in the Givetian, and in the Frasnian *Archaeopteris* trees had roots that penetrated to depths of over 100 centimetres. Both the evolution of land plants themselves, and their effects on soil formation (*see* **Soils:** Modern) and rock weathering on the land, were to have a profound impact on both the Earth's atmosphere (*see* **Atmosphere Evolution**) and hydrosphere, and precipitated the global climatic shift from the Early Devonian 'greenhouse' climate to the Late Devonian 'icehouse' climate.

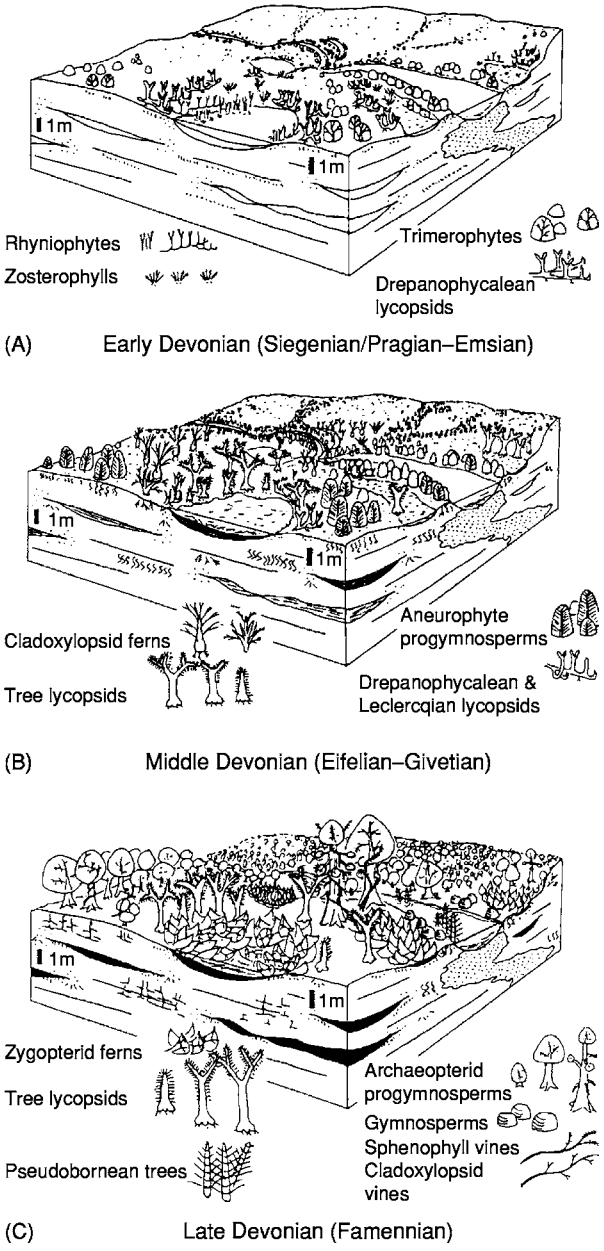


Figure 1 Palaeobotanical and palaeoecological reconstructions of Early Devonian (A), Middle Devonian (B), and Late Devonian (C) upland floodplains. (Adapted from Algeo, Scheckler and Maynard (2001), used with permission.)

The Animal Invasion

Plants are autotrophic organisms, producing their own food by photosynthesis. Plants are used as a food source by animals, and it is no coincidence that the evolution of land plants was quickly followed by the appearance of land animals.

Although millipedes may have made the transition to land as early as the Ordovician, it is in the Late

Silurian and Early Devonian that the animal invasion of the land began in earnest. Millipedes, centipedes, and extinct spider-like arachnids were present on land in the Late Silurian, and became much more numerous in the Early Devonian. Scorpions successfully invaded freshwater streams in the Late Silurian, and emerged onto dry land in the Early Devonian. The earliest known insect-like hexapods appeared in the Pragian, ancient ancestors of the insects that today comprise over 70% of all animal species.

Larger vertebrates were soon to follow, as many of these used the smaller arthropods (*see* **Fossil Invertebrates:** Arthropods) as a food source. The Devonian is justly called the 'Age of Fishes' due to the rapid diversification of numerous fish species that occurred during this period. Many of these species invaded freshwater habitats on land. The most important of these is *Eusthenopteron*, a lobe-finned fish that is anatomically very similar to – and has been cited as the direct ancestor of – the well-known Famennian amphibian *Ichthyostega*. Fossil trackways and fragmentary skeletal material suggests that the evolution of amphibians, the ancestors of all land dwelling vertebrates today, probably took place in the Frasnian.

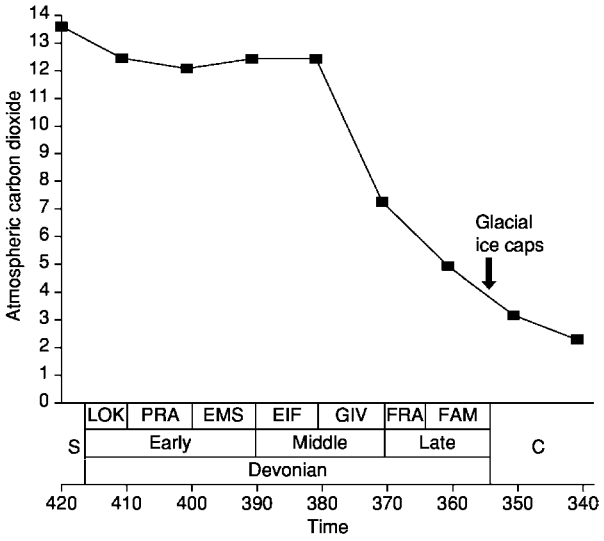


Figure 2 Atmospheric carbon dioxide concentrations during the Devonian, where the amount of carbon dioxide in the atmos phere is measured in factors of the carbon dioxide concentration in the atmosphere at the present (taken as the preindustrial level of 300 ppm). Timescale is in million of years before the present. Abbreviations: S Silurian, LOK Lochkovian, PRA Pragian, EMS Emsian, GIV Givetian, FRA Frasnian, FAM Famen nian. Data from Berner (2001), timescale after Gradstein and Ogg (1996).

The Devonian Global Climatic Shift

The Devonian is a critical period in the Earth's climate history. During this period the global climate switched from the hot greenhouse phase of the earlier Palaeozoic to the cold icehouse phase of the later Palaeozoic, a phase that would lead to continental glaciations in the succeeding Carboniferous and Permian.

The Devonian cooling trend was driven by a sharp decline in the partial pressure of carbon dioxide in the atmosphere from the Givetian to the Early Carboniferous (Figure 2), a decline driven by the establishment of extensive terrestrial land plant cover (discussed in the previous section) and associated soil weathering. Land plants themselves fixed atmospheric carbon in their tissues (*see* **Carbon Cycle**), and nutrient run-off produced by their activity on land triggered blooms of phytoplankton growth in the Devonian shallow seas. This, in turn, led to further atmospheric carbon fixing and organic carbon burial in extensive Middle and Late Devonian black-shale deposits. Chemical weathering of rocks by land plants also produced massive run-off of calcium and magnesium-rich minerals which precipitated as carbonates in the oceans, further drawing down the carbon dioxide partial pressure of the atmosphere.

From the Lochkovian through to the Eifelian the carbon dioxide concentration in the Earth's atmosphere was some 12 to 14 times greater than that of the modern atmosphere (Figure 2). By the end of the Givetian the carbon dioxide concentration had dropped to a little over seven times as great as the present, and in the Early Carboniferous the carbon dioxide content of the atmosphere was only about twice that of the modern atmosphere (Figure 2). Carbon dioxide is the most well-known of the greenhouse gasses, gasses that promote heat retention in the Earth's atmosphere. It is clear that the dramatic decline of carbon dioxide partial pressures in the atmosphere during the 40 million years from the Givetian to the Early Carboniferous must have made the atmosphere more transparent to heat loss, and it is no coincidence that glacial ice-caps formed and spread on the Earth's southern continent of Gondwana in the latest Devonian (Figure 2).

Late Devonian Biodiversity Crises

Life diversified and proliferated in the world's oceans during the first 30 million years or so of the Devonian (Figure 3). Then the trend reversed. Marine biodiversity began to decline in the Emsian and continued to decline, first slowly and then very abruptly, from the Eifelian through the Frasnian. Life rediversified briefly during the Famennian, but then marine biodiversity declined to a very substantial low at the close of the Devonian (Figure 3).

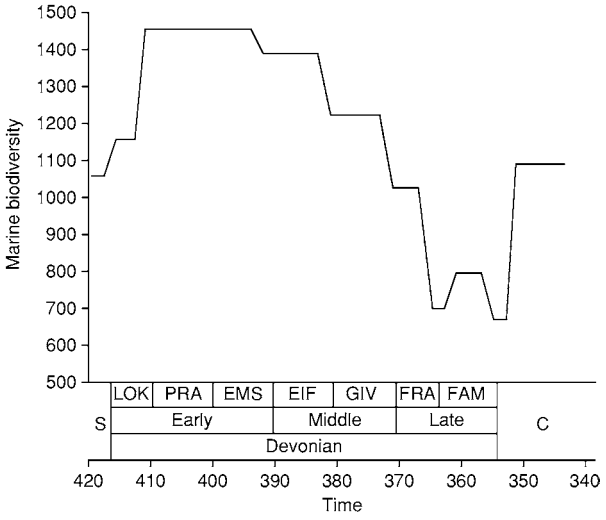


Figure 3 Marine biodiversity during the Devonian, where bio diversity is measured as the number of genera of marine organ isms that are well preserved in the fossil record. (Data from Sepkoski (1996).)

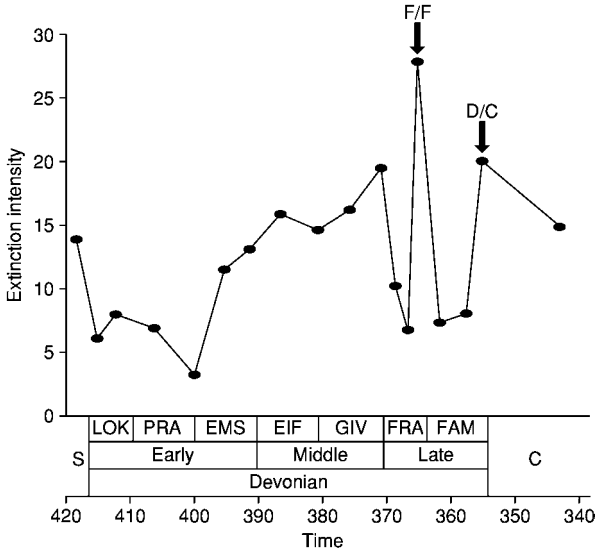


Figure 4 Extinction intensities during the Devonian, where intensity is measured as the percent of marine genera that go extinct per time interval. Positions of the Late Devonian (F/F) and end Devonian (D/C) extinction peaks are marked with arrows. (Data from Sepkoski (1996).)

The factors controlling the Devonian pattern of biodiversity gain and loss are complex, and not all factors are agreed upon to the present day. In general, the biodiversity gain seen in the Lochkovian was a recovery of biodiversity lost in extinctions that occurred in the Late Silurian. The Pragian saw the evolution of many new taxa, pushing Devonian marine biodiversity to levels higher than any that had existed previously in the Palaeozoic, but many of these new taxa were endemic, however, and having restricted geographic distributions and confined to local regions. Global sea-level began to rise in the Emsian and continued to rise through the Frasnian (perhaps due to the initiation of continental rifting in the Ukraine region of Europe), gradually breaking down barriers between marine regions and facilitating species migration. With the progress of rising sea-levels and increased species migration in the oceans, endemic local faunas successively became extinct and were replaced by more cosmopolitan migrants. By the Early Frasnian, biodiversity in the Devonian seas had dropped to levels comparable to those seen at the beginning of the Devonian (Figure 3), but these seas were now populated by cosmopolitan species with wide geographic distributions.

The gradual decline in marine biodiversity from the Emsian through to the Givetian was mirrored in the gradual increase in extinction intensity within marine faunas during this same interval (Figure 4). Extinction intensity reached a peak in the Late Givetian, with the elimination of most of the previously existing endemic faunas, and then fell to levels comparable with those seen in the Lochkovian and Pragian and within the cosmopolitan faunas of the Frasnian (Figure 4). Very abruptly, extinction intensity within these same cosmopolitan faunas jumped to a level seen only five times in Earth history, and precipitated an equally abrupt loss of marine biodiversity (Figure 3). This is the Late Devonian mass extinction, also termed the Frasnian/Famennian (F/F) mass extinction or the Kellwasser bioevent, and was the first of the Late Devonian biodiversity crises.

The second Late Devonian biodiversity crisis occurred at the very end of the Devonian, and is variously termed the end-Devonian extinction, the Devonian/Carboniferous (D/C) extinction, or the Hangenberg bioevent. The extinction intensity of the end-Devonian extinction was only slightly more severe than that seen in the Late Givetian (Figure 4), but was markedly different in that it was an abrupt jump in extinction, rather than the culmination of a long trend of increasing extinction magnitudes (Figure 4).

Causes of the Late Devonian Biodiversity Crises?

The End-Devonian Biodiversity Crisis

It is generally agreed that the end-Devonian biodiversity crisis, or Hangenberg bioevent, was triggered by global cooling and associated continental glaciation. In the very latest Famennian, just before the end of the Devonian, extensive glacial ice-caps formed in regions

of Gondwana (present day South America), ice-caps that extended to sea-level. Global sea level had begun to fall earlier in the Famennian, and continued to fall throughout the Famennian, indicating the probable formation of high-altitude mountain glaciers considerably before the end of the Famennian.

The end-Devonian extinctions began slightly earlier in the marine realm than the terrestrial, but the entire period of extinctions appears to have spanned only 100 000 years. Although the observed extinctions of marine animals were not as great as those in the Late Devonian mass extinction (Figure 4), the effect of the end-Devonian extinction on land plants and on phytoplankton in the sea was much more severe. In particular, the towering *Archaeopteris* trees that had been so widespread and characteristic of Late Devonian forests rapidly declined in abundance during the Late Famennian climatic cooling, barely surviving into the Early Carboniferous before dying out.

The end-Devonian ice caps were clearly the result of global cooling, which itself was most likely triggered by the dramatic decrease in atmospheric carbon dioxide throughout the latter half of the Devonian (Figure 2). By the end-Devonian, carbon dioxide concentrations in the atmosphere were only three to four times higher than those at present (Figure 2).

The Late Devonian Biodiversity Crisis

The principal cause of the Late Devonian extinction also appears to have been global cooling, but the pattern of cooling has a very different signature than that seen at the end of the Devonian, and the cause of the cooling remains controversial. Palaeobotanical evidence shows that the Frasnian world was quite warm, with extensive shallow seas covering much of the continental landmasses. In these warm seas the great Devonian reefs (*see* **Fossil Invertebrates:** Corals and Other Cnidaria), and other high-diversity marine ecosystems, proliferated. In contrast to Frasnian climates, the Early Famennian world appears to have been cold and arid. The great forests shrank in their areal extent, and were confined to low-latitude equatorial regions. The uplands and high-latitude regions of the Earth supported only very sparse vegetation in conditions that were not only very cold but were also very dry, without perennial snow. In the oceans, huge reef tracts died out all over the world during the latest Frasnian. The few reef organisms that managed to survive into the Famennian were confined to what few warm waters still existed on the Earth, in restricted regions along the equator.

It is still not clear what triggered the collapse of the Frasnian hot climates, and the rapid fall of global temperature into the cold Famennian world. The rapidity of the Late Frasnian cooling is particularly noteworthy, and quite different from the gradual pattern of cooling seen at the end of the Devonian that culminated in continental glaciation. Geochemical evidence indicates that tropical sea-surface temperatures fell by as much as 5° to 7°C in two abrupt pulses, separated by about 300 000 years, in the Late Frasnian to Early Famennian interval.

Three competing hypotheses are at present being debated concerning the ultimate cause of the Late Devonian biodiversity crisis. The 'Devonian plant hypothesis' proposes that both the Late Devonian and end-Devonian biodiversity crises were triggered by the evolution and spread of extensive land plant cover (Figure 1), and the effects of that new terrestrial plant biomass on the atmosphere (Figure 2) and on the oceans (particularly the massive increase in nutrient input to marine ecosystems). However, the Late Devonian biodiversity crisis occurred some 10 million years before the end of the Devonian (Figure 4). All attempts to find geologic evidence of glaciation in the Late Frasnian to Early Famennian interval to date have failed. Global sea-level was at an all-time high for the Devonian, indicating there was no major build-up of glacial ice in this time interval (although apparently rapid oscillations in sea-level during the Late Frasnian and Early Famennian may indicate the formation of small, high-altitude alpine-type glaciers). Atmospheric carbon dioxide concentrations were six times higher than at the present day (Figure 2), and approximately twice as high as those present during the formation of the end-Devonian glacial ice-caps. Many argue that these facts, particularly the rapidity of the temperature drops during the Late Devonian biodiversity crisis, indicate a substantially different triggering mechanism than that seen for the end-Devonian extinctions.

The 'volcanic winter hypothesis' proposes that an interval of catastrophic flood-basalt volcanism occurred during the Late Frasnian. One of the climatic effects of such extensive volcanism is hypothesized to have been rapid global cooling, due to the vast amount of debris, ash, and gasses injected into the Earth's atmosphere. A volcanic-produced global dust and gas cloud would have blocked light from the sun, from reaching the Earth's surface, triggering planet-wide lethally cold temperatures even at the equator.

The catastrophic volcanism scenario is based upon the recognition that there have been other flood-basalt episodes in Earth history, during which enormous amounts of lava, gasses, and volcanic dust have been produced in volcanic eruptions of almost unimaginable magnitude. These flood-basalt fissure eruptions are produced by gigantic plumes of molten rock that originate deep in the Earth's mantle, which then

slowly rise to produce paroxysms of volcanic eruptions over huge geographic areas when they intersect the Earth's surface. Several flood-basalt episodes are known to have occurred at other times of biotic crisis, such as the eruption of the Siberian flood basalts during the Permo-Triassic mass extinction and the Deccan flood basalts during the Cretaceous–Tertiary mass extinction.

Vast regions of the bottoms of the Earth's oceans were depleted in oxygen during the Late Devonian. It has been argued that the great geographic extent of these anoxic water masses were produced in part by extensive submarine volcanism, and hence might be evidence for catastrophic volcanic episodes during the Late Devonian. In addition, a major continental rift-system is now known to have been active in the Ukraine region of Europe during the Late Devonian (the Pripyat–Dnieper–Donet rift), and many of the Earth's flood-basalt fissure eruptions are associated with rifting and spreading of the Earth's tectonic plates.

However, the volume of volcanic material erupted in the Pripyat–Dnieper–Donet rift appears to be relatively small (less than 10 000 cubic kilometres), and thus the intensity of the volcanism associated with the rift was not near the magnitude associated with the Siberian or Deccan flood basalt fissure eruptions (both of which produced well over a million cubic kilometres of volcanic material). Likewise, the extensive spread of anoxic bottom waters in the Late Devonian seas, and the characteristic black-shale deposits within these seas, has been argued to have been due more likely to greatly enhanced organic productivity in the phytoplankton. The greatly increased influx of nutrients into the oceans, produced by the effects of extensive land plant cover on the land (**Figure 1**), is believed to have produced widespread eutrophication in Devonian shallow seas. Bacterial degradation of phytoplankton rain and burial of the massive amounts of organic carbon fixed by eutrophication is a more probable cause of the Late Devonian black-shale deposits than submarine volcanism.

Another hypothetical way to produce a rapid drop in global temperature would be to impact the Earth with a large asteroid or comet, in essence to produce an 'impact winter' instead of a 'volcanic winter'. As in the case of the hypothesised Late Frasnian glaciations, extensive searches for geologic evidence of major impact events in the Late Frasnian have produced very few hard data. In contrast, considerable evidence exists for several impact events in the Early Frasnian, but about three million years before the biodiversity crisis in the Late Frasnian. In particular, three impact events are currently dated to have occurred between 369 to 367 million years ago: the Alamo impactor, which produced ejecta debris that covers over 19 000 square kilometres of southern Nevada (USA) and which is estimated to have produced a crater 70 kilometres in diameter, the Siljan impactor, which produced a 52-kilometre-diameter crater in Sweden, and the Flynn Creek impactor, which produced a 3.5-kilometre-diameter crater in Tennessee (USA). These impacts all occurred on the Earth's continents; however, an unknown number of additional impacts may also have occurred in the Earth's oceans, for which we have no geologic record.

The 'lag-time multiple impacts hypothesis' proposes that these Early Frasnian impacts produced an abrupt increase in global temperature rather than a decrease. The Alamo impactor in particular struck carbonate target rock, and would have produced a massive injection of carbon dioxide into the atmosphere. The climatic effect of these known Early Frasnian impactors, and of other impacts as yet poorly dated but known to have occurred during the span of the Late Devonian, is argued to have produced an anomalous greenhouse interval that interrupted the Givetian to Carboniferous gradual decline in global temperatures (**Figure 5**). The collapse of this anomalous greenhouse interval, and the rapid drop in global temperatures produced by the resumption of the pre-existing global cooling trend, is proposed to have produced the observed rapidity of temperature fall seen during the Late Devonian biodiversity crisis (**Figure 5**). This hypothesis would also account for

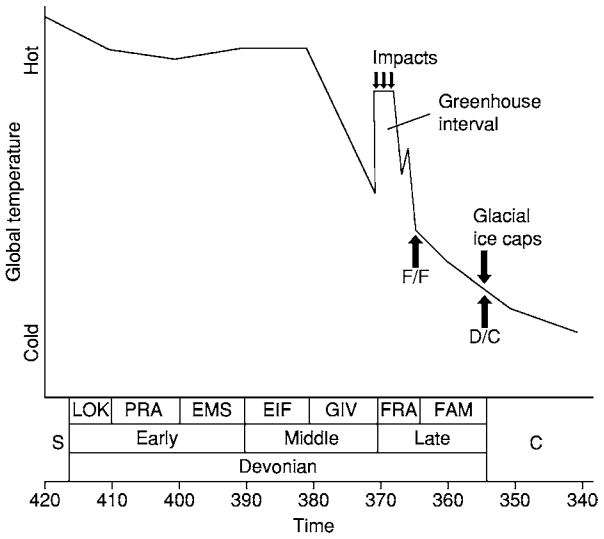


Figure 5 A proposed temperature curve for the Late Devonian, based upon the lag time multiple impacts hypothesis. Several known impacts that occurred in the Early Frasnian are proposed to have produced an anomalous greenhouse interval, interrupting the gradual cooling trend from the Givetian to the Carboniferous. Positions of the Late Devonian (F/F) and end Devonian (D/C) biodiversity crises are marked with arrows.

the apparent anomaly of hot Frasnian climates immediately preceding cold Famennian ones, indicated by the palaeobotanical evidence, whereas a gradual cooling trend produced by gradually decreasing atmospheric carbon dioxide levels (**Figure 2**) would predict Late Frasnian climates only slightly warmer, and different, from the Early Famennian.

The ultimate test of any of the proposed causal scenarios for the Late Devonian biodiversity crisis rests upon the construction of an accurate and detailed temperature curve for the entire Late Devonian period of time. Such a task will be laborious, but it is now technically feasible with oxygen-isotope analyses and laser-based sampling of conodont elements.

See Also

Atmosphere Evolution. **Carbon Cycle**. **Earth Structure and Origins**. **Famous Geologists:** Murchison; Sedgwick. **Fossil Invertebrates:** Arthropods; Brachiopods; Corals and Other Cnidaria; Cephalopods (Other Than Ammonites); Porifera. **Fossil Plants:** Gymnosperms. **Fossil Vertebrates:** Jawless Fish-Like Vertebrates; Fish. **Microfossils:** Conodonts. **Soils:** Modern. **Unidirectional Aqueous Flow**.

Further Reading

Algeo TJ, Scheckler SE, and Maynard JB (2001) Effects of the Middle to Late Devonian spread of vascular land plants on weathering regimes, marine biotas, and global climate. In: Gensel PG and Edwards D (eds.) *Plants Invade the Land*, pp. 213 236. New York: Columbia University Press.

Berner RA (2001) The effect of the rise of land plants on atmosphere CO_2 during the Paleozoic. In: Gensel PG and Edwards D (eds.) *Plants Invade the Land*, pp. 173 178. New York: Columbia University Press.

Dineley DL (1984) *Aspects of a Stratigraphic System: the Devonian*. New York: John Wiley & Sons.

Droser ML, Bottjer DJ, Sheehan PM, and McGhee GR (2000) Decoupling of taxonomic and ecologic severity of Phanerozoic marine mass extinctions. *Geology* 28: 675 678.

Filer JK (2002) Late Frasnian sedimentation cycles in the Appalachian basin possible evidence for high fre quency eustatic sea level changes. *Sedimentary Geology* 154: 31 52.

Gradstein FM and Ogg J (1996) A Phanerozoic time scale. *Episodes* 19: 3 5.

Joachimski MM and Buggisch W (2002) Conodont apatite $\delta^{13}C$ signatures indicate climatic cooling as a trigger of the Late Devonian mass extinction. *Geology* 30: 711 714.

Long JA (1995) *The Rise of Fishes*. Baltimore: The Johns Hopkins University Press.

McGhee GR (1996) *The Late Devonian Mass Extinction*. New York: Columbia University Press.

McGhee GR (2001) The 'multiple impacts hypothesis' for mass extinction: a comparison of the Late Devonian and the late Eocene. *Palaeogeography, Palaeoclimatology, Palaeoecology* 176: 47 58.

Murphy AE, Sageman BB, and Hollander DJ (2000) Eu trophication by decoupling of the marine biogeochemical cycles of C, N, and P: a mechanism for the Late Devonian mass extinction. *Geology* 28: 427 430.

Racki G (1998) Frasnian Famennian biotic crisis: under valued tectonic control? *Palaeogeography, Palaeoclimat ology, Palaeoecology* 141: 177 198.

Sepkoski JJ (1996) Patterns of Phanerozoic extinction: a perspective from global data bases. In: Walliser OH (ed.) *Global Events and Event Stratigraphy*, pp. 35 51. Berlin: Springer.

Streel M, Caputo MV, Loboziak S, and Melo JHG (2000) Late Frasnian Famennian climates based on palyno morph analyses and the question of the Late Devonian glaciations. *Earth Science Reviews* 52: 121 173.

Walliser OH (1996) Global events in the Devonian and Carboniferous. In: Walliser OH (ed.) *Global Events and Event Stratigraphy*, pp. 226 250. Berlin: Springer.

Carboniferous

A C Scott, Royal Holloway, University of London, Egham, UK

Introduction

The Carboniferous, spanning 60 Ma, was an important period in Earth history, with major changes in atmospheric composition (in terms of both oxygen and carbon dioxide), the development of an icehouse world, the formation of vast tropical peat (coal) deposits, and the radiation of plants and animals on land. In particular, vertebrates got a major foothold on land and the air was conquered by the insects. The vast supercontinent of Pangaea was also formed during this period. The significance of the Carboniferous can be demonstrated by the fact that it was one of the first geological periods to be recognized. Despite this and the fact that Carboniferous rocks and fossils

have been amongst the most widely studied, very few formal definitions of the units of the Carboniferous have been internationally ratified.

Historical Setting

The Carboniferous was named by William Conybeare and William Phillips in 1822, based upon their work in northern England. From their studies, particularly in Yorkshire, they recognized a distinctive coal-bearing sequence, which had an important limestone sequence below. The frequent coals, which were economically important, are reflected in the name. The lower limestone unit became known as the Mountain Limestone, and the coal series became known as the Coal Measures.

This sequence of rocks was found across much of Western Europe, and the Carboniferous was divided into the Lower Carboniferous, comprising predominantly marine limestones, and the Upper Carboniferous, comprising clastic and coal-bearing strata. In continental Europe these two major divisions were termed the Dinantian and Silesian. In North America, however, whilst a similar sequence of limestones and coals was found, the divisions were named the Mississippian by Winchell and the Pennsylvanian by Williams. Unfortunately, the boundary between them did not coincide with the boundary defined in western Europe. To complicate matters even further, a three-fold division of the Carboniferous was recognized in Russia.

Chronostratigraphy and Biostratigraphy

Through the Heerlen congresses, and later the International Carboniferous Congress, held in Europe, North America, South America, and China through the twentieth century, a number of attempts to divide the Carboniferous into stages and to correlate local stages worldwide were made. Local stages were erected with formal boundary stratotypes, but there is often confusion as to usage in the literature. Currently, whilst the Carboniferous stage nomenclature has been agreed, few boundaries have yet been defined by the International Commission on Stratigraphy (ICS) (Figure 1).

Devonian–Carboniferous Boundary

The base of the Carboniferous System (also the base of the Tournaisian Stage) has now been formally defined as the first appearance of the conodont *Siphonodella sulcata* within the evolutionary lineage from *Siphonodella praesulcata* to *Siphonodella sulcata*.

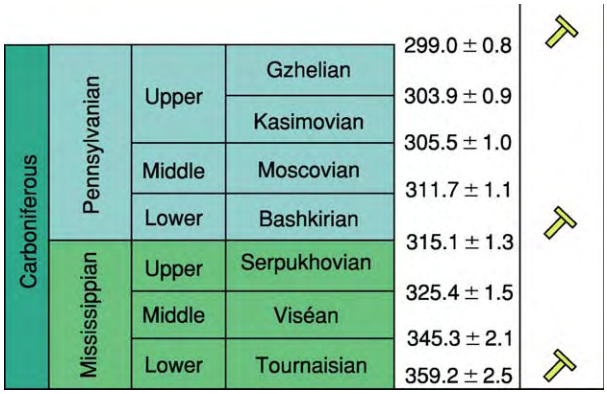


Figure 1 Current statigraphical nomenclature for the Carboniferous. Reproduced from the International Commission on Stratigraphy. Taken from ICS website.

The Global Standard Section and Point is at the base of Bed 89 in Trench E′ at La Serre in southern France. The rocks in the La Serre sequence are marine, belonging to the 'klippen of cabrières', comprising biodetrital oolitic limestones within a pelagic matrix of shale and cephalopod-bearing calcilutites. The fauna is varied and includes trilobites.

Carboniferous–Permian Boundary

The base of the Permian System (and hence the top of the Carboniferous System) is defined as the first occurrence of the 'isolated-nodular' morphotype of *Streptognathodus* '*wabaunsensis*' conodont chronocline, 27 m above the base of Bed 19, Aidarabsh Creek, northern Kazakhstan, which is in the southern Ural Mountains. The strata comprise hemipelagic silt and clay, with occasional lenses of sand and very course sand. The marine fauna, in addition to conodonts, includes ammonoids and fusilinid benthic foraminifera.

Mississippian–Pennsylvanian Boundary

Following considerable debate, the division of the Carboniferous System into two subsystems has been agreed. These two subsystems have been named the Mississippian and Pennsylvanian, although their usage may not always correspond with previous usage. The mid-Carboniferous boundary does not coincide with the boundary between either the Lower and Upper Carboniferous or the Dinantian and Silesian of previous schemes. In Europe the boundary is within the Namurian of older usage (Figure 2).

The base of the Pennsylvanian is also the base of the Bashkirian stage. The Global Standard Section and Point has been located at the lowest occurrence of the conodont *Declinognathodus nodiliferus* s.l., 82.9 m above the top of Battleship Wash Formation, Arrow Canyon, southern Nevada, USA.

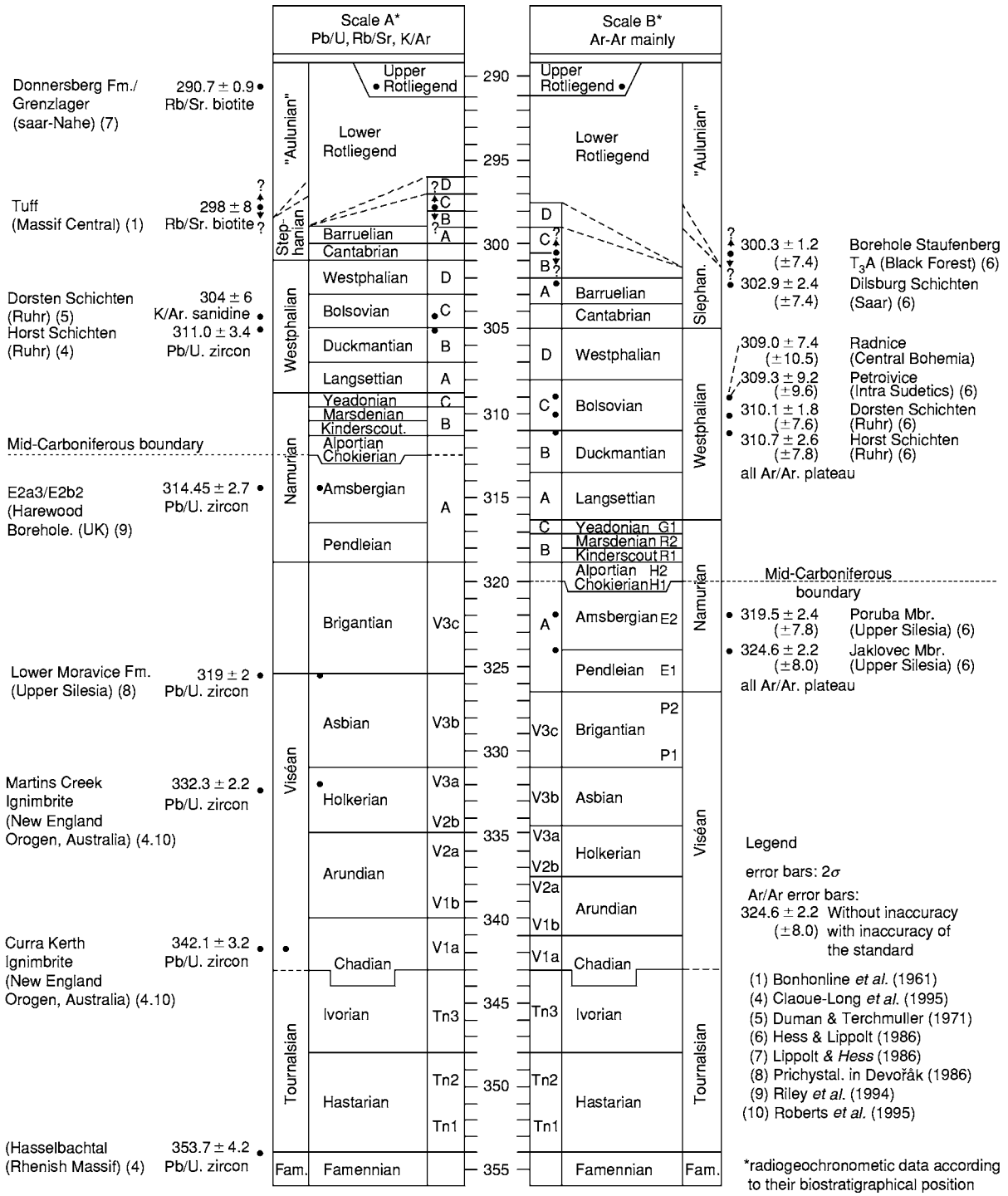


Figure 2 Absolute ages of Carboniferous strata obtained using different dating techniques and using a traditional Carboniferous chronostratigraphical nomenclature. Reproduced from Menning M, Weyer G, Drozdzewski G, Van Amerom HWJ, and Wendt I (2000) A Carboniferous time scale 2000: discussion and use of geological parameters as time indicators from central and western Europe. *Geologische Jahrbuch*, Hannover, A 156: 3 44.

Carboniferous Subdivisions

The Mississippian is now formally divided into three Stages (although these may become Series): Tournaisian, Viséan, and Serpukhovian. The Tournaisian–Viséan and Serpukhovian–Viséan boundaries have still to be formally placed. The Pennsylvanian is now formally divided into four: Bashkirian, Moscovian, Kasimovian, and Gzhelian. The Bashkirian–Moscovian,

Moscovian–Kasimovian, and Kasimovian–Gzhelian boundaries have yet to be decided. The division of the Carboniferous into formally internationally agreed series and stages may take a considerable time.

Absolute Ages

There is currently considerable debate over the absolute ages of Carboniferous sequences. Traditionally, most dates have been obtained from tonsteins (volcanic ash deposits) within coal seams. The ICS currently provide a number of dates. The base of the Carboniferous is stated by the ICS as 359 Ma and the top of the Carboniferous as 299 Ma, making the Carboniferous 60 Ma long. The mid-Carboniferous boundary is given as 318 Ma, making the Mississippian 41 Ma and the Pennsylvanian 19 Ma long. Menning, in his review of Carboniferous absolute time-scales, showed that there is a significant discrepancy between different radiometric-dating techniques (Figure 2). Traditional dates using uranium–lead dating techniques are in conflict with those obtained using argon–argon techniques. In scale A of Figure 2 the dates are from a series of techniques using uranium–lead, rubidium–strontium, and potassium–argon. A number of these involve the use of uranium–lead dates of zircons. Here the mid-Carboniferous boundary is placed at approximately 313.5 Ma. In contrast, scale B uses mainly argon–argon dates and the mid-Carboniferous boundary is placed at 320 Ma. The current trend is to place more reliance on the argon–argon dates, but it must be made clear that there are relatively few data points.

Biostratigraphy

Many groups of fossils have been used for biostratigraphical zonation and correlation. In the predominantly marine strata of the Mississippian, corals and brachiopods were widely used in Europe. Goniatites (cephalopods) have also been extensively used, particularly in North America. On land both vascular plants and non-marine bivalves have been widely used since the end of the nineteenth century. Advances in micropalaeontology in the mid-twentieth century brought the use of microfossils to the fore. Both foraminifera and conodonts became widely used in the later part of the twentieth century. Conodonts have, in marine strata, become the zonal fossils of choice, being used to define major chronostratigraphical boundaries. On land macrofossil plants and, more recently, pollen and spores (palynomorphs) have been widely used. Palynomorphs were first used to help correlate Pennsylvanian coal seams, and subsequent work on their distribution in clastic sediments has allowed them to be used to establish zones that can be correlated across continents. The use of pollen and spores for zonation and correlation was extended into the Mississippian, and they have the advantage not only of being found in non-marine strata but also of being washed or blown into marine rocks. This has allowed correlation between marine and non-marine sequences.

Lithologies and Environments

Within the Carboniferous equatorial belt, warm shallow-water carbonates dominate Mississippian sequences. This is what Derek Ager termed the 'persistence of facies.' Mississippian facies differ markedly in temperate and higher latitudes. Carboniferous sequences in South America, Australia, and South Africa are characterized by cold-water and often glacial deposits. In Europe, later Mississippian and Early Pennsylvanian deposits are commonly clastic shales and sandstones, which were lain down in thick deltaic sequences. By the later part of the Mississippian, commercial coals are found, such as in the Viséan of the Moscow Basin.

Widespread regression, possibly related to the onset of a major glaciation, allowed extensive mires to form throughout the tropics (Europe and North America) (Figure 3). Later Carboniferous deposits in Euramerica are characterized in many places by coal-bearing sequences. It was during the Pennsylvanian that many of the world's economic bituminous coals were formed. Some of these seams are several metres thick, taking tens of thousands of years to form. Many of the larger coal basins show evidence of regular marine influxes. The resulting marine bands, often containing goniatites, have been widely used for basinal and even intercontinental correlation. The regular sequences of non-marine rocks, in some cases with associated marine bands, have, on various scales, been called cyclothems and mesothems. The regular transgressive–regressive cycles, so well displayed in the mid-west of the USA, are thought to be the result of eustatic changes caused by the waxing and waning of the southern ice-sheet.

The later part of the Pennsylvanian is characterized by increasing aridity. Coal-forming mires shrank and peat-forming plants became extinct. In China, however, some of these environments and vegetation continued into the Permian. The drier environments supported very different vegetation from that which characterized the wet lowland habitats. Interbasin correlation, therefore, becomes difficult in some of these areas.

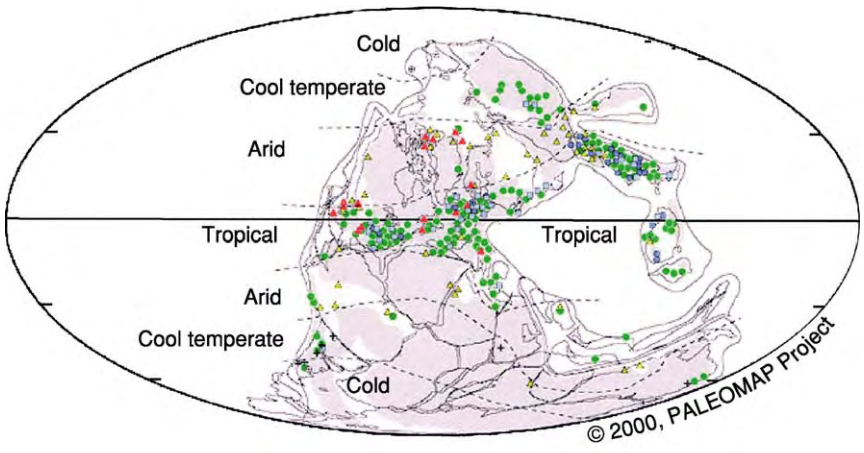


Figure 3 Lithological indicators of palaeoclimate in the Pennsylvanian. Reproduced from Scotese 2000 Paleomap project.

The Carboniferous World

Plate Reconstructions

The Carboniferous saw the gradual accretion of many continental landmasses. At the beginning of the Mississippian, two or three major continents had formed (Figure 4A). Europe and North America had collided to form Euramerica, which has also been termed Laurussia or Laurasia. Accretion was still active, with collision and uplift of the Appalachian mountain belt, which may have been a response to the northward migration of Africa. The Rheic Ocean, south of Britain, continued to close. The far-east Asian plate assemblies were more complex. Siberia was in the temperate northern hemisphere and is also known as Angara. China, comprising a northern block and a southern block, was also north of the equator, but there is dispute as to whether these two blocks had collided by the Carboniferous. This area is known as Cathaysia.

Many of the southern continents formed the supercontinent of Gondwana (*see* **Gondwanaland and Gondwana**). In the Carboniferous there continued to be continental accretion, for example in South America. Through the Carboniferous the northern and southern continents continued to amalgamate (Figure 4B) so that by the Permian the giant supercontinent of Pangaea (*see* **Pangaea**) had formed.

Palaeobiogeography

There have been several recent attempts to assess the palaeobiogeography of early Carboniferous vegetation using data from either macrofossils or pollen and spores. Taking the Mississippian as a whole, it is generally considered that the world's floras were predominantly cosmopolitan, characterized by assemblages including the lycopod genus *Lepidodendropsis* (Figure 5A), which was a widespread arborescent form that did not shed its leaves. The only regionally separate flora that has been recognized is the Angaran flora, which comprises mainly endemic genera. However, an additional temperate floral belt has recently been suggested. A larger number of regional floras have been recognized using mathematical techniques, but these analyses suffer from many problems, not least concerning the identity and age of the plants in question, some of which were first studied in the nineteenth century. These attempts of course have taken into consideration only compression plant genera, and there has been much recent work on anatomically preserved taxa, which can rarely be correlated with compression taxa. In addition, as might be expected, within any given region floras may vary because of ecological or preservational controls.

By the end of the Carboniferous, clear regional floras can be recognized (Figure 5B). The best-studied flora is that of Euramerica. This region comprised Europe (including eastern Europe) and North America and was, during this period, equatorial with a warm humid climate. The floras comprise diverse taxa typified by the species preserved in the Pennsylvanian Coal Measures, in particular the lycopods *Lepidodendron*, *Sigillaria*, and *Lepidophloios*, the sphenopsid *Calamites*, the ferns *Corynepteris*, *Renaultia*, and *Pecopteris*, the seed-ferns *Neuropteris*, *Alethopteris*, and *Mariopteris*, the cordiate *Cordaites*, and the conifer *Lebachia* (*Walchia*).

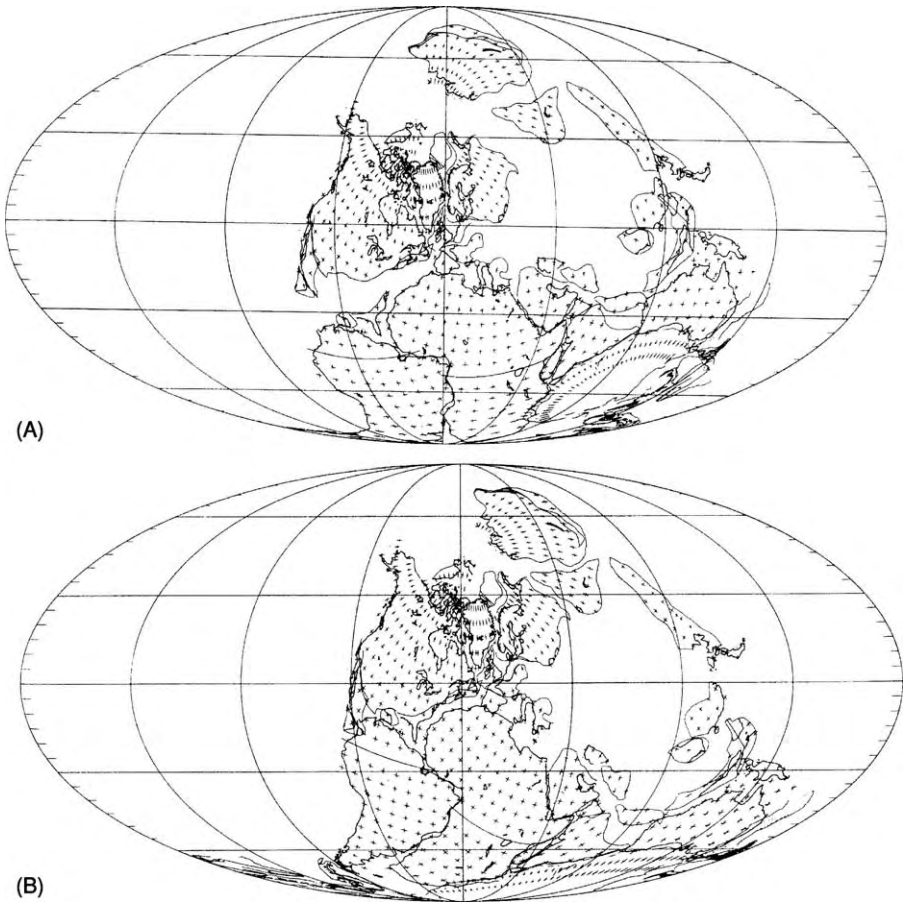


Figure 4 Plate reconstructions for (A) the Mississippian (Viséan) and (B) the Pennsylvanian (Moscovian). Reproduced from Scotese CR and McKerrow WS (1990) Revised world maps and introduction. In: McKerrow WS and Scotese CR (eds.) *Palaeozoic Palaeogeography and Biogeography*, pp. 1 21. Memoir 12. London: Geological Society.

The Gondwanan flora, characterized by the seed-fern *Glossopteris* and its relatives, is found during the Early Permian in Australia, Brazil, Africa, Antarctica, and India.

The Late Carboniferous flora from Australia is problematic. It comprises several genera with fern-like foliage, which have been linked to Early Carboniferous northern genera. There are two possible explanations for these occurrences: either they represent a relict population, which migrated southwards during the Carboniferous and survived in Australia after becoming extinct elsewhere, or, more likely, they belong to different plants with a similar leaf morphology. These plants include *Rhacopteris* and *Triphyllopteris*.

This regionalization of floras may partly reflect latitude – the *Glossopteris* flora is found in the high-latitude southern hemisphere, the Euramerican flora is found in equatorial areas, and the Angaran flora is found at high latitudes in the northern hemisphere – and may also reflect different climatic regimes (Figure 5C). This is supported by the fact that strata bearing the *Glossopteris* flora of Gondwana often overlie tillite horizons. The Cathaysian flora is also problematic, with the main regionalization in China occurring in the Permian.

The Carboniferous Atmosphere

The atmospheric composition changed markedly through the Carboniferous, and this period was one of the most unusual in Phanerozoic history (Figure 6).

Models of atmospheric oxygen for the Carboniferous by Berner and others show that the present atmospheric level of oxygen, 21%, was probably reached by the end of the Devonian. Oxygen levels are thought to have risen considerably during the Carboniferous, reaching a peak well above 30% by the Pennsylvanian. This rise in atmospheric oxygen may have been a result of the occurrence of

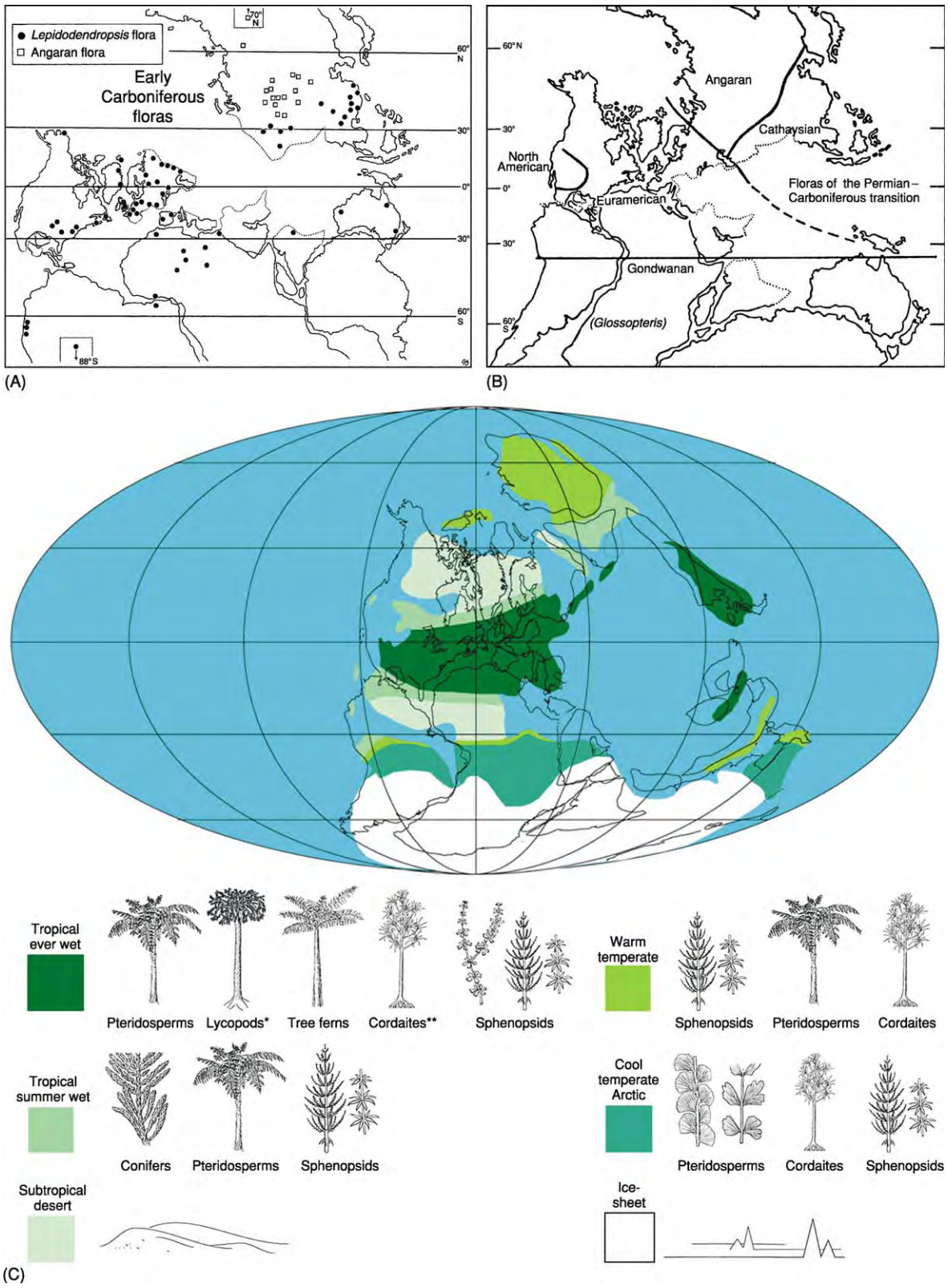


Figure 5 Palaeobiogeographical reconstructions of the Late Carboniferous showing (A) the floral provinces of the Mississippian (reproduced from Chaloner WG and Lacey WS (1973) *The Distribution of Late Palaeozoic Floras*. Special papers in Palaeontology 12. pp. 271 289), (B) the floral provinces of the Carboniferous Permian transition (adapted from Chaloner WG and Lacey WS (1973) *The Distribution of Late Palaeozoic Floras*. Special papers in Palaeontology 12. pp. 271 289.), and (C) Pennsylvanian biomes (reproduced from Willis KJ and McElwain JC (2002) *The Evolution of Plants*. Oxford: Oxford University Press).

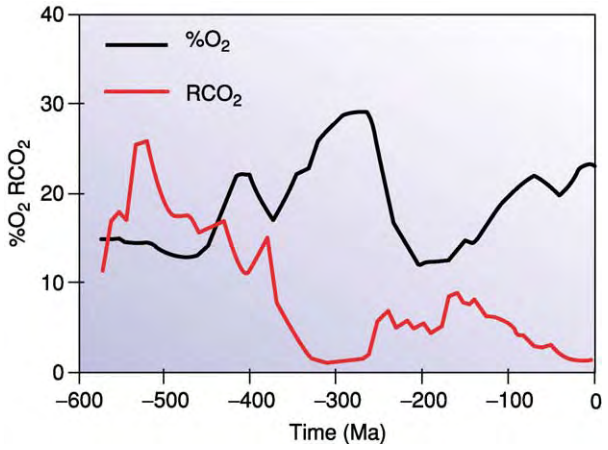


Figure 6 Changes in atmospheric composition over the Late Proterozoic and Phanerozoic, showing falling carbon dioxide and rising oxygen levels during the Carboniferous (360 300 Ma). Reproduced from Berner RA (2003) The long term carbon cycle, fossil fuels and atmospheric composition. *Nature* 426: 323 326.

widespread lowland tropical rainforest. The later postulated rapid fall in oxygen may have been a result of the demise of these extensive tropical coal-forming forests. The oxygen-rich atmosphere may have had a significant effect on animal evolution, allowing the growth of giant arthropods including insects. In addition, fires may have been more widespread and intense.

In contrast, levels of carbon dioxide fell rapidly at the end of the Devonian and during the Early Carboniferous. This coincided both with the evolution and initial spread of trees and with the onset of extensive peat formation and, hence, burial of carbon, in addition to the burial of extensive shallow-water carbonates. All of these factors combined to cause a major drawdown of carbon dioxide. The major climatic changes in the Late Carboniferous–Early Permian may be linked to increasing levels of carbon dioxide. The impact of these low levels of carbon dioxide on the Carboniferous world was severe, plunging it into a major icehouse phase, as evidenced by the spread of the polar ice-cap in the southern continents.

The Carboniferous Climate: Icehouse Conditions

One of the major glaciations in Phanerozoic history occurred in the Late Palaeozoic, beginning in the Carboniferous and ending in the Permian (Figure 7). The world, therefore, went from greenhouse conditions to icehouse conditions and back to greenhouse conditions. Striated pavements and glacial tillites both indicate the presence of a southern polar ice-cap across the Gondwanan continents (Figure 8). All other aspects of this glacial episode, however, are open to at least some debate.

- Onset. The onset of glacial conditions has been difficult to date. Palynological dating of South African sequences suggests that glaciation may have begun as early as the Devonian–Carboniferous boundary. Unequivocal evidence from tillites and inferences from carbon isotope excursions indicate that a significant ice-cap existed by Late Mississippian times. It is believed that the earliest glacial deposits are those of South America.
- Spread and extent. It has been shown that the onset of glaciation spread to South Africa and Antarctica and thence to Australia and India. Typically glacial tillites (e.g. Dwyka of South Africa) overlie striated pavements and are overlain by temperate coal-bearing strata. Whilst on palaeogeographical maps a large polar ice-cap is often shown across much of southern Gondwana, it is not clear whether this was the case or whether a smaller ice-sheet moved as the continents moved.
- Demise. Just as the bases of the glacial deposits across Gondwana are diachronous, so too are the overlying non-glacial deposits. Remnant ice-caps may have been present in upland areas, but Permian coal-bearing strata are found in South America, South Africa, Antartica, India, and Australia. The cessation of glaciation also appears to be diachronous, being older in South America and younger in Australia.
- Causes. Whilst a number of key facts about the Gondwanan glaciation and the Carboniferous world have been established, there is little agreement about cause and effect.

Climate change can be seen over several different time-scales, all of which have been well documented for the Carboniferous. The longest cycle, causing a change from greenhouse to icehouse cycles, may have been caused by the combination of plate movements, atmospheric change, orbital cycles, and the evolution of land plants. These factors continued to act throughout the Carboniferous. There are changes in both local and global climate through the Carboniferous. The occurrence of evaporates in some Mississippian deposits indicates a drier environment than that seen in the Pennsylvanian. Widespread wetland systems occurred in tropical regions during the Pennsylvanian (Figure 9).

Orbital-eccentricity cycles and axial-tilt and precession cycles have intermediate- and short-term climatic effects. However, these effects may be modified by local tectonic or orographic conditions. Studies of coals and vegetation show major wet–dry cycles

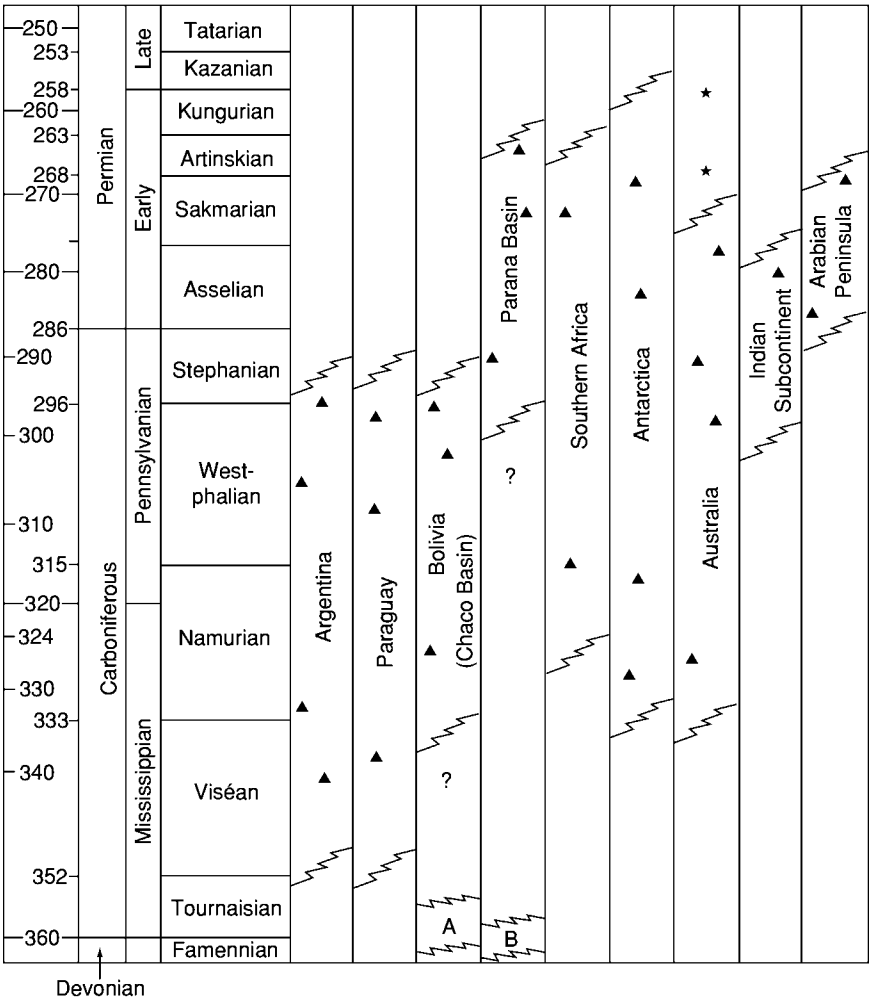


Figure 7 Timing of glaciation in the Late Palaeozoic across Gondwana. Reproduced from Eyles N (1993) Earth's glacial record and its tectonic setting. *Earth Science Reviews* 35: 1 248.

during the Pennsylvanian of Illinois. These changes in rainfall had a dramatic effect on vegetation and on peat (coal) formation. In particular, a major drying phase at the Westphalian–Stephanian boundary caused the extinction of a number of peat-forming lycophytes (Figure 9). Increasing aridity towards the end of the Carboniferous is also widely documented. The broad wet–dry cycles seen in Illinois are not always seen in other Euramerican basins, although the broad picture is confirmed (Figure 10).

Milankovitch cycles have been proposed to have caused waxing and waning of the ice-sheet over the southern continents, leading to rises and falls in sea-level (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). It has been suggested that these sea-level changes may have been responsible for the cyclothemic and mesothemic units that are seen particularly well in the Pennsylvanian.

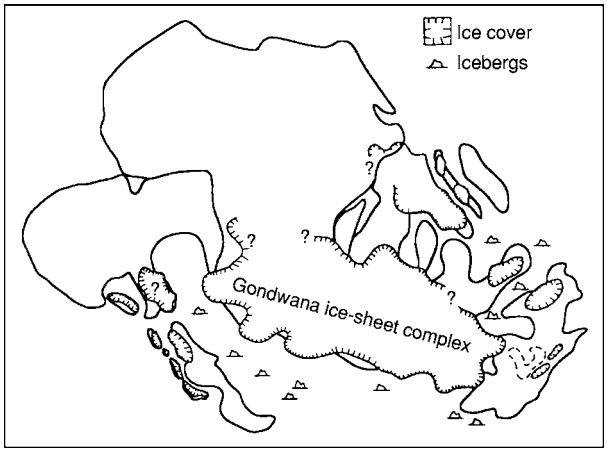


Figure 8 Maximum glaciation of Gondwana in the Carboniferous and Permian. Reproduced from Eyles N (1993) Earth's glacial record and its tectonic setting. *Earth Science Reviews* 35: 1 248.

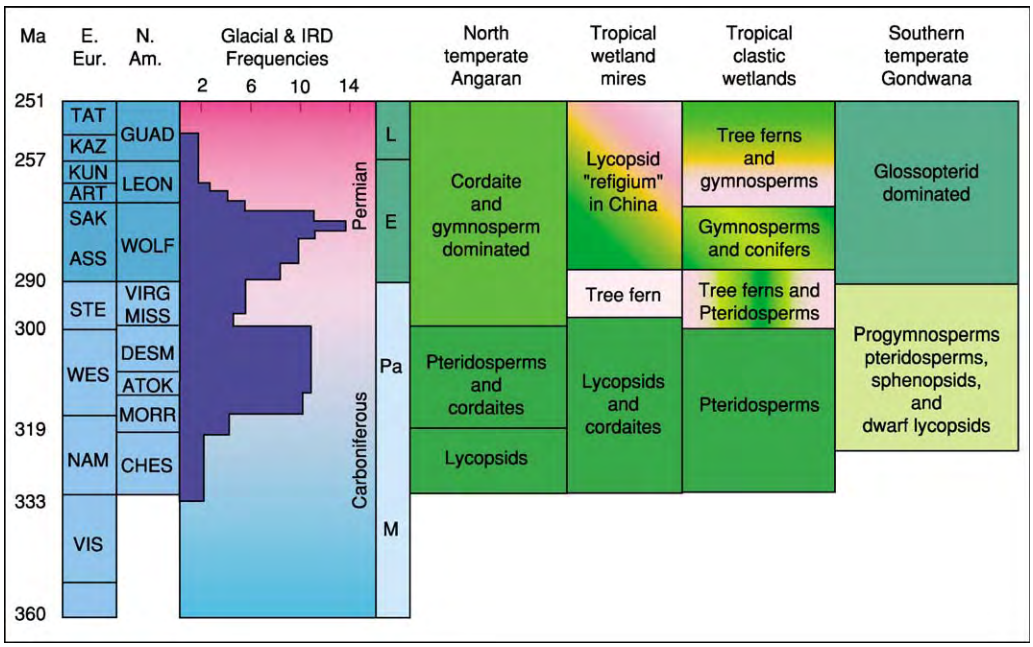


Figure 9 Vegetational change through the Late Palaeozoic in relation to climate change. IRD Ice rafted deposits; M Mississippian; P Pennsylvanian; E Early; L Late; VIS Visean; NAM Namurian; WES Westphalian; STE Stephanian; ASS Asselian; SAK Sakmarian; ART Artinskian; KUN Kungurian; KAZ Kazanian; TAT Tatarian; CHES Chesterian; MORR Morrowan; ATOK Atokan; DESM Desmoinian; MISS Missourian; VIRG Virgillian; WOLF Wolfcampian; LEON Leonardian; GUAD Gudalupian. Reproduced from Gastaldo RA, DiMichele WA, and Pefferkorn HW (1996) Out of the icehouse into the greenhouse: a Late Paleozoic analogue for modern global vegetational change. *GSA Today* 10: 1 7.

The Nature and Influence of Fire in Carboniferous Ecosystems

Fires may range from commonly seasonal to rare and catastrophic. Climate may play a significant role in controlling the frequency of fires. It is possible that, in addition to large-scale climatic changes, smaller-scale changes in climate may play a significant role in the occurrence of fires. It is possible that in some ever-wet environments fires are unusual. Disturbance of the normal climatic regime (e.g. El Niño events) may cause temporary changes in the rainfall pattern, for example, leading to a dry period when a catastrophic fire may occur. This may be the reason for the Great Fire of Borneo in 1982–1983. Small fluctuations in rainfall pattern may allow frequent small fires in a mire-forming environment. Storms may follow a major fire, and the increased erosion may give rise to the transportation and deposition of a sediment–charcoal mix at several sites. The occurrence of fusain (fossil charcoal) layers in the fossil record may be used to distinguish between events of different timescales.

The effects of large-scale fires are illustrated by the Lower Carboniferous marginal-marine fusain deposits at Shalwy Point, Donegal, Ireland. Here, there is no doubt that the fire was catastrophic. The widespread nature of the fire not only yielded significant amounts of charcoal but also would have led directly to increased soil erosion, which in turn would have led to a major increase in bed load in the alluvial system.

Life on Land

There were many major changes in the terrestrial biota during the Carboniferous.

Plants

The Early Carboniferous saw a major radiation in many groups of vascular land plants. Amongst spore-bearing plants, lycopsids underwent a major radiation into several new ecological niches. Extensive peat mires were developed for the first time. At this time sphenopsids also became abundant, with forms such as *Archaeocalamites* being particularly widespread. There was a dramatic diversification of ferns during the Mississippian, with zygopterid ferns being particularly abundant. It has been proposed that these plants were the primary colonizers of volcanic terrains throughout Europe.

The most important plant radiation at this time was in the seed plants. In particular, pteridosperms

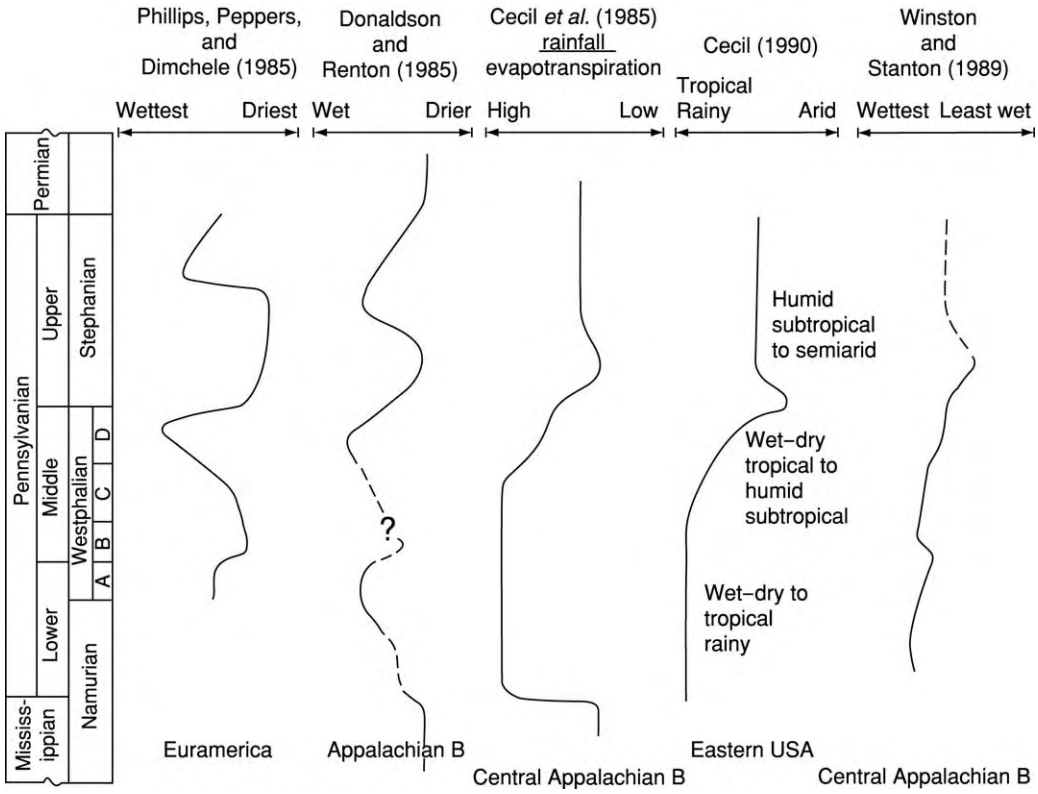


Figure 10 Interpreted palaeoclimatic trends for the Euramerican coal basins during the Pennsylvanian. Reproduced from Calder JH and Gibling MR (1994) The Euramerican Coal Province: controls on Late Paleozoic peat accumulation. *Palaeogeography, Palaeoclimatology, Palaeoecology* 106: 1 21.

underwent a rapid radiation into a variety of lowland habitats. The abundance of anatomically preserved specimens in the Mississippian of Scotland has allowed detailed phylogenetic studies of these plants.

Several plant groups diversified during the Pennsylvanian. The lycopods are the main coal-forming plants of the Carboniferous. They are also known as club mosses but are, of course, true vascular plants. Today only a few genera of herbaceous forms survive, but in the Carboniferous they showed a vast array of life habits from small herbaceous forms to large arboresent (tree) forms. Much of our recent data on these important plants has come from the study of coal balls – anatomically preserved fossils where the internal anatomy can be elucidated. As with the lycopods, Carboniferous sphenopsids are represented by both arborescent and herbaceous forms.

Ferns are plants with well-developed megaphyllous leaves. All the Carboniferous forms are homosporous. Fern-like foliage is abundant in the Pennsylvanian Coal Measures but most belongs to the pteridosperms (seed-ferns). The identification of true ferns in the compression fossil record requires fertile specimens. The Carboniferous was once known as the age of ferns because of the abundance of fern-like foliage amongst the compression fossils. It was shown, however, that most of this foliage belongs to an extinct group of plants that bore seeds and pollen organs, known as the pteridosperms or seed-ferns. Three major families are known from the Upper Carboniferous: Lyginopteridales, Medullosales, and Callistophytales. These were either shrubs or climbers in the vegetation.

The most advanced plants of the late Carboniferous were the closely related cordaites and conifers. It is still unclear whether the cordaites gave rise to the conifers or whether they had a common ancestor. Both, however, are gymnospermous and are represented by both trees and shrubs. The conifers were probably major elements in drier upland floras.

Terrestrial Invertebrates

The East Kirkton terrestrial Mississippian Lagerstätten has opened a major window on a terrestrial ecosystem. Several invertebrate groups are first found here or there is evidence of their terrestrial nature. The fauna includes eurypterids, scorpions, myriapods, and a spider, in addition to aquatic arthropods such as ostracods. The eurypterids were large and

include *Hibbertopterus*. The scorpions show unequivocal evidence of terrestrial adaptations and specimens over 30 cm long have been found; they may have reached up to 1 m in length and been a top predator. Other arthropods include millipedes and an opilionid or Harvestman spider.

No fossil insects are known from East Kirkton; the first unequivocal evidence of true insects (Hexapoda: Pterygota) does not occur until the Early Pennsylvanian, although there are recent claims of early Devonian insects. Insects of the Pennsylvanian had wings.

The terrestrial fauna of the Pennsylvanian Coal Measure forests included many carnivorous animals. Centipedes (Myriapoda: Chilopoda) were common, especially in the Mazon Creek fauna of Illinois, as were many spiders (Arachnida: Aranae), some of which had inward-striking (dianial) fangs, indicating an ability to kill insects and other groups of spiders that possessed downward-striking (paraxial) jaws for attacking prey on a firm substrate, either on the ground or on tree trunks.

The first insects, which make an appearance in the Namurian, were members of the extinct order Palaeodictyoptera. Together with five other orders they comprise the Palaeoptera, and all have their origins in the Pennsylvanian. Two orders, the Ephemeroptera (mayflies) and Odonata (dragonflies and damselflies) are extant.

The other Carboniferous orders are, like the Palaeodictyoptera, extinct, surviving only until the end of the Permian. The Protodonata were very similar to the first dragonflies except for a variation in wing venation and their great size. The Megasecoptera and Diaphanopterodea were very similar to the Palaeodictyoptera.

Five orders of exopterygote neopteran insects also appear in the Pennsylvanian. Three of these orders became extinct at the end of the Palaeozoic. These were the Protorthoptera, which was the most diverse of all the extinct orders, the Caloneurodea, which some workers place in the Endopterygota, and the Miomoptera, which were very small insects, perhaps related to the Protorthoptera. The other two orders, the Blattodea (cockroaches) and Orthoptera (grasshoppers, katydids, and crickets), are extant. The Blattodea were apparently the commonest group of insects during the Pennsylvanian. There is increasing evidence of plant–insect interactions in the Carboniferous.

Vertebrates

The Carboniferous saw the diversification and spread of vertebrates onto land. Earliest Mississippian deposits contain relatively few tetrapod fossils. Some late Tournasian deposits in Scotland and Canada have yielded a few isolated fragments. The later Viséan rocks, mainly of Scotland, have yielded many important new skeletons. A single small specimen of tetrapod, named *Casineria kiddi*, has been described from mid-Viséan deposits in eastern Scotland. The anatomical features of the specimen indicate that it was fully terrestrial. The most spectacular finds in recent years have come from the Late Viséan deposits of East Kirkton. Amongst a wide range of taxa, all less than 1 m long, are the temnospondyls. Other forms include early anthrocosaurs. This group is better known from Pennsylvanian strata. *Westlothiana lizziae* (Lizzie the lizard) is a species of particular note as it shows many of the characteristics of reptiles.

Several tetrapod groups show evidence of major diversification in the late Mississippian and early Pennsylvanian (*see* **Fossil Vertebrates:** Palaeozoic Non-Amniote Tetrapods). Most of the important tetrapod sites are found in tropical Euramerica and include the North American sites of Linton (Ohio), Mazon Creek (Illinois), and Joggins (Nova Scotia), several British and Irish sites, including Jarrow and Newsham, and continental European sites, such as Nyrany (Czech Republic).

The temnospondyls underwent a major radiation in the early Pennsylvanian and were the largest group of 'amphibians.' The most important of these was the small *Dendrerpeton*, which was first found in the tree stumps at Joggins. Amongst the temnospondyls were several aquatic forms. The anthracosaurs were another important group of tetrapods, also with a Mississippian origin. They included the crocodile-like embolomeres such as *Eoherpeton*. The most fearsome predator in this group was *Anthracosaurus*, which is known only from its skull.

Microsaurs also have a Mississippian origin; they include several important genera and have been confused with true amniotes. Other related groups include the lysorophids and aiistopods. Amniotes (including the reptiles for some authors) are first found in the Mississippian but radiated in the Pennsylvanian. Whilst several reptiles are known from the Pennsylvanian, it is clear that terrestrial tetrapod faunas were dominated by amphibians. It is significant to note that, with the possible exception of *Scincosaurus*, all the major Upper Carboniferous tetrapods were probably carnivores or insectivores.

The first true amniote (as opposed to ancestral group) was also found at Joggins. *Hylonomus* was amongst the animals found within the Joggins tree stumps. Other forms include *Captorhinus*, but it was not until the latest Carboniferous that herbivory evolved in larger forms, such as *Edaphosaurus*. In the Early Permian of Gondwana there is evidence of the

first aquatic amniote, *Mesosaurus*, which is found on several continents. It is in the Permian that the first major radiation of amniotes occurred.

Life in the Sea

Much of our knowledge about Carboniferous marine life comes from the extensive tropical shallow-water carbonates of the Mississippian. These warm shallow seas supported a rich benthos dominated by filter-feeding brachiopods and crinoids. Plankton diversity appears, however, to have been low, with the fauna dominated by only a few groups of acritarchs (marine aquatic algae). In some areas reefs dominated by rugose and tabulate corals together with sponges, bryozoans, and calcareous algae were important. Fusiline foraminifera became important for the first time. Nektonic animals included goniatites and such vertebrates as conodonts, which are important for biostratigraphical correlation. Undisputed conodont animals were discovered in the Mississippian rocks of Eastern Scotland associated with various aquatic arthropods. Fishes are also found in many marine and non-marine environments. A variety of sharks are known from Carboniferous deposits, and some of the best examples (such as *Stethacanthus* from Bearsden in Scotland) are from brackish marine deposits. Ray-finned and lobe-finned fishes also diversified in the oceans at this time.

The structure of Pennsylvanian marine communities was similar to that in the Mississippian. However, Pennsylvanian marine strata are not as widespread. In Europe the spread of marine facies across low-lying mires is well documented. The marine rocks are predominantly muddy and yield goniatites, bivalves, and conodonts, in contrast to the limestones that yield corals and brachiopods in addition to the goniatites and conodonts known from North America and Ukraine. Fusiline foraminifera became increasingly abundant through the Pennsylvanian in these carbonate environments.

The End of the Carboniferous and Mass Extinctions

Major changes took place at the end of the Carboniferous. Plate movements resulted in the continued collision of Laurasia and Gondwana to form the supercontinent Pangaea. The formation of Pangaea led to major climatic change, with tropical drying and the spread of continental red beds. The decrease in rainfall prevented peat formation in more tropical areas and was accompanied by the apparent extinctions of many terrestrial plant and animal groups. The floras of dryer upland habitats began to spread towards the lowlands. These floras were dominated by the newly evolved conifers. In the southern latitudes, the floras that had been typical of the Carboniferous, composed of progymnosperms, pteridosperms, spenopsids, and dwarf lycopsids, gave way to the glossopterid-dominated floras of the Permian. Recent research in China has shown that many of the extinctions in Euramerica at the end of the Carboniferous were local and not global. Several Carboniferous plant and animal groups are found in the Permian of China and become extinct only at the end of the Permian.

In the oceans there was continued faunal turnover, with some groups becoming extinct and others, such as the fusiline foraminifera, diversifying and becoming more important and widespread – they are particularly useful for defining the base of the Permian and for correlating subsequent Permian stages. There is no evidence, however, of a mass extinction at the end of the Carboniferous.

See Also

Analytical Methods: Geochronological Techniques. **Atmosphere Evolution**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Fossil Vertebrates:** Palaeozoic Non-Amniote Tetrapods. **Gondwanaland and Gondwana**. **Palaeoclimates**. **Pangaea**. **Sedimentary Processes:** Glaciers.

Further Reading

Ager DV (1984) *The Nature of the Stratigraphical Record,* 3rd edn. Basingstoke: MacMillan Publishers Ltd.

Calder JH and Gibling MR (1994) The Euramerican Coal Province: controls on Late Paleozoic peat accumula tion. *Palaeogeography, Palaeoclimatology, Palaeoecol ogy* 106: 1 21.

Clack JA (2002) *Gaining Ground: The Origin and Evolu tion of Tetrapods*. Bloomington: Indiana University Press.

Cleal CJ (1991) Carboniferous and Permian biostratigra phy. In: Cleal CJ (ed.) *Plant Fossils in Geological Investi gation: The Palaeozoic,* pp. 182 315. Chichester: Ellis Horwood.

Cleal CJ and Thomas BA (1991) Carboniferous and Per mian palaeogeography. In: Cleal CJ (ed.) *Plant Fossils in Geological Investigation: The Palaeozoic,* pp. 154 183. Chichester: Ellis Horwood.

Davydov VI, Glenister BF, Spinosa C, *et al.* (1998) Proposal of Aidaralash as Global Stratotype Section and Point (GSSP) for base of the Permian System. *Episodes* 21: 11 18.

DiMichele WA, Pfefferkorn HW, and Phillips TL (1996) Persistence of Late Carboniferous tropical vegetation during glacially driven climatic and sea level fluctuations. *Palaeogeography, Palaeoclimatology, Palaeoecology* 125: 105 128.

Gastaldo RA and DiMichele WA (eds.) (2000) *Phanerozoic Terrestrial Ecosystems*. Special Papers 6. Paleontological Society.

Gastaldo RA, DiMichele WA, and Pefferkorn HW (1996) Out of the icehouse into the greenhouse: a late Paleozoic analogue for modern global vegetational change. *GSA Today* 10: 1 7.

Hower J, Scott AC, Hutton AC, and Parekh BK (1995) Coal availability, mining and preparation. In: Bisio A and Boots SG (eds.) *Encyclopaedia of Energy Technology and the Environment,* pp. 603 684. New York: J Wiley & Sons.

Iannuzzi R and Pfefferkorn HW (2002) A pre glacial, warm temperate floral belt in Gondwana (Late Viséan, Early Carboniferous). *Palaios* 17: 571 590.

Labandeira CC (1998) Early history of arthropod and vas cular plant associations. *Annual Review of Earth and Planetary Sciences* 26: 329 377.

Lane HR, Brenckle PL, Baesemann JF, and Richards B (1999) The IUGS boundary in the middle of the Carboniferous: Arrow Canyon, Nevada, USA. *Episodes* 22: 272 283.

Menning M, Weyer G, Drozdzewski G, Van Amerom HWJ, and Wendt I (2000) A Carboniferous time scale 2000: discussion and use of geological parameters as time indi cators from central and Western Europe. *Geologische Jahrbuch,* Hannover, A 156: 3 44.

Paproth E, Feist R, and Flajs G (1991) Decision of the Devonian Carboniferous boundary stratotype. *Episodes* 14: 331 336.

Phillips TL and Peppers RA (1984) Changing patterns of Pennsylvanian coal swamp vegetation and implications of climatic control on coal occurrence. *International Journal of Coal Geology* 3: 205 255.

Rolfe WDI, Durant GP, Fallick AE, *et al.* (1990) An early terrestrial biota preserved by Viséan vulcanicity in Scotland. In: Lockley MG and Rice A (eds.) *Volcanism and Fossil Biotas,* pp. 13 24. Special Publication 244. Boulder: Geological Society of America.

Scheffler K, Hoernes S, and Schwark L (2003) Global changes during Carboniferous Permian glaciation of Gondwana: linking polar and equatorial climate evolution by geochemical proxies. *Geology* 31: 605 608.

Scott AC (1980) The ecology of some Upper Palaeozoic floras. In: Panchen A (ed.) *The Terrestrial Environment and Origin of Land Vertebrates,* pp. 87 115. Systematic Association Symposium Volume 15. London and New York: Academic Press.

Scott AC (2000) The Pre Quaternary history of fire. *Palaeogeography, Palaeoclimatology, Palaeoecology* 164: 281 329.

Scott AC and Galtier J (1996) A review of the problems in the stratigraphical, palaeoecological and palaeo biogeographical interpretation of Lower Carboniferous (Dinantian) floras from Western Europe. *Review of Palaeobotany and Palynology* 90: 141 153.

Scott AC, Stephenson J, and Chaloner WG (1992) Inter action and co evolution of plants and arthropods in the Palaeozoic and Mesozoic. *Philosophical Transactions of the Royal Society of London B* 335: 129 165.

Scott AC, Mattey D, and Howard R (1996) New data on the formation of Carboniferous coal balls. *Review of Palaeobotany and Palynology* 93: 317 331.

Wagner RH and Winkler PCF (1991) Major subdivisions of the Carboniferous system. In: *Compte Rendu XI^{e} Con grès International de Stratigraphie et de Géologie du Carbonifère, Beijing 1987,* pp. 213 245.

Wagner RH and Winkler PCF (1997) Carboniferous chronostratigraphy: quo vadis? *Proceedings of the XIII International Congress on the Carboniferous and Per mian*. In: *Prace Panstwowego Instytutu Geologicznego CLVII*, pp. 187 196.

Wnuk C (1996) The development of floristic provinciality during the Middle and Late Paleozoic. *Review of Palaeo botany and Palynology* 90: 5 40.

Permian

P B Wignall, University of Leeds, Leeds, UK

Introduction

The Permian Period, spanning the 43 Ma interval between 294 Ma and 251 Ma, witnessed several major turning points in Earth's history. Pangaea, the greatest supercontinent of all time, finally amalgamated in the Permian, and began to break up again almost immediately in a prolonged disintegration that would finally finish in the Cretaceous. Climate also changed markedly: the extensive southern-hemisphere glaciation that had begun in the Carboniferous concluded in the Middle Permian, to be replaced by a warming trend that was to peak in the earliest Triassic with some of the warmest conditions the world has ever experienced. The Permian was also a time of extremes in the fortunes of plants and animals. The Early to Middle Permian was generally a successful time for life, with low extinction rates and the evolution of the first large land animals and of plant communities that invaded the continental interiors. In contrast, the end of the Middle Permian was marked by an extinction crisis that primarily affected the inhabitants of shallow tropical seas. This was the first such crisis since the end of the Devonian, 100 Ma previously, and it brought to an end one of the most benign intervals of Earth history. Much worse was to follow with the end-Permian mass extinction, by far the biggest disaster in the history of life.

Subdivisions of Permian Time

Dividing the Permian into smaller time units has always proved surprisingly controversial. Whereas the subdivision of other intervals, such as the Jurassic and Cretaceous, has been stable and established for many decades, the current chronostratigraphical scheme for the Permian (Table 1) is only a few years old, and numerous other schemes can be found in the older literature. The problem stems principally from the lack of useful fossils available to correlate Permian strata in different regions of the world. In contrast, Ordovician stratigraphers can easily subdivide time using a succession of extremely widespread graptolite fossils, and Jurassic stratigraphers similarly use ammonites for subdivision and global correlation. For a long time Permian stratigraphers used ammonoids to define biozones, but these tend to be restricted to specific regions and in many areas are absent altogether. For example, there are extensive exposures of Permian rocks in western and Arctic Canada, but ammonoids are extremely rare in these deposits. In particular, since there are no ammonoids diagnostic of Late Permian time in this region, it was thought that Late Permian rocks were not present there. It now appears more likely that such rocks are present but the diagnostic fossils are not – a crucial distinction. Much Permian correlation is now achieved using conodonts (*see* **Microfossils:** Conodonts), a group of microfossils that are widespread in marine sediments of this age and therefore more useful than the ammonoids in establishing global time units.

Despite recent advances, there are still problems with correlating Permian rocks, primarily because different intervals of time are defined in different regions. Thus, the Early Permian (Cisuralian) is based on sections from Russia, the country where the period was originally defined; the Middle Permian (Guadalupian) is based on US sections, mostly in Texas; and the Upper Permian (Lopingian) is based on excellent marine sections in South China. Two Russian names, the Dzhulfian and the Dorashamian, were commonly used for the Late Permian interval, but these proved very difficult to identify outside the region, once again because of problems with the localized occurrences of ammonoids. As a result, the Dorashamian Stage – identified using ammonoids – was thought to be absent in most regions, but the essentially equivalent Changxingian Stage – identified using conodonts – is known to be widespread.

Tectonics

For nearly 200 Ma, from the Ordovician to the Permian, the Earth's continents gradually coalesced until eventually most continental crust had been amalgamated to form the supercontinent of Pangaea (Figure 1) (*see* **Pangaea**). The final stage of closure occurred during the Permian with collision of the Siberian continent with the Laurussian portion of northern Pangaea, the Eurasian Orogeny. This produced the Urals, a narrow and very long mountain belt that extends from the Volga region of southern Russia to Novaya Zemlya in the Arctic Ocean. During the Permian this newly forming mountain range shed vast quantities of sediment to the west of the advancing deformation front, into a deep-water basin in the eastern Barents Sea area, to the north of Norway, and into a series of terrestrial molasse

basins: the Volga-Urals and Timan-Pechora basins. To the east of the Urals the vast West Siberian Basin was also receiving large quantities of sediment and for much of the Late Permian sediment influx and subsidence kept pace so that the region remained around sea-level. In the latest Permian this area was the site of eruption of one of the largest volcanic provinces ever to form: the Siberian Traps. Subsidence rates continued to be high and the region remained around sea-level despite the eruption of several kilometres thickness of lava.

By the Middle Permian the areal extent of Pangaea was at its zenith, and it stretched in a giant arcuate shape from the north pole, which lay in eastern Siberia, to the south pole, which then, as now, lay in Antarctica (Figure 1). Pangaea was narrow at its waist owing to the presence of Tethys, an equatorially centred ocean that separated the northern and southern arms of the supercontinent. The eastern opening of Tethys was partially closed by the presence of some of the few continental masses that were not incorporated in Pangaea, notably South China and several smaller continents that are now part of southeast Asia. On the other side of the world from Tethys a vast ocean, called Panthalassa, stretched across 220° of longitude, making it 20% larger than the modern Pacific. Viewed from space in the Permian one side of the world would have looked entirely blue!

The break up of Pangaea began even as the Siberia–Laurussia collision was nearing completion, with the rifting of several small continents from the southern margin of Tethys, known as the Perigondwanan margin, in the Middle Permian (Figure 1). As these drifted northwards they opened up a narrow ocean behind them, known as Neotethys; the older Tethys to their north is often known as Palaeotethys. These slivers of continents included the present-day regions of Antalaya (Turkey), Armenia, parts of Iran and Afghanistan, and central Tibet; they are often depicted as forming an elongate narrow continent called Cimmeria. However, it is uncertain whether these ever formed a continuous continent. The Tibetan Block in particular seems to have begun its drift somewhat later than the other continents and to have lagged behind them during their northern passage. The Middle Permian rifting activity was accompanied by some voluminous volcanic episodes, notably that which formed the Panjal volcanics of Kashmir, India. This formerly extensive province of basaltic lavas was subsequently much deformed and metamorphosed during the formation of the Himalayas. However, the biggest Permian volcanic province is in the Emeishan region of south-western China and seems to have occurred during a poorly understood rifting episode around the Middle Permian–Late Permian boundary. Extensive sheets of basalt dominate the Emeishan

Table 1 Subdivisions of Permian time (names in brackets are those formerly used to describe Late Permian intervals)

Series	*Stages*
Lopingian base 262 Ma, top, 251 Ma (Tatarian)	Changxingian (Dorashamian) Wuchiapingian (Dzhulfian)
Guadalupian base 272 Ma	Capitanian Wordian Roadian
Cisuralian base 294 Ma	Kungurian Artinskian Sakmarian Asselian

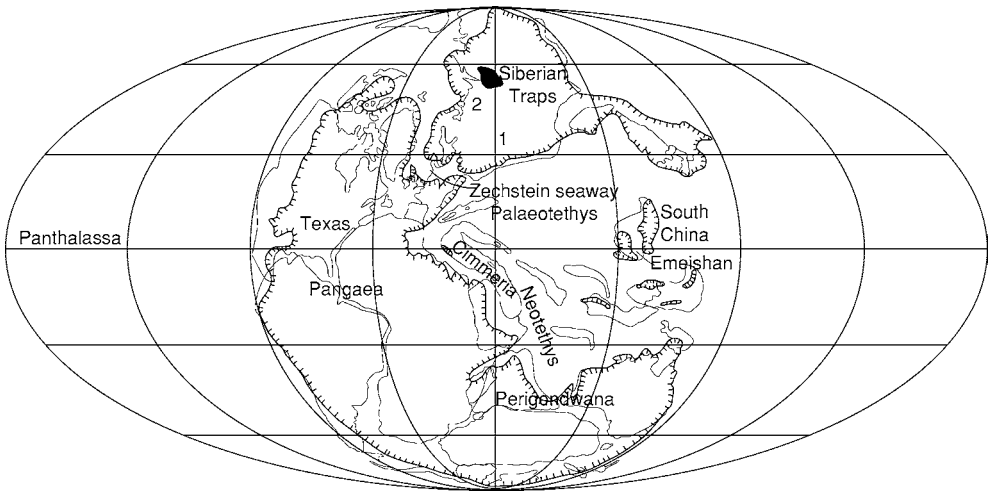


Figure 1 Global palaeogeography of the Late Permian showing the supercontinent of Pangaea. 1, Volga Urals Basin; 2, Timan Pechora Basin.

volcanic pile, and they may exceed a million cubic kilometres in volume. This basalt was extruded quietly from long fissures, but, during the later stages of the volcanism, these eruptions were replaced by explosive eruptions that produced more acidic lavas.

One consequence of supercontinent formation is that the area of fringing shallow shelf seas becomes minimal. This was indeed the case in the Permian, when the area of shallow-marine deposition was probably at its lowest level of the past 600 Ma. Truly extensive shallow seas were restricted to the continents that were not part of Pangaea, notably South China. However, several major marine embayments pierced the coast of Pangaea and brought shallow seas into the heart of the supercontinent. Two of these embayments were associated with the development of rift systems that would eventually contribute to the break up of Pangaea. Thus, down the east coast of Greenland and along the east coast of Africa (west coast of Madagascar) two long and narrow seaways developed.

A major invasion of the sea also occurred in the western and southern USA. In this region one of the most spectacular reef systems of all time developed. This is now superbly exposed in the Guadalupe Mountains of Texas. Much of the original topography of this reef system has been re-exposed, and it is possible to walk from the reef flat down the steep reef front to the deep basinal setting in front of the reefs – now a desert floor in front of a mountain escarpment. The other major area of Permian marine deposition in the USA occurs in the Idaho–Wyoming–Utah region. Here the rocks are of a very different character from the Texan rocks: dark shales and phosphates dominate the succession. Phosphates form today where cold deep waters upwell on western-facing continental margins, and a similar origin appears likely for the Permian phosphates.

At low to intermediate latitudes the deep interior of Pangaea was very arid (see below) and so it is not surprising that many of the seaways that penetrated into the supercontinent were subject to intense evaporation. The biggest of these evaporitic basins occurred in the north-west European area, where the Zechstein Sea was gradually infilled with a vast thickness of sea salts – one of the thickest evaporite deposits in the geological record.

Climate

Vast areas in the centre of Pangaea were a long way from any seaway and so it is not surprising that arid conditions were common over many areas of the continent. This was particularly true in the northern hemisphere, where much of Europe and North America lay in the rain shadow of the Appalachian–Variscan mountain chain (**Figure 1**). Many geologists live in these regions today and so a general impression of the Permian as a time of aridity has tended to enter the literature. However, the view from the Permian polar regions is rather different, and extensive coal deposits – testimony to humid climates – are known from the Permian of Siberia and Antarctica. Many of these coals formed in forested areas that were at latitudes of over 65° (e.g. eastern Australia, Antarctica, and eastern Siberia). These deposits indicate substantial rainfall and a lack of permanent snow cover at the Permian poles, conditions that were very different from those encountered in polar areas today.

The 43 Ma duration of the Permian also saw substantial changes in climate, the most important of which was the waning of the southern-hemisphere glaciation that had begun in the Early Carboniferous. At the beginning of the Permian extensive ice-sheets covered large areas of Antarctica and South Africa and deposited a jumble of boulders and clays known as the Dwyka Tillite. Ice also advanced as far north as southern Australia and parts of South America, at palaeolatitudes of approximately 45° S, before gradually shrinking back. By the end of the Middle Permian there is no evidence of permanent ice anywhere. The reason for the termination of this major glacial interval, known as the Permo-Carboniferous glaciation, is not understood but it may relate to the drift of southern Pangaea away from the south pole. At the start of the Permian the pole lay within Antarctica but by the Late Permian it was within the southernmost Panthalassa Ocean.

Marine Fossils

Life in the Permian seas and oceans was remarkably similar to that in the preceding Carboniferous. In the tropical carbonate seas of equatorial latitudes the bottom-dwelling fauna was diverse and consisted of brachiopods, stalked echinoderms (especially crinoids), bryozoans, corals, lesser numbers of molluscs, and the occasional rare trilobite. Foraminifera (single-celled organisms that secrete a tiny calcareous skeleton) were also common and they produced a group of 'giants' in the Permian – the fusulinids – that reached up to a centimetre in diameter. These are abundant and highly diverse in some Permian limestones. Reefs were common in equatorial locations and included the magnificent examples from the Guadalupe Mountains (mentioned above) and many others from Tethyan settings. Calcareous sponges and calcareous algae (especially a form known as *Tubiphytes*) were the principal constructors of these

reefs, although the associated fauna was diverse. This included a bizarre group of brachiopods known as the richthofenids, which evolved a conical valve with a lid-like upper valve. In gross morphology they resembled the contemporaneous rugose corals, and they were undoubtedly pursuing a similar life strategy.

Land Animals

Prior to the Late Carboniferous–Early Permian it is probable that very few land animals ever strayed far from water, but this changed dramatically during the Permian with the colonization of the continental interiors, a major ecological breakthrough that was almost certainly triggered by the evolution of eggs that did not have to be laid in water. The majority of Permian tetrapods (four-legged land dwellers) belong to two major groups, the synapsids, a group that ultimately gave rise to mammals (but not in the Permian), and the diapsids, a group that gave rise to reptiles and later (in the Triassic) to the dinosaurs.

Unlike the relatively stable Permian marine invertebrate faunas, the Permian tetrapod faunas underwent several phases of turnover. The initial radiation was dominated by the pelycosaurs, a synapsid group that included the large sail-backed predator *Dimetrodon*, one of the few 'famous' Permian tetrapods. The pelycosaurs included both herbivores and carnivores of a broad range of sizes although all had a characteristic sprawling pose and dragged long tails. They would have undoubtedly looked clumsy compared with many modern mammals. In the later Permian the pelycosaurs gave way to an even more diverse group of synapsids known as the therapsids. These included the dicynodonts, which were large (up to 3 m long) powerfully built herbivores, the cynodonts, which were smaller herbivores with distinctive tusks in their upper jaws, and the rather fearsome-looking gorgonopsids, which were predators of around a metre in length. The diapsids contributed the pareiasaurs, a group of large powerful herbivores, to the impressively diverse Late Permian tetrapod communities.

Land Plants

Just as the conquest of the land by animals was probably facilitated by the evolution of eggs, so the parallel conquest by plants was probably driven by the evolution of seed-bearing plants. Thus, the Permian saw the decline of tree ferns, which dominated Carboniferous floras, and the rise of several groups of seed plants (particularly cycads and ginkgos) and seed ferns (glossopterids). Four floral realms, with a clear climate-controlled distribution, are readily distinguished in the Permian. The flora of the Angaran Realm, centred on Siberia, was adapted to the cool-to-cold climates of this region. Cordaites, a primitive group of seed ferns, and the fern-like sphenopsids dominated the flora as they had done in this region in the preceding Late Carboniferous. These two groups were also common, together with ginkgos, in the Euramerican flora, which occupied the semiarid and possibly monsoonal locations of some low northerly palaeolatitudes. The Cathaysian flora, centred on northern China, was adapted to the warm ever-wet climate of this equatorial part of the world. It was similar in composition to the tree fern-dominated flora of the preceding Carboniferous, although with the addition of the large-leaved gigantopterids, another group of seed ferns. The most distinctive Permian flora occurred in the Gondwanan realm, where deciduous glossopterid tree fern forests were extensive (Figure 2). These thrived in the cool temperate conditions that developed in high-latitude locations such as Australia and South Africa after the melting of the ice-caps in the Early Permian. Most of the coal reserves of these countries, together with the less exploitable coals of Antarctica, formed at this time.

Extinction

For a long time the end-Permian mass extinction (*see* **Palaeozoic:** End Permian Extinctions) was thought to be a protracted affair that had begun well before the end of the Permian. However, recent work has demonstrated that there were in fact two separate extinction events in the latter part of the Permian, separated by a phase of radiation and recovery in the Late Permian. The end-Middle Permian extinction, sometimes known as the end-Guadalupian event, is poorly understood, primarily because it has only recently been discovered and has been little studied, but also because there are few complete successions recording the transition from Middle Permian to Upper Permian strata. The effects of this crisis are best known from equatorial sections, particularly in southern China where many species of fusulinid and brachiopod went extinct. Reef development also ceased in many areas of the world, including the Guadalupe reefs of Texas, at around this time. What is far from clear is whether this extinction event can be detected at higher latitudes and whether there were any corresponding extinctions on land. The plant record does not appear to show any major change around this time.

Despite its recent discovery, several theories have already been proposed to explain the end-Guadalupian mass extinction. As noted above, complete sections spanning the Middle Permian–Upper Permian boundary are rare. This is because of a major fall in sea-level

Figure 2 Slab of rock from the Permian of Antarctica, showing abundant fronds of the tree fern *Glossopteris*.

at this time, possibly to one of the lowest levels of the past 600 Ma. The result was that shallow-marine habitats became very limited, with drastic consequences for the inhabitants of such environments. Climate change triggered by volcanic activity is also a possible factor in this extinction event (and in many other extinction events). The Emeishan volcanic traps of southern China are roughly contemporaneous with the extinction event, raising the possibility that the emissions of volcanic gas associated with the lava eruptions may have changed the climate. However, precise synchronicity of the eruptions and the extinction has yet to be demonstrated. Recently, a Middle Permian–Late Permian succession was discovered in a limestone succession that formed on a small island that was formerly within the Panthalassa Ocean (the rocks have now been accreted to Japan). Remarkably, an acidic volcanic ash is developed at the boundary. This is a product of the kind of explosive volcanism normally characteristic of continental rather than oceanic locations. A huge volcanic eruption in a continental location (perhaps Emeishan) may have blown ash out into the far reaches of the Panthalassa Ocean. If ash could reach this distant location, then it is possible that it could also have triggered drastic climate change. Only further study will demonstrate the veracity of the currently proposed mechanisms of the end-Guadalupian extinction.

See Also

Fossil Invertebrates: Porifera. **Fossil Vertebrates:** Palaeozoic Non-Amniote Tetrapods. **Mesozoic:** Triassic. **Microfossils:** Conodonts. **Palaeoclimates**. **Palaeozoic:** Carboniferous; End Permian Extinctions. **Pangaea**. **Sedimentary Environments:** Reefs ('Build-Ups'). **Sedimentary Rocks:** Evaporites.

Further Reading

Benton MJ (1997) *Vertebrate Palaeontology,* 2nd edn. London: Chapman & Hall.

Hallam A and Wignall PB (1997) *Mass Extinctions and Their Aftermath*. Oxford: Oxford University Press.

Parrish JT (1993) Climate of the supercontinent Pangaea. *Journal of Geology* 101: 215 233.

Willis KJ and McElwain JC (2002) *The Evolution of Plants*. Oxford: Oxford University Press.

End Permian Extinctions

R J Twitchett, University of Plymouth, Plymouth, UK

Introduction

The extinction event at the close of the Permian period was the largest of the Phanerozoic. Understanding this event is crucial to understanding the history of life on Earth, yet it is only since the late 1980s that scientists have begun to study this event in detail. This is partly because of the discovery of many new geological sections spanning the time of the extinction crisis, partly because of a renewed public and scientific interest in extinction, and partly because of increasing dialogue, travel, and exchanges between geologists and palaeontologists from the West (Europe and America) and those from Russia and China. Over the years, a bewildering number of hypotheses have been suggested to explain the end-Permian extinction crisis, and new data are constantly appearing. There is currently no consensus as to the precise cause, and the current leading contenders are considered below.

Definition and Dating

The Permian–Triassic boundary is now defined as the point at which a specific species of fossil conodont (an extinct type of primitive vertebrate) called *Hindeodus parvus* first appears in a geological section at Meishan in South China (Figure 1). This definition was accepted only after many years of debate: its significance is immense. Palaeontologists from around the world are now able to correlate their local sections with this single reference section, thus unifying a confusing plethora of names that geologists working independently, in different countries, had erected to describe rocks of the same age. It also confirmed that the main extinction event occurred in the latest Permian, before the Permian–Triassic boundary (i.e. before the appearance of *H. parvus*), in the Changhsingian Stage (Figure 2).

The presence of layers of volcanic ash interbedded with the fossiliferous sediments at Meishan have also allowed absolute dating of the boundary interval through the precise measurement of radiogenic elements, such as ^{206}Pb and ^{238}U, in minerals such as zircons. In 1991, Claoue-Long and colleagues dated the Permian–Triassic boundary at 251.2 ± 3.4 Ma. Seven years later, Sam Bowring and his group reanalysed the same beds and arrived at a date of 251.4 ± 0.3 Ma, confirming the earlier result but with far higher resolution and cementing 251 Ma in the subsequent literature.

However, accurate dating is notoriously difficult, and the further back in time one goes the worse the resolution becomes and the greater is the potential error. In 2001, Roland Mundil and colleagues confirmed that there were problems with the Bowring dates. Mundil's group reassigned the ash bed above the Permian–Triassic boundary an (older) age of 252.7 ± 0.4 Ma. However, the ash bed below the Permian–Triassic boundary, approximating to the level of the mass extinction, proved far more difficult to date: that it is probably older than 254 Ma was the best that the geochronologists could achieve.

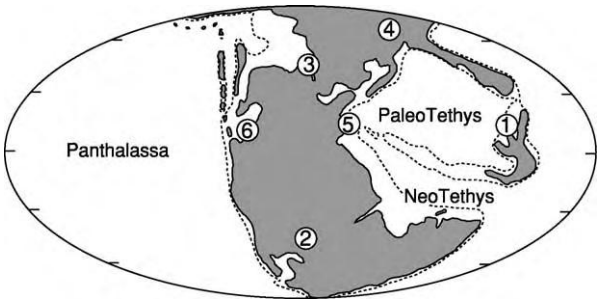


Figure 1 Sketch of Permian Triassic palaeogeography, showing the major regions discussed in the text: 1, South China; 2, Karoo Basin, South Africa; 3, eastern Greenland; 4, Siberia; 5, northern Italy; and 6, western USA.

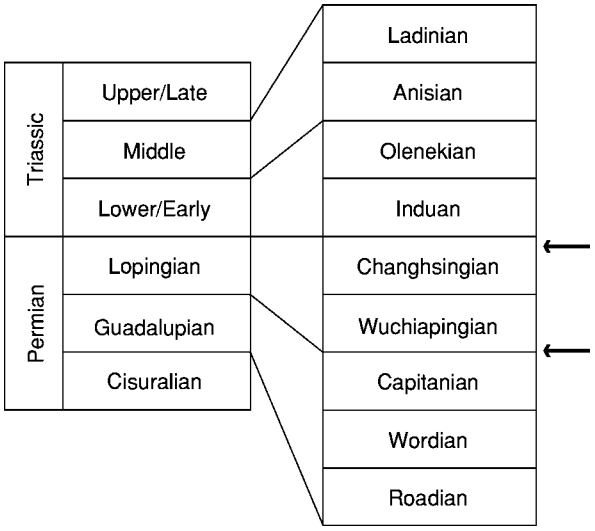


Figure 2 Permian and Triassic stratigraphy. Arrows indicate the stratigraphical positions of the extinction events discussed in the text.

Is this change in date and loss of resolution important? For many Permian–Triassic studies it is largely irrelevant. Correlation between marine sections relies on well-established methods of biostratigraphy and is unaffected by such results: the relative sequence of events can be deduced regardless. However, it is a significant blow for correlation between fossiliferous marine sediments and distant terrestrial sections, and for the hypothesis that the Siberian flood basalts were responsible for the extinction crisis (see below). It also means that absolute rates of change, both faunal and environmental, cannot be determined.

What Went Extinct?

Estimating the severity of past extinction events is not easy. Owing to the vagaries of preservation and fossil recovery and the difficulty of recognizing biological species from fossils, palaeontologists investigating extinctions tend to work at higher taxonomic levels. Thus, estimates of extinction magnitude are derived from large databases of family-level diversity through time, which show that 49% of marine invertebrate families disappeared during the Permian–Triassic interval.

Extinction at the species level is then estimated, rather than directly measured, using a statistical technique called reverse rarefaction. From this method, the often-quoted figure of a 96% loss of marine species is derived. However, such a calculation makes a number of assumptions, including, for example, the assumption that species extinction was random, with no selectivity against certain groups. The balance of evidence indicates that this assumption is incorrect, and, thus, the true magnitude of species loss may be closer to 80%. However, this still makes the end-Permian extinction event the most severe crisis of the Phanerozoic!

Marine Extinctions

Despite these shortcomings, it is clear that several groups were severely diminished in number in the Late Permian, some groups went completely extinct, and not all groups were affected equally. For example, the diverse and successful fusulinid foraminifera disappear suddenly during the Changhsingian stage with the loss of some 18 families, whereas other benthic foraminifera suffered much lower levels of extinction. The diversity of the stenolaemate bryozoans also decreased markedly through the latter half of the Permian, but only the Order Fenestrata became extinct. By contrast, other bryozoan groups, such as the gymnolaemate order Ctenostomata, maintained their diversity at the family level.

A pattern of gradual decline throughout the Permian, particularly during the latter half, followed by final extinction of the last few remaining taxa in the Changhsingian, is typical of many groups. Examples include the Palaeozoic corals (Rugosa and Tabulata), trilobites, and goniatites, all of which became extinct, and the articulate brachiopods, which were reduced to a handful of surviving genera. This pattern implies that longer-term (possibly climate driven) changes in the marine realm may have been largely to blame for the loss of diversity.

A few groups, such as the echinoderms, suffered dramatic losses prior to the end-Permian. All of the extant echinoderm groups experienced severe bottlenecks during the Permian–Triassic interval: only a couple of lineages of crinoids and echinoids survived into the Mesozoic. However, the major crisis interval appears to have been in the Late Guadalupian, when crinoids experienced over 90% loss and other groups of echinoderms, such as the Blastoidea, became extinct.

Terrestrial Extinctions

The extinction on land was just as severe as in the oceans, with some 77% of vertebrate families becoming extinct. In the Karoo Basin of South Africa (**Figure 1**), only seven out of 44 tetrapod genera (16%) cross the Permian–Triassic boundary. When all fossil terrestrial organisms are considered together, the Permian–Triassic event is the single largest extinction episode in the otherwise exponential rise of terrestrial diversity through the Phanerozoic.

A long-term (more than 20 Ma) change in terrestrial vegetation occurred during the Permian–Triassic interval. Termed the Palaeophytic–Mesophytic transition, it probably reflects gradual shifts in climate and palaeogeography. However, superimposed on this longer-term trend is a catastrophic collapse of the dominant gymnosperm forests in the latest Permian, as evidenced by the disappearance of pollen taxa such as *Vittatina*, *Weylandites*, and *Lueckisporites*. Studies in eastern Greenland (**Figure 1**) have shown that this ecological crisis occurred simultaneously with the marine extinction event.

Some workers have suggested that a peak in the abundance of *Reduviasporonites* (also called *Tympanicysta*), observed in some sections, marks the sudden destruction of these forests. They have interpreted this taxon as a saprophytic fungus, which thrived on the piles of dead and dying vegetation. Unfortunately, this attractive scenario must now be rejected. Recent geochemical, structural, and biomarker studies have shown that *Reduviasporonites* is definitely not a fungus and is probably a photosynthetic alga.

One Event or Two?

Early databases of global marine biodiversity recorded a broad peak of elevated extinction rates spanning the entire Late Permian. As the data improved, two distinct peaks were resolved: one in the latest Permian (the major event) and one at the end of the Guadalupian. This latter event marks the point at which many diverse groups (such as sponges and stenolaemate bryozoans) begin to decline in diversity. For the brachiopods and echinoderms, especially the crinoids, it marks the main episode of diversity loss.

While there is some evidence of climate change at this time, there are also good reasons why the end-Guadalupian event might not be a 'real' biotic extinction event. The Middle Permian was a time of incredible biodiversity, but the majority of the fossil taxa are endemic to the comprehensively monographed and exquisitely preserved silicified faunas of the Guadalupian reef belts of the southern USA, particularly western Texas. At the close of the Guadalupian, sea-level fell across that region, and the overlying sediments are unfossiliferous evaporites. The disappearance of the fossils is therefore a result of facies change and not real extinction.

Globally, Upper Permian fossiliferous marine rocks are relatively poorly known, and the quality (i.e. completeness) of the Late Permian fossil record is correspondingly poor. Many taxa that were common in the Guadalupian have not been found in Upper Permian rocks but must have been living at that time because they reappear in the Triassic. These missing taxa are called Lazarus taxa. The presence of many Lazarus taxa indicates that the completeness of the fossil record is low (Figure 3).

The reality of the end-Guadalupian event is still debated. If it is simply the result of bias in the rock record, then newly described sections should contain many of the taxa previously supposed to have become extinct. Certainly this is true of the brachiopods: newly described Upper Permian sections from Tibet contain some taxa that were thought to have vanished earlier. Further detailed analysis of this interval is clearly needed.

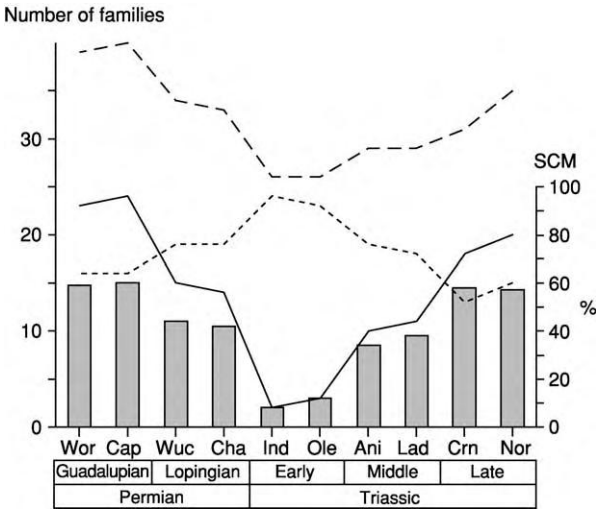


Figure 3 Fluctuations in fossil diversity and the quality of the fossil record of marine sponges through the Permian Triassic extinction and recovery interval: solid line, families of sponges represented by actual fossils; dotted line, number of Lazarus taxa; dashed line, total diversity (sum of fossil taxa and Lazarus taxa). Grey bars show the simple completeness metric (SCM), which is the number of actual fossil taxa as a percentage of the total diversity. The SCM gives an estimate of the completeness of the fossil record. Note that, as diversity declines after the Guadalupian, the number of Lazarus taxa increases and so the completeness of the fossil record decreases. Stratigraphy from Figure 2. Data from *The Fossil Record 2* (Benton, 1993).

Possible Causes

Extraterrestrial Impact

Initial attempts to find evidence of a Permian–Triassic impact were made in the early 1980s, following the proposal that a massive meteorite impact had caused the end-Cretaceous extinction event (*see* **Mesozoic:** End Cretaceous Extinctions). Three pieces of evidence are considered crucial in identifying a meteorite impact event: a crater, an enrichment ('spike') of the rare metallic element iridium, and a layer of impact debris containing shocked quartz and tektites. However, the few reports of these pieces of evidence in latest Permian rocks are considered to be, at best, equivocal. Certainly, there is no Permian–Triassic iridium spike comparable to that found at the Cretaceous–Tertiary boundary, no widespread impact layer, and no large crater that can be accurately correlated with the extinction horizon. For an impact event to be a possible cause of the extinction crisis, it must also be demonstrated that the extinctions were unequivocally instantaneous and coincident with the impact itself.

In 2001, Luann Becker and colleagues reported the presence of helium and argon trapped in the cage-like molecular structure of fullerenes from the Permian–Triassic boundary in China and Japan. Isotopic profiles indicated that the helium and argon, and hence the fullerenes, had to have come from an extraterrestrial source. However, there are persistent problems with the acceptance of Becker's data, as other scientists have consistently failed to replicate her results, despite using samples from exactly the same sites and exactly the same laboratory procedures. A Japanese expert, Yukio Isozaki, has also argued that the Permian–Triassic boundary is missing from the Japanese section studied by Becker and colleagues

and that their samples came from below the extinction level.

Later in 2001, Kaiho and colleagues described Permian–Triassic sediment grains that were supposedly formed by impact, as well as geochemical shifts that they interpreted as indicative of a huge impact. However, their data are far from conclusive and were rapidly and severely criticized by other geochemists. In 2003, Asish Basu and colleagues revived interest by claiming to have found 40 tiny (50–400 μm) unaltered fragments of meteorite in a sediment sample from a terrestrial Permian–Triassic boundary section in Antarctica. Although Basu and colleagues were quick to dismiss contamination as a source of these grains, meteorite experts were immediately sceptical of the findings, as meteoritic metals are highly reactive and, in terrestrial settings, oxidize extremely quickly.

Doubtless scientists will continue to produce evidence of extraterrestrial impact at or near the Permian–Triassic boundary, if only because such hypotheses readily appeal to the wider public and journal editors alike. However, current data are highly controversial and lack the key criteria of major impact that have been recorded time and again in Cretaceous–Tertiary sections the world over. Independent replication of results is crucial for scientific acceptance, especially when the data are unusual and controversial. So far, all the evidence proposed for an impact at the Permian–Triassic boundary has failed this necessary test.

Eruption

The largest outpouring of continental flood basalts in the Phanerozoic occurred in Siberia during the Permian–Triassic interval. Including both the Siberian Platform basalts and the newly discovered coeval deposits buried in the West Siberia Basin, described by Marc Reichow and colleagues, the flood basalts covered an area of $1.6 \times 10^6\,km^2$ to maximum depths of 3.5 km (Figure 1). If all other igneous rocks, such as pyroclastic flows, are included, then this coverage increases to $3.9 \times 10^6\,km^2$. Dating the top and bottom of the lava pile shows that the eruptions occurred over a relatively short period of time, maybe just 600 000 years. Was this huge volcanic event a cause of the Permian–Triassic extinction crisis?

Radiometric dating is the only way to answer this question because no fossils have yet been collected from sediments interbedded with the basalts that provide correlation with other regions. Early efforts at dating the Siberian Traps produced an array of ages, from 160 Ma to 280 Ma. By contrast, more recent results, by different scientists using a variety of methods, cluster around 250 ± 1 Ma. In the late 1990s, this was considered to be exactly the date required, and the flood basalts were thought to be the primary trigger for the catastrophic extinction. However, the recent redating of the Meishan beds now implies that the Permian–Triassic event occurred before 253 Ma, and the Siberian Traps are therefore several million years too young.

On a Phanerozoic time-scale, the excellent correlation between extinction episodes and flood-basalt provinces means that it is difficult to accept the possibility that the Siberian Traps played no role in the end-Permian extinction event. There is still a chance. The oldest date that Reichow's team have for the onset of volcanic activity (dating the emplacement of intrusive gabbros) is 253.4 ± 0.8 Ma.

Another intriguing consequence of the 253 Ma age for the Permian–Triassic boundary is that the Emeishan flood basalts of western South China now need to be considered. Dating by Ching-Hua Lo and colleagues (published in 2002) shows that the main eruptions, which were largely marginal marine events, occurred around 251–253 Ma, with an initial phase of activity at around 255 Ma; thus, they predate the Siberian Traps and are dated close to the new dates for the Permian–Triassic crisis. Interestingly, Ching-Hua Lo's new dates preclude the possibility that the Emeishan flood basalts played a role in the end-Guadalupian crisis (at about 256–259 Ma; see above). Mei-Fu Zhou and colleagues had championed this particular correlation following their dating of the Emeishan basalts to 259 ± 3 Ma, which was published just a few months before Ching-Hua Lo's results in the same journal! Absolute dating is a difficult business (*see* **Analytical Methods:** Geochronological Techniques).

Global Warming

Animals and plants do not curl up and die just because a volcano is erupting on the other side of the world. Species become extinct when their population falls below viable levels. Population decline in natural systems is a response to local changes, such as loss of habitat, drought, and temperature change, many of which are ultimately driven by climate. What climate changes are recorded around the extinction interval and how might they be related to the flood-basalt volcanism?

Global cooling was proposed early on as a cause of the Permian–Triassic event; this largely hinged on the identification of latest-Permian glacial deposits in Siberia and eastern Australia. However, reanalysis of the biostratigraphy proved that these deposits are,

in fact, Middle Permian in age. With evidence of extensive ice sheets confined to the Early Permian and the last vestiges of glaciation now dated as Middle Permian, it is evident that the Permian as a whole records a warming trend.

At the culmination of this long-term warming trend, there is also evidence for an additional rapid greenhouse episode at the time of the extinction crisis. A large drop in the proportion of heavy oxygen isotopes (^{18}O) in carbonates from the Gartnerkofel-1 core of southern Austria has been interpreted as representing an increase in temperature of 6°C. However, interpretation of the oxygen isotope record is problematic as it is very sensitive to alteration during burial and diagenesis. The limestones of Gartnerkofel-1 have been largely recrystallized, especially around the boundary interval, and thus the oxygen isotope data are surely suspect.

However, there is growing independent support for global warming at the boundary. Gregory Retallack has demonstrated, from studies of the stomatal index of fossil leaves, that peaks in atmospheric CO_2 (a proven greenhouse gas) occur at both the end-Permian and end-Guadalupian boundaries. His studies of fossil soils also indicate a change to humid greenhouse conditions in the Early Triassic. Finally, there is a large (usually 3–6‰) increase in the proportion of light ^{12}C, relative to heavier ^{13}C, in carbonates deposited at this time. This shift is interpreted by most geochemists to be the result of methane (CH_4) release, as this is the only mechanism currently known that could rapidly deliver the required amount of light ^{12}C to the atmosphere and oceans. This methane, which is also a potent greenhouse gas, is assumed to have been produced by the melting of methane hydrate deposits in shallow marine shelves and polar tundra. Thus, a runaway greenhouse scenario is envisaged, whereby warming led to melting of gas hydrates and release of methane, which fuelled further warming and further methane release, in a positive-feedback loop that presumably continued until the reserves of gas hydrate were depleted (Figure 4).

Flood basalts are assigned a triggering role in this runaway greenhouse model, by venting enough initial CO_2 to drive temperatures above the threshold for gas-hydrate breakdown. However, volcanic eruptions also reduce global temperatures by releasing particulates and SO_2. Some authors have incorporated a cooling episode into their extinction models, but regard it as a brief event, which was subsequently eclipsed by the later warming. There is currently no geological evidence to support an episode of global cooling at the onset of the Permian–Triassic extinction crisis.

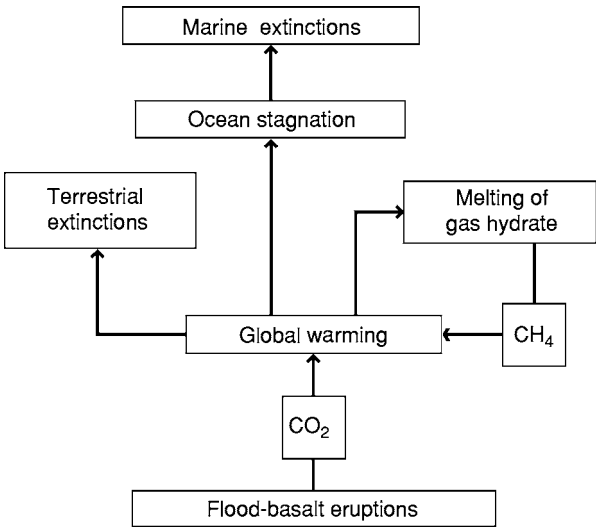


Figure 4 The runaway greenhouse model for the end Permian extinction event. See text for details.

Post-Extinction Recovery

Douglas Erwin has remarked that the duration of the post-Permian recovery interval is the greatest of any Phanerozoic mass extinction and is proportionally longer than would be expected given the magnitude of the species loss. Certainly, a literal reading of the fossil record shows that it took some 100 Ma (until the mid-Jurassic) for marine biodiversity at the family level to return to pre-extinction levels. However, ecological recovery was somewhat quicker, with complex communities such as reefs being re-established some 10 Ma after the Permian–Triassic boundary. Thus, how one views the duration of the recovery interval, measured on a local or global scale, depends on how one defines 'recovery'.

Details of the marine recovery are best known from the low palaeolatitudes of Panthalassa (the present-day western USA) and Palaeotethys (e.g. northern Italy) (Figure 1). Epifaunal benthic communities in the immediate aftermath of the event (the Early Induan) comprise low-diversity assemblages of small suspension feeders (typified by the bivalves *Promyalina* and *Claraia* and often the inarticulate brachiopod *Lingula*). Bedding planes are dominated by a single taxon, which may occur in prodigious numbers. These taxa are considered by many, with varying degrees of evidence, to be pioneering r-selected opportunists (i.e. animals that mature early and produce many offspring). Stromatolites and other microbial mats are commonly encountered and may have provided grazing opportunities for the depauperate microgastropod community. Occasional horizons of mm-diameter *Planolites* burrows attest to temporary colonization events by a scarce infauna

of small deposit feeders living just 1–2 cm below the sediment surface. In the higher palaeolatitude regions of the Boreal Ocean (Greenland, Spitsbergen) and NeoTethys (Madagascar, western Australia) a fairly diverse but small-sized nekton of fishes and ammonoids is recorded.

Sedimentary and geochemical data indicate that environmental conditions were particularly harsh. Most recent palaeoenvironmental studies have concluded that the Early Triassic oceans contained very low concentrations of dissolved oxygen and may have been anoxic or even euxinic for considerable periods of time. Low oxygen levels were probably the result of a combination of factors. First, levels of atmospheric oxygen had fallen steadily from 30% in the Early Permian to approximately 15% (i.e. 70% of present-day values) by the Permian–Triassic boundary. Second, global warming in the latest Permian (see above) would have increased sea-surface temperatures, and warm water holds less dissolved oxygen than cool water. Finally, climate models indicate that global warming would also lead to a reduction in the temperature gradient between the poles and the equator, thus slowing down thermohaline circulation. In present-day oceans, the temperature difference between the warm equator and the cold poles drives a relatively vigorous thermohaline circulation, which maintains oxygenation of the ocean floors. In the sluggish earliest Triassic oceans, oxygen would have been used up faster than it could be replaced, resulting in stagnation and anoxia.

The ecology of the post-extinction survivors and the long interval of low diversity are interpreted as being the result of this anoxic event. In the few places where the oceans were well-oxygenated in the immediate aftermath of the extinction (e.g. Oman), a far more diverse benthic community, comprising upwards of 20 genera, was soon established. Unfortunately, these localized experiments in rapid post-extinction recovery were snuffed out by the arrival of anoxia in the later Induan.

With the disappearance of benthic oxygen restriction later in the Early Triassic, larger organisms and more diverse communities reappeared. The ecological complexity of infaunal communities began to increase with the reappearance of deeper-burrowing suspension feeders (producing *Arenicolites*, *Skolithos*, and *Diplocraterion* burrows). Burrowing crustaceans (producing traces such as *Thalassinoides*) are the last significant component of the infauna to reappear. In the palaeotropics, *Thalassinoides* first reappears in the Middle Triassic. In higher palaeolatitudes (East Greenland, Spitsbergen), small *Thalassinoides* reappear in the later Induan, suggesting faster recovery in the Boreal Realm.

In most parts of the world, epifaunal complexity did not significantly recover until the Late Olenekian, with an increase in tiering (vertical stratification of organisms above the sediment surface) as crinoids and bryozoans returned. However, the most complex ecological structures – metazoan reefs – do not return until the Middle Triassic, accompanied by a significant increase in faunal diversity and the reappearance of many Lazarus taxa.

Our understanding of the recovery of terrestrial ecosystems is less refined, with most data deriving from the Karoo Basin of South Africa. Like the marine survivors, terrestrial vertebrate survivors tended to be small: in the Karoo Basin five small carnivorous therocephalians and one small anapsid (*Procolophon*) survived the Permian–Triassic crisis. The other survivor, and the dominant terrestrial vertebrate for several million years, was the herbivore *Lystrosaurus*, which, like *Claraia*, was globally widespread.

More is known of the recovery of the terrestrial vegetation. The most detailed studies are those of Cindy Looy and colleagues, who have analysed fossil spores and pollen from Europe and eastern Greenland. Following the collapse of the Late Permian gymnosperm forests, open herbaceous vegetation rapidly took over, with short-lived blooms of pioneering opportunistic lycopsids, ferns, and bryophytes – stress-tolerant forms that were subordinate members of the pre-crisis vegetation. Pollen from woody gymnosperms seems to indicate that a few surviving elements of the Permian forests lingered on for a while, but equally these records could just represent reworking of the pollen. Certainly, by the earliest Triassic no tree-like gymnosperms remained, and a stable low-diversity open-shrubland vegetation of cycads and lycopsids was established. Complex and diverse forest communities did not reappear, at least in Europe, until the latest Olenekian and early Middle Triassic. Thus, the return of ecological complexity on land closely mirrors that in epifaunal marine communities.

Conclusions

The greatest mass-extinction event of the Phanerozoic is receiving an unprecedented level of scientific interest. New sections are being discovered and described, allowing new hypotheses of causes and consequences to be tested and old ones to be modified or rejected. The 1990s witnessed a number of major advances, culminating in the acceptance of Meishan, China, as the appropriate locality to define the Permian–Triassic boundary and the subsequent formalization of Permian–Triassic stratigraphy. The

picture presented here is a current (2004) view of this crisis: a snapshot that doubtless will be refined in the decade to come.

See Also

Analytical Methods: Geochronological Techniques. **Carbon Cycle**. **Impact Structures**. **Large Igneous Provinces**. **Mesozoic:** Triassic; End Cretaceous Extinctions. **Palaeoclimates**. **Palaeozoic:** Permian. **Sedimentary Environments:** Anoxic Environments.

Further Reading

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Benton MJ (2003) *When Life Nearly Died: The Greatest Mass Extinction of all Time*. London: Thames and Hudson.

Benton MJ and Twitchett RJ (2003) How to kill all life: the end Permian extinction event. *Trends in Ecology and Evolution* 18: 358 365.

Erwin DH (1993) *The Great Paleozoic Crisis: Life and Death in the Permian*. New York: Columbia University Press.

Erwin DH (1994) The Permo Triassic extinction. *Nature* 367: 231 236.

Erwin DH (1996) Understanding biotic recoveries: extinc tion, survival and preservation during the end Permian mass extinction. In: Jablonski D, Erwin DH, and Lipps JH (eds.) *Evolutionary Paleobiology*, pp. 398 418. Chicago: University of Chicago Press.

Hallam A and Wignall PB (1997) *Mass Extinctions and Their Aftermath*. Oxford: Oxford University Press.

PANGAEA

S G Lucas, New Mexico Museum of Natural History, Albuquerque, NM, USA

Introduction

Geologists apply the term Pangaea (from the Greek words meaning 'all Earth', and usually pronounced pan-JEE-uh) to the supercontinent that existed during the Late Palaeozoic and Early Mesozoic, about 300–200 Ma ago. Pangaea was an amalgamation of continental blocks that were the precursors of the existing continents, though much of eastern Asia was an archipelago of islands loosely connected to eastern Pangaea.

Components of Pangaea

Pangaea (Figure 1) was accreted from continental blocks that differed from today's continents. The southern supercontinent of Gondwana (also called Gondwanaland) (*see* **Gondwanaland and Gondwana**) was the Palaeozoic amalgamation of South America, Africa, Antarctica, Australia, and the Indo-Pakistani subcontinent, as well as several smaller terranes. The northern supercontinent of Laurussia (or Laurasia) consisted of North America and most of Europe.

Asia, however, did not exist as a single block. The continental nucleus of present-day Asia was the Siberian block, but other blocks included Kazakhstan, Tarim, and a loose archipelago of blocks that were to become much of China and South-east Asia. Many of these eastern Asian blocks originated as slivers of northern Gondwana and drifted northwards during the Late Palaeozoic and earliest Mesozoic to accrete to the larger Asian blocks during the Late Triassic–Jurassic.

Pangaea was surrounded by a universal ocean called Panthalassa (from the Greek words meaning 'all sea'), which was the ancestor of today's Pacific Ocean. The Atlantic Ocean did not exist even in ancestral form, because of the fusions of North America to Europe and of South America to Africa.

An arm of Panthalassa formed a deep east–west embayment in the eastern edge of Pangaea (Figure 1). This was the Tethys Sea, which is named after the sister and consort of Oceanus in Greek mythology. Tethys was the ancestor of the present Mediterranean Sea. Successive ocean basins within it are termed Palaeotethys (Devonian–Permian) and Neotethys; the latter opened during the Late Permian as a result of rifting between Gondwana and the smaller central and southern Asian terranes.

Late Carboniferous Accretion of Pangaea

The Pangaean supercontinent came together (accreted) at the end of the Carboniferous (*see* **Palaeozoic:** Carboniferous), when Laurussia and Gondwana were sutured along what has been termed the Hercynian megasuture. Very old mountain ranges mark the

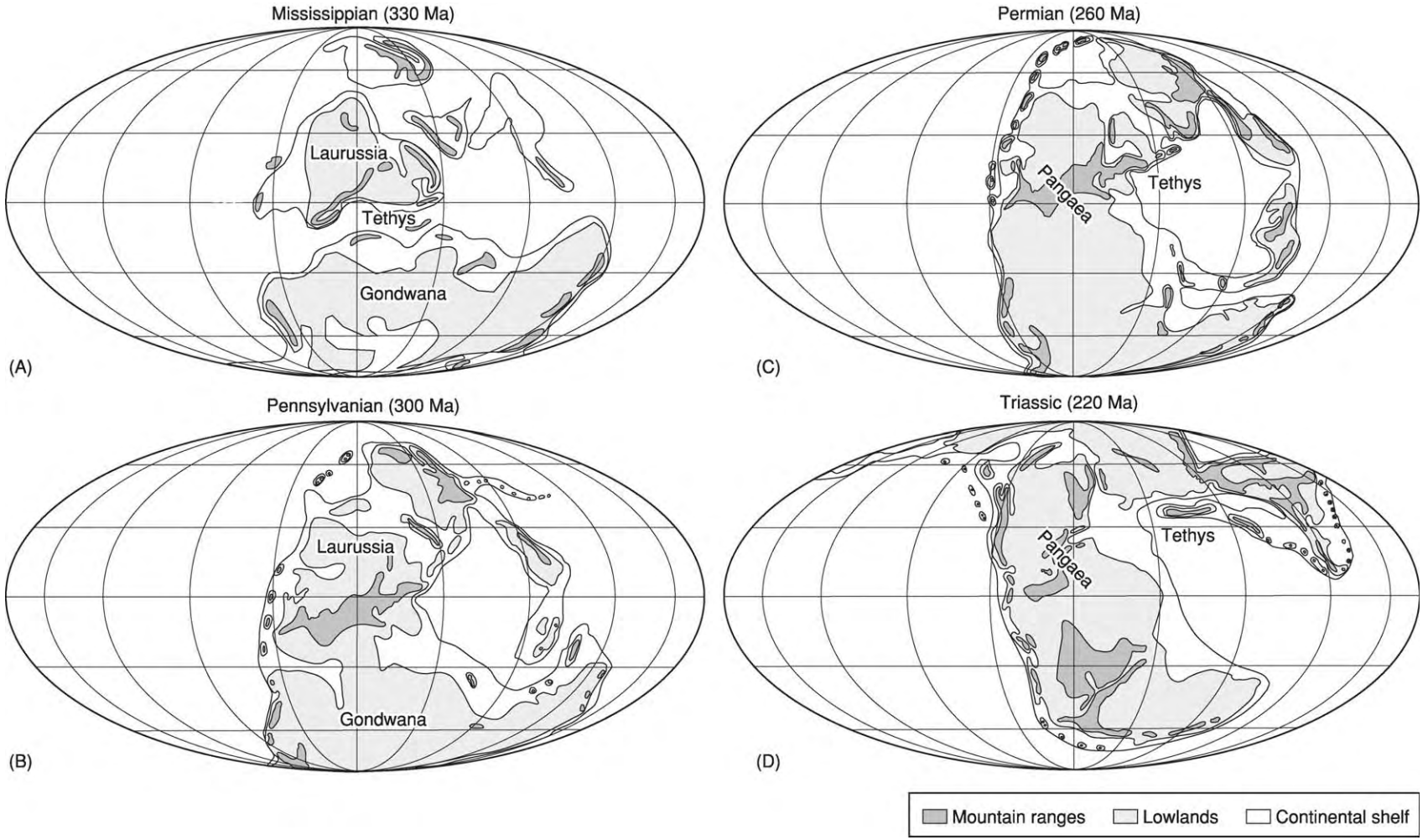


Figure 1 Palaeogeographical maps. (A,B) In the Carboniferous, Pangaea was accreted by the suturing of Gondwana and Laurussia near the palaeoequator (A) Early Carboniferous (Mississippian), and (B) Late Carboniferous (Pennsylvanian). Pangaea was fully united during (C) the Permian and (D) throughout much of the Triassic.

collisional boundaries of the Pangaean blocks. These are the Appalachian Mountains of eastern North America (*see* **North America:** Northern Appalachians; Southern and Central Appalachians), the Ural Mountains of European Russia (*see* **Europe:** The Urals), and the Hercynian Mountain ranges of southern Europe and parts of North Africa.

The Late Carboniferous was a time of vast coal swamps in the tropical latitudes. There was a steep temperature gradient during the Late Carboniferous, from icy poles to hot tropics, which was more similar to today's world than perhaps at any other time in Earth history. In Gondwana, ice ages pushed glacial ice to within 30° of the palaeoequator.

Permian Pangaea

Once assembled, at the beginning of the Permian, Pangaea stretched from pole to pole in a single hemisphere (Figure 1C). The ocean of Panthalassa covered the other hemisphere.

Permian Pangaea was a relatively diverse place in terms of climate and topography. Permian glacial deposits found in South America, Africa, India, and Australia provide evidence of continuing glacial ages in southern Gondwana during the Early Permian. Along the sutures of Pangaea, huge mountain ranges towered over vast tropical lowlands. Interior areas were dry deserts where dune sands accumulated. Evaporites (particularly gypsum and halite), deposited in the south-western USA and northern Europe, resulted from the evaporation of hot shallow seas and form the most extensive salt deposits in the geological record. Perhaps the best testimony to the diversity of Permian Pangaea are its fossil plants, which identify several floral provinces across the vast supercontinent.

The coastlines of the Pangaean tropics were warm, and the waters were teeming with life. The cold offshore waters of the poleward portions of Pangaea had a very different and less diverse biota. The biotic extinctions (turnover) at the end of the Permian were the largest of the entire Phanerozoic (*see* **Palaeozoic:** End Permian Extinctions); they were undoubtedly influenced by the vast volcanic outpourings in Russia (the Siberian traps) and South China (the Emeishan traps).

Triassic Pangaea

At the beginning of the Triassic, the Pangaean supercontinent still sat in a single hemisphere surrounded by the enormous Panthalassan Ocean. Subduction zones that dipped beneath the continents nearly encircled Pangaea, and elevations were relatively high owing to regional uplift across Pangaea during the Late Permian and Triassic. Intense mountain building and foreland-basin deformation took place around the Pangaean margins during the Late Permian and earliest Triassic. The maximum accretion (integration) of Pangaea occurred at about the end of the Middle Triassic, approximately 230 Ma ago.

Early–Middle Triassic Pangaea was a much more cosmopolitan world than Permian Pangaea. Triassic fossil plants indicate only two vast provinces, north and south of the palaeoequator. However, some land animals, such as the mammal-like reptile *Lystrosaurus* (Figure 2), roamed across much of Pangaea. Indeed, recognition during the 1970s of the widespread nature of Early Triassic fossils of *Lystrosaurus* (which are found in Russia, China, India, Africa, and Antarctica), a land animal, provided compelling evidence of the unity of Pangaea.

During the Triassic, the vast Pangaean supercontinent drifted northwards and rotated counterclockwise. The landmass was nearly symmetrical about the palaeoequator. However, no sooner had the maximum integration of Pangaea occurred, than the supercontinent began to fragment.

Thus, by the Late Triassic, separation of Laurussia and Gondwana had begun, with the onset of rifting in the Gulf of Mexico basin. Indeed, the breakup of Pangaea began in the western Tethys and severed the supercontinent into two pieces, north and south, by the Jurassic. The east–west rifting that opened up the Atlantic Ocean basin was not really achieved until the Cretaceous. However, around the nascent North Atlantic basin, a huge basaltic volcanic province, called the

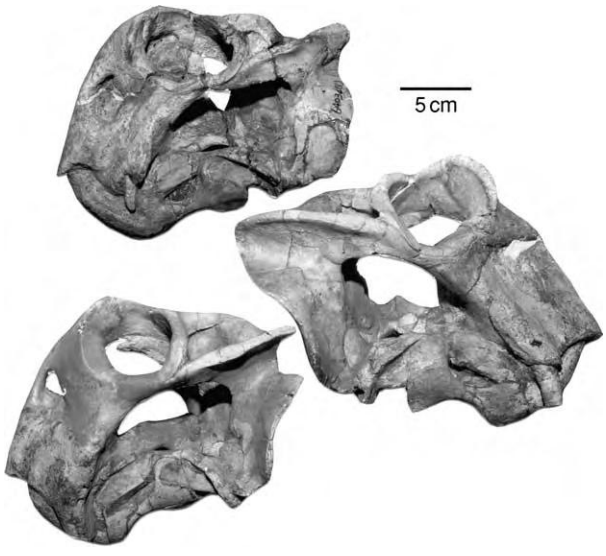


Figure 2 Skulls of the Early Triassic mammal like reptile *Lystrosaurus* from northern China. Fossils of *Lystrosaurus* are also known from Russia, India, Africa, and Antarctica, providing strong evidence of the unity of Pangaea.

circum-Atlantic magmatic province, appeared at about the beginning of the Jurassic, as lavas poured forth along fractures that formed as Pangaea rifted apart.

An important aspect of Triassic Pangaea was the large number of microplates (also called displaced or suspect terranes) that were to accrete to the Pangaean nucleus during the Late Triassic and Jurassic. These microplates were of variable size and can be identified by the sharp contrasts between their rocks and fossils and those of adjacent rocks. They are fragments of other continents, island arcs, or pieces of oceanic crust that moved long distances to be accreted to continents. More than 100 such terranes have been identified in the Cordilleran belt of western North America. Numerous terranes in the Panthalassa Ocean moved eastwards to join the western ends of the Americas. In eastern Tethys, as noted above, rifted slivers of northern Gondwana moved northwards to accrete to and form much of eastern and South-east Asia during the Late Triassic–Jurassic.

See Also

Asia: Central; South-East. **Europe:** The Urals. **Gondwanaland and Gondwana**. **Mesozoic:** Triassic. **North America:** Northern Appalachians; Southern and Central Appalachians. **Palaeozoic:** Carboniferous; Permian; End Permian Extinctions. **Terranes, Overview**.

Further Reading

Behrensmeyer AK, Damuth JD, DiMichele WA, *et al*. (eds.) (1992) *Terrestrial Ecosystems Through Time*. Chicago: University of Chicago Press.

Embry AF, Beauchamp B, and Glass DJ (eds.) (1994) *Pangea: Global Environments and Resources*. Memoir 17. Calgary: Canadian Society of Petroleum Geologists.

Lucas SG (2000) The epicontinental Triassic, an overview. *Zentralblatt für Geologie und Paläontologie Teil I* 7 8: 475 496.

Marvin UB (1973) *Continental Drift*. Washington DC: Smithsonian Institution Press.

Sherlock RL (1948) *The Permo Triassic Formations*. London: Hutchinson's Scientific and Technical Publications.

Smith AG and Briden JC (1977) *Mesozoic and Cenozoic Palaeocontinental Maps*. Cambridge: Cambridge Univer sity Press.

Tarling D and Tarling M (1971) *Continental Drift*. Garden City: Doubleday.

Ziegler PA (1989) *Evolution of Laurussia*. Dordrecht: Kluwer Academic Publishers.

PETROLEUM GEOLOGY

Contents

Overview

J Gluyas, Acorn Oil and Gas Ltd., Staines, UK

Introduction

Petroleum geoscience is defined as the disciplines of geology and geophysics applied to the understanding of the origin, distribution, and properties of petroleum and petroleum-bearing rocks.

Petroleum geoscience can be described as the study and understanding of five key components: source, seal, trap, reservoir, and timing (of petroleum migration). These are sometimes known as the 'magic five ingredients' without which a basin cannot become a petroleum province (Figure 1). This article examines each of these components.

Source Rock

Petroleum (oil and gas) forms from organic matter: dead plants and animals. Burial, and thus heating of such organic matter induces reactions leading to the generation of gas, then oil and gas, and, finally, gas alone as the temperature and residence time at high temperature increase. Continued burial and heating turn the residual organic matter into pyrobitumen and eventually into graphite (*see* **Petroleum Geology:** The Petroleum System).

Seal

Oil and gas are less dense than water and, following expulsion from the source rock, they rise towards the Earth's surface unless movement is arrested by a seal. Seals tend to be fine-grained or crystalline, low-permeability rocks, such as mudstone/shale, cemented limestones, cherts, anhydrite, and salt (halite). Seals can also develop along fault planes, faulted zones, and fractures.

The presence of seals is critical for the development of petroleum pools. In the absence of seals, petroleum will rise to the Earth's surface and be destroyed. Although seals are critical for the development of petroleum pools, none are perfect. All leak. Combinations of regionally extensive seals and underlying reservoir complexes are commonly referred to as 'plays', and the areas within which the quality of seals and reservoirs is such that petroleum accumulations could occur (given an appropriate trapping configuration) are commonly referred to as 'play fairways'.

The most common subdivision of seals distinguishes between seals in which petroleum is unable to force its way through the largest pores (membrane seals) and seals in which petroleum can escape only by creating fractures (hydraulic seals). Attributes which favour a rock as a seal include a small pore size, high ductility, large thickness, and wide lateral extent. The physical properties of the water and petroleum are also important. Water salinity, petroleum density, and interfacial tension between petroleum and water are the most important, and these properties will change according to changing pressure and temperature conditions.

The most common lithology that forms a petroleum seal is mudrock. Mudrocks are composed of either carbonate or siliciclastic minerals (or both), and mudrock sequences are often thick (>50 m) and laterally continuous (>1.0 km^2). Examples of mudrock seals are found in all deltaic settings (including the Gulf of Mexico, Niger, and Nile Delta petroleum

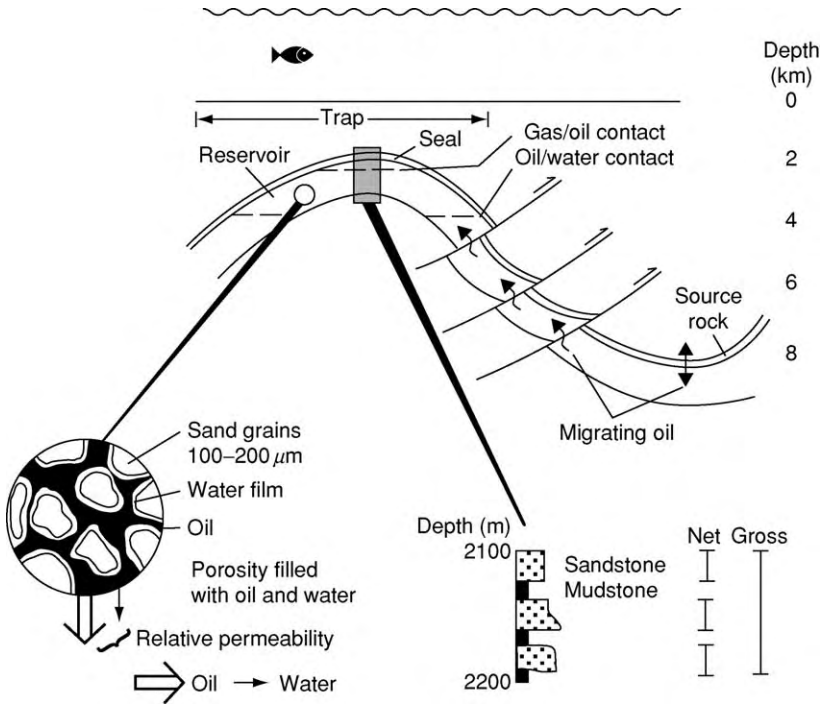


Figure 1 Diagram of a cross section of a petroleum bearing basin, illustrating the five key components: source rock, seal, reservoir, trap, and petroleum migration. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

provinces) and many interior, rift, and passive continental margin basins (including north-west Australian shelf, north-west Europe, North Sea, and south-east Asia). Halite can form a more effective seal, but is a relatively rare lithology found only where conditions of high evaporation of seawater have taken place. The Upper Permian rocks of north-west Europe contain Zechstein halite that is known to have trapped gas for long periods of geological time. Halite forms part of the sealing lithology in many of the large Middle East petroleum accumulations.

Membrane Seal

When petroleum is trapped beneath a seal, there is a buoyancy pressure (P_b). The magnitude of the buoyancy pressure is a function of the contrast in density between the water (ρ_w) and petroleum (ρ_p), and its height (h) above the free water level where both are in equilibrium (normally close to the petroleum–water contact). The relationship can be written as

$$P_b = (\rho_w - \rho_p)h$$

where P_b is expressed in units of pressure (e.g., bar, psi, MPa) and the fluid densities are expressed as pressure gradients (e.g., bar m ^{1}, psi ft ^{1}, Pa m ^{1}) (Figure 2). The maximum petroleum column is controlled by the capillary entry pressure of the petroleum into the largest pores in the seal. The capillary entry pressure (P_d) of a water-wet rock is given by the equation

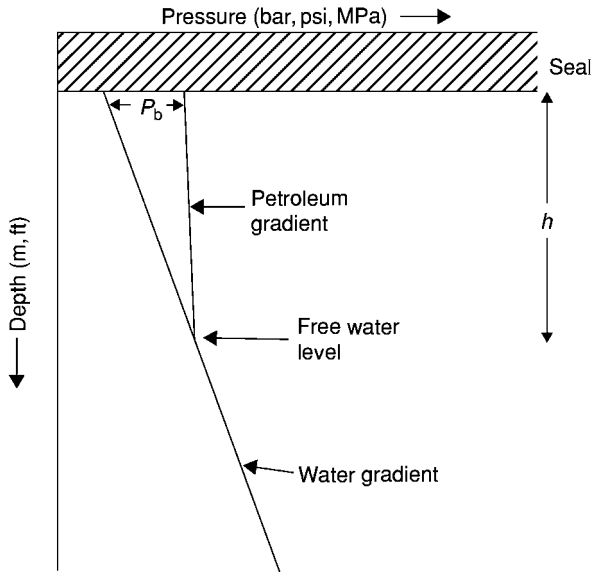


Figure 2 A pressure vs. depth plot, illustrating a typical water gradient (aquifer) supporting a petroleum column, whose steeper gradients lead to a pressure difference (P_b) at its maximum beneath the seal. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

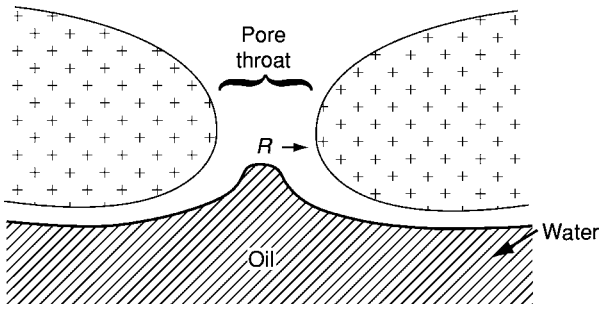


Figure 3 A schematic illustration of a pore throat between two grains. The radius of the pore throat and the buoyancy pressure, plus the interfacial angle and surface tension between oil and water, determine whether oil can migrate through the pore throat or remain trapped beneath. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

$$P_d = 2\gamma \cos\theta / R$$

where γ is the interfacial tension between the water and the petroleum, θ is the contact angle, and R is the radius of the largest pore (Figure 3). The interfacial tension and contact angle change with increasing temperature and pressure and are related to fluid type and density. These properties are routinely established from laboratory experiments on rocks. The seal capacity determines the height of a petroleum column that can be trapped beneath it, and the seal will be breached when the buoyancy pressure (P_b) exceeds the seal capillary entry pressure (P_d).

Hydraulic Seal

When the capillary entry pressure of the rock is extremely high, for example in evaporites, the failure of the seal is controlled by the strength of the rock that, if exceeded, creates natural tension fractures (*see* **Tectonics:** Faults; Fractures (Including Joints)). The rock will fracture when the pore pressure is greater than both the minimum stress and the tensile strength of the rock. Structural geologists describe the stresses in rock in terms of three orthogonal components of stress (Figure 4), one oriented vertically (S_v) and the other two oriented horizontally (S_{hmin} and S_{hmax}). In relaxed sedimentary rocks found in an extensional basin or a young delta, S_v is usually greater than both S_{hmin} and S_{hmax}, and so the minimum stress (σ_3) is horizontal. Under these conditions, the rock will fracture by creating vertically oriented fractures which will propagate horizontally. In rocks under horizontal push, or compression, the vertical stress (S_v) is the minimum stress (σ_3), and the rock will fail by the opening of horizontal fractures.

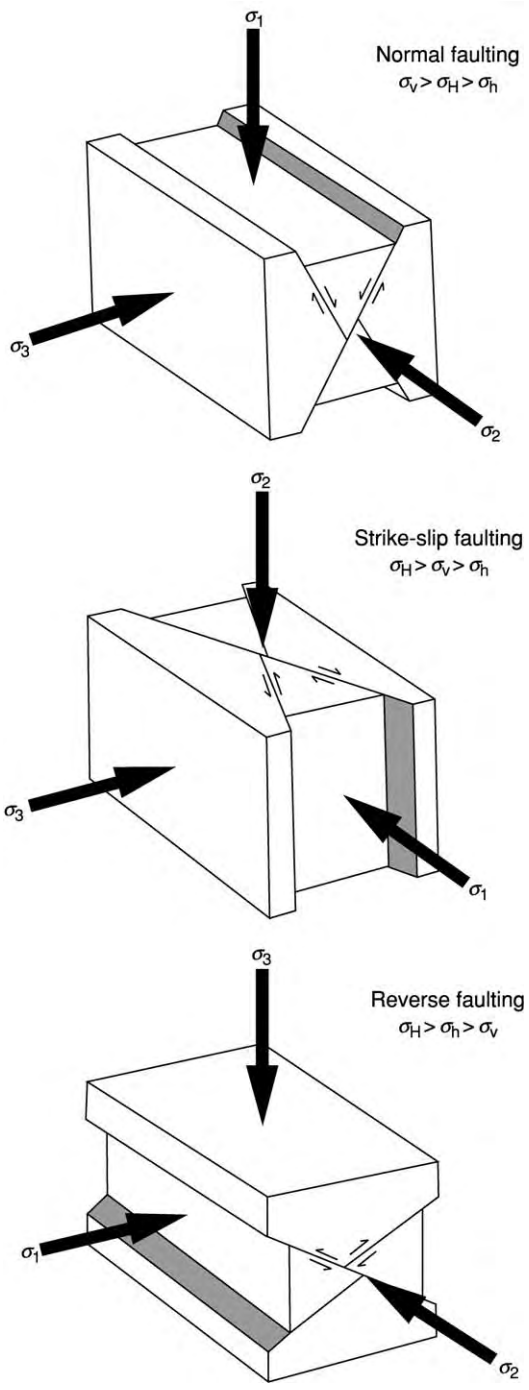


Figure 4 The relative magnitudes of the three principal stresses, one vertical (v) and two horizontal (H, h), acting in a rock mass, and the associated direction of shear, for normal strike slip and reverse faulting regimes. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

Fault

Faults can act as both conduits (migration pathways) and seals, depending on the hydraulic conditions, the rock properties of the faults, and the properties of the

rocks juxtaposed across the faults. The consideration of faults as seals follows the same reasoning as for cap-rock seals above, i.e. the sealing capacity of a fault relates to its membrane strength and hydraulic strength. Membrane fault seals fail when the pressure of the petroleum can exceed the entry pressure of the largest pores along the fault plane. Hydraulic fault seals fail when the fault is opened mechanically by high pore pressure which exceeds the minimum stress.

Reservoir

For a rock to be a petroleum reservoir, it need only be porous, i.e., capable of holding petroleum, and permeable, i.e., able to flow petroleum.

Intrinsic Properties

The following properties must be known or estimated in order for the petroleum volume to be calculated.

1. Net to gross.
2. Porosity.
3. Permeability.
4. Petroleum saturation.

The question regarding whether any discovered petroleum will flow from its reservoir into the well bore is only partially addressed in exploration. This is commonly because permeability estimations are rarely accurate.

Net to gross Net to gross is a measure of the potentially productive part of a reservoir, commonly expressed as the percentage of producible (net) reservoir within the overall (gross) reservoir package (Figure 5). The percentage net reservoir can vary from just a few per cent to 100%. Net pay is used to describe the petroleum-bearing net reservoir.

It is common to define net sand (or limestone) using a permeability cut-off (typically 1 mD for gas and 10 mD for light oil). Such information on permeability is only available when the reservoir has been cored or a petroleum flow test completed. For uncored intervals and uncored wells, a combination of data on lithology and porosity from wireline logs is used. These data are calibrated to permeability data in a cored interval (*see* **Petroleum Geology: Exploration**).

Porosity Porosity is the void space in a rock (Figure 6). It is commonly measured as a volume percentage (*see* **Sedimentary Rocks:** Sandstones, Diagenesis and Porosity Evolution). In the subsurface, this volume may be filled with petroleum, water, a range of non-hydrocarbon gases (CO_2, H_2S, N_2), or some combination of these. Most reservoirs have porosities in the range 20–30%.

Not all pores are alike; there are big pores and little pores, pores with simple shapes, and others with highly complex three-dimensional morphologies. A knowledge of the size and shape of the pores and the way in which they are interconnected is

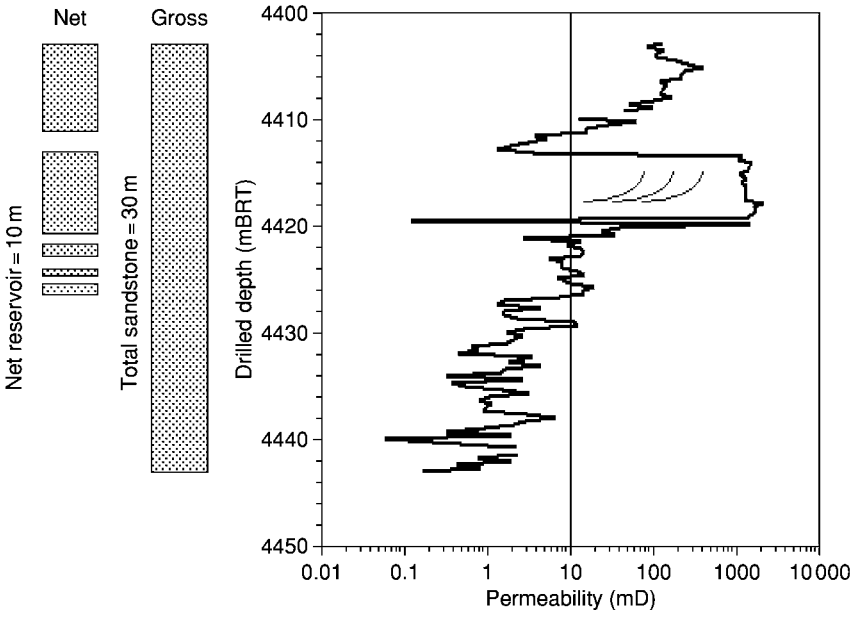


Figure 5 Net to gross is the ratio between reservoir rock capable of flowing petroleum and the gross reservoir interval. It is commonly defined using a single permeability cut off. The example here is a Jurassic oil bearing sandstone from the North Sea. Thirty metres of sandstone was cut in one core, but only 19 m had a permeability greater than the 10 mD cut off chosen, a net to gross of 63%. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

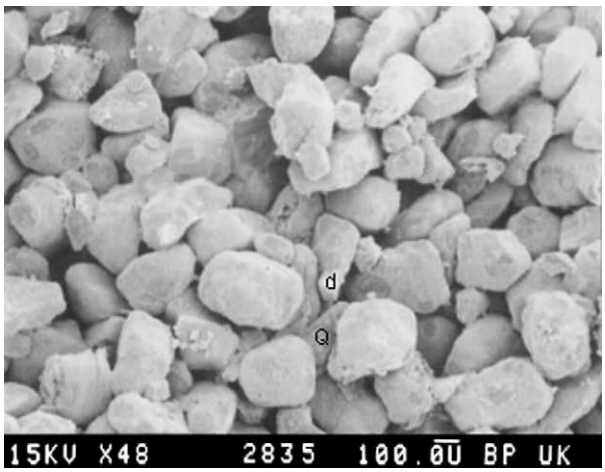


Figure 6 Scanning electron photomicrograph of a porous (28%) and permeable (2200 mD) Permian, Rotliegend reservoir sandstone, southern North Sea. Field of view, 2.7 mm × 1.8 mm. Photograph by A. J. Leonard, reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

important, because it is these factors that determine the permeability of the rock.

For sands and sandstones, a threefold description of porosity is used: intergranular, intragranular, and microporosity (Figure 7). Intergranular porosity occurs between grains. Individual pores in clean sand will occupy approximately 40% of the total volume. For coarse sands, the pores are larger than in fine sands. In most sandstones, the intergranular porosity is primary, a residuum of that imparted at deposition. Some intergranular porosity may be created in sandstones by the dissolution of mineral cements. Most intragranular porosity is secondary in origin, created on partial dissolution of grains. Microporosity simply means small pores, those associated with depositional or diagenetic clay or other microcrystalline cements.

Porosity development in limestones and dolomites is much more variable than that for sandstones. Both rock types are much more prone to mineral dissolution and precipitation than sandstones. This, coupled with the often varied suite of shell and other bioclastic material in the carbonates, makes for a wealth of pore types (Figure 8): intergranular, intragranular, intercrystalline, intracrystalline, biomouldic, vuggy, fracture, cavernous. The size range for pores is also much greater for limestones than for sandstones: from micropores (a few micrometres) in individual oolite grains to giant cave systems.

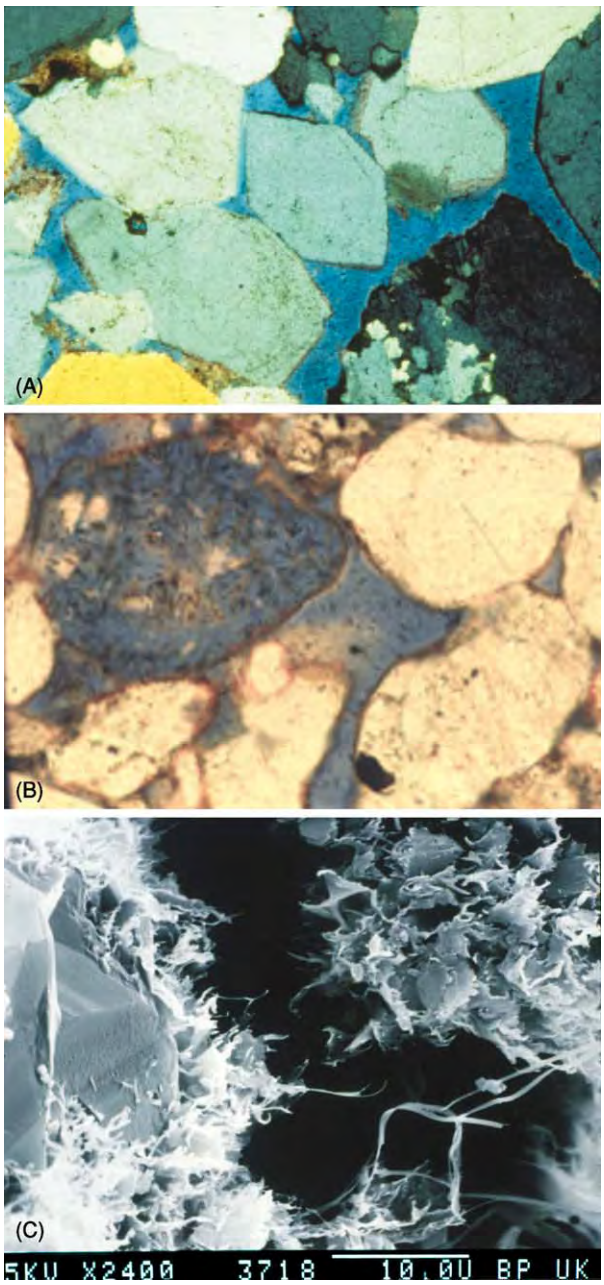


Figure 7 Porosity in sandstones. (A) Intergranular porosity (arrowed) between quartz grains with quartz overgrowths, Jurassic Brent sandstone, northern North Sea. Field of view, 1.3 mm × 0.9 mm. (B) Intragranular porosity within partially dissolved feldspar, Permian Rotliegend sandstone, southern North Sea. Field of view, 650 μm × 450 μm. (C) Microporosity (arrowed) between illitized kaolinite plates, Jurassic Brent sandstone, northern North Sea. Photographs by J. G. Gluyas.

Permeability Permeability is an intrinsic property of a material that determines how easily a fluid can pass through it. In the petroleum industry, the darcy (D) is the standard unit of permeability, but millidarcies (mD) ($1\,\text{mD} = 10^{-3}\,\text{D}$) are more commonly used. A darcy is defined as a flow rate of $10^{-2}\,\text{m s}^{-1}$ for a fluid of 1 cP (centipoise) under a pressure of $10^{-4}\,\text{atm m}^{-2}$. Permeability in reservoir rocks may range from 0.1 mD to more than 10 D. Permeability

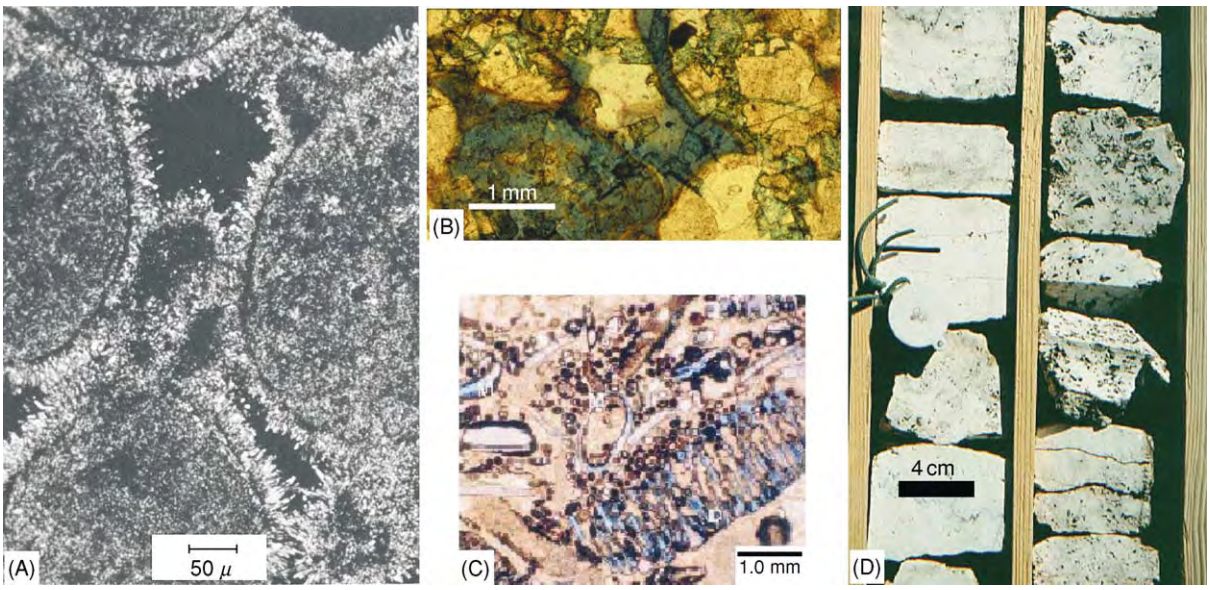


Figure 8 Porosity systems in carbonate reservoirs. (A) Intergranular porosity in limestone, beach rock, Bahamas. Reproduced from Bathurst RGC (1976) *Carbonate Sediments and Their Diagenesis, Developments in Sedimentology 12*. Oxford: Elsevier. (B) Intercrystalline porosity within dolomitized limestone, Permian Zechstein reservoir, southern North Sea, Dutch sector. Field of view, 3.25 mm × 2.50 mm. Photograph by J. G. Gluyas. (C) Biomoldic porosity within algal and mollusc moulds, Pennsylvanian limestone, Texas. Field of view, 5 mm × 4 mm. Reproduced with permission from Dickson JAD and Saller AH (1995) Identification of subaerial exposure surfaces and porosity preservation in Pennsylvanian and Lower Permian shelf limestones, eastern central Basin Platform, Texas. In: Budd DA, Saller AH, and Harris PM (eds.) *Unconformities and Porosity in Carbonate Strata, American Association of Petroleum Geologists Memoir 63*, pp. 239 258. Tulsa, OK: American Association of Petroleum Geologists. (D) Vuggy, oil stained porosity within Cretaceous Bangestan limestones, Zagros Mountains, Iran. Field of view, 10 cm × 8 cm. Photograph by J. G. Gluyas.

measurements made at the Earth's surface are commonly greater than those in the subsurface, and a pressure correction must be made to restore the value of permeability to reservoir conditions. This intrinsic rock property is called the absolute permeability when the rock is 100% saturated with one fluid phase.

Water, oil, and gas saturation It is rare in nature to find a reservoir entirely oil (or gas) saturated. More commonly, the pore system contains both oil and water. The proportions of each phase are commonly expressed as percentages linked to the abbreviations: S_w for water, S_o for oil, and S_g for gas. Water and petroleum saturations are not constant across a reservoir. They vary in response to the position in the oil column, the permeability of the rock, and the mineralogy of the rock. Oil and water saturations will also change as petroleum is produced.

Reservoir Lithologies

Sandstone and limestone (including dolomite) are the most common reservoir lithologies. Sandstones dominate as important reservoirs in the USA (including Alaska), South America, Europe, Russian Asia, north Africa, and Australia. Limestones form the dominant reservoirs in the Middle East. They are also important in the Far East, western Canada, and some of the former Soviet states.

Sandstones, limestones, and dolomites of any age can make fine reservoirs. However, most of the best reservoirs in the world are relatively young. Petroleum fields are more common in Cenozoic and Mesozoic sediments than in Palaeozoic reservoirs. Precambrian age reservoirs are rare. There is no intrinsic reason why old rocks are more or less likely to be reservoirs than younger ones; it is simply that older reservoirs have had greater chance to be involved in tectonism or cementation, so destroying their porosity or permeability.

In addition to sandstone and limestone, fractured rock of any type can form a reservoir. The fractures alone may form the total pore volume of the reservoir. Alternatively, the fractures may help drain petroleum from the intervening lower permeability rock.

Reservoir; sandstone depositional systems Sediments, including those which may one day form a petroleum reservoir; can accumulate in many environments on the Earth's surface (*see* **Sedimentary Environments:** Depositional Systems and Facies). This includes sands deposited both on land and beneath the sea (Table 1). The overall architecture and internal

Table 1 Clastic reservoirs

Depositional system	*Architectural elements*	*Size range*	*Reservoir properties*	*Example oil/gas field(s) or province*
Alluvial fan	Low angle half cones, linear and sheet sand bodies	1 10 km diameter	Heterogeneous, poorly sorted	Quiriquire Field, Pliocene E. Venezuela, 750 mmbbl
Aeolian deposit	Dune, sand sheet	100s km^2	Dune well sorted porous sands high quality sandsheet moderate quality	Permian Rotliegend, Europe; Jurassic Norphlet, US Gulf Coast
Lake deposit	Half cone (fan)	Few km diameter	Poor to good, function of sediment input	Thailand & China
Fluvial system	Channel fill, crevasse splay	Channel belts 10s km × few km, crevasse splay few km diameter	Channel fills in braided and meandering systems commonly good; braided net to gross > meandering net to gross	Prudhoe Bay Field, Triassic Alaska, >10 bn bbl; Wytch Farm Field, Triassic UK onshore, 300 mmbbl
Delta and coastline	Channel mouthbar, shoreface, beach	**Figure 9**	Commonly good in reworked sandstones, variable net to gross	Niger Delta, W. Africa; Brent system, North Sea; Mahakan Delta, Indonesia; multiple billion barrels fields
Shallow marine	Shoreface to offshore sandstone bars	**Figure 9**	Good to excellent, high net to gross	Shannon Sandstone, Cretaceous USA; Fulmar/Ula Sandstones, Jurassic UK and Norwegian North Sea; Toro Sandstone, Jurassic/Cretaceous Papua New Guinea
Deep marine	Fan lobe, fan channel	10 100s km long, 10s km across	Depends upon sediment source area	Tertiary formations, North Sea; Plio Pleistocene US Gulf Coast; Tertiary Congo Fan

bn bbl, billion barrels; mmbbl, million barrels.

geometry of the sand bodies (**Figure 9**) control the performance of a reservoir during petroleum production (*see* **Petroleum Geology:** Production).

Reservoir; carbonate depositional systems Limestones and dolomites form some of the largest petroleum reservoirs in the world. Many of the largest occur in the Middle East. Other areas in which carbonate reservoirs deliver large quantities of oil and gas are western Canada, Mexico, Texas (USA), Norway (central North Sea), Poland, Kazakhstan, western and south-eastern China, Iran, and Libya.

The sediment that forms most carbonate reservoirs accumulated in shallow marine environments (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves). The important exceptions are pelagic chalks (Ekofisk Complex of the North Sea) and deep-water resedimented carbonates of the Poza Rica Trend in Mexico.

Like their clastic counterparts, there is a clear link between the reservoir potential of a carbonate body and the environment in which the host sediment accumulated. High-energy ooid and shell shoals can make excellent reservoirs. Framework reef complexes are also prime reservoir targets. However, unlike siliciclastics, carbonates can undergo almost complete transformation on weathering and/or diagenesis to produce reservoirs from former seals and seals from former potential reservoirs.

Dolomite (*see* **Sedimentary Rocks:** Dolomites) Producing dolomites range in age from Precambrian to Tertiary. It is estimated that about 80% of the recoverable petroleum in carbonate-hosted reservoirs of the USA occurs in dolomite and only about 20% in limestone. The same ratio probably applies to the producible reservoirs in the Permian Zechstein of Europe, whilst older carbonate plays in Europe and Russian Asia are almost wholly dolomite. The dolomitization of limestones commonly leads to an increase in both porosity and permeability (**Figure 10**).

Karst (*see* **Sedimentary Processes:** Karst and Palaeokarst) Karstified limestones and dolomites represent the second major group of carbonate reservoirs not directly linked to depositional environments. Karst is a product of mineral dissolution (**Figure 11**). It develops where carbonates are exposed to meteoric water, often linked with episodes of sea-level fall. Karst features are well known to geologists and geographers alike: caves, collapse breccias, dissolution-enhanced joints and fractures, and vugs.

Fields producing from karstified limestones and dolomites include the Liuhua Field in the South China

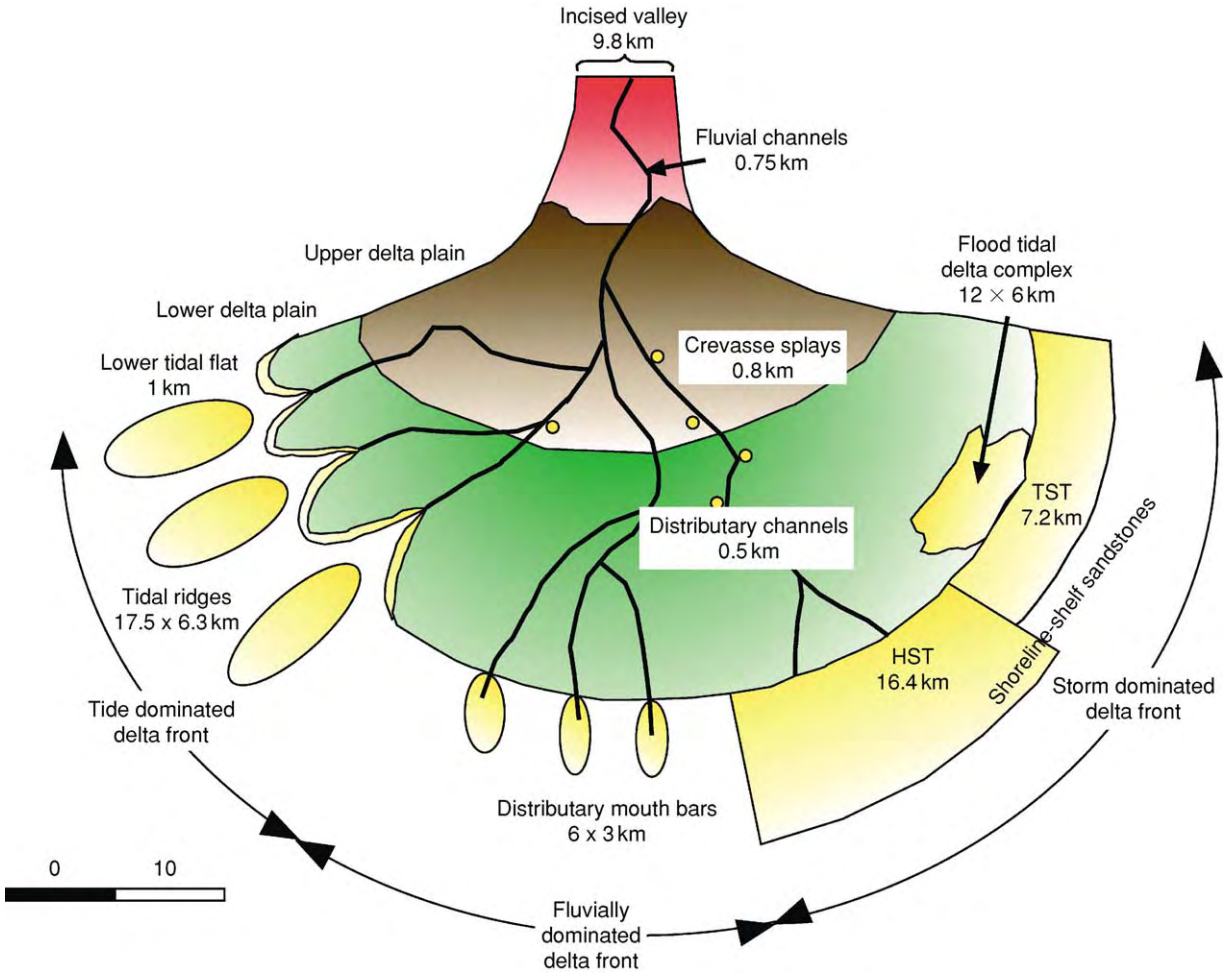


Figure 9 The average size, shape, and location of sand bodies in wave, tidal, and fluvially influenced reservoirs. Reproduced from Reynolds AD (1994) Sequence stratigraphy and the dimensions of paralic sandstone bodies. In: Johnson SD (ed.) *High Resolution Sequence Stratigraphy: Innovations and Applications*, pp. 69 72. Liverpool: Liverpool University.

Sea, the Permian reservoirs of Texas and New Mexico, and parts of the Upper Permian in the Zechstein Basin in Europe. Thermal karst, produced when hot fluids dissolve limestones at depth, may also become reservoirs. The Albion Scipio Field of Michigan is of this type.

Trap

Trap is the term to describe the body, bounded by seals and containing reservoir, in which petroleum can accumulate as it migrates from the source rock to the Earth's surface. There are many different trap geometries. These can be grouped into three categories: structural, stratigraphical, and hydrodynamic (Table 2). Structural traps are created by tectonic, diapiric, compactional, and gravitational processes (Figure 12). Almost the entire world's discovered petroleum is in traps that are largely structural. Stratigraphical traps are

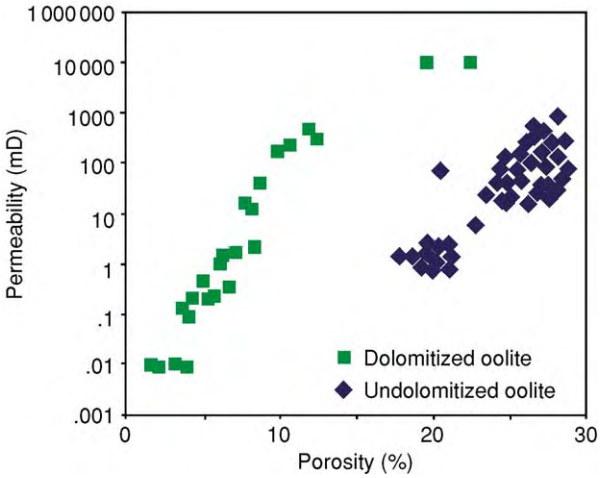


Figure 10 A comparison of porosity and permeability for dolomitized and undolomitized oolite, Cretaceous Middle East. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

formed by lithological variations or property variations generated by alteration of the sediment or fluid through diagenesis (Figure 13). Much of the world's remaining undiscovered petroleum will be found in stratigraphical traps. Purely hydrodynamic traps are rare. Such traps rely on the flow of water through the reservoir horizon to 'drag' the petroleum into a favourable trapping configuration, such as the plunging nose of a fold. The trapping mechanism for many fields is commonly a combination of structural and stratigraphical elements or, more rarely, structural elements and hydrodynamic conditions.

Figure 11 Tower karst containing fracture and cavernous porosity, Palaeozoic limestones, Zhaoquing, Guangdong Province, China. Photograph by J. G. Gluyas.

Structural Trap

Compressive tectonic regimes commonly lead to the development of large-scale contractional folds and thrusts. This is common at convergent plate boundaries and transpressional strike-slip plate boundaries (wrench systems). The El-Furrial Trend of eastern Venezuela is an example of such a system. The anticlinal traps of the trend were developed during convergence of the Caribbean and South American plates during the Neogene. Many of the traps are large ramp anticlines (Figure 14). They have oil columns of, on average, 400 m and reservoirs formed from high net to gross shallow marine sandstones.

In North America, thrust-linked rollover anticlines form the major trap type in the Wyoming–Utah thrust belt fields and the southern foothills of the Alberta Basin, Canada. Compressional anticlines also form giant traps within the Zagros fold belt of Iran.

Traps formed by extensional tectonics are common in rift basins. The East Shetland Basin of both the UK and Norwegian sectors of the North Sea contained about 15 billion barrels of recoverable oil. Much of this oil was trapped in tilted fault blocks formed during Late Jurassic rifting. In the pre-rift section, oil is reservoired in the sandstones of the Middle Jurassic Brent Group, together with other sandstones of both Jurassic and Triassic age (Figure 15). Traps formed through tectonic extension are also important in the Gulf of Suez, the Haltenbanken area, offshore mid-Norway, and in the pre-rift sections of the Gippsland Basin (Australia).

Traps can also be formed by diapiric processes. The specific gravity of salt (halite) is about 2.2 g cm ^{3} and that of fully consolidated rock is about 2.5–2.7 g cm ^{3}. Thus salt is buoyant relative to most other sediments

Table 2 Structural and stratigraphical traps

Structural	Tectonic	Extensional
		Compressional
	Diapiric	Salt movement
		Mud movement
	Compactional	Drape structures
	Gravitational	Listric fault movement
Stratigraphical	Depositional	Pinchouts (dunes, bars, reefs, channels, etc.)
		Unconformities (erosional, subcrop, karst, etc.)
	Diagenetic	Mineral precipitation
		Mineral dissolution (thermal karst, dolomitization)
		Tar mats
		Permafrost
		Gas hydrate crystallization

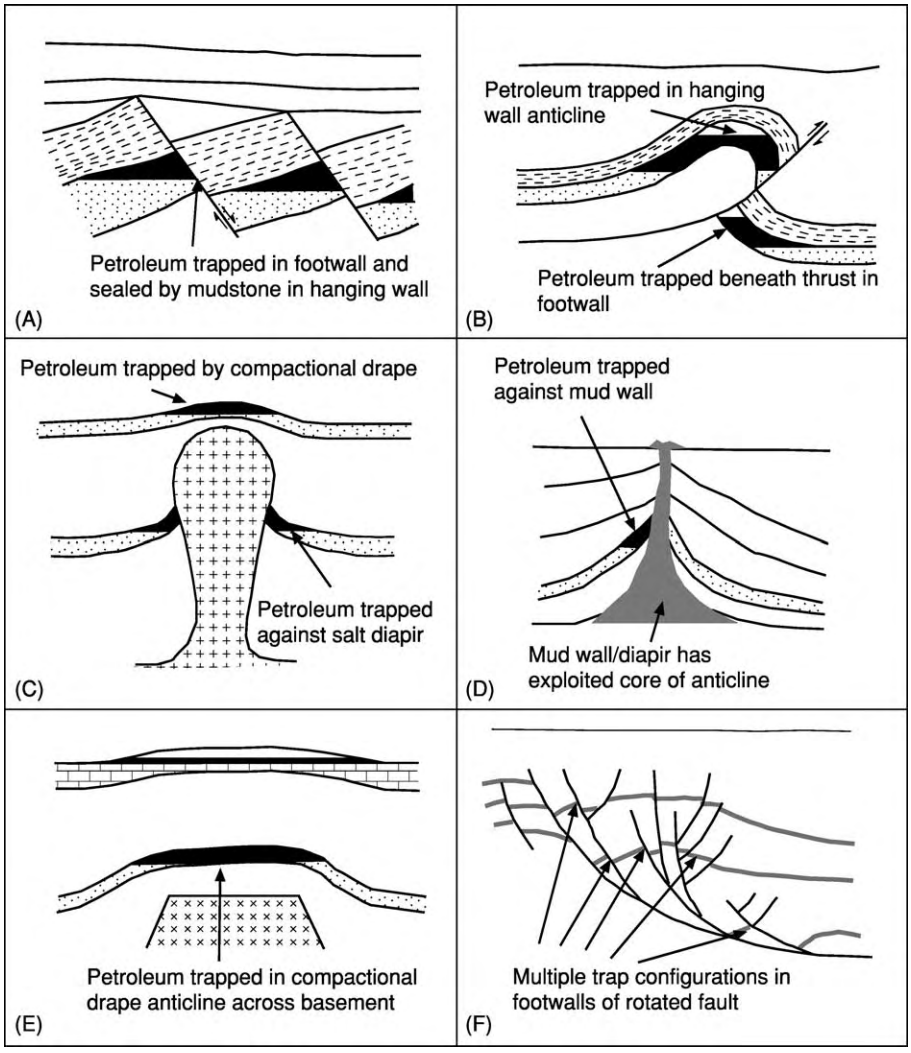


Figure 12 Structural traps. (A) Tilted fault blocks in an extensional regime. The seals are overlying mudstones and cross fault juxtaposition against mudstones. (B) Rollover anticline on thrust. Petroleum accumulations may occur on both the hanging wall and the footwall. The hanging wall accumulation is dependent on a subthrust fault seal, whereas at least part of the hanging wall trap is likely to be a simple, four way, dip closed structure. (C) Lateral seal of a trap against a salt diapir and compactional drape trap over the diapir crest. (D) Diapiric mudstone associated trap with lateral seal against mud wall. Diapiric mud associated traps share many common features with that of salt. In this diagram, the diapiric mud wall developed at the core of a compressional fold. (E) Compactional drape over a basement block commonly creates enormous low relief traps. (F) Gravity generated trapping commonly occurs in deltaic sequences. Sediment loading causes gravity driven failure and produces convex down (listric) faults. The hanging wall of the fault rotates, creating space for sediment accumulation adjacent to the fault planes. The marker beds (grey) illustrate the form of the structure that has many favourable sites for petroleum accumulation. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

and sedimentary rocks. Over geological time, salt deforms plastically. With loading caused by continued sedimentation, layers of salt may aggregate into swells and eventually pillows. Subsequently, a salt diapir may rise through the overburden. Very similar processes to those associated with salt diapirism can occur in association with muds. Rapidly deposited muds are commonly water rich, overpressured, and, in consequence, highly mobile. Mud lumps (Niger Delta), shale walls, diapirs, and mud volcanoes (Trinidad, Azerbaijan) are all products of mass mud movement.

Diapiric movement of both salt and mud can create anticlinal structures that could form petroleum traps. Trap configurations can also develop in the areas of salt withdrawal. The 'turtle' structure anticline develops via increased sedimentation in areas of salt withdrawal. Later, as salt continues to feed the diapir, the structure flounders and flips into an anticline.

Greater Burgan (Kuwait), the second largest oilfield in the world (>75 billion barrels of reserves), developed over a large, low-amplitude salt swell.

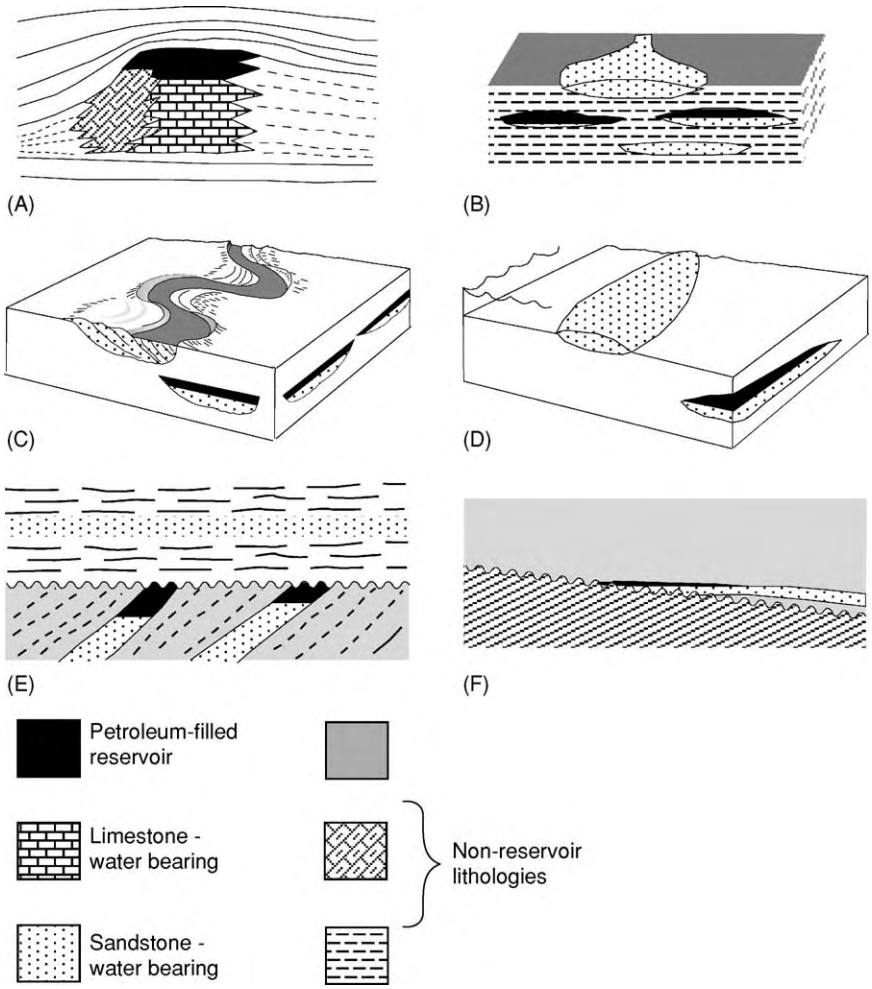


Figure 13 Stratigraphical traps. (A) 'Reef' oil is trapped in the core of the reef, with fore reef talus and back reef lagoonal muds acting as lateral seals and basinal mudstones as top seals. (B) Pinchout (sandstone) trap within stacked submarine fan sandstones. The upper surface of the diagram shows the plan geometry of a simple fan lobe. Lateral, bottom, and top seals are the surrounding basinal mudstones. (C) Channel fill sandstone trap. The oil occurs in ribbon shaped sandstone bodies. The top surface of the diagram shows the depositional geometry of the sandstone. Total seal may be provided by interdistributary mudstones or a combination of these and marine flooding surfaces. (D) Shallow marine sandstone bar completely encased in shallow marine mudstone. The upper surface of the diagram shows the prolate bar. (E) Subunconformity trap. The reservoir horizon is truncated at its up dip end by an unconformity and the sediments overlying the unconformity provide the top seal. Lateral and bottom seals, like the reservoir interval, pre date the unconformity. (F) Onlap trap. A basal or near basal sandstone onlaps a tilted unconformity. The sandstone pinches out on the unconformity and is overstepped by a top seal mudstone. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

Similar, simple anticlinal dome traps typify the Cretaceous Chalk fields of the Norwegian North Sea. As with the Middle East examples, the key controlling structures are the underlying salt pillows.

Traps associated with diapirs rather than swells tend to be much smaller in aerial extent than the giants described above. They also tend to be much more structurally complex, commonly containing both radial and concentric fault patterns. The Machar Field (STOOIP (standard barrels of oil originally in place) about 228 mmstb) of the North Sea is roughly circular in outline with a diameter of about 4 km (Figure 16).

Anticlinal traps created through compaction occur across basement highs, tilted fault blocks, carbonate shelf rims, reefs, or isolated sand bodies. Some of the simplest are also the largest. The world's biggest field, Ghawar in Saudia Arabia, is such a trap. Oil occurs in Jurassic carbonates draped over and compacted around a north–south-trending basement high.

Traps formed by gravity-driven processes are particularly important in large recent deltas. The best-described examples are from the US Gulf Coast and West African deltas (Niger, Congo). The gravity structures form independently of basement tectonics and owe their existence to shallow detachment along

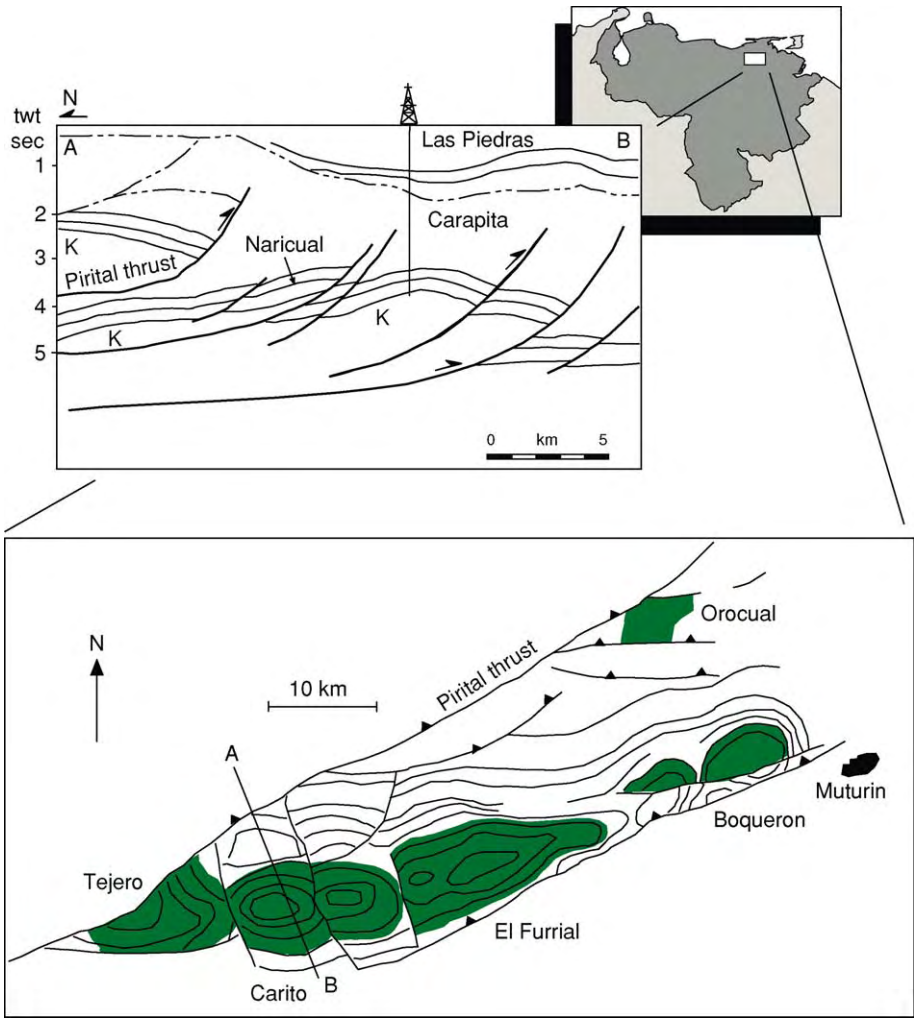


Figure 14 El Furrial Trend, eastern Venezuela. The petroleum traps are large rollover anticlines on the hanging walls of thrusts. Most are four way, dip closed structures, whilst some have a dependence on a fault component along their south eastern thrust margin. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

low-angle, basinward-dipping planes. The drive mechanism is provided by the weight of sediment deposited by the delta at the shelf–slope break or on the slope itself. In the Niger Delta, the detachment planes are highly mobile muds, whereas, in the Gulf Coast (Mississippi Delta), detachment occurs on both muds and the Louanne Salt (Jurassic). The key detachment surfaces are commonly listric, concave-up, and concave-basinward in plan view. The main faults are commonly large, being tens of kilometres tip to tip.

Stratigraphical Trap

From top to bottom of a systems tract, each depositional environment is capable of producing a juxtaposition of permeable and impermeable sediments which might one day form a stratigraphical trap for petroleum. In practice, the reservoir geometry becomes the trap geometry. Examples include aeolian dunes encased in lacustrine mudstone, sand-filled fluvial channels cut into mud-rich overbank deposits, shallow marine bar sandstones surrounded by marine shales, carbonate reefs isolated by enclosing marls, and submarine fan sands trapped within the domain of pelagic mud.

The Paradox Basin (Colorado and Utah, USA) contains a large array of small oil and gas fields in stratigraphical pinchout traps. Devonian reservoirs occur within shallow marine bar sandstones and Carboniferous reservoirs within carbonate mounds. The Paradox Basin traps are difficult to find, but have relatively simple shapes. Their geometries are either prolate bar forms or more equidimensional carbonate mounds. Pinchout traps formed in deltaic settings are often much more complex in outline and, because

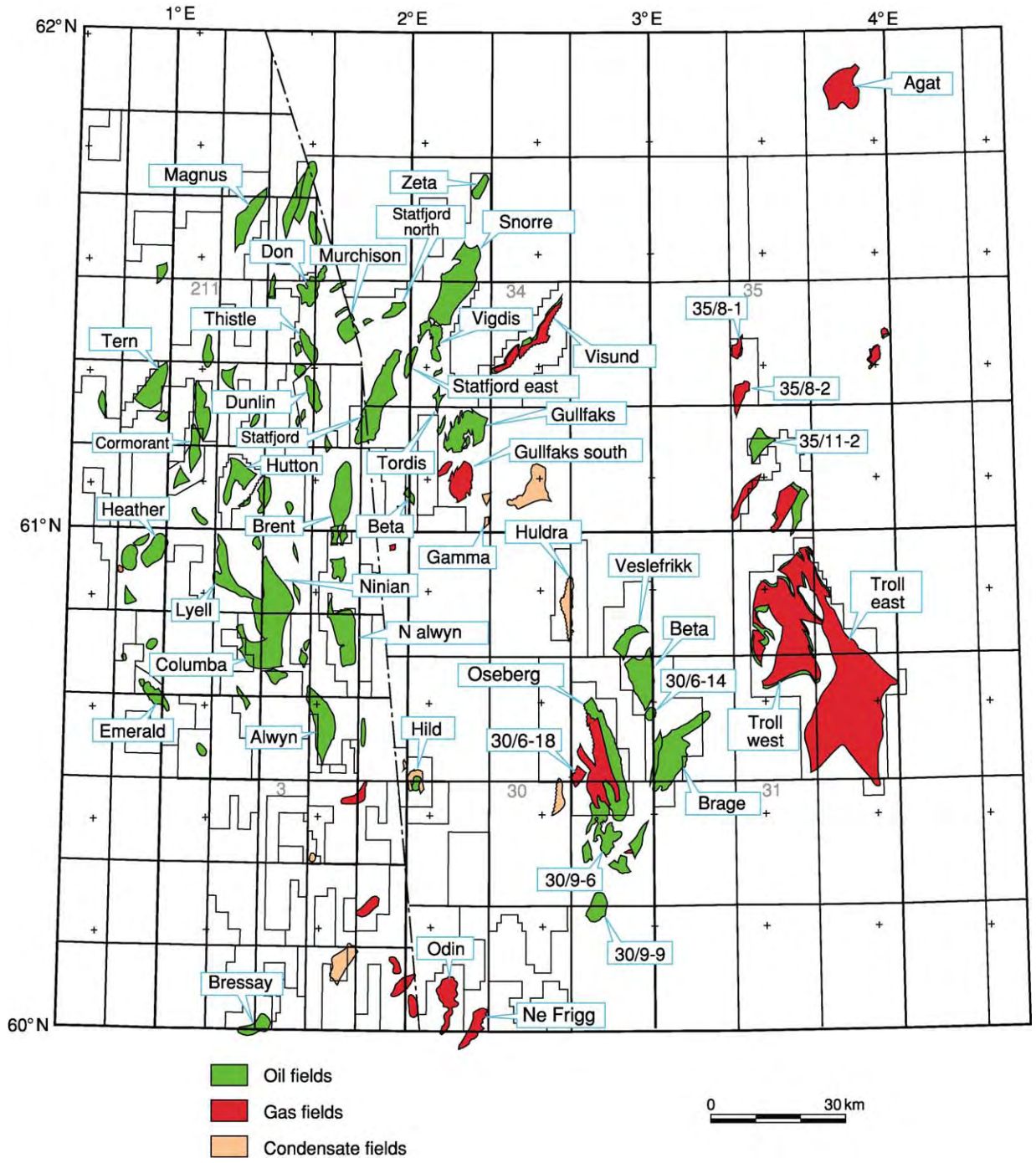


Figure 15 UK and Norwegian Brent Province. The elongate shapes reflect the geometry of the tilted fault blocks that form the traps. The reservoir is largely Middle Jurassic Brent sandstones together with Triassic and Upper Jurassic sandstones in some fields. The traps for neither Troll (Upper Jurassic reservoir in low relief anticline) nor Agat (Lower Cretaceous reservoir, stratigraphically trapped) are tilted fault blocks. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

potential reservoir sandstones are commonly discontinuous, multiple pools (clustered fields) are common (Figure 17).

Attenuation of the up-dip portions of a potential reservoir interval by an unconformity can create massive traps with enormous petroleum catchment (drainage) areas. The largest oilfield in North America, Alaska's Prudhoe Bay, is an unconformity trap. It has about 25 billion barrels of liquid and more than 20 trillion cubic feet of gas in place. East Texas, the

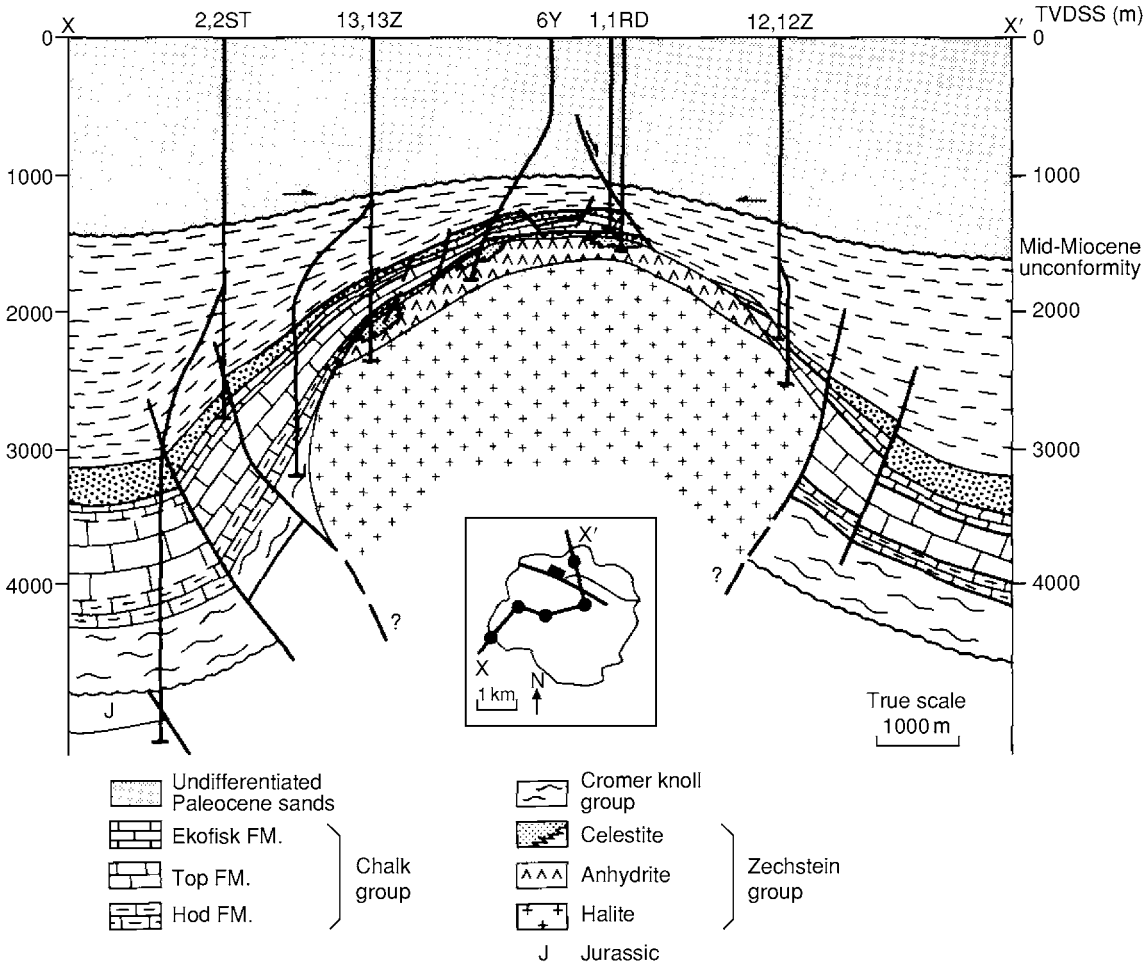


Figure 16 A structural cross section of the Machar Field, central North Sea, showing the circular outline of the field and the distribution of reservoirs around the head of the salt diapir. Reproduced from Foster PT and Rattey PR (1993) The evolution of a fractured chalk reservoir: Machar Oilfield, UK North Sea. In: Parker JR (ed.) *Petroleum Geology of Northwest Europe: Proceedings of the 4th Conference*, pp. 1445 1452. London: Geological Society.

largest oilfield in the USA Lower 48, is also a stratigraphical trap. The productive Woodbine Sandstone reservoir, with its initial reserves of about 6.8 billion barrels, is sandwiched between two unconformities. The sand rests upon the Washita Group mudstones and is itself truncated beneath the Austin Chalk. The field, some 40 miles long and 5 miles wide, is a simple homoclinal dip to the west.

Each of the unconformity traps described above relies on a combination of trapping mechanisms, which rely in large part on a planar or gently folded unconformity. Unconformities come in a variety of shapes. The most spectacular of the unconformity-bounded traps are those commonly referred to as 'buried hills'. Such hills are residual topography of a one-time land surface. Thus it is the unconformity surface that has the trapping geometry (Figure 18). Buried hill traps are most common in karstified areas, such as northern China.

Mineral cements are known to form top, lateral, and even bottom seals to reservoirs. Examples in carbonate systems are more numerous than those in clastic systems. In the Albion-Scipio Field of Michigan (USA), all surrounding rock to the trap is thoroughly cemented limestone and dolomite. A comparable situation exists for many of the carbonate-hosted oilfields of Abu Dhabi; porosity only exists where there is oil. Areas that at one time must have been the aquifers to the oilfields have been thoroughly cemented. For a few fields, such cementation has allowed trap integrity to be maintained despite tilting of the field after petroleum accumulated.

Tar mat seals are common in the shallow subsurface. They also act as cap rock for the largest single accumulation of heavy (viscous) oil in the world; the Faja of south-eastern Venezuela, which has about 1.2 trillion barrels of oil in place. Tar seals and tar sands

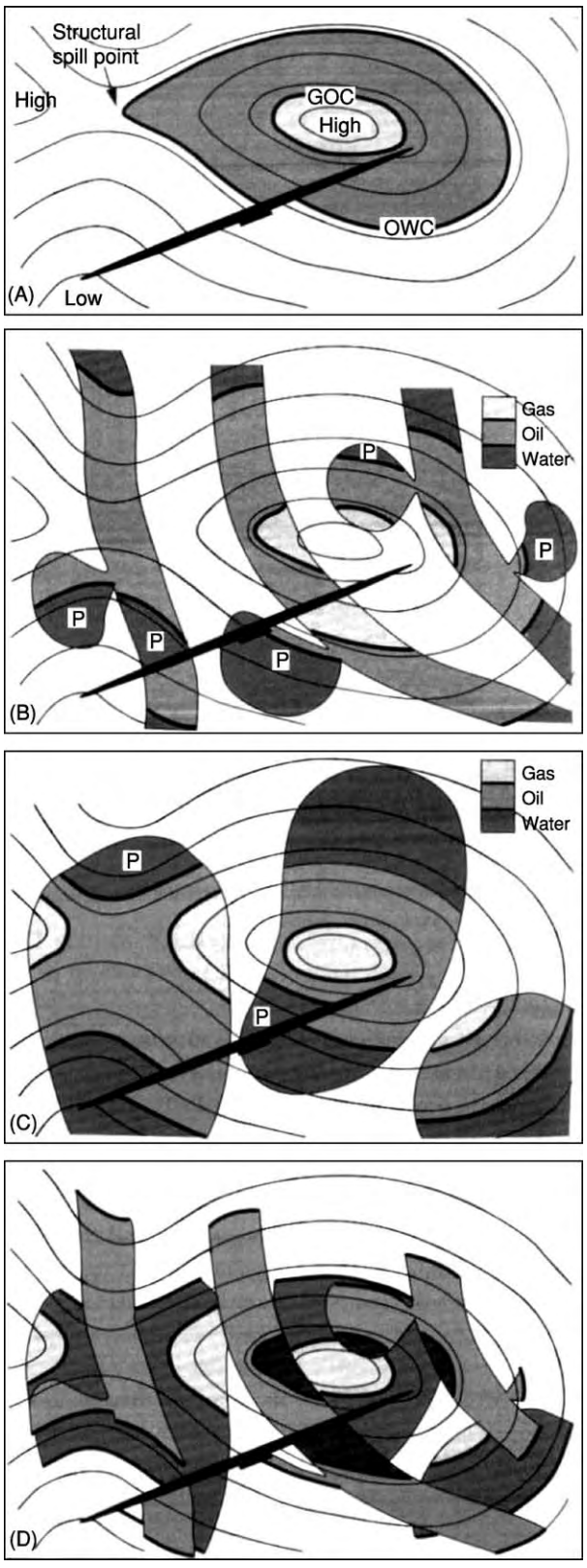


Figure 17 Paralic field outlines commonly have complex shapes because of the interaction between structure and sedi ment bodies. This complexity is multiplied because individual paralic sandstones tend to be stacked. The four examples show: (A) field shape on a simple faulted anticline for which the reservoir interval is much larger than the anticline; (B) the same structure as in (A), but with the reservoirs developed in channel and crevasse splay sandstones that are smaller in area than the structure; (C) the same structure as in (A), but with mouthbar sandstones which are also smaller than the structure; (D) a combination of channel and mouthbar sandstones at different levels. A. Reynolds, personal communication, 1994. Reproduced courtesy of BP.

are also common within the Western Canada Basin and Californian basins.

Gas trapped beneath permafrost forms large fields in the northern part of the West Siberia Basin, adjacent to the Kara Sea. In cold regions, gas (methane) is also trapped as gas hydrate.

Hydrodynamic Trap

The idea that moving water could and would control the distribution of both oil and gas traps was first advocated in 1909. The hypothesis had a number of supporters until the 1930s, when the number of publications on the topic dwindled and the anticlinal theory of petroleum accumulation reassumed its position as the only favoured theory. Twenty years later, the idea was resurrected, although it remains controversial.

Those traps with undoubted hydrodynamic credentials tend to be in foreland basins where subsurface reservoir units commonly crop out in adjacent mountain belts. The outcropping reservoir units are recharged with meteoric water and the hydraulic head drives the flow through the basin. Two of the best-documented examples are the Frannie Field of the Big Horn Basin, Wyoming, and the East Colinga Extension Field of San Joaquin Valley, California. In both instances, there is sufficient information to map the tilted oil–water contacts, rule out the possibility of significant permeability barriers in the systems, and explain the water flow in terms of the adjacent topography and subsurface structure.

Migration

Migration is the process (or processes) whereby petroleum moves from its place of origin, the source rock, to its destruction at the Earth's surface. Along the route, the petroleum's progress may be temporarily arrested and the petroleum may accumulate within a trap. The timing of trap formation relative to that of petroleum generation and migration is critical. The trap has to form at the same time or earlier than petroleum migration if it is to capture petroleum.

Migration may be divided into three stages (Figure 19).

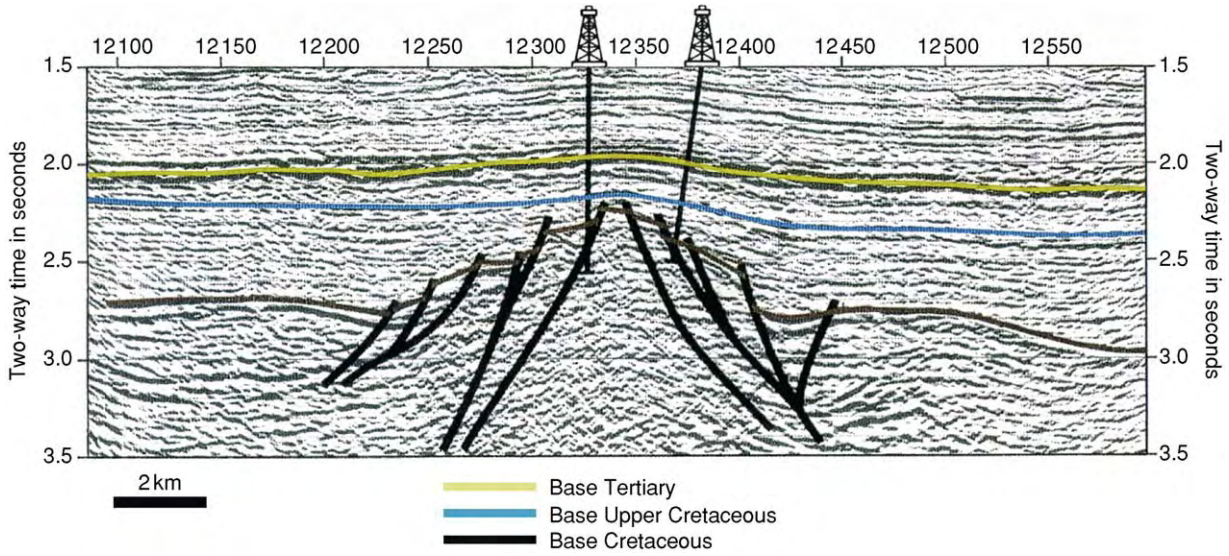


Figure 18 Subunconformity trap beneath the base Cretaceous unconformity, Buchan Field, UK North Sea (fractured Devonian sandstone reservoir). Reproduced from Abbots IL (1991) *United Kingdom Oil and Gas Fields, 25 Years Commemorative Volume, Geological Society Memoir No. 14*. London: Geological Society.

- Primary migration: expulsion of petroleum from the source rock.
- Secondary migration: the journey from source rock to trap.
- Tertiary migration: leakage and dissipation of the petroleum at the Earth's surface.

Primary Migration

There have been many hypotheses created to explain the migration of petroleum out of the source rock. Most researchers now favour processes whereby petroleum is expelled from the source rock as a separate phase within a water-wet rock matrix.

Analyses have been performed on a source rock (Kimmeridge Clay, North Sea) which is actively expelling petroleum. The aim was to elucidate the precise primary migration mechanisms. The analytical results could best be explained by invoking pressure-driven flow of a petroleum-rich phase as the main expulsion mechanism for source rocks. Specifically, it was demonstrated that petroleum was first expelled when the volume of generated petroleum approximately matched the volume of pore space within the mudstone. That is, the mudstone was almost fully saturated with petroleum before expulsion occurred. This supported earlier observations on lean source rocks. Those which yield less than 5 kg petroleum per tonne tend not to achieve sufficient saturation for expulsion to occur.

Gas expulsion may occur in a similar fashion to that of oil, albeit at higher temperatures. Clearly, the volume increase associated with gas generation is massive, be it directly from kerogen or from the thermal decomposition of previously formed oil. Pressure-driven expulsion will occur either through

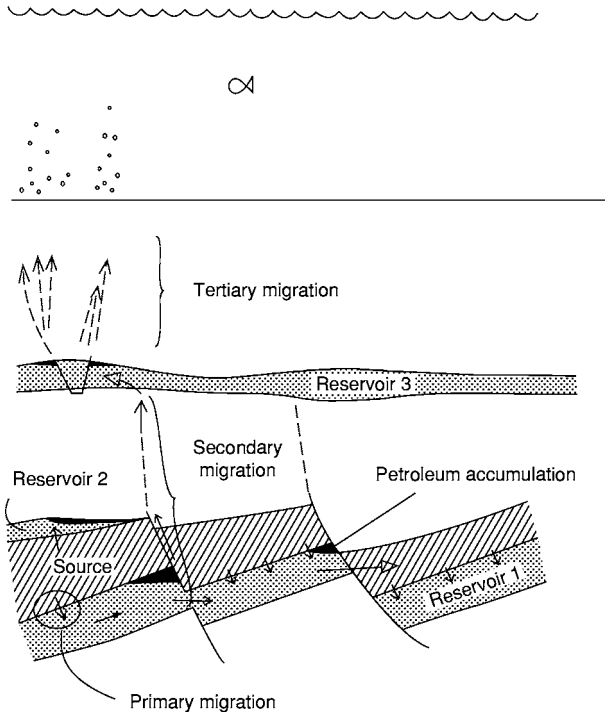


Figure 19 Diagram of the three stages of migration. Primary migration out of the source rock and into a trapped reservoir (reservoir 2) or a carrier bed (reservoir 1). Secondary migration in carrier reservoir 2 and up faults into reservoir 3. Tertiary migration (dissipation) from reservoir 3 to the surface. Reproduced from Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

the existing pore network or through induced fractures. During gas generation, previously generated, short-chain liquid hydrocarbons may become dissolved in the gas and expelled with it. This mechanism has been used to explain the production of condensate from Type III kerogen in overpressured mudstone.

Secondary Migration

Secondary migration takes petroleum from the source location to trap or traps via carrier beds. The defining aspect of secondary migration is that it concentrates or focuses the petroleum. On escape from the source rock, petroleum is dispersed over a large area. By the time petroleum reaches the relatively restricted area of a trap, it can occupy more than 90% of the pore volume in the reservoir. Secondary migration is temporarily arrested once the migrating petroleum enters a trap. Disruption of the trap or overfilling of the trap can lead to remigration of the petroleum to a higher structural level under the same secondary migration process. Such secondary migration ends when petroleum approaches the Earth's surface.

The medium through which the petroleum travels during secondary migration is also quite different from that of the source rock. The pore size and thus permeability in a carrier bed, be it a sandstone, carbonate, or fractured lithology, is much larger than that in a source rock. The driving mechanism for secondary migration is the density difference between the petroleum (less dense) and water (more dense). The density difference is expressed through the buoyancy force generated by the pressure difference between a point in a continuous petroleum column and the adjacent pore water.

The restricting force to petroleum migration is the capillary injection pressure. A slug of petroleum migrates from pore to pore in a carrier bed, squeezing through the intervening pore throats. The force required to move petroleum through a pore throat is a function of the radius of the pore throat, the interfacial tension between the petroleum and the water, and the wettability of the rock–petroleum–water system.

The buoyancy effect means that petroleum will tend to rise within the sediment column. The capillary effect dictates that, in the absence of other forces, petroleum will migrate from small pores to large pores. Furthermore, petroleum (and water) will attempt to equilibrate with respect to pressure. That is, flow can be induced by pressure differential (either overpressure or hydrodynamics).

It is possible to estimate the likely migration directions from source bed to reservoirs by mapping the orthocontours of the likely carrier systems (Figure 20). Orthocontours are simply lines constructed on a map at right angles to the contours. Instead of displaying areas of equal height (or depth), they depict lines of maximum dip. The buoyancy effect dictates that the rising petroleum will follow such orthocontours. Clearly, such an exercise must be attempted on the geometry of the carrier bed(s) as it was during the phase of petroleum migration. This clearly leads to attempts to reconstruct the basin history in terms of deposition, structuring, and source rock maturation.

The capillary effect controls how much of a carrier bed becomes petroleum saturated. Rarely are carrier beds of a uniform grain size distribution. Thus, petroleum will tend to migrate along the coarsest, high-permeability pathways (Figure 21). These may occupy 10% or less of any particular formation. Open fractures have the same effect as coarse beds. Petroleum will exploit them. Temporarily open fracture systems are commonly invoked as the mechanism whereby migrating petroleum 'jumps' upward in the stratigraphy of a particular basin.

The rate at which petroleum migrates can be calculated using Darcy's law

$$q = -(k/\mu)(\mathrm{d}\theta/\mathrm{d}z)$$

where q is the volume of flow rate ($m^3\,m^{-2}\,s^{-1}$), k is the permeability, μ is the viscosity (Pa s^{-1}), and $\mathrm{d}\theta/\mathrm{d}z$ is the fluid potential gradient.

Typical permeability values are: sandstones, 10^{-12}–$10^{-15}\,m^2$ (1 D to 1 mD); limestones, 10^{-14}–$10^{-17}\,m^2$ (10 mD to 10 μD).

From these data, it is possible to calculate that the migration rate for petroleum in sandstone will be 1–1000 km per million years and, in limestone, 0.01–10 km per million years.

Phase changes will occur in petroleum as a result of its migration upwards to regions of lower pressure and temperature. This is most important for high-temperature, high-pressure condensates, but any oil will exsolve some gas if the pressure in the formation drops below the bubble point. The residual petroleum and generated gas are then likely to behave differently with respect to subsequent migration.

At low temperatures (<70°C), and in regions in which there is significant water flow, petroleum may be degraded by bacterial action or by water washing. The bacterial process follows a systematic loss of the *n*-alkanes, branched alkanes, isoprenoids, alkylcyclohexanes, and polycyclic alkanes. This progressive destruction of the petroleum leads to increases in the pour point and viscosity of the oil and a lowering of the API (American Petroleum Institute) gravity.

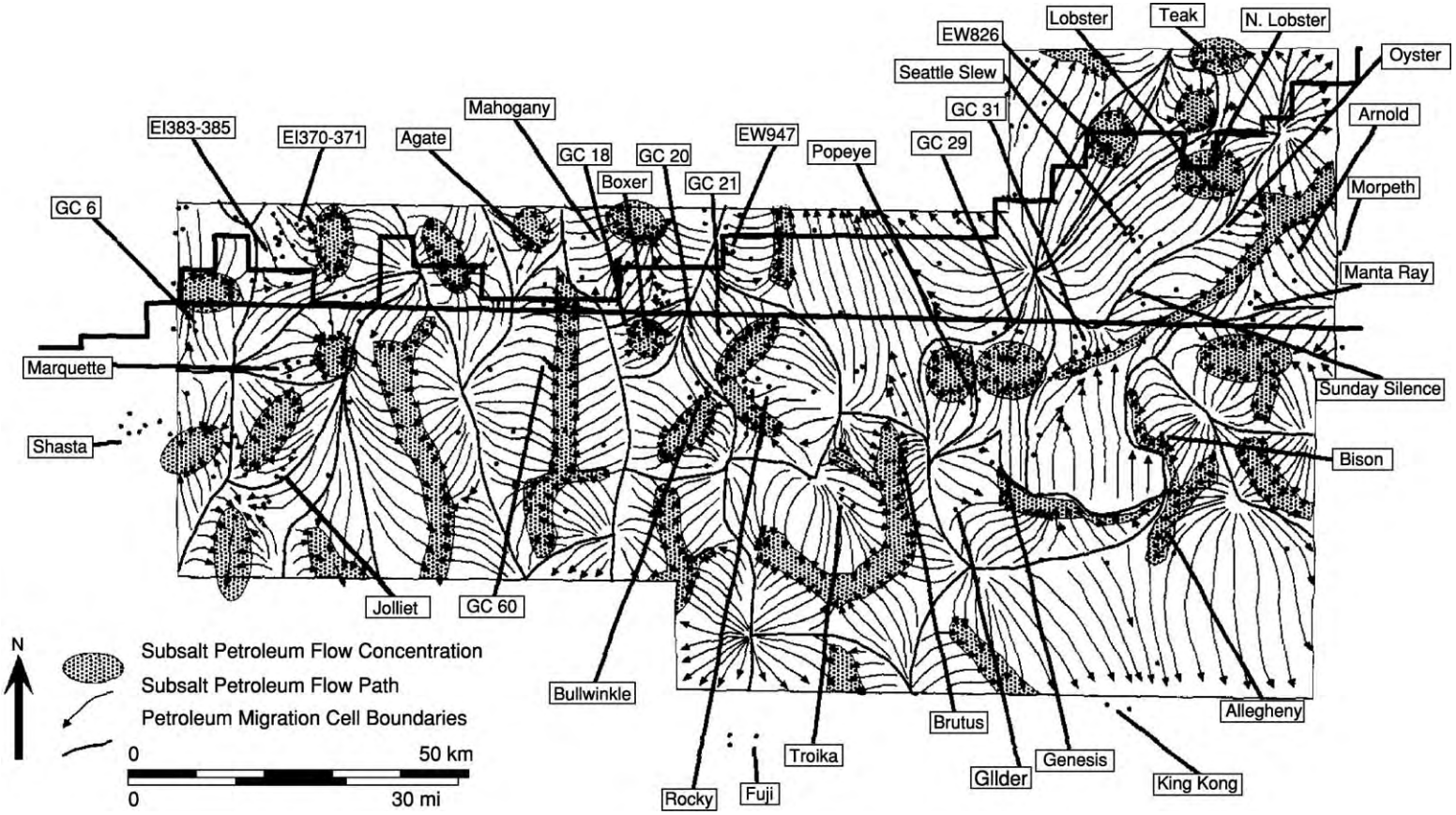


Figure 20 Orthocontours reconstructed subsalt petroleum migration pathways, Ewing Bank to Green Canyon areas, Gulf of Mexico, USA. Reproduced with permission from McBride BC, Weimer P, and Rowan MG (1998) The effect of allochthonous salt on the petroleum systems of the Northern Green Canyon and Ewing Bank (Offshore Louisiana), Northern Gulf of Mexico. *American Association of Petroleum Geologists Bulletin* 82: 1083 1112.

Figure 21 Petroleum migration along high permeability sand stone beds within a stacked sequence of turbidite sandstones and siltstones. The migration route was exposed during the excav ation of a road cutting in Ecuador. Photograph by M. Heffernan. Reproduced from England WA, Mackenzie A, Mann D, and Quigley T (1987) The movement and entrapment of petroleum fluids in the subsurface. *J. Geol. Soc.*, vol. 144, p. 327. London.

Tertiary Migration

Tertiary migration includes leakage, seepage, dissipation, and alteration of petroleum as it reaches the Earth's surface. The products of seepage may be gas chimneys in the shallow sediment, gas hydrate layers and mounds, cemented pock marks and mud volcanoes, effects on vegetation, and live oil and gas seepage at the surface.

The physical processes that drive tertiary migration are the same as those that operate during secondary migration. Buoyancy drives the petroleum to the surface. This may be helped or hindered by overpressure gradients or hydrodynamics. The only major difference that can be used to separate tertiary migration from secondary migration is the rate of petroleum supply. Trap failure, through capillary leakage, hydraulic fracture, or tectonism, supplies petroleum into a new carrier system much more rapidly than does a maturing source rock.

See Also

Petroleum Geology: The Petroleum System; Exploration; Production. **Sedimentary Environments:** Depositional Systems and Facies; Carbonate Shorelines and Shelves. **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Chalk; Dolomites; Sandstones, Diagenesis and Porosity Evolution; Limestones. **Tectonics:** Faults; Fractures (Including Joints).

Further Reading

Abbots IL (1991) *United Kingdom Oil and Gas Fields, 25 Years Commemorative Volume, Geological Society Memoir No. 14*. London: Geological Society.

Allen PA and Allen JR (1990) *Basin Analysis, Principles and Applications*. Oxford: Blackwell Science.

Archer JS and Wall PG (1986) *Petroleum Engineering, Principles and Practice*. London: Graham & Trotman.

Bathurst RGC (1976) *Carbonate Sediments and Their Dia genesis, Developments in Sedimentology 12*. Oxford: Elsevier.

Dickson JAD and Saller AH (1995) Identification of sub aerial exposure surfaces and porosity preservation in Pennsylvanian and Lower Permian shelf limestones, east ern central Basin Platform, Texas. In: Budd DA, Saller AH, and Harris PM (eds.) *Unconformities and Porosity in Carbonate Strata, American Association of Petroleum Geologists Memoir 63*, pp. 239 258. Tulsa, OK: Ameri can Association of Petroleum Geologists.

England WA and Fleet AJ (1991) *Petroleum Migration, Special Publication 59*. London: Geological Society.

Foster PT and Rattey PR (1993) The evolution of a frac tured chalk reservoir: Machar Oilfield, UK North Sea. In: Parker JR (ed.) *Petroleum Geology of Northwest Europe: Proceedings of the 4th Conference*, pp. 1445 1452. London: Geological Society.

Glennie KW (1998) *Petroleum Geology of the North Sea, Basic Concepts and Recent Advances*, 4th edn. Oxford: Blackwell Science.

Gluyas JG and Hichens HM (2003) *United Kingdom Oil and Gas Fields Commemorative Millennium Volume, Memoir 20*. London: Geological Society.

Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

McBride BC, Weimer P, and Rowan MG (1998) The effect of allochthonous salt on the petroleum systems of the Northern Green Canyon and Ewing Bank (Offshore Lou isiana), Northern Gulf of Mexico. *American Association of Petroleum Geologists Bulletin* 82: 1083 1112.

Reynolds AD (1994) Sequence stratigraphy and the dimen sions of paralic sandstone bodies. In: Johnson SD (ed.) *High Resolution Sequence Stratigraphy: Innovations and Applications*, pp. 69 72. Liverpool: Liverpool University.

Selley RC (1996) Elements of Petroleum Geology, 2nd edn. San Diego: Academic Press.

Chemical and Physical Properties

C Clayton, Eardiston, Tenbury Wells, UK

Definitions

Geochemists define petroleum as any subsurface material which, when produced, yields crude oil and/or gas. Most petroleum forms from the thermal degradation of kerogen during burial, although, strictly speaking, biogenic or bacterially formed gases should also be included. Hydrocarbons are molecules consisting only of carbon and hydrogen atoms. Although they are one of the principal constituents of petroleum, they are not the only components present, and so the two terms are not, strictly speaking, equivalent. Oil is the fraction of petroleum which is liquid at standard temperature and pressure, whilst the remainder is gas. Gas typically consists of compounds with up to five carbon atoms (C_1–C_5), whereas oil has six or more (C_{6+}), although some smaller molecules are sometimes present in oil and some C_{6+} compounds may be present in gas. Condensate is petroleum that is liquid at surface conditions, but exists in the reservoir in the gas phase as a gas condensate. There is no genetic connotation to this definition, although, in general, condensates tend to be of high thermal maturity.

Oils

Chemical Composition

Oil is a complex mixture, dominated by hydrocarbons, but with varying amounts of compounds containing sulphur, nitrogen, oxygen, and metals, principally nickel and vanadium. The chemistry of an oil is determined both by the nature and quality of the source rock and the depositional environment of the kerogen from which the oil formed, and by the thermal maturity of the source rock (i.e., the maximum temperature encountered in the source rock). For example, oil sourced from a marine carbonate is typically very heavy and sulphur rich, but becomes lighter and progressively sulphur depleted with increasing maturity. Clastic marine source rocks tend to produce lower sulphur, lighter oils. In addition, the composition of the petroleum may subsequently become altered in the reservoir by biological processes (biodegradation), thermal cracking, or physical separation (e.g., precipitation of high molecular weight components to form a tar mat).

At the most basic level, oil components can be divided into four main compound groups: saturated compounds, aromatic compounds, an NSO, Nitrogen-, Sulphur-, and Oxygen-bearing compounds, (sometimes called a residue or resin) fraction, and asphaltenes (Figure 1). In practice, the last two groups form a continuum, as their separation is based on solubility in the laboratory, rather than on a unique chemical composition. They are often referred to together as the polar fraction, as their molecules often have a significant surface charge, unlike saturated and aromatic compounds.

Saturated compounds These are generally the most abundant components of oil, comprising up to 90% of some light crudes. A large number of compounds are present within the saturated fraction, but these can be broadly divided into three groups: normal alkanes, branched alkanes, and cyclic compounds (Figure 2). All of these are said to be saturated as they contain no double chemical bonds.

Normal, or straight-chain, alkanes, denoted by the prefix '*n*', range in carbon number from five up to several hundred, but are dominantly in the range *n*-C_5 to *n*-C_{40}. Alkanes greater than about *n*-C_{20} are often referred to as the wax fraction, although here again the definition is based on laboratory solubility rather than on strict chemistry. In practice, industrial 'paraffin wax' also contains branched and cyclic alkanes. Within these higher molecular weight *n*-alkanes, oils derived from higher land plants usually show a dominance of odd carbon numbers over even, reflecting their origin from cuticular waxes of continental higher plants. With increasing maturity, the distribution of *n*-alkanes becomes progressively skewed towards shorter chain lengths and any generic information is lost. Normal alkanes are also particularly vulnerable to bacterial attack and are usually absent in biodegraded oils.

Branched alkanes contain one or more side-chains attached to the straight-chain backbone. These are quantitatively the most abundant fraction within the saturates. A special subgroup of the branched alkanes, called the isoprenoids, is worthy of note. These range from C_9 to C_{25} and are believed to be derived from the phytol side-chain of chlorophyll. The C_{19} and C_{20} compounds, pristane (Pr) and phytane (Ph), are of particular diagnostic value in determining the source of the oil. Coal-sourced oils

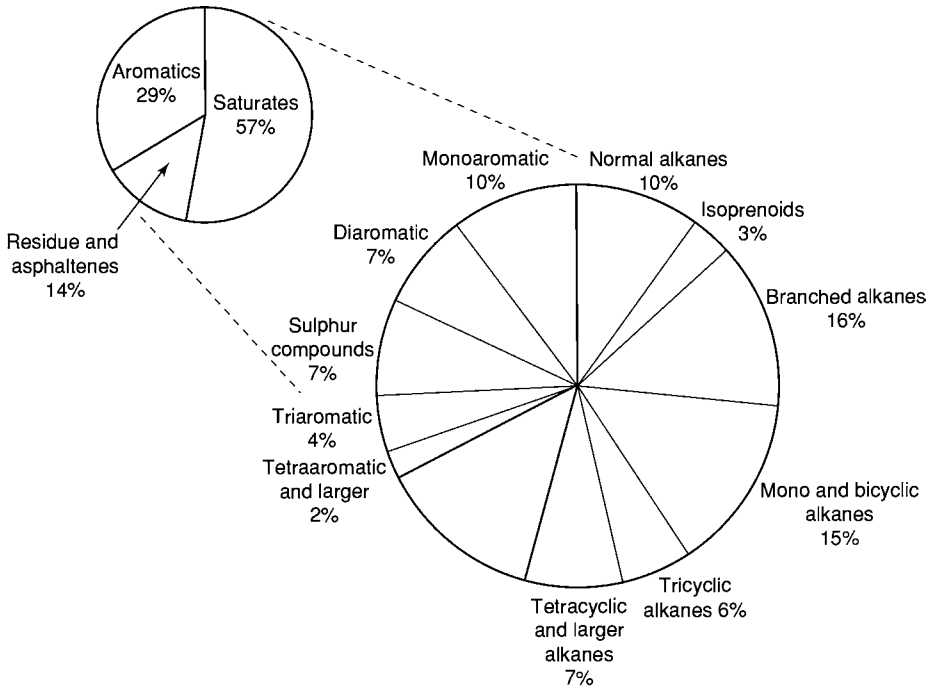


Figure 1 Average composition of oils based on between 100 and 500 samples depending on fraction.

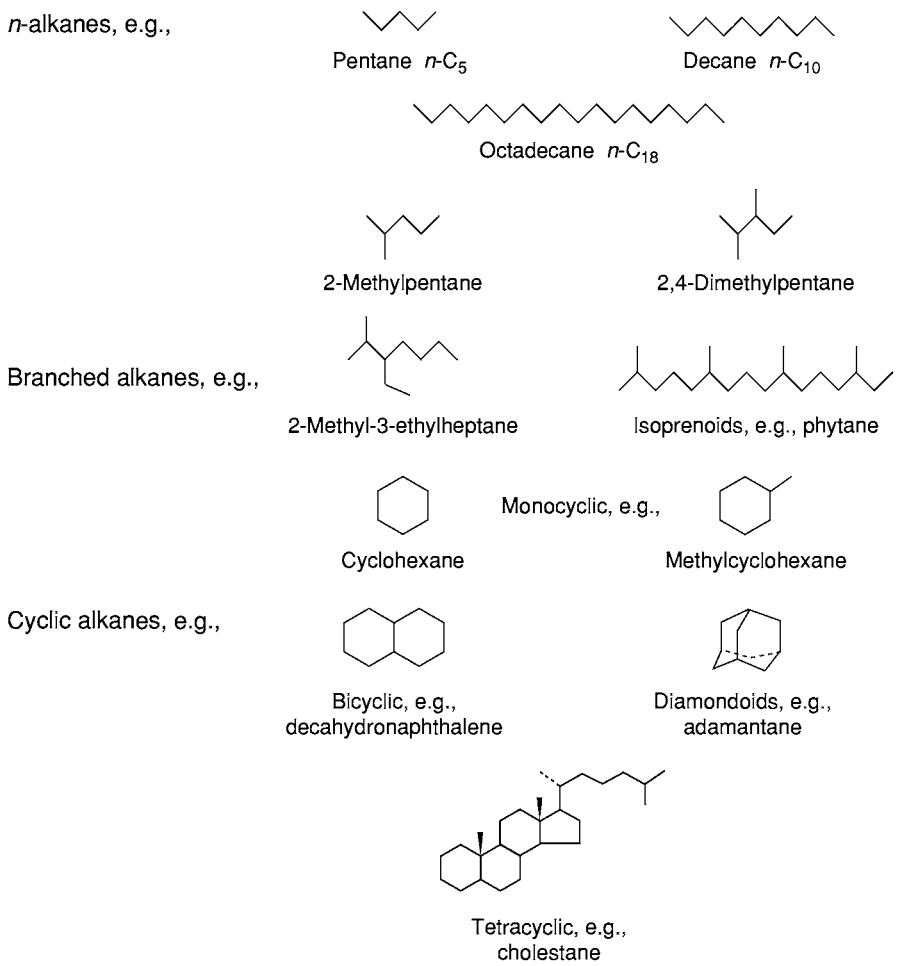


Figure 2 Typical saturated compounds in oil.

generally have a Pr/Ph ratio above three, whereas almost all aquatic kerogens produce oils with a Pr/Ph ratio less than this.

Cyclic alkanes contain carbon rings, usually with five or six carbon atoms in each ring. These are sometimes referred to as the naphthenes, in contrast with the straight and branched alkanes which are paraffins. These compounds may be monocyclic (one ring), or contain up to five or, in rare cases, more rings.

Amongst the saturated compounds (and, to a lesser extent, the aromatic and polar compounds too), some of the more complex molecules can have their origin traced back to specific biomolecules deposited in the source kerogen. They are therefore useful to determine the source of the oil. These so-called biomarkers (or geochemical fossils) are derived ultimately from bacterial or algal cell membranes, or from specific metabolites in algae, bacteria, or higher land plants, and a very large number are known. Some typical examples are shown in Figure 3. Particularly important amongst the polycyclic biomarkers are the pentacyclic steranes and tetracyclic triterpanes (hopanes), as, although present in only very low abundance, many of them are highly characteristic of the source input to the kerogen from which the oil was derived. In addition, subtle variations in the chemical structure of many of the biomarkers (such as their stereochemistry and degree of aromatization) are temperature sensitive, and hence can be used as molecular maturity indicators (see Figure 4).

Aromatic fraction Aromatic compounds contain one or more benzene rings within their structure (Figure 5). These generally consist of roughly equal amounts of one-ring, two-ring, and organosulphur compounds (mainly benzothiothenes and dibenzothiothenes; Figure 6), although lesser amounts of larger aromatic compounds are also present. Some of the more complex compounds, such as the so-called aromatic steranes, are formed by thermal alteration of the saturated equivalent, and hence their relative abundance is indicative of maturity (Figure 4).

Polar fractions The NSO and asphaltene fractions are made up of more complex compounds, usually containing nitrogen, oxygen, or sulphur (Figure 6). The larger molecules can be thought of as small fragments of kerogen. Asphaltenes are the highest molecular weight components present and are defined as the fraction that is not soluble in hydrocarbon solvents (usually hexane). In some heavy oils (such as those formed at low maturity or by bacterial alteration of an originally lighter oil), asphaltenes may comprise up to 50% of the total oil, but they are generally present as only a few per cent. It is the abundance of these polar fractions that determines in most part the viscosity of the oil.

The detailed chemistry of the polar fractions is less well understood than that of the other compound groups because of the complexity of many of the species present. However, recent research suggests that a number of the nitrogen- and oxygen-bearing molecules may be of value as indicators of migration distance, as their ratios potentially reflect the extent to which the oil has interacted with water since expulsion from the source rock.

Analytical Methods

Chemical analyses of oils are most frequently presented in graphical form rather than as tables of results, and so it is relevant to consider the most important analytical methods used to produce these plots. Although a very wide range of analytical tools are now used, by far the most important methods are based on chromatography.

Asphaltene separation and liquid chromatography The largest and most polar components are separated chemically as, by definition, asphaltenes are insoluble in hydrocarbon solvents (e.g., hexane or heptane). No further measurements are carried out on the asphaltene fraction other than perhaps carbon isotope measurements (see below) or occasionally pyrolysis to release smaller, more easily identifiable, components. After asphaltene separation, the remaining mixture is usually separated into compound groups using high-performance liquid chromatography (HPLC). This separates the saturated compounds, aromatic compounds, and residue fraction (the NSO fraction or resins).

Gas chromatography The saturated fraction (and sometimes the aromatic fraction) is then further analysed by gas chromatography (GC) (*see* **Analytical Methods:** Geochemical Analysis (Including X-Ray)). In this method, gas is used to force a sample through a long narrow tube. Small molecules elute first, whilst larger molecules are held up. It is therefore possible to resolve and quantify individual compounds of interest in the oil. The resulting gas chromatograms (Figure 7) show the concentration of each (resolved) compound plotted against the retention time in the chromatographic column. These show a number of features of interest, such as the distribution of the *n*-alkanes, the Pr/Ph ratio, the amount of wax (high molecular weight alkanes), and sometimes characteristic individual compounds of value in correlation studies. Gas chromatography is the method also used to determine natural gas composition.

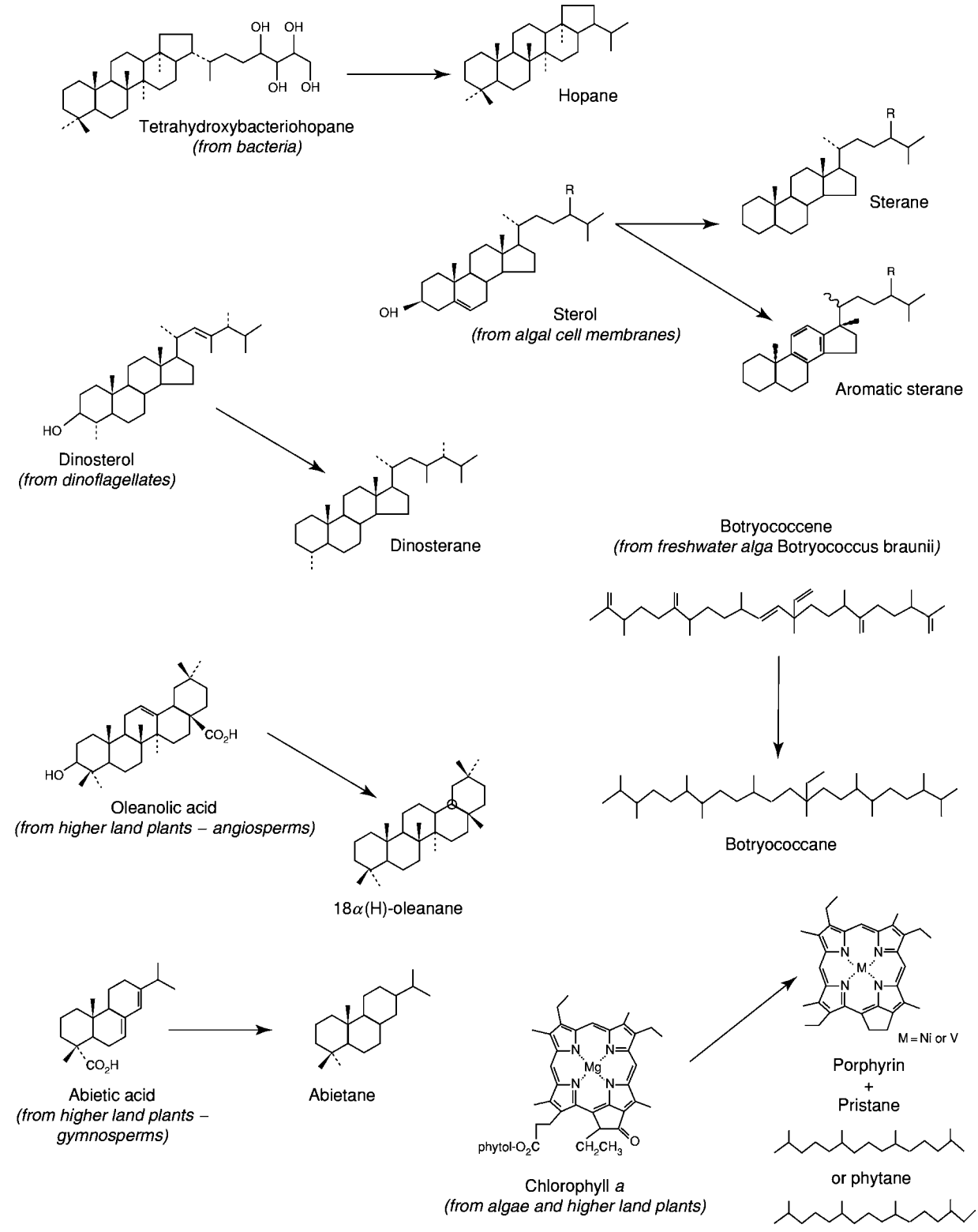


Figure 3 Origin of some common biomarkers.

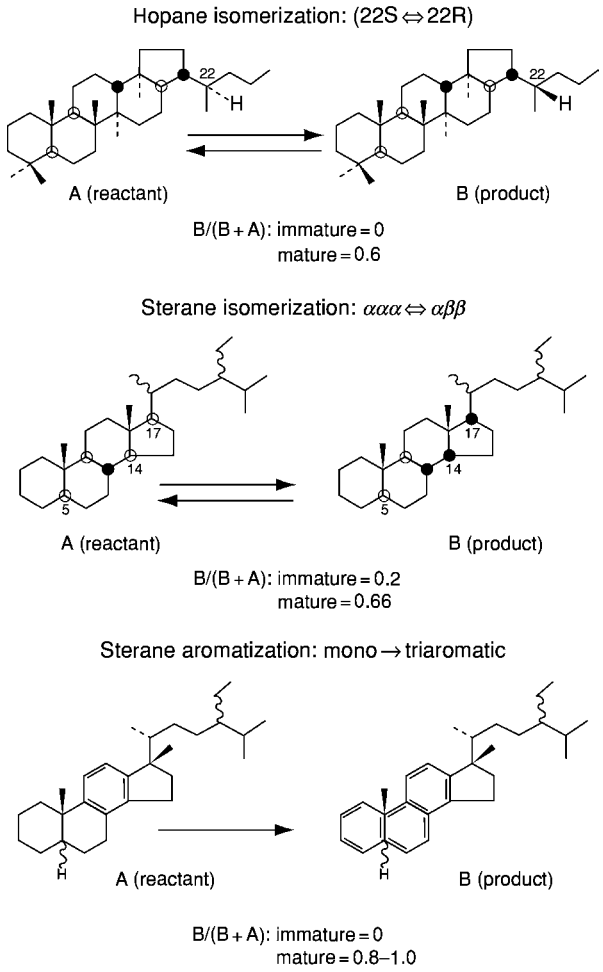


Figure 4 Examples of molecular maturity parameters.

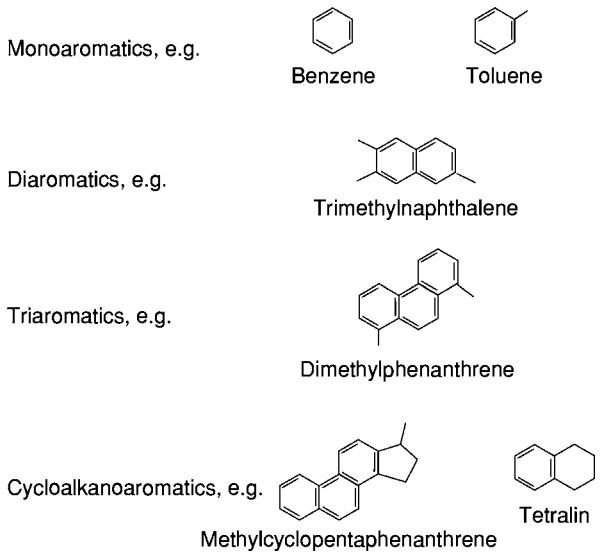


Figure 5 Typical aromatic compounds in oil.

Gas chromatography-mass spectrometry Most of the biomarkers of interest are present in only trace quantities, and so special methods are needed to measure their concentration. Central amongst these is gas chromatography-mass spectrometry (GCMS). This is similar to conventional GC, but a mass spectrometer is used as a detector. When a sterane, for example, enters the mass spectrometer, it fragments into many parts. However, all steranes produce a fragment with a characteristic mass (mass/charge ratio, *m*/*z*; *m*/*z* 217 in this case). Thus, by monitoring all fragments of *m*/*z* 217 as they come off the GC into the mass spectrometer, we are essentially resolving all of the steranes into individual compounds. The result is a plot similar to a GC trace but specific to the steranes. Similarly, the triterpanes (dominantly the hopanes) produce a characteristic fragment at *m*/*z* 191, and so on for all the other compound groups of interest. Figure 8 shows an example of mass 191 and 217 mass fragmentograms for a typical North Sea oil with some of the important peaks labelled.

Classification of Crude Oils

There are numerous schemes for classifying crude oils, but two main approaches dominate. The first is based on the relative concentrations of the main chemical classes and the most widely used of these was developed by the Institut Francais du Pétrole (Figure 9). Alternatively, and more useful for exploration purposes, oils can be classified according to the depositional environment in which their source kerogen was formed. A typical example of this is shown in Table 1, which includes some of the most relevant chemical characteristics for each environment. It should be noted, however, that, with increasing maturity, the chemical signatures are progressively lost as the characteristic compounds become progressively broken down to smaller, less diagnostic compounds.

Bulk Properties

The economic value of an oil is usually dictated by its bulk properties rather than its detailed chemistry, although the two are obviously inextricably linked. A number of standard measurements are used to define the bulk properties.

API gravity API gravity (from American Petroleum Institute) is a measure of the specific gravity of the oil at 60°F or 16°C. It is somewhat arbitrarily defined by the equation

$$\text{API}^\circ = (141.5/\text{specific gravity}) - 131.5$$

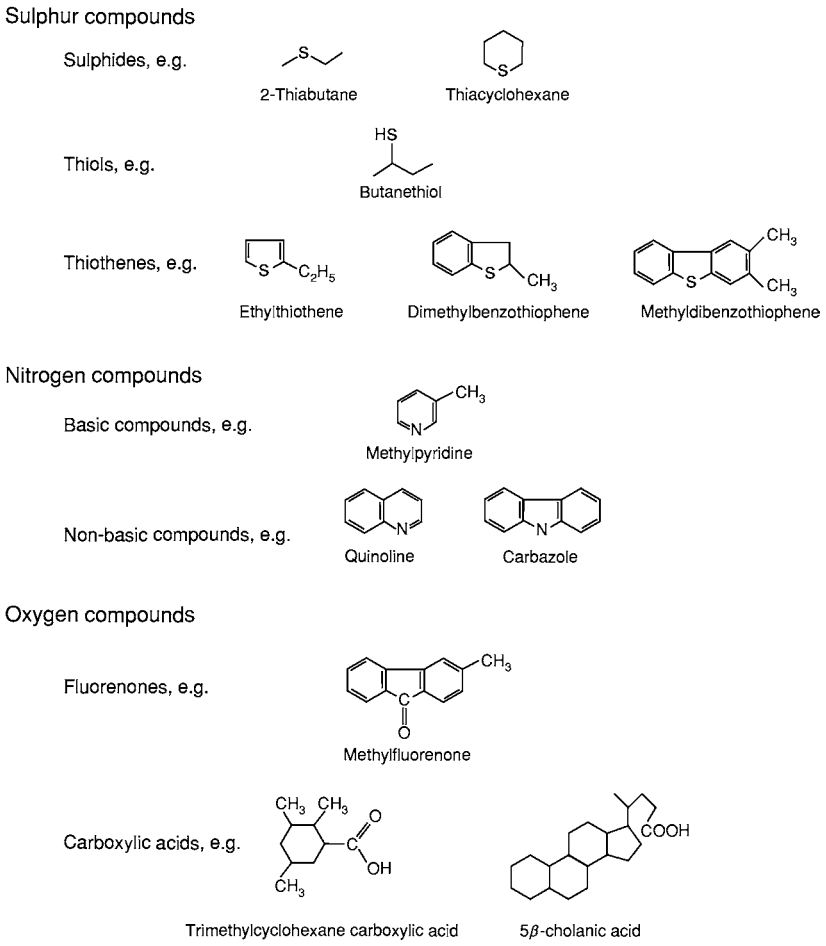


Figure 6 Typical sulphur , nitrogen , and oxygen bearing compounds found in oil.

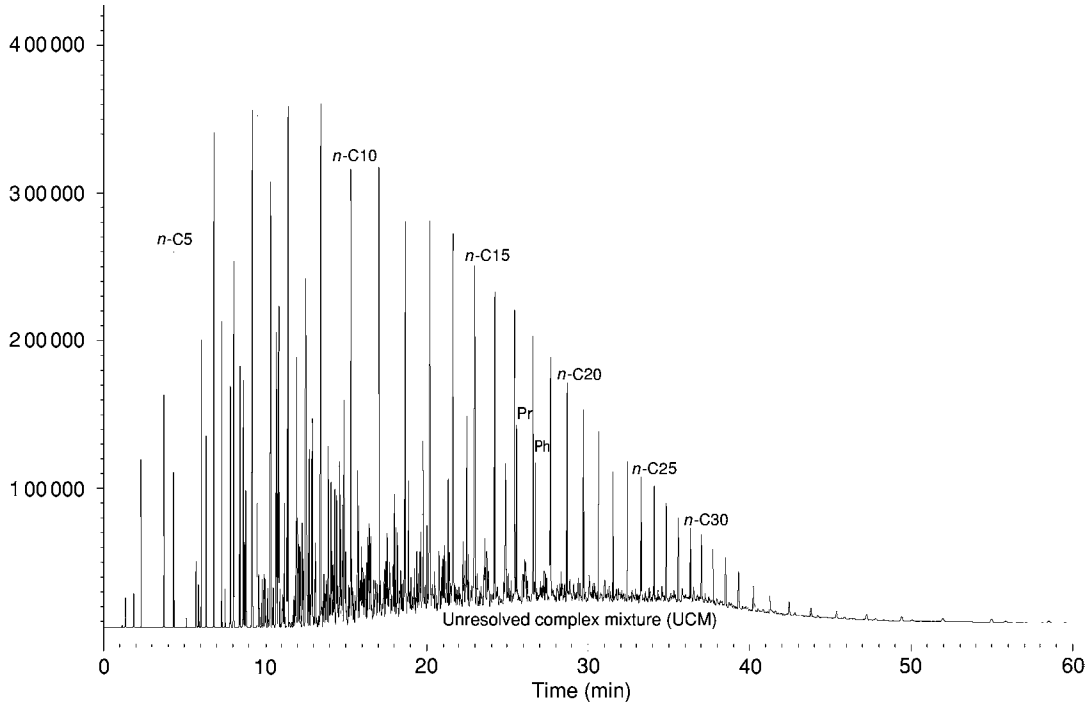


Figure 7 Typical gas chromatogram of saturates fraction. Ph, phytane; Pr, pristane.

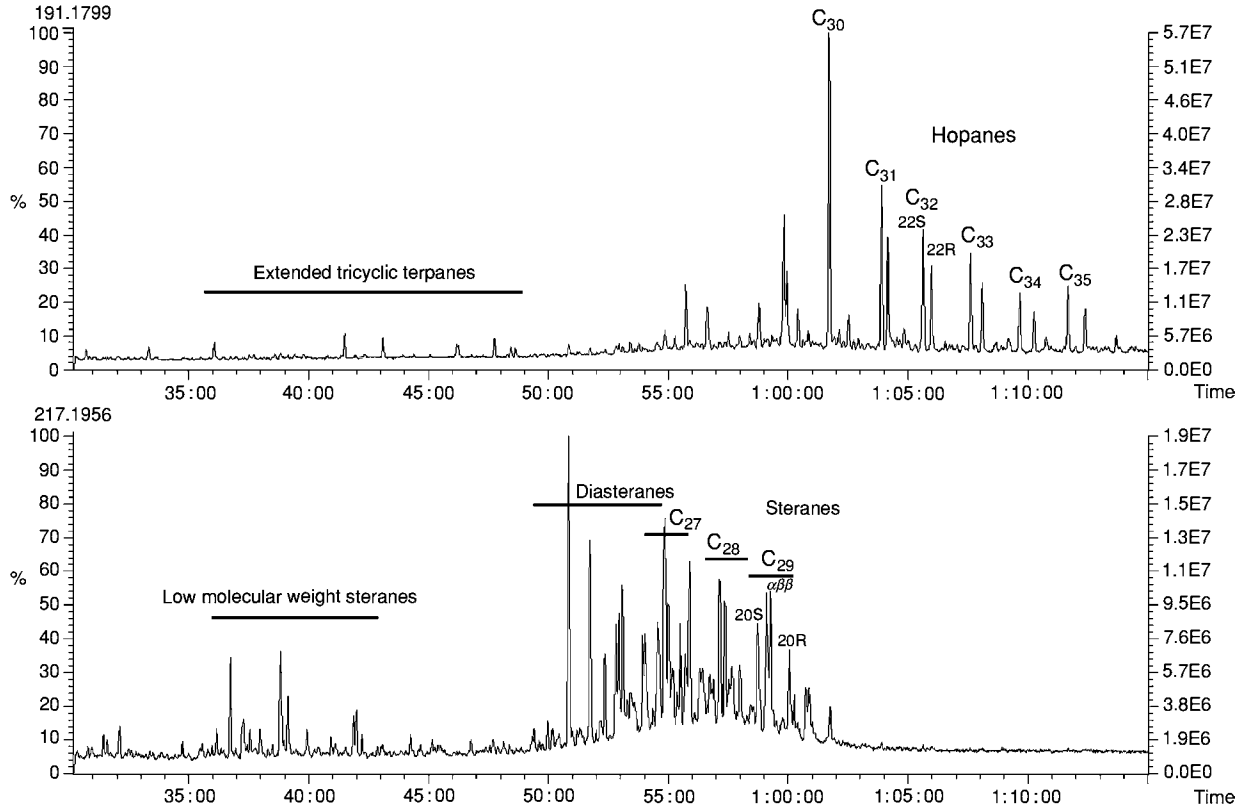


Figure 8 Typical gas chromatography mass spectrometry (GCMS) fragmentograms. Top part shows the *m/z* 191 trace; bottom part shows the *m/z* 217 trace.

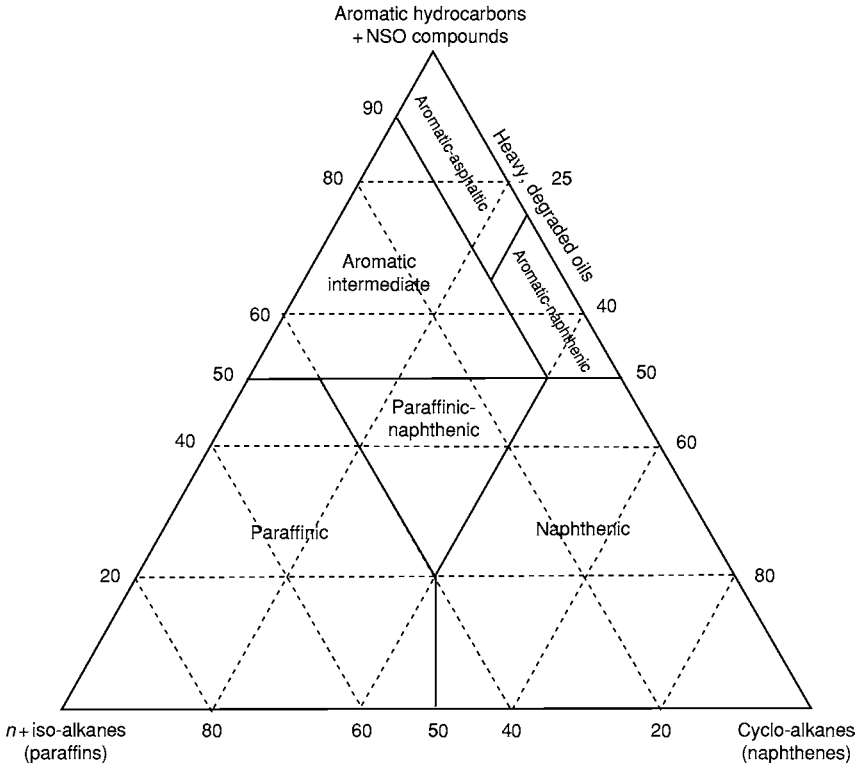


Figure 9 Classification of crude oils based on bulk composition. NSO, Nitrogen , Sulphur , and Oxygen bearing compounds.

Table 1 Oil classification according to source type

	Typical characteristics at moderate maturity[a]					
	API°	*Wax*	*Sulphur*	*Nitrogen*	*Ni + V*	*Ni/V*
Marine carbonate or siliceous	25 30	Low	High	High	50 300	<2
Marine siliciclastic	35 40	Low	Moderate low	Moderate	0 50	<2
Lacustrine	30 35 and 45 50[b]	High	Low	Moderate	0 50	>2
Coal	30 45	High	Low	Low	~0	

API, American Petroleum Institute.
[a]With increasing maturity, API increases, and wax, sulphur, nitrogen, and metal contents decrease.
[b]Bimodal distribution observed.

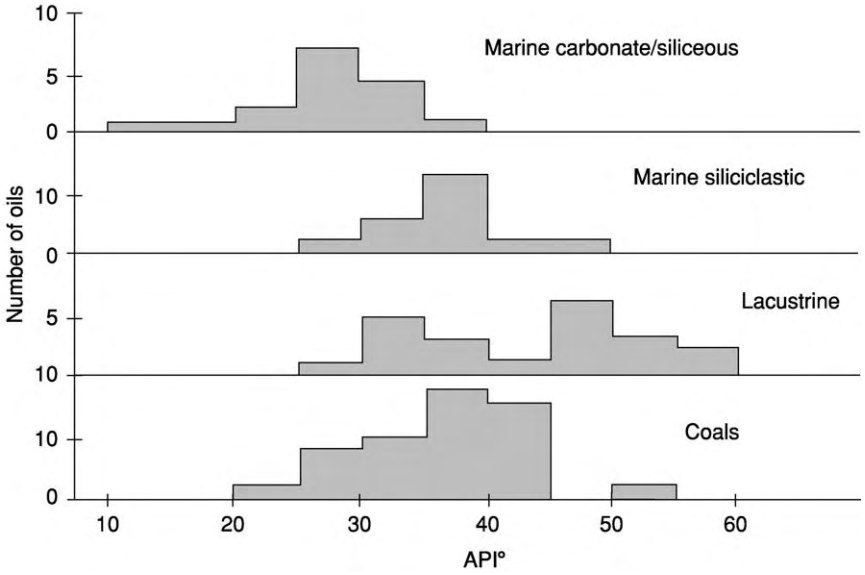


Figure 10 API (American Petroleum Institute) gravity variation of oils according to the depositional environment of the source kerogen.

This system is used as it gives an integer value that is easier to display than a four decimal place number (e.g., pure hexane with a specific gravity of 0.6594 is equivalent to an API of 83°). A heavy crude will generally have an API <25°, whilst a light crude will generally have an API between 35° and 45°. Most condensates have an API >45°. Pure water, by definition, has an API of 10°, and thus crudes with API <10° will not float on water.

The API gravity varies somewhat depending on the source of the oil (Figure 10), and also increases significantly with maturity regardless of source. Oils derived from carbonate or siliceous source facies are usually of lower API than those from other lithologies. In addition, bacterial alteration of oil in the reservoir can result in oils with especially low API values, and gravity segregation in thick homogeneous reservoir formations or highly compartmentalized reservoirs can lead to significant vertical and lateral heterogeneity, respectively.

Gas/oil ratio The gas/oil ratio (GOR) is the ratio of components that exist in the gas phase at surface conditions (dominantly compounds containing less than five carbon atoms) to the rest of the petroleum. It is traditionally expressed as standard cubic feet of gas per barrel of oil (sft^3/bbl), and 1 sft^3/bbl is equal to 0.178 m^3 m^{-3}. When dealing with gas condensates, it is often more convenient to express this ratio as the condensate/gas ratio (CGR) in terms of barrels of condensate per million standard cubic feet of gas (bbl/MMscf).

GOR varies depending on the source of the oil, but is influenced mainly by maturity. GOR increases enormously with increasing maturity.

Wax content The wax content is a reflection of the amount of high molecular weight saturated compounds (*n*-alkanes $>C_{20}$). Most oils range from 0% up to about 40% wax. There are two distinct sources of wax in oil. It may be derived ultimately

from cuticular waxes on the leaves of higher land plants or from long-chain unsaturated molecules synthesized by some freshwater algae. Consequently, many lacustrine or coal-sourced oils have a high wax content (Figure 11), although this decreases with increasing maturity. Wax deposits in pipelines and other equipment can cause mechanical problems.

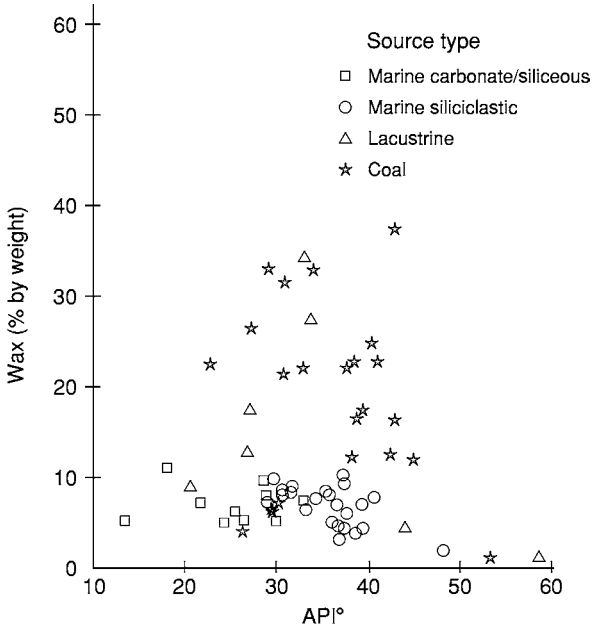


Figure 11 Wax content of oil types vs. API (American Petroleum Institute) gravity as a proxy for maturity.

Sulphur content The sulphur content of an oil has an important impact on its economic value as it causes problems during refining. The main control on the sulphur content is the kerogen type. Oils derived from carbonate or siliceous source rocks are higher in sulphur than those from other lithologies (Figure 12). This itself is a consequence of bacterial sulphate reduction: there is little free iron available in carbonate or siliceous sediments, and so the hydrogen sulphide formed reacts with the organic matter to give a sulphur-rich kerogen. All oils show a marked decrease in sulphur content with maturity.

Nitrogen content The nitrogen content of an oil is a reflection of the abundance of the NSO fraction and hence usually correlates with API gravity and viscosity. The nitrogen content is generally very low in oils sourced from coal, and highest in those derived from marine carbonate and siliceous sediments, but decreases markedly with increasing maturity (Figure 13).

Viscosity The viscosity of oils ranges from around 0.2 cP (2×10^{-4} Pa s) for a very light oil up to around 100 P (10 Pa s) for some very heavy oils. The viscosity is controlled dominantly by the abundance of the high molecular weight polar components (NSOs and asphaltenes), and hence shows a correlation with API gravity and sulphur content.

Distillation fractions Of interest to the refining process are the distillation fractions of an oil (i.e., the distribution of different boiling point fractions). The

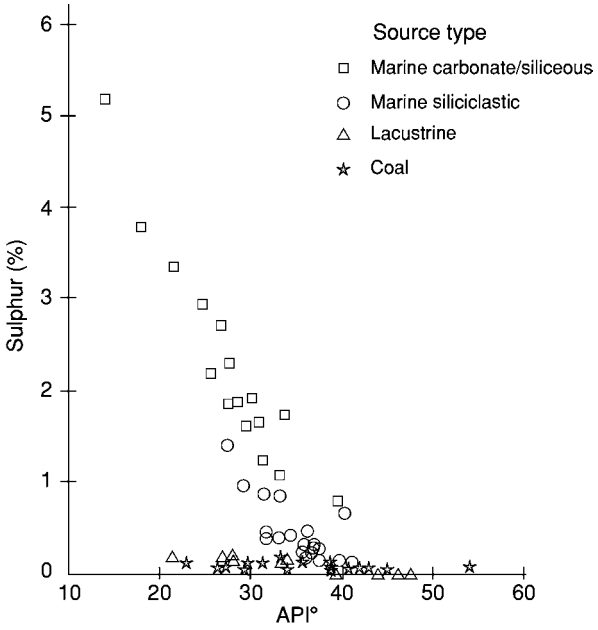


Figure 12 Sulphur content of oil types vs. API (American Petroleum Institute) gravity as a proxy for maturity.

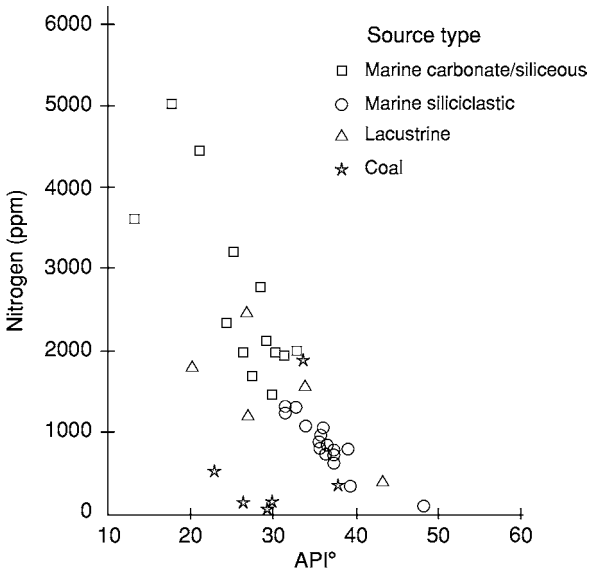


Figure 13 Nitrogen content of oil types vs. API (American Petroleum Institute) gravity as a proxy for maturity.

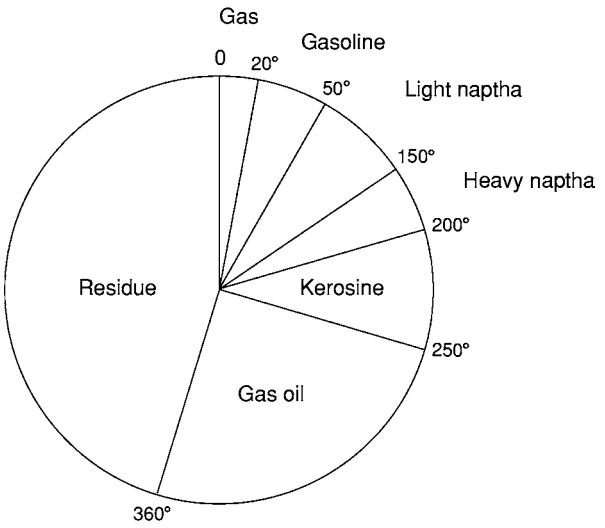


Figure 14 Distribution of boiling point cuts for a typical crude oil.

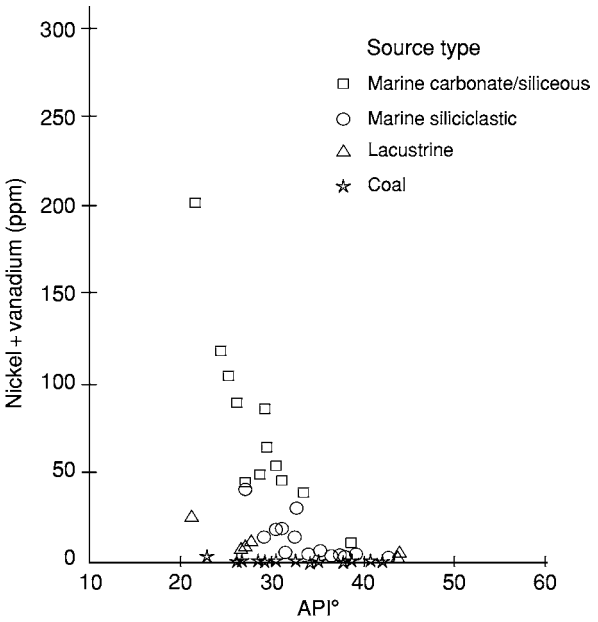


Figure 15 Metal content of oil types vs. API (American Petroleum Institute) gravity as a proxy for maturity.

higher the API gravity of an oil, the larger the yields of the lower boiling point cuts. Various names are applied to the different fractions, an example of which is summarized in Figure 14.

Metal content The metal content of oils is dominated by nickel and vanadium, although other metals, such as iron, zinc, copper, lead, molybdenum, chromium, manganese, cobalt, and arsenic, are also sometimes present in trace quantities. These metals are present dominantly in a specific group of cage-like compounds, called porphyrins (Figure 3), which are ultimately derived from chlorophyll. High nickel and vanadium contents significantly reduce the value of an oil, as they adversely affect the catalysts used in the cracking of oil at the refinery. Metal contents vary enormously, from less than 1 ppm in some Palaeozoic crudes up to 1200 ppm vanadium and 150 ppm nickel in the Boscan crude from Venezuela, and are again controlled by both the source type and maturity (Figure 15). In addition, the ratio of nickel to vanadium is of some diagnostic value in determining the source of the oil (Figure 16).

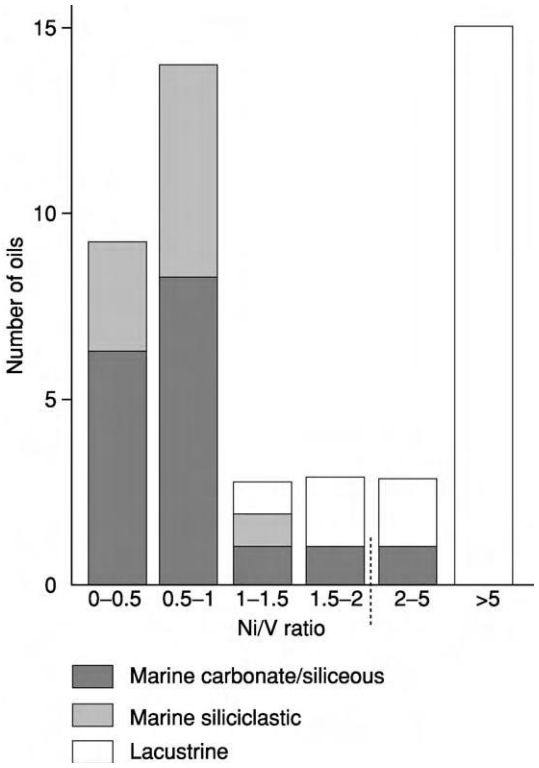


Figure 16 Nickel/vanadium ratios of different oil types.

Gases

Natural gases consist dominantly of hydrocarbons up to C_5 (pentane), together with varying amounts of inorganic components, principally nitrogen (N_2), carbon dioxide (CO_2), and hydrogen sulphide (H_2S).

Hydrocarbon Gases

Natural hydrocarbon gases can be either 'biogenic' (arising from bacterial processes, usually during early diagenesis) or 'thermogenic' (resulting from the effects of heat on oil or kerogen) in origin.

Biogenic (or bacterial) gases are formed in organic-rich sediments during diagenesis by a specialized

group of bacteria, the 'methanogens'. Methanogenic bacteria require anoxic, sulphate-free environments at temperatures below approximately 75°C. These are met, for example, in paddy fields, the intestinal tracts of higher organisms, and in marine sediments below the zone of sulphate reduction. Chemically, biogenic gases are almost pure methane.

Under the right conditions, biogenic gases can form substantial economic accumulations. They account for 15–20% of the world's gas resources, including the Cenomanian pay zones of the giant Urengoy field in western Siberia with ~180 trillion cubic feet recoverable reserves. They occur at depths down to 4600 m (Antonella, Po Valley) and in sediments from Holocene to Cretaceous age.

The thermogenic gases can be subdivided depending on their origin. 'Oil-associated gases' form from kerogen at the same time as oil is generated. They can be thought of as the short-chain component of the oil and are believed to be derived from the lipids (essential oils, fats, and waxes) of organisms. 'Non-associated gases' are derived from lignin, a component of higher land plants. They are the dominant component of gases derived from coals, although they are not strictly synonymous with coal-gas which may also contain cracked oil-derived gas, particularly where oil-prone coals are involved. 'Cracked oil gases' are formed from the thermal destruction of oil, either in the reservoir or by cracking of residual unexpelled oil in the source rock.

The chemistry of natural gases is basically simple. Hydrocarbon components are based on methane (CH_4) and the higher homologues ethane (C_2H_6), propane (C_3H_8), iso- and normal butane (C_4H_{10}), and iso- and normal pentane (C_5H_{12}). Hydrocarbon components containing more than five carbon atoms are generally referred to as the gasoline fraction and considered to be part of oil rather than gas. Unsaturated compounds, containing a double bond, such as ethene (ethylene), are rarely found in natural gases, although can form as a product of laboratory pyrolysis.

Because of the simplicity of gas composition, very little can be deduced from the composition alone, although two ratios are in common usage: the gas 'wetness' (strictly speaking the gas dryness!), $C_1/\Sigma C_{1-5}$, and the iso- to *n*-butane ratio, iso-C_4/*n*-C_4. Some authors also use simpler ratios, such as the methane/ethane ratio (C_1/C_2), or methane relative to ethane and propane ($C_1/C_2 + C_3$). In general, these are not as useful as the gas wetness index as they are open-ended scales (i.e., they extend to infinity). All of these ratios are affected by the gas origin, maturity, and sampling methodology, and generally also by secondary alteration processes, such as biodegradation. Dry gases are generally considered to contain less than 10% heavy hydrocarbons (i.e., $C_1/\Sigma C_n > 0.9$), although different authors use different cut-off points. It is therefore necessary to exercise caution when describing gases as 'wet' or 'dry'.

The best evidence of the source of a natural gas is from the stable carbon isotope ratios of the individual hydrocarbons (methane to butanes and sometimes pentanes). The bulk fractions of oils, and increasingly a number of individual compounds, are often also analysed for their stable carbon isotopic composition. The carbon isotope ratio is based on the ratio of the stable carbon isotopes ^{13}C and ^{12}C and is universally expressed in a standard 'del' (or delta) notation

$$\delta^{13}C = [(R_{sample} - R_{standard})/R_{standard}] \times 1000‰$$

in which R is the ratio $^{13}C/^{12}C$ and the units are parts per thousand or 'per mil'. The standard used for carbon is PDB, calcite from a fossil belemnite from the Cretaceous Peedee Formation of South Carolina, although secondary standards are now used in the laboratory. Positive values contain more ^{13}C than the standard, and are said to be isotopically heavier or enriched. Negative values are isotopically lighter or depleted.

When combined with basic chemical data for the gases, these allow the differentiation of the different sources of natural gas, estimates of their thermal maturity, and, under favourable conditions, the identification of the specific source rock involved (see examples in Figures 17 and 18).

Non-Hydrocarbon Gases

In many reservoirs, non-hydrocarbon gases (notably CO_2, N_2, and H_2S) are encountered in varying proportions. A number of sources for these may be proposed, although it is often difficult to be confident about their origin, even though isotope ratios of carbon, nitrogen, and sulphur can provide useful information. Table 2 shows composition data for a number of natural gases. The non-associated gases in particular contain large amounts of non-hydrocarbons. These reduce the economic value of an accumulation, partly by diluting the hydrocarbons and partly by causing problems in production (e.g., H_2S is both corrosive and highly toxic).

Nitrogen In some cases, nitrogen can be derived from basement rock, particularly from igneous sources. It can also be concentrated in biogenic

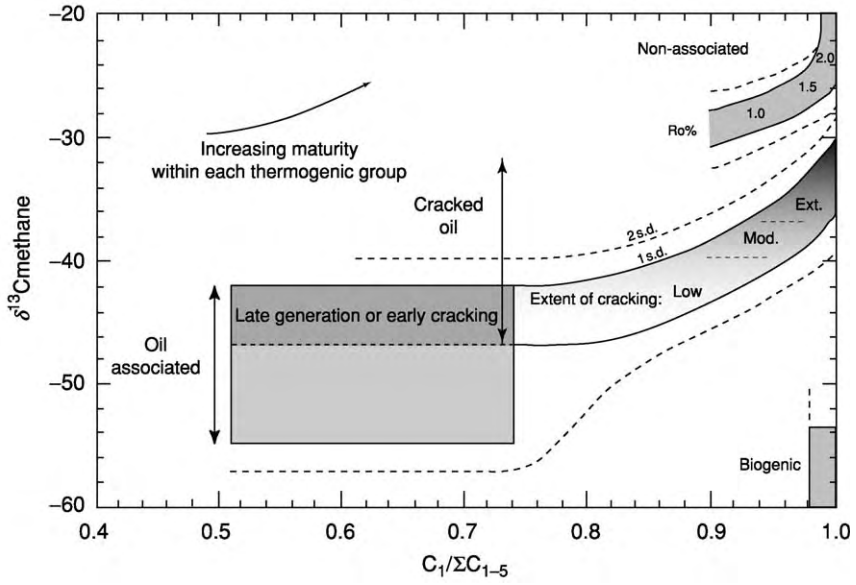


Figure 17 Differentiation of gas origins using gas wetness and carbon isotope ratios of methane.

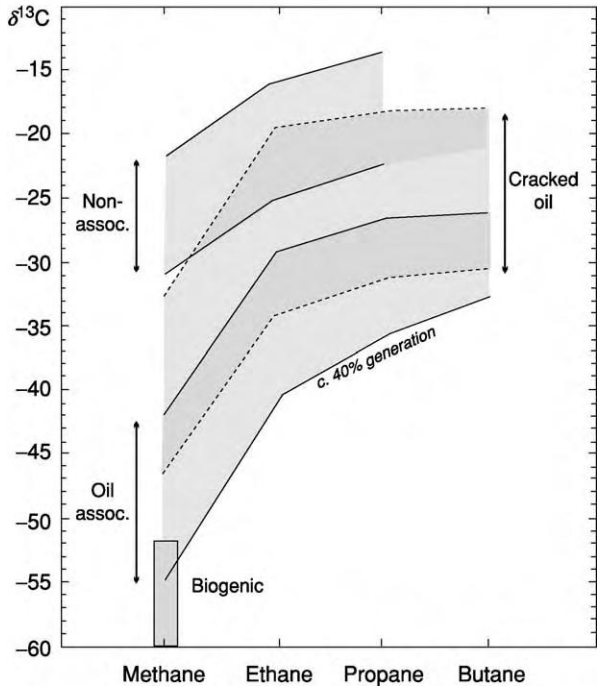


Figure 18 Isotope distributions in gases of different types.

gases by exsolution from flowing groundwater. However, by far the most important source is believed to be from kerogen breakdown (kerogen typically contains about 2% nitrogen by weight).

Table 2 Typical inorganic compositions of oil and gas fields (mol.%)

Area/field	CH_4	C_2 C_5	CO_2	N_2	H_2S
Oil associated					
Abu Dhabi/Zakum	76.0	20.3	2.3	1.1	0.3
Iran/Agha Jari	66.0	31.5	1.5	1.0	
N. Sea/Forties	44.5	53.6	0.6	1.3	
N. Sea/Brent	82.0	16.4	0.7	0.9	
Non associated					
Algeria/Hassi R'Mel	83.5	10.2	0.2	6.1	
France/Lacq	69.3	5.5	9.6	0.4	15.2
Holland/Groningen	81.3	3.5	0.9	14.3	Trace
New Zealand/Kapuni	46.2	7.9	44.9	1.0	
N. Sea/West Sole	94.4	4.0	0.5	1.1	

Nitrogen is generated from organic matter, particularly during the advanced stages of maturation. It has also been proposed that different types of kerogen yield different amounts of nitrogen. Experimental cracking of coals under laboratory conditions indicates that coals can generate up to 6% of nitrogen in total. Kinetic studies, based on laboratory experiments, suggest that gases containing more than 40–50% nitrogen are generated in excess of 300°C, after the main phase of hydrocarbon generation. These experiments imply that, in nature, gas with 50% nitrogen must have been formed by preferential trapping of late-stage generation products, whereby

the gas from the initial stages of generation has not been trapped.

Carbon dioxide Carbon dioxide (CO_2) has a number of possible sources, although most of these usually only contribute small quantities of gas. Although bacterial action produces some CO_2 under specific conditions, the most important sources are kerogen breakdown, inorganic reactions in the reservoir, and deep-seated CO_2.

CO_2 sourced from kerogen occurs as a result of thermal decarboxylation of kerogen. This occurs typically between about 50 and 100°C, before the main phase of oil and hydrocarbon gas generation. This reaction generally contributes only 1 or 2% CO_2 to the total gas. Inorganic reactions are limited to relatively hot reservoirs and are the result of chemical buffering between carbonates and silicates. These can either increase or decrease the CO_2 concentration, depending on the specific mineralogy involved and on the reservoir temperature and pressure. Nevertheless, many tens of per cent of CO_2 may result. Deep-seated CO_2 is often associated with igneous activity, but is also commonly the result of carbonate decomposition in the basement. This is probably the most important source of CO_2 in very high CO_2 (>30%) regions.

Hydrogen sulphide Hydrogen sulphide (H_2S) is an unpleasant, toxic gas occasionally found in large quantities in reservoirs. There are three main sources: bacterial sulphate reduction (of only limited occurrence), thermal cracking of kerogen or oil, and gas souring (thermochemical sulphate reduction). In addition, in an oil reservoir, H_2S preferentially resides in the water phase. This means that, even if there is little H_2S produced when the field is initially brought on stream, the field may subsequently sour as it goes into decline and the water cut increases.

A major source of H_2S is via the thermal cracking of kerogen or oil. Many compounds in oil contain sulphur, and the thermal degradation of these at very high temperatures leads to the formation of H_2S. Such a process becomes more important with increasing temperature. It is frequently associated with carbonate-sourced oils or kerogens as these are richer in sulphur.

The production of H_2S in deep, hot carbonate/evaporite reservoirs is probably the most important source of sour gas. Above approximately 140°C, the ratio of H_2S to methane often increases dramatically in proportion to temperature due to reaction between methane and gypsum or anhydrite, represented in its simplest form by

$$CaSO_4 + CH_4 \rightarrow CaCO_3 + H_2S + H_2O$$

However, in practice, numerous other parallel and intermediate reactions also occur.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Economic Geology**. **Petroleum Geology:** Overview; Gas Hydrates; The Petroleum System; Production; Reserves.

Further Reading

Hunt JM (1996) *Petroleum Geochemistry and Geology*, 2nd edn. New York: Freeman and Company.

Miles JA (1989) *Illustrated Glossary of Petroleum Geochemistry*. Oxford: Clarendon Press.

Peters KE and Moldowan JM (1991) *The Biomarker Guide*. Englewood Cliffs, NJ: Prentice Hall.

Tissot BP and Welte DH (1984) *Petroleum Formation and Occurrence*, 2nd edn. Berlin: Springer Verlag.

Vially R (ed.) (1992) *Bacterial Gas*. Paris: Edition Technip.

Gas Hydrates

M Hovland, Statoil, Stavanger, Norway

Introduction

Oceanic gas hydrates have attracted massive scientific and, not least, political attention over the past years. This is because of their dual role as a potential energy resource and as a geohazard/potential climate mediator. Interest in these compounds is reflected not only in the number of recent scientific projects aimed at studying and sampling oceanic hydrates, but also in the virtual flood of books, articles, and websites dedicated to these 'mysterious' substances.

Gas hydrates are crystalline, ice-like compounds composed of water and gas; the gas molecules are trapped within a cage-like framework of hydrogen-bonded water molecules. These structures are formed under very specific temperature and pressure conditions in an environment with adequate water and gas. The right conditions can be found on land (in polar regions, where surface temperatures are very cold) and within water or seabed sediments at depths exceeding 300–500 m and where the temperature is adequately low (generally <10°C). Mathews, in *The Log Analyst*, described hydrates as follows:

> Hydrates form when one species (known as the host) forms a crystal lattice containing voids or cavities that physically entrap other guest molecules. In natural gas hydrates, the host is water and the guest can be any common constituent of natural gas such as CH_4, C_2H_6, C_3H_8, C_4H_{10}, CO_2 or H_2S. Natural gas hydrates can have two possible structures, depending on the size of the guest molecules. Both structures consist of a network of water molecules hydrogen bonded to each other in a manner similar to water ice. (Mathews (1986))

Figure 1 shows a diagram of gas hydrates; the water molecules form a lattice containing 'guest' (gas) molecules – this is also called a clathrate substance in chemistry. The conditions under which gas hydrates are stable (Figure 2) are tightly controlled by temperature and pressure, but are also affected by the chemistry (i.e., including the salinity) of the pore water and the composition of the 'guest' gases. Chemicals (such as salts and methanol) that reduce the pressure/temperature envelope of hydrate stability are called hydrate inhibitors.

Because they tend to block pipes and flowlines in oil and gas fields, gas hydrates have been the focus of well-controlled laboratory studies for many years. Laboratory observations illustrate clearly how dynamic gas hydrates can be: at times these compounds are permeable and at times they present a complete barrier to flow. In a laboratory hydrate reactor at a Statoil facility in Norway, an initial attempt at forming hydrate plugs was carried out with a gas mixture of methane and isobutane, and tap water in the absence of hydrocarbon liquids. Hydrates were initially formed by injecting water droplets into the gas atmosphere at a pressure of 60 bar. A total of 13.5 litres of water was injected during 1.5 h at a subcooling temperature of 13°C. Gas was periodically circulated through the hydrate plug from the bottom of the reactor and the plug was periodically allowed to mature undisturbed for several days. After a static maturation period, an additional 4 litres of water was injected at the top of the plug. Most of the water was drawn into the plug, which acted as a sponge, and remained in the plug for the rest of the experiment. The plug remained porous throughout the entire 20-day-long experiment and grew to a length of 40 cm. When a pressure difference was introduced over the plug, the video recordings demonstrated that gas bubbles easily moved through the internal channels of the plug. Hence, the nature of the channels determined the permeability of the plug. Gas bubbled through the channels until more hydrates were formed and blocked further migration.

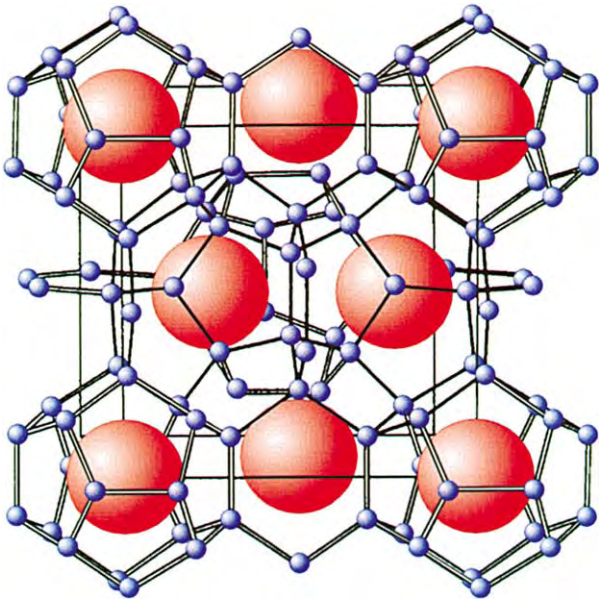


Figure 1 Diagram of a gas hydrate, showing how water molecules form a lattice around 'guest' gas molecules. The chemical name for this cage like solid structure is 'clathrate'. Diagram courtesy of Bjørn Kvamme.

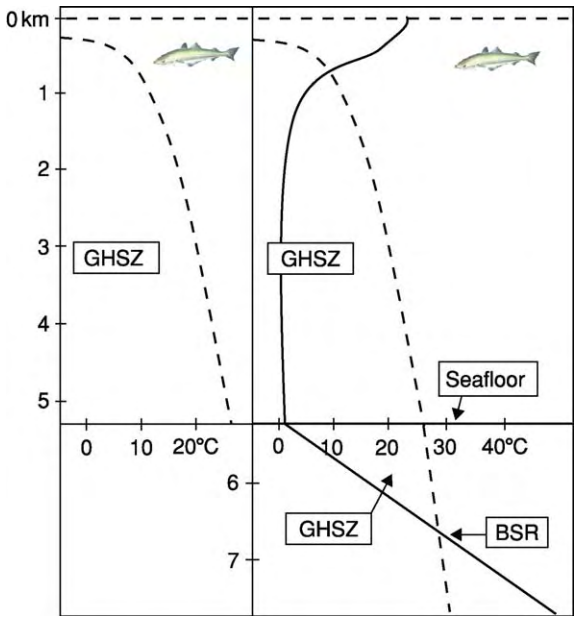


Figure 2 Methane hydrate stability field. The gas hydrate stability zone (GHSZ) in ocean water is determined by pressure and temperature. The stability envelope of gas hydrates in seawater is shown schematically on the left. Gas hydrates usually will not form at water depths of less than about 300 m and will not form in open seawater systems because of seawater circulation and the relatively high solubility of methane in seawater. However, in the confines of sedimentary pores in ocean bottom sediments (right hand diagram), gas hydrates will be stable and may form more or less permanently. Because of the increase in temperature with water depth (the geothermal gradient), hydrates will not be able to form beyond certain depths, indicated by the bottom simulating reflection (BSR), where the temperature gradient curve intersects the hydrate stability curve.

Figure 3 Core sample number 1249F 8h 2, from the Ocean Drilling Program Leg 204, drilling on Hydrate Ridge, the Cascadia Accretionary Prism, off the coast of Oregon, USA. The sample shows how gas hydrates (white) may occur in shallow, porous sediments. Photograph courtesy of Anne M Tréhu and the Shipboard Scientific Party, Ocean Drilling Program Leg 204, 2003.

The gas then found new channels until these were blocked as well. Thus, the plug gradually became less permeable. By measuring the gas released during melting and the water volume after complete dissociation, the plug content was calculated. The volume percentages of hydrates, free water, and free gas in the plug were 12, 38, and 50%, respectively. Hence, the actual hydrate content in the plug was low and the porosity of the plug was concluded to be as high as 88%.

In early work on gas hydrates, it was suspected that they were impervious to fluid flow, and that the entire sediment column within the gas hydrate stability zone (GHSZ) was cemented by gas hydrates. The active migration of free gas through a GHSZ and the porous nature of gas hydrates have since been amply demonstrated experimentally and are known to occur in nature. Gas hydrates may form rapidly when small-molecule gases leak into the sediments, and they may dissociate ('melt') rapidly when disturbed by heating or depressurization. When sediment-hosted hydrates dissociate, they return free gas and water to the environment, which may result in a total change of the physical sediment conditions and even of the seabed topography. Consequently, at water depths and sediment depths within the GHSZ, any planned interaction with the seabed has to take into consideration the possible existence and potential formation of gas hydrates.

When high-porosity sediments such as buried sands and gravels are present in regions with hydrocarbon flux, within the GHSZ, these sediment layers are likely to have developed into massive gas hydrate reservoirs. This has been demonstrated by the Mallik gas hydrate research well in the Mackenzie Delta, Canada. Also, with increasing hydrocarbon flux and increasing tectonic activity, the amount of gas hydrates within the near-surface sediments will increase, as has been demonstrated by, among others, the Geochemical Exploration and Research Group (GERG), in the Gulf of Mexico. In addition, the seafloor topography will be increasingly affected, and the presence of gas hydrates may even lead to very large slope failures and seafloor collapse events.

In regions with a very high fluid (hydrocarbon) flux through the seabed, masses of gas hydrates may form within the sediments, as found at Hydrate Ridge, off the coast of Oregon, USA (Figure 3). Here, seabed sampling showed that pure gas hydrates occurred in layers or in joints several millimetres to decimetres thick, generally oriented parallel to bedding, but sometimes cutting the bedding planes obliquely.

There was not one instance observed where the hydrates occupied the original sediment pore spaces and cemented the sediment framework on a microscale level. Rather, they filled large secondary pore spaces, either as fractures or joints, or seemed to create their own spaces during growth, by fracturing the sediment framework, most commonly along bedding planes. The internal fabric of the pure hydrates also showed a peculiar structure, with large, filled pores, resembling bubble-pack material.

Another important aspect of gas hydrates is that they chemically desalinate the ambient pore water. This characteristic is used by some geochemists to determine how much gas hydrate has been formed in low-permeability sediments, by studying the chlorinity of the pore waters. During Ocean Drilling Program (ODP) Leg 204 on Hydrate Ridge, this phenomenon was studied in detail. Among the most surprising findings of this leg was the rate at which hydrates form near the ridge summit. Salts are excluded when hydrates form, and the excess salt in the pore water diffuses away from the gas hydrates with time. However, if the hydrates form very rapidly, the salts do not have time to diffuse and the water in the sediment pore spaces will become saltier than seawater. Because of the hydrate inhibitor effect, gas hydrates will not be able to form within such brine-filled sediment pockets in the future, another factor that causes sedimentary inhomogeneities.

Detection and Distribution of Gas Hydrates

It is relatively easy to map the 'potential gas hydrate' distribution (i.e., the distribution of locations in which suitable physical conditions occur in the ocean sediments). However, the expected distribution frequency may be overestimated because, even within the GHSZ, gas hydrates will form only if there is adequate water within the sediments to form the 'host' molecules, and an adequate supply of gas 'guest molecules' (for example, methane). The origin of the gas is immaterial, so methane, being the most common gas, may be of either thermogenic or bacterial origin.

Indirect evidence for gas hydrates in ocean-floor sediments first came from anomalous reflections observed on seismic profiles recorded during the Deep Sea Drilling Program (DSDP), Leg 11, on the Blake Outer Ridge, off the eastern coast of the USA (*see* **North America:** Atlantic Margin). The special reflection event (Figure 2), called a bottom-simulating reflection (BSR), is an acoustic interface generated by a (diagenetic) phase boundary. It is anomalous owing to its highly reflective appearance, representing an abrupt change in acoustic impedance. It is also anomalous in that it cross-cuts bedding planes and is of reverse acoustic polarity. Sedimentary strata below the BSR are usually more strongly defined (i.e., enhanced) compared to those above it, where lithological layers tend to generate only weak reflections. This upper zone is therefore referred to as 'seismically transparent' or as 'blanking' and is suspected to be caused by hydrate 'cementation' and relatively high acoustic velocity.

Although the mapping of BSRs can aid in determining the presence of gas hydrates below the seafloor, some areas with gas hydrates do not have BSRs. It was long thought that a BSR was a necessary and adequate indicator of *in situ* gas hydrates, but one important result of the dedicated ODP BSR investigations is that this assumption was found to be partly false. In some places, drilling through BSRs has failed to find gas hydrates; in other cases, gas hydrates have been found where there is no BSR.

A prominent BSR occurring adjacent to the Storegga Slide scarp, off the coast of Norway, was investigated by drilling in 1997, as part of the geotechnical investigations for field development of the Ormen Lange field. However, no gas hydrates or temperature anomalies were found in the sampled sediments. Furthermore, only very little methane occurred in the pore waters below the BSR. These 'negative' results indicate that the existence of a BSR is no guarantee that there are detectable amounts of gas hydrates or free gases in the sediments. The question about what actually produces the seismically detectable BSR is therefore valid. Off the coast of Norway, it was concluded that the feature perhaps represents a palaeo-BSR, and that although the gases and gas hydrates may once have been present, they are no longer there in quantities that are of any practical consequence.

The true nature of BSRs and how they develop are therefore not yet fully understood. There is some consensus that they develop mainly as a consequence of the acoustic contrast between relatively high-velocity, gas-hydrate-charged sediments above low-velocity sediments. The low-velocity characteristic is attributed to free gas beneath a hydrate layer. A BSR may therefore be indicative of gas hydrates above the BSR, free gas beneath the BSR, and the diffusive migration of gas through the whole sediment package, i.e., from below the GHSZ to the seabed.

One of the most surprising results from ODP Leg 164, at Blake Ridge, was the finding of bacterial activity near the BSR. This finding is thought to demonstrate that the sediments near and below the BSR form a biogeochemically dynamic zone, with carbon cycling occurring through methane, acetate, and carbon dioxide. At Site 995, pore water acetate was present in very high concentrations, reaching

~15 mmol l^{-1} at 691 mbsf (metres below seafloor), ~1000 times higher than 'typical' near-surface concentrations (2–20 μmol l^{-1}). Potential rates of acetate metabolism were extremely high and could not be sustained without influx of organic carbon into the sediment; hence, *in situ* rates are likely to be lower than these potential rate measurements indicate. However, there is evidence for upward migration of high concentrations of dissolved organic carbon into the sediments at these sites. Based on these observations, it can be concluded that the BSR represents a physicochemical and perhaps even a biogeochemical 'reaction zone' brought about by the advection of light hydrocarbons; the hydrocarbons accumulate as free gas below the local GHSZ and form hydrates above – thereby rendering an acoustic reflector that can be imaged by seismic methods.

In 2001, Kvenvolden and Lorenson provided an updated overview of worldwide gas hydrate occurrences:

> An inventory of global occurrences shows 77 places in which the presence of gas hydrate is inferred by geophysical, geochemical, and geological methods. This inventory includes 19 places where samples of the substance have actually been recovered. Most natural gas hydrate contains methane that was generated by microbial processes. Methane with a thermal origin occurs in gas hydrate only where deeper deposits of thermogenic natural gas provide a source of the methane. The potential amount of methane in global gas hydrate occurrences is very large, with current estimates converging at about 10 teratonnes of methane carbon.

Gas Hydrates as an Energy Source

It has long been speculated that methane hydrates in ocean sediments may represent the greatest reservoir of methane on the planet, offering a clean and abundant energy source for the future. Historically, there has been a great disparity between various estimates of methane locked up as hydrates in ocean sediments. This can be explained by the fact that previous estimates had assumed that the gas hydrate-bearing sediments contained massive hydrates, whereas sampling has shown that hydrates normally occur as small, finely dispersed crystals and aggregates within the sediments; average concentrations may be as little as 3% by volume – more like the ore of a metallic mineral than a traditional hydrocarbon reservoir. But integrated over the world's oceans and Arctic tundras, the total methane stored in gas hydrates is about 10^{19} g, i.e., about 10 Mt, which represents an order of magnitude more carbon than exists in the biosphere, according to Nisbet.

In the Outer Niger Delta, potential BSR-associated gas hydrates have been considered for production

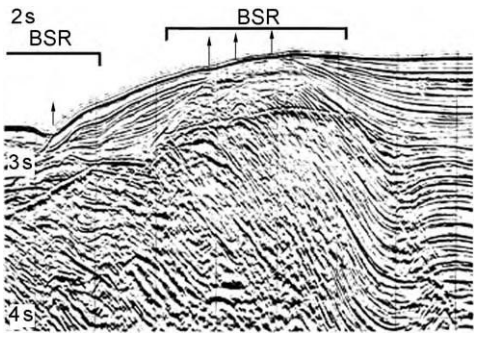


Figure 4 The bottom simulating reflection (BSR): Four of the main BSR characteristics are evident on this 2D seismic record from off Nigeria: (1) depth below the seafloor is more or less constant, (2) the BSR is anomalously strong, (3) the BSR cross cuts bedding planes, and (4) the BSR is of reverse polarity with respect to the seafloor reflection. These four aspects, respectively, mean that the BSR (1) is temperature dependent rather than pressure dependent, (2) is caused by a relatively large impedance contrast (i.e., jump in density and/or sound velocity with depth), and (3) is fluid/gas related (i.e., diagenetic rather than lithologically related), and that (4) the P waves reflect from a boundary with an inverse impedance contrast (i.e., where the density is reduced and/or the material's sound velocity is reduced with depth).

(Figures 4 and 5). However, their suspected dispersed nature suggests that they could be extracted only by open-pit mining, a method not currently viable, because the water depth is between 1000 and 2000 m! But attempts to find commercial oceanic gas hydrates have not yet been abandoned. There is hope of finding thick deposits of sandy turbidites containing massive methane hydrates at a 900-m water depth in the Nankai Trough off the coast of Japan. Japanese scientists and the Ministry of Economy, Trade, and Industry (METI) will be going ahead with a three-dimensional seismic acquisition and a full-scale exploration drilling campaign in the Nankai Trough over the next few years.

Oceanic methane hydrate deposits may be of strategic interest. Even if the growth of fossil-fuel consumption continues to slow in Europe, North America, and Japan as conservation and alternative energy sources come to play larger roles, the rest of the world inevitably will require much more energy than it now uses. Creating the means to exploit gas hydrates could be a matter of national security. This is why Japan has long devoted a substantial effort to studying natural-gas hydrates, including the drilling and analysis of a number of test wells. Strategic planning also accounts for the rising interest in hydrates within the US Department of Energy.

Several years ago, Japanese researchers teamed up with Canadian and US scientists to explore commercial exploitation of gas hydrates by drilling through

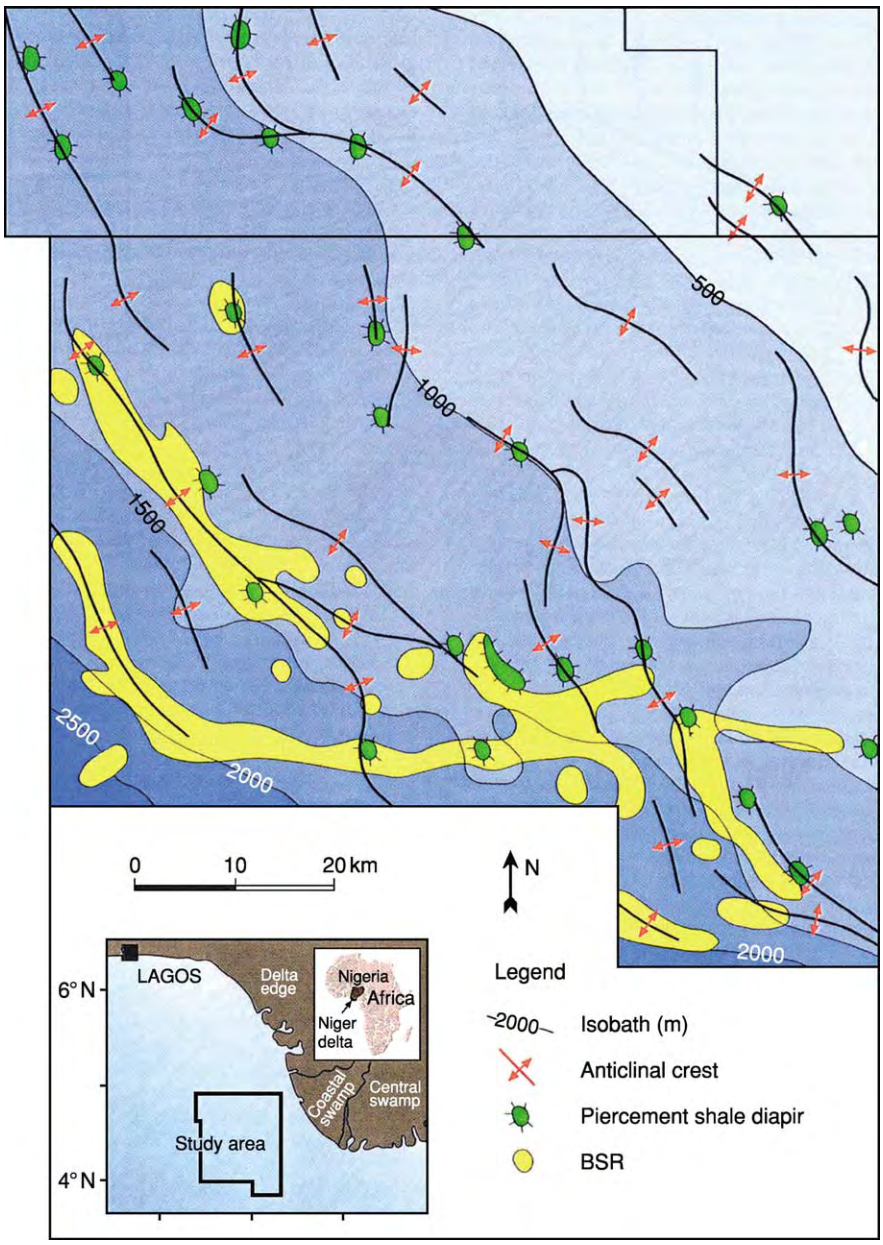


Figure 5 Map showing an area in the Outer Niger Delta Front where resource estimates of gas hydrates have been made. The yellow regions define areas with bottom simulating reflections (BSRs), i.e., strong gas hydrate indicators. The green circular features represent mud volcanoes, where fluids and mud erupt on the seafloor and build low lying sediments cones. Depth contours are in metres. Note that the BSRs occur only in water depths greater than about 900 m.

known, thick hydrate occurrences in Arctic Canada, the so-called Mallik well. In 2001, this project was expanded into a multinational campaign, including personnel from India and Germany. They have investigated both warm water circulation and pressure release as production means of freeing up the subsurface hydrate-locked gas. In 2002, Adam wrote that

> Those involved said that they were encouraged by the amount of methane produced enough to ignite a flare similar to those seen burning over oil rigs. But whether the yellow flame is merely symbolic or a genuine step forwards remains to be seen. The researchers are confi dent that they will get more energy out than they put in, but not [enough] to pipe it to a power station.

The most promising areas to look for available gas hydrate-associated energy are probably at active deep-water mud volcanoes. Large volumes ($\sim$25 Mt year ^{1}) of gas escape from deep-water mud volcanoes, suggesting that global gas flux from the

seafloor may be underestimated. The European Community is currently investing money into research on gas hydrates in order to assess the potential benefits of or hazards posed by these resources. As part of a European research programme (HYDRATECH), a specially designed laboratory apparatus (a resonant column) is to be constructed to simulate the conditions that allow gas hydrates to be grown artificially. The resonant column will also be used in experiments conducted to characterize the seismic signature of sediments containing gas hydrates. The results of the experiments will allow scientists to better identify and estimate the uses of or threats posed by gas hydrates. However, because even one of the world's best mapped and documented seabed areas still keeps its geological secrets well hidden, the race for exploitable hydrates may take a long time. As expressed by Lorenson and co-workers in 2002:

> Despite years of geophysical prospecting for hydrocar bons, the spatial and vertical distribution of deep gas hydrate in the Gulf of Mexico is not well known. This uncertainty hinders both the determination of the economic potential of gas hydrate resources in the region . . . and their potential as a geohazard.

Gas Hydrates as a Geohazard and Climate Mediator

Extensive data collection efforts on the Louisiana continental slope in the Gulf of Mexico have demonstrated that ambient gas hydrates in sediments are sensitive to small temperature and pressure changes. A temperature increase in the water column as small as 1.0–1.5°C caused distinct dissociation of the gas hydrate, leading to outgassing. Smaller fluctuations in fluid flow were also recorded at the tidal frequency, implying that very small perturbations in the ambient pressure conditions may lead to dissociation. The assessment of a potential drilling site or deep-water construction site, including potential pipeline routes, therefore, has to include an evaluation of the possible presence and formation of gas hydrates. It is therefore necessary to use and interpret all of the available signs of gas hydrate presence, including indirect means, such as interpretation of seismic, sonar, and topographic features, and direct ground-truthing means, such as visual observation and seabed sampling. Ground truth is required to calibrate wide-area surveys and derived results. Within a large dataset, small areas must be examined at the highest resolution and geotechnical standards. The interpretation of the wide-area dataset can then be constrained and the data made relevant within the whole of the wide-area survey. In addition, small areas that have been selected for exploration drilling, or because they constitute a representative seafloor, require optical inspection using video cameras; inspection should be made over relatively wide areas and for every sampling point.

An assessment of potential gas hydrates should also go hand in hand with a theoretical consideration and analysis. Because the regional and local diffusive and focused flux of light hydrocarbons through the sediments is of such importance, it may also be important to quantify the hydrocarbon flux from the seafloor into the ocean, and because of the dynamic nature of the hydrates, long-term monitoring of subcropping or outcropping hydrates could be useful at certain deep-water construction sites. Experiments for performing such studies in the Gulf of Mexico are being planned.

Recent interpretation of high-resolution seismic records of seafloor sediments on the Cascadia margin, off the coast of Vancouver Island, Canada, suggests a periodic disintegration of oceanic gas hydrates. Thus, vertically oriented acoustically anomalous 'chimneys', also called 'wipeouts', which are perturbations penetrating through the GHSZ, indicate fine-scale fluid flux through the upper seafloor. It has been suggested that these vertical intrusions are associated with upward flow of warmer fluids. Therefore, where seafloor fluid expulsion and methane hydrate deposits coincide, the base of the hydrate stability zone might exhibit significant roughness and increased surface area. Increased area implies that significantly more methane hydrate is close to being unstable and hence is closer to dissociation in the event of a lowering of pressure due to sea-level fall. On the Norwegian continental slope, north-east of the Storegga slide scarp, similar features are common on multichannel and high-resolution seismic records. In this region, many of the columnar 'chimneys' penetrate up to the seabed surface and are clearly associated with large pockmarks containing heaps of methane-derived limestone blocks.

'The clathrate gun' hypothesis, developed by Kennett and co-workers, is based on sediment sampling in the Santa Barbara Basin during ODP Leg 146B. Abrupt changes in the $\delta^{13}C$ values for marine organisms were observed in the samples, and it was speculated that the reason was abruptly changing methane contents in the atmosphere and oceanic water due to the dissociation of gas hydrates in ocean sediments. In association with the publication of the hypothesis, James Kennett made the following conclusions on this matter:

> Although no single, unequivocal proof exists for the hypothesis we present, a broad range of evidence suggests it. We expect reluctance by scientists to support such an

idea at this stage without further testing, but if the hypothesis holds, it will require a major shift in thinking about what drives Quaternary climate change.

However, it seems Kennett does not need to worry about reluctant scientists, because recent geological and palaeoclimatic results support the notion that methane released from oceanic gas hydrates is the main culprit in rapid warming events in the past. Smith and co-workers, for example, in 2003, reported suspected massive methane release during deglaciation, only about 14 000 years ago. Their conclusions are based on high-resolution stable isotope records from marine cores on the East Greenland shelf; these records exhibit three rapid light $\delta^{13}C$ events (-3 to -7%) in benthic and planktic foraminifera that are spatially and temporally transgressive. Smith and colleagues invoke that a combination of pressure release (due to ice melting) and temperature increase (associated with ice shelf retreat) has caused large masses of oceanic methane hydrates to dissociate suddenly. A similar but much more severe episode occurred about 55 million years ago, during the Paleocene/Eocene thermal maximum (PETM) global warming event. Based on seabed coring near Antarctica, researchers concluded that there must have been a massive release of carbon-12 into the ocean and atmosphere – an injection that must have occurred because of methane hydrate destabilization.

The main challenge now is to find out what triggers such destabilization. There are several possibilities, including ocean warming ('melting' of oceanic hydrates), lowering of mean sea-level and submarine landslides (depressurizing oceanic hydrates), and large impact events, all of which are equally difficult to document. Perhaps the triggering mechanism is quite feeble, such as a very slight warming or a very slight lowering of sea-level, but at some point, the balance is tipped, leading to a cascade of events. It is known, for example, that there are thousands of submarine mud volcanoes located on the ocean floor. These are known to represent 'pressure release valves' that may suddenly erupt violently, if disturbed by earthquakes, etc. The most dramatic and freak observation of an underwater mud volcano blowout (violent eruption) occurred in the Caspian Sea, on 18 November 1958. By chance, the event was photographed from the city of Baku, 20 km away from the eruption site, the Bachar Bank mud volcano, located at a water depth of 150 m. The flaming jet of gases at first shot several km into the sky. Later, it was estimated to stabilize at a height of 500 m and was topped by a large, 120-m-long horizontal flame. Judging from these dimensions, the jet must have been issuing at supersonic velocity from the sea surface, and the pressure must have been released directly from a depth several thousand metres below the seabed. It is known that the sea-level of the Caspian Sea is about 28 m lower than mean sea-level (MSL). Furthermore, the Caspian Sea drains the Volga River system and is known to adjust its level over long time-spans of 20 to 40 years, in response to suspected long-period, gradual, and cyclic climate changes. The Caspian Sea experienced a high-stand in the late 1930s, followed by a low-stand in 1974, which was followed by a new high-stand in 1996. Therefore, the triggering of the Bachar Bank natural blowout event could possibly have been the gradual (feeble) lowering of the Caspian sea-level. Although mud volcanoes are associated with gas hydrates, their formation mechanism is generally poorly understood.

All-in-all, it may be concluded that sediment-hosted gas hydrates and free gas located below capping layers of hydrates are dynamic, interesting, natural substances that may promise future wealth, but also wrath – if not treated with respect.

See Also

Atmosphere Evolution. **Economic Geology**. **Engineering Geology:** Natural and Anthropogenic Geohazards. **North America:** Atlantic Margin. **Palaeoclimates**. **Petroleum Geology:** Exploration.

Further Reading

Adam D (2002) Fire from ice. *Nature* 415: 913 914.

Austvik T, Li X, and Gjertsen LH (1999) *Hydrate Plug Properties Formation and Removal of Plugs*. Third International Conference on Natural Gas Hydrates, Salt Lake City, July, 1999.

Dallimore SR, Uchida T, and Collett TS (1999) Summary. In: Dallimore SR, Uchida T, and Collett TS (eds.) *Scientific Results from JAPEX/JNOC/GSC Mallik 2L 38 Gas Hydrate Research Well, Mackenzie Delta, Northwest Territories, Geological Survey of Canada Bulletin*, vol. 544, pp. 1 10. St. John's, NL: Geological Survey of Newfoundland and Labrador.

Ginsburg GD and Soloviev VA (1998) *Submarine Gas Hydrates*. St. Petersburg, Russia: VNIIOkeangoeolia.

Henriet J P and Mienert J (1998) *Gas Hydrates: Relevance to World Margin Stability and Climate Change. Geological Society Special Publication No. 137*. London: Geological Society of London.

Hovland M, Gallagher JW, Clennell MB, and Lekvam K (1997) Gas hydrate and free gas volumes in marine sediments: Example from the Niger Delta front. *Marine and Petroleum Geology* 14(3): 245 255.

Hovland M and Gudmestad OT (2001) Potential influence of gas hydrates on seabed installations. In: Paull CK and Dillon WP (eds.) *Natural Gas Hydrates, American*

Geophysical Union Monograph, vol. 124, pp. 300 309. Washington, DC: American Geophysical Union.
Kennett J, Cannariato KG, Hendy IL, and Behl RJ (2003) Methane hydrates in Quaternary climate change: The clathrate gun hypothesis. *American Geophysical Union Special Publication 54*. Washington, DC: American Geophysical Union.
Kleinberg RL and Brewer PG (2001) Probing gas hydrate deposits. *American Scientist* 89: 244 251.
Kvenvolden KA and Lorenson TD (2001) Introduction. In: Paull CK and Dillon WP (eds.) *Natural Gas Hydrates, American Geophysical Union Monograph*, vol. 124, pp. 5 7. Washington, DC: American Geophysical Union.
Lorenson TD, Winters WJ, Hart PE, and Paull CK (2002) Gas hydrate occurrence in the northern Gulf of Mexico studied with giant piston cores. *EOS, American Geophysical Union Transactions* 83(51): 601, 607.
Mathews M (1986) Logging characteristics of methane hydrate. *The Log Analyst* May June: 26 63.
Milkov AV and Sassen R (2002) Economic geology of off shore gas hydrate accumulations and provinces. *Marine Petroleum Geolology* 19: 1 11.
Nisbet EG (2002) Have sudden large releases of methane from geological reservoirs occurred since the Last Glacial Maximum, and could such releases occur again? *Philosophical Transactions of the Royal Society of London, Series A* 360: 581 607.
Sloan ED (1998) *Clathrate Hydrates of Natural Gas*, 2nd edn. New York: Marcel Dekker.
Soloviev VA (2002) Global estimation of gas content in submarine gas hydrate accumulations. *Russian Geology and Geophysics* 43: 609 624.
Suess E, Bohrmann G, Greinert J, and Lausch E (1999) Flammable ice. *Scientific American* Nov.: 52 59.
Wood WT, Gettrust JF, Chapman NR, Spence GD, and Hyndman RD (2002) Decreased stability of methane hydrates in marine sediments owing to phase boundary roughness. *Nature* 420: 656 660.

The Petroleum System

C Cornford, Integrated Geochemical Interpretation Ltd, Bideford, UK

Introduction and Definitions

The term 'petroleum system' implies an understanding of the generation of oil and gas within a source rock, followed by their expulsion, migration, entrapment, and survival to the present day. The petroleum system approach focuses on the 'genesis' of petroleum and hence the 'source rock', rather than merely the current location of oil and gas and hence the 'trap'. This differentiates the petroleum system approach of the 1990s from the 'play' concept of the 1970s and 1980s (Figure 1).

A simple definition of a petroleum system might be 'all commercial oil and gas deriving from a single pod of mature source rock'. With a better understanding of efficiencies and losses in the processes, this definition has developed to 'a petroleum system accounts, in terms of *volumes*, *composition*, and *process efficiencies*, for all commercial oil and gas expelled from a single pod of mature source rock'. What constitutes a 'commercial accumulation' varies with circumstances, such as oil composition, drilling costs, oil prices, political stability, and proximity to market.

Use of the petroleum system concept accepts that commercial oil and gas is generated from organic matter (kerogen) in sedimentary rocks. These rocks are at or have been buried to temperatures in the range 80–165°C (oil) or 145–220°C (gas), and the petroleum is expelled from the source rock, migrates, and then accumulates in traps. The term is inapplicable if an inorganic origin of petroleum is adopted and when exploring for biogenic methane, and unhelpful when exploring for coal-bed methane and shale-gas (gas trapped intimately within seam coal and shales, respectively).

A discussion of petroleum systems can be broken down into the contributing processes, viz:

- Source rock deposition and kerogen accumulation;
- Maturation of source rock upon burial;
- Generation of hydrocarbons in the source rock;
- Expulsion of oil and gas out of the source rock;
- Migration of oil and gas to the reservoir; and
- Alteration of reservoired hydrocarbons during survival to the present day.

These processes are illustrated gastronomically in Figure 2 and discussed in more detail in the subsequent sections.

In mapped terms (Figure 3), the prospect or trap is surrounded by a drainage area (polygon), from which oil generated within and expelled from mature source rock is focussed during migration towards the proposed accumulation. As illustrated, the actual amount of oil or gas generated will be determined by the richness of the source rock (type of organic matter, e.g., marine or terrigenous kerogen) and the maturity level reached (e.g., oil- or gas-mature). To undertake a full volumetric calculation, the source rock yield and thickness is also required.

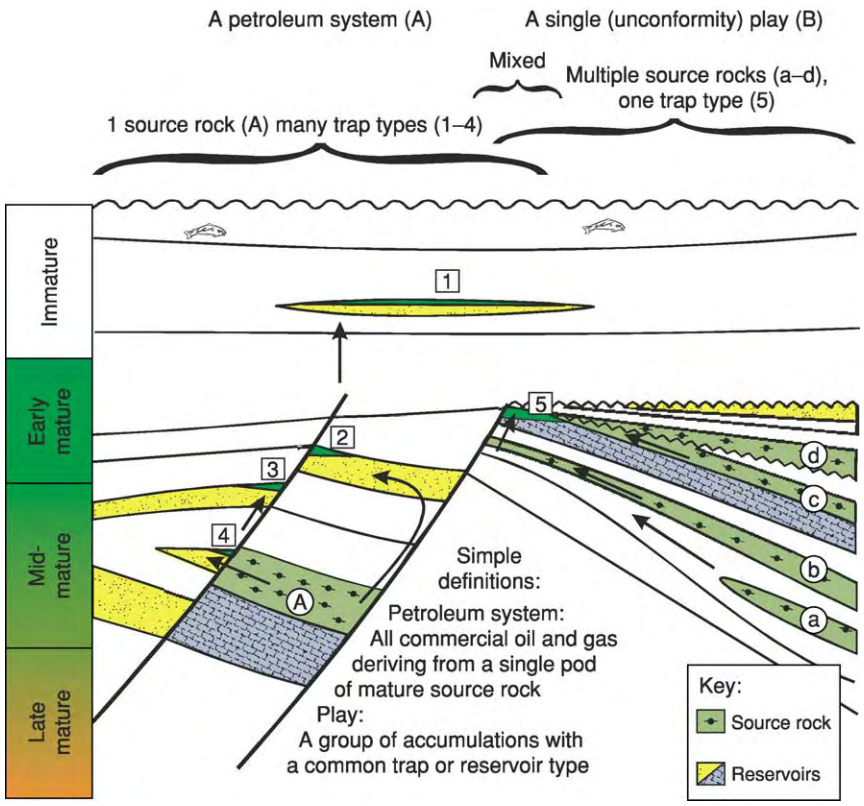


Figure 1 Contrasting the petroleum system with the play concept used for hydrocarbon exploration.

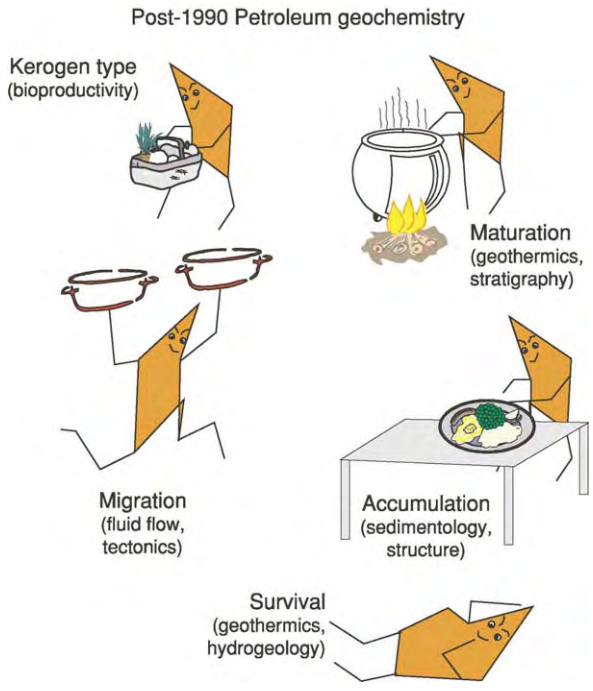


Figure 2 The petroleum system seen as a series of processes starting with (sedimentary) organic matter and resulting in a satisfyingly long lasting accumulation! Note the proceedings can be seen in terms of geological processes (text labels) or molecular activities (cartoon figures).

Source Rocks

Oil and gas are generated in organic-rich sediments, the source rock. As sediments accumulate, the soft body parts of organisms settle out with the mineral grains and carbonate or siliceous skeletal debris. The soft body parts are dominantly protein, carbohydrate, and fats, and in the case of the higher land plants, lignin. On bacterial attack and initial burial, these tissues condense to form kerogen. Most aquatic or littoral sedimentary environments have an abundant supply of organic matter, but petroleum source rocks require, in addition, a depositional environment promoting the preservation of that organic matter.

Source Rock Deposition

Depositional environments favouring the accumulation of petroleum source rocks are illustrated in Figure 4 in the context of a section through a tectonic plate. The amount of kerogen present in sediments is a balance between bioproductivity, survival, and dilution by inorganic grains. To take some extremes:

Low productivity—Aeolian (wind-blown) desert sandstones where land plant growth is restricted due to the extremely arid environment.

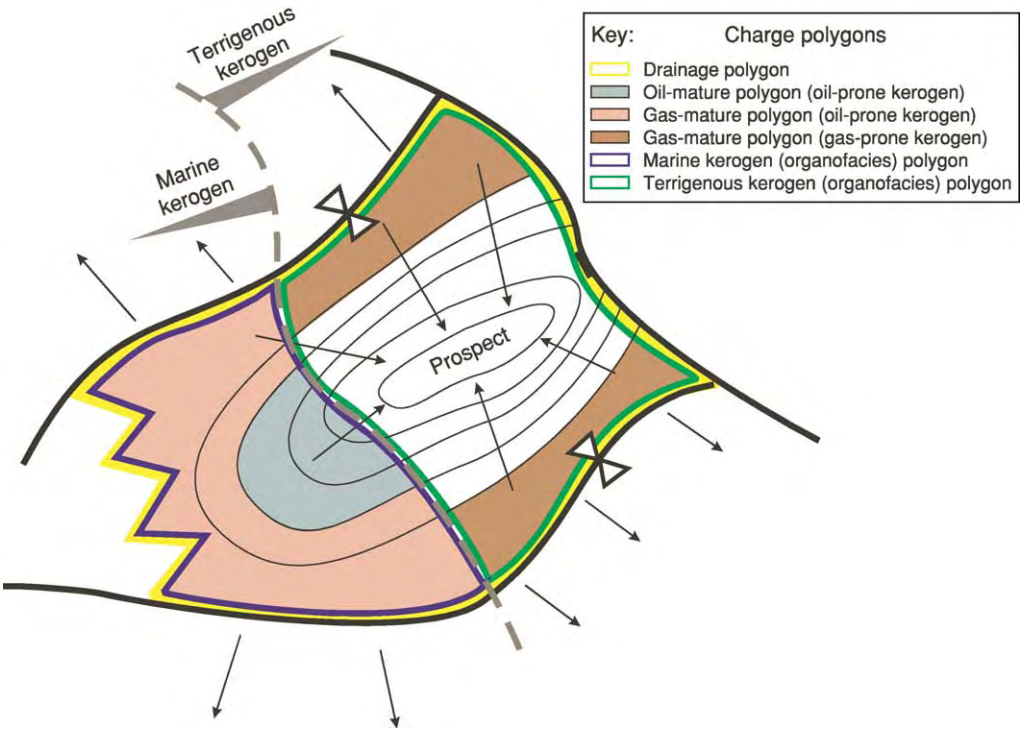


Figure 3 Mapping a trapping structure and the surrounding source rock drainage polygon supplying the oil and gas. The overall drainage area is bounded by faults and synclinal and distal boundaries and contains maturity kitchen and kerogen type polygons.

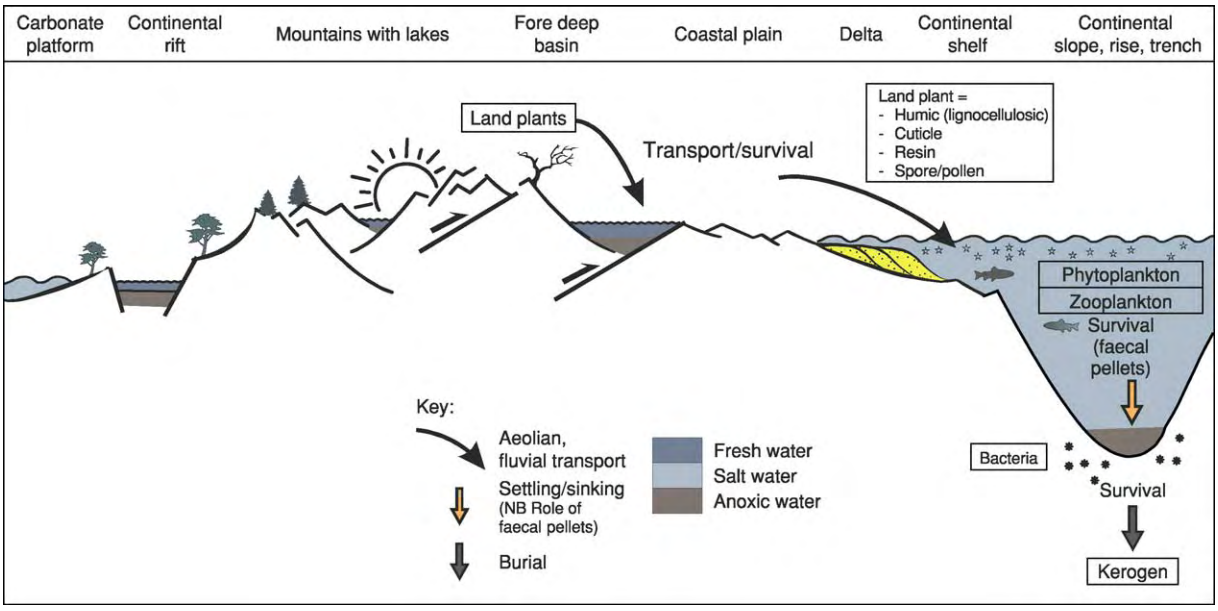


Figure 4 Environments where bioproductivity and depositional environment favour the accumulation of organic rich sediments.

High productivity—Areas of ocean upwelling where the enhanced supply of nutrients fuels explosive growth of phyto- and zooplankton.

Poor survival—Chalk composed almost exclusively of coccolith skeletal debris (high bioproductivity) but where all the coccolith body tissues are destroyed by bacterial activity under strongly oxic open marine conditions.

Excellent survival—Early rifting phases of oceans where both optimum sedimentation rates and anoxic conditions of the lakes and enclosed seaways promote high rates of organic matter preservation.

Strong dilution—The prolific sediment supply of major deltas produces organic lean delta-front and pro-delta sediments despite high bioproductivity.

Minimal dilution—Coals where the lignocellulosic and other tissues of high-productivity land plants are well preserved in a delta-top environment starved of (or bypassed by) mineral grains.

As illustrated above, the aspects of the depositional environment favouring organic preservation are anoxia and elevated sedimentation rates, though excessive sedimentation rates will eventually lead to dilution.

As implied in Figure 4, organic matter falling to the ocean or lake floor has to pass through various zones of bacterial degradation (Figure 5). The normal open-water oxic bacterial community is highly effective at destroying oil-prone organic matter, so time spent in this zone has disastrous effects on potential oil source rocks. These conditions are found in the water column and surface sediments in the open ocean, but in strongly stratified basins (e.g., the present-day Black Sea) only the top of the water column may be oxic.

Thus to summarize (Figure 6), the optimum oil-prone source rocks are deposited where anoxia develops in an aqueous environment enjoying high rates of sedimentation. The combination of high sedimentation rates and oxic environments favours the accumulation of gas-prone source rocks such as coals. The combination of depositional conditions and kerogen amount and type is termed 'organofacies', as discussed in the next section.

Source Rock Characterization

For a sediment to be considered a potential petroleum source rock it has to contain or have contained an appropriate amount and type of organic matter

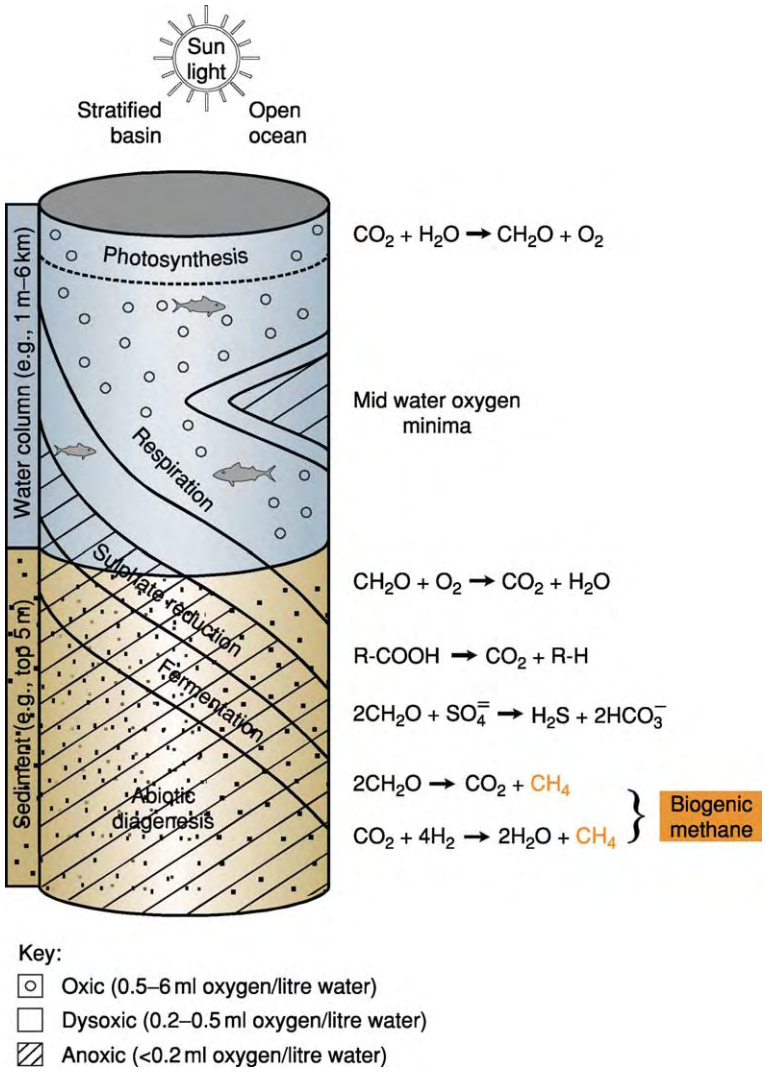


Figure 5 Processes affecting organic matter deposition and preservation.

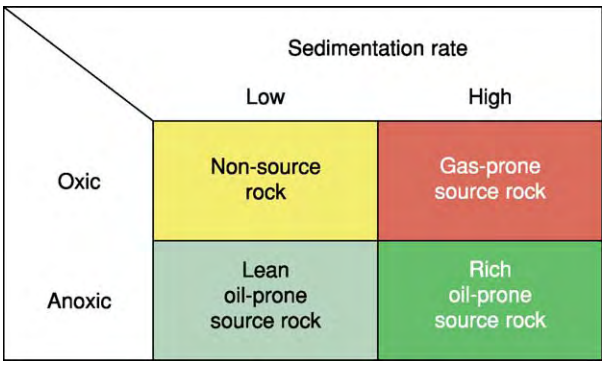


Figure 6 Summary of major processes controlling petroleum source rock accumulation (red gas prone; green oil prone).

(kerogen). In the same way that the mineral grains of a sediment can be linked to the conditions of deposition via the 'lithofacies', kerogens can be linked to depositional conditions via the 'organofacies'. An organofacies is characterized by a specific combination of organic matter abundance and type.

Abundance of organic matter The amount of kerogen is determined as the total organic carbon (TOC), measured as the weight percentage of the rock (Figure 7). The amount of organic matter can then be described as fair, good, or very good, as suggested in Figure 8. Experience has indicated that carbonate source rocks require lower TOC values to reach a certain level of source potential, probably due to more efficient expulsion of any generated hydrocarbon.

Type of kerogen—optical methods The most graphic way to determine the types of organic matter comprising kerogen is to view kerogen using optical microscopy. This technique is termed 'organic petrography' and is a combination of coal petrographic and palynological techniques and nomenclature. For a full organic petrographic characterization, three types of microscopy are needed:

- Reflected light (also used for vitrinite reflectance);
- Transmitted white light (also used for palynology); and
- Ultraviolet (UV) excitation fluorescence.

The reflected light technique is also the basis for the common maturity parameter, vitrinite reflectance, whereas kerogen colour in transmitted light and fluorescence colour can also contribute to maturity estimations (see next section).

A view of the same area of a bituminous coal thin section using all three methods is shown in Figure 9 with the logic for classification shown in Figure 10.

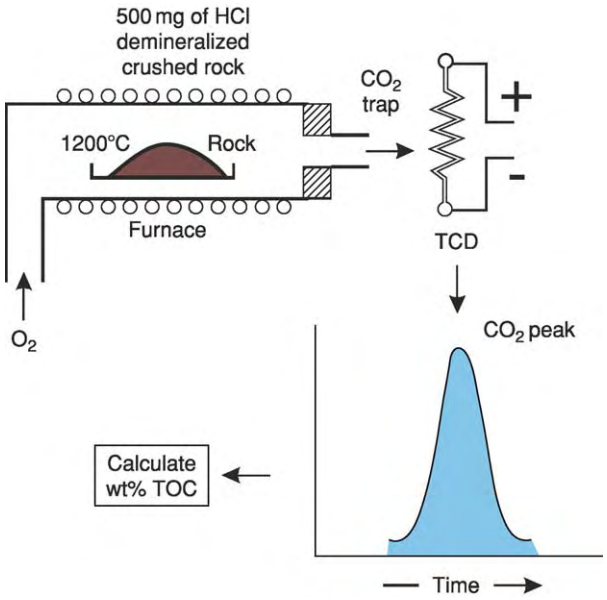


Figure 7 Schematic of a carbon analyser used to determine total organic carbon (TOC wt%) by burning acid demineralized rock powder in oxygen in a furnace at 1200°C, and determining the amount of carbon dioxide (CO_2) liberated.

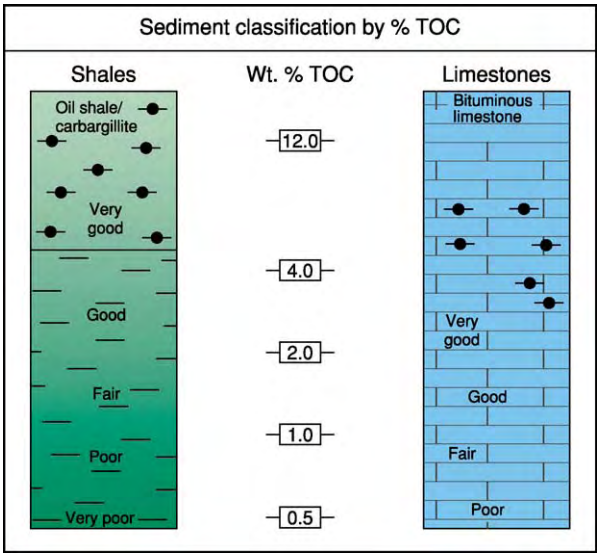


Figure 8 Quality of petroleum source rocks based on amount (but not type) of organic matter.

Discrimination using fluorescence can only be made for relatively shallowly buried sediments (e.g., <4 km burial) where the kerogen is in the immature to mid oil-mature levels of maturation. From Figure 9 it can be seen that this coal contains oil-prone spores and cuticle, gas-prone lignocellulosic vitrinite, together with inertinite (fungal sclerotinite and semifusinite)

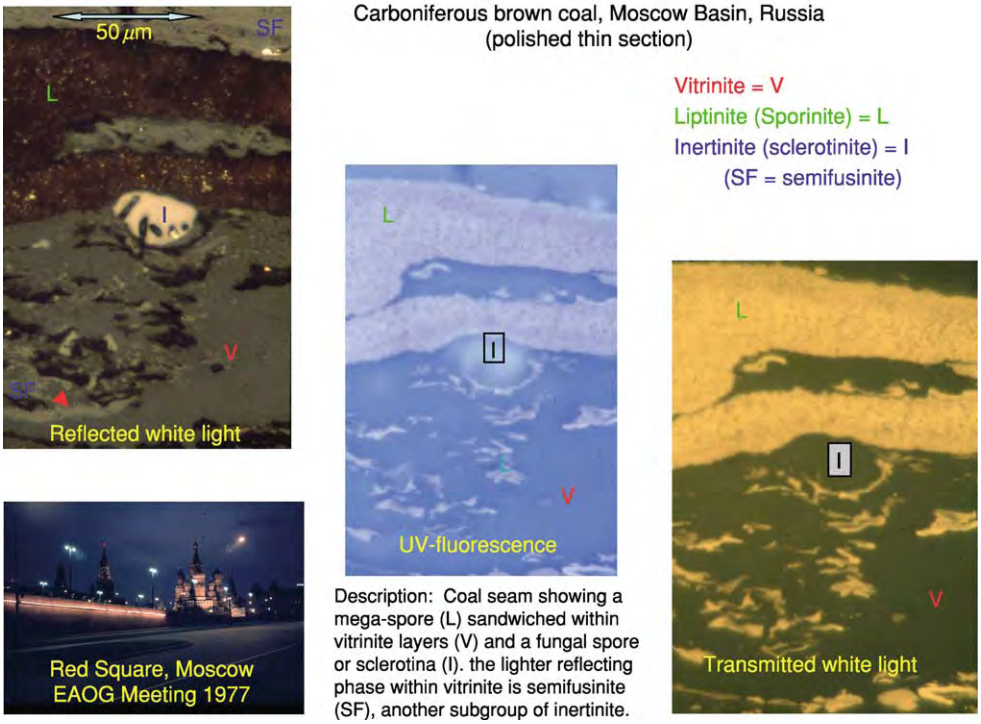


Figure 9 Reflected, transmitted, and fluorescent light views of a polished thin section of Russian carboniferous coal showing oil prone lipinite, gas prone vitrinite, together with inert kerogen macerals (inertinites).

with little or no hydrocarbon potential. The inert material is the result of oxidizing land plant material, this mainly being fossil charcoal resulting from natural forest fires and spontaneous combustion of humus.

The combination of these microscopy techniques allows the visible particulate organic matter to be classified into three groups:

- Oil-prone liptinites (fluorescent at low maturity and low reflectance);
- Gas-prone vitrinites (nonfluorescent and medium reflectance); and
- Inert or dead carbon (nonfluorescent and high reflectance at all levels of maturity).

The decision tree used by the organic petrographer is illustrated in Figure 10, where the initial distinction between fluorescent oil-prone liptinites and the nonfluorescent components is followed by high and low reflectance to split the latter into gas-prone and inert particles. Occasionally some vitrinites can exhibit low levels of flourescence. Some further breakdown of the three primary groups is also shown.

Visual descriptions (or a photograph) of a kerogen provides an excellent overview of the large variety of plant and animal tissues preserved in the sediment. However, this diversity of kerogen types also gives rise to a bewildering range of descriptive terms (Figure 11). Some of the commonest general terms are equivalences with respect to provenance (see comments in right column) and oil and gas potential (left column).

The coal micrographs shown in Figure 9 are unusual in that a polished thin section has been made to allow viewing in reflected and transmitted light together with fluorescence. As illustrated in Figure 12, visual characterization of the kerogen in sedimentary rocks may be undertaken on polished rock surfaces (reflected light and fluorescence) or on kerogen concentrates. The kerogen concentrate is made by removing carbonate with hydrochloric acid (HCl) and quartz and silicates (mica, etc.) with hydrofluoric acid (HF). The kerogen concentrate may still contain heavy minerals and pyrite, which can be removed by centrifuging in a high-density liquid. Some organic matter can be lost, so if a representative kerogen is required this step may be omitted.

Whole rock and kerogen concentrates are contrasted in Figure 13. The advantage of the 'whole rock' method is that the sedimentary context of the kerogen is preserved and the tedious acid digestion process—with associated health and safety issues—is omitted. The disadvantage is the dispersed nature of the kerogen makes some identifications difficult and quantification impossible. In contrast, viewing a representative kerogen concentrate allows for semiquantitative compositional estimates (%liptinite,

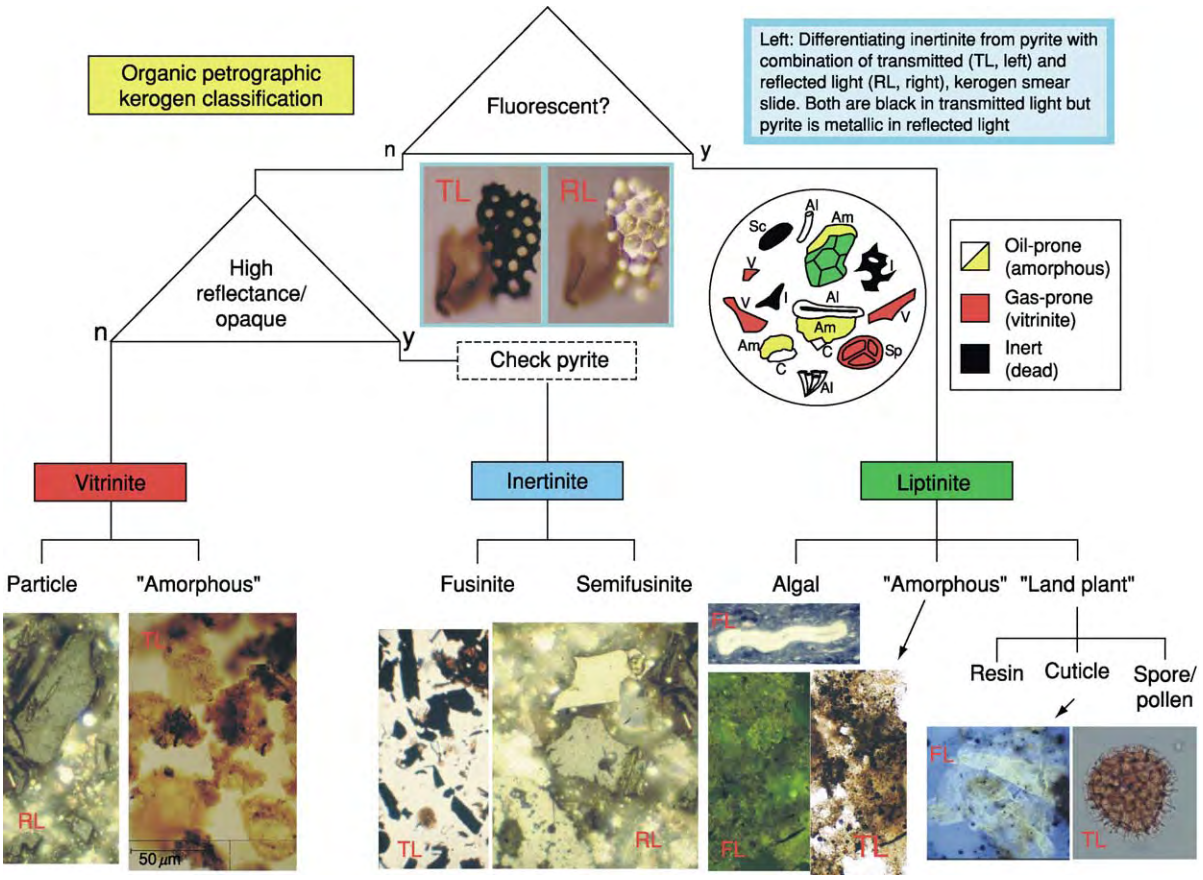


Figure 10 Classification of immature and mature kerogens using reflected (RL), transmitted (TL), and fluorescent light (FL) microscopy (y = yes, n = no).

Equivalent kerogen nomenclatures					
Hydrocarbon potential	Coal petrography	Palynology	Chemistry	Rock eval	Comments
Oil-prone (cracking to condensate andgas)	Liptinite	Algal/ Amorphous	Sapropel	Type I (Type IS)	Pure algal Includes bacterial biomass
Light oil, condensate-prone	Exinite	Herbaceous		Type II (Type IIS)	Cuticle gives waxyoil, Type II can be a mix of Types I, II land IV
Gas-prone	Vitrinite	"Woody"	Humic	Type III	Amorphous vitrinite and oil-impregnated fluorescent vitrinite exist
Dead carbon	Inertinite	"Coaly"	(not recognised)	Type IV (or IIIB)	Minor gas potential from semi-fusinite

Figure 11 Approximate equivalent terms used for the optical description of kerogen.

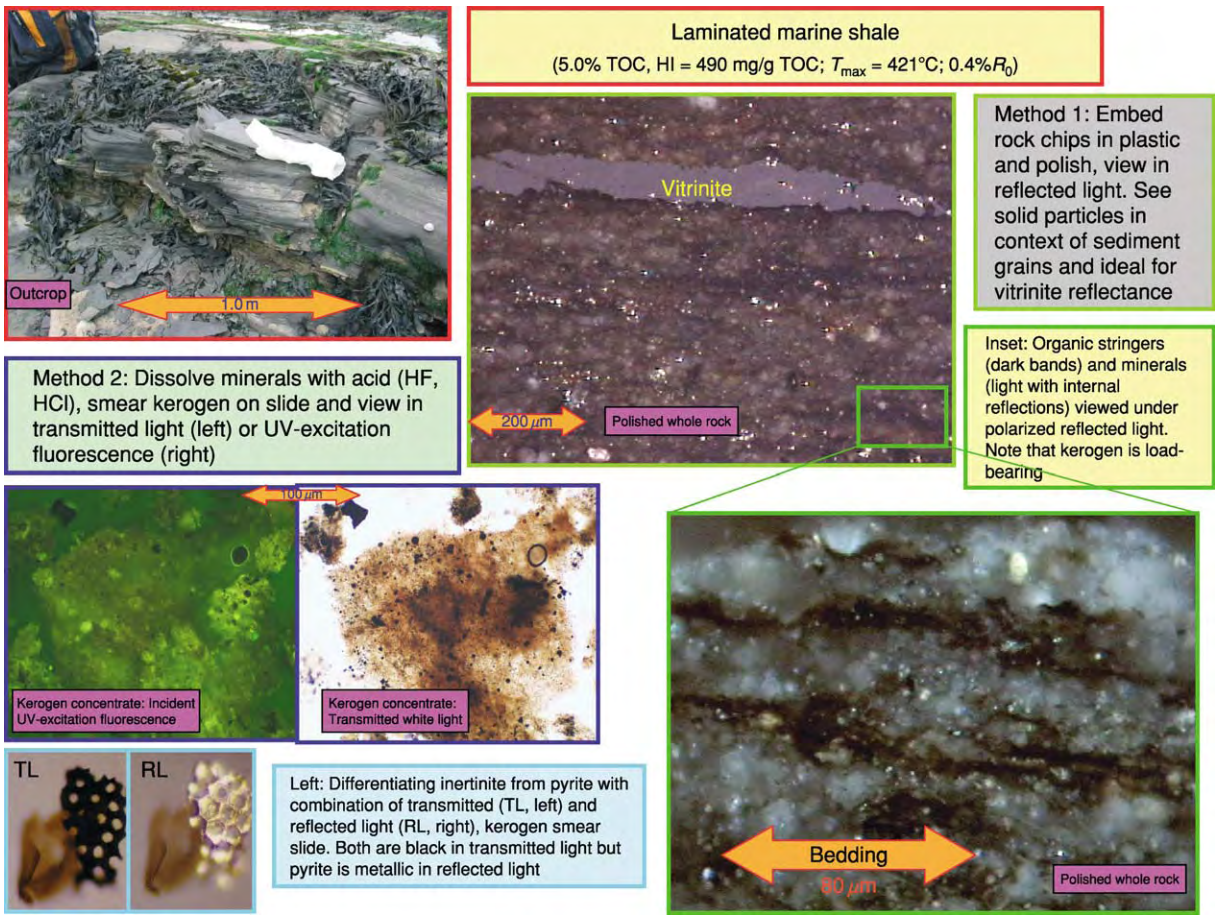


Figure 12 Organic petrographic methods applied to a sampled outcrop (top left) with rock chips polished and viewed in reflected light (right) and kerogen concentrates viewed in transmitted light together with UV excitation fluorescence (bottom left). Lower Lias oil shales, Kilve, Somerset, UK.

%vitrinite, %inertinite) as the particles are closely packed and may be viewed in all three microscopy modes (Figure 14).

Type of kerogen—pyrolysis methods Pyrolysis, particularly the widespread commercial Rock-Eval family of analysers, rapidly yields 'cheap and cheerful' information concerning kerogen type, maturity, potential yield of hydrocarbon, and the presence of migrated hydrocarbons. Developed by the Belgians and popularized by the French, Rock-Eval pyrolysis has become the most widespread apparatus for geochemical screening of drill samples. Its popularity derives from its modest sample requirements together with its relative cheapness and simplicity (though not necessarily reliability) of interpretation. For most applications, TOC data are required for the interpretation of Rock-Eval data. The recent versions of the Rock-Eval machine do, however, produce an 'organic carbon equivalent' value, arguably dispensing with the need for separate TOC data (*see* **Analytical Methods:** Geochemical Analysis (Including X-Ray)).

The conventional Rock-Eval machine progressively heats (pyrolyses) powdered rock in an inert atmosphere to produce the following sequence of events (Figure 15):

- Release of the 'free' hydrocarbons to give an S_1 peak (to about 300°C).
- Simulation of maturation and generation of new hydrocarbons from the kerogen (300–550°C) to yield a second peak, S_2, with maximum at temperature T_{max} (°C).
- Release of organically bound CO_2 over the temperature range 300–550°C to give an S_3 peak.

The interpretation of pyrolysis results in general and Rock-Eval in particular covers many aspects of petroleum geochemistry. Interpretation is normally undertaken using industry-standard cross-plots. The simplest, termed the pseudo-van Krevelen diagram, plots the oxygen index against the hydrogen index.

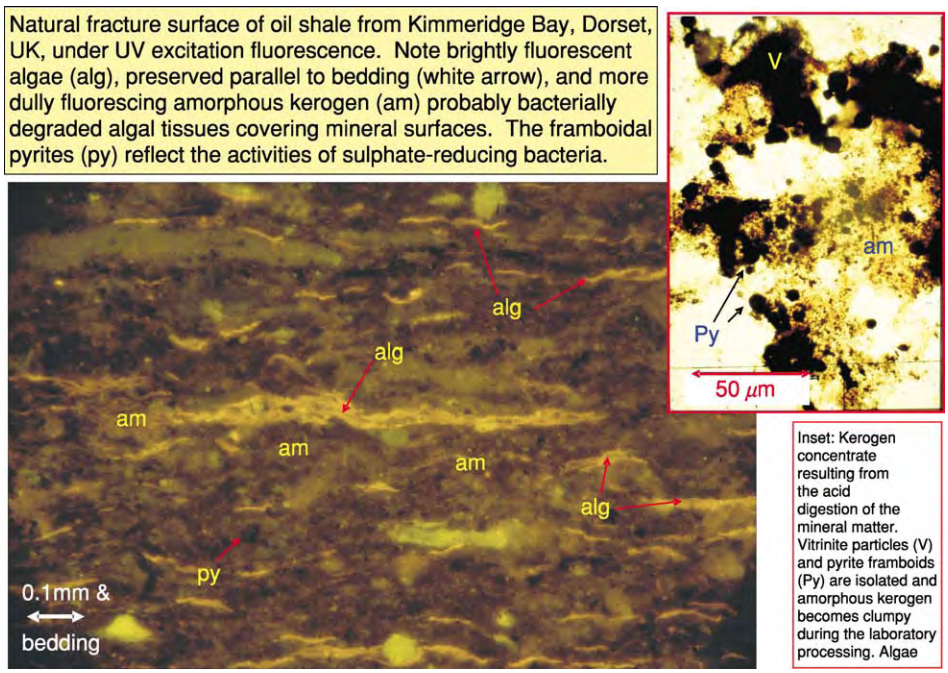


Figure 13 Visual kerogen descriptions can derive from whole rock surfaces viewed in reflected light or fluorescence (main picture) or as a kerogen concentrate in transmitted white light (inset).

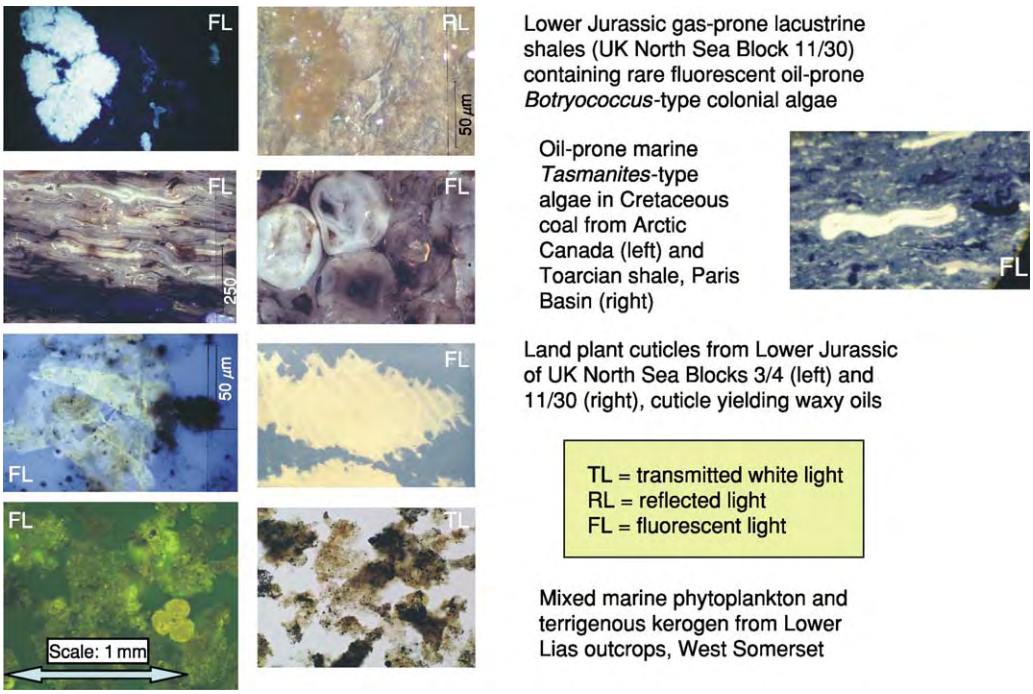


Figure 14 Some typical oil and gas prone kerogen associations illustrated as organic petrographic micrographs.

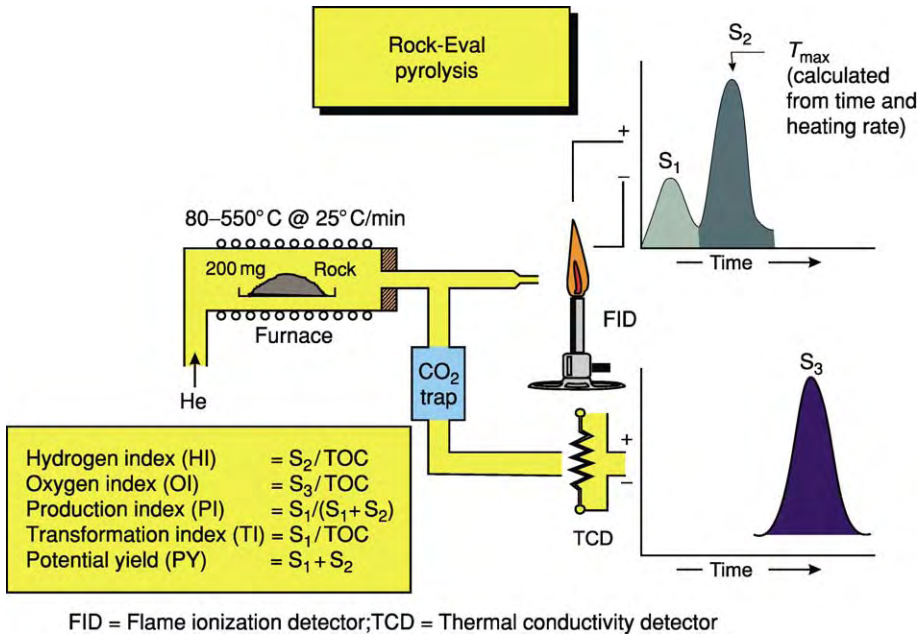


Figure 15 Schematic of the Rock Eval pyrolysis apparatus, output, and (with TOC) derived parameters.

This is shown in Figure 16, where four main kerogen types can be identified:

- Type I kerogen of algal origin, which is oil-prone;
- Type II kerogen deriving from a number of origins though all are oil-prone:
  - Land plant spores, exines, and resins;
  - Bacterially degraded algal (originally Type I) kerogen; and
  - A mixture of Type I (algal) and Type III (humic) particles.
- Type III kerogen of lignocellulosic or humic origin which is gas-prone;
- Type IIIb or IV kerogen defined as altered (oxidized) humic material or fossil charcoal: it has no oil or gas potential.

The reliability of oxygen indices has been questioned since some carbonates (e.g., iron carbonates) decompose at the temperatures where the 'organic CO_2' is being collected (see Figure 15). High oxygen indices are suspect if deriving from rocks with high carbonate contents.

A more detailed approach to kerogen identification is shown in Figure 17, where a plot of the Rock-Eval S_2 yield versus TOC is used to identify the average percentage of dead carbon (%DC) as the intercept of the best-fit data trend on the TOC axis. An incremental part of this intercept has also been interpreted as resulting from the adsorption of S_2 organics on active mineral surfaces.

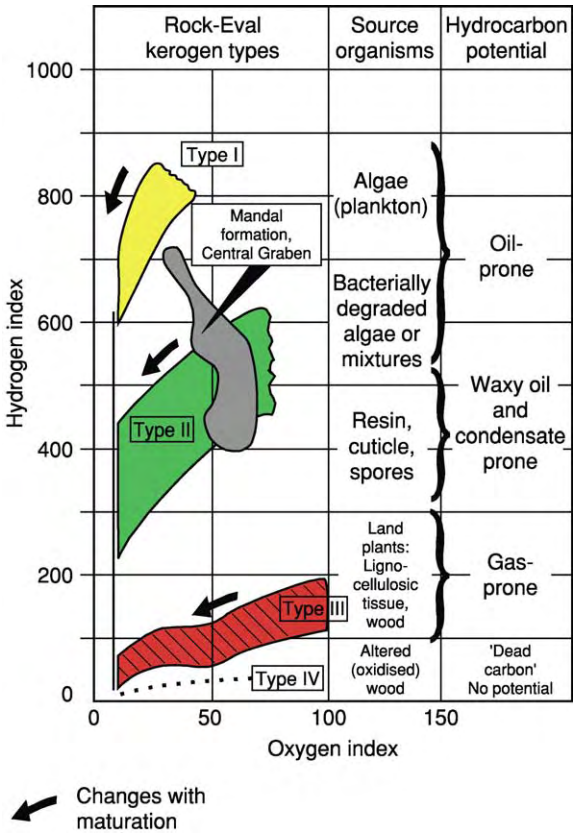


Figure 16 Kerogen type from Rock Eval hydrogen indices (mg Pyrolate/g TOC) and oxygen indices (mg CO_2/g TOC), with maturity trends indicated by the heavy black arrows.

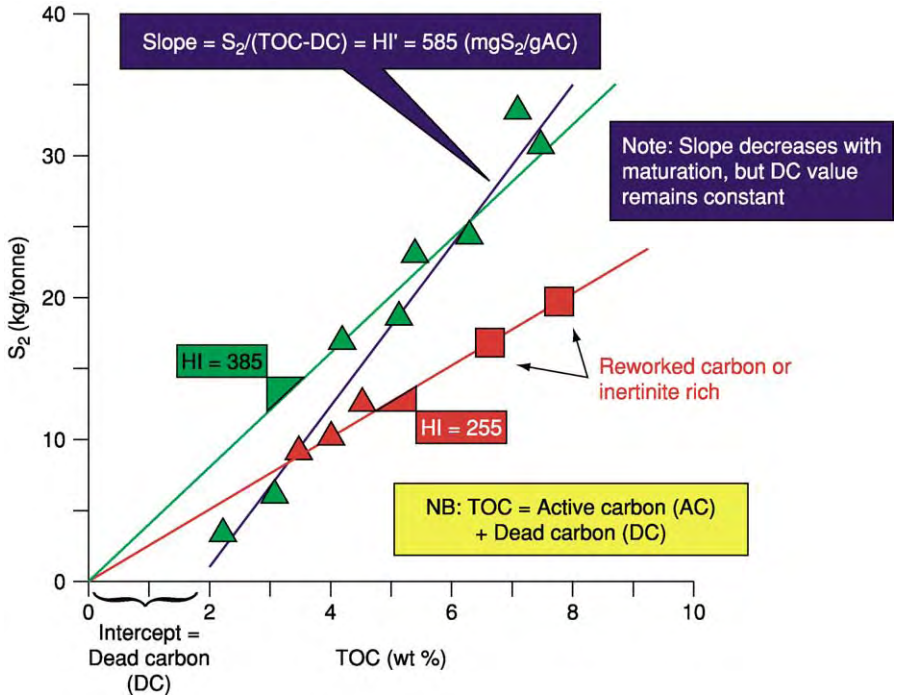


Figure 17 A more detailed determination of kerogen type in immature and early mature kerogens using Rock Eval S_2 and TOC data where the gradient is hydrogen index (S_2/TOC in units of mg Pyrolysate/g TOC).

Green and red trends (data plus lines) pass through the origin of the graph and define two types of kerogen with HI of 385 and 255 mg g TOC ^{1}, respectively. Ignoring two points, a single blue trend better fits all the values. The TOC can then be considered as the sum of dead carbon (DC) and active carbon (AC):

$$\%\text{TOC} = \%\text{DC} + \%\text{AC}$$

Note %TOC represents a percentage while TOC refers to a ratio.

It follows that the slope of the blue trend in Figure 17 is a modified hydrogen index (HI′), calculated on a dead-carbon-free basis (HI′ = 585 mg S_2/g AC ^{1}):

$$\text{HI}' = 100 \times S_2/(\%\text{TOC} - \%\text{DC}) = 100 \times S_2/\%\text{AC}$$

The above analysis is really only applicable to immature and early oil-mature kerogens (see next section). The determination of kerogen type for mid-, late-, and post-mature kerogens can be addressed with Figure 18, where Rock-Eval T_{max} is the maturity parameter. The typical bacterial-algal Type II kerogen shows a small increase in hydrogen index in the immature stage (up to $T_{max} = 430°C$), followed by a linear decrease, which reflects oil generation up to about $T_{max} = 465°C$. The gas-prone Type III kerogen with a lower initial hydrogen index falls slowly over a larger T_{max} maturity range.

For the highly oil-prone Type I algal kerogen the immature/early mature boundary can be as high as $T_{max} = 440°C$, but once reached the rapid generation of oil produces a swift decrease in hydrogen index. A singular kerogen type is likely to require its own T_{max} calibration: for example, generation from a sulphur-rich Type IIS kerogen produces a reduction in hydrogen index at an anomalously low maturity level. Thus, using the Rock-Eval pyrolysis approach to kerogen typing requires knowledge of source rock maturity.

The pyrolysis S_2 peak can be trapped and passed to a gas chromatograph (GC) to produce a characteristic fingerprint (Figure 19). This pyrolysis-GC fingerprint differentiates oil-prone kerogen dominated by the alkane–alkene doublets, from gas-prone kerogens where the aromatic compounds such as toluene, xylenes, and low-molecular-weight alkanes dominate.

Kerogen type can also be determined somewhat more indirectly from a number of other kerogen properties, such as elemental analysis, stable carbon, and hydrogen isotopes, and the molecular fingerprints from biomarkers, such as steranes and terpanes.

Using these more sophisticated properties a source rock kerogen can be attributed to an organofacies based on water chemistry (fresh, saline, hypersaline),

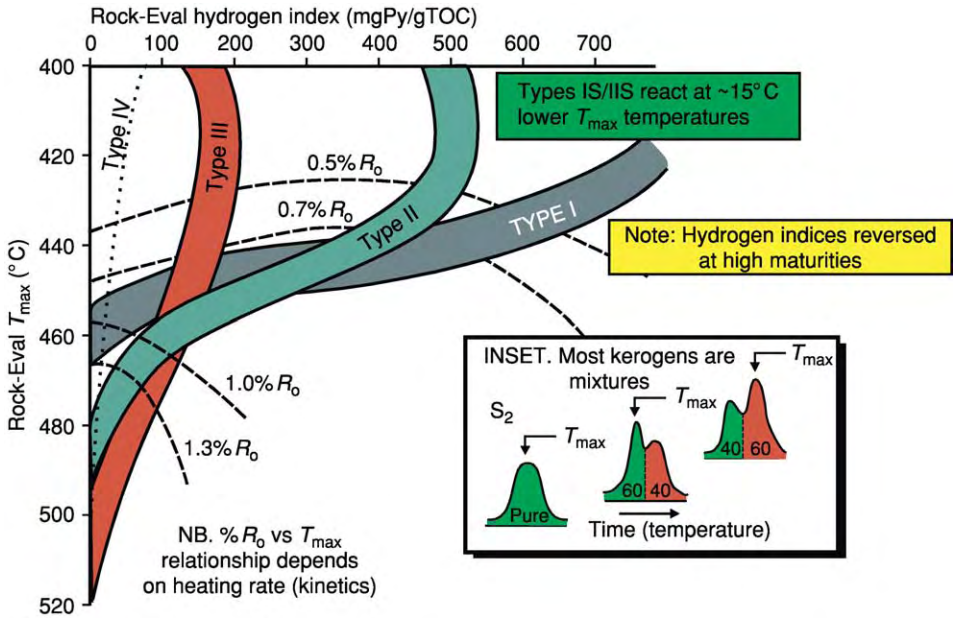


Figure 18 The effect of maturity changes on hydrogen index values for the three major kerogen types.

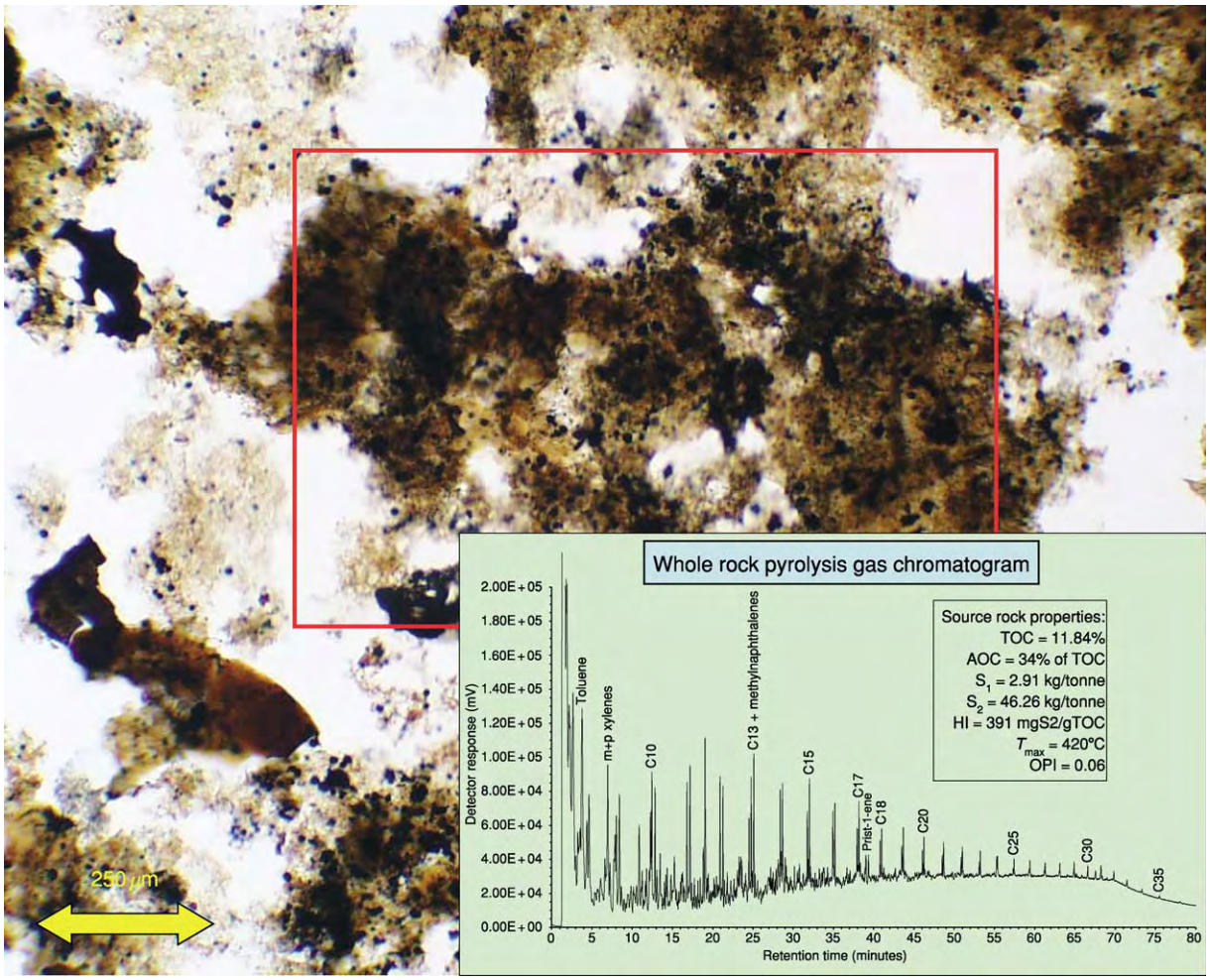


Figure 19 Kerogen typing using pyrolysis GC fingerprints to confirm the presence of oil prone kerogen (Lias shale from a well in the UK Western Approaches).

mineralogy (clastics, carbonates), and dissolved oxygen levels (anoxic, dysoxic, oxic) at deposition. In addition the molecular and isotopic signature can place limits on the stratigraphic age and types of organism contributing to the kerogen. The importance of these molecular and isotopic properties is that they are carried with the oil to the trap or surface seep. This allows the organofacies of the source to be estimated from migrated oil and hence address the question of oil–source rock correlation.

Maturation

As any sediment is buried the contained organic matter is matured as a function primarily of increasing temperature, with ancillary effects of time and possibly pressure. Maturity can be related to the stages of hydrocarbon generation and can be used to constrain geological models. In this section maturity is defined in terms of temperature and time, using measured indicators such as Rock-Eval T_{max}, vitrinite reflectance, and spore colour indices. Some common maturity parameters are related to oil generation levels in Table 1 and to gas generation levels in Table 2.

These equivalencies shown in Table 1 and Table 2 are subject to a number of caveats:

- Different responses to temperature and time for different kerogens (kinetics);
- Differences between sampling and laboratory analytical procedures; and
- Different concepts of what constitutes early, mid-, and late maturity.

The recommended procedure is to validate these equivalencies using locally generated data from within the project area.

The most common technique for measuring source rock maturation is vitrinite reflectance, developed for studies of coals in the 1960s and 1970s. It has subsequently been increasingly applied to sediments and derived kerogen concentrates. The reflectance of a material ($\%R$) is a function of its refractive and absorptive indices together with the refractive index of the observational medium (air, water, oil, etc.), as expressed by Beer's equation:

$$\%R = [(n - N)^2 + n^2k^2]/[(n + N)^2 + n^2k^2]$$

where n = the refractive index of the material, N = the refractive index of the measuring medium, and k = the absorptive index of the material.

The basic vitrinite reflectance technique is shown in Figure 20, where the upper left panel illustrates the reflected light polarizing microscope fitted with a photomultiplier detector. The reflectance is measured relative to a glass or crystal standard:

$$\text{vitrinite reflectance} = (\text{PMR of vitrinite}/\text{PMR of standard}) \times \text{reflectance of standard}$$

where PMR = photomultiplier response.

Reflectance, determined orthogonal to a polished surface, is expressed as a percentage of the incident light. It may be measured in air ($\%R_a$, typical of Russian data) or using oil-immersion objectives ($\%R_o$, typical in the West).

The first step for determining vitrinite reflectance is to identify vitrinite in general and the 'right type' of vitrinite (termed telocollinite) in particular (Figure 20A, inset). If all organic particles are measured, the low reflecting liptinite (L) may be differentiated from vitrinites (V) and the high-reflecting inertinites (I) (Figure 20B, upper histogram). This is rarely undertaken. In practice only particles believed

Table 1 Approximate[a] equivalence of maturity parameters with respect to oil generation

Generation stage (oil)	*%R_o (%)*	*T_{max} (°C)*	*SCI (1 10)*	*TAI (1 5)*	*LOM (1 20)*	*User*[a]
Immature						
	0.50	432	4	2.25	7.8	
Early mature						
	0.70	442	5	2.30	8.0	
Mid mature						
	1.00	455	9	2.60	10.5	
Late mature						
	1.30	465	>9	3.50	12	
Post mature						

$\%R_o$ vitrinite reflectance; T_{max} Rock Eval T_{max}; SCI spore colour index; TAI thermal alteration index; LOM Shell's level of organic maturation.

[a]To determine your own equivalences, produce your own cross plots.

Table 2 Approximate[a] equivalence of maturity parameters with respect to gas generation

Generation stage (gas)	*%R_o (%)*	*T_{max} (°C)*	*SCI (1 10)*	*TAI (1 5)*	*LOM (1 20)*	*User*[a]
Immature						
	1.00	455	9	2.6	10.5	
Early mature						
	1.30	465	>9	3.50	12	
Mid mature						
	2.20	525	10+	4.00	14.5	
Late mature						
	>3.00	>575	10+	5.00	18	
Post mature (onset greenschist facies metamorphism)						

$\%R_o$ vitrinite reflectance; T_{max} Rock Eval T_{max}; SCI spore colour index; TAI thermal alteration index; LOM Shell's level of organic maturation.

[a]To determine your own equivalencies, produce your own cross plots.

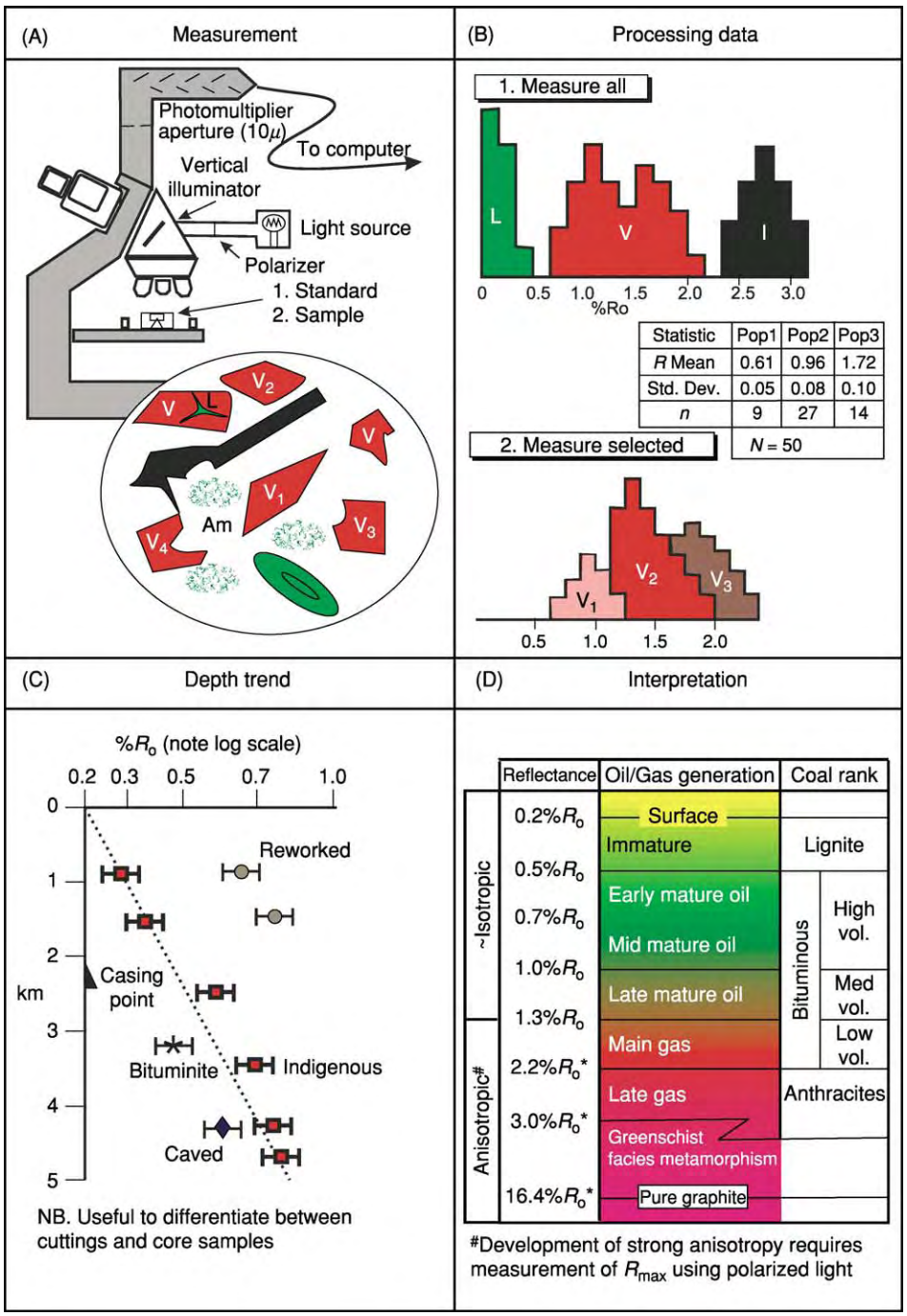


Figure 20 The measurement and interpretation of vitrinite reflectance for determining the maturity of sedimentary rocks in general and oil and gas source rocks in particular.

by the petrographer to be vitrinite are measured (Figure 20B, lower histogram): typically between 20 and 50 particles of vitrinite are measured, and an arithmetic mean value calculated. If multiple populations of vitrinite are recognized, then individual mean values are calculated for each population (Figure 20B, inset table).

If linearly polarized light is used (Figure 20A, see 'polarizer'), rotating the microscope stage and hence the sample with produce a maximum ($\%R_{o,\,max}$) and a minimum ($\%R_{o,min}$) if the material is optically anisotropic. As a result of maturing under overburden pressure, vitrinite develops bedding-parallel anisotropy at the late oil and early gas

generation levels ($\sim 1.0\%R_o$). If the vitrinite is anisotropic and maximum and minimum reflectances are determined, then the mean of the maximum values may be reported ($\%R_{o,\ \text{mean-max}}$). Theoretically the maximum $\%R_{o,\ \text{max}}$ is the most significant indicator of maturity.

The interpretation of the mean vitrinite reflectance values is normally undertaken by plotting $\%R_o$ on a log-scale against sample depth (Figure 20C), where anomalies may be recognized as plotting away from the linear trend. The interpretation then uses the industry-standard relationship with both oil and gas generation windows and coal rank (Figure 20D). Discontinuities in the maturity trend may be seen at unconformities (where the amount of uplift may be estimated from the offset), and excursions may relate to intrusions or hydrothermal flow. As well as defining the maturity of the sampled section, the reflectance trend can be projected to predict—in the absence of major unconformities—maturity levels ahead of the drilling bit.

In addition to vitrinite reflectance two other maturity parameters are commonly reported, Rock-Eval T_{max} (Figure 15), and kerogen or spore colour estimated on TAI (1–5) and SCI (1–10) scales, respectively. Acronyms are defined and equivalences detailed for oil and gas generation in Table 1 and Table 2, respectively. Although the T_{max} values are often available in large numbers, in a single well they often show substantial scatter with respect to depth. In contrast, fewer visual estimates of kerogen colour are determined, but they often produce a better maturity trend with depth. Spore colour estimates are particularly reliable since reworked or caved spores and spores introduced with drilling mud or other contamination can be readily identified as being out of stratigraphic sequence.

Finally the maturity involved in a petroleum system will normally be interpreted in the context of basin modelling and a burial history plot (Figure 21). The burial history plot (time and hence stratigraphy versus depth) traces the burial depth (and temperature) for each modelled strata (black lines) from deposition to the present day, including periods of uplift reflecting erosion at the palaeosurface. Compaction is modelled by various methods as the sediments are progressively buried, and the temperature grid (red lines) calculated from geothermal gradients or mantle heat flow. The heat flow calculation requires the knowledge of the thermal conductivity of each sediment unit—a function of mineralogy and porosity (compaction):

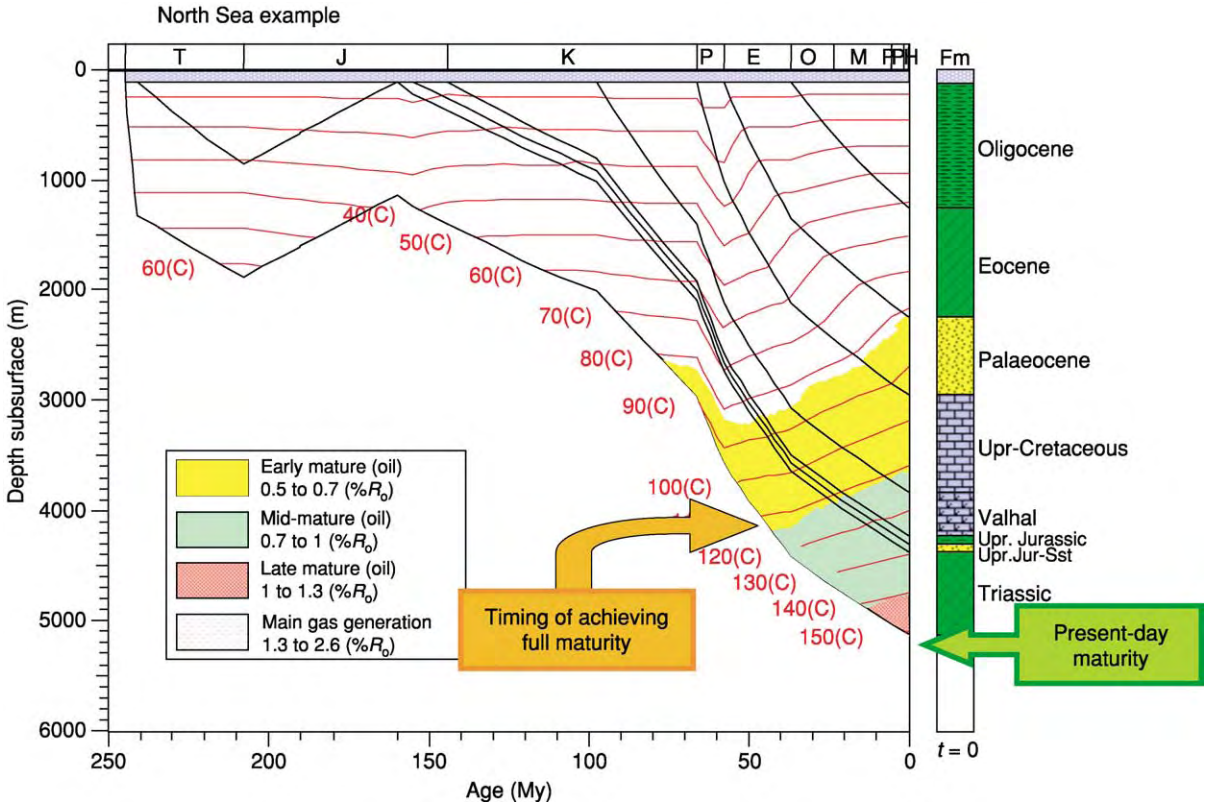


Figure 21 Burial history model showing present day and palaeotemperature and early , mid , and late mature oil windows defined by vitrinite reflectance.

$$\text{Heat flow}(\text{mW m}^{-2}) = \text{Geothermal gradient}\ (^{\circ}\text{C km}^{-1}) \times \text{Thermal conductivity}\ (\text{W m}^{-1}\,^{\circ}\text{C}^{-1})$$

The calibration step then minimizes the difference between measured and modelled temperature (by adjusting heat flow) and maturity parameters as given in Table 1 (by adjusting uplift, etc.). From knowledge of the temperature history for each stratum, the kinetics of decomposition of the kerogens (i.e., the generation of oil and gas) are modelled using the Arrhenius equation. Once calibrated, the model is used to predict the amount, composition, and timing of hydrocarbon generated as discussed below.

Generation and Expulsion

The process of generation describes the conversion of kerogen to petroleum (oil and gas):

$$\text{Kerogen} \xrightarrow{TR} \text{oil} + \text{gas} + \text{residue}$$

Generation, as opposed to maturation, is measured in terms of the extent of kerogen conversion (transformation ratio, TR) and is used to define the oil and gas windows (shown as yellow, green, and red areas in Figure 21).

The generation process can be monitored by laboratory measurement of the volumes (or masses) of generated oil or gas extracted from source rocks (Figure 22), or liberated during pyrolysis (Figure 23). Although the former is constructed from solvent extraction of the materials within source rocks as a fraction of the TOC (units of milligrams of extract per gram of TOC), the latter is reported as the Rock-Eval pyrolysis production index [$\text{PI} = S_1/(S_1 + S_2)$]. The PI is the ratio of the free hydrocarbon (S_1, kg t^{-1}) as a function of the remaining pyrolysate (S_2, kg t^{-1}). In both cases, the onset of generation (early generation) is indicated by an increase in yield. While the early generation phase starts the build-up of petroleum in the source rock, the continuation of the process saturates the source rock, leading to expulsion and eventually the cracking of oil to gas.

Assuming a typical geological heating rate (e.g., 1°C My^{-1}), the following maturity zones can be defined with respect to oil generation, expulsion, and cracking from Type II oil-prone kerogens:

Immature, <80°C (<175°F): No significant oil generation in the source rock: biogenic methane can be generated in recent and shallowly buried sediments.

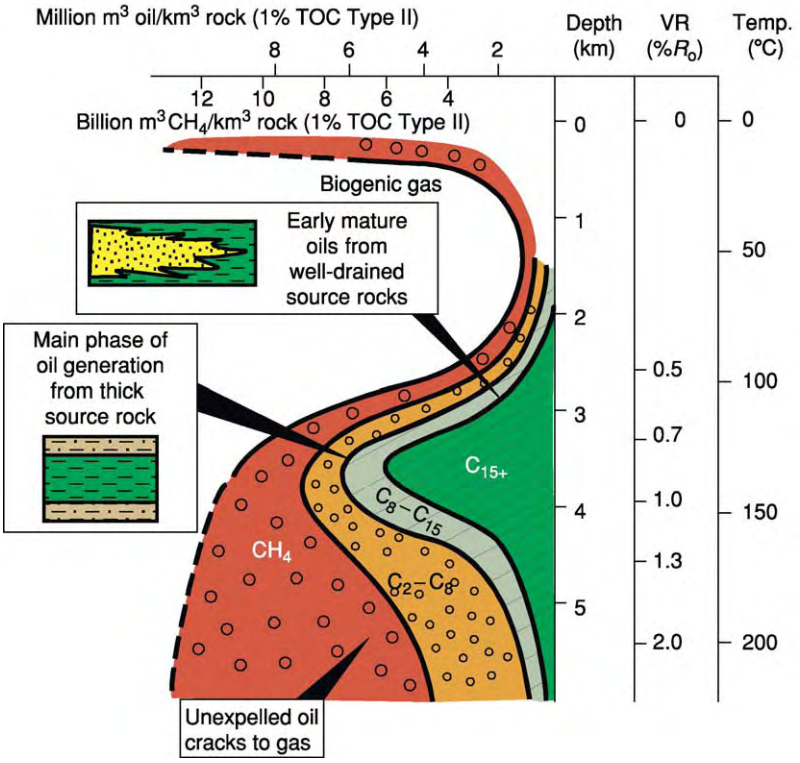


Figure 22 Illustration of generation based on a summation of methane, gas gasoline (C_2 C_8), light oil (C_8 C_{15}), and $C_{15}+$ fractions measured in marine source rocks containing 1% TOC of Type II kerogen at different levels of maturity.

Early mature, 80–115°C (175–240°F): Oil starts to generate in and saturate the source rock, hydrocarbons only being expelled in situations promoting highly efficient primary migration (e.g., where fault scarp fan sands, turbidites, or deltaic sands interdigitate with a source rock claystone or coals, or with fractured cherts and carbonates).

Mid-mature, 115–145°C (240–295°F): This is the main phase of oil generation and expulsion. Oil is expelled from 'massive' source rock units with efficiencies eventually in the range 60–80%.

Late mature, 145–160°C (295–330°F): Waning generation associated with cracking of unexpelled oil to give progressively lighter oils and condensate. Continued generation/expulsion phase particularly in poorly drained source rocks.

Post-mature, >165°C (>330°F): No remaining oil generative capacity exists in the source rock kerogen, and unexpelled or reservoired oil is cracking down to condensate and wet gas.

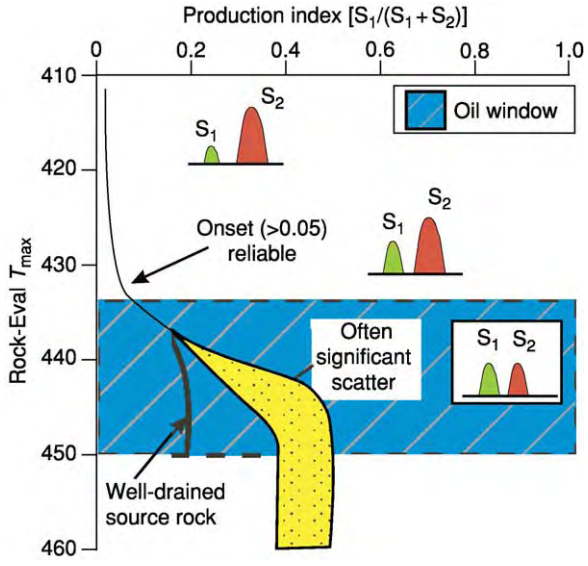


Figure 23 Oil generation as indicated by the Rock Eval production index (PI) where the maturity trend depends on both generation and expulsion (retention).

The temperatures given above relate to corrected wire-line log temperatures and Tertiary effective burial, as in the North Sea. Studies related to DST temperatures quote higher temperatures (up to 12°C higher) for these boundaries, while absence of temperature correction can produce substantially lower temperature boundaries. Lower temperature (e.g., 10°C lower overall) may be expected where a Mesozoic burial event is the controlling factor, while higher temperatures (up to 25°C higher) are reported where rapid Neogene–Quaternary burial controls generation.

Both extract yields and production index discussed above are measured parameters. The theoretical and hence modellable measure of kerogen degradation is the transformation ratio on a scale of $0 \rightarrow 1$:

$$\text{TR} = \text{Generated petroleum} / \text{Original petroleum potential}$$

where:

$$\text{Generated petroleum} = \text{Original petroleum potential} - \text{Residual petroleum potential}$$

The original petroleum potential is determined from immature source rocks of the same organofacies, whereas the residual petroleum potential is determined from the source rocks as it is buried through the oil and gas window. The equivalence of the transformation ratio as a measure of generation to the maturity parameters discussed in the previous section (Tables 1 and 2) is given in Table 3 for oil generation from Type II kerogen and Table 4 for gas generation from Type III kerogen.

Generation is a kinetically controlled process, being a function of the effects of both temperature and time on the breakage of chemical bonds in the kerogen present in the source rock. It can be modelled using simple reactions or networks of competing or sequential reactions, expressed in terms of the Arrhenius equation. The Arrhenius equation effectively expresses the ease of degradation of the kerogen to petroleum as a distribution of chemical bond strengths reflecting the C–C, C–H, C–O–C, C–S–C, and C–N–S bonds holding the kerogen network

Table 3 The generation defined oil window approximately related to some common maturity parameters

Parameters (oil & gas)	*Temperature (°C)*	*Transformation ratio[a] (0 1)*	*Vitrinite reflectance (%)*	*Spore colour index (1 10)*	*Rock Eval T_{max}(°C)*
Early mature	80 115	0.05 0.15	0.5 0.7	4.0 5.0	432 442
Mid mature	115 145	0.15 0.65	0.7 1.0	5.0 7.0	442 455
Late mature	145 165	0.65 0.95	1.0 1.3	7.0 9.0	455 465
Post mature	>165	>0.95	>1.3	>9.0	>465

[a]Transformation ratio of Type II kerogen (LLNL).

together. Each of these chemical bonds has a unique strength (activation energy), depending on the adjacent atoms and functional groups, together with stereochemistry. Using the information shown in a burial history plot (Figure 21) to solve the Arrhenius equation, the extent of kerogen degradation can be modelled.

The understanding of generation in both space and time is at the centre of the use of the petroleum system concept to explore for oil and gas, and this requires the use of computer calculation and modelling. The calculation of a theoretical transformation ratio using the Arrhenius equation can be undertaken using industry standard kinetics and basin modelling software, and can be used to calculate generation as a function of depth (Figure 24) and geological time (Figure 25).

The generation versus depth plot (Figure 24) is a simplification in that it assumes that all units within the modelled section contain the same kerogen type

Table 4 The generation defined gas window approximately related to some common maturity parameters

Parameters (gas)	*Temperature (°C)*	*Transformation ratio*[a] *(0 1)*	*Vitrinite reflectance (%)*	*Spore colour index (1 10)*	*Rock Eval* T_{max}*(°C)*
Early mature	115 145	0.05 0.45	0.7 1.0	5.0 7.0	455 465
Mid mature	145 220	0.45 0.85	1.0 2.2	7.0 9.0	465 525
Late mature	>220	0.85 0.95	2.2 >3.0	10	525 575
Post mature	??	>0.95	>3.0	10	>575

[a]Transformation ratio of Type III kerogen (LLNL).

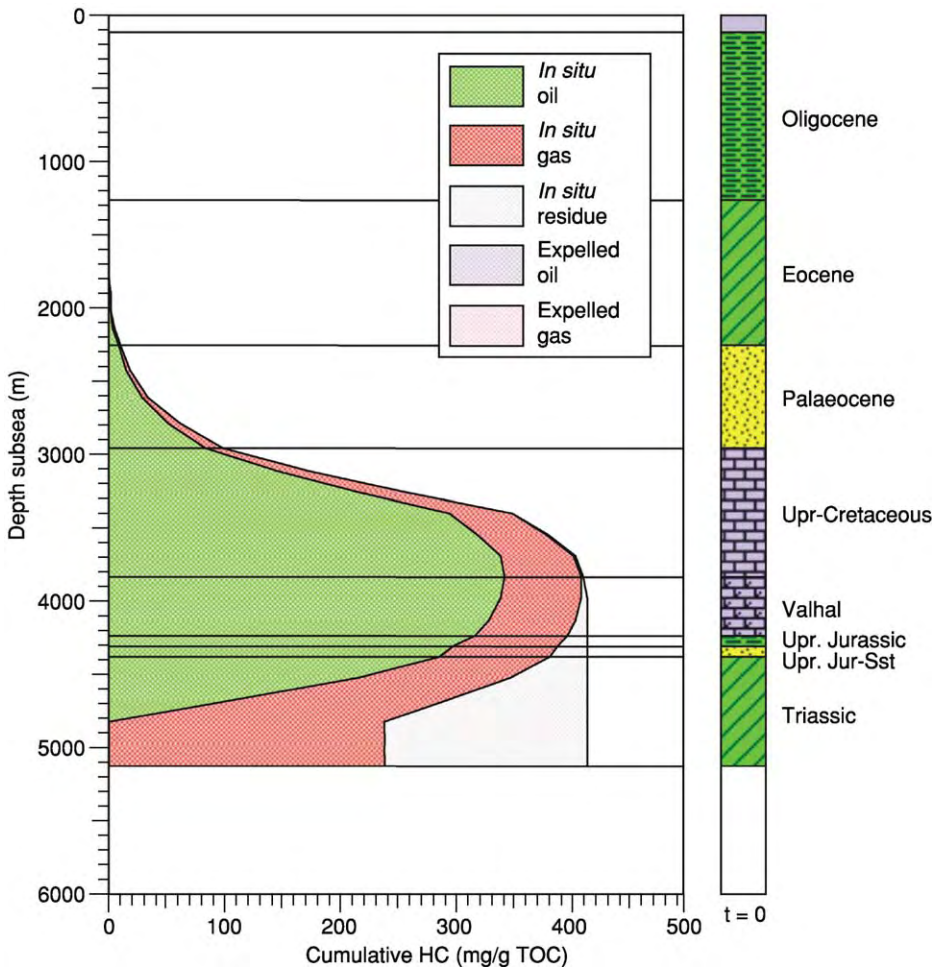


Figure 24 Computer modelled depth trend for the potential generation of oil and gas, assuming all rock units contain a common kerogen type (compare with Figure 22 based on measured data).

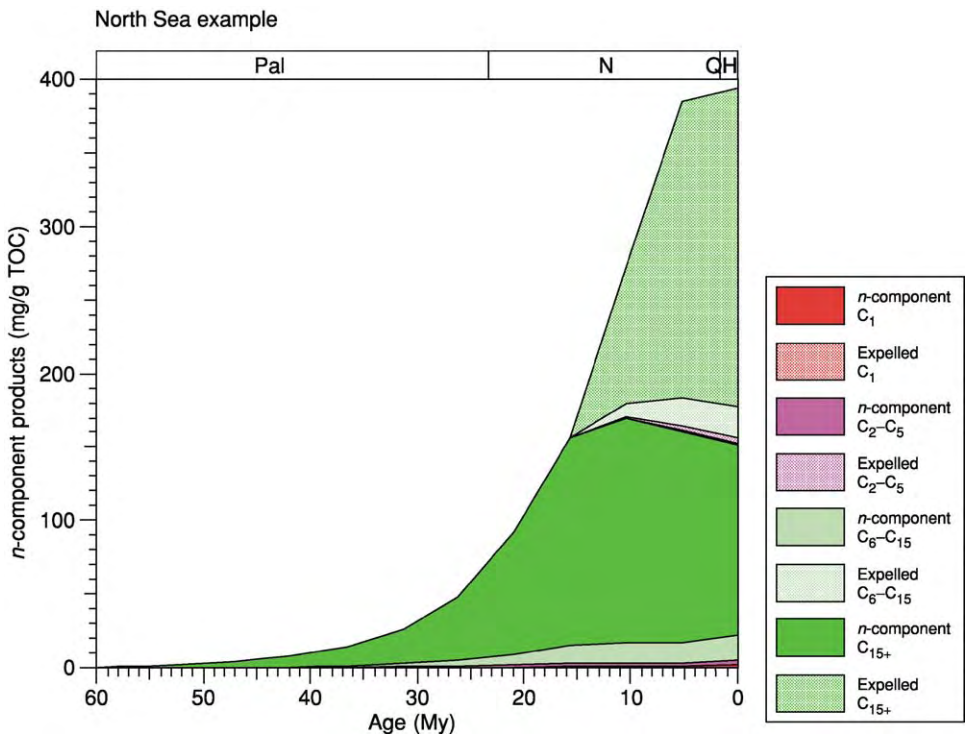


Figure 25 Computer modelled generation and expulsion of oil and gas through time from a single source rock unit containing kerogen.

(unlikely). In this case the kerogen is modelled to be broken down to oil and then oil potentially cracked to gas in a two-step procedure:

$$\text{Kerogen} \rightarrow \text{Oil} + \text{Primary gas}$$

$$\text{Oil} \rightarrow \text{Secondary gas}$$

Considering generation in terms of geological time, a single source rock unit will progressively generate a range of products, grouped somewhat arbitrarily as methane plus wet gas, light oil, and main oil ranges of molecules in Figure 25. The subdivision of oil into these molecular ranges, though subjective, is useful for estimating compositional risk in oil exploration.

In the source rock, these molecules form a single petroleum phase (monophasic mix), which can be modelled as being expelled from the source rock once the pore space in the source rock is saturated. Expulsion is poorly understood in terms of process, but well understood in terms of efficiency. Possible mechanisms for petroleum expulsion include movement:

1. From high to low pressure potential—with pressure also causing rock fracturing;
2. From high to low concentrations (diffusion);
3. Under capillary forces induced at mineral pore throats;
4. Under capillary forces through a continuous kerogen 'wick'; and
5. In aqueous solution during decompaction and clay dehydration.

A combination of these mechanisms possibly contributes to expulsion from the source rock, the relative contributions being the basis for the current debate. The favoured view envisages the dominant process as being based on pressure, where compaction coupled with the generation of hydrocarbon fluids from solid kerogen and temperature-induced expansion of pore fluids produces overpressure in the source rock. The overpressure eventually fractures the source rock, sporadically releasing the fluids into the migration pathway.

The expelled hydrocarbon yields shown in Figure 25 are calculated from the total generated less that required to saturate (a fraction of) the porosity of the source rock. The modelled porosity can be calculated using one of a number of possible compaction equations.

A quantitative understanding of expulsion efficiency is essential for successful prospect evaluation. Expulsion efficiency (EE) is defined in mass units as:

$$\text{EE} = \text{Petroleum expelled}/\text{Petroleum generated}$$

where

$$\text{Petroleum expelled} = \text{Petroleum generated} - \text{Petroleum retained}$$

and

$$\text{Petroleum generated} = \text{Initial potential} \times \text{Transformation ratio}$$

Within a basin such as the North Sea, which contains a uniform oil-prone source rock organofacies in the Upper Jurassic Kimmeridge Clay Formation, the initial potential is determined by analysis of immature source rock samples from the rim of the basin. The retained (i.e., generated but not expelled) petroleum can be determined by solvent extraction or pyrolysis (S_1 peak—see Figure 15). The transformation ratio can be determined from modelling, equated with vitrinite reflectance or pyrolysis T_{max} (e.g., Table 3) or from ratios of the initial to measured hydrogen index (e.g. initial HI = 600, present HI = 300, therefore TR = 0.5). Examples of the range of expulsion efficiencies calculated from North Sea source rock data are shown in Figure 26, where values rise from about 20% at 3.0–3.5 km to about 50% at 4.0 km with the deepest samples exhibiting values in the 75–95% range.

Considering dry gas (methane) generation in, and expulsion from seam coals requires an understanding of additional factors, in particular the very high

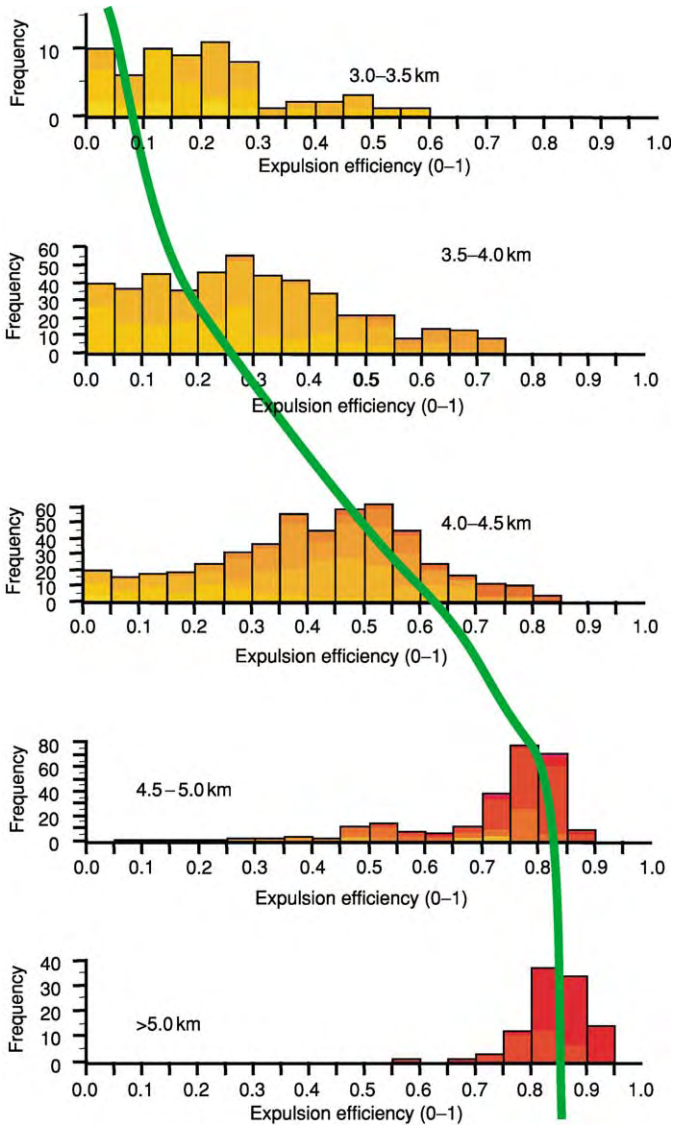


Figure 26 Expulsion efficiency as a function of depth in the North Sea based on solvent extraction (green trend line, from Larter (1988)) and Rock Eval pyrolysis data (histograms).

surface area (macro-, meso-, and microporosity) of coal in general and the vitrinite maceral in particular. The developing understanding seems to suggest a three-step process:

1. Generation of the methane from humic kerogen;
2. Adsorption of the methane on the abundant porosity of vitrinite; and
3. Expulsion of the methane at higher maturity as the porosity collapses.

In all cases the expelled monophasic petroleum leaves the source rock, but will only form commercial accumulations if migration provides a focus towards a trap. The processes controlling migration are discussed in the next section.

Migration

How petroleum travels long distances in the subsurface is a matter for speculation: facts are hard to come by.

Certainly oil being more buoyant than water, and gas being more buoyant than both, there is a natural tendency for the upward movement of separate gas and oil phases in a water-saturated permeable sediment. The fluids are essentially separating out within the porous media of sedimentary rocks. Attempting to understand the processes contributing to migration requires some knowledge over a large range of scales of the properties of the solid rock and the fluid phases flowing therein (Figure 27). The scales are not only

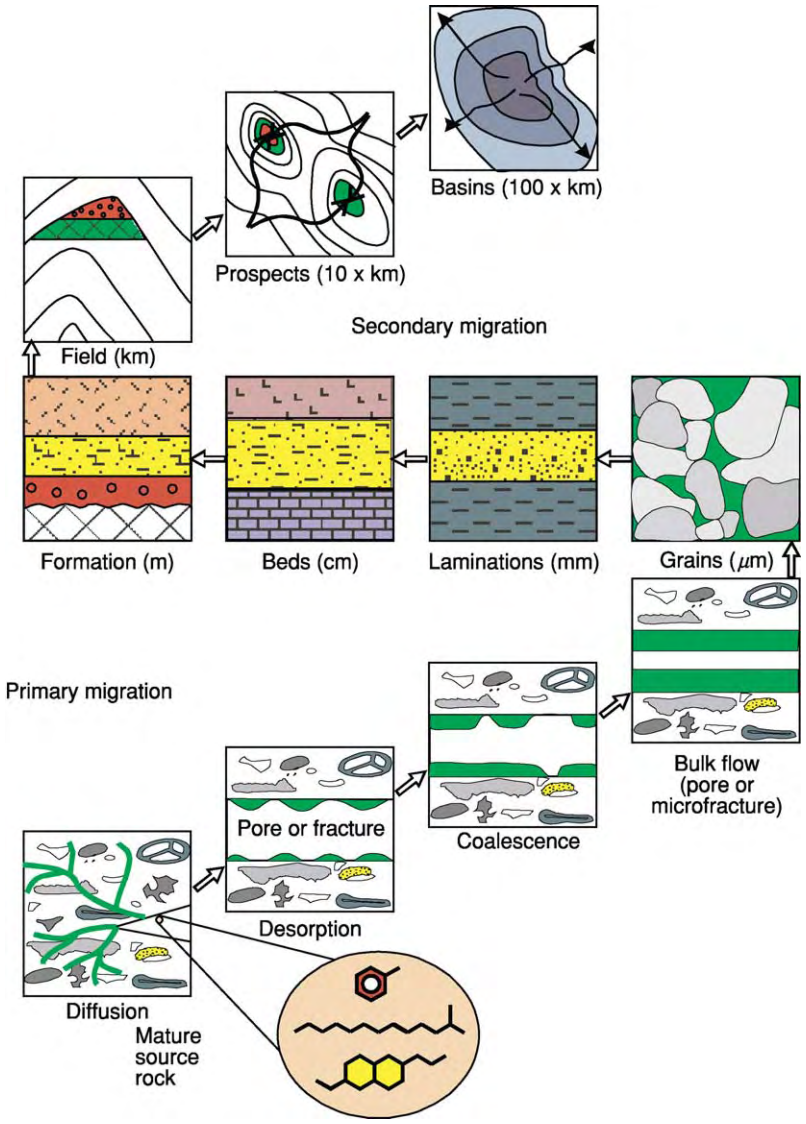


Figure 27 Our understanding of how oil and gas move through water saturated rock requires a study of flux over a large range of scales. Modified after Mann *et al.* (1991).

those from sedimentary grains to basins, but also from oil molecules to accumulations containing billions of barrels of oil.

At deposition, sedimentary rock is initially porous, but compaction and diagenesis rapidly reduces the porosity. Once sediments are indurated (compacted and cemented), fracture porosity may take over as the dominant conduit allowing fluid flow. Thus fluids can move through both intergranular or fracture porosity.

Once expelled into a more porous conduit such as a sandstone bed, a fault plane, or diagenetically altered carbonate, petroleum will start to move upwards under the force of buoyancy (Figure 28, inset). The buoyancy derives from the lower density of oil (typically 0.86 g cc $^{-1}$ or 36°API) relative to formation water (typically 1.03 g cc $^{-1}$), as in the kitchen where fat rises to the surface of milk. Once within the migration conduit the petroleum will rise to the roof of the rock layer and if the layers are tilted, the 'river of oil' will move up-dip and along zones of highest permeability from high point to high point. When seen at outcrop, this produces a braided network of oil-saturated rock, running under the top-seal of the migration conduit (Figure 28).

Migration can also be influenced by any water flux sharing the same migration route (hydrodynamic control) or by pore pressure gradients associated with the progressive increase in pore throat diameters (permeability control). The pore throat diameters together with wettability essentially control the permeability of the rock with respect to oil or gas.

Migration distances can vary widely. In situations where mature source rock is interbedded with reservoir sands the migration path may only be a number of metres. Large accumulations, however, demand the gathering of oil and gas over a wide area, and hence long-distance migration pathways. In the North Sea migration distances are mainly in the range 10–20 km with 65 km as a maximum. In large basins such as the West Canada Basin and West Siberian Basin migration pathways of many hundreds of kilometres have been proposed.

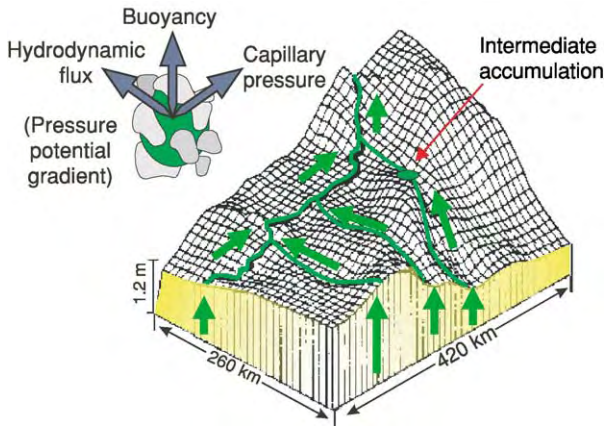


Figure 28 Migration occurs as 'rivers of oil' running up dip mainly under the influence of density driven buoyancy, though hydrodynamic flux and capillary pressure may exert ancillary controls. Modified after IES (1985).

As the expelled monophasic petroleum moves upwards under buoyancy, the pressure and temperature regime will change and the bubble-point or dew-point may be reached (Figure 29). Separation into a liquid (oil) phase and a gas phase will result, and these may migrate separately due to differential buoyancy and relative permeabilities.

Studying a petroleum system demands an understanding of the efficiency of the migration process. This is probably the least well-quantified element of a petroleum system, and is normally determined by difference:

$$\text{Accumulation} = \text{Generated} \times \text{Expulsion efficiency} \times \text{Migration efficiency}$$

Therefore

$$\text{Migration efficiency} = \text{Accumulated}/(\text{Generated} \times \text{Expulsion efficiency})$$

The above equation can be used to determine the migration efficiency if the size of an accumulation, the amount generated, and expulsion efficiency are all known. Calibrated against known accumulations, the migration efficiency can then be used to predict the petroleum charge to undrilled prospects. Migration efficiencies are not widely discussed or reported, but values seem to fall in the range of 50% (maximum) to 0 (i.e., no oil or gas reached the drilled prospect). It may be helpful to consider the inefficiency of migration as the residual oil left behind along the migration path. The recognition that oil migration is a focussed process affecting only a small percentage of the rock between kitchen and accumulation fits well with the calculated migration efficiencies.

Accumulation and Survival

Once charged to the trap, the survival of the petroleum depends on a good seal and avoiding the effects of various oil alteration processes. An approach to use bulk, molecular, and isotopic characterization of oil and gas to identify such reservoir processes is shown in Figure 30. As emphasized in the figure, the analytical approach to differentiating between oil alteration processes requires knowledge of the geometry and history of the trap.

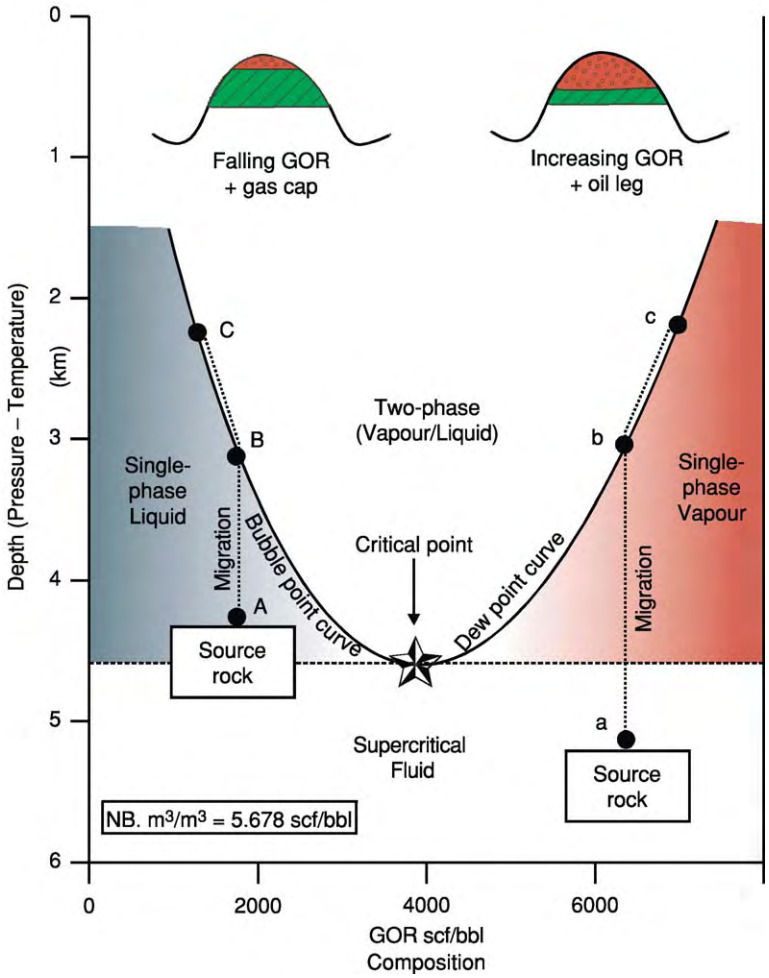


Figure 29 Phase separation of upward migrating petroleum depending on the initial gas/oil ratio (GOR) and reductions in temperature and pressure.

In terms of a seal, efficiencies range from near perfect in the case of intact salt (Halite) to 'highly leaky' in the case of thin and poorly compacted mudstone containing a significant portion silt and fine sand. Diagenetic cement can greatly enhance the sealing efficiency of more porous strata, even when relatively shallow. In essence the efficiency of a seal must be measured against the charge rate if generation continues, or the time since charge ceased if in the past. A more specific and dramatic way of breaching the seal of the trap is faulting.

For structural traps, the prediction of whether a fault is a seal or a migration conduit is complex, the controlling factors being only poorly understood at present. Some unusual seals comprise a hydrodynamic trap, where a counterflow of formation water opposes the buoyancy of the oil, and shallow seals in the Arctic involving the permafrost layer or gas hydrate (*see* **Petroleum Geology:** Gas Hydrates).

The concept of the 'half-life' of an oil or gas field has been proposed, this being defined as the time (millions of years) required for half of the charge volume to escape through the seal. Gas sealed by mudstone has a half-life in the range 30–60 My, and oil about an order of magnitude greater. A continuous and stable halite seal can be considered near perfect with an infinite half-life.

A number of intrareservoir processes are listed towards the base of Figure 30, and of these the most common are bacterial degradation and thermal cracking. These occur when the reservoir is unusually cool (<65°C) and unusually hot (>165°C), respectively.

Of the listed intrareservoir processes, bacterial degradation is arguably the most destructive. In essence it occurs when petroleum accumulates in a reservoir that is at a temperature allowing bacterial degradation of the oil or gas. As reported, the maximum

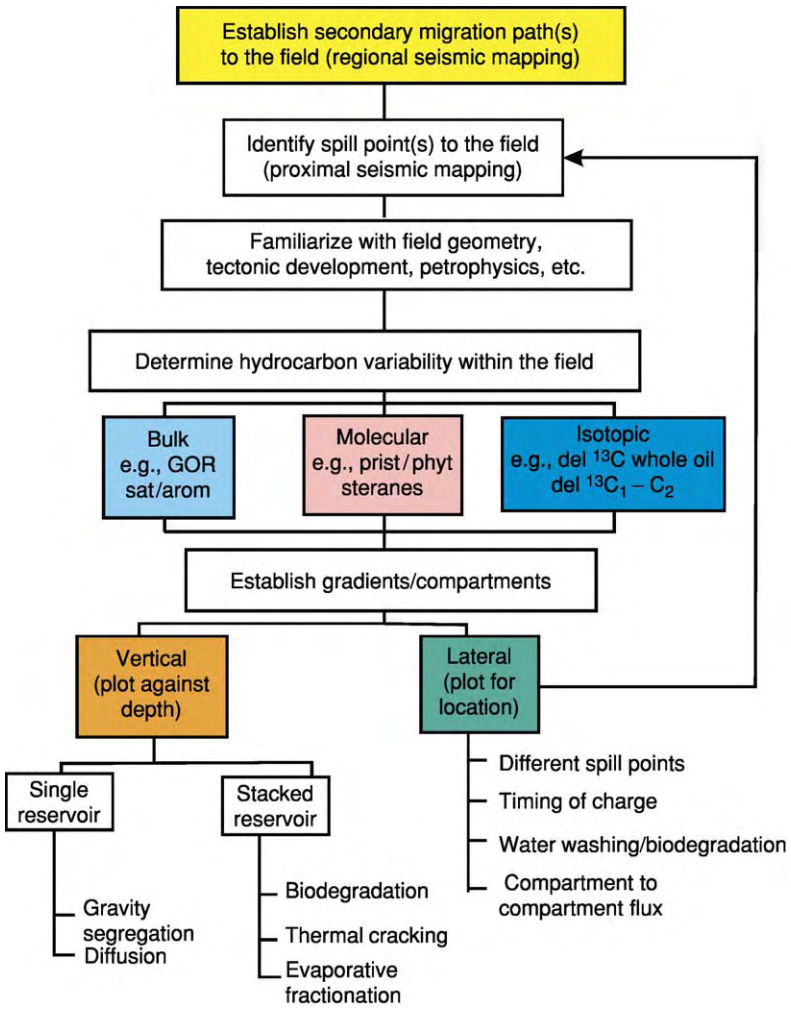


Figure 30 Steps required to identify and interpret vertical and lateral compositional gradients resulting from intrareservoir processes.

temperature boundary is in the range 65–75°C, the range reflecting different oil chemistries, formation waters, and hence active bacterial communities, together with different methods of measuring reservoir temperatures. In nature, viable bacteria will be preserved if the reservoir has never been buried to temperatures hotter than about 90°C. If hotter and then uplifted to a cooler regime, meteoric water influx into the reservoir is required to reintroduce bacteria.

The bacterial alteration of the oil progresses with the selective removal of various families of molecules (Table 5). The sequence of removal is broadly *n*-alkanes > branches alkanes > cyclic alkanes -> aromatics, this reflecting the 'digestibility' of hydrocarbon molecules by the bacteria.

Bacterial degradation of a gas accumulation removes the wet gas components (C_2–C_5) and in particular propane (C_3) with preferential removal of molecules with the ^{12}C isotope. Thus an initial wet gas is reduced to a dry gas accumulation of >95% methane with isotopically heavy residual propane. Bacterial activity is often confirmed by the thin rim of biodegraded heavy oil below the dry gas.

Most reservoirs are cooler than the source rocks from which they were filled. If the petroleum-filled reservoir is buried to higher temperatures, the oil will crack to condensate and gas (Table 6). This process of intrareservoir cracking starts at about 160°C and black oil is normally destroyed by 175°C. This has been termed the base of the oil preservation (as opposed to oil generation) window. An oil trap that has suffered intrareservoir cracking will be recognized as containing gas, with black flecks of pyrobitumen seen occupying some of the remaining porosity:

$$2 \times \text{oil}(CH_2) \rightarrow \text{pyrobitumen}(C) + \text{methane}(CH_4)$$

Table 5 The degree of bacterial degradation of oil on a 1 10 scale based on progressive loss of specific chemical groups

Degree of bacterial degradation	*Loss of named molecules indicates stage of biodegradation*			
1	C_{15}+ fraction	C_{15} C_7	Gasoline (C_4 C_7)	Gas (C_1 C_4)
			n Alkanes	*n* Propane[b]
2	$<C_{15}$ *n* alkanes	Alkyl aromatics		
3	Total loss of *n* alkanes, appearance of 10 demethyl hopanes			Ethane to *n* butane ... iso butane
4	C_{15} to C_{20} acyclic isoprenoids		iso alkanes	
5				
6	Regular steranes	Uncertain	Cycloalkanes	Dry gas only survives
7	Hopanes and			
8	diasteranes		Aromatics[a]	
9	Aromatized biomarkers			
10	Total loss of resolvable peaks ('hump' of UCM left as residue)			

UCM unresolved complex mixture.
[a]Earlier loss of aromatics (benzene and toluene) may result from concurrent water washing.
[b]Early loss of propane produces stable carbon isotope anomaly (isotopically heavy C_3).

Table 6 Intrareservoir cracking of oil: The relationship of transformation ratio of the kinetic cracking reaction to reservoir temperature based on kinetics from Platte River's BasinMod with heating rate of 1.25°C/My

Temperature (°C)	*Transformation ratio*[a]	*Comments*
110	0.00	Temperature of onset of oil window
130	0.00	Peak oil generation temperature
140	0.00	
145	0.01	Temperature of end of main oil window
150	0.03	Primary generation in source rock ceased, oil stable in reservoir
160	0.07	
165	0.16	Onset significant oil cracking
170	0.34	Main phase of oil cracking
175	0.65	
180	1.00	End of modelled black oil survival

[a]Transformation ratio of the oil → gas + residue cracking reaction.

System Efficiencies

Based on the definition of a petroleum system given at the start of this article a gross system efficiency may be defined as:

$$\text{System efficiency} = \text{Discovered petroleum}/\text{Generated petroleum}$$

Expressed as a percentage, reported system efficiencies fall in the range 0.5 to 30%. This range includes both natural variations attributable to the geology and errors in our understanding thereof. Based on the previous discussion and as illustrated in Figure 3, the simple equation given above can be expanded to give a more complete description of the controls:

$$\text{Discovered petroleum} = (\text{area of field} \times \text{thickness of reservoir} \times \text{porosity}) - \text{losses}$$

$$\text{Generated petroleum} = \text{source rock yield} \times \text{thickness} \times \text{area} \times \text{transformation ratio}$$

In addition to technical uncertainty in the above measurements, the reported values for 'discovered petroleum' may also be influenced by commercial and political considerations.

Uncertainty in the 'generated petroleum' value will include variations in total organic carbon, kerogen type, and its variation with source rock thickness: the transformation ratios (Table 3) for discrete areas of mature source rock will be determined from measured and modelled maturity data. These areas will be measured by planimeter on basin-scale mapping derived from the interpretation of a regional seismic grid. If the generated and discovered petroleum values are precisely defined, then the efficiency of the petroleum system is essentially the expulsion efficiency multiplied by the migration efficiency.

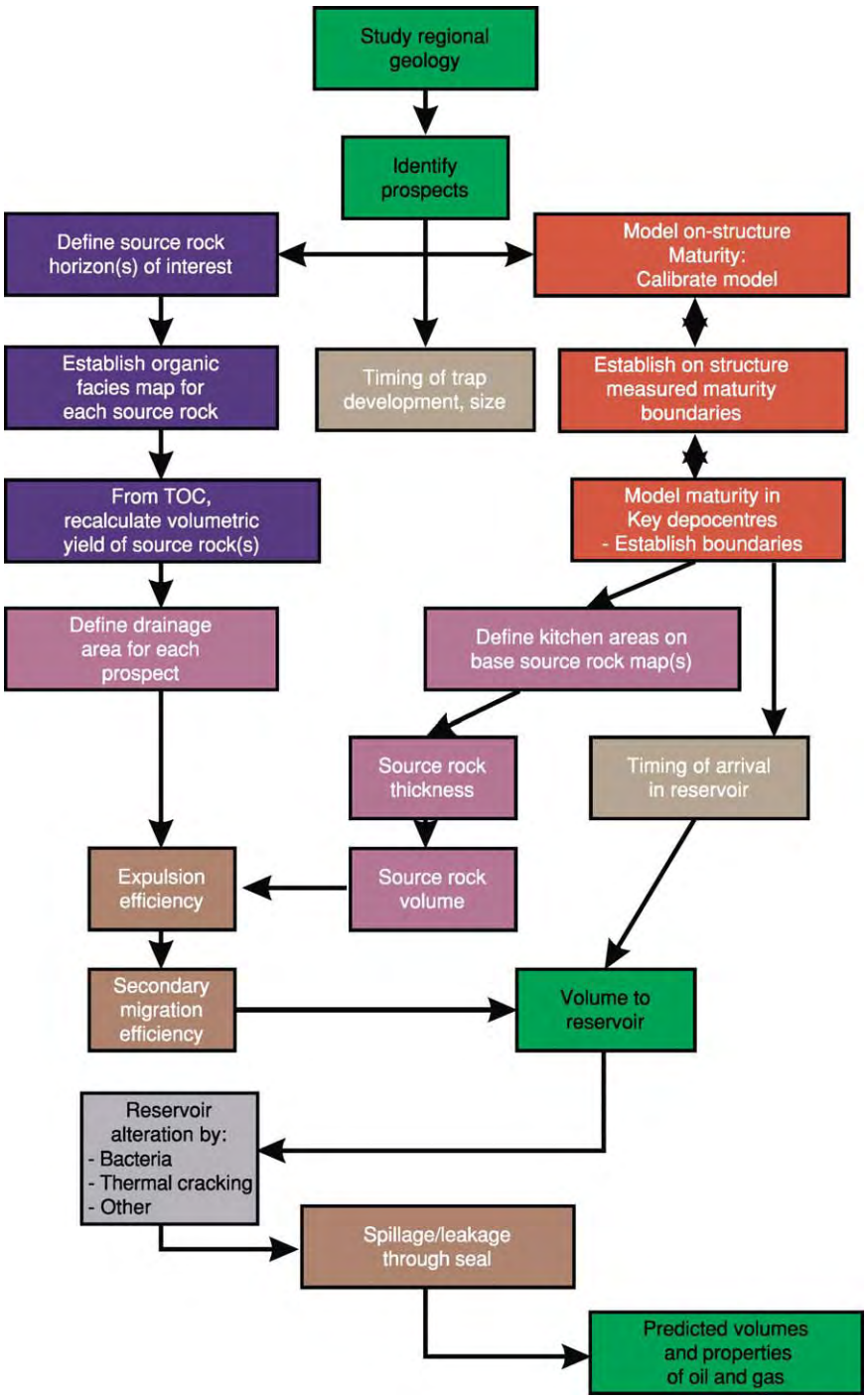


Figure 31 A summary of the steps leading to the determination of the efficiency of a petroleum system.

A possible workflow tying all these factors together is given in Figure 31 by way of a summary.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Petroleum Geology:** Overview; Chemical and Physical Properties; Gas Hydrates; Exploration; Production.

Further Reading

Biteau J J, de Janvry GC, and Perrodon A (2003) The petroleum system: a fundamental tool. *Oil and Gas Journal* 101(30): 34 39.

Biteau J J, Janvry GC, and Perrodon A (2003) How the petroleum system relates to the petroleum province. *Oil and Gas Journal* 101(31): 46 49.

Cornford C (1994) Mandal Ekofisk (!) petroleum system in the Central Graben of the North Sea. In: Magoon LB and Dow WG (eds.) *The Petroleum System from Source to Trap*, AAPG Memoir No. 60. AAPG., pp. 537 571. Oklahoma: Tulsa.

Demaison G and Huizinga BJ (1991) Genetic classification of petroleum systems. *American Association of Petroleum Geologists Bulletin* 75(10): 1626 1643.

Magoon LB and Dow WG (eds.) (1994) *The Petroleum System from Source to Trap*, AAPG Memoir No. 60. AAPG., p. 655. Oklahoma: Tulsa.

Selley RC (1996) *Elements of Petroleum Geology*, 2nd Edition. San Diego, California: Academic Press.

General Geochemistry

Cornford C (1998) Source rocks and hydrocarbons of the North Sea. In: Glennie KW (ed.) *Petroleum Geology of the North Sea*, pp. 376 462. Oxford: Blackwell Science.

Engel MH and Macko SA (1993) *Organic Geochemistry: Principles and Applications*, p. 861. New York: Plenum Press.

Hunt JM (1996) *Petroleum Geochemistry and Geology*, p. 743. New York: Freeman and Company.

Killops SD and Killops VJ (1993) *An Introduction to Or ganic Geochemistry*, p. 265. Harlow, Essex: Longman Scientific and Technical.

Peters KE and Moldowan JM (1993) *The Biomarker Guide*, p. 363. Eaglewood Cliffs, New Jersey: Prentice Hall Inc..

Tissot BP and Welte DH (1984) *Petroleum Formation and Occurrence*. Berlin: Springer Verlag.

Tyson RV (1995) *Sedimentary Organic Matter: Organic Facies and Palynofacies*, p. 615. London: Chapman and Hall.

Waples DW (1985) *Geochemistry in Petroleum Explor ation*, Geological Sciences Series. p. 232. Dordrecht, NL: D. Reidel Publishing Company (Kluwer Academic Publishers).

Basin Modelling

Beardsmore GR and Cull JP (2001) *Crustal Heat Flow. A guide to measurement and modelling*, p. 324. Cambridge: Cambridge University Press.

Duppenbecker SJ and Iliffe JE (eds.) (1998) *Basin Model ling: Practice and Progress*, Geol. Soc. Sp. Publ. No. 141. p. 245. London: The Geological Society.

Pepper AS and Corvi PJ (1995) Simple kinetic models of petroleum formation Part 1: oil and gas generation from kerogen. *Marine and Petroleum Geology* 12(3): 291 321.

Pepper AS and Corvi PJ (1995) Simple kinetic models of petroleum formation Part III: modelling an open system. *Marine and Petroleum Geology* 12(4): 417 452.

Welte DH, Horsfield B, and Baker DR (eds.) (1997) *Petroleum and Basin Evolution*, p. 524. Berlin: Springer Verlag.

Generation, Migration, Alteration and Volumetrics

Ahsan A, Karlsen DA, and Patience RL (1997) Petroleum biodegradation in the Tertiary reservoirs of the North Sea. *Marine and Petroleum Geology* 14(1): 55 64.

Cooles GP, Mackenzie AS, and Quigley TM (1986) Calculation of petroleum masses generated and expelled from source rocks. *Organic Geochemistry* 10(1 3): 235 245.

George S, Boreham CJ, Minifie S, and Teerman SC (2002) The effect of minor to moderate biodegradation on C_5 to C_9 hydrocarbons in crude oils. *Organic Geochemistry* 33(12): 1293 1318.

Johannesen J, *et al.* (2002) 3 D oil migration modelling of the Jurassic petroleum system of the Statfjord area, Norwegian North Sea. *Petroleum Geoscience* 8(1): 37 50.

Lewan MD, Henry ME, Higley DK, and Pitman JK (2002) Material balance assessment of the New Albany Chesterian petroleum system of the Illinois basin. *Ameri can Association of Petroleum Geologists Bulletin* 86(5): 745 777.

Leythaeuser D, Schwark L, and Keuser C (2000) Geological conditions and geochemical effects of secondary petrol eum migration and accumulation. *Marine and Petroleum Geology* 17(7): 857 859.

Larter SR (1988) Some pragmatic perspectives in source rock geochemistry. *Marine and Petroleum Geology* 5(3): 194 204.

Mann J, Duppenbecker SJ, Langen A, Ropertz B, and Welte DH (1991) Pore network evolution of the Lower Toar cian Posidonia Shale during petroleum generation and expulsion a multidisciplinary approach. *Zentralblatt fur Geologie und Palaeontologie* 1(8): 1051 1071.

Volkman JK, Alexander R, Kagi RI, Rowland SJ, and Sheppard PN (1984) Biodegradation of aromatic hydro carbons in crude oils from the Barrow sub basin of West ern Australia. In: Schenck PA, Leeuw JWD, and Lijmbach GWM (eds.) *Advances in Organic Geochemistry 1983*, pp. 619 632.

Exploration

J R Parker, Formerly Shell EP International, London, UK

Introduction

All commercial accumulations of hydrocarbons are found within, or in close proximity to, sedimentary basins, but not all sedimentary basins contain oil or gas. The presence of large quantities of hydrocarbons in a sedimentary basin requires (Figure 1)

- a source rock rich in organic carbon,
- sufficient heat over long periods of time to convert this organic carbon into hydrocarbons,
- migration pathways that enable the hydrocarbons to migrate away from the source rock,
- a suitable reservoir rock that is sufficiently porous for the hydrocarbons to accumulate and sufficiently permeable to allow the hydrocarbons to be produced at economic rates,
- an effective seal to prevent leakage from this reservoir, and
- a closed structure (a geometrical disposition of the reservoir and seal) to prevent further migration of the hydrocarbons.

Petroleum exploration depends primarily on the scientific understanding of these factors (*see* **Petroleum Geology:** Overview). However, exploration remains a risky business with an uncertain outcome, both in terms of the degree of confidence of realizing a successful outcome (probability of success) and in terms of the amount of hydrocarbons to be found in the case of success (success volume).

Thus, to justify an exploration programme, the expected benefits must exceed the expected costs. The principal benefit is the value of any oil or gas found, while the costs are those incurred in carrying out the programme. Hence, it is an important task of the petroleum geologist to develop quantitative estimates of the probability of success and of the likely recoverable volumes that may be present. In many countries the petroleum laws provide for governments and exploration companies to negotiate contracts that, typically, oblige a company to carry out an exploration programme of agreed scope and cost and entitle it to participate in the development and production of any oil and gas discovered in return for an agreed share of the rewards. Thus, even before exploration starts, a careful assessment of the risk and possible benefits is required.

The basis of this assessment is the so-called expected-monetary-value (EMV) calculation, in which the risk-weighted value of the potential reward should exceed (or at least equal) the risk-weighted cost of the potential loss, as risk without benefit is unacceptable:

$$\mathrm{NPV} \times \mathrm{POS} > \mathrm{AEC} \times (1 - \mathrm{POS})$$

where NPV is the net present value (monetary return in the case of a successful project), POS is the probability of success (chance that the project will be successful), AEC is the abortive exploration cost (the lost investment in the case of failure), and (1 – POS) is the probability of failure (complement of the probability of success).

An exploration programme is, therefore, a carefully planned series of activities aimed at discovering whether oil and gas fields are present in a particular area and, if present, whether they are sufficiently large and productive to be economical. Based on an analysis of the geological information, often involving sophisticated computer modelling of geological processes and an analysis of analogous information from petroleum provinces around the world, the probabilities of finding reserves of certain magnitudes can be assessed. In a well-defined area with fully appraised discoveries, for instance, there may be a high degree of certainty about the current reserves but little chance of finding major additions. In a speculative venture in a little-known area, on the other hand, the chances of finding any hydrocarbons may be low but there is an outside chance of making a very large discovery. However, even with the most sophisticated analysis, the discovery of any quantity of oil and gas cannot be anticipated with certainty, and petroleum exploration remains a classic example of 'decision-making under uncertainty'.

Exploration Methods

Exploration methods are the techniques employed in the search for oil and gas. Their primary purpose is not directly to find oil and gas but to provide evidence about the geology of the subsurface, the interpretation of which may eventually lead to drilling and the discovery of hydrocarbons.

Geological Analysis

The first stage in any exploration programme is to define a 'play' concept to explain how oil and gas may have accumulated in the basin. A play is defined

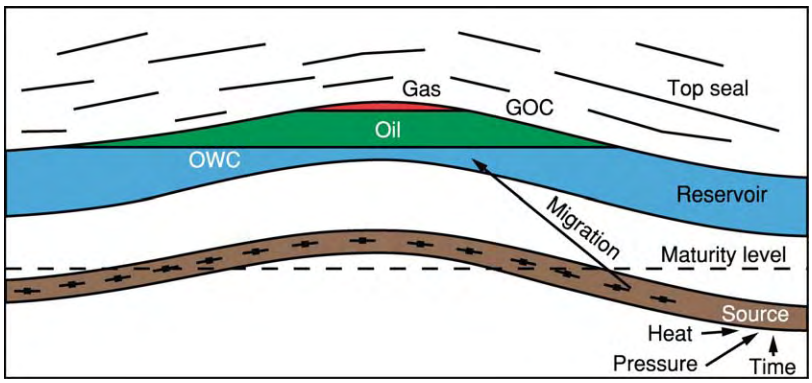


Figure 1 Geological conditions for hydrocarbon accumulation, showing the migration of oil and gas from mature source rock into an initially water filled reservoir. GOC, gas oil contact; OWC, oil water contact. Reproduced with permission of Nautilus Limited.

as a group of prospects (sites of potential hydrocarbon accumulation) and any related fields having common oil or gas sources, migration relationships, reservoir formations, seals, and trap types, thus sharing common elements of geological risk (Figure 2).

The analysis of hydrocarbon plays involves the synthesis and mapping of all the key geological parameters controlling the occurrence of oil and gas, which are principally source, maturation and migration, reservoir, trap, timing, seal, preservation, and recovery (Figure 3).

Geochemical techniques can help to establish the presence of a suitable source rock, while sedimentary geological models can predict the distribution of reservoir and seal rocks. Data are derived from previously drilled wells in the basin (although not necessarily testing an identical play), from traditional field geology (examining surface outcrops within or on the margins of the basin), and from remote sensing (Figure 4) (which may elucidate the geological structure of the area). For previously unexplored areas lacking in 'hard' data, concepts can be tested against analogous basins elsewhere in the world. Play analysis helps to estimate the risks and potential volumes of prospects, as the absence of any one of the factors listed above invariably means that there will be no hydrocarbons in the prospect; a poor development of source rock will mean a limited generation of hydrocarbons, and a poor development of reservoir rock will mean only limited opportunities for entrapment.

Geophysical Analysis

Several geophysical techniques are used to enhance the geological understanding of a basin. Gravimetric (*see* **Analytical Methods:** Gravity) and magnetic surveys (both surface and airborne) (Figure 5) can be used in a reconnaissance mode to delineate the deeper parts of the basin and the major highs before undertaking much more expensive seismic surveys (*see* **Seismic Surveys**). However, seismic surveys are the most important geophysical tool for obtaining a detailed understanding of the subsurface structure, including identifying likely migration paths, inferring the relative timing of trap formation and charge, delineating the geometry and size of potential traps, and even establishing the presence of hydrocarbons themselves.

Seismic surveys involve recording artificially generated shock waves that are reflected or refracted from the different rock strata, and they require a suitable sound source and an appropriate set of detectors. Dynamite has largely been replaced on land by special trucks called vibrators, which have metal plates on their undersides to shake the ground in a controlled manner (Figure 6). At sea, powerful air or water guns are the main energy sources. The reflections are picked up by receivers called geophones on land and hydrophones offshore (Figure 7), which are laid out in such a way as to form directional antennae, termed arrays. Geophones and hydrophones work in a similar way to seismographs, which register earthquakes, but all data are recorded in digital rather than analogue form.

When gathering data in new or little-explored areas, two-dimensional seismic surveys are generally acquired first; with lines spaced as much as 20 km apart, these can cover a great deal of ground at comparatively low cost. If these surveys show interesting structures, then more closely spaced two-dimensional lines can be shot to provide a more detailed picture. However, for an enhanced understanding of the structural and reservoir detail of a prospect, three-dimensional seismic surveys employing a dense grid of lines are commonly used, even in the exploration phase before any wells have been drilled or hydrocarbons discovered. This is

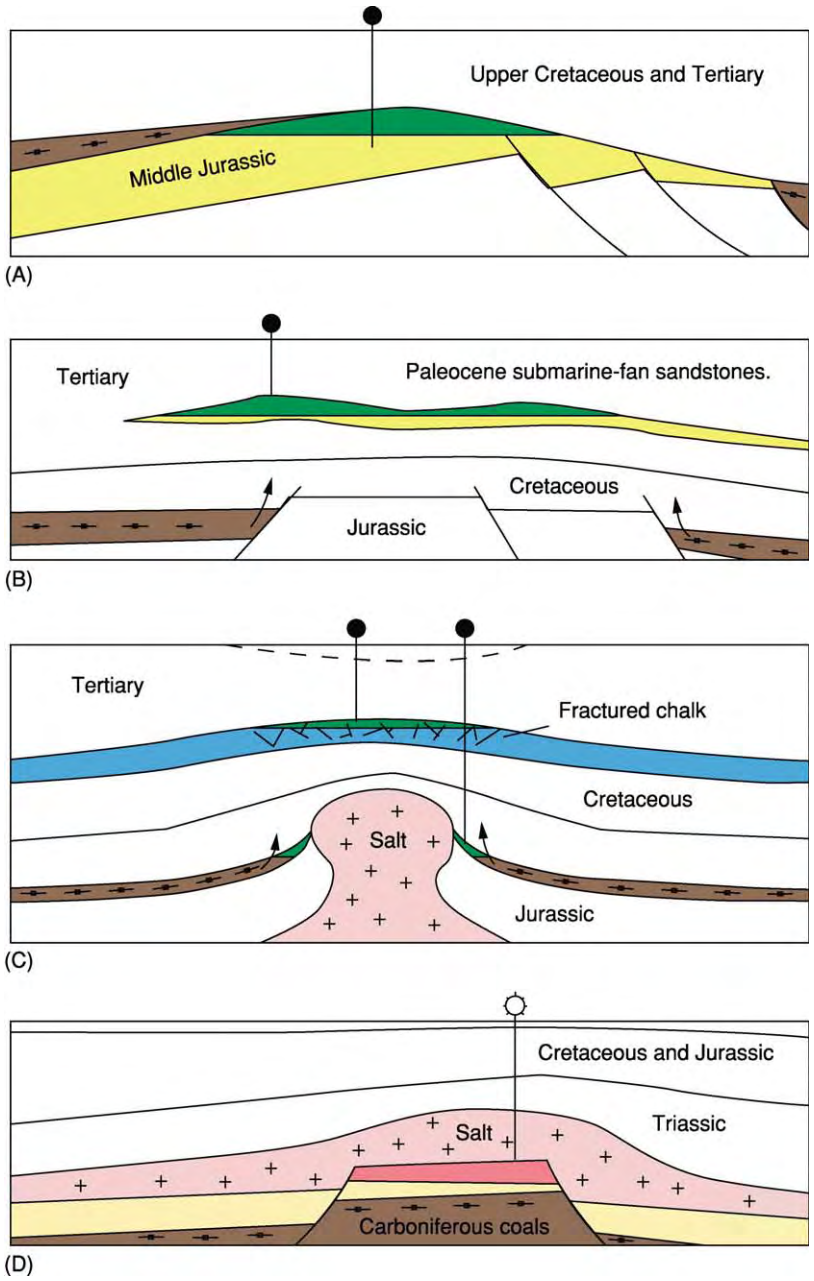


Figure 2 North Sea play types: (A) a truncated tilted fault block, typical of the northern North Sea; (B) a drape or pinchout trap, typical of the central graben; (C) a salt induced structure typical of the central graben; and (D) a horst block capped by salt, typical of the southern North Sea. Yellow, sandstone reservoirs; brown, with –▪– symbol, source rock; pink, salt; blue, chalk reservoir; red, gas; green, oil. Reproduced with permission of Nautilus Limited.

particularly true in geologically complex areas, where the cost of a three-dimensional survey can be considerably less than the cost of a mislocated well.

A traditional two-dimensional seismic survey is shot along single lines to provide vertical slices through the Earth's crust; from a series of parallel and intersecting lines, the three-dimensional structure of the subsurface can be interpreted (Figure 8). However, the assumption is that the reflections seen on each line originate within the plane of the vertical slice, whereas in reality many of the reflections originate from inclined rock layers that may be hundreds of metres outside the plane of the seismic line; also, the gaps between the lines may hide faults and other geological complexities. These limitations have been overcome by the use of three-dimensional seismic surveys, which employ multiple sound sources and simultaneous recording by lines of detectors laid out in a dense grid to

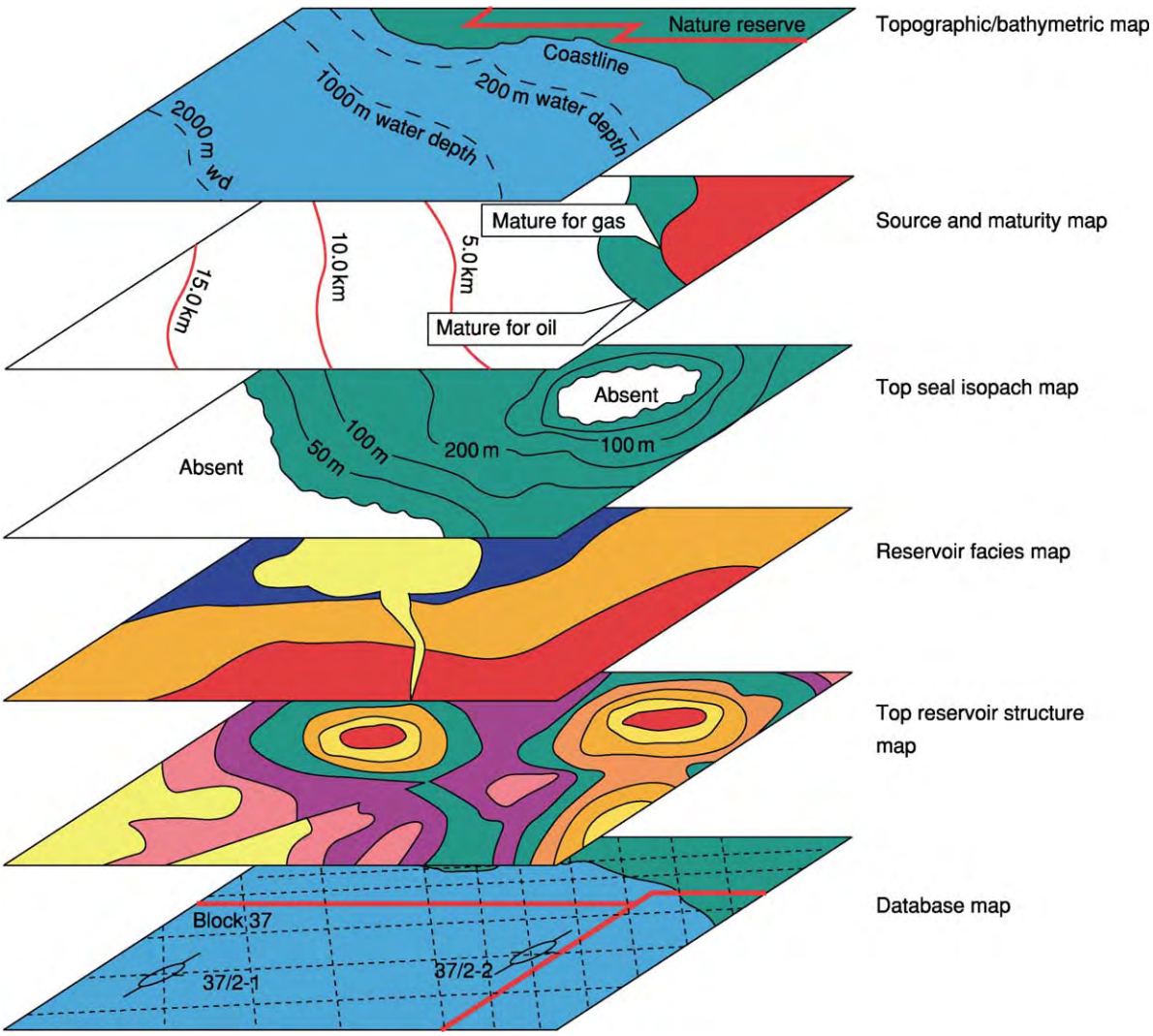


Figure 3 Elements of hydrocarbon play analysis, combining topographic/bathymetric data, source rock maturity, reservoir distribution and structure, and seal. Reproduced with permission of Nautilus Limited.

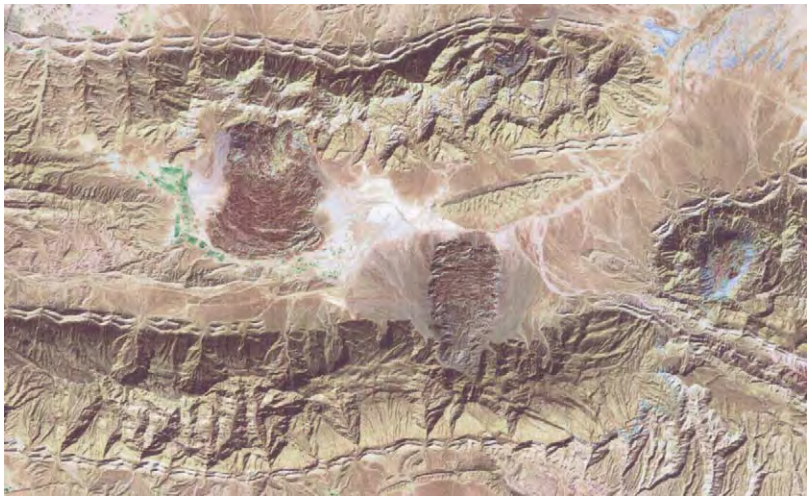


Figure 4 Landsat image of south eastern Iran, showing large, eroded anticlinal structures. The area shown is 45 km × 30 km. Image from Landsat 7, data available from US Geological Survey, EROS Data Center, Sioux Falls, SD, USA.

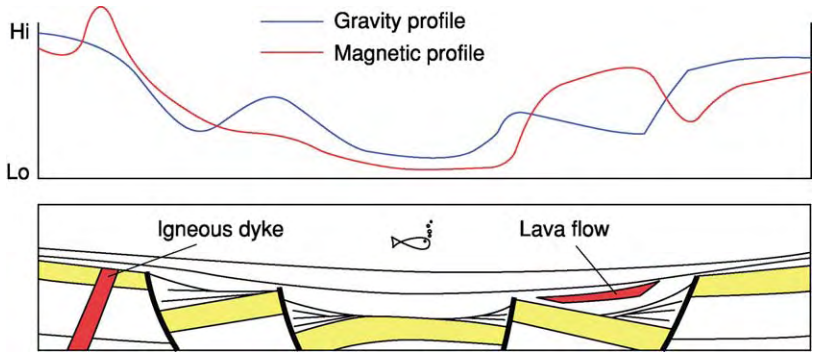


Figure 5 Gravimetric and magnetic profiles (above) which mirror the deeper parts of the basin and the major highs shown on the geological section (below). Note the effect of iron rich rocks, such as lava flows and igneous dykes on the magnetic profile. Yellow, (sandstone) reservoir. Reproduced with permission of Nautilus Limited.

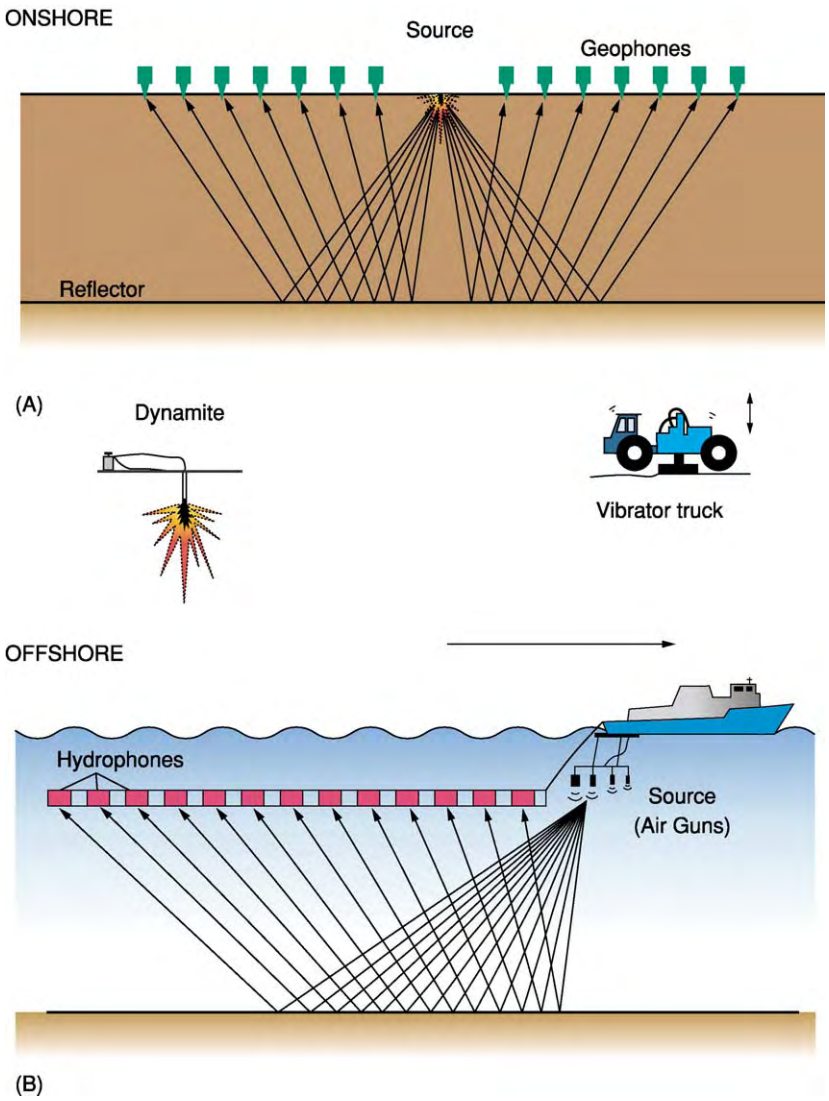


Figure 6 Principles of seismic surveying: (A) onshore, with diagrams showing the various methods of producing the signal, and (B) offshore. Reproduced with permission of Nautilus Limited.

Figure 7 Aerial view of an offshore seismic survey, showing the array of hydrophones being towed behind the ship. Reproduced with permission of Millenium Atlas Company Limited.

acquire a true three-dimensional picture (Figure 9). This development has been made possible by an almost exponential increase in computer power over the years coupled with a concomitant decrease in cost, which has enabled the vast amount of data acquired by three-dimensional surveys to be quickly processed, separating true signals from interference and providing a clearer picture by accurately locating all reflections in their true subsurface positions.

Increasing sophistication in both seismic acquisition and seismic processing techniques allows the direct identification of hydrocarbons in the subsurface. Each interface between separate rock layers will reflect some of the sound waves back towards the surface, and the amplitude of this reflection depends on the nature of that interface. For a porous rock, such as a sandstone, the reflectivity will depend in part on the fluid present in the rock pores, and from this it may be possible to identify oil, gas, and water in the subsurface (Figure 10).

An extension of this concept that is being increasingly applied is the so-called four-dimensional or time-lapse seismic survey. A three-dimensional seismic survey is re-acquired at regular intervals (say, every 3–4 years) over a producing field. The changes in the positions of the reflections indicating the positions of the oil–water contact and the gas–oil contact are used

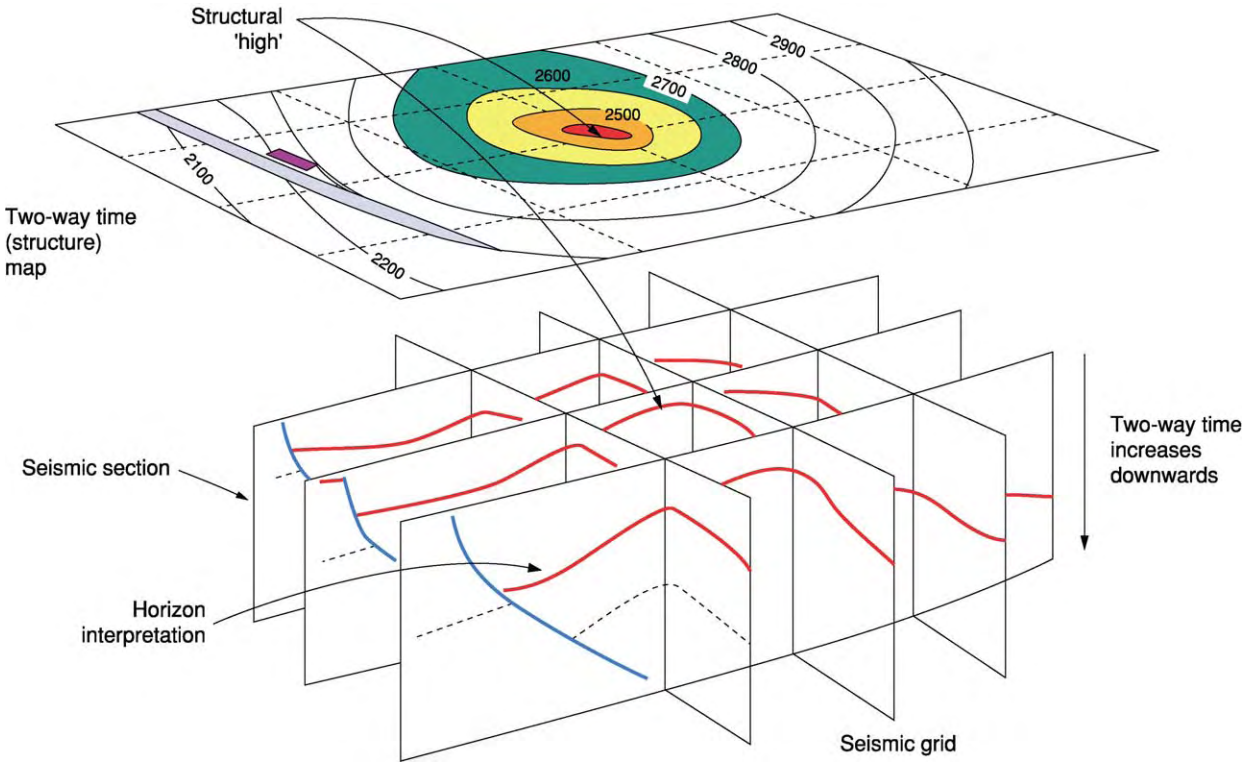


Figure 8 Interpretation of a two dimensional seismic survey. A geological horizon is traced around all the lines of the grid and then plotted in map view. Reproduced with permission of Nautilus Limited.

to monitor the efficiency with which oil and gas are being produced from the field and can identify areas that are not being drained for future infill drilling (Figure 11).

The increase in computer power that has allowed the acquisition and processing of the vast amounts of data resulting from three-dimensional seismic surveys has also revolutionized the interpretation of these data. Computer-based interpretation systems, termed workstations, allow the profiles to be viewed on colour screens and key geological horizons to be identified (Figure 12).

Once they have been identified through the use of appropriate algorithms, these horizons can be automatically 'picked' by the computer by recognizing their 'character' throughout the rest of the dataset. Information from wells and on the regional geological setting can also be integrated. The resulting interpretation can then be viewed and manipulated in a variety of ways: vertical sections can be viewed in any direction; time and horizon slices can be generated that can be horizontal, dipping, or undulating; maps of horizons can be produced; and the frequency and amplitude of the seismic reflectors can be analysed to give clues as to the nature of the lithology and the fluid content of the subsurface layers. The three-dimensional survey can be considered as a cube of data, and computer visualization techniques allow the data to be viewed from any direction and, indeed, to be rotated on screen (Figure 13). The latest development is the use of a so-called immersive 'visionarium' environment, in which observers are surrounded by the data, viewing it in actual three dimensions using special spectacles (Figure 14).

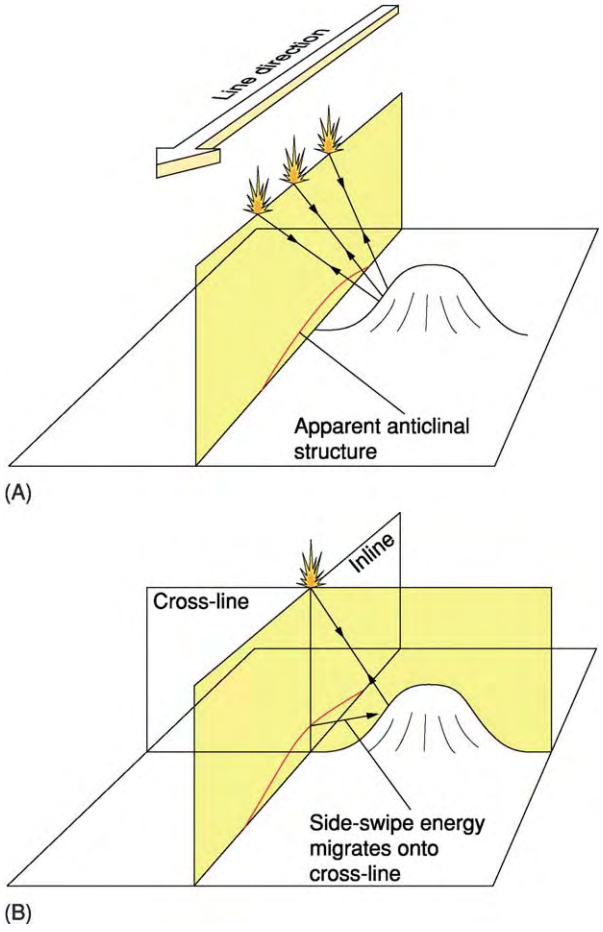


Figure 9 Correct positioning of reflections by the use of three dimensional seismic surveys. (A) A two dimensional seismic survey may contain spurious information generated from out of the plane of the profile being acquired. (B) A three dimensional seismic survey and its subsequent processing, enables all information recorded to be correctly located in the subsurface. Reproduced with permission of Nautilus Limited.

Prospect Appraisal Systems

A number of methods have been developed to attempt to translate complex geological data, derived from the geological and geophysical studies described above, into the set of numbers needed for decision-making. The particular appraisal method used depends on the size of the unit to be assessed (from global to part of a structure) and the amount and type of information available (from sparse regional data to detailed drilling results, or data of a more statistical nature such as the number of wells drilled and the amount of reserves found). The methods can be broadly grouped under three main headings: subjective methods, statistical methods, and deterministic models.

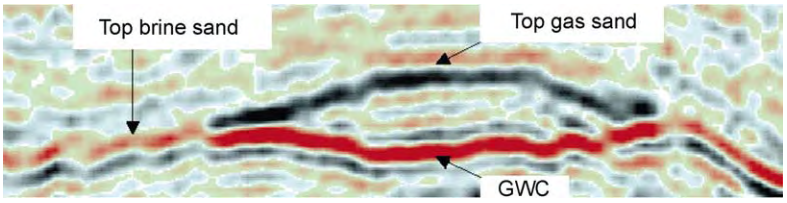


Figure 10 Gas effect seen on seismic showing the change in amplitude of the signal (so called 'bright spot'). There is a phase reversal at the gas water contact (GWC) and a time sag. Reproduced with permission of BG Group PLC.

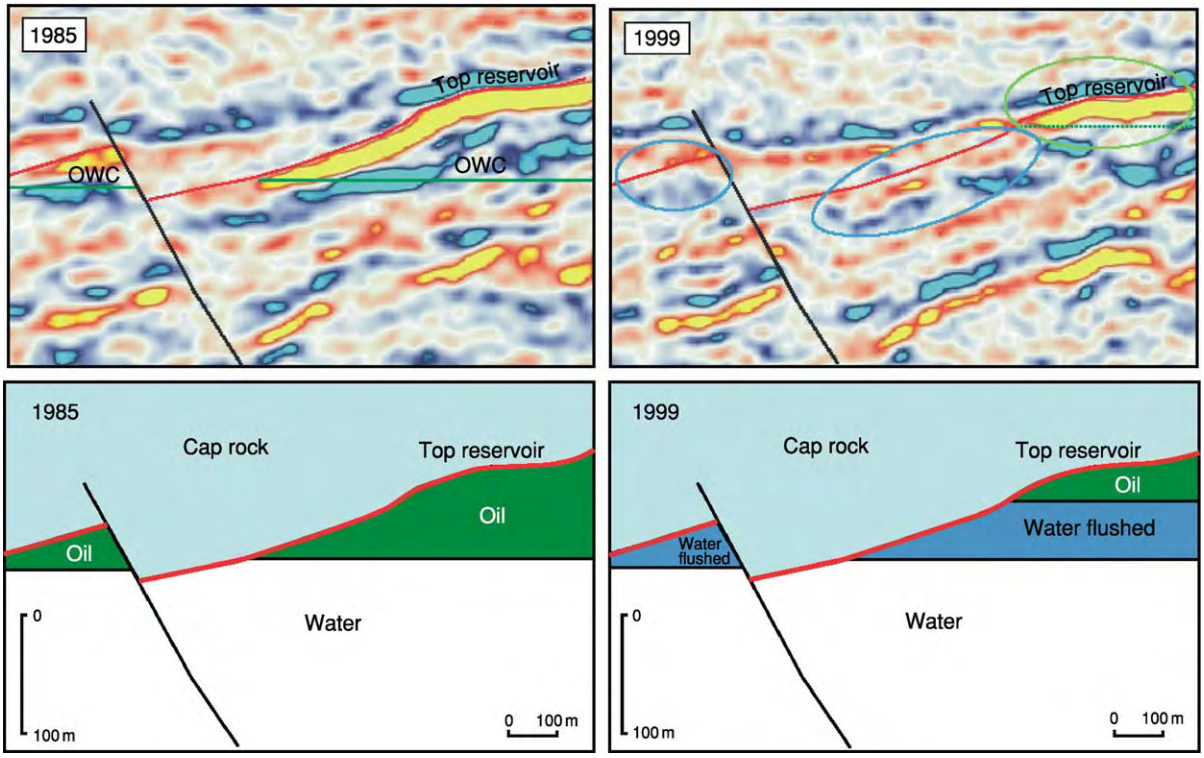


Figure 11 Time lapse (four dimensional) seismic survey over a North Sea field; the changes in amplitude result from a significant depletion of the oil due to extraction. OWC, oil water contact. Reproduced with permission of Millenium Atlas Company Limited.

Subjective Methods

Subjective methods result from the implicit intuitive thought processes of a single individual or group of individuals and can be formidable tools when a substantial body of expertise is available. However, there is the problem of bias: usually people are more certain about the unknown than they can afford to be on the basis of the information available, and some people are consistently pessimistic, while others are optimistic. There is also the problem of consistency of knowledge: local knowledge can override possibly more relevant but less well-known worldwide information.

Statistical Methods

Statistical methods attempt to extrapolate past experience using a variety of statistical techniques. Some depend on the validity of an essentially linear extrapolation, for example a hydrocarbon richness per unit volume of sediments in explored basins, which can then be applied to unexplored basins. Other methods attempt to improve this extrapolation by considering past exploration results, such as the decline of success ratio through time and the decline of field size as a function of exploration effort. In the example shown in Figure 15, the larger fields are discovered first (the so-called 'creaming' effect), but, although the average size of the fields found is decreasing, the number found is still increasing.

One of the best known of such studies is that of M King Hubbert, who in 1956 predicted that US oil production would peak in the early 1970s; this proved to be correct (the so-called 'Hubbert's Peak'; Figure 16).

Disadvantages of these methods are that a certain degree of exploration maturity is required for the necessary historical data to be available, the results may be strictly applicable only to the particular basin or province analysed, and the interrelationships of the variables chosen are not always known and correlations may be fortuitous.

Deterministic Models

Deterministic models attempt to integrate all the relevant geological concepts and research findings on the processes of hydrocarbon generation, migration, accumulation, and retention into a coherent model that will provide quantitative estimates of the hydrocarbon volumes likely to be present and the risks involved. These methods are now widely employed in the industry. The problem is that the information available is usually insufficient for precise estimates of the required variables to be made, and, in such

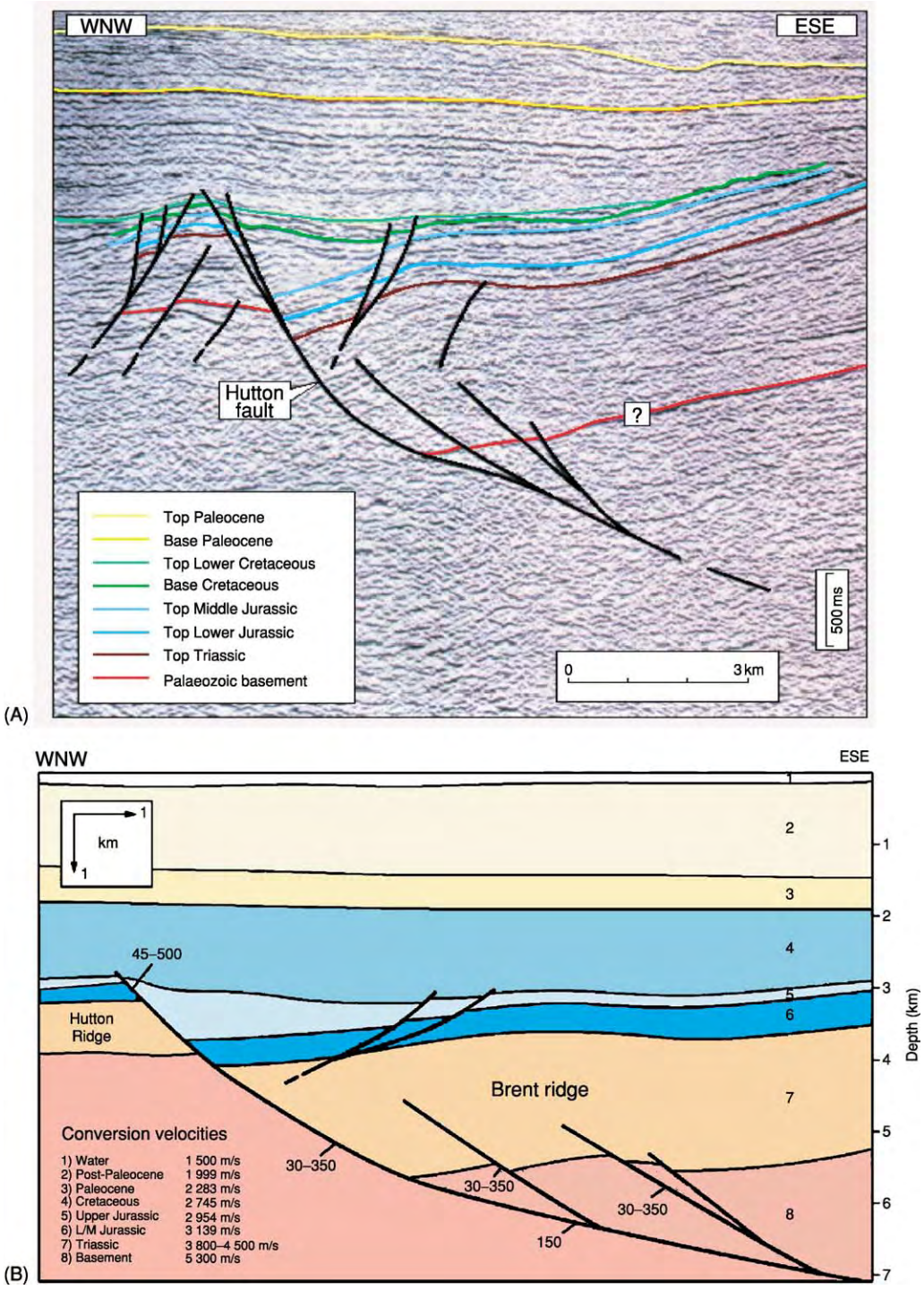


Figure 12 (A) Geological interpretation of a seismic line, and (B) depth converted to show the true fault geometry. Reproduced with permission of Millenium Atlas Company Limited.

cases, a single numerical estimate is often more misleading than helpful. Hence, the inputs will almost always be in the form of probability distributions, and the use of repetitive calculations, each time using an equally likely choice (the so-called 'Monte Carlo' technique; Figure 17) for the input parameters, results in a series of equally likely outcomes, reflecting the uncertainties of both the inputs and the applied formulae. In addition, despite many advances in the understanding of the geological processes involved, any model, although logically consistent and based on the best of present knowledge, is

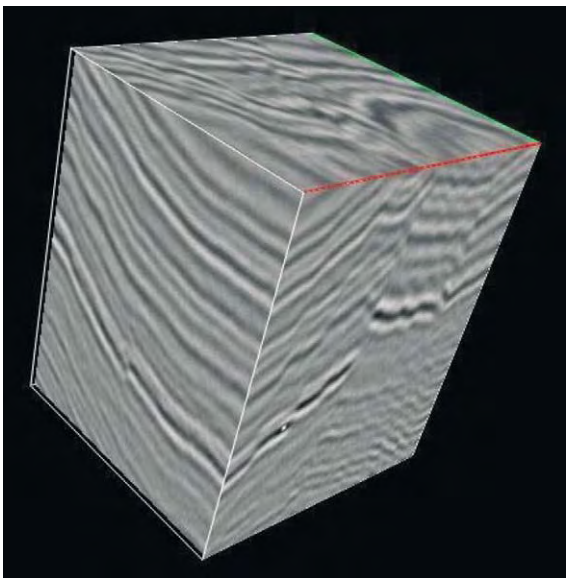


Figure 13 Three dimensional seismic cube which can be re viewed from any direction, or rotation, on the computer workstation screen. Reproduced with permission of Ikon Science Limited.

likely to be only partially correct because there are variables and processes that are unknown, cannot be measured with sufficient precision, or cannot be accommodated in the model. Hence, it is important to calibrate the output from these models against real-life situations.

Exploration Drilling

Assuming that the geological and geophysical work outlined above has identified a 'prospect' – a possible hydrocarbon-bearing structure – it has then to be decided whether to drill an exploratory well, which is the only way to prove the actual presence of hydrocarbons. Factors involved in this decision are the likelihood of hydrocarbons being present in the prospect, the extent – and hence economic viability – of any hydrocarbons found, the cost of development if hydrocarbons are found, and political and economic considerations relating to the concession and the country in which the prospect is situated.

Figure 14 Three dimensional seismic interpretation system (so called 'visionarium').

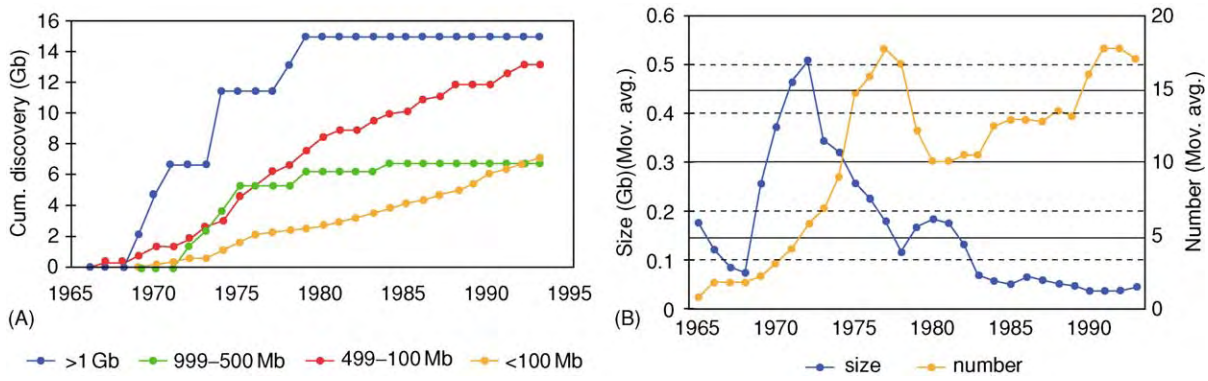


Figure 15 Discovery by (A) field size classes, showing that the larger fields are discovered first (so called 'creaming effect'), and (B) average size and number of fields showing that although the size of fields found is still decreasing. North Sea data, both plotted against time.

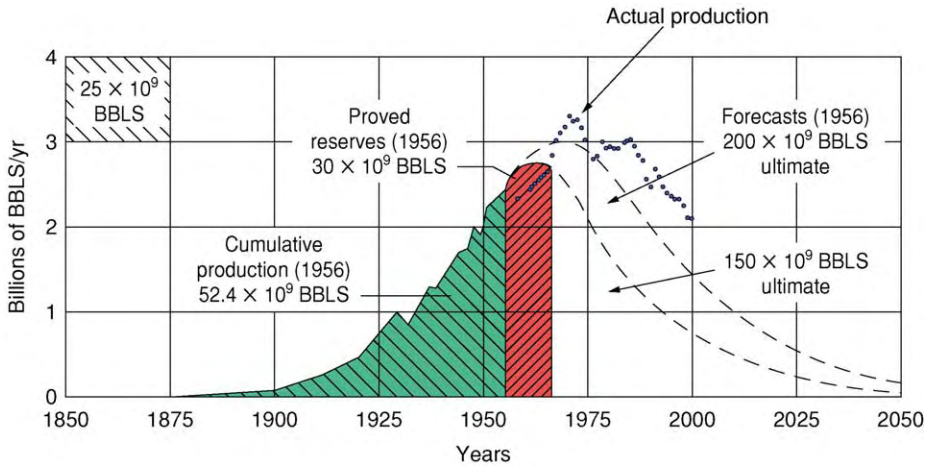


Figure 16 Hubbert's Peak: predicted and actual US oil production.

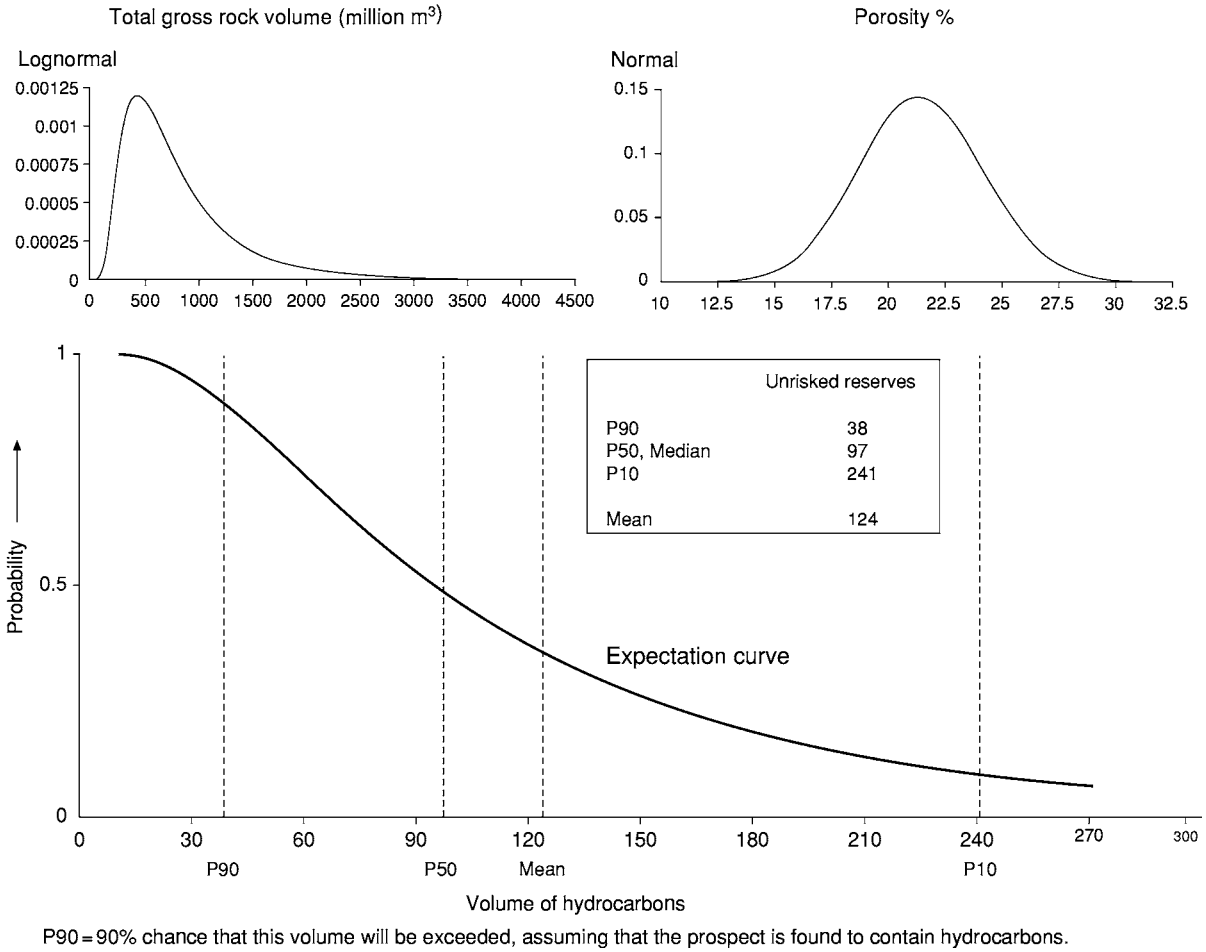


Figure 17 Monte Carlo technique, combining input probability distributions describing total gross rock volume and porosity (illustrated in the top diagrams) together with those describing net to gross ratio, hydrocarbon saturation, formation volume factor and recovery factor (not illustrated) to give an expectation curve (cumulative frequency distribution; bottom diagram). Reproduced with permission of Nautilus Limited.

In addition to confirming the presence of hydrocarbons, an exploration well provides information from which the size of the accumulation can be assessed, the development of the field planned, and further exploration in the area undertaken. Rock cuttings brought to the surface by the drilling mud and specially taken core samples contribute to an understanding of the geological history of the sedimentary basin and, specifically, of the nature and characteristics of the reservoir rocks. Wireline logging (Figure 18), using special sondes to measure the electrical, acoustic, and radioactive properties of the rocks, allows key physical properties of the rocks, such as porosity and the fluid content – of oil, gas, or water – within the pore space, to be determined. Other logging tools can detect the presence of fractures (which can be important for the flow characteristics of a reservoir), the dip (or inclination) of the rock layers, and sedimentary features (such as cross-stratification). If this information indicates a potentially commercial accumulation, then a flow test can be undertaken, in which the oil or gas is flowed to the surface: from flow and pressure measurements, further insights into the porosity and, particularly, permeability of the reservoir and the likely performance of wells under operational conditions can be gained.

Further surveys and additional drilling are then undertaken to appraise the size of the accumulation and reduce the geological uncertainties; this information is used to plan the development of the field.

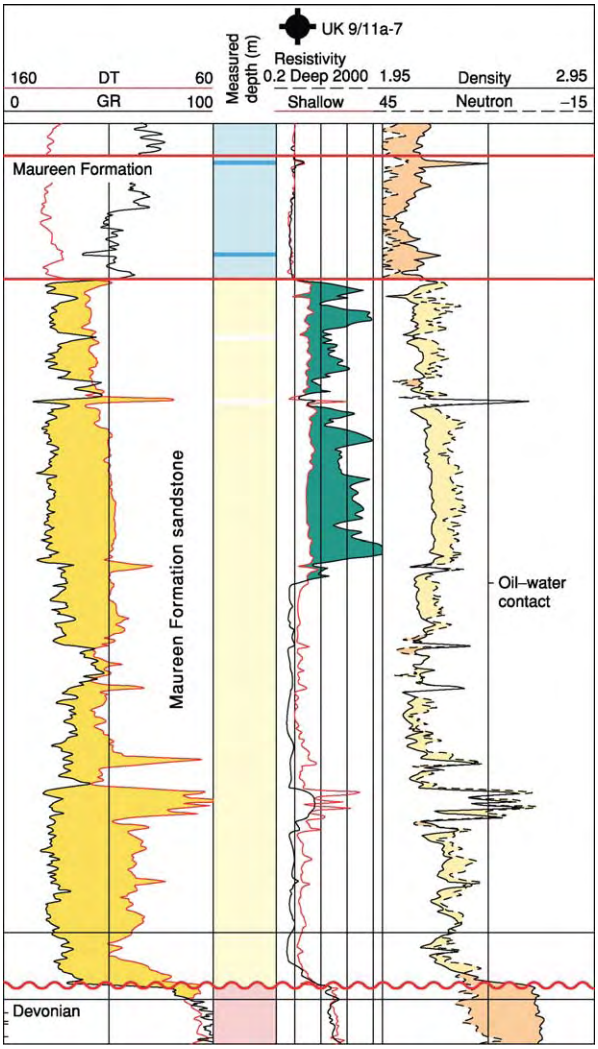


Figure 18 Wireline log response of a hydrocarbon bearing sandstone reservoir, with a shale seal above and water saturated reservoir below the oil. DT, interval transit time; GR, gamma ray. Reproduced with permission of Millenium Atlas Company Limited.

Exploration Costs

Onshore, the exploration costs depend very much of the nature of the terrain: operations in remote mountainous or jungle areas are much more expensive than those in flat desert country. Operations in urban and agricultural areas and in areas of environmental sensitivity are expensive, as access can be difficult and strict regulations relating to noise, pollution, and environmental protection must be adhered to. Offshore, costs are less variable, though remote and less climatically favourable locations inevitably attract higher costs (Table 1).

Petroleum Agreements

Apart from the onshore USA, where individuals can have title to the mineral rights, in most countries these rights are vested in the state, and exploration for oil and gas can be carried out only under licences or contracts granted by the state. Petroleum agreements can be divided into two main types: licences (including leases, permits, and concessions) and production-sharing contracts.

Under a licence agreement, a company is granted exclusive rights to explore particular areas. The company finances the exploration campaign and, if the exploration is successful, the development cost and in return is entitled to dispose of the production, sometimes subject to a deduction of a royalty in kind and an obligation to supply the domestic market. The state obtains revenues from the royalty and from taxation, often through special petroleum taxes. A variation on this type of agreement is a joint-venture agreement, where the state shares with the company the risk and expenses of the development phase; generally the company will carry the project at its own cost through the exploration phase.

Table 1 Indicative costs of exploration (for the year 2003) (all values are in US dollars)

Onshore jungle terrain	
Seismic cost per two dimensional kilometre	7 000
Seismic cost per three dimensional square kilometre	12 000 (but can be significantly higher)
Well cost (to 3500 m)	3 5 million
Onshore desert terrain	
Seismic cost per two dimensional kilometre	2 000
Seismic cost per three dimensional square kilometre	4 000
Well cost (to 3500 m)	1.5 2 million
Onshore urban or agricultural environment	
Seismic cost per three dimensional square kilometre	12 000
Offshore	
Seismic cost per two dimensional kilometre	600
Seismic cost per three dimensional square kilometre	5 000
Well cost (to 3500 m)	6 8 million (but can be significantly higher for deep wells encountering high pressure and high temperature, or in remote locations; up to 25 million)

In a production-sharing agreement, the company carries the exploration risk and funds all the development and operating costs. These costs are then recovered from part of the production, known as 'cost oil'. The remaining 'profit oil' is then split in a pre-determined manner between the state and the company.

A further type of agreement is the service contract, in which the state contracts for a service from the company for which the company receives a fee; this can be, for example, a set fee per barrel produced or a percentage of the hydrocarbons produced while providing the service.

Petroleum agreements take many forms, are increasingly complex, and are often a combination of elements from the different types. The main considerations for the company are the investment at risk, its ability to manage the operations, access to oil in the event of success, and the economic return on the investment.

Postscript

The development of new exploration techniques continues to improve geologists' and geophysicists' understanding of petroleum geology and to increase the efficiency of exploration by allowing wells to be sited more accurately and with a greater chance of success. However, even if the geological conditions for the presence of hydrocarbons are promising, exploration remains a high-risk business and investments are made in exploration many years before there is any prospect of producing the oil: those ventures that are successful must generate sufficient profit to pay for the unsuccessful ventures, both past and future. Thus, the fiscal framework established by states is vital to the commercial success of any exploration venture. The interests of governments wishing to develop their petroleum resources and the interests of companies as risk takers have much in common; petroleum-exploration strategies must take these mutual interests into account.

See Also

Analytical Methods: Gravity. **Petroleum Geology:** Overview; Chemical and Physical Properties; The Petroleum System; Production; Reserves. **Seismic Surveys**.

Further Reading

Deffeyes KS (2001) *Hubbert's Peak: the Impending World Oil Shortage*. Princeton: Princeton University Press.

Gluyas JG and Swarbrick RE (2004) Petroleum Geoscience. Oxford: Blackwell.

Johnson HD and Fisher MJ (1998) North Sea plays: geological controls on hydrocarbon distribution. In: Glennie KW (ed.) *Petroleum Geology of the North Sea: Basic Concepts and Recent Advances*, pp. 463 547. Oxford: Blackwell.

Poelchau HS, Baker DR, Hantschel Th, Horsfield B, and Wygrala B (1997) Basin simulation and the design of the conceptual basin model. In: Welte DH, Horsfield B, and Baker DR (eds.) *Petroleum and Basin Evolution,* pp. 5 70. Berlin: Springer Verlag.

Selley RC (1996) *Elements of Petroleum Geology,* 2nd edn. San Diego: Academic Press.

Shell Briefing Service (1994) *Upstream Essentials*. London: Shell International Petroleum Company Limited.

Steinmetz R (ed.) (1992) *The Business of Petroleum Exploration*. Treatise of Petroleum Geology. Handbook of Petroleum Geology. Tulsa: American Association of Petroleum Geologists.

Production

K J Weber, Technical University, Delft, The Netherlands
L C van Geuns, Clingendael International Energy Programme, The Hague, The Netherlands

Introduction

The geological activities following the discovery of a petroleum accumulation commence with a thorough analysis of the find in terms of structural definition, reservoir rock characteristics, and accumulation conditions. This is followed by the selection of appraisal well locations and the study of the results. If considered commercially viable, a multidisciplinary team, including specialists of all disciplines in petroleum engineering, designs an initial development plan. At this stage, the volume of petroleum-in-place is estimated and a dynamic model is built to test the production capacity and producible reserves of the field, which may be generated by a series of alternative development plans.

After executing the selected initial development plan, the reservoir model is updated with new data. The increasing information on field performance is used to calibrate the model by history matching. In certain favourable cases, fluid movements can be observed on additional three-dimensional (3D) seismic data (*see* **Seismic Surveys**). Further development activities, such as infill drilling, side tracking, recompletions, or fluid injection schemes, are planned on the basis of the updated model. Detailed field studies by integrated teams of petroleum engineers and production geologists are carried out periodically until the ultimate stages of the field life (Figure 1).

Reservoir performance is a function of the reservoir characteristics and the petroleum properties. The study of reservoir characteristics is the primary task of the production geologist. An overview is given of the influence of various reservoir heterogeneities on fluid flow. To analyse these characteristics, and to quantify the results, requires a wide range of techniques. These cover such a large field of subjects that the work has to be carried out by a group of specialists in seismology, sedimentology, structural geology, geostatistics, and petrophysics. In addition, a good understanding of the principles of reservoir and production engineering is required to be able to function successfully in a multidisciplinary team.

The production geological activities at the various stages of field development have a number of different purposes which all require geological input of some kind (Figure 2). This is illustrated by the different techniques and modelling methods that are employed at each stage (Figure 3). The pace of development of new techniques, in particular integrated computer systems, is very fast. As a result, the production geological activities are centred around workstations being served by a database containing all relevant information (Figure 4).

Historical Development of Production Geology

Production geology as a separate petroleum engineering discipline was only generally recognized in the 1950s. Early field development, with shallow, closely spaced wells and only cuttings and surface outcrop observations, provided little scope for reservoir studies. In the USA and Russia, where grid drilling was and still is in fashion, most companies did very little production geology work in the appraisal stage. In Indonesia, where complicated structures were encountered, the need was felt for more geologist involvement at an early stage.

Around 1915, the Bataafse Petroleum Company (Royal Dutch) started to engage mining engineers with the particular task to improve cooperation between production and geology to achieve a more efficient exploitation of the field. The reports from that time show that structural interpretation and detailed mapping of the producing horizons constituted the main tasks. In these pre-logging and pre-seismic days, correlations were carried out with the aid of cuttings glued to narrow planks.

The wireline logs developed by the Schlumberger brothers in 1927 progressively replaced these early methods. Just before World War II, effective reflection seismic methods came into use, although any sophisticated processing had to wait for the availability of more powerful computers after 1960. This is also the time that production geology, also called 'development geology', began to be more generally employed, although, in most companies, the geologists were still firmly attached to the exploration departments.

In the meantime, sedimentology had made great strides forwards and more attention was given to the internal architecture of the reservoir. Core descriptions and analysis were improved. Research studies of outcrops and recent sedimentation of both

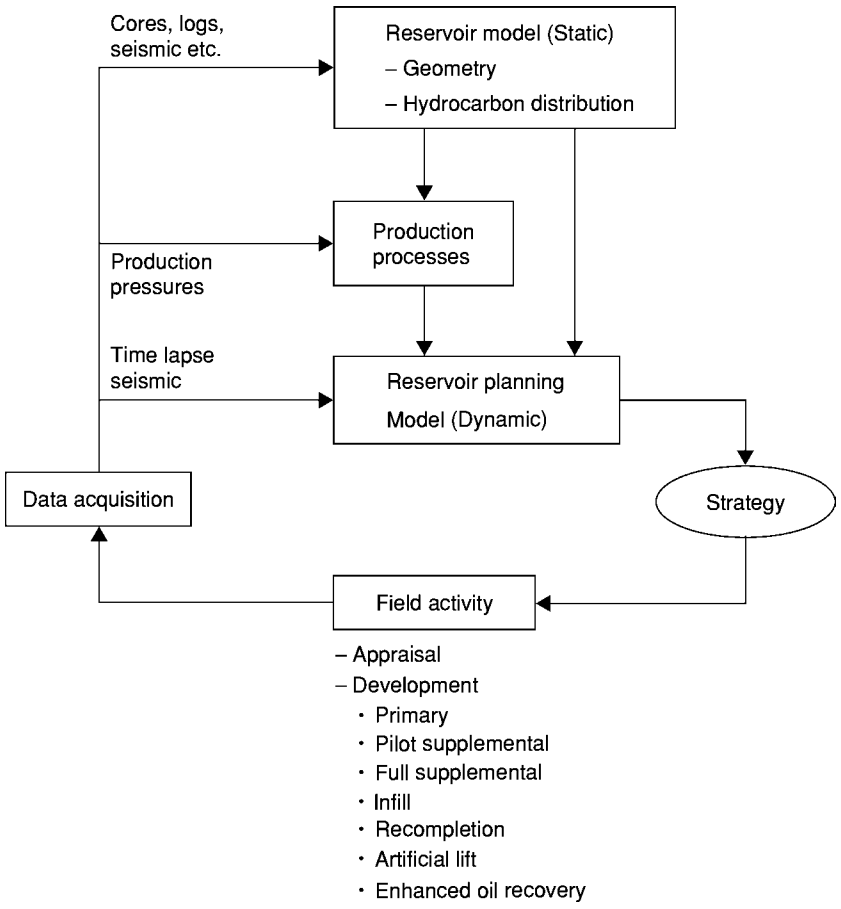


Figure 1 Field study cycles.

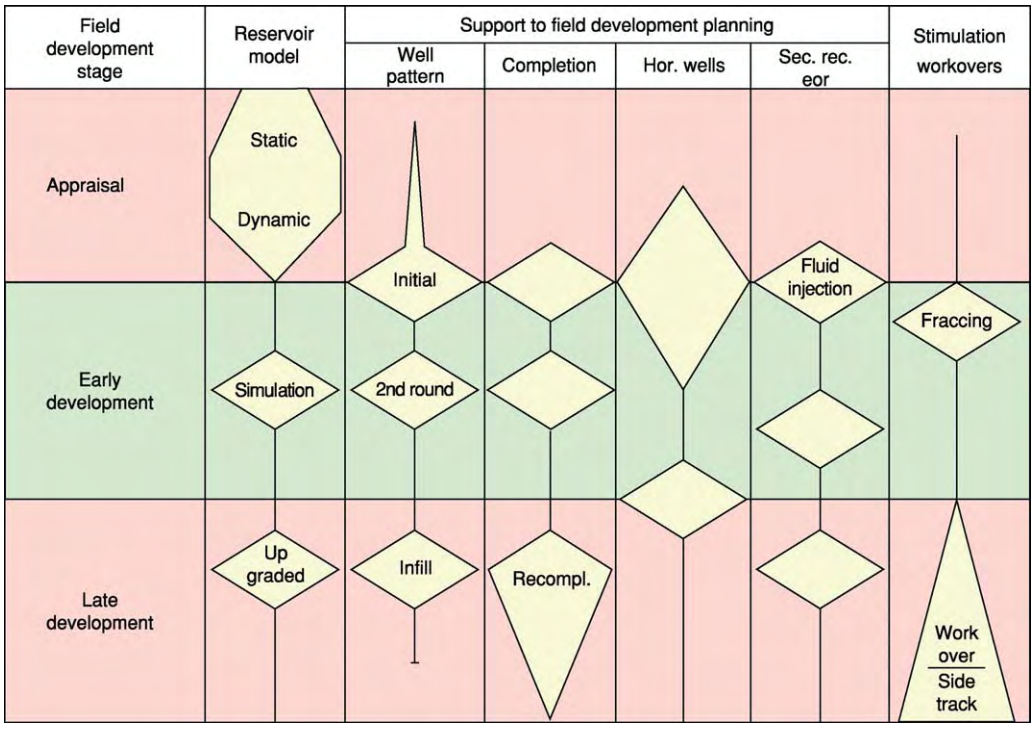


Figure 2 Purposes of reservoir characterization. Courtesy of Shell.

Field development stage	Seismic techniques	Core/log techniques	Modelling	
			2D	3D
Appraisal	Structure Fault sealing Hydrocarbon indications Sequence stratigraphy	Facies Poro/Perm Pore geometry Lithoclass (Cores) Lithoclass (Logs)	Stratigraphy Facies distribution	Probabilistic → Sensitivity → Connectivity
Early development	Reservoir architecture Lithology Porosity	Fracture types + orientation Borehole imaging	Well-to-Well Correlations	Hybrid model → Simulation input
Late development	Time lapse Well-to-Well Tomography	Sand body geometry Fault sealing	Distribution of saturation	Deterministic model → Remaining oil distribution

Figure 3 Reservoir modelling techniques during successive development stages. Reproduced with permission from Weber KJ (1995) Visions in reservoir management what next? In: *Reservoir Characterisation: Integration of Geology, Geophysics and Reservoir Engineering, The Third JNOC TRC International Symposium, February 2 23, Technical Research Centre, Chiba, Japan*, pp. 1 15. Chiba: Japan National Oil Corporation.

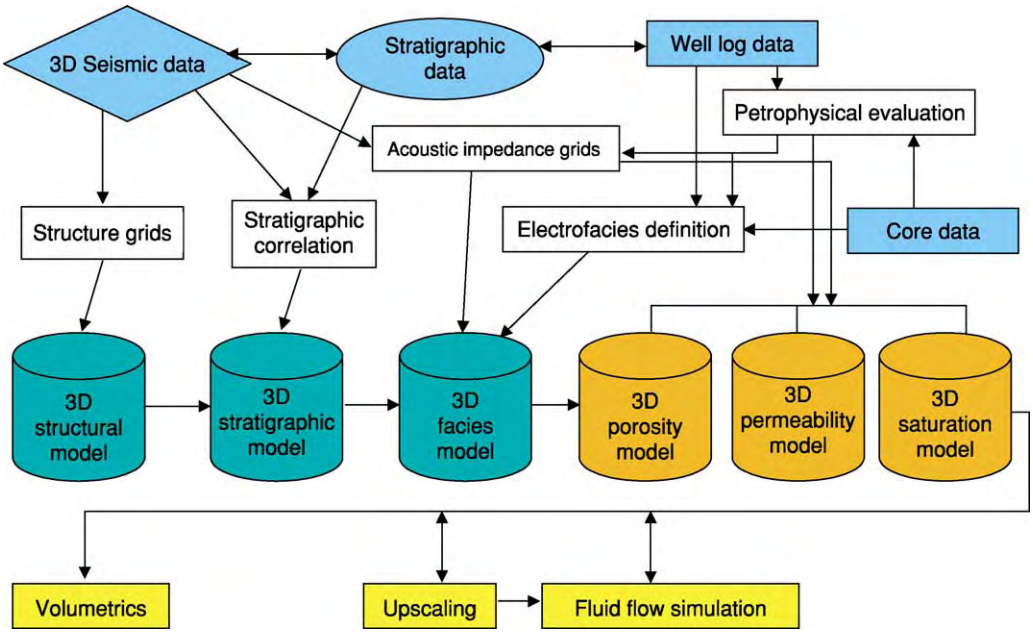


Figure 4 High level workflow for computerized reservoir modelling. Courtesy of Shell.

clastics and carbonates were carried out by several major oil companies, in combination with laboratory studies of diagenetic processes. Rock classification systems such as the 'Archie classification' appeared, providing the link between rock texture and permeability. The introduction of induction-, latero-, formation density- and dip-meter logs greatly improved the quantitative assessment of hydrocarbon volumes

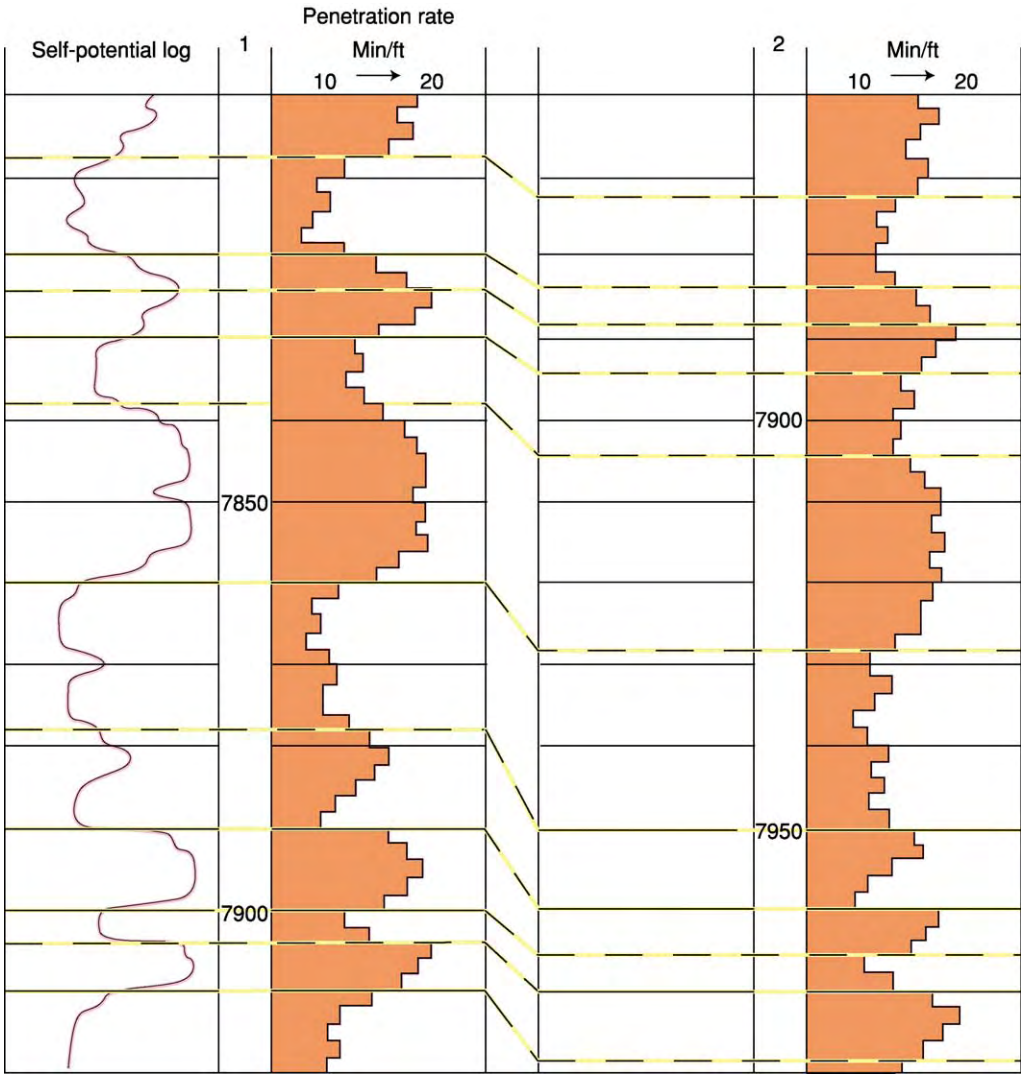


Figure 5 Penetration log correlation.

and distributions. Between 1960 and 1980, most of the presently used production geological methods were developed.

As a typical example of a classical production geological method, Figure 5 shows a penetration log correlation. The different rates of penetration of sandstones and shales (whilst drilling with constant speed and weight) can often be related to spontaneous potential (SP) or gamma-ray (GR) log curves, and thus the progress of wells could be monitored prior to the introduction of modern logging-while-drilling (LWD) much later. Further refinements in structural and reservoir imaging came with the spectacular advance of seismic acquisition, processing, and interpretation. Seismic attribute analysis is now one of the major production geological techniques. Sophisticated borehole imaging logs formed another new source of detailed information. The increase in computer capacity also led to reservoir simulation based on complex heterogeneous reservoir models comprising very large numbers of grid cells, which, in turn, required realistic well-calibrated geological models. This promoted the use of geostatistical methods and stochastic modelling. The present trend is to use a combination of all relevant reservoir data in integrated databases from which visualization can be generated, simulation models can be constructed, and well trajectories can be planned.

Reservoir Characteristics Influencing Fluid Flow and Storage Capacity

Controlling Reservoir Characteristics

The quality of a reservoir is directly related to the available pore space (porosity) and the capacity to

conduct fluid flow through the pores (permeability). Porosity depends on texture (grain size distribution, grain shape, and packing) and the degree of diagenesis (compactions, leaching, cementation). Permeability is a function of the size, shape, and distribution of pore spaces and their interconnectedness. Rocks with a similar texture that have been subjected to similar diagenetic processes will generally show a good correlation between porosity and permeability. However, for the same porosity, significantly different permeabilities may be found in the same reservoir because of facies-related textural differences and/or diagenetic variations.

There are a series of geological features that influence fluid flow and hydrocarbon recovery (Figure 6). Faults, fractures, reservoir architecture, internal zonation of genetic units, low permeability intercalations, sedimentary structures, and pore shape all have an impact on a different scale. The drive mechanism can also be influenced by these factors because they control the pressure support from the aquifer surrounding the reservoir. An accurate evaluation of a reservoir has to take into account all of the relevant heterogeneities at the right scale. The resulting 3D reservoir model forms the basis for volumetric estimates of the trapped hydrocarbons and simulations of

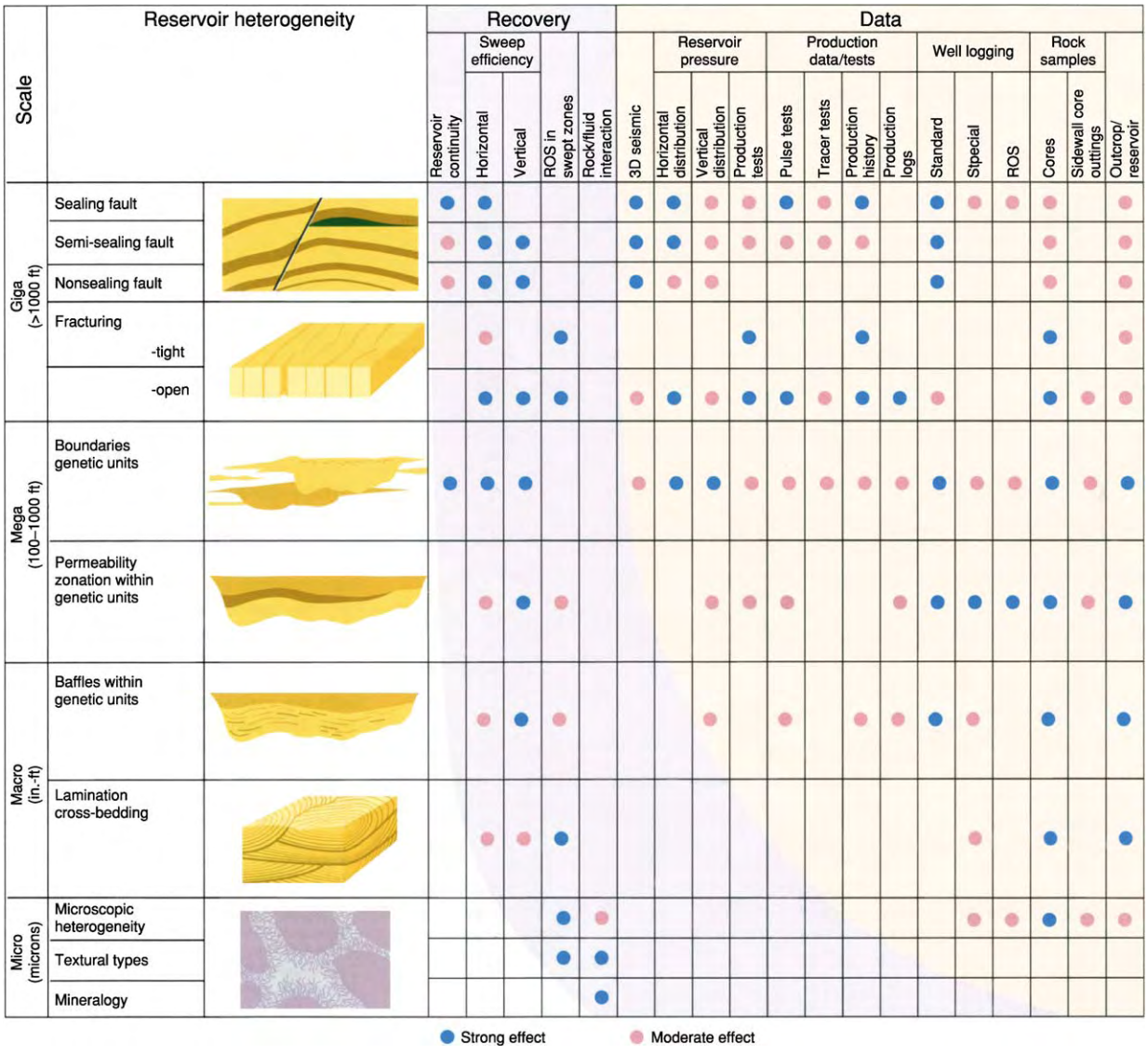


Figure 6 Reservoir heterogeneity influencing fluid flow and hydrocarbon recovery, and relevant data for analysis. ROS, Residual Oil Saturation. Modified with permission from Weber KJ and van Geuns LC (1990) Framework for constructing clastic reservoir simulation models. *Journal of Petroleum Technology* 42(10): 1248 1253, 1296 1297. © Society of Petroleum Engineers.

reservoir performance. In this way, the appraisal campaign can be planned and the optimum development scheme can be determined.

Faults and Fractures

Faults have a significant influence on hydrocarbon migration and trapping. The sealing capacity of a fault can be controlled by cataclasis and diagenesis in the fault zone, juxtaposition against a tight layer, or smearing of clay along the fault. In deltaic settings, where multiple reservoirs separated by shales are common, the determination of the clay smear potential along faults is particularly useful.

Fractures can have a positive or a negative effect on fluid flow. Cemented fractures compartmentalize the reservoir and reduce recovery. Swarms of shear fractures are common along many faults and can significantly reduce productivity over a wide zone. Open fractures, which are mainly associated with folded carbonate reservoirs, can increase productivity enormously, although they may also be the cause of early water breakthrough to the wells.

Clastic Reservoirs

Clastic reservoirs occur in a wide variety of sedimentary settings, ranging from almost homogeneous beds to an intricate 3D network of sinuous sand bodies. From an observation of the existing reservoir configurations, it appears that most clastic reservoirs can be classified into one of three basic architectural types – layer-cake, jigsaw, or labyrinth – which represent a decreasing order of connectivity (Figure 7). Correlation is comparatively easy in layer-cake reservoirs, such as barrier bars or shallow marine sheet sands. Thus, such reservoirs can already be modelled quite realistically in the appraisal stage. At the other end of the scale are the labyrinth-type reservoirs, such as low-sinuosity distributary channel complexes, which require a dense well spacing to achieve a deterministic correlation.

If we add common internal permeability distributions and discontinuity patterns, the impact of the reservoir types on sweep efficiency can be shown (Figure 8). In layer-cake reservoirs, not only the sands but also the shales can be very continuous, preventing vertical flow. High permeability at the base of point bars can lead to early water breakthrough, while the reverse, as in barrier bars, can have the opposite effect. Jigsaw reservoirs, with interlocking sand bodies with contrasting properties and local baffles to vertical and horizontal flow, often show uneven drainage and erratic flood fronts. Labyrinth reservoirs reflect their discontinuous nature in very uneven sweep efficiency and bypassing of major oil volumes.

Carbonate Reservoirs

Carbonate reservoirs are generally more heterogeneous than clastic reservoirs, because the original organic sediments can differ widely in their grain size and composition, but also because of the large influence of diagenesis. The original sedimentary setting can be split into four main types: the deeper chalks and turbidites; the shallow-water build-ups; platform/shoal sediments; and ramp/nearshore packages (Figure 9). These reservoirs are sometimes left with rock characteristics closely related to their original properties, but often considerable changes are effected by the various diagenetic processes. However, in the majority of cases, the original material with a similar texture and composition will be transformed into the same rock type. There may be a complete reversal of the original porosity and permeability distribution. A porous reef may become tight and a low-permeability, fine-grained back-reef may change into a permeable sucrose dolomite (Figure 9).

Several diagenetic processes can increase the heterogeneity markedly. Karstic leaching can be pervasive or restricted to certain zones. Some carbonate reservoirs consist of virtually tight rock in which leached fractures and voids comprise the only effective pore space. Another complicated type, which is very difficult to evaluate, is formed by reservoirs submitted to a mixed process of leaching and dolomitization, forming a network of moldic sucrose dolomite nodules and tunnels through a tight matrix.

For nearly all carbonate reservoirs, detailed microscopic studies are required to unravel their diagenetic history and to identify facies and composing organisms. The rock types have to be classified to be able to analyse and predict the permeability distribution, which, in carbonates, can cover an extremely large range even in a single reservoir (Figure 10).

Small-Scale Heterogeneity

Although the larger scale heterogeneities usually control the production performance, the sedimentary structures and rock texture can cause anisotropy of the reservoir and have a strong influence on residual oil distribution. In cross-bedded reservoirs which have been subject to severe diagenesis, it may be necessary to measure the permeability contrast between the laminae and to take the resulting anisotropy into account in the dynamic fluid flow model. It has also been shown that, even with permeability contrasts between the laminae as low as two, capillary forces can trap of the order of 10% additional oil in comparison with more homogeneous rock. Detailed studies of the capillary pressure curves and relative

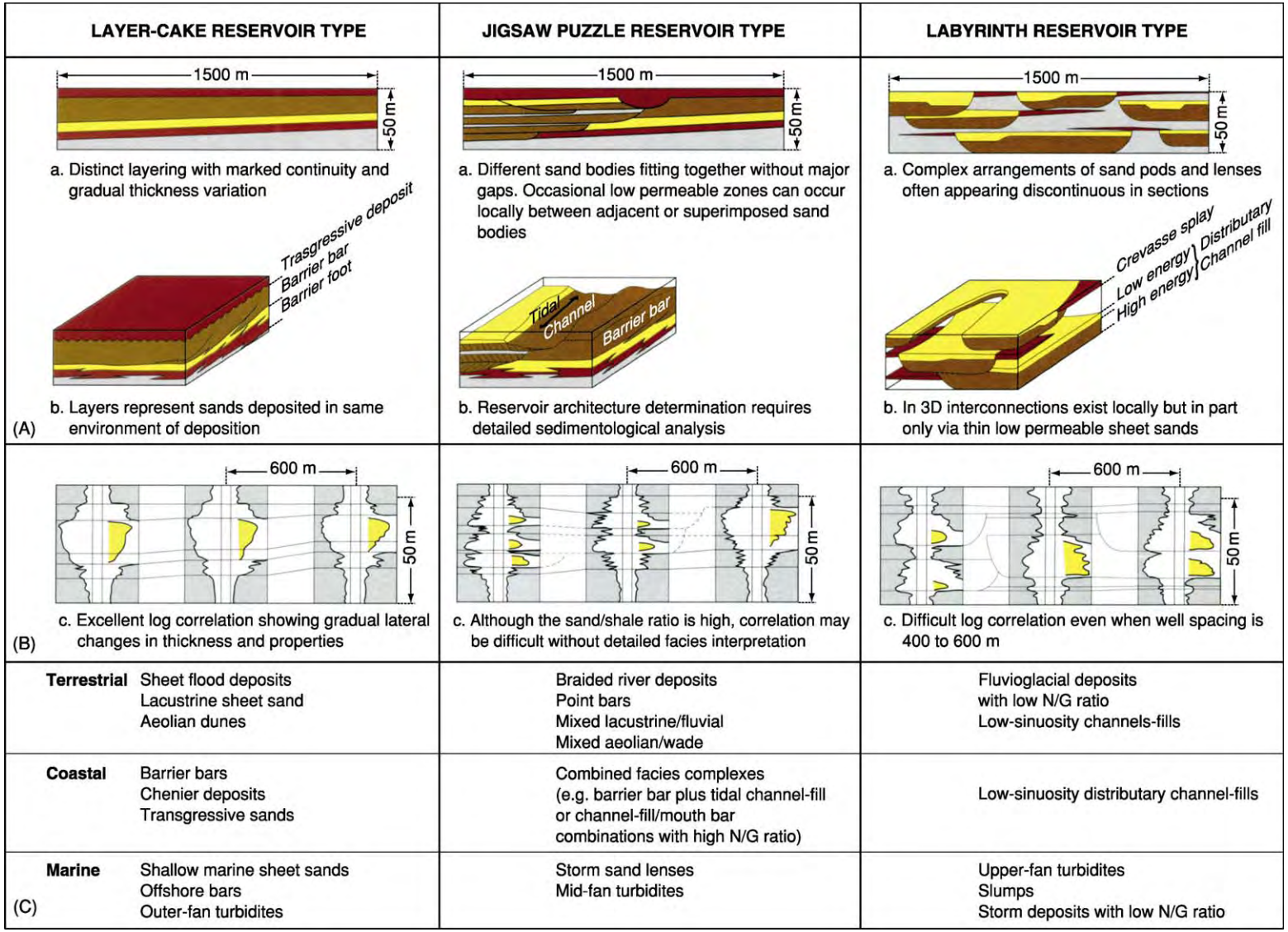


Figure 7 Clastic reservoirs classified into three architectural types: layer cake, jigsaw, and labyrinth. N/G ratio, Net/Gross. Modified with permission from Weber KJ and van Geuns LC (1990) Framework for constructing clastic reservoir simulation models. *Journal of Petroleum Technology* 42(10): 1248 1253, 1296 1297. © Society of Petroleum Engineers.

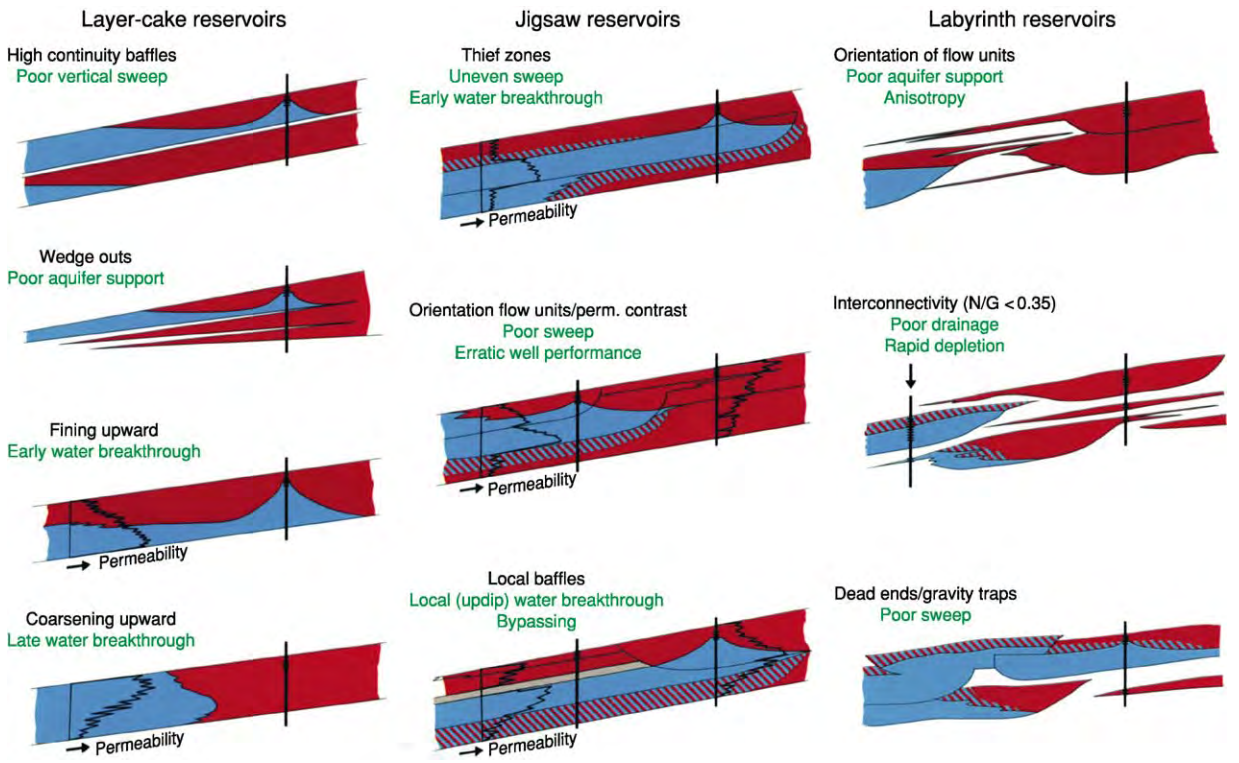


Figure 8 Connected permeable pathways through the major reservoir types. Courtesy of Shell.

permeability to oil and gas for each rock type are necessary. Major lateral differences in apparent oil–water contact can result from variations in rock type.

Techniques

Seismic Methods

The range of methods, techniques, and tools used in production geology has become very large. Seismic techniques are of primary importance in the appraisal stage, but are also used for reservoir monitoring (Figure 11). The traditional use of seismic techniques for structural definition has been much improved by 3D surveys and better processing techniques. Where seismic resolution is good and acoustic impedance contrasts are sufficiently large, amplitude measurements along the beds can reveal depositional patterns as well as oil and gas accumulations (Figure 12).

Besides acoustic impedance mapping, amplitude versus offset (AVO) techniques and shear-wave analysis are now being employed. These techniques are powerful methods to determine lithology, while shear waves are affected by open fractures. Repeat 3D surveys (so-called 4D seismic) can reveal movements of fluid interfaces and saturation changes (Figure 12). Analysis of changes in acoustic impedance and velocity can also be used to estimate pressure changes in the reservoir resulting from hydrocarbon production. This can reveal compartmentalization caused by sealing faults and uneven aquifer drive. Rock stresses can be derived from measurements of shear-wave polarization.

In favourable cases, porosity and pore-fill analysis is carried out via acoustic impedance measurements, calibrated with well data. In this way, volumetric estimates of hydrocarbon-in-place may be made. Of particular importance is to relate the 3D seismic reservoir models to models made from well data correlations with geostatistical methods. Improved calibration of seismic data, attributed using neural network techniques, can provide estimates of facies types and related properties throughout the reservoir. Seismic methods are frequently used for planning well trajectories (Figure 12). Three-dimensional displays and interactive virtual reality techniques are employed for this purpose. The developments in seismic methods for reservoir analysis have been so spectacular in the recent past that further research is likely to result in important advances.

Core Description and Analysis

Facies and rock types have to be identified and characterized from cores and wireline logs. Systematic

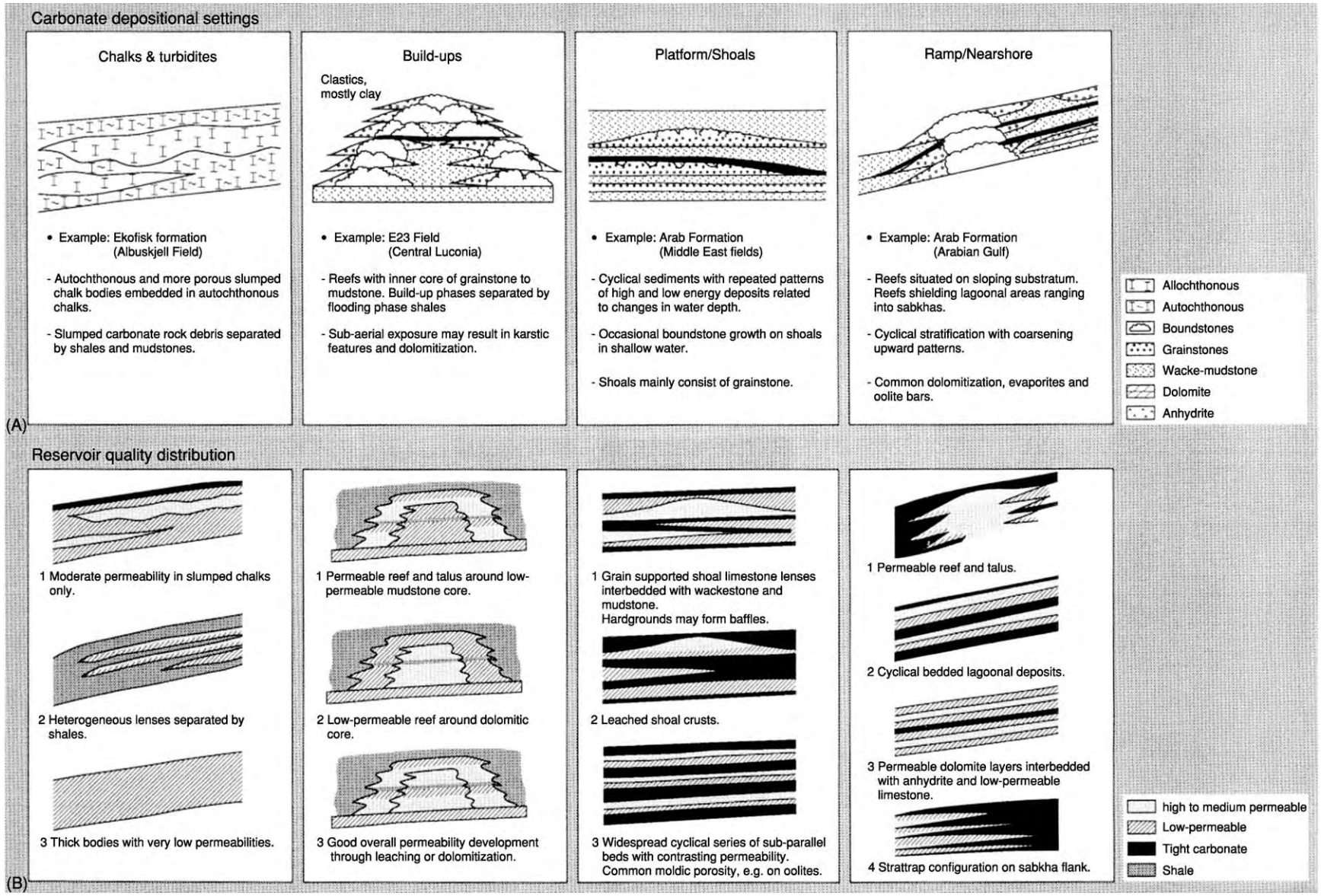


Figure 9 Scheme of the main carbonate depositional settings and common pattern of permeability distribution. Courtesy of Shell.

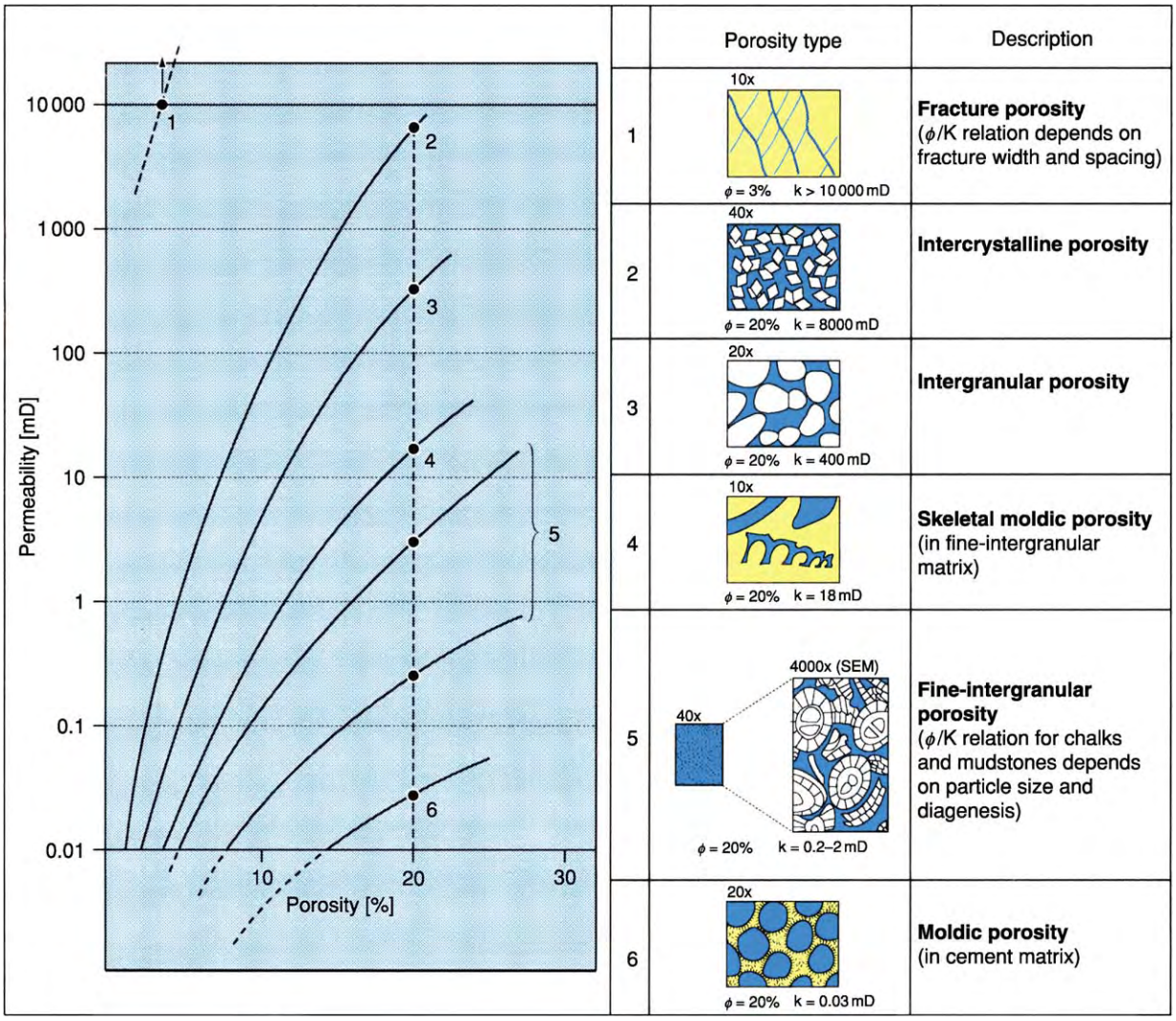


Figure 10 Generalized trends for porosity versus permeability of the main primary and secondary carbonate pore types. Courtesy of Shell.

geologically steered core analysis is the key to the proper understanding of the reservoir composition and the basis for log calibration, static and dynamic modelling, detailed seismic interpretation, and the identification of key heterogeneities. An overview of core analysis methods and applications is given in Figure 13. Core observations under normal light can be augmented by ultraviolet (UV) and infrared photography to reveal remaining oil and sedimentary structures. Standard petrophysical analysis should be carried out with a sample choice and spacing reflecting the reservoir heterogeneity. Probe permeability measurements or cube-shaped samples may be required to analyse permeability anisotropy (Figure 14). The influence of *in situ* stress on reservoir properties and rock compressibility has to be measured.

The core data are used to establish rock types with similar characteristics, such as porosity/permeability relationship, capillary pressure, and relative permeability curves. Each rock type usually represents a specific facies type having a similar lithology and texture. The core data are correlated with the logs to identify typical characteristics that allow the recognition of rock types in uncored wells. This is necessary to delineate facies distribution for well-to-well correlation, as well as to estimate permeability via porosity–permeability transforms. High-resolution borehole imaging logs are very useful for the identification of structural features, sedimentary structures and dips, and trends of individual sand bodies (Figure 15).

A series of electric logs is used to identify rock types in uncored wells. Linear multiple regression of, for example, GR, neutron and density log readings of

Field development stage	Modelling objectives	Seismic attributes	Techniques
Appraisal	– Structure	Reflection time/continuity acoustic velocities	3D survey dip/az plotting shaded relief
	– HC distribution	Acoustic impedance	Mapping impedance, avo
	– Reservoir architecture	Reflection cont./strength	Seismic horizon mapping
	– Porosity distribution	Acoustic IMP/velocities	Acoustic IMP, modelling
	– Volumetric estimates	Acoustic impedance	Finite element impedance distribution determination
	– Fault sealing capacity	Acoustic impedance along fault planes	Plotting impedance slices along faults
Early development	– High precision structure	Detailed acoustic velocity model	Pre-stack migration
	– Fracture distribution	Shear-wave polarization	Measuring shear-wave splitting and attenuation
	– Lithology	Poisson's ratio, velocities acoustic impedance	Shear and compressional wave analysis, avo
	– Facies identification	Acoustic impedance well data	Segmentation, neural network methods
	– Constraining stochastic reservoir models	Impedance amplitude patterns, well data	Geostatistical approach
	– Steering wells	Reflection time/continuity	Wsp (walkaway survey) look-ahead vsp, lwd
Late development	– Movements of HC contacts	Acoustic impedance	Time-lapse seismic, avo
	– Pressure distribution	Acoustic IMP/velocities	Time-lapse seismic
	– EOR steam drive monitoring CO2	Acoustic IMP/velocities poisson's ratio	Cross-well seismic tomography
	– Stress distribution	Shear-wave polarization	Shear-wave polarization analysis, neural networks

Figure 11 Overview of seismic techniques for reservoir delineation, analysis, and monitoring. AVO, amplitude versus offset; EOR, enhanced oil recovery; LWD, logging while drilling; VSP, Vertical Seismic Profile. Modified with permission from Weber KJ (1995) Visions in reservoir management what next? In: *Reservoir Characterisation: Integration of Geology, Geophysics and Reservoir Engineering, The Third JNOC TRC International Symposium, February 2 23, Technical Research Centre, Chiba, Japan*, pp. 1 15. Chiba: Japan National Oil Corporation.

known rock types in cored well is used to set up a diagnostic system for uncored wells. Alternatively, neural networks can be trained to carry out this identification. The advantage of the neural network techniques is that the influence of the input parameters is non-linear.

Permeability Distribution

Permeability measurements are amongst the most difficult to derive from well data and yet comprise one of the most useful and relevant measurements in production geology. The permeability distribution in the well is derived from cores and via rock type identification in uncored wells. However, this only provides local horizontal permeability data around boreholes. Cores and logs in horizontal holes can provide lateral permeability development, but only along narrow zones. There are a large number of techniques to augment the permeability database both horizontally and vertically on a range of scales (Figure 16). Production geologists must play an active role in identifying opportunities to measure the permeability, which provides an important control on fluid flow in the reservoir. Vertical permeability on the medium and large scale is particularly difficult to estimate from cores and logs. For this purpose, multiprobe well testers, pulse testing, production logging, and pressure build-up analysis are used.

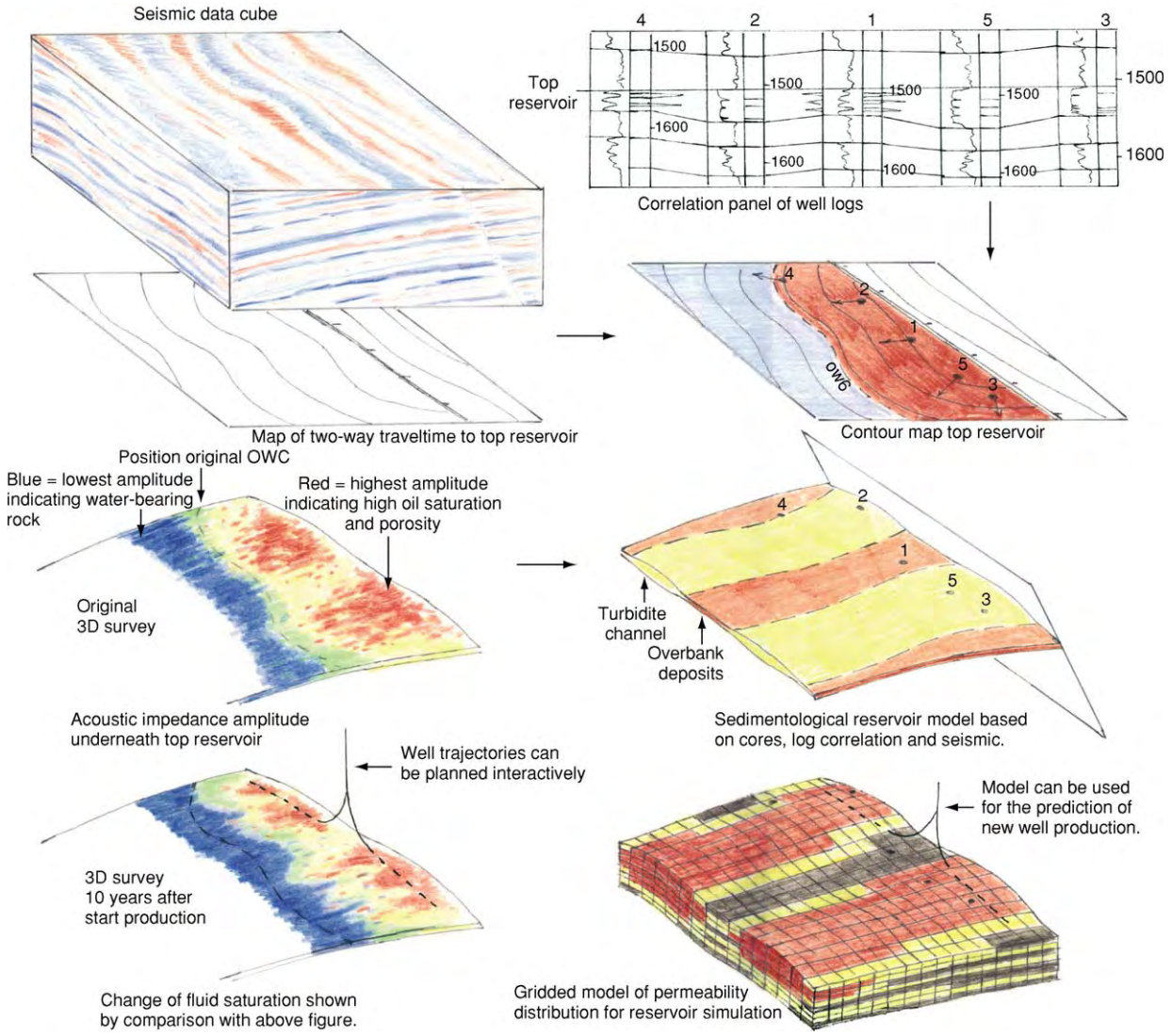


Figure 12 Integrated use of three dimensional (3D) seismic measurements for static modelling purposes.

Fault and Fracture Analysis

Faults are mainly analysed from seismic data, although well-to-well correlation can reveal fault cut-outs. Dip-meter logs can also indicate fault positions by showing drag patterns. The fault types have to be analysed from the deformation pattern seen on seismic measurements and the structural history of the sedimentary basin. The associated sealing capacity of the faults is closely related to the type of fault, fault throw, reservoir lithology, and depth. Cataclastic fault zones tend to be cemented and impermeable after burial to a depth of 1–2 km.

Clay smearing along normal faults is an important process in many deltaic reservoirs. Estimates of the sealing potential can be based on the shale thickness that passed along a fault plane (Figure 17). This can be achieved by a log analysis of wells near the fault and the fault throw derived from the cut-out observed in a well intersecting the fault. If good quality seismic data are available, however, the fault throw distribution can be established. With 3D seismic data, it is possible to sample the acoustic impedance values along both flanks of a fault and to translate these data into synthetic lithological logs. By comparing and correlating the hanging- and footwall stratigraphy, both the 3D distribution of the fault throw and the thickness of the shales that moved across any point of the fault surface can be obtained. The method requires local calibration because the ductility of the shales at the time of faulting can differ markedly.

Fractures are seen on borehole imaging logs. By the use of Stoneley waves (shear waves), open fractures can be identified and the fracture width can be estimated. This provides estimates of the usually very low

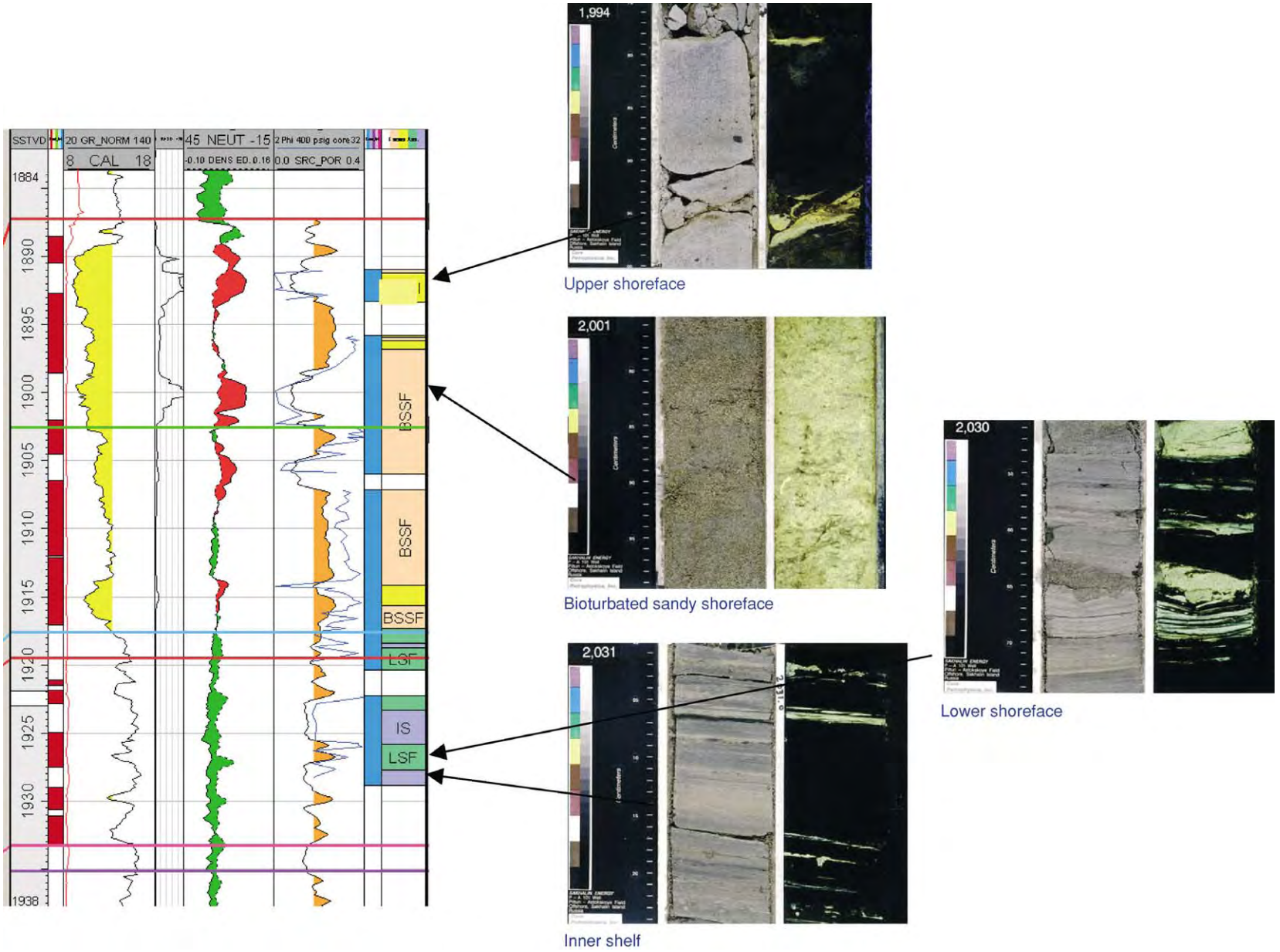


Figure 13 Facies definition from cores and logs. Courtesy of Shell.

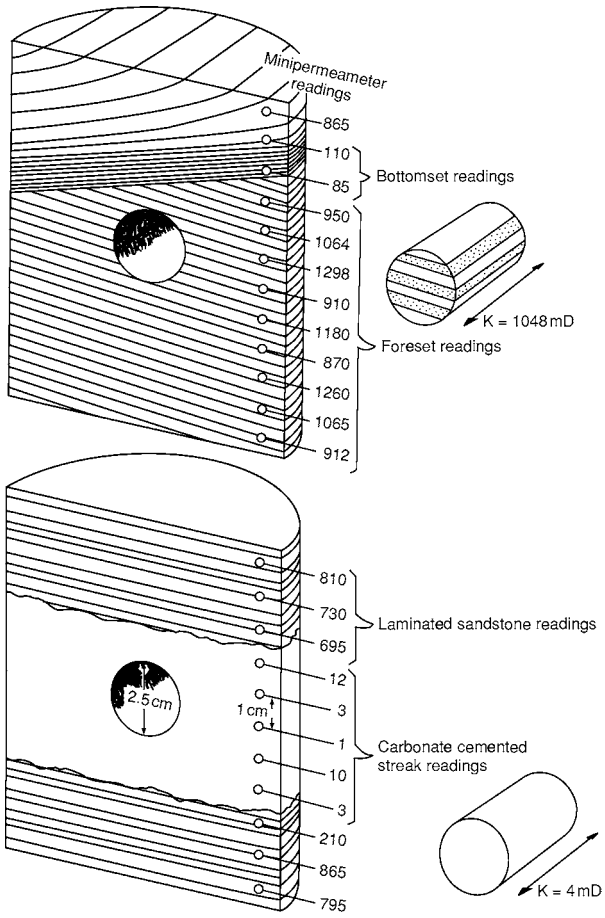


Figure 14 Probe permeability measurements showing small scale heterogeneities. Reproduced with permission from Weber KJ and van Geuns LC (1990) Framework for constructing clastic reservoir simulation models. *Journal of Petroleum Technology* 42(10): 1248 1253, 1296 1297. © Society of Petroleum Engineers.

fracture porosities. Flowmeter-type production logs can also show open fractures. The shear zones around faults can be analysed with borehole imaging logs in horizontal wells. Fracture density, as a result of tectonic deformation, can be estimated by determining the rate of change of the dip. In fault and fracture analysis, extensive outcrop studies have added enormously to the knowledge of fault zone compositions and fracture distributions (Figure 18).

Correlation and Use of Analogues

The correlation from well to well requires a detailed knowledge of the facies distribution and expectation values of the typical shape and dimensional relationships of genetic reservoir bodies. For this purpose, there is a need for a data bank of such information based on outcrop and field data, complete with their sedimentological setting, including type of basin and climate, to be able to choose a proper analogue example. Such data banks have been compiled by several oil companies, while many universities are engaged in reservoir analogue outcrop studies. The stratigraphy forming the major framework of the reservoir can be interpreted with sequence stratigraphical principles. These principles and, especially, the analysis of sea-level fluctuations can be used to delineate the sedimentary successions, and are a great help in correlating the wells (Figure 19). Sea-level changes have a large influence on the sedimentary processes along time slices of the reservoir and, in carbonates, they also have a strong impact on diagenesis (e.g., leaching and dolomitization).

The well-known techniques of comparing the floral and faunal contents of discrete intervals remain very useful, not only for stratigraphical purposes, but also for facies analysis. The palaeontological and palynological studies can be carried out on drill cuttings whilst drilling, and thus can provide important indications of drilling progress through a known stratigraphical sequence. In combination with LWD techniques, the correct casing setting depth can be determined with sufficient safety margins.

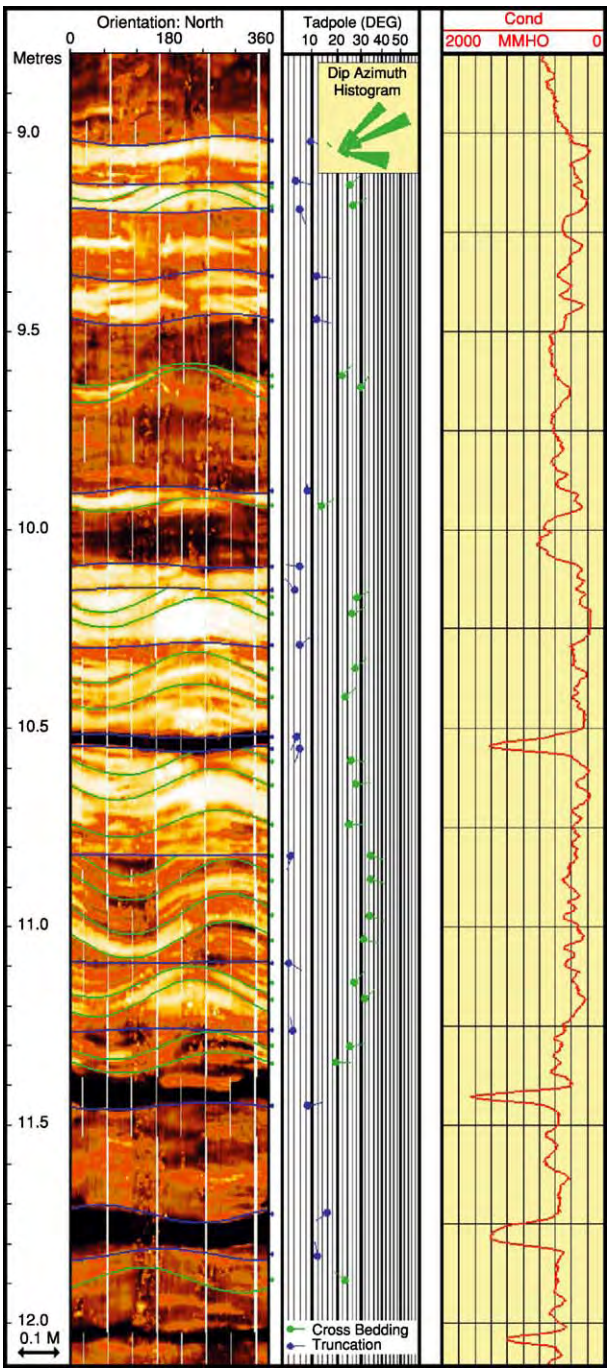


Figure 15 High resolution acoustic borehole images compared with core slab showing aeolian cross bed set. Reproduced with permission from Frikken HW (1999) *Reservoir Geological Aspects of Productivity and Connectivity of Gasfields in the Netherlands*. PhD thesis, TU Delft.

Another useful technique is oil typing via gas chromatography. Detailed comparison of oil samples has proven to be a sensitive indicator of reservoir continuity. Oils that are not in communication either vertically or laterally will show distinctive differences in their chemical fingerprints.

Volumetric Estimates

The static modelling carried out as the basis for volumetric estimates and dynamic modelling requires 3D infilling of the space between the wells. In an early stage, hydrocarbon-in-place estimates are made by

Objective	Scale (m)	Type of measurement
Distribution of horizontal permeability along borehole	0.001–0.01	Probe permeameter on core slabs
	0.01–0.05	Core plugs
	0.05	Grain-size distribution from sidewall samples combined with porosity from logs
	1.2	Nuclear magnetic resonance (NMR) log
	0.5–5	Rock typing in non-cored well and use of porosity/permeability relationship
	1.0–5.0	Pressure build-up from Repeat Formation Tester (RFT) combined with production log
Permeability within well drainage radius	100–500	Well test pressure build-up combined with geological model
	100–500	History matching of well performance after significant production
Vertical permeability of intervals	0.001–0.05	Cube-shaped samples and probe-permeameter
	0.1–0.7	Multi-probe formation tester
	5.0–50	Vertical pulse test
		Computation with geological model using outcrop shale continuity statistics
Lateral and vertical permeability continuity	1.0–50	RFT or multi-probe tester in infill wells
	100–300	Well-test analysis combined with geological model and vertical seismic profile
	200–1000	Horizontal pulse test
Permeability distribution around horizontal wells	0.001–1.0	Core plugs, probe-permeameter, log-derived estimates, production logging
	100–300	Well-test analysis combined with geological model and production logging results

Figure 16 Overview of methods to determine or estimate permeability distribution on a range of scales. Modified with permission from Weber KJ (1995) Visions in reservoir management what next? In: *Reservoir Characterisation: Integration of Geology, Geophysics and Reservoir Engineering, The Third JNOC TRC International Symposium, February 2 23, Technical Research Centre, Chiba, Japan*, pp. 1 15. Chiba: Japan National Oil Corporation.

combining the tentative probability distribution of the controlling parameters: reservoir volume, net-to-gross, porosity, petroleum saturation, and formation volume factor (FVF), the ratio between the net petroleum volume *in situ* and under standard conditions. The cumulative probability distribution curve or expectation curve for the petroleum-in-place is generated with Monte Carlo techniques. The probability distributions of the parameters are randomly sampled and the resulting petroleum-in-place volumes are plotted until the graph hardly changes by additional values (Figure 20).

As an interim step towards 3D modelling, Net-Oil-Sand (NOS) maps can be constructed, which yield better estimates than combining the field averages of all parameters. For thin reservoirs, such maps almost represent 3D models, if the properties of the individual sand bodies can also be plotted (Figure 21).

Geostatistics

Classic statistical methods for the analysis of numerical data sets to determine distributions, standard deviations, and correlations have been used in production geology for a long time. There are, however, a number of problems for which these methods are not suitable. Firstly, one is frequently confronted with sparse data sets which do not allow reliable contouring or correlation. Secondly, there is a need to quantify uncertainty in spatial distributions and also to estimate the possible reduction in uncertainty resulting from drilling an additional well in a given place. Thirdly, one wants to use spatial relationships established for specific types of reservoir body to constrain the models based on well-to-well correlations.

The method which has been developed to handle these problems is called 'kriging', after the originator Daniel Krige, a South African mining engineer. It is based on the reasonable assumption that the unknown spatial distribution of a geological property can be predicted on the basis of the spatial distribution of the measurement of that property. The main tool in this method is the 'variogram', which quantifies the spatial continuity of a property. It is a graph of the variance of the difference of two measurements as a function of their spacing. One can use variograms made on the basis of a data set. If the number of data points is insufficient to define a variogram, one can select a best-fitting variogram or use a variogram that is typical for a modelled property. Known trends can be honoured by using different variograms in different directions.

An example of the above is shown in Figure 22. A map is made of the thickness variation of a reservoir using a best-fitting variogram. Next, the residual

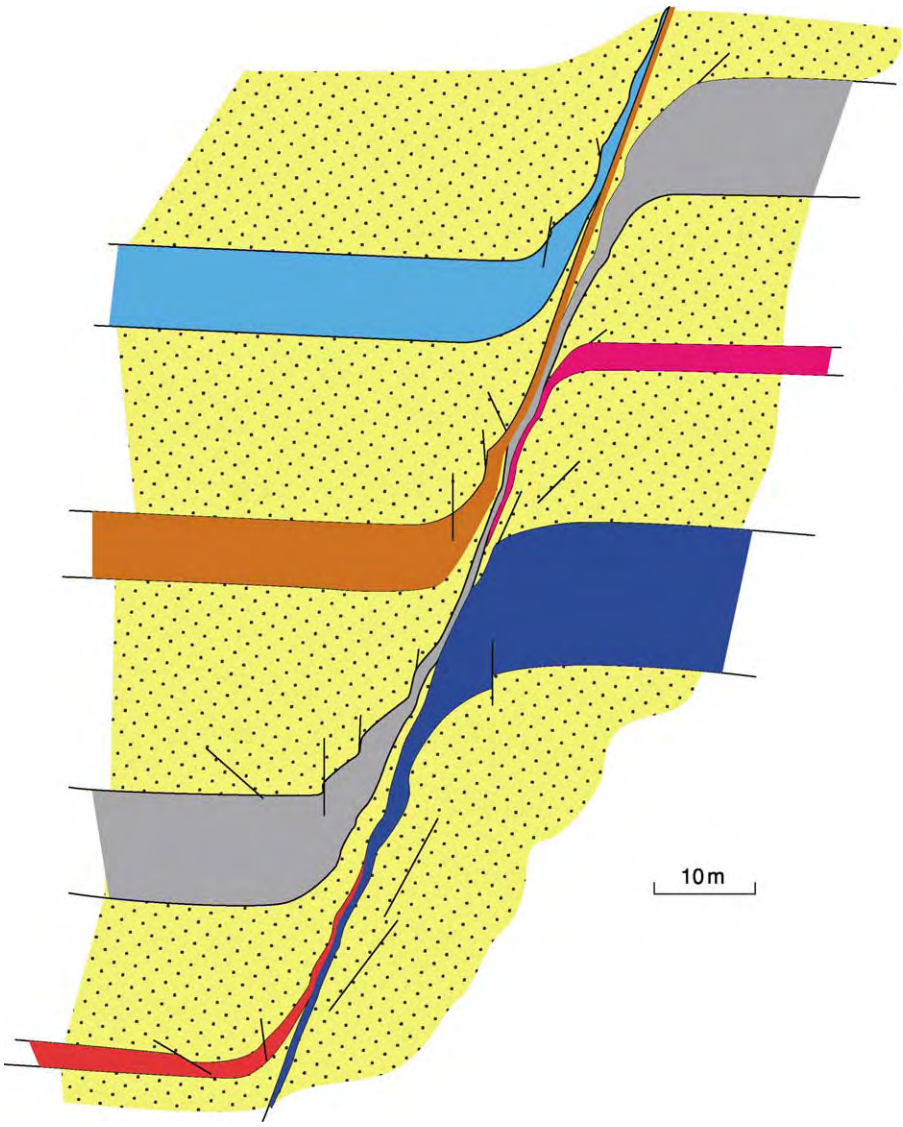


Figure 17 Clay smearing along fault as seen in outcrop. Note discontinuous clay smear of thin (red) shale.

variation between the variogram and the thickness map is computed to show the uncertainty at any point of the map. Finally, one can place a grid over the map and compute the reduction of uncertainty resulting from drilling a fictitious well in any of the grid points. This method can be used to select a next appraisal well.

Further applications of 'kriging' methods include the combination of data from different sources and with different accuracy to estimate a property. In this way, reservoir depth data from wells and from seismic measurements can be merged into one map. Probabilistic reservoir modelling also makes use of variograms to constrain the extent of the bodies within the reservoir. A series of equally probable models can be generated. Here lies one of the most useful applications of analogue data derived from outcrops and densely drilled fields.

Static Modelling

The particular technique used for 3D reservoir modelling depends on the reservoir type. For layer-cake reservoirs, deterministic models can often be made directly from the well-to-well correlations and the rock properties can be interpolated. For labyrinth reservoirs, on the other hand, deterministic modelling is rarely possible without a very close well spacing. Computer systems have been developed to generate a series of equally probable 3D models through 'probabilistic' modelling techniques. Most systems work in three steps. Firstly, correlatable reservoir bodies are determined and incorporated in the

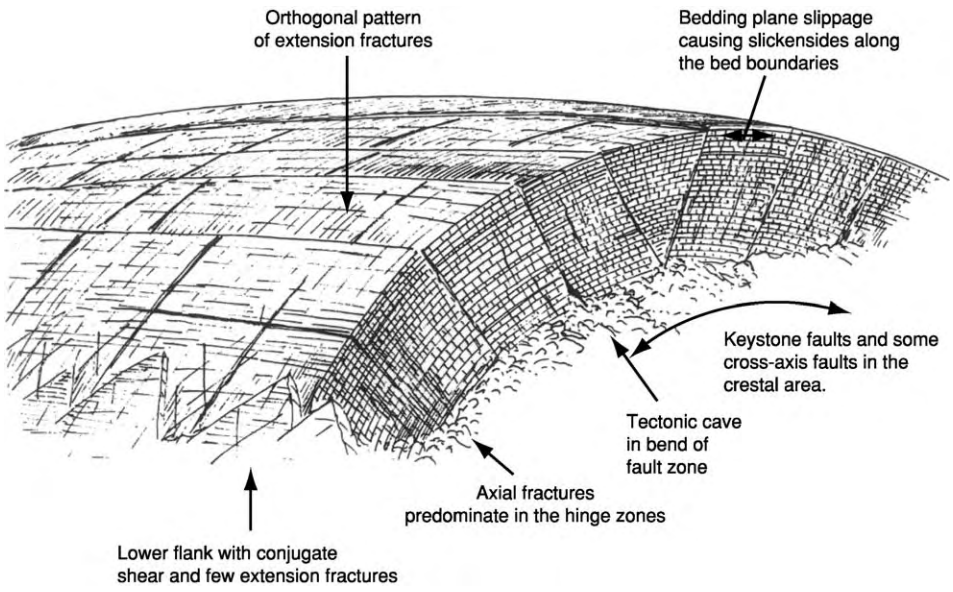


Figure 18 Faults and fractures in anticlinal structure of a thick limestone formation.

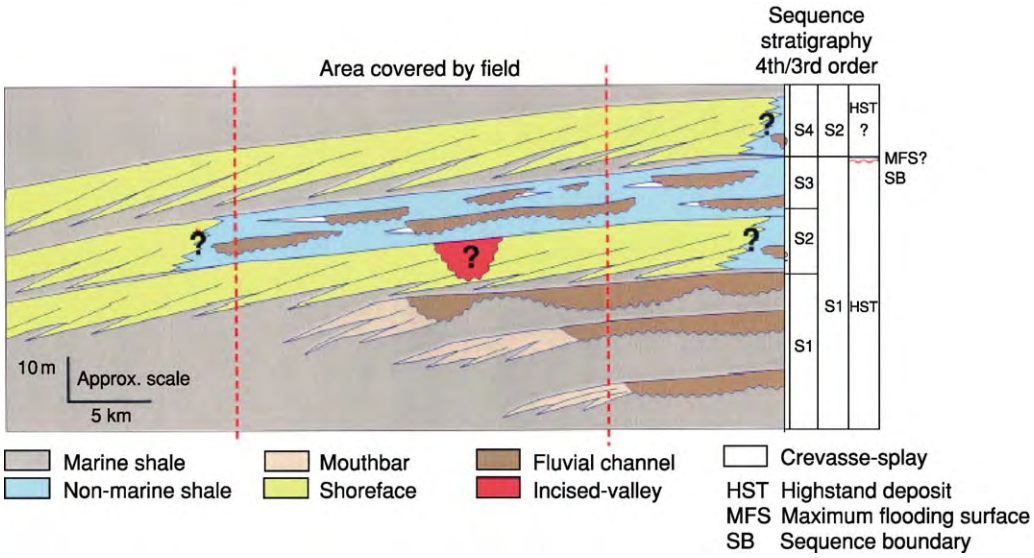


Figure 19 Sequence stratigraphical correlation in reservoir modelling. Courtesy of Shell.

model. Where extrapolations beyond control points are necessary, use is made of an analogue database for the relevant genetic type. Secondly, the uncorrelatable bodies are considered, for which the dimensions are also derived from the database, while body orientations are taken from borehole imaging logs or estimated on the basis of the general geological model. Characteristic variograms for the thickness distribution of genetic sand body types in different orientations relative to their expected trend are also used. Thirdly, especially when the well spacing is rather large, there are smaller or narrower reservoir bodies not penetrated by wells. Such bodies are added using statistical estimates of their occurrence from the wells. Their position and dimensions are conditioned by geological modelling rules and the analogue database (Figure 23). Jigsaw-type reservoirs are also difficult to correlate in the appraisal stage and probabilistic modelling is required. In a later stage, a large part of the architecture may be determined and only limited recourse has to be taken to probabilistic techniques.

The resulting architectural models can be compared with the seismic models and, in favourable

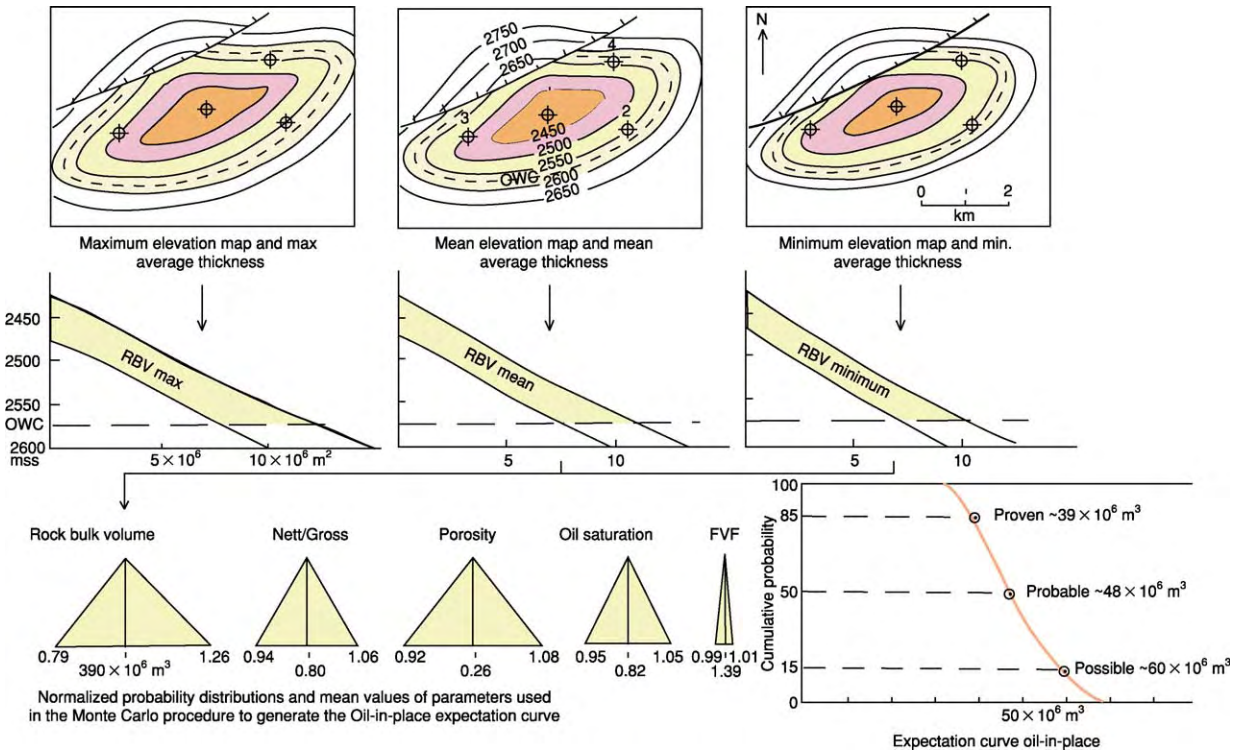


Figure 20 Approach to determine the cumulative probability (expectation) curve for the oil in place volume. RBV, Rock Bulk Volume.

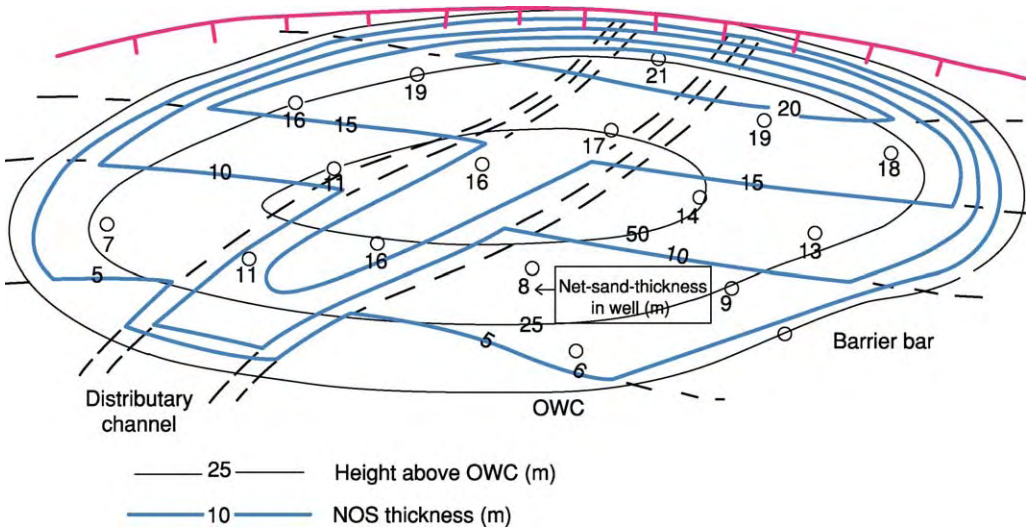


Figure 21 Net Oil Sand (NOS) map of deltaic reservoir.

cases, quantitative correlations can be made. This reduces the spread of possible model configurations. In any case, the permutation of the probabilistic models generated has to be carefully ranked in order of probability by screening them with regard to the presence of geological anomalies. A limited number of 3D models are selected as representing the likely range of variation in the reservoir with respect to hc-volume (hydrocarbon volume), connectivity, architecture, and permeability distribution.

Dynamic Modelling

The profitability of a hydrocarbon reservoir depends crucially on how its development is planned in

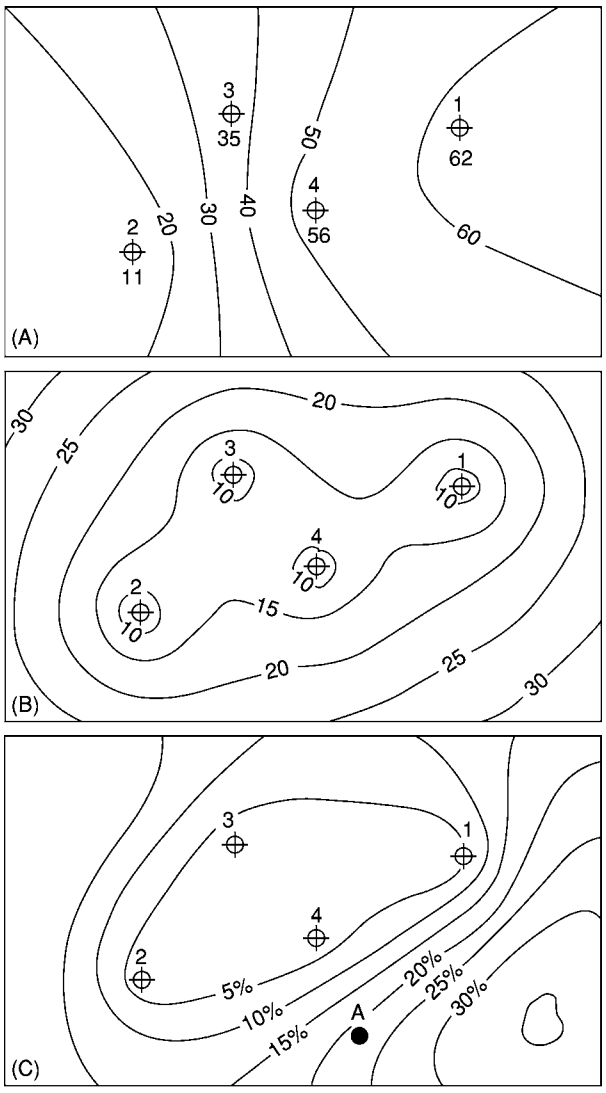


Figure 22 Geostatistical methods in mapping and well planning. (A) Estimated thickness map made with the best fitting variogram. (B) Thickness uncertainty (66% confidence limit) computed from the residual variation between the variogram and thickness map. (C) Expected reduction in estimation variance of the mean thickness after drilling well in location A.

the context of recovery mechanism(s) and infrastructure environment. A key element from the outset of planning is a conceptual model of the hydrocarbon reservoir (ranging from a simple analytical model to a multitude of equiprobable realizations). The reservoir model allows the evaluation of alternatives by providing an estimated range of hydrocarbon volumes that are potentially recoverable under various development schemes.

The majority of the field development plans are based on reservoir simulation models with varying degrees of complexity. Integrated dynamic reservoir models are constructed by reservoir engineers and require input on geological structure and architecture, rock and fluid properties and their distribution, displacement characteristics, and description of the wells and the surface facilities. Static and dynamic modelling programs differ from each other in terms of the number of volume elements they can handle. Like the 3D static modelling packages, the dynamic modelling programs require the reservoir to be segmented into unit cells, each of which has uniform properties, before it can perform computations. Static modelling computations involve the categorization, sorting, and counting of voxel cells; dynamic modelling computations involve the balancing of the fluid masses and pressures within and between 'gridblock' cells. Upscaling programs enable one to aggregate the geological detail (=3D static model) into the bigger elements of a reservoir engineering model (=3D dynamic simulation model) by using a method of permeability averaging (e.g., harmonic, arithmetic, geometric). These programs also produce a corresponding gridblock model of porosity through conventional averaging of voxel porosity values.

It is important that the 3D modeller identifies the key geological characteristics that should not be lost during upscaling (e.g., thin-bedded shales that may act as permeability barriers or baffles) (Figure 24). Not all geological architecture will have equal weighting in terms of its influence on reservoir performance, and therefore the geologist (with the help of a reservoir engineer) must filter through the geological detail in order to determine what is/is not important for reservoir simulation. Iteration between the static and dynamic models is recommended in order to ensure that the relevant detail (flow units and barriers) is not compromised during the upscaling procedure.

Conclusions

The scope of production geological activities and methods has become very wide, requiring an equally diversified group of specialists to cover all topics. Advanced seismic techniques are employed at every stage, from discovery to additional development planning late in the life of a field. The backbone of the profession is formed by expert sedimentologists who can understand and model the architecture of reservoirs, including the features that influence fluid flow. There is clearly a division in clastic and carbonate reservoir expertise because of the difference in sedimentological and diagenetic processes and relevant analysis techniques.

Structural geology is another speciality that should be handled by experts. Again, there are large differences between faults and fractures in clastic and carbonate reservoirs. The advances in seismic resolution

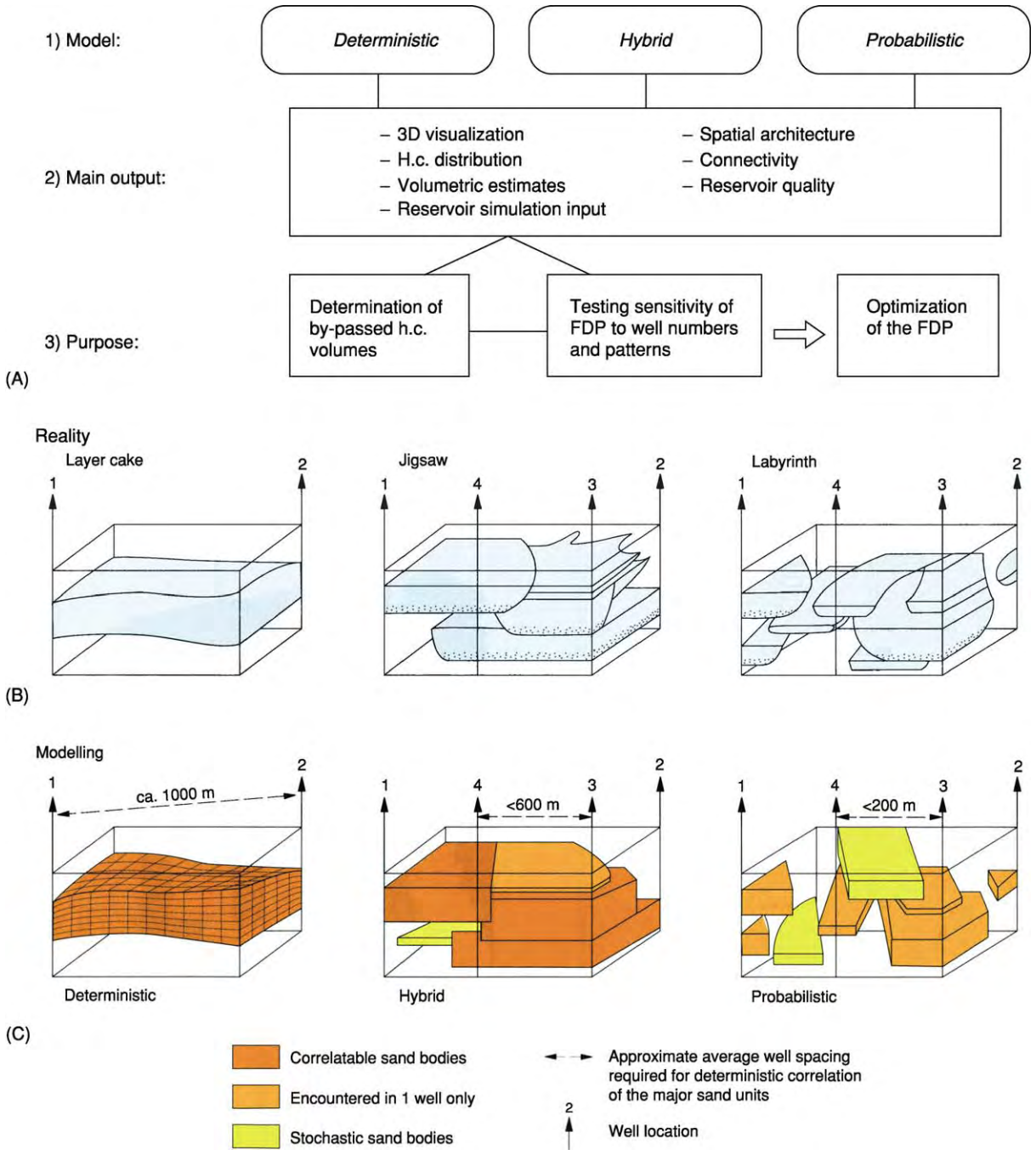


Figure 23 Summary of three dimensional (3D) reservoir modelling. Courtesy of Shell.

on the large scale and borehole imaging logs on the small scale have greatly improved the analysis of faults and fractures. Geostatistical methods are widely used in constructing maps and reservoir models, which reflect the level of uncertainty, but in which analogue information can be incorporated.

The present approach to production geology is heavily dependent on computer systems, and much of the time is spent behind a workstation or personal computer. Modern interactive computer techniques and integrated data systems are used to design models and well trajectories. Visualization plays an important part in this work and virtual reality techniques are already being used. Typical of the modern reservoir management approach is the integrated team composed of various production geology specialists, together with petrophysicists, reservoir engineers, and production technologists.

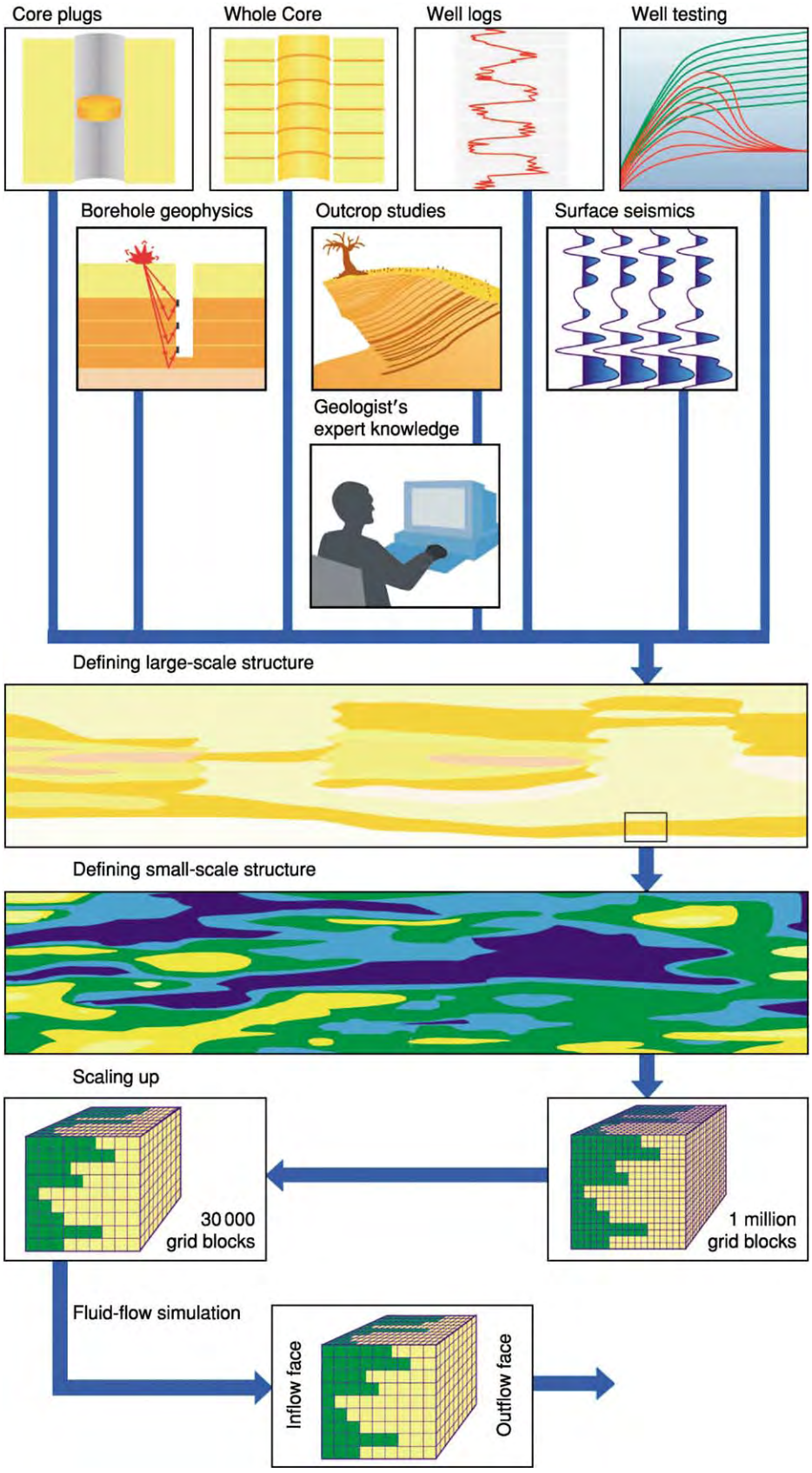


Figure 24 Building a reservoir model with measured data, expert knowledge, and statistics. © Schlumberger Limited, used with permission.

The future will undoubtedly bring important further improvements in seismic techniques. There is a need for more analogue data and sedimentological knowledge to aid in reservoir modelling. The reservoir models will benefit more and more from integration with seismic information, both with respect to reservoir architecture and to fluid content and movements. The high costs of deep-water field appraisal and development will place a premium on the timely construction of realistic reservoir models. On land, there is a large scope for enhanced oil recovery that equally requires reliable detailed reservoir models. Thus, it may be expected that there will be a significant further development in production geological techniques.

See Also

Diagenesis, Overview. **Petroleum Geology:** Overview; Chemical and Physical Properties; Exploration; Reserves. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Dolomites; Limestones; Sandstones, Diagenesis and Porosity Evolution. **Seismic Surveys**.

Further Reading

American Association of Petroleum Geologists (AAPG) (1993) *Development Geology Reference Manual. AAPG Methods Series No. 10*. Tulsa, OK: American Association of Petroleum Geologists.

Barwis JH, McPherson JG, and Studlick JRJ (eds.) (1990) *Sandstone Petroleum Reservoirs*. New York: Springer Verlag.

Bishop CM (1995) *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Brown AR (1999) *Interpretation of Three Dimensional Seismic Data. AAPG Memoir 42*. Tulsa, OK: American Association of Petroleum Geologists.

Dickey PA (1986) *Petroleum Development Geology*, 3rd edn. Tulsa, OK: PennWell Publishing Company.

Dikkers AJ (1985) *Geology in Petroleum Production. Developments in Petroleum Science 20*. Amsterdam New York: Elsevier Science.

Flint SS and Bryant ID (eds.) (1993) *The Geological Modelling of Hydrocarbon Reservoirs and Outcrop Analogues. Special Publication No. 15 of the IAS (International Association of Sedimentologists)*. Oxford: Blackwell Scientific Publications.

Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Jahn F, Graham M, and Cook M (1998) *Hydrocarbon Exploration and Production. Developments in Petroleum Science 46*. Amsterdam New York: Elsevier Science.

Jensen JL, Lake LW, Corbett PWM, and Goggin DJ (1997) *Statistics for Petroleum Engineers and Geoscientists*. Upper Saddle River, NJ: Prentice Hall.

Laudon R (1996) *Principles of Petroleum Development Geology. Petroleum Engineering Series*. Upper Saddle River, NJ: Prentice Hall.

Lowell JD (1985) *Structural Styles in Petroleum Exploration*. Tulsa, OK: OGCI Publications, Oil and Gas Consultants Inc.

Moller Pedersen P and Koestler AG (eds.) (1997) *Hydrocarbon Seals Importance for Exploration and Production. Norsk Petroleum Forening/NPF Special Publication No. 7*. Singapore: Elsevier Science.

Nelson RA (1985) *Geological Analysis of Naturally Fractured Reservoirs*. Houston, TX: Gulf Publishing Company.

Reading HG (ed.) (1996) *Sedimentary Environments and Facies*, 3rd edn. Oxford: Blackwell Scientific Publications.

Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) *Siliciclastic Sequence Stratigraphy in Well Logs, Core and Outcrops for High Resolution Correlation of Time and Facies. AAPG Methods in Exploration Series No. 7*. Tulsa, OK: American Association of Petroleum Geologists.

Weber KJ (1995) Visions in reservoir management what next? In: *Reservoir Characterisation: Integration of Geology, Geophysics and Reservoir Engineering, The Third JNOC TRC International Symposium, February 2 23, Technical Research Centre, Chiba, Japan*, pp. 1 15. Chiba: Japan National Oil Corporation.

Weber KJ and van Geuns LC (1990) Framework for constructing clastic reservoir simulation models. *Journal of Petroleum Technology* 42(10): 1248 1297.

Reserves

R Arnott, Oxford Institute for Energy Studies, Oxford, UK

Introduction

Fossil fuels can be broadly categorized as either resources or reserves. Resources include all fuels, both those identified and those as yet unknown. Reserves are that portion of the identified resources which can be economically extracted and exploited using current technology. Petroleum reserves can be labelled under a wide variety of physical, chemical, and geological circumstances. For example, the boundaries between crude oil as a liquid and condensates have long been the subject of controversy. In addition, there are issues of definition as to what to include or exclude from a particular production forecast, as there are as to what can and cannot be reported as reserves because of legal and political considerations. The only near-certainty on the supply side is the actual volume of oil that has been produced. This is because the definitions of petroleum reserves often include assumptions with regard to existing technology and present economic conditions. However, there is no uniformity or stated policy as to the time period over which the existing technology and present economic conditions are anticipated to prevail. As a consequence, there is often fierce debate about how long existing petroleum reserves are likely to last and the economic consequences if future production cannot meet demand.

Definitions of Reserves

Because of the considerable uncertainty surrounding the definitions of reserves, most authors prefer a probabilistic rather than a deterministic approach (Figure 1). An initial declaration of recoverable oil that can, with reasonable certainty, be recovered in the future under existing economic and operating conditions is the usual definition of the 'proven' reserves. However, all fields will also be declared as having additional volumes of 'probable' and 'possible' reserves. The definition of 'proven', 'probable', and 'possible' reserves varies from country to country. 'Proven' reserves are usually defined as being P90 reserves, indicating that there is a greater than 90% chance that the actual proven reserves base will be higher, and a 10% chance that it will be less. Similarly, 'probable' and 'possible' reserves can be defined as P50 and P10 reserves, respectively. The important point here is that, when oil reserve numbers are quoted in the literature, it must be realized that the numbers are probabilistic and, in the case of proven reserves, are more than likely to be exceeded. Therefore, it is wrong for commentators to argue that the volumes of discovered reserves are a fixed entity as, in the case of proven reserves, there will be a 90% chance that the initial reserve number will be exceeded.

Declarations of proven reserves are often only for a specific reservoir. Therefore, when such information is summed, eventually, to the regional and national level, the simple arithmetical addition of a large number of independent values, each representing the 90% probability of a specified volume in a specific reservoir, produces a higher joint probability of the total. It is for this reason that field growth and rates of field growth are well documented in practice.

Resources can be defined as existing reserves plus all of the accumulations of reserves that may eventually become available. These additional reserves might have already been discovered but be uneconomic, or may not yet have been discovered. The distribution of fossil fuels can therefore be viewed as a pyramid, with a small amount of higher quality resource at the top, but with increasingly large amounts of lower grade resource as we move down the pyramid (Figure 2). The costs of retrieving the resource increase lower down in the pyramid, making a larger amount of the resource available at higher prices. The issue as to what defines the total volume of petroleum resources depends on where the pyramid is sliced, and this is a very subjective decision.

There is often a curious circularity in the estimates of undiscovered and ultimate reserves. Undiscovered reserves are the difference between estimates of the discovered and ultimate reserves. However, ultimate resource estimates obviously depend on an estimate of undiscovered reserves. Although there are a number of ways in which to try and surmount this difficulty, the most commonly described method involves an examination of the discovery pattern. In general, there is some regularity to this pattern, with the largest fields being discovered first with attention then switching to the smaller fields. Therefore, in any one basin, the curve (commonly referred to as the creaming curve) relating cumulative discoveries to wildcat wells is usually hyperbolic with the asymptote revealing the level of ultimate reserves. However, this method is not immune to error, especially if larger

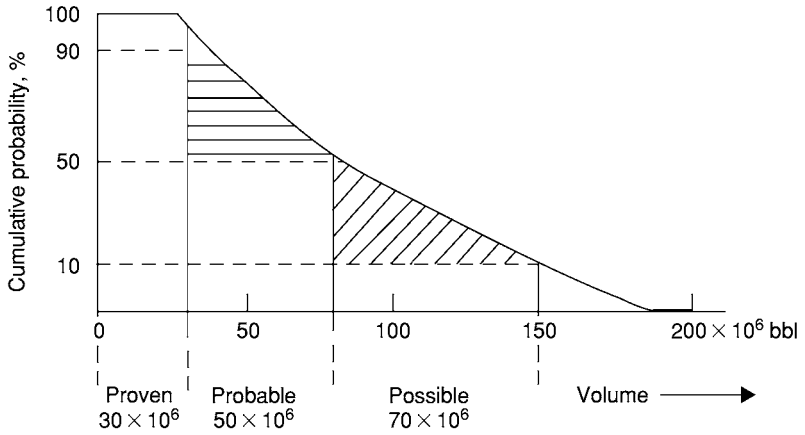


Figure 1 Expectation curve used to find approximate values of proven, probable, and possible reserves. bbl, barrels.

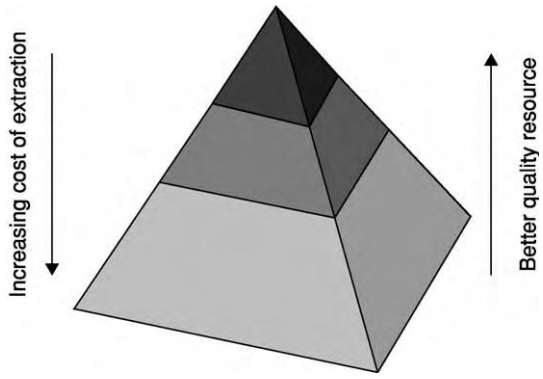


Figure 2 Resources visualized as a pyramid, with a small volume of prime resources that are of high quality and easy to extract, and a large volume of resources of lesser quality that are more difficult or expensive to extract.

fields are discovered late in the life of a particular basin. In the case of the UK sector of the North Sea (Figure 3), the asymptote has moved over time as the older fields have grown in size through better recovery techniques and reservoir definition. As a consequence, estimates of the total resource base have almost doubled in size from the initial values.

Predictions of Ultimate Recoverable Reserves

Numerous authors have attempted to predict the magnitude of the ultimate recoverable resource base. What is interesting about these estimates is that they all essentially identify those crude oil reserves that will be produced over the next 40–50 years. In other words, every assessment tends to slice lower through the resource pyramid (Figure 2). Estimates of the ultimate resource vary significantly depending upon the relative optimism or pessimism of the subsurface geotechnical assessment, as well as the definition of what is the minimum size of economic importance.

Recent estimates of the ultimate recoverable reserves of oil (Figure 4) vary between 1.5 and 3.8 trillion barrels. It is interesting to note that, despite the range of estimates, the mean estimate has remained stable at 2 trillion barrels for at least the past 20 years, suggesting that technology has reached an asymptote. Of course, not all of the estimates are based on the same set of data, as different estimates slice the resource pyramid at different levels. The most pessimistic estimate only includes conventional oil reserves and excludes all other oil types (for example, heavy oil, tar sands, secondary recovery). In contrast, the optimistic estimate by the United States Geological Survey (USGS) only excludes from its estimates oil from tar sands, oil from polar regions, and oil shales. However, nearly all estimates rely on existing technology and do not attempt to predict the impact of future technological developments. As a result, the estimates can only be regarded as static estimates at a fixed point in time under a particular set of technological and economic circumstances.

The actual volume recovered will depend on a number of factors, not least whether the reserves are economic to extract. Political, economic, and technological constraints all play a role in deferring or accelerating production. In addition, there is interplay with other sorts of fuel, which is likely to have a significant impact on determining the rate of extraction.

Historical evidence from producing fields highlights the fact that reserve estimates are dynamic as circumstances change and knowledge and technology progress. An understanding of the potential importance of the growth of reserves can be gained from considering the current estimate of the global oil resource base and the current average recovery factor. Using consensus estimates that the global average

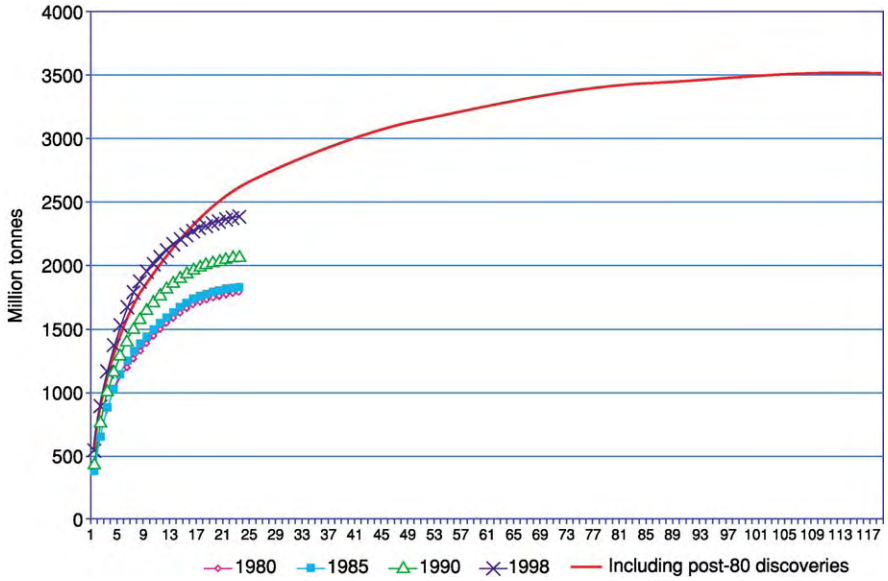


Figure 3 Creaming curves for the UK sector of the North Sea showing the impact of reserves creep on the older fields. The upper curve includes all fields that had been discovered and were producing by 1998. The symbols represent estimates of the same 25 fields made at different times (1980, 1985, 1990, and 1998). The figure shows the change in estimated size of these discoveries. The full line shows the addition of a further 100 discoveries made up to 1998 and which are currently in production.

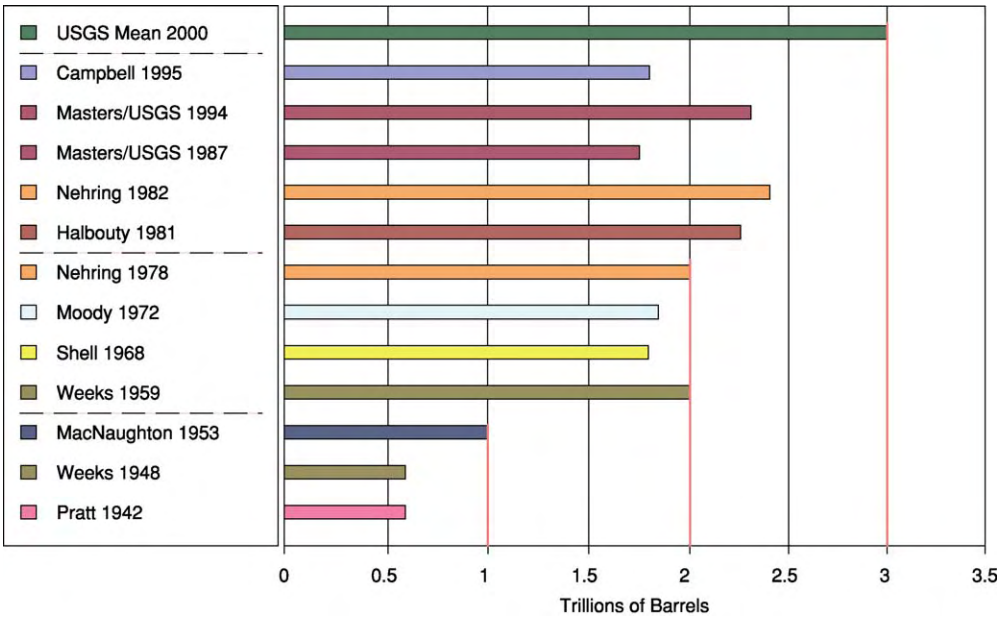


Figure 4 Published estimates of world ultimate oil recovery (trillions of barrels). This chart shows the various published estimates of ultimate oil recovery through time. Earliest estimates made in the 1940s suggested an ultimate recovery of around 0.5 trillion barrels. However, the average estimate since Halbouty (1981) has been around 2.0 trillion barrels. USGS, United States Geological Survey.

oil recovery factor is around 35% (Figure 5), and that the current resource base of reserves is around 2 trillion barrels, original oil in place would be nearly 6 trillion barrels. Therefore, every 1% increase in global oil recovery would lead to an additional 55 billion barrels of oil being produced. Moving the average global recovery factor up to 45%, a not unrealistic target, could increase global recoverable reserves by around 550 billion barrels.

Over the past 50 years, estimates of global recoverable oil reserves have risen annually, except for two small falls in the late 1970s, despite annual increases in oil production (Figure 6). This has led to a great deal of attention being paid to reserve reporting by

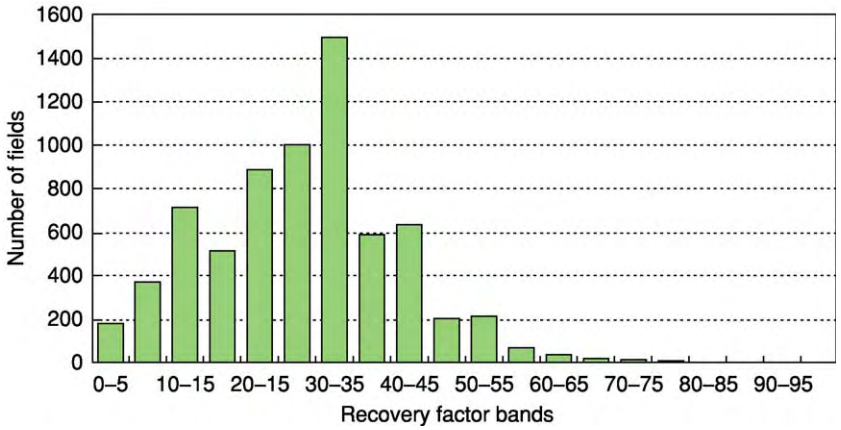


Figure 5 Distribution of recovery factors in all currently producing oil fields.

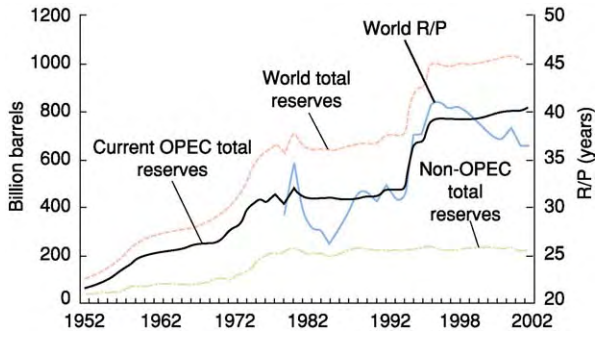


Figure 6 Global estimated reserves and reserves to production (R/P) ratios by year. OPEC, Organization of Petroleum Exporting Countries. R/P, reserves to production ratio.

companies and countries. Individual Organization of Petroleum Exporting Countries (OPEC) members have been accused of inflating reports of reserves because export quotas were originally based on estimates of recoverable reserves. The huge reserve revisions reported by many OPEC companies in 1987 have been used by some to undermine confidence in published numbers. In reality, it is possible that previous OPEC estimates were too conservative, having been inherited from private companies before nationalization.

Although the absolute level of reserves has risen only slightly over the past 5 years, within that total, upward revisions to existing fields and discoveries have exceeded 150 billion barrels, 30% more than actual consumption. The USGS recently published a mean estimate of 612 billion barrels from increased recovery, significantly increasing their estimate of the world's ultimate recoverable reserves. In the UK sector of the North Sea, reserves have grown by an average of 35% from the time of field approval (Annex B) to the present day or abandonment (Figure 7).

Estimates of the total resource base remaining depend not just on new discoveries being made and brought on stream, but also on the amount of reserves used due to depletion of existing fields. Over 70% of current world oil supply comes from oil fields that were discovered prior to 1970, and 14 of these fields produce over 20% of the world's total supply. Even though data for many of these fields are very hard to come by, the decline rates for these fields are likely to become of increasing importance in the future. However, the effect of depletion is often overstated as forecasters often assume a depletion rate of 10–20%, and no further investment. As an example, estimates of non-OPEC production have always tended to indicate that production is at or near its peak, and yet overestimation of depletion rates has led to successive estimates being proven to be incorrect (Figure 8).

The Peak Oil and Depletion Debate

Ever since oil was first produced in significant quantities over 140 years ago, debate has taken place on whether petroleum will run out within the foreseeable future. Historically, near-term supply concerns arise when the relative rate of production capacity growth falls short of expected rates of demand growth. Through time, however, changing attitudes to oil supply and demand have led to varying perceptions of whether the resource is in short supply or not. It is also noticeable that changes in perception can take place over a very short period of time depending on a particular event on either the supply side or the demand side. The discovery of super giant oil fields in Texas in the 1930s gave rise to perceptions of a huge glut in oil. This perception was sustained by the discovery of major reserves in the Middle East. However, by the early 1970s, there was a clear paradigm

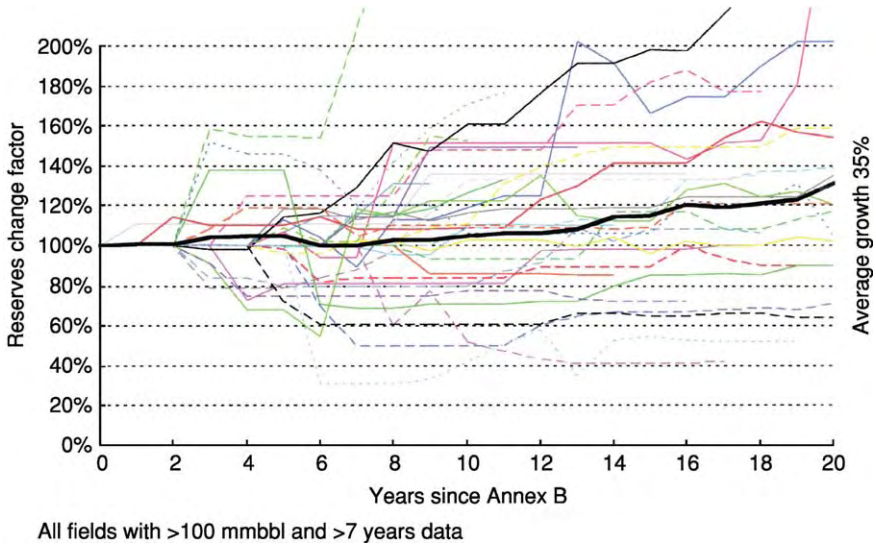


Figure 7 Growth of reserves in UK oil fields. This chart shows the change in reserve estimates of UK oil fields with time, since their 'Annex B' approval. Some field estimates have decreased, but most have increased very substantially. The average of all the fields (full line) has grown by 35% over 20 years.

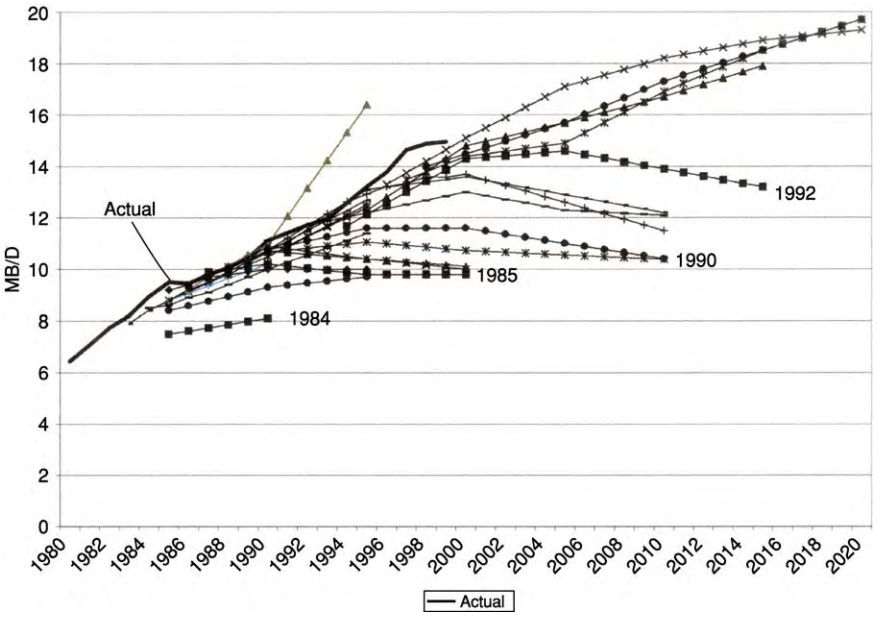


Figure 8 Forecast of non Organization of Petroleum Exporting Countries (OPEC) Third World oil production over time made by the United States Department of Energy. The chart shows that estimates of non OPEC production have consistently proved to be conservative through time. For example, the forecast made in 1990 for the year 2000 is now some 4 million barrels per day lower than actual production. MB/D, million barrels per day.

shift as part of a trend towards pessimism about resource availability as studies began to emerge that oil production outside the Middle East had peaked. This was triggered when US production levels started to fall from peak levels by the end of the 1980s. In the 1990s, the introduction of new development techniques to boost recovery rates, coupled with the revolution in computers which transformed seismic acquisition and processing, the technical innovations that enabled deep-water development, and the opening up of new petroleum provinces that were previously closed due to political constraints, all led to a period of optimism that weakened concerns about the exhaustibility of reserves. More recently, these

concerns have been raised again as the oil industry worldwide struggles to replace its reserves and achieve targeted rates of production growth.

Much of the published literature on the depletion debate falls into two camps. On the one hand, there are those who warn of the imminent danger of a collapse in oil supply and the economic consequences of that occurring. On the other hand, there are those who argue to varying degrees that there is no limit to oil supply in the near future. What is common to both camps is the extensive reference to the work of King Hubbert.

Hubbert gained prominence in the American geological community because he anticipated the peak and subsequent fall of US oil and gas production. At the heart of his prediction was the assumption that all resources are finite, and that eventually all resources will be depleted and exhausted over a period of time that is determined by the rate of production.

Hubbert argued that the complete cycle of exploitation must have the following characteristics. Beginning at zero, the rate of production tends initially to increase exponentially. Then, as difficulties of discovery and extraction increase, the production rate slows in its growth, passes one or more maxima and, as the resource is progressively depleted, declines eventually back to zero. He argued that this cycle of production would be bell shaped when plotted against time. The approximate rule here is that peak production will correspond to the mid-point of depletion and that this usually occurs about 20–25 years after the discovery of the mid-point. Although some production curves do have a shape that approximates such a 'normal' distribution, there is no inherent reason why production should follow such a pattern, and some production curves show very strong asymmetry. For example, when production is dominated by a small number of large fields, peak output tends to precede the depletion mid-point. Where the discovery pattern is more dispersed, or where offshore fields are significant, the production peak usually comes after the mid-point.

Hubbert's basic assumption was that, if known past and prospective rates of production are combined with a reasonable estimate of the amount of fuel initially present, one can calculate the probable length of time that the fuel can be exploited. In other words, the area under the complete production curve would be equal to the size of the resource. However, estimating the amount of oil and gas that will ultimately be discovered and produced in a given area is full of uncertainty, as we have seen. The Hubbert curve might have fitted US production, but, on a global scale, the modelled curve no longer fits the historical data (**Figure 9**). Up until the mid-1970s, the modelled production matched the actual production, but, after the supply shocks, which led to sharp oil price rises, annual increases in demand for oil fell from around 7% to just 2% per annum.

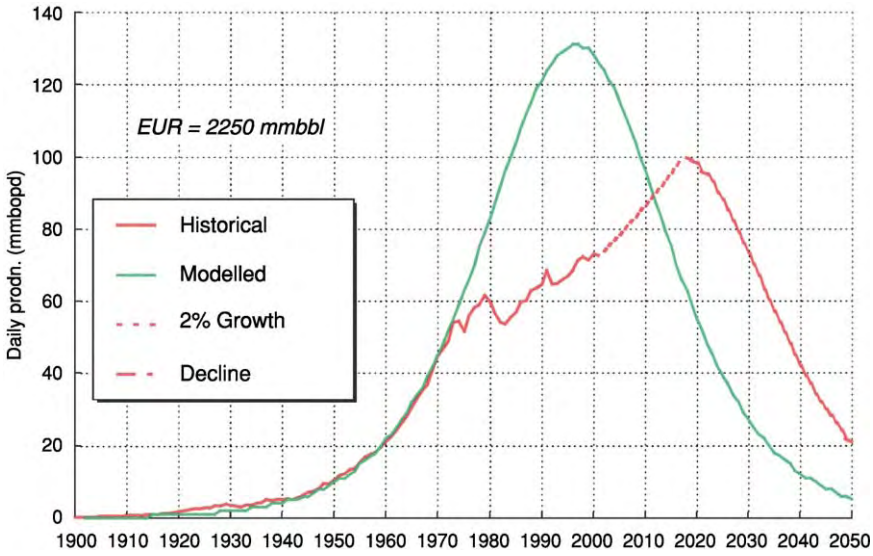


Figure 9 World production and the Hubbert peak. This chart shows a theoretical Hubbert curve based on production up to the oil price shocks of the 1970s. The resulting high oil prices led to a decrease in demand for oil. As a result, the slope of the oil supply curve departs from the predicted Hubbert curve after 1973. Assuming oil demand growth of 2% from 2003, and an ultimate reserve base of 2.25 trillion barrels, the peak in oil production could be reached in 2018. The chart highlights the impact that changes in demand can have on predictions of when global oil production will peak. mmbbl, billion barrels; mmbbpd, million barrels per day.

Models of Resource Depletion

There are, in essence, two main models that have been applied to the estimations of the volume of remaining reserves and the rate of depletion. On the one hand, there is the model for an open market and, on the other, the model for a closed market. The closed market model, in essence, treats the global oil reserves as a finite or 'fixed stock', whereby market depletion would ultimately result in higher prices. In the open market model, the cost of new energy sources is forecast to fall because of technological innovation, economies of scale, and the development of efficient systems of transportation. With lower costs, new energy sources become competitive with existing energy sources and ultimately lead to their replacement. The open market model could ultimately lead to lower energy costs through time through the potential substitution of fuels, in particular natural gas.

The open market model has stronger affinities to actual market examples. The US energy market is an open market, which has benefited from a long-term reduction in energy prices despite rising demand. Coal substituted for firewood; petroleum substituted for whale oil for illumination purposes, and eventually coal for transportation; gas substituted for coal in domestic heating; and gas and coal are currently competing in the electrical generation sector (Figure 10).

The End of the Petroleum Era?

With any finite resource, there will come a time when unfavourable economics will not permit further extraction. Worries concerning the long-term supply of petroleum began shortly after production started in 1859. The huge growth in energy markets after World War II also led to renewed fears that fossil fuels would be exhausted. In the 1970s, many people argued that the growth in world population and industrial production would lead to the total depletion of oil within 50 years of the forecast. The fact that the deadlines for these scenarios have now passed, and oil production and reserves continue to rise, shows that these forecasts were all overly pessimistic.

The resource exhaustion spectre was initially raised largely on the basis of a comparison of the current year's oil production with the so-called proven reserves of oil. However, this type of evaluation is inadequate as proven reserves represent nothing but the working stock of the oil industry. For example, the 'optimum' reserves life of 10 years is merely an expression of confidence from private and state oil companies as to where oil will come from in the future in order to fund investment in new infrastructure. Therefore, the current reserves life of oil is deemed to be irrelevant, especially in the context of a global reserves life that has risen to around 40 years from 1940 to 1975 despite the significant increase in global production.

The debate has now moved on from reserves life to a focus on the point in time when global production starts to decline. Supporters of this 'peak oil' school of thought forecast serious economic and social consequences in the aftermath of this occurring.

They highlight the fact that most of the world's major oil discoveries were made during the 1960s and 1970s and that, since that time, the discovery rate has declined (Figure 11). Using examples from various basins around the world, not least the USA, it is argued that peak oil production usually occurs around 25–30 years after peak exploration success. Using this as a proxy, it is predicted that oil production will peak within the next 10 years and, given that there are no near-term substitutes to oil, an economic crisis will ensue.

The Economic Viewpoint

One key assumption made by the supporters of 'peak oil' is that the resource base is finite. However, the amount of oil that can be recovered depends not only

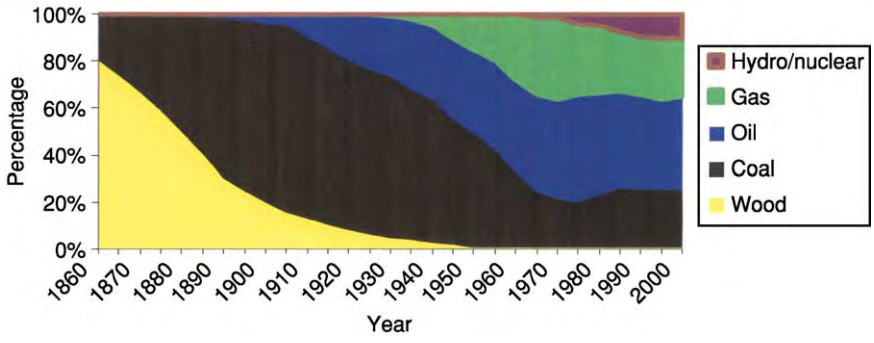


Figure 10 United States primary energy consumption by fuel, 1860 2000.

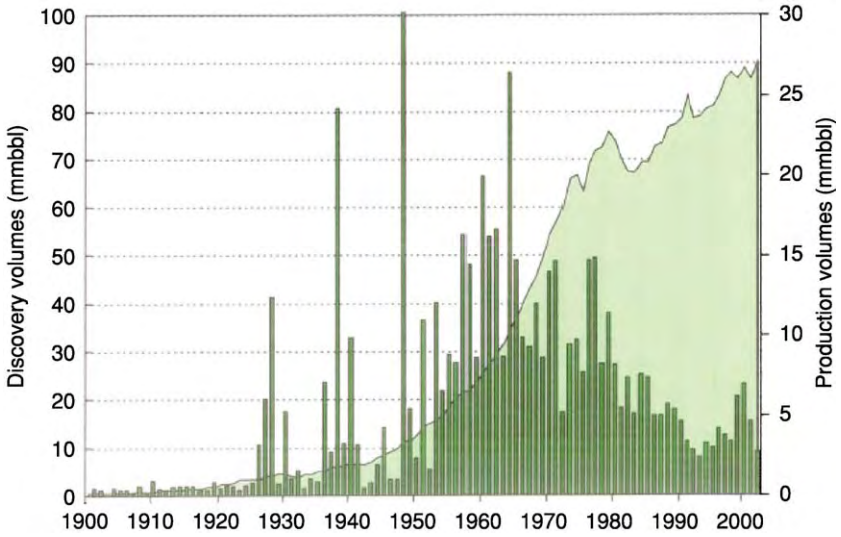


Figure 11 Discovery volumes and world oil production, 1900 2000. mmbbl, billion barrels.

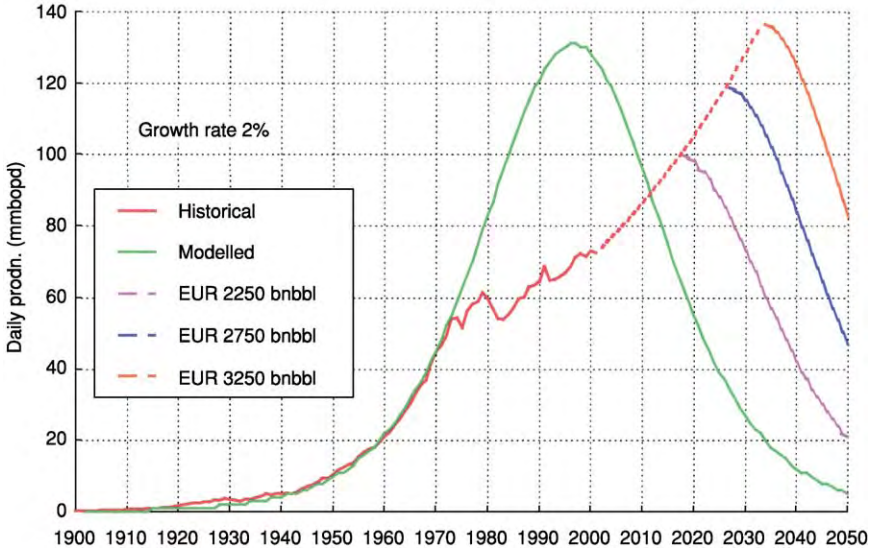


Figure 12 The effect of changing the resource base on the timing of peak oil. bnbbl, billion barrels; mmbopd, million barrels oil per day.

on the total resource base of oil, but also on dynamic variables, such as price, infrastructure, and technology. The level of peak production will also increase if recoverable reserves rise and, as a consequence, the timing of peak oil will be pushed further into the future (Figure 12).

Economists argue that oil reserves are not a 'fixed stock' and that energy has not been getting scarcer in basic economic terms, but rather has been getting more plentiful. The fall in the long run cost of oil production has been used in support of this argument. It is also argued that there is no such thing as an exhaustible natural resource and that the total petroleum reserves in the Earth are an irrelevant non-binding constraint. If oil becomes uneconomic to produce, the industry will disappear and whatever is left in the ground will remain unknown: a geological fact of no economic interest.

On the demand side, it has been proven wrong to project current rates of growth well into the future. For example, in the late 1970s, the history of nearly 30 years of an average 7% per annum rate of growth in oil demand still played a powerful role in determining attitudes to the longevity of oil resources. This was at a time when growth rates had already fallen to around 1% per annum and when most

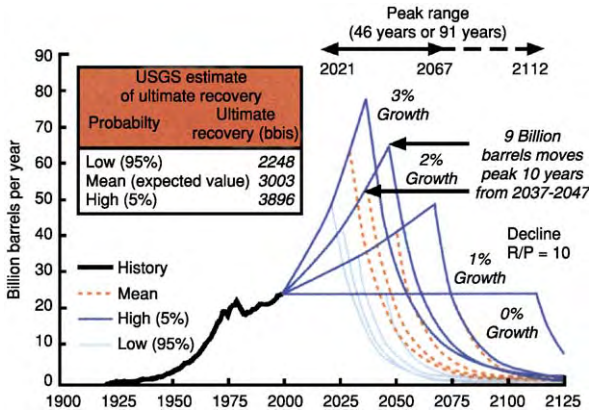


Figure 13 Conventional oil supply model showing the timing of peak oil with four rates of oil demand growth. R/P, reserves to production ratio; USGS, United States Geological Survey.

economists were forecasting a maximum growth rate of 3%. Changes in forecast to oil demand therefore have a major impact on the timing of peak oil; a difference of just 3% in average annual demand growth for oil gives a 46 year range in the timing of peak oil (Figure 13). Similar arguments can be applied to changes in the level of investment and the rate of depletion.

The Real Debate

The amount of fossil fuel on our planet is finite, certainly within the time-frame of civilization. However, does this necessarily mean that we will run out of petroleum reserves imminently or at all? Or will we develop alternative sources of energy long before the planet's petroleum reserves have been extracted? The changing pattern of world oil supply has meant that many of the predictions that oil would run out by the turn of the twentieth century have been proven wrong. Fears of impending scarcity of non-renewable energy have all been proven to be groundless as important changes on both the supply and demand sides have occurred.

The history of those who predict future trends in the depletion of petroleum reserves is one of persistent tension between those who assert that the best decisions are based on quantification and numbers, determined by the patterns of the past, and those who base their decisions on more subjective beliefs about the uncertain future. Ultimately, oil is a finite resource, but the real debate is not so much whether we are likely to significantly diminish the petroleum reserve base within the next few decades, but whether prices will fluctuate sufficiently to enable energy substitution. Volatility in the price of oil will sharpen the debate and could actually cause near-term unjustified changes in investment patterns by industry and governments alike.

See Also

Geological Surveys. **Petroleum Geology:** Overview; Chemical and Physical Properties; The Petroleum System; Exploration; Production. **Seismic Surveys**.

Further Reading

Adelman MA (1990) Mineral depletion, with special reference to petroleum. *The Review of Economics and Statistics* LXXII(1): 1 10.

BP (2003) *BP Statistical Review of World Energy 2003*. London: BP plc.

Campbell CJ (1997) *The Coming Oil Crisis*. Brentwood: Multi Science Publishing Company and Petroconsul tants.

Campbell CJ and Laherrere JH (1998) The end of cheap oil. *Scientific American* March: 80 86.

Deffeyes KS (2001) *Hubbert's Peak: The Impending World Oil Shortage*. Princeton, NJ: Princeton University Press.

Hubbert MK (1956) Nuclear energy and the fossil fuels. *Drilling and Production Practice*: 7 25.

Hubbert MK (1971) The energy resources of the Earth. *Scientific American* February: 31 40.

Laherrere JH (1999) World oil supply what goes up must come down, but when will it peak? *Oil and Gas Journal* 97: 57 64.

Lynch MC (1996) The analysis and forecasting of pe troleum supply: sources of error and bias. In: El Mallakh DH (ed.) *Energy Watchers VII*, pp. 51 71. Colorado: International Research Center for Energy and Economic Development.

Lynch MC (2001) Forecasting oil supply: theory and practice. *Quarterly Review of Economics and Finance* July: 1 28.

McCabe PJ (1998) Energy resources cornucopia or empty barrel? *American Association of Petroleum Geologists Bulletin* 82(11): 2110 2134.

Odell PR (2001) *Oil and Gas: Crises and Contro versies 1961 2000*. Brentwood: Multi Science Publishing Company Ltd.

Simmons MR (2001) *The World's Giant Oilfields. Energy Service Industry Research*. Houston: Simmons and Com pany.

Smith AJ and Lidsky BJ (1993) King Hubbert's analysis revisited: update of the lower 48 oil and gas resource base. *The Leading Edge* 12: 1082 1086.

Weeks LG (1958) Fuel reserves of the future. *American Association of Petroleum Geologists Bulletin* 42: 431 438.

PLATE TECTONICS

R C Searle, University of Durham, Durham, UK

Introduction

In the theory of plate tectonics, it is assumed that the Earth has a thin, brittle, outer layer, or 'lithosphere', which is broken up into a small number of tectonic 'plates'. These plates are further assumed to be rigid and undeformable, but capable of independent motion relative to each other on the surface of the Earth. The different relative motions of the plates are considered to be responsible for much of the varying geological activity on the Earth. The plate motions are describable in simple geometric terms and, once they have been determined, usually from a limited number of observations; they can be used to predict the type, magnitude, and direction of relative motions across plate boundaries anywhere in the world, both present and past. The power of plate tectonics theory arises both from its unifying view of global tectonics and from its powerful predictive ability. The subject of plate tectonics, *sensu lato*, is sometimes divided into 'plate kinematics', which deals with plate motions and their geometry, and plate tectonics *sensu stricto*, which deals with the geological consequences of the motions.

Tectonic Plates, Lithosphere, and Asthenosphere

The fundamental assumption underlying plate tectonics is that the outer part of the Earth consists of a small number of thin, rigid plates which are curved to fit the spherical shape of the Earth. The plates are considered to suffer no internal deformation, and only to deform and interact with each other along their boundaries, which are considered to be very narrow compared with the plate dimensions.

This view arose in the 1960s, largely as a result of the newly observed pattern of global seismicity provided by the World Wide Standard Seismograph Network. This showed that, over much of the world, earthquakes occur in narrow bands no more than a few tens of kilometres wide, whilst the large areas between these bands are essentially aseismic (Figure 1). The assumption was made that the aseismic areas are largely devoid of deformation and behave rigidly, and the bands of seismicity mark the plate boundaries. In simple plate tectonics theory, the plate boundaries are considered to have negligible width, although in practice they are associated with narrow but finite bands of geological activity (*see* **History of Geology Since 1962**).

Early workers on plate tectonics theory divided up the Earth's surface into about 12 major plates (Figure 2), although since then considerable numbers of smaller 'microplates' have been recognized.

The rigid outer layer of the Earth, which comprises the plates, is known as the 'lithosphere', from the Greek '*lithos*' meaning stone. Below this is a weaker layer, the 'asthenosphere' (Greek '*asthenia*', weakness), which deforms relatively easily and allows the plates to slide around. It is important to understand that these are terms describing the mechanical behaviour of the Earth and not its composition. In particular, the lithosphere does not normally correspond to the crust, but comprises both the crust and the uppermost part of the upper mantle, which usually behave together in a brittle manner (*see* **Earth:** Mantle; Crust). The lithosphere ranges in thickness from a few kilometres at mid-ocean ridge axes (where it may be entirely crust) to around a hundred kilometres in old ocean basins, and perhaps several hundred kilometres under continental cratons (where it is predominantly mantle). The asthenosphere constitutes a relatively weak layer of mantle below the lithosphere, capable of plastic deformation. The thickness of the asthenosphere is less well defined than that of the lithosphere, but may range from around 50 km in some places to several hundred kilometres, and may even encompass the whole upper mantle. Furthermore, except in the case of some subduction zones (see below), tectonic plate boundaries do not follow the boundaries between continents and oceans. An individual plate typically comprises regions of both continental and oceanic crust. For example, the South American plate consists of the continent of South America plus the western half of the South Atlantic Ocean (Figure 2).

The different mechanical behaviours of the lithosphere and asthenosphere are explained by the different deformation mechanisms that occur at different temperatures and pressures. At relatively low temperatures, most rocks deform elastically and suffer brittle failure (faulting) when their elastic limit or 'strength' is exceeded. This strength increases with pressure and hence depth (Figure 3). However, as the temperature increases, solid-state creep becomes increasingly important. Deformation by this mechanism becomes easier at higher temperatures, and the rock weakens exponentially (Figure 3). The depth at which brittle deformation gives way to ductile creep

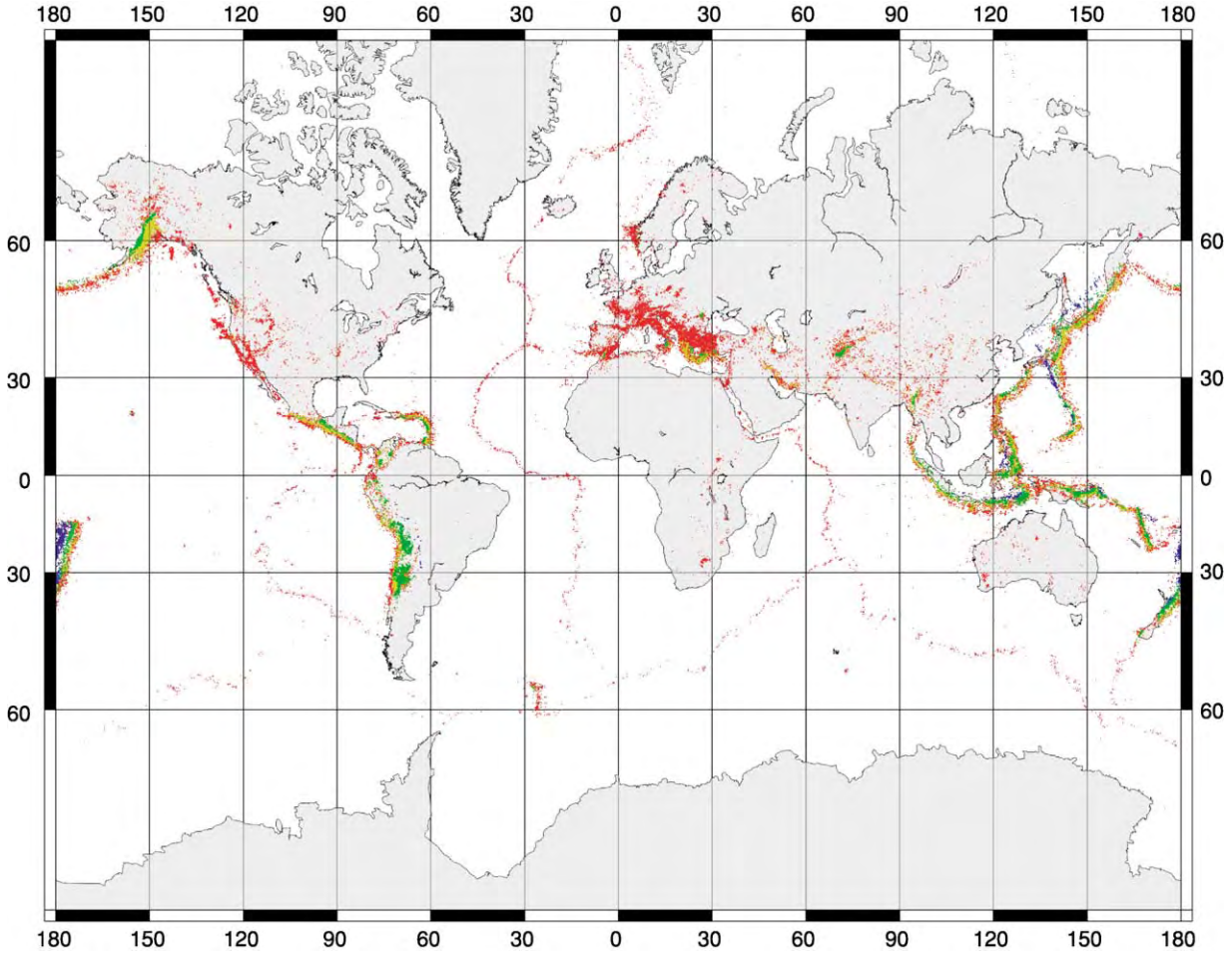


Figure 1 Distribution of shallow (0 33 km, red), intermediate (33 70 km, yellow; 70 300 km, green), and deep (300 700 km, blue) earthquakes of magnitude 4.0 or greater recorded between 1990 and 1996. Reproduced from D. Sawyer, http://www.geophysics.rice.edu/plateboundary/intro.html.

depends mainly on the composition of the rock, the rate of strain, and the geothermal gradient. Laboratory measurements of the mechanical properties of rocks and minerals predict depths of the brittle–plastic transition in the range suggested for the base of the lithosphere by observations of the Earth.

Evidence for the existence of the lithosphere and asthenosphere comes from a variety of sources. Seismic waves generated by earthquakes, particularly shear or S waves, show very little attenuation of their energy whilst travelling through the lithosphere, but are attenuated somewhat more in the asthenosphere (*see* **Tectonics:** Earthquakes). The velocities of these waves through the Earth show a marked decrease in the region of the asthenosphere. The effective viscosity of the Earth's mantle, as inferred from, for example, the rate of uplift of continental areas following the retreat of glaciers, shows a decrease in viscosity in the asthenosphere.

In general, crustal rocks are weaker than mantle, and, because oceanic crust is much thinner (around 6 km) than continental crust (approximately 35 km), oceanic lithosphere has a much greater component of mantle rocks and is correspondingly stronger. For this reason, oceanic areas tend to behave in a more rigid way than continents. In particular, continental mountain belts often display deformation over quite broad zones compared with oceanic plate boundaries (Figure 1).

Plate Tectonics

Although plate boundaries are primarily defined by the earthquakes along their boundaries, they are also marked by distinctive morphological features, narrow bands of tectonic activity, and, especially at transform boundaries (see below), by individual major faults. The active plate boundary may be only

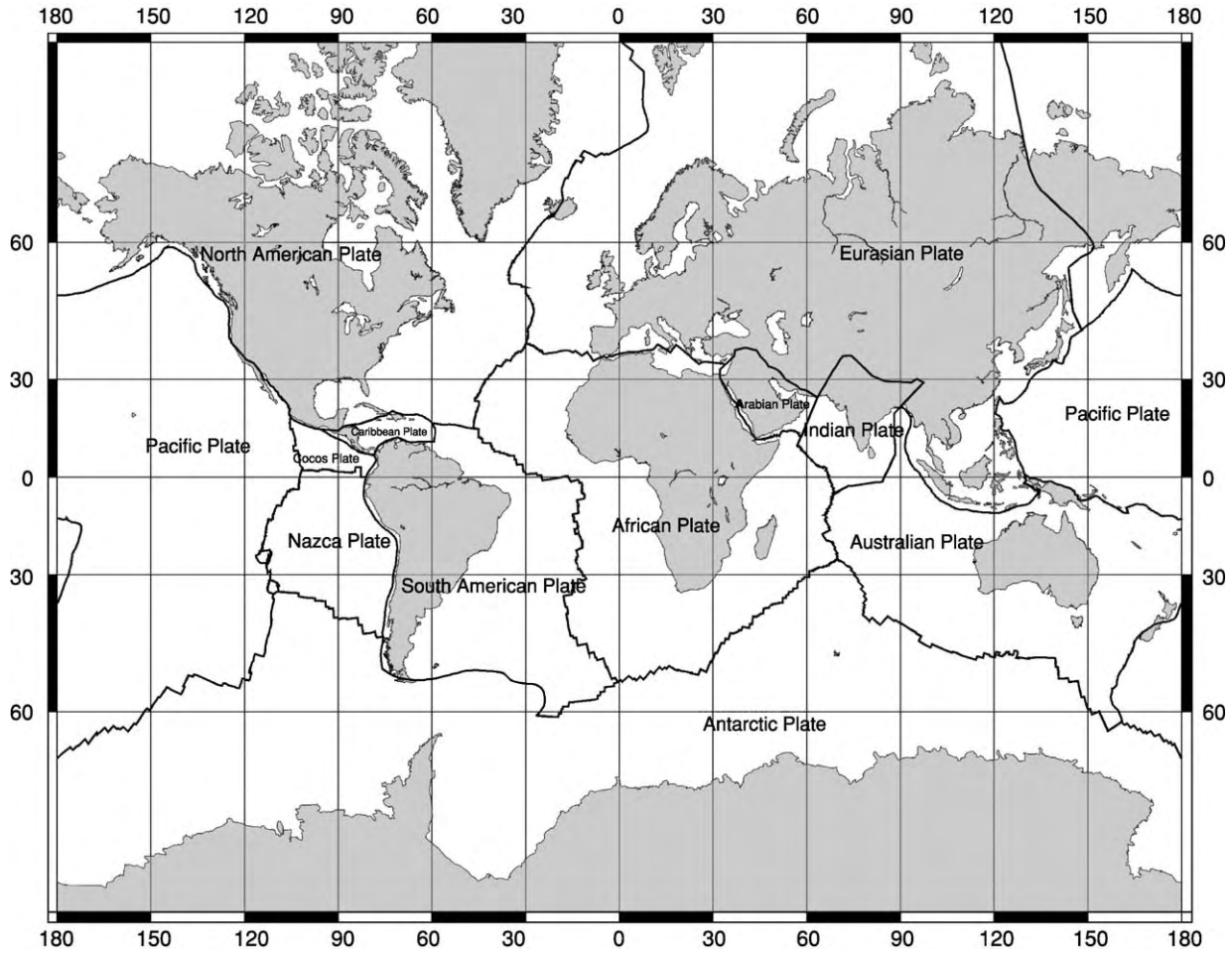


Figure 2 Major plates of the world and their boundaries. Reproduced from D. Sawyer, http://www.geophysics.rice.edu/plateboundary/intro.html, after D. Mueller, University of Sydney, Sydney, Australia.

a few kilometres wide, and rarely more than a few tens of kilometres.

Three types of plate boundary are recognized, corresponding to the three types of relative motion (Figure 4): 'divergent' (also 'ridge' or 'accretionary') plate boundaries, 'transform' (or conservative) plate boundaries, and 'convergent' ('subduction' or destructive) plate boundaries. In general, any type of plate boundary can join any other type.

Divergent (Ridge) Boundaries

Divergent boundaries occur only at mid-ocean ridges, and are the sites of creation of new lithosphere by seafloor spreading (*see* **Tectonics:** Mid-Ocean Ridges). The Mid-Atlantic Ridge and East Pacific Rise are typical examples. These boundaries are characterized by extensional tectonics dominated by normal faulting and extensional earthquake focal mechanisms. Generally, a divergent plate boundary is approximately orthogonal to the plate separation or spreading direction, but this is not essential, and ridges with greater or lesser extents of obliquity are fairly common.

The mid-ocean ridges associated with these plate boundaries arise because new plate is created at the boundary; thus, the plate gets older with increasing distance from the plate boundary and, as it does so, it cools, becomes denser, and thermally contracts. The plate boundary is thus the youngest, hottest, and shallowest region, forming the crest of the mid-ocean ridge. The detailed morphology of the plate boundary appears to depend on the temperature of the underlying mantle, which is mainly controlled by the spreading rate. For 'slow spreading ridges' (with a plate separation rate of less than about 60 km per million years), there is usually a 'median valley', a few kilometres deep and a few tens of kilometres wide, formed by rifting (Figure 5). At 'fast spreading ridges' (separation faster than 70 km per million years), the plate is too hot and weak to support rifting, and instead there is a volcanically built axial high.

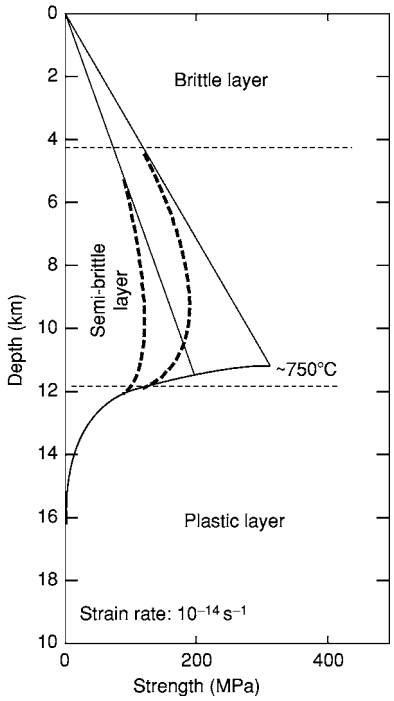


Figure 3 Calculated strength envelopes for material subject to brittle, semi brittle, and plastic deformation with increasing depth. Reproduced from Searle RC and Escartín J (2004) The rheology and morphology of oceanic lithosphere and mid ocean ridges. In: German C, Lin J, and Parson L (eds.) *Thermal Regime of Ocean Ridges and Dynamics of Hydrothermal Circulation. American Geophysical Union, Geophysical Monograph*, Fig. 8. Washington, DC: American Geophysical Union.

Ridges are formed by melting of dry mantle at depths of ~60 km, producing basaltic magma. Basaltic volcanism is widespread at mid-ocean ridge plate boundaries.

Plate tectonics theory allows for asymmetrical spreading (one plate accreting faster than the other). However, although temporary asymmetrical spreading is common, the net effect averaged over millions of years is usually approximately symmetrical. The precise reasons for this are not fully understood.

Transform Boundaries

Transform boundaries are those in which plates slide past one another with essentially no convergence or divergence; they thus conserve the areas of the adjacent plates. Because of this, they have the important property that they are exactly parallel to the direction of relative plate motion, and can be used for its estimation. Transform boundaries are characterized by strike-slip faulting and earthquakes with strike-slip mechanisms. A typical example of a transform boundary is the San Andreas Fault Zone in California, USA (Figure 6).

In simple plate tectonics theory, a transform boundary consists of a single fault called a 'transform fault', although, in practice, there is often a zone of such faulting some tens of kilometres wide. Because there is no large-scale convergence or divergence at these boundaries, there is also relatively little vertical relief,

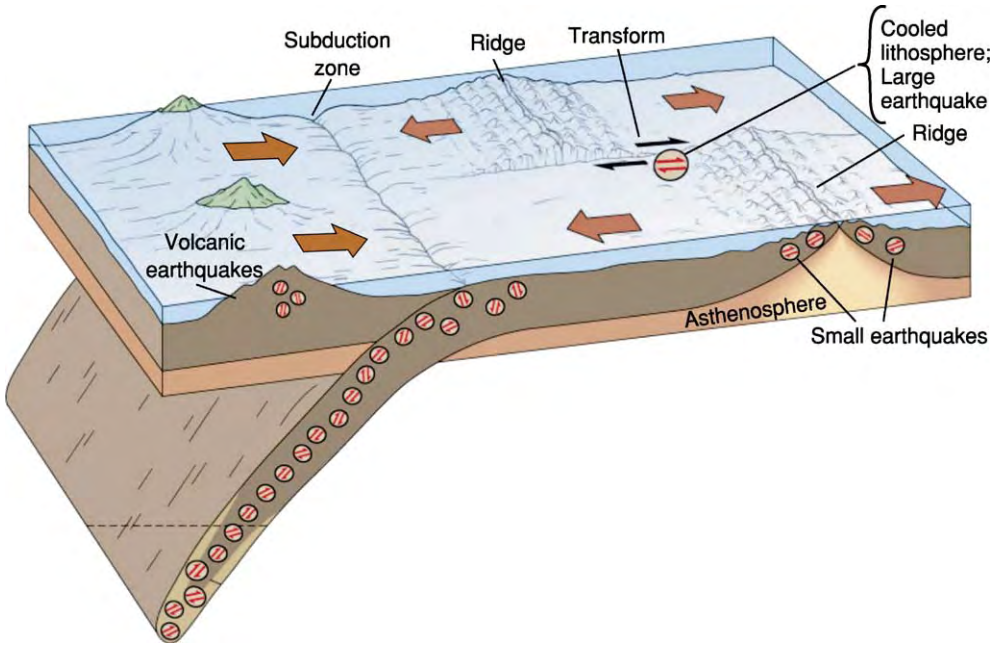


Figure 4 Diagram illustrating the three principal types of plate boundary (ridge, transform, and subduction), their general morphology, and the distribution of earthquakes. Large arrows show relative plate motion. Circles with double arrows indicate the general distribution of earthquakes. Modified from Davidson JP, Reed W, and Davis PM (2002) *Exploring Earth*, Fig. 8.13. Upper Saddle River, NJ: Prentice Hall.

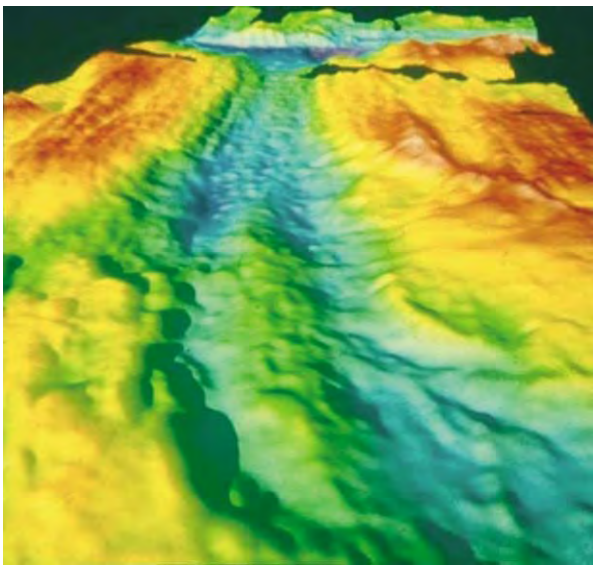


Figure 5 Oblique view of the topography of the Mid Atlantic Ridge Median Valley near 24° N, based on multibeam echosoun der measurements. The view is towards the south, with the cross cutting Kane Transform Fault in the distance. The width of the distant view is approximately 50 km, width in the foreground about 20 km, and distance from the foreground to the horizon about 100 km. Depths less than 2500 m, red; greater than 4200 m, blue. The median valley is approximately 30 km wide. There is evi dence of fault scarps on the valley walls. The ridge in the centre foreground is a linear volcano sitting above the plate boundary and representing the current axis of plate accretion. Image produced by M. Jones, University of Durham, Durham, UK, using data from Purdy GM, Sempere JC, Schouten H, Dubois DL, and Goldsmith R. (1990) Bathymetry of the Mid Atlantic Ridge, 24° N 31° N Amap series. *Marine Geophysical Researches* 12: 247 252; and Pockalny RA, *et al.* (1988) *Journal of Geophysical Research* 93: 3179 3193.

although small valleys and scarps usually mark the trace of the transform fault itself. Transforms link other plate boundaries, and can thus 'transform' one type of tectonic boundary (e.g., extensional) to another (e.g., compressional), hence the name. Many transform faults offset mid-ocean ridges, producing (in map view) a staircase pattern of alternating ridge and transform boundaries (Figure 2). Such offsets range from a few tens of kilometres to over a thousand kilometres in length. Transform boundaries occur in both continental and oceanic lithosphere.

Convergent Boundaries

Convergent plate boundaries in the strict plate tectonics sense occur only at the deep-sea trenches and related 'subduction zones' (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges). Areas of active mountain building on continents (e.g., the Alpine–Himalayan zone) are zones of plate convergence, but are characterized by broad and almost continuous zones of deformation. This is partly because continental lithosphere is weaker and more easily deformed than oceanic lithosphere, and partly because it is less dense and so is not readily removed by subduction (see below). Plate tectonics does not provide a very useful description of such broad continental convergence zones.

Where one of the converging plates consists of oceanic lithosphere, it will be overridden by the other and pushed down into the asthenosphere in a process known as subduction (Figure 7). This can occur because oceanic lithosphere (unlike continental lithosphere) has a similar density to the underlying asthenosphere. Where both plates are oceanic, either one may be subducted under the other, and sometimes, depending on local conditions, the 'polarity' of subduction may reverse (i.e., the subducted plate breaks off and subsequently becomes the overriding one, whilst that which was overriding begins to subduct).

The plate boundary at a subduction zone is marked by a deep trench, the deepest and probably most famous of which is the Mariana Trench in the western Pacific, whose base is over 11 km below sea-level (*see* **Tectonics:** Ocean Trenches). As the slab subducts, it is deformed and generates earthquakes. These occur in a narrow band, known as the Wadati–Benioff zone, that follows the position of the subducting slab, and constitutes some of the best evidence for the existence of subduction zones. Subduction zones are very complex, and contain local regions of both compressional deformation (mainly at the actual plate boundary and in the leading edge of the overriding plate) and extensional deformation (mainly where the subducting plate bends to begin its descent). Earthquake mechanisms reflect this, and also show variation in mechanism with depth in the subducting plate. Some of the world's largest earthquakes are associated with subduction zones.

As the slab descends, it heats up and gives off water trapped in the crustal rocks. The presence of this water lowers the melting point of the surrounding mantle and, at a depth of some 100 km, magma is generated that rises through the overriding plate to create a volcanic arc, a hundred or so kilometres behind the trench (depending on the angle at which the slab is descending).

As with ridges, oblique subduction is allowed by plate tectonics theory, and is relatively common.

Plate Kinematics

Rotation Poles

The great strength of plate kinematics is that it can describe plate motions in terms of simple geometry,

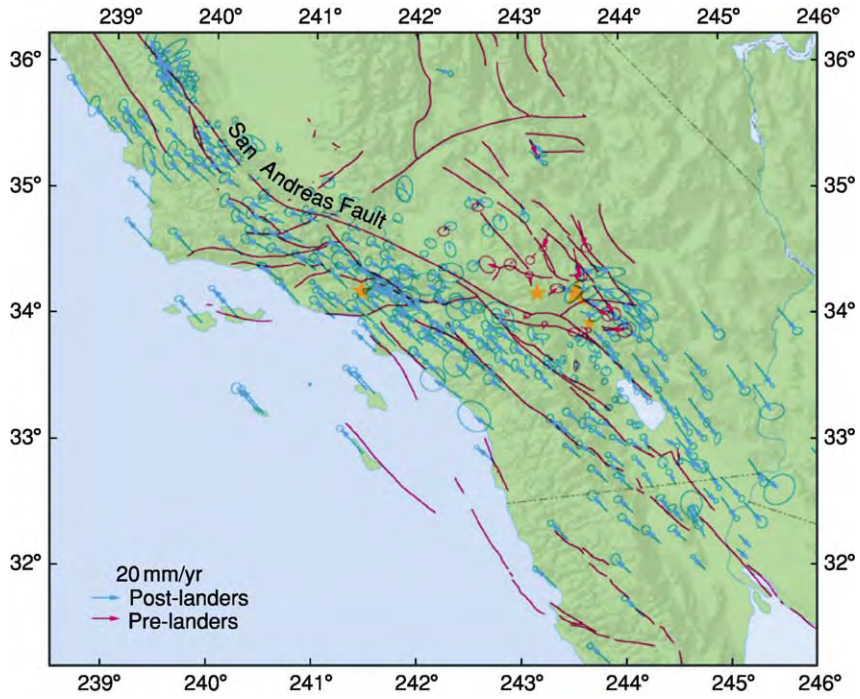


Figure 6 Map of major faults (red) and movement measured over 1 year by a Global Positioning System (GPS) (arrows) over part of the San Andreas Fault Zone in California, USA. Reproduced from Davidson JP, Reed W, and Davis PM (2002) *Exploring Earth*, Fig. 8.24. Upper Saddle River, NJ: Prentice Hall.

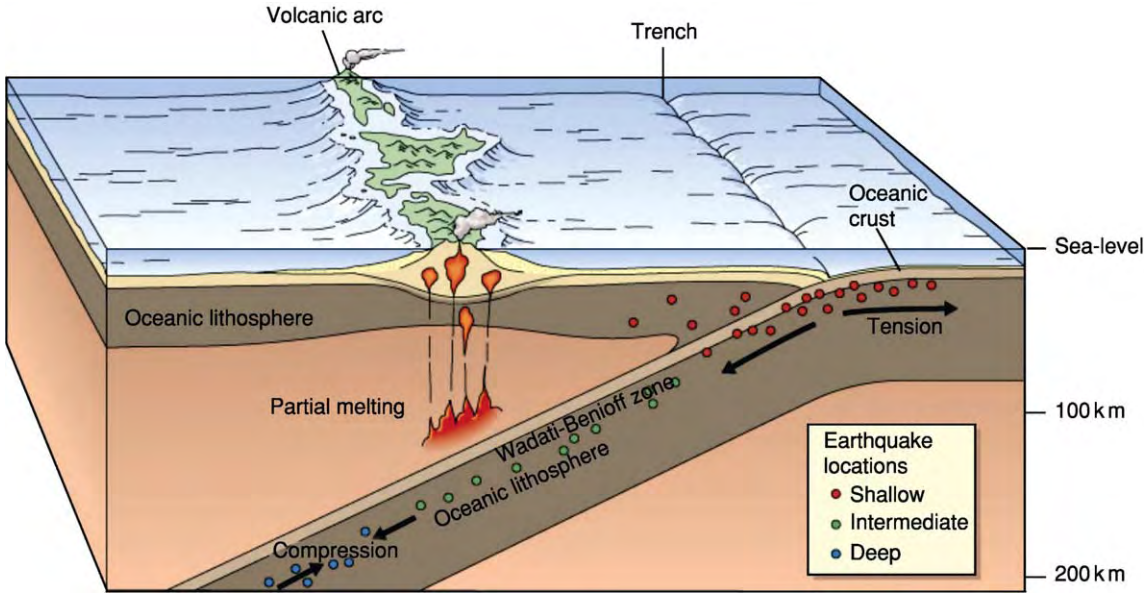


Figure 7 Diagram of a typical ocean ocean subduction zone. Reproduced from Davidson JP, Reed W, and Davis PM (2002) *Exploring Earth*, Fig. 10.3a. Upper Saddle River, NJ: Prentice Hall.

and hence make precise predictions of relative motions anywhere on the globe. At the heart of this geometry is a concept called Euler's theorem, which states that any displacement of a rigid body on the surface of a sphere can be described in terms of a single rotation about a specified axis (Figure 8). Such axes cut the Earth's surface at pairs of points called 'rotation poles' (or 'Euler poles'). Once the poles and the angular rotations are specified, the whole motion is completely determined. Thus, the

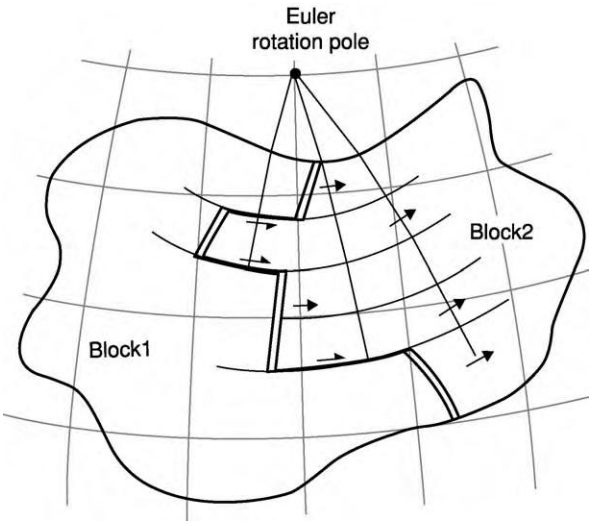


Figure 8 A Euler rotation pole describing the motion between Block 1 and Block 2. Plate boundary shown by double lines (mid ocean ridge) and bold single lines (transform faults). Circles (medium lines through transforms) and great circles drawn orth ogonal to them and passing through the pole are analogous to latitude and longitude in the geographic coordinate system. Light lines show the geographic coordinate system for comparison. Reproduced from Lowrie W (1997) *Fundamentals of Geophysics*, Fig. 6.30. Cambridge: Cambridge University Press.

motion of a given plate is specified in terms of its Euler pole and a corresponding angular rotation rate.

In practice, determining the motions of individual plates relative to some common reference frame is not easy, but determining 'relative' motions between pairs of plates is quite straightforward. These relative motions can also be described by rotation poles, although then the pole position must be specified relative to one or other of the pair of plates. Such relative rotation poles are the basis of most descriptions of plate kinematics.

As Euler poles define the 'directions' of the rotation axes, and the angles or rates of rotation define their 'magnitudes', rotations can also be described as vectors. Vector and matrix algebra can then be used to calculate plate motions, greatly simplifying and speeding up such calculations, which can be performed easily on computers.

It is important to distinguish between so-called 'instantaneous poles', which describe motion at an instant only, and 'finite rotation poles', which may describe the net result of motion over long periods of time. In descriptions of current plate motions, 'instantaneous' is usually taken to be about the past 1–3 million years. Motions over longer periods can be approximated by successions of so-called 'stage poles', each of which may describe the motion over a period of a few million years.

Like the geographic poles, we can imagine the Euler poles as centres of coordinate grids (Figure 8). 'Small circles' centred on the poles are the equivalent of lines of latitude, and plate relative motions are everywhere parallel to these small circles. Along any given small circle, the angular separation rate is constant, and increases with the distance from the rotation pole to the circle. 'Great circles' represent the shortest distance between two points on the surface of a sphere. Great circles that pass through the pole of rotation are equivalent to lines of longitude in the geographic system, and cut the small circles at right angles.

Measuring Plate Motions

Plates move at average rates of a few tens of millimetres per year. Thus, relatively indirect methods have usually been used to determine plate motions, although in recent years various geodetic measurements have been developed that are sufficiently precise to provide direct measurements of plate motions.

The most common way of determining plate separation rates is to use the linear magnetic anomalies (Vine–Matthews anomalies) produced during seafloor spreading to determine plate ages. These anomalies are produced as the Earth's magnetic field episodically reverses its direction, and the reversing field direction is recorded in the basalts of the oceanic crust as they cool after eruption at the mid-ocean ridge axis. This varying rock magnetization causes small variations, or 'anomalies', in the magnetic field above the seafloor, and these are readily observed by towing a magnetometer behind a ship or low-flying aircraft. Because the reversal process is irregular, the resultant magnetic anomalies have a characteristic pattern, which has been calibrated against crustal age through radiometric dating of samples of the magnetized rock. If the pattern of reversals can be recognized, the crust that is the source of the anomalies can thus be dated. The magnetic anomalies mark isochrons of crustal creation; therefore, by measuring the distance between a recent magnetic isochron on one plate and its conjugate on the other, the 'instantaneous' divergence rate, and hence the rotation rate, can be determined. (The commonly quoted 'spreading rate' is the rate at which a single plate accretes; for symmetrical spreading, it is half the divergence rate.) As the spreading or divergence rate varies with the distance from the Euler pole, in principle, the distance to the pole can be determined by measuring the spreading rate at several places along the plate boundary.

The best measure of the direction of relative plate motion is the azimuth of transform faults which, as stated above, are exactly parallel to the relative motion direction. Transform faults thus follow small

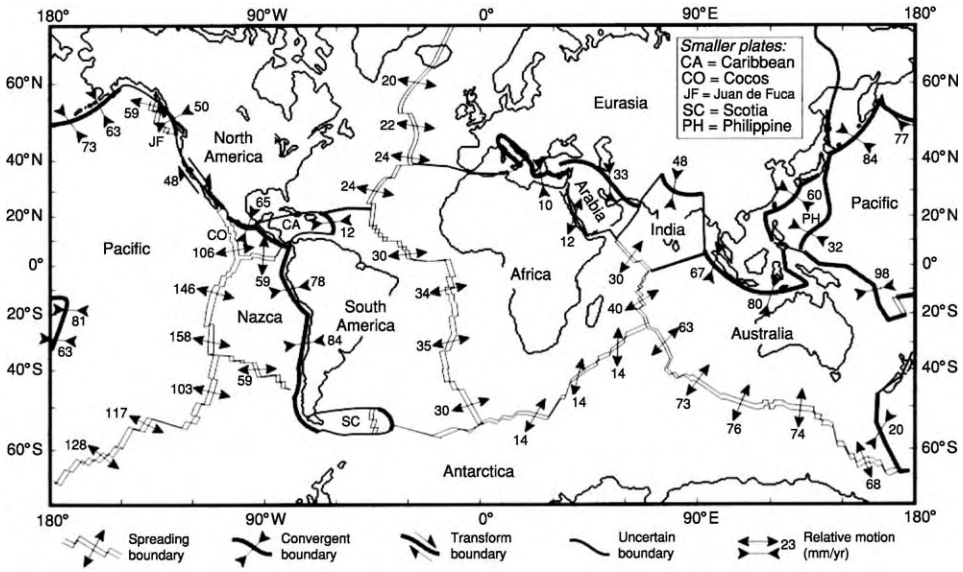


Figure 9 Map of plate boundaries showing the relative velocities across them from the data of DeMets C, Gordon RG, Argus DF, and Steins (1990). Reproduced from Lowrie W (1997) *Fundamentals of Geophysics*, Fig. 1.11. Cambridge: Cambridge University Press.

circles about Euler poles. A great circle at right angles to a small circle passes through the pole. Thus, if the azimuths of several transforms along a plate boundary are determined, great circles can be constructed normal to them, and should intersect at the Euler pole (Figure 8). Transform faults are readily recognized by their morphology (for example, a narrow linear valley for ridge–ridge transforms along mid-ocean ridges).

In practice, both spreading rates and transform azimuths are used together to solve for pole positions, sometimes supplemented by other data, such as earthquake focal mechanisms, to estimate relative motion directions. Euler poles may be calculated for individual plate pairs, for groups of plates, or for the global plate system (Figure 9). The most recent determination of global plate motions was performed by C. DeMets and others from Northwestern University, Illinois, USA, and provides a remarkably precise and self-consistent description of these motions (see 'Further Reading').

As stated above, an important attribute of plate kinematics is its ability to predict plate motions. An interesting example of this is the possibility of determining convergence rates across subduction zones. Even at ocean–ocean subduction zones, one plate is destroyed, together with the record of magnetic lineations carried on it. Thus, there was no direct way of measuring such motion until the recent development of sufficiently precise geodetic methods. However, the relative motions of the plate pair can be determined by global fits as described above, and then the motion at any point on the common plate boundary, including subduction zones, can be calculated from the Euler pole data.

In recent years, geodetic methods have been developed to the level at which they can begin to measure plate motions directly. Where plate boundaries exist on land (such as the Mid-Atlantic Ridge in Iceland or the San Andreas Fault in California, USA), standard geodetic methods, such as electronic distance measurement, can be used at a local scale (over ranges of a few kilometres). On a slightly larger scale of tens to hundreds of kilometres, precise relative position determinations (to precisions of a few millimetres) can be made by careful use of the Global Positioning System satellite network. Relative positions between widely separated continents can be determined by Very Long Baseline Interferometry, in which the variation in phase of radio signals from distant quasars is used. Repeat measurements by these methods over times of a few years can now resolve plate motions, and give results that, in general, agree well with the more traditional determinations.

The rotation rates of the major plates about their Euler poles range from about 2.1° per million years for the Cocos–Pacific pair, to about 0.1° per million years between Africa and Europe or Africa and Antarctica, and only 0.03° per million years for India–Arabia. Many of the minor plates (so-called 'microplates') rotate much faster than this, at tens of degrees per million years. In terms of linear rates, the fastest plate divergence rate at present is on the East Pacific Rise between the Pacific and Nazca plates, at

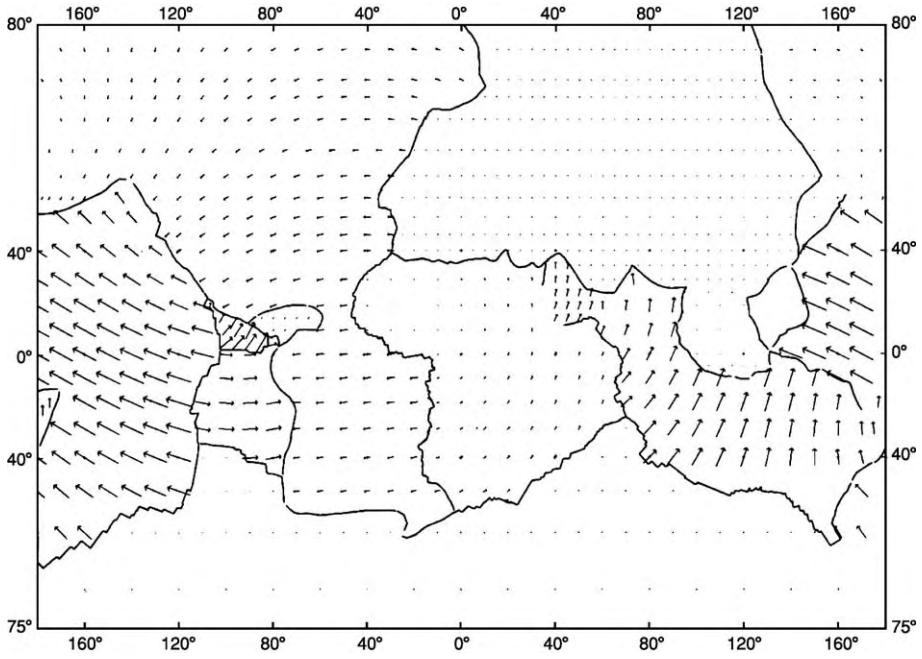


Figure 10 Map of plate boundaries and absolute motion vectors. Arrow lengths are proportional to the plate speed, with the longest (in the western Pacific) corresponding to speeds of about 100 mm per year. Reproduced from Fowler CMR (1990) *The Solid Earth: An Introduction to Global Geophysics*, Fig. 2.20. Cambridge: Cambridge University Press.

about 160 km per million years (or 160 mm per year). At the slow end, the North America–Eurasia plate boundary passes very close to its Euler pole in northern Siberia, where the relative motion becomes essentially zero.

Absolute Plate Motions

So far, we have dealt only with relative motions, which are fairly easy to determine. There is also interest in determining so-called 'absolute' plate motions, where the motions of all plates are related to some common reference frame. The possibility of doing this arises from the proposed existence of mantle plumes, which rise as narrow columns of relatively hot rock from deep in the mantle, possibly from the core–mantle boundary. They reach the Earth's surface in so-called 'hotspots' where they are manifest by clusters of intense volcanic and seismic activity. Well-known examples occur in Iceland and Hawaii, but there are thought to be many tens of such hotspots and associated plumes.

Hotspots leave clear trails on the Earth's surface, which comprise lines of volcanoes or volcanic seamounts and zones of thickened volcanically produced crust. The Hawaii–Emperor seamount chain, trending north-west from Hawaii, is an excellent example, but there are many other subparallel seamount trails in the south-western Pacific which are thought to have resulted from plumes. If points along the hotspot trails are dated (e.g., by radiometric dating of volcanic products), the relative motion between the plumes and the plate, and between given plumes, can be determined. We can also calculate the motions of individual plates relative to the average plume motion.

When this is performed, the relative motion between many of the hotspots turns out to be quite small, and significantly less than the average relative motions between plates. From this, and the fact that plumes are thought to rise through the mantle, it seems reasonable to assume that the average plume motion relative to the mantle is quite small, although it should be admitted that this has remained difficult to test in detail, and there are those who even question the existence of plumes. Nevertheless, if it is assumed that the average motion of hotspots relative to the mantle is zero, plate motions can then be given in the supposedly fixed mantle reference frame. These are referred to as absolute plate motions. They can also be described in terms of Euler poles, and are shown in Figure 10.

Mechanisms and Plate Driving Forces

Plates as Parts of the Mantle Convection Cycle

Plate motions are ultimately driven by the Earth's heat energy, and they are intimately related to the

mantle convection that is driven by this heat. One view of plates is that they simply represent the surficial parts of mantle convection cells: as hot, ductile mantle rises to the surface, it cools and becomes brittle – a plate – and then moves as a rigid block over the surface before being subducted, gaining temperature and becoming ductile again. Recent results from seismic tomography suggest that, around the rim of the Pacific, sheets of cold material descend below subduction zones deep into the lower mantle, implying a strong coupling of mantle motion and subducted plates.

However, the coupling is not perfect. There are some parts of the mid-ocean ridge (divergent plate boundaries) where it seems that the deeper mantle (below the asthenosphere) may be descending rather than rising. One such place is the so-called Australo-Antarctic Discordance south of Australia. Moreover, some plates, such as Africa, are almost entirely surrounded by ridges and have very few subduction zones on their boundaries. In such cases, a rigid coupling of plates to convection cells would imply the unusual scenario of upwelling along an expanding ring, with a downwelling column inside it. In fact, one of the advantages of plate tectonics is that it allows partial decoupling of plate motions from deeper mantle flow via the ductile asthenosphere.

The Forces Acting on Plates

Another way to look at the problem of the driving mechanism is to consider the forces acting directly on the plates. There are many possible forces, but amongst the most important are ridge push, slab pull, trench suction, and mantle drag. Ridge push arises from the tendency of the plates on a mid-ocean ridge flank to slide down the slopes of the wedge of thermally expanded asthenosphere that lies beneath the ridge. Slab pull arises from the negative buoyancy of the subducted plate which tends to drag the rest of the plate down with it. Trench suction is an additional force tending to pull plates together at subduction zones as a result of local convection driven by the subduction. Mantle drag is the frictional force between the base of the plate and the underlying asthenosphere.

It is possible to estimate some of these forces, at least approximately, and their relative importance can also be assessed by considering the observed stresses in plates and inferred absolute plate velocities. The latter is particularly instructive. Absolute velocities are largely independent of the total area of the plate, and so it is unlikely that mantle drag is an overall driving force, as was once thought (i.e., plates do not ride as passive passengers on top of mantle convection cells). However, plate velocities are inversely correlated with the area of continental lithosphere, suggesting that large areas of continent act as a brake (perhaps because such lithosphere is very thick or because subcontinental asthenosphere is cold and rather viscous). The fastest plates are those, mainly in the Pacific, which have large lengths of subduction zone along their boundaries, implying that slab pull and/or trench suction are important driving forces, as also suggested by calculation. There is a modest correlation with the effective length of ridge on the plate, indicating that ridge push is a driving force, but less strong than trench pull; this also tends to be backed up by calculations. Observations of intraplate stresses are also consistent with these conclusions.

Tests of Plate Tectonics

There have been numerous tests of plate tectonics theory. Its self-consistency, direct measurements of predicted plate motions, earthquake focal mechanisms, and distributions of earthquakes have all played their part in confirming the theory.

See Also

Earth: Mantle; Crust. **History of Geology Since 1962**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Earthquakes; Mid-Ocean Ridges; Mountain Building and Orogeny; Ocean Trenches.

Further Reading

Cox A and Hart RB (1986) *Plate Tectonics How It Works*. Oxford: Blackwell Scientific Publications.

DeMets C, Gordon RG, Argus DF, and Stein S (1990) Current Plate Motions. *Geophysical Journal International* 101: 425 478.

DeMets C, Gordon RG, Argus DF, and Stein S (1994) Effect of recent revisions to the geomagnetic reversal timescale on estimates of current plate motions. *Geophysical Research Letters* 21(20): 2191 2194.

Isacks B, Oliver J, and Sykes LR (1968) Seismology and the new global tectonics. *Journal of Geophysical Research* 73: 5855 5900.

Kearey P and Vine FJ (1996) *Global Tectonics*, 2nd edn. Oxford: Blackwell Science.

Le Pichon X (1968) Sea floor spreading and continental drift. *Journal of Geophysical Research* 73: 3661 3697.

McKenzie DP and Parker RL (1967) The North Pacific: An example of tectonics on a sphere. *Nature* 224: 125 133.

Morgan WJ (1968) Rises, trenches, great faults and crustal blocks. *Journal of Geophysical Research* 73: 1959 1982.

Wilson JT (1965) A new class of faults and their bearing on continental drift. *Nature* 207: 343 347.

PRECAMBRIAN

Contents

Overview

L R M Cocks, The Natural History Museum, London, UK

Introduction

Since the Precambrian–Cambrian boundary can now be dated at 543 Ma ago, and since the Earth is estimated to have been formed at about 4500 Ma ago, it follows that the Precambrian represents about seven eighths or 88% of geological time. However, that fact was not at all obvious to early geologists who, 200 or so years ago, had no idea of the age of the Earth. Between about 1820 and 1845, Earth history was divided into various named systems, with the Cambrian the oldest, based on the successive assemblages of distinct fossils contained in the sedimentary rocks and on the Law of Superposition, which states that rocks are older than other rocks now above them (assuming that they have not been structurally overturned). Igneous rocks were either undated or dated relatively as younger than the sedimentary rocks through which they had been intruded. Since no unambiguous fossils were then known from rocks below the Cambrian, all such rocks were and are simply termed the Precambrian. It was not until the invention and progressive refinement of radioisotopic dating in the twentieth century that the true ages of both the Precambrian and the rocks above it began to be understood.

Divisions of the Precambrian

Unlike the Phanerozoic (Cambrian–Holocene), there are no formal subdivisions of the Precambrian. However, the Precambrian is normally divided into two: the earlier Archaean and the younger Proterozoic. The Proterozoic has itself been divided into three, and the Precambrian is thus divided as follows (the ages are obviously approximate):

Neoproterozoic – 1000 Ma–543 Ma,
Mesoproterozoic – 1600 Ma–1000 Ma,
Palaeoproterozoic – 2500 Ma–1600 Ma, and
Archaean – *ca*. 4000 Ma–2500 Ma

No Earthly rocks are known that are older than about 4000 Ma, so there is no universally used name for the period between the formation of the Earth at about 4500 Ma and the formation of the oldest known rocks of the Archaean, although the Hadean is a term used by some. The Neoproterozoic is divided into the Riphean (1000–600 Ma) and the Vendian (600–543 Ma), and the Vendian is sometimes termed the Ediacaran.

The Precambrian in this Encyclopaedia

A large number of articles in this encyclopaedia deal in various ways with aspects of the Precambrian, and thus this article is devised to help the reader to locate the appropriate entry through cross-referencing. Thus there follows a brief guide to the key aspects of the Precambrian, under the headings the origin of the Earth, major Precambrian outcrops, Precambrian sediments and climate, the origin of life and Precambrian fossils, Precambrian orogenies and Precambrian terranes and palaeogeography.

The Origin of the Earth

Our galaxy and the universe appear to have been formed about ten thousand million years (10 Ga) ago, but our star – the Sun – and its Solar System do not seem to have been formed for a long time after that, at about 4.5 Ga. The article on Earth structure and origins (*see* **Earth Structure and Origins**) describes how this process is thought to have occurred. The oldest known individual minerals are found within

zoned zircons from Western Australia, which give an age of nearly 4.4 Ga. However, the oldest known rocks are the Acasta Gneisses of Canada, which are about 4.0 Ga old; this age is closely followed in other areas, for example the 3.9 Ga Napier Complex of Antarctica. The oldest cratons so far identified form parts of the South African and West Australian shields (*see* **Shields**) and are dated at about 3.2 Ga.

Major Precambrian Outcrops

Although slivers of Precambrian rocks are found in many terranes consisting mainly of younger rocks (*see* **Terranes, Overview**), the great majority of Precambrian outcrops are found in the old shield areas of North America (principally the Canadian Shield, North American Precambrian Continental Nucleus (*see* **North America:** Precambrian Continental Nucleus; Continental Interior), South America (largely underlying Brazil and the adjacent areas), Africa (several areas) (*see* **Africa:** Pan-African Orogeny), Northern Europe (the East European Craton (*see* **Europe:** East European Craton)), Siberia (the Angara Craton, Central Asia (*see* **Asia:** Central), and Russia (*see* **Russia**)), India (*see* **Indian Subcontinent**), Antarctica (*see* **Antarctic**), and Australia (*see* **Australia:** Proterozoic). Several of these shield areas are made up of more than one shield unit; for example, three individual shields can be identified within Australia, and another three are found within Antarctica. From late in the Proterozoic and for all of the Palaeozoic, the South American, African, Indian, Antarctic, and Australian shields were grouped together to form the supercontinent of Gondwana (*see* **Gondwanaland and Gondwana**).

Precambrian Sediments and Climate

The Earth's atmosphere in Archaean and pre-Archaean times was a reducing one, consisting largely of carbon dioxide, nitrogen, water vapour, and inert gases, with subsidiary amounts of hydrogen, methane, and ammonia. A critical event in Earth history was the change just before about 2.2 Ga in the atmosphere from reducing to oxidizing, termed the Great Oxidation Event. The sediments reflect that profound change; for example, a characteristic sediment of many Archaean and Early Palaeoproterozoic shields is the banded iron formation (*see* **Sedimentary Rocks:** Banded Iron Formations). This sediment type is first known from the Isua Supracrustal rocks of Greenland, dated at 3.8 Ga, and these chemically precipitated deposits, which contain up to 35% iron, reach their maximum distribution between 2.8 Ga and 2.5 Ga, the end of the Archaean. They are often associated with greywackes deposited by turbidity currents in 'greenstone belts', which are extensive sedimentary mudstones and coarser rocks (usually metamorphosed) with high percentages of iron and other minerals. Following the change in the atmosphere from reducing to oxidizing, there are no true banded iron formations, although comparable granular iron formations are known from a few areas, notably Lake Superior, Canada (dated at about 1.9 Ga), and, much later (Late Neoproterozoic), the Yukon and Namibia.

Each grain of clastic sediment at the bottom of the oceans today has been through several (an estimated average of five) cycles of erosion and sedimentation during Earth history. Since there was no vegetation on the land and soils only developed progressively with time, Precambrian erosion rates were much higher than those of today. Nearly all Early Archaean sediments were volcanogenic, usually basic or intermediate in chemical composition, but today these account for only about 20% of sediments, most of the earlier ones having broken down into their individual mineral components. Since quartz is more resistant to weathering than most other minerals, the proportion of quartz in sedimentary rocks has progressively increased, reaching today's figure of approximately 50%. Limestones became common for the first time in the Proterozoic and often contain algal stromatolites. These Precambrian limestones had a higher magnesium content than average modern limestones, which resulted in a higher proportion of dolostones. A form of coal termed schungite is found for the first time in rocks aged about 2.0 Ga, and, although now metamorphosed, it probably formed initially from local accumulations of blue-green algae. Chert is known from Archaean rocks onwards and occurs in the form of layers within banded iron formations and as nodules within dolomites; however, these cherts are of inorganic or secondary origin, and primary cherts, formed from siliceous organisms such as radiolaria, are not known until the Late Neoproterozoic.

Precambrian climates have formed the topic of many research projects. As can be seen from the desert sediments that many Precambrian rocks contain, the climates were often very hot; however, at other times, glacigenic deposits were widespread. These Precambrian glaciations occurred in several phases, and some workers believe that the Earth was at one time completely covered in glaciers, the so-called 'snowball Earth', but that concept is controversial, although palaeomagnetic evidence indicates that some glacigenic rocks in Australia must have been close to the then equator.

The Origin of Life and Precambrian Fossils

It is not known when life originated here on Earth; however, as far as can be deduced, life originated only once. Prokaryotes, which include bacteria (*see* **Biosediments and Biofilms**), are apparently simpler than eukaryotes, since they lack a nucleus, and were probably the first organisms. Because early organisms were largely soft-bodied, the early fossil record is tantalizingly incomplete. As detailed in the article on Precambrian prokaryotes (*see* **Precambrian:** Prokaryote Fossils), the changes in overall carbon isotopes seen after 3.5 Ga indicate that biological systems were in place by that time. This was well before the Great Oxidation Event at about 2.2 Ga, and thus early life forms must have lived in a reducing environment.

Laminated structures known as stromatolites have been described from rocks as old as 3.4 Ga in Australia and South Africa; however, it appears likely that the earliest stromatolites were inorganic in origin. In contrast, stromatolites deposited after about 3.0 Ga were apparently organic in origin. Although the range and diversity of these complex organisms peaked in the Middle Mesoproterozoic and Neoproterozoic, their descendants are still to be found living today, most famously at Shark Bay, Western Australia. The role of stromatolites and other early organisms in taking carbon dioxide out of the atmosphere and replacing it with free oxygen played a major part in the change of the atmosphere during the Great Oxidation Event at about 2.2 Ga.

As the Precambrian progressed, life forms became more diverse (*see* **Precambrian:** Eukaryote Fossils), although they rarely exceeded 1 mm in size. A major evolutionary event took place in the later Proterozoic Vendian period (*see* **Precambrian:** Vendian and Ediacaran), when much larger macroscopic fossils, albeit still without hard parts, evolved; these are known as the Ediacara fauna. That fauna has been found on most of the larger Late Precambrian terranes, but apparently became extinct at or near the Precambrian–Cambrian boundary.

Precambrian Orogenies

In the Archaean there were many orogenies, but in most cases their detailed ages and lateral extents have not yet been fully determined, and long periods of orogeny are often conflated under a single name. For example, various orogenies have been recognized as occurring during the formation and subsequent fusion of the three cratons that make up present-day Australia. However, the 2.6–2.4 Ga Sleafordian Orogeny of Australia marked the assembly of the South Australia craton and spanned the Archaean–Palaeoproterozoic boundary (*see* **Australia:** Proterozoic). In central Canada the Trans-Hudson Orogeny extended from 2.2 Ga to 1.7 Ga, peaking at 1.9 Ga (*see* **North America:** Precambrian Continental Nucleus). Some workers correlate the 1.9–1.8 Ga Barramundi Orogeny of the North Australian craton with the Trans-Hudson Orogeny of North America, and the two together form a key element in the assembly of the Rodinia supercontinent (see below). In Antarctica there are at least three different Archaean–Palaeoproterozoic cratons that formed between 3 Ga and 1.6 Ga (*see* **Antarctic**).

The best-known Precambrian orogeny in the Mesoproterozoic is the Grenvillian Orogeny (*see* **Grenvillian Orogeny**), which occurred between 1.3 Ga and 1.0 Ga and probably resulted in the assembly of the superterrane of Rodinia. The Grenvillian Orogeny is principally represented in eastern North America and consists of an earlier accretionary stage between 1.3 Ga and 1.2 Ga, widespread magmatism from 1.2 Ga to 1.1 Ga, and a continent–continent collision at about 1.0 Ga. The 1.3–1.1 Ga Albany Fraser Orogeny of the South and West Australian cratons has been interpreted as part of the Grenvillian Orogeny.

During the final part of the Neoproterozoic, the Timanide Orogeny (*see* **Europe:** Timanides of Northern Russia) affected today's northern Europe and resulted in the accretion of a substantial number of terranes to the old craton of Baltica. In addition, the Cadomian Orogeny affected most of the terranes now in southern Europe; this continued until about 530 Ma, in the Early Cambrian.

Precambrian Terranes and Palaeogeography

It is not known exactly when continental and oceanic crust became differentiated on the Earth's surface, but the process is thought to have started before the Archaean. The oldest known cratons are dated at about 3.2 Ga, and it is not until after the formation of these cratons that we can be sure that wide shallow continental-shelf seas actually existed. Although there are glimpses of early terranes, for example between 3.0 Ga and 1.6 Ga (Archaean–Palaeoproterozoic) several Antarctic cratons were stabilized, there is no clear image of global palaeogeography prior to the Late Mesoproterozoic.

The existence of a Precambrian superterrane was suggested in the 1970s, as geologists noted a number of 1.3–1.0 Ga mountain belts now on different continents, and later the name Rodinia was adopted for

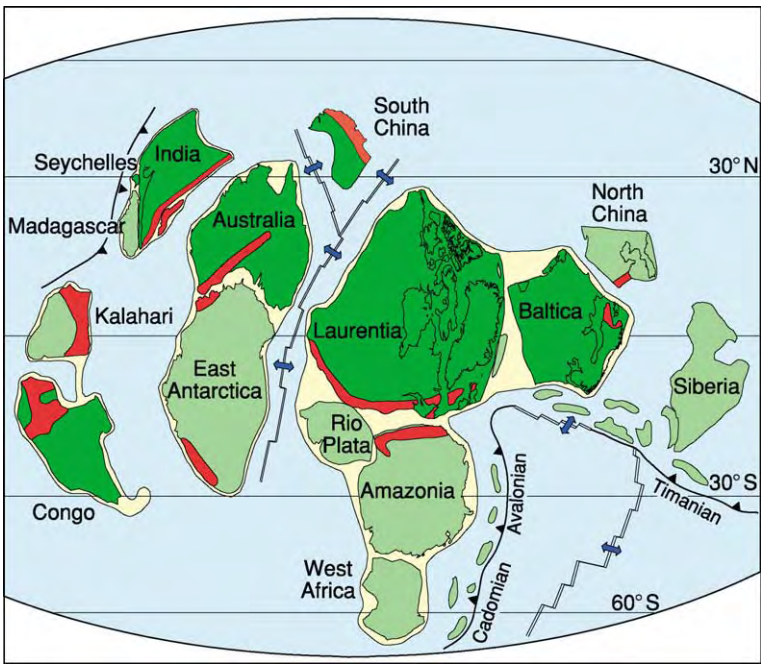


Figure 1 Palaeogeographical reconstruction at 750 Ma, just after the initial breakup of the Rodinia superterrane. The dark green shading represents terranes with good palaeomagnetic data, in contrast to the light green from with palaeomagnetic data is poor or absent. The red areas are the tectonic belts. Modified by Trond Torsvik, Trondheim, from Torsvik (2003).

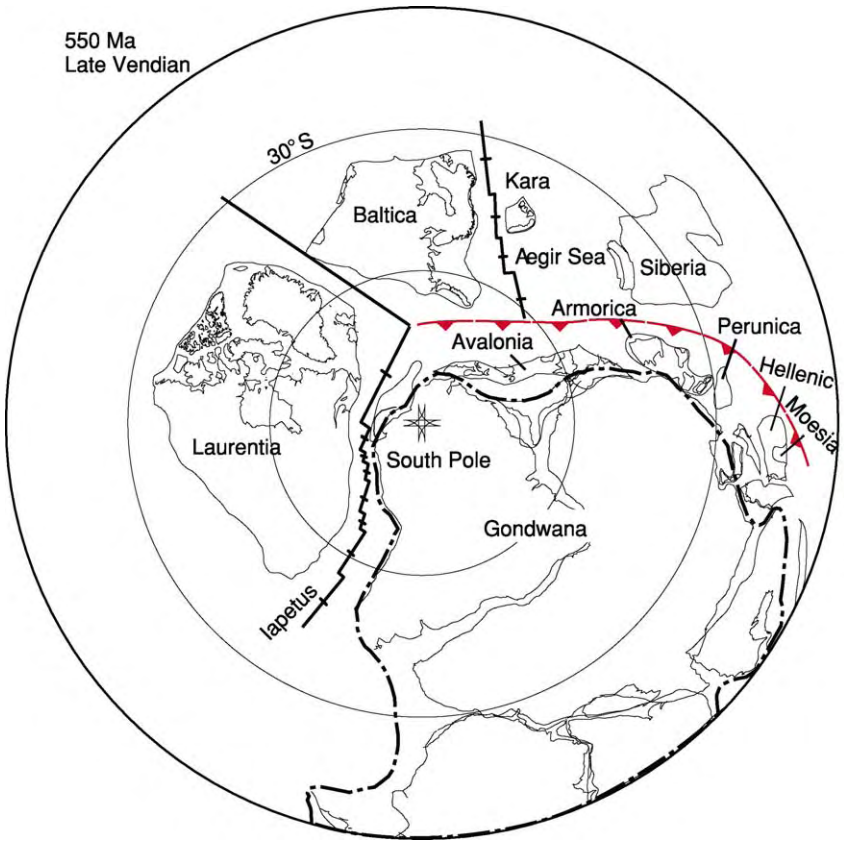


Figure 2 Palaeogeographical reconstruction at 550 Ma, near the end of the Neoproterozoic. New diagram by Trond Torsvik, Trondheim.

this terrane. There is now a progressively convincing scenario emerging of the slow assembly of the superterrane of Rodinia during the Mesoproterozoic between 1.3 Ga and 1.1 Ga. After relative stability for a long period, Rodinia started to break up at about 850 Ma. That breakup is documented by several events; for example, the 760 Ma Areyonga event in Australia may represent rifting between Australia and the new terrane of Laurentia. Similarly, Laurentia appears to have separated from Antarctica at about 750 Ma (**Figure 1**). However, South America, Africa, peninsular India, Australia, and Antarctica stayed together and, with further smaller accretions, remained together to form the superterrane of Gondwana in the Late Neoproterozoic, a process that was complete by the Early Cambrian. Between 580 Ma and 555 Ma there was massive plutonism in Scandinavia, and flood basalts in the Ukraine occurred at 576 Ma; these events may document the separation of Baltica from Laurentia. Gondwana remained a united superterrane until after the breakup of Pangaea in the Early Mesozoic.

Figure 2 shows a probable palaeogeography for the Late Neoproterozoic. The progressive rifting between Laurentia (North America) on the one hand and Gondwana and Baltica on the other resulted in the opening and steady widening of the Iapetus Ocean between the two sets of terranes, an opening that continued until about the end of the Cambrian. In addition the opening of the Ran Ocean marked the division between Baltica and Gondwana, and the Aegir Sea separated Baltica from Siberia. Most of the other side of the globe was occupied by the vast Panthalassic Ocean.

See Also

Africa: Pan-African Orogeny. **Antarctic**. **Asia:** Central. **Australia:** Proterozoic. **Biosediments and Biofilms**. **Earth Structure and Origins**. **Earth System Science**. **Europe:** East European Craton; Timanides of Northern Russia. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Indian Subcontinent**. **North America:** Continental Interior. **North America:** Precambrian Continental Nucleus. **Precambrian:** Eukaryote Fossils; Prokaryote Fossils; Vendian and Ediacaran. **Russia**. **Sedimentary Rocks:** Banded Iron Formations. **Shields**. **Terranes, Overview**.

Further Reading

Cocks LRM (ed.) (1981) *The Evolving Earth*. Cambridge: Cambridge University Press.

Edwards K and Rosen BR (2000) *From The Beginning*. London: The Natural History Museum.

Hancock PL and Skinner BJ (2000) *Oxford Companion to the Earth*. Oxford: Oxford University Press.

Hartz EH and Torsvik TH (2002) Baltica upside down: a new plate tectonic model for Rodinia and the Iapetus Ocean. *Geology* 30: 255 258.

Torsvik TH (2003) The Rodinia jigsaw puzzle. *Science* 300: 1379 1381.

Windley BF (1995) *The Evolving Continents*, 3rd edn. New York: John Wiley.

Eukaryote Fossils

S Xiao, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Introduction

Eukaryotes, archaeobacteria, and eubacteria are the three domains of living organisms (Figure 1). Most life forms that we can see with the naked eye are multicellular eukaryotes, which include animals, fungi, plants, and seaweeds. Many eukaryotes (for example, dinoflagellates, ciliates, and amoebae) are, however, single-celled and microscopic.

Eukaryotes are cytologically distinct from prokaryotes (archaeobacteria and eubacteria). A typical eukaryotic cell contains membrane-bound intracellular structures such as a nucleus, mitochondria, and, for photosynthetic eukaryotes, chloroplasts. Eukaryotes are also distinctively characterized by DNA-associated histone, eukaryotic gene regulation, and tubulin- and actin-based structures known as cytoskeletons, which help to maintain and manipulate the shape of the cell.

Nuclei, mitochondria, chloroplasts, histone, and cytoskeletons are rarely preserved in the fossil record, making it a serious challenge to recognize fossil eukaryotes, particularly single-celled ones. It has been suggested that cell size may be a useful guide: eukaryotes tend to have larger cells than prokaryotes. The size range of eukaryotic cells, however, overlaps considerably with that of prokaryotic cells. Cell size alone therefore is only indicative, not conclusive, evidence of eukaryotic affinity. Morphological complexity provides more reliable evidence. Because of the supportive function of cytoskeletons,

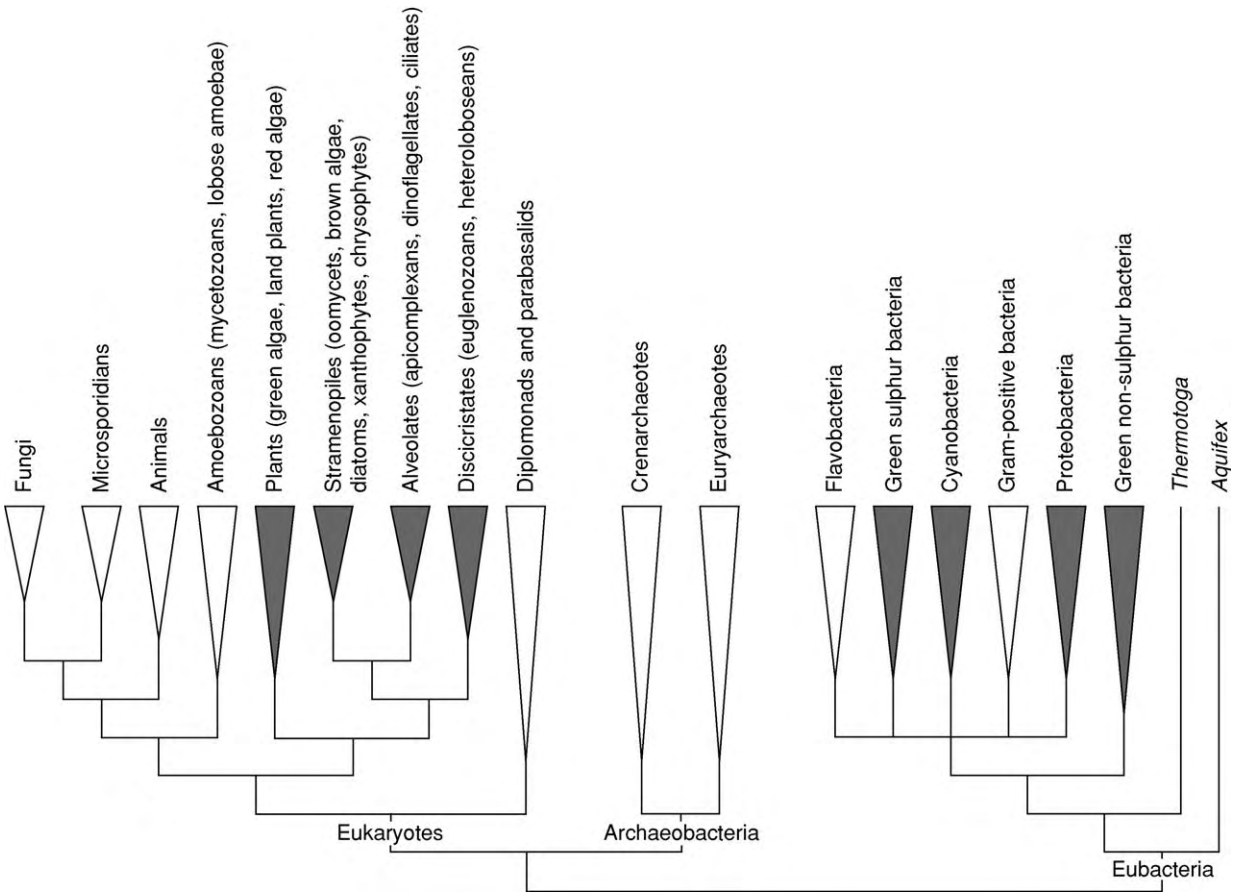


Figure 1 Phylogenetic relationships among the three major domains (eubacteria, archaeobacteria, and eukaryotes) and phylogen etic structure within the eukaryote domain. Shaded clades include photosynthetic members. Simplified from Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the USA* 87: 4576 4579 and Baldauf SL, Roger AJ, Wenk Siefert I, and Doolittle WF (2000) A kingdom level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972 977.

many eukaryotes maintain complex cell morphologies that can be preserved in the fossil record. For multicellular organisms, eukaryotic morphologies extend to cell organization and differentiation. Through intercellular interaction (e.g. via plasmodesmata and pit connections), controlled cell division, and genetic regulation, multicellular eukaryotes can have well-organized cell arrangements and some degree of cellular differentiation, which distinguishes them from prokaryotes. In addition, eukaryotic biochemistry may leave fingerprints in ancient rocks in the form of molecular fossils or biomarkers. Steranes, for example, are good eukaryote biomarkers. Steranes are derived from steroids, and the biosynthesis of steroids is almost exclusively a characteristic of eukaryotes.

Not only the size range but also the morphological disparity of eukaryotes overlaps with that of prokaryotes. Many eukaryotes, particularly single-celled ones, can be morphologically simple and volumetrically small. It can be very difficult to distinguish such eukaryotes from prokaryotes. This is a significant challenge because the majority of Precambrian microfossils are morphologically simple. Many of them are featureless filamentous or spheroidal structures, less than a few tens of micrometres in size, preserved in cherts or shales. The default interpretation of these microfossils is that they are prokaryotes. This is not an unreasonable interpretation for many, particularly if they can be shown to be microbial mat builders. For others, however, a eukaryote interpretation is equally plausible; there is simply not enough morphological detail to allow an unambiguous interpretation. These fossils are not the focus of this article. The discussion below takes a more conservative approach, focusing only on the Precambrian fossils that can be reasonably interpreted as eukaryotes. In addition, the discussion will concentrate on marine sediments; there are no known convincing eukaryote fossils in Precambrian terrestrial sediments.

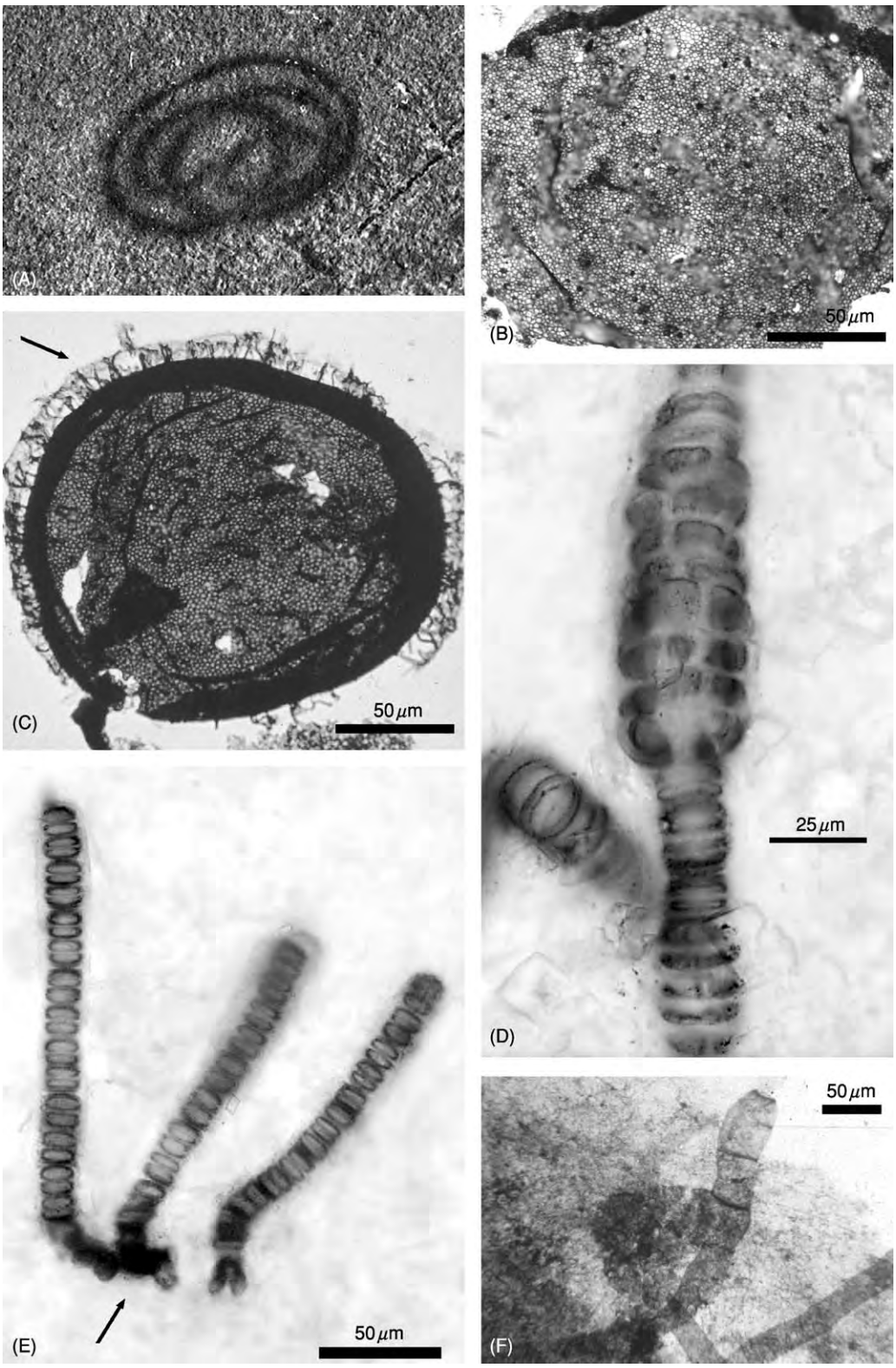


Figure 2 Palaeproterozoic (A) and Mesoproterozoic (B F) eukaryote fossils. (A) Coiled ribbons from the Negaunee Iron Formation, Michigan, USA. Specimen is about 2 cm in maximum dimension. Photograph courtesy of Bruce Runnegar. (B) *Dictyosphaera delicate* from the Ruyang Group, northern China. Notice the polygonal pattern (*ca.* 1 μm in diameter) on the vesicle surface. (C) *Shuiyou sphaeridium macroreticulatum* from the Ruyang Group, northern China. Arrow points to spines on the vesicle surface. This taxon also shows a polygonal pattern on the vesicle surface. Reproduced with permission from Xiao S, Knoll AH, Kaufman AJ, Yin L, and Zhang Y (1997) Neoproterozoic fossils in Mesoproterozoic rocks? Chemostratigraphic resolution of a biostratigraphic conundrum from the

Archaean (3800–2500 Ma) Eukaryotes

Phylogenetic relationships among the three domains (Figure 1) indicate that the eukaryotes may have a history that extends back as far as those of the other two domains. Since there are microbial fossils in the Archaean, the possibility of Archaean eukaryotes is intriguing. Indeed, we can follow the footprints of eukaryotes into the Archaean. Steranes, which are eukaryote biomarkers derived from steroids, occur abundantly in the ~2600 Ma Marra Mamba Formation and the ~2715 Ma Maddina Formation in northwestern Australia. But so far no eukaryote microfossils have been found in Archaean rocks.

The presence of steranes in Archaean rocks, however, does not necessarily imply that all of the features (mitochondria, cytoskeletons, and histone) that collectively define living eukaryotes evolved in the Archaean. It is possible that the biosynthesis of steroids appeared early in eukaryote history. These Archaean organisms may represent ancient branches that are more closely related to the eukaryotes than to the other two domains but that diverged before the last common ancestor of all living eukaryotes evolved. In other words, they may represent stem-group eukaryotes.

Palaeoproterozoic (2500–1600 Ma) Eukaryotes

Eukaryotic biomarkers continue to occur in Palaeoproterozoic and younger rocks, for example in the McArthur Group (approximately 1600–1700 Ma) in northern Australia. But it is in the Palaeoproterozoic rocks that the earliest morphological evidence for eukaryotes has been found. Such morphological evidence appears in two forms: macroscopic carbonaceous compressions and organic-walled microfossils.

Among Palaeoproterozoic carbonaceous compressions, the coiled ribbons (Figure 2A) of millimetric width and centimetric length from the *ca*. 2000 Ma Negaunee Iron Formation of Michigan are probably the most famous because of the ancient age of the formation. The Negaunee ribbons resemble the Mesoproterozoic fossil *Grypania spiralis*. Millimetre-sized discoidal to elliptical compressions resembling *Chuaria* and *Tawuia* are known from the *ca*. 1800–1900 Ma Changzhougou Formation and Chuanlinggou Formation in northern China. The slightly younger (*ca*. 1700 Ma) Tuanshanzi Formation in northern China contains millimetre- to centimetre-sized carbonaceous ribbons and blades. Although the cellular details of these fossils are not preserved, their macroscopic and stable morphologies suggest that they are probably the earliest eukaryotic fossils known so far.

The Changzhougou Formation and Chuanlinggou Formation in northern China also contain organic-walled microfossils (or acritarchs) and multicellular eukaryotes. The Chinese acritarchs are spherical vesicles with simple morphology but relatively large size (about 100 μm in diameter). They are commonly referred to the genus *Leiosphaeridia*. They are interpreted as single-celled resting cysts of ancient eukaryotes. The relatively thick and resistant vesicles of the resting cysts allowed their preservation in carbonaceous shales. The cellular details of another fossil, *Qingshania magnifica*, described from the Chuanlinggou Formation, are preserved, and the organism shows evidence of cellular differentiation – the expanded terminal cell at one end of the clavate filament was probably a reproductive cell. *Qingshania magnifica* may well be a multicellular eukaryote.

Palaeoproterozoic eukaryotic fossils do not have distinct morphological features that would allow them to be placed into extant eukaryotic clades. Like the Archean molecular fossils, the Palaeoproterozoic fossils may also reflect stem-group divergence in the early history of eukaryotes.

Mesoproterozoic (1600–1000 Ma) Eukaryotes

Morphologically simple acritarchs such as *Leiosphaeridia* continue to dominate Mesoproterozoic microfloral assemblages, and *Grypania spiralis*, *Chuaria*, and *Tawuia* are also known in Mesoproterozoic successions. However, several morphologically complex acritarchs first appeared in the Mesoproterozoic. In addition, the first crown-group eukaryotes (that is, fossils that can be phylogenetically resolved to an extant eukaryote clade) also occur in the Mesoproterozoic. This indicates that crown-group eukaryotes began to diversify in the Mesoproterozoic. By the

North China Platform. *Precambrian Research* 84: 197 220. (D, E) *Bangiomorpha pubescens*, interpreted as a bangiophyte red alga, from the Hunting Formation, arctic Canada. Arrow in (E) points to basal holdfasts. Specimen in (D) shows the transition from uniseriate to multiseriate growth, suggesting multiple ontogenetic phases and probably sexual reproduction in *Bangiomorpha pubescens*. Repro duced with permission from Butterfield NJ (2000). *Bangiomorpha pubescens* n.gen., n.sp.: Implications for the evolution of sex, multi cellularity, and the Mesoproterosoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* 26: 386 404. (F) *Palaeovaucheria clavata*, interpreted as a xanthophyte alga, from the Lakhanda Group, south eastern Siberia. Photograph courtesy of Andrew H. Knoll.

end of the Mesoproterozoic, several algal groups had already diverged.

Acanthomorphic acritarchs (or acritarchs bearing spines) and acritarchs with patterned vesicles are known in the 1400–1500 Ma Roper Group of northern Australia and the >1000 Ma Ruyang Group of northern China. The Roper and Ruyang assemblages include such acanthomorphs as *Tappania plana* and *Shuiyousphaeridium macroreticulatum* (**Figure 2C**), as well as acritarchs with polygonally patterned (*Dictyosphaera delicate* and *Shuiyousphaeridium macroreticulatum*; **Figure 2B** and **2C**) or striated (*Valeria lophostriata*) vesicles. Some of these fossils have wide geographical and long stratigraphical ranges. For example, *Valeria lophostriata* and *Tappania* occur in Mesoproterozoic and Neoproterozoic rocks in Laurentia, Australia, India, and northern China. As no prokaryotes are known to have comparable levels of morphological complexity, these acritarchs are probably eukaryotic. Some of these acritarchs (for example *Dictyosphaera delicate*) preserve an organic $\delta^{13}C$ signature that is consistent with eukaryotic photosynthetic biochemistry, but it is unclear which algal group (e.g. chlorophytes, rhodophytes, or stramenopile algae – a group that includes chrysophytes, xanthophytes, diatoms, and brown algae; **Figure 1**) they belong to.

One of the earliest eukaryotic fossils that has been confidently attributed to a modern algal group is *Bangiomorpha pubescens* (**Figures 2D** and **2E**) from the 1200 Ma Hunting Formation in arctic Canada. This is a multicellular filamentous fossil that shows evidence of holdfast differentiation and sexual reproduction. It is interpreted as a benthic bangiophyte red alga. Another phylogenetically resolved eukaryotic fossil is *Palaeovaucheria clavata* (**Figure 2F**), interpreted as a xanthophyte alga, from the upper Mesoproterozoic Lakhanda Group in south-eastern Siberia. Xanthophyte algae are members of the photosynthetic stramenopiles whose plastids were derived from a secondary endosymbiont (probably a red alga). The occurrence of *Bangiomorpha pubescens* and *Palaeovaucheria clavata* in Mesoproterozoic rocks suggests that not only must crown-group eukaryotes such as red algae have diverged but also the secondary endosymbiotic event leading to stramenopile algae must have occurred by the end of the Mesoproterozoic.

Neoproterozoic (1000–540 Ma) Eukaryotes

The Neoproterozoic era includes several major milestones in eukaryote evolution. The diversity and morphological complexity of eukaryotes increased appreciably in the Neoproterozoic, and several phylogenetically and ecologically important eukaryotic groups make their first appearance in the Neoproterozoic fossil record. These include heterotrophic protists, biomineralizing protists, and, towards the end of this era, animals. Molecular-clock estimates also indicate that land plants and fungi may have diverged in the Neoproterozoic, but so far this has not been confirmed by palaeontological evidence.

Major environmental crises occurred in the middle Neoproterozoic. Between about 720 and 600 Ma, the Earth experienced at least two global glaciations (also known as 'snowball Earth events'), during which glaciers reached the tropical oceans. More glaciations may have occurred in the Neoproterozoic, but these were not nearly as extreme. It is therefore convenient to divide the Neoproterozoic into three intervals: Early (1000–720 Ma), Middle (720–600 Ma), and Late (600–543 Ma) Neoproterozoic.

Early Neoproterozoic

A quick look at several fossiliferous units of Early Neoproterozoic age gives us a broad picture of eukaryote diversity at that time. The Early Neoproterozoic Little Dal Group (850–780 Ma) in northwestern Canada, the Chuar Group (>742 Ma) in the Grand Canyon, and the Huainan Group and Huaibei Group (*ca*. 740–900 Ma) in northern China contain some of the best-preserved carbonaceous compressions in the Early Neoproterozoic. *Chuaria* and *Tawuia* (**Figures 3A** and **3B**) are abundant in these successions. In addition, *Longfengshania stipitata* – a benthic alga with an ellipsoidal head, a stipe, and a simple holdfast – has been described from the Little Dal Group and from the Changlongshan Formation (*ca*. 800–900 Ma) in northern China. Individuals of *L. stipitata* sometimes occur in clusters (**Figure 3C**), like a bunch of inflated balloons tethered together. These carbonaceous compressions are probably multicellular eukaryotes, although *Chuaria* and *Tawuia* have been interpreted as colonial cyanobacteria by some palaeontologists.

Acritarchs, particularly those with more complex morphologies, became more diverse in the Early Neoproterozoic (**Figure 3D**). More than 20 acritarch species with complex morphologies have been described from Early Neoproterozoic successions. Except a few (e.g. *Tappania* and *Valeria lophostriata*), they are not known in the Mesoproterozoic. The Wynniatt Formation (*ca*. 800–900 Ma) of arctic Canada, the Svanberfjellet Formation (*ca*. 700–800 Ma) in Spitsbergen, and the Mirojedikha Formation in Siberia, for example, contain some of the best-preserved acritarchs in the Early Neoproterozoic. The Svanberfjellet

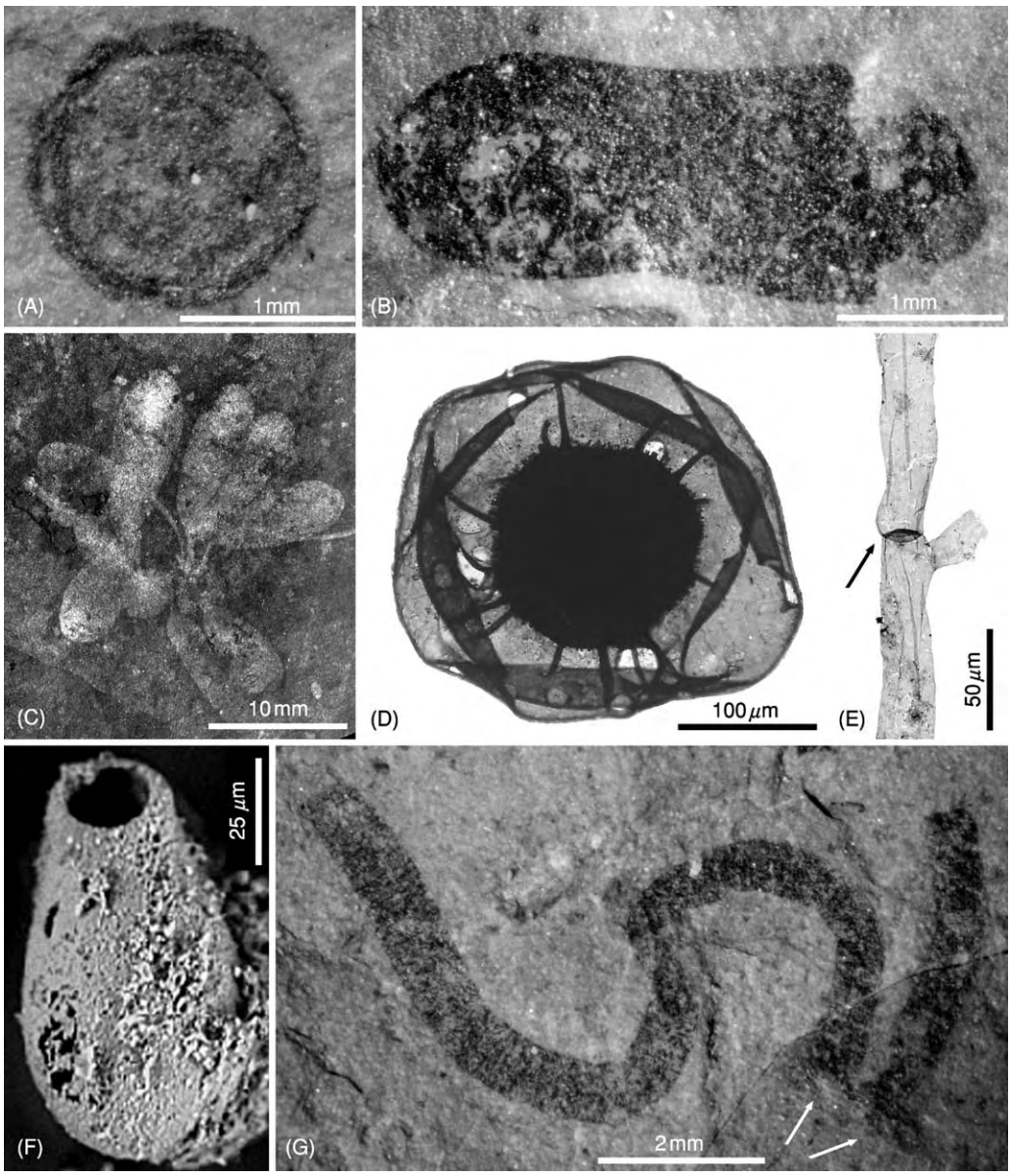


Figure 3 Early Neoproterozoic eukaryote fossils. (A) *Chuaria circularis* from the Huaibei Group, northern China. (B) *Tawuia dalensis* from the Huaibei Group, northern China. (C) *Longfengshania stipitata* from the Little Dal Group, north western Canada. Reprinted with permission from Butterfield NJ, Knoll AH, and Swett K (1994) Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. *Fossils and Strata* 34: 1 84. (D) *Trachyhystrichosphaera polaris* from the Svanberfjellet Formation, Spitsbergen. Reproduced from Hofmann HJ (1985) The mid Proterozoic Little Dal macrobiota, Mackenzie Mountains, north west Canada. *Palaeontology* 28: 331 354 by permission of Taylor & Francis AS. (E) *Proterocladus major*, interpreted as a coenocytic green alga, from the Svanberfjellet Formation, Spitsbergen. Reproduced from Butterfield NJ, Knoll AH, and Swett K (1994) Paleobiology of the Neoproterozoic Svan bergfjellet Formation, Spitsbergen. *Fossils and Strata* 34: 1 84 by permission of Taylor & Francis AS. (F) *Cyclocyrillium simplex*, a vase shaped microfossil interpreted as a testate amoeba, from the Chuar Group, Grand Canyon. Reproduced with permission from Porter SM and Knoll AH (2000) Testate amoebae in the Neoproterozoic era: evidence from vase shaped microfossils in the Chuar Group, Grand Canyon. *Paleobiology* 26: 360 385. (G) *Protoarenicola baiguashanensis*, probably a benthic alga, from the Huaibei Group, northern China. Arrows point to holdfast structures. Photograph courtesy of Xunlai Yuan.

Formation has also yielded the earliest known green algal fossil–*Proterocladus* (Figure 3E).

Biologically controlled mineralization and the formation of skeletons is an evolutionary event that significantly enhances fossil preservation. All fossils discussed in the preceding sections are non-biomineralizing eukaryotes. The earliest known biomineralizing protists are probably from the Early Neoproterozoic. Silicified scale-shaped microfossils (less than 100 μm in diameter) from the Tindir Group (age poorly

constrained, but possibly 620–780 Ma) of north-western Canada can be compared to chrysophyte skeletons. Vase-shaped microfossils (Figure 3F) from the Chuar Group and other Early Neoproterozoic successions (e.g. the Visingsö Formation in Sweden and the Draken Conglomerate Formation in Spitsbergen) have been interpreted as testate amoebae whose mineralized tests are typically preserved in casts and moulds.

The Chuar testate amoeba fossils add another dimension to our consideration of the Neoproterozoic biosphere, and that is heterotrophy. Because testate amoebae are heterotrophic protists, the Chuar vase-shaped microfossils suggest that the Early Neoproterozoic biosphere was ecologically complex. Of course, heterotrophic eukaryotes must have evolved earlier. In fact, the earliest eukaryotes may be heterotrophic, given that eukaryotic autotrophy evolved through primary and secondary endosymbiotic events. However, evidence for heterotrophy in the fossil record is scarce. Vase-shaped microfossils and ciliate biomarkers from the Chuar Group are probably the earliest known evidence for heterotrophic eukaryotes.

The best-known heterotrophic eukaryotes are perhaps the animals. There have been many reports of animal fossils from Early Neoproterozoic and Mesoproterozoic successions, but their interpretation has been controversial. *Sinosabellidites huainanensis*, *Pararenicola huaiyuanensis*, and *Protoarenicola baiguashanensis* (Figure 3G), from the Huainan Group and Huaibei Group (*ca*. 740–900 Ma) of northern China, are some of the often-cited Early Neoproterozoic animal fossils. They are carbonaceous compressions of tubes of millimetric diameter and centimetric length with transverse annulations. The transverse annulations superficially resemble animal metameric segmentation. A few specimens bear poorly defined terminal structures that have been interpreted as proboscis-like structures. However, recent study has shown that these carbonaceous compressions are probably benthic tubular algae.

Middle Neoproterozoic

The Middle Neoproterozoic is characterized by multiple global glaciations, and is unofficially labelled as the Cryogenian Period by some Precambrian geologists. On a broad scale, acritarchs and other eukaryotes suffered significant losses of diversity in the Middle Neoproterozoic. The documented diversities of several Middle Neoproterozoic assemblages are extremely low, and such assemblages are typically dominated by *Sphaerocongregus variabilis* (or *Bavlinella faveolata*). This Cryogenian drop in eukaryote diversity may be a true evolutionary pattern that was related to the glaciation events. Despite the loss of diversity, the occurrence of red algae, green algae, photosynthetic stramenopiles, and testate amoebae in Mesoproterozoic and Early Neoproterozoic rocks suggests that some members of these groups must have survived the Middle Neoproterozoic glaciations.

Late Neoproterozoic

Eukaryote diversity rose sharply in the Late Neoproterozoic. Both acritarchs and multicellular algae reached unprecedented levels of complexity and diversity in the Late Neoproterozoic. Some of the multicellular algae are preserved in anatomical detail, allowing them to be placed within the red algae.

One of the most important landmarks in Late Neoproterozoic eukaryote evolution is the emergence of animals and animal biomineralization. Molecular-clock studies suggest that the deepest (protostomes–deuterostomes) divergence within the crown-group bilaterian animals probably occurred in the Mesoproterozoic or Early Neoproterozoic. But, as discussed above, there is no convincing palaeontological evidence to support these molecular-clock estimates. Some have suggested that perhaps the earliest animals were microscopic in size and would not be well preserved in the fossil record.

The Doushantuo Formation and Dengying Formation of South China provide several taphonomic windows onto the Late Neoproterozoic biosphere. Carbonaceous shales, cherts, and phosphorites of the Doushantuo Formation (*ca*. 600–550 Ma) preserve some of the most extraordinary eukaryote fossils in the Neoproterozoic. More than 20 taxa of macroscopic carbonaceous compressions have been reported from the Yangtze Gorges area and elsewhere in South China. Most of these compressions can be unambiguously interpreted as multicellular algae (Figure 4A). Some of them show clear evidence of holdfast anchoring, dichotomous branching, apical meristematic growth, and specialized reproductive structures. A few of these compressions (for example *Calyptrina striata* and *Sinospongia typica*; Figure 4B), however, have been interpreted as sponges or cnidarians, but such interpretations are not unique and an algal interpretation cannot be falsified conclusively. In any case, it is safe to conclude that none of the Doushantuo compressions can be interpreted as macroscopic bilaterians (bilaterally symmetrical animals).

Multicellular algae also occur in the Doushantuo cherts and phosphorites. Cellular features are preserved (Figures 4C–F), so anatomical detail can be deduced from these fossils. Many of these silicified or phosphatized algal fossils show pseudoparenchymatous thallus construction, apical meristematic

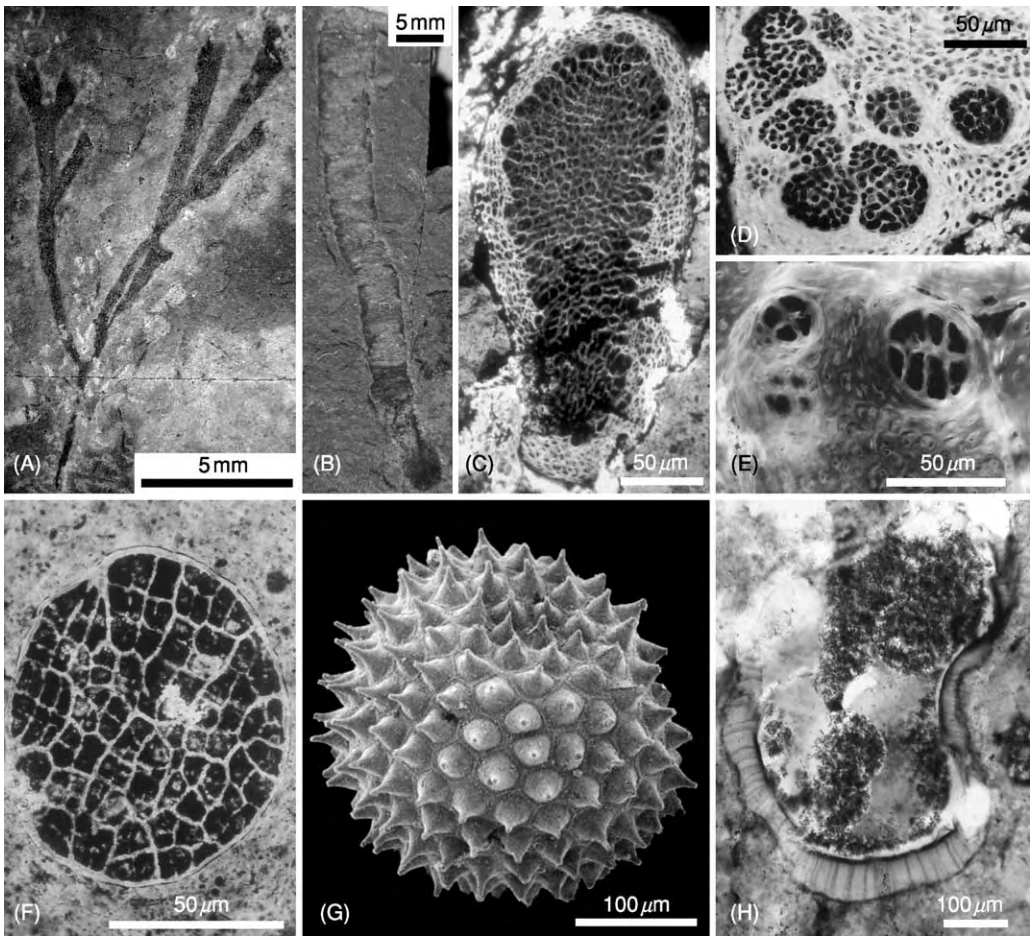


Figure 4 Late Neoproterozoic eukaryote fossils from the Doushantuo Formation, South China. (A) *Konglingiphyton erecta*, a dichot omously branching alga from the Doushantuo shales. (B) *Sinospongia typica*, possibly a benthic tubular alga from the Doushantuo shales. A, B, reproduced from Xiao S, Yuan X, Steiner M, and Knoll AH (2002) Macroscopic carbonaceous cpmpressions in a terminal Proterozoic shule: A systematic reassessment of the Miaohe biota, South China *Journal of Paleontology* 76: 345 374 with permission of The Paleontological Society. (C) A florideophyte red alga with medulla cortex thallus differentiation from the Doushantuo phosphorites. (D, E) Florideophyte red algae with possible reproductive structures from the Doushantuo phosphorites: (D) larger and darker cells arranged in clusters are interpreted as possible carposporangia; (E) tetrads and octads embedded in algal thallus are interpreted as possible tetraspores and octospores. (F) *Wengania globosa*, a possible stem group florideophyte red alga from the Doushantuo phosphorites. (G) *Meghystrichosphaeridium reticulatum*, an acanthomorphic acritarch from the Doush antuo phosphorites. (H) *Tianzhushania spinosa*, a large acanthomorph from the Doushantuo cherts. (A, C F) are courtesy of Xunlai Yuan.

growth, thallus differentiation (Figure 4C), and specialized reproductive structures (Figures 4D and 4E). Some of them have been interpreted as stem-group florideophyte red algae or stem-group coralline algae, suggesting the presence of advanced red algae in Late Neoproterozoic oceans.

In addition, the Doushantuo cherts and phosphorites contain an assemblage of large (several hundred micrometres in diameter) acanthomorphs that are morphologically complex and taxonomically diverse (Figure 4G and 4H). This assemblage includes nearly 30 species, some of which are also known to occur in other Late Neoproterozoic successions, for example the Pertatataka Formation of central Australia, the Scotia Group of Svalbard, the lower Krol Group (Krol A) of Lesser Himalaya, and lower Vendian rocks in eastern Siberia. In the Yangtze Gorges area, elements of this acritarch assemblage (for example *Tianzhushania spinosa*) first appear just metres above Cryogenian glacial deposits, suggesting that the eukaryote recovery occurred shortly after the last Cryogenian glaciations.

The Doushantuo diversification of complex acritarchs appears to be ephemeral, however. Available evidence suggests that most of the Doushantuo acritarchs disappeared when Ediacaran animals

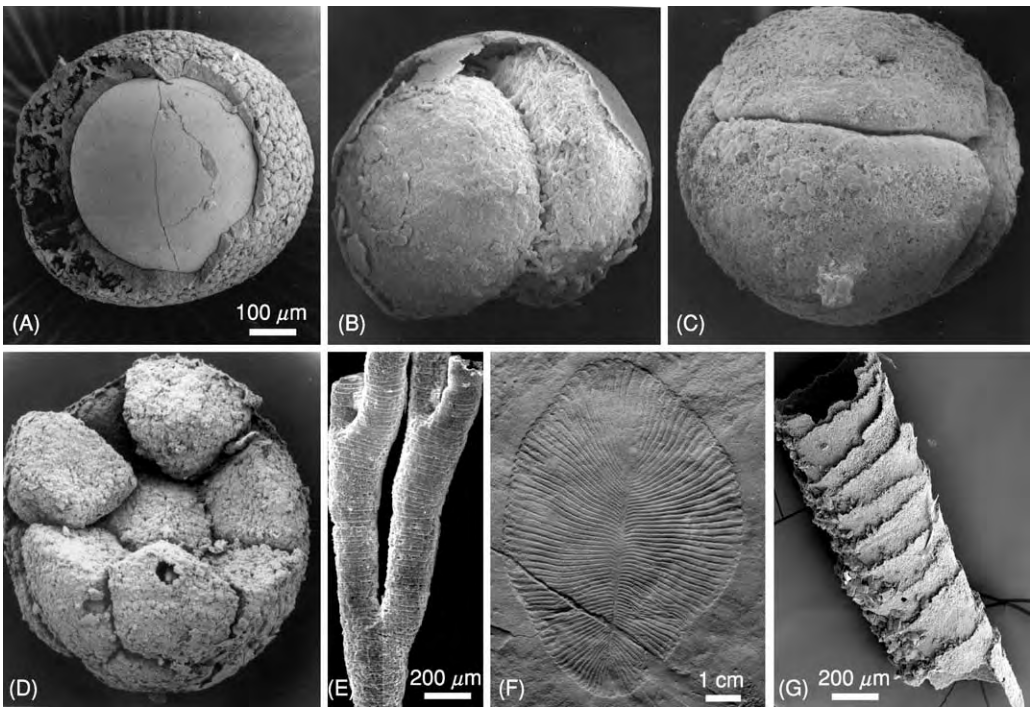


Figure 5 Late Neoproterozoic eukaryote fossils. (A D) Animal eggs and blastulas at (A) one , (B) two , (C) four , and (D) many cell cleavage stages, Doushantuo phosphorites, South China. Scale bar in (A) applies to (A D). Reproduced from Xiao S, Zhang Y, and Knoll AH (1998) Three dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. *Nature* 391:553 558 with permission from Nature Publishing Group, and from Xiao S and Knoll AH (2000) Phosphatized animal embryos from the Neoproterozoic Doushantuo Formation at Weng'an, Guizhou, South China. *Journal of Paleontology* 74:767 788 with permission from The Paleontological Society. (E) *Sinocyclocyclicus guizhouensis*, interpreted as a stem group cnidarian fossil, from the Doushantuo phosphorites, South China. (F) *Dickinsonia costata*, an Ediacaran fossil from the Flinders Ranges of South Australia. Photograph of a cast replicate preserved at the Botanical Museum, Harvard University. (G) *Cloudina riemkeae*, a biomineralized tubular animal fossil, from the Dengying Formation, South China. Photograph courtesy of Hong Hua.

began to diversify about 575 Ma ago. The disappearance is unlikely to be a preservational artefact: slightly younger phosphorites in the Dengying Formation contain no Doushantuo-type acritarchs, and basal Cambrian acritarchs are entirely different from those in the Doushantuo Formation. It is possible that the extinction of Doushantuo-type acritarchs is related to yet another small-scale glaciation in the Late Neoproterozoic.

Perhaps the most exciting fossils from the Doushantuo Formation are the submillimetric globular microfossils preserved in phosphorites at Weng'an in Guizhou Province. These globular microfossils have been interpreted as animal blastulae at successive cleavage stages, containing 2^n blastomeres within an envelope of roughly constant size (Figure 5A–D). The geometry of these microfossils is consistent with blastula cell division. It has not been determined, however, to which animal clade(s) these fossil embryos belong. Their chimeric combination of features that individually occur in crown-group sponges, cnidarians, and bilaterians suggests that these embryos may belong to stem groups at the animal, eumetazoan, or bilaterian levels. There are other microfossils in the Weng'an phosphorites that have been interpreted as adult sponges, adult cnidarians (Figure 5E), and putative bilaterian gastrulas. However, none of these are more than a few millimetres in size. Perhaps the earliest animals, at least bilaterian animals, were indeed microscopic.

Macroscopic bilaterian animals probably first evolved in the latest Neoproterozoic Ediacaran time. Among the best-known Ediacaran fossils, those in the Newfoundland assemblage (*ca*. 575–565 Ma) are probably the oldest, those in the White Sea and South Australia assemblages (*ca*. 555 Ma) are younger, and those in the Namibia assemblage (*ca*. 549–543 Ma) are the youngest. There do not appear to be any unambiguous macrobilaterians in the Newfoundland assemblage. The White Sea and South Australia assemblages, on the other hand, include body and trace fossils of macrobilaterians (e.g. *Kimberella quadrata*), as well as classic Ediacaran fossils such as *Dickinsonia* (Figure 5F) whose phylogenetic interpretations are still controversial. The Namibia assemblage and its equivalents (for

example the Dengying Formation in South China) include biomineralizing animals such as *Cloudina* (**Figure 5G**), *Namapoikia*, and *Namacalathus*. These Ediacaran assemblages, therefore, record a succession of evolutionary events leading to the rise of macrobilaterians and animal biomineralization. Most Ediacaran fossils, however, disappeared near the Precambrian–Cambrian boundary, closing the last chapter in Precambrian eukaryote evolution.

Acknowledgments

I would like to thank the National Science Foundation for funding my research on Proterozoic palaeontology. Nicholas J Butterfield, Hans J Hofmann, Hong Hua, Andrew H Knoll, Susannah M Porter, Bruce Runnegar, Leiming Yin, and Xunlai Yuan kindly provided the photographs used in this article.

See Also

Fossil Plants: Calcareous Algae. **Microfossils:** Acritarchs. **Origin of Life**. **Precambrian:** Prokaryote Fossils; Vendian and Ediacaran. **Sedimentary Rocks:** Chert. **Trace Fossils**.

Further Reading

Baldauf SL, Roger AJ, Wenk Siefert I, and Doolittle WF (2000) A kingdom level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972 977.

Brocks JJ, Logan GA, Buick R, and Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes. *Science* 285: 1033 1036.

Butterfield NJ, Knoll AH, and Swett K (1994) Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. *Fossils and Strata* 34: 1 84.

Fedonkin MA and Waggoner BM (1997) The late Precambrian fossil *Kimberella* is a mollusc like bilaterian organism. *Nature* 388: 868 871.

Hofmann HJ (1985) The mid Proterozoic Little Dal macrobiota, Mackenzie Mountains, north west Canada. *Palaeontology* 28: 331 354.

Knoll AH (1996) Archean and Proterozoic paleontology. In: Jansonius J and McGregor DC (eds.) *Palynology: Principles and Applications*, pp. 51 80. Salt Lake City: American Association of Stratigraphic Palynologists Foundation, Publishers Press.

Narbonne GM (1998) The Ediacara biota: a terminal Neoproterozoic experiment in the evolution of life. *GSA Today* 8: 1 6.

Porter SM and Knoll AH (2000) Testate amoebae in the Neoproterozoic era: evidence from vase shaped microfossils in the Chuar Group, Grand Canyon. *Paleobiology* 26: 360 385.

Schopf JW and Klein C (1992) *The Proterozoic Biosphere: A Multidisciplinary Study*. Cambridge: Cambridge University Press.

Vidal G and Moczydlowska Vidal M (1997) Biodiversity, speciation, and extinction trends of Proterozoic and Cambrian phytoplankton. *Paleobiology* 23: 230 246.

Waggoner B (2003) The Ediacaran biotas in space and time. *Integrative and Comparative Biology* 43: 104 113.

Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the USA* 87: 4576 4579.

Xiao S, Zhang Y, and Knoll AH (1998) Three dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. *Nature* 391: 553 558.

Yuan X, Xiao S, Yin L, *et al.* (2002) *Doushantuo Fossils: Life on the Eve of Animal Radiation*. Hefei: China University of Science and Technology Press.

Zhang Y, Yin L, Xiao S, and Knoll AH (1998) *Permineralized Fossils from the Terminal Proterozoic Doushantuo Formation, South China*. Memoir 50. The Paleontological Society, Lawrence, Kansas, USA.

Prokaryote Fossils

M D Brasier, University of Oxford, Oxford, UK

Introduction

Conditions for the development and survival of life on the young planet Earth are thought to have been extremely harsh. Reconstruction of those conditions requires multidisciplinary and interdisciplinary research (combining astronomy, planetary science, microbial and molecular biology, genetics, geochemistry, petrology, palaeobiology, and geology). These disciplines are united in the emerging field of astrobiology (exobiology), which covers the phenomenon of life within the solar system and beyond. Astrobiology takes a particular interest in modern and ancient prokaryotic ecosystems on Earth and in the prebiotic–biotic transition and its definition.

Meteoritic and lunar research shows that accretion of planet Earth took place around 4550 Ma ago (*see* **Earth Structure and Origins**). This accretion was soon followed by melting of the planetary surface owing to the energy released through gravitational collapse. A cataclysmic collision with a Mars-sized planet is

thought to have produced the Moon within the first 100 Ma or so. The Moon played a major role in setting the equable conditions for life on Earth through its influence in stabilizing our planetary motion about the rotational axis and by providing stable tidal and seasonal oscillations. Earth's surface had cooled and the oceans had probably condensed by about 4000 Ma, as shown by *ca*. 4200 Ma old (hydrous melt) zircons in the oldest rocks of Canada and by putative sediments in the 3800 Ma Akilia and Isua Groups of Greenland. Even so, the study of craters on the Moon and Mars shows that their surfaces were bombarded by large asteroids between 4550 Ma and 3800 Ma. Similar bombardment is likely to have affected the surface of the Earth, although any evidence has been destroyed by the mobility of the planetary surface. These bombardments would have had a devastating effect on any early biosphere (see Figure 1; the so-called 'impact frustration of the origin of life'). For example, the impact of an asteroid more than 350 km across would release enough energy to boil off the oceans, while that of

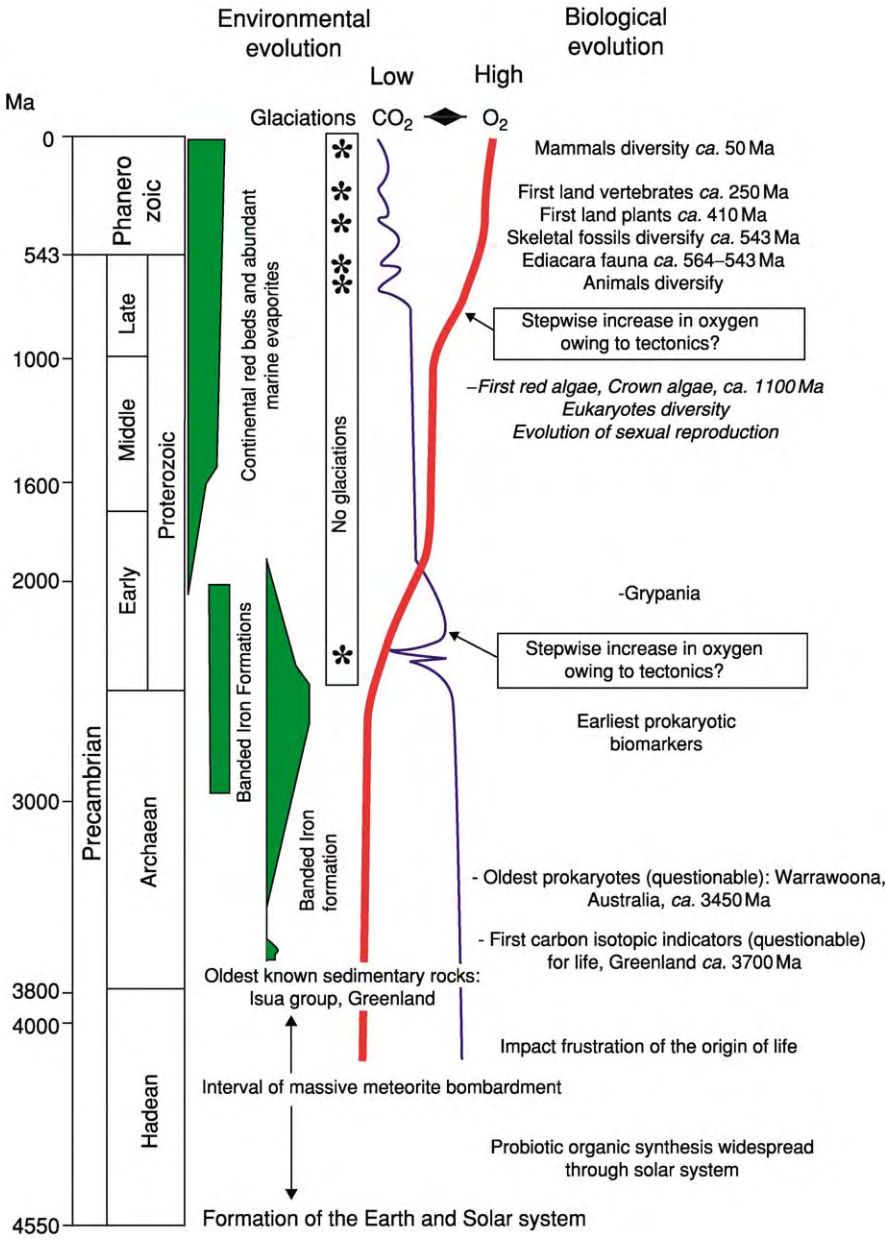


Figure 1 The main succession of events inferred for the evolution of the biosphere alongside geological evidence for changing levels of atmospheric oxygen and carbon dioxide during the Precambrian. The evidence for the first isotopic indicators (more than 3700 Ma) and the oldest prokaryotes (*ca*. 3450 Ma) is currently in dispute.

one about 150 km across could destroy the photic zone. Any life-forms synthesized prior to 3800 Ma could therefore have been repeatedly destroyed by episodic catastrophes. The only forms that could have survived such a holocaust would have been heat-tolerant hyperthermophile bacteria, which today are found living around volcanic and hydrothermal vents or deep in the Earth's crust. Such life-forms arguably originated from amino acid to RNA and DNA synthesis around hot alkaline hydrothermal vents. This is suggested not only by the antiquity of hyperthermophiles, as shown by molecular phylogeny (see below), but also by the ready availability of carbon sources (methane, carbon dioxide, carbon monoxide, aliphatics) and electron donors (metals, hydrogen sulphide, hydrogen) around such vents today, together with a plentiful supply of the phosphorus and transition metals that are needed to synthesize nucleic acids and enzymes. Against this, it has been argued that hydrothermal temperatures are too high to allow the necessary stability of complex organic molecules. But, on the early Earth, such hydrothermal vents were rich in metals (most of the world's metal resources come from the Archaean to Palaeoproterozoic crust) and clay minerals. These metals and clay minerals could arguably have allowed the stabilization of nucleic acids, even at relatively high temperatures. No other site on Earth could supply these essential building blocks in so many forms and so readily.

Asteroid bombardment of the Earth is thought to have declined progressively between about 3800 Ma and 3000 Ma. The hot, thick, and largely oceanic crust of this age can be envisaged as mantled by oceanic waters of moderate depth with a mean temperature of approximately 60°C. This means that aquatic habitats on Earth prior to about 3000 Ma may have been essentially hydrothermal in nature. It was not until about 3000 Ma that the wide shallow continental shelf seas, deep-sea troughs, and land areas began to develop. Crustal sources of free energy may therefore have begun to dwindle after about 3000 Ma, and the emerging biosphere may have been forced to switch to using solar energy, with one group perhaps converting from a heat-seeking catalyst to a light-seeking (photosynthetic) one.

Prokaryotes are considered to be the most primitive organisms on Earth and are assumed to have emerged early in the comparatively hot world of the Archaean. The cells of still-living forms are extremely small, generally less than 1 μm in diameter, and they lack a nucleus (hence the name prokaryote, which is derived from the Greek *pro* = before and *karyos* = nucleus). They may be single or colonial, the latter enclosed within a mucilaginous sheath called a capsule. Some living bacterial cells bear a whip-like thread (flagellum), and a few contain chlorophyll pigments for photosynthesis. Prokaryotes are important today in the formation of microbial sediments (*see* **Biosediments and Biofilms**), such as cyanobacterial mats and stromatolites, iron and manganese ores, carbonate concretions, and sulphide and sulphate minerals. They also yield important information about the early evolution of the cell and the histories of methanogenesis, photosynthesis, and biogeochemical cycles.

Molecular and Biochemical Evidence

Contrasts between the ribosomal RNA sequences of diverse bacteria, protists, fungi, plants, and animals indicate that life on Earth can be divided into three primary domains: the Archaea (or 'Archaebacteria'), which includes the methanogenic and sulphur bacteria; the Bacteria (or 'Eubacteria'), which includes cyanobacteria and other forms with photosynthetic pigments; and the Eucarya (or 'Eukaryota'), which includes all the protists, fungi, plants, and animals (Figure 2). It is notable that all the most deeply rooted branches of the tree of life are occupied by modern hyperthermophilic bacteria, which live at temperatures of 80–110°C and are seldom able to grow below 60°C. This has been taken to suggest that the

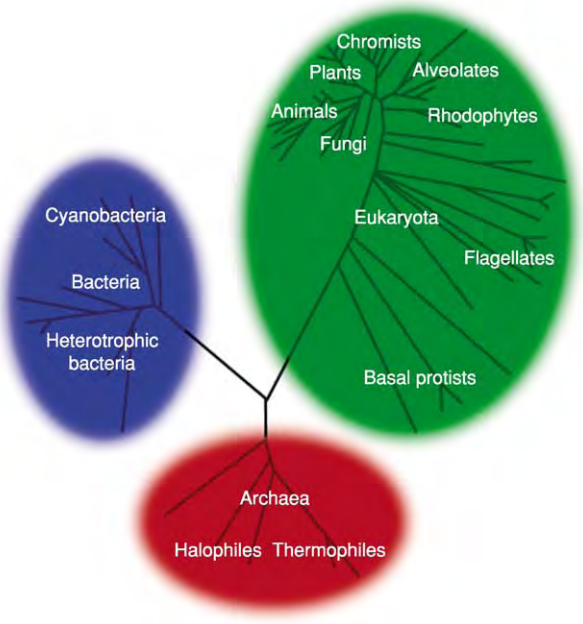


Figure 2 The three branches of the tree of life, in which all the deep seated branches are hyperthermophilic bacteria (shown in bold). Time increases along the branches, but not necessarily in a linear fashion or at the same rate in each branch. Longer branches relate to faster evolution.

last common ancestor of all living organisms was a hyperthermophile, adapted to hot hydrothermal springs or to life deep in the Earth's crust. One explanation for this is that life originated in such conditions. Another explanation is that hyperthermophiles were pre-adapted to survive the catastrophic period of meteorite bombardment between 4500 Ma and 3800 Ma.

Biochemical evidence can be taken to suggest the following evolutionary sequence of autotrophic prokaryotes, each of which used carbon dioxide as their sole source of carbon.

1. Anaerobic chemolithotrophic prokaryotes, which mainly use hydrogen produced from inorganic reactions between rock and water as their main electron source.
2. Anaerobic anoxygenic prokaryotes such as green and purple sulphur bacteria, which use photosynthesis to reduce carbon dioxide to form organic matter, with hydrogen sulphide as the electron source, in the absence of oxygen.
3. Oxygenic cyanobacteria, which use photosynthesis to reduce carbon dioxide to form organic matter, with water as the electron source, releasing oxygen. These must have had an enormous impact on Earth surface processes and the biosphere, and considerable interest has been focused by astrobiologists upon their first appearance in the rock record. At the time of NASA's Viking missions to Mars in 1976, it was such photosynthetic autotrophy that scientists were hoping to find.

Heterotrophic prokaryotes do not synthesize organic matter. Like us, they use preformed organic matter as their source of carbon and can use a range of oxidants to break it down and release the energy bonds. Methanogenic Archaea are among the most primitive heterotrophs alive today, living in highly reducing sediments (such as peat bogs) and releasing methane gas. Sulphate-reducing bacteria use seawater sulphate (SO_4) ions in the absence of oxygen, but require a highly oxidized form of sulphur (SO_4), which may not have been widely available in the early ocean. Aerobic heterotrophic bacteria use freely available atmospheric oxygen and are unlikely to have radiated before the so-called Great Oxygenation Event, 2450–2200 Ma ago, when various indicators of the weakly reducing planetary surface (banded iron formations (*see* **Sedimentary Rocks:** Banded Iron Formations), detrital pyrite, uraninite, and siderite) begin to disappear from the rock record and red beds start to appear. This oxygenation event may relate in part to increasing rates of carbon burial in expanding cratonic basins and subduction zones and in part to the irreversible loss of hydrogen to space from the upper atmosphere. While oxygen producers and consumers could have existed prior to 2450 Ma, they were probably restricted to rather local oases of oxygenation. This inferred evolutionary sequence of methanogenic to sulphate-reducing to aerobic heterotrophic prokaryotes is likely, on the basis of evidence from living bacteria, to have been accompanied by an increasing yield of energy from the same amount of carbonaceous 'food'. Significantly, this evolutionary succession closely resembles the modern distribution of prokaryotic populations within marine muds, with methanogenic Archaea lying deep within the sediment pile, aerobic heterotrophs and photoautotrophs in the upper layers of the sediment, and sulphate reducers in between.

Evidence for the Earliest Biosphere

Biogeochemistry

The fossil evidence for life on Earth gets increasingly scarce as the age of rock units increases. This is because older rocks have suffered more exposure to erosion and have experienced a greater degree of alteration by metamorphism. Hence, the oldest rocks on Earth (approximately 3800–3700 Ma), from Isua and Akilia in Greenland, have been too heavily metamorphosed to yield morphological evidence of life. Possible traces of life must therefore be explored using biogeochemical techniques. Stable isotopes of carbon from Isua and Akilia, for example, are somewhat lighter than usually expected from an inert world (*ca.* −18‰ δ^{13} CPDB cf. Pee Dee Belemnite standard). This has been taken to imply that life was able to self-organize and survive the period of catastrophic meteorite impacts before about 3800 Ma (**Figure 1**). Such a view is now controversial for a variety of reasons. The sedimentary origin of the carbonaceous grains is questionable: fractionation of carbon compounds could also have resulted from abiogenic processes or even from carbonaceous meteoritic debris. The carbon may also be younger than claimed. While light carbon isotopes (*ca.* −40‰ to −25‰ δ^{13} PDB) are commonly encountered in rocks younger than about 3500 Ma, some of these hydrocarbon compounds may also have an abiogenic origin, and precise discriminators, such as C–H ratios, H isotopes, and aliphaticity, are needed to discount this possibility. Even so, most scientists assume that this 25–40‰ difference in carbon isotopes of carbonates and organic matter seen after about 3500 Ma provides key evidence that biological metabolic pathways (i.e. autotrophic fractionation of carbon isotopes) were in place by this time.

A trend towards highly negative carbon isotope ratios about 2800–2200 Ma ago has been explained as the result of a bloom in heterotrophic methanogenic and methanotrophic Archaea at that time, prior to the Great Oxidation Event. Sulphur isotopes can also be used to trace the history of prokaryotic metabolism. In this case, ^{32}S is preferentially taken up by sulphate-reducing bacteria, leaving the water column enriched in the heavier isotope, ^{34}S. Recent studies of the ratio between ^{32}S and ^{34}S in sedimentary pyrite and in supersaturated barite, gypsum, and anhydrite have been used to argue that bacterial sulphate reduction was not in place before about 2450 Ma. Prior to this, sulphur isotope fractionation seems to have been of the abiogenic mass-independent kind, more like that seen in the Martian atmosphere. This lack of bacterial sulphate reduction before about 2450 Ma may be explained by a lack of sufficient atmospheric oxygen to form the sulphate ions and/or by surface water temperatures that were too high in the Archaean to produce a measurable fractionation.

Organic geochemical 'biomarkers' can also be sniffed out in well-preserved rocks, at least as far back as 2700 Ma. Biomarkers called 2-methylhopanes have been reported from rocks of this age in Western Australia and have been taken to indicate the presence of cyanobacteria at that time. This line of uniformitarian reasoning assumes, of course, that such biomarkers were not present in any other prokaryotic group, living or extinct, which is rather difficult to refute. Unfortunately, most rocks older than 2700 Ma appear to be too 'cooked' to preserve complex organic molecules.

Stromatolites

Laminated domical structures known as 'stromatolites' (Figure 3A) have been described from carbonate rocks as old as 3450 Ma in the Pilbara Supergroup of Western Australia (Figure 3C) and in the coeval Swaziland Supergroup of South Africa. Although an origin from the accretion of prokaryotic and even cyanobacterial mats has often been inferred for these early Archaean examples, they do not contain microfossils and they show some features that render their biogenicity rather questionable: an association with, and continuation down into, hydrothermal dyke systems; close association with epigenetic crystal fans (e.g. after aragonite, barite, gypsum) and directly precipitated carbonates; intergradation with ripple-like forms having rotational symmetry (i.e. they look the same upside down); lack of multiple

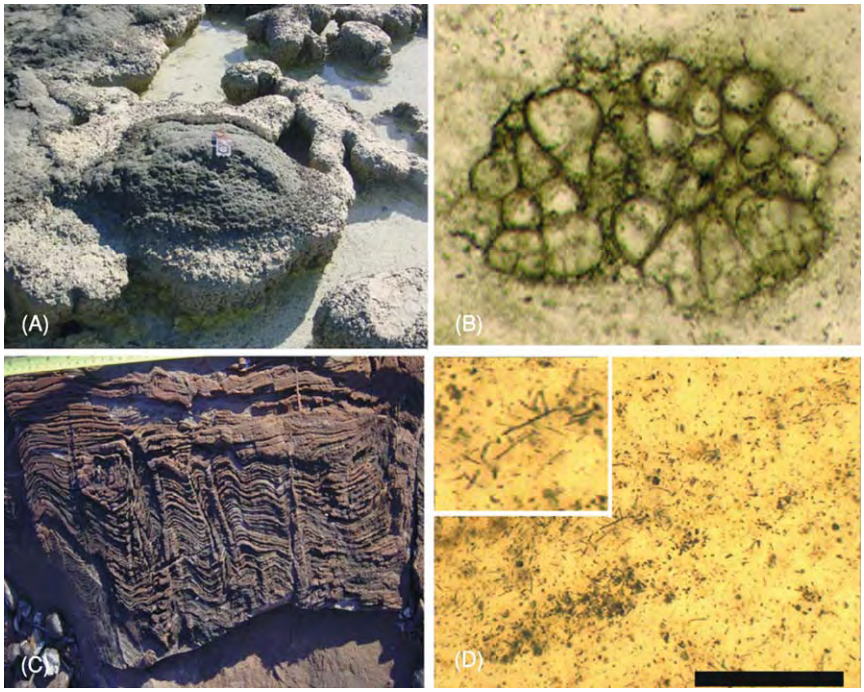


Figure 3 Stromatolites and Precambrian prokaryotes. (A) Living cyanobacterial diatom stromatolites from the hypersaline Shark Bay in Western Australia; compass is *ca.* 6.5 cm wide. (B) The coccoid cyanobacterium *Eoentophysalis* preserved in cherts of the *ca.* 800 Ma old Boorthanna Chert of Australia; scale bar below 100 μm. (C) A 30 cm wide exposure of 3400 Ma old stromatolites from the Strelley Pool Chert of Western Australia; it is unclear whether these are of microbial origin or are an abiogenic crystalline precipitate. (D) The 1900 Ma old filaments of cyanobacterial or iron bacterial origin preserved in the Gunflint Chert of Mink Mountain, Ontario, Canada; scale bar (main picture) 40 μm; scale bar (inset) 100 μm.

fractal dimensions; continuous laminae of barely varying thickness (typical of crystal growth); lack of wrinkle fabrics typical of biofilms; and lack of fenestrae or other signs of biological processing. Such simple stromatolites seem likely to have formed from largely physical processes such as direct chemical precipitation from seawater. It is only after about 3000 Ma that fractally complex stromatolites are found in carbonate platform settings, for which the presence of biofilms seems more plausible, if not always demonstrable. These tend to have wrinkle mat fabrics, discontinuous laminae, and laminar fenestrae (e.g. those from the 2700 Ma old Belingwe Group of Zimbabwe and the Fortescue Group of Western Australia). Microfossils associated with Archaean stromatolites remain questionable, however, until about 2530 Ma and are not really diverse until the 1900 Ma old Gunflint Chert of Canada (**Figure 3D**). Even here, the biogenicity of the associated stromatolites is open to question because they resemble abiogenic sinters.

Silicified Microbiotas

Precambrian oceans appear to have been supersaturated with silica (SiO_2) because it was not being removed from the water column by groups that evolved later, such as the diatoms, radiolaria, and sponges. In environments where early diagenetic silica was able to engulf prokaryotic populations, such as in peritidal hypersaline bacterial-mat settings, prokaryotic sheaths and even cell walls were sometimes (but still very rarely) well preserved in three dimensions. This prevalence of hypersaline settings may have produced an unfortunate bias towards prokaryotic assemblages in the early fossil record. In other words, the inferred dominance of prokaryotic microfossils within Precambrian cherts may be due to their restricted hypersaline setting rather than to evolutionary factors.

Silicified microbiotas are usually studied by means of standard (30 μm) to thick (300 μm) petrographic thin sections, at magnifications of up to about $\times$400. This thin-section technique is paramount because of the way in which it provides for contextual analysis, including three-dimensional morphology, mineralogy, rock fabric, and rock history. Other techniques, such as maceration (digestion in strong acids), etching of rock chips, scanning electron microscopy, and atomic force microscopy, are also used, but these do not provide the requisite information on context and are prone to the inclusion of structures that are later contaminants or 'artefacts' of the preparation process.

A classic example of a silicified Precambrian microbiota is the 1900 Ma old Gunflint Chert, which preserves about 12 taxa of prokaryotes, including forms that superficially resemble coccoid and filamentous cyanobacteria (**Figure 3D** and insert) but may be more closely allied to extant iron bacteria. The putative cyanobacteria *Eoentophysalis* (**Figure 3B**) and *Archaeoellipsoides* are thought to be present in the approximately 2100 Ma old cherts of West Africa and the 2000 Ma old Belcher Group cherts of Canada, respectively. The latter is claimed to preserve the specialist heterocyst cells used by cyanobacteria to help fix nitrogen in an otherwise oxidizing atmosphere. As one moves back into the Archaean, microfossils become both extremely rare and highly questionable, despite the great abundance of carbonaceous cherts and tufa-like carbonates. This may be explained partly by the inference that Archaean cherts were laid down in largely hydrothermal conditions that were often acidic and reducing, and partly by the scarcity of large and resistant cellular materials at that time. Bundles of silicified filaments and tiny calcified holes from stromatolites in the 2530 Ma old carbonates of the Transvaal Supergroup of South Africa may be the casts of coccoid and filamentous cyanobacteria but little of diagnostic significance is preserved. Intriguingly, such encrusted cyanobacterial filaments are rarely seen before about 1000 Ma, and endolithic microborings are not reported prior to 1500 Ma. A single microfossil-like structure from the 2700 Ma old Fortescue Group of Western Australia has been compared with a cyanobacterial filament but its biogenicity and context awaits full documentation.

Skeins of pyritic filaments found within carbonaceous cherts from 3200 Ma old black smokers of the Sulphur Springs Formation in Western Australia may be the remains of anaerobic hyperthermophile bacteria, though the indigenous and biogenic nature of these intriguing structures has yet to be demonstrated beyond question. The oldest cherts containing a supposed diverse prokaryotic microflora (**Figure 4A**) come from the 3450 Ma old Apex Cherts, which are intimately associated with ultrabasic and basaltic lava flows of the Apex Basalt in the Warrawoona Group of Western Australia. At least eleven different kinds of filamentous microfossil have been described from these rocks, some of which have been compared with cyanobacteria (**Figure 4A**). This has been taken to suggest that photosynthesis had begun to release oxygen into the atmosphere by 3450 Ma and that a substantial amount of evolution had taken place by this time. A critical re-examination of the context and fabric of these cherts suggests, however, that all these

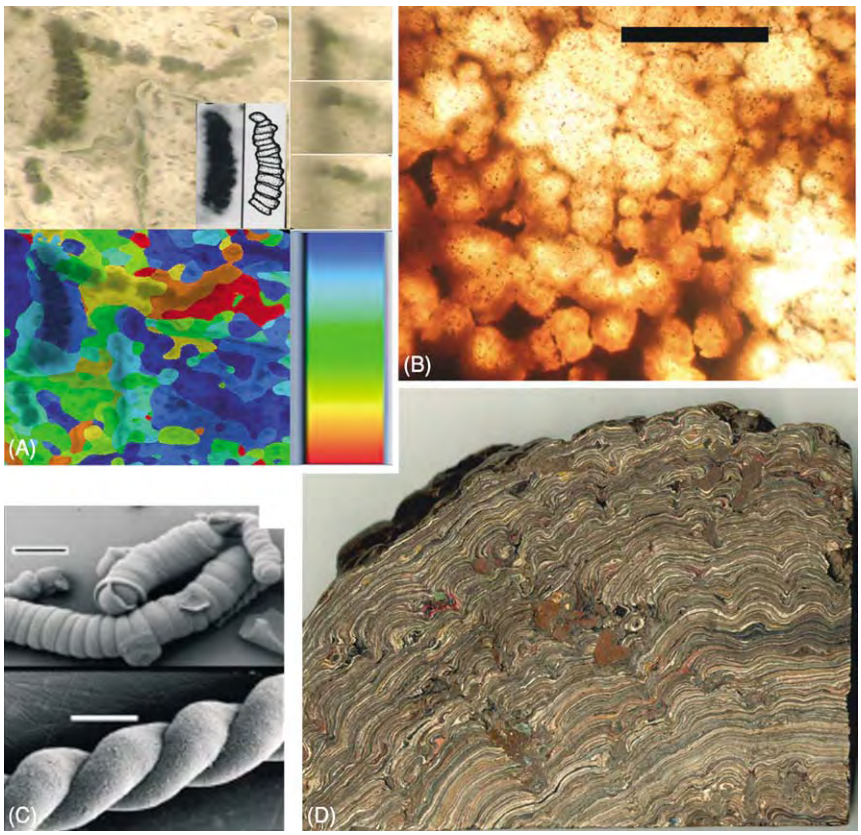


Figure 4 Pseudofossils. (A) Montage of the pseudofossil *Archaeoscillatoriopsis disciformis* holotype from the 3450 Ma old Apex Chert of Western Australia, showing (in black and white) its original interpretation and (in colour) its much more complex morphology when montaged (with topographic map) and its proximity to crystal growths. (B) Spherulites formed by the recrystallization of glassy hydrothermal silica with iron oxide impurities; recrystallization causes the formation of a spectrum of forms that range from abiogenic to microfossil like pseudofossils; scale bar in B 40 μm. (C) Complex carbonaceous artefacts formed in the laboratory; scale bars 40 μm. (D) Complex stromatolitic growths can easily be formed artificially by over spraying a surface with layers of paint; width of view 12 cm.

structures occur deep within a cross-cutting hydrothermal vein. They comprise a continuous spectrum of arcuate to complexly branched carbonaceous reaction rims, formed around the edges of spherulitic silica botryoids, that appear to have formed during the recrystallization of hydrothermal silica glass to chalcedony and around rhombic crystal margins. Doubt is also cast on other 'microfossils' of about this age from the Warrawoona Group of Australia and from the Barberton mountains of South Africa. Their origin as mineralic pseudofossils or contaminants remains an open question.

Criteria for Biogenicity

The search for the signals left in rocks by very ancient prokaryotes requires a checklist of criteria to help establish biogenicity, not least because prokaryotic cells are so easily mimicked by physical artefacts and so readily introduced by contamination. A previous formula for the recognition of *bona fide* Archean microfossils has been that the putative microfossil structures must be unquestionably biogenic remnants that are indigenous to and syngenetic with the primary deposition of a sedimentary rock, which is itself of known stratigraphic and geographical source and of established Archaean age (i.e. greater than 2500 Ma old).

Until recently, the above criteria have been used to contrast the questionable 'microfossils' in Martian meteorite ALH 84001 with the 3450 Ma old supposedly *bona fide* 'microfossils' from the Apex Chert. It is now necessary to acknowledge, however, that complex structures of organic composition, like those in the Apex Chert, can be assembled very easily (e.g. Figure 4C), including abiogenic artefacts of

isotopically light carbon formed through late-stage remobilization of hydrocarbons with hydrothermal silica. Complexity *per se* is not a valid criterion for the recognition of life. Biological populations may be distinguished from abiogenic ones by the *lack* of a continuous spectrum of morphological variation, limited by the DNA-controlled genetic constraints. This narrow range can be contrasted with complex abiogenic artefacts, which will tend towards a 'symmetry-breaking cascade' of forms, ranging from a few biological look-alikes with rotational symmetry (**Figure 4C**) to highly information-rich forms (**Figures 4A** and **4B**) with poor symmetry.

Research into the earliest biosphere is now served by a marvellous range of new tools, including molecular (RNA and DNA) analysis of living-prokaryote phylogeny, biogeochemical analysis of ancient Earth rocks, and digital-image analysis of putative microfossils. These techniques should not be allowed to prosper, however, at the expense of more traditional techniques such as mapping and thin-section petrography. The latter are also vital for providing a context for early-life studies at a range of scales, from satellite images of greenstone belts, through field mapping, to microfabric mapping of thin sections. Future research will also require a much better understanding of the nature of complexity itself and of the multitudinous ways in which seemingly complex structures can self-organize. (**Figure 4** shows a range of microfossil-like and stromatolite-like objects of inferred non-biological origin.) This better understanding of complexity is needed, not only to understand the ways in which pseudofossils and pseudo-stromatolites can form, but also to comprehend the conditions in which life-forms and their building blocks were able to assemble and develop on the early Earth, and perhaps beyond.

See Also

Biosediments and Biofilms. **Earth Structure and Origins**. **Origin of Life**. **Precambrian:** Eukaryote Fossils. **Pseudofossils**. **Sedimentary Rocks:** Banded Iron Formations. **Solar System:** Mars.

Further Reading

Brasier MD, Green OR, Jephcoat AP, *et al.* (2002) Questioning the evidence for Earth's oldest fossils. *Nature* 416: 76 81.

Farquhar J, Bao HM, and Thiemens M (2000) Atmospheric influence of Earth's earliest sulfur cycle. *Science* 289: 756 768.

Furnes HR, Banerjee NR, Muehlenbachs K, *et al.* (2004) Early life recorded in Archean pillow lavas. *Science* 304: 578 581.

Garcia Ruez JM, Hyde ST, Carnerup AM, *et al.* (2003) Self assembled silica carbonate structures and detection of ancient microfossils. *Science* 302: 1194 1197.

Hofmann HJ, Grey K, Hickman AH, and Thorpe RI (1999) Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia. *Bulletin of the Geological Society of America* 111: 1256 1262.

Kazmierczak J and Altermann W (2002) Neoarchean biomineralization by benthic cyanobacteria. *Science* 298: 2351.

Kerr R (2004) New biomarker proposed for earliest life on earth. *Science* 304: 503.

Knoll AH (2003) *Life on a Young Planet*. Princeton, NJ: Princeton University Press.

Mojsis SJ, Arrenhius G, McKeegan KD, *et al.* (1996) Evidence for life on Earth before 3,800 million years ago. *Nature* 384: 55 59.

Rasmussen B (2000) Filamentous microfossils in a 3,250 million year old volcanogenic massive sulphide. *Nature* 405: 676 679.

Rothman DH and Grotzinger JP (1996) An abiotic model for stromatolite morphogenesis. *Nature* 383: 423 425.

Schopf JW (1999) *The Cradle of Life*. New York: Princeton University Press.

Simpson S (2003) Questioning the oldest signs of life. *Scientific American* April 2003: 70 77.

Westall FM, de Wit MJ, Dann J, *et al.* (2001) Early Archean fossil bacteria and biofilms in hydrothermally influenced sediments from the Barberton greenstone belt, South Africa. *Precambrian Research* 106: 93 116.

Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: proposals for the domains of Archaea, Bacteria and Eucarya. *Proceedings of the National Academy of Sciences USA* 87: 4576 4579.

Vendian and Ediacaran

M A S McMenamin, Mount Holyoke College, South Hadley, MA, USA

Introduction

The geological, palaeobiological, and earth systems (i.e., global geochemical) events associated with the Vendian period set the stage for all subsequent Earth history. Unprecedented change took place in the biosphere at this time, and the Vendian witnessed the appearance of the first Ediacarans, the first shelly fossils, and the earliest known animals. The Vendian saw the continuation of breakup of the earliest known giant supercontinent (Rodinia) and, at the end of the period, continuing into the Cambrian, the amalgamation of the supercontinent Gondwana. The Vendian also witnessed the termination of the worst series of glaciations known. The climate change was accompanied by dramatic fluctuations in the records of carbon and strontium isotopes. Linkages between supercontinent breakup and extreme climate change, and between biotic diversification and ecological change (e.g., as driven by metazoan disturbance of the marine substrate), are being actively explored by researchers following diverse avenues of investigation.

The Vendian period and system (called by some authors Lipalian or Ediacaran) spans the interval from 600–543 Ma. It was the final period of Proterozoic time (Neoproterozoic). Some accounts place the beginning of the Vendian as far back as 670–620 Ma, but this would push it into what should be the preceding period (the Sinian, 680–600 Ma). In any case, the Vendian represents the last interval of Precambrian time, and its end marks the most important division in the geological time-scale, the Proterozoic–Cambrian boundary. The importance of this boundary has led to escalating interest amongst researchers, and great advances have recently been made in understanding the Vendian. Many questions remain, however, and amongst them are some of the most contentious issues in contemporary earth science. The primary questions of interest are threefold. First, what caused global climate to warm after the worst glaciation on record? Second, what are the Ediacarans? Third, what events triggered the so-called Cambrian Explosion (*see* **Palaeozoic:** Cambrian) and the appearance of familiar animal types? A host of unsolved secondary questions follow. These include: What was the makeup of supercontinent Rodinia? What was the timing sequence and geometry of its breakup? What were the controls on eustatic sea-level change during the Vendian? Was there a mass extinction at the Proterozoic–Cambrian boundary? If so, what was its cause? What was happening to global geochemistry during this time? The implications of these questions are far reaching, and the discussion that follows is divided into three sections: Geological Events, Palaeobiological Events, and Earth System Events.

Geological Events

It can be said that the prelude to the Vendian world began at 1000 Ma with the amalgamation of the supercontinent Rodinia. The continental collisions that led to the formation of the supercontinent are generally called the Grenville Orogenic Event (*see* **Grenvillian Orogeny**). Named for 1000 Ma rocks that record an ancient episode of mountain building and continental collision in North America, Grenvillian rocks have now been recognized in places as distant as northwest India and Antarctica.

Composed of all the large continents of the Precambrian planet, Rodinia began to split apart beginning at about 750 Ma. Some geologists argue for a failed attempt at breakup at 850 Ma. Rodinia is thus a remarkably long-lived supercontinent, lasting some 250 million years. The rifting event was a drawn out affair that seems to have proceeded in two stages. Stage one began with the opening of the Pacific Ocean, as parts of what now constitute eastern Gondwana (Australia, Antarctica) split apart from what is now the west coast of North America. In stage two, fragments connected to the north-eastern and eastern parts of North America split off and began to collide with the continental blocks already set free by continental drift. The net result of all this continental motion, which was not complete until about the time of the Cambrian boundary, was the reorganization of the continental plate geometry of the planet with the sundering of Rodinia and the amalgamation of Gondwana. Some geologists believe that a short-lived global supercontinent (Pannotia) existed between the times of Rodinia and Gondwana, but evidence for this supercontinent is tenuous at best.

Block faulting and volcanicity associated with tectonic sundering characterize many Vendian stratigraphical sections. Deposition of the Vendian Tindir Group of east-central Alaska was influenced by block faulting in North American basement rocks. The Vendian La Ciénega Formation of northern

Mexico is punctuated by layers of porphyritic basalt erupted from fissures presumably related to tectonic extension.

Palaeomagnetic evidence suggests that the rates of continental motion near the Cambrian boundary were amongst the fastest ever measured. At one point near the boundary, the continents appear to have been moving synchronously, prompting what has been called inertial interchange theory, in which the Earth's crust is thought to have detached from the mantle for a relatively short time.

As the fragments of Rodinia went their separate ways, they set into motion a sequence of events that ultimately resulted in the drowning by seawater of all or most continental margins. As Rodinia split apart, the Rodinia's counterpart superocean, Mirovia, began to vanish beneath the subduction zones along the leading edges of the dispersing continental fragments. This led, overall, to an exchange of old, cold Mirovian seafloor crust for the relatively hot, less dense, and hence more buoyant ocean crust of the newly developing rift basins on the trailing edges of the continents opposite the subduction zones. Seawater was consequently displaced on a massive scale by buoyant, new mid-ocean ridges, and this led to the great marine transgression that began the Palaeozoic era.

The transgression began very slowly, as the early stage of breakup involved only a single (albeit long) continental margin, namely the split between Australia–Antarctica and North America. This gradual rise was then suddenly punctuated by a rapid fall in eustatic sea-level due to the onset of the first of several (Sturtian and Marinoan) Late Proterozoic glaciations. These were the worst glaciations known, leading many to believe that the Earth must have passed through what has been called 'White Earth' or 'Snowball Earth' conditions. The timing of these glaciations is somewhat uncertain, although two groups, an older (Sturtian, including the Rapitan, Chuos, and Stuartian glaciations) and a younger (Marinoan, including the Ice Brook, Ghaub, and Elatina glaciations), appear to each consist of roughly synchronous glacial events.

Compelling evidence exists during both glacial episodes for glaciers on land at sea-level at the equator, and some geologists suggest that the ice cap extended into the equatorial ocean as well. Regardless of the true extent of the ice cap, the glaciation was tremendously severe. The climatic aberration is made all the more curious by the character of the sedimentary layers deposited directly above the glacial deposits. Also, sedimentary (banded) iron formations reappeared during these ice ages after having virtually disappeared from the record for 1.5 Ga. These and other geochemical anomalies are interpreted to suggest almost incredible changes in the oxygen and carbon dioxide cycles of the planet – anoxia under the marine ice cap, leading to banded iron deposition; build-up of sufficient carbon dioxide in a life-depleted Earth to eventually trigger greenhouse conditions and melt the glaciers; and supergreenhouse conditions and supersaturation, leading to massive deposition of cap carbonates under unusual carbonate depositional conditions.

The cap carbonates consist primarily of abiogenically precipitated calcite and dolomite. Carbonate rocks of this nature are usually associated with deposition under very warm climatic conditions. Therefore, the juxtaposition of these sediments directly above deposits of the worst known glaciation is unusual in the extreme. The last Proterozoic cap carbonates were deposited at approximately 600 Ma and define the base of the Vendian system.

With the final melting of the ice, the transgression resumed its flooding of continental shelf areas. The transgression continued essentially unabated until well into the Cambrian. In many stratigraphical sections throughout the world, the base of the Cambrian is marked by a basal unconformity, although this is by no means universal, particularly in regions with a more or less complete Vendian section as well. The Vendian is thus defined in many places by a cap carbonate at its base (a pronounced aid to lithostratigraphical correlation when not associated with an unconformity), but an unconformity at its top (a decided hindrance to correlation efforts). Nevertheless, a combination of radiometric dates, lithostratigraphy, biostratigraphy, palaeomagnetic stratigraphy, and carbon and strontium isotope stratigraphy has rendered preliminary correlations possible between Vendian sections on different continents. Much work remains to be done to refine these correlations.

The Vendian is not characterized by major extraterrestrial impact events, but this may be an artefact of a less well-understood record of mass extinction during the period. The Acraman impact site (570 Ma) of South Australia, associated with a 160 km diameter crater, shock metamorphism, and shattercone development, is comparable in size to the 214 Ma, Manicouagan impact structure of Canada.

Palaeobiological Events

Although eukaryotic organisms are thought to have existed for more than 1000 million years before the beginning of the Vendian, they are rather inconspicuous until the Vendian begins. Thus, the Vendian marks the beginning of the Phanerozoic, the age of

visible life. The reason for the rapid expansion of eukaryotic life at this time is unclear, but some researchers have linked it to the climatic amelioration following the Proterozoic glaciations.

Biogenic stromatolites, known from the oldest rocks bearing microbial fossils, dramatically changed texture during the Vendian. The concentric lamination that characterized more ancient stromatolites gave way to a clotted, thrombolitic texture. This transition from stromatolite to thrombolite has been attributed to the burrowing and lamination-disturbing activities of early animals.

Biological inferences regarding the origin of metazoa, based on molecular clock data, indicate that metazoa appeared by at least 1000 Ma, but these inferences are beset by controversy. The date of origin of the animal kingdom is controversial as well, and the rock record before the Vendian does not provide many unambiguous clues to metazoan origins.

One thing is clear, however; by the Middle to Late Vendian, animals were present, as indicated by their trace fossils (*see* **Trace Fossils**). These trace or ichnofossils show evidence of peristaltic burrowing and displacement of sediment by burrowing activity, indicating that, by the end of the Vendian, animals with hydrostatic skeletons (coelomic spaces) were well established in the marine biosphere.

A difficulty with the study of Vendian animals is the fact that, apart from some phosphatized embryos that cannot be identified confidently to phylum, and some fossil sponges with preserved spicules, actual body fossils of these animals are rare. A comb-like structure from Russia, called *Redkinia,* may represent the flexible, filter-feeding mouthpart of a Vendian animal. However, apart from this, there are very few Vendian animal body fossils, particularly body fossils of the burrowing tracemakers.

The Cloudinidae consists of Vendian shelly fossils. The cone-in-cone tubular shells of genera such as *Cloudina, Sinotubulites* (**Figure 1**), and *Wyattia* are presumed to have been formed by worm-like animals. Cloudinids are occasionally associated with a bizarre, weakly calcified, stalked, goblet-shaped organism called *Namacalathus*. The biological affinities of cloudinids and namacalathids are not well understood, but they do seem quite unlike the more familiar skeletonized animals of the Early Cambrian. An algal affinity for the two has been suggested and cannot be ruled out. Large pores in the calyx of *Namacalathus* were probably filled with soft tissue in life; the pores might have served to admit light into the organism to sustain photosymbionts.

Ediacarans are the most puzzling part of the Vendian biota. These bizarre creatures grew to enormous sizes by Vendian standards, with some of the frondose forms reaching 2 m or more in length. The Ediacaran body is non-skeletal and seems to have been formed of a tough integument, in some cases partitioned into modules of similar shape. Unlike most other soft-bodied creatures, Ediacarans were capable of being preserved in sandstones. Ediacarans appear to be multicellular, but even this inference has been subjected to dispute. The concept of metacellularity (i.e., a body composed of uni- or polycellular partitions called metacells) has been applied to Ediacarans with some success.

Many palaeontologists have assigned these forms to conventional animal phyla, such as the Cnidaria,

Figure 1 *Sinotubulites cienegensis.* An early shelly fossil, this cloudinid is from the La Ciénega Formation of Sonora, Mexico. Holotype specimen. Length of largest tube, 12 mm.

Echinodermata, or Annelida, but the arguments supporting these assignments have not convinced everyone. Sceptics point out that not a single uniquely animalian trait has been identified on any of the thousands of Ediacaran fossils collected so far. A consensus for at least some of the fossils may be emerging, however, with recent new evidence showing trilobitoid arthropod features in *Spriggina,* and with the association of *Kimberella* with grazing traces.

In addition to the difficulties with systematic placement, there is the thorny problem of Ediacaran preservation. None of the Ediacaran fossils have skeletons, with the possible exception of the cloudinids and the namacalathids (whose relationship to the other Ediacarans is not known). Their cuticle also appears to have been relatively soft and flexible. Many of the Ediacaran fossils are preserved in relatively coarse sandy sediment, not ordinarily considered to be a good substrate for the preservation of non-skeletal organic remains. Thus, the preservation of Ediacarans must be explained either by unusual properties of the cuticle itself, or by anactualistic processes on the Vendian seafloor that could account for the preservation of soft tissue in sandstone. The latter possibility has evoked what is called the 'death mask' hypothesis, namely the idea that microbial mats in some way hardened the surface of dead Ediacarans, allowing them to be preserved as fossils. Unfortunately, the death mask theory is invoked by its supporters to explain all aspects of Ediacaran taphonomy – for example, the death mask mats supposedly formed beneath layers of storm sand and both above and below living Ediacarans, three highly uncertain propositions.

A key bedding plane surface from Newfoundland provides a new perspective on the death mask controversy (**Figure 2**). This bedding surface was smothered by a volcanic ash fall, entombing specimens of an unnamed spindle-shaped Ediacaran and a stalked form called *Charniodiscus*. The *Charniodiscus* was superimposed over the spindle form, and yet the morphology of the spindle creature shows clearly through the cuticle of the *Charniodiscus*. Thus, the death mask hypothesis is falsified, as there was no space for a post-mortem microbial mat to form between the *Charniodiscus* and the spindle. If the death mask hypothesis were correct, only the overlying *Charniodiscus* morphology should be preserved, but this is clearly not the case in this example.

Consequently, we are left with the idea that Ediacarans bore an unusually resilient cuticle. There is much going for this concept, as a link between Ediacarans and a problematic group known as conulariids (first suggested in 1987) has been dramatically confirmed by the discovery of the probable conulariid *Vendoconularia triradiata* in Vendian strata of the Ust'-Pinega Formation of the Onega River region, Russia. Conulariids appear to be a group of Ediacarans that survived well into the Palaeozoic. As with the Ediacarans themselves, most attempts at classifying

Figure 2 Cast of specimen of *Charniodiscus* superimposed over a spindle shaped form from the Mistaken Point assemblage in Newfoundland. Fibreglass cast (Pratt Museum, Amherst College). Width of view, approximately 11 cm.

conulariids have tried to assign them to the Cnidaria, and have met with failure. Part of the problem seems to be that conulariids display, as do other Ediacarans, both triradial and fourfold radial symmetry. This translates in metacellular terms to the unipolar iteration of six, eight, twelve, or sixteen founding metacells. The phosphatic nature of the (also somewhat flexible) conulariid cuticle may provide an important clue to the nature of the Ediacaran cuticle.

The most difficult Ediacaran problem is that of body geometry. Attempts to classify Ediacarans as animals have been hindered by their unusual constitution, which can show combinations of triradial symmetry and glide symmetry (as in the genus *Pteridinium*; Figure 3); these have proven to be impossible to fit into a conventional metazoan body plan model. Some researchers have tried to break the impasse by arguing that Ediacarans are colonial communities of individual animals. This explanation is not adequate to the task either, as each Ediacaran seems to be a well-integrated individual holobiont rather than a loose collection of individuals in a colony. Furthermore, there is no evidence for the loss of morphological features in specialized individuals, as is the case in the highly integrated colonies of modern hydrozoans.

The concept of metacellularity may help to solve this morphological problem. Some Ediacarans are apparently modular partition creatures, and the individual units or partitions, referred to as metacells, are either uni- or polycellular pods that may be flattened, stretched, inflated, or repeated (to form an iterated chain of metacells) as required. Metacellular creatures of this style are known amongst the modern biota (e.g., characean pondweeds, such as *Nitella*, can have individual cells measuring up to 15 cm in length), but no other type of organism has explored the potential of metacellularity as thoroughly as have the Ediacarans. The concept of metacellularity seems to apply best to the Ediacarans of the Mistaken Point biota in Newfoundland, the morphologies of which (spindle forms, branch forms, pectinate forms) are strange even by Ediacaran standards.

The metacellularity concept does not solve the problem of whether Ediacarans are animals; for example, hexactinellid (glass) sponges can undergo a coenocytic or syncytial (plasmodial) stage that is quite uncharacteristic for animals. Metacellularity is unknown in living animals and, for this reason, a number of palaeontologists prefer to place at least some of the Ediacarans into their own Kingdom, the Vendobionta. These Ediacarans could conceivably bear a closer relationship to several protoctist groups (such as the xenophyophores) than they do to animals. A common ancestry with the protoctist group that gave rise to the animals is still plausible, though, and may even be likely. Even Ediacarans that would seem to have a more animalian character, such

Figure 3 *Pteridinium simplex*. Cast of specimen from the Nama Group, Namibia. The fossil is viewed from its underside; the undersides of two vanes are visible, and the basal part of the third vane is visible between them as a stitch like chain of tube ends. Scale bar in centimetres.

Table 1 Ediacaran body forms, classified by metacellular grade

Original number of metacells	*Founding metacell(s) only, no iteration*	*Unipolar iteration of metacells*	*Bipolar iteration of metacells*
One	*Cyclomedusa, Evandavia, Parvancorina (?)*	*Charniodiscus, Charnia*	Spindle form of Newfoundland
Two	*Gehlingia*	*Dickinsonia, Phyllozoon*	*Windermeria, Ernietta*
Three	*Tribrachidium, Anfesta, Albumares*	*Swartpuntia*	*Pteridinium simplex*
Four	*Conomedusites*	*Spriggina (?), Marywadea (?)*	*Pteridinium carolinaensis*
Five	*Arkarua*	*Rangea (?)*	Unknown

Figure 4 *Evandavia aureola*. This hiemalorid Ediacaran occurs in the Clemente Formation, Sonora, Mexico. It is perhaps the oldest known (approximately 600 Ma) specimen of a complex life form. Scale bar in centimetres.

as *Yorgia waggoneri* from the White Sea region in Russia, show strange asymmetries in the 'head' region that challenge the animal interpretation.

Many Ediacarans may be placed in the metacellular grouping (Table 1). These forms manifest both a smooth central metacell surface and a surface formed of radial (and often bifurcating) tubes. *Evandavia aureola* is a non-iterated discoidal form from Sonora, Mexico, with a single metacell and radial tubes in its outer ring (Figure 4). Figures 5 and 6 show *Parvancorina minchami* from Australia, displaying a single trifold metacell with numerous branches coming from each branch of the initial metacell. *Parvancorina* serves as the subunit counterpart to *Tribrachidium* and *Gehlingia*, with the 'thumb structures' in the bilaterally symmetric *Gehlingia* and in the triradiate *Tribrachidium* homologous to the main medial branch in *Parvancorina* (Figure 7). Similar metacellular rearrangements may be observed in another Ediacaran clade, which includes the Newfoundland spindle-shaped form (Figure 8) and a 2 m long Newfoundland frond ('*Charnia*' *wardii*) shaped like an extremely elongate primary feather of a bird's wing. An interesting variation on the discoid body theme is a Chinese fossil from the Xingmincun Formation, southern Liaoning Province, consisting of a single, elongate metacell helically coiled into a flat disc. These specimens reach up to 4 cm in diameter.

Earth System Events

Eukaryotic phytoplankton are known from strata approaching 2 Ga in age, but, by the Vendian, these organisms began to resemble the modern phytoplankton at least in terms of diversity of tests, skeletons, and sheaths. Siliceous chrysophyte algae are known from Vendian strata of the Tindir Group, Alaska, and genera such as *Chilodictyon* and *Characodictyon* share a superficially diatom-like aspect.

Figure 5 *Parvancorina minchami.* A cast of a specimen from the Pound Supergroup of South Australia. Greatest dimension of specimen, 10 mm.

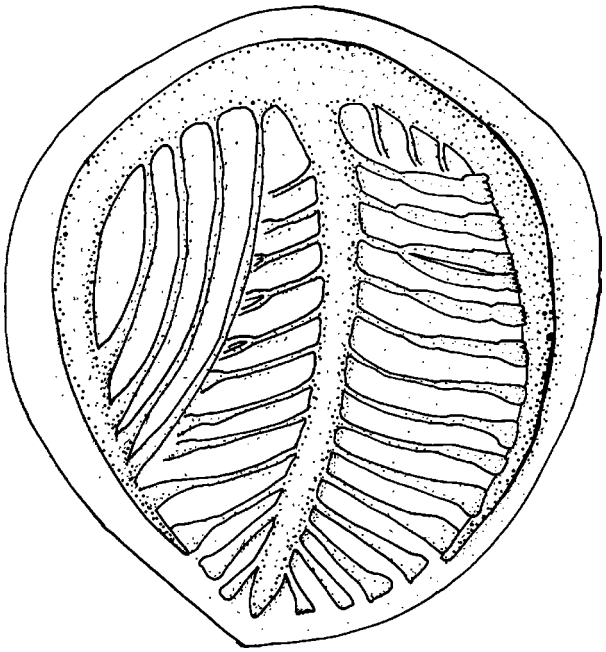


Figure 6 *Parvancorina minchami.* This Ediacaran poses difficulties for the arthropodan interpretation of this genus. Greatest dimension of specimen, approximately 10 mm.

In contrast, however, the Vendian seafloor had a distinctly primitive character. Microbial mats still blanketed the seafloor, as they had been doing since the Archaean, and stromatolites remained abundant even as burrowing began to alter their internal texture to the thrombolitic state. A wide variety of unusual sedimentary structures (with names such as *Arumberia*, *Kinneya*, and 'elephant-skin' texture) are known from fine clastic rocks of the Vendian. Lozenge-shaped structures (**Figure 9**) are often found on Vendian bedding surfaces; whether or not such structures might have use as a basis for interbasinal correlations is unknown. These primary sedimentary structures are puzzling, leading sedimentologists to invoke anactualist sedimentary processes to explain them. It seems quite reasonable to do so, as microbial mat carpeting of the seafloor would certainly influence the nature of marine sedimentation. Preston Cloud once called stromatolites 'organo-sedimentary structures', and the shelf, slope, and rise sediments of the Proterozoic may be thought of as a gigantic, connected, organo-sedimentary structure. Judging from the diversity of bedding plane texture types, never before or since have microbes had such a direct and intimate association with the basic processes of clastic sedimentation. Ediacarans evidently adhered to the surface of these mats or (in the case of forms with holdfasts and possibly in the case of *Pteridinium*) lived beneath them.

Metazoan burrowers apparently began their excavations beneath this mat surface, and evidence suggests that the earliest ichnofossil makers were submat burrowers. A good example is the Vendian trace fossil, *Vermiforma,* from the Carolina Slate Belt. This enigmatic and relatively large trace fossil, associated with other types of submat burrowers, consists of ten specimens that all follow the same rather tortuous track. As the traces are separated from one another by some distance, all underneath the mat, it is hard to

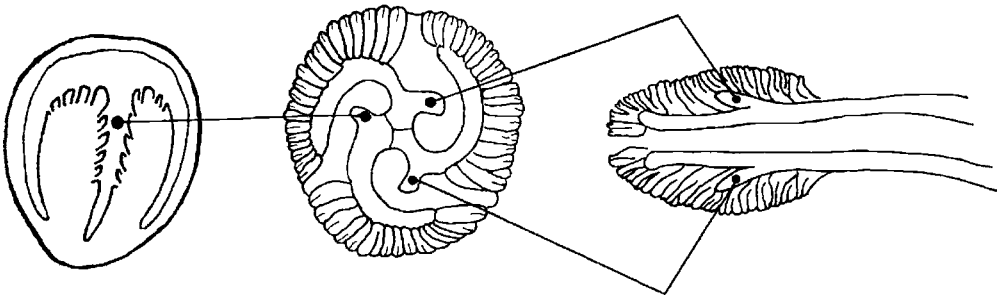


Figure 7 Ediacaran homology. Homologous structures ('thumb structures') linking the Ediacaran genera *Parvancorina* (left: length, 1 cm), *Gehlingia* (right: length, 8 cm), and *Tribrachidium* (centre: diameter, 1 cm). In spite of the vast differences in body symmetry type, the three genera are seen to be closely related.

Figure 8 Spindle shaped form from the Mistaken Point assemblage in Newfoundland. Fibreglass cast (Pratt Museum, Amherst College). Scale bar in centimetres.

imagine how and why the track paths were coordinated in shape. Arguments that *Vermiforma* is a pseudofossil are called into question by the presence of other types of trace fossils on the same bedding plane surface.

Throughout its duration the Vendian is marked by steadily increasing levels of bioturbation. The top of the Vendian system is currently defined by the appearance of the three-dimensional trace fossil *Trichophycus pedum*. The maker of this trace was able to excavate vertically as well as horizontally, and its activities (along with those of other vertical burrowers, such as the makers of *Skolithos*) tended to homogenize seafloor sediments. It would also presumably have punctured and shredded any microbial mats in the immediate vicinity.

It is thought that, towards the end of the period, the seafloor mat seal began to break down due to the intensity of metazoan burrowing activity. Such uncapping of the seafloor had dramatic consequences, both for the carbon budget of the planet (e.g., buried carbon was put immediately back into circulation) and for marine nutrient levels (e.g., sediment grains in suspension make wonderful substrates for nutritious bacterial growth). At about the same time, there was a tremendous flux of mineral nutrients to the oceans resulting from the fact that the Vendian saw a rare tectonic coincidence in Earth's history: the simultaneous occurrence of both divergent tectonics (final breakup of Rodinia) and convergent tectonics (formation of Gondwana, in an event known as the Pan-African Orogeny (*see* **Africa:** Pan-African Orogeny)), on a massive scale. As suggested by marine strontium isotopes of the Vendian, huge amounts of siliciclastic debris were shed from rift valley margins into the Vendian ocean. This would have been added to the volcaniclastic sediments derived from weathering of rift-associated basalts. The enhanced pool of igneous rock debris contributed mightily to an oceanic fertilization event that has been implicated in the emergence of skeletonized animals and the Cambrian explosion.

Figure 9 Lozenge shaped structures from the Clemente Formation, Sonora, Mexico, found in association with *Evandavia aureola* and other Ediacarans. Such structures are known from Ediacaran bearing strata on other continents. Scale bar in centimetres.

It appears that a massive palaeoecological reorganization of the marine biosphere took place at the end of the Vendian. With the exception of some very Late Vendian evidence for predatory activity in Sonora, Mexico, no evidence for large predators has been recognized in Vendian strata. In contrast, the Early Cambrian exhibits evidence for a great intensity of macropredation, involving forms such as the 2 m long predatory animal *Anomalocaris*. The Vendian biosphere has thus been referred to as a uniquely peaceful Garden of Ediacara, where large creatures, many with flattened bodies, partook of sunlight (via photosymbiosis), hydrogen sulphide (via chemosymbiosis), or osmotrophy (direct absorption of dissolved nutrients in seawater) on a mat-covered seafloor surface free of large predators. However, it was too good to last, and animals in the role of burrowers from below and predators from above triggered a rapid end to the Garden and the end of the Vendian.

The question of whether or not there was a mass extinction at the end of the Vendian is unresolved at present. A number of Ediacarans appear to have survived the Cambrian boundary, occurring with Cambrian fossils in Australia, Ireland, and elsewhere. Studies based on borehole data from the oil-producing strata in Oman have recently shown that cloudinids vanish from the record without a trace, and that this disappearance is not associated with any discernible lithological change. Thus, the disappearance of cloudinids from Oman cannot be attributed to environmental or facies change. Their loss has thus been interpreted as a major extinction event. Whether or not this localized loss of cloudinids represents a mass extinction, or an extinction event at all, is not known. Cloudinid-like fossils have been reported from the Cambrian Torneträsk Formation in northern Sweden.

Secular variations in secular isotopes in the Vendian are rather difficult to interpret, but, after a slight decrease from $-3‰$ $\delta^{13}C$ to $-5‰$ $\delta^{13}C$ after the Marinoan glaciation (Canadian and Namibian sections), the $\delta^{13}C$ values during the Vendian appear to undergo a fairly steady rise to a value of over $+3‰$ right before the Cambrian boundary (Dvortsy section, Aldan River, Siberia). The Vendian–Cambrian boundary itself is marked by a sudden $+3‰$ to $-1.4‰$ drop in $\delta^{13}C$ over a very short stratigraphical interval, as measured at the Dvortsy section.

This excursion has been linked to a variety of factors, including some sort of global environmental perturbation that also triggered the extinction of the cloudinids and many of the Ediacarans. Another way to look at it, though, is that the boundary excursion represents that moment in geological time when marine burrowing intensity crossed a threshold. At this critical point, the microbial mat seal on the seafloor may have been breached, resulting in previously immobile sediments (and their associated organic matter) becoming mobilized and injecting huge amounts of biogenic carbon into the water column. Owing to biogenic isotopic fractionation, this detrital organic matter was significantly depleted

in the heavy isotope of carbon and, as it went into circulation in marine water, caused the precipitous drop or boundary excursion in the $\delta^{13}C$ value.

It therefore seems reasonable to interpret the Vendian carbon isotope curve as a record of gradually increasing biotic productivity, with sequestering of much of the organic matter within and below the seafloor microbial mats, followed by a relatively sudden release of part of this organic matter deposited at the end of the Vendian as a result of increased burrowing intensity. Other factors, such as destabilization of gas hydrates in seafloor sediments, may also have been involved in these isotopic excursions.

A number of other phenomena can perhaps be traced to what has been called the Cambrian Substrate Revolution. The Vendian and Cambrian both saw an increase in the proportion of calcified filamentous microbes (such as *Girvanella*), which were perhaps less palatable to mat grazers than filamentous cyanobacteria and unprotected algae. With all the new sediment and organic matter in suspension, filter feeding probably became more possible throughout the water column, leading to the evolution of the first tiered filter feeders in the Cambrian. The only Vendian organisms that were likely to engage in suspension feeding were the cloudinids, which lived close to the sediment–water interface.

Finally, assuming that we are interpreting the secular carbon isotopic curve correctly, it is entirely possible that oxygen levels increased in the Vendian due to the sequestration of organic matter. Whether or not increasing oxygen levels influenced metazoan evolution is not known, although it seems fair to say that early burrowing animals would not have required high levels of oxygen. The earliest animal habitat appears to have been the submicrobial mat environment, where oxygen levels would probably have been rather low considering the relative abundance beneath the mats of hydrogen sulphide and other reduced compounds.

Glossary

abiogenically A term applied to rocks formed by processes not directly influenced by living organisms.

anactualistic processes Processes that occurred at one time in the Earth's past, but which are no longer operational today.

cloudinid A late Vendian calcareous shelly fossil consisting of closely nested, thin-walled tubes or cones. Thought to represent one of the earliest examples of a shelly animal fossil. Includes the genera *Cloudina* and *Sinotubulites*.

coelomic spaces The compartments that house the rigid, fluid-filled body cavity present in many animals. The coelom serves as a hydrostatic skeleton.

conulariid Any member of an enigmatic group of Vendian/Cambrian to Triassic shelly organisms. They formed conical, often pyramidal, tapering cones with transverse ribbing, composed of calcium phosphate.

Ediacaran Any member of a group of marine, megascopic fossils with a metacellular growth pattern. Found primarily in strata deposited before the Cambrian period. Assigned to extinct Kingdom Vendobionta.

frondose forms Ediacarans with a leaf, palm, or frond body form.

Garden of Ediacara A palaeoecological theory that holds that the marine ecosystems of the Vendian were largely free of megascopic predators and thus allowed organisms such as Ediacarans to survive unmolested using photosymbiotic, chemoautotrophic, and osmotrophic life styles.

holobiont A single integrated organism, as opposed to a colonial organism.

hydrostatic skeleton A fluid-filled internal organ or support structure within an animal's body that can be kept rigid or made limp by control of internal water pressure.

metacell A single or isolated modular unit of a metacellular organism; usually consists of a single enlarged cell.

metacellularity Term applied to organisms that are either multicellular (such as animals and plants) or consist of clusters or metacells (such as Ediacarans and certain types of aquatic algae).

Mirovia The Precambrian superocean that surrounded Rodinia.

molecular clock Any gene or gene sequence used by biologists in an attempt to determine the evolutionary time of divergence from a common ancestor between two or more groups of organisms belonging to different species.

osmotrophy A feeding strategy utilizing osmosis or direct absorption of nutrients.

peristaltic burrowing A burrowing strategy in metazoans that consists of rhythmic muscular contractions along the length of the body.

Rodinia A supercontinent consisting of all or nearly all of the continents. Consolidated one billion years ago (in an event referred to in North America as the Grenville Orogeny), this supercontinent broke up into smaller continents by the process of plate tectonics and continental drift before the Cambrian.

Sinian The Precambrian geological period immediately preceding the Vendian period.

Snowball Earth An extreme phase of glaciation in which glaciers reached tropical latitudes.

thrombolitic texture A texture in sedimentary rocks characterized by disrupted bedding lamination and a clotted fabric. Stromatolites develop this texture when influenced by burrowing metazoa.

Vendian The latest Precambrian geological period, immediately preceding the Cambrian period. Synonymous with Lipalian or Ediacaran.

Vendobionta The extinct kingdom to which Ediacarans are assigned.

See Also

Africa: Pan-African Orogeny. **Australia:** Proterozoic. **Biosediments and Biofilms**. **Grenvillian Orogeny**. **Palaeozoic:** Cambrian. **Precambrian:** Overview; Eukaryote Fossils. **Trace Fossils**.

Further Reading

Bottjer DJ, Hagadorn JW, and Dornbos SQ (2000) The Cambrian substrate revolution. *GSA Today* 10: 1 7.

Crimes TP (1999) Review of *Garden of Ediacara*. *Palaeogeography, Palaeoclimatology, Palaeoecology* 150: 357 358.

Crimes TP and Fedonkin MA (1996) Biotic changes in platform communities across the Precambrian Phanerozoic boundary. *Rivista Italiana di Paleontologia e Stratigrafia* 102: 317 332.

Crimes TP, Insole A, and Williams BPJ (1995) A rigid bodied Ediacaran biota from Upper Cambrian strata in Co. Wexford, Eire. *Geological Journal* 30: 89 109.

Donovan SK and Lewis DN (2001) The Ediacaran biota. *Geology Today* 17: 115 120.

Glaessner MF (1984) *The Dawn of Animal Life*. Cambridge: Cambridge University Press.

Hoffman PF and Schrag DP (2002) The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* 14: 129 155.

Ivantsov AYu (1999) A new dickinsoniid from the Upper Vendian of the White Sea Winter Coast (Russia, Arkhangelsk Region). *Paleontological Journal* 33: 211 221.

Ivantsov AYu and Fedonkin MA (2002) Conulariid like fossil from the Vendian of Russia: A metazoan clade across the Proterozoic/Palaeozoic boundary. *Palaeontology* 45: 1219 1229.

McMenamin MAS (1987) The fate of the Ediacaran fauna, the nature of conulariids, and the basal Paleozoic predator revolution. *Geological Society of America Abstracts with Program* 19: 29.

McMenamin MAS (1990) Vendian. In: Briggs DEG and Crowther PR (eds.) *Palaeobiology: A Synthesis,* pp. 179 181. Oxford: Blackwell Scientific Publications.

McMenamin MAS (1996) Ediacaran biota from Sonora, Mexico. *Proceedings of the National Academy of Sciences (USA)* 93: 4990 4993.

McMenamin MAS (1998) *The Garden of Ediacara: Discovering the First Complex Life*. New York: Columbia University Press.

McMenamin MAS (2000a) The antiquity of life: From life's origin to the end of the Vendian Period. In: Margulis L, Matthews C, and Haselton A (eds.) *Environmental Evolution,* 2nd edn. pp. 158 169. Cambridge, MA: MIT Press.

McMenamin MAS (2000b) Out of the shadows. *Notes and Records of the Royal Society of London* 54: 407 408.

McMenamin MAS (2001a) The Garden of Ediacara and the appearance of complex life. In: Guerzoni S, Harding S, Lenton T, and Ricci Lucchi F (eds.) *Proceedings of the International School of Earth and Planetary Sciences, Siena, Italy, 2001,* pp. 61 68. Siena, Italy: Consiglio Nazionale delle Richerche, University of Siena.

McMenamin MAS (2001b) *Paleontology Sonora: Vendian and Cambrian*. South Hadley, MA: Meanma Press.

McMenamin MAS (2003) Origin and early evolution of predators: The ecotone model and early evidence for macropredation. In: Kelley P, Kowalewski M, and Hansen T (eds.) *Predator Prey Interactions in the Fossil Record, Topics in Geobiology Series 20,* pp. 379 400. New York: Plenum Press/Kluwer.

McMenamin MAS (2003) *Spriggina* is a trilobitoid ecdysozoan. *Geological Society of America Abstracts* 35: 105.

McMenamin MAS and McMenamin DLS (1990) *The Emergence of Animals: the Cambrian Breakthrough*. New York: Columbia University Press.

McMenamin MAS and Weaver PG (1992) Proterozoic Cambrian paleobiogeography of the Carolina Terrane. *Southeastern Geology* 41: 119 128.

Seilacher A (1997) *Fossil Art*. Drumheller, Alta: Royal Tyrrell Museum of Paleontology.

Seilacher A and Pfluger F (1997) From biomats to benthic agriculture: A biohistoric revolution. In: Krumbein WE, Paterson DM, and Stal LJ (eds.) *Biostabilization of Sediments,* pp. 97 105. Oldenburg, Germany: Bibliotheks und Informationssystem der Carl von Ossietzky Universitat Oldenburg (BIS) Verlag.

PSEUDOFOSSILS

D M Martill, University of Portsmouth, Portsmouth, UK

Introduction

This entry discusses those objects that appear to be fossilized organic remains – often called pseudofossils – but which in reality are natural structures that merely resemble real fossils. Such structures occur commonly in sedimentary rocks and can easily deceive inexperienced, or even experienced, geologists and palaeontologists.

There are many types of pseudofossil, and they often occur associated with real fossils. They may be formed by both physicochemical and biochemical processes, some of which may be the same processes that are responsible for generating real fossils (*see* **Diagenesis, Overview**). Pseudofossils are highly variable in form, and the resemblance of a pseudofossil to a real fossil can be quite remarkable. To some degree an object's categorization as a pseudofossil is a reflection of the experience of the discoverer, and, as such, there is no strict definition of what is or is not a pseudofossil. Those with little or no experience of genuine fossils can easily believe that rocks and pebbles with striped or mottled patterning or unusual shapes might be genuine fossils; indeed, some pseudofossils are real fossils. Some calcareous algae can assume growth forms that resemble miniature versions of macro plant remains. A famous example of this is the so-called Landscape Marble from the Upper Triassic of the Bristol District of England. Here, intermixed arborescent and laminated growths of calcareous algae resemble scenes of forested hills (**Figure 1**).

Objects that are commonly confused with genuine fossils include both three-dimensional objects that represent some form of *in situ* mineral growth – called concretions or nodules – and largely two-dimensional entities occurring on rock surfaces that assume patterns resembling organic remains. In the latter category are the banded mineral growths called Liesegang bands (**Figure 2A**), dendritic mineral growths (**Figure 2B**), and various types of fracture surface (**Figures 3** and **4**).

Liesegang Banding

Liesegang banding (**Figure 2A**) is the result of the diffusion of soluble chemicals through micropores in a rock in differing concentrations and/or oxidation states and their precipitation to produce colourful (often orange, red, and brown) and variable banded patterns. Trace fossils (*see* **Trace Fossils**), such as filled burrows, can be prone to such diffusion and may resemble segmented worms and arthropods. In fact, such a structure would be both a genuine trace fossil (the burrow) and a pseudofossil (the imagined worm).

Liesegang banding occurs very commonly in chert nodules in the Cretaceous chalk formations of Europe. It also occurs commonly in many rocks where iron and manganese in a reduced state are weathering from the fresh rock to form oxides, such as goethite ($Fe_2O_3.H_2O$) and pyrolusite (MnO_2). Liesegang rings may penetrate the rock or may be restricted to surfaces such as bedding planes, joints, and microfracture planes.

Dendrites

Dendrites are arborescent patterns produced by the fractal growth of minerals as they precipitate from solutions that migrate through fractures in rocks. Commonly they occur as black traces of manganese dioxide (the mineral pyrolusite) or orange-brown precipitates of hydrated iron oxides such as goethite, although other mineral types may also occur. Such structures are commonly mistaken for the remains of soft foliage such as ferns and liverworts. Particularly beautiful examples occur in the Jurassic Solnhofen Limestone (**Figure 2B**) of Bavaria, Germany, and are often sold as ornaments.

Fracture Surfaces

The fractured surfaces of rocks may exhibit patterns that can be confused with organic remains. Very fine-grained and glassy rocks may split with conchoidal fracture – a series of concentric ripples emanating from the initial point of fracture (**Figure 3A**). Indeed, the very name suggests that the fracture resembles a shell. It is common for flint to fracture in this way, and such conchoidal fractures in flint may resemble the bivalve *Inoceramus* (**Figure 3B**), which is a genuine fossil that often occurs in Cretaceous flints.

Some fracture surfaces may exhibit feather-like traces, while others may show regular banding resulting from the intersection of cleavages or bedding planes with the fracture surfaces.

Figure 1 A real fossil resembling another. A cut surface through the Late Triassic 'Landscape Marble': both the laminated and the arborescent patterns were formed by calcium carbonate precipitation influenced by benthic photosynthetic microbes (probably blue green bacteria). Such structures are known as stromatolites. In former times, small village scenes were painted onto these slabs to create picturesque landscapes.

Figure 2 Superficial pseudofossils on bedding planes of the Late Jurassic Solnhofen Limestone Formation from Bavaria, Germany: (A) Liesegang banding as a series of iron oxide (goethite) diffusion fronts, and (B) dendritic patterns in pyrolusite and goethite. Photographs by Robert Loveridge.

In an unusual example (Figure 4), a piece of highly compacted fossil wood has numerous parallel and orthogonal evenly spaced cracks that have divided the wood into a series of small cubes. Such a blocky fracture pattern is common in fossil woods and can be seen in most bituminous coals. In the specimen shown in Figure 4, each small coaly cube generated a conchoidal fracture when the rock was split to reveal the fossil wood (the split propagated through the fossil wood rather than around it). As a consequence, each cube has a near-spherical surface with concentric ripples. This circular structure resembles the leaf-scar pattern seen on many stems of fossil plants, but the structures are an artefact of splitting.

Cone-In-Cone Structures

Cone-in-cone structures are the result of an unusual growth of minerals in which fibrous crystals assume a cone-like growth form (Figure 5A). They are frequently mistaken for fossils, particularly fossil corals, because the outer surface appears as a series of concentric rings or raised discs (Figure 5B). Cone-in-cone structures are usually composed of fibrous calcite crystals, which nucleate on the surfaces of limestone bands, shell beds, or even large fossils and grow orthogonal to that surface into fine-grained strata, such as clays and shales.

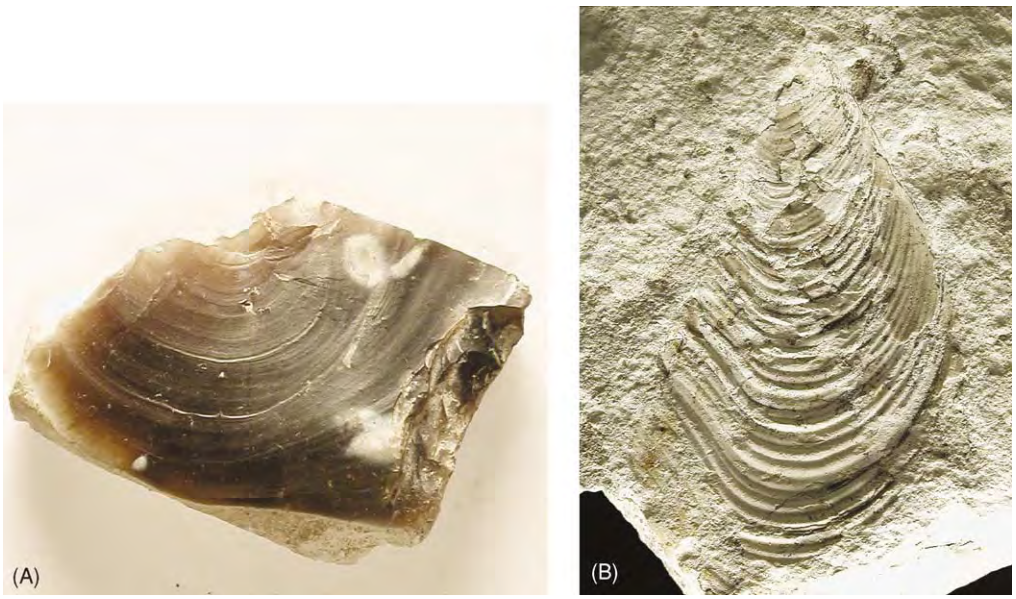


Figure 3 (A) Conchoidal fracture in flint. (B) The bivalve *Inoceramus*. *Inoceramus* occurs commonly in flint, and so conchoidal fractures may be easily confused with this bivalve. Photographs by Robert Loveridge.

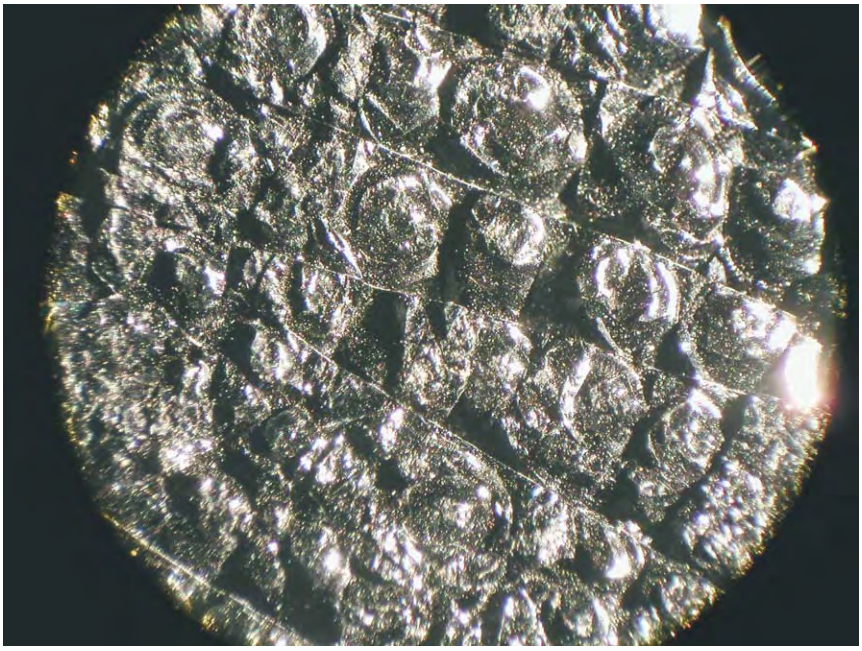


Figure 4 Conchoidal fractures in small cuboid segments of a fossil tree stem from the Lower Jurassic Posidonia Shale Formation of Dottenhausen, Germany. The conchoidal fractures resemble leaf scars but are a consequence of fracturing of the hard amorphous coal like material. Photograph by Robert Loveridge.

Nodules and Concretions

Many of the common diagenetic minerals (those minerals that form in sediment and cement that sediment into rock) form nodular growths that can easily be mistaken for organic remains. Such mineral growths may assume bizarre shapes, influenced by the rate and direction of diffusion of solutes through the sediment and fluctuations in pH and eH. They may contain genuine fossils, and in some cases the external shape of the nodule assumes the shape of the enclosed fossil as it grows around it. The external appearance of nodules and concretions can be

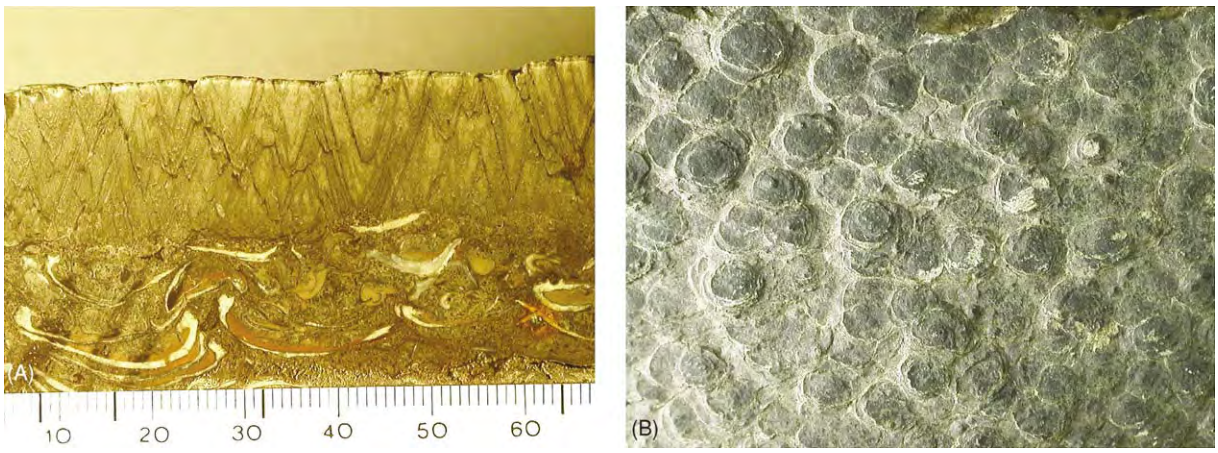


Figure 5 Cone in cone structures: (A) a cut section of a cone in cone that has grown on the surface of a shell bed, and (B) surface view each cone appears as a circle or disc that can easily be mistaken for a fossil coral at first glance. Photographs by Robert Loveridge.

Figure 6 Nodules and concretions. (A) A spherical concretion of mudstone can easily be mistaken for a fossil egg. (B) The concretion in **Figure 6A** is split in half to reveal internal septarian cracking. Such cracks are often mistaken for fossils. Photographs by Robert Loveridge.

highly irregular, and they frequently assume shapes resembling isolated bones or even whole animals when there is no fossil enclosed. Highly spherical concretions (**Figure 6A**) are often mistaken for fossil eggs. Internally, concretions may have radial cracks or a honeycomb network of crystal-lined cracks (**Figure 6B**). Such concretions are called septarian concretions, and they are frequently mistaken for fossils.

Common mineral species that occur as concretions and nodules include the iron sulphide minerals pyrite and marcasite (nodules of the latter are often mistaken for meteorites), silica in the form of chert and flint, and calcite where it cements clay. Nodules sometimes grow in bands or strings, and, en-masse, they can resemble rows of vertebrae or ribs of large skeletons. A nodule of ironstone found within the rib cage of a dinosaur was mistakenly thought to be a fossil heart by a team of Canadian palaeontologists.

Very recently, experiments in which crystals of barite have been grown in gels under laboratory conditions have produced morphologies and sizes that are remarkably similar to structures found in Precambrian rocks that have previously been interpreted as fossil bacteria. The laboratory-grown crystals also resemble structures found in Martian-derived meteorites that were interpreted as evidence of extraterrestrial life. It looks as though these ancient and extraterrestrial microfossils might also prove to be pseudofossils.

See Also

Biosediments and Biofilms. **Diagenesis, Overview**. **Fossil Plants:** Calcareous Algae. **Sedimentary Rocks:** Chert. **Trace Fossils**.

Further Reading

Garcia Ruiz JM, Hyde ST, Carnerup AM, *et al.* (2003) Self assembled silica carbonate structures and detection of ancient microfossils. *Science* 302: 1194 1197.

PYROCLASTICS

R J Brown, University of Bristol, Bristol, UK
E S Calder, Open University, Milton Keynes, UK

Introduction

Pyroclasts are formed by the explosive fragmentation of magma (molten rock) during volcanic eruptions (*see* **Volcanoes**). They are carried away from the vent by buoyant eruption plumes, extensive umbrella clouds, or by destructive ground-hugging pyroclastic density currents. The largest explosive eruptions can produce pyroclastic deposits many 1000's km^3 in volume, which can be emplaced on a regional scale in a matter of hours. The study of the physical characteristics of pyroclasts and pyroclastic deposits can reveal much about the dynamic processes involved during explosive eruptions and in pyroclast dispersal and deposition. Recognizing, observing, and understanding pyroclastic deposits are vital first steps in assessing and mitigating volcanic hazard (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards). This chapter summarizes the physical characteristics of the principal types of pyroclastic deposits and presents an introduction to their generation and emplacement mechanisms.

Generation of Pyroclastic Material

Explosive fragmentation of magma during volcanic eruptions can occur by two main mechanisms. The first involves the rapid exsolution of dissolved magmatic gases during rapid decompression events (magmatic eruptions), and the second results from the interaction of hot magma with external water sources (phreatomagmatic eruptions). Pyroclastic material can also be generated by rapid decompression and by autobrecciation processes during lava dome collapses.

Magma comprises three separate materials or phases: a viscous silicate melt (of varying composition), variable amounts of crystals (phenocrysts), and gas (volatiles) such as H_2O, CO_2, S, F, and Cl. There is a general positive correlation between the silica content of a magma and the degree of explosivity (Table 1). However, it is the quantity and behaviour of the gas phases that are critically important in determining the eruption style, because it is the rapid expansion of gas during decompression that drives explosive volcanic eruptions. Magma is stored at depth in magma chambers, under high temperatures and pressures. Magmatic eruptions are preceded by an increase in pressure and volume in the magma chamber. This is often attributed to the arrival of new magma into the chamber. The upper parts of many magma chambers are thought to contain a small volume fraction of gas bubbles (vesicles) due to supersaturation with volatiles, and seismic disturbance of these pre-existing bubbles can also lead to increases in magma chamber pressure (Figure 1). Crystallization, which enriches the melt in volatiles, can also act as a trigger. Once a critical point is reached, mechanical failure of the magma chamber roof occurs, allowing magma to rise, decompress, and exsolve gas in a runaway process (vesiculation) that can rapidly drive magma up the conduit at speeds of 200–400 m/s. Vesicle growth is controlled by the volatile content and by the physical properties of the magma (diffusivity rate, density, viscosity, and surface tension). The diffusivity rate is particularly important, and controls the rate at which gas bubbles escape from the magma: where escape is fast (in hot basic lavas), eruptions tend to be effusive or weakly explosive, but where escape is inhibited by high viscosities and low diffusivity rates (in intermediate and rhyolitic magmas), the exsolution of gas can explosively, and very violently, disrupt the magma. The expansion and coalescence of these bubbles forms a magma foam with radically different physical properties to that of the parent magma. During ascent, this rising vesiculated magma is fragmented into discrete particles and transforms into a gas-particle mixture, which accelerates up the conduit and is discharged into the atmosphere (Figure 1).

Phreatomagmatic fragmentation is driven by the volumetric expansion of external water after it has been rapidly heated by contact with magma. This mechanism is not restricted by magma type or vent type and it encompasses a spectrum of eruption styles

Table 1 Summary characteristics for the major types of explosive volcanic eruptions

Eruption type and examples	*C*	*H (km)*	*V (km^3)*	*D (km^2)*	*Common eruption processes*	*Pyroclast types*	*Duration*	*Deposits*
Magmatic								
Hawaiian Kilauea, Mauna Loa, Hawaii	Ba	$\ll$1	$\ll$1	<1	Sustained fire fountains; spatter fed lava flows; lava lakes; weak ash plumes	Scoria, ash, spatter, bombs, Pelee's hair and tears, reticulite	Days to years	Clastogenic lava; scoria cones; spatter cones; spatter ramparts; localized sheet like ashfall deposits
Strombolian Stromboli, Italy	Ba	$\ll$1	$\ll$1	5	Intermittent discrete explosions; ballistic clasts; weak ash plumes	Ash, scoria, bombs	Persistent over centuries	Scoria cones; scoria fall deposits; ballistic bombs and blocks localized sheet like ashfall deposits
Vulcanian Soufriere Hills (1995 present), Montserrat	Ba An, An, Da	<5 20	<1	<500	Numerous discrete and violent explosions; sustained eruption plumes; block and ash flows; dome extrusion; ballistic clasts	Ash, pumice, scoria, bombs	Minutes to hours	Ash fall deposits; ballistic blocks and bombs; small volume block and ash flow deposits; pumice flows
Plinian Vesuvius (1631), Italy Mt Pinatubo (1991) Philippines Taupo (AD 181) New Zealand	Rh, Ph, Da, Tr,	<20 >35	<1 3000	500 50 000	Sustained moderately to very high eruption plumes; widespread pumice fallout; sustained PDCs[1]; caldera collapse; commonly zoned	Ash, pumice	Hours to days	Extensive pumice fall deposits; widespread ignimbrites; ballistic blocks; pyroclastic breccias
Pheatomagmatic								
Surtseyan Surtsey (1963) Iceland	Ba	<20	<1	<500	Closely timed phreatomagmatic explosions; radially expanding PDCs[1]; ash plumes; peacock tail jets	Ash, pumice, scoria, glass	Days to months	Tuff cones; fall deposits; PDC deposits; ash fall deposits; ballistic clasts; accretionary lapilli
Taalian Taal (1965) Philippines	Ba	<20	<1	<500	Closely timed phreatomagmatic explosions; radially expanding PDCs[1]; ash plumes	Ash, pumice, scoria, glass	Days to months	Tuff rings; PDC deposits; ash fall deposits; ballistic clasts; accretionary lapilli
Phreatoplinian Askja (1875) Iceland Oruanui (26.5 ka BP) New Zealand	Rh(Ba)	<40	>1 >1000	500 50 000	Sustained high to very high eruption plumes; widespread pumice and ash fallout; sustained PDCs[1]; caldera collapse	Ash, pumice	Hours to days	Extensive bedded ash fall deposits; widespread ignimbrites

Abbreviations: C composition (Ba basalt; An andesite; Da dacite; Ph phonolite; Tr trachyte; Rh rhyolite), H plume height; V volume, D dispersal, PDC pyroclastic density current.
[1]Pyroclastic density cements.

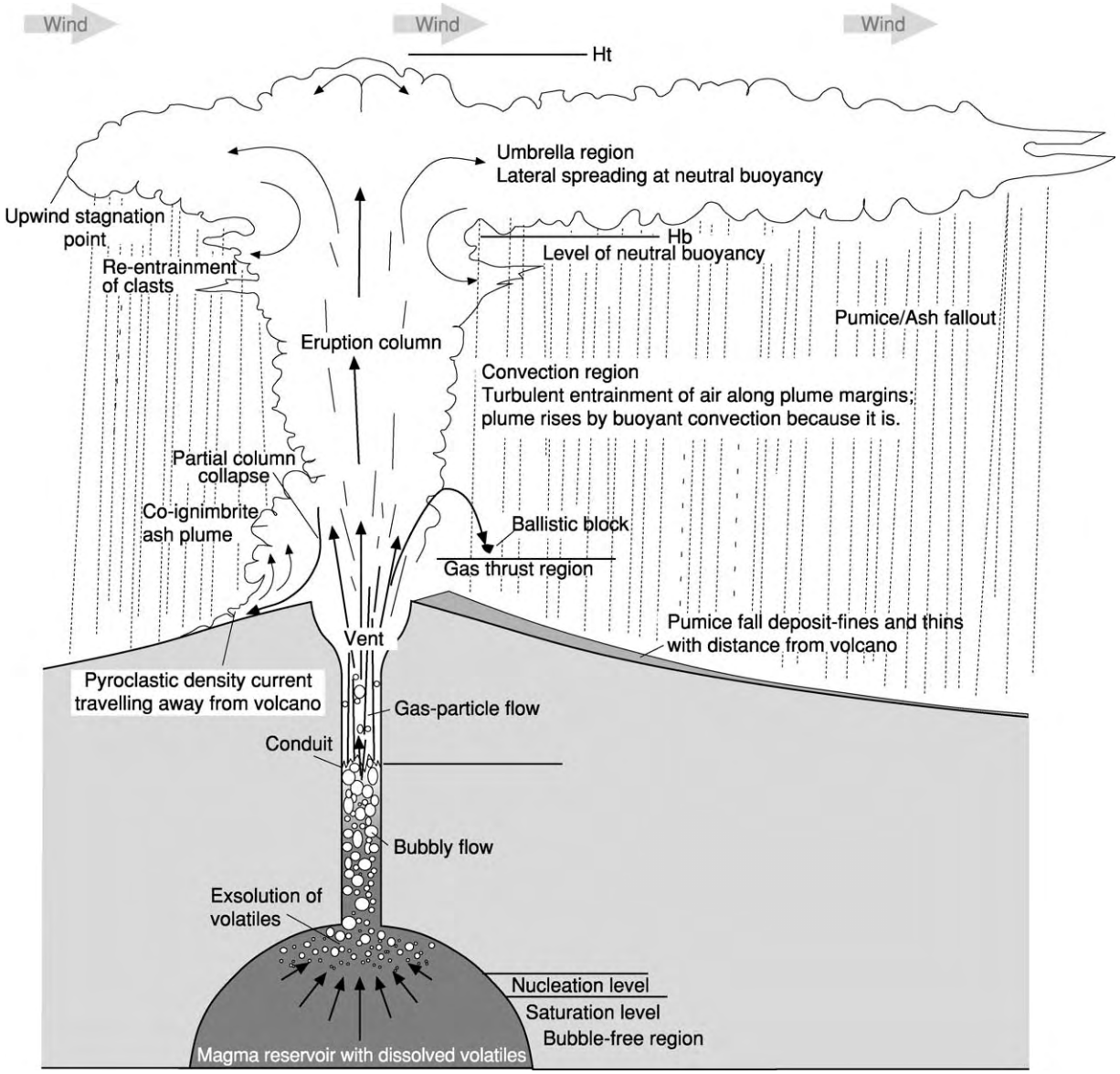


Figure 1 Cartoon illustrating the processes involved during an explosive eruption (in this case a Plinian eruption). Major processes in the magma chamber, conduit, and eruption plume are indicated.

from the shoreline fragmentation of lava (*see* **Lava**) entering the sea, to large volume phreatoplinian eruptions (Table 1). During a phreatomagmatic eruption, thermal energy is converted into mechanical energy (including seismic and acoustic energy) and kinetic energy (fragment motion). Fragmentation is controlled by the mixing ratio between water and magma, and by the magma viscosity, temperature/pressure, and the water/magma contact mode. As the eruption efficiency increases, more mechanical energy is released and fragmentation increases, resulting in a finer grain size and a greater distribution of pyroclasts.

The two fragmentation mechanisms are not mutually exclusive. Exsolving magmatic volatiles may contribute to expansion and fragmentation during phreatomagmatic eruptions, while phreatomagmatic phases are common during sustained magmatic Plinian and less extreme types of eruptions.

Eruption Plumes

Volcanic plumes are generated when explosive eruptions discharge a mixture of pyroclasts and gas into the atmosphere (Figure 2). These may either form buoyant convective plumes, which rise to great

Figure 2 Small Strombolian fire fountain feeding a weak ash plume (Etna, South Italy, 2003). (Photo courtesy of Tom Pfeiffer www.decadevolcano.com)

Figure 3 Pyroclastic density current generated during a Vulcanian eruption at Soufriere Hills volcano, Montserrat, West Indies. (Photo: E. Calder.)

heights in the atmosphere (<50 km) or they may collapse to form fountains (column collapse), which can generate flows of hot particles and gas that move laterally away from the volcano (pyroclastic density currents; Figure 3). Eruption plumes are termed 'maintained plumes' if the duration of discharge is sustained relative to the ascent time of the plume (e.g., Plinian eruptions) and 'instantaneous plumes' or 'thermals' if the discharge is short-lived relative to ascent time (e.g., Vulcanian explosions). The basic tripartite structure of these plumes is shown in Figure 1. In the gas thrust region the mixture of gas solids and liquid is expelled vertically from the vent as a jet with speeds up to a few hundred metres per second and a bulk density greater than that of the surrounding atmosphere. Air is mixed into the column, heated by the pyroclasts and expands, resulting in a decrease in the mixture density. If sufficient air is entrained and heated, the bulk density of the jet decreases below that of the surrounding atmosphere and the jet becomes buoyant, rising as a convective column. The plume continues to rise until it reaches a level of neutral buoyancy (Figure 1, 'Ht and Hb') at which it spreads laterally in an umbrella cloud. If however, sufficient air is not entrained, the jet decelerates until its velocity becomes zero and since the bulk density is greater than the surrounding atmosphere, the mixture begins to fall back to the Earth's surface.

Eruption Styles

Important measures of an explosive volcanic eruption are the magnitude (volume of erupted dense rock equivalent magma, in km^3), intensity (mass-flux), dispersal (areal extent of fall deposits, km^2), and driving mechanisms (magmatic or phreatomagmatic). These parameters are used to classify explosive eruptions; the characteristics of the major types are summarized in Table 1. Volcanic eruptions are extremely dynamic natural phenomena and many sustained eruptions undergo significant changes or fluctuations in eruption style through time, for example, changing from Surtseyan phreatomagmatism to Strombolian fire-fountaining. Likewise, the opening stages of Plinian eruptions can be phreatomagmatic or Vulcanian in style. Thus, the deposit of one eruptive sequence may comprise layers of contrasting pyroclasts that were formed by differing fragmentation mechanisms.

Pyroclast Types and Deposits

During explosive volcanic eruptions, magma can be transformed into pyroclasts (also called tephra) of widely varying physical properties, that range in size from submillimetric ash up to boulder size. Pyroclasts are classified according to their origin. Juvenile pyroclasts are derived from the erupting magma, and include pumice, scoria, volcanic glass, ash, crystals, and bombs (see Table 2 for descriptions). In magmatic eruptions, juvenile volcanic ash (vitric ash) comprises fine-grained cuspate, platy, or Y-shaped vesicle-wall shards. Vitric shards generated during phreatomagmatic eruptions can have blocky, fusiform, moss-like, platy, and spherical or drop-like shapes, reflecting the many different ways in which water and magma can explosively interact. Cognate pyroclasts are derived from the deposits of earlier eruptions at the same volcano, while accidental pyroclasts are those derived from the subvolcanic basement during the eruption or picked up by a pyroclastic density current during transport. Pyroclastic deposits typically comprise a mixture of juvenile, cognate, and accidental components. Due to the large differences in the physical properties of these components, and the sorting and segregation processes operating during pyroclast

Table 2 Characteristics of the main types of pyroclastic material

Pyroclast	*Description*	*Associated eruption styles*
Ash	Fine grained (<2 mm diameter); whole or broken crystals, accidental lithic clasts or glass (vitric) shards; vitric shards can be cuspate, platy (broken vesicle walls), blocky (unvesiculated glass), or pumiceous	All
Pumice	Highly vesiculated volcanic glass; (density <1.0 g cm^3 floats on water); fibrous or spherical/subspherical vesicles	Silicic explosive eruptions (particularly Plinian)
Scoria	Poorly vesiculated volcanic glass (density <1.0 g cm^3 sinks in water; spheroidal vesicles; vesicularities 70 85%	All explosive eruptions; particularly Strombolian, Vulcanian
Bomb	Fluidal clot of magma that partially cools during flight; types include spindle, cow pat, breadcrust, and cauliflower bombs	Hawaiian, Strombolian, Vulcanian
Spatter	Irregular fluidal clots of magma that can coalesce (weld) on impact	Hawaiian, Strombolian
Pelee's hair/ tears	Hair: long thin strands of glass Tears: irregular droplets with fluidal shapes (e.g., tear drop, ovoid and spheroid)	Hawaiian, Strombolian
Reticulite	Polygonal, lattice like networks of very fine glass rods; very highly vesicular (95 99%)	Hawaiian
Accretionary lapilli	Comprise an ash core surrounded by one or more concentric laminae of usually finer grained ash; also ash pellets, which lack concentric rims; Armoured or cored lapilli comprise a pumice or lithic lapilli core surrounded by a concentric rim of ash	All explosive eruptions that generate abundant fine ash; particularly phreatomagmatic (Surtseyan and Phreatoplinian); Magmatic eruptions during humid atmospheric conditions

Table 3 Particle size parameters for pyroclastic material

phi	*Mm*	*Class*
<3	<0.12	Very fine ash
3 1	0.12 0.5	Fine ash
1 to 0	0.5 1	Medium ash
0 to 1	1 2	Coarse ash
1 to 3	2 8	Fine lapilli
3 to 5	8 32	Medium lapilli
5 to 6	32 64	Coarse lapilli
> 6	>64	Blocks and bombs

dispersal, the lateral and vertical distributions of these components may vary significantly within a deposit.

The grainsize terminology for pyroclastic material is given in Table 3. The term ash, when used to describe a pyroclastic deposit, refers to a unit comprised dominantly of ash-grade particles. When such a deposit is indurated or lithified, it is commonly called tuff. Deposits comprising ash and lapilli-grade particles are referred to as lapilli-tuffs, and deposits containing an abundance of blocks are called pyroclastic breccias. Pyroclastic successions can be best studied by utilizing a lithofacies-based approach. Lithofacies are commonly defined on a two-tier system of sedimentary structure and grainsize, although many classification schemes have been proposed and there is perhaps a strong case to be made for the standardization of pyroclastic lithofacies.

Pyroclastic Fall Deposits

Pyroclastic material carried up in eruption plumes eventually rains out from umbrella clouds to form pyroclastic fall deposits. These are classified according to a spectrum of increasing dispersal and fragmentation (Hawaiian, Strombolian to Plinian, Figure 4), but can also be distinguished on broad lithological grounds; scoria-fall deposits, pumice-fall deposits, or ash-fall deposits. Pyroclastic fall deposits are commonly well sorted and they maintain an even thickness over local topography (mantle bedding). Small eruptions (e.g., Hawaiian and Surtseyan) tend to produce weak, low-level plumes that form steep-sided, high aspect ratio fall deposits (cones), whereas Plinian eruptions produce high level (<50 km) columns that deposit extensive low aspect ratios (sheets).

An important concept in the formation of these deposits is that of terminal fall velocity, which depends on particle size and density and determines at what distance downwind from the source individual particles will begin to settle out. Large pyroclasts are often explosively ejected on ballistic trajectories (ballistic clasts) and are unaffected by the wind. The

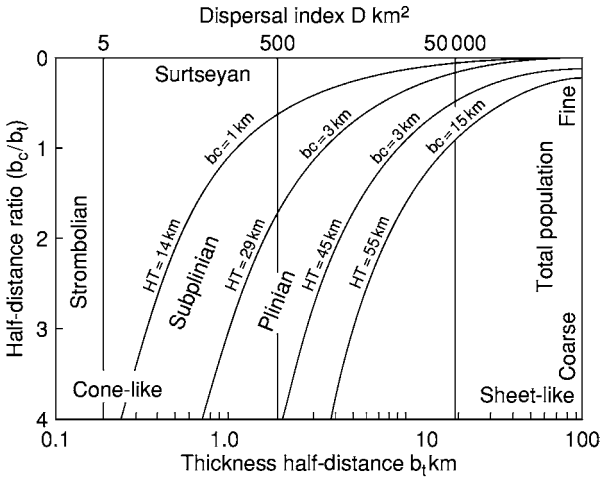


Figure 4 Classification scheme for fall deposits. The half distance ratio represents the total grainsize population and the thickness half distance represents the dispersal. The figure is contoured for clast half distance b_c and for total column height in km H_T. (From Pyle (1989).)

Figure 5 Parallel bedded Plinian pumice fall deposits (Bandas del Sur Group, Tenerife, Canary Islands) overlying palaeosol (level with man's chest). Thin ash layers are visible in the middle of the deposit. White unit at top of photo is an ignimbrite. (Photo: M. Branney.)

thickness and grainsize of fall deposits decays approximately exponentially away from source and the rate of this thinning or fining is conventionally described by the 'half-distance', a measure of the distance over which the thickness or grainsize of the deposit halves. Quantification of fall deposits relies on the accurate mapping of the distribution and on detailed granulometric analysis. Isopach (contours of equal thickness) and isopleth (contours of equal grain-size) maps can be constructed and used to determine erupted volumes and investigate eruption dynamics. Isopachs and isopleths are commonly elongated downwind due to wind shear, although varying wind velocities and directions at different levels in the atmosphere, or changes in wind direction through time, can complicate dispersal patterns during sustained fallout.

Fall deposits are sensitive indicators of eruption dynamics. Barring interference from local atmospheric phenomena, quasi-steady eruption conditions should produce texturally homogenous, massive fall deposits. Such conditions are rare in nature and many fall deposits exhibit features indicative of eruption unsteadiness, for example, bedding (Figure 5), stratification, vertical grainsize variation (grading), and/or compositional variations (e.g., in abundance of accidental lithic lapilli). Such features are usually better developed in proximal deposits, becoming less well defined toward distal regions due to increased transport time. Table 4 summarizes some common features of fall deposits and their possible interpretations.

Pyroclastic Density Currents and Their Deposits

During an explosive volcanic eruption, pyroclasts can be carried laterally away from the vent as pyroclastic density currents (PDCs). These are ground hugging (i.e., they are denser than air) particulate gravity currents comprised of gas, ash, lapilli, and blocks. Pyroclastic density current is a general term given to a wide variety of phenomena generated by a number of very different mechanisms, for example, the collapse of an eruption plume, sustained boil-over fountain eruptions and directed lateral blasts (Figure 6). They can travel at high velocities (30–200 ms), can have high temperatures (>550°C), and can reach over 100 km from their source. PDCs tend to be controlled by the topography over which they flow, often being preferentially channelled down valleys but they can also surmount topographic obstacles more than 1.5 km high and travel many kilometres across open sea.

The nomenclature relating to the types of pyroclastic density currents and their respective deposits is notoriously complex. Pyroclastic density current deposits can be subdivided and classified in a number of different ways. Obvious distinctions can be made based on composition, deposit volumes, and dominant lithofacies, or on generic interpretations such as whether they resulted from either short-lived transient currents (e.g., during phreatomagmatic eruptions) or from sustained currents (during Plinian eruptions).

The term 'ignimbrite' is generally used to describe an ash and pumice-rich deposit generated during a Plinian-type eruption. Ignimbrites are some of the most impressive deposits on the planet: vast landscape burying sheets, emplaced over a few hours that can be several hundred metres thick and cover huge areas (>45 000 km^2). Ignimbrite range from small volume

Table 4 Major features of pyroclastic fall deposits and possible interpretations

Fall deposit characteristic		*Interpretation*
Geometry		
High aspect ratio cone		Fallout from a low (<2 km) eruption column; restricted dispersal
Low aspect ratio sheet		Fallout from a high eruption column (>2 >45 km); widespread dispersal
Circular isopachs/ isopleths		Fallout during no wind conditions; fallout from a powerful eruption column undeformed by wind
Elongate Isopachs/ isopleths		Strong distortion of eruption column by wind (elongation in downwind direction)
Changing isopach/ isopleth distribution with height		Variable wind directions at different heights; changing wind directions with time
Secondary thickening of fall deposit		Downwind aggregation of ash and premature fallout of large aggregates with greater settling velocities
Structure		
Grading	Inverse	Fallout from a growing eruption column (waxing eruption intensity); gravitational (downslope) remobilization of deposits (e.g., Strombolian deposits)
	Normal	Fallout from a waning eruption column (waning eruption intensity)
	Bedding	Fallout during unsteady eruption column dynamics (fluctuating eruption intensity)
Cross stratification, lenses		Reworking of falling or deposited clasts due to surface winds, PDCs or gravitational remobilization (rolling down steep slopes)
Erosion surfaces (unconformities)		Removal of deposit due to mass wasting (landslides), PDCs, surface processes (alluvial/slope erosion)
Impact sags		Deformation by ballistic clasts
Compositional changes		
Chemical zoning of juvenile pyroclasts		Changes in the chemical composition of the erupted magma with time eruption from a zoned magma chamber
Layers rich in accidental clasts		Explosive fragmentation of vent walls following blocking, collapse or increase in erosive power
Vertical changes in accidental clast composition		Vertical or lateral (from fissure vents) shift in locus of vent erosion
Layers rich in dense juvenile clasts (e.g., obsidian)		Explosive clearance of a cognate volcanic plug
Intercalated thin ash beds		Short lived phreatomagmatic explosions which generate abundant fine ash; settling of fine ash from atmosphere during pauses in fallout activity
Pink ash coatings on pumice lapilli		Thermal alteration (oxidation) of pyroclasts
Accretionary lapilli; ash pellets; armoured lapilli		Moisture driven aggregation of ash common during phreatomagmatic activity and during humid atmospheric conditions; electrostatic attraction of fine ash particles
Texture characteristics		
Poor sorting (presence of an ash matrix)		Partial phreatomagmatic activity (hybrid eruption); coeval deposition of fallout and fine ash from a dilute ash rich PDC or a co ignimbrite ash cloud
Rounded pumice clasts		Emplacement by lateral currents (PDC); reworking by alluvial processes
Welding		Proximal high temperature agglutination of hot (fluidal) pyroclasts; contact fusing of pyroclasts by an overlying welded ignimbrite

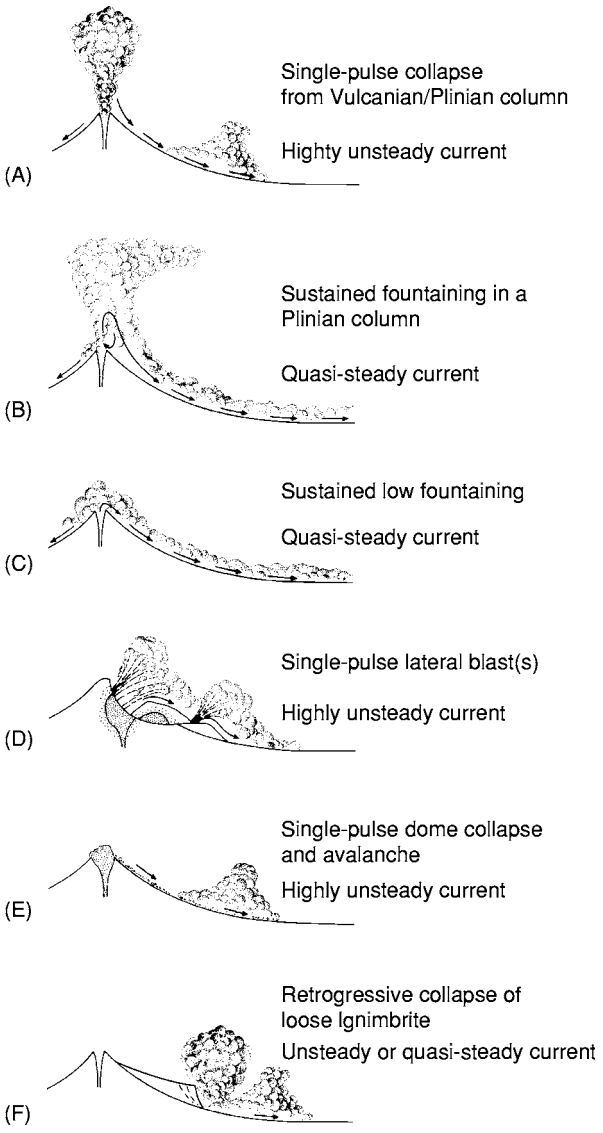


Figure 6 The origins of pyroclastic density currents. Most large volume ignimbrites are derived from B and C. Type D refers to transient pyroclastic density currents generated by lateral blasts or small scale phreatomagmatic activity (Surtseyan or Taalian eruptions). Type E commonly produces block and ash flows. (From Branney and Kokelaar (2002).)

ribbon-like deposits confined to valleys to extensive sheets distributed radially around the vent or across mountain ranges. Plinian eruptions are usually sustained for hours to days, and PDCs may be generated for all or part of that time, or intermittently. Ignimbrites exhibit a breath-taking diversity in their physical characteristics (e.g., geometry, bedding, sorting, grainsize, clast grading, geometry). Many are massive (Figure 7) and exhibit internal coarse-tail grading patterns the size grading shown by the coarsest particles in the deposit, but bedded, stratified, and cross-stratified lithofacies also occur. Detailed lithofacies

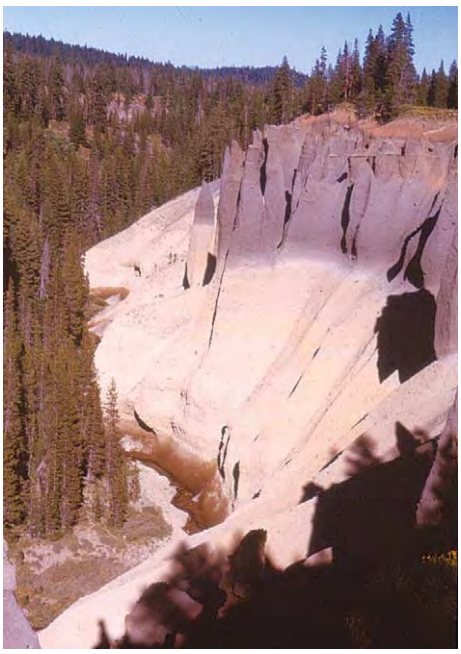


Figure 7 Compositionally zoned ignimbrite (Mount Mazama ignimbrite, Crater Lake, Oregon). (Photo: M. Branney.)

analyses, similar in approach to those applied to sedimentary deposits, have proved most successful in dealing with the complexity of ignimbrites and in elucidating important information about emplacement mechanisms. Moderate volume ignimbrites (1–10's km^3) often have complicated lateral, longitudinal, and vertical lithofacies transitions (ignimbrite architecture). Ignimbrites often show evidence for crystal enrichment relative to the juvenile pumice clasts, indicating significant loss of ash during transport (a process termed elutriation). This elutriated ash forms co-ignimbrite plumes, which rise vertically from moving flows and can later deposit extensive ash deposits.

Progressive evacuation of a chemically zoned magma chamber results in vertical chemical zoning of juvenile pyroclasts in an ignimbrite (Figure 7). Typically, juvenile pyroclasts become more basic with height (e.g., rhyolite to andesite; the inverse of vertical zoning in the magma chamber). Early phases of the eruption tap acidic magma (e.g., rhyolite) at the top of the chamber, while lower and more basic levels (e.g., andesite) are tapped during later phases. High temperature emplacement can lead to the agglutination of hot pyroclasts on deposition (welding). Pumice lapilli can flatten and deform to produce bedding-parallel *fiamme*, and columnar joints, similar to those developed in lavas, can develop. In extreme cases, the aggrading ignimbrite can start to flow as a dense viscous liquid and can end up looking superficially like lava.

Small volume ($<<1\,km^3$) well stratified pyroclastic density current deposits ('surge deposits') are typical products of monogenetic phreatomagmatic eruptions (e.g., Taal, Philippines) and directed lateral blasts (e.g., Mount St Helens, 1980). The pyroclastic density currents are typically transient single-surge types that travel out horizontally from source and rapidly wane. Each deposit can exhibit a wide range of sedimentary structures, and they characteristically show a proximal to distal lithofacies transition from massive coarse-grained lithofacies near source, stratified and cross-stratified deposits (with spectacular dune bedforms, Figure 8) further out, and planar-bedded deposits in more distal areas. Fine-grained ash fall deposits extend out from the distal limit of the sheet. Vertical lithofacies sequences often match the proximal–distal lithofacies transitions, reflecting deposition during strongly waning flow. Phreatomagmatic eruptions are characterized by numerous closely timed explosions, and during the course of the eruption, the deposits build up gently tapering cones with radii of several kilometres (called tuff rings or tuff cones). Due to the presence of external water, evidence for cool emplacement ($<100°C$) may be common in phreatomagmatic deposits, for example, abundant accretionary lapilli, vesicles in ash beds, rain-splash micro bedding, and the plastering of ash layers against objects.

Also included under the general term pyroclastic density currents, are hot avalanches of dense lava blocks and ash (block and ash flows) generated when portions of growing lava domes collapse due to gravitational instability. These events are short-lived and occur in discrete pulses. A classification has developed for these small volume pyroclastic flows depending on their generation mechanism. The term 'nuée ardent', favoured in some of the older literature, has also been applied in a rather general sense to most of these flow types. The generation of small pyroclastic flows by dome-collapse has been observed on numerous occasions (Montagne Pelée 1902; Fuego 1974; Mt Unzen 1990–1995; Merapi Volcano, and Soufrière Hills Volcano, Montserrat 1996–1998). These flows have been observed to move as far as a few kilometres from the vent at speeds of up to 60 ms. Due to the style of eruption, pyroclastic flows produced by gravitational collapse of lava domes occur repeatedly over significant time periods, typically, a few months to several years. This has enabled significant progress to be made in understanding the behaviour of these flow types in comparison to those of large volume ignimbrite eruptions. The deposits of these flows (block and ash flow deposits) vary from dense lava blocks and ash to dense semi-vesicular pumice and ash with only minor accidental components. Collapse volumes and local topography determine whether deposits are valley confined or sheet-like in nature. Small volume ($<0.5 \times 10^6\,m^3$) deposits tend to form well-defined lobate deposits with well-defined flow fronts and levees (Figure 9), while more extensive block and ash flow sheets tend to form deposits which thin laterally and have tapering margins.

Figure 8 Cross stratified pyroclastic density current deposit emplaced during a small volume phreatomagmatic eruption. Note spectacular dune bedforms interbedded with thicker mas sive deposits (Caldera del Rei tuff ring, Tenerife, Canary Islands). Current direction from right to left. (Photo: M. Branney.)

Pyroclastic Density Current Transport and Deposition

The literature on pyroclastic density currents has been dominated by discussions on alternative concepts of transport and emplacement mechanisms. The physical nature of these flows and the mechanisms of sediment transport and deposition are still not fully understood. Conceptual models have been developed, ranging from dense concentrated suspensions dominated by particle interactions and fluidisation effects to very dilute

Figure 9 Well developed lobes in block and ash flow deposits (Soufriere Hills Volcano, Montserrat). Small stripped trees for scale in left upper middle of photo. (Photo: E. Calder.)

turbulent suspensions. In high particle concentration models, pyroclastic flows are considered as avalanche-like granular flows analogous to debris flows and rock avalanches. Variants of models involving high particle concentration granular flows emphasize non-Newtonian rheology, fluidisation effects, and hindered settling. Conceptual models have also been proposed that to varying degrees bridge the gap between the two rheological extremes. These latter models, which perhaps best describe the dynamics of PDCs generated during larger ignimbrite forming eruptions, propose separate low-concentration flows and higher-concentration depositional systems (flow-boundary processes). The historical development and evolving ideas surrounding this still somewhat controversial topic are covered in detail in several textbooks and major reviews (see **Further Reading** section at end of this article).

Pyroclastic density currents transport components of widely differing grainsizes and hydrodynamic properties (low density pumice, ash, and dense lithic clasts). During transport, these pyroclasts may be supported by a number of different mechanisms (fluid turbulence, saltation, rolling and sliding along an interface, and clast collisions) and adjacent clasts in a deposit may have different transport and particle support histories. Several field investigations support the idea that ignimbrites can be rapidly aggraded increment-by-increment from the bases of pyroclastic density currents (a process termed 'progressive aggradation', Figure 10A). Vertical chemical zoning of juvenile pyroclasts, gradual changes with height in an ignimbrite in grain fabric orientation, and lateral transitions from massive to stratified lithofacies, provide good evidence for incremental deposition. Deposition accompanies transport and the thickness of a deposit reflects the durations and rates of deposition, with basal parts deposited earlier than upper parts. Each particle carried in the current must pass through a lower flow boundary zone before it can be deposited (Figure 10).

The massive, poorly sorted nature of many pyroclastic density current deposits (Figure 7) indicates deposition from high particle concentration flows or flow boundary zones in which particles are supported by collisions (in granular flows, Figure 10D) and the upward flux of dusty gas produced as larger particles are hindered in settling, Figure 10E). Hindered settling results in partial fluidisation (sedimentation fluidisation) and greatly increases a current's mobility. Stratification and cross-stratification result from traction sedimentation involving fully dilute flow boundaries in which particle interactions have little role in particle support or segregation or in current rheology (Figure 10C). Small particles are supported by the turbulence in the surrounding dusty gas (air and suspended ash), whereas larger particles move by saltation along the base (as a bedload). Fully dilute conditions in the flow boundary zone result in more efficient segregation mechanisms and the deposition of relatively well-sorted stratified and cross-stratified deposits ('surge deposits', Figure 8). Small volume phreatomagmatic eruptions tend to produce currents initially of these characteristics, but which lose energy rapidly as they travel away from the vent. This results in higher rates of suspended load fallout, increasing particle concentrations and the deposition of more massive deposits with distance from source. Flow-boundary conditions in between the two end members outlined above (for example, in terms of particle concentration) will result in deposits that are gradational in character (diffuse stratified and bedded ignimbrites).

Inferences made from the deposits about the structure (e.g., particle concentration, density stratification) within the moving pyroclastic density currents have to be made with caution. However, the ability of some flows to travel large distances across water and scale mountains up to 1500 m high, >45 km from source, indicates that the greater thickness of these currents were dilute, low particle suspensions. Fluid turbulence is likely to be the dominant particle support mechanism in the upper parts of the currents. Additional evidence on the vertical structure of PDCs can be gleaned from ignimbrites that were emplaced over rugged topography. Parts deposited over high ground tend to be thinner and stratified or bedded, while contemporaneous deposits in depressions and valleys are more typically thick and massive (Figure 11). Such evidence suggests that conditions in the flow-boundary zone (e.g., in terms of current velocity, turbulence, and particle concentration) of the PDC varied significantly across topography and that aggradation rates were significantly higher in the valleys. What remains unclear for most pyroclastic density currents, however, is the partitioning of material between a higher concentration lower flow boundary zone and an overlying dilute and turbulent zone, noting that the vertical structure and the vertical density gradients are likely to vary considerably, both between and within PDCs, due to current non-uniformity and unsteadiness (see below).

Detailed lithofacies analyses of ignimbrite sheets provide important information on spatial variations (non-uniformity) and temporal variations (unsteadiness) in PDC flow-boundary conditions, induced by changes in eruption dynamics and by interactions between the flow and the substrate topography. Many large volume ignimbrite sheets were emplaced over minutes to hours during the sustained collapse or

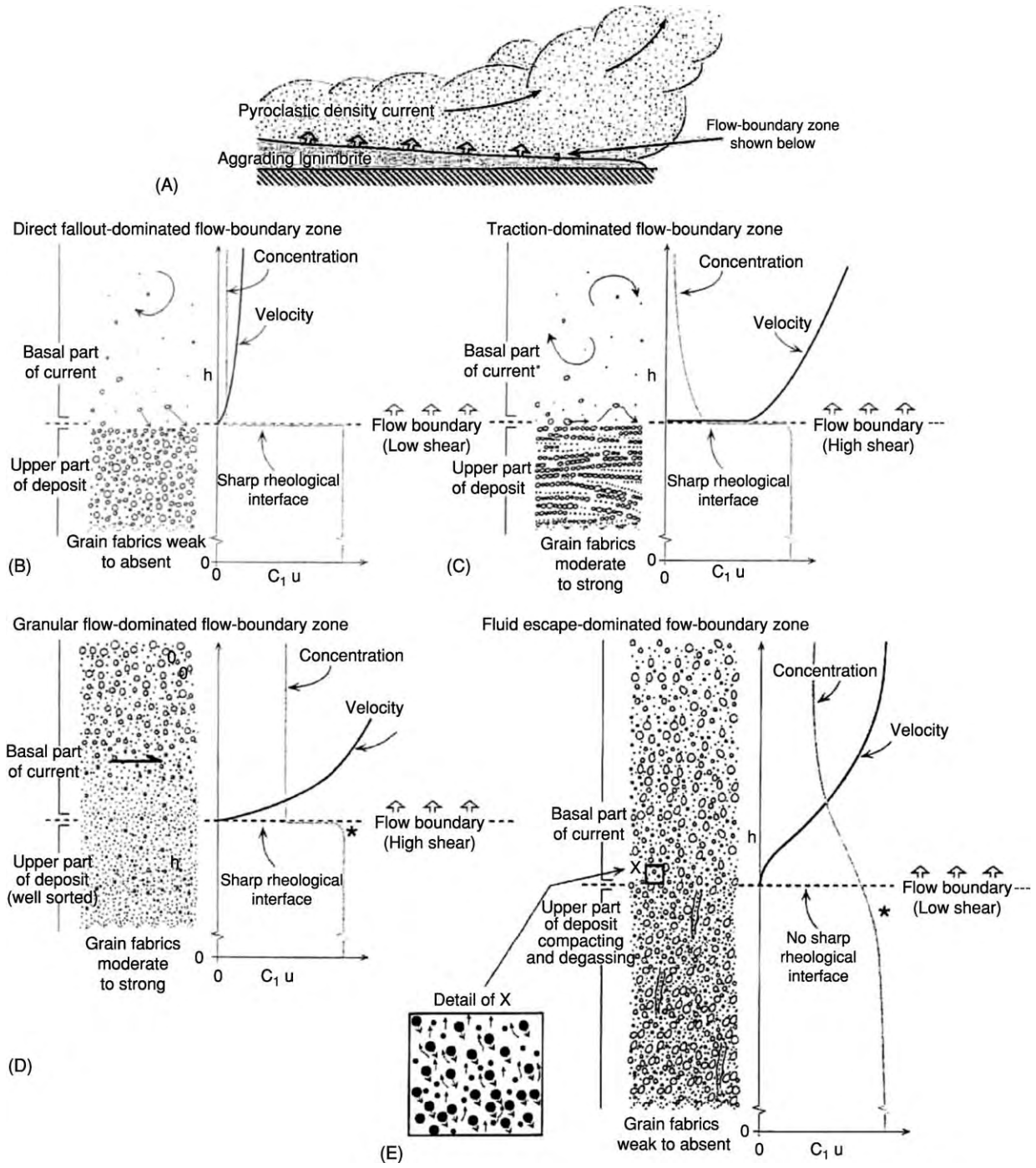


Figure 10 Four hypothetical types of pyroclastic density current flow boundary zones during steady conditions. (A) Scale and location of flow boundary zone emphasizing that depicted zones only include the basal part of the current and the uppermost part of the current. Most currents have flow boundaries intergradational between these types. (From Branney and Kokelaar (2002).)

fountaining of an eruption column. Sustained events are complex and can exhibit waxing and waning phases, pauses, and marked changes in eruption style. This can result in multiple flow-units and complex vertical lithofacies transitions and grading patterns. As the flows travel and deposit material, they substantially modify the substrate (e.g., by filling valleys, or by eroding loose material), and this, in turn, can affect sedimentation and flow boundary conditions up and downstream in subsequent parts of the current. Passage over uneven topography, changes in slope angle, and topographic blocking

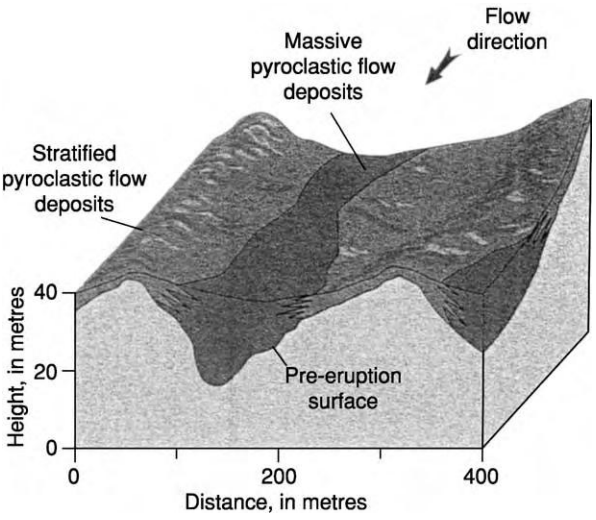


Figure 11 Lateral transitions from stratified ignimbrite (topographic highs) to massive ignimbrite (valleys) in the ignimbrite emplaced during the climactic 1991 eruption of Mount Pinatubo. (From Scott *et al.* in Newhall and Punongbayan (1996).)

(by mountain ranges), can all induce spatial changes in flow boundary conditions within a PDC and thus the deposit.

The dynamics of the small volume, end-member flows, such as those formed by fountain collapse during short-lived Vulcanian explosions and lava dome collapse are perhaps less controversial. Particle support mechanisms invoked for these small, relatively dense flows are commonly considered to be dominated by granular flow mechanisms (Figure 10D), where partial fluidisation may also play a role. The ash and gas components are considered important in friction reduction. Granular flow behaviour is controlled by a combination of the viscosity of the interstitial fluid and grain-to-grain interactions. In these dense small volume pyroclastic flows momentum is largely believed to be managed through grain-to-grain contacts. Low shear rate grain flows move as plug flows and become arrested en masse; high shear rate grain flows are smeared out during deposition as friction consumes momentum most strongly at the base while higher portions continue to move. Observations suggest separate mobile systems in which granular underflows or 'dense basal avalanches' occur beneath overriding turbulent ash clouds. The different properties of the two parts allow flow separation to occur, and the overlying ash cloud can move independently as a dilute pyroclastic density current.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards. **Igneous Processes**. **Lava**. **Volcanoes**.

Further Reading

Bonnadonna C, Ernst GGJ, and Sparks RSJ (1998) Thickness variations and volume estimates of tephra fall deposits: the importance of particle Reynolds number. *Journal of Volcanology and Geothermal Research* 81: 173 187.

Branney MJ and Kokelaar P (1992) A reappraisal of ignimbrite emplacement: progressive aggradation and changes from particulate to non particulate flow during emplacement of high grade ignimbrites. *Bulletin of Volcanology* 54: 504 520.

Branney MJ and Kokelaar P (2002) *Pyroclastic Density Currents and the Sedimentation of Ignimbrites*. London: Geological Society, London, Memoir 27.

Cas RAF and Wright JV (1987) *Volcanic Successions: Modern and Ancient*. London: Allen and Unwin.

Druitt TH (1998) Pyroclastic density currents. In: Gilbert JS and Sparks RSJ (eds.) *The Physics of Explosive Volcanic Eruptions*. Geological Society, London, Special Publications, 145, pp. 145 182.

Fisher RV and Schmincke H U (1985) *Pyroclastic Rocks*. Berlin: Springer.

Francis P and Oppenheimer C (2004) *Volcanoes*. Oxford: Oxford University Press.

Freundt A and Rosi M (1998) *From Magma to Tephra: Modelling Physical Processes of Explosive Volcanic Eruptions*. Amsterdam: Elsevier.

Gilbert JS and Sparks RSJ (1998) *The Physics of Explosive Volcanic Eruptions*. London: Geological Society, London, Special Publications, 145.

Newhall CG and Punongbayan RS (1996) *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*. Philippine Institute of Volcanology and Seismology and University of Washington Press.

Pyle DM (1989) The thickness, volume and grain size of tephra fall deposits. *Bulletin of Volcanology* 51: 1 15.

Sigurdsson H, Houghton BF, Rymer H, Stix J, and McNutt S (1999) *Encyclopedia of Volcanoes*. San Diego: Academic Press.

Sparks RSJ, Burnsik MI, Carey SN, *et al.* (1997) *Volcanic Plumes*. England, Chichester: Wiley.

QUARRYING

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

Quarries are resource extraction sites for removal of stone. The quarried stone is used as all manner of building materials, mainly as crushed-rock aggregate for concrete, road metal (crushed stone for road surfacing), asphaltic concrete pavement, and building construction stone. The stone is removed from 'pits' as cobbles, gravel, sand, and clay soil. Quarried stone is separated from outcrops along vertical working faces termed 'highwalls'. Highwalls are cut into a hillside of bedrock underlain at locations that have been selected to have little or no unconsolidated overburden (weathered rock, or 'soil'). Many quarries are developed as multisided rectangular excavations, and when the highwall exceeds a few tens of metres in height, the next level of extraction is 'benched' inward, leaving a stepped highwall, useful for control of inflowing groundwater. Where the host rock is not massive and may have some jointing dipping into the depression, bench heights are reduced to a general minimal height of 5–8 m. Stone removed from the quarry is hauled up a vehicle ramp for further processing at the company plant. Most stone is mechanically processed (broken and sized) to produce a size-graded product meeting engineering specifications of the buyer. When a quarry site is closed, environmental authorities typically require procedures to be followed as specified in a final closure plan; waste stone (spoil) is incorporated into the plan. Many urban-area quarries are closed out as disposal sites filled with environmentally inert construction and demolition wastes, mainly broken concrete and asphaltic concrete (Figure 1).

Creation of a geological quarry site model is essential for locating and proving quarry sites, in designing economical means for removing rock, and in devising environmentally sensitive ways for operation and closure. A number of geological factors and social/environmental considerations control the way in which quarries are located and operated:

- Stratigraphy and origin/mode of emplacement.
- Geomorphic setting.
- Postdepositional alteration of the host rock units and their rock fabric.
- Presence and orientation of discontinuities and potential layers of unacceptably altered weak rock.
- Nature and position of groundwater and management of its inflow, and ability to manage such water in terms of quarry operation and of mitigation of adverse impact on groundwater that is a neighbourhood resource.
- Ability to manage precipitation.
- Ability to manage rock extraction in such a way as to avoid instability of the highwalls, particularly in terms of quarry worker safety.
- Ability to manage waste rock (which typically amounts to about 5% of the extraction).
- Ability to operate in such a manner as to avoid danger to citizens and nearby residents, and to not degrade the environment.
- Development of a government-approved closure plan so that the closed quarry does not constitute a danger to the environment or to public health and safety.

Figure 1 Quarry operated for construction aggregate, road metal, and granular fill, in dolerite (diabase) and high grade crystalline metamorphic country rock of Prince George's County, Maryland, USA; the vertical face is approximately 15 m high. The pit has reached the lateral limits of the deposit of usable rock and also has come to the maximum viable depth for maintenance of the pit, removal of groundwater seepage, and precipitation runoff. The site was being considered for conversion to a long term secondary use compatible with environmental protection requirements. Photograph by the author.

Engineering Geology in Design and Operation of Quarries

Quarry development and operation has become a recognized field of practice requiring engineering geologists. As with many industries, quarry technology entered the machine age post-World War II, when excavation equipment became universally available to the family enterprises that owned most quarrying operations. Small operations are now rapidly disappearing and multinational companies control most of the stone that is mined today. In some instances, this equates to huge scales of magnitude (the 'superquarry'). Such are the economies of scale and the relatively low cost of bulk ocean transport that stone is imported over considerable distances, especially from quarries located close to coastal shipping.

Geological Factors Affecting Quarried Stone

Building stone is the traditional building material of mankind and still remains a critical construction material today, primarily as the aggregate, or bulking agent, for Portland cement and asphaltic concrete. Both cement and asphalt have been known to humans for several thousand years, yet their modern counterparts and their uses stem strictly from the industrial age.

Desirable Physical Properties of Quarried Stone

Quarried stone must have certain physical properties to serve as an effective building material. It must have strength, mainly compressional, to resist the load forces applied in engineered works; it must be durable and resistant to surface wear, especially when the use is for pavement; and it must be resistant to dynamic and chemical attacks, particularly from marine (saline) waters when the use is for port, harbor, and seacoast protective structures. Mined stone for building purposes must also be resistant to 'weathering' due to chemical attack by water and incident urban rainfall and must have consistency, be of uniform quality, and contain no deleterious minerals (e.g., pyrite, which causes staining of facing stone, and chert nodules, which cause eruptive pitting in concrete).

Desirable Mass Characteristics of Quarried Stone

Recovery cost for quarried stone is governed by several invariant cost factors inherent in the quarry site geology. The nature, pervasiveness, and orientation of discontinuities control the ease of extraction and the resultant stability of highwalls and benches. Stored tectonic stress may be a cost factor related to uneven responses to blasting patterns and explosive loading, due to release (relaxation) in strain-displacement relief ('pop-ups') of floor rock. The compressive strength of the quarry product, independent of geological structure, serves as a measure of the explosive force that will be needed (and thus the cost of explosives) to break the rock to requisite handling size.

Other Characteristics of Quarried Stone

Every stone quarry starts with an outcrop of durable stone or an excavated pit made in the overburden of surficial soils, often glacially deposited, so as to expose an artificial outcrop (working face) of stone. Once exposed, the stone must be subjected to efforts to dislodge it from the rock deposit. In most cases, the rock deposits selected for quarrying are cut by natural geological structural and depositional discontinuities such as bedding planes, foliation planes, and joints. Figure 2 depicts a rock face in which the supporting rock mass has three dominant joint 'sets', each representing a distinct array of joints, all of which will control the overall character and planar outline of a quarry developed in the rock mass.

Rock masses are virtually without strength in tension, and thus mechanical efforts to remove rock from a working face employ excavation machines to pull ('rip') discontinuity-bounded rock blocks from the quarry face (Table 1). Where the rock cannot be ripped, it is pulled from the working face by the judicial (planned) use of explosives set in geometrically arrayed and nominally vertical, drilled 'shot holes'. Discontinuities have value to the final product of the quarry only when the rock is destined to be recovered and sold as dimension stone, building stone, ornamental stone, or rip-rap. Rip-rap, used in shore protection works, is of great value when its quarry blocks are without further discontinuities, because the pieces can resist the constant pounding and surging action of ocean waves.

Many quarries also contain crystalline igneous and metamorphic rocks and the harder varieties of igneous and volcanic intrusive rocks. The many types of geological formations worldwide and their importance in quarrying operations are shown in Table 2.

Types of Rock Quarries

Quarries are established on the simple theme of locating and supplying a needed stone product at a competitive price in the free-enterprise system. There are some interesting relationships of quarry rock type and location worldwide (Table 2). Stone is intrinsically nearly valueless, and when obtained on a royalty

Figure 2 Rock highwall in rhyolite near Munsan, Republic of Korea, exhibiting three prominent joint sets of consequence to development of the site as a quarry. Such features should be taken into direct consideration in quarry dimensioning and in specification of the bearings of bounding highwalls; the horizon tal width of the exposed joint faces is about 5 m, left to right. Photograph by the author.

basis nets the landowner literally pennies per tonne. Thus the primary quarry costs are essentially those related to extraction machinery/explosives costs, labour costs, and haul costs (for transportation from quarry to buyer).

Planning Considerations

Planning, regulatory involvement, and engineered design are required to develop a new stone quarry. The overall effort (Table 3) is aimed at minimizing both the cost of production and the operational and postoperational impacts on the local environment. Proposed quarries, and even expansion of existing quarries, must go through a typical sequence of feasibility studies, engineered design reviews, permit applications, operational environmental compliance reviews, and eventual closure and postclosure land-use reviews in accordance with conditions established prior to initiation of operations.

Engineering Geological Inputs

In addition to the geological tasks that are required to open a quarry and begin quarry operations, engineering geologists are necessary to the continued operation of quarrying. A stone quarry is a dynamic venture in which previously unidentified geological situations develop on a regular basis. There is an ongoing need to identify, classify, describe, record, and update all quarry site data in a three-dimensional geological model. Regular auditing of product and spoil utilizes digitized geological data and geostatistical analyses of the potential reserves (and thus the ongoing viability of the operation). These measurements and computations generally are compiled on 30-cm contour intervals at lateral scales of 1:1000. Geologists are the ideal staff members for the entire spectrum of detection, evaluation, interpretation, recording, and notification responsibilities necessary to proper management of the site. Geologists are also the most logical choice to serve as quarry managers, officers, and corporate directors.

Environmental Aspects

The most prevalent environmental issues in quarry operations are related to nuisances experienced by or perceived by a site's 'neighbours' (Table 4). Typical concerns are the noise emitted by the crushing, washing, and sizing operations and the noise and traffic caused by stockpile movement and loading and departure of haul vehicles. Though these concerns are not geological in origin, monitoring the disruptive operations often is best handled by geologists because of their skills in observation, location, mapping, and recording. An important aspect of environmental compliance is blast vibration; monitoring utilizes portable seismometers, in order to meet permit obligations and to provide defence against claims for structural damage allegedly caused by transitory ground motions from production blasts.

Careers in the Quarrying Industry

Technical and professional geological aspects of stone quarrying are fascinating and offer a distinct challenge to the geologist interested in the day-to-day practical application of their scientific knowledge. Quarries live and die on optimization of costs of producing buyer-acceptable products, and in doing

Table 1 Geological characteristics relevant to stone quarrying

Characteristic	*Importance*	*Geological considerations*
Discontinuities	Provide natural rock break surfaces	Normally occur in predictable 'sets' of three dimensional orientation in space; usually two sets represent each historical tectonic deformation phase of the region around the quarry
Uniaxial compressive strength	An engineering property that basically relates to the ability of explosives to break the rock deposit at the quarry face, in the absence of appropriate discontinuities	Individual explosive arrays, employed at 'shots', serve to lift, displace, and move the drilled and loaded face mass of rock outward, into the quarry, in a pile on the quarry floor, for subsequent working and removal
Structural domains	Rock masses with a geological history of tectonic, intrusive, or geochemical alteration have a characteristic three dimensional variance in the orientation of discontinuities	Each 'domain' has definable geological contacts, within which the body of rock has predictable three dimensional attitudes of its discontinuities; these control the whole activity of 'pulling' rock at the quarry face
Fault blocks	Quarries operated in tectonically damaged or seismically active terrain often show block displacement of masses of quarry rock, incorporating 10^5 or 10^6 m^3 of stone in the fault bounded deposit selected for mining	Requires geological mapping and projection of geological structure, especially if the quality of fault displaced portions of the quarry mass have different engineering properties or characteristics
Perched water	Water of all sorts is a hindrance to quarry operation	Will drain in hours and days and can generally be engineered so as not to reoccur; where encountered in the quarry face, must be channeled to a floor sump and pumped free of the quarry
Groundwater	Rock masses generally contain some groundwater and this will 'bleed' into the quarry over its operational life	Place the quarry at the highest elevation of the groundwater shed and thereby limit the movement of groundwater towards the quarry, which will otherwise perform as an undesirable master sump
Dykes, sills, and other igneous intrusives	Undesirable in all respects; these features follow pre existing faults and shear zones and may foreshadow later discovery of fault block displacement of the quarry deposit; intrusions generally are associated with some degree of lithological damage from ancient geochemical alteration	Engineering geological exploration and ongoing operational mapping should be guided so as to detect and avoid these conditions, which represent a high probability of the presence of undesirable rock properties and/or characteristics detrimental to the value of the stone produced at the quarry

so in compliance with environmental controls. There is little room for error or mistake, because stone must be handled effectively the absolute minimal number of times and supplied on demand to the client. The industry is highly competitive in terms of price and reliability of product and delivery. Solution of potential and actual geological problems associated with production and environmental compliance is a high-value return based on employment of competent professional geological staff in quarry operations. Undergraduate geologists giving consideration towards career employment in the stone industry may thus be duly rewarded. In fact, any geologist experiencing family demands towards a reduction in travel and time away from home may well want to exploit the opportunities in the stone industry. Technical expertise, when blended with managerial capacities, tends to be appreciated by the new-breed of large-industry quarry operators.

Underpinnings in geology required for success in the technical aspects of quarrying include the ability to make field geological maps and to operate at mapping scales of less than 1:1000; sound knowledge of igneous, metamorphic, sedimentary, and volcanic petrology; skill at petrographic analysis; a fundamental knowledge of hydrogeology; an interest in environmental protection and the supporting regulatory, permitting, and compliance issues; and an understanding of the national standards governing stone quality as a material of construction.

Seeking and Gaining Stone Industry Employment

There are two pathways towards professional-grade employment in the geological engineering field related to quarrying; both routes ideally require an informed and caring university faculty. A direct approach to the industry can often be effected through faculty

Table 2 Regional generalizations

Location	*Generalization*	*Geological considerations*
Singapore	Bukit Timah Granite and Gombak Norite quarried in the central third of the island for all hard stone use	All pit and quarry space previously dedicated for follow up use as potable water storage in accordance with the Master Plan of the Republic
San Francisco peninsula, San Francisco, California, USA	All bedrock essentially weathered and/or altered ultrabasic subduction rock	Market requires long distance rail haul of suitable stone for aggregate
Los Angeles basin, Los Angeles, California, USA	Aggregate production mainly from coarse Pleistocene alluvial deposits; shortages made up from nearshore deposits of similar nature	Few quarries; mainly pits and these mainly converted to flood control storage in recognition of the endemic cloudburst nature of rainfall in the region; some utilized as infiltration sites for treated wastewater effluent as 'water factories'
Missoula, Montana, USA	Coarse postglacial displaced outwash deposits of the Clark Fork of the Columbia River	Crystalline igneous and metamorphic rock
Most of Nebraska, USA	Coarse Plio Pleistocene gravels of then rising Rocky Mountains	Crystalline igneous and metamorphic rock contaminated with noticeable content of vertebrate faunal bones
Lower Mississippi River valley, USA	Huge open pit quarries occupy the west bank of the Mississippi River, Missouri, below St. Louis; delivered by barge	Crushed Palaeozoic marine limestone of high carbonate content
Chicagoland: greater Chicago, Illinois, USA	Large, multisided quarries set below surrounding land surface	Palaeozoic dolomite; delivered by truck and rail
Greater Kansas City, Kansas, USA	Underground room and pillar limestone mines in dense, high strength Mississippian aged limestone	Modern mines engineered for long term stability and conversion to high value underground storage space
New York City/ Newark, New Jersey, USA	Triassic diabase (dolerite) quarries in New Jersey, with barge transport to New York City	High value for quarry end use as disposal sites for inert construction and demolition waste

and professional society contacts; the secondary pathway is to take quality control employment with transportation or environmental agencies and become involved, respectively, in the review of specification quality of quarried stone products or in the environmental permitting and compliance oversight of producing quarries.

Literature Support for the Quarry Geologist

Quarry geologists should act quickly to assemble a reference library of relevant published materials (*see* **Building Stone**). Though much of the literature is relatively old and therefore available primarily from dealers of used and rare books, older works are easily searchable on the Internet. Worldwide, the current standard literature staples are the outstanding books on aggregates compiled by the Engineering Working Group of The Geological Society, London (*see* **Aggregates**). Subscriptions to trade journals are also essential, as are library trips to recover helpful geologically oriented articles from the historical literature archives. Though stone recovery methods have improved vastly since the end of Word War II, the underlying scientific principles are timeless.

Quarry Restoration

With the advent of environmental awareness in the 1970s, quarries in First World nations generally became subject to closure requirements as an integral part of environmental and/or land-use permit regulations. These requirements are well in place both in the European Union, particularly in the United Kingdom, and in Canada and the United States. In most cases, the requirements have in common the following parameters:

- Mandatory termination of extraction activities at pre-established bounds, both lateral and vertical.
- Filing of a closure plan and a postclosure plan with a regulatory agency, including appropriate engineering details of the concept by which the site is to be made safe for residents of the surrounding area and by which precipitation and groundwater inflow are to be managed in an environmentally acceptable fashion, to include safeguarding against dumping of waste at the closed facility.
- Placement of closure and postclosure funds in an escrow account, with amounts and payments tied to quantity of rock extracted.

Table 3 Typical sequence of events

Event	*Action agency*	*Geological considerations*
Market feasibility study	Proposed operator	Nature, value, and cost to produce the end product(s)
Feasibility of prospective quarry	Proposed operator and financial partners	Estimated unit cost of product(s); cost to develop plant, as amortized over estimated quantity of stone reserves; estimated cost to close the site at termination of productive life; nature and value of end use of site
Development of topographic contour map	Consultant to proposed operator	Basis for planning of property geological exploration plan
Conceptual design of proposed quarry	Proposed operator, with assistance of consultants or corporate engineering staff	Location and bounds of the bedrock deposit; identification of potential geological constraints to operation and closure; management of groundwater and production spoil
Phase I geological exploration	Consultant to proposed owner or corporate engineering staff	Prove the existence of minimal reserve of suitable bedrock for production of stone product(s)
Tentative facility operational concept	Proposed operator, operator staff, or consultants to proposed operator	Identifies stages or phases of site development and product treatment; staging, groundwater, and surface water control features; and management of production spoil
Environmental impact report	By consultants to proposed quarry operator	Geological map of soil, bedrock, and groundwater conditions; hydrologic and hydrogeologic impact; control of dust, fugitive emissions, spoil, and runoff
Environmental impact statement	By state, county, or provincial public environmental agency	Identifies all physical and social environmental impacts and proposes appropriate mitigative measures; identifies and quantitifies risks to public health and welfare
Land use zoning application	Proposed owner	Identifies geological constraints related to nature land use, as such impact provisions of land use specified in the General Plan of the governmental administrative body; shows that proposed quarry is an acceptable land use in consequence of facility operational concept
Permit application	By proposed operator to the public agency holding regulatory control	Mainly equivalent to that of the environmental impact report
Operation, compliance, expansion, and/or closure	Proposed owner, staff, and consultants	All above noted geological considerations

Table 4 Major sources of environmental impact of quarrying

Impact	*Relevance*	*Geological considerations*
Alteration of groundwater regime	Quarries, as cuts and depressions, interrupt groundwater flow and create artificial flow gradients and discharge points/areas	Most of these 'interruptions' create, in fact, a 'shadow' of impact on the position and flow path of groundwater
Fugitive dust	Extraction of stone creates breakage and some comminution (size reduction), resulting in small particles subject to wind movement	Method of extraction, method of breakage, and method of size treatment for market product
Fugitive sediment	Results from dust caught up in the path of runoff	Reduction of amounts of dust and other small particles produced; methods of entrapment and sequestering from release to the environment
Loss of ground support to adjacent activities and structures	Slope instability could be induced here in the highwall, with the potential to ravel or migrate backward into adjacent land and buildings	Generally mitigated geologically by lateral buffer ground separating the highwall from adjacent activities and attention to the effects of adverse geological structure in the highwall rock (such as dips into the quarry)
Collateral damage from flying rock and ground motion of production blasting	Physical guidelines are well established for control[a]	Control related to empirical relationships of burden, ratio of explosive charge to mass of rock to be shot, and pattern and time delay of charges

[a]See guidelines of the United States Bureau of Mines, for example.

- An agreement on the postclosure land use for the property, including both the pit and highwall areas, the property roads, drainage, and the buffer zone surrounding the quarry.

In the twenty-first century, quarries have taken on a new postclosure value for several reasons, not the least of which includes the potential for new developable space following quarry backfilling with construction and demolition debris. Conversion to reservoir use for water supply storage, conversion to high-value residential use of the quarry rim when pit quarries are converted to recreational areas, and conversion to sanitary landfill use (if and where groundwater conditions and appropriate engineered design features can preclude environmental degradation) also are potentially valuable applications of former quarry sites. Conversion to some form of conservation use, often in agreement with local or regional public agencies, is another positive outcome for defunct stone mining operations.

See Also

Aggregates. **Building Stone**. **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills.

Further Reading

Department of the Environment, UK (1988) *Handbook on the Hydrogeology and Stability of Excavated Slopes in Quarries*. Geoffrey Walton Practice, Mining Geologists and Geotechnical Engineers. London: Her Majesty's Stationary Office.

Department of the Environment, UK (1988) *Technical Review of the Stability and Hydrogeology of Mineral Workings*. Geoffrey Walton Practice, Mining Geologists and Geotechnical Engineers. London: Her Majesty's Stationary Office.

Department of the Environment, UK (1992) *Landform Replication as a Technique for Reclamation of Limestone Quarries*. Limestone Research Group, Manchester Met ropolitan University. London: Her Majesty's Stationary Office.

Department of the Environment, UK (1992) *Revised Planning Guidelines for Extraction of Aggregates* (re places Master Planning Guidance Note 6). London: Her Majesty's Stationary Office.

Knill JL (ed.) (1978) *Industrial Geology:* Oxford, UK: Oxford University Press.

Latham J P (ed.) (1998) *Advances in Aggregates and Armourstone Evaluation: Geological Society of London. Special Publication 13*, London: Geological Society.

Lent FA (1925) *Trade Names and Descriptions of Marbles, Limestones, Sandstones, Granites and Other Building Stones Quarried in the United States, Canada and Other Countries*, New York: Stone Publishing Co. (ori ginally appearing in *Stone Magazine*, revised and corrected for the pamphlet).

Smith MR and Collis L (eds.) (2001) *Aggregates; Sand, Gravel and Crushed Rock Aggregates for Construction: Geological Society of London, Engineering Geology Special Publication No. 17,* 3rd edn. London: Geological Society.

United States Army (1967) *Pits and Quarries: Technical Manual TM 5 332*. Washington, DC: United States Government Printing Office.

REEFS

See **SEDIMENTARY ENVIRONMENTS: Reefs ('Build-Ups')**

REGIONAL METAMORPHISM

A Feenstra, GeoForschungsZentrum Potsdam, Potsdam, Germany
G Franz, Technische Universität Berlin, Berlin, Germany

Introduction

Regional metamorphic rocks form from other rocks (protoliths) by changes in mineralogy and texture in response to changing physical conditions (temperature, lithostatic pressure, and, in most cases, shear stress). These changes are essentially solid-state reactions, but very often a fluid phase is present, either participating in the reaction or as a reaction medium. Many regional metamorphic rocks have a chemical composition that is very similar to that of their sedimentary or igneous precursors, with the exception of removal or addition of volatiles (mainly H_2O and CO_2). This type of behaviour is termed 'isochemical metamorphism'. Metamorphism may also take place as a result of a change in chemical environment; this may occur by transport of elements between chemically contrasting rock types (e.g., formation of calc-silicate minerals at a quartzite–marble contact) or by circulation of fluids that dissolve some substances and precipitate others. This process of significant chemical change during metamorphism is known as 'allochemical metamorphism' or 'metasomatism', and rocks formed in this manner are metasomatic rocks. Metasomatism is, however, mostly of local significance, and the total volume of metasomatic rocks in regional metamorphic terranes is rather minor. The distinction between metasomatism and isochemical metamorphism is also a matter of scale. On the scale of individual grains, mass transport takes place during all phase transformations; on the scale of a thin section, it is probably the rule for regional metamorphism; on the scale of a hand (-sized) specimen, it can be observed frequently; and on a larger scale, it is the exception.

Regional metamorphism is nearly always accompanied by deformation and folding, which happens at converging plate boundaries and leads to orogenic belts. Therefore, regional metamorphic terranes follow the contacts between ancient lithospheric plates (suture zones), typically have an elongated shape that is parallel to the mountain belt, and cover areas of several 1000 to 10 000 km^2. Two basic types of regional (orogenic) metamorphism may be distinguished. The collision type relates to the collision of essentially continental plates and mainly produces metamorphic rocks at medium pressures and medium to high temperatures, often associated with partially molten (migmatites) and magmatic rocks. High- and ultra-high-pressure terranes ($P \approx 1.5$ GPa up to >3 GPa) from the deep roots of the collision zone can be preserved within the medium pressure (<1 GPa) belts. A variety of sediments from the continental margin and shelf (clay–shale, quartzofeldspathic rocks, limestone, marl), their metamorphosed equivalents, and acid magmatic rocks constitute the protolithic material of the collision type, whereas rock types of oceanic affinity are volumetrically unimportant. Typical examples include Mesozoic–Cenozoic metamorphic terranes in the Alps and Himalaya, Palaeozoic terranes in the Caledonides of Scotland and Scandinavia, as well as many Palaeozoic–Precambrian terranes in the shield areas of the continents.

The subduction type of regional metamorphism is associated with subduction of an oceanic plate underneath a continent and results in rocks formed at high pressure and relatively low temperatures (blueschist, eclogite). Basic igneous rocks (basalt, gabbro, volcanogenic rocks) together with clay-bearing clastic sediments are the dominant protolithic materials. Continental rocks may be included locally in the subduction-type terranes as well. Well-documented examples are found in the circum-Pacific region (Japan, California, New Caledonia, New Zealand), the Alps, and the eastern Mediterranean (Cycladic islands, Greece; south-western Turkey). In the case of a terminating oceanic plate, the oceanic subduction may be followed by a continent–continent collision, such that the two types may overlap each other in time.

Regional Metamorphic Processes

Large-scale geotectonic processes move rocks along particular paths in pressure–temperature–time (P–T–t) space. During movement along this path, the rocks re-equilibrate, and when chemical equilibrium is attained (when the reaction kinetics is fast enough), the mineral assemblage and the compositions of the individual minerals will react continuously by adapting to the new physical conditions. In addition to the mineralogy, the fabric of the rocks also changes, because regional metamorphism is a dynamic process typically accompanied by an anisotropic pressure field. Minerals with an elongated or platy habit will commonly align to the confining stress field. Mineral grains will also increase in size with increasing temperature, reducing the surface free energy (typically observed, for example, in metacarbonate rocks and quartzites). This process is termed 'static recrystallization'. Deformation in the anisotropic pressure field, however, reduces the grain size by various processes, depending on temperature and confining pressure; this is called 'dynamic recrystallization'.

Particularly at low-level metamorphic conditions, when deformation of the rock is non-penetrative, and in fluid-deficient rocks, equilibrium is often incomplete and minerals initially present in the protolith, and minerals formed during early metamorphic stages, may locally survive metastably. Such metastable persistence of minerals (e.g., as relict inclusions in newly formed minerals or as compositional zoning in minerals from core to rim) is fortunate for the petrologist in the way that it records the early geological history (the prograde path) of rocks. Petrological studies of metamorphic rocks traditionally aim to estimate the peak metamorphic conditions (pressure, temperature, and composition of coexisting fluid phase) that a rock has experienced, based on the assumption that the mineral assemblage essentially documents this peak equilibrium assemblage. Equilibrium among the minerals is commonly verified by textural evidence, i.e., that all minerals of the peak assemblage should be in mutual contact. Chemical evidence for equilibrium can be obtained from comparison of the composition of coexisting minerals in different bulk compositions; no crossing tie lines should be present and the tie lines (e.g., of garnet–biotite) should follow a systematic pattern. During the cooling and exhumation (the retrograde path), this peak metamorphic assemblage experiences only minor changes, because these processes are commonly associated with fluid-free conditions and minor deformation. A retrograde reaction, i.e., the formation of lower grade, mostly hydrous minerals from higher grade ones, thus typically occurs on a local scale. Examples include the replacement of garnet or biotite by chlorite or the replacement of plagioclase by calc-silicates such as (clino)zoisite, prehnite, or pumpellyite. Many occurrences of retrograde hydrous and carbonate phases are spatially connected with late (brittle) deformation, implying that deformation may facilitate local fluid introduction and thus formation of retrograde minerals.

In virtually all regional metamorphic terranes, mineral reactions are associated with deformation. Therefore, it is often possible to determine relative time relationships between growth of minerals and major deformation events characterized by development of a distinct foliation, folding, or shearing. Growth of a mineral can be specified as pre-, syn-, or post-tectonic with regard to specific deformation events. Relationships between mineral growth and deformation may be complex, however, because multiple or long-lasting deformation and mineral growth are the rule rather than the exception during orogenic cycles. Studies of regional metamorphic rocks conveniently separate the rocks' evolution into the three categories of prograde, peak, and retrograde stages using petrological and structural criteria. The researcher should be aware, however, that this is a simplified picture and the P–T–t paths of metamorphic rocks are in general quite complicated.

Regional Metamorphic Zones and Facies

It was recognized very early that metamorphic rocks can be classified according to different 'intensities' of metamorphism, or metamorphic grades. The concept of gradation is not applicable to the other families of sedimentary or magmatic rocks; a magmatic rock cannot be 'more magmatic' than another one, whereas a metamorphic rock can be metamorphosed more intensely. This concept also emphasizes that the processing of a sediment through diagenesis, then forming a metamorphic rock, is transitional, and it therefore is impossible to mark an unequivocal beginning of metamorphism (*see* **Diagenesis, Overview**).

The concept of metamorphic index minerals typical of metamorphic grade was first introduced by George Barrow in the south-east Highlands of Scotland in the late nineteenth and early twentieth centuries. Studying a sequence of metamorphosed clay-rich sediments, Barrow related the systematic change in mineralogy in these metapelitic rocks to changing metamorphic conditions. Barrow recognized chlorite, biotite, garnet, staurolite, kyanite, and sillimanite zones as representing increasing metamorphic grades

that are characterized by index minerals not present at lower grades. 'Mineral-appearance' isograds were mapped by drawing lines between locations where each index mineral first appeared. It later became common practice to replace these classic mineral isograds by 'reaction' isograds related to a particular metamorphic reaction and to use the equilibrium assemblage of that reaction as a zone boundary. The reaction isograd is defined by assemblages of reactant(s) on one side, of product(s) on the other side, and, if present, the equilibrium assemblage on the isograd line (surface).

It should be emphasized that in metamorphic petrology, an isograd is not exactly a line (surface) of constant temperature (and pressure). Various aspects of metamorphic reactions should be considered. (1) In cases in which minerals with variable chemical compositions (solid solutions) participate in the isograd reaction, the position of the isograd in pressure–temperature (P–T) space may depend on mineral compositions (and the bulk chemical composition of the rock). (2) The P–T position of reactions producing or consuming volatiles (H_2O, CO_2) may also depend strongly on the composition of the metamorphic fluid. The reaction kinetics may vary on a local or regional scale in metamorphic terranes, especially at the lower metamorphic grades. These aspects imply that mapped reaction isograds will only approximately reflect lines (surfaces) under identical peak temperature (and pressure) conditions.

Following the pioneering studies of Goldschmidt in southern Norway, in 1920, Pentii Eskola developed the concept of metamorphic facies. Studying mafic metamorphic rocks (mostly basaltic protoliths) in southern Finland, he recognized that "in any rock or metamorphic formation which has arrived at a chemical equilibrium through metamorphism at constant temperature and pressure conditions, the mineral composition is controlled only by the chemical composition". This concept has now been validated worldwide in innumerable regional field studies demonstrating that many rock compositions develop mineral assemblages typical of the physical conditions of formation. For a certain bulk composition, the mineral assemblages formed at these different metamorphic grades are then related by mineral reactions. Experimental determinations of the relative stabilities of metamorphic minerals and assemblages, followed by thermodynamic calculation of such phase equilibria, have allowed an approximate quantification of the pressure and temperature conditions of each metamorphic facies.

The metamorphic facies diagram shown in Figure 1 differs in some details from that proposed by Eskola, who named his metamorphic facies after

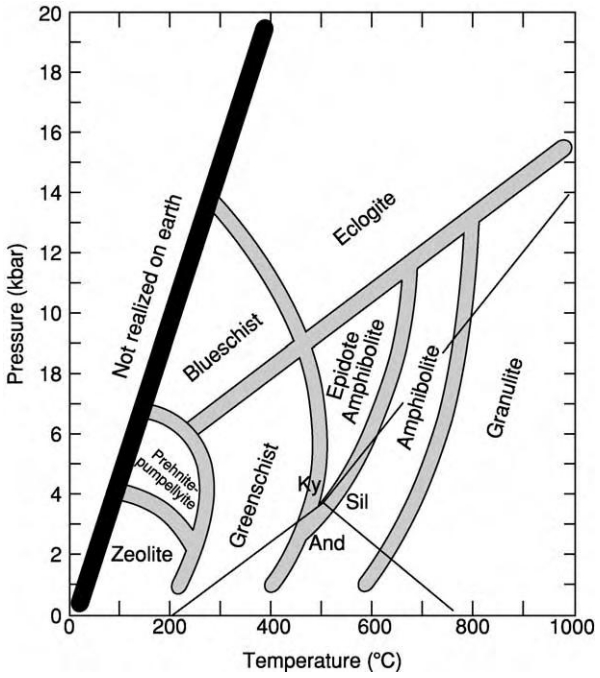


Figure 1 Pressure temperature diagram, showing principal metamorphic facies; approximate boundaries are represented by thick grey lines. Thin black lines designate the stability fields of the three Al_2SiO_5 polymorphs (And, andalusite; Ky, kyanite; Sil, sillimanite). Reproduced with permission from Spear FS (1993) *Metamorphic Phase Equilibria and Pressure Temperature Time Paths*. Washington, DC: Mineralogical Society of America.

metamorphosed mafic rocks. Nowadays, differences in facies classification exist mainly in the transition from greenschist facies to amphibolite facies at higher pressure; some authors prefer, and others do not, to introduce an epidote–amphibolite facies (Figure 1). In addition, the classification of the facies below the greenschist facies is summarized by several authors as 'subgreenschist facies'. In any case, the borders between different facies are not sharp, and the classification should be used in a broad sense. It is schematic, but there is consensus on the temperature and pressure limits for each facies. The boundaries between the facies represent the P–T conditions at which key minerals or mineral assemblages appear or disappear.

Extensive petrological studies were carried out in metamorphic terranes around the world during the twentieth century. These studies revealed that distinct metamorphic facies series (high, medium, and low pressure and very low grade) correlating with specific geotectonic settings could broadly be distinguished.

Facies of High Pressure

High-pressure rocks of the blueschist and/or eclogite facies were metamorphosed at a low geothermal gradient, i.e., small increase in temperature with

depth (Figure 1). They are typically formed during subduction of a cold oceanic plate underneath a continent. Mafic protoliths best document such high-pressure conditions by their diagnostic mineral assemblages in blueschist (sodic amphibole bearing; the term 'blueschist' derives from the blue sodic amphibole, glaucophane) and eclogite (dominated by garnet + omphacitic clinopyroxene). Metapelitic rock compositions may develop diagnostic minerals or assemblages such as carpholite, phengite, jadeite + quartz, and talc + kyanite.

Facies of Medium Pressure

Most metamorphic terranes, especially in the internal parts of continents, are composed of greenschist, amphibolite, and granulite facies rocks, including migmatites at the high metamorphic grades. Such a facies series is related to a geothermal gradient of 20–35°C km $^{-1}$, typical of the Barrovian-type metamorphism as defined by Barrow in south-east Scotland. A well-documented example of a Barrovian-type metamorphic zoning is exposed on the island of Naxos, Greece (see later). As already discussed, medium-pressure regional metamorphic zones are best developed in Al-rich metapelitic rocks, which may contain either kyanite or sillimanite, or both Al-silicates. The formation of staurolite and the breakdown of chloritoid are generally considered to reflect the greenschist–amphibolite facies transition in pelitic rocks. In mafic rocks, the greenschist–amphibolite facies transition is designated by the appearance of hornblende and Ca-rich plagioclase and the disappearance of actinolite and Na-rich plagioclase. In metapelites, the amphibolite–granulite transition is considered by some workers to be indicated by the assemblage K-feldspar + Al-silicate forming at the expense of muscovite + quartz. Other workers consider the formation of orthopyroxene to be typical of granulite facies conditions, although this occurs at significantly higher temperatures (100–150°C higher) than does formation of K-feldspar + Al-silicate. In mafic rocks, the presence of orthopyroxene, which forms at the expense of hornblende and/or biotite, is considered indicative of granulite facies conditions. The role of water activity, however, complicates characterization of the amphibolite–granulite facies transition. If rocks are water saturated, they will, with a eutectic or eutectoid chemical composition near to granite, partially melt at P–T conditions in the muscovite + quartz stability field. When water activity is low, pervasive partial (dry) melting will occur at much higher temperatures and hydrous minerals (muscovite, biotite, and amphibole) may decompose to anhydrous ones by local dehydration melting.

Facies of Low Pressure

Low-pressure facies are in many cases typical for contact metamorphism (*see* **Thermal Metamorphism**), but especially in cordilleran-type mountain belts this contact metamorphism can be of regional extent due to long-lasting intrusive activity over a large area, producing a transition to regional metamorphism (Buchan-type in Scotland, Abukuma-type in Japan). It also may follow the medium-pressure facies. Metapelites are characterized by andalusite and cordierite.

Facies of (Very) Low Grade

Facies of very low grade are present in the transition zones of an orogenic mountain belt to the sedimentary foreland (e.g., Swiss Alps), but are often missing in regional metamorphic terranes, where the metamorphic rocks are overthrusted onto the foreland. They are typically best developed in regions with burial or hydrothermal metamorphism (e.g., New Zealand, Central Andes).

The Barrovian-Type Metamorphic Complex of Naxos

The island of Naxos is part of the Attic–Cycladic Metamorphic Complex that stretches from mainland Greece to south-western Turkey. Isoclinally folded sequences of metasediments (pelitic and psammitic schists and gneisses, quartzites, and calcitic and dolomitic marbles) and metavolcanics (amphibolites and mafic schists) are the dominant lithologies on Naxos (Figure 2). Ultramafic rocks occur in a few stratigraphic horizons and as isolated lenses. Metacarbonate units are widespread and contain in certain stratigraphic horizons diaspore and/or corundum-bearing lenses (up to ≈8 m thick; Figure 2), for which protoliths are bauxite formed in a karst soil. These metabauxite occurrences offer a unique opportunity to investigate the prograde metamorphism of a rock composition with excess Al.

The metamorphic pattern of Naxos was shaped during a late Alpine medium-pressure event that culminated 15–20 Ma; it consists of concentric zones of decreasing metamorphic grade, outward from a high-grade migmatitic core (Figure 2; Table 1). In the western part of Naxos, a ≈13-My-old granodiorite intruded the metamorphic complex and induced andalusite–sillimanite-type contact metamorphism in about a 1-km-wide zone. In this same region of Naxos, a post-intrusive upper unit, comprising Tertiary sediments and ophiolite-suite rocks, tectonically overlies the metamorphic complex and granodiorite. Petrological and geochronological studies indicate that relict

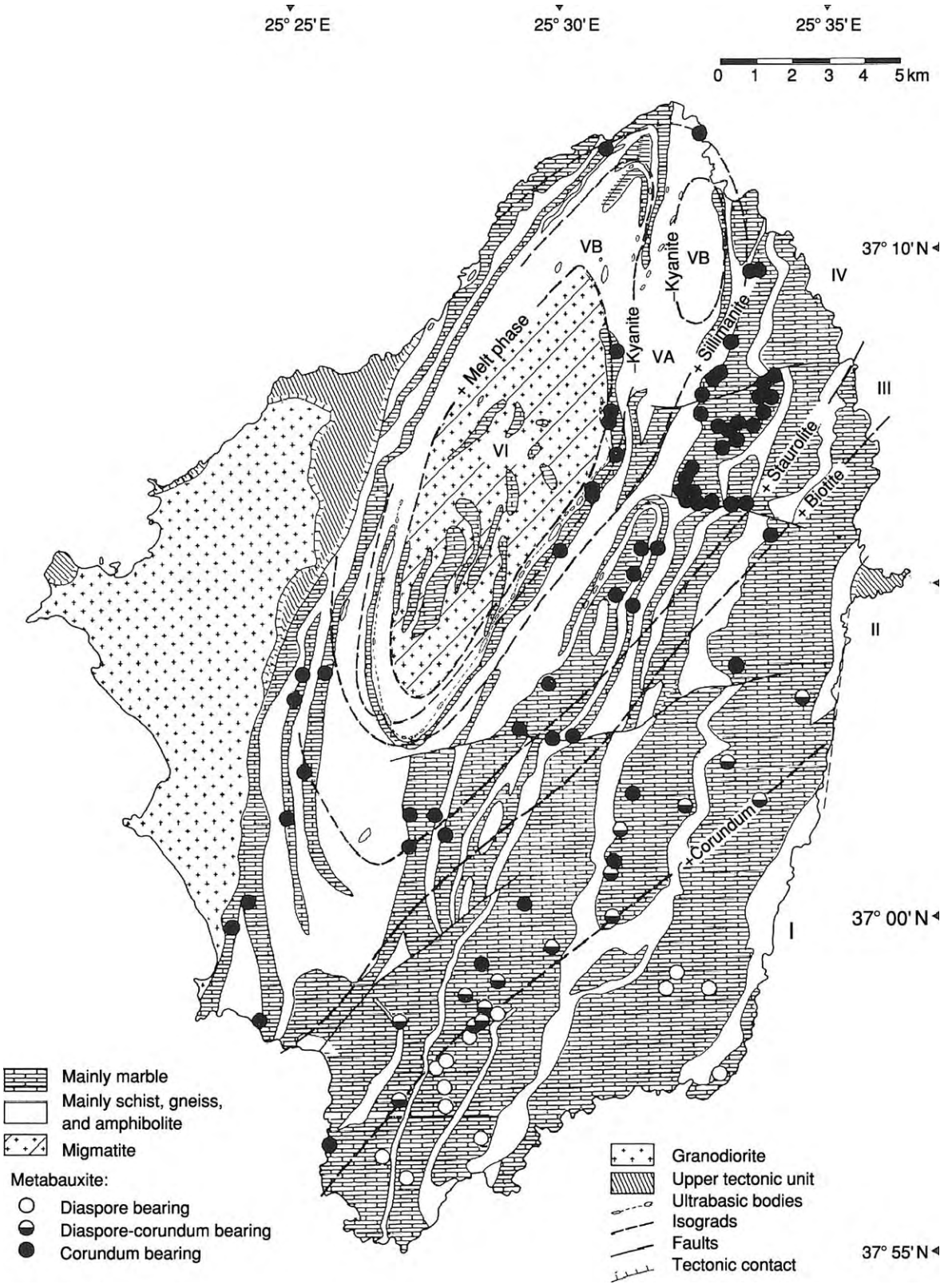


Figure 2 Geological metamorphic map of Naxos (Cycladic Islands, Greece), showing main lithologies, isograds, and metamorphic zones (I VI). The following isograds were mapped with increasing grade: corundum in (metabauxites); biotite in (metapelites); Fe rich staurolite in (metabauxites and metapelites); sillimanite in (metapelites); kyanite out (metapelites); melt phase in (metapelites). (Adapted from Feenstra A (1985).)

Table 1 Mineral assemblages as a function of metamorphic grade in different rock types in the metamorphic complex of Naxos, Greece. (Data from Jansen and Shviling (1976) and Feenstra (1985))

Metamorphic zone	*Pelitic (+Qtz)*	*Mafic*	*Utramafic*	*Carbonate (+Cal ± Dol)*	*Bauxitic*
I	Ab + Ms + Chl ± Ctd ± Pg ± Ep	Ab + Ep + Act + Chl + Ms + Hem	Srp + Chl + Tlc	Ab + Chl + Ep/Zo + Ms + Hem + Qtz	Dsp + Cld + Hem + Rt ± Prl ± Ky ± Ms ± Pg ± Cal
+Corundum					
II	Ab + Ms + Chl ± Ctd ± Pg ± Ep	Ab + Ep + Act + Chl + Ms + Hem ± Bt	Srp + Chl + Tlc ± Phl	Ab + Chl + Ep/Zo + Phl + Ms + Tlc + Hem + Qtz	Crn + Cld + Hem + Rt ± Ky ± Ms ± Pg ± Mrg
+Biotite					
III	Pl + Bt + Ms ± Chl ± Ctd ± Grt ± Pg ± Ep	Pl + Ep + Act + Chl + Bt + Ms + Hem + Mgt	Chl + Tlc + Phl ± Srp	Pl + Chl + Ep/Zo + Phl + Ms + Tlc + Hem + Mgt ± Qtz	Crn + Cld + Hem + Mgt ± Ky ± Ms ± Pg ± Mrg
+Staurolite					
IV	Pl + Bt + Ms + St + Ky + Grt ± Chl ± Ep	Pl + Ep + Hbl + Act + Bt + Mgt ± Chl ± Ms	Chl + Tlc + Ol + Ath + Mgs + Phl	Pl + Ep/Zo + Phl + Tlc + Tr + Mgt ± Grs ± Qtz	Crn + St + Hem/Ilm + Mgt + Ky + Mrg ± An ± Ms ± Bt
+Sillimanite					
VA	Pl + Bt + Ms + St + Ky + Sil + Grt	Pl + Ep + Hbl + Bt + Di + Mgt ± Chl ± Ms	Chl + Tlc + Ol + En + Ath + Phl	Pl + Ep/Zo + Phl + Tr + Di + Mgt ± Grs ± Scp ± Ves	Crn + St + Ilm + Mgt + Ky + An ± Mrg ± Ms ± Bt
Kyanite					
VB	Pl + Bt + Ms + Kfs + Sil + Grt ± St	Pl + Hbl + Bt + Di + Mgt + Grt ± Ms ± Ep	Ol + En + Spl + Ath + Phl + Di ± Tlc	Pl + Phl + Tr + Di + Mgt ± Grs ± Scp ± Ves ± Ep/Zo	Crn + Spl + Ilm + Mgt + An ± Ms ± Bt
+Melt phase					
VI	Pl + Bt + Kfs + Sil + Grt	Pl + Hbl + Bt + Di + Grt + Mgt	Not present	Pl + Phl + Tr + Di + Mgt ± Grs ± Scp ± Ves	Not present

Mineral abbreviations: Ab, albite; Act, actinolite; An, anorthite; Ath, anthophyllite; Bt, biotite; Cal, calcite; Chl, chlorite; Cld, chloritoid; Crn, corundum; Di, diopside; Dol, dolomite; Dsp, diaspore; En, enstatite; Ep, epidote; Grs, grossular; Grt, garnet; Hbl, hornblende; Hem, hematite; Ilm, ilmenite; Kfs, K feldspar; Ky, kyanite; Mgs, magnesite; Mgt, magnetite; Mrg, margarite; Ms, muscovite; Ol, olivine; Pg, paragonite; Phl, phlogopite; Pl, plagioclase; Prl, pyrophyllite; Qtz, quartz; Rt, rutile; Scp, scapolite; Sil, sillimanite; Spl, spinel; Srp, serpentine; St, staurolite; Tlc, talc; Tr, tremolite; Ves, vesuvianite; Zo, zoisite.

mineral assemblages of an early Alpine high-pressure/low-temperature metamorphism (about 45 Ma) locally occur in south-eastern Naxos (zones I–II) and that part of the migmatite core (zone VI) is of pre-Alpine origin (from ≈300 Ma).

Temperatures during the Barrovian-type metamorphism are estimated to have ranged from ≈700°C in the migmatite dome to ≈400°C in south-eastern Naxos at pressures of 0.5–0.7 GPa. Six isograds mapped in metapelitic and metabauxitic rocks divide the metamorphic complex into seven Barrovian-type metamorphic zones (Figure 2; Table 1). The first three zones (I–III) belong to the greenschist facies: their subdivision is based on the first occurrence of corundum, by the reaction diaspore = corundum + H_2O (in metabauxites), and of biotite, by the reaction chlorite + muscovite = biotite + H_2O (in metapelites). In mafic rocks, biotite already occurs locally in zone II, whereas the first biotite in bauxitic rocks appears in zone IV.

The presence of staurolite, both in pelitic and bauxitic rocks, indicates the beginning of amphibolite facies conditions. Staurolite formation is related to complex breakdown reactions of chloritoid and approximately coincides with the disappearance of chloritoid. Zones IV, VA, and VB belong to the amphibolite facies; whether zone VI, the beginning of which is defined by partial melting of metapelites, is amphibolite or granulite facies is a matter of debate. The subdivision of the amphibolite facies zones of Naxos is based on the appearance of sillimanite and the disappearance of kyanite in metapelitic rocks. The Naxos rocks clearly illustrate that bulk composition controls the formation of particular minerals or assemblages. In common with other Barrovian terranes, the first kyanite in metapelitic rocks is found in staurolite-grade zone IV, whereas kyanite in bauxitic rocks is already formed in zone I by the reaction pyrophyllite + diaspore = kyanite + H_2O (Table 1).

The upper amphibolite or granulite facies zone VI contains areas where partial melting of SiO_2-rich rocks was intensive (melt phase has a granitic composition) and areas showing little evidence of melting. Variations in the degree of migmatization may relate to local variations in water activity and rock composition. The typical mineral assemblage in metapelitic composition is K-feldspar + sillimanite + biotite. Muscovite often occurs, but textural evidence suggests it is mostly of retrograde origin. Mafic rocks are devoid of orthopyroxene but the presence of sapphirine + spinel + corundum at a few localities in zone VI argues for granulite facies grade conditions.

See Also

Diagenesis, Overview. **Metamorphic Rocks:** Classification, Nomenclature and Formation; Facies and Zones. **Minerals:** Other Silicates. **Plate Tectonics**. **Tectonics:** Mountain Building and Orogeny. **Thermal Metamorphism**. **Ultra High Pressure Metamorphism**.

Further Reading

Barrow G (1893) On an intrusion of muscovite biotite gneiss in the S.E. Highlands of Scotland and its ac companying metamorphism. *Quarterly Journal of the Geological Society, London* 49: 330 358.

Barrow G (1912) On the geology of lower Deeside and the southern highland border. *Proceedings of the Geologist' Association* 23: 268 284.

Bucher K and Frey M (2001) *Petrogenesis of Metamorphic Rocks,* 7th edn. Berlin Heidelberg New York: Springer Verlag.

Eskola P (1920) The mineral facies of rocks. *Norsk Geolo gische Tidsskrift* 6: 143 194.

Feenstra A (1985) *Metamorphism of Bauxites on Naxos, Greece*. PhD thesis, Rijksuniversiteit Utrecht, Geologica Ultraiectina 39.

Jansen JBH and Schuiling RD (1976) Metamorphism on Naxos: Petrology and geothermal gradients. *American Journal of Science* 276: 1225 1253.

Spear FS (1993) *Metamorphic Phase Equilibria and Pres sure Temperature Time Paths*. Washington DC: Miner alogical Society of America.

Winter JD (2001) *An Introduction to Igneous and Meta morphic Petrology*. Upper Saddle River, NJ: Prentice Hall.

Yardley BWD (1989) *An Introduction to Metamorphic Petrology*. Burnt Mill, Harlow, Essex: Longman Group UK Limited.

REMOTE SENSING

Contents

Active Sensors

G Wadge, University of Reading, Reading, UK

Introduction

The amount of solar radiation reflected from rocks at longer wavelengths (>1 mm) is very small. To properly exploit this microwave part of the electromagnetic spectrum for remote sensing, artificial sources of radiation are used. These artificial, or 'active' techniques allow measurements to be made at night as well as during the day. The longer wavelength of microwave radiation compared to that in the optical part of the spectrum also makes clouds transparent, allowing microwave radiation to 'see' the ground surface in all weathers, day and night.

In radar (Radio Detection and Ranging) remote sensing, a series of timed pulses of microwave energy are transmitted through the air, strike targets (e.g., the ground) and are reflected back to the radar which receives the returned signals. There are parallels between radar and seismic exploration techniques. The principle of 'ranging' is common to all active sensors in remote sensing described here. Because the energy source is under the control of the instrument designer, the frequency and wavelength of the radiation is chosen with a very narrow range so that it can be recognized more easily on reflection and its arrival timed accurately. Knowing the two-way (to the ground and back to the sensor) travel time (t) and the speed of travel through the medium (c), the distance or range (R) to the ground is given by:

$$R = ct/2 \qquad [1]$$

After briefly reviewing the three allied 'active' techniques of radar, lidar, and sonar, attention is focused on radar and the basic principles behind its use and interpretation. There are three main application areas relevant to geology: structural and geomorphological mapping, exploitation of the ability to measure fine surface roughness, and the measurement of surface motions including those associated with earthquakes (*see* **Tectonics:** Earthquakes) and volcanoes (*see* **Volcanoes**).

Radars, Lidars and Sonars

Radars send pulses of microwave radiation (wavelength usually 3–30 cm) from antennas through the atmosphere to the ground, in our case, and receive the return signal. Lidars do the same but with optical wavelength light (800–1000 nm) from lasers. Lidars do not work through cloud, as radars do. Sonars are the oceanographic equivalent to radars that send pulses of energy (1–15 cm wavelength) through the water column to strike the seabed. Figure 1 shows the typical sensing arrangement of the three techniques. Radars have made the most impact on geology.

Lidars have been relatively little used in geological applications to date but the technology is newer. Usually flown from low flying aircraft, they are used as a simple track or swath altimeter with a small footprint. With their high vertical (~10 cm) and horizontal (~1–2 m) resolutions, detailed digital elevation models (DEMs) can be created. They will become increasingly useful to studies requiring detailed geometric knowledge of the terrain, such as process-based simulations.

Sonars have a much longer history of sea-floor mapping. The sensors are towed behind a ship at various levels above the seabed. Absolute positioning is a more involved process than for space-borne sensors. The range information received by the sonar is used to form bathymetric maps with accuracies of a few tens of metres in some cases. As with radar, the magnitude information is a function of the geometrical relationship of the sensor, and the seabed and the surface roughness. Particularly valuable results have been achieved at spreading ridges, accretionary prisms, and in mapping the apron of gravity collapse deposits around oceanic islands.

Radar Techniques

Altimetry

Radar altimeters are the simplest form of radar. They are usually simple ranging devices recording the distance from the radar to the ground, pointing vertically downwards from a satellite. Their footprint on the Earth's surface is large (kilometre scale) and as a result their profiling measurements are only easily interpretable on surfaces that are relatively flat at this length scale (e.g., oceans and ice sheets). Their importance to geology has been that they have provided a measure of the shape of the geoid in the ocean basins that reflects the dynamic processes of plate tectonics (*see* **Europe:** Mediterranean Tectonics): subduction and spreading ridges.

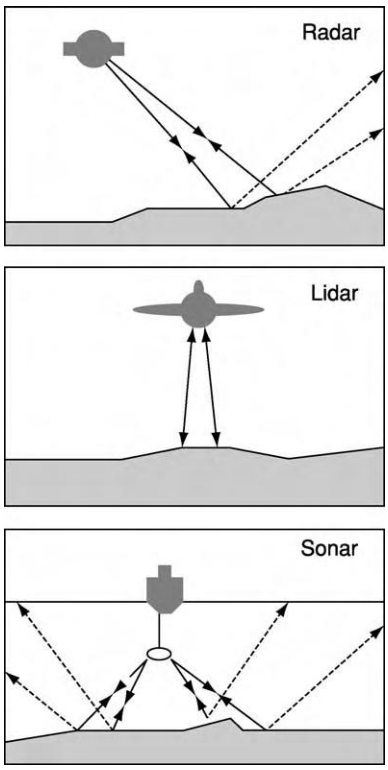


Figure 1 Schematic of the typical operating geometries of a spaceborne SAR, an airborne lidar and a multi swath sonar.

Doppler Radar

In an active remote sensing device, the engineer can control the rate of pulsing of signals transmitted to the surface and the range of frequencies of radiation transmitted and received. This means that the measurements can remain 'coherent' from one measurement time to another, provided that the external conditions remain the same. Such a coherent radar signal has three measurable quantities: timing, magnitude, and phase. Any relative movement between the radar antenna and the reflecting surface from one pulse to another can be detected as a change in the frequency of return of these pulses – a Doppler shift. The size of this shift allows the relative motion to be calculated. Doppler radars of this type have been designed to measure the dynamics of volcanic ash columns. Lower frequencies of radar signal are more sensitive to larger fragments. Using multiple frequency bands in the same instrument, it is possible to measure the relative motions of different particle size fractions. Because many ash eruption columns are usually obscured by fine turbulent cloud, Doppler radars offer the only practicable way of seeing through this and measuring the internal dynamics of the column.

Imaging Radar

The majority of Earth science applications of radar employ imaging techniques. Table 1 summarizes the characteristics of the main space-borne radar imaging systems to date and those planned in the next decade. In addition there are a number of airborne systems and ground-based systems. Future development of more compact systems and the exploitation

Table 1 Selected orbital SAR systems

	Seasat	*ERS 1/ 2*	*JERS 1*	*Radarsat*	*Envisat*	*Alos Palsar*	*Terrasar X*	*EVINSAR*[a]
Operation	1978 (3 months)	1991 2000 1995	1992 1998	1995	2002	2004	2007	2009
Country	USA	Europe	Japan	Canada	Europe	Japan	Germany	Europe
Frequency band	L	C	L	C	C	L	X	L
Repeat period (days)	17, 3	35 (1, 3, 176)	44	24	35	44	11	11
Incidence angle	23	23	26 41	20 60	20 50	8 60	20 55	30 46
Swath width (km)	100	100	85	10 500	100 485	70 350	10 100	250
Resolution (m)	25	28	18	10 100	25	10 100	1 16	45

[a]Proposed InSAR mission dedicated to ground motion measurement.

of the millimetre wave part of the spectrum will add considerably to these capabilities.

The motion of a plane or a satellite relative to the Earth's surface can be used to build up a 2D image of the radiation returning from its reflections, or backscatter. Such imaging radars are usually pointed to one side of the platform (**Figure 1**). This removes the ambiguity of signals returning from equal distances to the left and right of a downward pointing instrument. Most imaging radars are Synthetic Aperture Radars (SARs). In these instruments, the Doppler frequency information from radar returns received ahead of the moving radar (positive shift) relative to those behind (negative shift), are used to 'synthesize' electronically a larger antenna size. This is needed to improve the resolving power of the radar in the along-track or azimuth direction. The details of radar image formation and processing are complicated, but the resulting images essentially contain data on both the magnitude or amplitude and the phase (ϕ) of the returned signal reflected from the ground. Three main factors control these quantities in a radar image: the wavelength of the radar, the slope, orientation and roughness of the reflecting surface, and the dielectric properties of the surface.

The radar image amplitude is generally greater if the surface slope is close to perpendicular to the radar line-of-sight (LOS), which is often several tens of degrees from the vertical. It is also greater if the surface is 'rough' and if it is 'wet' (water content is the main determinant of dielectric behaviour). Within the instantaneous field of view of the radar there may be several individual surfaces or edges (e.g., rock facets) that return much stronger signals than intervening areas. Because the radiation is coherent, these strong signals may reinforce or destroy each other, depending on their relative positions. Depending on the arbitrary location of these strong reflectors on the surface then one radar pixel may have a high amplitude whilst its neighbour, comprising very similar surface material, has a low amplitude. This gives radar amplitude images their 'speckle' character. It also means that, in general, a meaningful measure cannot be retrieved of the surface properties from a single pixel of an amplitude image (unlike a passively acquired optical image). Instead a spatially averaged measure of backscatter amplitude is needed. A similar conclusion applies to the phase information measured. The cumulative sum of phase values from an individual pixel gives a cyclic angular measure ($0 < \phi < 2\pi$) that has no intrinsic significance. However, these values are not noise and the radar should retrieve exactly the same values on re-measurement. These phase data are used by taking temporal differences of phase in the technique of radar interferometry.

Radar Applications

Structural/Geomorphological Mapping

The ability to image large contiguous areas through cloud from airborne radars was first demonstrated in the 1960s. It enabled the first hundred kilometre scale reconnaissance structural mapping of cloud covered areas in the humid tropics. Slopes nearly perpendicular to the radar LOS appear bright in amplitude images and slopes away from this appear dark. This light and shade effect is the same as that seen in optical imagery. However, radar amplitude images suffer from distorting effects that complicate qualitative interpretation. Radar-facing slopes are foreshortened, whilst those facing away are extended. At very steep slopes in alpine-type terrane the foreshortening becomes 'layover' when the tops of peaks appear to overhang their valleys. Most of these distortions can be corrected quantitatively if a DEM is available. Suitably detailed surveys of several areas, particularly young fold and thrust belts in cloudy regions, have proven genuine exploration aids. The sensitivity of a radar amplitude image to a fault-controlled topography or any other geomorphological features is always a function of the angular relationship between the radar LOS and the feature on the surface. Features parallel to the LOS are hard to see. Airborne surveys can be designed with this in mind. The erosional morphology of different lithological terrains is also well expressed in radar amplitude imagery, permitting lithological discrimination (**Figure 2**). This is, of course, scale dependent, particularly where there is tree cover obscuring ground features less than about 100 m in size. With the increasing availability of high resolution DEMs for much of the globe then the value of imaging SAR at this reconnaissance scale is being superceded.

The most remarkable radar mapping survey yet undertaken was that performed by NASA's Magellan mission to Venus (*see* **Solar System:** Venus) from 1990 to 1992. Radar had to be used to penetrate the thick Venusian atmosphere. In addition to the huge scope of the global mapping undertaken, a lot of the geomorphological features seen had no obvious terrestrial analogues and hence stretched the interpretational skills of the Magellan science team.

Roughness Mapping

Radar backscatter is sensitive to the roughness of surfaces at length scales much less than the wavelength of the radar. For example, the C-band (wavelength = 5.8 cm) radar of the ERS/ENVISAT satellites are sensitive to surface roughness at the scale of millimetres. Thus, any changes in roughness in space or time that are relevant to geology can be utilized.

Figure 2 Radar amplitude image of part of north central Jamaica viewed by the SEASAT L band radar in 1978. The radar illumination is from the left (west south west) of the image at an incidence angle of about 20 degrees from vertical. Most of the scene is covered by ''dimpled'' pattern typical of tropical karst topography of Cenozoic limestones, except for the bottom left corner which is an inlier of Cretaceous volcanic and volcaniclas tic rocks. In the northernmost central part of the image a straight normal fault scarp downthrows the limestones to the north. This is cut by one of two, subparallel curving fault scarps. The east ernmost one shows the clear pattern of a scissor fault. To the south the fault shows a west facing escarpment that is bright in the image and futher north this changes to an east facing escarpment in shadow. The image measures about 36 × 24 km.

Satellite radar imagery can be used for offshore oil basin reconnaissance using this technique. Oil seeping from depth forms a film on the sea surface that dampens the backscatter roughness relative to surrounding areas. These areas of surface oil appear darker in amplitude images. Hydrocarbon-rich basins often have a variety of discrete oil seeps that help define the basin extent and character. These seeps can be detected by radar roughness mapping. Shorter wavelength radars (e.g., X-band, 3 cm) give a stronger roughness contrast than longer wavelength radars. The ambient sea-state also plays a role: contrast is enhanced by moderate wave formation and by viewing from perpendicular to the wave motion. In order to distinguish between natural seeps and ship pollution it is necessary to build up a time series of observations that establish the semi-permanent nature of seep locations.

Of course, most rock surfaces do not change their roughness over periods relevant to remote sensing. One environment that does, however, is arid playas. Here the evaporation of groundwater leads to the surface formation of evaporite minerals, particularly gypsum and halite. During the summer months, when evaporation is highest and rainfall negligible, the growth of these minerals, specifically halite, can produce a surface that becomes increasingly rough. The C-band radar of ERS is sensitive to this seasonal roughening of the surface of arid playas such that, using a time series of ERS radar amplitude images at 35-day intervals, it is possible to monitor quantitatively the evaporation rate from the backscatter values using models of roughening. Confidence in this approach only comes when the region is very flat and extensive, so that slope effects can be discounted.

Radar is very effective generally in deserts (*see* **Sedimentary Environments:** Deserts), because it can penetrate dry, unconsolidated materials. Where sheets of windblown metres-thick dry sand cover eroded surfaces in the Sahara, it is possible to image these buried surfaces and their subtle differences in roughness beneath their cover, even from space. This is best done with the long wavelength (24 cm) L-band radar which can penetrate several metres into the extremely dry sand by a process known as volume scattering. On reaching a buried rock surface, the radar signal it is reflected back through the sand. Shallow former river channels, eroded in wetter periods, form smoother reflectors relative to their surrounding eroded interfluves, revealing patterns totally hidden beneath their aeolian cover.

Ground Motion Measurement

The phase information from imaging radar can be used to measure the relative spatial motion of the ground surface between two separate images. This technique uses the principle of interferometric synthetic aperture radar (InSAR) and has made a major impact on the Earth sciences in the past ten years, since its practical value was first demonstrated using data from the ERS satellites. It relies on the ability to measure the change in phase at each pixel between two images. If the surface of part of the image moves towards or away from the radar LOS more than in another part of the image, then a phase gradient is measurable (Figure 3). Without independent control on motion, the technique known as differential InSAR only gives relative, within scene, motion-detection. The motion sensitivity is proportional to $4\pi/\lambda$, where λ is the wavelength. Typically for InSAR, using ERS data, LOS accuracies of about 1 cm can be achieved. After each 2π of phase change the phase wraps around and the interferogram image displays a fringe of interference. To retrieve a monotonic motion signal the

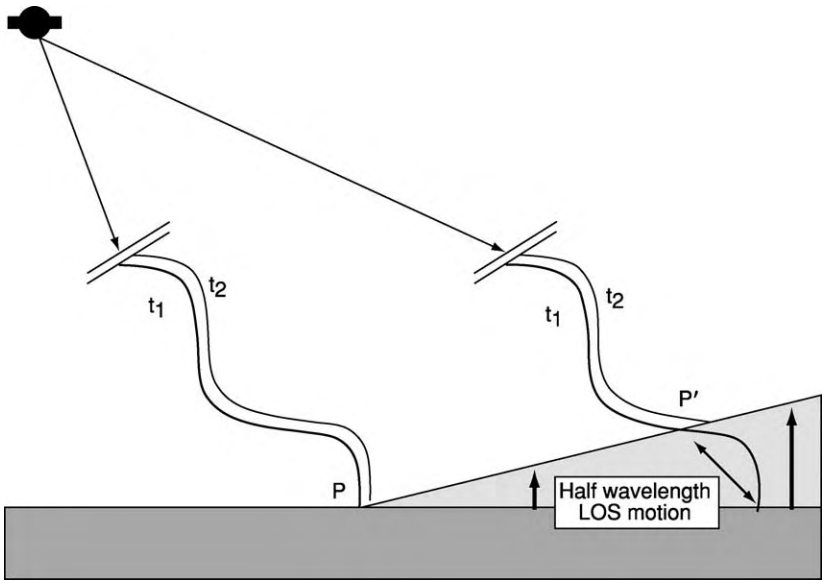


Figure 3 Schematic of the principle of differential InSAR. A radar images the earth's surface at time t_1 as represented by the thick cyclic lines reflecting from the surface at two places. After an interval the radar images the scene again, at time t_2, from the same orbit. During this time the ground surface has moved differentially as shown by the lighter shaded triangle such that there is now a difference in path length along the LOS of the radar of one half a wavelength of the radar between positions P and P′ on the ground. This corresponds to a gradient of 2π in phase, or one fringe in the interferogram made from the difference in phase of the two radar images.

phase fringing must be 'unwrapped' using automated algorithms.

InSAR has opened up a new form of space geodesy at the same time that GPS measurements of surface motion have become commonplace. The two techniques are largely complementary, InSAR has a dense spatial, but sparse temporal sampling, opposite to that of continuous GPS. InSAR also has the potential to work anywhere, without the need for ground instrumentation. However, there are a number of complicating factors that inhibit InSAR's practical usefulness at present. The technique requires the surface scatterers to remain unchanged between the two images of the interferogram. Vegetation change is a major inhibitor here at X- and C-band. Variable water vapour distribution in the troposphere produces unwanted path effects. Also, remarkably, none of the current space-borne radars were designed with InSAR in mind and so a lot of the radar parameters are not optimised for InSAR. Potentially, InSAR has a role to play in any Earth process that involves the motion of the Earth's surface at length scales of 100 m to 1000 km and at strain rates of 10^{-5}–$10^{-7}\,\mathrm{yr}^{-1}$. The Earth science fields that have made most use of it are the study of earthquakes and tectonics, volcanology, and the study of shallow subsidence processes.

Earthquakes and Tectonics

Seismology has dominated the study of earthquakes. In recent years, however, the ability to measure the surface ground motion field associated with fault rupture by InSAR has brought a new dimension to the study of shallow earthquakes. By inverting the co-seismic motion field of InSAR with numerical models of the fault plane motion in a 3D space, knowledge of the earthquake fault mechanism can be improved. Specifically, InSAR can greatly improve knowledge of exactly where the fault plane rupture was and estimates of its dip, whilst seismology is better at resolving the moment released. Immediately following a large earthquake the ground surface can move more slowly over the following months. This can be measured by InSAR to help constrain the physical processes potentially involved: afterslip on the fault plane, poroelastic pressure changes or viscoelastic relaxation in the lower crust/mantle. The other component of the earthquake cycle, the interseismic strain, can also be detected by InSAR, for example, on the San Andreas and North Anatolian Faults. This requires pushing the technique to its limits of sensitivity in measuring very slow strain buildup over several years across widths of about 100 km.

Not all Earth strain involves seismicity and sudden displacements. InSAR is starting to reveal a wealth of surface strain events in the continental plate boundary zones that whilst linked to fault systems, do not themselves display significant seismicity. Aseismic displacement events triggered by nearby seismic events has been measured and shallow folding deformation may be revealed in low competence

rocks. This raises the exciting prospect of being able to relate the current day surface deformation to neotectonic processes and geomorphological features of the past few hundred to thousands of years.

Volcanoes

Earthquakes and volcanoes can share a similar cycle of surface straining associated with the repeated charging of their systems by external forcing (plate motion and magma buoyancy) (*see* **Volcanoes**). In the volcanic case, dilational strain during charging of the magma reservoir is released suddenly, perhaps with shallower planar rupture deformation (dyking and surface fissuring). InSAR is having an impact on our knowledge of the volcano cycle as it has for earthquakes. In the volcano case, we also have the possibility of using the volume of erupted magma during an eruption to constrain the magma loss event in the reservoir. However, the complementarity of InSAR and seismicity in defining the source mechanism of earthquakes is not available in volcanoes. Because of this and the inherently more complex three-dimensional shape of pressured magma reservoirs, progress in refining their location and dynamics has been slower.

Many volcanoes are susceptible to gravitational instability and in particular to spreading under their own load if there is a suitable weak basal layer. This has been measured by InSAR. A complex mixture of spreading forces, local fault accommodation of these, regional tectonic movements, and pressurizing magma reservoirs mean that the INSAR-measured deformation at many volcanoes may reflect a much more complex system than hitherto realized (Figure 4).

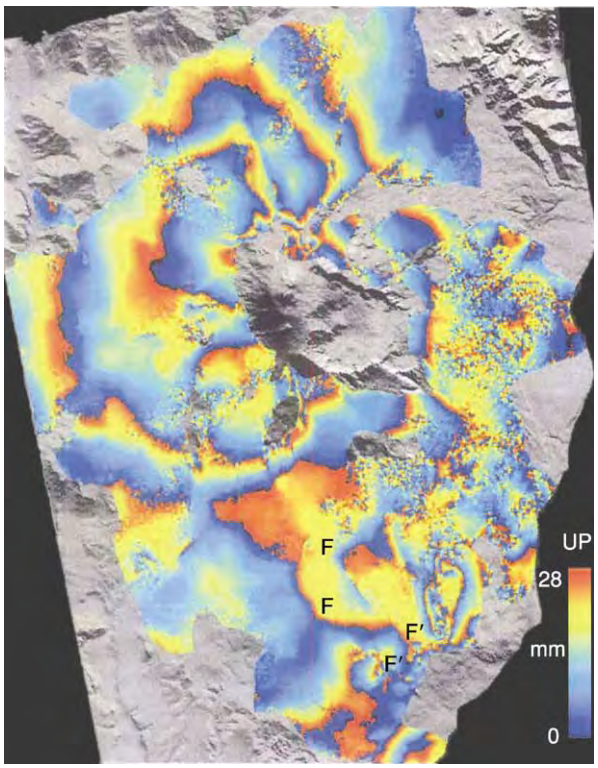


Figure 4 A differential radar interferogram of the Etna volcano in Sicily created using two ERS SAR images from 1995 and 2000. Each colour cycle from blue to red represents one fringe of relative motion of 28 mm. The parts of the scene where the phase data are too noisy are replaced by a shaded relief image of the topography. The displacement pattern of the ground surface shows apparent LOS uplift of >10 cm of the summit region of the volcano relative to the flanks. Also indicated by the F F′ symbols is evidence of relative movement along two faults on the southern flanks. The image is about 40 km wide. The ERS data were supplied by the European Space Agency.

Subsidence

There are a number of natural and anthropogenic or human-induced mechanisms that can produce shallow and hence local subsidence of the Earth's surface: for example, soluble rock dissolution, landslides, mining, groundwater/oil extraction, and tunnelling. InSAR has been shown to be valuable in monitoring all of these. Because of their small spatial extent, it is easier to calibrate the absolute motion.

Local subsidence events are of considerable concern in urban environments and a variant of the InSAR technique has been developed to take particular advantage of this. Buildings and roofs of a variety of shapes and orientations serendipitously provide a source of very strong radar scatterers. In the 'permanent scatterer' technique of InSAR, the phase change information from these scatterers alone is used in a time series of images (typically at least 20–30). By making some assumptions about the nature of the atmospheric effects and with good knowledge of the topography, these sources of error can be removed and a line of sight (LOS) motion history of the set of scatterers can be measured to accuracies of a few millimetres.

See Also

Engineering Geology: Natural and Anthropogenic Geohazards. **Europe:** Mediterranean Tectonics. **Remote Sensing:** Passive Sensors. **Sedimentary Environments:** Deserts. **Solar System:** Venus. **Tectonics:** Earthquakes. **Volcanoes**.

Further Reading

AGU (1992) *Magellan at Venus*. reprinted from Journal of Geophysical Research 97 (E8, E10), Washington, DC: American Geophysical Union.

Amelung F, Jonsson S, Zebker H, and Segall P (2000) Wide spread uplift and 'trapdoor' faulting on Galapagos volca noes observed with radar interferometry. *Nature* 407: 993 996.

Blondel P and Murton BJ (1998) *Handbook of Seafloor Sonar Imagery.* Chichester: J Wiley and Sons.

Elachi C, Roth LE, and Schaber GG (1984) Spaceborne radar subsurface imaging in hyperarid regions. *IEEE Transactions Geoscience and Remote Sensing* GE 22: 383 388.

Ferretti A, Prati C, and Rocca F (2001) Permanent scatter ers in SAR interferometry. *IEEE Transactions Geoscience and Remote Sensing* 39: 8 20.

Hanssen RF (2001) *Radar interferometry. Data Interpret ation and Error Analysis.* Dordrecht: Kluwer Academic Publishers.

Henderson FM and Lewis AJ (1998) *Principles and Appli cations of Imaging Radar.* Manual of Remote Sensing, 3rd edn. Vol. 2. New York: J Wiley and Sons.

Koopmans BN (1983) Side looking radar, a tool for geological surveys. *Remote Sensing Review* 1: 19 69.

Massonnet D, Rossi M, Carmona C, *et al.* (1993) The displacement field of the Landers earthquake mapped by radar interferometry. *Nature* 364: 138 142.

Massonnet D and Feigl KL (1998) Radar interferometry and its application to changes in the earth's surface. *Reviews of Geophysics* 36: 441 500.

Trevett JW (1986) *Imaging Radar for Resources Surveys.* New York: Chapman and Hall.

Ulaby FT, Moore RK, and Fung AK (1986) *Microwave Remote Sensing*, vols. I, II and III. Norwood MA: Artech House.

Wadge G and Archer DJ (2003) Evaporation of groundwater from arid playas measured by C band SAR. *IEEE Trans actions Geoscience and Remote Sensing* 41: 1641 1650.

Wehr A and Lohr U (1999) Airborne laser scanning an introduction and overview. *ISPRS Journal of Photogram metry & Remote Sensing* 54: 68 82.

GIS

P J Mason, HME Partnership, Romford, UK

Introduction

Geological phenomena vary in space and time and this basic trait makes geology a particularly suitable application field for Geographical Information Systems (GIS). A GIS is a computer-based tool for the mapping and analysis of things that exist on, and events that happen on Earth. It is, therefore, a sophisticated but general-purpose tool and one which can be applied in many ways to many problems, and especially so in geosciences.

There have been a number of technological advances, including the development of GIS, which have transformed the work of the geoscientist over the last 15–20 years. Until the 1980s, GIS was used and developed mainly by the geographical community and generally in government organizations and research institutions. The wider geological community has been, until relatively recently, a little sceptical about its operational success in solving complex geological problems.

The first GIS systems were pretty cumbersome hardware and software suites, with high-end functionality that required programming skills to operate. These have evolved into highly functional desktop systems, which anyone can learn to use, and which are part of a growing, multibillion dollar, worldwide industry. GIS functionality developed very quickly, with the advanced statistical and 3D analytical modules, which had been considered requirements only of the advanced GIS community, becoming the selling points for desktop GIS. Multivariate and geo-statistical methods laid the foundations of the geological application of GIS. Specialized exploration software, in the late 1970s and 1980s, had many features present in current GIS suites, such as those of Zycor, Z-Map Plus, Strat Works, GeoData Works, Earth Cube, and StratSoft's Statistica. The costs of these commercial software packages were generally prohibitive and the wider geological community continued to rely on the more readily and cheaply available geo-statistical packages, many of which were available in the public domain.

The technological developments in GIS were accompanied by, and to a large extent enabled by, huge leaps in computer power and development. Large mainframe machines gave way very rapidly to desktop, laptop, and now palmtop processing. Software evolved in parallel with these changes, and GIS mapping can now be done in your hand, in the field, enabling great savings in time and money. The growth of the World Wide Web (WWW) and wireless communications has enabled the generation of, and access to, accurate and timely information, in addition to fast and reliable transport of information from office to field and vice versa.

The end of GPS selective availability in 2000 and the reduced cost and increased availability of commercial satellite imagery, also make 'remote' digital mapping cheaper, more accurate and more accessible. One milestone in this part of the 'geospatial revolution' was the successful launch of the Earth observation satellite, Landsat 7 in 1997, whose data were

commercially available at unprecedented low prices and, for the first time ever, were browsable on the Internet only a short time after acquisition. The launch of ASTER on TERRA AM-1, in 1999, has broadened geological mapping capabilities and reduced costs even further. Sadly, Landsat 7 is no longer fully functional but ASTER is and this series of satellites is likely to evolve further.

This article not intended to provide an exhaustive technical explanation of what GIS is, how it functions or of the technical details of various software suites. Rather, it aims to give an overview of the state of the art in 'geo-GIS', i.e., those aspects of GIS which are specifically relevant to geoscientists and their activities. The Further Reading section at the end of this article includes a number of standard texts on geographical information systems and science, and on supporting subject matter, in addition to some rather more specific publications of interest to the geoscientist working with GIS.

The Geological Information System

Most geologists would agree that a geological map could be considered a database in its own right, as the information contained on the map and in the legend is, by definition, spatially referenced and attributed. To realize the full potential of that database, the map must be properly converted, and it is at this point that the power and tools of GIS come into play. They enable far more to be gained from the original map than that which first meets the eye. Fundamental to the conversion are the mapped spatial relationships, which are interpreted by the GIS through topological rules. Successful digital conversion requires the understanding of the geological and implied topological relationships described by the map.

Components of a geological map are stored as records, or objects, in a spatially referenced GIS database, which can be displayed spatially, as a map, or in graphical or tabular form (as attribute tables). Standard querying enables selection and extraction and/or the linking of related attribute tables.

Attributes, i.e., any kind of descriptive information, can be attached to a record or object in the database. Geological attributes, attached to a map object can be likened to the association of multi-element geochemical data to a single sample locality. GIS allows this attribute information to be viewed and edited in tabular form, where each object represents a record, and each column represents a field containing a different attribute. The objects can then be reclassified, according to the various attributes, and displayed in map form, revealing patterns and relationships not always detectable in tabular form.

Spatial Data, Models and, Structures

There are two basic methods of representing information in a GIS – raster and vector. Geological maps and data can be displayed using either method, but if a layer is to act as a database, it must be properly converted into or captured in vector format.

i. The raster or 'grid'
   The raster displays spatial information as a series of contiguous cells (or pixels) of uniform size, and each raster grid represents the variance within one attribute. This method of storage and display of information is appropriate for data that has high spatial variability, such as geophysical survey data or satellite images. As the raster grid (or lattice) is regular, spatial relationships are implicit and therefore, explicit storage of spatial (and topological) relationships is not necessary, as it is for the vector data model.
ii. Spatial or vector data comprises any or all of the following basic components:
   - Lines or arcs (e.g., representing contacts and boundaries)
   - Areas or polygons (e.g., representing closed areas and regions defined by surrounding lines)
   - Symbols, points or nodes (e.g., such as a sample point, borehole, or a strike and dip measurement at a specific location).

To correctly display a map in vector form, certain information about the geometric and spatial relationships of the map objects is required and this is where topological rules become necessary. For example, the GIS needs to be able to determine whether a fault is connected to another fault, or where one outcrop borders another. **Figure 1** illustrates the appearance of a simple digital geological map in vector and raster form, and the effect of varying raster resolution, on map detail.

There are advantages and disadvantages to both methods, but the fundamental requirement in both cases is that linear features (e.g., boundaries, contacts, and faults) are digitally captured only once. These linear features are then used to construct the areas or polygon objects (e.g., lithological outcrops) that comprise the map. The result is that all adjacent or connected objects have topological connectivity and contiguity, so that spatial relationships between them are understood by the GIS. Topologically correct data capture and database construction should be done from the outset.

The Open GIS Consortium, Inc. (OGC) leads in the development of open, interoperable, geo-processing computing standards and specifications. OGC is involved with government, private industry, and

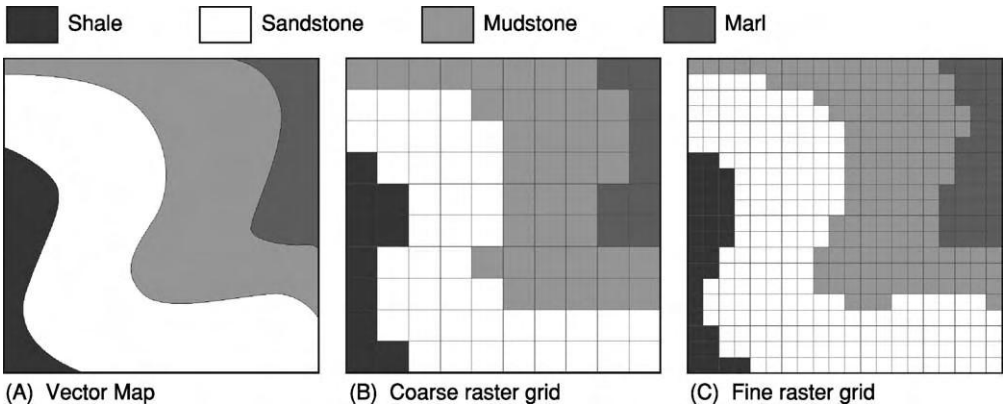


Figure 1 Digital representation of a simple digital geological map; (A) vector or spatial data map; (B) coarse raster grid; and (C) fine raster grid (illustrating the effect of spatial resolution or grid cell size on the detail stored in the representation).

academia, and its adopted GIS standards and specifications provide a valuable resource; they are available at the OGC's website (http://www.opengis.org/about/).

Database Design, Quality, Precision

The design and structure of the GIS database being developed should be governed by the application, and by the intended use of the database. The quality, accuracy, precision, and relevance of the captured data are all significant considerations. It has been said many times that the GIS analysis is only as 'good' as the information that is input.

Georeferencing

To make GIS work, there must be techniques for assigning 'coordinates' of time, location, and attribute in a way that is commonly understood. Basic tabulation, and calendar and time systems are not a problem, so the recording of location is the most vital reference information stored in GIS. Some georeferences are based on names, others on measurements, which are sometimes referred to as 'metric' georeferences. They include latitude and longitude and various other kinds of projected coordinate system, and provide infinitely fine resolution and enable accurate measuring capabilities.

Georeference, geolocation, georegistration, and geocoding are all common terms for the recording of an object's position and shape on a map, in the form of a map projection and datum. The need for map projections is fundamental – the world is a three-dimensional object, roughly oblately spherical in shape, which needs to be represented on a two-dimensional piece of paper or screen. We also need to measure positions, distances, and heights, all of which would be arbitrary and relative, if not made with reference to some known reference datum or ellipsoid. Many coordinate systems, projections, datums, spheroids, and ellipsoids have been developed over the years for various regions of the world and for the whole world. Integrating multisource, multiform data into the GIS presents the fundamental need to ensure coordinate reference system conformality. This can be time consuming but most GIS software contains tools for transforming datasets from one system to another. There are also some very comprehensive reference texts, and now Internet resources, to help deal with this issue.

Visualization

GIS visualization tools (two- and three-dimensional) within GIS enable and rely on the ability to share and integrate multisource data, models, and interpretations. The simplest form of visualization involves the display of two-dimensional images with conventional cartographic symbols. The geological map can be recreated using conventional geological symbology, which can easily be incorporated within the GIS, through the use of special fonts. Thus structural information can be presented to give the appearance of a published geological map (**Figure 2**).

The ability of GIS to bridge the gap between two- and three-dimensional display and analysis is especially useful in geosciences, because depth and 3D structure are fundamental considerations. The ability to integrate surface and sub-surface data is therefore an essential function of any digital mapping technology in geoscience.

GIS enables exploration and interrogation in several dimensions for improved and enhanced understanding of structures and relationships. It is possible to conduct virtual geological field visits, which can assist in the logistics and improve understanding prior to setting foot in the field, thus saving valuable time. Most GIS software provide such visualization tools

Figure 2 Data capture and conversion from a simple field sketch map (left) to the visualization of a digital geological map (right) within GIS.

for the manipulation of surfaces, images, and maps in pseudo three-dimensional space, and the mathematical derivation of other products, such as slope angle, aspect or azimuth, curvature, line-of-sight, watersheds, and catchments. Examples of software providing excellent 3D manipulation and viewing capabilities are ERDAS Imagine's Virtual GIS, ER Mapper, ArcGIS 3D Analyst, and Geomatica Fly.

Visualization in pseudo-3D requires a valuable commodity – a digital representation of a surface, whether it is topography or other feature data, such as an interpreted seismic horizon or geophysical survey data. The attribute information in this 'surface' layer provides the third dimension or 'z' value for the display.

There are many ways to obtain and derive digital surface topographic data or DEMs. Many datasets can now be browsed and obtained on the Internet from organizations such as the United States Geological Survey and the National Imagery and Mapping Agency (NIMA). In February 2000, the Shuttle Radar Topographic Mapping Mission (SRTM) carried out radar interferometric mapping of the Earth between +60 and −60 degrees latitude, in fewer than 11 days. SRTM is an international project, headed by NIMA and the National Aeronautics and Space Administration (NASA), with the objective of generating global DEMs. DEM data, gridded at 90 m are now available for much of the world, and at 30 m for the USA only, and are downloadable, free of charge, from ftp://edcsgs9.cr.usgs/pub/data/srtm. Sophisticated software exists (such as PCI Geomatica, Socket Set, ENVI ASTER module, Virtuozo) to generate DEMs from stereo airborne and space-borne photography and imagery. One such source is Terra-1 ASTER, whose stereo capability means that high-quality DEMs can be readily generated at very low costs and at spatial resolutions suitable for 1:50 000 scale mapping (http://asterweb. jpl.nasa.gov/).

Geological Application Fields

The application areas to which GIS makes a significant contribution are as varied as they are numerous. In geosciences there are some key areas where the necessity for and contribution of GIS are clear and significant. They range from the basic requirement of producing maps through the solving of complex geological challenges, to advanced spatial analysis.

Field Mapping

The most basic of activities for the geoscientist, mapping, has been changed fundamentally through the use of GIS. Though the hammer may remain ubiquitous, weatherworn field slips and notes are being replaced. Field mapping essentials now include laptop and palmtop computers (**Figure 3**), connected to GPS hand sets (and mobile telephones) so that positional data can be fed directly and dynamically into the GIS, to record current positions for waypoints or sample localities. Alternatively, tracks can be recorded as you walk along, so that you can re-trace your steps at any time, and boundaries and contacts can be walked out to provide direct mapping to the GIS database. Even sketches and notes can be immediately input through palmtop graphical interfaces and handwriting transcribers.

These processes are commonly facilitated by Earth observation image data (from both satellite and airborne platforms), which are easily integrated in the

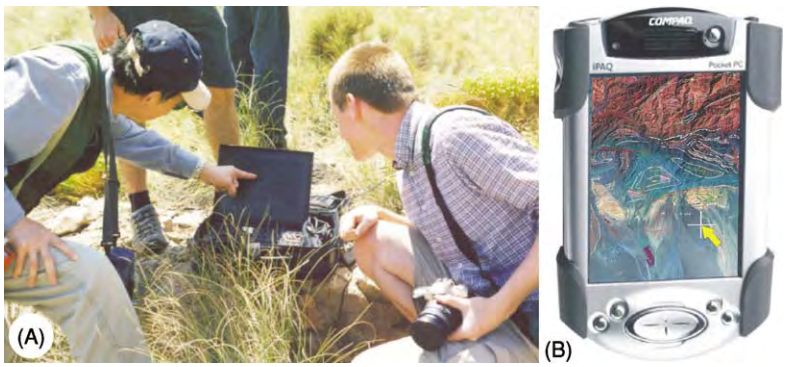


Figure 3 GIS in the field: (A) Staff and students record their observations on a digital map, in real time, using a GPS handset, GIS and laptop, during fieldwork in Spain; (B) Palm top computer, displaying a geo referenced Landsat 7 satellite image being used as a base map, within GIS, for field geological mapping.

GIS. They form ideal basemaps and provide hard evidence of Earth surface features and changes.

Natural Resources

A necessary and important area of growth for GIS is the promotion and employment of GIS in developing countries for the management of all natural resources. One of the most important global challenges, which will affect both developed and developing worlds, is water resource management. GIS can be used for almost all applications related to water resource management, such as groundwater targeting, resource estimation, and groundwater recharge estimation. Groundwater risk assessment can also be carried out; GIS enables mapping and analysis of contaminant plumes (caused by oil, chemical, and radioactive pollutants, for example), ground water vulnerability assessment, and environmental impact evaluation. The United States Geological Survey provides useful web based information on water resources and hydrological data (http://water.usgs.gov/maps.html).

Exploration and Production

The identification of potential new exploration targets is vital for exploration mapping and management and the use of GIS in mineral and petroleum exploration is now widespread. One of GIS's most useful contributions in this context is the capacity for integration of disparate digital datasets into a single database. All available data, geological, geochemical, and geophysical, can be compiled, analysed, and interpreted, to produce exploration potential maps and images (such as in **Figure 4**).

In quantifying prospectivity, through spatial analysis, the development of an analytical model is as important as the quality of the data; identifying significant factors, evaluating their significance, and establishing relevant weights for the layers used in the model are all fundamental considerations. The resultant prospectivity map reflects the depth and accuracy of the input information, the relevance of the model, and the geoscientists experience and local knowledge.

Environment and Geohazards

Environmental Quality

All aspects of environmental monitoring and management are particularly suited to the application of GIS. Using digital terrain models, geochemical and hydrological data, pathways for pollutants can be assessed and the probability of contamination calculated for areas where no sampling has been performed. The Interdepartmental Committee on the Redevelopment of Contaminated Land (ICRCL) publishes guidelines and trigger concentration values, based on appropriate usage in, for example, the UK, on the Internet (http://www.contaminatedland.co.uk/std-guid/icrcl-l.htm). Such values provide an ideal starting point for establishing 'decision rules' within the GIS model. Analytical tools, found in any GIS, enable generation of basic mapping, site selection, and advanced spatial analysis for environmental impact assessments, environmental audits and risk assessment, in addition to soil and groundwater investigation, remediation, and quality control.

Geohazards

The term geohazard is used in reference to those potentially dangerous phenomena that are caused by natural, geological processes, such as landslides, ground subsidence earthquakes, and volcanic eruptions. GIS has an important role to play in understanding and quantifying the spatial variability of the phenomena, in predicting its future behaviour,

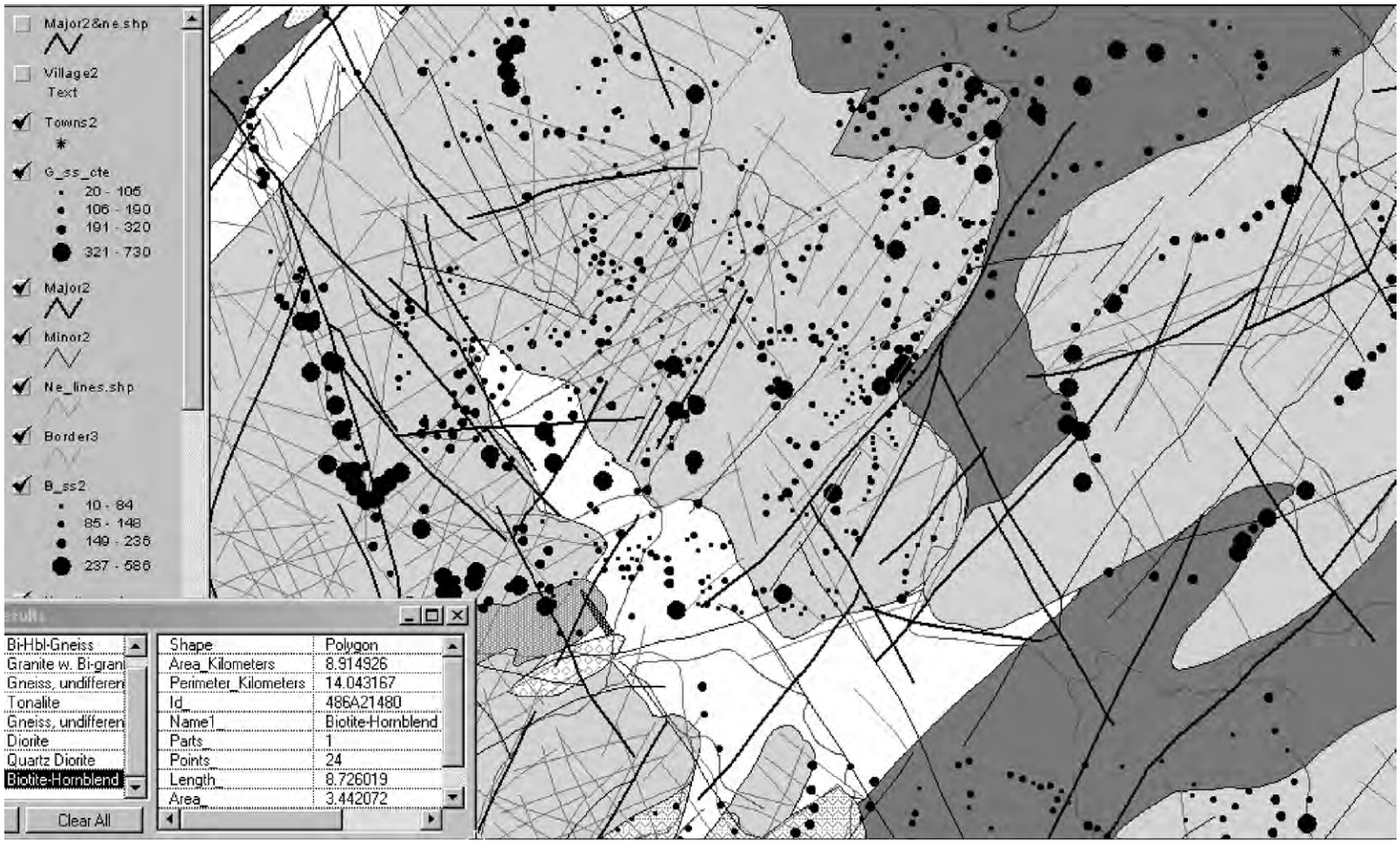


Figure 4 Display of vector polygon (lithological), line (structural features) and point (geochemical sample) data in an exploration GIS database.

and frequently in helping to identify appropriate remediatory activities.

The Integrated Global Observing Strategy (IGOS) is a strategic partnership linking research, and long-term monitoring and management with data users and suppliers, for global environmental observations of the atmosphere, oceans, and land. Its themes include all the main geohazards (earthquakes, volcanoes, subsidence, and landslides) and one of its key objectives is "to develop and integrate Earth observation data, *in-situ* spatial data organisation, spatio-temporal data modelling and analysis, using GIS, for a global, integrated geohazard monitoring system". Information about IGOS, its contributing partners and its goals, can be found on its website at http://www.igospartners.org.

Landslides Landslide hazard assessment involves the classification of the land surface into areas according to the degree of potential hazard posed by mass movements; an example of a landslide hazard assessment map is shown in Figure 5. Many methods and techniques have been proposed and used to quantify causative factors and GIS is commonly used to produce maps representing the probability of occurrences, on the basis of occurrences in the past. This 'direct' method of hazard assessment consists of geomorphological mapping of past and present landslides, and identification of factors leading to instability. GIS is then used for classification, or zonation, to reveal sites where future failures are most probable. The alternative 'indirect' method includes two different approaches, namely the didactic or statistical (data driven) and heuristic (knowledge driven) techniques. The didactic approach involves construction of an abstract model and proposal of hypotheses, followed by experiments and data collection, testing of the hypotheses and finally the construction of landslide instability rules. In the heuristic approach, landslide-influencing factors, such as slope angle, lithology, landform, and land-use, are ranked and weighted according to their expected significance in causing slope instability, and the success of the result is tested against known cases before rules are constructed (*see* **Sedimentary Processes:** Landslides).

Vulcanology GIS contributes many tools to manage field data (such as steam, water, and soil geochemistry), enabling patterns of output and geochemical changes to be identified effectively, and presented graphically. Another important role for GIS is in the modelling and prediction of the behaviour of ejected materials and mass movements after eruptions. Using a Digital Elevation Model (DEM) images, and maps, the existing topography can be analysed, enabling paths, discharge rates, and velocities of potential

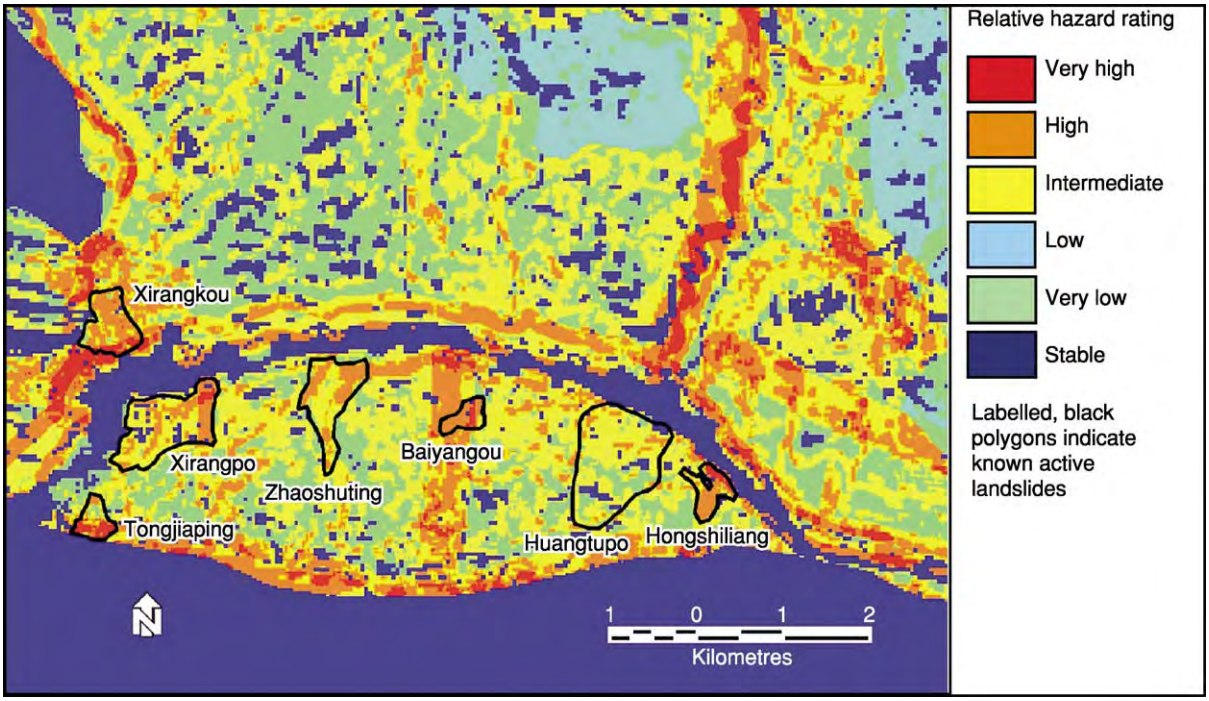


Figure 5 Map showing relative landslide hazard levels as derived from a multi criteria elimination and characterization model, of an area of the Three Gorges region in China. The locations of the landslides, which are known to be active in this area, are indicated by black polygons.

lahars and mudflows to be predicted. Pre- and post-eruption terrain modelling also allows the monitoring and detection of landform changes, and the preparation of susceptibility maps (*see* **Volcanoes**).

Earthquakes Seismic hazards caused by ground motion and failure are characterized by their spectral response, and by peak ground acceleration and velocity. The spatial distribution and patterns of ground motion can be determined using deterministic and probabilistic analysis within GIS. For an event of a certain magnitude, soil conditions and attenuation relationships can be used to predict the amount and distribution of ground shaking. Another significant role for GIS is the modelling and simulation of motion intensity and damage under various scenarios (*see* **Tectonics:** Earthquakes).

An example of the operational use of GIS in this field is the HAZUS programme (HAZards in the United States), under the control of the Federal Emergency Management Agency (FEMA), which has been developed to estimate losses caused by earthquakes, hurricane and wind erosion, and flooding. HAZUS takes the form of a software 'plugin' that exploits functionality within ArcView or MapInfo GIS packages. A new, HAZUS-MH (Multi-Hazard) version is due for release and more information can be obtained from http://www.hazus.org and http://www.fema.gov/hazus.

Tools for Spatial Analysis

Management, integration, manipulation, and visualization of high volumes of data require sophisticated tools for effective storage, interrogation, and analysis, to generate suitability maps based on a hypothesis. A GIS contains a variety of simple and complex tools for analysis of single data layers, pairs of layers, or multiple layers, enabling the extraction of a single composite result.

Individual Layers

Spatial analysis of individual maps and layers involves two-dimensional processing and geo-statistical methods, such as reclassification and thresholding, neighbourhood functions using spatial filters, distance, and buffer calculations, 2D spatial transformations and, importantly, gridding or interpolation.

Geo-statistical methods, involving the application of probabilistic methods to geographically related phenomena, can be used to highlight spatial correlation within a data layer. This idea is based on the assumption that points located close to one another, should also be close in value. Existing data are then used to interpolate into areas where no data exists.

Multiple Layers

Simple tools and processes can also be used effectively to manipulate a multisource, multilayer dataset, to produce a single, composite, and meaningful result. The term 'Map Algebra' is commonly used in reference to mathematical calculations carried out between two or more GIS layers, according to some expression, to produce a new layer. For example, to derive a map showing the depletion of an aquifer over a set period of time, layers representing groundwater levels as measured at set times are required. One layer is simply subtracted from the other to produce a layer representing the changes between the two dates (Figure 6). GIS enables the modification and classification of each layer and the combination of the modified layers, mathematically, to produce the final result. Through the combination of several, simple operations in this way, either collectively or in sequence, complex problems can be addressed using GIS.

Multicriteria Evaluation and Uncertainty

Multisource datasets are rarely complete and, in Earth sciences, are seldom defined by rigid, Boolean-type, observable boundaries. Uncertainties, arising from data and decision rules, are common. GIS provides a number of specialised, multicriteria, analytical tools to deal with these uncertainties, which draw on the principles of, for example, Bayesian probability and fuzzy logic.

Multicriteria evaluation is primarily concerned with the combination of multiple datasets to produce a single index of evaluation. In the simplest cases, such as those involving Boolean constraints, this combination can be achieved using union (logical OR) or intersection (logical AND) rules. For factors that display continuous variability, some form of ranking and weighting is required. As the criteria are measured and recorded on different scales, they also require standardization so that all factors are quantified on the same scale. It is also convenient if they are all positively correlated with 'suitability'. Deciding on ranking and weighting is the most difficult aspect, and it commonly requires discussion, field verification, and iterative modification, for which a common technique is the comparison matrix (illustrated in Table 1).

Quantifying uncertainty Ground truth data rarely provide pure examples of the classes they represent, and boundaries between one type of natural phenomenon and another are rarely crisp or clearly defined; there is usually a continuum between them. Modifications in the choice of function, and the replacement of 'hard' for 'soft' decision rules, are often

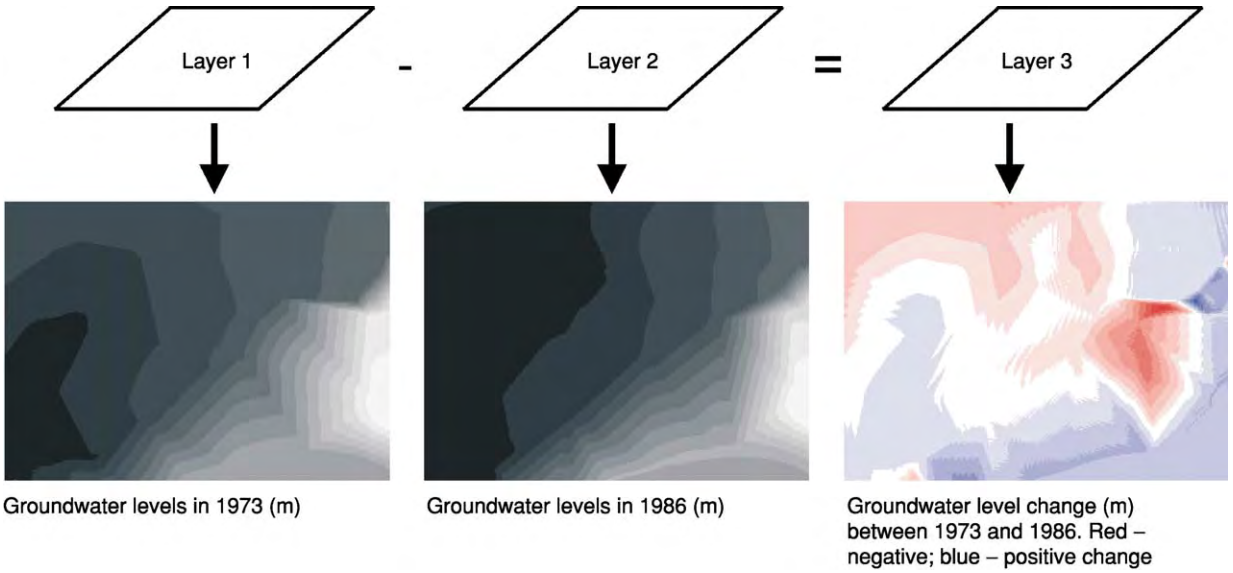


Figure 6 A simple map algebraic example, in which spatial changes in groundwater levels, over a fixed time period (between 1973 and 1986) are derived by layer subtraction.

Table 1 Example of a pairwise comparison matrix, used to assess relative factor significance and to calculate factor weights (for landslide hazard assessment in this case)

	Slope	*Aspect*	*FS*	*DEM*	*Distance (from roads)*	*Distance (from drainage)*	*Soil moisture*	*Iron oxides*	*Clay content*
Slope	1								
Aspect	1	1							
Factor of safety (FS)	1/3	1/3	1						
DEM	1/6	1/6	1/5	1					
Distance (from roads)	1/4	1/4	1/3	2	1				
Distance (from drainage)	1/7	1/7	1/4	1	1/5	1			
Soil moisture	1/5	1/5	1/3	2	1/2	2	1		
Iron oxides	1/5	1/5	1/3	1	1/2	1	1/2	1	
Clay content	1/7	1/7	1/5	1	1/2	1	1	1	1
Factor weights	0.278	0.278	0.151	0.037	0.086	0.034	0.057	0.041	0.039

necessary within the analysis. Bayesian probability theory, Dempster–Shafer theory, and fuzzy set membership are commonly used to incorporate uncertainty in GIS analysis in geological applications, for example, landslide hazard assessment and mineral prospectivity mapping.

Considerable research has been carried out into the extraction of mineral prospectivity maps from multisource datasets. In many cases, identification of the very lowest and very highest prospectivity has not been in dispute but uncertainty arises in definition of the regions of intermediate prospectivity, which then require further analysis and interpretation. Suitability or prospectivity must be treated as a continuous phenomenon representing a measure of confidence in the outcome.

A fuzzy set describes a continuous membership function where 0 represents a non-member and a value of 1 represents a member, and the values between them represent the increasing possibility of membership. The fuzzy set membership function can be most readily appreciated with reference to a simple linear function (Figure 7). The fuzzy function, may in fact be linear, sigmoidal or 'J' shaped, and monotonic or symmetric. The threshold values, which define it, will depend on the phenomenon and desired outcome

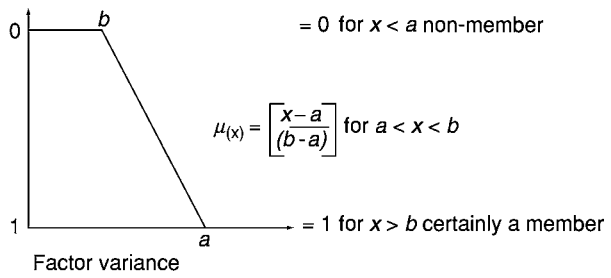


Figure 7 A simple, linear fuzzy set membership function, which can be used to standardize factors and to allow for a measure of uncertainty.

of the operation, i.e., the threshold values applied to each membership function reflect their influence on the result. The resultant fuzzy set layers can be combined in a number of ways, for example, using Boolean logic and fuzzy algebra or 'set theory'.

One novel technique, known as the Vectoral Fuzzy Logic method, produces two maps: one of calculated prospectivity and the other of confidence (represented by the similarity between the two inputs). Using the vectoral fuzzy logic method, null data and incomplete knowledge can be incorporated into the multicriterion analysis.

One of the most widely used statistical, multicriteria techniques is the Weights of Evidence Method. Here the quantitative spatial relationships between datasets representing significant criteria and known occurrences are analysed using Bayesian probability analysis. Predictor maps are used as input evidence and the end-products are maps representing the probability of occurrence of a phenomenon and of the uncertainty of those probability estimates.

The 'weighted linear factors in combination' method is also common, but can, in some instances, be considered too 'liberal' in its handling of the data in the system. Its factor aggregation method can be likened to a parallel connection system, which allows all input criteria to survive to the end. It is also possible that the relationships between the geological factors are not linear, in which case, a more complex model would be required. In many cases, this parallel system may be appropriate but in others a more strict system is required, which enables certain factors or combinations of factors to be eliminated completely from the system. A model based on the simple calculation of the geometric mean of scaled input factors (as opposed to arithmetic mean) can be thought of as a sequential connection system in which the occurrence of a zero rating, at a particular location, terminates the system and eliminates that location from the analysis. The method used to generate the landslide hazard map in **Figure 5**, incorporates this concept. Ideally, the method should incorporate both of these types of data treatment. Commonly the terms factor and 'constraint' are used to distinguish between information which is measured on a relative scale indicating a variable degree of likelihood, and that which is Boolean, respectively.

One difficulty with many of the approaches employed is the assumption that all factors contribute positively towards the suitability; consideration may also need to be given to evidence that appears contradictory. Dempster–Shafer theory provides a method of establishing suitability (or unsuitability) as well as a measure of plausibility for each hypothesis. Bayesian probability allows any new evidence to be incorporated, so updating the probability of suitability each time, but data are required for every possible occurrence, i.e., the absence of evidence is assumed to support the alternative hypothesis (unsuitability). The Dempster–Shafer method allows for 'Ignorance', i.e., the lack of evidence does not necessarily support the alternative hypothesis. A measure of plausibility (and of belief) is incorporated in the outcome. The theory recognizes a total of six concepts–Basic Probability Assignment, Ignorance, Belief, Disbelief, Plausibility, and a Belief Interval. Dempster–Shafer theory, therefore, provides a very useful basis for establishing the value of information and the design of further site investigation, including where new data would be most effective in further reducing the uncertainty.

Data Sharing and the Internet

The alliance between GIS and the Internet is both a natural and an extremely powerful one. Internet GIS databases allow access to vast resources of information, not only geological but also for public interest and attention. Database architectures and data formats, specially designed to be served through intranets and the Internet, enable huge volumes of information to be effectively served to computers anywhere.

Over the last few years, data compression software and specially designed internet servers have been developed, for example, Earth Resources Mapping's wavelet transform based data compression format (.ecw) and *Image Web Server*, which enables vast raster image datasets to be served and shared with incredible speed. Examples of the internet map data serving, through the use of intelligent databases and systems, include the NIMA Geospatial Engine (http://geoengine.nima.mil/), US based GIS Data Depot (http://data.geocomm.com/) and the British Geological Survey's Geoscience Data Index

Table 2 Commercially available GIS software suites, their manufacturers and application fields

Manfacturer	*Professional*	*Palmtop mapping*	*Internet mapping*	*Free viewer*
ESRI	ArcInfo, ArcGIS, & ArcView	ArcPAD	ArcIMS	ArcExplorer
Leica Geosystems	ERDAS Imagine		ESRI	Mapsheets Express
MapInfo	MapInfo	MapXtend	MapXtreme	ProViewer
Autodesk	AutoCAD/World	OnSite	MapGuide	AutoCad LT
Intergraph	GeoMedia & GeoMedia Pro	IntelliWhere	GeoMedia Web Map	GeoMedia Viewer

(http://www.bgs.ac.uk/geoindex/index.htm). These allow the user to search, view, and in the case of NIMA, actually download free any globally available geological, image, and map data (*see* **Geological Field Mapping**).

Software

The GIS software suites now commercially available provide a dazzling choice of functionality to suit all applications and budgets. The following table (Table 2) provides a brief overview of the main software products and their variants, as designed for different markets and applications.

Conclusions

GIS provides a very convenient framework for map-analysis and modelling functions which can be applied in many areas of the geosciences. It is expected that there will be continuing demand for GIS functionality to become faster, cheaper, and more complex, with continuing open standards. There will also be demand for the worlds national authorities to release more GIS data, free of charge, to the public. The vision of OGC is "a world in which everyone benefits from geographic information and services made available across any network, application or platform". Geoscience represents an expanding application area for GIS, both for research and commercial development.

A minor, but general, failing of standard GIS software functionality is its inability to properly deal with geological structures in three dimensions. There are, however, other software suites on the market, which are specially designed for this purpose and so the focus of GIS's contribution in geology lies in data integration, management, and spatial analysis.

There are more automated techniques for solving complex, spatial, multicriteria problems, such as neural networks and data mining, but these are in some respects at a disadvantage, compared with GIS, as they require very large and homogeneous datasets for training the systems. Data mining involves the development and use of effective tools for direct access and extraction of appropriate datasets and patterns from very large volume databases and represents a logical step in the evolution of effective data supply and analysis. Both data mining and neural networks have some overlap with GIS in functionality and objectives, and are anticipated to have some impact on GIS in the coming years. Commonly though, the geoscientist prefers a more intuitive, graphical, and less formal approach, so that even in complex analysis, interpretive skills are still important.

GIS owes its power and success to its flexibility and potential for cross-disciplinary application. GIS (and other related technologies) enable the analysis of vast and expanding datasets, allowing a better understanding of terrestrial and anthropogenic processes, the development of economic sustainability, and the improvement of environmental quality.

See Also

Geological Field Mapping. **Remote Sensing:** Active Sensors; Passive Sensors. **Sedimentary Processes:** Landslides. **Tectonics:** Earthquakes. **Volcanoes**.

Further Reading

Bonham Carter GF (1994) *Geographic Information Systems for Geoscientists*. Pergamon.

Bonham Carter GF (1996) *Geographic Information Systems for Geoscientists: Modelling with GIS*, 2nd edn. Pergamon.

Burrough PA (1986) *Principles of Geographical Information Systems for land resources assessment*. Oxford: Clarendon Press.

Eastman JR, Jin W, Kyem PAK, and Toledano J (1995) Raster Procedures for Multi Criteria/Multi Objective Decisions. *Photogrammetric Engineering and Remote Sensing* 61(5): 539 547.

Illife JC (2000) *Datums and projections for remote sensing, GIS, and surveying*. Caithness, Scotland: Whittle Publishing.

Isaaks EH and Srivastava RM (1989) *An Introduction to Applied Geostatistics*. New York, Oxford: Oxford Univ. Press.

Knox Robinson CM (2000) Vectoral fuzzy logic: a novel technique for enhanced mineral prospectivity mapping, with reference to the gold mineralisation potential of the Kalgoorlie terrane, Western Australia. *Australian Journal of Earth Sciences* 47(5): 929 941.

Liu JG, Mason PJ, Clerici N, Chen S, and Davis AM (2003) Landslide Hazard Assessment in the Three Gorges Area of the Yangtze River using ASTER Imagery: Zigui Badong. *Geomorphology* 61: 171 187.

Longley PA, Goodchild MR, Maguire DJ, and Rhind DW (eds.) (1999) *Geographical Information Systems Principles, techniques, applications and management*. New York: John Wiley.

Mason PJ and Rosenbaum MS (2002) Geohazard mapping for predicting landslides: the Langhe Hills in Piemonte, NW Italy. *Quarterly Journal of Engineering Geology & Hydrology* 35: 317 326.

Mather PM (ed.) (1994) *Geographical Information Handling research and application*. Wiley and Sons.

Zadeh LA (1965) Fuzzy sets, Information and Control. 8: 338 353.

Passive Sensors

J G Liu, Imperial College London, London, UK

What is a Passive Sensor?

A sensor detects and quantitatively records the electromagnetic radiation (EMR) from an object as 'sensing'. For an object to be sensed, it must either reflect or transmit the radiation impinging on it from an illumination source or radiate its own energy. If a sensor provides its own illumination source it is an active sensor. Synthetic Aperture Radar (SAR) is a typical active sensor system as it sends microwave radiation pulses to illuminate the target area and receives the returned signals to produce an image. If the sensor depends on an independent illumination source, such as the Sun, or the radiation from the target itself, such as the Earth's thermal emission, it is a passive sensor. The most commonly used sensors for remote sensing systems are passive sensors, which are the topic of this article.

A passive sensor images the Earth by recording either the reflected solar radiation or the emitted radiation from the Earth. An ordinary camera is a typical passive sensor system. A daytime outdoor colour photograph taken by a camera is a record of reflected solar radiation in the visible spectral range from the scene that is illuminated by the Sun. A camera can become an active sensor when it is used with a flash light in darkness. In this case, the camera provides its own light source to illuminate the object and then takes a picture.

The Sun is the primary illumination source for the Earth. For Earth observation remote sensing, most passive sensor systems operate under solar illumination during daytime, ranging from aerial photography to satellite multispectral scanners. These sensors pick up the reflected solar energy from the land surface to produce panchromatic and multispectral images. The features in these images are mainly influenced by two types of information: spatial pattern (e.g., topography) and spectral signatures. Ignoring minor factors, we can present such an image as

$$M_r(\lambda) = \rho(\lambda)E(\lambda) \quad [1]$$

where $M_r(\lambda)$ is the reflected solar radiation of spectral length λ by the land surface, or an image of spectral band λ, $E(\lambda)$ is irradiance, the incident solar radiation energy upon the land surface, while $\rho(\lambda)$ is the reflectance of land surface at wavelength λ.

If solar illumination is constant, $E(\lambda)$ is effectively the topography, as it governs the geometry of a surface in relation to illumination. In general, considering solar radiation as a parallel radiation source illuminating the Earth, with constant incident radiant flux density M_s, then the solar irradiance upon the land surface, E, varies with the angle between the land surface and the incident solar radiation, γ. When the land surface is perpendicular to the incident solar radiation M_s, E is at its maximum and equal to M_s. If the solar radiation has an incident angle θ_1 (from nadir) and azimuth angle ϕ_1, then the irradiance upon a land surface with slope angle θ_2 and aspect direction ϕ_2 is determined by:

$$\begin{aligned} E &= M_s \sin\gamma \\ &= M_s[\sin\theta_1\sin\theta_2\cos(\phi_1 - \phi_2) + \cos\theta_1\cos\theta_2] \end{aligned} \quad [2]$$

The spectral reflectance, $\rho(\lambda)$, is a physical property, quantitatively describing the reflectivity of the exposed materials on the land surface at wavelength λ. The selective absorption and reflection of a material result in variation of spectral reflectanc across a sensor's spectral range, giving a unique signature for this substance. It is, therefore, possible to work out the land cover types or mineral compositions of the land surface based on spectral signatures using multispectral image data. Reflective spectral remote sensing is among the most effective technologies to study the Earth surface composition and has become an essential tool for geological mapping.

A thermal sensor is also a passive sensor but the energy that a thermal sensor detects is radiation emitted from the Earth's surface, rather than reflection. Thermal sensing does not need an illumination source as the target itself is the illumination source. The Earth's surface can be approximated as a black body of 30 K and using Wien's law, we can calculate that the radiation peak of the Earth is at about 10 μm. In this spectral range, the radiation can be sensed and measured by temperature, rather than brightness, as measured for the visible spectral range; it is, therefore, called thermal sensing. Different materials on the land surface have different thermal radiation properties; thermal sensors are, therefore, useful tools for geological and environmental studies.

According to Wien's law, the higher the temperature of an object, the shorter the wavelength of its radiation peak. Thus, passive sensors designed for shorter wavelengths, though usually used to detect reflected solar radiation from the land surface, can also be used for detecting emitted energy from targets as long as the temperature of these targets is high enough. For instance, Landsat Thematic Mapper (TM) shortwave infrared band 5 (1.55–1.75 μm) and band 7 (2.08–2.35 μm) are far more effective to map fires on a land surface than the thermal band.

The land surface is also an effective radiator at microwave range, though the microwave emission is significantly weaker than thermal emission. As another type of passive sensor, radiometers are designed to image the emitted radiation from the Earth's surface at this spectral range.

For geological studies, the most effective passive sensor systems are of medium to high spatial resolutions with image pixel size ranging from a few to a few hundred metres and are often called Earth observation systems. The following sections use the most common Earth observation systems as examples for further discussion of passive sensors.

Passive Sensor Imaging Technology

With a few exceptions, most passive sensor systems are essentially an optical camera system. As shown in Figure 1, the incoming energy to the sensor system goes through an optical lens and is focused onto the rear focal plane of the lens where the energy is recorded by a sensor device of radiation sensitive media, such as film or CCD (Charged Coupled Device).

Aerial photography, using a large format camera, was the earliest operational remote sensing technology for topographic survey as well as for geological investigation. The space-borne remote sensing for Earth observation began with Landsat-1 with a MSS (Multi-spectral Scanner) and RBV (Return Beam Vidicon) on board, launched on 23 July 1972. This and subsequent landsat instruments collect, record, and transmit electronically to ground stations, images of the Earth. The images are then distributed to users in a digital format as digital imagery data. The concept and technology of digital imagery gave birth to satellite remote sensing. An Earth observation satellite images the Earth's surface continuously from its orbit and sends the images back electronically to the receiving stations on the ground.

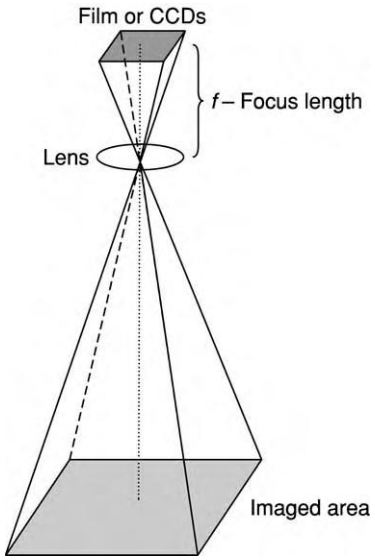


Figure 1 Basic structure of an optical sensor system.

The development of sensor technology is mainly focused on improving the spatial and spectral resolution. For a given sensor system, its spatial resolution is dictated by the minimal energy level of electromagnetic radiation (EMR) that can make a signal distinguishable from the electronic background noise, i.e., dark current, of the instrument. This minimal energy of EMR is proportional to the product of radiation intensity over a spectral range, IFOV (Instant Field Of View), and dwelling time.

The IFOV (Figure 2) is decided by spatial sample density of an optical sensor system and determines the pixel size of the image. For a given dwelling time (equivalent to exposure time) and spectral range, the larger the IFOV, the more energy will be received by the sensor, but the lower the spatial resolution. To improve the spatial resolution, IFOV needs to be reduced while maintaining the same energy level; either dwelling time or spectral range, or both have to be increased. When a sensor, which has its dwelling time fixed by the sensor design and platform orbit parameters, receives the reflected solar radiation from the Earth, it may record the energy in a broad spectral range as a single image. This generates a

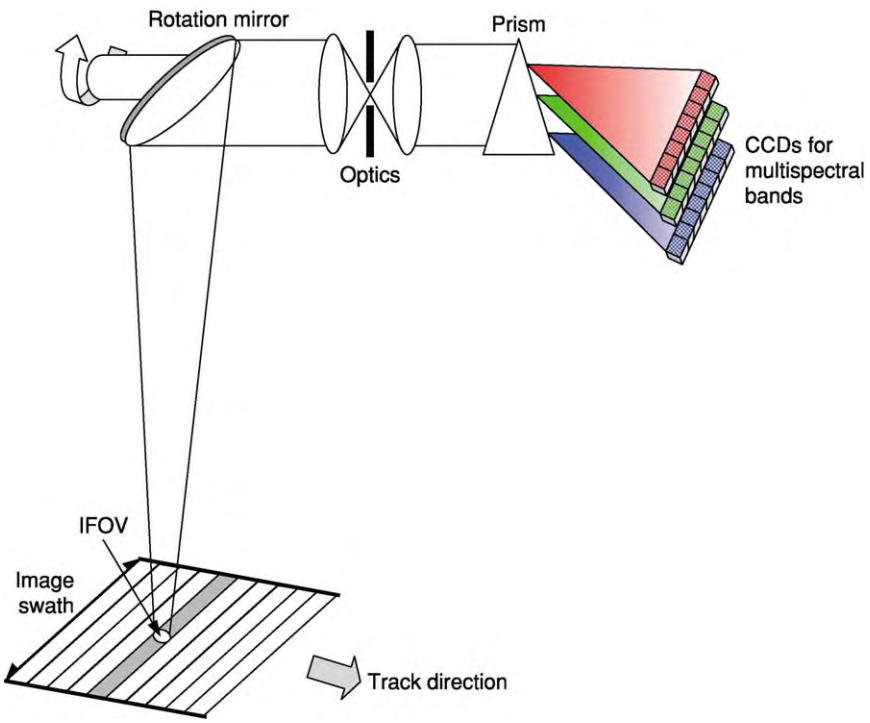


Figure 2 A schematic diagram of an across track mechanical multispectral scanner.

panchromatic image at a relatively high resolution. It may also split the light into several spectral bands and record them separately into several images of narrower spectral range: the multispectral images. In this case, the energy that reaches the CCDs of each narrow spectral range is significantly weaker than in the panchromatic mode. To achieve the same energy level, the only solution is to increase the size of IFOV, or reduce spatial resolution. This is the reason that nearly all the optical imaging systems achieve higher resolution in a panchromatic band than in multispectral bands (Tables 1 and 2). For instance, SPOT 1–3 HRV (High Resolution Visible) Panchromatic bands have 10 m resolution while its XS (multispectral) bands have 20 m resolution.

There is another way to improve the spatial resolution for both panchromatic and multispectral imagery; by increasing the dwelling (exposure) time. This has been an important consideration in sensor design, although the capacity for increasing the dwelling time is limited; for both airborne and space-borne remote sensing, the image is taken from a moving object and so a long exposure time will blur the image.

Across-Track Mechanical Scanner

The early design of space-borne passive sensor systems was constrained by the technology of the key sensor unit: the CCD. In many systems, a mechanical scanner has been an ideal solution to achieve multispectral imaging at relatively high resolution. Such scanners are based on a simple mechanical device using a few CCDs. Figure 2 is a schematic diagram of the principle of an across-track, mechanical multi-spectral scanner. The optical part of a scanner is essentially a camera but the image is formulated pixel by pixel, scanned across track by a rotating mirror, and line by line, as the platform (aircraft or satellite) flies along track over an area. The range of a scan line is called the swath. The light reflected from the land surface reaches the rotating mirror, which rotates at a designed speed and thus views different positions on the ground across a swath during a rotation scan cycle. The rotating mirror diverts the incident light through the scanner optics and then the light is dispersed into several spectral beams by a spectral splitting device (a prism or interference filters). The multispectral beams are then received by a group of CCDs which sample the light at regular time intervals. When the scanner finishes scanning one swath, the satellite or aircraft has moved forward along its track to the position for the next scan. One scan swath can be a single image line or several image lines depending on the sensor design and the synchronisation between flying speed, swath width, altitude of the satellite or aircraft, and required image resolution. In this way, a scanner can image a large area using a limited number of CCDs. A scanner, though mechanically complicated, relies less on CCD technology.

Table 1 Comparison of spectral bands of TM/ETM+, ASTER, and SPOT

	Sensor systems								
	Terra 1 ASTER			*Landsat 3 7 TM/ETM+*			*SPOT 1 3 HRV, SPOT 4 HRVI, SPOT 5 HRG*		
Spectral region	*Band*	*Spectral range (μm)*	*Spatial resolution (m)*	*Band*	*Spectral range (μm)*	*Spatial resolution (m)*	*Band*	*Spectral range (μm)*	*Spatial resolution (m)*
VNIR				1	0.45 0.53	30			
	1	0.52 0.60	15	2	0.52 0.60		1	0.50 0.59	20
	2	0.63 0.69		3	0.63 0.69		2	0.61 0.68	
	3N	0.78 0.86		4	0.76 0.90		3	0.79 0.89	
	3B	0.78 0.86		ETM+ Pan	0.52 0.90	15	Pan	SPOT 1 3: 0.51 0.73	10
								SPOT 4: 0.61 0.68	10
								SPOT 5: 0.48 0.71	2.5 5
SWIR	4	1.60 1.70	30	5	1.55 1.75	30	4	1.58 1.75	20
	5	2.145 2.185		7	2.08 2.35		Band 4 is only in SPOT 4 HRVI (High Resolution Visible Infrared) and SPOT 5 HRG (High Resolution Geometric)		
	6	2.185 2.225							
	7	2.235 2.285							
	8	2.295 2.365							
	9	2.360 2.430							
TIR	10	8.125 8.475	90	6	10.4 12.5	TM			
	11	8.475 8.825				120			
	12	8.925 9.275							
	13	10.25 10.95				ETM+ 60			
	14	10.95 11.65							

Table 2 Some satellite high resolution broadband sensor systems

		Spatial resolution (m)			*Spectral range (μm)*		
Satellite	*Launch time and Status*	*Pan*	*MS*	*Hyp*	*Pan*	*MS*	*Hyp*
Ikonos 2	24 Sept. 1999 In operation	1	4		0.45 0.90	0.45 0.53 0.52 0.61 0.64 0.72 0.77 0.88	
QuickBird	18 Oct. 2001 In operation	1	4		0.45 0.90	0.45 0.52 0.52 0.60 0.63 0.69 0.76 0.89	
Orbview 3	26 June 2003 In operation	1	4		0.45 0.90	0.45 0.52 0.52 0.60 0.625 0.695 0.76 0.90	
Orbview 4	Failed at launch in Sept. 2001	1	4	8	0.45 0.90	0.45 0.52 0.52 0.60 0.625 0.695 0.76 0.90	0.45 2.50 200 bands

The MSS with 4 spectral bands on board Landsat 1–3 is a classical example of a mechanical scanner. It is a one-way scanner that scans in one direction of mirror rotation only, and with an empty return run. Such a design makes compensation of the Earth's rotation easier, as the Earth rotates a fixed distance along the swath direction for each scanning cycle. However, the inactive return would waste valuable time for imaging, resulting in a shorter dwelling time in the active runs and thus reduce spatial resolution.

The Thematic Mapper (TM) on board of Landsat 4–7 is a significantly improved scanner with six reflective spectral bands and one thermal band (Table 1). It is a two-way scanner that scans in both directions. For the same width of swath, the two-way scan allows the mirror to rotate slower than a one-way scan, thus increasing the dwelling time of CCDs at each sample position. This configuration improves both spatial and spectral resolutions. The geometric correction, to compensate for Earth rotation effects, is more complicated for TM than for MSS, however, because one scan direction is for, and the other, against the Earth's rotation, respectively.

Along-Track Push-Broom Scanner

With the rapid development of CCD technology, a more advanced scanner, the so-called 'push-broom scanner', has become dominant in passive sensor design since the successful launch of SPOT-1 in 1986. As shown in Figure 3, the key difference between a push-broom scanner and a mechanical scanner is that a push-broom scanner does not have a mechanical mirror for pixel-by-pixel scanning across along the swath direction. In a push-broom scanner, a line array panel of CCDs, covering the whole imaging swath, is mounted at the rear of the spectral dispersal device. Push-broom scanner images an area line by line across the whole track when the sensor platform (a satellite or an aircraft) flies along track, like pushing a broom forward to sweep a floor.

Since one image swath is generated simultaneously, the dwelling time for each CCD representing an image pixel can be as long as a whole swath scanning time for a mechanical scanner. With significantly increased dwelling time, a push-broom scanner achieves much higher resolution. Based on advanced CCD technology, the push-broom scanner is simpler than the mechanical scanner in structure, and the geometric correction is less complicated. Without a mechanical part, the system is robust and reliable. The number of CCDs in a line array in swath direction decides the size and the resolution of the imagery data generated. For instance, the SPOT HRV has 6000 CCDs per line for its panchromatic band while 3000 CCDs per line for its multispectral bands. The system, therefore, produces panchromatic images with 6000 pixels per line at 10 m resolution and multispectral images with 3000 pixels per line at 20 m resolution.

The push-broom design also allows greater flexibility to manoeuvre the sensor to point to off-nadir positions. As shown in the Figure 3, a mirror is fixed in front of the lens like a mechanical scanner. This mirror is not for scanning; it has several fixed angle positions. At different angles, the mirror directs the sensor system to view and image different scenes off-nadir along the track.

Digital Cameras

In the last few years, digital cameras have started to compete in the market of traditional film photog-

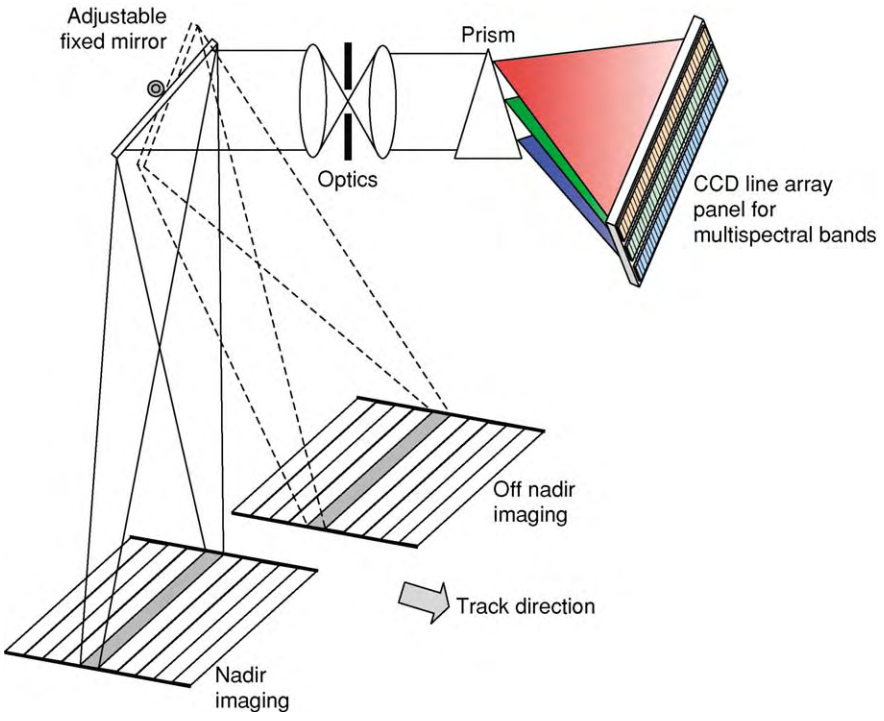


Figure 3 A schematic diagram of an along track push broom multispectral scanner.

raphy. A digital camera is built on a full 2D (two dimensional) CCD panel. The only difference between a digital camera and a film camera is that it records an image through a 2D CCD panel to a memory chip instead of using a film. The new generation of passive sensors will be largely based on digital camera technology that takes an image by an instantaneous frame rather than scanning line by line. The consequence is that the constraints on platform flight parameters can be relaxed, image resolution (spatial and spectral) can be improved, and image geometric correction processing can be further streamlined.

Broadband Reflective Multispectral Imagery

Broadband reflective multispectral sensors and thermal sensors are addressed separately being based on different physics. In practice, however, these two groups of passive sensors are often mounted in the same instrument as different bands. This is true for many sensor systems, such as TM, Enhanced Thematic Mapper Plus (ETM+), and ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) (Table 1), used for geological applications.

Broadband reflective multispectral image data are the most widely used for geological studies and regarded as an effective operational tool for mapping tectonic structures and lithology, for mineral exploration, for logistic planning, and for field survey navigation. Extending the scope of applications from Earth observation to planetary study, this group of sensors are often the major tools for data collection from other planets. A typical example is the exploration of Mars.

The American Landsat satellite family TM, ETM+, and French SPOT (Système Pour 1'Observation de la Terre) satellite family high resolution visible (HRV) are the most commonly used Earth observation systems, providing broadband multispectral and panchromatic imagery data of global coverage. As shown in Table 1, these types of sensor systems operate in: the visible spectral range with bands equivalent to three primary colours; blue (380–440 nm), green (440–600 nm), and red (600–750 nm); the near infrared (NIR) range (750–1100 nm), and the short wave infrared (SWIR) range (1550–2400 nm). The number of bands and the spectral width in VNIR (visible near infrared) and SWIR spectral ranges depend on the atmospheric windows and sensor design. For instance, the spectral width of SWIR bands needs to be much wider than visible bands if the same spatial resolution is to be achieved, as is the case for TM bands 5 and 7, because the solar radiation in the SWIR spectral region is significantly weaker than that in the visible spectral range.

In general, the 'broad' band means that the spectral range is significantly wider than a few nanometres, except in the case of hyperspectral sensor systems described below. Broadband reflective multispectral sensor systems are a successful compromise between spatial resolution and spectral resolution. With relatively broad spectral bands, such a sensor system offers reasonable spatial resolution with high SNR (Signal Noise Ratio) and meanwhile, operating in a wide spectral range from VNIR to SWIR, such a system can provide images of multispectral bands, enabling identification of major ground objects and discrimination of various land cover types. With the dramatic improvement of sensor technology, from mechanical scanners to push-broom scanners, and to CCD digital cameras, the spatial resolution of broadband multispectral imagery is improving all the time. For Sun synchronous near polar orbit satellites, the spatial resolution of this type of sensors has been improved from 80 m (Landsat MSS) in the 1970s to a few metres, on current systems, as shown by the examples in Table 2.

The VNIR spectral range is used by nearly all broadband reflective multispectral sensor systems. This spectral range is within the solar radiation peak and thus allows the generation of high resolution and high SNR images. It also covers diagnostic features of major ground objects such as the few examples below:

- *Vegetation*: minor reflection peak in green, absorption in red, and then significant reflection peak in NIR, often called the 'red edge'.
- *Water*: strong diffusion and penetration in blue and green, and nearly complete absorption in NIR.
- *Iron oxide (red soils, gossans, etc.)*: absorption in blue and high reflectance in red.

Many satellite sensor systems did not use the blue band, in order to avoid strong Rayleigh scattering effects occurring in the atmosphere that can make an image 'hazy'. A popular configuration is to offer three broad spectral bands in green, red, and NIR, such as the case of SPOT and the most recent commercial high spatial resolution space-borne sensors (Tables 1 and 2). In a computer graphic system, we can display the three band images as a colour composite with NIR displayed in red, red in green, and green in blue. Such a colour composite image is called a standard false colour composite. This image is the most effective for mapping healthy vegetation.

The SWIR spectral range is regarded as the most effective for lithological mapping and mineral exploration, because most rock types have high reflectance in 1.55–1.75 μm and clay minerals (often products of

alteration) related to mineralisation, have various absorption features in the spectral range 2.0–2.4 μm. These two SWIR spectral ranges, corresponding to Landsat TM band 5 and 7, are preferred by geologists. SWIR sensor systems are technically more difficult and complicated because the SWIR detectors have to operate at low temperatures, which therefore require a cooling system (a liquid nitrogen coolant or a cryocooler) to maintain the detectors at about 80 K.

With six broad reflective spectral bands, Landsat TM provided the best spectral resolution among broadband sensor systems for many years. The six broad reflective spectral bands are very effective for discrimination of various ground objects but they are not adequate to achieve specific identification in particular rock types and major mineral assemblies relating to mineral deposits. This requires a sensor system with much higher spectral resolution at a few nanometre bandwidth to resolve their subtle spectral signatures. This demand has led to the development of the hyperspectral system.

ASTER (a push-broom scanner for VNIR and SWIR bands), on board the Terra-1 satellite, is a representative of a transitional sensor system between broadband multispectral and hyper-spectral narrowband sensing. It is an integrated system of three scanners: a VNIR push-broom scanner with three broad spectral bands; a SWIR push-broom scanner with six narrow spectral bands; and a TIR across-track mechanical scanner with five thermal bands (Table 1). The system combines good spatial resolution in the VNIR bands and high spectral resolution in SWIR and thermal bands and was specifically designed for geological applications. The three 15 m resolution VNIR bands are adequate for distinguishing broad categories of land cover such as vegetation, water, red soils, urban areas, superficial deposits, and rock outcrops, while the six narrow SWIR bands of 30 m resolution provide potential for mapping major mineral assemblies of rock types and alterations. Another unique advantage of ASTER is that it has along track stereo capability. The VNIR scanner has a backward viewing telescope to take NIR images beside its nadir telescope for the three VNIR bands. Thus nadir and backward viewing NIR images are taken simultaneously, forming along track stereo image pairs. The along track stereo image pairs enable generation of DEM (Digital Elevation Model) data.

Thermal Infrared (TIR) Sensors

Various minerals, rock types, and other ground objects have different thermal properties, such as thermal inertia, thermal emission, and thermal absorption. For instance, quartz has distinctive high emissivity around 9 μm and strong absorption in 10–11 μm; dolomite can be distinguished from limestone by its significant lower emissivity in 8–11 μm.

There are several airborne thermal IR sensor systems for example, the Thermal Infrared Multispectral Scanner (TIMS) with 6 bands in the 8.2–12.2 μm spectral region developed in 1982. Of the satellite systems, the Landsat Thematic Mapper system has a broad thermal band, TM band 6, at a wavelength range of 10.4–12.5 μm. So far the only space-borne multispectral thermal system is ASTER on board of Terra-1 satellite, which has 5 thermal bands as shown in Table 1.

In general, a broadband TIR sensor operating in the 8–14 μm spectral range images the radiation temperature of the land surface while the narrower band of the multispectral thermal imagery data detect the thermal spectral signatures of materials on the land surface. It is important to note that daytime thermal images are fundamentally different from the night time thermal images. Daytime thermal images are dominated by topographic features, governed by the geometry between slopes and solar radiation described previously, while night time thermal images are solely determined by emission from the Earth surface. They therefore detect the thermal properties of ground materials more efficiently.

For both systems, TM and ASTER, the spatial resolutions of their thermal bands are significantly lower than their reflective multispectral bands, as shown in Table 1. One reason is that the interaction between thermal energy (or heat) and the atmosphere is more complicated than the case of VNIR and SWIR energy. Heat can be transmitted in air, not only by radiation, but also by air circulation. Both solar radiation to the Earth in TIR spectral range and direct thermal emission from the Earth are very weak, compared with the energy intensity of the Earth reflected solar radiation in the VNIR and SWIR spectral ranges.

Most thermal sensors are of cross-track mechanical scanner type, as shown in Figure 4. The major difference of a thermal scanner from a reflective multispectral scanner is that it needs a cooler system to maintain the TIR detector at very low temperature for maximum sensitivity. For instance, the thermal sensor of Landsat Thematic Mapper is surrounded by liquid nitrogen at 77 K stored in an insulated vessel. In the ASTER system, a cryocooler is used to maintain the detectors for TIR bands at 80 K. A blackbody plate is used as an on-board calibration reference that is viewed before and after each scan cycle, providing estimation of instrument drift. This is essential to maintain the accuracy and consistency of a TIR instrument. The temperature sensitivity of a modern TIR sensor system can be as high as 0.1 K.

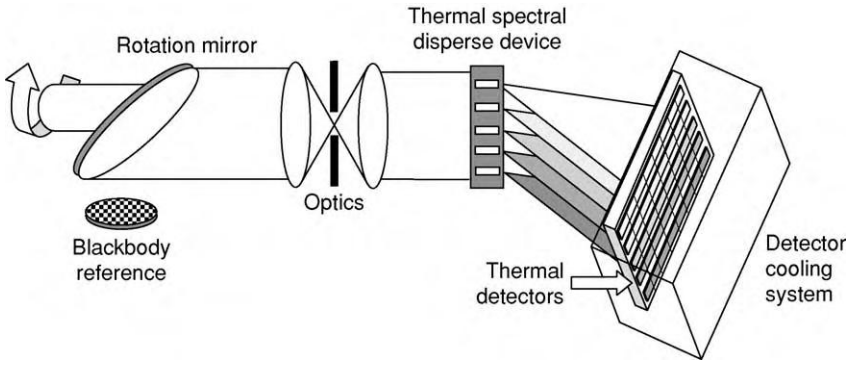


Figure 4 A schematic diagram of a thermal scanner.

To fully represent the sensitivity, many thermal IR sensors use 10–12 bit quantisation to record data, for example, ASTER multispectral thermal band images are 12 bit integer data.

Hyper-Spectral Sensors (Imaging Spectrometers)

The development of passive sensor technology is always aimed at higher spatial and spectral resolutions. Hyperspectral sensor systems represent a revolutionary development increasing passive sensor spectral resolution to as high as a few nanometres. This means that they are capable of generating nearly continuous spectral profiles for land surface materials. A hyperspectral sensor system is an imaging spectrometer. It combines the spatial imaging capacity of an imaging system with the spectral analytical capabilities of a spectrometer. Such a sensor system may have up to several hundreds narrow spectral bands with spectral resolution on the order of 10 nm or narrower. Imaging spectrometers produce a complete spectrum for every pixel in the image. The resulting dataset allows identification of materials rather than mere discrimination as with broadband sensor systems. The data processing methodology and strategy are, therefore, different from that employed for broadband images in many aspects. It is more important to analyse the spectral signature for each pixel, rather than to perform general image enhancement to improve visualization, though the latter is still essential for data overview.

One of the earliest and the most representative hyper-spectral systems is JPL's Advanced Visible Infrared Image Spectrometer (AVIRIS) (Table 3). Figure 5 shows the general principle of hyperspectral systems. The incoming EMR from the land surface goes through the sensor optics and is then split into hundreds (e.g., 224 for AVIRIS) of very narrow spectral beams by a complicated spectral dispersion device (e.g., interference filters) and finally the spectral beams are detected by arrays of CCDs corresponding to, say 224, spectral bands. A hyperspectral system can be either an across-track mechanical scanner with a small number of detectors for each band, or an along-track push-broom scanner with a panel of hundreds of line arrays of CCDs. Hyperspectral sensors are so far operating in either VNIR only or VNIR and SWIR spectral ranges (Table 3), although a similar evolution is taking place in TIR systems.

There is only one experimental, space-borne hyperspectral sensor, Hyperion. Several satellites with a hyperspectral system on board have failed, for various reasons, before the sensor system even started to operate (Table 2). Nevertheless, with the success of several airborne hyperspectral systems (Table 3), it is only a matter of time and money before hyperspectral satellite system become operational.

Table 3 Some airborne Hyper spectral sensors

Instrument	*Spectral range (nm)*	*Band width (nm)*	*No. bands*
AVIRIS	400 2400	9.6	224
AIS	1200 2400	9.6	128
SISEX	400 2500	10 20	128
HIRIS	400 2500	10 20	128
MIVIS	430 1270	20 54	102

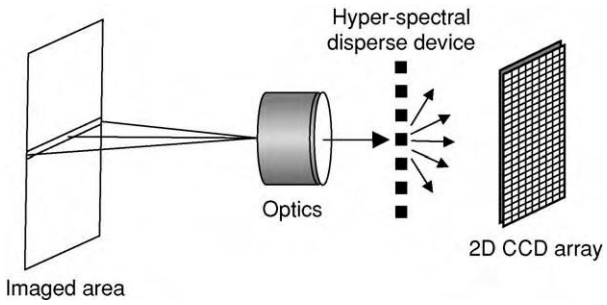


Figure 5 Principle of an imaging spectrometer (hyper spectral system).

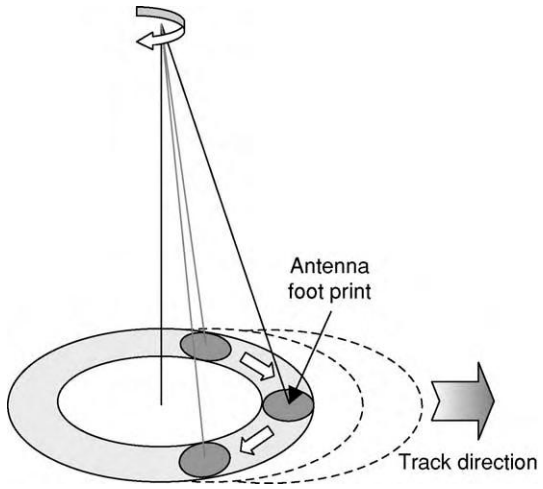


Figure 6 Conical scanning mechanism.

Passive Microwave Sensors

Thermal radiation from natural surfaces, such as the land surface, extends from its peak in the thermal infrared region into the microwave region. An Earth observation microwave imagining radiometer operates in this spectral region to receive the microwave radiation from the Earth. As a passive sensor system, it is important to understand that a microwave radiometer is fundamentally different from a radar sensor, that is a ranging system. The only resemblance between the two is that they both operate in the microwave spectral range. A passive microwave sensor system, a microwave imaging radiometer, works more like a thermal sensor system. It collects emitted energy radiated from the Earth in the microwave spectral range that provides useful information relating to surface temperature, topography, and material dielectric properties. This type of sensor has been used for global temperature mapping, polar ice mapping, and regional soil moisture monitoring.

A space-borne microwave imaging radiometer is often a multichannel scanner, such as the SMMR (Scanning Multichannel Microwave Radiometer), which was on board of Seasat and Nimbus in 1978 and the Microwave Imager on board the Tropical Rainfall Measuring Mission (TRMM) satellite in 1997. It is composed of an antenna together with its scanning mechanism, a receiver, and a data handling system. The received emitted microwave signals are highly related to the observation angle and the path length in the atmosphere. Ensuring that these scanning parameters are constant, thus they can significantly increase the accuracy of the derivation of surface parameters from the microwave brightness temperature. A conical scan configuration is, therefore, popular for passive microwave scanners. As shown in Figure 6, the antenna observation direction is offset at a fixed angle from a nadir rotating scan around the vertical (nadir) axis and thus sweeps the surface of a cone. If the scan is configured for a full 360°, double coverage fore and aft of the spacecraft is obtained. With the moving forward of a satellite along its orbit, a belt of land surface is imaged. Obviously, in a conical scanning geometry, the observation angle and distance to any scanned position are constants.

Space-borne passive microwave scanners are usually of low spatial resolution from several kilometres to several tens of kilometres, because of weak signals in the microwave spectral range.

See Also

Geological Field Mapping. **Remote Sensing:** Active Sensors.

Further Reading

Drury SA (2001) *Image Interpretation in Geology,* 3rd edn. Chelternam: Blackwell Science.

Elachi C (1987) *Introduction to the Physics and Techniques of Remote Sensing*. New York: John Wiley & Sons.

Lilesand TM and Kiefer RW (2000) *Remote Sensing and Image Processing,* 4th edn. New York: John Wiley & Sons.

Colwell RN (ed.) (1983) *Manual of remote sensing,* Vol.1, 2nd edn. Theory, instruments and techniques. Falls Church, Va: American Society of Photogrammetry.

Robinson N (1966) *Solar Radiation*. Amsterdam, London and New York: Elsevier.

Sabins FF (1996) *Remote Sensing: Principles and Interpret ation,* 3rd edn. Basingstoke: WH Freeman.

ASTER user Handbook v.2, on line at http://asterweb.jpl.nasa.gov/.

Landsat 7 handbook, on line at http://landsat.gsfc.nasa.gov/.

Remote sensing learning sources of Canada Centre for Remote Sensing, on line at http://www.ccrs.nrcan.gc.ca/ccrs/learn/learn e.html

The SPOT Payload, on line at http://www.spotimage.fr/automne modules files/standard/public/p229 file LINKEDFILE Spot payload.pdf.

RIFT VALLEYS

See **AFRICA: Rift Valley**

ROCK MECHANICS

JP Harrison, Imperial College London, London, UK

Introduction

Although rock has been used as an engineering material since time immemorial, rock mechanics as a specialist engineering discipline has existed only since about the middle of the twentieth century. The particular engineering endeavours in which rock mechanics is used run from traditional applications such as mining and tunnelling, through more modern applications such as large underground caverns for various civilian uses (e.g., fuel storage and transport infrastructure), to deep wellbores for petroleum extraction, and on to novel issues such as underground disposal of nuclear waste. There is also a growing realization that the principles of rock mechanics as used by engineers are applicable to problems in geological science, such as fault movement, seismicity, and plate tectonics. So, techniques that have their origin in engineering are more and more being applied to problems such as the stability of natural slopes, fracture development in extensional basins, and fluid flow through deforming rock.

Rock mechanics arose, and continues to develop, because, as far as engineers are concerned, rock is an awkward material. By this is meant that it is not like the usual well-specified and controlled engineering materials of steel, concrete, or polymers. Its natural basis means that it is likely to be variable in its character and to contain features that are weak; both of these properties present particular difficulties to an engineer. Thus, rock mechanics as an engineering discipline is concerned with the development of an understanding of rock as an engineering material, such that safe and economic design approaches can be identified. Generally, the geology of a particular rock is of secondary importance in rock mechanics: it is the behaviour of the rock, rather than its geological characteristics, that is of principal interest to the engineer. Notwithstanding this, an understanding of the geological setting of a rock is useful in rock mechanics, because a range of geological processes (e.g., the development of schistosity, or the form of a weathering profile) can impact greatly on the behaviour of rock as an engineering material.

It is now recognized that a key issue in understanding and applying rock mechanics is that, on the scale of most engineering projects, rock is fractured. As a result, rock often cannot be considered as a continuum, and a rational approach for dealing with the resulting discontinuum is to consider rock mechanics as a subject that comprises a number of clearly defined topics. To help understand the subject, these topics can be placed into three major categories: (1) the basic components of *in situ* stress, and the physical characteristics and behaviour of both fractures and intact rock; (2) rock masses as an amalgamation of these components; and (3) the interaction between rock engineering activities and rock masses.

The Basic Components: *In Situ* Stress, Fractures, and Intact Rock

The Earth's crust is almost everywhere subjected to a non-zero state of *in situ* stress. This state of stress is disturbed when any engineering operation takes place, and the perturbed stresses can cause a range of unwanted effects, such as stress-related failure of the rock. As a result, assessment of the *in situ* state of stress is a critical component of rock mechanics. Over the years, various techniques have been developed that allow engineers to determine, with varying degrees of completeness, the state of stress in a rock mass. Two of the most commonly used techniques are hydraulic fracturing and various so-called overcoring methods. In hydraulic fracturing, a short length of a borehole is filled with fluid and is pressurized. A fracture forms in the rock of the borehole wall when the fluid pressure reaches a critical value, and the magnitude of this pressure is used to calculate the *in situ* stress (Figure 1). Overcoring makes use of sophisticated strain measuring devices that are fixed with epoxy cement into a small-diameter borehole (Figure 2). A larger diameter borehole is then drilled coaxially with the first, and the device recovered. Continuous recording of the strains induced within device, and hence within the core of rock, then allows the *in situ* stress to be calculated.

Existing fractures are particularly important to engineers because they are usually much weaker than is the intact rock, and are hydraulically very conductive (i.e., readily permit fluid flow). Assessing the geometry, mechanical, and hydraulic characteristics of the fractures at a project site are thus key engineering activities. Simplified models are usually used to describe such characteristics. For example, fractures may be considered to be planar and of infinite extent, and to occur in sets in which the fractures are mutually parallel. The complete rock mass may then be considered to comprise a number of such sets. For

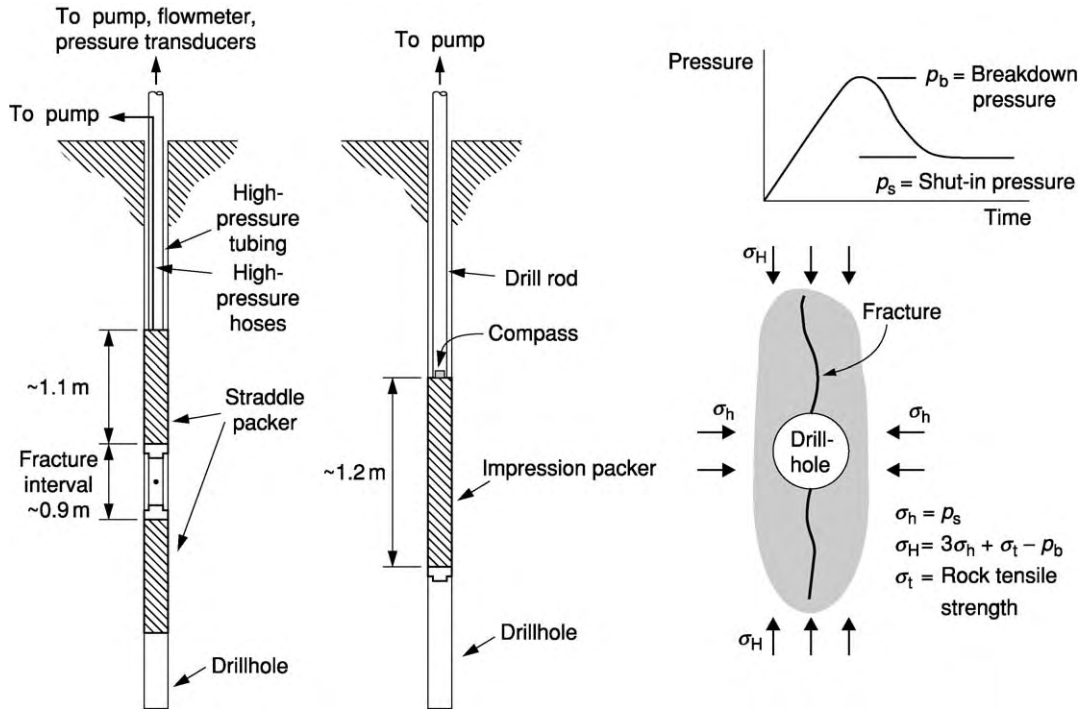


Figure 1 Hydraulic fracturing for estimation of *in situ* stress. Two longitudinal sections of the boreholes show the downhole equipment; the cross section shows the form of the induced fracture and the equations governing its development.

Figure 2 An overcoring device for stress estimation, sectioned after recovery from the test borehole. The control wires enter at the left; the part of the cell carrying the strain gauges, and which is cemented into the borehole, is at the right.

mechanical characteristics, a simple zero-cohesion Coulomb model is often used for frictional strength, with the shear and normal stiffnesses of a fracture being modelled as linear or non-linear springs. Hydraulic models range from those whereby each fracture is simply represented by a pair of smooth parallel plates, through to sophisticated statistical models in which the complex variation of fracture aperture is reproduced. Although many of the models of fracture geometry and strength are simple, they nevertheless allow valuable insight into the behaviour of fractured rock masses. For example, superposition of the frequencies of a number of fracture sets allows assessment of the variation of frequency with orientation through the rock mass (Figure 3), with the result that the direction in which a borehole, for example, would intersect most fractures can be seen. This has ramifications for optimal borehole orientation in endeavours such as water, petroleum, and geothermal energy extraction.

Most rock mechanics investigations involve determination of various mechanical properties of rock, particularly strength and elastic moduli. Laboratory testing of intact rock primarily makes use of cylindrical specimens, loading them in a state of triaxial compression in specialized servo-controlled machines that allow strain, rather than stress, to be the independent variable (Figure 4). Using strain as the independent variable allows the post-peak behaviour of a rock to be determined; this is critical for accurately establishing the engineering behaviour of structures, such as support pillars in underground mines, that undergo strain-controlled loading. Mechanical testing of fractures is usually undertaken using a shear box, in which a specimen of rock containing a through-going fracture is mounted and subjected to both normal and shear loads (Figure 5).

The strength of intact rock is often modelled using a linear Coulomb criterion, but because the relation

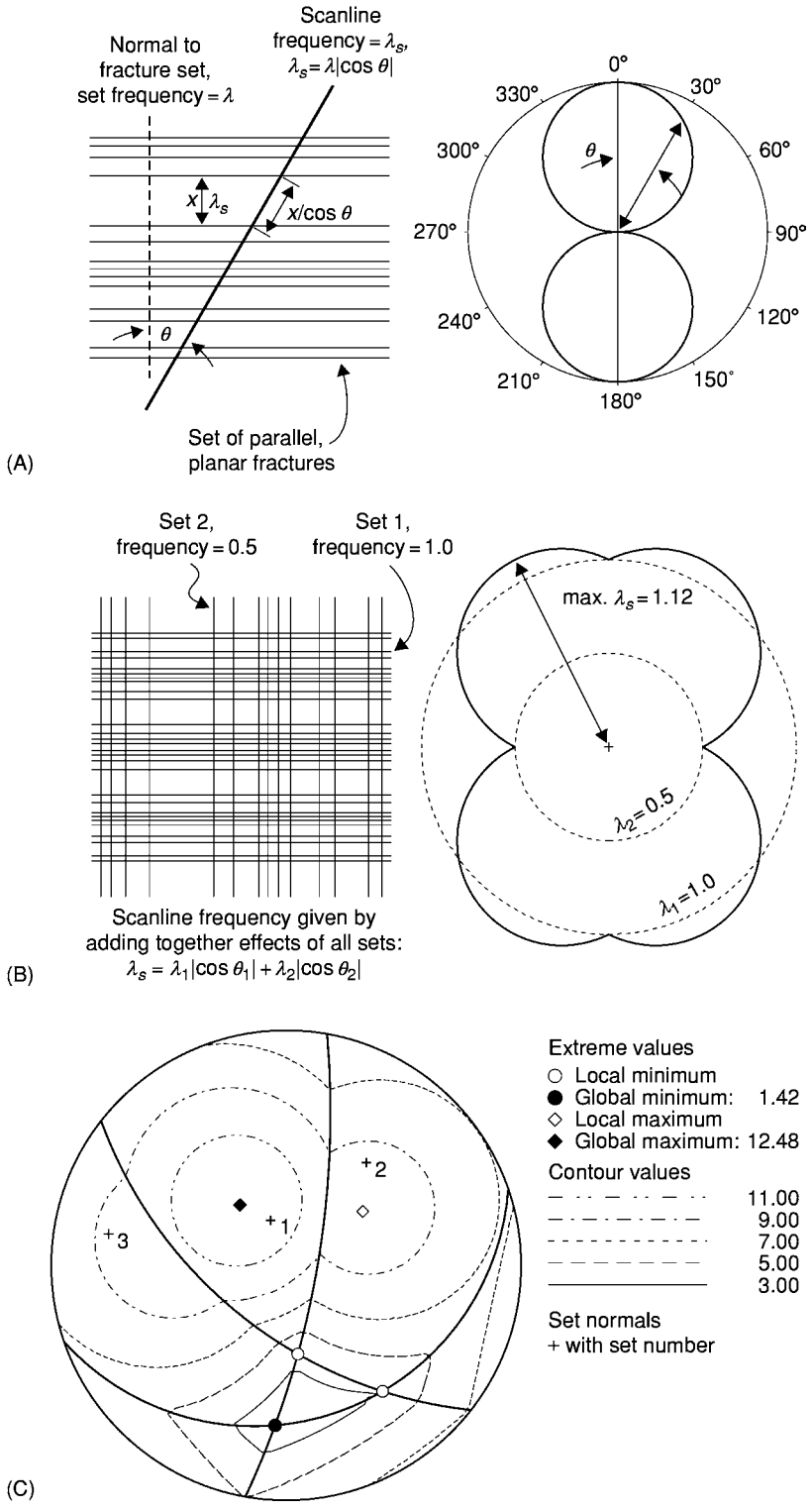


Figure 3 Predicting fracture frequency within a rock mass. (A) Frequency variation due to one set. (B) Frequency variation due to two orthogonal sets. (C) Frequency variation in three dimensions due to three sets. The fracture set orientation (dip direction/dip angle) and frequency values are as follows: set 1, 161/23 and 7.72 m ^{1}; set 2, 218/58 and 3.07 m ^{1}; set 3, 100/75 and 5.34 m ^{1}.

between strength and confining stress is known to be non-linear, increasing use is made of empirical non-linear strength criteria. One of the most popular of these is the Hoek-Brown criterion (Figure 6). Although the shear strength of fractures is similarly non-linear, for ease of calculation a simple linear Coulomb criterion is usually used for modelling.

The Engineering Material: Rock Masses

Rock masses can be usefully considered as being a composite material that comprises blocks of intact rock surrounded by fractures. Because both of these elements have their own mechanical characteristics, an estimate of the strength of the composite rock mass can be obtained by superposition of the component behaviours (Figure 7). This single-plane-of-weakness theory is widely used both for understanding the general strength of a rock mass and for determining the forces that tunnel support is required to supply. Although this approach can be used to examine the effect of more than one set of fractures, the usual approach for predicting the strength of fractured rock masses is through the use of non-linear empirical criteria. Over the years, many such criteria have been developed, but the one that appears to have become most widely used is the Hoek-Brown criterion. In its simplest form, this is a straightforward extension of the criterion used for intact rock, but with the addition of a parameter that describes the degree of fracturing of the rock mass (Figure 8).

Figure 4 A servo controlled testing machine. To the left is the stack of control electronics. In front of the loading frame (to the right) is an assembly comprising a triaxial test cell, axial strain transducers, and load cell.

It is often appropriate to consider intact rock as being impermeable, with the result that any fluid flow through a rock mass occurs along the fractures. A full, three-dimensional analysis of flow through a rock mass is not trivial, but a two-dimensional analysis can be undertaken relatively simply using a computer spreadsheet to solve simplified equations of flow. Figure 9 shows the results of an analysis undertaken for a two-dimensional network, in which the fractures have been modelled as pairs of parallel smooth plates, all with the same aperture. Rather than simply assessing the strength of a rock mass, an alternative approach is to assess its general engineering behaviour. Although this is a more nebulous concept compared to strength, through the use of a few, easily determined properties of a rock mass, a number of rock mass classification schemes have been developed that allow calculation of a single parameter that somehow uniquely quantifies the mass. Two of the most popular schemes in widespread engineering use are the so-called Q and rock mass rating (RMR)

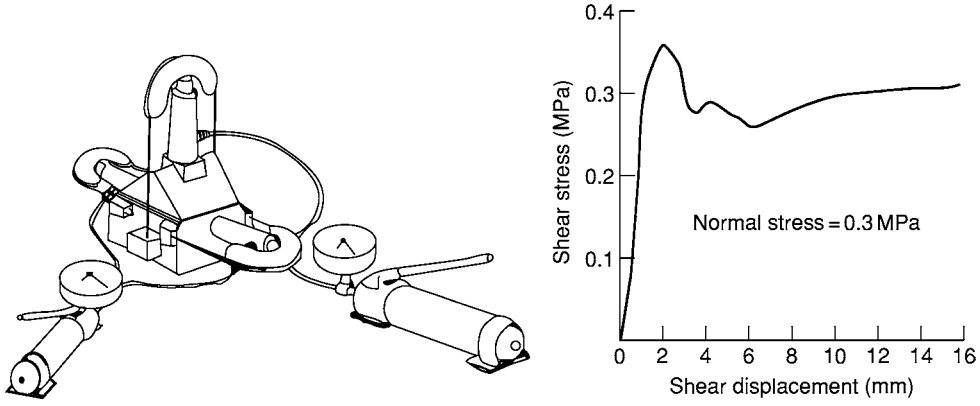


Figure 5 Schematic of a shear box for strength testing of fractures and a typical stress/shear displacement curve. The shear box is loaded by means of the hydraulic hand pumps (one for normal, the other for shear load).

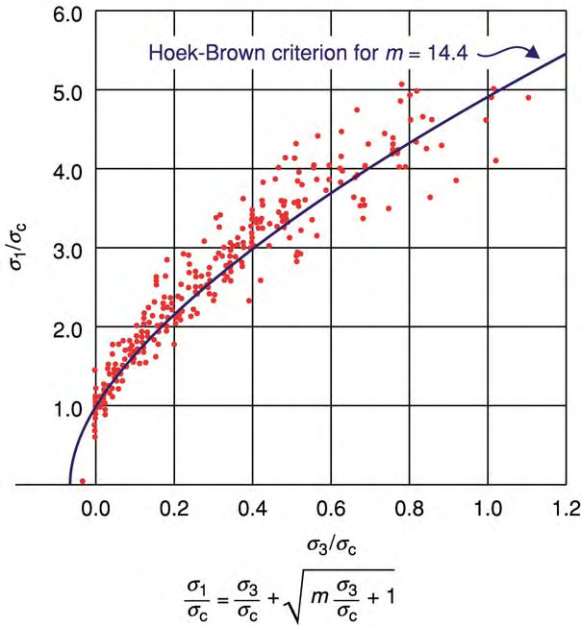


Figure 6 The equation and plot of the Hoek Brown criterion for intact rock, where σ_c is unconfined compressive strength of the particular rock under consideration.

systems (Figure 10); for both of these, extensive tables are readily available that indicate what parameter values should be used for particular conditions. Over the years, correlations have been developed between these rock mass classification values and various engineering aspects such as ease of mechanical excavation, blasting requirements, and tunnel support requirements. As a result, and due to their inherent simplicity, the use of rock mass classification schemes is widespread.

Basing an appreciation of rock masses on the concept of a composite material comprising intact rock and fractures as two separate materials leads automatically to an awareness of the variability of rock masses with respect to location and orientation. Techniques for the treatment and understanding of variability are still in their infancy, but two acronyms that are useful in helping to concentrate on the issues are CHILE and DIANE. The first of these represents 'continuous, homogeneous, isotropic, and linearly elastic' behaviour; the second, which is somewhat more contrived, represents 'discontinuous, inhomogeneous, anisotropic, and not elastic' behaviour. CHILE rock is a hypothetical material, but its existence is relied on extensively in rock mechanics. For example, the presence of a CHILE material may be

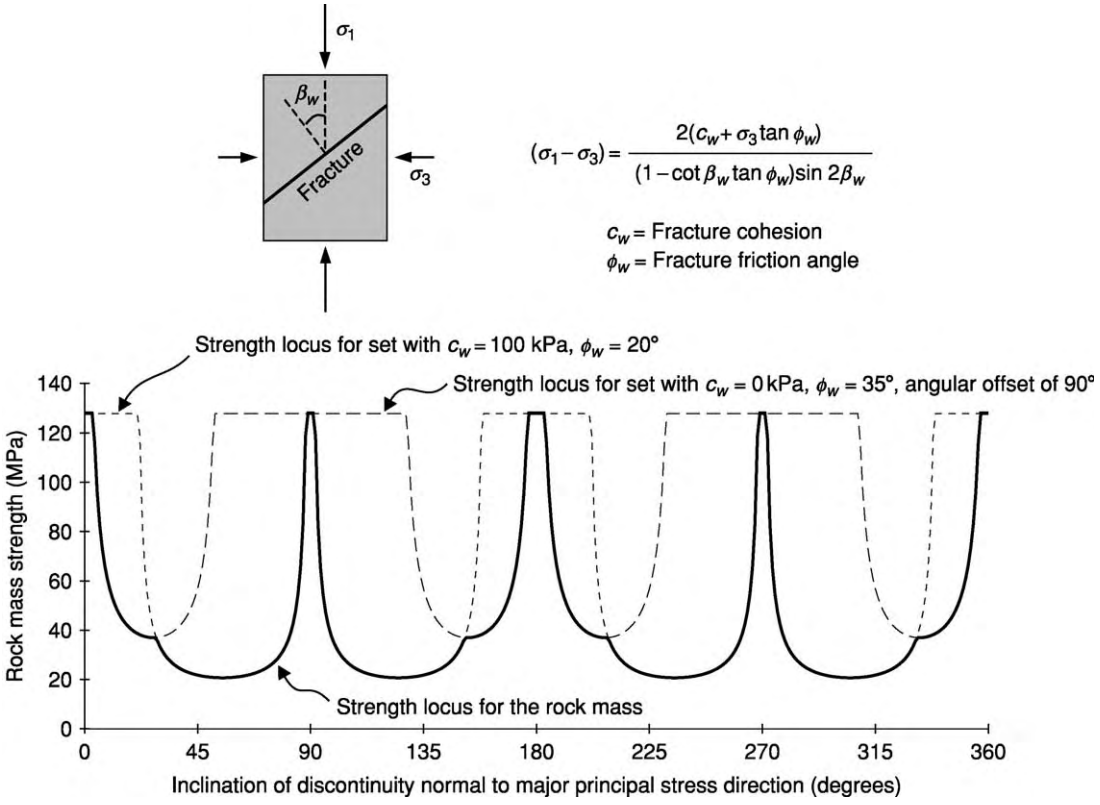


Figure 7 The single plane of weakness theory and its extension to the strength of a fractured rock mass.

The Hoek–Brown criterion for fractured rock masses $\frac{\sigma_1}{\sigma_c}=\frac{\sigma_3}{\sigma_c}+\sqrt{m\frac{\sigma_3}{\sigma_c}+s}$	Values of s	Dolomite, limestone, marble	Mudstone, siltstone, shale	Sandstone, quartzite	Andesite, dolerite, rhyolite	Gabbro, gneiss, granite
		Values of m				
Intact rock samples Laboratory-sized specimens free from fractures	1.0	7	10	15	17	25
Very good quality rock mass Tightly locking undisturbed rock with unweathered fractures, spacing 1 to 3 m	0.082	2.40	3.43	5.14	5.82	8.56
Good quality rock mass Fresh to slightly weathered rock, slightly disturbed fractures, spacing 1 to 3 m	0.00293	0.575	0.821	1.231	1.395	2.052
Fair quality rock mass Several sets of moderately weathered fractures, spacing 0.3 to 1 m	0.00009	0.128	0.183	0.275	0.311	0.458
Poor quality rock mass Numerous weathered fractures, spacing 30 to 500 mm and with some gouge	0.000003	0.029	0.041	0.061	0.069	0.102
Very poor quality rock mass Numerous heavily weathered fractures, spacing less than 50 mm, gouge filled	0.0000001	0.007	0.010	0.015	0.017	0.025

Figure 8 The Hoek Brown criterion for fractured rock masses, with typical values of the strength constants m and s for various qualities of different rock types.

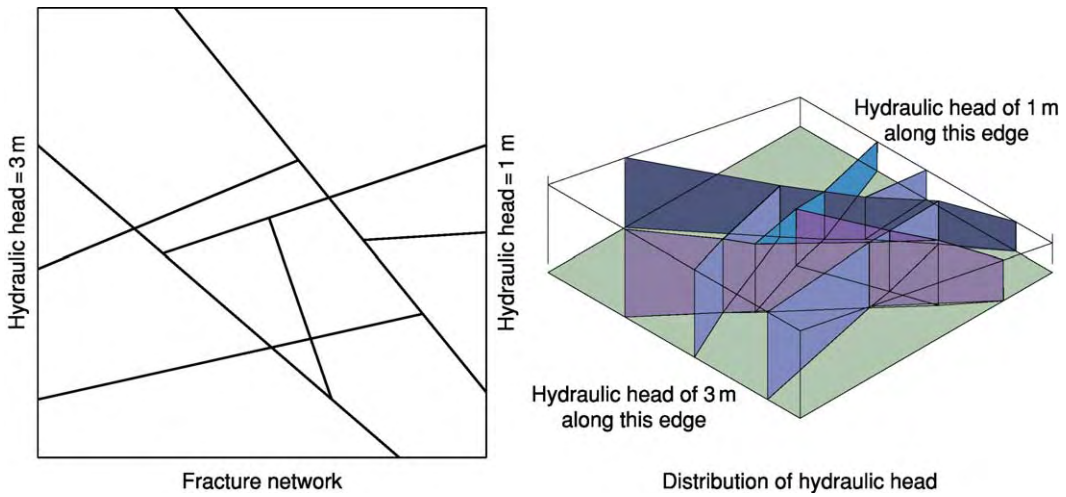


Figure 9 Distribution of hydraulic head in a two dimensional fracture network.

invoked in the determination of *in situ* stress, characterization of intact rock, and analysis of stresses induced around underground excavations. The DIANE nature of real rock masses means that the effects of fractures, variability with respect to location and orientation, and the fact that stress changes will usually lead to irreversible deformations in the material must always be considered. A further important consequence of the DIANE nature of rock masses is the scale effect. In general, because of the presence of fractures, it is found that large volumes of rock and small specimens of intact rock have different properties. This means that *in situ* rock mass strength and stiffness will be much lower than that determined in the laboratory. Similarly, the large-scale permeability of a fractured rock mass is very likely to be much greater than that of a specimen of intact rock. The scale effect is thought to apply to other properties, such as *in situ* stress, but in these cases the consequence of it is less predictable. For example, under

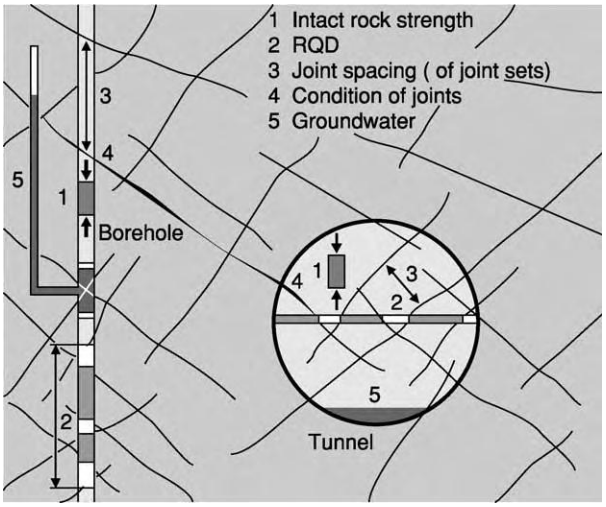


Figure 10 Schematic of typical rock mass properties included in a rock mass classification scheme, and how they may be assessed in both boreholes and tunnels. RQD, Rock quality designation.

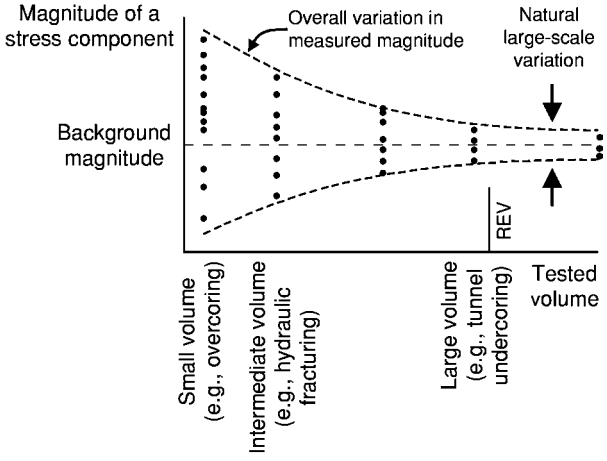


Figure 11 Illustration of scale effect and the representative elemental volume (REV) for *in situ* stress.

different circumstances, the measured state of stress in a small volume of rock could be less, the same as, or greater than the actual state of stress in a large volume. In all cases with the scale effect, the concept of a representative elemental volume (REV) is useful. As Figure 11 shows, there may be substantial variation in measured values or properties at small scales, but as the REV is approached the variation reduces until, at the REV and above, the variation is sufficiently small that it can be disregarded. Figure 11 also shows that the distribution of values at any particular scale may differ from those at another scale. This implies that it is usually not possible to use measurements of properties at one scale to estimate properties at another scale, and hence exemplifies the ideal requirement of performing rock mechanics tests on a volume of rock comparable to that of the final engineering structure. Unfortunately, this is seldom possible due to the constraints of cost and time.

Rock Engineering Topics

The development of rock mechanics concepts has in large part been driven by the requirement to solve problems in rock engineering. For example, although the rock engineering associated with mining and quarrying operations has a long history, the desire to exploit reserves at greater depths and hence in more onerous geomechanical environments has required a more rigorous approach than custom and practice can deliver. Endeavours such as the drilling of wellbores thousands of metres deep for petroleum extraction, or the requirement to ensure the long-term stability of steep rock slopes many tens of metres high adjacent to highways, has required additional developments in rock mechanics. Development efforts continue, driven by problems without precedent, such as those concerning the ultra-long-term stability of engineering designs for radioactive waste repositories. Regardless of the specific engineering application and its environment, the fundamental effects that an excavation has on a rock mass must always be borne in mind. These are summarized in Figure 12.

Fundamental techniques used in rock engineering are applicable to a range of engineering topics. These

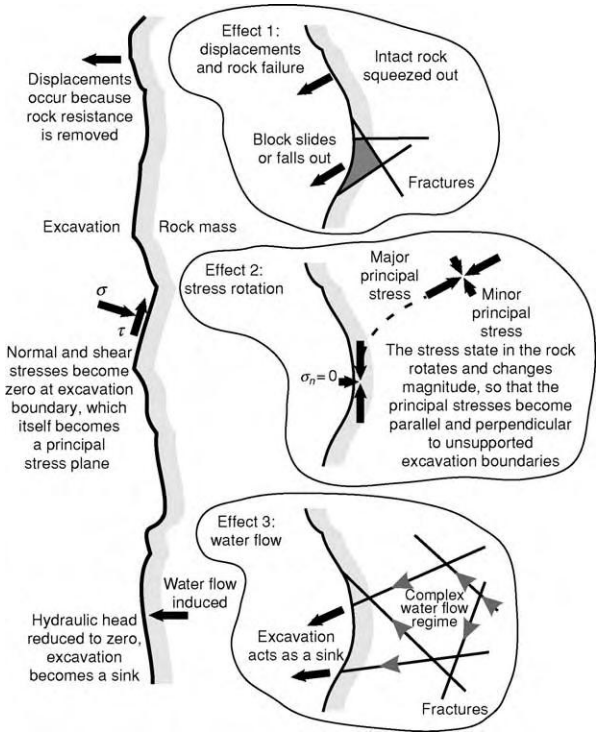
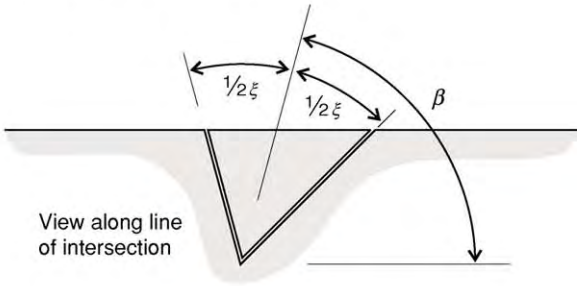


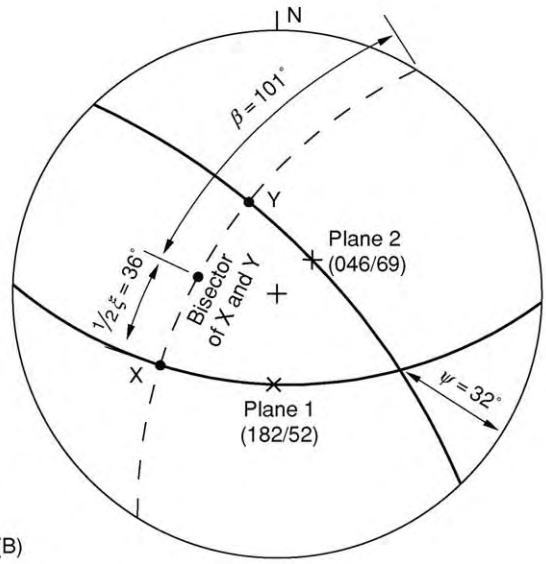
Figure 12 The fundamental effects of an excavation on a rock mass.

(A)

Factor of safety against sliding =

$$\frac{\sin\beta}{\sin \frac{1}{2}\xi} \times \frac{\tan\phi}{\tan\psi_i}$$





(B)

$$F_W = \frac{\sin 101}{\sin 36} \times \frac{\tan 29}{\tan 32} = \frac{0.982}{0.588} \times \frac{0.554}{0.625} = 1.48$$

Figure 13 Illustration and analysis of wedge instability in a rock face. (A) Basic equation for factor of safety against sliding, where β and ξ are shown in the sketch, ψ_i is the plunge of intersection and ϕ is the friction angle of the fractures. (B) The factor of safety against wedge sliding, F_W, is 1.48 for the case of two fractures of orientation 046/69 and 182/52 and a friction angle of 29°. Dashed line plane whose normal is the intersection of planes 1 and 2.

techniques can conveniently be classified in terms of those applicable to rock that is obviously discontinuous, and those that are strictly valid only for continuous rock. This effectively separates the applications into those for near-surface structures (such as slopes, foundations, and tunnels) and those for structures at depth (such as mine excavations and petroleum wellbores).

Engineering in Fractured Rock

Near-surface engineering is often controlled entirely by the strength and geometry of fractures. It is common, and conservative, to assume that the strength of fractures is purely frictional, i.e., the fractures have zero cohesion. This has an added advantage in that, in the absence of any external forces such as those generated by water pressure, an analysis of static equilibrium reduces to one that is governed by the orientation of the fractures. These orientations may conveniently be determined using standard hemispherical projection techniques. Of course, the blocks of rock that exist in a rock mass may possess a great many shapes and sizes. In order to develop a tractable problem, only tetrahedral blocks of the maximum size that can fall or slide into an excavation are analysed. Such tetrahedral blocks are generally formed from either two fracture surfaces and two excavation surfaces or from three fracture surfaces and one excavation surface. Polyhedral blocks (i.e., those possessing five or more surfaces) are usually ignored, both for analytical convenience and because such a block can be cut from, and hence will be smaller and less critical than, a tetrahedral block.

An example of how the hemispherical projection is used in rock block stability analysis is given by the case of a rock slope containing a wedge of rock that is delineated by two fracture surfaces, as shown in Figure 13. This figure also shows how the stability of this wedge is a function of the friction angle of the fractures and three angles that relate to the geometry of the fractures. These latter three angles can be measured directly and quickly from the hemispherical projection. Further simplification of this analysis is possible by ignoring the stabilizing effect of the wedge factor, i.e., the quotient $\sin\beta/\sin 1/2\xi$. This leads to kinematic analyses, in which only the feasibility of instability, given the interaction between the fracture and excavation geometry, is considered.

Kinematic analyses test for whether the geometry of the fractures and the excavation surface are capable of defining blocks that can be released into

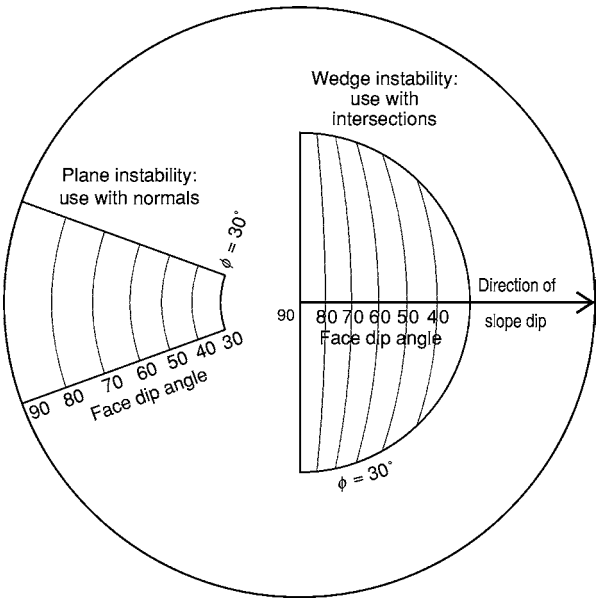


Figure 14 Example of kinematic assessment overlay for rock slope instability due to plane and wedge sliding.

the excavation, and therefore may be unstable. Again, for no external forces and fractures possessing only frictional strength, the kinematic criteria relate to orientations only, with the result that, once more, the hemispherical projection can be used rapidly to assess stability. Figure 14 shows an overlay that can be used for assessing instability due to plane and wedge sliding in a rock mass for which the friction angle is 30°. To use this overlay, the fracture data (both poles and great circles) are plotted on a sheet of tracing paper and placed on a hemispherical projection net, with a second sheet of tracing paper on which the assessment overlay is drawn placed over the top. The overlay is then rotated to identify those slope orientations that give rise to either poles or intersections falling within the unstable zones. Because the overlay also shows various slope angles, it can quickly be used to select a slope angle that avoids the kinematic conditions leading to instability.

A similar analysis can be undertaken for underground excavations. Figure 15 shows three hemispherical projections illustrating the unstable kinematic conditions that apply to a tetrahedral block formed by three fractures and the horizontal roof to a tunnel or cavern. These blocks are identified on the projection as areas bounded by three great circles. If the vertical direction (i.e., the centre of the projection) lies within the projection of the block, this indicates that the block is able to fall. In the case of sliding instability, the direction of sliding is the steepest of either a line of maximum dip or the intersection of two planes. For blocks that are identified as being

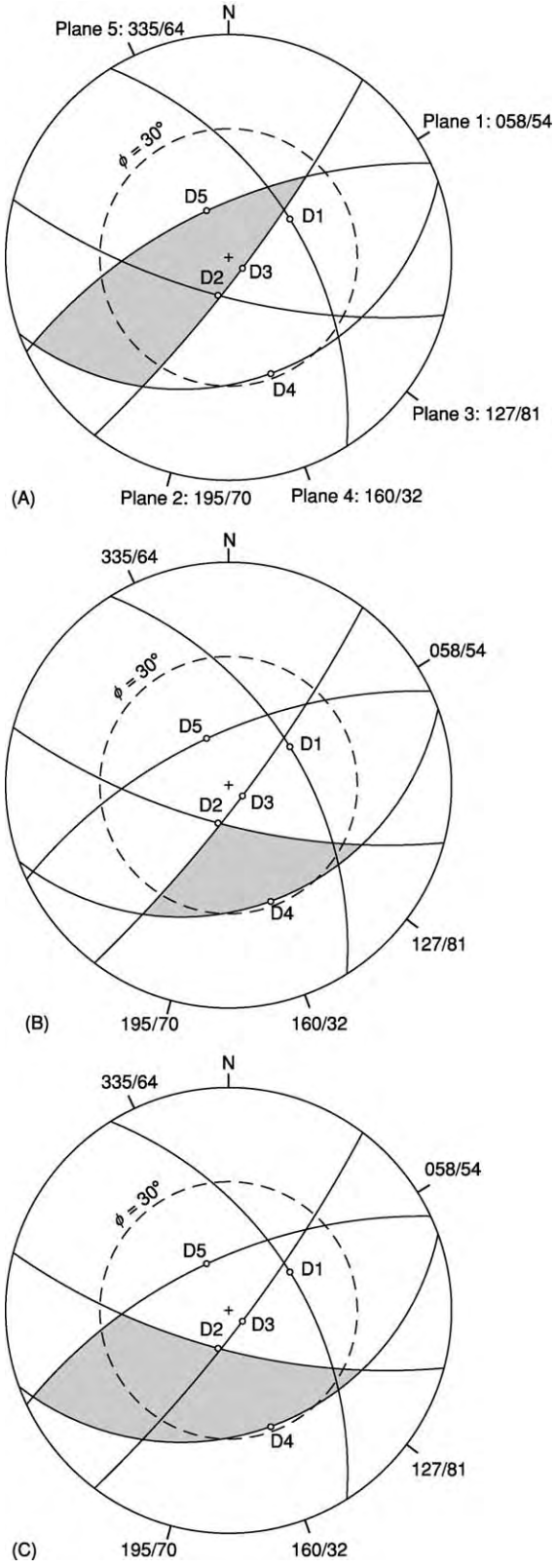


Figure 15 Identification of falling, sliding, and stable blocks in the roof of an underground excavation. (A) Block 345 falls verti cally. (B) Block 234 slides on the intersection of planes 2 and 3. (C) Block 245 slides on the line of maximum dip of plane 2.

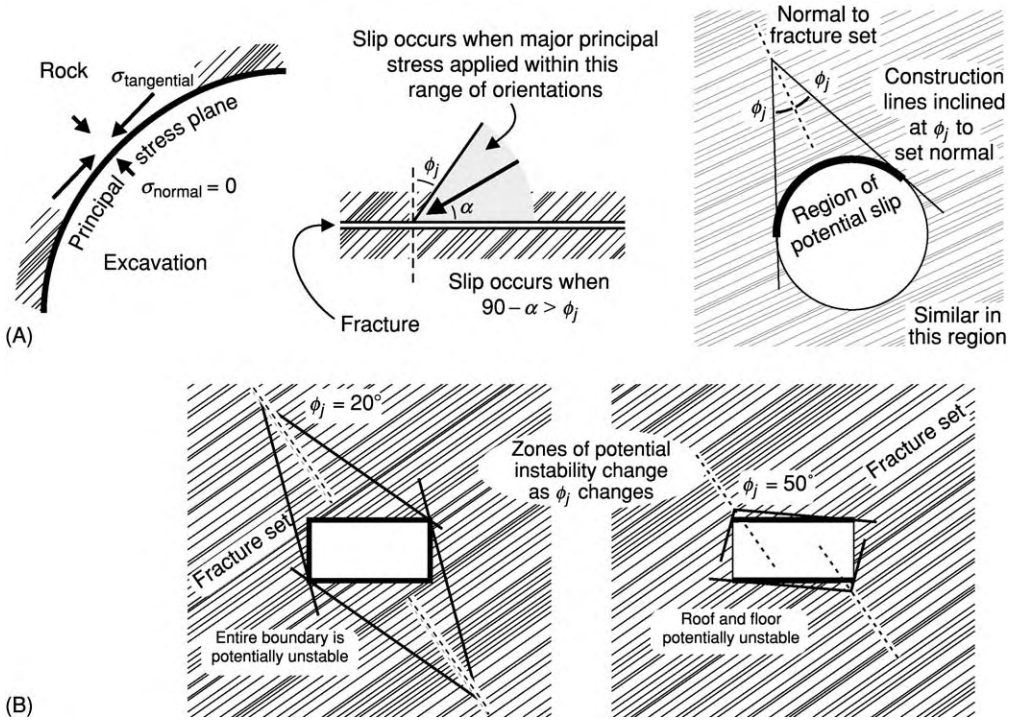


Figure 16 The ϕ_j theory for underground excavations. (A) The fundamental stress conditions leading to slip, and how these are identified in the ϕ_j theory. (B) Application to underground excavations in stratified materials with different frictional properties.

unstable, a further rigorous equilibrium analysis can be undertaken in order to determine their size and hence support requirements. These analyses can be time consuming and prone to error if undertaken by hand, and so are usually performed by computer programs.

Another form of kinematic analysis that is widely used is the so-called ϕ_j theory. This uses the principle that, if shearing along a fracture is to be avoided, the direction of action of the major principal stress acting on the fracture must lie within an angle ϕ_j, the friction angle, of the normal to the fracture. Now, close to an underground excavation the major principal stress is tangential to the boundary. This means that a geometric analysis can be used to identify those zones on the boundary where the orientation of the major principal stress is such as to induce shearing. This is illustrated in Figure 16, for the case of a rock mass that contains one dominant set of fractures, and hence can be assumed to be transversely isotropic.

Techniques for the analysis of foundations on fractured rock are not as straightforward as for slopes or underground excavations. However, with most engineering structures it is possible to use foundations that are sufficiently large as to impose stresses that are small enough to prevent instability from occurring. If it can be assumed that the rock beneath the foundation is continuous, then the induced stresses may be determined using analytical solutions developed from the theory of elasticity. This approach is widely used in soil mechanics, wherein the assumption of a continuum is often valid, but it is seldom used in rock mechanics for the reason that most near-surface rock masses are fractured. However, the concept of 'bulbs of pressure', or contours of major principal stress, from these analyses is useful in rock mechanics. As the ϕ_j theory shows, as the direction of action of the major principal stress is constrained by the presence of a fracture, the shape of the resulting stress contours is controlled by the fracture geometry. Figure 17 shows a simplified solution for the radial stress that is applicable to the case of rocks that contain one dominant set of fractures, and so can be assumed to be transversely isotropic. The contours of major principal stress are influenced strongly by the fracture geometry, showing that foundations on rock can induce high stresses in zones far removed from the foundation.

Engineering in Continuous Rock

For underground excavations in continuous rock, or in locations where the *in situ* stresses are sufficiently high as to allow a fractured rock to act as if it were continuous, it is the stress induced around the excavation that leads to instability. Simple analytical solutions for the induced stresses are available for only a restricted set of excavation geometries, namely, circular and elliptical openings that are in a state of plane strain. This last constraint requires the excavations to be long, of constant cross-section, and with

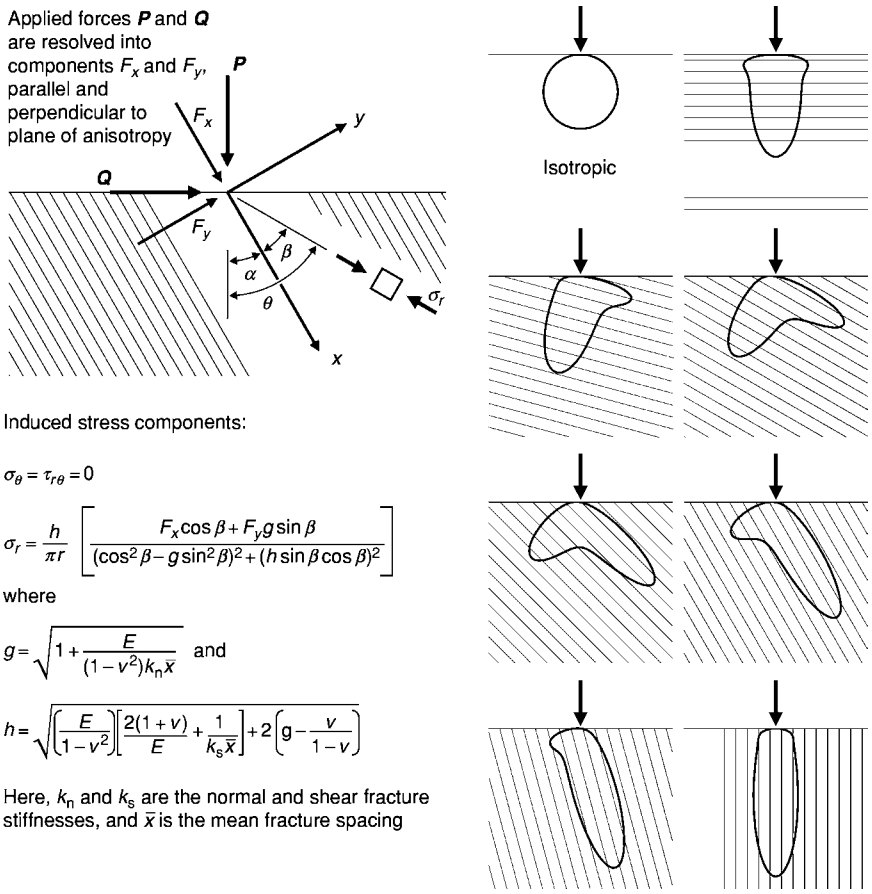


Figure 17 Simplified relations for contours of constant radial stress for a foundation in stratified material resembling a transversely isotropic rock.

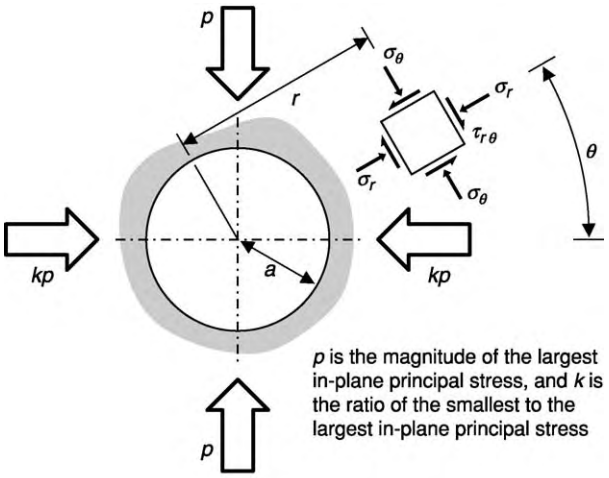
their axial direction parallel to one of the *in situ* principal stresses. Despite these severe restrictions on geometry, the fact that engineers are commonly required rapidly to assess the stability of underground excavations means that these solutions are used extensively.

Figure 18 shows a cross-section through a circular opening and presents the well-known Kirsch equations for the induced-stress components. These equations are used for many applications involving the analysis of stress around tunnels, wellbores, and mine shafts. As an example, Figure 18 also shows simplified relations for the specific case of the maximum and minimum values of tangential stress induced at the boundary of an excavation, and how these components vary as the *in situ* stress ratio varies. The figure shows how the maximum and minimum stresses are induced under uniaxial conditions, represented by $k = 0$, and that for isotropic conditions, represented by $k = 1$, the induced stress is constant around the boundary.

For the cases of stress induced around an elliptical excavation, a particularly elegant solution is available for the tangential stress induced at the boundary. This is shown in Figure 19. Also, for the particular case of an elliptical opening with a horizontal major axis, the extreme boundary stresses, which are induced in the side walls, roof, and floor, are also given. These final two equations can be equated, and doing so shows that the tangential stress is constant around the boundary in the case of an elliptical opening that has an aspect ratio equal to the *in situ* stress ratio. This elegant result is a valuable design guide for engineers wishing to minimize the range of stresses induced around the boundary of an excavation.

Numerical Analysis

With all of the fundamental analyses that are routinely used in rock mechanics and rock engineering, various simplifying assumptions are required in order to make the problem tractable. These simplified procedures are perfectly acceptable for the majority of applications, but for applications requiring an accurate understanding of the behaviour of a structure, more complex analyses are necessary. These are performed using computer programs that implement sophisticated numerical techniques. The techniques most often used in rock mechanics are finite element, boundary element, and finite difference methods for continua, and various discrete element methods



$$\sigma_r = \frac{p}{2}\left[(1+k)\left(1-\frac{a^2}{r^2}\right)-(1-k)\left(1-4\frac{a^2}{r^2}+3\frac{a^4}{r^4}\right)\cos 2\theta\right]$$

$$\sigma_\theta = \frac{p}{2}\left[(1+k)\left(1+\frac{a^2}{r^2}\right)+(1-k)\left(1+3\frac{a^4}{r^4}\right)\cos 2\theta\right]$$

$$\tau_{r\theta} = \frac{p}{2}\left[(1-k)\left(1+2\frac{a^2}{r^2}-3\frac{a^4}{r^4}\right)\sin 2\theta\right]$$

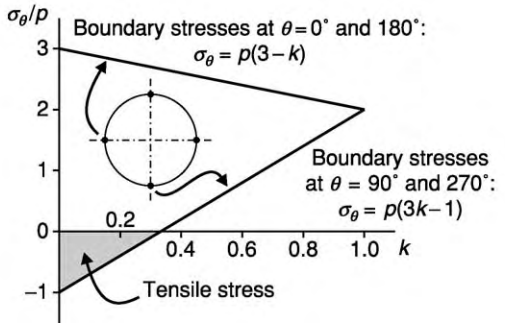
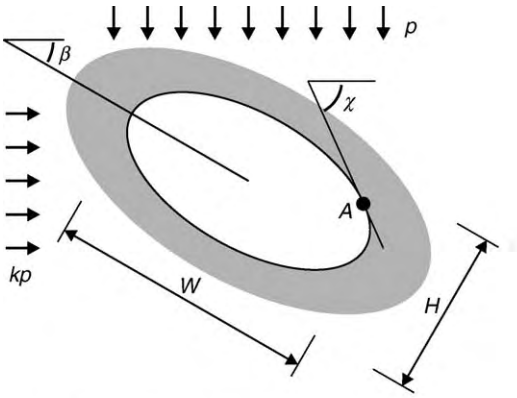


Figure 18 The Kirsch solution for stresses around a circular opening in isotropic continuous rock, under conditions of plane strain.



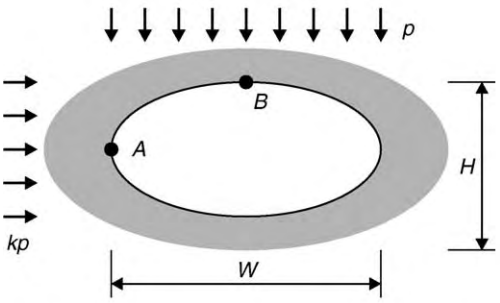
Boundary stress at point A:

$$\sigma_\theta = \frac{p}{2q}\left\{(1+k)\left[(1+q^2)+(1-q^2)\cos 2(\chi-\beta)\right]\right.$$

$$\left.-(1-k)\left[(1+q^2)\cos 2\chi+(1-q^2)\cos 2\beta\right]\right\}$$

where $q = W/H$,

p is the magnitude of the largest in-plane principal stress, and k is the ratio of the smallest to the largest in-plane principal stress



Extreme boundary stresses:

$$\sigma_A = p(1-k+2q)$$

$$\sigma_B = p\left(k-1+\frac{2k}{q}\right)$$

Figure 19 Simplified relations for the tangential stress at the boundary of an elliptical opening in isotropic continuous rock, under conditions of plane strain.

for discontinua. The background theory to these methods is complex, and their use is not straightforward. However, they are able to produce models and simulations of great accuracy and detail, and as researchers continue to implement new developments, the results obtained using them become ever more realistic. There is no doubt that the use of such procedures will become more widespread in the future.

See Also

Geological Engineering. **Geotechnical Engineering**. **Quarrying**. **Soil Mechanics**.

Further Reading

Amadei B and Stephansson S (1996) *Rock Stress and Its Measurement*. London: Chapman & Hall.

Harrison JP and Hudson JA (2000) *Engineering Rock Mechanics: Illustrative Worked Examples*. Oxford: Pergamon.

Hoek E and Bray JW (1981) *Rock Slope Engineering*, 3rd edn. London: Institution of Mining and Metallurgy.

Hoek E and Brown ET (1980) *Underground Excavations in Rock*. London: Institution of Mining and Metallurgy.

Hudson JA (ed.) (1993) *Comprehensive Rock Engineering*. Oxford: Pergamon Press.

Hudson JA and Harrison JP (1997) *Engineering Rock Mechanics: An Introduction to the Principles*. Oxford: Pergamon.

Jaeger JC and Cook NGW (1979) *Fundamentals of Rock Mechanics*. London: Chapman & Hall.

ROCKS AND THEIR CLASSIFICATION

R C Selley, Imperial College London, London, UK

Introduction: What on Earth is a Rock?

To a geologist the term 'rock' is applied to all the solid materials of the Earth, whether they are hard enough to hit with a hammer or soft unconsolidated gravel, sand, or mud. The nomenclature and classification of rocks is a prerequisite to understanding them. One of several definitions of science is that it is organized knowledge. It is the organization and classification of knowledge that provides insight into processes and adds predictive value. The Periodic Table of the Elements illustrates these points admirably. Classification is an integral part of geology in all its forms. It extends from palaeontology, where the classification of fossils according to the binomial nomenclature of Linnaeus is a prerequisite to demonstrating evolution and understanding palaeoecology, to mineralogy, where minerals are grouped into like families according to their chemistry and crystallography.

The object of this chapter is to review the fundamental classification of rocks as a foundation for discussing igneous, sedimentary, and metamorphic rock nomenclature and classification (*see* **Sedimentary Rocks:** Mineralogy and Classification, **Metamorphic Rocks:** Classification, Nomenclature and Formation).

Rocks are traditionally grouped into three main classes, according to their genesis. These will now be defined and described. All generalizations are dangerous, however, including this one. Thus, a detailed examination of the classification (*see* **Igneous Processes**) of rocks soon reveals discrepancies and inconsistencies (*see* **Metamorphic Rocks:** Classification, Nomenclature and Formation). It may, therefore, be helpful to examine these subsequently.

A Classification of Rocks

Rocks are of three genetic types: igneous (formed from cooling magma), sedimentary (formed by the breakdown of pre-existing rock), and metamorphic (formed by the action of heat and pressure on pre-existing rock) (Figure 1 and Table 1).

Igneous Rocks

Magma is molten material from deep beneath the surface of the Earth. Through a variety of processes discussed elsewhere in this encyclopaedia, magma may move up through the crust towards the surface. It may cool very slowly beneath the surface to form coarsely crystalline rocks. These include silica-rich rocks, such as granite (*see* **Igneous Rocks:** Granite), and ferromagnesian-rich rocks, such as gabbro. If the magma reaches the surface, it erupts in volcanoes (*see* **Volcanoes**), and the melt cools quickly to form microcrystalline or even glassy rocks. Depending on their chemistry and texture, volcanic rocks may be classified as silica-rich rhyolite or ferromagnesian-rich basalt.

Thus, owing to their mode of formation, igneous rocks tend to be structureless when viewed at outcrop or in a hand specimen. Under the microscope they are seen to consist of a random arrangement of interlocking crystals of a range of different minerals, which formed when the melt crystallized. Porosity is generally absent.

Sedimentary Rocks

Sedimentary rocks are formed from the detritus of pre-existing rocks, which may be igneous, metamorphic, or sedimentary. The way in which rock is

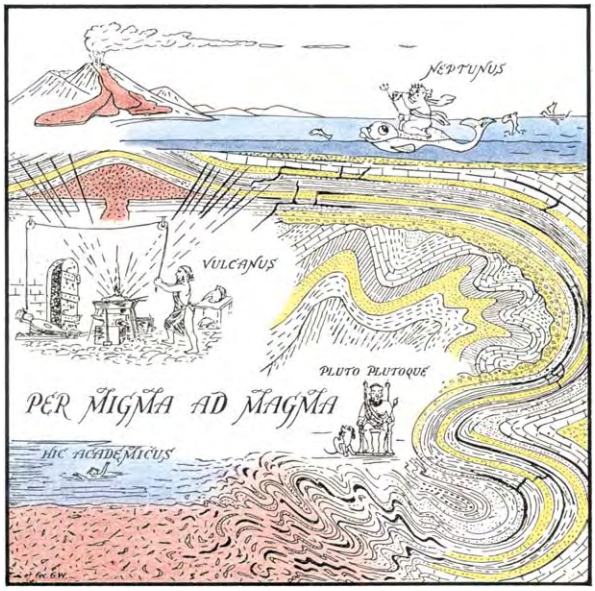


Figure 1 The Rocks Display'd. An illustration of the modes of formation of sedimentary, igneous, and metamorphic rocks, accompanied by their classical deities Neptune, Vulcan, and Pluto (together with his synonymous dog). Reproduced by permission of The Geologists' Association from *Proceedings of The Geologists' Association*, Wilson G, in Read HH, Meditations on Granite 2, volume 55, pp. 4 93. Fig. 1. © 1944 The Geologists' Association.

Table 1 The three main classes of rock, their origins, and examples

Class of rock	*Origin*	*Examples*
Sedimentary	Deposition of particles of pre existing rocks	Gravel, sand, and mudstone
	Chemical replacement and precipitation	Limestone dolomite and salt
Metamorphic	Heated by igneous intrusions	Hornfels
	Heated and subjected to high pressure by deep burial	Slate, schist, and gneiss
Igneous	Volcanic eruptions	Basalt, andesite, and rhyolite
	Deep cooling of magma	Granite, diorite, and gabbro

weathered, eroded, transported, and deposited is discussed in detail elsewhere (*see* **Weathering, Sedimentary Processes:** Erosional Sedimentary Structures, **Sedimentary Environments:** Depositional Systems and Facies). Sediments possess a wide range of particle sizes, ranging from boulders to clay, and of chemical compositions, including silica, lime, and ferromagnesian volcanic detritus. These parameters of particle size and composition are used to classify sedimentary rocks (*see* **Sedimentary Rocks:** Mineralogy and Classification). Sedimentary rocks commonly exhibit two properties that may be used to differentiate them from igneous and metamorphic rocks. Where they crop out at the surface of the Earth, sedimentary rocks generally show stratification. The strata indicate successive episodes of deposition. Layering is usually absent from igneous rocks, as discussed above, but is found in some metamorphic rocks, as discussed below. When examined under the microscope, sedimentary rocks are generally seen to consist of particles. Void space (porosity) is commonly present between the constituent grains. If there are interconnected pores, these give the rock permeability. Permeability allows fluids to migrate through the rock and enables rock and soil to drain. Fossils are found only in sedimentary rocks, some of which are, indeed, made up of nothing else.

Sedimentary rocks include sands, including terrigenous silica-rich sandstone, and carbonates, termed limestone. Sediments composed of rounded pebbles are termed conglomerates, while those composed of angular clasts (fragments) are termed breccias.

Metamorphic Rocks

Metamorphic rocks are so termed from the Greek *meta-*, meaning altered, and *morphos-*, meaning shape. The word is applied to rocks, of whatever previous origin, that have been changed as a result of being exposed to high temperature, high pressure, or both. Two types of metamorphism are recognized, regional metamorphism, and thermal or contact metamorphism. When magma moves up through the crust it creates a metamorphic aureole in the adjacent cover rocks. This is termed contact or thermal metamorphism. By contrast, rocks that are subjected to high temperatures and pressures during deep burial are said to have undergone regional metamorphism.

Common thermally metamorphosed rocks include hornfels and quartzite. Regionally metamorphosed rocks include schist and gneiss.

Thermally metamorphosed rocks are often crystalline and devoid of porosity. They may be massive or layered, according to the character of the original rock. Regionally metamorphosed rocks are also commonly crystalline and devoid of porosity. They are, however, often banded. The banding seldom resembles the regular stratification of sediments, but often takes the form of curvaceous foliation and cleavage (*see* **Regional Metamorphism**).

Rock Classification: A Simple Summary

Igneous and metamorphic rocks are both normally composed of a mosaic of interlocking crystals that formed at high temperatures and pressures; they thus usually lack porosity and permeability. In the petroleum industry igneous and metamorphic rocks are generally grouped together under the term 'basement', indicating that it is pointless to continue drilling deeper into such tight rocks. Igneous rocks are generally massive and structureless, while metamorphic rocks are often banded.

Sediments, on the other hand, are deposited with space between the constituent particles. Thus, most sedimentary rocks, by their very nature, tend to be porous and permeable. Sediments are deposited episodically and thus characteristically exhibit stratification. They are also the only one of the three main classes of rock to contain fossils. Igneous, metamorphic, and sedimentary rocks may therefore be differentiated from one another using macroscopic structure and microscopic texture as shown in Figure 2.

Complications and Anomalies

When the tripartite classification of rocks was introduced at the beginning of this article it was pointed out that, like all classifications, it has complications

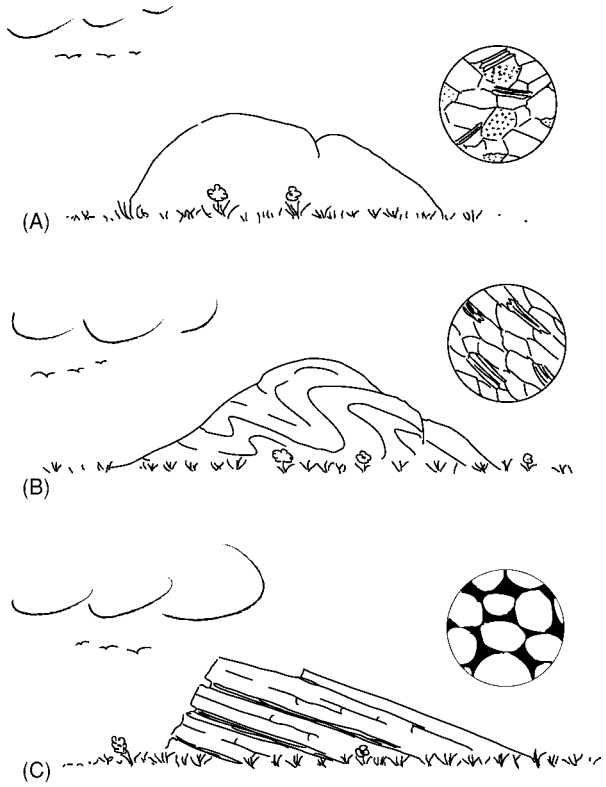


Figure 2 A rough guide to how igneous, metamorphic, and sedimentary rocks may be differentiated from one another by macroscopic structure (left) and microscopic texture (right). As discussed in the text, however, there are exceptions to these simple guidelines. (A) Igneous rocks are generally structureless in the field, with a random mosaic of crystals when viewed in thin section under the microscope. (B) Metamorphic rocks are gener ally banded (often curvaceous) in the field, with a linear arrange ment of crystals when viewed in thin section under the microscope. (C) Sedimentary rocks are generally stratified in the field and exhibit rounded grains with intergranular pores when viewed in thin section under the microscope.

and anomalies. These will now be examined for igneous, metamorphic, and sedimentary rocks in turn.

Igneous Rock Anomalies

When describing igneous rocks it was stated that they form from cooling magma and are generally composed of a mosaic of interlocking randomly arranged crystals. Thus, in hand specimen, and at outcrop, igneous rocks present a massive appearance. They generally lack the foliation of metamorphic rocks and the stratification of sediments. Now for the exceptions, which occur in both volcanic lavas and major intrusions.

Lava is extruded from a volcano episodically, on time-scales ranging from thousands of years to minutes. Major sequences of lava flows give rise to a very typical landform, with plateaus of lava bounded by cliffs in which successive lava flows can be discerned, separated from one another by weathered soil horizons that formed between individual flow events. Superficially this gives rise to landscapes analogous to those created by stratified sedimentary rocks. Plateau basalts are widespread around the world, from the Deccan Traps of India to the Tertiary volcanic province of the Scottish Hebridean islands. When examined closely, lavas commonly exhibit flow-banding on a variety of scales. This includes the 'ropy' structure of some basalt flows, the so-called 'pahoehoe' lavas of Hawaii, and a diverse range of structures that are morphologically similar to the post-depositional structures of sedimentary rocks (*see* **Sedimentary Processes:** Post-Depositional Sedimentary Structures). For example the Devonian rhyolites of the Glen Coe cauldron subsidence in Scotland exhibit excellent examples of slumps, slides, load-casts, and flame structures.

It was also noted earlier that igneous rocks generally lack porosity. Some gassy lavas, however, consolidate with contained bubbles of gas, giving rise to what are termed vesicular lavas. Vesicular lavas thus exhibit porosity, but, unless the bubbles are abundant and interconnected, there is no significant permeability.

Deep-seated igneous intrusions also sometimes show features indicative of sedimentation. These are most commonly found in ultrabasic ferromagnesian-rich intrusives, such as the gabbroic Skaergaard complex of Greenland and the Bushveld intrusion of South Africa. These intrusions show a gravity segregation of minerals, with the densest, such as chromite and iron ores, at the base, overlain by ferromagnesian minerals, and then feldspars above. It is clear that the minerals have crystallized out of a melt, and have settled under gravity. There is more. This gravity separation may occur not only on an intrusion-wide scale, but also on a scale of metres or centimetres, resulting in size-graded 'beds' analogous to those deposited by sediment gravity flows. The evidence suggests that convection currents within the magma chamber permit successive episodes of deposition from density flows. Inclined bedding analogous to that produced by traction currents in running water is also known (*see* **Unidirectional Aqueous Flow**).

Sedimentary Rock Anomalies

Sedimentary rocks are defined as those that are composed of fragments of pre-existing rocks, of a wide range of particle sizes, that are deposited under ice or water or in the atmosphere. They are thus commonly stratified and contain fossils, and usually exhibit porosity between the constituent sediment grains. This definition is adequate for gravels, sands, and muds and for the conglomerates, sandstones, and shales that are their lithified equivalents. There is,

however, a large group of sedimentary rocks where these characters are largely absent. These are the chemical sediments. The chemical sediments are a diverse group that includes limestones, dolomites, ironstones, evaporites, chert, hydrocarbons (peat, lignite, and coal), and phosphate (*see* **Sedimentary Rocks:** Chert; Dolomites; Evaporites; Ironstones; Limestones; Phosphates). Limestones are frequently composed of clasts and exhibit sedimentary structures analogous to those of sandstones. The other chemical rocks, however, owe their peculiar properties very largely to post-depositional chemical changes during diagenesis (*see* **Diagenesis, Overview**). In the extreme case of the so-called evaporites, it is now recognized that they should be more properly termed 'replacementites' since their original chemical composition has been completely changed, though primary sedimentary features may be discerned as ghostly relics. Indeed it could be argued that the evaporites/replacementites are really metamorphic rocks.

Stratification is sometimes absent in sedimentary rocks. Massive structureless sediments include 'loess' (wind-blown silt) and grain-flow and mass-flow deposits such as debrites and olistostromes.

Metamorphic Rock Anomalies

The previous paragraph on anomalous sedimentary rocks provides a useful introduction to problems of metamorphic-rock classification. Just as the boundary between chemical sediments and metamorphic rocks is gradational, so too is the boundary between detrital sediments and metamorphic rocks. During burial gravel lithifies to form conglomerate, sand forms sandstone, and mud forms shale. With increasing burial temperature and pressure, porosity is destroyed, and texture and mineralogy are modified. There is therefore a gradation between highly cemented sedimentary rocks and low-grade metamorphic rocks. Similar gradations occur when superficial igneous rocks undergo deep burial, with concomitant mineralogical and textural changes as they metamorphose. Historically nowhere has this problem been more acute than in the controversial theories about the origin of granite. Granites were believed to have formed from the slow cooling of magma that had intruded shallower rocks. It was pointed out, however, that, when considering the vast granite masses at the eroded roots of mountains, there was a space problem. This was resolved when it was realized that wholesale replacement of pre-existing rock had taken place. Intrusive granites were found only at shallow depths, having moved upwards from deep parent batholiths.

Rocks and their Classification: Conclusion

Rocks are of three genetic types: igneous, formed from cooling magma; sedimentary, formed mainly by the breakdown of pre-existing rock; and metamorphic, formed by the action of heat and pressure on pre-existing rock. Igneous and metamorphic rocks are both normally composed of a mosaic of interlocking crystals that formed at high temperatures and pressures; they thus usually lack porosity and permeability. In the petroleum industry igneous and metamorphic rocks are generally grouped together under the term 'basement', indicating that it is pointless to continue drilling deeper into such tight rocks. Igneous rocks are generally massive and structureless, while metamorphic rocks are often banded.

Sediments, on the other hand, are deposited with space between the constituent particles. Thus, sedimentary rocks, by their very nature, tend to be porous and permeable. Sediments are deposited episodically, and thus characteristically exhibit stratification. They are also the only one of the three main classes of rock to contain fossils.

All classifications have their illogicalities and inconsistencies. Thus sedimentary rocks grade with increasing burial into metamorphic rocks, and metamorphic rocks ultimately melt to form magma, which cools to form igneous rocks, which are weathered, eroded, transported, and redeposited as sediments.

See Also

Diagenesis, Overview. **Igneous Processes**. **Igneous Rocks:** Granite. **Metamorphic Rocks:** Classification, Nomenclature and Formation. **Regional Metamorphism**. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Processes:** Erosional Sedimentary Structures; Post-Depositional Sedimentary Structures. **Sedimentary Rocks:** Mineralogy and Classification; Chert; Dolomites; Evaporites; Ironstones; Limestones; Phosphates. **Unidirectional Aqueous Flow**. **Volcanoes**. **Weathering**.

Further Reading

Norton WH (2002) *The Elements of Geology*. Indypu blish.com, Cybercity.

Peters KE (1997) *No Stone Unturned*. San Francisco: WH Freeman.

Rothery D, Turner D, and Wilson RCL (2001) *Geology*. Milton Keynes: Open University Worldwide.

RUSSIA

A S Yakubchuk, The Natural History Museum, London, UK
A M Nikishin, Lomonosov Moscow State University, Moscow, Russia

Introduction

The Russian Federation is the world's largest country, spread across 10 time zones from west to east and from the Arctic to Subtropical climates from north to south. Her territory is mostly plains in the lowlands of Eastern Europe and western Siberia and highlands in eastern Siberia. The country's mountain ranges are the Urals, Verkhoyansk, Koryak, Sikhote-Alin, Altai, Sayan, and Caucasus.

Geological exploration of this vast area started in the sixteenth century, when Russia began its advance to the east. However, it became systematic only in the eighteenth century when Russia invited European scientists and many Russian people received education in Europe, many in mining centres. This stimulated development of base metal and silver deposits in the Urals, Altai, and Transbaikalia provinces. Gold and platinum were discovered only in the late eighteenth to early nineteenth century in the Urals. This boosted exploration, and many prospectors moved to Siberia and the Russian Far East, where they found many rich gold placers before the Golden Rushes of the nineteenth century in the West.

These discoveries required systematisation, and the Russian government established the Geological Committee in 1882. After the 1917 Revolutions, these works expanded across the Soviet Union, and its first geological map at a scale of 1:5 000 000 was compiled by 1937. In the 1940–1970s, regional tectonic maps were compiled and many world-class deposits of Ni, Au, diamonds, and oil were discovered. During this period, Russian geology evolved in isolation, but under the significant influence of the German geological school. As a result, the fixistic approach dominated in geology and many local names were proposed for orogenic events, such as Baikalides, Salairides, etc. As a result, Russia has its own system of Proterozoic subdivisions, such as Riphean (1600–680 Ma), that incorporates the Middle and Upper Proterozoic in the International Scale, and Vendian (650–540 Ma), roughly corresponding to the Eocambrian. Phanerozoic subdivisions generally correspond to the International Scale.

After the introduction of plate tectonics, Russian geologists remained mostly non-mobilistic in their thinking. Its regional geology was revised in line with plate tectonic theory only in the 1980s, which was also a period of active study of the Russian shelves and participation of many geologists in international projects. During these years, Russian geological knowledge advanced significantly, with superdeep drilling in the Kola Peninsula, Urals, Central Asia, and western Siberia.

After the breakup of the Soviet Union in 1991, there has been a significant reduction of geological work. However, interesting compilations were produced, as well as detailed petrological studies. The country is now an area of international exploration activity.

General Geology

Russian territory consists of Precambrian cratons and Neoproterozoic to Cenozoic orogenic collages, overlapped by the Mesozoic-Cenozoic sedimentary basins (Figure 1). The East European and Siberian cratons are the oldest structures in continental Russia. On the Arctic shelf, west of the Gakkel Ridge in the Arctic Ocean, there are the Svalbard and Kara blocks. The Hyperborean Craton is prominent in the Arctic shelf east of the Gakkel Ridge and north of Chukotka. The cratons are bordered by the Neoproterozoic orogens of the Baikalides and Timanides, which probably continue on to the Arctic shelf. Between these cratons is the Palaeozoic orogenic collage of the Altaids, mostly hidden under the West Siberian Mesozoic-Cenozoic sedimentary basin. The Altaids extend to Transbaikalia (Figure 2). To the south of the Siberian Craton is the Mongolide orogenic collage. Between the Siberian and Hyperborean cratons and along the Far East coast are the Mesozoic-Cenozoic Circum-Pacific orogens. The Alpine orogenic belt borders the East European Craton in the south.

East European Craton

The East European Craton consists of shields, basement rises, and sedimentary basins. The Baltic Shield outcrops in the Kola Peninsula and Karelia. In the south, the crystalline basement occurs in the so-called Voronezh Rise. There are several sedimentary basins – Baltic, Moscow, Mezen, Volga-Ural, Pachelma, Ulianovsk-Saratov, Peri-Caspian – and the Peri-Urals foredeep.

The craton formed prior to 1.6 Ga (Figure 3). In the Baltic Shield, there are the Karelia, Central

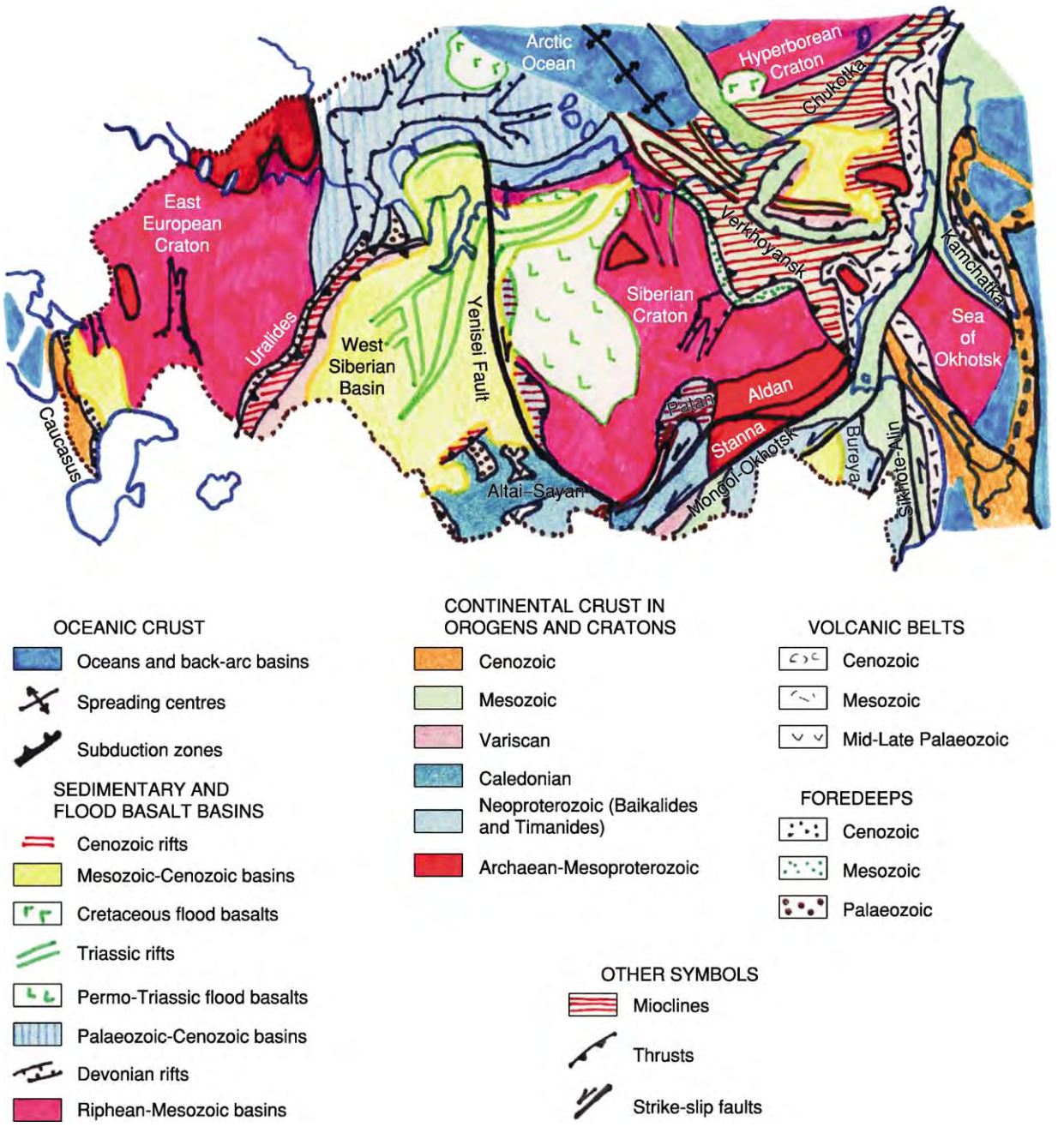


Figure 1 Outline Tectonic map of Russia and adjacent shelves (newly compiled by Nikishin and Yakubchuk). Note the presence of vast Mesozoic Cenozoic sedimentary basins superimposed onto the Palaeozoic Collage between the East European and Siberian cratons. Devonian rifts were superimposed onto cratons. Red cratons: deep blue, Ocean basins: light blue, Palaeozoic tectonic areas: horizontal red shading, Mesozoic to Cenozoic areas: yellow, Mesozoic Cenozoic sedimentary basins.

Kola, and Murmansk Archaean terranes. The Karelia Terrane is a typical 3.0–2.7 Ga granite-greenstone province. The Central Kola Terrane is a classic block with amphibolite metamorphism and it consists of various volcanic and sedimentary rocks with numerous granitoids and gneisses. The Murmansk Terrane consists of Late Archaean granitoid batholiths. Between the Karelia and Kola terranes is the linear White Sea Belt, in which the Archaean crust was reworked in the Early Proterozoic, producing 1.9 Ga granulites. The Archaean nuclei of the Baltic Shield are overlain by Early Proterozoic sedimentary basins and volcanics. Similar terranes exist in the unexposed part of the craton.

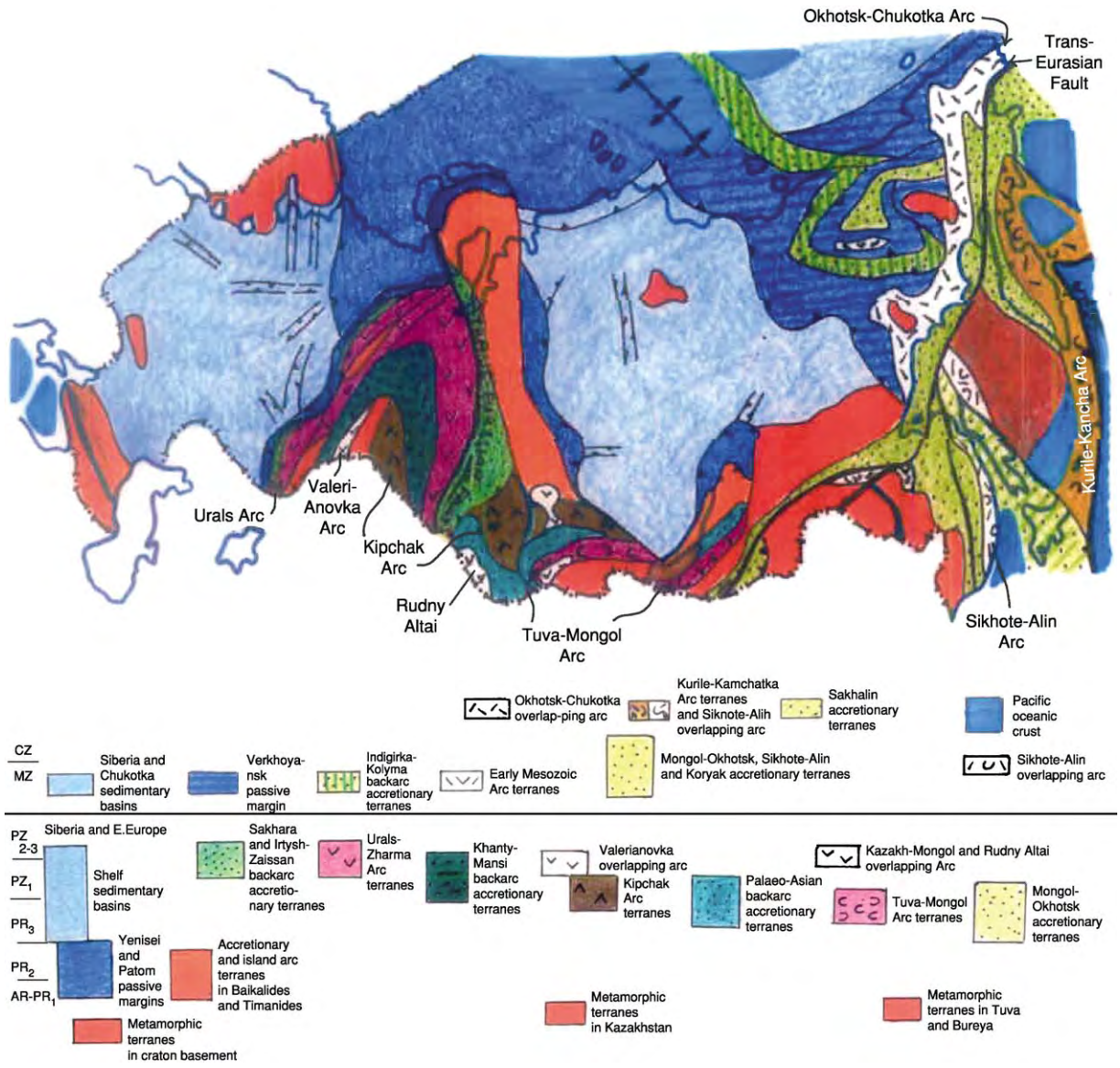


Figure 2 Tectonic map of Russia and adjacent shelves without Mesozoic Cenozoic sedimentary basins (newly compiled by Yakub chuk and Nikishin). Note the oroclinal structure in the Palaeozoic and Mesozoic orogenic collages. Riphean rifts are shown on the major cratons. Late Palaeozoic to Early Mesozoic Trans Eurasian strike slip faults separate the Altaid Collage into two domains.

The sedimentary basins of the East European Craton (Figure 4) consist of several stratigraphic units: (i) Riphean (1.6–0.65 Ga); (ii) Vendian–Early Cambrian; (iii) Middle Cambrian–Early Devonian; (iv) Middle Devonian–Permian; (v) Triassic; (vi) Jurassic–Lower Cretaceous; (vii) Upper Cretaceous–Eocene; (viii) Oligocene–Neogene; and (ix) Quaternary.

Riphean palaeorifts (or aulacogens) are filled with basalts, continental and marine clastic rocks, and carbonates. The Riphean stratigraphy is poorly studied. Three epochs of rifting at 1.6–1.3 Ga, 1.3–1.0 Ga, and 1.0–0.65 Ga have been identified, although some dates are approximate. Palaeorifts mainly occur on the flanks of the craton, but Central Russian palaeorifts crosscut the entire craton.

Vendian-Early Cambrian rocks form the sedimentary cover of the entire craton. The Lower Vendian (~650–600 Ma) consists of tillites, representing the Lapland glaciation. On the south-western edge of the craton are Lower Vendian Volynian plateau basalts (in Ukraine), which were formed prior to the opening of the Tornquist Ocean. Upper Vendian to Early Cambrian rocks are mostly shallow-water sandstone and clay. Upper Vendian rocks have many occurrences of the Ediacaran fauna (*see* **Precambrian:**

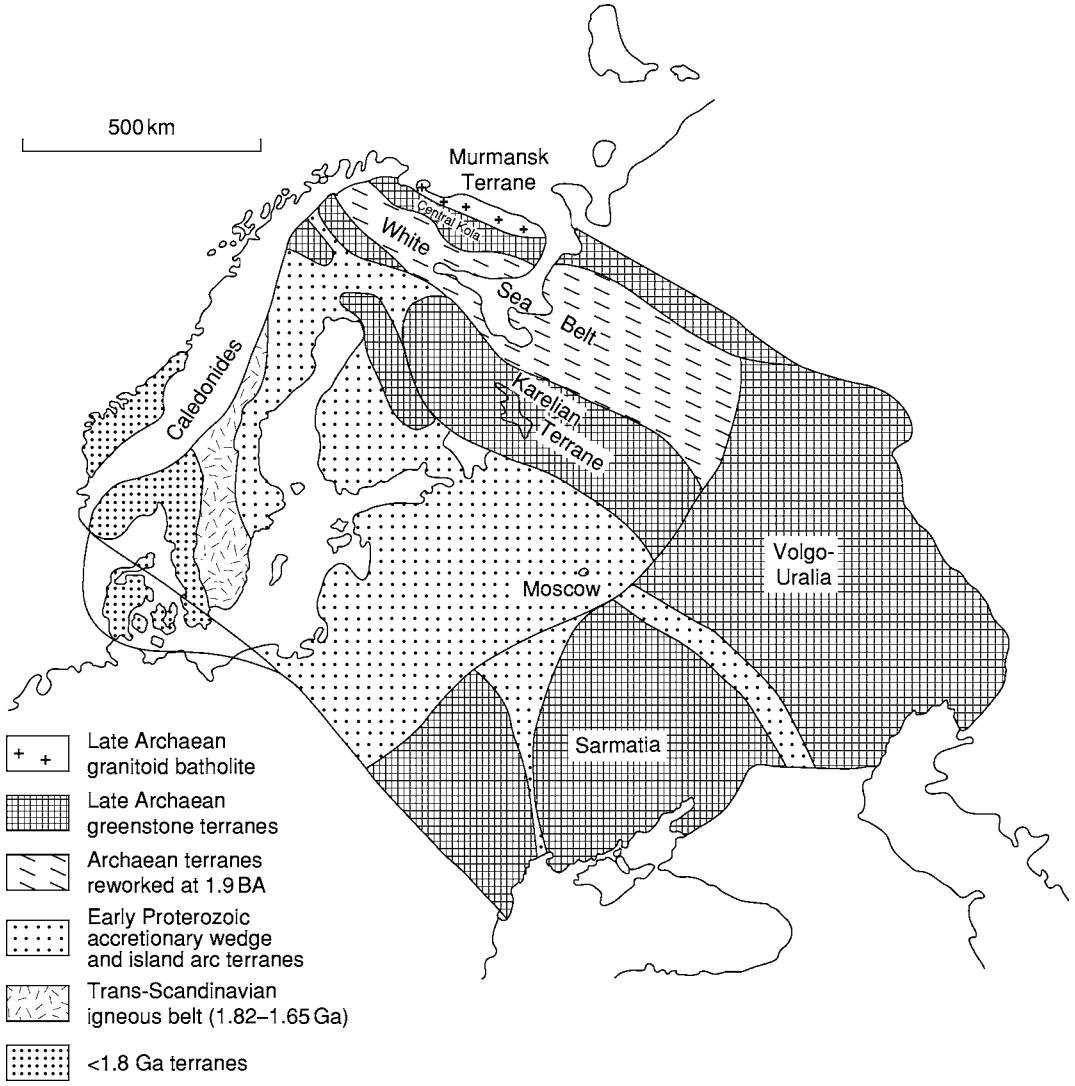


Figure 3 Crustal provinces of the East European Craton (compiled using data by Gorbatschev and Bogdanova).

Vendian and Ediacaran). Accumulation of the sedimentary basins was controlled by the Iapetus and Tornquist Oceans to the west (modern coordinates) and the foreland basin of the Timanides to the east (*see* **Europe:** Timanides of Northern Russia). The craton was elevated at the end of the Early Cambrian synchronously with the final orogenesis in the Timanides.

Middle Cambrian to Early Devonian rocks form a wide basin that might have fully covered even the Baltic Shield. The principal basins are Baltic, Moscow-Mezen-Pechora, Peri-Urals, and Peri-Caspian, but they formed a united system of epicontinental seas. In the Ordovician, to the east of the craton, was the Urals Ocean or oceanic back-arc basin, and the poorly studied Scythian Ocean was to the south. Marine sedimentation was controlled by climate: for the Middle Cambrian to Middle Ordovician strata, clays, sandstones, and carbonates are typical; in the Late Ordovician to Silurian, carbonate rocks are dominant; in Early Devonian times, red clastic rocks began to dominate against a background of arid climate and regression, culminating in total sea regression.

The Middle Devonian to Permian cover was formed in the epicontinental marine basin. The Middle Devonian rocks are typical clastic and carbonate sediments. At the Givetian-Frasnian transition was a magmatic event and rifting, with a system of rifts (Figure 4), such as Pripyat-Dnieper, Donets, Peri-Caspian, Barents Sea, Kola, Timan-Pechora, and Vyatka. During that time, diamondiferous kimberlites formed in the Arkhangelsk region, and the world's largest alkaline ultramafic massifs were emplaced in the Kola Peninsula. Middle to Late

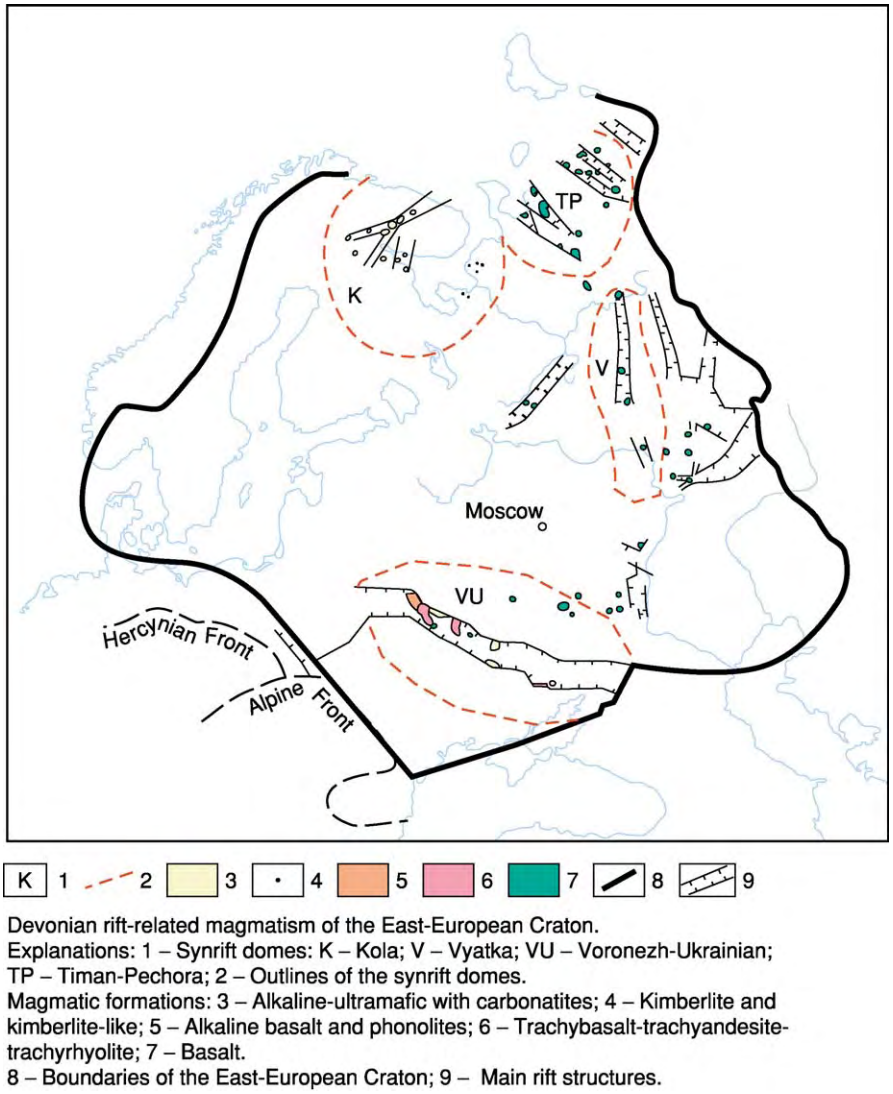


Figure 4 Sedimentary basins of the East European Craton (compiled by Nikishin and Fokin).

Devonian rifting and magmatism can be explained by the influence of mantle plumes, in combination with regional continental extension and possible back-arc processes.

The Middle Devonian to Permian rocks are mostly carbonates. An epicontinental marine basin changed its depth many times due to sea-level fluctuations. Carboniferous strata include typical buried river valleys and limnic coal. The Peri-Caspian Basin was a deep-water basin framed by reefs. In the Donets Basin, there was a large delta system that accumulated 15–20 km of sediments with numerous coal seams. In the Peri-Urals Basin, the Domanik Horizon, enriched in organic material, was formed during the Frasnian, and it is a regional oil-bearing stratum. In Kungurian time, the Peri-Caspian Basin was almost completely separated from the ocean by the Uralide Orogen, and in it was accumulated up to 5 km of evaporites. In the Late Carboniferous, the main Urals collision began, and the Peri-Urals foredeep formed in front of the Uralide Orogen in the Permian. From the mid-Permian, there was a major regression, and mostly shallow-water marine, limnic, and continental sediments were accumulated in the Late Permian. By the end of the Permian, most of the craton had experienced regression, but the giant Pechora-Barents Sea sedimentary basin survived in the north, accumulating 3–5 km of sediments.

At the Permian-Triassic transition, the East European Craton was affected by the Siberian Superplume (Figure 4). The main Triassic sedimentary basins formed on the flanks of the craton are the Peri-Caspian, Pechora-Barents Sea, and Baltic-Polish-German. On the craton, thin continental sediments accumulated in the Moscow-Mezen and Dnieper basins. At the Triassic-Jurassic transition, the

Caucasus-Black Sea-Turkish Orogen formed to the south of the craton; to the north was the Novaya Zemlya orogen. Almost complete regression took place synchronously with these deformations and it continued until Aalenian time.

In Middle Jurassic to Early Cretaceous times, the main sedimentary basins were in the south and north of the craton. In the south were the Peri-Caspian, Ulianovsk-Saratov, Dnieper, and Peri-Black Sea Basins of the Northern Peri-Tethys area. In the north was the Pechora-Barents Sea Basin of the Boreal area. The Boreal and Peri-Tethys basins were periodically linked to form a single basin. The Peri-Tethys area has typical shallow-water marine carbonate and clastic rocks, whereas the Boreal Basin has marine and continental clay and sandy rocks. In the internal part of the craton are shallow-water marine and continental clastic sediments with multiple hiatuses. Organic-rich sediments are typical in the Late Jurassic.

In the mid-Cretaceous, the Peri-Tethyan and Boreal basins were separated. Carbonate rocks dominated in the Late Cretaceous to Eocene in the Peri-Tethys basins, whereas sandy-clay sediments dominated in the Boreal Basin. In the Late Cretaceous-Paleocene there was an intracratonic inversion of palaeorifts.

Oligocene-Neogene shallow-water sediments accumulated in the south of the craton against a background of maximum orogeny in the Caucasus and a colder climate.

In the Quaternary period, there were five major glacial and interglacial epochs. The main glaciers were on the Baltic Shield. To the south, they expanded as far as the Ukrainian Shield, producing moraine and fluvioglacial accumulations.

Siberian Craton

Most of the Siberian Craton is covered by its sedimentary and flood basalt basins. Its basement outcrops are in the Anabar Shield in the north, the Aldan-Stanovoy Shield in the south, and in the Sayan-Yenisei area in the south-west. East of the Anabar Shield is the small Olenek Rise. All the basement rocks are older than 1.6 Ga, but the southern part of the Aldan-Stanovoy Shield was magmatically reworked in the Palaeozoic and Mesozoic. The Sayan-Yenisei area was deformed in the Neoproterozoic.

The Siberian Craton is the largest craton on Earth, with a pre-1.6 Ga crust. Its structure is dominated by Archaean terranes. Some of them were reworked during Early Proterozoic times, and juvenile Early Proterozoic crust occupies a relatively small area (Figure 5). It amalgamated as a single craton by the end of the Early Proterozoic.

There are three principal sedimentary basins (Figure 6): Vilyui in the east, Tunguska in the north-west, and Taseevo-Angara-Lena (or Irkutsk Amphitheatre) in the south-west. The thickness of the sedimentary cover is up to 5–8 km and it consists of several rock complexes: (i) Riphean (1.6–0.65 Ga); (ii) Vendian–Early Devonian; (iii) Middle Devonian–Early Carboniferous; (iv) Carboniferous–Permian; (v) Triassic; (vi) Jurassic–Cretaceous; (vii) Tertiary; and (viii) Quaternary.

Riphean palaeorifts (aulacogens) are widely distributed and are filled with carbonates, continental clastic rocks, and basalts. Their stratigraphy is poorly studied, but there are at least three epochs of rifting distinguished at 1.6–1.3 Ga, 1.3–1.0 Ga, and 1.0–0.65 Ga. Most of the palaeorifts are perpendicular to the craton margins: their palaeotectonic setting is not clear, but some data indicate that 0.8 Ga rifts in the east of the craton could be related to the opening of the ocean. In addition to the palaeorifts, wider Riphean cover overlaps northern parts of the craton.

Vendian to Early Devonian rocks cover almost the entire craton. The Vendian, known in Siberia as the Yudomian, consists of clastic rocks at its base which were overlain by carbonates. During the Early Cambrian, carbonate sedimentation dominated in the Yudoma-Olenek shelf in the east, and a carbonate-evaporite shallow-water basin existed in the Tunguska-Irkutsk Basin. These two basins were separated by carbonate reefs with archaeocyathid-algae buildups. From the Middle Cambrian to the Late Ordovician, carbonates dominated in all basins. Silurian rocks are carbonates, clastic sediments, and evaporites. The maximum regression occurred in the Late Silurian-Early Devonian and coincides with orogeny in Transbaikalia, where the Patom Highlands were thrust onto the Siberian Craton, causing thin-skinned deformations of its cover north of Lake Baikal.

In the Middle Devonian to Early Carboniferous, there was rifting and magmatism. The main rifts formed in the east (Figure 6). Vilyui was the largest system, consisting of the Ygyata, Kempendyai, and other rifts. Rifting started with flood basalt magmatism and the intrusion of basalt dyke swarms. Later, there was the main phase that continued until Tournaisian times. In the Early Carboniferous, the Verkhoyansk branch of the rift system was transformed into the passive margin of the oceanic back-arc basin. As a result, the continental rifts became extinct, but the passive continental margin continued to evolve until Jurassic times. Middle-Late Devonian rifts possibly formed in the Tunguska Basin and

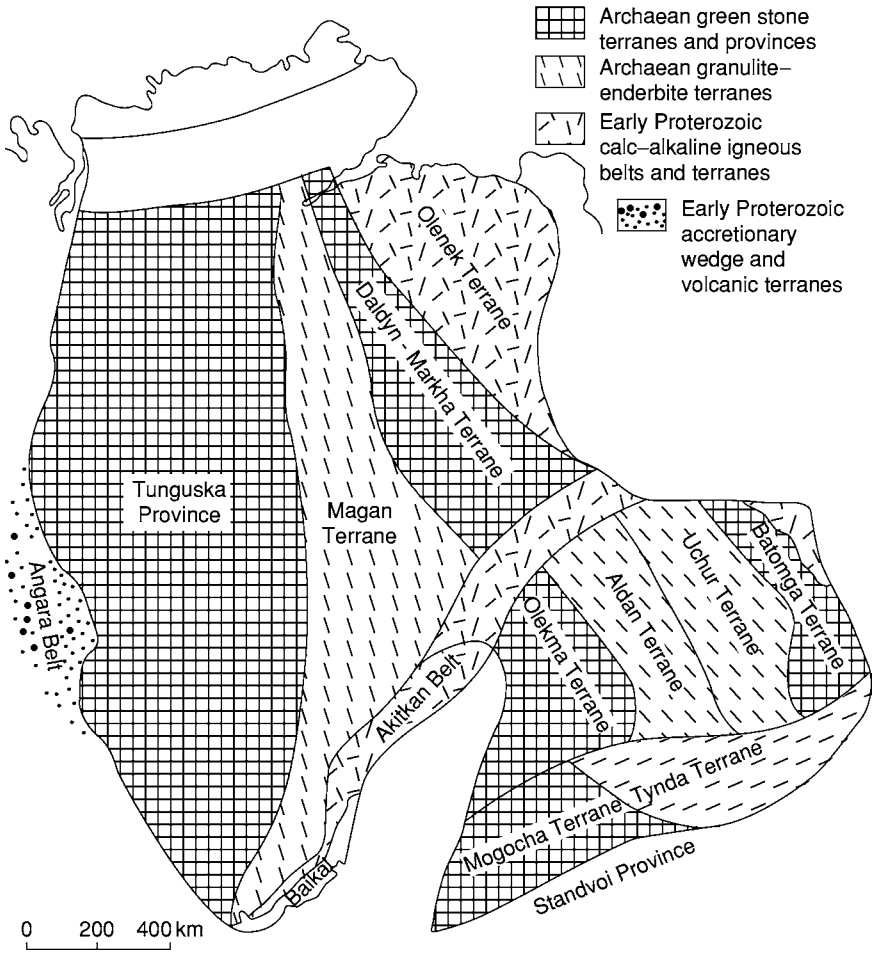


Figure 5 Crustal provinces and terranes of the Siberian Craton (simplified and modified after Rozen).

along the northern flank of the craton. In addition, numerous diamondiferous kimberlites were emplaced into Archaean nuclei, reflecting the impact of Middle-Late Devonian plumes, in a similar way to the East European Craton.

Carboniferous and Permian rocks occur in the Vilyui and Tunguska basins. In the Vilyui Basin, Middle Devonian rocks form thick grey clastic shallow-water and limnic-alluvial formations. This setting persisted until the end of Jurassic time, when numerous coal beds accumulated in the Vilyui Basin. At the end of the Jurassic and in the Early Cretaceous, orogenesis occurred in the Verkhoyansk Orogen, and a foredeep basin formed in its front. In the Carboniferous to Permian, the Tunguska Basin was filled with clastic continental and shallow-water marine sediments with numerous coal beds.

The latest Permian and Triassic was dominated by mantle plume tectonics. Large-scale flood basalt magmatism (often termed the Siberian Traps) took place in the north of the Siberian Craton and southern Taimyr (a probable northern continuation of the craton). The total thickness of basalts in the Tunguska Basin reached 2–4 km. This was the largest intracontinental Phanerozoic magmatic event (Figure 6), and it affected an area extending from the northern Urals to Verkhoyansk and from Taimyr to Central Asia, revealing it as a superplume event.

Jurassic rocks occur in the Vilyui and Irkutsk Basins. They also form a chain of foredeep basins between the Aldan and Stanovoy blocks. Typical rocks are continental coal-bearing sediments. The Stanovoy Block was an Andean-type margin in Jurassic times. Cretaceous clastic rocks are known in the Verkhoyansk foredeep.

During Tertiary times, the Siberian Craton was relatively elevated, possibly in response to the India-Asia Himalayan collision. The altitude of its plateaux was 1–1.5 km. Tertiary sediments occur locally and consist mostly of continental clastic rocks. The Popigai meteorite crater of Eocene age (80 km in diameter) occurs north of the Anabar Shield.

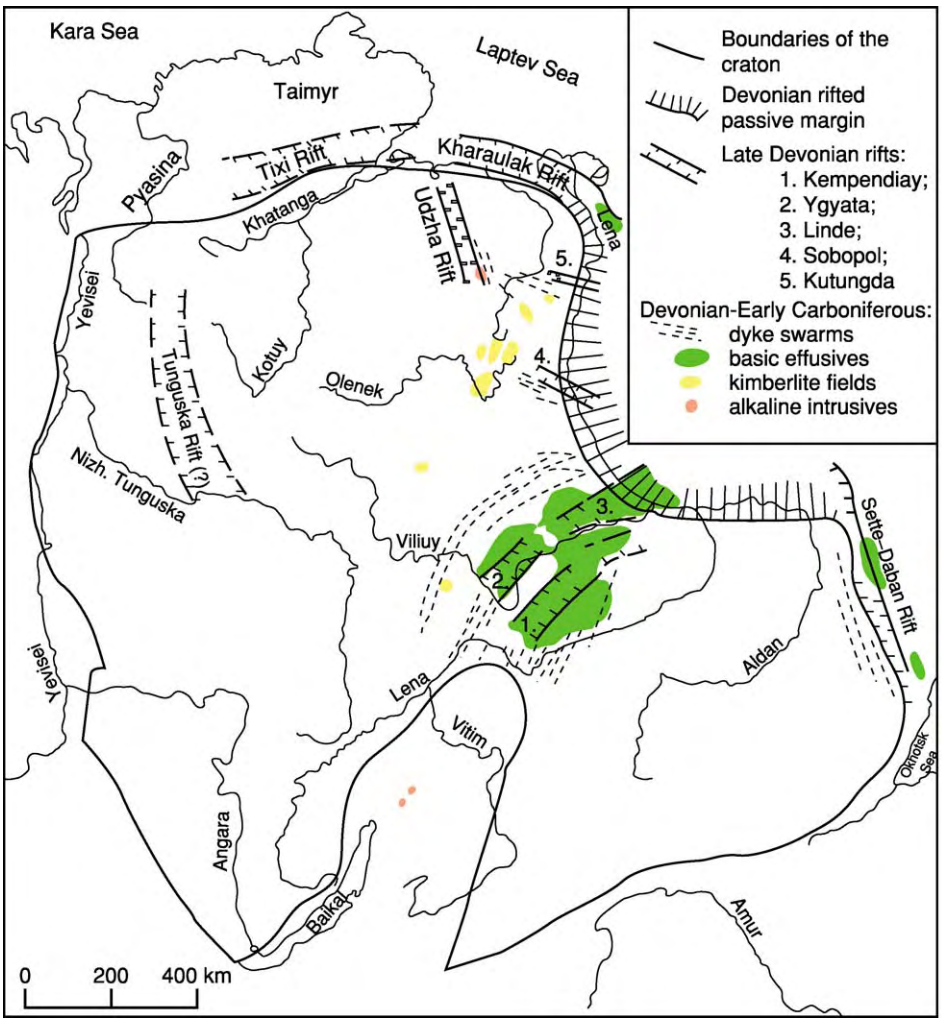


Figure 6 Sedimentary basins of the Siberian Craton (compiled by Nikishin and Fokin).

In Quaternary times, the craton was glaciated many times, but thick glaciers formed only in its northern part and the rest of the area was affected by thick permafrost that survives until the present day.

Neoproterozoic Orogens

Shatsky proposed the term Baikalides for the Neoproterozoic orogens on the flanks of the major cratons of Russia. Their type locality is in the Patom Highlands in northern Transbaikalia, but they were also distinguished in the Yenisei ridge, northern Taimyr, and Timan. Their slivers are recognized inside the Palaeozoic to Cenozoic orogens.

Until the 1980s, the interpretation of Baikalide age was mostly based on stratigraphic data. Recently, the Neoproterozoic age of the type locality in northern Transbaikalia was questioned. Some researchers suggested that the Baikalides are Early or Middle Proterozoic structures, whereas others proposed a Caledonian age. However, the deformed sequences in the Patom Highlands are unconformably overlapped by less deformed Vendian to Cambrian sedimentary rocks, indicating Neoproterozoic deformations. In addition, isotopic dating of syn- and postmetamorphic granitoids in the Yenisei ridge, Taimyr, and Timan also demonstrate the validity of Neoproterozoic deformations.

Patom Highlands This is an oroclinal embayment in the south of the Siberian Craton, extending from Lake Baikal towards the Patom Highlands (Figure 1). In the 1970–1980s, the Baikalides were interpreted as an Early Proterozoic orogen. Subsequent regional mapping re-confirmed that the orogen is Neoproterozoic. Most of this area is occupied by the Barguzin granitoid batholith. In the 1960s, it was considered as Proterozoic. In the 1970–1980s, the first K-Ar data showing its Early Palaeozoic age were obtained. However, application of more advanced techniques

in the 1990s revealed a Carboniferous age. In addition, isotopic dating of some regionally metamorphosed rocks in the Patom Highlands also revealed a Carboniferous age. This raised speculations that the Baikalides might be a Caledonian Orogen.

Geologically, the Patom Highlands consists of two major parts. In the north is a Baikal-Patom passive margin of Middle to Upper Riphean carbonate-terrigenous rocks deformed into linear folds. These rocks are unconformably overlain by Vendian-Cambrian molasse. Together they were thrust onto the Siberian Craton in the Middle Devonian. To the south is the Barguzin-Vitim Terrane which is separated from the passive margin by a suture with Riphean ophiolites. The Barguzin-Vitim Terrane is exposed in small fragments, consisting of metamorphosed Riphean accretionary wedge and island arc rocks.

Yenisei Ridge The Yenisei Orogen in the west of the Siberian Craton (Figure 1) is considered as an equivalent and continuation of the Baikalides. Deformed passive margin rocks occupy the bulk of the ridge. It has been demonstrated that in the west there is a suture with Late Riphean ophiolites and an island arc terrane. The folded structure is intruded by Neoproterozoic granites. Passive margin rocks were thrust onto the Siberian Craton, possibly in mid-Palaeozoic times.

Northern Taimyr, Arctic Shelf and Timanides The east–west trending Taimyr Orogen is an imbricated structure produced during Mesozoic thrusting to the south (Figure 7). It has been suggested that this orogen was a result of collision between the Kara Plate to the north and the Siberian Craton to the south. Northernmost Taimyr is a deformed Neoproterozoic to Early Palaeozoic passive margin of the Kara Plate, consisting of flysch and carbonate sequences which were regionally metamorphosed. They were earlier considered as Precambrian rocks: however, recent study has showed that metamorphism is Late Palaeozoic in age and might have taken place during the emplacement of 300–264 Ma collisional and post-collisional granitoids. The Main Taimyr Thrust separates northern and central Taimyr. The latter hosts two terranes (Mamont-Shrenk and Faddey) with Neoproterozoic island arc rocks and ophiolites (740–820 Ma) which were obducted 650 Ma ago onto Neoproterozoic carbonates, which are also thrust on top of Vendian to Early Carboniferous terrigenous-carbonate rocks of central Taimyr, part of the Siberian Craton. Another ophiolite belt inside northern Taimyr might be a root zone where the actual suture might be located. If so, northern Taimyr should be separated into two terranes, the northern one representing the Kara Plate and the southern one a deformed passive margin of the Siberian Craton.

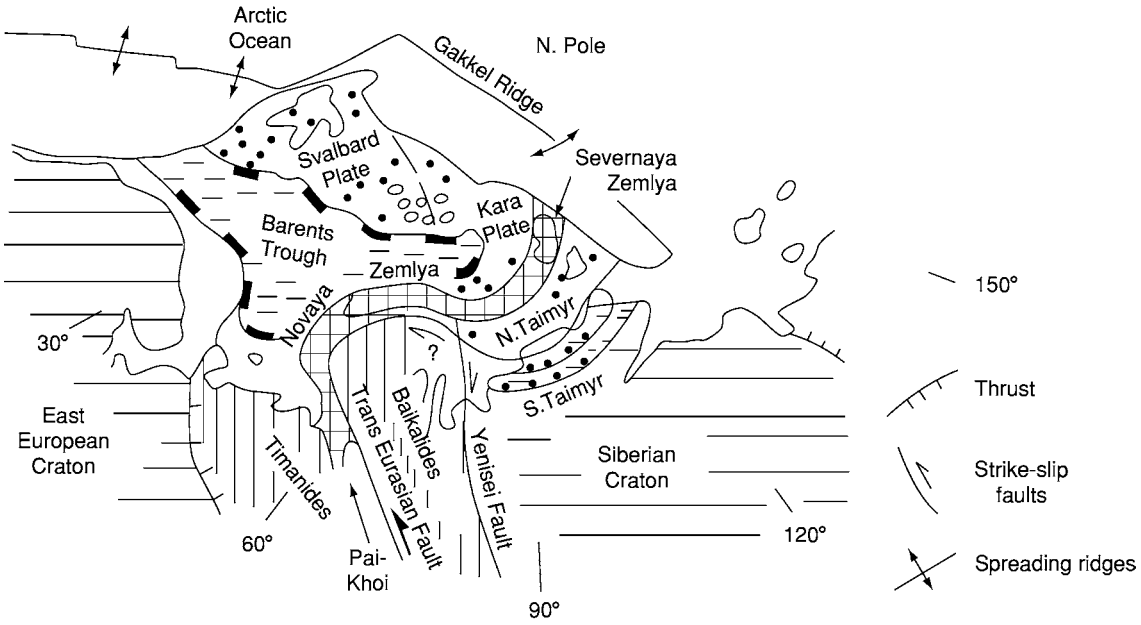


Figure 7 Tectonics of the Arctic shelf and adjacent regions (compiled by Yakubchuk, using data by Bogatskiy *et al.* and Bogdanov *et al.*). The Yenisei Fault separates the Arctic shelf into the Kara and Svalbard plates. Severnaya Zemlya is probably a continuation of Novaya Zemlya. The Barents extensional trough bounds the Svalbard Plate in the south. A major strike slip fault might separate the west east trending structures of the Arctic shelf and Siberian Craton.

Central Taimyr was thrust southward to form the southern Taimyr thrust zone which consists of an imbricated Middle Carboniferous to Triassic sequence. Its stratigraphy is generally similar to the Siberian Craton in the Tunguska Basin.

The position of Taimyr relative to other structures has been controversial. It has been proposed that it might represent a continuation of Novaya Zemlya. However in the 1980–1990s, the Arctic shelf was actively explored for oil. New data were obtained, and it was suggested that the northern Taimyr is an exposed part of the Kara Plate, which occupies the adjacent Arctic shelf. To the east, the Taimyr structures are truncated by Arctic oceanic floor. Its western continuation is far from clear, but magnetic data suggest correlation of Pai-Khoi and Novaya Zemlya with Severnaya Zemlya and not with Taimyr (Figure 7).

The Severnaya Zemlya islands consist of 6–7 km of Ordovician to Devonian deformed shallow-water sedimentary rocks with evaporites and felsic alkaline volcanics (456 Ma). They conformably overlie Cambrian rocks and their fossils are similar to those of the Novaya Zemlya archipelago. Folded Ordovician to Devonian rocks are unconformably overlain by Carboniferous-Permian redbed sequences. The timing of deformation is assumed to be Late Triassic, as in Taimyr, but no structural relationship has been recorded.

Novaya Zemlya and Pai-Khoi consist of Riphean to Lower Cambrian metamorphic rocks unconformably overlain by deformed Ordovician-Devonian terrigenous-carbonate rocks. These are also unconformably overlain by Carboniferous-Triassic continental rocks deformed prior to Jurassic times.

Pai-Khoi, Novaya Zemlya, and Severnaya Zemlya reveal similar lithologies and similar unconformities. The two regions are dextrally offset relative to each other along the Yenisei strike-slip fault that separates the Arctic shelf into the Kara and Svalbard plates and bounds the western margin of the Siberian Craton. Tectonically, Severnaya Zemlya and Novaya Zemlya might represent an inverted rift structure.

West of Pai-Khoi and Novaya Zemlya are the Timanides, where Puchkov reported relic Late Vendian fold belts (600–540 Ma). The Timanides are poorly exposed, and an understanding of their structure is based on geophysical and drilling data. There are two major terranes: in the south-west is the Izhma-Pechora passive margin with Riphean shales, felsic volcanics, and minor ultramafic intrusions; in the north-east is the Bolshaya Zemlya Terrane with Riphean intermediate volcanics and granitoids. The passive margin of the terrane can be traced along the northern coast of the Kola Peninsula, whereas the island arc terrane is truncated by the East Barents Trough and Pai-Khoi structures.

The Timanides are overlain by Devonian to Jurassic rocks of the Pechora sedimentary basin. The whole orogen and its sedimentary basin were thrust onto the East European Craton prior to Jurassic times. Numerous listric compressional faults in the Pechora basin reflect the Urals collision.

In the middle of the Arctic shelf is the 1500 km long East Barents depression, extending parallel to the proposed Pai-Khoi–Novaya Zemlya–Severnaya Zemlya inverted rift. Aplonov and his colleagues proposed that this depression might have a Palaeozoic oceanic crust over which lies up to 15–17 km of Middle Palaeozoic to Cretaceous sediments. In the south, the East Barents depression continues as continental rifts in the base of the Pechora Basin. In the north-east, it is truncated by the Arctic oceanic floor.

Altaid Collage

The area between the East European and Siberian cratons and south of the Arctic shelf is occupied by the Altaid Orogenic Collage (Figure 2). The Russian portion of the Altaid Collage is concealed under the Mesozoic-Cenozoic sedimentary basin of western Siberia, being exposed in isolated orogens in the Urals, Altai-Sayan, and Transbaikalia. Other exposures are in adjacent Central Asia (*see* **Asia:** Central; **China and Mongolia**), Mongolia, and China.

The Altaid Collage is separated into two domains along the giant Trans-Eurasian strike-slip fault, which strikes from Novaya Zemlya towards western Mongolia and then to the Mongol-Okhotsk suture, south of the Siberian Craton. To the west is the Kazakhstan-Khingan Domain and to the north-east is the Altai-Mongol Domain. Both domains have similar Terranes which were sinistrally offset during the Late Palaeozoic to Triassic by 1000 km. The internal structure of the domains is oroclinal, and their axes are also sinistrally offset by 1000 km.

Altai-Mongol Domain

The Altai-Mongol Domain consists of the 'Caledonides' and 'Variscides' (these terms are used here in different ways from their West European counterparts). The Caledonides, which occur in Altai-Sayan and Transbaikalia, represent a portion of the Altaid Collage attached to the Siberian Craton and Neoproterozoic orogens (Figure 8). In the south-west, it is truncated by the Trans-Eurasian Fault. The 'Variscides' occur in the Tom-Kolyvan Terrane in Siberia and the Khangai-Khentei Superterrane in Transbaikalia.

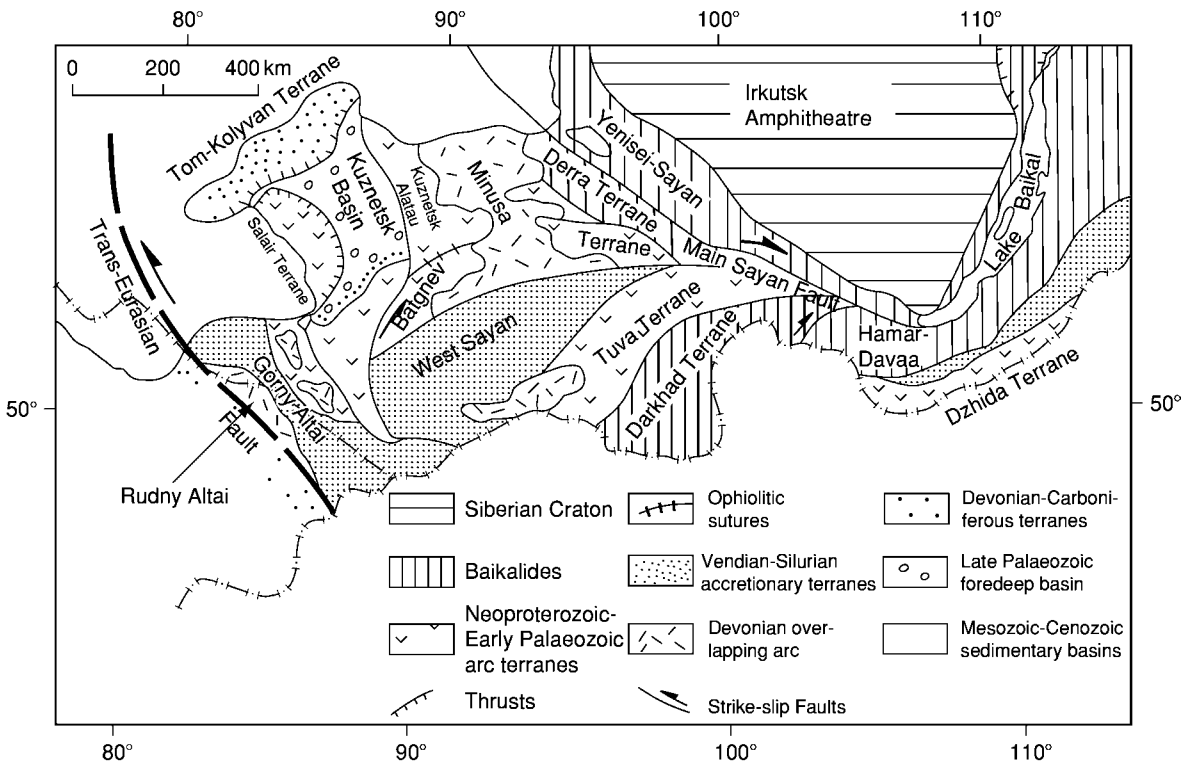


Figure 8 Tectonics of the Altai Sayan area (compiled by Yakubchuk). The mosaic pattern reflects significant strike slip faulting in the area. The Main Sayan fault is a southern continuation of the Yenisei fault in the north that separates the Baikalides and Siberian Craton (**Figure 7**). Offset equivalents of the West Sayan and Tuva terranes are proposed in southern Transbaikalia.

The 'Caledonides' incorporate Neoproterozoic metamorphic terranes and Vendian to Early Palaeozoic magmatic arc and turbidite terranes. In Russia, there are terranes of two Neoproterozoic to Early Palaeozoic arcs – Kipchak and Tuva-Mongol, overlapped by the Middle Palaeozoic Kazakh-Mongol and Rudny Altai and Late Palaeozoic to Early Mesozoic Orkhon-Selenga-Khanka arcs.

The Kipchak Arc includes the Salair and Batenev immature arc terranes separated by sutures with Vendian-Early Cambrian ophiolites from the adjacent Neoproterozoic terranes of Kuznetsk Alatau and Derba. These structures are stitched by Cambrian to Silurian granitoids, but they are offset by the Kuznetsk-Teletskoye sinistral strike-slip fault for as much as 200 km. In the north, the Variscan Tom-Kolyvan Terrane is thrust southward onto the 'Caledonides'. It consists of flysch that might represent an accretionary wedge of an arc, presently hidden under the West Siberian sedimentary basin.

To the south and west, the 'Caledonian' island arc terranes are bound by the accretionary and forearc complexes of Gorny Altai and West Sayan. They consist of Vendian to Silurian flysch with slivers of cherts, basalts, and Vendian–Early Cambrian ophiolites. In the east, they have a T-shaped junction with the Main Sayan Fault, a continuation of the above-mentioned Yenisei Fault that strikes for more than 3000 km. It truncates the accretionary terranes of West Sayan, and their dextrally offset equivalents can be found 600 km to the south-east in Transbaikalia.

South of the accretionary terranes are the 'Caledonian' terranes of Central Mongolia. These are Precambrian metamorphic and Vendian-Cambrian immature arc terranes, known as the Tuva-Mongol Arc. The arc starts near the Stanovoy ridge of the Siberian Craton and continues to the west. It then turns to the south and east, forming the Mongol and then the Khingan oroclines. The 'Caledonian' terranes are stitched by Cambrian-Ordovician and Silurian granites and Devonian calc-alkaline and alkaline volcanicsin Minusa, Rudny Altai, and other locations.

In the core of the Mongol Orocline are Neoproterozoic to Early Mesozoic accretionary wedges of the Khangai-Khentei Superterrane. In Russia the superterrane is exposed in southern Transbaikalia, from where its analogues can be traced to the Circum-Pacific orogens via the Mongol-Okhotsk Suture. In

Mongolia, these terranes are overlain by the Carboniferous to Early Mesozoic Orkhon-Selenga-Khanka Arc. The huge Barguzin Batholith was emplaced behind this arc in Transbaikalia. The northern edge of this arc is present in Transbaikalia and its eastern edge overlaps the Bureya Terrane, but this latter structure belongs to the Kazakhstan-Khingan Domain.

Kazakhstan-Khingan Domain

In Russia, the Kazakhstan-Khingan Domain includes the Urals and buried continuation of the Kazakh uplands. Outside Russia, it is exposed in Kazakhstan, western China, and southern Mongolia, appearing again in the Russian Far East.

The Urals is a 2000 km long isolated Variscan Orogen (Figure 9). In the Urals, there was no 'Caledonian' deformation. Its tectonic cycle started in the Late Cambrian-Ordovician. The Urals is traditionally subdivided into western and eastern slopes. The western slope consists of a deformed passive margin, which records the Urals history from Cambrian to Devonian times.

The eastern slope consists of ophiolite sutures and island arc terranes of Ordovician to Carboniferous age. The Sakmara suture separates the East European Craton and Magnitogorsk and Tagil arc terranes thrust westward onto its deformed passive margin. Due to this position, the island arc sequences represent one of the world's best preserved ancient arcs, with well pronounced internal zonation and famous volcanic-hosted massive sulphide deposits.

To the east is the East Uralian Megazone, considered as a north–south trending Late Palaeozoic 'granite axis of the Urals'. However, its Middle

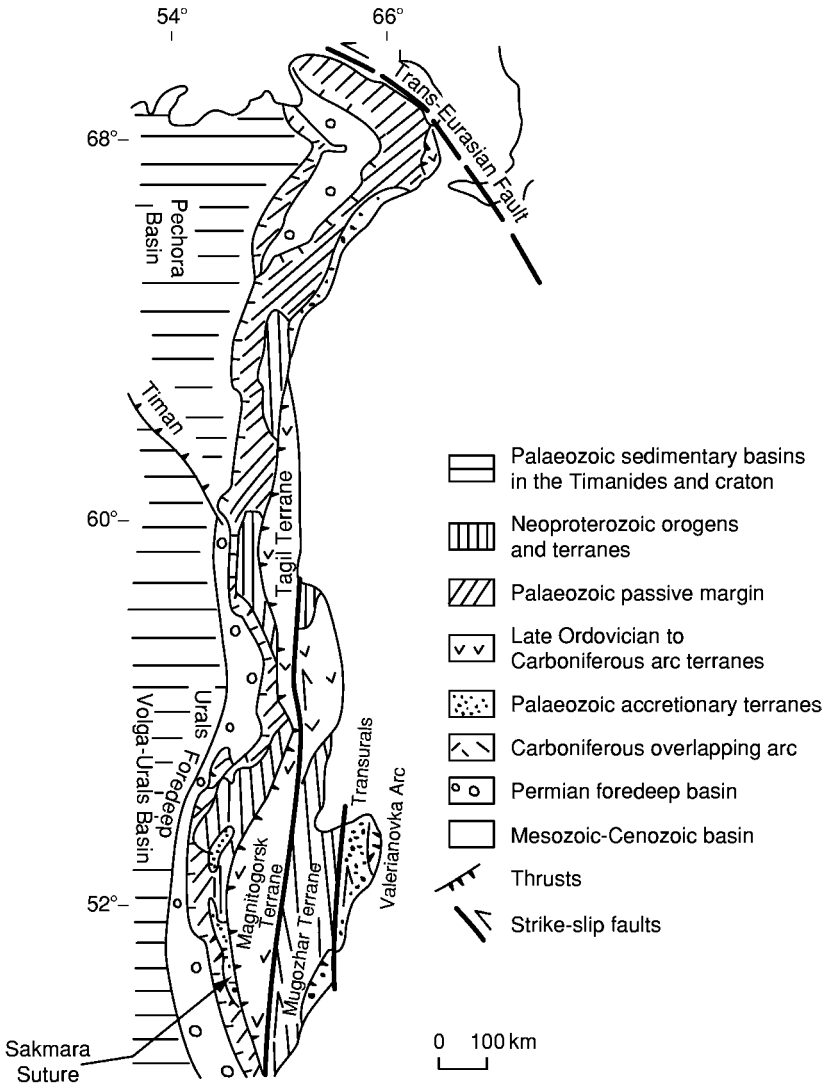


Figure 9 Tectonics of the Urals Orogen (compiled by Yakubchuk, using data by Puchkov and Sobornov). The traditional East Urals Zone consists of several NE trending terranes truncated in the west and east by strike slip faults. The sinistral amplitude of their displacement is at least 300 km.

Palaeozoic pre-granitic terranes strike obliquely relative to its western and eastern fault limits. These are two Precambrain metamorphic terranes and one Palaeozoic island arc terrane inbetween. The latter is principally similar to the Magnitogorsk Terrane, but it is offset sinistrally for 300 km. This suggests that the boundary faults are likely to be strike-slip faults. Seismic studies revealed that they dip to the west, producing an apparent divergent structure of the orogen.

Further to the east is the Trans-Uralian Terrane, consisting of Ordovician to Devonian accretionary wedge rocks. The easternmost structure in the southern Urals is the poorly exposed Carboniferous Valerianovka Volcanic Arc. It has been considered as an accreted arc, but regional airborne magnetic maps suggest that it can be traced to the Kurama Arc of Uzbekistan. This indicates that the Uralides can be interpreted as the result of arc-arc-continent collision.

This collision was accompanied by the formation of flysch in the Famennian to Early Carboniferous, which continued on into the Middle Carboniferous. This collision progressed northward, because in the northern Urals the accumulation of flysch and thrust deformations started in the Early Visean and ended by Kungurian time, when accumulation of salt and molasse took place. The collisional deformations terminated at the end of the Permian. In the Early Triassic, graben structures filled with clastic rocks and basalts formed in the Urals. They were deformed in Early Jurassic times.

The northern continuation of the Urals is a long-lasting debate. Some researchers suggested that Pai-Khoi and Novaya Zemlya represent its 'degraded continuation'. However, airborne magnetic data clearly suggest that magmatic arcs in the Polar Urals turn to the south-east under Mesozoic-Cenozoic sediments and might also be exposed in eastern Kazakhstan, forming a giant orocline. In the core of this orocline are Early and Middle Palaeozoic accretionary, magmatic arc and 1.0 Ga metamorphic terranes, identified on the basis of geophysical and drilling data. They are well exposed in the Kazakh uplands. The bulk of the Palaeozoic orogenic structures in Russia lie under the West Siberian sedimentary basin.

Mesozoic-Cenozoic Sedimentary Basins of Western Siberia

The West Siberian Basin is one of the three world's largest supergiant oil and gas reservoirs. It occurs between the Urals Orogen and the Siberian Craton (Figure 1). Its basement consists of various Neoproterozoic to Palaeozoic terranes that were amalgamated in the Permian. It has steep eastern and western borders and a flat bottom that dips to the north.

At the Permo-Triassic transition, all western Siberia experienced orogenic collapse which was followed by Triassic rifting and voluminous basalt magmatism of the Siberian Superplume. Rifts formed between the Urals and the Siberian Craton and in the western Kara Sea Basin. Urengoi is the largest rift, which was drilled down to a depth of more than 7 km. At the Triassic-Jurassic transition was a weak inversion and uplift, followed by post-rift subsidence for 4–7 km in the Jurassic-Cenozoic.

In Early-Middle Jurassic times, continental sedimentation dominated in the south, whereas there was a shallow-water marine environment in the north. The shoreline frequently migrated north or south. The dominant rocks are clastic alluvial, limnic, and marine sediments. In the Late Jurassic, the proportion of marine sandy and clayey sediments increased. At the Jurassic-Cretaceous transition, West Siberia subsided for 500 m, accompanied by the accumulation of the Bazhenovo oil-bearing facies which are enriched in organic matter. This subsidence was possibly facilitated by synchronous dextral strike-slip faulting along the western flank of the Siberian Craton.

In the Neocomian it became a relatively deep basin, which was quickly filled with sand-clay clinoforms. The main flux of clastic rocks was from the Siberian Craton and the Altai-Sayan Orogen. In the Aptian-Albian, the sedimentation environment became shallow-water. In the Late Cretaceous, Paleocene, and Eocene, the typical rocks are clay and sand with some cherty sediments. At the end of the Eocene, the northern part of the basin was uplifted and its link with the Arctic Seas was terminated. The Oligocene-Neogene in the central and southern parts of the basin is represented by shallow-water marine, limnic, and alluvial facies. Quaternary sediments are widely distributed. These are various sandy-clay facies which accumulated during several episodes of glacial and interglacial events and development of limnic-alluvial systems.

Circum-Pacific Orogenic Collages

The Circum-Pacific orogenic collages started to form in the Palaeozoic and their growth has continued until the present time. They occupy the Russian North-east and Far East. In the Russian North-east is the Verkhoyansk-Chukotka Orogenic Collage, bordered by the Nipponides.

Verkhoyansk-Chukotka Orogenic Collage The Verkhoyansk-Chukotka Orogenic Collage (Figure 10)

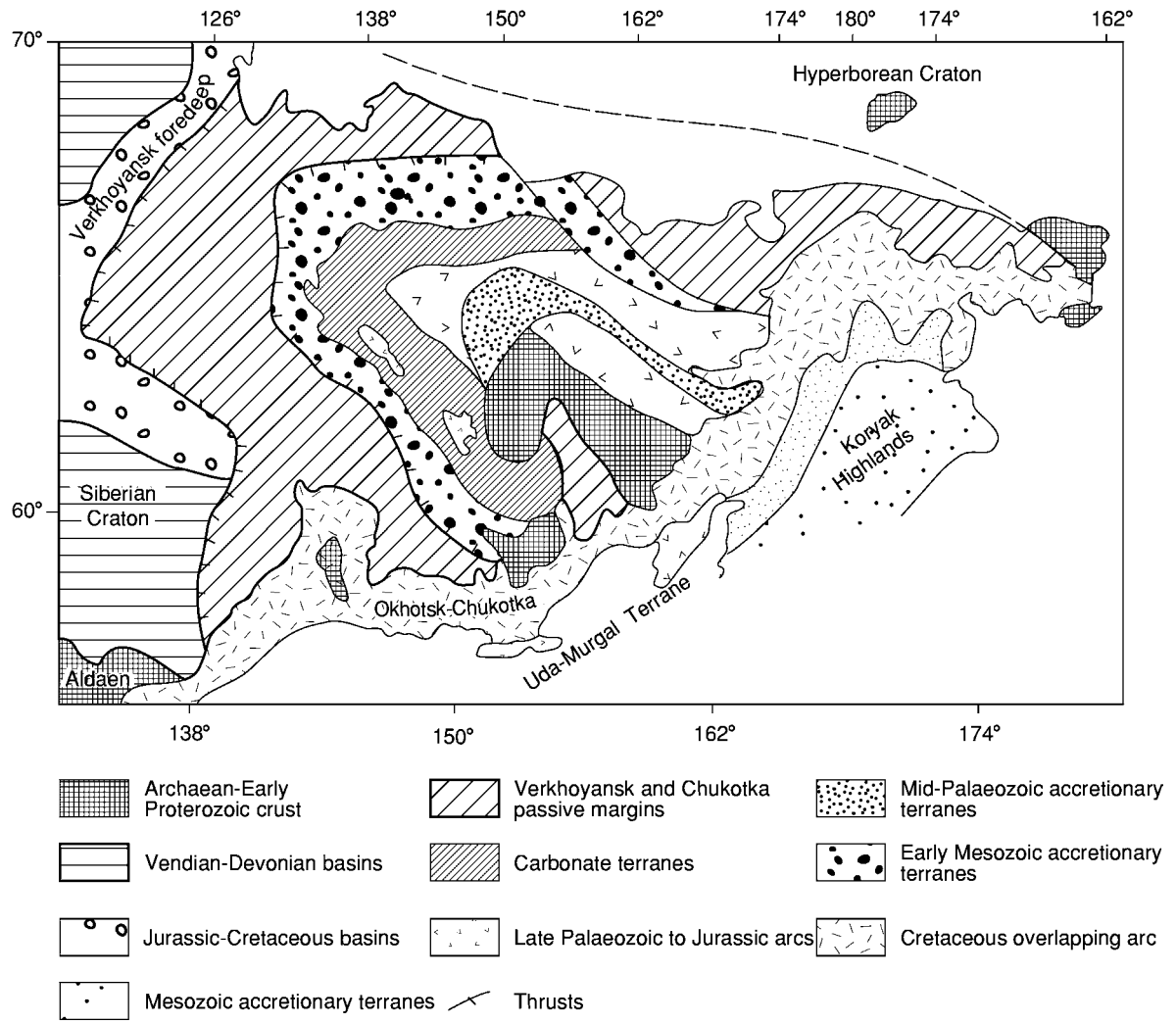


Figure 10 Tectonics of the Verkhoyansk Chukotka Orogenic Collage (compiled by Yakubchuk and Nikishin). Note oroclinal structure of the collage. Mid Cretaceous Okhotsk Chukotka Arc stitched the terranes that constitute the collage. Terranes of the Koryak Highlands represent an accretionary wedge that formed in front of this arc.

evolved from the Early to Middle Palaeozoic until the Cretaceous. In the north, it is bounded by the Hyperborean Craton. To the south-west is the Siberian Craton. The eastern boundary of the collage coincides with the Okhotsk-Chukotka Andean-type arc that stitched various terranes in the Late Cretaceous.

The Hyperborean Craton is a possible fragment of the North American Craton. It was incorporated into the collage after spreading in the Canadian basin in the Late Jurassic to Early Cretaceous. Plate tectonic reconstructions suggest that its arrival in the Early Cretaceous was the major reason for the amalgamation of the Verkhoyansk-Chukotka Collage. This craton occurs almost completely in the shelf zone, and its small fragments are exposed in eastern Chukotka, Novaya Sibir, and the Wrangel Islands. In the south, it is bounded by the deformed passive margin terrane of Chukotka and Arctic Alaska. It consists of Triassic-Jurassic terrigenous rocks deformed in the Early Cretaceous. In the south, the Chukotka Terrane is bounded by the South Anyui Suture, hosting Late Palaeozoic to possible Early Mesozoic ophiolites and accretionary wedges.

The Verkhoyansk passive margin is a mirror image of the Chukotka Terrane, but it includes Ordovician to Middle Jurassic turbidite sequences which are present in the imbricated structure thrust onto the Siberian Craton. These assemblages extend along the north-east margin of the Siberian Craton and can also be traced along the western flanks of smaller cratonal terranes of Okhotsk, Omolon, and Prikolyma, forming the Kolyma Orocline.

The Verkhoyansk passive margin rocks are structurally superimposed by the Jurassic accretionary terranes. Inside them, in the core of the Kolyma Orocline, are the Early Palaeozoic carbonate terranes

and Middle to Late Palaeozoic terranes of the Alazeya Arc which are overlain by the Middle–Late Jurassic to Early Cretaceous Uyandina-Yasachnaya magmatic arc. The opposite ends of this arc collided with each other in the Early Cretaceous to form the Oloy Suture which extends to the east towards the Nipponides.

The 2500 km long Okhotsk-Chukotka continental arc stitched this collage in the mid-Cretaceous. This arc was active from the Albian until the Eocene due to subduction of the oceanic crust under the Siberian Craton. It reveals classic transverse zonation and Andean-type retreat of its magmatism towards the continent in time.

Nipponide Collage The term Nipponides was proposed to describe the structures whose tectonic style is similar to the Japanese islands. In Russia, the Nipponide Collage occupies Sikhote-Alin, Sakhalin, Sea of Okhotsk, Koryak Highlands, Kamchatka, and the Kurile Islands, where it is growing due to subduction of the Pacific oceanic plate (Figure 11).

In Sikhote-Alin, these are Middle Palaeozoic to Early Cretaceous accretionary terranes and forearc

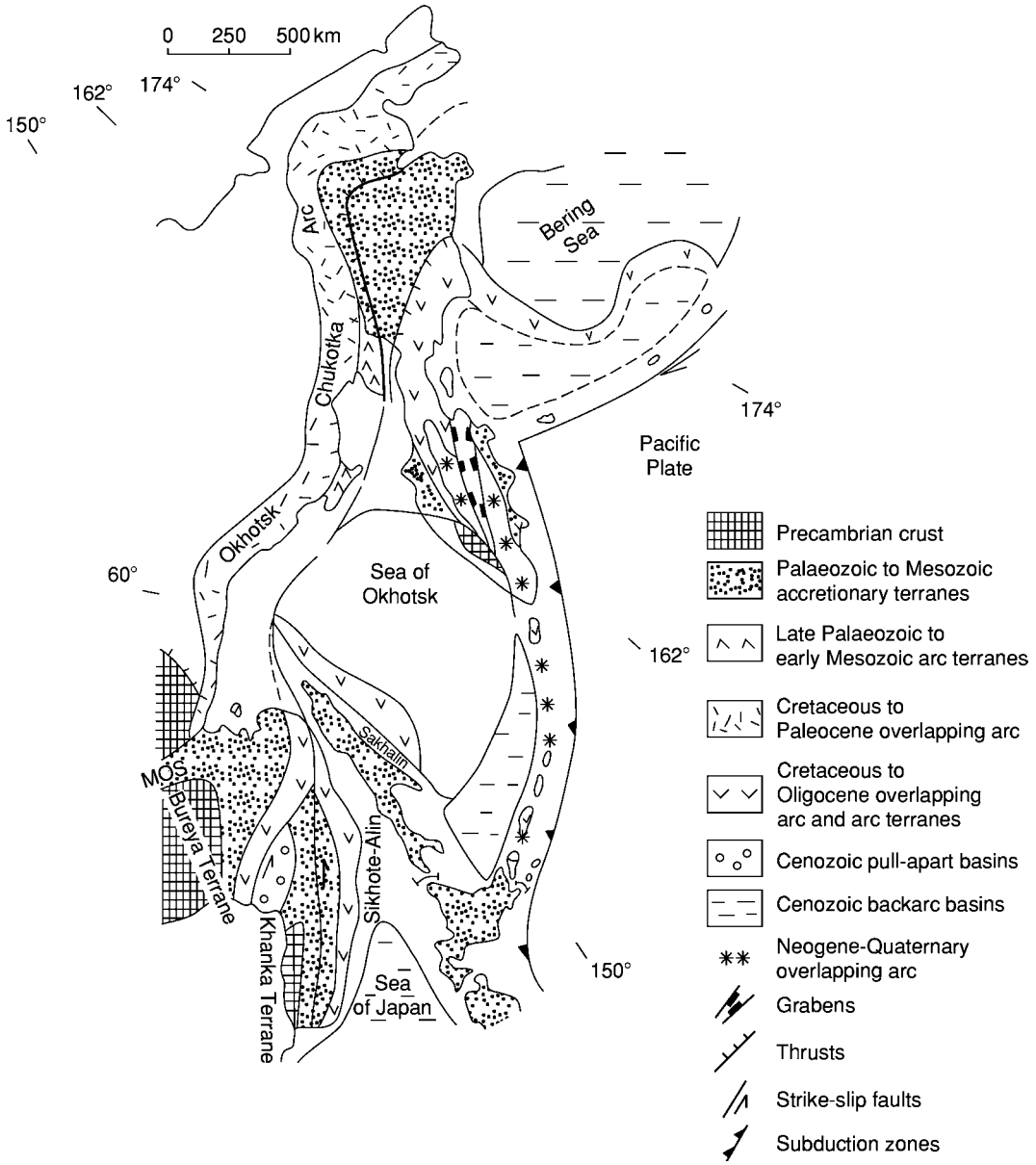


Figure 11 Tectonics of the Nipponides (compiled by Yakubchuk). Oroclinal bending and strike slip duplication is typical for the Nipponides. Behind the active subduction zone of the Pacific plate, new oceanic crust formed in the Sea of Japan and southern Sea of Okhotsk basins. The Central Kamchatka Graben represents back arc rifting.

turbidites. Similar accretionary terranes occur in front of the Okhotsk-Chukotka Arc in the Koryak Highlands. In Sikhote-Alin, they are overlain by another Late Cretaceous to Oligocene magmatic arc that extends towards the Sea of Okhotsk. Magnetic data and outcrops on the islands indicate that this arc turns to the south, so that the Mesozoic accretionary terranes of Sakhalin are squeezed between its two branches. The arc then turns to the north-east and continues to Kamchatka, where it forms several island arc terranes, continuing into the Aleutian island arc.

Inside this orocline is the Sea of Okhotsk. In the 1980s, a Precambrian cratonal terrane was proposed as its basement, but more recent work suggests that this is an accreted Cretaceous oceanic plateau. Whichever is correct, this structure is a major undeformed block inside the Nipponides.

Along the eastern flank of the Nipponides is the Kurile-Kamchatka chain of active volcanoes that have been superimposed on the older arc since the Neogene. In front of them are the actively growing accretionary wedges in the deep subduction-related trenches. Behind this arc are active back-arc basins with newly formed oceanic crust in the south of the Okhotsk Sea and Sea of Japan.

Alpine Belt

The Alpine Orogenic Belt intersects Russian territory south of the East European Craton, in the Scythian Platform located near the Sea of Azov and the northern Caucasus (Figure 1). Its sedimentary cover is 3–5 km thick. Its basement is intersected in numerous drillholes, but because of poor dating there are many problems in the interpretation of this region.

In its basement are the Karpinsky Swell and Scythian Orogen. The Karpinsky Swell is a buried eastern continuation of the Donets Basin that accumulated up to 15 km of Carboniferous shales. Devonian rifting was followed by formation of a Carboniferous back-arc deepwater basin was proposed to explain the Karpinsky Swell. This basin was deformed and thrust northward onto the Peri-Caspian Basin in the Early Permian. The structure of the Palaeozoic Scythian Orogen is poorly known, but in the Palaeozoic it was a continuation of the Greater Caucasus Orogen. The basement of the Scythian Orogen consists of Devonian and Carboniferous schists, Precambrian terranes, Palaeozoic ophiolites, and island arc terranes. Some interpretations suggest that this area was a Cordilleran-type orogen framed by the Palaeotethys Ocean to the south at the end of the Palaeozoic. In the Carboniferous to Early Permian, along the southern flank of the orogen in the Greater Caucasus, an Andean-type magmatic arc accompanied by molasse basins was formed. The other fragments of the orogen are in the Dzirula Massif in Georgia, the Eastern Pontides in Turkey, and Dobrudja in Romania. It was proposed to call it the Euxinian Orogen. To the west, this orogen continued into the Variscides of western Europe.

In Mesozoic to Cenozoic times, the Caucasus-Black Sea region had a complex history of a supra-subduction continental marginal system. In the Triassic, a back-arc basin represented by shales in Georgia and Dobrudja opened. In the Late Triassic, a volcanic arc formed in the north of the Greater Caucasus. At the Triassic-Jurassic transition, the entire north Caucasus region was deformed and its rifts were inverted. In the Early Jurassic, a Greater Caucasus deep back-arc trough was formed on the rifted and thinned continental crust. This trough was inverted in pre-Callovian times. In the Callovian to Late Jurassic, a new deep back-arc rift on the continental crust was formed along the present southern slope of the Greater Caucasus. This basin survived as a deep-water structure until the end of the Eocene, but it also experienced several phases of extension and compression. In the Middle Cretaceous, continental back-arc rifting occurred in the Black Sea and transformed into spreading in the West and East Black Sea Basins in the Late Cretaceous. At the Eocene to Oligocene transition, the main collision in the Greater Caucasus took place. Since that time Cis-Caucasus foredeeps – Kuban in the west and Terek-Caspian in the east – began to form. The collisional shortening in the Greater Caucasus is estimated to be at least 200–300 km, and the orogen was uplifted by 5 km. This elevation was a result of Arabian Plate collision, and it was accompanied by intrusive and explosive volcanism in such volcanoes as Elbrus. In contrast, the molasse basins subsided by 5–7 km.

Modern Plate Tectonics

Russian territory occupies two major modern plates. These are the Eurasian and North American Plates, and their boundary goes from the Arctic Ocean towards Sakhalin Island (Figure 12). In the ocean it is a spreading ridge, whereas in the continent it is a system of diffused continental rifts. It behaves as a transform (strike-slip) fault boundary near Sakhalin island. In the east, the North American Plate is bounded by a subduction zone of the Pacific Plate that produced the active volcanoes of the Kurile-Kamchatka Arc.

Some researchers subdivide the North American Plate into smaller plates, suggesting identification of

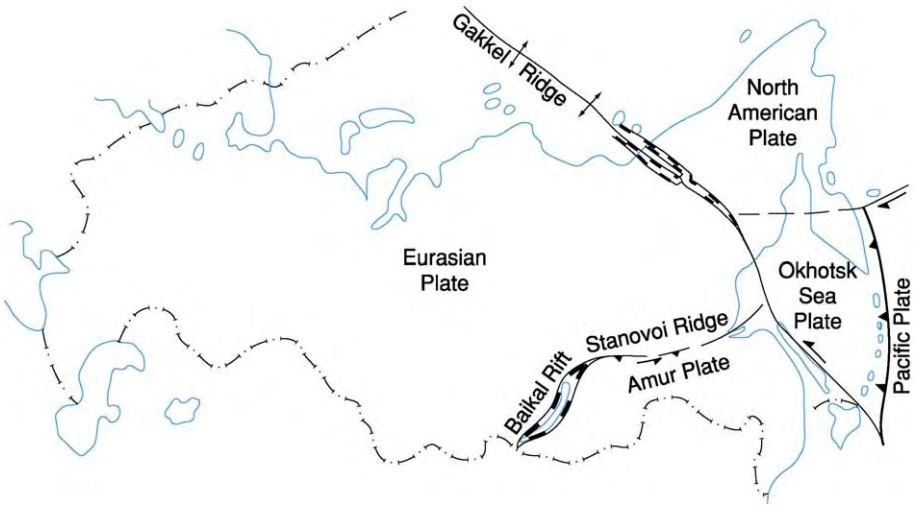


Figure 12 Modern plates in the territory of Russia (compiled by Yakubchuk and Nikishin). Parts of the two major plates of Eurasia and North America can be separated as smaller plates of Okhotsk and Amur.

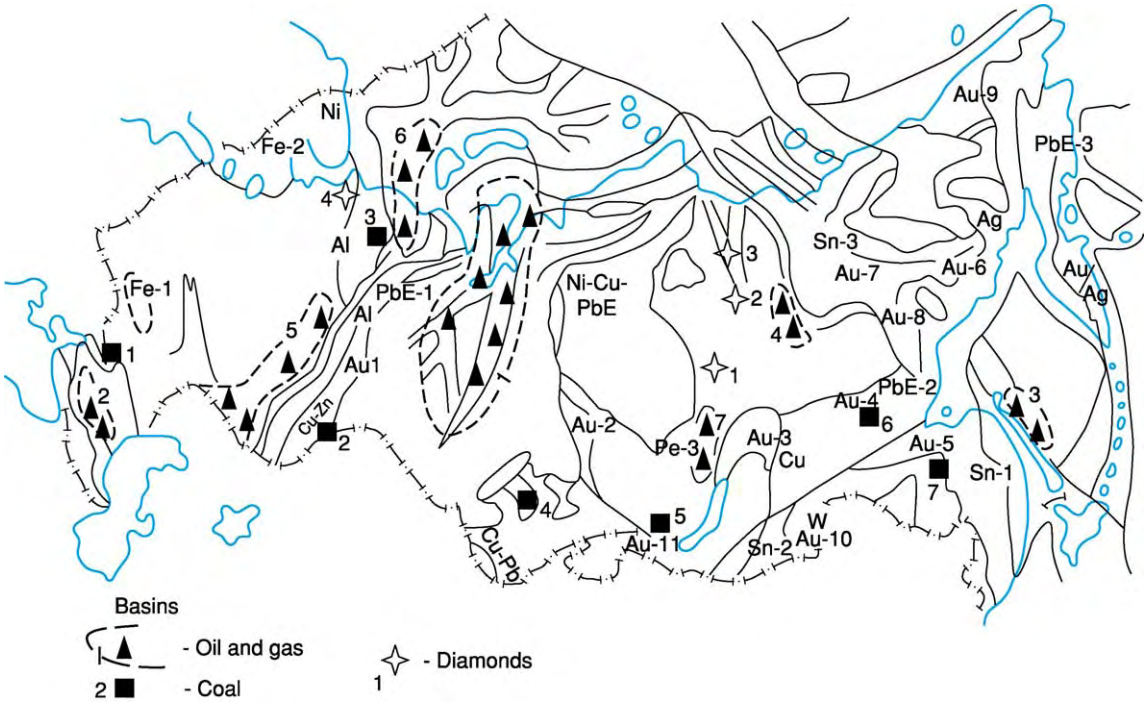


Figure 13 Mineral resources of Russia (compiled by Yakubchuk). Commodities and deposits: U Transbaikalian province, Fe iron ore deposits and their numbers in the text, Ni Pechenga Ni deposit, Ni Cu PGE Norilsk Talnakh deposits, Cu Udokan deposit, Cu Zn Urals VMS deposits, Cu Pb Rudny Altai VMS deposits, Au gold deposits and their numbers in the text, PGE platinoid deposits and their numbers, Al bauxite deposits, Sn tin deposits and their numbers in the text, Au/Ag emerging epithermal gold province in Kamchatka. Solid lines correspond to the tectonic units as shown on **Figure 1**.

a separate Okhotsk Sea Plate. Following the active seismic belt extending along Lake Baikal and the Stanovoy Ridge, it was proposed to recognize the area to the south as the Amurian Plate that includes most of (China *see* **China and Mongolia**). The Eurasian-Amurian plate boundary in this case is a zone of active continental rifting produced in response to India-Asia collision.

Mineral Resources

Russia has a large variety of mineral resources, many world class. Numbers on Figure 13 will be referred to as a basis for the following brief description.

Energy Resources

Russia is the world's second largest producer of oil and the world's largest producer of gas. Its supergiant oil and gas fields are in the Mesozoic-Cenozoic West Siberian Basin (Oil basin N1 in Figure 13). Smaller Cenozoic provinces are in the Scythian Basin (2) and in Sakhalin (3). Mesozoic oil is known in the Verkhoyansk foredeep (4). Middle Palaeozoic provinces are in the Cis-Urals (5), Pechora (6), and Lena (7) basins. Russia's significant coal resources occur in several basins. Palaeozoic coal is in the Donets (1) and Kuznetsk Basins. Most of the Mesozoic coal is in the Urals (2), Pechora (3), Kuznetsk (4), Bureya (5) and Aldan (6) provinces. The supply for generation of nuclear energy comes from a domestic source in Transbaikalia as well as from other CIS countries.

Mineral Deposits

Russia has enormous mineral wealth due both to the size of its territory and world-class quality of some of its deposits. Russia is ranked fourth or fifth in the world in terms of iron ore reserves. Its most productive deposits are in the Early Proterozoic Kursk Magnetic Anomaly (Fe-1), Palaeozoic intrusions in Karelia (Fe-2), and Triassic plume-related pipes in the Siberian Craton (Fe-3). Russia is the world's largest producer of nickel, whose ores are mined from the Early Proterozoic Pechenga deposit in the Kola Peninsula and lateritic deposits in the Urals, but most of the nickel production (as well as 50% of Russian Cu, 60% of the world's Pd and 20% of the world's Pt) is extracted from the Norilsk-Talnakh supergiant deposit formed during the activity of the Siberian Superplume.

Russia's other producing base metal (Cu-Pb-Zn) deposits are volcanic-hosted massive sulphides in the Middle Palaeozoic arcs in the Urals and Altai. Despite significant aluminium production, Russia has only small- to medium-size bauxite deposits, such as in Timan. Domestic tin production is almost suspended, but in recent years Mesozoic tin was extracted from Sikhote-Alin (Sn-1), Transbaikalia (Sn-2), and Yakutia (Sn-3). Russia is famous worldwide due to its precious metals and diamond production. 25% of the world's raw diamonds are extracted annually from the Palaeozoic and Mesozoic kimberlite pipes in the Siberian Craton, such as Mir (1), Udachnaya (2), Aikhal, and Ebelyakh (3).

In terms of gold resources, Russia is second in the world after South Africa. Russia's first gold mine was established in the Urals at the Berezovskoe deposit (Au-1) in the nineteenth century. Today most of production comes from the Yenisei Ridge (Au-2), Patom Highlands (Au-3), Aldan (Au-4), Far East (Au-5), Kolyma (Au-6), Indigirka (Au-7), Sette-Daban (Au-8), Chukotka (Au-9), Transbaikalia (Au-10), and eastern Sayan (Au-11). Russia's largest silver mine is at Dukat in the north-east. A new epithermal gold–silver province was discovered in Kamchatka. In addition to Norilsk, Russian Pt production comes from placer deposits in the Urals (PGE-1), Kondyor (PGE-2), and the Koryak Highlands (PGE-3).

See Also

Asia: Central. **China and Mongolia**. **Europe:** East European Craton; Timanides of Northern Russia; The Urals. **Precambrian:** Vendian and Ediacaran.

Further Reading

Khain VE (2001) *Tectonics of the Continents and Oceans*. Moscow: Nauchnyi Mir(in Russian).

Milanovskiy EE (1996) *Geology of Russia*. Moscow: Moscow University Press(in Russian).

Nikishin AM, Ziegler PA, Stephenson RA, *et al*. (1996) Late Precambrian to Triassic history of the East European Craton. *Tectonophysics* 268: 23 63.

Nikishin AM, Ziegler PA, Abbott D, Brunet M F, and Cloetingh S (2002) Permo Triassic intraplate magmatism and rifting in Eurasia: implications for mantle plumes and mantle dynamics. *Tectonophysics* 351: 3 39.

Mezhelovsky NV, Morozov AF, Gusev GS, and Popov VS (eds.) (2000) *Geodynamics and Metallogeny: Theory and Implications for Applied Geology*. Moscow: Ministry of Natural Resources of the Russian Federation, Inter Re gional Center for Geological Cartography (GEOKART).

Rozen OM (2003) Siberian craton: tectonic zonation and evolution stages. *Geotektonika* 3: 3 21.

Sengor AMC and Natal'in BA (1996) Palaeotectonics of Asia: fragments of a synthesis. In: Yin A and Harrison MT (eds.) *Tectonic Evolution of Asia*, pp. 486 640. Cambridge: Cambridge University Press.

Yakubchuk A (2002) The Baikalide Altaid, Transbaikal Mongolian and North Pacific orogenic collages: similar ity and diversity of structural patterns and metallogenic zoning. In: Blundell DJ, Neubauer F, and von Quadt A (eds.) *The Timing and Location of Major Ore Deposits in an Evolving Orogen*, pp. 273 298. London: The Geological Society.

Yakubchuk AS and Nikishin AM (2004) Noril'sk Talnakh Cu Ni PGE deposits: a revised tectonic model. *Minera lium Deposita* 39: 125 142.

Zonenshain LP, Kuzmin MI, and Natapov LM (1990) *Geol ogy of the USSR: A Plate Tectonic Synthesis*. Washington, DC: American Geophysical Union.

SATURN

See **SOLAR SYSTEM: Jupiter, Saturn and Their Moons**

SEAMOUNTS

S M White, University of South Carolina, Columbia, SC, USA

Introduction

Maps of the seafloor reveal a surface that is dotted with a vast number of cones and truncated cones, arranged in lines, grouped in clusters, or isolated. These are known as seamounts, the name given to any steep-sloped more-or-less conical feature on the seafloor. The overwhelming majority of seamounts are volcanoes. In fact, seamounts are the most common volcanic landform on Earth, but one of the least studied, owing to their wide dispersal and relative inaccessibility. As a consequence, much of the information presented here is derived from a small number of well-studied seamounts and applied to the general population of seamounts.

Seamounts are officially defined as volcanic cones on the seafloor with at least 1 km of relief. This distinction arose during the days of seafloor mapping with wide-beam echosounders when only these relatively large features could be unambiguously identified. Technological advances in remote sensing have improved our ability to locate and image seamounts of increasingly smaller size. Modern shipboard multibeam echosounders provide high-resolution maps showing the morphology and distribution of small (>50 m high) seamounts. Deep-towed sonars reveal tiny (<20 m high) seamounts along the axes of mid-ocean ridges. Satellite altimetry has enabled complete mapping of the ocean basins, but is capable of imaging only relatively large (>2 km high) seamounts. There is little scientific justification for a 1 km height cut-off for seamounts. All volcanoes start growing from the seafloor, regardless of their current size. For the purpose of this work, seamounts are defined broadly to include submarine volcanic edifices of any size.

Seamounts, ocean islands, and guyots form a natural continuum in the process of submarine volcanic construction from submerged, through emergent, to erosional. This article focuses on the processes that build and erode seamounts and their distribution in the oceans. The quasi-conical shape and volcanic origin of seamounts distinguish them from other relief-forming features on the seafloor, such as abyssal hills and carbonate reefs, which are not dealt with here. One exception is the process of serpentine mud volcanism, which builds seamounts in convergent-margin settings.

Geochemical and Geophysical Characteristics

Nearly all seamounts, aside from some in island arcs, are composed of basalt. Geophysical studies also suggest that seamounts have densities, magnetizations, and seismic velocities that are consistent with porous basalt. Most seamounts consist of a base layer of alkali basalt that grades into voluminous edifice-building tholeiitic basalt, which is capped by alkali basalt. Occasionally differentiation products of alkali basalt (hawaiite, mugearite, benmoreite, and trachyte) are observed in the latter stages of growth. However, not all seamounts exhibit all these growth stages, and small seamounts in particular may cease activity before reaching the main tholeiite stage. Smaller seamounts located near mid-ocean ridges have a composition that strongly resembles that of the lava erupted at the mid-ocean ridge but have a tendency to be both more primitive and more depleted than ridge-axis basalt.

Seismic, geodetic, and gravity studies have identified magma chambers beneath several ocean islands. Geophysical studies of the Kilauea volcano on Hawaii provide perhaps the clearest picture of a magma system anywhere. These studies define a partly molten conduit that rises to about 5 km beneath the summit caldera, with large blade-like dikes running laterally down the rift zones. Large positive gravity anomalies centred beneath many of the islands of French Polynesia suggest large (100 km^3) frozen magma chambers at shallow levels (2–4 km). At Krafla on Iceland, attenuation of seismic shear waves suggests a large

magma body (10–100 km^3) at a depth of 3 km below the surface. Small melt bodies that have been detected seismically in the uppermost mantle near the EPR (East Pacific Rise) may be evidence that even very small near-ridge seamounts have ephemeral magma chambers.

Distribution

Global Distribution and Spatial Arrangement

The earliest studies of seamount distribution and abundance in the ocean basins relied on estimates of height and volume obtained from single-beam echosounder crossings of seamounts. Since single-beam data provide only a cross-sectional view of the topography, with no guarantee of crossing the summit, approximations were required to estimate true seamount height and shape. Nevertheless, these early studies provided valuable indications of the distribution of seamounts in the ocean basins. Modern multi-beam echosounders have revealed the complexity and variety of seamount shapes and distributions, but still only cover regional areas (a few hundreds of square kilometres). The development of satellite altimetry has allowed the complete mapping of the larger seamounts in the ocean basins (Figure 1). Studies of seamounts suggest that satellite altimetry reliably detects seamounts that are at least 2 km high.

The global distribution of seamounts is best approximated by power-law models. Power-law models predict that there are in the order of 10^5 seamounts over 1 km high in the ocean basins and in the order of 10^7 seamounts including seamounts that are much less than 1 km high. Over 50% of all seamounts are found in the Pacific Ocean. Statistical studies of seamount distribution find very clear differences in the distributions of seamounts in different oceans and in different tectonic settings (Table 1).

Seamounts are distributed in several spatial patterns. The most widely known pattern is the seamount chain (Figure 2). In seamount chains, several very closely spaced individual volcanoes occur in a line. The Hawaiian–Emperor Seamounts are a classic example of a seamount chain. In this chain, as in many others, the bases of individual seamounts nearly touch or may completely overlap, but each volcano creates an individual peak. When seamounts are even more closely spaced, so that individual seamounts along the chain are difficult to recognize, the arrangement is called a volcanic or aseismic ridge. Aseismic ridges often occur along the track of a hotspot, for example the Ninety

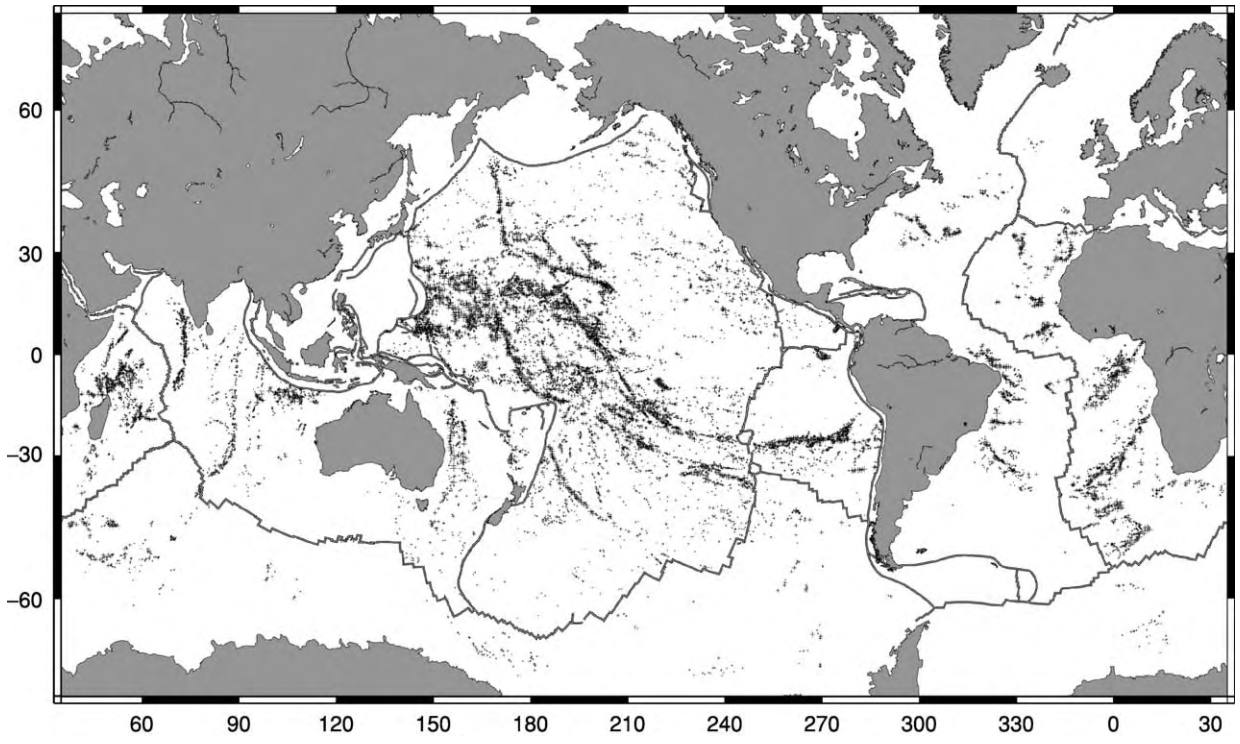


Figure 1 Global distribution of seamounts (indicated by plus symbols scaled according to the size of the seamount edifice) detected by Geosat/ERS 1 altimetry. Approximately 14 675 seamounts ranging in size from approximately 1 km to 7 km were identified. Reproduced by permission of American Geophysical Union from Wessel P (2001) Global distribution of seamounts inferred from gridded Geosat/ERS 1 altimetry. *Journal of Geophysical Research* 106: 19 431 19 441. Copyright 2001 American Geophysical Union.

Table 1 Seamount abundances in intraplate and mid ocean ridge settings

Region	Inclusive area (km^2)	Lithospheric age (Ma)	Observed seamount abundance (10^3 km ^{2})	Minimum height cutoff (m)	Reference
Mid Atlantic Ridge, 24° 30° N	6027	0 0.5	79.8	50	Smith and Cann (1992)
Reykjanes Ridge, 57° 62° N	5075	0 0.6	106	50	Magde and Smith (1995)
Mid Atlantic Ridge, 25° 27° N	73 000	0 28	18.35	70	Jaroslow *et al.* (2000)
Galapagos Ridge, 91° 97.6° W	14 100	0 0.5	18.4	50	Behn *et al.* (2003)
East Pacific Rise, 8° 17° N	200 400	0 2.9	0.9	200	Schierer and Macdonald (1995)
East Pacific Rise, 15.5° 20° S	113 000	0 1.9	9.5	50	White *et al.* (1998)
Easter Seamount Chain	243 000	0 9.8	1.6	200	Rappaport *et al.* (1997)
Pacific plate	Pacific plate	0 169	Total 8882	~1500	Wessel and Lyons (1997)
Global	All ocean basins	0 180	Total 14 675	~1500	Wessel (2001)

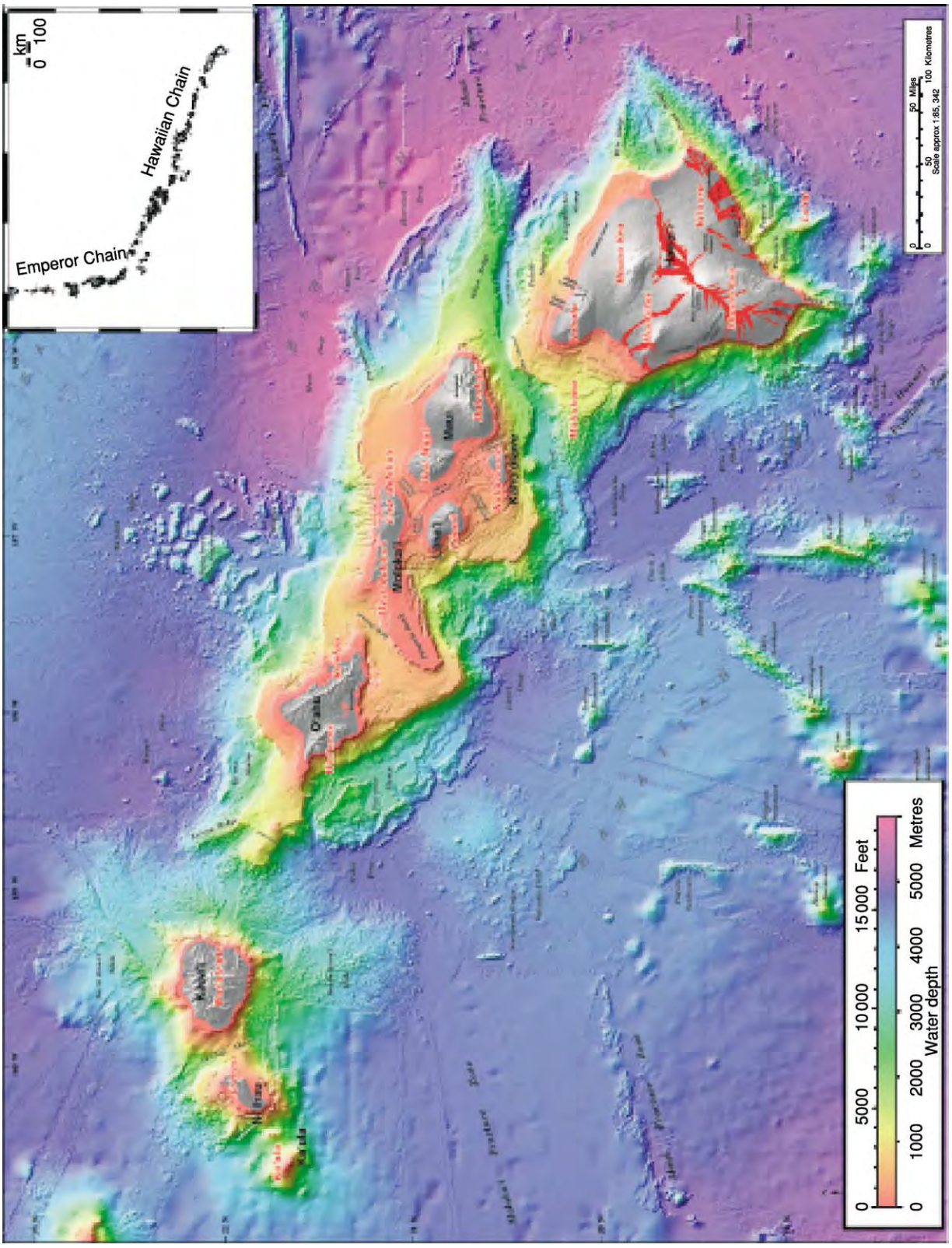
Data based on raw seamount counts and seafloor survey areas selected on the basis of complete seafloor coverage with modern mapping technology in the following references: Behn MD, Sinton JM, and Detrick RS (2003) Effect of the Galapagos Hotspot on seafloor volcanism along the Galapagos Spreading Center (90.9° 97.6° W). *Earth and Planetary Science Letters*, vol 217, pp. 331 347 (2004); Jaroslow GE, Smith DK, and Tucholke BE (2000) Record of seamount production and off axis evolution in the western North Atlantic Ocean, 25°25 27°10 N. *Journal of Geophysical Research* 105: 2737 2760; Magde L and Smith DK (1995) Seamount volcanism at the Reykjanes Ridge: relationship to the Iceland Hotspot. *Journal of Geophysical Research* 100: 8449 8468; Scheirer DS and Macdonald KC (1995) Near axis seamounts on the flanks of the East Pacific Rise, 8° 17° N. *Journal of Geophysical Research* 100: 7871 7885; Smith DK and Cann JR (1992) The role of seamount volcanism in crustal construction at the Mid Atlantic Ridge (24°30 N) *Journal of Geophysical Research* 97: 1645 1658; Rappaport Y, Naar DF, Barton CC, Liu ZJ, and Hey RN (1997) Morphology and distribution of seamounts surrounding Easter Island. *Journal of Geophysical Research* 102: 24 719 24 728; Wessel P (2001) Global distribution of seamounts inferred from gridded Geosat/ERS 1 altimetry. *Journal of Geophysical Research* 106: 19 431 19 441; Wessel P and Lyons S (1997) Distribution of large Pacific seamounts from Geosat/ERS 1: implications for the history of intraplate volcanism. *Journal of Geophysical Research* 102: 22 459 22 475; White SM, Macdonald KC, Scheirer DS, and Cormier M H (1998) Distribution of isolated volcanoes on the flanks of the East Pacific Rise, 15.3° 20° S. *Journal of Geophysical Research* 103: 30 371 30 384.

East Ridge in the Indian Ocean. However, some aseismic ridges seem to emerge from mid-ocean ridges, such as the Sojourn Ridge and the Pukapuka Ridges in the southern Pacific. Seamounts also appear in clusters. Classic examples of the cluster pattern include the Mid-Pacific Mountains and the Hawaiian North Arch volcanic field. Finally, isolated seamounts are fairly common around fast-spreading mid-ocean ridges, where large numbers of very small seamounts (less than 1 km high) occur randomly on the seafloor.

Tectonic Setting: Intraplate Seamounts

The largest seamounts and the longest seamount chains known on Earth are intraplate seamounts. In the Atlantic, several intraplate seamount chains can be traced back to the edge of the continent, indicating 80–120 Ma of continuous activity. Probably the most famous example of an intraplate seamount chain, the Hawaiian–Emperor chain, stretches for 5000 km from its source to 80 Ma seamounts at its end (Figure 2). The oldest parts of this chain may have been subducted, so its true longevity is unknown. The Hawaiian chain also boasts the largest volcano on Earth: Mauna Loa (with an estimated volume of 42 500 km^3). What causes this great volume of volcanic activity is still a subject of debate.

The most widely accepted hypothesis for the origin of intraplate seamounts is that they are formed above hotspots supplied by mantle plumes (*see* **Mantle Plumes and Hot Spots**). The widely accepted model for hotspots states that they are fixed with respect to the Earth's centre of mass and thus leave a trail of consecutively older seamounts as the plate moves over the stationary hotspot. The trend and age progression of the hotspot seamount chain should record the absolute (relative to the Earth's centre of mass) plate-motion vector. The strongest evidence supporting the hotspot model is the linear arrangement and age progression along seamount chains. When a hotspot occurs underneath a mid-ocean ridge it can send out seamount chains or ridges on both sides of the plate boundary. The Carnegie and Cocos ridges from


km
0 100
Hawaiian Chain
Emperor Chain
0 20 Miles
0 50 100 Kilometres
Scale approx 1:85, 342
0 5000 10000 15000 Feet
0 1000 2000 3000 4000 5000 Metres
Water depth

the Galapagos hotspot, and the Walvis Ridge and Rio Grande Rise from the St Helena hotspot are two examples of this dual-ridge phenomenon.

However, many intraplate seamount chains fail to show the simple age progression predicted by the hotspot model. Cocos Island on the Cocos Ridge has lavas that are much younger than predicted by a simple age-progression model consistent with a hotspot origin. Likewise, the Line Islands, the Cameroon Line, and the Pratt–Welker chain show no regular age progression. The lack of age progression in many, even most, seamount chains has resulted in the advancement of several hypotheses to the explain them. Recent palaeomagnetic studies suggest that the hotspot source of the Emperor seamount drifted south at a rate of approximately 30 mm a $^{-1}$ before becoming stationary at 43 Ma. Several other seamount chains in the Pacific exhibit changes in direction near 43 Ma (the Gambier–Tuamotus, Gilbert–Marshall, and Austral chains) but the significance of these kinks in the seamount chains is not yet clear. Lack of age progression and hotspot fixity make it unlikely that all intraplate seamounts result from mantle plumes rising from the deep mantle. Other hypotheses for the origin of seamount chains are extensional stress across zones of pre-existing weakness in the lithosphere, feedback between plate flexure and tectonic stress, and small-scale convective rolls in the upper mantle.

Tectonic Setting: Mid-Ocean Ridges

Seamounts are most abundant near mid-ocean ridges (*see* **Tectonics:** Mid-Ocean Ridges). These seamounts are predominately small; thus, high-resolution bathymetric mapping is necessary to detect them. Most of the studies mapping seamounts near mid-ocean ridges rely on morphometric assumptions that they have an aspect ratio of 2 : 1 or less and that their heights are several times the resolution of the mapping instrument to distinguish them from abyssal hills or other tectonic seafloor topography. The near-ridge population consists mostly of seamounts less than 1 km high, with the overwhelming majority being small isolated seamounts less than 100 m high (Table 1). On slow-spreading ridges, such as the Mid-Atlantic Ridge, seamounts occur directly over the spreading axis within the rift valley of the ridge, while on fast-spreading ridges, such as the East Pacific Rise, there is no rift valley and no seamounts along the spreading axis. The largest on-axis seamount yet discovered (more than 5 km high) is on the ultra-slow spreading Gakkel Ridge, while recent surveys have found tiny (less than 20 m high) pillow-lava mounds as the largest edifices on the axis of the fast-spreading East Pacific Rise. Thus, seamounts over the spreading axis grow larger but become less abundant as the spreading rate decreases. However, small (less than 200 m high) seamounts are ubiquitous away from the axis of fast-spreading ridges and increasingly rare on slower-spreading ridges. Small but ubiquitous seamounts form on lithosphere younger than 0.2 Ma, regardless of spreading rate. This suggests that the thickness of the brittle lithosphere controls the near-ridge seamount population.

Many seamount chains originate at mid-ocean ridges. Near-ridge chains are typically shorter and composed of smaller volcanoes than the intraplate seamount chains. Near-ridge chains become more common as the spreading rate increases, with the largest seamount chains being associated with locations where the ridge axis exhibits signs of a high magma budget. For example, the Axial seamount on the Juan de Fuca ridge is a place where excess magma has produced the Cobb–Eikelberg seamount chain. The highest concentration of seamount chains is adjacent to the most inflated and shallowest part of the super-fast-spreading southern East Pacific Rise (Figure 3). Rarely seamount chains are found near fracture zones, but these are smaller lower-volume seamounts than those found near the middle of ridge segments. Unlike the hotspot chains discussed above, most near-ridge seamount chains are orientated between the direction of absolute plate motion and that of relative plate motion at the ridge. Most of the chains are aligned with the direction of shallow asthenospheric flow beneath the ridge flanks.

Tectonic Setting: Island Arcs

Seamounts occur in island arcs as a product of subduction-zone volcanism. Seamounts are distributed in a linear or curved pattern following the curvature of the trench. Tectonic segmentation of the arc by cross-faults creates discontinuities in the seamount distribution, illustrating the strong tectonic control on the distribution of volcanoes in island-arc settings. Seamounts in island arcs are more diverse in both morphology and composition than seamounts in the other two tectonic settings. The Izu–Ogasawara Arc

Figure 2 Bathymetry surrounding the Hawaiian Islands, a prominent intraplate seamount chain stretching across half of the Pacific plate (see inset), reveals several long rift zones and smaller seamounts. Volcanically and hydrothermically active Loihi, to the south east of Hawaii, is the youngest volcano in the chain. Reproduced with permission from Eakins BW, Robinson JE, Kanamatsu T, *et al.* (2003) *Hawaii's Volcanoes Revealed*. Geological Investigations Series I 2809. US Geological Survey.

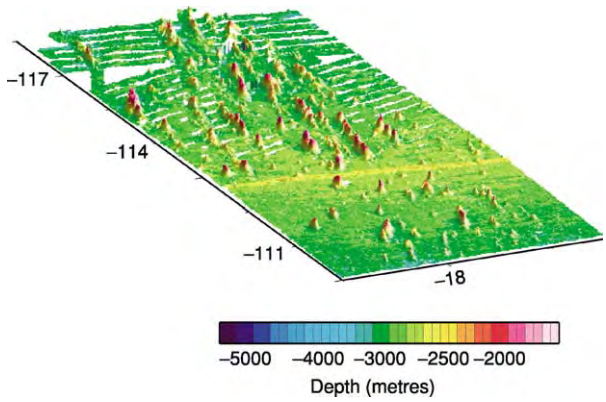


Figure 3 Seamount chains in the Rano Rahi Field adjacent to the southern East Pacific Rise. The shaded relief bathymetry reveals a remarkable abundance of seamounts near the part of the spreading axis with the greatest inferred magma budget in the region.

contains eight submarine calderas that have erupted large (more than 100 km^3) volumes of rhyolite pumice. In contrast, Kick-Em-Jenny, the southernmost island in the Lesser Antilles, is an island-arc seamount that has erupted 10 times since 1939, forming lava domes of basaltic magma within the summit caldera of its very steep-sided edifice. Almost every island arc has a number of seamounts that are smaller versions of the emergent volcanic islands making up the arc.

Mud volcanoes are a unique amagmatic type of seamount found in the fore-arc regions of several subduction zones. The widely accepted explanation for mud volcanism is the upwelling of overpressured serpentinite diapirs from the subduction décollement through normal faults in the fore arc. One of the largest and best-studied mud volcanoes is the Conical Seamount (25 km in diameter and 2 km high), which is located in the fore arc of the Marianas Arc. As more seafloor mapping is done along outer fore arcs, more examples of mud volcanoes are discovered.

Distribution through Time: Cretaceous Seamounts

During the Cretaceous (85–120 Ma) incredibly intense seafloor volcanism occurred throughout what is now the western Pacific (Figure 1). Not only were many of the large oceanic plateaus emplaced in the Pacific during the Cretaceous (Ontong–Java, Manihiki, Shatsky), but also the topography of the seafloor reveals a striking abundance of large seamounts (Mid-Pacific Mountains, Magellan Seamounts, Wake Guyots, etc.). Both the highest density of seamounts per unit area and the highest density of large (more than 3.5 km high) seamounts occur in the Pacific. Satellite altimetry data reveal that the concentration of Cretaceous seamounts is almost double that of Cenozoic seamounts.

Morphology

The vast majority of seamounts are predominately constructed of effusive lava flows. Thus, the overall shapes of seamounts are primarily controlled by the geometry of the magma plumbing system, the eruption rate, the lava type, and the local stresses on the volcano edifice. The cooling effect of water leads to the construction of much steeper edifices than those on land. Most seamounts have steep (10° – 30°) sides but nearly flat summit areas, in contrast to the low slopes of shield volcanoes (5° – 10°). As a seamount grows, it is subject to more complex stresses and tends to develop a more complex shape (Figure 4). In map-view, seamount shape evolves from nearly circular small seamounts to the more complex stellate forms of large seamounts and ocean islands. As a seamount grows larger, mass wasting from gravitational instability or wave erosion becomes more important in modifying its morphology.

Seamount Growth and Development

Deep-water stage Basaltic magma erupted under high hydrostatic pressure produces non-explosive eruptions forming edifices built from lava flows (Figure 4A). Repeated eruptions build extrusive edifices composed of pillow lavas with minor sheet lava flows. Intrusion of magma into the pelagic sediments on the seafloor produces a mixture of lava and wet sediment called peperite. Small seamounts (less than 1 km high and less than 10 km^3 in volume) usually form edifices that can best be described as low lava domes. Many seamounts even of this small size have relatively steep slopes and flat summits, unlike subaerial lava shields. Studies of small isolated seamounts near mid-ocean ridges show that a few seamounts of this size develop complex shapes, including relatively large summit craters. The seamount shape becomes generally more complex as the seamount grows. When seamounts reach a somewhat larger size (*ca*. 1 km high and 10–100 km^3 in volume) they more often have an irregular outline and an 'overturned soup bowl' or truncated-cone shape.

Shoaling stage Large seamounts (more than 2 km high and with volumes of over 200 km^3) may develop a stellate shape as a result of long rift zones concentrating lateral eruptions of lava from a central feeding system (Figure 4B). Not all large seamounts develop rift zones (e.g. Galapagos), but the rift zones on some seamounts become very pronounced ridges extending for more than 100 km (e.g. Hawaii). The presence of a magma chamber above the level of the seafloor and

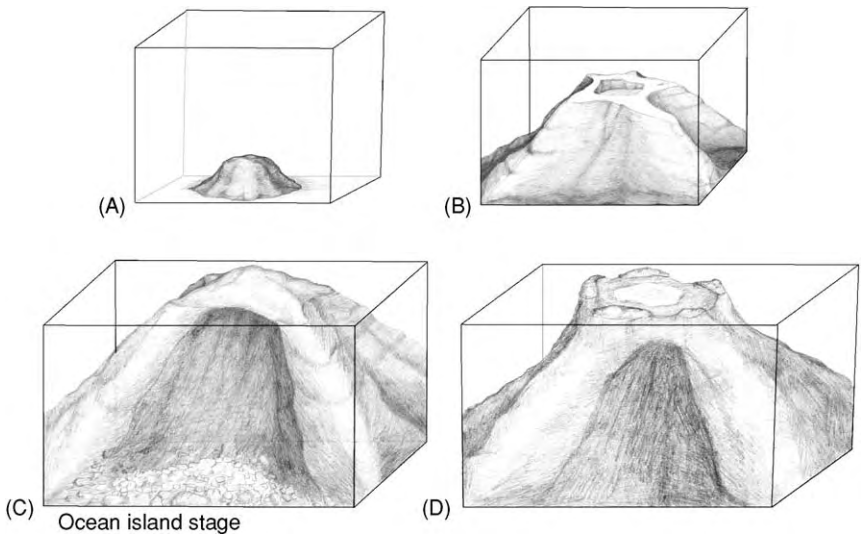


Figure 4 The growth stages of seamounts from (A) early deep water stage, through (B) shoaling stage, showing the development of rift zones, and (C) ocean island stage, including the development of a subaerial shield volcano and mass wasting, to (D) an atoll in the guyot stage.

within the edifice itself aids the development of rift zones on large seamounts by enabling lateral dikes and flank eruptions in focused zones of weakness. Within about 1 km of sea-level, reduced hydrostatic pressure allows exsolution of gas, and explosive eruptions become frequent. Lava flows of scoriaceous pillows and lapilli breccia become important lava types.

Emergent stage At sea-level, eruptions become surtseyan in type, producing fine-grained base surge and air-fall tephra deposits. Besides the famous and well-described eruption of Surtsey off Iceland (1963–1967), the only recorded observations during the emergent stage are from Graham Island (1831), Metis Shoal (1968, 1979), Myujin-sho (1952–1953), and Shin-Iwo-Jima (1986). Phreatomagmatic eruptions form tuff rings and cones consisting of easily eroded deposits of loose lapilli and ash. If eruptions stop at this stage, wave action will quickly erode the seamount. The small islands of Myujin-sho disappeared within months of formation.

Ocean island stage Once the volcanic vent rises above sea-level, phreatomagmatic eruptions become rare and effusive eruptions resume (Figure 4C). Pahoehoe and aa are the important lava types, and in most cases a lava shield is created from numerous low-relief lava flows. Giant landslides may cover the flanks of the island with debris from slumping or debris avalanches. Some of the largest landslides on Earth (involving volumes of 5000 km^3) have occurred around the Hawaiian Islands (Figure 2), and some studies speculate that up to 50% of the original volume of the older Hawaiian Islands has been lost owing to mass wasting.

Guyot stage Thermal subsidence and erosion combine to drown all islands after volcanic activity ceases. Wave erosion planes off a flat summit plateau, but the sides of the volcano remain steep, creating the distinctive guyot cross-section. Atolls (including the Marshall, Tuamotu, Kiribati, and Maldive Islands) are a type of guyot found in tropical oceans, where coral forms a fringing reef around an ocean island. If the upward growth of the coral reef keeps pace with the subsidence of the seamount, a ring of coral islands form, encircling a lagoon, and a thick layer of carbonate sediment develops, which caps the old seamount basement rocks (Figure 4D).

Why Seamounts Have Flat Tops

The characteristic steep-sided flat-topped shape of seamounts is a morphology rarely seen in subaerial volcanoes. The flat tops of guyots can be explained by wave erosion, but many seamounts with flat tops have never been near the wave base. Several models have been advanced to explain why so many seamounts have a summit plateau or flat-topped shape. It is still unclear whether one model can explain all the occurrences of flat-topped volcanoes, or whether each of various models is appropriate under certain conditions.

In the ring-dike model proposed by Batiza and Vanko, eruptions occur from a series of circumferential dikes encircling the summit area rather than from

a central feeder vent. The steep slopes and flat summits of the Galapagos volcanoes, which essentially have the same form as flat-topped seamounts, are attributed to lateral extension by ring-dike intrusion. Circumferential fractures observed in the summit areas of some near-ridge seamounts support the ring-dike model. Ring dikes are thought to form when magma stalls within the crust, creating a small laccolith. Ring dikes shoot upwards from the edge of the laccolith as it inflates. A variation on this model proposes dipping cone sheets instead of more vertical ring dikes as the intrusive body.

In the caldera-fill model proposed by Clague and others, large calderas form early in seamount growth and are filled by later eruptions. A repeated sequence of caldera collapse and re-filling allows the seamount to grow upwards and outwards while maintaining a flat top. This model is based on new high-resolution mapping that suggests that lava shields (erupted from central vents) formed on the summits of flat-topped volcanoes in the north-eastern Pacific.

In the neutral-buoyancy model proposed by Barone and Ryan, low-density basaltic crust counterbalances the buoyancy of the magma so that eruptions will rise only to a limited height. Seamounts will grow only to the height that balances the magma pressure in their magma chambers and will then grow laterally instead of upwards, forming a flat top. Other studies, however, have questioned why this model would lead to circular volcano outlines and what continues to drive eruptions once the critical height is reached.

Broader Effects of Seamounts

Subduction Asperities

Seamounts being subducted erode the fore arc and may also cause erosion at the base of the overriding plate. Huge furrows, mass wasting, and structural disruption of the fore arc are the effects attributed to seamount subduction (Figure 5). There are many areas that exhibit the morphological signs of seamount subduction, but proving the existence of buried seamounts depends on interpreting seismic and magnetic data. However, seamounts being subducted are sometimes directly visible, for example Daiichi-Kashima Guyot in the Japan Trench and the Bannock (Seamount) Structure in the Mediterranean. Seamount subduction may also be responsible for seismic asperities on the subduction thrust-fault interface. Seamount tunnelling, the process by which a seamount going down on the subducting plate scrapes material off the overriding plate, has been inferred to have an important role in crustal development and fluid flow in the margin. The subduction of large seamount chains and aseismic ridges coincides with a reduced dip angle in the subducting plate, for example in the case of the Juan Fernandez chain and the flat slab in northern Chile.

Hydrothermal Circulation

Active seamounts may host high-temperature focused hydrothermal flow. The Loihi seamount in the Hawaiian chain hosts high-temperature hydrothermal vents within its summit caldera, which are very similar to vents found at mid-ocean ridges (Figure 2). Most seamounts have neither the magma supply volume nor the longevity of Loihi, but sulphide hydrothermal deposits have been found on the summits of several other seamounts during submersible dives. These hydrothermal deposits indicate that seamount and mid-ocean ridge hydrothermal systems are broadly similar and contribute the same ions to seawater. Possibly more significant is the role that inactive seamounts play as basement outcrops that provide easy paths for the escape of fluid and heat.

Oceanographic Circulation

The topography created by seamounts has a large effect on the local water masses, primarily generating upwelling and enhancing currents. Anticyclonic circulation over the tops of seamounts creates a so-called 'cold dome' as a result of enhanced upwelling along the sides of the edifice. Also, long aseismic ridges and seamount chains are barriers to deep- and mid-water circulation. Aseismic ridges standing 2–4 km above the seafloor and stretching for hundreds of kilometres across the Pacific deflect north–south currents at depth.

Critical Habitat

Seamounts play a key role in ocean environments by providing habitats for fishes and suspension feeders. The enhanced upwelling combined with the potential for a reef environment in very shallow seamounts make large seamounts sites of high primary productivity. Seamounts have been known as productive fishing grounds for centuries, but their role in oceanic biodiversity has been appreciated only in the last 50 years. Seamounts host a relatively large percentage (estimated at 15–35%) of endemic species and may be important sites of speciation for deep-sea fauna. Seamounts in the south-west Pacific show highly localized species distributions and fairly limited recruitment among seamounts, even between closely spaced seamounts. The formation of atolls by coral growth on subsiding seamounts creates an important habitat niche for filter feeders and reef fishes. Finally, many humans find ocean islands and

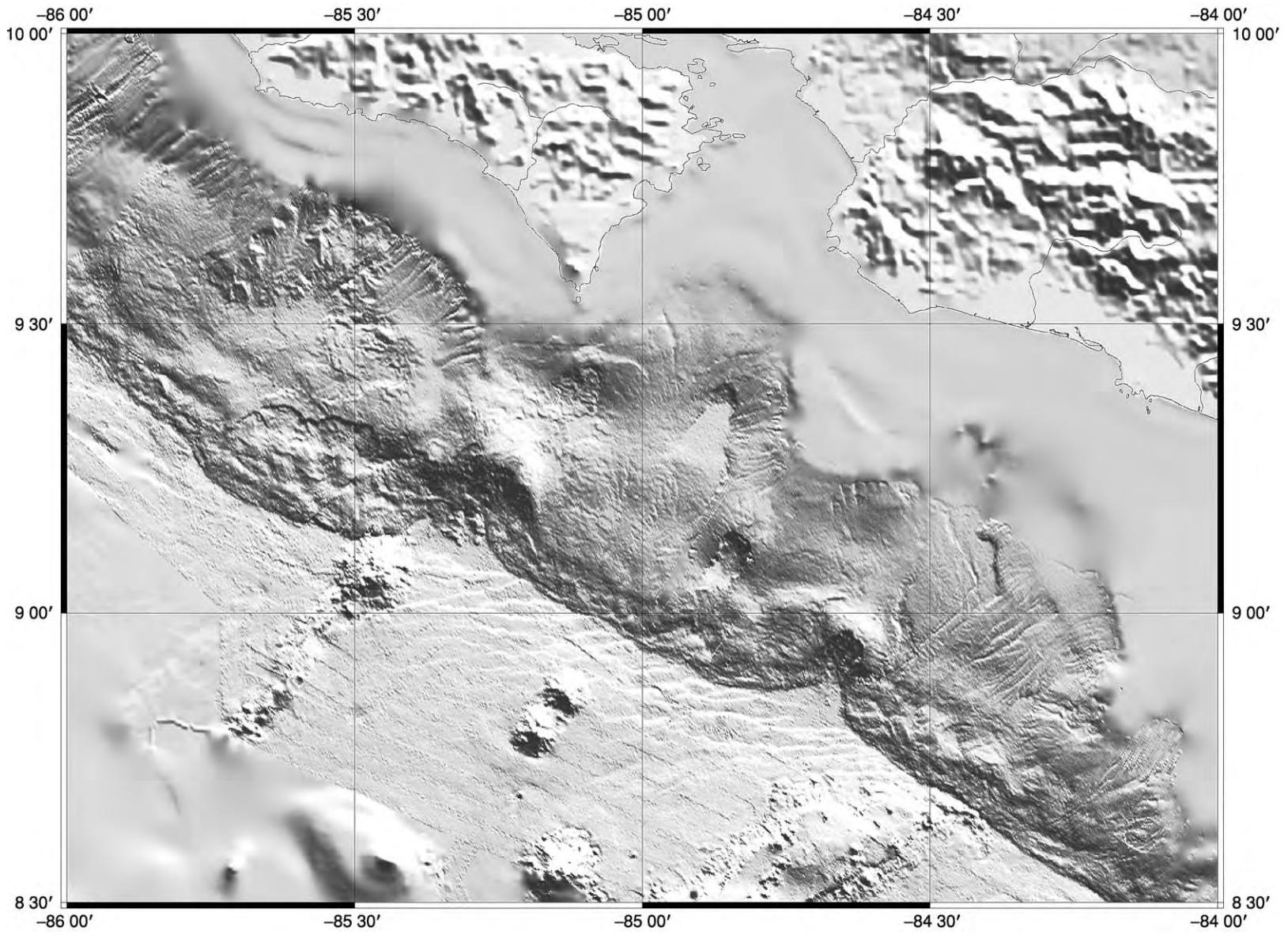


Figure 5 Seamount subduction has produced huge furrows and re entrants in the fore arc margin of Costa Rica. Shaded relief bathymetry reproduced from Dominguez S, Lallemand SE, Malavieille J, and von Huene R (1998) Upper plate deformation associated with seamount subduction. *Tectonophysics* 293: 207 224, copyright 1998, with permission from Elsevier.

atolls highly desirable locations for settlement and vacation.

Glossary

Abyssal hill A low-relief elongate hill formed by horst-and-graben style faulting around a mid-ocean ridge.

Alkali basalt A variety of basalt that has relatively high sodium and potassium contents.

Atoll A ring of carbonate reef islands enclosing a lagoon built atop a subsiding seamount in tropical latitudes.

Pillow lava A type of subaqueous lava flow consisting of spherical or elongate tubes formed by the rapid quenching of lava. Tubes seen in cross-section often resemble pillows, with a flat base and a convex upper surface.

Scoria Glassy porous pyroclastic lava.

Surtseyan A type of volcanic eruption, based on observations at Surtsey in Iceland, which is the phreatomagmatic equivalent of a Hawaiian or quiescent lava flow eruption.

Tholeiitic basalt A variety of basalt characterized by the presence of orthopyroxene in addition to clinopyroxene and calcium-plagioclase.

See Also

History of Geology Since 1962. **Large Igneous Provinces**. **Lava**. **Mantle Plumes and Hot Spots**. **Sedimentary Environments:** Reefs ('Build-Ups'). **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes. **Tectonics:** Hydrothermal Activity; Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Barone AM and Ryan WBF (1990) Single plume model for asynchronous formation of Lamont Seamounts and adjacent Pacific Rise terrains. *Journal of Geophysical Research* 95 N. B7: 10801 10827.

Batiza R (1989) Seamounts and seamount chains of the eastern Pacific. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America, Volume N, The Eastern Pacific Ocean and Hawaii*, pp. 289 306. Boulder: Geological Society of America.

Batiza R and Vanko D (1984) Volcanic development of small oceanic central volcanoes on the flanks of the East Pacific Rise inferred from narrow beam echo sounder surveys. *Marine Geology* 54: 53 90.

Clague DA, Reynolds JR, and Davis AS (2000) Near ridge seamount chains in the north eastern Pacific Ocean. *Journal of Geophysical Research* 105: 16 541 16 561.

Davis EE (2000) Earth science: volcanic action at Axial Seamount. *Nature* 403: 379 480.

Johnson HP and Embley RW (eds.) (1990) Axial Seamount: an active ridge axis volcano on the central Juan de Fuca Ridge. *Journal of Geophysical Research* 95: 12 689 12 966.

Keating BH and McGuire WJ (2000) Island edifice failures and associated tsunami hazards. *Pure and Applied Geophysics* 157: 899 955.

Keating BH, Fryer P, Batiza R, and Boehlert W (eds.) (1987) *Seamounts, Islands, and Atolls*. Geophyiscal Mono graph Series Volume 43. Washington DC: American Geophysical Union.

Mitchell NC (2001) Transition from circular to stellate forms of submarine volcanoes. *Journal of Geophysical Research* 106: 1987 2004.

Moore JG, Normark WR, and Holcomb RT (1994) Giant Hawaiian landslides. *Annual Review of Earth and Planetary Sciences* 22: 119 144.

Ranero CR and Von Huene R (2000) Subduction erosion along the Middle America convergent margin. *Nature* 404: 748 752.

Schmidt R and Schminke H U (2000) Seamounts and island building. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, and Stix J (eds.) *Encyclopedia of Volcanoes*, pp. 383 402. San Diego: Academic Press.

Takahashi E, Lipman PW, Garcia MO, Naka J, and Aramaki S (eds.) (2002) *Hawaiian Volcanoes: Deep Un derwater Perspectives*. Geophysical Monograph Series Volume 128. Washington DC: American Geophysical Union.

Thorarinsson S (1967) *Surtsey: The New Island in the North Atlantic*. New York: Viking.

Wessel P (2001) Global distribution of seamounts inferred from gridded Geosat/ERS 1 altimetry. *Journal of Geophysical Research* 106: 19 431 19 441.

SEDIMENTARY ENVIRONMENTS

Contents

Depositional Systems and Facies

J Collinson, John Collinson Consulting, Beech, UK

Introduction

The concept of facies as applied to sediments is a means of classifying and grouping deposits in such a way that objective differences, usually with genetic significance, are highlighted. In present-day settings, the relationships between processes and environments are self-evident, but in sequences of ancient sediment various analytical procedures are needed to reconstruct environments. 'Facies analysis' is such a suite of techniques and is based on an assemblage of facies and the relationships of the facies to one another. The results of facies analysis may be expressed as facies models, which summarize and, to some extent, idealize the relationships and allow the facies of a stratigraphical interval to be predicted in areas where it cannot be directly observed. Facies analysis also provides a major input to 'sequence stratigraphy' (*see* **Sequence Stratigraphy**), whereby facies successions are placed within a stratigraphical framework defined by key stratal surfaces that are related to changes in relative sea-level or base-level.

Facies

The concept of facies dates back to the early nineteenth century, when A Gressly introduced the term to summarize those descriptive features of a rock that distinguish it from other rocks. Historically, the concept has been applied at a wide range of scales and, therefore, with varying significance. At the largest scale, 'facies' has been applied to major stratigraphical units in particular tectonic contexts. 'Flysch' and 'molasse' are examples of this and are used particularly in the description of the Tertiary fold belts of Europe. At a large scale, 'facies' is also used to describe regional differences of depositional environment within a particular stratigraphical interval. For example, in Britain, one might contrast the 'basinal facies' of the Silurian of Central Wales with the coeval 'shelf facies' of the Welsh Borders. These usages are generally accepted and understood but, in order to arrive at an interpretation of the depositional environment, it is necessary to go through a process of facies analysis. This involves establishing an appropriate facies scheme for the succession being studied and then analysing the spatial relationships of those facies.

Facies schemes are established using a variety of physical, chemical, and biological criteria, and there is no such thing as a universally applicable facies system. Different sedimentary successions demand different schemes because of their contrasting character or because the intended end products of facies analysis may differ. A simple facies scheme may suffice for a rather basic interpretation, whilst a more elaborate scheme may be appropriate for a more refined interpretation. The facies scheme therefore depends on the scale and detail of the study.

In successions of dominantly clastic sediments, a facies scheme would commonly be based largely on

the physical properties of the sediments. Grain size is a common starting point for subdivision. Not only dominant grain size but also secondary features, such as sorting, might be considered. Any differences in composition would also be used for primary subdivision. A subdivision based on grain size may be refined by considering sedimentary structures (e.g. styles of lamination), so that facies such as 'fine-grained, ripple cross-laminated sandstone' or 'horizontally laminated siltstone' might result. Facies defined in this way are called lithofacies, and they can commonly be interpreted in terms of depositional processes by applying knowledge of the relationships between hydrodynamics, particle size, and sedimentary structures. However, many depositional processes occur in a range of sedimentary settings, and, therefore, few lithofacies are fully diagnostic of depositional environment.

Other features, formed soon after deposition, permit more refined facies schemes. Bioturbation by animals or plant roots can provide clues about the environment, and highly distinctive facies such as coal may also allow a specific environment of deposition to be deduced. Where burrowing is important, the assemblage of burrow types and the intensity of burrowing may necessitate schemes of ichnofacies. Early-formed concretions and textures in palaeosols can indicate the drainage conditions of emergent surfaces soon after deposition and can lead to pedofacies. If sediments are fossiliferous, the types and diversities of the fossils may lead to biofacies, and consequent facies analysis may incorporate palaeoecological considerations. Analysis of very fine plant debris or spores may lead to palynofacies.

For carbonate sediments, the nature of the constituent grains is important in establishing the depositional setting, and their characterization through petrographical description may lead to microfacies. In complex sequences, combinations of these different approaches may be appropriate.

Where two contrasting lithologies are interbedded, it may be appropriate to recognize event deposition (Figure 1). In many successions of interbedded muds and sands, sand beds commonly have sharp bases and may show internal grading and vertically changing styles of lamination. They were clearly deposited by sudden high-energy events in otherwise low-energy settings. Establishing the nature of the events may depend on interpreting quite subtle features of both the sand beds and the finer interbeds. Where the muddy component lacks evidence of higher energy and the sands show only current-generated structures and lamination, the events are likely to be turbidity currents in a deep-water setting. If the finer sediments are strongly bioturbated and show signs of fluctuation energy and the sand beds show evidence of wave action, perhaps reworking their tops, then storm events on a shelf may be the appropriate inference. If the finer sediments show evidence of emergence, with soil textures, and the sands are current dominated, then river floods (crevasse events) on a floodplain may be the appropriate interpretation. In all these cases, it would be valid to regard the finer and coarser components as separate facies, but it may prove practical to use composite facies based on the thicknesses and proportions of the coarse and fine components. This approach is particularly applicable to turbidite successions.

Figure 1 Interbedded sandstones and mudstones showing the distinction between the quiet background sedimentation of muds and the high energy events that delivered the sands. In this case, the high energy events were turbidity currents in a deep water marine setting. Paleocene, Zumaia, Spain.

Facies Analysis

Most facies defined according to the criteria outlined above may be interpreted in terms of processes of deposition, but these are seldom unique to a

particular environment. In order to identify the depositional settings of ancient sediments, it is necessary to understand the processes active in present-day environments and the ways in which sediments are deposited and preserved there. In other words, it is necessary to interpret the patterns of facies seen in the rocks by comparison with the patterns of processes established in modern settings. Understanding the interrelationships of facies and processes allows environments of deposition to be inferred with greater precision.

Process-based facies analysis, which began in earnest in the early 1960s, concentrated initially on understanding the vertical succession of facies observed in a borehole core or stream section. Key to this approach was an appreciation of the importance of the nature of the contacts between vertically adjacent facies. In some cases, the contact is a gradational transition, whilst in other cases it is sharp or even erosive. These distinctions are important in interpretation, which is usually based on the application of Walther's Law of the Correlation of Facies, which was first formulated in 1884, but has been applied critically only since the late 1950s. This law, simply stated, points out that facies that are laid down side-by-side in their depositional environment will come to lie vertically on top of one another in the same order as their lateral distribution, or, conversely, only those facies that are superimposed one upon the other without a break can have been deposited in laterally adjacent settings (Figure 2). The importance of the nature of the vertical facies contacts is apparent in Figure 2, which shows that an erosion surface can allow the superposition of facies deposited in widely separated settings. Sedimentary hiatuses can have the same result. A rapid transgression can lead to shallow-water facies being directly overlain by deep-water facies, with facies representative of intermediate environments missing across the contact because of a lack of deposition or condensation, though not necessarily requiring a temporal break in sedimentation.

Some facies successions show systematic trends of vertical change, expressed through grain size or other properties. In other cases, a less organized succession may occur. Trends of vertically increasing and decreasing grain size (coarsening-up and fining-up units) are amongst the most common patterns encountered in clastic successions, and these may, for example, be equated with prograding shorelines and laterally migrating channels, respectively (Figure 3). Such facies trends may occur as single examples or as vertically repeated patterns. Where repetition occurs, successions are commonly described as cyclic, and the idea of cyclic sedimentation has a history extending back to the early nineteenth century. In cyclic successions, it is important to distinguish between interpreting the facies sequence within cyclothems (i.e. the environment) and explaining the cyclicity (i.e. the repetition). Facies patterns may repeat vertically because of inherent properties of the depositional environment, such as the tendency for river or distributary channels to switch positions on a floodplain or delta top. Such processes are termed autocyclic. Where external controls, such as fluctuating sea-levels or subsidence rates, are thought to cause the repetition, the processes are termed allocyclic.

In cyclic successions, it is rare for all cyclothems to show an identical facies sequence. Rather, they are variations on a theme (Figure 4). Various statistical techniques can be employed to extract a 'complete' or 'ideal' cyclothem, the interpretation of which can be developed into a facies model. However, it is also instructive to consider the variations in the facies successions of cycles and their relationships with the ideal cycle, since the variability itself may help to refine the model. For example, a series of coarsening-up cycles may record successive progradations of a delta, whilst the differences between individual cycles may reflect differences between subenvironments of the delta, stacked in vertical succession through autocyclic delta-lobe switching. Comparison of the facies in the upper parts of several

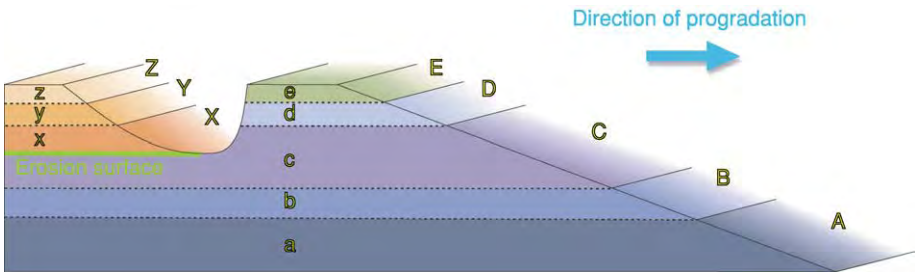


Figure 2 An illustration of the operation of Walther's Law of the Correlation of Facies. The subenvironments A E lie side by side and, through progradation, generate a vertical facies succession a e. The development of an erosion surface through the migration of a channel triggers a new facies sequence, x z, which is independent of the earlier one. The erosion surface separates facies that were generated in settings that have no predictable lateral relationship.

Figure 3 A large scale coarsening up unit generated by the progradation of a major delta. The unit ends abruptly at a flooding surface (arrowed), above which are stacked smaller scale coarsening up units (parasequences). Namurian, County Clare, Ireland.

cycles may lead to inferences about the variability of the delta front and delta plain and hence allow a more refined characterization of delta type. The validity of such an approach depends, of course, on the successively prograding deltas having been of broadly similar character.

In successions where no cyclicity is apparent and no patterns of vertical facies change can be identified, mathematical investigation such as Markhov chain analysis may help to detect an underlying pattern beneath the random noise. However, the demonstration of statistical significance may be limited by insufficient data.

Architectural Elements and Bounding Surfaces

Where sedimentary successions are very well exposed in extensive desert or coastal cliffs, it is clearly not sensible to rely solely on the vertical sequence of facies to reconstruct the depositional environment. Lateral facies relationships, where they can be

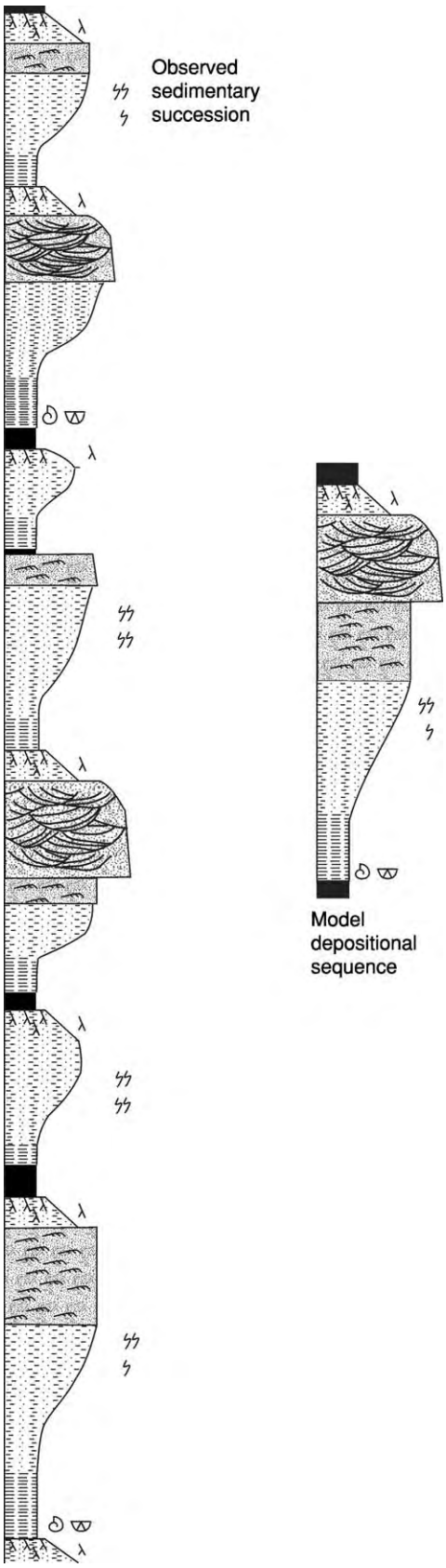


Figure 4 The facies succession of an idealized model cyclothem compared with the observed facies successions upon which it is based.

observed, are highly instructive, and the two- and three-dimensional shapes and sizes of facies units may be highly diagnostic. Lateral changes in facies or in the thicknesses of facies units (architectural elements) may help to localize particular elements of the depositional environment. For example, the thickening of crevasse splay sands within a floodplain succession may be linked to the proximity of a channel sandstone.

Laterally extensive exposures may reveal subtle low-angle dipping surfaces that would remain undetected in one-dimensional sections such as borehole cores. Such surfaces may be confined within channel sand bodies, indicating lateral accretion of the channel or downstream accretion of a channel bar. They may also occur on a larger scale within coarsening-up units such as progradational clinoforms deposited in a delta-front setting. Such surfaces contribute to facies definition and, more importantly, they aid interpretation by constraining the context of the facies.

Within complex channel sandstones, the usefulness of Walther's Law is often limited by abundant erosion surfaces. In a narrow one-dimensional section, the significance of a particular erosion surface may be difficult to establish. The basal erosion surface of a channel sand body, where it rests on contrasting, probably finer-grained, facies, is likely to be extensive, but erosion surfaces within the sand body, perhaps picked out by pebble-lag or intraformational conglomerates, may have more local significance. Within large complex channel units, laterally extensive exposures commonly show that erosion surfaces fall naturally into a hierarchy of scale, lateral extent, and genetic significance. The bodies of sediment that they bound are termed architectural elements, facies building blocks of differing genetic significance that are linked to the geomorphology of the channel system (Figure 5). Within aeolian dune deposits, similar hierarchies of bounding surfaces are critical to the discrimination of dunes and larger compound bedforms.

It is clear that the more fully the three-dimensional patterns of facies are established, the more precise can be the environmental interpretation. Where only subsurface data are available, the one-dimensional facies information provided by boreholes may sometimes be suitably supplemented by three-dimensional seismic data. However, in most cases, the lateral extent of facies units and lateral facies relationships must be predicted intuitively using Walther's Law and judiciously chosen analogues, usually from extensive outcrop exposures.

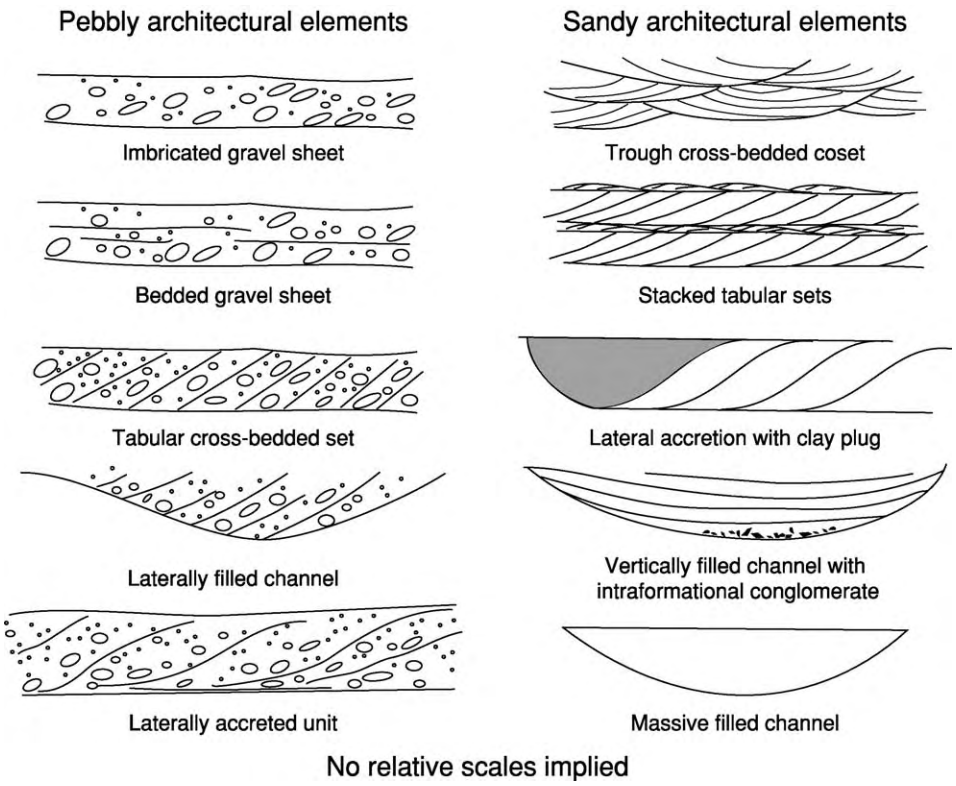


Figure 5 Examples of the various 'architectural elements' that can commonly be identified in alluvial channel deposits. The architectural elements are the products, usually only partially preserved, of various morphological features (bedforms, bars, channel fills) that occupied the channel floor.

Facies Models

Facies models are attempts to summarize the essential aspects of facies sequences and relate them to inferred depositional environments. They are usually most effectively conveyed graphically as three-dimensional block diagrams that relate the environment and its behaviour to the facies pattern. Facies models have been proposed for most major depositional environments, and there is a large measure of consensus about their general aspects, particularly for fluvial, aeolian, deltaic, and other shallow-water settings. For deep-water sand settings, differences of opinion about what constitutes a 'submarine-fan model' have largely been resolved by recognizing that there is a whole spectrum of fan types and, accordingly, a spectrum of valid models. General rather idealized models of major depositional settings have their place as educational tools. They also form a basis with which to compare an actual sedimentary succession, where differences between the ideal and the actual may be important in refining the interpretation. However, it is important to recognize that some settings are inherently more organized and predictable than others and that lateral variation takes place on a variety of scales. This is important when a model is used to predict the nature of a succession between or beyond known data points. In a well-organized setting, it may be appropriate to build a predictive model that is deterministic, and predicted interpolations and extrapolations can be expected to reflect reality closely. In more complex, less organized settings, the stochastic component means that such predictions must be generalized and expressed in probabilistic terms.

Sequence Stratigraphy and Key Surfaces

At outcrop or in a borehole section, the establishment of a sequence stratigraphical framework depends on distinguishing key surfaces of potentially widespread significance that can be related to episodes of rising or falling base level, which in marine settings is relative sea-level, in lacustrine settings is lake level, in fluvial settings is the graded fluvial profile, and in aeolian settings is the water table.

Erosion surfaces at the bases of channel units or elsewhere in a sequence are, in many cases, simply consequences of the depositional processes operating within the environment, for example the migration of a river channel on a flood plain. Such surfaces are autocyclic and have only local significance. Other channel bases may have much more widespread stratigraphical significance and may be related to valley incision caused by a fall in base level (i.e. allocyclic control) (**Figure 6**). These are one expression of sequence boundaries in sequence stratigraphical terminology. Distinguishing between the two types of erosion surface can be difficult, based on features of the erosion surface alone, especially

Figure 6 A major erosion surface (arrowed) at the base of a thick multistorey channel complex. The erosion surface is regional in extent and has a mapped relief of several tens of metres. It is inferred to be the base of an incised palaeovalley and would be regarded as a sequence boundary, related to a major fall in base level (relative sea level). Namurian, County Clare, Ireland.

where local channel bases are nested within a larger-scale palaeovalley. However, deeper incision as a result of base-level fall may cause the erosion surface to cut down into facies that differ from those beneath autocyclic channel bases. Outside the confines of an incised valley, the long-term emergence of the sedimentary surface and the resulting nondeposition may allow highly mature soil profiles to develop, so that recognition of palaeosol facies in interfluve areas may have widespread stratigraphical significance. Non-erosive sharp contacts at the bases of coarser facies may also indicate that regressions were forced by falling base level. Falling base level or relative sea-level, therefore, may be manifested in a variety of different ways leading to uncertainty in discrimination and correlation.

Rises of base level or relative sea-level produce more uniform facies expressions, as they tend to be associated with transgression and a cut-off of sediment supply to more distal settings owing to sediment being trapped closer to its source. The results are flooding surfaces (**Figure 3**) characterized by condensation, often to the point of nondeposition. In marine settings, such surfaces are commonly characterized by intense bioturbation, reflecting the increased time available for animals to rework the sediment before burial. When flooding occurred in humid near-emergent settings, peat developed as plant productivity balanced increasing accommodation space. Coal seams, the compacted products of peat, may be of local or regional significance. Where flooding occurred through autocyclic processes, for example by cut-off of sediment supply through distributary switching on a delta plain, the seam extent reflects the area of an interdistributary bay or floodplain lake. Where the control was allocyclic, coal seams may be regionally important stratigraphical markers. Flooding surfaces, however expressed in the sediments, commonly mark the bases of coarsening-up units, termed parasequences, that record the progradations that eventually filled the resulting bodies of water. Where flooding is associated with the transgression of a high-energy shoreline, the transgressive surface may be an erosional pebble lag ravinement surface, separating nearshore sediment below from more offshore sediment above.

See Also

Sedimentary Environments: Alluvial Fans, Alluvial Sediments and Settings; Deltas. **Sedimentary Processes:** Depositional Sedimentary Structures; Post-Depositional Sedimentary Structures; Aeolian Processes. **Sedimentary Rocks:** Mineralogy and Classification. **Sequence Stratigraphy**. **Soils:** Palaeosols. **Trace Fossils**.

Further Reading

Anderton R (1985) Clastic facies models and facies analysis. In: Brenchley PJ and Williams BPJ (eds.) *Sedimentology: Recent Developments and Applied Aspects*, pp. 31 47. Special Publication 18. London: Geological Society.

Galloway WE (1989) Genetic stratigraphic sequences in basin analysis. 1. Architecture and genesis of flooding surface bounded depositional units. *Bulletin of the American Association of Petroleum Geologists* 73: 125 142.

Leeder MR (1999) *Sedimentology and Sedimentary Basins*. Oxford: Blackwell Scientific.

Nichols G (1999) *Sedimentology and Stratigraphy*. Oxford: Blackwell Scientific.

Reading HG (ed.) (1996) *Sedimentary Environments: Processes, Facies and Stratigraphy*, 3rd edn. Oxford: Blackwell Scientific Publications.

Reading HG (2001) Clastic facies models, a personal perspective. *Bulletin of the Geological Society of Denmark* 48: 101 115.

Scholle PA and Spearing D (eds.) (1983) *Carbonate Depositional Environments*. Memoir 33. Tulsa: American Association of Petroleum Geologists.

Scholle PA, Bebout DG, and Moore CH (eds.) (1983) *Sandstone Depositional Environments*. Memoir 31. Tulsa: American Association of Petroleum Geologists.

Selley RC (1996) *Ancient Sedimentary Environments and their Subsurface Diagnosis*, 4th edn. London: Chapman & Hall.

Tucker ME (1985) Shallow marine carbonate facies and facies models. In: Brenchley PJ and Williams BPJ (eds.) *Sedimentology: Recent Developments and Applied Aspects*, pp. 147 169. Special Publication 18. London: Geological Society.

Walker RG and James NP (eds.) *Facies Models; Response to Sea Level Change*. Waterloo, Canada: Geological Society of Canada.

Alluvial Fans, Alluvial Sediments and Settings

K D Jäger, Martin Luther University, Halle, Germany

The term alluvia (alluvions) comprises clastic sediments, deposited by the running water, on river bottoms as well as on the surface of adjacent riverside areas inundated temporarily by floods. This material has to be attributed to the particular catchment area. Within the respective boundaries the sedimentary material supply is dominated by the area shared by open ground where protection by the vegetation cover has been reduced or entirely removed. Consequently, forest clearance, as well as arable land, have caused soil erosion by the clearance, development, and use of this open ground.

The varying intensity of soil erosion during different historical periods following each other reflects – especially in manmade landscapes – the alternating intensity of settlement and agrarian landuse over the course of time. This interrelation may be exemplified by numerous archaeological datings of alluvial layers in stratigraphical sequences of riverside exposures. Frequently, the dated development of such layers is linked with events structuring the development of settlement and landuse.

The petrographic composition of alluvial deposits reflects the superficial substrates (as rocks or loose earth masses) of the particular catchment area. However, the composition is dominated by components from superficially exposed open ground.

Consequently, since the first occurrence of agriculture, more and more alluvial deposits came into being. However, regional chronological differences of the global advance of agriculture have to be considered. As a rule, alluvions have been originated and developed during the Holocene to date. Moreover, usually the deposits are still unconsolidated.

The composition of alluvions is determined substantially, but not exclusively, by the material supply from the particular catchment area. Other decisive factors consist in continuity, discontinuity, and velocity of water flow. Their effect is reflected by the granulation of the sediment: high flow velocity results in course granulation, less velocity in fine granulation. Therefore, mainly gravels and sands occur more often at the upper courses of running waters provided with steeper inclination, for example, in mountain terrains. On the other hand, courses characterised by low inclination, as in the lowlands and near to the mouth of rivers, result in clay deposition. Frequently, the granulation of the deposits is mixed, and this feature is reflected by the nomenclature of alluvions in many languages as German ('Auelehm'), Russian ('пойменная глина'), Czech ('nivní hlína'), whereas other terms try to avoid a concrete relation to the granulation of the deposit as English ('alluvium'), French ('alluvion'), or the proposed German term 'Klock' (adapted from popular use in Lower Lusatia). Independent of granulation, is the term alluvial substrates, soil types in pedological nomenclatures using the term 'vega' (from the Spanish meaning of 'riverside').

On riversides, a two-dimensional differentiation of the distribution of alluvial deposits frequently occurs and corresponds to regular patterns. The pattern of superficial relief differentiation is determined by the buildup of embankments and channels. The spatial differentiation of processes and patterns of the relief, as of the sediment distribution on riversides, has been modified frequently, especially as a consequence of extreme flooding of the respective rivers. Therefore, a high- and narrow-scale heterogeneity characterises the area of stratigraphical subdivision of alluvions. A temporal modification of the spatial pattern of the riverside has been exemplified in the case of the lower Vistula in Poland.

Frequently, the alluvions are linked with alluvial and colluvial fans. The former occur at the confluence of tributaries to their main valley, the latter mainly in slope foot positions of the riverside margins. Floods from tributaries can rearrange the superficial shape of the stratigraphy of alluvions in the respective main valley, too, which has been exemplified by detailed investigation in the region of Ulm (Germany), where the River Iller flows into the Danube (so-called 'Iller-Schwemmkegel'). The appearance of colluvial fans covering flood-plain soils, concealing alluvial deposits, has been proved recently by examples from the Upper Rhine Lowlands region surrounding the Kaiserstuhl massif: Endingen, Emmendingen. A common feature of these investigations, with other detailed stratigraphical studies devoted to the alluvions in the valleys, is provided by archaeological observations proving precise Holocene origin. These datings include two other important indications. They are evidence of man's presence and impact on the landscape by irrigation and drainage works, as well as soil erosion, as a consequence of land use. On the other hand, the datings prove the alternation of sedimentation and erosion, respectively, incision into previous valley bottoms with their alluvions, as well as into alluvial fans.

The same situation for determining datings and the sequence of processes and events is valid regarding the stratigraphy of alluvions in more extended riversides and main valleys. The detailed regional examination and the comparability of the results may be exemplified by the studies described in the further reading section at the end of this chapter.

Sometimes the archaeological datings of single layers within alluvial strata evidence the covering of ancient sites of human presence and settlement by subsequent floods, as is the case of urban centres connected with the medieval Slavonic state of Greater Moravia (actually in Czechia: Pohánsko, Mikulčice, Staré Město, and Uherské Hradiště).

Other chances of achieving more specific and precise dating of the sequence of Holocene events and processes in the stratigraphy of alluvions, are provided by trunks of trees engulfed by the fluvial deposits. Trunks have been found frequently, for instance, in many central European valleys, as Elbe and Spree, Rhine and Main, or the Danube with its tributaries. In the German literature, they are called 'Rannen'.

On a regional scale, there are numerous trunks within the sediment. In the brown coal seams of central Germany (Saxonia, Saxony-Anhalt, Lusatia) the respective fluvial deposits are frequently well-exposed. However, independent of the conditions of exposure, these trunks have provided the opportunity of dendrochronological datings. Using this method can be applied to examination of the valley development in many cases. This may be exemplified for some catchment areas in central Europe as that of the Danube and its tributaries, or that of the River Main.

The regional development of valley-net systems is reflected by the stratigraphy of alluvions. A detailed documentation of varying stratigraphical observations in different parts of one and the same valley-net illustrates the regional differentiation in the course of events. This differentiation has been exemplified carefully in the case of the River Vistula (Weichsel) with its tributaries in Poland. The chronological foundation of this investigation has been completed by radiocarbon datings and pollen analysis.

However, in other parts of temperate Europe, careful and detailed stratigraphical examinations have covered different sections of the same catchment area. Such investigations may be exemplified by the valley system of the River Severn in the British Isles (counties of Shropshire and Worcestershire, England).

Alternating layers in the stratigraphy of alluvions reflect temporally changing conditions of accumulation, as in single valleys or valley segments, such as completely examined valley-net systems. These changes record climatic consequences, with the varying amount of precipitation in the respective catchment area and its temporal peaks caused by floods. They also reflect the changing intensity of human impact influencing the sedimentary yield of the respective river or creek. This yield comprises effects resulting from the extension of open grounds, forest clearance, and utilisation of arable land in the respective catchment area.

The same causes have ruled the origin, development, and stratigraphy of alluvions in the Mediterranean. Evidence has been provided previously for the Mediterranean and more recently, in particular, from the Iberian Peninsula.

In all the investigated catchment areas, the effect of human impacts depends on the respective regional course of settlement and land-use intensity. The climatic causes take effect in extended regions of the European continent in a comparable way. However, the detection of palaeoclimatic fluctuations, and especially of oscillations of the water balance in the course of the Holocene, is based on other stratigraphic observations as in exposures of freshwater lime deposits. Among geostratigraphical, palaeontological (especially palynological and palaeomalacological), and archaeological evidence supporting the resulting concept concerning the palaeoclimatic development during Holocene, the stratigraphy of alluvions claims an important but not essential role.

Related to the youngest alluvions, historical record archives especially support the linking of distinct stratigraphical facts with particular palaeofloods.

However, irrespective of all other sources clarifying the palaeoenvironmental development as a whole and on the valley bottoms in particular, significant contributions to the reconstruction of previous conditions during the Holocene are provided by the alluvions and their stratigraphy.

See Also

Europe: Holocene. **Sedimentary Processes:** Fluvial Geomorphology. **Sedimentary Rocks:** Rudaceous Rocks.

Further Reading

Andrzejewski L and Juśkiewicz W (2003) Lithofacies diversification of the alluvia in the area of Kepa Dzikowska, Kepa Bazarowa, Kepa Strońska and of the Vistulian floodplain near Toruń. *Prace Geograficzne* 189 (Warszawa): 159 176.

Becker B (1982) *Dendrochronologie und Paläoökologie subfossiler Baumstämme aus Flußablagerungen Ein Beitrag zur nacheiszeitlichen Auenentwicklung im südlichen Mitteleuropa.* Wien. Mitt. Komm Quartär forschung d. Österr. Akad. d. Wiss. No. 5.

Becker B (1983) Postglaziale Auwaldentwicklung im mittleren und oberen Maintal. In: Schirmer W. (ed.) *Holozäne Talentwicklung Methoden und Ergebnisse.* Geol. Jb., Reihe A, 71: 45 59.

Brown AG (1983) Floodplain deposits and accelerated sedimentation in the lower Severn basin. In: Gregory KJ (ed.) *Background to Palaechydrology A Perspective,* pp. 375 397. Chichester: John Wiley & Sons.

Dury GH (1983) Osage type underfitness on the River Severn near Shrewsbury, Shropshire, England. In: Gregory KJ (ed.) *Background to Palaechydrology A Perspective,* pp. 399 412. Chichester: John Wiley & Sons.

Graul H and Groschopf P (1952) Geologische und morpho logische Betrachtungen zum Iller Schwemmkegel bei Ulm. *Ber. Naturforsch. Ges. Augsburg* 5: 3 27.

Havlíček P, Galuška, and Poláček (2004) Die geologische Situation im Bereich des großmährischen Zentrums von Staré Mesto Uherské Hradiště. *Studien zum Burgwall von Mikulée* 7: Brno, 1 17.

Heine K and Niller HP (2003) Human and climate impacts on the Holocene landscape development in southern Ger many. *Geografia Polonica* 76(2): Warszawa, 109 122.

Jäger K D (1962) Über Alter und Ursachen der Auelehma blagerung Thüringer Flüsse. *Praehist. Z.* 60(1/2): 1 59.

Jäger K D (1967) Anthropogene Ablagerungen im Holozän der südöstlichen Thüringer Triasmulde. In: Kliewe H (ed.) *Probleme und Befunde der Holozänstratigraphie in Thüringen,* Sachsen und Böhmen, Berlin and Prague, pp. 31 51.

Jäger K D (1986) Reliefmustertypen in Talauen im nördli chen Mitteleuropa. *Acta Univ. Nicolai Copernici, Geo grafia XXI (Nauki matermatyczno przyrodnicze* 67: *Toruń*, 109 117.

Jäger K D (1997) Anthropogene Ablagerungen im Holozän Mitteleuropas. *Brandenburgische Geowissenschaftl. Beiträge* 4(3): Kleinmachnow, 89 96.

Jäger K D (2002) Oscillations of the water balance during the Holocene in interior Central Europe features, dating und consequences. *Quaternary International* 91: 33 37.

Jäger K D (2003) Die Oberfläche: Uferwälle und Rinnen im Oderbruch. In: Schroeder JH and Brose F (eds.) *Oderbruch Märkische Schweiz Östlicher Barnim.* Führer Geol. Berlin u. Brandenburg 9. Berlin, Techn. Univ., 66 77.

Kubiena W (1953) *Bestimmungsbuch und Systematik der Böden Europas,* p. 392. Stuttgart: Ferdinand Enke.

Limbrey S (1983) Archaelogy and palaeohydrology. In: Gregory K J (ed.) *Background to Paleohydrology A Per spective,* pp. 190 121. Chichester: John Wiley & Sons.

Litt Th (1988) Stratigraphische Belege für anthropogen ausgelöste Bodenlagerungen vom Neolithikum bis zur frühen Eisenzeit im circumhercynen Raum. *Ethnographisch Archäologische Z.* 29: 129 137.

Litt Th, Hiller A, and Eissmann L (1991) Zur Entwicklung der jungquartären Tieflandstäler im Saale Elbe Raum unter besonderer Berücksichtigung von C^{14} Daten. *Eiszeitalter u. Gegenwart* 41: 26 46.

Litt Th (1994) Holozäne Talentwicklung am Beispiel der Elster Luppe Aue, Tagebau Merseburg Ost. In: Eissmann L and Litt Th (eds.) *Das Quartär Mitteldeutschlands Ein Leitfaden und Exkursionsführer,* pp. 333 337. Altenburger Naturwiss. Forschungen 7, Altenburg.

Mäckel R, Schneider R, and Seidel J (2003) Anthropogene Impact on the Landscape of southern Badenia (Germany) during the Holocene documented by colluvial and alluvial sediments. *Archaeometry* 45(3): 487 501.

Pfister Chr (1999) *Wetternachhersage 500 Jahre Klimavar iationen und Naturkatastrophen,* p. 304. Bern: Stuttgart and Wien.

Pfister Chr, Brázdil R, Glaser R, *et al.* (1999) Documen tary evidence on climate in sixteenth Europe. *Climatic Change* 43: 55 110.

Scamoni A (1954) Die Waldvegetation des Unterspree waldes. *Arch. Forstwesen* 2: 232 244.

Scheffer F and Schachtschabel P (1988) *Lehrbuch der Bod enkunde,* 14th edn., p. 494. Stuttgart: Ferdinand Enke.

Schulte L (2000) Climatic and human impact on river systems in Southeast Spain. In: Díaz del Olmo F, Faust D, and Porras AI (eds.) *Environmental Changes during the Holocene,* pp. 45 49. Sevilla.

Schulte L (2002) Climatic and human influence on river systems and glacier fluctuations in southeast Spain since the Last Glacial Maximum. *Quaternary International* 93/94: 85 100.

Schulte L (2003) River response and terrace aggradation in the Mediterranean Iberian Peninsula during historical times. In: Thorndycraft VR, Benito G, Barriendos M, and Lasat C (eds.) *Palaeofloods Historical floods and Climatic Variability: Applications in flood Risk Assess ment,* Proc. of the PHEFRA Workshops, Barcelona 2002, pp. 67 72.

Starkel L (ed.) (1982 and 1987) *Evolution of the Vistula river valley during the last 15000 years.* 2 vols. Wroclaw etc.: Polish Ac. Sci., Geographical Studies, Special Issues.

Vita Finzi Cl (1969) *The Mediterranean Valleys Geo logical Changes in Historical Times,* p. 139. Cambridge: Cambridge University Press.

Vollrath H (1963) Die Morphologie der Itzaue als Ausdruckhydro und sedimentologischen Geschehens. *Mitt. d. Fränkischen Geograph. Ges. (Erlangen)* 10: 297 309.

Anoxic Environments

P B Wignall, University of Leeds, Leeds, UK

Introduction

Nearly all modern environments, on both land and sea, are well oxygenated, thanks to the ample levels of oxygen available in the atmosphere. This also seems to have been the case for much of Phanerozoic history. Lack of oxygen appears to have been a critical feature for life only during specific time intervals (see below). However, you do not have to look very far to find anoxic environments in many terrestrial and marine settings. Oxygen is consumed by respiration, particularly by bacteria that use this powerful oxidant to break down organic matter. Therefore, in settings where abundant organic matter accumulates (e.g. swamps, estuarine muds, deep lakes, and shelf sea sediments) the oxygen is rapidly consumed, producing an oxygen-free (anoxic) zone that may, in some sediments, be only a few centimetres beneath the surface. Within the anoxic zone anaerobic bacteria continue to oxidize organic matter by using a series of increasingly less efficient oxidants, starting with nitrates, then sulphates, and finally fermentation processes that use carboxyl groups in the organic matter as oxidants. In the marine realm, sulphate ions are the most readily available oxidant after oxygen, and a sulphate reduction zone is therefore often well developed. The by-product of this activity is hydrogen sulphide, which rapidly reacts with iron oxides in the sediment to produce the mineral pyrite, which therefore often provides tell-tale evidence of sulphate reduction zone activity. In non-marine anoxic environments, such as swamps and lakes, sulphate is often not available, with the result that the surface oxic layer of sediment is often directly underlain by a methane-generating (methanogenic) zone.

Modern Anoxic Environments

The development of open-water anoxia is a relatively rare phenomenon in the modern world. Anoxia develops where the supply of oxygen is exceeded by the demand for oxygen by decaying organic matter, with the result that the chemical zones summarized above move up out of the sediment and into the water column. This occurs in two distinct modern environments: silled basins and upwelling zones (Figure 1). Silled basins arise where a body of deep water is semi-isolated from the world's oceans by the presence of a shallow-water sill that can restrict deep-water circulation. This occurs particularly if the basin receives a lot of freshwater run-off and thus exports a low-density surface layer of water across the sill into the open sea. The basin is therefore said to have a positive water balance and to display an estuarine circulation. This can effectively seal the deeper waters from vertical mixing and oxygen supply. The present-day Black Sea is a giant example of a silled anoxic basin, although many fjords provide smaller examples. These modern examples owe their density stratification to the differences in density caused by variations in salinity. However, temperature contrasts can also produce stratification, since warmer waters are less dense than colder waters, and it is often mooted that some ancient examples of silled anoxic basins may have been thermally stratified.

Silled anoxic basins are essentially the product of a failure of oxygen supply to deeper waters. Upwelling-zone anoxia, in contrast, is the product of an overabundance of organic matter and therefore an excessive oxygen demand. West-facing continental margins are often characterized by prevailing offshore winds, which drive surface waters offshore (Figure 1B). This water is replaced by water that upwells from a considerable depth and has usually originated in polar areas. Because this deep water has not resided in the photic zone for some considerable time, it is often nutrient rich, and its arrival at the surface of the ocean stimulates very high plankton productivity. It is when this plankton dies and descends that decay processes rapidly consume oxygen, producing an oxygen-minimum zone (OMZ) within the water column. Intense OMZs are known off the coasts of California and South America, although truly anoxic conditions are not developed. The zone of oxygen-poor deposition beneath an upwelling zone is at an intermediate water depth, where the OMZ intersects the continental slope (Figure 1B). This contrasts with the silled-basin environment, where the anoxic deposition is in the deepest-water locations.

Identifying Ancient Anoxic Environments

Recognizing anoxic environments in sedimentary rocks is relatively straightforward because of the distinct range of attributes produced by anoxic conditions. Probably the most diagnostic feature of

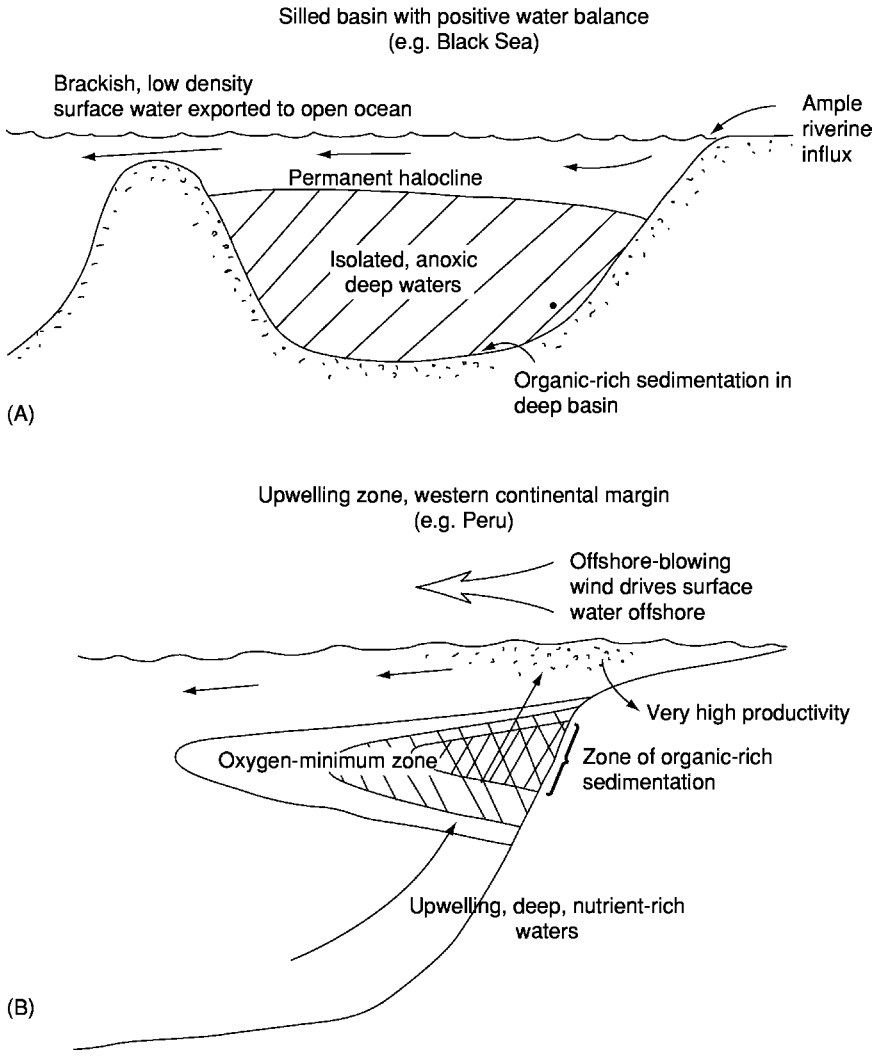


Figure 1 Models of the production of open water anoxic environments. (A) The modern Black Sea is the type example of silled basin anoxia, whilst (B) the high productivity upwelling environments are encountered along the western seaboard of several countries, including Peru and the USA. Note that in silled basins the site of organic rich deposition is in the deepest waters, whereas beneath upwelling zones organic rich deposition occurs at intermediate water depths, often on continental slopes, as depicted here, or in deep shelf locations.

marine anoxia is the presence of abundant pyrite – particularly its framboidal form – produced in the sulphate reduction zone. Framboids are named after the French word for raspberry (because that is what they look like, at least superficially); they are spherical aggregates of tiny pyrite crystals (microcrysts) a few microns in size (Figure 2). The framboids themselves range from a few microns to a few tens of microns in size, although a diameter of 5–6 μm is typical. Framboids form today in weakly anoxic conditions above well-developed sulphate reduction zones where reduced iron reacts rapidly with the hydrogen sulphide produced in the zone beneath. When these conditions are developed within the water column only small (up to 6 μm in diameter) framboids are able to form before they sink out of the water column. Populations of tiny framboids are therefore produced when sulphide is present in the lower water column. This is sometimes called a euxinic environment. Note that very minor amounts of crystalline (non-framboidal) pyrite also form in some sediments beneath oxygenated bottom waters, but pyrite-rich sediments are highly diagnostic of anoxic environments.

The change from oxygenated to anoxic conditions causes a valence change in most metals with the result that anoxic sediments contain different concentrations of these elements from normal shales. Many metals (e.g. vanadium, molybdenum, and nickel) often precipitate as impurities in pyrite with the result that they are relatively enriched in black shales. Other elements occur in a more soluble state in anoxic conditions (e.g. phosphorus and manganese) and are therefore lost from the anoxic environment, although they often precipitate at the anoxic–oxic boundary.

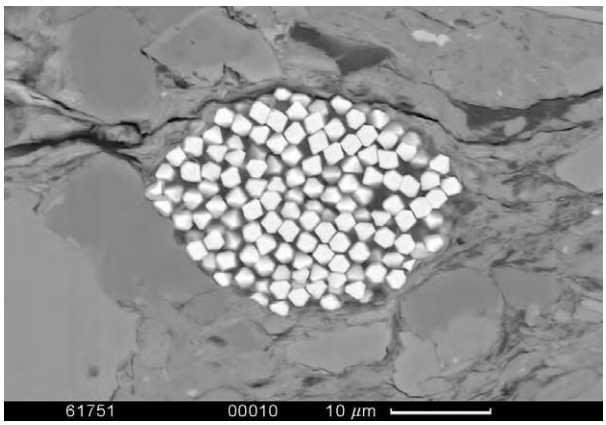


Figure 2 Scanning electron microscope image of a pyrite framboid, from the *annulata* Shale (Upper Devonian) of northern Germany. Such spherical accumulations of tiny pyrite microcrystals (microcrysts) form in prolific numbers in anoxic marine environments.

This is particularly the case for manganese, which is often concentrated as a carbonate on the margins of anoxic areas.

Although a subject of controversy, anoxic conditions are thought by many to favour the preservation of organic matter, and many anoxic sediments are organic-rich. This is manifest in the dark colour of the sediments, and the epithet 'black shale' is often used to describe anoxic sediments. Such sediments often also display very fine lamination, and they characteristically weather into thin sheets to produce a rock known as a paper shale. The preservation of fine laminae is testimony to the fact that animals cannot live in anoxic conditions. Thus, the worms and other creatures that normally disrupt delicate laminae by their burrowing activities are absent. On the face of it, fossils should also be absent from anoxic sediments because the anoxia inhibits all but microbial life. However, on the contrary, many black shales contain abundant and well-preserved fossil remains. Many of these belong to creatures that swam, such as fishes (*see* **Fossil Vertebrates:** Fish) and ammonites (*see* **Fossil Invertebrates:** Ammonites), and so their presence is readily explained: they presumably lived in the oxygenated surface waters and sank into the anoxic deeper waters only after death (Figure 3A). However, many fossils in anoxic sediments belong to ostensibly bottom-living forms, notably bivalves (*see* **Fossil Invertebrates:** Bivalves) and brachiopods (*see* **Fossil Invertebrates:** Brachiopods) (Figure 3B). For a long time such occurrences have been a subject of debate, but most workers now agree that they are dysaerobic fossils. The term 'dysaerobic' was first coined for modern forms found living on the seafloor within the OMZ of the Californian Borderland. Low-oxygen bottom waters provide harsh conditions for animals with the result that only a low-diversity, but often abundant, community of creatures can survive. Low diversity–high abundance is a common attribute of black-shale fossil assemblages too, and they are generally regarded as representing a community tolerant of low oxygen. Ancient dysaerobic assemblages are often dominated by bivalves that display a shell morphology of broad, flat, often circular valves, sometimes with dense, fine radial ribs. These are loosely termed 'paper pectens', and this characteristic morphology has evolved again and again in black-shale depositional environments (Figure 3B).

By implication, the presence of fossils in a black shale suggests the presence of oxygen during deposition, although levels need not have been very high. However, the geochemistry of anoxic sediments implies that there was no oxygen at all during deposition. For example, uranium is concentrated in sediments only in the complete absence of oxygen; otherwise it occurs as a highly soluble ion that is not precipitated. Thus, there is often a discrepancy between geochemical and palaeontological evidence for oxygen levels. This can be resolved if it is appreciated that black shales record a range of depositional environments in which the average depositional conditions were anoxic but seafloor oxygen was present during some, probably brief, intervals. These oxygenation events would allow transient colonization by rapidly dispersing species such as paper pectens. This is a distinctly different concept from that originally proposed for dysaerobic faunas, which envisaged persistently low seafloor oxygen levels, and as a result the alternative names 'episodically dysaerobic' and 'poikiloaerobic' have been proposed. However, neither has really caught on in the literature, and the term dysaerobic continues to be used. In order to get away from these generic terms, a simple descriptive oxygen-restricted biofacies scheme has been provided based on fossil and sediment attributes (Figure 4). Based on British Jurassic examples, the scheme recognizes a gradient of features thought to record improving seafloor oxygen levels. Oxygen-restricted biofacies 3 and 4 are very common, and they typically combine attributes of anoxic deposition (lamination and trace-metal enrichment) and dysaerobic deposition (presence of low-diversity benthic fossils); they undoubtedly formed in the variably oxygenated conditions described above.

Oceanic Anoxic Events

The modern oceans are extremely well ventilated, and deep waters are everywhere supplied with high levels of oxygen, thanks primarily to a vigorous thermohaline circulation regime. It therefore came as

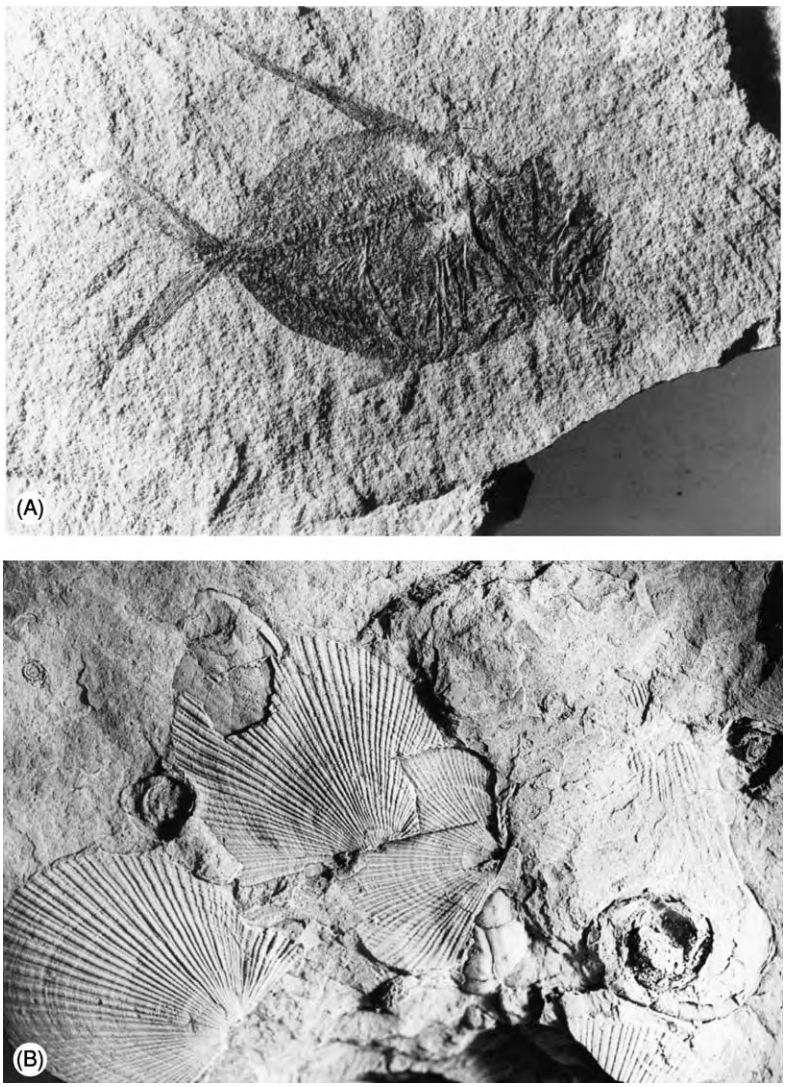


Figure 3 Fossils from ancient anoxic environments. (A) Well preserved skeleton of the deep bodied fish *Dorypterus*, from the Upper Permian Marl Slate of north east England. The lack of predation in anoxic environments often favours the preservation of fragile skeletons such as this magnificent example. The fish is 10 cm long. (B) Typical dysaerobic bivalve assemblage dominated by a single species of the radially ribbed bivalve *Dunbarella* from the *Reticuloceras subreticulatum* zone, Upper Carboniferous, Fisherstreet Bay, County Clare, western Ireland. Flat valved bivalve morphologies such as this are common in black shales from the Devonian to the Jurassic. The circular impressions on this slab are poorly preserved specimens of the zonal goniatite *R. subreticulatum*. The left hand bivalve is 3 cm high. Both specimens were coated in ammonium chloride for photography.

something of a surprise in the 1970s when the cores of the Deep Sea Drilling Program revealed the presence of thin black-shale layers sandwiched between the more normal oxygenated deep-ocean sediments. Many of these oceanic black shales were found to be contemporaneous with black-shale horizons encountered in shelf sections. Based on these observations, the concept of global oceanic anoxic events (OAEs) was proposed. Several are known from the Cretaceous, with the Bonarelli Event (named after a shale in Italy) providing the best-known and best-studied example. This event occurred at the transition between the Cenomanian and Turonian stages in the Late Cretaceous and saw anoxic or oxygen-poor waters develop over a broad area of the globe and over a broad range of water depths from the ocean floor in the Atlantic to many deep shelf sections. Ample evidence indicates that this event and probably most other OAEs coincided with phases of extreme global warmth (the Bonarelli Event may coincide with the peak of Cretaceous warmth). Many models for the origin of OAEs are therefore linked with this climatic observation. Potentially, warming of waters in high latitudes can shut down the supply of cold

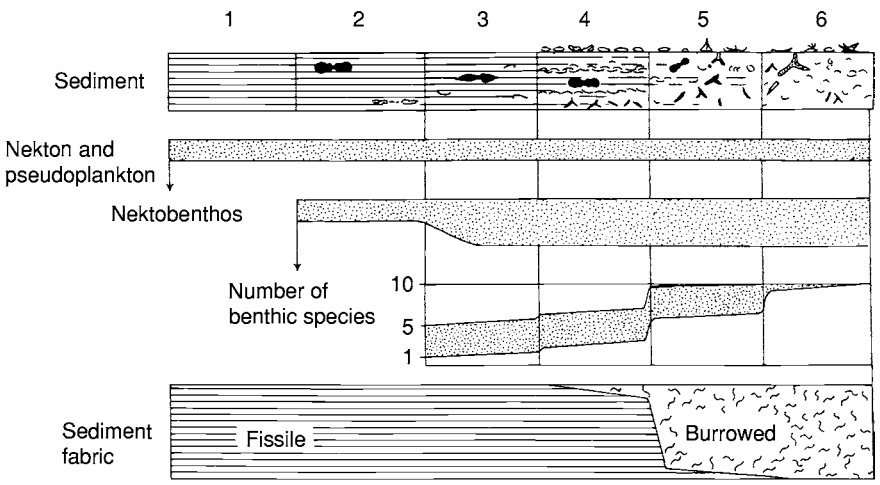


Figure 4 Oxygen restricted biofacies based on fossil content and sediment properties. The fossils have been divided into three categories. First, nekton are free swimming animals and pseudoplankton are forms that attach to floating objects, such as driftwood, in the surface waters. Such species are not affected by bottom water oxygen levels and so are found in a broad range of sediments. Second, nektobenthos are swimming forms, such as ammonites, that probably lived near the seafloor and so were absent from environments developing bottom water anoxia. Third, benthic species live on the seafloor and cannot tolerate bottom water anoxia. Reproduced from Wignall PB (1994) *Black Shales*. Oxford Monographs in Geology and Geophysics. Oxford: Oxford University Press.

dense water that drives much of the present-day oceanic circulation. Oceanic deep waters are therefore envisaged to have been replaced by dense, warm, saline waters generated in evaporitic tropical-shelf seas. Warm water holds considerably less dissolved oxygen than colder water, and this factor alone will tend to encourage deep-sea anoxia. Some modelling experiments suggest that an ocean with warm saline deep waters will circulate more rapidly than one with colder waters, with the result that upwelling may have been more vigorous. This in turn would stimulate more plankton productivity, thereby intensifying the mid-water OMZ. Intense oceanic volcanism, which is also correlated with the OAEs, may also have supplied increased levels of nutrients to the oceans and fostered further productivity.

The Cretaceous OAEs were originally intended to denote intervals when oxygen-poor marine deposition was widespread, without implying that all oceans, beneath the surface waters, were simultaneously anoxic. In essence they refer to intervals when anoxic environments were very widespread. However, there is evidence to suggest that true global oceanic anoxia may have in fact happened, with the most notable example occurring at the transition from the Permian to the Triassic. Appropriately enough, this has been termed a superanoxic event, and it is probably no coincidence that this interval is also marked by the greatest marine mass extinction of all time (*see* **Palaeozoic:** End Permian Extinctions). Like the Bonarelli Event, the Permo-Triassic superanoxia coincides with an extreme greenhouse climate and with the eruption of a giant volcanic province. In fact, the marine anoxia–global warming–massive volcanism triumvirate is seen several times in Phanerozoic history, and invariably coincides with extinction events, although of considerably different magnitudes (the Bonarelli Event is marked by only a minor extinction event). The Early Jurassic provides another classic example, with volcanism in the Karoo region of South Africa being correlated with a significant marine extinction event and widespread deposition of black shales, particularly in north-west Europe, where they are variously known as the Jet Rock (England), Schistes Cartons (France), and the Posidonienschiefer (Germany). Linking these various phenomena into a cause-and-effect scenario is a key goal of much current geological research.

Productivity versus Preservation

All black shales appear to have formed in anoxic environments, and black shales are the source all the world's oil and much of its natural gas; therefore understanding and predicting the occurrence of anoxic environments is a key goal in hydrocarbon exploration. However, it is remarkably difficult to constrain the key attributes of anoxic-environment development. As noted above, study of modern environments indicates that there are two routes to producing organic-rich sediments, and they are distinctly different. In high-productivity settings, supplied with abundant nutrients, the seafloor is overwhelmed by the flux of organic matter, and there is insufficient oxygen to decay it all. As a consequence oxygen-poor or anoxic conditions develop, but this is merely due to

the abundance of organic matter. The low oxygen levels do not cause the organic enrichment. Herein lies the key difference between the productivity and preservation models. In the latter case, the development of anoxic conditions in, for example, a silled basin with a positive water balance causes enhanced preservation of organic matter because less is lost to efficient degradation by oxygen-consuming microbes. The two models also differ in terms of their water circulation: high-productivity models require a good supply of nutrients to the surface waters from, for example, vigorous upwelling, whereas high-preservation models imply stagnation or, at best, sluggish circulation.

Distinguishing between the two alternatives is considerably more difficult than it might at first appear. Both models produce anoxic conditions; therefore merely proving that anoxia occurred is not particularly useful. Similarly, high levels of organic carbon in the sediment cannot be held as evidence of originally high organic-matter input from the water column because these high levels can be achieved by the high-productivity or the high-preservation route. Indirect proxies of organic-matter flux to the sediments have been developed, but these are contentious. Barium occurs as barite in trace concentrations in organic matter. Therefore the flux of barite to the sediment is held as a proxy for organic productivity. This assumption has the advantage that barite is considered to be less labile than organic matter. Therefore, whilst only a few percent of the organic carbon settling on the seafloor is likely to be preserved within the buried sediment (even beneath upwelling zones), perhaps the majority of the barite flux is preserved. However, the total barium content of the sediments also includes a detrital component, which has to be removed from the total to leave the biogenic component. Unfortunately, this detrital barium concentration can only be guessed at (it is controlled by factors such as sediment grain size and sediment source). A typical barium content for sediments is often assumed, but this is a rather unsatisfactory approach. Furthermore, during early sediment burial it appears that barium is often lost owing to dissolution with the result that maybe less than half the flux is preserved in the sediment. Not surprisingly, barium contents have been used successfully (i.e. convincingly) to calculate productivity levels only in young sediments not subject to significant burial.

Biogenic silica produced by certain types of plankton, particularly diatoms and radiolarians, also has a higher preservation potential during sedimentation and burial than organic matter, with perhaps in excess of 90% surviving to be incorporated into the geological record. Therefore, concentrations of biogenic silica have been used as a proxy for organic-carbon productivity, and it is significant that many organic-rich sediments forming beneath modern upwelling zones also have high levels of biogenic silica. This is primarily thanks to the abundance of diatoms in upwelling settings. The problem for geologists is that diatoms have become a common component of plankton populations only in the past 50 Ma, whilst several other factors also control siliceous plankton productivity, not least the water temperature. High-latitude plankton populations have proportionately more siliceous plankton than lower-latitude ones, and unravelling this cold-water constraint from productivity constraints in ancient sediments is extremely difficult.

Many sediments accumulating beneath high-productivity upwelling zones are rich in phosphates (much of this material is derived from fish bones sourced by the prolific fisheries encountered in these productive waters), and many phosphatic black shales are therefore assumed also to have formed beneath ancient upwelling zones. The Phosphoria Formation, a thick Upper Permian rock unit seen in Idaho and adjacent states in the western USA, is a classic example of a black shale that probably formed beneath an upwelling zone. However, the simple relationship between phosphate content and productivity is confounded by the fact that phosphate-rich accumulation also develops under conditions of very low sedimentation. This is because, in the absence of any other form of sediment accumulation, biogenic phosphate is often one of the few sources of sediment. Black shales are classically very slowly accumulated sediments.

It is something of a paradox for geologists that, whereas it is very easy to identify ancient anoxic environments, it is one of the most difficult endeavours to determine the cause of this oxygen deficiency. The recent history of the eastern Mediterranean provides a classic example of this problem. Organic-rich sediments, called sapropels, have formed in the deepest waters of this region within the recent geological past. The youngest sapropel, labelled S1, ceased forming only around 6000 years ago, almost within historical times. Unfortunately, early civilizations have left little clue to contemporaneous water conditions in the eastern Mediterranean, and evaluating the cause of the anoxia has proved almost as hard as for more ancient examples. All the evidence indicates that the eastern Mediterranean water column became truly anoxic during sapropel deposition. These periods of deposition appear to coincide with glacial intervals, when increased rainfall in the Nile's hinterland

caused increased run-off into the Mediterranean. The consequences of this were probably two-fold. First, it may have generated a positive water balance, like that in the present-day Black Sea, with low-density lowered-salinity surface waters lying on top of denser saltier deep waters. This would have resulted in the lower water column becoming relatively isolated from surface waters and more prone to oxygen deficiency. Therefore, the sapropels could be a product of increased preservation. Second, the increased run-off would also have increased the nutrient flux into the eastern Mediterranean and stimulated higher plankton productivity. Therefore, the productivity model is generally favoured by sapropel workers. However, the most recent research suggests that cause-and-effect relationships are not always clear cut when it comes to understanding sapropel generation. In some instances, anoxic deposition appears to predate slightly the onset of organic-rich sediment accumulation. This is not the correct way around if the anoxia is caused by the increased flux of organic matter into the bottom waters. Within some of the thicker sapropels, palaeoproductivity indices show substantial fluctuations from high to low levels, which again suggests that productivity is somehow decoupled from the oxygenation history. Finally, some of the older sapropels are developed in the western as well as the eastern Mediterranean, a long way removed from the regional effects of the Nile's discharge.

See Also

Fossil Invertebrates: Brachiopods; Bivalves; Ammonites. **Fossil Vertebrates:** Fish. **Mesozoic:** Jurassic; Cretaceous. **Minerals:** Sulphides. **Palaeozoic:** End Permian Extinctions. **Petroleum Geology:** Overview.

Further Reading

Fenchel T and Finlay BJ (1995) *Ecology and Evolution in Anoxic Worlds*. Oxford Series in Ecology and Evolution. Oxford: Oxford University Press.

Jenkyns HC (1980) Cretaceous anoxic events: from contin ents to oceans. *Journal of the Geological Society* 137: 171 188.

McManus J, Berelson WM, Klinkhammer GP, *et al.* (1999) Geochemistry of barium in marine sediments: impli cations for its use as a paleoproxy. *Geochimica et Cosmochimica Acta* 62: 3453 3473.

Meyers PA and Negri A (eds.) (2003) Paleoclimatic and paleoceanographic records in Mediterranean sapropels and Mesozoic black shales. *Palaeogeography Palaeocli matology Palaeoecology* 190.

Rhoads DC and Morse JW (1971) Evolutionary and eco logic significance of oxygen deficient marine basins. *Lethaia* 4: 413 428.

Wignall PB (1994) *Black Shales*. Oxford Monographs in Geology and Geophysics. Oxford: Oxford University Press.

Wilkin RT and Barnes HL (1997) Formation processes of framboidal pyrite. *Geochimica et Cosmochimica Acta* 61: 323 339.

Carbonate Shorelines and Shelves

D W J Bosence, Royal Holloway, University of London, Egham, UK

Introduction

Shorelines and shelves that accumulate sediment composed of calcium carbonate form a large, but not major, part of the world's continental-shelf seas. The best-known examples of these sediments are tropical coral reefs and the white sandy beaches of tropical and temperate coasts (Figure 1).

Most carbonate sediments are formed by the accumulation of skeletons and shells constructed by marine organisms through the precipitation of calcium carbonate (e.g. corals, molluscs, and foraminifera). These are generally known as skeletal, or bioclastic, carbonate sediments. Skeletal sediments occur in both the warm and the cold waters of the world's shelf areas (Figure 2). They are commonest on those continental shelves where they are not diluted by large amounts of eroded and transported siliciclastic sediment, as is the case near major river mouths and mountainous coastal regions. The composition of skeletal carbonates varies because of the different ecologies of the organisms that build skeletons and shells. Tropical waters (Figure 1A and B) are characterized by seafloor communities of corals (including shallow-water coral reefs), green calcareous algae (seaweed), molluscs, and foraminifera. Cold-water carbonate-forming environments (Figure 1C) are dominated by molluscs, bryozoans, red calcareous algae, and deep-water coral reefs. Other carbonate sediments are formed by the direct precipitation of crystals of calcium carbonate from seawater in the warm saline waters of some tropical seas. This

Figure 1 Examples of commonly seen environments of carbonate shorelines and shelves. (A) Coral reef on the Great Barrier Reef, Australia, constructed by the prolific growth of calcium carbonate secreting corals (coral colonies in the foreground are approximately 1 m across). (B) Tropical beach in southern Java composed of sand derived from the breakdown of the carbonate skeletons of organisms living in nearby coral reef communities. (C) Carbonate beach from the cold temperate climate of Connemara, western Ireland, composed of the skeletal debris of molluscs, coralline algae, bryozoans etc. that live in these shallow coastal waters.

process forms calcium carbonate mud and spherical sand-sized grains known as ooids (see below). These are called non-skeletal carbonate sediments and are found only in warm-water areas (Figure 2).

Shelf seas are a major environment of carbonate-sediment formation owing to the environmental conditions, which favour both the organisms that precipitate calcium carbonate skeletons and the direct precipitation of calcium carbonate from seawater. For skeletal carbonates to be produced, most organisms require well-lit oxygenated marine waters of normal salinity for the energy-consuming process of precipitating a skeleton. For non-skeletal carbonates, the chemical conditions must favour the precipitation of calcium carbonate from seawater; this occurs under conditions of increased temperature and salinity and reduced levels of dissolved carbon dioxide. All these conditions favour the faster accumulation of carbonate sediments in shallow tropical waters than in cooler temperate waters (Figure 2).

Tropical shelves typically reach rates of accumulation on the seafloor of about 1 m year ^{3} for both skeletal and non-skeletal carbonates. Cold-water shelf environments accumulate only skeletal carbonates, and these accumulate at generally lower rates (0.1–1 m year ^{3}) than their warmer-water counterparts. Both these rates of sediment accumulation are several orders of magnitude higher than rates of continental-margin subsidence (0.01–0.1 m year ^{3}), which means that carbonate sediments generally infill shallow-water areas and build up to sea-level. In ancient limestone successions this is often demonstrated by the occurrence of hundreds of metres of vertically stacked limestone beds that were all deposited in water depths of a few meters. Because the shallow-water areas are infilled, any additional carbonate accumulation must then take place in deeper more offshore regions. By this mechanism the carbonate shelf builds outwards into deeper waters, or progrades. This is well illustrated by the seismic section of the carbonate shelf area of Great Bahama Bank in the Caribbean (Figure 3), where carbonate sediments have accumulated during the last 10 Ma and are still forming today. The sedimentary layers imaged in the seismic profile (*see* **Seismic Surveys**) show how the sediments have built up to the sea surface on the Bahama Bank to the east-north-east and are building out (prograding) into the deeper waters of the Santaren Channel to the west-south-west. Thick accumulations of shelf carbonate sediments with this characteristic flat-topped morphology are known as carbonate platforms, and these are the major sites of shelf carbonate accumulation both today (Figures 1 and 3) and in the geological past (Figures 4 and 5).

The term carbonate platform is both a stratigraphical term for thick deposits of shallow-water carbonate rocks and used to describe the following carbonate shelf morphologies (Figure 4). The terms carbonate platform and carbonate shelf are often used synonymously.

- Attached carbonate platforms are carbonate shelves that are tied to a nearby continental landmass, which may mean that at times land-derived siliciclastic sediments are mixed in with the carbonate sediments.

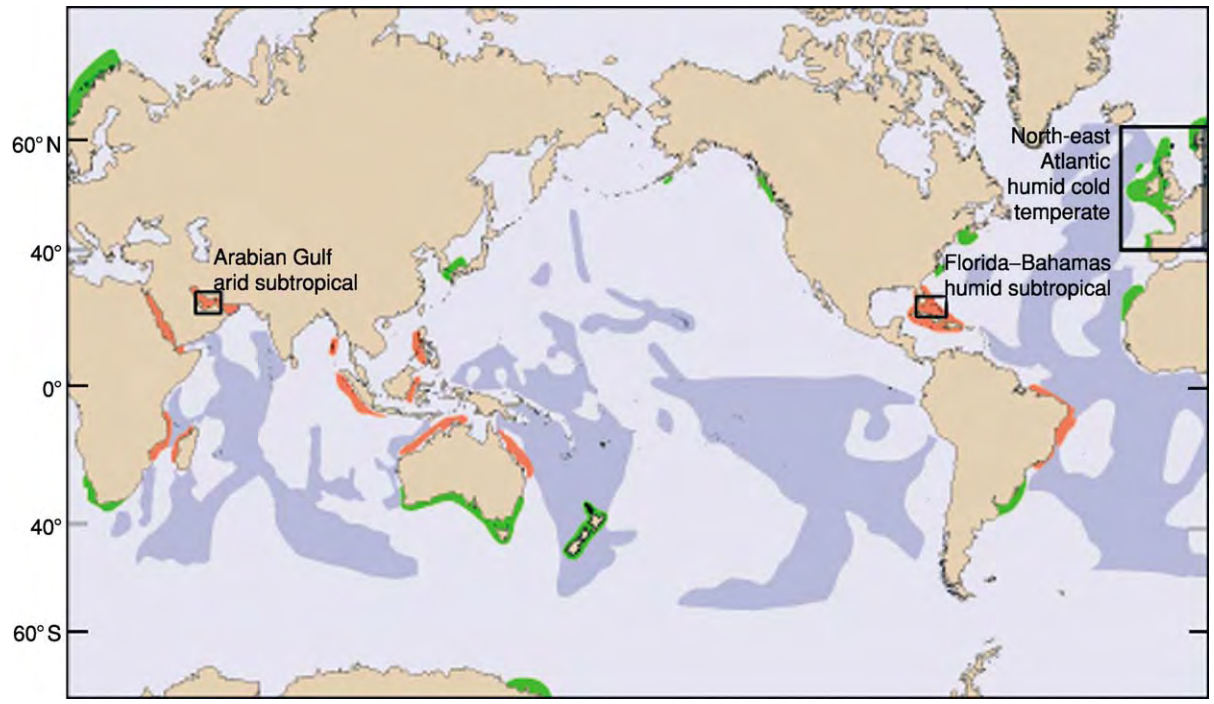


Figure 2 The distribution of calcium carbonate rich sediments in the world's oceans, and the locations of the shelf areas described in this article. Red, warm water shelf carbonates; green, cool water shelf carbonates; dark blue, pelagic carbonates; brown, land; pale blue, ocean. (After Bosence DWJ and Wilson RCL (2003) Carbonate depositional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.)

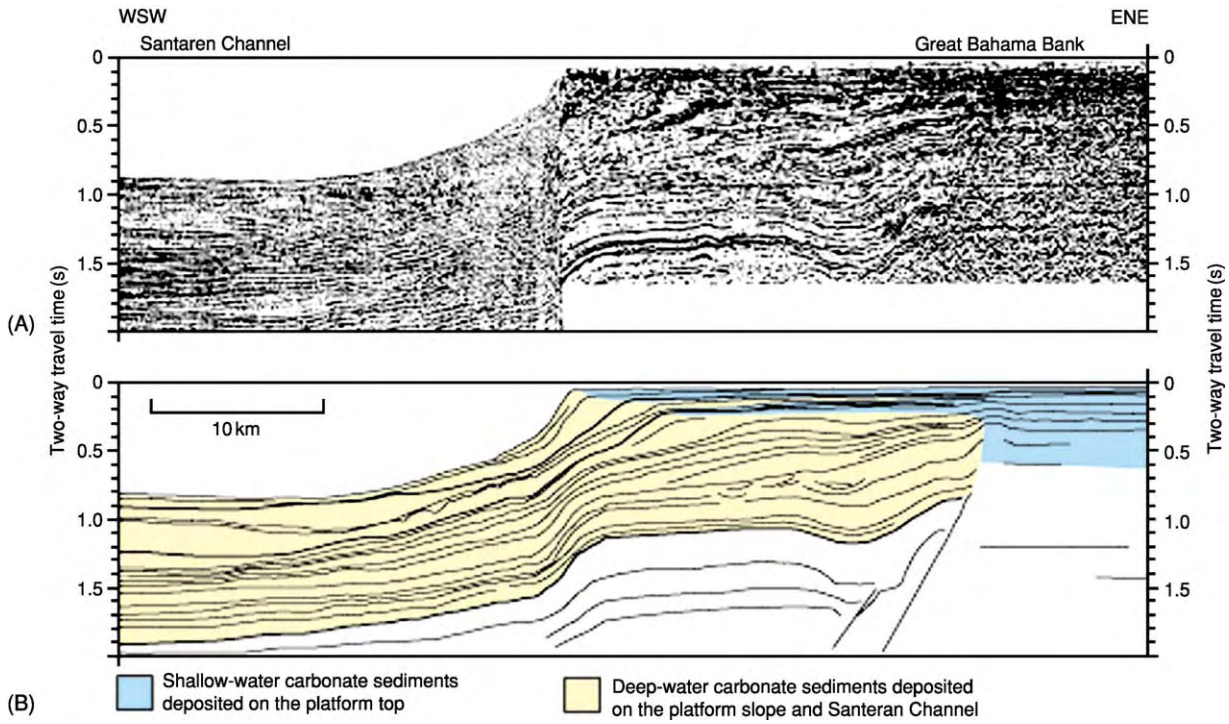


Figure 3 Seismic section through the carbonate shelf of the Great Bahama Bank (for location of section see **Figures 2** and **6**). (A) Seismic reflectors within the bank. These are where sound waves, sent from a source towed behind a boat, bounce off the sediment layers within the bank. (B) An interpretation of these layers, which are all carbonate sediment. The horizontal (blue) layers were deposited on the platform top and the inclined (yellow) layers were deposited on the platform slope and channel. Note how the reflectors preserve the morphology of the platform back through time. (Reproduced from Anselmetti FS, Eberli GP, and Ding Z D (2000) From the Great Bahama Banks into the Strait of Florida: a margin architecture controlled by sea level fluctuations and ocean currents. *Geological Society of America Bulletin* 112: 829 844.)

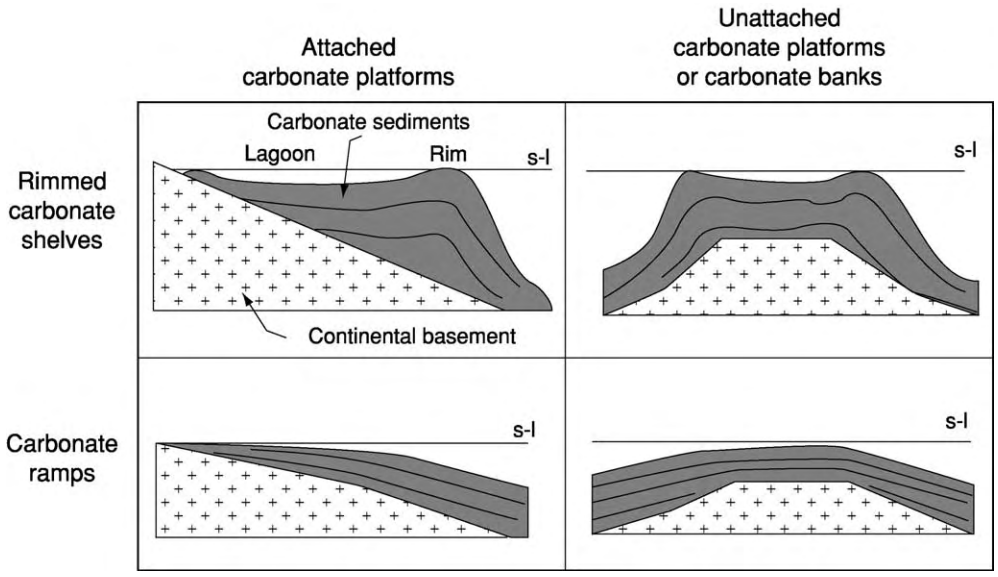


Figure 4 Morphological classification of carbonate shelves and platforms; s l, sea level.

- An isolated platform (sometimes also referred to as a carbonate bank, e.g. Great Bahama Bank) is isolated from continental landmasses (Figure 4), and an atoll is a small isolated platform that is commonly formed over a subsiding volcano.

Attached and isolated platforms may have different types of morphological profile, which fundamentally affect the accumulation of sediment on the shelf.

- A rimmed carbonate platform has a shelf margin rim or barrier, such as a reef or sand shoal, that partially isolates an inner platform or lagoon (Figure 4).
- A carbonate ramp is a platform or shelf that is gently inclined (<1°) towards an open sea or ocean with no major build-up of reefs or steep slopes (Figure 4).

When deposited in ancient shelf seas and lithified over time, carbonate reefs and sediments become limestones (*see* **Sedimentary Rocks:** Limestones) and dolomites (*see* **Sedimentary Rocks:** Dolomites), which are known collectively as carbonate rocks. The abundance of carbonate rocks formed in shelf-sea environments has varied considerably over geological time (Figure 5). Not surprisingly they were most abundant when global sea-levels were high and large areas of continents were flooded by shallow shelf seas. This occurred several times in Earth history when there were prolonged periods with no major ice-caps covering the polar regions as there are today. This meant that sea-level was higher all over the globe. These generally warmer waters formed extensive shelf seas in the past and were ideal environments for

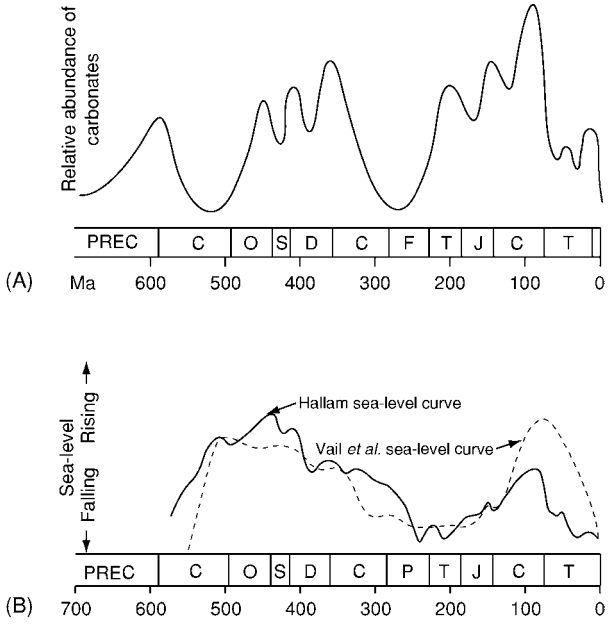


Figure 5 (A) The variation in the abundance of carbonate rocks through time, and (B) how this correlates positively with global sea level through time. (After Tucker ME and Wright VP (1990) Carbonate Sedimentology. Oxford: Blackwell Scientific Publications.)

the formation of carbonate sediments that later formed limestones and dolomites. Much of this article focuses on the sedimentary processes that operate on the different morphological types of carbonate platform, and the different types of sediment that accumulate on them. This information comes from studies of present-day carbonate shelves and is used

extensively in the interpretation of how and where ancient limestones and dolomites accumulated. These carbonate rocks have important economic uses as they host oil and gas reservoirs and heavy-metal accumulations; they are also widely used in building and in the pharmaceutical industry.

Principal Types of Carbonate Shelf

Attached Rimmed Carbonate Shelf

South Florida is an attached carbonate platform and shelf with a large low-lying hinterland (including the Everglades Swamps) set in a semi-humid subtropical climate (Figure 6). Carbonate sediments accumulate throughout this shelf area, and deep boreholes in the region tell us that this has been the case since Jurassic times. The platform comprises four main elements: a semi-enclosed lagoon (known as Florida Bay) that is partially isolated from open-marine waters by a chain of islands (the Florida Keys), a back-reef area of normal marine waters protected from high wave energy by the rimmed margin, the reef and sand shoal margin of the platform, and the fore-reef slope that descends into the deep waters of the Starit of Florida (Figures 6 and 7).

The semi-humid climate is seasonal, with a wet season from July to December (100–150 cm of annual precipitation). During this time there is considerable runoff of freshwater into the shallow-marine coastal environments, which can reduce salinities to 6‰ in Florida Bay (normal marine waters have on average a salinity of 34–38‰). By contrast, in the warmer conditions of spring and summer, temperatures increase and salinities of up to 60‰ are found in the restricted waters of Florida Bay. South Florida is within the north-easterly tradewind belt and, because coral reefs preferentially grow in shallow-water open-marine windward settings, rimmed shelf margins are found on the east-facing margin open to the waters of the Atlantic Ocean (Figures 6 and 7). The region has a very low tidal range, reaching 0.8 m ^{1} in open-water settings and reducing to zero in the semi-enclosed lagoons of Florida Bay. The reef areas experience normal marine salinities of 35–38‰, and water temperatures vary from 18°C to 30°C. In shallow-water settings Holocene carbonate sediments, which are locally lithified, rest on the eroded and cemented Pleistocene Key Largo Limestone (Figure 7). Many of the islands of the Florida Keys are limestones formed from reefs or carbonate sand shoals that accumulated during the last interglacial period, when the sea-level was slightly higher than it is today, and have since been cemented by diagenetic processes to form a rocky carbonate shoreline. The protected shorelines of Florida Bay are dominated by mangroves. These

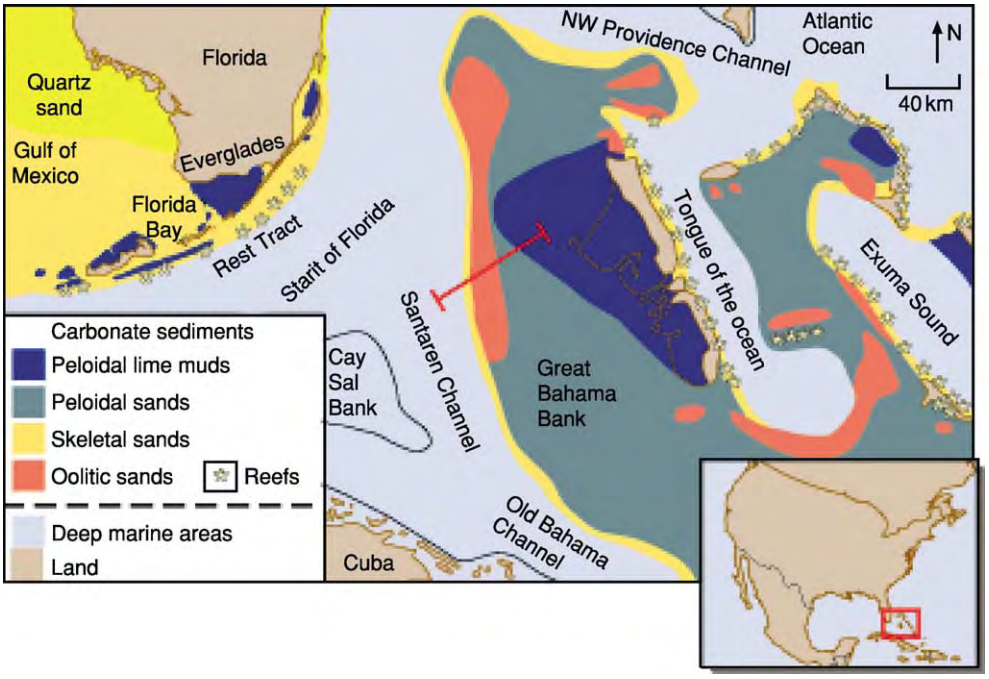


Figure 6 The distribution of major carbonate sediment types in the Florida Bahamas shelf region (for location see Figure 2). (After Jones B and Desrochers A (1992) Shallow platform carbonates. In: Walker RG and James NP (eds.) *Facies Models Response to Sea Level Change*, pp. 227 301. St Johns: Geological Association of Canada.)

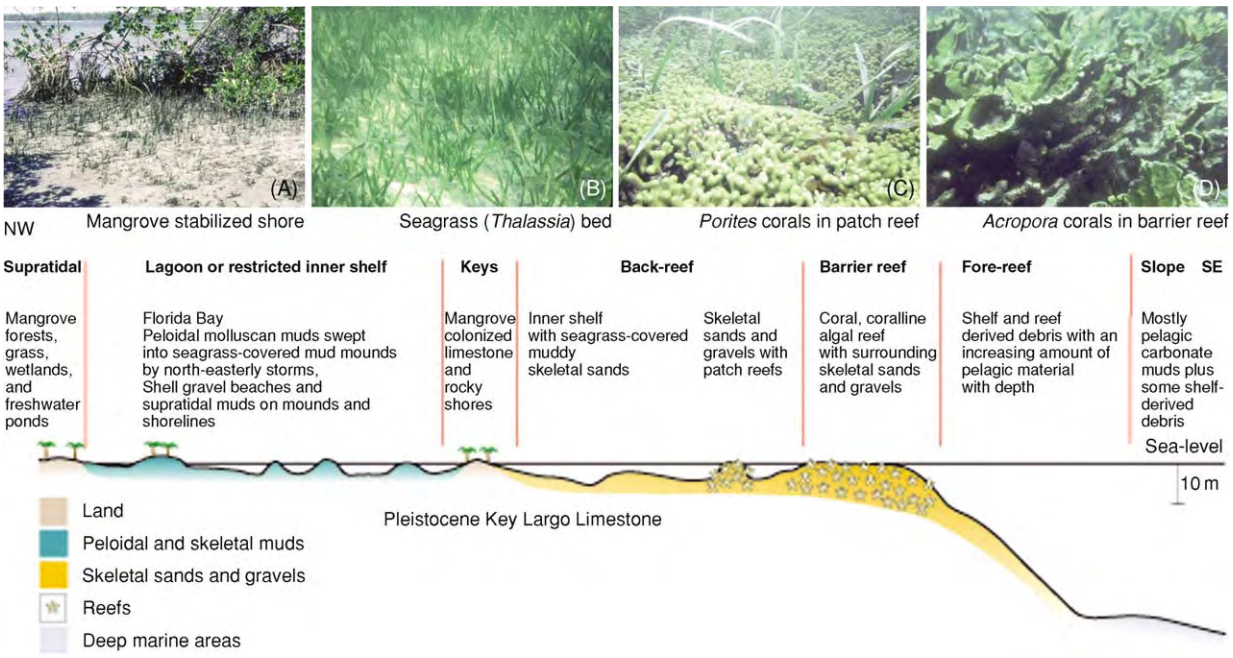


Figure 7 Cross section through the Florida shelf region normal to the coast and shelf margin, showing the main sedimentary environments. (After Bosence DWJ and Wilson RCL (2003) Carbonate depositional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.) The photographs illustrate the fundamental control that organisms have on carbonate sediment production and accumulation. (A) Mangrove roots stabilizing the shoreline of Florida Bay (foreground 4 m across). (B) *Thalassia* seagrass bed; the grass leaves (20 30 cm high) act as a substrate for carbonate secreting organisms, whose debris is trapped and stabilized within the grass beds. (C) *Porites* corals (branches are 1 cm across) are common reef builders in the back reef zone. Note also the encrusted seagrass blades. (D) The shallow zone of the shelf margin barrier reef is mainly constructed by the elk horn coral, *Acropora palmata* (branches are 15 20 cm across).

trees are adapted to live in coastal waters of variable salinity and are unusual in having aerial root systems, which have a useful sedimentological function in that they trap and bind sediment so as to resist coastal wave erosion (Figure 7A).

The occurrence of fully marine organisms in semi-isolated shallow coastal waters of Florida Bay is restricted by the variable salinities, dissolved oxygen levels, and fine carbonate sediment suspended within the bay waters. Seagrass (e.g. *Thalassia*) beds (Figure 7B) and the calcareous organisms that encrust the grass blades (foraminifera, serpulid worms, and bryozoans) are common, as are other skeletal organisms such as molluscs, foraminifera, and calcareous green algae (*Penicillus* and *Udotea*). When these organisms die their skeletal debris accumulates as shelly muds, which would form bioclastic mudstones and wackestones (*see* **Sedimentary Rocks:** Limestones) if lithified into limestones. Shells are concentrated on beaches and storm ridges as bioclastic gravels and sands. Identifying the origin of the carbonate mud is a complex problem because lime mud may be produced by the breakdown of the lightly calcified tissue of green algae, by organisms that graze or bore into rocky carbonate shorelines and skeletal carbonate material (known as bioerosion), or by chemical precipitation directly from seawater. Mud production has been extensively studied in South Florida, and mineralogical (calcite and aragonite) and geochemical (calcium, magnesium, and iron levels, etc.) analyses of the muds show that they are similar to the carbonate minerals of the skeletal organisms. This implies that most of the mud is produced by the breakdown of skeletons of algae and benthic invertebrates. Once formed, the mud is reworked by burrowing organisms (mainly shrimps and worms) that digest what they can from the mud and defecate the remainder as sedimentary faecal pellets or peloids.

Within Florida Bay the peloidal muds are swept by storm waves into shallow-water mud mounds (piles of mud that accumulate from the seafloor to sea-level), which are stabilized by seagrass (*Thalassia*) (Figure 7B). These mounds have distinct windward erosional margins on their north-east sides and accretionary *Thalassia*-stabilized leeward (south-westerly) margins. Mounds therefore migrate in a south-westerly direction towards the Gulf of Mexico. Between the mounds there may be a thin cover of bioclastic (molluscan, algal, and foraminiferan) sands and muddy sands or alternatively areas may be bare of

sediment and expose the underlying Pleistocene Key Largo Limestone.

Back-reef areas are of normal marine salinity and are more exposed to waves, which generate well-oxygenated waters. These conditions favour a more diverse community of carbonate-secreting organisms, such as corals, encrusting coralline algae, calcified plate-like green algae (*Halimeda*), molluscs, echinoderms, and foraminifera. The corals are only abundant enough to form small patch reefs, which are built up by corals and coralline algae overgrowing each other to form a rigid reef framework (Figure 7C). This framework forms an attractive habitat for many other organisms, including those that eat or bore into the coral, such as sponges, worms, parrot fishes, echinoids, and molluscs. The coral debris they generate forms much of the skeletal carbonate sand and gravel that accumulates around the patch reefs. These, if lithified, would form skeletal or bioclastic packstones or grainstones (*see* **Sedimentary Rocks:** Limestones) in the geological record.

The rimmed margin of this platform comprises a barrier reef (because the reefs are detached from the shoreline and have back-reef and lagoonal areas behind them) with intervening high-energy shallow-water shoals of coarse skeletal sand and gravel. The patchiness of the reefs arises because Florida is near the northern limit of warm-water or tropical carbonates (Figure 2) and reef growth is not as prolific as it is further south in the Caribbean, where more continuous reefs occur. The shallowest zone of the reef is formed by the branching coral *Acropora palmata* (Figure 7D). The reef is built by a series of spurs and grooves that face out into the ocean. Spurs are built by seaward growth of the branching *A. palmata* and by some round-shaped (or head) corals and encrusting coralline algae. The grooves are gravel- and pebble-floored erosional channels through the reef. Deeper down the ocean-facing front of the reef, more massive and rounded coral growth forms are found, which give way to deeper-water flattened or dish-shaped corals. Lower areas of this fore-reef slope are characterized by coarse-grained reef talus deposits formed of broken and bioeroded debris from the reef. Similar reef morphologies and coral zones are preserved in ancient reef limestones, where they are called coral boundstones and are composed mainly of fossil corals.

Attached rimmed carbonate shelves are common in the geological record. In the Oligocene and Miocene of the Mediterranean, reef-rimmed shelves built out from coastlines to landmasses generated in the Alpine Orogeny. Mesozoic examples are found in the Jurassic limestones of the Jura Mountains (France and Switzerland), and Palaeozoic examples include reef-rimmed shelves (the Capitan Limestone) in the Midland Basin (USA) of West Texas and New Mexico in North America. In the Midland Basin these carbonate shelves are buried to a depth of several kilometres and contain oil and gas reservoirs in the porous limestones and dolomites that formed originally on a Permian carbonate shelf.

Unattached Rimmed Carbonate Shelf

The Great Bahama Bank is the best-studied present-day example of a large unattached or isolated carbonate platform with a rimmed platform margin. The early research of the 1950s and 1960s concentrated on the shallow platform-top environments and sediments; later work in the 1970s and 1980s studied the deeper slopes, and more recently the internal structure is being studied with seismic surveys (e.g. Figure 3) and drilling.

The climatic setting of the Bahamas is similar to that of Florida, and they are also dominated by north-easterly tradewinds. Therefore reef-fringed margins form on the east-facing windward platform margins (Figures 6 and 8). As in South Florida, the distribution of sediments across the platform is influenced by water depth, wave and tidal energy, and variations in salinity. The waters adjacent to the Bahamas Banks are open-marine waters (36‰ salinity and 22–31°C) that descend to oceanic depths. The large shallow bank-top environments are seasonally arid, and salinities up to 46‰ are recorded. The area is microtidal, with a range of about 0.8 m at the ocean margin, diminishing in the platform interior.

Many of the bank-top environments (Figure 8) are similar to those of South Florida, but there are also some important differences. The large sheltered interior of the platform accumulates an aragonite-rich mud (unlike Florida Bay mud, which is of mixed mineralogy) whose origin has long been disputed. Current theories focus on two interpretations, which are not mutually exclusive. Surveys of the skeletal marine communities of Little Bahama Bank (just north of Providence Channel; Figure 6) indicate that most of the mud is produced by the breakdown of calcareous green algae, as is the case in Florida Bay. The other proposal is that high salinities in the bank interiors may result in direct precipitation of aragonite from seawater. This is similar to the process of lime mud production in the lagoons of the Arabian Gulf (see below). In the leeward environments of Andros Island extensive mud flats are developed, which are cut by sinuous tidal channels. The channels and the mud flats are both active sites of sediment accumulation as they build out into the peloidal muds of the platform interior (Figure 8).

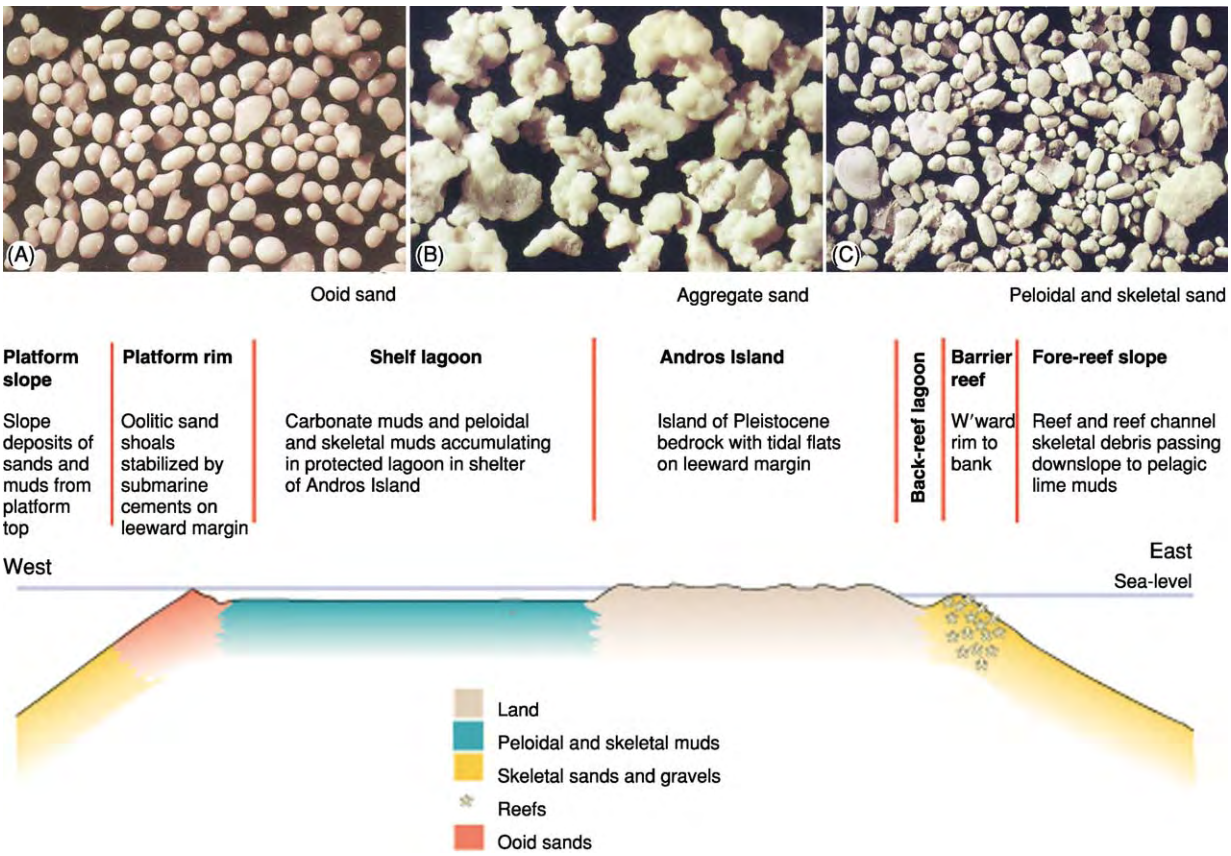


Figure 8 East west cross section through the Great Bahama Bank (for location see **Figure 6**) showing the main sedimentary environments on this rimmed platform. Photographs of sediment types (courtesy R. Till) illustrate the range of non skeletal grains on this isolated bank: (A) ooids; (B) aggregate grains; and (C) peloidal and skeletal grains (note foraminifera bottom left). All photographs are 2 cm across. (After Bosence DWJ and Wilson RCL (2003) Carbonate depositional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.)

Whilst most windward margins are reef dominated and, as such, are similar to those described above from south Florida, other margins are rimmed by sand shoals. These may be formed of oolitic, peloidal, or bioclastic sands. Aragonite precipitation onto fine sedimentary particles forms ooids. These are spherical concentrically structured grains of carbonate sand (**Figure 8A**). The precipitation of aragonite crystals to form the concentric layers is favoured by the high salinity of platform interior waters, which are swept off the banks by tidal currents, and by the degassing of carbon dioxide from deeper waters off the platform margin. Tidal currents sweep the ooid sands into large submarine dunes cut by channels and interdune areas. The sand grains may also become cemented together on the seafloor by crystal growth to form grain aggregates (**Figure 8B**). Large areas may become cemented to form what are known as hardgrounds. This early cementation helps to stabilize the sand shoals and the steep margins of the platform.

The Bahaman platforms are bounded by slopes, some almost vertical, that plunge to depths of 4 km on the ocean-facing margins. These slopes may be indented by gulleys or canyons formed by slope failure followed by abrasion by sediment gravity flows. Whilst there is observable sediment production down to around 200 m on these slopes from deep-water coral and coralline algal communities, much of the sediment on the slopes and moving downslope is of shallow-water origin. In addition some is derived from erosion on the slopes themselves, and, in areas of little sediment redeposition, accumulation from pelagic organisms is important. The pelagic component, characterized by calcitic muds derived from pelagic organisms (in contrast to the aragonitic muds generated on the platform top), also increases from the toe of the slope towards the oceanic environments of the Atlantic Ocean (**Figure 2**).

The broad pattern of sediment distribution shown in **Figure 8** has probably prevailed for many millions of years, but the leeward margin of the Bank has

prograded many kilometres westwards into the Florida Strait (Figure 3). The windward (eastern) side (not illustrated in Figure 3) has remained as a steep reef-rimmed escarpment for much of this time.

Unattached rimmed shelves or carbonate banks are known from many parts of the geological record. Tertiary examples are best known from the subsurface studies of the foundations of the Bahama Banks, as discussed above and in Figure 3. Mesozoic examples come from the ancient Tethys Ocean and now form extensive Jurassic limestones in the Betic–Alpine–Apennine mountain chains of Europe. In particular, the Jurassic limestones of the Apennines of Italy once formed large unattached platforms that were isolated from the margins of the Tethys Ocean. Palaeozoic examples come from the Silurian and Devonian inland seas of Canada. Here, communities of corals and stromatoporoids helped to build reef-rimmed margins to unattached platforms known locally as pinnacle reefs. In the subsurface these platforms are encased in shales or evaporites and contain oil and gas reservoirs hosted in the porous and permeable carbonate rocks.

Attached Carbonate Ramp in an Arid Tropical Environment

The southern coast of the Arabian Gulf offshore from the Gulf Coast States (Abu Dhabi, Dubai, and Qatar) is a fine example of a present-day carbonate ramp. The shelf slopes gradually from the low-relief desert of the coastal plain through the coastal waters down to a maximum depth of about 100 m over a distance of a couple of hundred kilometres (Figure 9). This gently sloping shelf morphology is the distinctive feature of a carbonate ramp, and there are no major reef systems or rimmed shelf margins. The prevailing winds ('Shamal') blow onshore from the north-west, which makes this a storm-wave dominated coastline (Figure 9). Owing to the restricted opening to the Indian Ocean through the Strait of Hormuz and the arid climate, salinities are elevated in the Gulf to 40–45‰ and are even higher in coastal areas, where evaporite minerals are precipitated from these waters.

The depth of the wave base, the water depth at which wave-generated currents affect the seafloor, has an important control on sediment texture in

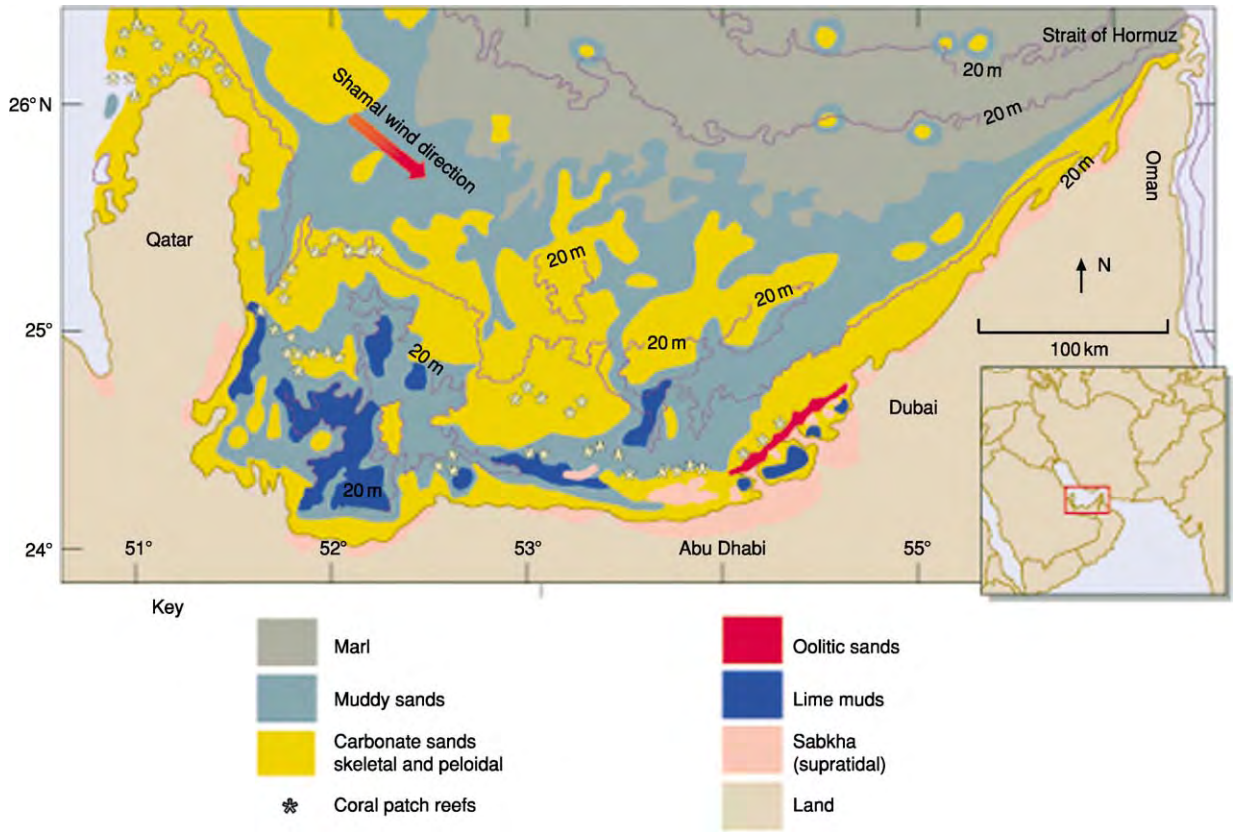


Figure 9 The south eastern region of the Arabian Gulf, showing bathymetry and the distributions of the main carbonate sediment types. (After Bosence DWJ and Wilson RCL (2003) Carbonate depositional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.)

gently sloping ramps and is used to subdivide ramps into zones (Figure 10). Above the fair-weather wave base (the depth at which waves impinge on the seafloor during normal or fair-weather conditions) is known as the inner ramp (Figure 10). The zone between the fair-weather wave base and the storm wave base (wave base for storm-generated waves) is the mid-ramp, and the outer ramp is the deep-water zone below the effect of waves.

In the Gulf, the outer-ramp areas accumulate muds with a mixed composition (called marl), which are part carbonate of pelagic origin and part siliciclastic material brought into the Gulf by rivers. These marls grade up the gentle ramp slope into muddy and then clean skeletal and oolitic sands at the fair-weather wave base at around 10–20 m (Figures 9 and 10). Areas of clean skeletal sand are found in inner-ramp environments, and these are commonly cemented on the seafloor to produce extensive hardgrounds. Small patch reefs (Figure 10) occur but the coral fauna is of low diversity because of the elevated salinities and sometimes low winter temperatures.

The inner ramp is dominated by skeletal sand beaches or ooid barrier beaches in the Abu Dhabi area (Figure 9). Aragonitic ooids are precipitated in the relatively high-salinity marine waters with a moderate tidal and wave-energy regime. Ooid barrier islands are situated up to 20 km from the shoreline, and broad shallow lagoons have developed behind the barriers (Figure 9). The lagoons have elevated salinities of 40–50‰ and so have a reduced diversity of marine faunas comprising gastropods and ostracods that occur in peloidal muddy sands. The lagoons accumulate aragonitic lime muds, and both the

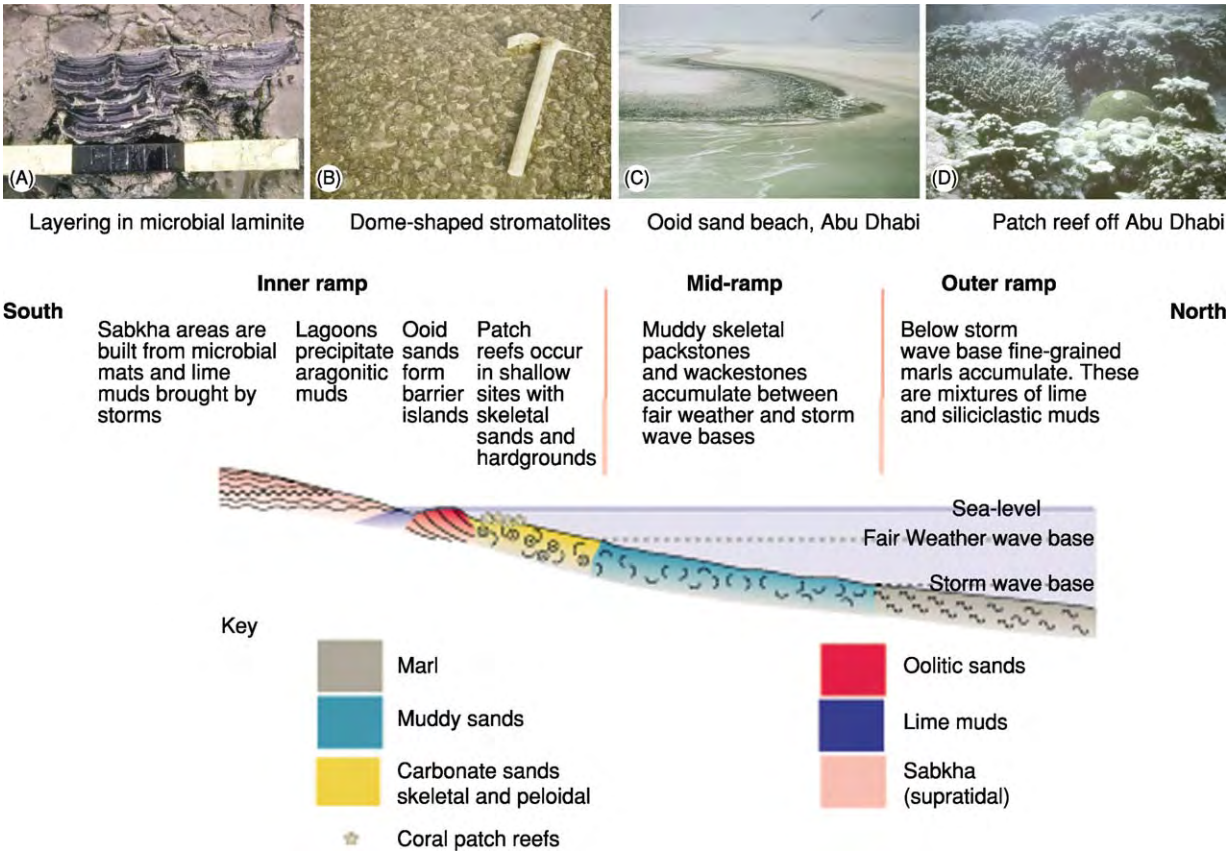


Figure 10 Cross section through the south eastern margin of the Arabian Gulf to illustrate depth related sediment types on a carbonate ramp. Note the changes in sediment type at the storm wave base and fair weather wave base. (Modified from Bosence DWJ and Wilson RCL (2003) Carbonate depositional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.) The photographs (courtesy R. Till) illustrate the diversity of inner ramp sediment types in this arid climate. (A) Vertical slice though intertidal flats, showing alternating microbial (dark) and sediment (light) layers. White patches within the layering are precipitates of the calcium sulphate salt anhydrite (scale in inches). (B) Dome shaped stromatolites in the intertidal sabkha. These result from the upward growth of microbial mats and bound sediment, with lighter coloured sediment accumulating between the domes. (C) Light coloured beaches of ooid sand occur on a string of barrier islands off the coast of Abu Dhabi. The dark area by the water's edge is microbial mat. (D) Patch reefs offshore of the barrier islands have a low diversity compared with reefs of the nearby Indian Ocean because of the high salinity of the Gulf waters.

absence of calcified green algae (such as those present in South Florida and the Bahamas) and geochemical evidence from the lagoon waters and the deposited muds indicate that these muds were precipitated chemically within the lagoon rather than having an organic origin, as is the case in Florida Bay.

The sabkhas (Arabic for area of low-lying salty ground) are broad intertidal areas and coastal plains (Figures 9 and 10) that may be flooded by lagoon waters when storm winds blow landwards. Intertidal areas have extensive microbial populations on the tidal flats growing on and within sediment deposited on the sabkhas when they are flooded by storms. These processes give rise to sediments with alternating microbial-rich and sediment-rich layers, known as stromatolites (Figure 10). High aridity in the area leads to a net evaporation of floodwaters and the drawing up of saline groundwaters into the sabkha. This leads to the precipitation of evaporite minerals, such as dolomite, gypsum, and anhydrite, within the microbial mats, including the stromatolites.

With time all these depth-related sedimentary zones will build out or prograde into the Gulf, forming a sedimentary succession that has been recognized in many ancient limestones throughout the world. Ancient carbonate ramp successions are known from the Carboniferous of north-western Europe and the USA, where they build out carbonate shelves with facies similar to those described from the present-day Arabian Gulf. However, in addition, the deep-water areas are characterized by carbonate mud mounds that do not have an exact modern-day analogue. Very extensive Jurassic and Cretaceous carbonate ramps are known from the subsurface of the eastern part of the Arabian Plate. Here they host the rich oil and gas reserves of the Middle East in prograding units of inner ramp carbonate sand shoals and patch reefs. The oil and gas is preserved in these porous carbonate rocks and sealed by overlying evaporites, which formed in sabkha environments similar to those that occur today in the southern Arabian Gulf.

Attached Carbonate Ramp in a Cold Temperate Environment

The shelf regions of the north-eastern Atlantic form a large area of cold-water shelf carbonate sediments (Figures 2 and 11). These extend from the narrow Spanish shelf to broad shelf regions within the Arctic Circle near Spitzbergen. Similar extensive areas of cold-water carbonates are found in the southern hemisphere, particularly off the south coast of Australia (Figure 2). Cold-water carbonates differ from their warm-water counterparts in that they are formed only of skeletal material: there are no precipitated grains (ooids etc.) in these cooler waters. In the north Atlantic temperatures vary from an average of 7°C in winter to 15°C in the summer, and salinity levels are relatively constant at around 35‰. Within these waters skeletal sands and gravels are formed from a variety of different seafloor communities that contrast with those of the tropics.

The shelves of north-western Europe form what are termed distally steepened ramps in that they have very gentle gradients from the shoreline to a shelf margin at around 200 m water depth (Figure 11). Here, the seafloor drops away more steeply on what is known as the continental slope. The shelf environment is subject to a number of different currents that have an important effect on sediment types and their distribution. The North Atlantic Drift is an oceanic current that flows north-eastwards along the continental slope and is strong enough to transport sand along the upper slope in large seafloor sand dunes. The shelf is also affected by westerly winds that have blown across the Atlantic Ocean and generate significant storm waves (i.e. a maximum of 30 m during a 50 year period) that can rework the seafloor sediments

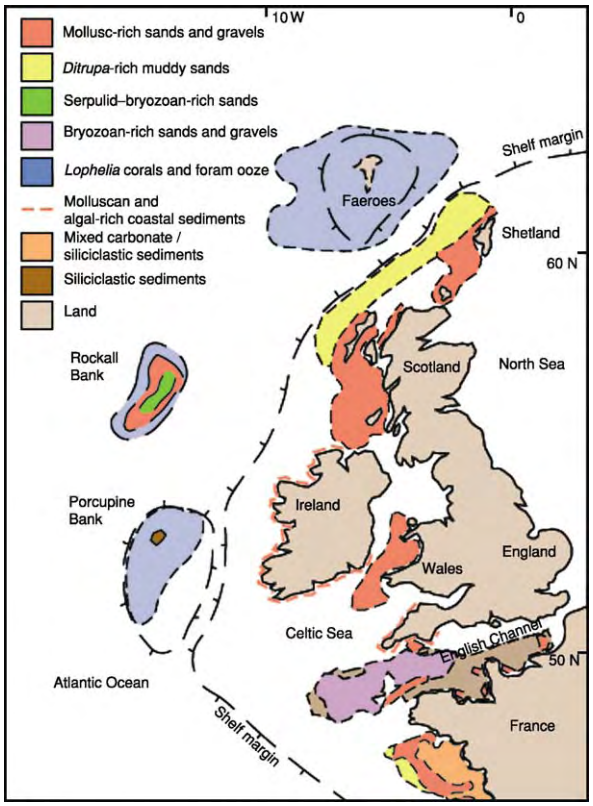


Figure 11 The north east Atlantic Ocean, showing the distribu tions of the main types of shelf and slope carbonate sediments. Note that carbonate sediments are abundant on the attached shelves of north western Europe and on the isolated platforms of the Faeroes, Rockall Bank, and Porcupine Bank. White areas of shelf region are those that have not been studied in detail.

right down to the shelf edge at 200 m depth. In addition, tidal currents are strong and can rework sediments and prevent mud deposition over much of the shelf. Therefore the entire shelf is above storm wave base and much of it is within the mid-ramp zone as the nearshore areas slope steeply off the coastal cliffs. Locally, for example in western Ireland and north-west Scotland, carbonate beaches are formed of skeletal sands and gravels (Figure 1C). By way of example, a shelf profile is shown across the shelf north of Scotland (Figure 12). Inner to mid parts of the shelf are swept by tidal currents that restrict any mud sedimentation, and these areas are characterized by mollusc-rich sands and gravels (Figure 12). Such facies are common throughout this carbonate province (Figure 11). Locally they are enriched in siliciclastic sand grains that have been reworked from nearby rocky cliffs and islands. These shell gravels not only represent the remains of a diverse molluscan-rich seafloor community but also form a substrate for encrusting calcareous organisms such as tube worms and bryozoans.

The outer part of the shelf and the shelf margin are inhabited by a quite different seafloor community dominated by an unusual free-living or non-encrusting tube worm known as *Ditrupa* (Figure 12). These conical tubes occur in great abundance (upto 1000 individuals per square metre) in the outer shelf, associated with minor contributions from molluscs, bryozoans, and solitary corals that use the *Ditrupa* tubes as a hard substrate to grow on. With increasing depth, pelagic foraminifera become more important and the sediments have a muddy texture below wave base and the shelf edge. Many areas of the deep shelf and continental slope are covered by a colonial coral known as *Lophelia* (Figure 12). Although not present on the selected transect, it is common around the unattached carbonate shelves of the Faeroes, Rockall, and Porcupine Banks of the north-east Atlantic (Figure 11) where it forms extensive thickets on the seafloor. On the Norwegian shelf it forms deep-water reefs tens of metres high and hundreds of kilometres long. Living at depths of 200–500 m, these coral thickets exist by suspension feeding. They are not

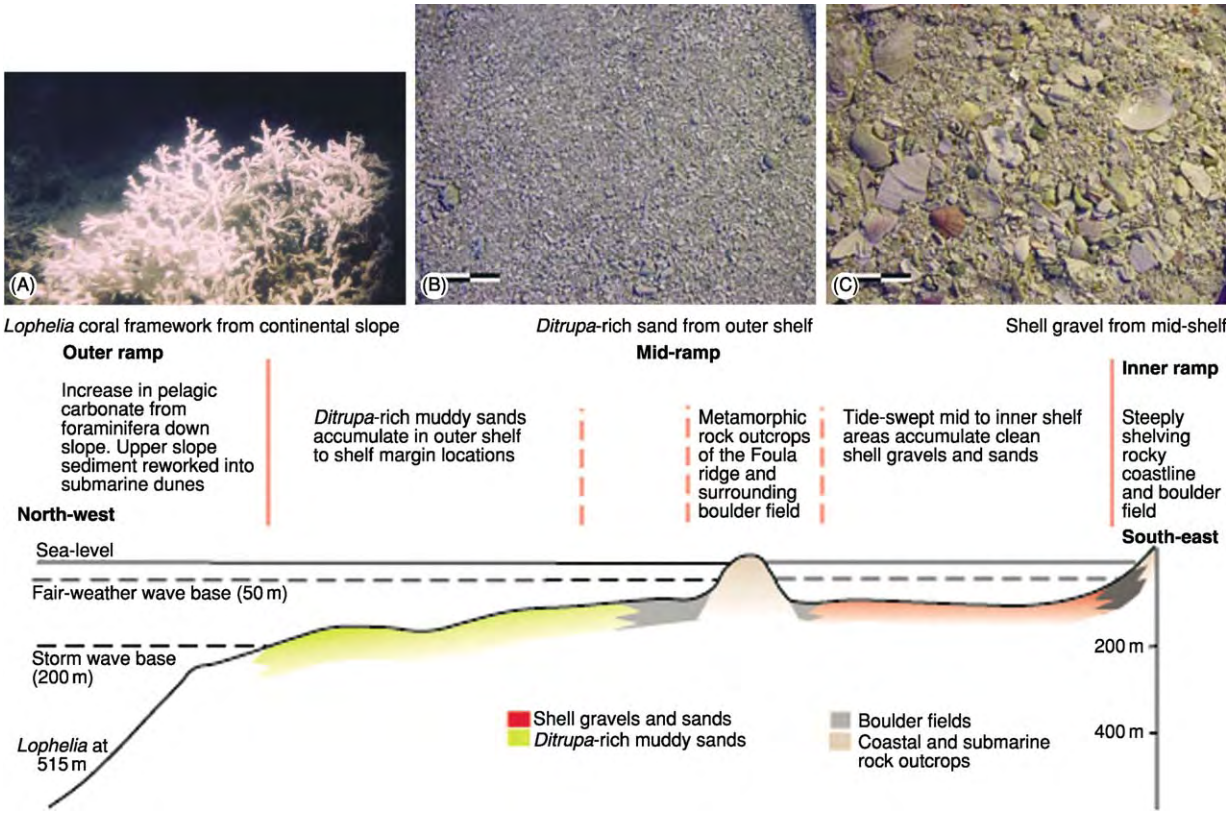


Figure 12 North west south east section across the shelf to the west of Shetland, illustrating the occurrence of the main carbonate facies with respect to shelf regions. Note the considerable depth of the fair weather and storm wave bases compared with the sheltered Gulf (Figure 10) and the importance of the rocky shoreline and islands (Foula) in this region. Photographs illustrate the skeletal nature of the shelf carbonates: (A) scale: foreground 50 cm across; (B, C) scale bars 2 cm. (After Light JL and Wilson JB (1998) Cool water carbonate deposition on the West Shetland Shelf: a modern distally steepened ramp. In: Wright VP and Burchette TP (eds.) *Carbonate Ramps*, pp. 73 105. Special Publication 149. London: Geological Society.

restricted to the photic zone like their warm shallow-water counterparts such as those from the Caribbean described above. However, they have an analogous complex associated fauna. The extent and thickness of these cold-water carbonates in the northern and southern hemispheres has become known only in the last 20 years or so, and the full geological impact of their discovery has yet to be established. Many Miocene limestones from southern Europe are now believed to be of temperate origin as are some Late Palaeozoic limestones of North America and southern Australia. Because of their more limited extent, cold-water carbonates have fewer economic uses than their tropical counterparts. In the coastal areas of north-western Europe the calcareous algae are dredged for use as agricultural and horticultural fertilizer and for liming the acidic soils of the region. However, because of the low growth rates of the carbonate producers, this is a non renewable resource.

Conclusions

From this review it can be seen that continental shelf seas have extensive areas of carbonate sediment accumulation. The nature and distribution of these areas are mainly controlled by climate and morphology. Although pelagic sediments forming in deep ocean waters are more abundant today, their usual geological fate is to disappear into subduction zones. This explains why limestones formed in shelf environments are the commonest type of limestone in the geological record. If we are to understand how limestones formed and what controls their distribution and composition, we must study their modern counterparts.

To conclude, carbonate shorelines and shelves are important features of the world's sedimentary and biological systems; they contain the biologically diverse and economically important habitat of tropical coral reefs, analogues for the interpretation of ancient limestones, and some of the most attractive coastlines of the world.

See Also

Diagenesis, Overview. **Fossil Invertebrates:** Corals and Other Cnidaria; Bivalves. **Microfossils:** Foraminifera. **Minerals:** Carbonates. **Sedimentary Environments:** Reefs ('Build-Ups'). **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Dolomites; Limestones. **Seismic Surveys**.

Further Reading

Bosence DWJ and Wilson RCL (2003) Carbonate depos itional systems. In: Coe A (ed.) *The Sedimentary Record of Sea level Change*, pp. 209 233. Milton Keynes and Cambridge: The Open University and Cambridge University Press.

Jones B and Desrochers A (1992) Shallow platform carbon ates. In: Walker RG and James NP (eds.) *Facies Models Response to Sea Level Change*, pp. 227 301. St Johns: Geological Association of Canada.

Light JL and Wilson JB (1998) Cool water carbonate de position on the West Shetland Shelf: a modern distally steepened ramp. In: Wright VP and Burchette TP (eds.) *Carbonate Ramps*, pp. 73 105. Special Publication 149. London: Geological Society.

Purser BH (ed.) (1973) *The Persian Gulf: Holocene Car bonate Sedimentation and Diagenesis in a Shallow Epicontinental Sea*. Berlin: Springer Verlag.

Tucker ME and Wright VP (1990) *Carbonate Sedimentol ogy*. Oxford: Blackwell Scientific Publications.

Wright VP and Burchette TP (1996) Shallow water car bonate sediments. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 325 394. Oxford: Blackwell Science.

Contourites

M Rebesco, Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS), Italy

Introduction

Contourites are sediments deposited or substantially reworked by the powerful action of bottom currents.

The study of contourites is important in at least three main respects: palaeoclimatology, hydrocarbon exploration, and slope stability. The continuous and relatively high-resolution record included in these sediments is crucial for palaeoclimatology since it holds the key for priceless information on the variability in circulation pattern, current velocity, oceanographic history, and basin interconnectivity. Bottom currents are a crucial factor in oil and gas reservoir development since weak

flows may favour the accumulation of source rocks and robust flows may represent a viable mechanism capable of forming 'clean' sands in the deep sea. Low-permeability, fine-grained contourites have been found recently to play a critical role in slope stability by providing potential overpressured glide planes, either when their high-water-content is rapidly loaded or when their rigid biosiliceous microfabric is collapsed by diagenesis.

Despite its significance, little is still known about this group of sediments. There are three reasons for the long-lasting disregard: the inherent elusive nature of these complex deposits, the 50-year dominance of the turbidite paradigm, and the controversy that surrounds these sediments since they were first recognized. The elusive and very subtle characteristics of these slowly and continuously accumulated sediments and their occurrence within a spectrum of deep-water deposits (Figure 1) does not allow them to be easily recognized and decoded. The monumental efforts to promote the turbidite systems, the simplicity and predictability of turbidity current concepts, and the sense of confidence given by the routine use of widely applicable models has induced the geologic community to ignore alternative, more complex deep-water models. The never-ending disputes regarding the occurrence of sandy contourites versus reworked turbidites and the early errors in recognition of fossil examples have prevented the establishment of a widely shared consensus on valid diagnostic criteria for the identification of these deposits.

The growing level of interest and research in contourites is shown by the recent publication of several special volumes dealing with such systems, covering large parts of present ocean floors and continental margins. An increasing number of fossil occurrences in ancient sediments exposed on land have also been documented. Nevertheless, wide multidisciplinary approaches and the integrated work of different international specialists (e.g., deep-water sedimentologists, seismic interpreters, physical oceanographers, and palaeoclimatic modellers) are still needed to improve our knowledge of these systems and to help tackle the problems that remain.

History

The systematic study of deep-sea sediments began at the end of the nineteenth century. At that time the ocean floor was perceived as a tranquil realm receiving only pelagic clays. The possibility that thermohaline bottom currents may influence sedimentation in the deep oceans was suggested by the German oceanographer Wust in 1936. However, the contourite concept was not accepted in marine science until the second half of the 1960s after the American team of Heezen and Hollister had provided geological and oceanographic evidence of this process along the eastern North American continental margin. During the next two decades standard facies models for contourite sediments were developed from coring contourite accumulations. Other important steps forward were achieved by addressing the link between current strength and grain size and by confronting the problem of distinguishing contourites from other deep-sea facies like hemipelagites and fine-grained turbidites. Among the projects that have provided significant contributions to contourite research have been the HEBBLE

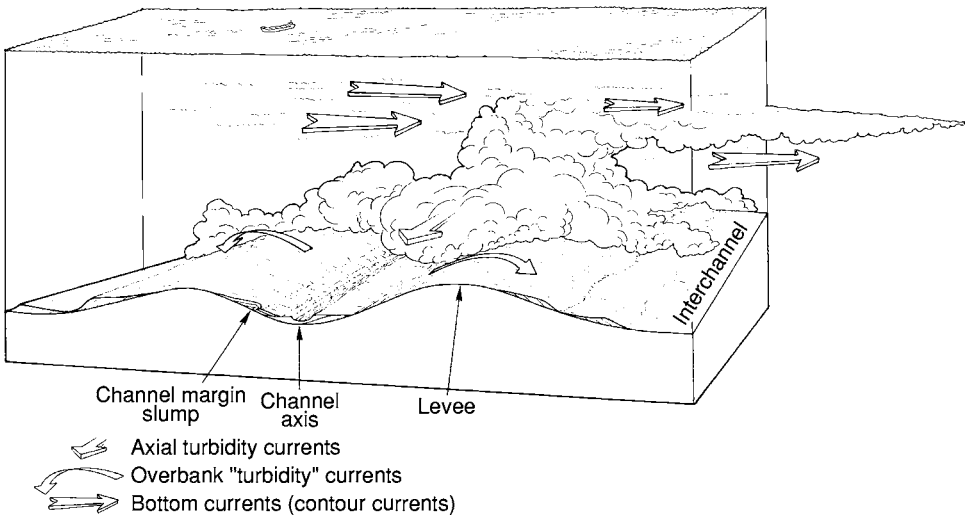


Figure 1 Schematic model of downslope gravity driven currents and alongslope bottom currents interacting within the continuous spectrum of deep water processes. Reprinted from Shanmugam (2000) 50 years of the turbidite paradigm (1950s 1990s): Deep water processes and facies models A critical perspective, *Marine and Petroleum Geology* 17: 285 342, with permission.

(High-Energy Benthic Boundary Layer Experiment) in the Nova Scotia Rise and the international drilling programs (ODP and predecessors) in most contourite deposits of the world's oceans (Figure 2). At the very end of the 1990s the International Geological Correlation Programme (IGCP) Project 432, named 'Contourite Watch', was launched to facilitate research on contourites and bottom currents. It included the establishment of a global network of contourite workers, the organization of international workshops, and the publication of selected special volumes, including an atlas of contourite systems.

Terminology

Most workers currently agree on very broad definitions and on the use of most terms related to contourite research. The need to maintain a certain degree of flexibility to allow for development in understanding has been acknowledged; nevertheless the present state of contourite terminology usage is possibly still too loose to allow for a clear description of the products (sediments) and consequently for a clear understanding of the depositional processes.

The term 'contourite' was originally introduced to define the sediments deposited in the deep sea by contour-parallel bottom currents. The term was subsequently widened to embrace a larger spectrum of sediments that at diverse depths are affected to various extents by different types of currents. Since both the depth and direction of the bottom currents that influenced nonrecent deposits are seldom precisely identifiable, a rigorous restriction of contourite to its original sense would severely limit its application. Such a long-established term should hence be maintained for nonspecific use, and the employment of modifying terms to qualify the deposit in terms of depth, current type and action, and interacting process is recommended (see Table 1).

The term 'contour current' refers to those thermohaline-driven, deep-water currents that flow approximately parallel to the bathymetric contours. It is widely used synonymously for 'bottom current', which in contrast includes all deep currents not driven by sediment suspension that are capable of eroding, transporting, and depositing sediment on the sea floor. Although bottom current is suitably applicable to different types of currents flowing alongslope as well as upslope and downslope (see below), the term contour current should be applied only to currents whose flow can be recognized as being parallel to contours.

Contourites were first described as being at great depths beneath deep-water bottom currents. Subsequently, contourites and bottom current-controlled deposits were recognized in several settings including shallow gateways, outer shelf and slopes, and even lakes. The depth ranges for a three-category subdivision was hence proposed: deep water ≥ 2000 m; mid-water = 300–2000 m; shallow water = 50–300 m. However, a rigid definition of the depth confine is of limited value, especially for nonrecent and fossil contourites for which the deposition depth is largely unknown. Nevertheless, the use of a modifying term to qualify the deposit in terms of depth is recommended when a distinction can be made.

'Sediment drift' is the generic term for a sediment accumulation that was appreciably controlled by the action of a bottom current. It is often used synonymously for 'contourite drift', which in contrast should be specifically used for sediment accumulations deposited principally (though not exclusively) by contour currents. Therefore, the most appropriate term for contourite accumulations is in most cases the wide-ranging designation 'sediment drift' or simply 'drift'. The occurrence of mixed system is the norm rather than the exception with sediment drifts. For this reason, a qualifying modifier like bottom current-controlled (modified, reworked, etc.) is recommended for turbidite systems and hemipelagic deposits that were significantly affected by interaction with bottom currents. Similarly, when mounds, levees, fans, lobes, channels, and other terms closely associated with downslope systems are applied to contourite systems, they must be preceded by the term contourite.

Bottom Currents

There is a wide spectrum of bottom currents that operate in deep water. The bottom currents sensu stricto (contour currents) are those that are part of either the thermohaline- or wind-driven major circulation patterns (Figure 3). They are usually persistent for long time intervals as testified by the thick drifts that took million of years to develop. Such currents generally have an overall alongslope flow, but in detail their velocity and direction are extremely variable in both time and space. In fact they are affected by seafloor morphology (obstacles, gateways, and changes in slope direction and steepness), Coriolis force, circular motions (gyres and eddies) unrelated to contours, and eddy kinetic energy changes (seasonal and at different scales). Substantial changes in kinetic energy at the seafloor may be produced by sea-surface topographic variations, resulting in episodic high bottom current velocities, referred to as 'benthic storms'. Such storms may resuspend large amount of sediment that nourish the bottom 'nepheloid layer' (several hundred metres thick, long-living, low-concentration (50 to 100 $\mu g\,l^{-1}$) clouds of sediment particles close to the seafloor).

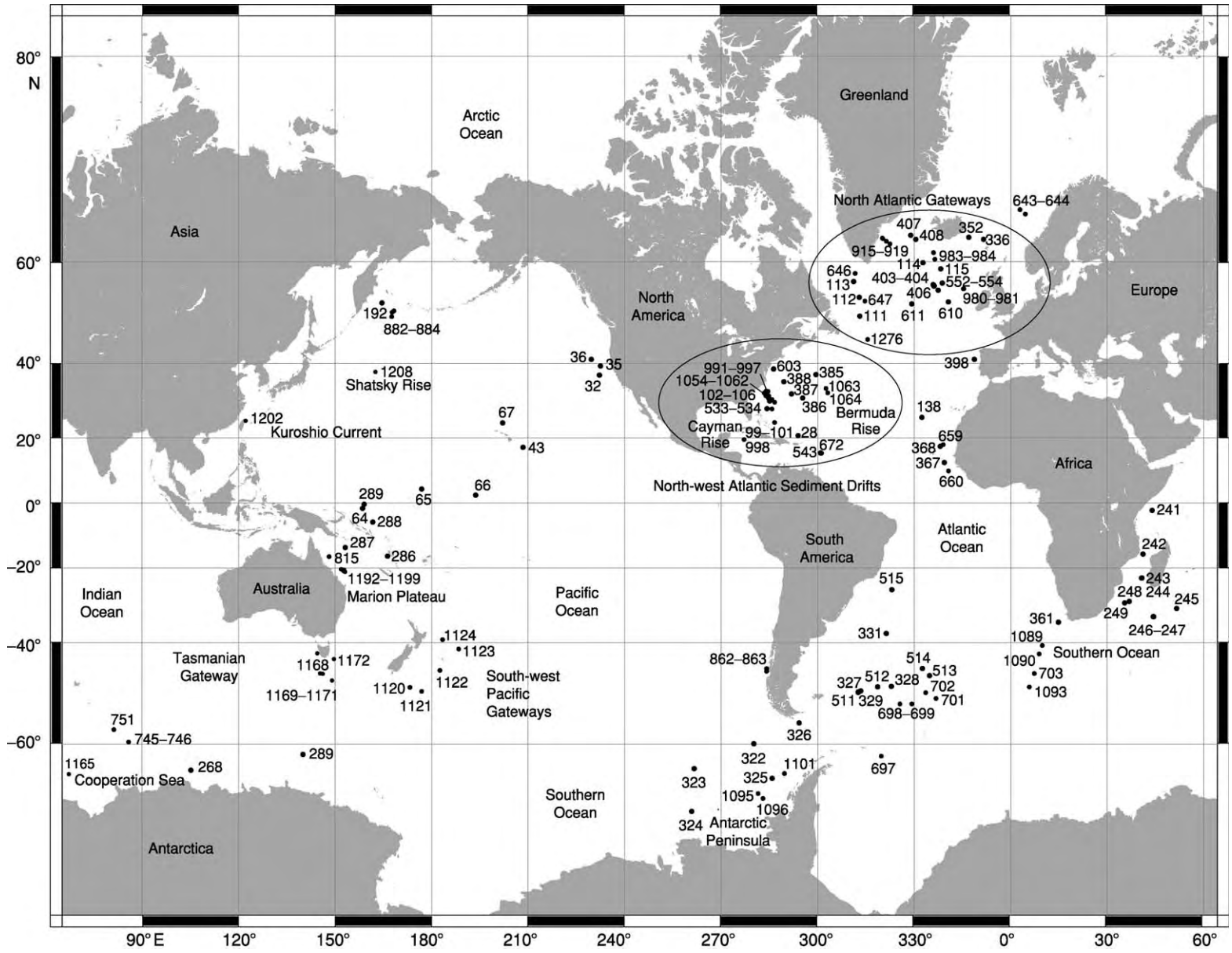


Figure 2 ODP, IPOD, and DSDP sites drilled in contourite deposits of the world's oceans. Data sources: 1998 to present from ODP Web site http://www odp.tamu.edu/sitemap/sitemap.html; and before 1998 from Stow DAV, Faugères JC, Viana A, and Gonthier E (1998) Fossil contourites: A critical review, in: Stow DAV and Faugères JC (eds.) Contourites, turbidites and process interaction, *Sedimentary Geology* 115: 3 31.

Table 1 Examples of qualifying terms for contourite related deposits

Depth	Deep water	Mid water	Shallow water	Shelf	Slope	Lacustrine
Type of current	Contour parallel	Thermohaline driven	Wind driven	Up and downwelling	Up and down canyon	Internal waves and tides
Type of action	Resuspended	Pirated	Far off transported	Condensed	Polished	Winnowed
Interacting process	Turbidite sourcing	Pelagic settling	Overbanking	Faulting	Creeping	Downslope resedimentation

N.B.: Not all kinds of terms are to be used concurrently (e.g., deep water turbidites winnowed by wind driven bottom currents, or mid water turbidite sourced contourites).

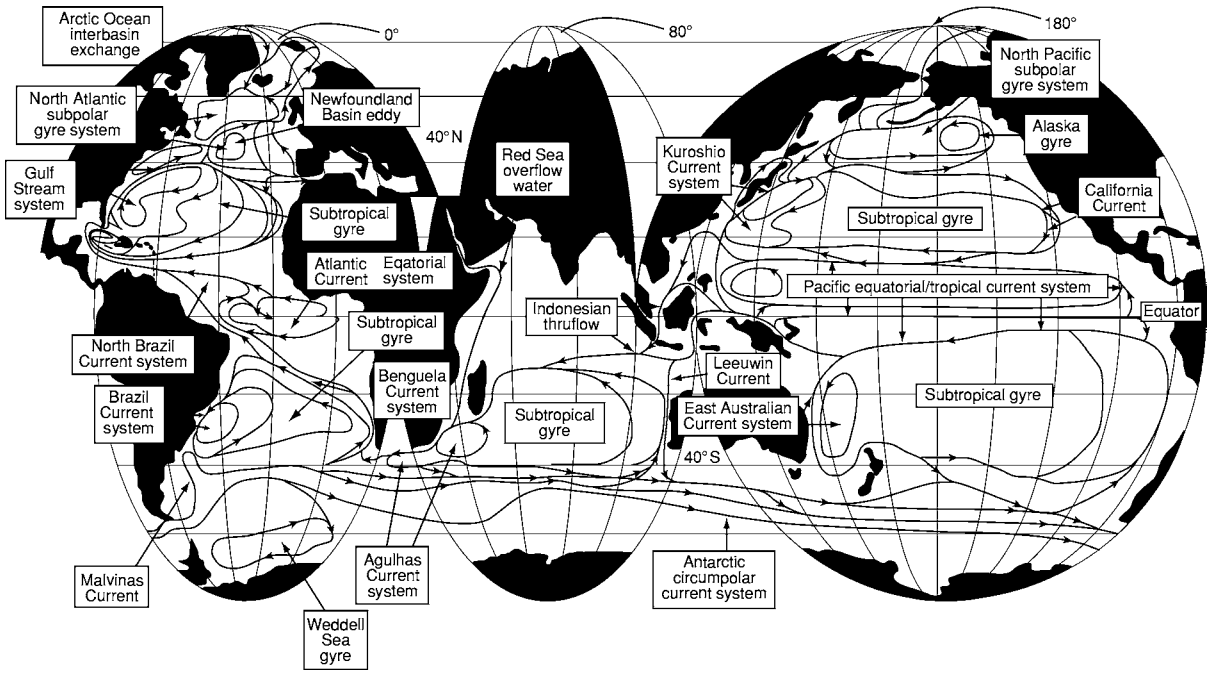


Figure 3 Global ocean circulation pattern. Reprinted from Steele JH, Turekian KK, and Thorpe SA (eds.) *Encyclopedia of Ocean Sciences*, Academic Press, London, with permission.

Additional types of deep (bottom) currents not driven by sediment suspension have been observed to flow mostly perpendicular to the slope. These include currents related to internal waves and tides (formed between subsurface water layers of varying density), canyon currents (frequently reversing flow of clear or very low-density turbid waters), and down- and upwelling flows (currents on the continental slopes that may be generated by density, wind, storms, or obstructions).

The present global thermohaline circulation of deep water is originated by the cooling and sinking of surface water masses in the polar areas. The Antarctic bottom water (AABW) forms beneath Antarctic floating ice shelves (especially in the Weddell Sea), descends the continental slope, and hence circulates clockwise in the Southern Ocean from which it escapes northwards into the world's oceans. The Arctic bottom water (ABW) forms in the Norwegian and Greenland Seas, flows intermittently south through basin gateways, and mixes with the North Atlantic bottom water and Labrador seawater. Subsequently, AABW and ABW gradually mix with other stratified water masses of their own density and circulation pattern (e.g., intermediate depth water produced by evaporation in semienclosed seas, like the Mediterranean overflow water—MOW).

The present global surface circulation is governed by dominant wind systems. In many cases the action of such currents affects the entire water column down to the seafloor. This is the case for the powerful Antarctic circumpolar current (ACC) and western boundary currents like the Gulf Stream, Brazil, and Agulhas and Kuroshio Currents.

All the above currents impinge upon the seafloor where they may erode, transport, and deposit sediment

in the form of contourites. The velocities of these bottom currents are generally very slow (few centimetres per second) and hence just able to slightly affect the action of other depositional processes (hemipelagic settling and low-density fine-grained turbidity currents). However, bottom currents accelerate when constrained by Coriolis force within western boundary undercurrents restricted against the continental slope (commonly with velocities ranging between 10 and 100 cm s $^{-1}$) or when constricted in narrow gateways (velocities may even exceed 200 cm s $^{-1}$). In many cases the bottom currents are strong enough to profoundly affect deep-water sedimentation by a number of processes including (in order of increasing current strength) the prevention of fine sediment deposition, the entrainment (pirating) and transport of suspended sediments within the nepheloid layer, the reworking of the coarse fraction introduced by episodic down-slope processes, the erosion and transport by traction of silt and fine sand, and sedimentary lag production by polishing and winnowing the seafloor. Such energetic processes take place chiefly beneath the core of the main bottom current flow. The velocity decreases away from the axis of the current, and deposition of various types of sediment drifts takes place in the relatively slack waters to the side of sediment-laden bottom currents. In most palaeocurrent reconstructions, the axis of the current is hence generally considered subparallel to the elongated direction of the drift and to its crest (if traceable). However, such a relationship between current and accumulation is in most case highly speculative since the associated time-scales are extremely dissimilar. In fact, the time required for a drift development is commonly one order of magnitude larger than the glacial–interglacial cyclicity, which is widely accepted to profoundly affect bottom-water circulation, and exceeds by over four orders the time length of our direct oceanographic observations.

Sediment Drifts

There is a large variety of contourites, ranging from those that occur closely interbedded with other deep-water facies to those that build up individually distinct bodies (mounded drifts). Due to their large dimensions, commonly ranging between a few tens and several hundred thousand square kilometres, drifts are essentially identifiable by means of seismic profiles. Although in some instances sediment drifts may be clearly recognized on the basis of seismic profiles alone according to a widely agreed-upon criteria, a rigorous identification generally requires independent supporting data. The large-scale features of the drifts (their morphology and overall geometry) are controlled by a number of interrelated factors including bathymetric setting (physiography, depth, morphology), current conditions (velocity, time length of activity, variability, frequency of changes), sediment availability (amount, type, source type, frequency, and variability of input), and the interaction with other depositional processes (in time and space, both collaboratively or antagonistically). Accumulation rates in sediment drifts are rather variable, from a few tens to hundreds metres per million years, depending on regional conditions, specific age, and evolutionary stage. However, for most drifts their average growth rate is commonly several tens of metres per million years.

Repeated attempts to categorize the essentially continuous spectrum of contourite accumulations into several types of drift have been made (Figure 4); however, the categorizations are not totally distinct such that some drifts may be classified as more than one type:

1. A ‘sheeted drift’ is characterized by a very broad low-mounded geometry. It has a very low relief (up to a few hundred metres) over a very large area (up to 1×10^6 km^2) and shows a slight decrease in thickness towards the margins. Two kinds of sheet drift may be identified:
   a. ‘abyssal sheet’, which fills basin plains whose margins trap the bottom currents within a complex gyratory circulation, and
   b. ‘slope plastered sheet’, which is coated (plastered) against continental margins where a gentle and smooth topography favours a broad nonfocussed current.
2. An ‘elongated, mounded drift’ is characterized by distinctly elongated and mounded geometry (Figure 5). This kind of drift is the most spectacular and was among the first to be identified. It may have a relief of several hundred metres above the surrounding seafloor (total thickness up to 2 km) and a very variable extent (10^3–10^5 km^2) with an elongation ratio of at least 2:1. Elongation is generally parallel to the margin, which means that the crest of the drift is parallel to the current axis. Nevertheless, elongation and progradation trends are variable, being controlled by at least three interacting factors: current system, bathymetry, and Coriolis force. Two types of mounded drift may be identified:
   a. A ‘separated drift’ is kept apart (separated) from the adjacent continental slope by a distinct moat in which the main core (axis) of the bottom current is flowing as testified by local nondeposition and/or erosion (Figure 6). Such drifts show an evident upslope progradation with a minor alongslope migration in the downstream direction.

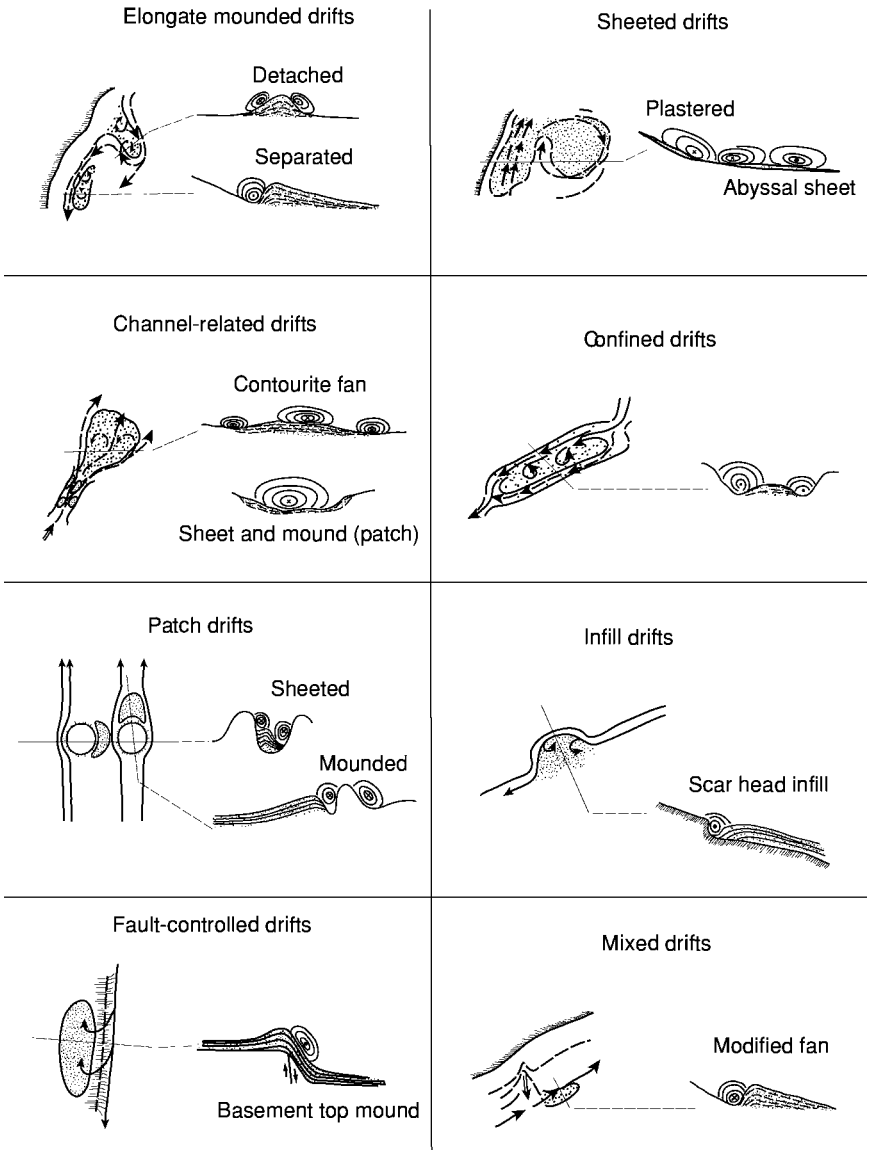


Figure 4 Sediment drift types and inferred bottom current paths. Adapted from Rebesco M and Stow DAV (2001) Seismic expression of contourites and related deposits: A preface, In: Rebesco M and Stow DAV (eds.) Seismic expression of contourites and related deposits, *Marine Geophysical Researches* 22: 303 308, with kind permission of Kluwer Academic Publishers; and: Stow DAV, Faugères JC, Howe J, Pudsey CJ, and Viana A (2002) Bottom currents, contourites and deep sea sediment drifts: Current state of the art, in: Stow DAV, Faugères JC, Howe J, Pudsey CJ, and Viana A (eds.) Deep Water Contourite Systems: Modern Drifts and Ancient Series, Seismic and Sedimentary Characteristics, *Geological Society, London, Memoir* 22: 7 20, with permission.

b. A 'detached drift' is removed (detached) from that part of the continental slope where it originally enucleated and has oppositely directed currents flowing along its two flanks. It is generated by the interaction of the bottom current with a distinct current system or with a prominent change in the slope orientation, and as a consequence in plain view it generally has an arched shape produced by the dominant downslope progradation.

3. A 'channel-related drifts' is characterized by its relationships with narrow conduits (deep channels, gateways, trenches, moats, etc.) where bottom currents are constricted and drastically accelerated. In addition to significant erosion/nondeposition, two quite distinct types of channel-related drift may develop:
   a. A drift deposited within the conduits, either as axial mounds on the floor or as lateral sheets on the flanks of the channel, is known as a 'patch drift' (though this name may cause confusion with patch drifts unrelated to channels) and 'subsidiary drift', where it occurs within the moat, separating a giant elongated drift

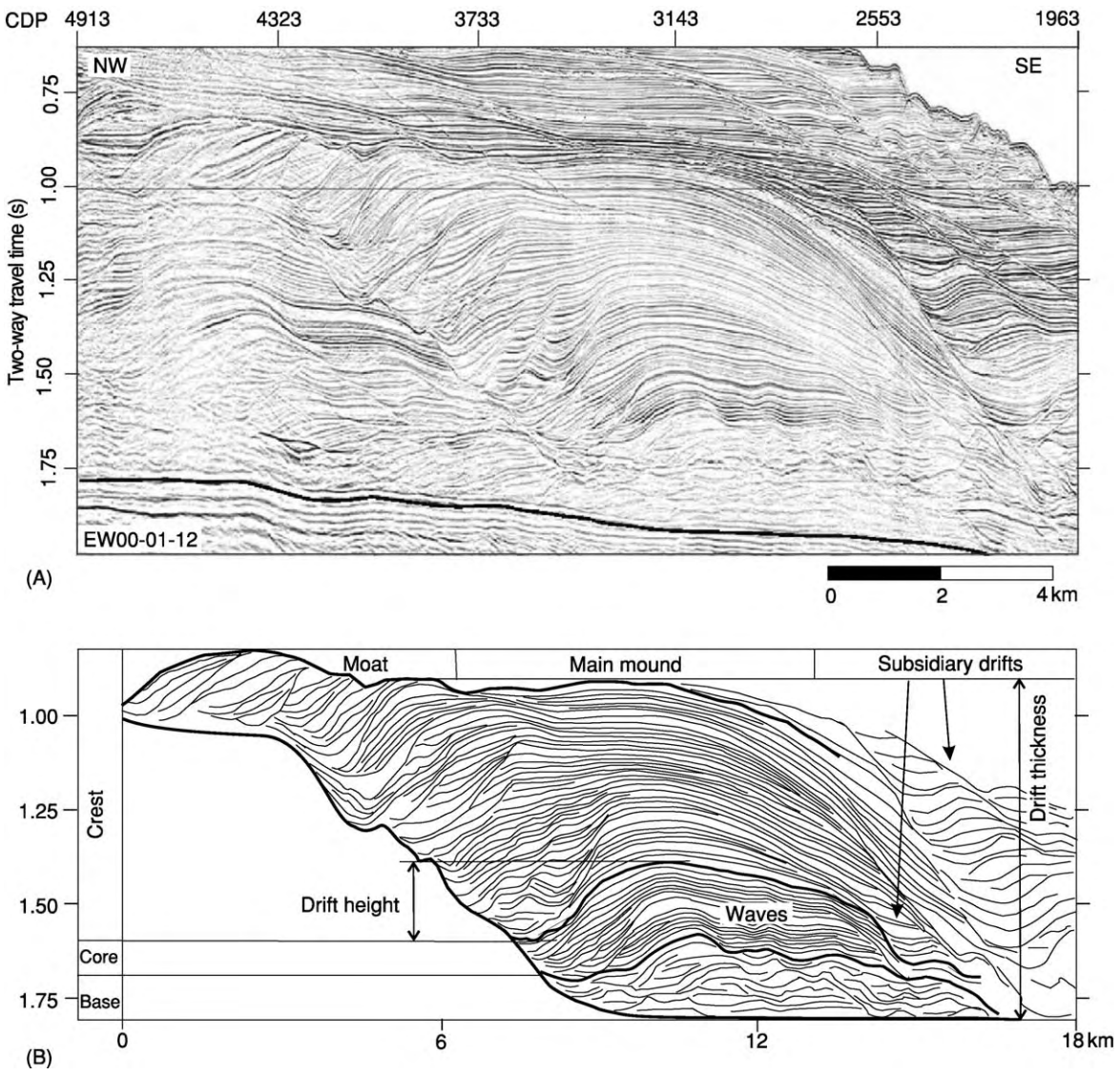


Figure 5 Seismic profile (A) and interpretation (B) showing an elongated, mounded drift in the north eastern Canterbury Bight. Adapted from Lu H, Fulthorpe CS, and Mann P (2003) Three dimensional architecture of shelf building sediment drifts in the offshore Canterbury Basin, New Zealand, *Marine Geology* 193: 19 47, with permission.

from the continental slope. This irregular, discontinuous channel-related drift generally elongates in the flow direction.

b. A 'contourite fan' is a much larger, cone-shaped deposit developed at the downcurrent exit of the conduits. It usually overlays major basal erosive disconformities, but for the most part shares many characteristics with medium-sized turbidite fans.

4. A 'patch drift' is characterized by a random (patchy) distribution controlled by the intricate interaction of the bottom current system with a complex seafloor morphology. Such a small-scale (a few tens of square kilometres), elongated, irregular drift may be either relatively mounded or thinly sheeted. It occurs plastered against reliefs or within a trough where the irregular topography may modify both direction and velocity of the local current flow.
5. A 'confined drift' is characterized by a mounded shape elongated parallel to the basin axis and distinct moats along both flanks. It occurs within relatively small enclosed basins, eventually actively subsiding, and generally shows a complex staking of upward-convex lenticular depositional units.
6. An 'infill drift' is characterized by a mounded geometry that progressively infills the topographic depression or irregularity in which it occurs. It typically forms at the head of the scar or at the

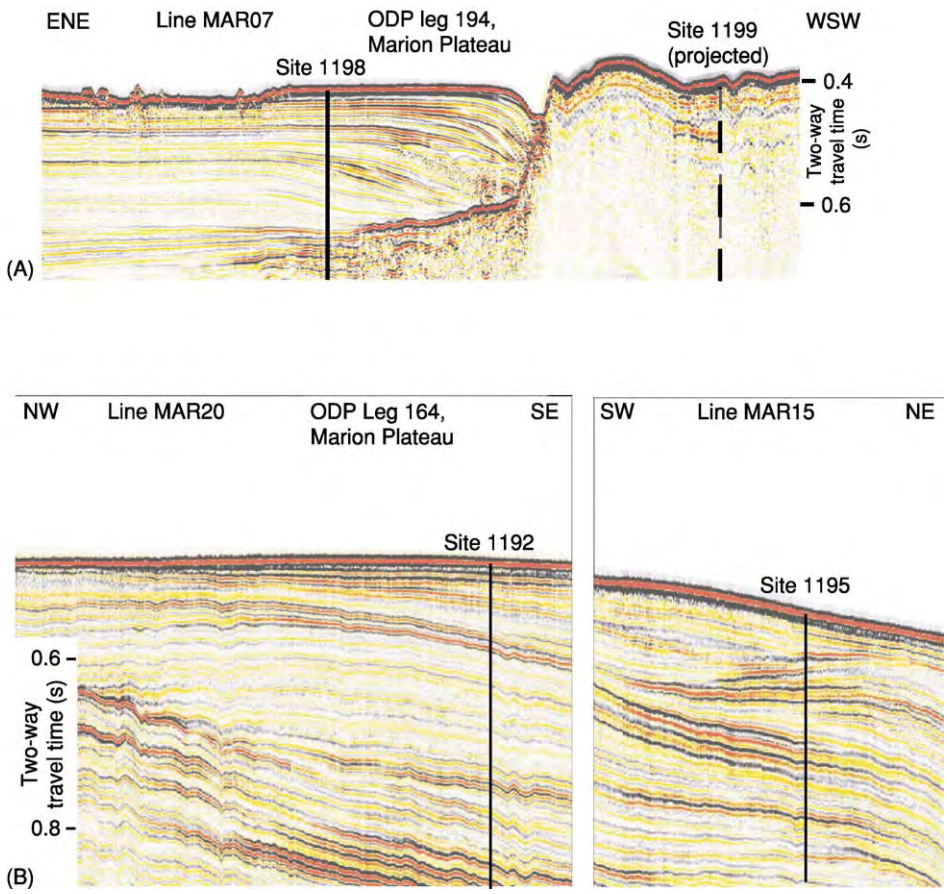


Figure 6 Seismic profiles showing a separated drift close to the Marion Platform margin (A) and a mounded drift complex, with location of some ODP leg 194 sites. Adapted from Shipboard Scientific Party (2002) Leg 194 summary, In: Isern AR, Anselmetti FS, Blum P, *et al.* (eds.), *Proc. ODP, Initial Reports*, Vol. 194. Available from: http://www odp.tamu.edu/publications/194_IR/chap_01/chap_01.htm.

margin of the toe of slumps developed beneath the path of a bottom current, and generally shows moderate relief and extent, variable shape, and downcurrent progradation.

7. A 'fault-controlled drift' is characterized by a certain influence in its development from faulting. This recent addition lacks many well-documented examples and is too little known to allow any generalization regarding deposit geometry and bottom current nature. However, it appears to develop either at the base or at the top of a fault-generated basement relief in response to perturbations in the bottom current flow pattern. A supplementary characteristic may be (subsequently reactivated) syndepositional faulting that affects the relatively steeper side of such drifts.
8. 'Mixed drift systems' are characterized by the interactions of alongslope currents with other depositional processes. The most effective interplay is between contourites and turbidites, but drift development may be variously affected by the association with debrite, hemipelagite, and glacigenic systems. On the basis of a large variety of different types of interactions (interfingering, intercalation, imbrication, incorporation, winnowing, entrainment, etc.), a wide assortment of terms has been adopted, including: companion drift-fan, fan-drift, levee-drift, composite slope-front fan, etc.

Seismic Characteristics

Although much progress has been made in determining the combination of criteria that best represent contourite deposits, seismic data without additional supporting evidence should not be used to make a firm identification of contourites. This is especially true when contourites are closely interbedded with other deep-water sediment facies (e.g., turbidites, hemipelagites, debrites, or glacigenic deposits). In addition, the vital information derived from detailed bathymetric images (either by multibeam or 3D

seismic surveys) of contourite deposits is still relatively rare (Figure 7). Nevertheless, there is a wide consensus on a three-scale set of seismic diagnostic criteria for confident contourite identification.

Large Scale (i.e., Sediment Body)

1. Drift geometry (one of those described in the previous section) beneath a bottom current system;
2. Large dimensions (up to $1 \times 10^6\,km^2$), larger on average than those of turbidite channel–levee systems;
3. Asymmetric mound-and-moat geometry in contrast to the symmetric gull-wing geometry typical of turbidite channel–levee systems (Figure 8);
4. Overall downcurrent elongation; and
5. Widespread regional discontinuities, especially at the base of the drift.

Medium Scale (i.e., Unit)

1. Overall downcurrent migration of the stacked lenticular, upward-convex, seismic units;

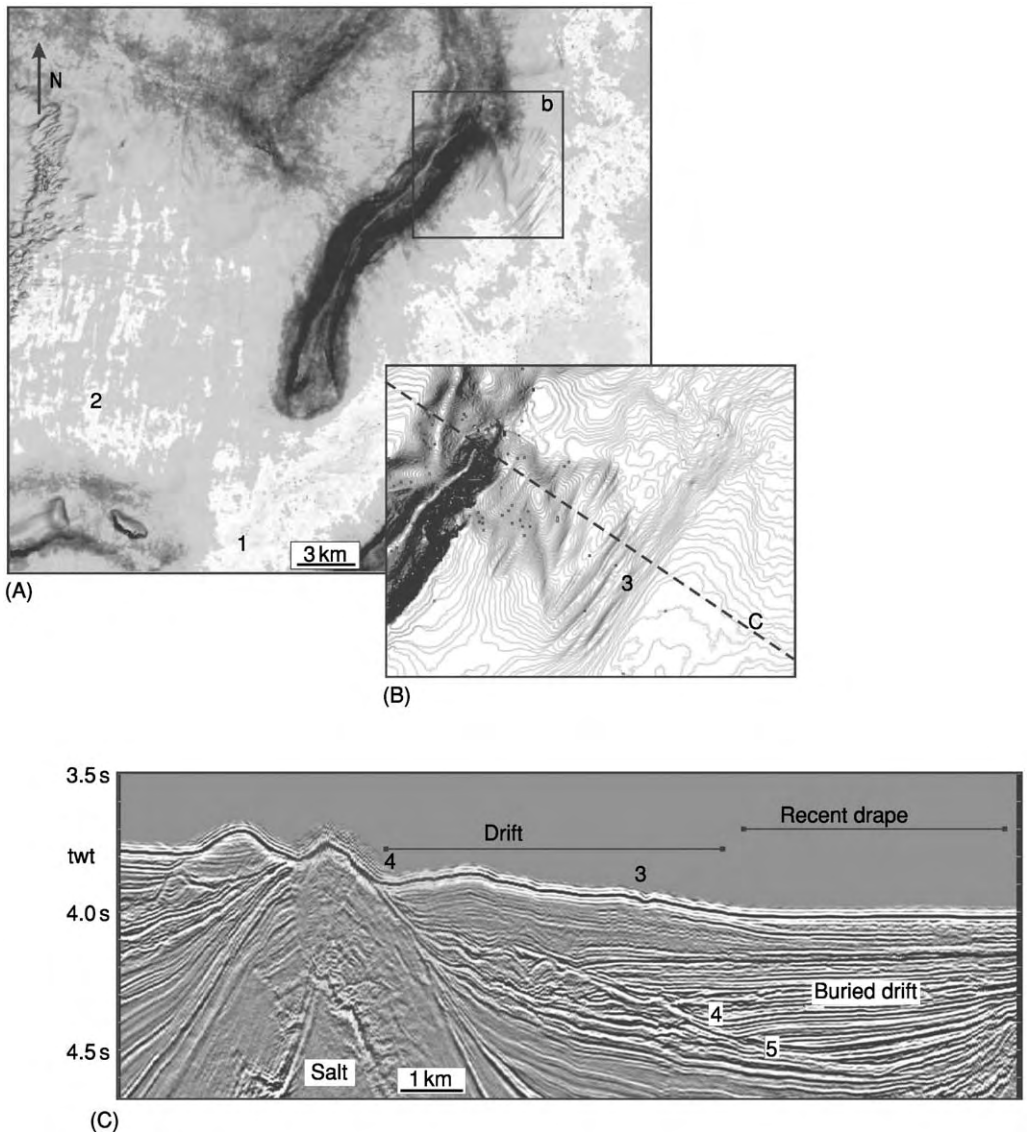


Figure 7 Grey scale seafloor map derived from 3D seismic amplitude (A), detail of the contour map (B), and seismic line (C) showing a mounded drift and bottom current related deposits adjacent to a NE SW trending topographic high. (1) Sheet like coarse grained deposit; (2) nonconfined, string like deposit; (3) erosional furrows; (4) moat between upslope migrating separated drift and topographic high; (5) basal erosional discontinuity. Reprinted from Viana A (2001) Seismic expression of shallow to deep water contourites along the south eastern Brazilian margin, In: Rebesco M and Stow DAV (eds.) Seismic expression of contourites and related deposits, *Marine Geophysical Researches* 22: 509 521, with kind permission of Kluwer Academic Publishers.

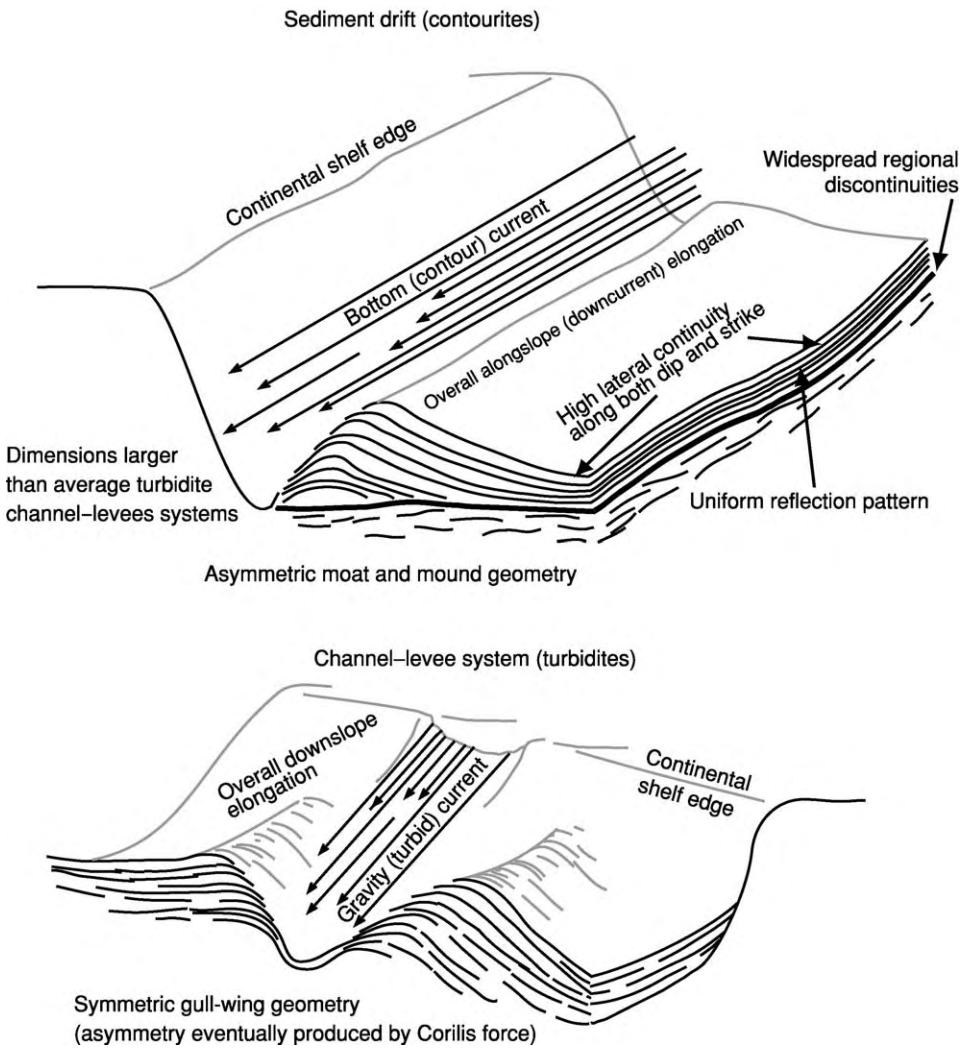


Figure 8 Schematic model showing some typical seismic characteristics of the sediment drifts and the ideal difference with a channel levee system.

2. Downlapping (onlapping on steep slopes) and sigmoidal progradational reflector patterns where downstream and upslope migration occurred;
3. Uniform reflection pattern, commonly found with extensive subparallel moderate to low amplitude reflectors;
4. Well-layered units with high lateral continuity along both strike and dip; and
5. Distribution of the depositional/erosional areas and lateral migration influenced (significantly at high latitudes) by the Coriolis force.

Small Scale (i.e., Facies)

1. Transparent layers and low-amplitude parallel reflectors, though seismic characteristics are very dependent on acquisition and processing methods;
2. Continuous, discontinuous, chaotic, and wavy reflectors in contourites and in turbidites, though seismic facies exclusively associated with contourites have not been defined yet;
3. Sediment waves, in both contourite and turbidite systems, though initially considered diagnostic of contourites; and
4. Bedforms, including longitudinal furrows, depositional tails in the lee of obstacles, and dunes.

Facies Model

Most typically, contourites are composed of fine-grained, structureless, highly bioturbated mud. However, they show a wide range of grain-sizes, composition, and preserved sedimentary structures (Figures 9 and 10).

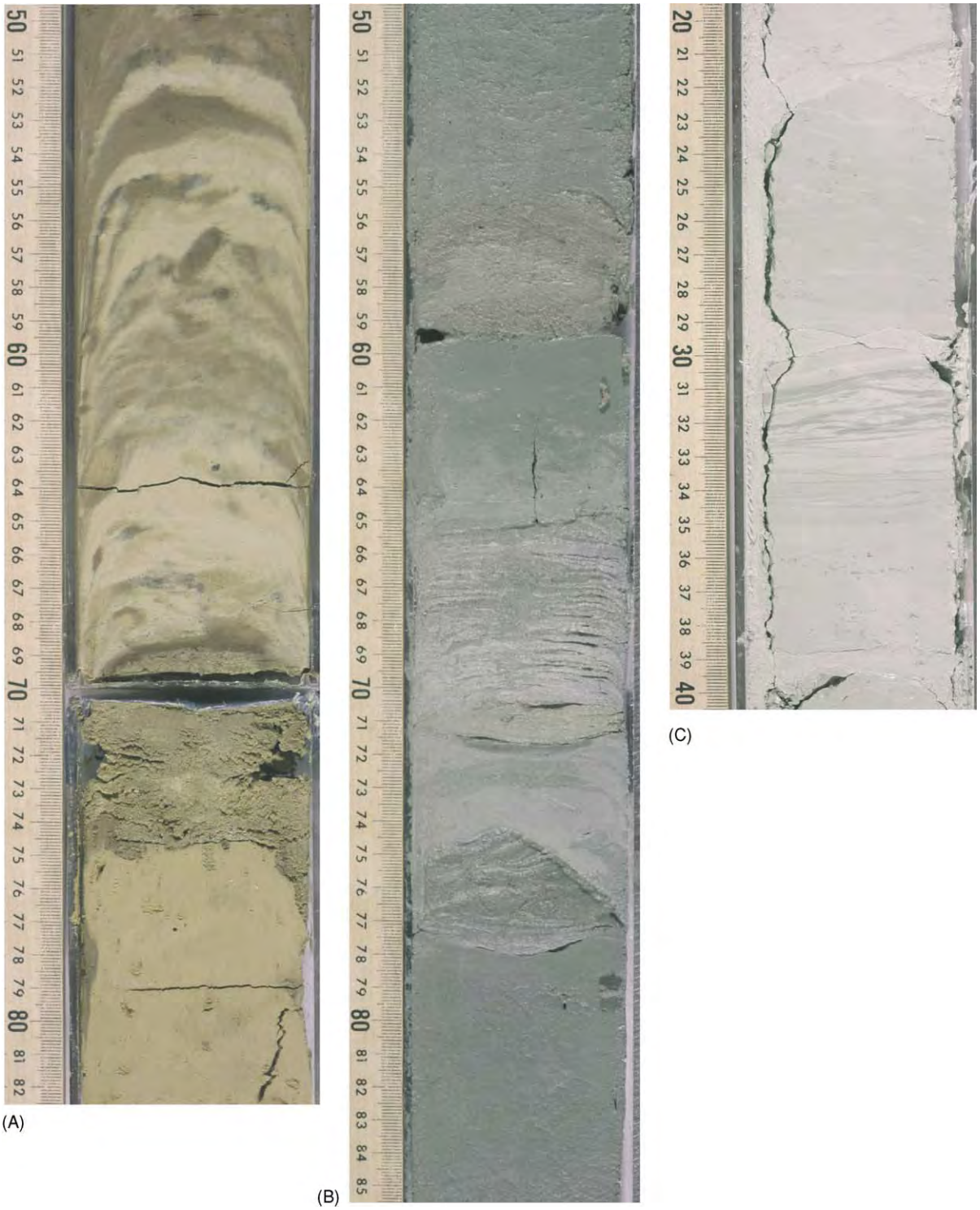


Figure 9 Core photos from ODP Leg 181 in the SW Pacific Gateway. From Carter RM, McCave IN, Richter C, *et al.* (eds.) (1999) *Proc. ODP, Initial Reports*, Vol. 181. Available from: http://www odp.tamu.edu/publications/181_IR/181ir.htm. (A) Core 181 1121B 1H CC, 50 82 cm from the Campbell Drift showing brownish yellow sand with extremely sharp bottom contact (75 cm) with the underlying yellow clay. This sand bed exhibits no grading and is interpreted as a contourite deposite left behind after that an intense winnowing removed the fine material. (B) Core181 1122C 44X 2, 50 85 cm in the left (north) bank levee of the abyssal Bounty Fan showing greenish grey fine sand and silt beds that commonly exhibit sharp scoured top and bottom contacts with the interbedded mottled, dark greenish grey pelagic/hemipelagic bioturbated silty clay. The conspicuous planar and cross laminations, representing concen trations of foraminifera and carbonate debris, suggest a stronger, episodic benthic flow regime, in contrast to that of the decelerating turbidity currents. (C) Core 181 1124C 37X 2, 20 40 cm from the Rekohu Drift showing light greenish grey clay bearing nannofossil chalk with flaser like interbeds and laminae suggesting the presence of sediment moving bottom currents.

They normally occur as thick (tens to hundreds of metres), uniform, fine-grained sequences (including thin to medium coarser-grained beds), or interbedded with hemipelagites, turbidites, and other resedimented facies, or as coarse lags within gateways. They are dominantly homogeneous (Figure 9A), poorly bedded (Figure 10A), and mostly bioturbated (Figure 10B) and mottled throughout, with little primary structures preserved. Bioturbation is generally considered an essential characteristic of contourites based on the belief that an active bottom current would increase oxygen concentration in the water and in turn increase organic activity, thus allowing contourites to be distinguished from episodic turbidites. However, bioturbation potential is also inferred to be a function of current intensity and other life-favourable environmental conditions. As a matter of fact, nonbioturbated contourites have been described as well (Figure 10C). Well-developed but somewhat irregular, fine lamination with indistinct bioturbation may be evident both in shallow-water and in high-latitude facies (Figure 10C), possibly as a result of a hybrid turbidite–contourite deposition. Although cyclicity is common, primary silty parallel and cross lamination (Figures 9B, 10C), where present, shows no regular sequence of facies as in turbidites. Contourites may show reverse grading, with coarse lag concentrations and sharp or erosive contacts (Figures 9A and

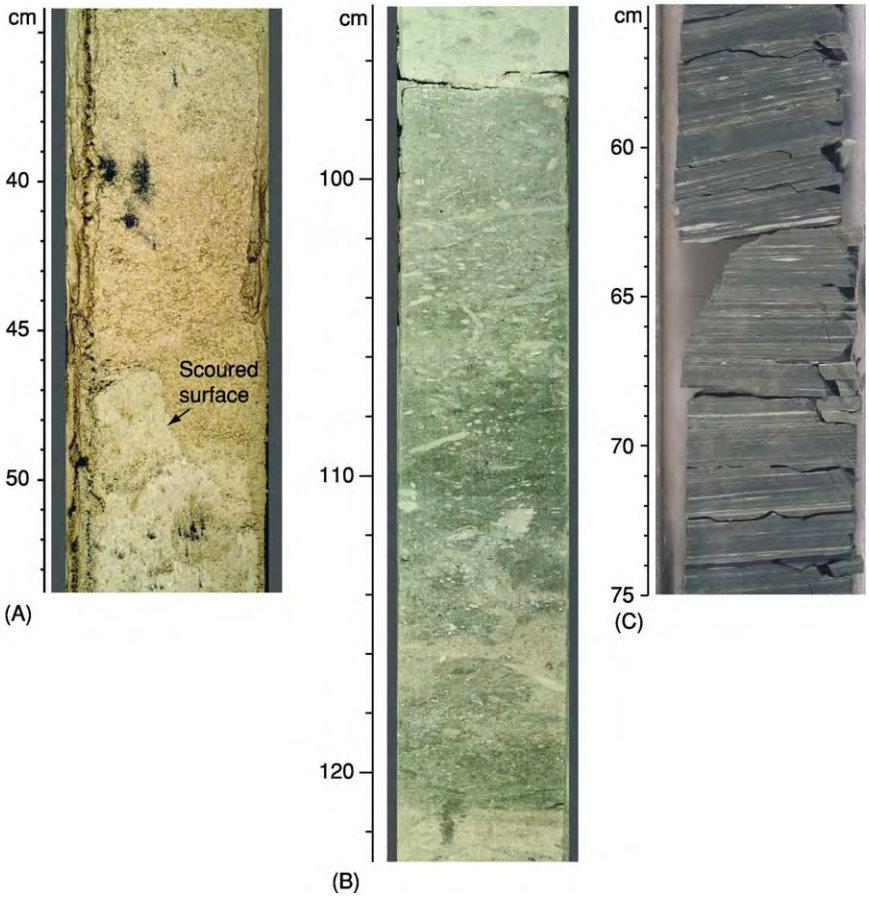


Figure 10 Core photos from two different ODP Legs. (A, B) ODP Leg 194 in the sediment drifts adjacent to Marion Plateau. From Isern AR, Anselmetti FS, Blum P, *et al.* (2002) *Proc. ODP, Initial Reports*, Vol. 194. Available from: http://www odp.tamu.edu/publications/194_IR/194ir.htm. (C) ODP Leg 188, Wild Drift in Cooperation Sea. From Shipboard Scientific Party (2001) Site 1165, in: O'Brien PE, Cooper AK, Richter C, et al. (eds) (2001) *Proc. ODP, Initial Reports*, Vol. 188. Available from: http://www odp.tamu.edu/publications/188_IR/VOLUME/CHAPTERS/IR188_03.PDF. (A) Core 194 1195B 21H 5, 34 54 cm showing a surface scoured into a light greenish grey wackestone by strong bottom currents in a high energy hemipelagic setting. The sediments above the scoured surface consist of alternations between light grey silt and very fine sand sized wackestone and very fine to fine sand sized skeletal packstone dominated by broken planktonic foraminifera. (B) Core 194 1195B 44X 1, 94 122 cm showing alternation of silt sized light grey skeletal packstone and greenish gray mudstone with well preserved *Chondrites* burrows and top and basal sharp scoured surfaces. (C) Core 188 1165C 3R 3, 55 75 cm showing a dark grey claystone with <1 cm beds and silt partings along bedding planes interpreted as contourites deposited by relatively slow moving, near continuous bottom currents entraining episodic turbidity currents, whereas bioturbation occurring in discrete decimetre sized intervals reflect comparatively low deposition rates.

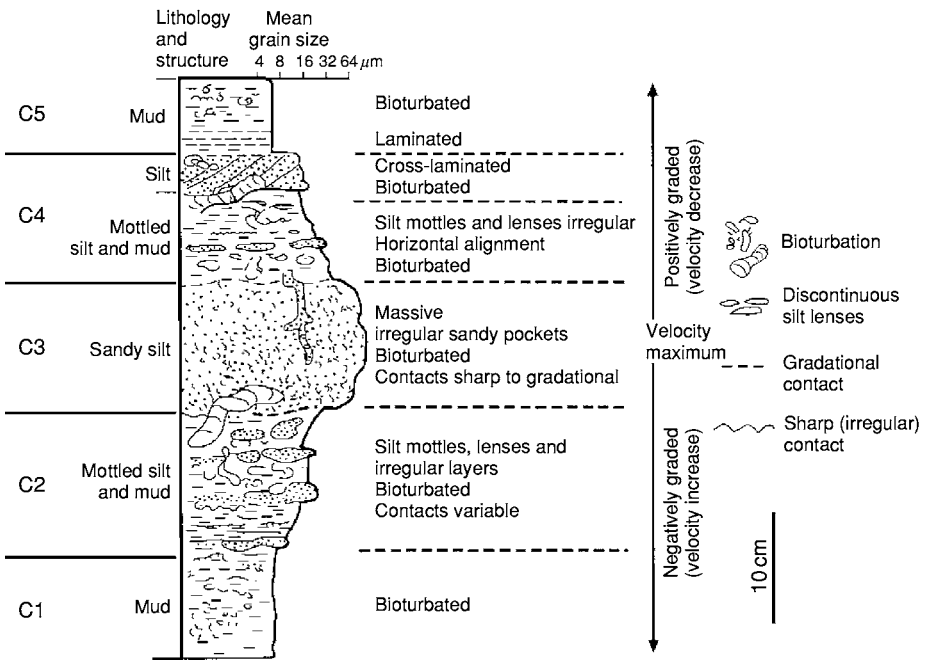


Figure 11 Conjectural standard composite (muddy + sandy) contourite facies model. Adapted with permission from Stow DAV, Faugères JC, Howe J, Pudsey CJ, and Viana A (2002) Bottom currents, contourites and deep sea sediment drifts: Current state of the art, In: Stow DAV, Faugères JC, Howe J, Pudsey CJ, and Viana A (eds.) Deep Water Contourite Systems: Modern Drifts and Ancient Series, Seismic and Sedimentary Characteristics. *Geological Society, London, Memoir* 22: 7 20, with permission.

9B, 10A and 10B). Grain size may vary from (silty) mud to sand. Gravel-rich contourites are common in glacigenic environments due to the presence of ice-rafted debris. Bottom current winnowing and erosion in narrow gateways may also produce gravel lags or shale-chip layers. Sorting is generally medium to poor, but some sandy contourites may be well sorted and relatively free of mud, showing low or negative skewness values. No offshore textural trend may be observed. Composition is generally mixed with a combination of biogenic and terrigenous material. High sand-sized content in muddy hemipelagic contourites is often formed by bioclasts (Figures 9B, 10B). The terrigenous component largely reflects the nearby local sources, but alongslope mixing and far-off contributions also occur. In contrast the composition is essentially pelagic in biogenic contourites developed in open oceans or in upwelling areas. Manganese dioxides or iron oxides may occur as micronodules, stains, coating, and metal-enriched crusts within the manganiferous contourites. Fabric (when not exceedingly altered by bioturbation) is indicative of alongslope grain orientation, parallel to the bottom current flow. Contourites (Figure 11) are normally arranged in few decimetre-thick sequences showing gradual grain size and/or compositional changes. Five divisions may be eventually identified: (C1) lower mud, (C2) lower mottled silt, (C3) middle sand, (C4) upper mottled silt, and (C5) upper mud. A complete C1–5 sequence is interpreted to represent a gradual long-term current velocity variation and/or sediment supply change. Partial base-only or top-only sequences are also common.

Facies Continuum and Distinguishing Criteria

Deep-sea sediments may be interpreted as the product of three main processes: gravity-driven downslope resedimentation, alongslope bottom current activity, and slow pelagic settling through the water column. However, such processes are in fact end members belonging to a continuum spectrum. Deep-sea sedimentary facies are often hybrids resulting from interactions within such process continuum (Figure 1) rather than a mere stacking of interstratified deposits produced by discrete processes. Imaginably, the distal part of very-low-concentration (0.025 to 2.5 g l ^{1}), low-velocity (10–50 cm s ^{1}), fine-grained (silt and clay) turbidity currents may be easily deflected by Coriolis force from its natural downslope path to an alongslope direction and then entrained in the indigenous contour current. Concurrently, the tests of calcareous and siliceous planktonic organisms and the associated organic matter biosynthesized in the surface layers of the ocean may be significantly contributed by slow vertical

settling through the water column. The current originally driven by the gravitational effect of its terrigenous sedimentary load would hence imperceptibly grade into a thermohaline- or wind-driven, contour-parallel bottom current capable of depositing a hemipelagic, 'resedimented' contourite.

Although the traction effects of the bottom currents within an otherwise tranquil pelagic realm are fairly visible and may easily lead to the interpretation of the resulting deposits as contourites, a much more difficult and disputable work is required in trying to distinguish contourites from turbidites. Different sets of criteria have been proposed to characterize bottom current-reworked sandy sediments that occur where strong, permanent bottom currents have been active within a turbidite setting. In general, reworked turbidites are constituted by predominantly fine-grained sand and silt and are specially characterized by thin-bedded to laminated sand (usually less than 5 cm) in deep-water mud. Rhythmic occurrences of sand and mud layers are common (Figure 10C), with numerous sand layers (50 or more per 1 m of core). In many cases, the lower divisions of turbidites may be preserved and the upper ones entirely removed or partially modified by bottom current reworking. Reworked tops may show common bioturbation and burrowing, reverse grading, and lag concentrations. Horizontal and bidirectional low-angle cross-laminations and sharp erosive contacts may occur within the turbidite sequence. Cross-bedding is believed to be produced by steady and equlibrium flow conditions due to the bottom current action and, in contrast, is, envisaged with difficulty in episodic turbidity current events (Figure 9B). With respect to the adjoining turbidites, they may be cleaner, better sorted, more negatively skewed, and with a more bimodal fabric. In addition, the following diagnostic sedimentary structures have been proposed, although not universally agreed upon: lenticular bedding and starved ripples; mud offshoots; flaser bedding (Figure 9C); and the occurrence of sand layers with traction structures in discrete units, but not as part of a vertical sequence of structures, such as the Bouma sequence with basal graded division.

See Also

Petroleum Geology: The Petroleum System. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Processes:** Erosional Sedimentary Structures; Depositional Sedimentary Structures; Deposition from Suspension. **Sedimentary Rocks:** Deep Ocean Pelagic Oozes. **Seismic Surveys**. **Unconformities**. **Unidirectional Aqueous Flow**.

Further Reading

Faugères JC, Stow DAV, Imbert P, and Viana A (1999) Seismic features diagnostic of contourite drifts. *Marine Geology* 162: 1 38.

Gao ZZ, Eriksson KA, He YB, Luo SS, and Guo JH (eds.) (1998) *Deep Water Traction Current Deposits*. Beijing/New York: Science Press.

Lu H, Fulthorpe CS, and Mann P (2003) Three dimensional architecture of shelf building sediment drifts in the offshore Canterbury Basin, New Zealand. *Marine Geology* 193: 19 47.

Maldonado A and Nelson HC (eds.) (1999) Marine geology of the Gulf of Cadiz. *Marine Geology* 155.

McCave IN and Tucholke BE (1986) Deep current controlled sedimentation in the western North Atlantic. In: Vogt P and Tucholke B (eds.) *The Geology of North America, Vol. M, The Western North Atlantic Region*, pp. 451 468. Boulder, CO: Geological Society of America.

Mienert J (ed.) (1998) European North Atlantic margin (ENAM): Sediment pathways, processes and Flux. *Marine Geology* 152.

Nowell ARM and Hollister CD (eds.) (1985) Deep ocean sediment transport Preliminary results of the high energy Benthic boundary layer experiment. *Marine Geology* 66.

Pickering K, Hiscott R, and Hein F (1989) *Deep Marine Environments: Clastic Sedimentation and Tectonics*, pp. 218 245. London: Unwin Hyman.

Rebesco M and Stow DAV (eds.) (2001) Seismic expression of contourites and related deposits. *Marine Geophysical Researches* 22(5 6).

Shanmugam G (2000) 50 years of the turbidite paradigm (1950s 1990s): Deep water processes and facies models A critical perspective. *Marine and Petroleum Geology* 17: 285 342.

Steele JH, Turekian KK, and Thorpe SA (eds.) *Encyclopedia of Ocean Sciences*. London: Academic Press.

Stoker MS, Evans D, and Cramp A (eds.) (1998) *Geological Processes on Continental Margins: Sedimentation, Mass Wasting and Stability*. Special Publication 129. London: Geological Society.

Stow DAV and Faugères JC (eds.) (1993) Contourites and bottom currents. *Sedimentary Geology* 82(1 4).

Stow DAV and Faugères JC (eds.) (1998) Contourites, turbidites and process interaction. *Sedimentary Geology* 115(1 4).

Stow DAV and Mayall M (eds.) (2000) Deep water sedimentary systems: New models for the 21st century. *Marine and Petroleum Geology* 17.

Stow DAV, Faugères JC, Howe J, Pudsey CJ, and Viana A (eds.) (2002) Deep water contourite systems: Modern drifts and ancient series, seismic and sedimentary characteristics. *Geological Society, London, Memoir* 22.

Wynn RB and Stow DAV (eds.) (2002) Recognition and interpretation of deep water sediment waves. *Marine Geology* 192(1 3).

Deltas

T Elliott, University of Liverpool, Liverpool, UK

Introduction

Deltas form where sediment-laden rivers enter oceans, enclosed seas, lakes, or lagoons and supply more sediment than can be distributed by processes that operate within the receiving basin. They produce a protrusion of the shoreline and typically advance, or prograde, into the basin. Major deltas such as the Amazon, Ganges-Brahmaputra, Nile, and Mississippi are served by well-defined, often long-lived, continent-scale drainage basins that result in an alluvial trunk stream (Figure 1). Deltas are important sites of sediment deposition in their own right and they also create basin-scale depocentres, in that deltas in turn supply sediment to shelves, slopes, and deep-basin submarine fans. It is no accident that many present-day submarine fans occur basinwards of major deltas.

The social, cultural, and economic importance of deltas is considerable. They account for a significant proportion of the world's wetlands and are important in terms of habitation, agriculture, fishing, and navigation. They are extraordinarily sensitive to fluctuations in sea-level, whether caused by subsidence or eustatic sea-level change, and therefore feature prominently in studies of land loss, water management, and coastal stability. For example, in the United States, wetlands in Louisiana have been experiencing compaction-induced rates of sea-level rise in excess of $1\,\mathrm{cm\,year^{-1}}$ for several decades, causing land loss rates to exceed $100\,\mathrm{km^2\,year^{-1}}$. Current measurements and predictions of the rate at which eustatic sea-level is presently rising will have a major impact not only in Louisiana, but also on deltas globally. In view of their links to continent-scale drainage basins, deltas are also important when considering pollutant dispersal to the margins of oceans. Finally, the deposits of former deltas host significant fossil fuel reserves, providing both source and reservoir rocks in many oil- and gas-bearing basins.

Controls on the Variability of Deltas

Deltas are extremely diverse in appearance and behaviour. The formation and ultimate character of a delta are determined by variables that stem from the hinterland that is responsible for creating the sediment supply and from the fluvial system that serves the delta, and from the basin that receives the sediment. Major controls, such as climate and tectonics, act on both the hinterland and the receiving basin, and sea-level is critical at the interface between them. The means by which the variability of deltas is presently summarized involves a classification scheme based on the sediment calibre delivered to the delta and the interaction between fluvial, wave, and tidal processes that takes place in the shoreline or delta front sector (Figure 2). This allows interpretation of the Mississippi delta as a fine-grained, fluvial-dominated delta and the Ganges-Brahmaputra delta as a more sand-rich, tide-dominated delta. Despite their complexity and variability, deltas comprise two major components: the delta front, which includes the river mouth, adjacent shoreline, and seaward-dipping profile, and the delta plain, which is the low-lying, largely subareal part of the delta behind the delta front; the delta plain is dominated by

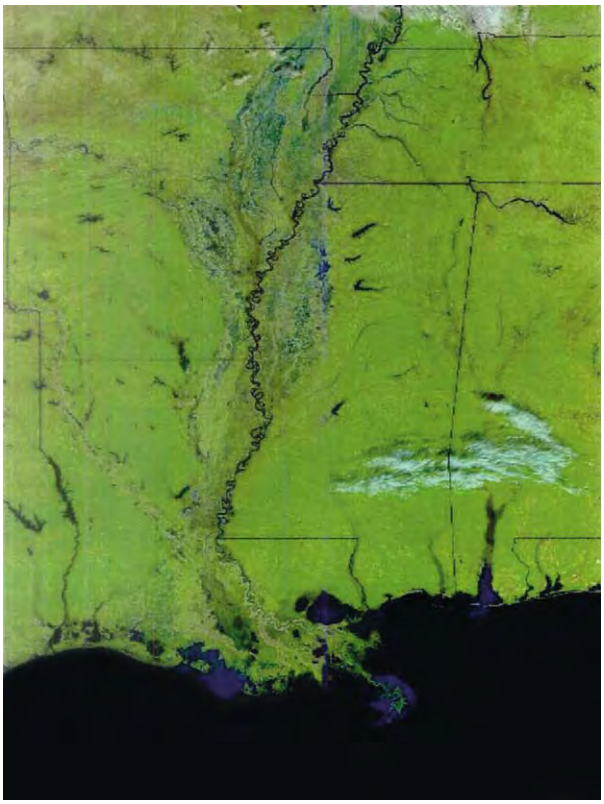


Figure 1 The Mississippi River and delta complex; the Mississippi River drainage basin extends over two thirds of the United States and is the alluvial trunk stream of the delta. The modern birdsfoot delta is flanked by the now abandoned, premodern shoal water deltas. The light blue sediment rich embayment west of the modern delta is where the new Atchafalaya and Wax river deltas are located.

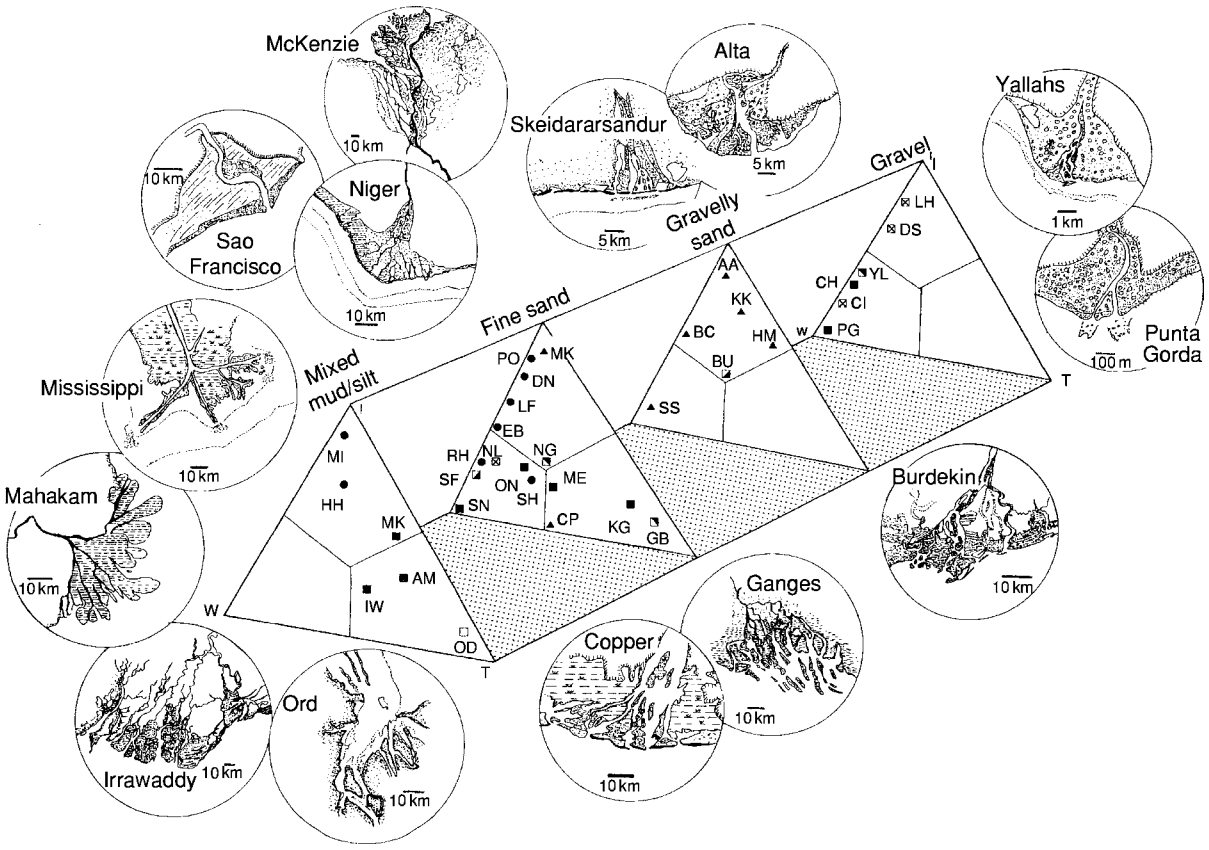


Figure 2 A classification scheme for deltas based on sediment calibre and the process regime of the delta front. In each triangular cross section, the top portion represents an input dominated (I) process, the left bottom portion represents a wave dominated (W) process, and the right bottom portion represents a tide dominated (T) process. Reproduced with permission from Orton GJ and Reading HG (1993) Variability of deltaic processes in terms of sediment supply, with particular emphasis on grain size. *Sedimentology* 40: 475 512.

distributary and crevasse channels, levees, bays, lakes, and tidal flats in tide-influenced deltas.

River Mouth Processes and Plumes in Deltas

The most important subenvironment of deltas is the mouth of the river, or the distributaries of the river, because this is where the interaction takes place between the sediment-laden river waters derived from the hinterland and the receiving basin. When river discharge reaches the mouth of a channel, it decelerates due to loss of gradient, expands due to loss of confinement by the channel margins, is required to mix with the waters of the receiving basin, and comes under the influence of basinal processes such as waves and tides. The outcome of this complex interaction will determine, first, whether a delta forms, and, if so, the characteristics the delta. The first-order control is the relative density of the incoming sediment-laden river waters and the waters of the receiving basin (Figure 3). Where the water masses are equal density (homopycnal flow), immediate mixing occurs at the river mouth and the sediment load is deposited locally at the river mouth. This can lead to the formation of localized, steep-fronted deltas referred to as 'Gilbert-type' deltas, which form commonly where bedload-rich rivers enter lakes. They can also form where gravel-grade rivers or alluvial fans enter the margins of marine basins, creating fan deltas. In this case, deposition results from the loss of competence to transport the gravel-grade material beyond the river mouth, rather than the density contrast of the water masses.

Where sediment-laden river waters are less dense than the basin waters (hypopycnal flow), then the outflow from the river is buoyantly supported and extends into the basin as a surface jet or plume (Figure 3). This process operates in major marine deltas because salt water is denser than fresh water is, and suspended sediment concentrations in river waters are typically insufficient to overcome this difference. Satellite images of major deltas invariably show murky, sediment-laden

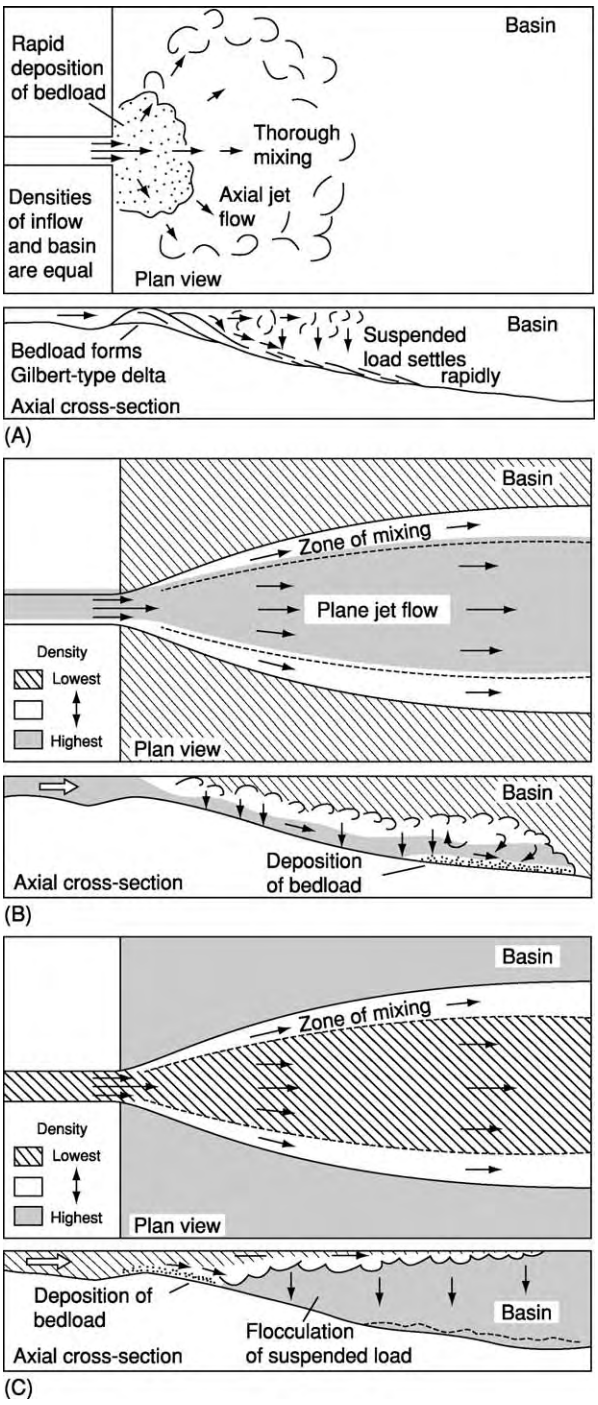


Figure 3 Fundamental processes that operate at the mouths of rivers or distributary channels based on the relative densities of the sediment laden fluvial discharge and the water of the receiving basin, distinguish ng homopycnal (A), hypopycnal (B), and hyper pycnal (C) flow. Reproduced with permission from Bates (1953).

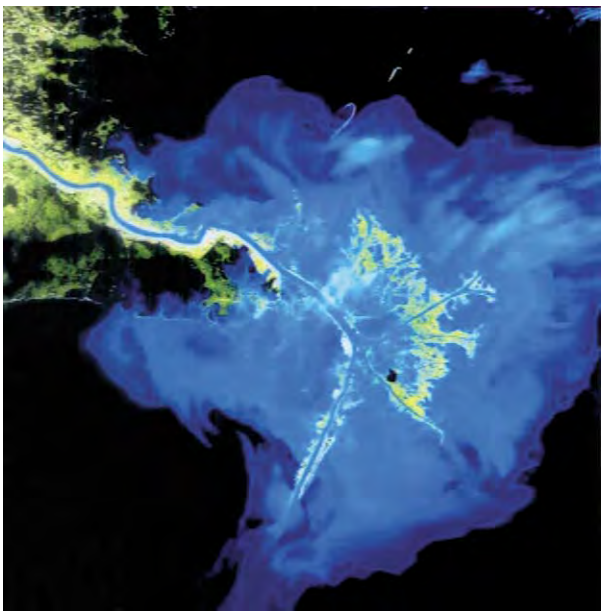


Figure 4 Plumes of sediment laden water (light blue) issuing from the Mississippi delta during a high discharge, river flood period. The plumes are buoyantly supported because the saline basin water is denser than the river water, despite the presence of suspended sediment in the latter.

plumes issuing from river and distributary mouths across the darker, sediment-free waters of the basin (Figure 4). In this situation, relatively coarse, sand-grade sediment is often deposited close to the channel mouth, either as bedload or from suspension, whereas finer material is transported further offshore in suspension. Turbulent mixing gradually reduces the density contrast between the water masses, causing the suspended load of the plume to be deposited with a trend of decreasing grain size offshore.

Where the sediment-laden river waters are more dense than the receiving basin (hyperpycnal flow), an underflow occurs, causing the sediment-laden flow to move discretely beneath the basin waters down the gradient of the basin margin (Figure 3). The extent to which this process operates in deltas is debatable. Clearly, if it operated exclusively, then a delta would not form, because all of the sediment load of the river would bypass the shoreline to be deposited subaqueously. However, the process may operate intermittently during major river floods. In a freshwater lake, suspended sediment concentrations of $\sim 1\,\mathrm{kg\,m^{-3}}$ can produce a hyperpycnal flow, but in marine settings, concentrations of $35–45\,\mathrm{kg\,m^{-3}}$ are required due to the greater density of salt water. The Huanghe River delta (China) regularly exceeds the critical value and produces hyperpycnal flows. Predictive modelling indicates that numerous small rivers entering marine basins also have the potential to produce hyperpycnal flows, because their catchments produce high sediment yields but limited water discharge. However, the modelling also argues that major river deltas cannot produce hyperpycnal flows,

Figure 5 The delta plain and upper delta front of the strongly tidally influenced Ganges Brahmaputra delta. The distributaries are tidal estuaries, the lower delta plain comprises tidal flats and creeks, and tidal processes contribute to the seaward dispersal of the plumes of suspended sediment at the distributary mouths.

because their ratio of sediment yield to water discharge is too low, even during extreme floods. A residual question is the extent to which these arguments apply during periods of sea-level lowstand, when the physiography of the fluvial systems is different (see later, Shelf-Edge Deltas in Quaternary Lowstand Periods).

The dispersal of sediment at the river mouth is also influenced by processes within the receiving basin. Moderate to high levels of wave or tidal processes promote mixing of the two water masses and also influence the orientation of the resultant plume. Wave processes limit the offshore extent of the plume and extend shore-parallel transport. Continuous beach sands rim the delta front, connecting the distributary mouths. In contrast, tidal currents extend the offshore component of sediment transport and elongate the plume into the basin (Figure 5). The tidal currents produce a network of tide-dominated channels separated by long, linear sand ridges.

The Initiation and Abandonment of Deltas

Deltas have a natural life cycle of initiation, growth, and abandonment. As deltas prograde into the receiving basin, the length of the lower part of the river feeding the delta is increased and the gradient is reduced. Shorter, steeper gradients therefore develop adjacent to stretches of the coastal plain. During major floods, the banks of a river can be breached and a new channel can branch, or avulse, off the main channel. If the breach is deep and is reused during subsequent floods, the new channel, with its steeper gradient, develops at the expense of original channel by diverting increasing amounts of the river discharge into the new channel. On reaching the margin of the receiving basin, the new channel initiates a delta, providing that it supplies more sediment than the receiving basin can redistribute. The delta at the mouth of the original river subsides because it no longer receives sufficient sediment to counter compaction of previously deposited sediments. The waters of the receiving basin progressively drown the former delta. The original delta dies and new delta is born.

This life cycle is particularly pronounced where fluvial-dominated deltas form on broad, unconfined coastal plains, such as the Gulf of Mexico. During the past 6000 years, the site of the Mississippi delta has varied along the margin of the Gulf of Mexico. Lobes and major delta complexes have switched across the margin of the Gulf of Mexico to the west and, to a lesser extent, to the east of the modern delta (Figure 6). These lobes and complexes that have been abandoned sequentially during the past 6000 years are now in differing states of abandonment. The present-day birdsfoot delta was established 600–800 years ago and new deltas are presently forming at the mouth of the Atchafalaya River and the Wax River, west of the modern delta (Figure 7). The Atchafalaya River branches off the Mississippi River and has a course to the Gulf of Mexico that is more than 300 km shorter than that of the Mississippi River. The tendency for sediment to be diverted along the Atchafalaya River has been stemmed for several decades by man-made controls because of the dire consequences for land loss in Louisiana if the avulsion is allowed to proceed unabated. The continued progradation of the modern delta towards the shelf edge is therefore artificially maintained.

The Abandonment of Deltas

What happens when a delta is abandoned? In view of current concerns of rising global sea-level, this question is germane. Rising sea-level not only inundates low-lying delta plains but also reduces gradients in the lowest parts of the river course and promotes avulsion. Abandonment of a delta therefore involves a reduction in sediment supply, whether via a lowering of river gradients or diversion of the river. Compaction of previously deposited sediments and related subsidence continues and land loss results because the supply can no longer maintain the delta plain.

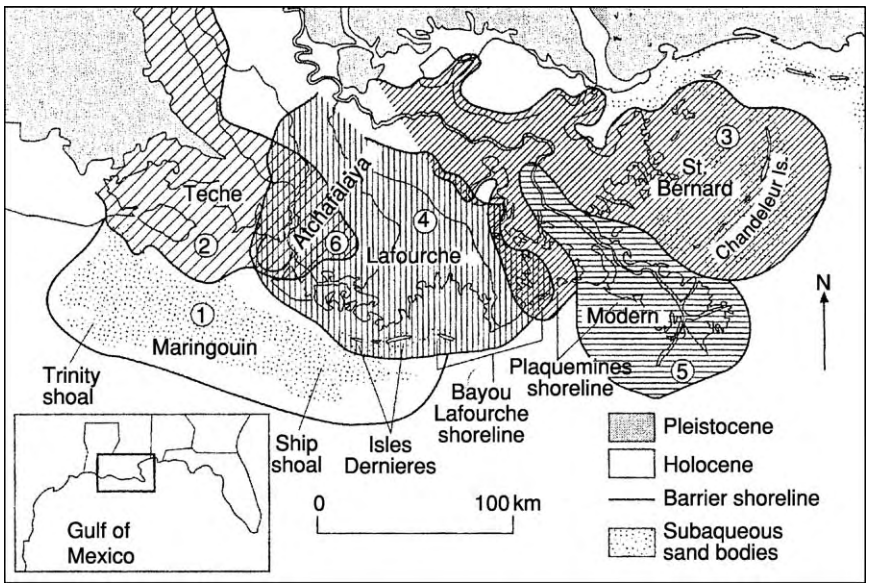


Figure 6 The major lobe complexes of the Mississippi delta system. The lower numbers indicate progressively older lobes that were active during the past 6000 years but are now in varying stages of abandonment (compare with **Figure 1**).

Figure 7 The Atchafalaya (right) and Wax River (left) deltas, forming in the Atchafalaya basin, are the new deltas of the Mississippi complex, west of the modern birdsfoot delta (see **Figure 1** for location of the Atchafalaya basin).

Additionally, with a reduction in sediment supply to the delta, the effectiveness of basinal processes is enhanced and these processes begin to rework the abandoned delta, particularly the former delta front. New shorelines are created and migrate landwards over the former delta plain as sea-level continues to rise. Precisely what happens during abandonment is therefore determined by the rate and magnitude of relative sea-level rise, the physical regime of the receiving basin, and the nature of the delta being abandoned.

In the Mississippi delta, the premodern, more lobate deltas experienced moderately high rates of relative sea-level rise related to compactional subsidence (**Figure 8**). Delta front areas are partially reworked initially into land-attached beaches (**Figure 9**) and subsequently into detached, arcuate barrier islands (**Figure 10**) that migrate landwards and are ultimately submerged as subsidence continues. The lower part of the former delta plain is converted into a shallow lagoon that progressively inundates the former delta plain (**Figure 11**), and the upper part becomes an extensive peat-forming swamp that blankets the underlying delta plain deposits. Offshore, the lower parts of the delta front accumulate slowly deposited, faunally rich muds or even carbonates as a consequence of the reduced sediment supply. In overall terms, the abandoned delta is draped by a suite of facies that record slow sedimentation rates by comparison with the facies that accumulate during active delta progradation. These slowly deposited abandonment facies produce distinctive marker horizons that are invaluable aids to identifying abandonment events and mapping the extent of former lobes and delta complexes.

Syn-Sedimentary Deformation in Deltas

Deltas are inherently unstable. This tendency has been known for some time in deltas worldwide, but as hydrocarbon exploration in the Gulf of Mexico extended onto the delta front area in the 1960s, the extent of these processes suddenly became clear. As soon as drilling rigs were placed on the delta front, they encountered problems. In 1969, as Hurricane

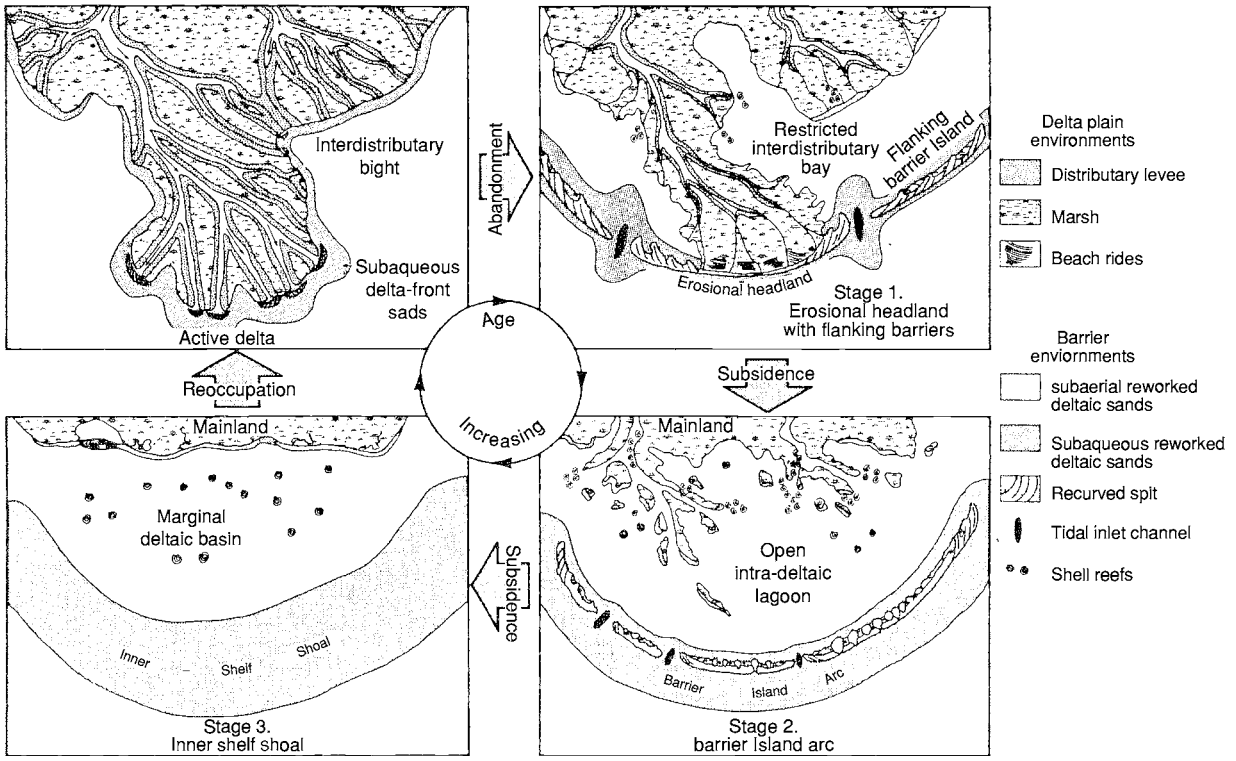


Figure 8 The evolution of the premodern lobe complexes as they are abandoned. In Stage 1 of abandonment, the former delta is reworked into a beach and barrier island complex as the effectiveness of the waves is enhanced and sediment supply diminishes. In Stage 2, transgression causes the initial beach to detach and migrate landwards and the former delta plain to be converted into a lagoon. In Stage 3, continuing transgression causes the barrier island to drown. At this stage, the only exposed part of the former delta is the upper delta plain, which is now blanketed beneath an extensive peat blanket (see **Figures 9** **11** for illustrations of Stages 2 and 3).

Figure 9 Formation of land attached, transgressive beaches during the early stages of abandonment of the Lafourche delta complex, Mississippi delta. Note the sealing of a former distributary channel by the beach sands and the storm induced wash over lobes that project onto the former delta plain (compare with Stage 1 of **Figure 8**).

Camille passed west of the delta, one rig was lost and another moved downslope. This prompted a thorough re-evaluation of the delta front, including seafloor mapping and renewed geotechnical investigations. During the 1970s, the diversity and areal density of deformational processes operating on the delta front were appreciated for the first time Figure 12). The Mississippi delta is still the most well-understood delta in this regard, but the processes are known to occur on many other deltas, and syn-sedimentary deformation should be regarded as an integral part of delta systems, particularly if the deltas have a significant fine-grained component to their sediment calibre. These processes are important in understanding patterns and rates of subsidence in deltas and also because they can have a profound effect on the final facies patterns that are preserved in deltaic deposits. The latter point is confirmed in studies showing that ancient deltaic successions commonly exhibit these features.

Aside from the possibility of seismic activity where deltas are located in tectonically active basins, there are three main mechanisms for syn-sedimentary deformation:

1. Oversteepening of the front of the delta related to high sediment accumulation rates can cause the delta to collapse, releasing large volumes of sediment as slides and slumps to deeper waters areas offshore and leaving a large detachment scar that is repaired by later sedimentation.

Figure 10 Chandeleur Island, a transgressive barrier island associated with abandonment of the St Bernard lobe, Mississippi delta complex. The lagoon on the landward side is established on the former delta plain; note the irregular margin on the barrier island on the lagoonal side due to multiple washover fans that are associated with the landward migration of the island across the lagoon (compare with Stage 2 of **Figure 8**; see also **Figure 1**, in which Chandeleur Island can be seen in an apparent NNE direction from the modern delta).

Figure 11 Levees bordering a former distributary channel in the delta plain of an abandoned delta, Mississippi delta complex. The remainder of the delta plain has been transgressed and is submerged, leaving only the highest former levee exposed.

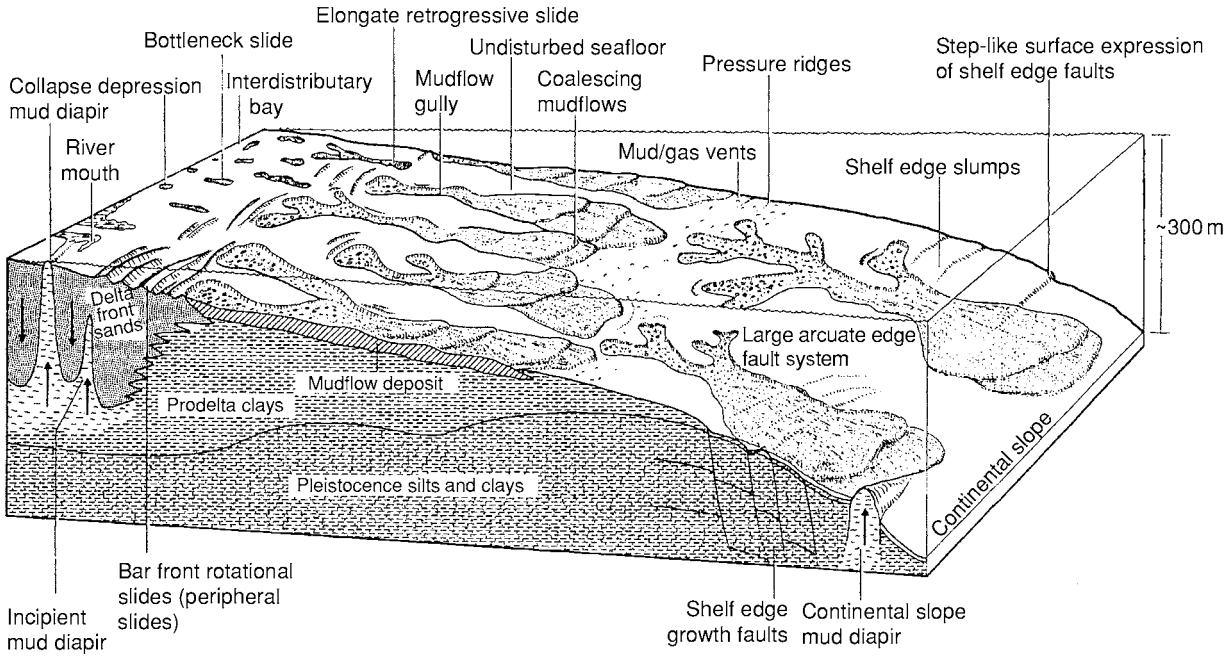


Figure 12 Summary of syn sedimentary deformational features in the delta front of the modern Mississippi delta; note the distinction between superficial and deeper seated features. Reproduced with permission from Coleman JM, Suhayda JN, Whelan T, and Wright LD (1974) Mass movement of Mississippi River delta sediments. *Transactions of the Gulf Coast Association of Geological Societies* 24: 49 68.

2. Repetitive pounding of the delta front by storm or hurricane waves can reduce the shear strength of near-surface sediments, leading to a coupling between surface waves and bottom and to sediment surface waves that can penetrate up to 20 m below the sediment surface.
3. High rates of sediment accumulation lead to thick piles of sediment that are water rich and in some cases also include early methane and carbon dioxide gases. During early burial of these sediments, pore pressures often increase as the sediments begin to compact and permeability is reduced.

Eventually, pore pressures begin to support part of the sediment load, thereby slowing compaction and leading to the sediments being overpressured and undercompacted for their depth. These sediments are unstable in relation to further sediment loading, and they creep basinwards, creating a wide range of deep-seated deformational features.

Superficial Deformational Features

Superficial deformational features include circular collapse depressions, linear gullies and mudflows, and rotational slides (Figure 13). It is estimated that 50–60% of the front of the Mississippi delta has failed via these mechanisms and has been transferred basinwards in recent times. Collapse depressions are created by wave pounding of shallow upper delta front areas between distributary mouths during storm or hurricane periods. As waves cyclically pound the sediment surface, they cause sediments to expel water and early methane gas, to re-pack, and to locally collapse, creating depressions 100–200 m across and 1–3 m deep. The downslope margins of the depressions often fail and viscous sediment/water mixes flow downslope as mudflows. The mudflows shear a linear pathway that is enlarged by rotational failures of the margins after the mudflow has passed. This creates delta front gullies 1000 m wide and 3–20 m deep, at the mouths of which are lobate mudflows that can be 10–15 m thick and can include outsized blocks of sediment that were rafted in the viscous flows.

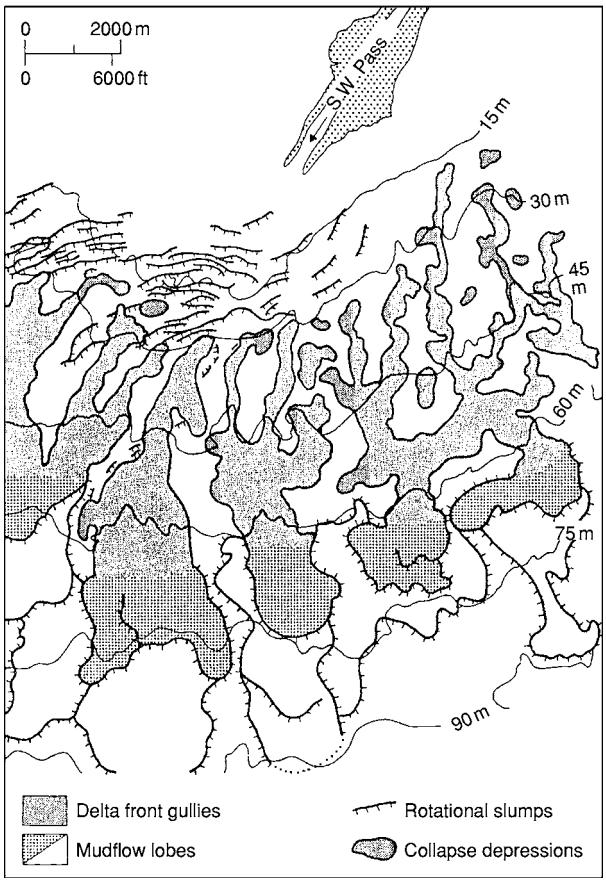


Figure 13 Map of collapse depressions, gullies, mudflow lobes, and rotational slumps in the delta front of the modern Mississippi delta. Reproduced with permission from Suhayda and Prior (1978).

Low-angle rotational slides, analogous to subareal landslides, are also common features, and it is this process that led to the loss of the drilling rig in 1969 (see earlier). Detachment surfaces dip at 1–4° in the headward regions then flatten into slope parallel surfaces. Blocks of sediment slide basinwards and may be preserved either as coherent, undeformed slides or as slumps that are deformed internally, particularly as they come to rest.

Deeper Seated Deformational Features

Basinward creep and gravitational sliding of overpressured/undecompacted fine-grained facies create an extensional regime along the basin margin and can lead to the formation of growth faults. Farther offshore in the toe regions of these structures, shale ridges and diapirs can occur (e.g., the Niger delta; Figure 14). The faults are listric extensional faults that mainly downthrow towards the basin, though counter-regional, landward-dipping faults also occur. Displacement varies from zero at the upper fault tip, to a maximum value in the middle part of the fault, decreasing again as the fault flattens with depth. In plan view, faults have curved traces that are concave to the basin and are of limited lateral persistence, though they may coalesce to form more extensive linked faults. The faults intersect the sediment surface and the hangingwall sides form preferred sites of deposition for delta front sediments. Thick, expanded, often sand-rich sections accumulate in the hangingwall of the faults and cause the faults to become self-perpetuating, because loading by the hangingwall section promotes further displacement. Anticlines offset by minor synthetic and antithetic faults form in the hangingwall succession. Through time, the site of active faulting migrates basinwards in concert with the overall migration of the delta system. Earlier faults experience a reduction in displacement and are eventually overlain by unfaulted strata. The vertical scale of the faults can vary from several tens of metres to kilometres. Sand-rich hangingwall anticlines are prime reservoirs in many deltaic hydrocarbon provinces (e.g., Niger delta).

Shale ridges and diapirs are compressional structures that form in the distal, offshore parts of deltas or the upper continental slope. They form where the toes

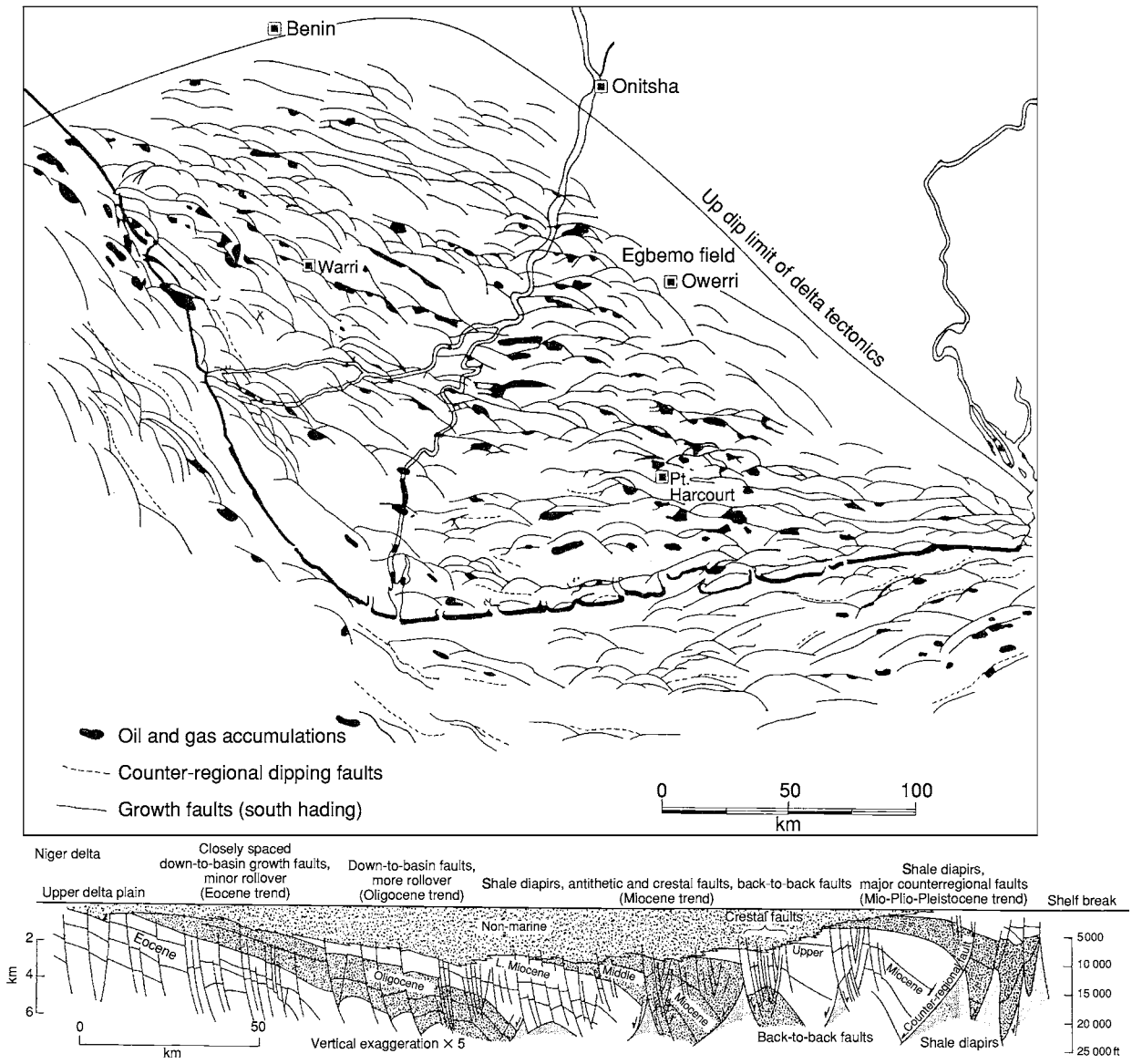


Figure 14 Map view of (top) extensional growth faults in the Niger delta and (bottom) a cross section through the delta from the upper delta plain (left) to the shelf break (right). Note the lateral impersistence of the faults, their curved traces, and general parallelism to the delta front in the map. In the cross section, note the down to basin displacement on most of the faults and the shale diapir province in the toe region of the deformation zone. Reproduced with permission from (top) Weber and Daukoru (1975) and (bottom) Winker CD and Edwards MB (1983) Unstable progradational clastic shelf margins. In: Stanley DJ and Moore GT (eds.). *The Shelfbreak: Critical Interface on Continental Margins. Special Publication the Society of Economic Paleontologists and Mineralogists*, Vol. 33, pp. 139 157. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

of listric faults cut up-section as they intersect the sediment surface, or where basinward-creeping overpressured shales intersect the surface. In the Mississippi delta, mud diapirs occur, somewhat unusually, in the delta front area, due probably to earlier shale ridges being point-loaded by distributary mouth bar sands. The diapirs appear as mudlump islands that project above sea-level by several metres. They are elevated shortly after major flood periods in response to high sediment accumulation rates and related sediment loading in the distributary mouth area and are eroded slowly by wave processes between flood periods. Mudlumps often survive between flood periods and are reactivated by the major flood, therefore becoming semipermanent features of the delta front. Sediment uplift of several hundred metres can be demonstrated and uplift rates of 100 m in 20 years have been documented. Subsurface investigations

Figure 15 A delta front gully (lower left) filled by mudflow deposits in an Upper Carboniferous deltaic succession in the Clare Basin, western Ireland. One margin of the gully is exposed on the right side, truncating earlier strata; the gully is 3 4 m deep and extends over at least 20 25 m laterally to the edge of the cliff.

Figure 16 A mud diapir in an Upper Carboniferous deltaic succession in the Clare Basin, western Ireland. Note that the bedded sandstones that overlie the diapir are thin above the crest and thicker on the flank of the diapir, indicating that the diapir grew during deposition of the sandstones. Cliff height is 25 m.

reveal that the mudlumps are growth anticlines related to reverse faults. The growth nature of the structures is depicted by thickness variations in the overlying mouth bar deposits that are thin above the diapir crest and thick in areas between diapers. Syn-sedimentary deformational features can also be observed in exposed, ancient deltaic successions (Figures 15 and 16).

Shelf-Edge Deltas in Quaternary Lowstand Periods

Presently, because of the Holocene transgression and the current highstand of sea-level, most major, ocean-facing deltas are located at the inner margin of a broad continental shelf. In contrast, during periods of sea-level lowstand associated with peak glacial in the Quaternary, these deltas were located at the outer edge of the continental shelf where they border on the continental slope. Falling sea-level forced the deltas into this position and they remained there until they retreated back across the shelf during sea-level rise associated with the onset of interglacial conditions. These shelf-edge deltas have been examined extensively in recent years using shallow, high-resolution seismic data that have demonstrated several distinctive features. The deltas are fed by widespread fluvial systems that are often incised into the former shelf (incised valleys). These systems supply large volumes of sediment to the shelf-edge deltas at a time when the accommodation space available for sediment accumulation is limited. As a result, the delta front slopes of lowstand deltas, seen as dipping clinoforms in seismic data, are typically steep (~2–4°) and frequently display evidence of large gravitational failures. Slump scars can cut 200–300 m of strata and extend laterally for several kilometres. Instantaneous bathymetric relief associated with the scars can be as great as 30 m and the scars may grow further by subsequent faulting. Slump scars therefore have the potential to release significant volumes of sediment to the deep basin as slides, slumps, debris flow, or turbidity currents, and to define minibasins that are repaired by subsequent deltaic sedimentation. Seismic data from these slump scar minibasins indicates that the repair of the scars is rarely simple. Healing by sediment accumulation alternates with renewed sediment failure, and the evolution of the scar fill can be critical in understanding the delivery of sediment to the deeper water setting during lowstand periods.

Economic Aspects

Deltaic deposits are commonly preserved in the geological record as thick, expanded sections of predominantly clastic sedimentary facies. As already mentioned, these deposits are of great economical importance. They contain vast reserves of hydrocarbons, both solid, in the form of coal, and fluid, as gas and oil (*see* **Petroleum Geology:** Overview). Because deltaic sediments are commonly rapidly deposited, with little reworking, organic matter is preserved in them. When buried and heated, this organic matter may generate oil and gas. In humid climates, the waterlogged surfaces of deltas are colonized by lush vegetation that is preserved in anaerobic marshes and swamps. These environments are sites of formations of peat that can accumulate to significant

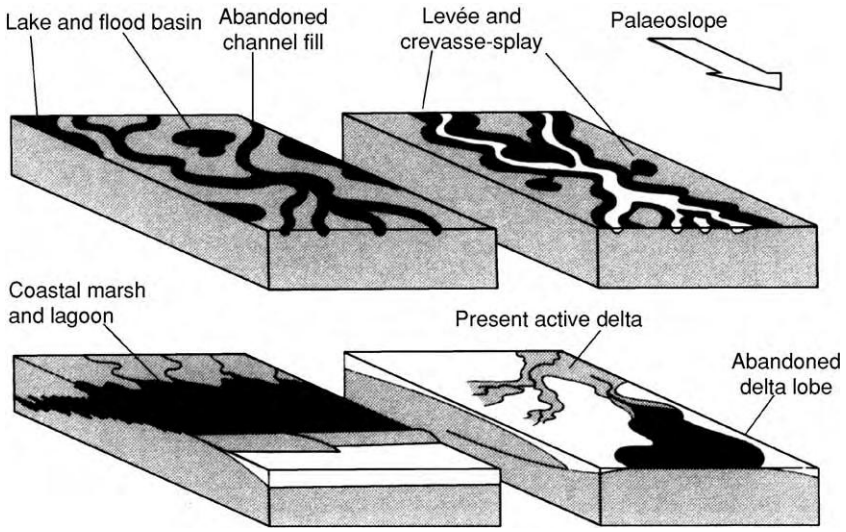


Figure 17 Sketches illustrating the common occurrence of beds of coal in deltaic settings. Note that to predict the geometry of an individual coal bed it is important to diagnose the subenvironment, rather than the overall depositional environment of the associated sediments. Channel abandonment coals, for example, occurs in fluvial, deltaic, and intertidal settings. Note also that coal geometry is controlled not only by its depositional environment, but also by subsequent erosion. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

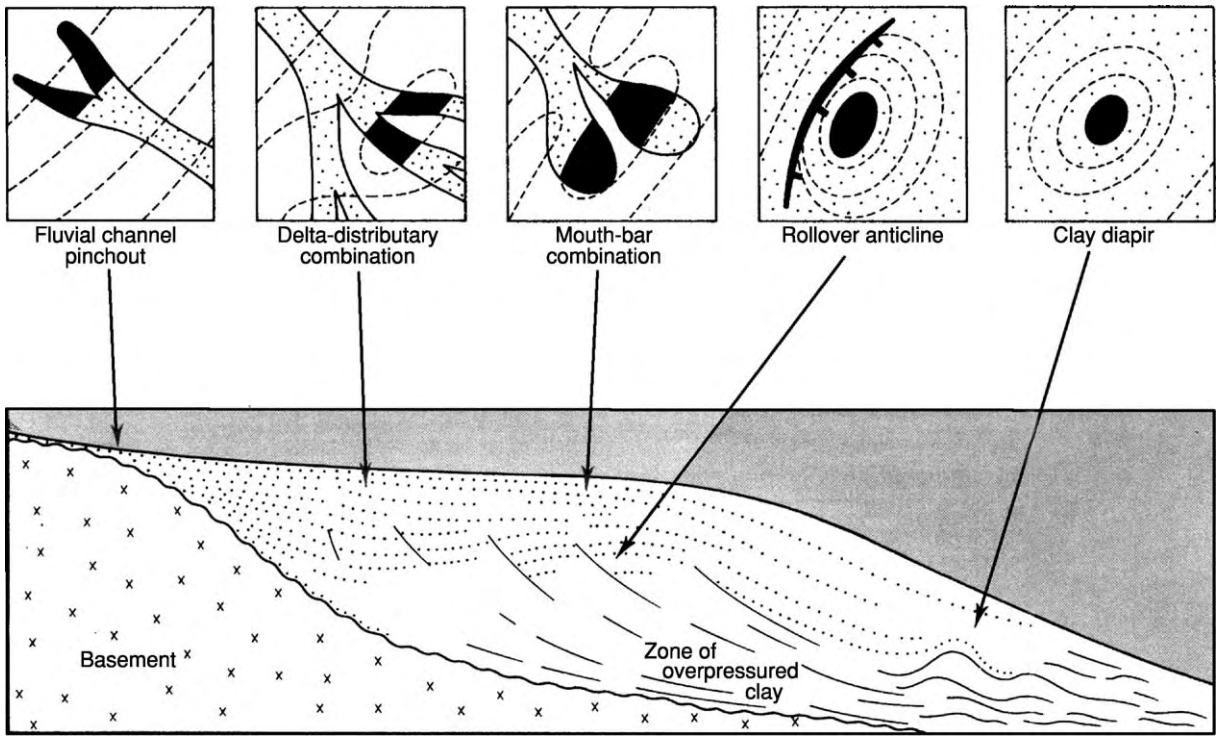


Figure 18 Diagrammatic cross section to illustrate common trapping styles in deltaic petroleum systems. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

thicknesses due to the high subsidence rates that prevail in deltas. When beds of peat are buried and heated, they metamorphose into lignite (brown coal), then bituminous coal, and finally anthracite. The conversion of peat into moderate- to high-grade coal involves a compaction ratio of ~10:1. Most coal beds are laterally continuous and relate to episodes of delta abandonment. In plan view, the geometry of coal beds may be irregular and is closely controlled by the depositional environment of the parent marsh (Figure 17). However, major coal beds in deltaic successions often relate to episodes of delta abandonment and are therefore laterally continuous coal beds that blanket underlying deposits. Most of the coal deposits of the world occur in deltaic deposits of Carboniferous age, and most of the lignite (brown coal) deposits of the world occur in Tertiary deltas. Deltas are often major petroleum systems (*see* **Petroleum Geology:** The Petroleum System). This is because a delta is a natural mechanism for depositing large volumes of sand (potential petroleum reservoirs) that prograde out over organic-rich muds (potential source rocks). Furthermore, deltas generate their own petroleum traps. These include the diapiric mud lumps, growth faults, and roll-over anticlinal structures. Deltas also contain a wide range of potential stratigraphic traps, such as sand-filled channels and delta front bars (Figure 18). The Tertiary–Holocene Niger and Mississippi deltas are major petroleum systems that illustrate these juxtapositions of source and reservoir rocks, and this range of trapping styles.

See Also

Petroleum Geology: Overview; The Petroleum System. **Sedimentary Environments:** Depositional Systems and Facies; Alluvial Fans, Alluvial Sediments and Settings.

Further Reading

Boyd R and Penland S (1988) A geomorphologic model for Mississippi delta evolution. *Transactions of the Gulf Coast Association of Geology Societies* 28: 443 452.

Coleman JM, Suhayda JN, Whelan T, and Wright LD (1974) Mass movement of Mississippi River delta sedi ments. *Transactions of the Gulf Coast Association of Geological Societies* 24: 49 68.

Morton RA and Suter JR (1996) Sequence stratigraphy and composition of shelf margin deltas, northern Gulf of Mexico. *American Association of Petroleum Geologists* 80: 505 530.

Mulder T and Syvitski JPM (1995) Turbidity currents generated at river mouths during exceptional dis charges to the World oceans. *Journal of Geology* 103: 285 299.

Nemec W and Steel RJ (eds.) (1988) *Fan Deltas: Sedimen tology and Tectonic Settings*. London: Blackie.

Orton GJ and Reading HG (1993) Variability of deltaic processes in terms of sediment supply, with particular emphasis on grain size. *Sedimentology* 40: 475 512.

Reading HG and Collinson JD (1996) Clastic coasts. In: *Sedimentary Environments and Facies*, 3rd edn., ch. 6, pp. 154 231. Oxford: Blackwell Scientific Publications.

Roberts HH, Rosen NC, Fillon RH, and Anderson JB (2003) *Shelf Margin Deltas and Linked Down Slope Pet roleum Systems: Global Significance and Future Explor ation Potential*. Proceedings of the 23rd Annual GCSSEPM Foundation Bob F. Perkins Research Confer ence. Available on CD ROM at www.sepm.org.

Winker CD and Edwards MB (1983) Unstable prograda tional clastic shelf margins. In: Stanley DJ and Moore GT (eds.) *The Shelfbreak: Critical Interface on Continental Margins. Special Publication the Society of Economic Paleontologists and Mineralogists*, Vol. 33, pp. 139 157. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Wright LD (1977) Sediment transport and deposition at river mouths: a synthesis. *Bulletin of the Geological Society of America* 88: 857 868.

Deserts

N P Mountney, Keele University, Keele, UK

Introduction

Deserts are arid regions in which the annual precipitation is less than half of the annual potential evapotranspiration (a measure of moisture loss). Approximately 30% of the present-day land surface of the Earth is characterized by desert conditions, with hyper-arid regions restricted primarily to parts of the Saharan, Arabian, Atacama, and Namib Deserts (**Figure 1**). The world's largest deserts, including the Saharan, Arabian, and Australian, are located in tropical tradewind belts that are characterized by the atmospheric circulation of relatively moisture-free air. Other deserts, for example the Sonoran of the

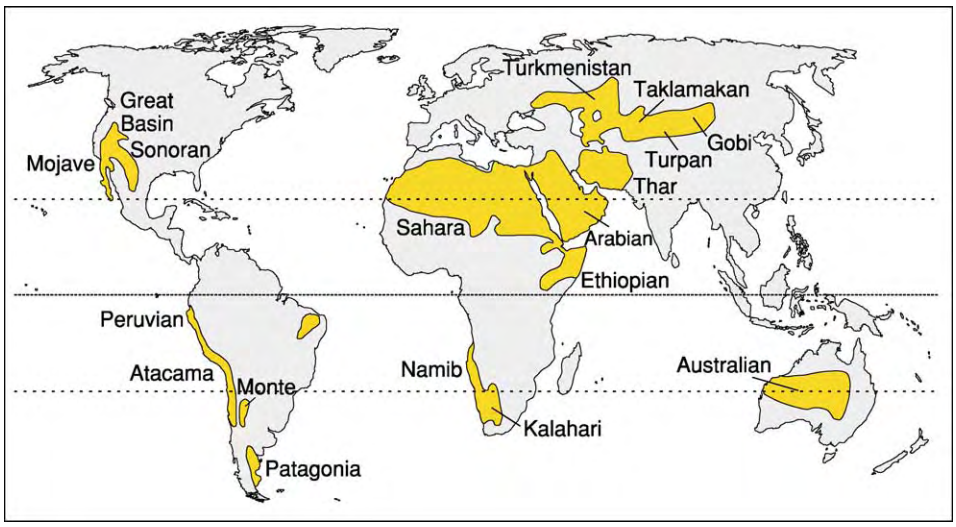


Figure 1 Distribution of the world's major climatic deserts.

south-western USA, are located in continental interiors away from humid maritime climates. The Namib and Atacama are coastal deserts in which aridity is promoted by cold offshore currents, whilst rain-shadow deserts, such as the Turpan of China, occur in the lee of mountain ranges. Deserts are not necessarily characterized by high temperatures; high-atitude polar deserts cover significant parts of Antarctica.

Deserts are composed of a variety of erosional and depositional geomorphic elements (Figure 2). Predominantly erosional mountainous and badland regions cover in excess of one-third of desert areas, whilst predominantly depositional regions composed of sand dunes, interdunes, sandsheets, playa lakes, sabkhas (salt flats), ephemeral and occasionally perennial rivers, and alluvial fans cover the remaining area. Where actively migrating sand dunes occur collectively, they form 'sand seas' or 'ergs', which cover approximately 20–30% of the surface area of the world's major deserts. The erosion, transport, and deposition of sediment in deserts are dominated by aeolian and fluvial processes, although chemical processes, such as salt precipitation, may be important locally.

The morphology and sedimentary characteristics of modern and recent desert environments, together with an understanding of the mechanisms by which these systems accumulate and become preserved, provide the basis for the recognition and interpretation of ancient outcrop examples. The architecture of ancient desert sedimentary systems can be summarized in the form of a depositional model.

Modern and Recent Desert Systems

Dunes and Sand Seas

Aeolian (wind-blown) dune bedforms are common in deserts in which the airflow is saturated with sediment. Dunes adopt a variety of morphologies (Figure 2) depending on factors such as the strength and directional variability of the wind, the size of sediment grains in transport, and whether the airflow undergoes downwind deceleration. Dune flanks and stoss (windward) slopes are characterized by wind-ripple strata, whilst steeper lee slope 'slipfaces', which are subject to periodic gravitational collapse, are characterized by grainflow (avalanche) and grainfall strata. Dunes occur either as individual, spatially isolated sand bodies, or they form collections ('trains') of bedforms that are arranged into sand sea accumulations, the morphologies of which are determined by the migration of hierarchies of bedforms of differing sizes and shapes, moving at varying rates and in varying directions relative to one another. Sand seas range in size from a few square kilometres to 560 000 km^2 in the case of the Rub al Khali erg in Arabia. Downwind changes in dune type across sand seas are common. For example, the Great Sand Dunes of Colorado are characterized by a zone of small, spatially isolated, partly cemented dunes at their upwind margin, a central zone of undulating barchan, parabolic, and transverse dunes with intervening interdunes, and a zone of large, actively accumulating transverse and star dunes separated by

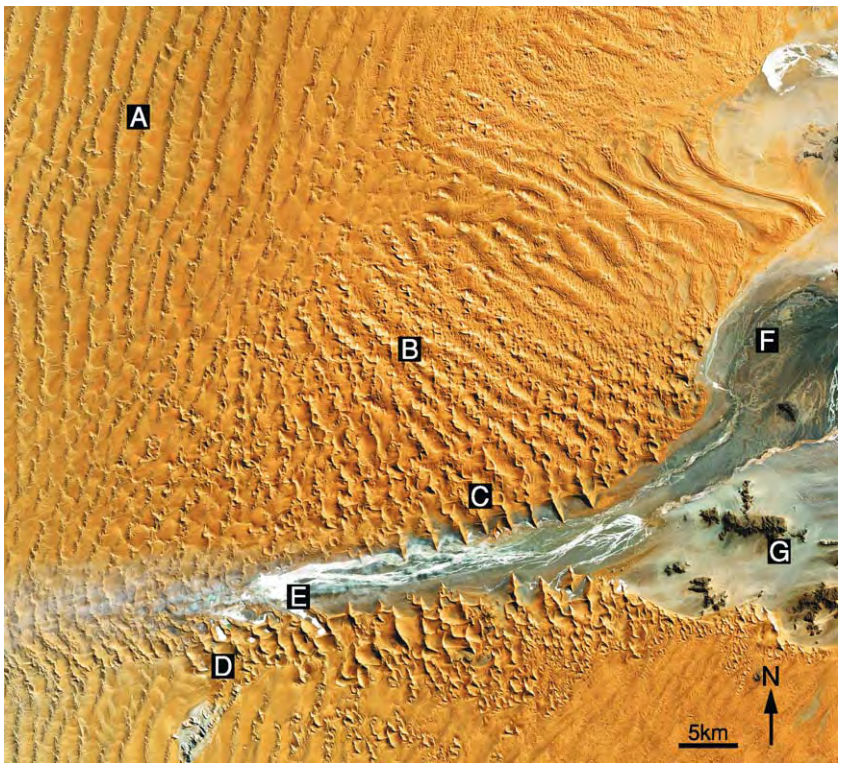


Figure 2 The Sossusvlei region of the eastern Namib sand sea, Namibia. (A) Complex linear bedforms (draa) with superimposed transverse ribs. Net sand transport is from SSW to NNE. (B) Complex star (pyramid) bedforms with radiating arms and superimposed sinuous dunes. (C) Linear dune ridges. (D) Mosaic of star (pyramid) bedforms with isolated interdune hollows. White colour represents salt and calcrete deposits precipitated following the evaporation of ponded water from wet interdunes. (E) Salt flats in broad interdune corridor. (F) Ephemeral alluvial braid plain. (G) Eroding basement with an alluvial fan apron. Image from NASA Earth Observatory Collection, with permission.

only minor interdune depressions at their downwind margin.

Interdunes

Interdune flats and hollows occur between aeolian dunes (Figure 2) and are classified as dry, damp, or wet. 'Dry interdunes' occur where the depositional surface exhibits no evidence of sedimentation that is influenced by moisture. Dry interdunes are dominated by wind-ripple strata, aeolian plane-beds, and granule and pebble lags left behind after aeolian 'deflation'. 'Damp interdunes' occur where the depositional surface is in contact with the capillary fringe of the water table and sedimentation is influenced by the presence of moisture. Damp interdunes are dominated by a range of adhesion structures where sand sticks to the damp surface and by minor salt precipitation structures. 'Wet interdunes' occur where the depositional surface is continuously or periodically inundated by water due to elevation of the water table, flash flooding within the dune-field, or overbank flooding from beyond the margins of the dune-field. Mud-, silt-, and sand-grade sediment may be supplied to these interdunes via a combination of aeolian and fluvial processes. Sedimentary structures associated with such settings include subaqueous current and wave ripples, wavy and contorted bedding, desiccation cracks, raindrop impressions, hardpan crusts, and mud flakes. Carbonate sediments are known to accumulate in some long-lived ponds. Near-surface moisture in damp and wet interdunes encourages plant and animal colonization, and a variety of bioturbation structures, including roots, burrows, and trackways, are common in these settings. The geometry of interdune elements is determined by the spacing and plan-view shape (morphology) of adjoining dunes, and varies from spatially isolated hollows that are completely surrounded by dunes, to narrow but elongate corridors that extend for several kilometres between rows of transverse or linear dunes, to extensive interdune flats completely surrounding isolated bedforms. Interdunes typically decrease in

size and degree of interconnectedness from the margin to the centre of a sand sea as surrounding dunes increase in size.

Sandsheets

'Sandsheets' are areas of wind-blown sand that lack high-relief bedforms, but are instead characterized by wind ripples, plane-beds, and sometimes by low-relief ridge- and dome-like features called 'zibar'. Sandsheets develop where aeolian sand supply and its availability for transport are restricted to the extent that dunes do not develop. Additionally, the presence of stabilizing factors, such as vegetation or a shallow water table, can act to restrict sand movement and promote accumulation. Sandsheets subject to erosion are often characterized by an armoured lag of granules and pebbles left behind following aeolian deflation of finer sand-grade sediment. Individual sandsheets vary in extent from localized sand patches ($<1\,km^2$) to major geomorphic features, such as the $>100\,000\,km^2$ Salima Sandsheet of the eastern Sahara.

Sabkhas, Playa Lakes, and Salt Flats

'Sabkhas' are low-relief flats in which accumulation occurs wholly or partly as a result of evaporite precipitation (and in some cases carbonate sedimentation). The term 'sabkha' was originally used exclusively for the description of salt flats in coastal desert settings, but is now also widely used for the description of inland salt flats (Figure 2), which often border evaporitic 'playa lake' basins. Sabkha sedimentation usually involves interactions between chemical (precipitate) and aeolian processes, and results in the generation of a variety of wavy and crinkly laminae that are often disturbed by salt growth structures such as 'teepees'. Salt precipitation in sabkhas requires periodic wetting and subsequent desiccation of the surface, and is often controlled by subtle water table changes which, in coastal settings, may be driven by sea-level change, as is the case in the Dhahran Sabkha of Saudi Arabia.

Desert Fluvial Systems

Ephemeral rivers that typically flow for no more than a few days each year are common in many desert systems and are usually fed by high-intensity but short-duration and low-frequency rainfall events. Flash flood events mean that the infiltration capacity of the substrate is rapidly exceeded, resulting in overland flow. Desert stream flood hydrographs are markedly peaked, exhibiting a rapid rise to peak discharge, followed by a rapid fall. High peak discharge rates, together with a lack of vegetation cover and a substrate that is often covered with loose sediment, mean that fluvial transport and the deposition of sediment in desert regions are important geological processes. A lack of vegetation and a high proportion of non-cohesive sandy sediment result in a low degree of bank stability, which, when coupled with high rates of bedload transport, tends to favour the development of unconfined, braid-plain fluvial systems, such as those encountered in Namibia (Figure 2). Many desert rivers exhibit a downstream reduction in discharge due to transmission losses through a porous, sandy substrate. 'Endogenic' desert streams have catchment areas within the boundaries of the desert region, whilst the catchment area of 'allogenic' streams lies beyond the confines of the desert region, and such rivers may therefore be perennial in nature and provide a valuable water resource in an otherwise arid environment.

Alluvial Fans

'Alluvial fans' are depositional landforms that occur where confined streams fed by mountain catchments emerge, often via a narrow feeder canyon, onto a low-relief plain. A decrease in gradient causes an abrupt reduction in stream power at the mountain front, which promotes deposition and the build-up of a radial fan or cone of sediment. Desert alluvial fans are ephemeral systems with sediment transport only occurring during and in the immediate aftermath of flash flood events. Sediment transport may occur either as a sheetflood or as channellized flow, in which case channels on the fan surface often develop a distributary (downstream splitting) architecture. Alluvial fan sediments are generally coarse grained and exhibit a proximal to distal fining that reflects a downstream loss of stream power and resultant transport capacity. Proximal fan deposits are typically chaotic and structureless, often being the product of debris flows, whilst more distal fan deposits usually form more organized cross-bedded units that reflect a stream flow origin. At any given time, alluvial fan sedimentation is typically restricted to one particular part of the fan surface called a 'depositional lobe'. However, over time, lobes become choked with sediment, and eventually it becomes more efficient for the sedimentation process to 'avulse' (jump) to a new lobe. Alluvial fans often develop adjacent to active normal faults, the movement of which provides the mechanism that maintains the high gradient necessary for fan development. Where faults control fan activity, as is the case in Death Valley, California, a line of overlapping fans develops into a linear apron or 'bajada'.

Accumulation and Preservation of Aeolian Strata

The preservation of desert aeolian systems in the rock record is a three-stage process that requires: (1) the construction of a sand sea, (2) its accumulation to form a succession, and (3) its long-term burial.

Sand Sea Construction

Aeolian sand sea construction is a function of sediment supply, sediment availability, and the transport capacity of the wind. The 'sediment supply' is the volume of sediment of a suitable grain size for aeolian transport generated per unit time. This sediment may form either a contemporaneous or time-lagged source of material with which to construct an aeolian system, and may be derived from a variety of pre-existing aeolian or non-aeolian sources. 'Sediment availability' is the susceptibility of surface grains to aeolian entrainment, and is influenced by stabilizing factors, such as the presence of vegetation, mud drapes, coarse-grained lags, an elevated water table, or surface binding and cementing agents. The 'transport capacity of the wind' is a measure of the potential sand-carrying capability of the wind, which increases with wind power. Where sediment availability is limited, the airflow will be undersaturated with respect to its potential sediment load, and the wind will be potentially erosional. Conversely, an airflow that is fully saturated with sediment and undergoes deceleration must drop some of its load, thereby encouraging aeolian bedform growth.

Accumulation

'Accumulation' to generate a body of strata occurs when an aeolian system experiences a positive 'net sediment budget', such that upstream sediment 'influx' exceeds downstream 'outflux' and the accumulation surface rises over time. Accumulation of migrating bedforms occurs as a consequence of 'bedform climb' with respect to the accumulation surface. The 'angle of climb' is determined by the ratio between the rate of downwind bedform migration and the rate of rise of the accumulation surface (Figure 3). For most aeolian dune systems, the accumulation rate is small compared with the migration rate and the resultant angle of climb is low, such that, as bedforms move through space, they truncate the upper parts of the preceding bedforms in a train, and only the basal parts of the bedforms accumulate to form sets of cross-strata, a process called 'subcritical climbing'.

Bounding Surfaces in Aeolian Strata

'Bounding surfaces' are erosional surfaces that are generated as an intrinsic product of aeolian dune migration and climb, such that bedforms (or parts thereof) scour into pre-existing deposits as they move through space. Three broad types of aeolian bounding surface are recognized to occur as a product of 'autocyclic' (intrinsic) migratory bedform behaviour. 'Reactivation surfaces' result from periodic erosion of a dune lee slope, followed by renewed sedimentation as a consequence of a change in bedform migration direction, migration speed, asymmetry, or steepness (Figure 4). These changes are common because the airflow on lee slopes is often subject to modification and is rarely steady. In some cases, the period of the flow fluctuation is regular and generates cyclic reactivation surfaces, for example due to diurnal or seasonal wind reversals. Nested reactivation surfaces on two or more scales occur when cyclic sets of cross-bedding separated by reactivation surfaces are generated by the interaction of two or more forcing parameters operating with different periodicities. 'Superimposition surfaces' result from either the migration of superimposed dunes over a larger parent bedform, or the

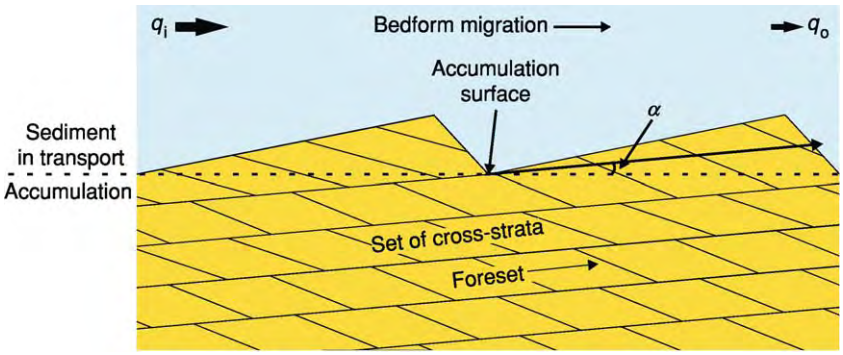


Figure 3 Generation and accumulation of sets of cross strata by migrating and climbing bedforms. Accumulation requires a positive net sediment budget, whereby the influx (q_i) exceeds the outflux (q_o) and the accumulation surface rises through time. The ratio between the rate of rise of the accumulation surface and the rate of downwind migration of the bedforms determines the angle of climb (α) of the accumulated sets of cross strata.

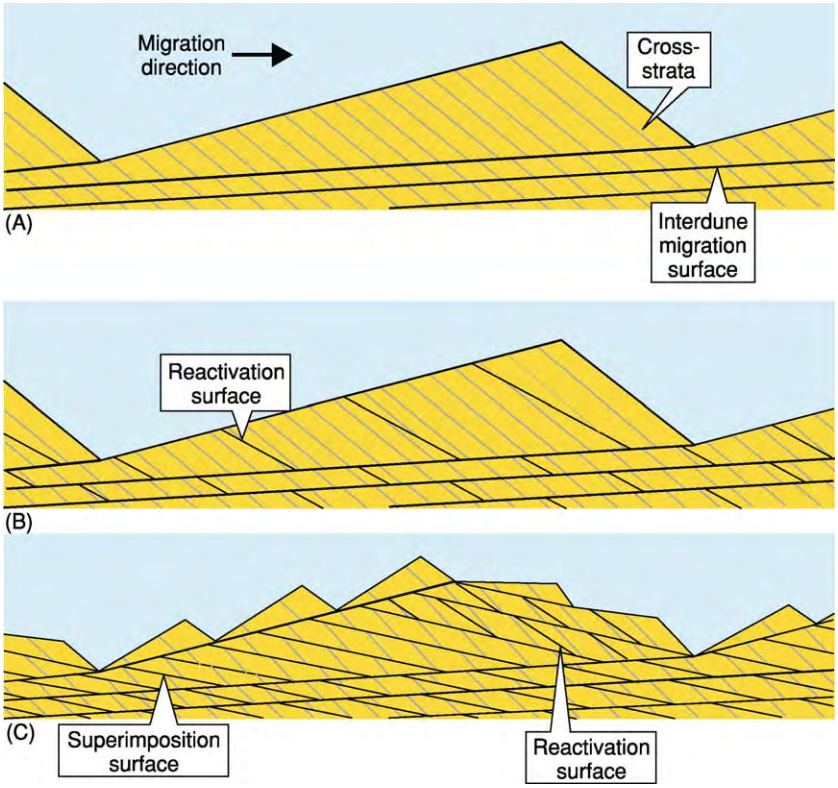


Figure 4 Generation of bounding surfaces in aeolian strata as a consequence of bedform migration. (A) Interdune migration surfaces. (B) Reactivation surfaces. (C) Superimposition surfaces arising from the migration of small bedforms over more slowly migrating parent bedforms. Reactivation surfaces may be nested within sets bounded by superimposition surfaces.

migration of scour troughs on the lee slope of a bedform (Figure 4). Although superimposed dunes and scour troughs can theoretically migrate directly up or down the lee slope of a parent bedform, oblique migration is more common because of secondary airflow along the lee slope of the parent bedform. Where both reactivation and superimposition surfaces are developed, the latter always truncate the former. 'Interdune migration surfaces' result from the migration of bedforms separated by interdunes (Figure 4). The surfaces are carved by the passage of an erosive scour that defines the interdune trough between successive bedforms. The depth to which the interdune trough scours as it migrates influences the extent to which deposits of the preceding bedform are eroded. Interdune surfaces, which truncate both superimposition and reactivation surfaces, typically extend continuously downwind for distances of hundreds of metres to several kilometres.

Types of Desert Aeolian System

In 'dry aeolian systems', the water table lies substantially below the accumulation surface, such that moisture plays no role in influencing sedimentation, and deposition occurs as a result of aerodynamic configuration alone. Bedforms in dry systems only commence climbing (i.e. accumulating) once they have grown to the point at which interdune flats have been reduced to isolated interdune depressions. In 'wet aeolian systems', the water table or its capillary fringe is in contact with the accumulation surface, such that moisture influences sedimentation, and deposition occurs as a result of both aerodynamic configuration and moisture content determined by the level of the water table. The angle of climb in wet aeolian systems is determined by the ratio between the rate of water table rise and the rate of downwind bedform migration, and such accumulations tend to be characterized by downwind climbing dune strata separated by climbing damp interdune units. In 'stabilizing aeolian systems', factors such as vegetation and surface cementation influence sedimentation, and deposition occurs as a result of both aerodynamic configuration and the degree and type of surface stabilization. For example, vegetation acts to disrupt the primary airflow, leading to a reduction in sand transport capacity and thereby a promotion of accumulation.

Supersurfaces and the Preservation of Aeolian Sequences

Aeolian accumulation ceases when the net sediment budget switches from positive to neutral or negative. Under such circumstances, a positive angle of bedform climb is replaced either by a zero angle of climb ('bypass'), when the accumulation remains static over time, or by a negative angle of climb ('deflation'), when an undersaturated airflow cannibalizes the dunes and the accumulation surface falls over time (Figure 5). Both bypass and deflation result in the generation of a 'supersurface', a fourth type of bounding surface that caps the underlying accumulation. The accumulation defines a 'sequence' and its bounding supersurface defines a 'sequence boundary'. Distinguishing supersurfaces in aeolian strata from other types of bounding surfaces is often difficult. Sedimentary features associated with supersurfaces include desiccation cracks, polygonal fractures, bioturbation, rhizoliths (calcified roots), and salt structures, all of which yield important palaeoenvironmental information regarding the nature of the desert surface at the time of supersurface formation. Supersurfaces often have great lateral extent and continuity and may bound entire sand sea successions. Thus, supersurfaces, which often result from changes in 'allogenic' (external) controlling factors, such as climate or tectonic setting, are considered to be of a higher order than the autogenically generated bounding surfaces that they truncate.

Long-term preservation of an aeolian succession requires that the body of strata is placed below some regional 'baseline of erosion', beneath which erosion does not occur, the principal agents promoting preservation being subsidence, water table rise, sea-level rise, and surface stabilization.

Ancient Desert Systems

Ancient aeolian desert systems exposed in outcrop are widely recognized from all seven continents, and a sizeable database of case examples now exists. In particular, the extensive Permian to Jurassic outcrops of the Colorado Plateau in the western USA (Figure 6) have been the focus of intensive study in recent years.

Dry Aeolian Systems

Dry aeolian systems, in which accumulation is controlled solely by aerodynamic configuration, are widely documented in the rock record. Accumulation in such systems occurs once dunes have grown to cover former interdune flats. The angle of climb of the system is determined by the ratio between the rate of accumulation (i.e. rise of the accumulation surface) and the rate of bedform migration. Steep angles of climb enable larger proportions of the original bedforms to be preserved, although climb angles rarely exceed 1° and typically only the basal 10–15% of a climbing bedform is preserved as a dune set (Figure 3). Thus, the facies preserved within aeolian dune sets record the processes that operated on the lower part of the dune lee slope at the time of accumulation, and are dominated by wind-ripple, grainflow, and grainfall strata. Additionally, superimposition and/or reactivation bounding surfaces may be present within climbing dune sets bounded by interdune migration surfaces. Supersurfaces within dry aeolian systems develop as a consequence of changed aerodynamic conditions, with deflation usually occurring when the airflow passing over a dune-field is not fully laden with sediment. The Page Sandstone of Utah and northern Arizona is a dry aeolian system represented by separate supersurface-bounded sequences, each composed of climbing dune strata and an

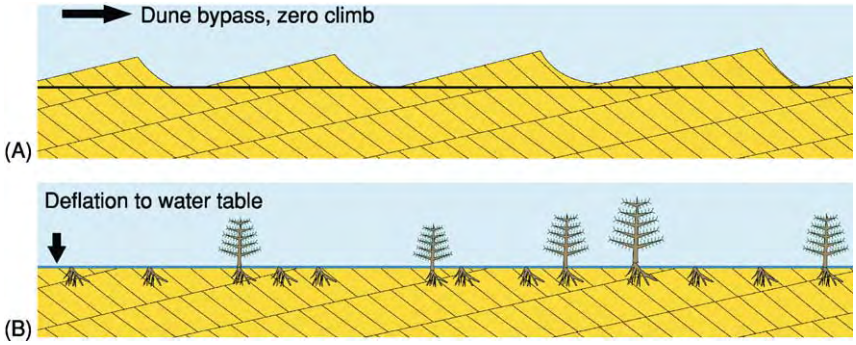


Figure 5 Models for the generation of supersurfaces in aeolian systems. (A) Bypass supersurfaces occur where the angle of dune climb reduces to zero, although dune migration may continue. (B) Deflationary supersurfaces occur as a result of erosion, often to the level of the water table. Such surfaces may exhibit evidence of widespread plant and/or animal colonization.

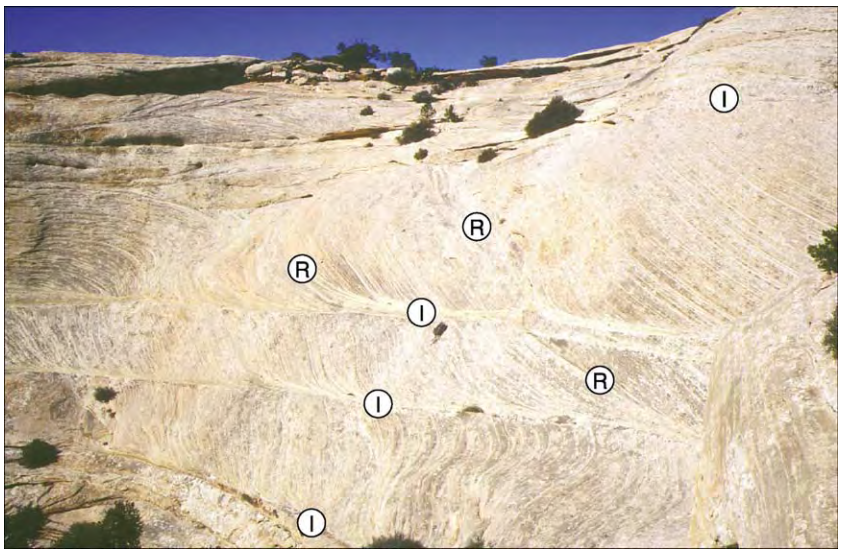


Figure 6 Sets of climbing aeolian dune strata bounded by interdune migration surfaces (I). Sets are composed of grainflow and grainfall strata that pass down into low angle inclined wind ripple strata at the set base. Numerous reactivation surfaces (R) are evident. Permian Cedar Mesa Sandstone, south east Utah, USA. Rucksack is 0.8 m.

absence of extensive interdune flat strata. The aeolian succession interfingers with transgressive marine units of the neighbouring Carmel Formation, indicating a coastal dune setting. In many cases, the supersurfaces are overlain by sabkha and subaqueous strata and exhibit great lateral extent. They are characterized by polygonal fractures and a corrugated, erosional relief. The aeolian-dominated sequences record episodic, punctuated accumulation and, rather than being stacked vertically, occupy laterally variable palaeo-depocentres. Episodes of accumulation of dry aeolian strata occurred because of a plentiful sand supply during periods of marine regression when the exposed coastal shelf acted as the sand source. Deflation to the water table and supersurface generation occurred because of reduced aeolian sand supply during marine transgression.

Wet Aeolian Systems

Wet aeolian systems, in which accumulation is controlled by a progressive water table rise coinciding with ongoing aeolian activity, are less widely recognized in the rock record than dry systems. Ancient wet aeolian systems are characterized by cross-stratified aeolian dune sets that are separated by interdune units exhibiting evidence of accumulation at or close to the water table. A progressive water table rise that occurred synchronously with dune migration in such systems is indicated by intertonguing of the toesets of advancing dune strata and strata that accumulated in the adjacent interdune. Accumulation controlled by a water table rise is indicated by a downwind rise (climb) through the stratigraphy of both dune sets and intervening interdune units, as seen, for example, in the Jurassic Entrada Sandstone, Utah, USA (Figure 7). Changes in either the rate of water table rise or the rate of dune migration cause the angle of climb to vary. The relative level of the water table is determined by both absolute water table changes, as governed by the climatic regime, and by subsidence, whereby the accumulation can subside through a static water table. Increases and decreases in the rate of relative water table rise cause interdunes to expand, and contract, respectively, at the expense of dunes. Where climatic regimes fluctuate on a cyclical basis, interdune units preserved in wet aeolian successions reflect this change (Figure 7), as observed, for example, in the Triassic Helsby Sandstone of north-west England. As interdune units expand and contract through time, so conditions at the depositional surface vary from dry, through damp, to wet and back in response to subtle variations in the level of the water table relative to the accumulation surface. Bypass supersurfaces are generated in wet aeolian systems when the relative water table remains static, whilst deflationary supersurfaces may occur as a consequence of relative water table fall.

Stabilizing Aeolian Systems

Stabilizing aeolian systems, in which accumulation occurs because agents such as vegetation or cementation act to restrict the availability of sediment for

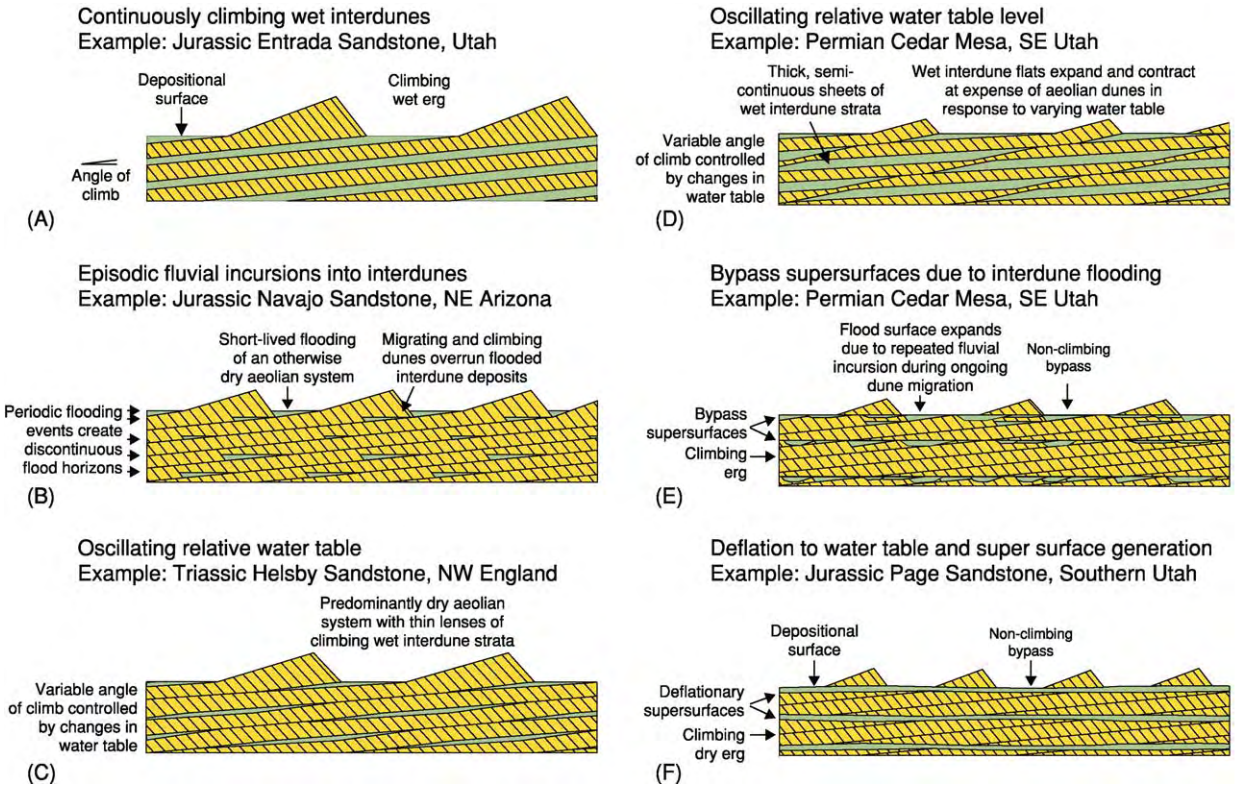


Figure 7 Common styles of dune interdune architecture in preserved aeolian desert successions. Interactions can potentially occur on a variety of scales. Adapted with permission from Mountney NP and Thompson DB (2002) Stratigraphic evolution and preservation of aeolian dune and damp/wet interdune strata: an example from the Helsby Sandstone Formation, Cheshire Basin, UK. *Sedimentology* 49: 805 833.

transport, are not widely recognized in the rock record. Accumulation of such systems is determined by the rate of emplacement of the agent responsible for stabilization. Colonization of dunes by vegetation is the most common modern stabilizing agent and is also considered to be responsible for enabling the accumulation of some ancient successions. Part of the Tertiary Tsondab Sandstone that underlies much of the modern Namib Desert is characterized by sets of cross-stratified dune strata that contain abundant plant root structures, which are considered to have both decreased near-surface wind speed (thereby promoting sediment accumulation) and acted as a sand-binding agent (thereby preventing deflation). Prior to the development of widespread land-based vegetation in the Devonian, chemical and physical factors, rather than biogenic factors, were the main stabilizing agents. Surface and near-surface cementation by chemical precipitates, a high water table, periodic flooding, and the presence of coarse-grained lags all encourage accumulation in warm-climate stabilizing aeolian desert systems, whilst permafrost may play an important role in enabling accumulation in cold-climate systems.

Depositional Models for Desert Systems

Preserved desert accumulations exhibit complex transitions on a variety of scales between the various subenvironments that comprise the arid climate depositional system for two main reasons (Figure 8). Firstly, at a relatively small scale, the various neighbouring depositional elements present within desert systems often undergo contemporaneous accumulation, resulting in the generation of complex interfingering architectures. These types of interactions are examples of 'autocyclic' (intrinsic) behaviour within the sedimentary system, reflecting episodic switching of sedimentary processes. For example, aeolian erg margin systems often interfinger with ephemeral fluvial and alluvial fan systems, such that aeolian and fluvial processes are in direct interaction. Interdune corridors periodically act as conduits for flood waters and, in turn, dry river beds provide a source of sand for the construction of aeolian dunes. Secondly, at a larger scale, sedimentary architecture is governed by shifts in key controlling factors that influence

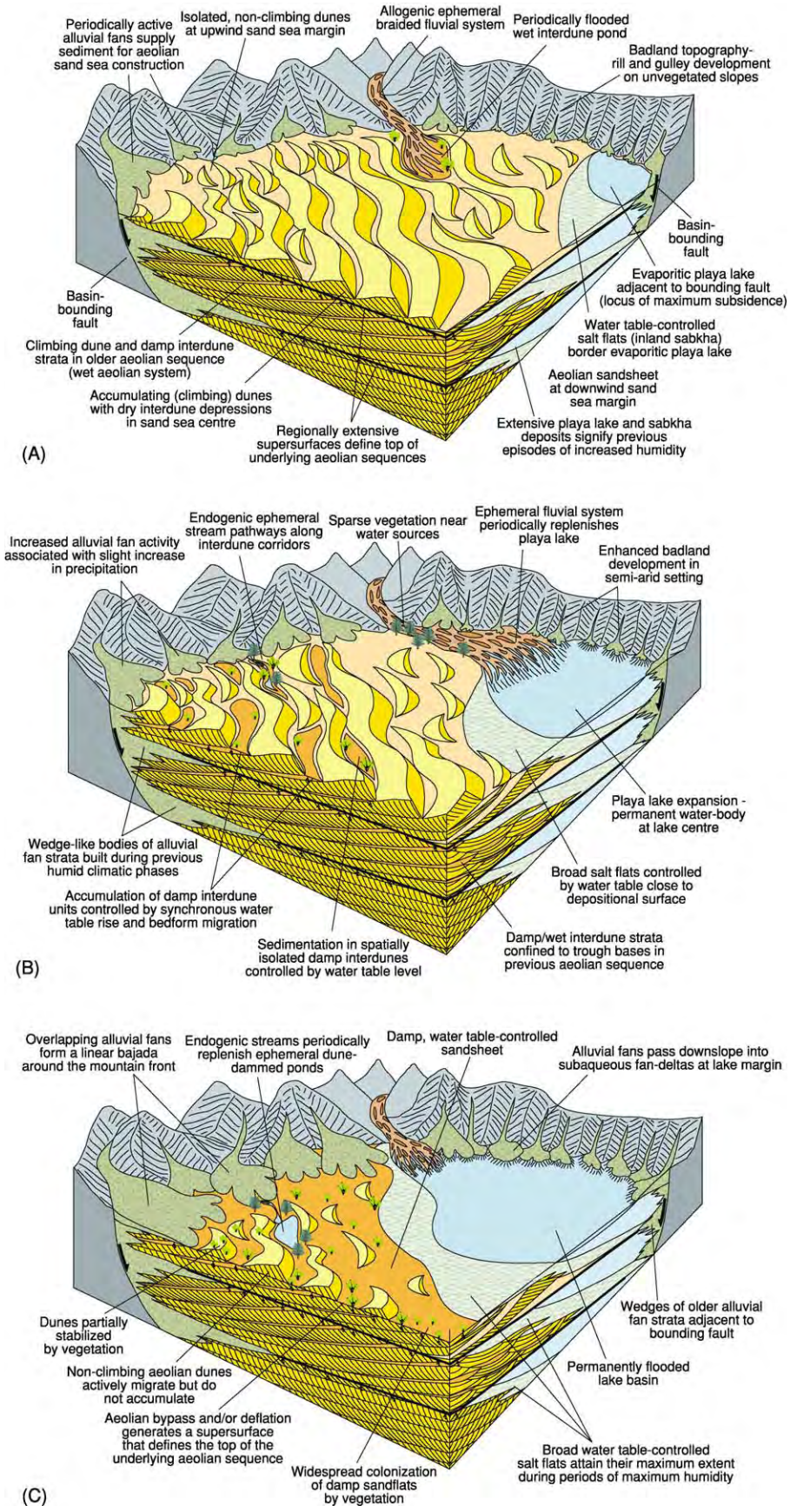


Figure 8 Depositional models for a conceptual desert system.

the gross-scale preserved architecture of the system. These 'allocyclic' (extrinsic) controls impose widespread change on the behaviour of the system. The major external controls on sedimentary architecture in desert systems are climate and tectonic basin setting. Together, these also determine the generation, supply, and availability of sediment for transport. Temporal changes in these factors will have different effects on different parts of the desert system (Figure 8). For example, a switch to a more humid climate may result in playa lake expansion due to a rise in water table. At the same time, elevation of the water table may restrict the availability of sediment for aeolian transport, causing dunes within an erg setting to stop accumulating and interdunes to expand or become flooded. Thus, various subenvironments within the desert system expand or contract, and accumulate or become eroded, over time as external controls vary. This is reflected in the preserved sedimentary architecture by large-scale shifts in the position of facies belts and the generation of surfaces of non-deposition and/or erosion (supersurfaces). In attempting to understand the complexity present within desert successions and to be able to differentiate between the extrinsic and intrinsic controls, preserved sedimentary architecture needs to be investigated on a variety of scales.

See Also

Arabia and The Gulf. **Palaeoclimates**. **Sedimentary Environments:** Depositional Systems and Facies; Alluvial Fans, Alluvial Sediments and Settings; Lake Processes and Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures; Aeolian Processes. **Sedimentary Rocks:** Evaporites.

Further Reading

Alsharhan AS, Glennie K, Whittle GL, and Kendall CGStC (1998) *Quaternary Deserts and Climate Change*. Rotterdam: Balkema.

Brookfield ME (1992) Eolian systems. In: Walker RG and James NP (eds.) *Facies Models: Response to Sea Level Change*, pp. 143 156. St John's: Geological Association of Canada.

Brookfield ME and Ahlbrandt TS (1983) *Eolian Sediments and Processes. Developments in Sedimentology 38*. Amsterdam: Elsevier.

Cooke R, Warren A, and Goudie A (1993) *Desert Geomorphology*. London: UCL Press.

Frostick L and Reid I (1987) *Desert Sediments: Ancient and Modern. Geological Society of London Special Publication 35*. London: Geological Society of London.

Kocurek G (1991) Interpretation of ancient eolian sand dunes. *Annual Review of Earth and Planetary Sciences* 19: 43 75.

Kocurek G (1996) Desert aeolian systems. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 125 153. Oxford: Blackwell Science.

Kocurek G and Havholm KG (1993) Eolian sequence stratigraphy a conceptual framework. In: Weimer P and Posamentier HW (eds.) *Siliciclastic Sequence Stratigraphy. American Association of Petroleum Geologists Memoir 58*, pp. 393 409. Tulsa, OK: American Association of Petroleum Geologists.

Kocurek G and Lancaster N (1999) Aeolian system sedi ment state: theory and Mojave Desert Kelso dune field example. *Sedimentology* 46: 505 515.

Lancaster N (1995) *Geomorphology of Desert Dunes*. New York: Routledge.

McKee ED (1979) *A Study of Global Sand Seas. Geological Survey Professional Paper 1052*. Washington: US Government Printing Office.

Mountney NP and Thompson DB (2002) Stratigraphic evo lution and preservation of aeolian dune and damp/wet interdune strata: an example from the Helsby Sandstone Formation, Cheshire Basin, UK. *Sedimentology* 49: 805 833.

Pye K (1993) *The Dynamic and Environmental Context of Aeolian Sedimentary Systems. Geological Society of London Special Publication 72*. London: Geological Society of London.

Pye K and Lancaster N (1993) Aeolian sediments, ancient and modern. *International Association of Sedimentologists* 16. Oxford: Blackwell Science.

Rubin DM (1987) Cross bedding, bedforms and palaeo currents. *SEPM Concepts in Sedimentology and Paleontology* 1.

Lake Processes and Deposits

M R Talbot, University of Bergen, Bergen, Norway

Introduction

Lakes range in area from less than 1 km^2 to the *ca.* 374 000 km^2 of the modern Caspian Sea and in depth from less than 1 m to 1741 m in Lake Baikal. The salinity of lake waters ranges from fresh to salt concentrations several times that of seawater. Today, about 75% of large lakes (areas of over 500 km^2) are fresh; the rest are brackish to hypersaline. In view of this variety, it is not surprising that lakes share several sedimentary processes with the oceans but are also subject to other processes that, although not unique to lakes, are of much greater significance here than in marine environments. We can recognize a number of fundamental processes that directly or indirectly influence the types and distributions of deposits found in lakes.

- Physical processes are related to currents and waves generated by inflowing water or within the lake itself. Of particular importance are processes associated with rivers and with vertical and horizontal lake currents, and the physical structure of the water column. Physical processes, in particular the processes associated with slope failure and turbidity flows, are responsible for sediment mass movement and deposition.
- Biological processes may be directly responsible for the production of sediment or may play an important role in sediment accumulation and stabilization.
- Chemical processes: chemical sediments are common in some sorts of lakes, the most important being calcareous deposits and the evaporite accumulations characteristic of saline lakes.
- Hydrothermal processes: many lakes in tectonically active or volcanic regions are influenced by hydrothermal springs, which may affect the water chemistry or precipitate mineral deposits.
- Tectonic processes: most large long-lived lakes occur in basins of tectonic origin. In such basins tectonic effects may have a significant impact on patterns of sediment distribution and accumulation.

Although each process is listed separately, it should be borne in mind that some are closely interrelated. The production of biogenic sediments, for example, may depend on the presence of an essential nutrient, the supply of which is controlled by a lake's mixing regime. The inorganic precipitation of calcium carbonate often occurs during periods of high biological production and in some settings is probably controlled by it (see below).

Water-Column Structure, Currents, and Waves

Vertical Mixing

The degree to which surface and deeper waters mix is of fundamental importance to several aspects of lake sedimentation and biological production. Mixing is largely controlled by the turbulence associated with wind-driven surface waves and by the density contrasts between different parts of the water column. In shallow lakes wind stress may be sufficient to keep the water column more-or-less permanently well mixed, but in deeper lakes the degree of mixing between surface and bottom waters is largely determined by density differences. Although sediment and salt content influence the density of lake water, the primary control in freshwater lakes is temperature. Water becomes denser as it cools, achieving its maximum density at 4°C, so many temperate and high-latitude or high-altitude lakes go through an annual mixing cycle. In summer, when insolation is strongest, the surface water is warm and has a lower density than the deeper cooler waters. The water column is density stratified, with a warmer epilimnion and a cooler hypolimnion. These are separated by a relatively thin zone, the metalimnion, where there is typically a marked temperature gradient, the thermocline (Figure 1). During autumn and winter, surface waters cool until they become denser than the deep water, at which time the density stratification breaks down and vertical mixing occurs (Figure 1). In warm climates, particularly the subtropics and tropics, where temperatures are always high, surface waters may never cool sufficiently to become denser than the deep waters, so complete vertical mixing cannot occur. Such lakes are permanently stratified and are said to be meromictic. The role of climate in controlling mixing regime is clearly illustrated by the contrast between Lake Baikal and the deep East African rift lakes. The former is subject to a mid-latitude continental climate, with hot summers and frigid winters. Despite its great depth, Lake Baikal is well mixed. Lakes Tanganyika and Malawi, on the other hand, are meromictic tropical lakes, with an anoxic hypolimnion that constitutes a significant proportion of the total water mass.

Density stratification has several important sedimentological consequences. It determines the depth

at which inflowing water enters the lake (see below), determines the degree of oxygen renewal to the hypolimnion, and has a major influence on the nutrient supply to the surface waters, where most primary production takes place. A prolonged or permanent absence of complete vertical mixing may result in oxygen deficiency in the hypolimnion. This anoxia is inimical to higher forms of life (*see* **Sedimentary Environments:** Anoxic Environments). The hypolimnion can also become a sink for nutrients. Primary production is limited to the photic zone, which typically has a maximum depth of *ca*. 40 m and is often much less, depending on the clarity of the water. Dead organic matter produced in the epilimnion sinks into the hypolimnion where it is broken down by microbes; the nutrients released by this decomposition will be unavailable for new production unless they are returned to the surface waters. Recycling of nutrients in this way is of major significance to primary production and, ultimately, to economically important fisheries in many large lakes. Under anoxic conditions, some of the dead organic matter may survive decomposition to become incorporated into the bottom sediments. The deep-water deposits of stratified lakes are commonly notably rich in organic matter. The absence of higher organisms also means that the sediments can accumulate undisturbed by bioturbation. As well as having a high organic-matter content, sediments that accumulate in the hypolimnions of meromictic lakes are often characterized by a very regular fine lamination.

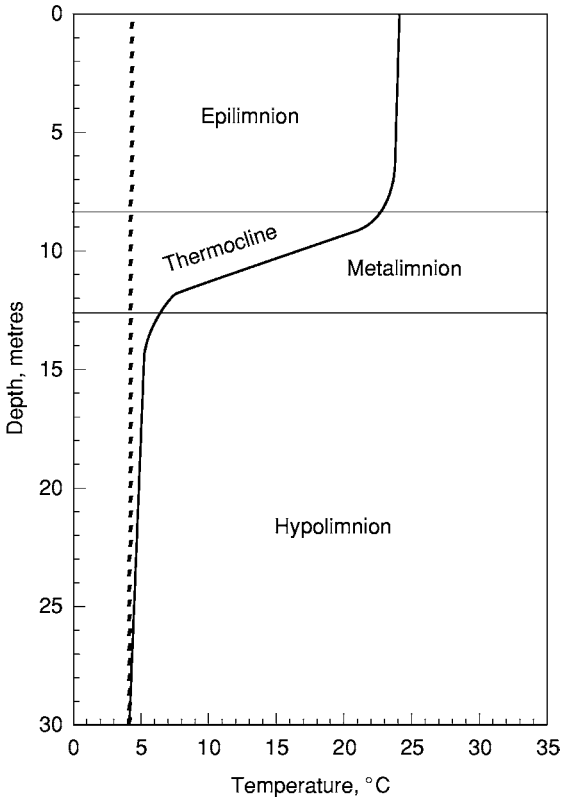


Figure 1 Temperature profiles through a typical temperate lake in summer (solid line) and winter (dashed line). Note how the thermocline separates the lake into two distinct water masses in summer. Isolation of the hypolimnion at this time can lead to severe oxygen deficiency or even anoxia. The winter profile shows the lake in an isothermal condition when it can mix to the bottom.

In addition to turbulent and density-related effects, other processes may enhance the mixing between deep and surface waters, even in lakes with permanent stratification. If a strong wind blows continuously from one direction, the surface waters can be pushed downwind, leading to the development of an epilimnetic wedge, which is separated from the hypolimnion by a tilted thermocline (Figure 2A). At the upwind end of this wedge, surface waters may eventually be completely displaced, to be replaced by upwelled water from the hypolimnion (Figure 2A). When the wind drops or changes direction, the thermocline tends to return to the horizontal; typically this does not occur in a single simple motion, but instead the thermocline oscillates in a seesaw fashion, setting up an internal wave or seiche (Figure 2B) along the length of the lake. The seiche will eventually decay, but, before doing so, may repeatedly bring cool nutrient-rich waters into the photic zone. Upwelling related to the development of an epilimnetic wedge and seiching is an important aspect of nutrient cycling in large tropical water bodies such as Lakes Malawi and

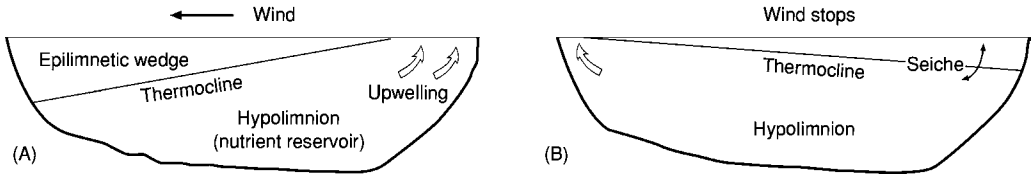


Figure 2 (A) Development of an epilimnetic wedge due to a strong unidirectional wind. Downwind displacement of the epilimnion causes the thermocline to tilt in an upwind direction and can eventually lead to the upwelling of cooler nutrient rich water from the hypolimnion. (B) When the wind stops, the thermocline rebounds, setting up an oscillatory motion or internal wave, known as a seiche. The seiche eventually decays because of frictional effects, but while it persists it may cause periodic upwelling at both ends of the lake.

Tanganyika. The upwelling zones are typically areas of very high primary production and may be characterized by the accumulation of sediments that are notably rich in biogenic material (see below).

River Inflow

Virtually all lakes, apart from those in areas of karst, are maintained by surface inflow from rivers and streams, which are also the principal suppliers of detrital sediment. How the inflowing water interacts with the lake water depends on the relative densities of the two water masses. It is not uncommon for river water to be significantly cooler than the water at the lake surface, such that the inflow tends to sink, forming either an interflow or an underflow (Figure 3). High sediment loads or salt contents can also produce dense river water. When such waters enter a lake they will plunge below the surface until they reach a depth where the lake water has a similar density. Interflows and underflows can be very efficient distributors of detrital sediment to offshore areas of lakes (Figure 3). One of the best-known examples occurs in Lake Mead (USA), which was formed by the damming of the Colorado River. Relatively salty sediment-laden river water enters the head of Lake Mead as an underflow current, which persists until it reaches the dam *ca.* 150 km downstream.

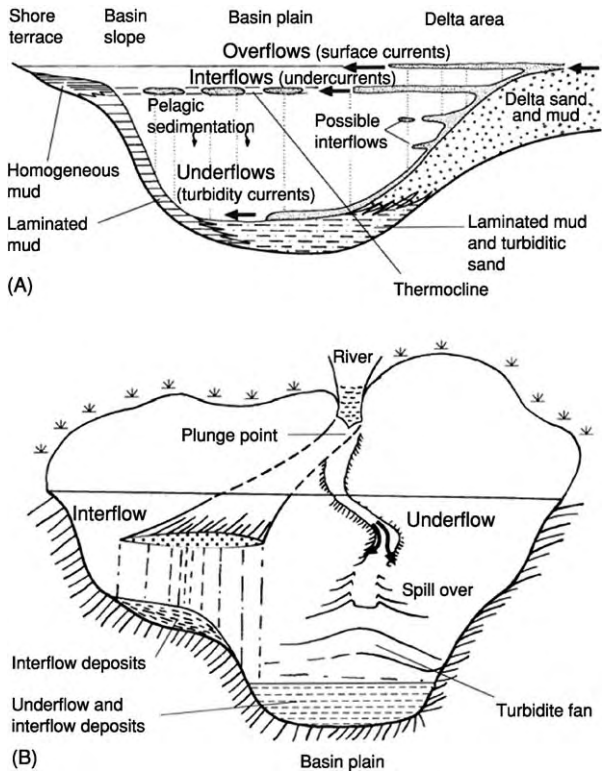


Figure 3 (A) Various types of density driven currents and their characteristic deposits as they relate to the flow of water into a lake. Overflows occur when the inflowing water has a lower density than the lake water; interflows have a density similar to that of a particular level in the water column, and underflows have a density that is at least as great as that of the deepest water in the lake. (B) Sediment distribution as it relates to interflows and underflows. The former can transport sediment to areas of the lake floor that would otherwise be starved of sediment. Persistent underflows can lead to the development of turbidite fans with well developed levées. Reproduced with permission from Talbot MR and Allen PA (1996) Lakes. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 83 124. Oxford: Blackwell.

The most characteristic deposits associated with river inflow are those that accumulate in deltas and fan deltas, which are a typical feature of many lake shorelines. Lacustrine deltas share many features with marine deltas, but the absence of tidal and, in many lakes, longshore currents means that sediment may not be as rapidly dispersed from the river mouth as it is in the ocean. These are some of the reasons why there is a tendency for lacustrine deltas to develop foresets with a large angle of repose, particularly where they build into relatively deep water (Figure 4). Such deltas are commonly referred to as Gilbert-type deltas after G K Gilbert, who first described these features in the abandoned shorelines of Lake Bonneville, a giant Pleistocene lake that formerly occupied the Great Salt Lake basin in Utah. The deposits of lacustrine deltas vary from coarse gravels to fine sands and silts, depending on flow conditions and the nature of the available sediment (Figure 5A). The relatively steep slopes of Gilbert-delta foresets can make them prone to collapse (see below), and, if underflows are frequent, the foresets can be reworked by them to produce ripples and imbricated clasts. Delta topsets merge with river-channel and floodplain deposits, and the bottomsets merge with the lake's deeper-water offshore deposits.

Waves

In addition to their role in the turbulent mixing of surface waters (see above), waves can be important sedimentological agents. In small or sheltered lakes where the wind's fetch is restricted, wave action may be limited. Indeed, shorelines that are little affected by waves may be marked by beds of reeds or other macrophytes, which act as very effective baffles, further reducing the potential impact of waves. On exposed shorelines, however, wave action can be significant, producing beaches and berms of well-sorted sand or gravel (Figures 5B and 6). In the absence of tides, lacustrine beaches may be relatively narrow in comparison with their marine counterparts and may show a rapid transition into offshore deposits. Nevertheless, seasonal or longer-term variations in lake level due to changes in water balance can cause large areas of the shoreline to be affected by wave action. In

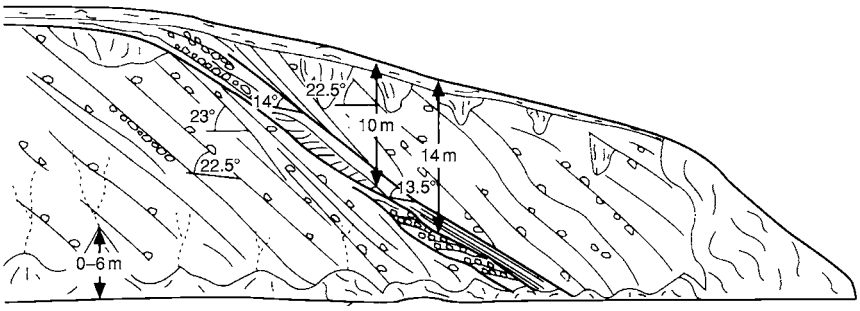


Figure 4 Cross section through a Gilbert delta from the type area along the eastern shore of the Great Salt Lake. The unit with low angle cross bedding in the middle of the section is due to wave reworking of the delta foresets, possibly related to a fall in lake level. The length of the section is *ca*. 150 m. See **Figure 5A** for a close up of the foresets showing the typical deposits. Reproduced with permission from Talbot MR and Allen PA (1996) Lakes. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 83 124. Oxford: Blackwell.

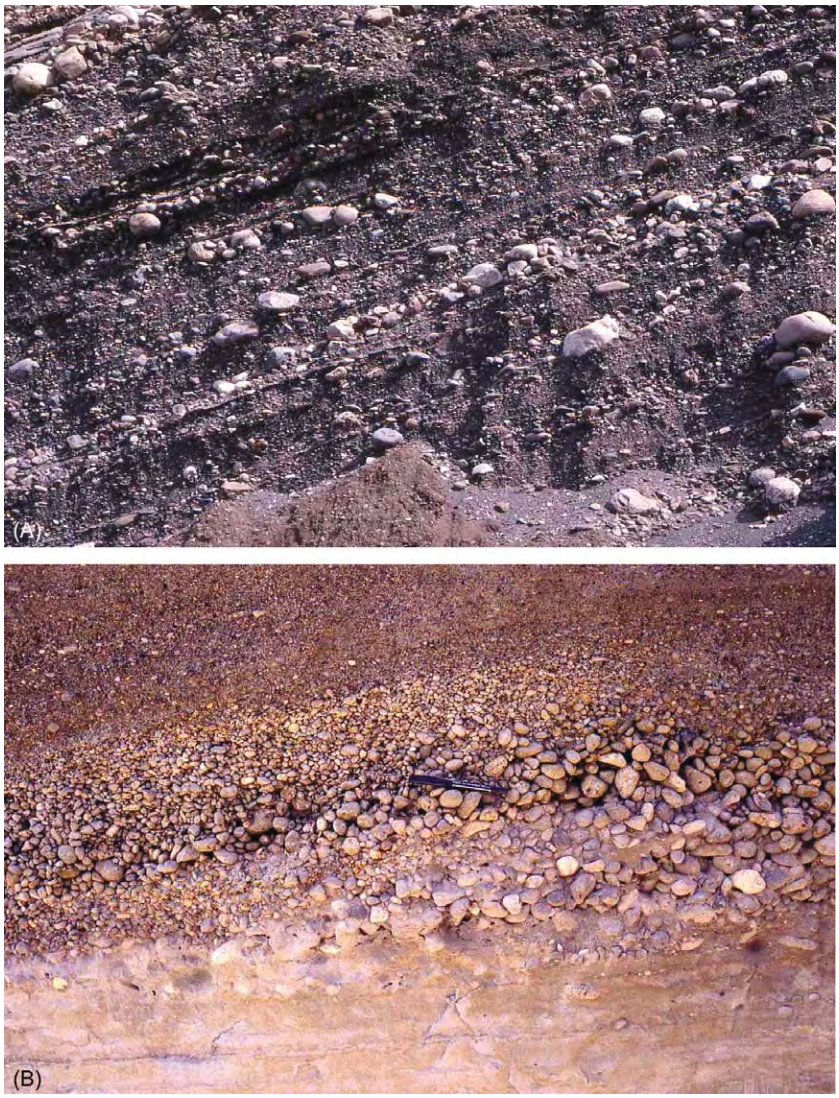


Figure 5 (A) Typical foreset deposits in a Gilbert delta. Note the variable grain size, due to changes in discharge, and the tendency to downslope imbrication in some of the layers of cobbles. The height of the section is *ca*. 2.5 m. Late Pleistocene, Willard, Utah. (B) Beach gravel consisting of well rounded lava and pumice clasts. The coarser zones, in particular, show the typical open framework texture (absence of matrix) characteristic of wave washed beaches. Holocene, Lake Langano, Ethiopia.

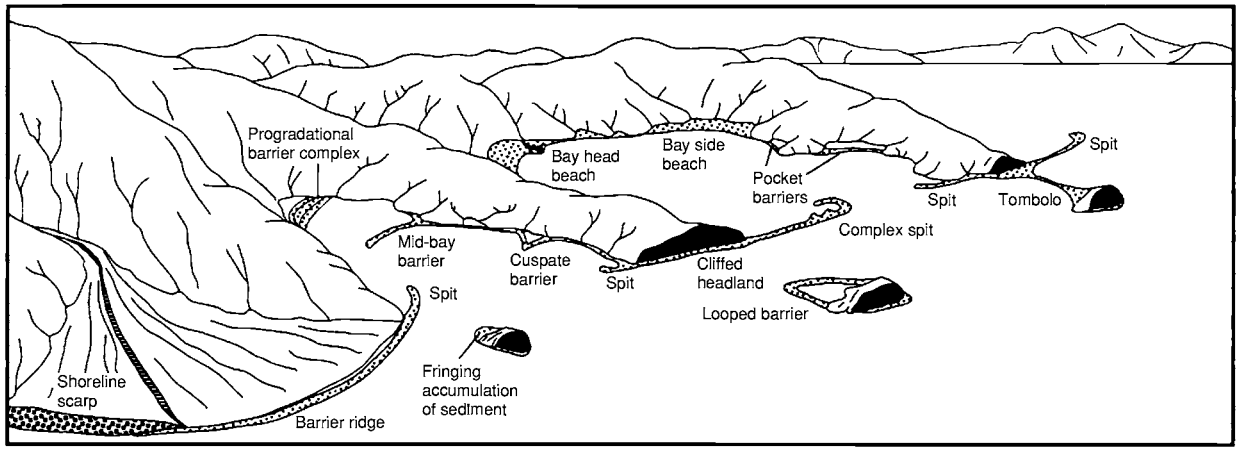


Figure 6 Depositional features formed along the shoreline of a wave prone lake (see **Figure 5B** for an example of a gravelly beach deposit). Reproduced with permission from Adams KD and Wesnousky SG (1998) Shoreline processes and the age of the Lake Lahontan highstand in the Jessup embayment, Nevada. *Geological Society of America Bulletin* 110: 1318 1332.

some lake basins, stranded beach ridges provide important evidence of major changes in lake level related to regional climate change.

Surface Currents

Horizontal currents occur in all lakes, and in large lakes they can be a major feature of the water's circulation. Wind stress is the commonest cause of surface currents, but they can also occur in association with river inflow. Strong persistent winds can set in motion longshore currents of sufficient strength to transport significant volumes of sediment, leading to the formation of spits (Figure 6). Where surface flow is restricted by topographical obstacles such as islands, the resulting intensification of the current can also affect deeper waters – sandy current-reworked deposits have been observed at depths in excess of 200 m in Lake Superior, for example.

Subsurface Currents

As indicated above, cooler or sediment-laden river water may plunge beneath the surface of a lake to become an interflow or underflow; in this way river-derived sediment can be spread over large areas of a lake. It is also not uncommon for significant quantities of suspended sediment to be temporarily trapped at the equilibrium depth of the interflow (which is often the thermocline), to be released gradually through simple settling or rapidly as a result of a breakdown in thermal stratification. In both cases, large areas of the lake floor can be blanketed with silt- and mud-rich sediment (Figure 3). Persistent sediment-laden underflows can lead to the development of sublacustrine canyons and deep-water fans with well-developed channels and levees (Figures 3 and 7) and can deposit beds of graded sand over large areas of the deep lake floor. The accumulation of subaqueous fans from underflows and turbidity currents is a characteristic feature of deep-water sedimentation in lakes at high and low latitudes.

Underflows can also play an important role in the formation of one of the most characteristic lacustrine deposits, varves (Figure 8). These are composed of sand rhythmically interbedded with clay or silt and, as originally defined, preserve one annual cycle of sediment accumulation on a lake floor. Each sand lamina represents summer deposition, when abundant sediment is transported into the lake, and the cold sediment-laden river water forms an underflow that transports the sand to the lake floor. The clay or silt lamina accumulates in winter, when rivers are frozen and the lake is covered by ice. However, detailed studies of rhythmically laminated sand and clay or silt deposits have shown that by no means all such couplets are the product of an annual cycle of deposition. Some may be graded beds produced by turbidity currents during minutes to hours of sedimentation; others are the product of an individual flood event (the sand) and its waning flow (the clay or silt) with a duration of hours to days. Given that varve counting is widely used for chronological purposes, it is clearly essential to exercise care in the interpretation of varve or varve-like deposits. Ideally, the inferred annual cyclicity should be confirmed by an independent dating method.

A variety of rhythmically interlaminated deposits reflecting annual climatic variations other than those related to a freeze–thaw cycle are known from lakes. Some of these are described below.

Mass Failure

As indicated above, the foreset slopes of lacustrine deltas may be prone to collapse. In addition, many lake

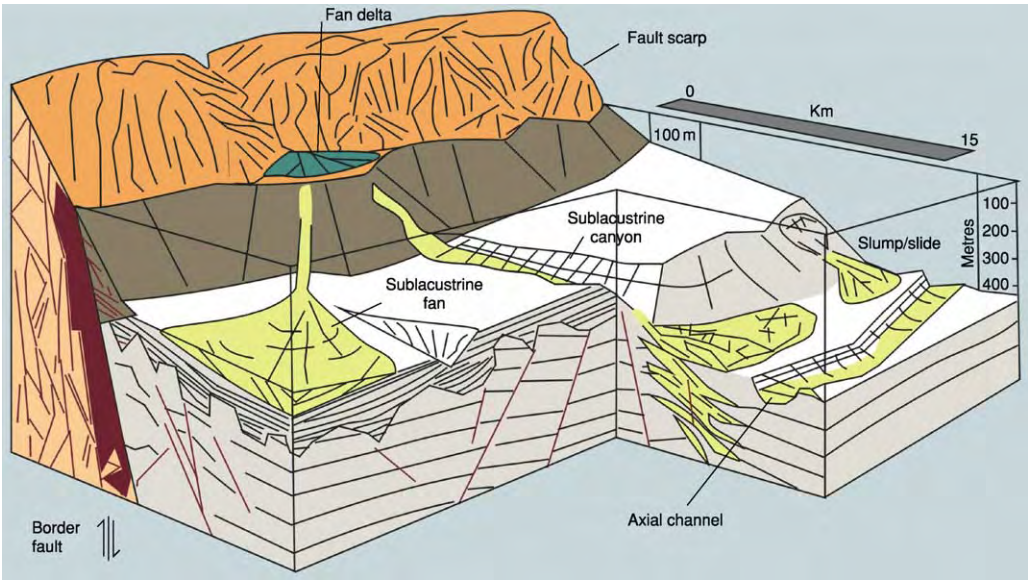


Figure 7 Sublacustrine depositional and erosional features along the margins of a deep lake subject to density flows and slumping. Based on the western margin of Lake Malawi, Africa. Adapted with permission from Wells JT, Scholz CA, and Soreghan MJ (1999) Processes of sedimentation on a lacustrine border fault margin: interpretation of cores from Lake Malawi, East Africa. *Journal of Sedimentary Research* 69: 816 831.

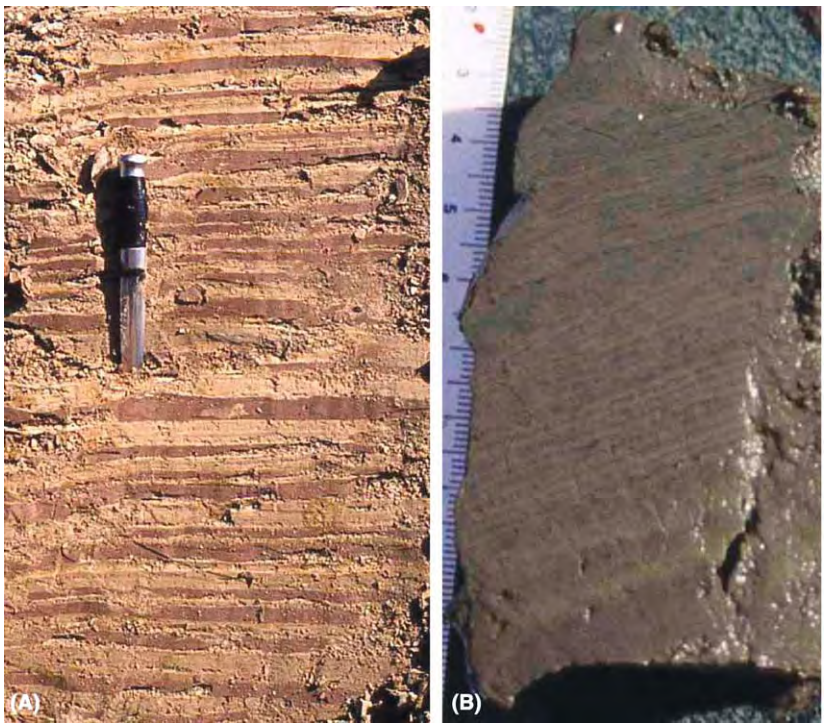


Figure 8 (A) Varves from a glacial lake. The darker brown layers are clay and silt (winter); the lighter layers are sand (spring summer). Note that, in detail, some of the sand layers are distinctly laminated, suggesting that they accumulated during a number of discrete depositional events. Late Pleistocene, Bredåkra, Sweden. Photograph by Jan Mangerud. (B) Recent varves from Lake Malawi. The lighter laminae are rich in diatoms and accumulated during the dry windy season. The darker laminae contain abundant terrigenous plant and mineral material and were deposited from the abundant runoff generated during the rainy season. The thicker more homogeneous layers are thin turbidites.

basins have steep margins adjacent to high mountains. Both settings favour mass failure and the generation of slumps, slides, and turbidity currents (Figure 7). Such events can result in the redistribution of large volumes of sediment and leave extensive slump scars on delta tops (*see* **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes). In addition to producing characteristic deposits, such slumps can have catastrophic consequences for lakeside communities. Sediment cores and historical records document the occurrence of damaging surges (tsunamis) associated with major slumps in some Swiss lakes. Slumps may be triggered by earthquakes or major floods. Typically, they produce turbidite beds that are exceptionally thick in relation to the normal deepwater sediments and may be characterized by a relatively homogeneous texture due to the mixing and rapid deposition of sediment from a number of sources. Seismic shock may also deform lacustrine deposits through liquefaction, without causing mass movement. Zones of seismically induced deformation in the cores of lake sediments have been used to date prehistoric earthquakes and determine earthquake recurrence intervals.

Mass-flow processes are also of major importance in glacial lakes, where glaciers deposit large volumes of unconsolidated till and other sediment at the ice margin. Slumping of this material into deep water can be widespread. In addition, large masses of sediment can be dumped in the lake by the calving and overturning of icebergs.

Biological Processes

Organic remains can make a significant contribution to lake sediments; the most important types are calcium carbonate in the form of shells, opaline silica predominantly produced by diatoms, and organic matter.

Calcium Carbonate

In lakes there are no planktonic organisms equivalent to the calcium carbonate-secreting foraminifera and green algae (coccoliths) responsible for producing the extensive chalk deposits of Mesozoic to modern oceans. However, there are many species of freshwater mollusc, and, under suitable conditions, these can be important sediment producers. In Lake Tanganyika, for example, some shallow marginal areas are characterized by very extensive accumulations of gastropod shells, which are concentrated by the winnowing action of waves in this large and windy lake. Ostracod carapaces can also contribute significant amounts of biogenic calcium carbonate to lake sediments, but it is rare for them to be the dominant component of a sediment.

Some macrophytes actively precipitate calcium carbonate. Of particular significance are the charophytes (stoneworts), which can form dense 'meadows' in the shallows of hard-water lakes. Charophytes produce calcified reproductive bodies, and the plants themselves are often encrusted with calcite. Both forms of calcium carbonate can make important contributions to the sediments of shallow lakes.

Stromatolites are another type of calcium-carbonate accumulation that is widespread in modern and ancient lacustrine deposits. Stromatolites are formed through the development of mats and films of algae, bacteria (particularly cyanobacteria), and other microbes. In some situations the role of these micro-organisms has been primarily to trap and bind sediment particles, but many lacustrine stromatolites are calcified, giving the structures considerable rigidity and resistance to erosion. These calcified stromatolites are often referred to as lake tufa. Internally, the stromatolites typically show a fine wavy lamination, and in many the laminae are convex-up or hemispherical, reflecting the generally upward and outward growth of the structure. Some stromatolites reach spectacular proportions, and examples several metres high and tens of metres in diameter are known from the fossil record; in some cases they form reef-like accumulations around the margins of former lakes (*see* **Biosediments and Biofilms**).

Silica

During the Tertiary, diatoms emerged as important nonmarine algae, and from the Miocene onwards they have been major producers of lacustrine sediment. Diatoms secrete a porous skeleton (frustule) of amorphous opaline silica, the majority of which are between 5 μm and 200 μm in diameter. Where supplies of dissolved silica and other nutrients are abundant, diatoms often play a major role in primary production, even in offshore pelagic zones, thereby supplying abundant opal to the sediment. During the spring warming of temperate lakes, and at times of upwelling in stratified lakes, diatoms can occur in huge numbers, producing blooms that turn the surface waters cloudy. Deposits rich in diatoms are called diatomites and are characterized by a chalky appearance and very low density (some float when dry). Because of their low bulk density, it may seem surprising that so many diatoms sink to the bottom. It is likely that another biological process – grazing by zooplankton and the subsequent excretion of diatom-rich faecal pellets, which sink relatively quickly – plays an important part in the sedimentation of diatom remains. Variation in the rate of accumulation of biogenic silica has proved to be a useful proxy indicator of palaeoproductivity and thus palaeoclimate.

Where diatom blooms are a seasonal event, the resulting diatom-rich sediment layer may contribute to the development of varves, as is the case in some areas of Lake Malawi. Diatom production here is highest in the dry windy season, when turbulent mixing and upwelling recycle nutrients from the anoxic hypolimnion. At this time of the year runoff from the land is at a minimum, so a relatively pure diatomaceous lamina accumulates. In the rainy season, on the other hand, winds are slacker and productivity declines, but runoff is at a maximum, bringing abundant mineral and organic debris into the lake. Thus, the two seasons are represented by a light diatom-rich lamina and a dark detrital-rich layer (Figure 8B).

Organic Matter

Many lake sediments contain notable quantities of organic matter. This material has two principal sources.

1. It may consist of the remains of higher plants that populate the lake shallows, shore, and surrounding landscape. Accumulation of large quantities of such plant remains leads to the formation of freshwater peats and coals.
2. It may consist of the remains of planktonic organisms. In addition to diatoms, a number of planktonic algae and cyanobacteria can be major primary producers in lakes. Under some conditions these can make a major contribution to the sediment. In oxygenated environments, dead algal and bacterial-cell material quickly breaks down, but its chances of preservation are much greater if it accumulates in an anoxic environment, such as the hypolimnion of a stratified lake. This environment, allied with high primary production in the epilimnion, can result in the deposition of offshore muds with organic-matter contents in excess of 10%. Ancient organic-rich lacustrine shales are of major economic importance as hydrocarbon source rocks; in countries such as China, Brazil, and Angola a large proportion of the oil has been generated from such sources.

Chemical Processes

A number of important lacustrine sediments are of chemical origin, formed by direct precipitation from the lake water. The most important in terms of sediment production are calcium carbonate and evaporite minerals.

Calcium Carbonate

Lakes fed by rivers draining areas of limestone or other rocks rich in calcium, such as some basic lavas, can have high concentrations of dissolved calcium carbonate, which may precipitate as calcite, or, less commonly, aragonite. In temperate lakes, precipitation occurs during the warm season, but it may be a year-round phenomenon at lower latitudes. Although the precipitates typically comprise simple crystals, 10–100 μm long, of clearly inorganic origin, it is now recognized that their formation is often closely tied to organic processes. At times of high primary production, the photosynthetic consumption of dissolved carbon dioxide may greatly exceed the rate at which it can be replenished. Under these conditions, surface waters can rapidly become supersaturated with calcium carbonate, leading to widespread calcite or aragonite precipitation. Recently, it has also become clear that micron-sized plankton (picoplankton) may play a crucial role in the initial nucleation of the calcite crystals. During phytoplankton blooms precipitation can occur over large areas, resulting in events known as whitings (Figure 9). In Europe, deposits rich in calcium carbonate of this sort are sometimes referred to as lake marl. Seasonal precipitation of significant amounts of calcium carbonate can result in the formation of varved sediments, the light carbonate-rich lamina representing spring and summer seasons, while a darker detrital-rich layer is deposited in autumn and winter. A number of Swiss lakes have accumulated such deposits through much of the Holocene.

Another form of inorganic calcium carbonate can occur along the shores and in the shallows of wave-agitated carbonate-precipitating lakes. Here characteristic concentrically laminated sand-sized grains known as ooids may develop and, under favourable conditions, produce extensive deposits. They are, for example, widespread in the modern Great Salt Lake and on some of the beaches of Lake Tanganyika. Calcite cement may also precipitate in the pores of these deposits to produce beach-rock.

Evaporites

Evaporite minerals are deposited from saline lakes, where they form as a result of the evaporative concentration of lake water (*see* **Sedimentary Rocks:** Evaporites). Definitions of what constitutes a saline lake vary, depending on the criteria used. The most practical definition is a biological one, based on the observation that typical freshwater organisms are generally absent from waters whose salinity exceeds 5‰. A bewildering number of different salts have been recorded from saline lake basins, including some, such as borates, nitrates, and sodium carbonate minerals, that are of major economic importance. The commonest lacustrine evaporites are gypsum ($CaSO_4 \cdot 2H_2O$), halite (NaCl), and a variety of sulphates, carbonates, and bicarbonates of sodium.

Figure 9 Satellite image of Lake Michigan taken on 20 August 1999. The surface waters over much of the southern half of the lake have a milky appearance ('whiting') due to the extensive development of phytoplankton blooms and the accompanying precipitation of calcite (Part of NASA Visible Earth image #ev4338_S1999254181550.png).

Which minerals are precipitated is determined by the chemistry of the inflowing water. This is in marked contrast to the formation of marine evaporites, where the uniform composition of seawater ensures that only a limited suite of minerals can occur.

Because the precipitation of salts requires the evaporation of large volumes of lake water, their formation and accumulation tend to occur in lakes that are hydrologically closed (i.e. they have no outflow) and in regions with a dry or arid climate. In these settings, water loss due to evaporation can exceed inflow. Such lakes tend to alternate between brief periods of flooding, following rainfall, and longer periods of saline or even completely desiccated conditions. Because of their large variations in volume, many saline lakes are surrounded by extensive salt flats, where salt precipitation may continue owing to groundwater seepage and loss through capillary effects (Figure 10).

Hydrothermal Processes

Lake Baikal and many lakes in the East African rift system, the western USA, and the Andes receive a portion of their inflow from the subsurface. These springs may enter the lake from above or below the lake surface, and their temperature may be anywhere between ambient and boiling. The water may be rich in dissolved ions, which can have a measurable influence on the chemistry of the lake water and, in saline lakes, the nature of any evaporite minerals that may precipitate. In addition, effects induced by changes in physicochemical conditions at the point where a spring vents into a lake may trigger mineral precipitation and the accumulation of hydrothermal deposits. Silica and carbonates are the commonest precipitates; the latter may form spectacular deposits (Figure 11). Sulphide mineralization has also been recorded in some of the East African rift lakes.

The spring deposits typically take the form of irregular masses of travertine and sinter, which in many cases are rather porous and show complex growth patterns. One reason for these features is that many springs, even those that emerge at close to boiling point, have dense growths of filamentous and encrusting microbial communities associated with them. These colonies clearly influence the manner in which the minerals accrete. The larger voids represent conduits through which fluids passed; some of the sublacustrine springs in Lake Tanganyika are marked by chimney-like structures of aragonite, many of them carrying fronds of milky white filamentous bacteria, swaying in the stream of hot water exiting the conduit.

Tectonic Processes

Although lake basins have a variety of origins, outside areas of glaciation most large long-lived lakes occupy basins formed by tectonic processes. The morphology of such a basin is to a large degree determined by its tectonic setting. The Lake Baikal and East African rifts, and many of the basins in the western USA, for example, are the result of crustal extension, which typically produces half-graben structures characterized by marked topographical asymmetry (Figure 12) (*see* **Tectonics:** Rift Valleys). The form of the basin determines the position of the lake, which will occur at the topographically lowest point of the basin; typically, this means that the deepest water is close to a border fault. Sediment accumulation is also greatest here, where maximum subsidence rates produce the space needed to accommodate new deposits. Continuous Miocene–Holocene subsidence has resulted in the accumulation of several kilometres of lacustrine sediment beneath Lake Baikal and in parts of the East

Figure 10 (A) Surface of desiccated salt lake, Death Valley, California. (B) Close up view showing upturned edges of desiccation cracks with salt efflorescence precipitated from seeping groundwater. The salt is mainly halite (NaCl).

Figure 11 Mounds of calcium carbonate ('The Pinnacles') formed by the discharge of carbonate rich water through the floor of a Late Pleistocene lake. The mounds are aligned along a fault zone. Searles Lake, California.

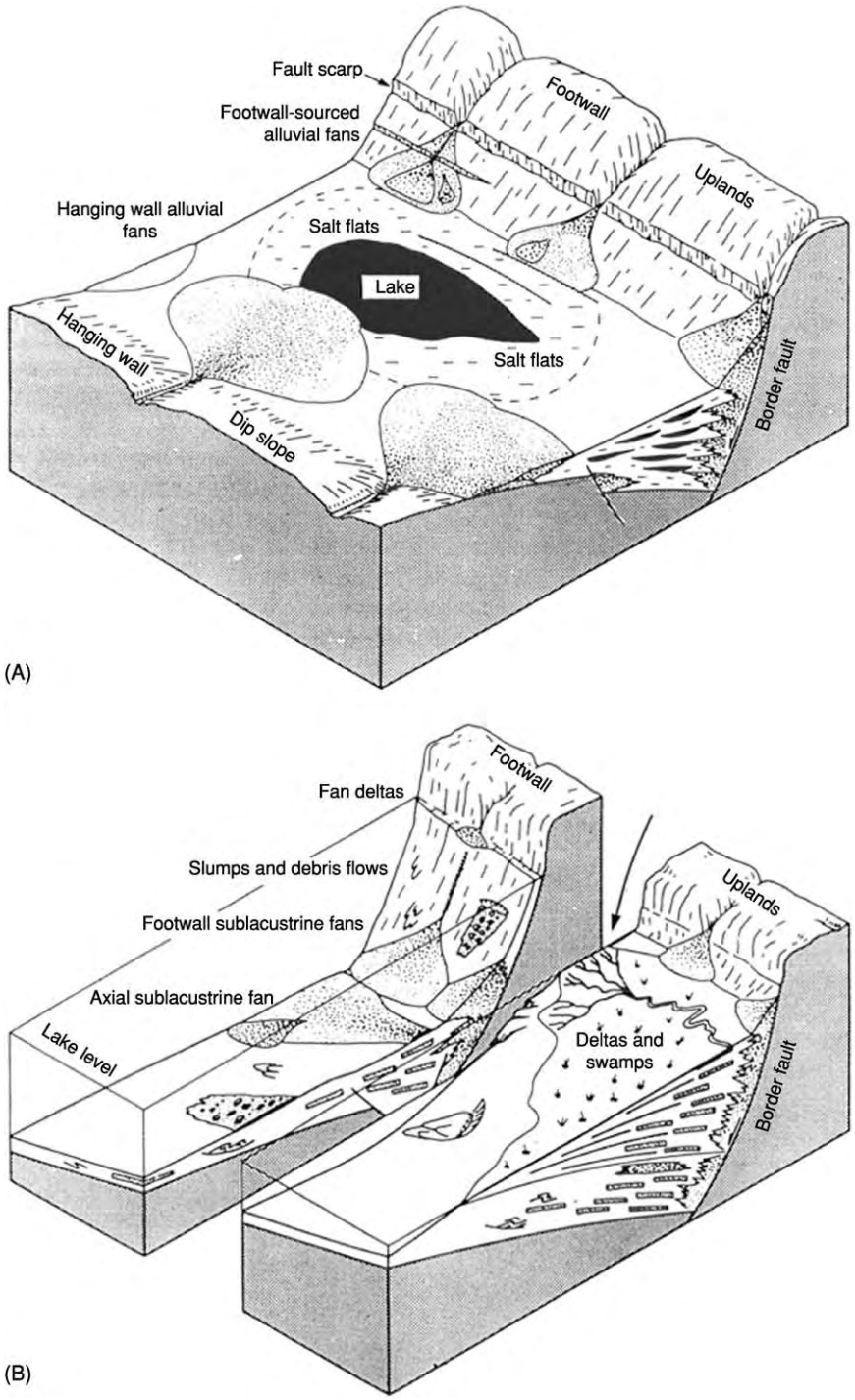


Figure 12 Schematic distributions of deposits in and around lakes developed in a half graben (rift) basin under (A) dry to arid and (B) humid climatic conditions. (A) The basin is hydrologically closed and the lake is probably saline; extensive salt flats surround the lake and are flooded only following rare rainstorms. (B) A deep rift lake. In both cases the asymmetry of the basin has a strong influence on sediment distribution. This is particularly apparent in (B), where the deepest water is close to the border fault and the axial drainage system also tends to be located here owing to maximum subsidence in this part of the basin. Adapted with permission from Leeder MR and Gawthorpe RL (1987) Sedimentary models for extensional tilt block/half garben basins. In: Coward MP, Dewey JF and Hancock PL (eds.) *Continental Extensional Tectonics*, pp. 139 152. London, Geological Society Special Publication No. 28.

African rift. The distribution of different sediment types closely reflects the asymmetry and varied bathymetry of the basin (Figure 12), while structurally determined sublacustrine topography exerts a major influence on deep-water sedimentation. Fault scarps form steep slopes, which tend to be unstable, generating debris and turbidite flows and leading to widespread sediment redeposition (Figures 7 and 12). Furthermore, studies of Lake Baikal and the East African rift lakes clearly demonstrate that many of the principal sublacustrine canyons along which sediment is channelled from shallow water to deep water are fault controlled. In addition to these structural effects, sedimentation in Lake Baikal has been strongly influenced by variations in the types and amounts of sediment delivered to the basin resulting from changes between glacial and non-glacial conditions, while the East African lakes have been subject to major climate-driven changes in sediment supply and lake level (on the scale of hundreds of metres). The impact of climate change and the active tectonism have resulted in complex depositional histories in all these basins.

See Also

Biosediments and Biofilms. **Sedimentary Environments:** Anoxic Environments; Deltas; Shoreline and Shoreface Deposits; Storms and Storm Deposits. **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes. **Sedimentary Rocks:** Evaporites. **Tectonics:** Rift Valleys.

Further Reading

Beadle LC (1981) *The Inland Waters of Tropical Africa*. London: Longman.

Cohen AS (2003) *Paleolimnology: The History and Evolution of Lake Systems*. Oxford: Oxford University Press.

Colman SM, Karabanov EB, and Nelson CH (2003) Quaternary sedimentation and subsidence history of Lake Baikal, Siberia, based on seismic stratigraphy and coring. *Journal of Sedimentary Research* 73: 941 956.

Eugster HP and Hardie LA (1978) Saline lakes. In: Lerman A (ed.) *Lakes: Chemistry, Geology, Physics*, pp. 237 293. Berlin: Springer.

Gawthorpe RL and Leeder MR (2000) Tectono sedimentary evolution of active extensional basins. *Basin Research* 12: 195 218.

Gilbert GK (1885) *The Topographic Features of Lake Shores*. 5th Annual Report. Washington: United States Geological Survey, USA.

Johnson TC, Brown ET, McManus J, *et al.* (2002) A high resolution paleoclimate record spanning the past 25 000 years in southern East Africa. *Science* 296: 113 132.

Katz BJ (2001) Lacustrine basin hydrocarbon exploration current thoughts. *Journal of Paleolimnology* 26: 161 179.

Lezzar KE, Tiercelin J J, De Batist M, *et al.* (1996) New seismic stratigraphy and Late Tertiary history of the North Tanganyika basin, East African rift system, de duced from multichannel and high resolution reflection seismic data and piston core evidence. *Basin Research* 8: 1 28.

Lowenstein TK and Hardie LA (1985) Criteria for the recognition of salt pan evaporites. *Sedimentology* 32: 627 644.

Owen RB (2002) Sedimentological characteristics and origins of diatomaceous deposits in the East African rift system. In: Renaut RW and Ashley GM (eds.) *Sedimentation in Continental Rifts*, pp. 233 246. Special Publication 73. Tulsa: Society for Sedimentary Geology.

Rosendahl BR (1987) Architecture of continental rifts with special reference to East Africa. *Annual Review of Earth and Planetary Sciences* 15: 445 503.

Siegenthaler C, Finger W, Kelts K, and Sumin W (1987) Earthquake and seiche deposits in Lake Lucerne, Switzerland. *Eclogae Geologicae Helvetica* 80: 241 260.

Sturm M and Matter A (1978) Turbidites and varves in Lake Brienz (Switzerland): deposition of clastic detritus by density currents. In: Matter A and Tucker ME (eds.) *Modern and Ancient Lake Sediments*, pp. 147 168. Special Publication 2. London: International Association of Sedimentologists.

Talbot MR and Allen PA (1996) Lakes. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 83 124. Oxford: Blackwell.

TANGANYDRO (1992) Sublacustrine hydrothermal seeps in northern Lake Tanganyika, East African rift: 1991 TANGANYDRO Expedition. *Bulletin des Centres de Recherches Exploration Production Elf Aquitaine* 16: 55 81.

Thompson JB, Schultze Lam S, Beveridge TJ, and Des Marais DJ (1997) Whiting events: biogenic origin due to the photosynthetic activity of cyanobacterial picoplankton. *Limnology and Oceanography* 42: 133 141.

Wells JT, Scholz CA, and Soreghan MJ (1999) Processes of sedimentation on a lacustrine border fault margin: interpretation of cores from Lake Malawi, East Africa. *Journal of Sedimentary Research* 69: 816 831.

Wetzel RG (2001) *Limnology: Lake and River Systems*. San Diego: Academic Press.

Reefs ('Build-Ups')

B W Sellwood, University of Reading, Reading, UK

Introduction

Reefs are biologically generated accumulations of carbonate (*see* **Sedimentary Rocks:** Limestones) that grow upwards from the adjacent seafloor and are entities of their own making. They are generally considered to be robust structures (i.e., rigid or wave-resistant), although biogenic mud mounds are treated as reefs by many authors. There have been periods in the geological past when the main contributors to biogenic mounds appear to have been organisms lacking a rigid frame, and their constructions have had lime mud (matrix-rich) compositions. Thus a wide spectrum of structures have been loosely classified as reefs. It has therefore generally been considered useful to draw a distinction between reefs with a rigid calcareous frame and those without (i.e., mud mounds). As communities growing on the accumulated debris of their own forebears, reefs are essentially unbedded constructions. The debris derived from the reef-dwellers fills the interstices between *in situ* constructors and also forms aprons of bedded resedimented material on the flanks of the edifice, and in adjacent hollows. Such aprons provide critical evidence that biogenic activity produced an elevated mound in the first place, and may indicate that upward growth of the mound had taken place into shallow waters, where strong currents and waves were active.

Growing reefs are often strongly zoned ecologically and usually have a core, the main carbonate factory within the system, showing a distinct vertical succession from initial colonization through to culmination. The vertical zonation can frequently be related to upward growth in response to the filling of accommodation space (i.e., formed in response to 'transgression' with the bulk of the reef pile accumulating during a 'highstand'). Reef successions are often stacked one on top of the other, separated by lowstand emersion surfaces (often 'karstic' in character).

Modern Reefs

Modern reefs (Figure 1), dominated by scleractinian corals (*see* **Fossil Invertebrates:** Corals and Other Cnidaria), reflect a balance between growth and destruction. Coral growth produces a rigid framework ('framestone'(Figure 2)), sheeted encrustations generate 'bindstone'(Figures 3 and 4), whereas meadows of upstanding but flimsy (generally noncoral) organisms may passively trap fines to form mud mounds ('bafflestone' after lithification (Figure 5)). Attached

Figure 1 Modern patch reefs with intervening grass anchored carbonate sands as seen from the air. Great Bahamas Bank, field of view approximately 1 km across.

Figure 2 Framestone with *in situ* coral heads and unbedded character. Key Largo Limestone, Quaternary, Florida, USA. Height of outcrop about 2 m.

Figure 3 Bindstone predominantly composed of sheeted stromatoporoids, Devonian, Belgium. Height of outcrop approximately 40 cm.

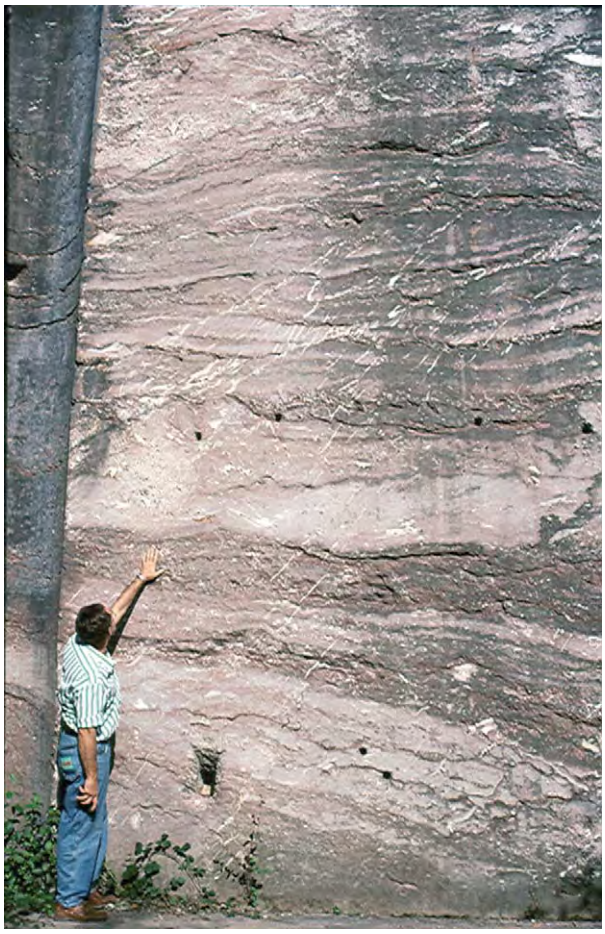


Figure 4 Bindstone sheets of stromatoporoids, Devonian, Belgium.

Figure 5 Bafflestone composed predominantly of lime mud probably trapped by mangroves. Messinian Miocene, Mallorca, Spain.

organisms and those living in any cavities and crevices contribute to the overall community, and some will have fossilization potential, whereas many will not. Destructive processes may be both physical (waves and currents) and biological (raspers, borers, and grazers). *In situ* collapse (following bioinfestation) and downslope relocation, as gravel-grade rubble, produces 'rudstones'(Figures 6, 7 and 8), whereas mud slides and debris flows, with blocks entrained or embedded within slurries of carbonate fines, produce 'floatstones' (Figure 9). Sand-grade material generated by grazers may itself be rendered still finer by biological degradation.

Corals generally remain in place after death. Cores through coral heads reveal regular annual growth layers, which provide a precise year-by-year record ('sclerochronology') of climate-related, anthropogenic, volcanogenic, and other events that leave a geochemical signature within the coral frame. Bioinfestation by macro- and microborers degrades skeletal structures, leading to the obliteration of original, and often characteristic, textures in grains and generates micritized 'peloids'. These processes also contribute to reef sediment, particularly the fines, as

Figure 6 Rudstone of coral debris. Quaternary, Barbados. Scale at centre 10 cm.

Figure 7 Rudstone of stromatoporoids debris, Devonian, Devon, England. Field width approximately 50 cm.

Figure 8 Rudstone of stromatoporoids, Devonian, Belgium.

do faecal and pseudofaecal material generated by the reef dwellers. However, many reef-dwelling organisms are microcarnivores and suspension-feeders highly sensitive to the amount of turbidity in the water, so reefs grow best on windward platform margins where wave surge cleanses the reef of the inhibiting fines. In such sites daily and seasonal temperature fluctuations are small, as are salinity fluctuations, the reef community generally thriving best in well-oxygenated waters within tight limits of normal marine salinities (~35‰ (range 27–40‰)), warm temperatures (18–36°C), and low nutrient content (phosphates, etc.). Investigations, using dynamite, have shown that the robustness of many reefs results from early marine cementation of the reef structure. Wave pumping forces vast quantities of seawater saturated with calcium carbonate through the interstices of the reef, and 'isopachous cements' of 'Mg-calcite' (and 'aragonite', which is often 'botryoidal' in form) grow actively and rapidly.

Figure 9 Floatstone. A reef derived boulder set in deep water mudstones and packstones, Permian Capitan Reef Complex, McKittrick Canyon, New Mexico, USA.

Reefs formed immediately adjacent to shorelines (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves) are termed 'fringing reefs'. Offshore reefs form a barrier, behind which is a 'lagoon'. Barrier reefs, generally circular in plan and with a central lagoon, are termed 'atolls', which, in the open waters of the Pacific and Indian Oceans, were initiated around volcanic islands and continued to grow upwards following the submergence of the original edifice. This mechanism was originally suggested by Darwin (1842) and was confirmed by the Funafuti boring, funded by the Royal Society of London. Lagoons behind barrier reefs may have patch reefs, which can form miniatolls, such as those behind the Great Barrier Reef off eastern Australia. Patch reefs (e.g., Figure 1) rise from depths of a few metres, and between them are usually extensive spreads of muddy lime sand anchored by sea grasses (e.g., *Thalassia*).

Platform margin reefs drop off, often precipitously, into deep water where the 'fore reef' comprises a zone predominantly receiving reef-derived debris, carbonate sand, and tumbled blocks (talus). The 'reef front' rises from this deep water. One hundred metres, in the clearest waters, is about the maximum depth in which coral colonization occurs but depths are usually less than this. The plate-like growth forms of the pioneers 'stabilizing' reef foundations in these gloomy waters are an adaptation for capturing the maximum amount of light. Coral polyps live symbiotically with zooxanthellae algae within their tissues, so enough light for their photosynthetic activities is an essential component in reef colonization and growth. Into shallower depths platy forms give way to 'colonization' by branching forms as architectures favouring food collecting take over in habitats where light is no longer a limiting factor. Massive, and finally encrusting, forms 'diversify' and proliferate in wave-activated depths approaching the reef crest. The reef crest is the highest energy zone impacted by breaking waves. Encrusting forms often 'dominate' there, but if wave energy is of moderate intensity they may be accompanied by forms with abbreviated branches and robust hemispherical shapes. The 'reef flat', often exposed at low tide and as such inhospitable to corals, is crossed by channels through which the surf surges. These are usually floored by sand that is carried lagoonwards. James (1983) suggests that reefs of many ages show a similar succession in growth forms (stabilization, colonization, diversification, domination) even though the organisms generating the reefs are often from phyla different from those constructing modern reefs. In contrast to the seaward zone the 'back reef' is a tranquil place, receiving much of the mud-grade carbonate generated by the reef. Abundant sand- and mud-producing organisms live there, particularly calcareous green algae (e.g., codiacians such as *Halimeda*). Stubby and dendroid growth forms dominate this habitat with species adapted to rapid upward growth in muddy environments.

Reefs through Time

This topic has been extensively reviewed. Prior to the advent of grazing metazoans in the Early Cambrian stromatolites were the prime reef builders. Stromatolitic reefs in Western Australia, formed between 3.0 and 3.5 Ga, and through the Archaean and Proterozoic large and increasingly complex structures and ecosystems, were formed by *Cyanobacteria* and other microorganisms ('microbes' such as noncalcified and calcified *Cyanobacteria*, Eubacteria, fungi, and other protozoans). These same communities are considered by many to be responsible, through their photosynthesizing activities, for changing the character of the atmosphere from its initial reducing (CO_2-rich state) to one with free oxygen (around 2.32 Ga).

The first metazoan reef mounds a few metres across were constructed in the Early Cambrian by 'Archaeocyathans' (sponge-like organisms) encrusted by calcitized algae and surrounded by carbonate sands rich in pelmatozoan debris. They have an apparently wide palaeolatitudinal spread (60° S to 30° N) and are particularly well described from Newfoundland, Labrador, and the Siberian Platform. Archaeocyathans became extinct at the end of the Early Cambrian and reef structures disappeared until the mid-Ordovician. At that time colonial corals and sponge-like 'stromatoporoids' made their first appearance and dominated palaeotropical reef associations, accompanied by an increasingly diverse array of sponges (*see* **Fossil Invertebrates:** Porifera), bryozoans, and calcareous algae, until the Late Devonian. Brachiopods (*see* **Fossil Invertebrates:** Brachiopods) and pelmatozoans contributed in a major way to these build-ups. Diversification continued through the Silurian and Devonian, a time of exceptionally high global sea-levels, which generated vast tropical epicontinental seaways within which grew abundant reefs dominated by sponges (stromatoporoids) and corals; reefs extended from 40° to 60° poleward of the equator. By Late Devonian times (**Figures 7** and **8**) reef systems probably rivalled modern reefs in terms of both their scale and complexity (some belts exceeding 2000 km in length). Many of the organisms contributing to these structures produced calcitic rather than aragonitic skeletons, and internal cavities within the reef systems were elongate lensoids with smooth flattish floors and complex, often crenulate ceilings. Termed 'stromatactis', and filled with radiaxial calcite spar, these cavities have long been a topic of controversy, as to whether they represent a long extinct organism group, dilation cracks caused by down-slope movement of partially cemented sediment, methane seepage, or strange burrow structures. That they were open cavities within the mounds, and preserved by early cementation processes, is confirmed by their internal sediment fillings and the lining cement phases.

A major extinction (*see* **Palaeozoic:** Devonian) event took place within the Late Devonian, considered by many to rival that at the End Permian in terms of its severity (*see* **Palaeozoic:** End Permian Extinctions). This caused the end of the coral-stromatoporoid reef systems, with mud mound systems apparently lacking rigid frame builders and dominated

by sponge-algal-microbial-foraminiferal communities continuing. The Early Carboniferous was a time of mud-mound communities (e.g., Waulsortian bank facies), the absence of major reef systems at this time being attributed to an overall decrease in global temperatures. The pattern of a relative absence of large frame builders continued through the Late Carboniferous and Early Permian when mound communities were dominated by calcareous algae, sponges, and bryozoans. During the Permian tropical and subtropical carbonate systems became dominated by aragonite, rather than calcite, a situation mineralogically similar to the Cenozoic and Recent.

Many mid- and Late Permian reefs, like the classic Capitan complex on the margins of the Delaware basin (Figures 9 and 10), most probably accumulated as a series of marginal mud mounds anchored by very early marine cements. Sponges, phylloid algae, and microbial precipitates compose the major components and debate has continued over many years about the architecture, palaeobathymetry over the reef, and other aspects of palaeoecology and geometry. It is clear that the evolution of the system was influenced significantly by sea-level changes, and considerable palaeobathymetry between reef and basin can be demonstrated by the presence of rudstones and floatstones that can be traced up to the reef from the basin floor. Another set of famous, but still controversial reefs are those of the Late Permian Zechstein basin of northern Europe. Communities within these reefs are highly zoned, containing bryozoans and stromatolites, with reef base, reef core, reef flat, reef crest, and patch reef communities all being recognized, the latter growing upon a distally steepened ramp. During the Late Permian reefs with corals (rugose corals that may have lived symbiotically with zooxanthellae and tabulates) are confined to the Australasian rim of the Pacific and the Middle East. Other biota within these reefs were calcareous algae, bryozoans, foraminiferans, and brachiopods. With the exception of coral communities in East Timor and the Salt Range of Pakistan, most of these build-ups occur within the palaeotropics.

There was a sudden and complete disappearance of high diversity reef biota at the end of the Permian (*see* **Palaeozoic:** End Permian Extinctions), and a global hiatus of about 10 Ma occurs before the renewal of reef development in the early Triassic. Triassic reefs were generated by a range of organisms including microbes; scleractinian corals (direct ancestors of modern corals); sponges; calcareous algae (dasyclads); bivalves; and serpulids. Microbial reefs dominate the earlier Triassic, sponge reefs the mid-Triassic, and coral reefs the Late Triassic. Throughout the Triassic great changes occur in the latitudinal distribution of reefs, expanding from an equatorial and tropical distribution in the mid-Triassic to a maximum of around 35° N and S in the Late Triassic. A mass extinction event at around the Triassic/Jurassic boundary caused a near extinction of reef biota and earliest Jurassic (Hettangian) reefs are extremely rare. Reef complexes, regarded now as classic locations, grew in the Tethys. Those in the Dolomites of northern Italy (Figure 11) were strongly controlled by sea-level changes, volcanic eruptions, and palaeobathymetry, and exhibit a wide range of facies and at a seismostratigraphic scale. Allochthonous (reef-derived) material is abundant; back-reef lagoonal and intertidal facies are also well developed. Some reefs have atoll-like geometries. The principle constructors in some earlier cases were encrusting and binding organisms, often termed 'problematica', of

Figure 10 Permian Capitan Reef Complex, McKittrick Canyon, New Mexico. Reef facies (massive face at upper left); fore reef talus (comprising lower ground and area to the right).

Figure 11 Triassic reef associated facies, talus, and bedded back reef lagoonal sediments. Dolomites of northern Italy.

which one such is 'Tubiphytes'. Later reefs were made by calcareous sponges, corals, red algae, and problematic calcareous binders. Flügel emphasizes that although Late Triassic reefs exhibit some 'modern' traits as would be found in recent coral thickets, there are some important differences too. Modern corals are primarily constructors, whereas Triassic ones seem to have been bafflers and binders. There would have been fewer nooks and crannies in Triassic reefs, and there appears to have been significantly less by way of bioerosion, predation, and grazing. Finally, significant numbers of Triassic corals were adapted to colonizing unstable sandy and muddy substrates, whereas modern species require a firm seafloor.

Following their initial setback resulting from the end-Triassic mass extinction reefs, often with corals as a significant component, reappeared in the Early Jurassic, firstly in Morocco (Figure 12). As sea-level rose eustatically through the period, to reach a peak in the Oxfordian, so reef fortunes rose too. With the break-up of Pangaea and the formation of new seaways, reefs proliferated. Coral reefs are confined between about 30° N and S during the Early Jurassic, expanding to 35° N in the mid-Jurassic and to 45° N in the Late Jurassic. Resedimented coral boulders occur within Kimmeridgian deep-water debris flows on the margins of the North Sea in northern Scotland. Jurassic reefs were constructed by a range of organisms (sponges, microbes, bivalves (e.g., oysters and ancestors to the Cretaceous rudistids)), as well as corals.

The Cretaceous (Figures 13 and 14) saw the rise of rudists, heterodont bivalves that became very important reef builders that underwent a great proliferation after the Early Aptian (caprinids and caprotinids), particularly through the Tethyan region. They range overall from end-Jurassic to end-Cretaceous, once again tracking the general trend in eustatic change

Figure 12 Lower Jurassic reef facies, Morocco. Photograph by Professor A Hallam.

Figure 13 Cretaceous rudist reefs limestones, Mount Carmel, Israel.

Figure 14 View of *in situ* rudists within Carmel reef wackestones, Israel.

from Early Cretaceous lowstand to Late Cretaceous highstand and subsequent fall. Another group of rudists, the radiolitids, appear in the Aptian and underwent a major proliferation through the Late Cretaceous, despite a setback at the Cenomanian–Turonian boundary. From the Cenomanian there is a great expansion in the geographical range occupied by rudistids with a new group, the hippuritids, markedly pipe-like forms, contributing in a major way to build-ups. In the Early Cretaceous the rudistids are associated with corals, but there is an increasing tendency for ecosystems to become rudist-specific, to the extent that biological mechanisms have often been invoked to explain the dominance of single species within communities. Rudist reef complexes are associated with high-energy shelf grainstones and platform margin talus, and around the margins of the Gulf of Mexico and in the Middle East (Arabian/Persian Gulf) areas these build-ups, and their associated facies, are associated with prolific hydrocarbon production. Rudist reefs were confined to between about 30° N and S during the Late Cretaceous, whereas reefs with corals extended further polewards (by about 10° further of latitude). However, reef communities suffered badly at the Cretaceous–Tertiary boundary and, although mud-mound and bryozoan build-ups occur in the Danian (e.g., in Denmark) and through the Eocene, reef systems did not fully recover until the Oligocene when they took on a nearly modern character.

In the Tertiary reef systems reached their maximum northern extent during the Miocene (Burdigalian (**Figure 15**) to Messinian (**Figures 16, 17 and 18**)) when extensive reefs grew in the Mediterranean area. Tortonian reefs have a diverse coral biota, but species were progressively lost into the Messinian in this area where monospecific associations (dominated by the tolerant genus *Porites*) prevailed shortly

Figure 15 Fossilized fringing reef (wedge shaped unit thickening seaward from below small tower). Reef prograded over talus and grew from a fossil cliff line comprising Triassic Muschelkalk Dolomites (foreground). Burdigalian (Miocene), Banyalbufar, Mallorca, Spain.

Figure 16 Messinian (Miocene) prereef carbonate sands (white clinoforms just above sea level); reef (massive brownish limestone in mid section of cliff) and well bedded postreef sediments above. Cap Blanc, Mallorca, Spain.

Figure 17 Finger *Porites* framestone; lower reef front. Messinian (Miocene), Cap Blanc, Mallorca, Spain.

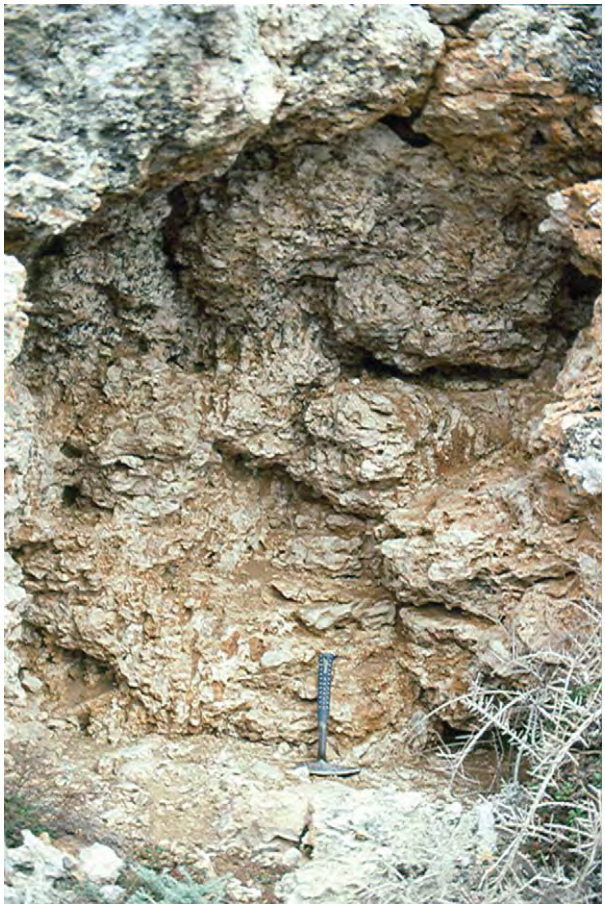


Figure 18 Massive *Porites* framestone, upper reef front. Mes sinian, Cap Blanc, Mallorca, Spain.

before the onset of the 'salinity crisis'. Debate still continues as to whether lowering temperatures or rising salinity caused these restrictions, but following the resumption of normal marine sedimentation in the Pliocene coral reefs were never again developed at these latitudes. Today, Bermuda and the Gulf of Suez represent the most northerly outposts of shallow, warm-water reefs in the northern hemisphere, and in the latter reefs grow on the maritime fringes of alluvial fans, a situation which also occurred along this rift during the mid-Miocene.

In deep cold waters lacking a high turbidity, and with slow sedimentation rates, coral reefs do grow today, but these are made by the ahermatypic (non-zooxanthellate) coral *Lophelia*. These form major edifices along the Atlantic margins, but their growth rates are slow and they are extremely susceptible to disturbances from fishing and other offshore activities, which is why they are protected by nature conservancy measures.

See Also

Fossil Invertebrates: Brachiopods; Corals and Other Cnidaria; Porifera. **Mesozoic:** Triassic; Jurassic; Cretaceous; End Cretaceous Extinctions. **Palaeoclimates**. **Palaeozoic:** Cambrian; Silurian; Devonian; Carboniferous; Permian; End Permian Extinctions. **Sedimentary Environments:** Carbonate Shorelines and Shelves. **Sedimentary Rocks:** Limestones.

Further Reading

Bekker A, Holland HD, Wang PL, *et al.* (2004) Dating the rise of atmospheric oxygen. *Nature* 427: 117 120.

Bonney TG (ed.) (1904) *The Atoll of Funafuti*. London: Royal Society.

Cestari R and Sartorio D (1995) *Rudists and Facies of the Periadriatic Domain*. San Donato Milanese: Agip S.p.A.

Darwin C (1842) *Structure and Distribution of Coral Reefs*. Reprinted 1962 by University of California Press with foreword by HW Menard.

Embry AF and Klovan JE (1971) The Late Devonian reef tract on northern Banks Island, N.W.T. *Bulletin of Can adian Petroleum Geology* 19: 730 781.

Geldsetzer HHJ, James NP, and Tebbutt GE (1988) Reefs, Canada and adjacent areas. *Memoir of the Canadian Society of Petroleum Geologists* 13.

Heckel PH (1974) Carbonate buildups in the geological record: A review. In: Laporte LF (ed.) *Reefs in Space and Time*. Special Publication, vol. 18, pp. 90 154. Tulsa: Society of Economic Paleontologists and Mineral ogists.

Hollingworth NTJ and Tucker ME (1987) The Upper Per mian (Zechstein) Tunstall reef of north east England: Palaeoecology and early diagenesis. In: Peryt TM (ed.) *The Zechstein Facies in Europe*. Lecture Notes in Earth Science, vol. 10, pp. 23 50. Berlin: Springer Verlag.

James NP (1983) Reef environment. In: Scholle PA, Bebout DG, and Moore CH (eds.) Carbonate depositional envir onments. *Memoirs of the American Association of Petroleum Geologists* 33: 345 440.

James NP and Ginsburg RN (1979) *The Seaward Margin of Belize Barrier and Atoll Reefs*. Special Publication of the International Association of Sedimentologists, 3.

Kiessling W, Flügel E, and Golonka J (eds.) (2002) *Phaner ozoic Reef Patterns*. Special Publication. Tulsa: Society of Economic Paleontologists and Mineralogists.

Longman MW (1981) A process approach to recognizing facies of reef complexes. In: Toomey DF (ed.) *European Reef Models*, Special Publication, vol. 30, pp. 9 40. Tulsa: Society of Economic Paleontologists and Mineral ogists.

Runnalls LA and Coleman ML (2003) Record of natural and anthropogenic changes in reef environments (Barba dos West indies) using laser ablation ICP MS and scler ochronology on coral cores. *Coral Reefs* 22: 416 426.

Scott RW (1990) *Models and stratigraphy of mid Cret aceous reef communities, Gulf of Mexico*. Concepts in

Sedimentology and Paleontology, 2. Tulsa: Society of Economic Paleontologists and Mineralogists.
Sellwood BW (1986) Shallow marine carbonate environments. In: Reading HG (ed.) *Sedimentary Environments and Facies*, pp. 283 342. Oxford: Blackwell Scientific.
Wood R (1999) *Reef Evolution*. Oxford: Oxford University Press.
Wright VP and Burchette TP (1996) Shallow water carbonate environments. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, pp. 325 394. Blackwell Science.

Shoreline and Shoreface Deposits

J Howell, University of Bergen, Bergen, Norway

Introduction

A shoreline is the interface between a standing body of water, typically the sea, and a landmass. Over geological time, shorelines are highly transient features and the depositional and erosional systems associated with them are extremely dynamic. The shoreface *sensu stricto* is defined as the zone that lies between the low tidemark and the fair weather wave base (the maximum depth at which everyday wave movements affect the seabed). The shoreface is part of a linked depositional system that includes all of the coastal region affected by wave action, from the top of the beach, seaward, to the mean storm wave base (the maximum typical depth at which the storm waves move sediment on the seafloor) (*see* **Sedimentary Environments:** Storms and Storm Deposits).

At the shoreline, wave processes are the most significant agents of erosion, sediment transport, and deposition, although fluvial, tidal, and biogenic processes are also important and may be locally dominant. Sediments deposited in the system may be siliciclastic, carbonate, or a mixture of the two, and grain sizes can range from those of clay particles to boulders. The position of the shoreline is highly sensitive to changes in sea-level and sediment supply, and understanding shoreline systems is central to sequence stratigraphy and basin analysis. Shoreface deposits form important hydrocarbon reservoirs and aquifers in many parts of the world, and some contain placer mineral and gem deposits. Modern systems are environmentally sensitive areas and many of the world's coastlines are heavily populated and susceptible to sea-level change. The study of shoreline processes and deposits includes aspects of geology, geomorphology, biology, and oceanography.

Both physical and biogenic processes (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves) occur at shoreline systems. All clastic, coastal depositional systems can be classified within a triangular diagram that summarizes the relative importance of fluvial, tidal, and wave processes (Figure 1A). Away from fluvial input points, the key controlling factor is the tidal range. In areas with a moderate to low tidal range, narrow beaches dominated by wave action occur. Conversely, in areas with a high tidal range, broad low-lying, muddy tidal flats are created (Figure 1B). The wave-dominated, beach-related systems that make up the majority of the world's coastlines are the focus here (but *see* **Sedimentary Processes:** Fluvial Geomorphology).

Wave Processes

Surface water waves are formed by the movement of the wind across a standing body of water. Wind shear effectively drags the upper layers of the water and horizontally compresses them, producing waves. The key features of a wave are its length and height (Figure 2). Within the water body, the vertical motion at the water surface sets up circular orbital motion that is translated downward. This downward translation of energy typical moves the water in a series of orbital motions to a depth that is approximately half of the wave length. This depth is termed the 'wave base' (Figure 2), and waves that occur in water that is deeper than the wave base will not move sediment on the seafloor. Because the depth to the wave base is controlled by the size of a wave, it varies both between basins and over time within an individual basin. In large, open bodies of water with a large fetch (such as the major oceans), the wave base may be in excess of 100 m, whereas in enclosed seas and lakes, it may be less than 10 m. Wave size and the depth to the wave base also vary within an individual setting on a daily and seasonal basis, depending on changes in the weather. During storms, short-term, larger waves result in a deep wave base, whereas during fair weather periods, the wave base is much shallower. It is common to consider marine depositional systems in terms of two hypothetical surfaces,

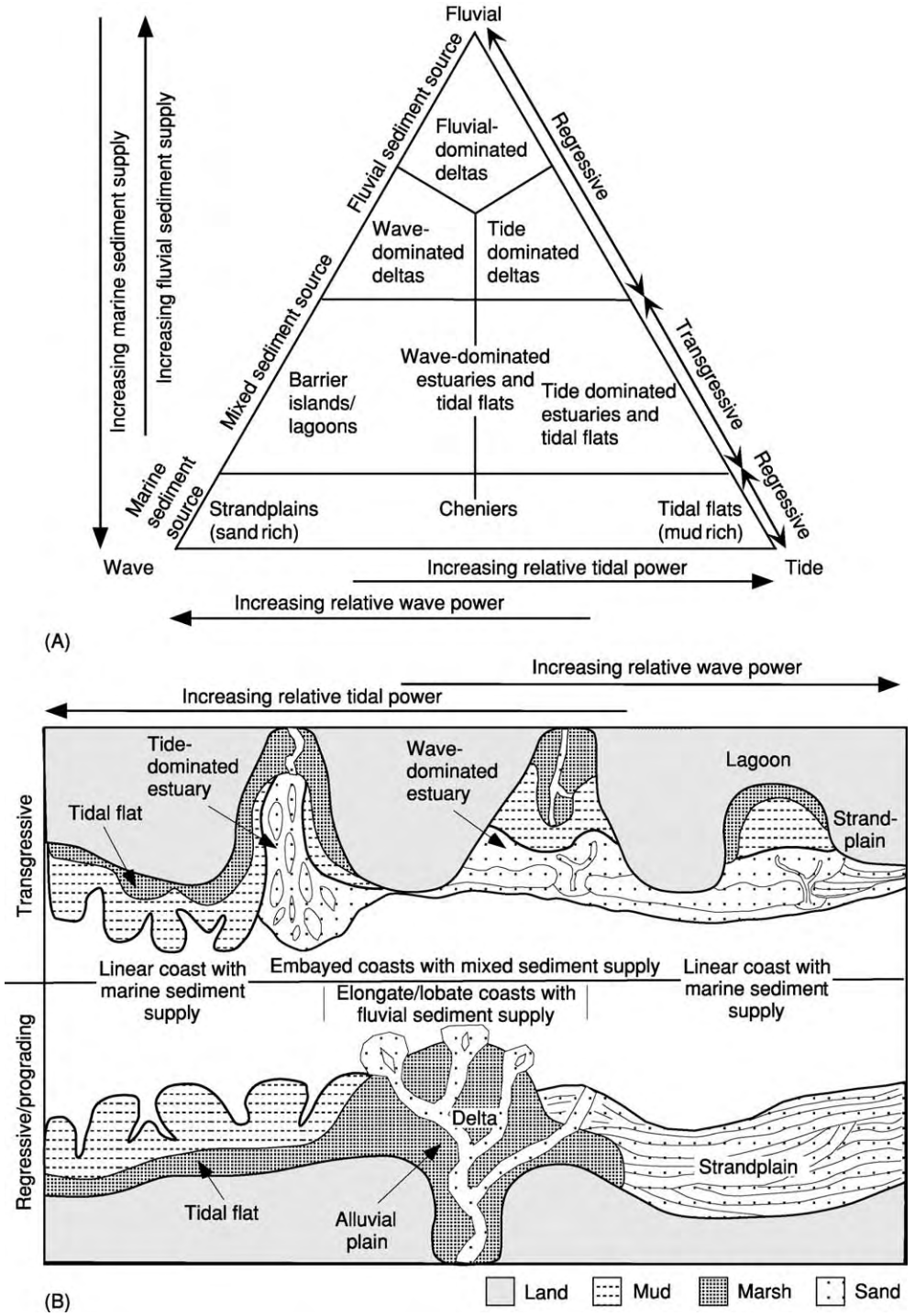


Figure 1 Classification of clastic shallow marine systems. (A) The relative importance of fluvial, tidal, and wave processes. (B) Transgressive and progradational coastal systems with varying degrees of fluvial, tidal, and wave influence. Reproduced with permission from Dalrymple RW, Zaitlin BA, and Boyd R (1992). Estuarine facies models: conceptual basis and stratigraphic implications. *Journal of Sedimentary Petrology* 62: 1130 1146.

the mean storm wave base (SWB) and the mean fair weather wave base (FWWB). The mean SWB is the depth at which typical storms affect the seabed for short periods of time (in the order of hours to days); the FWWB reflects the typical wave climate that is present within the basin between storm events.

Waves that are formed in the open ocean are pushed towards the land. As waves approach the coastline,

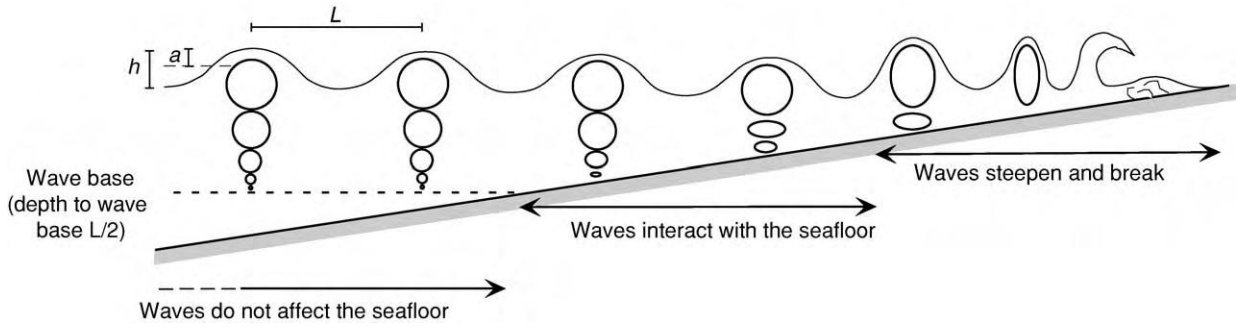


Figure 2 Key features related to water waves approaching a smooth dipping coast; *L* is the wave length, *h* is the wave height, and *a* is the wave amplitude.

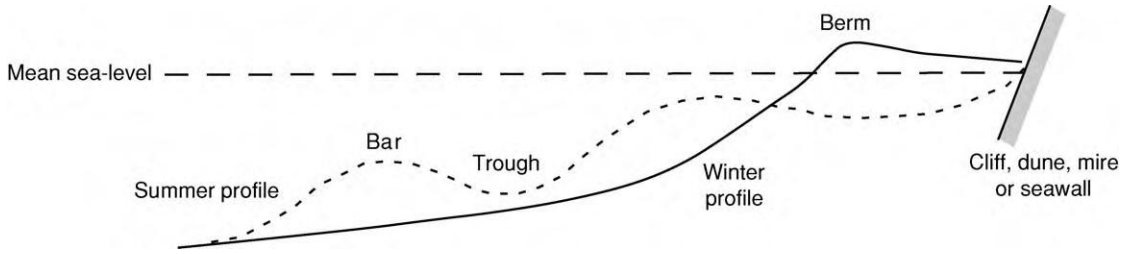


Figure 3 Typical summer and winter beach profiles. During the summer months, fair weather waves move sediment onshore and build up a berm. Winter storms erode the beach and deposit the sediment in the shoreface and offshore transition zone.

the water depth becomes less than the wave base and the waves interact with the seafloor. Waves hitting the coast may be classified as reflective, dissipative, or intermediate forms. Reflective waves strike very steep coastlines (with features such as cliffs or a seawall) and the majority of the energy is reflected directly back into the water body. Purely dissipative waves gradually lose their energy through frictional interaction with a very gently dipping seafloor and do not break at all. A majority of the waves that affect active shorelines are intermediate between these two forms. In intermediate cases, interaction with the seafloor slows the waves and pushes them up; the waves steepen until they eventually collapse or break. If the shoreline is steep, the waves will break as plunging forms, whereas if it inclines more gently, the waves will tend to surge in a landward direction. Steeper shorefaces with plunging waves tend to be simple systems in which the waves break directly on the beach. Most shorefaces dip gently and have a complex topography, including numerous bar forms over which the waves may break before reaching the shoreline, to reform and then break again. These systems include complex sets of plunging and surging waves.

During fair weather periods, waves tend to move sediment in an onshore direction because water loss due to percolation into permeable beach sediments reduces the efficiency of the backflow. Large storms tend to erode the beach and to deposit sediment offshore, or in the case of barrier systems, in the lagoon behind the beach. During the summer months, when few storms occur, there is a net onshore migration of sediment; during the winter storm period, the beach is frequently eroded and does not have time to be repaired. Consequently, typical summer and winter profiles for beaches may be defined (Figure 3). The preserved rock record is a complex mixture of the depositional products of the repeated seasonal changes.

If waves approach the shoreline perpendicularly, there is no net movement of sediment along the coast. However, in many cases, waves from the open ocean approach the shoreline obliquely. This results in an oblique onshore movement of sediment with the incoming wave. The returning swash tends to be more perpendicular to the shore, which in turn leads to migration of sediment parallel to the shore; this is termed 'longshore sediment transport', or longshore drift (Figure 4). Transport of sediment by longshore drift is a key factor in moving volumes of sediment and in controlling the morphology of wave-dominated shoreline systems.

The sedimentary structures produced by waves show an increase in bedform asymmetry towards the shoreline, passing from oscillatory, symmetrical wave ripples near the wave base into combined-flow wave ripples as the water becomes shallower (Figure 5). In the shallowest part of the system, shear between the water mass and the seabed can

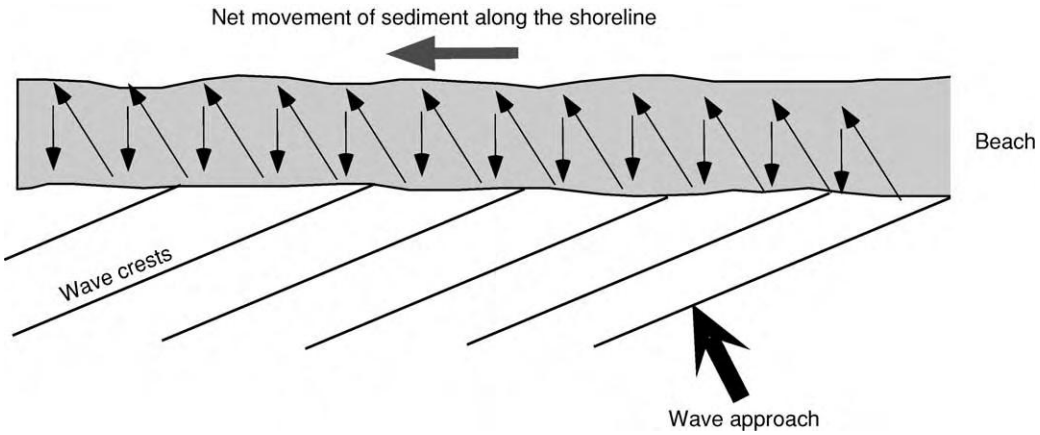


Figure 4 Longshore sediment transport (longshore drift). Waves approach the shoreline at an oblique angle. Sediment is moved up and across the beach by the incoming wave; the returning flow is weaker and tends to be straight back (i.e., shore normal). The net result is that sediment is moved along the coastline.

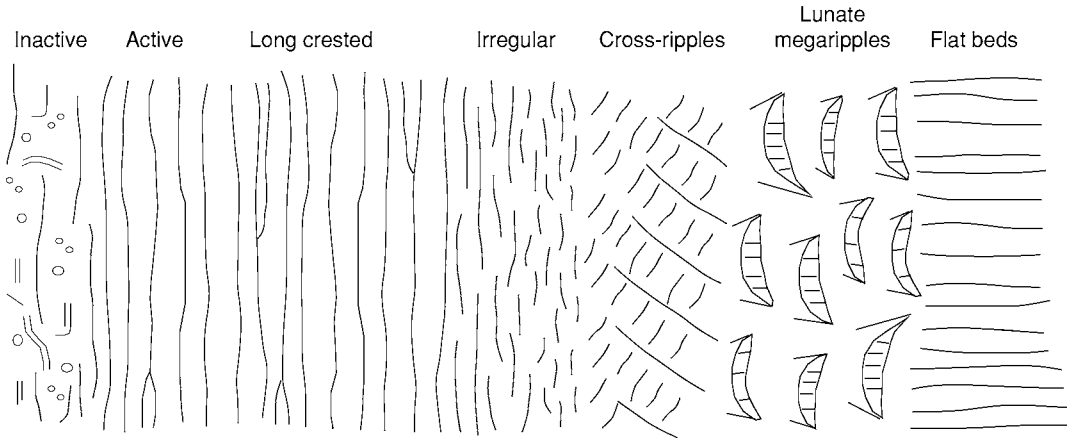


Figure 5 Plan view of the shoreface showing the distribution of bedforms resulting from the onshore approach of shoaling waves. Reproduced with permission from Clifton NE, Hunter RE, and Phillips RL (1971) Depositional structures and processes in the non barred high energy nearshore. *Journal of Sedimentary Petrology* 41: 651 670.

produce a variety of current-generated structures, including ripples, plane beds, and dunes, producing trough and tabular cross-bedding. Breaking waves generally produce upper-phase plane beds, although the returning swash current may also generate current ripples. In the deeper part of the system, combined flow associated with very large storm-related waves produces compound, three-dimensional low-angle bedforms called hummocks, which in turn produce hummocky cross-stratification (*see* **Sedimentary Environments:** Storms and Storm Deposits).

Classification and Geomorphology of Shoreline Systems

Geomorphologically, coastal systems can be characterized by the medium- to long-term movement of the shoreline (Figure 1B). In transgressive systems, the shoreline moves towards the hinterland and land is lost to the sea, whereas in regressive systems, the shoreline moves basinward and new land is created on the coastal plain. These movements are controlled by changes in sea-level and by the amount of sediment that is supplied to the system. Sea-level change may occur due to global changes in sea-level (eustasy) or may be driven by local uplift or subsidence within the basin. Commonly, a combination of the two forces is in play; this is termed 'relative sea-level change'. If relative sea-level is falling, the shoreline will move basinward (termed a 'forced' regression). Sea-level rise does not always lead to transgression, because the movement of the shoreline also depends on whether enough sediment is supplied to the system to fill the new space that is created. If more sediment is supplied than space is created, then the shoreline will prograde (i.e., will be regressive). If, however, the

rate of space creation is greater than the rate of sediment supply, then the shoreline will move landward (i.e., will be transgressive).

Transgressive shorelines may be further subdivided into low- and high-relief forms. Low-relief transgressive shorelines generally occur when relative sea-level rise floods a former delta or coastal plain. In such cases, systems may contribute to the geological rock record and their deposits will either be barrier islands in systems with a moderate to low tidal range, or tidal flats in systems with a higher tidal range. High-relief transgressive shorelines occur where the sea-level rise is flooding bedrock. In such cases, the systems are erosional. High-relief transgressive shorelines do not contribute to the geological rock record, although they are an important part of modern systems (see later).

Progradational Wave-Dominated Shoreline Systems

Progradational shoreline systems, which contribute significantly to the geological rock record, are classified by the dominant depositional mechanism: wave, fluvial (*see* **Sedimentary Processes:** Fluvial Geomorphology), or tidal. Progradational wave-dominated systems include shorefaces and wave-dominated deltas. In a shoreface, the sediment is introduced by longshore drift; in a wave-dominated delta, the sediment is originally introduce by a fluvial system and is then reworked by wave processes. In reality, especially in the rock record, it can be very difficult to distinguish the two, and because the depositional processes are the same, it is useful to consider them as largely synonymous. In both cases, the system includes a beach that lies between the low tidemark and the next significant feature in a landward direction, which may be dunes, a coastal forest, a mire, a cliff, or in many modern systems a seawall. Between the low tidemark and the FWWB is the shoreface, which includes the sediment that is being reworked by fair weather wave action. Seaward of the FWWB is the offshore transition zone where, for the majority of time, deposition consists of very fine-grained suspended sediment, with brief influxes of storm-derived coarser deposits. Seaward of the SWB is the offshore shelf where waves no longer play a role in the introduction or deposition of sediment (**Figure 6**).

Within a wave-dominated system, sediment is introduced by longshore drift, either from the reworking of a fluvial system (wave-dominated delta) or from wave erosion of the bedrock or older shoreline deposits. In plan view, a variety of features may exist within the depositional system. Offshore, seaward of the point at which the storm wave base interacts with the seafloor, deposition is restricted to mud deposited from suspension or to coarser sediments introduced by sediment gravity flows. In the zone between the SWB and FWWB, storms deposit hummocky cross-stratified

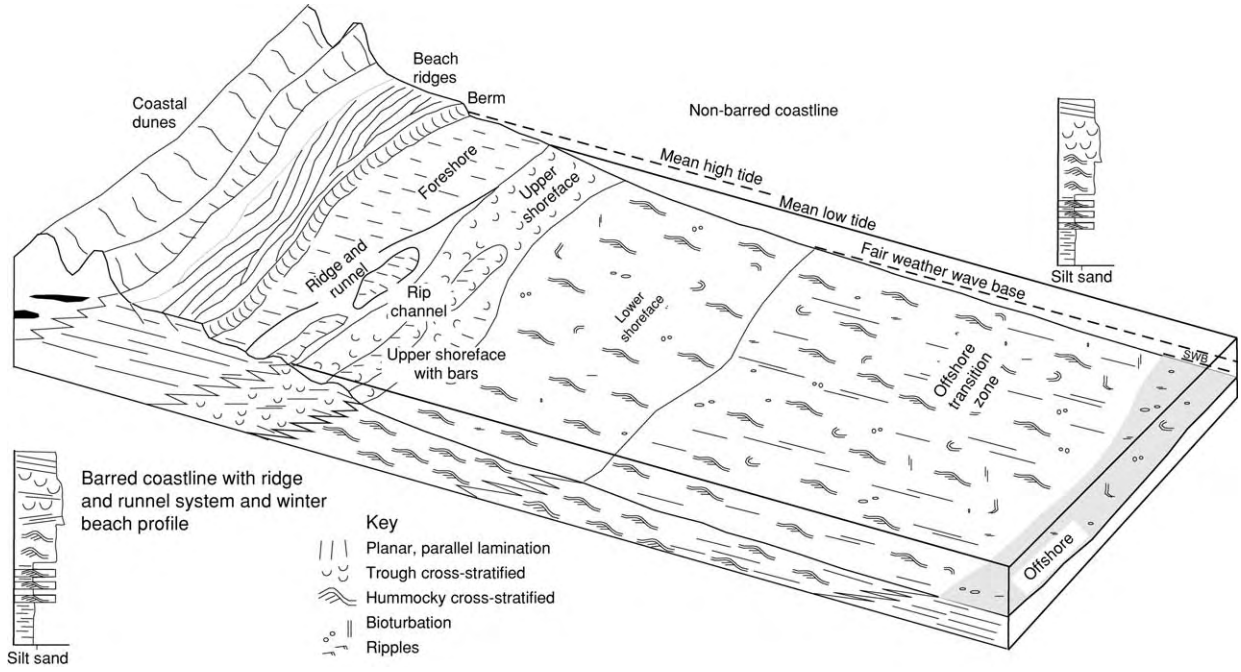


Figure 6 Facies model for a progradational wave dominated shoreline system. View shows both barred and non barred coastlines. Sediment is introduced by longshore drift. No scale is implied because the depths to the fair weather wave base and the storm wave base (SWB) are controlled by the size of the basin and the local climate.

(HCS) sheets of sandstone, typically less than 0.5 m thick; thin layers of coarser material (up to granule grade) may also be deposited if such material is available on the beach. These sheets have erosive bases whereas the top surfaces commonly exhibit wave ripples formed during the waning part of the storm. Bioturbation of the storm bed by opportunistic organisms is common. Between storm events, mud is deposited from suspension.

The shoreface is the interval that lies landward of the point at which the fair weather wave base interacts with the seafloor. Within the lower part of the shoreface, sediment is deposited as sheets of HCS sand during storms, and fair weather wave action prevents the deposition of mud between the storm events, resulting in a sand-rich depositional system. The point at which the fair weather waves actively move sediment in a landward direction marks the base of the upper shoreface. The upper shoreface may be a simple seaward-dipping surface or it may contain one or even several large bar forms. Simple, 'non-barred' shorelines are generally steeply dipping and coarse grained. They are covered with small dune forms that migrate in a landward and longshore direction; plunging waves generally produce seaward-dipping planar laminations on a relatively simple beach (**Figure 7**). Non-barred shorelines prograde through the steady addition of sediment to the beach and shoreface. Modern examples include the Pacific coast of North America. The Book Cliffs of Utah display superb Cretaceous examples of non-barred shoreline deposits (**Figure 7D, E, H,** and **J**).

Barred shorelines are more common than non-barred systems and contain large (up to 50 m wide and hundreds of metres long) shore-parallel barforms separated by narrow rip-channels that run perpendicular to the shoreline. Water and sediment are carried in an onshore direction by incoming waves that break over the bars and in an offshore direction by currents running through the rip-channels. The combination of these processes generates circulatory cells (termed 'longshore cells') within the shoreface. Barred shorelines are extremely dynamic and change markedly on a seasonal basis. Bars are generally built during the winter months as storms move sediment offshore from the beach. Bars can be symmetrical or asymmetrical and contain a series of seaward-dipping planer-beds deposited by storms on the seaward side and trough cross-bedding and landward-dipping planar-beds deposited by fair weather wave reworking. Rip-channels are either erosional or contain seaward-directed trough cross-bedding. During the summer months, the bars migrate landward and may emerge above the low tidemark, where they become part of the beach.

Within barred shorelines, the beach is more complex, compared to the simple seaward-dipping planar surface that exists in non-barred systems. Beaches within barred coastlines comprise a series of subzones, not all of which are always present. The lower intertidal portion includes a series of low, shore-parallel barforms, or 'ridges', separated from one another by hollows, or 'runnels' (**Figure 6**). Sedimentary features within this ridge and runnel system include low-angle dipping planar laminations on the ridges and shore-parallel current ripples formed by the draining of the runnels as the tide recedes. These ridge and runnel systems generally develop during the winter months when storms occur most frequently, but may also be related to the emergence of landward-migrating bars during the summer period.

The main part of the intertidal zone is termed the 'foreshore' and comprises a series of seaward-dipping planar surfaces formed by the breaking waves. The mean grain size of the sediment will strongly control the dip of the foreshore. If the grain size is coarse, the water brought onshore by the breaking waves will rapidly percolate into the sediment, reducing the backwash and producing a steep beach. Conversely, if the sediment is fine and less permeable, the beach will be gently dipping. The top of the foreshore is marked by a ridge called a 'berm'. The berm builds during the summer months as the fair weather waves push sediment onto the beach. In the absence of large storms, beaches may develop a series of parallel berms.

Beyond the berm lies the backshore. This is the area above the mean high tidemark, where sediment is introduced only during extremely high tides or by storms. If a lot of coarse (cobble to pebble grade) material is available, a storm ridge may be formed. This is typically composed of well-rounded, well-sorted, discoid clasts with a seaward imbrication that are thrown up during storms and are coarse enough to resist fair weather reworking. In the absence of a storm ridge, the backshore may comprise partially vegetated sand flats in which aeolian processes rework the sediment that is infrequently washed in.

Vertical Succession through a Wave-Dominated Shoreface Succession

The vertical succession deposited during the progradation of a wave-dominated shoreline succession reflects the lateral facies relationships as described. The onshore coarsening of grain size and increasing bedform asymmetry are recorded vertically in the rock record (**Figure 6**). A typical vertical succession from base to top will exhibit mudstones and siltstones deposited in the offshore setting at the base. Moving upward, the appearance of the first significant

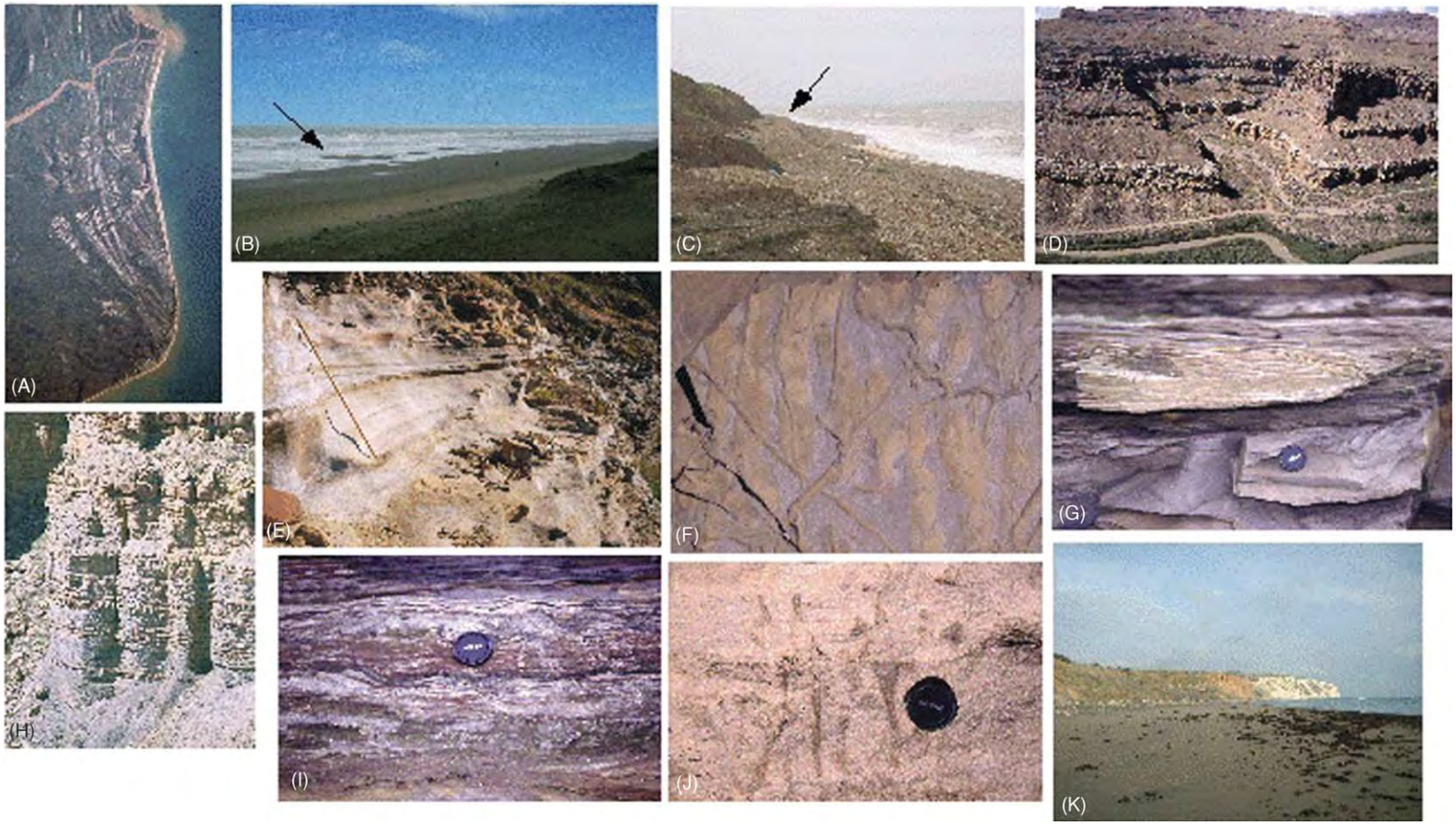


Figure 7 Modern and ancient wave dominated coastlines. (A) Space shuttle photo showing the linear shoreline and well developed beach ridges in the wave dominated Paraiba Delta, Brazil. Image courtesy of Earth Sciences and Image Analysis Laboratory, National Aeronautics and Space Administration Johnson Space Centre (see website at http://eol.jsc.nasa.gov). (B) Modern wave dominated beach (Dutch coast); note the low relief of the foreshore, the dunes on the left hand side, and the ridge and runnel system (arrow) at the lower part of the beach. Photograph by Katharine Mathers; reproduced with permission. (C) Modern coarse grained, non barred wave dominated shoreline (Isle of Wight, UK); note the well developed berm (arrow) and the steep dip of the foreshore. (D) Book Cliffs of eastern Utah, USA, showing world class outcrops of wave dominated shoreface deposits. Laterally continuous, white sandstones are foreshore and upper shoreface deposits of successive progradational episodes. (E) Upper shoreface and foreshore deposits from the Book Cliffs of eastern Utah. The low angle dipping planar laminations were deposited by waves breaking on a beach, and the trough cross stratified sandstones (base of picture) were deposited by shoaling waves in the upper shoreface (pole = 1.2 m). (F) Bioturbated and rippled lower shoreface sandstones (Wasatch Plateau, Utah). (G) Hummocky cross stratified sandstones from the offshore transition zone. (Carboniferous; Berwick, UK). (H) Offshore transition zone deposits comprising interbedded hummocky cross stratified sandstones (brown) and siltstones (grey), from the Book Cliffs of eastern Utah. (I) Bioturbated, wave rippled heteroliths from the offshore transition zone (Berwick, UK). (J) Bioturbated shoreface deposits. Low angle, swaley cross stratification cut by large, vertical *Rosselia* burrows (Wasatch Plateau, Utah). (K) Modern erosive beach from the Isle of Wight. Note the landslip deposits in the midground and the steep, retreating cliffs of chalk in the background. The foreshore is covered in a thin veneer of sand reworked from the cliffs.

hummocky cross-stratified storm beds represents the SWB and marks the base of the offshore transition zone (OTZ), which is characterized by interbedded sandstones with HCS and bioturbated mudstones. There is an upward increase in the thickness and the proportion of sandstone beds within the OTZ to the point at which they become amalgamated and no significant mudstones interbeds are observed. This point in the succession represents the position of the FWWB. The overlying amalgamated storm deposits are deposits of the lower shoreface. The lower shoreface also locally includes intervals that are heavily bioturbated, and top-truncated hummocks sometimes termed 'swaley cross-stratification' (SCS).

As previously noted, the upper part of the shoreface in a non-barred coastline will comprise simple landward and longshore migrating dunes that produce trough cross-bedding; this cross-bedding passes up into the seaward-inclined planar cross-bedding that represents the foreshore. In a barred coastline system, the vertical succession will be more complex, with a series of landward- and seaward-dipping planar stratified beds, interbedded with trough and ripple cross-stratified units and frequent erosion surfaces. In such cases, it may be difficult to distinguish the upper shoreface from the beach, except by the vertical transition to the overlying backshore deposits.

Muddy Coastlines and Cheniers

The majority of depositional wave-dominated shorelines are sand rich because waves will rework mud-grade material, moving it offshore in suspension. However, if mud is the only material present and is available in large enough quantities, muddy wave-dominated coasts can occur. Such systems build very large, low-angle dipping muddy barforms that are held together by the cohesiveness of the sediment. The very low angle dissipates all wave action and prevents wave erosion. Periodic influxes of sand and/or concentrations of shell debris can produce seaward-dipping beds associated with prominent ridges on the beach. These are termed 'cheniers'. Such systems are seen on the Surinam coast of South America and in areas adjacent to the Mississippi delta.

Low-Relief Transgressive Coastlines

Low-relief transgressive shorelines occur when the rate of sediment supply is less than the rate of relative sea-level rise and the coastal zone is a low-lying feature, typically a former depositional surface such as an abandoned delta top or coastal plain. In systems with a moderate to low tidal range, the shoreline will generally develop barrier islands. Barrier islands are linear, shore-parallel strips of land that separate the open ocean from a sheltered lagoon or intertidal flat area; the intertidal area may be connected to the marine realm through tidal inlets. There are extensively developed barrier islands on the Texas gulf coast (Figure 8). Many of the processes that are active on the seaward side of the barrier are identical to those described for the shoreface and beach systems in the previous section. The key difference is that

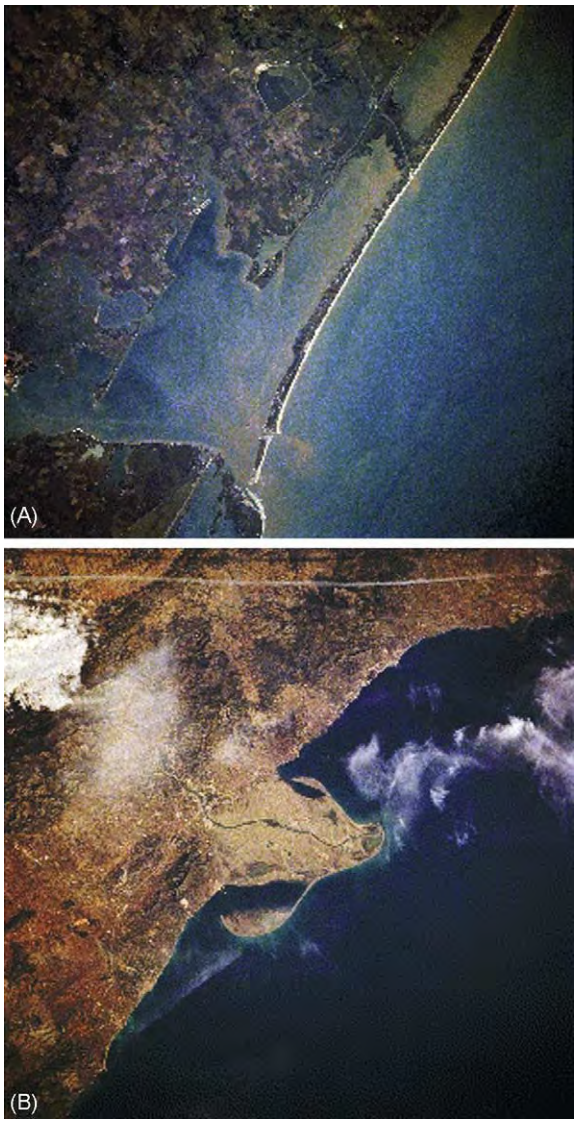


Figure 8 Space shuttle photographs of transgressive wave dominated shorelines. (A) Matagorda Bay, Texas, USA. A narrow barrier island system shelters the lagoon from the open ocean. The irregular left hand side of the barrier island is due to washover fans. Note the small tidal inlets towards the base of the photo. (B) Ebro Delta, Spain. Thirty five dams along the river have signifi cantly reduced the sediment supply to the delta, which has subse quently been transgressed due to relative sea level rise caused by compactional driven subsidence. Note the development of spits and barriers on the wings of the delta. Both images courtesy of Earth Sciences and Image Analysis Laboratory, National Aeronautics and Space Administration Johnson Space Centre.

during large storms, some of the sediment is washed into the lagoons and is deposited as sheets of planar bedded and rippled deposits; these are termed 'washover fans' (Figure 9). Because sediment is removed from the seaward side of the barrier and is deposited on the landward side, the shoreline moves landward. This process, termed barrier 'rollover', can be relatively rapid. Parts of the modern-day Galveston Island in Texas have migrated 80 m in the past 100 years. Generally, the migrating barrier system leaves very little sediment behind and the geological record of the transgression is an erosive surface, often associated with a lag of pebble or shell debris, concentrating the coarsest material from the underlying unit. The barrier island will be preserved at the point at which it stops migrating landward and the system starts to prograde. If the sea-level rise is very rapid, the entire barrier complex may be preserved by drowning.

Barrier island systems that occur in mesotidal settings are cut by tidal inlets through which tidal flows connect the lagoon with the open marine realm. In such cases, the barrier is also locally reworked by tidal currents passing through inlets. Deposition associated with the tidal inlet will include small subaqueous deltas on the landward and seaward sides, termed 'flood' and 'ebb' tidal deltas, respectively. These deltas are composed of rippled and planar cross-stratified sandstones that interfinger with the shoreface deposits on the seaward side and with the lagoonal muds and washover deposits on the landward side. Deposition of trough and tabular cross-bedding, often showing tidal draping, may also occur within the inlet channel. This will frequently be reworked but can be preserved if the position of the inlet migrates laterally.

Although numerous authors have discussed progradational barrier systems, especially in modern settings, geologically such features are very short lived; they represent the transition from transgressive barrier systems to land-attached shorelines because, if the barrier progrades, the lagoon will rapidly (over geological time) be infilled. Progradational systems with a strong component of longshore sediment transport can also develop beach-attached spits and even barrier systems that also shelter a lagoon on their landward side; however, these lagoons may be quickly filled by sediment that is introduced from the landward side by runoff and is washed over by storms from the seaward side.

High-Relief Transgressive Coastlines

High-relief coastlines may occur when the system is in net erosion and the wave action produces a cliff. The dominant processes are wave action and mass wasting. Fluvial and tidal processes are negligible. Wave action undercuts the cliff, which eventually collapses onto the foreshore; the cliff sediment is then reworked by wave processes and is transported either offshore or along the coastline by longshore drift. In its simplest form, a rocky coastline typically comprises a cliff and a wave-cut platform (Figure 10). The nature of the bedrock geology, rates of relative sea-level fall, and the wave climate will characterize any particular example. Variations in the bedrock geology due to lithological variations and structural geology (presence of faults and fractures) will result in differential erosion, leading to a variety of landforms, including stumps, stacks, arches, and caves

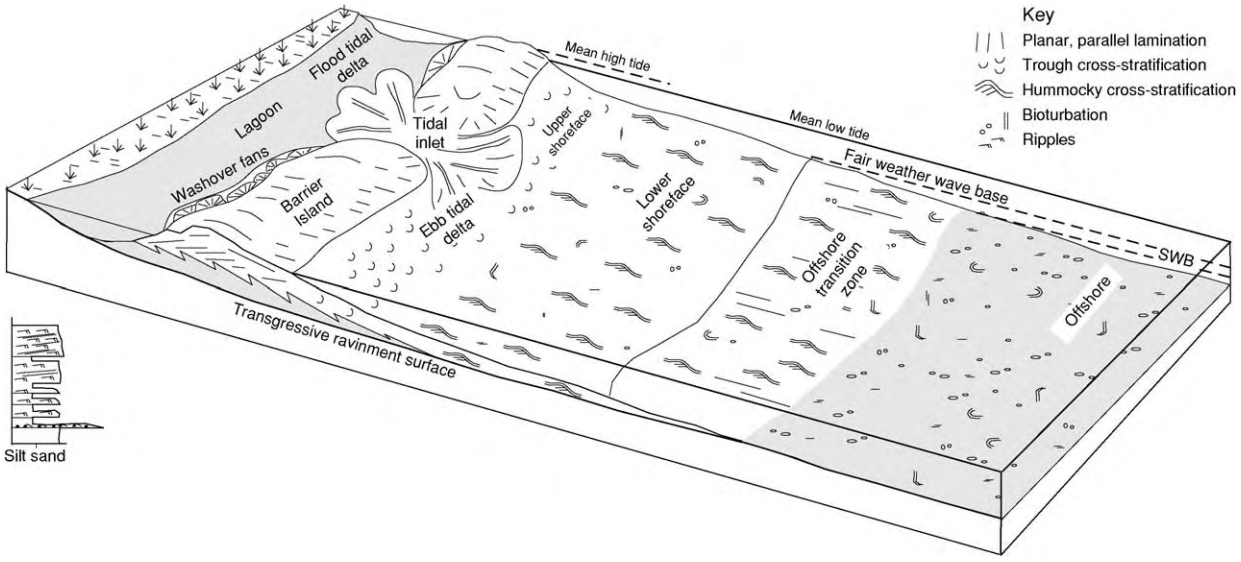


Figure 9 Facies model for a transgressive wave dominated shoreline system. The barrier island migrates landward through erosion on the seaward side and by deposition in washover fans and flood tidal deltas. SWB, Storm wave base.

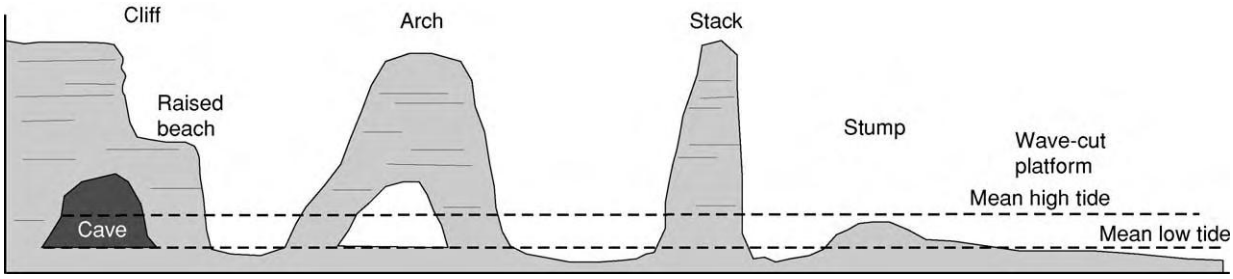


Figure 10 Geographical features of a high relief erosional coastline.

(Figure 10). Changes in relative sea-level may lead to the formation of wave-cut terraces or marked wave-cut notches.

The geomorphological character of a cliff is a product of the bedrock geology and the climate. Hard, resistant bedrock will result in high, steep cliffs that retreat slowly and show pronounced wave-cut notches at the base; softer deposits will result in lower, less steep, rapidly retreating clifflines. The process of cliff retreat is enhanced in a wetter climate and is retarded in more arid regions. The wave-cut platform is a horizontal to low-angle seaward-dipping surface that occurs between mean high watermark and wave base. The wave-cut platform may be covered in a thin veneer of reworked sediment or may be bare. Where present, sediment may range in grain size from sand particles to boulders. Clasts are generally well rounded due to frequent reworking by waves. Wave movement of the retreating cliffline-derived sediment is a key process in the scouring of the platform.

As erosional features, rocky coastlines have little preservation potential in the geological rock record, although they may account for significant amounts of geological time. Preservation will be limited to an erosional surface that records the final position of the shoreline prior to a change in conditions that led to the onset of deposition. Immediately above this surface a thin veneer of sediment may be preserved.

See Also

Sedimentary Environments: Carbonate Shorelines and Shelves; Storms and Storm Deposits. **Sedimentary Processes:** Fluvial Geomorphology.

Further Reading

Boyd R, Dalrymple RW, and Zaitlin BA (1992) Classification of clastic coastal depositional environments. *Sedimentary Geology* 80: 139 150.

Coe A (ed.) *The Sedimentary Record of Sea Level Change*. Cambridge: Cambridge University Press.

Clifton HE, Hunter RE, and Phillips RL (1971) Depositional structures and processes in the non barred high energy nearshore. *Journal of Sedimentary Petrology* 41(3): 651 670.

Davidson Arnott RGD and Greenwood B (1976) Facies relationships on a barred coast, Kouchibouguac Bay, New Brunswick, Canada. *Beach and Nearshore Sedimentation. SEPM Special Publication No. 24*, pp. 149 168. Tulsa, OK: Society for Sedimentary Geology.

Davis RA and Fitzgerald DM (2004) *Beaches and Coasts*. London: Blackwell Science.

Davis RA, Fox WT, Hayes MO, and Boothroyd JC (1972) Comparison of ridge and runnel systems in tidal and non tidal environments. *Journal of Sedimentary Petrology* 40(2): 413 421.

Dominguez JML, Bittencourt ACSP, and Martin L (1992) Controls on Quaternary coastal evolution of the east northeastern coast of Brazil: roles of sea level history, trade winds and climate. *Sedimentary Geology* 80: 213 232.

Elliott T (1986) Clastic shorelines. In: Reading HG (ed.) *Sedimentary Environments and Facies*, pp. 143 177. London: Blackwell Scientific Publications.

Heward AP (1981) A review of wave dominated clastic shoreline deposits. *Earth Science Reviews* 17: 223 276.

Walker RG and Plint GA (1992) Wave and storm dominated shallow marine systems. In: Walker RG and James NP (eds.) *Facies Models: Response to Sea Level Change*, pp. 219 238. Ottawa: Geological Society of Canada.

Storms and Storm Deposits

P Myrow, Colorado College, Colorado Springs, CO, USA

Introduction

Numerous studies of modern shelves indicate that storms have wide-ranging effects in a spectrum of environments from outer shelves to coastal plains. Early oceanographic studies led to interest in finding analogues in the rock record. Resulting facies models for storm-influenced depositional systems were not published until the late 1970s and early 1980s, later than for many other depositional systems (e.g. fluvial, tidal-influenced shoreline, submarine fan). There are a number of reasons for this. First, the oceanographical study of modern storms is extremely difficult. Direct observations are challenging and dangerous. Quantitative analysis also required the development of current meters and other devices that could survive the effects of strong storm-generated currents and waves. Second, storms are extremely complex systems that include strong currents and waves, both of which vary both temporally and spatially during an individual storm. Third, although tempestites, individual storm-generated sandstone beds, were attributed to deposition from storm-generated currents fairly regularly by the mid-1970s, features considered to be diagnostic of deposition by storm processes, such as hummocky cross-stratification (HCS), were not identified early in the study of sedimentary structures. Many case studies of ancient storm-influenced deposits and modern settings indicate that storms have wide-ranging effects in a spectrum of environments from outer shelves to coastal plains. The most common storm deposits in the rock record are interbedded sandstone and bioturbated shale (Figure 1). The sandstone beds were deposited as a result of powerful near-bottom water motions produced by waves and currents, and these are the main focus of the following discussion. Additional storm-generated features include gutter casts, pot casts, gravel and conglomerate accumulations, and reworked or transported shell beds.

Figure 1 The deposits of storm influenced shelves consist of tempestites or storm generated sandstone beds and shale beds deposited between storm events. Note how the sandstone beds tend to pinch and swell in thickness, a common feature of tem pestites. These strata, exposed in the central Transantarctic mountains, were deposited along the East Antarctic continental margin during the Cambrian. Hammer for scale.

Early Facies Models

Early studies established the basic idea that storms result in the transport of sediment from generally well-sorted shoreline sources onto shelves, as deduced from lithofacies patterns of decreasing bed thickness and evidence for less powerful currents from onshore to offshore. Despite more than 30 years of work on storm-influenced shorelines and ancient deposits, there remains considerable debate concerning the processes by which sediment is transported seawards during storms. Early discussions of storm processes produced two radically different perspectives from oceanographers and from geologists who worked on ancient deposits. Stratigraphical and sedimentological data indicated a potential role of density-induced forces, in part because many tempestites closely resemble turbidites. The first facies model of storm-influenced shelves invoked the mobilization of sand at the shoreline, the creation of dense dispersions, and subsequent gravity-driven offshore transport. Data from many ancient deposits indicate transport nearly perpendicular to the shore for long distances. In this view, storm-deposited sandstone beds are indicative of nonuniformitarian processes, and the very rare storm events responsible for these beds (with recurrence intervals of tens of thousands of years) are presumably significantly different (quantitatively if not qualitatively) from those studied by oceanographers.

Oceanographic Studies

Although the results of major storms on modern shorelines were first described in the late 1960s, the dynamics of modern storm-generated flows were not outlined until the 1980s by oceanographers who described the effects of storms on shelf and nearshore circulation. Winds generated by storms move surface water, which is in turn deflected by Coriolis

acceleration. This deflection affects much of the water column and results in net transport at 90° to the wind direction (to the right in the northern hemisphere). One common consequence of such transport is that under certain wind conditions surface waters may move towards, and bank up against, the shoreline. The resulting sloping sea surface causes an onshore–offshore pressure gradient, which tends to drive flow offshore along the sea bottom. Coriolis force affects water particles moving in this flow at right angles to the flow direction, and an equilibrium is eventually established between the offshore-directed pressure gradient and the Coriolis force, such that the majority of storm-generated flows are alongshore. The surface flows are more landward oriented, and the bottom flows, which are affected by bottom friction, tend to be oriented at a slightly higher angle offshore.

The conflict between the oceanographical observations of shore-parallel geostrophic storm flows and the geological data that indicates both cross-shelf transport and the deposition of turbidite-like beds has never been fully resolved. Theoretical work and numerical simulations of combined-flow boundary layers (Figure 2) are important in addressing this problem and have been used by some workers to suggest that offshore transport of sand could take place during storms under combined flow. Since geostrophic currents are oriented slightly oblique to shore, the offshore component of the flow is added to the seawards strokes of storm waves and acts against the shorewards strokes, resulting in an asymmetry of shear stress. It was thus postulated that this asymmetry was responsible for the offshore sediment transport of sand and for the unimodal offshore-directed sole-mark orientations on the bases of tempestites. Studies in the ancient sedimentary record have affirmed the importance of geostrophic combined flows in some shoreface and nearshore storm deposits, but it remains unclear to what degree such flows are responsible for cross-shelf transport in modern systems or to what degree they were responsible for the deposition of ancient tempestites.

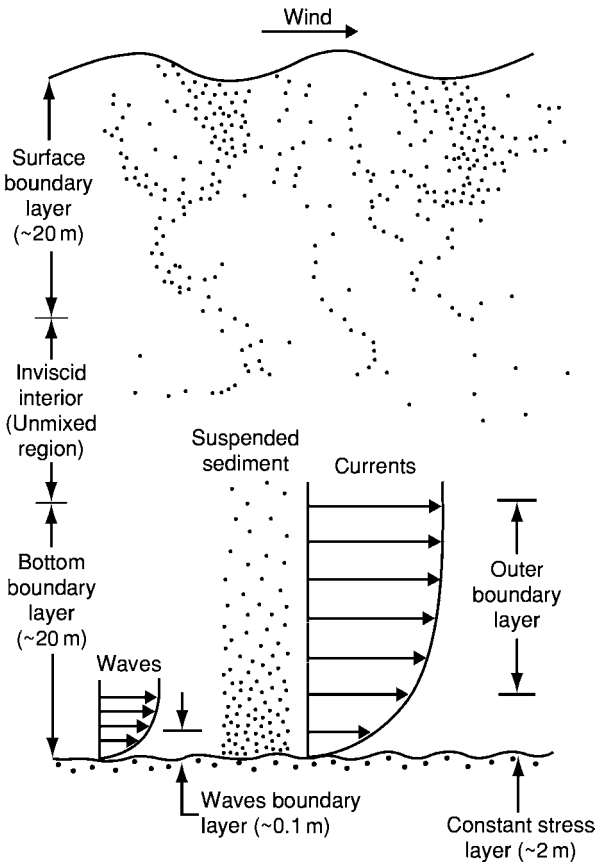


Figure 2 Subdivision of the water column into three layers during storms: a highly turbulent surface boundary layer, an inviscid interior or core flow layer, and a bottom boundary layer. The bottom boundary layer consists of a thin wave bound ary layer and a thicker outer boundary layer produced by cur rents. Modified from Trowbridge and Nowell (1994).

Interpreting Ancient Tempestites

The large number of case studies of tempestites has led in recent years to a greater appreciation of the potential variety of storm processes. A wide range of storm deposits result from spatial and temporal variability in the degree to which three fundamental processes operate: geostrophic currents, storm-generated waves, and excess-weight forces (Figure 3). For instance, some studies have shown that geostrophic flows are in fact agents of transport and deposition in tempestites, as determined in part by shore-parallel palaeocurrents. Such flows cannot explain successions with offshore-directed palaeocurrents or sole markings, such as 'classic' flute marks, that are inconsistent with formation in the presence of waves. In fact, the wide range of sole-mark geometries alone supports genesis from various types of flows. Bidirectional and multidirectional orientations of prod marks and groove marks indicate strong wave activity. Flutes, unimodal prods, and well-oriented grooves probably reflect erosion by powerful unidirectional flow. Finally, a range of sole marks including curvilinear and quasi-sinusoidal forms record erosion by a variety of combined flows.

The internal structures of tempestites are equally diverse and relate to the temporal and spatial histories of the three basic components of storm-related flow described earlier. Most tempestites contain evidence of flow deceleration in the form of either normal grading or particular successions of stratification. They also contain stratification representative of a wide range of bedforms, from purely unidirectional

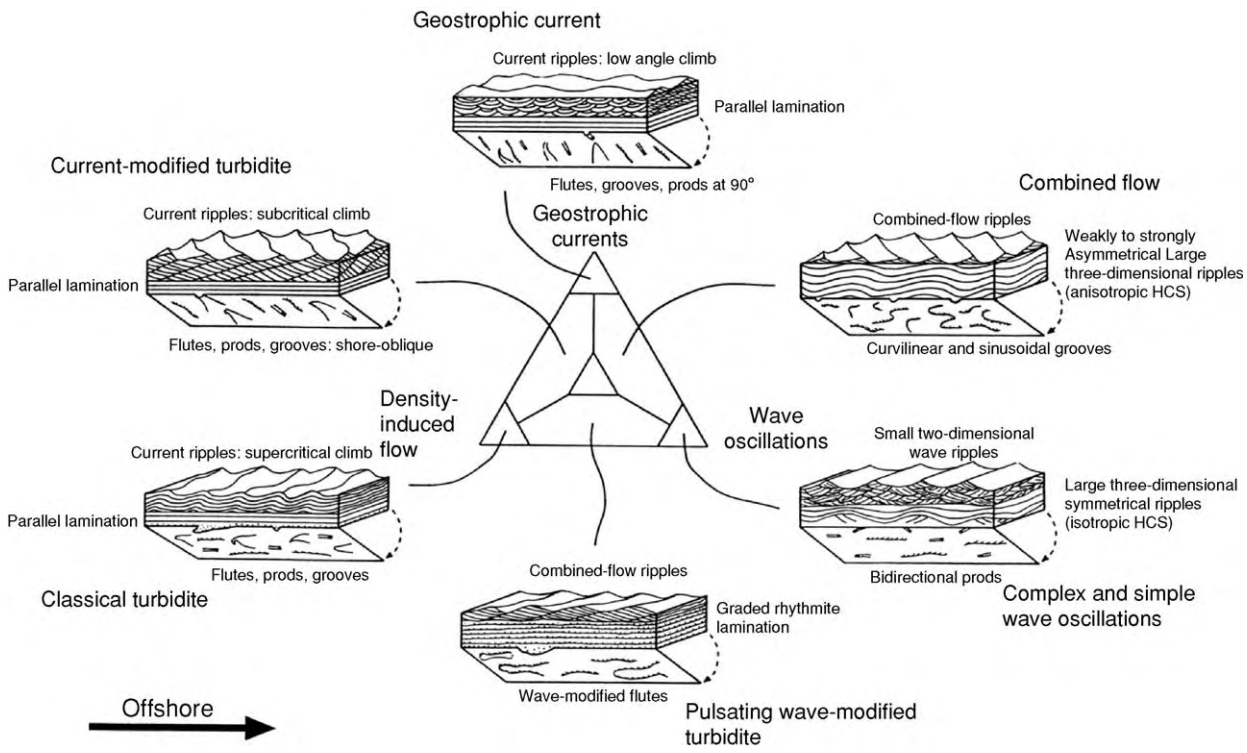


Figure 3 Heuristic model for the origin of stratification and sole markings for various possible combinations of storm processes. The diagram has three corners, representing density induced flow due to excess weight forces, geostrophic currents, and wave oscilla tions. The potential history of a storm event, in this three part system, is extremely variable. The deposits of flows near the corners are well understood, but the central triangle and intermediate fields are poorly understood since so little is known experimentally about these fields, except for some information on simple rectilinear combined flows. HCS, hummocky cross stratification. Reproduced from Myrow PM and Southard JB (1996) Tempestite deposition. *Journal of Sedimentary Research* 66: 875 887.

to purely oscillatory flow, including a spectrum of combined flows (Figure 4). The most diagnostic feature of storm deposition is HCS at various scales (Figures 5 and 6). The characteristics of HCS are well described and have been studied experimentally and with grain fabric analysis. There does not appear to be a size break in the bedforms that create HCS, namely three-dimensional, symmetrical to weakly asymmetrical, subaqueous dunes. Bedforms on modern shelves with HCS-like lamination have been attributed to combined flows, and experimental work has shown that hummocky bedforms are stable in complex oscillatory flows and wave-dominated combined flows (i.e. low velocities of current component).

Details of the geometry and internal structure of ancient tempestites indicate that beds range from those deposited in single events to those representing amalgamated accumulations produced by multiple events. The latter include thick beds that grade laterally into thinner beds separated by thin shale beds. These suggest that transport of sand onto shelves is in some cases incremental and involves a succession of storms. Some beds that appear to result from single events are potentially remobilized sand that simply records the last stage of transport and deposition.

Other tempestite beds show little or no evidence of amalgamation and are suggestive of powerful offshore-directed unidirectional flow during the erosional and initial depositional conditions of the storm. The nature of such flows is difficult to determine from the ancient. In many cases they may have been driven, at least in part, by excess-weight forces, caused by high near-bed suspended-sediment concentrations (SSCs). 'Sediment equilibrium' conditions exist when SSCs are caused by resuspension of sediment by storm-generated currents and waves. Very high SSCs (up to 4 g l ^{1}) and resulting powerful seaward-directed sediment-laden flows have been documented from shorefaces and inner-shelf settings. It is less clear to what distances such flows can carry sand offshore on the relatively low gradients common to modern continental shelves.

Excess-weight forces can also be generated extrinsically ('sediment disequilibrium' conditions), for example when rivers discharge sediment-laden floodwater into standing bodies of water and thus produce hyperpycnal (bottom-hugging) flows termed oceanic floods. Oceanic floods on the continental shelf of northern California adjacent to the Eel River have high SSCs (up to 2.5 g l ^{1}) of silt and

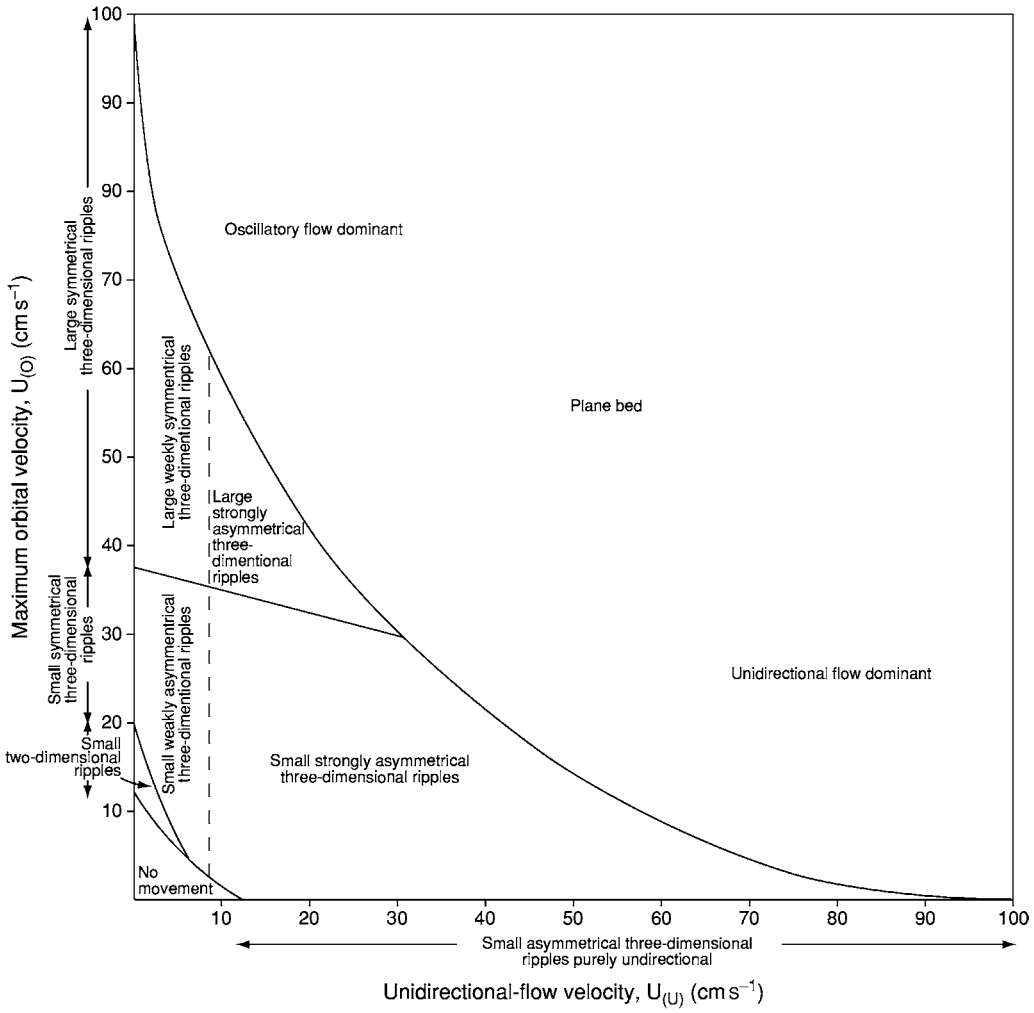


Figure 4 Bed configuration stability diagram for combined flows in fine sand. Based on published data and on data from Arnott RW and Southard JB (1990) Exploratory flow duct experiments on combined flow bed configurations, and some implications for interpret ing storm event stratification. *Journal of Sedimentary Petrology* 60: 211 219 and Southard *et al.* (1990). The left side of the diagram (up to U_U of 25 cm s^{-1}) is drawn from the data in Arnott and Southard (1990), and the right side is extrapolated.

clay even at mid-shelf depths (50–70 m). Such floods are potentially a geologically important process for cross-shelf sediment transport, although it is unknown whether they could transport very fine and fine sand, the grain sizes that dominate most ancient tempestites. Facies models for some ancient sandstone tempestites cite nearshore sediment bypass (Figure 7), which is consistent with powerful unidirectional flows in very proximal positions, possibly associated with hyperpycnal oceanic floods.

In both sediment equilibrium and sediment disequilibrium conditions sediment may be transported in combined flows produced by a combination of excess-weight forces and storm waves. Such wave-modified turbidity currents differ from classic turbidity currents in that storm waves enhance boundary-layer shear stress and eddy viscosity, which reduces deposition from suspension and thus enhances gravity-driven transport. As long as turbulent diffusion by waves does not destroy the density stratification, it provides the energy necessary to maintain gravity-driven flow on relatively low gradients. This process has been documented in the cross-shelf movement of modern fluid mud layers, but not for sand-sized sediment. Ancient examples of wave-modified turbidites have internal sedimentary structures, including combined-flow ripples, that reflect the influence of both waves and gravity-driven flow (Figure 8).

Whereas the preceding discussion has emphasized the range of possible processes responsible for tempestite deposition and thus has explained the wide variety of sole marks and internal sedimentary structures in the ancient, there are two important aspects of tempestites that remain controversial. First, beds in the ancient that are interbedded with shale and were therefore deposited below the fair-weather wave base

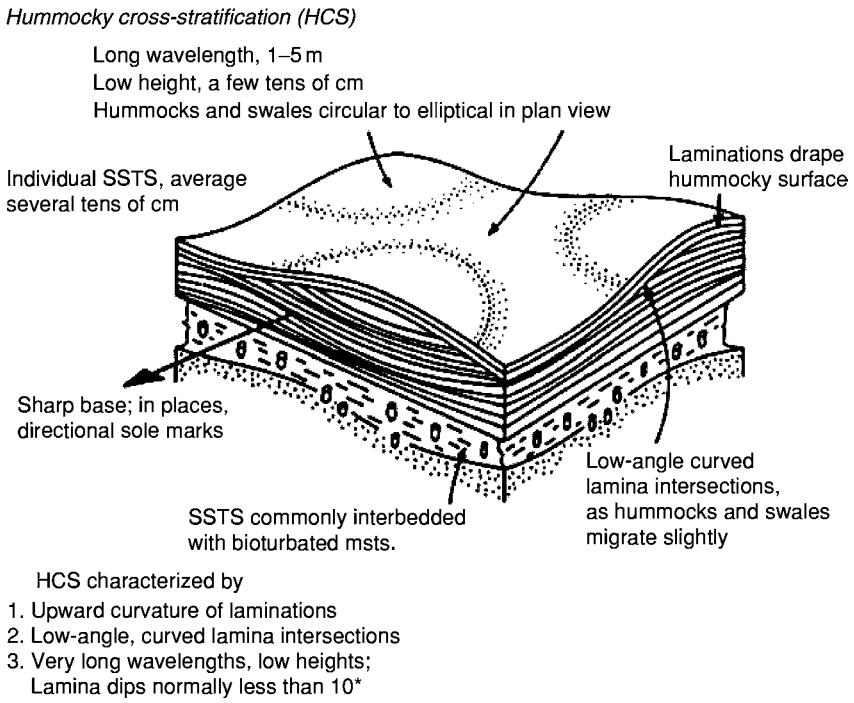


Figure 5 Block diagram illustrating the diagnostic features of hummocky cross stratification (HCS). Modified from Walker RG (1984) Shelf and shallow marine sands. In: Walker RG (ed.) *Facies Models*, 2nd edn., pp. 141 170. Reprint Series 1. Toronto: Geoscience Canada.

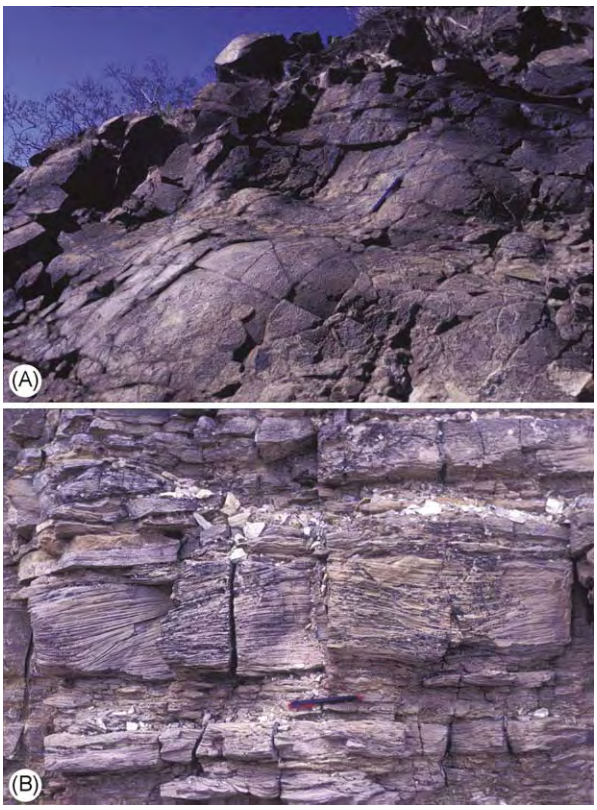


Figure 6 Hummocky cross stratification (HCS). (A) Bedding plane view of exhumed HCS bedforms. Note the rounded sym metrical hummocks and swales. From Neoproterozoic Johnny Formation, Death Valley, California, USA. Pencil for scale. (B) Bed showing classic HCS features including upward doming laminae, fanning of laminae, low angle lamination, and low angle curved erosion surfaces. From the Triassic Moenkopi Formation, central Utah, USA. Pencil for scale.

are considerably thicker than sand beds deposited on the shelf as a result of modern storms and appear to have a greater cross-shelf distribution than their modern counterparts. In unusual settings, such as the shallow (less than 20 m), flat, epicontinental shelf off the Yukon Delta in the Bering Sea, sand is carried offshore up to 100 km, although the beds range in (precompaction) thickness from 10–20 cm in the nearshore to less than 5 cm offshore. In this case, strong friction forces may have aided offshore transport, and thus such a setting may be a good analogue of epicontinental deposits, including those of the Cretaceous interior seaway of North America. In many modern settings sand beds of any appreciable thickness do not extend much below 5–10 m water depth. The thicknesses of modern storm beds fall in the small-to-average range for ancient tempestites, which commonly reach (compacted) thicknesses of many tens of centimetres and locally up to a metre or more. It is unclear to what degree thinner modern tempestites may simply reflect insufficient study of the modern, particularly since their record is commonly rapidly lost offshore owing to bioturbation. It

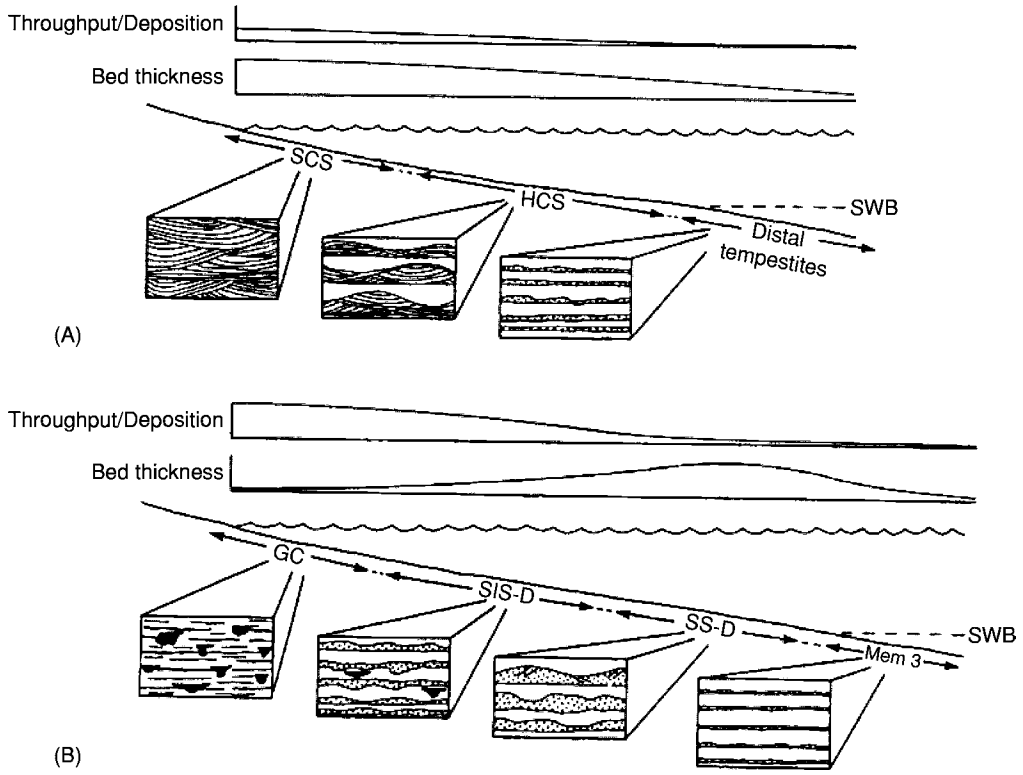


Figure 7 Comparison of (A) standard tempestite model and (B) bypass model (Myrow 1992). The spatial distribution of facies in each model is illustrated with onshore offshore profiles and associated qualitative changes in bed thickness. The degree of sediment bypass (throughput to deposition ratio) is plotted above each profile. (A) Proximality trends in the standard model include constantly decreasing bed thickness away from the shoreline with a shift from swaly cross stratified sandstone to hummocky cross stratified (HCS) sandstone to distal turbidite like beds. (B) The bypass model shows a bed thickness trend that first increases then decreases away from the shoreline. The proximal setting is one of bypass, with erosion being the dominant process during storms (many gutter casts). Passing seaward, gutter casts die out and bed thickness increases. HCS is formed in the thicker sandstone beds farther out on the shelf. Below the storm wave base (SWB) distal tempestites are turbidite like in character. The shelf gradient for the two models is shown to be the same simply for convenience.

is also difficult to establish accurate water depths of deposition from ancient shelf deposits, so the degree to which the ancient is nonuniformitarian in terms of bed thickness and cross-shelf transport is still unresolved. Finally, comparison of ancient shelf deposits with modern shelves is hampered by the fact that the latter may be strongly out of equilibrium owing to Holocene sea-level rise.

A second problem, seemingly forgotten in the literature, is that the time between the deposition of successive tempestite beds in many deposits is seemingly many orders of magnitude too large to explain simply as a series of storm events. Data tabulation suggests that preservation recurrence intervals (time between deposition of successive tempestite beds) in a small sample of ancient successions are of the order of 400–15 000 years. High-precision geochronological studies have never been done on tempestite successions, but they are necessary to establish a reasonably accurate picture of the recurrence interval of tempestite beds in the ancient. If one assumes that the recurrence intervals calculated for the ancient are reasonably accurate, then two possibilities arise for their explanation. The simple explanation would be that only extremely large and rare storm events are preserved in the rock record and that, given the recurrence intervals mentioned above, such events would potentially be qualitatively or quantitatively different from any documented in recent deposits. Such a stance might be difficult to justify on the basis of the physics of storms or on basic oceanographical principles, although changes in the Earth's climate, ocean circulation, and palaeogeography could certainly lead to nonuniformitarian atmospheric circulations including extremely large storms.

An alternative explanation is that the apparent recurrence intervals are artefacts of processes that bias the stratigraphical record. Given the range of recurrence intervals, there exists the possibility that relative sea-level changes may have played a role by shifting the source of sandy sediment seaward or landward. In other words, shoreline migration may

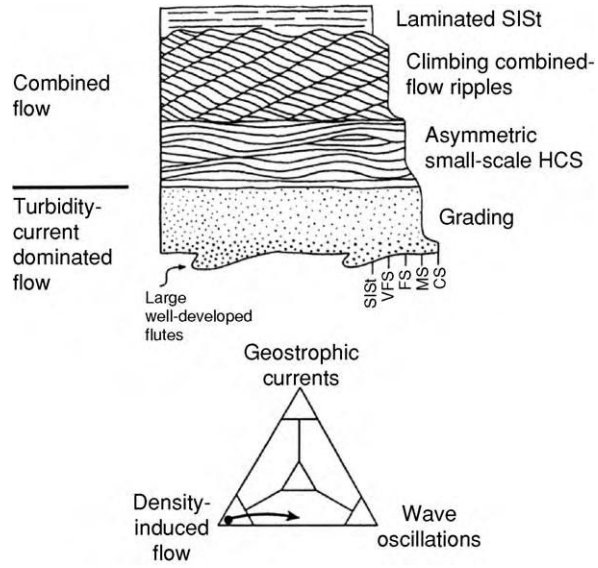


Figure 8 Generalized wave modified turbidite bed. Temporal history of generalized bed is shown in the triangular diagram of storm related effects (see also Myrow PM and Southard JB (1996) Tempestite deposition. *Journal of Sedimentary Research* 66: 875 887). HCS, hummocky cross stratification.

have enhanced apparent cross-shelf transport. Sedimentological data must therefore be combined with sequence-stratigraphical data to understand better the bed geometries and transport distances of ancient tempestites. The stratigraphical distribution of tempestites may also be understood within a sequence-stratigraphical framework, particularly as sea-level has a strong control on shelf profile. For instance, geostrophic flow may be more common during transgressive to early highstand conditions, when accommodation space is higher and there is greater potential for Coriolis deflection, whereas more offshore-directed flow may take place during late highstand conditions characterized by greater bottom friction. Tempestites are probably more abundant during late highstand to early lowstand systems tracts of third- or higher-order sequences, when sand is mobilized from the coast during relative sea-level fall.

Carbonate Systems

Although the previous discussion applies to both carbonate and siliciclastic systems, there are also depositional features that are specific to, or most common in, carbonate systems. The most abundant such features include flat-pebble conglomerate beds and shell-bed accumulations. Flat-pebble beds are an important component of carbonate deposits, particularly in Cambrian and Lower Ordovician strata, and are generally considered to be accumulations of intraclasts produced by the storm reworking of early-cemented carbonate in shoreline and shallow subtidal environments. Detailed sedimentological analyses indicate that some flat-pebble beds were deposited from storm-generated combined flows, although it is likely that there are many other processes that could produce such beds.

Shell-rich grainstone tempestites accumulate as a result of winnowing by storm waves, commonly in shoreline and nearshore settings, and they have characteristics that are different from those of tempestites produced in condensed sections within maximum flooding intervals. In shallow epicratonic seaways, winnowing and concentration of shell debris occurs during lowstand conditions, and further reworking and basinward transport occurs during early transgression. Pronounced trends in the thickness, number, and character of shell beds occurred over the Phanerozoic, and these relate to evolutionary trends and complex taphonomic variables.

See Also

Sedimentary Environments: Depositional Systems and Facies; Carbonate Shorelines and Shelves. **Sedimentary Processes:** Depositional Sedimentary Structures; Particle-Driven Subaqueous Gravity Processes. **Sedimentary Rocks:** Mineralogy and Classification; Sandstones, Diagenesis and Porosity Evolution. **Unidirectional Aqueous Flow**.

Further Reading

Arnott RW and Southard JB (1990) Exploratory flow duct experiments on combined flow bed configurations, and some implications for interpreting storm event stratification. *Journal of Sedimentary Petrology* 60: 211 219.

Dott RH Jr and Bourgeois J (1982) Hummocky stratification: significance of its variable bedding sequences. *Geological Society of America Bulletin* 93: 663 680.

Duke WL (1990) Geostrophic circulation or shallow marine turbidity currents? The dilemma of paleoflow patterns in storm influenced prograding shoreline systems. *Journal of Sedimentary Petrology* 60: 870 883.

Einsele G (1996) Event deposits: the role of sediment supply and relative sea level changes overview. *Sedimentary Geology* 104: 11 37.

Einsele G and Seilacher A (1982) *Cyclic and Event Stratification*. Berlin: Springer Verlag.

Kidwell SM (1991) The stratigraphy of shell concentrations. In: Allison PA and Briggs EG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record*, pp. 211 290. Topics in Geobiology, 9. New York: Plenum Press.

Kreisa RD (1981) Storm generated sedimentary structures in subtidal marine facies with examples from Middle and Upper Ordovician of southwestern Virginia. *Journal of Sedimentary Petrology* 51: 823 848.

Leckie DA and Krystinik L (1989) Is there evidence for geostrophic currents preserved in the sedimentary record of inner to middle shelf deposits? *Journal of Sedimentary Petrology* 59: 862 870.

Myrow PM (1992) Bypass zone tempestite facies model and proximality trends for an ancient muddy shoreline and shelf. *Journal of Sedimentary Petrology* 62: 99 115.

Myrow PM and Southard JB (1991) Combined flow model for vertical stratification sequences in shallow marine storm deposited beds. *Journal of Sedimentary Petrology* 61: 202 210.

Myrow PM and Southard JB (1996) Tempestite deposition. *Journal of Sedimentary Research* 66: 875 887.

Snedden JW, Nummedal D, and Amos AF (1988) Storm and fair weather combined flow on the Central Texas continental shelf. *Journal of Sedimentary Petrology* 58: 580 595.

Swift DJP and Niedoroda AW (1985) Fluid and sediment dynamics on continental shelves. In: Tillman R, Swift DJP, and Walker RG (eds.) *Shelf Sands and Sandstone Reservoirs*, pp. 47 133. Society of Economic Paleontologists Mineralogists Short Course Notes 13. Albuquerque: Society of Economic Paleontologists Mineralogists.

Walker RG (1984) Shelf and shallow marine sands. In: Walker RG (ed.) *Facies Models*, 2nd edn., pp. 141 170. Reprint Series 1. Toronto: Geoscience Canada.

Walker RG (1985) Fluid and sediment dynamics on continental shelves. In: Tillman RW, Swift DJP, and Walker RG (eds.) *Shelf Sands and Sandstone Reservoirs*, pp. 243 295. Society of Economic Paleontologists Mineralogists Short Course Notes 13. Albuquerque: Society of Economic Paleontologists Mineralogists.

SEDIMENTARY PROCESSES

Contents

Erosional Sedimentary Structures

J Collinson, John Collinson Consulting, Beech, UK

Introduction

The accumulation of sediments into the rock record is not a process of continuous deposition, but rather the net result of interaction between episodes of deposition, non-deposition, and erosion. The erosional episodes are commonly recorded by distinctive sedimentary structures whose recognition and interpretation enhance the overall understanding of the sedimentary record. This article deals with some of the processes involved in sediment erosion and also reviews the range of distinctive structures that are produced both on present-day sediment surfaces and in the rock record. It deals solely with features that are present where overall accumulation has occurred or could occur. It does not deal with erosional features that form in subaerial landscapes undergoing long-term erosion.

Erosive Processes

Critical Erosion Velocity

When water or air flows over a bed of loose particles, the boundary shear stress caused by the moving fluid tends to initiate particle movement. There is a critical boundary shear stress (sometimes expressed as a critical erosion velocity) above which particle movement occurs. For sediment of rather uniform grain size, the critical shear stress will increase as the grain size increases. However, for small grain sizes (silt and mud), the situation is less straightforward, as the combined effect of increased cohesive strength and lower surface roughness means that higher velocities are needed to initiate movement. This relationship is illustrated for water in the Hjulström–Sundborg diagram (Figure 1), which operates ideally for well-sorted sediments. A poorly sorted substrate may be winnowed as the critical shear stress of the finer particles is exceeded, leading to the development of an armoured lag on the surface that inhibits further erosion.

Sediment Cohesion

The cohesive strength of damp or wet, fine-grained sediment results from the increasing importance of intergranular electrochemical forces as the surface area to volume ratio of the grains increases and as clay minerals make up an increasing proportion of the sediment. The cohesive strength of fine-grained sediments increases with the time following deposition and is augmented by shallow burial. As well as influencing the critical shear stress of fine-grained sediment, the cohesive strength also complicates the way in which fine-grained sediments are eroded. The particle-by-particle erosion that characterizes non-cohesive sediment is replaced by the erosion of grain aggregates of cohesive sediment. These range from small rip-up clasts to larger blocks of eroded mud (Figure 2). A cohesive substrate is also able to retain detailed erosional relief and to stand at steep, even overhanging, angles. This is critical to the preservation of the many erosional structures discussed below.

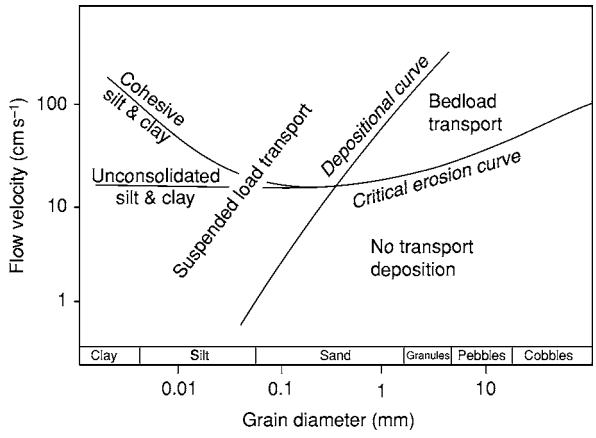


Figure 1 Plot of grain size against stream power showing the velocity at which sediment is set in motion or eroded. Note the different behaviour of fine grained sediment depending on whether it has had time to develop cohesive strength.

Role of Abrasion

The relationships between the critical boundary shear stress and particle size, as expressed in the Hjulström–Sundborg diagram, are complicated further by the fact that natural currents capable of erosion are likely to be carrying material in suspension. Sand and silt particles within this suspended load will enhance the erosive capability of the current by providing an abrasive supplement to the boundary shear stress.

Erosional Sole Marks

Erosional sole marks are relatively small-scale structures preserved as casts on the bases ('soles') of sandstone or, more rarely, limestone beds usually in interbedded sandstone/mudstone sequences. They vary in style and are valuable indicators of erosional and depositional processes. They also provide palaeocurrent information and act as 'way-up' indicators in deformed successions. Their common occurrence on the bases of sandstone beds in interbedded sequences suggests that the sands were deposited as episodic, high-energy events in settings in which quiet conditions normally prevailed, with background sedimentation being from suspension. The fact that erosional sole marks occur on the bases of the sandstones suggests that the high-energy events were erosional in their initial stages, prior to the onset of sand deposition. Such surges, which decelerated through time, typically occur as turbidity currents in deep water, as

Figure 2 Blocks of cohesive mud occurring as clasts in the base of a channel. Recent terrace deposits, Tana Valley, Norway.

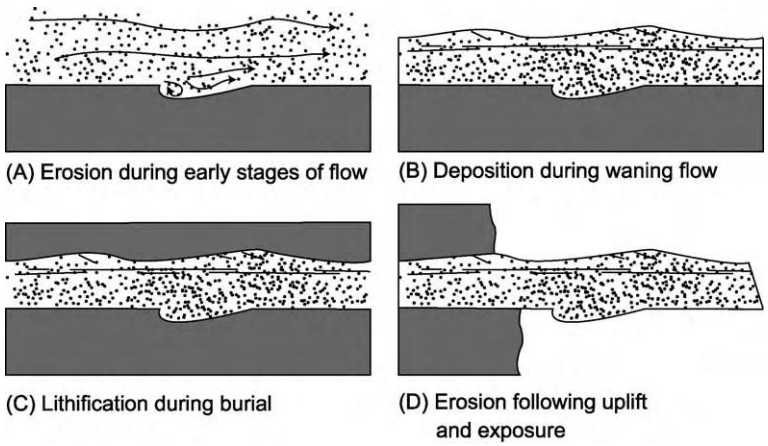


Figure 3 Schematic diagram to show the main stages in the development and preservation of erosional sole marks in an interbedded succession of cohesive sediment and coarser, high energy event beds.

storm surges in shelf settings, and as flood events in deltaic and alluvial settings.

The erosive initial stages of such flows are likely to be highly charged with suspended sediment, which may include coarse particles and, consequently, erosion may result from a combination of both fluid shear and grain abrasion. If the transported clasts are sand and silt, abrasion will augment and accelerate fluid scour. If, on the other hand, the clasts are quite large, their interaction with the bed may create distinctive structures that relate directly to that interaction. Such structures are 'tool marks'. Erosional forms created by both types of process occur initially as negative relief on the bed, and their preservation and later recognition depend on that surface being rapidly overlain by coarser sediment, giving a positive cast of the relief (Figure 3). The survival of the surface relief through the transition from erosion to deposition depends on the cohesive strength of the substrate, which typically preserves steep and detailed features.

Sole Marks due to Fluid Turbulence

These structures relate to the pattern of fluid turbulence close to the bed and are essentially smooth features related to fluid scour, probably augmented by abrasion. They may be classified into three main groups: obstacle scours, flute marks, and longitudinal furrows.

'Obstacle scours' occur around an object lodged on the bed. This is most often a pebble or a shell fragment and the scour is preserved as a horseshoe-shaped ridge that wraps around the upstream side of the obstacle (Figure 4). The ridge was a trough during erosion and its form results from the locally enhanced boundary shear stress as flow was forced to accelerate around the object. Spiral eddies which are shed off either side of the obstacle give the tails that gradually die out downstream.

Figure 4 Obstacle scour around a pebble on the base of a sandstone bed. The horseshoe shaped ridge is the cast of a scour trough eroded around the upstream side of the pebble as the flow accelerated past the obstacle. Eocene, Ainsa, Spanish Pyrenees.

'Flute marks' occur in isolation and in groups, sometimes in linear arrays parallel to flow. They are characterized by a rounded nose at the deepest part of the scour from which the structure flares away in a V-shaped form (Figure 5). The nose points upstream and the shape of the flute reflects the action of a local eddy at the bed, enhanced by the growing flute. Their shape in plan ranges from very sharp V-forms to features with a straighter transverse upstream edge. Flutes may have been initiated at some initial irregularity on the bed surface that caused flow separation and heightened bed shear stress. The irregularity responsible may sometimes be apparent, but in most cases is not. The separation eddy, in turn, led to local scour, enhancing the relief, so that, once initiated, a positive feedback loop was set up between the turbulence and the bed topography. The downstream flaring of the flute away from the nose reflects the mixing of the eddy back into the body of the flow. At the head

Figure 5 Flute casts on the base of a turbidite sandstone bed showing the noses pointing in the upstream direction. Eocene, San Sebastian, Spain.

of the flute, the axis of the separation eddy would have been transverse to flow and the axes would have rotated parallel to flow along the downstream flaring margins.

'Longitudinal furrows' occur on the bases of sandstone beds as patterns of broadly parallel ridges that have a rounded profile and are separated by narrower grooves. This reflects an erosional morphology of rounded parallel furrows separated by cuspate ridges. Along the furrows, it is quite common to find rounded noses that all point in the same direction, similar to small flutes. Indeed, there is a continuous series of erosional forms between clear flutes at one end and smooth parallel furrows at the other. The furrows are thought to result from the erosive effects of paired spiral eddies with their axes parallel to flow and with opposed senses of rotation. The furrows correspond to divergence zones where eddies direct flow down on to the bed, whilst the cuspate ridges lie beneath the convergence zones where flow components lift off from the bed.

'Gutter casts' are features that seem to occur preferentially in shallow marine settings where storm currents are typically the agent responsible. They differ from longitudinal furrows in being larger and separated from one another by areas of flat bedding surface rather than cuspate ridges (Figure 6). They are typically many centimetres wide and a few centimetres deep and they may be separated by flat sectors up to metres in width. Their fills sometimes carry a coarse lag in the base, often made up of fossil fragments. In plan view, they tend to be rather continuous, may be somewhat sinuous, and may split and rejoin, suggesting an anastomosing pattern. They probably result from rather sustained currents, possibly driven by wind or storm effects.

All the scours related to fluid scour have a strong orientation parallel to the current, which can be used to infer palaeoflow direction. With longitudinal scours without clear noses, it may be impossible to infer the sense of movement.

Figure 6 Lower bedding surface of a limestone bed from within a shallow marine interbedded limestone mudstone succession showing gutter casts. Ordovician, north west Greenland.

Tool Marks

These constitute a diverse suite of structures that are produced by the interaction of objects ('tools') carried by the flow with a soft cohesive substrate. The shape and scale of such structures vary with the nature of the tools and the way in which they impinge on the bed (Figure 7). Tools marks can generally be distinguished from structures produced by fluid scour by the fact that they tend to have sharp, often angular relief, reflecting the shape of the tool, compared with the much smoother shapes of fluid scour marks. The tools themselves are seldom seen associated with the tool marks that they created, and one can only imagine or infer what they may have been. They are most likely to have been pebbles, mud rip-up clasts, fossils, or wood fragments. In some cases, delicate relief on the marks can be related to features of the tool, such as ribbing on a shell.

Tools may interact with the bed in a variety of ways, and a loose terminology is applied to the results. Where tools remain in contact with the bed for some distance, they gouge 'grooves', usually with sharp margins and sometimes changing in form as the tool rotates slowly. Where the tools are in intermittent contact with the bed, a string of similar impact marks

Figure 7 A variety of tool marks on the base of a thin turbidite sandstone bed showing prod marks and bounce marks. The series of distinctive bounce marks towards the top of the slab is thought to have been made by a fish vertebra, unlike most tool marks where the nature of the tool is unknown. Oligocene, Polish Carpathians.

Figure 8 Chevron cast on the base of a turbidite sandstone. Ordovician, Kirkudbrightshire, Scotland.

with more or less regular spacing may form 'bounce marks'. These may be symmetrical or may have an asymmetry reflecting the differing angles of impact and rebound. Where tools make isolated impact with the bed, single impressions occur. These 'prod marks' commonly have a clear asymmetry, gently inclined on the upstream side, recording the low-angle approach trajectory, and steeper on the downstream end where the tool was lifted back into the flow.

Less common are 'chevron marks' where the tool appears to have been dragged across a rather viscous muddy substrate so that the sediment has been rucked into a series of chevron-like folds that face in a downstream direction (Figure 8).

Most tool marks are good palaeocurrent indicators and, in most cases, both direction and sense of movement can be deduced. Highly elongate forms, such as grooves and bounce mark sequences, give the best directional measure, whilst the asymmetry of prod marks provides the most reliable indication of sense of movement. It is often the case that a variety of forms occur together on the same surface and it is not uncommon for a range of directions to be present, suggesting that the current responsible changed direction whilst in the erosive mode.

Erosional Surface Forms

On present-day subaerial sediment surfaces, e.g., tidal flats, beaches, river beds, and desert flats, it is possible to find structures that record erosion by recent currents, both water and wind. Examples of these features also occur in the rock record, but they are less common than erosional sole marks because many are developed in sand, which is likely to be reworked by subsequent currents. Erosional forms are also detected and mapped on the seafloor by remote sensing techniques on both sandy and muddy substrates. Ancient small examples are seen in the rock record at outcrop, but some larger examples probably go undetected.

'Obstacle scours' occur around pebbles and shells on sandy river beds and on beaches. They are curved troughs that wrap around the upstream side of the obstacle and flare away downstream. They relate to the local acceleration of the current around the obstacle and the generation of a local eddy system. On stream beds, the sand may be rippled and the obstacle scour may, in effect, be a distortion of the ripple pattern. On beaches, the current responsible is the backwash of the waves and often they occur on an otherwise flat sediment surface.

'Wind ridges' occur where strong winds blow across damp sand as on a beach. As the sand dries out differentially, dry sand is set in motion by the wind, whilst damper patches retain some coherence and resist erosion. These will typically become streamlined as erosion goes on around them, producing ridges with blunt noses at their upwind ends and tails that flare away downwind. Rare examples are known in some ancient lake deposits where the surface was inundated and draped with finer sediment.

'Seafloor flutes and lineations' have been detected by remote sensing techniques, particularly side-scan sonar. They occur on both sandy and muddy seafloors and can result from both tidal currents and deep-sea turbidity currents. Lineations eroded in muddy seafloors may be many hundreds of metres in length and form spaced arrays. Large-scale flutes, which are

known from the surfaces of some submarine fans, are tens to hundreds of metres in length and seem particularly to be features of areas where large flows have spilled over from fan channels.

The recognition of ancient examples of such features requires exceptional exposure. The largest forms would commonly be categorized as channels. With extensive bedding surface exposure, large, flute-like forms (megaflutes) occur on the upper surfaces of sand units in turbidite sequences and are typically overlain by mud drapes and fills. It may well be the case that, where erosion was followed by the deposition of sand, the features usually go undetected.

Channel Forms

Channels occur in a wide range of settings as an integral part of the depositional process. In some cases, channels and their fills constitute only a small part of the succession; in other settings, the successions may be dominated by channel fills. Given good outcrop, the recognition of channels is often very straightforward, as an erosional surface, with significant relief, separating two discrete bodies of sediment, the overlying unit commonly being coarser grained than that below (Figure 9). The shapes of channel erosion surfaces are quite varied, ranging from strongly curved, concave-upwards forms to horizontal surfaces. The observed shape records the interaction of the active channel and its behaviour during its active life. Channels that maintain a stable position generate a cross-section that is very close to that of the active channel and the resultant sediment body is lenticular. Actively migrating channels generate an erosion surface with an extensive flat sector and steeper margins, so that the channel sediment body has a tabular shape. Channel stability reflects the energy of the flow within the channel and the stability of the material in the channel banks. With cohesive bank material, channels are less likely to migrate. For non-cohesive bank material, the greater tendency for migration means that channel margins are less likely to be seen in the rock record, and erosion surfaces have to be inferred from the concentration of larger clasts, both exotic and intraformational, in a lag conglomerate directly above the erosion surface. Where channel margins are cut into dominantly cohesive sediment, but with some interbedding of lithologies, it is quite common for the interbedding to be reflected in the morphology of the erosion surface. Stepped profiles and even overhangs may occur. These smaller features are often valuable for the measurement of channel trend directions, particularly where accompanied by sole marks, such as flutes.

Figure 9 Part of a channel margin within a turbidite succession. The stepped nature of the junction in part reflects the lithological differences in the underlying thin bedded succession. The thick beds of sand that make up the channel fill reflect the stronger currents that flowed through it. Namurian, North Derbyshire, England.

Other Erosion Surfaces

Many bedding surfaces have an element of erosion in their genesis, but hardly qualify as 'sedimentary structures' and will not be dealt with in detail. Marine transgressions may often be erosive in nature, and the landward advance of a shoreface may produce a horizontal and relatively flat surface (ravinement surface) recognized by a sharp upwards change from more nearshore to more offshore facies, possibly associated with a lag conglomerate. In aeolian settings, major bounding surfaces (supersurfaces) may be associated with erosion down to the water table and are recognized in some cases only on the basis of their wide extent.

Channel-like features may result from processes other than fluid scour, especially in finer grained sediments and particularly those deposited on subaqueous slopes. Instability, often due to rapid deposition, may give rise to failure and the initiation of slumps and slides. The surfaces over which such movements take place may have concave-upwards forms, especially in the up-slope slump scar area, and these can superficially resemble channels. Generally, the context and the associated sediments allow the differences to be recognized.

Other erosional forms result from the dissolution of lithified sediments, typically limestones, so that irregular palaeokarstic surfaces may be incorporated into the rock record.

See Also

Sedimentary Processes: Depositional Sedimentary Structures; Post-Depositional Sedimentary Structures; Aeolian Processes; Karst and Palaeokarst; Particle-Driven Subaqueous Gravity Processes. **Sedimentary**

Rocks: Mineralogy and Classification. **Unidirectional Aqueous Flow**.

Further Reading

Allen JRL (1982) *Sedimentary Structures: Their Character and Physical Basis. Developments in Sedimentology 30A & B*, pp. 593 663. Amsterdam: Elsevier.
Allen JRL (1985) *Principles of Physical Sedimentology*, ch. 8. London: Allen & Unwin.
Collinson JD and Thompson DB (1989) *Sedimentary Structures*, 2nd edn., ch. 3 4. London: Chapman & Hall.
Dzulynski S and Walton EK (1965) *Sedimentary Features of Flysch and Greywackes*. Amsterdam: Elsevier.
Leeder MR (1982) *Sedimentology; Process and Product*, ch. 5, 6, 9. London: Allen & Unwin.
Leeder MR (1999) *Sedimentology and Sedimentary Basins*, ch. 5, 6, 10. Oxford: Blackwell Science.
Mutti E (1992) *Turbidite Sandstones*. Milan: Agip.
Pettijohn FJ and Potter PE (1964) *Atlas and Glossary of Primary Sedimentary Structures*. Berlin: Springer.
Ricci Lucchi F (1970) *Sedimentografia*. Bologna: Zanichelli.
Selley RC (2000) *Applied Sedimentology*, 2nd edn., ch. 5. San Diego: Academic Press.

Depositional Sedimentary Structures

J Collinson, John Collinson Consulting, Beech, UK

Introduction

Depositional sedimentary structures reflect closely the processes by which sediment was transported immediately prior to deposition, and therefore provide an important starting point for environmental interpretation through facies analysis (*see* **Sedimentary Environments:** Depositional Systems and Facies). Their interpretation requires an appreciation of the hydrodynamics or aerodynamics of fluid–sediment interactions.

Sediment Transport and Deposition

Excluding highly concentrated sediment–water mixtures, sediment is transported by fluid in two distinct ways. Finer grained material, commonly clay and silt, but sometimes including sand and coarser material, is carried along with the fluid in 'suspension', when particles are supported by the upwards component of turbulence. Coarser grained material tends to move in intermittent contact with the bed, once its critical boundary shear stress has been exceeded. This movement, referred to generally as 'bedload transport', takes place through rolling or, more commonly, bouncing of grains on the bed, a process called 'saltation'. At its most vigorous, saltation grades into suspension. Sand particles in saltation, on colliding with a bed of similar sand, set other grains in motion as their kinetic energy is dissipated, a process especially important in wind-blown transport.

Bedload Transport

Bedload transport in water results from unidirectional currents, wave action, and combinations of the two. Clearly, for waves to affect sediment movement, the bed has to be above wave base (i.e., within the depth range of the waves), and it is important in many settings to recognize the distinction between fair weather and storm wave base.

Suspension

Sediment is carried in suspension provided that the intensity of turbulence within the fluid remains high. A current that is underloaded with suspended sediment can take more material into suspension up to a certain critical capacity. Sediment is deposited when the level of turbulence can no longer support all the suspended grains. This usually occurs when the current decelerates, often as a result of flow expansion at a channel mouth, but also through the waning of a high-energy event, such as a storm or flood. Coarser particles, with higher settling velocities, fall from suspension first. Sedimentation is commonly accelerated in river mouth and estuarine settings where a freshwater suspension meets saline water and particles coagulate into larger flocs. Deposition from suspension occurs in many settings and produces a variety of types of bedding and lamination. 'Bedding' is used for units of centimetres or larger thickness, whilst 'lamination' is used for features at the scale of millimetres.

At coastal delta fronts, a floating plume of turbid freshwater floats over denser seawater and may extend many kilometres offshore before it disperses. In freshwater lakes, the reverse situation may apply, with denser turbid water hugging the lake floor as a density underflow. Such relatively sustained underflows contrast with more episodic underflow surges, which are features of deep marine settings. These turbidity currents carry a wide variety of grain sizes in suspension and the interaction of density contrast, down-slope movement, resultant turbulence, and

consequent sediment load leads to a phenomenon referred to as 'autosuspension'. Changes to the balance amongst the variables, mainly through changes in bottom gradient, lead to accelerations or decelerations and to phases of erosion and deposition, respectively.

Lamination in Fine-Grained Sediments

In settings in which low energy prevails, only the finest sediment, carried in suspension, settles as mud. If the bottom waters are well oxygenated, sediment will usually be disturbed by burrowing organisms and the depositional lamination destroyed. If the bottom waters are anaerobic, a fine depositional lamination may be preserved. In all cases, burial compaction flattens the original fabric and creates a foliation or lamination of later origin. Only where early cementation occurs, usually as concretions, is depositional and early postdepositional fabric and lamination preserved.

In areas of higher sediment supply, such as a coastal river mouth, lamination and bedding in sediments deposited from suspension often result from changes in river discharge. In marine settings, deposition from a floating turbid plume can give gradational thin bedding or lamination due to gradual changes in discharge or in the position of the plume. In lakes, where underflows are more common, bedding and lamination may be more sharply defined, particularly in proglacial settings where seasonal discharge variations are shown by varves (rhythmites; *see* **Sedimentary Environments:** Lake Processes and Deposits).

In certain tidal settings, particularly estuaries, where fine-grained suspended sediment occurs due to both river supply and erosion by waves and tidal currents, the deposition of mud is particularly associated with slack water periods between tidal flows. These occur four times a day in subtidal settings and twice a day in intertidal areas. High concentrations of suspended mud and the effects of flocculation cause the deposition of discrete 'mud drapes' within the space of a few hours. When associated with sand deposition, reflecting higher energy tidal flows, these help to create 'heterolithic facies', which are discussed below.

Aqueous Sandy Bedforms and their Internal Structures

Once sand or coarse silt begins to move under a unidirectional water current, the sediment surface takes on a succession of dynamic morphologies in response to changing current strength. These morphologies are termed 'bedforms', and their hydrodynamic significance and relationships to internal lamination in sands and sandstones are key elements in interpreting processes and environments of deposition.

Current Ripples and Ripple Lamination

When sand starts to move, the surface becomes covered by 'current ripples', small-scale, repetitive, asymmetrical bedforms, a few centimetres high and a few tens of centimetres in wavelength (Figure 1). Their dimensions scale with grain size rather than with the properties of the flow. Their asymmetrical profile parallel to flow has a steep 'lee side' facing downstream and a gently inclined 'stoss side' facing upstream. Soon after formation, current ripples have rather straight crest lines, orientated transverse to flow, but this pattern soon breaks down into more complex three-dimensional 'linguoid' geometries, whereby the ripples have strongly curved crest lines, usually convex downstream and with scour pits in front of gaps between adjacent ripples. These shapes are intimately related to the patterns of fluid turbulence over the ripples, with flow separation at the crests and downward-directed eddies converging to create lee-side scour pits. As an array of ripples moves downstream, sand is swept from the scour pits over downstream stoss sides and deposited on lee sides. Ripples and dunes (dealt with below) are together referred to as the 'lower flow regime' of bedload transport (Figure 2).

The deposition of sand on the lee sides of ripples leads to inclined lamination that records successive positions of the lee side. This cross-lamination is commonly defined by micas and other platy grains. It is preserved only when ripple migration takes place during vertical aggradation of the bed (Figure 3).

Figure 1 Small scale asymmetrical current ripples formed by a unidirectional flow from left to right. Gently inclined stoss sides are inclined upstream and steeper lee sides face downstream. Note the variability in the curvature of the crest lines. Tana River, Norway.

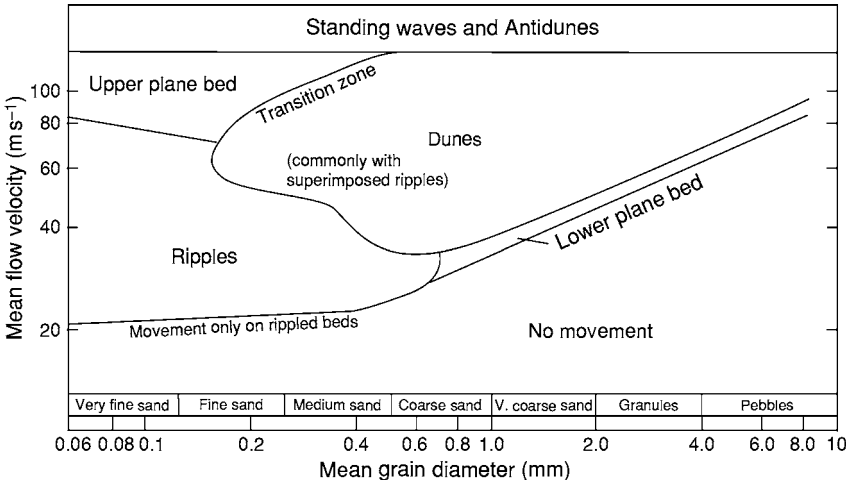


Figure 2 A plot of stream power vs. grain size showing the fields of occurrence of the main classes of bedform produced by unidirectional water flows.

Figure 3 Climbing ripple cross lamination in fine sand. The trajectories of the ripple crests reflect the balance between the rate of bed accretion and rate of ripple migration. Where the angle of climb exceeds that of the ripple stoss sides, stoss side lamination is preserved. Holocene terrace deposits, Tana Valley, Norway.

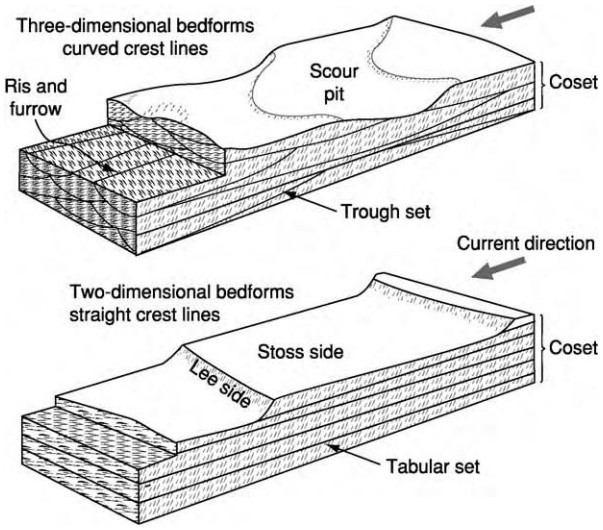


Figure 4 Block diagram to show the main types of three dimensional geometries of cross bedding and cross lamination and the terminologies used.

Where migration occurs without aggradation, lee-side scour removes the cross-laminated sand deposited by the preceding ripple. When aggradation occurs, the trajectory of the scour zone climbs downstream, allowing the partial preservation of cross-laminated sand. The angle of climb determines the extent of preservation, but, in many cases, the angle is not apparent because of the random component of erosion created by migrating scour pits. Where climb angles are high, due to high aggradation rates, the inclination of erosion surfaces between cross-laminated sets may be apparent, and the style is referred to as 'ripple drift' or 'climbing ripple lamination' (Figure 3). Where climb angles exceed the inclination of the stoss sides, stoss-side laminae are preserved with no erosion between sets.

In three dimensions, the geometry of ripple cross-lamination most commonly occurs as a series of intersecting troughs, each filled with curved cross-laminae that are concave both upwards and downstream (Figure 4). The troughs are the curved erosion surfaces cut by the migration of the scour pits and the laminae record their filling by the migration of the succeeding ripples. Usually, only the lower parts of lee-side laminae are preserved. In plan view, trough cross-lamination appears as a pattern of flow-parallel zones of curved laminae, known as 'rib and furrow', which is an important palaeocurrent indicator. Palaeocurrents are most accurately measured from the axes of the troughs. Rib and furrow is a more

common record of current ripples than the preservation of ripple forms on bedding surfaces. Where ripple forms are preserved, they are usually overlain by a drape of finer grained sediment due to a rapid reduction in current strength.

Dunes, Sand Waves, and Cross-Bedding

As current velocities increase above those appropriate for ripples, sand beds deform into larger bedforms, termed 'dunes' or 'sand waves', provided that the water is deep enough (Figure 2). These have a similar asymmetry to current ripples, although the wavelengths are greater, seldom less than 1 m and extending to many tens of metres. Heights are in tens of centimetres or even metres. There is a clear hydrodynamic distinction between ripples and dunes, illustrated by the superimposition of ripples on the larger forms under equilibrium conditions. Dunes, therefore, are not just large ripples. The orientations of superimposed ripples reflect the direction of flow close to the bed over and around the dunes, recording patterns of eddy separation and reattachment, especially those associated with scour pits.

Dunes and sand waves occur in a wide range of plan forms, from straight crested through highly three dimensional, and also span a wide range of height to length ratios (Figures 5 and 6). Steep, strongly three-dimensional forms are mainly a response to strong and relatively deep flows, whilst straight-crested or low-relief forms, which tend to lack scour pits, are responses to less strong or more shallow flows. The dimensions of dunes and sand waves reflect the properties of the flow, rather than the sediment. They seem particularly susceptible to changes in flow depth. However, it is quite difficult to interpret large bedforms in terms of flow parameters as they are commonly out of equilibrium with the prevailing current. Large volumes of sand must be moved to remould large forms and, in rivers where discharge varies quickly or in tidal settings, large-scale bedforms are often in a state of disequilibrium. In intertidal settings or on emergent river beds, larger bedforms are commonly modified by the process of emergence. Changing flow patterns over and around the forms and the action of waves both rework ripple patterns and reduce the angle of the lee-side slipfaces. In subtidal settings, sand waves often are present on the seabed where they are susceptible to semi-diurnal reversals of flow. Asymmetrical profiles may reflect quite a small asymmetry in strengths of ebb and flood tides.

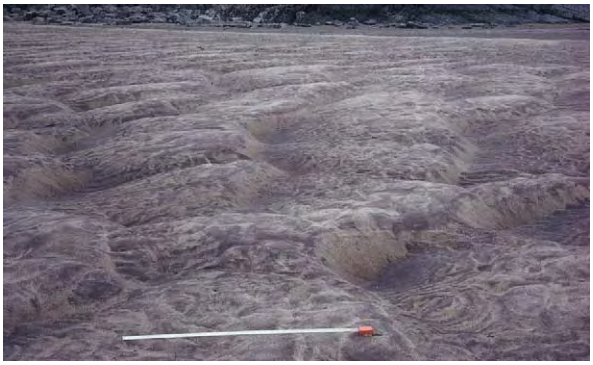


Figure 5 Small, three dimensional dunes with well developed scour pits in their lee side areas. Note how the superimposed ripples fan out to reflect the pattern of eddying over the dunes close to the bed. Note also the difference in scale of the dunes and the ripples. Tana River Delta, Norway.

Figure 6 A wide sandy braided bed of the Tana River, Norway, showing large, low relief sand wave bedforms. These examples have curved crest lines (convex downstream), superimposed ripples, and, locally, smaller dunes on their upper surfaces. Whilst their spacing is in hundreds of metres, the heights of their lee faces seldom exceed 1 m. The larger areas of sand, which separate the river into discrete channels, are compound bedforms or bars. Tana River Delta, Norway.

As with ripples, internal lamination produced by dune-scale bedforms is dominated by inclined lamination, generally referred to as 'cross-bedding' (Figure 7). The individual dipping laminae are termed 'foresets', whilst packages of foresets produced by the migration of individual dunes are 'sets'. Sets range in thickness from around 10 cm up to several metres, and foresets are generally defined by slight differences in grain size, related to avalanches of grains (grain flows) on the lee side and to fluctuating turbulence in that area. In vertical section, some foresets are straight with an angular basal contact (termed 'planar'), whilst others have a tangential or asymptotic contact. The sets fall into one of two classes depending on the shape of the set boundaries. Where these are broadly subparallel and flat-lying, the sets are described as 'tabular'. Where set boundaries are concave-upwards erosion surfaces, they compare, except in scale, with ripple cross-lamination, and the term 'trough cross-bedding' is used. The foresets of trough sets are predominantly concave upwards and tangential, whilst the foresets of

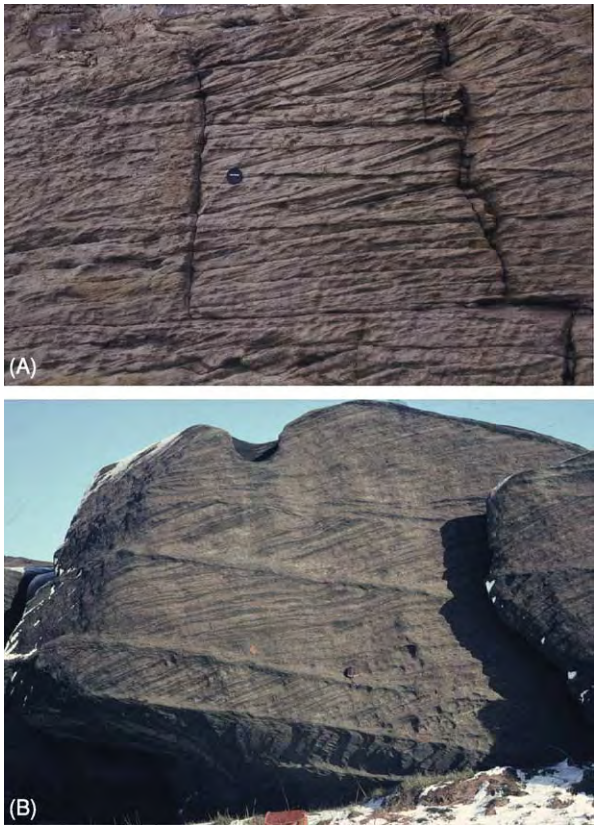


Figure 7 Examples of cross bedding in sandstones. (A) Coset of small sets with slightly asymptotic foresets. The bounding surfaces between sets are themselves inclined downstream, sug gesting deposition on the downstream side of a larger compound bar form. Westphalian, Northumberland, England. (B) Tabular sets with thin units of ripple lamination between sets. The foresets are essentially planar. Namurian, Staffordshire, England.

tabular sets vary in shape. This reflects the association between trough sets and lee-side scour pits and the fact that steep, three-dimensional dunes, which commonly have such pits, are associated with stronger currents. Tabular sets record the migration of dunes or sand waves without scour pits, probably with straight or only gently curved crest lines.

Changes in the shape of foresets along a set record changes in flow strength during the life of the bedform. They are a particular feature of tidal sediments and the deposits of rivers with large discharge variations. Wave action on emergence during the falling stage may create low-angle erosion surfaces, truncating the foresets. Succeeding foresets re-establish the higher angle dips above these 'reactivation surfaces' (Figure 8). Flow reversals in subtidal settings produce similar structures and tidal slack water can cause mud drapes to be interbedded with the foresets. In exceptional cases, the spacing of drapes can be interpreted in terms of spring-neap tidal cycles ('tidal bundles').

Figure 8 Reactivation surface within a tabular cross bedded set produced by the migration of a low relief sand wave. The discontinuity reflects modification of the lee side of the bedform during emergence during falling water discharge. Tana River, Norway.

Compound Bedforms and Bars

In many wide, sandy rivers, the large repetitive bedforms described above are organized into compound forms, which commonly appear to scale with the channel itself. Some of these features are attached to channel banks, whilst others are located in mid-channel. Some examples have a clear asymmetry to their stream-parallel profile with a slipface at the downstream end. In other cases, dunes descend a more gently inclined lee side. These compound features are commonly referred to as 'bars', of which there are many types. The important point, in terms of internal structure and its interpretation, is the relative movement directions of dunes and of the accretion of the bar. Where the bar advances downstream, in the same direction as the dunes, either a large, cross-bedded set will be present beneath smaller cross-beds, or the bounding surfaces between sets will be inclined downstream. Where bars move laterally as a result of dunes migrating along their flanks, bounding surfaces dip in a direction highly divergent from that of the cross-bedding (lateral accretion surfaces).

Upper Flow Regime Bedforms and Lamination

In situations in which water flow is very rapid and/or very shallow, saltating sand close to the bed moves continuously over an essentially flat surface. This phase of movement is known as 'plane bed transport' with the moving grain layer termed a 'traction carpet'. When the flow is particularly strong, the bed develops gentle undulations, forming ridges transverse to flow or three-dimensional domes. When this happens, the water surface also takes on a wave form which parallels that on the sediment surface. These undulations are 'standing waves' (Figure 9). Where the standing waves on the water

surface grow beyond a critical height, they move rapidly upstream and break, a case described as 'antidunes'. Breaking antidunes temporarily wipe out the undulations and the cycle of growth and breaking is repeated. This whole family of transport mechanisms is referred to as 'upper flow regime' transport.

Standing waves and antidunes are highly ephemeral features and preservation of their deposits is rare. However, where plane bed conditions coincide with bed aggradation, distinctive parallel and horizontal lamination forms. This may be accentuated by mica and gives rise to flaggy sandstones that split easily parallel to bedding. The bedding surfaces themselves show a distinctive lineation ('parting lineation') (Figure 10) parallel to the current, which is a useful palaeocurrent indicator. Rarely, parting lineation occurs on the bedding surfaces of undulatory lamination, indicating the chance preservation of standing waves.

Figure 9 Standing waves on the surface of a shallow, rapidly flowing stream. In phase undulations on the underlying, rapidly moving sand bed may produce undulating lamination with low preservation potential.

Figure 10 Upper bedding surface of parallel laminated sand stone showing a clear parting lineation or primary current linea tion. This lies parallel to the direction of the shallow, rapid flow that produced it. Middle Jurassic, Yorkshire Coast, England.

Wave Ripples and their Lamination

The to-and-fro movement of water associated with waves is able to move sediment, both on its own and in conjunction with unidirectional currents. Sand moved by waves is readily moulded into ripples, which are distinguished in several ways from current ripples. Wave ripples, most distinctively, have straight crests, which may be rounded or quite cuspate in profile (Figure 11). Their profile is also commonly symmetrical, although not always so. Where the strength of the wave surge is stronger in one direction than in the other, or where waves coexist with currents, asymmetry results. Wave ripples vary quite widely in spacing, depending on the grain size, the strength of the waves, and the water depth. Less common are ripples produced by coexisting wave trains with differing propagation directions. These show interference patterns, reflecting the wave sets responsible.

As well as producing ripples in their own right, waves also modify pre-existing bedforms, particularly under emergent or near-emergent conditions. Wave-modified current ripples are particularly common on tidal flats.

Internally, wave ripples produce small-scale cross-lamination which is not always easy to distinguish from that of current ripples. Interfingering bundles of laminae with opposed dips, draping laminae over symmetrical ripple crests, and discordances between internal lamination and ripple forms all suggest wave influence, but many cases are difficult to fully diagnose (Figure 12).

As well as producing ripples, waves also create plane bed conditions. This occurs most readily in the swash zone of beaches where the run-up and backwash of waves produce the shallow, rapid flows

Figure 11 Small, straight crested wave ripples with symmet rical, rather rounded crest lines. The straightness of the crest lines and the tuning fork junctions along the crests are most characteristic of wave influence. Tana River Delta, Norway.

Figure 12 A wave ripple bedding surface with clear, symmetrical ripple profiles, underlain by a unit of ripple laminated sand showing bundled laminae and trough forms in a view normal to the crest lines. Namurian, County Clare, Ireland.

Figure 14 Heterolithic interlaminated sand and mud typical of many tidal settings where energy levels fluctuate between strong currents capable of transporting sand and still conditions where mud falls from suspension to drape the bed morphology. Lower Cretaceous, Isle of Wight, England.

Figure 13 Sandstone bed with undulating lamination with convex upwards sectors and undulatory erosion surfaces. This assemblage of hummocky and swaley cross stratification is typical of deposition from storms under conditions of coexisting, intense current and wave activity. Upper Jurassic, Dorset, England.

needed. Parallel lamination and parting lineations result if the sand is preserved.

Wave–Current Interaction

During storms, shallow-water shelf or nearshore areas are commonly subjected to energy regimes involving a combination of waves and currents. Under such conditions, sand falling from suspension may be reworked into low-relief, rounded forms and may be re-eroded by higher energy surges. The result is an undulatory lamination, characterized by gentle convex-upwards patterns and gentle scours. Where the convex-upwards laminae dominate, the structure is called 'hummocky cross-stratification', and where the scours dominate, it is termed 'swaley cross-stratification' (Figure 13).

Heterolithic Lamination

In settings in which energy levels fluctuate and in which there is a significant amount of sediment in suspension, it is common for mixed sand–mud sediments to be deposited. These record the alternation of suspended and bedload sediment deposition and are a particular, although not exclusive, feature of tidal settings. The sandy components are often ripple laminated with ripple profiles commonly preserved, so that current and wave forms can be distinguished. Collectively, these interlaminated sediments are termed 'heterolithic' (Figure 14) and are subdivided using the proportions of sandy and muddy components. Where mud dominates and sand forms isolated ripple lenses, the term 'lenticular bedding' is applied, whilst ripple-laminated sand with discrete mud drapes is called 'flaser bedding'. More or less equal components are referred to as 'wavy bedding'. Some confusion surrounds the application of these terms. Laminae rich in mica and carbonaceous debris in normal ripple cross-laminated sand can look superficially like flaser bedding, but the term is best reserved for cases with a clear separation of sand and mud.

Aeolian Bedforms and Internal Bedding

Wind, blowing over dry sand, initiates movement in much the same way as water. Above a certain critical velocity or shear stress, grains begin to move as bedload. However, in comparison with water, wind moves a much narrower range of grain sizes in this way. The buoyant effect of water is much greater than that of air and, in addition, the viscosity of water dampens collision impacts between grains in saltation. The result is that sands moved by air tend to be very well sorted and also well rounded. The first bedforms to develop on dry sand are ripples. These have low relief compared with aqueous current

ripples and are less clearly asymmetrical. They are commonly straight crested with the bifurcation of crest lines.

The vigorous saltation leads to grains following asymmetrical trajectories: a steep uplift from the bed, and a lower angle approach and impact. The impact throws other grains into motion, but the extent of this depends on the inclination of the bed at the point of impact. If the bed slopes upwind, the impact is steep and energy is dissipated strongly into grains near the impact point, throwing some into saltation and advancing others by a process of creep. If the bed slopes downwind, energy is dissipated more gently and grains accumulate. Wind ripples reflect this behaviour and their wavelength scales to the saltation path length and, hence, to the wind strength. Where wind blows dry sand across a surface that is immobile, such as a rock surface, a gravel bed, or a damp surface, wind ripples may be widely spaced and may begin to cluster together into larger forms that are incipient dunes. Where the surface is damp, saltating grains may adhere on impact through surface tension and patterns of wind 'adhesion ripples' or 'adhesion warts' may form. These irregular warty forms have steeper sides pointing upwind. All wind ripples have quite a low preservation potential, but examples do occur in the rock record. Wind ripples commonly produce subparallel lamination, 'pinstripe' lamination with inverse grading, and rare cross-lamination. It is an important component in the cross-bedding produced by aeolian dunes.

Vigorous wind transport commonly leads to the development of larger bedforms, 'aeolian dunes' (*see* **Sedimentary Processes:** Aeolian Processes), which occur in isolation on hard substrates or as parts of larger accumulations of sand. Dunes vary greatly in both scale and shape, and it is quite common for dunes of several scales to occur superimposed upon one another. An important distinction is between dunes with their own slipfaces and those without such surfaces. Slipfaces occur downwind of crest lines where sand accumulates through grain flow avalanches. Such processes commonly occur in conjunction with grainfall, whereby grains are thrown over the dune crest and accumulate on the lee side. Where grainfall dominates, the inclination of the surfaces may be lower and reworking by wind ripples may occur, particularly in the lower, more gently inclined parts of the dunes ('plinths').

Where sand supply is insufficient for the substrate to be fully covered, 'barchan dunes' or, more rarely, 'linear (seif) dunes' occur. In sand seas, a variety of more complex forms occurs. These include transverse dunes, barchanoid forms, and star-shaped dunes. The largest forms, termed 'draa', commonly have superimposed smaller dunes, which leads to complex morphologies. The size of the largest forms means that they change shape only over long periods of time and thus reflect a wind regime rather than a particular wind episode. Some are out of phase with the current regime as a result of lag effects.

Internally, dune sands show complex cross-bedding at scales up to many metres thick. The inclined laminae are commonly well defined and it may be possible to differentiate grainfall, grain-flow and wind-ripple laminae. Wind-ripple laminae generally occur in more gently inclined intervals, whilst grain flow is commonly close to the angle of rest (*ca*. 30°). Aeolian dune cross-bedding is characterized by discordances or bounding surfaces at several scales. Individual cross-bedded sets are separated from one another by erosion surfaces, and internally they may also be punctuated by low-angle erosion surfaces similar to the reactivation surfaces of aqueous cross-beds. These record the complexity of the wind regime whereby the lee sides of dunes become sites of erosion during particular wind episodes. They may also result from the migration of dunes over slipfaceless draas or the oblique migration of scour pits along the flanks of dunes.

Structureless Sand and Sandstone

Not all sand and sandstone has visible internal lamination. This happens for several reasons. First, it may have been deposited in that way. The dumping of sediment from suspension from, for example, a decelerating current, may have been so rapid that there was no time for the sand to be reworked into bedforms and laminae. On the other hand, sand that was initially laminated may have lost its lamination through later remobilization, perhaps due to liquefaction. Finally, lamination may be lost through the activities of burrowing organisms, which have the ability to totally homogenize laminated sediments (*see* **Trace Fossils**).

Decelerating Flows and the Bouma Sequence

Many sandstone beds in interbedded sandstone–mudstone successions show an internal vertical sequence of lamination types that are diagnostic of decelerating flows. The sandstones are sharp based and may have erosional sole marks (*see* **Sedimentary Processes:** Erosional Sedimentary Structures). Above their bases, they show one or more intervals of five different lamination types that occur in a constant relative order even though they are rarely all present in the same bed. These five types are: A, structureless

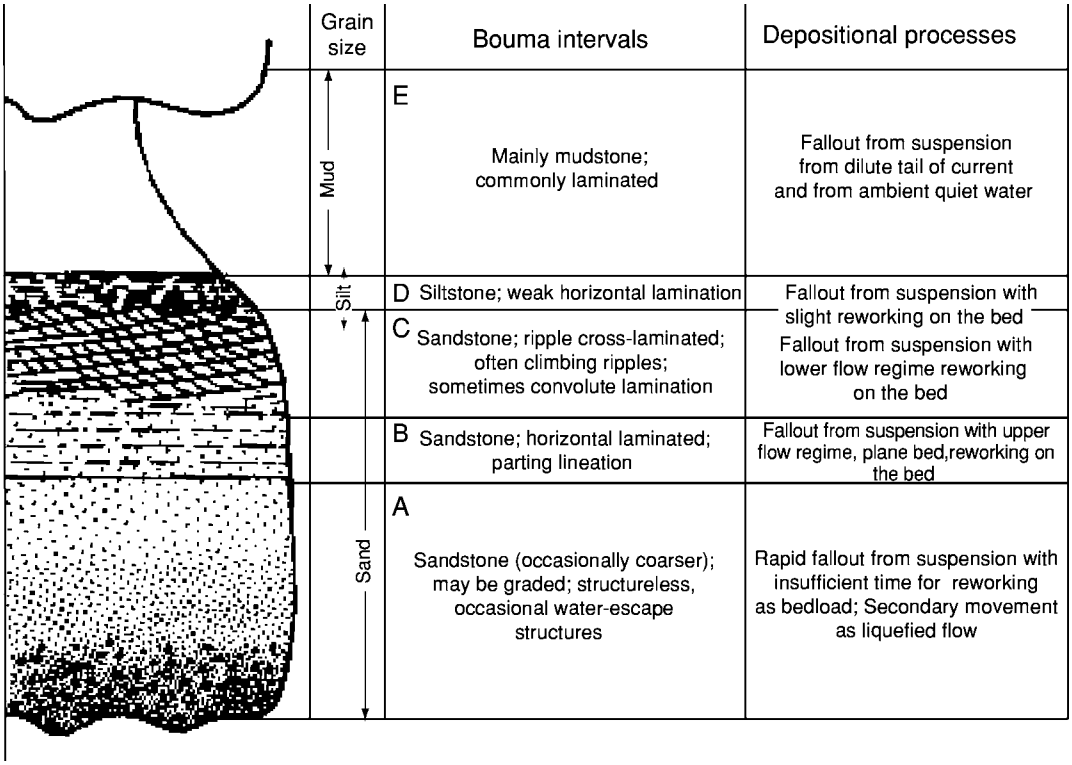


Figure 15 A schematic profile through a turbidite sandstone bed showing a complete sequence of Bouma intervals and their hydrodynamic interpretation.

(commonly graded) sandstone; B, parallel laminated sandstone; C, ripple cross-laminated sandstone; D, horizontally laminated (commonly silty) sandstone; E, mudstone. The complete succession of lamination types is known as the 'Bouma sequence' (Figure 15). The A interval records rapid deposition from suspension, with graded bedding reflecting sorting within the flow of a mixed grain size population. The B interval records upper flow regime plane bed transport, whilst the C interval records ripples moving in the lower flow regime. Climbing ripple lamination is quite common, recording rapid rates of bed aggradation, and convolute lamination may also result from rapid deposition. The D interval, often poorly displayed in weathered exposures, records fallout from suspension with weak reworking on the bed, whilst the E interval grades into the background sedimentation from suspension.

The overall succession, however complete or incomplete, records an episodic current, possibly erosional in its early stages, and progressively decelerating thereafter. Most commonly, the Bouma sequence is associated with deep-water settings where the episodic flows are turbidity currents (*see* **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes), but similar decelerating flows occur in shallow-water settings of floodplains, shallow lagoons, and lakes due to crevassing of river banks during floods. Similar sands also occur in shallow marine settings as a result of storms, although the Bouma sequence may then be modified to include hummocky cross-stratification and wave ripples.

Figure 16 Coarse fluvial gravels showing clear pebble imbrication. The flat surfaces of the clasts dip upstream, indicating that flow, in this case, was from left to right. Holocene, Pyrenees, Spain.

Gravel Forms and Fabric

Whilst the preceding section deals mainly with sands and finer grained sediments, powerful water currents also transport gravel, and its deposition can lead to distinctive structures and fabrics (*see* **Sedimentary**

Rocks: Rudaceous Rocks). Where gravel moves as a traction load, on a stream bed or on a beach, it may become lodged on the bed, particularly where clasts land amongst similar sized clasts. If the clasts are flattened, they will tend to be deposited in an imbricated fabric with an upstream dip. This is the most stable position beneath the prevailing flow and, when seen in the rock record, 'imbrication' (**Figure 16**) is a valuable palaeocurrent indicator.

Gravel is also deposited in bars on river beds and these commonly have slipfaces at their downstream ends. Cross-bedding results from the avalanching that takes place there. Depending on the overall sediment population, sandy and gravel-rich foresets may be intermixed.

See Also

Sedimentary Environments: Depositional Systems and Facies; Lake Processes and Deposits. **Sedimentary Processes:** Erosional Sedimentary Structures; Post-Depositional Sedimentary Structures; Aeolian Processes; Particle-Driven Subaqueous Gravity Processes. **Sedimentary Rocks:** Rudaceous Rocks. **Trace Fossils**. **Unidirectional Aqueous Flow**.

Further Reading

Allen JRL (1982) *Sedimentary Structures: Their Character and Physical Basis. Developments in Sedimentology 30A & B*, pp. 593, 663. Amsterdam: Elsevier.

Allen JRL (1985) *Principles of Physical Sedimentology.* ch. 4. London: Allen & Unwin.

Brookfield ME and Alhbrandt TS (eds.) (1983) *Eolian Sedi ments and Processes. Developments in Sedimentology 38*, p. 660. Amsterdam: Elsevier.

Collinson JD and Thompson DB (1989) *Sedimentary Struc tures*, 2nd edn., ch. 5 7. London: Chapman & Hall.

Leeder MR (1982) *Sedimentology; Process and Product*, ch. 6 8. London: Allen & Unwin.

Leeder MR (1999) *Sedimentology and Sedimentary Basins*, ch. 4 12. Oxford: Blackwell Science.

Mutti E (1992) *Turbidite Sandstones*. Milan: Agip.

Pettijohn FJ and Potter PE (1964) *Atlas and Glossary of Primary Sedimentary Structures*. Berlin: Springer.

Pye K and Lancaster N (eds.) (1993) *Aeolian Sediments Ancient and Modern. Special Publication of the Inter national Association of Sedimentologists 16*, p. 167. Oxford: Blackwell Scientific Publications.

Reineck HE and Singh IB (1980) *Depositional Sedimentary Environments*, 2nd edn. Berlin: Springer.

Selley RC (2000) *Applied Sedimentology*, 2nd edn., Ch. 5. San Diego: Academic press.

Post-Depositional Sedimentary Structures

J Collinson, John Collinson Consulting, Beech, UK

Introduction

Primary sedimentary structures and lamination in clastic sediments result from deposition under different hydrodynamic regimes. However, there are other secondary structures, which originate during or very soon after deposition as a result of sediment instability, extreme climatic effects, or chemical changes during burial. The last group involves the precipitation of mineral cements within the sediment. Some of the early physical structures deform primary depositional structures and lamination, whilst others originate entirely as a result of the deformation processes. Those associated with cementation often help to preserve original lamination and other delicate features, particularly in fine-grained sediment, that would otherwise have been destroyed by compaction.

Soft-Sediment Deformation Processes

Early formed soft-sediment deformation structures are typically features of sandstones and finer sediment. Deformation in gravels and conglomerates is rare, except in glacial and periglacial settings, or where there is abundant muddy matrix. The structures have an entirely physical origin and must be distinguished from organic deformation structures, such as burrows and trails of animals and disturbance by plant roots (*see* **Trace Fossils**). In addition, tectonic structures are often geometrically identical with syndepositional structures, and the boundaries between widespread tectonic deformation and more local, sediment-induced deformation may be blurred.

Deformation structures are quite difficult to classify as they result from rather diverse processes. However, all attempts at interpreting physical soft-sediment deformation structures must consider two discrete questions. First, why did the sediment lose its strength and become susceptible to deformation? Second, what forces operated on the weakened sediment to cause the deformation?

In most cases, the sediment loses frictional shear strength and behaves, briefly, as a plastic or a fluid. In other cases, a loss of tensile strength can lead to brittle deformation as soft-sediment faults. Loss of frictional shear strength may result in total remobilization of the sediment and transformation into a gravitational

mass movement, leading to a phase of resedimentation. However, this article deals only with situations in which the loss of strength was short-lived and the sediment regained its strength before it moved downslope, thus preserving structures that record more or less *in situ* deformation.

Deformation occurs within sediment when intergranular forces are unable to resist applied stresses, which are usually, but not exclusively, gravitational. The shear strength of sediment is normally expressed by the equation:

$$\tau = C + (\sigma - p)\tan\varphi$$

where τ is the shear strength, C is the grain cohesion, σ is the pressure normal to shear, p is the excess pore fluid pressure, and φ is the angle of internal friction. This means that sediment will fail when the applied stress exceeds τ; such a condition will be favoured by a lower cohesive strength C, by changes in grain packing to reduce $\tan\varphi$, or by an increase in pore fluid pressure p beyond a certain critical value. Such conditions can occur for several reasons, discussed below. Strain, when it occurs, may be isotropic, dispersed throughout the body of sediment, or concentrated on discrete slip surfaces.

Loss of Strength

In the shear strength equation set out above, the presence of fine-grained, especially muddy sediment tends to increase the role of the cohesive strength, which is largely a function of grain size, whilst in better sorted, mud-free sediment, frictional forces dominate. In both cohesive and non-cohesive sediments, high pore water pressure commonly leads to a loss of shear strength, but this condition can have different causes and durations in different sediments.

In fine-grained, muddy sediments, excess pore fluid pressure is often caused by relatively rapid deposition. Low permeability inhibits the escape of pore water and hence retards sediment compaction, giving overpressured conditions. In addition, the decay of organic matter within the sediment can generate gases, which, if unable to escape, also contribute to pore pressure. Such overpressured conditions occur close to the sediment surface where highly mobile mud may be susceptible to deformation or flowage. They also develop at a larger scale, and in a longer time frame, within thick piles of sediment, such as those deposited by prograding deltas or deltaic margins (e.g., Mississippi, Niger). In such cases, increasing pore water pressure with burial may eventually exceed lithostatic pressure, leading to failure. This results in plastic flowage of a thick, overpressured layer. These movements set up extensional stresses in the overburden, which can lead to the development of discrete slip surfaces that cut up-section as listric faults when the tensile strength of the sediments is exceeded. Such surfaces may also become the bounding surfaces of sediment gravity slides. These conditions, and the resultant movements, develop and persist over long intervals of time; in the case of the largest deltaic margins, over millions of years.

Rapidly deposited sands, lacking significant fine-grained sediment, commonly have rather loose grain packing. They are particularly susceptible to shock, such as by an earthquake, heavy wave action, further rapid deposition, or a sudden rise in water level. Shock breaks grain contacts and induces tighter packing and, as a result, excess pore water is present, raising the pore water pressure and causing temporary liquefaction, which will last until the extra pore water escapes. In the case of sands, the high permeability usually means that liquefaction is short-lived. Whilst liquefied, the sediment–water mixture deforms readily as a Newtonian fluid in response to gravitational and other applied stresses. As pore fluid escapes, usually upwards to the sediment surface, grain contacts are re-established and a rising front of reconsolidation 'freezes' the deformed sediment. Deformation is usually identified through the distortion of depositional lamination but, with extreme or sustained liquefaction and deformation, depositional lamination is totally obliterated and structureless sand or slurried textures are produced. The upwards movement of escaping pore water may also create deformation, both by distorting depositional laminae and by creating new vertical structures, such as dishes, pipes, and sheets, within which sediment–water mixtures move as fluidized flows.

Deforming Forces

Virtually all the forces that act upon sediment weakened by the above processes are gravitational. These can be divided into two major classes: a down-slope component of gravity and the action of gravity on inverted density gradients. In addition, shear at the sediment surface due to flowing water or ice, or to sediment gravity flows, may also deform the sediment.

The down-slope gravitational component may be sufficient to overcome the cohesive strength of muddy sediment on a slope. This can lead to the detachment of a slab of sediment on a basal shear surface and to internal deformation within the moving layer. The style of deformation varies depending upon the position within the moving sheet. Stretching, particularly at the up-slope end, may lead to tensile failure and to the development of brittle extensional structures, whilst compression, commonly in a down-slope setting, causes plastic folds. These styles of deformation are associated with a family of mass

movements dealt with in detail elsewhere (*see* **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes). Also partly driven by down-slope gravitational components are sedimentary growth faults. These are related to failures in deeply buried overpressured muds. Extension on the faults is also driven by lateral flowage of these muds due to vertical loading by the overlying sediment pile.

Density inversions occur as a result of both depositional processes and the early loss of pore water. A sand layer, deposited on mud, will commonly have a higher density than the mud and, if the mud is weak, the instability may lead to loading of sand into the mud. Where the sand is of uneven thickness, as with ripple lenses or erosional sole marks, loading may be concentrated where the sand is thickest. Extreme cases of loading show detached balls of sand, usually encased in slurried mud or silt: so-called ball and pillow structures.

Within liquefied sand layers, which commonly result from shock, patterns of density inversion may be more subtle. Internal depositional lamination may be contorted in convolute lamination, some of which reflects the upward escape of water or low-density sediment–water mixtures. Very rapidly deposited sands may lack depositional lamination and experience liquefaction during the depositional process. The dewatering of such sands and the elutriation of fines by escaping water can produce subtle features, such as dish structures and water-escape pillars and pipes.

Deeper burial and associated differential compaction lead to the deformation of larger scale depositional geometries. The cross-sections of channel sand bodies, in particular, are susceptible to change from planar-convex forms to concavo-convex patterns where encased in muds. Where channel sands are encased in mud, as in some deep-water depositional settings, dewatering of the muds during burial can lead to high pore water pressures in the sands. This can eventually lead to rupture of the mud seal and to the injection of liquefied sands into the muds as sills and dykes. When such processes occur near the sediment surface, liquefied sand may be extruded as slurry flows or as sand volcanoes.

Where sands are deposited in large bedforms, such as dunes and sand waves, they may be liquefied through rapid deposition on the lee slope. If this happens whilst the current is still active, the cross-bedded sands, developed by lee-side deposition, may be sheared and the foresets dragged into folds, giving overturned cross-bedding.

The movement of ice over unconsolidated sediment also generates large-scale and intense deformation. This occurs near the front of a glacier, where glacial advance operates as a bulldozer to fold and thrust proglacial deposits. In addition, the basal layers of a mobile glacier may be rich in sediment. Shearing across the basal layers can lead to intense deformation, although the rheology mainly involves plastic deformation of ice rather than sediment liquefaction or cohesion.

Soft-Sediment Deformation Structures

This section deals with the descriptive aspects of soft-sediment deformation and relates the structures to the deforming processes. The structures themselves do not fall naturally into well-defined classes, but Figure 1 shows a scheme which relates all soft-sediment structures to processes. This diagram illustrates the possible interactions between different causes of a loss of strength and deforming forces, and shows that processes are gradational with tectonic deformation and with resedimentation. It also shows the wide range of scales across which deformation occurs, from microloading to growth faults at the scale of continental slopes, and highlights the differences in timing and duration of deformation. Structures formed during deposition, such as overturned cross-bedding, contrast with structures that require significant later burial, such as growth faults. Commonly occurring deformation structures are described below, in order of increasing scale, but moderated by considerations of process.

Load Casts and Pseudonodules

These usually occur at the interface between a mud layer and an overlying sand layer. Downward-facing bulbous load casts of sand are separated by upward-pointing flame structures of mud (Figure 2). Some occur on otherwise flat bedding surfaces, whilst others accentuate erosional or depositional irregularities. In extreme cases, sand may be detached from its source bed and occur as isolated load balls, or pseudonodules, within highly disturbed muds or silts, a features sometimes referred to as a 'ball and pillow' structure. Internal lamination within both sand and mud is highly contorted. Loading is caused by gravity acting on unstable density stratification where the sand layer is denser than the muds and where both layers are weakened by excess pore fluid.

Convolute Lamination

This commonly occurs within sands or silts and involves the distortion of depositional lamination by folding of variable intensity (Figure 3). In some cases, the folding is chaotic, whilst in others upward movement of escaping water has dragged lamination into upright folds. Cuspate folds with sharp anticlines and

<table>
<tr><th colspan="2" rowspan="2">Loss of strength
Nature of deforming force</th><th colspan="3">Exceed strength of sediment</th><th colspan="2">Liquidize</th></tr>
<tr><th>Internal tensile (Brittle)</th><th>Internal cohesive (Plastic)</th><th>External surface cohesive (Plastic)</th><th>Liquefied</th><th>Fluidized</th></tr>
<tr><td colspan="2">Gravitational body force on slope</td><td>Slides</td><td>Slumps</td><td>Slumps and slides</td><td colspan="2">Debris flows</td></tr>
<tr><td colspan="2" rowspan="2">Unequal confining load</td><td rowspan="2">Growth faults</td><td colspan="3" rowspan="2">Loaded ripples shale ridges and diapirs</td><td colspan="2">Loaded ripples and sole marks</td></tr>
<tr><td></td><td>Clastic dykes sand volcanoes</td></tr>
<tr><td rowspan="4">Gravitationally unstable density gradient (density inversion)</td><td>Continuous</td><td rowspan="6">Soft sediment faults</td><td colspan="3" rowspan="3"></td><td>Convolute lamination</td><td></td></tr>
<tr><td>Within a single layer</td><td>Dish structures</td><td>Water-escape pipes and pillars</td></tr>
<tr><td>Multiple layer, not pierced</td><td colspan="2">Bedding surface load casts</td></tr>
<tr><td>Multiple layer, pierced</td><td colspan="3">Shale ridges and mud diapirs</td><td colspan="2">Ball and pillow/pseudonodules
Isolated load balls</td></tr>
<tr><td rowspan="2">Applied shear stress</td><td>Current drag</td><td colspan="3" rowspan="2"></td><td>Overturned cross-bedding</td><td></td></tr>
<tr><td>Vertical</td><td></td><td>Water-escape pipes and pillars</td></tr>
</table>

Figure 1 The relationship of various post depositional structures to the nature of the loss of strength and the type of deforming force. Slumps and slides, which involve lateral resedimentation of material, are dealt with in detail in (*see* **Sedimentary Processes:** Particle Driven Subaqueous Gravity Processes). Modified from Collinson JD (1994) Sedimentary deformational structures. In: Maltman A (ed.) *The Geological Deformation of Sediments*, pp. 95 125. London: Chapman & Hall; and Collinson JD (2003) Deformation structures and growth faults. In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 193 195. Dordrecht, Kluwer: Academic Publishers.

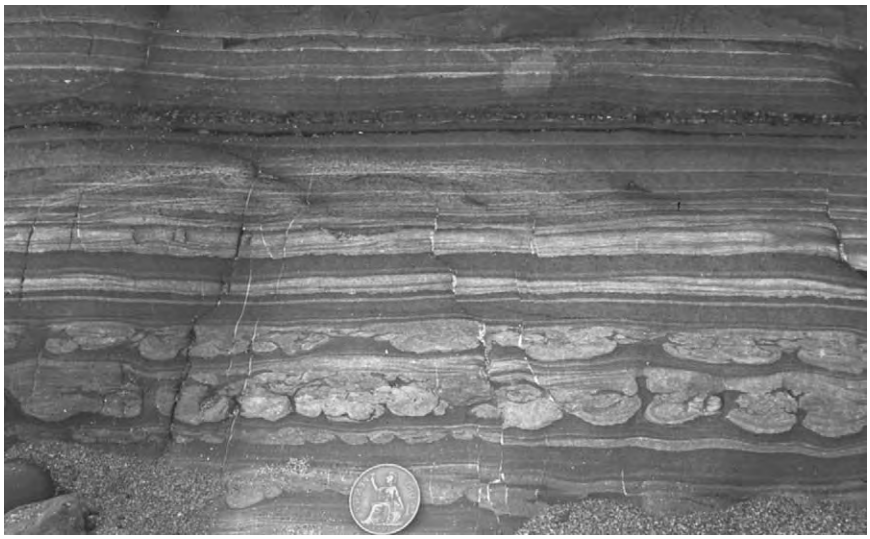


Figure 2 Small scale load casts within thin bedded turbidites. The downward sinking sand lobes are separated by upward pointing flames of mud. Carboniferous, North Cornwall, England.

more rounded synclines are typical. In many cases, depositional lamination can still be identified, commonly cross-bedding or cross-lamination. The good sorting and high porosities of aeolian sands make them particularly susceptible to liquefaction when the water table rises rapidly through them, giving convolute lamination on a large scale. Convolute lamination is common in turbidite sandstones, often

Figure 3 Convolute bedding developed in coarse cross bedded fluvial sandstones as a result of the sediment experiencing a short lived period of liquefaction shortly after deposition. Namurian, Yorkshire, England.

associated with the ripple laminated (Bouma C) interval (*see* **Sedimentary Processes:** Depositional Sedimentary Structures; Deep Water Processes and Deposits).

All examples reflect liquefaction of the sand, which can happen for a variety of reasons, including seismic shock, rapid sedimentation, breaking waves, or a shift in water table. Where seismic shock caused liquefaction, deformed layers may be very widespread and have stratigraphical significance. Upward escape of excess pore water, internal density inversions, and down-slope components of gravity may all contribute to deformation. Dewatering leads to reconsolidation of the sediment from the bottom up, 'freezing' the deforming laminae. Protracted liquefaction may lead to total homogenization.

Overturned Cross-Bedding

This is a special case of convolute lamination in which liquefaction occurred in an actively migrating bedform (Figure 4). Current shear on the sediment–water interface dragged the liquefied sand in a down-current direction in an essentially laminar style. An upward-migrating front of reconsolidation allowed higher parts of the sediment to be sheared longer, giving down-current-facing recumbent folds.

Figure 4 Overturned cross bedding in two small sets within a shallow marine sandstone. This was caused by ongoing bed shear by a current during a short lived loss of strength. The deformed sand was progressively 'frozen' as a front of reconsoli dation moved upwards through the sediment. Late Precambrian, north east Greenland.

Dish and Pillar Structures

These are most common in thick, otherwise massive sands and result from dewatering following rapid

Figure 5 A thick bed of sandstone in a turbidite setting showing intense development of dish structures due to rapid escape of pore water shortly after rapid deposition. Eocene, Fuenterrabia, Spain.

deposition. Dish and pillar structures are direct products of dewatering and are not distortions of pre-existing lamination. Dish structures are thin concave-upward, subhorizontal zones of slight clay enrichment produced by local filtering out of the elutriated fines (Figure 5). Their upturned edges merge with pillar structures, which are vertical conduits of water escape and record fluidized or near-fluidized conditions.

Sand Injection Structures

Where a sand body is encased in mud and progressively buried, compaction and dewatering of the muds can lead to overpressure in the sands. If the fluid pressure exceeds the tensile strength of the overlying mud, rupturing occurs and complex networks of dykes and sills of liquefied sand are intruded into the muds. This appears to be quite common in subsurface channel sand bodies in deep-water and slope settings, although outcrop examples are rare.

Small-scale injection dykes occur in mudstones of interbedded sandstone and mudstone successions. These may be folded ptygmatically because of differential compaction. Sand within dykes is typically structureless, although some examples show foliation parallel with dyke margins. Remobilized sands are increasingly being recognized as important petroleum reservoirs in the North Sea and elsewhere in the world.

Sand Volcanoes and Extruded Sheets

Sometimes upward-intruding liquefied sand, both dykes and pipes, penetrates to the free sediment surface where the sand is extruded. Where the escape conduits are pipes, sand volcanoes form, at scales from a few centimetres to a few metres (Figure 6). If the intrusions are dykes, fissure extrusion occurs and liquefied sand may flow as sheets before dewatering and 'freezing'. Where sand dykes penetrate muddy debris flows, extruded sand may be interfolded with the upper levels of the debris flow.

Mud Diapirs

Overpressure develops where muds are buried quite rapidly, as in a prograding delta. The resultant loss of strength and the significant overburden lead to both vertical and horizontal flowage of the muds. Where the overlying sediment is denser than the mud, as in an upward-coarsening deltaic succession, mud may rise vertically as a diapir or mudlump, pushing aside or penetrating mouth bar sands. In the Mississippi delta, mudlumps rise to sea-level and create short-lived islands offshore from distributary mouths. Mud diapirism may lead to great variability in the thickness of mouth bar sands, which thicken into withdrawal synclines between diapirs (Figure 7).

Horizontal flowage of overpressured muds leads to the extrusion of muds at the base of the prograding slope as imbricate thrusts and folds. The same motion is partly responsible for extensional stresses that drive extensional sedimentary growth faults in deltaic successions. Mud diapirs are widespread in Tertiary deltas around the world and may help to trap petroleum.

Slumps and Slides

Gravitational body forces acting on sediments lying on a slope lead to slumps and slides when cohesive (plastic) and tensile (brittle) strengths are exceeded. In both types, shearing is concentrated on discrete basal surfaces. They differ principally in their internal deformation. Slides are largely undeformed, but may have internal faults, whilst slumps show plastic folding. The two styles may coexist in the same deformed unit. The fold orientation may indicate palaeoslope direction, but analysis must be carried out with caution. Gravitational mass movements are dealt with more fully in (*see* **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes).

Figure 6 Sand volcanoes on top of a slump sheet. Dewatering of the slump led to rapid upwards injection of a sand slurry which was extruded at a point vent to give the volcano. Namurian, County Clare, Ireland.

Figure 7 Mud diapir penetrating delta front mouth bar sands. Overpressuring of the buried muds led to a loss of strength which, combined with a density inversion, led to the upwards flow of mud. Namurian, County Clare, Ireland.

Sedimentary Growth Faults

These occur in deltaic successions. They are large-scale features and require exceptional exposures to be identified at outcrop (Figure 8). In small exposures, they may be confused with later tectonic faults. They are caused by shear failure at discrete dislocations within the sedimentary pile, and they commonly occur as listric surfaces that sole out on a basal '*décollement*'. In small examples, within a single progradational unit, the basal surface may be within particularly fine-grained mudstone directly above a basal flooding surface. In larger examples, penetrating several progradational units, the *décollement* may be within a thick interval of overpressured muds, whose lateral flowage towards the free surface slope combines with gravitational forces on the hanging wall block to drive that block down-slope. Growth faults are characterized by slow and continual movement

Figure 8 Small, synsedimentary growth fault within deltaic sediment. Note the thickening of sands into the hanging wall of the fault and the roll over within these sands. The synsedimentary nature of the fault is shown by the fact that the uppermost beds pass over the fault without disturbance. Namurian, County Clare, Ireland.

and are aided by progressive sediment loading. Their hanging wall areas act as local depocentres in which thickened delta-front sediments are deposited and preserved. Continued progradation of individual deltas or of the continental slope causes the active faulting to shift progressively basinwards. Large growth faults are important traps for petroleum in deltas such as the Mississippi and the Niger.

Climatically Induced Structures

In addition to the soft-sediment structures described above, in which a temporary loss of strength causes the disturbance, there are some rather diverse structures which result from climatic effects.

Desiccation and Other Cracks

When surface layers of mud or silt dry out, they contract and create isotropic horizontal tension. If this exceeds the tensile strength of the mud, polygonal (commonly hexagonal) cracks develop which may be filled with wind-blown or water-lain sand (Figure 9). The dimensions of the polygons are a function of the mud layer thickness, and several scales may coexist. Subaqueous shrinkage (synaeresis) of surface mud layers also leads to cracks, although these often form less regular patterns. They are thought to result from volume changes of clays within the surface layer. Synaeresis cracks may commonly be differentiated from subaerial desiccation cracks because they are infilled with mud, not sand. Large-scale polygonal cracks also develop through protracted freeze–thaw processes in periglacial settings.

Figure 9 Desiccation cracks on an upper bedding surface in ephemeral lacustrine sediments. Carboniferous, New Brunswick, Canada.

Evaporite Pseudomorphs

Evaporation of bodies of saline water leads to saturated solutions and to the precipitation of minerals on the floor of the lake or lagoon. If saturated conditions persist, large, euhedral crystals may grow at or just below the muddy sediment surface. When a subsequent flood or storm introduces both fresher water and coarse sediment, the crystals may be dissolved and their forms buried and cast by the coarser material. The bases of sandstone interbeds in playa lake sequences are often decorated by cubic pseudomorphs of halite, commonly with hopper faces (Figure 10), whilst in marine or lagoonal settings, gypsum pseudomorphs may occur.

Figure 10 Halite pseudomorphs on the base of a fine grained sandstone bed. Triassic, Cheshire, England.

Raindrop Impressions

These occur as small circular or elliptical depressions on the upper surfaces of beds of fine sediment. They record the impact of raindrops or hailstones and their preservation depends on the cohesive strength of the sediment. They are quite often associated with desiccation cracks.

Periglacial Deformation

This results mainly from the expansion of water on freezing and the contraction of frozen sediment during extreme freezing. This takes place on a seasonal basis in an active layer close to the surface and creates metre-scale patterns of polygonal cracks. Open cracks are filled with wind-blown sediment and material collapsing from the crack sides during periods of thaw. Expansion on freezing leads to the pushing up of ramparts on the edges of the cracks. In vertical section, frost polygons appear as sharply tapering wedges with complex, poly-phase fills. Convolute

Figure 11 Large, discoid, calcite cemented concretions developed in mudstones. Lower Jurassic, Yorkshire, England.

Figure 12 Septarian nodule developed in mudstone with calcite filled cracks which demonstrate shrinkage of the concretion. Kimmeridge Clay, Dorset, England. Courtesy of R Selley.

bedding also occurs in periglacial sediment due to freeze–thaw action termed cryoturbation.

Concretions

During the burial of sediment, pore water may precipitate minerals in the pore spaces. In some cases, this is as pervasive cement, but, in other cases, it occurs in discrete patches around nuclei (Figure 11). In some cases, these cemented concretions or nodules occur apparently at random, whilst others follow bedding or are nucleated on obvious earlier inhomogeneities, such as fossils, burrows, or root traces. Some concretions, which form at shallow burial depths in fine-grained sediment, predate major compaction and, by resisting compaction, help to preserve lamination and delicate fossils that would otherwise be flattened beyond recognition. Other early concretions (e.g., calcrete nodules) are an integral part of soil textures and are discussed elsewhere (*see* **Soils:** Modern; Palaeosols). Most concretions occur during deeper burial and many have distinctive internal features. 'Septarian nodules' have internal cracks filled with calcite, reflecting shrinkage during their development (Figure 12). Other concretions have complex nested conical fractures, termed 'cone-in-cone structure'.

See Also

Sedimentary Processes: Depositional Sedimentary Structures; Deep Water Processes and Deposits; Particle-Driven Subaqueous Gravity Processes. **Soils:** Modern; Palaeosols. **Trace Fossils**.

Further Reading

Allen JRL (1982) *Sedimentary Structures: Their Character and Physical Basis. Developments in Sedimentology 30A & B*, pp. 593, 663. Amsterdam: Elsevier.

Allen JRL (1985) *Principles of Physical Sedimentology*, ch. 10. London: Allen & Unwin.

Collinson JD (2003) Deformation structures and growth faults. In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 193 195. Dordrecht, Kluwer: Academic Publishers.

Collinson JD (1994) Sedimentary deformational structures. In: Maltman A (ed.) *The Geological Deformation of Sediments*, pp. 95 125. London: Chapman & Hall.

Collinson JD and Thompson DB (1989) *Sedimentary Structures*, 2nd edn., ch. 9. London: Chapman & Hall.

Jones ME and Preston RMF (eds.) (1987) *Deformation of Sediments and Sedimentary Rocks. Special Publication of the Geological Society, London, 29*. London: Geological Society.

Leeder MR (1982) *Sedimentology; Process and Product*, ch. 11. London: Allen & Unwin.

Leeder MR (1999) *Sedimentology and Sedimentary Basins*, ch. 12. Oxford: Blackwell Science.

Mutti E (1992) *Turbidite Sandstones*. Milan: Agip.

Pettijohn FJ and Potter PE (1964) *Atlas and Glossary of Primary Sedimentary Structures*. Berlin: Springer.

Reineck HE and Singh IB (1980) *Depositional Sedimentary Environments*, 2nd edn. Berlin: Springer.

Selley RC (2000) *Applied Sedimentology*, 2nd edn., ch. 5. San Diego: Academic Press.

Aeolian Processes

N Lancaster, Desert Research Institute, Reno, NV, and United States Geological Survey, Reston, VA, USA

Introduction

Aeolian processes, involving erosion, transportation, and deposition of sediment by the wind, occur in a variety of environments, including beaches, semi-arid and arid regions (e.g., cold and hot deserts), agricultural fields, and some terrestrial planets, notably Mars and possibly Venus. Common features of these environments are a sparse or non-existent vegetation cover, a supply of fine sediment (clay, silt, and sand size), and strong winds. Aeolian processes are responsible for the emission and/or mobilization of dust and formation of areas of sand dunes. They largely depend on other geological agents (e.g., rivers, waves) to supply sediment for transport. Despite assertions by many early workers, erosion of bedrock by the wind is not a significant process in most regions.

Sediment Movement by the Wind

Most naturally occurring airflow is turbulent. Mixing of turbulent eddies, each with a different velocity and direction, together with the effect of friction at the surface, results in a wind profile with a logarithmic increase in velocity with height above the aerodynamic roughness length (z_o) (Figure 1). The aerodynamic roughness length is determined by the particle size and roughness of the surface. For sand surface, z_o is typically 1/30th of the mean particle size. For surfaces with scattered isolated roughness elements such as plants or rocks, aerodynamic roughness scales with the roughness density (the volume of roughness elements per unit area). The existence of roughness elements plays an important role in the interaction of the wind with the surface because the available wind shear stress is partitioned between the roughness elements and the intervening surface. A high density of roughness elements will therefore reduce the wind shear stress on the surface so that sediment mobilization and transport may be reduced or eliminated.

Particle movement is achieved via a combination of direct wind shear stress on the surface and atmospheric turbulence. There are three modes of sediment transport by wind: creep or reptation, saltation, and suspension (Figure 2). The mode of transport depends primarily on the ratio between particle settling velocity, and hence particle size, and wind shear stress and turbulence intensity. Very small particles (<20 μm) are transported in suspension (tens of kilometres or greater) and are kept aloft by turbulent eddies in the wind. True suspension occurs when the particle settling velocity is very small compared to the turbulence intensity of the wind. Larger particles (20–63 μm) undergo short-term suspension for distances of tens to hundreds of metres; material of sand size (63–1000 μm) is transported mainly in a series of short hops (saltation), in which the vertical component of wind velocity (turbulence) has no effect on particle trajectories. Material coarser than 500 μm diameter is transported in contact with the surface by reptation and creep. The modes of transport are interdependent: saltating sand particles eject silt- and

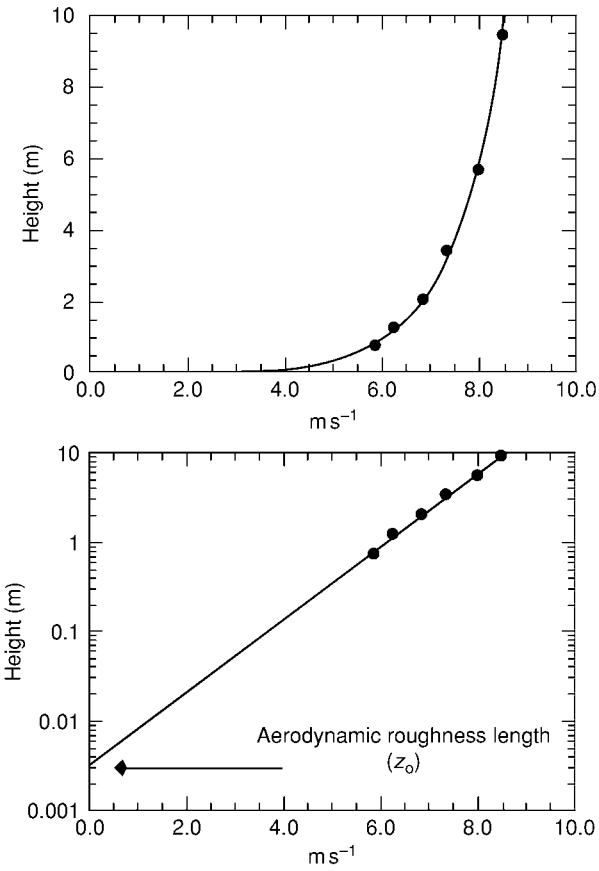


Figure 1 Development of wind profiles. The wind profile equation is $u_* \quad u/\kappa \ln(z/z_o)$, where u_* is wind shear velocity, u_* is wind speed at height z, and κ is the von Karman constant (~0.4). Field data from Lancaster N, Rasmussen KR, and Greeley R (1991) Interactions between unvegetated desert surfaces and the atmospheric boundary layer: a preliminary assessment. *Acta Mechanica, Suppl.* 2: 89 102.

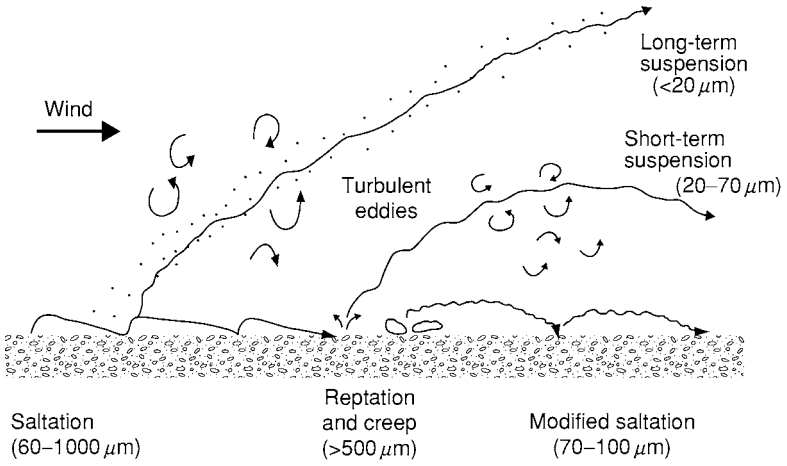


Figure 2 Modes of sediment transport by the wind. Reproduced with permission from Pye K (1987) *Aeolian Dust and Dust Deposits*. London: Academic Press.

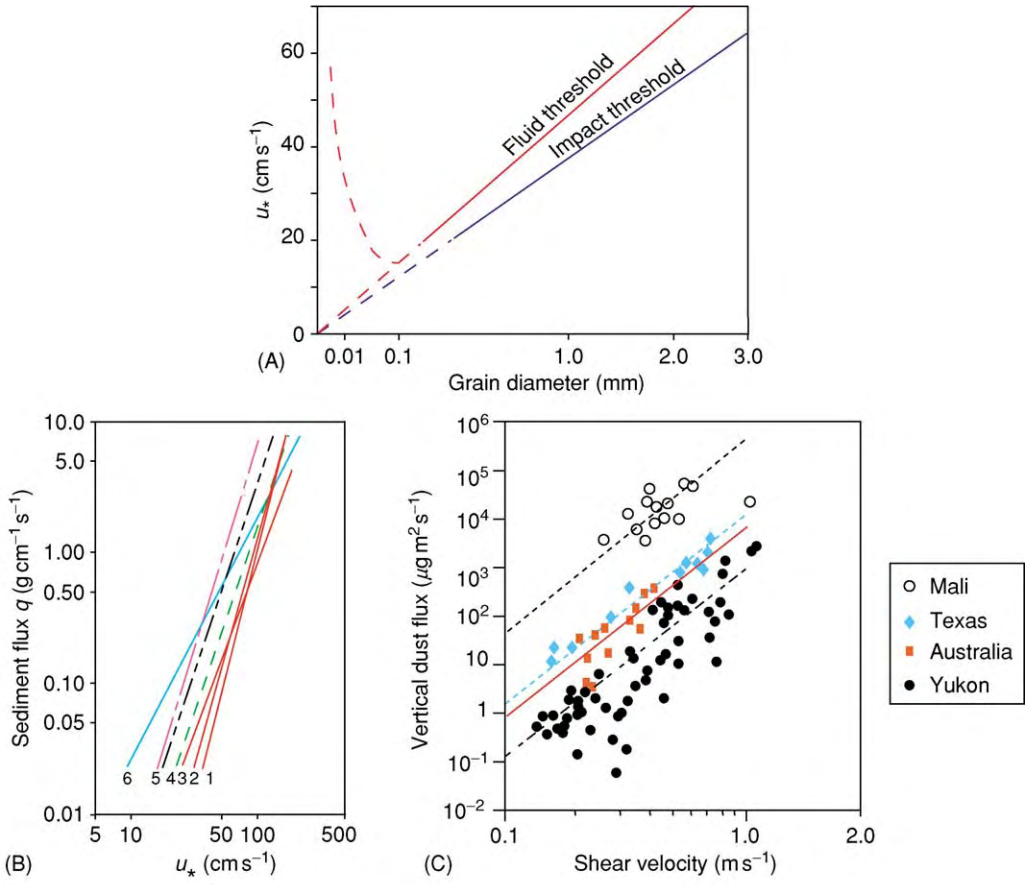


Figure 3 Transport of sediment by the wind. (A) Relation between threshold wind shear velocity and particle size (from Bagnold RA (1941) *The Physics of Blown Sand and Desert Dunes*, p. 265. London: Chapman and Hall). (B) Mass flux of sand as a function of wind shear velocity. Data from laboratory wind tunnel experiments (from Lancaster, 1995). (C) Relations between horizontal flux of sand sized particles and vertical flux of dust (from Nickling WG, McTainsh GH, and Leys JF (1999) Dust emissions from the Channel Country of western Queensland, Australia. *Zeistchrift für Geomorphologie, Supplement Bd.* 116: 1 17.) Data from field experiments.

clay-size particles into the wind and impact coarse grains that are rolled along the bed.

Grains begin to move and sediment is entrained by the wind when fluid forces (lift, drag, and moment) exceed the effects of the weight of the particle and any cohesion between adjacent particles as a result of moisture, salts, or soil crusts. There is a cascading effect as the first particles to move dislodge or impact other grains, so that the number of particles in transport increases exponentially. These particles extract momentum from the wind, reducing its near-bed velocity, so that the transport rate reaches a dynamic equilibrium state over a period of 1 to 2 s. The threshold wind speed at which grains begin to move is strongly dependent on particle size (**Figure 3A**). As in water, there is a particle size at which the transport threshold is at a minimum. For quartz sand in air, the minimum threshold velocity is associated with fine sand (approximately 100 μm diameter).

Most sand particles are moved by saltation in a layer close to the bed, with an exponential decay in the number of particles moved with height above the bed. Particles in saltation move in a parabolic path with a steep ejection limb and a gradual return to the bed. The impacting grains may rebound directly (successive saltation), deform the bed, eject fine particles, or move coarse grains a short distance by reptation or surface creep. The mass flux of sand has been determined by numerous laboratory wind tunnel and field studies to be proportional to the cube of wind shear velocity above a threshold value (**Figure 3B**). For any wind shear velocity, there is a potential rate of sand transport or transport capacity; this is reached only when the availability of sediment is unrestricted (e.g., most loose sand surfaces). In these conditions, the wind is saturated with respect to transport capacity. If the actual transport rate is less than the potential rate, then the wind is undersaturated. Very fine grains (silt and clay size) are inherently resistant to entrainment, yet are readily transported by the wind. Studies have shown the critical role of impacting sand grains in the mobilization of silt- and clay-size particles and have demonstrated the close relations between the horizontal flux of sand-size particles and the vertical flux of fine particles. In these situations, the horizontal mass transport rate is directly related to shear velocity (**Figure 3B**), so dust emissions scale to the fourth power of wind shear velocity (u_*^4; **Figure 3C**). Where there is a limited supply of particles able to abrade soil clods or playa crusts, dust emissions are limited by the supply of particles rather than by the wind shear velocity, and the vertical flux of dust is almost independent of wind shear velocity.

Based on the principle of sediment continuity, winds are erosional if transport rate (or wind shear velocity) increases downwind; deposition occurs when transport rates decrease in the direction of transport, and sediment bypass occurs when there is no

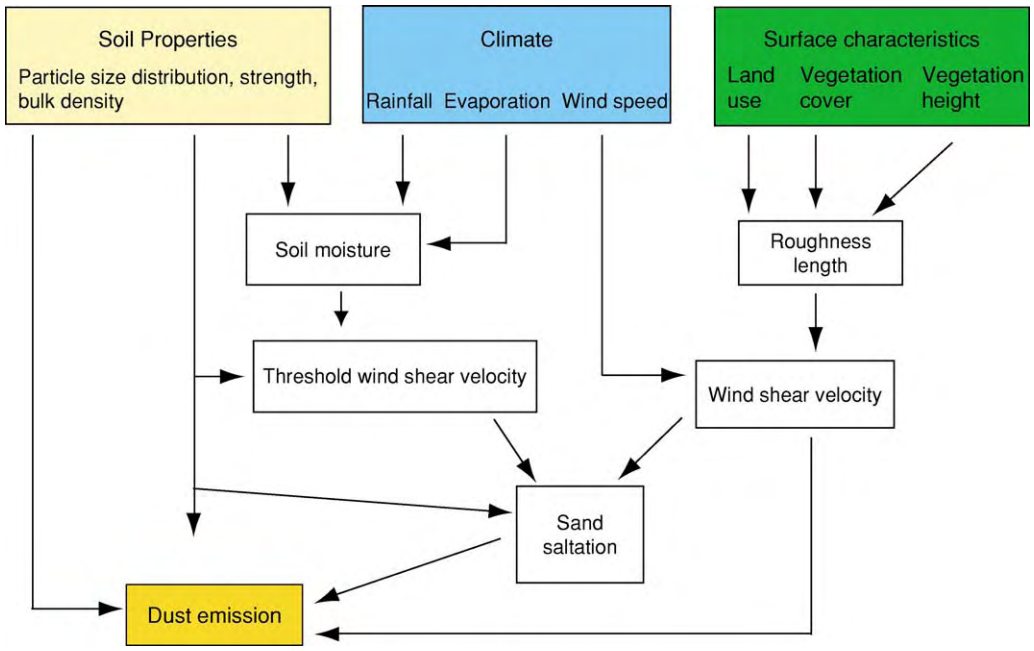


Figure 4 Controls of wind erosion processes on agricultural lands. Reproduced with permission from Leys JF (1999) Wind erosion on agricultural land. In: Goudie AS, Livingstone I, and Stokes S (eds.) *Aeolian Environments, Sediments, and Landforms*, pp. 143 166. Chichester and New York: Wiley.

change in transport rates. These principles are of vital importance to understanding of the dynamics of dunes and sand seas.

Wind Erosion

Erosion by wind involves two linked processes: abrasion (mechanical wearing of coherent materials, including playa crusts and clods created by tillage) and deflation (removal of loose material). Considerable attention has been devoted to the processes and rates of wind erosion because of their impact on agriculture, especially in semi-arid regions, and the implications of dust emissions for air quality. Wind erosion abrades crops, removes organic matter, nutrients, and fertilizer, and changes soil texture. The products of wind erosion (especially dust particles) impact air quality, atmospheric radiative properties, and human health, causing respiratory illnesses and transporting pathogens. Rates of wind erosion vary widely and for a given wind shear velocity are dependent on soil or sediment texture and the degree of crusting and cohesion. The highest emission rates for fine-grained sediment are associated with soils of loamy texture, especially those that have been disturbed by vehicular traffic and/or animals.

Empirical studies of wind erosion rates in relation to environmental conditions and agricultural practices have resulted in the development of predictive models, including the widely used wind erosion equation (WEQ) and the new Wind Erosion Prediction System (WEPS) developed by the United States Department of Agriculture (USDA). The considerable uncertainties associated with such empirically derived models have stimulated development of physically based models of wind erosion in which relations between the physical parameters of the wind and the surface are used to predict dust emissions (Figure 4).

Wind Erosion Landforms

Wind erosion landforms have not received the same level of research attention as have those associated with aeolian deposition. Landforms created by wind erosion can occur on several scales. Large-scale landforms include yardangs (Figure 5A), some of which may occur in systems of ridges and swales aligned with the wind. Yardangs are streamlined forms, with a blunt upwind face and a tapering leeward projection that appears to have evolved to minimize drag. Many yardangs are formed in Quaternary lacustrine deposits; others occur in bedrock, including Precambrian dolomite, Cambrian sandstone, and Mesozoic limestone. Large-scale systems of yardangs are also observed in several areas of Mars. Formation of yardangs appears to involve winds from a very limited range of directions and, at least for bedrock examples, long-term erosion. In some areas (e.g., Kalahari, West Texas), wind erosion has created enclosed depressions with their depth controlled by the water table or resistant strata (Figure 5B). Small-scale aeolian erosional landforms involve wind abrasion of clasts

Figure 5 Wind erosion landforms. (A) Yardangs in Dahkla Oasis, Egypt; (B) Kalahari pans; (C) Ventifacts in the Namib Desert.

and bedrock outcrops to form ventifacts (Figure 5C). Smoothing and polishing, faceting, fluting, and grooving of rock surfaces occur in areas of high-intensity sand transport and affect a variety of lithologies, including granite, marble, and basalt.

Dust Storms

Dust storms are defined as events in which visibility is reduced to 1 km or less as a result of blowing dust (Figure 6). Interest in air quality and in the impact of aerosols on atmospheric properties has promoted intensive study of dust storms and their occurrence in space using meteorological records and satellite images. The Advanced Very High Resolution Radiometer (AVHRR) and the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite images and the Total Ozone Monitoring Sensor(TOMS) have been used to identify major dust source areas and to track the dispersion of particulate matter (Figure 7). Major dust source areas include the Bodelé Depression in Tchad, the Aral Sea area, south-east Iran, and the loess plateau of China. These areas are characterized by seasonally strong winds and areas of fine-grained sediments (including those of playa lakes or previously deposited aeolian materials). The interannual frequency and magnitude of dust storms are strongly linked to climatic variability, so that in many areas dust emissions are inversely correlated to rainfall. Although dust storm intensity in some areas is affected by human impacts, including water diversion for irrigated agriculture and poor land use practices, the largest sources of dust emissions in the Sahara are in uninhabited regions.

Aeolian Deposits

Aeolian deposits include sand seas and dune fields, deposits of silt (loess), and fine-grained material that forms a significant component of desert margin and other soils. Many marine deposits, especially those downwind of major desert areas, also include significant proportions of wind-transported material.

Silt and Clay Size

Terrestrial deposits of wind-transported silt-sized quartz particles, termed loess, cover as much as 10% of Earth's land surface. Loess deposits are widespread in areas of northern China, central Russia, the Ukraine, Kazakhstan, central Europe, Argentina, Alaska, and the central United States. Much of the material was thought to be derived from silt particles produced by glacial grinding and supplied to aeolian processes by glacial outwash ('glacial loess'), but aeolian abrasion and other processes, including frost shattering, salt weathering, and fluvial comminution, are important, especially in the formation of 'desert loess'. At least some loess (e.g., in Colorado and Nebraska) has its origins in weathered bedrock (e.g., siltstones, volcaniclastic deposits). Loess deposits and their interstratified buried palaeosols provide a long record of Quaternary climate change (Figure 8). In these sequences, the rate of loess deposition varies

Figure 6 Reduced visibility from major dust storm in West Texas, 15 December 2003. Compare with clear day (inset). Photographs by J Lee, Texas Tech University.

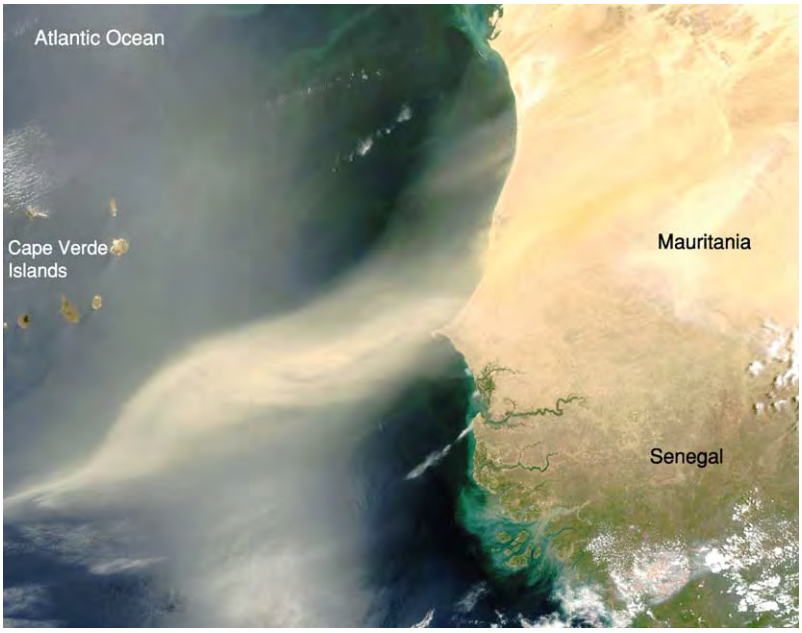


Figure 7 Dust plumes extend up to 200 km offshore from the western Sahara in Senegal and Mauritania, 30 April 2003. This was the fourth day of the storm. Details and many other dust storm images are available from the National Aeronautics and Space Administration website at http://www.visibleearth.nasa.gov.

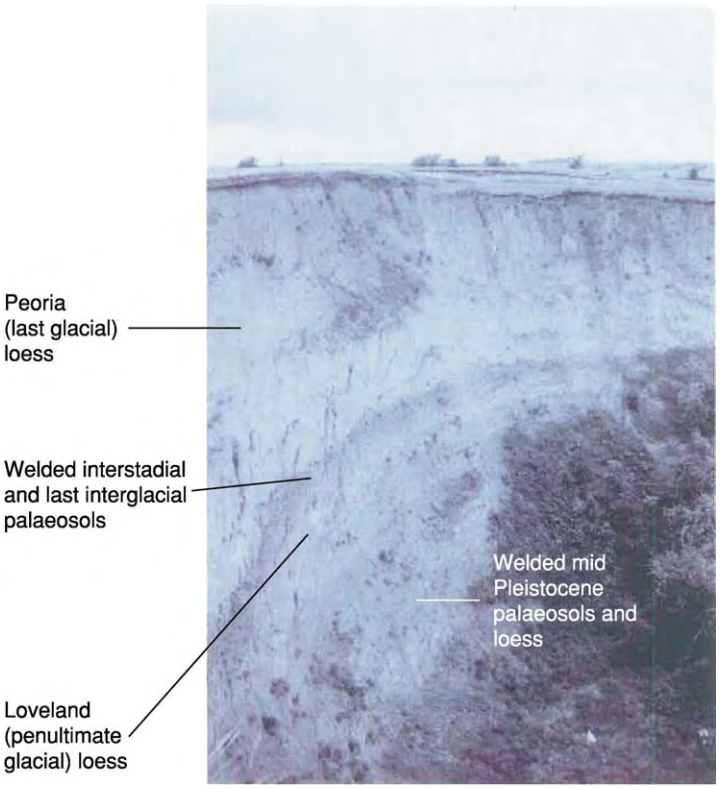


Figure 8 Loess deposits in Nebraska. Photograph and annotations by D Muhs, USGS.

over time, being greater in glacial periods (more arid, windier, less vegetation) compared to interglacial periods (less arid and windy, more vegetation), when the influx of particles was low enough to allow soils to form.

Silt- and clay-sized material of aeolian origin is also an important component of many desert margin soils. Deposition of silt plays a role in the formation of many stone pavement surfaces in desert regions (desert pavement, or 'reg'). These surfaces are characterized by a surface layer of gravel or larger clasts that overlie fine-grained materials. Older work argued that desert pavement surfaces were deflationary and that the mantle of coarse clasts resulted from winnowing and removal of fine material. Detailed studies of these surfaces now show, however, that the surface layer of gravel-sized clasts rests on an accretionary mantle of soil-modified dust that may be a metre or more thick. The dust is trapped by the clasts and deposited between them. The fine material is incorporated into the mantle by pedogenic processes (e.g., shrinking and swelling of clay minerals), so that the clasts remain at the surface as it inflates over periods of thousands of years. Thus desert pavement surfaces developed on bedrock and alluvial surfaces are actually depositional features.

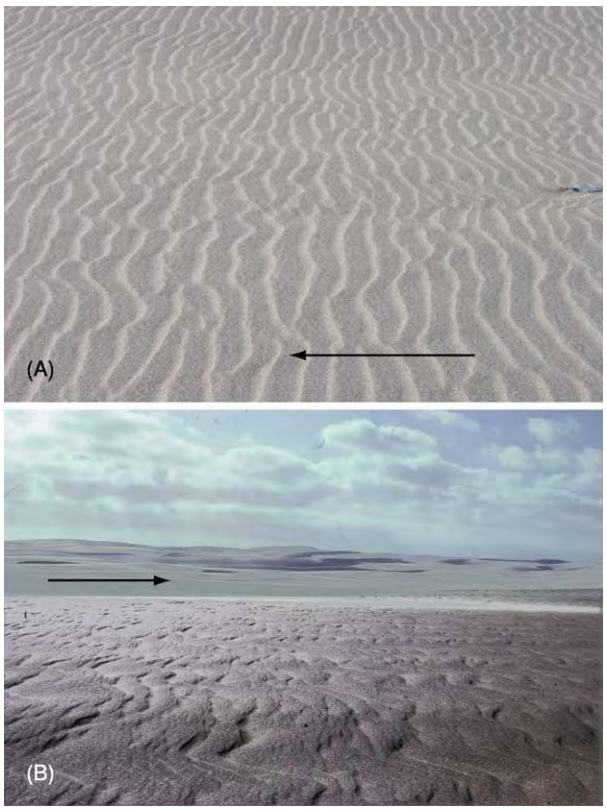


Figure 9 (A) Wind ripples in the Victoria Valley, Antarctica; (B) granule ripples (megaripples) at Skeleton Coast, Namibia. Arrows indicate formative wind direction.

Sand Dunes

Aeolian dunes form part of a hierarchical system of self-organizing bedforms developed in wind-transported sand; this system comprises wind ripples (spacing 0.05–2 m), individual simple dunes or superimposed compound and complex dunes (spacing 50–500 m), and compound and complex megadunes, or draa (spacing >500 m). Dunes occur wherever there is a sufficient supply of sand-sized sediment, winds to transport that sediment, and conditions that promote deposition of the transported sediment. These requirements are satisfied in two main environments: coastal areas with sandy beaches and onshore winds, and desert areas. Most dunes occur in contiguous areas of aeolian deposits termed 'sand seas' (>100 km^2) or 'dune fields'.

Wind ripples (Figure 9) typically have a wavelength of 0.05–0.2 m and an amplitude of 0.005–0.010 m. They are ubiquitous on sand surfaces except for those undergoing very rapid erosion or deposition, and form because a flat sand surface over which sand transport by saltation and reptation occurs is dynamically unstable. Field and laboratory wind tunnel experiments and numerical modelling suggest that ripple wavelength is a function of grain transport distance and that ripples develop as a result of stochastic variations in the flux of reptating coarser grains driven by the impacts of finer saltating grains. Once the initial variation in bed topography develops, variations in microtopography influence the intensity of saltation impacts and therefore the local mass flux of reptating grains, and the pattern becomes self-organizing by the merging and growth of a characteristic wind ripple wavelength and amplitude.

Aeolian dunes occur in a self-organized pattern that develops as sand surfaces respond to the wind regime (especially its directional variability) and the supply of sand. Sand dunes occur in four main morphological types (Figure 10): crescentic (transverse), linear, star, and parabolic. The simplest dune types and patterns form in areas characterized by a narrow range of wind directions (unidirectional wind regime). In the absence of vegetation, crescentic dunes will be the dominant form. Isolated crescentic dunes, or barchans, occur in areas of limited sand supply; they coalesce laterally to form crescentic (barchanoid) ridges as sand supply increases (Figure 10A and B). Linear dunes, which form in areas of bimodal or wide unimodal wind regimes, are characterized by their length (often more than 20 km), sinuous crestline, parallelism, and regular spacing (Figure 10C and D). The essential mechanism for linear dune formation is

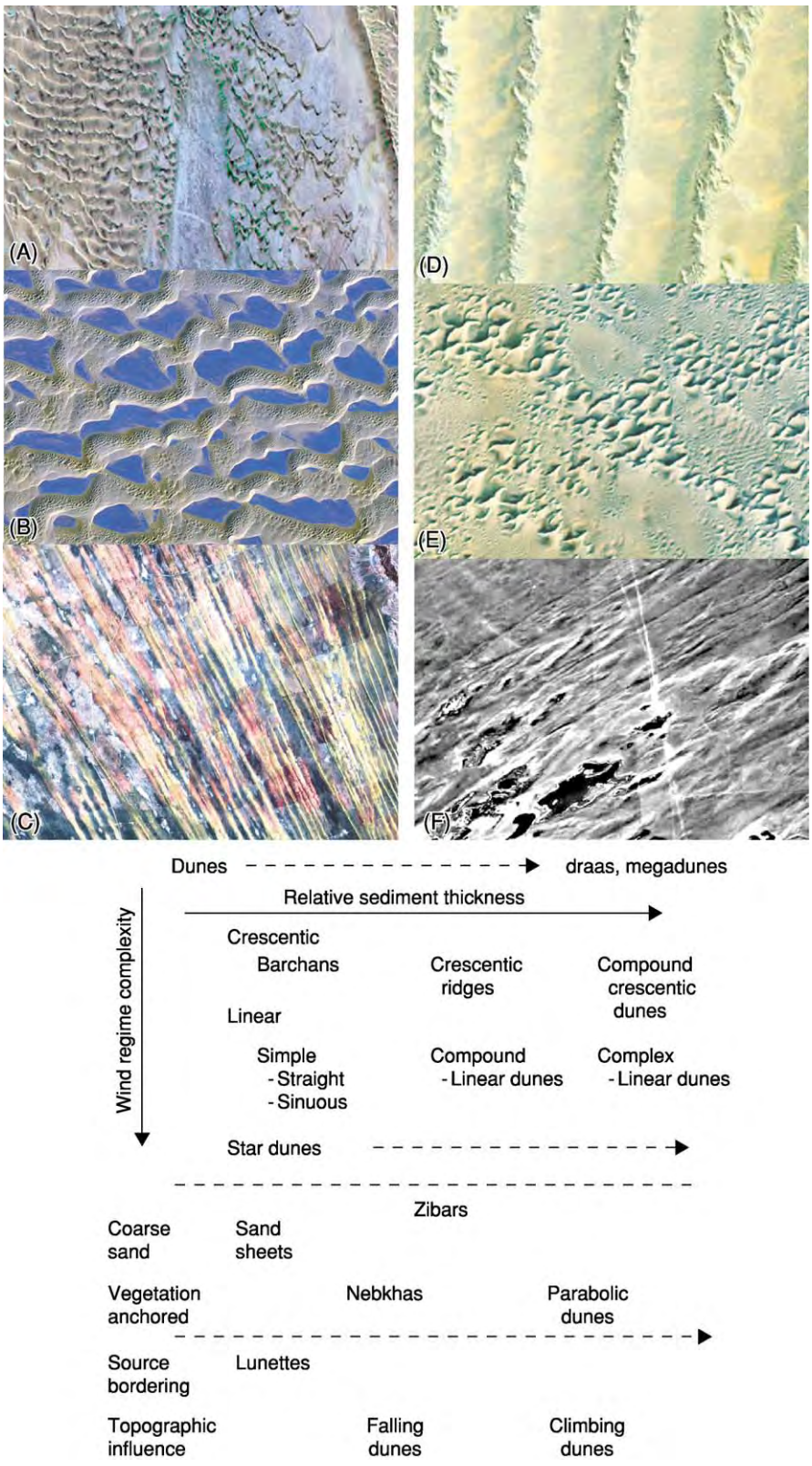


Figure 10 Morphological classification of aeolian sand dunes, and satellite images and aerial photographs of major dune types: (A) barchans and crescentic dunes, Namib Sand Sea; (B) compound crescentic dunes, Liwa, United Arab Emirates; (C) simple linear dunes, Kalahari Desert; (D) complex linear dunes, Namib Sand Sea; (E) star dunes, Gran Desierto, Mexico; (F) parabolic dunes, Casper, Wyoming.

the deflection of winds that approach the crest at an oblique angle, then parallel wind flow to the lee side and transport of sand along the dune, so elongating it. Star dunes have a pyramidal shape, with three or four sinuous sharp-crested arms radiating from a central peak and multiple avalanche faces (Figure 10E). Star dunes occur in multidirectional or complex wind regimes and are the largest dunes in many sand seas, reaching heights of more than 300 m. The development of star dunes is strongly influenced by the high degree of form–flow interaction, which occurs as a result of seasonal changes in wind direction, and the existence of a major lee-side secondary circulation. Parabolic dunes (Figure 10F) are characterized by a U or V shape with a 'nose' of active sand and two partly vegetated arms that trail upwind. They are common in many coastal dunefields and semi-arid inland areas and often develop from localized blowouts in vegetated sand surfaces. Other important dune types include zibars (low-relief dunes lacking discernible slip faces; commonly occur on sand sheets), nebkhas (hummock dunes; anchored by vegetation; common in may coastal dunefields), lunettes (form downwind of small playas; often composed of sand-sized clay pellets), and a variety of topographically controlled dunes (climbing and falling dunes, echo dunes).

Relations between dune types and wind regimes indicate that the main control of dune type is the directional variability of the wind regime (Figure 11). Grain size, vegetation cover, and sediment supply play subordinate roles.

Dune Processes and Dynamics

The development and morphology of all dunes are determined by a complex series of interactions between dune morphology, airflow, vegetation cover, and sediment transport rates. As dunes grow, they project into the atmospheric boundary layer so that streamlines are compressed and winds are accelerated up the stoss (windward) slope by up to 2.5 times, depending on the size and steepness of the dune. Flow acceleration results in an exponential increase in sediment transport rates towards the dune crest, resulting in erosion of the stoss slope. In the lee of the crest of dunes, wind velocities and transport rates decrease rapidly as a result of flow expansion between the crest and brink of the lee or avalanche face and flow separation on the avalanche face. Flow separation causes fallout of previously saltating sand grains and development of avalanche (slip) faces. Secondary flows, including lee-side flow diversion, are also important when winds approach the dune obliquely, and are an important process on linear and many star dunes. Migration of crescentic dunes occurs as a result of erosion of the stoss slope and deposition on the lee. Their movement can be described by $c = Q/yh$, where c is the migration rate, Q is the bulk volumetric sand transport rate, y is the bulk density of sand, and h is dune height. Linear dunes

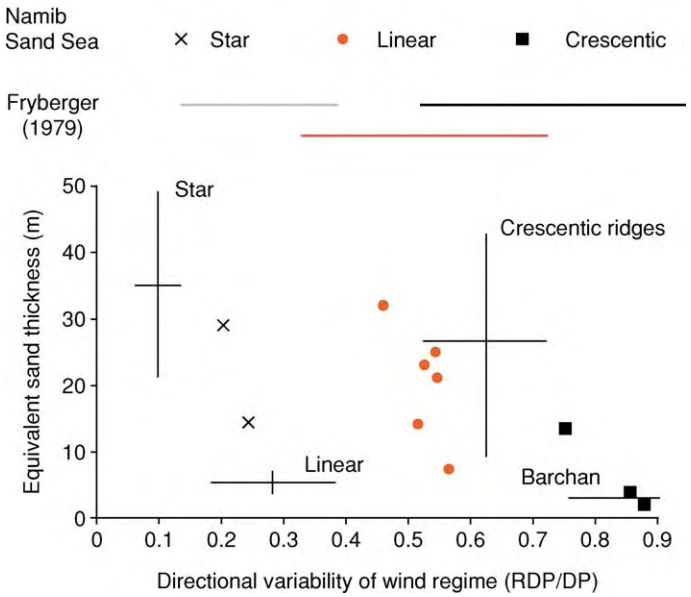


Figure 11 Relations between dune types and wind regimes. The directional wind variability index is the ratio of the resultant drift potential (RDP) to the drift potential (DP). Data from Fryberger SG (1979) Dune forms and wind regimes. In: McKee ED (ed.) *A Study of Global Sand Seas: United States Geological Survey, Professional Paper, 1052*, pp. 137 140. Reston, VA: USGS, Lancaster (1983), Wassan and Hyde (1983).

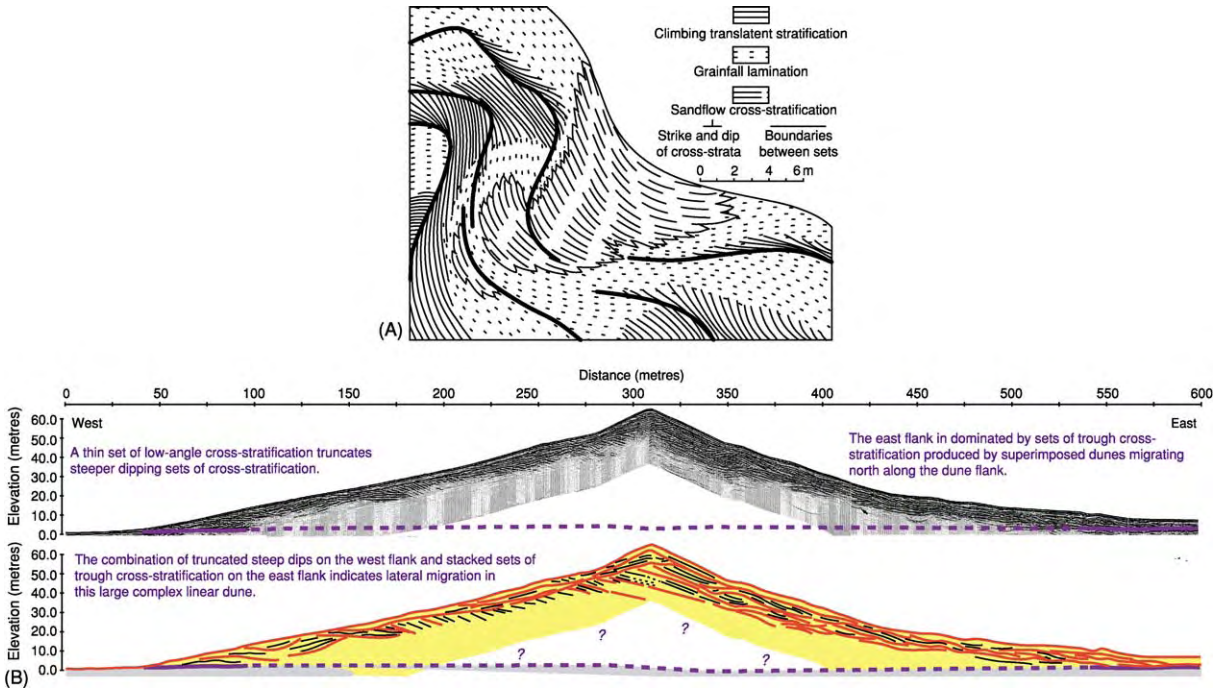


Figure 12 Dune sedimentary structures. (A) Crescentic dunes. Reproduced with permission from Hunter RE (1977) Basic types of stratification in small eolian dunes. *Sedimentology* 24: 361 388. (B) Linear dunes. Courtesy of CS Bristow, N Lancaster, and SD Bailey. Poster presented at British Sedimentological Research Group Annual Meeting, 1999.

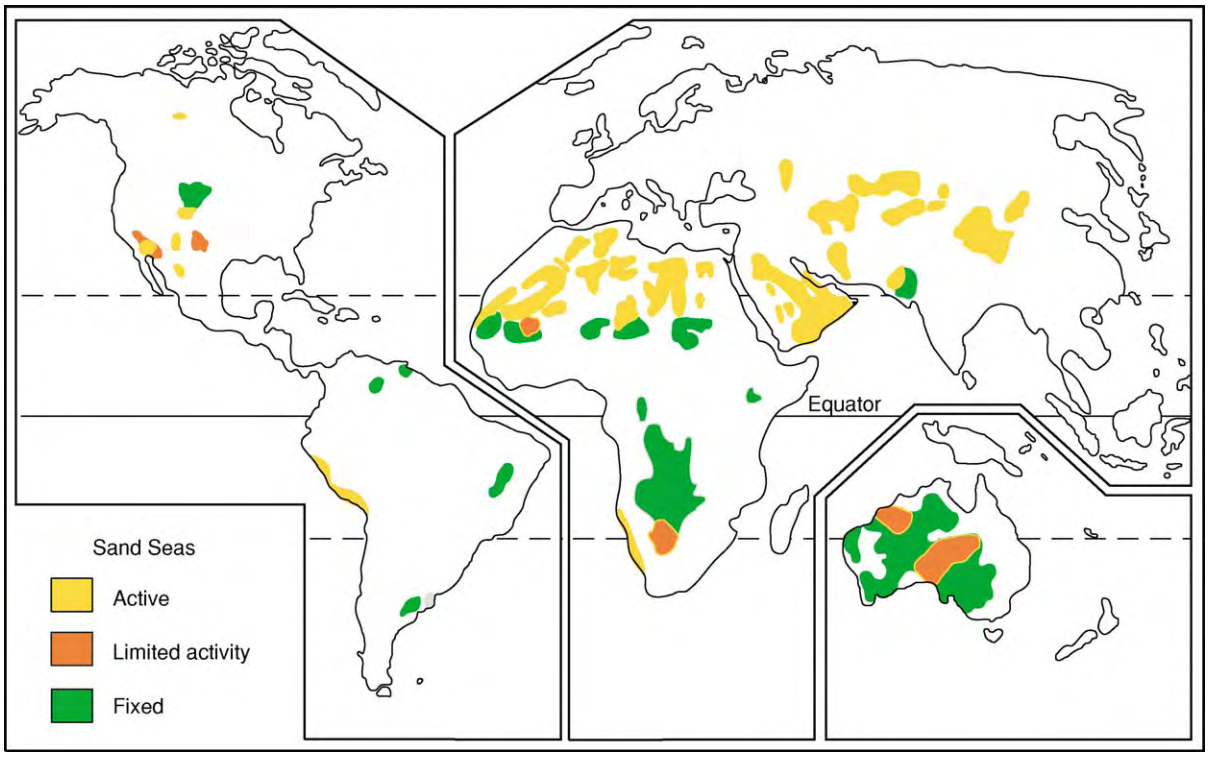


Figure 13 Distribution of major sand seas.

migrate laterally at a very slow rate and also extend downwind. Star dunes are essentially stationary forms.

Dune Sediments

Dunes are composed of moderately to well-sorted sands (63–1000 μm), with a mean grain size in the range 160–300 μm. Most dune sands are composed of quartz, but may include significant quantities of feldspar. Dunes composed of volcaniclastic materials occur in some areas. Carbonate-rich dune sands are found adjacent to many subtropical coastlines and gypsum dunes occur close to some playas.

Three primary modes of deposition occur on dunes: migration of wind ripples, fallout from temporary suspension of grains in the flow separation zone (grainfall), and avalanching of grains initially deposited by grainfall (grainflow). These deposition modes create three basic types of aeolian sedimentary structures: wind ripple laminae, grainfall laminae, and grainflow cross-strata. The packages of laminae may be separated by a hierarchy of bounding surfaces representing episodes of reactivation, dune growth, and migration of dunes in conditions of bedform climbing.

Crescentic dunes are dominated by grainfall and grainflow cross-strata (Figure 12A). Linear dune structures imaged using ground-penetrating radar show a change in the dominant type of sedimentary structures as the dune develops. Small dunes are dominated by stacked sets of cross-strata that dip in opposite directions and are formed by the migration of the elements of a sinuous crestline up the dune. The structures of larger dunes are dominated by trough cross-strata created by the migration of superimposed bedforms (Figure 12B). Sedimentary structures of star dunes are not well known, but are likely complex.

Sand Seas

Desert sand seas are major depositional landforms that comprise dunes of varying size, spacing, and morphological type; sand seas also have areas of sand sheets as well as interdune deposits (e.g., playas, lacustrine deposits) and extradune fluvial, lacustrine, and marine sediments. Major areas of sand seas lie in

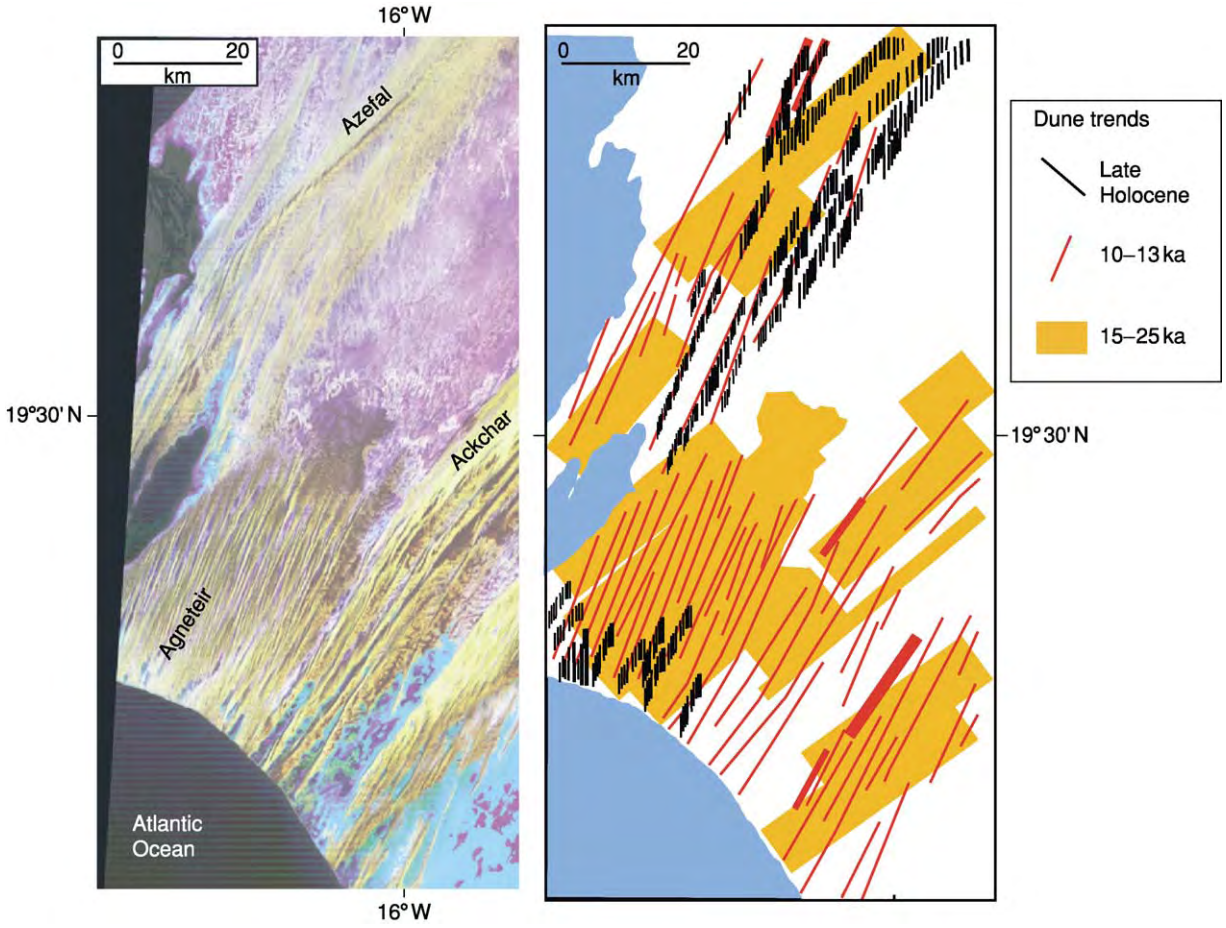


Figure 14 Linear dune generations in western Mauritania; LANDSAT image (left) and interpreted dune trends (right).

the tectonically stable desert regions of the Sahara, Arabian Peninsula, Australia, and southern Africa and in the enclosed basins of central Asia (Figure 13). Sand seas and dune fields form part of well-defined regional- and local-scale sediment transport systems in which sand is moved by wind from source areas (e.g., distal fluvial deposits, sandy beaches) via transport pathways to depositional sinks. Sand seas accumulate downwind of source zones wherever sand transport rates decrease as a result of regional and/or local changes in wind speed and directional variability, so that the influx of sand exceeds outflux.

Most sand seas exhibit significant spatial variations in dune type and sediment characteristics. These may be the result of regional changes in wind regime, but can also be the product of multiple periods of sand influx and stabilization, controlled by changes in climate and sea-level. Areas of different dune types therefore often represent distinct generations of dunes. Advances in dating techniques (e.g., optically stimulated luminescence) are beginning to provide absolute chronologies of dune generations in sand seas (Figure 14).

Preservation of Aeolian Deposits in the Rock Record

Formation of an aeolian rock record requires three sets of processes: sand sea construction, accumulation of aeolian strata, and preservation of the

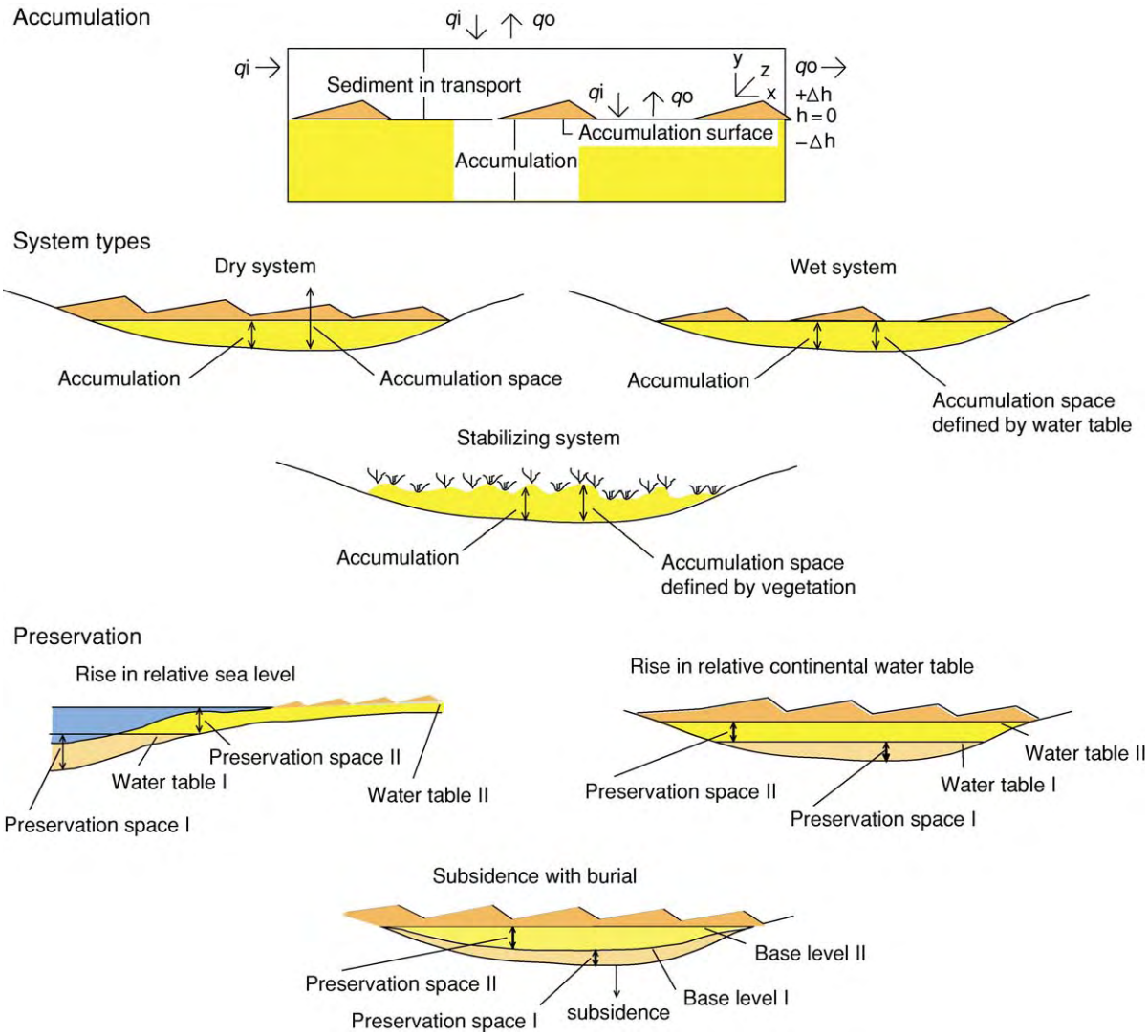


Figure 15 Accumulation and preservation of aeolian sands. Accumulation depends on the transport rate *qi* into and *qo* out of the system. Reproduced with permission from Kocurek G (1998) Aeolian system response to external forcing factors a sequence stratigraphic view of the Saharan region. In: Alsharan AS, Glennie KW, Whittle GL, and Kendall CGSC (eds.) *Quaternary Deserts and Climatic Change*, pp. 327 338. Rotterdam/Brookfield: Balkema.

accumulation. Accumulation of aeolian deposits occurs when that sediment is transferred from bedforms to the accumulation such that its surface rises over time (Figure 15). In dry aeolian systems, the water table is sufficiently far below the surface that it plays no role in the dynamics of the accumulation, which is instead dominated by the wind. In wet aeolian systems, the water table is at or near the surface and so controls accumulation dynamics. Where vegetation controls accumulation dynamics, the system is known as a stabilizing aeolian system. In all cases, accumulation occurs when sediment transport rates decrease downwind and the excess sediment is deposited; and/or when the concentration of sediment in transport (approximately indicated by dune size) decreases over time. In dry systems, spatial or temporal decreases in wind strength give rise to accumulation of sediment by bedform climbing; in wet systems, transport rates

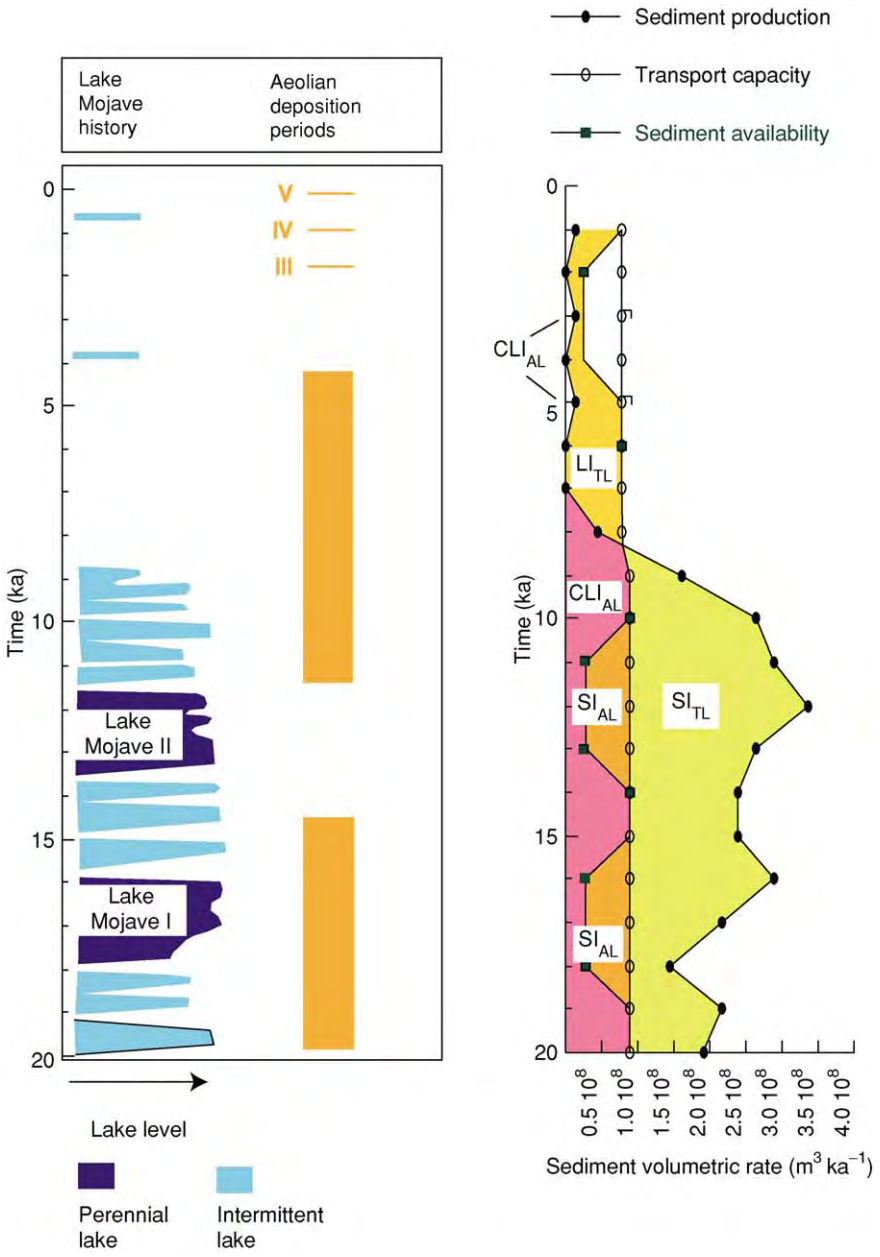


Figure 16 Example of controls on an aeolian system state in the Mojave Desert, both transport limited (TL) and availability limited (AL). CLI, Contemporaneous and lagged input; LI, lagged input; SI, stored input. Reproduced with permission from Kocurek G and Lancaster N (1999) Aeolian sediment states: theory and Mojave Desert Kelso Dunefield example. *Sedimentology* 46(3): 505 516.

decrease downwind because sediment is trapped by dunes migrating into a wet area, resulting in a decrease in dune size, or dunes shrink because the water table rises under the whole system.

For these accumulations to be preserved, there must be subsidence and burial and/or a rise of the water table through the accumulation. Because wind has the potential to erode any sand accumulation above the water table if it is undersaturated with respect to transport capacity, exhaustion of sediment supply may promote erosion unless the accumulation is buried by non-aeolian strata or the water table is at

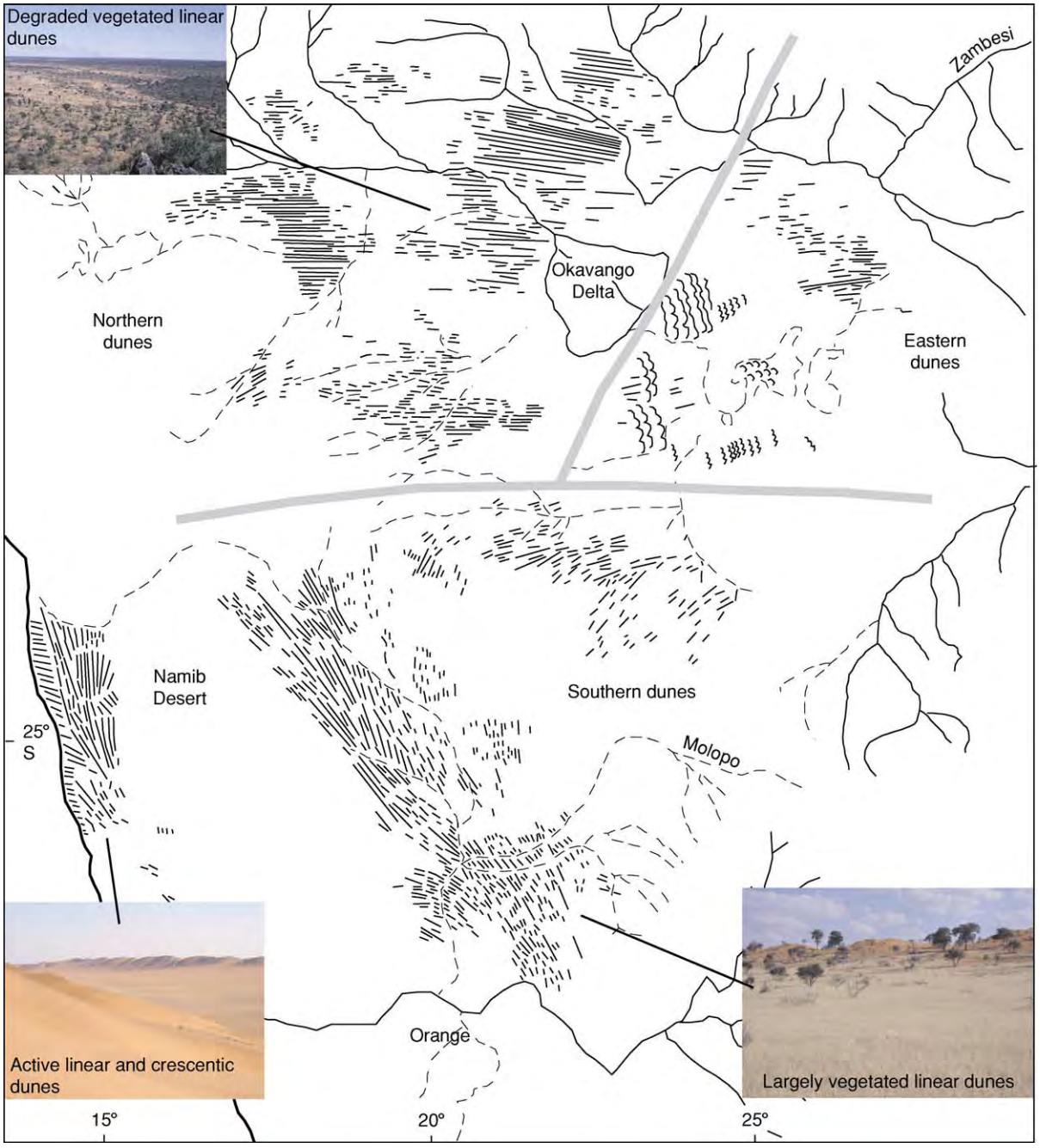


Figure 17 Relict dune systems in the Kalahari Desert of southern Africa. Dune patterns reproduced with permission from Lancaster N (2000) Eolian deposits. In: Partridge TC and Maud RR (eds.) The Cenozoic of Southern Africa, pp. 73 87. Cape Town, South Africa: Oxford University Press.

or near the surface. Preservation of aeolian accumulations is favoured in tectonic basins subject to a long history of sediment influx and subsidence, in coastal areas in which marine transgressions can bury aeolian accumulations or cause a rise in the regional water table, and in continental areas in which the water table rises permanently because of tectonism or climate. These conditions were satisfied during the formation of the extensive Jurassic aeolian sandstones of the western United States, which were deposited in a subsiding foreland basin supplied with large amounts of sediment from the North American Craton. Marine transgressions from the west preserved parts of the record via sabkha and marine strata. By contrast, most aeolian accumulations in the Sahara and many other deserts are unfavorably located with respect to preservation. One exception is that along the coast of Mauritania and its continental shelf, where aeolian sands sourced from inland craton areas are preserved by the postglacial rise in sea-level and the associated rise in water tables.

Effects of Climate and Sea-Level Change on Aeolian Systems

Many aeolian deposits and landforms formed during the Quaternary Era, thus they have been affected by climatic and sea-level changes associated with glacial–interglacial cycles. The effects of changes in external forcing factors are manifested by changes in the supply of sediment of a size suitable for transport by the wind, the availability of this sediment for transport (determined by vegetation cover and soil moisture), and the mobility of this sediment (controlled by wind strength). The interactions between these variables can be evaluated in terms of the state of the aeolian system and the limiting factors

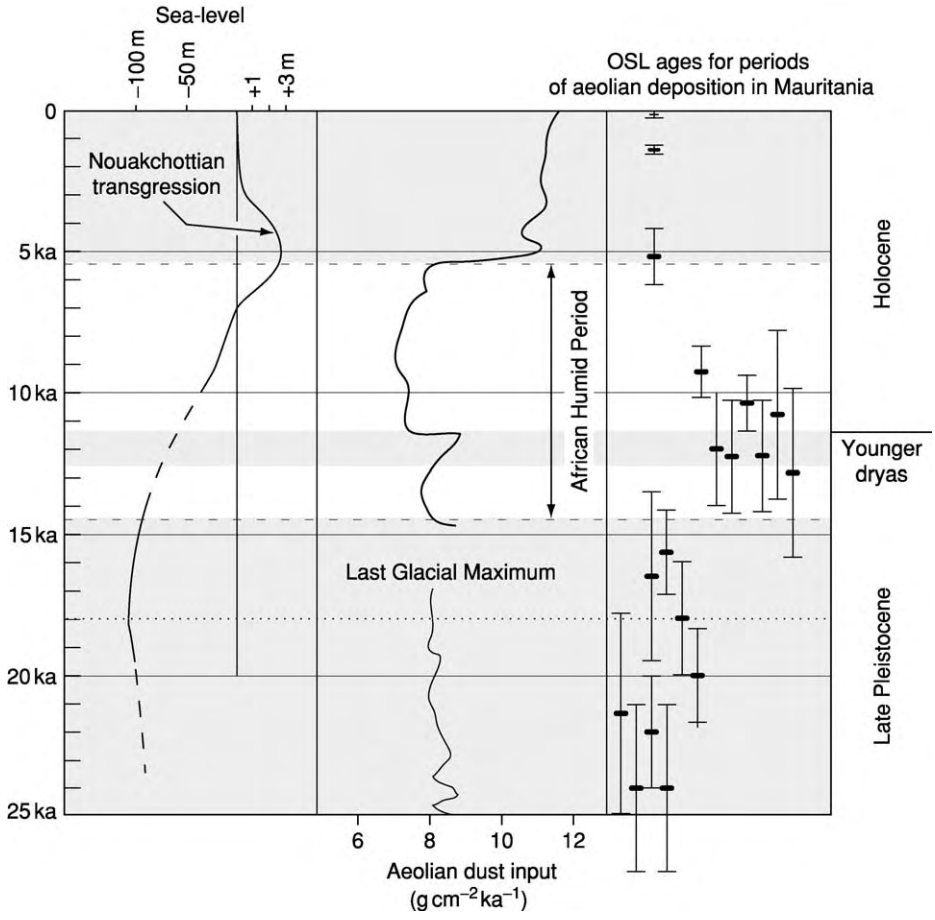


Figure 18 Relations between periods of dune construction, sea level, and dust flux to the Atlantic Ocean (dust flux after deMenocal P, *et al.* (2000) Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quaternary Science Reviews* 19(1 5): 347 361). OSL ages for aeolian deposition in Mauritania from Lancaster N, *et al.* (2002) Late Pleistocene and Holocene dune activity and wind regimes in the western Sahara of Mauritania. *Geology* 30(11): 991 994.

identified (Figure 16). The changes in system state are manifested as episodic accumulation of aeolian deposits, resulting, for example, in sequences of loess accumulation interspersed by palaeosols (Figure 8), or formation of different dune generations in aeolian sand seas and dune fields (Figure 14). Consideration of variations in the supply of sediment for dune construction or loess accumulation, together with changes in the availability and mobility of this sediment, has prompted new explanations of regional variations in loess stratigraphy in North America and periods of aeolian accumulation in the Mojave and Arabian deserts. For example, loess accumulation was low during the last glacial maximum in Alaska, despite high rates of sediment supply and mobility, because tundra vegetation is ineffective at trapping sediment. This is in contrast to areas in the Great Plains and Central Lowlands, where accumulations of thick sequences of loess during the last glacial period were favoured by high rates of sediment supply from glacial outwash and high rates of accumulation in the boreal forest that then covered these areas. In the Mojave Desert, high rates of aeolian accumulation and dune construction took place during times when palaeolake systems were drying out seasonally or annually. Sediment deposited in these basins during wetter periods was mobilized during periods of drier climates. Aeolian construction ceased when these supplies of sediment were exhausted (Figure 16).

In addition to being influenced by climate and sea-level change, aeolian sediments and landforms may also provide evidence for past climate changes. Examples include loess–palaeosol sequences, as well as dune systems, now vegetated and inactive, indicating the former extent of arid conditions and past wind regimes in what are today areas of semi-arid or sub-humid climates (Figure 17), and increased rates of aeolian dust deposition in ocean sediments and ice cores (Figure 18).

See Also

Sedimentary Environments: Deserts. **Sedimentary Processes:** Erosional Sedimentary Structures; Depositional Sedimentary Structures; Deposition from Suspension. **Soils:** Palaeosols.

Further Reading

Bagnold RA (1941) *The Physics of Blown Sand and Desert Dunes*. London: Chapman and Hall.

Bristow CS, Bailey SD, and Lancaster N (2000) Sedimentary structure of linear sand dunes. *Nature* 406: 56–59.

Fryberger SG (1979) Dune forms and wind regimes. In: McKee ED (ed.) *A Study of Global Sand Seas: United States Geological Survey, Professional Paper, 1052*, pp. 137–140. Reston, VA: USGS.

Goudie AS, Livingstone I, and Stokes S (eds.) (1999) *Aeolian Environments, Sediments, and Landforms*. New York: Wiley.

Kocurek G (1998) Aeolian system response to external forcing factors – a sequence stratigraphic view of the Saharan region. In: Alsharan AS, Glennie KW, Whittle GL, and Kendall CGSC (eds.) *Quaternary Deserts and Climatic Change*, pp. 327–338. Rotterdam/Brookfield: Balkema.

Kocurek G and Lancaster N (1999) Aeolian sediment states: theory and Mojave Desert Kelso Dunefield example. *Sedimentology* 46(3): 505–516.

Laity JE (1994) Landforms of aeolian erosion. In: Abrahams AD and Parsons AJ (eds.) *Geomorphology of Desert Environments*, pp. 506–535. London: Chapman and Hall.

Lancaster N (1983) Controls of dune morphology in the Namib sand sea. In: Ahlbrandt TS and Brookfield ME (eds.) *Eolian Sediments and Processes. Developments in Sedimentology*, pp. 261–289. Amsterdam: Elsevier.

Lancaster N (1995) *Geomorphology of Desert Dunes*. London: Routledge.

Lancaster N (2000) Eolian deposits. In: Partridge TC and Maud RR (eds.) *The Cenozoic of Southern Africa*, pp. 73–87. Cape Town, South Africa: Oxford University Press.

Livingstone I and Warren A (1996) *Aeolian Geomorphology: An Introduction*. Harlow: Addison Wesley Longman.

Muhs DR and Bettis EA III (2003) Quaternary loess–paleosol sequences as examples of climate driven sedimentary events. In: Chan MA and Archer AW (eds.) *Extreme Depositional Environments, Mega End Members in Geologic Time*, pp. 53–74. Boulder, Colorado: Geological Society of America.

Pye K (1987) *Aeolian Dust and Dust Deposits*. London: Academic Press.

Pye K and Tsoar H (1990) *Aeolian Sand and Sand Dunes*. London: Unwin Hyman.

Wasson RJ and Hyde R (1983) Factors determining desert dune type. *Nature* 304: 337–339.

Wiggs GFS (2001) Desert dune processes and dynamics. *Progress in Physical Geography* 25(1): 53–79.

Catastrophic Floods

A J Russell, University of Newcastle upon Tyne, Newcastle upon Tyne, UK

Introduction

At present there is no widely recognized definition of a flood other than a 'notable flow of water'. Since the diluvial theory of the biblical flood, cataclysms and catastrophic flooding have drawn the attention of geologists throughout the development of the subject and to the present day. Catastrophic floods occur worldwide within a wide range of depositional environments and hydroclimatic settings. Catastrophic floods represent a major natural hazard within riverine environments, generating substantial loss of life and infrastructure. As a surface process, catastrophic floods are capable of generating immense landscape change and the transfer of huge sediment volumes. Large meltwater outbursts associated with continental deglaciation have recently been invoked as a trigger mechanism for changes in ocean circulation and consequent climate changes.

Catastrophic floods have been defined in a number of ways. From an applied perspective, catastrophic floods can be defined in terms of human and monetary loss. Floods can also be judged truly catastrophic, depending upon their impact on channel morphology and sedimentology or the work they do in transporting sediment. The extent to which a flood is catastrophic also depends upon the time-scale over which one wishes to consider its impacts. Preservation of erosional landforms and sedimentary successions associated with floods over geological time-scales are likely to reflect truly catastrophic events and the presence of appropriate depositional environments for preserving evidence of high-magnitude floods. Over shorter time-scales, a flood may be judged catastrophic if its impacts result in an irreversible change in fluvial geomorphology. High-energy flows are known to result in a suite of 'catastrophic processes', which do not operate under non-flood conditions and which achieve substantial and distinctive impacts. Catastrophic processes can, however, also occur under nonflood flow conditions within, for example, steep bedrock fluvial systems and subglacial drainage systems characterized by high stream powers.

Causes of Catastrophic Floods

Catastrophic floods require the sudden input of large volumes of water to a flood routeway. Flood generation mechanisms provide a major control on flood magnitude, frequency, and characteristics (Figure 1). Traditional classification schemes of flood causes examine the relationship between floods and climatological variables. More recently, attention has focussed on identifying the physical limits on the magnitude of floods generated by various mechanisms.

Rainfall is the most common cause of flooding in rivers around the globe. Rainwater is a major input to the fluvial system, often having travelled through hillslopes or more rapidly as overland flow. Rainfall intensities tend to be high over short time periods and on a relatively local scale. Meteorological and climatological controls on flood-producing precipitation events can be identified over a range of spatial and temporal scales (Figure 2). This hydroclimatological approach uses understanding of the controls on flood-generating precipitation events to characterize the distribution of floods in space and time. Other approaches have considered maximum flood discharges, which can be yielded from a certain catchment area. Empirically derived 'envelope curves' illustrate major differences in rainfall–runoff relationships between different climatic regions. Envelope curves do not, however, consider the relative role of specific drainage basin characteristics in generating flood flows (Figure 3).

Snowfall generally reduces the immediate potential for flooding by putting water into storage for more gradual release upon melting. However, the progressive or rapid melt of a snow-pack can generate substantial floods, especially as a 'nival' flood in springtime. Nival floods commonly affect rivers with large catchments, which allow synchronous water input from all parts of the catchment area. Snowmelt rates are controlled by the amount of available energy. Heavy rainfall onto a snow-pack will enhance peak flood discharges from a catchment, although modest precipitation onto a snow-pack may delay flows, resulting in lower peak discharges. Nival flows are also associated with secondary storage-release floods resulting from the failure of temporary snow dams or river ice jams.

Melt of glacier ice is a major control on discharge within proglacial river channels. Although melt-induced flows commonly exhibit discharge variations on diurnal and synoptic time-scales, they are not normally considered floods unless associated with unusually high ablation. This may occur, for example, during subglacial volcanic eruptions or where there are enhanced levels of geothermal activity. More commonly, however, glacial meltwater is stored

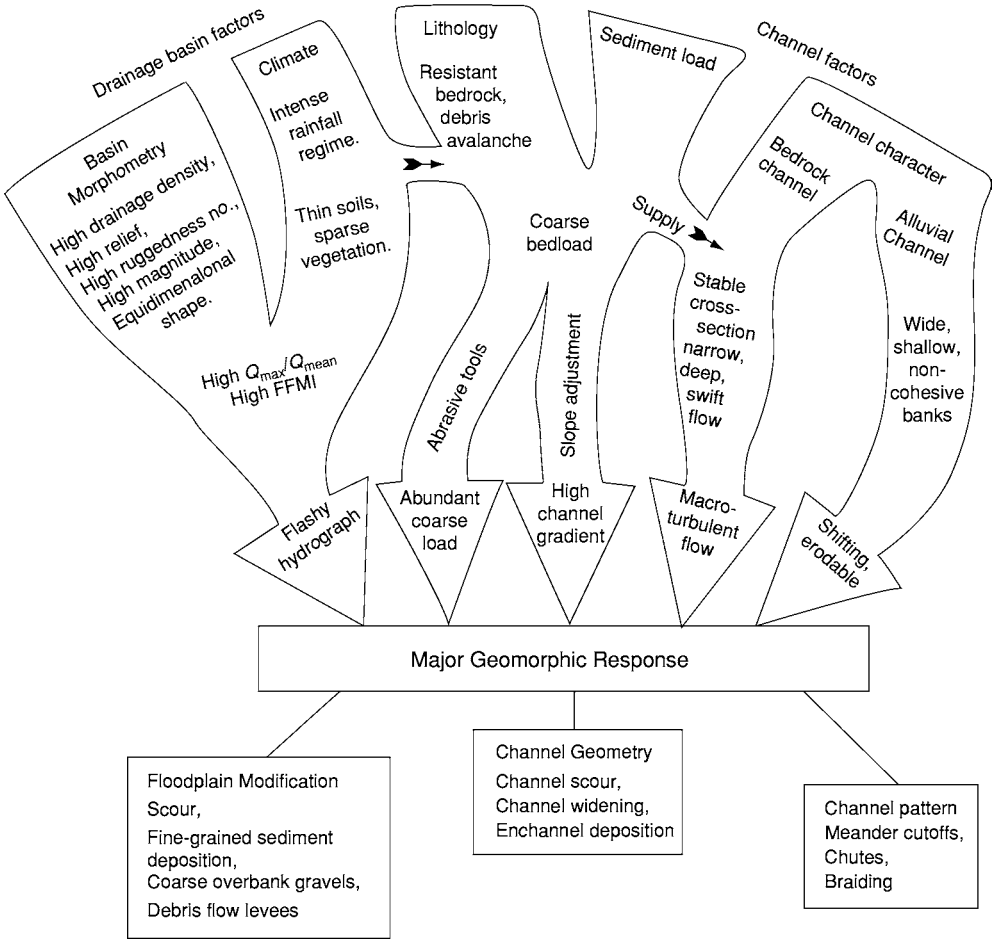


Figure 1 Controls on flood magnitude, frequency, and characteristics. After Kochel RC (1988) Geomorphic impact of large floods: review and new perspectives on magnitude and frequency. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 169 187. London: Wiley.

on top of, against, within, and underneath glaciers, providing the potential for storage-release events (Figure 4).

Natural or man-made dam failures can release outbursts of large volumes of water over relatively short time-spans. Natural dams commonly comprise vegetation debris, river ice, glacier ice, or sediment. Dam formation mechanism, composition, and geographic setting control dam longevity, failure mechanisms, and flood magnitude. The world's largest floods have resulted from the failure of glacier dams during the Late Quaternary. Man-made dams are also subject to failure, with the capability of releasing catastrophic floods many times greater in magnitude than could physically be generated by natural processes within the same catchment (Figure 5).

Accumulations of vegetation debris or coarse woody debris within river channels can obstruct flood flows, temporarily and partially damming floodwater. Infrequent high-magnitude flows within ephemeral rivers, for example, would allow sufficient time for coarse woody debris to accumulate within the channel between the highest magnitude events. Sudden dam failure commonly generates an enhanced flood peak as water released from storage is superimposed on the main flood hydrograph.

Although snow accumulation with river channels is a common occurrence in many parts of the world, there are relatively few reports of flooding produced by snow-dam failure. Outburst floods have been described from snow-dammed lakes in Antarctica and nival flood peaks enhanced by small snow-dammed lake outbursts in Canada. Snow damming mainly affects small low-order streams and can only pond water for very short time periods.

River ice break-up during the spring melt can lead to ice-jams at flow constrictions, resulting in considerable upstream inundation as well as downstream flooding on ice-jam failure. Unlike the snow-dammed lake bursts, ice-jam-related flooding occurs on major

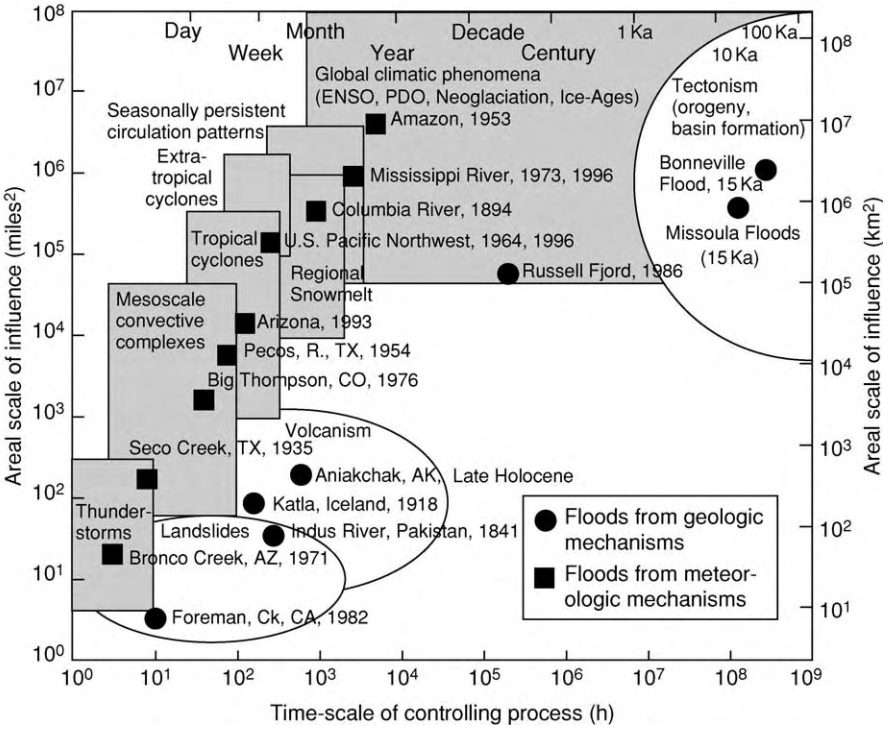


Figure 2 Space and time domain for flood producing hydroclimatological conditions and flooding. After Hirschboeck (1988) Flood hydroclimatology. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 27 49, London: Wiley; and O'Connor *et al.* (2002) The geology and geography of floods. *Water Science and Application* 5: 359 385.

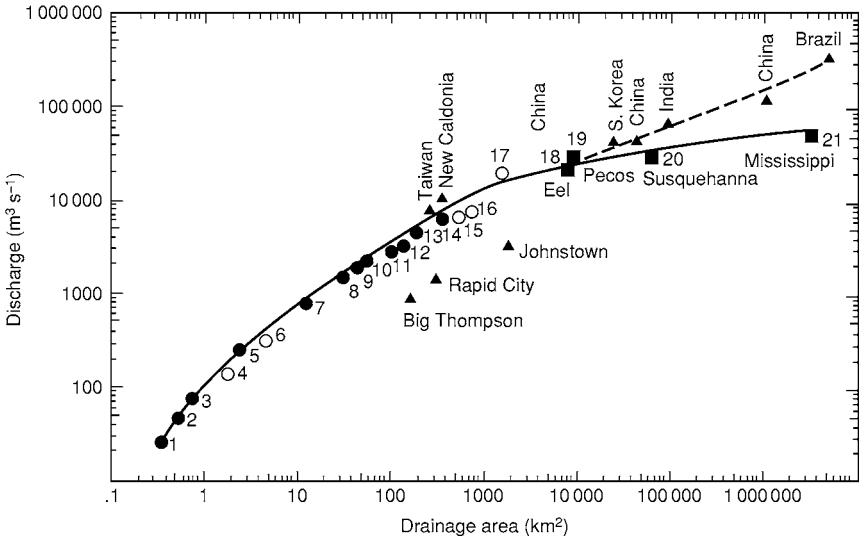


Figure 3 Envelope curve for maximum discharge generation in relation to catchment area. After Hirschboeck KK (1987) Cata strophic flooding and atmospheric circulation anomolies. In: Mayer LD and Nasch D (eds.) *Catastrophic Flooding* pp. 23 56. Allen & Unwin.

river systems. Within the northern hemisphere, northward draining rivers are particularly prone to ice-jams as the spring melt is initiated within the catchment headwaters before the lower reaches have thawed.

Within glaciated catchments, ice-cliff or subglacial tunnel collapse can temporarily dam meltwater rivers with a breccia of ice blocks. Damming commonly lasts only for periods of up to one hour, as meltwater flows are easily able to penetrate, float, and bulldoze

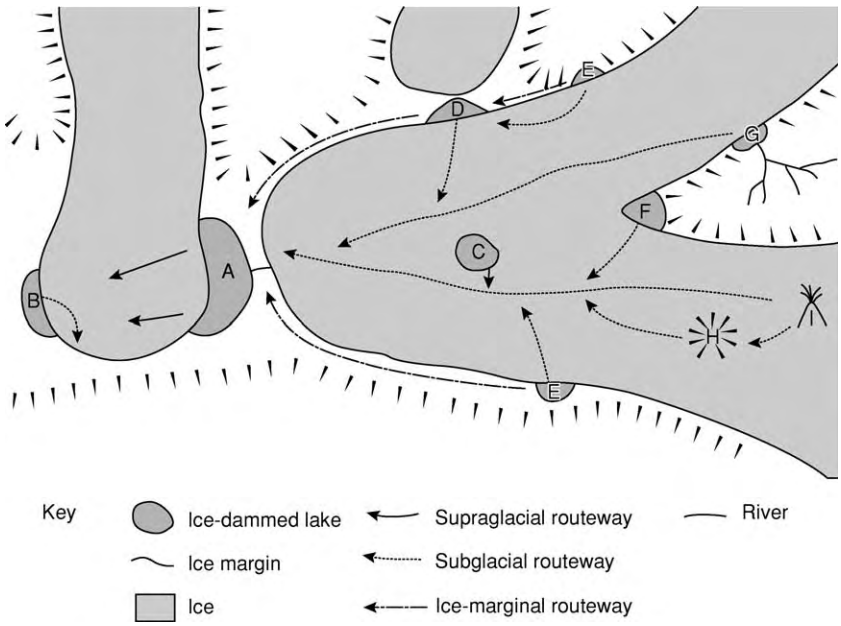


Figure 4 Ice dammed lake storage locations. Letters A I represent different meltwater storage locations, each associated with characteristic lake drainage triggers and jokulhlaup routing characterics. 'A' represents a tributary glacier which dams a trunk valley and which is subject to subaerial breach widenin. 'B' 'D' 'E' 'F' and 'G' represent proglacial and ice marginal ice dammed lakes in a variety of settings where drainage is triggered by mechanisms such as lake overspill, ice dam flotation, syphoning and subglacial cavity formation. 'C' represents a supraglacial lake which taps through the entire thickness of the glacier. 'H' is a subglacial lake within either a suitable storage location such as a volcanic caldera, or as an up turned bowl shaped lake often associated with locations with high geothermal acitivity. Drainage triggers postulated for subglacial lakes include ice dam flotation as well as Darcian flow through the glacier substrate. 'I' represents a subglacial eruption site, which may result in little or short lived meltwater storage during the course of an eruption. High meltwater temperatures associated with the drainage of geothermal and volcanic events characteristically result in jokulhlaups with rapid rising stages. Jokulhlaup hydrographs have a longer rising stage and lower peak discharge the further the ice dammed lake is from the glacier snout. After Tweed and Russell (1999) Controls on the formation and sudden drainage of glacier impounded lakes: implications for jokulhlaup characteristics. *Progress in Physical Geography* 23: 79 110.

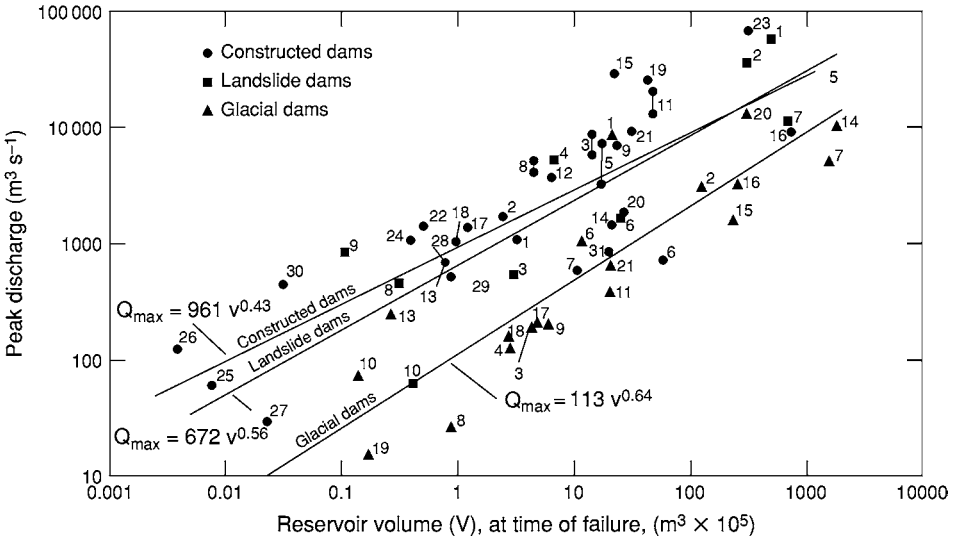


Figure 5 Relationships between lake volume and peak discharge for the failure of different dam types. After Costa (1988) Rheologic, geomorphic and sedimentological differentiation of water floods, hyperconcentrated flows, and debris flows. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 113 122. London: Wiley.

the breccia dam. Such events have been reported from locations such as the Alps, Greenland, and the Canadian Arctic and commonly result in floods with peak discharges up to an order of magnitude higher than 'normal' diurnal meltwater flows. These floods occur within either ice-marginal river systems or englacial tunnels, and they are more frequent earlier in the melt season when increasing meltwater discharges undermine and remove glacier ice from winter advance positions.

Glaciers are able to store water within ice-dammed lakes over time-periods of 10^1–10^2 years, generating glacier outburst floods or 'jökulhlaups' when they drain. Lakes can form in a number of locations: beneath, within, on top of, and at the edge of glaciers (**Figure 4**). Subglacial lakes can exist either within pre-existing topographic depressions, such as volcanic calderas and valleys, or within up-turned bowl-shaped lakes protruding upwards into the glacier. Such subglacial lakes are common in Iceland where high geothermal heat fluxes and frequent volcanic activity are common. In Iceland, subglacial Lake Grímsvötn drains every few years as meltwater takes time to accumulate, raising the lake to a level where drainage will occur. In 1996, rapid influx of meltwater from the nearby Gjálp subglacial eruption site resulted in the filling of Grímsvötn within one month. Approximately 3.4 km^3 of meltwater drained subglacially from the lake as a jökulhlaup, which lasted for three days.

Although Grímsvötn is one of the largest modern subglacial lakes known to drain, it is dwarfed by Late Quaternary jökulhlaups from glacial Lake Missoula, which drained over 2000 km^3 of water within just a few days. The release of water from sub- or englacial pockets has also been invoked as an explanation for sudden outburst floods or 'debacles' from Alpine glaciers. Large supraglacial lakes are also known from modern glaciers and ice-sheets and can involve the drainage of up to ~0.1 km^3 of meltwater subglacially over distances of tens of kilometres. Ice-marginal, ice-dammed lakes are relatively common at modern glaciers, and the presence of shorelines, deltas, and lacustrine deposits provide widespread evidence for the existence of large ice-marginal lakes associated with Palaeo ice-margins. Large proglacial lakes formed during the deglaciation of North America, and drained to the North Atlantic and Arctic oceans via a series of enormous spillway channels. For example, glacial Lake Agassiz is known to have drained 160 000 km^3 of meltwater within only a few years.

Although ice-dammed lakes occupy a variety of locations relative to the impounding glacier, the most significant control on jökulhlaup characteristics is whether the glacier enters a main valley as an outlet and has the potential to pond a lake within the main valley upstream of the glacier dam. In this case, the ice-dammed lake volume would be much larger than that of ice-marginal, ice-dammed lakes relative to the glacier dam. A number of mechanisms to account for the drainage of ice-dammed lakes have been proposed. In general terms, lake drainage will take place if there is a hydraulic gradient allowing lake water to flow into the glacier. However, a connection must also exist between the lake and the glacier for drainage to commence. In most cases ice-dammed lake drainage leaves the glacier intact, but complete ice-dam removal may take place where the lake volume is large relative to ice dam volume. The timing of ice-dammed lake drainage is frequently controlled by the retreat and advance of glaciers. Glacier thinning and retreat often allow the initiation of cycles of lake drainage (**Figure 6**). Jökulhlaup magnitude decreases during each cycle as impounded lake volume decreases during glacier retreat (**Figure 6**).

Glacier retreat within high mountain regions often creates highly unstable moraine-dammed lakes, which are prone to sudden failure and flood generation. Calving of glacier ice into these lakes can generate waves, which overtop and then incise the moraine dam, allowing the extremely rapid development of a breach, facilitating catastrophic evacuation of lake water. Melt out of buried ice within moraine dams may take decades, and it is common for moraine-dammed lakes to exist for considerable periods before drainage. Once a moraine dam has been breached, only a subsequent glacier advance can reinstate the dam.

Landslides can generate floods in two ways. Firstly, rapid input of landslide debris to lakes can generate displacement waves. The best-known example of such a flood was at the Vaiont reservoir in Italy, where a landslide generated a large displacement wave that washed up a hillside and over the dam. In Iceland, a rockfall generated a displacement wave known as the Steinholtshlaup, which constituted one of the most rapid flood-rising stages experienced in Iceland. Landslides frequently dam major valleys allowing the formation of large lakes. Dam failure may occur immediately after formation, or may be delayed, possibly for decades, until reservoir volume has built up to a sufficient level for drainage to occur.

Bedrock and sediment dams are also known to fail via catastrophic enlargement of their spillways, leading to large outbursts. The drainage of Lake Bonneville in the USA during the Late Quaternary generated an outburst with a peak discharge of approximately 1×10^6 m^3 s^{-1}, whilst draining 4800 km^3 of water over several months. In New Zealand the

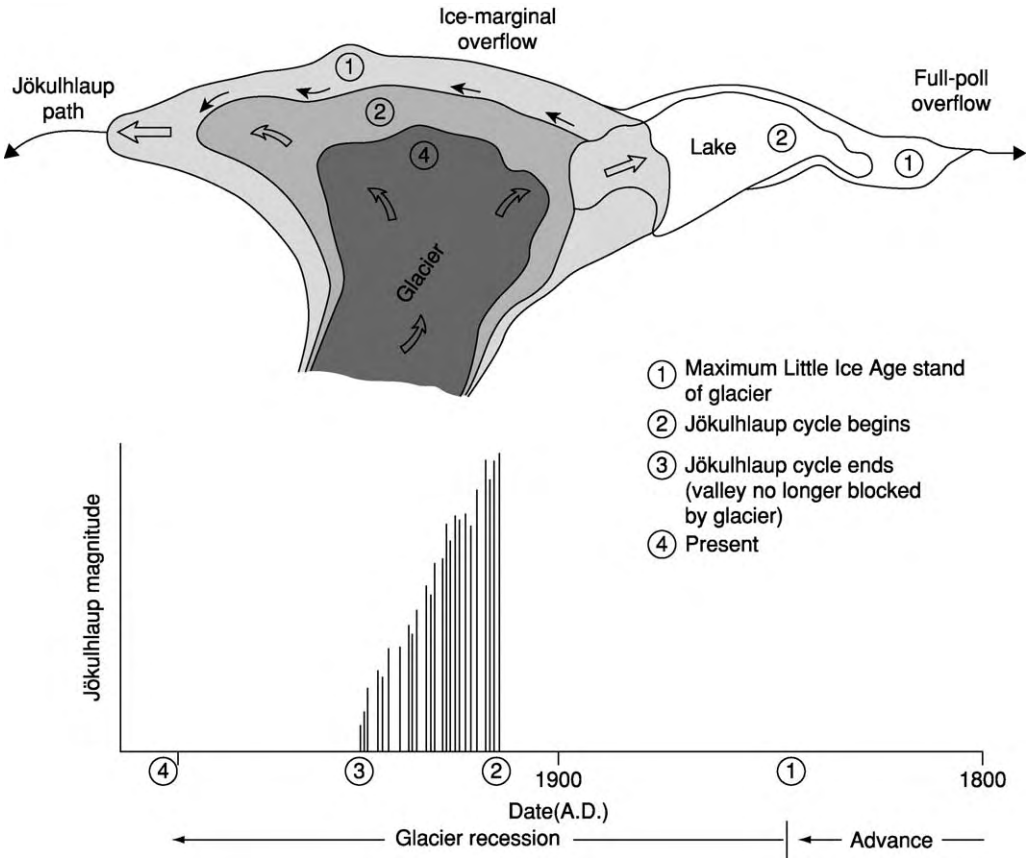


Figure 6 The jokulhlaup cycle. The upper diagram illustrates the advance and retreat of a glacier during the Little Ice Age. Ice positions 17 4 in the upper diagram correspond with the timescale illustrates in the lower diagram. The cycle of jokulhlaup drainage occurs as the glacier is thinning and retreating from its advance. Jokulhlaup magnitude within the cycle decreases as ice dammed lake volumes decrease progressively during glacier retreat. After Evans SG and Clague JJ (1994) Recent climatic change and catastrophic geomorphic processes in mountain environments. *Geomorphology* 10: 107 128.

failure of caldera Lake Taupo generated a peak discharge of $1.7–3.5 \times 10^5\,m^3\,s^{-1}$, draining a total of $20\,km^3$ of water. The failure in Alaska of Aniakchak caldera ($3.7\,km^3$) generated a peak discharge of $1 \times 10^6\,m^3\,s^{-1}$. Deposition of alluvial fans at tributary junctions within ephemeral systems can block subsequent flood flows, resulting in enhanced flood peaks due to sudden dam failure.

Catastrophic Flood Characteristics

The discharge hydrograph is the most important characteristic of any flood as it defines flood duration and magnitude. Flood magnitude can be described by the peak discharge and total flood volume. Hydrograph shape reflects flood generation mechanisms and drainage basin characteristics. Different flood generation mechanisms result in different hydrograph types (Figure 7). Dam failure due to rapid breach development generates very rapid rises to peak discharge (Figure 7A). Volcanically generated (volcanoglacial) jökulhlaups also produce extremely rapid rises to peak discharge due to extremely high ice melt rates and glacier hydrofracturing. Rainfall-generated floods in ephemeral river-systems also display rapid rises to peak discharge followed by more sedate waning flow stages. Many jökulhlaup hydrographs exhibit gradual rising stages characterized by an exponential shape (Figure 7B). These jökulhlaup hydrographs are controlled by a positive feedback between meltwater flow and tunnel enlargement (Figure 7B and 7E). The rate of exponential discharge increase is controlled by ice-dammed lake water temperature, the length of the tunnel routeway, and height difference between the lake surface and the outlet. Flood hydrographs with multiple peaks can represent either nonsynchronous response of various subcatchments, complex channel networks to runoff, or temporary channel blockage and subsequent release. For volcanoglacial jökulhlaups, multiple peaks may be generated by peaks in volcanic activity. Similarly 'heartbeat' events associated with temporary tunnel blockage are characterized by a sudden fall and then rise in discharge (Figure 7C). Sudden onset

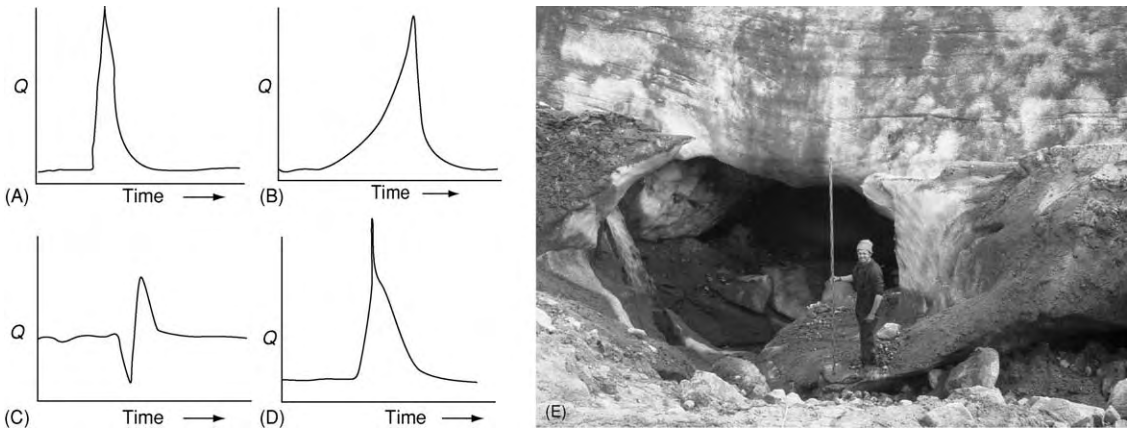


Figure 7 Flood hydrographs with contrasting shapes. (A) Rapid rise to peak discharge associated with sudden dam failure and glacier outburst floods associated with sheet flow and hydrofracturing. (B) Exponential rise to peak discharge is typical of floods where rising stage is controlled by progressive enlargement of ice tunnels. (C) 'Heartbeat' event where temporary channel blockage results in a reduction in discharge followed immediately by a flood peak associated with dam failure. (D) Composite flood hydrograph characterized by the superimposition of one flood wave on top of another. (E) Jokulhlaup tunnel inlet generated by the drainage of a small ice dammed lake in West Greenland.

jökulhlaups may fill ice-marginal basins within minutes, creating short-lived ice-dammed lakes, which in turn drain to produce a composite flood hydrograph where one flood peak is superimposed upon another (Figure 7D).

Catastrophic floods are often characterized by a dearth of direct process measurements. Most hydraulic variables are extrapolated from geomorphological and sedimentary evidence using hydraulic modelling techniques developed for lower magnitude flows. Calculated velocities are often extremely high ($<40\,m\,s^{-1}$), often an order of magnitude greater than that for nonflood flows. Flow depths in excess of 500 m have been reconstructed for the largest floods on earth. Flood flows commonly vary between sub- and supercritical states. Shear stress and unit stream power provide useful indicators of a flood's capacity to erode channel boundaries and to transport sediment. Flows are commonly macroturbulent, and involve considerable aeration of the flow (Figure 8A). Large-scale roughness generates tornado-like vortices known as 'kolks', which have the capability to pluck large fragments from the bedrock surface. High-flow velocities and violent flow separation can generate local instantaneous reductions in the vapour pressure of water, allowing the formation of water vapour bubbles within the flow. Violent collapse of the vapour bubbles releases large amounts of energy, enabling the pitting of bedrock surfaces. Flood flows commonly superelevate towards the outside of channel bends and exhibit cambered surface profiles (Figure 8B). In many cases it is very difficult to define the flood surface due to unsteady and macroturbulent flow conditions (Figure 8A and 8B). Flood peak discharges commonly attenuate downstream due to resistance effects and transmission losses within some systems. Floods and displacement waves with high peak discharges and relatively modest volumes show the highest rates of downstream attenuation.

Figure 8 (A) Macroturbulent flow during a jokulhlaup in West Greenland. (B) Superelevated flood flow during a jokulhlaup in West Greenland.

High flood powers and bed shear stresses allow grain sizes that would normally be transported as bed load to be transported as either suspended or

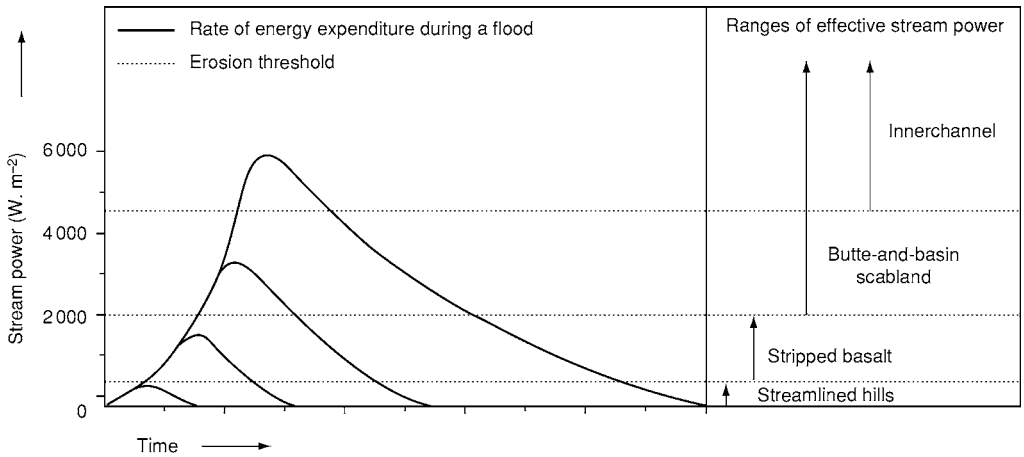


Figure 9 Progressive increase in stream power for various erosional channel morphologies in the Columbia River gorge. After Benito (1997) Energy expenditure and geomorphic work of the cataclysmic Missoula flooding in the Columbia River gorge, USA. *Earth Surface Processes and Landforms* 22: 457 472.

Figure 10 (A) Cataract erosion in the Jokulsá á Fjollum system, northern Iceland. (B) Erosional bedforms sculpted by fluvially driven abrasion in high resistant bedrock, Kangerlus suaq, West Greenland.

wash load within the flow. A jökulhlaup in Greenland with a discharge of 1100 $m^3 s^{-1}$ and powers of 3000 $W m^{-2}$ transported sediment of up to 10 cm in diameter in suspension, whilst the November 1996 jökulhlaup in Iceland with a peak discharge of 45–53 000 $m^3 s^{-1}$ and a power of 40 000 $W m^{-2}$ was capable of moving 2-m-diameter boulders in suspension. These modern jökulhlaups only maintained such high stream powers locally and for a short duration. Giant Quaternary outbursts in the Altai Mountains of Siberia and from glacial Lake Missoula have generated flood powers of 10^5–10^6 $W m^{-2}$ over larger areas.

As erosional and sediment transport capacities are so high during floods, large amounts of sediment can alter flow conditions. Typical alterations include reduction of resistance to flow as sediment in transport blankets roughness elements, reductions of fall velocities of sediment in transport, and transition to upper flow regime bed conditions.

Sediment concentrations often exceed the values normally associated with fluidal flow conditions and become hyperconcentrated. Sediment bulking during floods may transform floods into debris flows, which have their own strength. Debris flow deposits directly reflect the entire flow thickness in contrast to water flood deposits, which only represent what was deposited at the bed of the flow. Sediment concentrations within catastrophic floods may also increase towards the base of the flow, leading to hyperconcentrated near-bed conditions, often referred to as a traction carpet. This may account for the deposition of units more typical of hyperconcentrated flow conditions by seemingly turbulent high-energy fluvial flows.

High sediment concentrations within floods can only be achieved if there is sufficient sediment readily available for entrainment. Sediment availability to catastrophic floods varies, depending upon flood generation mechanism, lithology, and weathering environment. Sediment bulking is usually most efficient within catchment headwaters, where precipitation inputs are high and stream and hillslope systems are

Figure 11 Subglacially eroded rip up clast deposited by the November 1996 jokulhlaup near to the glacier margin.

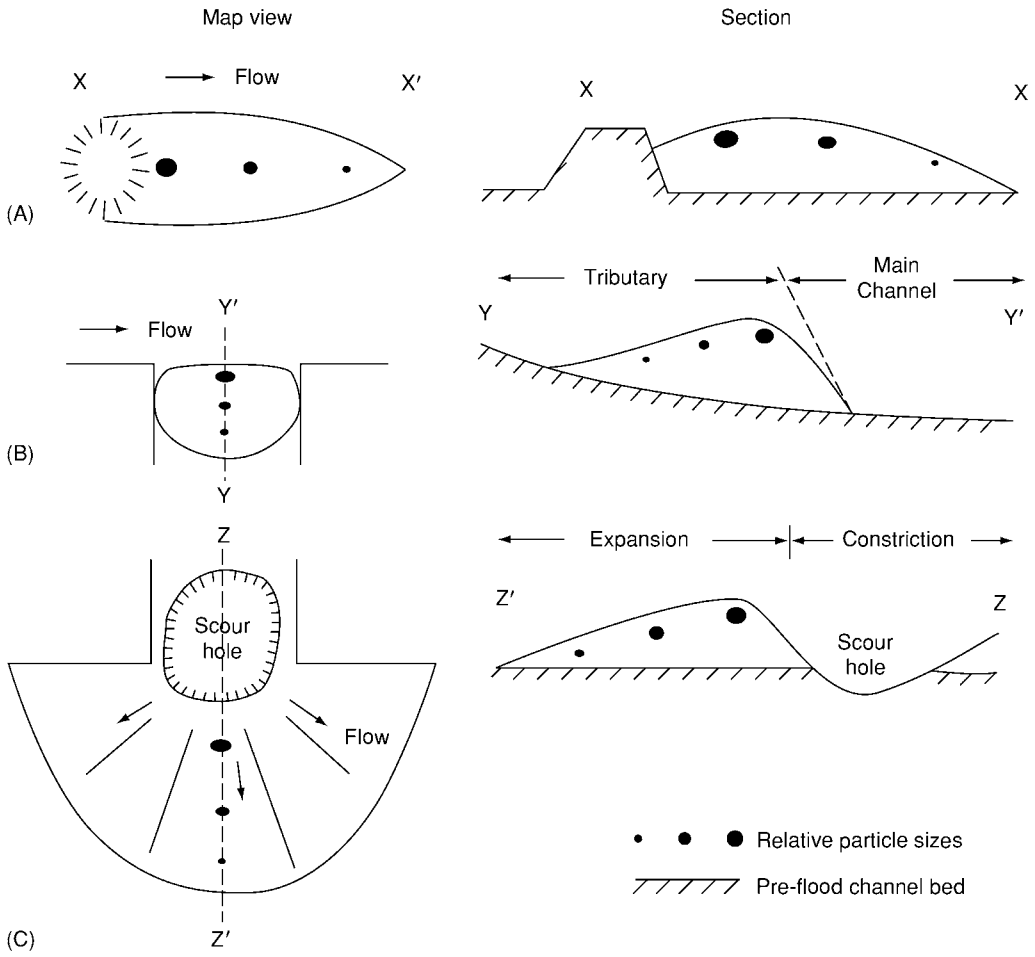


Figure 12 Role of channel geometry in controlling the morphology and sedimentology of flood deposits. Adapted from Baker (1984) Flood sediments in bedrock fluvial systems. In: Koster EH and Steel RJ (eds.) *Sedimentology of Gravels and Conglomerates*. Canadian Society of Petroleum Geologists Memoir 10, pp. 87 98.

coupled. In many cases, hillslope failure may transform directly into debris flows. Some floods may acquire large volumes of sediment at source. Jökulhlaups associated with subglacial volcanic eruptions contain large volumes of primary eruption products, and dam-failure outbursts are often heavily freighted with sediment eroded from the dam breach. Sediment concentration during nonvolcanic glacier outburst floods is usually supply limited, as large volumes of water dilute a finite volume of available sediment. Catastrophic floods can also bulk-up by eroding their channel boundaries.

Sediment transport during catastrophic floods exceeds nonflood transport rates by orders of magnitude. Such flows are able to access sediment sources beyond the reach of lower magnitude flows and are capable of exceeding thresholds for bedrock erosion. However, infrequent high-magnitude events may not be the dominant transport mechanism in catchments dominated by high productivity of fine-grained sediment.

Geomorphic and Sedimentary Impact

Flood impacts can be either erosional or depositional, with the occurrence, distribution, and character of specific impacts controlled by the spatial and temporal distribution of stream power. Erosional landforms are generated when flood power exceeds channel resistance to erosion. Flood flows not only have to exceed local erosional thresholds, but must persist long enough to generate distinctive suites of erosional landforms. Repeated Late Quaternary outburst floods from glacial Lake Missoula have created the world's most spectacular erosional landscape. Within the Columbia gorge system erosional landforms have been associated with various stream power ranges. Only the largest of the Missoula floods had sufficient power and duration to develop the entire suite of erosional features: streamlined hills, stripped basalt, butte and basin scabland, and inner channels, in order of ascending stream power (Figure 9). Although landscapes of flood erosion owe their detailed morphology to the interaction between hydraulic processes and local bedrock strength and structure, bedrock channels exhibit numerous common characteristics. Cataracts are common where flood flows excavate a more resistant cap-rock, allowing for rapid headward migration of a plunge pool (Figure 10A). Large-scale potholes are common where deep flood flows allow kolks to form. Where bedrock surfaces are stable, they can be abraded by sediment in transport to produce a range of erosional bedforms (Figure 10B). The widespread distribution of bedrock bedforms in parts of North America has been interpreted as the product of catastrophic subglacial sheet floods. Large channels and lake overspills containing large-scale and widespread erosional evidence are common in many

Figure 13 (A) Oblique aerial photograph of valley confined outwash plain subject to jokulhlaups showing proximal expansion bars. (B) Photograph of 1987 jokulhlaup flows across a outwash plain in West Greenland 1987. Note high energy proximal flows expanding from a flow constriction and relatively tranquil flows in the distal area related to backwater ponding upstream of a flow constriction.

parts of the world subject to catastrophic floods. As well as eroding bedrock, floods have the potential to incise unconsolidated sediment. The November 1996 jökulhlaup in Iceland excavated large volumes of subglacial sediment, providing enormous numbers of rip-up clasts to the ice proximal outwash plain (**Figures 11** and **15B**). Large buried tunnel valleys in former glaciated regions have been attributed to Quaternary catastrophic floods. Large tracts of the channelled scablands inundated by the Late Quaternary Missoula outburst floods have eroded into loess, generating large-scale streamlined islands.

Catastrophic flood deposition is controlled by spatial and temporal variations in stream power. Downstream variations in flood channel morphology control stream power variations, resulting in zones of erosion, nondeposition, or deposition. Flood deposition within confined bedrock channels is limited either to the flood-waning stage or to specific locations such as the mouths of tributary channels, embayments, or the lee of obstacles to flow (**Figure 12**). Large bars radiate downstream from flow expansions and commonly surround a scour hollow or zone of nondeposition. Boulder ridges or 'berms' are also frequently deposited at sudden flow expansions at the boundary between the main and secondary flows (**Figures 12** and **13A**). Flow constrictions commonly generate considerable backwater effects, which build up progressively during the rising stage and subside during the waning stage (**Figure 13B**). Deposition takes place into the backwater zone under progressively lower energy flow conditions (**Figure 13A** and **13B**). High-stage backwater deposits are reworked by higher energy waning stage flows. If during the flood the channel constriction is widened by erosion, the backwater effect will reduce substantially and subsequent backwater effects will be reduced. Unconfined flood flows tend to be shallower with high rates of flow dissipation consequently, resulting in thinner flood units.

Although catastrophic flood deposits are commonly identified by the presence of anomalously large clasts, indicative of high sediment transport capacities, this only applies where coarse-grained sediment is available for entrainment. However, for many fluvial systems coarse-grained sediment is supply limited. Other criteria are therefore needed to characterize flood deposits. Landforms and deposits may be interpreted to be of flood in origin if they are sufficiently large in scale relative to those associated with normal flows within the same fluvial system. Giant flood bars often display large-scale cross-stratification, which could easily be misinterpreted as delta foreset beds. Flood units are usually laterally extensive and contain a relatively uniform suite of clast lithologies compared with nonflood units. Flood deposits often contain rip-up clasts composed of blocks of sediment eroded from the river bed or banks (**Figures 11** and **15B**). Mud clasts are common within ephemeral river systems where silts and clays are ripped up during the flood-rising stage. Glacier proximal jökulhlaup deposits commonly contain rip-ups derived from subglacial fluvial excavation, which comprise sheared and folded subglacial diamicton or older outwash sediments (**Figures 11** and **15B**). Large-scale ice block obstacle marks and their associated secondary melt-out structures are diagnostic of high-magnitude glacier outburst floods (**Figure 14**). Simultaneous progradation, aggradation, and backfilling within the ice-marginal zone are

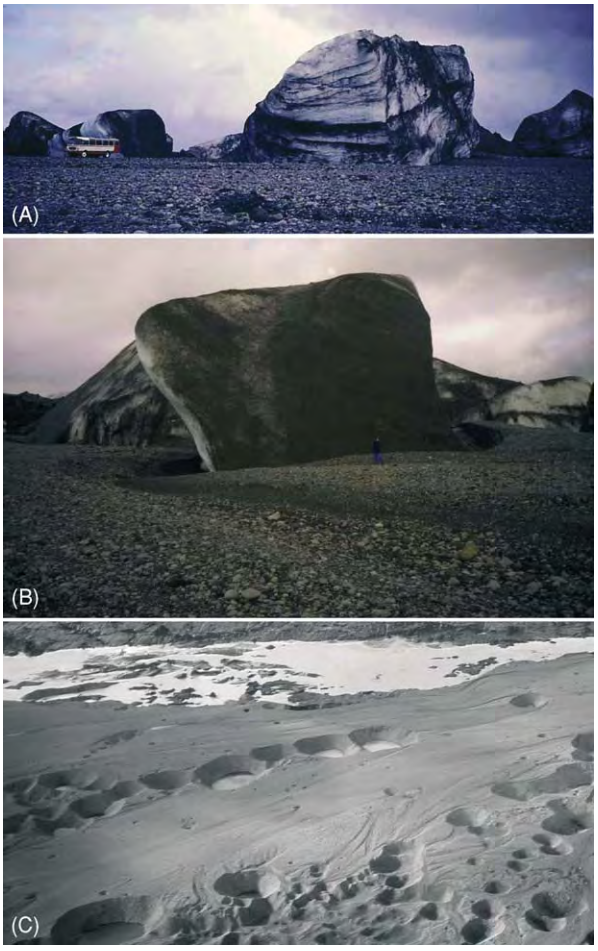


Figure 14 (A) Large ice block transported over 1 km by the November 1996 jokulhlaup on Skeiðarársandur, Iceland. (B) Deeply embedded ice block acts as an obstacle to the flow generating lateral scour channel and a leeside 'tail' of sediment. (C) Aerial view of kettle holes and obstacle marks four years after the jokulhlaup.

characteristic of high-magnitude events, capable of simultaneous channel erosion and infill.

Catastrophic flood units vary in sorting, texture, and matrix content, depending upon the size distribution of available sediment as well as its rate and mode of emplacement. Traction deposition generally results in very poorly sorted, cohesionless, clast-supported deposits, which commonly display poorly defined imbrication (Figure 15A). Stream powers calculated for many catastrophic floods are high enough to carry boulder-sized sediment in suspension. Rapid rainout and deposition from suspension at the mouth of tributary valleys commonly results in more matrix-rich deposits, which are typical of proximal eddy bar deposition (Figure 15B). More distal suspension sedimentation forms stacked upward-fining sequences similar to turbidites within lacustrine and marine environments (Figure 15C). Coarsening-upward successions are commonly deposited on the rising flood stage, reflecting deposition from transport-limited flood flows (Figure 15D).

Floods commonly access large volumes of sediment, allowing their transformation from water flood to hyperconcentrated flows (Figure 16). The presence of laterally extensive deposits consistent with deposition from hyperconcentrated flows indicates catastrophic floods. Although debris-rich floods may be commonplace within alluvial fan and ephemeral systems, within most river systems only exceptional floods have the capacity to generate hyperconcentrated flow conditions. Bedforms within fluvial systems dominated by repeated floods contain reactivation surfaces, which indicate progressive deposition during a series of high-magnitude floods interrupted by interflood pauses. Reactivation surfaces can also be identified by dateable palaeosols and tephra horizons, which can help in the reconstruction of Palaeo flood magnitude and frequency.

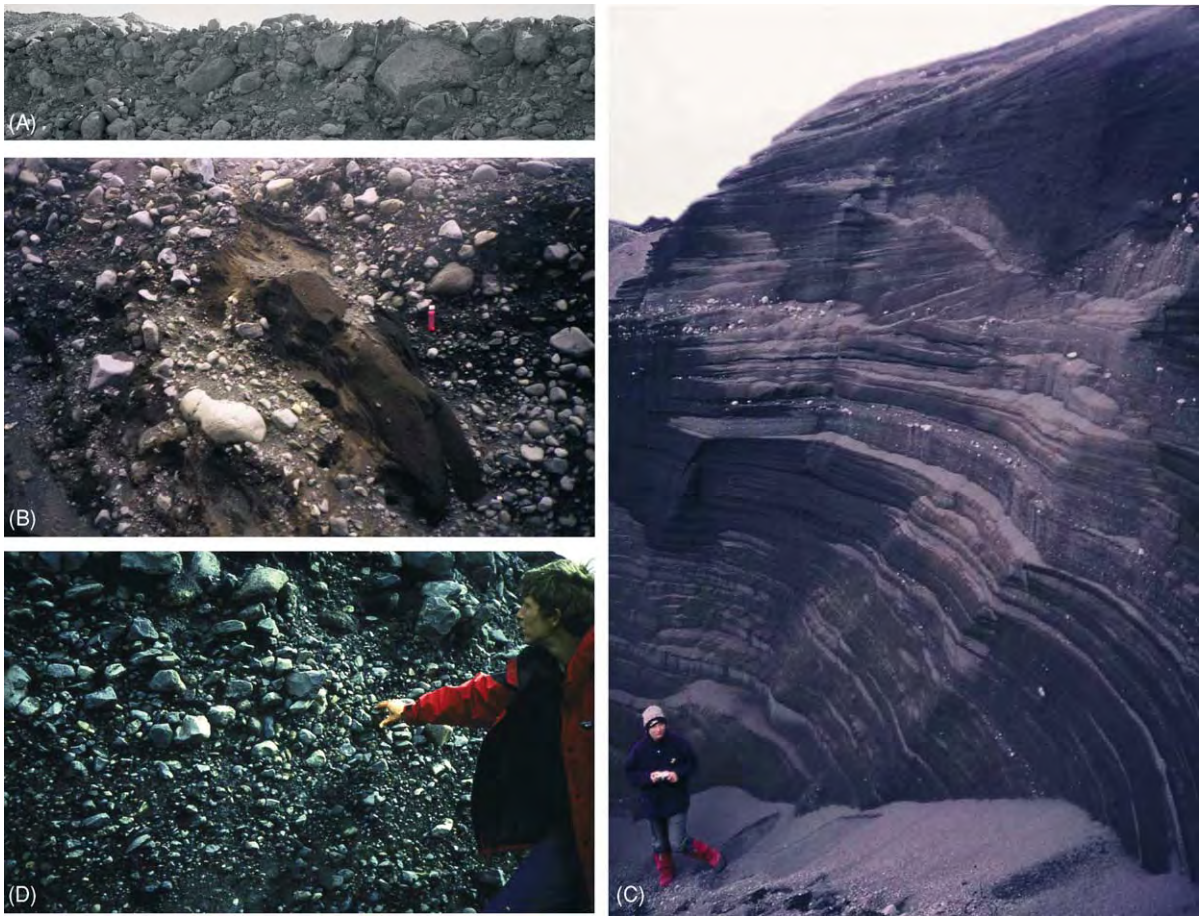


Figure 15 (A) Flood deposits from traction load; (B) proximal eddy bar deposits displaying noncohesive poorly sorted, matrix supported deposits surrounding a large rip up clast; (C) rhythmically deposited fine gravel beds; (D) photograph of a coarsening upward sequence reflecting rising stage deposition.

Figure 16 Massive fine grained unit on Sólheimasandur, Iceland interpreted as the product of a hyperconcentrated jokulhlaup.

Catastrophic floods deliver large volumes of sediment to the marine environment where they generate turbidity currents capable of transporting sediment offshore for up to 10^3 km. Turbidite units of up to 60 m in thickness identified in the offshore sedimentary record of the north-eastern Pacific Ocean have been attributed to outbursts from glacial Lake Missoula. In Iceland, volcanogenic jökulhlaups play an important role in the supply and transport of sediment to southern Iceland shelf and adjacent marine basins. Giant, sediment-rich jökulhlaups from Katla subglacial volcano dominate the offshore sedimentary record along most of Iceland's southern coast.

Controls on Catastrophic Flood Impact

Flood generation mechanisms, sediment availability, and routeway characteristics are the main controls on flood impact (Figure 1). The long-term impact of any flood may, however, depend upon the ability of more frequent, lower magnitude events to rework flood deposits, thereby allowing channel recovery. If flood recurrence interval is greater than the time required for recovery, then floods may be considered to have not had a long-term impact. Conversely, if flood recurrence interval is less than recovery time, then the channel will not be able to recover from the impact of one flood by the time the next flood occurs. Some flood impacts may, however, be regarded as being irreversible, in instances where high flood sediment calibre makes between-flood reworking impossible or where bedrock erosion occurs. Large-scale flood excavation of bedrock channels, such as those associated with glacial Lake Missoula outbursts, represent a long-lasting impact. Similarly, slackwater deposits associated with extremely high-magnitude flows will also have a high preservation potential if intervening flows with tributary streams are incapable of reworking flood deposits.

Conclusions

Catastrophic floods can be generated by a range of mechanisms within a number of hydroclimatic environments. Although the impacts of floods can be defined in a number of ways it is crucial that we assess carefully whether a flood is truly catastrophic. Continental-scale catastrophic flood landscapes can be identified in association with the growth and decay of ice-sheets during repeated glacial and interglacial cycles. Some floods were large enough to influence ocean circulation, thereby impacting on the climate system. Better understanding of the causes, characteristics, and impacts of catastrophic floods is crucial for the identification and mitigation of hazards and for interpretation of the sedimentary record.

See Also

Biblical Geology. **Engineering Geology:** Natural and Anthropogenic Geohazards. **Sedimentary Environments:** Depositional Systems and Facies; Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Fluvial Geomorphology; Glaciers; Particle-Driven Subaqueous Gravity Processes. **Sedimentary Rocks:** Rudaceous Rocks. **Unidirectional Aqueous Flow**. **Volcanoes**.

Further Reading

Baker VR (1973) Paleohydrology and sedimentology of Lake Missoula flooding in Eastern Washington. Geological Society of America Special Paper 144.

Baker VR (1984) Flood sediments in bedrock fluvial systems. In: Koster EH and Steel RJ (eds.) *Sedimentology of Gravels and Conglomerates*. Canadian Society of Petroleum Geologists Memoir 10, pp. 87 98.

Baker VR and Costa JE (1987) Flood power. In: Mayer L and Nash D (eds.) *Catastrophic Flooding*, pp. 1 25. London: Allen and Unwin.

Baker VR, Benito G, and Rudoy AN (1993) Palaeohydrology of Late Pleistocene superflooding, Altay Mountains, Siberia. *Science* 259: 348 350.

Benito G (1997) Energy expenditure and geomorphic work of the cataclysmic Missoula flooding in the Columbia river gorge, USA. *Earth Surface Processes and Landforms* 22: 457 472.

Björnsson H (1992) Jökulhlaups in Iceland: Prediction, characteristics and simulation. *Annals of Glaciology* 16: 95 106.

Björnsson H (2002) Subglacial lakes and jökulhlaups in Iceland. *Global and Planetary Change* 35: 255 271.

Bull WB (1988) Floods degradation and aggradation. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 157 165. London: Wiley.

Church M (1988) Floods in cold climates. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 205 229. London: Wiley.

Clague JJ and Evans SG (2000) A review of catastrophic drainage of moraine dammed lakes in British Columbia. *Quaternary Science Reviews* 19: 1763 1783.

Costa JE (1988) Rheologic, geomorphic and sedimentological differentiation of water floods, hyperconcentrated flows, and debris flows. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 113 122. London: Wiley.

Costa JT and Schuster RL (1988) The formation and failure of natural dams. *Geological Society of America Bulletin* 100: 1054 1068.

Evans SG and Clague JJ (1994) Recent climatic change and catastrophic geomorphic processes in mountain environments. *Geomorphology* 10: 107 128.

Fisher TG, Smith DG, and Andrews JT (2002) Preboreal oscillation caused by a glacial Lake Agassiz flood. *Quaternary Science Reviews* 21: 873 878.

Hirschboeck KK (1988) Flood hydroclimatology. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 27 49. London: Wiley.

Hirschboeck KK (1987) Catastrophic flooding and atmospheric circulation anomolies. In: Mayer LD and Nasch D (eds.) *Catastrophic Flooding*, pp. 23 56. Allen & Unwin.

Kochel RC (1988) Geomorphic impact of large floods: review and new perspectives on magnitude and frequency. In: Baker VR, Kochel RC, and Patton PC (eds.) *Flood Geomorphology*, pp. 169 187. London: Wiley.

O'Connor JE (1993) *Hydrology, hydraulics, and geomorphology of the Bonneville Flood*. Geological Society of America Special Paper 274.

O'Connor JE, Grant GE, and Costa JE (2002) The geology and geography of floods. *Water Science and Application* 5: 359 385.

Shaw J (2002) The meltwater hypothesis for subglacial bedforms. *Quaternary International* 90: 5 22.

Tweed FS and Russell AJ (1999) Controls on the formation and sudden drainage of glacier impounded lakes: implications for jökulhlaup characteristics. *Progress in Physical Geography* 23: 79 110.

Walder JS and Costa JE (1996) Outburst floods from glacier dammed lakes: The effect of mode of drainage on flood magnitude. *Earth Surface Processes and Landforms* 21: 701 723.

Ward R (1978) *Floods*. Macmillan: London.

Wolman MG and Gerson R (1978) Relative scales of time and effectiveness of climate in watershed geomorphology. *Earth Surface Processes and Landforms* 3: 189 203.

Wolman MG and Miller JP (1960) Magnitude and frequency of forces in geomorphic processes. *Journal of Geology* 68: 54 74.

Deep Water Processes and Deposits

D J W Piper, Geological Survey of Canada, Dartmouth, NS, Canada

Introduction

Deep water beyond the continental shelves covers 63% of the Earth's surface. About one-seventh of this deep-water area, on the continental margins, receives a direct supply of terrigenous sediment from the continents, and this accounts for the majority of deep-ocean sediment deposits, including the largest Quaternary sediment accumulations on the planet, which occur beneath continental slopes and on submarine fans (Figures 1 and 2). Beyond the continental margins, the pelagic realm slowly accumulates biogenic material and far-travelled clays (Figure 3). The deep oceans are primarily a depositional environment, and that deposition is strongly influenced by ocean-floor morphology. Deep trenches at subduction zones and deep abyssal plains at passive continental

margins trap terrigenous sediment, preventing turbidity currents from reaching the greater part of the ocean floor. The large-scale morphology of the ocean is controlled principally by tectonic processes: seafloor spreading and its associated thermal effects, faulting, and volcanism; the complex rifting of continents in the early stages of ocean formation; and the even more complex effects of subduction and eventual closure of oceans. This article reviews the processes that lead to sediment distribution in the ocean and the large-scale distribution of these sediment deposits.

Technological advances in swath bathymetry in the past decades have greatly improved our understanding of the morphology of the seafloor. Given the vastness of the ocean and the technological challenges involved in monitoring seafloor processes, morphology is an important guide to the processes acting on the seafloor. Seismic-reflection profiling is an essential tool for evaluating the distribution of seafloor facies and hence inferring processes: the growing availability of three-dimensional seismic has revolutionized our understanding of continental-margin sedimentation. Direct sampling of deep-water sediments is limited to piston cores, which rarely penetrate more than 20 m, and some 1000 scientific boreholes, which are hundreds of metres long, drilled by the Ocean Drilling Program and its predecessors.

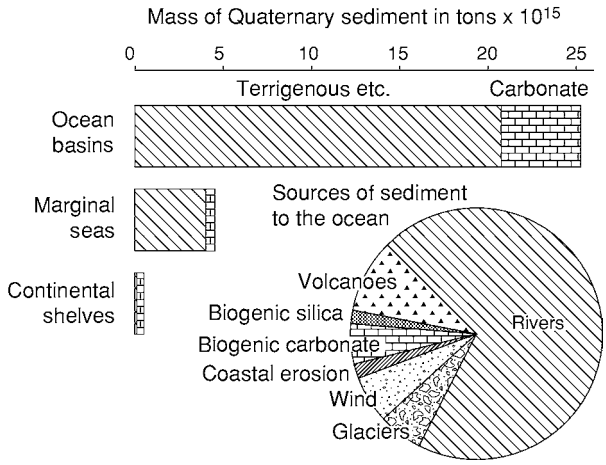


Figure 1 The volume of Quaternary sediment in the ocean, showing that the bulk of the deposition is in the deep ocean. The principal sediment sources are rivers, with volcanoes, glacial and wind transport, and biogenic carbonate also being important.

Sediment Sources

Rivers provide the principal input of sediment into the oceans (Figure 1). At high latitudes, terrigenous

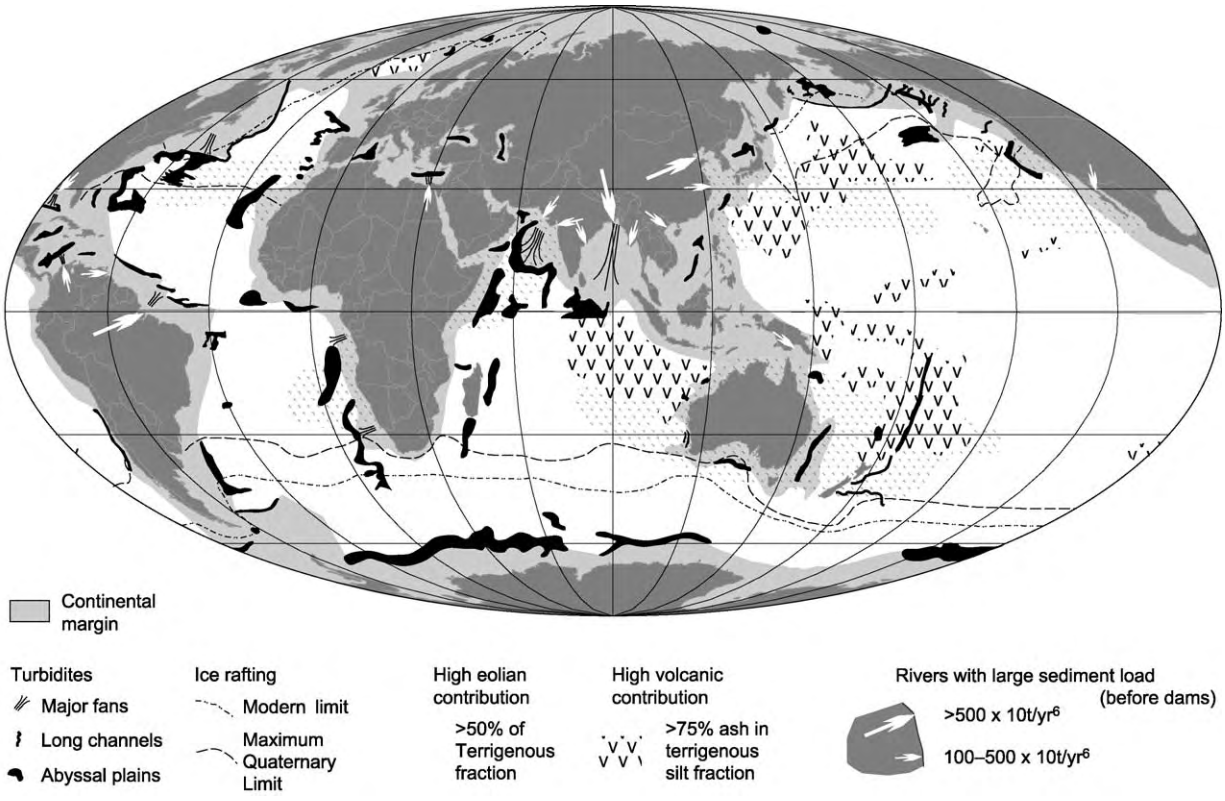


Figure 2 Terrigenous sediment in the deep ocean.

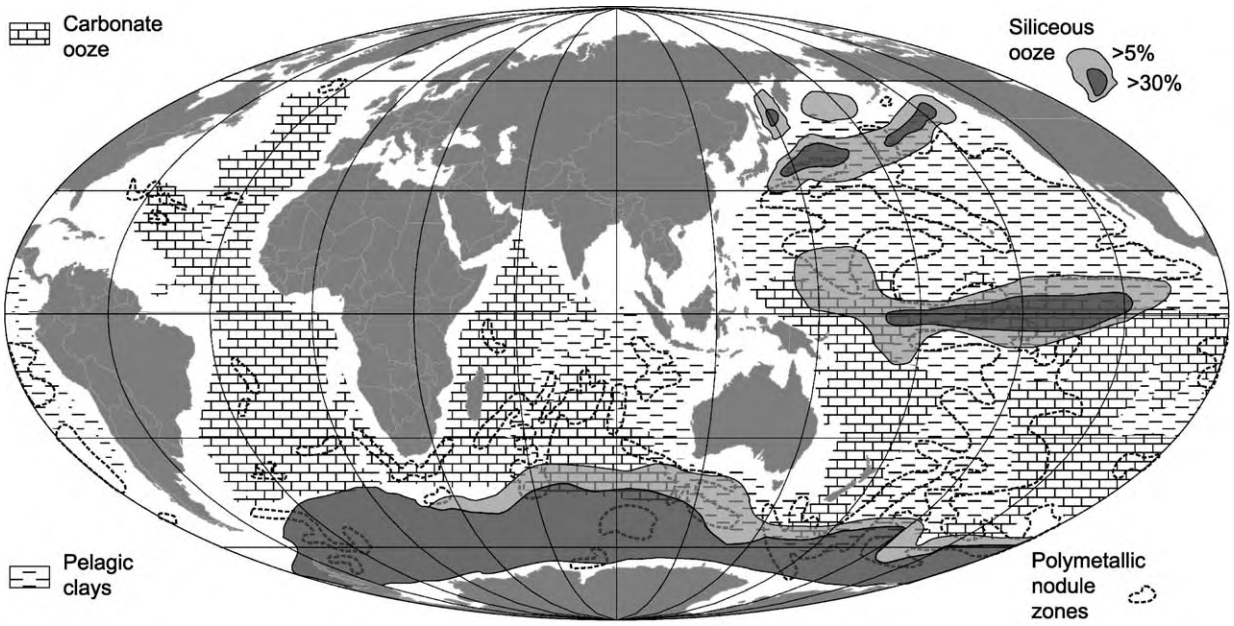


Figure 3 Distribution of the principal types of pelagic sediment.

sediment is supplied by sediment derived from subglacial melt water discharge and directly from ice, including meltout from icebergs thousands of kilometres from their source. In mid-latitudes, wind-blown dust (exceptionally including sand) is a significant component of deep-ocean sediment. Locally, volcanic material is a major component of seafloor sediment. Pelagic sediments (*see* **Sedimentary Rocks:** Deep Ocean Pelagic Oozes) consist principally of the skeletal material of pelagic organisms, notably calcium carbonate from Foraminifera, Coccolithophoridae (coccoliths or nannofossils), and pteropods, and opaline silica from Radiolaria, diatoms, and silicoflagellates. In oceanic areas with low sedimentation rates, various authigenic minerals form *in situ* in the sediments or on the ocean floor and may become significant components of the sediment. These include: phosphorites formed in slow-sedimentation environments in areas of active upwelling; metal-rich sediments and iron oxides associated with hydrothermal discharge in areas of active volcanism, particularly mid-ocean ridges (*see* **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges); manganese nodules and crusts of iron–manganese oxides, which precipitate in areas of slow sedimentation on the deep ocean floor (*see* **Sedimentary Rocks:** Oceanic Manganese Deposits); zeolites, which form principally from the alteration of volcanic ash and glass; and barite, which may be related to either hydrothermal activity or high organic productivity.

The dispersal of terrigenous sediment from various sources can be tracked by studying sediment petrology. The gross distribution of clay minerals within the ocean has long been used to infer sediment source and dispersion. In recent years, the signature of radiogenic isotopes (particularly lead and neodymium) and mineral geochronology in detrital terrigenous sediment have been used to understand further source and dispersion, including the tracking of ice-rafted detritus and aeolian dust. Tephra beds have been characterized geochemically and mineralogically, both to identify source and as stratigraphic markers.

Sediment Transport Processes in Deep Water

The transport of sediment to deep water involves both the normal processes of oceanic circulation and episodic sediment gravity flows. The resulting terrigenous deposits depend on the supply of sediment from the continent and shallow continental margin and on the complex interaction between flows and ocean morphology.

Surface currents in the ocean are primarily wind-driven, whereas overall circulation in the oceans is a consequence of density differences caused by temperature and salinity variations in the ocean water. Deep circulation is driven by the sinking of cold saline water in the North Atlantic Ocean and around Antarctica, with net upwelling in the North Pacific Ocean. The effect of the Coriolis force is to intensify ocean circulation on the western sides of oceans, where the thermohaline circulation generally

parallels the continental margin. Thus, for example, on the western side of the North Atlantic Ocean a powerful surface current (the Labrador Current) strongly influences the seafloor to water depths of 1500 m, and the deep Western Boundary Undercurrent of cold saline water derived from the Norwegian and Greenland Seas flows southwards at water depths of 3500–5000 m off the coast of Canada and the USA. Constricting topography results in the intensification of currents, for example around and over seamounts, through major fracture zones in the mid-ocean ridge system, and where microcontinental fragments such as Rockall Bank or the Campbell Plateau create constrictions.

The mean flow in both surface currents and the thermohaline circulation is generally a few centimetres per second, but locally the flow can be complex and energetic, for example in the large eddies in the Gulf Stream and Kuroshio, the former creating the 'abyssal storms' measured at the Hebble site in the Western Boundary Undercurrent of the North Atlantic, where velocities exceed $0.2\,m\,s^{-1}$. Similarly high velocities are found in topographical constrictions, for example in the Gulf of Cadiz, where Mediterranean Overflow Water exits the Straits of Gibraltar. Such flows are capable of eroding fine-grained seafloor sediment, may leave lag sands and gravels, and produce a variety of abyssal bedforms. Suspended sediment derived from the continental margin or suspended by ocean currents may be advected long distances as nepheloid layers – zones of higher amounts of particulate matter in the ocean circulation. In lower energy areas, this suspended matter settles to form sediment drifts. Boundaries between different thermohaline water masses are commonly marked by zones of higher turbulence and internal waves, which lead to higher amounts of suspended particulate matter and, in some cases, to erosion where they interact with the continental slope or with seamounts.

Organic matter sinking from the surface waters of the ocean is progressively oxidized, thereby depleting mid-level waters in oxygen and at the same time returning many nutrients to seawater. These nutrients become available in areas of upwelling. Deeper waters therefore have higher concentrations of dissolved carbon dioxide, with the highest concentrations being found in the older deep water of the Pacific Ocean. Deep waters of the Pacific Ocean are consequently undersaturated in calcium carbonate, leading to the dissolution of biogenic carbonate. Most dissolution of opaline silica occurs during oxidation of the protoplasm, but further dissolution takes place in undersaturated deep waters.

The wind-driven surface circulation is sedimentologically important in dispersing surface plumes derived from rivers and glacial margins. In addition, icebergs are transported in the surface circulation and, as they reach warmer waters and melt, they deposit ice-rafted detritus on the seabed. During Pleistocene Heinrich events in the North Atlantic Ocean, surface melt-water plumes were dispersed 2500 km south-eastwards from Hudson Strait and icebergs were dispersed even farther by the surface circulation.

Atmospheric circulation is a lesser agent of sediment transport to the deep ocean. Volcanic eruptions contribute fine volcanic ash through both troposphere plume transport, which may extend for thousands of kilometres, and worldwide distribution of fine ash that reaches the stratosphere. Dust derived from deserts is the principal source of terrigenous material in much of the mid-latitude pelagic realm.

Infrequent gravity-driven processes, including turbidity currents, debris flows, and submarine landslides, transport the largest volumes of sediment from the shallow continental margins to the deep sea (*see* **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes). A turbidity current is a density current in which the denser fluid is a sediment suspension. Turbidity currents may be initiated by many processes: all that are needed are a mechanism to put sediment into suspension and a steep slope to maintain the flow. Thus, hyperpycnal flow of rivers, glacial outburst floods (jökulhlaups), storm resuspension of shelf or beach sediment (*see* **Sedimentary Environments:** Storms and Storm Deposits), and in-mixing of ambient water into a debris flow will all initiate turbidity currents on steep slopes. Potential energy is converted to kinetic energy as the flow accelerates down the continental slope, commonly eroding the seafloor and eventually depositing sediment on the deep ocean floor. Processes within an individual turbidity current may range from hyperconcentrated flow at the head and base of the flow to normal turbulent sediment suspension in the upper part and tail of the flow, ultimately producing a low-density suspension, which contributes to nepheloid-layer dispersion by the thermohaline circulation. Turbidity currents occur every 1–10 years in steep river-fed basins with narrow shelves and off large shelf-crossing rivers, but only every 100–1000 years on most continental margins and every 1000–10 000 years on distal abyssal plains.

The largest submarine landslides involve debris avalanches of blocks tens to hundreds of metres in size, which flow in a laminar fashion and scour deep grooves into the seafloor. Slide blocks overlying sands liquefied by seismic shaking and muddy rotational slumps may break up and in-mix water to evolve into debris flows. Large coherent slides may plough

into seafloor sediments, and outrunner blocks may extend far beyond the main slide; the movements of both are facilitated by hydroplaning.

Subaerial volcanic islands, most importantly in island arcs, contribute sediment to the marine realm by proximal fall-out of ash and by pyroclastic flows entering the sea and forming turbidity currents. Flank collapse of volcanoes, both subaerial and submarine, produces widespread and voluminous debris avalanches around volcanic islands, the largest exceeding 5000 km^3. Submarine volcanic eruptions may be effusive or explosive, producing lava flows, large quantities of hyaloclastite, and pumice. Loose material will tend to move downslope by mass-transport processes and may also be modified by bottom currents.

The biota plays an important role in deep-ocean sedimentology. The skeletons of planktonic organisms are the major component of pelagic sediments. Pelagic sediments are generally thoroughly bioturbated, as are hemipelagic muds. Microbial processes are important in controlling the chemistry of seawater and in the development of authigenic pelagic sediments. Coral reefs and both cold and hot vent communities create large biogenic rock masses in deep water.

Sediment Nomenclature

Different nomenclature schemes are applied to terrigenous sediments of the continental margin and to pelagic sediments of the open ocean. The standard nomenclature for terrigenous sediment is based on grain size, with the Shepard and Folk systems (Figure 4) being the most widely used. The term 'mud' is widely used for sediment comprising principally silt and clay. Standard pyroclastic terms are applied to submarine pyroclastic deposits, and terminology applicable to shallow-water limestones is generally also applied to deep-continental-margin limestones. For pelagic sediments, the Ocean Drilling Program scheme is widely used (Table 1). The term 'ooze' is applied to sediment with a biogenic component exceeding 60% and is preceded by the names of all components making up more than 30% of the sediment, for example foraminifer–nannofossil ooze (often abbreviated to foram–nanno ooze). Minor components (10–30%) are prefixed to the sediment name with the term rich, e.g. zeolite-rich radiolarian ooze. Where the biogenic content is between 30% and 60%, the name is based on the predominant fossil and the appropriate textural name for the siliciclastic component, e.g. foraminiferal clay.

Several genetic names are widely and loosely used to describe marine sediments. Hemipelagic muds are highly bioturbated muds with less than 30% biogenic material that accumulate slowly from suspension

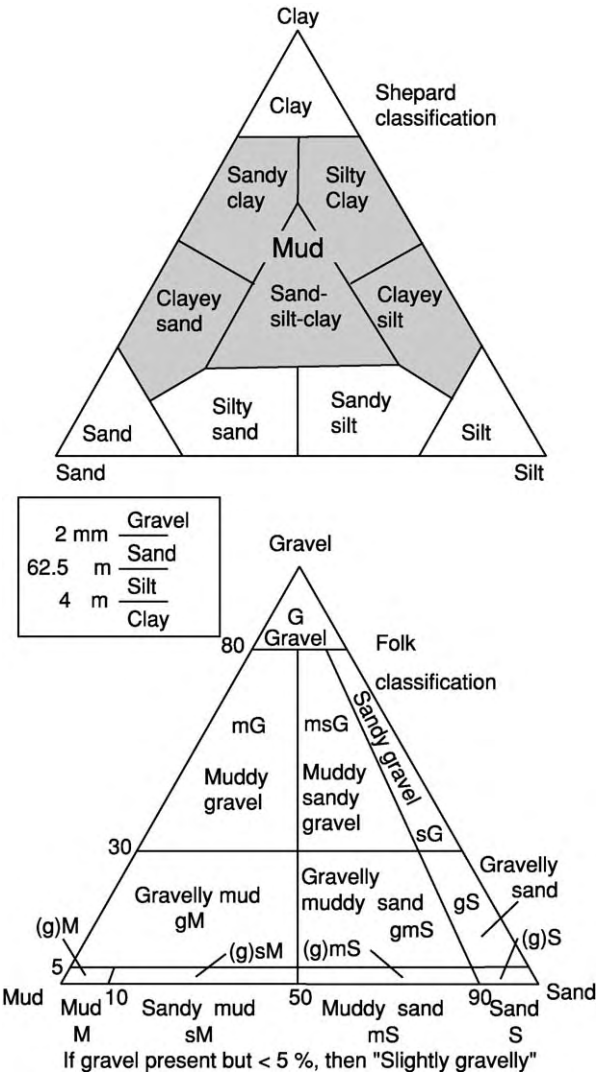


Figure 4 Shepard and Folk nomenclatures for terrigenous sediments.

fall-out of sediment that is either resuspended on the adjacent continental shelf or transported in nepheloid layers. The term turbidite is applied to sands, silts, and muds that are deposited principally from turbidity currents; their deposition is rapid, and bioturbation is sparse or absent. Sorted sands and gravels inferred to be deposited from hyperconcentrated flows are generally included in the term turbidite, although some authors have argued that the transport process resembles that of debris flows. Debrite is used for the blocky and poorly sorted deposits of debris flows, debris avalanches, and similar mass-transport processes. Deposits of contour-following bottom currents are generally called contourites (*see* **Sedimentary Environments:** Contourites); such deposits are most readily recognized from their morphological setting and from their character in seismic-reflection

Table 1 Nomenclature of pelagic sediments

EXAMPLES OF SEDIMENT NAMES					TYPE OF SEDIMENT		PRINCIPAL NAMED COMPONENTS
0	5	10	30	60	100 % biogenic material		
Silty clay, Clayey silt, Clayey sand, Clay (mud = silty clay and clayey silt)					TERRIGENOUS (continental margin)		Sand, silt clay
Clay	Zeolite-bearing clay	Nanno-rich clay			PELAGIC	Mainly abyssal brown clay	Clay, zeolite, ash, nannofossils, Radiolaria
	Foram-bearing clay	Foram-rich clay	Foram clay, nanno clay	Foram ooze, Foram-nanno ooze		Mainly biogenic carbonate	Foraminifera (forams), Nannofossils (nannos), Pteropods
	Diatom-bearing clay	Diatom-rich clay	Diatom clay	Diatom ooze, radiolarian ooze		Mainly biogenic silica	

profiles. The actual deposits are generally hemipelagic muds, although, in areas of strong current flow, coarser bioturbated winnowed deposits occur.

Continental Slopes

Continental slopes are regions of steeply sloping seafloor that lie between continental shelves and the deep ocean basins (Figure 2). Regional gradients are typically 2–5°, but locally slopes may be much steeper. Their large-scale morphology is a consequence of tectonic processes: the different elevations of continental crust and oceanic crust, the details of the original rift tectonics on passive continental margins, and the styles of subduction and accretion on convergent margins. This tectonic frame has been modified, most profoundly on older margins, by the deposition of continental-margin sediment, which locally may be more than 10 km thick over the basement. The position of the shelf break – the generally abrupt increase in gradient where the continental shelf passes into the continental slope – is in many places a direct consequence of sea-level lowering of around 110 m at the Last Glacial Maximum 21 000 years ago and marks the limit of progradation of shallow-marine sediments. In many cases the shelf break represents the seaward limit of delta-top progradation. In high latitudes, the shelf break may mark the seaward limit of continental ice-sheets. Where shelf-edge reefs are developed, the elevation of the shelf break is a consequence of a balance between sea-level rise and vertical reef accretion and is commonly close to present sea-level.

Continental slopes show a complex balance between erosional and depositional processes. The steep gradients promote sediment gravity processes, and the ocean-margin topography leads to strengthening of the oceanic circulation, particularly on the western sides of oceans. Deposition is principally of sediment that crosses the continental shelf and is therefore most important where shelves are narrow, shallow, and energetic and where sediment supply is high, notably seaward of major rivers and in areas of high carbonate productivity, including reefs. Regional deposition on the continental slopes off river mouths is from low-salinity surface plumes of fine-grained suspended sediment and from nepheloid layers developed from the settling of such sediment and resuspension of seafloor sediment. Storms and, on some shelves, tidal currents resuspend shelf sediment and advect it in suspension to the upper slope, where it is deposited as hemipelagic mud.

Slope morphology strongly influences sedimentological processes and products. Submarine canyons are widespread on many continental slopes, ranging in scale from the mighty Monterey Canyon, which is of similar dimensions to the Grand Canyon of the Colorado River, to small-scale slope gullies. Submarine canyons seaward of wide continental shelves are relatively inactive: the morphology appears to be a relict from sea-level lowstands. The canyons concentrate open-ocean water circulation, so that tidal flows may be sufficient to transport fine sands. These flows and the related upwelling also result in higher biogenic productivity than on the open slope. Sediment gravity flows, however initiated, will tend to flow within canyons, where they will accelerate on steep slopes.

Many canyons have their origins in river-mouth processes, both contemporary, as in the case of the Var Canyon in south-eastern France, and more generally at times of sea-level lowstands, when many rivers

discharged near the present shelf break. Direct hyperpycnal flow of sediment-laden floodwaters is the principal erosional process for smaller rivers with high bedload discharge; the role of hyperpycnal flows in very large rivers is uncertain. Glacial outburst floods erode particularly large submarine canyons, such as on the Laurentian Fan off south-eastern Canada. Studies of fjord deltas, a readily accessible modern analogue of shelf-edge rivers, suggest that failure of rapidly deposited river-mouth sediment may be equally important for initiating erosive turbidity-current flows. Canyons initiated by river-mouth processes at sea-level lowstands may undergo headward erosion as sea-level rises, as in the La Jolla Canyon off southern California and the Zaire/Congo Canyon of West Africa. In the La Jolla Canyon, beach sands accumulate in the canyon head under fair-weather conditions and are resuspended by storm waves that set up down-canyon turbidity flows that accelerate and erode canyon-filling sediment. Storm-wave transport of carbonate-platform sediment, especially through tidal passes in reefs, is important in initiating erosive turbidity-current flows on low-latitude carbonate margins. In general, turbidity currents will accelerate and erode on slopes of more than 2°, resulting in erosional undercutting of canyon walls, which promotes sediment failure. On some continental slopes, retrogressive sediment failure is an important process in the development of slope gullies. Slope gullies may also form as a result of density flows, derived from the fall-out of suspended sediment near river mouths and at ice margins, progressively coalescing to produce a badland-like drainage system.

The effects of sediment failure are widespread on many continental slopes, and failures are some of the most prominent morphological features on modern swath bathymetry of the continental slope. Many large-scale features, such as the 200 km wide Storegga Slide on the Norwegian margin, which failed 8000 years ago, are thought to have been triggered by earthquakes, by analogy with the 1929 Grand Banks failure on the Canadian Atlantic margin and the 1998 Papua New Guinea event. In many cases, including Storegga and the Cape Fear Slide on the US Atlantic margin, there is evidence that sediment strength was reduced by excess pore pressure due to the release of gas from gas hydrates. A particular type of large-scale sediment failure occurs on the slopes of oceanic volcanic islands, such as the Canary and Hawaiian islands, where progradation of steep volcanic edifices over weak deep-water sediments results in episodic catastrophic failure of the volcano flanks. Smaller-scale sediment failure is also widespread on continental slopes. On sediment-starved slopes, such as off New England (USA), failure may take place in Tertiary strata and principally involves slides. Retrogressive rotational slumps commonly evacuate many tens of metres of soft sediment on continental slopes producing blocky mass-transport deposits and debris flows. Sediment creep, such as that described from the South Korea Plateau, may eventually lead to failure along décollement zones tens to hundreds of metres below the seafloor.

Larger-scale gravity sliding of the upper few kilometres of the sediment column is the dominant feature of slope evolution seaward of some of the largest deltas in the world, notably those of the Niger and the Amazon, producing slope-parallel ridges and a prominent slope-toe escarpment. A similar large-scale morphology is produced on slopes with active salt tectonics, most spectacularly in the Gulf of Mexico, where the Sigsbee Escarpment is located at the toe of the slope and the slope morphology is dominated by salt diapirs and salt-withdrawal basins. Accretionary prisms at sediment-dominated convergent margins show large-scale morphological similarities to these complex passive-margin slopes, with thrust-controlled ridges and basins. In all of these complex slope settings, submarine canyons and valleys have irregular paths and variable degrees of incision. Slope failures are particularly common, as a result of tectonic oversteepening. Gas hydrates are particularly abundant on many accretionary margins, and fluctuations in the stability field with changes in bottom-water temperature or sea-level may trigger failures.

Overthrusting results in underconsolidated sediment with excess pore pressures, which commonly migrates to the surface to form mud volcanoes. Mud volcanoes are also common in prodeltaic environments with high sedimentation rates. Smaller-scale pockmarks – crater-like seafloor depressions a few tens of metres in diameter – indicate locations where gas or formation fluids are released, either catastrophically or quasi-continuously. In many areas, such pockmarks are developed along faults that leak formation fluids. Leaking hydrocarbons may also provide the energy source for biogenic 'cold-seep' communities on the continental slope.

Low-latitude carbonate platforms supply two main types of sediment to the continental slope. Where the platforms are fringed by rapidly growing reefs, collapse of the reef front results in debris avalanches. Storm transport of shelf sediment off a carbonate platform, particularly if it is focussed through passes in reefs, can produce turbidity currents, depositing well-sorted carbonate-rich sands and muds. At sea-level lowstands, with a narrow continental shelf, carbonate margins may behave like terrigenous margins, with the cutting of slope canyons that lead to small fan channels and lobes. At sea-level highstands,

carbonate production is generally much higher and may lead to progradation of the continental slope, channels on the lower slope, and turbidite deposition of carbonate sediment.

In contrast, the effect of lowered sea-level on mid-latitude continental margins was to increase sediment supply to the deep sea. In many cases, rivers discharged onto and prograded across a much narrower continental shelf, so that sediment was supplied directly to the high-gradient continental slope, down which it was transported by turbidity currents and mass-transport processes. At high latitudes, continental ice streams crossed many continental shelves, delivering glacial diamict directly to trough-mouth fans. Sediment delivery was particularly high during ice retreat in subarctic regions and was aided by abundant melt water. Sediment failures were triggered by glacio-isostatic earthquakes.

The Deep Continental Margin

The continental rise on a passive continental margin is a zone of sediment deposition on slopes that are typically between 1 : 50 and 1 : 500 and occurs beyond the steeper continental slope, which is commonly incised by canyons. The continental rise consists principally of submarine fans. An erosional submarine canyon leads to a submarine fan valley or channel, generally with depositional levees and a downslope decrease in channel depth. Depositional lobes are developed at the downslope termination of the channel and pass gradually into the flat abyssal plain, where the gradient of less than 1 : 1000 would be imperceptible to an observer on the ground. In more complex convergent continental margins there is generally more morphological control of deposition, but the same architectural elements are present: erosional canyons lead to leveed channels, channel-termination lobes, and, finally, basin or trench floors with more sheet-like stratigraphy. Channels may link two morphologically distinct basins, as in the case of the Biscay and Iberia abyssal plains. Indeed, these elements are also present on complex continental slopes with intraslope basins.

The channel systems are pathways for turbidity currents. Many channels are highly meandering; whether a channel is meandering or straight appears to be related to the gradient. On many passive-margin submarine fans, such as the Amazon, Zaire, and Bengal fans, channels change course abruptly through avulsion. Where channels are topographically constrained, for example Bounty Channel east of New Zealand, Surveyor Channel in the north-east Pacific, and the North-west Atlantic Mid-Ocean Channel in the Labrador Sea, an almost constant path has been maintained over thousands of kilometres. The lower parts of the turbidity currents transport sand through the channels and, owing to flow expansion, deposit it rapidly as channel-termination lobes. Only fine-grained sediment is transported to the distal parts of basins. The upper parts of the current entrain water, thicken, and spill over the levees, depositing overbank silts and muds. The cross-sectional areas of channels (kilometres wide, tens to hundreds of metres deep) and the velocities of turbidity currents estimated from cable breaks (2–19 m s ^{1}) testify to the size and power of the turbidity currents.

Sediment drifts are large sediment bodies constructed by deposition, generally of hemipelagic sediment, from the deep-water thermohaline circulation. The most powerful currents, in constrictions or adjacent to steep slopes, will commonly winnow or erode sediment, some of which may be of turbidite origin. This sediment, together with material introduced into nepheloid layers from the tails of turbidity currents or from resuspension on the shelf, is deposited slowly in areas of lesser current velocity. The North Atlantic Ocean, the Mediterranean Sea outlet, the eastern sides of the North and South American continents, the New Zealand margin, and the west side of the Antarctic Peninsula have particularly well-developed sediment drifts. Most have accumulated over many millions of years and have undergone alternating phases of deposition and winnowing or erosion as bottom-water circulation has fluctuated with changing climate. Many sediment drifts support mesoscale bedforms, including erosional furrows and sediment waves, with wavelengths of kilometres and heights of tens of metres.

Sediment Distribution in Pelagic Realms

Half of the Earth's surface is covered by pelagic sediment, yet study of its sedimentology is challenging because of its slow sedimentation rates and intense bioturbation. Some 47% of the pelagic realm is floored by foraminiferal ooze, 15% by siliceous ooze (mostly diatom ooze around Antarctica), and 38% by abyssal brown clay, in areas where there is total dissolution of biogenic material.

The distribution of biogenic oozes is a consequence of surface productivity, which is greatest in areas of upwelling, notably at the equator, on the western margins of continents, and at the Antarctic Convergence. Most biogenic material that reaches the

deep-sea floor does so through accelerated settling in the form of faecal pellets from zooplankton. Calcium carbonate is dissolved in deeper waters, particularly in the Pacific Ocean. The depth at which seafloor dissolution is complete is known as the carbonate compensation depth and fluctuates spatially and through time with changes in the thermohaline circulation.

Thus, foraminiferal and nannofossil oozes are widespread in the Atlantic and Indian oceans and on the East Pacific Rise but are absent over much of the northern and south-western Pacific Ocean (Figure 3). Sedimentation rates are generally 10–50 mm ka ^{1}. Diatom oozes occur beneath the high-productivity areas of the Antarctic Convergence and the extreme North Pacific Ocean, and radiolarian oozes occur beneath the Equatorial Divergence in the Pacific Ocean, with sedimentation rates of 2–10 mm ka ^{1}. Abyssal brown clays are most abundant in the Pacific Ocean, below the carbonate compensation depth, and have sedimentation rates of approximately 1 mm ka ^{1}. The authigenic components of pelagic sediments have been described above. Metalliferous sediments are generally restricted to small areas adjacent to sites of active volcanism but do cover a large area of the East Pacific Rise near the Galapagos Islands. Manganese nodules are particularly common in parts of the Indian and Pacific oceans with slow sedimentation rates. Poorly sorted ice-rafted detritus is an important component of pelagic sediment in the Arctic Ocean, the southern Ocean around Antarctica, and the northern North Atlantic Ocean.

The progressive subsidence of oceanic crust away from spreading centres results in a systematic stratigraphical succession in pelagic sediments. Thin metalliferous sediment and iron–manganese oxides may directly overlie oceanic crust. They are overlain by foraminiferal ooze deposited on the shallower seafloor of the mid-ocean ridge, with the sedimentation rate and the proportion of siliceous organisms depending on latitude. Eventually, subsidence of the oceanic crust may bring the seafloor below the carbonate compensation depth, and abyssal brown clays accumulate. Finally, subsidence and perhaps subduction will bring the ocean floor within reach of terrigenous sediment from the continent.

See Also

Sedimentary Environments: Contourites; Storms and Storm Deposits. **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes; Deposition from Suspension. **Sedimentary Rocks:** Deep Ocean Pelagic Oozes; Oceanic Manganese Deposits. **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges; Ocean Trenches.

Further Reading

Fisher RV and Smith GA (1991) *Sedimentation in Volcanic Settings*. SEPM Special Publication 45. Tulsa: Society for Sedimentary Geology.

Gardner JV, Field ME, and Twichell DC (1996) *Geology of the United States' Seafloor: The View from GLORIA*. Cambridge: Cambridge University Press.

Ginsburg RN (2001) *Subsurface Geology of a Prograding Carbonate Platform Margin, Great Bahamas Bank*. SEPM Special Publication 70. Tulsa: Society for Sedi mentary Geology.

Hay WW (1994) *Material Fluxes on the Surface of the Earth*. Washington, DC: National Academy of Sciences.

Hovland M and Judd AG (1988) *Seabed Pockmarks and Seepages*. London: Graham and Trotman.

Ivanov MK, Kenyon NH, Suzyumov AE, and Woodside JM (eds.) (2003) Special Issue: Sedimentary processes and seafloor hydrocarbon emissions on deep European con tinental margins. *Marine Geology* 195: 1 318.

Kennett JP (1982) *Marine Geology*. Englewood Cliffs, NJ: Prentice Hall.

Lisitizin AP (2002) *Sea ice and Iceberg Sedimentation in the Ocean: Recent and Past*. New York: Springer.

Lisitzin AP, Kennett JP, and Woolhiser C (1996) *Oceanic Sedimentation: Lithology and Geochemistry*. Washington, DC: American Geophysical Union.

Mulder T and Syvitski JPM (1995) Turbidity currents gen erated at mouths of rivers during exceptional discharges to the world oceans. *Journal of Geology* 103: 285 299.

Pickering KT, Hiscott RN, and Hein FJ (1989) *Deep Marine Environments: Clastic Sedimentation and Tectonics*. London: Unwin Hyman.

Rea DK (1994) The paleoclimate record provided by eolian deposition in the deep sea: the geologic history of wind. *Reviews of Geophysics* 32: 159 195.

Stoker MS, Evans D, and Cramp A (eds.) (1998) *Geological Processes on Continental Margins: Sedimentation, Mass Wasting and Stability*. Special Publication 129. London: Geological Society.

Stow DAV and Mayall M (2000) Deep water sedimentary systems: new models for the 21st century. *Marine and Petroleum Geology* 17: 125 135.

Tappin DR (ed.) (2004) Special Issue: Submarine slump generated tsunamis. *Marine Geology* 203: 199 386.

Weaver PPE, Wynn RB, Kenyon NH, and Evans J (2000) Continental margin sedimentation with special refer ence to the north east Atlantic margin. *Sedimentology* 47(Suppl. 1): 229 256.

Wynn RB and Stow DAV (2002) Special Issue: Recognition and interpretation of deep water sediment waves. *Marine Geology* 192: 1 333.

Fluvial Geomorphology

J Lewin and P A Brewer, University of Wales, Aberystwyth, UK

Introduction

Fluvial geomorphology is concerned with the creation of landforms by river processes through the removal and transfer of materials on Earth's surface. Process studies are rooted in a number of disciplines, the earliest identifying 'processes' as evolutionary time sequences in landforms. In later work, this has continued in such fields as the study of meander dynamics. Hydrology-based approaches have involved the identification of formative (generally high river flow) events and their frequency, together with empirical relationships between river channel dimensions and river flow parameters. Work in engineering hydraulics, including the need to design stable channels, has also been a source of process understanding. Such work has led on to detailed analysis of water flow structures in river channels and prediction of thresholds for, and amounts of, sediment entrainment, transport, and deposition. Physical modelling both of channel reaches and of the treelike networks of rivers in their drainage basins has also now been joined by numerical modelling at a number of scales: flow in meander bends and confluences, the development of channel patterns, the evolution of drainage networks, and simulation of alluvial stratigraphy and sediment transfers within catchments as a whole. All this is particularly helpful because landform development, responding perhaps to only a few formative events in a year or to a rare flood, is only fractionally observable. In effect, modelling allows both a 'speeding up' of processes and experimentation with presumed controlling factors such as gradient and sediment sizes and mixtures. Results may then be checked against field observations. On the other hand, some field circumstances (such as the effect of needle ice or animal trampling on river bank stability, or real-time discharge fluctuations and their effects on groundwater and sediment flux on floodplains) are less easily modelled.

Fluvial systems exist over a range of scales, from centimetre-wide intertidal channels and other drainage networks on a sandy beach, to the more than 6 million km^2 drainage basin of the Amazon. A fine-textured system is shown in Figure 1. All such systems are complex, self-organizing, and hierarchical in structure. At the microscale, processes involve the movement of individual particles. These may be arranged in bodies of sediment (as in increments of bank erosion, moving gravel sheets on a river bed, or as sedimentation layers following a large flood), and these in turn become part of 'form units' such as channels, bed material forms (dunes or bars), levees and the crevasse channels that may cut across them, and backswamps. These units are combined into what have been called 'architectural ensembles', or types of rivers and their floodplains. These have form units that are present to differing extents according to contrasting process environments, each with their differing hydrological regimes, stream

Figure 1 The Zabriskie Point badlands in Death Valley, California, USA. The landscape is dissected by a hierarchy of channel networks ranging in scale from rills a few centimetres wide to a braided channel belt over 10 m wide.

energies, vegetation, and sediment sizes and supply rates.

At the highest level, drainage basins incorporate a range of ensembles from head to mouth. Higher level dynamic behaviour includes the evolution of drainage networks in the context of changing sea-levels, climate, and tectonic activity. This can become very complex, with forms retaining a kind of 'memory' of past conditions, which in part may affect their active behaviour. For example, cold-climate weathering and erosion products of the last Ice Age make up a major part of channel bed materials in some present-day temperate environment rivers. Elsewhere, bedrock channels may be more apparent, and in fact bedrock erosion has recently been more closely studied, notably in headwater channels that may be the major source of transported catchment sediment, and in the rocky channels and gorges found in steeplands, some semiarid environments, and elsewhere.

Overall, fluvial geomorphology is therefore concerned with a wide range of time and space scales, ranging from rock particles to river catchments. At the lowest levels, studies may be concerned with equilibrium responses to quasi-instantaneous forces; with increasing level, however, response time increases to a length that may parallel global climate change. Even without a change in external factors of this kind, it is likely that there is a need to understand evolutionary (and technically chaotic) behaviour, as well as equilibrium behaviour (as in the response of a particle of given shape and mass to applied fluid force). However, modelling complex behaviour in hierarchical systems does commonly incorporate lower level 'process rules', and this underlies the need in fluvial geomorphology for improved understanding of process mechanisms at all levels. Catchment factors (varying the gradient, sediment, and discharge at individual sites) affect local processes, which in turn may operate to change the character and extent of catchments.

Material Transfer

Lower level material transfer in alluvial systems is illustrated diagrammatically in Figure 2A, which emphasizes that periods of storage and reactivation are involved. Coarser material moves incrementally in reach-length steps associated with flood events, whereas fine sediment may move long distances during single events. Solute transfer is continuous, if varying in concentration. Solutes constitute an extremely significant part of the total load carried by rivers, especially, of course, in catchments with soluble limestone and shale lithologies (Table 1). Weathering processes (surface and subsurface) are important both in the general reduction of land surfaces and for making possible the input and transport of rock materials in particulate or solute form. For rivers, cold-climate environments (including glacierized ones) may produce large quantities of coarse, physically weathered materials; low-relief and vegetated tropical environments may contribute clay-rich sediment and solutes to their river systems from weathered soils. Fine sand may dominate river loads in sandstone terrains. Generally speaking, sediment transfer is intermittent; thresholds for movement both in channel and related to sediment input from hillslopes are differentially achieved within catchments, leading to what is called 'complex response' to forcing events.

For submerged or partially submerged particles, it is possible to calculate the physical forces that must be exerted on the particles to move them on a river

Table 1 Suspended and solute load data for selected rivers[a]

River	*Drainage area (km^2)*	*Mean specific suspended sediment load ($t\,km^{-2}\,yr^{-1}$)*	*Mean specific dissolved load ($t\,km^{-2}\,yr^{-1}$)*	*Percentage of total load carried in solution*
Congo	4 000 000	13	12	48
Nile	3 000 000	0.7	4	86
Brahmaputra	580 000	1370	130	9
Ganges	975 000	537	78	13
Danube	817 000	102	65	39
Mississippi	3 267 000	94	40	30
Colorado	635 000	870	23	3
St Lawrence	1 025 000	5	53	91
Amazon	6 000 000	79	46	37

[a]Data from Milliman JD and Meade RH (1983) World wide delivery of sediment to the oceans. *Journal of Geology* 91: 1 21; and Meybeck M (1976) Total mineral dissolved transport by world major rivers. *Hydrological Sciences Bulletin*, *International Association of Scientific Hydrology* 21: 265 284.

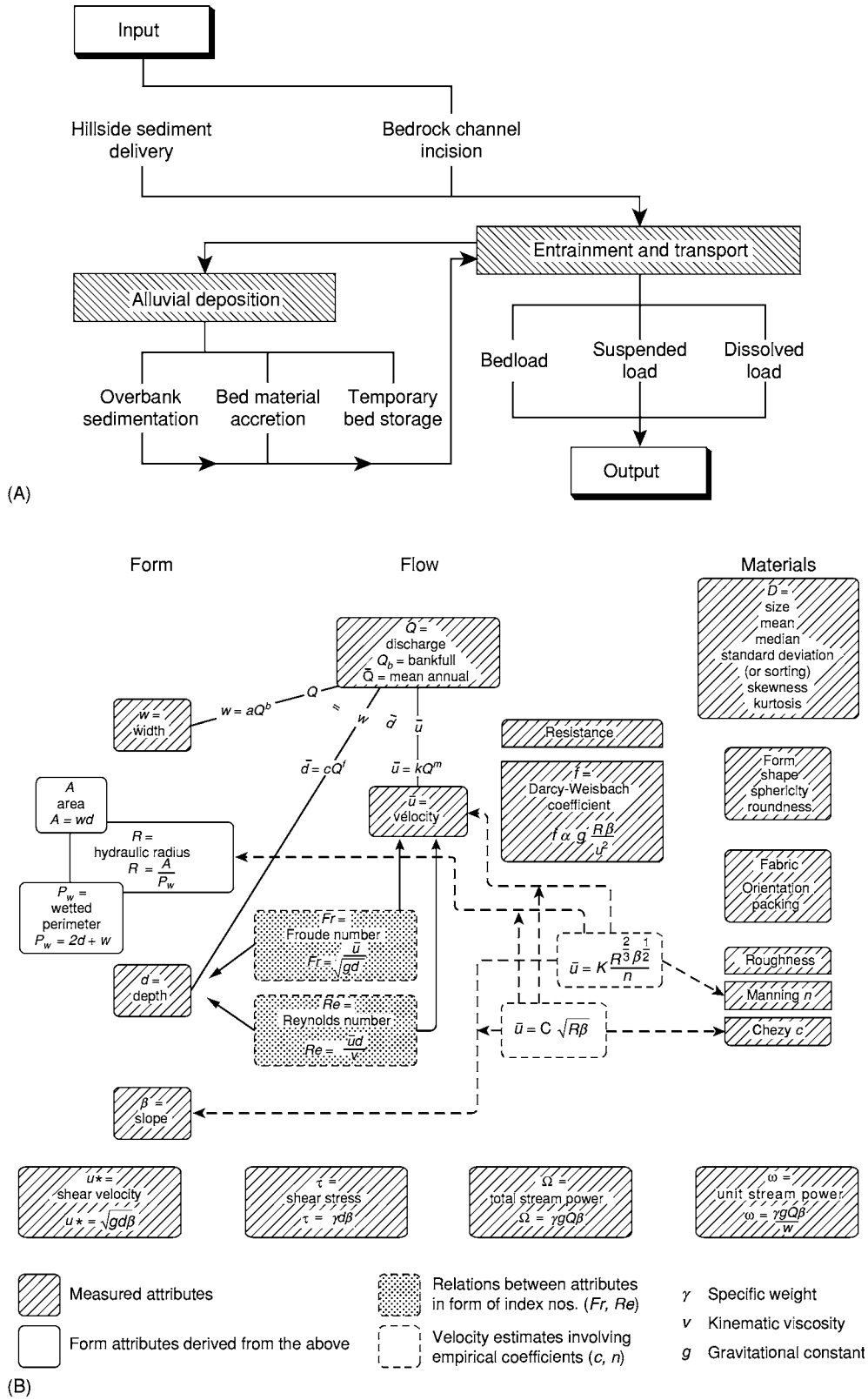


Figure 2 (A) Material transfer in an alluvial river system. (B) Relationships between river form, flow, and materials. Reprinted with permission of Taylor and Francis.

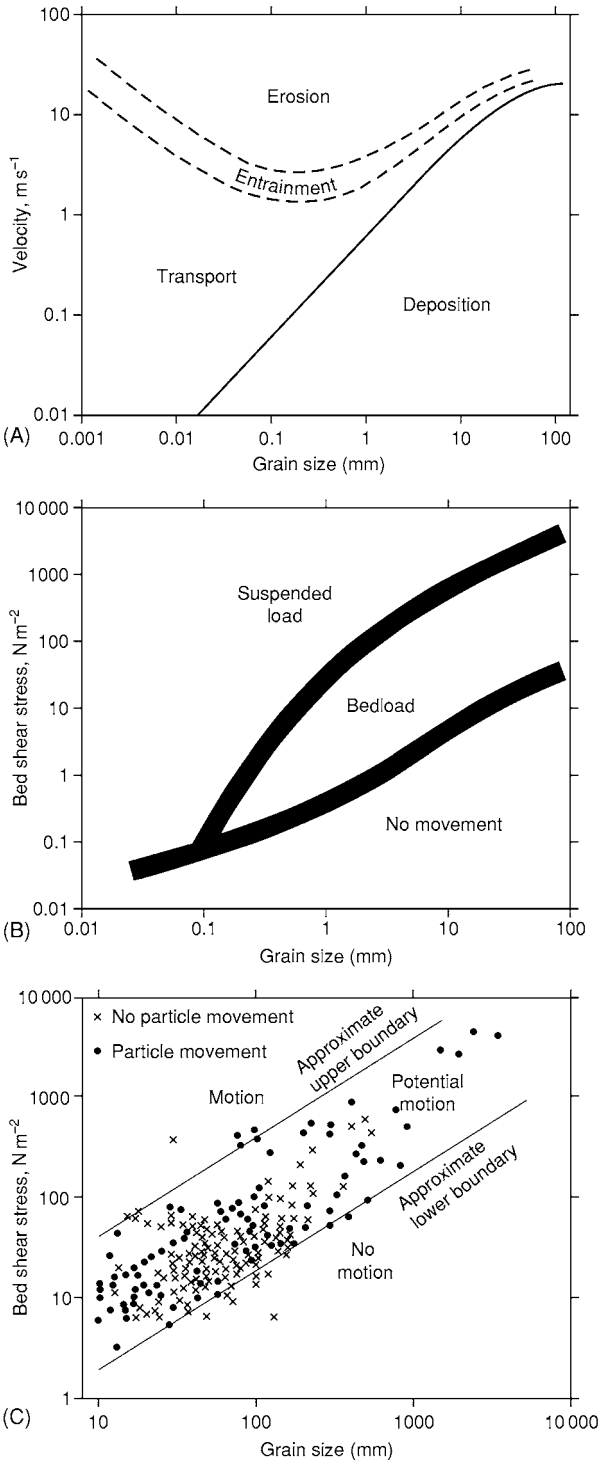


Figure 3 Relationships between sediment transport and bed velocity/shear stress. (A) Hjulstrom's (1935) critical velocity fields for entrainment, erosion, transport, and deposition. Note that for fine grained cohesive materials, much higher velocities may be required to erode materials than to transport them in disaggregated form. (B) Summary of experimental observations on the forms of sediment transport. Reproduced with permission from Bridge JS (2003) *Rivers and Floodplains*. Blackwell: Oxford. (C) The more complex relationships actually observed in field situations. Reproduced with permission from Knighton D (1998) Fluvial Forms and Processes. Arnold: London.

bed, but these forces are usually expressed in terms of mean flow parameters (which are more easily estimated) such as mean velocity, bed shear stress, or unit stream power for a given river level (see Figures 2B and 3).These can then be related to movement thresholds, for particles of given size, as bedload or suspension load. However, field data show great complexity because of factors such as the presence of mixed grain size and the packing and armouring (in which finer material at depth is protected by coarser material above) of bed materials (Figure 3). The actual entrainment of particles also involves eddy, sweep, and burst flow phenomena in flowing water. Nevertheless, in general, turbulent flow (turbulence being expressed in terms of the Reynolds number) is capable of transporting fine material in suspension, with amounts set by supply rates, whereas coarser materials (available to the river from its bed and banks) are more characteristically competence limited, such that a threshold of river flow has to be reached for coarser sediment entrainment and transport. For coarser load, it has proved possible to develop a range of bedload formulas (for example, relating transport rate to excess shear stress, discharge, or stream power, the 'excess' being the value above the threshold at which particle motion begins). For gravel rivers, such equations have not in general proved very successful when checked against the limited amount of direct bedload transport observations available. Equipment may be of the kind installed on the East Fork River, Wyoming, where moving sediment was trapped in a set of open slots on the river bed, moved to the bank on a conveyor belt to be weighed, and then returned to the river further downstream (Figure 4).

In general, suspended and solute loads have to be measured following water sampling, and yield measures are much more generally available. Data obtained during flow events over periods of years show results that may be complex in detail: for example, sediment concentrations may be higher as waters rise during a flood, compared to when waters recede, and there is no precise and unique relationship between discharge or water stage and concentration. However, it appears that events of moderate frequency (around annual) and stage (about bankfull or channel-full) achieve the most work, though for bed materials, large floods may be necessary both to generate slope failure and coarse-material input and to move the larger particles on a river bed. Floods that are extreme, compared to the annual average, occur every few years in some environments, whereas in other environments the flood expected, say, every 10 years is not that much larger than the annual one (Figure 5). Globally, sediment yields may vary greatly and are high particularly in high-relief/active tectonic

Figure 4 A field facility for bedload sampling and measurement installed by the United States Geological Survey on the East Fork River, Wyoming.

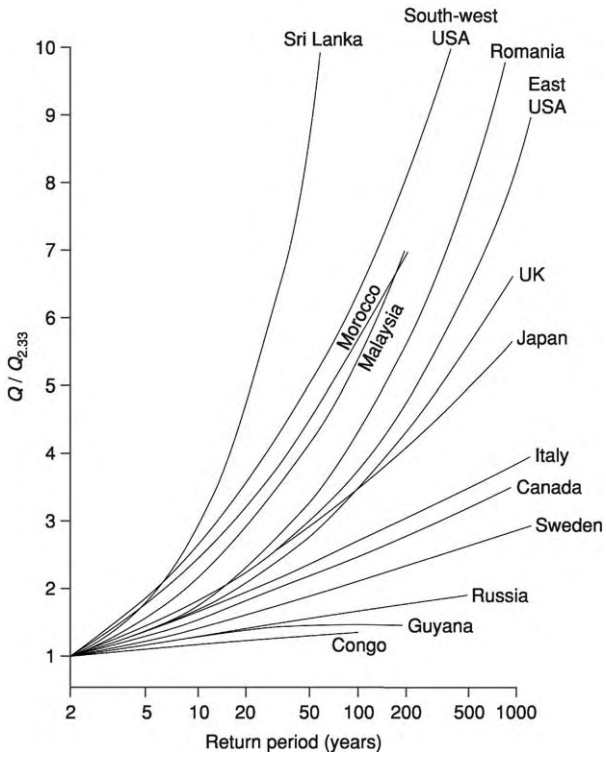


Figure 5 Curves showing flood frequency and magnitude relative to the mean annual flood ($Q_{2.33}$).

environments, on highly erodible terrains (including loess areas), and in regions of human disturbance.

Empirical relationships have also been established between channel measures and flow parameters. Flow velocity can be related to hydraulic radius (A/P; where A is the cross-sectional area of water flow and P is the wetted perimeter) and estimates of bed roughness, whereas river width and depth and velocity are differently related to variation in discharge, either as discharge rises or falls at a given cross-section ('at a station') or for constant flow conditions as discharge increases downstream (Figure 2B). These statistical average 'hydraulic geometry' relations (derived for single-thread rivers and so called because they are power function or 'geometrical' in form) are less apparent for downstream gradient and grain size, though both in general decrease, the former giving concave-up river 'long profiles', and the latter showing the effectiveness of in-channel size sorting and abrasion of particles as they move. In fact, research has rather moved away from detecting empirical relationships of this kind towards physically and theoretically underpinned modelling and field monitoring of bank erosion and channel shape processes. Field and laboratory studies of sorting and abrasion processes have explored topics such as abrupt discontinuities in channel bed grain sizes, and the nature of abrasion processes (see Figure 6).

Form Units

Alluvial rivers generally possess a set of landforms, self-formed in river-transported materials. There is a basic contrast between channels and their floodplains, the latter inundated only during flood flows producing a blanketing of flood deposits, which may

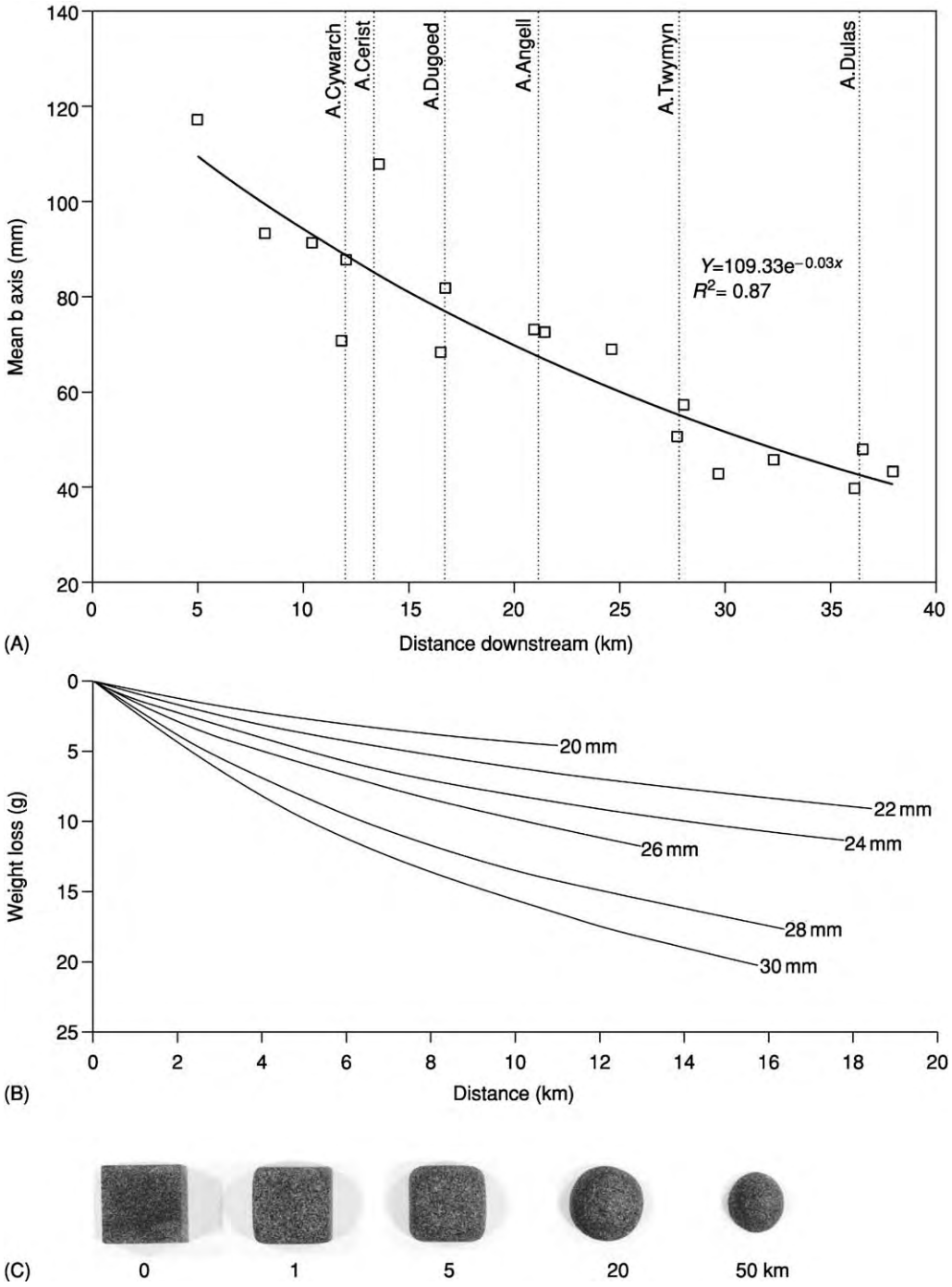


Figure 6 Field and laboratory data illustrating downstream fining and abrasion processes. (A) Downstream decrease in mean grain size on the Afon Dyfi, mid Wales, UK; abrupt, but generally localized, increases in grain size are attributed to relatively coarse sediment inputs from tributary channels. (B) Weight loss of rock cubes (ranging in size from 20 to 30 mm) in an abrasion tank; weight loss per unit distance is greater for larger cubes because of kinetic energy and surface area effects. (C) Progressive size reduction and roundness increase of a 20 mm chinastone cube during 50 km of travel in an abrasion tank.

be on top of coarser materials laid down in-channel by migrating streams. In more detail, the set of landforms includes channel bars of various types (point bars on the inside of meander bends, together with side- and mid-channel bars exposed at low flows); eroding banks characterized by various types of failure (including particulate removal or mass failure, or beam failure where overhanging banks are undercut); levees (channel-side ridges) and crevasses (dissecting channels) at the channel margin; abandoned channel fills; and backswamp-type floodplain environments in which organic processes may dominate. Each of these has been studied intensively in recent years, with documentation of the style and rate of deposition/erosion and the flow conditions responsible. This has, in the case of in-channel flow and sediment monitoring and modelling, involved painstaking field measurement (sometimes under hazardous flood conditions) and theory development. In other situations, historical records and maps have been used to reconstruct channel changes over a century and longer.

The series of case studies now available exemplifies the general point that hydraulic processes operating in contrasted energy (set by gradient and incident flows) and sedimentary (set by the type and supply volume of sediment of particular size) environments produce contrasted landform developments. For example, rivers with laterally stable courses and high sand load may generate large levees and ponded backswamps behind them, whereas gravel bed streams may have a diversity of bars and bedforms but less well-developed levees and backswamps – especially where rivers are laterally mobile, such as to eliminate the incipient development of such form units.

Architectural Ensembles

Qualitatively recognized channel-pattern types initially included braided, meandering, and straight; anabranching varieties of each (in which the river has more than one branch, with islands or swamps between) are also possible (Figure 7A). The term 'anastomosing' is also used for laterally rather stable anabranching styles. These commonly develop in channels with levees and channel-bed aggradation, with avulsion periodically relocating the river course through a backswamp zone. Braided and meandering channels have also proved to be diverse in style; some of the former have more stable islands between

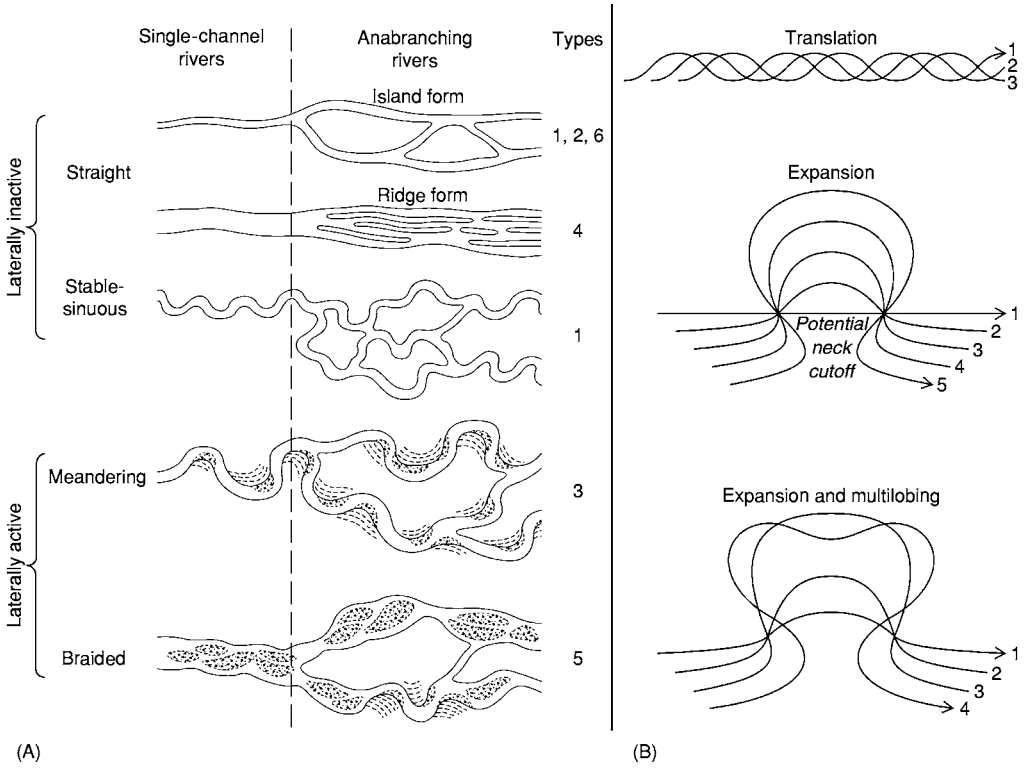


Figure 7 Alluvial river channel patterns. (A) Classification of types. Numbers refer to the styles of anabranching channels plotted in Figure 9. Reproduced with permission from Nanson GC and Knighton AD (1996) Anabranching rivers: their cause, character and classification. *Earth Surface Processes and Landforms* 21: 217 239 © John Wiley and Sons Limited. Reproduced with permission. (B) Forms of development in meandering patterns over time.

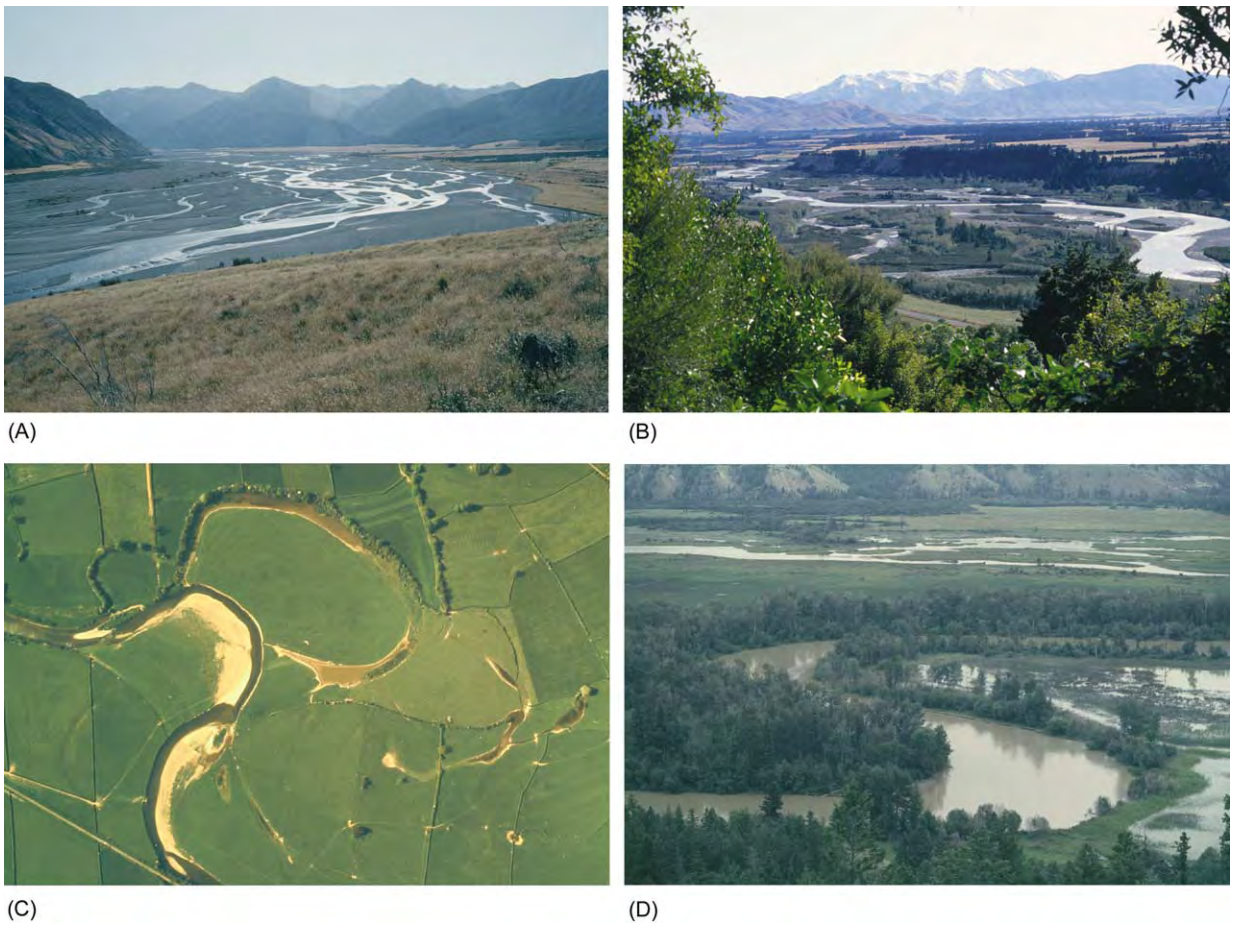


Figure 8 Channel and floodplain styles. (A) The braided course of the Upper Waimakariri River, New Zealand; (B) the Rangitata River at Peel Forest, New Zealand, showing vegetated islands and braid bars; (C) the meandering River Tywi, Wales, UK, showing gravel point bars and water filled cut off channels being filled by fine grained flood sediment; (D) the anastomosing Columbia River, British Columbia, Canada, showing channels, levees with trees, and backswamps (a recent avulsion has created a new disorganized channel in the background).

bar-filled channels (the bars being submerged and migrating at high flows, with scour holes where flow threads join, and deposition at zones of flow divergence) and the latter vary in sinuosity, symmetry, and migration pattern (Figure 7B). Illustrations of four channel-style prototypes are shown in Figure 8.

A similar kind of approach may be taken with architectural ensembles as a whole. Table 2 shows the order and suborder classification of floodplains. The classes are broadly based on energy and sediment size, with subdivision according to factors such as confinement in relatively narrow valleys and the presence of particular sedimentation styles. The observed styles have specified ranges of stream power and grain size, but these may overlap or leave gaps on a stream power/grain size plot (Figure 9). The same also applies to data for rivers designated as braided, meandering, and straight. This is probably because factors other than bankfull power and bed material size are important. Such factors include bank strength (related also to vegetation), the role of overbank flows in bank breaching (as in cut-offs, crevasse formation, and avulsion), and the rate of supply of finer sediment. It is not easy to produce a set of simple process/environment criteria to underpin a qualitative subdivision of ensemble or floodplain types. Such discrimination has been noted, however, on gradient/discharge plots for channel types (Figure 10).

Drainage Basins

Despite local complexities, drainage basins as a whole exhibit general trends in controlling factors from source to mouth (or with increasing drainage area) and therefore in fluvial styles. These are illustrated schematically in Figure 11A, though it should be appreciated that individual catchments may be complex as a result of factors such as lithological

Table 2 A classification of floodplains[a]

Class	*Order*	*Description*
A: High energy non cohesive floodplains	A1	Confined coarse textured
	A2	Confined vertical accretion sandy
	A3	Unconfined vertical accretion sandy
	A4	Cut and fill
B: Medium energy non cohesive floodplains	B1	Braided river
	B2	Wandering gravel bed river
	B3	Meandering, lateral migration
		3a Non scrolled
		3b Scrolled
		3c Backswamp
		3d Counterpoint
C: Low energy cohesive floodplains	C1	Laterally stable single channel
	C2	Anastomosing
		2a Organic rich
		2b Inorganic

[a]Data from Nanson GC and Croke JC (1992) A genetic classification of floodplains. *Geomorphology* 4: 459 486.

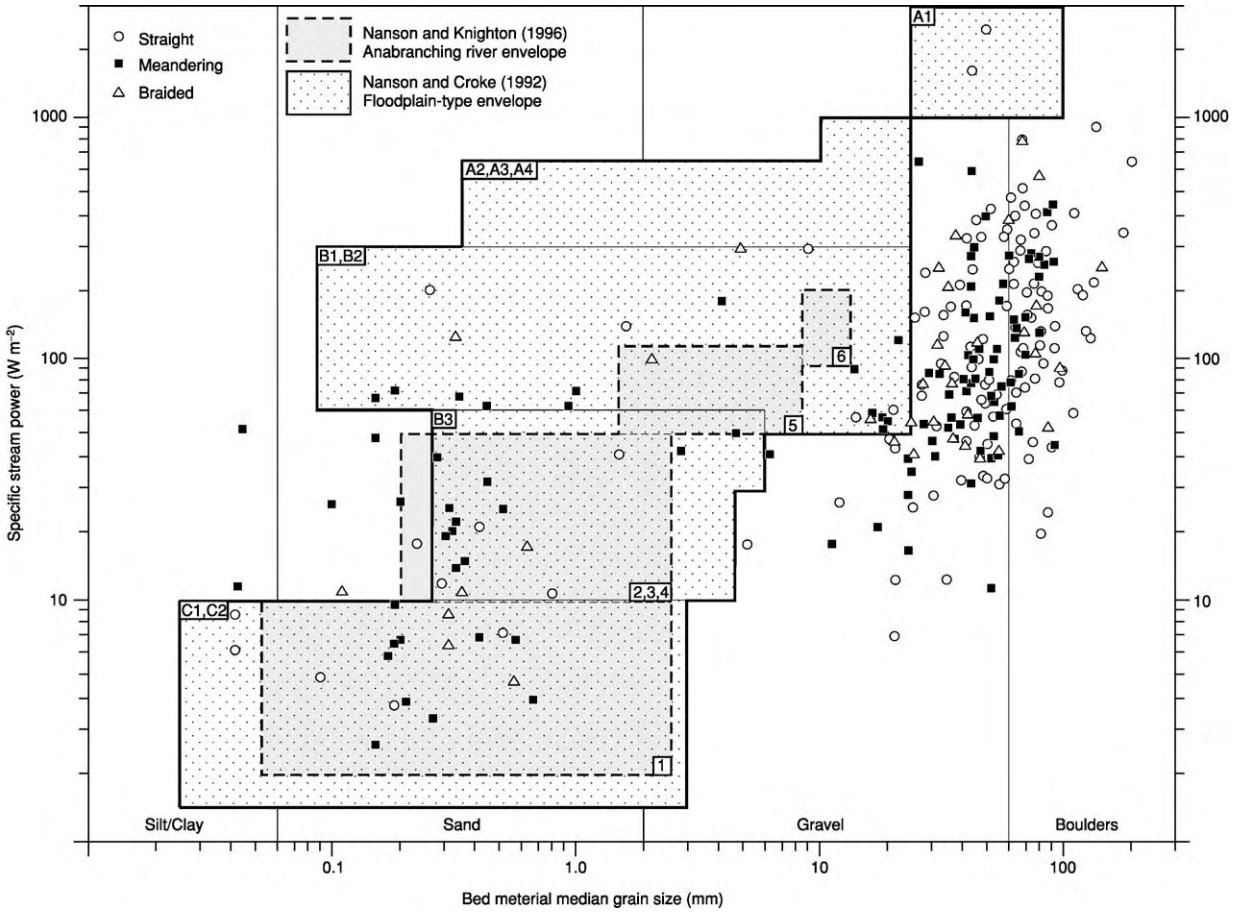


Figure 9 Floodplain types and channel patterns on a bed grain size and unit stream power plot. The envelope boxes overlap and show both the floodplain and the anabranching pattern case study based types. Many of the channel type symbols plot in areas outside the floodplain types, where there may be limited floodplain development, though in part this also reflects a bias in available data towards smaller, steeper streams. The alphanumeric codes and the numbers correlate to the types shown in **Figure 7A** and **Table 2**.

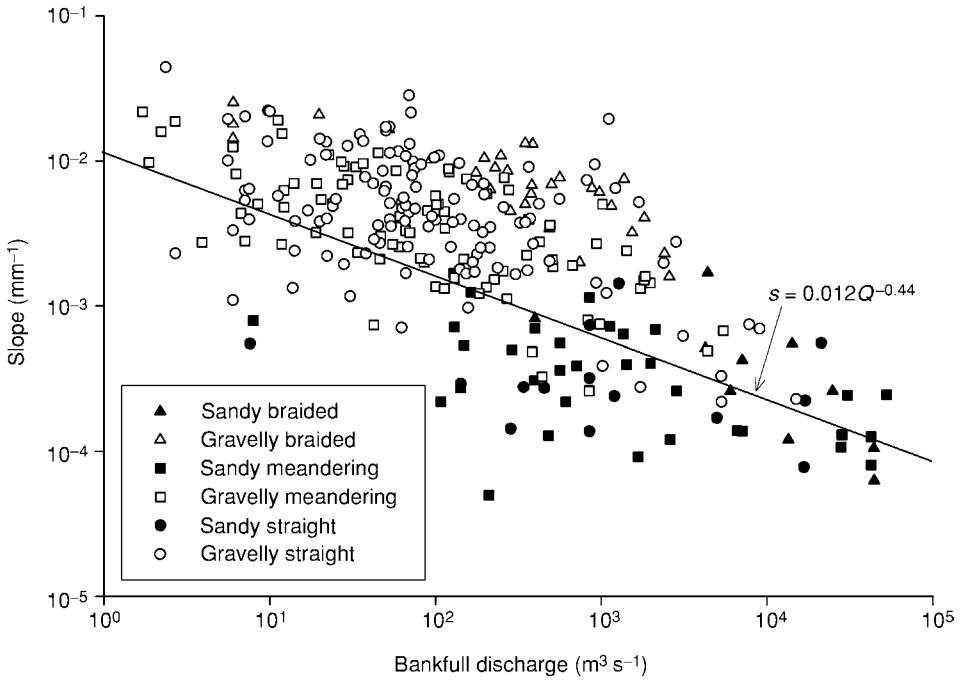


Figure 10 Channel pattern types on a plot of gradient and bankfull discharge in the manner proposed in 1957 by Leopold and Wolman. Their original, but smaller, dataset showed better discrimination, especially between meandering and braided channels. On this plot, sandy channels occur mainly on larger rivers at lower gradients; the threshold between braiding and meandering is higher (and a little more distinct) for gravel bed channels than for sand bed channels. The original Leopold and Wolman threshold is plotted here.

variation or localized tectonic activity leading to gradient modification. But in general, stream gradients decrease and discharge increases with increasing catchment area; stream power (as the product of gradient and discharge) may peak in mid-catchment. Bed grain size decreases downchannel, though this is in practice complicated by factors such as tributary sediment inputs and the existence of step changes in grain size with an abrupt gravel/sand transition. Morphological trends are also apparent (Figure 11B). In headwaters, eroding channels primarily transport hillslope and bedrock erosion materials, which may be coarse (as large or larger than the depth of the channel) and unsorted by river activity. Channels take the form of rapids, pools, and boulder jams. Rare flood events can have dramatic and immediate effects (Figure 12). Lower down-valley, the river sorting and abrasion of transported materials and the increasing dominance of storage and alluvial reworking begin to take effect. Valleys become wider and hillslopes decouple from the river at their foot. In headwaters, channel widths may occupy the whole valley floor, but there is room lower down-valley for rivers to migrate across an increasing body of alluvial storage, with architectural style changing in sympathy with energy, grain size, and other factors. Sediment-transporting events occur more frequently with finer sediment and steadier discharge, though channels may become more fixed in position. In practice, also, the widest valleys may relate to subsiding tectonically defined sedimentation zones in which very thick alluvial (and other) sediments may accumulate, and rivers occupy or migrate across only a fraction of wide valley floors. Here sediment build-up along restricted alluvial ridges allows the periodic relocation of channel belts by avulsion.

Developments in Time

At different scales, it is possible for landforms to achieve quasi-equilibrium with formative factors, exhibit perturbed states in which forms respond and then recover from events/periods of high activity, or evolve unidirectionally towards end-states (in which a highly-organized landscape is reduced to a low-energy and low-relief one following removal of rock mass). These states have been variously approached – initially via qualitative and theoretical proposals (such as WM Davis' influential 'cycle of erosion' proposed more than a century ago, in which uplift initiated drainage development and relief reduction to

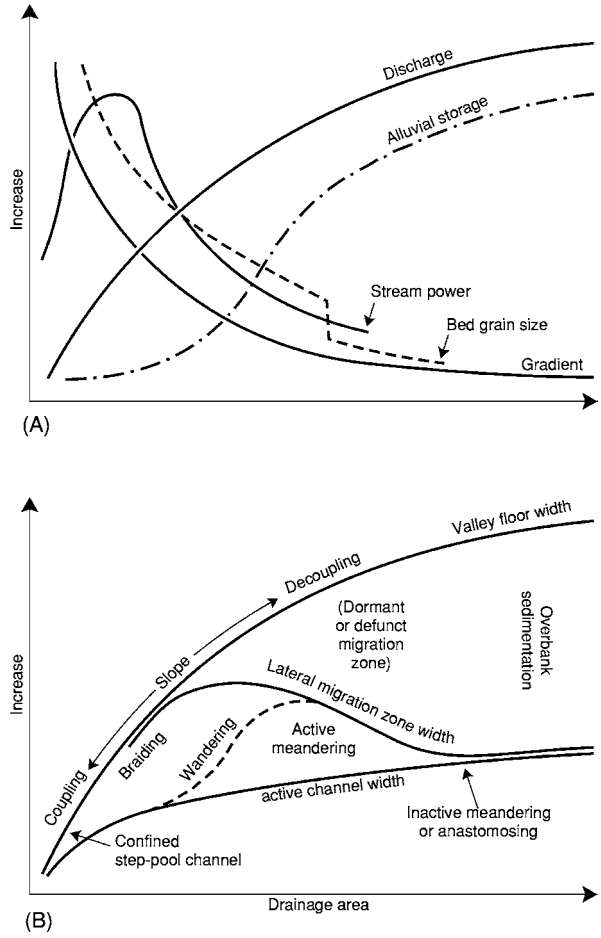


Figure 11 Trends in river catchments (A) in hydraulic factors, based on concepts developed by SA Schumm (in 1977) and M Church (in 2002), and (B) in channel types and activities.

achieve an end-state 'peneplain'). More recently, quantitative modelling has allowed exploration of such developments. Figure 13 illustrates models of drainage evolution; the results can be compared with patterns achieved over very long time-scales in natural landscapes. Such studies may use 'process laws' (based on empirical and/or physically justified relationships) applied iteratively to construct evolving forms. These may be shown to achieve stable pattern states during experimental runs, even though the prediction of individual forms may not be possible (as in the case of individual river bend form and location).

Though unidirectional trajectories in forms have been proposed for higher level landforms, at intermediate scales there has been great concern with perturbations produced by episodic changes in environments (response to a rise or fall in sea-level, tectonic effects, or climate fluctuations that may alter river discharges, sediments, and vegetation cover). Such 'external' changes have also been produced by human agency (including the effects of deforestation and agriculture, mineral extraction, urbanization, and river impounding and channelization). Perturbations over decades to centuries have produced effects that have been closely studied, and these may be ameliorated by good management.

Fluctuations in the rate of catchment erosion and local rates of sediment transfer can result in phases of sediment accumulation and storage, but also of river incision and stored sediment removal. Enhanced delivery may follow climatic change, tectonic uplift, or

Figure 12 Results of an extreme flood in Big Thompson Canyon, Colorado, USA, in July 1976. Photograph by WC Bradley.

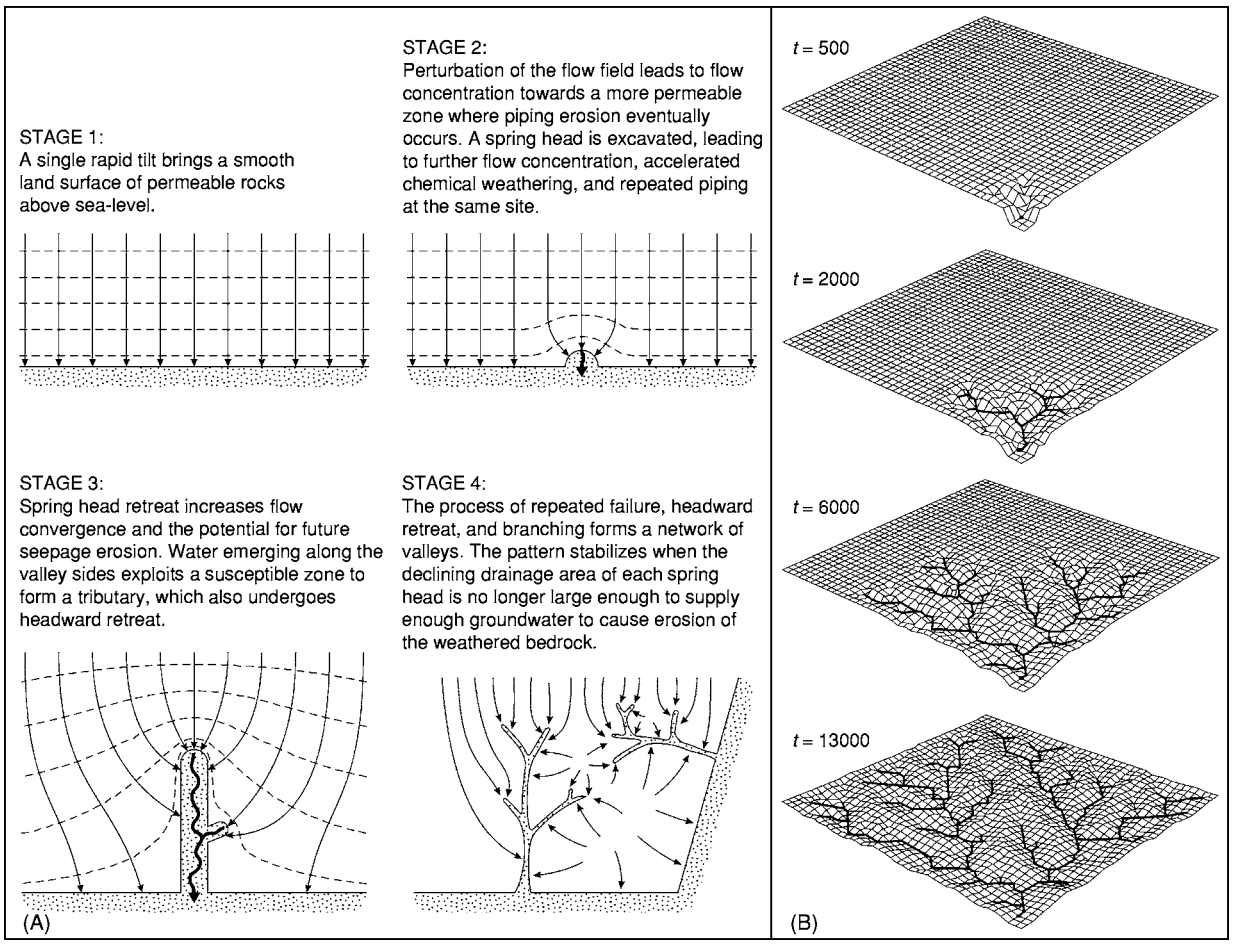


Figure 13 Models of river network development by (A) seepage erosion and (B) a numerical simulation of headward growth. (A) Reproduced with permission from Dunne T (1980) Formation and controls of channel networks. *Progress in Physical Geography* 4: 211 239. (B) Reproduced with permission from Willgoose GR, Bras RL, and Rodriguez Iturbe I (1991) Results from a new model of river basin evolution. *Earth Surface Processes and Landforms* 16: 237 254.

human activity over a range of time-scales. Channel incision into alluvial deposits or planar bedrock valley floors may produce terraces and inset channels and floodplains. Incision episodes have been related to periods of decreased sediment supply from upstream, to tectonic river steepening, and to sea-level fall. Characteristically, channel steepening tends to propagate up-valley, with a knickpoint or zone moving headward at the expense of the former low-gradient valley floor or floodplain. Downcutting streams can produce terraces at a variety of levels as they migrate ('unpaired terraces'), whereas episodes associated with the occurrence of forcing events may give 'paired' or regionally developed terraces that can be matched over long distances. Major research efforts have been devoted to relating Quaternary climatic fluctuations (as in glacial–interglacial cycles) to the flights of terraces associated with many mid-latitude rivers, and to the relationships with cyclical sea-level change over rather longer time-scales (see **Figure 14**). Explanatory models have gradually become more complex as the interrelationships between climatic and vegetation change, sea-level fluctuation and tectonics, and river channel activity have been explored.

Studying the effects of environmental change, whether human or natural, has been greatly aided by improving dating techniques for alluvial sediments and by the availability of historical maps and remote-sensing imagery. Episodes of activity can be pinpointed over appropriate time-scales so that, for example, episodes of Holocene and to some extent Pleistocene river activity can be reasonably closely identified. However, alluvial sedimentation is seldom continuous; rivers also rework their own materials to the extent that what is left is only a 'partial archive of remnants' that leaves much activity unrecorded. Nevertheless, evidence is becoming available to allow change/evolutionary models to be compared not simply with static patterns (for example, drainage networks or meander trains) but

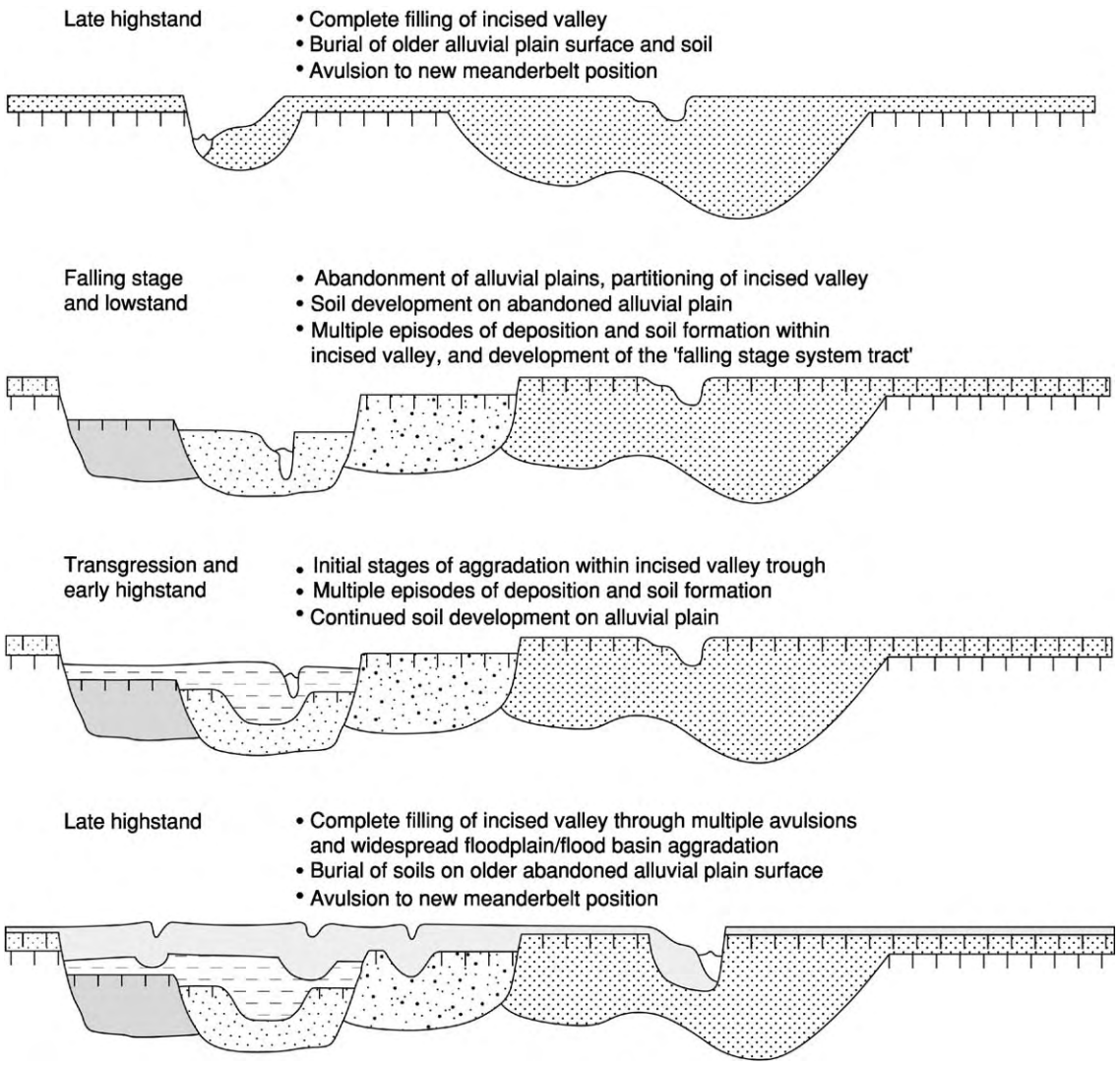


Figure 14 MD Blum and DM Price's 1998 conceptual model of river incision and aggradation in relation to a cycle of sea level change. Reproduced with permission from Bridge JS (2003) *Rivers and Floodplains*. Blackwell: Oxford.

also with development evidence (as in the case of dated alluvial materials in terraces).

See Also

Europe: Holocene. **Geomorphology**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Depositional Sedimentary Structures; Catastrophic Floods; Karst and Palaeokarst; Landslides. **Tertiary To Present:** Pleistocene and The Ice Age. **Weathering**.

Further Reading

Bridge JS (2003) *Rivers and Floodplains*. Blackwell: Oxford.

Brown AG and Quine TA (eds.) (1999) *Fluvial Processes and Environmental Change*. Chichester: Wiley.

Church M (2002) Geomorphic thresholds in riverine landscapes. *Freshwater Biology* 47: 541–557.

Dunne T (1980) Formation and controls of channel networks. *Progress in Physical Geography* 4: 211–239.

Gomez B and Church M (1989) An assessment of bedload sediment transport formulae for gravel bed rivers. *Water Resources Research* 25: 1161–1186.

Hjulström F (1935) Studies of the morphological activities of rivers as illustrated by the River Fyris. *Bulletin of the Geological Institute University of Uppsala* 25: 221–527.

Knighton D (1998) *Fluvial Forms and Processes*. London: Arnold.

Leopold LB and Wolman MG (1957) River channel patterns – braided, meandering and straight. *United States Geological Survey Professional Paper* 282B: 39–85.

Lewin J (1989) Floods in fluvial geomorphology. In: Bevan KJ and Carling PA (eds.) *Floods: Hydrological, Sedimentological and Geomorphological Implications*, pp. 265 284. Chichester: Wiley.
Lewin J and Brewer PA (2002) Laboratory simulation of clast abrasion. *Earth Surface Processes and Landforms* 27: 145 164.
Meybeck M (1976) Total mineral dissolved transport by world major rivers. *Hydrological Sciences Bulletin, International Association of Scientific Hydrology* 21: 265 284.
Milliman JD and Meade RH (1983) World wide delivery of sediment to the oceans. *Journal of Geology* 91: 1 21.
Nanson GC and Croke JC (1992) A genetic classification of floodplains. *Geomorphology* 4: 459 486.
Nanson GC and Knighton AD (1996) Anabranching rivers: their cause, character and classification. *Earth Surface Processes and Landforms* 21: 217 239.
Petts GE, Möller H, and Roux AL (eds.) (1989) *Historical Change of Large Alluvial Rivers: Western Europe*. Chichester: Wiley.
Schumm SA (1977) *The Fluvial System*. New York: Wiley.
Thorne CR, Hey RD, and Newson MD (eds.) (1997) *Applied Fluvial Geomorphology for River Engineering and Management*. Chichester: Wiley.
Willgoose GR, Bras RL, and Rodriguez Iturbe I (1991) Results from a new model of river basin evolution. *Earth Surface Processes and Landforms* 16: 237 254.
Williams GP (1983) Palaeohydrological methods and some examples from Swedish fluvial environments. *Geografiska Annaler* 65A: 227 244.

Glaciers

M J Hambrey and N F Glasser, University of Wales, Aberystwyth, UK

Introduction

Glaciers are amongst the most powerful geological agents on the Earth's surface. They are not only responsible for carving out some of the world's finest landscapes, but also have left widespread deposits that influence agriculture and groundwater flow and provide abundant reserves of sand and gravel for the construction industry. At present, approximately 10% of the Earth's land surface is covered by glacier ice, whilst during Pleistocene glaciations the figure was at least 30%. Continental-scale glaciations occurred in Permo-Carboniferous, Ordovician–Silurian, Neoproterozoic, and Palaeoproterozoic times. The greatest of these glaciations was during the Neoproterozoic, and this has given rise to the concept of the 'snowball Earth', with some geologists arguing for a near-total freeze-up of the planet, followed by severe greenhouse warming, which might have triggered the explosion of life. When the idea of a great Ice Age was first mooted by the Swiss natural historian Louis Agassiz in 1837 (*see* **Famous Geologists:** Agassiz) it met with fierce opposition. In the following decades, increasing numbers of geologists were won over to the concept, as the evidence for glaciation, in the form of characteristic landscapes and extensive surficial deposits ('drift'), throughout Europe and North America became irrefutable. Towards the latter part of the nineteenth century, glacial deposits were being discovered in rocks as old as the Precambrian, although the validity of a glacial interpretation of rocks of Neoproterozoic age was challenged by many until as recently as the 1970s. It was only after glaciologists began examining the processes of erosion and deposition in modern glaciers, from the late 1960s onwards, that the older records of glaciation could be evaluated systematically.

Global Distribution of Glaciers

The global extent of glacier ice has been documented by the World Glacier Monitoring Service (Table 1). The two largest ice-masses, Greenland (10.9%) and Antarctica (85.7%), account for the bulk of glacier ice on Earth, although it is the remaining 3.4% that impinges directly on human activity (Figure 1). The remaining ice-masses occur throughout the Arctic islands and on high mountains, even in the tropics. It is the major ice-masses that are most likely to

Table 1 Distribution of glaciated areas of the world (data from the World Glacier Monitoring Service, 1989)

Region	*Area (km^2)*
Africa	10
Antarctica	13 593 310
Asia and Eastern Europe	185 211
Australasia (i.e. New Zealand)	860
Europe (Western)	53 967
Greenland	1 726 400
North America excluding Greenland	276 100
South America	25 908
World total	15 861 766

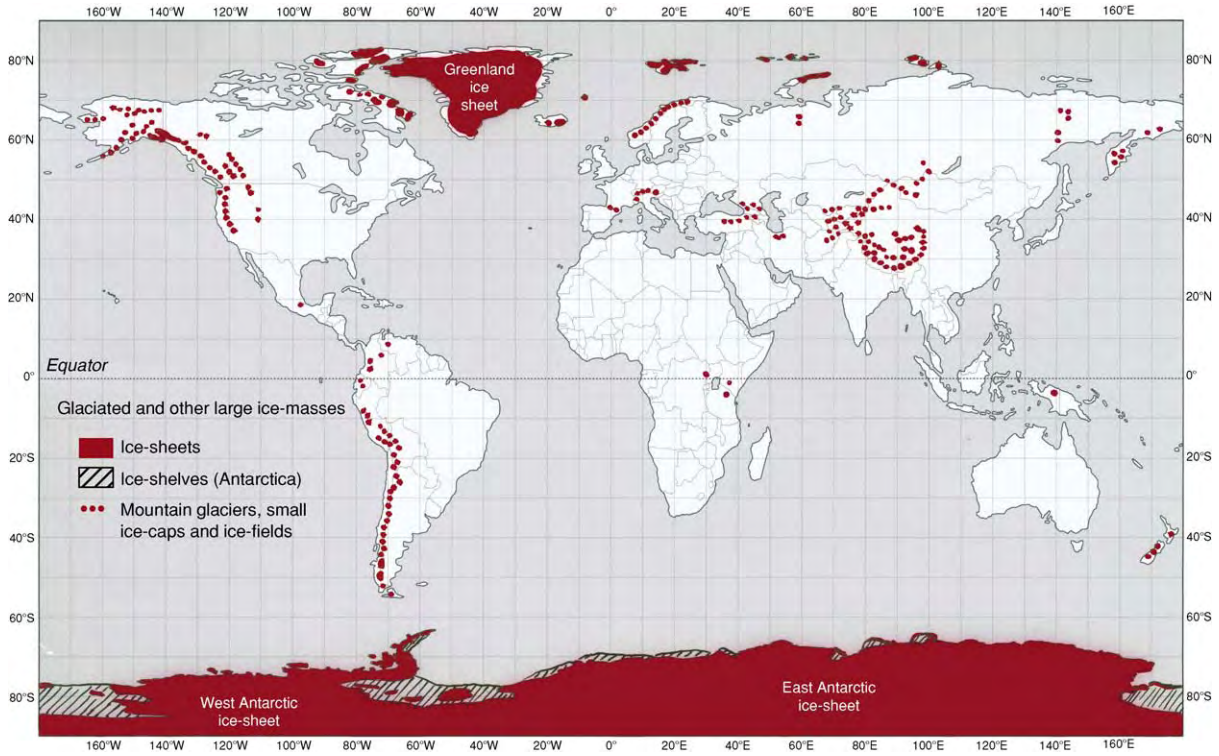


Figure 1 The present day distribution of glaciers; locations are based in part on *Satellite Image Atlas of Glaciers of the World* (US Geological Survey, (2002) Satellite Image Atlas of Glaciers of the World).

contribute to sea-level rise over the next 100 years as the climate warms. Published figures for the potential contribution of glaciers to sea-level rise have large errors, but Table 2 gives an indication of the relative importance of the different ice-masses. The polar ice-sheets will continue to grow and decay in different places at the same time, until the increase in temperature exceeds the threshold at which they can remain in equilibrium.

Table 2 Potential contribution of the world's ice masses to sea level rise (US Geological Survey website, 2001); note that there are considerable uncertainties in the accuracy of these estimates

Ice masses	*Sea level rise (m)*
West Antarctic Ice Sheet	8
East Antarctic Ice Sheet	65
Antarctic Peninsula	0.5
Greenland Ice Sheet	6.5
Mountain glaciers	0.5

Characteristics of Glaciers and Ice-Sheets

Glacier Morphology

Glaciers range in size from the great ice-sheets of Greenland and Antarctica to tiny ice-masses just a few hundred metres across. The generally accepted definition of an ice-sheet is an ice-mass exceeding 50 000 km^2 in area. The Greenland Ice Sheet forms an elongated dome, with mountains projecting through it near the periphery, whilst the Antarctic Ice Sheet consists of three dynamically separate regions, of which the dome of East Antarctica has by far the greatest volume (Figure 2). Ice-caps are also dome-like masses, but measure less than 50 000 km^2 (Figure 3). Both ice-sheets and ice-caps discharge the bulk of their ice via ice-streams – zones of accelerated flow with slow-moving ice on either side. As ice-streams enter the sea, coalescence leads to the formation of ice-shelves – floating slabs of ice several hundred metres thick. Like ice-caps, highland ice-fields also bury much of the topography, but mountains project through them as 'nunataks'. Valley glaciers are typical of mountain regions, such as the Alps, and are the best studied of all ice-masses (Figure 4). Above them are upland amphitheatre-like basins where cirque (or corrie) glaciers accumulate. Lastly, hanging glaciers are perched on steep mountain rock faces and are so prone to avalanches that below them are regenerated glaciers.

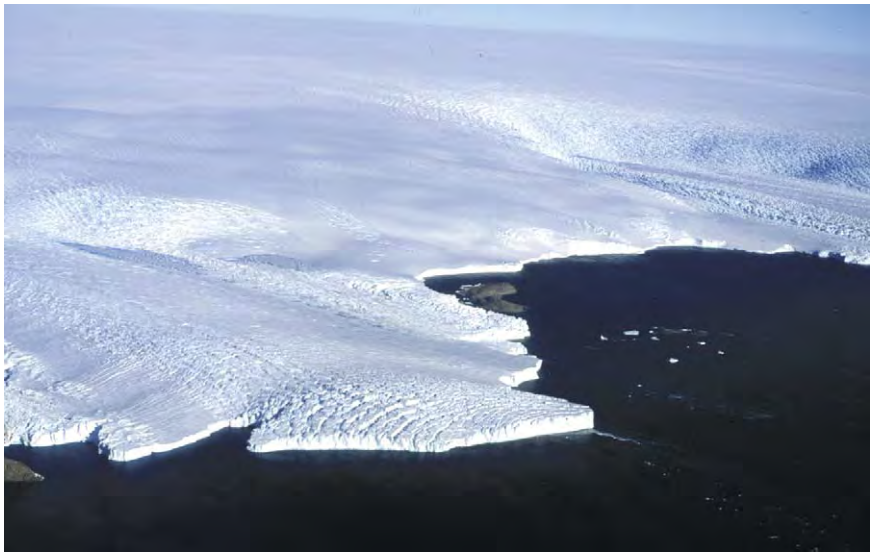


Figure 2 The East Antarctic Ice Sheet near Casey Station, with two small ice streams (zones of crevassed fast flowing ice) discharging into the sea. © 2005, Photograph, M J Hambrey.

Figure 3 A small ice cap on the north east side of James Ross Island, northern Antarctic Peninsula. © 2005, Photograph, M J Hambrey.

Glacier Mass Balance

Mass balance is the study of the inputs and outputs to the glacial system, i.e. the net gain or loss of snow and ice across a glacier (Figure 5). The inputs to the glacial system are known as accumulation and include all materials that add mass to the glacier, such as snow, refrozen slush, hail, frost, rain, and avalanched snow or ice. The outputs from the system are known as ablation and include anything that removes mass from the glacier, such as direct ice melt, iceberg calving, wind erosion, and sublimation. Mass balance is largely dependent on the prevailing climate since this determines the rates of accumulation and ablation. In general, ablation tends to be dominant in the warm summer months, and accumulation is dominant in the winter months (Figure 5B). The exception to this rule is where iceberg calving occurs throughout the year where an ice-sheet or glacier terminates in water. The rate of growth or recession of a glacier is determined by the difference between the accumulation and ablation rates (called the net mass balance and usually measured over a 1 year period). The point

Figure 4 An alpine valley glacier, the Mer de Glace, in the French Alps. © 2005, Photograph, M J Hambrey.

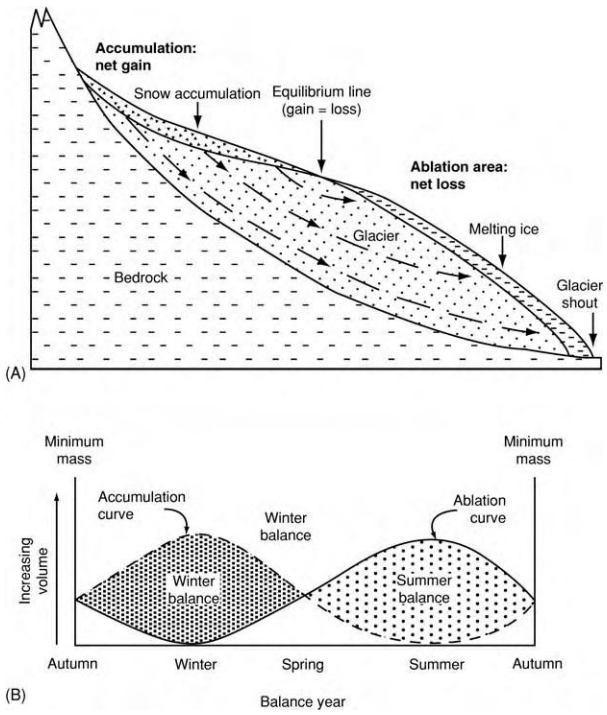


Figure 5 Mass balance regime of a typical glacier in the tem perate regions of the Arctic. (A) Schematic cross section through the glacier, illustrating zones of accumulation and ablation. (B) How the volumes of accumulation and ablation vary over the year, assuming steady state conditions.

on a glacier where accumulation equals ablation is known as the equilibrium line and is normally expressed as an altitude in metres above sea-level. Accumulation exceeds ablation in the upper part of a glacier, whereas the converse is true in the lower part (Figure 5A). Exceptions to this rule occur in parts of Antarctica, where costal snowfall compensates for losses in the middle (drier) reaches of a glacier.

Glacier Thermal Regime

Temperatures within glacier ice vary both vertically and horizontally. Variations in temperature at the base of a glacier are known as the basal thermal regime and are determined by the balance between the heat generated at the base of the glacier and the temperature gradient within the ice, which governs the rate at which heat is drawn towards the ice surface (Figure 6). The most important factors in determining the basal thermal regime are ice thickness, ice surface temperature, geothermal heat, and frictional heat. Heat is generated beneath a glacier by geothermal heat, which enters the base of the glacier from the rocks beneath, and by frictional heat, which is given off as a result of basal sliding and internal deformation. The distribution of ice temperature is vital in understanding the landscapes produced by glaciers because the basal thermal regime is a primary factor controlling the patterns of erosion and deposition.

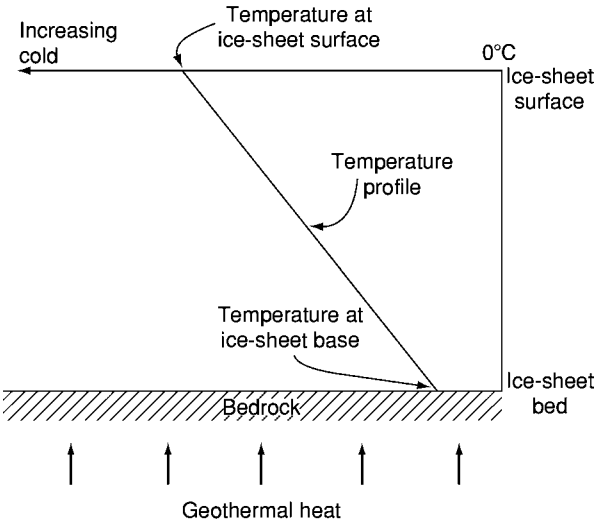


Figure 6 The vertical variation in thermal regime through a polar ice sheet.

This is because the temperature at the base of the glacier controls which of the flow processes (ice creep, basal sliding, and subglacial deformation) can occur and therefore the dynamics of the glacier.

There are three fundamental types of glacier: cold-based, warm-based, and polythermal. Cold-based glaciers are frozen to their beds and produce little basal melt water. Basal sliding cannot occur, and these glaciers flow by internal deformation alone. In contrast, warm-based glaciers produce abundant meltwater that finds its way to the bed; thus basal sliding is an important component of their flow. Warm-based glaciers therefore have the potential for greater flow rates than cold-based glaciers and as a result have a greater ability to modify the underlying landscape. Polythermal glaciers have basal melting in their accumulation zones, basal freezing at their margins, and a transition zone between the two.

Glacier Flow

Glaciers flow primarily because the ice within them deforms under the influence of gravity. Glacier flow is achieved by three mechanisms: internal deformation, basal sliding, and subglacial bed deformation (Figure 7). Internal deformation is achieved by the processes of ice creep, large-scale folding, and faulting. Ice creep is the mutual displacement of ice crystals relative to one another in response to the applied shear stress and results in slow forward motion in the direction of the ice-surface slope. Folding and faulting are common in situations where ice creep alone cannot adjust sufficiently rapidly to the stresses set up within the ice. Basal sliding is the name given to the process whereby a glacier slides over its bed, often on a thin film of lubricating melt water. Sliding is accompanied by a combination of enhanced basal creep and regelation (refreezing) (Figure 8). Subglacial bed deformation is the process by which unfrozen sediment beneath the glacier deforms under the weight of the overlying ice. This occurs when the water pressure in the pore spaces between the sediment grains rises sufficiently to overcome the resistance between individual grains, allowing them to move relative to one another. This is an important process beneath temperate glaciers that overlie poorly consolidated sediments. For example, it has been estimated that 90% of the velocity of Breiðamerkurjökull in Iceland can be accounted for by subglacial bed deformation.

The fastest moving glaciers are typically ice-streams or outlet glaciers terminating in the sea, which may have velocities of several kilometres a year. Mountain-valley glaciers commonly move a few hundred metres a year, whilst small cirque glaciers may move only a few metres a year. Some glaciers alternate between slow-moving ('quiescent') phases and surge phases (when velocities are several orders of magnitude greater than normal). A surge may last for just a few months (as in Alaska) or for several years (as in the high Arctic).

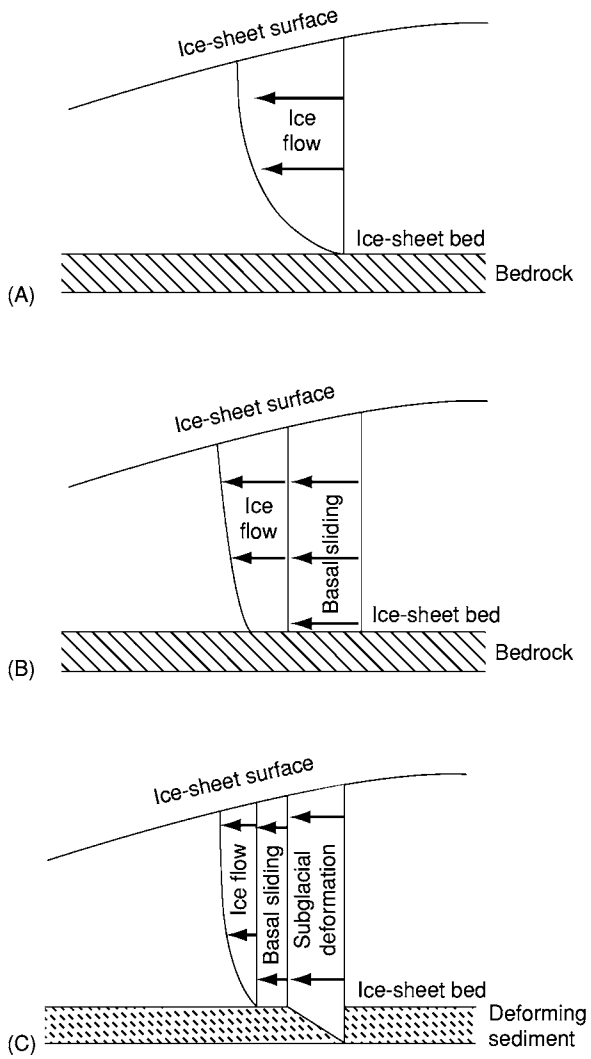


Figure 7 Relative vertical velocity profiles through (A) a cold based ice sheet on bedrock, (B) a warm based ice sheet on bedrock, and (C) a warm based ice sheet on deformable sediment, showing the different components of glacial flow under the different thermal regimes.

Glacier Structure

Glacier ice resembles other geological materials in that it starts as a stratified sequence, in this case mostly of snow, and then undergoes diagenesis (to firn and then to ice) and deformation, producing a wide range of structures. Primary structures include sedimentary stratification, unconformities created in years of

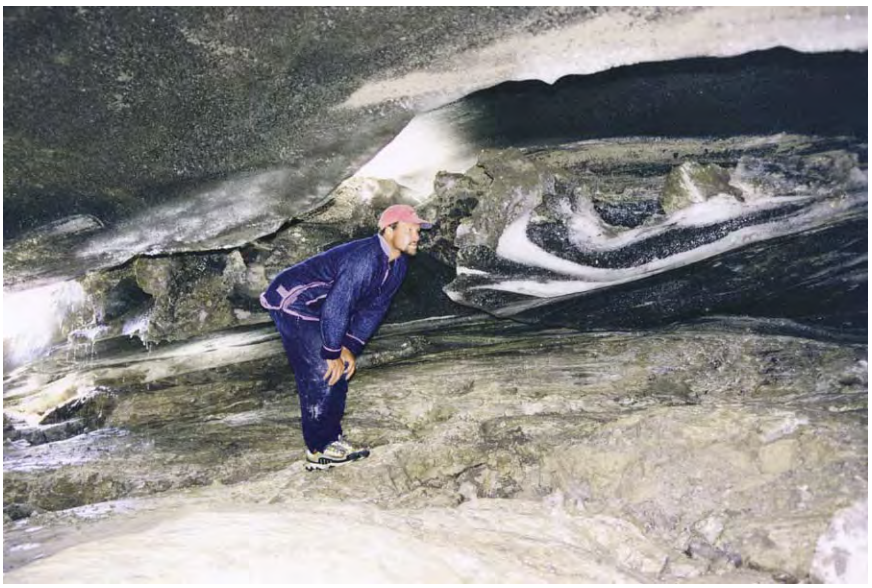


Figure 8 The debris bearing basal ice layer in a cave beneath Glacier de Tsanfleuron, Switzerland, formed by regelation and subsequently modified by folding. © 2005, Photograph, M J Hambrey.

heavy ablation (Figure 9A), and regelation layering resulting from pressure melting and refreezing at the base of the glacier (Figure 8). Secondary structures are the result of internal deformation and include brittle structures, such as crevasses (Figure 9B), crevasse traces, faults, and thrusts, and ductile structures, typified by foliation (Figure 9C), folding (Figure 9D), and boudinage. Structural development in glaciers can be examined in the context of strain rates (usually measured over a 1 year period) and cumulative strain (estimated from velocities and flow-line maps or by finite-element modelling) over several decades.

Glacial Erosion

Processes of Glacial Erosion

Glacial erosion involves the removal and transport of bedrock or sediment by three main processes: quarrying (also known as plucking), abrasion, and melt water erosion. Quarrying involves two separate processes: the fracturing or crushing of bedrock beneath the glacier, and the entrainment of this fractured or crushed rock. Fracturing of bedrock takes place when a glacier flowing over bedrock creates pressure differences in the underlying rock, causing stress fields that may be sufficient to induce rock fracture. Fluctuations in basal water pressure may help to propagate bedrock fractures beneath a glacier. Evacuation of rock fragments or sediment is possible where localized basal freezing occurs, where the tractive force is sufficient to entrain individual clasts, or where debris is elevated along thrusts.

Abrasion is achieved by bodies of subglacial sediment sliding over bedrock or by individual clasts contained within the ice. The progressive breakdown of material in traction provides a mechanism for the continued replenishment of debris that can carry out erosion beneath ice-masses, even where they flow across hard rocks of uniform lithology. Thus, glacial erosion is most efficient beneath temperate ice. Although cold glaciers may achieve limited amounts of erosion, the release of melt water to lubricate basal sliding beneath temperate ice is a prerequisite for widespread and efficient quarrying and abrasion.

Erosion by glacial melt water beneath glaciers is the result of either mechanical or chemical processes. The effectiveness of melt water as an agent of erosion depends on the susceptibility of the bedrock involved (in particular the presence of structural weaknesses or susceptibility to chemical attack), the discharge regime (in particular the water velocity and the level of turbulent flow), and the quantity of the sediment in transport.

Landforms of Glacial Erosion

The landforms produced by glacial erosion are commonly described on three spatial scales: the microscale, the mesoscale, and the macroscale.

Microscale landforms of glacial erosion are those that are below 1 m in size. They are of low-relief amplitude and are often found superimposed on larger landforms. Striae or striations (Figure 10A) are lines or scratches on a rock surface that are produced by glacial abrasion as debris or clasts entrained

Figure 9 Structures in glacier ice: (A) stratification and unconformity in ice cliffs bordering Granite Harbour, Ross Sea, Antarctica; (B) crevasses in Glacier des Bossons, French Alps; (C) foliation in Gornergletscher, Swiss Alps, with the Matterhorn in the background; and (D) 'similar style' folding associated with foliation in Griesgletscher, Swiss Alps. © 2005, Photographs, M J Hambrey.

in the basal ice are dragged over the rock. Microscale crag-and-tails are small elongated bedrock protuberances that form where the rock is preferentially protected from glacial abrasion in the lee of resistant grains or mineral crystals on the surface. Crescentic gouges (Figure 10B) and chattermarks are small cracks, gouges, and indentations created on bedrock surfaces. A variety of different morphologies have been recorded, varying in shape from simple fractures or cracks in the bedrock to larger crescentic gouges in which small chips of bedrock have been physically removed from the bed. Finally, p-forms, s-forms, and microchannel networks include a variety of smooth sinuous depressions and grooves sculpted on glaciated bedrock surfaces (Figure 10C).

Mesoscale landforms of glacial erosion are between 1 m and 1 km in size. Streamlined bedrock features (also known as whalebacks or rock drumlins) are bedrock knolls that have been smoothed and rounded on all sides by glacial action. Stoss-and-lee forms are bedrock features with a pronounced asymmetrical profile: they have an abraded slope on the up-glacier or stoss side and a steeper rougher quarried slope on the down-glacier or lee side. Individual landforms

Figure 10 Microscale erosional landforms: (A) striated and ice moulded limestone at the margin of Glacier de Tsanfleuron, Swiss Alps; (B) crescentic gouges in dolerite beneath Neogene glacigenic sediments, Roberts Massif, Antarctica (ice flow was from left to right); and (C) p forms (melt water eroded and glacially striated channels) cut in limestone, Glacier de Tsanfleuron, Swiss Alps. © 2005, Photographs, M J Hambrey.

are commonly known as *roches moutonnées* (Figure 11A). Rock grooves are similar in morphology to striae, but have greater size and depth (Figure 11B). They range from tens of metres to hundreds of metres in length and may be up to several metres wide and a few metres deep. Rock basins are individual depressions carved in the bedrock, often developed along lines of structural weakness. Melt-water channels span a large spatial scale from centimetre- or metre-scale channels cut in soft sediment or bedrock to large valley-scale melt-water channel systems. Landforms associated with glacial melt water form wherever sufficient quantities of water are released in subglacial, ice-marginal, and proglacial settings.

Macroscale landforms of glacial erosion are those features that are at least 1 km in size. They are significant landscape components, which may contain within them many of the smaller landforms described above. Zones of areal scouring are scoured bedrock consisting of an assemblage of streamlined bedrock features, rock basins, and stoss-and-lee forms. Glacial troughs are deep linear features carved into the bedrock, which represent the effects of glacial erosion where ice flow has been confined by topography and channelled along a valley or fjord (Figures 12A and 12B). Cirques are large bedrock hollows that open down-slope and are bounded up-slope by a cliff or steep slope known as a headwall (Figure 12C). Headwall erosion that continues from two opposing sides until they meet produces a sharp ridge – an arête (Figure 12D). If the process takes place from three or more sides, a sharp isolated peak – a horn – may be the product (Figure 12E). Large stoss-and-lee forms (giant *roches moutonnées*) are between hundreds of metres and several kilometres in size. These valley spurs and hills have a pronounced asymmetric profile, with an abraded slope on the up-glacier (stoss) side and a steeper rougher quarried slope on the down-glacier (lee) side.

Bathymetric Forms Resulting from Glacial Erosion

Various erosional phenomena, mainly associated with grounded ice or subglacial melt water, are found in marine settings. The larger-scale forms are filled by sediment and may be recognizable only in

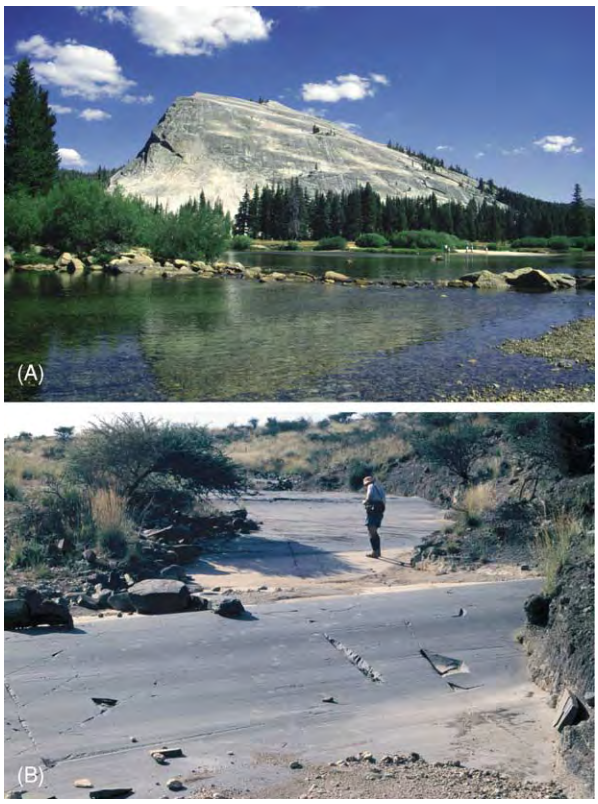


Figure 11 Mesoscale landforms of glacial erosion: (A) *roche moutonnée*, the Lembert Dome, Yosemite, California, USA; (B) shallow striated grooves from the Permo Carboniferous glaci ation, near Douglas, Karoo, South Africa. © 2005, Photographs, M J Hambrey.

seismic profiles. Submarine troughs formed by ice-streams are found on continental shelves and are genetically equivalent to fjords and other glacial troughs but are generally much broader. The largest occur in Antarctica, where they can reach over 400 km in length, 200 km in width, and 1100 m in depth. Steep-sided channels 2–3 km wide, over 100 m deep, and up to 100 km long, carved out by subglacial melt water and subsequently filled by sediment, are known as tunnel valleys. These are well known from the north-west European continental shelf around Britain, the Scotian Shelf off Canada, and Antarctica. Icebergs can also cause considerable erosion if they become grounded on the seafloor. Large tabular icebergs can scour the bed of the sea for several tens of kilometres, leaving impressions up to 100 m wide and several metres deep.

Debris Entrainment and Transport

Glacial erosion and subaerial erosion above a glacier produce a range of sediment types that can be entrained and transported in a variety of ways. The manner in which sediment is transferred through the glacial system is related to the dynamic and thermal characteristics of the glacier. Broadly speaking, polythermal glaciers tend to carry a high basal debris load but little on their surfaces, whereas the reverse is true for temperate glaciers.

As noted above, entrainment of debris at the bed is a product of regelation. Basally derived debris is subject to comminution at the ice–bedrock interface and typically consists of clasts up to boulder size, with subangular and subrounded shapes and faceted and striated surfaces. In addition, there are varying proportions of sand, silt, and clay. Debris on the surface of the glacier (supraglacial debris) is mainly derived from frost-induced rock fall from the crags above and is mainly angular with few fines.

During flow, both supraglacial and basal debris is reorganized. In addition to regelation, the following mechanisms, some of which are illustrated in Figure 13, have been identified

- incorporation of angular rock-fall material within the stratified sequence of snow, firn, and refrozen slush, followed by folding (this debris emerges near the snout at the hinges and upper limbs of the folds, forming medial moraines);
- falling or washing of supraglacial debris into crevasses and moulins;
- incorporation of both supraglacial and basal debris within longitudinal foliation;
- thrusting, allowing debris-rich basal ice and subglacial sediments to be uplifted into an englacial or even supraglacial position; and
- squeezing of subglacial debris into basal crevasses.

Glacial Deposition

Processes of Glacial Deposition

Glacial deposition involves the release of sediment from supraglacial, englacial, and basal-glacial positions. Debris that has been in contact with the bed is heavily modified by intraclast collision, subglacial deformation, and contact with water. Debris that follows a passive transport path (supraglacially or englacially) tends to retain its original characteristics.

Sediment may be deposited directly beneath the glacier or at its margins, or it can be transported long distances away from the glacier by other agents, such as rivers (producing outwash plains) or icebergs. During release from the ice, sediment is reworked by numerous processes, including fluvial processes, debris-flowage, marine or lacustrine currents, and aeolian processes (Figure 14A). The resulting sedimentary facies associations are therefore exceedingly complex, but particular combinations of facies can be

Figure 12 Macroscale glacial erosional landforms: (A) Yosemite glacial valley, California, USA; (B) Kejser Franz Josef Fjord, central east Greenland; (C) a pair of deeply incised cirques on Cuernos del Paine, Chilean Patagonia; (D) arêtes carved in granite peaks, Stauning Alper, East Greenland; and (E) the horn like peak of Ama Dablam, Khumbu Himal, Nepal. © 2005, Photographs, M J Hambrey.

used to determine the glacial thermal regime and tectonic setting. A temperate glacial system in an area of high relief, especially if it is tectonically active, will have a high proportion of supraglacial debris that on release is modified by streams; little sediment of basal character will be preserved. A polythermal glacier tends to produce a mainly basal glacial deposit, only part of which may be reworked by outlet streams.

Glaciers that enter water bodies behave differently according to their thermal regime. Polythermal glaciers tend to have a floating tongue if the water is deep enough, as in the fjords of Greenland, and the dominant facies produced is poorly sorted sediment ('diamicton') derived from the base of the glacier. The ice in temperate glaciers is usually too weak to float, and the glacial terminus is a cliff resting on the seabed. Sediments produced in these situations, typified by the fjords of Alaska, are dominated by subglacial water output (Figure 14B). For the coldest glaciers, terminating as ice-shelves on the continental shelf of Antarctica, direct deposition is restricted to a narrow grounding zone because freeze-on of saline ice at the base of the ice-shelf seals in any remaining sediment. Melt-water production in the grounding zone is generally limited, and the dominant sediment

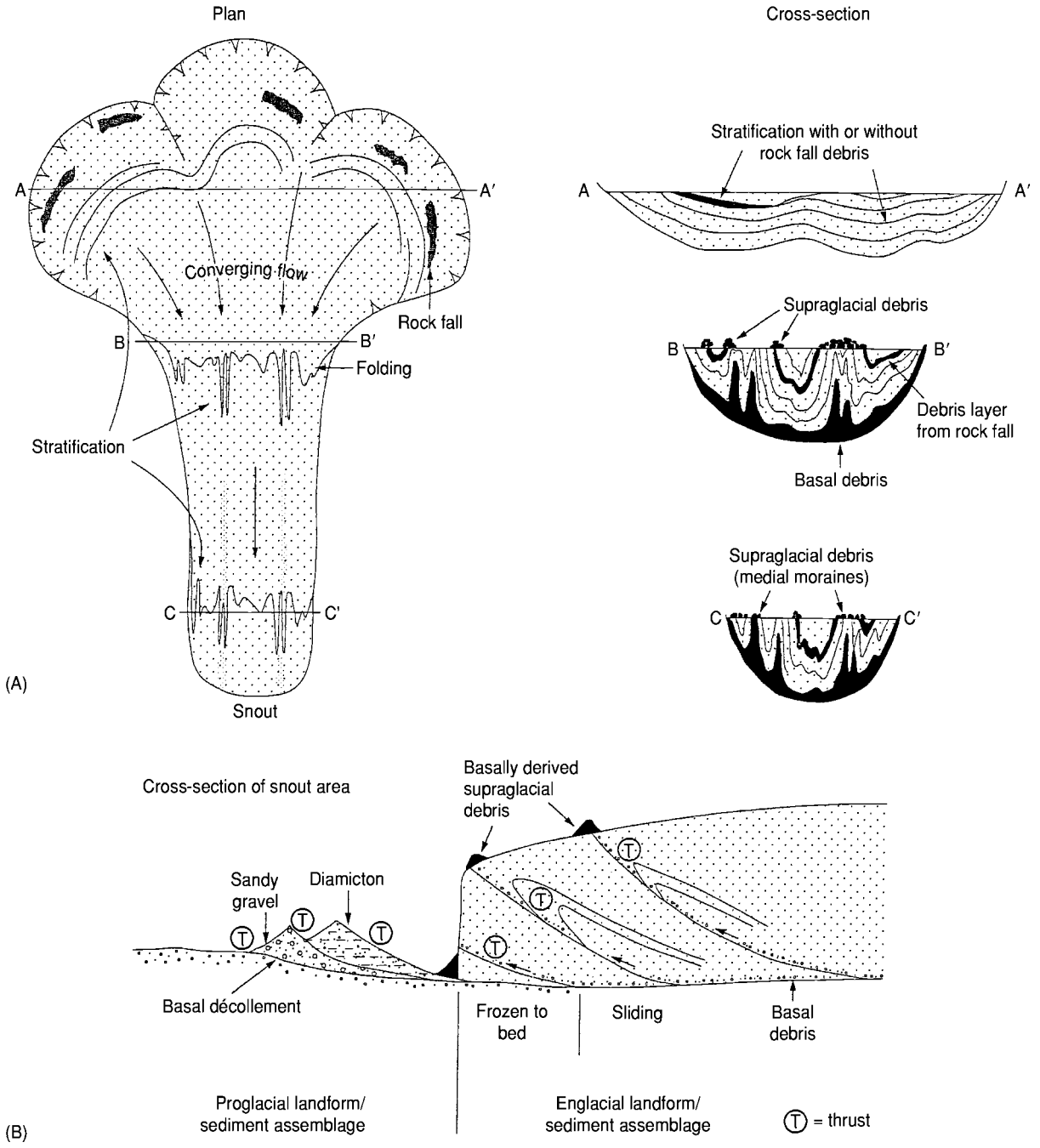


Figure 13 Conceptual models of debris entrainment by (A) folding and (B) thrusting, based on glaciers with multiple accumulation basins in Svalbard. (Reproduced from Hambrey MJ and Glasser NF (2003) Glacial; sediments: processes, environments and facies. In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 316 331. Dordrecht: Kluwer.)

immediately beyond this area is biogenic and is typically derived from the settling out and current transfer of diatom remains (Figure 14C). This sediment is carried away from the Antarctic coast within icebergs, and some may be released onto the ocean floor hundreds of kilometres from the source.

Glaciotectonic deformation is a widespread phenomenon that affects materials for up to several hundred metres adjacent to and beneath glaciers. Processes such as thrusting and folding within the ice are mirrored by the same processes in subglacial and proglacial settings. All types of material are affected, including wet, dry, and frozen sediments and bedrock. Deformation may detach blocks of rock and frozen sediment that are sometimes hundreds of metres across, leaving depressions in the

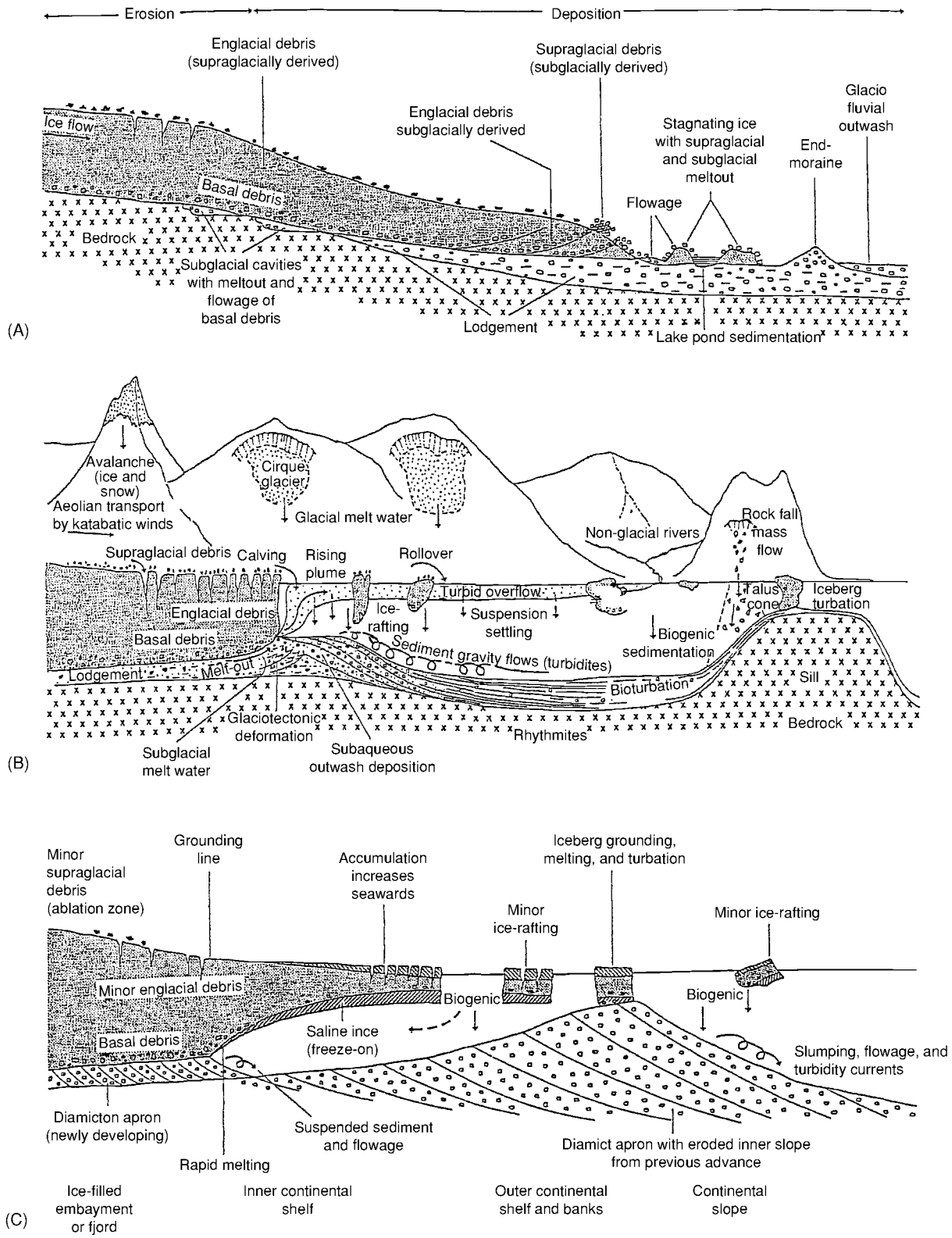


Figure 14 Schematic cross sections through glaciers showing deposition in different geographical and climatic settings: (A) a terrestrial temperate polythermal glacier; (B) a temperate tidewater glacier in a fjord; and (C) an ice shelf and continental shelf (Antarctica). (Reproduced with permission from Hambrey MJ (1999) The record of Earth's glacial history during the last 3000 Ma. In: Barrett PJ and Orombelli G (eds.) *Geological Records of Global and Planetary Changes*, pp. 73 107. Terra Antartica Report 3. Siena: Museo Nazionale dell'Antartide.) Kluwer Academic Press.

Figure 15 Typical glacigenic facies: (A) a diamicton interpreted as a basal till on the Isle of Arran, Scotland; and (B) sand and gravel interpreted as a glaciofluvial sediment in Rhosesmor Quarry, North Wales (the reddish colour derives from eroded New Red Sandstone in the Irish Sea). © 2005, Photographs, M J Hambrey.

ground and producing complex ridge forms in front of the glacier. Much of the material is brecciated, but, if it is wet and fine grained, it may deform in a ductile manner.

Glacigenic Sediments

A number of general terms have been applied to deposits left behind by glaciers and the associated melt water. 'Glacigenic sediment' means 'sediment of glacial origin' in the broad sense, embracing sediments that have been influenced by glacier ice even if only slightly. 'Glacial drift' refers to all material being transported by glacier ice, all deposits released by glaciers, and all deposits of predominantly glacial origin deposited from glacial melt water or in the sea by icebergs.

Traditionally, Quaternary researchers have tended to give genetic terms to glacigenic sediments in the field, but this has led to difficulties where the meanings of the terms have not been agreed and where the origin of the sediment is disputed. The increasingly preferred approach is to describe sediments objectively using textural terms, and classifications are available for the typical range of poorly sorted sediments that exist in glacial settings. The terms 'diamicton' (unlithified), 'diamictite' (lithified), and 'diamict' (both) are used for 'a non-sorted or poorly sorted terrigenous sediment that contains a wide range of particle sizes'. In other words diamicts contain particles resulting from weathering and erosion ranging in size from clay to boulders in various proportions (Figure 15A). We use a range of 1% to 50% for the proportion of clasts within diamicts (from a detailed textural classification presented elsewhere).

The International Quaternary Association (INQUA) has proposed a genetic classification that, although somewhat dated, embraces the terminology used by most glacial geologists. The principal terms are as follows.

- Lodgement till is deposited by 'plastering' of glacial debris from the sliding base of a moving glacier as a result of pressure melting or other mechanical processes. Texturally, the deposit is a diamicton or a muddy or sandy gravel.
- Meltout till is deposited by the slow release of glacial debris from ice that is not moving. Typically, meltout tills inherit the characteristics of the sediment within the basal ice layer and are diamictons.
- Basal till embraces both the above.
- Sublimation till is released by the sublimation of debris-rich ice in subzero temperatures. Texturally this is diamicton, but it is found only in arid parts of Antarctica.
- Deformation till comprises weak rock or unconsolidated sediment that has been deformed or disaggregated, with some foreign material admixed. Texturally this is variable and often breccia-like; rafts of rock or sediment may be many metres in length.
- Flow till is till that has been remobilized by gravity flow, although it is better known as glacigenic sediment flow.

The above INQUA classification refers to modern terrestrial glacigenic sediments. The lithified equivalent of till is 'tillite'. All these deposits are prone to reworking by melt water and wind. Thus extensive well-sorted sand and gravel deposits are a major feature of glacial environments (Figure 15B). There is no agreed genetic classification for glaciomarine sediments. The following terms, however, are commonly used.

- Ice-proximal glaciomarine sediment is sediment that is released either from floating basal glacier

ice or by continuous rainout from icebergs, without subsequent winnowing by currents and waves. Texturally this is normally a diamicton.

- Ice-distal glaciomarine sediment is sediment that is released from icebergs, but is subject to winnowing and admixed with other marine sediment, including biogenic components. Texturally these sediments are usually sandy muds or muddy sands with dispersed stones, and in modern environments they are often rich in diatoms.
- Iceberg turbate is sediment that is deposited on the seafloor and subsequently reworked by grounded icebergs. The end product is commonly a massive diamicton.
- Cyclopels (silt–mud couplets) and cyclopsams (graded sand–mud couplets) are rhythmically laminated sediments derived from the turbid overflow plumes of subglacial discharge, especially in fjords. Iceberg rafting may produce dropstone structures.

Glaciolacustrine facies are similar to those in glaciomarine environments. In addition, lakes commonly have rhythmically laminated sands and muds called varves (or varvites if lithified). Sand laminae represent input to the lake during summer, and silt or clay laminae represent settling out from suspension in winter. Dropstones may disrupt the laminae. Other laminated sediments in lakes or the sea may be deposited from turbidity currents and also include dropstone structures.

Depositional Landforms

Glaciers and ice-sheets produce a large range of depositional landforms that are commonly grouped into ice-marginal and subglacial landforms.

Ice-marginal landforms Ice-marginal landforms are produced by advancing, static, or receding ice-margins, as well as during seasonal fluctuations of an ice-front. They may be deposited directly from glacier ice or be composed of sediments previously deposited by other processes. Ice-marginal landforms are commonly used to reconstruct changes in glacier size, morphology, and extent over time.

Glaciotectonic moraines encompass a broad range of different types of moraine formed by the deformation of ice, sediment, and rock. Of these, the most common are push moraines, which are formed when a glacier flows into sediment and bulldozes material into a ridge.

Dump moraines are formed at stationary or near-stationary ice-fronts, where debris accumulates along the margin or front of the glacier to form a ridge of sediment (Figure 16A).

Ablation moraines (sometimes referred to as ice-cored moraines) form wherever ice-melt is retarded beneath a cover of supraglacial debris. This supraglacial debris may be derived from rock falls and avalanches in the accumulation area, in which case it is often arranged in flow-parallel medial moraines, or it may be composed of subglacial or englacial debris elevated to the ice surface by folding, thrusting, or upward-directed flow lines in the ablation area. Uneven melting gives rise to hummocky moraines (Figure 16B).

Outwash fans and outwash plains (or sandar) are formed where glacial melt water emerges from the glacier and sediment is deposited at or beyond the ice margin, producing extensive braided river tracts composed of sand and gravel. Kettle holes form where sediment is deposited over buried ice.

Kame terraces are formed when sediment is deposited by melt water flowing laterally along an ice-margin. Kames are more fragmentary features, which are formed in a similar manner but often in ice-walled tunnels and against steep valley sides.

Subglacial landforms Subglacial landforms (sometimes referred to as subglacial bedforms) are produced beneath actively flowing ice. The most common of these are a family of ice-moulded landforms, all of which are elongated parallel to the direction of ice flow.

Flutes are low flow-parallel ridges that are rarely continuous for less than 100 m (Figure 16C). Megascale lineations are much larger (tens of kilometres in length and hundreds of metres in width) and are often visible only on satellite imagery. Drumlins are typically smooth oval or elliptical hills that are tens of metres high and long. The origin of drumlins is unclear, but they have been attributed to deposition by actively moving ice and to subglacial sheet floods.

Ribbed moraines (also known as Rogen moraines) are large closely spaced moraine ridges, which are often curved or anastomosing but generally orientated transversely to the direction of ice flow. They may represent the fracturing of a pre-existing subglacial till sheet at the transition from cold-based to warm-based ice.

Geometric ridge networks and crevasse-fill ridges are low subglacial landforms that are composed of material derived from the basal debris layer of the glacier, or material beneath the glacier. They form as a result of the squeezing of subglacial material into basal crevasses or former subglacial tunnels, commonly during surges. Alternatively, geometric ridge networks can also form beneath glaciers as a result of the intersection of foliation-parallel ridges and englacial thrusts.

Figure 16 Glacial depositional landforms: (A) terminal dump moraine (left) of Chuchung Glacier (breached by a stream) and lateral moraine (right) of Ama Dablam Glacier, Khumbu Himal, Nepal (an older moraine with person for scale in the foreground); (B) hummocky moraine formed by the thrusting of subglacial sediment into an englacial position before release, Kronebreen, Svalbard; (C) flutes on proglacial area of Austre Levénbreen, north west Spitsbergen; (D) esker (sinuous ridge of sand and gravel) emerging from beneath Comfortlessbreen, a tidewater glacier in western Spitsbergen. © 2005, Photographs, M J Hambrey.

Eskers are glaciofluvial landforms that are created by the flow of melt water in subglacial, englacial, or supraglacial channels. They are usually sinuous in plan view and composed of sand and gravel (Figure 16D).

Bathymetric Forms Resulting from Glacial Deposition

The morphology and sediment composition of subaquatic features, particularly in fjords and on continental shelves, are less well known than their terrestrial counterparts. As on land, subaqueous depositional assemblages reflect the interaction of a wide range of processes.

Ice-contact features form when a glacier terminus remains quasi-stationary in water, particularly in fjords. Morainal banks form as a result of a combination of lodgement, meltout, dumping, push, and squeeze processes combined with glaciofluvial discharge; poorly sorted deposits are typical of such features. Grounding-line fans extend from subglacial tunnels that discharge melt water and sediment into the sea, and are typically composed of sand and gravel. Developing out of grounding-line fans are ice-contact deltas, which form when the terminus remains stable long enough for sediment to build up to the surface of the fjord. Where a glacier becomes disconnected from the water body, alluvial sediments may prograde to form fluviodeltaic complexes, again dominated by sand and gravel. In addition to these large-scale forms, there are small-scale features that are found particularly on beaches, such as iceberg toolmarks, ridges, depressions from the melting of buried ice, bounce marks, chatter marks, and roll marks. Icebergs can also churn up submarine sediment, particularly on shoals, producing iceberg turbates.

Depositional forms on continental shelves are best known from a combination of deep drilling and seismic profiling. Some, as in Alaska, are simply larger-scale analogues of fjordal features, such as fan-delta complexes, but others show evidence of having formed under floating tongues that are typical of the colder ice of the polar regions. These include subglacial deltas, till tongues, diamicton aprons, and the immense (up to 400 km wide) trough-mouth fans.

Other features are similar to those on land and include shelf moraines, flutes, and transverse ridges.

Conclusions

Glaciers and ice-sheets can be highly effective agents of landscape change, eroding, transporting, and depositing large amounts of material at a variety of spatial scales, from a valley scale to a continental scale. The rates of operation of these processes are strongly controlled by glacier mass balance and thermal regime, which together determine glacier velocity, melt water production, ice deformation styles, and sediment routing through the ice-mass. Glaciers produce a diverse range of landforms and sediments. The classic glacial erosional landforms, such as cirques, glacial troughs, stoss-and-lee forms, and areas of striated bedrock, are generally carved in solid bedrock, although glacial erosional landforms such as melt water channels and tunnel valleys can also form in unconsolidated sediments. Glacial depositional landforms take a variety of forms and include ice-marginal landforms, glaciotectonic moraines, hummocky moraines, outwash fans and plains, kames and kame terraces, and subglacial bedforms. These landforms are typically composed of a range of sediment types, ranging from poorly sorted sediments such as diamicton to well-sorted sand and gravel. Glacial sediments may be deposited in both terrestrial and marine settings, although those deposited on continental shelves and slopes probably have the best preservation potential in the geological record.

See Also

Famous Geologists: Agassiz. **Geomorphology**. **Palaeoclimates**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings; Deltas; Lake Processes and Deposits. **Sedimentary Processes:** Erosional Sedimentary Structures. **Sedimentary Rocks:** Mineralogy and Classification; Rudaceous Rocks.

Further Reading

Anderson JB (1999) *Antarctic Marine Geology*. Cambridge: Cambridge University Press.

Benn DI and Evans DJA (1999) *Glaciers and Glaciation*. London: Arnold.

Bennett MR and Glasser NF (1996) *Glacial Geology: Ice Sheets and Landforms*. Chichester: John Wiley and Sons.

Dowdeswell JA and O'Cofaigh C (eds.) (2002) *Glacier Influenced Sedimentation in High Latitude Continental Margins*. Special Publication 203. London: Geological Society.

Dreimanis A (1989) Tills, their genetic terminology and classification. In: Goldthwait RP and Matsch CL (eds.) *Genetic Classification of Glacienic Deposits,* pp. 17 84. Rotterdam: Balkema.

Eyles N (1993) Earth's glacial record and its tectonic set ting. *Earth Science Reviews* 35: 1 248.

Hambrey MJ (1994) *Glacial Environments*. Vancouver and London: University of British Columbia Press and UCL Press.

Hambrey MJ (1999) The record of Earth's glacial history during the last 3000 Ma. In: Barrett PJ and Orombelli G (eds.) *Geological Records of Global and Planetary Changes,* pp. 73 107. Terra Antartica Report 3. Siena: Museo Nazionale dell'Antartide.

Hambrey MJ and Alean J (2004) *Glaciers,* 2nd edn. Cam bridge: Cambridge University Press.

Knight P (1999) *Glaciers*. Cheltenham: Stanley Thornes (Publishers) Ltd.

Maltman AJ, Hubbard B, and Hambrey MJ (eds.) (2000) *Deformation of Glacial Materials*. Special Publication 176. London: Geological Society.

Menzies J (ed.) (1995) *Modern Glacial Environments: Pro cesses, Dynamics and Sediments*. Oxford: Butterworth Heinemann.

Menzies J (ed.) (1996) *Past Glacial Environments: Sedi ments, Forms and Techniques*. Oxford: Butterworth Heinemann.

Paterson WSB (1994) *Physics of Glaciers,* 3rd edn. Oxford: Pergamon.

Post A and LaChapelle ER (2000) *Glacier Ice*. Seattle and Cambridge: University of Washington Press and The International Glaciological Society.

Karst and Palaeokarst

M J Simms, Ulster Museum, Belfast, UK

Introduction

Karst is a term applied to terrain with distinctive landforms and underground drainage systems that form through the greater solubility in water of certain rock types, particularly limestone. Karst landscapes are sculpted largely by solution, other rock types largely by mechanical erosion. The word 'karst' has its origins in pre-Indo-European languages, from *kar*, meaning 'rock'. In Slovenia the word 'kras' (or 'krs'), subsequently germanicized as 'Karst', derives from the name of a barren stony limestone area near Trieste, which is still considered the type area for limestone karst. Gypsum and rock salt may form karst,

with extensive gypsum karsts known from Russia and the Ukraine, but their greater solubility renders such landforms more dynamic and, for rock salt, ephemeral in all but the most arid climates. Karst features may also develop, though rarely, on very weakly soluble rocks, such as basalt, granite, or quartzite.

Rock solubility and water are the primary factors in karst development. Arid climates, whether hot or cold, support little karst. Physical rock properties also are important. Highly porous rocks seldom support well-defined karst features, which instead are favoured by low porosity and good secondary permeability, in the form of fractures, focusing the drainage into specific conduits through the karst rock. The removal of rock in solution allows the development of drainage through the rock, rather than just across its surface as happens largely with rocks removed by mechanical erosion. Consequently, karst landscapes generally lack well-developed surface drainage but have underground drainage conduits, or caves. Hence a significant component of karst terrains typically lies beneath the surface, sometimes extending to depths of hundreds, or even thousands, of metres. Intimately associated with the dissolutional aspects of karst are depositional ones. The latter include clastic sediments within the caves and, particularly, minerals deposited by precipitation from karst waters both above and below ground.

Many subdivisions of karst have been proposed. Relict karst is used to denote landforms inherited from earlier climatic or drainage regimes but still subject to modification by the current conditions. Palaeokarst refers to karst features buried by younger rocks and so largely isolated from current karst modification; where uncovered by later denudation, this isolated karst is called exhumed karst. Biokarst encompasses small-scale sculpting of limestone by animals and plants, although the distinction between dissolutional sculpting (true biokarst) and mechanical excavation (bioerosion) is seldom made. Pseudokarst is, as its name implies, 'false karst'. Such features superficially resemble karst but form by quite different processes, such as lava tubes, soil piping, and thermokarst, or cryokarst, formed by localized melting of permafrost.

Often karst geomorphology is regarded as a specialist discipline that is of limited general application to geology or geomorphology. However, ~12% of Earth's terrestrial, ice-free surface is composed of limestone, with 7–10% supporting some form of karst landscape. Furthermore, as much as 25% of the world's population may depend to some extent on karst water supplies. Consequently, the study of karst is crucial to understanding landscape and drainage development over a significant area of Earth's surface.

Karst Processes

The basic process of karst dissolution involves ion dissociation. For rock salt (NaCl), gypsum ($CaSO_4{\cdot}2H_2O$), and quartz (SiO_2), this requires only the presence of water but, on a global scale, the outcrop area of evaporites and the low solubility of quartz render them of only minor and localized significance for karst development. Limestone ($CaCO_3$) and dolomite ($CaMg[CO_3]_2$) are by far the dominant karst rocks but experience very low rates of dissociation in pure water. The addition of free H^+ ions greatly increases the rate of carbonate dissociation and hence even weak acids become effective solvents. In most karst environments, carbonic acid, derived from atmospheric or soil CO_2, is the main source of H^+ ions, although other organic or inorganic acids may be significant locally. The solubility of CO_2 increases with decreasing temperature, in common with other gases. The same is true also for limestone, in marked contrast to most solids for which solubility increases with temperature. Nonetheless, the availability of liquid water and biogenic CO_2 is far more significant for karst development than are low temperatures.

Water with an excess of H^+ ions is commonly referred to as 'aggressive', and continues to take up HCO_3 ions until an H^+/HCO_3 equilibrium is reached at saturation, as in the following equation:

$$CO_3^{2} + 2H^+ \rightleftarrows HCO_3 + H^+ \rightleftarrows H_2CO_3$$

Different karst waters may reach saturation at different concentrations, determined by the initial CO_2 concentration, but this is not a simple straight-line relationship and mixing of different karst waters may increase aggressivity. This phenomenon, called mixing corrosion, may be significant in certain karst environments. Carbonate solubility also is increased by the foreign ion effect, the addition of ions such as Na^+, K^+, and Cl . Seawater is saturated and cannot directly dissolve limestone, but mixing with freshwater can considerably increase carbonate solubility and is of major significance in certain environments. In the same way that an increase in CO_2 concentration increases carbonate uptake, so degassing of CO_2 from a saturated solution causes reprecipitation of calcite.

Dissolution can and does occur in static or laminar flow conditions, though constrained by diffusion rates through the boundary layer. Permeable soil or sediment cover, even when vegetated, may offer only limited resistance to downward percolation of water to the limestone beneath. Although this subsoil water movement may be slow, its dissolutional efficacy is enhanced by higher CO_2 concentrations generated by

respiring plant roots and soil organisms. With faster flow, generally associated with increased water depth and/or conduit width, laminar flow gives way to turbulent flow and there is a rapid (~fivefold) increase in dissolution rate. This laminar–turbulent transition, or turbulent threshold, is very significant for the development of many karst features.

Time is a major factor to consider in the development of karst, reflecting dissolution rates. Features on a millimetre to centimetre scale may form within a few years on limestone, whereas the largest scale landforms may require hundreds of thousands, or even millions, of years to reach their present form. Timescales for karst development on gypsum are much faster, whereas on quartz they are very much slower.

Figure 1 Solution flutes (rillenkarren) on gypsum, Devon coast, England. These may form on gypsum within a few months or years in a humid climate.

Surface Karst (Exokarst)

At the surface of the Earth, karst features are manifest at scales from tens of kilometres down to millimetres. Large-scale karst landscapes and landforms invariably support a diversity of smaller scale features. The type and scale of karst features that develop at a particular location are determined by several factors that can be grouped broadly: (1) intrinsic rock properties such as composition, texture, porosity, and fracturing; (2) extrinsic factors, particularly climate (especially the amount of rainfall), specific environment (terrestrial, lacustrine, or marine), the type and extent of plant cover, sediment cover (if any), and tectonism (uplift or subsidence); and (3) time. In general, karst is best developed when pure limestones with low porosity but good secondary permeability are subject to frequent, rather than heavy, rainfall. Hence karst landforms are better developed in humid temperate or tropical climates than in strongly seasonal ones. Duration of exposure to dissolution is critical; for particular intrinsic and extrinsic conditions there is an optimum time-scale for specific karst features to develop. Beyond this time-scale, dissolution may continue to modify and, ultimately, even destroy the features.

Small-Scale Karst (Karren)

Small-scale solutional sculpting (millimetres to metres), often termed 'karren', develops purely as a function of rock solubility and run off, independently of the underground drainage characteristic of karst terrains. The development time-scale for most karren ranges from a few years to tens of thousands of years under optimum conditions. On exposed limestone solution pans (kamenitzas) form by ponding of water on level surfaces, while solution flutes (rillenkarren; Figure 1), runnels (rinnenkarren), and ripples develop by water flow across inclined surfaces. Fractures, commonly joints, act as initial conduits for flow into the karst aquifer and are widened by dissolution to form grikes (kluftkarren). The limestone between grikes may, depending on factors such as lithology, soil cover, and climate, form flat-topped blocks (clints or flachkarren), rounded blocks, domes or teeth (rundkarren), or sharp pinnacles (spitzkarren), or may break up into fragments (shillow, felsenmeer, or trümmerkarren). Glacially stranded boulders may protect limestone directly beneath them from direct dissolution by rainfall, eventually leaving the boulder perched on a pedestal surrounded by a limestone surface lowered by solution. The height of such pedestals provides a measure of the surface lowering rate (Figure 2). Clints, grikes, solution runnels, and solution pans are common features beneath mineral or organic soils, though typically they are more rounded than those formed subaerially.

Freshwater lake shores support their own unique karren assemblages, strongly influenced by lake water chemistry. Small hemispherical solution pits (Figure 3) are particularly characteristic on flat limestone surfaces around lakes, and notches (swamp notches) may indicate former water levels. Even if the lake water is permanently saturated with carbonate, the limestone can dissolve if seasonal changes in water level trap air pockets in which condensation corrosion occurs; this process forms upward-tapering tubes (röhrenkarren). Similarly shaped, but larger, features called bell holes are found in the entrance zone of some tropical caves and are considered to form by condensation corrosion associated with daily changes in air temperature.

Most karren are essentially abiogenic in origin, but living organisms may contribute directly to limestone

Figure 2 Limestone pedestal beneath a sandstone erratic. The pedestal provides a measure of postglacial (~14 ka) surface lowering of the surrounding limestone. Conspicuous solution runnels (rinnenkarren) have formed on the top of this sandstone erratic despite its low solubility. County Cavan, Ireland.

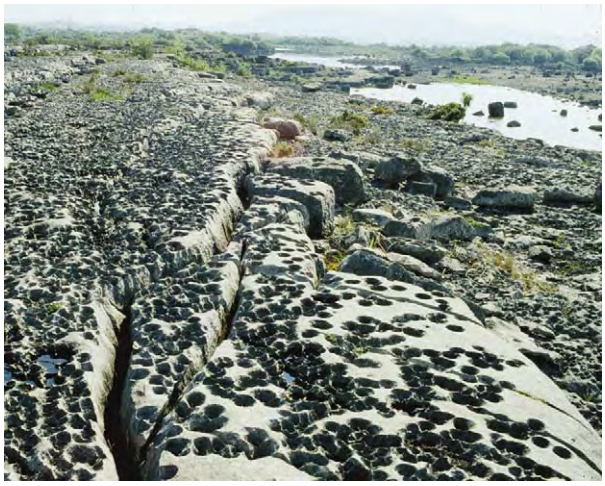


Figure 3 Abundant solution pits, typical of lacustrine karst, on limestone surfaces exposed in the zone of seasonal submergence. Lough Mask, Ireland.

dissolution on a small scale. Certain algae and invertebrates adopt an endolithic habit, living within cavities that they excavate in the rock. Cyanophytes constitute a significant karstic agent in some climates. Where created with acids, these cavities are true biokarst (Figure 4). However, many organisms excavate cavities mechanically, by bioerosion, and these are not strictly biokarst. Algal biokarst has been termed phytokarst and is ubiquitous in intertidal marine environments, where endolithic organisms are the dominant karst agent. In cave entrances or beneath overhangs, light-orientated phytokarst pinnacles, or

Figure 4 Temperate marine biokarst, encrusted with *Mytilus*, in the intertidal zone. The Burren, Ireland.

photokarren, may develop where biokarstic algae are illuminated from just one direction.

Medium-Scale Karst (Karst Landforms)

Medium-scale features (metres to kilometres) can perhaps be grouped together as karst landforms and develop mainly over time-scales of thousands to hundreds of thousands of years. In any karst landscape, dissolution is greatest near the surface where water first comes into contact with the rock, creating a shallow zone of epikarst above less karstified rock. Over time, and dependent on climate, the depth and dissection of this epikarst zone increases. The most characteristic surface karst landform is the doline (sometimes also known as a sinkhole, especially in North America and by engineers everywhere), an enclosed depression from which drainage is underground; on non-karst lithologies such depressions would form lakes. Dolines can form by one or a combination of mechanisms, e.g., by enhanced dissolution through focusing of drainage, by collapse into a cavity beneath, or by downwashing of a weakly consolidated soil or sediment cover into cavities beneath (Figure 5). Collapse dolines may intercept the regional water table to form cenotes; they are particularly characteristic of some lowland karst plains. Irregular, flat-floored depressions may develop by the lateral solution and planation of limestone inundated by seasonal rises in the regional water table. Known as poljes in their type area of the Dinaric Alps, they are fed by springs and drain via sinks, or may fill and empty via the same conduit, known as an estavelle, depending on flood stage or groundwater level.

Drainage in karst terrains is not exclusively subterranean and elements of fluvial geomorphology are common. Valleys and gorges may develop where rivers cross from adjacent non-karst (allogenic) rocks and travel across the limestone surface before sinking.

Figure 5 Suffosion dolines on glacial till, formed by downwashing of sediment into cavities in the underlying limestone. Black Mountain, South Wales.

Many are now largely dry and were formed under different climatic regimes, reflecting either greater rainfall or the blocking of underground drainage by permafrost; Cheddar Gorge in south-west England is a fine example of the latter. So-called blind valleys end abruptly downstream against an often steep slope where a surface stream sinks underground, though the former surface continuation may continue beyond at a higher level. Conversely, headward excavation by a spring may form a valley ending abruptly upstream at a steep slope, called a steephead or pocket valley. Emerging saturated karst waters may precipitate calcite, often under biogenic mediation, to form sometimes extensive deposits of tufa or travertine, such as those of Plitvice, Croatia.

Residual limestone hills at scales from a few metres to hundreds of metres high are an inevitable consequence of focused drainage from surrounding limestone, particularly where this is sufficient for turbulent flow, and the enhanced dissolution rates associated with it, to occur. Although individual hills can be regarded as landforms, they are more logically considered here as elements of the largest scale of karst development, karst landscapes.

Large-Scale Karst (Karst Landscapes)

Karst landscapes can be broadly defined as areas, typically covering many square kilometres, encompassing characteristic assemblages of karst landforms and karren. Generally, though not invariably, they reflect prolonged karstification over several million years and large-scale interactions between geology, tectonics, and climate. The youngest distinct type of karst landscape, characteristic of many cool temperate regions subjected to repeated glaciation during the Pleistocene, is termed glaciokarst (Figure 6). The Burren of western Ireland is a classic example and shows features typical of glaciation, such as ice-plucked

Figure 6 Temperate glaciokarst landscape with typical joint controlled clints and grikes, rinnenkarren, rillenkarren, and kamenitzas. The Burren, Ireland.

crags, scoured rock surfaces, limestone pavements, and erratic boulders. Commonly these are superimposed on pre-existing karst landforms/landscapes, but smaller scale karst landforms and karren, particularly grikes (kluftkarren), small dolines, and caves, have commonly been superimposed on the glacial landforms since the ice retreated.

Fluviokarst landscapes are dominated by valleys cut by surface rivers, either before significant underground drainage had developed or under different climatic conditions, notably periglacial environments, which favoured surface rather than underground drainage. Many fluviokarst valleys are now largely dry because drainage has been captured underground.

Cone karst (Fencong) is dominated by conical or hemispherical hills, typically a few tens of metres high and several hundred metres across, separated by dolines and other irregular depressions. Cone karst develops only in wet tropical climates such as in parts of Java, some Caribbean islands, and particularly in China's Ghizou Province. Tower karst (Fenglin),

Figure 7 Relict karst towers, accentuated by marine undercut ting, in Halong Bay, Vietnam. These testify to active tectonic uplift in a terrestrial environment, prior to drowning by sea level rise or regional subsidence. Photo by Sally Hunt.

perhaps the most spectacular of all karst landscapes, is dominated by often precipitously steep limestone towers, from tens to hundreds of metres high, and is restricted to areas in south-east Asia and the adjacent islands (Figure 7). Cone and tower karsts develop where rates of tectonic uplift exceed, for the former, or are equalled, for the latter, rates of surface lowering at points of focused dissolution, leaving residual hills or towers between. In contrast, prolonged denudation with minimal uplift will, ultimately, produce a low-relief corrosion plain, as typified by the Gort lowlands to the south of Galway, Ireland.

Various classifications describe stages of 'maturity' in karst landscapes; the most useful, and comprehensive, scheme (Figure 8) was developed primarily for engineers and encompasses a range of parameters, including karst class (a qualitative measure based on the scale and diversity of karst landforms), mean doline density, rockhead relief, and typical cave size.

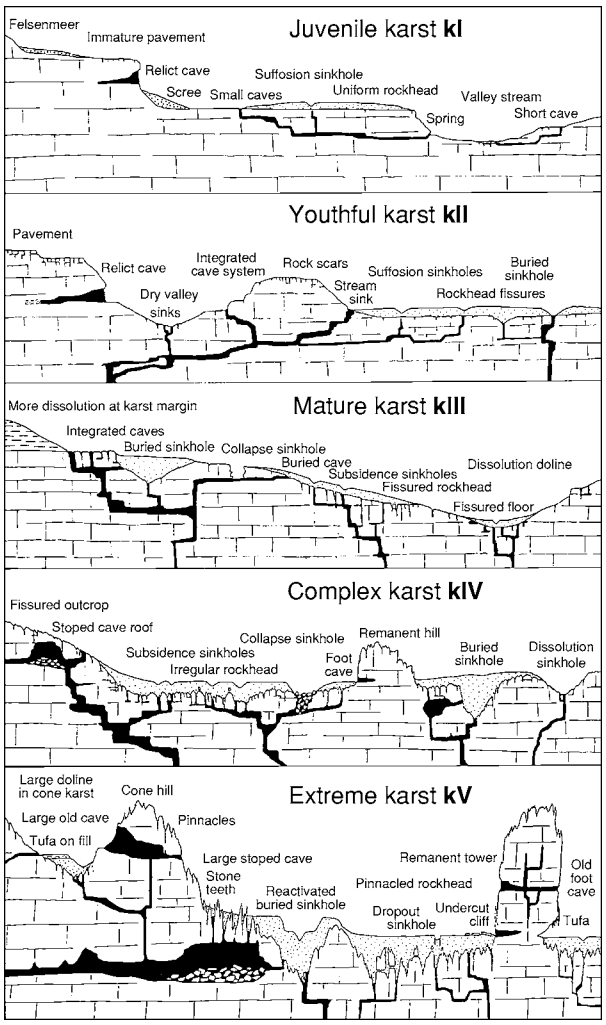


Figure 8 Typical landforms encountered within the five cat egories of ground conditions recognized by Waltham and Fookes in their 2003 engineering classification of karst.

Karst Denudation and Landscape Development

In landscapes encompassing both soluble and insoluble lithologies, differences between erosional and dissolutional processes have a considerable effect on landscape development in tropical and temperate regions. Karst dissolution is a low-energy process that may occur even in static water and, in the case of limestone, is enhanced by a vegetation cover, which increases soil CO_2 levels. Analogous weathering of non-karst rocks may alter their composition and mechanical strength without significant volume change or transport, forming saprolites sometimes many metres thick. These are removed only by erosion, a higher energy process requiring fast water flow but hindered by vegetation, which binds and protects weathered rock and soil. In temperate and tropical environments, limestone may be removed at a steady rate, proportional to mean annual rainfall, whereas erosion of rock and saprolite is more sporadic and depends on the frequency of intense rainfall events and the extent of vegetation cover. In arid regions, limestones commonly prove more resistant to erosion than do other, less well-cemented, rocks. Conversely, in tectonically stable humid climates, the long-term denudation rate of limestones may be significantly greater than is the overall erosion rate on adjacent rocks, reducing limestone outcrops to lowlands flanked by hills of more slowly denuded lithologies.

Karst Drainage

A significant element of drainage in karst landscapes is underground. Drainage into a karst aquifer is either

autogenic, where derived from a karst catchment, or allogenic, originating on non-karst rocks. Commonly, input to a karst aquifer is mixed. Autogenic recharge is diffuse, as rainfall across the entire karst outcrop, whereas allogenic recharge is concentrated at discrete stream sinks or swallets. Most negotiable caves are associated with such swallets. Water re-emerges from a karst aquifer at a rising (also known as a resurgence or spring). Typically, this occurs at the lowest point that the karst rock is exposed at the surface, although impermeable layers within a karst aquifer may cause a rising to be perched at a higher level. Resurgences may be free-draining streams or may rise from depth under hydrostatic pressure. The more than 300 m-deep Fontaine de Vaucluse in France is an example of the latter; the term 'vauclusian spring' has been coined to refer to the point of emergence of water under pressure.

Caves (Endokarst)

Underground drainage conduits are a fundamental element of karst. Those large enough for a person to enter are called caves, an artificial definition constrained by the size of the caver! In fact, water can move slowly along submillimetre fractures, but at widths of ~5 mm laminar flow gives way to turbulent flow (the turbulent threshold, or hydraulic jump) and dissolution rates increase significantly. Such 'protocaves', often anastomosing networks of small tubes, commonly are found on fracture planes adjacent to larger caves that have developed from them. The development of a cave, from inception to destruction, is called speleogenesis.

Fractures commonly represent the initial routes for water flow, and the configuration of many passages is strongly influenced by the spatial distribution and orientation of joints, faults, and bedding planes. Rectilinear passage networks arise where inception is dominated by vertical fractures, whereas meandering passages imply inception on subhorizontal ones. Passage shape may be influenced by these fractures but also is strongly controlled by passage location with respect to the 'water table'. Above, in the vadose zone (see Figure 9), water occupies only the lower part of the passage and flows downwards, as in surface streams and waterfalls; dissolution here forms canyonlike passages and vertical shafts. Sequences of deep vadose shafts and canyons, sometimes descending a kilometre or more, are characteristic of high mountain karst regions such as in the Alps. Below, in the phreatic, or 'saturated', zone, fractures and

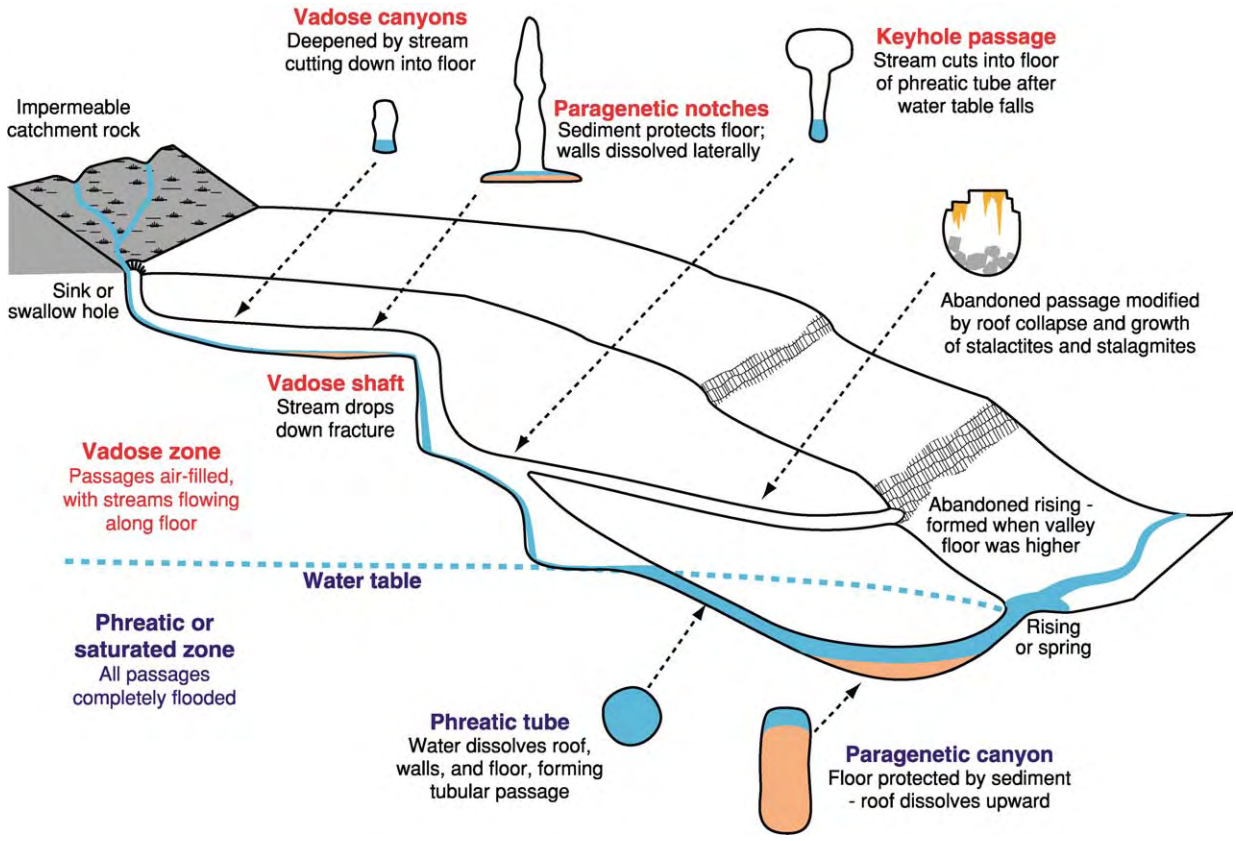


Figure 9 Diagrammatic section through a cave system to show the various features discussed in the text.

Figure 10 Relict paragenetic canyon, with sediment removed, extending upwards from the bedding plane on which the anastomosing 'proto caves' first developed. The Burren, Ireland.

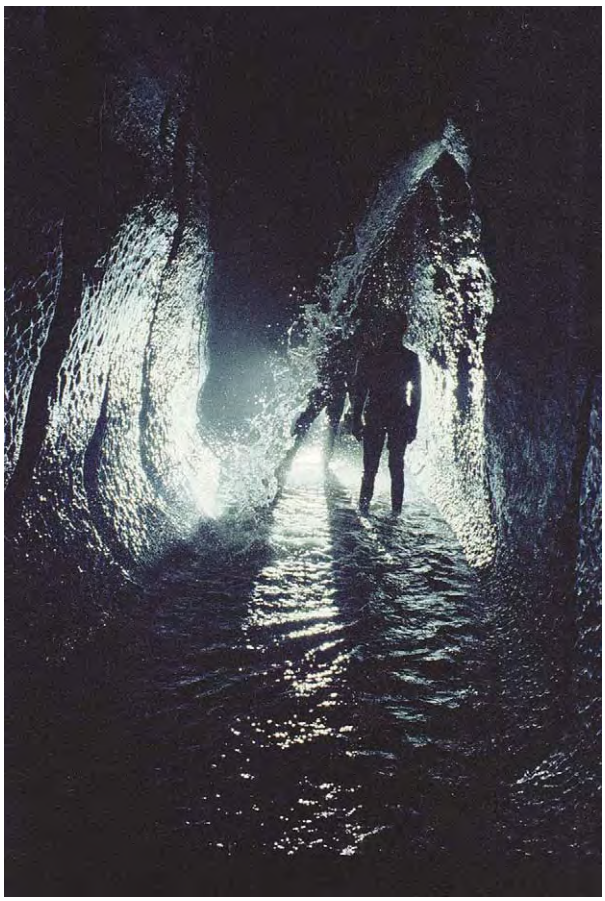


Figure 11 Active vadose canyon incised below the inception level at which the cascade is entering. Doolin River Cave, The Burren, Ireland.

cavities are water filled, and water flow follows the path of least resistance, often flowing upwards if necessary. Dissolution here occurs in all directions, forming passages with circular or elliptical cross-sections, as seen in the drained tubes of Kentucky's Mammoth Cave and still forming in the active phreas of Mexico's Yucatan, often with blind pockets formed by eddies or by mixing corrosion where percolation water emerges from a minor fracture into a flooded passage. Mixing of karst waters from separate sources may enhance dissolution locally, influencing both conduit configuration and shape. Sediment aggradation is common both in vadose and in phreatic passages and may armour the floor against downward dissolution, causing lateral or upward dissolution by paragenesis (this is a process quite distinct from mineral paragenesis). In vadose passages, lateral dissolution may form paragenetic notches at stream level, spectacularly seen in some of the huge river caves of Mulu, Sarawak. In phreatic environments, dissolution may raise the roof, forming a largely sediment-filled paragenetic canyon. Paragenetic canyons (Figure 10) extend upwards from their initiation level, such as a bedding plane, whereas vadose canyons (Figure 11) cut downwards from it. A drop in aquifer base level may flush out sediment fill from paragenetic canyons and/or form 'keyhole' passages by vadose incision in the floor of former phreatic tubes. The two types of canyon passages can be distinguished by the presence of either undulating wall notches, indicating phreatic paragenesis associated with undulating sediment banks, or horizontal notches, indicating vadose conditions. Keyhole passages, with a vadose trench incised in the floor of a phreatic tube, are particularly significant for interpreting the history of karst aquifers, indicating rejuvenation when base level drops below the cave passage.

Water in a karst aquifer may be confined by an impermeable rock (aquifuge), whereas impermeable or weakly permeable layers (aquicludes or aquitards) may influence water movement through the rock, allowing the development of perched aquifers above regional base level. Hence the altitude of a particular phreatic passage may reflect only local conditions. Eddies caused by turbulent water flowing across limestone in caves and surface streams commonly form asymmetric concavities, called scallops, which are invaluable flow indicators in relict cave passages (Figure 12). Their steeper slopes face downstream, and their size is a function of flow velocity. Smaller scallops indicate faster flow than larger ones; vadose scallops are sometimes only 1–2 cm across, whereas phreatic scallops may be 1 m across! An approximate rule of thumb for calculating flow velocities (in metres per second) is to divide 3.5 by the mean length of the scallops in centimetres.

Figure 12 Heavily scalloped relict phreatic pressure tube entering a larger cave passage. The strongly asymmetric scal lops provide clear evidence of upward flow, under pressure, into a larger water filled passage. Deflection of these scallops above the pressure tube indicates flow from left to right in the main passage. Ogof Llyn Du, North Wales.

Figure 13 Typical cave fills in an active vadose cave. Spe leothems (stalactites) are deposited by carbonate saturated per colation water emerging along joints on roof. Coarse clastic sediments are carried into the cave by fast flowing water. Marble Arch Cave, Northern Ireland. Photo by Tony Waltham.

Most caves form by an overall downward movement of water through the karst aquifer from sink to rising, but a few form by very different processes. Some are created instead by hydrothermal springs or by the action of sulphuric acid formed from hydrogen sulphide rising from hydrocarbon deposits beneath. Encompassing some of the largest and longest caves in the world, such as the Carlsbad and Lechaguilla caves in New Mexico, they contain some of the most extraordinary arrays of speleothems (cave mineral deposits) known, with gypsum a major component.

Over time, cave passages may collapse or be infilled by clastic sediment, transported in from the surface, or by speleothems (Figure 13). Clastic sediments provide important information on environmental histories and potentially can be dated using techniques such as optically stimulated luminescence. Calcite and aragonite speleothems (stalactites, stalagmites, and flowstone) are the most common mineral deposits, precipitated in response to CO_2 degassing of carbonate-saturated percolation water emerging into air-filled cavities and hence providing evidence of vadose conditions. They are immensely important in karst research because the calcite can be dated radiometrically, particularly by uranium-series disequilibrium, and can provide information on palaeoclimate and hydrology.

Figure 14 Spectacular palaeokarst shaft developed in, and capped by, marine Carboniferous limestones. The size of this pothole, filled with non marine clastics, provides evidence of the scale and duration of this intra Carboniferous emergent episode. Cookstown, Northern Ireland.

Palaeokarst and Interstratal Karst

Palaeokarst is defined as karst that has been buried by younger rocks. It is a common, though often unrecognized, component of successions in which limestones are present (Figure 14), and serves as a clear indicator of terrestrial environments and, to some extent, duration of emergence. Interpretation

of palaeokarst may be complicated by two factors; it is usually visible only in two, rather than three, dimensions; and burial by younger rocks does not prevent palaeokarst being modified or even destroyed by subsurface dissolution. Small-scale karst microsculpting with relief of a few centimetres, such as rillenkarren or solution pits, is readily destroyed by pressure solution associated with rock overburden. Lithological contrasts commonly act as a focus for water movement and the development of interstratal karst, which may resemble palaeokarst but is younger than the overlying rock. Interstratal karst is especially prevalent in highly soluble rocks such as gypsum and rock salt. Cave passages may remain open as potential conduits for water flow long after burial of the karst surface, and sometimes are invaded and infilled by mineralising fluids. Such occurrences may be difficult to distinguish from syngenetic mineral deposits, in which conduit formation and mineralisation were penecontemporaneous, but mineralized palaeocaves commonly include some clastic sediment infill beneath the mineral deposits. Remains of organisms within caves are protected from destruction by weathering and scavengers; hence palaeokarst caves encompass many important fossil lagerstätten. Although some animals are adapted for life in caves, most remains found are of animals and plants washed in or fallen in from the surface.

See Also

Engineering Geology: Geomorphology; Subsidence; Ground Water Monitoring at Solid Waste Landfills. **Minerals:** Carbonates. **Mining Geology:** Hydrothermal Ores. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Lake Processes and Deposits. **Sedimentary Processes:** Erosional Sedimentary Structures. **Sedimentary Rocks:** Dolomites; Evaporites; Limestones. **Unidirectional Aqueous Flow**.

Further Reading

Ford DC and Williams PW (1989) *Karst Geomorphology and Hydrology*. London: Unwin Hyman.

Gunn J (ed.) (2003) *Encyclopedia of Caves and Karst Science*. London: Routledge.

Klimchouk AB, Ford DC, Palmer AN, and Dreybrodt W (eds.) (2000) *Speleogenesis: Evolution of Karst Aquifers*. Huntsville, AL: National Speleological Society.

Jennings JN (1985) *Karst Geomorphology*. Oxford: Black well.

Lowe DJ and Waltham AC (2002) *Dictionary of Karst and Caves*. Bridgewater, UK: British Cave Research Assoc.

Palmer AN (1991) Origin and morphology of limestone caves. *Bulletin of the Geological Society of America* 103: 1 21.

Simms MJ (1994) Emplacement and preservation of verte brates in caves and fissures. *Zoological Journal of the Linnean Society* 112: 261 283.

Simms MJ (2004) Tortoises and hares: Dissolution, erosion and isostasy in landscape evolution. *Earth Surface Pro cesses and Landforms* 29: 477 494.

Waltham AC and Fookes PG (2003) Engineering classi fication of karst ground conditions. *Quarterly Journal of Engineering Geology and Hydrogeology* 36: 101 118.

Zhang Z (1980) Karst types in China. *GeoJournal* 4: 541 570.

Landslides

S F Burns, Portland State University, Portland, OR, USA

Introduction

Landslides are a major form of geohazard that cause millions of dollars of damage and many deaths each year. They result from the movement of masses of rock, earth, and debris down a slope under the influence of gravity. They can occur at all scales, from a few metres across to several kilometres, and at all rates, from metres per year to metres per second. The term landslide is really a misnomer because the material can move not only by sliding, but also by flowing and/or falling. Landslides are important superficial processes for geomorphology, moving sediment down the slope where water in streams can eventually carry it away. They are generally triggered by water, but sometimes earthquakes and even volcanic eruptions can lead to landsliding. On land the water comes mainly from heavy and persistent rainfall but also can come from rapid snowmelt. Submarine landslides may be triggered by cyclic wave loading as well as by earthquakes, and frequently follow marine trangressions.

Possibly the largest landslide on land occurred some 10 000 years ago in south-western Iran. Called the Saidmarreh Landslide, a mass of limestone 15 km long, 5 km wide, and 300 m thick slid off the underlying marl bedrock and travelled a distance of 18 km while dropping only 1000 m. One of the largest underwater slides has been described to have

occurred on the continental margin of the Atlantic Ocean off the coast of Spanish Sahara in water depths over 2000 m and covering an area of 18 000 km^2. It probably originated after the last glacial advance when sea-level was 100 m lower than it is today. Landslides have also been identified on other planets like Mars, where one with a volume of 100 km^3, 60 km long, and 50 km wide that has travelled 2 km, leaving a scarp 6 km high, has been identified.

Loss of Life and Property due to Landsliding

Human and economic losses from landslides are considerable worldwide and continue to grow as pressures of expanding population cause development to extend into less stable hillslope regions. It is estimated that each year landslides kill nearly 600 people around the world, and cause over $12 billion (US) damage (direct and indirect losses). Direct losses include the cost of rebuilding the structures, whereas indirect losses are concerned with commercial loss, reduced real estate value, and loss of productivity. Japan is one of the most affected countries, suffering over $4 billion (US) loss per year. The United States, Italy, and India each have total annual losses of between $1 and $2 billion (US) each.

The worst landslide disaster in historical times occurred on December 16, 1020 in Kansou, China when over 180 000 people were killed. Many of these people were living in homes excavated in or built on the slopes of loess (silt). This soil has a metastable fabric and collapsed during an earthquake-triggered event. One of the worst disasters in recent years happened in 1998 when an ancient landslide reactivated in Kelso, Washington (USA), and 60 houses on it were destroyed, causing over $25 million (US) of damage to homes and utilities (Figure 1).

Figure 1 Home destroyed in 1998 in Kelso, Washington, by a thrust of soil under it as an ancient landslide reactivated, destroying 60 similar homes over a four month period. Note the main street in the foreground has also been destroyed.

Factors Determining Slope Stability

The stability of a slope can be considered as two forces working against each other: Driving forces (or shear stress) work to cause slope materials to move down-slope, whereas resisting forces (shear strength) act to keep the materials on the slope. When the ratio of resisting forces over driving forces (called the factor of safety) is greater than 1, the slope is stable. When it is less than 1, the slope usually fails.

'Water' can both increase the driving forces and reduce the resisting forces, thereby reducing stability. Saturation from rainfall can increase the slope mass, thereby increasing the driving forces. Filling soil pores with water also reduces soil strength by increasing pore pressure and reducing the effective stress, and possibly allowing particles to pass by one another so further reducing the internal resistance of the soil to sliding. To reduce the risk of landsliding on a slope, the first consideration should therefore be the removal of excess water.

As the angle of slope increases, the driving forces also increase. Cutting a road into a slope could easily oversteepen it, making that slope more prone to landsliding unless a retaining wall is built. Each geological material has a characteristic 'angle of repose', which is its maximum stable angle, related to its frictional strength. For instance, the angle of repose for an uncemented sand is about 35°, and slopes steeper than this are likely to fail. Clay, shale, serpentinite, and uncompacted fill are all slope materials prone to failure, having significantly lower angles of repose.

'Vegetation' is also important to slope stability for it increases resisting forces through its roots that bind the soil. Trees also act as natural suction pumps that remove water from the soils through evapotranspiration, thereby further increasing slope stability.

If the bedrock is inclined somewhat parallel to the slope, it is called a 'dip slope', and the stability will be determined by the bedding plane strength, which may well be lower than the mass strength. Shales are particularly weak, as are buried 'fossil' soils on these dip slopes. The Gros Vente Landslide that dammed the Gros Vente River in 1925 in Wyoming (USA) and the Vaiont disaster in Italy in 1963 are examples of dip slope failures. Adding a stockpile or a structure to a slope can also increase the driving forces on a slope, reducing its stability.

Types of Landslide

Various classification schemes and names have been given to describe landslides, depending upon the

process that brings the soil and/or rock down the slope. Landslides are generally classified into process groups based on their rates of movement and corresponding amounts of water contained within the landslide mass (Figure 2).

'Falls' refers to the free fall of detached materials, usually rocks, that descend down a steep slope. Frost heave or heavy rains generally dislodge rocks from cliffs, causing a free fall to the bottom of the slope. A collection of this rockfall material at the bottom of the slope is called 'talus'. Rockfalls can be spectacular, such as the 9000-m^3 rockfall that plunged 500 m into the valley at Yosemite National Park in California on 10 July 1996, travelling at 300 km h^{-1}. It generated an air blast that leveled 2000 trees! Another example of a large rockfall happened near the village of Randa in the Pennine Alps of Switzerland in the early 1990s (Figure 3).

A 'topple' is a type of rock displacement where there is forward movement of a block rotating around a fixed hinge. This can occur where jointing in a slope is steeply inclined and parallel to the cliff-face. In some places in the world, this is an important factor in cliff retreat (e.g., Tasmania) and also in recently glaciated upland regions (e.g., Scotland). Toppling is also very common in basalt cliffs where columnar jointing is present.

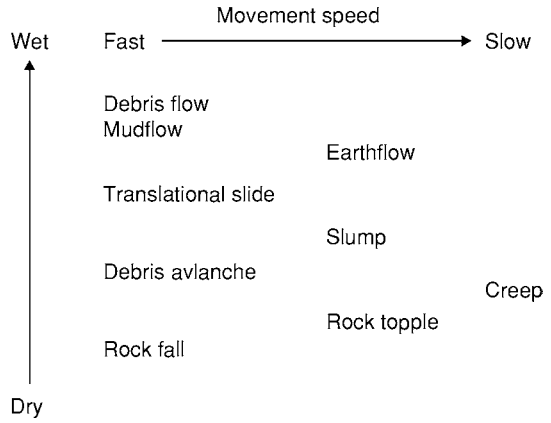


Figure 2 Landslides classification that relates water content to movement speed.

Figure 3 Large rockfall that occurred in the early 1990s in the village of Randa in the Pennine Alps of Switzerland.

'Translational slides' occur along a failure surface in the bedrock and move parallel to the surface. Most of the time the slide fractures at the base and breaks into many pieces. When the landslide moves as a uniform piece, it is called a slab failure or block slide. The failure surfaces upon which such landslides move are generally clay layers, shale layers, or palaeosols (old, buried soils) that parallel the slope. A classic translational slide occurred on 23 June 1925 in Gros Vente, Wyoming, USA. Over many years this slope of sandstone overlying shale (that dips at 20°) had been undercut by the stream at the bottom of the slope. Heavy snowmelt and heavy rains saturated the sandstone that consequently slid on the shale bed into the stream, damming it up and forming Gros Vente Lake.

If the sliding mass occurs along a curved-upward plane and shows rotation, it is called a 'rotational slide', which may develop into a 'slump' if the integrity of the failed mass is destroyed. Such movements are common in mudstones and clay soils or where there is a resistant geologic layer overlying a weaker layer. Such landslides tend to be small, producing gently inclined surfaces on the sides of steep slopes as the mass rotates (backtilts). Many large slumps occurred in Panama between 1914 and 1920 during the construction of the Panama Canal; one was so large that the whole of one side of the canal rotated into the canal itself. The slide was so large that the toe of the slide stuck up as an island in the center of the canal. Note the rotational slide movement affecting a house built on the toe of an ancient landslide that reactivated in 1996 in Tigard, Oregon, USA (Figure 4).

'Flows' occur when material moves down-slope as a slurry, generally due to continuous, permanent deformation of the sediment. 'Debris flows' consist of material ranging in size from clay to boulders mixed with water, displaying plastic flow behaviour. They form steep, lobate fronts with marginal levees. Mudflows are very similar to debris flows but contain finer-grained particles and tend to exhibit more viscous flow characteristics.

Debris flows and mudflows originate in upland source areas from many small soil slips that deliver sediment rapidly to create a stream, often the result of persistent rainfall. Deposition forms a fan-shaped deposit when the gradient becomes flat. Figure 5 shows a house in Dodson, Oregon, along the Columbia River Gorge, USA, that was buried by eight debris flows over a four-day period in February 1996.

Figure 4 Slump destroys a house in Tigard, Oregon, USA; note deformation of supporting pillars. This home was built on the toe of an ancient landslide that reactivated in February 1996.

Figure 5 Debris flow burying a house in Dodson, Oregon, USA, alongside the Columbia River Gorge; eight debris flows occurred here in February 1996.

Velocities can reach 55 km h^{-1}. Debris flows pose a significant hazard to life and property within the debris fans and along the stream valleys that connect the source areas to the debris fans. Debris flows are most common in arid regions, in alpine areas, and on the side of volcanoes. Deposits from debris flows can be differentiated from stream deposits by their high density, due to an abundance of clay and silt in the matrix. 'Debris torrents' are special types of debris flow that commonly consist of large volumes of logs and woody debris in addition to sediment in the slurry, occurring mainly in main drainage channels.

Debris flows that originate during volcanic eruptions are called 'lahars'. One of the world's most devastating debris flow disasters involving a lahar occurred on 13 November 1985 in Armero, Colombia, when a lahar originating 50 km away on the volcano of Nevado del Ruiz roared down the Lagunilla River Valley. It consisted of a 40-m-high wall of mud by the time it hit the town of Armero, burying 23 000 people. Debris flows are so dangerous because they travel at high velocities over long distances, and occur with little warning. America's volcano, Mt. Rainier, is the most dangerous because of its lahar risk: 55 lahars have originated from Mt. Rainier over the past 10 000 years, travelling up to 40 km from the mountain. Thousands of people live within a 40-km radius of Mt. Rainier and face this potential danger. The 1980 eruption of Mt. St. Helens also produced lahars that traveled more than 120 km from the mountain, destroyed 27 bridges over rivers, buried about half of Highway 504 to the west of the volcano, and delivered 34 million m^3 of mud to the Columbia River, 45 km from the volcano, which blocked shipping in the main channel.

'Earthflows' travel more slowly and generally originate on hillslopes as large tongues that break away from scarps; they slow down as they lose water. Heavy and persistent precipitation on steep slopes underlain by weak soils produce such flows. Rates range from 0.17 to 17 km h^{-1}. Once the main movement has occurred, they continue to show creep movement for years afterwards.

'Quick clay landslides' are a special type of flow, characterized by thixotropic clay, especially glaciomarine deposits formed in coastal areas of Norway, Sweden, and eastern Canada. Isostatic rebound of these clays following glacial melting has exposed the sediment to weathering and flushing out of the salts by rainwater-fed groundwater. Sudden vibrations can subsequently cause the deposit to liquefy and flow. A classic example occurred in Rissa, Norway, in April 1978 as 5 million m^3 of quick clays failed, destroying seven farmhouses and moving at a rate of 30 km h^{-1}. One person was killed in this flow, which was triggered by a simple excavation in a local farmyard.

The fastest and most destructive landslides are called 'debris avalanches', or 'sturtzstroms'. They travel at speeds up to 300 km h^{-1}, travel long distances (sometimes in excess of 50 km), and may be very large (some have been in excess of 80 million m^3). Why do these landslides travel such large distances at such high velocities and large volumes? An early idea was that the debris moved on a trapped layer of air. More recent hypotheses stress semielastic collisions between landslide particles: the body moves until all of the original (potential) energy is consumed by friction between the particles. Another hypothesis stresses internal melting and basal melting of the debris, thereby decreasing friction. This last idea comes from a study of the 2.1-km^3 Kofels Landslide in the Otz Valley in Austria where fused rock, resembling obsidian (and called frictionite), is found in layers throughout the base of the debris avalanche,

Figure 6 The May 18, 1980, eruption of Mt. St. Helens in Wash ington, USA, seen here in 2004, was initiated by a 5.2 magnitude earthquake that also triggered a large debris avalanche seen in the foreground. It has been dissected by the Toutle River. Note Castle Lake, which was formed as the landslide dammed Castle Creek, to the centre right.

Figure 7 Soil creep expressed in trees with bent trunks called 'pistol butt' trees or 'trees with knees'. The trees are Radiata pines in New Zealand's Southern Alps.

and pumice-like rocks abound at the edges, presumably being shed by this prehistoric landslide as it moved at high velocities.

The world's largest historical landslide, which occurred when Mt. St. Helens erupted in 1980, was a debris avalanche. It traveled over 25 km down the Toutle River Valley at a velocity over 150 km h ^{1} (Figure 6). In Yungay, Peru 22 000 people were killed in 1970 when a debris avalanche of over 80 million m^3 descended from the volcano, Nevado Huascaran, traveling a distance of 14.5 km in 7 minutes at a velocity of about 300 km h ^{1}. It jumped 300 m out of the valley to bury the village that was perched supposedly out of harm's way on a shelf. Another debris avalanche occurred in Elm, Switzerland, in 1881 after heavy rains and caused 10 million m^3 to move 5.5 km in less than a minute at 310 km h ^{1}, killing 115 people. Another 70 people were killed in 1903 in Frank, Alberta, when 30 million m^3 of limestone debris cascaded down the slope above the town at a velocity of 100 km h ^{1}. One the earth's largest debris avalanches dammed the river at Gohna, India, in 1893 during the monsoon rains with 3.6 billion m^3 of debris.

Significant movement on slopes can also occur but at rates almost imperceptible. This is called 'creep', with rates in the range from 0.1 to 100 mm yr ^{1}. Causes can be freeze–thaw or shrink–swell processes operating on the soil sediments, assisting movement down-slope by gravity. Evidence of movement is found in cracked walls, leaning telephone poles and walls, and trees with knees (pistol-butt trees) (Figure 7). Movement occurs mainly in the soil on the slope. Creep may also be a precursor to the failure of a slope during a landslide. It is also a common feature of a landslide mass after it has come to rest, and slopes within the failed mass regain equilibrium. Ancient earthflows have been known to exhibit creep thousands of years after the initial movement of the landslide. Hoofed animals grazing on the sides of slopes when they are wet cause small-scale terraces to form from creep, and these are called 'terracettes'.

Reactivation of Ancient Landslides

If land has moved once, there is a high chance of it moving again given the right conditions. As population pressures increase, people move onto marginal land, which may be more prone to landslides. It is therefore imperative to recognize ancient landslides and assist the planning process, for instance, with maps to raise awareness of their presence prior to building.

Geomorphologists are particularly interested in reading the landscape for signs that tell of past landslides, in order to identify active processes and understand the evolution of the landscape. First indications are scarps (arcuate steep slopes cut into the hill where the landslide has torn away). Second is hummocky topography ('bumpy' ground produced by the landslide mass weathering over time). A hummocky slope with a steep, half-moon slope at the top usually indicates that an old landslide has occurred there. Younger landslides will generally have different vegetation on the slide mass than the surrounding slopes. In coniferous forests, landslides generally are covered with deciduous trees for the first hundred years after the failure. If one can see the actual sediments making up the landslide, the colluvium there is likely to be unsorted and angular in shape as all particles have been mixed together randomly.

Prevention of Landslides and Mitigating Their Impact

Prevention of landsliding or repairing the damage occurring as the result of a landslide necessitates understanding the basic relationship of driving forces and resisting forces. The need is to reduce the driving forces and/or increase the resisting forces on the slope. Decreasing the slope angle to less than the angle of repose for the slope will significantly increase stability. Removing water from the slope through drains or temporary water diversion by the use of tarps in full will decrease the driving forces and increase the resisting forces. Building a buttress at the bottom of the slope by erecting a wall of boulders, building a concrete wall with weep holes (to prevent build-up of water pressure behind it), or building a gabion basket wall will increase the resisting forces of the slope.

Recognizing and mapping old landslides is important. Planning regulations recommend not building on ancient landslides, so preventing disasters like Kelso, Washington. Replacing slope vegetation aids the slope stability and increases the resisting forces through the removal of water. In rockfall-prone areas, the use of screens or sprayed-on shotcrete will keep the rock fragments on the slope, as will rock bolts for the larger boulders. Avalanche shelters will provide additional protection, aided by electronic fences to give warning of rockfall events.

A landslide killed 2600 people in Longarone, Italy, on 9 October 1963. The second highest concrete dam in the world had just been built to a height of 265 m—the Vaiont Dam at the base of Mt. Toc. The reservoir area was prone to landslides, and it was known that Mt. Toc had sedimentary beds that dipped steeply into the valley below and that the slopes had been oversteepened by the river at the base. Filling of the reservoir aggravated the situation, raising the groundwater level (and water pressure) at the base, aggravated by persistent heavy rains that raised the reservoir level 20 m higher than expected within a short period. The slope began to creep at a rate reaching 400 mm day^{-1} the day before failure. On the day of failure, 240 million m^3 of rock broke away on the south side of the valley and slid down into the reservoir, rapidly displacing the water and creating a huge wave 100 m high, which topped over the dam and flooded the Piave River Valley. A total of 2600 people were drowned. Today, the dam still stands, intact. Its foundations were adequate, but the location with respect to the reservoir impounded behind was quite inappropriate. The landslide could have been prevented if the planners had paid attention to the broader context of the location and, in particular, the effect of impounding on slope instability of the reservoir margins, and the movements that had been noted from the past.

See Also

Engineering Geology: Overview; Aspects of Earthquakes; Geological Maps; Natural and Anthropogenic Geohazards; Site and Ground Investigation; Site Classification. **Geological Engineering**. **Geomorphology**. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Particle-Driven Subaqueous Gravity Processes. **Solar System:** Mars. **Tectonics:** Earthquakes. **Urban Geology**. **Volcanoes**.

Further Reading

Brabb EE and Harrod BL (eds.) (1989) *Landslides: Extent and Economic Significance*. Rotterdam: AA Balkema.

Carson MA and Kirkby MJ (1972) *Hillslope Form and Process*. Cambridge, UK: Cambridge University Press.

Chacon J, Irigaray C, and Fernandez T (eds.) (1996) *Landslides*. Rotterdam: AA Balkema.

Chorley RJ, Schumm SA, and Sugden DE (1984) *Geomorphology*. London: Methuen.

Cruden DM and Fell R (1997) *Landslide Risk Assessment*. Rotterdam: AA Balkema.

Easterbrook DJ (1999) *Surface Processes and Landforms*, 2nd edn. Upper Saddle River NJ: Prentice Hall.

Ritter DF, Kochel RC, and Miller JR (2002) *Process Geomorphology*, 4th edn. Dubuque IA: William C Brown.

Selby MJ (1982) *Hillslope Materials and Processes*. Oxford: Oxford University Press.

Slosson JE, Keene AG, and Johnson JA (eds.) (1992) *Landslides/Landslide Mitigation*, Reviews in Engineering Geology, vol. IX. Boulder: Geological Society of America.

Summerfield MA (1991) *Global Geomorphology*. London: Longman Scientific and Technical.

Turner AK and Schuster RL (eds.) (1996) *Landslides: Investigation and Mitigation*. Special Report 247. Washington DC: Transportation Research Board, National Research Council.

Walker BF and Fell R (eds.) (1987) *Soil Slope Instability and Stabilisation*. Rotterdam: AA Balkema.

ENCYCLOPEDIA OF
GEOLOGY

ENCYCLOPEDIA OF GEOLOGY

EDITED BY

RICHARD C. SELLEY
L. ROBIN M. COCKS
IAN R. PLIMER

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Particle-Driven Subaqueous Gravity Processes

M Felix and W McCaffrey, University of Leeds, Leeds, UK

Introduction

Particulate subaqueous gravity flows are sediment-water mixtures that move as a result of gravity acting on the sediment-induced density excess compared with the ambient water. The mixtures can range from densely-packed sediment flows, that are essentially submarine landslides, to very dilute flows carrying only a few $kg\,m^3$ of sediment. Gravity flow can take place in lakes and oceans, but some dense flows also occur in rivers. Sediment volumes transported by individual events can range up to thousands of cubic kilometres, although most events are of much smaller magnitude. Due to their infrequent occurrence and destructive nature, much information about subaqueous gravity processes comes from the study of their deposits and from laboratory experiments. Flow initiation mechanisms, sediment transport mechanisms, and flow types are described here separately, to emphasise the sense of process continuum needed to appreciate the development of most natural subaqueous gravity flows. This is followed by a description of internal and external influences on flow behaviour. Finally, the influence of flow regime on individual deposits is outlined.

Flow Initiation Mechanisms

A variety of processes can generate subaqueous gravity currents, with varying initial concentrations.

Direct Formation From Rivers

Currents can be formed when turbid river water flows into bodies of standing water such as lakes or oceans. If the bulk density of the turbid river water (sediment plus interstitial fluid) is higher than that of the receiving body of water, the river outflow will plunge, travelling along the bed as a hyperpycnal flow (or plume) beneath the ambient water. Such sediment-laden underflows may mix with the ambient water and transport sediment oceanward as particulate gravity currents. Although sometimes these river-derived flows are of high concentration (e.g., the Yellow River hyperpycnal plume), mostly they are dilute. Direct formation of subaqueous gravity currents in this way is, however, the exception rather than the rule. More commonly, the bulk density of the turbid river outflow is less than that of the ocean, and turbid surface plumes are generated. Nevertheless, particulate gravity flows can also form from surface plumes if material settling out collects near the bed at high enough concentrations to begin moving. A similar effect results from flow generated by glacial plumes where the sediment is slowly released into the water body.

Where the interstitial fluid in a hyperpycnal plume is of lower density than that of the ambient fluid, as is the case when freshwater rivers flow into brackish or fully saline bodies of water, ongoing sedimentation may induce buoyancy reversal. Thus, the gravity current will loft, in a manner similar to some subaerial pyroclastic density flows, and the flow will essentially cease to travel forwards, resulting in the development of abrupt deposit margins.

Sediment Resuspension

Loose sediment on the seafloor can be resuspended if bed shear stress is high enough. This can occur during storms or during passage of flows caused by density differences as a result of temperature or salinity. The resulting suspended sediment concentrations can be high enough to allow the mixtures to flow under the influence of gravity. As in the case of river-derived flows, resuspension usually generates initially dilute currents.

Slope Failure

Flows of much higher concentration may form as a result of slope failure. Sediment on submarine slopes can become unstable as a result of slope oversteepening during ongoing sedimentation, and during sea-level falls, as a result of high inherited pore fluid pressures and gas hydrate exsolution. Slope failure can alternatively be triggered by externally applied stresses, due to earthquakes, or as a result of loading induced by internal waves in the water column above (which chiefly occur in oceans). Initially, the failing mass becomes unstable along a plane of instability and a whole segment of the slope starts moving. Retrogressive failure and/or breaching can continue, adding material following the initial loss of stability. The concentration of this mass is at packing density but can become more dilute as flow continues.

Terrestrial Input

Not all subaqueous gravity flows need originate under water. Landslides, pyroclastic flows, and aeolian sediment transport originating on land can enter

lakes or oceans and continue flowing underwater if the rates of mass flux are sufficiently high.

Grain Transport Mechanisms

Matrix Strength and Particle-Particle Interactions

Within dense flows, grains can be prevented from settling as a result of matrix strength (Figure 1). This strength may arise if some or all of the particles are cohesive. The resulting cohesive matrix prevents both cohesive and non-cohesive particles from settling out. In addition, particles can be supported by matrix strength within flows of non-cohesive grains if the particles are in semi-permanent contact, as is the case for flows whose densities are close to that of static, loose-packed sediment. For slightly lower concentrations, inter-particle collisions will help keep particles in suspension.

Hindered Settling and Buoyancy

Settling of particles can be slowed down by water displaced upwards by other settling particles (Figure 1). Such hindered settling is especially effective in dense mixtures with a range of grain sizes so that the smaller particles are slowed down by settling of the larger particles. The presence of smaller particles also increases the effective density of the fluid that the particles are settling in and thus enhances the buoyancy of the suspended particles and reduces settling rates.

Turbulence

The motion of sediment-laden flows can generate turbulence through shear at the bed, internally in the flow or at the top of a dense layer. The turbulent bursts generated at the bed tend to have an asymmetrical vertical velocity structure, with slower downward sweeps and more rapid upward bursts. This turbulence pattern counteracts the downwards settling of particles, moving them higher up in the flow (Figure 1). Turbulence generation is hindered and dissipation increased, however, if the particle concentration is high, or if the flow is very cohesive or highly stratified.

Flow Types

Broadly speaking, flows can be divided into three main types, depending on density:

Dense, Relatively Undeformed Flows, Creeps, Slides and Slumps

Flows of this type essentially have the same density as the pre-failure material. In each case the sediment moves as one large coherent mass, but with varying amounts of internal deformation. Grains remain in contact during flow and thus matrix strength is the main sediment transport mechanism. Such flows will stop moving or shear stress becomes too low to overcome friction, at which point the entire mass comes to rest. Flow thickness and deposit thickness are essentially the same, although flows may thicken via internal thrusting or ductile deformation as they decelerate prior to arrest. Slope creep caused by gravity moves beds slowly downslope with gentle internal deformation of the original depositional structure. Slides undergo little or no pervasive internal deformation, while slumps undergo partial deformation but the original internal structure is still recognisable in separate blocks. Thicknesses of slides and slumps range from several tens of metres to 1–2 km and travel distances can be up to about 100 km, with displaced volumes of up to 10^{12} m^3, although most flows are considerably smaller.

Dense, Deformed Flows: Rockfalls, Grain flows, Debris Flows and Mudflows

In flows of this type, sediment still moves as one coherent mass, but concentrations can be lower and the mass is generally well mixed, with little or no preservation of remnant structure from the original failed material. Sediment support mechanisms are matrix strength, buoyancy, hindered settling, and grain-grain collisions. Rheologically such flows are plastic (i.e., they have a yield strength). Clast types generally range from purely cohesive in mudflows, to cohesive and/or non-cohesive in debris flows (Figure 2) and purely non-cohesive for grain flows and rockfalls (where movement is by freefall on very steep slopes). These types of flow are formed as a

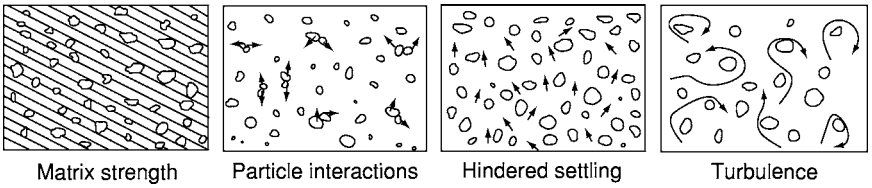


Figure 1 Schematic illustration of the principal grain transport mechanisms, shown in decreasing order of concentration from left to right.

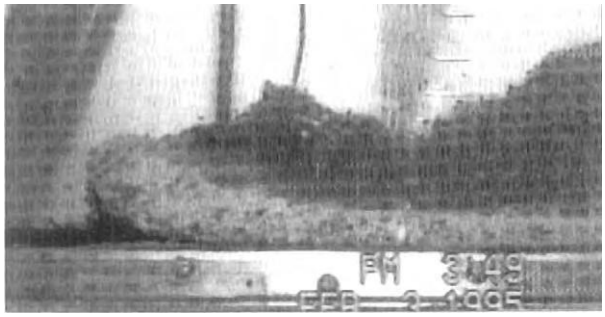


Figure 2 A laboratory debris flow from right to left. Note: a dilute turbidity current has been generated on the upper surface of the debris flow due to erosion of material by fluid shear. (After Mohrig *et al.* (1998) *GSA Bulletin* 110: 387 394.)

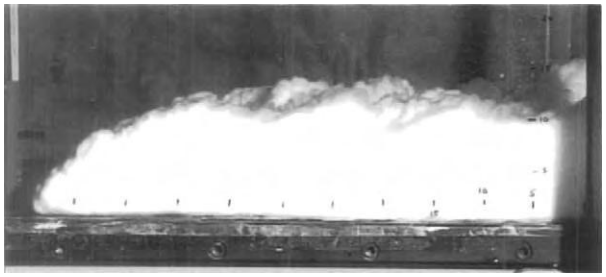


Figure 3 A laboratory turbidity current flow from right to left. Field of view is 55 cm wide. (After McCaffrey *et al.* (2003) *Marine and Petroleum Geology* 20: 851 860, with permission from Elsevier.)

result of rapid internal deformation following slope failure, from high concentration river input or from reconcentration of dilute flows (described below). Flow and deposit thicknesses can be up to several tens of metres with travel distances of several hundreds of kilometres. Erosion can add material to the flow and thus extend both travel distances and size of deposit – neither of which, therefore, necessarily relate to the initial flow mass. Motion will stop once friction is too high and flows will generally deposit *en masse*. Debris flows may develop a rigid plug of material at the top of the flow, where the applied stress falls below the yield strength. Such flows move along a basal zone of deformation, and may progressively 'freeze' from the top downwards, ultimately coming to rest when the freezing interface reaches the substrate.

(Partly) Dilute Flows: Turbidity Currents

In flows of this type, the sediment does not move as one coherent mass (Figure 3). These flows are generally dilute although parts of these flows can be of high concentration, especially near the bed. In the dilute parts of these flows, sediment is transported in either laminar or turbulent suspension. In higher concentration areas additional sediment transport mechanisms, such as grain-grain interactions, hindered settling, and buoyancy effects may also play a role. Rheologically, the dense parts of such flows can behave plastically, but the dilute parts are Newtonian. Concentrations in turbidity currents range from only a few kg m^3 to concentrations approaching those of static, loose-packed sediment. The dilute parts of these flows are commonly strongly vertically density-stratified. Turbidity currents can be formed via dilution of debris flows (see below), directly from river input or from resuspension of sediment.

Turbidity current thicknesses can be up to several hundreds of metres and can increase during flow due to turbulent entrainment of ambient water. Velocities can be up to tens of m s, but more commonly are around 1 m s or less. Larger flows, such as the well-documented Grand Banks event of 1929, may travel distances of a few thousand kilometres, even on nearly flat slopes, although distances of tens to hundreds of kilometres are more common. Sediment eroded during flow can add to the driving force and will increase flow duration and travel distance. Flows will gradually slow down as sediment settles out, with coarse material being deposited proximally and fine material distally. Deposit thicknesses generally are significantly smaller than flow thickness and are on the order of cm to dm, but can be up to multi-metre scale for large flows. However, ongoing sedimentation from flows of long duration can result in deposits whose thickness relates principally to flow longevity rather than flow thickness. Consequently, it is generally more difficult to interpret flow properties from analysis of turbidity current deposits (turbidites) than it is for the denser flow types.

Flow Transformations

Transformations of one flow type into another are common. Initially-dense slide masses may be disrupted due to internal shear, liquefaction, and disaggregation on various scales. If this deformation is sufficiently vigorous all the original structure of the failed material will be lost and the slides transformed into debris flows. In turn, these can transform into turbidity currents by erosion of sediment from the front and top of the dense mass due to ambient fluid shear (Figure 2), by disaggregation and dilution, and by deposition of sediment, diluting the flow. Turbidity currents can be transformed into debris flows if they reconcentrate, for example when mud-rich flows slow down. Further transformation into slides is not possible once the original internal structure is broken up.

The extent of transformation depends on flow size, velocity, and sediment content. Variable degrees of

transformation can lead to the development of different flow types within one current, both vertically and from front to back. This co-occurrence of different flow types is especially common in flows with a dense basal layer and more dilute upper part. Thus, classification schemes which subdivide flows on the basis of discrete flow types do not recognise the diversity of natural flows, in which different types of flow may occur simultaneously and vary in relative importance in time and space as the flows evolve.

Internal and External Influences on Flow Behaviour

Flow behaviour is influenced both by internal factors such as concentration and grain size distribution and external factors such as input conditions and topography.

Flow Velocity

The driving force, and hence velocity of subaqueous gravity currents increases with both concentration and flow size. However, resistance to internal shear will increase with increasing viscosity due to increasing particle concentrations, and with increasing yield strength caused by cohesive particles. This will inhibit the increase of flow velocities. However, because concentration-induced resistance to shear does not scale with flow size, it can more readily be overcome by the higher gravitational driving forces of larger flows, which are, therefore, faster than smaller flows.

Flow Duration and Run-Out Length

Slope failure-induced slumps and slides that do not transform into debris flows and/or turbidity currents will generally be of short duration and have run-out lengths on the order of the initial failure size. If the failed sediment mass does transform into a debris flow, the duration and run-out length depend on the mobility as described above, with larger flows travelling further. However, because debris flows stretch out as they are flowing and because they may incorporate material by erosion, their run-out length may not be directly related to the initial failure size.

The duration and run-out length of turbidity currents depend on their size and sediment content, and hence also on their formation mechanism. Sustained input from rivers or glacial plumes can result in long duration flows, even if the input concentration is low. Turbidity currents that are generated from slope failures can have a short duration input, but tend to stretch considerably due to turbulent mixing and will thus increase in flow duration provided the transported sediment is kept in suspension. The ability of a flow to keep sediment in suspension, known as the flow 'efficiency', directly affects flow run-out lengths. Flow efficiency depends on flow magnitude, with larger flows being more efficient, and on grain size, as finer grains settle out more slowly than coarser grains. The presence of fine sediment in the flow also increases the ability to carry coarse sediment so both types of sediment will be carried further and both flow duration and run-out length will be increased.

Spatial and Temporal Changes to Flow

Flows are influenced both by the input conditions and by the terrain over which flow takes place. Flow behaviour therefore varies both temporally and spatially, causing local areas of erosion and deposition that lead to a deviation from a simple decelerating depositing flow and complicate the depositional pattern. Both spatial and temporal changes in flow behaviour can be caused by changes in sediment content of the flow: erosion adds driving force to the flow and increases velocity, while deposition slows flows down. Temporal changes to flow can also be caused by changing input conditions. River input from floods leads to flows that initially have a progressive increase in velocity followed by a long period of decreasing velocity. In retrogressive failure ongoing detachment of discrete sediment masses will result in pulsed sediment input; the rate of input generally tends to peak rapidly, and then diminish as successive slope failures reduce in size.

Local spatial changes in flow are caused by changes in the topography (**Figure 4**). The angle of the slope on which flow takes place is obviously important for gravity driven flows; when slope angle increases, the flow will go faster although the velocity increase will be diminished by the increase of friction with the ambient water. Nevertheless, small changes in slope angle can change flow behaviour. If the slope angle decreases, very dense flows can be stopped as the basal friction becomes too high. More dilute flows may undergo hydraulic jumps, in which they abruptly thicken and decelerate. This deceleration can cause coarser sediment to be deposited. Local changes to flow can also be caused by changes in the constriction of the flow path. When a flow goes into a constriction, velocity will increase. Where a flow can expand, as at the end of submarine canyons, velocity will decrease.

Momentum Loss

The evolution of flow behaviour can be different along flow-parallel and flow-transverse directions. Momentum will be greater in the direction of flow

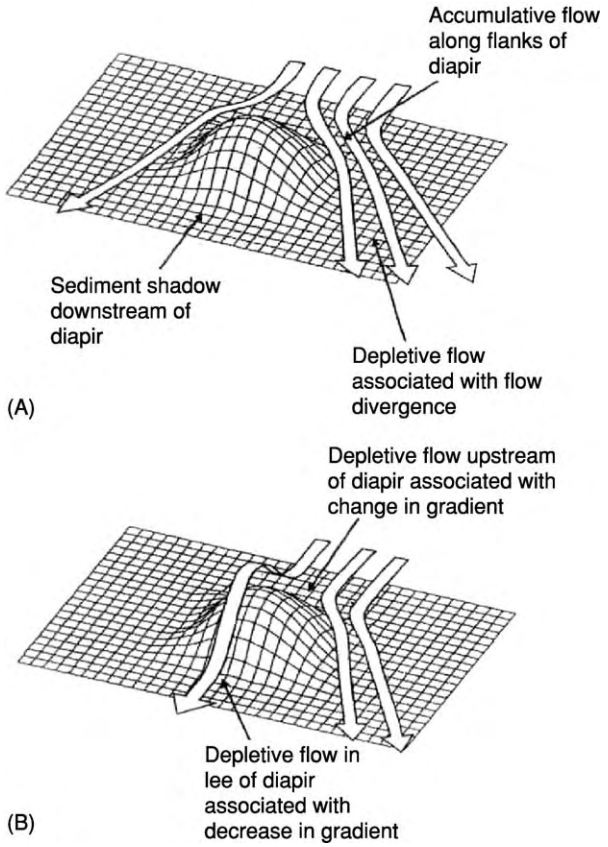


Figure 4 Schematic illustration of the interaction of turbidity currents with (A) high amplitude and (B) low amplitude bathymetry. Flows are uniform if the velocity does not change with distance and are non uniform if the velocity does change. Accumulative flows have spatially increasing velocity while depletive flows have decreasing velocity. (After Kneller and McCaffrey (1995) SEPM, Gulf Coast Section, 137 145.) Published with the permission of the GCSSEPM Foundation; Further copying requires permission of the GCSSEPM Foundation.

than in the transverse direction. For coarse sediment in dilute flows, this means transport is principally in the main flow direction as rapid transverse momentum loss results in rapid deposition. This is less the case for fine-grained sediment, which will stay in suspension more easily and will thus generate momentum for flow in the transverse direction. These differences are not so important in restricted parts of the flow path, such as in canyons, but are important in less confined settings.

Channelised flow

If flows are erosive they can create conduits (incisional channels) both for themselves and for later flows. In aggradational systems, dense flows such as debris flows will start to form levees at their edges where flow becomes too thin to overcome the matrix strength. Sideway expansion of coarser-grained turbidity currents may lead to loss of momentum in the transverse direction, and thus greater rates of off-axis than on-axis deposition. This incipient levee formation may lead to the development of aggradational channels (Figure 5). These channels, which are generally sinuous, and often meandering, partly confine flow and can carry sediment downstream for long distances. Dilute parts of the flow can overtop the levee crests resulting in overspill and deposition of thin sheets of relatively fine-grained sediment that decrease in thickness away from the channel. This winnowing process causes the flows progressively to become relatively depleted in fine grained material, resulting in the development of sandy lobe deposits at the end of relatively muddy channel-levee systems. Levee height decreases downstream and flows become less confined. Like subaerial channels,

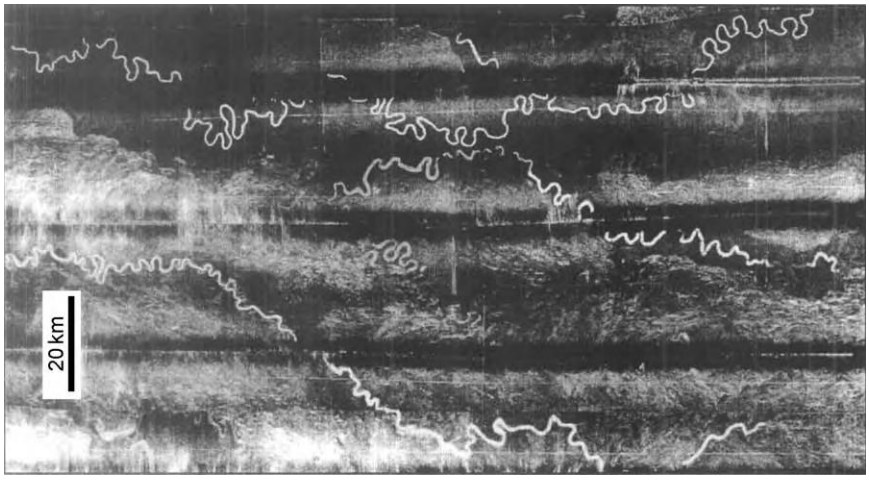


Figure 5 GLORIA image of sinuous submarine channels on the Indus fan. (From Kenyon *et al.* (1995). In: Pickering *et al.* Atlas of Deepwater Environments: architectural style in turbidite systems: London: Chapman and Hall, 89 93.)

aggradational submarine channels may undergo avulsion, resulting in the formation of internally-complex sedimentary fan deposits. Although channels are largely formed by the flows themselves, they can be influenced by pre-existing topography.

Flow in Unconfined Basins

When the basin size is very large compared to the flow, the flows are effectively unconfined. Flows that are not strong enough to erode and that are not captured by antecedent channels can start to spread out. Fine-grained, efficient turbidity currents spread out more evenly in all directions than their coarser-grained counterparts as a result of differing rates of momentum loss. Such unconfined flows can be influenced by Coriolis forces, being deflected to the right in the northern hemisphere, and to the left in the southern hemisphere. Unconfined flows deposit sediment in lobes, with deposit thicknesses decreasing in all directions away from the depocentre. Development of depositional topography may cause subsequent flows to be steered away from depocentres of previous flows and to deposit relatively more of their sediment load in offset positions in a process of autocyclic compensation. Deep-sea fan systems can form in unconfined basins settings through this process.

Flow in Confined Basins

When the basin size is smaller than or of the same size as the flow, the basin margins will prevent flow from expanding and the basin is said to be confined. Processes of topographic interaction induce spatial changes to flow, as detailed above. Flows can overcome small topographic obstacles, but as obstacle height increases relative to the flow height, part or all of the flow will be diverted. Flows in confined basins can be reflected back and forth between different basin margins if enough energy is available, which can result in reworking of the part of the deposit laid down during a previous pass of the flow. If the basin walls are sufficiently high to prevent any of the flow escaping, the basin is said to be ponded. In this case all the sediment is retained in the basin, and any mud present in the flow will be distributed in suspension evenly across the basin and will slowly settle out. The spatial restriction created by confined or ponded basins will hinder flow expansion. Thus, although autocyclic processes can play a role in dictating sedimentary architecture, in general basin fill patterns will be dominated by the confinement. Successive deposits can gradually fill up a basin completely. This can result in flows being able to partially bypass the basin, and enter the next basin downstream, in a process known as fill-and-spill.

Flow Regime Recorded in Depositional Sequences

Erosion and Bypass

If flow power is large enough, erosion can take place, which can remove significant volumes of sediment. This material adds to the driving force of the flow and can lead to acceleration (a process called ignition), and increased flow duration and travel distance. Smaller-scale erosion can form structures that indicate palaeoflow direction, including grooves, where an object is dragged along the bed, and flutes, where turbulent motions erode a characteristic shape that is deeper upstream, and both flares and shallows downstream. Erosion can take place beneath both debris flows and turbidity currents, although flutes require turbulence for their formation, a condition more likely to be met in turbidity currents. Not all flows are capable of erosion, but this does not necessarily mean they deposit their transported load. Bypass of sediment is common in upstream areas and may leave no record in the deposit. This behaviour is closely related to the process of autosuspension, in which sediment is transported by turbulence generated by flow caused by the density difference due to the sediment itself. Strictly speaking, such flows neither erode nor deposit.

Deposition

Eventually all flows, whether they start out as dense or as dilute flows, will lose their momentum and deposit their sediment. Dense flows such as slumps and debris flows will leave deposits whose structure more or less corresponds to that of the flows themselves. This is not the case for turbidity currents, which generally deposit their sediment progressively. Whether deposition takes place at all in turbidity currents depends on local flow competence and capacity. Flow competence indicates which grain sizes can be transported by a flow of a given velocity and flow capacity indicates how much sediment can be carried by a flow of a given velocity. The depositional structures of turbidity current deposits (turbidites) are influenced by the grain sizes carried in the flow, the velocity of the flow, and the sediment fall-out rate. High sediment fall-out rates cause suppression of primary sedimentary structures and lead to the formation of massive (structureless) deposits. The grains in these deposits tend not to be packed at maximum density and commonly re-organise themselves post-depositionally, expelling pore water in the process. This process commonly produces structures that overprint any primary depositional fabric. If fall-out rate is low enough, structures such as ripples and

laminations can be formed, depending on grain size and flow regime. Deposit thickness is influenced both by flow size, with larger flows resulting in thicker deposits, and also by flow duration, with sustained flows being able to deposit thick beds, even if the flows themselves are not particularly large.

Various models have been proposed to describe the vertical succession of features in an idealised turbidite. The most widely applied is the model of Bouma, which describes a sequence deposited by a gradually decelerating turbidity current. Because all flows must eventually wane, full or (more commonly) partial Bouma sequences are developed quite frequently, particularly in relatively distal locations. However, the assumption that flows gradually decelerate over the entire flowpath is unlikely to be met, and many deposits will not look like this or other standard sequences. The influence of temporal changes and spatial changes on deposits will be reflected in terms of bed thickness and grain size distribution (grading). These are schematically presented in the diagram of Kneller for turbidity current deposits (Figure 6). This scheme is strictly valid only for flow where concentration does not change, which limits the applicability of the approach, but it illustrates the idea well. Finally, it should be borne in mind that depositional sequences may be reworked by surface currents, dewatering, and/or bioturbation. These processes may obscure any evidence of flow character that was originally recorded in the deposit.

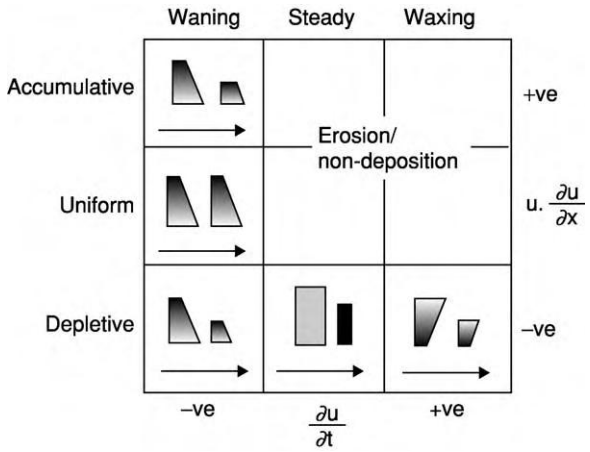


Figure 6 Schematic representation of vertical and lateral grain size variation within single beds as a function of the combined effects of flow steadiness and uniformity. The two logs in each field represent relatively proximal and distal configurations, respectively (arrow indicates flow direction). Flows are steady if velocity does not change with time and are unsteady if the velocity does change with time. Waxing flows have temporally increasing velocity while waning flows have decreasing velocity. Non uniformity definitions are given in Figure 4. (After Kneller and McCaffrey (1995) SEPM, Gulf Coast Section, 137 145.) Published with the permission of the GCSSEPM Foundation; Further copying requires permission of the GCSSEPM Foundation.

Summary

Subaqueous particulate gravity currents may exhibit a wide range of concentrations, magnitudes, grain size, and type and flow velocities, all of which may change as flow develops. Flow behaviour is dictated both by input conditions (affecting flow magnitude, grain size distribution, and duration), and by the flow pathway (including its bathymetry, and the erodibility of the substrate). Thus, subaqueous particulate gravity currents form a complex and variable range of flow types, which together constitute the principal means by which coarser-grained clastic material is transported into the deep ocean.

Further Reading

Allen PA (1997) *Earth surface processes*. Blackwell Science, Oxford.

Hampton MA, Lee HJ, and Locat J (1996) Submarine landslides. *Reviews of Geophysics* 34(1): 33 59.

Kneller BC (1995) Characters of Deep Marine Clastic Systems. *Geological Society Special Publication 94*.

Kuenen Ph H (1950) Turbidity currents of high density. 8th International Geological Congress, London 8: 44 52.

Kuenen Ph H (1952) Estimated size of the Grand Banks turbidity current. *American Journal of Science* 250(12): 874 884.

McCaffrey WD, Kneller BC, and Peakall J (eds.) (2001) Particulate Gravity Currents *IAS Special Publication* 31: 302.

Schwarz HU (1982) Subaqueous slope failures experiments and modern occurrences. *Contributions to Sedimentology* 11: 116.

Simpson JE (1997) Gravity currents in the environment and the laboratory, 2nd edn. Cambridge: Cambridge University Press.

Stow DAV, Reading HG, and Collinson JD (1996) Deep seas. In: Reading HE (ed.) *Sedimentary environments: processes, facies and stratigraphy,* 3rd ed., chapter 10, pp. 395 453. Blackwell Science, Oxford.

Walker RG (1992) Turbidites and Deep Sea Fans. In: Walker RG and James NP (eds.) Facies Models, 3rd edition, ch.13, pp. 239 263, *Geol. Soc.* Canada, St John's Canada.

Deposition from Suspension

I N McCave, University of Cambridge, Cambridge, UK

Introduction

Geological treatments of sediment dynamics generally lose sight of the fact that the last event of dynamic importance that happened to the sediment was that it was deposited. Instead, most accounts concentrate on the process of transport. Of course, a fair amount of work deals with the creation of bedforms, many of which are depositional, but occur in the transport regime of 'steady' flow, as well as in the 'unsteady' regime of flow deceleration which leads to deposition. It is not possible to deal sensibly with the topic of deposition from suspension without some mention of how material is transported, and so this article deals briefly with this aspect after giving an outline of the controlling factors and before describing the processes of deposition.

Almost any material, even boulders, can be transported in an aqueous turbulent suspension if the flow is large and sufficiently rapid. Even gravels were in suspension in the flood following the bursting of the glacial Lake Missoula in western Washington State (USA). However, most material deposited from suspension is mud and fine sand. Indeed, most (>50%) of the sedimentary geological record is of silt and finer sizes (<63 μm). Fine silt and clay, material of $\leq$10 μm, has the peculiar property that it can stick together, thereby transforming its settling velocity distribution. This, in turn, affects its response to changes in factors controlling its transport and deposition, such as the boundary shear stress and turbulence intensity.

The term 'suspension' is normally applied to material supported by turbulence in a boundary layer. However, in the oceans, much past work has referred to 'suspended particulate matter' (SPM) or 'total suspended matter' (TSM) obtained by filtration. This material, unless in the ocean bottom mixed layer, is not suspended but sinking, and thus, in a sense, is being deposited, although it may have several kilometres to go to reach the bottom. This material, comprising 'pelagic flux', is also affected by settling velocity transformations and is included in this article.

This article deals with controlling factors, entry into and maintenance in suspension, and aspects of deposition: pelagic flux and deposition from boundary layers on to flat beds and bedforms. It mainly concerns deposition from water, but some of the diagrams in non-dimensional form are applicable to air, and some comparisons are made with dust deposition from wind.

Controlling Factors

Particle Settling Velocity w_s

The still-water settling velocity of spheres collapses nicely on to a single curve when plotted as a dimensionless Reynolds number Re_p ($= w_s d/\nu$) vs. another dimensionless number used by (amongst others) M.S. Yalin in 1972, and here called Yalin's number: $\Xi = (\Delta\rho_s g d^3/\rho\nu^2)$ (Figure 1) (w_s is the settling velocity, d is the diameter, $\Delta\rho_s = \rho_s - \rho$ is the solid minus fluid density, and ν is the kinematic viscosity). It should be noted that the abscissa Ξ contains material variables only, i.e., particle size and density and fluid viscosity and density. It should also be noted that the curve has two straight line segments and a curved transition joining them. The upper segment is $Re_p = 1.75\Xi^{0.5}$, with a lower limit around $d = 2$ mm, and is thus applicable to gravel in water. The lower segment is Stokes' law, $w_s = \Delta\rho_s g d^2/18\mu$, where μ is the molecular viscosity ($\mu = \rho\nu$), applicable below $Re_p \approx 0.5$ or $d < 100$ μm in water and air. A further order of complexity is introduced by the fact that particles are not often spheres. Dietrich has developed empirical relations that deal with the varying shapes of solid particles which always sink more slowly than spheres. We can note in passing that Figure 1 divides into a coarse end (gravel), in which $w_s \propto d^{1/2}$, and a fine end (silt and clay), in which $w_s \propto d^2$, whilst the transition is occupied by sand for which, roughly, $w_s \propto d$. Density exerts a strong control, but only for a minority of particles – most solids are quartz-carbonate density (2500–2900 kg m^{-3}). The molecular viscosity of water ranges mainly from 0.9×10^{-3} to 1.5×10^{-3} Pa s (25 to 2°C), and thus gives nearly a factor of two variation. For air, the viscosity is $\sim 1.8 \times 10^{-5}$ Pa s, but as its density is only 1.2 kg m^{-3}, the kinematic viscosity is $\sim 1.33 \times 10^{-5}$ m^2 s^{-1}.

Much greater variation is due to the bulk density variations of particles that are not solid: aggregates, hollow particles, and grains containing gas bubbles. There has been little systematic study of the latter, although they are clearly important in hot volcanic dust suspensions (e.g., ignimbrites). The most important classes of hollow particles are foraminifera, diatoms, and radiolaria. Forams are often partially sediment filled, resulting in saturated bulk densities

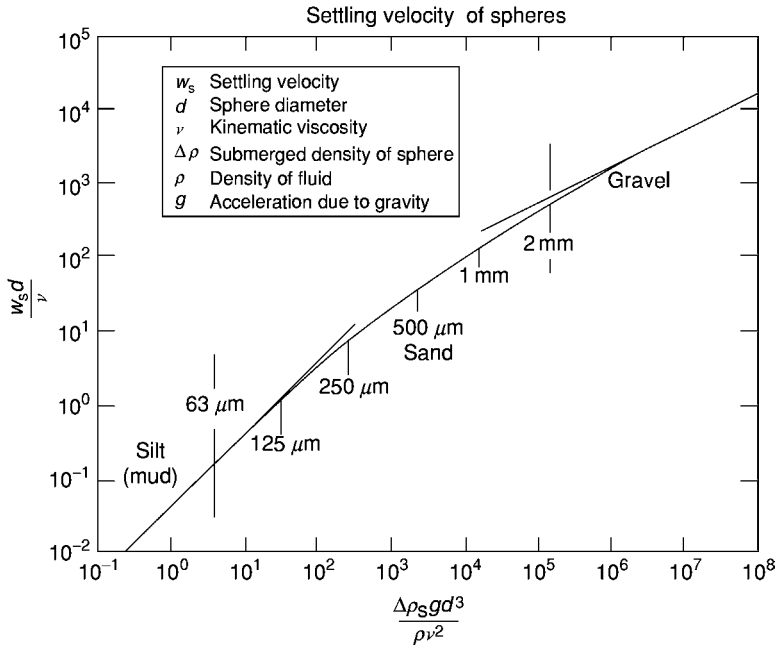


Figure 1 Universal settling velocity curve for spheres. Axes are the particle settling Reynolds number, $Re_p \quad w_s d/\nu$, and Yalin's number, $\Xi \quad \Delta\rho_s g d^3/\rho\nu^2$. This allows variable density particle and size fluid viscosity to be accommodated.

(that is the mean density of the foram shell plus sediment and water in the cavities) of 1150–1550 kg m ^{3}. (For simplicity, a deep seawater density of 1050 kg m ^{3} is used.) This gives a large range of $\Delta\rho_s$ from 100 to 500, which translates straight into a settling velocity straddling the Stokes' boundary, but giving a settling velocity of 125–500 m per day for 200 μm forams. With variable size, density, and viscosity, sinking rates can be from 50 to 1000 m per day. In air, the settling velocities are much faster but, because the viscosity is less, the viscous-dominated Stokes' settling region also persists up to 100 μm. However, the fast falling speed means that sand is almost never suspended by wind.

Aggregation

Most fine silt and clay is deposited as aggregates. These aggregates may be formed by physical (often referred to as flocculation, sometimes coagulation) or biological processes, generally involving the feeding and production of real or pseudo-faeces. The finest particles (d below 1 μm) are brought into contact by molecular buffeting known as Brownian motion. Here, for similar sized (diameters d_i, d_j) spherical particles, the probability of an encounter (i.e., number of collisions per cubic metre of suspension per second) is proportional to $TN_0{d_{ij}}^2$ and inversely proportional to $\mu d_i d_j$ (where d_{ij} is the sum of the diameters of the two colliding particles). This is clearly favoured by high concentration (N_0) and temperature (T) and opposed by viscosity (μ). Larger particles are brought together either by turbulent shear, where the collision probability is proportional to $N{d_{ij}}^3 \, du/dz$, or by larger, fast-sinking particles sweeping up finer ones, like rain falling through mist. In both cases, larger particles grow more rapidly because, when d_i is large and d_j is small, the sum cubed is large, whereas, when the size difference is not great, ${d_{ij}}^3$ is not large either. (This also takes into account the fact that the number concentration distribution of particles is such that there are far fewer large than small particles. The simplest standard is a flat log volume vs. log diameter distribution, which is equivalent to the cumulative number (log) vs. size (log) with a slope of -3. This means a ratio of 1000 1 μm particles to just one of 10 μm diameter.) The rainfall analogy above is particularly apposite for airborne dust, because one method of deposition is washout in which falling rain droplets form aggregates with dust particles by collision and carry them to earth. The other mechanisms of aggregation, Brownian motion, and shear also act on fine particles in air.

A key feature of much aggregation is the presence of organic mucus which acts as 'glue', allowing particles which get close to stick. Many organisms produce mucopolysaccharides, notably bacteria which sit on particles. Although there has been progress in implementing schemes to calculate particle aggregation, they are not yet simple or robust and represent the

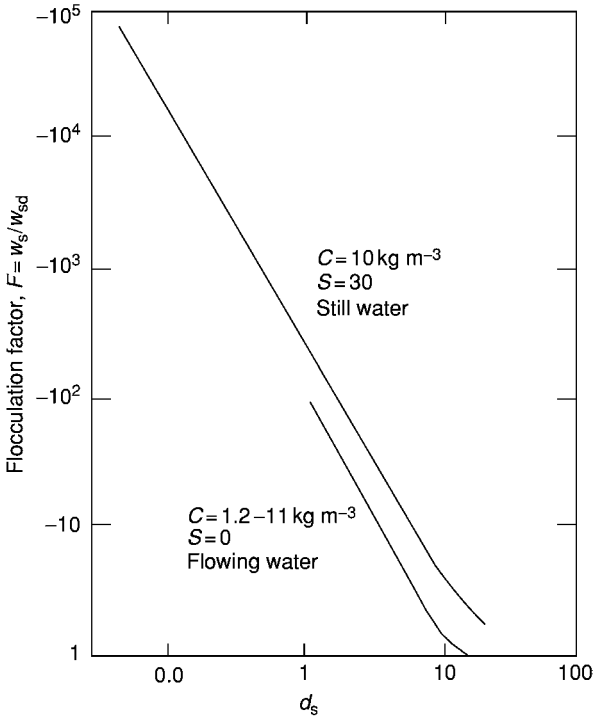


Figure 2 Flocculation factor F (ratio of floc settling velocity to settling velocity of the primary particles from which it is made) vs. diameter of the primary particles. It should be noted that F is negligible in flowing water for $d_{50} > 10\,\mu$m. Data from Migniot C (1968) Etude des proprietes physiques de differents sediments tres fins et de leur comportement sous des actions hydrodynamiques. *La Houille Blanche* 7: 591 620 and Dixit JG (1982) *Resuspension Potential of Deposited Kaolinite Beds*, MS thesis, U. of Florida, Gainesville.

frontier of research in cohesive sediment dynamics. The importance of aggregation is shown in Figure 2, where the 'flocculation factor', the ratio of the settling velocity of an aggregate to the settling velocity of the primary particles from which it is made, can be up to 10^5.

Aggregates are not stable entities. The organic membrane covering faecal pellets decays and the mucus that holds aggregates together also degrades, and so particles fall apart whilst sinking. Aggregates assembled by moderate levels of turbulence in the outer part of the boundary layer may be broken up by more energetic turbulent eddies close to the boundary. The relationship between aggregate size and boundary shear stress is very poorly known and, if the floc diameter $d_f \propto \tau^{\ n}$, then n ranges from 0.25 to 1 (τ is the shear stress in the fluid). The larger aggregates, which can be up to 5 mm in diameter, are thus found in the moderately turbulent, high-concentration environment of the estuarine turbidity maximum above the region within a metre of the bed. Closer to the bed, high shear breaks these large sloppy aggregates into smaller pieces. The density of flocculated sediment decreases as the flocs increase in size. A simple expression based on field data is $\Delta\rho_f = 4.9 d_f^{\ 0.61}$, where the floc excess density $\Delta\rho_f$ is in kg m^{-3} and the floc diameter d_f is in millimetres. This yields floc excess densities of less than 10 kg m^{-3} for 300 μm aggregates (compared with solid particles, where it is $\sim$1600 kg m^3). Nevertheless, because the settling velocity increases as the square of the diameter, large flocs settle at the same speed as fine quartz sand grains, or at >200 m per day, which means that they reach the bottom quickly, resulting in significant clearing of the water in a 10 m deep estuary in a slack-tide period of 2 h.

Boundary Layer Turbulence

A fluid flowing over a surface exerts a drag force on it. The drag at the boundary slows the fluid down, but some distance out, known as the boundary layer thickness, the average flow speed no longer changes much with distance. Most rivers are completely boundary layer as are shallow marine tidal flows. In the atmosphere and deep sea, the boundary layer extends several tens of metres above the surface. Boundary layers are intensely turbulent, and the drag force τ_0 exerted on the bed is related to that intensity because the stress is transmitted by eddies. In the vertical plane, $\tau_0 = -\rho\overline{uw}$, where u is the turbulent component in the flow direction and w is the up and down component (actually perpendicular to stream lines which may not be quite vertical). This expression is very important because u and w are related, so that $\tau_0 \propto \overline{w}^2$. This vertical turbulent velocity is responsible for keeping particles in suspension, and the turbulent stress $\overline{uw}$ either causes aggregation or, at higher values, disaggregates fine particles. The term (τ_0/ρ) has the dimensions of a velocity squared and that velocity is called the shear or friction velocity: $U_* = (\tau_0/\rho)^{1/2}$. From the above, it can be seen that $U_* \propto \overline{w}$.

Regions of the Boundary Layer

Flows may be distinguished as laminar or turbulent on the basis of their Reynolds number, Ul/ν, where l is some relevant length scale (e.g., depth of a river, diameter of a particle) (*see* **Unidirectional Aqueous Flow**). Low Reynolds number flows are laminar, high Reynolds number flows are turbulent. In a turbulent flow, the speed decreases towards the bed because of the drag, so that very close to the bed the flow becomes laminar, or at least dominated by viscosity, in a layer known as the 'viscous sublayer' of the turbulent boundary layer. This is very thin. In water, for a flow that just moves very fine sand ($U_* = 0.01$ m s^{-1}, $\nu = 10^{-6}$ m^2 s^{-1}), $\delta_v = 10\nu/U_*$ is just 1 mm thick. (In

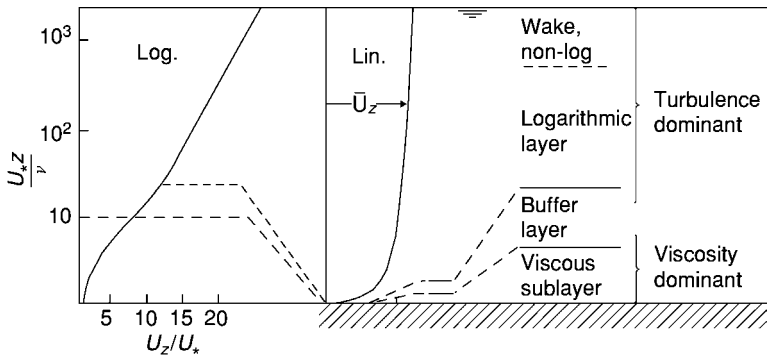


Figure 3 Regions of the turbulent boundary layer for a flow 1 10 m deep. In the centre, the linear representation of flow speed vs. height cannot resolve the viscous sublayer, but the speed vs. log height (z, expressed as a Reynolds number) shows it very well.

air, for the same stress, its thickness is similar, about 0.5 mm.) However, this is ten times the diameter of very fine sand. The shear across this layer is very large; for $U_* = 0.01\,\mathrm{m\,s^{-1}}$, the speed goes from 0 to $0.1\,\mathrm{m\,s^{-1}}$ in just 1 mm. Weak aggregates cannot survive this shear and break up. Above this sublayer, there is a transition ('buffer layer') to a region in which the flow speed varies as the logarithm of height above the bed (Figure 3). As the flow speed decreases, U_* decreases and δ_v increases, so that, in deposition, most particles that are going to become part of the geological record have to get through the viscous-dominated layer. Although viscous dominated, this layer is actually not laminar. Spatially, it has a structure of high- and low-speed streaks, and temporally very high-speed 'bursts' of fluid out of the layer and 'sweeps' of fluid into it from outside. These are associated with stresses typically up to 10 times the average (and extremes of 30 times), and so the mean shear example given above is a minimum, and even strongly bound particles may find themselves ripped apart just as they were getting within sight of the bed and posterity. Above the viscous sublayer, the 'buffer layer' is overlain by a zone in which the flow speed varies as the logarithm of distance from the bed (the 'log layer'). This zone is fully turbulent with eddies becoming longer with height above the bed and turbulence intensity becoming smaller.

The roughness of the bed positively influences the drag and turbulence, but also provides quiet regions in between large grains where fine particles can settle. Fine sediment can thus be deposited in the interstices of gravel, affecting several processes, e.g., the spawning of salmon.

Critical Conditions for Suspension

Two views of the critical suspension condition are as follows: (1) at critical movement conditions, the turbulent intensity can hold particles up, and so suspension depends on whether the particles are ejected from the viscous sublayer; and (2) sublayer ejections are fast, and so suspension depends on whether the vertical turbulent velocity can hold the particles up after injection into the flow. The second view was held by many, but recent work suggests that the first view may be correct. This view is based on high-speed video observations of particles close to the bed, which show that there is a threshold level of shear stress for the particles to respond to turbulent ejections of fluid from the viscous sublayer. The second view would mean that fine to very fine sand would immediately go into suspension as soon as it moved. For example, for 100 μm sand, the critical erosion shear velocity U_* is $0.012\,\mathrm{m\,s^{-1}}$, and the settling velocity of this very fine sand is $0.008\,\mathrm{m\,s^{-1}}$, and so it is capable of being held up by the flow, but video data show that it is not suspended. This means that there is a region of bedload transport for all particles of settling velocity, at least down to $\sim 30\,\mu$m silt. This is shown on a conventional non-dimensional erosion diagram in Figure 4.

The significance of this is that, in a decelerating flow, below the suspension threshold, material may continue to move, but not in suspension. Experimentally it has usually been found easier to determine the critical suspension condition with increasing flow, rather than failure of suspension on decreasing flow. It is generally assumed that the two views are equivalent.

Transport in Suspension

Once material is moved out of the near-bed region, it is held in suspension by the action of fluid turbulence. For this, because the vertical turbulent component of velocity is about the same as the shear velocity U_*, the normal suspension criterion is that $w_s/U_* \leq 1$.

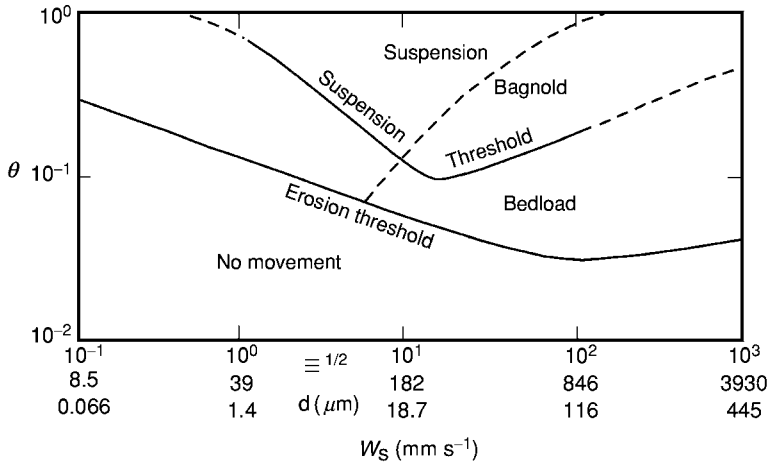


Figure 4 A critical erosion diagram on non dimensional axes with a critical suspension line added. This divides the diagram into regions of suspension, bedload, and no movement. Below the suspension threshold, material falls out and, as the capacity of a flow to carry bedload is limited, deposition will ensue. Two suspension lines are shown, ''suspension threshold'' results from view (1), while ''Bagnold'' expresses view (2); see text $\theta \quad \tau_0/\Delta\rho_s g d$, $\Xi \quad \Delta\rho_s g d^3/\rho\nu^2$.

Particles in 'steady' transport diffuse up from the source at the bed and sink back down under gravity with a balance in steady state. This is expressed as

$$Cw_s + \varepsilon_s dC/dz = 0$$

where the first term is gravity settling and the second is upward diffusion (ε_s is the sediment diffusivity). The result of this is that, for a given value of U_*, the faster settling grains are found closer to the bed and the finer slower settling particles are more uniformly distributed over the flow depth (Figure 5). In the bottom of a deep flow, the concentration at height z in the flow is $C_z = C_a(a/z)^{\zeta}$, where C_a is the concentration at height a (the point near the bed at which a measurement is made) and $\zeta = w_s/\kappa U_*$, where κ is von Karman's constant (0.4). This means that, with our suspension criterion, $w_s \leq U_*$, we would not expect much material in suspension for $\zeta > 2.5$. Figure 5 shows this. Here, it can be seen that relatively fine material (with $\zeta < 0.125$) is distributed throughout the whole flow. This is the fine silt and clay of river 'washload'. Closer to the bed, the relatively coarser sediment is concentrated. Clearly, if the flow slows down, the coarser material will be rapidly deposited because it is only just above the suspension threshold and has very little distance to reach the bed.

In air, there is very little suspension of sand. Above camel height in a 'sandstorm', the suspended material is virtually all silt and clay-sized dust. Saltation (which is bedload) is confined to the lower ~1.5 m, and very little material of >70 μm is carried in suspension.

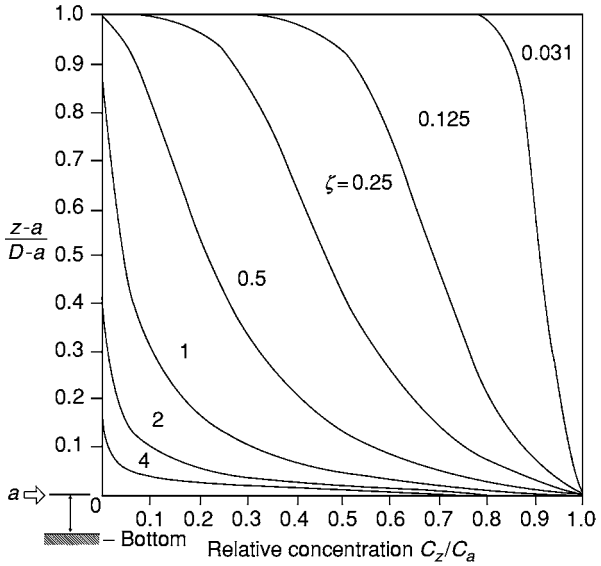


Figure 5 Variation of concentration with height of particles with increasing ratio $\zeta \quad 2.5w_s/U_*$. This shows material with high values of ζ close to the bed. This could be quite fine grained material if the flow has become very slow (low U_*).

Sinking Deposition: Pelagic Flux

The oceans are full of particles that are sinking, some slowly, some fast. The origin of most of this material is from biological production in the upper ocean. Thus, it comprises organic matter, calcium carbonate, and opal. It has been shown theoretically and by the use of sediment traps that most of this material would not reach the seabed were it not for the process of

aggregation which transforms its settling velocity spectrum. The aggregates are faecal pellets and 'marine snow', loose aggregates based on mucus and gelatinous structures made by zooplankton. This material sinks at ~100 m per day, a huge increase over the 2 m per day of a 5 μm coccolith. Aeolian dust rained out on to the sea surface also becomes incorporated into these aggregates, providing a rapid route to the bed.

Only if there is very slow flow at the bed in the bottom boundary layer will these aggregates plummet down directly on to the bed. This is true for most of the ocean most of the time, but some areas have fast currents which can break up the aggregates and control deposition.

Close to continental margins, the action of waves on the outer shelf and internal waves on the shelf-break and slope leads to the resuspension of material. This resuspended sediment spreads out on surfaces of density contrast as intermediate nepheloid ('cloudy') layers and flows down-slope in bottom nepheloid layers. These turbid layers are found all over the sea bottom, some more concentrated than others. Most fine sediment deposition involves some transport and removal from nepheloid layers, except for coarser (sand-sized) components which simply sink to the bed. Continental margins thus contain much material that is rained out of suspension and moved in bottom nepheloid layers during deposition.

Deposition from Turbulent Boundary Layers

The 'deposition' of bedload is rather straightforward: it stops moving. This occurs in water at a shear stress only slightly lower than the critical erosion stress. In fact, we cannot measure the erosion stress precisely enough to distinguish between erosion and deposition stresses, and so they are effectively the same. In air, high-speed grain bombardment of the bed keeps the bedload moving until the stress has been reduced to ~80% of the critical value. For suspended sediment, once the stress has decreased below the suspension threshold (see Figure 4), material will sink into the near-bed region, thereby increasing the concentration and causing some material to be deposited.

One well-documented consequence of the concentration increase is that the flow becomes density-stratified. This reduces turbulence intensity by absorbing turbulent energy in order to keep grains up. A reduction in turbulence means a reduction in shear stress, and deposition ensues. An extreme case is the suppression of the ability of a highly concentrated flow to sort sediment, resulting in the deposition of the massive graded A division at the base of Bouma-type turbidites. As the concentration decreases, turbulence is sufficient to sort the succeeding B division into laminae.

Deposition from different modes of transport is reflected in sediment size distributions. In a typical S-shaped cumulative size frequency curve for sands, the coarse tail reflects material that was always carried as bedload, the central part of the distribution reflects material carried intermittently in suspension, and the fine tail is made up of the 'washload' – long-distance suspensions. However, for air, the coarse tail is the 'creep' part of the bedload, the central part is the saltation component, and the fine tail is sand and dust from suspension.

Processes of Deposition

Fine sediment may reach the bottom in one of three ways. It may settle to the bed under gravity, it may impinge on the bottom as a result of molecular agitation in Brownian diffusion, or it may be transported downwards by eddy diffusion. The critical region of particle transport for all three of these processes lies within the viscous sublayer of a turbulent boundary layer, because, in most cases in which deposition occurs, a large part of the bed is covered by the sublayer. Simple calculations show that settling, even of 1 μm particles, is several orders of magnitude greater than the diffusive deposition rate. As diffusion is only likely to be important for the smallest particles, we can safely neglect it, and consider deposition to be controlled by particle settling through the sublayer to the bed. In the case of rain washout of dust and falling of 'marine snow' in areas of slow deep-sea currents, particles reach the bed at relatively high speed with no intervention of turbulence or the viscous sublayer.

Rate of Deposition

Experiments in flowing water show that, below a certain shear stress τ_d, for $C_0 < 0.30\,\mathrm{kg\,m^{-3}}$, the concentration in suspension C_t decreases exponentially with time t

$$C_t = C_0 \exp(-w_s p t/D)$$

where D is the depth of flow (or thickness of the boundary layer), C_0 is the initial concentration, and p is the probability of deposition; the probability is given by $p = (1 - \tau_0/\tau_d)$ (this includes nearly all normal marine conditions; only intense storms, estuaries, and mass flows have higher values). In this expression, τ_d is the limiting shear stress for deposition, the stress below which all the sediment will eventually deposit. This yields $R_d = C_b w_s (1 - \tau_0/\tau_d)$ for the rate of deposition R_d ($\mathrm{kg\,m^{-2}\,s^{-1}}$). Here, C_b

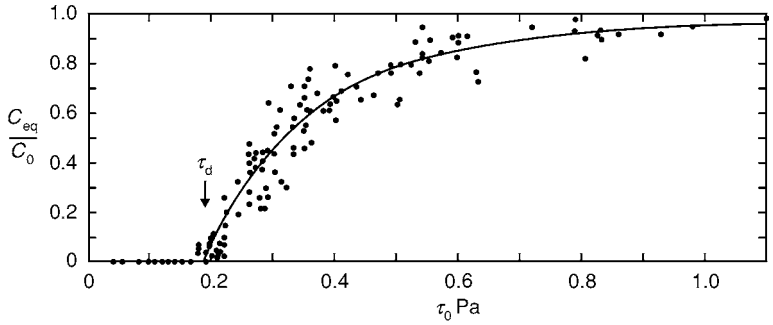


Figure 6 Diagram showing the equilibrium concentration (C_{eq}) of material in suspension as a ratio with the amount initially in suspension (C_0, 1.25 kg m ^{3} in this case) as a function of the bottom shear stress (τ_0). This shows that some material is deposited at shear stresses as high as 0.6 Pa, but that, below τ_0 0.2 Pa, all is deposited, thus defining the limiting stress for deposition (τ_d) of this material. Reproduced from McCave IN (1984) Erosion, transport and deposition of fine grained marine sediments. In: Stow DAV and Piper DJW (eds.) *Fine Grained Sediments: Deep Sea Processes and Facies, Special Publication, Geological Society of London 15*, pp. 35 69.

is the value near the bed. If there is no flow, this reduces simply to the settling flux $C_b w_s$.

At higher concentrations, $C = 0.3$–$10\,\mathrm{kg\,m^{-3}}$, C declines logarithmically with the logarithm of time, $\log C = -K \log t + \text{constant}$, in which $K = 10^3(1-\tau_0/\tau_d)/D$. It has also been noted that, at high concentration, some material is deposited at $\tau_0 > \tau_d$ (Figure 6); an equilibrium concentration is attained, which reduces as τ_0 is reduced to τ_d. This means that some mud can be deposited from relatively fast flows, as long as the concentration is fairly high ($>1\,\mathrm{kg\,m^{-3}}$). The significance of this is that, after storms or under turbidity currents, fine sediment may be deposited under flow speeds of several tens of centimetres per second. The deposition of mud at $\tau_0 = 0.4$–0.5 Pa in Figure 6 is occurring under conditions capable of moving coarse sand of 0.5–1 mm. It is emphatically not the case that mud is only deposited under 'quiet water' conditions when sand cannot be moved.

Limiting Shear Stress for Deposition τ_d

The value of this is not well known, but is probably related to the diameter or, more properly, the settling velocity of the particles, whether aggregates or single grains. The safest assumption is that it is given by the critical erosion stress for non-cohesive grains because, below this value, movement ceases and any grain reaching the bed would be removed from the transport system. This is shown in the critical erosion diagram (Figure 4). An alternative, based on measurements in a laminar flow cell, is $\tau_d = 0.048\Delta\rho_s g d$.

Deposits Formed from Currents

Often, fine sediment deposition may occur from flows of, for example, 0.1–0.2 m s ^{1}. What influence might this have on the character of deposits? A current of 0.1 m s ^{1} in a deep boundary layer having a shear velocity of 4×10^{-3} m s ^{1} allows the deposition of particles larger than 20 μm, but the deposition of finer particles is suppressed. Work on suspended material in nepheloid layers has shown that it comprises aggregates made of silt and clay-sized particles and organic matter. The suppression of the deposition of the finer particles must result in a more silty deposit, but does not eliminate clay completely, because some is caught up in larger fast-settling aggregates, the strongest of which survive stresses in the buffer layer and are deposited. Thus, there is fractionation of a suspension during deposition to yield a more (or less) silty deposit. Some people tend to think that this results from 'winnowing', but this is a process of selective erosion not deposition. Two sorting processes occur: (1) fractionation during deposition, yielding more silty accumulations by deposition under higher shear stress; and (2) fractionation during intermittent erosion, yielding a more compact deposit with (micro) erosion surfaces marked by thin lag layers of terrigenous coarse silt and sand grains and foraminifera.

Above about 10 μm, the flocculation factor in flowing water becomes quite small, as many aggregates are broken up by flow in the buffer layer, particularly by strong flow (but may re-aggregate when away from the bed) (Figure 2). This means that, under stronger flows, this material can be size sorted according to its primary grain size. This part of the size spectrum (10–63 μm) comprises what has been called 'sortable silt' (as opposed to the cohesive material finer than 10 μm), and its mean size has been used as an index of the flow speed of the depositing current. The method gives results showing a striking correspondence to climate change-driven deep circulation changes.

Bedforms from Suspension

Bedforms from sandy suspensions are covered in (*see* **Sedimentary Processes:** Depositional Sedimentary Structures). Some structures seen in the deep sea, but rarely preserved in mudstones, are worth noting.

Mud Waves

Mud waves are regular undulations of the sediment surface with wavelengths of about 0.5–3 km and heights of 10–100 m. Most mud waves are very nearly symmetrical, but they contain subsurface layering that indicates migration. Observed migration usually is up-slope and up-current, but instances of down-current migration have been observed. Under a simple flow, maximum shear stress is expected on the up-stream face of a wavy bedform, and lower shear stress, with a greater deposition rate, on the downstream side. This would give downstream migration of the wave. Commonly observed upstream migration has suggested to some that mud waves are analogous to fluvial antidunes developed under a supercritical flow.

An alternative is that the mud waves form under internal lee waves initially triggered by an upstream topographical disturbance, without the necessity for a supercritical flow speed. Temperature data over mud waves show that such an upstream phase shift does occur, that the implicit internal-wave phase velocity is $0.05\,m\,s^{-1}$, and that this is very similar to the measured flow velocity required for the internal wave to be stationary. The flow pattern over the waves has widely spaced stream lines, giving a small velocity gradient and shear stress (= high deposition rate) on the upstream slope and the opposite on the downstream slope. This would give the observed upstream migration.

Longitudinal Ripples

Longitudinal ripples are elongated features parallel to the depositing flow, probably with helical secondary circulation involved in their formation. In the deep sea, they are 5–15 cm high, 0.25–1 m wide, spaced at 1–5 m apart, and up to 10 m long, and have a generally symmetrical cross-section with sides slightly concave upwards (Figure 7A). In many cases, the ripples have a mound of biological origin at the upstream end. Surface markings on some ripples demonstrate the action of oblique flows, with flow separation and a zone of helical reversed flow on the lee side.

Dating by ^{234}Th (half-life, 22 days) suggests that longitudinal ripples form by deposition from suspension, occurring in a few episodes of very rapid deposition following deep-sea storms. Subsequently, the ripple is scoured by flows that may be oblique to its trend, giving the surface markings seen in deep-sea photographs. These structures have not been recorded from shallow marine or estuarine muds, although closely spaced features up to a few tens of centimetres apart have been seen on tidal flats.

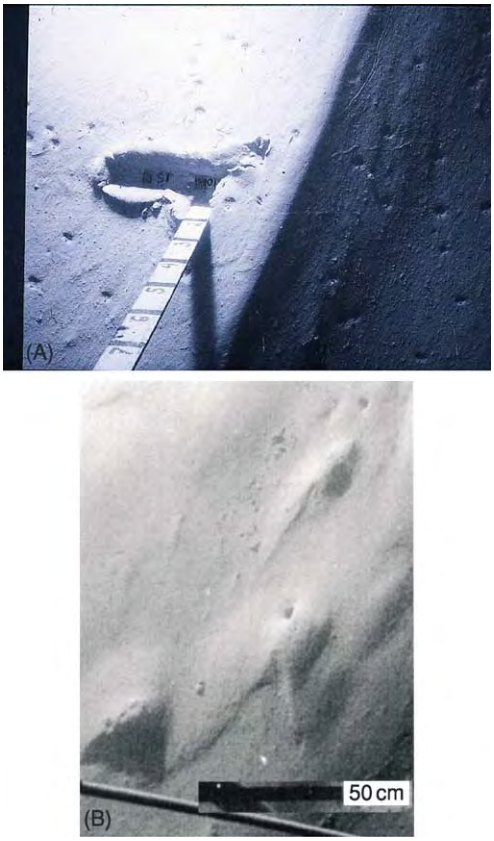


Figure 7 Photographs of bedforms in mud from the deep sea. (A) Longitudinal ripple. Scale bar has penetrated the mud; width of view, ~40 cm; relief of ripple, ~12 cm. (B) Barchan ripples. Both from the Nova Scotian Rise at 4800 m depth.

Smaller, Current-Controlled Bedforms

Smaller, current-controlled bedforms are also revealed by deep-sea photography. The photographed features of the seabed can be arranged in a sequence indicative of increasing flow speed. The progression is from tranquil seafloor (biological mounds, tracks, trails, and faecal pellets), through increasing overprinting by current effects, to features showing clear evidence of erosion. Biological activity is almost ubiquitous, so that a smoothed surface is indicative of an appreciable current, sufficient to remove the surface effects of biota. The most common features are actually biologically produced faecal mounds, tracks and trails, and pelleted surfaces. Mounds are often modified by current activity, the most frequent structure being mound-and-tail formed by lee-side deposition. As suggested above, it may be that longitudinal ripples are very large tails on mounds. Both structures are

excellent current direction indicators. Transverse ripples are sometimes seen in muds, but barchan-shaped (crescentic) ripples are more common (Figure 7B). Some are formed rapidly on deposition from high-concentration suspensions in deep-sea storms, and others are winnowed crescentic silt ripples. Both tend to nucleate around biogenic mounds. Unfortunately, the preservation potential of these structures as distinctive stratification patterns is negligible. Mud is so nutritious that it is populated by a rich burrowing infauna (producing the mounds) and the main structure is pervasive bioturbation.

Nomenclature

Parameter

$\Delta\rho_s$	density difference($=\rho_s-\rho$); unit, kg m^{-3}; dimension, M L^{-3}; value, 1650 (water), 2650 (air)
δ	boundary layer thickness; unit, m (mm); dimension, L
ε	eddy diffusivity; unit, m^2 s^{-1}; dimension, L^2 T^{-1}
μ	dynamic viscosity; unit, Pa s; dimension, M L^{-1} T^{-1}; value, 1×10^{-3} (water), 1.8×10^{-5}(air)
ν	kinematic viscosity ($=\mu/\rho$); unit, m^2 s^{-1}; dimension, L^2 T^{-1}; value, 1×10^{-6} (water),1.5×10^{-5} (air)
ρ	fluid density; unit, kg m^{-3}; dimension, M L^{-3}; value, 1000 (water), 1.2 (air)
ρ_s	sediment density; unit, kg m^{-3}; dimension, M L^{-3}; value, 2650 (water and air)
τ	shear stress; unit, Pa ($=$ N m^{-2}); dimension, M L^{-1} T^{-2}
C	concentration (by mass/volume); unit, kg m^{-3}; dimension, M L^{-3}
D	flow depth; unit, m; dimension, L
d	grain size; unit, m, mm, μm; dimension, L
g	acceleration due to gravity; unit, m s^{-2}; dimension, L T^{-2}; value, 9.8 (air)
J	collision probability; unit, m^3 s^{-1}; dimension, L^3 T^{-1}
N	concentration (by number/volume); unit, m^{-3}; dimension, L^{-3}
U_*	shear velocity ($=\sqrt{(\tau/\rho)}$); unit, m s^{-1}; dimension, L T^{-1}
u,v,w	velocity components (w, vertical); unit, m s^{-1}; dimension, L T^{-1}
x,y,z	space coordinates; unit, m; dimension, L

Subscripts

0	at the bed, or at $t=0$, e.g., τ_0, U_0, N_0
50	50 percentile value, median (e.g., d_{50})
B	Brownian
b	near bed (e.g., C_b)
c	critical value (e.g., τ_c for critical erosion stress)
d	deposition (e.g., τ_d)
f	floc
i,j	i and j particles
p	particle
S	shear
s	sediment (e.g., ρ_s)
v	viscous sublayer (e.g., δ_v)
z	at height z above the bed (e.g., U_z)

Superscripts

— above	time-averaged value parameter

Constants and Dimensionless Numbers

ζ	Rouse number; $w_s/\kappa U_*$
κ	von Karman's constant; 0.4
Ξ	Yalin's number; $[(\rho_s-\rho)gd^3]/(\rho\nu^2)$
k	Boltzman's constant; 1.38×10^{-23} J°K^{-1}
Re	Reynolds number; U_*d/ν, w_sd/ν, etc.

Equations

$\delta_v = 10\nu/U_*$	Viscous sublayer thickness
$w_s = \Delta\rho g d^2/18\mu$	Stokes' settling speed

See Also

Sedimentary Environments: Contourites; Storms and Storm Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures; Aeolian Processes; Deep Water Processes and Deposits; Particle-Driven Subaqueous Gravity Processes. **Unidirectional Aqueous Flow**.

Further Reading

Allen JRL (1985) *Principles of Physical Sedimentology*, ch. 6 and 7. London: Allen & Unwin.

Dade WB, Hogg AJ, and Boudreau BP (2001) Physics of flow above the sediment water interface. In: Boudreau BP and Jorgensen BB (eds.) *The Benthic Boundary Layer: Transport Processes and Biogeochemistry,* pp. 4 43. New York: Oxford University Press.

Dietrich WE (1982) Settling velocity of natural particles. *Water Resources Research* 18: 1615 1626.

Friedlander SK (1977) *Smoke, Dust and Haze*. New York: Wiley Interscience.

Heezen BC and Hollister CD (1971) *The Face of the Deep*, ch. 9. New York: Oxford University Press.

Hill PS and McCave IN (2001) Suspended particle trans port in benthic boundary layers. In: Boudreau BP and

Jorgensen BB (eds.) *The Benthic Boundary Layer: Transport Processes and Biogeochemistry*, pp. 78 103. New York: Oxford University Press.
Leeder MR (1999) *Sedimentology and Sedimentary Basins*, ch. 4 6. Oxford: Blackwell Science.
McCave IN (1984) Erosion, transport and deposition of fine grained marine sediments. In: Stow DAV and Piper DJW (eds.) *Fine Grained Sediments: Deep Sea Processes and Facies, Special Publication of the Geo logical Society of London 15*, pp. 35 69. London: Geological Society.
McCave IN (2001) Nepheloid layers. In: Steele JH, Thorpe SA, and Turekian KK (eds.) *Encyclopaedia of Ocean Sciences*, vol. 4, pp. 1861 1870. London: Academic Press.
Mehta AJ (ed.) (1993) *Nearshore and Estuarine Cohesive Sediment Transport. Coastal and Estuarine Studies*, vol. 42. Washington DC: American Geophysical Union.
Miller MC, McCave IN, and Komar PD (1977) Threshold of sediment motion under unidirectional currents. *Sedi mentology* 24: 507 527.
Pye K (1987) *Aeolian Dust and Dust Deposits*. London: Academic Press.

Fluxes and Budgets

L Frostick, University of Hull, Hull, UK

Introduction

The word 'flux' when applied to sediments has come to mean the movement of particles of rock from upland areas down to a receiving basin, the fundamental processes of landscape evolution, and the geological cycle (Figure 1). The basin which receives the sediment can be terrestrial, either a lake, inland sea, or valley, but the most important basins are the seas and oceans which eventually claim more than 99% of the sediment produced on land. Understanding how, where, and why sediment moves into these basins is fundamental to interpreting and predicting the way in which a basin was formed or might evolve in the future and is, therefore, central to the economically important discipline of basin analysis.

Controls on Sediment Fluxes

The rate of sediment flux to basins is governed by a complex series of interactions amongst the physical and chemical processes which bring about rock uplift, weathering, erosion, and transportation (*see* **Sedimentary Rocks:** Mineralogy and Classification). The quantities of sediment that arrive at a receiving basin are the product of these processes integrated across the total area of supply. As the contributing processes can and do alter in both space and time, sediment fluxes also vary at a range of scales in response to geological and climatic changes. Understanding the controlling processes is, therefore, central to predicting fluxes which are, themselves, key to developing accurate models of basin development.

Weathering

The reason why rock fragments are worn away is inextricably linked with the tectonic processes which produce uplift. As rocks rise up to form hills or mountains and successive surface layers are stripped off, the minerals contained in them move from a dry, hot, high pressure environment to one with an abundant water supply from rainfall and where the temperature and pressure are relatively low (*see* **Sedimentary Environments:** Depositional Systems and Facies). Under these new conditions many minerals become unstable and begin to break down. This process is called weathering and it causes what were originally solid rocks to fragment into smaller fragments which are more easily moved. It is these particles, along with the dissolved products of weathering which, once transported and deposited, make up all sediments and sedimentary rocks (*see* **Weathering**). The way in which a rock breaks down is directly related to its composition. Rock forming minerals are stable at different temperatures and pressures and those that form under conditions most unlike those at the Earth's surface are the most unstable. Quartz is the rock forming mineral that is most stable during weathering and this is the reason why quartz is the predominant mineral in present day beach and river sands and is also common in most ancient sandstones. The rate at which weathering occurs depends on climate, with rapid breakdown favoured by the high rainfall and temperatures of tropical areas and slow weathering occurring where water is absent or solid, i.e., in deserts and arctic zones (*see* **Sedimentary Environments:** Deserts).

Transport

The breakdown products of weathering are transported away from their site of formation and downslope in several ways. On steeper slopes, and for the

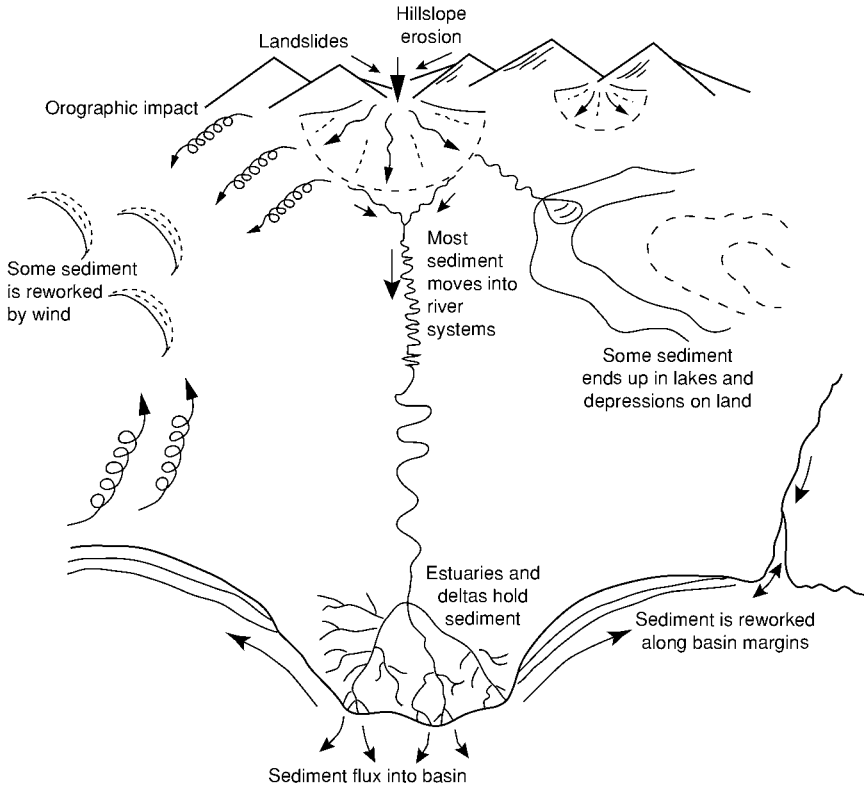


Figure 1 Diagrammatic representation of the way in which sediment moves through the landscape from the mountains to the ocean basins. (Adapted from Frostick LE and Jones SJ (2002) Impact of periodicity on sediment flux in alluvial systems: grain to basin scale. *Geological Society of London Special Publication* 191: 81 95.)

larger particles this may occur as a result of gravity alone acting on the particles, which cause rolling, sliding, and avalanching. Such processes are episodic and often linked with the development of instability as rainfall wets the slope (Figure 2). Some material may be removed by wind but this is important only in desert areas and steep rocky slopes which are devoid of the protection offered by vegetation. In all areas with flowing water it is the movement of that medium that induces sediment to move down the slopes and into adjacent valleys. The water sometimes flows as a shallow sheet over wide areas (called overland flow) but more often collects into small channels, generally known as rills. From here the sediment moves into streams and rivers to complete its journey to the receiving basin. Over the majority of the continental landmasses, where temperatures are high enough for water to remain liquid for most of the time, rivers are the main transporting medium for sediment and flux rates, therefore, vary with the character of the river network. Large, long-lived and integrated networks, such as those of the Mississippi, Congo, and Amazon rivers, deliver large volumes of water and quantities of sediment over long periods. The Mississippi, for example, brings to the Gulf of Mexico one tonne of sediment for every 400 tonnes of water during periods of flooding. When the river is flowing less fast, this may diminish to 0.35 tonne. At the present time, nearly 70% of the total sediment supplied by rivers to marine basins comes from five large rivers, the Ganges/Brahmaputra, Amazon, Huang Ho, Irrawaddy, and Mississippi (Table 1).

In arctic and subarctic areas and where mountains are sufficiently high to generate significant quantities of snowfall, glaciers become a major agent for sediment transport. In the past, shifts in the climatic balance have led to major glacial events when ice-sheets spread out from the polar caps and engulfed large areas of previously temperate landmasses. During these periods, soil and other surface deposits are scoured from the landscape and accumulate in basins.

Climate and Tectonism

Both the character and quantity of sediment carried by a river will reflect the climate and geology of the drainage basin. Even in adjacent river basins, differences in these factors can lead to huge contrasts in sediment fluxes. One example of this can be seen

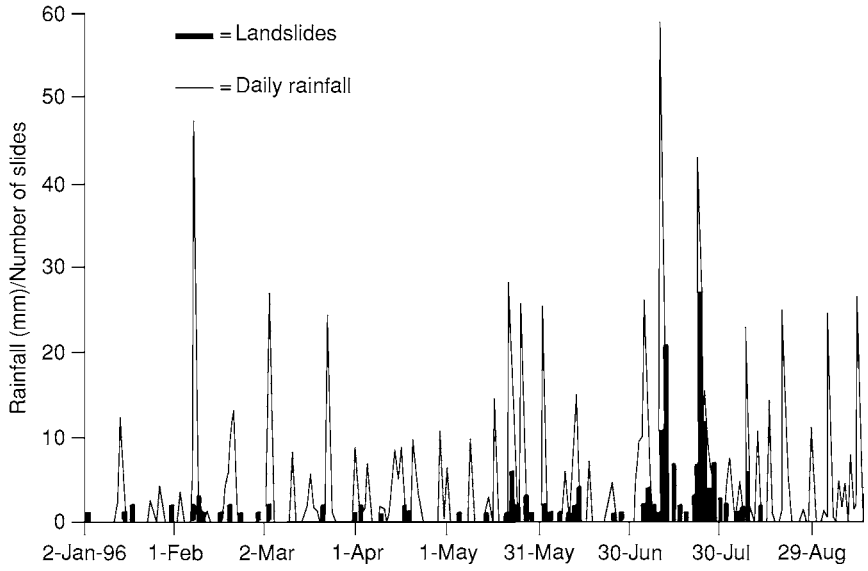


Figure 2 Diagram of the relationship between avalanche frequency and the occurrence of rainfall in the mountains near Wellington, New Zealand. The process is very episodic and each event will produce a 'slug' of sediment which will move down adjacent streams. (Adapted from Crozier MJ (1999) Prediction of rainfall triggered landslides: a test of the antecedent water status model. *Earth Surface Processes and Landforms* 24: 825 833.)

Table 1 The 24 rivers of the world with the largest annual sediment fluxes. Major Rivers Ranked by Sediment Load

Ranking	*River*	*Mean water discharge, 10^3 m^3 s^{-1}*	*Mean sediment load, 10^6 ta^{-1}*
1	Ganges/Brahmaputra	31	1821
2	Amazon	200	1190
3	Huang Ho	1	922
4	Irrawaddy	13.5	356
5	Mississippi	18	352
6	Magdalena	7	240
7	Mekong	21	219
8	Orinoco	36	181
9	Indus	7.5	179
10	Mackenzie	8	144
11	Danube	6.5	136
12	Paraná	15	118
13	Rhone	2	96
14	Yukon	7	94
15	Congo	41	85
16	Volga	8	81
17	Yenisei	17.5	75
18	St Lawrence	13	75
19	Lena	160	68
20	Ob	14	59
21	Zambezi	2.5	45
22	Niger	5	39
23	Murray Darling	1	38
24	Columbia	6	36

in Brazil, where the sediment laden Rio Solimoes meets the sediment poor Rio Negro to form the Amazon River in a spectacular river confluence where the contrast is so marked that it can be seen from space (Figure 3). Sediment fluxes are at a maximum where tropical weathering results in rapid release of particles, large rivers are generated by high rainfall, and tectonic activity promotes uplift and steep slopes. One such area is the Himalayas where the monsoon rains sweep sediment into the vast Indus, Ganges, and Bramaputra rivers. The courses of these rivers and the locations of their outlets into the Indian Ocean are also controlled by tectonic activity, areas of uplift shedding water into adjacent lowland areas. At the mouths of these rivers much of the sediment load is deposited and deltas may build out into the adjacent basin. Large accumulations of sediment on the continental shelf can also lead to the development of turbidity currents which sweep material offshore into large submarine fans (e.g., the Bengal fan) (*see* **Sedimentary Environments:** Shoreline and Shoreface Deposits). Such localised deposits are characteristic of all areas where rivers debouch into larger water bodies and their morphology and sedimentology are largely controlled by the nature of basin processes actively redistributing the sediment.

Basin Processes

The character of the basin receiving the material removed from the land surface will control its dispersion. The balance between subsidence rates and sediment supply is particularly important, since if a deposit is buried rapidly it is less likely to be removed and reworked by waves and currents. In some areas sediments form large river deltas, e.g., the Mississippi (Figure 4), in others the deposit is submerged and

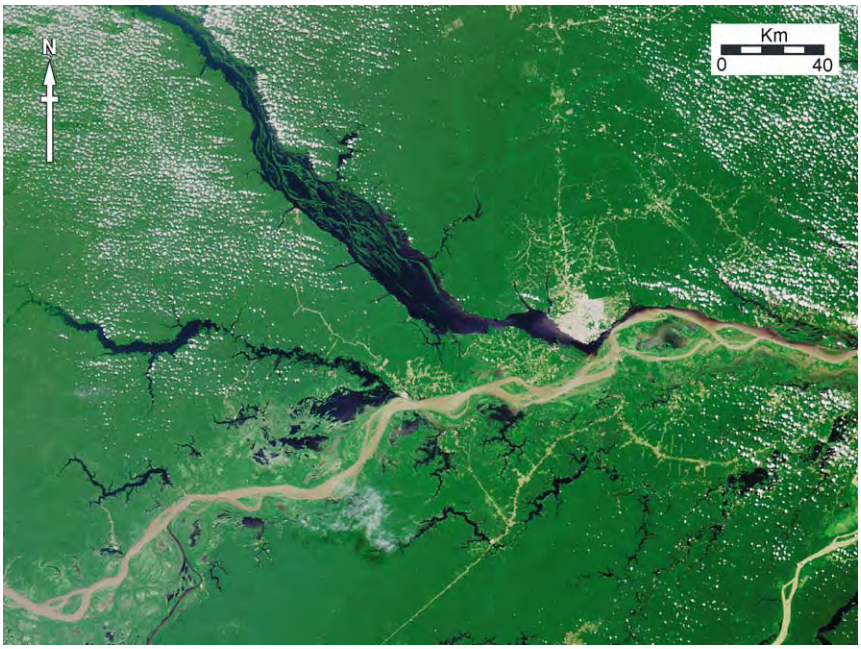


Figure 3 Satellite remote sensing image of the confluence between the sediment laden Rio Solimoes (showing brown) and the clear waters of the Rio Negro (showing black) to form the Amazon River. The Solimoes crosses soft glacial silt and sand, whereas the Negro passes through old hard rocks resistant to weathering. The two water bodies remain essentially separate for tens of kilometres downstream of the confluence. (Picture taken from the TERRA satellite using the multi angle imaging spectroradiometer on 23/7/2000.)

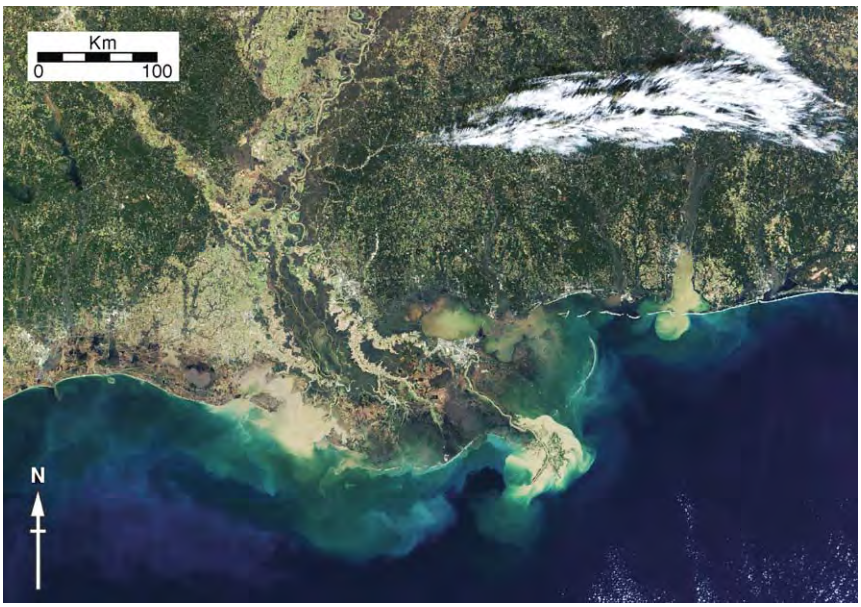


Figure 4 Satellite remote sensing image of the Mississippi Delta on the south east coast of the USA, showing a sediment plume issuing from the mouth of the main channel. This sediment will settle out in the adjacent ocean basin. (Picture taken from the TERRA satellite using the moderate resolution imaging spectroradiometer.)

Figure 5 Satellite remote sensing image of tidal sand banks off the coast of Guinea in West Africa. Note the plumes of sediment moving along the channels between the banks. The high tidal velocities are reworking sediment supplied by local rivers. (Picture taken from LANDSAT 4TM using bands 3, 2, and 1 for red green and blue colour channels, respectively.)

forms either a single fan or breaks up into a series of sand-banks (see, e.g., the coast of Guinea, West Africa, shown in Figure 5). As with the supply of sediment, the intensity of basin processes, such as tides, waves, and currents will also vary over distances, giving a complex spectrum of interactions that can lead to a plethora of deposit types.

Wind Blown Sediment

The flux of sediment in wind may be a minor factor in most areas today, except from in our major deserts, e.g., the Sahara, but in the past it has been significant and has led to the accumulation of extensive loess deposits, e.g., in China. Dust loadings in the atmosphere have changed in response to climate change, both on the scale of millennia, e.g., during the last Ice Age (Late Glacial Maximum) and at a decadal scale, e.g., in response to natural oscillations in climate associated with changes in ocean currents (e.g., the North Atlantic Oscillation). At present, most dust is derived from the worlds desert areas where reduced vegetation allows high-speed winds to pick up predominantly silt-sized particles and transport them for thousands of kilometres before finally depositing them downwind. The Sahara is the world's major source of wind-blown dust, producing between 400 and 700×10^6 tons.a ^{1}, approximately half the total amount of wind blown or aeolian dust estimated as being supplied to all the worlds oceans (*see* **Sedimentary Processes:** Aeolian Processes). At times of major dust storms, the dust being blown out of the Sahara can travel long distances and penetrate far out into the Atlantic Ocean. The composition of this dust is important as it adds nutrients to the deep oceans and influences their productivity (*see* **Sedimentary Processes:** Deep Water Processes and Deposits). Although quartz is the dominant mineralogy, dusts may contain compounds with appreciable proportions of aluminium, iron, magnesium, and calcium.

Flux Variations Over Time

Sediment fluxes at a point will vary over both short and long time scales as a result of temporal changes in any of the controlling variables. At the longer time-scales of geology, both climate change and tectonic uplift can bring about large-scale changes in flux rates. One example of the impact of climate change is the large-scale fluctuations in sediment movement during glacial and interglacial periods of the Quaternary, approximately 1.8 million years before present. Evidence of the fluctuations is found in the preserved deposits of this time, particularly the oceanic deposits, which received rock fragments ranging in size from flour to the size of a house carried by ice as it ground its way in glaciers across the barren landscape.

Periodicity in tectonic uplift has also been linked to major shifts in sediment fluxes. In the Himalayas, for example, there have been four major periods of uplift over the past 12 million years, each linked with a higher rate of sediment accumulation in adjacent basin areas (Figure 6). Local changes in surface elevation, as a result of tectonic activity, will impact on both where a river flows and how fast it flows. Higher flow speeds associated with steeper slopes will allow a river to pick up more sediment from its bed, thereby causing it to cut down and increasing fluxes. When slopes get less steep, less sediment can be carried by lower energy flows and material accumulates in the river valleys.

At shorter time-scales, sediment fluxes change in response to fluctuations in water discharge (Figure 7), the majority of the material being carried during flood periods when the rivers are more energetic. Any changes which impact on the delivery of water to the river system, for example, changes in rainfall patterns surface vegetation and land use, can result in shifts in flood frequency and impact on sediment fluxes. One example is in the American midwest during the late nineteenth and early twentieth centuries where the removal of vegetation, as a result of intensifying agriculture, led to floods becoming more intense (*see* **Sedimentary Processes:** Catastrophic Floods). This, combined with the removal of the protection from erosion offered by plants, led to rapid erosion and the development of 'badlands'.

The Importance of Geology

The rock types within the river system will control the rate at which sediment can be produced and supplied to the river system. This can lead to the development of very different environments in otherwise identical basins. For example, in the African rift, lakes in areas where the rocks are mainly old, hard, and resistant to weathering are deep and sediment starved (e.g., Lake Tanganyika), whereas those in areas containing volcanic rocks and old sediments are sand- and mud-rich and generally more shallow (e.g., Lake Baringo). The same factors are important in controlling the way in which sediments filled up ancient basins. For

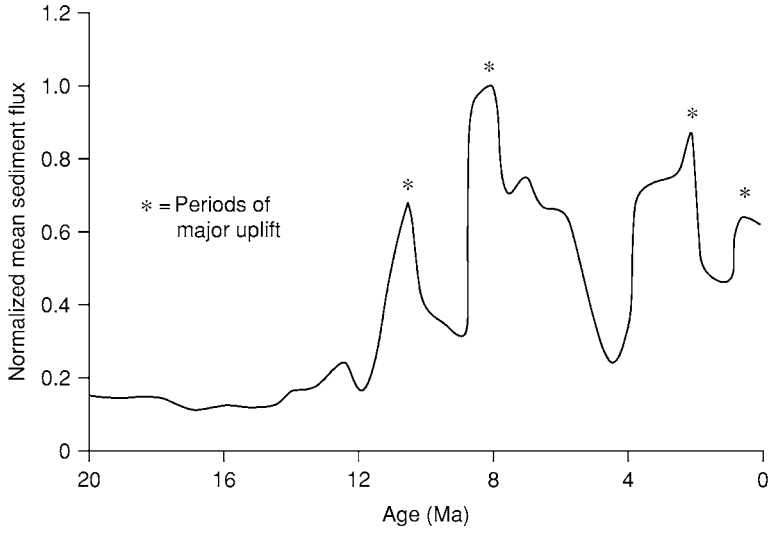


Figure 6 Variations in sediment fluxes from the Himalayas. High fluxes relate to periods of active uplift. (Adapted from Hovan SA and Rea DK (1992) The Cenozoic record of continental mineral deposition on Broken and Ninetyeast ridges, Indian Ocean: southern African aridity and sediment delivery from the Himalayas. *Paleoceanography* 7: 833 860.)

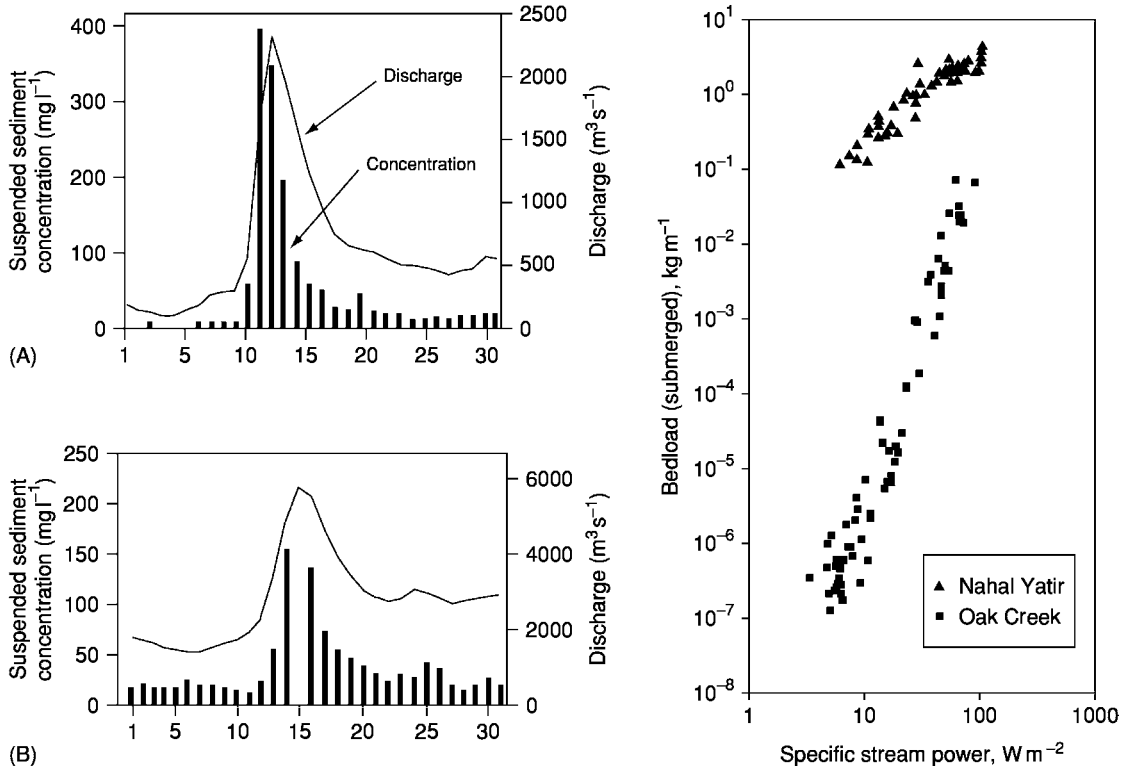


Figure 7 Variations in suspended sediment transport for the Meuse and the Rhine Rivers (left) showing how sediment flux varies over flood events. (Adapted from Asselman NEM and Middlekoop H (1998) Temporal variability of contemporary floodplain sedimen tation in the Rhine Meuse delta, The Netherlands. *Earth Surface Processes and Landforms* 23: 595 609.) and bed material transport for Nahal Yatir, Israel and Oak Creek, USA (right) showing the difference between sediment fluxes in ephemeral (Nahal Yatir) and perennial (Oak Creek) rivers. (Adapted from Laronne JB and Reid I (1993) Very high rates of bedload sediment transport by ephemeral desert rivers. *Nature* 366: 148 150.) plotted against water discharge over flood periods. Fluxes generally increase with increasing water discharge, reaching maxima at flood peaks.

example, the Cretaceous to Tertiary Tucano and Reconcavo basins of Brazil are similar in size and structure but whereas the Reconcavo has a low sediment flux and is oil retaining the Tucano Basin contains interconnected sands that allow any oil generated to escape (*see* **Petroleum Geology:** The Petroleum System). When a geological map of the area is consulted, it is evident that the Reconcavo Basin was formed on ancient and hard gneisses which could only be broken down very slowly, whereas the Tucano Basin was cut into soft sedimentary rocks. This example highlights the economic importance of understanding the controls on sediment flux.

Sediment Budgets: Modelling the Past and Predicting the Future

The complexity of controls on sediment fluxes conspires with changes in rates that occur at all timescales to make the interpretation of the past and predictions of the future equally problematic. Added to this, the role of river drainage basin in storing and releasing sediment, thereby modulating the impact of environmental change on sediment flux, is still poorly understood. Even the direct measurement of present-day fluxes is fraught with difficulties since the costly nature of making such measurements usually means that data are collected only from a small number of points and for very short periods of time. As rivers are the main means of sediment delivery to basins, many of the flux calculations that exist are based on short periods of measuring water and sediment discharges which are then used to extrapolate to larger areas and longer periods, ignoring the limitations of the data. This makes the calculation of sediment budgets for any area very difficult; even under ideal circumstances with good data they generally only give a crude approximation of what is really happening.

Estimations of sediment fluxes and the understanding of sediment budgets are central to the science of

basin analysis. Basin analysts endeavour to determine the controls on where and how sediments are deposited in a basin and they provide essential information to those wishing to exploit natural mineral resources, for example, oil, gas, and diamonds. In the case of oil and gas, it is important to locate sands and other permeable and porous rocks that might act as oil and gas reservoirs (*see* **Petroleum Geology:** The Petroleum System). In addition, identifying parts of the basin and periods of time when sediment fluxes are low and the organic debris from which oil is formed can accumulate undiluted by other sediments, is a vital part of exploration. For placer deposits such as diamonds, rivers are the main transport system and they deliver these important heavy minerals to the basin for reworking into economic beach deposits, e.g., in South Africa diamonds are delivered by the Orange River to the coast of Namibia.

See Also

Geomorphology. **Petroleum Geology:** The Petroleum System. **Sedimentary Environments:** Depositional Systems and Facies; Deserts; Shoreline and Shoreface Deposits. **Sedimentary Processes:** Aeolian Processes; Catastrophic Floods; Deep Water Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification. **Weathering**.

Further Reading

Allen PA and Allen JR (1990) *Basin Analysis Principles and Applications*. Oxford: Blackwells.

Burbank D, Leland J, Fielding E, *et al.* (1996) Bedrock incision, rock uplift and threshold hill slopes in the northwestern Himalayas. *Nature* 379: 505 510.

Descroix L and Mathys N (2003) Processes, spatiotemporal factors and measurements of current erosion in the French southern Alps: a review. *Earth Surface Processes and Landforms* 28: 993 1011.

Hovius N (1996) Regular spacing of drainage outlets from linear mountain belts. *Basin Research* 8: 29 44.

Hovius N and Leeder MR (1998) Clastic sediment supply to basins. *Basin Research* 10: 1 5.

Jansson MB (1988) A global survey of sediment yields. *Geografiska Annaler* 70: 81 98.

Jones SJ and Frostick LE (eds.) (2002) Sediment Flux to Basins: Causes, controls and consequences. *Geological Society Special Publication* 191: 284.

Macklin MG (1999) Holocene river environments in prehistoric Britain: human interaction and impact. *Journal of Quaternary Science* 14: 521 530.

Middleton NJ and Goudie AS (2001) Saharan dust: sources and trajectories. *Transactions of the Institute of British Geographers* 26: 165 181.

Milliman JD and Meade RH (1983) Worldwide delivery of river sediments to the ocean. *Journal of Geology* 91: 1 21.

Milliman JD and Syvitski JPM (1992) Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *Journal of Geology* 100: 525 544.

Shanley KW and McCabe PJ (eds.) (1998) Relative role of Eustacy, climate and tectonics in continental rocks. *Society of Economic Palaeontologists and Mineralogists Special Publication,* 59.

Studies in Geophysics (1994) *Material Flux on the Surface of the Earth*. Washington, USA: National Academy Press.

Syvitski JP, Morehead MD, and Nicholson M (1998) HYDROTREND: a climatically driven hydrologic transport model for predicting discharge and transport load to lakes and oceans. *Computers and Geosciences* 24: 51 68.

Tipper JC (2000) Patterns of stratigraphic cyclicity. *Journal of Sedimentary Research* 70: 1262 1279.

Tucker GE and Slingerland R (1996) Predicting sediment flux from fold and thrust belts. *Basin Research* 8: 329 350.

SEDIMENTARY ROCKS

Contents

Mineralogy and Classification

R C Selley, Imperial College London, London, UK

Introduction

Rocks and Their Classification defined the three main classes of rock: igneous, metamorphic, and sedimentary. It described the main features by which the three types of rock may be distinguished, and presented anomalous examples of each. The classification of igneous and metamorphic rocks is described in **Igneous Processes** and **Metamorphic Rocks:** Classification, Nomenclature and Formation, respectively. This article describes the mineralogy and classification of sedimentary rocks.

Sedimentary rocks are formed from the detritus of pre-existing rocks: igneous, metamorphic, or sedimentary. The way in which rock is weathered, eroded, transported, and deposited is discussed in detail elsewhere (*see* **Weathering, Sedimentary Processes:** Fluxes and Budgets, **Unidirectional Aqueous Flow**, and **Sedimentary Environments:** Depositional Systems and Facies). Sediments possess a wide range of particle size, ranging from boulders to clay, and of chemical composition, including silica, lime, or ferromagnesian volcanic detritus. These parameters of particle size and composition are used to classify sedimentary rocks. Sedimentary rocks commonly exhibit two properties that may be used to differentiate them from igneous and metamorphic rocks.

1. Where they crop out at the surface of the Earth, sedimentary rocks generally show stratification (layering). The strata indicate successive episodes of deposition. Layering is usually absent from igneous rocks, but is found in some metamorphic rocks.
2. When examined under the microscope, sedimentary rocks are generally seen to consist of particles. Void space (porosity) is commonly present between the constituent grains. Interconnected pores give the rock permeability. Permeability allows fluids to migrate through rock, and enables rock and soil to drain. Additionally, fossils are only found in sedimentary rocks, some of which are, indeed, made up of nothing else.

Mineralogical Basis for Sedimentary Rock Classification

In the earliest classifications of rock, such as that proposed by Charles Lyell (*see* **Famous Geologists:** Lyell), four classes were recognized: volcanic, plutonic (these two are now grouped as igneous), metamorphic, and aqueous. The aqueous rocks were subdivided into three groups: arenaceous, argillaceous, and calcareous. The term 'aqueous' has long

been abandoned, as it is now known that sediments are deposited by aeolian, gravitational, and glacial processes, as well as by purely aqueous ones.

The processes of weathering, erosion, transportation, and deposition are nature's way of chemically fractionating the Earth's surface, and lead to a logical classification of sedimentary rocks based largely on their chemistry and mineralogy. Serendipitously, this fractionation correlates broadly with their mode of formation. Although geologists broadly agree about the definition of the main classes of sedimentary rocks, there is no unanimity about all of them.

In 1937, Goldschmidt proposed a classification based on five chemical groups, namely: (1) resistates, (2) hydrolysates, (3) oxidates, (4) carbonate precipitates, and (5) evaporites. In 1950, Rankama and Sahama added a sixth: reduzates. The resulting classification highlights the chemical fractionation that results from sedimentary processes, but produces some very strange and uncouth names.

Meanwhile, a practical field-based classification had been widely adopted, although with some variation in the fine detail. It had long been noted that fractionation on the surface of the earth naturally divided rocks into those that had never gone into solution, and might therefore be termed 'allochthonous' or 'detrital', and those formed from minerals that had been dissolved in surface water, and had precipitated out. These are termed 'autochthonous' or 'chemical' rocks. The allochthonous or detrital sediments are subdivided by grain size. The autochthonous or chemical sediments are subdivided by mineralogical (chemical) composition (Table 1).

Geopedants will already notice the inconsistency of this classification. The detrital sediments are composed of a wide range of minerals, and thus exhibit a diversity of chemistry, which is ignored for the purposes of their classification. Similarly, the chemical rocks may occur in a wide range of particle size, from boulders of limestone to sapropelic muds. Nonetheless, the classification displayed in Table 1 is not for the benefit of geopedants, but for practical use by geologists. There is no consensus on the classification of sedimentary rocks proposed in Table 1. Note that the caption reads 'A classification of sedimentary rocks' not 'The classification of sedimentary rocks'.

The main groups of sedimentary rocks are now described briefly, pointing the way to articles in this encyclopedia that describe them in more detail.

Table 1 A classification of sedimentary rocks

Allochthonous or detrital sediments
Classified by grain size
Gravel/conglomerate
Sand/sandstone
Silt/siltstone
Clay/claystone (sometimes also termed 'mudrocks' or 'shales')
Autochthonous or chemical sediments
Classified by mineralogy
Carbonates (limestone and dolomite)
Evaporites (gypsum/anhydrite, halite, etc.)
Residual (bauxite, laterite, kaolinite)
Kerogenous (peat, lignite, coal)
Ironstones (haematitic, chamositic, and sideritic)
Phosphates (guano)
Siliceous (chert, opal)

Allochthonous or Detrital Sediments

As defined earlier, the allochthonous or detrital sediments are the insoluble residue of weathering of pre-existing rocks: igneous, metamorphic, or sedimentary. The mineralogy is very varied, depending on the source material and the type and duration of the weathering process (*see* **Weathering**). The mineralogy also correlates crudely with the grain size. Conglomerates tend to be polymineralic, sandstones are dominated by quartz, and mudrocks are dominated by clay minerals. Table 1 shows the subdivision of the allochthonous or detrital rocks by grain size, gravel, sand, silt, and clay being the basis for conglomerate, sandstone, siltstone, and mudstone, respectively. The terms 'rudaceous', 'arenaceous', and 'argillaceous' have also been applied to conglomerates, sandstones, and shales, but are little used now. The main groups of detrital sediments are now described briefly.

Conglomerate

Conglomerate is composed of particles of gravel, that is to say of particles of greater than 2 mm in diameter, consisting, with increasing size, of granules, pebbles, cobbles, and boulders. Collectively, conglomerates have also been known as rudaceous rocks. Conglomerates are distinguished from breccias by the fact that the clasts are rounded, whilst those of breccias are angular. Because of their large size, conglomerate clasts are composed of many grains or crystals (depending on whether they were derived from earlier sediments or from crystalline igneous or metamorphic rocks). They may thus be composed of a wide range of minerals. When derived from igneous or metamorphic rocks, conglomerates may be composed of the wide range of minerals found in the parent rock. By contrast, conglomerates derived from sediments will reflect their source mineralogy, but will tend to be composed of a higher percentage of minerals that are stable at the Earth's surface, rather than in the parent rock. The concept of sediments as the insoluble residue of pre-existing rocks is again a useful one to recall. This is illustrated

by the vast volumes of flint (chert) gravels to be found on the beaches of north-west Europe. These flints originate in rare horizons in the Cretaceous Chalk. As Charles Lyell wrote: "The entire mass of stratified deposits in the Earth's crust is at once the monument and measure of the denudation which has taken place." Gravels and conglomerates are described in greater detail in **Sedimentary Rocks:** Rudaceous Rocks.

Sandstones

Sandstones are composed of particles with an average size of between 2.00 and 0.0625 mm in diameter. They have four constituents: grains, matrix, cement, and, sometimes, porosity (Figure 1). Sand-sized particles form the framework of the rock. Matrix, the finer grained material that may infill space between the framework grains, was deposited at the same time as the framework grains. Cement is the term that describes minerals precipitated in pores after the deposition of the sediment. Thus, matrix is syndepositional and cement is postdepositional. Cement and porosity are described in **Sedimentary Rocks:** Sandstones, Diagenesis and Porosity Evolution, which deals with the diagenesis of petroleum reservoirs. Framework grains and matrix are described below.

The framework grains of sandstones are normally composed of varying amounts of the mineral quartz (silica, SiO_2). In order of decreasing abundance, sandstones also contain feldspar (a suite of calcium, potassium, and sodium silicates), micas (sheet silicates, with varying amounts of iron, magnesium, and aluminium), a complex of ferromagnesian minerals, informally termed 'mafics', and heavy minerals (those with densities significantly greater than that of quartz (2.65 g/cc), examples of which include iron ores, mica, siderite, zircon, and apatite). Sandstone may also contain sand-sized grains composed of more than one mineral or crystal. These are termed 'rock fragments' or 'lithic grains'. Sand-sized rounded green grains of the complex mineral glauconite are a common constituent of shallow marine sands (glauconite is described in **Minerals:** Glauconites). Sandstones often contain fossil fragments. Teeth, fish scales, and bone are largely phosphatic. The most common fossils, however, are shells, mainly composed of lime, calcium carbonate ($CaCO_3$). With increasing lime content, sandstones grade into calcareous sandstones, then to sandy limestones, and finally to pure limestone, composed entirely of calcium carbonate, and with negligible quartz. Thus, although typically composed of quartz, sandstones also contain a range of other minerals. These are used as a basis for naming and classifying sands and are important because of their impact on geophysical well-log interpretation.

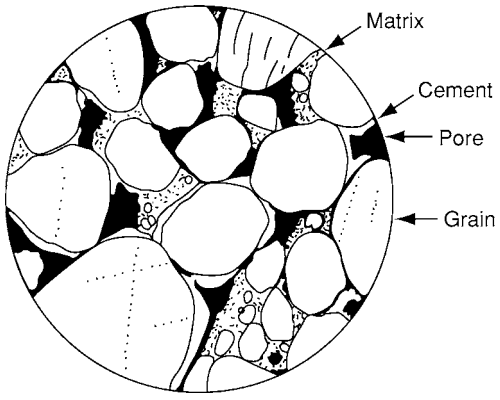


Figure 1 Diagram of a thin section of sandstone showing the four components of framework grains, matrix, cement, and pores. Only the first of these is always present. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

The syndepositional matrix that may occupy some of the space between the framework sand grains consists of silt, clay, and organic matter. Heavy mineral grains are commonly silt sized, and so technically they may form part of the matrix.

The composition of typical sandstone may be as follows. Framework grains: quartz, 45%; rock fragments, 5%; feldspar, 10%; mafics, 5%; mica, 5%; heavy minerals, 2%. Matrix: clay, 7%. Cement: calcite, 5%. Porosity, 16%. Total: 100%.

Two parameters are used to name and classify sandstones: chemical mineralogy and physical texture. When sediment is first eroded from its parent outcrop, it is generally immature in both its composition and texture. That is to say, it will still contain a range of chemically unstable mineral grains that surface processes have yet to break down and dissolve. Similarly, the debris first transported down a hillside will be very poorly sorted, consisting of a range of particles, varying in size from boulders to clay. When looking at an ancient lithified sandstone, its maturity may be described in terms of its chemical and physical properties (mineralogy and texture). Four main types of sandstone may thus be recognized as shown in Table 2. This table also employs four commonly used names to describe sandstones.

Table 2 A classification of sandstones based on textural and mineralogical maturity

	Mineralogical maturity	
	Immature	*Mature*
Texturally immature	Greywacke **Figure 2**	Quartz wacke **Figure 3**
Texturally mature	Arkose **Figure 4**	Quartzite **Figure 5**

'Greywackes' are poorly sorted sandstones with a large component of chemically unstable grains, not only feldspars, but also rock fragments and ferromagnesian minerals (Figure 2). 'Quartz wackes' are also texturally immature, but the framework grains are composed largely of quartz and lithic (rock) fragments (Figure 3). 'Arkoses' are texturally mature, but contain a large percentage of chemically unstable grains, principally feldspar (Figure 4). Quartzites, also termed quartz arenites (from the Greek '*arenos*' for sand), are texturally and mineralogically mature, being well sorted, and composed of little but quartz (Figure 5).

Mudrocks

There is little unanimity over the terminology for the argillaceous detrital sedimentary rocks. The easiest to name objectively are siltstone and claystone, being

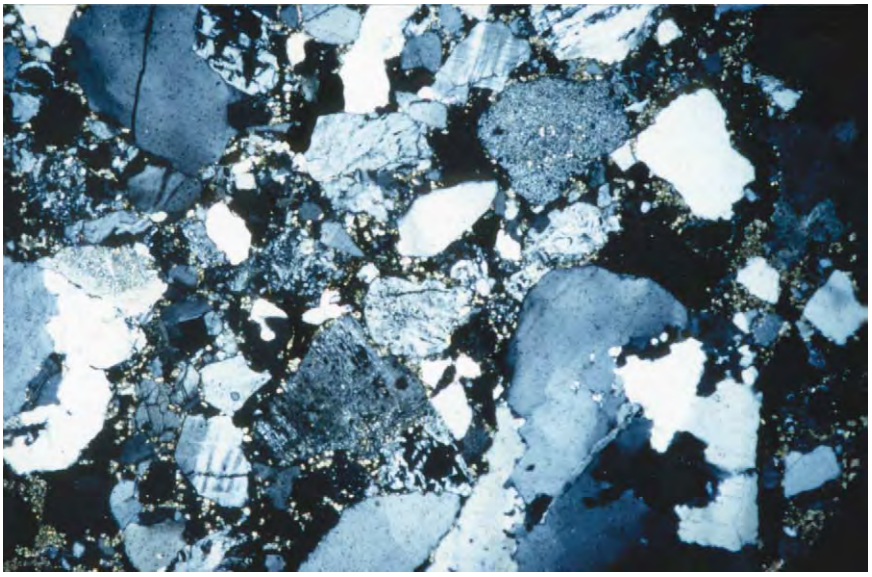


Figure 2 Photomicrograph of a greywacke under polarized light. Jurassic, UK North Sea. Note the poorly sorted texture and abundance of matrix and twinned feldspar. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

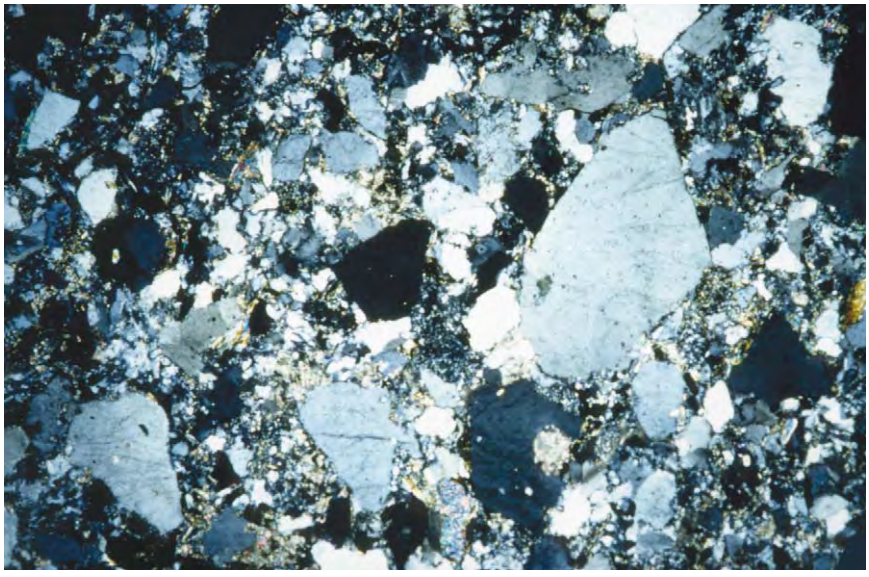


Figure 3 Photomicrograph of quartz wacke under polarized light. Carboniferous, Chios, Greece. Note the poorly sorted texture and abundance of matrix. The framework grains are almost entirely composed of quartz and chert. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

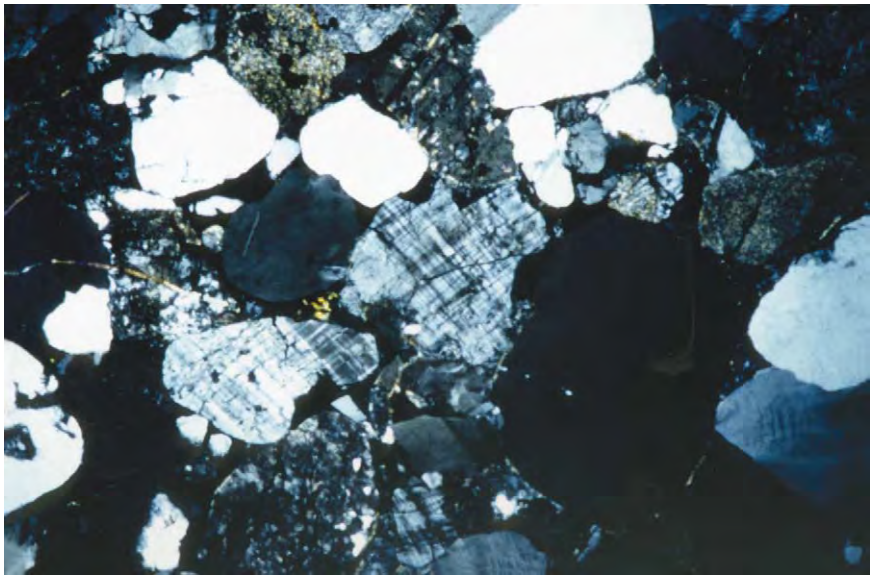


Figure 4 Photomicrograph of arkose under polarized light. Torridonian, Precambrian, Scotland. Note the abundance of twinned feldspar and the better sorted texture than in **Figures 2** and **3**. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

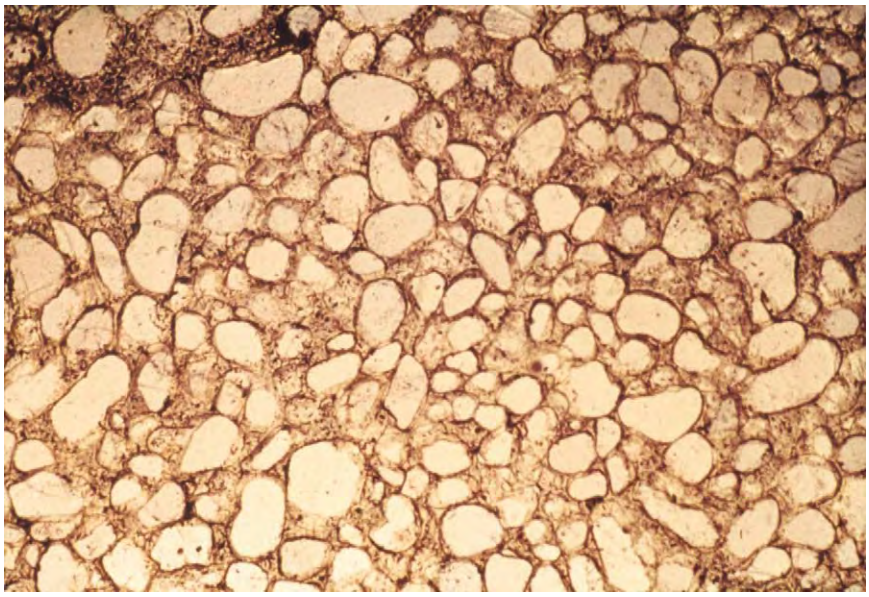


Figure 5 Photomicrograph of a quartz arenite under ordinary light. Simpson Group, Ordovician, Oklahoma, USA. Note the well sorted texture. The framework grains are almost entirely composed of well rounded and well sorted quartz. There is neither matrix nor cement. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

composed, by definition, of predominantly silt or clay, respectively. They are easy to identify in the field. Siltstones have an unpleasant gritty feel between the teeth, clays a pleasing plastic texture. When sufficiently indurated, both siltstones and claystones may be termed 'shales'. The term 'shale' was defined by Pettijohn as "a laminated or fissile claystone or siltstone". This is a widely used term, and with good reason: siltstones are commonly argillaceous and claystones silty, and so the term shale covers all variations. The term 'mudrock' or 'mudstone' has been applied to siltstones and claystones (and their admixtures) that do not possess the fissility of shale. Siltstones are composed of detrital quartz, shell fragments, assorted heavy minerals, and, often, mica, which imparts the fissility to shale. Siltstones

also commonly contain varying amounts of clay matrix and organic matter.

Claystones are composed largely of the clay minerals, kaolin, illite, montmorillonite, and chlorite. Clay mineralogy is described in detail in **Clay Minerals**. Lime and organic matter may also be present in claystones. With increasing lime content, shales grade into marls, argillaceous limestones, and limestones. With increasing organic content, shales grade into sapropelite. Detrital grains of silt, mica, plant debris, and shell fragments may occur in mudstones as impurities. Clays and their diagenesis are described in **Sedimentary Rocks:** Clays and Their Diagenesis.

Autochthonous or Chemical Sediments

It has already been noted that the classification of sediments into detrital and chemical categories is somewhat artificial. Both are composed of chemical or biochemical components, and of particles of varying sizes. Nonetheless, it is a convenient grouping. The chemical sediments are those that principally precipitate out of solution, although thereafter they may become detrital in some instances. Table 1 shows that seven types may be recognized: carbonates, evaporites, residual deposits, kerogen, ironstone, phosphate, and silica. These are now described briefly.

Carbonates

The carbonate chemical sediments include a wide range of rocks, of which limestone and then dolomite are volumetrically the most important, whilst siderite and magnesite, although rare, are economically important. The mineralogy of carbonates is described in detail in **Minerals:** Carbonates. Limestone is composed largely of the mineral calcium carbonate ($CaCO_3$). Limestones may be made up of many different types of grain that originate in different ways in a range of depositional environments (Figure 6). Thus grain type is one of the keys to interpreting their depositional environment. Limestones form, almost without exception, from the aqueous precipitation of calcium carbonate, aided by some organic process or other, most obviously as shells secreted by invertebrates, but also as nodules, laminae, and clouds whose origins owe much to biochemical reactions. Limestones are described in greater detail in **Sedimentary Rocks:** Limestones.

Dolomite is composed of the mineral dolomite, $CaMg(CO_3)_2$, named by and from the eponymous French Count Dolomieu (1750–1801). Geopedants restrict the term dolomite to the mineral, and dolostone to the rock. This was not the Count's original intent. Like limestone, dolomite forms in several different ways, from penecontemporaneous cryptocrystalline mudstones in sabkhas, to coarsely crystalline varieties during late diagenesis. In the latter case, dolomite is virtually a metamorphic rock (Figure 7). Dolomite is an important petroleum reservoir, and occurs as a gangue mineral with lead–zinc sulphide ores (*see* **Mining Geology:** Hydrothermal Ores). It is described more fully in **Sedimentary Rocks:** Dolomites.

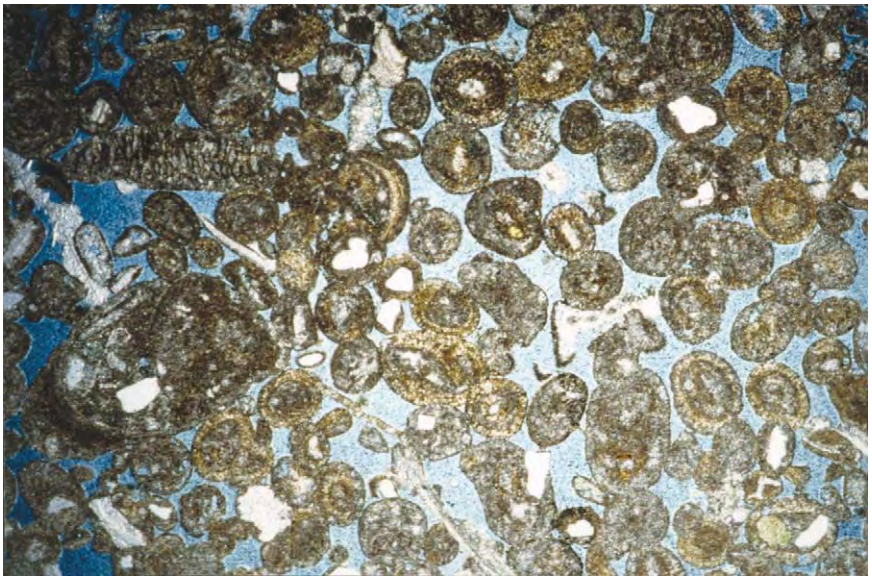


Figure 6 Photomicrograph of limestone under ordinary light. This is a well sorted oolite grainstone from the Upper Jurassic Portland Limestone, Dorset, UK. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

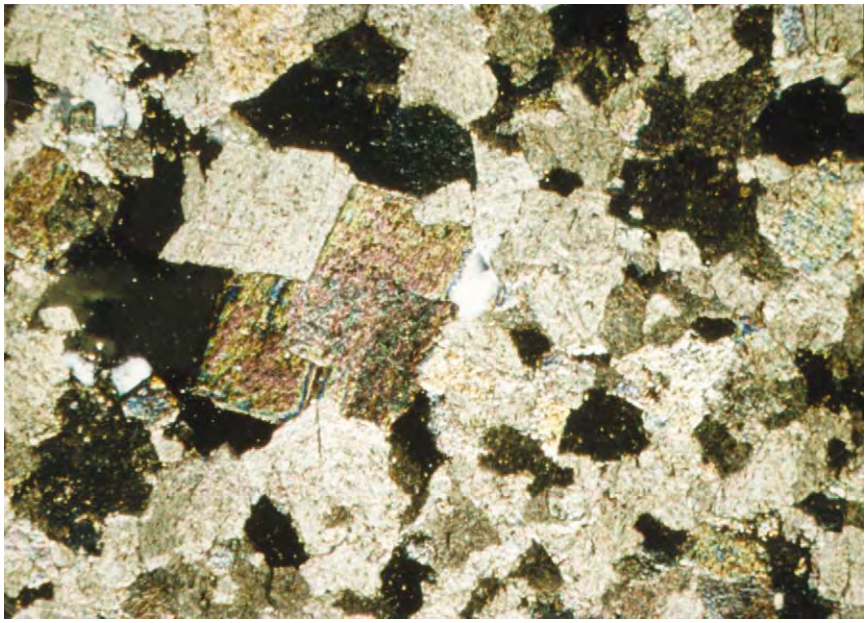


Figure 7 Photomicrograph of dolomite under ordinary light. This is a coarsely crystalline variety from the Zechstein (Upper Permian) of the UK North Sea. Some porosity (pale blue) is visible. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Siderite is composed of iron carbonate ($FeCO_3$). It occurs commonly in shales as early cement and as concretions. It occurs as crystalline cement in sandstones, and occasionally in spherulites (sphaerosiderite) in lacustrine deposits. Here, it may be sufficiently abundant to become an iron ore (described below).

Magnesite ($MgCO_3$) is the name for the mineral magnesium carbonate, as well as the rock. It forms both as an alteration product of dolomite, and from the action of magnesium-rich fluids on limestone. Magnesite is rare, but occurs in commercially important deposits at Radenhein (Austria), Liaotung (China), and Clark County, Nevada (USA).

Evaporites

The evaporite chemical sedimentary rocks are rare, but extremely important commercially as the raw materials for the chemical industry. As the name suggests, the evaporites consist of a suite of minerals formed from the evaporation of sea water. They tend to occur in restricted sedimentary basins in cyclic sequences that begin with carbonates (limestone and/or dolomite), overlain by sulphates (gypsum and/or anhydrite), halite (sodium chloride), and then a range of potassium salts, including carnallite and polyhalite (Figure 8).

As the name suggests, it was once thought that evaporites formed exclusively from the drying out of enclosed marine basins. This required improbably large volumes of sea water to provide the resultant evaporites. It is now realized that many evaporites actually form in sabkhas (Arabic for salt marsh) from the replacement of pre-existing rocks, principally carbonates, by circulating brines. Evaporites should thus more correctly be termed 'replacementites'. Evaporites are described in more detail in **Sedimentary Rocks:** Evaporites.

Residual Deposits

Residual deposits are a variety of rocks produced by *in situ* chemical alteration or weathering (*see* **Weathering**). They include three economically important rocks: laterite, china clay (kaolin), and bauxite. These are now described in turn.

The word 'laterite' is derived from the Latin '*later*', a brick, as this rock has been widely employed for this purpose, being soft when quarried, but hardening on exposure. The term was first employed by a British geologist of the Raj, working in India, where laterites are exceptionally well developed. Laterites result from the intense weathering in many parts of the world of rocks of diverse ages and types, but particularly iron-rich rocks such as basalts. Laterites thus occur as laterally extensive residues up to 10 m in thickness above the bedrock. They require thousands of years to form, a humid climate, and a well-drained terrain. The resultant laterite is rich in hydrated iron and aluminium oxides, and low in humus, silica, lime, silicate clays, and most other minerals. Laterites are red and argillaceous in appearance, but often possess

Figure 8 Photomicrograph of nodular anhydrite ($CaSO_4$) from the sabkha of Abu Dhabi, United Arab Emirates. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

a nodular pisolitic texture. They are often termed ferricrete (etymologically, although not genetically, comparable with silcretes and calcretes). Laterites are important sources of iron in West Africa and Western Australia, and of nickel in Cuba. Further details are given in **Soils:** Modern and **Soils:** Palaeosols.

China clay, or kaolin, is the name given to rock composed almost entirely of the clay mineral kaolinite, $Al_2Si_2O_5.(OH)_4$ (*see* **Sedimentary Rocks:** Clays and Their Diagenesis). Kaolin and kaolinite are occidental corruptions of Kauling, a hill in China, from whence the first samples to enter Europe were shipped by a Jesuit missionary in 1700. Some kaolinite is produced by the *in situ* hydrothermal alteration of feldspar in granites, as for example that of south-west England. Kaolinite may then be reworked from such a source, and re-deposited in lacustrine environments, as for example the Oligocene 'Ball Clay' deposits of Bovey Tracey, Devon.

Kaolin also forms, however, as a residual deposit due to the intense weathering of aluminosilicate-rich rocks. These include feldspar-bearing igneous rocks, such as granites and gneisses. Kaolin can also be produced from sedimentary rocks, including arkosic sandstones and shales. The general chemical reaction leading to the production of kaolinite is:

$$KalSi_3O_8 + H_2O \rightarrow Al_2Si_2O_5(OH)_4$$
$$\text{Feldspar} + \text{Water} \rightarrow \text{Kaolinite}$$

Kaolin forms as a residual deposit in the soil horizons of warm humid climates, where erosion rates are low, and there is plenty of time for leaching to take place. Kaolin has many important industrial uses (see **Clays, Economic Uses**). Notable commercial deposits occur in China, naturally, south-west England, Saxony (Germany), Bohemia (Czech Republic), and Georgia, USA.

The third type of residual deposit is bauxite, hydrated aluminium hydroxide ($Al_2O_4.nH_2O$). Bauxite takes its name from Le Baux, near Arles in France. Bauxite is the end result of the intensive and prolonged weathering of soils that commences with laterite, and proceeds, via kaolin, to bauxite. These changes reflect the progressive leaching of silica, iron, and kaolinite (Figure 9).

The chemical reaction that finally leads to the formation of bauxite is:

$$H_2O + Al_2Si_2O_5(OH)_4 = Al_2O_4.nH_2O + 2SiO_2.2H_2O$$

$$\text{Water} + \text{Kaolinite} = \text{Aluminium Hydroxide} + \text{Silicic Acid}$$

Bauxites tend to be reddish or pink in colour due to some residual iron oxide. They may also possess a pisolitic texture inherited from an earlier lateritic phase (Figure 10).

Bauxite is very important as the ore for aluminium. Bauxite occurs as residual deposits on limestone, as for example in France and Jamaica. It also occurs as a residual soil on Precambrian igneous and metamorphic rocks, as in Surinam.

Kerogenous Chemical Sediments

Kerogen is defined as hydrocarbons that are insoluble in normal solvents, such as carbon tetrachloride, but which yield liquid or gaseous petroleum when heated. Chemically, kerogen includes a range of complex hydrocarbons, with traces of many other elements, including sulphur, nitrogen, and various radioactive and heavy metals. Kerogen is generally deposited in anoxic reducing stagnant conditions, most commonly found in marshes, swamps, meres, salt marshes, and lagoons, and is particularly characteristic of deltas (*see* **Sedimentary Environments:** Deltas). In these environments, vegetation may accumulate as laterally extensive horizons of peat many metres thick. During subsequent burial, peat undergoes extensive compaction and diagenesis, changing first into brown coal (lignite), then bituminous coal, then anthracite, and finally graphite, as it enters the realm of metamorphism. The variety of kerogen termed coal is, of course, a very important source of energy. Less conspicuously, although equally important, kerogen occurs in varying amounts in mudstones. Originating as plant and animal detritus, this disseminated kerogen is the mother of petroleum (Figure 11). When kerogen constitutes >1.5%, or thereabouts, of a shale, the shale becomes a potential petroleum source rock, subject to sufficient thermal maturation. Depending both on its chemistry and the level of maturation, kerogen generates petroleum gas and oil (*see* **Petroleum Geology:** The Petroleum System).

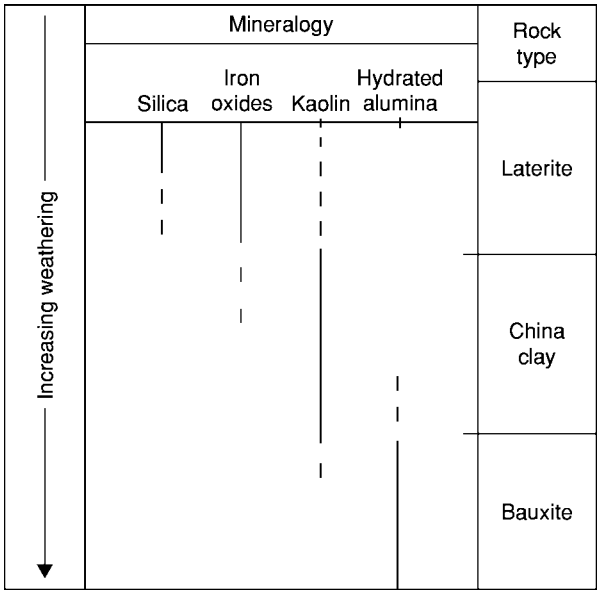


Figure 9 Diagram to show the progressive formation of the residual deposits laterite, kaolin, and bauxite that result from prolonged chemical weathering. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Ironstones

Iron is widespread in sedimentary rocks, but is concentrated in economic amounts in very few. A distinction is made between the Banded Ironstone Formations (colloquially referred to as BIFs), and ironstones *sensu stricto*. Banded ironstones are widespread around the Earth, but they are all of Precambrian age, and are curiously interbedded with

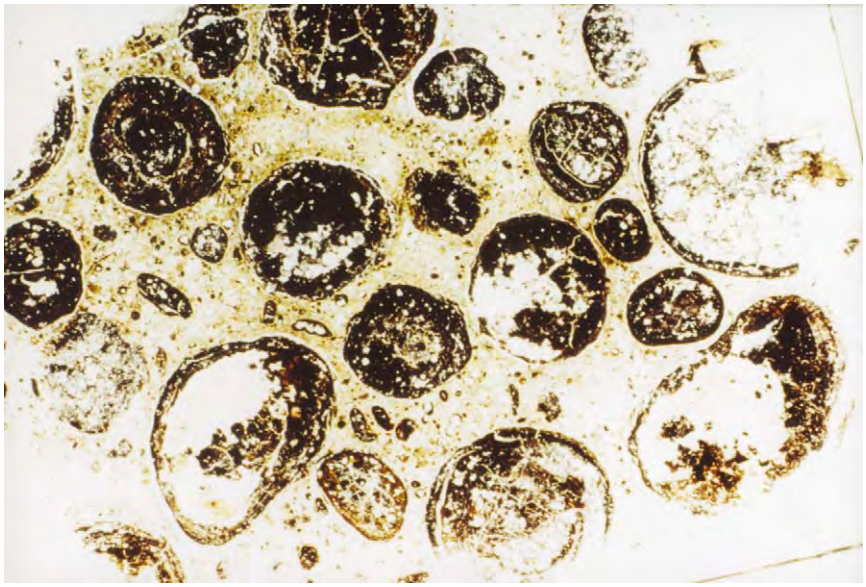


Figure 10 Photomicrograph of pisolitic bauxite under ordinary light. Individual pisoids are approximately 0.5 cm in diameter. Le Baux, France. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

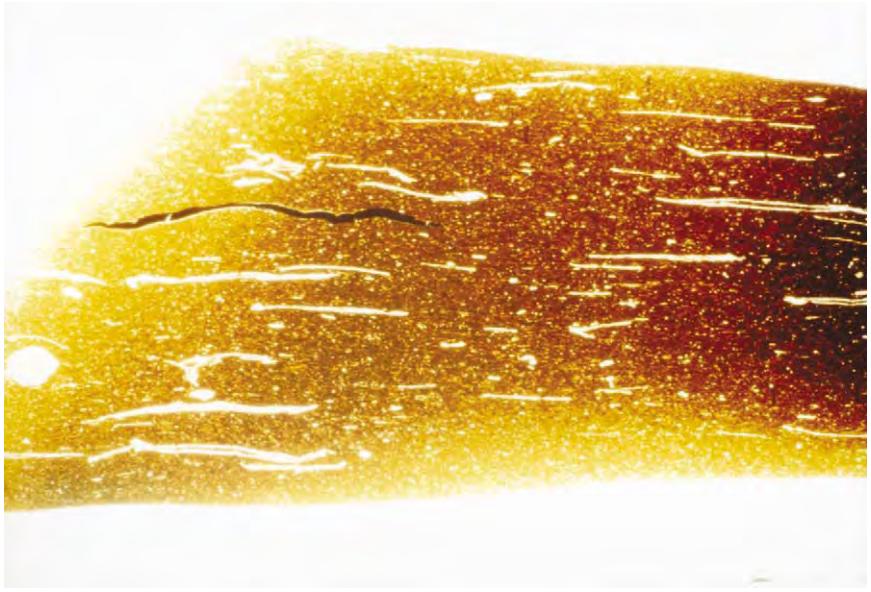


Figure 11 Photomicrograph of kerogen. This is the sapropelic Kimmeridge Coal (Upper Jurassic) from Dorset, UK. Cross sections of bivalves are ubiquitous, and carbonized plant detritus is also visible. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

chert. They formed when the Earth's atmosphere was significantly different from that of today (*see* **Sedimentary Rocks:** Banded Iron Formations). The term ironstone is now restricted to Phanerozoic sedimentary rocks consisting of at least 15% by weight of iron, either 19% FeO, or 21% Fe_2O_3, or an equivalent admixture.

Ironstones consist of a range of iron minerals, including oxides (magnetite, haematite, and goethite/limonite), carbonates (siderite), and silicates (chamosite and berthierine). Three main types of ironstone are recognized: blackbands, claystones, and ooidal. Blackband and claystone ironstones are organic-rich sideritic mudstones. They commonly occur in deltaic deposits associated with peat and coal. Intraformational conglomerates composed of ironstone clasts and horizons of brittle fractured ironstone in slumped delta slope shales indicate that the ironstones formed during early shallow burial.

The origin of the third type of ironstone is more controversial. The ooidal ironstones are composed of several types of iron mineral (Figure 12). Ooid formation is normally associated with high-energy depositional environments (*see* **Sedimentary Rocks:** Limestones). The ooidal ironstones, however, are often poorly sorted wackes. Thus, argument has raged as to whether ferruginous ooids formed in high-energy environments, and were then dumped as poorly sorted wackes. Alternatively, did iron minerals replace quotidian lime ooids during subsequent diagenesis?

The data and arguments are examined more fully in **Sedimentary Rocks:** Ironstones, but it is probably as true today as it was in 1949, when Taylor wrote in his seminal memoir on the Northampton Sand Ironstone, that: "Conditions of deposition of the sedimentary iron ores are still to some extent a mystery".

Phosphates

The penultimate group of chemical sediments to consider are the phosphates. Phosphates are an extremely important mineralogically complex group of rocks that are essential as plant fertilizers. About three-quarters of the world's supply of phosphates comes from sedimentary deposits. Sedimentary phosphate deposits are of three types: bedded, placer, and guano. Bedded phosphates are formed by the replacement of limestone, and of teeth, bones, and coprolites, to form the mineral phosphorite. Factors that favour phophoritization include a broad shelf adjacent to an ocean, slow shallow marine sedimentation, low terrigenous input, and high organic productivity. Once phosphorite has formed, it is stable in sea water, and sparingly soluble in fresh water. Thus, phosphate pebbles occur, not only as intraformational conglomerates intimately associated with bedded phosphate, but also as placer deposits that are sometimes far removed from their parent body. Bedded phosphates formed during several geological periods on ancient shelves around the world. The Cretaceous phosphate belt of the southern shores of Tethys is noteworthy. This stretches

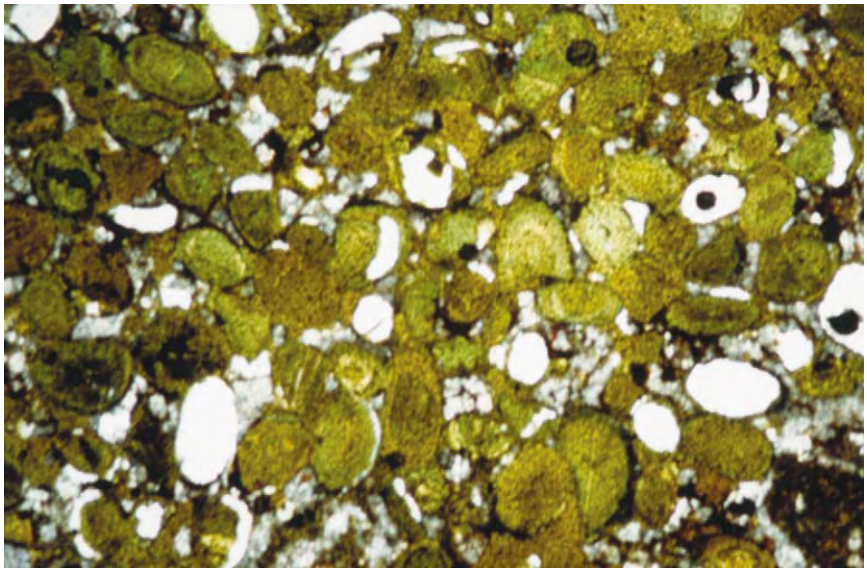


Figure 12 Photomicrograph, under ordinary light, of Northampton Sand Ironstone, Middle Jurassic, UK. This ironstone is composed of chamosite and siderite ooids with concentric growth rings. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

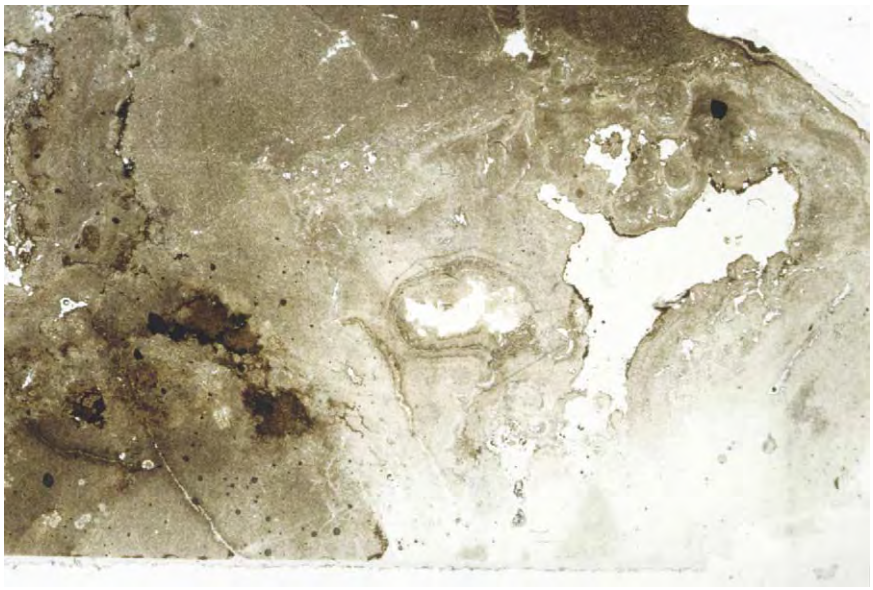


Figure 13 Photomicrograph, under ordinary light, of guano phosphate deposit. This is formed by the phosphatization of bird droppings. St Helena, South Atlantic. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

from Bu Craa in the western Sahara, through Morocco and Algeria to the flanks of the Red Sea, Jordan, and Jabal ash-Sharki in Syria. Younger placer phosphate deposits of note include the alluvial phosphate gravels of South Carolina and Florida in the USA.

Guano is the youngest phosphate rock (Figure 13). This is a fertilizer rich in phosphates and nitrates that forms from the accreted excreta of birds and bats. Whole Pacific Islands, such as Nauru, are, or rather, were composed of guano that has subsequently been quarried away.

Phosphates are described in greater detail in **Sedimentary Rocks: Phosphates.**

Siliceous Deposits

The last group of chemical sediments to describe are those composed of silica (SiO_2), not the detrital sands, but those that formed by organic secretion, replacement, or, possibly, by direct precipitation from water.

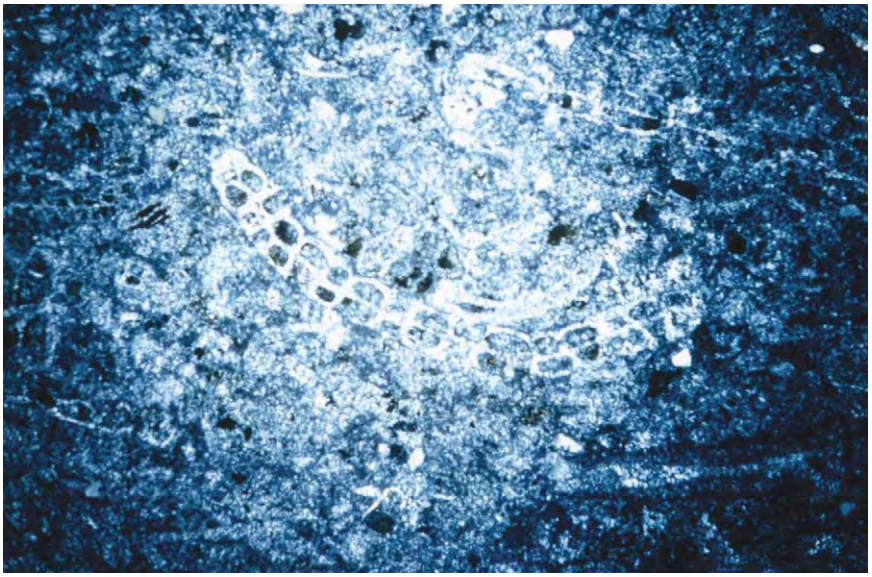


Figure 14 Photomicrograph, under polarized light, of chert (colloquially, flint) from the Chalk (Upper Cretaceous), Dover, UK. This formed by replacement of pre existing limestone. Some silicified bioclasts can be detected. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Various organisms secrete silica. Radiolaria and diatoms secrete silica shells, whilst some sponges secrete internal spicules of silica. Deposits of radiolarian and diatom shells may accumulate under water in sufficient amounts to constitute sedimentary rocks in their own right, termed 'radiolarite' and 'diatomite', respectively. Organically precipitated silica disseminated in sediments may undergo dissolution and reprecipitate as cryptocrystalline silica. This variety of silica is termed 'chert', often referred to colloquially as 'flint' by the aboriginal inhabitants of the Cretaceous Chalk downlands of south-east England. Chert is a non-clastic cryptocrystalline variety of silica that occurs as nodules and horizons in limestones and sandstones (Figure 14). Chalcedony is a radial or fibrous variety of chert that infills fossils and other pores. Opal, sometimes termed opaline silica, is a hydrated variety of silica, amorphous and isotropic, with the chemical formula $SiO_2.nH_2O$, sometimes also written $SiO_2.nSi(OH)_4$. It occurs only in Tertiary to Recent sediments, having dehydrated to chert in older rocks. Opal is a semiprecious stone.

Cherts pose three main problems. What is the source of the silica? What is the environment of deposition of bedded cherts? How do nodular cherts replace their host sediments? These questions are discussed in **Sedimentary Rocks:** Chert.

Conclusion

The primary aim of this overview of the sedimentary rocks has been to simultaneously point the reader in two different directions: to the tranche of articles dealing with mineralogy, and to the tranche of articles that describe the various sedimentary rocks in more detail. The article offers a simple classification of the sedimentary rocks, but at the same time points out the problems and inconsistencies of any scheme of rock classification. Sediments can be broadly grouped either by their physical attributes, principally grain size, but also by their chemistry or mineralogy. A detailed analysis of any scheme shows its inconsistencies. To call a group of rocks 'chemical' is nonsensical, because it implies that others are 'non-chemical', whatever that might mean. Evaporites should really be called 'replacementites,' but who would dare to argue for such a change. Siliceous chemical rocks include detrital sediments, primary precipitates, and also some of diagenetic origin, yet they are placed in one single group. Classification is purely a pedagogical framework, like the chrysalis from which exegesis may emerge like a butterfly.

See Also

Clay Minerals. **Clays, Economic Uses**. **Famous Geologists:** Lyell. **Metamorphic Rocks:** Classification, Nomenclature and Formation. **Minerals:** Carbonates; Glauconites. **Mining Geology:** Hydrothermal Ores. **Petroleum Geology:** The Petroleum System. **Rocks and Their Classification**. **Sedimentary Environments:** Depositional Systems and Facies; Deltas; Storms and Storm Deposits. **Sedimentary Processes:** Fluxes and Budgets. **Sedimentary Rocks:** Banded Iron Formations; Chert;

Clays and Their Diagenesis; Dolomites; Evaporites; Ironstones; Limestones; Phosphates; Rudaceous Rocks; Sandstones, Diagenesis and Porosity Evolution. **Soils:** Modern. **Unidirectional Aqueous Flow**. **Weathering**.

Further Reading

Leeder MR (1999) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Science.

Lyell C (1842) *(and many subsequent editions) Elements of Geology*. London: John Murray.

Pettijohn FJ (1957) *Sedimentary Rocks*, 2nd edn. New York: Harper Geoscience.

Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Taylor JH (1949) *Petrology of the Northampton Sand Iron stone Formation. Memoirs of the Geological Survey*. London: HMSO.

Banded Iron Formations

A Trendall, Curtin University of Technology, Perth, Australia

Introduction

Banded Iron Formation, normally abbreviated to BIF, is a type of sedimentary rock commonly present in Precambrian sedimentary successions of specific ages. Apart from this time restriction it has a number of characteristics which make it unique among sedimentary materials. Firstly, there is nothing chemically similar to it being laid down in the waters of the modern Earth, so that its origin cannot be deduced by directly observing the formation of similar materials in the present-day environment. Secondly, it has virtually none of the textural features common in other sedimentary rocks that give a clue to the conditions under which it was laid down; theories for its origin, therefore, have to be argued from various indirect lines of evidence, and there is still no complete agreement on its significance in the stratigraphic record. A third feature is that it is exceptionally hard, tough, and dense, making it instantly recognisable in field exposures, where it commonly forms resistant ridges standing out from more easily eroded rocks. Finally, it is a sedimentary rock in which fine details of stratification have been shown to persist over exceptionally large distances, arguing for depositional conditions free from disturbance by currents. BIF also has great economic importance because it is the source, either directly or indirectly, of most of the iron ore presently mined; and as the main source of raw material for the world's iron and steel industries, underpins the physical fabric of the developed world.

Nomenclature, Classification, Definition

Before describing the characteristics of BIF in more detail, it is worth reviewing the different names that have been applied to it on different continents and at different times, so that there is a clear understanding of how the name is applied in this article. The first formal geological descriptions of BIF were made in the latter part of the nineteenth century, focusing on occurrences in parts of the USA south of Lake Superior, where iron ore mining was established in areas of outcrop known as 'ranges', such as the Mesabi, Marquette, Cuyuna, Gogebic, and Menominee Ranges. The rock was then called 'jasper', 'jaspilite', or 'iron-bearing formation', which was later shortened to 'iron-formation'. The name 'taconite' was also used, and because many of the rocks had conspicuously multicoloured banding the term 'banded iron-formation' also became common. During the early twentieth century, other names came to be used for similar rocks of other continents. For example, 'itabirite' was used in Brazil, 'ironstone' in South Africa, and 'BHQ' ('banded haematite quartzite') in India. Studies in all these places tended to assume that the iron-rich rocks to which these names were applied were identical with those that had been well documented from the Lake Superior ranges.

As more detailed studies were made later in the twentieth century, it was realized that many extensive occurrences of BIF, and especially those of South Africa and Australia, were distinctively different from those of the Lake Superior area. In particularly, they lacked the granular structure that was generally present in the latter, and had a different pattern of banding. The term 'granular iron-formation', or GIF, is now preferred for the type of BIF to which the name was first applied in the USA. In a perfectly rational classification it would

probably be best to restrict the names BIF and GIF to two types of iron-rich sedimentary rock with separate textural features, and to use the name 'iron-formation' (IF) as a generic name for both. But because the use of BIF is so strongly entrenched, it is used here as the general term, and GIF is regarded as a particular type of BIF.

BIF is therefore defined as a chemically precipitated sedimentary rock containing at least 15% of iron, typically thin-bedded or laminated, and commonly containing layers of chert.

Chemical and Mineralogical Composition

Although the definition above requires a minimum iron content of 15%, the great majority of rocks that would be called BIF contain between 25% and 35% Fe. About half of the rock by weight consists of iron oxides, while the remainder is mainly silica. Carbon dioxide (as carbonate) is present as a minor constituent in many BIFs, and is a major component in some, but all other oxides (e.g., Al_2O_3, MgO, Na_2O, K_2O, P_2O_5) are relatively minor, and trace elements are insignificant. Haematite (Fe_2O_3) and magnetite (Fe_3O_4) are the most abundant iron minerals. Others that may be present are carbonates (ankerite, siderite) and silicates (stilpnomelane, greenalite, riebeckite). The silica normally occurs as microcrystalline quartz, usually called chert. Neither the chemical nor mineralogical composition of GIF differs significantly from that of BIF.

The Banding

The banding of BIF, with the exception of the type known as GIF, which is dealt with later, may best be described by reference to that of the well preserved BIFs which were laid down in the Hamersley Basin of north-western Australia between about 2650 Ma and 2450 Ma. Apart from very low-grade metamorphism, open folding, and gradual uplift and exposure, the rocks of this basin remain essentially unaffected by post-depositional events; the basin had a depositional area of at least $10^5\,km^2$. The 1.2 km (compacted thickness) of BIF present within the 2.5 km-thick median sequence of this basin forms five main units interstratified with shales and carbonate sediments. One of these units, the *ca*. 140 m-thick Dales Gorge Member, is particularly useful for stratigraphic study in that it has 33 internal alternations, termed macrobands, of BIF and shale. The banding of the BIF of each of the 16 BIF macroband is termed 'mesobanding', and is defined by sharply defined alternations, on a centimetric scale, of dark iron-rich (silica-poor) mesobands and light silica-rich (iron-poor) mesobands; the latter are usually called chert. Chert mesobands of the Dales Gorge Member BIF are normally between 5 mm and 15 mm thick, with an average thickness about 8 mm; they constitute about 60% of the total BIF volume; the intervening iron-rich mesobands have a mean thickness of about 10 mm, and make up about 20% of the total BIF volume. A minor proportion of mesobands of magnetite and/or carbonate make up the remainder of the volume. All the mineral components of all mesoband types are made up of fine-grained closely crystalline material that shows no evidence of a clastic contribution to the formation of the rock.

The monotonously repetitive alternation of iron-rich and iron-poor mesobands, on the same centimetric scale as those of the Dales Gorge Member, is an easily identifiable characteristic of BIF wherever it occurs, and is the scale of banding from which the name BIF derived. In field exposures the alternating bands may be coloured black and white, red and white, or red and black, largely according to the weathering of the rock, and often give it a spectacularly striped appearance (Figure 1).

Within many chert mesobands of the Dales Gorge Member there is a regular small-scale lamination defined by a concentration of some Fe mineral within the pervasive matrix of microcrystalline quartz. The Fe mineral may be either hematite, magnetite, carbonate (siderite or ankerite), stilpnomelane, or some combination of these. This lamination within individual chert mesobands has been called microbanding. The iron-rich laminae that define microbands are normally separated by thicknesses of between 0.2 mm and 1.6 mm of virtually iron-free chert. Within a microbanded chert mesoband the microband thickness tends to vary only slightly, but from one such mesoband to another there may be significant variation. Since its identification in the Dales Gorge

Figure 1 Hand specimen (15cm wide) of folded Archaean BIF from the Yilgarn Craton, Western Australia. The conspicuous red and black banding is mesobanding; the red mesobands are chert and the black mesobands consist mainly of iron oxides and microcrystalline quartz. (Photograph by GJH McCall).

Member, it has been identified in a number of other well-preserved BIF units elsewhere.

The characteristic mesobanding of BIF is not present in GIF. GIF also has alternations of iron-rich and iron-poor material, but these are typically coarser and much less regular. The coarsely crystalline chert bands are commonly wavy or lenticular. Both iron-rich and silica-rich bands may be granular, more particularly the latter. The iron-rich bands of GIF, as the name implies, often consist of a close-packed and lithified mass of granules or ooliths, averaging about a millimetre in diameter. They are made up of iron oxides, with or without quartz, and the intergranular material consists mainly of the same minerals, but usually with a lower iron content.

Continuity of Banding

A remarkable stratigraphic feature of Hamersley Basin BIF is its exceptional lateral continuity. Thus the Dales Gorge Member, used above as a model for the description of banding, and the other main BIF units, are easily identifiable over the entire basin area. BIF macrobands within the Dales Gorge Member are similarly recognisable through their constant relative thicknesses, and within BIF macrobands both individual centimetre-scale mesobands, as well as sub-millimetre microbands within them, have been correlated over 300 km. This degree of fine-scale lateral stratigraphic continuity is reminiscent of evaporites, particularly in the Permian of Europe and the United States.

Metamorphic and Tectonic Modification

Because of their exceptional freedom from post-depositional metamorphism and tectonic deformation, the BIFs of the Hamersley Basin have been used as 'type examples' for the introductory descriptions above; the BIFs of the Transvaal and correlative Griqualand West basins of South Africa, which are of similar age, also have the same excellent preservation. But in this respect they are atypical of the majority of BIF occurrences. Many Early Precambrian examples have suffered significant metamorphic modification, which begins with coarse-grained recystallisation (annealing) of both the initially microcystalline chert and fine-grained iron oxides, and continues with the growth of iron-rich silicates (e.g., grunerite, ferrohypersthene, fayalite). Even in early metamorphic stages the delicate fine textures within the banding tend to become blurred, but the essential iron-rich/iron-poor alternation of the mesobanding shows a robust resistance to complete obliteration. Most metamorphism of BIF appears to be isochemical, but there is some evidence of chemical modification at higher grades.

Older Precambrian BIFs also tend to have undergone significant tectonism, particularly those of Archaean greenstone belts of all continents, where the BIFs often form curvilinear steeply dipping units which are useful in deciphering complex structures. It is characteristic of deformed BIF in these belts that it appears to have reacted sensitively to tectonic stress, forming complex internal flowage folds defined by the banding.

Distribution Over the Earth

BIFs are widely distributed throughout the Precambrian areas (cratons and shields) of all continents except Antarctica, where only one occurrence is so far known. BIFs of the older cratons include the oldest known example, at Isua, in Greenland, aged about 3.8 Ga. They are consistently present in the greenstone belt sequences of all the main old cratons. Examples include the Abitibi Belt of the Superior Province of Canada, the greenstone belts of the Yilgarn and Pilbara Cratons of Australia, and the those of the Baltic Shield, the North China Craton, the Amazon Craton of Brazil, the Kaapvaal and West African Cratons, and the BIFs of the Ukraine Craton, notably at Krivoi Rog; most of these are relatively thin, tectonised, and metamorphosed. Most of these greenstone-associated BIFs have ages between 2.8 Ga and 2.5 Ga. A different style of BIF occurrence is present in four of the Gondwana continents (South America, southern Africa, India, and Australia). In this type the BIFs occur as well preserved, gently dipping supracrustal sequences, which may form conspicuous topographic plateaus. The Carajás Formation of the Amazon Craton, the Cauê Itabirite of the São Francisco Craton (both in Brazil), the Kuruman Iron Formation and Penge Iron Formation of South Africa, and the Mulaingiri Formation of the Indian Karnataka Craton all belong to this group, which includes the BIFs of the Australian Hamersley Basin, already mentioned above; these occurrences have been called the Great Gondwana BIFs, and are mostly younger than those present in greenstone belt occurrences. Further mention of a final distinct category of BIF occurrences, which includes Rapitan in the Yukon, and those of the Damara Belt, in Namibia, is made below the next heading.

Distribution in Time

Mention has already been made not only of the general restriction of BIFs to the Precambrian, but also to

specific intervals within that Eon. This topic needs closer attention, since it is critical for understanding the genesis of BIF, and has been a subject of strong controversy. A relevant point to be made is, of course, that the topic could not be discussed at all without the availability of the precise isotopic dating techniques that have established, only in the last couple of decades, a reliable time-scale for the entire span of Precambrian time.

An age of about 3.8 Ga has already been noted for the oldest BIF, from Isua. This is still the oldest known sedimentary (or more accurately metasedimentary) succession, so that BIF is included among the oldest sedimentary rocks. The abundance of BIF in Archaean greenstone belts indicates that it continued to be deposited intermittently until the end of that Era, which by convention is accepted as 2.5 Ga. By far the two largest basins of BIF deposition on Earth, in terms of contained Fe, are the Hamersley Basin and the Transvaal/Griqualand West Basin, and in these, massive deposition of BIF continued until ~2450 Ma. There is then a paucity of BIF occurrences until the Lake Superior GIFs, at ~1850 Ma, which represents a global peak for this type. A long interval in which no BIF deposition is known then followed, and it was not until the end of the Precambrian, in the Late Neoproterozoic, that there is a final burst of BIF deposition, which includes the Rapitan and Damara examples.

In summary, the overall picture is of a long early period of intermittent BIF deposition, culminating in a peak at ~2500 Ma, after which there was a smaller burst at ~1800 Ma, and finally, after a long hiatus, another significant depositional event at the end of the Precambrian. It has been pointed out that some iron-rich Phanerozoic rocks closely resemble Precambrian BIFs, but these exceptions only prove the rule, and indicate that there were special environmental conditions during the Precambrian, and particularly during the Early Precambrian, which led to the deposition of BIF.

Association with Other Rocks, and with Volcanism

BIFs do not, of course, exist in isolation, but occur in association with other lithologies as components of sedimentary, or volcanosedimentary, sequences. Two general stratigraphic points are worth making: in the first place, BIF tends to form discrete, clearly demarcated, units of significant thickness and wide lateral extent – it does not form thin or discontinuous beds interlaminated, or interdigitated, with other lithologies; and in the second place, BIF has never been demonstrated to grade imperceptibly into another lithology, whether laterally or vertically. It is as though the conditions for BIF deposition, in any given basin, were intermittently switched on and off abruptly, rather than by any gradual change in the depositional environment.

Within those two overarching generalisations, BIF has not been demonstrated to occur in immediate sedimentary contact with any preferred lithology. Examples can be found of association with both fine-grained (shale) and medium-grained (quartzite) epiclastic sediments, with turbidites and with carbonates. For BIFs in Archaean greenstone belts, there is a common association with mafic volcanic rocks, but an immediate and local genetic link between the two has never been unequivocally demonstrated. While it is true that volcanic rocks, both felsic and mafic, are commonly present as components of depositional basins in which BIF occurs, the fact remains that no volcanic rocks at all are closely associated with one major BIF of Brazil – the Cauê Itabirite of the Quadrilátero Ferrífero. There is abundant trace element and isotopic evidence that BIFs have an igneous connection, but an immediately local association cannot be inferred from this.

Theories of Origin

From direct observation of the present environment it is easy to see that lithostratigraphic units of, say, sandstone are likely to have been formed by the transport of sand in rivers to the sea. There is no such direct way to determine how BIF formed, and a credible hypothesis for its origin must be built up step-by-step, from various lines of evidence. An early suggestion in the Lake Superior area was that BIF may be an acid lava, and more recently it has been proposed that BIF was initially a carbonate rock which had been replaced by iron and silica.

Most workers now accept that BIF is a chemically precipitated sedimentary rock, but that acceptance does not remove the need to explain many associated problems, including:

i. what was the source of the materials (both iron and silica)?
ii. how were they transported to the basin?
iii. did deposition occur in a lake or in the ocean?
iv. what caused precipitation?
v. what does the banding represent?
vi. to what degree does the final rock differ from the precipitate?

The first two of these questions were at one time debated in terms of two radically opposite concepts. One suggested that the iron and silica were derived from deep weathering of continental crust, and were

selectively carried to the adjacent basin by rivers; the other proposed that the source of the the iron and silica was fumarolic volcanism very close to or within the basin, obviating the need for significant transport. As the debate progressed both ideas were generally abandoned, in favour of a model in which both iron and silica were in very dilute solution in ocean water, and were precipitated from the water of marginal basins kept supplied with both materials by appropriate circulation; in this model the primary source of iron and silica was ocean-floor volcanism or volcanic rocks, either at mid-ocean ridges or more generally from the ocean floors. As far as the third question is concerned, the abundance of BIF in the Precambrian stratigraphic record, its frequent association with clearly marine lithologies, and the difficulty of forming such immense thicknesses of iron-rich material in lakes, jointly make the lacustrine hypothesis untenable.

The fourth question then arises from a model involving precipitation from the water of an ocean margin basin which has low levels of dissolved iron and silica. Attention here has focused on the iron component, and specifically on the fact that dissolved iron will be in the ferrous state, and that oxidation will lead to precipitation in some ferric form. A variety of mechanisms for this oxidative precipitation has been proposed, including algal photosynthesis, anoxygenic bacterial photosynthesis, photo-oxidation by sunlight, and decomposition of water by ^{40}K radiation. Current research increasingly involves attention by microbiologists to the detailed mechanisms by which early biota could have effected precipitation. From a sedimentological viewpoint, the key point is that such mechanisms are quantitatively viable for precipitation of iron from small basin water concentrations.

The fifth question is one which received little early attention. Although it was generally agreed to represent the stratification of the BIF there was virtually no analysis of how the mesobanding was generated. A structured model was first proposed in respect to the BIFs of the Hamersley Basin. There it was argued that the microbands, whose presence has been noted within the chert mesobands, probably represent annual layers, or chemical varves: the thin iron-rich laminae of microbands may be presumed to represent summer seasons of high photosynthetic activity. Acceptance of this hypothesis permits calculation of the quantity of iron precipitated in the Hamersley Basin per unit area each year ($225\,t \cdot km^{-2} \cdot yr^{-1}$), and hence, from knowledge of the bulk composition of the rock, an estimate of the compacted depositional rate. But these arguments have still not addressed the fifth question, which asks about mesobanding, not microbanding. So an additional step of this model added the proposal that the initial precipitate was a collloidal silica ferrihydrate gel which was compressed to about 10% of its initially deposited thickness during burial and diagenesis; and its final step was the suggestion that the mesobands developed during this stage by differential compaction. These last two steps, of course, address the sixth question also.

Some aspects of this depositional model have been challenged. Most workers on BIF agree that an annual (varve) significance for microbands is likely, although a diurnal or even tidal origin has been suggested. Others have preferred to see a direct link between mesobanding and iron supply from mid-ocean ridge activity, the mesobands representing pulsed variations in supply, with microbanded chert mesobands representing stable periods of perhaps of tens of years of relatively low iron availability. The origin of mesobanding still remains uncertain.

Unsolved Problems

Apart from the origin of mesobanding, the uncertainty of whose origin has just been noted, three other still-unsolved questions of BIF genesis are worth emphasizing in conclusion. The possible involvement of biotic processes in the precipitation of the iron of BIFs has already been noted; further studies related to this question are clearly not only significant for understanding BIF but for understanding important processes of biochemical evolution. The distribution of BIF in time has already been described, but its significance was not discussed. An early hypothesis tied the Late Archaean peak closely into a step in the biochemistry of photosynthesis, but additional geochronological work has lessened support for that proposal. A recent suggestion that the deposition of BIF is related to deep-water phases in basin development, and that this development shows systematic tectonically controlled secular changes, has yet to be fully debated. And finally, the place of BIF in the chemical evolution of the atmosphere and oceans has yet to be fully understood. Research on this uniquely puzzling sedimentary rock still holds many challenges.

See Also

Precambrian: Overview. **Rocks and Their Classification**. **Sedimentary Rocks:** Chert; Ironstones.

Further Reading

Appel PWU and La Berge GL (eds.) (1987) Precambrian iron formations. Athens: Theophrastus Publications.

Beukes NJ (1980) Lithofacies and stratigraphy of the Kuruman and Griquatown Iron Formations, Northern Cape Province, South Africa. *Transactions of the Geological Society of South Africa* 83: 69 86.

Beukes NJ and Klein C (1992) Models for iron formation deposition: Section 4.3. In: Schopf JW and Klein C (eds.) *The Proterozoic Biosphere: a multidisci plinary study,* pp. 146 151. Cambridge: Cambridge University Press.

Isley AE (1995) Hydrothermal plumes and delivery of iron to banded iron formations: *Journal of Geology* 103: 169 185.

James HL and Sims PK (eds.) (1973) Precambrian iron formations of the world. *Economic Geology* 68(7): 913 1179.

Klein C and Beukes NJ (1992) Time distribution, stratig raphy and sedimentologic setting and geochemistry of Precambrian banded iron formations: Section 4.2. In: Schopf JW and Klein C (eds.) *The Proterozoic Biosphere: a multidisciplinary study,* pp. 139 146. Cambridge: Cambridge University Press.

Morris RC (1993) Genetic modelling for banded iron formation of the Hamersley Group, Pilbara Craton, Western Australia. In: Blake TS and Meakins A (eds.) *Archaean and Early Proterozoic Geology of the Pilbara Region, Western Australia, Precambrian Research:* 60 243 286.

Simonson BM (1985) Sedimentological constraints on the origins of Precambrian iron formation. *Geological Soci ety of America Bulletin* 96: 244 252.

Trendall AF (2002) The significance of iron formation in the Precambrian stratigraphic record. In: Altermann W and Corcorane PL (eds.) *Precambrian Sedimentary envir onments: a modern approach to depositional systems.* International Association of Sedimentologists Special Publication 44: 33 66.

Trendall AF and Blockley JG (1970) The iron formations of the Precambrian Hamersley Group, Western Australia, with special reference to the associated crocidolite. *Western Australia Geological Survey Bulletin* 119: 365.

Trendall AF and Morris RC (eds.) (1983) *Iron formation: Facts and Problems.* Elsevier: Amsterdam.

Chalk

J R Ineson and L Stemmerik, Geological Survey of Denmark and Greenland, Geocenter Copenhagen, Copenhagen, Denmark
F Surlyk, University of Copenhagen, Geocenter Copenhagen, Copenhagen, Denmark

Introduction

Chalk is a familiar rock type, particularly amongst Europeans, forming spectacular white cliffs along coastlines flanking the North Sea, the English Channel and the Baltic Sea. The essential characteristic of a true chalk is its microscopic composition – being composed predominantly of the skeletal remains of tiny calcareous marine algae known as coccolithophorids (Figure 1). Following their appearance in the Jurassic, these haptophycean algae became a common constituent of marine sediments and remain important components of the marine ecosystem today. Only during specific periods of Earth history and in certain palaeogeographic areas, however, were conditions such that pure carbonate oozes accumulated and were preserved on continental shelves and in vast epeiric seas. Chalk is thus most characteristic of the Upper Cretaceous (and in places the Danian) of north-west Europe and North America. This review focuses on these typical chalks, particularly from north-west Europe (Figure 2).

Chalk as a Sediment

Composition

A typical chalk is a fine-grained carbonate rock (a lime mudstone or micrite), the lithified equivalent of pelagic carbonate oozes recorded from ODP boreholes in present-day oceans. The sediment is dominated by debris derived from coccolithophorid algae that comprise a spherical calcareous test (coccosphere) up to several tens of microns in diameter. The test is made up of overlapping circular or elliptical discs or rings (coccoliths), which are 1–20 μm in diameter and are, in turn, constructed of tiny calcite platelets, laths or rays ranging from 0.1 to 2.5 μm across (typically 0.5–1 μm; Figure 1). The complete coccosphere is only rarely preserved and the chalk consists largely of coccoliths and their disaggregated platelets and spines (rhabdoliths). Despite being dominated by coccolithophorid debris, planktonic foraminifers and calcispheres are also common in chalks, together with coarser skeletal elements from both pelagic and benthic organisms such as belemnites, bryozoans, echinoids, bivalves, brachiopods, serpulids and sponges.

Mineralogically, pure chalks are composed of low magnesium calcite; this is the stable form of calcite

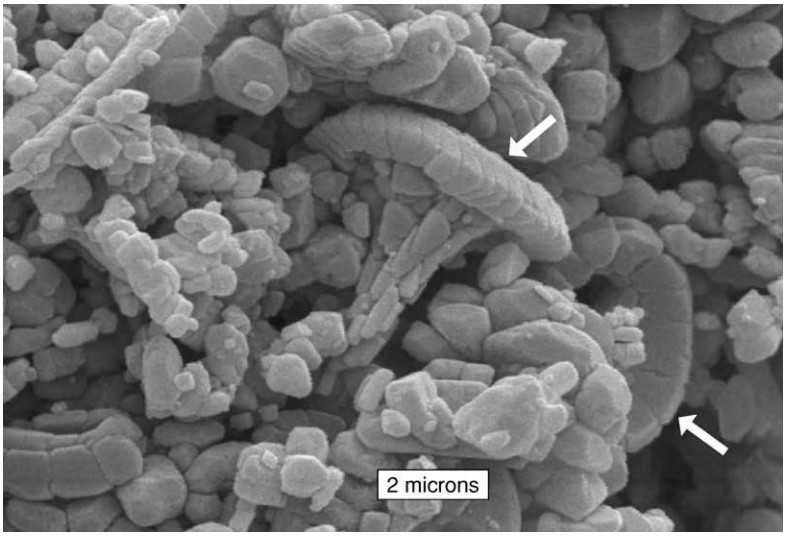


Figure 1 Scanning electron micrograph of chalk showing disc shaped coccoliths (arrowed) in a mass of disaggregated platelets and cement (e.g., top right). Uppermost Maastrichtian, eastern Denmark. Photo: P Frykman.

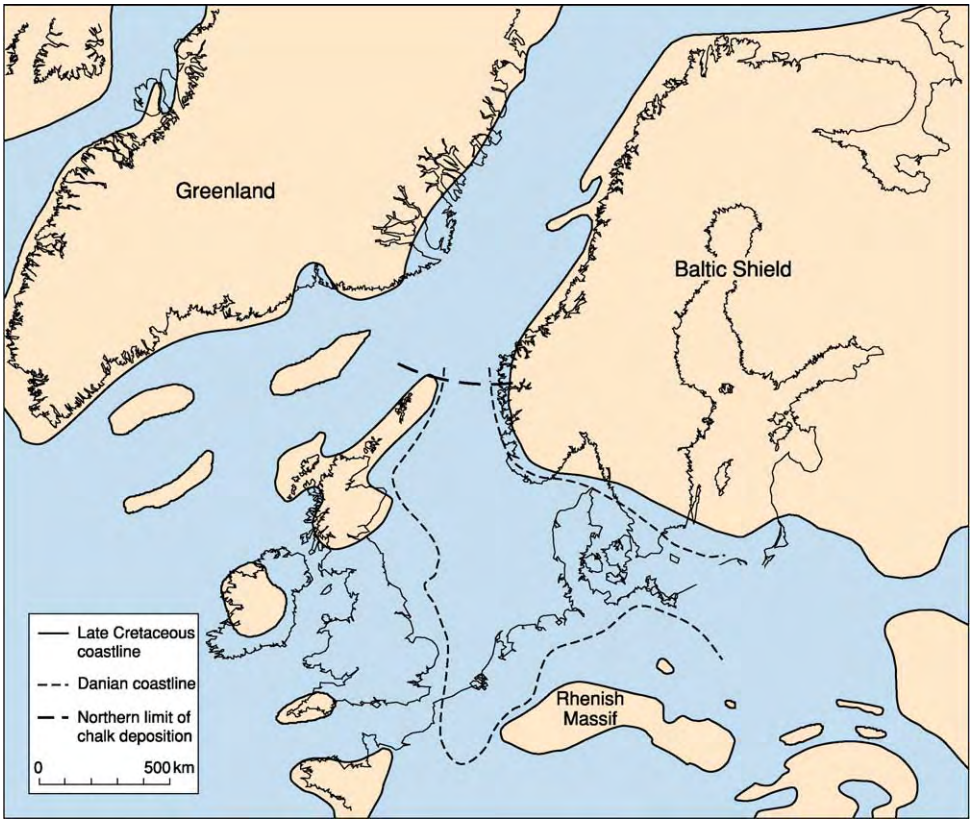


Figure 2 Late Cretaceous Danian palaeogeography of the North European North Atlantic region showing the land:sea distribution and the northward limit of the chalk facies in the North Sea region (at a palaeolatitude of *ca.* 50° N).

seen in most ancient limestones and is commonly the result of diagenetic modification of unstable carbonate precursors (e.g., high magnesium calcite, aragonite). In the case of chalks, however, this composition is essentially primary since coccolithophorids secrete low magnesium calcite. The chalk is thus not prone to significant diagenesis under normal marine conditions, imparting a chemical stability to the chalk that

is unique amongst limestones. This is a major factor in its utility as a reservoir both for hydrocarbons and water (see below). The content of siliciclastic detritus in chalk is typically low, but clay-rich chalks are observed at certain stratigraphic and palaeogeographic positions and clay content is often instrumental in highlighting a distinctive cyclicity in the chalk succession. Primary biogenic silica may be present in the form of opaline radiolarians, diatoms and sponge spicules, but more characteristically silica in the chalk is represented by bands of nodular flint or its precursor, cristobalite lepispheres. Glauconite and phosphorite are also important authigenic minerals in chalks, particularly over structural highs or associated with hiatal surfaces.

Facies and Processes

Typical fine-grained coccolith-dominated chalk forms one, deeper-water variant within a spectrum of facies referred broadly to the chalk family (Figure 3). The faunal content increases towards the basin margins and over structural highs, and in the shallowest parts of the north-west European chalk sea coarse-grained, shallow marine carbonates were deposited. These are dominated by skeletal carbonate sands with abundant bivalves, brachiopods and echinoderms. Basinwards, this facies belt may pass into a zone dominated by bryozoan mounds before entering the area of true chalk deposition (Figure 3). The marginal facies of the chalk sea are known mainly from onshore outcrops in Denmark and southern Sweden; the Danian-age bryozoan mound complexes exposed in eastern Denmark are particularly impressive (Figure 4).

The pelagic chalk was initially deposited as an ooze consisting of coccoliths with a variable content of foraminifers and calcispheres, and a landward increasing content of invertebrate fossils, including bryozoans, echinoderms, brachiopods, and bivalves. The chalk seafloor was a unique, long-lived macrohabitat and a remarkably well-adapted, highly specialized fauna gradually developed, dominated by millimetre-sized suspension-feeding invertebrates. This reached a climax in the Late Campanian–Maastrichtian with a diversity of several thousand benthic species. Most epifaunal species were very small allowing attachment to very restricted hard substrates such as individual skeletal fragments. Other organisms developed 'snowshoe' strategies (a flat profile, often with long marginal spines, or hemispherical with the convex valve downwards) permitting the organism to 'float' on the soft substrate.

The coccolithophorid algae lived largely within the photic zone near the sea surface, and their skeletal debris settled slowly to the seafloor from suspension, most likely in the form of faecal pellets. At the sea floor, the ooze was watery with a primary porosity of 70–80%. The grain size was extremely fine, probably about 1 μm, since the coccolithophorid tests readily disaggregate into their component coccoliths and platelets. The pelagic ooze typically accumulated under well-ventilated conditions on the sea floor where sufficient oxygen was available to support a diverse fauna of burrowing benthic invertebrates – the pelagic chalks are thus characteristically intensely bioturbated. Studies of the onshore chalk exposures have revealed composite ichnofabrics that reflect the succession of diverse benthic communities that occupied the uppermost layers of the ooze as it experienced gradual dewatering and changed from a soupground to a softground (Figure 5). The trace fossils in pelagic chalks reveal much information about substrate conditions, sedimentation rates and oxygenation as well as evidence of non-deposition and the development of firmgrounds and hardgrounds. This is based on the recognition of characteristic groups of trace fossils, known as tiers, that characterize different levels in the ooze from the sea bed down to a metre or more below the sediment surface (Figure 6). The shallowest tier completely obliterated the primary fabric and only rarely are discrete trace fossils recognizable (e.g., diffuse *Planolites*). Downwards, trace fossils are better preserved and define a succession of tiers characterized by forms such as *Thalassinoides*, *Zoophycos* and *Chondrites*. Nodular chalks and hardgrounds are distinctive features of shelf-sea chalks and record decreasing rates of sedimentation and consequent increasing intensity of cementation at or near the sea floor. True hardground surfaces (i.e., cemented layers exposed on the seafloor) may show evidence of encrustation of the hardened surface by bivalves, serpulids and bryozoans and boring by sponges, algae and bivalves. The hardground surface may also be impregnated by phosphorite or glauconite.

Resedimentation of pelagic chalks has been documented on all scales from both onshore and offshore areas, and it is widely recognized that intrabasinal slides, slumps, turbidity current and debris flow deposits form an important part of the chalk depositional system (Figure 3). The largest slides occur close to tectonic inversion or salt structures and involved downslope movement of slabs of semi-lithified chalk, tens of metres thick. They may be identifiable on seismic data, but can be difficult to recognize in core since there is little or no internal deformation within the slide masses. Such allochthonous sediment slices have, however, been recognized on the basis of anomalous biostratigraphic data, a Maastrichtian interval sandwiched between Danian chalks for example. In contrast, slumps are more readily recognized by the presence of pervasive deformation

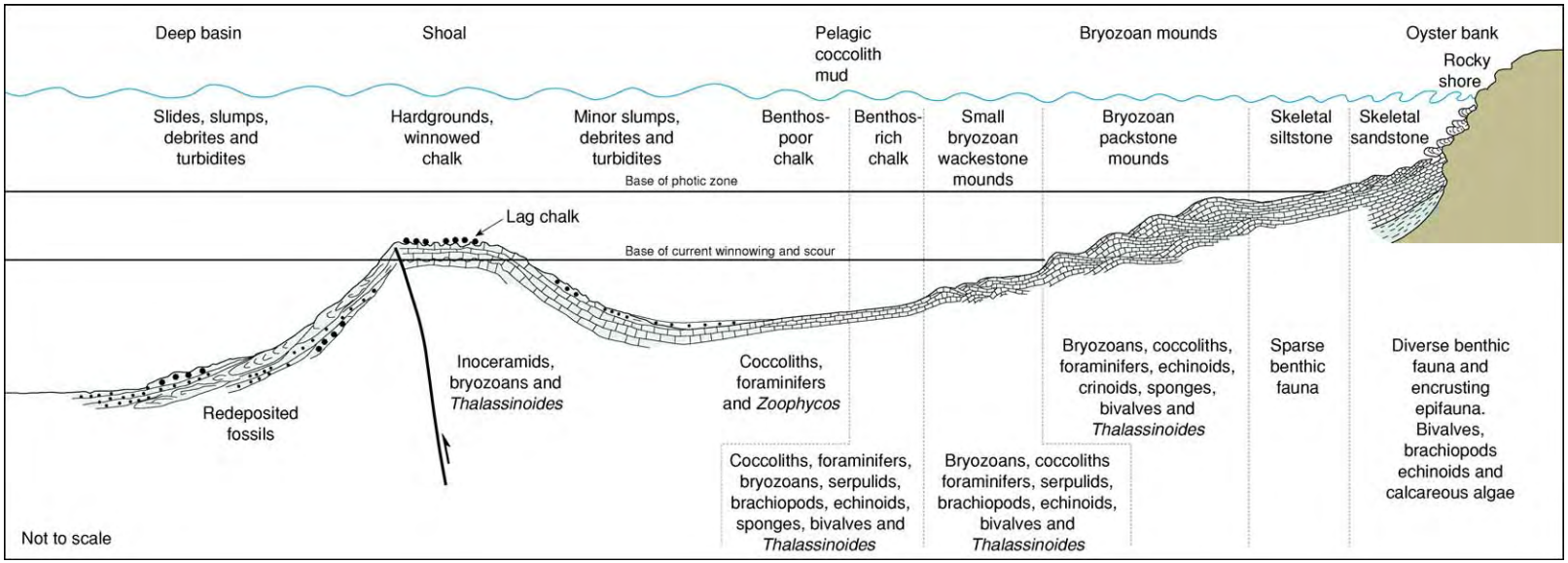


Figure 3 Schematic Late Cretaceous Danian facies model for the NW European chalk sea showing the main facies belts passing from shoreline skeletal sands through bryozoan mounds to the area of 'true' chalk deposition. Based mainly on a NE SW transect from the western margin of the Baltic Shield to the central North Sea. Reproduced from Surlyk F, Dons T, Clausen CK, and Higham J (2003) Upper Cretaceous. In: Evans D, Graham C, Armour A, and Bathurst P (eds.) *The Millenium Atlas: Petroleum Geology of the Central and Northern North Sea*, pp. 213 233. London: Geological Society.

Figure 4 Upper Maastrichtian Danian chalks exposed at Stevns Klint, eastern Denmark. The lighter coloured, lower third of the cliff (up to the prominent overhang) is the uppermost Maastrichtian. The K/T boundary (arrow) is gently undulating and the boundary clay layer is only preserved in the depressions. The Danian bryozoan rich succession above shows well developed mounds. Height of cliff *ca.* 40 m. Photo: F Surlyk.

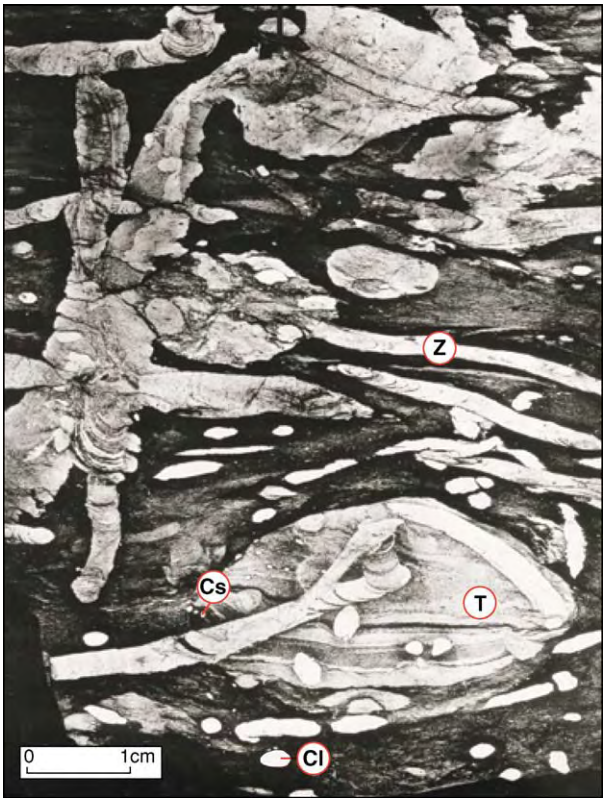


Figure 5 Photograph of slabbed core (Maastrichtian, Denmark) in which the ichnofabric is enhanced by oil staining. Note the complex cross cutting relationships recording the overprinting of successive tiers (see **Figure 6**). Cl, large *Chondrites*; Cs, small *Chondrites*; T, *Thalassinoides*; Z, *Zoophycos*. Photo courtesy of RG Bromley.

structures such as isoclinal folds and stratigraphically inverted successions. Chalk debrites, comprising chalk pebbles or slabs supported in a fine-grained chalk matrix, form a significant part of the Maastrichtian–Danian succession in the North Sea Central Graben. Most resedimented chalk clasts are plastically deformed, implying that they were poorly lithified at the time of deposition. However, the presence of angular clasts in some debrites indicates that some of the material originated from lithified chalk, either from penecontemporaneous firmgrounds/hardgrounds or from exhumed more deeply buried chalks, for example at fault scarps. Sand-grade 'classical' turbidites are uncommon in the chalk, most likely due to the scarcity of sand- and silt-sized material, although dilute low-density turbidity currents were important in the redistribution of mud-grade sediment.

The Chalk Sea

In the Late Cretaceous, pelagic carbonate oozes extended far onto the European craton and formed the dominant facies for tens of millions of years. This was the result of a unique coincidence of global and regional factors. The chalk sedimentary record attests both to such long-term controlling factors as eustatic sea-level and regional tectonics and to the influence of short-term climatic variation controlled by orbital forcing mechanisms.

Palaeogeography

The chalk sea of north-west Europe existed for more than 35 Ma, from the Cenomanian to the Danian, at a time when global sea-level was at its highest during the Phanerozoic and relative tectonic stability prevailed in the region. Much of the north-west European craton was flooded to depths in excess of 50 m. Hinterland relief was low and potential source areas were restricted in extent so siliciclastic supply was limited and a pelagic carbonate drape accumulated, extending from a palaeolatitude of 35° N northwards to 50° N where the carbonates passed into siliciclastic muds (**Figure 2**). The biogenic components largely belonged to the heterozoan association that today characterizes cool-water, temperate carbonate systems; typical Cretaceous tropical organisms, such as reef corals, large foraminifers and rudist bivalves are absent or rare in the chalk of north-west Europe. However, direct latitudinal comparison with present-day seas are invalid since the Cretaceous was one of the 'greenhouse' phases of Earth history when equable temperatures extended further poleward than in our present 'icehouse' situation. The chalk sea is thus probably best characterized as ranging from warm temperate to sub-tropical, despite its mid-latitude

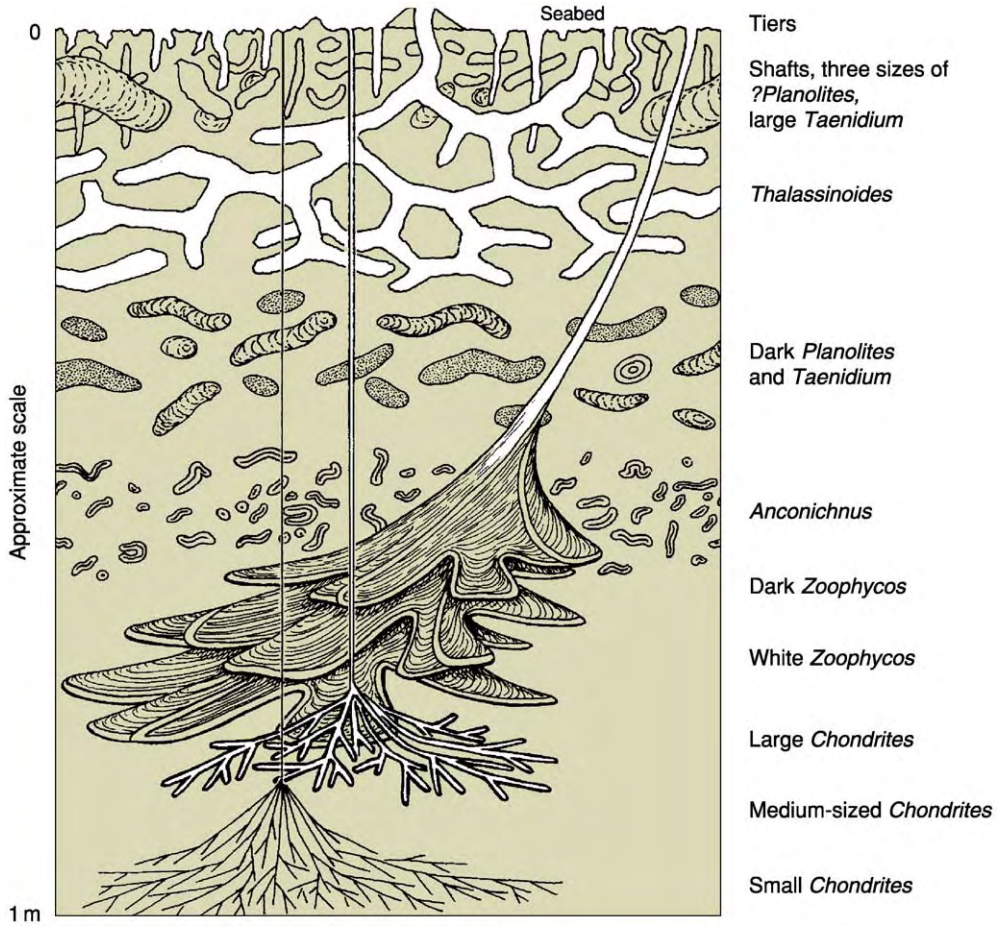


Figure 6 Schematic 'snapshot' of a Cretaceous ooze profile showing the various trace fossil tiers at different levels beneath the sea floor. Modified from Ekdale AA and Bromley RG (1991) Analysis of composite ichnofabrics: an example in uppermost Cretaceous chalk of Denmark. *Palaios* 6: 232 249.

setting. The overwhelming dominance of coccolithophorid skeletal material suggests that overall the chalk sea was a low nutrient (oligotrophic) setting. Today, shelf seas are separated from the open ocean by shelf break fronts that isolate inshore waters from the open ocean. During maximum sea-level highstand in the Late Cretaceous, the high water depths over the shelf break precluded the development of an effective shelf front and oceanic conditions extended far onto continental shelves and into epeiric seas.

Late Cretaceous pelagic sedimentation rates are estimated at 2–2.5 cm per thousand years. The Upper Cretaceous–Danian chalk succession is typically a few hundred metres thick where exposed in countries bordering the North Sea (Figure 2), but can be over a kilometre thick within major graben structures in the central North Sea and thicknesses in excess of 2 km are found in the Danish Basin. Chalk deposition was interrupted at the Maastrichtian–Danian boundary due to the mass extinction, which included coccolithophorids and foraminifers with only a few surviving species. The mass extinction severely affected all carbonate-shelled organisms and some, such as the ammonites, became totally extinct. The boundary is marked by a thin clay layer in all complete sections (Figure 4); this layer shows enrichment in iridium, forming the basis for the hypothesis that the mass extinction owed its origins to the impact of a giant meteorite. Carbonate deposition, however, rapidly resumed; the surviving microplankton and benthic invertebrates soon regained high diversities and the Danian ecosystem closely resembled that of pre-extinction, Cretaceous times.

Sea-floor relief in the NW European chalk sea was subdued and the carbonate system is best considered overall as a gently shelving ramp (Figure 3). However, significant depositional relief was developed along structures inherited from Jurassic rift events or related to localized Cretaceous inversion or salt movements. The North Sea Central Graben, for example, was a north–south-trending trough with a complex morphology formed both by the marginal slopes and by

Figure 7 Cliff section, *ca.* 50 m high, at Port d'Amont, NE of Etretat, France, showing prominent chalk flint cycles. Note slump sheet (right) and coalescing hardgrounds at beach level. Photo: F Surlyk.

intra-basinal ridges and domes along inversion axes and atop salt structures, respectively. Such relief led to sediment instability and instigated sediment slumps and gravity flows, resulting in redeposition of the coccolith ooze in deeper depocentres. The depositional relief may also have inhibited bottom water circulation and promoted the periodic development of anoxia/dysoxia in the deeper parts of the Central Graben.

In marginal settings, over intrabasinal highs and in areas of focussed, amplified bottom currents, the chalk sea floor locally developed marked depositional relief, both in the form of aggradational mounds and ridges, and erosional features. Thus, seismic data from the Danish Basin, the North Sea and onshore UK reveal ridges and valleys with a relief of up to 150 m and width of several kilometres. These features are combined constructional/erosional structures and were probably controlled by long-lived contour current systems. On a smaller scale, well-developed bryozoan-rich mounds in the Maastrichtian–Danian of eastern Denmark show amplitudes of 50–100 m, heights of 5–9 m and flanks dipping up to 20°; they show a marked asymmetry recording lateral (southwards) migration of the mounds (Figure 4). At Etretat, in northern France, Coniacian–Santonian chalks exposed in dramatic cliff sections (Figure 7) display a complex array of erosional and constructional architectures that record enhanced current activity in a tectonically constrained setting close to the margin of the chalk sea.

Cyclic Sedimentation and Orbital Forcing

Chalk successions are often overtly cyclic in nature, the typical decimetre- to metre-scale cyclicity being picked out either lithologically, as in chalk/flint and chalk/marl cycles (Figure 7), or due to changes in the fabric of the chalk, as in laminated/bioturbated cycles and those revealed by variation in the intensity or type of bioturbation. Detailed correlation of such small-scale cycles in the Cenomanian, constrained by biostratigraphy, has demonstrated their lateral persistence across the basin – at certain levels, individual cycles have been correlated from southern England, over northern Germany to southern Crimea, a distance of nearly 4000 km! This small-scale cyclicity records recurrent change in a number of interrelated factors including carbonate productivity, the balance between productivity and siliciclastic input, and bottom-water conditions, factors that are thought to have been ultimately controlled by subtle fluctuations in climate dictated by orbital fluctuations in the Milankovitch frequency band. The precession signal (mode at 21 ka) dominates, at least in the Cenomanian where the most detailed studies of small-scale cyclicity have been undertaken. Sequence-scale sea-level changes were driven by the long eccentricity cycle of 400 ka, allowing sequence stratigraphic correlation from north-west Europe to Kazakhstan and south-east India.

Chalk as a Hydrocarbon Reservoir and Aquifer

The Upper Cretaceous–Danian chalk forms significant reservoirs both for hydrocarbons, as in the North Sea Central Graben, and for groundwater, for example in Denmark, England, France and Belgium. Hydrocarbons are also produced from older, Barremian–Aptian marly chalk facies in the Danish Valdemar Field and from the Upper Cretaceous Austin Chalk in Texas. Indeed, the first hydrocarbon discovery in the North Sea, in 1966, was in chalk – the A-1 well of the Danish Kraka Field. Since that first discovery, the number of chalk fields in the North Sea has increased to nearly 30, containing almost 5 billion barrels of recoverable oil and more than 16 000 billion cubic feet of gas. Production of hydrocarbons from these fields is still a major challenge, since the chalk forms a unique family of very fine-grained reservoir rocks, characterized by high matrix porosity and low permeability, differing from most other carbonate reservoirs. The minute coccolith platelets making up the chalk are typically 0.5–1 μm across, and pores and pore throats are on the order of a micron in size, reducing the matrix permeability to a few millidarcy even at porosities of more than 35%.

The North Sea chalk is composed almost entirely of the stable low magnesium variant of calcite and has not been subjected to freshwater diagenesis.

Diagenetic modifications of the chalk were controlled by early processes at or near the sea floor and by the later burial history. The chalk ooze had an initial porosity of 70–80%, but dewatering due to bioturbation rapidly reduced porosity to about 50%. The porosity declined further to 35–40% at depths of around 1000 m due to compaction. At greater burial depths, the effects of pressure solution became more important and under normal conditions the matrix porosity is around 10% at burial depths of 2000 m. The permeability of the chalk is directly related to the porosity, so reduced porosity also means lowering of permeability and thereby hydrocarbon productivity (Figure 8). The relationship between porosity and permeability is not constant, but varies with the stratigraphic age of the chalk (Figure 8). The best chalk reservoir properties in the North Sea are found in the Maastrichtian Tor Formation where the matrix permeability for a given porosity is almost 10 times that of the Danian Ekofisk Formation. As seen in the figure, Lower Cretaceous chalks form even poorer reservoirs due to their high content of clay. The differences in reservoir properties between the Maastrichtian and Danian chalks are not fully understood, but it has been suggested that a change in the coccosphere flora across the Cretaceous–Palaeogene boundary resulted in changes in the detailed texture of the coccolith-rich sediment thereby affecting the size/geometry of pores and pore throats.

Other factors that are known to influence the quality of the chalk reservoirs are the content of clay and silica, and the mode of deposition. It was long a common belief that reworking of the chalk was the key to preservation of anomalously high porosities, a belief possibly driven by the fact that allochthonous chalks form the main reservoirs in the Norwegian North Sea sector. This view has changed, however, and although resedimentation locally has a positive effect on porosity preservation this is not always the case. On a metre-scale, however, it has been shown that facies have a major control on porosity in cyclically interbedded successions of bioturbated and laminated chalk, the highest porosities being in the laminated units. The development of firmgrounds and hardgrounds is also important since early cementation reduces primary porosity, so these layers form characteristic low-porosity zones in the chalk and may create barriers to fluid flow in the reservoir. On structural highs, several hardgrounds may coalesce, leading to a major negative effect on reservoir quality. This is particularly true for the complex hardground that developed at the Maastrichtian–Danian boundary over most of the North Sea highs. The best reservoir properties are found in the purest chalks, and the Maastrichtian Tor Formation is particularly pure with less than 5%, commonly only 1–3%, non-carbonate fraction. The Danian Ekofisk Formation has a more variable content of clay and silica throughout the North Sea; the lower Danian forms a more clay- and silica-rich (up to 20%), non-reservoir interval known as the 'Danian tight zone'. The reason for the strong negative influence of clay on reservoir quality seems to be that its presence inhibits the growth of early cement between carbonate grains, thereby preventing early lithification. As a consequence, the more weakly lithified clay-rich intervals are more easily compacted during deeper burial.

North Sea chalks are extremely fine-grained and have low permeabilities so that porosities of more than 25% are required to allow commercial

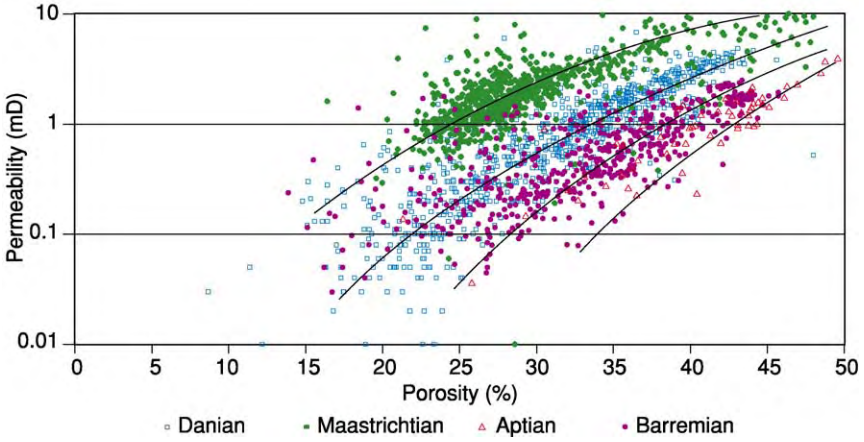


Figure 8 Plot showing the relationship between porosity and permeability for four different chalk units in the North Sea (Danish sector): the Barremian Tuxen Formation, the Aptian Sola Formation, the Maastrichtian Tor Formation and the Danian Ekofisk Formation. Reproduced with permission from Jakobsen F, Ineson JR, Kristensen L, and Stemmerik L (2004) Characterization and zonation of a marly chalk reservoir: the Lower Cretaceous Valdemar Field of the Danish Central Graben. *Petroleum Geoscience* 10: 21 33.

production. Since matrix porosity is reduced to 10% at burial depths of around 2000 m, preservation of such substantial porosity requires unusual conditions, and the North Sea chalk fields are all located in areas with significant overpressure. Another process that may result in the retention of high porosities is early oil migration into the reservoir, since the oil prevents further cementation during burial. The overpressured chalk reservoirs maintain porosities of up to 45–50% and matrix permeabilities in the 3–10 mD range at burial depths of 1700–3300 m. The effective permeability is commonly much higher due to fracturing since many North Sea chalk fields are localized over salt structures.

During the last two decades, the recovery factors in the chalk fields have increased significantly as the result of horizontal drilling, water injection and stimulation of artificial fractures. In the Danish Dan Field, the initial recovery factor of 10% has increased to 35%, and in the Norwegian Ekofisk Field estimates have increased from 18% to 38%.

In north-west Europe, chalks and other Upper Cretaceous–Danian limestones form important aquifers, both where they outcrop at the surface and in areas where they are covered by thin Cenozoic or Quaternary deposits. The chalk is the most important unconfined aquifer in the Paris Basin, both in terms of areal extent and size of resources, and production from the aquifer is *ca.* $10^9\,m^3$ per year. In Denmark, about 35% of the annual water consumption (*ca.* $0.4 \times 10^9\,m^3$) is derived from the chalk, largely in the north and east of the country, whereas south-east England is particularly dependent on chalk aquifers, accounting for 55% of the groundwater utilized in the UK. The uppermost 50–60 m of the saturated chalk forms the principal aquifer since water flow at reasonable rates relies on the presence of open, commonly solution-modified fractures and fissures. The permeability of fissured chalk is 10^{-5}–$10^{-3}\,ms^{-1}$ whereas that of the chalk matrix is in the order of 10^{-9}–$10^{-8}\,ms^{-1}$ and thus has a negligible contribution to the transmissivity of the aquifer. As in hydrocarbon reservoirs, therefore, the matrix porosity provides the volume for storage of groundwater and the fractures provide the distribution system that drains the matrix and allows the water to flow.

See Also

Diagenesis, Overview. **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills. **Mesozoic:** Cretaceous. **Petroleum Geology:** Overview. **Sedimentary Environments:** Reefs ('Build-Ups'). **Sedimentary Rocks:** Chert. **Trace Fossils**.

Further Reading

Bromley RG and Ekdale AA (1986) Composite ichnofabrics and tiering of burrows. *Geological Magazine* 123: 59–65.

Downing RA, Price M, and Jones GP (eds.) (1993) *The Hydrogeology of the Chalk of North West Europe,* p. 300. Oxford: Clarendon Press.

Ekdale AA and Bromley RG (1984) Comparative ichnology of shelf sea and deep sea chalk. *Journal of Paleontology* 58: 322–332.

Ekdale AA and Bromley RG (1991) Analysis of composite ichnofabrics: an example in uppermost Cretaceous chalk of Denmark. *Palaios* 6: 232–249.

Gale AS (1995) Cyclostratigraphy and correlation of the Cenomanian stage in Western Europe. In: House MR and Gale AS (eds.) *Orbital Forcing Timescales and Cyclostratigraphy,* Geological Society Special Publication 85, pp. 177–197. London: Geological Society.

Hancock JM (1976) The petrology of the chalk. *Proceedings of the Geological Association* 86: 499–535.

Hancock JM (1993) The formation and diagenesis of chalk. In: Downing RA, Price M, and Jones GP (eds.) *The Hydrogeology of the Chalk of North West Europe,* pp. 14–34. Oxford: Clarendon Press.

Hay WW (1995) Cretaceous paleoceanography. *Geologica Carpathica* 46: 257–266.

Kennedy WJ (1987) Late Cretaceous and early Palaeocene Chalk Group sedimentation in the Greater Ekofisk Area, North Sea Graben. *Bulletin du Centre Recherche Exploration Production Elf Aquitaine* 11: 91–126.

Kennedy WJ and Garrison RE (1975) Morphology and genesis of nodular chalks and hardgrounds in the Upper Cretaceous of southern England. *Sedimentology* 22: 311–386.

Quine M and Bosence D (1991) Stratal geometries, facies and sea floor erosion in Upper Cretaceous chalk, Normandy, France. *Sedimentology* 38: 1113–1152.

Scholle PA (1977) Chalk diagenesis and its relation to petroleum exploration: oil from chalks, a modern miracle? *American Association of Petroleum Geologists Bulletin* 61: 982–1009.

Surlyk F (1997) A cool water carbonate ramp with bryozoan mounds: Late Cretaceous Danian of the Danish Basin. In: James NP and Clarke JAD (eds.) *Cool Water Carbonates. SEPM (Society for Sedimentary Geology) Special Publication,* 56, pp. 293–307. Tulsa, Oklahoma: SEPM.

Surlyk F, Dons T, Clausen CK, and Higham J (2003) Upper Cretaceous. In: Evans D, Graham C, Armour A, and Bathurst P (eds.) *The Millenium Atlas: Petroleum Geology of the Central and Northern North Sea,* pp. 213–233. London: Geological Society.

Chert

N H Trewin and S R Fayers, University of Aberdeen, Aberdeen, UK

Introduction

The term chert is currently used for any microcrystalline siliceous rock containing only minor impurities. In the early nineteenth century, hornstone and chert were names used for rather non-descript splintery siliceous rocks, but chemically similar material with specific features of colour and texture received a plethora of names, largely based on ornamental value. It is usually the minor impurities that impart colour, such as the red of haematite in jasper, and the green of chrysoprase assigned to nickel. Differing textural features produce the banding seen in agates. Porcellanite is a white variety containing clay inclusions and resembling porcelain, and flint refers to the nodules from the Cretaceous Chalk that produce superb conchoidal fractures, and were certainly one of the earliest geological industrial materials, being used by Palaeolithic man.

Cherts occur in low abundance in a variety of geological settings, and have a variety of origins. In many cases, the chert product is the result of maturation by time and the diagenesis of precursor silica phases. The diagenetic pathway of silica is controlled by phase solubility, a function of crystal structure and size, and usually proceeds from amorphous opal through intermediate stages to quartz by dissolution–reprecipitation reactions. Amorphous silica may have its origin in biogenic skeletal material (radiolaria, diatoms, siliceous sponge spicules), or in volcanic glass, or in sinter deposited from hydrothermal solutions.

Chert Composition

Chert is seldom uniform in texture and, in young material, transitions from opal are frequent. The grain size and texture are related to the origin of the silica and the diagenetic history of the rock.

The Main Constituents of Chert

Microquartz Microquartz is the main constituent of chert. It occurs as equant crystals in a solid mosaic (Figure 1), and often has a porous, spongy texture. Spongy forms are frequently dark in thin section due to included pores. Crystals are generally less than 5–20 μm in diameter. Microcrystalline quartz comprising individual crystals below the resolution of a standard petrographical microscope is termed cryptocrystalline.

Megaquartz Megaquartz occurs as a mosaic quartz fill to cavities and veins in chert, often displaying a drusy fabric (Figure 1A and B). It may also occur as a replacement fabric in carbonate fossils or fossil wood. Crystals are tens to hundreds of micrometres in diameter.

Chalcedony Chalcedony has a fibrous texture, with fibres tens to hundreds of micrometres in length. It occurs in radiating spherulitic textures and has overlays, frequently brown and zoned in thin section. Chalcedonic quartz is length-fast. Quartzine is length-slow. Leutecite is intermediate between chalcedonic quartz and quartzine, with the fibre axis oriented approximately 30° to the crystallographic *c*-axis. Chalcedonic quartz and quartzine both form cement and replacement fabrics (Figure 1C and D), whereas leutecite typically occurs as a replacement fabric in carbonate fossils. Together with quartzine, other fibrous types, namely zebraic chalcedony and microflamboyant quartz, are most common in chert-replaced evaporites.

Opal Opal occurs as amorphous to cryptocrystalline forms. The crystallinity increases from opal-A to opal-CT to opal-C; these phases can be distinguished by X-ray diffraction. Opal-A is amorphous, opal-CT comprises disordered interlayers of cristobalite and tridymite, and opal-C comprises cristobalite.

Silica Solubility and Precipitation

Biogenic opal-A has a solubility of 120–140 ppm in normal marine sediment pore water, cristobalite 25–30 ppm, and quartz 6–10 ppm. The dissolution of opal-A generally results in supersaturation with respect to opal-CT, which is precipitated in preference to quartz. Opal-CT typically consists of small bladed crystals that form spherical lepispheres 5–10 μm in diameter. Quartz precipitation generally takes place from dilute solutions over longer periods of time. The solubility in water of both amorphous silica and quartz increases rapidly above pH 9, and thus a sharp reduction in alkalinity from pH $>$ 9 results in silica precipitation.

The precipitation of silica, particularly around hot springs, occurs due to the evaporation of silica-saturated water. Silica solubility is higher in hot water, and thus cooling results in silica deposition. Boiling is also a trigger for silica precipitation in hydrothermal systems.

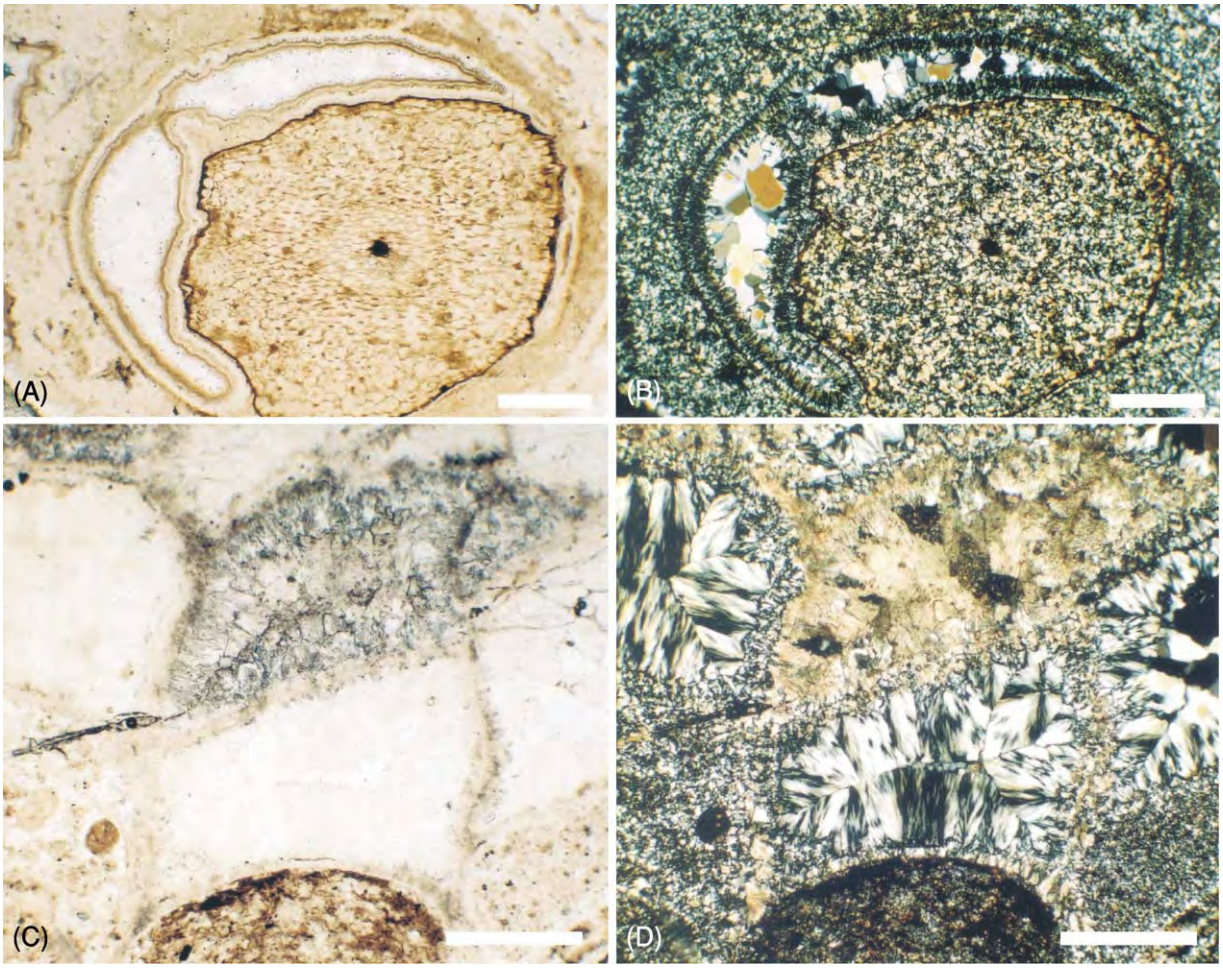


Figure 1 Examples of chert constituents and textures in thin section. (A) Transverse section through a shrunken plant stem (*Rhynia gwynne vaughanii*) in Early Devonian Rhynie Chert (Rhynie, Scotland), observed in PPL (plane polarized light) (plane polarized light). (B) Same view as (A) in XPL (cross polarized light). The plant and surrounding organic rich matrix is preserved in microcrystalline quartz. The void created by the shrunken stem is lined by zoned chalcedonic quartz and occluded by later megaquartz cement. (C) Cross section through dissepiments of the coral *Dibunophyllum* viewed in PPL. Hob's House Coral Bed, Viséan, Carboniferous, Derbyshire, UK. (D) Same view as (C) in XPL. Sediment filled voids are replaced by microcrystalline quartz (below); those above are occluded by chalcedonic quartz cement. One dissepiment is occluded by calcite spar (top, centre). Scale bars 500 μm.

Textures

In many cherts, there is a clear textural sequence in which microquartz, a transformation of original silica gel or opal, is overlain by chalcedonic layers, and the remaining space is filled by megaquartz. The full maturation sequence of opal to quartz is not the only route to chert; amorphous opal can transform directly to quartz and, in diagenetic cherts, microquartz or chalcedony may be the primary crystallization phase.

Sources of Silica

Biogenic

Two groups of common micro-organisms and one group of sponges build skeletons of opal, thus fixing silica from solution. The main groups with their ranges and environments are as follows:

- Radiolaria – Cambrian–Holocene – plankton, marine.
- Diatoms – Jurassic–Holocene – plankton, marine.
- Diatoms – Tertiary–Holocene – planktic–benthic, non-marine.
- Siliceous sponges – Cambrian–Holocene – benthic, dominantly marine.

These organisms can extract silica from water that is undersaturated with respect to silica by one or two orders of magnitude. For example, diatoms have been recorded to reduce the Si content of aquarium water from 0.95 ppm to 0.075 ppm. Thus, these organisms are important in fixing silica in a particulate form that

can be sedimented, and provide a concentration of metastable biogenic opal. This opal can then be converted to chert *in situ* to produce bedded cherts, or be dissolved and transported to a site of diagenetic deposition to produce chert cement and replacement nodules.

Volcanic

There is a strong association between chert and submarine volcanics in the geological record. Thus, it has been postulated that silica is derived from the devitrification of volcanic glass, leading to the production of smectite and silica. However, it is likely that the higher Si contents of water in volcanically active areas result in population explosions ('blooms') of diatoms and radiolaria, and that the silica is fixed by organisms.

Hydrothermal Silica

Hydrothermal systems, developed in the waning phase of volcanicity, produce hot springs and geysers that frequently deposit silica in the form of amorphous siliceous sinter. The silica is dissolved from hot rocks at depth and, as water is circulated through the convecting hydrothermal system, hot silica-saturated water is brought to the surface where silica is deposited due to cooling.

Silica Precipitation in Lakes

The sources of silica resulting in cherts in non-marine lakes are various. In sediment-starved Tertiary to Holocene lakes, diatoms can accumulate to form a siliceous sediment (diatomite) that can be converted to chert through time and diagenesis. Lakes (e.g., Lake Magadi) in volcanically active areas of the African Rift Valley contain sodium carbonate brines with $pH > 10$. Silica is leached from volcanic rocks and Si concentrations can rise to 2500 ppm. Seasonal evaporation and dilution of the brine by river waters causes the deposition of hydrated sodium silicates that are converted to chert during diagenesis. In the Coorong region of South Australia, the pH in some Mg-rich carbonate lakes can rise above pH 10 due to algal photosynthesis. Silica is derived by the corrosion of detrital minerals, resulting in Si supersaturation of the lake waters. Subsequently, the silica is deposited in lake carbonates as a gel, giving the potential for conversion to chert during diagenesis.

Occurrence of Chert

There are a number of modes of occurrence of chert, the most common, and volumetrically the most important, being bedded cherts and nodular cherts in limestone sequences.

- Bedded cherts in ocean basins.
- Nodular cherts in limestone sequences.
- Cherts of hydrothermal origin, both surface and subsurface.
- Cherts in lake basins.
- Silcrete, chert in palaeosols.
- Silicified wood.

Bedded Cherts

Bedded cherts have been formed through the burial and diagenesis of siliceous oozes throughout Phanerozoic time (Figure 2). However, the Palaeozoic is dominated by radiolaria, and diatoms do not make a significant contribution until the Late Mesozoic. There are also extensive bedded cherts in the Precambrian, at a time from which no silica-secreting organisms are known. Thus, it is pertinent to consider the mechanisms and environments of accumulation of siliceous oozes through time.

At the present time, siliceous oozes are accumulating in deep ocean basins, in areas starved of detrital supply. A broad band of siliceous diatom-dominated deposits surrounds Antarctica, and similar deposits are accumulating between North America and Asia in the northern Pacific to the south of the Aleutian Island chain. An equatorial belt of radiolarian-dominated ooze is present in the Pacific and Indian Oceans. Drilling by the Deep Sea Drilling Project has shown that, in some oceanic areas, the siliceous oozes are converted at depth to bedded cherts, the chert generally being of Tertiary age.

The conditions considered to be favourable for the accumulation of siliceous ooze are summarized below (Figure 3).

- High organic productivity in surface waters due to upwelling of nutrient-rich oceanic currents.
- Lack of significant input of land-derived detritus that would dilute the deposit. Such material is carried by the wind from deserts, and by ocean currents from sources of clastic input.
- Limited presence of calcareous plankton. Calcareous oozes are accumulating at present at rates of 10–50 mm Ka^{-1}, compared with <10 mm Ka^{-1} for siliceous oozes. Modern siliceous oozes tend to occur in oceans below the carbonate compensation depth (CCD), the depth at which carbonate dissolution balances the oceanic carbonate rain. The CCD varies in the oceans, but is generally at about 4 km. Siliceous oozes are associated with, and diluted by, red clays derived from aeolian desert dust and volcanic fallout.

Siliceous oozes can also accumulate in shallower oceanic settings if conditions are suitable. Small,

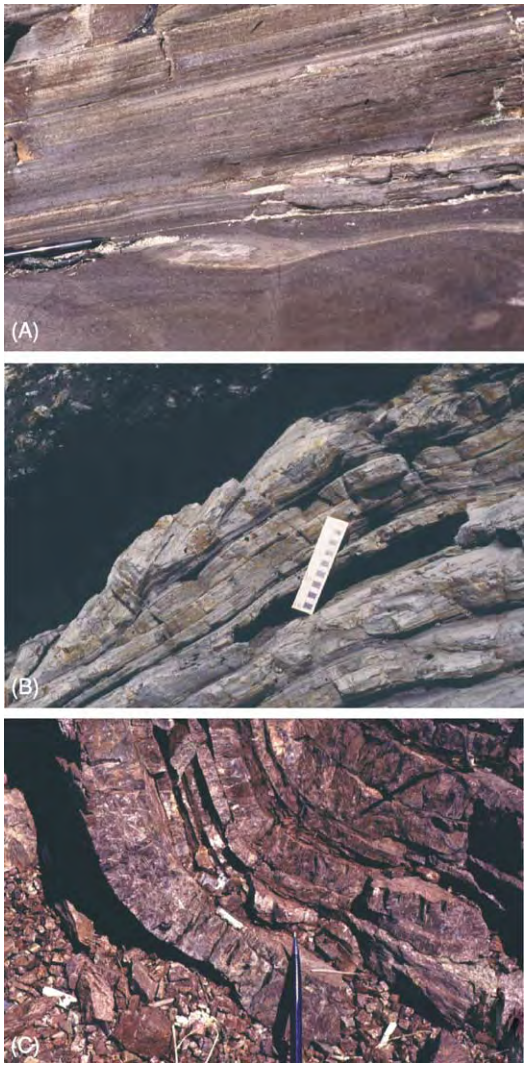


Figure 2 Bedded cherts. Laminated diatomite (A) and laminated diatomite with slump fold (B). Monterey Formation, Miocene, Gaviotta Beach, near Ventura, California, USA. (C) Deformed bedded radiolarian cherts within the Khabarovsky Complex, exposed on the banks of the Amur River in the city of Khabarovsk, Russian Far East. These Upper Triassic Lower Jurassic cherts are associated with black mudstone and metabasalt units of the same age, and are tectonically interleaved with Cretaceous turbidite units. This complex association formed in a Late Mesozoic accretionary complex during collision of the North China Block with Siberia. Photographs courtesy of David Macdonald.

sediment-starved basins and submerged carbonate platforms on passive continental margins are potential sites for deposition when upwelling currents bring nutrients and calcareous plankton is scarce. The Gulf of California (<1.5 km deep) is an area in which siliceous oozes are accumulating in association with distal turbidites and organic-rich shales.

Tertiary bedded chert Tertiary bedded chert derived from diatomaceous ooze occurs in the Miocene–Pliocene of the Pacific margin, and formed in small rifted and back-arc basins where there was strong nutrient supply, high phytoplankton activity, and a lack of detrital sediment input. The Monterey Formation (Miocene, California) contains chert derived from diatomaceous sediments (Figure 2A and B), and is associated with hydrocarbon source rocks. The maturation process from diatomaceous sediment to chert involves porosity loss to the extent that the cherts can form a diagenetic reservoir seal at depth.

In the Mediterranean, cherts occur in the Late Miocene at the time when the area was characterized by small, restricted basins in the build-up to the Messinian salinity crisis at the end of the Miocene. In general, cherts of Tertiary age are widespread in oceanic areas, and occur in a wide variety of settings satisfying the conditions described above.

Mesozoic and Palaeozoic bedded cherts Mesozoic and Palaeozoic bedded cherts can be divided into those that are associated with volcanics, usually of ocean floor origin, and those that have no relationship to volcanicity. Prior to the expansion of planktonic diatoms in the Late Mesozoic, the radiolaria were the main contributors to siliceous ooze. The major calcareous planktonic organisms, coccoliths and planktonic foraminifera, did not appear in abundance until the Late Mesozoic. Thus, it can be postulated that siliceous radiolarian ooze would have been more abundant in the Palaeozoic and Early Mesozoic, and that radiolaria provided the main source for biogenic silica. It is likely that radiolarian ooze accumulated over a greater depth range, extending into shallower water above the current level of the CCD, in the absence of, or at least the reduction of, the diluting effect of calcareous plankton.

Volcanic association There is a common association of bedded cherts with black shales, pillow lavas, and volcaniclastic rocks. In some cases, sheeted dykes and ultramafic rocks are present, and the whole assemblage is typical of the ophiolite suite: a preserved ocean floor succession. The cherts are usually dark in colour, and may contain recognizable ghosts of radiolaria (Figure 4). The presence of radiolaria points to a biogenic origin for the silica, although elevated Si in seawater from the volcanics may have been responsible for the proliferation of radiolaria. Mesozoic examples occur in the Troodos Massif in Cyprus, the Franciscan of California, and the Khabarovsk area of the Russian Far East (Figure 2C). Palaeozoic examples associated with the margin and closure of the Iapetus Ocean occur in the Ordovician of Scotland, from the Girvan–Ballantrae area in the

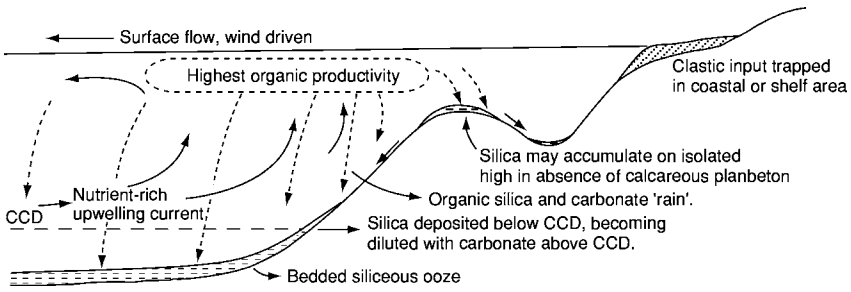


Figure 3 Diagram illustrating the production and deposition of organic silica under the influence of an upwelling nutrient rich current at an ocean margin. CCD, carbonate compensation depth. Not to scale.

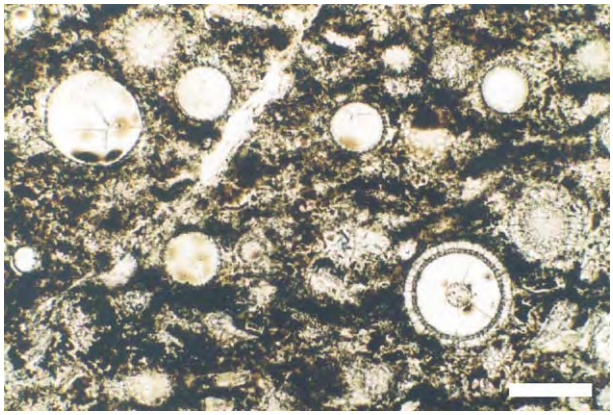


Figure 4 Photomicrograph of radiolarian chert in thin section. Viewed in PPL (plane polarized light). Scale bar 250 μm.

west, through to Stonehaven in the east, where the chert is present as jasper. Similar developments occur in Newfoundland (Canada) and Maine (USA). The jasper in these instances lacks recognizable biogenic material and is frequently attributed to the hydrothermal alteration of sediments and hydrothermal vent deposits. It is not always possible to distinguish biogenic and volcanic/hydrothermal origins.

No volcanic association Many Palaeozoic and Mesozoic cherts have no volcanic association, and occur with black shales, pelagic limestones, and turbidites. Passive continental margins with small rifted basins are a common setting. Water depths were typically less than oceanic. Radiolarian cherts in the Lower Carboniferous of south-west England and Germany provide an example. Submarine rises, starved of land-derived detritus, formed a typical location. Resedimentation of siliceous material from shallow-water areas by turbidity currents provided a concentration mechanism in deeper basinal areas. The Caballos Formation of the Marathon Basin, Texas, USA, is a widespread bedded chert (part termed novaculite). It is of Devonian age and marine in origin, containing sponge spicules and radiolaria. The chert is interbedded with shale, and some sandstone beds and chert conglomerates are present. Arguments have been presented for both shallow marine to restricted lagoonal and deep marine environments.

Precambrian cherts Precambrian cherts are abundant, particularly in association with Precambrian 'iron formations'. However, no silica-secreting organisms are known from the Precambrian and it cannot be assumed that Precambrian cherts have a biogenic origin. There is a concentration of these deposits dated at around 2 Ga, and these cherts contain fossils of coccoid and filamentous bacteria. In the absence of silica-fixing organisms, it is likely that Si concentrations in water were high, and Si precipitation may have occurred initially as a gel in the large iron formation basins. It is possible that photosynthetic cyanobacteria played a significant role in silica deposition as in Coorong-type lakes. Texturally, the presence of algal stromatolites, ooids, and intraclasts implies shallow water, and Si replacement of carbonates. Shallow extensive shelves and large lake basins are postulated environments.

Classic examples at around 2 Ga are the Gunflint Chert of Ontario (Canada) (Figure 5), and the Biwabik Chert of South Africa. These deposits may mark the period in the Earth's history when oceans were changed from reducing to oxygenated conditions by the action of photosynthetic bacteria, and consequently vast amounts of iron were deposited that had previously been held in solution in ocean waters. Even older, at about 3.5 Ga, are the cherty rocks of the north pole region in Western Australia containing silicified stromatolites that are possibly the oldest morphological fossils on Earth (there is older geochemical evidence).

Nodular Cherts

Chert is common as a nodular and, occasionally, 'bedded' replacement feature in limestone sequences,

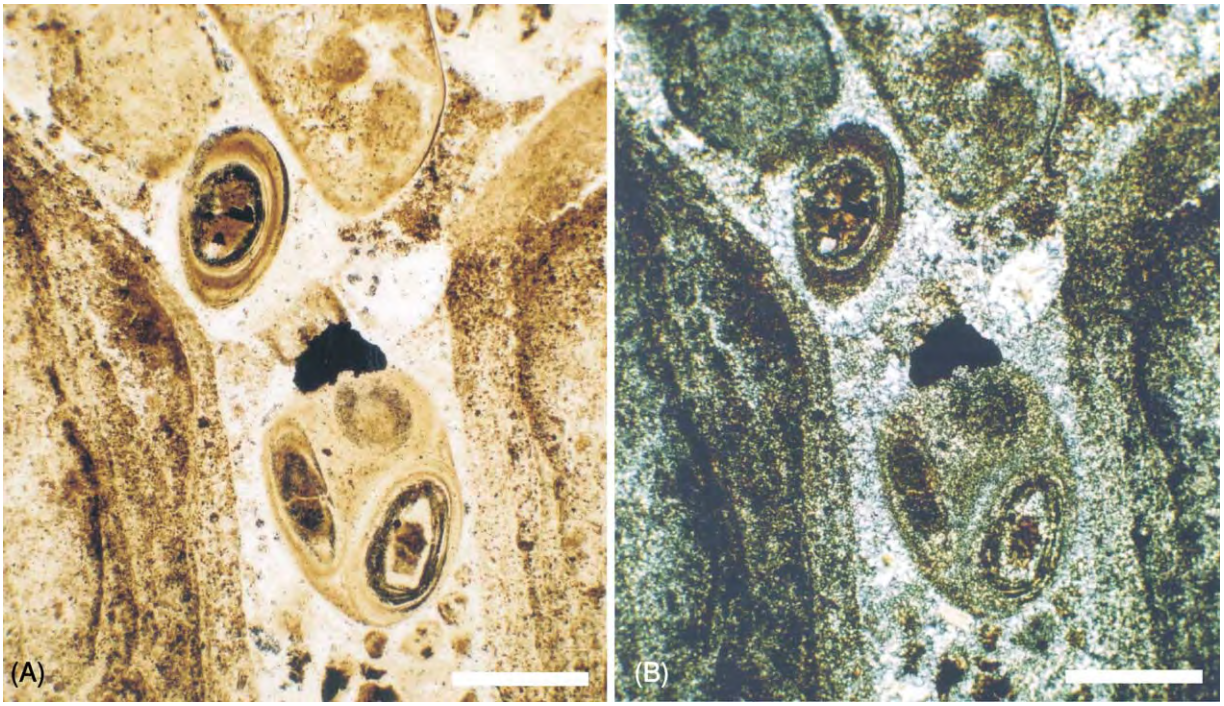


Figure 5 Photomicrographs of Gunflint Chert in thin section showing chert replaced intraclasts and ooliths within a crack between stromatolites. (A) Viewed in PPL (plane polarized light). (B) Viewed in XPL (cross polarized light) showing replacement mainly in the form of microcrystalline quartz. Precambrian, Ontario, Canada. Scale bars 500 μm.

and occurs to a lesser extent in mudstones and evaporites. The apparently bedded cherts in this situation are the result of the replacement of original sedimentary beds by chert (Figure 6). The source of the silica is considered to be biogenic, with the dissolution products of biogenic opal being redistributed in solution and precipitated as cement and replacement during diagenesis. The replacement nodules show a great variety in form, ranging from irregular forms with smooth curved margins (Figure 7), to more tabular and diffuse cherts seen in Carboniferous limestones, and the generally spherical nodules (geodes) representing the replacement of original anhydrite nodules.

Both chalcedony and microquartz are present (Figure 6), and there is frequently evidence for the direct precipitation of quartz in the rock in the form of isolated bipyramidal crystals. The general process of formation of the nodules involves the dissolution of biogenic opal, present in low abundance in the deposited sediment (*ca*. 1%). The mobilized silica is then deposited at suitable nucleation sites, probably as opal-CT. Such sites are controlled by rock texture and biogenic content; hence, silica deposition may favour specific beds. The opal-CT fills the pore space and replaces carbonate, and, with burial, is converted to chert. The silicification appears to be a relatively early diagenetic event, taking place during shallow burial.

In marine phreatic conditions, silica precipitation and replacement of carbonate tend to occur along redox boundaries between aerobic surface sediments and underlying sediments dominated by sulphate-reducing bacteria. The degradation of organic material by sulphate-reducing bacteria releases carboxyl and sulphide ions. Many carbonate sediments contain very little iron; therefore, very little sulphide is precipitated as pyrite. The rest is hydrolysed to hydrogen sulphide which then diffuses to more oxic conditions. Oxidation produces sulphate and hydrogen ions; the former diffuse back into the sulphate reduction zone, whilst the increased acidity causes carbonate dissolution at the redox boundary. The high concentration of carbonate ions, organic matter, and the reduced pH promote silicification.

Early silicification may also take place in emergent areas where marine pore waters in carbonate sediments mix with meteoric pore waters. In these 'mixing zones', the mixing of waters with suitable differences in PCO_2 provides ideal conditions for carbonate dissolution with contemporaneous silica replacement and precipitation.

The concentration of chert replacement nodules at specific horizons, often on a basin-wide scale, can be

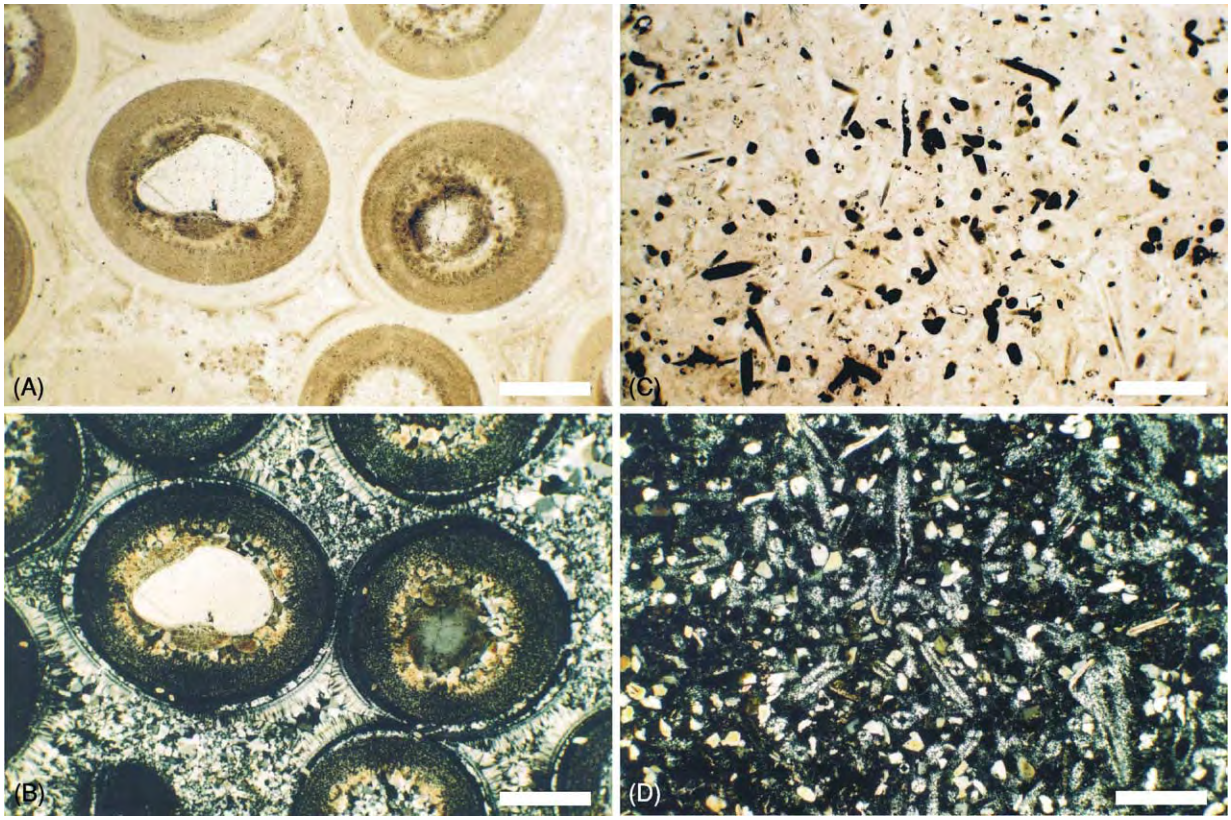


Figure 6 Photomicrographs of chert replaced and cemented sedimentary rocks in thin section. (A) Silicified oolite in PPL (plane polarized light). (B) Same view as (A) in XPL (cross polarized light) showing a variety of silica cement and replacement fabrics after the original carbonate sediment. Trenton Series, Ordovician, Centre County, Pennsylvania, USA. (C) Silicified bioclastic sandstone with abundant sponge spicule tetraxons, viewed in PPL. (D) Same view as (C) in XPL. Both the matrix and sponge spicules are replaced primarily by microcrystalline quartz. Upper Greensand, Cretaceous, Ventnor, Isle of Wight, UK. Scale bars 500 μm.

related to redox boundaries in the original sediment (Figure 8). The spacing of chert bands reflects abrupt, stepwise rises of the redox boundary related to pulses in sedimentation and hiatuses. The geometry of the redox boundary (governed by permeability and porosity contrasts in the sediment) generally determines the chert morphology, accounting for the spectrum of burrow-form and tabular cherts commonly observed in the field.

The replacement by chert can be remarkably selective, with preferential replacement of limestone matrix, or of biogenic debris of a particular original composition, such as high-Mg calcite. Very often the earliest stages of silicification in carbonate sequences occur in shell material in which localized silica precipitation/carbonate dissolution is promoted by the bacterial breakdown of organic matter, particularly conchiolin within the shell matrix. Thus, as well as the nodular form of chert, selective silicification may result in scattered silicified fossils within limestone.

In the Cretaceous Chalk of Europe and the USA, flint nodules of irregular form occur at specific widespread stratigraphical horizons, but also in sheets and pipes that cross-cut bedding. Flint is generally dark grey, and contains carbonate inclusions, particularly of bivalves and echinoderms. Flint nodules have a thin white crust, or patina. Many echinoids from the Upper Chalk are filled with flint, the silica having nucleated within the urchin, but without replacing the shell. Sponges and burrows are also selectively silicified, with the shape of many flints in the chalk reflecting the morphology of *Thalassinoides* burrows in which they nucleated (Figure 9). Siliceous sponge spicules probably provided much of the biogenic silica for the formation of flint.

There are also nodular chert-bearing beds within the Portlandian (Late Jurassic) limestones of southern England (Figures 7 and 8), where bioturbation textures and diagenetic redox boundaries controlled silica precipitation and replacement. Siliceous sponges were the main biogenic silica source.

Figure 7 Nodular cherts replacing bioclastic and spicule wack estones. Cherty Beds, Portland Stone Formation, Upper Jurassic, Isle of Portland, Dorset, UK.

Figure 8 Laterally persistent beds of nodular and tabular chert (dark bands) concentrated within the Cherty Beds (Ch) of the Upper Jurassic Portland Stone Formation. Isle of Portland, Dorset, UK.

The Carboniferous limestones of Europe and the USA contain abundant chert as nodules and as silicified fossils (Figures 1C,D, and 10). The chert is generally black and has a splintery fracture, rather than the conchoidal fracture of flint. Silicification can affect specific beds, such that chert nodules may link up to form a diagenetically bedded chert.

Figure 9 Flint nodules after *Thalassinoides* burrows. Chalk, Upper Cretaceous, UK. Scale bar 25 mm.

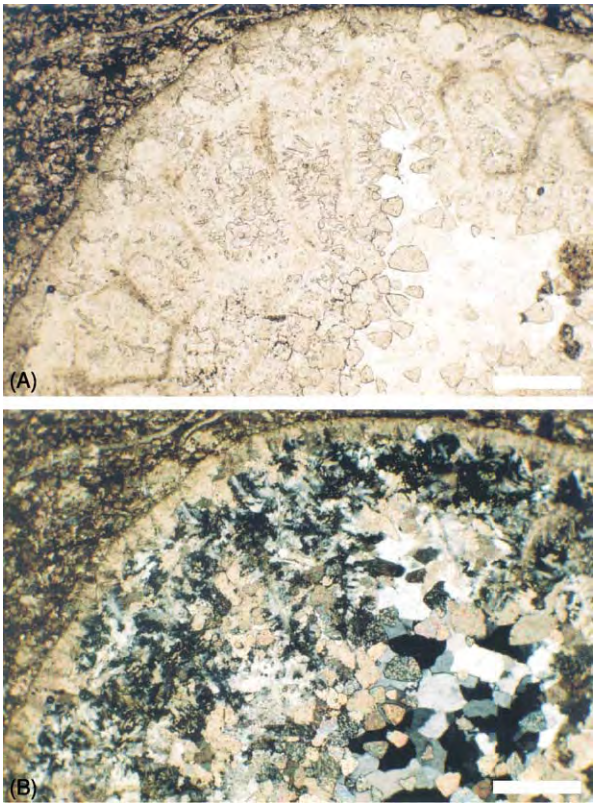


Figure 10 Transverse cross section through a partially silici fied corallite of *Siphonodendron junceum*. (A) Viewed in PPL (plane polarized light). (B) Viewed in XPL (cross polarized light). Hob's House Coral Bed, Viséan, Lower Carboniferous, Derbyshire, UK. Scale bars 500 μm.

Chert in Lakes

Modern examples of lacustrine chert deposition comprise the Lake Magadi type, where the lake is Na-rich and alkaline and has pH > 9 in the dry season, leading to silica dissolution, and a pH that fluctuates below

pH 9 in the wetter months, resulting in silica precipitation. Silica is initially deposited as magadiite (hydrated sodium silicate), which is subsequently replaced by silica. Thus, the controlling factors are evaporation and freshwater input to the lake.

In the Coorong type from South Australia, Mg-rich carbonate lakes acquire a high pH due to the seasonal activity of photosynthetic algae, resulting in the dissolution of silicates; with a seasonal reduction in pH, direct precipitation of mixed opal and cristobalite takes place.

The Cretaceous Uhangri Formation of southwest Korea was deposited in an alkaline lake surrounded by alkaline volcanics. The sequence includes couplets of sandstone overlain by chert, and of laminated chert with black shale. The sandstone/chert couplets were deposited following episodic influxes of fresh, less alkaline water. The influxes carrying sand produced density-current underflows in the stratified lake, depositing sand followed by opaline silica, caused by the fall in pH due to the influx of freshwater. The laminated cherts are interpreted to be the result of interflows causing silica precipitation. The chert beds show soft-sediment deformation and injection features, indicating a gelatinous consistency for the deposited silica. Thus, with regard to the feature of direct silica deposition, this example has similarities with the Coorong type.

Ancient deposits interpreted as belonging to the Magadi type are more common, and range in age from the Precambrian Reitgat Formation, Hartbeesfontain, South Africa, to the present day. In typical examples, there is an association with contemporaneous volcanics, and evidence for evaporite minerals.

Chert of Hydrothermal Origin

Silica-rich fluid expulsion from basins Basin marginal faults are commonly the site of chert deposition as veins and porosity-filling cement. Chert is deposited as a result of the cooling of silica-rich water expelled from the basin and rapidly rising up marginal fault zones. Silica is more soluble at high temperatures, and hence cooling results in silica precipitation. Chert may seal a fault, and subsequent fault movement may result in new fractures, which themselves become sealed; the result is a chert-cemented and veined fault zone.

Cherts resulting from hydrothermal systems Hydrothermal systems associated with volcanic activity are seen today at Geysir in Iceland, Yellowstone National Park in the USA, and North Island, New Zealand. At these, and many other localities, hot springs and geysers deposit large quantities of silica both in the subsurface and at the point of eruption, which may be on land or under water. Silica is deposited from cooling waters that have dissolved silica from hot rocks at depth (**Figure 11**).

In the subsurface, the result is the silicification of country rocks, particularly along fluid pathways such as faults. Cherty rock may develop on a large scale in the subsurface above a hydrothermal system, resulting in chert cement and cherty veins. The silica is initially deposited as amorphous silica, and this matures to chert with time, heat, and burial.

Hot springs and geysers bring hot water to the surface that cools rapidly on eruption, resulting in the instant deposition of amorphous silica in the form of sinter. Sinter may form mounds around geyser vents, or the outflow from a hot spring may result in sinter terraces or a low-angle sinter outwash apron (**Figure 12**). Under water, sinter chimneys may form above vents as occurs in Lake Yellowstone. The silica is deposited as highly porous amorphous opal-A, which is transformed to opal-CT, and later to chert, with a loss of porosity.

In New Zealand, the stages of mineral transformation are well documented. The Umikiri sinter is up to 15 m thick, can be dated to between 27 000 and 200 000 years BP, and shows a preserved silica maturation stratigraphy of opal-CT to opal-C to quartz with depth, all original opal-A having already been converted to opal-CT. Thus, the textural features associated with phase changes and solution–precipitation phenomena occur in a geologically short period of time in near-surface environments.

Probably the best-known fossil hot spring deposit is the Early Devonian Rhynie Chert of north-east Scotland (**Figures 1A,B, and 13**). The chert beds were deposited as sinters on a low-angle run-off apron from hot springs fed along a marginal fault to the Rhynie Basin of Old Red Sandstone. The beds are up to 0.5 m thick, laterally non-persistent, and with interbedded shale and sandstone of an alluvial plain environment. The chert is generally bluish to brown in colour, and is remarkable for the early terrestrial and freshwater biota it contains. The plants in some beds are preserved in three dimensions, with perfect cellular preservation, with plant axes still in the position of growth (**Figure 13**) to a height of 15 cm. This chert has yielded the most diverse terrestrial and freshwater arthropod fauna of any locality of similar age in the world. The detail of preservation is remarkable, including germinating plant spores and even sperm in the process of release from the male fertile organ of a gametophyte plant. Such features require virtually instant preservation, and point to a silica gel as the primary silica deposit. The presence of framboidal pyrite and the preservation of organic matter suggest reducing conditions during silicification. The textures within the Rhynie Chert are closely comparable with

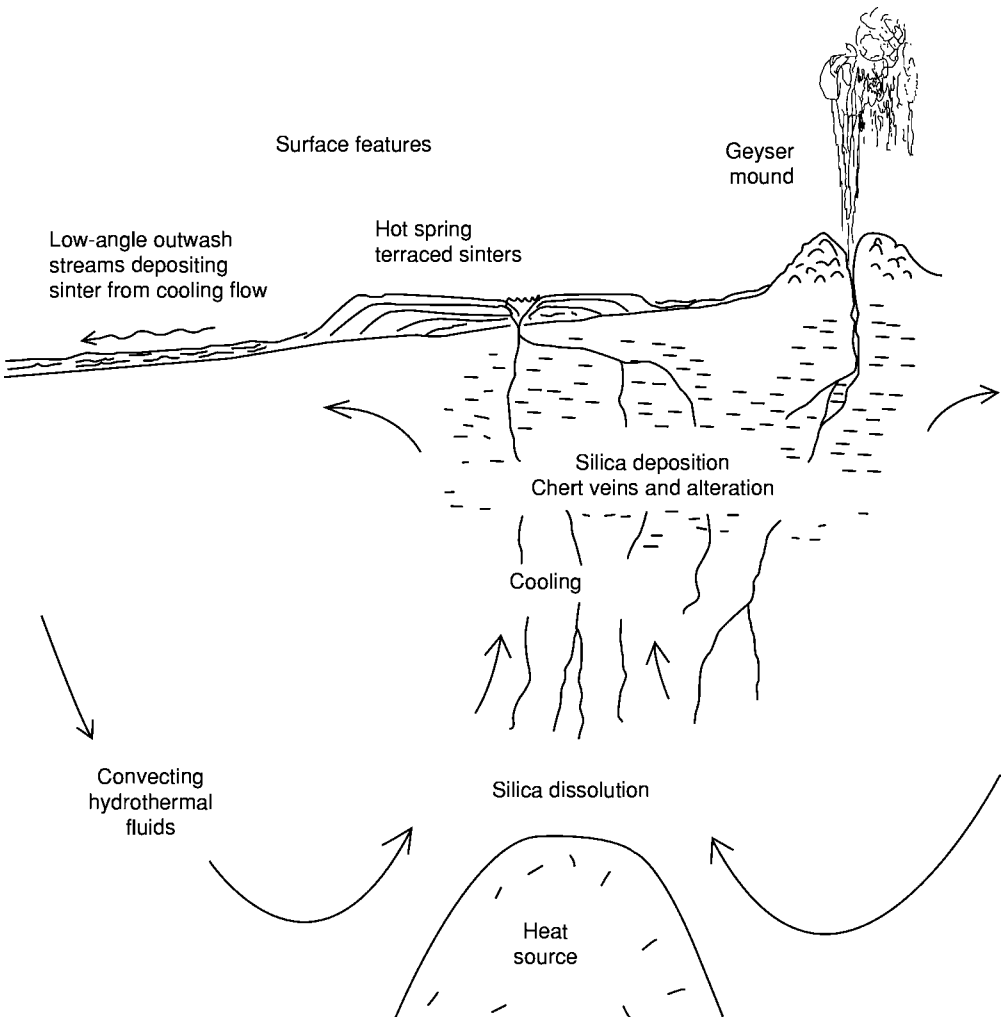


Figure 11 Diagram of a convecting hydrothermal system above an igneous heat source to illustrate surface and subsurface deposition of silica. Not to scale.

Figure 12 Geyserite mound and outlying sinter apron surrounding a small active geyser vent. Shell Spring, Lower Geyser Basin, Yellowstone National Park, Wyoming, USA.

those of modern siliceous sinters, but the maturation process to quartz is complete, and the chert comprises microcrystalline quartz, chalcedony, and macroquartz (Figure 1A and B).

Also found in the same area is a remnant of a geyser vent, with the typical geyserite texture preserved in chert. The country rocks in the area of this ancient hot spring system are also silicified, and a cherty breccia occupies the hot spring feeder zone along the marginal fault. More uncommon are silica deposits resulting from submarine exhalations; examples are the Cretaceous ochres of Cyprus, and the cherty ironstones of Tynagh, Ireland.

Agates with concentric and layered textures of microcrystalline quartz, chalcedony, and megaquartz

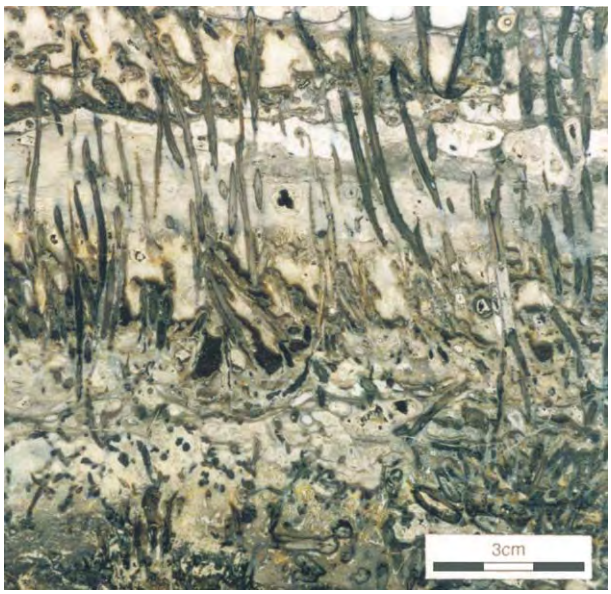


Figure 13 Vertical section through a bed of Rhynie Chert showing abundant stems of the plant *Rhynia gwynne vaughanii* preserved in a three dimensional, upright, growth position. Vague horizontal laminae, draped between the plant axes, represent silicified microbial mats. Chert filled fenestrae within the laminae and between the plants represent gas bubbles trapped within the microbial layers. Pragian, Early Devonian, Rhynie, Scotland.

are mainly found in gas cavities (amygdales) in lavas. The silica is sourced from the volcanic rock, and deposited from hydrothermal and later diagenetic solutions migrating through the rock.

Chert in Soil Profiles – Silcrete

Silica is precipitated in soil profiles forming a hard chert-cemented rock known as silcrete. Most silcretes form in arid to semiarid climatic regions in which silica-saturated, alkaline groundwater with $pH \geq 9$ is evaporated from the surface, or mixes with surface water of lower pH. The presence of iron, aluminium, and magnesium oxides, and also NaCl, appears to encourage silica precipitation. The microquartz occurs as a mosaic cementing any detrital material present. Some replacement by chert is usual, particularly affecting carbonate and micaceous minerals. Silcretes are present in parts of Australia and in both northern and southern Africa. Silica is present in the form of microcrystalline quartz, chalcedony, and, to a lesser extent, opal. Silcrete horizons up to 10 m thick occur; the degree of silicification decreases with depth, with isolated silcrete nodules in the lower part of the profile, and massive silcrete at the top.

The 'Hertfordshire Puddingstone' of southern England is a Tertiary silcrete containing rounded pebbles of flint derived from the chalk, and cemented with chert. It closely resembles examples from the Lake Eyre region of Australia. Silcrete has also been described from the Proterozoic of north-west Canada, where it formed from the weathering of acid volcanics.

Silicified Wood

At many localities throughout the world, fossil wood is preserved in microcrystalline quartz with excellent preservation of the cellular structure of plant tissue. Woody material is a favoured site for silica deposition; in some cases, organic cell walls are preserved; in others, all organic material is lost. The Eocene fossil forests of Yellowstone National Park comprise a succession of 27 forests that were buried by volcanic ash, and occur in a 400 m thick sequence. The silicified trees are preserved as upright stumps several metres high. The silica source was the volcanic ash. In contrast, the Petrified Forest in the Painted Desert region of Arizona represents logs transported to the depositional site, where they occur in alluvial mudstones of Late Triassic age. Silica was probably derived from migrating groundwater, and nucleated in the acidic environment of the decaying wood structure.

Acknowledgments

SRF acknowledges funding from the Lyon Bequest to the University of Aberdeen. We would like to thank Professor David Macdonald (University of Aberdeen) for providing the images and caption information for Figure 2.

See Also

Geysers and Hot Springs. **Minerals:** Quartz. **Sedimentary Environments:** Lake Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification; Deep Ocean Pelagic Oozes; Evaporites. **Tectonics:** Hydrothermal Activity.

Further Reading

Carson GA (1991) Silicification of fossils. In: Allison PA and Briggs DEG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record*, pp. 25 70. New York: Plenum Press.

Hesse R (1989) Silica diagenesis: origin of inorganic and replacement cherts. *Earth Science Reviews* 26: 253 284.

Knauth LP (1979) A model for the origin of chert in limestone. *Geology* 7: 274 277.

McBride EF (compiler) (1979) *Silica in Sediments: Nodular and Bedded Chert. Society of Economic Palaeontologists and Mineralogists, Reprint Series No. 8*. Tulsa, OK: Society of Economic Palaeontologists and Mineralogists.

Sieveking G de G and Hart MB (eds.) (1986) *The Scientific Study of Flint and Chert*. Cambridge: Cambridge University Press.

Trewin NH, Fayers SR, and Anderson LI (2002) *The Rhynie Chert: A Web Based Teaching and Learning Resource*. http://www.abdn.ac.uk/rhynie.

Tucker ME (1991) *Sedimentary Petrology: An Introduction to the Origin of Sedimentary Rocks*. London: Blackwell Scientific Publications.

Williams LA and Crerar DA (1985) Silica diagenesis, II. General mechanisms. *Journal of Sedimentary Petrology* 55: 312 321.

Williams LA, Parks GA, and Crerar DA (1985) Silica diagenesis, I. Solubility controls. *Journal of Sedimentary Petrology* 55: 301 311.

Clays and Their Diagenesis

J M Huggett, Petroclays, Ashtead, UK and The Natural History Museum, London, UK

Introduction

Clay diagenesis is the process of clay transformation (layer by layer replacement) and authigenesis (or neoformation) in buried sediments. Diagenesis commences with the onset of burial and ends with the onset of metamorphism. These boundaries are defined in a variety of ways, including clay mineral crystallinity. Although, in the nineteenth century, microscopists were able to observe euhedral stacks of kaolinite platelets in sandstones, it was not until the early 1970s that it became apparent just how widespread and significant are clays that form after burial. Until this time, geologists had argued over 'the greywacke problem', i.e., how was it possible that clay, and sometimes quite high proportions of clay, could be present in sandstones deposited in high-energy environments? The realization that many sandstones contain authigenic clays came about largely as a result of the arrival of the scanning electron microscope in geological research. This allowed the examination of rocks in three dimensions at magnifications (typically $<1000\times$) ideal for imaging clay particles. Clay diagenesis is not restricted to sandstones; similar processes occur in mudrocks, although transformation reactions and reaction pathways may differ. At around the same time, X-ray diffraction (XRD) was being used to measure changes with depth in clay and non-clay mineralogy with depth in mudrocks, and it was through the pioneering work of John Hower that this was shown to be due to diagenesis (Figure 1). This difference in approach resulted in some rather different perceptions regarding the nature of diagenesis in sandstones and mudrocks.

Factors controlling clay diagenesis include the detrital sediment composition, environment of deposition, temperature, permeability, and burial history (rate of burial, overpressuring, faulting, uplift). Diagenesis typically involves a simplification of the mineralogical suite of a sedimentary rock unit. As the temperature and pressure increase, so too does the tendency towards an equilibrium assemblage, provided that the pore fluid remains aqueous. Most diagenesis occurs at less than 160°C (although the cut-off is not determined by temperature; it is more likely to be determined by the exhaustion of reactive minerals, loss of permeability, or the cessation of movement of aqueous fluids for other reasons).

The movement of ions in solution from argillaceous sediment to coarser sediment has been widely invoked to account for the lack of an obvious internal source for the authigenic minerals present. This has partly come about through a lack of petrographical studies of mudrocks. However, it is now apparent that interbedded mudrock/sandstone can contain the same authigenic minerals, but in very different proportions. This is because, although the same or similar detrital minerals are present in both lithologies, their proportions can be quite different. In argillaceous rocks, there may be more organic matter diagenesis driving particular reactions, and the higher proportions of clay will result in much lower fluid flow rates and water/rock ratios than exist in sandstones. In Tertiary sediments from the North Sea, qualitative mass balance calculations have demonstrated that cross-lithology (sandstone/mudrock) flow can be insignificant for either clay or quartz diagenesis. Backscattered scanning electron microscopy (SEM) provides unequivocal evidence for coarsely crystalline authigenic clays in mudrocks, as well as in sandstones, and with the high resolution and magnifications possible by field emission SEM, it is possible to image overgrowths on clays and small packets of authigenic clay enclosed in detrital clay.

Clay Diagenesis in Mudrocks

Detailed investigations of diagenesis in argillaceous sediments in a wide variety of sedimentary basins have shown some consistent patterns of clay diagenesis. Variations in diagenetic clay assemblages result from differences in detrital assemblages and burial history. The bulk composition of most mudrock results in illite as the predominant end product of diagenesis. For

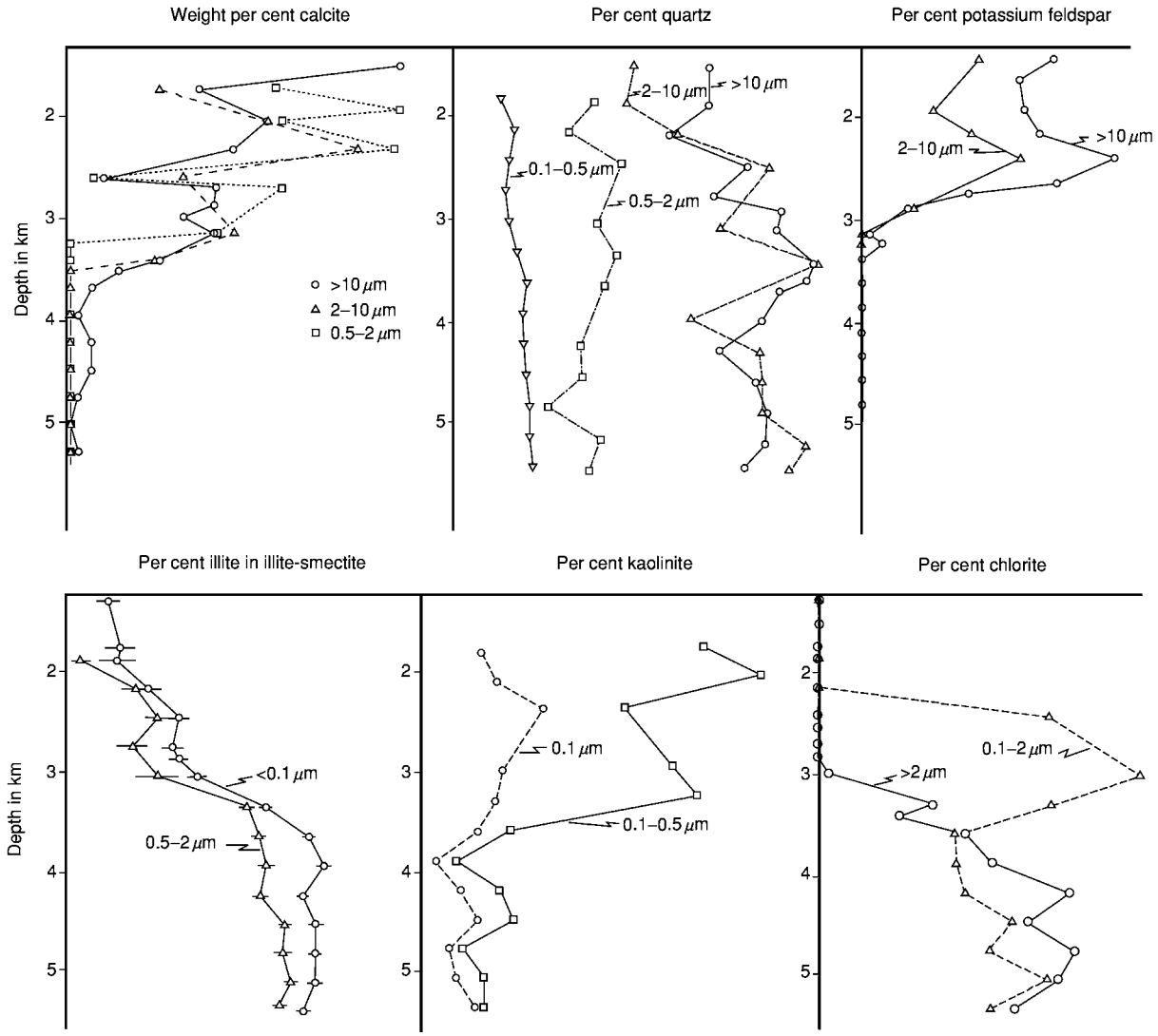


Figure 1 Change in percentage of calcite, quartz, K feldspar, illite, kaolinite, and chlorite with depth. Adapted from Hower J, Eslinger EV, Hower ME, and Perry EA (1976) Mechanism of burial metamorphism of argillaceous sediments: 1: Mineralogical and chemical evidence. *Geological Society of America Bulletin* 87: 725 737.

kaolinite or chlorite to be the predominant clay, special conditions are required, whilst the preservation of smectite at depth requires the inhibition of diagenesis, usually through overpressuring.

Illitization of Smectite

The most studied and first recognized aspect of clay diagenesis is the illitization of smectite with increasing depth. This has been most intensively studied in the Gulf Coast region of the USA, where Tertiary smectite-rich argillaceous sediments have undergone progressive burial. With increasing depth (i.e., increasing temperature), K feldspar, kaolinite, and smectite decrease, whilst illite and chlorite increase. In its simplest form, this reaction can be written as the dissolution of K feldspar to yield Al^{3+} and K^+, which react with smectite to form illite–smectite and ultimately illite:

$$\text{smectite} + Al^{3+} + K^+ \rightarrow \text{illite} + Si^{4+} + Fe^{2+} + Na^+ + Mg^{2+}$$

or

$$\text{smectite} + K^+ \rightarrow \text{illite} + Si^{4+} + Fe^{2+} + Na^+ + Mg^{2+}$$

The second reaction conserves aluminium and requires the dissolution of some smectite. The rock evidence suggests that both reactions are possible. K feldspar may also react with kaolinite to form illite. It is widely claimed that the K^+ for these reactions is a

result of acid dissolution, especially by organic acids. However, feldspar dissolution can occur at low, neutral, or high pH. Indeed, the rate of feldspar dissolution is kinetically controlled. Hence feldspar (and mica) dissolution increases with increasing temperature, and therefore depth. Moreover, organic acids have low buffering capacities and therefore do not influence pH greatly. It should be noted that the reaction results in the release of silica and, through smectite dehydration, the release of water. The source of aluminium is probably mostly feldspar, but the means of transporting sufficient dissolved aluminium to the reaction site has not been entirely resolved, as the aluminium solubility varies enormously with pH, but in most geological situations is rather low. Carboxylic acids are claimed to have the capacity to complex with aluminium, thereby increasing the amount that can be held in solution. However, such acids are unlikely to have much effect on aluminium solubility in complex (i.e., natural) systems. It should also be noted that the reaction yields a potential source of quartz cement. It is largely agreed that the reaction is kinetically controlled, although, if the proposed kinetic equations are applied to the older sedimentary basins, the amount of illitization is vastly overestimated; this may be because the total heating to which they have been exposed has been overestimated. How close K feldspar needs to be to the site of illitization probably depends on the fluid flow, the degree of sandstone/mudrock interbedding, and the overpressure. In the Mahakam Delta Basin in Indonesia, it has been shown that the K feldspar alteration in both sandstone and mudrock is restricted to the upper 2 km of sediment, whereas illitization occurs at greater depths, thus necessitating an open system for K^+ transfer at depth. In contrast, in the Gulf Coast and the Tertiary mudrock/sandstones of the North Sea, diagenesis may be a nearly closed system.

The illitization of smectite commences at approximately 70°C, and peaks at approximately 120–130°C. However, in sedimentary basins with high geothermal gradients, this will occur at shallower depths than in those with lower geothermal gradients. Time, overpressure, pore fluid composition, and hydrothermal activity are also important factors in clay burial diagenesis. In general, there is sufficient K feldspar and mica for this not to be an inhibiting factor; clay diagenesis in clay-rich basins is most likely to be inhibited by overpressure which restricts fluid movement. If a source of K^+ is lacking, illite will not be formed, except where intense organic diagenesis releases NH_4, which is able to form ammonium illite. The illitization of smectite in mudrocks proceeds via random, mixed-layer, smectite-rich, illite–smectite to ordered, illite-rich, illite–smectite. Ordering commences at about 35% expandable layers. The maximum illite content is typically 80%. This sequence has been recognized in a wide variety of settings. The analysis of the expandability and thermal histories of basins ranging in age from Precambrian to Quaternary has indicated that the composition of illite–smectite in mudrocks is primarily controlled by the maximum palaeotemperature. Hence, illite–smectite may be used as a geothermometer for mudrocks, although, as pore fluid overpressure and a lack of K^+ may occasionally be more than minor controls on the illitization process, this should always be performed with caution. The interpretation of mixed-layer illite–smectite in terms of fundamental particles and interparticle diffraction, in the early 1980s, triggered much research into the true nature of illite–smectite. The concept of interparticle diffraction implied that, during illitization, mixed-layer crystals were not two chemically distinct clay minerals, but single illite layers 10 Å thick (Figure 2). When these fundamental particles were analysed by XRD, diffraction between particles created the illusion of smectite interlayers. With increasing diagenetic maturity, these particles grow in three dimensions and the apparent smectite layers decrease. This interpretation is not intended to imply that smectite does not exist! The mechanism implies the dissolution of smectite and the precipitation of illite, which initially exists as fundamental particles. However, it is still argued from transmission electron microscopy (TEM) data that layer-by-layer replacement of smectite by illite occurs. High-resolution investigations of the illitization of smectite have shown coherently scattering domains of interstratified illite-expandable layers (Figure 3), expandable layers within coherently scattering domains of illite, and domains of illite within

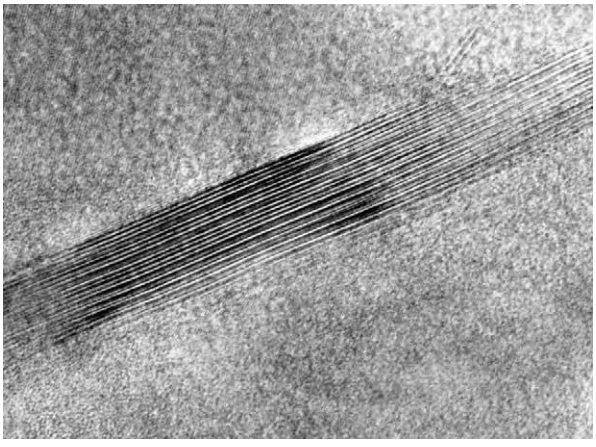


Figure 2 Lattice fringe transmission electron microscopy (TEM) image of interstratified illite expandable clay (the expandable layers have been fixed to prevent collapse in the electron beam).

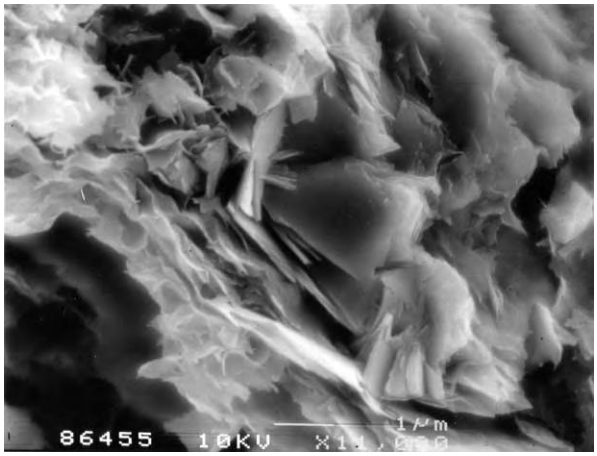


Figure 3 Field emission scanning electron microscopy (SEM) image of authigenic chlorite enclosed by detrital clay in mudrock.

smectite. It is also probable that both dissolution/precipitation and layer transformation occur in different settings. In both bentonites and mudrocks, most smectite is dioctahedral, which appears to be more prone than trioctahedral smectite to illitization. The latter is more likely to react to form chlorite via interstratified chlorite/smectite minerals.

Illitization in Bentonites

Bentonites are ash falls that have undergone extensive devitrification to dioctahedral smectite (usually montmorillonite). Because they have a very simple mineralogical assemblage (most mudrocks contain not only more than one clay type, but a mixture of smectites and illite–smectites), and are often almost monomineralic, ancient bentonites have been extensively used to study the process of illitization of smectite. Comparison of different bentonites, or single bentonites which have undergone variable heating during burial, shows that Si^{4+}, Ca^{2+}, and Na^{+} are lost from the bed and K^{+} is gained as the smectite is illitized. It should be noted that the supply of K^{+} is the rate-limiting step in the illitization of most bentonites because they are K^{+} deficient. Thus, the most potassic (illitized) portions of many bentonite beds are frequently the margins. Where the enclosing sediment is limestone, illitization will be restricted to any K^{+} present within the bentonite bed.

Illite Crystallinity and Illite Sharpness Ratio

With increasing temperature, illite in mudrocks undergoes an increase in crystallinity, as measured by the 001 reflection sharpness on XRD traces. This is due to the loss of smectite layers, increased particle size, and a reduction in abundance of crystal lattice defects. This property has been widely used as an indicator of diagenetic grade, and the results may be correlated with vitrinite reflectance. The Kubler crystallinity index is a measure of the width at half height of the glycol-solvated illite 001 reflection. The Weaver sharpness ratio is the ratio of the illite 001 (10 Å) reflection to the height of the low-angle side of the reflection at 10.5 Å, also on the glycol-solvated trace. The validity of using illite crystallinity or the sharpness ratio has been much debated, but the recent finding that the thickness of fundamental illite particles follows a unique evolution has permitted the refinement of illite crystallinity into a precise measurement of mean crystal thickness. It should be noted that the inclusion of detrital metamorphic mica in the analysis will result in an overestimation of crystallinity.

Chlorite

Si^{4+}, Al^{3+}, Fe^{2+}, and Mg^{2+}, released from the dissolution of smectite and kaolinite, may react to form chlorite. This is usually detected as a down-hole increase in chlorite, although it can be argued that mineralogical changes may also result from a shift in provenance or climate. To obtain unequivocal evidence of authigenesis, it is usually necessary to use an imaging technique, such as SEM or TEM analysis, to demonstrate the face-to-edge arrangement of euhedral platelets. The chlorite shown in Figure 4 was investigated because XRD patterns for a Tertiary mudrock sequence showed unusually high chlorite concentrations at around 2.6 km burial depth. A sample from the same depth, analysed by TEM, confirmed the presence of 14 Å and 7 Å lattice fringes, and X-ray analyses confirmed that it was an iron-rich chlorite. Authigenic chlorite in mudrocks can also form by the replacement of biotite; commonly, replacement is partial, resulting in chlorite–biotite 'stacks'.

Kaolinite

Authigenic kaolinite in mudrocks is much more abundant than was previously thought before backscattered electron imaging made mudrock petrography a real possibility. Previously, a high kaolinite content in an argillaceous rock was assumed to indicate that the clay was formed through tropical weathering, and that it was consequently a climatic indicator. Kaolinite in mudrocks typically replaces muscovite and phengite mica, and cements microfossil cavities. Replacement of detrital mica by kaolinite is a hydrolysis reaction that releases K^{+}. It characteristically occurs during early diagenesis, whilst the

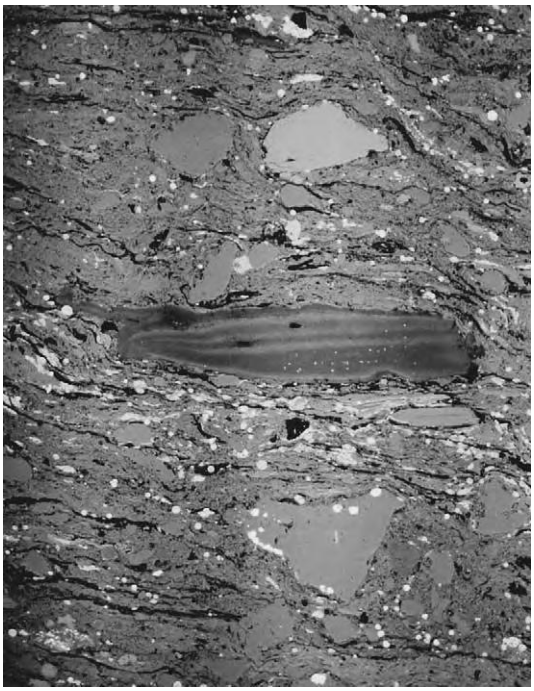


Figure 4 Backscattered scanning electron microscopy (SEM) image of kaolinite pseudomorph after mica in the London Clay Formation (Eocene).

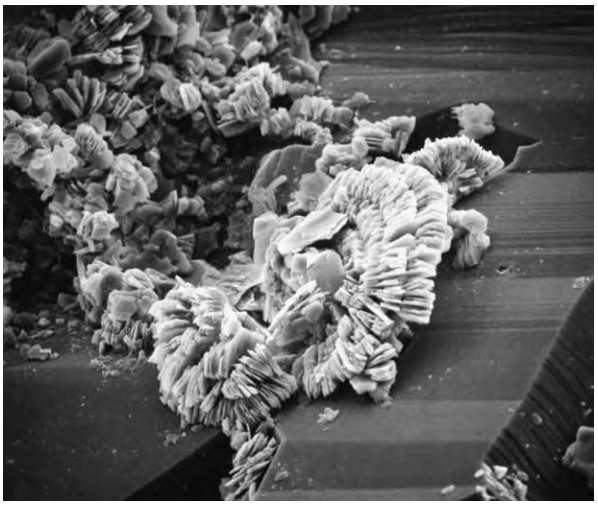


Figure 5 Fracture surface scanning electron microscopy (SEM) image of pore filling kaolinite.

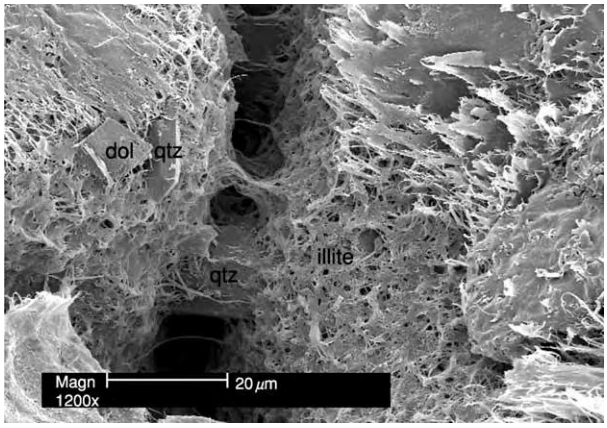


Figure 6 Fracture surface scanning electron microscopy (SEM) image of wispy illite in Rotliegendes Sandstone (Permian). The clay has constricted the pore throat.

fluid/rock ratio is still high, and the K^+ can be removed from the site of dissolution, allowing the reaction to continue. A tentative reaction is deduced from the common association of kaolinite pseudomorphs after mica with mica/quartz pressure solution contacts:

$$\text{muscovite} + H_4SiO_4 + K^+ + H^+ \rightarrow \text{kaolinite} + \text{K feldspar} + Na^+ + Mg^{2+} + Fe^{2+} + H_2O$$

However, in the Eocene of the London Basin, where burial has never been greater than 1 km, probably much less, kaolinite pseudomorphs after mica are widespread (Figure 5), and evidence for pressure solution is minimal. This particular reaction mechanism may be more applicable to sandstones than mudrocks. The fate of the K^+ that is released at shallow depths is not clear; certainly, it is not normally needed for the illitization of smectite at such low temperatures.

Clay Diagenesis in Sandstones

Kaolin Clays

Authigenic kaolinite in sandstones forms stacks of euhedral pseudohexagonal platelets, with the *c* axis parallel to the stacking direction (Figure 6). Very long stacks are called vermicules (Figure 6).Typically, this clay has a pore-filling habit. Kaolinite also forms pseudomorphs after detrital mica, usually muscovite and chlorite. It should be noted that this kaolinite forms particles far larger than the 2 μm maximum defined for clay particles. Pore-filling authigenic kaolinite is often linked to the dissolution of feldspar at temperatures of less than 100°C. Although it is commonly assumed that a high aH^+ is necessary for significant feldspar and mica leaching, the stability field of kaolinite extends to the greatest range of $[K^+]/[H^+]$ values at a pH close to neutral. It should be noted that, for significant kaolinite to precipitate, K^+ and Na^+ from feldspar and mica need to be removed, or illite rather than kaolinite will become the stable clay mineral. The implication is that kaolinite will form at higher fluid flow rates (or a higher water/rock ratio) than illite, i.e., in the most porous parts

of a sandstone, in the most coarse-grained beds, and at a time when a sandstone is less cemented. This is likely to be an important reason why kaolinite almost invariably precipitates before illite in a paragenetic sequence, and why it is often interpreted as a product of meteoric flushing. Meteoric flushing can be an effective mechanism for the replacement of feldspar by kaolinite if the water is sufficiently acidic, which effectively restricts the mechanism to the tropics. Early authigenic kaolinite in sandstones is most abundant in nearshore sediments, because these are most prone to meteoric flushing. Late (or relatively late) authigenic kaolinite is often associated with unconformities, and uplifted fault blocks. A meteoric origin for authigenic kaolinite can be demonstrated through $\delta^{18}O$ stable isotope measurements.

Dickite in sandstones typically forms larger crystals than kaolinite, which are thicker in the direction of the *c* axis. Dickite has been observed to replace kaolinite in reservoir sandstones at approximately 120°C (typically at 2500–4000 m burial depth), most notably where a high water/rock ratio has been preserved. The scarcity of dickite relative to kaolinite is probably due to the rarity of a high water/rock ratio in deeply buried sediment.

Smectite

Authigenic smectite in sandstones is fairly uncommon. Most smectite forms in surface sediments or through the alteration of ash layers. Where it does occur, it characteristically forms early diagenetic rims of crenulate, 'honeycomb', interlocking crystallites. The fusing of adjacent crystallites and the undulose morphology serve to distinguish smectite rims from chlorite rims in SEM images. It should be noted that SEM qualitative X-ray analyses are not always sufficiently accurate to distinguish between smectite and Mg chlorite. Smectite formation is associated with the dissolution of acid volcanic rock fragments and biogenic silica, because smectite formation is favoured by a high Si^{4+} activity. During burial, dissolved Si^{4+} is removed by quartz precipitation (greater than approximately 65°C), making smectite unstable.

Illite

Illite in sandstones has a range of morphologies, from undulose platelets ('cornflake' texture) at one end of the spectrum to laths, fibres, wisps, or ribbons ('hairy' illite) at the other. In fact, the elongate form is not fibrous, but sometimes the particles are so long that they appear so. Wispy illite particles are typically only a few 100 Å thick and 0.1–0.4 μm wide. Their length varies enormously, but can be tens of micrometres. There is no general agreement on any relationship between authigenic illite morphology, the timing of precipitation, chemistry, or structure (although this does not preclude the existence of such relationships). Certainly, with increasing burial depth, the morphology becomes increasingly that of rigid laths, i.e., the width and thickness increase (reflected in an increase in illite crystallinity). Wispy illite is bad news for hydrocarbon reservoir quality. Its high surface area and pore-throat constricting habit (Figure 7) can

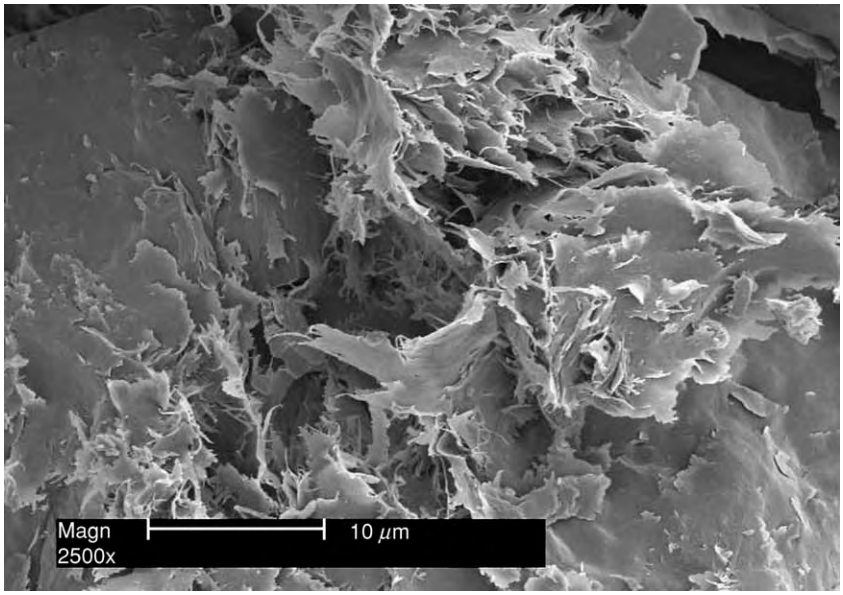


Figure 7 Fracture surface scanning electron microscopy (SEM) image of illite lath overgrowths on infiltrated illitic clay in Rotliegendes Sandstone (Permian).

drastically reduce the permeability (without causing much porosity reduction, because the actual volume of illite is small in proportion to the pore volume). Furthermore, injection wells may suffer a loss of permeability if the injected water breaks up the illite, which then migrates to pore-throats. Not uncommonly, authigenic illite forms short wispy overgrowths on platey detrital illite or illite–smectite, whilst, in some non-marine settings, authigenic illite preferentially nucleates on infiltrated illitic clay (Figure 8). With increasing temperature, pH, and $[K^+]/[H^+]$, kaolin and smectite minerals become unstable and are replaced by illite (Figure 9). In simple terms, the reactions may be shown as:

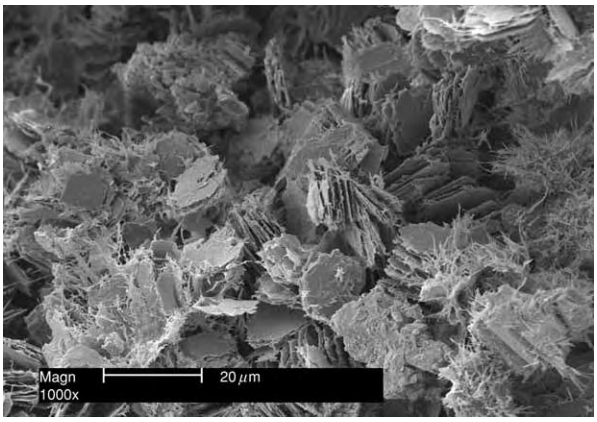


Figure 8 Fracture surface scanning electron microscopy (SEM) image of illite pseudomorph of kaolinite.

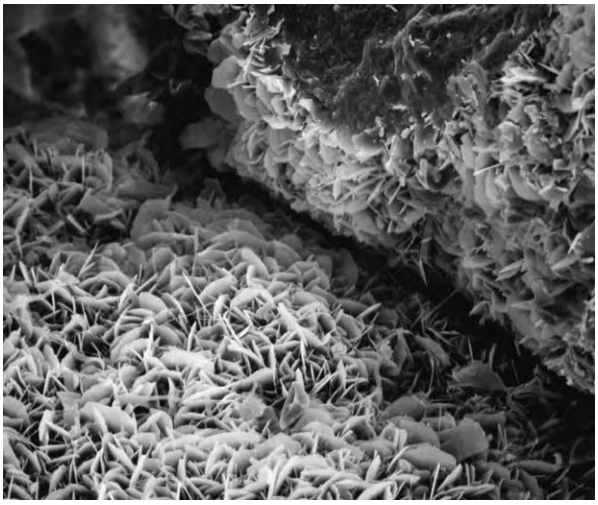


Figure 9 Fracture surface scanning electron microscopy (SEM) image of grain rimming Fe chlorite.

$$\text{smectite} + K^+ \rightarrow \text{illite} + \text{silica (approximately } 80 - 100^\circ\text{C)}$$

$$\text{kaolinite} + \text{K feldspar} \rightarrow \text{illite} + \text{quartz (greater than } 120^\circ\text{C)}$$

This reaction involves an increase in layer charge to accommodate the K^+ in place of the more weakly bonded exchangeable cations in the smectite. This is achieved through substitution of Al^{3+} for Si^{4+} in tetrahedral sites and a reduction in octahedral iron. In fact, many sandstones have pore fluid in thermodynamic equilibrium with illite, but little precipitation occurs due to an extremely low kinetic precipitation rate at temperatures of less than 120°C. Kaolinite is illitized at burial depths in the region of 3–3.5 km, whilst, at greater depths, the thermodynamically more stable dickite is also replaced by illite. Sometimes the mass balance (based on petrographical data) shows that sufficient K^+ can be obtained from locally dissolved K feldspar; in other instances, insufficient K feldspar dissolution has occurred (K feldspar may even be absent) to account for all the illite present, and external sources need to be invoked. It should be noted that the illitization of kaolin minerals is not inevitable above 120°C; there are many instances in which kaolinite coexists with authigenic illite and unleached K feldspar at burial depths ranging from 3 to 4 km. This is because significant illite precipitation requires a higher $[K^+]/[H^+]$ value than that which can occur in a closed system, in which $[K^+]/[H^+]$ is controlled by K feldspar solubility. Hence, the fact that sufficient K^+ can be locally derived does not mean that it is. In the North Sea, authigenic illite abundance is often highest close to major faults, suggesting that the K^+ is derived from the dissolution of K feldspar in deeper parts of the basin. In parts of the North Sea Basin, extensive illite cementation of the Rotliegend Sandstone has been linked to the dissolution of Zechstein salt deposits. Hence, in sandstones, the degree of illitization may reflect the temperature of migrating fluid rather than, as is the case with mudrocks, the maximum burial temperature. Consequently, illite geothermometry is not so reliable a tool for sandstones as it is for mudrocks. The dissolution of K feldspar releases more silica than is required for either kaolinite or illite precipitation. This is thought to be one of the main sources of quartz cement in sandstones. Indeed, textural relationships between kaolinite, illite, and quartz are frequently suggestive of the coprecipitation of one or other clay with quartz cement.

Table 1 Differences between Fe chlorite and Mg chlorite. Adapted from Hillier (1984)

	Fe chlorite	*Mg chlorite*
Morphology	Pseudohexagonal, planar	Cornflake like
Arrangement	Individual plates and rosettes, face to edge contacts	Boxwork, face to face and edge to edge contacts
XRD	Interstratified with 7 Å layers at <200°C	Often interstratified with corrensite or smectite
Polytype	Ib transforming to IIb	IIb only
Occurrence	Marginal marine sandstones offshore from major river systems in tropical climates	Coastal aeolian dunes and sandy sabkhas, any facies associated with evaporite brines
Facies associations	Oolitic ironstone	Evaporites
Associated early authigenic minerals	Siderite, calcite, phosphates	FeO rims, anhydrite, K feldspar, calcite, dolomite
Precursor		

XRD, X ray diffraction.

K/Ar dating allows the determination of the mean age of an authigenic illite occurrence. Although contamination of data by detrital illite and K feldspar can be a problem, this technique is much used in unravelling burial histories. The method is based on the known rate of decay of radioactive ^{40}K to stable ^{40}Ar. Important advantages of this method are that it is generally reasonable to assume that no argon was present in the mineral at the time of its formation, whilst illite has excellent Ar retention at burial temperatures of less than 175°C. Illite ages generally decrease with increasing burial depth. In sandstones, smaller sized fractions may give younger dates than coarser ones, as contamination by detrital illite decreases with decreasing particle size. However, in bentonites, smaller sized fractions can give older dates than coarser ones, which is consistent with the growth of fundamental particles as described above. In the Jurassic of the North Sea, most illite dates are in the range 50–30 Ma, coincident with the onset of rapid burial in the Late Cretaceous–Early Tertiary.

Glauconite

Although glauconite forms only at the sediment–water interface, with increasing burial temperature, aluminium is partly substituted for iron.

Chlorite

In sandstones, authigenic chlorite typically forms grain-coating rims of radial, interlocking platelets. Such rims have been the focus of petroleum company research, due to the inhibiting effect they have on quartz overgrowth cementation. The result of this inhibition can be sandstones with excellent reservoir quality at 4–5 km burial depth. Textural relationships indicate unequivocally that such chlorite forms very early in diagenesis. As with chloritic green clay pellets, early Fe chlorite rims are associated with sandstones offshore from major river systems within the tropics. Examples of this occur in the Jurassic Åre Formation in the Norwegian sector of the North Sea and the Miocene off the Niger Delta. It is now thought that Fe chlorite (chamosite) originates as odinite or a similar 7 Å iron-rich clay, although recent examples of odinite are rare. One line of evidence for a 7 Å clay precursor is that the proportion of 7 Å interlayers in 14 Å chlorite decreases with depth and burial temperature. This is not the only change exhibited by chlorite with depth: a gradual Ib to IIb polytype transition has been demonstrated for Fe chlorite, and there is a strong linear relationship between tetrahedral Al in authigenic chlorites and present-day temperature in hydrothermal systems. However, there is no simple relationship between chlorite polytype and temperature, and attempts to develop a universal chlorite geothermometer have largely failed as authigenic chlorite composition is influenced by detrital sediment composition. Chlorite in sandstones also occurs as a pore-filling replacement of ferromagnesian igneous rock fragments; the composition of this type of chlorite will reflect the mineral being replaced. Chlorite–smectite and corrensite can form by a similar process. These clays typically have more undulose platelets than true chlorite and are most frequently found in volcaniclastic sediments. Mg-rich chlorite, chlorite, and corrensite also form by diagenetic replacement of Mg smectite in evaporite basins. Non-chemical differences between Mg chlorite and Fe chlorite are summarized in Table 1.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray); Geochronological Techniques; Mineral Analysis. **Clay Minerals**. **Colonial Surveys**. **Sedimentary Rocks:** Ironstones; Sandstones, Diagenesis and Porosity Evolution.

Further Reading

Bjørlykke K (1998) Clay mineral diagenesis in sedimentary basins a key to the prediction of rock properties. Examples from the North Sea. *Clay Minerals* 33: 15 34.

Burley SD and MacQuaker JHS (1992) Authigenic clays, diagenetic sequences and conceptual diagenetic models in contrasting basin margin and basin centre North Sea Jurassic sandstones and mudstones. In: Houseknecht DW and Pittman ED (eds.) *Origin, Diagenesis and Petrophysics of Clay Minerals in Sandstones, Society of Economic Paleontologists and Mineralogists Special Publication 47*, pp. 81 110. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Ehrenberg SN and Nadeau PH (1989) Formation of diagenetic illite in sandstones of the Garn Formation, Haltenbanken area, mid Norwegian continental shelf. *Clay Minerals* 24: 233 253.

Hower J, Eslinger EV, Hower ME, and Perry EA (1976) Mechanism of burial metamorphism of argillaceous sediments: 1: Mineralogical and chemical evidence. *Geological Society of America Bulletin* 87: 725 737.

Huggett JM (1995) Formation of authigenic illite in Palaeocene mudrocks from the central North Sea: a study by high resolution electron microscopy. *Clays and Clay Minerals* 43: 682 692.

Huggett JM (1996) Aluminosilicate diagenesis in a Tertiary sandstone mudrock sequence from the central North Sea, U.K. *Clay Minerals* 31: 523 536.

Kisch HJ (1990) Calibration of the anchizone: a critical comparison of illite 'crystallinity' scales used for definition. *Journal of Metamorphic Geology* 8: 31 46.

Lanson B, Beaufort D, Berger G, Bauer A, Cassagnabere, and Meunier A (2002) Authigenic kaolin and illitic minerals during burial diagenesis of sandstones: a review. *Clay Minerals* 37: 1 22.

Longstaffe F (1989) Stable isotopes as tracers in clastic diagenesis. In: Hutcheon IE (ed.) *Short Course in Burial Diagenesis*, pp. 201 277. Toronto: Mineralogical Association of Canada.

Nadeau PH, Wilson MJ, McHardy WJ, and Tait J (1984) Interstratified clays as fundamental particles. *Science* 225: 923 925.

Pollastro R (1993) Considerations and applications of the illite/smectite geothermometer in hydrocarbon bearing rocks of Miocene to Mississippian age. *Clays and Clay Minerals* 41: 119 133.

Srodon J (1999) Use of clay minerals in reconstructing geological processes: recent advances and some perspectives. *Clay Minerals* 34: 27 38.

Deep Ocean Pelagic Oozes

R G Rothwell, Southampton Oceanography Centre, Southampton, UK

Introduction

Deep-ocean pelagic (from the Greek *pelagios*, meaning 'of the sea') sediments are areally and volumetrically the dominant sediment type found on the ocean floor. They comprise three main types, depending on their primary composition: deep-sea siliceous oozes, calcareous oozes, and deep-water red clays (Figure 1). Pelagic sediments mixed with terrigenous material derived from continental weathering are termed 'hemipelagic'. Siliceous and calcareous oozes are composed largely of test and test debris of planktonic micro-organisms such as foraminifera, coccolithophores, pteropods, diatoms, and radiolaria. The formation of pelagic sediments involves settling of material, that is commonly derived from biological surface productivity, but also includes wind-derived material that travels through the water column ('pelagic rain') to the seafloor (Figure 2). This process occurs throughout the world's oceans and true pelagic deposits tend to blanket seafloor topography. Local remobilization of pelagic sediments on topographic highs due to slope instability may result in pelagic turbidites (redeposited units) that pond in adjacent lows. However, the distribution of pelagic sediments is strongly depth controlled, because calcium carbonate shows increasing solubility with depth. In contrast to terrigenous sediments (the other main type of deep-sea sediment, largely composed of detrital material derived from continental weathering), pelagic sediments are characterized by low sedimentation rates and frequently contain a high proportion of authigenic minerals, extraterrestrial material, and, where physicochemical conditions allow, a substantial biogenic component.

History of Research

Although mention of marine sediments has been found in Greek and Roman texts, it was not until 1773 that the recovery of sediment from the deep sea was first recorded. In that year, Captain John

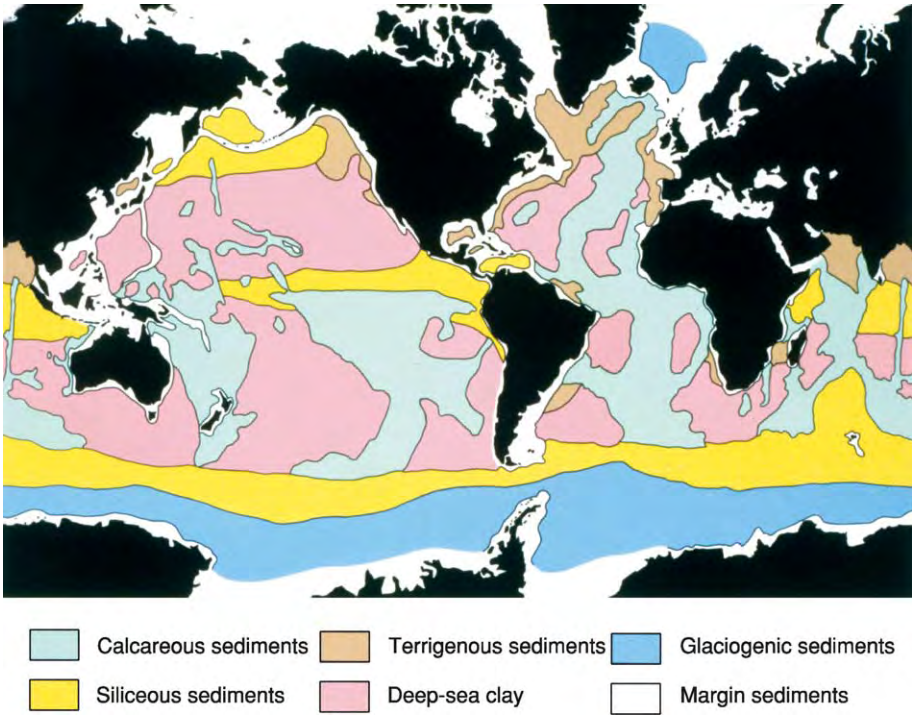


Figure 1 Distribution of the main sediment types on the ocean floor. Reproduced with permission from Davies TA and Gorsline DS (1976) Oceanic sediments and sedimentary processes. In: Riley JP and Chester R *Chemical Oceanography*, vol. 5, 2nd edn., pp. 1 80. London: Academic Press.

Phipps, on *HMS Racehorse*, bought up "fine soft blue clay" from a depth of 1250 m, the first recorded successful deep-sea sounding. The sample was taken in water on the southern edge of the Voring Plateau in the Norwegian Sea. In 1818, Sir John Ross recovered 6 lb (2.7 kg) of greenish mud from a depth of 1920 m in Baffin Bay, offshore Canada, using a deep-sea grab, representing one of the first recorded successful substantial deep-sea sediment recoveries. The laying of the first functioning submarine telegraph cable across the Straits of Dover in 1851 led to rapid expansion in the collection of deep-sea soundings and samples, driven by the prospect of the new means of intercontinental communication. However, it was not until the voyage of *HMS Challenger* (1872–76) that enough deep-sea samples were recovered to produce the first global seafloor sediment map. The voyage of *HMS Challenger*, led by Professor Charles Wyville Thomson, professor of natural history at the University of Edinburgh, was the first large-scale expedition devoted to oceanography. A wealth of seafloor samples were recovered from 362 observing stations, spaced at uniform intervals, along the 128 000-km track traversed during the voyage. John Murray, a naturalist on the *Challenger* expedition, oversaw the initial analysis of the recovered samples; Murray later edited the *Challenger Reports*, following the death of Wyville Thomson in 1882. The milestone *Challenger Report* on 'Deep-Sea Deposits' represented the first comprehensive volume on sediments of the deep-ocean seafloor. Published in 1891 with the assistance of Murray's co-worker AF Renard, this volume introduced many of the descriptive terms used today, such as 'red clay' and '*Globigerina* ooze', and provided the basis for further deep-ocean sediment studies. Murray also correctly related the distribution of shell-bearing plankton in the surface waters to the calcareous and siliceous sediments of the deep-ocean seafloor.

A major step forward in the investigation of seafloor sediments was the invention of the gravity corer by German researchers; the corer allowed continuous samples of extended lengths of sediment to be collected (but generally restricted to 1–2 m in length). The German South Polar Expedition (1901–03) collected several 2-m cores that were described by E Philippi in 1910. These cores showed that some deep-water sediments were stratified. During the period 1925–38, Germany ran a series of oceanographic expeditions using the ship *Meteor*, which recovered several 1-m-long cores from the southern Atlantic and Indian oceans. These cores, studied by Wolfgang Schott, showed changes in foraminifer species with depth and initiated the first

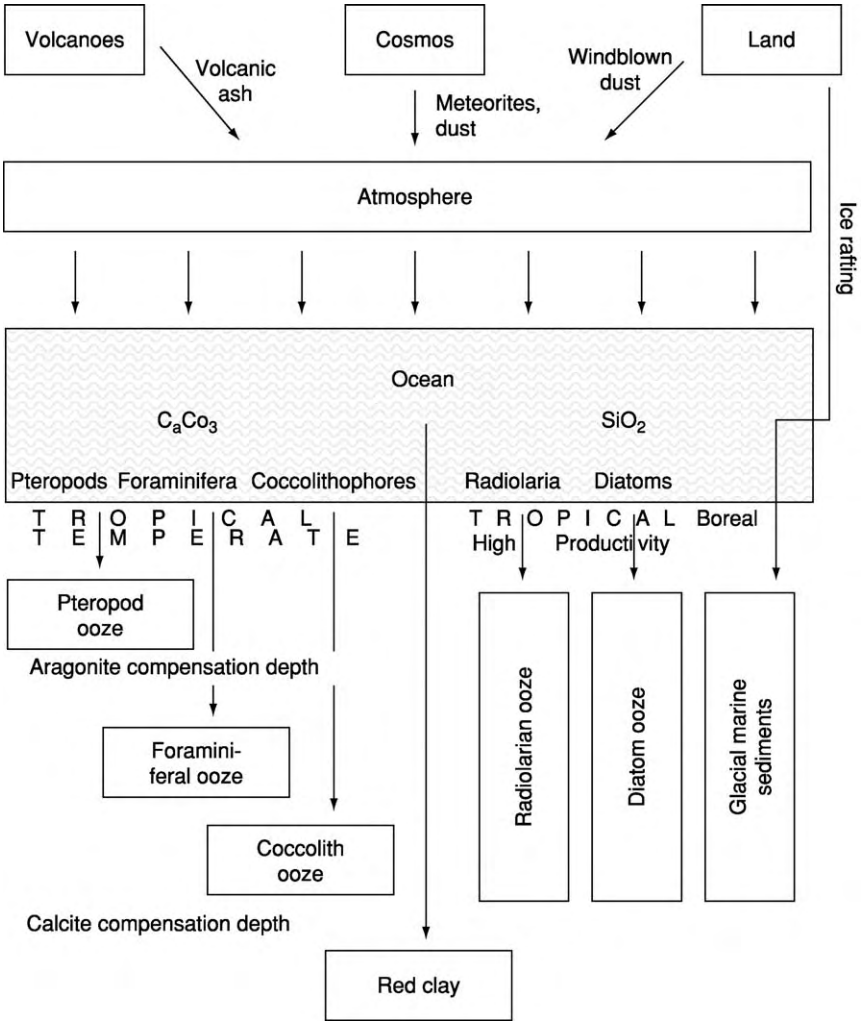


Figure 2 Sources and pathways for pelagic sedimentation in the oceans. Reproduced with permission from Hay WW (1974) Studies in Paleoceanography. Tulsa, Oklahoma: Society of Economic Paleontologists and Mineralogists.

palaeoceanographic studies. The Swedish Deep-Sea Expedition (1947–49) provided another fundamental advance in the study of deep-sea sediments. This expedition on the *Albatross* deployed a new kind of coring device, called a piston corer, which was developed by Börje Kullenberg, a marine geologist working in Gothenburg, Sweden. An innovative modification of the traditional gravity corer allowed the coring tube to fall past a stationary piston at the end of the wire, so that water was expelled from the falling tube above the piston and sediment was admitted from below. This allowed retrieval of much longer (10 m or more) and much less disturbed sediment cores. Acquisition of such long piston cores made possible the study of Pleistocene ocean history. The era of modern deep-sea sediment sampling had begun. Piston coring remains the main contemporary method of sampling the deep-sea sedimentary record, although sizing-up of the coring apparatus has led to development of giant piston corers, now capable of obtaining sediment cores up to 60 m in length. A giant piston core providing a continuous 54-m-long sediment record covering 4 million years of sedimentation was recovered from the Indian Ocean by the French ship *Marion Dufresne* in 1990; this is one of the longest piston cores ever recovered.

The advent of the Deep Sea Drilling Project (DSDP) in 1968 took deep-sea sediment sampling technology further. This programme, using the dynamically positioned drillship *Glomar Challenger*, set to recover continuous or semicontinuous sediment records from the ocean basins and heralded another new era in the exploration of the deep-ocean sedimentary record. Prior to the DSDP, the global inventory of cores recovering pre-Quaternary sediments

numbered fewer than 100. The DSDP and its successor, the Ocean Drilling Program (ODP), which began drilling operations in 1985 with a new drillship *JOIDES Resolution*, with improved capabilities, have led to major advances in our understanding of the processes of plate tectonics, of Earth's crustal structure and composition, of past changes in climate, and of conditions in ancient oceans. A vast amount of core material has been collected by the DSDP and ODP from all ocean basins (except the Arctic Ocean) and continuous sediment sequences back to Early Jurassic times have now been recovered.

Control of the Distribution of Pelagic Deposits

Pelagic sediments are defined as those formed of settled material that has fallen through the water column; their distribution is controlled by three main factors, distance from major landmasses (which affects their dilution by terrigenous, or land-derived, material), water depth (which affects sediment preservation), and ocean fertility (which controls surface water productivity). Pelagic sediments are composed largely of the calcareous or siliceous remains of planktonic micro-organisms or wind-derived material or mixtures of these. Several types of pelagic deposits can be identified on compositional grounds, and because seawater is increasingly corrosive with pressure and depth, the distribution of pelagic sediment types is strongly controlled by the calcite compensation depth (CCD), which is that depth at which the rate of supply of biogenic calcite equals its rate of dissolution (Figure 3). Therefore, below the CCD, only carbonate-free sediments accumulate. Thus the calcite compensation depth marks a major boundary defining the deposition of pelagic clays and calcareous sediments.

Another important depth in the water column is the lysocline, which lies above the CCD and is the depth at which the degree of undersaturation with respect to calcium carbonate is sufficient for dissolution of calcareous particles to become significant (Figure 3). Therefore, the lysocline is a depth that separates well preserved from poorly preserved solution-etched calcareous particles. The depth of the lysocline varies but it generally lies between 3000 and 5000 m. At water depths less than the lysocline, calcareous particles accumulate without loss through dissolution. The depth of the CCD varies as a function of a number of variables that reflect oceanic productivity patterns and the shoaling of the lysocline near continental margins. The CCD varies between 3500 and 5500 m in the Atlantic and Pacific oceans but has a mean depth of around 4500 m. The difference between the lysocline and CCD depths is also not constant, because this depends on the gradient of the concentration of carbonate ions in the water column overlying the sediments. However, the depth at which the calcite content of sediments falls to only a few percent is typically about 700 m deeper than the lysocline.

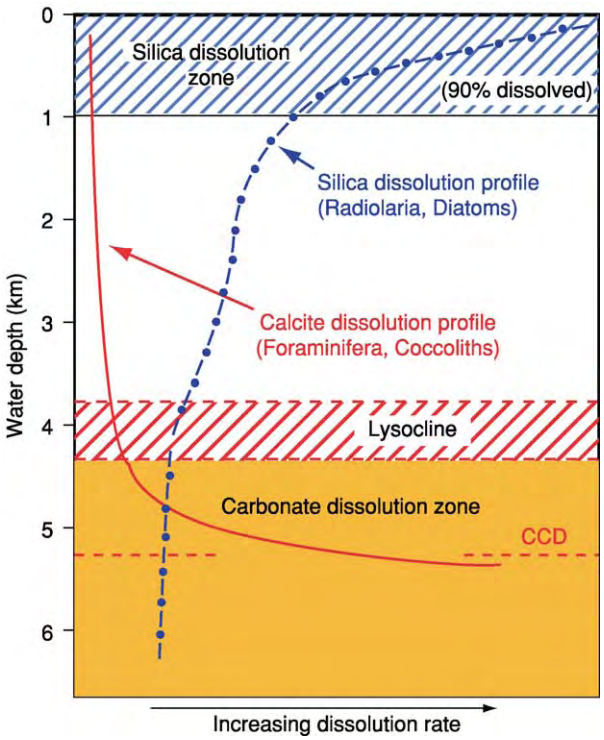


Figure 3 Generalized dissolution profiles of silica and calcite. Note that the depth of the lysocline, taken as the level below which there is rapid increase in calcite dissolution, and the calcite compensation depth (CCD), which is the depth at which the rate of supply of biogenic calcite equals its rate of dissolution, vary within and between ocean basins. Reproduced with permis sion from Douglas RG (2003) Oceanic sediments (Figure 5). In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 481 492. Dordrecht: Kluwer Academic Publishers.

Types of Pelagic Deposits

Pelagic sediments fall into two broad groups based on composition, deep-water pelagic clay and biogenic oozes. Deep-water pelagic clay is found in deep-ocean areas far from land, where solution has removed the biogenic component and only insoluble inorganic material, much of it wind-derived, remains. Biogenic oozes are composed largely of biogenic planktonic debris derived from surface water productivity (they contain more than 30% biogenic debris). Biogenic oozes are largely composed of the remains of zoo- and phytoplankton such as foraminifera,

coccolithophores, pteropods, diatoms, and radiolaria. In the upper water column, these remains are biologically 'packaged' and 'repackaged' into larger particles, which hastens their descent to the seafloor (e.g., as faecal pellets or phytoplankton aggregates). Indeed, most of the organic and skeletal matter produced in the euphotic zone is consumed and only a fraction is exported, and a fraction of this reaches the deep seafloor, where more is destroyed by dissolution. The distribution of biogenic oozes is strongly depth controlled due to dissolution of calcium carbonate with depth. Two main types are recognized, calcareous oozes, the composition of which is dominated by the remains of calcareous plankton, and siliceous oozes, which are dominated by the remains of siliceous plankton. Siliceous oozes lithify into radiolarites, diatomites, and cherts, whereas calcareous oozes lithify into pelagic chalks and limestones, and examples are well known from the geological record, well-documented examples occurring in the Troodos Massif, Cyprus, and the Ligurian Apennines in Italy.

Calcareous Oozes

Calcareous oozes may be dominated either by the tests and test debris of planktonic foraminifera (termed 'foraminiferal ooze') (Figure 4) or by the remains of planktonic plants (coccolithophores; termed 'nannofossil ooze'). In either type of calcareous ooze, the other component will often be the second most important constituent. In the modern world ocean, ~50% of the seafloor is blanketed by foraminiferal ooze (Table 1). Calcareous oozes commonly also contain a terrigenous fraction (which may amount to 10–15%), composed mainly of quartz and clay minerals, but may contain trace amounts of pyrite, iron and manganese precipitates, mica, chert, rock fragments, glauconite, feldspar, ferromanganese minerals, detrital carbonate, zeolites, volcanic glass, and cosmic spherules. Minor biogenic components may include benthonic (bottom-dwelling) foraminifera, ostracods, echinoid remains, radiolaria, silicoflagellates, diatoms, sponge spicules, pteropod shells and shell debris (in shallow water), phosphatic vertebrate remains and fish teeth.

Pteropods (pelagic gastropods) are relatively common zooplankton, especially in warm-water latitudes, and some forms secrete delicate aragonitic shells. Pteropod shells may range up to 30 mm in length, although most are in the range 0.3 to 10 mm. Aragonite is unstable and dissolves as ocean waters become undersaturated in respect to carbonate with depth. Consequently, pteropod-rich oozes are only found at depths shallower than 2500 m in the Atlantic Ocean and shallower than 1500 m in the Pacific Ocean.

Foraminifera comprise a group of protozoans characterized by a test of one to many chambers composed of secreted calcite or agglutinated grains. Test sizes are generally in the range 0.05–1 mm. Forms with agglutinated tests are typically benthonic (bottom-dwelling) and make only a very minor contribution to pelagic sediments, which are overwhelmingly dominated by the remains of globular planktonic forms. Modern species show clear latitudinal distribution patterns related to water temperature. Oxygen isotope analysis of planktonic foraminifera tests can provide estimates for past

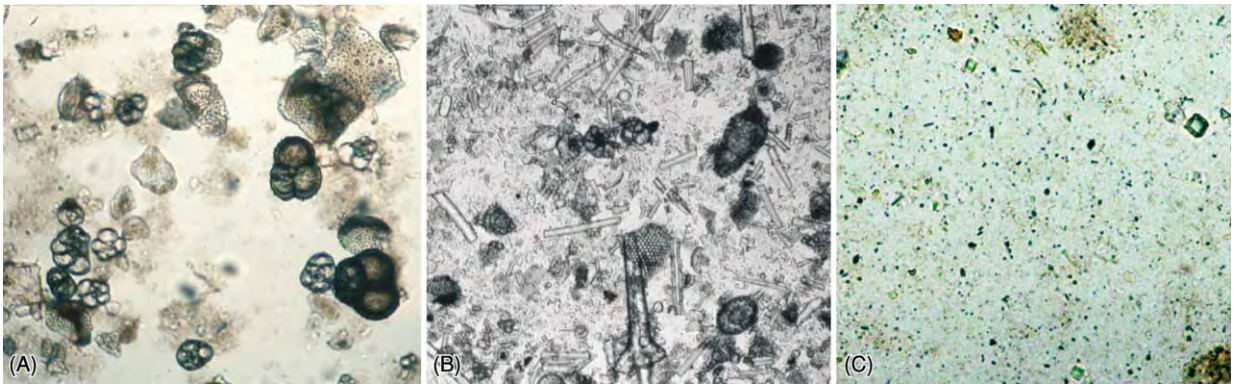


Figure 4 Illustration showing the three main types of pelagic sediments as seen under the microscope in plane polarized light. Left: Calcareous ooze from the North Atlantic Ocean, comprising mainly planktonic foraminifer tests and test fragments. The larger complete foraminifer tests are about 0.1 mm across. Centre: Siliceous ooze from the South Atlantic Ocean, comprising mainly silica sponge spicules (tubular forms), radiolaria (high relief bell shaped and circular forms, right of centre), and broken centric diatom frustules (lower left and centre). Two planktonic foraminifera can be seen in the upper centre field. The foraminifera are about 0.05 mm across. Right: North east Atlantic Ocean pelagic red clay containing rhomboid dolomite crystals. The red colour is due to the presence of amorphous or poorly crystalline iron oxide minerals and grain coatings. The largest dolomite rhomb (upper right) is about 0.01 mm across.

Table 1 Coverage of the deep ocean floor by pelagic sediments[a]

Sediment type	Seafloor coverage (%)			
	Atlantic Ocean	*Pacific Ocean*	*Indian Ocean*	*Total World Ocean*
Foraminiferal and nannofossil ooze	65.1	36.2	54.3	47.1
Pteropod ooze	2.4	0.1		0.6
Diatom ooze	6.7	10.1	19.9	11.6
Radiolarian ooze		4.6	0.5	2.6
Pelagic clay	25.8	49.0	25.3	38.1

[a]Data from Open University (1991) *Ocean Chemistry and Deep Sea Sediments*, 2nd edn. Oxford: Pergamon Press.

sea-surface temperatures and salinities. Isotope data from benthonic forms allow reconstruction of bottom-water mass histories. Foraminifera hence can provide important information on thermohaline structure and circulation patterns in ancient oceans.

Coccoliths are minute, usually oval, calcite plates produced by unicellular planktonic algae (family Coccolithophoridae); because of their small size, coccoliths are referred to as nannoplankton. In life, coccolith plates, eight or more in number, depending on the species, are attached to a membrane surrounding a living cell. Each organism (i.e., the cell surrounded by coccolith plates) is termed a 'coccosphere'. Coccospheres are generally spherical, usually 5–30 μm in diameter. The individual coccolith plates are usually around 3 μm in diameter, although some forms can be as large as 35 μm. On the death of the organism, the membrane holding the coccolith plates disintegrates, releasing the coccoliths to contribute to calcareous oozes. Coccoliths are single calcite crystals and are more resistant to dissolution than the tests of foraminifera or pteropods are. Globally, their diversity increases from a minimum in subpolar seas to a maximum in tropical and equatorial waters; and species distribution is closely linked to water masses.

Siliceous Oozes

Siliceous oozes are largely composed of the opaline silica tests and test fragments of siliceous plankton (Figure 4). Again, there are two main varieties: radiolarian ooze, composed mainly of the tests of radiolarians, and diatom ooze, dominated by the siliceous remains of unicellular plants (diatoms). Both types may contain minor amounts of silicoflagellates. Some sediments (for example, in some high-latitude abyssal environments and near spreading ridges) may also contain significant numbers of siliceous sponge spicules (Figure 4). Typically, siliceous ooze is present only in regions of high biological surface water productivity (such as the equatorial and polar belts and areas of coastal upwelling), where the depth of the seafloor is deeper than the calcite compensation depth. In the North Pacific and Antarctic belts of siliceous oozes, diatoms make up as much as 95% of the bulk sediment. The mineralogical composition of the detrital fraction of siliceous oozes is commonly similar to that of calcareous oozes, with quartz and clay minerals being the dominant detrital minerals.

Radiolaria are a diverse group of planktonic, pseudopod-bearing protozoans characterized by transparent opal skeletons. These exquisitely structured lattices are often of great complexity. Radiolarian tests show a great variety of shapes, but most are based on conical, spherical, or helmet-shaped forms. Most radiolarians are within the size range 20–400 μm. They are particularly abundant and diverse in equatorial latitudes (especially in areas of upwelling) and in subpolar seas. Radiolarian oozes occur mainly in the equatorial Pacific.

Diatoms are single-celled algae that secrete a test (called the frustule) of opaline silica. They are a major part of the phytoplankton and typically occur as pinnate (spindle-, rod-, or wedge-shaped) or centric (discoidal, spherical, elliptical, or oblong) forms. Most planktonic diatoms are centric types, although in Antarctic Seas, planktonic pinnate forms occur. They generally fall within the size range 10–100 μm. Diatoms represent most of the suspended silica in the water column and are the main contributors to deep-sea siliceous sediments. However, although in productive areas diatom concentrations are many millions of frustules per cubic metre, most tests are redissolved in the water column, because surface waters are greatly undersaturated in respect to silica due to high biological demand (Figure 3). Indeed, in areas of low silica supply, diatom assemblages in sediments are commonly biased to dissolution-resistant robust forms rather than to more fragile species. Diatoms are particularly abundant in regions of high productivity, especially in high latitudes and areas of upwelling.

Silicoflagellates are small unicellular flagellated marine plankton with internal skeletons of opaline silica. These skeletons consist of hollow rods arranged in a lattice, a common arrangement comprising a basal ring from which rods arise on one side to form an arch or dome, resulting in an overall hemispherical shape. Most silicoflagellates are in the size range of 20–50 μm. Although silicoflagellates are widespread in sediments, they are seldom abundant, so do not make a significant contribution to marine sediments.

Pelagic Clays

Deep-water pelagic clays (sometimes called 'red clay') are found only in deep-ocean areas, generally below water depths of 4000 m, far from land. Such clays cover large areas of the seafloor, particularly in the Pacific, southern Atlantic, and southern Indian oceans (Figure 1), in areas remote from terrigenous sources and below the calcite compensation depth. The reddish-brown appearance of these clays, first noted on the *Challenger* expedition, is due to the presence of amorphous or poorly crystalline iron oxide minerals and grain coatings. Pelagic clays usually contain less than 10% biogenic material and are mainly composed of fine-grained quartz and clay minerals, the bulk of which is derived from aeolian fallout and has been slowly deposited from fine suspensions. Typically, 75–95% of pelagic clay deposits consists of clay minerals with a grain size of less than 3 μm (Figure 4). These clay minerals are dominated by illite, smectite, kaolinite, and chlorite, with illite as the main type. Illite is, in fact, largely land derived and is transported to the ocean by rivers and glaciers and as windblown dust. Both kaolinite (a product of humid tropical weathering) and chlorite (typically derived from low-grade metamorphic rocks) are also mainly land derived. Smectite, however, is a low-temperature alteration product of volcanic ash and is particularly widespread on the Pacific Ocean floor. Wind transport is the major mechanism by which land-derived clay, fine-grained silt (commonly quartz), and dust reach the ocean surface, ultimately to be deposited in pelagic clays. The highest rates of aeolian dust deposition (up to 1000 mg cm $^{-2}$ ky $^{-1}$) are in the north-western Pacific downwind of far-east Asia. Substantial fluxes of windblown dust also enter the deep ocean offshore of the Sahara, South Africa, the Arabian peninsula, and the Horn of Africa and around Australia. The origin of wind-derived material in pelagic clays can be determined by rare earth geochemistry and study of Sr and Nd isotopes. Pelagic clays also commonly contain significant amounts of authigenic minerals, such as chert, zeolites, apatite, phosphorite, volcanic glass, and manganese micronodules, as well as indicators of slow sedimentation, such as fish debris and cosmic spherules. Pelagic clays may also contain varying amounts of feldspar, pyroxenes, and mica. In total, pelagic clays cover about 38% of the modern seafloor (Table 1).

Ferromanganese Deposits

New mineral phases may be formed on the seafloor (a process known as authigenesis) either by direct precipitation from seawater or by the alteration of pre-existing minerals or grains. Ferromanganese deposits are the most common and probably the most widely known authigenic deposits found on the deep-ocean seafloor. They occur as encrustations or crusts on submarine rock outcrops, or as discrete nodules and concretions on the seafloor. Ferromanganese crusts, which grow on exposed rock surfaces, acquire the elements necessary for their growth directly from seawater. Ferromanganese nodules occur throughout the sediment column, but the greatest concentrations are found on the surface; the nodules range in size from the microscopic (called 'micronodules', usually in the silt-sand size range) to the macroscopic, reaching several centimetres across (Figure 5). There are two main controls on nodule abundance: (1) the rate of accumulation of the host sediment, with the highest number of nodules being found on sediments with low accumulation rates (e.g., a few millimetres per thousand years), and (2) the presence of suitable accretion nuclei for the nodules to grow around; the nuclei may be small clumps of sediment, fragments of volcanic rock, shark teeth or teeth fragments, or even foraminifer tests. Ferromanganese nodules can show a great variety of shape and size and are found in all oceans, but are particularly common on the deep Pacific seafloor. Nodules may form by precipitation from the overlying seawater and from elements supplied from interstitial porewaters below the sediment surface, or through a combination of both element sources. The shape of the nodule may reflect the dominant source of elements available for its precipitation and growth. For circular nodules, it is thought that the dominant supply of metals is from the overlying seawater, whereas for discoidal, flattened nodules the dominant supply of metals may be via interstitial porewaters below the seabed. Nodule growth rates are slow, varying from a few to a few hundred millimetres per million years. Ferromanganese nodules are rich in iron and manganese as their name implies, but also contain relatively high concentrations of a number of trace metals, including cobalt, molybdenum, thorium, nickel, silver, iridium, and lead. Deep-sea ferromanganese nodules may someday become an important economic resource.

Metalliferous sediments, including iron and manganese-rich mudstones (termed 'umbers') and iron-rich sediments (termed 'ochres'), are frequently associated with ophiolites (fragments of oceanic crust that have been tectonically emplaced onto continental margins), well-known examples of which occur in Cyprus and Oman. These record past oceanic sediments that have contained hydrothermal minerals or authigenic ferromanganese deposits.

Figure 5 Deep sea photograph, showing a field of ferromanganese nodules over part of the Madeira Abyssal Plain, north east Atlantic Ocean. Individual nodules are a few centimetres across; water depth is 5400 m.

Biogenic Sedimentation in the World Ocean

Early researchers such as John Murray believed that pelagic sediments must accumulate slowly, but it was Wolfgang Schott, studying cores collected by the German *Meteor* expeditions of 1925–27, who was first able to demonstrate that Atlantic calcareous oozes had accumulated at rates of several centimetres thickness per thousand years. He found that the distinctive tropical foraminifer *Globorotalia menardii* was absent in glacial-period sediments in the North Atlantic, and that its appearance correlated with the start of the Holocene. Because the age of the base of the Holocene was known from land sections, Schott was able to determine accumulation rates for Atlantic pelagic calcareous oozes for this time period. Today, a wide variety of dating techniques (for example, radiocarbon dating and uranium series dating) can be used to determine accumulation rates. Pelagic sedimentation rates do vary considerably, but pelagic clays accumulate the most slowly (typically $0.1–0.5\,\mathrm{cm\,ky^{-1}}$), whereas sedimentation rates for calcareous oozes are typically in the range $0.3–5\,\mathrm{cm\,ky^{-1}}$ and siliceous oozes are in the range $0.2–1\,\mathrm{cm\,ky^{-1}}$.

Biogenic sediments show considerable variation in both space and time. In the present-day Atlantic Ocean, pelagic sediments are predominantly calcareous and siliceous sediments are virtually absent in the North Atlantic. In the Pacific, however, calcareous sediments are limited to oceanic ridges, plateaus, and seamounts (at water depths less than 3500 m) and also occur as a broad belt in the southern central Pacific. Siliceous sediments are widespread in the North Pacific, along the equator and adjacent to Antarctica. Calcareous sediments occur along the mid-ocean ridges in the Indian Ocean and siliceous sediments are widespread in the northern and southern Indian Ocean (Figure 1). Pelagic sediment distribution reflects both seafloor depth and ocean fertility. Where nutrient supply is low and surface waters are nutrient poor (especially in dissolved silica), sinking particles deliver to the seafloor more carbonate than silica (low Si/Ca ratio), which will be preserved, providing the seafloor lies above the CCD. Low nutrient supply favours production of coccolithophorids, which are fed on by small foraminifera, and long food chains develop in the euphotic zone. Foraminifera and coccoliths therefore dominate export to the seafloor. Where nutrient concentrations in surface waters are high, such as at upwelling areas and ocean divergence zones, diatoms will be the primary producers. Diatoms can reproduce rapidly and produce dense blooms (10^7 frustules per cubic metre). Food chains in these regions tend to be short because large diatoms are eaten by large zooplankton and fish (high trophic-level consumers). Export to the seafloor is high in silica and organic carbon, and flux rates are high, leading to siliceous sediment deposition. Further, bacterial decomposition of the organic carbon results in production of carbonic acid, which dissolves carbonate grains. In this way, carbonate is removed and the siliceous content as a proportion of total sediment increased. Deep-ocean circulation also leads to fractionation of silica and carbonate between ocean basins. In the North Atlantic, deep-water outflow is

exchanged for surface water inflow and bottom waters are young, nutrient poor, well oxygenated and saturated in respect to calcium carbonate. Sediments deposited here tend to be calcareous. In the North Pacific, in contrast, deep water is old and poorly oxygenated but nutrient rich, because surface water outflow is exchanged for deep-water inflow. Before reaching the North Pacific, the deep water has flowed through the southern Atlantic and Indian oceans, hence its age and low oxygen content. However, during this long passage, microbial breakdown of organic matter (which has depleted oxygen) produces CO_2 and regenerates nutrients. These waters therefore become undersaturated in regard to calcium carbonate but are enriched in nutrients and dissolved silica. Upwelling of this water will cause high surface productivity and diatom production, resulting in deposition of siliceous oozes with little calcareous content. Thus, pelagic sediment distribution is determined by bottom water circulation, which controls both the rate of particle dissolution and the productivity of surface waters through upwelling. In this way, in the modern ocean, the Atlantic is depositing carbonate and exporting silica, whereas in the Pacific, the reverse is happening. However, changes in climate and continental positioning and ocean connectivity, caused by plate motion, will affect ocean chemistry and fertility, and hence pelagic sediment deposition and distribution.

Data from the DSDP and ODP have shown that the distribution and relative abundance of seafloor sediment types have changed with time. Biogenic sediments were even more widely distributed in Cretaceous and Early Tertiary time. The deep-ocean sedimentary record provides a most important source for our knowledge of the past Earth, particularly regarding ocean fertility, geochemistry, evolution of marine biota, and past wind regimes and patterns.

See Also

Fossil Plants: Calcareous Algae. **Microfossils:** Foraminifera. **Sedimentary Processes:** Deposition from Suspension. **Sedimentary Rocks:** Oceanic Manganese Deposits.

Further Reading

Burton JD (1996) The ocean: a global geochemical system. In: Summerhayes CP and Thorpe SA (eds.) *Oceanography An Illustrated Guide*, pp. 165 181. London: Manson Publishing.

Chester R (2000) *Marine Geochemistry*, 2nd edn. [particularly chs. 13, 15, and 16]. Oxford: Blackwell Science.

DeMaster DJ (2004) The diagenesis of biogenic silica: chemical transformations occurring in the water column, seabed and crust. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry*, vol. 7, pp. 87 98. Oxford: Elsevier Pergamon.

Douglas RG (2003) Oceanic sediments. In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 481 492. Dordrecht: Kluwer Academic Publishers.

Li Y H and Schoonmaker JE (2004) Chemical composition and mineralogy of marine sediments. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry*, vol. 7, pp. 1 35. Oxford: Elsevier Pergamon.

Martin WR and Sayles FL (2004) The recycling of biogenic material at the seafloor. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry*, vol. 7, pp. 37 65. Oxford: Elsevier Pergamon.

Morse JW (2004) Formation and diagenesis of carbonate sediments. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry*, vol. 7, pp. 67 85. Oxford: Elsevier Pergamon.

Open University (1991) *Ocean Chemistry and Deep Sea Sediments*, 2nd edn. Oxford: Pergamon Press.

Rothwell RG (1989) *Minerals and Mineraloids in Marine Sediments An Optical Identification Guide*. Barking: Elsevier Applied Science.

Seibold E and Berger WH (1993) *The Sea Floor An Introduction to Marine Geology*, 3rd edn. [particularly chs. 3 and 6 9]. Berlin: Springer Verlag.

Stow DAV, Reading HG, and Collinson JD (1996) Deep seas. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, 3rd edn., pp. 395 453. Oxford: Blackwell Science.

Whitmarsh RB, Bull JM, Rothwell RG, and Thomson J (1996) The evolution and structure of ocean basins. In: Summerhayes CP and Thorpe SA (eds.) *Oceanography An Illustrated Guide*, pp. 113 135. London: Manson Publishing.

Dolomites

H G Machel, University of Alberta, Edmonton, Alberta, Canada

Introduction

Dolomite was first described in 1791 as a rock by Deodat de Dolomieu, who investigated samples from the Italian Alps. Dolomites are of special interest because they often form hydrocarbon reservoir rocks. Despite intensive research for more than 200 years, the origin of dolomites is subject to considerable controversy. This is because some of the chemical and/or hydrological conditions of dolomite formation are poorly understood, and because the available data often permit more than one viable genetic interpretation. This article covers the thermodynamic and kinetic conditions that favour dolomitization, mass balance considerations for the generation of massive dolostones, dolomite textures and pore spaces in dolostones, geochemical methods that are used in dolomite case studies, an overview of the various dolomitization models, and a brief section on secular variations in dolomite abundance.

Basics

Ideal, ordered 'dolomite' has a formula of $CaMg(CO_3)_2$ and consists of alternating layers of Ca^{2+}–CO_3^{2-} –Mg^{2+}–CO_3^{2-} –Ca^{2+}, etc., perpendicular to the crystallographic c axis. Most natural dolomite contains up to a few per cent Ca surplus (and a corresponding Mg deficit), as well as less than ideal ordering. 'Protodolomite' contains about 55–60% Ca, is poorly ordered, i.e., the alternating cation layer structure is poorly developed, and is common as a metastable precursor of well-ordered, nearly stoichiometric dolomite in both laboratory experiments and in nature. Good arguments have been made to abandon the term protodolomite or to restrict it to laboratory products, yet the term is useful to describe metastable precursors of dolomite in nature. The term 'dolostone' refers to a rock that consists largely (>75%) of the mineral dolomite. This term has been rejected by some, but has gained wide acceptance during the last 20 years. The term 'dolomites' is the best term to use to refer to types of dolomite that vary in texture, composition, genesis, or a combination thereof.

Two types of 'dolomite formation' are common, i.e., 'dolomitization', which is the replacement of $CaCO_3$ by $CaMg(CO_3)_2$, and 'dolomite cementation', which is the precipitation of dolomite from aqueous solution as a cement in primary or secondary pore spaces. Dolomites and dolostones that originate via replacement of $CaCO_3$ are called 'replacement dolomites' or 'secondary dolomites', especially in the older literature. A third type of dolomite formation is direct precipitation from aqueous solution to form sedimentary deposits. Dolomites that form in this way may be called 'primary dolomites'.

Genetically, all natural dolomites can be placed into two major families, i.e., 'penecontemporaneous' dolomites and 'postdepositional' dolomites. Penecontemporaneous dolomites may also be called 'syndepositional' dolomites. They form while a carbonate sediment or limestone still resides in the original environment of deposition as a result of the geochemical conditions that are 'normal' for that environment. Such dolomites are also called 'primary' or 'early diagenetic', although these terms are not strictly synonymous with penecontemporaneous. True penecontemporaneous dolomites appear to be relatively rare. Most known cases are of Holocene age, and are restricted to certain evaporitic lagoonal and/or lacustrine settings. It is quite possible, however, that such dolomites are much more common in the geological record than presently known, but their presence is hard to prove because of later diagenetic overprinting.

'Postdepositional' dolomites may also be called 'postsedimentary'. They form after a carbonate sediment has been deposited and removed from the active zone of sedimentation, which may happen via progradation of the sedimentary surface, burial and subsidence, uplift and emergence, eustatic sea-level fluctuations, or any combination of these. Such dolomites and dolostones are often called 'late diagenetic', although this term is not synonymous with postdepositional. Almost all known examples of massive, regionally extensive dolostones are postdepositional.

One aspect that transcends the above genetic grouping is that of hydrology. Whether syndepositional or postdepositional, the formation of large amounts of dolomite requires advection, i.e., fluid flow, because of chemical mass balance constraints. On the other hand, small amounts of dolomite can be formed without advection. In such cases, the Mg for dolomite formation is locally derived and redistributed, or supplied via diffusion. Examples include dolomite formed from Mg that was contained in (high-)Mg calcite, adsorbed to the surfaces of minerals, organic substances, or biogenic silica, or

that was contained in older primary or secondary dolomites.

Thermodynamic and Kinetic Constraints

The thermodynamic conditions of dolomite formation have been known quite well since at least the 1970s. The kinetics, however, i.e., the catalysts and inhibitors of dolomite formation, are relatively poorly understood and continue to be a source of controversy.

According to the present state of knowledge, dolomite formation is favoured chemically, i.e., thermodynamically and/or kinetically, under the following conditions: (1) low Ca^{2+}/Mg^{2+} ratios, (2) low Ca^{2+}/CO_3^{2} ratios (high carbonate alkalinity), (3) high temperatures, and (4) salinities substantially lower or higher than that of seawater. These constraints translate into four essential and common conditions for the formation of dolostones in natural settings:

- Settings with a sufficient supply of Mg^{2+} and CO_3^{2}. This condition favours marine settings and burial diagenetic settings with pore fluids of marine parentage, because seawater is the only common Mg-rich natural fluid in sedimentary/diagenetic settings.
- Settings with a long-lasting and efficient delivery system for Mg^{2+} and/or CO_3^{2} (also exporting Ca^{2+} in the case of calcite replacement). This favours settings with an active and long-lasting hydrological drive.
- Carbonate depositional settings and/or limestones that can be replaced.
- Settings in which fluids suddenly release CO_2, i.e., from hydrothermal solutions that ascend rapidly via fault systems.

Considering that the above chemical constraints allow dolomite formation in almost the entire range of surface and subsurface diagenetic settings, the question arises as to why there are so many undolomitized limestones. The essential conditions for the common lack of dolomitization appear to be:

- Ion pair formation (especially hydration), inactivating much of the Mg^{2+} and CO_3^{2} in solution.
- Insufficient flow because of the lack of a persistent hydraulic head, too small a hydraulic head, or insufficient diffusion, resulting in insufficient Mg^{2+} and/or CO_3^{2} supply.
- The limestones are cemented and not permeable enough, inhibiting or prohibiting the throughput of Mg-rich waters.
- The diagenetic fluids are incapable of forming dolomite because of kinetic inhibition, e.g., because the environment is too cold (most kinetic inhibitors of dolomite nucleation and growth are rather potent at temperatures below about 50°C), and the Ca^{2+}/Mg^{2+} ratio of many cold diagenetic fluids is not low enough for dolomitization.

The last point leads to kinetic factors, three aspects of which deserve special mention. Firstly, almost all researchers agree that most kinetic inhibitors that lower the nucleation rate and growth rate of dolomite are especially potent at temperatures below about 50°C. Hence, dolomite formation is easier at higher temperatures. Secondly, it is also generally acknowledged that dolomite forms via metastable precursors, but the significance of this phenomenon for the formation of massive dolostones is not clear. Thirdly, sulphate has been shown to increase as well as decrease the rate of dolomitization, and thus the role of sulphate is not clear, and may vary from place to place, depending mainly on fluid composition and temperature.

Mass Balance Constraints

Within the chemical constraints outlined in the previous section, the amount of dolomite that can be formed in a given diagenetic setting depends on the stoichiometry of the reaction, temperature, and fluid composition. Dolomitization can be represented by two equations, i.e.,

$$2CaCO_3(s) + Mg^{2+}(aq) \rightarrow CaMg(CO_3)_2(s) + Ca^{2+}(aq) \quad [1]$$

where '(s)' is solid and '(aq)' is aqueous, or by

$$CaCO_3(s) + Mg^{2+}(aq) + CO_3^{2}\ (aq) \rightarrow CaMg(CO_3)_2(s) \quad [2]$$

Reactions [1] and [2] are end members of a range of possible reaction stoichiometries, i.e.,

$$(2-x)CaCO_3(s) + Mg^{2+}(aq) + xCO_3^{2}\ (aq) \rightarrow CaMg(CO_3)_2(s) + (1-x)Ca^{2+}(aq) \quad [3]$$

Reaction [3] can be used to represent dolomitization in general, as it encompasses reactions [1] and [2]. For $x = 0$, reaction [3] becomes reaction [1] and, for $x = 1$, reaction [3] becomes reaction [2]. Dolomite cementation is most simplistically represented by

$$Ca^{2+}(aq) + Mg^{2+}(aq) + 2CO_3^{2}\ (aq) \rightarrow CaMg(CO_3)_2(s) \quad [4]$$

If dolomitization proceeds via reaction [1], and if the dolomitizing solution is normal seawater, about 650 m^3 of solution is needed to dolomitize 1 m^3 of limestone with 40% initial porosity at 25°C. Dolomitization may not take place with 100% efficiency, however, and some Mg in excess of saturation is carried away by the dolomitizing solution. In such cases, larger water/rock ratios are required for complete dolomitization. If seawater is diluted to 10% of its original concentration, as is the case in a typical seawater–freshwater mixing zone, ten times as much water is needed. On the other hand, only about 30 m^3 of brine is needed per cubic metre of limestone at 100% dolomitization efficiency in the case of a halite-saturated brine. The role of increasing temperature in the underlying thermodynamic calculations is to reduce the amount of Mg necessary for dolomitization, because the equilibrium constant (and hence the equilibrium Ca/Mg ratio) is temperature dependent. For example, at 50°C, only about 450 m^3 of seawater is needed for complete dolomitization of 1 m^3 of limestone with 40% initial porosity at 100% efficiency. The amounts of dilute and hypersaline waters change accordingly.

These calculations have two major implications. Firstly, large water/rock ratios are required for complete dolomitization, and the more dilute the solution, the larger the water/rock ratio. This necessitates advection for extensive and pervasive dolomitization, which is why all models for the genesis of massive dolostones are essentially hydrological models. The exceptions are natural environments in which carbonate muds or limestones can be dolomitized via diffusion of magnesium from seawater rather than by advection. Secondly, variable reaction stoichiometries result in variable porosity development during dolomite formation (see below).

Rock and Pore Classifications

Crystal size distributions are classified as 'unimodal' and 'polymodal', whereas crystal shapes are classified as 'planar-e' (euhedral), 'planar-s' (subhedral), and 'nonplanar-a' (anhedral). Using this semantic scheme, almost all other dolomite texture types can be named, i.e., planar-c (cement), planar-p, and nonplanar-p (both porphyrotopic). Saddle dolomite, with its distinctive warped crystal faces, is simply called nonplanar (Figure 1). A complete textural description includes recognizable allochems or biochems, matrix, and void fillings. Particles and cements may be unreplaced, partially replaced, or completely replaced. Replacement may be mimetic or non-mimetic, which can be added to a rock description, such as 'unimodal, non-mimetic, planar-s dolomite'.

Pores in dolostones are commonly addressed using the same classification as for limestones, with pore types such as mouldic, vuggy, shelter, etc. This

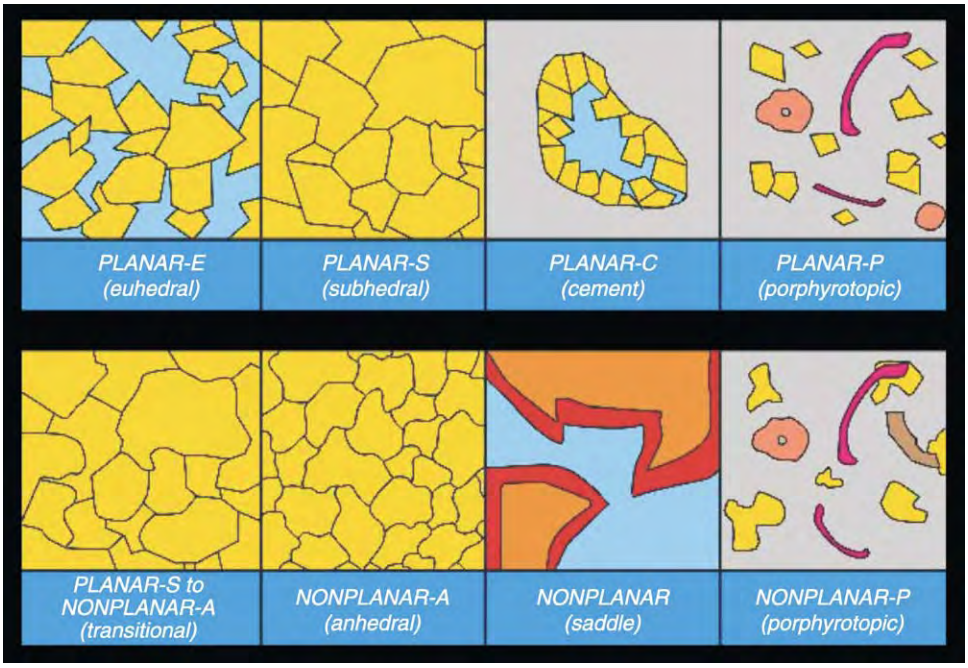


Figure 1 Dolomite textural classification combined from Gregg and Sibley (1984) and Sibley and Gregg (1997), supplemented by a 'transitional' form. Reproduced with permission from Wright WR (2001) Dolomitization, fluid flow and minerlization of the Lower Carboniferous rocks of the Irish Midlands and Dublin Basin. Unpub. Ph.D. thesis, Univerisity College Dublin, Belfield, Ireland, 407 p.

classification is independent of pore size. The latter, however, is of special interest for the petroleum industry. Another classification contains categories in size/magnitude from the very smallest to the very largest, i.e., from mercury injection capillary measurements (MICPM) and scanning electron microscopy (SEM) to karst caverns, respectively (Figure 2).

Textural Evolution

The textures and reservoir characteristics of natural dolostones are highly variable. On the microscopic scale, a unimodal size distribution generally results from a single nucleation event and/or a unimodal primary (pre-dolomite) size distribution of the substrate. Polymodal size distributions result from multiple nucleation events and/or a differential nucleation on an originally polymodal substrate. Planar crystal boundaries tend to develop at growth below about 50°C (the so-called 'critical roughening temperature'), whereas nonplanar boundaries tend to develop at $T > 50$°C and/or high degrees of supersaturation.

Within this framework, observations from many dolostone occurrences show that dolomitization often proceeds in a certain sequence of steps that correspond to certain textural types on the macroscopic scale. Within limits, these steps correspond to certain types of dolomitizing fluids (especially seawater and its derivatives) and/or meteoric water incursion. The most common sequence includes:

1. *Matrix-selective dolomitization*. Dolomitization begins as a selective replacement of the matrix (Figure 3).
2. *Vugs and moulds*. Holes resulting from the dissolution of undolomitized fossils and allochems (Figures 4 and 5).
3. *Emplacement of calcium sulphate*. Commonly anhydrite, both as a replacement and as a cement during advanced dolomitization from seawater (Figures 5 and 6).
4. *Development of two dolomite populations*. A smaller sized population with 'cloudy' centres with or without clear rims (overgrowths), and a larger population (Figures 7 and 8) resulting either from recrystallization or inherited from primary textural features.
5. *Dolomite cementation ('overdolomitization')*. Dolomite cement as overgrowth on the earlier formed dolomite crystals.

Furthermore, outcrop evidence shows that there is a distinct difference in the textures resulting from 'low-temperature' versus 'high-temperature' dolomitization of limestones. Empirical evidence suggests that the range of 50–80°C approximately marks the boundary between these two temperature realms. In the low-temperature settings, dolomitization commonly is matrix selective and at least partially fabric retentive, as discussed earlier, whereas dolomitization tends to be fabric destructive in the high-temperature settings (Figures 9 and 10). However, there are counterexamples.

Saddle dolomite (Figures 11 and 12) is a special type of dolomite. Its crystallographic, geochemical, and paragenetic characteristics suggest formation at temperatures above about 80°C. Saddle dolomite forms from stylolitization of older dolomites, as a

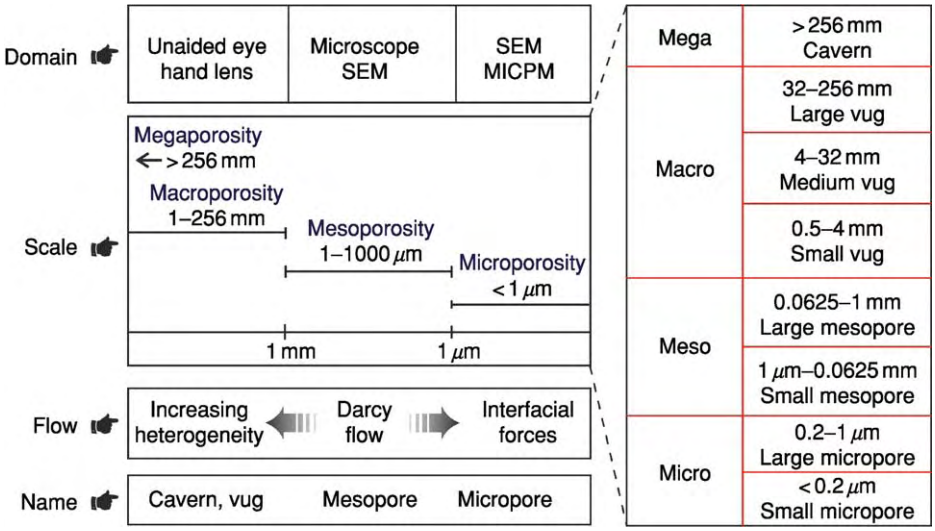


Figure 2 Pore size classification for carbonates. Measurements under 'scale' refer to pore diameters. MICPM, mercury injection capillary measurements; SEM, scanning electron microscopy. Reproduced from Luo and Machel (1995). Reprinted by permission of the AAPG whose permission is required for further use. AAPG ©1995.

Figure 3 Uncemented *Smithiphyllum* and *Phacelophyllum* with calcite preservation of the delicate chamber walls (trabeculae) in partially dolomitized matrix. Sample is from the Devonian Nisku Formation, Alberta, Canada.

Figure 4 Vuggy dolostone that resulted from (macro)dissol ution of unreplaced calcite matrix and fossils, similar to the sample shown in **Figure 3**. Connection of pores is intercrystalline pervasive. Sample is from the Devonian Nisku Formation, Alberta, Canada.

by-product of thermochemical sulphate reduction, and from hydrothermal fluids. Saddle dolomite commonly occurs as gangue in MVT-type metal sulphide deposits.

Porosity and Permeability

Comparison of the molar volumes of calcite and dolomite reveals that about 13% of porosity is generated in the so-called 'mole-per-mole' replacement of calcite by dolomite according to reaction [1] (whereby two moles of calcite are replaced by one mole of dolomite). However, several other processes are involved. As a generalization, dolostones can have higher, the same, or lower porosity and permeability than their precursor limestones, and the poroperm evolution has to be investigated on a case-by-case basis. Many/most dolostones have higher porosities than limestones, and this fact may be the result of one or several of six processes (**Figure 13**): (1) mole-per-mole replacement; (2) dissolution of unreplaced calcite (solution undersaturated for calcite after all Mg in excess of dolomite saturation is exhausted); (3) dissolution of dolomite (without externally controlled acidification); (4) acidification of the pore waters (via decarboxylation, clay mineral diagenesis, etc.); (5) fluid mixing ('*Mischungskorrosion*'); and (6) thermochemical sulphate reduction, which may generate porosity under certain circumstances.

Dolomitization almost invariably involves the reorganization of permeability pathways. Commonly, permeability increases along with porosity, and vice versa, such as in the Upper Devonian Grosmont Formation in eastern Alberta, which hosts a giant heavy-oil reservoir, and in the Cambrian–Ordovician Bonneterre Formation of Missouri, USA, which hosts one of the world's largest MVT-type sulphide deposits. Planar-e dolomites tend to have the highest porosities and permeabilities, the latter caused by well-connected pore systems with low pore to throat size ratios (as indicated by mercury injection curves);

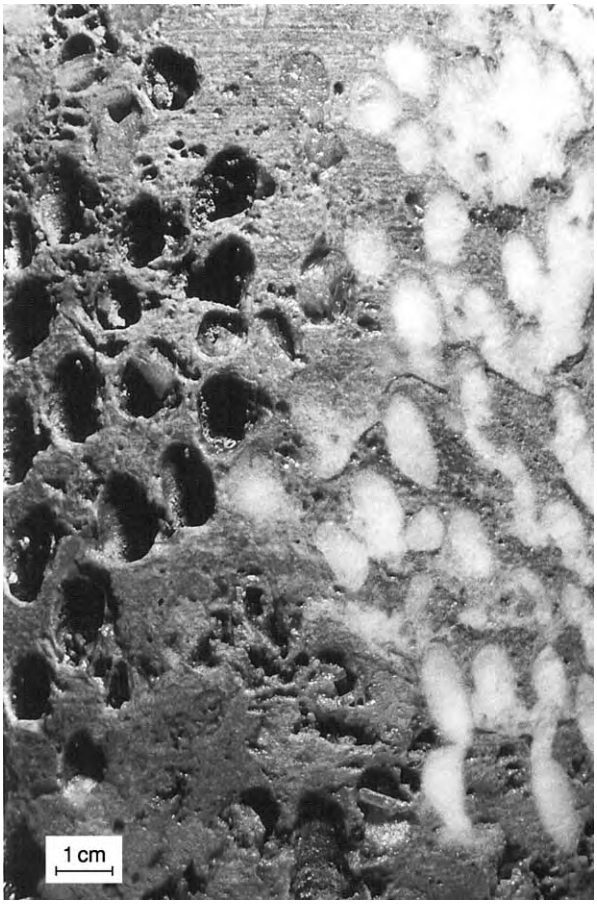


Figure 5 Coral mouldic porosity in tight matrix dolomite. White anhydrite occurs partially as a replacement (top right) and partially as a cement in coral mouldic porosity (centre and bottom right). From the Devonian Nisku Formation, Alberta, Canada.

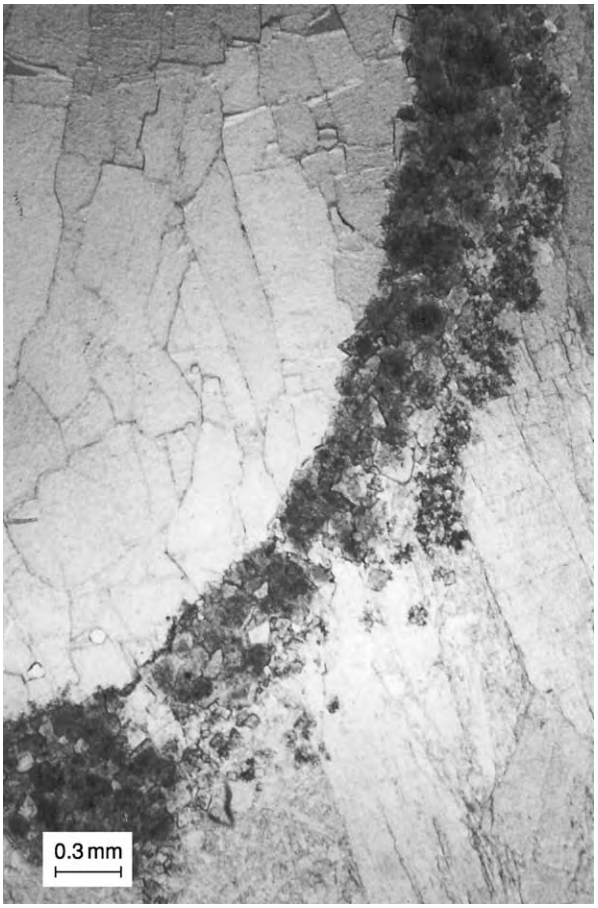


Figure 6 Thin section photomicrograph of the sample shown in Figure 5.

in planar-s dolomite, the permeabilities do not increase as rapidly with increasing porosity, corresponding to relatively large pore to throat size ratios; nonplanar dolomites have a statistically insignificant porosity–permeability relationship, whereby the pore systems have a high tortuosity and large pore to throat size ratios. Some authors have disputed that there is a systematic correlation between porosity and permeability in dolostones, or that these two petrophysical parameters are enhanced in dolostones relative to limestones. The Grosmont and the Bonneterre clearly show, however, that there is a relationship between porosity and permeability in at least some major and economically important dolostone sequences.

In cases of mole-per-mole replacement, the fabrics of the original limestone must be at least partially obliterated in order to account for the volume change during the replacement process. On the other hand, limestones dolomitized in a volume-per-volume replacement should not contain secondary intercrystal pores or dolomite cements, and the primary textures may be partially or largely, even mimetically (if the crystal size is very small), preserved. Partial or complete obliteration of primary textures can occur even in a volume-per-volume replacement, however, if there is a marked change in crystal size (usually an increase, due to Ostwaldt ripening) and/or porosity redistribution.

Dolomite Geochemistry

A wide range of geochemical methods may be used to characterize dolomites and dolostones, and to decipher their origin. One aspect of particular interest is the determination of the type of the dolomitizing fluid(s), i.e., marine, evaporitic, subsurface brine, etc., and the identification of the direction of fluid flow during dolomitization. The latter can often be ascertained by mapping a gradient in dolomite abundance, i.e., complete dolomitization near the upflow direction and decreasing abundance downflow. However, this approach necessarily fails where

Figure 7 Dolostones consisting of domains of relatively tight, light to medium grey dolomite intergrown with domains of highly porous, brownish dolomite. From the Devonian Nisku Formation, Alberta, Canada.

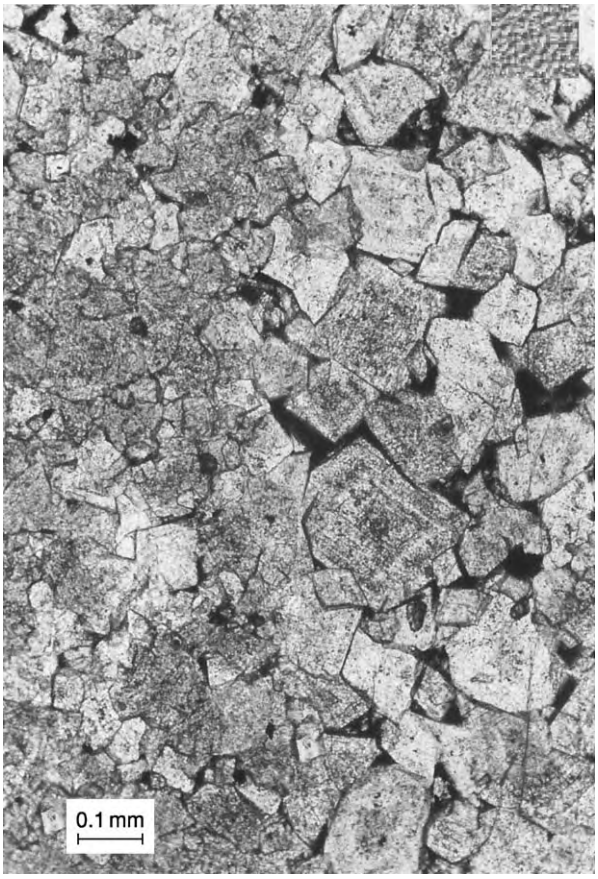


Figure 8 Thin section photomicrograph of the sample shown in Figure 7 from the boundary region between the two dolomite types.

dolomitization is 'complete' or where exposure and/or core are insufficient. In such cases, the geochemical composition of dolomites can be used, within limits, to determine the fluid composition and/or the flow direction during dolomitization.

Oxygen and carbon isotope ratios ($\delta^{18}O$ and $\delta^{13}C$) are the most widely applied and probably best understood geochemical methods in dolomite research. $\delta^{18}O$ values of carbonates can be used, within limits, to determine the $\delta^{18}O$ value and/or temperature of the fluid present during crystallization, including a possible distinction between meteoric, marine, and/or evaporitic waters.

Fluid inclusion homogenization temperatures arguably are the best method to determine the temperature of formation of dolomites (and other minerals), in addition to the highly desirable information on fluid compositions that can be gained from freezing experiments. Unfortunately, the vast majority of fluid inclusions in dolomites are too small for standard heating–freezing runs, as phase transitions within the inclusions are not observable. This is especially true of matrix-selective, replacive dolomites. On the other hand, the sparry saddle dolomite cements found in late-diagenetic dissolution vugs, but also as a replacement, commonly yield excellent fluid inclusion data.

Where possible, fluid inclusion homogenization temperatures are used in conjunction with $\delta^{18}O$ values to further characterize the conditions of dolomite formation. This type of analysis can reveal the direction(s) and temperature gradient(s) of the dolomitizing fluid flow on a local scale (a few kilometres) or on a regional scale (over several hundred kilometres). Mapping and contouring of the oxygen isotope and/or fluid inclusion homogenization temperatures have shown clear, spatially resolved gradients in some locations. Unfortunately, such gradients do not appear to be common.

The $\delta^{13}C$ values of the carbonates can be used to identify whether meteoric water (carrying soil CO_2) was involved, whether thermogenic or biogenic CH_4 was oxidized, whether CO_2 from microbial processes or organic matter maturation was available,

Figure 9 Outcrop photograph of Upper Carboniferous carbonates from the southwestern Cantabrian Zone, Spain: high temperature dolomitization of limestones. The dolostone appears dark where covered with lichen (upper right corner), yet light beige where cleaned of lichen (centre). The limestone (left) has a medium grey colour. Note the sharp contacts between limestones and dolostones, and that sedimentary and diagenetic textures visible in the limestones are obliterated in the dolostones. Hammer for scale.

Figure 10 Outcrop photograph of Upper Carboniferous carbonates from the southwestern Cantabrian Zone, Spain: high temperature dolomitization of limestones. The dolostone is light coloured and at the bottom. Note the sharp contacts between limestones and dolostones, and that sedimentary and diagenetic textures visible in the limestones are obliterated in the dolostones. Hammer for scale.

or whether thermochemical sulphate reduction contributed carbon to the carbonates. Also, there is a secular carbon isotope trend that may be used in the dating of marine dolostones, but only under very favourable circumstances.

Radiogenic isotopes are less commonly used in studies of carbonate diagenesis, mainly because they are analytically much more expensive. Yet, strontium isotopic compositions (usually quoted as $^{87}Sr/^{86}Sr$ ratios) are an excellent parameter to deduce compositional changes and, especially, flow directions of the fluids from which the diagenetic carbonates have formed. This is because strontium isotopes, as opposed to the more commonly used stable isotopes of oxygen and carbon, are not fractionated by pressure, temperature, and (as in the case of carbon) microbial processes.

The direction of fluid flow can also be determined using trace elements, which is especially attractive because trace element analysis is the cheapest of all the common geochemical methods. Trace element trends have been documented in several Phanerozoic dolostone sequences.

For all practical applications, i.e., the determination of fluid composition and/or fluid flow direction, the absence, presence, and/or degree of recrystallization is important. If changes via recrystallization in texture, structure, composition, and/or palaeomagnetic properties are so small that the total data range after recrystallization is the same as when the dolomite first formed, a dolomite/dolostone is said to be 'insignificantly recrystallized' (**Figure 14**, top), and its properties are still representative of the fluid and environment of dolomitization. On the other hand, if these changes result in data ranges that are larger than

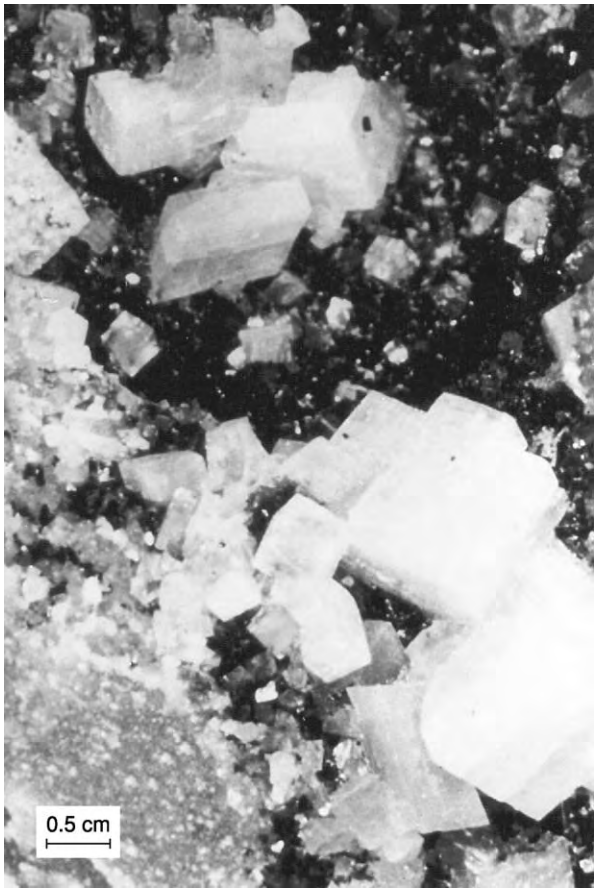


Figure 11 Core specimen of milky white saddle dolomite cement in a vug that is coated with solid bitumen ('dead oil'). Host rock is grey matrix dolomite. Saddle dolomite appears as a large crystal in the centre and lower right, with undulous extinc tion. From the Devonian Nisku Formation, Alberta, Canada.

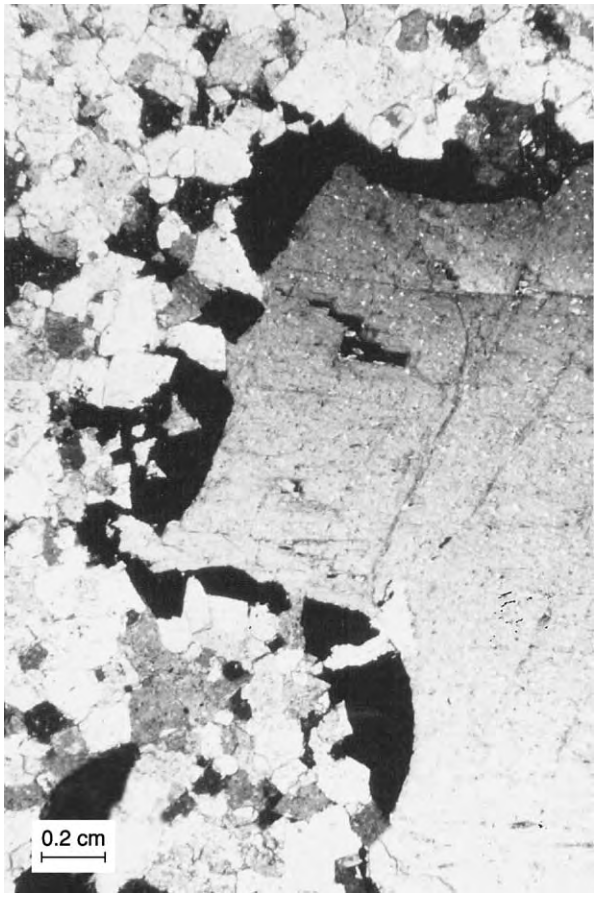


Figure 12 Thin section photomicrograph (transmitted light with crossed polarizers) of the sample shown in Figure 11.

the original ones, a dolomite/dolostone is said to be 'significantly recrystallized' (Figure 14, bottom), and its properties are no longer representative of the fluid and environment of dolomitization. In this case, the measured properties are reset and they characterize the last event of recrystallization. Furthermore, not all measurable properties must be reset during recrystallization. For a dolomite to be recognized as significantly recrystallized, only one of the measurable properties has to be modified to a range larger than the original one. In this case, inherited properties may still represent the event of dolomitization, whereas reset properties represent recrystallization.

Most dolomites that originally form very close to the surface and/or from evaporitic brines tend to recrystallize with time and during burial, because they form as metastable protodolomite phases that become thermodynamically highly unstable as a result of increasing temperature, increasing pressure, and changing fluid composition. On the other hand, dolomites that form at several hundred to a few thousand metres of depth are not or hardly prone to recrystallization, because these dolomites tend to form as rather stable (nearly stoichiometric, well-ordered) phases, whose stability does not change much during further burial and with increasing time.

Another important aspect of dolomite research requiring the application of geochemical methods is the recognition of hydrothermal activity. In many studies, the presence of saddle dolomite has been taken as an indication of elevated heat flow and/or increased temperatures during dolomite formation. However, the presence of saddle dolomite merely indicates a temperature of formation that is relatively high in the context of diagenetic studies. Its presence in uplifted dolomites, or in structurally inverted subsurface systems currently at lower temperatures, may merely reflect processes formerly operating at depths (and temperatures) at or around maximum burial and with normal geothermal gradients. Saddle dolomite may be hydrothermal, geothermal, or hydrofrigid (Figure 15). A distinction between these

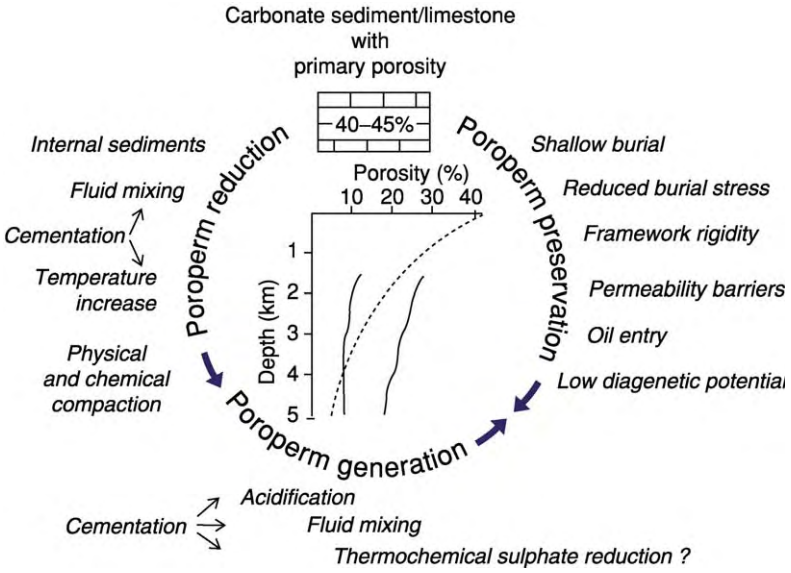


Figure 13 Major processes of porosity and permeability ('poroperm') generation, preservation, and reduction in carbonates. The inset contains averaged porosity/depth data from Mesozoic and Cenozoic limestones and dolostones in south Florida (broken curve; from Schmoker and Halley (1992)) and from the Jurassic Smackover oolite carbonate reservoirs in the southern USA (full curves, which envelop the measured maximum and minimum values below depths of about 1.5 km; from Scholle and Halley (1985) and Heydari (1997)). The Florida trend can be considered typical for most carbonates elsewhere. The large variations in Smackover carbonates at any given depth reflect highly variable degrees of porosity generation, preservation, and reduction due to various competing diagenetic processes. Reproduced with permission from Machel HG (1999) Effects of groundwater flow on mineral diagenesis, with emphasis on carbonate aquifers. *Hydrogeology Journal* 7: 94 107, ©1999 Springer Verlag.

alternatives can only be made if the temperature of formation of saddle dolomite is considered relative to the temperature of the surrounding rocks at the time of saddle dolomite formation, e.g., via fluid inclusion data in silicates or other carbonates, vitrinite reflectance data, reconstruction of maximum burial and geothermal gradient, etc.

Environments and Models of Dolomitization

In near-surface and shallow diagenetic settings, dolomitization models are defined and/or based on water chemistry, but on hydrology in deeper burial diagenetic settings. This poses an obvious dilemma when some type of near-surface diagenetic fluid moves into the deeper subsurface, and when deeper subsurface fluids (commonly brines) ascend into shallow diagenetic settings. Research over the last 15–20 years has revealed several such 'cross-overs' or 'overlaps' between models, which has resulted in unnecessary ambiguities in semantics and classification.

Penecontemporaneous Dolomites and the Microbial/Organogenic Model

In shallow marine to supratidal environments, penecontemporaneous dolomites commonly form in quantities of <5 vol.%, mostly as Ca-rich and poorly ordered, microcrystalline to fine crystalline cements and/or directly from aqueous solution. These occurrences include: lithified supratidal crusts (e.g., Andros Island, Sugarloaf Key, Ambergris Cay); thin layers in salinas (e.g., Bonaire, West Caicos Island) and evaporative lagoons/lakes (e.g., Coorong); and fine crystalline cements and replacements in peritidal sediments (e.g., Florida Bay, Andros Island). The dolomite-forming fluids are normal seawater and/or evaporated seawater, in some cases with admixtures of evaporated groundwater. There are also cases of penecontemporaneous dolomite formation in association with volcanics or volcanic activity, dolomite as fine crystalline supratidal weathering products of basic rocks, and hydrothermal dolomite forming at submarine vents. One especially important type in this group, commonly classified under hypersaline dolomites, forms lenses and layers of up to 100 vol.% dolomite in sabkhas.

Penecontemporaneous dolomites in hemipelagic to pelagic settings commonly form in very small quantities as microcrystalline protodolomites, generally with less than 1 wt.%. However, under favourable circumstances, the amount of dolomite can reach up to 100 vol.% locally. For example, Miocene hemipelagic carbonate sediments from the margin of the

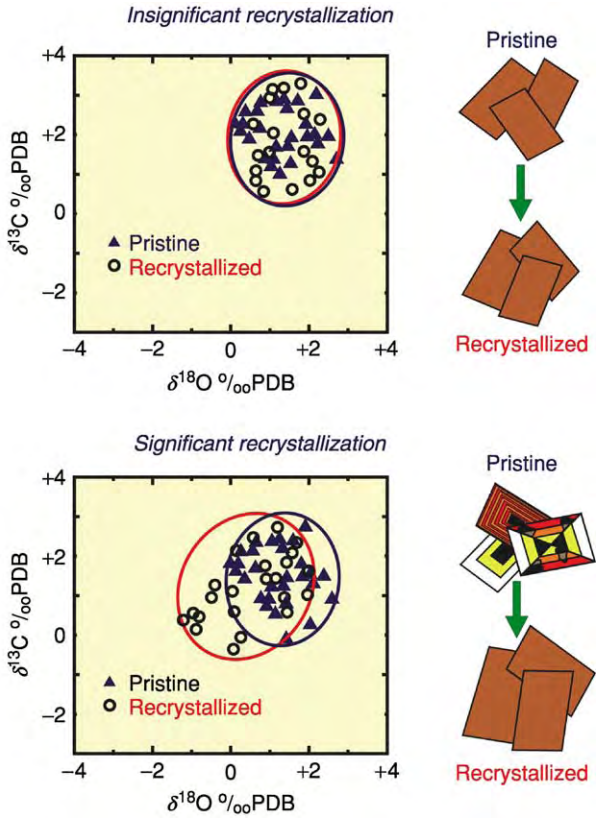


Figure 14 Schematic illustration of insignificant and significant recrystallization. For the properties shown, i.e., $\delta^{13}C$ and $\delta^{18}O$ values, crystal sizes, and luminescence, the pristine and recrystallized samples are the same in the case of insignificant recrystallization, but different, despite some overlap, in the case of significant recrystallization, i.e., at least some isotope values of the recrystallized samples fall outside the range of the pristine samples. The crystals also have increased in size and lost their respective zonations (concentric, sector, oscillatory). Reproduced with permission from Machel HG (1997) Recrystallization versus neomorphism, and the concept of 'significant recrystallization' in dolomite research. *Sedimentary Geology* 113: 161–168.

Great Bahama Bank are partially to completely dolomitized over a depth range of about 50–500 m subsea. In this setting, dolomite forms as a primary void-filling cement and by replacing micritic sediments, red algae, and echinoderm grains.

Both settings of penecontemporaneous dolomite formation mentioned above appear to be linked to the 'microbial model' or 'organogenic model' of dolomitization. According to this model, dolomite may be formed syndepositionally or early postdepositionally, i.e., at depths of a few centimetres to a few hundred metres, under the influence of, or promoted by, bacterial sulphate reduction and/or methanogenesis. The latter is indicated by organogenic $\delta^{13}C$ values. The exact role of microbial activity in reducing the notorious kinetic barriers to dolomitization is unknown, although it seems likely that the reduction of Mg and Ca hydration barriers, an increase in alkalinity, and/or changes in pH are involved. Most microbial/organogenic dolomites are cements; some are replacive, typically fine crystalline to microcrystalline (less than 10 μm), calcic and poorly ordered protodolomites. The chief modes of Mg supply are diffusion from the overlying seawater and/or release from Mg calcites and clay minerals, which obviously places severe limits on the amounts of dolomite that can be formed. Microbial/organogenic dolomites may act as nuclei for later, more pervasive dolomitization during burial.

Hyposaline Environments and the Mixing Zone Model

Hyposaline environments are those with salinities below that of normal seawater (35 g l^{-1}). These environments include coastal and inland freshwater/seawater mixing zones, marshes, rivers, lakes, and caves. Virtually all hyposaline environments are near-surface to shallow burial diagenetic settings at depths of less than about 600–1000 m.

One hyposaline environment, the coastal freshwater/seawater mixing zone (often simply called mixing zone), has given rise to one of the oldest and most popular models, i.e., the 'mixing zone model' (also called the Dorag model) for dolomitization. However, the mixing model has been overrated with regard to its potential to form massive dolostones. Not a single location in the world has been shown to be extensively dolomitized in a freshwater/seawater mixing zone, in recent or in ancient carbonates, and several lines of evidence indicate that massive dolomitization in mixing zones is so unlikely as to be virtually impossible. Rather, mixing zones tend to form only very small amounts of dolomite, commonly along with substantial dissolution porosity, up to the scale of caves. A striking example is the vast cave system with essentially no dolomite generated by mixing zone diagenesis along the coastline of the Yucatan Peninsula, Mexico. The main role of coastal mixing zones in dolomitization may be as a hydrological pump for seawater dolomitization, rather than a geochemical environment favourable for dolomitization.

Most true mixing zone dolomites are petrologically and geochemically distinct. The crystals tend to be relatively clear, planar-e or planar-s, stoichiometric, well-ordered rhombs. However, some mixing zone dolomite is protodolomite. Crystal sizes commonly range from 1 to 100 μm, but may reach several millimetres in some cases. Most mixing zone dolomite occurs as cements in microscopic interstices and macroscopic voids, moulds, vugs, and caverns, and

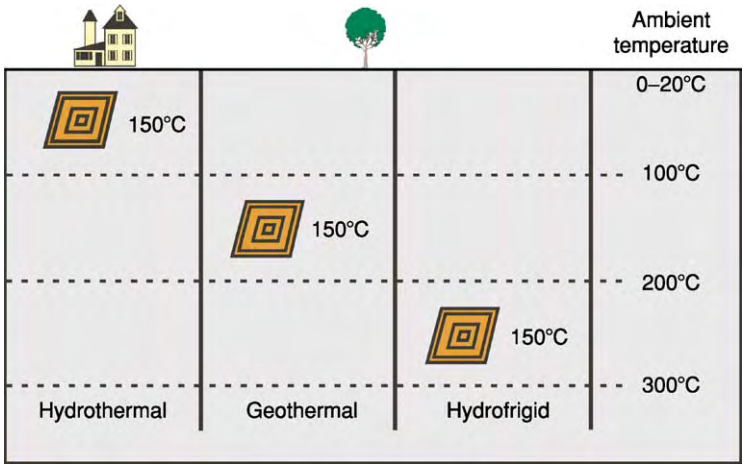


Figure 15 Hydrothermal, geothermal (formed in thermal equilibrium with the surrounding rocks), and hydrofrigid mineral formation. Reproduced with permission from Machel HG and Lonnee J (2002) Hydrothermal dolomite a product of poor definition and imagination. *Sedimentary Geology* 152: 163 171.

subordinately as a replacement. Alternating generations or growth zones of calcite/dolomite are common in coastal mixing zones with rapid and cyclical changes of salinity.

Hypersaline Environments and the Reflux and Sabkha Models

Hypersaline environments have salinities greater than that of normal seawater and are widespread at latitudes of less than about 30°, while some occur at even higher latitudes. Hypersaline environments thus defined include the so-called mesohaline (also called penehaline) environments, which are mildly hypersaline, i.e., between normal seawater salinity ($35\,g\,l^{-1}$) and that of gypsum saturation (about $120\,g\,l^{-1}$). In all these types of environment, dolomite is formed from water whose salinity is controlled by surface evaporation, that is, in near-surface and shallow burial diagenetic settings.

The (evaporative) reflux model of dolomitization was originally proposed for seawater evaporated beyond gypsum saturation in lagoonal and shallow marine settings on a carbonate platform behind a barrier, such as a reef. Surficial water circulation on the platform is severely restricted because of the barrier, which leads to evaporation and a landward salinity gradient behind the barrier. The evaporated seawater flows downwards into and seaward through the platform sediments because of its increased density (active reflux), thereby dolomitizing the penetrated sediments. Platforms can be penetrated by mesohaline reflux to depths of several hundred metres. Furthermore, numerical modelling predicts that 'latent reflux' flows the a platform after the evaporative generation of brine ceases at the platform top, which can be expected after a significant rise in sea-level with concomitant freshening of the waters on the platform. However, a platform can be dolomitized completely only if it has very high permeability and does not contain aquitards (such as shale or evaporite layers), and if reflux is permitted to persist for a very long time. The published literature provides several examples of localities that probably were dolomitized by evaporative reflux, including the Permian carbonates of west Texas and New Mexico, and the peritidal Jurassic carbonates of Gibraltar. Whether active or latent, all refluxing brines exit or even landward of the platform margin, which confines reflux dolomitization to the platform interior.

The sabkha model of dolomitization is hydrologically and hydrochemically related to the reflux model. Sabkhas are intertidal to supratidal deflation surfaces that are flooded episodically. In the sabkha of the Trucial Coast of Abu Dhabi, the type location of the sabkha model for dolomitization, Mg for dolomitization is supplied synsedimentarily/penecontemporaneously by seawater that is propelled periodically onto the lower supratidal zone and along remnant tidal channels by strong onshore winds. The seawater has normal to slightly elevated salinity (up to about $38\,g\,l^{-1}$), but becomes significantly evaporated beyond gypsum saturation on/within the supratidal flats, through which it refluxes via its increased density, similar to flow in the reflux model. Sabkha dolomite appears to form via evaporative pumping in a narrow (1–1.5 km) fringe next to the strandline, and in the flooded tidal channels that extend more landward. In

the Abu Dhabi sabkha, the best dolomitized parts contain 5 to about 65 wt.% protodolomite. Dolomite forms as a cement and aragonite is replaced, but lithification does not occur, or only partially. Furthermore, dolomitization is restricted to the upper 1–2 m of the sediments, and appears to be most intense where the pore waters become chemically reducing, which leads to enhanced carbonate alkalinity via sulphate reduction and/or microbial methanogenesis. Therefore, sabkha dolomites are texturally and geochemically similar to organogenic dolomites in some respects, i.e., they tend to form as protodolomite and may have reduced carbon isotope ratios. In most respects, the Abu Dhabi sabkha appears to be a good recent analogue for dolomitization in many ancient intertidal to supratidal flats, such as landward of the famous Permian carbonates of Texas and New Mexico. Rather than forming reservoir rocks, these dolostones – including the associated evaporites – generally form tight seals for underlying hydrocarbon reservoirs. More generally, sabkhas and similar intertidal to supratidal environments in more humid climates typically form small quantities of fine crystalline protodolomite in thin beds, crusts, or nodules, either within the upper 1–2 m of sediment, or at the sediment surface. Repeated transgressions and regressions may stack such sequences upon one another to cumulative thicknesses of several tens of metres.

Seawater Dolomitization

Most postdepositional formation of massive dolostones probably results from 'seawater dolomitization'. There are a group of models of seawater dolomitization, whose common denominator is seawater as the principal dolomitizing fluid, and which differ in hydrology and/or depth and timing of dolomitization. All dolomites belonging to this group are postdepositional.

The Cenozoic dolostones of the Bahama platform, often used as an analogue for older dolomitized carbonate platforms elsewhere, can be considered the type location for seawater dolomitization. Extensive petrographical and geochemical data indicate that seawater and/or chemically slightly modified seawater was the principal agent of dolomitization in the Bahama platform at shallow to intermediate depths and commensurate temperatures of less than about 60°C.

The hydrology of and during seawater dolomitization is still very much contested. Various hydrological systems have been invoked to drive the large amounts of seawater needed for pervasive dolomitization through the Bahama platform, i.e., thermal convection, a combination of thermal seawater convection and reflux of slightly evaporated seawater derived from above, or seawater driven by an overlying freshwater/seawater mixing zone during partial platform exposure, possibly layer by layer in several episodes.

The Bahamas' dolostones actually represent a hybrid with regard to the traditional, conventional classifications of models. Petrographical and geochemical data indicate that seawater was the principal dolomitizing agent, yet thermal convection, as a hydrological system and drive for dolomitization, is better classified under the burial (subsurface) models discussed below. Analogously, the regionally extensive Devonian dolostones in Alberta, western Canada, are also a hybrid with regard to the conventional dolomitization models. These Devonian dolostones probably formed at depths of 300–1500 m at temperatures of about 50–80°C from chemically slightly modified seawater, and have been classified as burial dolostones. Another example is represented by the regionally extensive dolostones of the Carboniferous of Ireland, which are petrographically and geochemically very similar to the Devonian dolostones of Alberta, and whose genesis has been interpreted in an analogous manner. In all cases, the hydrology that facilitated dolomitization is unclear, with thermal convection, reflux, compaction, tectonic expulsion, or a combination thereof, as alternatives. The regionally extensive dolostones of the Cretaceous Soreq Formation in Israel represent a Mesozoic example of this type of dolomitization. These Palaeozoic and Mesozoic dolostones can be (re-)classified along with the Cenozoic Bahama dolostones as 'seawater dolomites'. This classification dilemma arises from the historical evolution of our understanding of these dolostones, rather than invalidating the earlier 'burial' interpretations.

Intermediate to Deep Burial (Subsurface) Environments and Models

Burial (subsurface) environments are those removed from active sedimentation by burial, and in which the pore fluid chemistry is no longer entirely governed by surface processes, i.e., where water–rock interaction has modified the original pore waters to a significant degree, or where the fluid chemistry is dominated by subsurface diagenetic processes. The textures, porosities, and permeabilities of dolostones formed in intermediate and deep burial settings are variable. These textures in themselves are not indicative of the depth of burial. However, three specific textural characteristics may be used to indicate considerable burial: dolomites cross-cut by stylolites suggest burial of at least 600 m (stylolites in dolostones appear to require at least 600 m of burial); the absence of planar

crystals suggests temperatures of formation or recrystallization in excess of the critical roughening temperature of about 50°C; and the presence of saddle dolomite suggests temperatures of formation in excess of about 80°C.

All burial (subsurface) models for dolomitization essentially are hydrological models. They differ mainly in the hydrological drives and direction(s) of fluid flow. Four main types of fluid flow take place in subsurface diagenetic settings: (1) compaction flow; (2) thermal convection; (3) topography driven flow; and (4) tectonically driven flow. Combinations of these flow regimes and associated fluid compositions are possible under certain circumstances.

The oldest burial model of dolomitization is the compaction model, according to which seawater and/or its subsurface derivative(s), that were buried along with the sediments, are pumped through the rocks at several tens to several hundreds of metres as a result of compaction dewatering.

The compaction model in its original form was never especially popular because burial compaction can only generate fairly limited amounts of dolostone due to the limited amounts of expelled water. However, despite this mass balance constraint, the compaction model remains a viable alternative for burial/subsurface dolomitization where funnelling of the compaction waters is/was possible.

Thermal convection is driven by the temperature gradient prevailing across sedimentary strata, which is vertical in most geological situations, except in cases of vigorous advection, igneous intrusions, or in the proximity to plate boundaries and/or orogenic fronts, all of which can 'distort' the normal subvertical temperature gradient. Where the temperature gradient and average rock permeability are high enough, convection cells may become established. In principle, there are two types of convection, i.e., open and closed, although mixed cases are possible. Open convection cells (also called half-cells) may form in carbonate platforms that are open to seawater recharge and discharge laterally and at the top, respectively. Numerical modelling has shown that the magnitude and distribution of permeability are the most important parameters governing flow and dolomitization, and that this type of convection can be active to a depth of about 2–3 km, provided that the sequence does not contain effective aquitards, such as (overpressured) shales or evaporites. The amounts of dolomite that can be formed are theoretically very large, i.e., dolomite can be formed as long as convection is sustained, because Mg is constantly (re-)supplied from the surrounding seawater. However, even at a moderate width of only 40 km, complete dolomitization in a 2 km thick sequence takes about 30–60 million years, which is much longer than the time during which most carbonate platforms remain laterally open to seawater recharge. Hence, most carbonate platforms, even if subjected to thermal convection by seawater, would at best become only partially dolomitized during the time that they were open to seawater recharge.

Closed convection can occur, in principle, in any sedimentary basin over tens to hundreds of metres of thickness, provided that the temperature gradient is high enough relative to the permeability of the strata. As a rule of thumb, however, such convection cells will only be established, and capable of dolomitizing a carbonate sequence of interest, if a sequence is highly permeable and not interbedded with aquitards. Such conditions are rarely fulfilled in typical sedimentary basins, most of which do contain effective aquitards. Furthermore, even if closed thermal convection cells are established, the amounts of dolomite that can be formed are limited by the pre-convection Mg content of the fluids, even more so than in the case of compaction flow, as no new Mg is supplied to the system and 'compaction funnelling' is not possible. Therefore, extensive, pervasive dolomitization by closed cell thermal convection is highly unlikely.

Convection cells invariably have rising limbs that penetrate the overlying and cooler strata, linking thermal convection to hydrothermal dolomitization. There are well-documented examples of hydrothermal dolomite on a local and regional scale. Most cases of hydrothermal dolomitization are rather small and restricted to the vicinity of faults and fractures and/or localized heat sources. One striking case of this type is the Pb–Zn mineralized Navan dolomite plume in Ireland, and another is a dolomitized plume in the Latemar in the Italian Alps. There are also some cases of larger scale, even regionally extensive hydrothermal dolomitization, such as the Middle Devonian Presqu'ile barrier, which forms an aquifer in northwestern Canada that contains abundant saddle dolomite as a replacement and as a cement, including MVT-type mineralization near the discharge area at Pine Point. Texturally, most true hydrothermal dolomite is saddle dolomite.

Topography driven flow takes place in all uplifted sedimentary basins that are exposed to meteoric recharge. With time, topography can drive enormous quantities of meteoric water through a basin, often concentrated by water–rock interaction (especially salt dissolution), and preferentially funnelled through aquifers. Volumetrically significant dolomitization can only take place, however, where the meteoric water dissolves enough Mg *en route* before encountering limestones. This does not appear to be common.

Another type of flow that has been suggested to result in pervasive dolomitization is tectonically driven squeegee-type flow. In this type of flow system, metamorphic fluids are expelled from crustal sections affected by tectonic loading so that basinal fluids are driven towards the basin margin. Metamorphic fluids could be injected into compaction and/or topography driven flow, with attendant fluid mixing. However, it is doubtful whether tectonically induced flow, or the related fluid mixing, can form massive dolostones. Diagenetic studies on this type of flow system suggest that the fluxes are low and short lived.

Secular Distribution of Dolostones

The relative abundance of dolostones that originated from the replacement of marine limestones appears to have varied cyclically through geological time, commonly referred to as secular variation. Early data suggested that dolomite was most abundant in rocks of the Early Palaeozoic systems and decreased in abundance with time. Relatively recent reassessments of the dolomite distribution throughout time revealed two discrete maxima of 'significant early' dolomite formation (massive early diagenetic replacement of marine limestones) during the Phanerozoic, i.e., the Early Ordovician/Middle Silurian and the Early Cretaceous.

Various explanations have been proposed for this phenomenon, i.e., that periods of enhanced early dolomite formation were related to or controlled by plate tectonics and related changes in the compositions of the atmosphere and seawater, such as an increased atmospheric CO_2 level, high eustatic sea-level, low saturation state of seawater with respect to calcite, changes in the marine Mg/Ca ratio, or low atmospheric O_2 levels that coincided with enhanced rates of bacterial sulphate reduction. It appears possible that a combination of two or more of these factors was involved. Perhaps the most elegant explanation is that the secular dolomite variations are the result of the lengthy induction period for dolomite formation that was observed in laboratory experiments. Marine carbonates in prolonged contact with seawater may be dolomitized because they remained in contact with the dolomitizing solution (seawater) long enough to exceed the induction period. On the other hand, undolomitized carbonates were not in contact with seawater for long enough, and metastable precursors to dolomite that may have formed were destroyed by freshwater diagenesis during intervening periods of exposure. The secular variations in marine dolomitization may thus reflect periods of seawater contact longer or shorter than the induction period.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Diagenesis, Overview**. **Minerals:** Carbonates. **Petroleum Geology:** Production. **Sedimentary Environments:** Carbonate Shorelines and Shelves. **Sedimentary Rocks:** Mineralogy and Classification; Evaporites; Limestones.

Further Reading

Allen JR and Wiggins WD (1993) *Dolomite Reservoirs Geochemical Techniques for Evaluating Origin and Distri bution. AAPG Continuing Education Course Note Series, 36*. Town: American Association of Petroleum Geologists.

Braithwaite C, Rizzi G, and Darke G (eds.) (2004) *The Geometry and Petrogenesis of Dolomite Hydrocarbon Reservoirs. Special Publication of the Geological Society* (in press).

Budd DA (1997) Cenozoic dolomites of carbonate islands: their attributes and origin. *Earth Science Reviews* 42: 1 47.

de Dolomieu D (1791) Sur un genre de Pierres calcaires très peu effervescentes avec les Acides, & phosphorescentes par la collision. *Journal de Physique* 39: 3 10. Transla tion with notes of Dolomieu's paper reporting his discov ery of dolomite by: Carozzi AV and Zenger DH (1981) *Journal of Geological Education* 29: 4 10.

Gregg JM and Sibley DF (1984) Epigenetic dolomitiza tion and the origin of xenotopic dolomite texture. *Journal of Sedimentary Petrology* 54: 908 931.

Hardie LA (1987) Dolomitization: a critical view of some current views. *Journal of Sedimentary Petrology* 57: 166 183.

Land LS (1985) The origin of massive dolomite. *Journal of Geological Education* 33: 112 125.

Luo P and Machel HG (1995) Pore size and pore throat types in a heterogeneous Dolostone reservoir, Devonian Grosmont Formation, Western Canada Sedimentary Basin. *American Association of Petroleum Geologists Bulletin* 79: 1698 1720.

Machel HG (1997) Recrystallization versus neomorphism, and the concept of 'significant recrystallization' in dolomite research. *Sedimentary Geology* 113: 161 168.

Machel HG (2004) Concepts and models of dolomitization a critical reappraisal. In: Braithwaite C, Rizzi G, and Darke G (eds.) *The Geometry and Petrogenesis of Dolo mite Hydrocarbon Reservoirs. Special Publication of the Geological Society* (in press).

Machel HG and Mountjoy EW (1986) Chemistry and environments of dolomitization a reappraisal. *Earth Science Reviews* 23: 175 222.

Mazzullo SJ (1992) Geochemical and neomorphic alter ation of dolomite: a review. *Carbonates and Evaporites* 6: 21 37.

Morrow DW (1982a) Diagenesis 1. Dolomite Part 1: The chemistry of dolomitization and dolomite precipitation. *Geoscience Canada* 9: 5 13.

Morrow DW (1982b) Diagenesis 2. Dolomite Part 2: Dolomitization models and ancient dolostones. *Geoscience Canada* 9: 95 107.

Morrow DW (1999) Regional subsurface dolomitization: models and constraints. *Geoscience Canada* 25: 57 70.

Nordeng SH and Sibley DF (1994) Dolomite stoichiometry and Ostwald's step rule. *Geochimica et Cosmochimica Acta* 58: 191 196.

Pray LC and Murray RC (eds.) (1965) *Dolomitization and Limestone Diagenesis A Symposium. Society of Economic Paleontologists and Mineralogists Special Publication 13*. Town: Society of Economic Paleontologists and Mineralogists.

Purser B, Tucker M, and Zenger D (eds.) (1994) *Dolomites A Volume in Honour of Dolomieu: Special Publication 21*. Town: International Association of Sedimentologists.

Sibley DF and Gregg JM (1987) Classification of dolomite rock textures. *Journal of Sedimentary Petrology* 57: 967 975.

Van Tuyl FM (1914) *The Origin of Dolomite. Annual Report 1914,* vol. XXV, pp. 257 421. Town: Iowa Geological Survey.

Wright WR (2001) Dolomitization, fluid flow and mineralization of the Lower Carboniferous rocks of the Irish Midlands and Dublin Basin. Unpub. Ph.D. thesis, University College Dublin, Belfield, Ireland, 407 p.

Zenger DH, Dunham JB, and Ethington RL (eds.) (1980) *Concepts and Models of Dolomitization. Society of Economic Paleontologists and Mineralogists Special Publication 28*. Town: Society of Economic Paleontologists and Mineralogists.

Evaporites

A C Kendall, University of East Anglia, Norwich, UK

Deposits Produced by the Evaporation of Seawater

Seawater is considered to be the major or the only feedstock capable of generating large bodies of evaporite. All deposits of potash salts are associated with large basin-central evaporites and, consequently, are believed by most to have been formed by the evaporation of seawater. The problem with this marine origin is that the chemical and mineralogical characters of most potash deposits depart significantly from those that would be expected from simple seawater evaporation. If the marine origin is correct, then other processes must have been involved to cause the differences.

Seawater becomes progressively more concentrated as it evaporates until it is supersaturated with respect to a particular mineral phase, which then precipitates. Precipitation of a salt preferentially extracts chemical components from the seawater-derived brine, altering its overall composition. Initially, calcium combines with bicarbonate, but, after seawater has been concentrated approximately 3.5 times, gypsum ($CaSO_4 \cdot 2H_2O$) saturation is reached and calcium and sulphate are extracted from the brine. Seawater contains abundant sulphate, and, after the greater part of the calcium has been extracted (as carbonates and as gypsum), fully two-thirds of the original sulphate remains in the brine. At this stage, and in all subsequent stages of evaporation, a marine-derived brine will be impoverished in calcium and should contain abundant sulphate.

The next mineral to precipitate by continued seawater evaporation is halite (NaCl), and this extracts sodium and chloride from the brine. At about 60 times seawater concentration, the sulphate remaining in the brine should begin to be removed in the form of various magnesium sulphate minerals. Only after considerable amounts of sulphate have been eliminated from the brine will the next mineral – carnallite (hydrated magnesium and potassium chloride) – precipitate. Finally, at the last stages of concentration, the mineral bischoffite ($MgCl_2$) is precipitated.

Typical Composition of Evaporite Deposits

Only about 10% of all potash-bearing evaporites contain significant quantities of magnesium sulphate, which would be expected from simple seawater evaporation. Of these 10%, all differ from direct seawater precipitation sequences in having different magnesium sulphate minerals in different amounts from those expected. These differences can be explained in two ways. First, most magnesium sulphates precipitated during experimental seawater evaporation are highly unstable hydrous phases. These alter to less hydrous minerals upon burial. Second, during evaporation, the concentrated brines may react with previously precipitated calcium sulphate to form the mineral polyhalite. This reaction removes sulphate and potassium from the brine, so changing its composition. Further evaporation of this modified brine is capable of generating the mineral sequences found in 10% of potash salt deposits.

The majority of potash salts, however, differ substantially from those expected to result from seawater

evaporation. After halite has precipitated, the next mineral to appear is carnallite (hydrous magnesium and potassium chloride) or sylvite (potassium chloride); usually there are no magnesium sulphates. The uppermost parts of some deposits, especially from the Atlantic-marginal Cretaceous salt basins of Brazil and Gabon, also contain the mineral tachyhydrite (hydrous calcium chloride). This highly soluble salt (which dissolves in atmospheric moisture) must represent the final evaporative stages of a brine. However, if this brine were concentrated seawater, then all the available calcium should have been extracted at a much earlier evaporative stage, during gypsum and early-stage halite precipitation.

Sylvite should not precipitate during simple seawater evaporation. Some sylvite can be explained as a later alteration product of carnallite, but in other cases textural evidence indicates that sylvite is a primary mineral. The absence of magnesium sulphates cannot be explained by later diagenetic changes (for instance, converting them to carnallite) because textural evidence also suggests that much carnallite is primary.

The Missing Sulphate

The brines that precipitated potash deposits low in magnesium sulphate did not lack magnesium (carnallite and bischoffite contain this element), and so the problem is to explain why marine-derived evaporites are impoverished in sulphate in their later evaporative stages.

Explanations of this missing marine-derived sulphate are unsatisfactory. One hypothesis suggests that sulphate is removed from the brine by the addition of river water containing additional calcium. This calcium strips the brine of its remaining sulphate by the precipitation of additional gypsum. Simple calculations indicate that the amount of river water needed would be enormous and more than enough to dilute the brines so that no evaporites would form in the first place.

A commonly proposed explanation is that sulphate is removed by the activity of sulphate-reducing bacteria. The sulphate is reduced to hydrogen sulphide, which is then lost to the atmosphere. Hite convincingly argued, by analogy with Holocene environments, that sulphate reduction would be confined to the uppermost metre of sediment, and so the sulphate-reducing capabilities of evaporite basins would be limited. His calculations showed that bacteria would be unable to remove all the seawater sulphate and, furthermore, that the amount of organic carbon required to reduce the marine sulphate could not be supplied, even if evaporite basins were extraordinarily productive.

The problem of the missing sulphate exists only if seawater was the original feedstock. It is clear from the presence of tachyhydrite in some sequences that either the seawater was substantially modified or some other water was evaporated to generate the potash salts. In essence, no brine containing more sulphate than calcium (including all modern seawater-derived brines) can generate tachyhydrite and evaporites deficient in magnesium sulphate, whereas these sequences can be generated if waters are used where more calcium than sulphate is present. By definition, any concentrated water that has excess calcium (calcium $>$ sulphate $+$ bicarbonate) is a calcium chloride brine.

Hardie believed that evaporites that are impoverished in magnesium sulphate formed by the evaporation of calcium chloride brines that were generated in rift or transtensional basins – basins with high heat flows and active hydrothermal groundwater circulations. Calcium chloride brines are being expelled today in these types of basin. Groundwater circulates deep in the crust and reacts with hot host rocks. Hydration of host-rock minerals removes water, concentrating the groundwater into a brine. The hot brine reacts with calcium-bearing minerals, commonly feldspars (exchanging sodium for calcium), and some of the expelled calcium reacts with any groundwater sulphate to precipitate anhydrite, thus stripping the brine of its contained sulphate. The final product is a brine that is depleted in sulphate and enriched in calcium. When heated, these brines become buoyant and may be expelled to the surface, where they evaporate.

Evaporites as Hydrothermal Deposits in Rift Basins

Rift and transtensional basins are ideal locations for evaporites to form. Commonly they are isolated or have very restricted access to the ocean. Even if they are not located in arid climatic zones, they may develop evaporites by virtue of uplifted rift shoulders or transpression ridges causing orographic aridity. Calcium chloride waters entering such basins as hot springs can evaporate and generate evaporite sequences without magnesium sulphate minerals. Many evaporites deficient in magnesium sulphate are located in present-day and ancient rift basins.

When hydrothermal waters are the only feedstock, the resultant evaporites may be entirely deficient in calcium and other sulphates. Evaporite sequences should have little, if any, gypsum and anhydrite. It is also possible, however, that the main water in the basin is seawater but that this is substantially modified by the addition of relatively small volumes of

hydrothermal brines, which nevertheless carry large amounts of dissolved materials. Hardie calculated that modern seawater could be modified by the addition of only just over 3% Salton Sea brine into water that would not precipitate evaporites containing magnesium sulphate.

Evaporites in Non-Rift Basins

Hardie's hydrothermal brine explanation is not convincing for some evaporite sequences because they occur in non-rift basins. Kendall provided an alternative explanation for some of these evaporites. Desiccation of large evaporite basins produces large and deep depressions. This induces a hydrodynamic drive, which causes subsurface waters to migrate into the basins, where they evaporate. Dolomitization of limestones by migrating formation waters with seawater-like compositions would release calcium and generate calcium chloride waters. Basin desiccation thus provides both the drive that allows formation waters to enter the basin and a mechanism to generate waters with more calcium than sulphate. Middle Devonian evaporites in western Canada provide evidence to support this desiccation–drive model. Where calcium-rich formation waters entered the Devonian evaporite basin, spring-associated carbonates were precipitated, and there was mixing with seawater-derived brines that had already precipitated gypsum. The addition of spring-water calcium stripped the remaining sulphate from the basin brine, precipitated anomalous concentrations of calcium sulphate far into the evaporite basin, and led to the formation of a sulphate-depleted brine, which may have caused later potash salts deficient in magnesium sulphate to form in the basin.

Halite-saturated brines, refluxing beneath large evaporite basins, react with all types of sediments (not just limestones) by exchanging sodium for calcium, to generate calcium chloride brines. The main problem in understanding how these deep dense brines could form potash salts is to explain how the brines are transported to the surface. This could occur at times of basin inversion, by heating (creating buoyant hydrothermal brines) or by evaporative draw into a later evaporite basin.

Past Composition of Seawater

A more exciting alternative that explains the seemingly anomalous compositions of most ancient potash salts is that seawater compositions were substantially different in the past. Secular variations in the distributions of magnesium sulphate and evaporites deficient in magnesium sulphate are in phase with better-known variations in the mineralogy of ancient marine carbonates. These secular variations can be attributed to changes in the major-ion composition of seawater over time.

A model that explains how seawater can change over time was used to predict periods when aragonite and evaporites containing magnesium sulphate are dominant, and episodes when calcite and evaporites deficient in magnesium sulphate are favoured. Seawater chemistry is controlled by steady-state mixing of river water and mid-ocean ridge hydrothermal brines (coupled with calcium carbonate and silica precipitation). Relatively small changes in mid-ocean ridge fluxes cause significant changes in magnesium : calcium, sodium : potassium, and chloride : sulphate ratios in seawater. Such changes are believed to be responsible for variations in the primary mineralogies of marine evaporites and carbonates. Variations in mid-ocean ridge flux correspond to variations in the production rate of oceanic crust (seafloor spreading rate), and this can be estimated using various proxies, such as areas of ocean floor of different ages, the global sea-level curve, and granite-pluton emplacement rates.

Predictions of the variation in past seawater chemistry produced by variation in mid-ocean-ridge flux rates are in agreement with the known age distribution of primary marine carbonate and evaporite mineralogies. The coherence of the datasets strongly suggests that past variations in evaporite and carbonate mineralogy were largely caused by secular variations in seawater chemistry.

The idea that varying seawater chemistry can explain potash salt composition has been challenged. Holland *et al*. predict similar changes of seawater composition but of much smaller magnitude. They argue that an apparent near constancy of the level of potassium in seawater during the Phanerozoic (demonstrated by the compositions of brines trapped in ancient halites) supports this view: Hardie's model predicts significant changes in the sodium : potassium ratio. Instead, Holland *et al*. suggest that changes in past evaporite mineralogy are due to differences in the extent to which dolomitization of carbonate sediments occurred before or during seawater evaporation. During times of rapid seafloor spreading, sea-levels are higher and large carbonate platforms are more abundant. Changes in seawater chemistry (caused by increased mid-ocean ridge flux) coupled with increases in the extent of dolomitization of carbonate platforms are believed to be responsible for the formation of potash deposits that are impoverished in magnesium sulphate. This explanation resembles, in part, that suggested earlier by Kendall.

Variations in the chemistry of primary fluid inclusions from ancient halite deposits are significant. They also imply that seawater chemistry has changed significantly. Variations are in phase with inferred seafloor spreading rates, global changes in sea-level, and the primary mineralogies of ancient marine carbonates and evaporites. Of particular significance is the fact that inclusions in halites of the same age from different geological basins exhibit similar compositions. This suggests that the association with dolomitization (proposed by Holland *et al*.) is incorrect: more interbasin variation in the amount of dolomitization would be expected, resulting in a greater variation in the chemistry of fluid inclusions than that observed. It is surprising, however, that the question of whether or not variations in sodium : potassium ratios match model predictions was not addressed. More recent, unpublished, work suggests that Cretaceous and Permian seawaters were enriched in potassium and relatively depleted in sodium, as would be expected from the Hardie hypothesis.

See Also

Minerals: Sulphates. **Sedimentary Environments:** Lake Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification; Dolomites. **Tectonics:** Hydrothermal Activity; Hydrothermal Vents At Mid-Ocean Ridges; Rift Valleys.

Further Reading

Hanor JS (1996) Variations in chloride as a driving force in siliciclastic diagenesis. In: Crossey LJ, Loucks R, and Totten MW (eds.) *Siliciclastic Diagenesis and Fluid Flow: Concepts and Applications*, pp. 3 12. Special Publication 55. Tulsa: Society for Sedimentary Geology.

Hardie LA (1990) The roles of rifting and hydrothermal $CaCl_2$ brines in the origin of potash evaporites: a hypoth esis. *American Journal of Science* 290: 43 106.

Hardie LA and Spencer RJ (1990) Control of seawater composition by mixing of river waters and mid ocean ridge hydrothermal brines. In: Spencer RJ and Chou I M (eds.) *Fluid Mineral Interactions: A tribute to H P Eugster*, pp. 409 419. Special Publication 2. San Antonio: Geochemical Society.

Hite RJ (1983) The sulphate problem in marine evaporites. In: Schreiber BC (ed.) *Proceedings of the 6th Inter national Salt Symposium, Toronto*, pp. 217 230. Alexandria, VA: Salt Institute.

Kendall AC (1989) Brine mixing in the Middle Devonian of western Canada and its possible significance to regional dolomitization. *Sedimentary Geology* 64: 271 285.

Lowenstein TK and Spencer RJ (1990) Syndepositional origin of potash evaporites: petrographic and fluid inclu sion evidence. *American Journal of Science* 290: 1 42.

Ironstones

W E G Taylor, University of Lancaster, Lancaster, UK

Introduction

Ironstones have been critical to industry and industrial revolutions. They have been essential raw materials since the dawn of the Iron Age (about 700 BC). Without iron-rich deposits many of the manmade structures and utensils that we take for granted today – tall urban buildings, power pylons, bridges, ships, cutlery, hammers, saws, and the seemingly indispensable motor car – could not exist.

The point at which an ironstone deposit is considered to be an ore has changed considerably over the years and depends upon the particular economic, technological, social, and political circumstances at the time. Nowadays deposits need to have an iron content in excess of 60% by weight to be worked, whilst in the mid-twentieth century ironstones with 28% iron by weight were regularly extracted as ores. The quality of the potential ore, and in particular the proportion of phosphatic material, is an important factor that has to be considered in the mining of iron.

Initially, the availability of water power and the proximity of coal were the factors controlling production. The middle of the nineteenth century saw a change from coal-fired furnaces producing cast iron to the Bessemer process, which produced steel. Later in the same century, the open-hearth process and various refinements were developed. Each of these new production processes demanded ores of a particular quality.

Records of global production are scarce before the latter half of the twentieth century, and certainly in Europe much of the exploitation predates that century. In Great Britain the maximum annual output was in the order of 18 Mt (Milliontonnes) and occurred during two main periods – 1870–1890 and 1940–1945. In the former period the main type of ore extracted was from the blackband and claystone ironstones (see below) of the Carboniferous rocks of various coalfields, whilst in the latter period the

Figure 1 Ordovician ironstones, Betws Garmon, North Wales, UK (inset showing the steeply inclined stoping method of under ground mining for the deposit and a residual pillar of magnetite rich material).

ooidal ironstones of Northamptonshire and Lincolnshire were the dominant ores. Although Ordovician ooidal ironstones from North Wales were extracted until early part of the twentieth century (Figure 1).

The terms used to describe both the processes of ironstone formation and the ironstones themselves have been many and varied. Attempts to simplify and standardize the terminology have recently met with some success, mainly through the International Geological Correlation Programme (IGCP). For example, the terms 'Clinton' (from the Silurian Clinton Group of New York State, USA) and 'Minette' (from the Jurassic Minette oolitic ironstones of northeastern France and adjacent areas) as descriptions of ironstone types have proved to be unsatisfactory and have now fallen into disuse.

Definition

Largely because iron may invade and impregnate a wide range of rocks, defining what constitutes an ironstone has proved difficult. Exhortations to merge the nomenclatures of the various iron-rich deposits, such as the banded iron formations and ironstones, have been resisted on the basis that the mineralogy, petrology, and genesis of these deposits are distinct and separate. A precise definition of 'ironstone' was agreed only in the last decade of the twentieth century and stems from the work of the IGCP 277 (Phanerozoic Ooidal Ironstones). Ironstones are sedimentary rocks consisting of at least 15% iron by weight, which may be quoted as 19% FeO or 21% Fe_2O_3 or an equivalent admixture in a chemical analysis. They occur almost exclusively in the Phanerozoic Era and are distinguished from the mainly Precambrian banded iron formations (*see* **Sedimentary Rocks:** Banded Iron Formations) by their lack of both regular banding and chert and by their age: banded iron formations were produced when there was a deficiency of oxygen in the Earth's atmosphere. The ferruginization (iron enrichment) may be the result of either direct deposition or subsequent chemical changes.

Ironstone Mineralogy

The iron-ore minerals may be oxides (including goethite, haematite, and magnetite), carbonates (usually siderite), or silicates (normally berthierine or chamosite). They may be associated with other carbonate minerals, sulphides and/or phosphatic minerals.

Goethite – FeO(OH) – is commonly formed by oxidation during weathering. Also, in many ooidal ironstones, it can result from the oxidation of berthierine; the two minerals may be found intermixed, often in alternate concentric layers, in ooids. Limonite was formerly thought to be a distinct mineral with the composition $2Fe_2O_3 \cdot 3H_2O$ but is now considered to have a variable composition (and properties) and to consist of several iron hydroxides (commonly goethite) or a mixture of iron minerals. Generally, it occurs as a secondary alteration product. Haematite – Fe_2O_3 – can be an important mineral in some ironstones, where it is usually formed as a late-stage

diagenetic product of the alteration of goethite. Experimental synthesis indicates that this transformation occurs at a temperature above 80°C and at a depth of about 2 km. Magnetite – Fe_3O_4 – normally forms during the low-grade metamorphism of ironstones, although the mineral has been reported from unmetamorphosed deposits in Libya.

Siderite – $(Fe,Mg,Ca,Mn)CO_3$ – is a very important mineral in ironstones. It is the only iron-bearing mineral in many claystone and blackband ironstones. Substitution of magnesium, calcium, or manganese for iron in the structure of siderite has been hypothesized to be related to the environment of formation. Substitution appears to have been greatest in marine sediments and in those ironstones formed during the later diagenetic stages of non-marine sediments.

Berthierine – $(Fe^{2+},Fe^{3+},Al,Mg,Mn)_2(Si,Al)_2O_5(OH)_4$ – is a 0.7 nm repeat serpentine. Reported variations in the chemical composition may reflect the analytical difficulties of dealing with very fine-grained samples. The formation of berthierine in ironstones is the subject of some debate and will be considered later. Chamosite – $(Fe_5^{2+},Al)(Si_3Al)O_{10}(OH)_8$ – is a 1.4 nm repeat chlorite with a very similar chemical composition to berthierine. Experimental work has shown that berthierine may be transformed into chamosite at a temperature of about 150°C and a depth of about 3 km. The phyllosilicate glauconite – $(K,Na)(Al,Fe^{3+},Mg)_2(Al,Si)_4O_{10}(OH)_2$ – is generally thought to be restricted to marine environments and occurs in some ironstones.

Other carbonate minerals, such as calcite, aragonite, dolomite, and ankerite – $Ca(Mg,Fe)(CO_3)_2$ – may be particularly common in ironstones both as a constituent of the cement and as discrete bioclasts. Phosphate minerals, such as francolite (carbonate fluorapatite) and vivianite – $Fe_3(PO_4)_2 \cdot 8H_2O$ – can be major components of ooidal ironstones. They can be a detrimental contributory factor to the viability of a deposit as an ore, particularly since, in most cases, the mineral grains are very small and difficult to separate.

Types of Ironstones

Extensive use of high-precision microscopy and analytical techniques has allowed detailed insights into the composition and formation of ironstones. A formal classification of ironstones has not yet been universally accepted, but three distinctive categories have been recognized (Figure 2).

Blackband Ironstones

Blackband ironstones are, typically, fossiliferous sapropel-rich (usually with an organic content in excess of 10%) finely laminated sideritic ironstones. Although non-laminated types are known, more frequently they are formed of alternating siderite- and organic-rich laminae. They are found almost exclusively above coal seams in a lacustrine parasequence

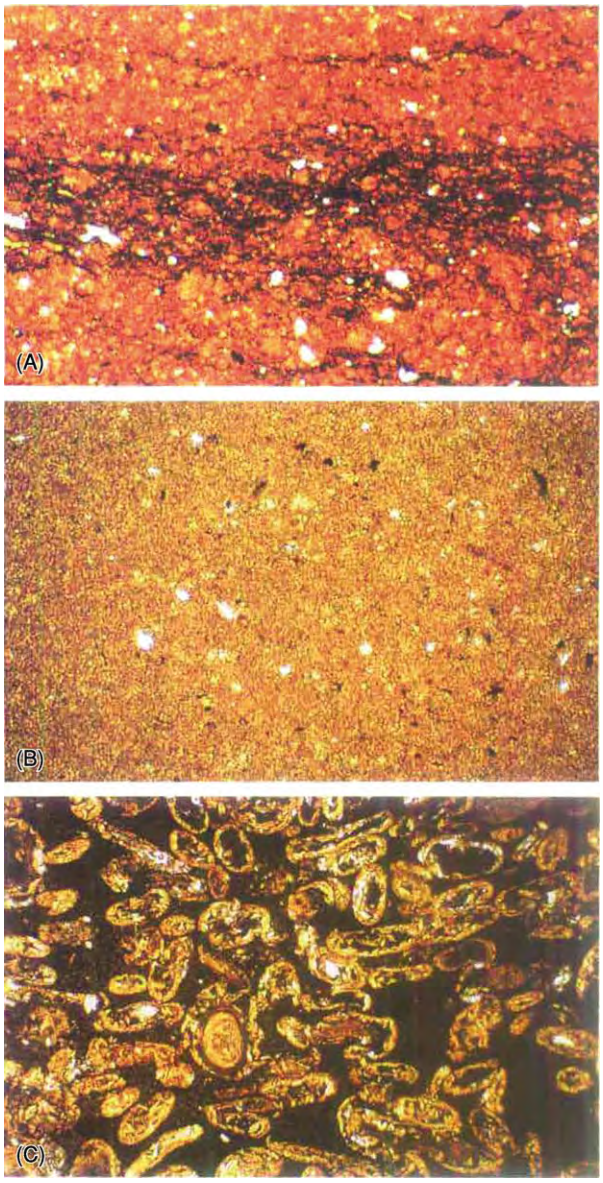


Figure 2 Typical photomicrographs of the three categories of ironstone. (A) Blackband ironstone. Note the dark organic rich lamination in a mainly sideritic matrix. Plane polarized light; horizontal dimension is 1.3 mm. (B) Claystone ironstone. Note the lack of organic material in the mainly sideritic matrix. Plane polarized light; horizontal dimension is 1.3 mm. (C) Ooidal ironstone. Ooids, showing selective replacement by siderite and phosphatic minerals, are set in a berthierine rich mudstone matrix. Plane polarized light; horizontal dimension is 5.2 mm (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446 489. London: Chapman & Hall, plates 14b, 14a and 14e).

with mudstone and seat earth deposits (Figure 3). Palaeontological and mineralogical evidence indicates that these ironstones were formed during freshwater inundation. Unlike non-marine clayband ironstones, there is an absence of early diagenetic pyrite, and the occurrence of coal balls (calcite–pyrite concretions) is an indication of marine incursion. The ironstones typically form thin (less than 10 cm) sheets of less than 10 km^2 extent and often change laterally into limestones with similar textural characteristics. Bog iron ores, which occur as lenses of ferruginous concretions within peat deposits, are thought to be the modern analogues of blackband ironstones.

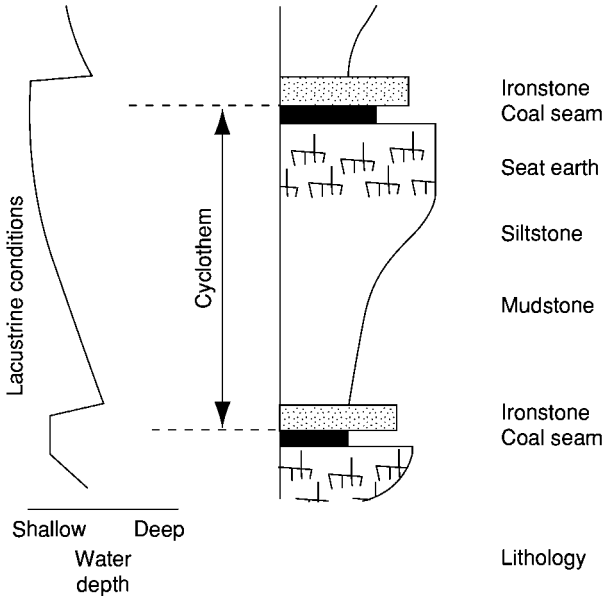


Figure 3 Idealized stratigraphical column for blackband iron stones showing relative water depths of sedimentation, not to scale.

Claystone Ironstones

Claystone or clayband ironstones have been the basis of the steel industry in many industrialized countries, largely because of their association with coalfields. Essentially, they are accumulations of iron carbonates (usually siderite) that have replaced the non-marine shales of coal-measure cyclothems (parasequences) and occur as either thin sheets or, more commonly, layers of concretions (Figure 4). Occasionally these sheets may extend over several hundred square kilometres. Normally, each concretion is unlaminated and does not contain high amounts of organic material, and the siderite grains are usually microscopic or sub-microscopic in size (less than 10 μm). Marine claystone ironstones are predominately rich in ankerite with pyrite, and production of siderite is suppressed. Irregularly shaped sphaerosiderites (ball ironstones), which usually occur at the base of palaeosols, are composed of siderite cement in the form of distributed spherulites (0.5–1 mm in diameter).

Ooidal Ironstones

Ooidal ironstones are characterized by the presence of ooids and/or pisoids and are very diverse, with a

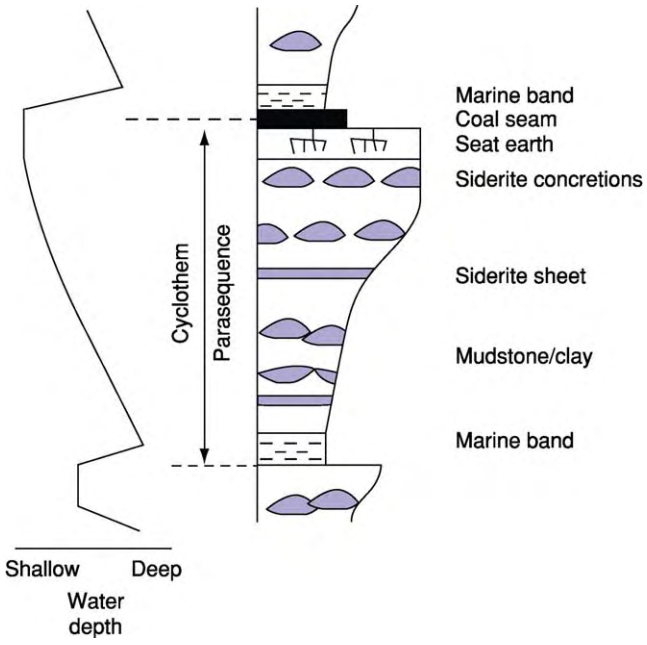


Figure 4 Idealized stratigraphical column for claystone ironstones showing relative water depths of sedimentation, not to scale.

wide range of mineralogy, textures, and chemical compositions. Because they possess oolites, shell fragments, and mud matrices in various admixtures similar to limestones (*see* **Sedimentary Rocks:** Limestones), most researchers in the field use the petrographic terminology advocated by Young to describe and classify ooidal ironstones. Most ooidal ironstones are less than a metre in thickness and are laterally persistent over approximately 100 km^2. A few deposits are in excess of 15 m thick (e.g. the Gara Djebilet Ironstone in Algeria). Although an idealized stratigraphical model for this type of ironstone consists of an upward shoaling sequence from black shales at the base, through progressively coarser deposits, to the ironstone at the top (Figure 5), in practice there are many deviations from this standard. Ironstones develop during periods of reduced sediment input (starvation), with abundant bioturbation, and often exhibit signs of storm reworking to form tempestites. The earliest-formed minerals are usually iron oxides and silicates. Iron-rich carbonates may be generated subsequently, often during early diagenesis.

Modern Examples of Ironstone Development

Bog iron ores are found associated with peat deposits in swampy conditions. Typically they contain hydrated ferric-oxide and manganese-oxide cements but, below the water table, they may be cemented by siderite. It has been suggested that microbial activity in tropical climates particularly promotes the direct precipitation of siderite.

A possible present-day analogue of ancient ooidal ironstones appears to be the verdine facies. In this facies, iron-rich aluminous green clay minerals replace bioclasts and pellets. Ferruginous peloids, in many cases altered faecal pellets, are known to be forming today in sediments deposited in front of equatorial deltas, such as those on the continental shelves off Senegal, Guinea, Nigeria, Gabon, Sarawak, and east Kalimantan. Present-day examples of ferruginous ooid accumulation are rare. In the interior of Africa, along the southernmost parts of Lake Malawi, amorphous ferric-oxide ooids have been found associated with geothermal springs, and, in the brackish open water of southern Lake Chad, goethitic brown ooids are being formed. In the shallow seas of northern Venezuela, berthierine-rich green-brown muddy ooidal sediments with peloids have been discovered.

Environment of Deposition and Subsequent Alteration during Lithification

Very few generalizations can be made about the sedimentary environment of ironstones. Ironstones may be deposited in shallow-marine, interdeltaic, nonmarine lacustrine, and alluvial environments and may interfinger or replace sandy and shelly marine deposits laterally. They are frequently associated with

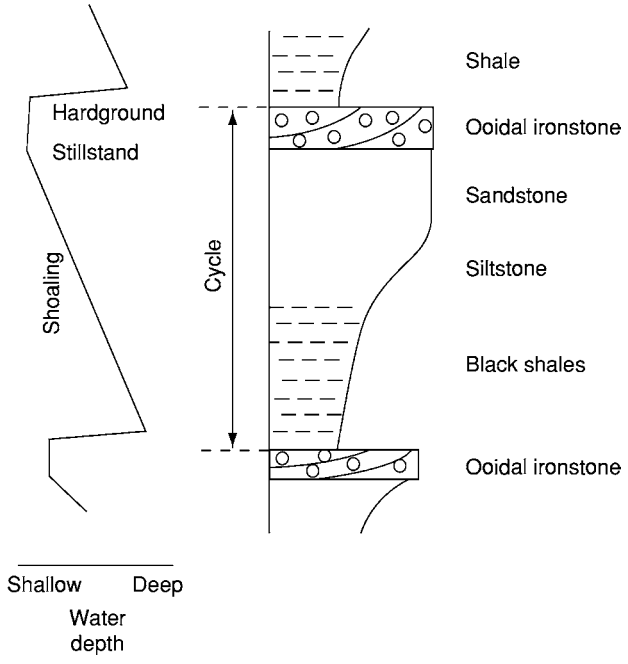


Figure 5 Idealized stratigraphical column for ooidal ironstones showing relative water depths of sedimentation, not to scale.

phosphates, coals, evaporites or laterites, and most have no direct relation to volcanism.

Blackband ironstones have many of the characteristics of bog iron ores, which are developed *in situ*, soon after deposition, by a reaction between organic material and underground colloidal iron-rich solutions under a thick vegetative cover. Progression of the process could yield siderite by reduction. Alternative evidence has been put forward suggesting that these deposits could form by direct sideritic precipitation from tropical swamp waters that are already rich in iron. Blackband ironstones are always developed in close proximity to coal seams, so either process could be feasible.

The diagenesis of the fine-grained claystone ironstones has been studied in great detail (Figure 6). Most became enriched in iron during very early diagenesis along or near the sediment–water interface. Based upon distinct chemical reactions involving the oxidation of organic matter buried within the sediment, diagenetic zones have been established. Although the zones can be considered as due to burial, their development is especially dependent upon the availability of oxidizing agents and organic matter, the sedimentary environment, the nature and amount of organic material, the composition of the inorganic sediment, the hydrological regime of the sedimentary pile, and the composition of the overlying water. The reactions below the oxic zone may be complicated by kinetic controls, which could explain the occasional appearance of residual ferric iron in an anoxic environment. Because some siderite concretions are developed early and are associated with many nonsequences, the sedimentation rate must have been relatively low (less than 40 m Ma ^{1}). Whilst claystone ironstones are formed during diagenesis by the growth of siderite in the pore spaces of argillaceous materials, sphaerosiderites form by the direct precipitation of siderite from pore fluids, and their size and shape probably reflect a higher growth rate. They can occur in a variety of environments, including the deep sea, but are usually products of a waterlogged zone below a leached soil profile.

The exact genesis of ooidal ironstones remains controversial. Particularly, the origin of the ooids is the subject of a long-lasting debate. The original constituents of ooids and how they vary from deposit to deposit are not known with any certainty. It is debatable whether the ooids grew from solutes, colloidal particles in solutions, or gels. The ferrous ion in bicarbonate form survives only in an anoxic or reducing environment, so this would place a severe constraint on its presence in solution. Ferruginous ooids are commonly built of alternating ferric oxide and berthierine sheaths of submicroscopic thickness. Whether the initial crystalline phase was berthierine

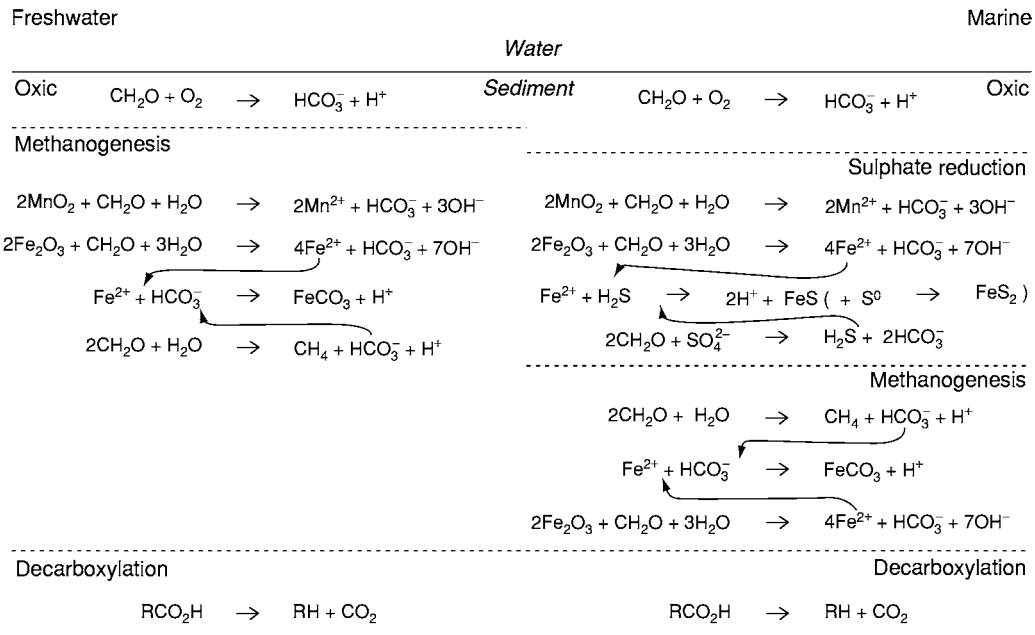


Figure 6 Summary of reactions and zonation that may occur during the diagenesis of sediments in marine and non marine conditions (after Curtis and Coleman 1986, Spears 1989 and reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446 489. London: Chapman & Hall, Figure 9.5 after Curtis and Coleman 1986 and Spears 1989).

or a precursor ferric mineral (e.g. odinite) is uncertain. Some feel that it was crystallized at the earliest stage, probably from a gel; others have suggested that it formed during the early stages of burial diagenesis. Alternatively, it could be a product of the transformation of either a mixture of kaolinite and hydrous ferric oxide or a complex synthesis of silicic, ferric, and aluminous substances. The processes involving micro-organisms (such as bacteria) are not understood, particularly in terms of how they promote the growth of ooids. Reworking of ooidal sediments in shallow-water environments often separates, concentrates, and highly sorts the ooids, forming lenses, which probably accumulated in shallow depressions. Often zonation of ironstones may be observed when the body is less affected by redistribution.

The variable nature of the nuclei of ooids and the trapping of marine microflora during growth indicate that ooids are probably generated within the host sediment. Ferruginous ooids could have grown on the seafloor, at the water–sediment interface, by either concentric growth due to precipitation of mineral matter, frequently around heterogeneous nuclei, or mechanical accretion by rolling (like a snowball). Alternatively, they could have grown inside the sediment at shallow depths below the water–sediment interface either as early diagenetic microconcretions or by replacement or addition of iron to peloids. Fluviatile examples do exist (e.g. the Late Oligocene deposits of Aral Lake, Russia), in which ooids have been developed on land and then moved, but this has not been convincingly demonstrated to be of general application.

The IGCP 277 came to the conclusion that ferruginous deposition must have been due to the interplay of a number of different processes and hence that there is rarely a single genetic explanation. The salinity of seawater, the carbon dioxide and oxygen contents of the atmosphere, the action of organisms, the sources and availability of iron compounds, seasonal or long-lasting climatic conditions, specific physicochemical conditions, the marine water depth, diagenetic processes, and tectonism are all potential factors. However, the dominant influences seem to be the local hydrodynamic conditions and the topographical relief of the land and seafloor, which may help to protect the ooidal deposits from excessive dilution by clastics. Paradoxically, the rates of deposition of the stratigraphical equivalents of many ooidal ironstones do not always correspond to the periods of lowest detrital input. Changes during burial are numerous and complex and include the formation of phosphatic minerals, iron oxides, siderite, pyrite, and quartz. In most cases, these are followed by alterations due to the effects of meteoric waters.

The Ferruginization Process

Although ironstones are generally considered to be the products of ferruginization during diagenesis, the physical sedimentary environment is thought to control the style of diagenesis in ironstones. Blackband ironstones are geochemically and isotopically homogeneous, suggesting stability of conditions during growth. They were probably formed close to the sediment surface, with precipitation of siderite, and not during progressive burial (**Figure 7**). The high manganese content of siderite, the relatively low calcium and magnesium contents, and the high carbonate content support this. Studies of carbon isotopes show that calcareous shells from limestones and ironstones have similar $\delta^{13}C_{PDB}$ values (from +4‰ to −6‰), indicating that the siderites replaced original calcite or aragonite and precluding the domination of methanogenesis. As has been previously noted, very early siderite could be precipitated directly from swamp waters, but could precipitation have occurred from primitive freshwater too?

Claystone ironstones usually have lower manganese and ^{13}C enrichment than blackband ironstones, which can be related to slightly later ferruginization, which takes place below the oxic zone in non-marine waters by diagenetic distribution of iron within the sediment (**Figure 8**). The relationships leading to the precipitation of iron minerals are complex and are susceptible to slight shifts in the concentrations and availability of reactants especially S and organic C. The thermodynamics of the reactions predict the observation of manganese enrichment within the concretion cores. The iron and manganese would have been present in the silicate minerals of the sediment. Sulphate reduction would be inhibited, as the sediments were isolated from potential sources of sulphate (e.g. seawater), and changes in organic matter would be methanogenic, giving rise to bicarbonates rich in ^{13}C ($\delta^{13}C_{PDB}$ values in the order of +10‰). Also precipitation would be enhanced by an increase in alkalinity resulting from the combination of changes in organic matter and the reduction of Fe^{3+} and Mn^{4+}. Any growth at deeper levels of burial would be slow under decarboxylation conditions with bicarbonate depleted in ^{13}C. In marine claystone ironstones, sulphate and iron reduction would proceed broadly simultaneously, leading to the production of iron pyrites. Siderite is normally rare in marine sediments because iron can become incorporated in carbonates only below the zone of sulphate reduction.

Over the past two decades there have been significant developments in research into the environmental conditions under which ooidal ironstones are formed.

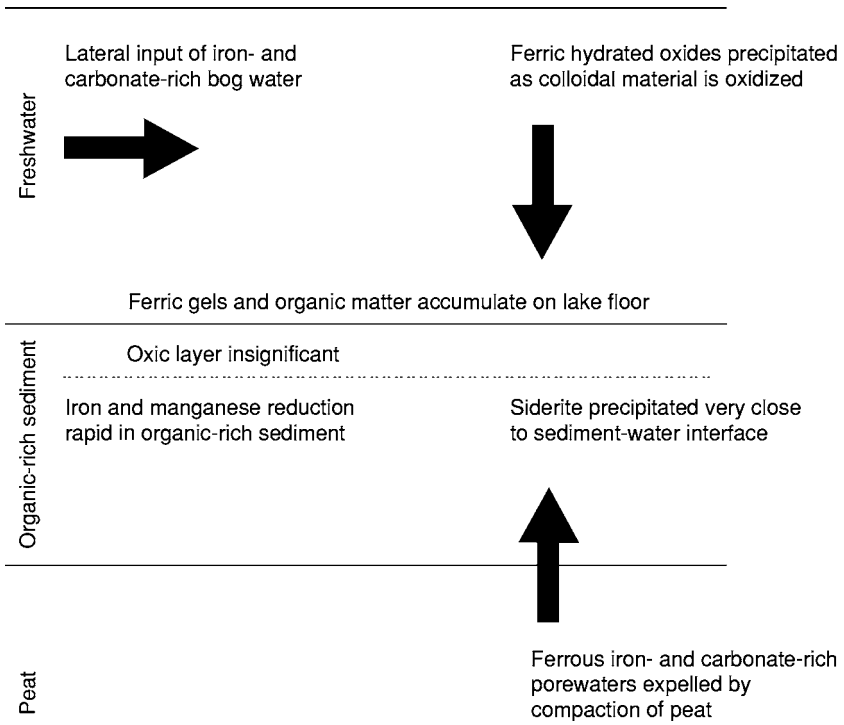


Figure 7 Model of mineralization for blackband ironstones (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446 489. London: Chapman & Hall, Figure 9.9).

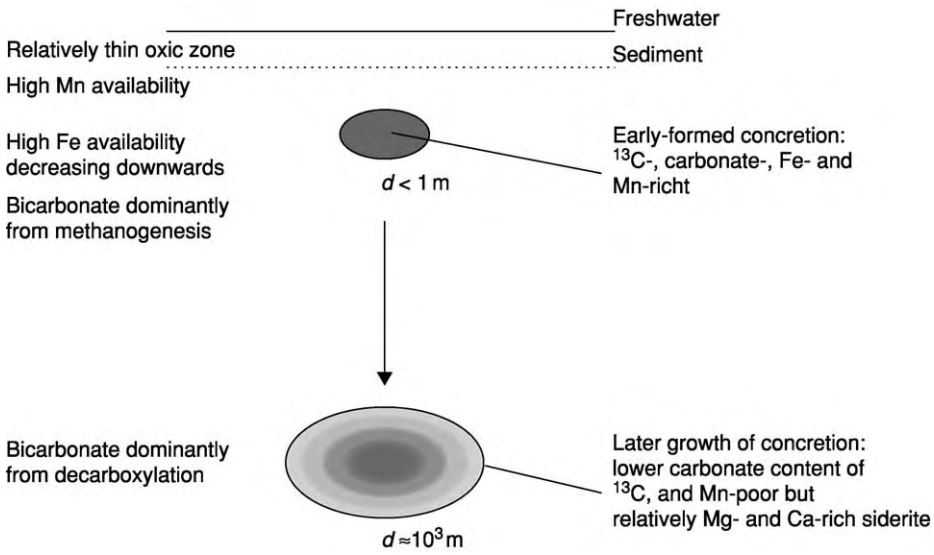


Figure 8 Model of mineralization for claystone ironstones (*d*, diameter) (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446 489. London: Chapman & Hall, Figure 9.7).

Sea-level change may be the most significant genetic control since it can generate very low accumulation rates within basins with low overall sedimentation rates (Figure 9). Widespread sediment starvation could be produced by relative sea-level rise in shallow epeiric seas with a topographically low hinterland. There is a dispute as to whether these conditions appertain to the end of regression or to the beginning

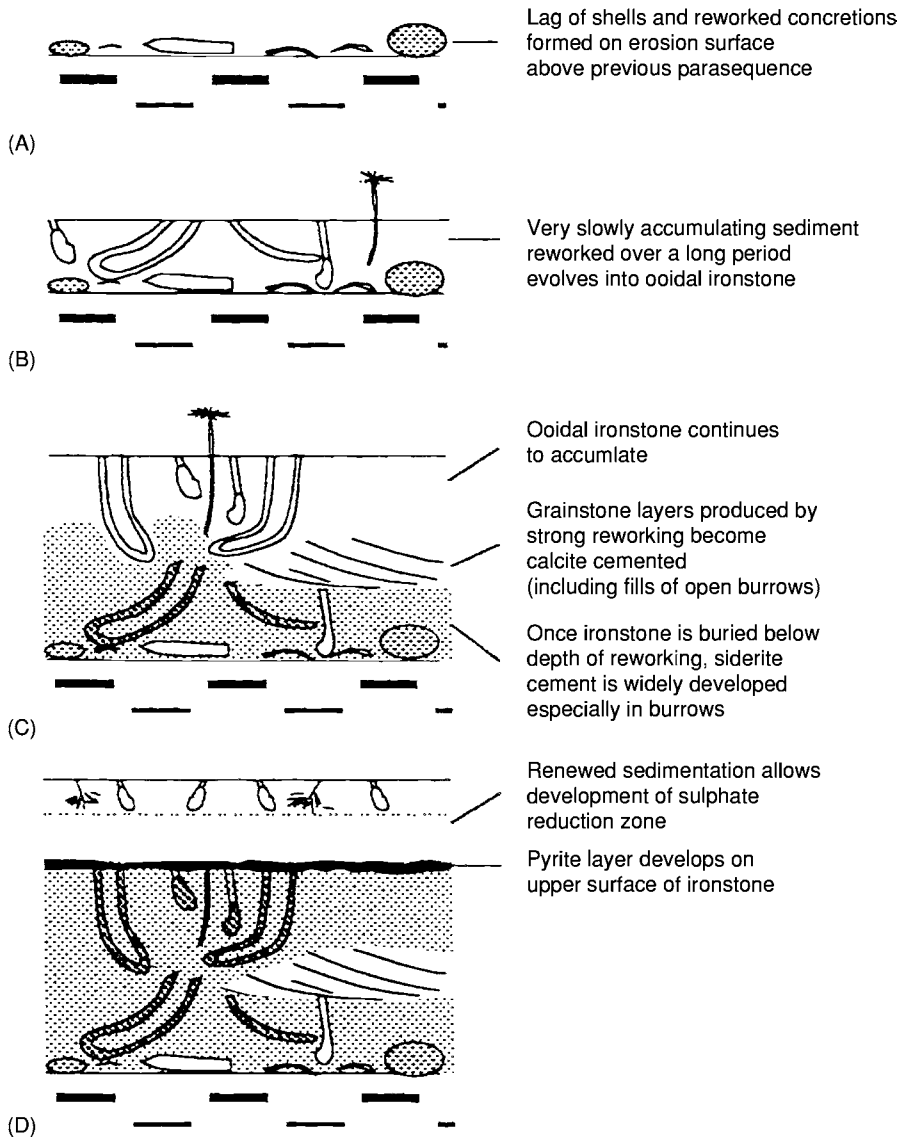


Figure 9 Model of mineralization for ooidal ironstones; (A), earliest phase; (B) and (C), middle phases; (D), latest phase (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446 489. London: Chapman & Hall, Figure 9.13).

of transgression. Similarly, the origin of early sideritic units can be related to a decrease in the role of sulphate reduction, a low sedimentation rate (less than 40 m Ma ^{1}), and oxygenated and carbon-poor sediments (values of $\delta^{13}C_{PDB}$ for various cements vary from -3 to $-22‰$).

The source and influx of iron is a subject of much controversy. In 1856, Sorby proposed that extensive ferruginization occurred during later diagenesis, but this idea is no longer accepted, since most ferruginous ooids were formed within the depositional environment. There are three proposals for the origin of iron enrichment:

1. iron-rich exhalative fluids, supplying the sediment–water interface (some examples do seem to be related to the episodic reactivation of faults involving exhalative hydrothermal or seep sources);
2. mechanical accretion of lateritic terrestrial weathering products (e.g. kaolinite and iron oxides) or lateritic soils to form ooids in a marine environment, with subsequent transformation to berthierine (this does not seem to be generally applicable since unaltered primary ooids with mixed iron oxide–kaolinite composition have not been found in marine ooidal ironstones); and

3. leaching from underlying sediments, especially organic-rich shales, during very early marine alteration of detrital material (the diagenetic redistribution of iron within sediments is difficult to demonstrate in ancient ooidal ironstones since the process would probably require considerable time).

The role of clay minerals in effecting ferruginization is unknown, particularly in respect of the transformation of non-iron-bearing phyllosilicates into iron-bearing ones and the role of iron-rich green trioctohedral clay minerals of warm seas as possible precursors of later ooidal minerals.

Stratigraphical Record (Temporal Occurrences) and Tectonic Settings

It has puzzled geologists that some geological periods have significant numbers of ironstones, whilst other periods are devoid of them. Ironstones are almost completely restricted to the Phanerozoic. Blackband and claystone ironstones are particularly prevalent in the Carboniferous, when the depositional basins occupied near-tropical locations. Ooidal ironstones are particularly common in the Ordovician, Devonian, Jurassic, and Cretaceous periods. Most ironstones were formed in warm climates, although some were deposited in cooler climates (e.g. the Late Ordovician and Late Permian ironstones). Palaeolatitude data has shown that the Ordovician and Devonian ironstones formed in a zone of the Gondwanan shelf seas ranging from 45°N to 65°S of the palaeoequator. In the Jurassic and Cretaceous, ironstones formed between 70°N and 10°S. For this reason climate cannot be the major contributory factor in their formation.

Ironstones are largely confined to three types of cratonic setting.

1. Many developed in anorogenic basins dominated by prolonged stability and sometimes with complex extensional faulting that involved the formation of marine basins and swells in areas of subdued relief.
2. Some developed along the margins of cratons during initial convergence or divergence of plates.
3. Other ironstones accumulated on the inner sides of mobile belts at times of diminished deformation.

Glossary

Condensed deposit A relatively thin but uninterrupted sedimentary sequence representing a significant period of time during which the deposits have accumulated very slowly. It is generally correlated with a thicker time-equivalent succession elsewhere.

Ferruginization A synonym of ferrification and the preferred term by IGCP 277 to describe the processes of iron-enrichment of various Earth materials.

Hardground A zone at the seafloor a few centimetres thick, where the sediment is lithified to form a hardened surface, which is often encrusted and bored.

IGCP The International Geological Correlation Programme.

Neoformation A synonym of neogenesis, the formation of new minerals.

Ooid A synonym of oolith and the preferred term to describe a spherical or ellipsoidal accretionary sand-sized (diameter of 0.25–2 mm) particle in a sedimentary rock (mainly limestones and ironstones). Ooids usually consist of successive concentric layers (often carbonates) around a central nucleus.

Pellet A small, usually ellipsoidal, aggregate of accretionary material (mainly micrite) that has, in most cases, formed from the faeces of molluscs and worms.

Peloid An allochem composed of micrite, irrespective of size or origin, without internal structure. Includes both pellets and intraclasts.

Pisoid A synonym of pisolith and the preferred term to describe small round or ellipsoidal particles (diameter of 2–10 mm) in a sedimentary rock (mainly limestones and ironstones). Pisoids are larger than ooids and usually consist of concentric layers around a central nucleus.

Stillstand A period of time when an area of land is stable relative to mean sea-level (or some other global measure), leading to a relatively unvarying base level of erosion.

Verdine facies Green marine clay characterized by the authigenesis (neoformation *in situ*) of iron-rich aluminous clay minerals, especially 0.7 nm repeat odinite, but not berthierine or glauconite.

See Also

Economic Geology. **Palaeozoic:** Carboniferous. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Mineralogy and Classification; Banded Iron Formations; Clays and Their Diagenesis; Limestones.

Further Reading

Boardman EL (1989) Coal measures (Namurian and West phalian) blackband iron formations: fossil bog iron ores. *Sedimentology* 36: 621 633.

Curtis CD and Coleman ML (1986) Controls on the precipitation of early diagenetic calcite, dolomite and siderite concretions in complex depositional sequences. In: Gautier DL (ed.) *Roles of Organic Matter in Sedi ment Diagenesis,* pp. 23 33. Special Publication 38. Denver: Society of Economic Palaeontologists and Min eralogists.

Curtis CD and Spears DA (1968) The formation of sedimentary iron minerals. *Economic Geology* 63: 257 270.

Kearsley AT (1989) Iron rich ooids, their mineralogy and microfabric: clues to their origin and evolution. In: Young TP and Taylor WEG (eds.) *Phanerozoic Ironstones,* pp. 141 164. Special Publication 46. London: Geological Society of London.

Kimberley MM (1994) Debate about ironstone: has solute supply been surficial weathering, hydrothermal con vection, or exhalation of deep fluids? *Terra Nova* 8: 116 132.

Odin GS (ed.) (1988) *Green Marine Clays, Oolitic Iron stone Facies, Verdine Facies, Glaucony Facies and Celadonite Bearing Facies A Comparative Study.* Devel opments in Sedimentology 45. Amsterdam: Elsevier.

Petranek J and Van Houten F (1997) *Phanerozoic Ooidal Ironstones.* Special Papers 7. Prague: Czech Geological Survey.

Spears DA (1989) Aspects of iron incorporation into sedi ments with special reference to the Yorkshire Ironstones. In: Young TP and Taylor WEG (eds.) *Phanerozoic Ironstones,* pp. 19 30. Special Publication 46. London: Geological Society of London.

Taylor JH (1949) *The Mesozoic Ironstones of Britain: Petrology of the Northampton Sand Ironstone.* Memoir of the Geological Survey of Great Britain. London: Geo logical Survey of Great Britain.

Van Houten FB and Arthur MA (1989) Temporal patterns among Phanerozoic oolitic ironstones and oceanic anoxia. In: Young TP and Taylor WEG (eds.) *Phanero zoic Ironstones,* pp. 33 49. Special Publication 46. London: Geological Society of London.

Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles,* pp. 446 489. London: Chapman & Hall.

Young TP and Taylor WEG (eds.) (1989) *Phanerozoic Ironstones.* Special Publication 46. London: Geological Society of London.

Limestones

R C Selley, Imperial College London, London, UK

Introduction

Limestones are one of the most important of all the sedimentary rocks introduced in (*see* **Sedimentary Rocks:** Mineralogy and Classification). Limestones are composed largely of calcium carbonate ($CaCO_3$) in the mineral form calcite, but there are several other important carbonate minerals with which limestones are associated. This article opens by discussing important differences between limestones and sandstones, and continues by outlining the mineralogy, classification, and rock names of limestones. This is followed by a brief account of limestone depositional environments, and, logically, by their postdepositional diagenesis. The article concludes with a description of the economic importance of limestones, which is considerable, and a selected reading list.

Differences between Limestones and Sandstones

Limestones and sandstones are the two most important groups of sedimentary rocks. However, limestones pose a completely different set of problems to those of sandstones, the solutions of which require the application of different concepts and techniques.

First, limestones, unlike sandstones, are intrabasinal in origin. That is to say they form in the environment in which they are deposited. The source material of sandstones, by contrast, has been weathered, eroded, transported, and may finally be deposited hundreds of kilometres from its point of origin. Sandstones (or siliciclastic rocks) therefore often contain many different minerals. Limestones, by contrast, have a much simpler mineralogy, generally consisting of only calcite and two or three others (which will be mentioned shortly). Siliciclastic sand grains may hold clues to their source, but tell little of their depositional environment. Limestone grains, by contrast, although largely monomineralic, occur in a wide range of sizes and shapes, reflecting their multiple origins. These grains form in specific environments from which they are seldom transported. Limestone grains thus give important clues about their environment of deposition.

When studying sandstones, vertical profiles of grain size and analysis of sedimentary structures are the keys to environmental diagnosis. With limestones, however, it is the analysis of grain type and texture that aids environmental diagnosis.

The second large difference between sandstones and limestones lies in their chemistry. Sandstones are

composed largely of quartz (SiO_2) sand, whilst limestones are composed largely of the mineral calcite ($CaCO_3$). In the subsurface environment, silica is chemically relatively inert, whereas calcite is much more reactive. This means that diagenesis in sandstones is relatively less important. Primary intergranular porosity may be preserved as it was when the sand was first deposited. In limestones, however, primary intergranular porosity is often quickly infilled by cement, even before the sediment has been buried. In the subsurface though, limestones are more vulnerable to the effects of acid solutions which can selectively leach out the rock and generate secondary pores. These may just be where individual fossil shells have leached out (Biomouldic porosity). These pores may enlarge to cross-cut the fabric of the rock (vuggy porosity), or even form caves, described by geologists as cavernous porosity (cavernous pores are defined as those that are large enough to contain a crouched geologist, or for the drill string to drop by 1 m or more). Fenestral porosity is a less common but significant type of pore system found in intertidal lagoonal muds. It forms from the buckling of laminae or the trapping of gas bubbles in carbonate mud when exposed to hot sunshine. It is characterized by thin, horizontally elongated pores, thus giving good horizontal and poor vertical permeability. The replacement of limestone by dolomite (dolomitization) may create intercrystalline porosity.

Through the creation of pore systems of diverse shapes and sizes, diagenesis may completely destroy the original depositional fabric of the sediment, a feature unknown in sandstones whose diagenetic overprint has little impact on the primary features.

These differences between limestones and sandstones will become clearer as this article unfolds.

Limestone Mineralogy, Grains, and Rock Names

Limestone Mineralogy

Limestones are principally composed of calcium carbonate in the form of calcite ($CaCO_3$). They may also contain several other carbonate minerals, listed in Table 1, and several non-carbonate impurities. There are two varieties of calcium carbonate ($CaCO_3$): 'aragonite', which has an orthorhombic crystal system, and 'calcite', which has a hexagonal crystal system. Aragonite is an important component of carbonate mud and of many shells. It is, however, relatively unstable in the subsurface, and soon goes into solution, often generating mouldic porosity, either at the surface or during shallow burial. It is very rarely preserved in old and/or deeply buried limestones. Calcite also occurs in many shells and other carbonate grains. It is more stable than aragonite. 'Dolomite' ($CaMg(CO_3)_2$) is the third and most important mineral associated with limestones (described in detail in (*see* **Sedimentary Rocks:** Dolomites)). It rarely forms on the Earth's surface, but commonly does in the subsurface. With increasing abundance of dolomite, limestones grade, via dolomitic limestones, into limey dolomites, and finally dolomite rock. Geopedants restrict the name dolomite to the mineral, and dolostone to the rock. Not everyone is so particular. 'Magnesite', 'ankerite', and 'siderite' are rare constituents of limestones.

Limestone Grains and Matrix

Just like sandstones, limestones consist of framework grains, matrix (syndepositional), cement (postdepositional), and, sometimes, pores. There are many types of carbonate grain. They are briefly described here, and are illustrated in Figure 1. Probably the most common grain type in limestones is shell debris. Indeed, many limestones are made up of nothing but fossils, whole or fragmented. These are termed bioclastic or biogenic limestones. Because of their origin, palaeoecology is an important tool in the diagnosis of the depositional environment. Not only whole fossils but even fragmented bioclasts may be identifiable, and hence of diagnostic value. Some limestones are composed of rounded grains termed ooids, or ooliths, and the rock oolite (named from '*oos*', the Greek for egg, the rock having the appearance of the roe of a fish). Internally, ooids show a concentric growth ring structure around a nucleus of a quartz grain or shell

Table 1 Summary of the minerals commonly associated with limestones

Mineral	*Formula*	*Crystal system*	*Occurrence*
Aragonite	$CaCO_3$	Orthorhombic	Some shells and mud, unstable during burial
Calcite	$CaCO_3$	Hexagonal	Some shells and mud, relatively stable during burial
Magnesite	$MgCO_3$	Hexagonal	Rare surface mineral
Dolomite	$CaMg(CO_3)_2$	Hexagonal	Rarely at the surface, more common as a subsurface replacement
Ankerite	$Ca(MgFe)(CO_3)_2$	Hexagonal	A rare cement
Siderite	$FeCO_3$	Hexagonal	As ooliths and cement

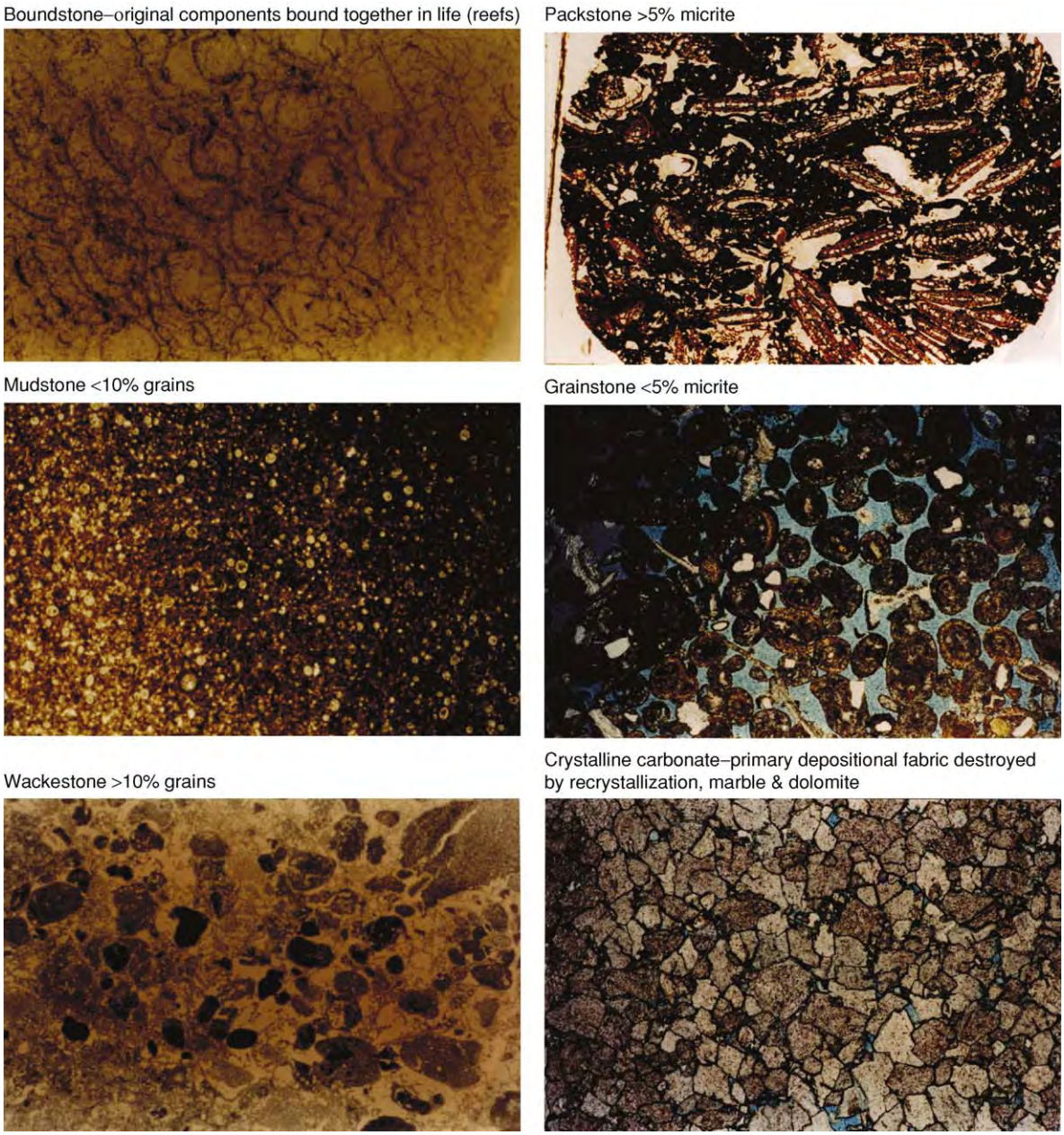


Figure 1 Illustrations of carbonate grain types, rock types, and names set within the Dunham classification of limestones. Boundstone: Colonial '*Halysites*' from Wenlock (Silurian) reef, Welsh Marches. Mudstone: Chalk (Upper Cretaceous), Beer, Devon. Wackestone: Rounded pellets and angular intraclasts in a micrite matrix, Marada Formation (Miocene), Jebel Zelten, Libya. Pack stone: Foraminifera in a modest micrite matrix (Oligocene), west of Marada Oasis, Libya. Grainstone: Ooids with quartz and shell fragment nuclei. Blue is preserved primary intergranular porosity. Portland Limestone (Upper Jurassic), Dorset. Crystalline carbon ate: Dolomite with minor intercrystalline porosity (blue), Zechstein (Upper Permian), UK North Sea. Once upon a time, this was probably a bryozoan reef. All illustrations from Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

fragment. Ooids form in shallow, high-energy marine environments with elevated temperatures and salinity, where carbonate precipitates episodically around an agitated nucleus. Larger sized concentric carbonate grains are known as pisoliths and oncolites; these are algally coated clasts. Some limestones are composed of structureless, bullet-shaped, sand-sized grains of lime mud and comminuted shell fragments. These grains are faecal pellets, the excreta of diverse burrowing aquatic creatures. Bizarre as it may seem, whole rock formations are composed of such material. Faecal pellets are the characteristic grain type of inner shelves, sheltered bays, and lagoons. Intraclasts are irregular, generally platy-shaped, carbonate grains

of various lengths and size. They are formed by the penecontemporaneous erosion of lithified carbonate sediment. Intraclasts are typically found in continental shelf and slope environments.

Between the framework grains briefly described above, there may be a finer grained syndepositional matrix. In sandstones, this is generally composed of clay minerals. In limestones, the matrix is more usually composed of lime mud, termed micrite. Micrite is sometimes aragonitic, sometimes calcitic. Micrite has several origins. It forms when calcareous algae decompose to liberate skeletal aragonite needles into the water. Waves and tidal currents, together with shell-munching predators, also play a part in disaggregating structured shells into comminuted lime mud. There is some evidence in modern warm shallow seas for the direct precipitation of aragonite mud in seawater. Limestone is commonly cemented by calcite, referred to as 'spar' or 'sparite' in this context. Several other carbonate and evaporite minerals precipitate out in limestone pore spaces as postdepositional cement.

Limestone Classification and Nomenclature

There are several different classifications of carbonate rocks. The one most widely used was proposed by Dunham in 1962 (Figure 1) and is briefly described below. 'Boundstone' is the term applied to limestone formed from organic skeletal material that grew bound together at the Earth's surface: in other words, reef rock. 'Mudstone' is composed of micrite with less than 10% grains; 'wackestone' is composed of micrite with over 10% grains. Both mudstone and wackestone are mud supported. That is to say the grains appear to 'float' within the micrite. In contrast, 'packstone' is grain supported, and the space between the grains is partly or completely filled with micrite matrix. 'Grainstone' is grain supported with negligible micrite matrix. The sequence mudstone–wackestone–packstone–grainstone reflects increasing depositional turbulence and energy, and is therefore useful in palaeoenvironmental reconstruction. Dunham's rock names can be qualified by grain type. For example, faecal wackestone, bioclastic packstone, ooidal grainstone, and so forth. The last rock name in Dunham's classification is 'crystalline carbonate', which would normally include dolomite and marble.

Limestone Depositional Environments

All carbonate sediment is precipitated by organic processes, either directly, as animals and plants secrete lime skeletons, or indirectly, as biochemical changes in water cause carbonate to precipitate as individual crystals. Except in a few deep marine environments, all ecosystems are based on plants, and all plants require sunlight to photosynthesize and grow. Plants provide the food for higher life forms to develop. Thus carbonate precipitation, caused or aided by plants, occurs in shallow water, and most of it takes place on the seafloor. Carbonate skeletal development decreases with increasing water depth, as darkness inhibits photosynthesis. Over time, therefore, a carbonate shelf will develop on a gently sloping seafloor (Figure 2). If sea-level remains constant, this shelf will gradually build out or prograde into deeper water. In certain situations, this gently sloping ramp may have an abrupt break in slope. This may occur in one of two ways. A fault may downthrow the seabed into deeper water. Rapid deepening may also occur if sea-level drops, erodes a sea cliff, and rises again, whereupon the rim will be oversteepened by rapid carbonate growth on the crest of the drowned sea cliff. These processes give rise to two types of carbonate setting: the gently sloping accretionary ramp, and the rimmed carbonate platform (Figure 3). This figure also shows the grain types and textures of the carbonate lithologies in these settings, and elegantly illustrates how carbonate rock type correlates with depositional environment. Considered in more detail, the following range of carbonate sediments may be found in sequence from the deep basin across the shelf towards the land. Basinal lime mud may form from the settling of organic detritus of plant and animal plankton that drifted near the surface. In this manner, many lime mudstones, including chalk, formed. These basinal muds may be interbedded with shallow-water carbonate sediment that was transported downslope as turbidity flows, submarine debris flows, and slides. Such transported carbonates, referred to sometimes as 're-deposited' or 'allodapic' limestones, are particularly common on the steep flanks of rimmed platforms and reefs. In warm, clear, shallow water, organic reefs may form by the *in situ* growth of corals,

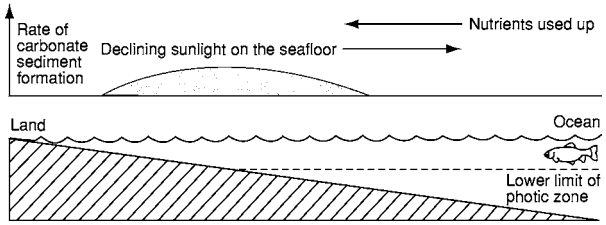


Figure 2 Diagram to show how carbonate sediment forms optimally in a zone between deep water, where the seafloor is too dark for photosynthesis to occur, and shallow water, where all the nutrients from the open sea have been used up. If sea level remains constant, the carbonate factory will gradually accrete seaward across the shelf into deeper water.

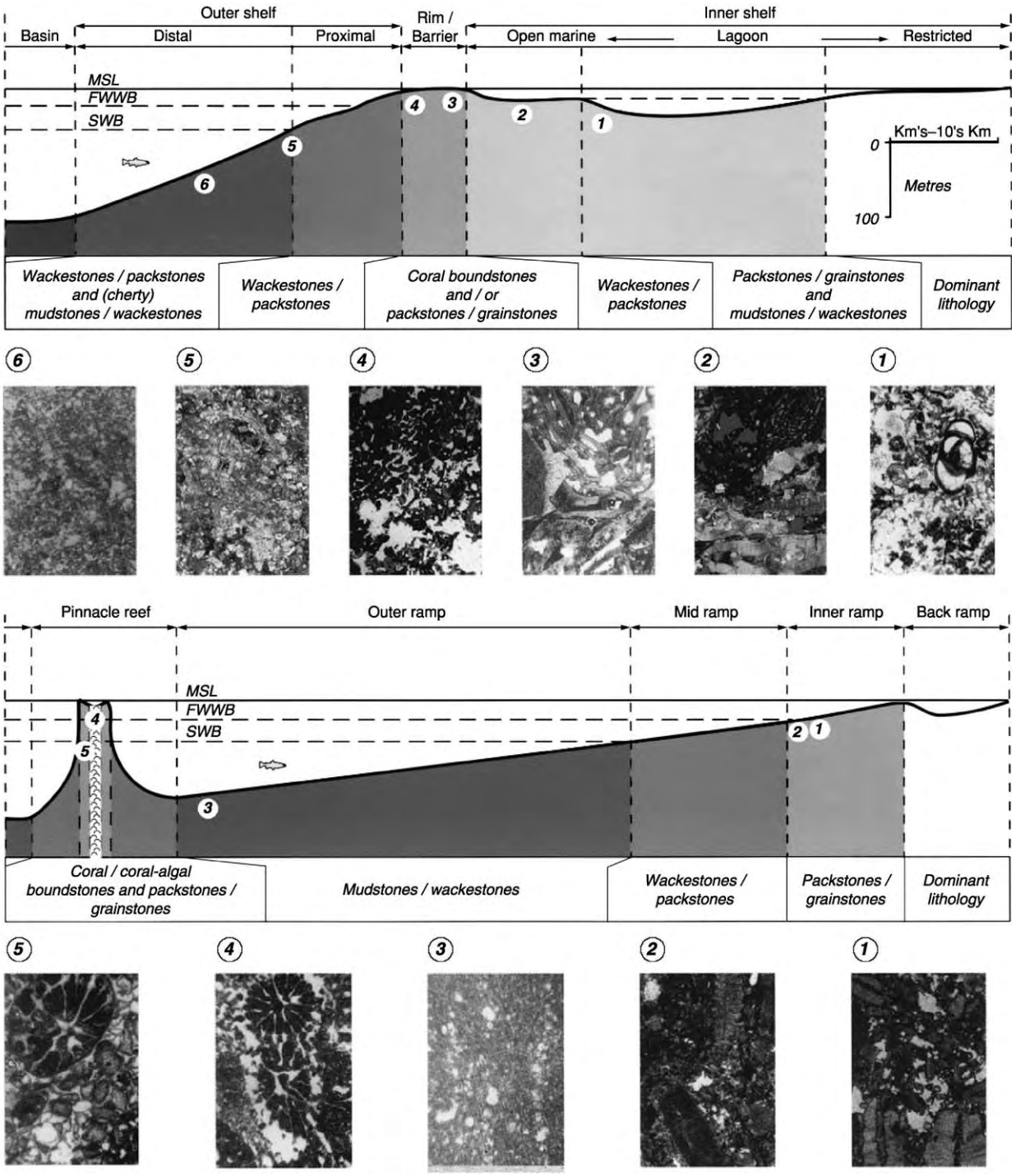


Figure 3 Cross sections to illustrate the correlation between depositional environments and carbonate rock types (grains and textures) for a rimmed carbonate platform (top) and a carbonate ramp (bottom). Reproduced with permission from Spring D and Hansen OP (1998) The influence of platform morphology and sea level on a carbonate sequence: the Harash Formation, Eastern Sirte Basin, Libya. In: McGregor DS, Moody RTJ, and Clark Lowes DD (eds.). *Special Publication of the Geological Society of London 132*, pp. 335 353. London: Geological Society of London.

bryozoa, algae, and many other sedentary biota. In turbulent conditions, shoals of oolitic and skeletal grainstone may form, as seen in the modern carbonate banks of the Bahamas, as described in more detail in (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves). In sheltered lagoons behind the high-energy environments of reefs and shoals, burrowing marine animals may excrete faecal pellets

to deposit thick formations of peloidal packstones and wackestones. In arid climates, these sediments may, in turn, pass into 'sabkha' ('*sabkha*' is Arabic for salt marsh) where dolomite and evaporite minerals may form. In humid climates, where terrigenous sediment runs off from the land, the carbonate lagoons may interfinger with siliciclastic sand and mud.

The depositional environments of carbonate shorelines and shelves in general, and reefs in particular, are described in greater detail in (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups')), respectively.

Limestone Diagenesis

As noted earlier, the minerals that form limestones are far less stable in the subsurface than are those that form sandstones. Recent carbonate sediment at the Earth's surface is composed of the two isomorphs of calcium carbonate: aragonite and calcite. Recent lime mud is largely aragonitic, but skeletal material is composed of both varieties, which vary in importance between different animal and plant groups.

The change of unconsolidated lime sediment into limestone happens very quickly, and with negligible burial. The 'fossilized' beer bottles and other anthropogenic detritus found in modern 'beach rock' prove this. These early cements are of both calcite and aragonite. In skeletal sands, one of the first diagenetic reactions is the dissolution of aragonite shells. This generates biomoldic porosity.

During burial, aragonitic muds undergo a reordering of the crystal lattice to form calcite. This change is concomitant with a volumetric increase of 8%, and a corresponding loss of porosity. This is why most ancient lime mudstones, certainly those of pre-Mesozoic age, are normally hard, tight, splintery rocks. By contrast, many Cretaceous and younger lime mudstones are light, porous, and chalky. Chalk consists mainly of the fossils of planktonic algae, termed the Coccolithophoridae, together with their disaggregated skeletal plates, termed coccoliths, coccolith-rich faecal pellets, calcispheres, and unicellular planktonic foraminifers. Coccoliths are not composed of unstable aragonite, but of the stabler calcite. Thus, during burial, these lime muds do not undergo expansive diagenesis like aragonitic muds. They maintain their chalky texture, being highly porous, but normally impermeable unless fractured. Chalks are described in greater detail in (*see* **Sedimentary Rocks:** Chalk).

Returning to the diagenesis of carbonate sands, during shallow burial, early cementation may destroy some porosity, but aragonite dissolution may enhance it. With continued burial, calcite cement may infill both biomolds and any remaining intergranular porosity. There are, however, several other diagenetic processes to which a cemented limestone may be subjected.

Limestones may undergo recrystallization, during which some or all of the primary fabric may be destroyed. Individual carbonate grains, generally bioclasts or ooids, may undergo pressure solution. This is a process whereby dissolution occurs at grain contacts due to overburden pressure. Concomitantly, the dissolved mineral matter may be precipitated as cement in adjacent pores. Additional evidence of dissolution is provided by stylolites. These are sutured surfaces, generally subparallel to bedding, where extensive dissolution has left an insoluble residue of clay, kerogen, and other matter along the suture. Stylolites occur in both pure limestones and quartzose sandstones.

Limestone diagenesis must not be thought of as a 'one-way street' that leads to the total loss of porosity and permeability. Limestones may be flushed through with acidic pore fluids, whose leaching properties may generate secondary porosity and permeability. The acidic fluids may come from adjacent compacting clay beds, conveniently generating secondary porosity ahead of petroleum invasion. More usually, however, secondary solution porosity is the result of uplift and erosion, and the flushing of limestone by acidic meteoric water (there is nothing new in acid rain). Solution may form moldic and vuggy pores. It may enlarge fractures and, in extreme cases, develop karstic caverns with concomitant collapse breccias (*see* **Sedimentary Processes:** Karst and Palaeokarst). Many of the best carbonate petroleum reservoirs occur where solution porosity has been developed and preserved beneath unconformities. The best way of preserving porosity in a limestone is for petroleum invasion to occur and expel cementing connate fluids. Renewed burial, without the benefit of petroleum invasion, may, of course, result in total recementation of the limestone as it makes its way to a completely cemented and recrystallized rock, termed marble.

The last important diagenetic process to which limestones are subjected is dolomitization, a process of such complexity and importance that it merits an article to itself (*see* **Sedimentary Rocks:** Dolomites).

Economic Importance of Limestones

Limestones are of great economic importance for many reasons. First, limestones contain lime, an essential ingredient for plant growth, and so limestone quarries are ubiquitous adjacent to farmland with lime-poor acid soil. Hard cemented limestones make excellent building stone and aggregate. Porous and permeable limestones, by contrast, serve as aquifers.

Limestone is used in the manufacture of cement and as a flux in the smelting of iron. Limestones are the host of several metallic minerals, including the eponymous Mississippi Valley telethermal Pb–Zn sulphide ores described in (*see* **Mineral Deposits and Their Genesis**). About 45% of the known petroleum reserves in the world occur in carbonate reservoirs (limestones and dolomites). Six main settings are recognized that preserve large volumes of porous and permeable limestone which have the potential to serve as petroleum reservoirs. These are: oolite grainstone shoals; reefs (often dolomitized); fore-reef talus; grainstone shoals sealed up-dip by evaporites; subunconformity traps, with extensive secondary porosity; and chalk, uplifted and fractured over salt diapirs. Small wonder, then, that limestones, their depositional environments, and diagenesis have been so intensively studied by geologists.

See Also

Building Stone. **Diagenesis, Overview**. **Mineral Deposits and Their Genesis**. **Minerals:** Carbonates. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Mineralogy and Classification; Chalk; Dolomites.

Further Reading

Dunham RJ (1962) Classification of carbonate rocks according to depositional texture. In: Ham WE (ed.) *Classification of Carbonate Rocks*, American Association of Petroleum Geologists. Tulsa. Ok: pp. 108 121.

Jordan CF and Wilson JL (1994) Carbonate reservoir rocks. *Memoir of the American Association of Petroleum Geologists* No. 60.

Leeder MR (1999) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Science.

Lucia FJ (1999) *Carbonate Reservoir Characterization*. Berlin: Springer Verlag.

Reading HG (ed.) (1996) *Sedimentary Environments, Processes, Facies and Stratigraphy*, 3rd edn. Oxford: Blackwell Science.

Selley RC (1996) *Ancient Sedimentary Environments and Their Subsurface Diagnosis*, 4th edn. London: Chapman & Hall.

Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

Spring D and Hansen OP (1998) The influence of platform morphology and sea level on a carbonate sequence: the Harash Formation, Eastern Sirt Basin, Libya. In: McGregor DS, Moody RTJ, and Clark Lowes DD (eds.) *Special Publication of the Geological Society of London 132*, pp. 335 353. London: Geological Society of London.

Tucker ME and Wright VP (1990) *Carbonate Sedimentology*. Oxford: Blackwell Scientific Publications.

Oceanic Manganese Deposits

D S Cronan, Imperial College London, London, UK

Introduction

Manganese nodules and encrustations (crusts) together with micronodules are ferromanganese oxide deposits which contain variable amounts of other elements (Table 1). They occur throughout the oceans, although the economically interesting varieties have a much more restricted distribution. Manganese nodules are spherical to oblate in shape and range in size from less than 1 cm in diameter up to 10 cm or more. Most accrete around a nucleus of some sort, usually a volcanic fragment but sometimes biological remains. Crusts are usually tabular.

The deposits were first described in detail in the Challenger Reports. This work was co-authored by J. Murray and A. Renard, who between them initiated the first great ferromanganese oxide controversy. Murray believed the deposits to have been formed by submarine volcanic processes whereas Renard believed that they had precipitated from continental run-off products in seawater. This controversy remained unresolved until it was realized that they could obtain their metals from either or both sources. The evidence for this included the finding of abundant nodules in the Baltic Sea where there are no volcanic influences, and the finding of rapidly grown ferromanganese oxide crusts associated with submarine hydrothermal activity of volcanic origin on the Mid-Atlantic Ridge. Subsequently, a third source of metals to the deposits was discovered, diagenetic remobilization from underlying sediments. Thus, marine ferromanganese oxides can be represented on a triangular diagram (Figure 1), the corners being occupied by hydrothermal (volcanically derived), hydrogenous (seawater derived), and diagenetic (sediment interstitial water derived) constituents.

There appears to be a continuous compositional transition between hydrogenous and diagenetic deposits, all of which are formed relatively slowly at normal deep seafloor temperatures. By contrast,

Table 1 Average abundances of elements in ferromanganese oxide deposits

	Pacific Ocean	*Atlantic Ocean*	*Indian Ocean*	*Southern Ocean*	*World Ocean average*	*Crustal abundance*	*Enrichment factor*	*Shallow marine*	*Lakes*
B	0.0277					0.0010	27.7		
Na	2.054	1.88			1.9409	2.36	0.822	0.81	0.22
Mg	1.710	1.89			1.8234	2.33	0.782	0.55	0.26
Al	3.060	3.27	2.49		2.82	8.23	0.342	1.80	1.16
Si	8.320	9.58	11.40		8.624	28.15	0.306	8.76	5.38
P	0.235	0.098			0.2244	0.105	2.13	0.91	0.15
K	0.753	0.567			0.6427	2.09	0.307	1.30	0.40
Ca	1.960	2.96	2.37		2.47	4.15	0.595	2.40	1.14
Sc	0.00097					0.0022	0.441		
Ti	0.674	0.421	0.662	0.640	0.647	0.570	1.14	0.212	0.338
V	0.053	0.053	0.044	0.060	0.0558	0.0135	4.13	0.012	0.001
Cr	0.0013	0.007	0.0029		0.0035	0.01	0.35	0.002	0.006
Mn	19.78	15.78	15.10	11.69	16.02	0.095	168.6	11.88	12.61
Fe	11.96	20.78	14.74	15.78	15.55	5.63	2.76	21.67	21.59
Co	0.335	0.318	0.230	0.240	0.284	0.0025	113.6	0.008	0.013
Ni	0.634	0.328	0.464	0.450	0.480	0.0075	64.0	0.014	0.022
Cu	0.392	0.116	0.294	0.210	0.259	0.0055	47.01	0.002	0.003
Zn	0.068	0.084	0.069	0.060	0.078	0.007	11.15	0.011	0.051
Ga	0.001					0.0015	0.666		
Sr	0.085	0.093	0.086	0.080	0.0825	0.0375	2.20		
Y	0.031					0.0033	9.39	0.002	0.002
Zr	0.052			0.070	0.0648	0.0165	3.92	0.004	0.045
Mo	0.044	0.049	0.029	0.040	0.0412	0.00015	274.66	0.004	0.003
Pd	0.602^{6}	0.574^{6}	0.391^{6}		0.553^{6}	0.665^{6}	0.832		
Ag	0.0006					0.000007	85.71		
Cd	0.0007	0.0011			0.00079	0.00002	39.50		
Sn	0.00027					0.00002	13.50		
Te	0.0050								
Ba	0.276	0.498	0.182	0.100	0.2012	0.0425	4.73	0.287	0.910
La	0.016					0.0030	5.33		0.027
Yb	0.0031					0.0003	10.33		
W	0.006					0.00015	40.00		
Ir	0.939^{6}	0.932^{6}			0.935^{6}	0.132^{7}	70.83		
Au	0.266^{6}	0.302^{6}	0.811^{7}		0.248^{6}	0.400^{6}	0.62		
Hg	0.82^{4}	0.16^{4}	0.15^{6}		0.50^{4}	0.80^{5}	6.25		
Tl	0.017	0.0077	0.010		0.0129	0.000045	286.66		
Pb	0.0846	0.127	0.093		0.090	0.00125	72.72	0.002	0.063
Bi	0.0006	0.0005	0.0014		0.0008	0.000017	47.05		

Note: Superscript numbers denote powers of ten, e.g., $^{6} \times 10^{6}$.
(Reproduced with permission from Cronan (1980).)

although theoretically possible, no continuous compositional gradation has been reported between hydrogenous and hydrothermal deposits, although mixtures of the two do occur. This may be partly because: (i) the growth rates of hydrogenous and hydrothermal deposits are very different with the latter accumulating much more rapidly than the former, leading to the incorporation of only limited amounts of the more slowly accumulating hydrogenous material in them; and (ii) the temperatures of formation of the deposits are different, leading to mineralogical differences between them which can affect their chemical composition. Similarly, a continuous compositional gradation between hydrothermal and diagenetic ferromanganese oxide deposits has not been found, although again this is theoretically possible. However, the depositional conditions with which the respective deposits are associated i.e., high temperature hydrothermal activity in mainly sediment-free elevated volcanic areas on the one hand, and low-temperature accumulation of organic rich sediments in basin areas on the other, would preclude much mixing between the two. Possibly they may occur in sedimented active submarine volcanic areas.

Internal Structure

The main feature of the internal structure of nodules and crusts is concentric or tabular banding which is

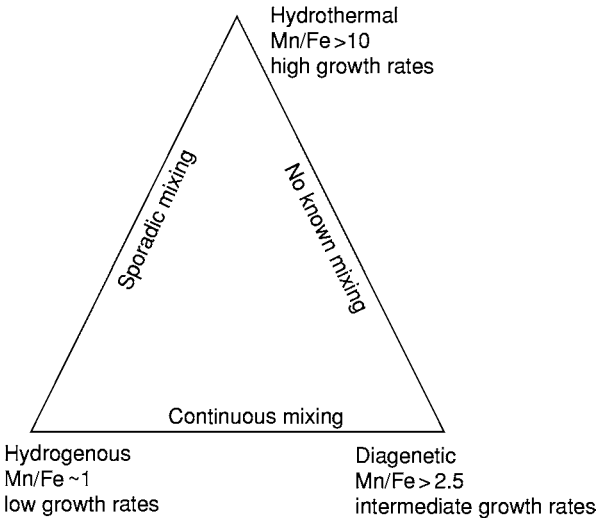


Figure 1 Triangular representation of marine ferromanganese oxide deposits.

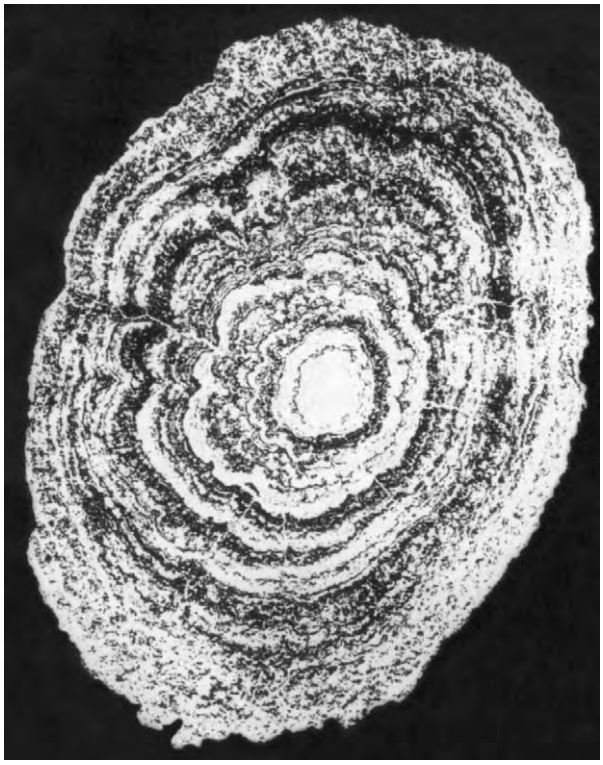


Figure 2 Concentric banding in a manganese nodule. (Reproduced by kind permission of CNEXO, France.)

developed to a greater or lesser extent in most of them (Figure 2). The bands represent thin layers of varying reflectivity in polished section, the more highly reflective layers being generally richer in manganese than the more poorly reflective ones. They are thought to possibly represent varying growth conditions.

On a microscopic scale, a great variety of structures and textures are apparent, some of them indicative of postdepositional alteration of nodule and crust interiors. One of the most commonly observed and most easily recognizable is that of collomorphic globular segregations of ferromanganese oxides on a scale of tenths of a millimetre or less, which often persist throughout much of the nodule or crust interior. Often the segregations become linked into polygons or cusps elongated radially in the direction of growth of the deposits. Several workers have also recognized organic structures within manganese nodules. Furthermore, cracks and fissures of various sorts are a common feature of nodule and crust interiors. Fracturing of nodules is a process which can lead to their breakup on the seafloor, in some cases as a result of the activity of benthic organisms, or of bottom currents. Fracturing is an important process in limiting the overall size of nodules growing under any particular set of conditions.

Growth Rates

It is possible to assess the rate of growth of nodules and crusts either by dating their nuclei, which gives a minimum rate of growth, or by measuring age differences between their different layers. Most radiometric dating techniques indicate a slow growth rate, from a few to a few tens of millimeters per million years. Existing radiometric and other techniques for dating include uranium series disequilibrium methods utilizing ^{230}Th ^{231}Pa, the ^{10}Be method, the K-Ar method, fission track dating of nodule nuclei, and hydration rind dating.

In spite of the overwhelming evidence for slow growth, data have been accumulating from a number of sources which indicate that the growth of nodules may be variable with periods of rapid accumulation being separated by periods of slower, or little or no growth. In general, the most important factor influencing growth rate is likely to be the rate at which elements are supplied to the deposits, diagenetic sources generally supplying elements at a faster rate than hydrogenous sources (Figure 1). Furthermore, the tops, bottoms, and sides of nodules do not necessarily accumulate elements at the same rate, leading to the formation of asymmetric nodules in certain circumstances (Figure 3). Differences in the surface morphology between the tops, bottoms, and sides of nodules *in situ* may also be partly related to growth rate differences. The tops receive slowly accumulating elements hydrogenously supplied from seawater and are smooth, whereas the bottoms receive more rapidly accumulating elements diagenetically supplied from the interstitial waters of the sediments

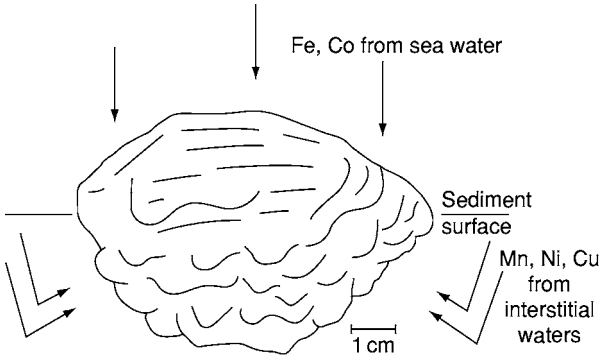


Figure 3 Morphological and compositional differences between the top and bottom of a Pacific nodule. (Reproduced with permission from Cronan, 1980.)

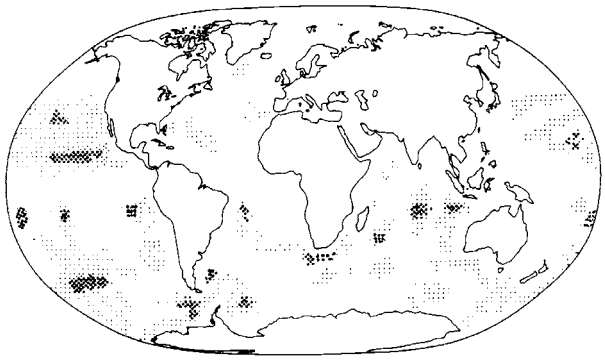


Figure 4 Distribution of manganese nodules in the oceans (updated from Cronan, 1980 after various sources.) ⋮⋮⋮, Areas of nodule coverage; ●●●, areas where nodules are locally abundant.

and are rough (Figure 3). The 'equatorial bulges' at the sediment-water interface on some nodules have a greater abundance of organisms on them than elsewhere on the nodule surface, suggesting that the bulges may be due to rapid growth promoted by the organisms.

It is evident, therefore, that growth cannot be regarded as being continuous or regular. Nodules and crusts may accrete material at different rates at different times and on different surfaces. They may also be completely buried for periods of time during which it is possible that they may grow from interstitial waters at rates different from those while on the surface, or possibly not grow at all for some periods. Some even undergo dissolution, as occurs in the Peru Basin where some nodules get buried in suboxic to reducing sediments.

Distribution of Manganese Nodules

The distribution and abundance of manganese nodules is very variable on an oceanwide basis, and can also be highly variable on a scale of a kilometre or less. Nevertheless, there are certain regional regularities in average nodule abundance that permit some broad areas of the oceans to be categorized as containing abundant nodules, and others containing few nodules (Figure 4), although it should always be borne in mind that within these regions local variations in nodule abundance do occur.

The distribution of nodules on the seafloor is a function of a variety of factors which include the presence of nucleating agents and/or the nature and age of the substrate, the proximity of sources of elements, sedimentation rates, and the influence of organisms. The presence of potential nuclei on the seafloor is of prime importance in determining nodule distribution. As most nodule nuclei are volcanic in origin, patterns of volcanic activity and the subsequent dispersal of volcanic materials have an important influence on where and in what amounts nodules occur. Other materials can also be important as nodule nuclei. Biogenic debris, such as sharks' teeth, can be locally abundant in areas of slow sedimentation and their distribution will in time influence the abundance of nodules in such areas.

As most nuclei are subject to replacement with time, old nodules have sometimes completely replaced their nuclei and have fractured, thus providing abundant nodule fragments to serve as fresh nuclei for ferromanganese oxide deposition. In this way, given sufficient time, areas which initially contained only limited nuclei may become covered with nodules.

One of the most important factors affecting nodule abundance on the seafloor is the rate of accumulation of their associated sediments, low sedimentation rates favouring high nodule abundances. Areas of the seafloor where sedimentation is rapid are generally only sparsely covered with nodules. For example, most continental margin areas have sedimentation rates that are too rapid for appreciable nodule development, as do turbidite-floored deep-sea abyssal plains. Low rates of sedimentation can result either from a minimal sediment supply to the seafloor or currents inhibiting its deposition. Large areas in the centres of ocean basins receive minimal sediment input. Under these conditions, substantial accumulation of nodules at the sediment surface is favoured.

Worldwide Nodule Distribution Patterns

Pacific Ocean As shown in Figure 4, nodules are abundant in the Pacific Ocean in a broad area, called the Clarion–Clipperton Zone, between about 6° N and 20° N, extending from approximately 120° W to 160° W. The limits of the area are largely determined by sedimentation rates. Nodules are also locally abundant further west in the Central Pacific Basin.

Sediments in the northern part of the areas of abundant nodules in the North Pacific are red clays with accumulation rates of around 1 mm per thousand years, whereas in the south they are siliceous oozes with accumulation rates of 3 mm per thousand years, or more.

Nodule distribution appears to be more irregular in the South Pacific than in the North Pacific, possibly as a result of the greater topographic and sedimentological diversity of the South Pacific. The nodules are most abundant in basin environments, such as those of the south-western Pacific Basin, Peru Basin, Tiki Basin, Penrhyn Basin, and the Circum-Antarctic area.

Indian Ocean In the Indian Ocean the most extensive areas of nodule coverage are to the south of the equator. Few nodules have been recorded in the Arabian Sea or the Bay of Bengal, most probably because of the high rates of terrigenous sediment input in these regions from the south Asian rivers. The equatorial zone is also largely devoid of nodules. High nodule concentrations have been recorded in parts of the Crozet Basin, in the Central Indian Ocean Basin, and in the Wharton Basin.

Atlantic Ocean Nodule abundance in the Atlantic Ocean appears to be more limited than in the Pacific or Indian Oceans, probably as a result of its relatively high sedimentation rates. Another feature which inhibits nodule abundance in the Atlantic is that much of the seafloor is above the calcium carbonate compensation depth (CCD). The areas of the Atlantic where nodules do occur in appreciable amounts are those where sedimentation is low. The deep water basins on either side of the Mid-Atlantic Ridge which are below the CCD and which accumulate only limited sediment, contain nodules in reasonable abundance, particularly in the western Atlantic. Similarly, there is a widespread occurrence of nodules and encrustations in the Drake Passage-Scotia Sea area, probably due to the strong bottom currents under the Circum-Antarctic Current inhibiting sediment deposition in this region. Abundant nodule deposits on the Blake Plateau can also be related to strong bottom currents.

Buried nodules Most workers on the subject agree that the preferential concentration of nodules at the sediment surface is due to the activity of benthic organisms which can slightly move the nodules. Buried nodules have, however, been found in all the oceans of the world. Their abundance is highly variable, but it is possible that it may not be entirely random. Buried nodules recovered in large diameter cores are sometimes concentrated in distinct layers. These layers may represent ancient erosion surfaces or surfaces of non-deposition on which manganese nodules were concentrated in the past. By contrast, in the Peru Basin, large asymmetrical nodules get buried when their bottoms get stuck in tenacious sediment just below the surface layer.

Compositional Variability of Manganese Nodules

Manganese nodules exhibit a continuous mixing from diagenetic end-members which contain the mineral 10Å manganite (todorokite) and are enriched in Mn, Ni, and Cu, to hydrogenous end-members which contain the mineral δMnO_2 (vernadite) and are enriched in Fe and Co. The diagenetic deposits derive their metals at least in part from the recycling through the sediment interstitial waters of elements originally contained in organic phases on their decay and dissolution in the sediments, whereas the hydrogenous deposits receive their metals from normal seawater or diagenetically unenriched interstitial waters. Potentially ore-grade manganese nodules of resource interest fall near the diagenetic end-member in composition. These are nodules that are variably enriched in Ni and Cu, up to a maximum of about 3.0% combined.

One of the most striking features shown by chemical data on nodules are enrichments of many elements over and above their normal crustal abundances (Table 1). Some elements such as Mn, Co, Mo, and Tl are concentrated about 100-fold or more: Ni, Ag, Ir, and Pb are concentrated from about 50- to 100-fold; B, Cu, Zn, Cd, Yb, W, and Bi from about 10- to 50-fold; and P, V, Fe, Sr, Y, Zr, Ba, La, and Hg up to about 10-fold, above crustal abundances.

Regional Compositional Variability

Pacific Ocean In the Pacific, potentially ore-grade nodules are generally confined to two zones running roughly east–west in the tropical regions, which are well separated in the eastern Pacific but which converge at about 170°–180° W (Figure 5). They follow the isolines of intermediate biological productivity, strongly suggestive of a biological control on their distribution. Within these zones, the nodules preferentially occupy basin areas near or below the CCD. Thus, they are found in the Peru Basin, Tiki Basin, Penrhyn Basin, Nova Canton Trough area, Central Pacific Basin, and Clarion–Clipperton Zone (Figure 5). Nodules in all these areas have features in common and are thought to have attained their distinctive composition by similar processes.

The potentially ore-grade manganese nodule field in the Peru Basin, centred at about 7°–8° S and 90° W

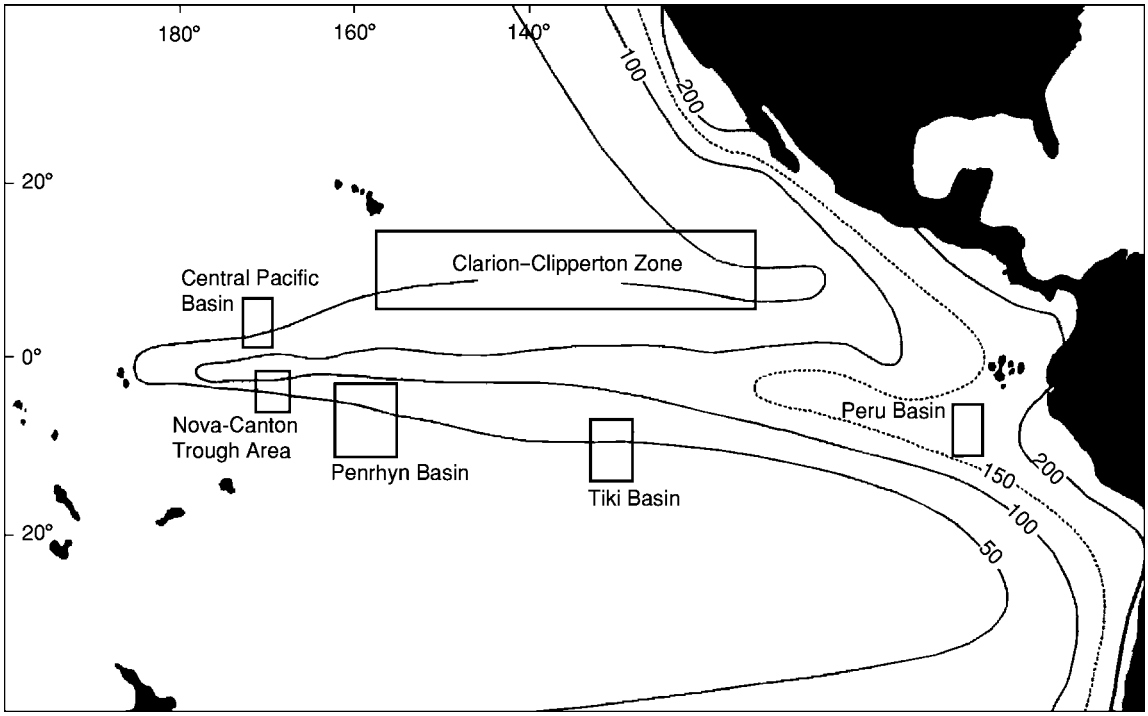


Figure 5 Approximate limits of areas of nickel and copper rich nodules in the subequatorial Pacific referred to in the text (productivity isolines in g Cm ^{2} y ^{1}).

(Figure 5), is situated under the southern flank of the equatorial zone of high biological productivity on a seafloor composed of pelagic brown mud with variable amounts of siliceous and calcareous remains. Nodules from near the CCD at around 4250 m are characterized by diagenetic growth and are enriched in Mn, Ni, and Cu, whereas those from shallower depth are characterized mainly by hydrogenous growth. The Mn/Fe ratio increases from south to north as productivity increases, whereas the Ni and Cu contents reach maximum values in the middle of the area where Mn/Fe ratios are about 5.

In the Tiki Basin there is also an increase in the Mn/Fe ratio of the nodules from south to north. All Ni + Cu values are above the lower limit expected in diagenetically supplied material.

The Penrhyn Basin nodules fall compositionally within the lower and middle parts of the Mn/Fe range for Pacific nodules as a whole. However, nodules from the northern part of the Basin have the highest Mn/Fe ratios and highest Mn, Ni, and Cu concentrations, reflecting diagenetic supply of metals to them, although Ni and Cu decrease slightly near the equator. Superimposed on this trend are variations in nodule composition with their distance above or below the CCD. In the Mn-, Ni-, and Cu-rich nodule area, maximum values of these metals in nodules occur within about 200 m above and below the CCD. The latititudinal variation in Mn, Ni, and Cu in Penrhyn Basin nodules may be due to there being a hydrogenous source of these metals throughout the Basin, superimposed on which is a diagenetic source of them between about 2° and 6° S at depths near the CCD, but less so in the very north of the Basin (0–2° S) where siliceous sedimentation prevails under highest productivity waters.

In the Nova Canton Trough area, manganese concentrations in the nodules are at a maximum between the equator and 2.5° S, where the Mn/Fe ratio is also highest. Manganese shows a tendency to decrease towards the south. Nickel and copper show similar trends to Mn, with maximum values of these elements being centred just south of the equator at depths of 5300–5500 m, just below the CCD.

In the central part of the Central Pacific Basin, between the Magellan Trough and the Nova Canton Trough, diagenetic nodules are found associated with siliceous ooze and clay sedimentation below the CCD. Their Ni and Cu contents increase south-eastwards, reaching a maximum at about 2.5°–3° N and then decrease again towards the equator where productivity is highest.

The Clarion–Clipperton Zone deposits rest largely on slowly accumulated siliceous ooze and pelagic clay below the CCD. The axis of highest average Mn/Fe ratio and Mn, Ni, and Cu concentrations runs roughly south-west–north-east, with values of these elements decreasing both to the north and south as productivity declines respectively to the north and increases towards the equatorial maximum in the south.

Indian Ocean In the Indian Ocean, Mn-, Ni-, and Cu-rich nodules are present in the Central Indian Ocean Basin between about 5° and 15° S. They are largely diagenetic in origin and rest on siliceous sediments below the CCD under high productivity waters. The deposits show north–south compositional variability with the highest grades occurring in the north.

Atlantic Ocean In the Atlantic Ocean, diagenetic Mn-, Ni-, and Cu-rich nodules occur most notably in the Angola Basin and to a lesser extent in the Cape/Agulhas Basin and the East Georgia Basin. These three areas have in common elevated biological productivity and elevated organic carbon contents in their sediments which, coupled with their depth near or below the CCD, would help to explain the composition of their nodules. However, Ni and Cu contents are lower in them than in areas of diagenetic nodules in the Pacific and Indian Oceans.

Distribution and Compositional Variability of Ferromanganese Oxide Crusts

Crusts generally accumulate on sediment free hard rock substrates, and thus their regional distribution is related to that of seamounts, plateaux, and other sediment free areas. In a major study on crusts by Hein *et al*. (2000) it is pointed out that the main substrata on which crusts form include basalt, phosphorite, and limestone. However, other than serving as nucleating surfaces for precipitation to occur, the substrata do not contribute to the formation or the composition of the crusts to any significant degree.

Ferromanganese oxide crusts (excluding hydrothermal ones) are generally less variable in composition than manganese nodules. In a large-scale study on crusts in the South Pacific, Verlaan *et al*. (in press) have shown that over the depth range from which the analysed crusts were sampled (650–5853 m), Co, Mn, and Ni increase as depth decreases, while Fe and Cu increase as depth increases. However, the relationship between crust composition and depth may be more complex than this, as analysis of crust composition versus depth in 500 m depth intervals, shows that in certain intervals the correlations between individual elements and depth differ from their overall correlations with depth. These differences are mainly found between three depth segments, above 1500 m (shallow), 1500–3000 m (middle), and below 3000 m (deep). Particularly noteable are the relationships (or lack of them) between elements in crusts and depth in the shallow segment in comparison with those in the deeper segments. There is an absence of any correlation with depth in the shallow segment for Co and Cu, and there is an opposite correlation with depth in the shallow segment for Mn compared with that in the middle and deep segments. Also noteable is the disappearance in the deep segment of any depth correlation for Ni and Fe, and the weakening of the correlation between Cu and depth.

Investigations on the regional variability in crust composition in the South Pacific by Verlaan *et al*. (in press), show that Co increases overall towards the equator. Manganese also increases from south to north and is generally low south of the 12th parallel. Nickel likewise increases northwards towards the equator, while Fe increases to the south-west, away from the equator. Copper shows little regional variation in crusts in the South Pacific. Regionally, Co, Mn, and Ni maintain an opposite behaviour to that of Fe throughout the South Pacific, over the full depth range of the samples collected. Furthermore, the overall equator-ward increase in Co, Mn, and Ni remains evident in each depth segment. The opposite trends in Co, Mn, and Ni enrichment, on the one hand, and Fe enrichment on the other, start from about the 10th parallel, which is the approximate latitude dividing the the high from the low biological productivity regions in the area studied, suggesting that the latitudinal compositional variations in crusts are at least partly productivity influenced. Longitudinally, Co, Mn, and Ni show a tendency to increase to the north-west and Fe towards the south-west, but these variations are much less pronounced than the latitudinal variations.

Economic Potential

Interest in manganese nodules commenced around the mid-1960s and developed during the 1970s, at the same time as the Third United Nations Law of the Sea Conference. However, the outcome of that Conference, in 1982, was widely regarded as unfavourable for the mining industry. This, coupled with a general downturn in metal prices, resulted in a lessening of mining company interest in nodules. About this time, however, several government-backed consortia became interested in them and this work expanded as evaluation of the deposits by mining companies declined. Part 11 of the 1982 Law of the Sea Convention, that part dealing with deep-sea mining, was substantially amended in an agreement on 28 July 1994, which ameliorated some of the provisions relating to deep-sea mining. The Convention entered into force in November 1994.

During the 1980s, interest in manganese nodules and crusts in exclusive economic zones (EEZs) started to increase. An important result of the Third Law of the Sea Conference, was the acceptance of a

200-nautical-mile EEZ in which the adjacent coastal state could claim any mineral deposits as their own. The nodules and crusts found in EEZs are similar to those found in adjacent parts of the International Seabed Area, and are of greatest economic potential in the EEZs of the South Pacific.

At the beginning of the twenty-first century, the outlook for deep sea mining remains rather unclear. It is likely to commence some time in this century, although it is not possible to give a precise estimate as to when. The year 2015 has been suggested as the earliest possible date for nodule mining outside of the EEZs. It is possible, however, that EEZ mining for nodules might commence earlier if conditions were favourable. It would depend upon many factors; economic, technological, and political.

Conclusions

Manganese nodules and crusts, although not being mined today, are a considerable resource for the future. They consist of ferromanganese oxides variably enriched in Ni, Cu, Co, and other metals. They generally accumulate on or around a nucleus and exhibit internal layering on both a macro- and microscale. Growth rates are generally slow. The most potentially economic varieties of the deposits occur in the subequatorial Pacific.

See Also

Mineral Deposits and Their Genesis. **Mining Geology:** Exploration. **Sedimentary Processes:** Deep Water Processes and Deposits. **Sedimentary Rocks:** Deep Ocean Pelagic Oozes.

Further Reading

Cronan DS (1980) *Underwater Minerals*. London: Academic Press.

Cronan DS (1992) *Marine Minerals in Exclusive Economic Zones*. London: Chapman and Hall.

Cronan DS (ed.) (2000) *Handbook of Marine Mineral Deposits*. Boca Raton: CRC Press.

Cronan DS (2000) Origin of manganese nodule 'ore provinces'. *Proceedings of the 31st International Geological Congress*, Rio de Janero, Brazil, August 2000.

Earney FC (1990) *Marine Mineral Resources*. London: Routledge.

Glasby GP (ed.) (1977) *Marine Manganese Deposits*. Amsterdam: Elsevier.

Halbach P, Friedrich G, and von Stackelberg U (eds.) (1988) *The Manganese Nodule Belt of the Pacific Ocean*. Stuttgart: Enke.

Hein JR, Koschinsky A, Ban M, Manhein FT, Kang J K, and Roberts L (2000) Cobalt rich ferromanganese crusts in the Pacific. In: Cronan DS (ed.) *Handbook of Marine Mineral Deposits,* pp. 239 279. Boca Raton: CRC Press.

Nicholson K, Hein J, Buhn B, and Dasgupta S (eds.) (1997) *Manganese Mineralisation: Geochemistry and Mineralogy of Terrestrial and Marine Deposits*. Geological Society Special Publication 119, London.

Roy S (1981) *Manganese Deposits*. London: Academic Press.

Teleki PG, Dobson MR, Moore JR, and von Stackelberg U (eds.) (1987) *Marine Minerals: Advances in Research and Resource Assessment*. Dordrecht: D. Riedel.

Verlaan P, Cronan DS, and Morgan C (in press) A comparative analysis of compositional variations in and between marine ferromanganese nodules and crusts and their environmental controls. *Progress in Oceanography*.

Phosphates

W D Birch, Museum Victoria, Melbourne, VIC, Australia

Introduction

Phosphorus is the tenth most abundant element on Earth and plays a key role in geological and biological processes. In the mineral kingdom, phosphates are amongst the most complex and diverse, with approximately 460 recognized species. Over the past five years about twenty new phosphate minerals have been recognized.

Phosphates are found in diverse geological environments and in many associations or assemblages. In igneous and metamorphic rocks, members of the apatite group, in particular fluorapatite, are the dominant phosphates. Because the solubility of phosphorus

is generally low in silicate minerals, fluorapatite and other ubiquitous phosphates such as monazite and xenotime usually occur as accessory minerals. An exception to this general rule is the wide diversity of late-crystallising phosphate minerals found in some granite pegmatites (*see* **Igneous Rocks:** Granite). Other important environments include sedimentary rocks, in which phosphates reach their maximum abundance in the form of phosphorite, and the oxidised zones of sulphide-bearing ore deposits. About ten phosphates, including four not found on Earth, have been recorded from meteorites.

Studies of phosphate minerals are important for scientific, environmental, agricultural, and health reasons. For mineralogists and crystallographers, the crystal chemistry of phosphate minerals embraces novel and diverse structures. To mineral collectors, the better crystallized and more colourful phosphates provide an unlimited source of intriguing and attractive specimens. The geochemistry of many phosphate minerals is important, as they commonly include trace amounts of uranium, thorium, and the rare earth elements. Apatite, monazite, and xenotime are significant for geochronology and thermochronology, utilising the trace amounts of radioactive elements uranium and thorium contained in their crystal structure. Phosphate enrichment of soils and discharge of phosphorus-bearing waste are human processes that have significant environmental impact requiring monitoring and control. As well, calcium phosphates are important constituents of human tissue.

There are thousands of references on phosphate minerals, which means that this review cannot be exhaustive. It deals mainly with the classification of phosphate minerals (*see* **Minerals:** Definition and Classification) and the major geological environments in which they occur.

Classification of Phosphate Minerals

The structures of phosphate minerals are almost exclusively built on the tetrahedral anionic unit $(PO_4)^3$, in which the P atom is central to the four O atoms. P–O bond lengths may vary, leading to distortions in the tetrahedra, but the average distance is 1.537 Å. From bond valence considerations, (PO_4) groups link easily with a range of non-tetrahedrally coordinated cations, such as Al^{3+}, Fe^{3+}, Mg^{2+}, Fe^{2+}, Mn^{2+}, Ca^{2+}, Sr^{2+}, Na^+, and K^+. Many phosphates are also hydrated and/or hydroxyl-bearing. Even though the anionic radius of P^{5+} (0.25 Å) is smaller than that of As^{5+} (0.42 Å) and V^{5+} (0.44 Å), a number of near-ideal solid solutions series are observed between phosphates and arsenates and to a more limited extent between phosphates and vanadates.

The crystal structures of most phosphate minerals have been well characterized, thereby facilitating classification. The simplest schemes divide the phosphates into classes based on whether they are anhydrous or hydrated, contain hydroxyl and/or halogen, or contain another anion such as $(SO_4)^2$, $(CO_3)^2$, $(CrO_4)^2$, $(AsO_4)^3$, and $(VO_4)^3$. The well-known Dana system adopts such an approach, as do James Ferraiolo and Hugo Strunz. Alexander Povarennykh combined this approach within divisions based on crystal chemical features. Perhaps because of the great diversity and complexity shown by phosphate structures, overall classification schemes based on crystal chemistry alone have been attempted by only a few researchers, notably Paul Moore, Frank Hawthorne, and Ivan Kostov. These schemes are generally based on the recognition that the PO_4 tetrahedra can polymerize in a number of ways, leading to a broad three-fold subdivision:

i. Polymerization of TO_4 tetrahedra, where T may be P, Be, Zn, B, Al, and Si
ii. Polymerization of PO_4 tetrahedra and MO_6 octahedra. This grouping covers a very large number of species and, within it phosphates can be further subdivided on the basis of whether polyhedra are unconnected, in finite clusters, or in infinite chains, sheets, or frameworks
iii. Polymerization of PO_4 tetrahedra and polyhedra that contain large cations coordinated by more than 6 oxygen atoms.

In many of these structures, OH and H_2O may provide one or more of the oxygen atoms in the PO_4 and MO_6 groups (for simplicity, this is not always indicated in the terminology used in the following review).

Structures with Polymerized TO_4 Tetrahedra

There are about 30 minerals in this category that have structures based mainly on polymerization of two or three PO_4 tetrahedra, or of PO_4 groups with BeO_4, ZnO_4, and AlO_4 tetrahedra. The structures can be based on finite clusters, such as in the rare isostructural zirconium-bearing species gainesite, $Na_2Zr_2[Be(PO_4)_4]$, mccrillisite, $NaCsZr_2[Be(PO_4)_4]$, and selwynite, $NaKZr_2[Be(PO_4)_4]$, or on infinite chains, sheets, and frameworks of tetrahedra. Examples of minerals with structures based on chains of PO_4–BeO_4 linkages included moraesite, $Be_2(PO_4)(OH)$ and roscherite, $CaMn_3[Be_2(PO_4)_3(OH)_3]$, while spencerite, $Zn_4(PO_4)_2(OH)_2$ involves chains of alternating $ZnO_2(OH)(H_2O)$ and PO_4 tetrahedra. Both PO_4–ZnO_4 and PO_4–BeO_4 linkages are present in the sheet-like structures, which include hopeite, $Zn_3(PO_4)_2 \cdot 4H_2O$, scholzite, $CaZn_2(PO_4)_2 \cdot 2H_2O$,

Figure 1 Hydroxylherderite crystals (largest crystal 18 mm long) on muscovite from the Xanda mine, Minãs Gerais, Brazil. Museum Victoria specimen M43389, photography by J Broomfield. Reproduced with permission from Museum Victoria.

Figure 2 Wavellite sprays (up to 30 mm across) from Montgomery County, Arkansas, USA. Museum Victoria specimen M27840, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

the isostructural herderite, $CaBe(PO_4)F$, (Figure 1) and hydroxylherderite, $CaBe(PO_4)OH$. In the infinite framework structures, all except berlinite, $AlPO_4$, which is isostructural with quartz, are based on PO_4–BeO_4 linkages, including beryllonite, $NaBePO_4$ and pahasapaite, $Li_8Ca_8Be_{24}(PO_4)_{24} \cdot 38H_2O$, which has a complex, zeolite-like framework structure.

Structures with Linked TO_4 Tetrahedra and MO_6 Groups

About 180 minerals are known to be represented in this structural grouping, with most being infinite frameworks. Only five minerals are known that have structures based on isolated or finite clusters of tetrahedra and octahedra, linked together by hydrogen bonding. They include struvite, $NH_4Mg(PO_4) \cdot 6H_2O$, anapaite, $Ca_2Fe^{2+}(PO_4)_2 \cdot 4H_2O$, and morinite, $NaCa_2Al_2(PO_4)_2(F,OH)_5 \cdot 2H_2O$. There are twenty-two minerals with structures based on infinite chains of tetrahedra and octahedra. These can be subdivided further into five topologically distinct types, depending on how the TO_4 and MO_6 groups are linked. A group of minerals that includes collinsite, $Ca_2(Mg,Fe)(PO_4)_2 \cdot 2H_2O$, and fairfieldite, $Ca_2(Mn,Fe)(PO_4)_2 \cdot 2H_2O$, consists of chains of alternating $(M^{2+}O_4\{H_2O\}_2)$ octahedra and pairs of (PO_4) tetrahedra, with the chains linked by 7-coordinated Ca atoms and by hydrogen bonding. In childrenite, $(Fe,Mn)Al(PO_4)(OH) \cdot H_2O$, and members of the jahnsite group, $CaMn(Fe,Mn,Mg)_2Fe_2(PO_4)_4(OH)_2 \cdot 8H_2O$, chains are based on corner-sharing octahedra with bridging (PO_4) groups, while in bearthite, $Ca_2Al(PO_4)_2(OH)$, adjacent octahedra share an edge to build chains linked by $(PO_4)_4$ groups and Ca cations.

Nearly 50 minerals are known to have structures consisting of infinite sheets of (PO_4) tetrahedra and (MO_6) octahedra. These structures can also be grouped, depending on how the octahedra and tetrahedra are linked. However, due to their complexity, it is not feasible to describe or summarise them here. Some of the more important phosphates with sheet-like structures include members of the crandallite group, based on $[Al_3(PO_4)(PO_3\{OH\})(OH)_6]$, and vivianite, $Fe_3^{2+}(PO_4)_2 \cdot 8H_2O$. About 110 phosphate minerals are known to have framework structures, by far the largest group. Important minerals in this category include wavellite, $Al_3(PO_4)_2(OH,F)_3 \cdot 5H_2O$ (Figure 2), variscite, $AlPO_4 \cdot 2H_2O$, pseudomalachite, libethenite, $Cu_2(PO_4)(OH)$ $Cu_5(PO_4)_2(OH)_4$, members of the turquoise $[CuAl_6(PO_4)_4(OH)_8 \cdot 4H_2O]$ group, the iron phosphates dufrenite, $Fe^{2+}Fe_4^{3+}(PO_4)_3(OH)_5 \cdot 2H_2O$, and rockbridgeite, $(Fe^{2+},Mn)Fe_4^{3+}(PO_4)_3(OH)_5$.

Structures with TO_4 Groups and Large Cations

In the phosphate minerals within this broad grouping, the main cations, which may be either monovalent, divalent, or trivalent, are coordinated to varying numbers of oxygen atoms to form polyhedra, which

are then linked in various ways, commonly through (PO_4) tetrahedra, to form chains. Chains are then linked to form sheets, which stack in various crystallographic directions. It is not possible here to describe or summarize individual structures, but some important minerals represented include xenotime, (Y,Yb)(PO_4), and members of the monazite (REE,Ce,Ca,Th)(PO_4) group, as well as a suite of hydroxyl-bearing and hydrated ammonium, sodium, and potassium-bearing species typically found in cave environments. However, the apatite group is the most significant in this structural category and is briefly outlined below. There is also an important and widespread suite of phosphate minerals whose structures are dominated by the uranyl $(U^{6+}O_2)^{2+}$ group (Figures 3 and 4).

Figure 3 Crystals of meta autunite, $Ca(UO_2)(PO_4)_2\ 6H_2O$ (up to 7 mm wide) from Autun, Burgundy, France. Museum Victoria specimen M27680, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Figure 4 Crystals of saleeite, $Mg(UO_2)(PO_4)_2\ 8H_2O$ (up to 4 mm across) from the Ranger mine, Northern Territory, Australia. Museum Victoria specimen M45060, photograph by F Coffa. Reproduced with permission from Museum Victoria.

These include about thirty species in two main groups related to torbernite, $Cu(UO_2)(PO_4)_2 \cdot 10H_2O$, and phosphuranylite, $KCa(H_3O)_3(UO_2)_7(PO_4)_4O_4(UO_2)(PO_4)_2 \cdot 8H_2O$, respectively.

The Apatite structural group The apatite group contains ten species, including pyromorphite, $Pb_5(PO_4)_3Cl$, and belovite, $Sr_3Na(Ce,La)(PO_4)_3(OH)$. However, fluorapatite, $Ca_5(PO_4)_3F$, chlorapatite, $Ca_5(PO_4)_3Cl$, and hydroxylapatite, $Ca_5(PO_4)_3OH$, are the most widespread and influential in geological and biological processes. The essential atomic arrangement for these three species consists of (PO_4) tetrahedra and two Ca polyhedra (Figure 5). Ca1 is coordinated to nine oxygen atoms, and Ca_2 bonds to six oxygen atoms and one anion (F, Cl, or OH) situated in channels running parallel to the *c*-axis. The Ca1 polyhedron shows little response to the effect of different channel anions, whereas in the Ca_2 polyhedron there are significant shifts in the positions of the channel anions, arising from their markedly different sizes. The structure permits a very wide range of substitutions in all cation and anion sites in natural and synthetic apatites. For example, the monovalent ions in the *c*-axis channel sites can be replaced by divalent anions such as $(CO_3)^2$ (eg., in carbonite–fluorapatite) and O^2 . Vacancies may also occur in the *c*-axis channels. A large number of divalent cations (for example Pb^{2+}, Ba^{2+}, Mn^{2+}, and Sr^{2+}) can substitute for Ca. The (PO_4) group is commonly replaced by other tetrahedral anion groups, such as $(AsO_4)^3$, $(SO_4)^2$, $(SiO_4)^4$, and $(VO_4)^3$. Apatites may also take up rare earth elements but the mechanisms are complex and beyond the scope of this discussion.

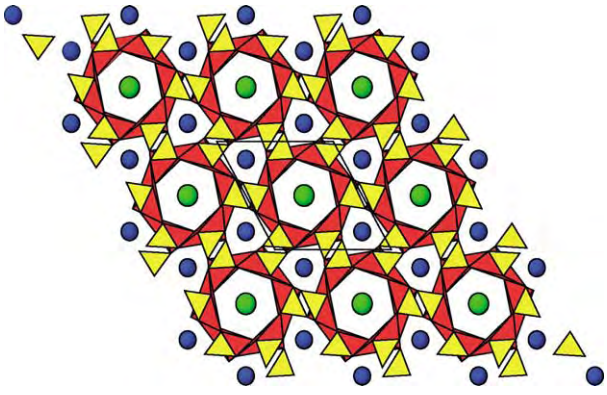


Figure 5 Apatite crystal structural diagram, viewed down the c axes and with the unit cell outlined in black. The PO_4 tetrahedra are yellow, the smaller Ca site is drawn as red polyhedra, the second Ca is shown as blue balls and the (F, Cl, OH) ion is green. Diagram prepared by A Pring.

Geological Environments for Phosphate Mineral Occurrences

Phosphate minerals are found in a wide range of igneous, metamorphic, and sedimentary rocks. In general however, only three minerals—apatite *per se*, monazite, and xenotime—are ubiquitous in typical igneous and metamorphic rocks. With a few exceptions, however, such as granitic pegmatites, alkaline intrusive rocks and some calc-silicate skarns and marbles (Figure 6), these minerals occur as primary accessory phases only. Their abundance is a general indication of the phosphorus content of the rock, as phosphorus has low solubility in most silicate minerals. Extreme phosphorus enrichment in magmas in represented by carbonatites, in which apatite is the most widespread economic mineral, notably in two deposits, Phalaborwa in South Africa and Khibiny on the Kola Peninsula in Russia.

In sedimentary rocks, phosphate minerals are represented throughout the geological time-scale to the present day, and occur in a wide range of host rocks. Many are hosted by metasediments, in which mobilization and recrystallization of original primary phosphate have yielded diverse assemblages of species.

Figure 6 Fluorapatite crystal (2 cm long) in marble from Wilberforce, Ontario, Canada. Museum Victoria specimen M39659, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Granite Pegmatites

Apatite varieties dominate the phosphate suites found in many granite pegmatites and may crystallize at different stages. For example, primary apatite may occur intergrown with feldspar, quartz, and tourmaline, or it may crystallize later as a druse or miarolitic-cavity mineral during the hydrothermal stage (Figure 7). At lower temperatures, crusts of carbonate-bearing apatite may form.

Paul Moore has constructed a 'paragenetic tree' of pegmatite phosphates. Primary phosphates are found as giant crystals or lenticular masses which crystallized near the core of the pegmatite, usually embedded in massive quartz. As well as apatite, the triphylite–lithiophilite, triplite–zwieselite, and amblygonite–montebrasite series are significant at the primary stage, with the last series in places constituting an ore of lithium. Metasomatic alteration during the final stages of pegmatite formation may result in these primary phases being partially replaced by such species as alluadite, triploidite–wolfeite, and purpurite–heterosite, which may be nodular and fine-grained. Typical reactions during alteration involve Na, K, and

Figure 7 Fluorapatite crystal (12 mm across) showing colour zonation, in granite from Lake Boga, Victoria, Australia. Museum Victoria specimen M29944, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Ca displacing Li, and the addition of OH. The largest diversity of phosphate species in granite pegmatites arises from the oxidation of these primary phases. This may take place under 'hydrothermal' conditions, at temperatures less than 250°C during the cooling history, or much later during atmospheric weathering. The great diversity of these secondary species reflects the mixed valence states of Fe and Mn; the different configurations in which oxygen atoms from H_2O molecules and OH and PO_4^2 groups may cluster around these metal cations in octahedral coordination; and the different ways the octahedra can combine (polymerise) to form clusters in solution. Such clusters may be represented in the crystal structures of secondary phosphates. Amongst these are barbosalite, rockbridgeite, beraunite, phosphosiderite, strengite, leucophosphite, bermanite, strunzite, laueite, cacoxenite, cyrilovite, eosphorite–childrenite, and ludlamite (Figure 8), which are derived from Fe- and Mn-bearing primary phosphates. Species such as hurlbutite, herderite, brazilianite, morinite, crandallite, and whitlockite are derived from primary phosphates with low Fe and Mn and/or high Li contents. Many of these secondary species crystallise as small crystals in open cavities, resulting from removal of much of the PO_4^2 and most of the alkalis.

Figure 8 Ludlamite crystal group (12 mm across) from the San Antonio mine, Chihuahua, Mexico. Museum Victoria specimen M37681, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

The best known and most prolific localities for phosphate minerals include the famous pegmatites of the Black Hills, South Dakota; the Palermo mine, New Hampshire; Hagendorf Sud, Bavaria; Tsaobismund, Namibia; Buranga, Rwanda; Viitaniemi, Finland; Sapucaia, Brazil, and occurrences in southern California and Maine in the USA.

Sediment-Hosted Phosphate Deposits

Sedimentary phosphates—phosphorites—are the most important of the world's sources of phosphate rock (*see* **Sedimentary Rocks:** Mineralogy and Classification). They occur on every continent and range in age from Precambrian to Holocene, with nearly all having a marine origin. Modern phosphorites are mineralogically monotonous, consisting of grains of cryptocrystalline or amorphous carbonate-fluorapatite (variously referred to as collophane or francolite), occurring as beds ranging in thickness from a few centimetres up to tens of metres. Other forms of phosphorite include nodules and concretions. Phosphorites are commonly observed in shallow seas, along the edges of continental shelves, and on ocean plateaus. The phosphorus is believed to be derived from faecal matter, bone material, and decaying marine organisms that accumulate locally or are carried into shallow coastal regions by upwelling deep ocean currents. These nutrients encourage a diverse biota to flourish, ultimately producing organic-rich sediments. During early diagenesis, collophane precipitation occurs within the upper layers of these sediments from pore waters rich in phosphorus leached from the organic remains; precipitation is enhanced where phosphatic nuclei are already present. A changing depositional environment with periods of reduced deposition and reworking of sediments in shallow seas favours phosphogenesis. This model is generally applicable to old phosphorites that remain recognizable, such as those of the Cambrian–Ordovician Georgina Basin, in Queensland, Australia. However, settings and methods of deposition (including transport of phosphate grains) vary widely and are subject to debate. Study of the age and global distribution of phosphorites has led to the identification of major phosphogenic episodes as far back as the Proterozoic.

Older phosphorites are more likely to have undergone diagenesis, deformation, and metamorphism, to the extent that the original nature of the deposits may become obscured. Phosphorus may be mobilized in solution and distributed into surrounding rocks, where diverse suites of well-crystallized secondary phosphates may form. Perhaps the most notable such

occurrence is in the Cretaceous Rapid Creek Formation in the Canadian Yukon. Here, a marine sequence of ironstone and shale containing unusual Fe- and Mn-bearing phosphates instead of collophane has been deformed and uplifted. This resulted in sets of fractures in which a wealth of well-crystallized phosphate minerals have formed. Four major assemblages characterized by the predominance of specific elements and related to a specific host rock have been identified, with at least five new phosphate species recognized (baricite, garyansellite, gormanite, kulanite, and penikisite). As well, remarkable crystals of arrojadite, augelite, and lazulite, amongst others, occur in this fracture-filling paragenesis.

In south-eastern Australia, a variety of settings for sedimentary phosphate deposits has been recognized, with several producing a range of unusual, in some cases new species. Some deposits have been exploited for phosphate, but all are low grade. The oldest deposits are in South Australia, where there are two main phosphatic horizons, one Late Precambrian, the other Early Cambrian, associated with limestones. There has been local leaching and intermittent concentration of phosphate by replacement. The Moculta deposit has been affected by regional metamorphism, which has recrystallised and brecciated the phosphatic rock. A range of secondary phosphate minerals, such as wavellite, beraunite, cyrilovite, leucophosphite, variscite, crandallite, and aldermanite (for which Moculta is the type locality) has been recorded in veinlets and small cavities and probably formed during near-surface weathering. An intense and prolonged weathering origin can probably be ascribed to a suite of phosphate minerals found in metamorphosed Early Proterozoic iron-rich sediments at Iron Monarch, in the Middleback Ranges. Over thirty phosphate species, including bermanite, collinsite, cyrilovite, fairfieldite, kidwellite, montgomeryite, turquoise, and wavellite, have been identified. A number of vanadates also occur in the assemblage. In central Victoria, small, low-grade phosphate deposits within Ordovician black slate–chert host rocks exhibit a number of mineralization styles, such as phosphorite bands, intraformational breccias, and vein networks. Secondary minerals resulting from weathering of the primary phosphorites include wavellite, turquoise, variscite, cacoxenite, and fluellite.

Amongst the world's largest phosphate deposits are those of Morocco, where Late Cretaceous marine sediments occur on the plains fronting the Atlas Mountains (*see* **Africa:** North African Phanerozoic). These are nodular and sandy deposits riddled with fish teeth and fit well into the upwelling nutrient-rich current model outlined above. Other significant world producers of sedimentary phosphate are the USA, Brazil, and China.

Guano Deposits

Phosphate deposits derived from bird and bat guano represent only a small proportion of the total world reserves of phosphate rock. Insular deposits are common in warm-arid or semi-arid regions with large bird populations either at the present day or in the recent past. The most important deposits, now essentially worked out are on larger islands over 50 metres above sea-level, such as Nauru and Christmas Island, and are thought to be older than about one million years old. In these deposits, solutions derived from overlying bird droppings have percolated into the bedrock, where minerals such as apatite, whitlockite, crandallite, and millisite have crystallized. This phosphatized bedrock forms much of the resource.

Cave phosphate deposits derived from bat droppings are of more interest for the unusual minerals they may contain than for their economic value. Such deposits are mostly in limestone caves, with a minority in lava-tube caves. The chemical reactions involved in forming phosphate minerals are complex, but usually begin with leaching of very soluble nitrogen from the guano. This leaves phosphorus to combine with whatever cations are available from the surrounding rocks. The resulting sequence of minerals may be well stratified within the guano. Typical cave phosphates include brushite, carbonate-hydroxylapatite carbonate-fluorapatite, taranakite, and variscite, generally occurring as powdery nodules within the guano or as coatings on bedrock or cave walls. Distinct crystals of phosphate minerals, such as newberyite and struvite are rare, with a notable occurrence in lava caves at Skipton, Victoria.

Phosphates in Oxidized Metal Sulphide Deposits

Large numbers of phosphate minerals occur in the oxidized zones of base metal orebodies. Solubility phenomena play the most important role in determining which phosphates crystallize in these low-temperature environments, where generally acidic groundwaters dominate. Phosphates of Pb^{2+} are generally the least soluble, so these minerals, particularly pyromorphite (**Figure 9**), are prominent in oxidized zones above galena-bearing ores. As primary ores commonly contain a mix of lead, copper, and zinc sulphides, as well as arsenopyrite, a diverse suite of secondary phosphates and arsenates can form in oxidized zones. Whether phosphates will be prominent over arsenates depends on the availability of phosphorus, usually as apatite, in either the primary ore or the host rocks – it can vary widely. These differences are illustrated by the two most mineralogically diverse oxidized zones known. At Tsumeb, Namibia, arsenates generally dominate the secondary assemblage, whereas at Broken Hill,

Figure 9 Pyromorphite crystals (up to 5 mm long) from Yang Shao mine, Guangxi Province, China. Museum Victoria specimen M48184, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

in New South Wales, Australia, both phosphates and arsenates occur. The phosphate assemblage libethenite–pseudomalachite is particularly widespread above copper-bearing sulphides in arid regions. Notable occurrences of zinc phosphates, including parahopeite, hopeite, tarbuttite, and scholzite, occur at Broken Hill, Zambia, and at Reaphook Hill, in South Australia. The great chemical diversity shown by secondary phosphates is reflected in their often-spectacular colours and crystal habits, making them much prized by mineral collectors.

Geochronological and Thermochronological Applications of Phosphate Minerals

Apatite, monazite, and xenotime commonly contain between a few tens and hundreds of ppm U and Th in their crystal structures. As a result, several different isotopic dating techniques can be applied to them. While the underlying principles, assumptions, and counting methods for each technique are complex and beyond the scope of this review, a brief summary of each is useful. Fission track dating (*see* **Analytical Methods:** Fission Track Analysis) uses damage tracks in apatite arising from the spontaneous fission of ^{238}U, which occurs at a known rate. Measuring the number of tracks that have accumulated since a crystal formed, along with estimating the amount of uranium it contains, means that a geological age can be calculated. Because fission tracks in apatite are 'healed' or annealed at temperatures over about 120°C, only rocks which have not undergone subsequent heating events can be dated this way. However, the annealing properties of apatite fission tracks have led to a growing number of opportunities to model significant thermal processes in the upper parts of the Earth's crust. These include reconstructing the thermal histories of sedimentary basins (*see* **Sedimentary Environments:** Depositional Systems and Facies) and evaluating their potential for oil and gas resources, and estimating the timing and magnitude of erosional and tectonic denudation of mountain ranges (*see* **Plate Tectonics**).

The U–Th–Pb and (U–Th)/He dating methods applied to apatite, monazite, and xenotime have as their basis the decay series of the long-lived isotopes of uranium, ^{238}U and ^{235}U, and of thorium, ^{232}Th. These decay at a known rate through a series of short-lived radionuclides ultimately to Pb isotopes. Determination of the ratios of ^{206}Pb/^{238}U and ^{207}Pb/^{235}U enables a concordia plot to be drawn, which provides an age for the crystals being analysed. However, there are many complicating factors involved in interpreting these plots and in measurement of the data. As well, different methods of determining isotopic compositions are available and need to be selected, depending on which mineral is involved, the precision required, and other factors. The assumption behind the (U–Th)/He method is that all three phosphates appear to retain He, which is produced during the alpha decay of ^{147}Sm. By measuring U, Th, and He contents, an apparent He age can be calculated on the assumption that the initial He content of the mineral was zero. Both the U–Th–Pb and (U–Th)/He methods are still being developed and refined but offer great scope for accurate dating of Earth processes.

Phosphate Biomineralization

The main inorganic constituent of bones and teeth in vertebrate animals, including human beings, is an apatite-like mineral similar to carbonate-fluorapatite. Small amounts of other elements such as sodium, potassium, magnesium, and zinc are present in the structure. The precipitation of apatite takes place after secretion of certain proteins by specialized cells. Other phosphate minerals such as whitlockite, struvite, and brushite, as well as a number of amorphous calcium and/or magnesium-bearing phosphates, have been found in pathological tissue calcifications, such as dental and urinary calculi. Formation of these and other biophosphates is sensitive to conditions such as temperature and pH, so that transformation by dissolution and recrystallization, especially of apatites, may take place. A range of synthetic apatites, in the form of cements and porous ceramics, is now being developed and trialled in order to repair defects and damaged tissue and to correct deformities. These have the capacity to considerably improve both the quality and span of human life.

Environmental Significance of Phosphate

Phosphorus is essential for all forms of life. Considerable cycling of phosphorus takes place within the biosphere and interchange occurs between ecosystems. While the overwhelming amount of phosphorus fluxing takes place between marine organisms and ocean water, human activities play a significant role at the ecosystem scale. The widespread use of phosphate-based fertilisers and insecticides, the disposal of sewerage sludge and industrial waste, including some derived from nuclear reactors, are examples of larger-scale processes that can have serious environmental impacts. Perhaps the best known involves the overload of phosphorus in streams and lakes, leading to an explosion of plant life, especially algae, which upon decay uses up most of the dissolved oxygen. This process, known as eutrophication, results in fish kills (*see* **Fossil Vertebrates:** Fish) and degradation of water quality. On a more restricted scale, there is some evidence for the formation of lead phosphates such as plumbogummite–crandallite and pyromorphite–apatite in soils verging on roads and highways used by vehicles burning leaded gasoline. Increasing awareness of all these problems has meant that control programmes are in place in many regions. Phasing out of lead-based fuels and phosphate-based detergents, together with possible use of crystalline phosphates and phosphate glasses for nuclear waste immobilization, are also helping to improve environmental outcomes.

See Also

Africa: North African Phanerozoic. **Analytical Methods:** Fission Track Analysis. **Fossil Vertebrates:** Fish. **Igneous Rocks:** Granite. **Minerals:** Definition and Classification. **Plate Tectonics**. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Mineralogy and Classification.

Further Reading

Anthony JW, Bideaux RA, Bladh KW, and Nichols MC (2000) *Handbook of Mineralogy.* Volume IV: arsenates, phosphates, vanadates. USA, Tucson: Mineral Data Publishing.

Birch WD and Henry DA (1993) *Phosphate Minerals of Victoria.* Australia, Melbourne: Mineralogical Society of Victoria Inc.

Cook PJ (1984) Spatial and temporal controls on the for mation of phosphate deposits a review. In: Nriagu JO and Moore PB (eds.) *Phosphate Minerals,* pp. 242 274. Germany, Berlin: Springer Verlag.

Filippelli GM (2002) The global phosphorus cycle. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phos phates: Geochemical, Geobiological and Materials Im portance,* Reviews in mineralogy and geochemistry 48, pp. 391 425. Washington, DC: Mineralogical Society of America.

Gleadow AJW, Belton DX, Kohn BP, and Brown RW (2002) Fission track dating of phosphate minerals and the ther mochronology of apatite. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiolo gical and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 579 630. Washington, DC: Mineralogical Society of America.

Hill C and Forti P (1997) *Cave Minerals of the World.* USA, Alabama: National Speleological Society Inc.

Huminicki DMC and Hawthorne FC (2002) The crystal chemistry of the phosphate minerals. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochem ical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 123 253. Washington, DC: Mineralogical Society of America.

Kostov I and Breskovska V (1989) *Phosphate, Arsenate and Vanadate Minerals. Crystal Chemistry and Classifi cation.* Bulgaria, Sofia: Kliment Ohridski University Press.

Moore PB (1973) Pegmatite phosphates: descriptive miner alogy and crystal chemistry. *The Mineralogical Record* 4: 103 130.

Nash WP (1984) Phosphate minerals in terrestrial igneous and metamorphic rocks. In: Nriagu JO and Moore PB (eds.) *Phosphate Minerals,* pp. 215 241. Germany, Berlin: Springer Verlag.

Piccoli PM and Candela PA (2002) Apatite in igneous systems. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 255 292. Washington, DC: Mineralogical Society of America.

Robinson GW, Velthuizen J van, Ansell HG, and Sturman BD (1992) Mineralogy of the Rapid Creek and Big Fish River area, Yukon Territory. *The Mineralogical Record* 23(4): 3 47.

Spear FS and Pyle JM (2002) Apatite, monazite and xeno time in metamorphic rocks. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiolo gical and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 293 255. Washington, DC: Mineralogical Society of America.

Williams PA (1990) *Oxide Zone Geochemistry.* England, Chichester: Ellis Horwood Limited.

Rudaceous Rocks

J McManus, University of St. Andrews, St. Andrews, UK

Introduction and Terminology

Sedimentary rocks in which coarse particles are dominant are termed 'rudites'. They consist of broken fragments, clasts, of pre-existing rocks, and have formed in a wide range of conditions, such as in scree, in landslides, as tills, on alluvial fans and in many sites along river courses, on beaches, in offshore reef-fringing areas and in deep-water environments. Characteristic inter-relationships between the rudites, and other environmentally significant features of these rocks and their associated sediments provide clues to their modes of origin. The clasts themselves provide additional evidence from their shapes and composition. The particle shapes evolve during transport, and textural sorting by size, shape or form may characterize certain depositional conditions. The composition of the particles often indicates the nature of the source from which they were derived. The rudaceous deposits, therefore, provide a stimulating variety of geological challenges at all levels.

The term conglomerate is applied to rudaceous rocks composed of rounded pebbles, and breccia to those composed of angular clasts. A distinction is made between rudaceous rocks where the clasts are in contact with one another, and those in which the clasts 'float' in a finer matrix of sand and clay. These are termed 'clast-supported' and 'matrix-supported' conglomerates respectively. Intraformational rudaceous rocks are composed of clasts of penecontemporaneously cemented sediment; limestone 'beach rock', for example, or intraformational shale pellet conglomerates. Most rudaceous rocks, however, are composed of 'extraformational' clasts derived from outside the formation in which they occur. A further distinction is made between rudaceous rocks composed of many or one rock type. These are termed 'polymictic' and 'ologomictic' conglomerates respectively.

Rudaceous Rock Textures and Fabrics

The size of the particles is of primary importance. Pebbles are defined as particles between 4 mm and 64 mm in diameter and cobbles up to 256 mm. Coarser materials are boulders or blocks. No natural deposits of clastic materials consist of clasts with a single size of particles and a range of diameters are always present. This is partly a function of the material supplied and partly due to variations in the dynamics of the transporting medium and in the conditions during deposition. As large quantities of sediment (often tens of kg) need to be analyzed to obtain statistically meaningful information on coarse grain size populations, such information is relatively rare. In many cases the coarseness quoted is related to the diameter of the largest clast observed, or the average of the largest clasts.

When clasts are released from their source rocks their shapes are defined initially by the distribution of weaknesses in the parental rocks. Fractures such as bedding planes, joints, or cleavages exert a major influence in both the size and shape of the materials produced. Likewise the composition of the bedrock determines the ease with which the large fragments become broken and rounded in transport. Soft, poorly cemented sandstones and limestones break apart more readily and form better rounded clasts than schists, quartzites or granites.

Three aspects of particle shape need to be considered characterizing pebbly materials, namely roundness, sphericity, and form. The three measures may appear related but they address totally different aspects of the clasts. In a numerical sense the roundness is the relationship between the radius of curvature of the sharpest edge and the length of the longest or intermediate axes or a combination of the two. It is conventionally expressed:

$$\text{Roundness} = \frac{\text{average radius of corners and edges}}{\text{radius of maximum inscribed circle}}$$

For speed of processing, most workers assess roundness with the aid of visual comparator charts, as shown in Figure 1.

The sphericity of a clast is the ratio of the diameter, D, of a sphere having the same volume as the clast to that of a circumscribing sphere (i.e., the longest axis, A). It may also be defined as a triaxial ellipsoid based on the product of the lengths of the three diameters of the particle to the volume of the circumscribed sphere, i.e., BC/A^2, where B and C are the intermediate and short axes respectively. Sphericity is also expressed as:

$$\text{Sphericity} = \frac{\text{surface area of the particle}}{\text{surface area of a sphere of equal volume}}$$

Since particles settle through any transporting medium with their maximum projection area

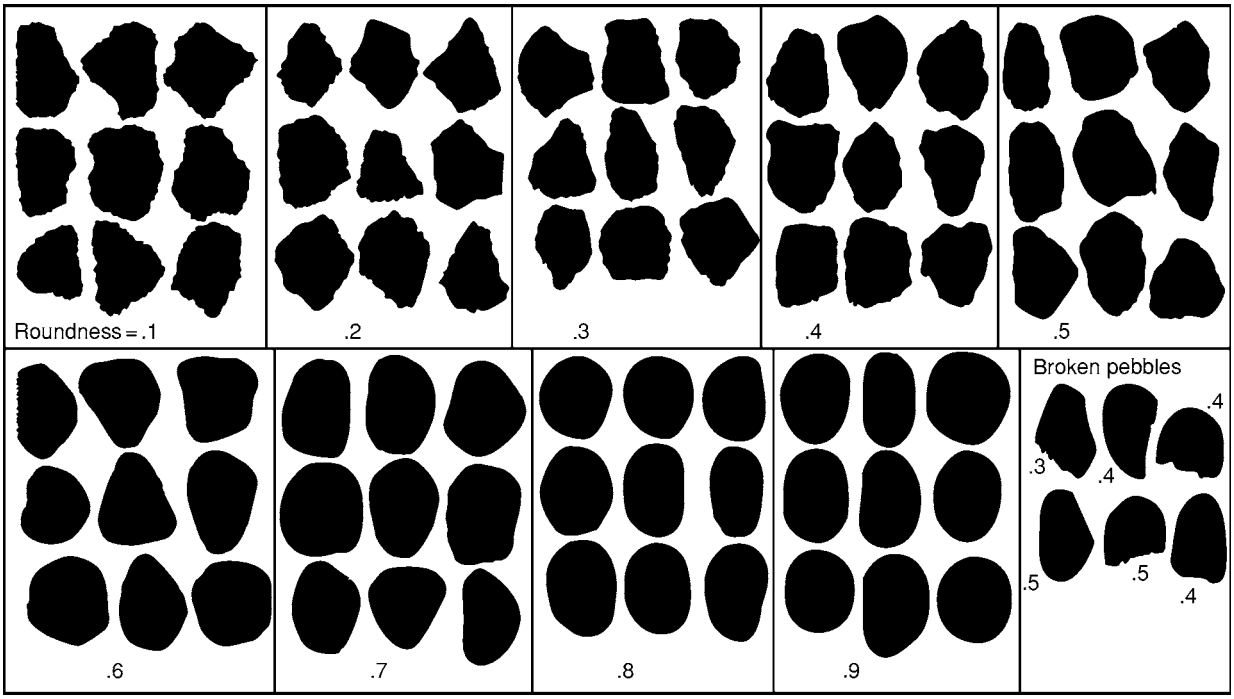


Figure 1 Roundness chart for particles 16 32 mm diameter (Reproduced from Source; Krumbein WC (1941) Measurement and geological significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Petrology* 11: 64 72).

perpendicular to the direction of settling a more dynamically related measure may be obtained by comparing that value with the projection area of the maximum circumscribing sphere as (C^2/AB) 1/3. The reciprocal of this value is of importance, relating to the ease of transport.

Clast form notation (Figure 2) is based on two ratios (2/3) of the particle axes, B/A and C/B to define four form fields: spheres (equant), discs, blades and rods. Using the same three axial lengths it is possible to create a triangular, ten-field form diagram (Figure 3). Although various other combinations of axial ratios have been suggested none has achieved the lasting impact of the above methods of characterizing pebble form. In any attempt to relate the particle form to particular environments of deposition it is necessary to measure a significant number of pebbles (over 200), so that a definitive spread of values may be obtained.

Plotting the values of A against A/B enabled Moss (1962) to identify three particle populations, which he termed framework, interstitial and contact, in a range of gravel deposits. The framework consists of a pebble population graded in size from small and equant particles to relatively large and elongated clasts over a small size range. The interstitial population is subsidiary to and always associated with the framework. Its coarsest pebbles are the same size as, but more elongated than the finest framework material. The contact population, which may be very minor in proportion or may dominate the deposit, is normally coarser than the coarsest part of the framework and characteristically is more equant in form. The contact population is of materials that are unable to fit into the stable gravel bed and commonly move more rapidly along the river than the bulk of the bed material. The value of 1.5 for A/B provides a separation of blades and rods from the discs and spheres, but as indicated above, in natural systems the rods and spheres generally behave similarly. In order to provide a dynamically meaningful plot, while retaining the use of ratios of the A and B axes, the use of D, the volumetrically determined nominal diameter of the clasts, again enables the four particle forms to be recognized (Figure 4).

Since each of the three properties roundness, sphericity, and form are at least partly defined from the particle diameter they are not entirely independent. Along a sediment transport path, such as a river, the mean clast diameter decreases with distance traveled, an exponential relationship, in which the most rapid changes occur near the source and progressively lesser changes in more distal locations (Figure 5).

Both sphericity and roundness increase as the particles decrease in size, and again the changes are most

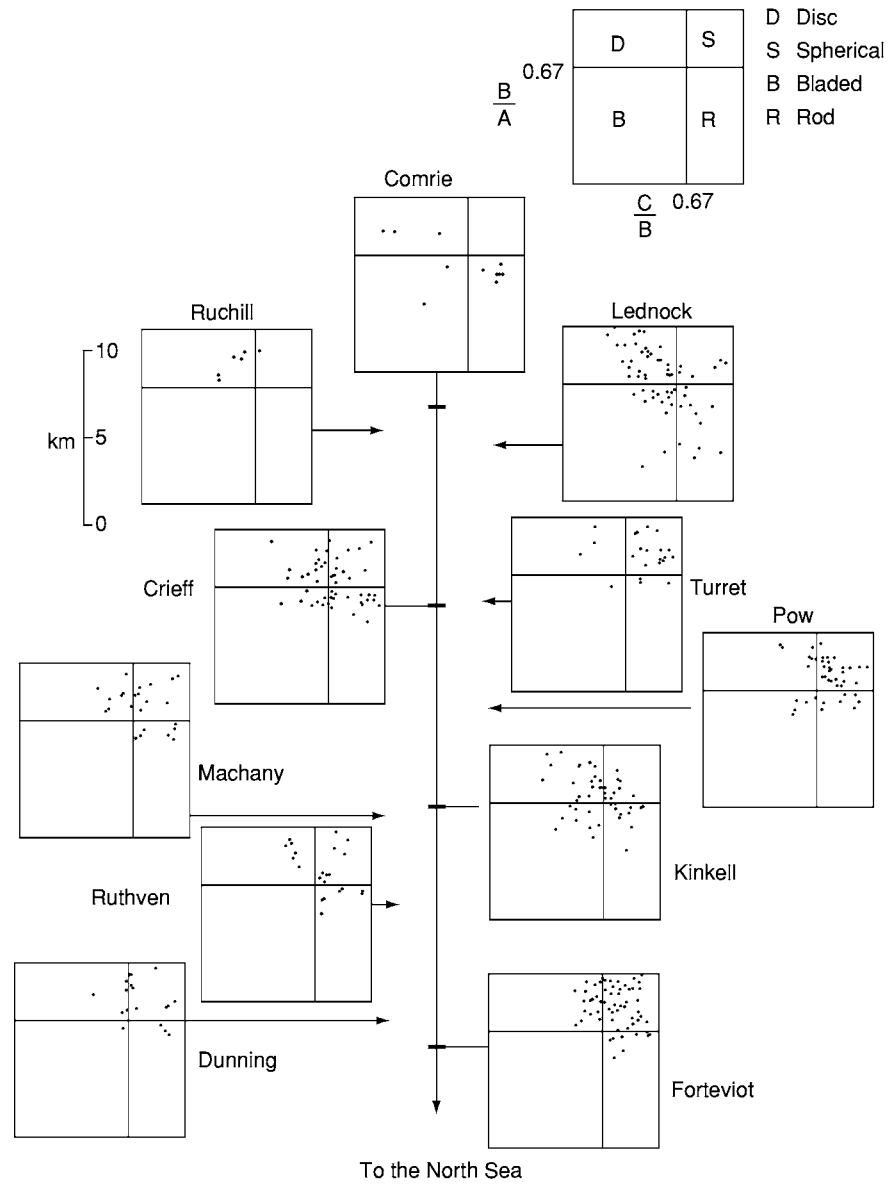


Figure 2 Variation of form of quartz pebbles along the River Earn, Scotland, using the Zingg (1935) plot (after Al Jabbari, Al Ansari and McManus (1982) *Journal of Water Resources* 1, 81 110).

rapid near the source. The main controls of roundness are a) distance traveled; b) composition of the pebble; c) clast size; d) initial clast form; e) nature of the bed material, and f) the dynamics of the transporting medium. Whilst these controls are readily isolated in the laboratory they are not so readily assessed in the field, where most gravel-transporting streams receive detritus derived from tributary catchments that includes pebbles having different transport histories that are added along the length of the stream. Furthermore, erosion of stream bank terraces or tills may lead to the addition of clasts at any point along the stream.

Clasts in Natural Environments

When the coarse particles are released from the exposed rock surfaces to form scree move down slopes debris flows or landslides, or enter streams to be carried into lakes or the sea, where they may form beaches, gravity is the prime motive force. Gravity is all-important in the initial stage, whether the clasts are released from the rock face by frost wedging or by a combination of physical and chemical weathering processes. Within scree, debris flow or landslide particle motion is enhanced by the lubricating and hydraulic effects of water, or ice, working in

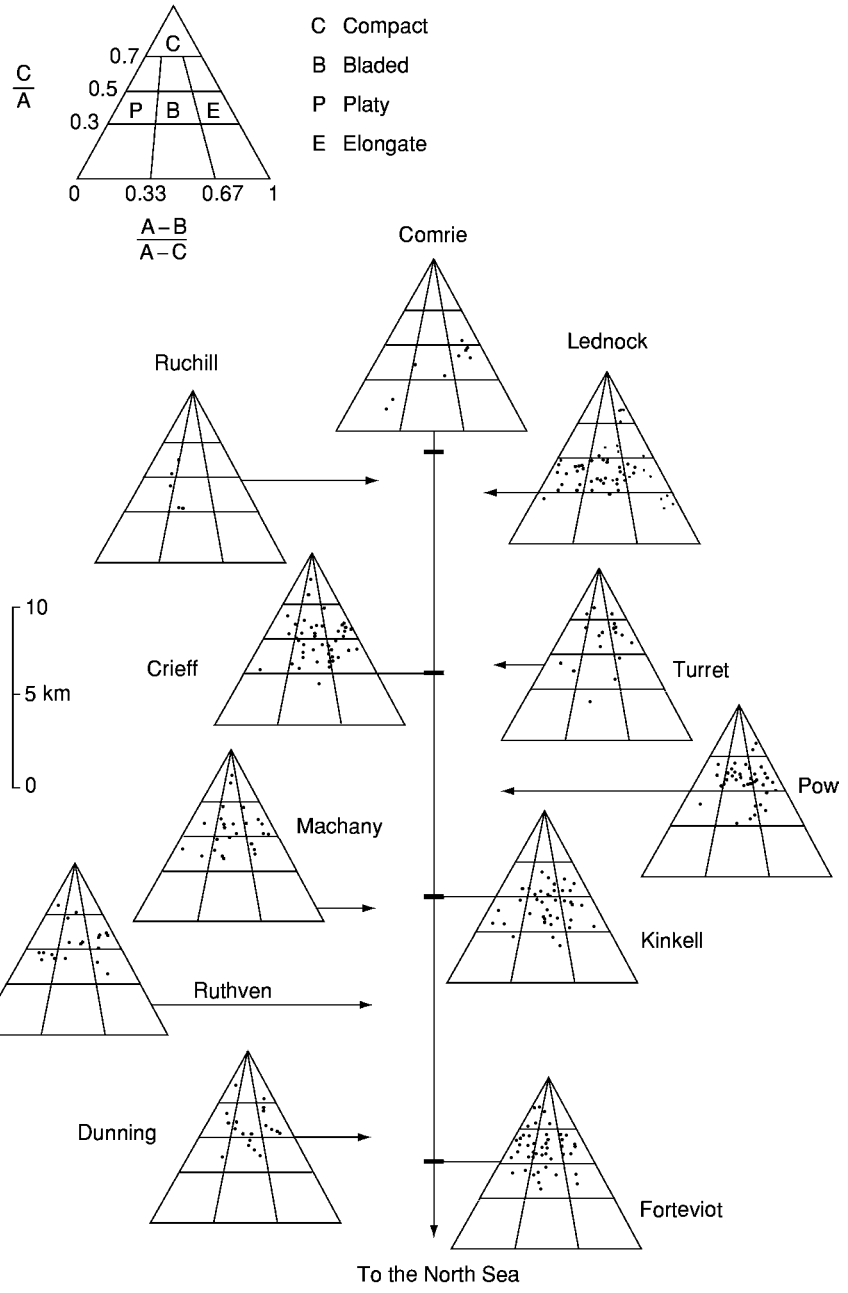


Figure 3 Variation of form of quartz pebbles along the River Earn, Scotland, using the Sneed and Folk (1958) plot. (After Al Jabbari, Al Ansari and McManus (1982) *Journal of Water Resources* 1, 81 110).

combination with gravitational downhill pull, although frictional forces generated by neighbouring materials impede free movement.

Measurement of clasts from screes of different rock types in Scotland revealed that very few of the original clasts fall into the form field for rods or spheres. Virtually all were either discs or blades when released into the environment. Both the composition and structure of the materials, as well as physical weathering by frost action, are believed to exert major controls on form.

Clasts in Streams

When a mixed population of particles occurs on a stream bed the finest, silt-sized particles are carried away in suspension, and the sands saltate downstream, bouncing along with the flow. The pebbles normally remain on the bed, often with their upper parts extending through the boundary layer of the flow. Once flow strength exceeds some critical value the smaller pebbles begin to slide or roll along the bed, and as flow strength increases so increasing quantities

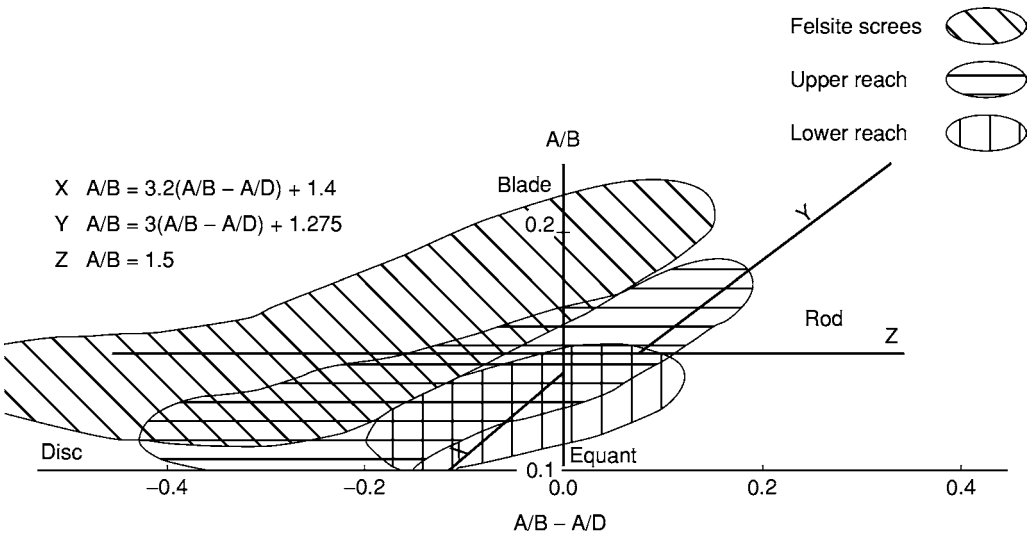


Figure 4 Variation of pebbles form from scree, and the upper and lower reaches of the Shee Water, Scotland, using the A, B and D diameters (after Dhiab IH 1979, unpublished M.Sc. Thesis, University of Dundee).

migrate until, in extreme floods, most of the bed is in motion. As the power of the flow decreases so particles become deposited and the bed aggrades, as material accumulates. During floods the entire bed down to bedrock may become removed from a reach of the river, to be replaced by new material as the peak of flooding passes. There are records of over 8 m of bed removal and replacement during individual storms in the western USA. The various interrelationships between water flow, sediment size, sediment load, and stream slope are summarized in Figure 6.

Under normal conditions the 'vibration' of the turbulent flow induces minor movement within the bed, leading to the upward migration of the coarser particles, while finer materials move downwards into the deposit. In this way the stream bed develops a residual lag of coarse pebbles, and there is a progressive downward fining of the gravels. Frequent inter-pebble contacts induce progressive rounding of edges through abrasion in the bed and surfaces exposed on the streambed become scoured through collisions with the saltating sand grains.

Within the stream bed the clasts may become arranged such that an internal texture, related to both size and form of the particles is developed. The texture results from burial of blades or discs whose intermediate axes slope upstream (up current). This is termed imbrication. During motion the more equant particles roll along the bed once they have been disturbed, whereas the flatter clasts are more likely to slide across the other pebbles, turning over as they encounter immobile clasts. The ability of the clast to find a place into which it can fit in the bed often determines the distance that it travels during any single displacement from the bed. The more nearly spherical the particle the greater the difficulty it experiences and the further it is likely to move once displaced from its resting place.

Attrition of the sliding and turning particles increases the probability of their splitting to form smaller more equant or rod-shaped pebbles. There is commonly a size gradation along the length of a gravel bar within a braided stream reach, with the coarsest clasts at the upstream end and progressively smaller ones downstream. As breaking of the pebbles occurs along the river so there is a downstream increase in the proportion of spherical clasts within the bed load.

Clasts on Beaches

Once the pebbles reach the sea, either from a river, or cliff collapse, they become subject to the motion of waves, which carry them along the coast by longshore drift, and deposit them on beaches. Beaches formed entirely of gravel occur on dynamic coasts. The presence of sand is an indication of quieter conditions. The latter used the 10-fold form diagram to differentiate between beach and river gravels with some success.

The gravel beaches of South Wales develop a fourfold structure, recognized on the basis of pebble size, form and the texture of the deposit. Between an upper zone of coarse discs and a lower zone principally of spheres lies a zone with imbricate discs, which passes down slope into a infill zone comprising particles of many shapes, and often incorporating sand (Figure 7).

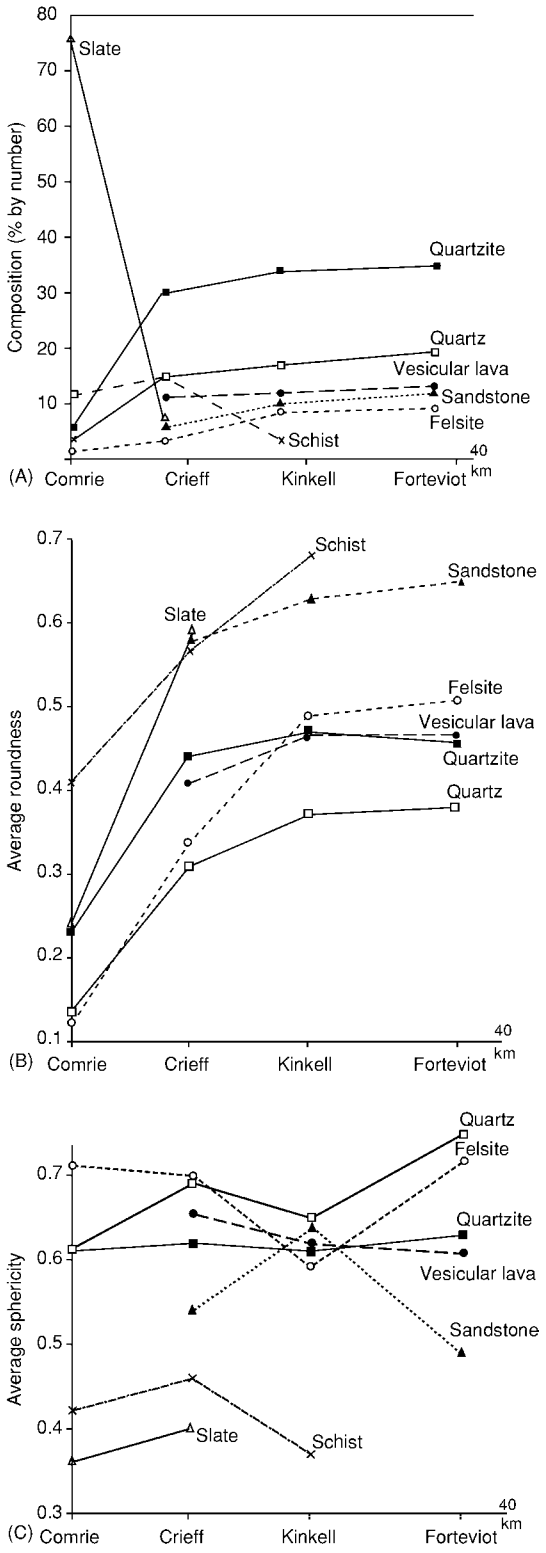


Figure 5 Variation of A) the pebble composition, B) roundness, and C) sphericity along the River Earn, Scotland (after Al Jabbari, Al Ansari and McManus (1982) *Journal of Water Resources* 1, 81 110).

As clasts are brought to a beach they are carried by waves that wash over the surface many times during transport of the particles. The mobile clasts interact with the underlying gravels. If they are larger than the pore spaces between the settled clasts the pebbles bridge the gap and continue to migrate along or up and down the beach. If the clasts are too small they enter the open spaces, but wash through to continue migrating. A clast of similar size will fully occupy the space provided. In this way clasts of similar size and form characteristics gather together usually on the lower parts of beaches and provide 'selection pavements', effectively providing a gauntlet through which clasts must pass if they are to move up or down the beach face. Noting the internal structure of many beaches in western Scotland, Bluck examined gravel beaches forming and migrating through time, drawing attention to the prevalence of imbricate structural zones in almost all beaches examined (Figure 8).

Particle tracing techniques using dyes, paint or creating artificial electronically tagged pebbles have enabled the motion of individual pebbles to be tracked sometimes for periods exceeding ten tides. Disturbance of the sediment often penetrates to 10–20 cm below the beach surface under moderate sea conditions. Large clasts migrate along the beach face more rapidly than do their smaller counterparts, and under identical conditions (on the same days) clasts of ironstone migrate more slowly than similar clasts of sandstone, and they in turn are more sluggish than matched coal fragments. The density of the particles provides the control in this case. On many British beaches rates have been measured of longshore movement of 5–8 cm diameter clasts of flint, chert, sandstone and ironstone of up to 10 m per tide under moderate wave conditions. This suggests that many of the features illustrated by Bluck may be essentially short lived, although regularly regenerated in the same locations.

Where steep rocky cliffs lie behind the coast, direct cliff fall contributes boulders and cobbles to the beach. The large clasts become rounded through attrition but remain on the beach.

Ancient or 'fossilized' cliffed coasts are rarely preserved, but at Enard Bay, north-west Scotland, a Precambrian coastline has been exhumed showing cliffs cut into Lewisian Gneiss, and cut into Lewisian gneisses with Torridon Group marginal fanglomerate and beach deposits banked against it (Figure 9).

Isostatic uplift following deglaciation has allowed many former coasts to rise above current sea level, and the raised beach features preserved around many northern European and north American coasts display most of the internal textures explored by Bluck.

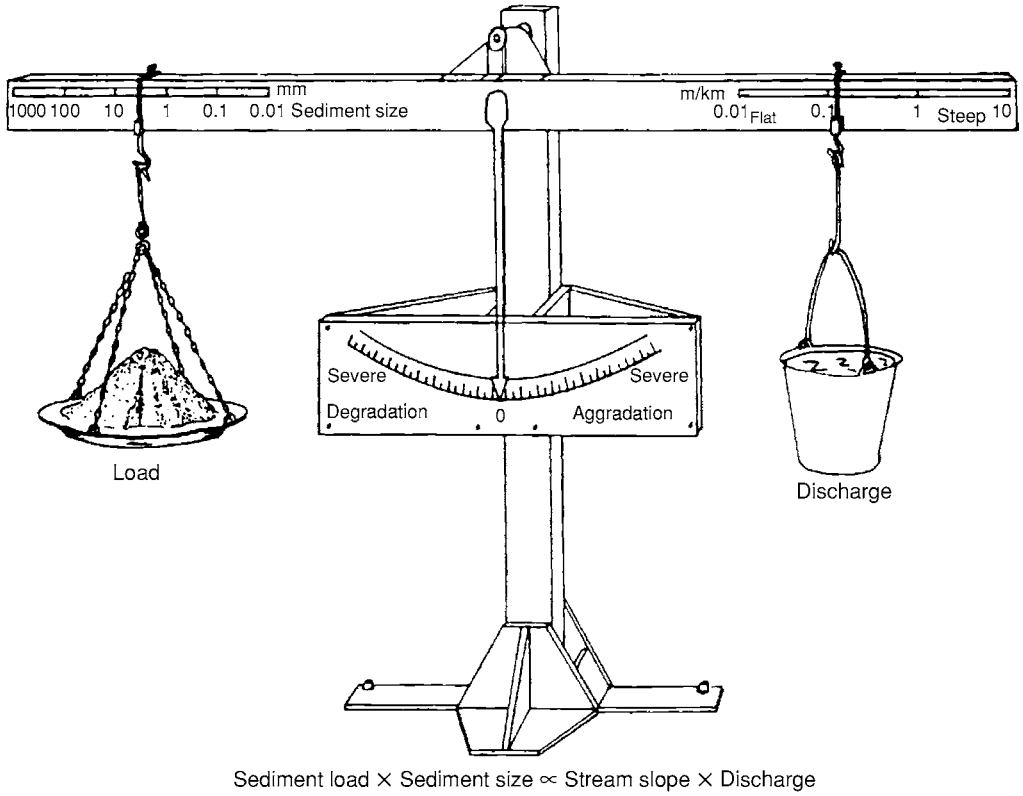


Figure 6 Interrelationships between factors controlling a stable channel bed in a river (after Lane 1955).

Lakeshore gravel beaches are common, but the low power in the waves generated on small lakes ensures that the beaches rarely achieve a fully mature condition. The large clasts are commonly essentially static, show limited size and shape sorting, and are often poorly rounded. Rods and spheres are generally more common in the coarser particles than in the smaller clasts.

In shallow tropical or subtropical seas carbonate sediment dominates. These include in situ reef rock, and carbonate sands and muds. All of these may become quickly cemented, and subsequently eroded during heavy storms generating large amounts of rudaceous detritus. An apron of limestone breccia may extend into deep waters at the foot of reefs and other abrupt carbonate shelf margins. Such reef apron breccias have been recognized in the Permian Capitan Reef of Texas and New Mexico, around many Carboniferous structures in Britain and Belgium, and round Cretaceous reefs in the French Juras.

Clasts on Alluvial Fans

Adjacent to newly emergent mountains, or uplifting fault blocks rapid changes of stream bed gradient ensure that deposition of sediment occurs adjacent to the source, creating alluvial fans of coarse detritus. These sediments are particularly well preserved in arid or semi-arid areas providing important sites in which rudaceous materials accumulate. In the geological past, before evolution of extensive land-living plants in the Devonian, fans were widely developed and thick accumulations of pebbly sandstones mark the margins of areas of active uplift.

The evolving alluvial fans, or the coalescent adjacent fans forming a bajada, produce wedges of sediment, commonly but against a fault. Repeated uplift generates successions of wedges of coarse materials stacked above each other, with thicker beds of coarse sediments near the source, grading to thinner beds of finer materials in more distal areas. The uppermost parts of the fans are characterized by the presence of debris- and mud-flow deposits, each of which contain large clasts, but the latter also contain much fine sediment. In essence the water drains from the moving sediment-enriched flow to induce deposition. The fan head is dominated by ribbons of the mass flow deposits, through which the streams erode as they flow towards the lower, outer parts of the fan, where slopes decrease from 5° to less than 1° and the waters become divided into many distributaries. The sediments of the outer part of the fan are dominated by sheets of sand or gravel from the often ephemeral, braided streams. Characteristically, the fan sediments

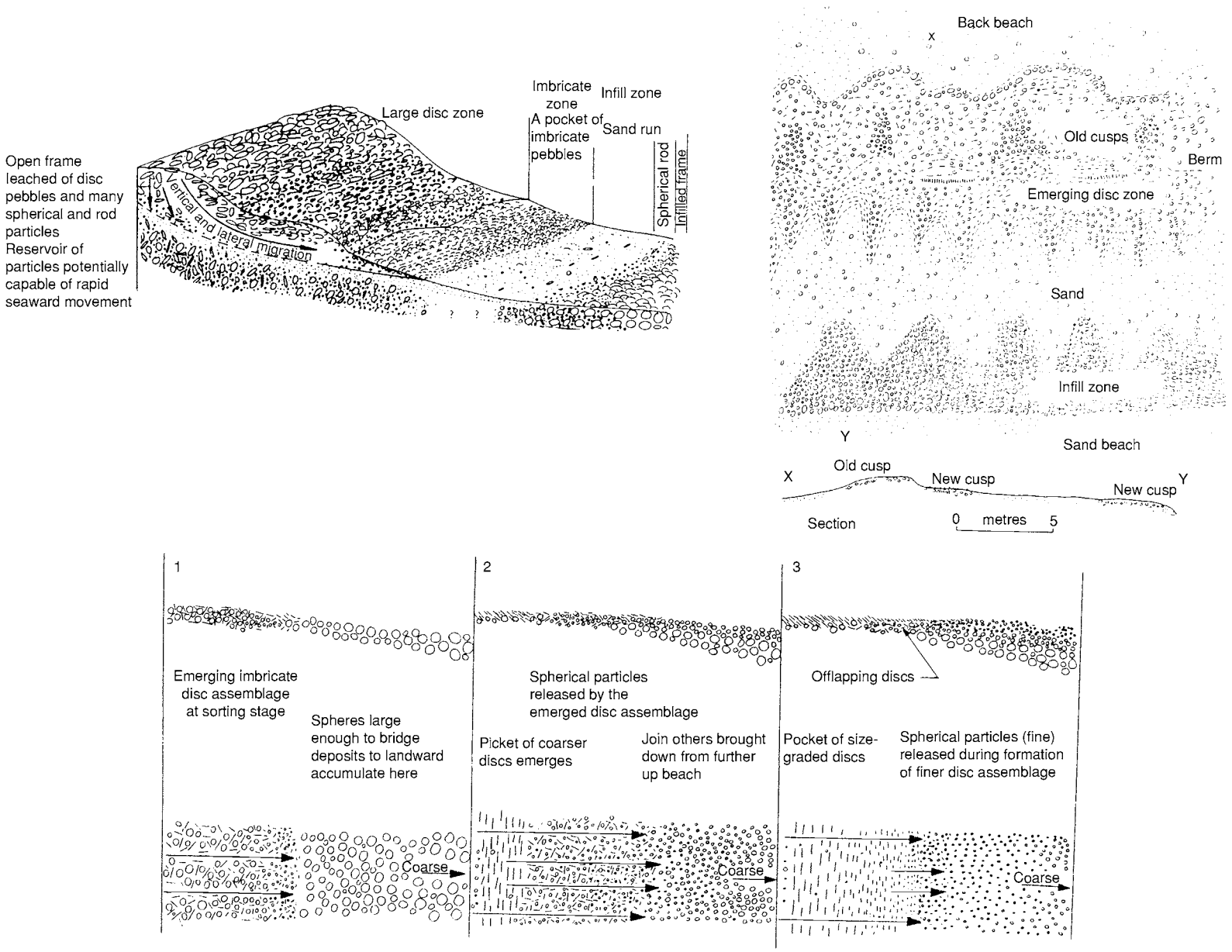


Figure 7 Arrangement of clasts on gravel beaches (after Bluck BJ (1999) *Transactions of the Royal Society of Edinburgh* 89, 291 323).

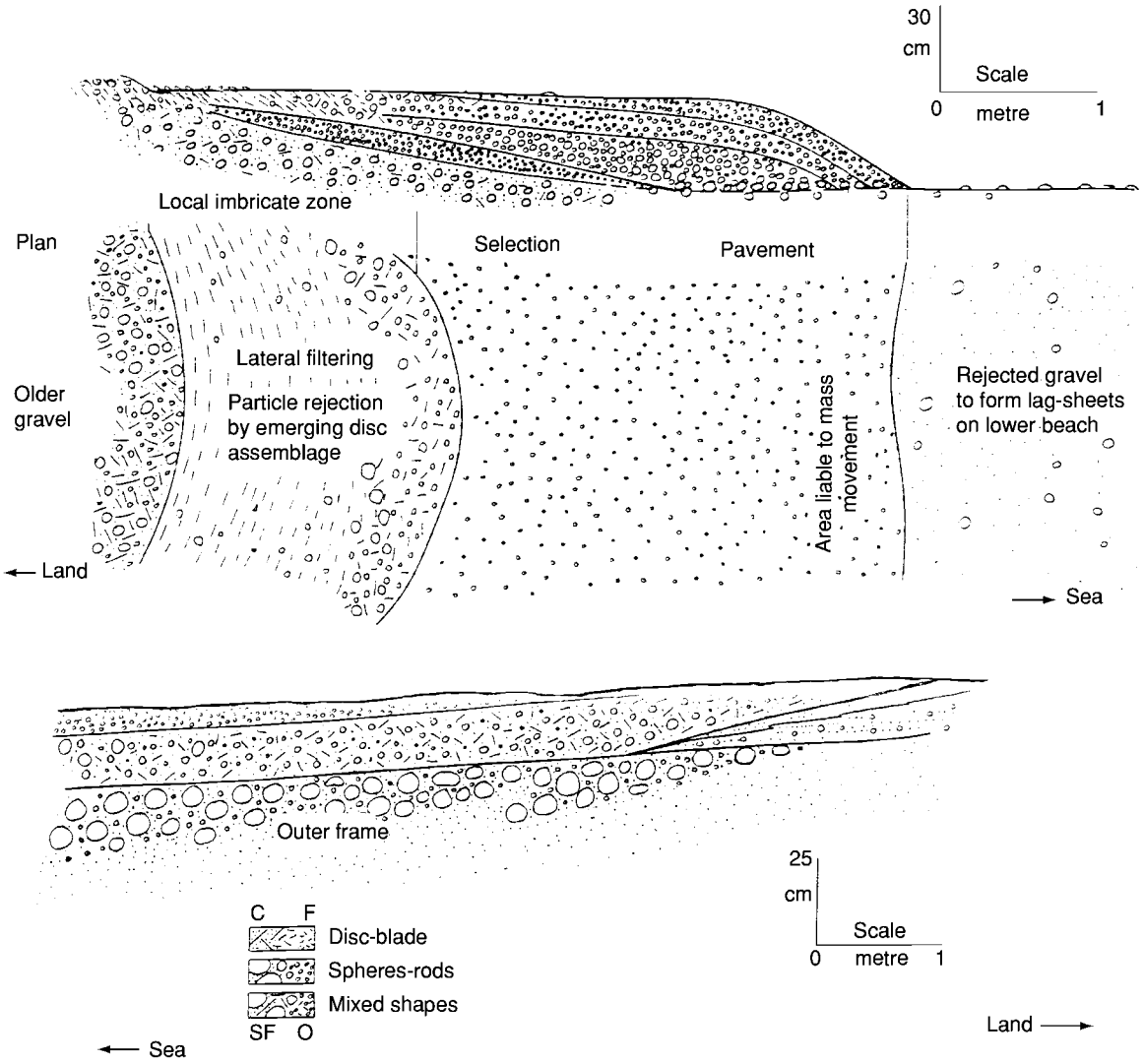


Figure 8 Schematic cross sections of beaches from south west Scotland (after Bluck BJ (1999) *Transactions of the Royal Society of Edinburgh* 89, 291 323).

are coarse, and dominated by gravels in the upper reaches, becoming more fine grained in the lower areas.

Sheet-like conglomerates with pebbles of similar sizes result from the formation of temporary selection pavements during sheet flows of flood discharge. Other deposits fill shallow channels scoured into pre-existing sediments. Sometimes their steeply inclined imbricate clasts serving to identify channel margins, while more gently inclined clast axes occur in the central parts of the channels. The maximum clast size decreases exponentially down fan, as does the thickness of the individual conglomeratic beds.

The alluvial fan-bajada systems may extend for 25–30 km from the fan head on to alluvial plains, for tens of kilometres along active mountain fronts, and, where long-lived geologically the deposits may reach several kilometres in thickness. Such dimensions are recorded from modern systems in Nevada, ancient fans in Texas, and from Neogene deposits of Italy and Switzerland, related to the rising Apennines and Alps respectively. In Britain large ancient fans have been identified from the Applecross Formation of the Torridonian (Late Precambrian) of northwest Scotland, with smaller fans in the Devonian successions of the Midland Valley and Orcadian basins of Scotland, and of Permian age in south-west England.

Clasts in Braided Rivers

The outermost parts of some fans are dominated by braided streams, whose normally shallow channels of low sinuosity, become subdivided by mid-channel

Figure 9 Exhumed cliffline and beach sediments of Early Torridonian age, Enard Bay, north west Scotland.

bars. The braided rivers beyond the fans typically develop in areas with greatly varying water discharge histories. The wadi-floor streams of desert and semi-desert areas, for example, receive little water for many months before floods sweep through the area carrying sediment-charged waters capable of transporting material up to the size of large boulders. Thus more people drown in deserts than die of thirst. As the waters subside the sediments are rapidly deposited to give a layer of coarse, matrix-supported gravel.

Another important site for the formation of braided systems is in the periglacial sandurs and plains associated with glacial retreat, such as those of Iceland. The rivers carry little water during the winter months, but in summer may carry large quantities of glacial melt water, accompanied by the transport of high sediment loads, which become deposited as the bed gradient and flow velocities fall. These streams drain areas in which sediments of all size are available and movement is minimally restricted by vegetation. In rift valleys rudaceous marginal fault-bounded fanglomerates may pass out into braided river sands and gravels on the floor the central parts of basin (Figure 10).

The deposits of braided rivers, explored by Miall and Bluck, typically show successions of fillings of stacked channels, some of the major stream, and others of second or third-order channels (Figure 11). Typically upward fining sheets of sand and gravel result from the migration of mid-channel and overbank bars that are the principal sites of deposition. Miall showed that in the upper reaches of some

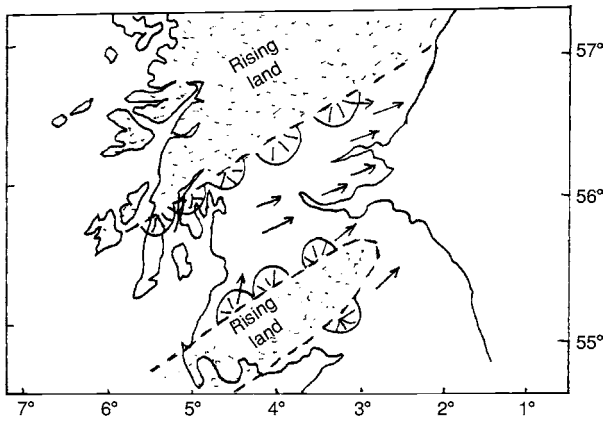


Figure 10 Suggested distribution of Upper Old Red Sandstone alluvial fans in central Scotland (after Trewin NH and Thirlwall MF (2003) Old Red Sandstone. In: Trewin NH (Ed.) *The Geology of Scotland*, 4th edn, pp. 213 251. London: The Geological Society of London).

braided systems debris flows occur within the more normal upward fining cycles of the flood deposits. He identified three pebble-rich assemblage types, dominated by channel gravels with intervening debris flows, by superposed channel bars or by channel floor gravels passing upwards into current bedded sands. Pebbles also occurred less frequently in other sand-rich braid deposits. In the British geological column the braided systems have been recognized from the Precambrian Torridonian, in both the upper and lower Old Red Sandstone (Figure 12) of the Devonian and in the New Red Sandstones of the Permo-Trias.

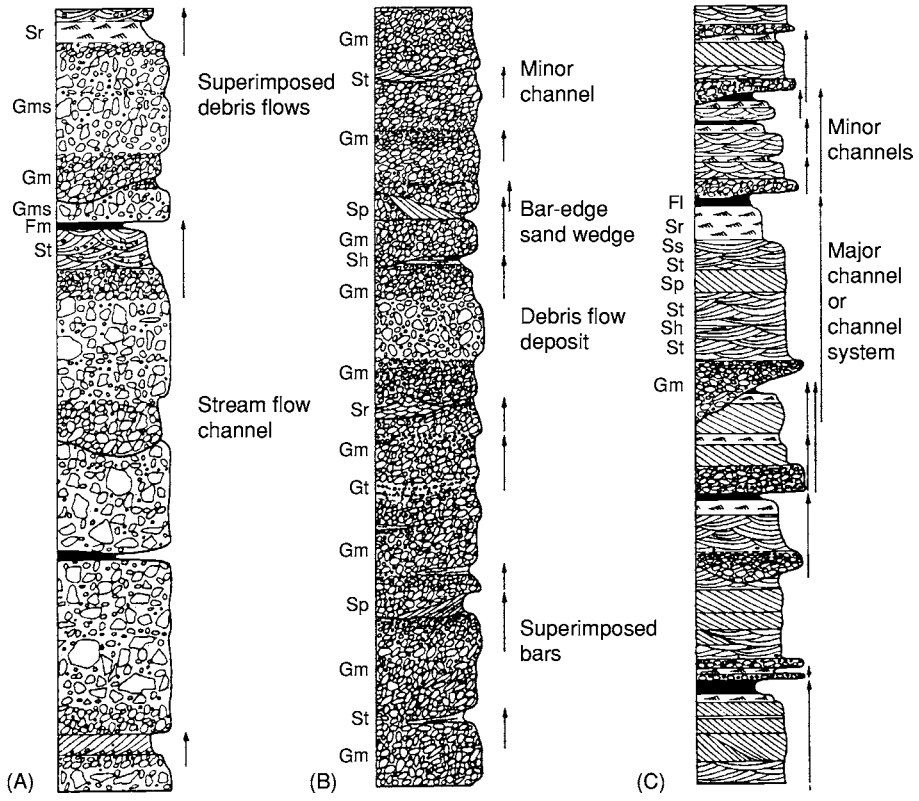


Figure 11 Upper Old Red Sandstone braided river deposits at Whiting Ness, Arbroath, Scotland, showing the flood plain deposits against a buried unconformity, with associated debris flows, the presence of secondary channels and the dominance of bar head deposits.

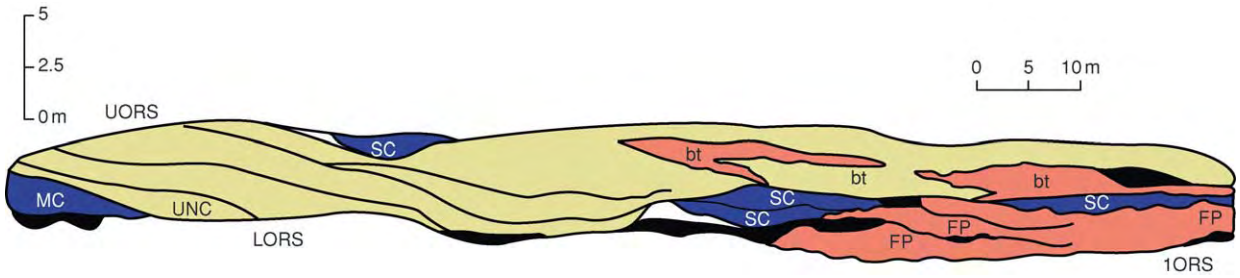


Figure 12 Coarse Lower Old Red Sandstone conglomerates at Stonehaven, Scotland. The larger clasts are over 1 m in diameter.

Clasts in Tills

Rudaceous sediment is found in glacial tills (*see* **Sedimentary Processes:** Glaciers). Tills may be composed of any mix of coarse and fine materials. Whereas in the mountains large boulders are common in clast-supported tills, in more distal areas the more readily transported fine materials dominate. Two forms of till are recognized. Lodgment tills are in direct contact with the underlying rocks, and have usually been deposited beneath the moving glacier, and been partly compressed by the weight of overlying ice. Particles in this deposit may be striated. The overlying ablation till is structurally weaker, having been deposited from down-melting ice. Imbrication, indicating the direction of ice flow, may form in clast-supported or matrix-supported tills. Often in areas away from the mountain sources the large Pleistocene ice sheets became enriched with locally derived materials in addition to those derived from upper catchment areas. Tills containing glacial erratic clasts have been recorded from the Precambrian and Permo-Carboniferous and Cretaceous glaciations in many parts of the world.

Deep Water Rudaceous Deposits

Rudaceous deposits, both terrigenous and carbonate, occur in deep-water environments, ranging from turbidites to mega-boulder complexes. Gravels in turbidites are restricted to the basal part of the graded bed, and increase in size and abundance towards the source. Isolated clasts occur in debris and grain flows. At the other extreme is the 'Wild-Flysch' of Alpine geologists, termed 'olostostroma' by Italians. These are irregularly shaped formations that contain clasts the size of skyscrapers and jumbo-jets. Such deposits usually occur at the foot of submarine fault scarps, and are associated with tectonic disturbances of violent intensity.

Conclusions

The clasts of rudaceous sediments hold important information about not only the rock types of the hinterland from which they were derived, but also about its geological history. As a terrain is unroofed it will shed progressively older and more lithified clasts into the depositional system. With continuous or discontinuous uplift erosion unmantles progressively older or more changed, often more highly metamorphosed materials, which are transported and deposited in the resultant conglomerates and breccias. Structurally or compositionally weak rocks do not preserve as well as stronger materials and allowance must be made in attempting to reconstruct unroofing histories. Furthermore, it is generally the more chemically stable silica-rich rocks that contribute to rudaceous deposits. Thus, of all of the sediments it is sandstones and cherts that are preserved at the expense of shales and limestones, of metamorphic rocks quartzites are preserved at the expense of slate and schist. Of all of the igneous rocks, pebbles of rhyolite are more usually preserved than those of basalt or gabbro.

Now that our understanding of the processes leading to the formation of rudaceous deposits is fairly advanced, much present research is moving into the field of exploring the geological characteristics of ancient catchments, even to the level of distinguishing separate phases of advance of thrust sheets into an area during orogeny.

The rudaceous rocks have much to offer the sedimentologist and the geological historian. The ability to recognize particular depositional environments in the ancient record and to recreate the conditions at the land surface during mountain-building enables the geologist to postulate the locations of potential metalliferous and hydrocarbon economic resources. Gold and uranium occur in Precambrian rudaceous rocks in Canada, the USA, Brazil and South Africa, wherein those of the Witwatersrand basin are probably the best known. Rudaceous rocks host placer ores in many parts of the world. Because they are composed of clasts, which of their very nature are tough, and therefore of low porosity, Rudaceous rocks are seldom good petroleum reservoirs. But it is as aggregates for road building and construction that unconsolidated rudaceous sediments are economically most important.

See Also

Sedimentary Environments: Alluvial Fans, Alluvial Sediments and Settings; Lake Processes and Deposits; Shoreline and Shoreface Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures; Fluvial Geomorphology; Glaciers; Landslides. **Weathering**.

Further Reading

Bluck BJ (1967) Sedimentation of beach gravels; examples from South Wales. *Journal of Sedimentary Petrology* 37: 128 156.

Bluck BJ (1980) Structure, generation and preservation of upward fining braided stream cycles in the Old Red Sandstones of Scotland. *Transactions of the Royal Soci ety of Edinburgh, Earth Sciences* 71: 29 46.

Bluck BJ (1999) Clast assemblages, bed forms and structure in beach gravels. *Transactions of the Royal Society of Edinburgh, Earth Sciences* 89: 291 323.

Bluck BJ (2000) Old Red Sandstone basins and alluvial systems of Midland Scotland. In: Friend PF and Williams BPJ (eds.) *New Perspectives on the Old Red Sand stone,* 180, pp. 417 437. London: Geological Society of London.

Bray M, Workman M, Smith J, and Pope DJ (1996) Field measurements of shingle transport using electronic tracers. In: Proceedings of 31st Ministry of Agriculture, Fisheries and Food Conference on River and Coastal Engineering. Keele: University of Keele.

Bull WB (1977) The alluvial fan environment. *Progress in Physical Geography* 1: 222 270.

Cailleux A (1945) Distinction des galets marins et fluviatiles. *Bulletin of the Geological Society of France* 5: 125 138.

Dobkins JE and Folk RL (1970) Shape development on Tahiti Nui. *Journal of Sedimentary Petrology* 40: 116 203.

Glennie KW (2002) Permian and Triassic. In: Trewin NH (ed.) *Geology of Scotland,* 4th ed, pp. 301 322. London: The Geological Society of London.

Griffiths JC (1967) *Scientific Method in the Analysis of Sediments,* p. 508. New York: McGraw Hill.

Krumbein WC (1941) Measurement and geological signifi cance of shape and roundness of sedimentary particles. *Journal of Sedimentary Petrology* 11: 64 72.

Lane EW (1955) Design of stable channels. *Transactions of the American Society of Civil Engineers* 120: 1234 1279.

Laming DJC (1966) Imbrication, paleocurrents and other sedimentary features in the lower New Red Sandstone,

Devonshire, England. *Journal of Sedimentary Petrology* 36: 949 959.
Miall AD (1977) A review of the braided river depositional environment. *Earth Science Reviews* 13: 1 62.
Moss AJ (1962) The physical nature of common sand and pebbly deposits 1. *American Journal of Science* 262: 337 373.
Moss AJ (1963) The physical nature of common sand and pebbly deposits 2. *American Journal of Science* 263: 297 343.
Selley RC (1965) Diagnostic characters of fluviatile deposits of the Torridonian. *Journal of Sedimentary Petrology* 35: 366 380.
Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.
Sneed ED and Folk RL (1958) Pebbles in the lower Color ado River, Texas, a study in particle morphogenesis. *Journal of Geology* 66: 114 150.
Trewin NH and Thirlwall MF (2002) Old Red Sandstone. In: Trewin NH (ed.) *The Geology of Scotland*, 4th edn, pp. 213 251. London: The Geological Society of London.
Wadell H (1935) Volume, shape, and roundness of quartz particles. *Journal of Geology* 27: 507 521.
Williams GE (1968) Neoproterozoic (Torridonian) alluvial fan succession, northwest Scotland, and its tectonic set ting and provenance. *Geological Magazine* 138: 161 184.

Sandstones, Diagenesis and Porosity Evolution

J Gluyas, Acorn Oil and Gas Ltd., Staines, UK

Introduction

Sand comprises particles of rock and mineral with a mean grain size between 0.0625 and 2 mm and deposited by sedimentary processes on the Earth's surface (*see* **Sedimentary Rocks:** Mineralogy and Classification).

The composition of sand is highly variable, depending on the source of the sediment and the extent to which weathering and erosion during transport have removed unstable minerals (*see* **Weathering**). As a general rule, sands derived, first cycle, from igneous and metamorphic terrains tend to contain more mineral phases that are unstable under surface and shallow burial conditions than do sands that have been involved in many cycles of erosion, transport, and deposition.

Following deposition, sand may become buried. It may also be lithified (indurated) into sandstone. The process whereby sand becomes sandstone is known as diagenesis (*see* **Diagenesis, Overview**). It includes three distinct components: one mechanical (compaction) and two chemical (cementation and dissolution).

When sands are deposited, they are commonly highly porous and highly permeable. Any given volume of newly deposited sand will contain between 40% and 50% pore space (Figure 1). The permeability of loose sand is enormous, measured in tens to hundreds of darcy. Sandstones are less porous and less permeable, there being a continuous range from the values for sand shown above to sandstones that are non-porous and impermeable.

Grain Size and Sorting

Sand having grain sizes between 0.0625 and 2 mm is further divided into a series of subcategories, from very fine sand at the lower end of the size range to very coarse sand at the upper end of the range. Smaller grains (silt and clay grade) and larger grains (granules to boulders) are defined in **Sedimentary Rocks:** Mineralogy and Classification. Grain size is governed by the grain or crystal size in the provenance area and the degree of abrasion suffered by the sediment *en route* from the source area to deposition. Sorting is a measure of the range of grain sizes in a given sand sample. Well-sorted sand has a narrower range of grain sizes than poorly sorted sand. Sorting within sand is controlled by both provenance and sedimentary process. Surface processes which constantly rework sediment, such as in shallow marine settings (wave

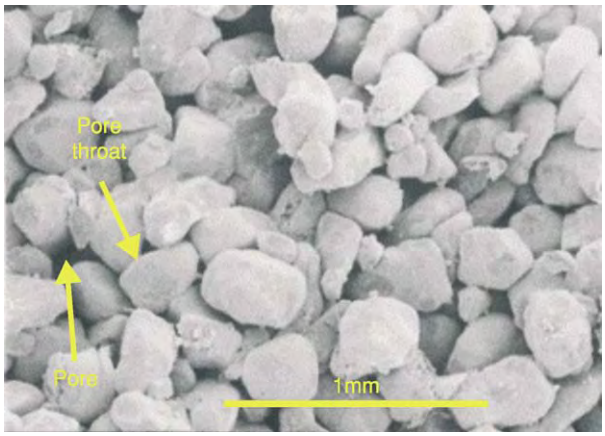


Figure 1 Well sorted, medium grained, aeolian, uncemented sandstone from the Cleeton Field, UK southern North Sea, showing well developed pores and pore throats; scanning electron photomicrograph. From Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

and tidal action), are likely to produce better sorted sand than, for example, gravity-driven processes such as debris flow. However, if the sediment provenance area comprises well-sorted sand, so too will the deposition area, irrespective of the specific sedimentary process responsible for deposition.

The grain size and sorting of sand control its initial permeability and the sorting of sand controls its initial porosity. Moreover, as compaction and diagenesis proceed, the 'memory' of the depositional characteristics can be retained, such that the sands that were the most permeable at deposition become the most permeable sandstones after compaction and cementation.

Compaction

Loose sand compacts easily. During initial burial, much of the compaction is taken up by the rearrangement of grains. Simple burial combined with seismic shock will turn loose sand into consolidated sand. The amount of porosity lost will depend largely on how well sorted the sand is. In poorly sorted sand, more porosity will be lost than in well-sorted sand – small grains fill in between larger ones. As burial continues, rough edges tend to be knocked off grains, so aiding greater compaction. At deeper levels (about 1–4 km), the sand begins to behave like a linearly deformable solid. Deeper still, plastic deformation is probably more common. The boundaries between the occurrence of these processes will vary from sand to sand and basin to basin and, in some instances, may be gradational.

The net outcome of all the above processes is that sands compact when stressed, but decompact very little when the stress is released. This means that, for compacted but uncemented sandstone at the Earth's surface, it is possible to calculate the maximum stress suffered in any previous burial phase. Such a stress calculation can be used to provide an estimate of the maximum burial depth.

In the geological literature, there are a large number of so-called compaction curves for sandstones. Alas, most of these curves are porosity/depth plots, rather than porosity/stress plots. As such, the great swathes of data on these plots include, but do not differentiate, the effects of fluid overpressure and sandstone cementation. However, experimental data are available on the way in which sands compact and, for clean quartzose or arkosic sandstones, these data have been used to formulate a compaction equation

$$\Phi = 0.5 \exp\left(\frac{-10^{\ 3}z}{2.4 + 10^{\ 4}z}\right)$$

where z is in metres.

In this equation, porosity is expressed as a fraction (i.e., <1) and the equation is calibrated to a normal hydrostatic pressure gradient. If the system is overpressured, the pressure borne by the grains is less than in a hydrostatic system and an effective depth must be calculated. As a simple rule of thumb for typical burial depths of 2–4 km, 1 MPa of overpressure is equivalent to about 80 m less burial. The equation is well tested, predicting porosity to within $\pm 3\%$ at 95% confidence limits.

Sands that contain easily squashed grains, such as glauconite or mica, and those rich in matrix clay lose porosity much more readily at a given applied stress. Empirical curves linking porosity to applied stress have also been constructed for sands with various quantities of easily deformed grains.

Detrital Mineralogy

Quartz is the most common mineral found in sands and sandstones (*see* **Minerals:** Quartz). Feldspar and lithic (rock) fragments are also common in most sandstones and, as a consequence, these three components are often used to classify sandstones. The QFL plot sums the three components (quartz, feldspar, lithics) to 100% on a triangular diagram (Figure 2). The triangle is divided into fields: quartz arenite, sublithic arenite, arkosic arenite, etc. There is no strict convention as to whether polycrystalline quartz is included with (monocrystalline) quartz or with lithic fragments, although it is common to label the diagram so as to show where the polycrystalline quartz has been included. The feldspar component includes both alkali and plagioclase, whilst the lithic component can include sedimentary, igneous, and metamorphic

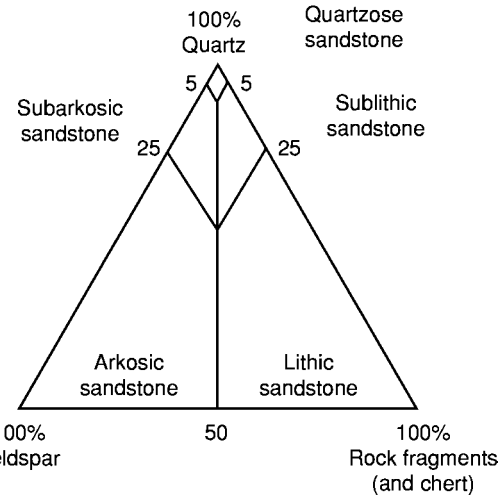


Figure 2 Sandstone classification using the QFL (quartz, feld spar, lithic fragments) system.

Table 1 Summarized mineralogy of the Upper Jurassic Brae Formation (Miller Field) and Middle Jurassic Etive Formation (Columbia Field), UK North Sea

	Brae formation	*Etive formation*
Quartz + polycrystalline quartz	86.3	74.1
Total feldspar	2.0	3.3
Mica	1.0	1.9
Other detrital minerals	1.6	1.1
Matrix clay	0.9	1.2
Organic matter	1.1	0.4
Calcite cement	0.2	1.8
Siderite cement	0.0	0.0
Quartz cement	5.3	3.9
Pyrite cement	0.2	0.4
Kaolinite cement	0.1	6.1
Illite cement	1.3	5.8
Number of samples	56	18

rock fragments. Although the QFL diagram is widely used, it may not be adequate for some sandstones, in which case alternative classifications and descriptions may be employed, e.g., micaceous sandstone, glauconitic sandstone, shelly sandstone, and tuffaceous sandstone. Most sandstones contain between about 10 and 20 distinct mineralogical and rock components (Table 1).

Diagenetic Mineralogy

Minerals that precipitate during diagenesis are commonly referred to as cements. A wide variety of cements have been identified in sandstones. Some are common, others are rare. The most common cements, in decreasing order of abundance, are quartz (Figure 3), carbonates, zeolites, clays, and evaporite minerals (Figure 4). Less common cements include barite, celestite, opal, amorphous silica, albite, haematite, pyrite and other sulphides, apatite, and many more. A systematic study of more than 100 case histories of diagenesis from a range of sandstones worldwide has revealed several recurring patterns, in addition to demonstrating the relative abundance of the five mineral groups (quartz to evaporites) listed above (Figure 5). There appear to be five common styles of diagenesis that can be seen in sandstones of different ages from across the globe. The mineral associations that form these styles are as follows.

- Quartz-dominated diagenesis with lesser quantities of clay minerals and carbonate minerals that precipitated after the quartz.
- Clay mineral-dominated diagenesis with lesser quantities of carbonate minerals and quartz or zeolite that precipitated after the quartz.

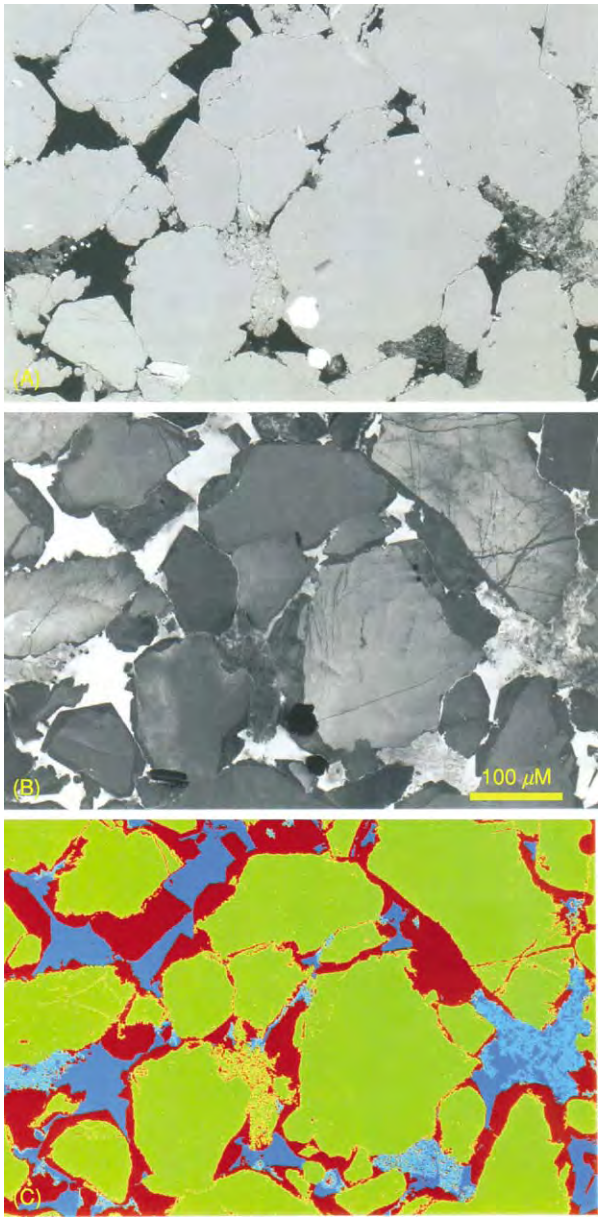


Figure 3 Quartz cemented quartzose sandstone, Miller Field, North Sea. (A) Backscattered scanning electron microscopy (BSEM) photomicrograph. Minerals with highest mean atomic number appear white and those with lowest mean atomic number appear black. Pore space is filled with a low mean atomic num ber resin, so appearing black. (B) Scanning electron microscopy (SEM) cathodoluminescence (CL) image of the same field of view as in (A). There are a few impurities and lattice defects in the syntaxial quartz cement and fracture fills, and so these areas appear darker than the detrital grains. (C) Combined BSEM and CL images with false colour added. Green, quartz grains; red, quartz cement; blue, pore space. The areas of pale blue are resin impregnated kaolinite plates and partially dissolved feld spar grains. From Gluyas JG, Garland CR, Oxtoby NH, and Hogg AJC (2000) Quartz cement; the Miller's tale. In: Worden RH and Morad S (eds.) *Special Publication of the International Association of Sedimentologists 29*, pp. 199 218. Oxford: International Associ ation of Sedimentologists.

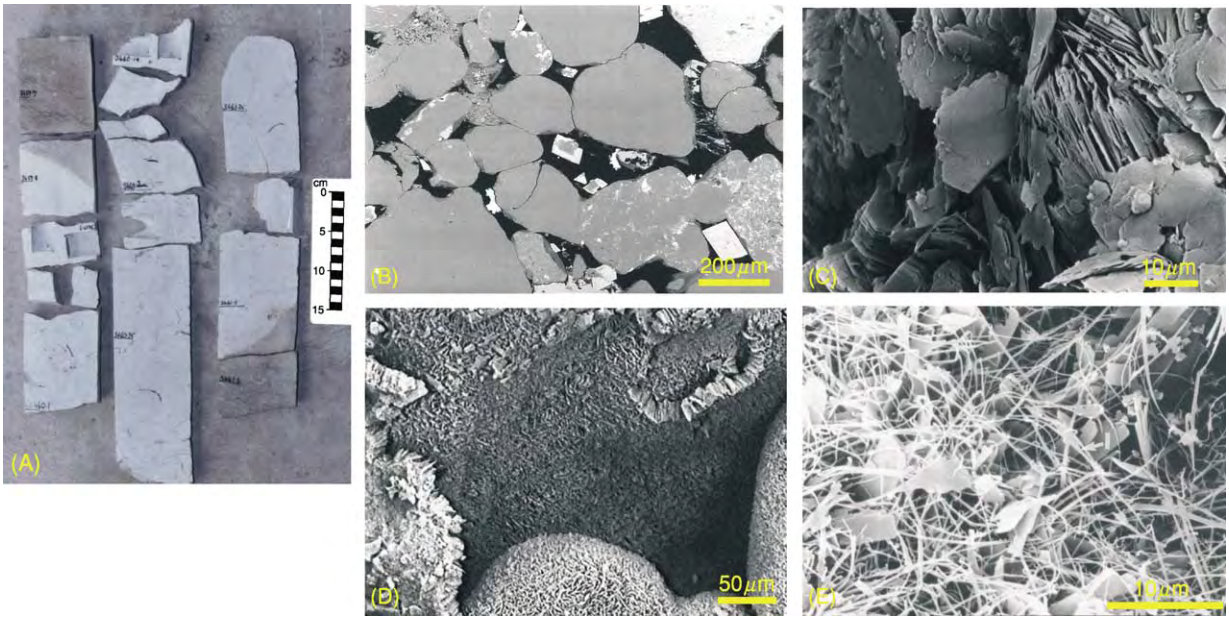


Figure 4 Common mineral cements in sandstones. (A) Spherical calcite concretion in core, Upper Jurassic, Ula Formation, North Sea; scale, 15 cm. (B) Rhombs of dolomite cement (with ferroan dolomite rims), Lower Permian Rotliegend Sandstone, North Sea; backscattered scanning electron microscopy (BSEM) photomicrograph. (C) Pseudohexagonal plates of kaolinite, Upper Jurassic, Magnus Member, North Sea; scanning electron microscopy (SEM) photomicrograph. (D) Grain coating chlorite cement, Cretaceous, Tuscaloosa Sandstone, Louisiana, USA; SEM photomicrograph. (E) Pore bridging illite cement, Triassic Skagerrak Formation, North Sea; SEM photomicrograph. (C) (E) Secondary electron microscope images. Photographs reproduced courtesy of BP.

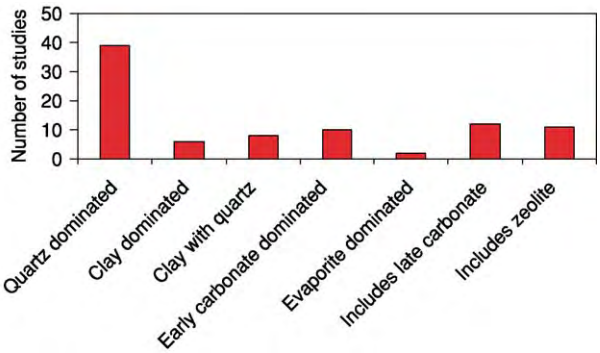


Figure 5 Styles of diagenesis summarized from a worldwide survey. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

- Grain coating clay precipitated soon after deposition and wholly or partially inhibiting subsequent precipitation of quartz and carbonates.
- Carbonate cements precipitated soon after deposition.
- Zeolites precipitated with clays, followed by carbonates and opal or quartz.

The reasons why such associations are common are investigated in the following sections.

Diagenetic Sequence

From the associations listed above, it is clear that diagenesis has a chronology. Observations made under the microscope (optical microscopy, scanning electron microscopy (SEM), transmission electron microscopy (TEM), backscattered scanning electron microscopy (BSEM), cathodoluminescence (CL), **Analytical Methods:** Geochemical Analysis (Including X-Ray)) allow mineral precipitation (and dissolution) events to be arranged in a temporal sequence. It is also possible to include the relative timing of compaction within such sequences. An example of the mineral precipitation sequence for the Middle Jurassic Brent Sandstone from the North Sea is shown in Figure 6. The sequence of diagenetic events was deduced from observations made using thin sections and SEM. Although such diagrams are useful in conveying the sequence of events, they often accidentally convey two other impressions, neither of which is likely to be true. In the absence of quantitative data on when and where cements precipitated, it is common to display the high-abundance cements as having taken the longest to precipitate. This is probably an error. It is also common for the sequence of events to fill up all the available time from the deposition of the sand to the present day. This is certainly an error. A similar diagenetic sequence diagram is

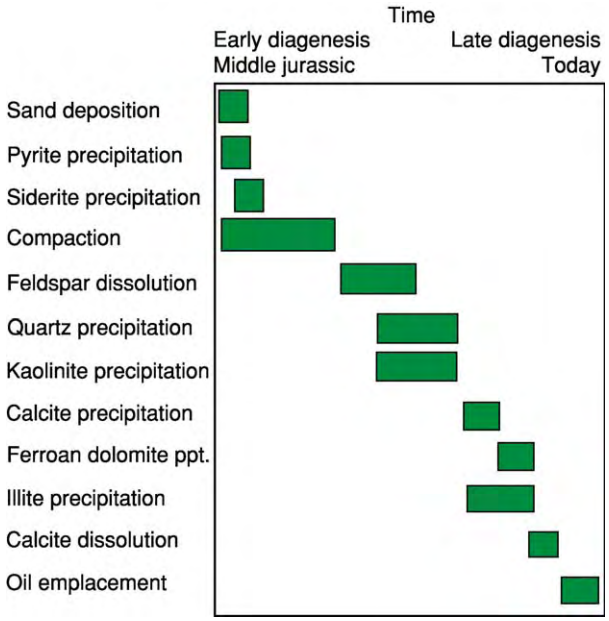


Figure 6 Diagenetic sequence deduced from thin section and scanning electron microscopy (SEM) analysis, Brent Group, North Sea. Adapted from Eglington G, Curtis CD, McKenzie DP, and Murchison DG (1985) Geochemistry of buried sediments. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences* 315.

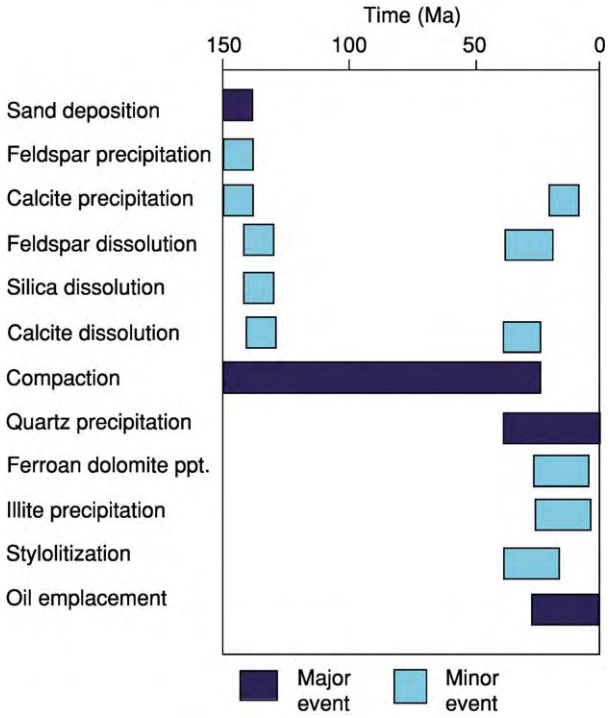


Figure 7 Diagenetic sequence calibrated to absolute time using geochemical and isotopic analysis in addition to conventional petrographical observations, Ula Formation, North Sea. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

shown in Figure 7, although here the duration of events has been constrained using additional data from geochemical, isotopic, and petrographical analyses.

There are many descriptive terms used to refine the qualitative description of diagenesis. Thus, it is possible to read of 'early carbonate', 'mesogenic quartz', 'burial cements', 'late ferroan dolomite', and so on. It is all too easy to be confused by the plethora of terms, particularly when some are contractions of long, although better, descriptive terms. For example, 'early carbonate' is often used to describe calcite or dolomite that precipitated soon after the deposition of a sand, before significant compaction, and whilst the sand was still in contact with surface or near-surface formation water.

Mineral Dissolution

Dissolution of either grains or cements in a sandstone leads to the development of secondary porosity. In the 1970s and 1980s, many publications suggested that mineral dissolution was a key process whereby significant porosity could be created at depth. Such porosity could then be occupied by petroleum. A range of dissolution mechanisms were proposed to explain this. More recently, new work has indicated that none of these mechanisms is likely to be capable of generating significant secondary porosity in the deep subsurface. That is to say, secondary porosity is rarely so extensive as to significantly improve reservoir quality.

Many minerals will dissolve during deposition and subsequent diagenesis. The only requirement is that the connate (formation) water that surrounds the grains is undersaturated with respect to the mineral in question. However, proof that a particular mineral has dissolved during diagenesis is often more difficult to come by. Grains that have partially dissolved are positive proof that secondary porosity has been created, as is mouldic porosity within otherwise tight rock (Figure 8). However, so-called oversized pores are commonly cited as evidence for the complete dissolution of grains and, although such claims are sensible, proof of secondary porosity creation is lacking.

Advocates of secondary porosity often claimed the wholesale dissolution of mineral cements (particularly calcite) during deep burial, rendering once cemented, low-porosity sandstones highly porous and permeable. Popular amongst the various processes invoked for such widespread dissolution was appeal to

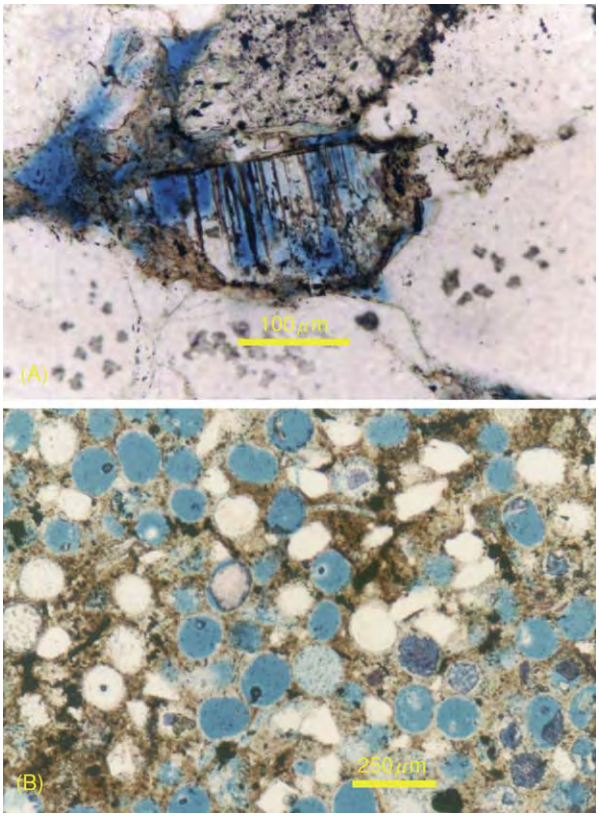


Figure 8 (A) Skeletal feldspar grain (blue, porosity), Upper Jurassic Fulmar Sandstone, North Sea; plane polarized light photomicrograph. (B) Sponge spiculite sandstone in which many of the spicules have dissolved (blue pore space) and microcrystalline quartz has precipitated in the original pore space (stained brown by oil), Jurassic Alness Spiculite, North Sea; plane polarized light photomicrograph.

organic acids created during the initial phases of oil source rock maturation. The hypothesis invoked such acids racing ahead of the migrating oil, leaching carbonates as they went. Oil then followed in the newly created porosity. In an anthropomorphic twist, this became known as the 'John the Baptist Hypothesis' – porosity created ahead of the oil coming. Although appealing and superficially elegant, there is scant evidence to support such a hypothesis. Quite apart from the difficulties of creating sufficient acid and getting it to the reservoir where secondary porosity is required, it remains difficult to find convincing evidence of large-scale, large-volume mineral dissolution in the deep subsurface. A partial exception to this rule occurs in association with unconformities. There is commonly ample evidence of porosity creation due to reaction between rock and meteoric water beneath unconformity surfaces. The improved porosity is then commonly (partially) retained during reburial of the sequence (Figure 9).

Diagenesis Quantified

The foregoing text describes diagenesis in terms of minerals that can precipitate and others which dissolve. It also investigates the relative timing of diagenetic events. However, in order to understand how diagenetic processes operate, it is important to determine when and where minerals precipitate and dissolve and the quantities of matter involved in such reactions.

Before about 1990, there were few published examples in which the absolute date of precipitation, temperature of precipitation, and isotopic composition of the host fluid were known. A wide range of analytical techniques are now available which enable quantitative or semi-quantitative data to be gathered on the 'when' and 'where' of diagenesis. The most commonly used techniques for obtaining quantitative data are listed below. A description of the methods is given in **Analytical Methods:** Geochemical Analysis (Including X-Ray).

- Fluid inclusion analysis: homogenization temperature data obtained from aqueous inclusions within diagenetic minerals can be used to estimate trapping temperatures and hence the precipitation temperature of the minerals (Figure 10) (*see* **Fluid Inclusions**).
- Radiogenic dating: potassium–argon, argon–argon, and several other methods can be used to give absolute precipitation dates for a few diagenetic minerals, such as illite (clay) and feldspar (Figure 11).
- Stable isotope analysis: $\delta^{13}C$, $\delta^{18}O$, $\delta^{24}S$, and δD (deuterium) are in common usage for helping to identify the source of elements, together with the temperature and composition of formation water during the precipitation of carbonates, sulphates, clay minerals, quartz, and sulphides (Figures 12 and 13).

The techniques outlined above allow some attempt to be made at quantifying when, at what temperature (and depth), and from what formation water a particular mineral precipitated. None of the methods addresses directly how much material moved and over what scale such movement took place during diagenetic processes. There have been many papers questioning whether sandstone diagenesis occurs in open or closed systems. There is no consensus. It is, however, important to try to answer the question because it has a direct bearing on the prediction of porosity (and permeability) ahead of drilling petroleum exploration wells.

One method which has been used to address the movement of matter during diagenesis is whole rock geochemistry. The basic premise of the method is to compare portions of the sandstone that have and have

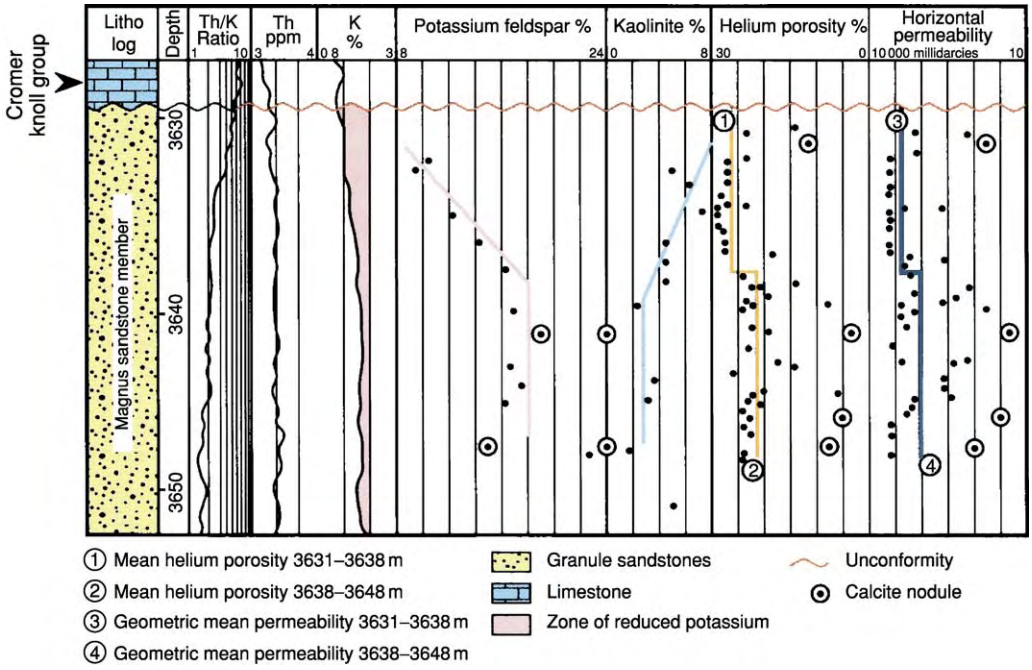


Figure 9 Secondary porosity creation beneath an unconformity during exposure and weathering, Upper Jurassic Magnus Member Sandstone, North Sea. Reproduced courtesy of BP.

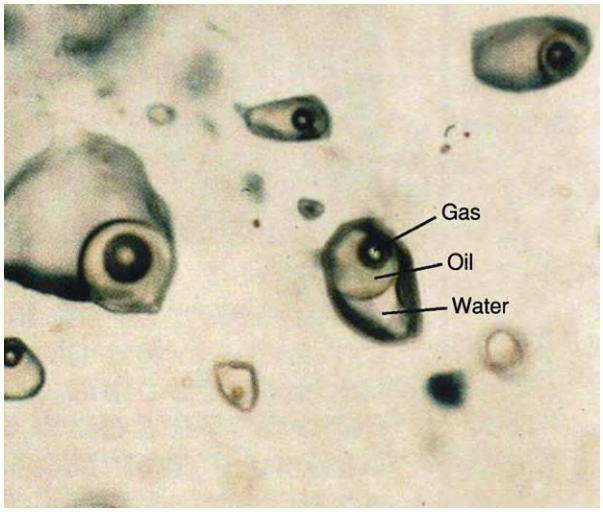


Figure 10 Fluid inclusion within mineral cement. On heating, the fluid phases within the inclusion homogenize. For aqueous inclusions, the homogenization temperature commonly equates to the minimum trapping temperature. Freezing the same inclu sions yields a measure of the salinity of the trapped fluid. For those inclusions containing petroleum, ultraviolet fluorescence analysis can give a measure of the maturity of the oil. Moreover, if abundant, oil can be extracted from the inclusions and analysed. Reproduced courtesy of BP.

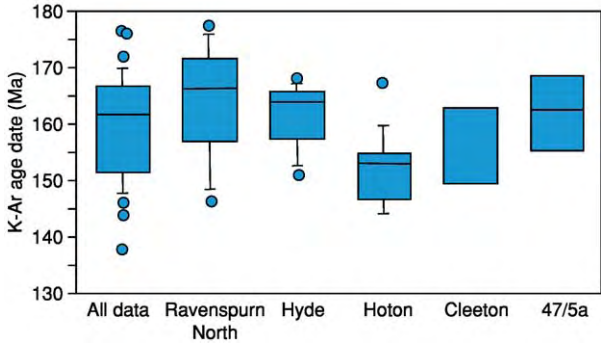


Figure 11 Potassium argon (K Ar) age distributions for illite extracted from the Permian Rotliegend Sandstone within the North Sea Gas Fields. The box plots show modes, 10th, 25th, 75th, and 90th percentiles. Modified from Emery D and Robinson AG (1993) *Inorganic Geochemistry: Applications to Petroleum Geology*. Oxford: Blackwell Science.

not undergone diagenesis. There are several possibilities for sand which has been unaffected by deep diagenesis. Part of the formation may never have been significantly buried or, alternatively, part of the formation may have been preserved from the effects of diagenesis. The Garn Formation from the Middle Jurassic of the Norwegian Sea area provides a good example, where little buried sandstone from the Draugen Field at 1.6 km can be compared with deeply buried sandstone in the Smørbukk Field at >4 km. The deep samples are relatively enriched in silica and depleted in potash compared with the shallow samples when normalized to TiO_2 content which is essentially immobile. The same results have been obtained when comparing sand trapped within calcite

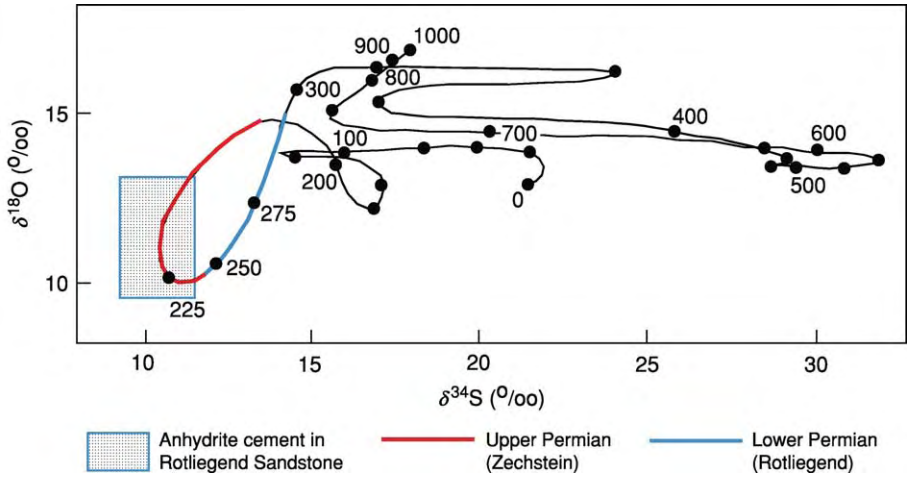


Figure 12 Global secular curve for sulphur and oxygen isotope covariance in marine derived sulphate minerals. The Permian interval is highlighted, as is the distribution of data obtained from anhydrite and barite cements in the Rotliegend Sandstones of the Amethyst Field (North Sea). The sulphur isotope data clearly indicate derivation from the overlying Zechstein. Modified from Gluyas JG, Jolley EJ, and Primmer TP (1997) Element mobility during diagenesis: sulphate cementation of Rotliegend sandstones, Southern North Sea. *Marine and Petroleum Geology* 14: 1001 1012.

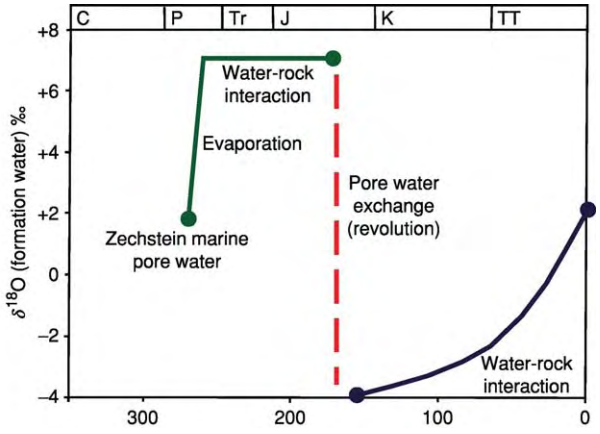


Figure 13 Pore water evolution for the Permian Rotliegend Sandstone of the Village Fields Area (North Sea) deduced from analysis of stable isotope ratios, fluid inclusion homogenization temperatures, and radiometric dating in and of mineral cements.

concretions, which precipitated soon after deposition, with surrounding sandstones. It is tempting to deduce from such studies that (some) sandstones import silica and export potash during diagenesis. Critics of such studies point to the data obtained from formation water isotopic analysis, which have been used to suggest that the water budget is severely limited in the deep subsurface and there is insufficient water to transport the observed cement volumes to the site of precipitation. Others researchers invoke local sources of silica from pressure dissolution along stylolite seams, although this too is not a panacea, as many sandstones are without such pressure dissolution phenomena.

Diagenesis and Petroleum Emplacement

A possible relationship between diagenesis and petroleum emplacement has already been touched upon in the section on 'Mineral Dissolution'. Here, the likelihood of significant porosity creation by organic acids was questioned. Much more controversial is the effect on diagenesis caused by oil emplacement. There are two extreme viewpoints: oil emplacement halts diagenesis by displacing the formation water, and diagenesis continues unaffected by oil emplacement. It is probable that the truth lies somewhere between these two extremes.

Ample evidence exists of continued diagenesis in the presence of (possibly) low oil saturations. Oil-filled fluid inclusions occur in many mineral cements (Figure 10). However, quantitative analysis of these same inclusion distributions often indicates that the presence of petroleum inhibits mineral precipitation. Studies on several sandstones, including those from the Upper Jurassic of the North Sea, have shown that cementation and petroleum migration commonly occur at the same time. In some papers, this has been referred to by the acronym SMAC (synchronous migration and cementation) and in others as the 'Race for Space'. Oilfields so affected have highly porous sandstone at their crest and low-porosity sandstone at the oil–water contact. The rate of porosity decline as a function of depth is perhaps twice that of the regional porosity gradient determined from water-bearing sandstones. In the instance of the North Sea sandstones mentioned above, the regional gradient is

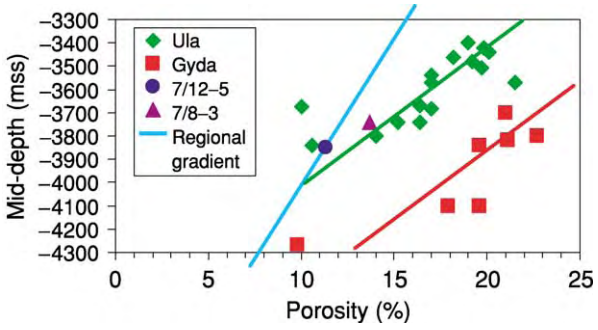


Figure 14 Porosity/depth relationships for fields within the Ula Trend (Norwegian Central Graben). Intrafield porosity gradients are about twice those observed for water bearing sandstones (regional gradient) in the same area. It is possible that oil em placement limited cementation within oil bearing reservoir inter vals. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

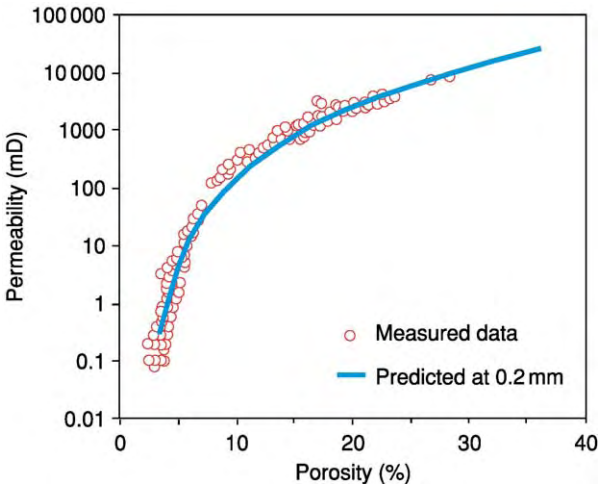


Figure 15 Measured relationship between porosity and per meability for the Fontainbleau Sandstone, and modelled relation ship for a monodisperse sphere pack with a grain size of 0.2 mm. Reproduced from Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491 501.

8% porosity loss per extra kilometre of burial depth, whilst that seen in the Ula and Gyda Fields is 16% km ^{1} (Figure 14). This same pattern of porosity loss also occurs within individual coarsening-up sequences within the reservoir interval, particularly in the direction of known mature oil source. Detailed observations on the distribution of petroleum-filled fluid inclusions indicate an exponential decline in such inclusions from field crests to field flanks, with the same sort of distribution occurring in the individual retrogradational cycles. The sympathetic patterns of porosity and fluid inclusion distribution are most easily explained by considering that diagenesis was progressively retarded as the fields filled with oil. The coarse, permeable tops of the retrogradational cycles formed the natural migration pathways of oil into the structures and these, too, had retarded diagenesis.

Impact of Diagenesis on Porosity and Permeability

From a physical perspective, sands and sandstones comprise two basic components: solid and void. In the preceding sections, the intrinsic properties of the solid component, its grain size and sorting, and its mineralogy have been examined. The void space is now examined. The void in a sand or sandstone is porosity, an intricate network of pores connected by pore throats (Figure 1). At the Earth's surface, the void space is commonly filled by a combination of water and air (depending on the elevation of the sandstone relative to the local water table). In the subsurface, the void space can, in addition to water, contain petroleum (oil and/or hydrocarbon gas) and possibly non-petroleum gas (CO_2, H_2S, N_2, and He_2).

Porosity is measured as a percentage (or fraction) of the rock plus void. For sands, porosity commonly lies in the range 35–50%. Well-sorted sands are more porous than poorly sorted sands, and loosely packed sands are more porous than tightly packed sands.

Sandstones commonly have a lower porosity than sands. This is because compaction and mineral precipitation (diagenesis) reduce the pore space between grains. In extreme instances, the porosity of sandstone can be close to 0%.

The permeability is a measure of the rate at which fluid can be transmitted through a porous medium. It's unit is the Darcy (D), such that a rock has a permeability of 1 D if a potential gradient of $1\text{atm} \times 10^{-2}\text{m}$ induces a flow rate of $10^{-6}\text{m}^3\text{s}^{-1}10^{-4}\text{m}^{-2}$ and a liquid viscosity of 1cP. For loose sands, the unit of permeability is the darcy, whereas, for sandstones, a more convenient unit is the millidarcy. There is no particular reason why porosity and permeability should be related, other than that, for a rock to have non-zero permeability, it must also have non-zero porosity. However, for individual sands and sandstones, porosity and permeability are commonly positively correlated (more porous sandstones tend to be more permeable than less porous sandstones). Where it does exist in granular porous media, the correlation between porosity and permeability commonly reflects the variation in one or possibly more of the components. For example, the Fontainbleau Sandstone of the Paris Basin is essentially monodisperse (perfectly sorted) and uncompacted. However, the quantity of cement varies between 0% and about 40%. For this

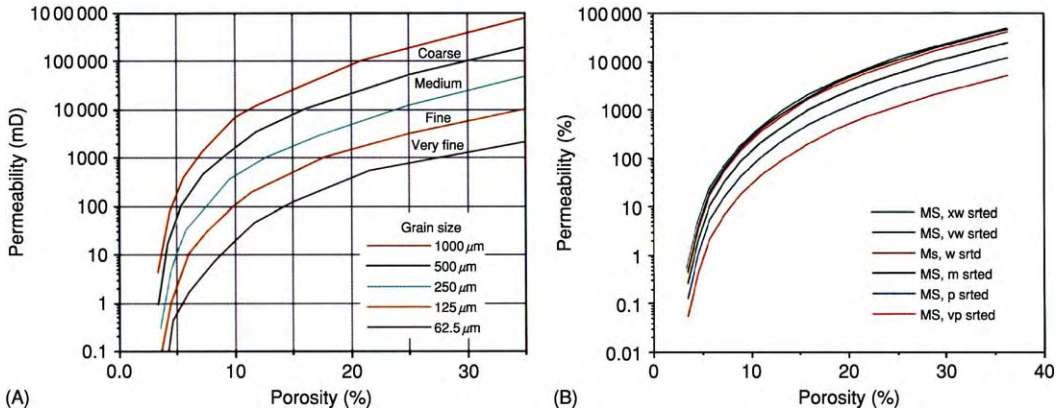


Figure 16 (A) Modelled relationship between porosity and permeability for monodisperse (perfectly sorted) sands of different grain size. (B) Modelled relationship between porosity and permeability for a medium grained sand with different sorting characteristics (xw, vw, w, m, p, and vp srted denote extra well, very well, well, medium, poorly, and very poorly sorted, respectively). Reproduced from Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491 501.

sandstone, there is a non-linear correlation between porosity and permeability on a semi-logarithmic plot (Figure 15). The increasing rate of decline in permeability at low porosity is due to progressive closure of the pore throats between pores.

The relatively simple relationship between porosity and permeability for the Fontainbleau Sandstone has been used as the foundation for a comprehensive predictive model for permeability based on a real physical model of a porous medium. The model combines data derived from the perfectly sorted porous medium with empirical curves linking porosity and permeability for less well-sorted sands (Figure 16). Cements are then modelled as grain rimming or pore filling, and solid (such as quartz or carbonate) or microporous (clays).

Controls on Diagenetic Processes

In broad terms, near-surface diagenetic processes are much better understood than those occurring at depth. Geochemical and isotopic studies have revealed the importance of bacterial reactions in modifying pore water and inducing the precipitation of carbonates, oxides of iron and manganese, and sulphides.

For the deep subsurface, we know less about what triggers diagenesis, although we can, as shown above, determine when and under what conditions diagenetic reactions occurred. A recurrent observation is that major diagenetic events commonly accompany or follow immediately after significant geological events. This is almost self-evident in the case of mineral dissolution beneath unconformities, but in other situations it is a little more subtle. For example, most of the clay and carbonate minerals in the Permian Rotliegend Sandstone of the southern North Sea precipitated towards the end of the Jurassic, a time of major rifting in the area. At the same time, there was a fundamental change in the connate water from Zechstein (evaporated seawater) derived to meteoric, yet saline, water. There was also a dramatic loss of overpressure from the reservoir system (the reservoirs are normally pressured today). It is tempting to conclude that the rifting led to failure of the salt seals above the Rotliegend, and massive pore water revolution, so causing cementation. In contrast, the Middle Jurassic Brent Sandstone over much of the northern North Sea was cemented at around the Paleocene to Eocene boundary. This, too, may have been associated with the ingress of meteoric water as the rift shoulder became elevated. Although it is possible that such external factors were the cause of cementation in these two sequences, it seems probable that the degree and style of cementation was controlled by the conditions within any particular sandstone (temperature, pressure, mineralogical composition).

See Also

Analytical Methods: Fission Track Analysis; Geochemical Analysis (Including X-Ray); Geochronological Techniques. **Diagenesis, Overview**. **Fluid Inclusions**. **Minerals:** Feldspars; Quartz. **Petroleum Geology:** The Petroleum System. **Sedimentary Rocks:** Mineralogy and Classification. **Sedimentary Processes:** Fluxes and Budgets. **Weathering**.

Further Reading

Burley SD and Worden RH (2003) *Sandstone Diagenesis Recent and Ancient, Reprints Series, International Association of Sedimentologists,* vol. 4. Oxford: Blackwell Science.

Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491 501.
Eglington G, Curtis CD, McKenzie DP, and Murchison DG (1985) Geochemistry of buried sediments. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences* 315.
Emery D and Robinson AG (1993) *Inorganic Geochemistry: Applications to Petroleum Geology.* Oxford: Blackwell Science.
Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience.* Oxford: Blackwell Science.
Gluyas JG, Garland CR, Oxtoby NH, and Hogg AJC (2000) Quartz cement; the Miller's tale. In: Worden RH and Morad S (eds.) *Special Publication of the International Association of Sedimentologists 29,* pp. 199 218. Oxford: International Association of Sedimentologists.
Gluyas JG, Jolley EJ, and Primmer TP (1997) Element mobility during diagenesis: sulphate cementation of Rotliegend sandstones, Southern North Sea. *Marine and Petroleum Geology* 14: 1001 1012.
Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69.* Tulsa: American Association of Petroleum Geologists.
MacDonald DA and Surdam RA (1984) *Clastic Diagenesis, American Association of Petroleum Geologists' Memoir 37.* Tulsa: American Association of Petroleum Geologists.
Morad S (1998) *Carbonate Cementation in Sandstones, Special Publication of the International Association of Sedimentologists 26.* Tulsa: Blackwell Science.
Selley RC (2000) *Applied Sedimentology*, chap. 8. San Diego: Academic Press.
Worden RH and Morad S (2000) *Quartz Cementation in Sandstones, Special Publication of the International Association of Sedimentologists 29.* Oxford: Blackwell Science.

SEISMIC SURVEYS

M Bacon, Petro-Canada, London, UK

Introduction

Seismic methods study the subsurface by generating seismic waves and observing the way that they propagate through the Earth. Various methods of field acquisition and data processing are used, mainly with the objective of producing cross-sections through the subsurface that can be interpreted in geologically meaningful ways. The methods are particularly widely used in the oil and gas industries.

The type of wave most often used for seismic investigation is a low-frequency sound wave. This is usually called a P wave; during its passage, individual particles oscillate backwards and forwards in the direction that the wave is travelling, so that the wave consists of alternating compressions and rarefactions. The velocity at which the wave travels depends on the rock through which it is passing, and is related to the mineral constituents, the amount and geometry of the porosity, and the type of fluid contained in the pore space. Another type of wave sometimes used is the shear (or S) wave, where the particles vibrate at right angles to the direction in which the wave travels. This type of wave cannot travel through fluids. In rocks, its velocity is affected by similar factors to those that influence P-wave velocity, except that it is relatively insensitive to the type of fluid in the pore space.

Seismic reflection is the method most commonly used. The basic idea is shown in Figure 1. Seismic P waves are generated by a source (such as a small explosive charge) at the ground surface. They travel down through the Earth, are reflected at boundaries between rock layers, and travel back to the surface, where they are detected by a receiver (similar to a microphone, but sensitive to low frequencies down to 5 Hz) and recorded. The time taken for the wave to travel from source to receiver tells us the depth of the reflecting boundary, and, by repeating the measurement at a series of points, it is possible to map the reflecting surface. The principle is similar to the way a

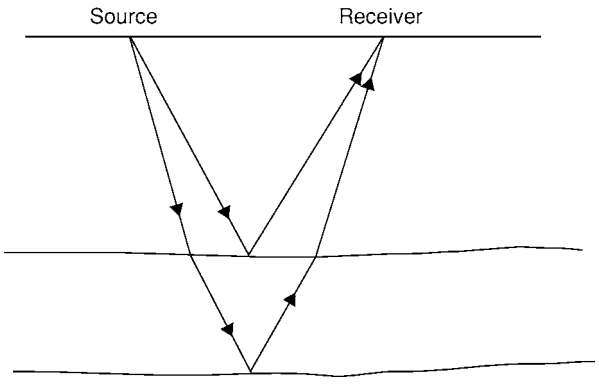


Figure 1 Principle of seismic reflection: seismic waves are generated by a surface source, are reflected at boundaries between rock layers, and are detected and recorded by a receiver at the surface.

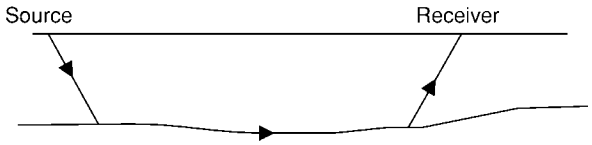


Figure 2 Principle of seismic refraction: a seismic wave travels between a surface source and a receiver along a subsurface boundary across which there is an increase in seismic velocity.

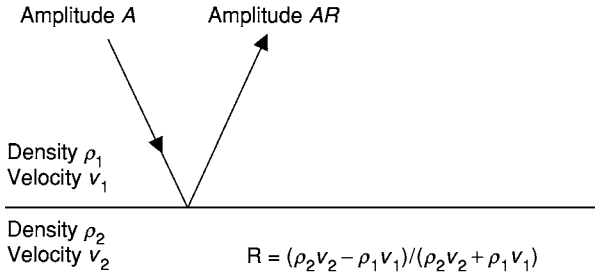


Figure 3 The amplitude of the reflection from an interface depends on the contrasts in density and seismic velocity across it.

ship's echo sounder is used to measure the distance to the seabed.

Other types of seismic wave can travel along boundaries between layers where there is an increase in wave velocity, and this is the basis of the seismic refraction method (Figure 2). A distance much larger than the depth of investigation separates the source and receiver. The travel time of the seismic signal is used to map the deep interface. This method is usually able to detect only a few such surfaces, across each of which there is a large velocity increase. It is used both for large-scale crustal studies and for shallow engineering investigations. The seismic reflection method is able to detect much more detail, typically allowing us to see many tens or even hundreds of reflecting surfaces. For this reason, it has become the method of choice for the subsurface investigation of sedimentary basins, particularly for petroleum exploration.

Most seismic reflection work uses sources and receivers at or near the surface, and this is what I shall proceed to discuss in detail. It is also possible to acquire data using a surface source and receivers in a borehole. The advantage of such a vertical seismic profile (VSP) is that a particularly detailed image of the subsurface is obtained, which can be closely tied to the drilled succession.

Reflecting Interfaces

What determines how well an interface between two layers will reflect seismic waves? There is a characteristic of a material called its acoustic impedance; this is the product of the seismic velocity and the density of the material. The amplitude of the reflected signal is proportional to the contrast in acoustic impedance across the interface. The formula shown in Figure 3 applies to P waves at an incident angle of 0°, when the seismic wave travels perpendicularly to the reflecting interface; at larger angles of incidence the formula is more complicated and involves a combination of S-wave and P-wave velocities. The acoustic impedance of a rock type depends on many factors. In a sandstone, for example, it will depend on porosity, cementation, and clay content; in a shale it will depend on the degree of compaction. In practice, many interfaces between different lithologies will have enough acoustic impedance contrast to cause appreciable seismic reflection. The interface needs to be sharp rather than gradational. This requirement is not onerous, however; the transition has to take place over a distance very much less than the seismic wavelength. Since the seismic signal typically has a frequency of 30 Hz and the seismic velocity in a sand or shale is typically $3\,\mathrm{km\,s^{-1}}$, a typical seismic wavelength is about 100 m. A lithological transition over a vertical distance of a few metres will therefore be seen as sharp by the seismic wave. A further requirement is that the reflecting interface should be laterally continuous over distances similar to the seismic wavelength. In sedimentary basins, there are often many interfaces that meet these criteria and are therefore a suitable target for the seismic reflection method. It is sometimes possible to use the method to investigate the internal structures of metamorphic or igneous rocks, but results are often poor owing to a lack of suitable reflecting surfaces or to scattering of seismic energy by internal complexity.

Data Acquisition and Processing

The simple geometry shown in Figure 1 is not usually an adequate approach to acquiring seismic data. The reflections are weak and easily swamped by noise. Increasing the power of the source will help. Modern seismic sources include airguns (which are used at sea and work by releasing a bubble of compressed air into the water) and vibrator trucks (which are used on land and vibrate a metal pad held in contact with the ground). Cost, practicality, and concern about possible environmental damage place limits on the energy that can be put into the ground. The solution is to use an array of receivers to make the best use of the available energy. Figure 4 shows how, in the marine case, a long array of receivers is towed behind a ship, which fires the source at regular intervals along a line. After acquisition is complete, the

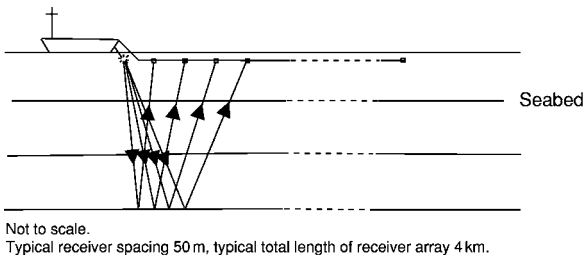


Figure 4 Schematic geometry for acquiring marine seismic reflection data.

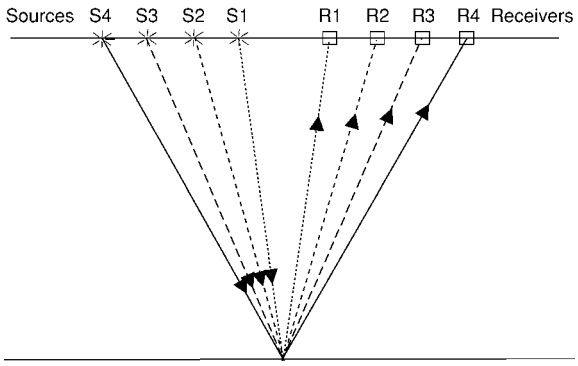


Figure 5 Acquisition geometry of traces sharing a common reflection point.

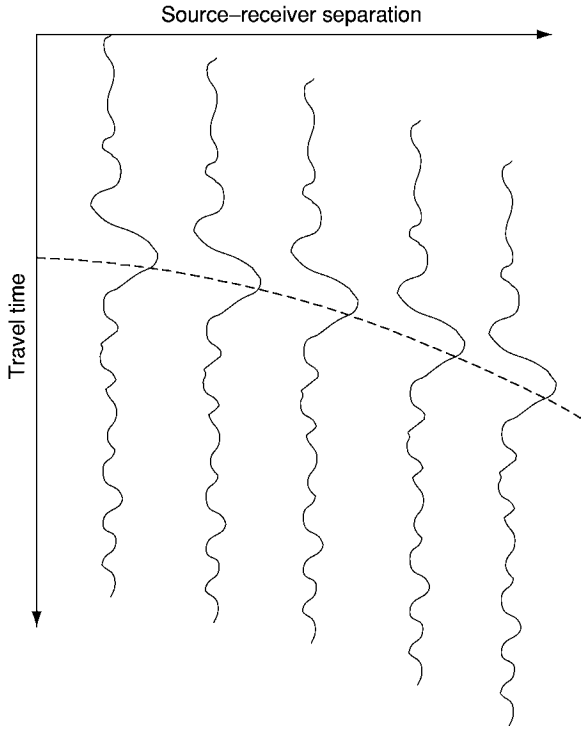


Figure 6 Schematic plot of traces acquired with the geometry shown in Figure 5.

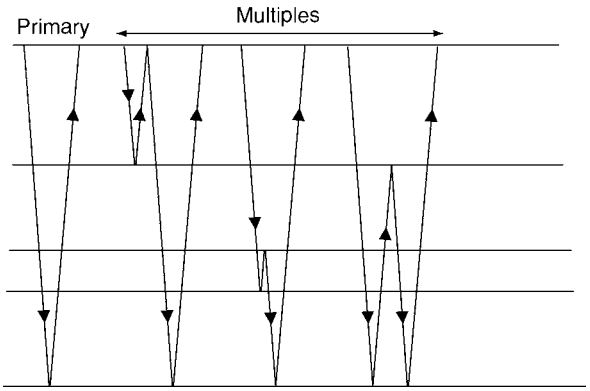


Figure 7 Travel paths for primaries and multiples.

recorded data from different shots can be reordered to bring together traces corresponding to a single reflection point in the subsurface (Figure 5). Of course, the travel time increases as the source–receiver separation becomes larger (Figure 6), but this can be corrected so as to line up all the signal peaks at the same travel time. They can then be added together (stacked) to create a signal with a much higher amplitude. The correction required to align all the traces contains information about the average velocity of the seismic waves, which is useful in later processing.

However, the medium above a target reflector is usually strongly layered. This means that signals can bounce back and forth between these shallow layers, and may perhaps arrive at the receiver at much the same time as a genuine reflection from a deeper layer (Figure 7). There are several ways to remove these 'multiples'. Many of these methods depend on the difference in average velocity along the travel path between the primary and the multiples, caused by the general increase of velocity with depth due to compaction. The multiples have spent more time at shallow depths, so their average velocity is lower than that of the deeper-penetrating primaries. Correction for variable source–receiver distances will thus line up the primaries but not the multiples, which will be attenuated in the stacked result. Various algorithms exploit this velocity difference to improve the discrimination against multiples further. The same process of multiple bounces, on a smaller layer-thickness scale, acts to blur the crisp initial seismic signal on its passage through the Earth. This combines with the effect of the absorption of seismic energy (which is more pronounced over a given distance for the higher frequencies) to reduce the content of high-frequency energy in deep reflections. Commonly, reflections from a depth of 3000 m will have peak energy at a frequency of 25–30 Hz. As we shall see, this reduces the resolution that can be achieved.

Seismic data are often acquired along a straight line, with the objective of producing a cross-section showing the subsurface reflectors along the line (2D seismic). At first sight, this just requires each stacked trace to be plotted in the correct place along the line; the wiggles corresponding to each reflector will then line up to produce a cross-section through the Earth, though the vertical axis will be travel time rather than depth. There is, however, a complication, which is illustrated in Figure 8. This shows the travel paths of the seismic signals for various source and receiver positions along a line over a schematic buried syncline. Since the data have been corrected for variable source–receiver distances and stacked, we can assume that they are equivalent to the data that would have been recorded with zero source–receiver separation at each point along the line. In that case, the seismic travel paths must hit the reflector at right angles, so that the reflected path is the same as the incoming path. We see from Figure 8A that at some surface locations it is possible to receive reflections from both sides and from the bottom of the syncline, so that the stacked section will present a 'bow-tie' appearance (Figure 8B). For other reflector geometries, the distortion would be less dramatic, but would still be present; it arises wherever the travel paths are not vertical, but we ignore this and plot the reflection traces vertically below the relevant surface point. To correct for the distortion, the reflector segment seen on each trace needs to be moved laterally by the correct amount. This process is known as migration. It requires knowledge of the seismic velocities in the subsurface, which have already been obtained for use in the variable source–receiver distances correction prior to stacking. The effect of migration in transforming the image into a recognizable picture of subsurface structure can be dramatic.

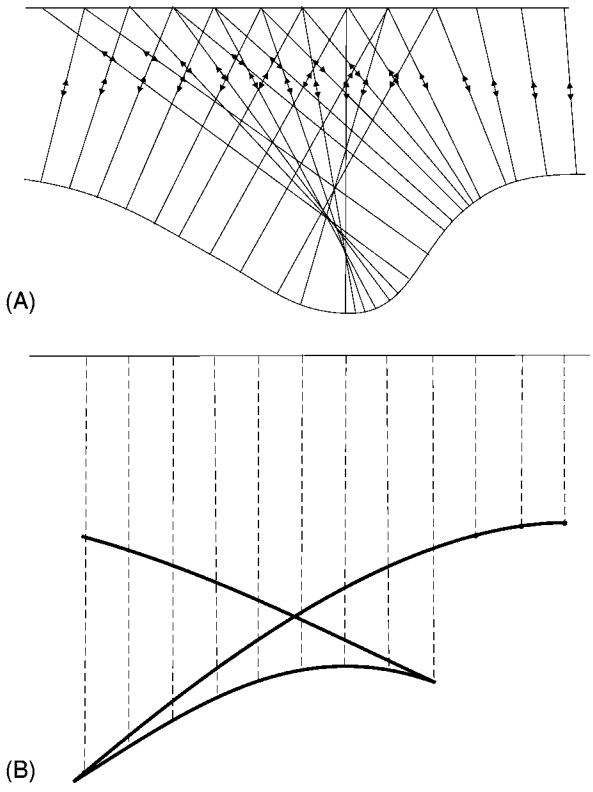


Figure 8 A (B) travel time section can be more complicated than (A) the real depth section. In (B) time is plotted vertically below the surface point concerned.

3D Seismic

Migration of seismic data along a 2D line does not perfectly position the reflectors in the right place, however. The problem is that, if the line is not exactly along the dip direction, reflection points may be laterally offset from the line (Figure 9). Standard processing has no way of detecting that this is so, and the final migrated section will be plotted vertically below the surface line. To minimize this effect, 2D lines are acquired along the dip direction where possible; however, the dip direction may change with stratigraphic level. A big improvement in subsurface imaging can be gained by the use of 3D seismic. Suppose we acquire a large number of parallel 2D lines, at close spacing (perhaps 50 m). Then information about the structure off to the side of each line is available from other lines in the dataset (except for the lines at the edge of the survey). We can reposition (migrate) the data in 3D, so that, when we plot a vertical section along one of the 2D lines, it contains only information about reflectors that are vertically below it.

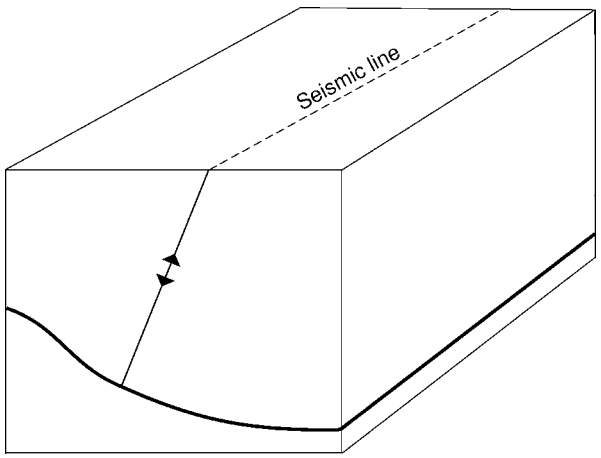


Figure 9 Reflection points for a seismic line may not be verti cally below it.

Such 3D surveys are routinely being undertaken, primarily for petroleum exploration, which can support the high cost of acquisition. The output from the processing of such a survey will be a cube of data, made up of traces plotted vertically below points on a square grid. At a typical trace spacing of 25 m × 25 m, a surface area of 200 km^2 would contain 320 000 traces. There are two further benefits from such a survey, besides the 3D imaging.

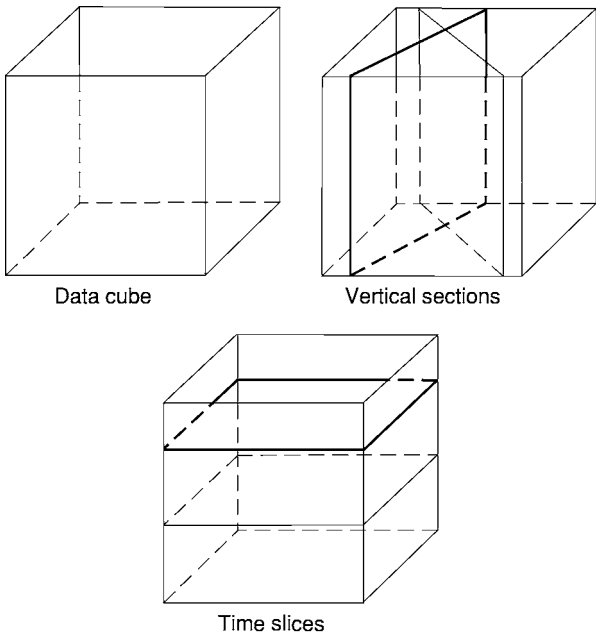


Figure 10 A 3D data cube can be viewed as slices in any direction.

1. The density of the data makes it easy to follow features from line to line across the cube. For example, the development of a fault can be studied across the cube as its throw grows and diminishes, or the detailed geometry of a sedimentary channel can be mapped.
2. Given sufficient computational power, it is easy to construct slices through the data cube in any direction (Figure 10). Vertical sections can be chosen in any direction; for example, sections perpendicular and parallel to a given fault may be helpful in understanding its geometry. The plan view (time slice) can be particularly helpful in understanding depositional systems. It is also possible to view a depositional body in 3D (Figure 11).

These displays extracted from the cube contain a great deal of information. There are, however, two limitations that need to be borne constantly in mind. One is the limited vertical and horizontal resolution. The vertical traces consist of seismic 'wiggles'; each reflecting interface is marked by a signal that represents the source signal, modified by its passage through the Earth and modified further by data processing. Figure 12 shows a typical response for a thin layer, representing perhaps a sand encased in shale. As the layer thins, there comes a point where the reflections from the top and base of the layer start to coalesce. Beyond this point, the layer is thinner than the separation of the apparent top and base wiggles would suggest. The amplitude response is a maximum when the bed thickness is one-quarter of

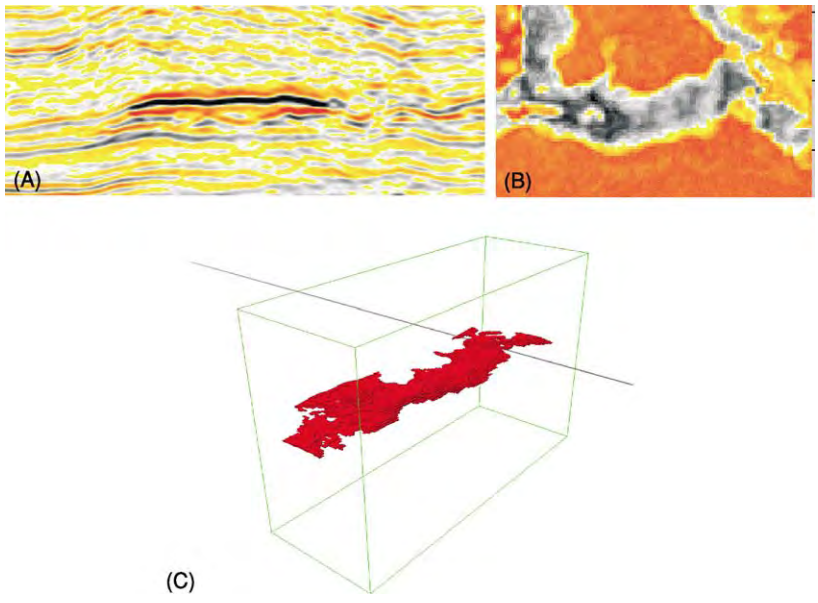


Figure 11 Displays created from a 3D data cube: (A) vertical section, (B) map view and (C) 3D perspective view.

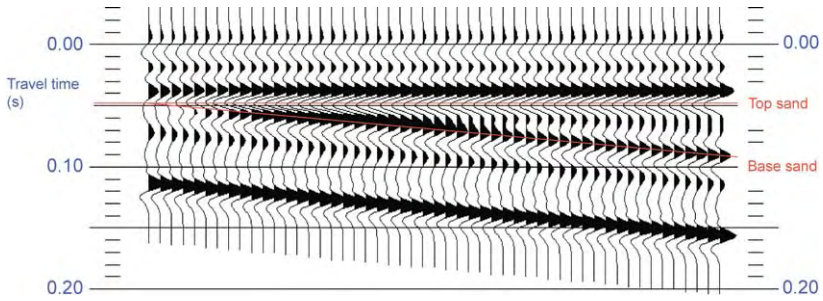


Figure 12 Modelled traces showing the seismic response of a sand of varying thickness.

the seismic wavelength: if a typical wavelength is about 100 m, the amplitude maximum is at a thickness of 25 m. This will be the approximate limit of vertical resolution. Various processes can be applied to the trace data to try to sharpen up the wiggles, but at depths of a few thousand metres it is hard to achieve a better resolution than about 12 m. Horizontal resolution is also limited. The resolution achievable depends on the accuracy with which seismic velocities are known: errors in the velocities degrade the focusing of the migrated seismic image. In practice, the resolution might be 50–100 m at a depth of a few thousand metres.

The second limitation is that the vertical axis of the traces represents travel time, not distance. If we know the seismic velocities, we can of course convert the travel times into depths. However, if there is no well control the velocities may be fairly uncertain. The resulting errors may not be important for mapping on a basin scale, but are often critical in the detailed work of oil and gas prospecting. Even when no detailed depth conversion is intended, it is essential to have a rough idea of the depth scale corresponding to the travel-time scale whenever a seismic section is being interpreted. This is because displays with considerable vertical exaggeration are often used: true-scale seismic sections are usually much wider than they are high, leading to display problems on the typical workstation computer screen with an aspect ratio near to one. If unrecognized, this distortion will hinder the understanding of depositional and tectonic features.

Interpretation

Seismic reflection allows us to see and map layering within the subsurface. We usually need to put some stratigraphic label on the mapped interfaces. Sometimes, distinctive interfaces such as a major angular unconformity are easy to recognize (Figure 13). If some boreholes have been drilled and wireline logs have been run in them to record seismic velocity and density, we can calculate the acoustic impedance of each layer and hence the expected seismic response. We also know the travel time from the surface to the reflecting interfaces, either from direct observation (e.g., in a VSP) or by integration of the sonic log. The interfaces that give rise to the largest reflection amplitudes can thus be related to the sequence drilled by the well. If there are several boreholes, reflectors can be tracked from one to another to establish a consistent identification scheme. Usually, seismic reflectors are time-lines, at least on the broad scale. The overall depositional setting can be inferred from interpretation of seismic sequences. Relative sea-level fall and rise can be inferred from variation in the pattern of onlap, and this provides information about the overall depositional environment.

Figure 13 Seismic section showing a prominent angular unconformity.

Structure mapping is often quite easy provided deformation is not extreme. Reflecting interfaces can be followed through a cube of 3D seismic, or around a grid of intersecting 2D lines, and a map constructed. Seismic reflection works best for interfaces with dips of up to 30° or so. Fault planes are therefore seldom imaged directly; they are recognized from the displacement of sedimentary layering across them

(Figure 14). Steeply dipping bedding, for example in an overthrust zone or against the flank of a salt or mud diapir, will often not be imaged.

Depositional environments can often be recognized and mapped from the external geometry of a feature (the shape of its envelope) and from the geometry and character of the reflections within it. For example, within a fluvial system it may be possible to recognize channels by mapping reflection amplitude on a slice through a 3D cube, parallel to the regional dip; the channel fill often has a different acoustic impedance from the rest of the unit. By making a series of such slices, it is possible to follow the evolution of the channel system through time. If seismic resolution permits, it may be possible to see internal depositional geometry, such as the downlap geometry of a laterally accreting point bar. Discrimination between sand and shale infill may be possible: sands often have a mounded appearance due to differential compaction.

In carbonate systems, it is often possible to recognize reefs from seismic reflections. A reflection is usually obtained from the top of a reef, though it may be discontinuous if the topography is complex. The interior is usually quite transparent. Often the reef has separated different depositional environments, so there is a sharp change in the reflection character of the contemporaneous package from one side to the other.

Salt and shale diapirs are often inferred from the deformation of the layered sediments around them. The salt or shale itself is usually acoustically homogeneous and therefore appears as a transparent body on seismic sections. Imaging sedimentary layering below the overhanging top of a mushroom-shaped salt diapir is difficult, because of the complicated paths that seismic waves follow, owing to the much higher seismic velocities in the salt.

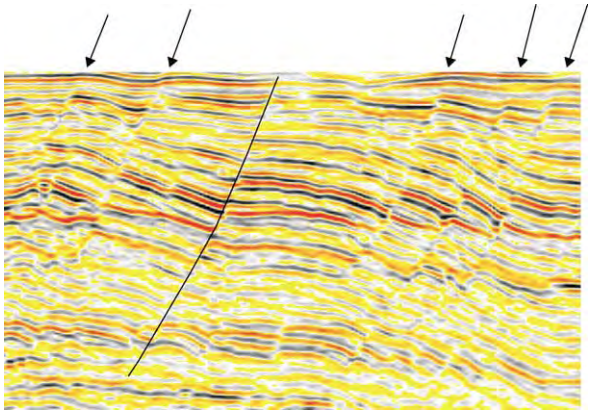


Figure 14 Seismic section showing faults, imaged as discontinuities in reflectors. One fault is shown interpreted; some others are indicated by arrows.

Seismic Reflection in the Oil and Gas Industry

The primary use of seismic reflection in the oil and gas industry is for mapping structure. In exploration **Petroleum Geology:** Exploration, this is mainly a matter of looking for the closed anticlinal features that form potential hydrocarbon traps. These may be either pure (four-way) dip closures or combination fault–dip closures. If a possible trap is fault bounded, it will be necessary to look into the juxtaposition of beds across the fault, to see whether the reservoir is always juxtaposed against a seal (e.g., a shale), or whether it is in places in contact with another reservoir (e.g., a sand). In the latter case, it may still be possible for the trap to work if the fault plane itself provides a seal, but this is inherently more risky. 3D seismic, with its high trace density, is well suited to making a juxtaposition analysis along the whole length of a fault. When a hydrocarbon discovery has been made, seismic reflection can be used to define the internal geometry of the reservoir. Small faults or thin (but laterally extensive) shales may be significant barriers to hydrocarbon flow when the reservoir is put on production, so it is important to plan development wells with this in mind.

Sometimes, it is possible to see directly from the seismic reflections whether hydrocarbons are present in a particular reservoir. When oil or gas replaces brine in the pore space of a reservoir, the acoustic impedance of the material is reduced. This can be a large effect (Figure 15), particularly for gas in high-porosity reservoirs. If the brine sand has an impedance that is less than that of the overlying shale, then the impedance of an oil or gas sand will be much less. The consequence is that the reflection amplitude from the top of the reservoir will be higher where the hydrocarbon is present. Such a 'bright spot' is

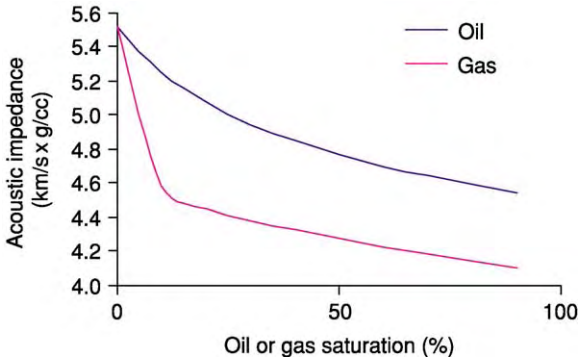


Figure 15 Effect of oil or gas on the acoustic impedance of a high porosity sandstone.

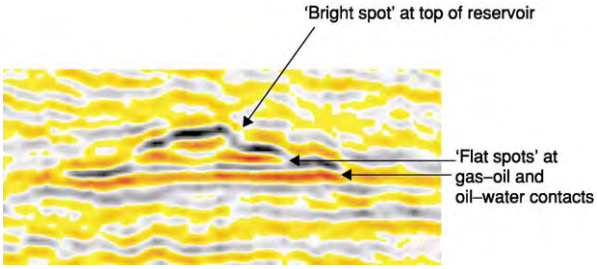


Figure 16 Examples of bright spot and flat spot direct hydro carbon indicators.

a direct hydrocarbon indicator (DHI). Of course, the brightening might be due to a lateral change in lithology, perhaps an increase in the porosity of the reservoir sand. If the brightening is a DHI, then the amplitude increase will follow a particular depth contour, at the hydrocarbon–water contact. Another type of DHI is the 'flat spot', which is a reflection from a fluid (e.g., oil–water) contact. This should be flat and consistent with the amplitude change at the top of the reservoir (Figure 16).

When these effects of fluid fill were first recognized, they were used to reduce the risk of exploration prospects. More recently, it has become possible to use this principle to follow the way that fluids are moving through an oil or gas reservoir as the hydrocarbons are being produced. This is done by acquiring a survey before production starts, and then repeating the survey after some production has taken place. The difference between the two surveys (after careful matching to remove the effects of any differences in acquisition) will be due to fluid movement, for example an oil–water contact rising because some of the oil has been produced. This is useful information if the reservoir has internal barriers to fluid movement, for example due to faulting. In such a case, oil may be produced from some fault blocks and not others; the seismic differences would be confined to the blocks where production is occurring. Such seismic surveys are almost always shot as 3D surveys, because the high trace density is important in identifying small DHI effects, and the repeat survey is often called 4D seismic (elapsed time being the fourth dimension).

Huge amounts of 3D seismic reflection data are acquired by the oil and gas industry. In the year 2000, some 300 000 km^2 were shot offshore, and 30 000 km^2 onshore. This is targeted on hydrocarbon resources, of course, and therefore confined to sedimentary basins with proven or potential hydrocarbon generation and migration. Within such basins, seismic reflection has revealed a great deal about subsurface geology.

See Also

Economic Geology. **Engineering Geology:** Seismology. **Petroleum Geology:** The Petroleum System; Exploration; Production; Reserves. **Sedimentary Processes:** Depositional Sedimentary Structures; Post-Depositional Sedimentary Structures.

Further Reading

Blake BA and Figueroa DE (1999) Interpretation strategy drives acquisition of 2 D seismic in sub Andean Bolivia. *The Leading Edge* 18: 1360 1365.

Brown AR (1999) *Interpretation of Three Dimensional Seismic Data*. Tulsa, OK, USA: Society of Exploration Geophysicists.

Gersztenkorn A, Sharp J, and Marfurt K (1999) Delineation of tectonic features offshore Trinidad using 3 D seismic coherence. *The Leading Edge* 18: 1000 1008.

Payton CE (ed.) (1977) *Seismic Stratigraphy Applications on Hydrocarbon Exploration*. Memoir 26. Tulsa, OK, USA: American Association of Petroleum Geologists.

Sheriff RE and Geldart LP (1995) *Exploration Seismology*. Cambridge: Cambridge University Press.

Story C, Peng P, Heubeck C, Sullivan C, and Lin JD (2000) Liuhua 11 1 Field, South China Sea: a shallow carbonate reservoir developed using ultrahigh resolution 3 D seis mic, inversion, and attribute based reservoir modeling. *The Leading Edge* 19: 834 844.

Telford WM, Geldart LP, and Sheriff RE (1991) *Applied Geophysics*. Cambridge: Cambridge University Press.

Wescott WA and Boucher PJ (2000) Imaging submarine channels in the western Nile Delta. *The Leading Edge* 19: 580 591.

Yilmaz O (2000) *Seismic Data Analysis*. Tulsa, OK, USA: Society of Exploration Geophysicists.

Zeng H, Tucker FH, and Wood LJ (2001) Stratal slicing of Miocene Pliocene sediments on Vermilion Block 50 Tiger Shoal Area, offshore Louisiana. *The Leading Edge* 20: 408 418.

SEQUENCE STRATIGRAPHY

P P McLaughlin Jr, Delaware Geological Society, Newark, DE, USA

Introduction: What is Sequence Stratigraphy?

Sequence stratigraphy is one of the major unifying concepts of the geosciences to arise in the twentieth century. Rooted in the cross-fertilization of regional facies mapping and geophysics, sequence stratigraphy provides an invaluable approach to practical problems in applied geology and fundamental scientific questions in Earth history. It incorporates a variety of disciplines of stratigraphic geology (i.e. lithofacies analysis, biostratigraphy, and chronostratigraphy) and is intrinsically related to a number of other areas of Earth history, notably sea-level change, tectonics, and palaeoclimate.

Sequence stratigraphy is defined as the study of rock relationships within a chronostratigraphic framework of repetitive genetically related strata bounded by surfaces of erosion or deposition or their correlative conformities. The fundamental starting point for sequence stratigraphy is the sedimentary facies, which is a lithostratigraphic body characterized by distinct lithological or fossil characteristics, generally reflecting a certain origin. A group of sedimentary facies genetically linked by common processes and environments comprises a depositional system. These depositional systems can be grouped together within a framework of unconformity-bound relatively conformable stratigraphic packages called sequences.

Early publications on sequence stratigraphy emphasized the relationships between global sea-level change, or eustasy, and large-scale stratigraphic patterns. This work provoked serious debate about the importance of global sea-level change as a genetic control on stratigraphy. Recent work integrating sequence-stratigraphic analysis with isotope data is providing new insights into the relationships between ice-sheets, climate, and sea-level, and is helping to clarify the role of eustasy in the evolution of stratigraphic successions.

However, sequence stratigraphy is more than a record of global sea-level: it is a practical stratigraphic tool. Sequences are a product of the interplay of eustasy, tectonics, and sediment supply. As a result, they can be recognized and correlated regionally, regardless of whether global sea-level change was the dominant control. With an understanding of these controls, the sequence concept provides a framework for understanding the evolution of depositional systems through time, making it a powerful predictive tool for stratigraphic analysis.

Development of the Concept

Sequence stratigraphy has seen major growth and development since the 1970s. However, the roots of the field extend back to the 1940s, when LL Sloss coined the term 'stratigraphic sequence' in his regional facies mapping of the Palaeozoic of North America. Sloss defined stratigraphic sequences as "rock stratigraphic units of higher rank than group, megagroup, or supergroup, traceable over major areas of a continent and bounded by unconformities of interregional scope". He recognized six sequences and gave them Native American names derived from localities where they are well developed: Sauk, Tippecanoe, Kaskaskia, Absaroka, Zuni, and Tejas.

In the 1960s and 1970s, the concept of the stratigraphic sequence was applied to the geophysical data collected by oil companies that were using new seismic-imaging tools to obtain a picture of basin and stratigraphic architecture. Under the leadership of former Sloss student Peter Vail, researchers at Exxon and its predecessors recognized stratigraphic patterns on seismic lines that they believed corresponded to the same types of sequences and unconformities mapped by Sloss. In addition, they identified 'onlap unconformities' within marine successions in basins on different continents and inferred global sea-level control and a worldwide extent for 'onlap cycles'.

These seismic-derived sequence-stratigraphic concepts were brought into the public domain with the 1977 publication of seminal papers by Vail and collaborators in AAPG Memoir 26. The 'depositional sequence' was defined as a stratigraphic unit composed of a relatively conformable succession of genetically related strata bounded at its top and base by unconformities or their correlative conformities. Conceptually, depositional sequences resemble Sloss cratonic sequences, but represent much shorter time intervals. The AAPG memoir also provided the first published documentation of the Exxon group's view of the relationship between inter-regional unconformities and global cycles of sea-level, including documentation of coastal onlap curves established from seismic stratigraphic records on different continental margins. Other papers detailed methods for determining sea-level change from coastal onlap and for

interpreting sequence stratigraphy and facies from seismic-reflection patterns. The second major treatise on the subject, SEPM Special Publication 42, was published in 1988. This volume elevated the significance of sequence stratigraphy as a means of understanding Earth history and as a practical tool in petroleum geology. Included articles further defined the key elements of sequence stratigraphy, documented the ages of sequences, examined theoretical aspects, and discussed both outcrop and subsurface examples.

Although these two volumes helped to bring sequence stratigraphy into the mainstream of geological thought, there was notable criticism of the concepts and cycle charts. One of the more contentious issues was the hypothesis that eustasy is the primary control on the timing and patterns of deposition of sequences. Critics argued that the accuracy and precision of chronostratigraphic control are inadequate to demonstrate synchronicity of sequences from around the world and thereby establish the uniqueness of eustatic control.

The late 1980s and 1990s saw the evolution of sequence stratigraphy into a tool for investigating increasingly detailed stratigraphic problems. Studies examined factors controlling sedimentation in specific basins or time intervals and dealt with increasingly finer-scale stratigraphic problems in a variety of depositional environments beyond the marine continental margins. This approach can be used for finer-scale reservoir- and aquifer-scale stratigraphy problems integrating well log, core, and outcrop-based datasets.

Parasequences: The Building Blocks of Sequences

The parasequence is the fundamental building block of a sequence. A parasequence is a relatively conformable succession of genetically related strata bounded by marine flooding surfaces. It is normally a progradational or aggradational package that reflects a shoaling-upwards trend. The succession of facies within a parasequence generally follows Walther's Law, which states that a normal vertical facies succession mirrors the lateral distribution of facies in a sedimentary environment.

The parasequence boundary is the key to correlation in a sequence-stratigraphic framework. It is an approximately planar marine flooding surface commonly characterized by non-deposition or minor erosion. It may be marked by significant burrowing by organisms and may have an associated lag deposit of coarse material such as shells, gravel, authigenic minerals, or rip-up clasts formed by erosion and winnowing in the course of the flooding event.

In shallow-marine successions, parasequences typically coarsen upwards with an increase in sand content and a general increase in the thicknesses of the sand beds. Sedimentary facies trace a regular succession of shallower-water sedimentary environments. For example, in a river-dominated deltaic environment, facies could reflect shoaling from prodelta to delta front to stream-mouth bar (**Figure 1A**); for a wave-dominated shoreline, a parallel succession from offshore to lower-shoreface to upper-shoreface environments might be expected (**Figure 1B**).

In some cases, fining-upwards parasequences can be recognized. For example, in marginal-marine settings, the base of the parasequence may be marked by the abrupt appearance of marine sand above marginal-marine muds, above which the percentage of sand decreases and the sand beds become thinner. The facies trace a succession of shallower-water environments, in this example shoaling from subtidal to intertidal to supratidal non-marine facies (**Figure 1C**).

Parasequence-Stacking Patterns

Just as a normal succession of genetically related beds make up a parasequence, so a normal succession of parasequences can be grouped into a unit called a parasequence set. The pattern of changes between successive parasequences in a parasequence set is termed the parasequence-stacking pattern. The concept of accommodation is fundamental to understanding parasequence-stacking patterns. Accommodation is the space available for potential sediment to accumulate and is a function of eustasy and subsidence (**Figure 2**). Sediment influx controls the rate at which this space is filled. The interplay between accommodation and sedimentation rates controls whether the shoreline advances or retreats and the resulting vertical facies changes.

Three types of parasequence-stacking pattern are progradational, retrogradational, and aggradational. A progradational parasequence set is recognized where parasequence stacking reflects overall shoaling and basinwards advance of a depositional system. Progradation occurs when the rate of deposition exceeds the rate of accommodation: the lack of vertical space for sediment accumulation forces sedimentation basinwards (**Figure 3A**). A retrogradational parasequence set reflects the opposite case, in which parasequences are stacked in a pattern that reflects overall deepening and a landwards retreat of the depositional system. Retrogradation reflects a sedimentation rate that is lower than the rate of accommodation: the inability of sedimentation to fill the available vertical space shifts sedimentation landwards (**Figure 3B**). An aggradational stacking

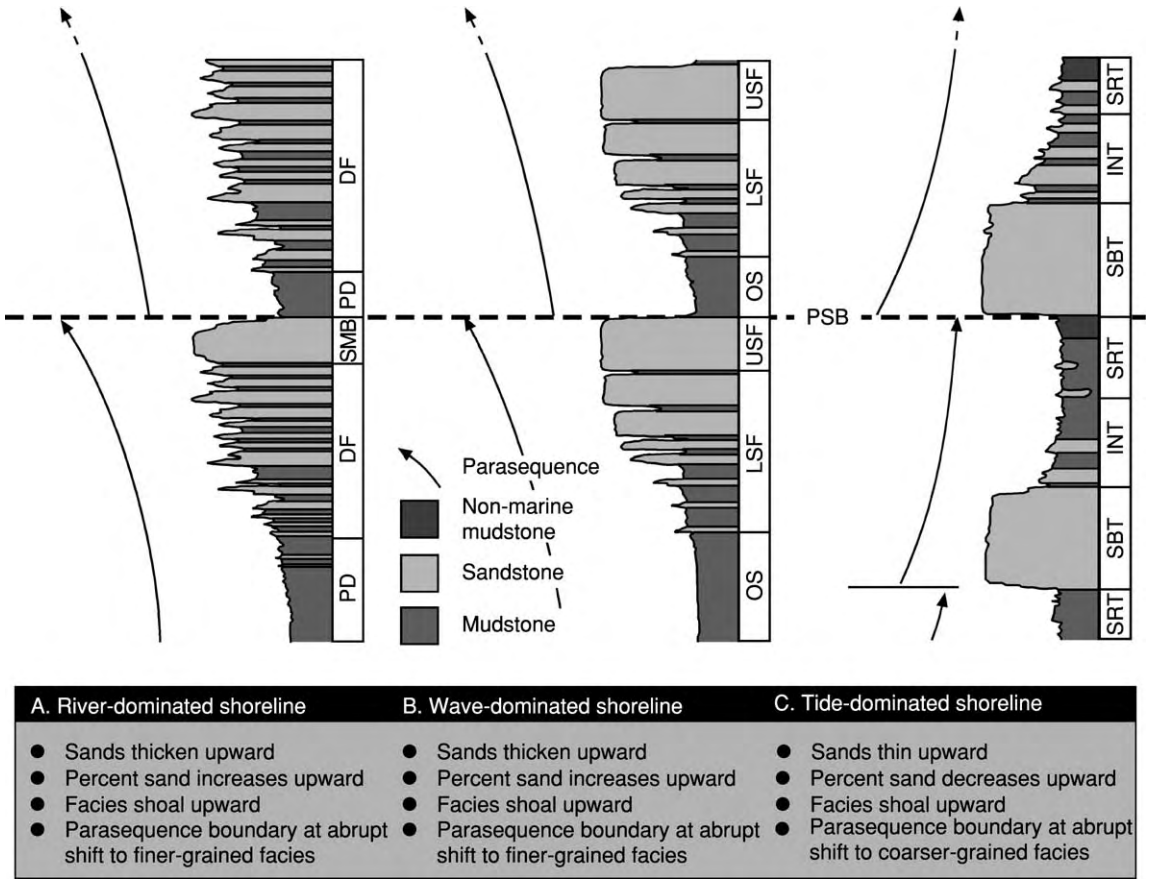


Figure 1 Characteristics of parasequences in various coastal environments: (A) deltaic parasequence on river dominated shore line; (B) offshore to shoreface parasequence on wave dominated shoreline; and (C) subtidal to intertidal parasequence on muddy tide dominated shoreline. PSB, parasequence boundary; SMB, stream mouth bar; DF, delta front; PD, prodelta; USF, upper shoreface; LSF, lower shoreface; OS, offshore; SRT, supratidal; INT, intertidal; SBT, subtidal. (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1 55.)

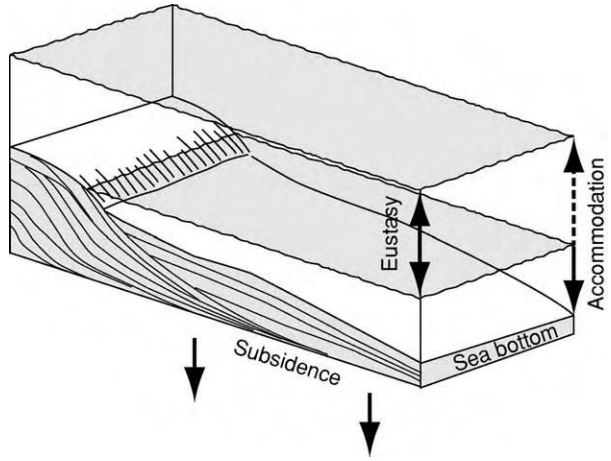


Figure 2 Accommodation as a function of eustasy (sea level changes) and subsidence (tectonics). Horizontal grey surfaces represent positions of sea level. Dashed arrow indicates the additional accommodation at the higher sea level.

pattern is recognized where the facies reflect steady accumulation of sediments without significant shifts basinwards or landwards. Aggradation occurs where the sedimentation rate and accommodation rate are approximately in balance (Figure 3C).

Parasequences are useful tools for chronostratigraphic correlation. Correlations based on sequence stratigraphy may differ significantly from lithostratigraphic correlations. Conventional lithostratigraphic correlations typically emphasize the linkage of similar lithologies, with the underlying philosophy of tracing mappable rock stratigraphic units; in the lithostratigraphic-correlation example in Figure 4A, the sand intervals in each well are correlated. In contrast, correlation of parasequences places the sedimentary section in a chronostratigraphic reference frame. Parasequence boundaries are correlated as local time-lines and provide a basis for connecting genetically linked strata. In the example of chronostratigraphic correlation shown in Figure 4B,

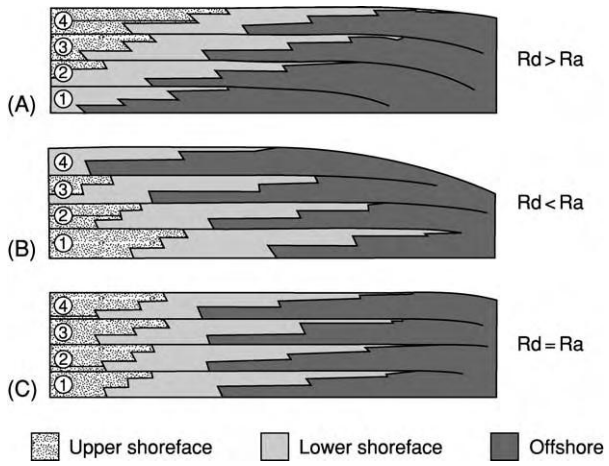


Figure 3 Parasequence stacking patterns for (A) progradational, (B) retrogradational, and (C) aggradational parasequence sets. Stacking patterns are indicated by the movement of facies in successive parasequences (numbered) in each diagram. Rd, rate of deposition; Ra, rate of accommodation. (Adapted from Van Wagoner JC, Posamentier HW, Mitchum RM Jr, *et al.* (1988) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hasting BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach*, pp. 39 45. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

the resulting stratigraphic interpretation provides an improved understanding of stratigraphic geometry by demonstrating the connection of the thinner shallow-marine sands to thicker sand beds in a landwards direction.

Recognition of Sequences and Systems Tracts

A sequence is a succession of genetically related relatively conformable strata bounded by unconformities or their correlative conformities. Based on parasequence-stacking patterns and facies trends, a number of distinct sequence components called systems tracts can be defined. A systems tract is a linkage of contemporaneous depositional systems, with a depositional system defined as a three-dimensional assemblage of lithofacies. The five systems tracts most commonly recognized are lowstand fan, lowstand wedge, shelf-margin wedge, transgressive, and highstand (Table 1 and Figure 5). These systems tracts are separated by significant stratigraphic surfaces, the most important being the sequence boundary, transgressive surface, and maximum flooding surface.

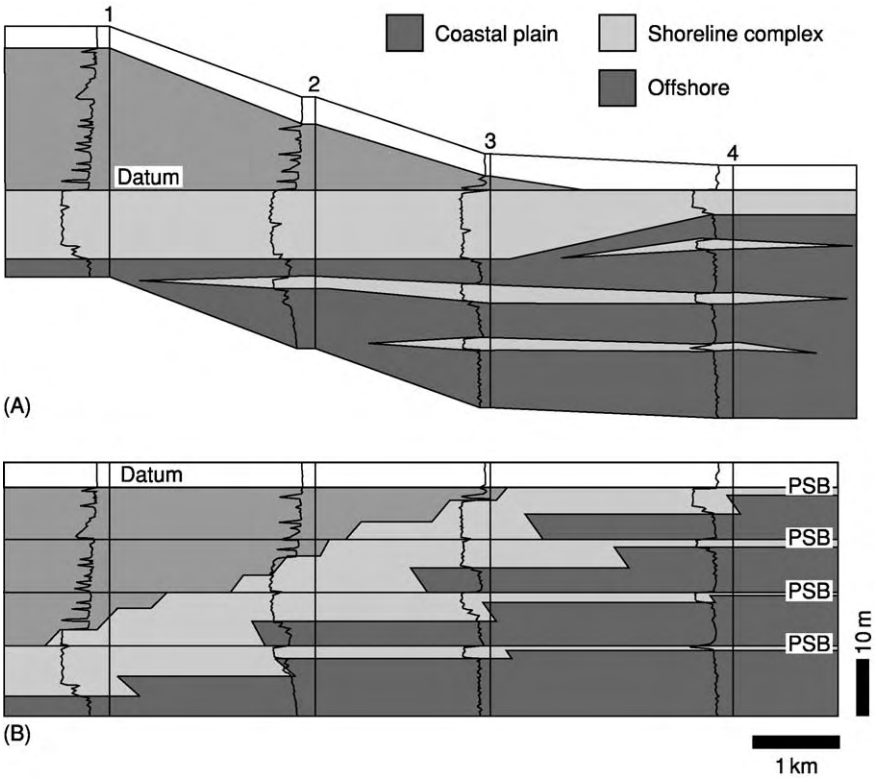


Figure 4 Comparison of (A) lithostratigraphic and (B) chronostratigraphic correlation styles for a prograding parasequence set. The correlation datum for the lithostratigraphic section is the top of the sandstone; the datum for the chronostratigraphic section is the uppermost parasequence boundary. PSB, parasequence boundary. (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1 55.)

Table 1 Characteristics of systems tracts

Systems tract	*Stacking pattern*	*Bounding surfaces*	*Stratal terminations*	*Location of best development*
Highstand	Aggradational to progradational	Base: maximum flooding surface Top: sequence boundary	Downlap basinwards, toplap or truncation at top landwards	Landwards of the offlap break
Transgressive	Retrogradational	Base: transgressive surface Top: maximum flooding surface	Downlap basinwards, onlap landwards	Landwards of the offlap break
Shelf margin wedge	Weakly progradational	Base: sequence boundary Top: transgressive surface	Onlap landwards, downlap basinwards	Near the offlap break
Lowstand wedge	Progradational to aggradational	Base: sequence boundary or downlap surface at top of underlying lowstand fan Top: transgressive surface	Onlap landwards, downlap basinwards	Basinwards of offlap break where one exists; or in incised valleys and basinward end of ramp
Lowstand fan	Aggradational	Base: sequence boundary Top: flooding surface at top of fan	Onlap landwards, or bidirectional downlap	Deep water basinwards of offlap break

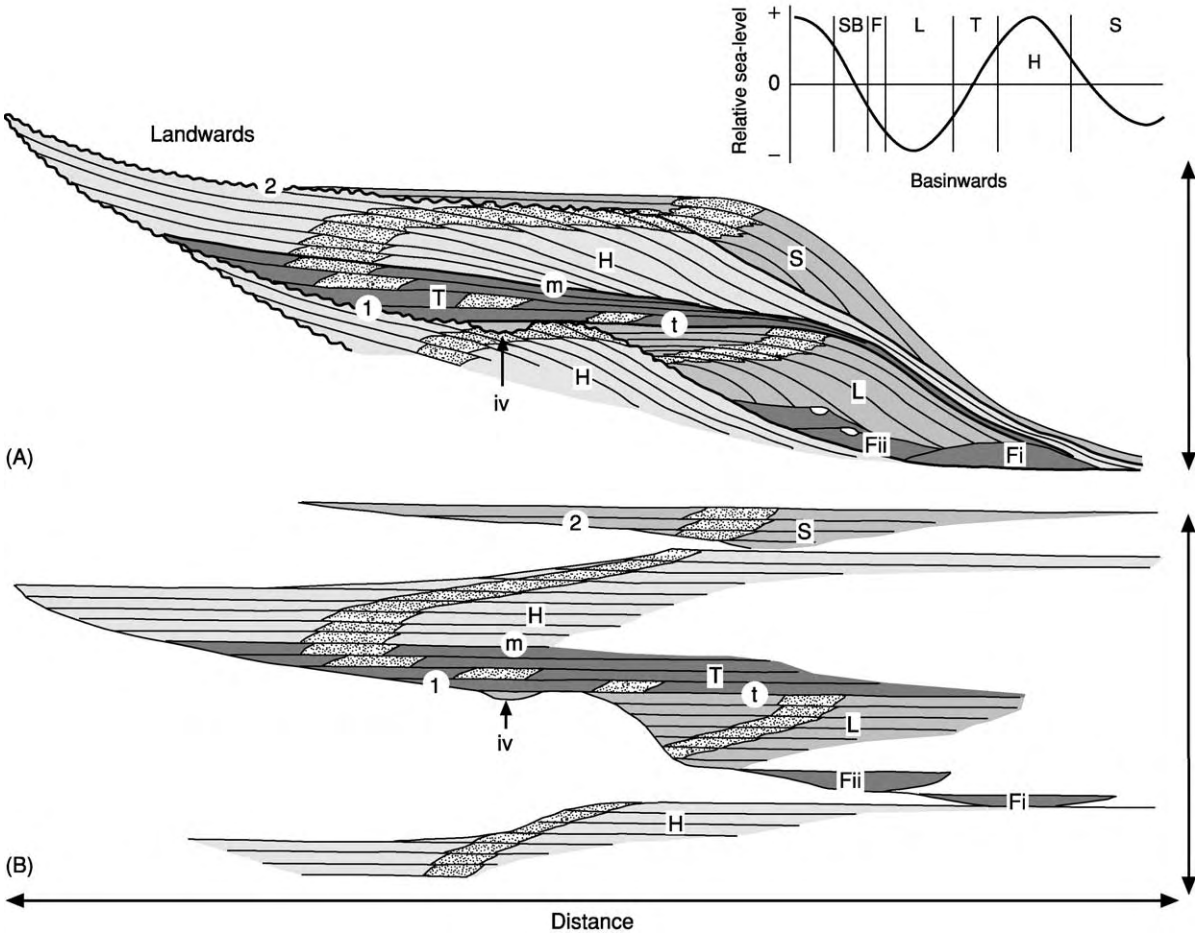


Figure 5 (A) Stratigraphic cross section and (B) chronostratigraphic section through a conceptual clastic sequence. Systems tracts, in white boxes: Fi, lowstand basin floor fan; Fii, lowstand slope fan; L, lowstand wedge; T, transgressive; H, highstand; S, shelf margin wedge. Surfaces, in white circles: 1, type 1 sequence boundary; t, transgressive surface; m, maximum flooding surface; 2, type 2 sequence boundary. Other features: iv, incised valley. Stippled pattern represents sandy shoreline complex. Relative sea level curve indicates period of deposition for each systems tract. (Adapted from Christie Blick N and Driscoll NW (1995) Sequence stratigraphy. *Annual Review of Earth and Planetary Sciences* 23: 451 478.)

Descriptive Terminology

Most of the descriptive terms for larger-scale features are derived from seismic stratigraphy (Figure 6). One of the more common geometries for a sequence is a wedge-shaped slug of sediments, with a thin zone of gently dipping strata on the landwards end, a thicker zone of more steeply seawards-dipping strata in the middle, and another thin zone of gently dipping strata on the basinwards end. The term topset is applied to the relatively flat zone of sediments on the proximal part of the basin margin. More steeply inclined strata called clinoforms characterize the thicker zone, and the reflection pattern is termed offlap. The relatively flat thinner zone basinwards of the clinoforms is referred to as the bottomset.

A fundamental principle of sequence stratigraphy is that seismic reflections are produced by contrasts in sonic velocity at chronostratigraphically significant stratal surfaces and unconformities; therefore, they are considered to approximate time-lines in the sedimentary record. Identifying terminations of these reflections is fundamental to the definition of systems tracts and key surfaces (Figure 6). Some reflection types terminate against an underlying surface. Onlap is defined by the termination of a reflection against a more steeply inclined underlying reflection, most commonly in a landwards direction. Downlap is interpreted where an inclined reflection terminates against a less inclined underlying reflection, for example the basinwards termination of prograding clinoforms. Other reflection types terminate against an overlying surface. Toplap is a subtle low-angle termination where a seismic reflection terminates against an overlying reflection without significant erosional truncation. Toplap may reflect the disappearance of an interval in a landwards direction due to sediment bypass or thinning of the bed to below seismic resolution. In contrast, erosional truncation is more abrupt, where a reflection exhibits an angular truncation against a younger surface. This generally signifies an erosional contact.

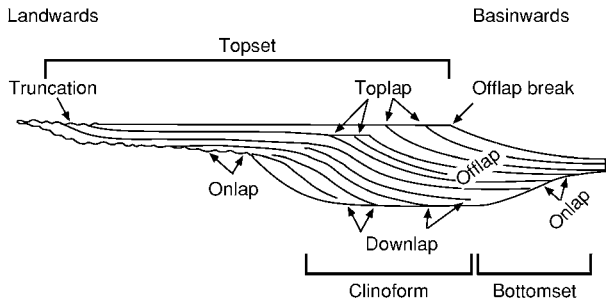


Figure 6 Seismic terminology used in sequence stratigraphic analysis. Truncation, toplap, offlap, onlap, and downlap are seismic terminations; topset, clinoform, and bottomset are zones of basin margin succession. (Adapted from Mitchum *et al.* (1977) Copyright © 1977 by The American Association of Petroleum Geologists; used by permission of AAPG whose permission is required for further use.)

An important point of reference for the description of sequences is called the offlap break. Offlap is a term sometimes used to describe clinoforms. The offlap break is the main break in slope in the depositional profile and is located at the boundary between the topset and the clinoform. In many sequence-stratigraphy publications, this is referred to as the shelf edge; however, this has created some confusion with the actual topographical break at the edge of a continental shelf, and so the term offlap break is a clearer term for this feature.

Surfaces

Sequence boundary The sequence boundary is the defining surface in sequence stratigraphy. A typical sequence boundary is an areally extensive unconformity above which there is a basinwards shift in facies, a downwards shift in coastal onlap, and onlap of underlying strata (Figure 5).

The facies shift at a sequence boundary commonly does not follow the order predicted by Walther's Law and may have a gap of an environment or two. In such cases, the sequence boundary reflects a significant basinwards shift produced by a rapid decrease in accommodation. It is commonly expressed as an unconformity produced by subaerial erosion that occurs across extensive areas both landwards and basinwards of the offlap break. Such a surface is termed a type 1 sequence boundary (Figure 7A).

In other cases, the characteristics of a sequence boundary are not as distinct, and the area affected by exposure and subaerial erosion is minimal. Although such sequence boundaries generally exhibit some onlap of underlying strata, a downwards shift in coastal onlap, and a change in facies-stacking patterns, the resulting unconformity usually has limited areal extent and the basinwards shift in facies is minor. This is termed a type 2 sequence boundary (Figure 7B). The modest reduction in accommodation permits accumulation of subsequent lowstand sediments landwards of the offlap break, in contrast to the shift of sedimentation off the shelf that occurs at the more significant type 1 sequence boundaries. In early sequence-stratigraphical publications, global sequence charts labelled individual sequence boundaries as either type 1 or type 2 and considered this to be a function of the rate of global sea-level fall; however, it is now understood that a given sequence

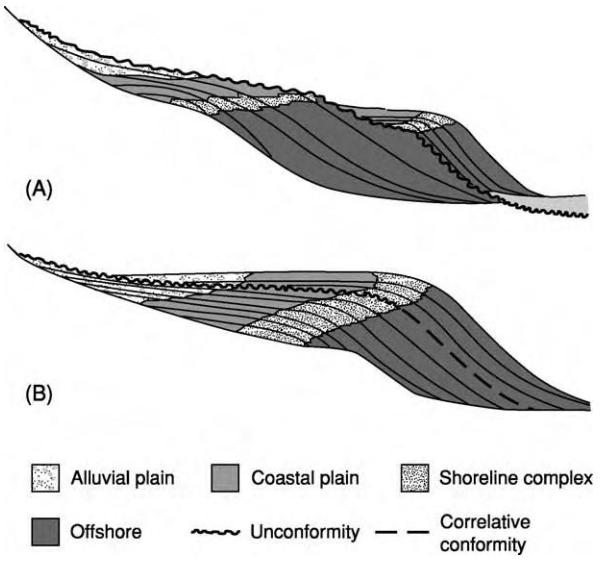


Figure 7 Physical expressions of (A) type 1 and (B) type 2 sequence boundaries. The type 1 boundary exhibits an abrupt downward shift in facies; the type 2 boundary exhibits more a gradual downward shift. (Adapted from Posamentier HW and Vail PR (1988) Eustatic controls on clastic deposition II: sequence and systems tract models. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach*, pp. 125 154. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

boundary may be of either type depending on local accommodation and sedimentation rates.

The subaerially produced unconformity that occurs at a sequence boundary passes at some basinwards point into a genetically related conformable stratigraphic contact. Some workers contend that a sequence boundary should be recognized only where an unconformity exists. However, many workers correlate this as the same chronostratigraphic surface, potentially allowing a sequence to be recognized over an entire basin.

Transgressive surface The transgressive surface is the first marine flooding surface of significant areal extent landwards of the offlap break (Figure 5). In a complete sequence, it marks a change from progradational or aggradational parasequence stacking at the bottom of the sequence to retrogradation in the middle of the sequence. It may be somewhat diachronous, with the onset of transgression appearing earlier in more basinward areas and reaching more landward areas later.

Maximum-flooding surface The maximum-flooding surface reflects the maximum landward extent of transgression and is marked by a stacking-pattern change from retrogradational in the middle part of a complete sequence to progradational in the upper part (Figure 5). On seismic profiles, it is commonly marked by the downlap of younger horizons onto it and, as such, is sometimes termed the downlap surface. The maximum-flooding surface may have an associated condensed section characterized by strongly burrowed intervals or hardgrounds, in some cases with an associated marine hiatus, as well as enrichment of authigenic minerals such as glauconite or phosphate, a high organic content, and a peak in the abundance of deeper-marine fossils. The condensed section reflects slow sedimentation rates in basinward areas when the peak of the transgression focuses sedimentation in the heads of estuaries and in other landward areas. The elevated concentration of fossils in condensed sections commonly makes them important intervals for the occurrence of age-significant fossils such as ammonites, planktonic foraminifera, and calcareous nannofossils.

Systems Tracts

Lowstand systems tracts Three systems tracts can be deposited during a lowstand of relative sea-level: lowstand fan, lowstand wedge, and shelf-margin wedge (Figure 5). These lowstand systems tracts overlie the sequence boundary and express a basinwards shift in facies produced during periods of relative sea-level fall. The lowstand fan systems tract is the most basinward of the lowstand systems tracts and forms by the accumulation of clastic deposits in a deep-basin setting. During times of relative sea-level fall, large areas of the basin margin are exposed and subjected to erosion. Sedimentation mostly bypasses the basin margin and is fed directly to the basin through incised valleys and submarine canyons. As a result, lowstand fans are commonly detached from the depositional system that built the preceding highstand complex upslope. They may onlap the underlying sequence boundary in the landwards direction; the relief built during fan formation may also produce bidirectional downlap in other directions (Table 1). Aggradational to slightly retrogradational stacking patterns are the most common. The top of the fan may be marked by a shift in deposition to the overlying lowstand wedge systems tract, which produces a downlap surface between the units. In some deep-water systems, the lowstand fan can be divided into two parts: a basin-floor fan, which occurs on the basin floor and may be detached from the depositional system that built the preceding highstand complex, and a slope fan, which develops along the middle or lower part of the slope.

The lowstand wedge systems tract is composed of a prograding wedge basinwards of the offlap break and a thinner unit of incised-valley fill in the landwards direction (Figure 5). The base of the unit is marked by onlap onto the sequence boundary along the landward end of the wedge and in areas of lowstand incised-valley fill (Table 1). In a basinwards direction, it commonly exhibits downlap onto the underlying sequence boundary or the lowstand fan systems tract. The top of the unit is defined by the transgressive surface. The facies-stacking pattern is progradational to aggradational. Because lowstand wedge deposition is generally focused basinwards of the offlap break formed by the preceding sequence, landward areas may be subaerially exposed and subject to fluvial incision, especially during sea-level fall. However, during later parts of the lowstand, the rebound of relative sea-level may result in some sediment accumulation in the incised valleys.

The shelf-margin wedge systems tract represents an accumulation of lowstand deposits near the offlap break of the preceding sequence. Like other lowstand systems tracts, it is produced when a fall in relative sea-level causes a basinwards shift in facies, but the reduction in accommodation is not rapid enough to force sedimentation into the basin. The base of this systems tract is defined by a type II sequence boundary, onto which it onlaps in a landwards direction and downlaps in a basinwards direction (Table 1). The top is marked by the transgressive surface. Facies-stacking patterns are typically weakly progradational.

Recent works have proposed several unique systems tracts for deposits produced during periods of sea-level fall. One of the more commonly cited, the forced regressive wedge systems tract, describes a complex of downstepping shorelines and subaerial erosion overlying a sequence boundary. Successive shorelines partly cannibalize sand through erosion of previous shorelines, producing stranded shoreline sand bodies. Another type, the falling-stage systems tract, is similar but differs in part in the placement of the sequence boundary at the top of the unit. However, these systems tracts are not yet consistently used by sequence stratigraphers.

Transgressive systems tract The transgressive systems tract traces a landward shift in depositional environments that reflects a rise in relative sea-level (Figure 5). The base of this unit is the transgressive surface, and the top is defined by the maximum-flooding surface (Table 1). Because sediment input is overwhelmed by accommodation, a retrogradational facies-stacking pattern is produced. The deposits of transgressive systems tracts are best developed landwards of the underlying offlap break; they onlap the merged sequence boundary–transgressive surface in a landwards direction and downlap onto the transgressive surface in a basinwards direction. They may comprise a thin sheet of deposits that reflect the landwards migration of drowned shoreface complexes; in cases where they directly overlie a sequence boundary with significant erosive relief, they may comprise more laterally variable incised-valley fill.

Highstand systems tract The highstand systems tract traces the basinwards march of deposition over the transgressive systems tract as the rate of sediment input overtakes a slowing rate of relative sea-level rise (Figure 5). It is bounded by the maximum-flooding surface below and the sequence boundary above (Table 1) and is commonly characterized by a prograding topset–clinoform system, the toes of which downlap onto the maximum-flooding surface. The top of the highstand systems tract may be marked by toplap or truncation under the overlying sequence boundary. The facies-stacking patterns change upward from aggradational to progradational, reflecting decreasing accommodation; as a result, accumulation patterns are increasingly driven basinwards rather than vertically, producing a regressive succession (Figure 5).

Variations by Depositional System

The descriptions of sequence-stratigraphic elements in this article have been principally focused on shallow-marine clastic successions. However, sequence concepts are applicable to a variety of other depositional systems. Carbonate systems are very responsive to changes in relative sea-level, but differ from clastic systems in that they generate most of their own sediment from biological sources in the photic zone. When sea-level rises, carbonate systems build upwards to fill the available space, commonly creating thick transgressive systems tracts (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves); in some cases, very rapid transgression may drown the carbonate system, resulting in a thinner transgressive interval. During highstands, carbonate sedimentation typically continues to be vigorous, but the waning increase in accommodation causes sediment accumulation to prograde basinwards, a phenomenon termed 'highstand shedding'. When sea-level subsequently falls, little carbonate sediment is produced and little is physically eroded; in wet climates, meteoric diagenesis may produce karst and/or cementation, whereas, in arid climates, evaporites may form in the basin.

Sequence concepts can also be applied to nonmarine depositional systems. In alluvial environments, the base level (the surface to which erosion

and deposition respond), subsidence (or uplift), sediment supply, and climate all affect the stratigraphic evolution of the system. Sequence interpretation is complicated by the abundance of erosion surfaces, which may be difficult to differentiate from a sequence boundary. Falling base level typically results in sediment bypass and erosion. Lowstand deposits may be characterized by high-gradient stream deposits with sand-rich amalgamated alluvial facies. Where low base level in the early lowstand is accompanied by a low water table, heavily weathered palaeosols may be formed in interfluvial areas; the rise in the water table that accompanies base-level rise in the late lowstand may be conducive to widespread peat formation. Transgressive deposits may reflect decreasing fluvial gradients, which can be expressed as more isolated fining-upwards fluvial sands and crevasse splays upward, with a shift from aggradational to retrogradational facies successions. If accompanied by a rising water table, peat-forming conditions would wane as mires are flooded, and poorly drained gleyed palaeosols would typify interfluvial areas. Highstand deposits may be thick and trace a shift back to aggradational and even progradational facies successions. Fluvial sands are mostly isolated, but amalgamation is increasingly common upwards. Slowing of the water-table rise can provide good conditions for peat formation, but the beginning of base-level (and probably water-table) fall at the end of the highstand would make conditions less favourable.

Sequence stratigraphy of lacustrine deposits has some parallels with marine sequence stratigraphy but is generally independent of changes in sea-level. Instead, tectonically driven accommodation and climate-driven variations in lake level and sediment supply can combine in many different ways to control lacustrine sequence expression (*see* **Sedimentary Environments:** Lake Processes and Deposits). Reduced precipitation reduces both lake level and sediment supply, creating lowstand deposits that are characterized by evaporites in the basin centre surrounded by an exposure surface with little erosion. However, where sediment influx is significant, the lowstand systems tract may be composed of erosive fluvial systems feeding deep-lacustrine turbidite successions. Where lake-level is raised by increased precipitation, increased sediment input may produce a transgressive systems tract of back-stepping lacustrine deltas; where relative lake-level is raised owing to basin subsidence, it may instead be characterized by low clastic input and a thin interval of fine-grained deposits. Highstand deposits may be composed of deltas and associated deep-lacustrine turbidites if high precipitation creates a large influx of clastic sediments; when precipitation is lower, low sediment influx may produce aggradational carbonate packages in shoreline areas and thin fine-grained successions in the lake basin.

Palaeontological Expression of Sequences

Palaeontology is essential to sequence-stratigraphic analysis, and sequence stratigraphy is a useful frame of reference for understanding the fossil record. Fossils provide information on two essential elements of sequence stratigraphy: environment and age. Understanding facies change is also an essential element of sequence-stratigraphic analysis. Biofacies provide palaeoenvironmental constraints that, like lithofacies, trace the facies changes that define sequence-stratigraphic surfaces and systems tracts. Transgression and regression can be readily demonstrated by palaeontological data, providing important information when lithofacies criteria are not definitive. Although coarse clastics reveal palaeoenvironmental information through sedimentary structures that reflect unique hydrodynamic regimes, environmental differences may be more difficult to discern in mudstones. Biofacies analysis can be especially useful in such fine-grained facies, providing critical criteria for differentiating mudstones from different water depths or different depositional systems (e.g. marine versus freshwater).

In addition, some sequence elements have unique biofacies signals. The maximum-flooding surface is one of the most distinctive. Because the maximum-flooding surface commonly has an associated condensed section on basin margins, it typically exhibits an especially high concentration of fossils due to slow sedimentation rates. The fauna and flora typically reflect the culmination of transgression and may indicate the maximum water depth. The condensed section is also commonly marked by a peak in diversity of common marine microfossils (such as foraminifera), the highest abundance of oceanic planktonic microfossils, and the greatest foraminiferal planktonic–benthic ratios. In some cases, a peak of low-oxygen benthic microfossils occurs in the condensed section. Maximum-flooding surfaces can be ideal locations for age control because of the abundance of fossils, particularly of more age-diagnostic oceanic types, such as planktonic foraminifera and ammonites, and of shell material suitable for isotopic analyses. The maximum-flooding surface is especially important in continental margin sequences because it commonly contains the most landward occurrences of biostratigraphically useful open-marine fossils.

The lowstand systems tract may exhibit a distinctly different biofacies signal. Reworked assemblages may be common: falling relative sea-level exposes older sediments higher on the margin and subjects them, and the microfossils in them, to erosion and

basinwards redeposition. Lowstand deposits can also have relatively high abundances of terrestrial microfossils, such as pollen, owing to the more direct transport of terrestrial material to the ocean basin.

Chronostratigraphic Aspects of Sequence Stratigraphy

Sequence stratigraphy and chronostratigraphy are intimately entwined. Sequence stratigraphy provides a framework for understanding the relationships between depositional systems in both time and space.

Sequences are chronostratigraphically significant units. Early papers on the subject considered sequence boundaries to be globally synchronous time lines corresponding to times of sea-level fall and provided detailed ties of biostratigraphic schemes to the global sequence record. Although the ages of some sequence boundaries can be established using biostratigraphy, they are often difficult to date because of a paucity of age-significant open-marine fossils in the associated regressive intervals. In contrast, maximum flooding surfaces mark the landwards incursions of open-marine environments and can be well dated because of the abundance of open-marine fossils in the associated condensed sections.

Global and regional sequence records are shown on a type of chronostratigraphic chart referred to as a cycle chart (Figure 8). Cycle charts commonly show the ages of the sequences, the magnitudes of coastal onlap, and the interpreted eustatic changes. Coastal onlap is defined as the progressive landwards onlap of coastal deposits in a depositional sequence; by definition, it excludes marine onlap such as the onlap of lowstand fan deposits. Coastal-onlap curves have their origin in the analysis of seismic-reflection profiles and trace the migration of the point of coastal onlap across a basin margin for each seismic reflection. The migration of this point reflects rises and falls in relative sea-level through time. Sequence boundaries stand out prominently on cycle charts as the horizontal lines at the jagged edge of the saw-tooth curve. The ages of the sequences are indicated by their positions on the time axis, and on some charts the sequences are also tied to biochronostratigraphic zonations. Global and regional cycle charts are derived by comparing coastal-onlap profiles from multiple basin margins. With some understanding of the tectonic history of each basin and a consistent age framework, differences in subsidence can be accounted for in creating a composite global or regional coastal-onlap curve. The first such global cycle chart, produced by Vail *et al.*, assumed that the coastal onlap curves were actually sea-level curves. These workers assumed that the saw-tooth pattern of the curve represented asymmetric rates of sea-level change, with slow sea-level rises and nearly instantaneous sea-level falls. It is now understood that the abrupt basinwards shift of coastal onlap at a sequence boundary reflects the jump of deposition from the topsets of the highstand systems tract, over the offlap break, and into the basin for the lowstand. Later versions of the cycle charts added a more symmetric eustatic curve to represent estimated global sea-level changes.

Our understanding of the chronostratigraphy of global sequences continues to improve. Integration of isotope stratigraphy and biostratigraphy from marine and onshore boreholes drilled as part of the Ocean Drilling Program has resulted in significant refinements to the dating of the sequence record of the Cenozoic and Late Cretaceous. For many Cenozoic sequences, ages of sequence boundaries can be determined at a resolution of 0.5 Ma or better.

Genesis of Sequence-Stratigraphic Units

To understand the genesis of sequence-stratigraphic units, three essential factors need to be considered: sea-level change, tectonics, and sediment supply.

Sea-Level Change

The importance of sea-level as a causal mechanism in the development of sequences is well understood, and it is important at different periodicities and scales (Table 2A). Supercontinent cycles operate at a scale of hundreds of millions of years and exercise a fundamental control on the volume of the ocean basins and hence on sea-level. These drive first-order sea-level cycles. Second-order cycles may be influenced in part by changes in rates of seafloor spreading on a scale of tens of millions of years. Faster seafloor spreading creates a greater volume of hot and more buoyant mid-ocean ridge material, decreasing the volume of the ocean basins and raising sea-level; slower seafloor spreading creates less mid-ocean ridge material, increasing the volume of the ocean basins.

At the period of third-order sequences, generally 1–10 Ma, mechanisms of sea-level change are more problematic. In times of significant continental glaciation, gross patterns of ice volume may provide a mechanism. Growth of continental glaciers decreases the volume of water in the oceans, lowering sea-level; melting of ice-sheets releases water into the oceans, raising sea-level. Continental glaciation was the major factor controlling sea-level in the Pleistocene, and its importance as early as the Miocene is widely accepted. Recent results from the study of

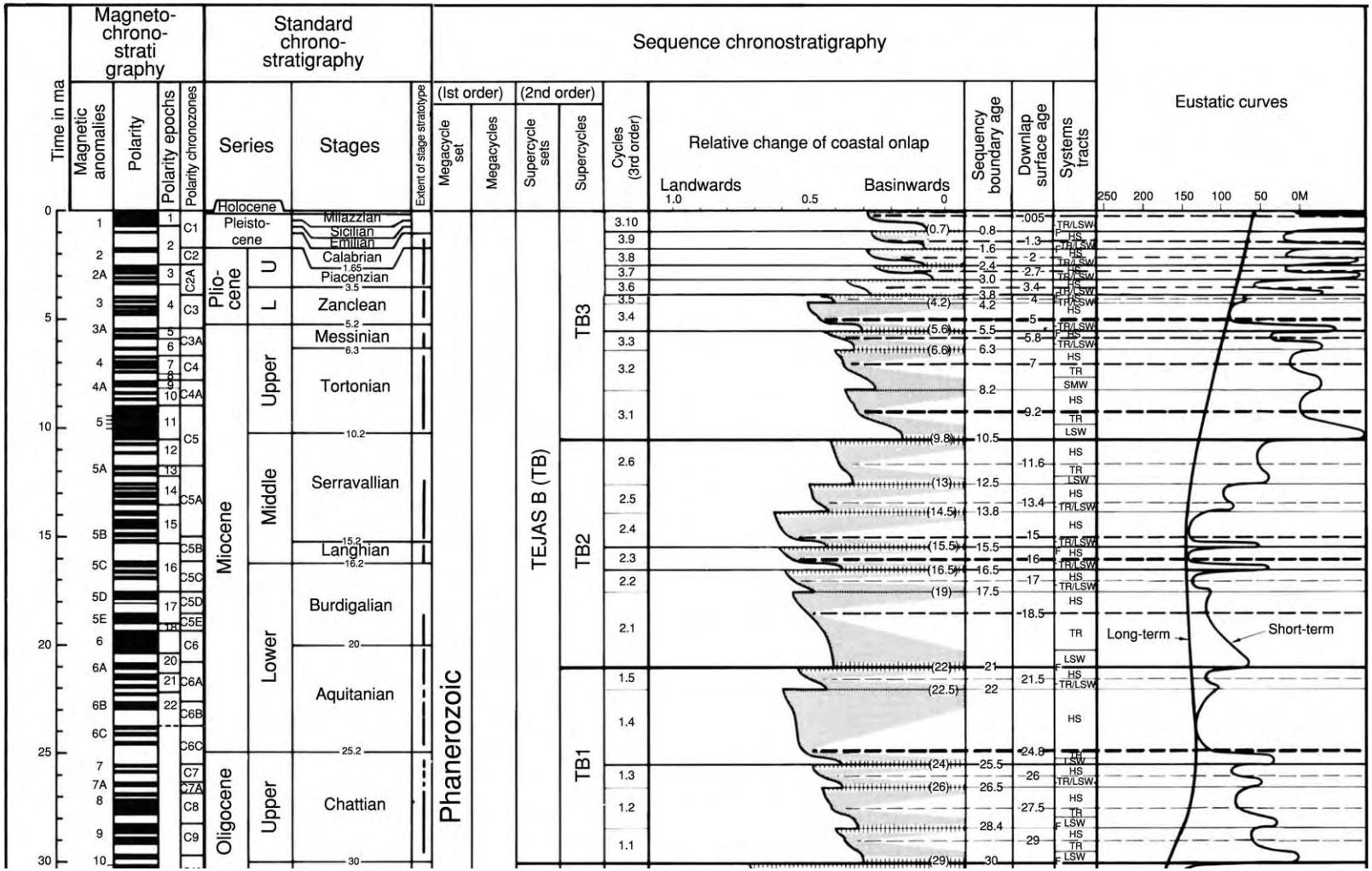


Figure 8 Cenozoic chronostratigraphic and eustatic cycle chart. Systems tracts abbreviations: HS, highstand; TR, transgressive; LSW, Lowstand wedge; SMW, shelf margin wedge. (Reproduced from Haq B, Hardenbol J, and Vail PR (1988) Mesozoic and Cenozoic chronostratigraphy and eustatic cycles. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach*, pp. 71 108. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

Table 2 Mechanisms for sea level change. (A) Orders of sea level cyclicity. (B) Characteristics of Milankovitch astronomical cycles

(A)

Order	*Duration*	*Mechanism*
First	200 400 Ma	Breakup of continents
Second	10 100 Ma	Volume of mid ocean ridges
Third	1 10 Ma	Glacioeustasy, possibly tectonics
Fourth	200 500 Ka	Astronomical forcing of glacioeustasy or climate
Fifth	20 200 Ka	Astronomical forcing of glacioeustasy or climate
Sixth	1 10 Ka	Astronomical forcing of glacioeustasy or climate

(B)

Cycle	*Duration*	*Mechanism*
Eccentricity	100 Ka and 400 Ka	Variations in degree of roundness of orbit of Earth around Sun
Obliquity	41 Ka	Variations in angle of tilt of axis of Earth relative to Sun
Precession	19 23 Ka	Variations in wobble of rotation of Earth

Ocean Drilling Program sites suggest that continental ice-sheets may have existed even earlier, with significant volumes as early as the middle Eocene and small- to moderate-sized sheets as far back as the Late Cretaceous. However, during periods of Earth history without significant glaciation, mechanisms for eustatic change are less clear. Variation in intraplate stress has been proposed as a mechanism for inducing apparent sea-level changes of as much as 100 m on the flanks of passive margins. Variation through time in the irregularities of the geoid (equipotential surface of the gravitational field) has been postulated to cause sea-level changes at different times in different parts of the globe. A more exotic mechanism invoked is an asteroid or comet impact that induces the global release of stress at plate boundaries and a resultant isostatic response of continental margins.

Fourth-order and higher cycles have periods of hundreds of thousands of years or less. Cyclic variations in the tilt and wobble of the Earth's axis, called Milankovitch cycles, cause variations in the intensity of solar radiation, which can strongly influence climate on the scale of fourth- and fifth-order cycles (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). Milankovitch cycles include three components with different periods: eccentricity, obliquity, and precession (Table 2B). Milankovitch-related climate changes affect the size of the polar ice-caps and thus global sea-level change. They can also influence monsoonal fluctuations and hence vary the amount of water delivered to and stored in lakes, aquifers, and soils over periods as short as tens of thousands of years. During the opening of the South Atlantic Ocean in the Early Cretaceous, the Parana-Benue basin is estimated to have been able to store enough water to vary sea-level by 3.46 m. It is thought that obliquity is more important at high latitudes, and precession is more important in tropical latitudes.

Small eustatic changes associated with higher-order cycles may be overprinted by larger-scale changes associated with lower-order cycles, resulting in quite different stratigraphic expressions of the same order sequence. Figure 9 shows that the effect of a small (10 m amplitude) fourth-order sea-level fall can be significant if it occurs during a period of overall limited accommodation (e.g. on the falling leg of a third-order sea-level cycle), enhancing the expression of sequence boundaries (Section I). In contrast, during a period of overall high accommodation (e.g. on the rising leg of a third-order sea-level cycle), the fourth-order fall is muted and difficult to detect in the section, while flooding surfaces are enhanced (Section II).

The early cycle charts of Vail *et al.* proposed sea-level falls for third-order sequences from as little as tens of meters to more than 300 m. Second-generation charts by Vail *et al.* indicated smaller variations, but still with some pre-Pleistocene changes of more than 100 m. More recent work arising from the Ocean Drilling Program suggests that these earlier estimates may be too high. Based on analysis of oxygen isotope records and backstripping of continental-margin sites, Cenozoic sea-level changes are estimated to be less than 100 m, approximately one-quarter the older estimates.

The relative importance of self-regulating (autocyclic) and externally forced (allocyclic) cycles in the formation of fourth- and higher-order sequences is an important consideration in sequence-stratigraphic analysis. In a clastic system, a delta lobe normally produces an upward shoaling package as it progrades

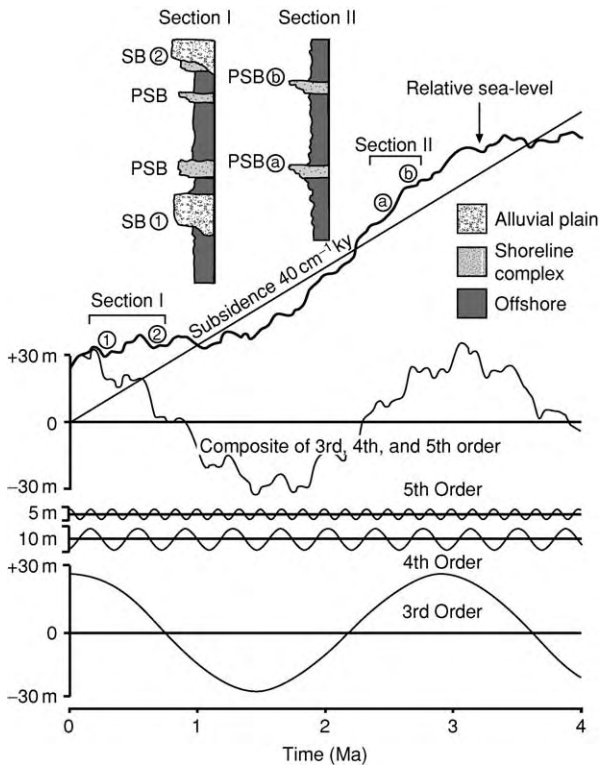


Figure 9 Interaction between subsidence and multiple orders of eustatic cyclicity, and the effect on sequence stratigraphy. Third , fourth , and fifth order eustatic changes combine with subsidence to produce a relative sea level curve, which controls the character of the sequence expression. Section I reflects de position during a period of lower accommodation related to the third order sea level fall; points 1 and 2 identify fourth order sea level falls corresponding to sequence boundaries (SB). Section II reflects deposition during a period of greater accommo dation related to the third order sea level rise; points a and b designate fourth order sea level rises corresponding to para sequence boundaries (PSB). (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1 55.)

into an area, and is capped by a flooding surface after sedimentation shifts to another lobe and the abandoned lobe sinks through subsidence and compaction. In a carbonate system, upward growth of carbonate sediments normally produces an upward shoaling pattern until all the available accommodation space is used; carbonate production is then shut off until subsidence provides space for carbonate growth to resume. These successions may occur repeatedly at scales of tens of thousands of years, similar to the time scales of higher order sea-level cycles. Thus, differentiating autocyclicity from allocyclicity in high-resolution sequence analysis may be difficult.

Tectonics

Tectonics is another major factor that contributes to accommodation; subsidence creates space for sediment accumulation, and uplift takes it away. Subsidence patterns vary according to tectonic setting. In an extensional basin, subsidence may be a response to lithospheric thinning and cooling following rifting. Subsidence due to extensional faulting is typically greatest in the early phases and decreases with time. In a foreland basin, subsidence may be a response to lithospheric flexure due to loading in the adjacent fold-thrust belt. Subsidence in foreland basins commonly accelerates over time as the load on the lithosphere increases during thrusting, followed by post-orogenic rebound. In cratonic basins, subsidence may be due to a subtler regional warping.

Tectonics is considered to be generally too slow a process to cause third- and higher-order sequences. Certainly, movement on individual faults can be significant at shorter time scales. However, overall, changes in rates of basin subsidence tend to occur on a time scale of tens of millions of years, which is a broader period than third-order sea-level variations. As a result, tectonics exerts a broader-brush influence on the nature of expression of the sequences, upon which higher-frequency sea-level events are superimposed (Figure 9).

Sediment Supply

Sediment supply controls how rapidly the accommodation space is filled, and the balance between accommodation and sediment supply controls sediment stacking patterns and thus the expression of sequences. In a setting with a high input of sediments, such as a delta, sedimentation may fill the available space in the two systems tracts with lower accommodation, the lowstand wedge and the highstand systems tract. As a result, these intervals will be characterized by aggradational and progradational packages. The sequence boundary reflects low accommodation relative to sediment input. Retrogradation of the depositional system will probably occur only at the peak of transgression, and condensed sections will be developed only in offshore locations. In settings with a low influx of sediments, accommodation may exceed the ability of the sedimentary system to fill it. In such cases, only during the part of the sequence when sea-level is falling and accommodation is lowest will sediment advance basinwards. The transgressive and highstand systems tracts will probably be characterized by sediment starvation and represented as a condensed section. Climate changes arising from Milankovitch cycles can exert a strong influence on sediment supply.

Climate variations can affect precipitation rates and vegetation patterns, which in turn influence erosion rates and sediment supply.

Conclusions

Sequence concepts provide a genetic frame of reference that has made stratigraphy a more dynamic process-based field of study compared with the static nomenclatural emphasis of traditional approaches. Sequence stratigraphy encourages an integrative approach, bringing together lithofacies analysis, geophysics, biostratigraphy, and chronostratigraphy. It provides a stratigraphic record of changes in sea-level, tectonics, and climate on local to global scales. The significant volume of jargon common in sequence-stratigraphic literature can be a barrier to understanding, and the key to wider understanding of sequence-stratigraphic concepts is to focus on the processes that create these genetically related rock units rather than on the nomenclature.

The historical association of the sequence concept with a universal explanation of global sea-level change may also hinder the acceptance of the discipline. It is important to recognize that accommodation, rather than sea-level, is the key concept in sequence stratigraphy. Because the interplay of accommodation and sediment supply controls sequence expression, the concepts of sequence-stratigraphic interpretation can be used to correlate and understand genetically related stratigraphic units regardless of whether the units are global or regional in extent.

Research drilling on coastal plains and margins will provide important new data to advance our understanding of Earth history. Continuous cores provide complete records of sequence expression, together with age data, in different geological settings. The isotopic records from such continuous-core material can provide unique insights into ice volume and palaeoceanographic changes, which shed light on the relationship between global sea-level change and sequence stratigraphy.

The nature of lithologic heterogeneities in reservoirs and aquifers and their impact on fluid flow will be better understood through the integration of sequence concepts with process-based geological models and increasingly powerful computer-based 3D-visualization technology. This detailed information allows numerical characterization of vertical and lateral heterogeneities that influence fluid flow. Sequence stratigraphy hold great growth potential as a tool for detailed stratigraphic analysis in petroleum exploration and production and in ground water resource management.

See Also

Earth: Orbital Variation (Including Milankovitch Cycles). **Palaeoclimates**. **Sedimentary Environments:** Depositional Systems and Facies; Carbonate Shorelines and Shelves; Lake Processes and Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures. **Seismic Surveys**. **Stratigraphical Principles**. **Unconformities**.

Further Reading

Bohacs K and Suter J (1997) Sequence stratigraphic distri bution of coaly rocks: fundamental controls and paralic examples. *American Association of Petroleum Geologists Bulletin* 81: 1612 1632.

Bohacs KM, Carroll AR, Neal JK, and Mankiewicz PJ (2000) Lake basin type, source potential, and hydrocar bon character: an integrated sequence stratigraphic geochemical framework. In: Gierlowski Kordesch EH and Kelts KR (eds.) *Lake Basins Through Space and Time*, pp. 3 34. Studies in Geology 46. Tulsa: American Association of Petroleum Geologists.

Christie Blick N and Driscoll NW (1995) Sequence stratig raphy. *Annual Review of Earth and Planetary Sciences* 23: 451 478.

Christie Blick N, Mountain GS, and Miller KG (1990) Seismic stratigraphic record of sea level change. In: National Research Council (ed.) *Studies in Geophysics: Sea level Change*, pp. 116 140. Washington, DC: National Academy Press.

Coe AL, Bosence DWJ, Church KD, *et al.* (2003) *The Sedimentary Record of Sea Level Change*. Cambridge: Cambridge University Press.

Emery D and Myers KJ (1996) *Sequence Stratigraphy*. Oxford: Blackwell Science.

Haq B, Hardenbol J, and Vail PR (1988) Mesozoic and Ceno zoic chronostratigraphy and eustatic cycles. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach*, pp. 71 108. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Loutit TS, Hardenbol J, Vail PR, and Baum GR (1988) Condensed sections: the key to age dating and correl ation of continental margin sequences. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach*, pp. 183 213. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Miall AD (1997) *The Geology of Stratigraphic Sequences*. Berlin: Springer Verlag.

Miall AD and Miall CE (2001) Sequence stratigraphy as a scientific enterprise: the evolution and persistence of conflicting paradigms. *Earth Science Reviews* 54: 321 348.

Miller KG (2002) The role of ODP in understanding the causes and effects of global sea level change. *JOIDES Journal* 28: 23 28.

Posamentier HW, Jervey MT, and Vail PR (1988) Eustatic controls on clastic deposition I: conceptual framework. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach,* pp. 109 124. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Posamentier HW, Allen GP, James DP, and Tesson M (1992) Forced regressions in a sequence stratigraphic frame work: concepts, examples, and exploration significance. *American Association of Petroleum Geologists Bulletin* 76: 1687 1709.

Sarg JF (1988) Carbonate sequence stratigraphy. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach,* pp. 155 181. Special Publication 42. Tulsa: Society of Economic Paleontolo gists and Mineralogists.

Shanley KW and McCabe PJ (1994) Perspectives on the sequence stratigraphy of continental strata. *American Association of Petroleum Geologists Bulletin* 78: 544 568.

Sloss LL (1963) Sequences in the cratonic interior of North America. *Geological Society of America Bulletin* 74: 93 114.

Vail P, Mitchum RM Jr, Todd RG, *et al.* (1963) Seismic stratigraphy and global changes of sea level. In: Payton CE (ed.) *Stratigraphic Interpretation of Seismic Data,* pp. 49 212. (in 11 parts). Memoir 26. Tulsa: American Association of Petroleum Geologists.

Van Wagoner JC, Posamentier HW, Mitchum RM Jr, *et al.* (1963) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea Level Changes: An Integrated Approach,* pp. 39 45. Special Publication 42. Tulsa: Society of Economic Paleontolo gists and Mineralogists.

Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1963) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1 55.

SHIELDS

K C Condie, New Mexico Tech, Socorro, NM, USA

General Features

Precambrian shields are stable parts of the continents composed of Precambrian rocks with little or no sediment cover (Figure 1). Rocks in shields range in age from 0.5 to >3.5 Ga. Metamorphic and plutonic rock types dominate, and temperature-pressure regimes recorded in rocks now exposed at the surface suggest burial depths ranging from as shallow as 5 km to as deep as 40 km or more. Shield areas, in general, exhibit very little relief and have remained tectonically stable for long periods of time. They comprise about 11% of the total crust by volume, with the largest shields occurring in Africa, Canada, and Antarctica (Figure 1). Platforms are also stable parts of the crust with little relief. They are composed of Precambrian basement similar to that exposed in shields, but overlain by 1 to 3 km of relatively undeformed sedimentary rock. Sedimentary rocks on platforms range in age from Precambrian to Cenozoic and reach thicknesses up to 5 km, as, for instance, in the Williston basin in the north-central United States. Platforms comprise most of the crust in terms of volume (35%) and most of the continental crust in terms of both area and volume. The largest platform is the Eurasian platform (Figure 1). Shields and the Precambrian basement of platforms are collectively referred to as cratons. A craton is an isostatically positive portion of the continent that is tectonically stable relative to adjacent orogens.

Seismic Characteristics

Shields and platforms have similar seismic wave velocities and layers (Figure 2). The difference in their mean thickness reflects primarily the presence of the sediment layer in the platforms with P-wave velocities $\leq 5\ \mathrm{km\ s^{-1}}$. Upper layer thicknesses range from about 10 to 25 km and each of the lower layers ranges from 16 to 30 km. The P-wave velocities in both layers are rather uniform, generally ranging from 6.0 to

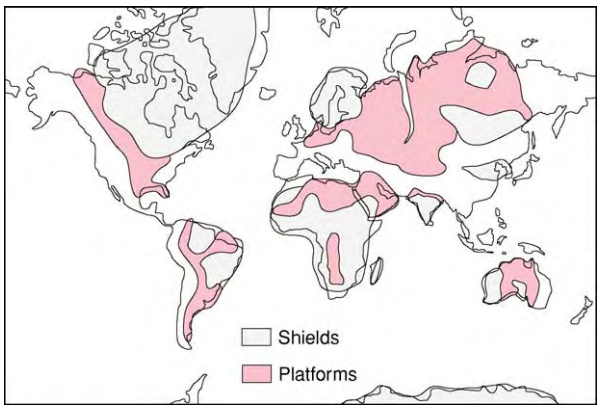


Figure 1 Map showing the distribution of Precambrian shields and platforms.

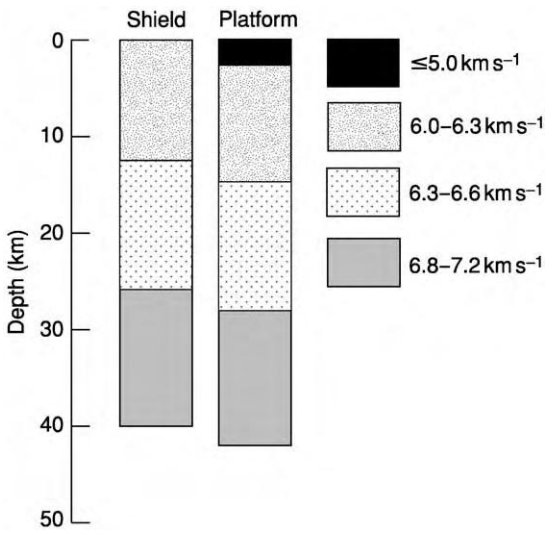


Figure 2 Seismic cross sections of shield and platform crust with typical P wave velocity distributions.

6.3 km s ^{1} in the upper layer, 6.3 to 6.6 km s ^{1} in the middle layer, and 6.8 to 7.2 km s ^{1} in the lower layer. Upper mantle velocities beneath shields and platforms are typically in the range of 8.1–8.2 km s ^{1}, rarely reaching 8.6 km s ^{1}. Seismic reflection studies show an increase in the number of reflections with depth, and generally weak, but laterally continuous, reflections at the Moho, the seismic discontinuity defining the base of the crust.

Table 1 Average chemical composition of continental crust

	Crust composition[a]			
Component	*Upper*	*Middle*	*Lower*	*Total*
SiO_2	66.3	60.6	52.3	59.7
TiO_2	0.7	0.8	0.54	0.68
Al_2O_3	14.9	15.5	16.6	15.7
FeOT	4.68	6.4	8.4	6.5
MgO	2.46	3.4	7.1	4.3
MnO	0.07	0.1	0.1	0.09
CaO	3.55	5.1	9.4	6.0
Na_2O	3.43	3.2	2.6	3.1
K_2O	2.85	2.0	0.6	1.8
P_2O_5	0.12	0.1	0.1	0.11
Rb	87	62	11	53
Sr	269	281	348	299
Ba	626	402	259	429
Th	9.1	6.1	1.2	5.5
Pb	18	15.3	4.2	13
U	2.4	1.6	0.2	1.4
Zr	162	125	68	118
Hf	4.4	4.0	1.9	3.4
Nb	10.3	8	5	7.8
Ta	0.82	0.6	0.6	0.7
Y	25	22	16	21
La	29	17	8	18
Ce	59.4	45	20	42
Sm	4.83	4.4	2.8	4.0
Eu	1.05	1.5	1.1	1.2
Yb	2.02	2.3	1.5	1.9
V	86	118	196	133
Cr	112	150	215	159

[a]Major elements in weight percent of the oxide; trace elements in parts per million.

Composition of the Crust in Cratons

Several approaches have been used to estimate the chemical and mineralogical composition of the crust. One of the earliest methods to estimate the composition of the upper continental crust is based on chemical analysis of glacial clays, which were assumed to be representative of the composition of large portions of the upper continental crust. Probably the most accurate estimates of the composition of the upper continental crust come from extensive sampling of rocks exhumed from varying depths in Precambrian shields and from the composition of Phanerozoic shales. Because the lower continental crust is not accessible for sampling, indirect approaches must be used. These include (1) measurement of seismic-wave velocities of crustal rocks in the laboratory at appropriate temperatures and pressures, and comparing these to observed velocity distributions in the crust, (2) sampling and analysing rocks from blocks of continental crust exhumed from middle to lower crustal depths, and (3) analysing xenoliths of lower crustal rocks brought to the surface during volcanic eruptions.

The average chemical composition of the upper continental crust is reasonably well known. An average composition is similar to granodiorite (Table 1), although there are differences related to the age of the crust. The composition of the middle and lower continental crust is much less well constrained. Uplifted crustal blocks, xenolith populations, and seismic velocity data suggest that the middle crust is intermediate (andesitic) in composition and that the lower crust is mafic (basalt-like) in overall composition. The estimate of total continental crust composition in Table 1 is a mixture of upper, middle, and lower crustal averages in equal amounts. The composition is similar to other published total crustal compositions indicating an overall intermediate composition. Incompatible elements, which are elements that are strongly partitioned into the liquid phase on melting, are known to be concentrated chiefly in the continental crust. During melting in the mantle, these elements are enriched in the magma, and thus are transferred upward into the crust as magmas rise. Between 35 and 65% of the most incompatible elements (such as Rb, Th, U, K, and Ba) are contained in the continents, whereas

continents contain <10% of the least incompatible elements (such as Y, Yb, and Ti).

Cratonization

Cratonization refers to the process of craton formation. Collisional orogenesis occurs when plates carrying blocks of continental crust collide with each other, producing major mountain chains such as the Himalayas. Cratons are the end product of collisional orogenesis, and thus they are the building blocks of continents. Collisional mountain chains are eroded away in 200–400 Ma, leaving the roots exposed at Earth's surface as 'sutures' between pre-existing crustal blocks. The complex amalgamation of crustal blocks and orogen roots comprises today's cratons.

Using a variety of radiogenic isotopic systems and estimated closure temperatures in various minerals, it is possible to track the cooling and uplift histories of cratons. Results suggest a wide variation in cooling and uplift rates, with most orogens having cooling rates of $<2^{\circ}C\,My^{-1}$, whereas a few (such as southern Brittany) have cooled at rates of $>10^{\circ}C\,My^{-1}$ (Figure 3). In most cases, it would appear to take a minimum of 300 My to make a craton. Some terranes, such as Enderbyland in Antarctica, have had very long, exceedingly complex cooling histories lasting for more than 2 Ga. Many orogens, such as the Grenville Orogen (*see* **Grenvillian Orogeny**) in eastern Canada, have been exhumed as indicated by unconformably overlying sediments, and then reheated during subsequent burial and then re-exhumed. In some instances, postorogenic thermal events such as plutonism and metamorphism have thermally overprinted earlier segments of the cooling history of an orogen, such that only the very early high-temperature history (>500°C) and, perhaps, the latest exhumation record (<300°C) are preserved. Fission track ages suggest that final uplift and exhumation of some orogens, such as the 1.9-Ga Trans-Hudson Orogen in central Canada, may be related to the early stages of supercontinent fragmentation.

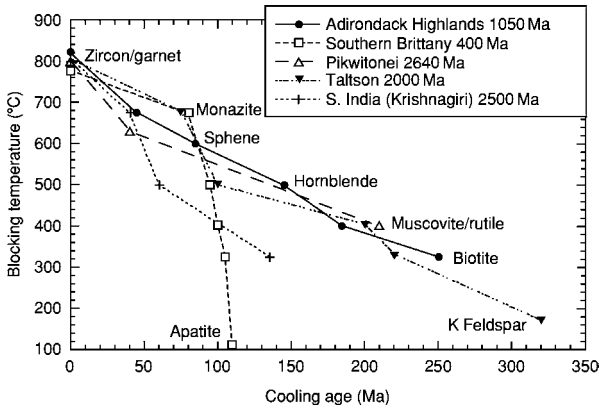


Figure 3 Cooling histories of several orogens leading to the production of stable cratons. Blocking temperature is the temperature at which the daughter isotope is captured by the indicated host mineral.

Crustal Provinces and Terranes

The Canadian Shield can be subdivided into structural provinces based on differences in structural trends and style of folding. Structural trends are defined by foliation, fold axes, and bedding, and sometimes by geophysical anomalies. Boundaries between the provinces are drawn where one trend cuts across another, along either unconformities or structural-metamorphic breaks. Large numbers of isotopic ages from the Canadian Shield indicate that structural provinces are broadly coincident with age provinces. Similar relationships have been described on other continents and lead to the concept of a crustal province (Figure 4).

Terranes are fault-bounded crustal blocks that have distinct lithologic and stratigraphic successions and that have geological histories different from neighbouring terranes (*see* **Terranes, Overview**). Most terranes have collided with continental crust, either along transcurrent faults or at subduction zones, and have been sutured to continents. Many terranes contain faunal populations and palaeomagnetic evidence indicating they have been displaced great distances (thousands of kilometres) from their sources prior to continental collision. For instance, Wrangellia, which collided with western North America in the Late Cretaceous, had travelled many thousands of kilometres from what is now the South Pacific and is now represented in Vancouver Island. Results suggest that as much as 30% of North America was formed by terrane accretion in the past 300 Ma and that terrane accretion has been an important process in the growth of continents.

Terranes form in a variety of tectonic settings, including island arcs, oceanic plateaus, volcanic islands, and microcontinents. Continental crust may be fragmented and dispersed by rifting or strike-slip faulting. In western North America, dispersion is occurring along transform faults such as the San Andreas and Fairweather faults, and in New Zealand movement along the Alpine Transform Fault is fragmenting the Campbell Plateau from the Lord Howe Rise. Baja California and California west of the San Andreas Fault were rifted from North America about 4 Ma, and today this region is a potential terrane moving northwards, perhaps on a collision course with Alaska. Terranes may continue to fragment and

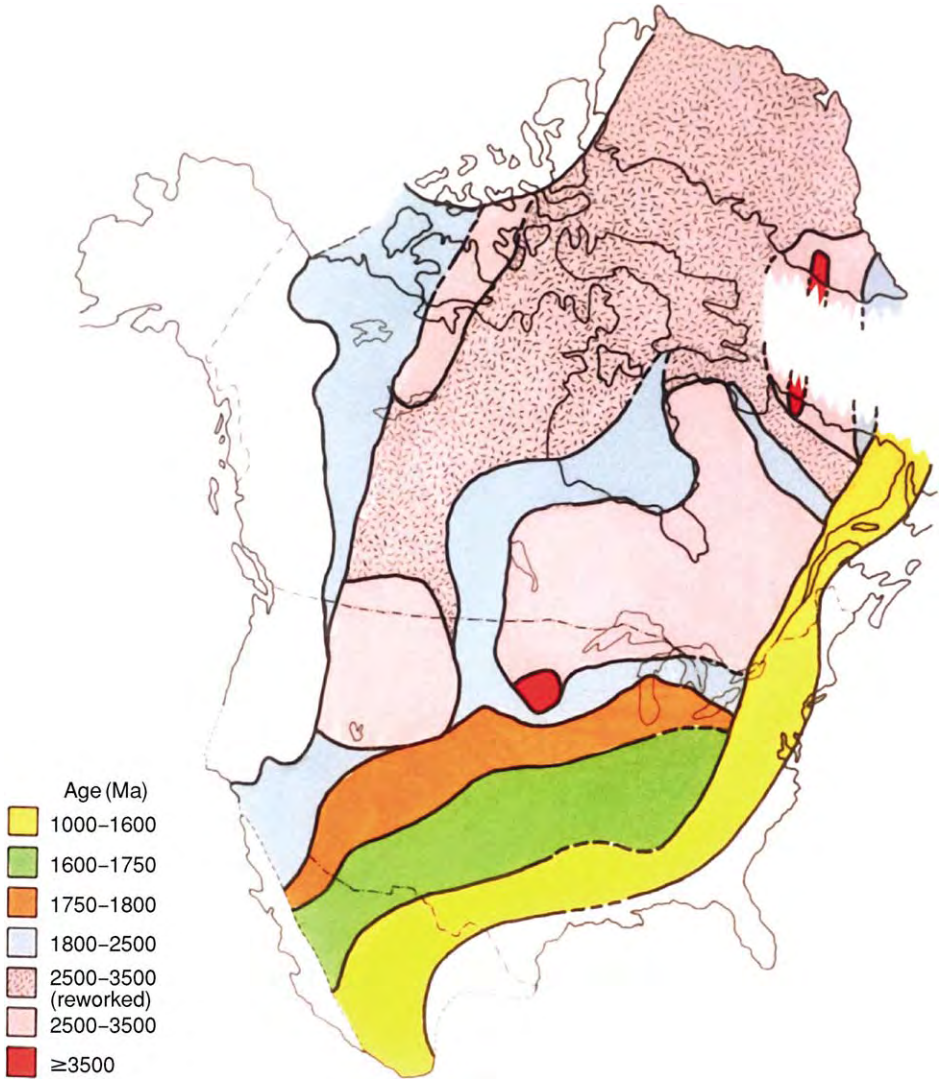


Figure 4 Crustal provinces in North America.

disperse after collision with continents, as did Wrangellia, which is now distributed in pieces from Oregon to Alaska. The 1.9-Ga-old Trans-Hudson Orogen in Canada and the 1.65- to 1.75-Ga-old Yavapai Orogen in the south-west United States are examples of Proterozoic orogens composed of terranes, and the Alps, Himalayas, and American Cordillera are Phanerozoic examples of orogens composed of terranes. Most crustal provinces are composed of terranes, and in turn, cratons are composed of exhumed orogens. In fact, terranes might be considered as the basic building blocks of continents, and terrane collision as a major means by which continents grow in size.

A crustal province is an orogen, active or exhumed, composed of terranes, and it records a similar range of isotopic ages and exhibits a similar postamalgamation deformational history (Figure 4). Shields and platforms are composed of exhumed crustal provinces. Structural trends within provinces range from linear to exceedingly complex swirling patterns reflecting multiple deformation superimposed on differing terrane structural patterns. Exhumed crustal provinces that have undergone numerous episodes of deformation and metamorphism are old orogens. Isotopic dating using several isotopic systems is critical to defining and unravelling the complex, polydeformational histories of crustal provinces.

The definition of 'crustal province' is not always unambiguous. Most crustal provinces contain rocks of a wide range in age and record more than one period of deformation, metamorphism, and plutonism. For instance, the Trans-Hudson orogen in North America includes rocks ranging in age from about 1.7 to 3.0 Gy and records several periods of

complex deformation and regional metamorphism. Likewise, the Grenville province in eastern North America records a polydeformational history, with rocks ranging in age from 1.0 to 2.7 Ga. Some parts of crustal provinces are new mantle-derived crust, known as juvenile crust, whereas other parts represent reworked older crust. Reworking, also known as overprinting or reactivation, describes crust that has been deformed, metamorphosed, and partially melted more than once. There is increasing evidence that crustal reworking results from continental collisions, and large segments of continental crust appear to have been reactivated by such collisions. For instance, much of central Asia at least as far north as the Baikal Rift, which is in a craton, was affected by the India–Tibet collision beginning about 50 Ma. Widespread faulting and magmatism at present crustal levels suggest that deeper crustal levels may be extensively reactivated. In Phanerozoic collisional orogens where deeper crustal levels are exposed, such as the Appalachian and Variscan orogens, there is isotopic evidence for widespread reactivation.

Sediments Deposited on Cratons

Rock assemblages deposited on cratons are mature clastic sediments, chiefly quartz arenites and shales, and shallow marine carbonates. In Late Archaean and Palaeoproterozoic successions, banded iron formation may also be important. Cratonic sandstones are relatively pure quartz sands, reflecting intense weathering, low relief in source areas, and prolonged transport across subdued continental surfaces. Commonly associated marine carbonates are deposited as blankets and as reefs around the basin margins. Transgression and regression successions in large cratonic basins reflect the rise and fall of sea level, respectively.

Depositional systems in cratonic basins vary depending on the relative roles of fluvial, aeolian, deltaic, wave, storm, and tidal processes. Spatial and temporal distribution of sediments is controlled by regional uplift, the amount of continent covered by shallow seas, and climate. If tectonic uplift is important during deposition, continental shelves are narrow and sedimentation is dominated by wave and storm systems. However, if uplift is confined chiefly to craton margins, sediment yield increases into the craton, and fluvial and deltaic systems may dominate. For transgressive marine clastic sequences, shallow seas are extensive and subtidal, and storm-dominated and wave-dominated environments are important. During regression, fluvial and aeolian depositional systems become dominant.

The rates of subsidence and uplift in cratons are a function of the time interval over which they are measured. Current rates are of the order of a few centimetres per year, whereas data from older successions suggest rates 1–2 orders of magnitude slower. In general, Phanerozoic rates of uplift appear to have been 0.1–1 cm year ^{1} over periods of 10^4–10^5 years and over areas of 10^4–10^6 km^2. Craton subsidence can be considered in terms of two stages: in the first stage, subsidence rate varies greatly, whereas the second stage subsidence is widespread. After about 50 Ma, the depth of subsidence decreases exponentially to a constant value.

Several models have been suggested to explain cratonic subsidence. Sediment loading, lithosphere stretching, and thermal doming followed by contraction are the most widely cited mechanisms. Although the accumulation of sediments in a depression loads the lithosphere and causes further subsidence, calculations indicate that the contribution of sediment loading to subsidence must be minor compared to other effects. Subsidence at passive margins may result from thinning of continental crust by progressive creep of the ductile lower crust towards the suboceanic upper mantle. As the crust thins, sediments accumulate in overlying basins.

Supercontinents and Cratons

Supercontinents are large continents that include several or all of the existing cratons. Matching of continental borders, stratigraphic sections, and fossil assemblages are some of the earliest methods used to reconstruct ancient supercontinents. Today, in addition to these methods, we have polar wandering paths, seafloor spreading directions, hotspot tracks, and correlation of crustal provinces. The use of computers in matching continental borders has resulted in more accurate and objective fits. One of the most definitive matching tools in reconstructing plate positions in a former supercontinent is a piercing point. A piercing point is a distinct geologic feature such as a fault or terrane that strikes at a steep angle to a rifted continental margin, the continuation of which should be found on the continental fragment rifted away.

The youngest supercontinent is Pangaea (*see* **Pangaea**), which formed between 450 and 320 Ma and includes most of the existing continents (Figure 5). Pangaea began to fragment about 160 Ma and is still dispersing today. Gondwana (*see* **Gondwanaland and Gondwana**) was a southern hemisphere supercontinent composed principally of South America, Africa, Arabia, Madagascar, India, Antarctica, and Australia. It formed in the latest Proterozoic and was largely completed by the Early Cambrian (750–550 Ma). Later it became incorporated in Pangaea. Laurentia, which is also part of Pangaea,

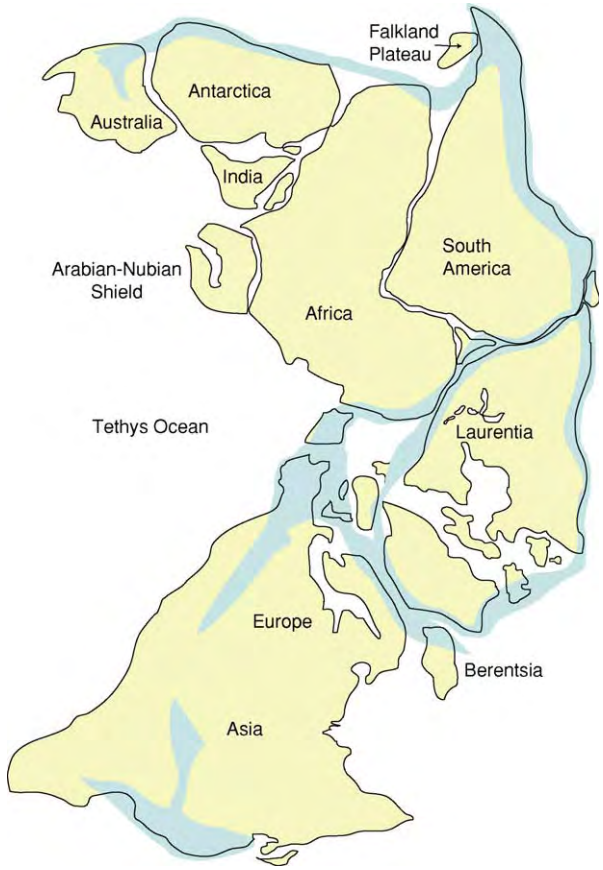


Figure 5 Pangaea, a supercontinent that formed between 450 and 320 Ma and began to fragment about 160 Ma. The major collisional orogens are indicated.

includes most of North America, Scotland, and Ireland north of the Caledonian suture, and Greenland, Spitzbergen, and the Chukotsk Peninsula of eastern Siberia. Although the existence of older supercontinents is likely, their configurations are not well known. Geological data strongly suggest the existence of supercontinents in the Proterozoic and in the Late Archaean. Current thinking is that supercontinents have been episodic, giving rise to the idea of a supercontinent cycle. A supercontinent cycle consists of rifting and break up of one supercontinent, followed by a stage of reassembly in which dispersed cratons collide to form a new supercontinent, with most or all fragments in different configurations, compared to the older supercontinent. The supercontinent cycle provides a record of the processes that control the formation and redistribution of cratons throughout Earth history.

See Also

Analytical Methods: Fission Track Analysis. **Antarctic**. **Earth:** Crust. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Pangaea**. **Precambrian:** Overview. **Tectonics:** Mountain Building and Orogeny. **Terranes, Overview**.

Further Reading

Beardsmore GR and Cull JP (2001) *Crustal Heat Flow.* Cambridge, UK: Cambridge University Press.

Brown M and Rushmer T (eds.) (2003) *Evolution and Differentiation of the Continental Crust*. Cambridge, UK: Cambridge University Press.

Condie KC (ed.) (1992) *Proterozoic Crustal Evolution*. Amsterdam: Elsevier.

Fountain DM, Arculus R, and Kay RW (1992) *Continental Lower Crust*. Amsterdam: Elsevier.

Juteau T and Maury R (1999) *The Oceanic Crust, from Accretion to Mantle Recycling*. New York: Springer Verlag.

Kleine E (2003) The ocean crust. In: Rudnick RL (ed.) *The Crust, Treatise on Geochemistry,* vol. 3, pp. 433 463. Amsterdam: Elsevier.

Leitch EC and Scheibner E (eds.) (1987) *Terrane Accretion and Orogenic Belts*. Geodynamics Series 19. Washington DC: American Geophysical Union.

Meissner R (1986) *The Continental Crust, A Geophysical Approach*. New York: Academic Press.

Moores EM and Twiss RJ (1995) *Tectonics*. New York: WH Freeman.

Rudnick RL and Fountain DM (1995) Nature and compos ition of the continental crust: a lower crustal perspective. *Reviews of Geophysics* 33: 267 309.

Taylor SR and McLennan SM (1985) *The Continental Crust: Its Composition and Evolution.* Oxford: Black well Scientific Publication.

Windley BF (1995) *The Evolving Continents,* 3rd edn. New York: John Wiley & Sons.

SHOCK METAMORPHISM

P S DeCarli, SRI International, Menlo Park, CA, USA

Introduction

The term 'shock metamorphism', synonymous with 'shock wave metamorphism' or 'impact metamorphism', refers to the range of effects produced by the collision of two bodies, e.g., by the collision of an asteroid with the Earth. These effects include fracturing, the formation of planar deformation features (PDF), the formation of high-pressure phases, melting, and vaporization. Our knowledge of shock metamorphism, currently quite incomplete, is derived from laboratory shock experiments, static high-pressure experiments, studies of naturally impacted materials, theoretical analyses, and numerical computations.

It is generally accepted that the history of the solar system is one of repeated collisions between orbiting bodies. Lunar craters, now widely accepted as impact craters, provide a partial record of that history. Only during the past 50 years has it become evident that the Earth, because of its higher gravity, should have experienced about twice as many large craters per unit area as the Moon. Most of these craters on the continental crust have been deformed, modified by erosion, and buried by sediments or volcanism.

Between 1960 and 2003, about 170 terrestrial impact craters were identified, and three to five newly identified craters are added to the list each year.

The minimum velocity of an encounter between the Earth and a body within our solar system is 11.2 km s $^{-1}$, the escape velocity of the Earth. Photographic measurements of meteors, the familiar shooting stars, indicate that they enter the atmosphere at velocities in the range 13–30 km s $^{-1}$; this is an appropriate velocity range for encounters with asteroids. Comets encounter the Earth at velocities in the range of 30 km s $^{-1}$ (short period comets) to 70 km s $^{-1}$ (very long period comets).

The fate of a body entering the Earth's atmosphere at very high velocity depends on such details as the strength and density of the body, its velocity, and its angle of entry. If the body is non-spherical, details of shape and orientation must also be considered, e.g., whether an elongated body enters the atmosphere in point-first or side-first orientation. To simplify further discussion, only vertical impacts of spherical bodies that are strong enough to survive passage through the atmosphere are considered. The interaction of a fast-moving fragile body with the atmosphere can produce an effect equivalent to a large nuclear explosion at an altitude above 20 km. Resultant pressures at the surface may knock down trees, but are too low to produce shock metamorphic effects in minerals.

A very large body, greater than 10 km in diameter, will not be sensibly retarded by the atmosphere. Impact with the Earth will result in the formation of a large crater, greater than 100 km in diameter. Only two craters larger than 100 km in diameter are known to have formed within the past 150 million years. The larger of the two, the 65-million-year-old Chicxulub crater, Yucatan, Mexico, is buried under more than 300 m of carbonate rocks, and was identified in 1981 by the recognition of circular patterns in gravity and magnetic field data. Shock metamorphic features in drill cores have confirmed the identification. The iridium-rich Cretaceous–Teritary K–T boundary layer, which contains shock metamorphosed minerals, coincides with the mass extinction (including dinosaurs) at the end of the Cretaceous, and is believed to be associated with this impact event. Many impact specialists are convinced that the environmental effects of the Chicxulub impact were the primary cause of an abrupt mass extinction. However, many palaeontologists disagree. They argue that the extinction was not abrupt and that there is evidence for other causes. The one matter on which all agree is that the iridium-rich boundary layer serves as an excellent worldwide common time marker that will be essential to further studies of K–T extinctions. Impact specialists continue to search for evidence of large impacts that could be related to other mass extinctions.

Smaller impact events are much more frequent, but the resultant craters are more easily eroded or obscured. Four craters having diameters in the range 7–18 km have been identified as less than 6 million years old. These craters were formed by the impact of bodies in the diameter range of about 300 m to 2 km, sufficiently large to minimize retardation by the atmosphere. Atmospheric retardation becomes significant only for bodies having diameters less than about 10 m, corresponding to masses less than about 1000 tons.

Thus, the velocity of a stony object of 10 m diameter might be reduced from approximately 15 km s $^{-1}$ on atmospheric entry to 10 km s $^{-1}$ on impact with the Earth. The impact would deposit the energy equivalent of approximately 36 000 tons of trinitrotoluene (TNT), and the resultant crater would have a diameter in the range of 100–200 m.

As noted by Melosh in his book on impact cratering (see Further Reading), relationships between crater dimensions and impact parameters are poorly constrained. There are a variety of empirical scaling relations extrapolated from small-scale laboratory impact experiments, high explosive and nuclear experiments, and large-scale computer calculations. Here, we apply the observation that, for a variety of impact conditions, many scaling relations predict D/d, the ratio of the crater diameter to the impactor diameter, to be in the range of 10–20.

Small objects with a mass in kilograms are slowed by atmospheric drag to terminal velocities in the region below about 200 m s^{-1}. The resultant impact pressure of about 1 GPa for an impact on rock is too low to produce shock metamorphic effects other than fracture. Thus, 20 GPa shock metamorphic effects found in some small meteorites may be interpreted as the result of impacts on a meteorite parent body. The exception to this general rule is when there is evidence that a small meteorite is a fragment of a much larger body that impacted the Earth at high velocity. Iron meteorites found in the vicinity of the Meteor Crater (northern Arizona, USA) (1.3 km in diameter) are interpreted as fragments of the rear surface of a 100 000 ton (approximately 30 m in diameter) iron meteorite that is estimated to have impacted the Earth at approximately 20 km s^{-1}.

The bulk of this meteorite was melted or vaporized as a result of very high shock pressures. Intuitively, it might be expected that the entire meteorite would be exposed to the same peak pressure, as predicted by some low-resolution calculations. However, the most recent high-resolution calculations predict that rarefactions originating at free surfaces (the meteorite–air interface) will interact to create low-pressure regions near the rear surface of the meteorite. At least 20 tons of meteorite fragments have been recovered from the vicinity of Meteor Crater. Some of these fragments have shock metamorphic features indicative of peak pressures of less than 10 GPa. Other fragments have shock metamorphic features, including shock-synthesized diamond, indicative of pressures in excess of 100 GPa.

Shock Waves and Large Impacts

Pressure scale definitions: the modern unit of pressure, the pascal, is defined as 1 N m^{-2}. Atmospheric pressure on the Earth at sea-level is approximately 10^5 Pa (100 000 Pa); shock pressures are usually stated as gigapascals (GPa), 10^9 Pa. Earlier literature may refer to bars, kilobars (kb or kbar), atmospheres (atm), dynes per square centimetre (dyn cm^{-2}), kilograms per square centimetre (kg cm^{-2}), and pounds per square inch (psi).

$$1 \text{ bar} = 10^5 \text{ Pa} = 10^6 \text{ dyn cm}^{-2} = 0.9869 \text{ atm} = 1.0197 \text{ kg cm}^{-2} = 14.504 \text{ psi}$$

$$1 \text{ GPa} = 10 \text{ kbar} = 10^{10} \text{ dyn cm}^{-2} = \sim 145\,000 \text{ psi}$$

A collision between two bodies produces a high pressure (shock wave) at the point of impact. The shock wave propagates into both bodies and is attenuated by rarefaction waves originating at free surfaces. The magnitude of the peak pressure depends on both the impact velocity and the relative stiffness of the impacting bodies, as shown in Table 1. The pressure calculations are based on material properties extrapolated from much lower pressures.

Table 1 Parameters of typical Asteriod Earth and Comet Earth Impacts

Impactor target	*Velocity (km s^{-1})*	*Peak pressure (GPa)*	*Fate of impactor*	*Fate of target*
Iron water (ice)[a]	20	~360	Completely molten	Total vaporization of water or ice
Iron alluvium (1.5 g cm^{-3})	20	~400	Completely molten	Total vaporization
Iron granite	20	~750	Partial vaporization	Total vaporization
Iron peridotite	20	~950	Partial vaporization	Total vaporization
Peridotite water (ice)	20	~280	Partial vaporization	Total vaporization
Peridotite alluvium	20	~300	Partial vaporization	Total vaporization
Peridotite alluvium over granite[b]	20	~300, then ~400, shock reflection	Partial vaporization	Total vaporization of alluvium Partial vaporization of granite
Peridotite granite	20	550	Total vaporization	Total vaporization
Peridotite peridotite	20	650	Total vaporization	Total vaporization
Snow (0.6 g cm^{-3}) ice[c]	40	~500	Total vaporization	Total vaporization
Ice alluvium[c]	40	~600	Total vaporization	Total vaporization
Ice ice	40	~650	Total vaporization	Total vaporization

[a] At these very high pressures, the properties of ice and water are indistinguishable.
[b] Shock interactions occur at interfaces between materials having different properties. The 300 GPa shock in alluvium is reflected at the granite interface as a 400 GPa shock moving back into the alluvium. A 400 GPa shock is transmitted into the granite.
[c] The properties of a comet are approximately bracketed by the properties of snow and ice.

The maximum duration of the pressure peak occurs at the point of impact and depends on such details as the relative sizes and shapes of the two bodies, their properties at very high pressure, the impact velocity, and the angle of impact. One popular approximation is that the duration is equal to the impact velocity divided by the diameter of the smaller body. The peak pressure duration for the 20 km s $^{-1}$ impact on the Earth of an asteroid of 20 km in diameter would thus be about 1 s, i.e., between about 0.5 and 2 s, depending on the material properties and the geometry of the impact.

To put these very high impact pressures and durations into perspective, it should be noted that the pressure at the centre of the Earth is about 350 GPa and the effective high-pressure duration of a very large buried thermonuclear explosion is less than 1 ms.

The very high pressures at the point of impact decay in amplitude as they propagate into the Earth. There are two causes of pressure decay. The first is geometric. If the shock front is considered as an expanding hemispherical shell, the peak pressure would be expected to decay as the radius of the hemisphere increases. The second cause of pressure decay is that rarefaction waves originating from the free surface overtake the shock front and reduce the pressure. Melosh's book cites a number of approximate methods for estimating pressure decay in a homogeneous geological environment. Another approach is to perform large-scale computer calculations with 'hydrocodes', computer programs designed to calculate shock wave propagation in various media. These codes were developed for national defence-related purposes to calculate the attenuation of the shock waves produced by large chemical or nuclear explosions. Numerous comparisons of calculations with experimental measurements have shown that agreement within about $\pm 20\%$ can now be achieved over the pressure range between about 1 and 100 GPa, the range over which geological media have been well characterized by Hugoniot (compressional) and release adiabatic measurements. In order to achieve such agreement, the details of geology (including faults and layering) and accurate material properties must be incorporated into the calculation. It is particularly important to include the details of dynamic phase transitions, including the pressure hysteresis between the forward transitions on loading and the reverse transitions on release. The most detailed calculations require extraordinary computer power.

There are many pitfalls in the use of hydrocodes to calculate large impacts on Earth. One problem is the scale of the event. The 20 km s $^{-1}$ impact of a stony asteroid of 10 km in diameter on the Earth releases an energy of approximately 6.3×10^{28} J, about a million times greater than the sum of all nuclear arsenals. The scope of the calculation must be reduced to bring it within the range of even our fastest of current computers. The details of the geological setting must be ignored and the spatial resolution of the calculation must be very coarse, about 1 km. A second problem is that the calculation requires a knowledge of detailed material properties at very high pressures, well above the range of current experimental data. Data on the compression and release behaviour of geological media are virtually non-existent for the pressure range above 100 GPa. Furthermore, most existing data for the range below 100 GPa were obtained in experiments of submicrosecond duration. A few experiments at longer duration, up to 10 μs, have been performed to explore the effects of pressure duration on dynamic phase transitions; kinetic effects were not observed. However, it may be inferred from static high-pressure studies that kinetic effects could possibly be significant in large impact events where the shock duration may exceed 1 s.

Although the equations governing shock wave propagation are simple, the details of shock wave propagation through a typical polymineralic rock are extraordinarily complex. The shock properties of individual minerals can differ substantially; shock interactions occur at mineral boundaries and in the vicinity of pores and cracks. Recent high-resolution hydrocode calculations have studied the details of pressure and temperature equilibration in a rock-like material. On a time-scale of nanoseconds (10^{-9} s) and a distance scale of micrometres, the shock front appears chaotic. Shock collisions around a pore produce nanosecond duration pressure spikes that may be ten times the amplitude of the pressure within an adjacent millimetre-sized grain. These initial pressure inhomogeneities equilibrate within less than a microsecond (for a mineral grain size of about 1 mm) to a uniform pressure. In general, these initial pressure inhomogeneities are ignored. The term 'peak pressure', as it is commonly used (e.g., in Table 1), refers to the pressure after equilibration of the nanosecond duration spikes. However, a knowledge of the complexity of pressure equilibration may be very useful to the researcher who uses microscopic techniques to study shock metamorphic effects on a micrometre scale.

Although accurate high-resolution calculations of a specific large impact may be presently beyond reach, there is nevertheless an excellent qualitative understanding of generic impacts from the study of numerous low-resolution calculations, from small-scale laboratory experiments, and from theoretical considerations, as summarized by Melosh. Figure 1 shows a

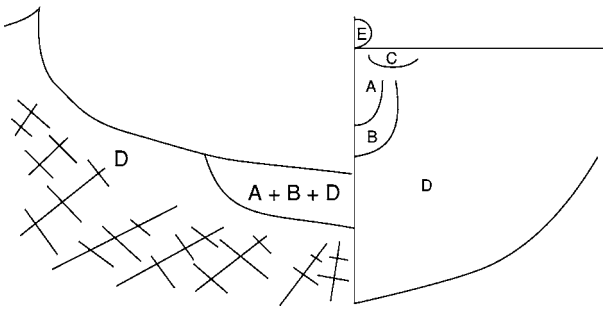


Figure 1 Diagram showing zones of damage by the high velocity impact of a large meteorite. Right: at moment of impact. Left: after cratering is complete. Zone A: very high pressure region above approximately 100 GPa; all minerals melted or va porized. Zone B: high pressure region, approximately 7 100 GPa; most minerals show distinctive shock metamorphic effects. Zone C: near impactor near surface region. Together with the atmos phere, material is ejected at high velocity into a high angle trajec tory. The ejected material can range from strongly shocked (even molten) to weakly shocked. This is the probable source region for tektites, natural glasses of crustal composition. Zone D: weakly shocked fractured region, approximately 0.1 7 GPa; radial and circumferential fractures. Most of the material ejected from the crater originates from this zone and is not strongly shocked. However, block motion in this region can produce localized melting caused by frictional heating. The resultant sheet like formations are called pseudotachylites. Zone E: the impacting body. Melted or vaporized, for the most part, but a small fraction may survive in a relatively unshocked condition. The melted and vaporized material can be ejected into the upper atmosphere and subsequently deposited over a large area. Left: A + B + D, breccia lens; an intimate mixture of shock metamorphosed clasts with melt and weakly shocked material. If melt predominates, the breccia is called tagamite. If the melt is less dominant, the breccia is called suevite.

diagram depicting the effects of an impact, referenced to the pre-impact setting and approximately scaled to the diameter of the impacting body.

Shock Metamorphic Effects

The impact event depicted in Figure 1 can produce numerous shock metamorphic effects. Ejected material, including solid, solidified melt, and condensed vapour, can serve as a stratigraphical marker. The breccia lens provides most of the evidence to distinguish an impact crater from a volcanic crater. As noted above, the breccia lens is a heterogeneous mixture of high-pressure and lower pressure material. This is advantageous in that the lower pressure clasts have been subjected to only minor shock heating and serve to quench the more strongly shock-heated and shock-melted material. Quenching helps to preserve more fragile high-pressure phases, such as stishovite. Some common shock metamorphic effects are presented in Table 2.

Controversial Issues

There are numerous controversies in the literature on shock metamorphism, as should become evident to the person who reads more than one of the items in the Further Reading section at the end of this article. The controversies are, in general, a measure of the incompleteness of our knowledge. Almost without exception, published experimental data and observations are trustworthy and reproducible. The majority of controversies centre on the interpretations of the data, inferences from hydrocode calculations, or the validity of often unstated assumptions.

For example, tektites (distinctive forms of natural glass showered down on the Earth) show unequivocal signs of sculpturing by aerodynamic forces during high-speed entry into the Earth's atmosphere. Although the detailed chemical composition of tektites implies an Earth origin, simple physical arguments indicate that it is impossible to propel tektites into space through the Earth's atmosphere. Those who therefore argued against a terrestrial origin of tektites failed to consider the possibility that a large impact on Earth could melt crustal material and propel it into space, together with a portion of the atmosphere. Large-scale hydrocode calculations indicate that this latter possibility is plausible. The chemical evidence for a terrestrial origin of tektites is so strong that it overwhelms concerns that a calculation having a resolution of 1 km is used to predict the fate of centimetre-sized objects.

There is a related controversy over the minimum shock pressure to which Martian meteorites have been exposed. There is very strong chemical evidence that certain meteorites found on the Earth actually originated on Mars. Some scientists have argued that any meteorite ejected from Mars by a large impact would have melted on release from the requisite high pressure, about 150 GPa. Although various investigators have disagreed about the peak pressure implied by shock metamorphic effects, they have agreed that the meteorites were not melted by their ejection from Mars. Subsequent calculations have indicated that the Martian meteorites could have been accelerated to the Martian escape velocity of 5 km s ^{1} by shock pressures as low as 65 GPa. Some investigators have interpreted these calculations as evidence that all Martian meteorites must necessarily have been exposed to shock pressures of 65 GPa or higher. However, one Martian meteorite shows remanent magnetism that would have been destroyed by shock pressures exceeding about 20 GPa. If the magnetic data and the evidence for a Martian origin are accepted, there must be an even lower pressure mechanism for accelerating a Martian rock to escape

Table 2 Shock metamorphic effects

Effect	*Source material*	*Pressure (GPa) (single shock)*	*Comments*
Melting	Iron	>170	Melts on release of pressure
Melting	Olivine, pyroxene	>100	Melts on release of pressure
Melting	Quartz, granite	>50	Melts on release of pressure
Melting	Sand, soil	>20, possibly as low as 7	Energy increase on shock compression much greater for porous materials
Diaplectic glass	Quartz, feldspars	>15, possibly as low as 7	Diaplectic glass forms by solid state transformation. It is amorphous, but retains original crystal form and usually has a higher refractive index than melt glass. Lower bound pressure from PDF formation
Stishovite, hollandite	Quartz, feldspars	>15, possibly as low as 7	Stishovite, hollandite, polymorphs of quartz, and feldspar found in impact craters and meteorites. Lower bound pressure from PDF formation
Coesite	Quartz	>15, possibly as low as 3	Coesite found in impact craters in association with diaplectic glass, implying that it formed on release of pressure. Could conceivably be found in a pseudotachylite that solidified under pressure
Ringwoodite, wadsleyite	Olivines	>15	Found in meteorite melt veins; pseudotachylite like structures that were quenched at high pressure
Akimotite, majorite	Pyroxenes	>17	Found in meteorite melt veins; pseudotachylite like structures that were quenched at high pressure
Diamond, cubic and hexagonal mixture	Graphite	>25	Found in meteorites; $P \sim 100$ GPa from graphite in iron meteorites. Found in impact craters; $P \sim 30$ GPa from graphite in gneiss. Made in laboratory shock experiments
Cubic diamond	Porous carbon	>15	Made in laboratory shock experiments
Planar deformation features (PDFs)	Quartz, feldspar predominantly. Also other minerals	>7	PDFs in quartz are a primary diagnostic for impact. A PDF is a lamellar feature aligned with a low index crystallographic plane. A number of different orientations may appear in the same grain. There is evidence that the lamellae contain high pressure phases that invert to low pressure forms during electron microscopy
Fractures	All rocks	>~0.1	Laboratory shock experiments show dynamic fracture strength comparable (~1.5 times) to static strength
Pseudotachylite formation	All rocks	>~0.1	Pressure estimate based on observation of pseudotachylites in the fractured zone

Numerous other high pressure minerals have been observed in meteorites and impact craters. The most common and readily observed are listed. The book by French (see Further Reading) contains numerous micrographs of shock metamorphosed quartz.

velocity. Melosh has suggested one such mechanism: entrainment of the rock in the vapour plume formed by strongly shocked material (see **Table 1**).

Finally, there are long-standing controversies over the peak pressures associated with various metamorphic effects. Shock metamorphic effects in rocks and minerals have been studied in numerous laboratory shock experiments over the past 45 years. It was initially hoped that a peak shock pressure calibration could be established based on the presence of various metamorphic effects. The assumption that the peak shock pressure is the only significant parameter seems to be incorrect, as may be inferred by analyses of apparent conflicts in research reports. These conflicts can usually be resolved by considerations of experimental differences between the experiments. Samples loaded to the same peak pressure via different loading paths (single shock vs. a sequence of shock reflections) often show marked differences in metamorphic effects. Although the pressure duration of laboratory shock experiments is in the range of a microsecond, the shock pressure duration for a large natural impact may exceed a second. The interpretation of metamorphic effects on the basis of laboratory static high-pressure data may be more appropriate in this regime. Some controversies over the precise peak pressure to which a given natural sample has been exposed may not be resolvable on the basis of present knowledge. However, there is usually no argument about whether a given sample has been shock metamorphosed at all.

See Also

Impact Structures. **Solar System:** Meteorites; Mercury; Moon; Mars. **Tektites**.

Further Reading

Desonie D (1996) *Cosmic Collisions, A Scientific American Focus Book*. New York: Henry Holt and Co.

French BM (1998) *Traces of Catastrophe: A Handbook of Shock Metamorphic Effects in Terrestrial Meteorite Impact Structures*. LPI Contribution 954. Houston, TX: Lunar and Planetary Institute.

Koeberl C and Martinez Ruiz F (eds.) (2003) *Impact Markers in the Stratigraphic Record*. Berlin, Heidelberg: Springer Verlag.

McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

Melosh HJ (1989) *Impact Cratering: A Geologic Process*. Oxford: Oxford University Press and Oxford: Clarendon Press.

Rubin AE (2002) *Disturbing the Solar System: Impacts, Close Encounters, and Coming Attractions*. Princeton: Princeton University Press.

SOIL MECHANICS

J Atkinson, City University, London, UK

Soil and Mechanics

Engineering Soils

Soil mechanics describes the mechanical behaviour of granular materials. Mechanical behaviour covers strength, shear stiffness, volumetric compressibility, and seepage of water. Granular materials include powders, grain, and other foodstuffs, mineral ores and concentrates, as well as natural soils.

The simple theories of soil mechanics are intended for collections of grains which are uncemented or only very slightly cemented and which contain fluid, usually water or air, in the pore spaces. This covers dense and loose sands and soft and stiff clays. Rock mechanics describes the behaviour of strongly bonded grains whose overall behaviour is governed by joints and fractures. There is a range of materials between these, including weathered rocks, weak rocks, and cemented soils for which simple theories of soil mechanics have limited application.

The theories of soil mechanics apply equally to sands (coarse-grained soils) and clays (fine-grained soils). Figure 1 shows samples of sand and clay under load in unconfined compression. In each case the strength arises from suctions in the pore water. The clay is stronger than the sand because it can sustain larger suctions: otherwise their behaviour is fundamentally the same.

In describing theories for the behaviour of materials some mathematics is unavoidable. In the following, the mathematics is kept as simple as possible and does not extend beyond simple algebra. Only the most basic and fundamental equations and parameters are included.

Mechanics: Strength, Stiffness, Compressibility, and Permeability

Soils are highly compressible. The volume decreases significantly as it is compressed under an isotropic stress state. This is illustrated in Figure 2C. Soils also change volume when they are sheared and distorted.

Strength is basically the maximum shear stress a soil can sustain before it fails. Stiffness is the distortion which occurs as the soil is loaded before it fails. These are illustrated in Figure 2(D). G is the shear modulus and describes stiffness: τ_f is the shear stress after large distortion and it is the strength. In soils both strength and stiffness increase with increasing mean stress.

The frictional nature and the coupling between shear and volume change are the two main differences between the mechanical behaviour of granular materials and the mechanical behaviour of metals and other similar materials.

Figure 1 Unconfined compression of sand and clay.

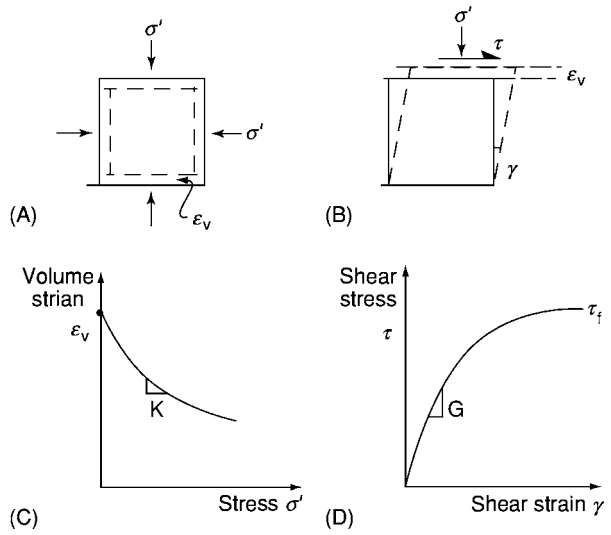


Figure 2 Compression and distortion.

A Brief History of Soil Mechanics

Coulomb and Soil Strength

Theories for soil mechanics originated around the middle of the eighteenth century. Coulomb was a military engineer and he was concerned with calculating soil loads on masonry retaining walls. He carried out experiments on the strength of soils and he found that the resistance of soil to shear loading had two components, one cohesive and the other frictional. Coulomb tested unsaturated samples and his analyses were in terms of forces, not stresses. Methods for analysis of stress discovered later by Mohr were incorporated into Coulomb's results and this is the basis of the well-known Mohr–Coulomb strength equation:

$$\tau_f = c + \sigma \tan \phi \qquad [1]$$

Neither Coulomb not Mohr had a clear understanding of the importance of pore pressures and effective stresses and the original Mohr–Coulomb equation is in terms of total stress. It is now known that the Mohr–Coulomb equation for soil strength is limited but it is still widely used.

Terzaghi and Effective Stress

Karl Terzaghi was an Austrian civil engineer. The major contribution which he made to soil mechanics was to set out a clear theory in the 1920s for accounting for the influence of pore pressure on soil strength and deformation. He proposed an effective stress σ' which controls all soil behaviour and he discovered that for saturated soil this is related to total stress σ and pore pressure u by:

$$\sigma' = \sigma - u \qquad [2]$$

The Terzaghi effective stress equation has been found to apply for a very wide range of loadings and soils, and it is used universally for geotechnical analysis of saturated soils.

Plasticity and Cam Clay

In the 1960s Andrew Schofield and Peter Wroth were lecturers at Cambridge University. They applied the then relatively new theories of plastic flow to frictional materials and created a complete stress-strain theory for soils. The model they developed they called Cam Clay and this remains the basis for many of the current constitutive equations for soils.

These theories of frictional strength, effective stress, and plastic flow are the basic buiding blocks for modern soil mechanics.

Effective Stress and Drainage

Principle of Effective Stress

The Principle of effective stress first proposed by Terzaghi in 1923, states that the stress which is effective in determining strength, stiffness, and compressibility, the effective stress, σ' is given by eqn [2].

Total stresses arise from external loads due to foundations and walls and loads from self weight. Pore pressures are the pressures in the fluid in the pore spaces. For dry soils the pore pressure is the pressure in the air in the pores. For saturated soil it is the pore water pressure. For soils which are not fully saturated and which contain both air (or gas) and water in the pores the equivalent pore pressure is some combination of the air and water pressure. At present there is no simple and robust theory for determining the equivalent pore pressure and effective stress in unsaturated soils.

So far as is known, the principle of effective stress and the effective stress equation (eqn [2]) holds for all dry or saturated soils over a very wide range of stress and pore pressure up to several tens of MPa. The strength and stiffness of soil 1 m below the bed of the deep ocean, where the depth of water may exceed 5 km, will be the same as that of soil 1 m below the bed of a duck pond.

Drainage and Consolidation

Because water is relatively incompressible in comparison with soil, volume changes in soil can occur only if water can flow into or out from the pore spaces. Whether or not this happens depends on the rate of drainage and the rate of loading.

If water cannot drain from the soil it is said to be undrained: its volume must remain constant but pore pressures will change in response to the loading. If water has time to drain freely from the soil it is said to be drained: pore pressures remain constant and volume changes occur. Hence:

$$\text{Undrained loading}: \quad \delta V = 0 \text{ and u changes}$$

$$\text{Drained loading}: \quad \delta u = 0 \text{ and volume changes}$$

where the symbol δ means 'a change of.' In many cases soil in the ground is neither fully drained not fully undrained but simple soil mechanics theories are applicable only for fully drained or fully undrained cases.

If soil is loaded undrained the resulting pore pressures will not be in equilibrium with the long-term groundwater pressures. As the excess (out of balance) pore pressures dissipate under constant total stress there will be changes of effective stress and volume changes. This process is known as consolidation. Because the rate of drainage during consolidation depends on hydraulic gradient, which decreases as excess pore pressures dissipate, the rate of consolidation diminishes with time.

Description and Classification of Engineering Soils

There are standard schemes for description and classification of soils for engineering purposes. These essentially classify soils under the two main headings: the nature of the grains and how they are packed together. For natural soils, descriptions are added for structure including bonding, bedding, and discontinuities.

The Nature of the Soil: Characteristics of the Grains

The most important characteristic is the grain size or grading. Figure 3 shows the range of grain sizes commonly found in natural soils and their descriptions (e.g., sand is 0.06 m to 2 mm). The range is very large. Clay grains are of the order of 1000 times smaller than coarse sand grains. Since permeability is related to the square of the size; sands are of the order of 1 million times more permeable than clays. If a soil is essentially single sized it is poorly graded (or well sorted). If it contains a range of sizes it is well graded (or poorly sorted). In a well graded soil it is usually the 10% smaller than (D_{10}) size which governs drainage.

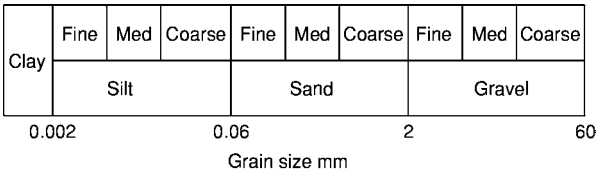


Figure 3 Grain size descriptions.

Grains of silt size and larger normally consist of rock fragments. They may be rounded or angular, rough or smooth. Grains of clay size are normally made of a clay mineral belonging to one of the major families which are kaolinite, illite, and montmorillonite (smectite). These may be distinguished by their Atterberg Limits and Activity (see below). The characteristics of the grains do not effect the fundamental behaviour of soils but they do influence numerical values of strength and stiffness parameters.

Rates of Loading and Drainage

In soil the rate of drainage depends primarily on the permeability which itself depends on the grading of the soil. The Hazen formula for coefficient of permeability k is

$$k \propto D_{10}^2 \qquad [3]$$

where D_{10} is the size of the grains with 10% smaller. Typical values for coefficient of permeability range from greater than 10^{-2} ms^{-1} (approx. 1.5 m in a minute) for coarse-grained soils to less than 10^{-8} ms^{-1} (approx 1 m in 3 years) for fine-grained soils. This very large difference (more than 1 million times) in rate of drainage between coarse-grained and fine-grained soils accounts for many of the differences in observed behaviour of sands and clays.

The rate of loading also varies widely. Some natural processes, such as deposition and erosion, occur relatively slowly (over decades) while others, such as earthquakes, occur relatively rapidly (over a few seconds). In construction, a shallow trench might be dug in a few hours and a large dam built in a few years.

In determining whether a certain event applied to a certain soil is drained or undrained, it is necessary to consider both the rate of drainage and the rate of loading. During earthquakes, coarse-grained sandy soils may be undrained causing liquefaction failure. In construction it is usual to take fine-grained clay soils as undrained and coarse-grained sand soils as drained.

Atterberg Limits

If a clay soil has a very high water content it will flow like a liquid; if it has a low water content it will become brittle and crumbly. For intermediate water contents it will be plastic. The Atterberg Limits, the Liquid Limit, W_l, and the Plastic Limit, W_p, define the range of water content over which a clay soil is plastic. The Plasticity Index, I_P, is the difference between the Liquid and Plastic Limits:

Table 1 Typical values for some characteristic soil parameters

Parameter	*Symbol and units*	*Kaolinite clay (China clay)*	*London clay*	*Alluvial sand (Thames)*	*Carbonate sand*	*Decomposed granite (Dartmoor)*
Liquid Limit	W_l	65	75			
Plastic Limit	W_p	35	30			
Plasticity Index	I_p	30	45			
Activity	A	0.4	1			
Maximum specific volume	V_{max}	2.72	2.98	2.2	3.2	
Minimum specific volume	V_{min}	1.92	1.80	1.5	2.0	
Coefft of compressibility	Cc	0.44	0.37	0.37	0.78	0.21
Coefficient of swelling	Cs	0.11	0.14	0.03	0.01	0.01
Specific volume on NCL at σ' 1 kPa	Vn	3.26	2.68	3.17	4.8	2.17
Specific volume on CSL at σ' 1 kPa	Vc	3.14	2.45	2.99	4.35	2.04
Critical state friction angle	ϕ'c degrees	25	23	32	40	39
Very small strain shear modulus at σ' 100 kPa on the NCL	G′o MP_a	40	15	60	60	60

(Data from research at City University, London.)

$$I_P = W_l - W_p \quad [4]$$

For a natural clay soil, which may contain silt and sand sized grains, the Activity is

$$A = \frac{I_P}{\%clay} \quad [5]$$

and this is related to the mineralogy of the clay, as shown in Table 1. Many numerical values for soil parameters are related to the clay mineralogy and to the Atterberg Limits.

State: Liquidity Index and Relative Density

Grains in a soil may be densely packed or loosely packed or in an intermediate state of packing. The packing influences strength and stiffness, as shown in Figure 4. In a clay soil, the loosest packing corresponds to the Liquid Limit and the densest to the Plastic Limit. Intermediate states are described by the Liquidity Index:

$$I_l = \frac{w - W_p}{I_P} \quad [6]$$

where w is the water content. At the Liquid Limit, the Liquidity Index is 1.0 and at the Plastic Limit it is 0, as shown in Figure 4.

In a coarse-grained soil the loosest packing corresponds to the maximum water content, w_{max}, and the densest to the minimum water content, w_{min}. Intermediate states are described by the relative density:

$$I_d = \frac{w_{max} - w}{w_{max} - w_{min}} \quad [7]$$

At the loosest state the Relative Density is 0 and at the densest state it is 1.0, as shown in Figure 4.

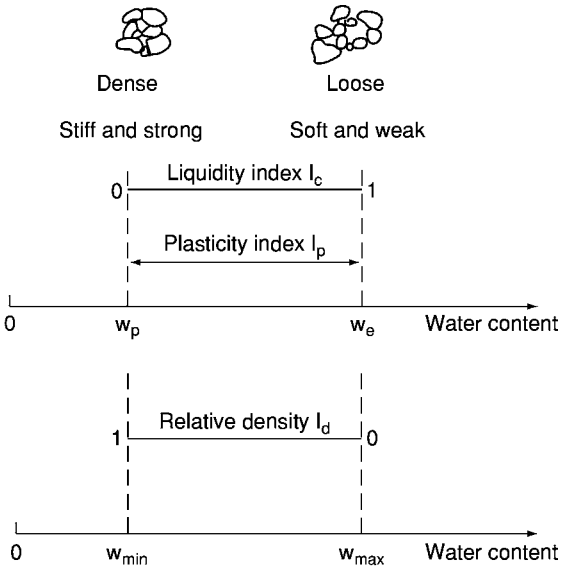


Figure 4 Packing: plasticity index and relative density.

Because soil strength and stiffness are essentially frictional they depend on the current effective stress. Packing, described by Liquidity Index or Relative Density, is not sufficient itself to describe soil behaviour. Soil state will be defined by a combination of packing and stress, as discussed later.

Behaviour in Compression: Change of Size

Isotropic Compression and Swelling

As saturated soil is loaded and unloaded under drained conditions water flows from and into the soil as it compresses and swells, rather like a sponge. The change in volume with changing effective stress is

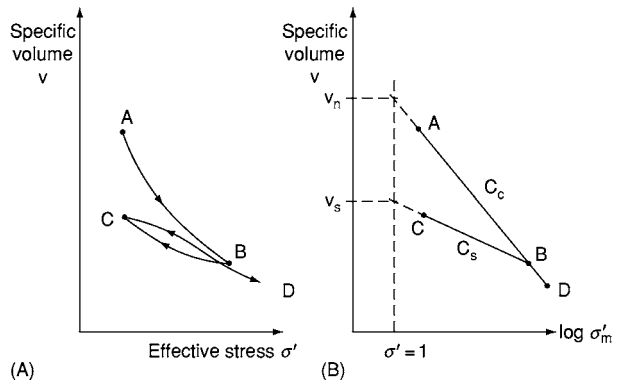


Figure 5 Isotropic compression and swelling.

illustrated in Figure 5(A). The soil is first loaded from A to B and it compresses. The compression of the soil skeleton is due mostly to particles rearranging and also to weak coarse grains fracturing or clay grains bending. The soil is unloaded from B to C and reloaded back to B. Some strains are recovered and there is a hysteresis loop. Volume changes in coarse-grained soils will be small because grains do not 'un-rearrange' or 'unfracture' but may be significant in clay soils as the grains can unbend.

In Figure 5B, effective stresses are plotted to a logarithmic scale and the compression and swelling curves have been idealised. The volume axis is the Specific Volume defined as

$$v = \frac{V}{V_s} \qquad [8]$$

where V_s is the volume of soil grains in a volume V of soil. For many soils v will range from about 1.2 if the soil is dense to over 2 if it is loose. The linear normal compression line ABD is given by

$$v = v_n - C_c \log \sigma' \qquad [9]$$

and the linear swelling and recompression line CB is given by

$$v = v_s - C_s \log \sigma' \qquad [10]$$

The Compression Index, C_c, the Swelling Index, C_s, and the Specific Volume, v_n, at unit stress are material parameters and are related to the characteristics of the grains. Typical values are given in Table 1. The location of a swelling line is given either by the maximum stress, σ'_m, or the specific volume, v_s, at unit stress.

A soil whose state lies on the line ABD is said to be normally compressed and ABC is the normal compression line (NCL). Soil whose state is on the NCL has not experienced a larger stress. A soil whose state is on a swelling line, such as CB, is said to be overconsolidated; it has experienced a greater stress, σ'_m. The overconsolidation ratio is

$$R_0 = \frac{\sigma'_m}{\sigma'} \qquad [11]$$

Equations [9] and [10] relate volume to stress for isotropic loading and unloading. Since the stress scale is logarithmic, the stress-strain behaviour is non-linear; the bulk modulus is not a constant but varies with both stress and overconsolidation.

The idealization of the hysteresis loop in Figure 5A to the line CB in Figure 5(B), common in simple soil mechanics theories, is unrealistic for many soils. Soil stiffness will be discussed later.

One-dimensional Compression in the Ground

Below level ground, the state of stress is not isotropic but one-dimensional, with zero horizontal strain during deposition and erosion; the vertical and horizontal effective stresses σ'_v and σ'_h are related by the coefficient of Earth pressure, K_o, given by

$$K_o = \frac{\sigma'_h}{\sigma'_v} \qquad [12]$$

For normally consolidated soil ($R_o = 1$) K_{onc} is given by

$$K_{onc} = (1 - \sin \phi'_c) \qquad [13]$$

where ϕ'_{cs} is the critical state friction angle. K_o increases with overconsolidation ratio. Horizontal effective stresses given by eqn [12] are for level ground with zero horizontal strain. Near slopes, foundations, and other underground construction stresses will be modified by the stresses imposed by the slope and the structure.

Calculations of settlement in the ground are often carried out in terms of a coefficient of compressibility, m_v, or a one-dimensional modulus, M, given by

$$M = \frac{1}{m_v} = \frac{\Delta\sigma'_v}{\Delta v/v} \qquad [14]$$

where Δv is the change of specific volume observed in a laboratory test on a soil sample with initial specific volume, v, when subjected to an increment of vertical stress, $\Delta\sigma'_v$. Since soil stiffness is non-linear, M is not a soil constant and the increment of stress applied in the test should correspond to the expected change of stress in the ground.

State: Stress and Packing

The behaviour of a particular soil depends on both the current effective stress and on the Relative

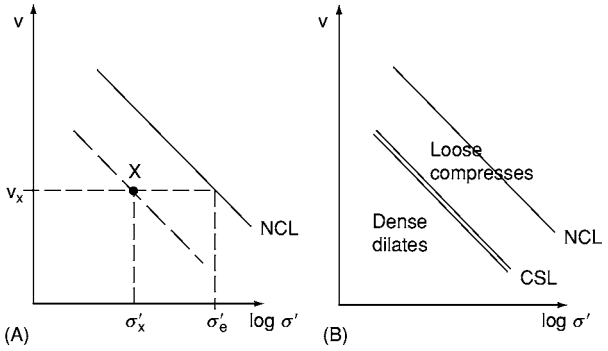


Figure 6 States and state parameters.

Density or Liquidity Index. These may be combined into a state parameter.

Figure 6A, which is similar to Figure 5B, shows the state of an overconsolidated sample at X and the normal compression line. All samples with states on the broken line through X parallel to the NCL will behave in a similar way. These states can be described by a stress state parameter S_σ given by

$$S_\sigma = \frac{\sigma'_x}{\sigma'_e} \qquad [15]$$

where σ'_e is the equivalent stress on the NCL at the same specific volume as that at X. The state parameter describes the distance of the state from the NCL. If the swelling index C_s is small, S_σ is approximately equal to the overconsolidation ratio, R_0. The concept of state is of fundamental importance for soils which are both frictional and which change volume during loading, as it combines both relative density and stress into a single parameter.

Dense and Loose States

After shearing, soils reach ultimate or critical states in which they continue to distort at constant state (i.e., at constant stress and volume). The relationship between specific volume and effective stress gives a critical state line (CSL) parallel to the normal compression line, as shown in Figure 6(B). The critical state line is given by

$$v = v_c - C_c \log \sigma' \qquad [16]$$

Soil states which are above the CSL are known as 'loose of critical' and the soil will compress on shearing. Soil states which are below the CSL are known as 'dense of critical' and the soil will dilate on shearing.

The CSL separates regions of fundamentally different behaviour of the same soil. A soil which has a relatively low specific volume and the grains are relatively closely packed will compress if the effective stress is very large. Similarly, a soil which has a relatively high specific volume and the grains are relatively loosely packed will dilate if the effective stresses are very small.

Sediments at great depth deform plastically. Near-surface soils often behave in a brittle manner and crack. Relative Density, or Liquidity Index, on its own is not sufficient to predict the behaviour on subsequent shearing; the effective stress must be taken into account as well.

Strength of Soil

Behaviour of Soil During Shearing

Figure 7A shows a block of soil with a constant normal effective stress σ' subjected to an increasing shear stress τ'.

The soil is drained and it distorts with a shear strain γ and a volumetric strain ε_v. If the soil is undrained, there are no volume changes but the pore pressures change. The block of soil represents conditions inside a slip zone in the slope illustrated in Figure 7B or in a foundation illustrated in Figure 7C. If the slope is created by excavation or erosion the normal stress decreases and, since soil strength is frictional, it will weaken, whereas below the loaded foundation the normal stress increases and the soil becomes stronger.

The behaviour of soil initially loose and initially dense of critical is illustrated in Figure 8. The loose soil (marked L) compresses during shearing even

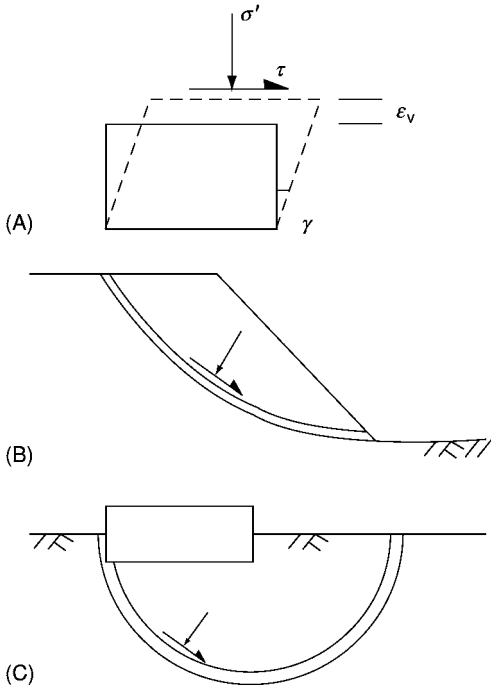


Figure 7 Shearing of soil.

though the normal stress remains constant and the dense soil (marked D) dilates. The rate of dilation is given by an angle of dilation ψ, given by

$$\tan\psi = -\frac{d\varepsilon_v}{d\gamma} \qquad [17]$$

(The negative sign is required as ψ is positive for negative (dilation) volumetric strains.)

Critical State Strength

The samples shown in Figure 8 have the same effective stress and they reach the same critical shear stress and the same critical specific volume after relatively large strains. Figure 9 shows critical states for a number of samples. There are unique relationships between the critical shear stress τ'_f, the critical normal stress σ'_f, and critical specific volume v_f, given by

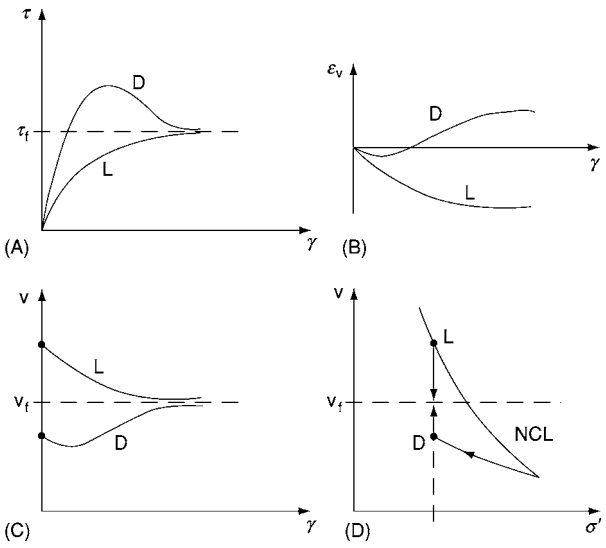


Figure 8 Stress and volume change in shearing soil.

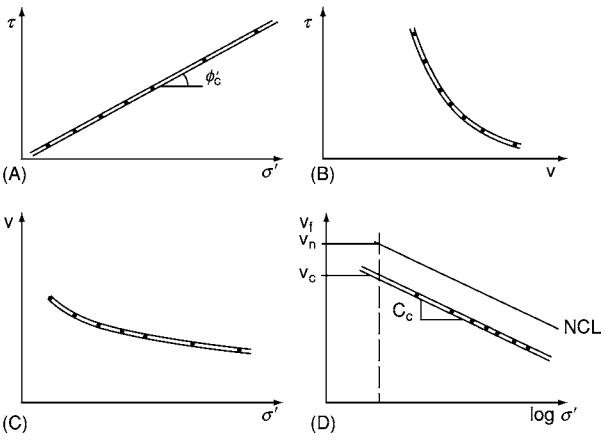


Figure 9 Critical states.

$$\tau_f = \sigma'_f \tan\phi'_c \qquad [18]$$

$$\tau_f = v_c - C_c \log\sigma'_f \qquad [19]$$

These equations define a critical state line and the parameters ϕ'_c, C_c, and v_c are material parameters. (Critical state lines are usually shown as double lines, as in Figure 9). Typical values are given in Table 1. During shearing distortions, all soils will ultimately reach a critical state; if they did not they would continue to change state indefinitely, which is impossible. In simple soil mechanics theories, the critical states reached by a particular soil, given by eqns [18] and [19], are independent of the starting state and whether the soil is drained or undrained.

Undrained Strength

Figure 9B shows that the shear stress at failure, which is the shear strength, decreases as the specific volume at failure increases. If soil is undrained the water content and the undrained strength remain unchanged for any changes in total normal stress. The undrained strength.

$$\tau_f = s_u \qquad [20]$$

depends on the water content. In practice, samples are taken from the ground and tested without change of water content. The undrained strengths measured can be used for design so long as the water content in the ground does not change.

It is common knowledge that soils become weaker as their water content increases. This is shown in Figure 10 in which the undrained strength, with a logarithmic scale, decreases linearly with water content. The strength of soil at its Liquid Limit is approximately 1.5 kPa and the undrained strength

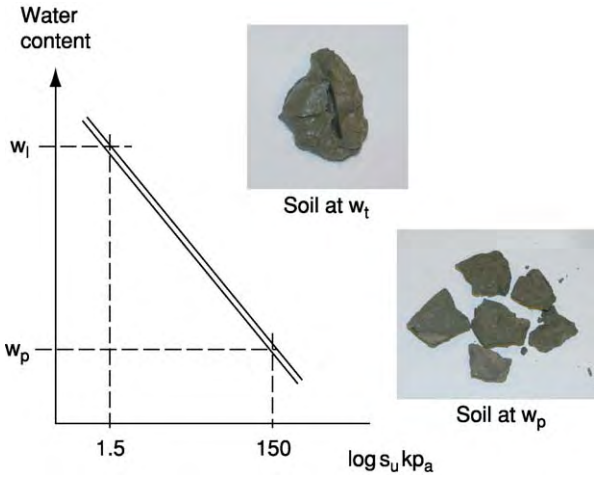


Figure 10 Undrained strength and water content.

of soil at its Plastic Limit is approximately 150 kPa (i.e., the strength of soil changes by about 100 times as the water content changes from the Liquid Limit to the Plastic Limit).

Peak Strength

Soils whose initial states are dense of critical have a peak strength before they reach a critical state, and they dilate during drained shear, as shown earlier in Figure 8. The peak strengths vary with effective normal stress and specific volume, as shown in Figure 11.

Samples which reach their peak states at the same specific volume have peak strengths on an envelope shown in Figure 11A. The envelope is often approximated by a straight line, shown in Figure 11A given by

$$\tau_p = c'_p + \sigma'_p \tan \phi'_p \qquad [21]$$

The peak friction angle, ϕ'_p, is a material parameter and, from Figure 11A $\phi'_p < \phi'_c$. The cohesion intercept, c'_p, is not a material parameter and its value depends on the specific volume. Moreover c'_p is not the strength at zero effective normal stress, as this must be zero for an uncemented granular material.

The linear approximation for peak strength given by eqn [21] is applicable only within the range for which data are available. Figure 11B shows additional data at smaller normal effective stresses; there the envelope is now distinctly curved and passes through the origin. The curved peak failure envelope, shown in Figure 11C, can be represented by a power law of the form.

$$\tau_p = A\sigma'^b \qquad [22]$$

where b is a material parameter and A depends on the specific volume.

From analyses of the stresses and strains in the soil block, shown in Figure 7A, peak shear strength is given by

$$\tau_p = \sigma' \tan (\phi'_c + \psi) \qquad [23]$$

At the critical state, $\psi = 0$ and τ'_c is given by eqn [18]. At the peak state, the angle of dilation is at a maximum. The maximum rate of dilation is governed by the state parameter so the peak strength increases as the initial state moves away from the critical state line.

Equations [21, 22 and 23] are alternative theories for the peak strength of soils. They all contain a combination of material parameters and state dependent parameters. Equations [22 and 23] correctly give zero strength at zero effective stress. Equation [21] is most commonly applied in practice.

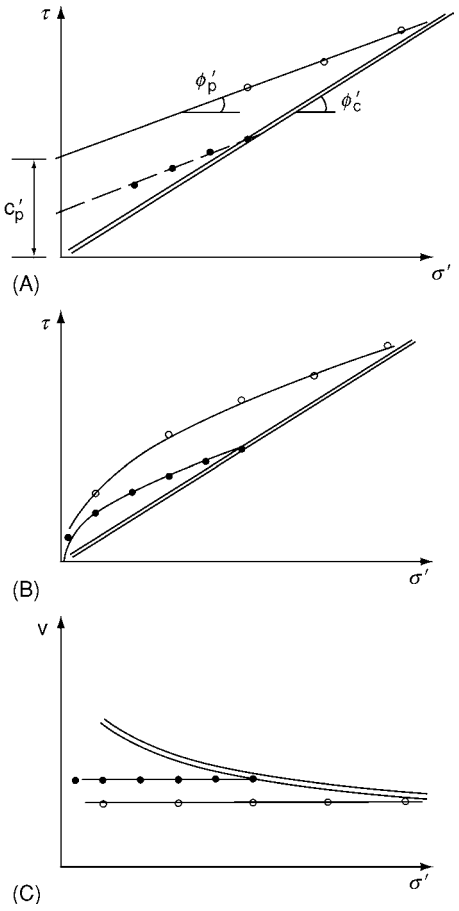


Figure 11 Peak strength.

Stiffness of Soil

Figure 5A shows non-linear isotropic unloading and reloading behaviour. Similar non-linear behaviour occurs during shearing, as shown in Figure 12A. The tangent shear modulus G′ is the gradient of the stress-strain curve given by

$$G' = \frac{d\tau}{d\gamma} \qquad [24]$$

At the start of shearing near the origin the shear modulus is G'_o and at failure the shear modulus is zero.

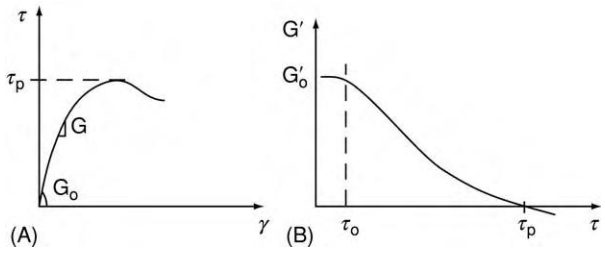


Figure 12 Stiffness and shear modulus.

Figure 12B shows the variation of shear modulus G′ with the progress of loading. There is a very small range up to a shear stress τ_o, in which G'_o is constant and the soil is linear, but over the remainder of loading the shear modulus decays with loading. For a particular soil the value of G'_o and the shear modulus at a particular strain, vary with the effective stress and with the state parameter.

For modest compression the bulk modulus, K′, and the one-dimensional compression modulus, M, both decay with normal stress in a manner similar to the decay of shear modulus with shear stress, shown in Figure 12B. At large compressive stresses the stiffness is the modulus corresponding to states on the NCL. At very large compressive stresses, the stiffness becomes very large as the specific volume approaches 1.0.

Consolidation

As soil is loaded or unloaded undrained, there are no volume changes but there are changes of pore pressure. These create excess pore pressures which are not in equilibrium with the surrounding pore pressures and so they dissipate with time. As they dissipate, under constant total stress, there are changes of effective stress which cause volume changes accompanied by drainage of water.

The basic theories for consolidation are for one-dimensional loading and drainage, illustrated in Figure 13A, in which all movements of soil and water are vertical. In practice this corresponds to conditions below a wide foundation or embankment.

Solutions for the rate of consolidation are given in terms of the degree of consolidation, U_t, and the time factor, T_v, given by

$$U_t = \frac{\rho_t}{\rho_\infty} \qquad [25]$$

$$T_v = \frac{c_v t}{H^2} \qquad [26]$$

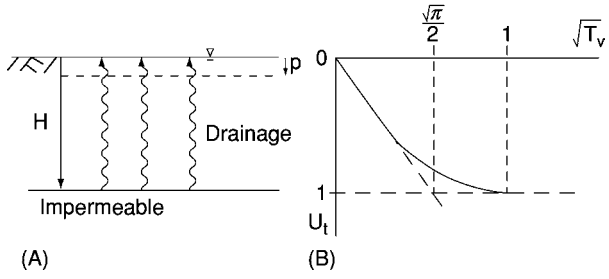


Figure 13 Consolidation.

where ρ_t is the settlement at time t, ρ_∞ is the settlement after a very long time, H is the length of the drainage path, and c_v is the coefficient of consolidation given by

$$c_v = \frac{Mk}{\gamma_w} \qquad [27]$$

where M is the one-dimensional modulus, k is the coefficient of permeability, and γ_w is the unit weight of water. The relationship between degree of consolidation and time factor is shown in Figure 13B.

The rate of consolidation depends on soil characteristics of stiffness and permeability and also on the geometry of the consolidating layer. This is given by the drainage path length H which is the greatest distance water must move to reach a drainage layer. Consolidation times can be significantly reduced by installing drains into the ground to reduce H.

Consolidation is the principal cause of the settlement of foundations and embankments on clays long after construction is complete.

Normalization and a State Boundary Surface

Figure 14A and B shows some different soil states. There are peak states corresponding to two different specific volumes; these are the same as those shown in Figure 11. There are paths for shearing of normally consolidated loose samples: path LD is for drained shearing and path LU is for undrained shearing.

These soil states involve three parameters, shear stress τ, normal stress σ', and specific volume v. Soil states can be represented by a three-dimensional surface using these axes. They may be represented on a two-dimensional graph using an appropriate normalizing procedure. There are several possibilities and one is to divide the shear and normal stresses by the equivalent stress σ'_e, shown in Figure 6A.

Figure 13C shows the states normalized by the equivalent stress. The NCL and the CSL reduce to single points. The peak states fall on a unique curve. The state paths for drained and undrained shearing of normally consolidated samples fall on a unique curve. The full curve represents a boundary to all possible states, known as a state boundary surface.

The concept of a state boundary surface is employed in advanced soil mechanics theories to develop complete constitutive relationships for soils. The surface is taken to be a yield surface and as a plastic potential surface from which plastic strains are determined. For states inside the boundary surface, the behaviour is taken to be elastic. One such theory is known as Cam Clay, for which the state boundary surface is represented by a logarithmic spiral curve.

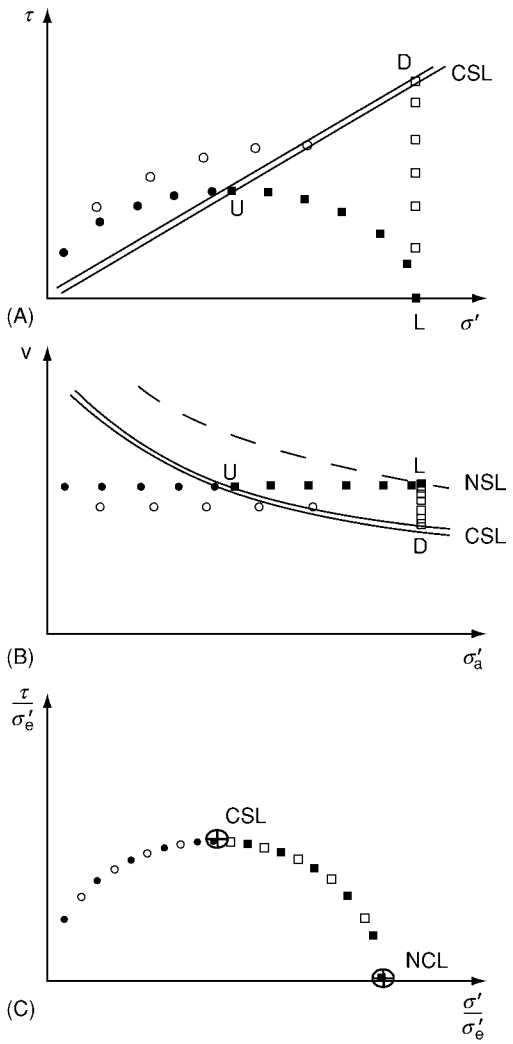


Figure 14 A state boundary surface.

Applications

The simple theories presented above for granular materials form the basis for analysis and design of engineering works which interact with the ground such as foundations, slopes, tunnels and retaining walls.

The basic theories for the mechanical behaviour granular materials are applicable equally to coarse-grained soils (sands and gravels) and fine-grained soils (clays). The principle factor to consider is the relative rate of loading and drainage. For routine analysis a particular case must be taken to be either fully drained or fully undrained.

If the soil is assumed to be fully drained, pore pressures can be determined and effective stresses calculated. Analyses are then carried out using *effective stresses* with effective stress strength and stiffness parameters. If the soil is assumed to be undrained, there are no changes in volume but there are changes in pore pressure which cannot be easily determined. In this case analyses have to be carried out using *total stresses* with undrained strength and stiffness parameters.

The critical state strength should be used to investigate ultimate failures. The peak strengths, with appropriate factors, should be used to investigate designs which are required to limit movements.

Simple analyses of foundation settlement are often carried out assuming one-dimensional conditions using the one-dimensional modulus, M, or using simple elastic theories using a shear modulus, G, and a bulk modulus, K. In all cases, it is necessary to take account of non-linear stress-strain behaviour and the appropriate drainage conditions. Simple analyses of rate of settlement due to consolidation can only be carried out assuming one-dimensional conditions.

The advanced soil mechanics theories, such as Cam Clay, are not used in simple analysis and design except for extremely simplified cases. Instead they form the basis for analyses using finite element or other comparable numerical methods.

See Also

Engineering Geology: Liquefaction; Made Ground; Problematic Soils; Subsidence. **Soils:** Modern; Palaeosols.

Further Reading

Atkinson JH (1993) *The Mechanics of Soils and Foundations*. London: McGraw Hill.

Goodman RE (1999) *Karl Terzaghi: the Engineer as Artist*. American Society of Engineering Press, Reston, Virginia.

Heyman J (1972) *Coulomb's Memoir on Statics*. Cambridge: Cambridge University Press.

Lancellotta R (1995) *Geotechnical Engineering*. Balkema, Rotterdam.

Muir Wood DM (1990) *Soil Behaviour and Critical State Soil mechanics*. Cambridge: Cambridge University Press.

Powrie W (2004). *Soil Mechanics*, 2nd edn. Spon Press: London.

Schofield AN and Wroth CP (1968) *Critical State Soil Mechanics*. McGraw Hill.

SOILS

Contents

Modern

G J Retallack, University of Oregon, Eugene, OR, USA

Introduction

There are many soil-forming processes, which in varying combinations create the large array of soils forming at the surface of the Earth. The study of soils is aided by the observation that soil-forming processes are slow and seldom go to completion. The parent materials of soils are modified over thousands of years by physical, chemical, and biological influences. However, few of these processes can be observed directly. Podzolization is one of the few soil-forming processes that is rapid enough to be recreated in the laboratory. Soil-forming processes that operate over thousands of years are studied using a space-for-time strategy (that is, studying soils of differing ages that are subject to the same soil-forming regime). A set of soils of different ages with comparable climates, vegetation, topographical positions, and parent materials is called a chronosequence (Figure 1). Mathematical relationships between the development of particular soil features and time are called chronofunctions, and include the increased clayeyness produced by the soil-forming process of lessivage (Figure 2). While specifying the rate and progress of soil formation, chronofunctions can also be used to infer the ages of landscapes from undated soils by comparison with dated soils. Such estimates of soil age can be important in the study of the neotectonic deformation of landscapes and their suitability for long-term installations such as dams and nuclear power plants. Soil fertility also varies with soil age, and chronofunctions can guide agricultural use and rehabilitation of soils.

Soil-forming processes vary not only with time but also with parent materials, topographical relief, vegetation, and climate. For example, the fragments of volcanic glass in certain kinds of air-fall tuff are distinct from the minerals of most soils, and they bestow high fertility and low bulk density on some volcanic soils (the process of andisolization). Waterlogging in low-lying parts of the landscape prevents the rusting of iron minerals and imparts a grey-green colour to the soil (the process of gleization). Leachates from highly acidic vegetation, such as pine forest, create soils in which clays are destroyed but quartz and haematite accumulate (the process of podzolization). Finally, climate is also an important factor in

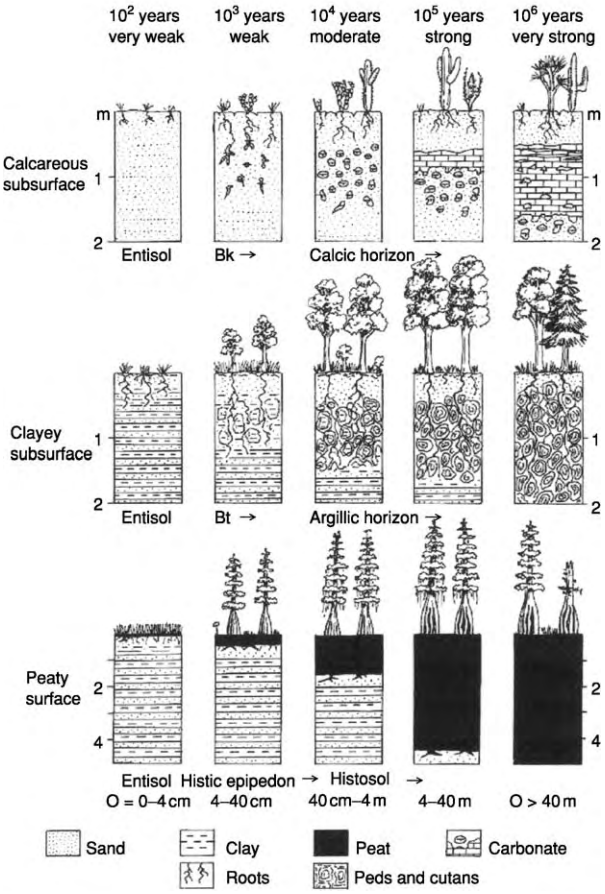


Figure 1 Soil development stages involving progressive calci fication (top), lessivage (middle), and paludization (bottom). Re produced with permission from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

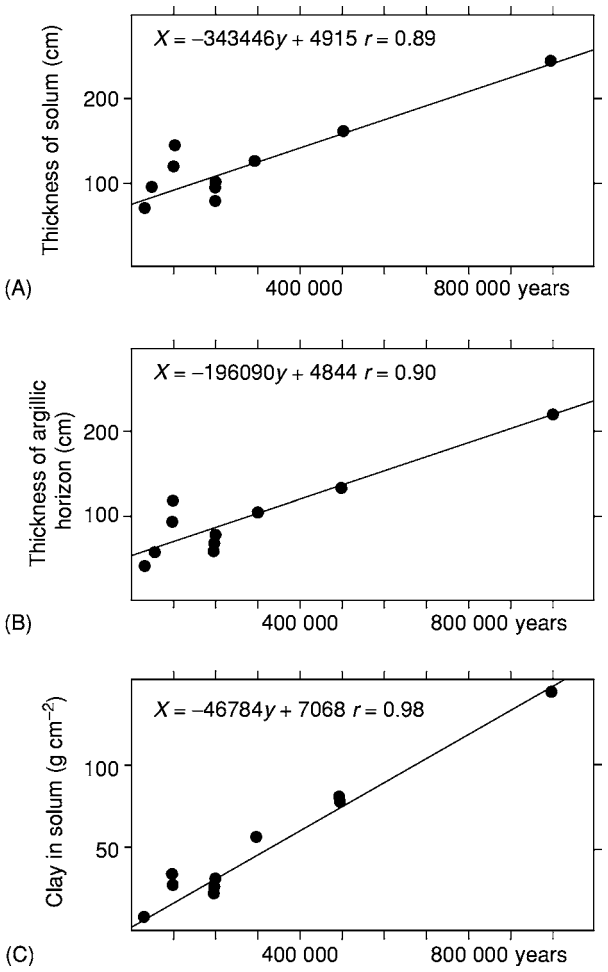


Figure 2 Chronofunctions for the progress of lessivage in soils of the Coastal Plain and Piedmont of south eastern USA over time: (A) solum thickness; (B) thickness of the argillic horizon; and (C) the amount of clay in the solum. The solum is the A and B horizons; the argillic horizon is the Bt horizon; and the total profile is the A, B, and C horizons as defined in **Table 2**. Reproduced with permis sion from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell, using data from Markewich HW, Pavich MJ, and Buell GR (1990) Contrasting soils and landscapes of the Piedmont and Coastal Plain, eastern United States. *Geomorphology* 3: 417 447.

soil-forming processes, encouraging deeper and more thorough weathering in wetter and warmer climates (Figure 3).

The study of soil-forming processes has informed both soil taxonomy (Table 1) and soil-profile terminology (Table 2). The following outlines of soil-forming processes are presented in the order in which they would be encountered from warm wetlands to cold arid lands.

Gleization

Gleying or gleization is a process that produces and maintains unoxidized minerals in soils and is a term derived from a Russian term for the grey clay of swamps and bogs. Waterlogged peat-covered stagnant groundwaters allow the preservation of ferrous iron in clay minerals, such as grey smectite, carbonates, such as the siderite of freshwater bogs, and sulphides, such as the pyrite of mangrove swamps and salt marshes. In normally drained soils these minerals rust to produce red and brown clays, hydroxides such as goethite, and oxides such as haematite (Table 3). Goethite and haematite also form within gleyed soils when a short-term depression of the water table allows the atmospheric penetration of oxygen. Despite these red nodules and concretions, the dominant colour of gleyed soils is bluish or greenish grey (Figure 4).

Paludization

Paludization is literally ponding, but a pond would not be commonly understood as a soil. Paludization is soil flooding that is tolerated by swamp trees but not by most soil decomposers. Paludization is thus an accumulation of undecayed plant debris as peat in the waterlogged surface layer (O horizon of Table 2) of Histosols (Table 1). This process requires a balance between plant production and decomposition. If ponding is intermittent and the soil is moderately oxidized, usually because of a low subsidence rate, then fungal and other decay prevents the accumulation of plant debris. If, on the other hand, ponding is too deep or prolonged, because of high subsidence rates, then soil stagnation kills the roots of woody plants, thus cutting off the supply of vegetation for further peat accumulation. As swamp forests die from anoxia at the roots, peaty soils become overwhelmed by lakes, bayous, or lagoons. The rate of subsidence and accumulation of woody peats is generally between 0.5 mm and 1 mm per year, because of constraints on the growth rate of woody plants in low-fertility peaty substrates and the depth of penetration of air and decomposers within woody peats. Herbaceous plants and mosses are less constrained in their growth rates and form domed peats that rise well above the water table. Peat accumulation in both cases involves addition from the top, in the same way as sediment accumulation, and thus differs from soil-forming processes that modify pre-existing materials. The progress of paludization leads to progressively thicker peat (Figure 1).

Podzolization

Podzol in its original Russian means ‘under ash’ and refers to the light-coloured quartz-rich (E) horizon immediately beneath the humus. Many podzolic

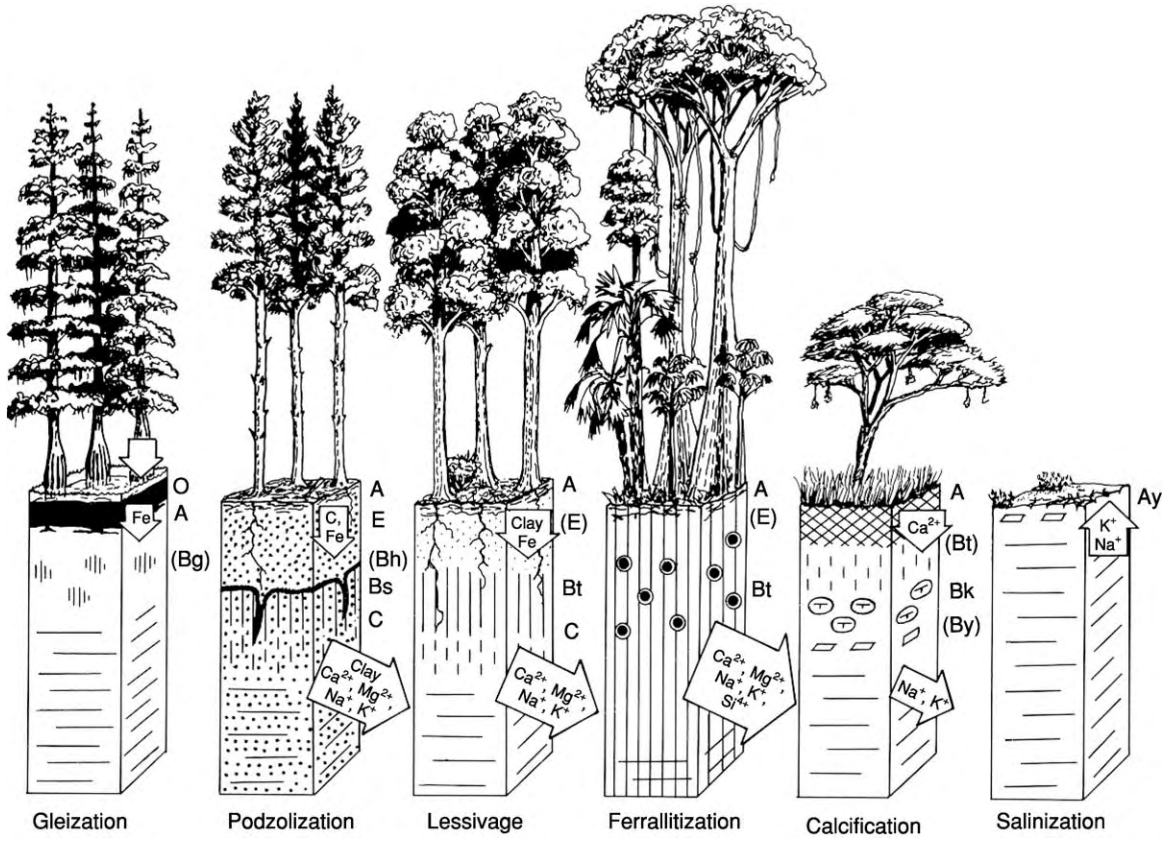


Figure 3 Selected common soil forming processes arranged along a climatic gradient. The ecosystems depicted are (from left to right): bald cypress swamp, spruce forest, oak forest, tropical rain forest, *Acacia* savannah, and saltbush scrub. Horizon nomenclature is described in **Table 1**, and the large arrows indicate the movement of key soil components. Reproduced with permission from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Table 1 Outline of soil taxonomy

Order	*Description*
Entisol	Very weakly developed soil with surface rooting and litter (A horizon) over weathered (C horizon) sediment with relict bedding or weathered igneous or metamorphic rock with relict crystals
Inceptisol	Weakly developed soil with surface rooting and litter (A horizon) over somewhat weathered (Bw horizon) clayey (Bt horizon) or calcareous (Bk horizon) subsurface
Andisol	Soil composed of volcanic ash with low bulk density and high fertility
Histosol	Peat (O horizon) over rooted grey clay (A horizon)
Spodosol	Quartz rich clay poor soil with bleached subsurface (E horizon) above a red black iron aluminia organic cemented zone (Bs horizon)
Vertisol	Very clayey profile with common swelling clay (smectite), laterally variable thickness of surface (A horizon), and strongly slickensided subsurface (Bt horizon)
Mollisol	Grassland soil with thick crumb textured carbon rich surface (A horizon)
Gelisol	Permafrost soil with frost heave and other periglacial features
Aridisol	Desert soil with a shallow subsurface accumulation of pedogenic carbonate (Bk horizon) and soluble salts (By horizon)
Alfisol	Fertile forest soil with clay enriched subsurface (Bt horizon) and high amounts of Mg, Ca, Na, and K
Ultisol	Infertile forest soil with clay enriched subsurface (Bt horizon) and low amounts of Mg, Ca, Na, and K
Oxisol	Deeply weathered tropical soil, often highly ferruginous and aluminous, but with very low amounts of Mg, Ca, Na, and K

For technical limits of soil orders see Soil Survey Staff (2000) *Keys to Soil Taxonomy*. Blacksburg: Pocahontas Press.

soils are now included in the USDA (United States Department of Agriculture) soil order Spodosol (Table 1), which refers to the red, brown, or black (Bs) horizon below the light coloured near-surface layer. This striking differentiation between white near-surface and dark subsurface horizons is created by podzolization, which effectively leaches iron and organic matter from the upper horizons and

reprecipitates them in a lower horizon. The resulting effect is as striking as the chromatographic separation of organic compounds, and podzolization is one of the few soil-forming processes that is rapid enough to have been recreated under controlled laboratory conditions. The process is particularly helped by highly acidic soil solutions (with a pH of less than 4) in well-drained soils of humid climates under acid-generating litter such as that of conifer forest (Figure 3). Under highly acidic conditions clay minerals are destroyed, so Podzols and Spodosols usually have a sandy texture.

Ferrallitization

The term ferrallitization is derived from iron (Fe) and aluminium (Al), which become enriched in minerals such as haematite, kaolinite, and gibbsite during intense weathering of well-drained tropical soils such as Oxisols (Figure 3). Much of the loss of major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}) by hydrolysis requires carbonic acid derived from the carbon dioxide of soil respiration, yet the soil pH remains above 4, so that clays are not destroyed. Mitigation of acidity and deep oxidation of these soils may in part be due to the activity of termites and tropical trees, as ferrallitization is primarily found in soils under tropical rainforest. The broad-leaved trees of tropical rainforests produce less acidic litter than conifers and other plants, and litter decomposition rates are high on humid and warm forest floors. Furthermore, ferrallitic soils commonly contain abundant microscopic (125–750 μm) spherical to ovoid pellets of oxidized clay, like the faecal and oral pellets of termites. Some ferrallitic soils appear to have passed through the guts of termites many times. Termites are unique in having extremely alkaline midguts (with a pH of 11–12.5).

Table 2 Standard acronyms for soil horizon description

Acronym	*Description*
O	Surface accumulation of peaty organic matter
A	Surface horizon of mixed organic and mineral material
E	Subsurface horizon rich in weather resistant minerals, e.g. quartz
Bt	Subsurface horizon enriched in washed in clay
Bs	Subsurface horizon enriched in organic matter, or iron or aluminium oxides
Bk	Subsurface horizon enriched in pedogenic carbonate
Bn	Subsurface horizon with domed columnar structure and sodium clays
By	Subsurface horizon enriched in salts such as gypsum and halite
Bo	Subsurface horizon deeply depleted of Ca, Mg, Na, and K
Bw	Subsurface horizon mildly oxidized and little weathered
C	Mildly weathered transitional horizon between soil and substrate
R	Unweathered bedrock

For technical limits of soil orders see Soil Survey Staff (2000) *Keys to Soil Taxonomy*. Blacksburg: Pocahontas Press.

Biocycling

Biocycling includes a variety of processes in which nutrient elements are exchanged by soil biota without reincorporation into soil minerals. In tropical soils such as Oxisols (Table 1) this is a very efficient process in which the decay of leaves and wood is orchestrated by waves of bacteria, fungi, ants, and termites, which excrete and die to feed a copious network of epiphytes and tree roots. Effective biocycling explains the spectacular luxuriance of tropical-rainforest ecosystems despite their extremely nutrient-depleted and humus-poor mineral soils (Oxisols). Comparable mechanisms operate in swamp forests growing in peat (Histosols), which also experience severe mineral-nutrient limitations. These mineral nutrients include the major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}), but these are seldom as limiting as nitrogen, which is derived largely from the microbial recombination of atmospheric nitrogen, or phosphorous, which is derived largely from the weathering of apatite. Biocycling of

Table 3 Common kinds of chemical reactions during weathering

Reaction	*Example*
Hydrolysis	$2NaAlSi_3O_8 + 2CO_2 + 11H_2O \rightarrow Al_2Si_2O_5(OH_4) + 2Na^{+} + 2HCO_3^{-} + 4H_4SiO_4$ albite + carbon dioxide + water → kaolinite + sodium ions + bicarbonate ions + silicic acid
Oxidation	$2Fe^{3+} + 4HCO_3^{-} + 1/2O_2 + 4H_2O \rightarrow Fe_2O_3 + 4CO_2 + 6H_2O$ ferrous ions + bicarbonate ions + oxygen + water → haematite + carbon dioxide + water
Dehydration	$2FeOOH \rightarrow Fe_2O_3 + H_2O$ goethite → haematite + water
Dissolution	$CaCO_3 + CO_2 + H_2O \rightarrow Ca^{2+} + 2HCO_3^{-}$ calcite + carbon dioxide + water → calcium ions + bicarbonate ions

Figure 4 Red and brown mottles of goethite in the upper part of the profile and dark stains of pyrite formed by gleization in the lower part of the profile of a gleyed Inceptisol, excavated as a soil column from a salt marsh on Sapelo Island, Georgia, USA. Hammer handle is 25 cm long.

Figure 5 Light brown near surface (E) and dark brown subsur face (Bt) horizons of an Alfisol produced by lessivage near Killini, Greece. Hammer handle upper right is 25 cm long.

nitrogen is especially important during the early development of soils such as Entisols and Inceptisols, which are developed over decades or centuries. Biocycling of phosphorous becomes increasingly important in very old soils such as oxisols and ultisols, which are depleted in apatite over thousands or millions of years.

Lessivage

Lessivage or argilluviation is the process of clay accumulation within a subsurface (Bt or argillic) soil horizon (**Figures 1, 2** and **3**). This is a common and widespread soil-forming process in the forested soils of humid climates, particularly Alfisols and Ultisols (**Figure 5**). The clay is primarily derived from a hydrolytic weathering reaction in which clays remain as a residuum and dissolved cations are removed in groundwater during the incongruent dissolution of feldspars and other minerals by carbonic acid (**Table 3**). Driving the reaction are abundant rainfall and high soil respiration rates fuelled by high primary productivity. Clay forms rinds around mineral grains of the sedimentary, igneous, or metamorphic parent material, but is also washed down cracks in the soil created by desiccation, roots, and burrows. This washed in or illuvial clay has a very distinctive banded appearance, which is obvious in petrographic thin sections. The clay is not washed any lower than the water table, where percolating rainwater ponds. Clay is less common near the surface of the soil, where unweathered grains are added by wind and water, and grains are leached of clay by plant acids. The net effect is a subsurface clayey horizon that becomes more clayey over time (**Figures 1, 2** and **3**).

Lixiviation

Lixiviation is a process of leaching of major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}) from soil minerals and their loss from the soil in groundwater. Lixiviation is a component of ferrallitization, podzolization, and lessivage, and represents the progress of the hydrolysis chemical reaction, in which hydronium ions (H^{+}) of a weak acid (usually carbonic acid) displace cations into solution and thus convert primary minerals such a feldspars into soil minerals such as clays (**Table 3**). The term lixiviation is primarily used to describe the beginnings of this process in soils such as Entisols and Inceptisols that have developed over

only decades or centuries. Such young soils have not yet acquired the distinctive deeply weathered and oxidized horizons produced by ferrallitization in Oxisols, the distinctive leached (E) and enriched (Bs) horizons produced by podzolization in Spodosols, or the distinctive clay-enriched subsurface (Bt) horizons produced by lessivage in Alfisols and Ultisols.

Melanization

Melanization is a process of soil darkening due to the addition of soil organic matter. The process is best known in Mollisols, the fertile dark crumb-textured soils of grasslands (Figure 6). In these soils melanization is largely a product of the activities of grasses and earthworms. Earthworms produce faecal pellets rich in organic matter and nutrients such as carbonate. Earthworms also produce slime, which facilitates their passage through the soil. Root exudates from grasses are also added to soil crumbs. Many soils have dark humic near-surface horizons, but a peculiarity of grassland soils is that dark organic fertile crumb-textured soil extends to the base of the rooting zone, which can be more than a metre deep in soils under tall-grass prairie. Melanization also occurs in swamp and marsh soils (gleyed Inceptisols and Entisols), where the decay of humus is suppressed by poor oxidation and waterlogging. Unlike the alkaline crumb-textured melanized surface of grassland soils, the melanized surface of wetland soils is nutrient-poor, acidic, and has a massive to laminated fabric. Melanization is not usually applied to the precipitation of amorphous Fe–Mn oxides (birnessite) in gleyed soils, which can also produce dark soil. The creation of these Fe–Mn-stained (placic) horizons is a process of gleization rather than melanization.

Figure 6 Dark organic rich surface (mollic epipedon) of a Mollisol formed by melanization near Joliet, IL, USA. The shovel handle is 15 cm wide at the top.

Andisolization

Andisolization is the formation of fertile mineralogically amorphous low-density horizons within soils of volcanic ash (Andisols). Many volcanic ashes are composed largely of small angular fragments (shards) of volcanic glass. Unlike soil minerals such as feldspar, volcanic glass weathers, not to crystalline minerals such as clay, but to non-crystalline compounds such as imogolite. The loosely packed angular shards and colloidal weathering products create a soil of unusually low bulk density (1.0–1.5 g cm ^{3}, compared with 2.5–3.0 g cm ^{3} for most common minerals and rocks). Furthermore, these colloidal compounds contain plant-nutrient cations, and particularly phosphorous, in a form that is more readily available to plants than those of other kinds of soils dominated by crystalline minerals such as apatite. Andisolization is not sustainable for more than a few thousand years unless there are renewed inputs of volcanic glass, because glass and other colloids (such as imogolite) weather eventually to oxides and clay minerals.

Vertization

Vertization is the physical soil overturning and mixing by means of the shrink–swell behaviour of clays. It occurs mostly in Vertisols but also in Entisols, Inceptisols, Mollisols, and Alfisols. It is especially characteristic of soils rich in swelling clays (smectites), which swell when wet and shrink when dry. Also characteristic is a climate with a pronounced seasonal contrast in precipitation. During the wet season the clays swell and buckle under the pressure of their inflation. During the dry season they open up in a system of cracks, which are then partly filled by wall collapse. This fill exacerbates the buckling in the next wet season so that the soil develops ridges or mounds with intervening furrows or pits, called gilgai microtopography. In a soil pit, the cracks of mounded areas divide areas of festooned slickensides under the furrows and pits in a distinctive arrangement called mukkara structure (Figure 7). Vertization is mainly a phenomenon of semiarid to subhumid regions. Soils of arid regions are generally not sufficiently clayey, whereas soils of humid regions are generally too deeply weathered to contain abundant smectite and are also stabilized by massive plant and animal communities.

Figure 7 Gilgai microrelief (low to left, high to right) and its subsurface mukkara structure (festooned intersecting slicken sided cracks) produced by vertization in the Branyon clay soil, a Vertisol, near New Braunfels, TX, USA. Scale to left shows 50 cm and 100 cm; red and white bands on pole to right are 10 cm wide.

Anthrosolization

Anthrosolization is the alteration of soil by human use, such as buildings, roads, cesspits, garbage dumps, terracing, and ploughing. Archaeological ruins and artefacts are important clues to prior occupation of a site, but many sites also contain impressive amounts of mollusc shells and mammal and fish bones. A distinctive soil structure of subsoil pockets of laminated clay between large soil clods is produced by moldboard ploughs. The primitive or ard plough also tends to disrupt the natural crumb structure to a fixed depth (plough line). Phosphorous content is an indicator of human use. Many soils have trace amounts of phosphorus (10–20 ppm by weight), but occupation floors and long-used garden soils and middens have large amounts of phosphorous (1000–2000 ppm). Anthrosolization is locally common worldwide in cities and fields, both ancient and new, but is scattered and local in deserts, polar regions, and high mountains.

Calcification

Calcification is the accumulation of calcium and magnesium carbonates in the subsurface (Bk) horizons of soils (Figures 1 and 3). The carbonate is usually obvious, appearing as soft white filaments, hard white nodules, and thick white benches within the soil. Calcification is largely a soil-forming process of dry climatic regions, where evaporation exceeds precipitation. It is characteristic of Aridisols but is also found in some Mollisols, Andisols, Vertisols, Inceptisols, and Alfisols. The source of the carbonate is the soil respiration of roots, soil animals, and microorganisms. Calcification requires soil respiration at levels greater than those in hyperarid soils, where halite and gypsum formed by salinization prevail, and less than those in humid soils, where lessivage prevails. The source of the cations of calcium and magnesium, which create the soil minerals calcite and dolomite, respectively, is the weathering of soil minerals by hydrolysis (Table 3). Some of these cations originate from feldspars and other minerals of the parent material, but dry regions of calcification have open vegetation and are often dusty, so that carbonate and feldspar dust is an important source of cations. Dissolved cations from hydrolytic weathering are commonly lost downstream in the groundwater in humid regions, but in arid lands the water table is commonly much deeper than the soil profiles, which are seldom wet much beyond the depth of rooting. The subsurface zone of groundwater evaporation and absorption is where the wisps of soil carbonate form, then coalesce into nodules and, eventually, thick layers.

Solonization

Solonization is a process by which clays rich in soda are formed within the soils of dry climates (Aridisols), where the hydrolytic mobilization of major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}) is weak. Hydrolysis removes cations from soils by lixiation in humid climates, but in dry climates the acidity created by soil respiration after rain storms is sufficient to remove cations from minerals such as feldspar without leaching them from the profile. Solonized soils commonly contain carbonate nodules of dolomite or low-magnesium calcite, formed by calcification, as well as salts of halite and gypsum, formed by salinization. Solonized soils have illitic clays rich in potassium and smectite clays rich in sodium, and the progress of solonization can be assessed by measuring the pH (which is usually around 9–10), by chemical analysis, or by X-ray diffraction to determine the mineral composition. A field indicator of solonization is the presence of domed columnar peds that run through most of the subsurface (natric or Bn) horizon of the soil (Figure 8). The sodium-smectite clays of solonized soils have some shrink–swell capacity, meaning that they form prismatic cracks as the soil dries out and swelling or domed tops to the prisms when the soil is wet. Solonization is common around desert playa lakes and salinas and in coastal soils affected by saltwater spray.

Solodization

Solodization is intermediate between solonization and lessivage, and creates profiles with acidic-to-neutral near-surface horizons but alkaline subsurface

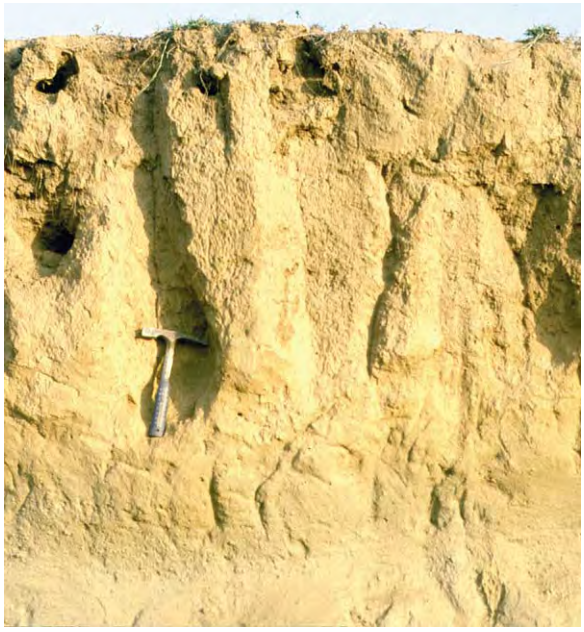


Figure 8 Domed columnar peds produced by solonization in an inceptisol near Narok, Kenya. Hammer handle is 25 cm.

horizons dominated by sodium-smectite. Solodized soils have domed columnar clayey peds in a subsurface (Bn) horizon, but these are sharply truncated by a granular leached (E) horizon. Solodization occurs in desert soils (Aridisols) with better vegetative cover and a more humid climate than solonized soils.

Salinization

Salinization is the precipitation of salts in soils (Figure 3) and is found mostly in desert soils (Aridisols). The most common salts are halite and gypsum, which can form either as clear crystals within soil cracks or as sand crystals that engulf the pre-existing soil matrix. Salts are easily dissolved by rain and so accumulate in regions where there is a marked excess of evaporation over precipitation, which is generally less than 300 mm per year. There is a strong relationship between mean annual precipitation and the depth of leaching of salts in soils. Salinized soils are sparsely vegetated or lack vegetation, and occur in playa lakes, sabkhas, and salinas. Although these are commonly regarded as depositional environments, they are significant soil environments as well.

Cryoturbation

Cryoturbation is the mixing of soils by the freezing and thawing of ground ice. The ice can form disseminated crystals, hair-like threads, thin bands, thick benches, or vertical cracks depending on the local climatic conditions. Soil mixing results from the expansion of water to ice during winter freezing and the relaxation of the deformation on summer melting. Ice-wedge polygons, for example, are wide polygonal cracks that are filled with ice in winter but can be filled with layered sediments in water during the summer in climates where the mean annual temperature is less than −4°C. Sand-wedge polygons form in colder climates where the mean annual temperature is less than −12°C; here, summer melting of ice is limited and sediment fills cracks between the ice and soil in a series of near vertical layers.

Conclusion

Soil-forming processes are varied and complex, and our understanding of them guides the classification, description, and management of soils. The processes are also of interest in simplifying the vast array of chemical reactions, biological processes, and physical effects that create soil. Some processes are more common and widespread than others. Lixiviation and its underlying hydrolysis chemical reaction is perhaps the most important weathering process on Earth, affecting geomorphology, sedimentation, ocean chemistry, and climate. Other processes are restricted to more specific climatic, biotic, geomorphological, geological, and temporal environments, but are no less important in their local environments.

Glossary

Alfisol A fertile forested soil with subsurface enrichment of clay.
Andisol A volcanic-ash soil.
Andisolization A soil-forming process that creates low-density non-crystalline fertile soil from volcanic ash.
Anthropic epipedon A soil surface modified by human use.
Anthrosolization A soil-forming process involving modification by human activities.
Argillic horizon A subsurface horizon of soil enriched in clay.
Argilluviation A soil-forming process that involves creating clay and washing it into a subsurface clayey horizon.
Aridisol A soil of arid regions, usually containing carbonate nodules.
Biocycling The recycling of nutrient elements by biota.
Birnessite A non-crystalline mixture of iron and manganese oxides.
Entisol A very weakly developed soil.

Ferrallitization A soil-forming process involving intense weathering that removes most elements other than iron and aluminium.
Gelisol A soil of permafrost regions, usually containing ground ice.
Gibbsite An aluminium hydroxide mineral ($Al(OH)_3$).
Gilgai A soil microrelief consisting of ridges or mounds alternating with furrows or pits.
Gleization A soil-forming process involving chemical reduction of the soil due to waterlogging.
Halite A salt mineral (NaCl).
Haematite An iron oxide mineral (Fe_2O_3).
Imogolite A colloidal weathering product of volcanic-ash soils.
Inceptisol A weakly developed soil.
Lessivage A soil-forming process that creates clay and washes it into a subsurface clayey horizon.
Lixiviation A soil-forming process that involves leaching nutrient cations from the soil.
Melanization A soil-forming process that involves darkening the soil with organic matter.
Mollic epipedon A humic fertile crumb-textured soil surface typical of grassland soils.
Mollisol A grassland soil with a humic fertile crumb-textured surface.
Mukkara A soil structure consisting of festooned and slickensided cracks between uplifted parts of the soil; the subsurface structures below gilgai microrelief.
Natric horizon A subsurface horizon of soil enriched in sodium-clay.
Oxisol A deeply weathered soil of tropical humid regions.
Paludization A soil-forming process involving peat accumulation in waterlogged soils.
Ped A clod, a unit of soil structure.
Placic horizon Iron- and manganese-stained bands and nodules in soils.
Plaggen epipedon A ploughed surface horizon of soils.
Podzol A sandy soil with a bleached near-surface horizon.
Podzolization A soil-forming process in which acid leaching creates a bleached sandy upper horizon and an iron- or organic-rich subsurface horizon.
Siderite An iron carbonate mineral ($FeCO_3$).
Solonization A soil-forming process that creates soda-rich clays and domed columnar peds in arid regions.
Spodosol A sandy clay-poor soil with an iron- or organic-rich subsurface horizon.
Ultisol A deeply weathered forest soil with subsurface enrichment in clay.
Umbric epipedon A humic acidic clayey massive-to-laminar soil surface found in wetland soils.
Vertisol Swelling clay soil.
Vertization A soil-forming process involving deformation and mixing due to the shrink–swell behaviour of clay during drying and wetting cycles.

See Also

Carbon Cycle. **Clay Minerals**. **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills. **Sedimentary Environments:** Deltas; Deserts. **Sedimentary Processes:** Glaciers. **Soils:** Palaeosols. **Weathering**.

Further Reading

Bockheim JG and Gennadiyev AN (2000) The role of soil forming processes in the definition of taxa in soil tax onomy. *Geoderma* 95: 53 72.
Bohn H, McNeal B, and O'Connor G (1985) *Soil Chemis try*. New York: Wiley.
Eisenbeis G and Wichard H (1987) *Atlas on the Biology of Soil Arthropods*. Berlin: Springer.
Jenny H (1941) *Factors of Soil Formation*. New York: McGraw Hill.
Lündstrom US, Van Breeman N, and Bain D (2000) The podzolization process: a review. *Geoderma* 94: 91 107.
McFadden LD, Amundson RG, and Chadwick OA (1991) Numerical modelling, chemical and isotopic studies of carbonate accumulation in arid soils. In: Nettleton WD (ed.) *Occurrence, Characteristics and Genesis of Car bonate Gypsum and Silica Accumulations in Soils*, pp. 17 35. Special Publication 26. Madison: Soil Science Society of America.
Markewich HW, Pavich MJ, and Buell GR (1990) Con trasting soils and landscapes of the Piedmont and Coa stal Plain, eastern United States. *Geomorphology* 3: 417 447.
Marshall TJ, Holmes JW, and Rose CW (1996) *Soil Physics*. Cambridge: Cambridge University Press.
Paton TR, Humphreys GS, and Mitchell PB (1995) *Soils: A New Global View*. London: UCL Press.
Retallack GJ (1997) *A Colour Guide to Paleosols*. Chiches ter: Wiley.
Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.
Richter DD and Markewitz D (2001) *Understanding Soil Change*. Cambridge: Cambridge University Press.
Sanford RI (1987) Apogeotropic roots in an Amazon rain forest. *Science* 235: 1062 1064.
Soil Survey Staff (2000) *Keys to Soil Taxonomy*. Blacks burg: Pocahontas Press.
Washburn AL (1980) *Geocryology*. New York: Wiley.

Palaeosols

G J Retallack, University of Oregon, Eugene, OR, USA

Introduction

Palaeosols are ancient soils, formed on landscapes of the past. Most palaeosols have been buried in the sedimentary record, covered by flood debris, landslides, volcanic ash, or lava (Figure 1). Some palaeosols, however, are still at the land surface but are no longer forming in the same way that they did under different climates and vegetation in the past. Climate and vegetation change on a variety of time-scales, and the term relict palaeosol for profiles still at the surface should be used only for such distinct soil materials as laterites among non-lateritic suites of soils (Figure 2). Thus, not all palaeosols are fossil soils or buried soils.

An alternative spelling of paleosol has been adopted by the International Quaternary Association. Other terms such as pedoderm and geosol refer to whole landscapes of buried soils. These soil stratigraphical units are named and mapped in order to establish stratigraphical levels. The terms pedotype and soil facies are more or less equivalent and are used to refer to individual palaeosol types preserved within ancient buried landscapes. These terms are used to distinguish one type of palaeosol from another in environmental interpretations of palaeosols. Pedolith, or soil sediment, describes a sediment, as indicated by bedding and other sedimentary features, with distinctive soil clasts, such as ferruginous concretions. Pedoliths are uncommon in sedimentary sequences, because soils are readily eroded to their constituent mineral grains, which retain few distinctive soil microfabrics.

Recognition of Palaeosols

Palaeosols buried in sedimentary and volcaniclastic sequences can be difficult to distinguish from enclosing sediments, tuffs, or lavas and were not widely recognized before about 20 years ago. Three features of palaeosols in particular aid their identification: root traces, soil horizons, and soil structure.

Soil is often defined as the medium of plant growth. Geological and engineering definitions of soil are broader, but fossilized roots and traces of their former paths through the soil are universally accepted as diagnostic of palaeosols. Not all palaeosol root traces are permineralized or compressed original organic matter: some are tortuous infillings of clay with discoloured haloes or mineralized alteration (Figure 3). Both fossilized roots and root traces show the downward tapering and branching of roots. Soils also contain fossil burrows, but these are usually more sparsely branched and parallel-sided than root traces. The distinction between burrows and roots can be blurred in cases where soil animals feed on roots and where roots find an easier passage through the soft fill of burrows. For very old rocks, predating the Early Devonian evolution of roots, the criterion of root traces is of no use in identifying palaeosols.

Figure 1 The subtle colour banding in these cliffs is the result of a sequence of 87 Eocene and Oligocene palaeosols in 143 m of nonmarine silty claystones exposed in the Pinnacles area of Badlands National Park, South Dakota, USA.

Figure 2 The red rock exposures to the left on the beach are a lateritic palaeosol of Middle Miocene age. Even though these horizons are at the surface, they are considered to be palaeosols because soil horizons of this type are not currently forming in this area. The red rock in the background is a sequence of Early Triassic palaeosols in Bald Hill Claystone, near Long Reef, New South Wales, Australia.

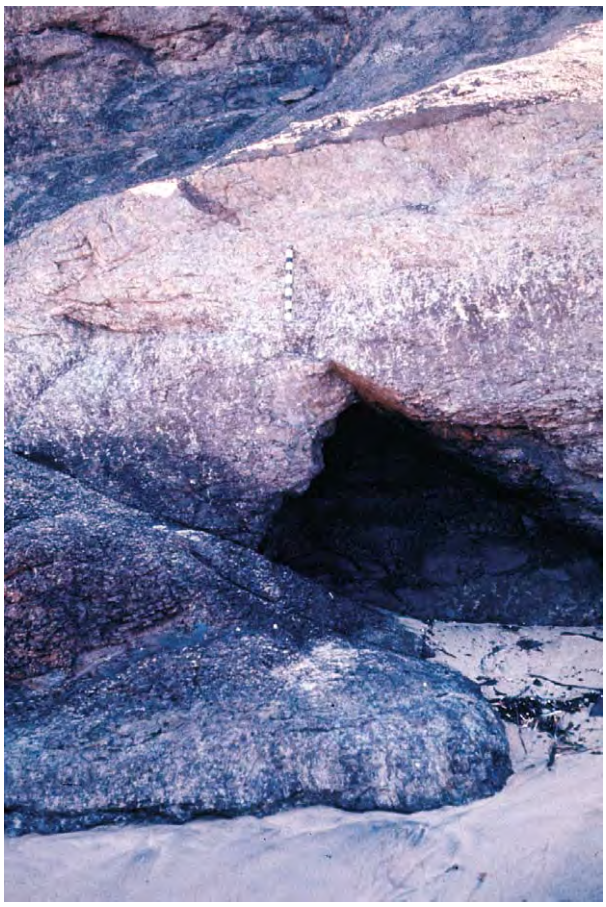


Figure 3 The sharply truncated top and abundant drab haloed root traces (A horizon) petering out downwards into red claystone (Bt horizon) are soil horizons of a palaeosol (Long Reef clay palaeosol, Early Triassic, Bald Hill Claystone, near Long Reef, New South Wales, Australia).

Palaeosols also have recognizable soil horizons, which differ from most kinds of sedimentary bedding in their diffuse contacts downwards from the sharp upper truncation of the palaeosol at the former land surface. Palaeosol horizons, like soil horizons, are seldom more than a metre thick and tend to follow one of a few set patterns. Subsurface layers enriched in clay are called Bt horizons in the shorthand of soil science (Figure 3). Unlike a soil, in which clayeyness can be gauged by resistance to the shovel or plasticity between the fingers, clayeyness in palaeosols that have been turned to rock by burial compaction must be evaluated by petrographic, X-ray, or geochemical techniques. Subsurface layers enriched in pedogenic micrite are called Bk horizons in the shorthand of soil science and are generally composed of hard calcareous nodules or benches in both soils and lithified palaeosols (Figure 4).

A final distinctive feature of palaeosols is soil structure, which varies in its degree of expression and

Figure 4 Two successive palaeosols overlain sharply by volcanic grits show crumb structured organic surfaces (A horizon) over calcareous nodule studded subsurfaces (Bk horizon). In the upper right corner is a comparable modern soil (Middle Miocene fossil quarry near Fort Ternan, Kenya).

replaces sedimentary structures such as bedding planes and ripple marks, metamorphic structures such as schistosity and porphyoblasts, and igneous structures such as crystal outlines and columnar jointing. Because they lack such familiar geological structures, palaeosols are commonly described as featureless, massive, hackly, or jointed. Palaeosols, like soils, have distinctive systems of cracks and clods. The technical term for a natural soil clod is a ped, which can be crumb, granular, blocky, or columnar, among other shapes. Peds are bounded by open cracks in a soil and by surfaces that are modified by plastering over with clay, by rusting, or by other alterations. These irregular altered surfaces are called cutans, and they are vital in recognizing soil peds in palaeosols that have been lithified so that the original cracks are crushed. The rounded 3–4 mm ellipsoidal crumb peds of grassland soils and palaeosols (Figure 4) are quite distinct from the angular blocky peds of forest palaeosols (Figure 3). Common cutans in soils and palaeosols include rusty alteration rinds (ferrans) and laminated coatings of washed-in clay (argillans). Cutans and other features of lithified palaeosols are best studied in petrographic thin sections and by electron microprobing and scanning electron microscopy. Some petrographic fabrics, such as the streaky birefringence of soil clays when viewed under crossed Nicols or sepic plasmic fabric, are diagnostic of soils and palaeosols.

Alteration of Soils after Burial

Palaeosols are seldom exactly like soils because of alteration after burial or exposure to additional weathering, and this can compromise their interpretation and identification with modern soils. Palaeosols,

like sediments, can be altered by a wide array of burial processes: cementation with carbonate, haematite, or silica; compaction due to pressure or overburden; thermal maturation of organic matter; and metamorphic recrystallization and partial melting. These high-pressure and high-temperature alterations of palaeosols are not as difficult to disentangle from processes of original soil formation as are three common early modifications: burial decomposition, burial reddening, and burial gleization.

Some soils are buried rapidly by chemically reducing swamps or thick lava flows, preserving most of their organic matter. In contrast, many palaeosols are covered thinly by floodborne silt or colluvium, and their buried organic matter is then decomposed by aerobic bacteria and fungi deep within the newly forming soil of the palaeosol sequence. For this reason many palaeosols have much less organic carbon (fractions of a weight per cent) than comparable modern soils (usually 5–10% by weight of carbon at the surface). Thus palaeosol A horizons are seldom as dark as soil surface horizons, and must be inferred from the abundance of roots rather than from colour and carbon content.

Soils vary considerably in their degree of redness, but most palaeosols are red to reddish brown from haematite (iron oxide) or occasionally yellowish brown from goethite (iron hydroxide). Soils become redder from the poles to the tropics, from moderately drained to well drained sites, and with increasing time for development, as iron hydroxides are dehydrated to oxides. The dehydration of iron hydroxides continues with the burial of soils, so that red palaeosols are not necessarily tropical, unusually well drained, or especially well developed.

In river-valley and coastal sedimentary sequences with abundant palaeosols, formerly well-drained soils can find themselves subsiding below the water table with root traces and humus largely intact. Burial gleization is a process in which organic matter is used by microbes as a fuel for the chemical reduction of yellow and red iron oxides and hydroxides. Comparable processes of biologically induced chemical reduction are common in swamp soils, but superimposition of this process on the organic parts of formerly well-drained soils produces striking effects in some palaeosols. The whole A horizon is turned grey, with grey haloes extending outwards from individual roots, which diminish in abundance down the profile (Figure 3). Burial gleization is especially suspected when the lower parts of the profile are highly oxidized and have deeply penetrating roots, as in well-drained soils, and when there is no pronounced clayey layer that would perch a water table within the soil.

The combined effect of burial decomposition, dehydration, and gleization can completely change the appearance of a soil. The gaudy grey-green Triassic palaeosol shown in Figure 3, for example, was probably modified by all three processes from an originally dark brown over reddish brown forest soil.

Palaeosols and Palaeoclimate

Many palaeosols and soils bear clear marks of the climatic regime in which they formed. The Berkeley soil scientist Hans Jenny quantified the role of climate in soil formation by proposing a space-for-climate strategy. What was needed was a carefully selected group of soils, or climosequence, that varied in climate of formation but were comparable in vegetation, parent material, topographical setting and time for formation. He noted that mean annual rainfall and the depth in the profile to calcareous nodules decline from St Louis west to Colorado Springs, in the mid-western USA, but that temperatures and seasonality at these locations are comparable. Also common to all these soils is grassy vegetation on postglacial loess that is about 14 000–12 000 years old. From these soils he derived a climofunction or mathematical relationship between climate and soil features. A 1994 compilation of comparable data showed a clear relationship between the depth from the surface of the soil of carbonate nodules (D in cm) and the mean annual precipitation (P in mm) according to the formula:

$$P = 139.6 + 6.388D - 1.01303D^2$$

Such climofunctions can be used to interpret palaeoclimate from the depth within palaeosols of calcareous nodules (Figure 4), once allowance is made for reduction in depth due to burial compaction.

Climatic inferences also can be made from ice deformation features, concretions, clay mineral compositions, bioturbation, and chemical analyses of palaeosols. The thick clayey palaeosol shown in Figure 5 is riddled with large root traces of the kind found under forests and is very severely depleted in elemental plant nutrients such as calcium, magnesium, sodium, and potassium. Comparable modern soils are found at mid-latitudes, yet this palaeosol formed during the Triassic at a palaeolatitude of about 70° S. This palaeoclimatic anomaly indicates pronounced global warming, in this case a postapocalyptic greenhouse effect following the largest mass extinction in the history of life at the Permian-Triassic boundary.

Palaeosols and Ancient Ecosystems

Just as soils bear the imprint of the vegetation and other organisms they support, so many aspects of

ancient ecosystems can be interpreted from palaeosols. The palaeosols shown in Figure 4, for example, have a dark crumb-textured surface horizon with abundant fine (1–2 mm) roots, comparable to the modern grassland soil seen forming on the outcrop to the upper left. Forest soils, in contrast (Figure 3), have large woody root traces, a blocky structure, and thick subsurface clayey horizons (Bt).

In some cases root traces in palaeosols are identifiable, although the species *Stigmaria ficoides* (Figure 6) is a form genus for roots of a variety of extinct tree lycopsids and not a precisely identified ancient plant. The tabular form of the roots of *Stigmaria* indicates a poorly drained soil, because roots do not photosynthesize, but rather respire using oxygen from soil air. Tabular, rather than deeply reaching, root traces (Figure 3) are characteristic of swamp palaeosols.

Some palaeosols also contain fossil leaves, fruits, wood, stones, bones, and teeth. These are direct evidence of soil ecosystems. Unlike fossils in deposits of lakes and shallow seas, fossil assemblages in palaeosols have the advantage of being near the place where the organisms lived. However, the preservation of fossils in palaeosols is seldom as ideal as complete skeletons in river-channel deposits or compressed leaves in carbonaceous shales. The carbon and carbonate contents of palaeosols can be used to evaluate the Eh and pH, respectively, of the palaeosol preservational environments of the fossils.

Figure 5 An unusually warm palaeoclimate is indicated by this palaeosol, which is unusually thick, clayey, and deeply weathered for its palaeolatitude of 70°S and is comparable to soils now forming no further south than 48°S (Early Triassic Feather Conglomerate, Allan Hills, Victoria Land, Antarctica).

Figure 6 Swamp forests of tree lycopsids (*Stigmaria ficoides*) grew in waterlogged soils, in which lack of oxygen forced the roots to form planar mats rather than reaching deeply into the soil (Carboniferous Lower Limestone Coal Group, Victoria Park, Glasgow, Scotland).

Palaeosols and Palaeogeography

Just as soils vary from mountain tops to coastal swamps, so do palaeosols give clues to their ancient topographical setting. Many palaeosols within sedimentary sequences show clear relationships with deposits of palaeochannels and levees, so that their depositional subenvironment can be inferred from context. Water tables are close to the ground surface in many sedimentary environments, and palaeosols yield important information on their position relative to ancient water tables. Palaeosols formed below the water table include peats and are grey with chemically reduced minerals such as pyrite and siderite. Burrows of crayfish and other aquatic organisms are locally common in waterlogged soils, but burrows of most rodents and beetles are not. Root traces also do not penetrate deeply into waterlogged soils or palaeosols (Figure 6). Deeply penetrating roots and burrows and red oxidized minerals of iron or aluminium are common in formerly well-drained palaeosols (Figure 3). Palaeosols may also reveal upland sedimentary environments such as alluvial and colluvial fans, glacial moraines, river terraces, and erosional gullies (Figure 7).

Major geological unconformities often mark erosional landscapes of the past. Rocky cliffs and bedrock platforms are found along geological unconformities, but so are upland palaeosols. For example, the hilly erosional landscape of Lewisian Gneiss in northern Scotland had 1 km of relief (Figure 8).

Palaeosols and their Parent Materials

The parent material of a soil or palaeosol is the substance from which it formed and can usually be inferred from the less-weathered lower parts of the profile. The parent material may be precisely known if the palaeosol is on metamorphic or igneous rocks (Figure 8), because pedogenic minerals are easily

Figure 7 A palaeogully in a strongly developed sequence of palaeosols (dark coloured) is filled with alluvium including weakly developed palaeosols (Late Triassic Chinle Formation, Petrified Forest National Park, Arizona, USA). The hill in the foreground is 11 m high. Photograph courtesy of Mary Kraus.

Figure 8 The bleached pink palaeosol formed on gneiss to the right (Sheigra palaeosol) is thicker and more deeply weathered than the light green palaeosol formed on amphibolite to the left (Staca palaeosol). Both palaeosols are overlain by red quartz sandstones of the Torridonian Group (Late Precambrian, near Sheigra, Scotland).

distinquished from igneous and metamorphic minerals. Parent material is more difficult to find in palaeosols that are developed from sedimentary parent materials, especially if sedimentary facies reveal erosional relief (Figure 7). In such settings, the sediment is derived from pre-existing soils, whose degree of weathering can be quite varied. The kinds of soils of sediment and rock also can be very different. If soil were a commercial product, economy would dictate manufacturing it from materials that are already similar in chemical composition and physical characteristics. Soils form more readily from sediments than from rocks. Perhaps the most distinctive of parent materials is volcanic ash, because it may consist of more volcanic glass than minerals. Volcanic glass weathers to noncrystalline amorphous substances such as imogolite, which confer high fertility from loosely bound phosphorous, potassium, and other plant nutrients. Such soils also have low bulk density and good moisture-retaining properties. Such soils around tropical volcanoes support intensive agriculture, despite the hazards of the nearby active volcano, because they are so much more fertile than surrounding soils. Comparable palaeosols are commonly associated with volcanic arcs of the past (Figure 1).

Palaeosols and their Times for Formation

Soils develop their profiles over time, although some soils, such as peats, also accumulate layer-by-layer in the manner of sediments. Each palaeosol within a sedimentary or volcanic sequence represents a short break in sedimentary accumulation, or diastem, whose duration can be calculated from key features of the soil. The peats that become coal seams in the geological record, for example, cannot accumulate at rates of more than 1 mm year ^{1} because the roots will be suffocated by stagnant water. Nor can they accumulate at rates of less than 0.5 mm year ^{1} because aerobic decay will destroy the organic debris as fast as it accumulates. Thus, the durations of coal-bearing palaeosols can be calculated from coal thickness, once compaction is taken into account. Calcareous soils and palaeosols accumulate carbonate at first in wisps and filaments, and later in nodules, which become larger and larger (Figure 4). The size of the nodules thus gives us an idea of the time over which they formed. The development of clayey subsurface horizons is comparable (Figure 3) in that clay becomes more and more abundant over time. The amount of washed-in clay can thus be a guide to the time over which palaeosols formed.

From the times for palaeosol formation and the thickness of rock for successive palaeosols it is possible to calculate rates of sediment accumulation. In the badlands of South Dakota, for example, the clayey lower part of the section accumulated at a slower rate than the ashy and silty upper part of the section (Figure 1). Variations in the rate of sediment accumulation can be used to address a variety of tectonic, volcanic, and sequence stratigraphical problems using palaeosols.

Glossary

Argillan Clay skin, a kind of planar feature in a soil or cutan formed of clay.

Burial decomposition An early diagenetic modification of a palaeosol in which buried organic matter is decayed microbially.

Burial gleization An early diagenetic modification of a palaeosol in which buried organic matter fuels microbial chemical reduction of iron oxides and oxyhydraes to ferrous clays, siderite or pyrite.

Climofunction A mathematical relationship between a soil feature and a measure of climate.

Climosequence A set of soils formed under similar vegetation, topographic setting, parent material and time, but varied climate.

Concretion A seggregation of materials in a soil, harder or more cemented than the matrix, with prominent internal concentric banding, for example iron-manganese concretion.

Cutan A planar feature within a soil formed by enrichment, bleaching, coating or other alteration, for example a clay skin (argillan).

Ferran Ferruginized surface, a kind of planar feature in a soil (cutan) formed by chemical oxidation.

Geosol A mappable land surface of palaeosols, a soil stratigraphic unit in the North American Code of Stratigraphic Nomenclature.

Nodule A segregation of materials in a soil, harder or more cemented than the matrix, with massive internal fabric, for example caliche nodule.

Palaeosol A soil of a landscape of the past: a past surficial region of a planet or similar body altered in place by biological, chemical or physical processes, or a combination of these.

Ped A natural aggregate of soil: stable lumps or clods of soil between roots, burrows, cracks and other planes of weakness.

Pedoderm A mappable land surface of palaeosols, a soil stratigraphic unit in the Australian Code of Stratigraphic Nomenclature.

Pedolith Soil sediment: a seadimentary rock dominated by clasts with the internal microfabrics of soils.

Pedotype A kind of palaeosol: an ancient equivalent of soil series of the United States Soil Conservation Service.

Perched water table Level of water ponded in a soil by an impermeable subsurface layer.

Sepic plasmic fabric Birefringence microfabric: appearance of the fine grained part of a soil or palaeosol in petrographic thin sections viewed under crossed Nicols of wisps or streaks of highly oriented and highly birefringent clay in a less organized dark matrix.

See Also

Carbon Cycle. **Clay Minerals**. **Palaeoclimates**. **Sedimentary Environments:** Depositional Systems and Facies; Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Evaporites. **Soils:** Modern. **Weathering**.

Further Reading

Delvigne JE (1998) *Atlas of Micromorphology of Mineral Alteration and Weathering*. Canadian Mineralogist Special Publication 3. Ottawa: Mineralogical Association of Canada.

Follmer LR, Johnson GD, and Catt JA (eds.) (1998) Revisitation of concepts in paleopedology. *Quaternary International* 51/52: 1 221.

International Subcommission on Stratigraphic Classifica tion (1994) *International Stratigraphic Guide*. Boulder: Geological Society of America.

Jenny HJ (1941) *Factors in Soil Formation*. New York: Wiley.

Martini IP and Chesworth W (eds.) (1992) *Weathering, Soils and Paleosols*. Amsterdam: Elsevier.

Ollier C (1991) *Ancient Landforms*. London: Belhaven.

Ollier C and Pain C (1996) *Regolith, Soils and Landforms*. Chichester: Wiley.

Reinhardt J and Sigleo WR (1988) *Paleosols and Weathering through Geologic Time: Principles and Ap plications*. Special Paper 216. Boulder: Geological Society of America.

Retallack GJ (1983) *Late Eocene and Oligocene Paleosols from Badlands National Park, South Dakota*. Special Paper 193. Boulder: Geological Society of America.

Retallack GJ (ed.) (1986) Precambrian paleopedology. *Precambrian Research* 32: 93 259.

Retallack GJ (1991) *Miocene Paleosols and Ape Habitats of Pakistan and Kenya*. New York: Oxford University Press.

Retallack GJ (1997) *A Colour Guide to Paleosols*. Chiches ter: Wiley.

Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Retallack GJ, Bestland EA, and Fremd TJ (2000) *Eocene and Oligocene Paleosols of Central Oregon*. Special Paper 344. Boulder: Geological Society of America.

Thiry M and Simon Coinçon R (eds.) (1999) *Palaeo weathering, Palaeosurfaces, and Related Continental Deposits*. Oxford: Blackwell.

Wright VP (ed.) (1986) *Paleosols: their Recognition and Interpretation*. Oxford: Blackwell.

SOLAR SYSTEM

Contents

The Sun

K R Lang, Tufts University, Medford, MA, USA

Physical Characteristics of the Sun

Distance to the Sun

The mean distance of the Sun from the Earth sets the scale of our Solar System and enables us to infer, from other observations, the luminosity, radius, effective temperature, and mass of the Sun. This distance is called the astronomical unit, or AU for short, with a value of $1\,\mathrm{AU} = 1.49597870 \times 10^{11}\,\mathrm{m}$. At that distance, light from the Sun takes 499.004 782 s to travel to the Earth. By way of comparison, light from the Sun's nearest stellar neighbour, Proxima Centauri (part of the triple star system Alpha Centauri), takes 4.29 years to reach us.

Absolute Solar Luminosity

The Sun's absolute, or intrinsic, luminosity is designated by the symbol $L_\odot$, where the subscript $\odot$ denotes the Sun. We can infer the Sun's luminosity from satellite measurements of the total amount of solar energy reaching every square centimetre of the Earth every second, obtaining $L_\odot = 3.854 \times 10^{26}$ W, where a power of $1\,\mathrm{W} = 1\,\mathrm{J\,s^{-1}}$.

Radius of the Sun

The Sun's radius, which can be inferred from its distance and angular extent, has a value of $R_\odot = 6.955 \times 10^8\,\mathrm{m}$. That is about 109 times the radius of the Earth.

The Sun's Effective Temperature

We can use the Stefan–Boltzmann law, together with the Sun's size and luminous output, to determine an effective temperature of 5780 K. The temperature of the Sun increases below and above the visible disk (Table 1).

Mass of the Sun

The Sun's gravitational pull holds the solar system together. That is why we call it a solar system: governed by the central Sun with its huge mass. This gravitational attraction keeps the planets in orbit around the Sun, with longer orbital periods at increasing distances from the Sun. And since we know the Earth's orbital period and mean distance from the Sun, we can weigh the Sun from a distance, obtaining its mass $M_\odot = 5.9165 \times 10^{11}\ (\mathrm{AU})^3/P^2 = 1.989 \times 10^{30}\,\mathrm{kg}$, where the constant is equal to $4\pi^2/G$, the universal constant of gravitation is G, the semi-major axis of the Earth's orbit is $1\,\mathrm{AU} = 1.4959787 \times 10^{11}\,\mathrm{m}$, and the orbital period of the Earth is $\mathrm{P} = 1$ year $= 3.1557 \times 10^7\,\mathrm{s}$.

The Sun does not just lie at the heart of our solar system; it dominates it. Some 99.8% of all the matter between the Sun and halfway to the nearest star is contained in the Sun. It is 332 946 times the mass of the Earth. All the objects that orbit the Sun—the planets and their moons, the comets, and the asteroids—add up to just 0.2% of the mass in our solar system.

Composition of the Sun

When the intensity of sunlight is displayed as a function of wavelength, in a spectrum, it exhibits numerous fine dark gaps of missing colours called

Table 1 The Sun's physical properties[a]

Mean distance, AU	1.4959787×10^{11} m
Light travel time from Sun to Earth	499.004782 s
Radius, $R_\odot$	6.955×10^{8} m (109 Earth radii)
Volume	1.412×10^{27} m^3 (1.3 million Earths)
Mass, $M_\odot$	1.989×10^{30} kg (332 946 Earth masses)
Escape velocity at photosphere	617 km s^{-1}
Mean density	1409 kg m^{-3}
Solar constant, $f_\odot$	1366 J s^{-1} m^{-2} 1366 W m^{-2}
Luminosity, $L_\odot$	3.854×10^{26} J s^{-1} 3.854×10^{26} W
Principal chemical constituents	(By number of atoms) (By mass fraction)
	Hydrogen 92.1% X 70.68%
	Helium 7.8% Y 27.43%
	All other 0.1% Z 1.89%
Age	4.566 billion years
Temperature (center)	15.6 million K
Temperature (effective)	5780 K
Temperature (photosphere)	6400 K
Temperature (chromosphere)	6000 to 20 000 K
Temperature (corona)	2 million to 3 million K
Rotation period (equator)	26.8 days
Rotation period (60° latitude)	30.8 days
Magnetic field (sunspots)	0.1 to 0.4 T 1000 to 4000 G
Magnetic field (polar)	0.001 T 10 G

[a]Mass density is given in kilograms per cubic metre (kg m^{-3}); the density of water is 1000 kg m^{-3}. The unit of luminosity is joules per second, power is often expressed in watts, where 1.0 W 1.0 J s^{-1}.

absorption lines. Each chemical element, and only that element, produces a unique set, or pattern, of wavelengths at which the dark absorption lines fall. So these lines can be used to determine the chemical ingredients of the Sun. They indicate that hydrogen is the most abundant element in the visible solar gases. Since the Sun is chemically homogenous, except for its core, a high hydrogen abundance is implied for the entire star, and this was confirmed by subsequent calculations of its luminosity. Hydrogen accounts for 92.1% of the number of atoms in the Sun, and it amounts to 70.68% by mass.

Helium, the second-most abundant element in the Sun, accounts for 7.8% of the number of atoms in the Sun, and it amounts to 27.43% by mass. Helium is so rare on Earth that it was first discovered on the Sun. All of the heavier elements in the Sun amount to only 0.1% of the number of atoms, and just 1.89% by mass.

Rotation of the Sun

The Sun rotates, or spins, around a rotational axis whose top and bottom mark the Sun's north and south poles. Like Earth, the Sun rotates from west to east when viewed from above the north pole, but unlike Earth, different parts of the Sun rotate at different rates. We know from watching sunspots that the visible disk of the Sun rotates faster at the equator than it does at higher latitudes, decreasing in speed evenly towards each pole. Also, because the Earth orbits the Sun, we observe a rotation period that is about a day longer than the true value. The synodic rotation period of the visible solar equator, as observed from Earth, is 26.75 days, while the equatorial region of the visible solar disk is intrinsically spinning about the Sun's axis once every 25.67 days.

Scientists have used sound waves, generated inside the Sun, to show that the differential rotation of the Sun persists to about one-third of the way down inside the Sun, or 220 000 km from the visible disk. Lower down the rotation speed becomes uniform from pole to pole and the rotation rate remains independent of latitude. The Sun's magnetism is probably generated at the interface between the deep interior, which rotates with one speed, and the overlying gas that spins faster in the equatorial middle.

Solar Magnetic Fields

Detailed scrutiny indicates that the visible solar disk often contains dark, ephemeral spots, called sunspots, which can be as large as the Earth. The sunspots appear and disappear, rising out from inside the Sun and moving back into it. Most sunspots remain visible for only a few days; others persist for weeks and even months.

Sunspots contain magnetic fields as strong as 0.3 T, or 3000 G, thousands of times stronger than the Earth's magnetic field. The intense sunspot magnetism chokes off the upward flow of heat and energy from the solar interior, keeping a sunspot thousands of degrees colder than the surrounding gas.

The total number of sunspots visible on the Sun varies over an 11-year cycle. At the maximum in the cycle we may find 100 or more spots on the visible disk of the Sun at one time; at sunspot minimum very few of them are seen, and for periods as long as a month none can be found. Since most forms of solar activity are magnetic in origin, they also follow an 11-year cycle. Thus, the sunspot cycle is also known as the solar cycle of magnetic activity.

Sunspots are usually found in pairs or groups of opposite magnetic polarity. The magnetic field lines emerge from a sunspot of one polarity, loop through the solar atmosphere above it, and enter a neighbouring sunspot of opposite polarity. The highly magnetized realm in, around, and above bipolar sunspot pairs or groups is a disturbed area called an active

region; it consists of sunspots and the magnetic loops that connect them.

Sunspots are usually oriented roughly parallel to the Sun's equator, in the east–west direction of the Sun's rotation. Moreover, sunspot pairs in either the northern or southern hemisphere have the same orientation and polarity alignment, with an exact opposite arrangement in the two hemispheres.

The Outer Solar Atmosphere

The visible photosphere, or sphere of light, is the level of the solar atmosphere from which we get our light and heat, and it is the part that we can see with our eyes. The thin chromosphere and extensive corona lie above the visible sharp edge of the photosphere. They can both be seen during a total solar eclipse, when the Moon blocks the intense light of the photosphere.

Telescopes called coronagraphs allow us to see the corona by using occulting disks to mask the Sun's face and block out the photosphere's glare. Modern solar satellites, such as the *Solar and Heliospheric Observatory* (*SOHO*), use coronagraphs to get clear, edge-on views of the corona.

The solar corona has a temperature of millions of degrees kelvin, hundreds of times hotter than the underlying visible solar disk whose effective temperature is 5780 K. Very hot material—such as that within the corona—emits most of its energy at X-ray wavelengths. Also, the photosphere is too cool to emit intense radiation at these wavelengths, so it appears dark under the hot gas. As a result, the million-degree corona can be seen all across the Sun's face, with high spatial and temporal resolution, in X-rays.

Since X-rays are totally absorbed by the Earth's atmosphere, they must be observed through telescopes in space. This has been done using a soft X-ray telescope on the *Yohkoh* spacecraft (**Figure 1**). *Yohkoh's* soft X-ray images have demonstrated that the corona contains thin, bright, magnetized loops that shape, mold, and constrain the million-degree gas. Wherever the magnetism in these coronal loops is strongest, the coronal gas in them shines brightly at soft X-ray wavelengths.

Not all magnetic fields on the Sun are closed loops. Some of the magnetic fields extend outward, within regions called coronal holes. These extended regions

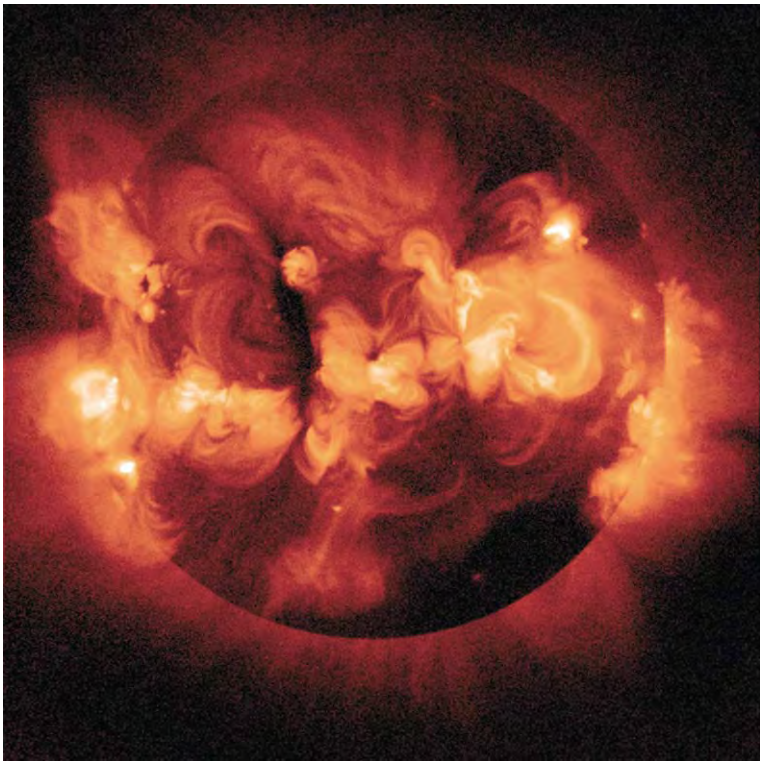


Figure 1 The Sun in X rays. The bright glow seen in this X ray image of the Sun is produced by ionized gases at a temperature of a few million degrees kelvin. It shows magnetic coronal loops which thread the corona and hold the hot gases in place. The brightest features are called active regions and correspond to the sites of the most intense magnetic field strength. This image of the Sun's corona was recorded by the Soft X ray Telescope (SXT) aboard the Japanese *Yohkoh* satellite on 1 February 1992, near the maximum of the 11 year cycle of solar magnetic activity. Courtesy of Gregory L Slater, Gary A Linford, and Lawrence Shing, NASA, ISAS, Lockheed Martin Solar and Astrophysics Laboratory, National Astronomical Observatory of Japan, and University of Tokyo.

have so little hot material in them that they appear as large dark areas seemingly devoid of radiation at X-ray wavelengths. Coronal holes are nearly always present at the Sun's poles, and are sometimes found at lower solar latitudes. The open magnetic fields in coronal holes do not return directly to another place on the Sun, allowing charged particles to escape the Sun's magnetic grasp and flow outwards into surrounding space.

Explosions on the Sun

Solar Flares

Sudden and brief explosions, called solar flares, rip through the atmosphere above sunspots, releasing an incredible amount of energy, amounting to as much as a million, billion, billion (10^{24}) joules in just a few minutes. All of this power is created in a relatively compact explosion, comparable in total area to an Earth-sized sunspot.

For a short time, usually about 10 minutes, a flare is heated to tens of millions of degrees kelvin. The explosion floods the solar system with intense radiation across the full electromagnetic spectrum, from the shortest X-rays to the longest radio waves, and hurls high-energy electrons and protons out into interplanetary space.

Despite the powerful cataclysm, most solar flares are only minor perturbations in the total amount of emitted sunlight. Routine visual observations of solar explosions are only made possible by tuning into the red emission of hydrogen alpha, designated Hα, at a wavelength of 656.3 nm, and rejecting all the other colours of sunlight.

Since solar flares are very hot, they emit the bulk of their energy at X-ray wavelengths, and for a short while, a large flare can outshine the entire Sun in X-rays. The energetic electrons that produce the impulsive, flaring X-ray emission also emit radio waves known as a radio burst to emphasize its brief, energetic, and explosive characteristics. A solar flare can also outshine the entire Sun at radio wavelengths.

There are more flares near the peak of the 11-year cycle of magnetic activity, but this does not mean that sunspots cause solar flares. They are instead energized by the powerful magnetism associated with sunspots. When these magnetic fields become contorted, they can suddenly and explosively release pent-up magnetic energy as a solar flare, with a main energy release in the corona just above sunspots. The energy is apparently released when magnetized coronal loops, driven by motions beneath them, meet to touch each other and connect. If magnetic fields of opposite polarity are pressed together, an instability takes place and the fields partially annihilate each other, releasing energy to power the explosion.

Coronal Mass Ejections

A coronal mass ejection (CME) is a giant magnetic bubble that rapidly expands to rival the Sun in size. Each time a mass ejection rises out of the corona, it carries away up to 50 billion tons (5×10^{13} kg) of coronal material. Its associated shocks also accelerate and propel vast quantities of high-speed particles ahead of them.

CMEs release about as much energy as a solar flare. However, most of the energy of a mass ejection goes into the kinetic energy of the expelled material, whereas a flare's energy is mainly transferred into accelerated particles that emit intense X-ray and radio radiation.

Coronal mass ejections are detected during routine visible-light observations of the corona from spacecraft such the *SOHO*. With a disk in the centre to block out the Sun's glare, the coronagraph is able to show huge pieces of the corona blasted out from the edge of the occulted photosphere (**Figure 2**).

Like sunspots, solar flares, and other forms of solar activity, coronal mass ejections occur with a frequency that varies in step with the 11-year cycle. A few coronal mass ejections balloon out of the corona per day, on average, during activity maximum, and the rate decreases by about an order of magnitude by sunspot minimum.

The triggering mechanism for CMEs seems to be related to large-scale interactions of the magnetic field in the low solar corona. This magnetism is continuously emerging from inside the Sun, and disappearing back into it, driven by the Sun's 11-year cycle of magnetic activity. The release of a coronal mass ejection appears to be one way that the solar atmosphere reconfigures itself in response to these slow magnetic changes.

The Sun's Winds

Basic Properties of the Solar Wind

The tenuous solar atmosphere expands out in all directions, filling interplanetary space with a ceaseless wind that is forever blowing from the Sun. This solar wind is mainly composed of electrons and protons, set free from the Sun's abundant hydrogen atoms, but it also contains heavier ions and magnetic fields. This perpetual solar gale brushes past the planets and engulfs them, carrying the Sun's atmosphere out into interstellar space at the rate of a million tons (10^6 tons $= 10^9$ kg) every second.

The relentless wind has never stopped blowing during the more than three decades that it has been

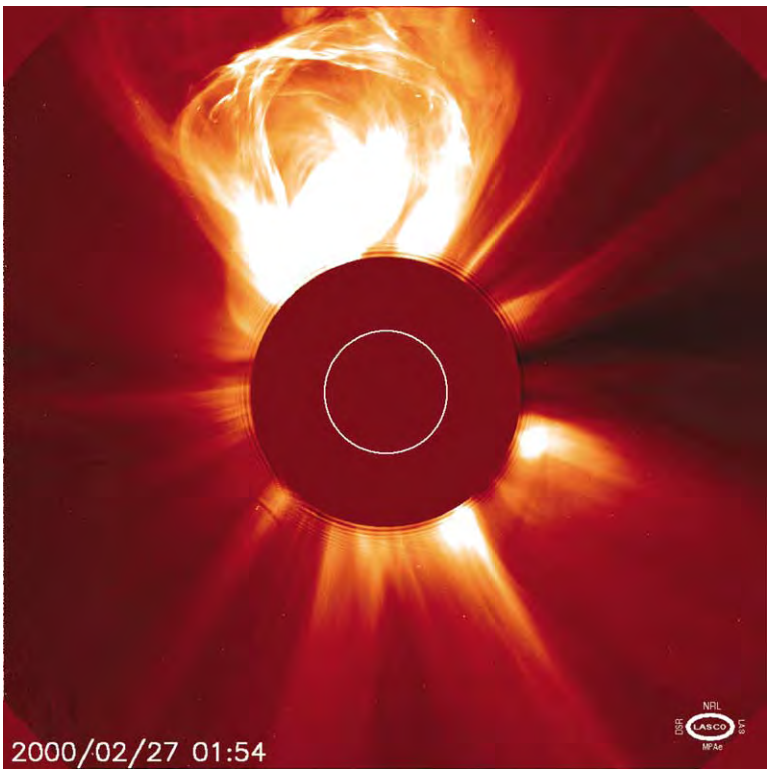


Figure 2 Coronal mass ejection. A huge coronal mass ejection is seen in this coronagraph image, taken on 27 February 2000 with the Large Angle Spectrometric Coronagraph (LASCO) on *the Solar and Heliospheric Observatory* (*SOHO*). The white circle denotes the edge of the photosphere, so this mass ejection is about twice as large as the visible Sun. The black area corresponds to the occulting disk of the coronagraph that blocks intense sunlight and permits the corona to be seen. About one hour before this image was taken, another *SOHO* instrument, the Extreme Ultraviolet Imaging Telescope (EIT), detected a filament eruption lower down near the solar chromosphere. Courtesy of the *SOHO* LASCO consortium. *SOHO* is a project of international cooperation between ESA and NASA.

Table 2 Mean values of solar wind parameters at the Earth's orbit[a]

Parameter	*Mean value*
Particle density, N	$N \approx 10$ million particles m ^{3} (5 million electrons and 5 million protons)
Velocity, V	$V \approx 375\,000$ m s ^{1} and V $\approx 750\,000$ m s ^{1}
Flux, F	$F \approx 10^{12}$ to 10^{13} particles m ^{2} s ^{1}
Temperature, T	T $\approx 120\,000$ K (protons) to 140 000 K (electrons)
Magnetic field strength, H	$H \approx 6 \times 10^{9}$ T 6 nT 6×10^{5} G

[a]These solar wind parameters are at the mean distance of the Earth from the Sun, or at one astronomical unit, 1 AU, where 1 AU 1.496×10^{11} m.

observed with spacecraft. Two winds are always detected—a fast, uniform wind blowing at about 750 km s ^{1} and a variable, gusty slow wind moving at about half that speed.

By the time it reaches the Earth's orbit, the solar wind has been diluted by its expansion into the increasing volume of space to about 5 million electrons and 5 million protons per cubic meter (Table 2).

The charged particles in the solar wind drag the Sun's magnetic fields with them. While one end of the interplanetary magnetic field remains firmly rooted in the photosphere and below, the other end is extended and stretched out by the radial expansion of the solar wind. The Sun's rotation bends this radial pattern into an interplanetary spiral shape within the plane of the Sun's equator. This spiral pattern has been confirmed by tracking the radio emission of high-energy electrons emitted during solar flares (Figure 3), as well as by spacecraft that have directly measured the interplanetary magnetism.

Origin of the Sun's Winds

The million-degree coronal gas creates an outward pressure that tends to oppose the inward pull of the Sun's gravity. At great distances, where the solar gravity weakens, the hot protons and electrons in the corona overcome the Sun's gravity and accelerate away to supersonic speed.

Instruments aboard the *Ulysses* spacecraft conclusively proved that a relatively uniform, fast wind pours out at high latitudes near the solar poles, and

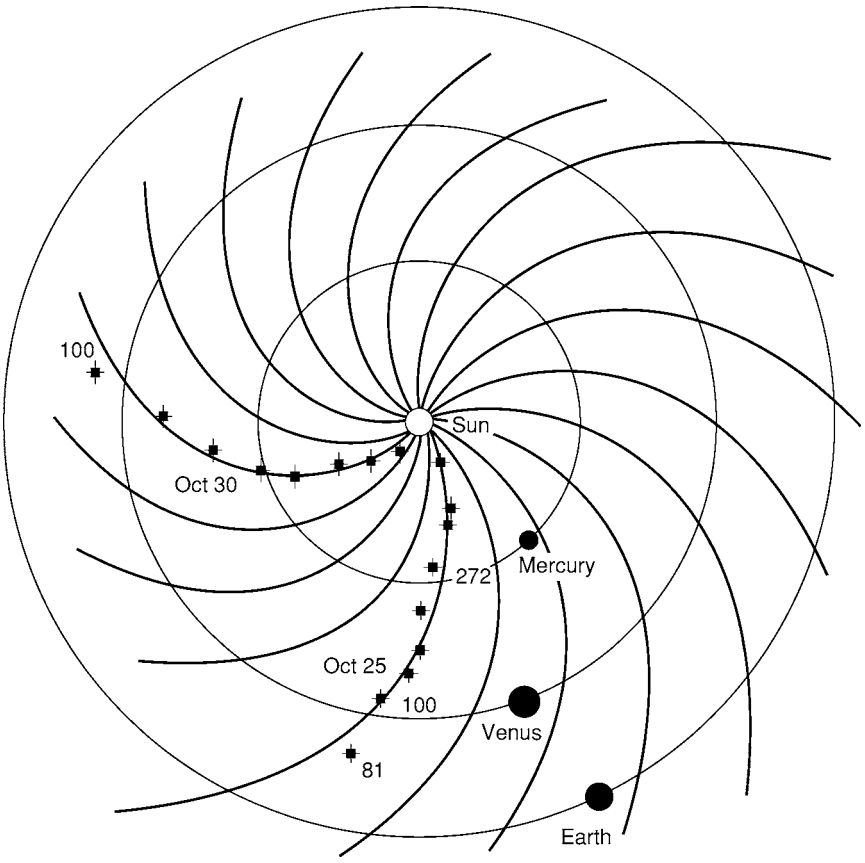


Figure 3 Spiral path of interplanetary electrons. The trajectory of flare electrons in interplanetary space as viewed from above the polar regions using the *Ulysses* spacecraft. As the high speed electrons move out from the Sun, they excite radiation at successively lower plasma frequencies; the numbers denote the observed frequency in kilohertz (kHz). Since the flaring electrons are forced to follow the interplanetary magnetic field, they do not move in a straight line from the Sun to the Earth, but instead move along the spiral pattern of the interplanetary magnetic field, shown by the solid curved lines. The squares and crosses show *Ulysses* radio mea surements of type III radio bursts on 25 October 1994 and 30 October 1994. The approximate locations of the orbits of Mercury, Venus, and the Earth are shown as circles. Courtesy of Michael J Reiner. *Ulysses* is a project of international collaboration between ESA and NASA.

that a capricious, gusty, slow wind emanates from the Sun's equatorial regions at activity minimum.

Comparisons with *Yohkoh* soft X-ray images showed that much, if not all, of the high-speed solar wind flows out of the open magnetic fields in polar coronal holes, at least during the minimum in the 11-year cycle of magnetic activity. In addition, instruments aboard *SOHO* have shown that the strongest high-speed flows gush out of a magnetic network at the bottom of coronal holes near the Sun's poles. Comparisons of *Ulysses* data with coronagraph images pinpointed the equatorial coronal streamers as the birthplace of the slow and sporadic wind during the minimum in the 11-year cycle.

The Heliosphere

A solar gale carries the Sun's rarefied atmosphere past the planets and out to the space between the stars, creating a large cavity in interstellar space called the heliosphere—from the Greek word 'helios' for 'Sun'. Within the heliosphere, physical conditions are dominated, established, maintained, modified, and governed by the magnetic fields and charged particles in the solar wind.

The solar wind's domain extends out to about 150 times the distance between the Earth and Sun, marking the outer boundary of the heliosphere and the edge of our solar system. Out there, the solar wind has become so weakened by expansion that it is no longer dense or powerful enough to repel the ionized matter and magnetic fields coursing between the stars.

The Sun–Earth Connection

Radiation from the Sun

The Sun emits radiation that carries energy through space as waves. Different types of solar radiation differ in their wavelength, although they propagate

at the same speed—299 792 458 m s $^{-1}$, the velocity of light.

Our eyes detect a narrow range of these wavelengths that is collectively called visible radiation. From long to short waves, the colours of visible sunlight correspond to red, orange, yellow, green, blue, and violet. Red light has a wavelength of about 7×10^{-7} m, or 700 nm, and violet waves are about 400 nm long.

Although the most intense radiation from the Sun is emitted at visible wavelengths, it emits less luminous radiation at invisible wavelengths that include the infrared and radio waves, with wavelengths longer than that of red light, and ultraviolet (UV), X-rays, and gamma (γ) rays, which have wavelengths shorter than that of violet light.

Radio waves have wavelengths between 0.001 and 10 m, and they pass easily through the atmosphere, even on cloudy days or in stormy weather. The infrared part of the Sun's spectrum is located between the radiowave region and the red part of the visible region. Atmospheric molecules, such as carbon dioxide and water vapor, absorb infrared radiation from the Sun.

The short-wavelength, ultraviolet radiation from the Sun is sufficiently energetic to tear electrons or atoms off many of the molecular constituents of the Earth's atmosphere, particularly in the stratosphere where ozone is formed.

The X-ray region of the Sun's spectrum extends from a wavelength of one-hundred billionth (10^{-11}) of a meter, which is about the size of an atom, to the short-wavelength side of the ultraviolet. X-ray radiation is so energetic that it is usually described in terms of the energy it carries. The X-ray region lies between 1 and 100 keV (kiloelectron volts) of energy, where 1 keV $= 1.6 \times 10^{-16}$ J. The atmosphere effectively absorbs most of the Sun's ultraviolet radiation and all of its X-rays.

Varying Solar Irradiance of Earth

The total amount of the Sun's life-sustaining energy is called the 'solar constant', and it is defined as the total amount of radiant solar energy per unit time per unit area reaching the top of the Earth's atmosphere at the Earth's mean distance from the Sun. Satellites have been used to accurately measure the solar constant, obtaining a value of $f_{\odot} = 1366.2\ \mathrm{W\,m^{-2}}$, where the power of one watt is equivalent to one joule per second and the uncertainty in this measurement is $\pm 1.0\ \mathrm{W\,m^{-2}}$.

The total power received at any square metre of the Earth's surface, known as the solar insolation, is much less than the solar constant. This is due to the absorption of sunlight in the terrestrial atmosphere, as well as the time of day.

The solar constant is almost always changing, in amounts of up to a few tenths of a per cent and on time-scales from 1 s to 20 years. This inconstant behaviour can be traced to changing magnetic fields in the solar atmosphere, and its variation tracks the 11-year cycle of magnetic activity (Figure 4).

There are enormous changes in the Sun's radiation at the short ultraviolet and X-ray wavelengths that contribute only a tiny fraction of the Sun's total luminosity. The ultraviolet emission doubles from the minimum to maximum of the 11-year cycle, while the X-ray brightness of the corona increases by a factor of 100.

Global Warming and Cooling by the Sun

The brightening and dimming of the Sun dominated our climate for two centuries, from 1600 to 1800. Cooling by hazy emission from volcanoes next played an important role, but the Sun noticeably warmed the climate for another century, from 1870 to 1970. After that, heat-trapping gases apparently took control of our climate. Global warming by the greenhouse effect is probably responsible for this recent, unprecedented rise in temperature. If current emissions of carbon dioxide and other greenhouse gases go unchecked, the average surface temperature of the globe will rise by about 2°C, making the Earth hotter than it has been in millions of years.

The varying Sun may offset some of this warming. Observations of other stars indicate that the Sun luminosity could vary by much more than that observed by satellites so far, producing dramatic changes in the Earth's climate on time-scales of hundreds and thousands of years. Radioactive isotopes found in both tree rings and ice cores indicate that the Sun's activity has fallen to unusually low levels at least three times during the past one thousand years, each drop apparently corresponding to a long, cold spell on Earth of roughly a century in duration.

Further back in time, during the past one million years, our climate has been dominated by the recurrent ice ages, each lasting about 100 000 years. These glaciations begin and end in a relatively short interval of unusual warmth, called an interglacial, lasting 10 000 or 20 000 years, when the glaciers retreat. We now live in such a warm interglacial interval. The rhythmic alteration of glacial and interglacial intervals is related to periodic alterations in the amount and distribution of sunlight received by Earth over tens of thousands of years.

Our Sun-Layered Atmosphere

Our thin atmosphere is pulled close to the Earth by its gravity, and suspended above the ground by

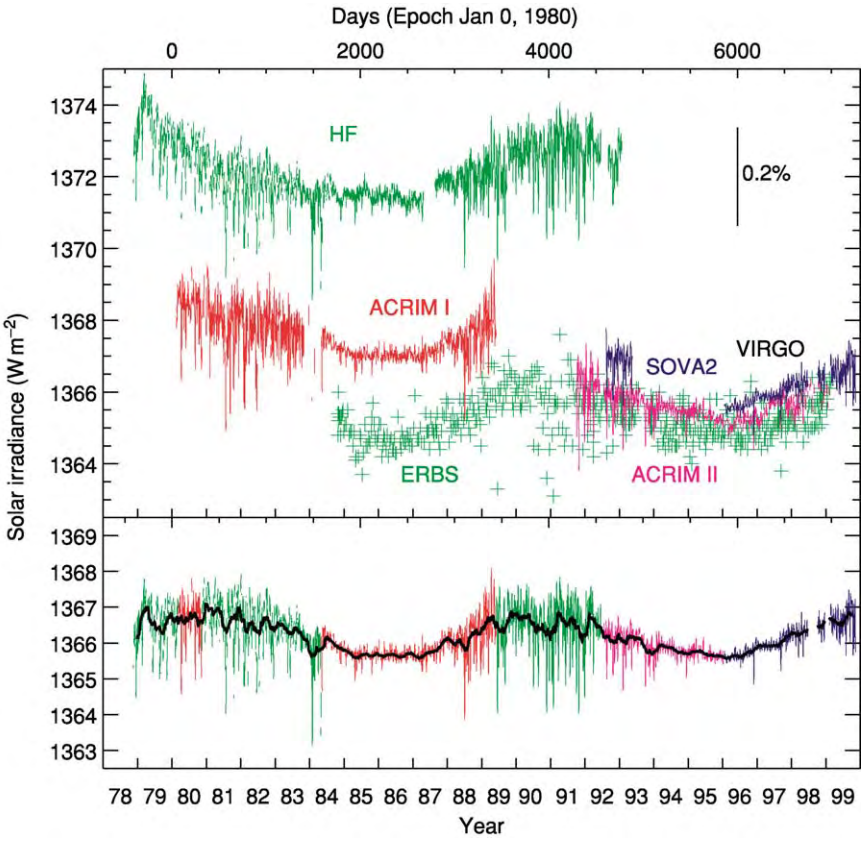


Figure 4 Variations in the solar constant. Observations with very stable and precise detectors on several Earth orbiting satellites show that the Sun's total radiative input to the Earth, termed the solar irradiance, is not a constant, but instead varies over time scales of days and years. Measurements from five independent space based radiometers since 1978 (top) have been combined to produce the composite solar irradiance (bottom) over two decades. They show that the Sun's output fluctuates during each 11 year sunspot cycle, changing by about 0.1% between maximums (1980 and 1990) and minimums (1987 and 1997) in magnetic activity. Temporary dips of up to 0.3% and a few days' duration are due to the presence of large sunspots on the visible hemisphere. The larger number of sunspots near the peak in the 11 year cycle is accompanied by a rise in magnetic activity that creates an increase in luminous output that exceeds the cooling effects of sunspots. Here the total irradiance just outside our atmosphere, called the solar constant, is given in units of watts per square metre, where 1 W 1 J s ^{1}. The capital letters are acronyms for the different radiometres, and offsets among the various datasets are the direct result of uncertainties in their scales. Despite these offsets, each dataset clearly shows varying radiation levels that track the overall 11 year solar activity cycle. Courtesy of Claus Frohlich.

molecular motion. The atmosphere near the ground is compacted to its greatest density and pressure by the weight of the overlying air. At greater heights there is less air pushing down from above, so the compression is less and the density and pressure of the air falls off into the near vacuum of space. The temperature of the air falls and rises in two full cycles at increasing altitudes, and the temperature increases are caused by the Sun's energetic radiation (Figure 5).

The temperature above the ground tends to fall at higher altitudes where the air expands in the lower pressure and becomes cooler. The average air temperature drops below the freezing point of water (273 K) about 1 km above the Earth's surface, and bottoms out at roughly 10 times this height.

The temperature increases at greater heights, within the stratosphere, where the Sun's invisible ultraviolet radiation warms the gas and helps make ozone. This ozone layer protects us by absorbing most of the ultraviolet and keeping its destructive rays from reaching the ground, where it can cause eye cataracts and skin cancer.

Due to the Sun's variable ultraviolet radiation, the total global amount of ozone becomes enhanced, depleted, and enhanced again from 1 to 2% every 11 years, modulating the protective ozone layer at a level comparable to human-induced ozone depletion by chemicals wafting up from the ground. Monitoring of the expected recovery of the ozone layer from outlawed, man-made chemicals will therefore require careful watch over how the Sun is changing the layer from above.

The temperature declines rapidly with increasing height just above the stratosphere, reaching the lowest levels in the entire atmosphere, but the temperature

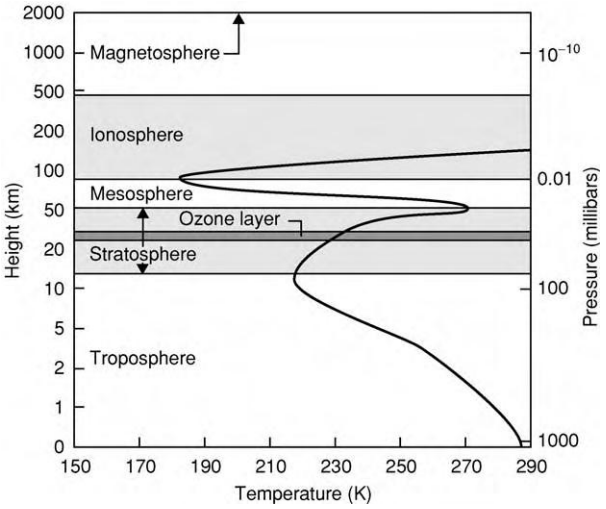


Figure 5 Sun layered atmosphere. The pressure of our atmosphere (right scale) decreases with altitude (left scale). This is because fewer particles are able to overcome the Earth's gravitational pull and reach higher altitudes. The temperature (bottom scale) also decreases steadily with height in the ground hugging troposphere, but the temperature increases in two higher regions heated by the Sun. They are the stratosphere, with its critical ozone layer, and the ionosphere. The stratosphere is mainly heated by ultraviolet radiation from the Sun, and the ionosphere is created and modulated by the Sun's X ray and extreme ultraviolet radiation.

rises again within the ionosphere, reaching temperatures that are hotter than the ground. The ionosphere is created and heated by absorbing the extreme ultraviolet and X-ray portions of the Sun's energy. This radiation tears electrons off the atoms and molecules in the upper atmosphere, thereby creating ions and free electrons not attached to atoms.

At a given height in the ionosphere, the temperature, the density of free electrons, and the density of neutral, unionized atoms all increase and decrease in synchronism with solar activity over its 11-year cycle.

The Earth's Magnetosphere

Invisible magnetic fields, produced by currents in the Earth's molten core, emerge out of the Earth's south geographic polar regions, loop through nearby space, and re-enter at the north polar regions. The surface equatorial field strength is 0.000031 T, or 31 000 nT, and the field strength decreases at greater distances from the Earth.

Yet, the Earth's magnetism is strong enough to deflect the Sun's wind away from the Earth, forming the magnetosphere (Figure 6). The magnetosphere of the Earth, or any other planet, is that region surrounding the planet in which its magnetic field dominates the motions of energetic charged particles such as electrons, protons, and other ions. It is also the volume of space from which the main thrust of the solar wind is excluded.

The solar wind pushes the terrestrial magnetic field towards the Earth on the dayside that faces the Sun, compressing the outer magnetic boundary and forming a bow shock at about 10 times the Earth's radius. Also the Sun's wind drags and stretches the Earth's magnetic field out into a long magnetotail on the night side of our planet. The magnetic field points roughly towards the Earth in the northern half of the tail and away in the southern. The field strength drops to nearly zero at the centre of the tail where the opposite magnetic orientations lie next to each other and currents can flow.

Some of the energetic particles outside the magnetosphere do manage to penetrate it, especially in the magnetotail. When the solar and terrestrial magnetic fields touch each other in the magnetotail, it can catapult the outer part of the tail downstream and propel the inner part back towards Earth.

The inner magnetosphere is always filled with electrons and protons, trapped within two torus-shaped belts that encircle the Earth's equator but do not touch it. These regions are often called the inner and outer Van Allen radiation belts, named after James A Van Allen (1914–) who discovered them in 1958. The inner belt is about 1.5 Earth radii from planet centre, and the outer belt is located at about 4.5 Earth radii, where the Earth's radius is 6378 km.

Intense Geomagnetic Storms

Significant variations in the Earth's magnetic field, lasting seconds to days, are known as geomagnetic storms. The great, sporadic geomagnetic storms, which shake the Earth's magnetic field to its very foundations, can produce magnetic fluctuations as large as 1.6% at mid-terrestrial latitudes, or 500 nT, compared with the Earth's equatorial field strength of 31 000 nT.

Solar wind disturbances driven by exceptionally fast coronal mass ejections produce the most intense geomagnetic storms. The Earth intercepts about 70 coronal mass ejections per year when solar activity is at its peak, and less than 10 will have the punch needed to produce large, geomagnetic storms. These mass ejections plow through the solar wind, driving a huge shock wave far ahead of them. When directed at the Earth, these shocks ram into the terrestrial magnetic field and trigger the initial phase, or sudden commencement, of an intense geomagnetic storm a few days after the mass ejection leaves the Sun.

Strong interplanetary magnetic fields are also generated by fast coronal mass ejections (*see* **Magnetostratigraphy**). It is their intense magnetism and high speed that account for the main phase of a powerful

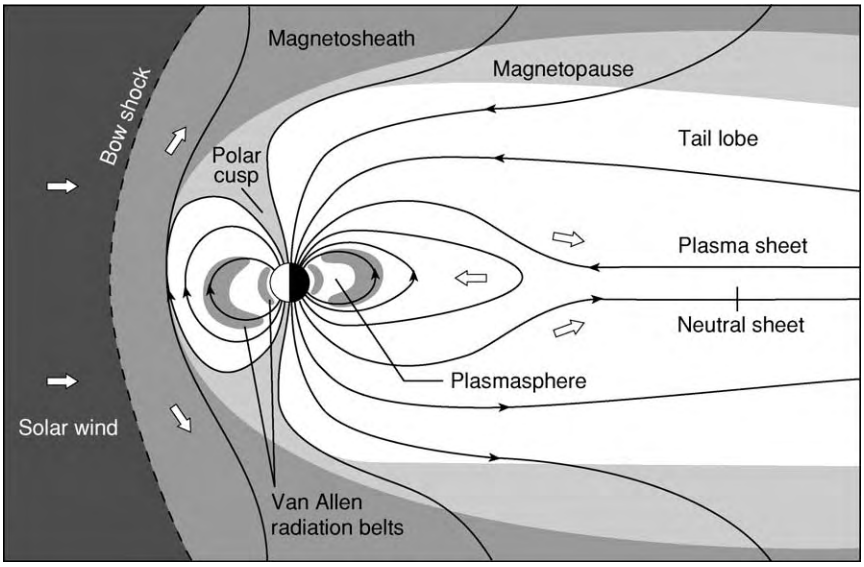


Figure 6 Magnetosphere. The Earth's magnetic field carves out a hollow in the solar wind, creating a protective cavity, called the magnetosphere. A bow shock forms at about 10 Earth radii on the sunlit side of our planet. Its location is highly variable since it is pushed in and out by the gusty solar wind. The magnetopause marks the outer boundary of the magnetosphere, at the place where the solar wind takes control of the motions of charged particles. The solar wind is deflected around the Earth, pulling the terrestrial magnetic field into a long magnetotail on the night side. Plasma in the solar wind is deflected at the bow shock (left), flows along the magnetopause into the magnetic tail (right), and is then injected back towards the Earth and Sun within the plasma sheet (centre). The Earth, its auroras, atmosphere, and ionosphere and the two Van Allen radiation belts all lie within this magnetic cocoon.

geomagnetic storm, provided that the magnetic alignment is right. The Earth's field is generally directed northwards in the outer dayside magnetosphere, so a fast coronal mass ejection is more likely to merge and connect with the terrestrial field if it points in the opposite southward direction.

Moderate Geomagnetic Activity

Moderate mid-latitude magnetic fluctuations of about 0.1%, or tens of nanoTesla, last a few hours, and they are most noticeable near the minimum of the 11-year solar activity cycle. They have a 27-day repetition period, corresponding to the rotation period of the Sun at low solar latitudes when viewed from the moving Earth.

The recurrent activity is linked to long-lived, high-speed streams in the solar wind that emanate from coronal holes. When this fast wind overtakes the slow-speed, equatorial one, the two wind components interact, producing shock waves and intense magnetic fields that rotate with the Sun, and periodically sweep past the Earth, producing moderate geomagnetic activity every 27 days.

The Auroras

The northern or southern lights, named the 'aurora borealis' and 'aurora australis' in Latin, are one of the most magnificent and earliest-known examples of solar–terrestrial interaction. They illuminate the cold, dark Arctic and Antarctic skies with curtains of green and red light that flicker across the night sky far above the highest clouds.

Spacecraft look down on the auroras from high above, showing an oval centred at each magnetic pole (Figure 7). An observer on the ground sees only a small, changing piece of the aurora oval.

The reason that auroras are usually located near the polar regions is that the Earth's magnetic fields guide energetic electrons there. The high-speed electrons move down along the Earth's magnetic field lines into the upper polar atmosphere, colliding with oxygen and nitrogen. The pumped-up atoms or molecules fluoresce, giving up the energy acquired from the electrons and emitting a burst of light.

The electrons that cause the auroras come from the Earth's magnetic tail and are also energized locally within the magnetosphere. The rare, bright, auroras seen at low terrestrial latitudes only become visible during very intense geomagnetic storms.

Space Weather

Down here on the ground, we are shielded from the direct onslaught of solar explosions and the solar wind by the Earth's atmosphere and magnetic fields, but out in deep space there is no protection. Energetic charged particles hurled out from intense solar flares

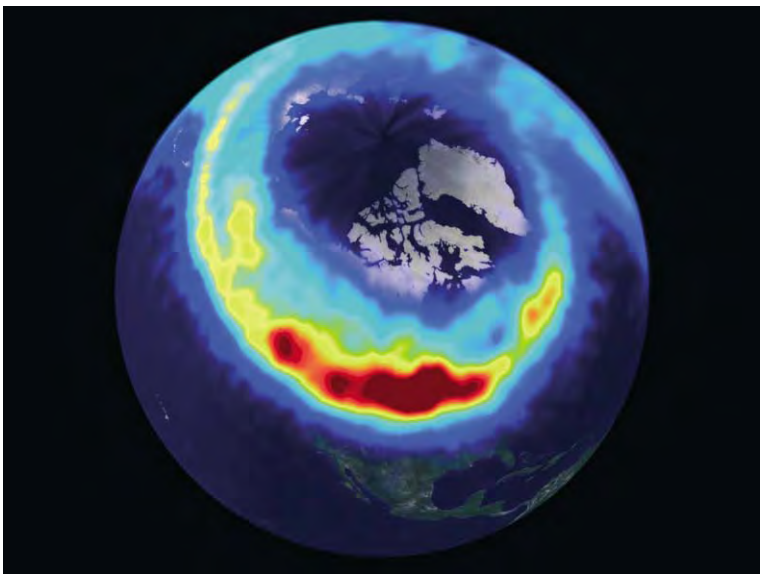


Figure 7 The auroral oval. The *POLAR* spacecraft looks down on an aurora from high above the Earth's north polar region on 22 October 1999, showing the northern lights in their entirety. The glowing oval, imaged in ultraviolet light, is 4500 km across. The most intense aurora activity appears in bright red or yellow. It is typically produced by magnetic reconnection events in the Earth's magnetotail, on the night side of the Earth. Courtesy of the Visible Imaging System, University of Iowa and NASA.

can seriously damage satellites, including their solar cells and electronic components, and even kill an unprotected astronaut.

The high-speed protons and electrons follow a narrow, curved path once they leave the Sun, guided by the spiral structure of the interplanetary magnetic field, so they must be emitted from active regions near the west limb and the solar equator to be magnetically connected with the Earth. Solar flares emitted from other places on the Sun are not likely to hit Earth, but they could be headed towards interplanetary spacecraft, the Moon, Mars, or other planets. The most energetic flare particles can travel from the Sun to the Earth in just 8 minutes, moving at nearly the velocity of light.

Coronal mass ejections move straight out of the Sun, energizing particles over large regions in interplanetary space. Mass ejections are most likely to hit the Earth if they originate near the centre of the solar disk, as viewed from the Earth, and are sent directly towards the planet. They take about 4 days to travel from the Sun to the Earth, moving at a typical speed of about 400 km s $^{-1}$.

The strong blast of X-rays and ultraviolet radiation from a solar flare alters the Earth's atmosphere, transforming the ionosphere, which reflects radio waves to distant locations on Earth. During moderately intense flares, radio communications can be silenced over the Earth's entire sunlit hemisphere, disrupting contact with airplanes flying over oceans or remote countries.

The enhanced ultraviolet and X-ray radiation from solar flares also heats the atmosphere and causes it to expand, and similar or greater effects are caused by coronal mass ejections that produce major geomagnetic storms. The expansion of the terrestrial atmosphere brings higher densities to a given altitude, increasing the drag exerted on a satellite, pulling it to a lower altitude, and causing a premature and fatal spiral towards the Earth.

When a coronal mass ejection slams into the Earth, the force of impact can push the bow shock, at the dayside of the magnetosphere, down to half its usual distance of about 10 times the Earth's radius, compressing the magnetosphere below the orbits of geosynchronous satellites and exposing them to the damaging effects of the full brunt of the gusty solar wind.

During an intense geomagnetic storm, associated with a colliding coronal mass ejection, strong electric currents flow in the ionosphere. They induce potential differences in the ground below them, and produce strong currents in any long conductor such as a power line. These currents can blow circuit breakers, overheat and melt the windings of transformers, and cause massive failures of electrical distribution systems. A coronal mass ejection can thereby plunge major urban centres, like New York City or Montreal, into complete darkness, causing social chaos and threatening safety.

Our technological society has become so vulnerable to the potential devastation of these storms in space that national centres employ space weather

forecasters to continuously monitor the Sun from ground and space to warn of threatening solar activity.

See Also

Earth: Orbital Variation (Including Milankovitch Cycles). **Gaia**. **Magnetostratigraphy**. **Palaeoclimates**. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Bone N (1996) *The Aurora: Sun Earth Interactions*. New York: Wiley.

Calowicz MJ and Lopez RE (2002) *Storms from the Sun: The Emerging Science of Space Weather*. Washington, DC: Joseph Henry Press.

Golub L and Pasachoff JM (2001) *Nearest Star: The Surprising Science of Our Sun*. Cambridge, MA: Harvard University Press.

Kaler JB (1992) *Stars*. Scientific American Library. New York: WH Freeman.

Lang KR (1995) *Sun, Earth and Sky*. New York: Springer Verlag.

Lang KR (2000) *The Sun from Space*. New York: Springer Verlag.

Lang KR (2001) *The Cambridge Encyclopedia of the Sun*. New York: Cambridge University Press.

Lang KR (2003) *The Cambridge Guide to the Solar System*. New York: Cambridge University Press.

Odenwald S (2001) *The 23rd Cycle: Learning to Live with a Stormy Sun*. New York: Columbia University Press.

Phillips KJH (1992) *Guide to the Sun*. New York: Cambridge University Press.

Asteroids, Comets and Space Dust

P Moore, Selsey, UK

Introduction

Asteroids and comets must be regarded as minor members of the Solar System. Compared with planets they are of very low mass, and they have even been referred to as cosmic debris. The asteroids, dwarf worlds most of which are well below 1000 km in diameter, are found mainly between the orbits of Mars and Jupiter, though some stray from this 'main belt'; comets have been described as 'dirty snowballs', and though they may become very conspicuous in the sky they are very insubstantial. This article reviews the asteroids and comets, together with the large amount of thinly-spread material lying in the Solar System.

Distribution of the Asteroids

The Solar System is divided into two well-defined parts. There are four relatively small, rocky planets: Mercury, Venus, the Earth, and Mars. Then come the four giants: Jupiter, Saturn, Uranus, and Neptune. Between the orbits of Mars and Jupiter thousands of asteroids, otherwise known as minor planets, make up what is known as the main belt (Figure 1). Of the main belt asteroids, only one (Ceres) is as much as 900 km in diameter, and only one (Vesta) is ever visible with the naked eye. Some of the larger main belt asteroids are listed in Table 1.

Some small asteroids can leave the main belt, and swing closer to the Sun; they may even approach the Earth, and collision cannot be ruled out (it may even be that the impact of an asteroid, some 65 million years ago in Mexico, caused a climatic change and mass extinction, which included the dinosaurs). All of these Near Earth Approach (NEA) asteroids are very small indeed. There are asteroids known as Trojans which share the orbits of major planets; others have very eccentric orbits which take them into the far reaches of the Solar System, and recently it has been found that there are asteroid-sized bodies near and

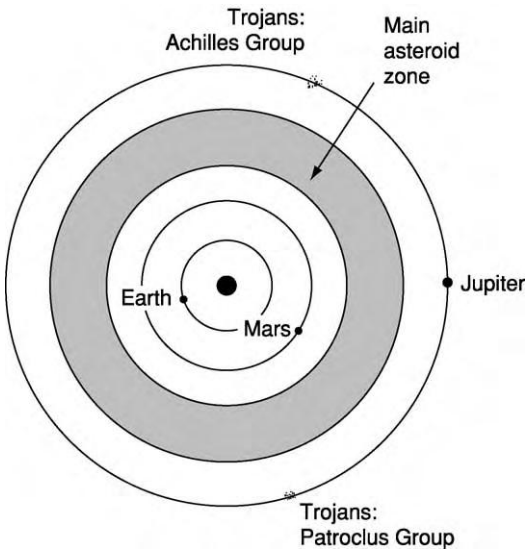


Figure 1 Distribution of asteroids.

Table 1 Some of the larger Main Belt ateroids

Asteroid		*q*	*Q*	*Period, years*	*Orbital eccentricity*	*Orbital inclination*	*T*	*Diameter, km (max)*	*M*
1	Ceres	2.55	2.77	4.60	0.078	10.60	C	960	7.4
2	Pallas	2.12	2.77	4.62	0.234	34.80	CU	571	8.0
3	Juno	1.98	2.67	4.36	0.258	13.00	S	288	8.7
4	Vesta	2.15	2.37	3.63	0.090	7.14	V	525	6.5
5	Astrea	2.08	2.57	4.13	0.190	5.36	S	120	9.8
6	Hebe	1.94	2.43	3.77	0.202	14.79	S	204	8.3
7	Iris	1.84	2.39	5.51	0.229	5.51	S	208	7.8
8	Flora	1.86	2.20	3.27	0.156	5.89	S	162	8.7
9	Metis	2.10	2.39	3.69	0.121	5.59	S	158	9.1
10	Hygeia	2.76	3.13	5.54	0.120	3.84	C	430	10.2
72	Feronia	1.99	2.67	3.41	0.120	5.42	U	96	12.0
87	Sylvia	3.19	3.48	6.50	0.083	10.87	P	282	12.6
253	Mathilde	1.94	3.35	5.61	0.262	6.70	C	66	10.0
153	Hilda	3.10	3.97	7.91	0.142	7.83	P	222	13.3
279	Thule	4.22	4.27	8.23	0.011	8.23	D	130	15.4
511	Davida	2.61	3.18	5.66	0.177	15.93	C	324	10.5
704	Interamnia	2.61	3.06	5.36	0.148	17.30	D	338	11.0

q perihelion distance, in astronomical units.
Q aphelion distance, in astronomical units.
M mean magnitude at opposition.
T type (see **Table 2**).

well beyond the orbits of Neptune and Pluto. These make up what is known as the Kuiper Belt.

Discovery

A mathematical relationship, known as Bode's Law, led astronomers to believe that there should be another planet moving between the orbits of Mars and Jupiter. From 1800, a systematic search was carried out by a group of observers who called themselves the 'Celestial Police', and on 1 January 1801, the first asteroid, Ceres, was discovered by G Piazzi (who was not then a member of the group, though he joined later). Three more small bodies were found during the next few years: Pallas, Juno, and Vesta. It was not until 1845 that the next asteroid, Astræa, was found; others followed, and by now many thousands are known. When a new asteroid is discovered, it is given a temporary designation and then, when its orbit has been reliably worked out, a number. At first mythological names were used, but the supply of these names soon ran out; today the discoverer is entitled to suggest a name, which is then ratified by the International Astronomical Union.

Origin and Orbits

It is no longer thought that the asteroids are fragments of a large planet which broke up. It seems that no planet of appreciable size could form in this part of the Solar System, because of the disruptive influence of Jupiter. The asteroids in the main belt are not evenly distributed; Jupiter's gravitational pull tends to produce groups or 'families', made up of numbers of asteroids moving at around the same distance from the Sun (**Figure 2**). A family is named after one of its most prominent members, and does seem to be due to the disruption of a larger body, The Flora family has at least 400 members. There are also gaps in the main belt (the Kirkwood Gaps) which are almost empty, because of regular gravitational interactions with Jupiter. For instance, there is a gap at a distance of 375 million km from the Sun, where an asteroid would complete three orbits for every one orbit of Jupiter (**Figure 3**).

Types of Asteroids

Asteroids are divided into various types, according to their physical and surface characteristics. The main types are listed in **Table 2** (**Figure 4**). There is certainly a link between comets and small asteroids; thus a tailed comet discovered in 1951 (Wilson-Harrington) was lost for years before being recovered in 1979 in the guise of an asteroid. It was given a number (4015) and now shows no sign of cometary activity.

Asteroid Surfaces and Composition

Details on some asteroids have been recorded. 3 Vesta has been imaged by the Hubble Space Telescope, and

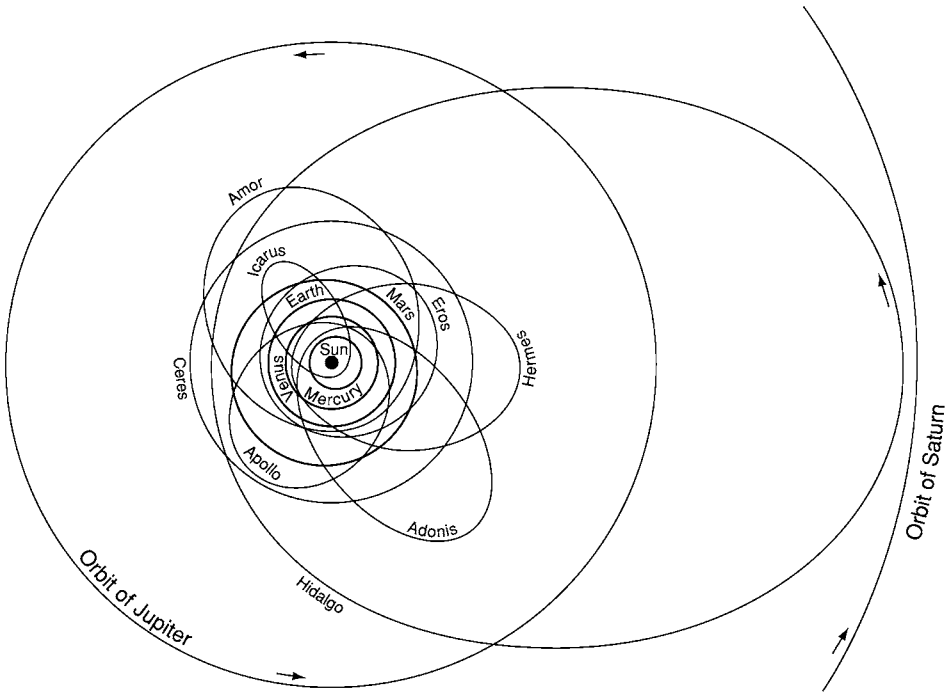


Figure 2 Orbits of some asteroids.

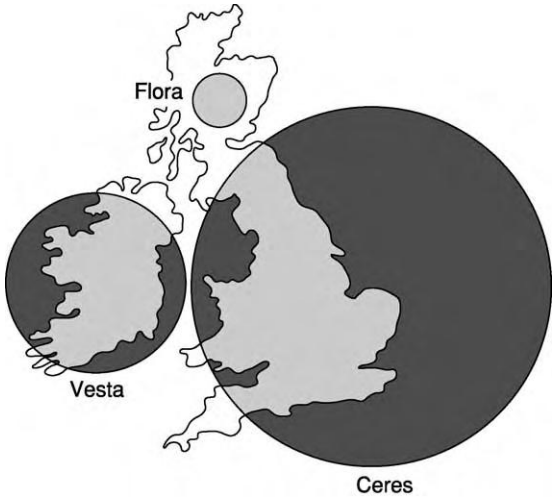


Figure 3 Sizes of same asteroids compared with British Isles Diameters: Ceres 970 km Vesta 288 × 230, Flora 204.

Table 2 Types of asteroids

C (Carbonaceous)	Most numerous, increasing in number from 10% at 2.2 a.u. up to 80% at 3 a.u. Low albedo; spectra resembles carbonaceous chondrites
S (Silicaceous)	Most numerous in the inner part of the main zone. Generally reddish, spectra resemble those of chondrites
M (Metallic)	Moderate albedoes; may be the metal rich cores of larger parent bodies
E (Enstatite)	Relatively rare, high albedos; enstatite ($MgSiO_3$), is a major constituent
D	Low albedo; reddish; surfaces are 90% clays, with magnetite and carbon rich substances
A	Almost pure olivine
P	Dark and reddish; not unlike Type B.
V	Igneous rock surfaces, very rare; 4 Vesta is the only large example
U	Asteroids which are regarded as unclassifiable

is geologically of great interest; there are two distinct hemispheres, covered with different types of solidified lava, and there is one huge impact crater. Some asteroids have been imaged from passing space-craft; 253 Mathilde (Figure 5) is very dark and irregular, and has been described as 'a heap of rubble', while 243 Ida is cratered and is accompanied by a tiny satellite, Dactyl. 216 Kleopatra has two lobes of similar size, and looks remarkably like a dog's bone!

Asteroids closer-in than the Main Belt

These are of various types. Details are given in Table 3. All are small, usually only a few kilometres across, and are irregular in shape. The first to be discovered (in 1898) was 433 Eros; it is an Amor asteroid, so that its orbit crosses that of Mars but not that of the Earth. It can approach Earth at a distance of 23 million km. On 12 February 2001,

Figure 4 511 Davida, a Main Belt asteroid 320 km in diameter. This sequence of images was taken at the WM Keck Observatory on 28 December 2002, The rotation period is just over one hour; here Davida is seen from above its north pole as it spins counter clockwise.

Figure 5 Asteroid 253 Mathilde, imaged by NEAR space craft on 27 June 1997, from a range of 2400 km. There are large craters. The asteroid is very dark, with an average albedo of 4%. Mathil de's diameter is 50 × 50 × 70 km, rotation period 418 hours.

the space-craft NEAR-Shoemaker made a controlled landing on it; Eros proved to be a very primitive body, (Figure 2) and very ancient. Craters were plentiful, as well as rocks and boulders of all kinds, and superficial 'landslides' in the surface material were recognized.

Some small asteroids pass between the Earth and the Moon, and collision cannot be ruled out, and there are more potentially hazardous asteroids (PHAs) than used to be thought.

Asteroids Beyond the Main Belt

The Trojan asteroids move in the same orbit as Jupiter, though they keep either well ahead of or well behind the Giant Planet and are in no danger of being engulfed. Mars has several Trojans, and Neptune one. No true Earth Trojans are known, though 3753 Cruithne has almost the same orbital period and describes a curious sort of 'horseshoe' path with respect to the Earth. There are also asteroids, such as 944 Hidalgo and 5335 Damocles, with very eccentric orbits, very like those of comets. For example, Damocles has a period of 40.9 years; its orbit crosses those of Mars, Jupiter, Saturn, and Uranus, but is in no danger of collision as its orbital inclination is 61°. It is no more than 15 km in diameter.

Table 3 Asteroids closer in than the Main Belt

Apohele type	Orbit entirely within that of the Earth, only one example is known, the tiny 2003 CR20
Aten type	Average distance from the Sun less than 1 a.u., though they may cross the Earth's orbit. All very small
Apollo type	Orbits cross that of the Earth; average distance from the Sun over 1 a.u.
Amor class	Orbits cross that of Mars, but not that of the Earth

The 'Centaur' asteroids remain well beyond the Main Belt; the first to be found (in 1977) was 2060 Chiron, which moves mainly between the orbits of Saturn and Uranus, in a period of 50 years. It shows traces of a coma at times, but seems much too large to be classed as a comet, even though it has been given a cometary number.

The Kuiper Belt

Many asteroidal bodies have been found near and beyond the orbits of Neptune and Pluto; the existence of such a belt was suggested by GP Kuiper (and earlier, less positively, by K. Edgeworth). Some are larger than any Main Belt asteroids; 50 000 Quaoar has a diameter of about 250 km, more than half that of Pluto. Other large Kuiper Belt objects are 28 978 Ixion (1200 km), 20 000 Varuna (900 km), and 38 093 Rhadamanthus (320 km). There are also asteroid-sized bodies which recede to immense distances from the Sun, and have orbital periods of hundreds of years. There are excellent reasons for suggesting that Pluto should be regarded as merely an exceptionally large Kuiper Belt object rather than as a bonafide planet. The Kuiper belt also includes some comets.

Comets

Comets are the most erratic members of the Solar System. They were once regarded as unlucky, and descriptions of them go back for thousands of years. Certainly a brilliant comet may look really spectacular, but by planetary standards all comets are of very low mass. They are true members of the Solar System, but in general their orbits are very eccentric, and their movements are quite unlike those of the planets.

Nature of Comets

The only fairly substantial part of a comet is the nucleus, made up of rocky fragments held together by frozen ices such as H_2O methane, carbon dioxide,

and ammonia. When a comet is warmed as it approaches perihelion the rise in temperature leads to evaporation, so that the comet develops a head or coma, often with a tail or tails. Cometary tails always point away from the Sun, and are of two main types ion and dust tails. A gas or ion tail is due to particles being repelled by the solar wind, while with a dust tail the particles are driven out by the pressure of sunlight; this means that when a comet is moving outward, after perihelion, it travels tail-first. However, not all comets develop tails of any kind, and even a large comet will lose its tail when it has receded into the far part of the Solar System.

Nomenclature

Traditionally, a comet is named after its discoverer or discoverers; thus the brilliant comet seen in 1995 and 1996 was known as Hale-Bopp, since it was found independently by two American observers, Alan Hale and Tom Bopp. Occasionally the comet is known by the name of the mathematician who first computed its orbit, as with Comets Halley and Encke. There is also an official numbering system which relates to the date of discovery.

Orbits

Many comets have short periods – for example 3.3 years for Comet Encke. These short-period comets can be predicted, and some can be followed all round their orbits. Many have their aphelia near the distance of the orbit of Jupiter, making up what is termed Jupiter's comet family. Most of them are faint, and few attain naked-eye visibility. The only reasonably bright comet with a period of less than 100 years is Halley's (76 years), which last returned to perihelion in 1986–1987.

Long-period comets recede to great distances, and since their periods amount to many centuries they cannot be predicted, Hale-Bopp (Figure 6) will be back in about 2350 years, but for the next return of Comet Hyakutake, which was bright for a few weeks in 1996, we must wait for around 14 000 years. These orbits are almost parabolic, and indeed some comets are thrown into parabolic orbits after passing perihelion, so that they will never return. Arend-Rolànd, the bright comet of April 1957, is one example of this.

Figure 6 Comet Hale Bopp, 1997. Note the straight ion tail, and the curved dust tail. This was the most spectacular comet for many years.

Origin of Comets

It seems that short-period comets come from the Kuiper Belt. In general, their orbits are not highly inclined to the ecliptic, though some, notably Halley's Comet, have retrograde motion. Comets of much longer period are thought to come from the Oort Cloud, a huge spherical cloud of debris surrounding the Sun at a distance of over one light-year; its existence was suggested in 1950 by the Dutch astronomer JH Oort. It is, of course, unobservable from Earth. If an Oort Cloud comet is perturbed for any reason, it may swing in towards the Sun; it may then be perturbed into a short-period orbit, it may fall into the Sun and be destroyed, or it may simply return to the Oort Cloud. The orbital inclinations may be very high, and many long-period comets have retrograde motion.

It may be that the Oort Cloud comets were formed closer to the Sun than the Kuiper Belt objects. Low-mass objects formed near the giant planets would have been ejected by gravitational encounters. While Kuiper Belt objects, formed further out, were not so affected. Details of some notable comets are given in Table 4 (*see* **Solar System:** Meteorites).

Comets and Meteors

As a comet moves, it leaves a 'dusty trail', and if the Earth passes through such a trail we see a meteor shower. In many cases the parent comets are identifiable, for example the Orionid meteors, seen every October, come from Halley's Comet, while the August Perseids come from Comet Swift–Tuttle.

Some comets have been seen to disintegrate; thus Biela's Comet, which had a period of 6.6 years, broke in two during its return in 1846, and has not been seen since 1852, though for many years meteors appeared from the position where the comet ought to have been. Other periodical comets have been lost because their orbits have been so violently perturbed by planetary encounters. One comet, Shoemaker-Levy 9, in captured orbit around Jupiter, was destroyed in 1994 when it impacted Jupiter.

Halley's Comet

Named for Edmond Halley, who observed it in 1682 and was the first to realize that it was periodical

Table 4 Some Notable Comets

Periodical Comets

Comet		*P*	*q*	*Q*	*E*	*I*	*M*
2	Encke	3.28	0.33	2.21	0.850	11.9	11
26	Grigg Skjellerup	5.10	0.99	2.96	0.664	6.6	12
10	Tempel 2	5.47	1.48	3.10	0.552	12.0	10
46	Wirtanen	5.46	1.07	3.10	0.657	11.7	16
9	Tempel 1	5.51	1.50	3.12	0.502	10.5	9
7	Pons Winnecke	6.37	1.26	3.44	0.634	22.3	14
6	D'Arrest	6.51	1.35	3.49	0.614	19.5	6
21	Giacobini Zinner	6.61	1.03	3.52	0.706	31.9	10
19	Borrelly	6.80	1.37	3.59	0.623	30.2	13
15	Finlay	6.95	1.09	3.64	0.699	3.7	13
4	Faye	7.34	1.59	3.78	0.578	9.1	8
36	Whipple	8.53	3.09	4.17	0.239	9.9	9
8	Tuttle	13.51	0.997	5.67	0.824	54.7	8
27	Crommelin	27.4	0.74	17.4	0.919	19.1	11
13	Olbers	69.6	1.18	32.6	0.930	44.6	5
1	Halley	76.0	0.59	35.3	0.967	162.2	4
109	Swift Tuttle	135.0	0.96	51.7	0.964	113.4	4
153	Ikeya Zhang	341	0.51	60	0.99	28.1	5

q perihelion distance, astronomical units.
Q aphelion distance, astronomical units.
E orbital eccentricity.
I orbital inclination, degrees.
M absolute magnitude (the magnitude which the comet would have if seen from a distance of 1 a.u. from the Sun and 1 a.u. from the Earth.)
P period, years.

(Figure 7). It was probably record by the Chinese as early as 1059 BC, and since 240 BC it has been seen at every return; it came to perihelion in 1066, and is shown in the famous Bayeux Tapestry. During the 1986 return several space–craft were sent to it, and one of these, Giotto, passed within 605 km of the nucleus. The nucleus was shaped rather like a peanut, and measured 15 × 8 × 8 km. Over 60 000 million comets of this mass would be needed to equal the Earth. The nucleus was dark-coated, and made up largely of water ice; dust-jets were active, though only from a small area on the sunward wide (Figure 8). The comet is now too faint to be detected, though it should be recovered before the next perihelion passage, due in 2061.

Figure 7 Halley's Comet, March 1986, (Photo by Tom Polaks with a 100 mm lens at f/2.8.) The faint globular cluster M75 is also shown. From the most left of the three conspicuous stars left and above Halley's head, go to the fainter star above and left. This star forms a fainter, nearly rectangular triangle with the other stars above and left of it. On the line connecting with the far left edge a star like spot anneals; this is M75.

Great Comets

Really brilliant comets were seen fairly frequently during the nineteenth century, but were less common in the twentieth century (Figure 9). The brightest comet of near-modern times was probably that of 1843, which cast shadows and was visible in broad daylight. The last two really spectacular comets were those of 1910 – the Daylight Comet, seen shortly before Halley's – and 1965 (Ikeya–Seki), which faded quickly. Its period has been given as 880 years. Some Great comets are listed in Table 5.

Comet Hale-Bopp was not so brilliant as these, but was exceptionally beautiful, and was visible with the naked eye for over a year, from July 1996 to October 1997. It was enormous by cometary standards, with a 40 km nucleus, but unfortunately it did not come close to the Earth. There were both ion and dust tails, plus a third inconspicuous tail made up of

sodium. It was last at perihelion about 4200 years ago, but planetary perturbations mean that it should return in about 2350 years, though of course all periods of this kind of length cannot be given accurately. Its orbital inclination is 89°, so that its path lies at almost a right angle to that of the Earth. During its period of visibility there were marked changes in the tails, and a spiral structure developed in the coma. Comet Ikeya-Zhang of 2002 was much less striking – it became no brighter than the fourth magnitude – but is notable because it was found to be a return of the comet of 1661, and is therefore the longest-period comet to have been seen at more than one apparition. It will be back once more in 2343.

Life in Comets?

The 'panspermia' theory was due to the Swedish scientist Svants Arrhenius, whose work won him the Nobel Prize for Chemistry in 1903. Arrhenius believed that life on Earth was brought here in a meteorite, but the theory never became popular, because it seemed to raise more problems that it solved. The same sort of theme has been followed up much more recently by Sir Fred Hoyle and C Wickramasinghe,

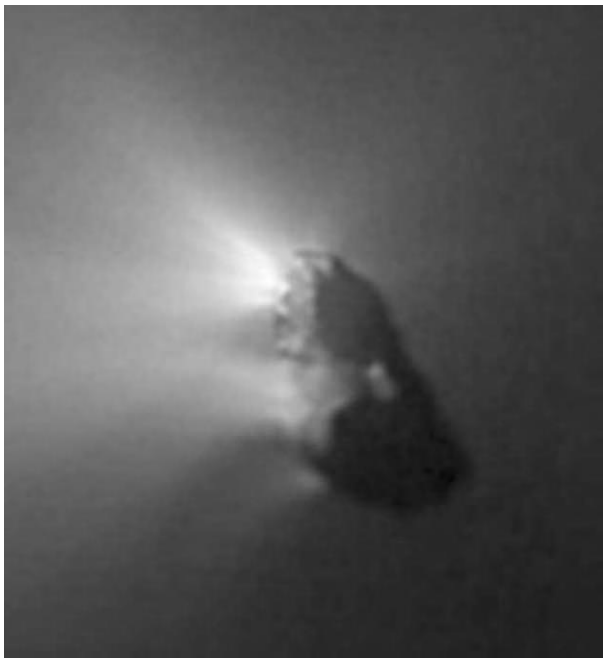


Figure 8 Head of Halley's Comet, imaged from the Giotto space craft. The dark coating and the active dust jets are well seen. (Photograph from the HMC [Halley Multi colour Camera]), Giotto passed 605 km from the nucleus on the night of, 13 14 March 1986.

Figure 9 Comet Hyakitake, C/1996 B2. This beautiful comet was conspicuous object briefly in April May 1996; it was obvi ously greenish, and had a long tail. It was in fact a small comet, but made a fairly close approach to the Earth. It will next come to perihelion in 14 000 years! time, look out for it then.

Table 5 Some Great Comets

Year	*Name*	
1744	de Chéseaux	Multi tailed comet; max. magnitude 7
1811	Flaugergues	Mag. 0; 24 degree ion tail. Period 3096 years
1843	Great Comet	Mag. 6. Sun grazer. Period 517 years
1858	Donati	Mag, 1. Most beautiful of all comets, with ion and dust tails. Period 1951 years
1882	Great Comet	Reached mag. 4. Period 760 years
1910	Daylight Comet	Magnitude 4. Immensely long period
1927	Skjellerup Maristany	Magnitude 6; 35 degree tail
1947	Southern Comet	Magnitude 5, 25 degree tail
1965	Ikeya Seki	Magnitude 10; seen very near the Sun
1976	West	Magnitude 2. Multiple dust tail
1996	Hyakutake	Briefly reaches magnitude Q. Very long tail. The green comet
1997	Hale Bopp	Magnitude 0.5; naked eye object for over a year

Figure 10 The Zodiacal Light. A typical display, photographed on 19 November 1998 over the Qinghai Radio Observatory near Delinghom Qinghai, Central China. (M Langbroek).

who claimed that comets can actually deposit harmful bacteria in the Earth's upper atmosphere, causing epidemics. Again there has been little support.

Space Dust

There is a great quantity of 'dust' in the Solar System, particularly near the main plane. It is the cause of the Zodiacal Light, (Figure 10) which may be seen as a cone of light extending upwards from the horizon for a fairly brief period either after sunset or before sunrise. Since it extends along the ecliptic, it is best seen when the ecliptic is nearly vertical to the horizon, in February to March and again in September–October. Cometary debris is a major contributary factor. It was first correctly explained by the Italian astronomer, GS Cassini, in 1683.

Another glow due to cosmic dust is the Gegenschein, seen as a faint patch exactly opposite to the Sun in the sky. It is extremely elusive, and is visible only under near-ideal conditions. The best opportunities occur when the anti-Sun position is well away from the Milky Way, from February to April and from September to November. Generally it is oval in shape, measuring about 10° by 22°, so that its maximum diameter is 40 times that of the full moon.

The Zodiacal Band is a very dim, parallel-sided band of radiance which may extend to either side of the Gegenschein, or prolonged from the apex of the Zodiacal Light Cone to join the Zodiacal Light with the Gegenschein. It also is due to sunlight being reflected from interplanetary particles near the main plane of the Solar System.

See Also

Solar System: The Sun; Meteorites; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus.

Further Reading

Bone N (1986) *Meteors*. London: Philip.
Bhandt G and Chapman D (1982) *Introduction to Comets*. Cambridge: Cambridge University Press.
Burnfam R (2000) *Great Comets*. Cambridge: Cambridge University Press.
Krishna S (1997) *Physics of Comets*. Singapore: World Scientific.
Kronk G (1988) *Comet Catalogue*. Enslow: Hillside NJ and Aldershot.
Kronk G (1988) *Meteor Showers*. Enslow: Hillside NJ and Aldershot.
Moore P (2001) *Astronomy Data Book*. London: Institute of Physics, Publishing.
Moore P (2003) *Atlas of the Universe*. London: Philip.
Norton CR (1992) *Rooks from Space*. Montana: USA Mountain Press Publishing.
Schmadel L (2002) *Dictionary of Minor Planet Names*. Berlin, Heidelberg, New York: Springer verlag.
Kowal CT (1996) *Asteroids*. Wiley.
Whipple FL (1985) *The Mystery of Comets*. Cambridge: Cambridge University Press.
Yeomans K (1991) *Comets*. New York: Wiley Science Editions.

Meteorites

G J H McCall, Cirencester, Gloucester, UK

Introduction

Meteorites are bodies of metal or stony material mixed with metal which fall to Earth in sporadic and random arrival events, characterized by entry of a fireball or bolide streaking, often with punctuated explosive bursts, through the sky on their frictional passage through the Earth's atmosphere (Figure 1). The history of the gradual scientific acceptance of the reality of such events is followed by a brief description of the classification of various types of meteorite; the four age and time interval measurements significant for any meteorite; and the known or likely provenance in the bodies of the Solar System of the various types are then considered. After a brief mention of impact cratering and tektites, and also 'fossil' meteorites enclosed in ancient rocks, an account is given of the revolution in 'Meteoritics' (essentially an extension of geology, geochemistry, metallurgy, and physics into the realms of astronomy and planetology) during the latter half of the twentieth century. This is a result of space exploration and the recognition of hitherto unknown optimum collection regions (icebound Antarctica; the Nullarbor Plain, Australia; and other desert regions). Examples of some extensions of research into meteoritics in modern state-of-the-art science are listed.

Figure 1 A painting by P.V. Medvedev of the fireball accompanying the Sikhot Alin fall of 1949 (reproduced from McCall 1973).

Historical: the Fall of Stones and Metal from the Sky

Records of shooting stars, bright objects seen to dart across the night sky, go back to Egyptian papyrus writings of *ca*. 2000 BC and records of actual meteorites falling to Earth out of the sky go back almost as far – the fall of a black stone in the form of a cone, circular below and ending in an apex above, was reported in Phrygia about 652 BC, the familiar image of a stony meteorite such as the Middlesborough Meteorite (Figure 2) coming to us from the distant past. The Parian chronicle records falls of stones in Crete in 1478 BC and in 1168 BC of iron. In 618 BC, a fall of stones is reported to have broken

Figure 2 The Middlesborough, England, stone (fell 1881) showing the dark fusion crust and anterior surface in flight, the apex of the cone being in the direction of flight and the radiating flutings being produced by atmospheric ablation (from McCall 1973).

several chariots and killed ten men, a unique fatality. The sacred stone built into the Kaaba at Mecca is reported to have been long known prior to Islam and to have fallen from the sky. Such falls were given a religious significance, and officers of the Geological Survey in India had to go hot-foot to the site of a fall or the mass was either enshrined or broken into pieces to release evil spirits. American Indians confused later scientists by transporting masses long distances and burying them in cysts. Particularly pleasing is the custom in mediaeval France of chaining meteorites up to prevent them departing as swiftly as they arrived or from wandering at night. The earliest material from a fall preserved in western Europe is believed to be at Ensisheim, Alsace, stored in the local church since it fell in AD 1462.

Despite all these early records (and there are many more, in particular from Russia and China), scientists were slow to accept the process of rocky or metal material falling from the sky. Though there are records of the finds of irons and the falls of stones much earlier and the problem had been solved – Diogenes of Apollonia wrote "meteors are invisible stars that fall to Earth and die out, like the fiery, stony star that fall to Earth near the Egos Potamos River (in 465 BC): and natives in northern Argentina had led the conquistadors to buried masses of exotic iron, of supposed celestial origin in 1576 – scientific acceptance was widely achieved only in the last years of the eighteenth century Age of Enlightenment and the earliest years of the nineteenth century, with natives leading the explorer Pallas in Siberia to a buried stony-iron mass reputedly fallen from the sky; also falls were followed by material recovery at Wold Cottage, near Scarborough, Yorkshire and L/Aigle France.

The fall at Albareto, Italy, in 1766, had been well described by the Abbé Dominico Troili, but dismissed as the product of a subterranean explosion which hurled it high in the sky from a vent in the Earth. The stone which fell at Lucé, France in 1768, the first to be chemically analysed, was dismissed as neither from thunder, nor fallen from the sky, but as a piece of pyritiferous sandstone by a panel of august scientists! So it was the Pallas stony-iron meteorite (700 kg, 'Krasnojarsk'), the subject, together with the Otumpa iron from South America, of a book published by E.F.F. Chladni in Riga in 1794, which really established the scientific reality of meteorite falls. Both were exotic, being found far from any known volcanic province, and by a process of elimination, he reached a single possible answer and further connected them with the phenomenon of fireball meteors. Russian scientific circles were distant from western Europe, and the English were really convinced only by the fall of a stone at Wold Cottage near Scarborough in 1795. This came into the possession of Joseph Banks, who recognized the similarity of the black fusion crust to the Siena fall material of 1794 in his possession. Edward Howard studied both and the presentation of his findings to the Royal Society in 1802–1803 convinced sceptics in England. Presentation to the Institut de France convinced several important scientists, but resistance to the idea was not finally overcome in that country until 3000 stones showered down on L'Aigle, Normandy and were described by Biot. Chladni's work then received belated international acknowledgement, but decades would elapse before the connection with fireballs was completely established and a century before the origin of most of them through impacts between asteroids would be established.

Classification

The classification of meteorites has developed over the years and some new types and revisions of the system have inevitably arisen in the last half of the twentieth century with the prolific collection from optimum Antarctic and desert regions; despite this, the system remains workable though some revision might in time be necessary.

There are three principal classes:

Irons (**Figures 3, 4**)
Stony-irons (**Figure 5**)
Stones (**Figure 2**)

The latter are subdivided into (i) Chondrites, which display rounded bodies (chondrules) (**Figures 6, 7**),

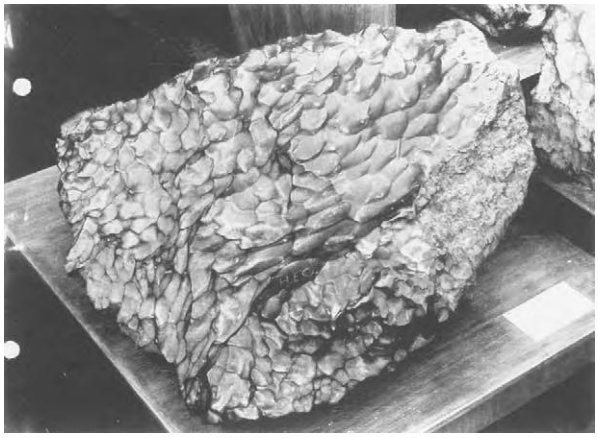


Figure 3 The Haig, Western Australia, iron (find 1951, 480 kg, III AB) with typical hackly markings on the surface.

Figure 4 Cut and etched surface of the Mount Edith iron, Western Australia (find 1913, 160 kg, III AB) showing the Widmanstatten pattern and dark troilite (sulphide) nodules.

Figure 5 Cut surface of the Brenham, Kansas, pallasite stony iron (find 1962, 22 and 9 kg), showing nickel iron (light grey) and olivine (dark) (from McCall 1973).

believed to be relics of a very early accretionary stage in the formation of the asteroidal parent bodies (the chondrules may be wholly obliterated by recrystallization); and (ii) Achondrites, without chondrules, having textures resembling those of terrestrial igneous rocks (Figure 12). The classification used worldwide, as at 2003, is shown in Table 1a and 1b and the statistics of meteorite falls and finds in Table 2.

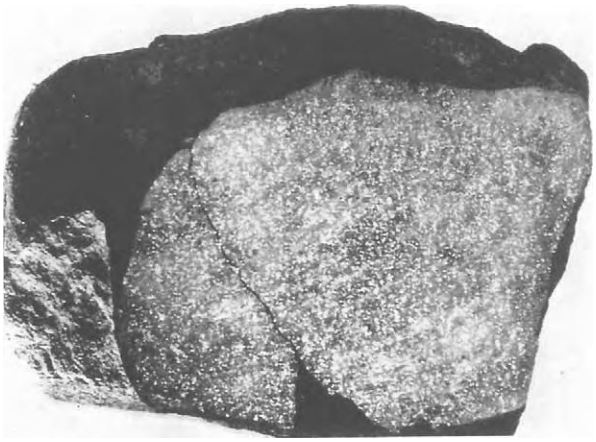


Figure 6 The Cocklebiddy, Western Australia, ordinary chondrite (fall 1949, 0.794 kg), cut face showing specks of light grey nickel iron disseminated in a dark grey silicate matrix: the rounded chondrules are microscopic and thus not visible (from McCall 1973).

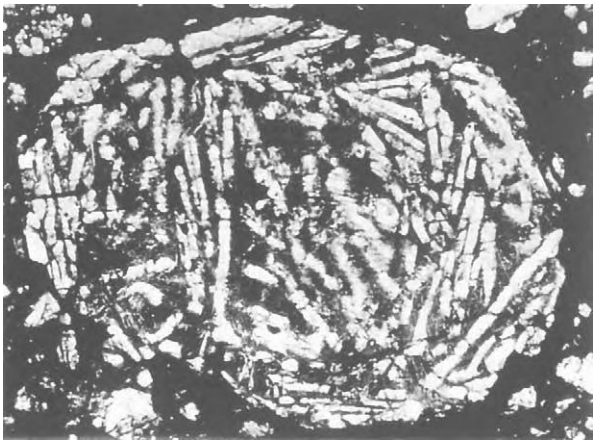


Figure 7 View in a microscope thin section across a chondrule (2 mm diameter) showing elongated olivine crystals and dark glass, within the rounded chondrule, which is set in an aggregate of olivine, pyroxene, and feldspar grains, opaque nickel iron, sulphide and products of secondary weathering (Mulga South ordinary chondrite, Western Australia (from McCall 1973)).

Meteorites within Meteorites

Many meteorites are brecciated, probably mainly due to shock processes through collision with other meteorites in space, but some also carry other meteorite types as fragments within them. Chondrites may occur as fragments within dissimilar host chondrites. Even more spectacular are shocked eucrite achondrite bodies within the Mount Padbury stony iron (mesosiderite) and enstatite and carbonaceous and ordinary chondrite bodies within the Bencubbin stony iron meteorite, both found in Western Australia.

Table 1a Undifferentiated meteorites

Class	*Symbol*	*Example*
Carbonaceous chondrites	CI	Orgueil
	CM	Murchison
	CO	Ornans
	CV	Allende
	CK	Karounda
	CR	Renazzo
	CH	ALH 85085
Rumurutiites	R	Rumuruti
Kakangari type chondrites	K	Kakangari
Ordinary chondrites	LL	Saint Mesmin
	L	L'Aigle
	H	Wiluna
Enstatite chondrites	EL	Eagle
	EH	Saint Sauveur

Carbonaceous chondrites: characterized by sparse to abundant chondrules set in a dark, friable matrix of carbon rich compounds, phyllosilicates, mafic silicates, Ni Fe metal, and glass. The letter symbols separate groups based on different mineralogy, relative abundance of different lithophile and siderophile elements, relative abundance and size of chondrules, and oxygen isotope signatures. Numerical suffixes 3, 2, and 1 denote progressive aqueous alteration and 3, 4, 5, and 6 progressive thermal alteration.
Rumurutiites: a new rare group of chondrites.
Kakangari type chondrites: a small group of chondrites now separately defined.
Ordinary chondrites: chondrules are embedded in a finely crystalline matrix of mafic minerals, pyroxene, and olivine, together with NI Fe metal and glass. Some are recystallised thermally and lose the definition of chondrules and the glass. The H, L, and LL groups differ in the magnesian/iron ratio in the ferromagnesian silicate minerals. The number suffixes 3 7 denote degree of thermal alteration (loss of original texture and recystallization).
Enstatite chondrites: these are chondrites with the Mg rich pyroxene enstatite. The EL and EH groups have different relative abundances of silicates and metals. The numerical suffixes above (3 6) may be applied.

Table 1b Differentiated meteorites*

Class	*Symbol*	*Example*
Irons	I AB	Campo del Cielo
	I C	
	II AB	Sikhot Alin
	II C	
	II D	
	II E	
	II F	
	III AB	Cape York
	III CD	
	III E	
	III F	
	IV A	Gibeon
	IV B	
Stony irons	Mesosiderites	Mount Padbury
	Pallasites	Krasnojarsk
Stones (Achondrites)	Eucrites	Camel Donga
	Diogenites	Johnstown
	Howardites	Kapoeta
	Angrites	Angra dos Rios
	Ureilites	Novo Urei
	Aubrites	Aubres
	SNC Meteorites (Mars sourced?)	
	Shergottites	Shergotty
	Naklites	Nakhla
	Chassignite (Orthopyroxenite)	Chassigny ALH 84001
	Basaltic and anorthositic achondrites (Lunar sourced)	ALH 85085
Primitive achondrites*	Brachinites	Brachina
	Winonaites	Winona

*The primitive achondrites have igneous textures with no chondrules, but their mineralogy and bulk chemistry shows little difference from ordinary chondrites. They are supposed to have undergone igneous processes but with no fractional crystallization, but partial melting and segregation of the phases to varying degrees.
The irons were formerly separated into *octahedrites* (kamacite plus taenite; on etching yield criss crossing Windmanstatten patterns) (**Figure 4**): *hexahedrites* (mostly kamacite, yield only narrow thin Neumann lines on etching) and *ataxites* (no etch pattern). The Symbol classification above which replaced this metallurgical classification is still being modified and I AB and III CD have recently been grouped as I AB III CD. These symbols reflect the differences in chemistry (nickel, gold, iridium content, etc.).
The eucrites have basaltic textures.
Many meteorites defy classification and are listed as unclassified. For example, the Bencubbin (find, Australia) meteorite appears to be a stony iron but is in fact a mixture of four types, an iron, an enstatite achondrite, and two chondrites, one carbonaceous. It would seem to be the result of more than one collision, the first mixing occurring very early in its history (*ca.* 4500 Ma) and causing heating and melting.

Age

There are four periods of time that are significant in the history of any meteorite:

Terrestrial age: the time spent on Earth since fall. Obviously, the material from an observed fall has an immediately known terrestrial age. Cosmic-ray induced isotopes are used to obtain this age from such finds. We know from observed fall meteorites how much of these isotopes are in a meteorite when it arrives. A meteorite found later will have less isotopes because the Earth's atmosphere protected it after arrival, and unstable products of cosmic radiation, such as ^{14}C will decay, so that the difference between the normal content on arrival and that

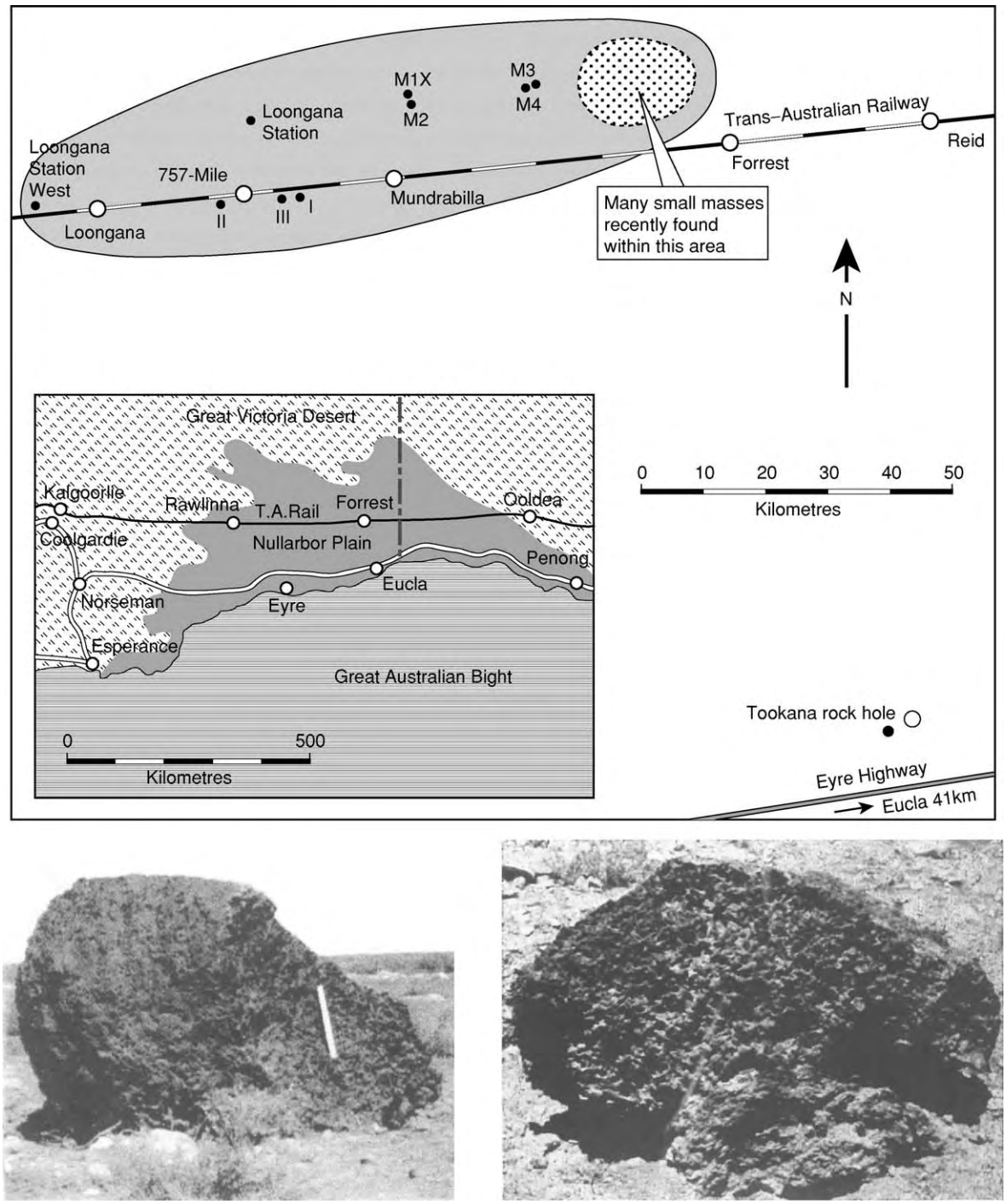


Figure 8 The distribution of the Mundrabilla irons on the Nullarbor Plain, Western Australia (rediscovered 1964 onwards) showing the typical dispersion ellipse. Below left: the M1 mass (est. 11 tonnes), as found, showing the space capsule shape with striations on anterior surface in atmosphere descent: also the curved face where the M2 mass separated. Below right, the M2 mass (est. 5 tonnes), showing the 10 cm pad of iron shale below, the product of a million years weathering by surface agents since fall (from McCall 1999, reproduced with permission of Palgrave Macmillan).

Table 2 The total of known meteorites up to the end of 1999

Class	*Fall*	*Find*	*Total*
Stones	940	20574	21514
Stony irons	12	104	116
Irons	48	817	865
Unknown	5	7	12
Total	1005	21502	22507

(After MM Grady (2002)).

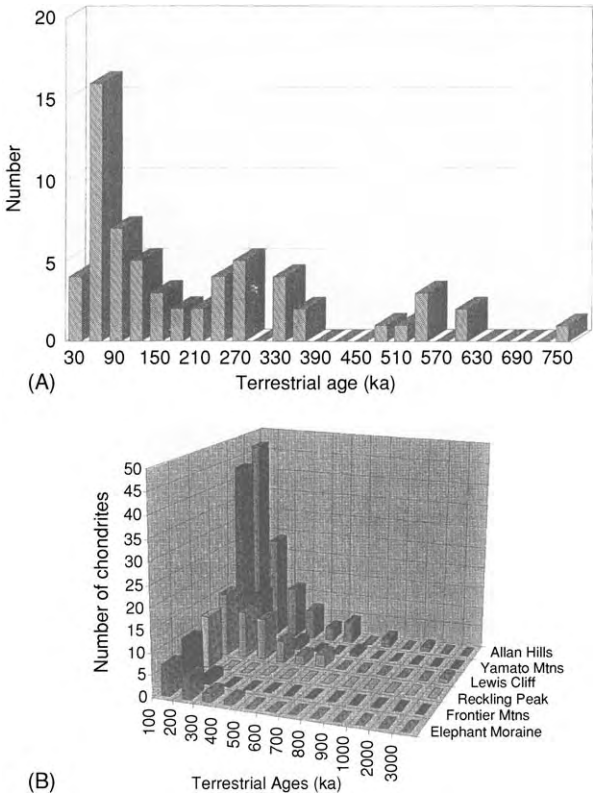


Figure 9 (A) Terrestrial age distribution for meteorites from the Allan Hills main icefield, Antarctica. (B) Terrestrial age distribu tion for 280 Antarctic meteorites sorted by stranding site. (A) from AJT Jull, S Cloudt, and E Cielaszyk; and (B) from ME Zolensky, in Grady *et al.* (1998). Published with the permission of the Geological Society Publishing House, Bath.

measured after the find can be used to determine the terrestrial age. As meteorites decay through natural weathering processes, these ages are usually values of tens of thousands of years, but in arid regions such as the Nullarbor Plain they are likely to be more, even a million years in the case of the large Mundrabilla iron (Figure 8); and in the Antarctic the ages taper off about 300 000 years though a very few have ages of one to three million years (Figure 9).

Cosmic ray exposure age: the time spent as a metre-scale meteoroid orbiting the Sun. Cosmic rays react with some atoms in iron or stony meteoroids and the quantity of gases formed depends on the chemical nature of the meteoroid and the duration of exposure to cosmic rays in space. The most usual measurements are of the quantity of neon gas resulting from this cosmic ray exposure. The evidence suggests that few stony meteorites survive in space without further collisional destruction and pulverisation for more than 40 million years, but iron meteorites are more robust, surviving up to 1000 million years.

Formation age: the age between the last high temperature episode in the parent body and the present. In the case of basaltic achondrites, this represents the time of crystallization from the liquid in a magma: chondrites, which have slightly greater formation ages, did not melt but were hot and recrystallised as solids soon after formation. The method involves the normal radioactive 'clocks' used by geologists, such as uranium-lead, the amount of lead produced by radioactive decay being an indicator of formation age. Values for chondrites are near to 4550 million years; some parent bodies were then heated and melted with fractional crystallization during the next 100 million years.

Formation interval: the time of the formation of the elements in stars (where almost all the elements except H and He were formed) and their incorporation in the parent body. This is done by measurement of the decay products of plutonium, an element which, because of its short half life, does not occur naturally. Plutonium was formed in a star about 150 million years before the formation of the asteroidal parent bodies of meteorites, but other elements were formed at different times.

Provenance

Asteroidal

Meteorites are nowadays accepted as fragments of strays from the asteroid belt between Mars and Jupiter. Prior to the mechanism being established of producing (due to collisions) eccentric elongated orbits for asteroids – replacing their quasi-circular orbits beyond Mars – the nucleii of comets, impoverished in volatiles by repeated passage round the Sun, were long favoured as their source, but petrological and mineralogical evidence is against this. The Farmington fall in Kansas in 1890 seems to have heralded the firm establishment of asteroidal source. Sixty reports of visual observations of this fireball, at 12.50 pm on

a midsummer day and reportedly rivalling the Sun, were selected by scientists who deduced an orbit indicating that the parent asteroid was 1862 Apollo, Hermes, or 1865 Cerberus. Direct observation of fireballs by astronomers of the Sikhot-Alin, Siberia, 1949 and Pribram, Czechoslovakia, 1959 fireballs again strongly supported asteroidal sources and there have been many further supporting observations since (Figure 10). In recent years there have been numerous attempts to use optical and spectrographic methods to equate the reflectance and chemistry of asteroids with different classes of meteorites, but results seem to be inconclusive, possibly because of the operation of little understood space-weathering processes which affect the regolith surface of asteroids. Even a direct exploration mission to Eros in 2000–2001 (Figure 11) yielded no correlation and it must be borne in mind that there must be asteroids of classes never sampled by meteorites falling on the Earth. Several thousand asteroids are now known and it is estimated that there may be as many as 10 000 out there.

Even in these small parent bodies, though some did not reach 100°C, others heated to more than 1200°C, the temperature needed to form a basaltic-textured eucrite. The heat sources in these small bodies are not known for certain, but a source in extreme early heating of the Sun or internal short-lived radioactive isotopes such as ^{26}Al is favoured.

Martian Achondrites?

Some meteorites apparently do not originate from asteroids. The 'SNC' group of achondrites (Shergottites, Nakhlites, Chassignite) (Figures 12 and 13)

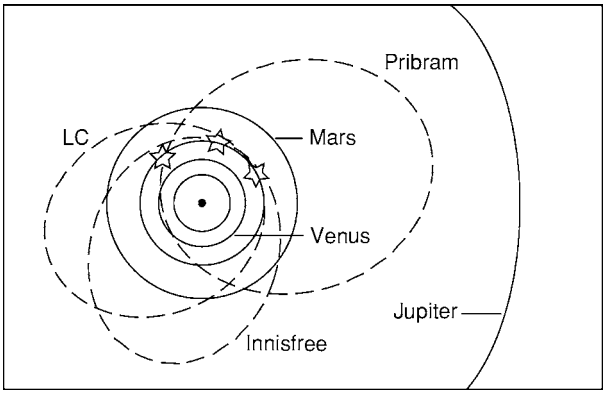


Figure 10 Orbits crossing that of the Earth derived photographically from the falls of the Pribram (Czechsoslovakia), Innisfree (Canada), and Lost City (USA) meteorites. (New figure, after Hutchison and Graham (1992).)

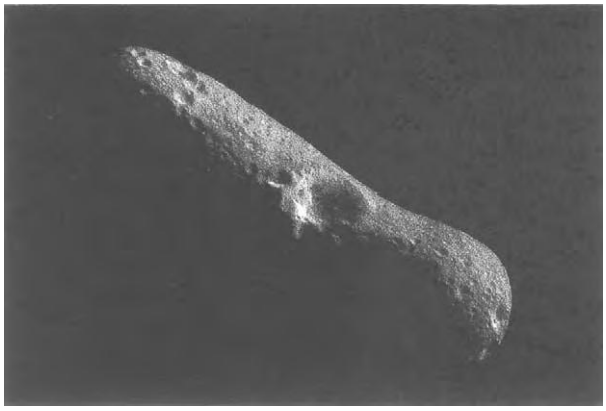


Figure 11 Asteroid 433 Eros (NEAR Shoemaker multispectral NASA image). The large crater, Psyche, has a diameter of 5.3 km.

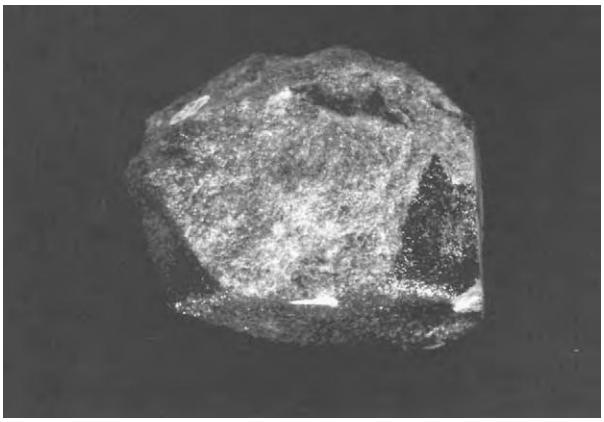


Figure 12 The Nakhla achondrite (fell 1911, Egypt, one of 40 stones, totalling 40 kg); one of the SNC (?Mars sourced) meteorites (from McCall 1999, reproduced with permission of Palgrave Macmillan).

Figure 13 Thin section view of the microtexture of the Nakhla meteorite, a typical achondritic texture resembling that of terrestrial igneous rocks, formed by diopside pyroxene, olivine, and a few plagioclase crystals (×10) (from McCall 1999, reproduced with permission of Palgrave Macmillan).

were first thought to come from Mars because of the presence of oxidised iron and hydrated minerals. Later in the twentieth century, entrapped gases in these meteorites were found to be similar to the Martian atmosphere sampled by Viking missions. The ages of formation of these meteorites (see below) are not those of the asteroidal meteorites (*ca*. 4550 Ma), but fall into two groups – Nakhlites 180 and Shergottites 1300 Ma (equivalent to Earth's Jurassic and mid-Proterozoic). The widely accepted source of these meteorites is Mars – the source must surely be a planet, and the mechanism the spalling off the surface by large impacts (there are theoretical objections to volcanic ejection). However there are problems: the trapped atmosphere should be the planet's atmosphere 180 and 1300 Ma ago, not the present atmosphere, and atmosphere's change with time: also, why are the 26 SNC meteorites recovered to date all a limited range of familiar igneous rocks – Mars is a very diverse surfaced planet? A hypothetical geological history of Mars has been built up by scientists on the basis of these 26 meteorites, an edifice which direct exploration may surely demolish?

The joker in the pack is the famous ALH 84001 from Antarctica, a unique orthopyroxenite, which has a formation age similar to the asteroidal meteorites and contains the famous putative microfossils, the evidence about which seems now to favour inorganic rather than organic origin.

Lunar Achondrites

Lunar achondrite meteorites (Figure 14) so completely match lunar surface rock samples obtained by Apollo and Luna missions that there is no doubt as to their provenance. First found in Antarctica, they have been later recognized in an existing collection from Western Australia and also new finds in the Libyan desert. Volcanic ejection can be ruled out; isotopic evidence suggests that all were spalled off by geologically quite recent and relatively minor impacts on the surface of the Moon, but here there is a glaring unresolved problem. There is widespread scientific acceptance of a major impact bombardment of the Moon 3.9 Ma ago, forming innumerable and immense craters: this must have hurled vast volumes of rock out into space, sampling deep below the regolith and surficial breccia (which is all that has yet been directly sampled), there is no trace of this material in the varied log of meteorites. Where has it gone?

Cratering and Tektites

Meteorites normally land with little effect on the ground – even the 11 tonne Mundrabilla iron left no dent in the limestone surface – but multiple showers may produce small, simple craters (the 1947 Sikhot-Alin shower produced 106 associated with nickel-iron fragments). Larger masses have, in the quite recent geological past, produced kilometre-scale simple craters associated with nickel-iron (e.g., Canyon Diablo, Arizona; Wolfe Creek, Western Australia) and about 170 larger simple craters and more complex ring structures in the geological record are attributed to impact explosion processes involving larger masses, even asteroids. The largest, at 180 km diameter (Chicxulub, Yucatan, Mexico) has been associated with the Cretaceous-Tertiary boundary extinction of life (*see* **Impact Structures**). Only geochemical traces of the impactor have been discovered at such sites. Tektite showers were associated with a very small minority of such structures, but tektites are not meteorites, but are glassy objects melted from the impacted surface rocks, and spread over strewn fields at long distances from the impact sites (*see* **Tektites**).

Figure 14 Lunar sourced achondrite meteorite, ALH 81005 from Antarctica, discovered in 1981, after the first such discovery in 1979 by Japanese scientists in the Yamato Mountains. The structure is that of the lunar regolith breccias and a large white fragment of highlands anorthosite is visible. The cube has sides of 1 cm length (from McCall 1999, reproduced with permission of Palgrave Macmillan).

Fossil Meteorites

The only recorded case of a meteorite being recorded in ancient rocks relates to limestone strata at a quarry near Goteborg, Sweden, where there are 12 horizons crowded with ordinary chondrite meteorites, which must have been derived from rains of stones 480 Ma ago, in the Ordovician, the stones falling onto the limey mud bottom of shallow sea. Meteorites do not fall repeatedly at the same place because of the Earth's rotation and this repetition is astonishing, as it implies repeated globally spread rains of meteorites over a period of about 1.75 million years.

The Twentieth-Century Revolution in Meteoritics

Up to the orbital flight of Sputnik heralding the space age in 1957, the study of meteorites was a quiet museum occupation. The scientific interest in meteorites then exploded because of what they might tell us about planets, satellites, and asteroids.

By coincidence, the year of Apollo XI, 1969, saw a Japanese party find nine meteorites on an area of bare ice in the Yamato Mountains. Antarctica. This 5×10 km area subsequently yielded 1000 meteorites. Blue ice areas and moraines in Antarctica have now yielded approximately 30 000 specimens representing some 20 000 falls. Two principle factors produce the optimum conditions for recovery: weathering is virtually nil in the arid climatic conditions and low temperatures prevailing, and the 'conveyor belt' situation on the ice sheet, snow falling and being buried and compacted to ice together with any meteorites on the surface, the snow moves coastwards and where nunataks (rocky peaks) obstruct its passage, the entrained and buried meteorites are excavated by wind action which removes the ice above (Figures 15 and 16).

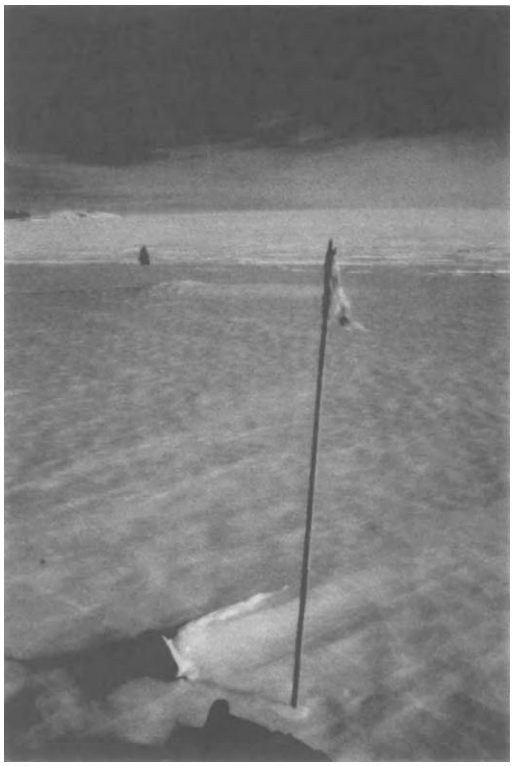


Figure 15 A meteorite as found on blue ice, its position flagged, Antarctica (from McCall 1999, reproduced with permission of Palgrave Macmillan).

By coincidence again, in the 1960s, rabbit trappers kept bringing in meteorite finds strewing the limestone surface of the arid Nullarbor Plain in Western Australia, and the writer of this entry, then working at the Western Australian Museum, wrote prophetically, "that the Nullarbor Plain must be littered with meteorites of all types". This was indeed so and systematic collection has so far yielded about 300 individual meteorites including two shower groups of more than 500 meteorite masses. Other desert areas of the world were then searched and Libya, Algeria, Morocco, and Oman have yielded several hundred finds, while desert areas in Roosevelt County, New Mexico have also proved productive. Neither Antarctica nor the desert areas are 'worked out' and many more finds will undoubtedly be made in the next years of this century. There are some desert areas in Asia, including the Gobi, that are not even searched so far, but a reconnaissance in the Gobi proved disappointing.

State-of-the-Art Research

Meteoritics is a major area of scientific research nowadays and as many as 500 scientists may attend the yearly meetings of the Meteoritical Society. Research topics are extremely varied and besides such related topics as impact processes; tektites; planetary, lunar, satellite, cometary, and asteroid exploration, topics bearing directly on meteorites may include:

- Ca-Al rich inclusions (CAIs) in meteorites, believed to be survivals from the accretion of the Solar System
- Isotope fractionation in pre-solar graphite in carbonaceous chondrites
- Isotope studies of chondrules and CAIs
- Modelling conditions for the launch-window of ?Martian meteorites
- Aqueous alteration of carbonaceous chondrites
- Presolar nano-diamonds in meteorites
- Xenon isotopes in nano-diamonds
- Trapped gases in ordinary chondrites
- Trace elements trapped in lunar meteorites

This random sample illustrates the diversity of research: the revolution in meteoritics described above has produced enough subject material to keep science busy for many decades, if not centuries, and more keeps coming in. The important point to remember that meteorites come in free of charge – they have been called the poor man's 'space probe'. Even the cost of searching after major bolide events, searching Antarctica and systematic searching of the

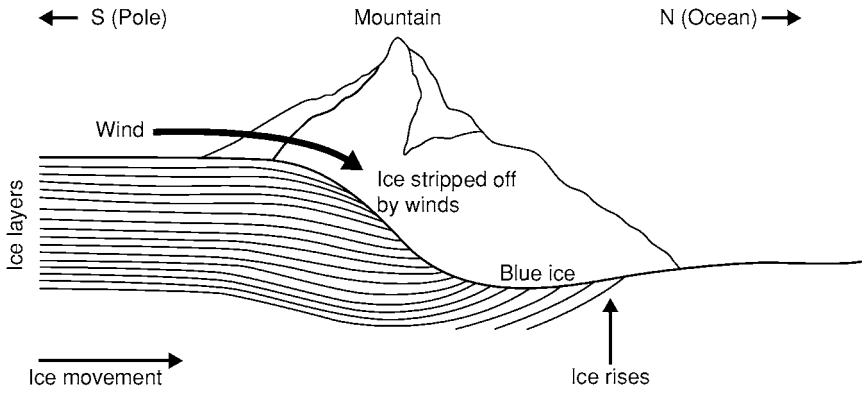


Figure 16 Diagram showing how ice, moving very slowly towards the coastal ice front, is arrested by a rock nunatak, is stripped by wind action while stationary, to reveal entrained meteorites (new figure, after Hutchison and Graham 1992).

Nullarbor are infinitesimal when compared with the costs involved in direct space exploration.

See Also

Impact Structures. **Solar System:** Asteroids, Comets and Space Dust. **Tektites**.

Further Reading

Bevan AWR and Dehàeter JR (2002) *Meteorites: A Journey Through Space and Time*. Washington DC: Smithschian Institution Press.

Grady MM (2002) *Catalogue of Meteorites*, 5th edn. London: Natural History Museum.

Grady MM, Hutchison R, McCall GJH, and Rothery DA (1998) *Meteorites: Flux With Time and Impact Effects*, Special Publication No. 140. Bath: Geological Society Publishing House.

Hey MH (1966) *Catalogue of Meteorites*, 3rd edn. London: British Museum (Natural History).

Hutchison R and Graham A (1992) *Meteorites*. London: Natural History Museum.

Krinov EL (1960) *Principles of Meteoritics*. Oxford, London, New York, Paris: Pergamon Press.

Mason B (1962) *Meteorites*. New York, London: John Wiley & Sons.

McCall GJH (1973) *Meteorites and their Origin*. Newton Abbot: David and Charles.

McCall GJH (1999) The Mundrabilla iron meteorite from the Nullarbor Plain, Western Australia: an update. In: Moore P (ed.) *2000 Yearbook of Astronomy*, pp. 156 166. London: Macmillan.

McCall GJH (1999) Meteoritics at the millennium. In: Moore P (ed.) *2000 Yearbook of Astronomy*, pp. 153 177. London: Macmillan.

McCall GJH (2001) *Tektites in the Geological Record*. Bath: Geological Society Publishing House.

McCall GJH and de Laeter JR (1965) *Catalogue of Western Australian Meteorite Collections*, Special Publication No. 3. Perth: Western Australian Museum.

Norton OR (2002) *The Cambridge Encyclopedia of Meteorites*. Cambridge: Cambridge University Press.

Olson RJ and Pasachoff JM (1998) *Fire in the Sky*. Cambridge: Cambridge University Press.

Zanda B and Rotaru M (2001) *Meteorites*. Cambridge: Cambridge University Press.

Mercury

G J H McCall, Cirencester, Gloucester, UK

Historical

Mercury, the closest planet to the Sun, was something of a mystery to ancient watchers of the sky, being visible to the naked eye only low down on the horizon close to sunset or sunrise – it is never seen more than 28° of arc from the Sun and is never seen against a dark sky. It was also some time before 'morning Mercury' and 'evening Mercury' where identified as the same planet. Nothing was known of its physical appearance until the advent of telescopes. Its phases and the blunting of its 'horns' (an optical effect) were then recognized (Figure 1). Johann Schroter (1745–1815) and W F Denning (1848–1931) claimed to have detected light and dark configurations, but their sketches bear no resemblance to the real surface as revealed by Mariner 10 in 1974. Denning also claimed to have detected a 25 h rotation period, now known to be erroneous. In 1953 A Dollfus recorded the presence of a tenuous atmosphere, which was later confirmed by Mariner 10, although it is even more tenuous than he supposed. The largest telescope cannot show Mercury as well as the Moon can be seen with the naked eye. Thus, accurate representation of a large part of its surface had to await Mariner 10, which reached a distance of 470 miles from the planet and transmitted images with a resolution of approximately 100 m showing a surface remarkably like that of the Moon, predominantly cratered with scarps, ridges, and plains.

'Vulcan': An Inner Neighbour Planet?

In 1958, Le Verrier received a report that a French amateur astronomer had discovered an innermost planet, closer to the Sun than Mercury. He had earlier found the movements of Mercury to suggest that such a planet existed, but in fact the anomalous movements have since been explained, and it is certain that 'Vulcan', the putative inner planet, does not exist, although some asteroids may pass closer to the Sun than Mercury on their orbits oblique to the ecliptic.

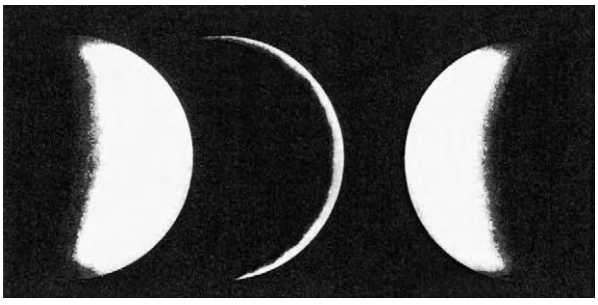


Figure 1 Phases of Mercury showing the optical effect of blunted 'horns'. Reproduced from Cross CA and Moore P (1977) *The Atlas of Mercury.* London: Mitchell Beazley Publications.

Statistics

Mercury is situated within the Solar System 57 850 000 km from the Sun. Its orbital eccentricity is 0.206, as determined by Antoniadi (1870–1943), the largest eccentricity of any planet except Pluto. It is, unlike Venus, brightest when gibbous. The equatorial diameter is 4880 km, intermediate between those of the Moon and Mars, more or less equal to that of Jupiter's satellite Callisto, and less than those of Ganymede (Jupiter) and Titan (Saturn). The escape velocity is 4.3 km s ^{1}, meaning that very little atmosphere is likely to be retained. Its density is surprising, at 5.4 g cm ^{3}; this high value compared with the Moon requires that a heavy iron-rich core takes up a higher relative proportion of the globe than in the case of the Earth. The mass of Mercury is 0.055 times that of the Earth, and its volume is 0.056 times that of the Earth. Its orbital period of 87.969 Earth days is not, as in the case of the Moon, synchronous with its rotation around its axis, which takes 58.65 Earth days. On Mercury there is no area of permanent daylight or night and no twilight zone. It has no satellite. There is a suggestion in the literature that it may have once been a satellite of Venus – the diameter ratio is not unlike that of the Earth and Moon.

Mariner 10 Mission

Technical Summary

All we know in any detail of Mercury is derived from the remarkable Mariner 10 mission, which visited both Venus and Mercury in 1974 on a gravity-assist trajectory. The mission lasted 17 months, and the same instruments were used throughout to obtain information about Earth, Moon, Venus, and Mercury – an advantage in making comparisons. There were two daylight-side encounters with Mercury as well as a night-side encounter, for orbital-change reasons, which allowed measurement of the night-surface temperature, the atmosphere, and the magnetic field. During 17 days of encounter only 17 h were spent

close enough to obtain high-resolution images: 647 pictures were taken during the first daytime encounter and 300 during the second. The peculiar relation between the rotation period and the orbital period of Mercury meant that the same hemisphere was studied during both encounters. The sun rises and sets once during a Mercury day, which is two Mercury years long. During a Mercury day the planet rotates three times with respect to the stars.

Results

Mariner 10 imaged 40% of the lunar-like landscape, covering virtually a complete hemisphere. Despite the startling similarity, the high density of Mercury means that similarity to the Moon is only skin deep. In the 1970s the Moon was considered to supply a 'paradigm' for the understanding of other planets, but the high density and geochemical properties (volatiles, refractory minerals, FeO content in the crust) of Mercury revealed by Mariner 10 suggested that Mercury is the end member of an inner–outer progression of planets, whereas Mercury is anomalous. Of course, if the suggestion that Mercury is a displaced satellite of Venus is correct, then both the Moon and Mercury are anomalous. Only further missions to Mercury will answer this question.

The surface revealed by Mariner 10 has all the features of the Moon, except that it lacks extensive dark smooth plains (e.g. Imbrium) but there are quite substantial areas of lighter smooth-plain terrain, and the circular Caloris Planita feature (the largest single feature so far recognized at 1300 km in diameter) is of comparable extent to some maria and does show resemblance to lunar maria (Figure 2). The smooth-plain material does appear to have lapped over, obscured, and infilled large as on the Moon craters (Figure 3), which were formed in an older, rougher surface formation, analogous to the lunar highlands, although probably not of the same composition.

Craters dominate the entire mapped surface, and, as on the Moon, when one crater interferes with another the smaller crater is usually the intruder. Beethoven, the largest crater on Mercury, has a diameter of 625 km. Tolstoj (Figure 4), at 400 km, is about the same size as Mare Crisium on the Moon and is larger than any lunar crater. Some craters are double or have distorted circular outlines – Bach (225 km in diameter) shows both these features (Figure 5). Other craters have double walls. There are crater-sized rings outlined by annular grooves in otherwise flat plains. There are even circlets of small craters. There are prominent rayed craters like those of the Moon, which are apparently late-introduced features (e.g. Copley, Kuiper, Snori, Mena): Copley (Figure 6) is clearly later than the smooth plains; Mena (Figure 7) displays two anomalous features – one sector totally lacks rays and the rays are both curved in places and do not all emanate from a shared point focus, a characteristic seen at other rayed craters. These features are more consistent with rays being due to deposits along fracture lines than being ejection rays. Central peaks are common and may be single central peaks or off-centre single and clustered peaks.

There are few 'Montes' on Mercury, the only such feature recorded so far being the edge of Caloris Planita. Linear scarps called 'Rupes' are, however, widespread. The albedo is different from that of the Moon – on the Moon iron-rich plain basalts and light feldspathic highland anorthosites make for a dark–light contrast, whereas the surface rocks of Mercury are all relatively light coloured because of their

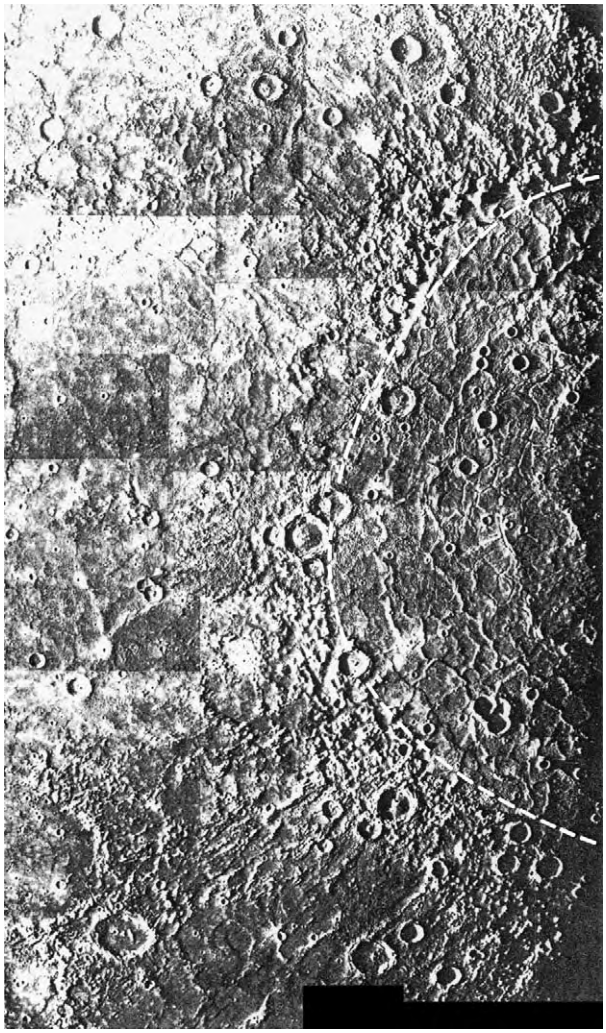


Figure 2 Caloris Planitia (dashed line) showing the concentric ridge pattern. The basin has a diameter of 1300 km. Mountain blocks at the margin rise to 1 2 km above the surrounding terrain and the peripheral linear ridge terrain extends to 100 km from the outer edge. Photograph from NASA image bank.

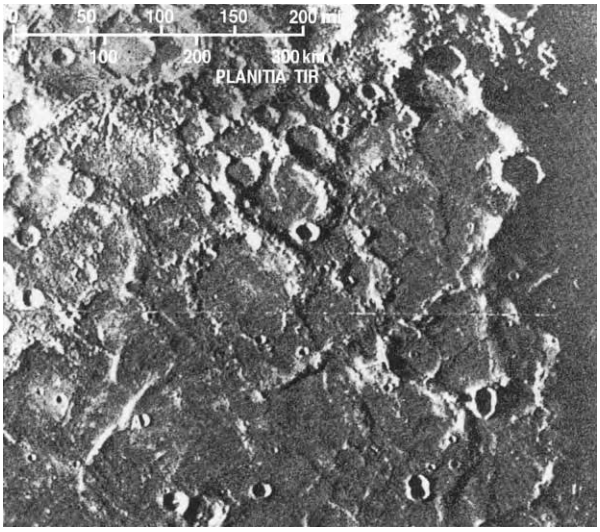


Figure 3 Part of Tir Planitia, on Mercury, showing the flooding of older large craters by smooth plain material.

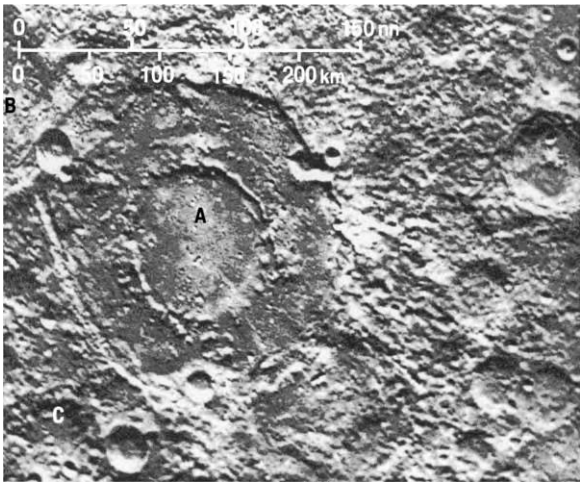


Figure 5 Bach, a double ring crater with plain material in its floor; the shape of the outer ring is subpolygonal and one side has a wall formed by an almost straight groove.

Figure 4 The large crater Tolstoj (outlined by the dashed line), which is comparable in size to the lunar Mare Crisium.

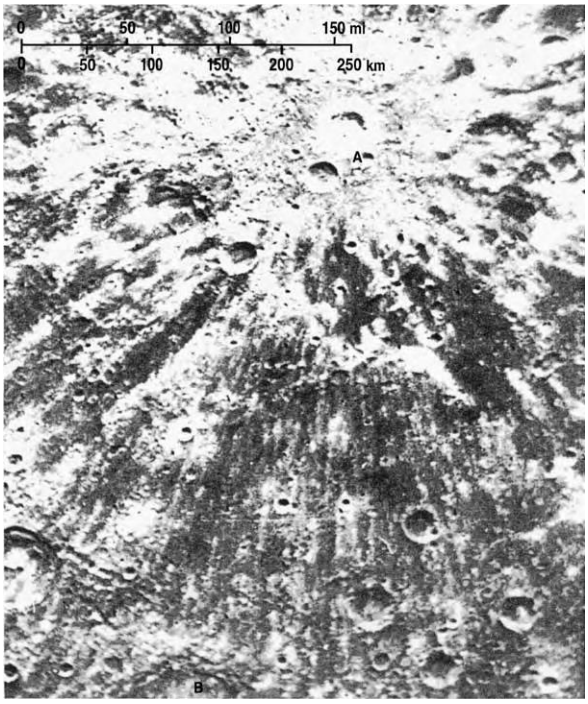


Figure 6 The rayed crater Copley, which is clearly younger than the smooth plain material. The rays extend out into the south east sector for 400 km. Note the irregularity and curved trace of the rays and the fact that rays overprint smooth plain terrain.

iron-poor nature (Figure 8). At the Mercury conference in Chicago in 2001 there was a consensus that the FeO percentage in the rocks of Mercury averages around 3%. This is consistent with models in which the planet was assembled from planetesmals that were formed near the planet's current position. The magmas of Mercury may be similar in composition to the aubrites (enstatite achondrite meteorites), though these are demonstrably from their isotopic character asteroidal, not Mercurian, in provenance.

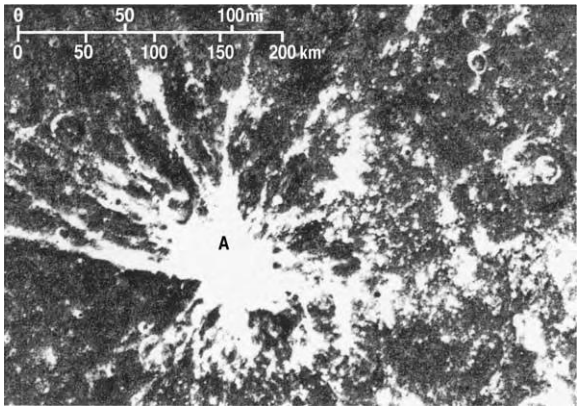


Figure 7 The rayed crater Mena, the rays of which neither emanate from a single focal point nor are straight; they extend outwards for more than 250 km. Photograph from NASA image bank.

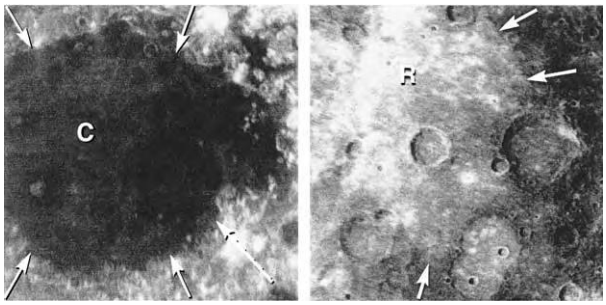


Figure 8 The contrast between the dark floored Mare Crisium on the Moon and the similar smooth Rudaki plains of Mercury; both have embaying boundaries (arrowed). Both images were taken by Mariner 10 and are reproduced from Robinson MS and Taylor GJ (2001) Ferrous oxide in Mercury's crust and mantle. *Meteoritics and Planetary Science* 36: 842 847. © 2001 by the Meteoritical Society.

There are lobate scarps on the surface that may be due to shrinkage (thermal models predict 4–6% shrinkage), and these have been suggested to be the result of thrust faulting.

One entire side of Mercury remains to be seen, and this may be either similar to or very different from the cratered known surface.

The magnetic field was the biggest surprise revealed by Mariner 10. Though only amounting to 1% of the strength of our own planet's field, it is enough to indicate the existence of a core dynamo. Only the strength of the dipole component is at present known, but a solid inner core and liquid outer core are required by the present evidence. Convection in the outer core becomes more complex as the inner core grows. Thermal models suggest that the inner core of Mercury, if it exists, cannot be pure metal, and a non-metal such as sulphur is required to lower the crystallization temperature and density.

Volcanism on Mercury?

The impact-cratering paradigm is part and parcel of NASA's exploration and interpretation philosophy. For example, a presentation by Potts and others in 2002 made the assumption that the overall cratering results from 'bombardment time'. It remains possible, however, that many supposed impact craters, especially simple craters and some very complex structures, on surfaces of space bodies may have been too summarily dismissed as due to impact. Past volcanism is manifest on Mars, the Moon, and Venus, and there is active volcanism on Io. The widespread plains material on the surface of Mercury, as revealed by Mariner 10, though not as extensive as the larger lunar Maria, could be the result of primary volcanic flows or lobate crater ejecta. Study of the theoretical possibilities by Milkovich and others, published in 2002, indicates that widespread volcanism or no volcanism whatever or something in between are all possible. No volcanic features can be identified in the Mariner 10 images, although at the same image resolution few, if any, volcanic features would be identified on the Moon without the prior knowledge obtained by Apollo on-the-ground examination. High-resolution low-sun-angle images from Mariner 10 do show what appear to be flow fronts on Mercury; these could be volcanic lava flow fronts or ejecta flows. It seems likely that Mercury may show the same 'freezing' of the surfaces of lava flows as seen on the Moon, with lava of a basaltic (*sensu lato*) or a peculiar Mercurian petrological character preserving a plains surface among the craters and flooding some craters that were formed in the early history of the planet, perhaps around 4 Ga ago, with, as in the case of the Moon, very little having happened since then except for minor-scale impacts. This is, however, no more than informed speculation – Milkovich and others are correct in concluding that a clear assessment of the role of volcanism and whether it is primary or secondary (impact generated) must wait for better data.

Nevertheless, there are many features in the Mariner 10 database that appear to be incompatible with impact origin – for instance, the sprinkling of an area about 400 km across centred on the crater Zeami with a myriad of small craters, all virtually the same size. These have been dismissed as 'secondaries', but this explanation appears facile. The scalloped walls of the largest crater, Beethoven (Figure 9), are anomalous in an impact crater, and it lies in the centre of a cluster of similarly scalloped-walled craters that include double and triple rings. Unfortunately,

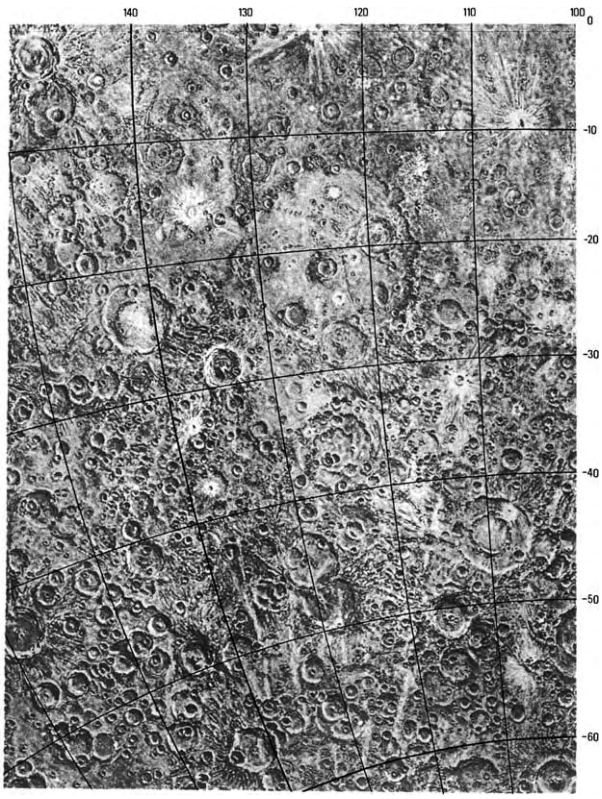


Figure 9 The area around Beethoven, the largest Mercurian crater, showing its scalloped walls and its position within a cluster of scalloped walled craters including doubles and triples. Photograph from NASA image bank.

rigorous analysis of cratering features appears to have lapsed amongst planetologists, with the convenient assumption that impact craters can be secondarily modified in every conceivable way – and so all can be dismissed as products of the 'Great Bombardment'.

It must be concluded at the post-Mariner 10 state of knowledge that Mercury probably has had no volcanic activity, like the Moon, for nearly 4 Ga, but that the heavily cratered 'lunar-like' terrain cannot, as yet, be entirely dismissed as impact generated, and volcanism (either primary and endogenous or secondary, exogenous, and impact-generated) may have contributed significantly to the early formation of the crater dominated surface of Mercury. Future Mercury-directed missions will hopefully resolve this problem.

The success of Mariner 10

Mariner 10 told us just enough for us to realize how interesting it would be to plan return missions, more technically equipped and specifically designed, and building on Mariner 10, to answer outstanding questions, but Mercury remained the elusive planet until 2001 when the two new 'Messenger' and 'Bepi Colombo' missions were proposed.

The Future: 'Messenger' and 'Bepi Colombo'

It requires considerable energy to put a spacecraft into orbit around the innermost planet. NASA's 'Messenger' will use multiple gravity-assist encounters when it is launched in 2004 to reach Mercury in 2009. The European Space Agency and Japan Institute of Space and Astronomical Science's 'Bepi Colombo', to be launched in 2009 and arrive in 2012, will use propulsion technology, which is costlier and riskier but reduces the transit time from 5 years to 2.5 years. Messenger will study the nature of the surface, geochemistry, the Space environment, and ranging. Bepi Combo's remit is not fully worked out but will include geochemistry. It will deploy two orbiters and a lander. Thus, the nature of the unexplored side of the planet at least will be revealed when the orbiters send back the data.

The thermal environment provides the biggest challenge, with the Sun being more than 11 times as intense as on Earth and temperatures reaching up to 400°C on the sunlit side. The high temperature causes stress in equipment and may inhibit full uptake of geochemical data. The collection of mineralogical data is similarly inhibited by the blocking out of some infrared bands. Bepi Colombo will carry an actively cooled infrared spectrometer to counter this. Use of conventional solid-state detectors is impossible without power-hungry active cooling. Solar panels decay under such high temperatures. Despite these constraints, both missions will use photon (gamma, X-ray, optical) and neutron spectrometers to provide impressive geochemical information.

These two missions will research the magnetic field and its implications for the core configuration. Experiments will determine whether there is an inner core, decoupled from the rest of the planet. These experiments are important in understanding the geophysical properties of the planet and its volatile inventory, sulphur being a volatile element.

Space weathering by micro-impacts is likely to be greatly enhanced on Mercury compared with the Moon because of the flux of incoming particles close to the Sun (*ca.* X 20% has been suggested) but, because of the magnetic field, this effect may be limited to an equatorial belt. Mercury has a long comet-like sodium tail, which is probably caused by particle sputtering.

Ground-based radar observations suggest that there may be water-ice in the polar regions – high

radar reflectivity suggests ice, possibly mantled by dust. There is similar evidence from craters where the surface is in 'shadow' from direct solar heating. Thermal models, however, predict that ice would not be stable there. The two missions will investigate this problem.

The tenuous atmosphere will also be the subject of investigation. Ground-based spectroscopy has detected a high sodium to potassium ratio. This is the case in the Moon's exosphere, and there it is related to the ratio found in the returned lunar-surface rocks. However, the Mercurian ratio is very large and shows diurnal variations; it appears to be related to the magnetic field rather than to the surface rocks.

Data return to Earth will also be constrained, because Mercury and the Sun interfere with radio transmission during part of the duty cycle. Antenna design is critical and constrained by weight limits and the fact that pointing antennae will tend to fail owing to the thermal cycling. Two fixed-phase array antennae on Messenger will limit the data return rate, and the same constraints will affect Bepi Colombo's planetary orbiters.

The two missions will overlap in their remits. Unfortunately, there seems to be little scope for short-term adjustment of the remit of the later Bepi Colombo mission on the basis of Messenger findings as the launch of the former in 2009 will coincide with the duty of the latter.

The orbits of both missions will, of necessity, be highly eccentric, and periapsis for Messenger will be over the north pole, so the southern hemisphere will be less well mapped. Thus, it would be ideal if Bepi Colombo had its periapsis at the south pole.

Conclusion

The idea that Mercury is a displaced satellite of Venus, though perhaps unlikely for astronomical reasons, must, unlike the 'Vulcan' concept, be taken seriously, in view of the surficial similarity to our Moon. If true, it would relegate the popular but criticized 'big planet collision' model for the origin of our Moon to the outer limits of credibility, for two such collisions of like dimensions are beyond the realms of probability.

The surface rocks of Mercury have been likened to the aubrite meteorites (enstatite achondrites) in their low FeO content (though we know from isotopic evidence that these achondrites do not come from Mercury). If the surface of Mercury was moulded by a giant bombardment as is widely supposed for the surface of the Moon, then the vast amount of internal rock material ejected into Space must be somewhere. Like the ejecta from the Moon's supposed Great Bombardment, this is as yet unrecognized among Space bodies. Because the Mercury ejecta was sent out closer to the Sun, total spiralling into the Sun is more conceivable as a reason for this absence, but it is still to be expected that some of it would have formed breccias by collision with asteroids. Once we know more about the nature of Mercurian rocks it will be possible to search for such foreign material in asteroidally sourced brecciated meteorites.

The renewed interest in Mercury is welcome, for the astonishing resemblence of its surface to that of the Moon revealed by Mariner 10 does suggest that when we know more about Mercury we will be able to extend this knowledge to the prime conundrum of the Earth–Moon system, and we may have to reject out of hand models for the Earth–Moon system that are at present widely supported (as we had to throw out the concept of lunar tektites after the Apollo and Lunar missions). Geologists would value more than anything a Luna-style retrieval and return of a rock sample from Mercury, but, alas, it appears that the technical obstacles are overwhelming. Yet there seem to be no limits to the ingenuity of Man. One thing is certain, there will never be a 'one step for mankind' on Mercury, such are the physical constraints.

See Also

Earth Structure and Origins. **Impact Structures**. **Solar System:** Meteorites; Venus; Moon; Mars; Jupiter, Saturn and Their Moons. **Volcanoes**.

Further Reading

Cross CA and Moore P (1977) *The Atlas of Mercury.* London: Mitchell Beazley Publications.

Hunten DM and Sprague AL (2002) Diurnal variation of sodium and potassium at Mercury. *Meteoritics and Planetary Science* 37: 1191 1195.

Koehn PL, Zurbuchen TL, Gloeckler G, Lundgren RA, and Fisk LA (2002) Measuring plasma environment at Mercury: the fast plasma spectrometer. *Meteoritics and Planetary Science* 37: 1173 1189.

Kracher A (2002) Mercury 2001 conference Field Museum, Chicago, 2001, October 4 5. Illinois. *Meteoritics and Planetary Science* 37: 307 309.

McCall GJH (2000) The Moon's origin: constraints on the giant impact theory. In: Moore P (ed.) *2001 Yearbook of Astronony,* pp. 212 217. London: Macmillan.

McCall GJH (2002) Back to the elusive planet. *Geoscientist* 12: 19.

Milkovich SM, Head JW, and Wilson J (2002) Identifica tion of Mercurian volcanism. *Meteoritics and Planetary Science* 37: 1209 1222.

Moore P (1961) *Astronomy*. London: Oldbourne.
Peale SJ, Phillips RJ, Solomon SC, Smith DE, and Zuber MT (2002) A procedure for determining the nature of Mercury's core. *Meteoritics and Planetary Science* 37: 1269 1283.
Potter AE, Killen RM, and Morgan TH (2002) The sodium tail of Mercury. *Meteoritics and Planetary Science* 37: 1165 1172.
Potts LV, von Freese RRB, and Shum CK (2002) Crustal properties of Mercury by morphometric analysis of multi ring basins on the Moon and Mars. *Meteoritics and Planetary Science* 37: 1197 1207.
Robinson MS and Taylor GJ (2001) Ferrous oxide in Mercury's crust and mantle. *Meteoritics and Planetary Science* 36: 842 847.
Sprague AL, Emery JP, Donaldson KL, *et al.* (2002) Mercury: mid infra red (3 13.5 μm) observations show heterogeneous composition, presence of intermediate and basic soil types, and pyroxene. *Meteoritics and Planetary Science* 37: 1255 1268.

Venus

M A Ivanov, Russian Academy of Sciences, Moscow, Russia
J W Head, Brown University, Providence, RI, USA

Introduction

Venus, similar to Earth in many ways, also shows many differences and provides insight into different paths of evolution that can be taken by Earth-like planets. The atmosphere of Venus is predominantly carbon dioxide; surface temperatures exceed the melting point of lead and surface pressures are almost 100 times that of Earth's atmosphere. The crater retention age of the surface of Venus is very young geologically, similar to that of the Earth; however, plate tectonics does not seem to be recycling the crust and lithosphere at present. The surface is dominated by regional volcanic activity and vertical crustal accretion, and regional tectonism appears to have been much more pervasive in the earliest part of the preserved stratigraphic record, dating from less than a billion years ago. The characteristics and distribution of superposed impact craters suggest that a major resurfacing event, perhaps catastrophic in nature, occurred on Venus in its relatively recent geological history. Venus may thus be characterized by relatively recent episodic heat loss, rather than the more monotonic loss thought to be typical of the other Earth-like planets. Despite the fact that the majority of the preserved geological record on Venus dates from the last ~20% of its history, Venus may provide insight into processes, such as the formation of continents, that operated in the first half of Earth history.

Venus is the second largest terrestrial planet by size after Earth, and in major characteristics is close to our planet: The radius of Venus is 6051.8 km (0.95 of Earth's radius), its mass is 4.87×10^{27} g (0.81 of Earth's mass), bulk density is 5.24 g cm ^{3} (0.95 of Earth's density), and surface gravity is 8.87 m s ^{2} (0.91 of Earth's gravity). For decades, Venus was considered as a 'twin' planet to Earth. Current knowledge of Venus geology is derived from several interplanetary missions, including landers and orbiters, as well as Earth-based observations. In the mid-1970s, the former Soviet Union conducted a series of successful landings; the Soviet landers transmitted panoramas of the surface of Venus, in addition to data on the near-surface environment, on surface rocks, and the chemistry of the atmosphere. The Pioneer Venus was the first American orbiter of Venus; launched by the United States in 1978, the Pioneer Venus collected data on global topography and gravity. The fundamental findings of this mission were that the global Venus hypsogram, in contrast to that of Earth, is characterized by one peak corresponding to the mean planetary radius (MPR), about 6051 km (**Figure 1A**), and that the gravity and topography of Venus are highly correlated. Three major topographic provinces characterize the surface of Venus (**Figure 1B**): lowlands (below MPR, ~11% of the surface), midlands (0–2 km above MPR, ~80% of the surface), and highlands (>2 km above MPR, ~9% of the surface). The spatial resolution of the Pioneer Venus imaging radar was too low to describe morphology of the surface in detail.

The systematic photogeological study of Venus began when high-resolution radar images were collected by the Soviet Venera-15/16 orbiters and by Earth-based radar observations from Arecibo Observatory. At a resolution of ~1–2 km, Venera-15/16 mapped the surface in the northern hemisphere above ~30° N; images from the Arecibo telescope covered a large area between 65° S–65° N and 270° E–30° E. In the early 1990s, the United States Magellan orbiter provided almost complete coverage (~97% of the surface) of Venus, providing high-resolution images (100–200 m) and medium-resolution altimetry

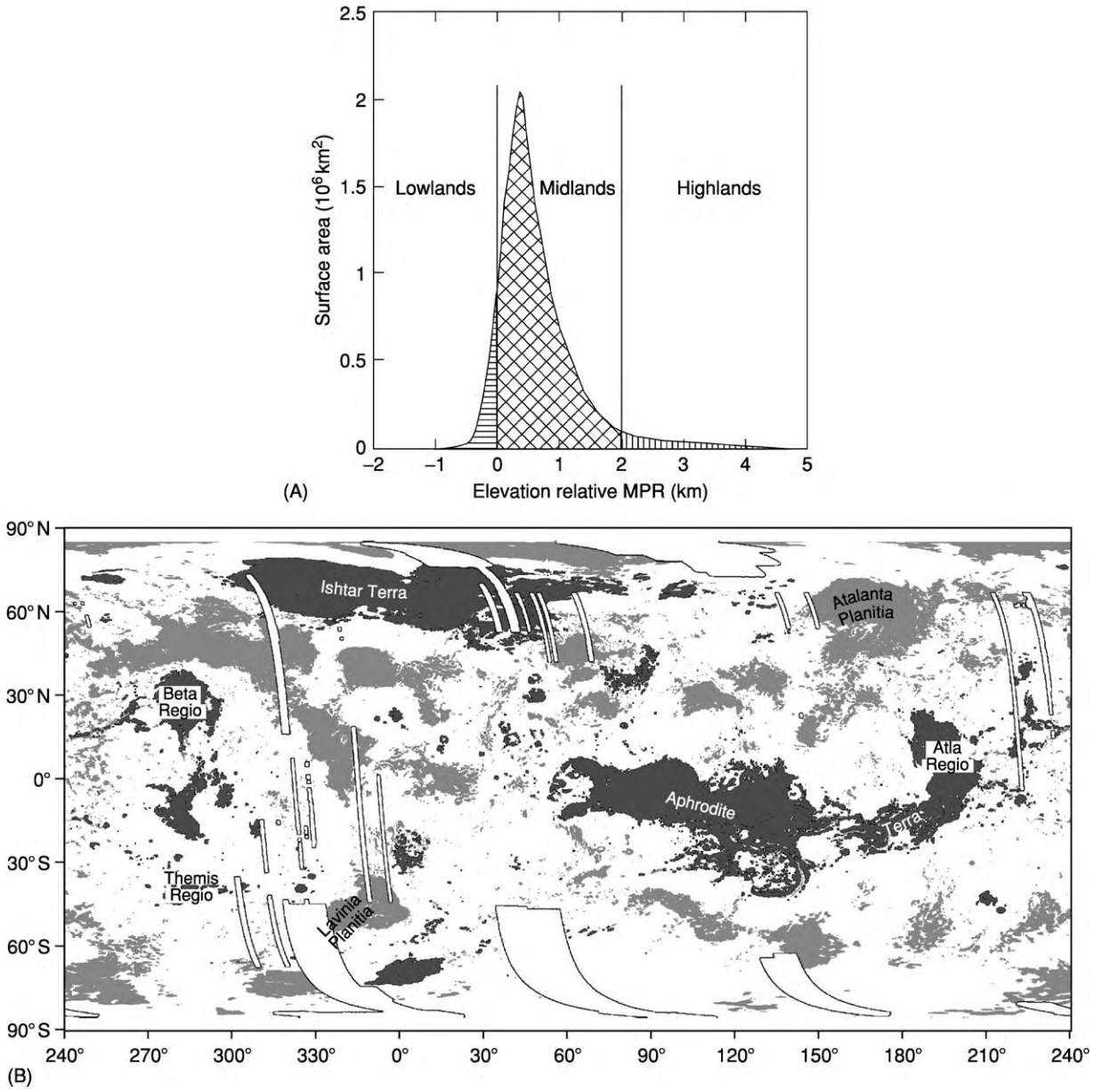


Figure 1 Characteristics of the global altimetry of Venus. (A). The Venus hypsogram (the fraction of surface area vs. elevation) has one peak, implying the absence of the surface elevation dichotomy that characterizes the distribution of the surface elevation on Earth (high standing continents and low lying ocean floor). The mean planetary radius (MPR) of Venus is 6051.84 km. (B) The map in simple cylindrical projection, showing the areal distribution of the three major topographic provinces of Venus. Lowlands (light grey, below 0 km), midlands (white, 0 2 km), and highlands (dark grey, above 2 km). The majority of the surface is within the midlands.

(~20-km footprint). Magellan also collected data on the Venus gravity field.

Surface Conditions and Rock Composition

In contrast to Earth, a very dense atmosphere (the pressure at the surface is ~95 bar) consisting primarily of CO_2 (Table 1) blankets Venus. The relative role of three major contributors to the atmosphere, primordial nebular material, volcanic outgassing, and influx of volatiles by comets, in the formation of the present atmosphere is an open question. Although the current atmosphere is very dry, a minute quantity of water is still detectable. An important feature of water in the Venusian atmosphere is that the deuterium/hydrogen (D/H) ratio is 150 ± 30 times higher than is found in terrestrial water. If water on Earth represents a sample of primitive water on Venus, the Venusian D/H ratio suggests that, depending on the

escape flux of hydrogen and deuterium, originally Venus had 260 to 7700 times the current amount of water. Such an amount is equivalent to a global layer of water 4 to 115 m deep.

The dense and dry atmosphere on Venus results in a strong greenhouse effect that governs the conditions on the surface, leading to very high near-surface temperatures (~740 K) and equalizing the temperatures over the planet. Important consequences are the absence of both free water and water erosion, along with significantly reduced wind activity. Thus, the principal factors resurfacing Venus are volcanism and tectonics. Volcanism is the main mechanism driving the growth of the Venusian crust. The chemical compositions of the surface rocks have been measured at seven landing sites (Tables 2 and 3). The rock chemistry correlates with the compositions of terrestrial basalts, suggesting that volcanism on Venus is mostly basaltic.

Surface Population of Impact Craters

A study of the spatial density and distribution of impact craters is the principal means of understanding the age of the surface and the history of resurfacing of planetary bodies. There are 968 impact craters on the Venusian surface, making the mean crater density ~2 craters per 10^6 km^2. This value implies that the surface of Venus is relatively young; depending on the models of the flux of projectiles crossing the orbit of Venus, the surface is estimated to be from 300 to 750 My old.

A discovery of fundamental importance is that Venus lacks the densely cratered terrain characterizing significant portions of the surface of the Moon, Mars, and Mercury (Figure 2). Thus, Venus does not display the surface age dichotomy typical of the smaller terrestrial planets. Detailed study of the spatial distribution of surface craters, either by Monte Carlo simulations or by near-neighbor analyses, has shown that the distribution of impact craters is very close to a completely spatially random distribution (Figure 2). Another important characteristic of impact craters on Venus is their state of preservation. A global survey of craters has revealed that a majority appear to be morphologically pristine (Figure 3A), with only a small percentage either embayed by volcanic flows (about 2.5%; Figure 3B) or deformed tectonically (about 12%; Figure 3C). The small total number of impact craters on Venus, their predominantly pristine morphology, and their almost completely random spatial distribution may be explained by two alternative end-member models. In the first model, the catastrophic resurfacing model, the craters on Venus are considered to belong to a production population, with the planet now in the stage of accumulation of craters. This model proposes that the observable geological history of Venus (the last ~10–15% of the total history) began after a major planet-wide and relatively short 'catastrophic' episode of resurfacing that reset the crater record. The present crater population is thought to be accumulating since that time, in an environment of sluggish endogenous activity. Alternatively, the equilibrium-resurfacing model proposes that the actual crater population is in equilibrium with the current volcanic and tectonic activity that occurs within small areas ~400 km across. These two models imply fundamentally different histories and modes of geology on Venus. In the catastrophic model, the geological history is considered to be episodic, whereas the equilibrium-resurfacing model implies a steady-state (more Earth-like) history.

Table 1 Composition of the atmosphere of Venus

Atmosphere component	*Elevation above the surface (km)*		*Content*
	From	*To*	
CO_2	1.5	63	97 ± 4 vol. %
N_2	1.5	63	1.35 4.5 vol. %
H_2O	45	54	500 10000 ppm
H_2O	25	45	~500 ppm
H_2O	0	15	~20 ppm
O_2	0	42	<20 ppm
CO	0	42	28 ± 14 ppm
SO_2	0	42	130 ± 60 ppm
Ar	1.5	24	~100 ppm

Table 2 Contents of radiogenic elements in rocks on the surface of Venus[a]

Component	*Lander*				
	Venera 8	*Venera 9*	*Venera 10*	*Vega 1*	*Vega 2*
K_2O (wt.%)	4.8 ± 1.4	0.6 ± 0.1	0.4 ± 0.2	0.54 ± 0.26	0.48 ± 0.24
U (ppm)	2.2 ± 0.7	0.6 ± 0.2	0.5 ± 0.3	0.64 ± 0.47	0.68 ± 0.38
Th (ppm)	6.5 ± 0.2	3.7 ± 0.4	0.7 ± 0.3	1.5 ± 1.2	2.0 ± 1.0

[a]Determined by gamma spectroscopy during lander missions.

Major Surface Tectonic, Volcanic, and Volcano-Tectonic Features

Image and altimetry data reveal a rich array of volcanic and tectonic features on Venus, implying a complex geological history. In the practical absence of erosion on Venus, the features shaping its surface are well preserved and directly portray the volcanic and tectonic processes that have formed them. The most important terrains, tectonic structures, and volcanic material units making up the surface of Venus are listed in Table 4.

Table 3 Contents of major petrogenic elements in rocks on the surface of Venus[a]

	Lander		
Component (wt.%)	*Venera 13*	*Venera 14*	*Vega 2*
SiO_2	45.1 ± 3.0	48.7 ± 3.6	45.6 ± 3.2
TiO_2	1.59 ± 0.45	1.25 ± 0.41	0.2 ± 0.1
Al_2O_3	15.8 ± 3.0	17.9 ± 2.6	16.0 ± 1.8
FeO total	9.3 ± 2.2	8.8 ± 1.8	7.7 ± 1.1
MnO	0.2 ± 0.1	0.16 ± 0.08	0.14 ± 0.12
MgO	11.4 ± 6.2	8.1 ± 3.3	11.5 ± 3.7
CaO	7.1 ± 0.96	10.3 ± 1.2	7.5 ± 0.7
K_2O	4.0 ± 0.63	0.2 ± 0.07	0.1 ± 0.08
S	0.65 ± 0.4	0.35 ± 0.31	1.9 ± 0.6
Cl	0.3	0.4	0.3

[a]Determined by X ray fluorescence during lander missions. Note: sodium is not analyzed by the X ray fluorescence technique.

Major Tectonic Features

Tessera terrain The tessera terrain (Figure 4) is among the most deformed regions on Venus. Tesserae are defined as radar-bright, topographically elevated, equidimensional or elongated areas that are complexly deformed by at least two sets of coupled contractional (ridges) and extensional (grooves) tectonic structures. Tesserae cover ~8% of the surface of Venus; the size of individual tessera occurrences varies from a few hundred kilometres up to several thousand kilometres. The largest tesserae occur on the surface of major plateau-shaped highlands of Venus (Figure 1B). The smaller tesserae are within the midlands and only a few small fragments of a tessera occur on the floor of large lowland basins. Tessera terrain is concentrated in the equatorial zone and at high northern latitudes; there is a distinct paucity of tessera terrain south of about 30° S (Figure 5), where the vast low-lying plains dominate the surface of Venus. The elevated regions where the major rift zones are associated with coronae and large volcanoes represent another type of territory where

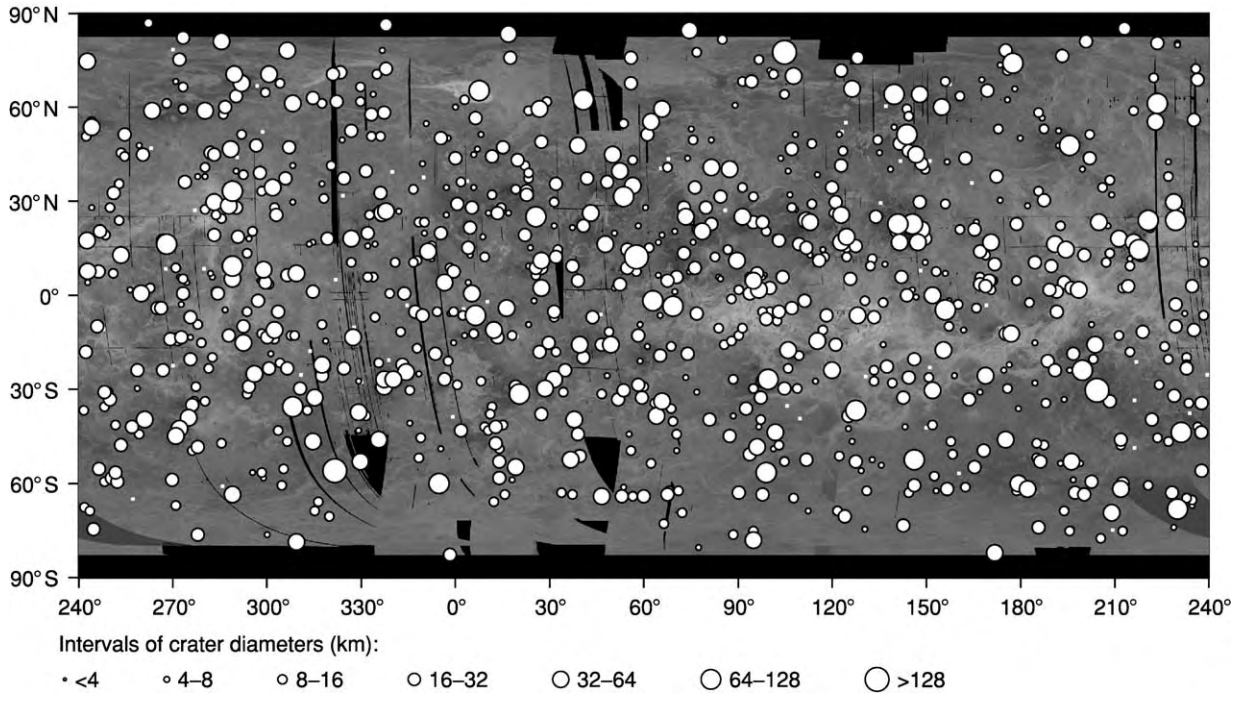


Figure 2 The spatial distribution of impact craters (white dots) on the surface of Venus is very close to a spatially random distribution. There is no evidence for the bimodal distribution of crater density characterizing the surface of Mercury, Moon, and Mars, where the heavily cratered terrains are in contact with lightly cratered plains. Crater size on Venus corresponds to the different intervals of diameters indicated at the bottom of the map. The map is in simple cylindrical projection.

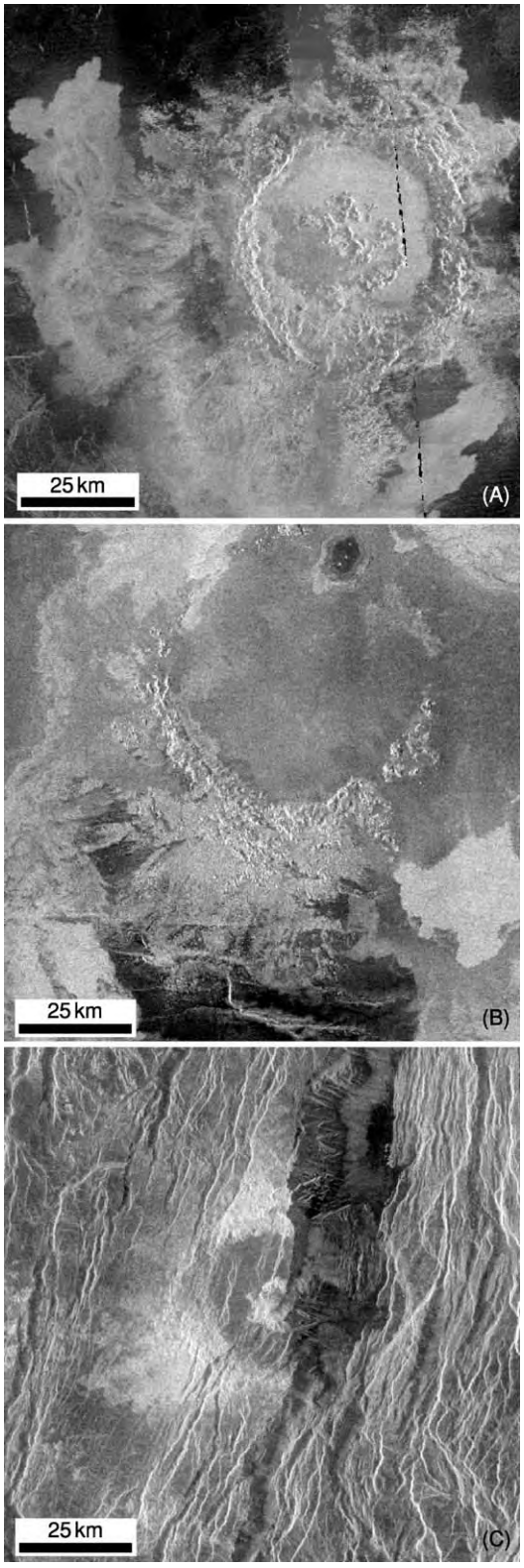


Figure 3 Morphology of impact craters on Venus. (A) Impact crater Caldwell (23.6° N, 112.4° E, 51 km) is characterized by a pristine morphology, a radar bright floor, and extensive outflows (south and west of the crater) superposed on surrounding plains. Fragment of C1 MIDR.30N117; illumination is from the right. (B) Impact crater Gautier (26.3° N, 42.8° E, 59.3 km) is heavily embayed by volcanic flows; only the southern half of the crater rim is visible. Dark material at the bottom of the image is regional wrinkle ridged plains on which ejecta from the crater are superposed. Younger lobate plains embay both the crater materials and the regional plains. Fragment of C1 MIDR.30N045; illumination is from the left. (C) Impact crater Balch (29.9° N, 282.9° E, 40 km) is severely deformed by tectonic structures of the Devana Chasma rift zone in Beta Regio. Only the western half of the crater is visible (center of the image); the eastern half is almost completely destroyed. A small fragment of the easternmost part of the rim is visible among individual graben of the rift zone. The position of the eastern rim suggests horizontal extension across the rift. Fragment of C1 MIDR.30N279; illumination is from the left.

tessera terrain is rare or absent. The surrounding plains embay practically all tessera massifs, meaning that the tessera is the oldest recognizable unit on Venus.

Densely fractured areas On the surface of Venus small areas are heavily dissected by numerous subparallel, densely packed, narrow and short lineaments (Figure 6). The lineaments that are wide enough to be imaged appear as fractures and imply a tectonic environment of extension during their formation. The dense fractures typically deform low-relief areas with an overall smooth surface, suggesting that these were initially lava plains. Densely lineated plains predominantly occur within midlands and there are almost no outcrops of the unit within vast areas of regional plains in lowland basins. The densely fractured areas and lineated plains occupy a small percentage of the Venusian surface and typically form small (tens of kilometres across) equidimensional, elongated, and arc-like occurrences (Figure 7) that sometimes form rims of coronae and corona-like features. Where the plains and tesserae are in contact, there is evidence for embayment of the tesserae by the material of densely lineated plains (Figure 8), suggesting that the plains are younger. Other plains units embay occurrences of densely lineated plains, which means that this unit represents one of the oldest terrains on Venus.

Ridge belts and ridged and grooved plains In some areas on Venus there are distinct belts consisting of swarms of contractional ridges (Figure 9). The linear and curvilinear ridges, 10–15 km wide and several tens of kilometres long, have smooth surfaces, rounded hinges, and appear to be symmetrical in cross-section. The ridge belts and individual ridges deform ridged and grooved plains and gradually merge with the surface of these plains, suggesting that they are tectonic facies of the same material unit. The ridged and grooved plains and ridge belts comprise less than 5% of the surface of Venus and

Table 4 Classification of the major features and terrains on Venus

Feature/terrain	Shape of occurrences	Dimensions	Areal distribution	Topographic characteristics		Associated tectonics	Associated volcanism
				Background	Relief		
Regional tectonic features							
Tessera	Equidimensional, elongated, irregular shape	A few 100s up to ~1000s of km across	Megaclusters, arcs; mostly in northern hemisphere	Midlands and highlands; rare in lowlands	Major regions are a few km high	Crisscrossing ridges and grooves	No evidence
Areas of dense fractures	Equidimensional, elongated, arc like	10s to a few 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Little relief	Dense narrow and short fractures	No evidence for contemporaneous volcanism
Ridge belts	Compact, elongated belts	100s to 1000s of km long, 10s of km wide	Major area is in a fan shaped area in the northern hemisphere	Midlands and lowlands; rare in highlands	Rises, several 100s of m	Linear ridges a few km wide and tens of km long	Associated with specific volcanic plains
Fracture belts	Belts and arc and star like occurrences	100s to 1000s of km long, 10s of km wide	Belts, swarms in both hemispheres	Midlands to lowlands	Rises (100s of m high) in places; central valley	Dense fractures and graben up to 100s of km long	In places, evidence for contemporane ous volcanism
Rift zones	Elongated zones	1000s of km long, 100s of km wide	Equatorial zone, Beta Atla Themis regiones	Highlands to midlands	Deep (a few km) depressions	Fractures and graben	Volcanic plains in places
Regional volcanic plains							
Densely lineated plains	Equidimensional, elongated, arc like	10s to a few 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Low relief	Dense narrow and short fractures	Plains volcanism, no discernible sources
Ridged and grooved plains	Elongated, belt like, equidimensional	100s (up to 1000s) of km long, 10s of km wide	Major area is in a fan shaped area in the northern hemisphere	Midlands to lowlands	Low relief	Deformed by ridges of ridge belts	Plains volcanism, no discernible sources
Shield plains	Equidimensional patches	10s to 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Low relief	Moderately deformed by wrinkle ridges	Plains volcanism with abundant small edifices
Wrinkle ridged plains (regional plains)	Vast extensions	Up to 1000s of km across	Equidimensional and elongated regional lowlands	Lowlands and midlands	N/A	Moderately deformed by wrinkle ridges	Plains volcanism, no discernible sources
Lobate plains	Equidimensional and elongated	100s of km across	Patchy regional distribution	Highlands; rarely in midlands	N/A	In places, cut by rift structures	Plains volcanism through distinct sources
Regional volcano-tectonic features							
Lakshmi Planum	Circular highland plateau	A few 1000s of km across	Unique feature in the northern hemisphere	Midlands	Several km high	Highest mountain ranges outside	Abundant volcanism inside
Coronae and arachnoids	Circular, equidimensional	100s to 1000s of km across; average is ~300 km	Chains, clusters, isolated features	Midlands; some are in highlands	Relief varies from local lows to local highs	Annulus of fractures; ridges are subordinate	Lava flows and lava fields
Novae	Star like, equidimensional	10s to 100s of km across; average is ~200 km	Isolated occurrences	Midlands; some are in highlands	Predominantly, topographic highs (100s of m)	Radial pattern of fractures and graben	Individual lava flows
Large shield volcanoes	Equidimensional	100s of km across	Isolated occurrences	Midlands to highlands	Several km high	In places, rift related graben	Abundant lava flows on flanks

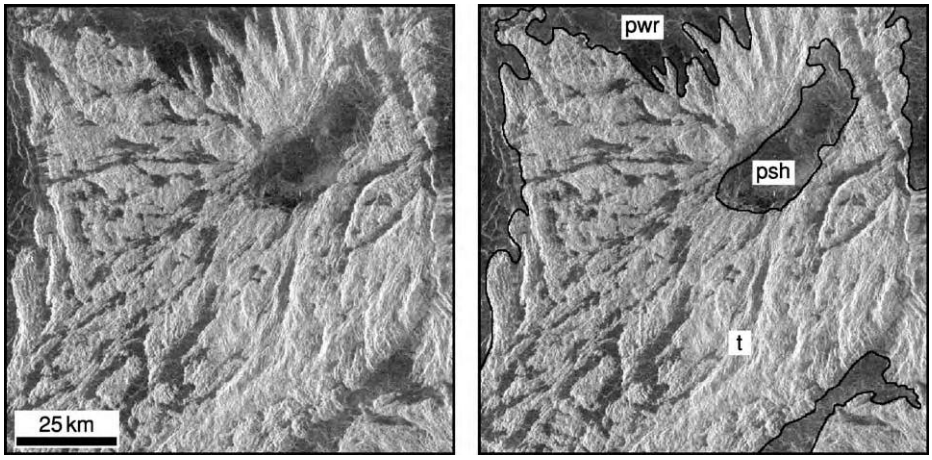


Figure 4 The surface of a tessera terrain (t) is deformed by several sets of tectonic features that completely obscure the morphological nature of the pre existing terrain. The tessera is embayed by all plains units on Venus. The units in this example are shield plains (psh) and wrinkle ridged plains (pwr). Fragment of C1 MIDR.30N009; illumination is from the left.

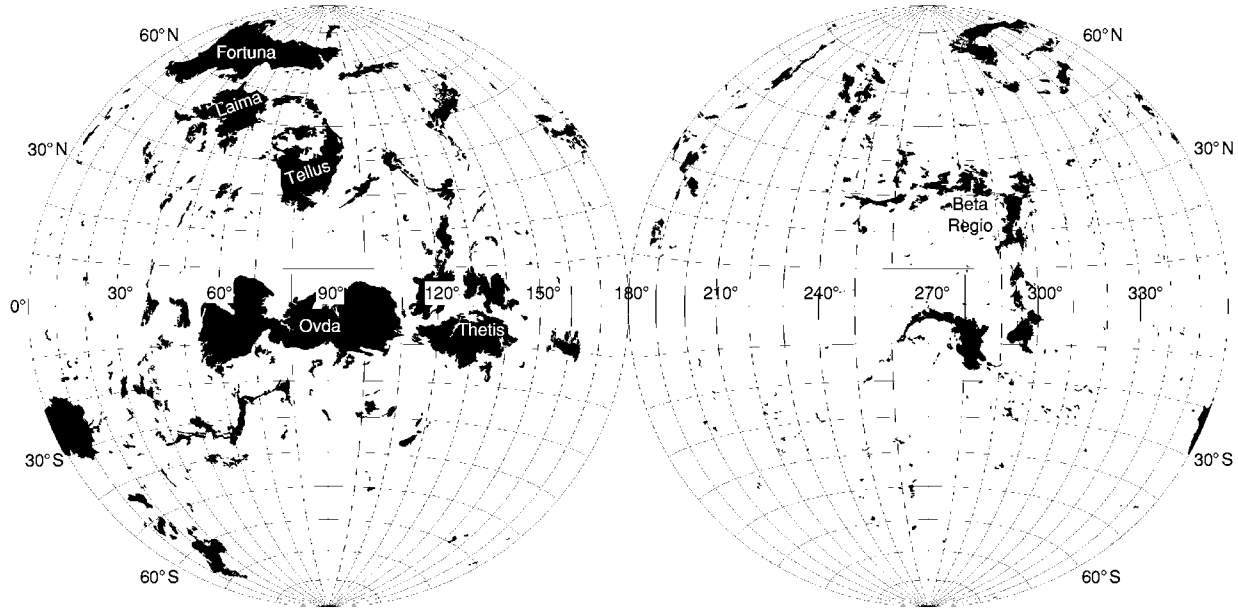


Figure 5 Map showing the global distribution of all occurrences of tessera terrains on Venus. The equatorial zone (Ovda and Thetis regiones) and high northern latitudes (Fortuna, Laima, and Tellus tesserae) have a higher density of tessera massifs. The hemisphere centred at about 230° E is dominated by major rift zones and large volcanoes and shows the clear scarcity of tesserae. The paucity of tesserae is also noticeable south of about 30° S. The map is in Lambert equal area projection.

occur predominantly in midlands, but in some cases are found on the floor of lowlands (e.g., Lavinia and Atalanta planitiae) (Figure 10). The most prominent zone of ridge belts extends for thousands of kilometres in the northern hemisphere, where individual belts are several hundreds of kilometres long and tens of kilometres wide (Figure 11). The plains embay tesserae and densely lineated plains, suggesting the younger age of both emplacement and deformation (Figure 12A and B).

Groove belts Groove belts are swarms of curvilinear lineaments that are usually wide enough to be resolved as fractures and graben (Figure 13), manifesting the tectonic environment of extension across long (hundreds of kilometres) and broad (up to a few hundred kilometres) zones. Individual structures of the belts can reach several tens of kilometres in length and are up to 1–2 km wide. The belts occupy a small percentage of the surface and in places form prominent zones thousands of kilometres long (Figure 10).

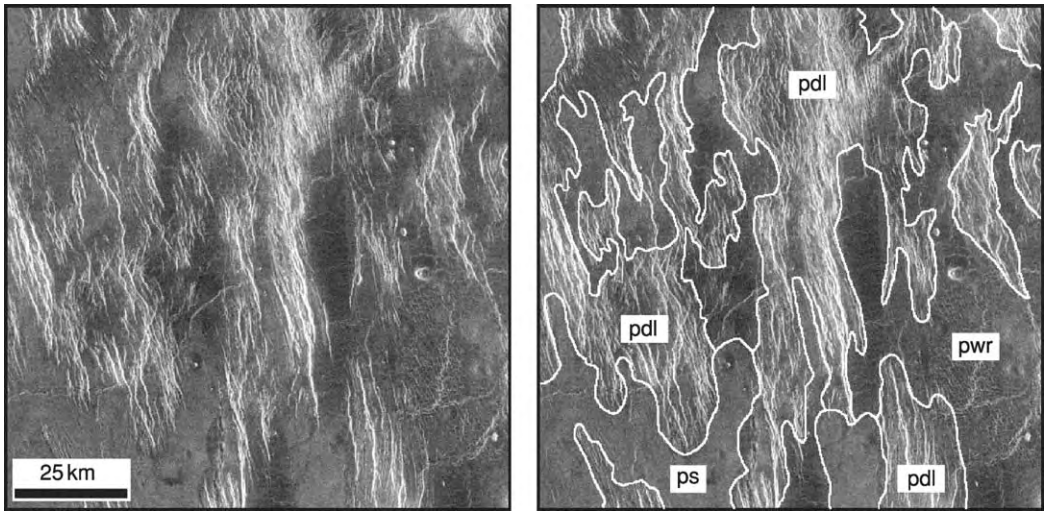


Figure 6 Areas of dense fracturing. Typically, small occurrences of these densely lineated plains (pdl) are heavily deformed by numerous narrow and densely packed lineaments (very narrow fractures). Fragments of densely lineated plains are embayed by varieties of lava plains, such as smooth plains (ps) and regional wrinkle ridged plains (pwr). Fragment of C1 MIDR.30N027; illumination is from the left.

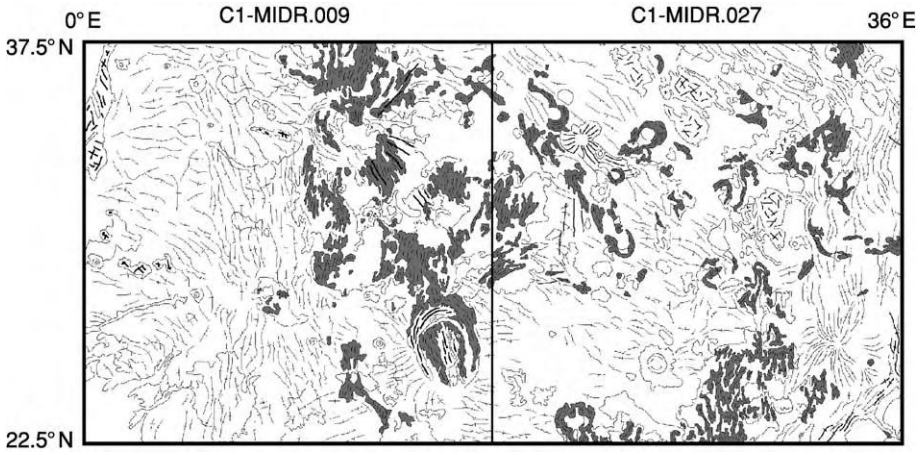


Figure 7 An example of the spatial distribution of densely lineated plains. Small fragments of densely lineated plains (dark grey) occur in clusters and as isolated patches within regional plains. In places, fragments of densely lineated plains occur within and near coronae and corona like features (north and south east of C1 MIDR.30N009).

Within the belts, the fractures are often anastomosing and sometimes form elliptical and circular coronae and corona-like features (Figure 14). Where groove belts are in contact with other units, they cut tesserae, densely lineated plains, and ridge belts, but are mostly embayed by younger plains such as shield plains and regional wrinkle-ridged plains (Figure 15).

Rift zones The most spectacular deformational belts on Venus are zones a few hundred kilometres wide and up to thousands of kilometres long, consisting of fractures and wide graben that can reach hundreds of kilometres in length and tens of kilometres in width (Figure 16). These features imply that the zones were formed by tensional stresses and in many aspects they resemble continental rifts on Earth. Topographically, the rift zones on Venus are troughs up to several kilometres deep that usually occur within regionally elevated areas. Preferentially, the rifts occur in a giant triangle-like area thousands of kilometres across, between Beta, Atla, and Themis regiones (the BAT province; Figure 17), where relatively young volcanic and tectonic activity on Venus is concentrated. The rift zones tend to occur with large dome-shaped rises; they are in close spatial association with lava plains and individual large lava flows, which appear to be

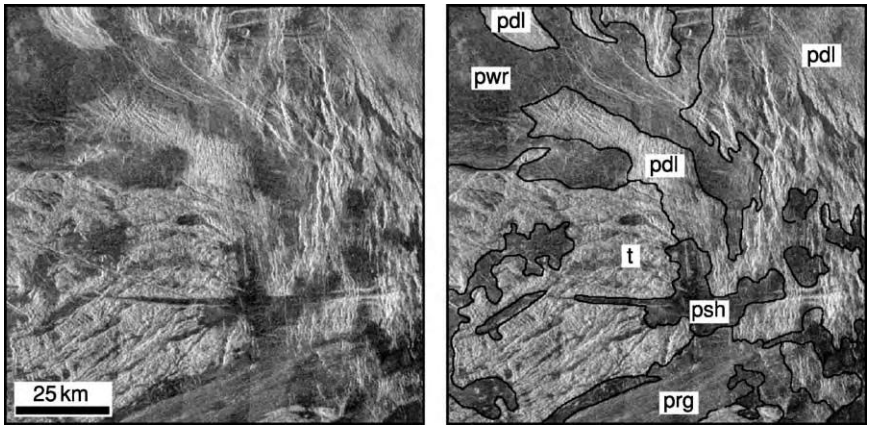


Figure 8 The relationships between tesserae (t) and densely lineated plains (pdl). A tessera has multiple sets of tectonic structures and a pdl is dissected by one set of narrow parallel lineaments. The complex pattern of deformation of tesserae is confined within its occurrences and does not penetrate into fragments of densely lineated plains. The pdl type structures deform tessera massifs (centre of the image). The younger shield plains (psh), wrinkle ridged plains (pwr), and ridged and grooved plains (prg) embay both tesserae and densely lineated plains. Fragment of C1 MIDR.30N125; illumination is from the left.

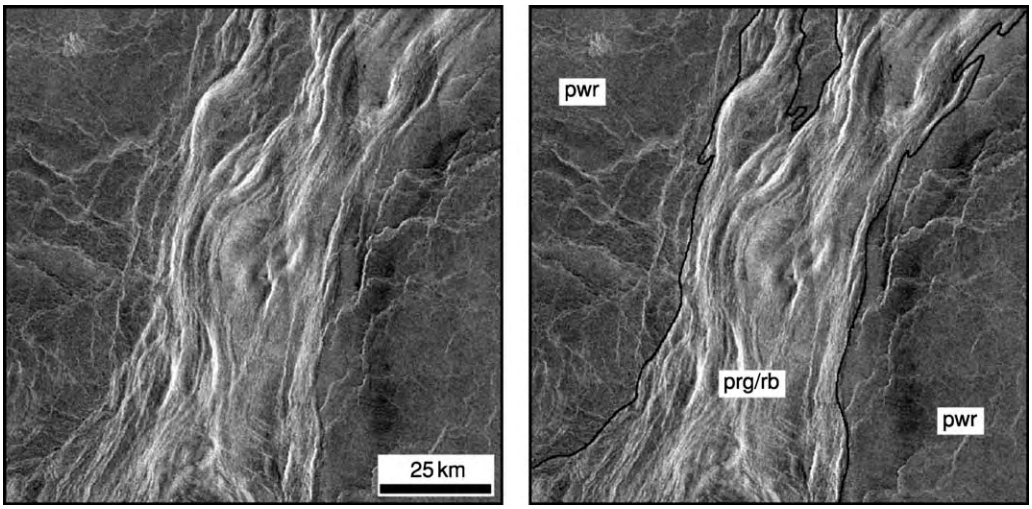


Figure 9 This typical ridge belt (rb) represents a small fragment of a larger elongated occurrence of ridged and grooved plains (prg). The curvilinear ridges typical of ridge belts are broader and less sinuous than wrinkle ridges within regional plains (pwr) are. Regional plains embay the ridge belt, implying that the belt was formed by contractional deformation of the material of ridged and grooved plains before emplacement of regional plains. Fragment of C1 MIDR.30N153; illumination is from the left.

the youngest volcanic features on Venus, and there is evidence for partly contemporaneous formation of the rifts and young volcanic plains.

Regional Plains on Venus

Due to conditions on Venus, volcanism is the prime factor contributing to growth of the crust on the planet, and extensive lava plains make up the vast majority of the surface. Several distinct units form extensive regional plains that are moderately deformed by tectonic structures.

Shield plains Shield plains are characterized by numerous small (up to 10 km across) shield- and cone-like features that are interpreted to be volcanic edifices (Figure 18). The surface of the plains is morphologically smooth but is sometimes deformed by wrinkle ridges. Shield plains cover about 10–15% of the surface of Venus and typically occur as small equidimensional areas several tens of kilometres across. Less frequently, occurrences of the plains are larger and can reach a few hundred kilometres (Figure 19A and B). The overall relief of the plains

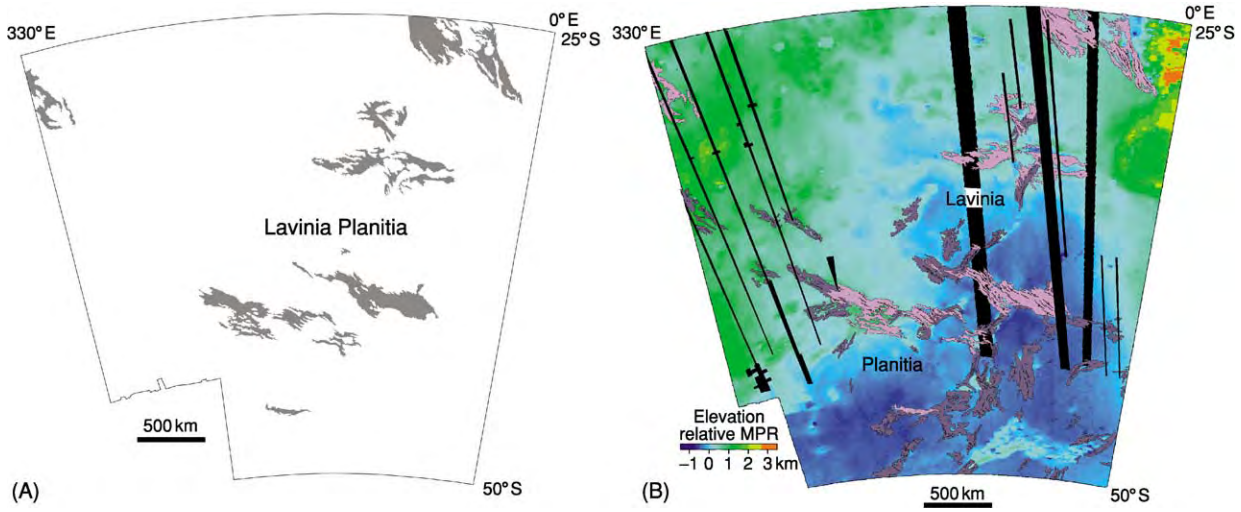


Figure 10 (A) An example of the spatial distribution of groove belts. Groove belts (dark grey) make major deformational belts on the floor of Lavinia Planitia. Individual occurrences of the belts can reach a 1000 km in length and several hundred kilometres in width. The map is in Lambert conformal projection. (B). An example of the spatial distribution of major deformational belts within the lowland of Lavinia Planitia. Occurrences of ridge belts (dark purple) are oriented in a north eastern direction. Groove belts (pink) are orthogonal to the strike of ridge belts. Both types of belts are concentrated within the deepest portion of Lavinia Planitia and the ridge belts are parallel to the elongation of the floor of the lowland. Colours show distribution of elevation relative to mean planetary radius (MPR; 6051 km). Black strips are data gaps. The map is in Lambert conformal projection.

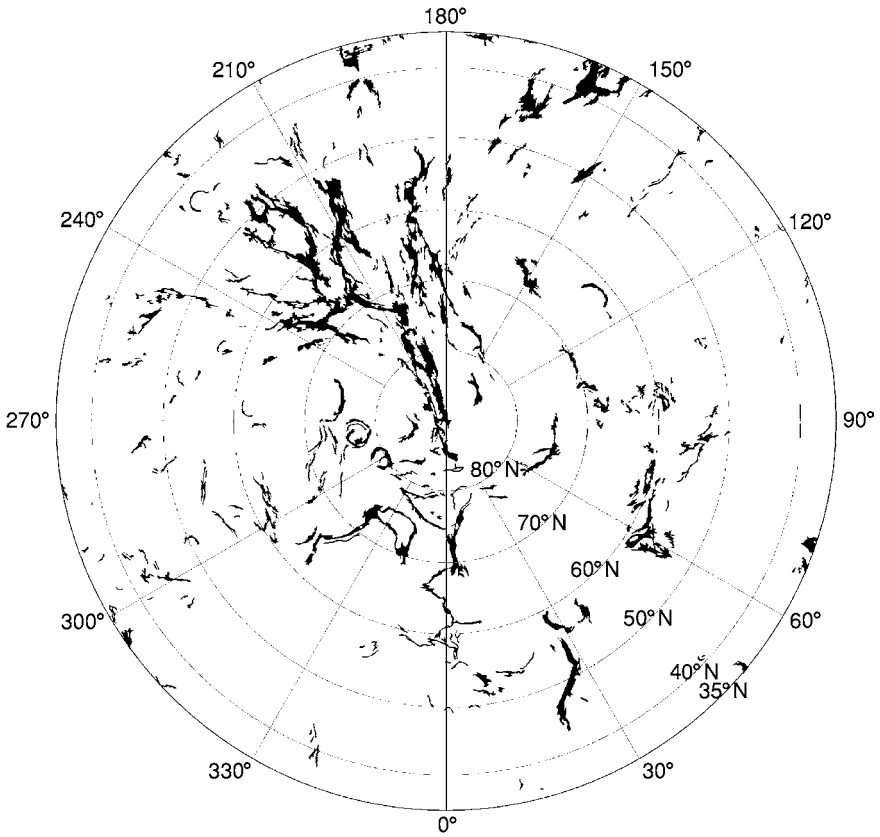


Figure 11 Distribution of ridged and grooved plains and ridge belts (black areas) in the northern hemisphere of Venus above 35° N. The most important occurrence of ridge belts is fan shaped and centred at about 210° E. The map is in polar stereographic projection.

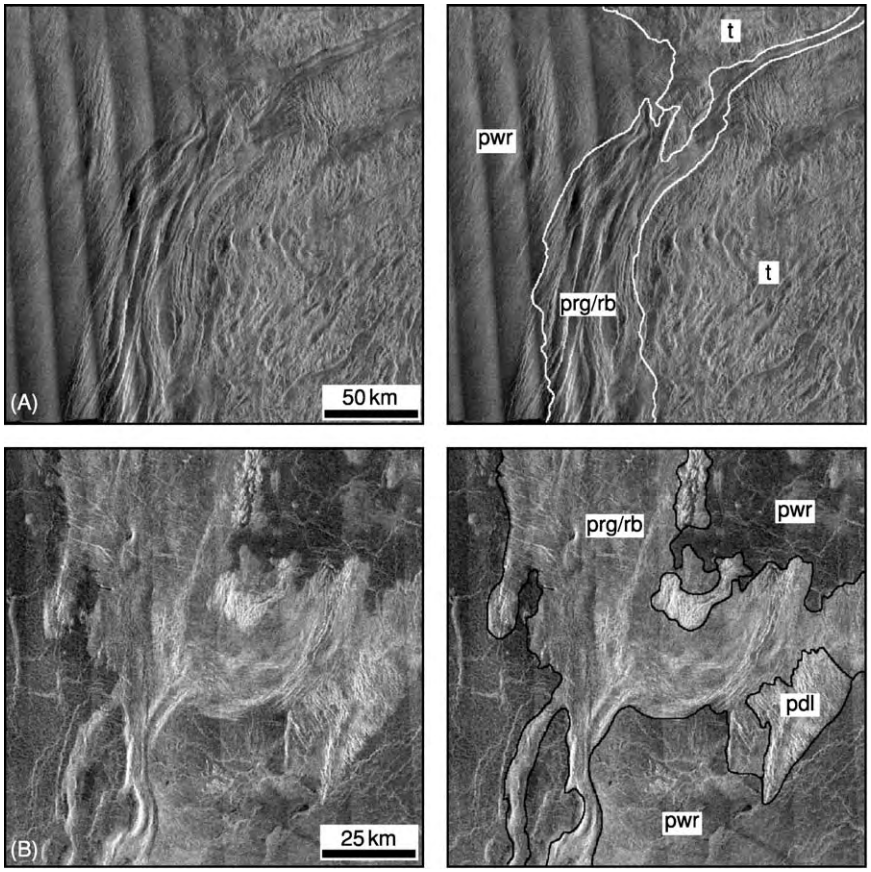


Figure 12 The relationships of ridge belts and ridged and grooved plains with other units. (A) Occurrences of ridged and grooved plains deformed into ridge belts (prg/rb) run along the edge of extensive tessera regions (t). The complex pattern of deformation within a tessera is confined within its massifs and appears to be cut by the system of ridges of the ridge belt. The surface of the regional plains (pwr) is moderately deformed, and plains material embays ridges of the belt. Fragment of C1 MIDR.00N112; illumination is from the right. (B) Material of the ridges and grooved plains, which is deformed in places into ridge belts (prg/rb), embays small fragments of densely lineated plains (pdl). Regional wrinkle ridged plains (pwr) broadly embay both prg/rb and pdl. Fragment of C1 MIDR.30N153; illumination is from the left.

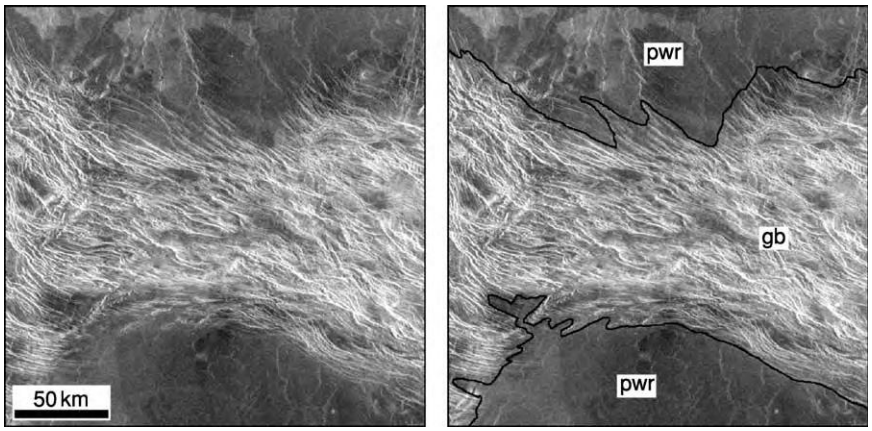


Figure 13 Typical groove belt consists of numerous linear and curvilinear fractures and graben that almost completely destroy pre existing materials (pwr, wrinkle ridged plains; gb, groove belt). Fragment of C1 MIDR.45S350; illumination is from the left.

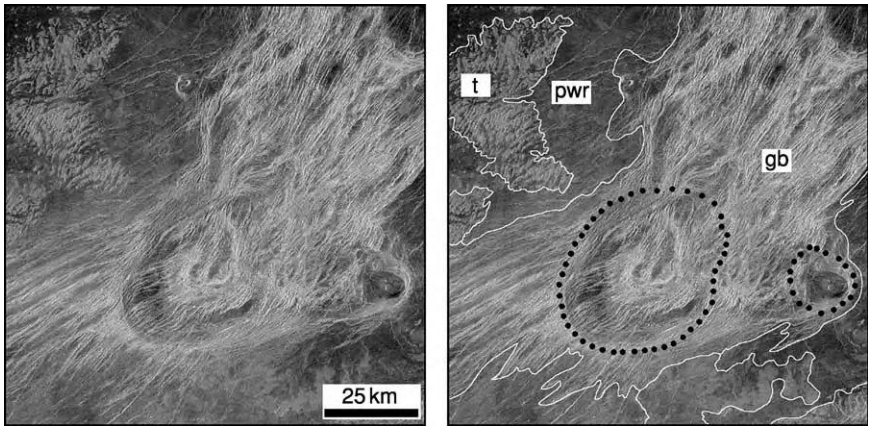


Figure 14 A fragment of a groove belt (gb), the structures of which outline coronae and corona like features (dotted lines in the right image). t, tessera; pwr, wrinkle ridged plains. Fragment of C1 MIDR.30N261; illumination is from the left.

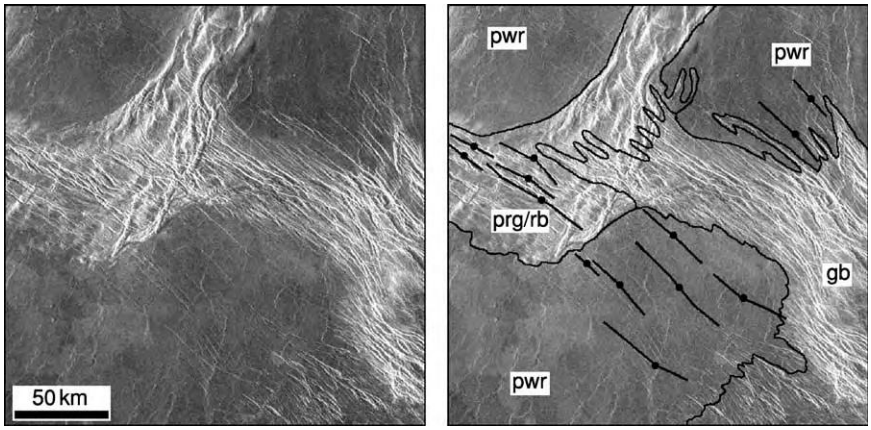


Figure 15 Relationships between groove belts (gb) and ridge belts (rb) typically show that fractures and graben of groove belts cut structures of ridge belts. Material of regional wrinkle ridged plains (pwr) embays both types of deformational belts. prg, ridged and grooved plains. Fragment of C1 MIDR.45S350; illumination is from the left.

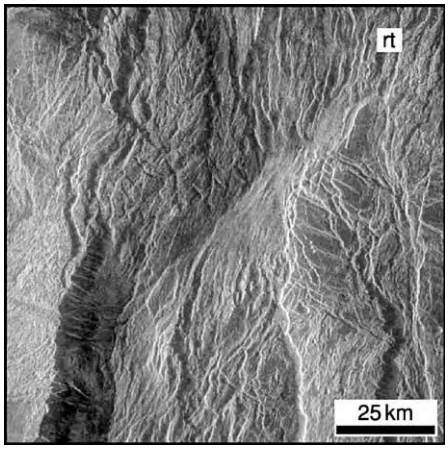


Figure 16 A portion of the Devana Chasma rift zone (rt) that cuts through the central portion of Beta Regio. The rift zone consists of a great number of fractures and graben, between which remnants of pre existing terrains are visible. Fragment of C1 MIDR.30N279; illumination is from the left.

appears to be hilly due to abundant shield features; occurrences of shield plains tend to be slightly higher compared to the surrounding regional plains. Although shield plains occur at different elevations, they preferentially occupy regional slopes away from old elevated terrains such as tesserae or ridge belts. The plains embay older, heavily tectonized units (Figure 20A) but are predominantly embayed by regional wrinkle-ridged plains (Figure 20B).

Regional wrinkle-ridged plains Regional wrinkle-ridged plains have morphologically smooth surfaces that are moderately deformed by numerous low, narrow, and sinuous wrinkle ridges (Figure 21). Wrinkle-ridged plains make up ~50–55% of the surface, appear as a regional background (with other units and structures being either older or younger), and can be traced almost continuously around the planet. The surface of the plains usually has a

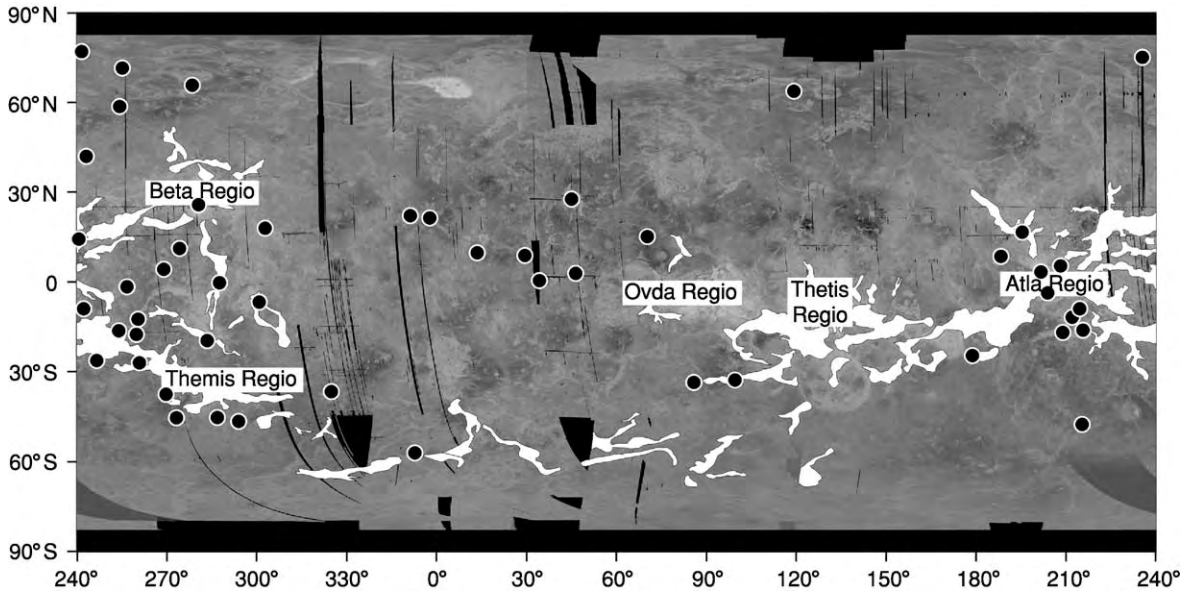


Figure 17 The global distribution of major rift zones (white) on Venus. The main concentration of rifts is within a giant triangle shaped area between Beta, Atla, and Themis regiones. The photobase is a low resolution synthetic aperture radar image of the Venus globe in simple cylindrical projection.

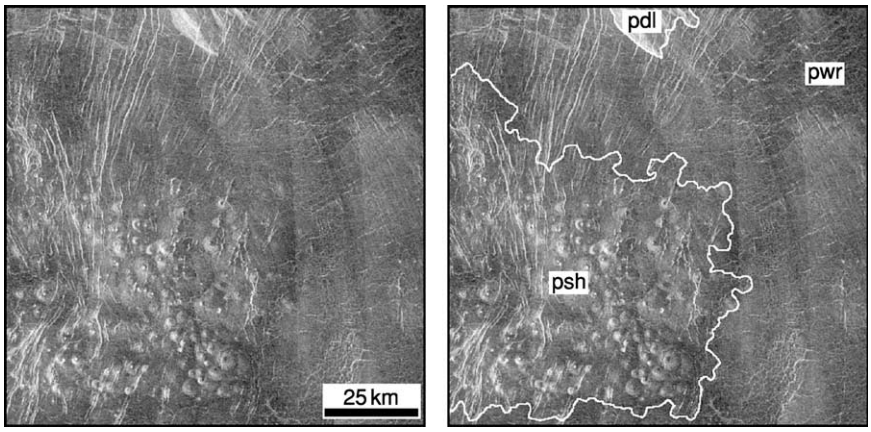


Figure 18 The surface of shield plains (psh) is characterized by a large number of small shield like features interpreted as volcanic edifices. Many of the shields have a bright dot in the centre; this is interpreted as a central pit (crater). The occurrence of shield plains is visible in the central part of Boala Corona. pdl, Densely lineated plains; pwr, wrinkle ridged plains. Fragment of C1 MIDR.30N135; illumination is from the left.

relatively low and uniform radar albedo with no visible flowlike features, which precludes identification of the sources of the plains material. A specific characteristic of the plains is the presence of long and narrow channels on their surface. The longest channel, Baltis Vallis, is about 7000 km long (Figures 21 and 22). In some places, the radar albedo of regional plains is distinctly higher and plains occur as relatively bright areas hundreds of kilometres across. These areas often surround distinct volcanic centers, such as large volcanoes and some coronae, and they form a distal apron of volcanic materials around them. Although in many cases there is evidence for embayment of the darker plains by the material of the brighter ones (Figure 23), the same family of wrinkle ridges appears to deform both varieties of regional plains (Figure 23). Regional plains cover the surface of large equidimensional basins (Figure 1) and make up the majority of midlands, but are noticeably less abundant within either plateau-shaped or dome-shaped highlands such as Ovda or Beta regiones.

Large volcanoes and lobate plains The large volcanoes are equidimensional mountains several hundred

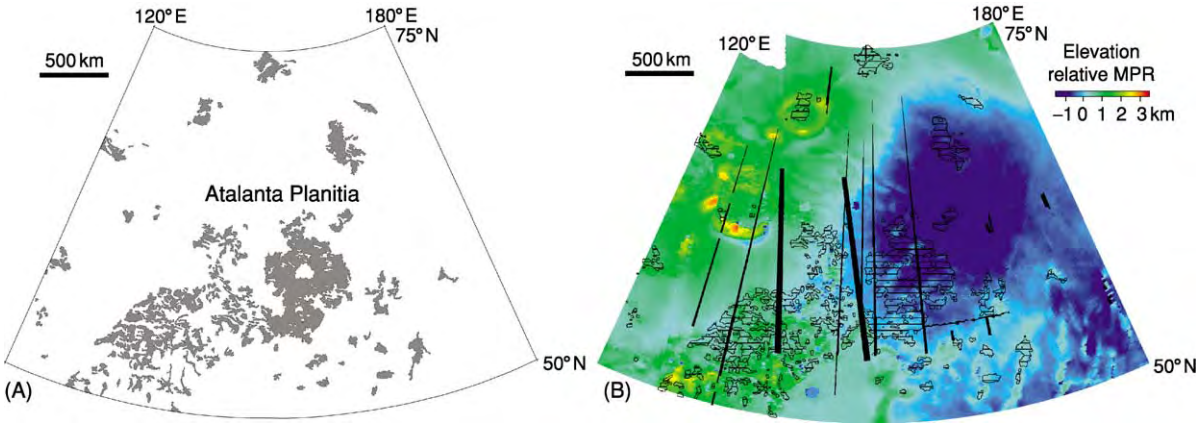


Figure 19 (A) The spatial distribution of shield plains (dark grey) within Atalanta Planitia. Fragments of the plains vary in size from several tens of kilometres up to a few hundred kilometres and occur in clusters and as individual patches. The map is in Lambert conformal projection. (B) The spatial distribution of shield plains (ruled pattern) within the Atalanta Planitia basin. Fragments of the plains vary in size from several tens of kilometres up to a few hundred kilometres and occur mostly on the regional slope of the Atalanta lowland, where groups of older units such as tesserae and densely lineated plains collectively form local highs. Colours show distribution of elevation relative to mean planetary radius (MPR; 6051 km). Black strips are data gaps. The map is in Lambert conformal projection.

kilometres across and a few kilometres high. Sometimes, a broad caldera-like feature is present at the summit of the volcano; in these cases, numerous radar-bright and dark flow-like features interpreted as lava flows always cover the slopes (Figure 24). The individual lava flows are superposed on each other and collectively form extensive lava plains (lobate plains), occurrences of which have lobate boundaries and can be several hundred kilometres across. There are 168 large volcanoes on Venus and their diameters vary from 100 to 1000 km. Typically, the large volcanoes associate with major rift zones and many of them occur within the Beta–Atla–Themis region (Figure 17).

Major Volcano-Tectonic Features

Lakshmi Planum Lakshmi Planum is a high-standing (2–4 km above MPR) plateau almost completely surrounded by mountain ranges (Figure 25). These ranges, the highest mountains on Venus, average 7–8 km in height (some reach 11 km). Lakshmi Planum, which is a few thousand kilometres wide, is so dissimilar to other types of highlands on Venus that it can be considered a specific class of topographic province. The interior of the Planum, flat and slightly tilted to south, is covered by smooth volcanic plains that are morphologically similar to vast regional wrinkle-ridged plains elsewhere on Venus. These plains embay both the tessera-like terrain within the plateau and the individual ridges at the base of the surrounding mountains. Two major volcanic structures, Colette and Sacajawea paterae, dominate the interior of Lakshmi and are the centers of the abundant lava flows that are superposed on the wrinkle-ridged plains inside the Planum; these plains are similar to the lobate plains on the slopes of large volcanoes.

Coronae, arachnoids, and novae Coronae, arachnoids, and novae are circular or quasi-circular features tens to hundreds of kilometres across. Coronae (Figure 26) and arachnoids (Figure 27) are characterized by concentric deformational annuli that predominantly consist of extensional structures (fractures and graben) and sometimes ridges (contractional features) and novae form starlike patterns of radial fractures; graben (Figure 28). In the catalogue of volcano-tectonic landforms compiled by Crumpler and Aubele in 2000, 209 coronae, 265 arachnoids, and 64 novae are listed. All of these features are thought to be the surface manifestations of mantle diapirs at different stages of evolution. Coronae and other circular volcano-tectonic features occur predominantly in the midlands and only a few of them are either within the high-standing plateaulike highlands and dome-shaped rises or in the lowlands (Figure 29). The topographic configuration of these structures varies from rimmed depressions to plateaus to dome-shaped features. Many coronae and novae are surrounded by prominent lava flows, suggesting that these features are distinct volcanic centers. However, there are neither medium-sized nor large volcanoes in association with these features. Lava flows are rarely associated with arachnoids and these structures appear mostly as tectonic structures. In many cases, regional wrinkle-ridged plains embay tectonic

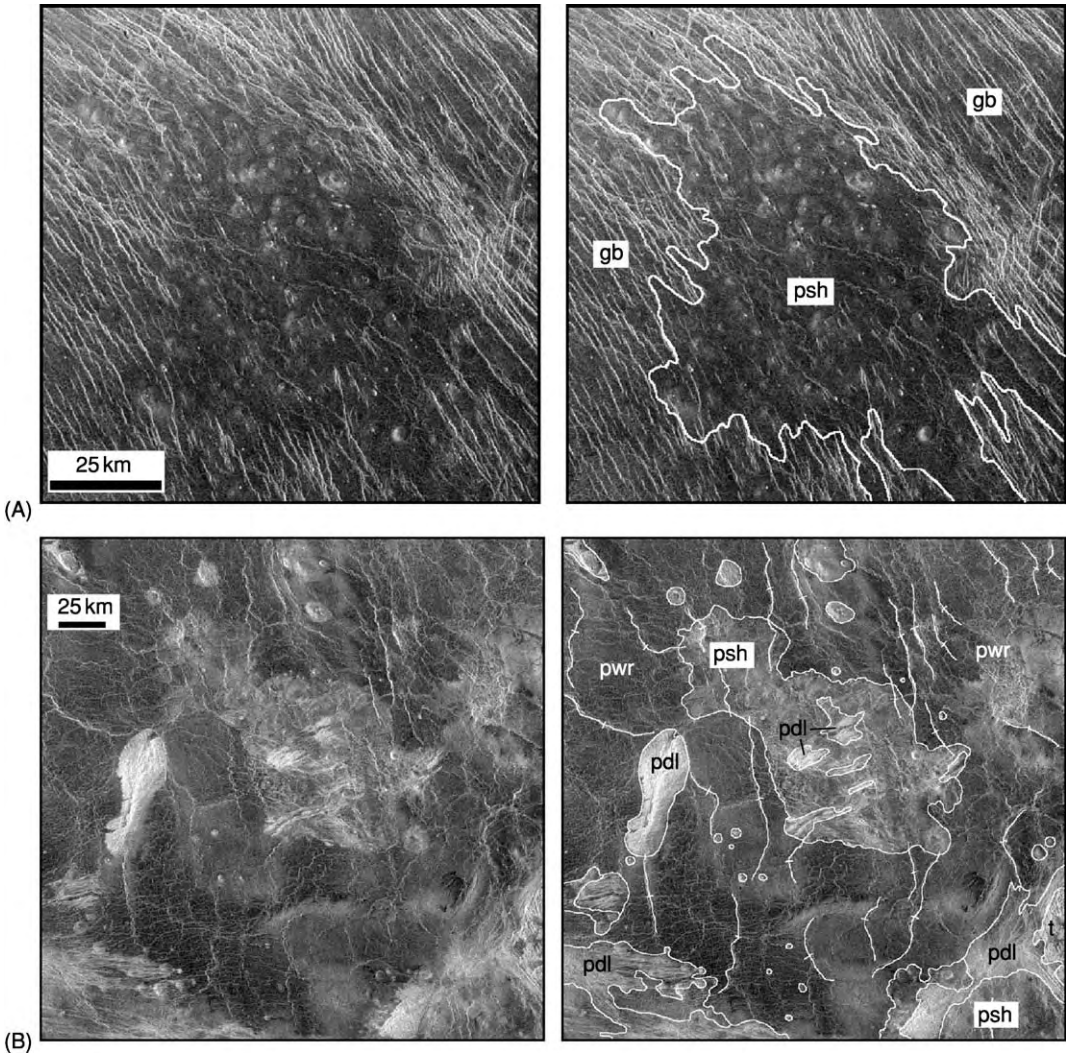


Figure 20 Relationships of shield plains (psh) with other units and structures. (A) Some portion of a groove belt (gb) is covered by deposits of shield plains that are obviously younger. A few shields, however, are cut by the fractures of the belt, suggesting that the formation of groove belt and shield plains partly overlapped in time. Fragment of C1 MIDR.30N333; illumination is from the left. (B) Occurrences of shield plains within regional plains are often characterized by a specific pattern of deformation confined within shield plains and a radar albedo that is different from the albedo of regional plains. Small individual shields that are seen within regional plains have morphological characteristics similar to these, typical of the main occurrences of shield plains. This suggests that the individual shields represent kipukas of more widespread shield plains covered by a mantle of wrinkle ridged plains (pwr).

structures of coronae and arachnoids, and fractures and graben of novae commonly cut the plains.

Major Topographic Features

The fact that the gravity and topography fields of Venus are highly correlated suggests that Venus may not have a low-viscosity layer, as in the asthenosphere of Earth, and that the Venusian mantle is strongly coupled with the lithosphere. Thus, the mantle circulation on Venus could be directly responsible for the formation of large-scale tectonic and topographic features. In the almost complete absence of erosion, on the other hand, the large-scale topographic features on Venus should much better reflect the balance between lithospheric buoyancy and mantle dynamics, compared to Earth. Thus the distribution of the major topographic features on Venus combines the present pattern of mantle convection with contributions from extinct patterns. The global altimetry data collected by Pioneer Venus, Venera-15/16, and Magellan show that three principal topographic provinces characterize the surface of the planet (Figure 1B). Lowlands (<0 km) make up $\sim$11% of the surface and consist of equidimensional basins and elongated depressions thousands of kilometres across. Their surface

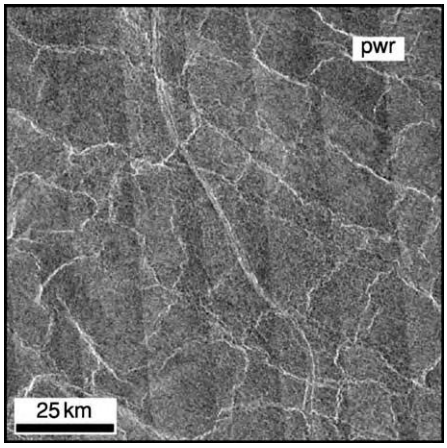


Figure 21 Regional plains with wrinkle ridges (pwr). The plains have a generally smooth surface with a relatively low and uniform radar albedo. Numerous narrow and sinuous ridges deform the surface of the plains. In the centre of the image, a narrow channel like feature that cuts the surface of the plains is visible. These channels are typical features on the surface of regional plains. Fragment of C1 MIDR.30N153; illumination is from the left.

is predominantly covered with regional wrinkle-ridged plains. Midlands constitute the majority of the surface of Venus (~80%), occur at elevations between 0 and 2 km, and host the richest variety of terrains, units, and structures.

Highlands are above 2 km and comprise ~9% of the surface. The highlands include two distinct classes of first-order features that are thousands of kilometres across. The first class consists of relatively steep-sided plateaulike features, the surface of which is typically covered by tesserae (e.g., Ovda Regio and Fortuna Tessera in eastern Ishtar Terra). These features appear to be isostatically compensated at relatively shallow depth, several tens of kilometres, suggesting that they are areas of thickened crust and probably relate to ancient regimes of mantle convection. This is consistent with the stratigraphic position of tesserae, which are the oldest terrain on the surface of Venus. The second class of highlands includes dome-shaped rises that are typically rifted and topped by large volcanoes (e.g., Atla and Ulfrun regiones).

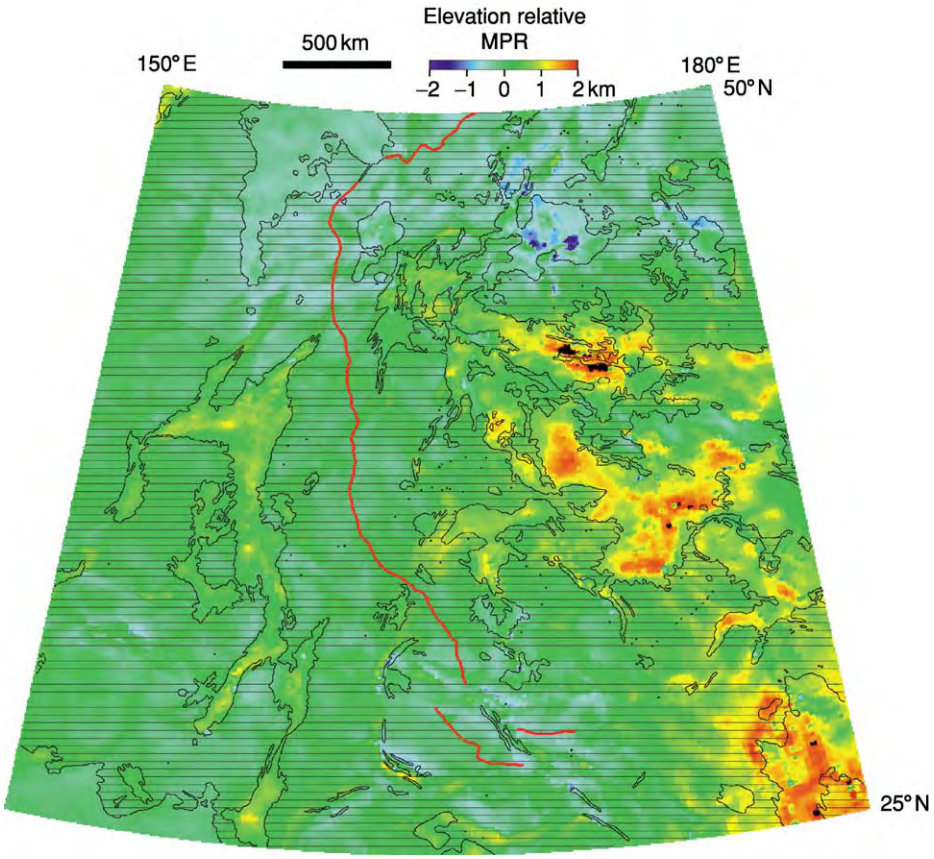


Figure 22 Spatial distribution of regional wrinkle ridged plains (ruled pattern). The plains occupy the relatively low portion of the area (Ganiki Planitia) between elevated territories to the west and east, where the older units and structures such as tesserae and ridge belts are exposed. The thick red line in the centre of the map is Baltis Vallis, which runs along the major continuous lowland. Colours show the distribution of elevation relative to mean planetary radius (MPR; 6051 km). The map is in Lambert conformal projection.

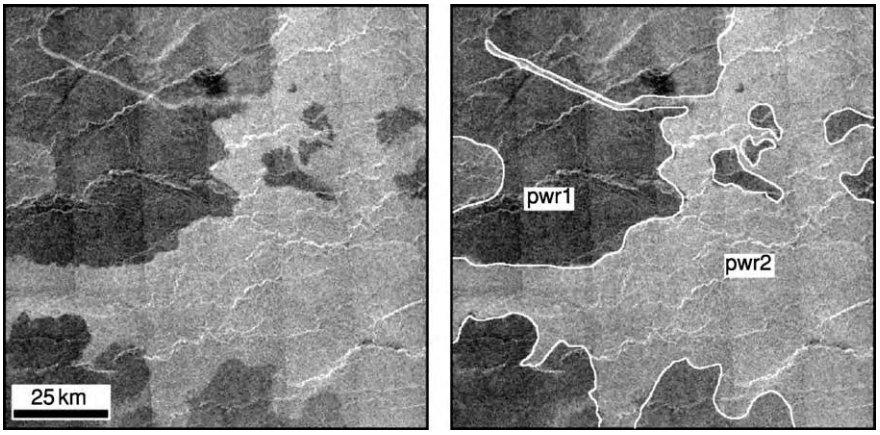


Figure 23 Varieties of regional wrinkle ridged plains. The upper member of the plains (pwr2) has a uniform and relatively higher radar albedo compared to the lower member of the plains (pwr1). Material of the brighter plains fills a portion of the lava channel (upper left), implying that pwr2 plains are younger. The same pervasive network of wrinkle ridges, however, deforms both varieties of the plains. Fragment of C1 MIDR.30N153; illumination is from the left.

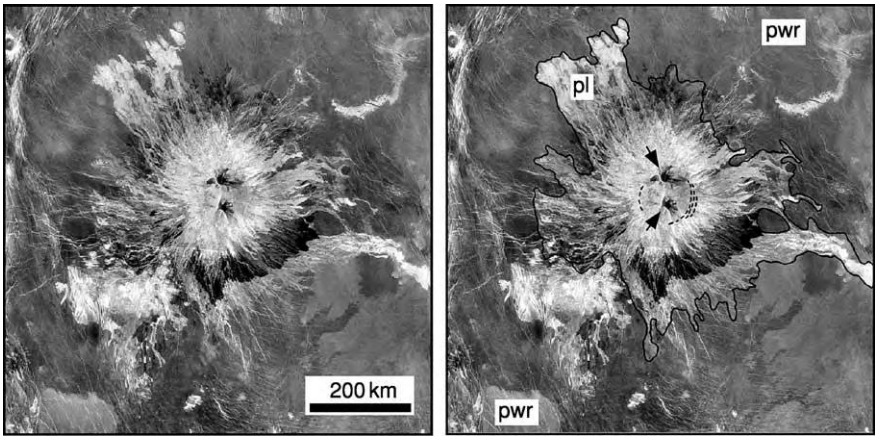


Figure 24 A large volcano (Sapas Mons) and lobate plains. Sapas Mons is a distinct volcanic centre from which issue a large number of radar bright and radar dark lava flows. In places, the flows merge with each other and form extensive lobate lava plains (pl) that are clearly superposed on the background of regional wrinkle ridged plains (pwr). Dashed lines (right image) show a series of arcuate graben in the summit area of Sapas Mons and black arrows indicate two steep sided domes. Mosaic of C1 MIDR.15N180, C1 MIDR.15N197, C1 MIDR.00N180, and C1 MIDR.00N197; illumination is from the left.

These rises appear to be compensated at much deeper levels, hundreds of kilometres, suggesting their dynamical support by active mantle upwelling. This is consistent with the geological characteristics of the rises, such as young rift structures and abundant young volcanism emerging through distinct sources.

Heat Loss Mechanisms

The style of volcanic and tectonic activity and the distribution of major topographic provinces are the specific manifestations of heat loss mechanisms operating on a planetary body. The global survey of the surface of Venus by Magellan has showed that except for a few possible sites, evidence for subduction is absent on the surface. Thus, the principal heat loss mechanism of Earth-like plate tectonics (lateral crust recycling, or "lateral" heat loss) apparently does not work on Venus. The alternative is a different orientation of the principal vector of heat loss mechanisms, vertical instead of horizontal. This means that Venus should be characterized by vertical crust accretion/recycling, or 'vertical' heat loss mechanisms. The manifestation of these is mantle upwelling and downwelling. The question of crucial importance in this context is the continuous or discontinuous nature of these mechanisms. Did they operate in a steady-state mode or did the temporal pattern of heat loss consist

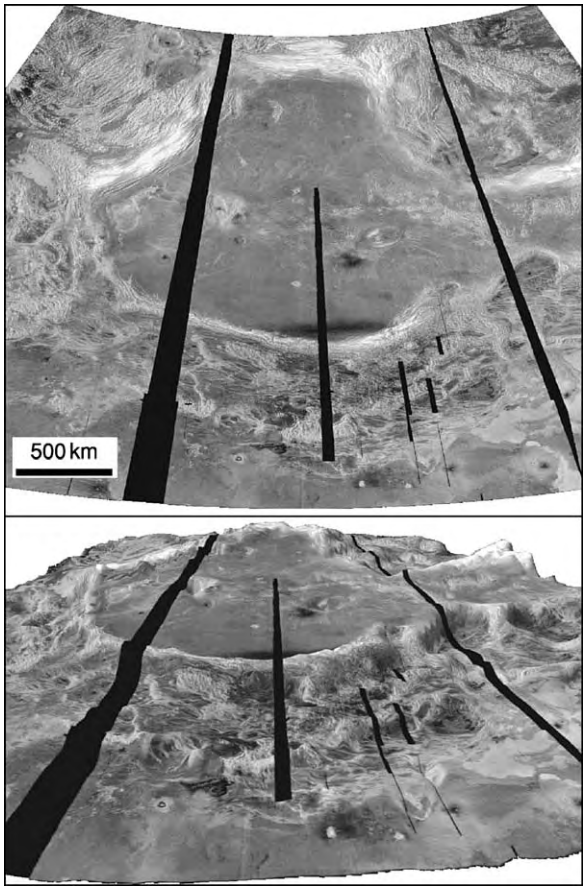


Figure 25 Lakshmi Planum, plan view (top) and perspective view (bottom). High mountain ranges (Danu Montes, south; Akna Montes, west and north west; Freyja Montes, north) almost com pletely border the interior of Lakshmi. Two large volcanic centres, Colette and Sacajawea Paterae, are distinct sources of relatively young volcanic materials (lobate plains) within the Planum. The plan view is in Lambert conformal projection.

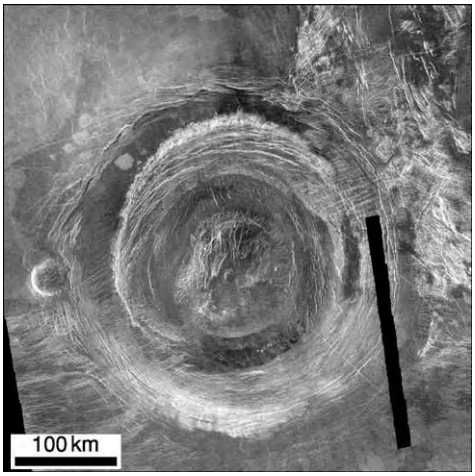


Figure 26 Corona Aramaiti is a typical example of this class of volcano tectonic structures. The corona has an outer and inner rim and a relatively flat floor populated with small shields. Within the outer rim, contractional ridges (northern portion of the rim) and extensional fractures (southern portion of the rim) are seen. Extensional features dominate the inner rim. A swarm of narrow lineaments (fractures and graben) appears to cut the outer rim (lower left and upper right) but disappears within the inner rim and the floor of the corona. Fragment of C1 MIDR.30S082; illumination is from the left.

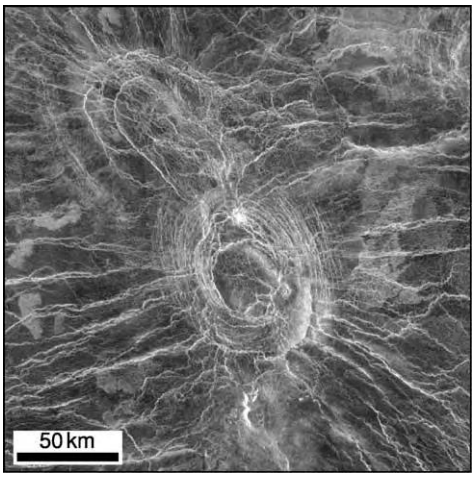


Figure 27 An example of an arachnoid. Swarms of concentric arcuate lineaments interpreted as fractures outline the doubled core of this feature (centre and upper left). Wrinkle ridges within regional plains appear to be focused at the arachnoid and form a radial pattern of structures around it. Fragment of C1 MIDR.45N011; illumination is from the left.

of a series of 'bursts' of endogenous activity intermittent with epochs of volcanic and tectonic quiescence, or was there one major change from a vigorous to a sluggish character of mantle convection?

The spatial distribution and morphology of impact craters place important constraints on the possible mode of heat loss/crust recycling on Venus. The catastrophic model of resurfacing is consistent with the characteristics of the crater population whereas the model of equilibrium resurfacing requires geological activity within small, ~400-km-diameter spots corresponding to the mean crater-to-crater distance. Thus, the characteristic horizontal scale of the mantle convection is also small, much smaller that the typical dimensions of many major features, both topographic and morphologic, on the surface. Thus, although the hypothesis of catastrophic resurfacing is an end-member model and almost certainly is incorrect in some details, it appears to describe better the geological situation on the surface of Venus.

There are two variants of the catastrophic resurfacing. In the first, the vertical accretion of crust and growth of the lithosphere lead finally to gravitational instability and large-scale delamination within the

lithosphere. Depending on the rheological properties of the material, horizontal scales of the instabilities, and the time-scale at which the instabilities exist, this process may lead either to transient plate tectonics or to large-scale mantle overturn. Both scenarios imply a cyclic nature of the heat loss mechanisms and may lead to a planet-wide resurfacing event on the surface.

The second variant of the catastrophic resurfacing is based on the secular cooling of the interior of the planet during most of the geological history of Venus. According to this scenario, the observable geological history of Venus begins after the transition from vigorous mantle convection under thin lithosphere to stagnant lid convection under thick lithosphere. In this model, tessera terrain is the remnant of the previous, thin-lithosphere regime, and the rifted dome-shaped rises topped with large volcanoes are the manifestation of the current regime of mantle convection under thicker lithosphere.

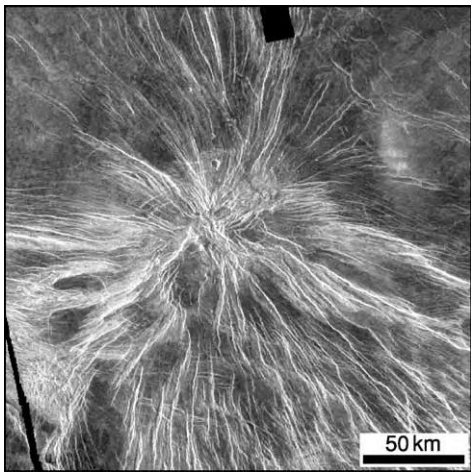


Figure 28 A typical nova is characterized by a star like pattern of broad radial fractures and graben originating at its centre. Fragment of C1 MIDR.30S279; illumination is from the left.

Models of Geological History on Venus

The high quality and global coverage of the Magellan data provide the possibility of detailed geological mapping of the surface based on defining distinct units and structures and establishing their relative ages. The results of the mapping efforts have led to two proposed end-member models for the correlation of regionally observed sequences of units and structures. In the first model, sequences of distinctive units mapped in different regions appear to have similar repetitive sequences in different places. This model has been called a "directional" geological history, implying a specific set of global trends in the evolution of Venus. For example, the consistently oldest relative age of tesserae suggests that a tectonic style of tessera formation has changed, with subsequent tectonic styles that led to the formation of other types of terrains. The important attribute of this model is its 'synchronism', implying that the sequences of events observed in different regions are broadly synchronous globally. For instance, regional wrinkle-ridged plains

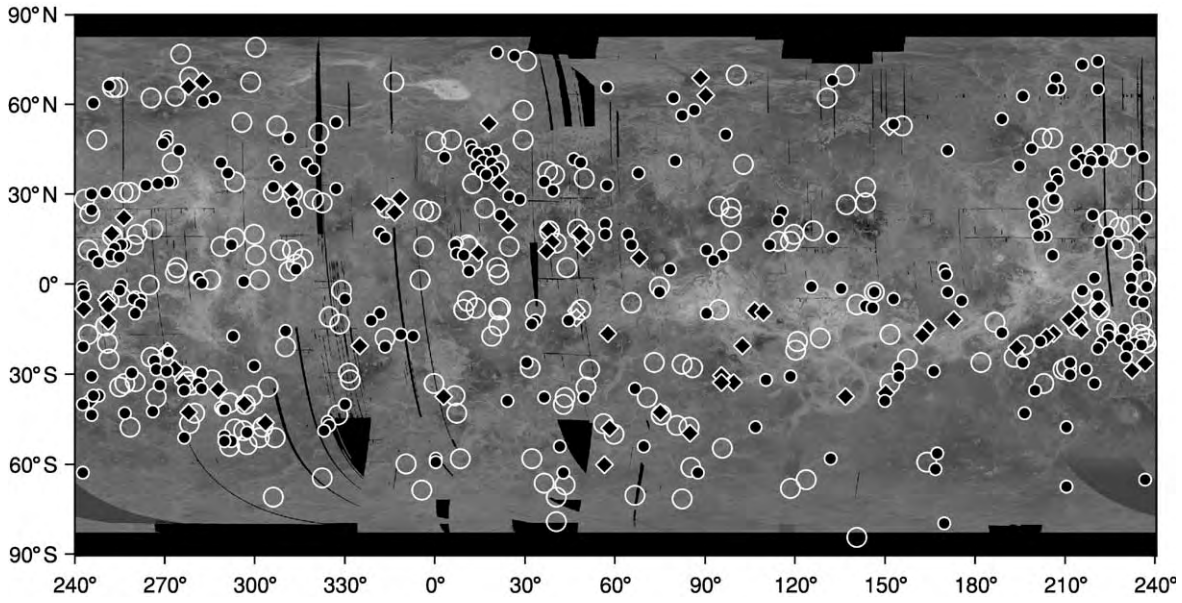


Figure 29 The spatial distribution of coronae (large circles), arachnoids (black dots), and novae (black diamonds) on the surface of Venus. The majority of these features, especially the novae, is concentrated in the Beta, Atla, and Themis regiones province and tends to be off of the large plateau shaped tessera highlands, such as Ovda and Thetis regiones. The map is in simple cylindrical projection.

appear to represent a broadly similar unit with a distinct stratigraphic position (either postdating or predating groups of other units) that can be traced continuously around the globe of Venus.

In the alternative model of geological history, the observed sequence of units is interpreted to be due to specific volcano-tectonic regimes that occur at different times on different parts of Venus, similar to Wilson cycles on Earth (individual plate-tectonic cycles that are repeated at different times and in different places on Earth). In this model for Venus, the local sequence of units represents only local or regional time-dependent sequential styles of endogenous activity. This is a 'non-directional' model of geological history, implying that the individual sequences represent local conditions occurring at different times in different places. Because the sequences of units and structures are almost the same in different regions on Venus, this model indicates that similar sequences of events resulted in similar stratigraphic columns occurring in these areas throughout the visible part of geological history. Another aspect of this model is its 'non-synchronous' nature, implying that the sequences of units/events are non-synchronous globally and that similar stratigraphic columns in specific regions are shifted relative to each other in terms of their age.

Further Investigation

The data collected during the exploration of Venus reveal the uniqueness of this planet. Venus does not have the surface age dichotomy characterizing the Moon and smaller planets such as Mercury and Mars. This fundamental characteristic implies that Venus, like Earth, has a prolonged history of geological activity that did not significantly decrease in intensity early in the evolution of the planet. In contrast to Earth, where the global heat loss mechanism is governed by plate tectonics, vertical crust accretion/recycling appears to be the principal style of geological activity on Venus.

Although current knowledge of Venus is great, there are still several major issues about its geology that are open to debate and further investigation. What is the evolution of the heat loss mechanisms on Venus? What are the paths of the evolution of the large-scale topographic features on the planet? How did the properties of the Venus lithosphere change as a function of time? Why are the Earth and Venus, the 'twin' planets, so different? What is the role of water in the evolution of both planets? Why is there little evidence on Venus for the presence of non-basaltic continental crust, which constitutes the major part of crustal material on Earth? How and when did the present atmosphere form and how has it evolved with time? How has the atmosphere interacted with the surface in recent and more ancient history of Venus? Obtaining answers to these questions requires continued exploration and key datasets, including seismic data, global high-resolution topography, *in situ* analysis of ancient terrains such as tesserae, and samples returned to terrestrial laboratories.

See Also

Earth Structure and Origins. **Solar System:** Mercury; Moon; Mars.

Further Reading

Barsukov VL, Basilevsky AT, Burba GA, *et al.* (1986) The geology and geomorphology of the Venus surface as revealed by the radar images obtained by Venera 15 and 16. *Journal of Geophysical Research* 91: D399 D411.

Bougher SW, Hunten DM, and Philips RJ (eds.) (1997) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment*. Tucson: University of Arizona Press.

Crumpler LS and Aubele JA (2000) Volcanism on Venus. In: Sigurdsson H, Houghton BF, McNutt SR, Rymer H, and Stix J (eds.) *Encyclopedia of Volcanoes,* pp. 727 769. San Diego: Academic Press.

Ford PG and Pettengill GH (1992) Venus topography at kilometer scale slopes. *Journal of Geophysical Research* 97: 13103 13114.

Hansen VL, Willis JJ, and Banerdt WB (1997) Tectonic overview and synthesis. In: Bougher SW, Hunten DM, and Phillips RJ (eds.) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment,* pp. 797 844. Tucson: University of Arizona Press.

Hauck SA, Phillips RJ, and Price MH (1998) Venus: Crater distribution and plains resurfacing models. *Journal of Geophysical Research* 103: 13635 13642.

Head JW, Crumpler LS, Aubele JC, Guest JE, and Saunders RS (1992) Venus volcanism: classification of volcanic features and structures, associations, and global distribu tion from Magellan data. *Journal of Geophysical Re search* 97: 13153 13197.

Ivanov MA and Head JW (2001) Geology of Venus: map ping of a global geotraverse at 30° N latitude. *Journal of Geophysical Research* 106: 17515 17566.

Masursky H, Eliason E, Ford PG, *et al.* (1980) Pioneer Venus radar results: geology from the images and altime try. *Journal of Geophysical Research* 85: 8232 8260.

McKinnon WB, Zahnle KJ, Ivanov BA, and Melosh HJ (1997) Cratering on Venus: models and observations. In: Bougher SW, Hunten DM, and Phillips RJ (eds.) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment,* pp. 969 1014. Tucson: University of Arizona Press.

Parmentier EM and Hess PC (1992) Chemical differenti ation of a convecting planetary interior: consequences for a one plate planet such as Venus. *Geophysical Research Letters* 19: 2015 2018.

Phillips RJ, Raubertas RF, Arvidson RE, *et al.* (1992) Impact craters and Venus resurfacing history. *Journal of Geophysical Research* 97: 15 923 15 948.

Schaber GG, Strom RG, Moore HJ, *et al.* (1992) Geology and distribution of impact craters on Venus: what are they telling us? *Journal of Geophysical Research* 97: 13 257 13 301.

Schubert G and Sandwell TD (1995) A global survey of possible subduction sites on Venus. *Icarus* 117: 173 196.

Simons M, Solomon SC, and Hager BH (1997) Localization of gravity and topography: constraints on the tectonics and mantle dynamics of Venus. *Geophysical Journal International* 131: 24 44.

Solomon SC, Smrekar SE, Bindschadler DL, *et al.* (1992) Venus tectonics: an overview of Magellan observations. *Journal of Geophysical Research* 97: 13 199 13 255.

Stofan ER, Sharpton VL, Schubert G, *et al.* (1992) Global distribution and characteristics of coronae and related features on Venus: implications for origin and relation to mantle processes. *Journal of Geophysical Research* 97: 13 347 13 378.

Strom RG, Schaber GG, and Dawson DD (1994) The global resurfacing of Venus. *Journal of Geophysical Research* 99: 10 899 10 926.

Sukhanov AL (1992) Tesserae. In: Barsukov VL, Basilevsky AT, Volkov VP, and Zharkov VN (eds.) *Venus Geology, Geochemistry, and Geophysics (Research Results from the USSR)*, pp. 82 95. Tucson: University of Arizona Press.

Surkov YA, Moskalyova VP, Kharyukova AD, Dudin AD, Smirnov GG, and Zaitseva SE (1986) Venus rock composition at the Vega 2 landing site. *Proceedings of the Lunar and Planetary Science Conference, Part 1, Journal of Geophysical Research* 9(supplement): E215 E218.

Turcotte DL (1995) How does Venus lose heat? *Journal of Geophysical Research* 100: 16 931 16 940.

Moon

P Moore, Selsey, UK

Introduction

The Moon is our companion in space. It is usually regarded as the Earth's satellite, though since it has 1/81 the mass of the Earth it may be better to class the Earth–Moon system as a double planet. The Moon is a world of craters, mountains, and wide plains always referred to as seas (maria), though there has never been any water in them; the craters are generally accepted as having been produced by impacting meteorites, and some are well over 200 km in diameter. There have been six successful manned missions to the Moon, and many unmanned probes have been sent there. This article presents a general survey of the Moon, and summarises what has been learned from the lunar space-craft.

Origin

The Moon is so close to the Earth that even with the naked eye the surface markings are obvious. Physical and orbital data are given in Table 1.

For many years it was believed that the Earth and the Moon were one body, and that rapid rotation resulted in a portion being flung off to form the Moon – leaving the hollow now filled by the Pacific Ocean. This theory has long since been rejected, and only two serious theories remain. It is possible that the Earth and the Moon were formed at the same time and in the same region of the 'solar nebula', the cloud of material surrounding the young Sun; certainly the Earth and Moon are of the same age – about 4.6 thousand million years. However, most authorities now favour the 'giant impact' theory; the original

Table 1 Lunar data

Distance from Earth	
centre to centre:	max 406 697 km (apogee)
	mean 384 400 km
	min 356 410 km (perigee)
surface to surface:	max 398 581 km (apogee)
	mean 376 284 km
	min 356 410 km (perigee)
Orbital period:	27.321661 days
Axial rotation period:	synchronous
Synodic period:	29.53 days (29 d 12 h 44 m 2 s.9)
Mean orbital velocity:	1.023 km/s
Orbital eccentricity:	0.0549
Mean orbital inclination:	5° 9′
Diameter:	
equatorial	3746 km
polar	3470 km
Oblateness:	0.002
Mean apparent diameter from Earth:	31′5″
Reciprocal mass, Earth 1:	81.301
Density, water 1:	3.342
Volume, Earth 1:	0.0203
Escape velocity:	2.38 km/s
Surface gravity, Earth 1:	0.1653
Mean albedo:	0.067
Atmospheric density:	10^{-14} that of the Earth's air at sea level.

body was struck by a 'planetary-sized body larger than Mars', so that the Moon condensed from the debris ejected during the collision.

Movements and Rotation

It is usually said that 'the Moon revolves round the Earth'. In fact the two bodies revolve together round the barycentre (the centre of gravity of the Earth–Moon system), but as the barycentre lies 1700 km below the Earth's surface the simple statement is good enough for most purposes.

The lunar orbit is not circular; the distance from Earth (centre to centre) ranges between over 400 000 km at furthest recession (apogee) and less than 360 000 km at closest approach (perigee). The orbital period is 27.32 days, but this is not the same as the synodic period, or interval between successive full moons or successive new moons, because the two bodies are moving together around the Sun; the synodic period is 29.53 days. It is often said that the Moon is 'new' when it appears as a slender crescent in the evening sky, but this is not strictly true; new moon occurs when the Moon lies between the Sun and the Earth, and its dark side is facing us, so that the actual new moon cannot be seen at all unless it passes directly in front of the Sun and produces a solar eclipse. A solar eclipse does not happen every month, because the lunar orbit is inclined at an angle of over 5°, and most new moons pass unseen either above or below the Sun.

During the crescent stage, the unlit hemisphere can usually be seen shining dimly. This is because of 'Earthshine' – light reflected on to the Moon from the Earth.

The axial rotation period of the Moon is exactly the same as its orbital period, so that the Moon always keeps the same hemisphere turned toward the Earth, and part of the surface is permanently turned away from us. This is not mere coincidence, it is the result of tidal effects over the ages. Originally the Moon spun much more quickly, but the tidal pull of the Earth slowed it down until the rotation had become 'captured' or synchronous. Note, however, that the Moon does not keep the same hemisphere turned sunward, so that day and night conditions there are the same all over the globe – apart from the fact that from the far side, the Earth can never be seen.

From Earth it is in fact possible to see 59% of the Moon's surface, though of course no more than 50% at any one time. This is because the Moon rotates at a constant speed, but the orbital velocity varies; the Moon moves quickest when near perigee. This means that the rotation and the position in orbit become 'out of step', and the Moon seems to rock very slowly to and fro. This libration in longitude means that narrow zones are brought alternately in and out of view. There are other minor librations, and only 41% of the surface is visually inaccessible, so that before the Space Age nothing definite was known about it.

Because of tidal effects, the Moon is receding from the Earth at the rate of 3.8 centimetres per year, and on average the Earth's rotation period is lengthening at a rate of 0.0000002 second per day.

Structure and Atmosphere

The outer surface layer of the Moon is termed the regolith; it is a loose layer or debris blanket probably, up to 10 metres deep in places, continually churned up by the impacts of micrometeorites (**Figure 1**). It is often referred to as 'soil', but this is misleading, as it contains no organic material. It overlies the rocky crust, which is on average just over 60 km in depth – thicker on the far side of the Moon than on the Earth-turned hemisphere. However, no unbrecciated rock outcrop has been encountered on Apollo missions and it remains uncertain how deep the zone of brecciation penetrates. At a fairly shallow level there are areas of denser material, which have been located because they affect the movements of orbiting spacecraft, they are called mascons (mass concentrations), and lie below some of the regular maria and very large basins. Below the crust lies the mantle, the structure of which may be fairly uniform; at a depth of around 1000–1200 km there seems to be a region where the rocks are hot enough to be molten. Finally, there may be a metallic core, no more than 1000 km in diameter; results from the Lunar Prospector spacecraft of 1998–1999 led to an estimate of an iron-rich core between 440 and 900 km in diameter. Most of what we know about the lunar interior comes from studies of moonquakes, which do occur and have been recorded by instruments left on the surface by the Apollo astronauts; some are shallow, but most occur in a zone from 800 to 1000 km below the surface. By terrestrial standards they are very mild, and never exceed a value of 3 on the Richter scale (*see* **Tectonics:** Earthquakes).

The Moon's low escape velocity means that there is only a trace of atmosphere, made up chiefly of helium and argon. If the entire atmosphere were compressed to a density equal to that of the Earth's air at sea-level, it would just about fill a cube with a side length of 65 metres. There is no detectable magnetic field, though the remnant magnetism of some rocks indicates that a definite field may have existed over 3.5 thousand million years ago.

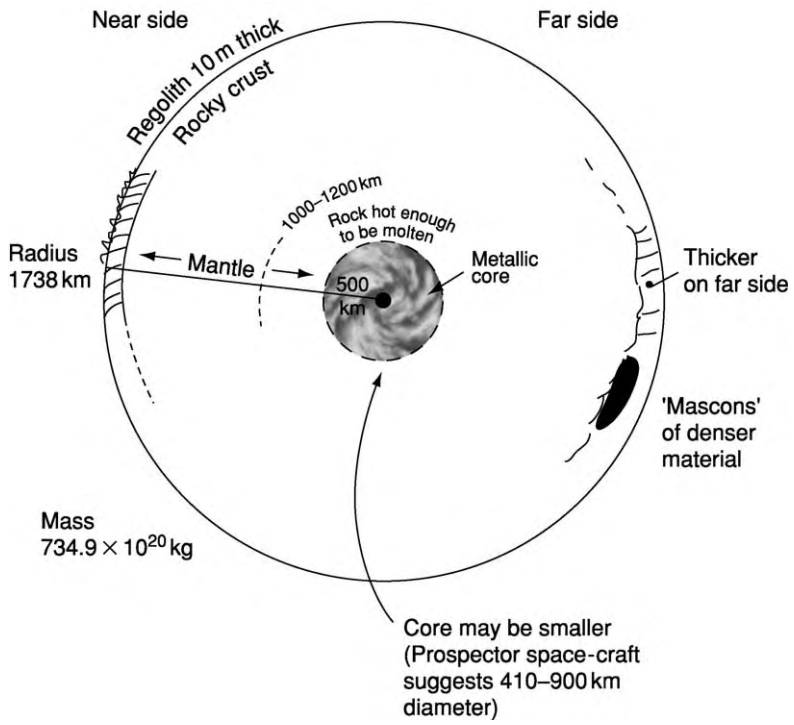


Figure 1 Structure of the Moon.

Table 2 Selected list of successful lunar missions

	Launch date	*Landing date*	
Name	*United States*		*Area*
Ranger 7	28 July 1964	31 July 1964	Mare Nubium hard lander, 4306 images returned.
Surveyor 1	30 May 1966	2 June 1966	Mare Nubium, near Flamsteed. Controlled landing. 11/150 images returned.
Surveyor 2	7 Jan 1968	9 Jan 1968	Nrim of Tycho. Controlled landing: 21 274 images
Orbiter 1	10 Aug 1966		207 images. (Uncontrolled impact, 29 Oct 1966.)
Orbiter 5	1 Aug 1967		212 images returned. (Impact, 31 Jan 1968.)
Clementine	25 Jan 1994		Mapping. Left orbit 3 May 1994.
Prospector	6 Jan 1998		Mapping, analysis. Uncontrolled impact, 11 July 1999.

Apollo manned missions

Number	*Landing date*	*Area*	*Astronauts*
11	19 July 1969	M. Tranquillitatis	N. Armstrong, E. Aldrin, J. Collins
12	19 Nov 1969	Oc. Procellarum	C. Conrad, A. Bean, R. Gordon
14	2 Feb 1971	Fra Mauro	A. Shepard, E. Mitchell, S. Poosa
15	10 July 1971	Hadley Apennines	D. Scott, J. Irwin, A. Worden
16	21 Aug 1972	Descartes formation	J. Young, C. Duke, T. Mattingly
17	11 Dec 1972	Taurus Littrow	E. Cernan, H. Schmitt, R. Evans

Lunar Missions

Many space-craft have been sent to the Moon, quite apart from the manned Apollo missions. A list of the most important probes is given in Table 2.

The first successful missions were Russian, in 1959; in October of that year Luna 3 sent back the first images of the far side, always turned away from the Earth. On 3 February 1966, Luna 9 made the first controlled landing, showing that the surface was firm and disposing of an earlier theory that the maria at least were covered with deep dust. Controlled landings were made by the American Surveyors

(1966–1968) and the entire surface was mapped by the five Orbiters (1966–1967). The Apollo Programme extended from 1968 to 1972; of the seven planned landings, only Apollo 13 was unsuccessful. The Russians sent two automatic 'rovers', the Lunokhods (1970 and 1973). The latest lunar mapping probes have been the American Clementine (launched 1994) and Prospector (launched 1998, deliberately crashed on to the surface in 1999). Altogether 382 kg of samples have been returned to Earth, mainly by the Apollo astronauts but with small amounts from four Russian sample-and-return missions.

Surface Features

The most obvious features are, of course, the 'seas' (maria), which cover about 17% of the surface. A list of the major maria is given in Table 3. Most of these form a connected system, though the well-defined Mare Crisium is separate. Some of the maria are fairly regular in outline; others are very irregular. The largest of the regular maria is the Mare Imbrium, with a diameter of over 1000 km, bounded in part by the mountain ranges of the Apennines, Alps, and Carpathians, though the irregular Oceanus Procellarum has a longest diameter of over 2500 km and an area of well over 2 million square km. Of the major seas, only the Mare Orientale extends on to the far hemisphere of the Moon; there are no large maria wholly on the areas never visible from Earth. A small part of the Mare Orientale can be seen from Earth under favourable conditions of libration.

The whole lunar scene is dominated by craters, ranging from tiny pits up to huge enclosures well over 200 km in diameter (Figure 2). Basically they are circular, though often distorted by later formations, and near the limb they are foreshortened so much that it is sometimes hard to distinguish a crater wall from a ridge. A typical crater has a sunken floor, often with a central peak, and walls which rise to only a modest height above the outer surface. In profile, a lunar crater resembles a saucer rather than a steep-sided mine-shaft, and large formations would be better known as walled plains. Some have dark, relatively smooth floors, such as Plato in the region of the Alps (Figure 3), others have massive central structures which, however, never equal the height of the surrounding rampart. Some craters, notably Tycho in the southern uplands and Copernicus in the Oceanus Procellarum, are the centres of systems of bright rays which extend for hundreds of kilometres, and cross all formations in their path. A crater is at its most spectacular when seen at the terminator (the boundary between the sunlit and night hemispheres), as its floor will be wholly or partly filled with shadow. Near full moon there are almost no shadows, and the scene is dominated by bright rays, so that even a large crater will be difficult to identify unless it has a dark floor or very bright walls.

Large craters are often found in chains, such as the prominent. Ptolemæus, Alphonsus, and Arzachel, near the centre of the Earth-turned hemisphere, and Theophilus, Cyrillus, and Catharina, near the border of the Mare Nectaris. Chains of small craters are

Table 3 Selected list of Lunar maria

Name		*Diameter, km*	
Mare Australe	Southern Sea	600	Irregular, patchy area near SE limb
Mare Crisium	Sea of Crises	500	Well defined, separate
Mare Fœcunditatis	Sea of Fertility	900	Irregular; confluent with M. Tranquillitatis
Mare Frigoris	Sea of Cold	1600	Elongated, irregular, in places narrow
Mare Humboldtianum	Humboldt's Sea	270	Limb sea beyond Endymion
Mare Humorum	Sea of Humours	390	Regular; leads off Mare Nubium
Mare Imbrium	Sea of Showers	1120	Regular: area 863 000 sq km.
Mare Nectaris	Sea of Nectar	330	Leads off Mare Tranquillitatis. Fairly regular
Mare Nubium	Sea of Clouds	750	Ill defined border
Mare Orientale	Eastern Sea	340	Vast ringed structure; mainly on far side
Oceanus Procellarum	Ocean of Storms	2570	Irregular
Mare Serenitatis	Sea of Serenity	700	Regular. Few craters
Mare Smythii	Smyth's Sea	370	Well defined limb sea
Mare Tranquillitatis	Sea of Tranquillity	870	Adjoins M. Serenitatis
Mare Vaporum	Sea of Vapours	250	SE of the Apennines
Lacus Mortis	Lake of Death	150	Adjoins Lacus Somniorum
Lacus Somniorum	Lake of the Dreamers	380	Irregular darkish area leading off Mare Serinitatis
Palus Putredinis	Marsh of Decay	160	Part of Mare Imbrium
Palus Somnii	Marsh of Sleep	290	Curiously coloured area near Mare Crisium
Sinus Æstuum	Bay of Heats	290	Fairly dark area leading off Mare Nubium
Sinus Medii	Central Bay	260	Almost central on the disk
Sinus Roris	Bay of Dew	400	Joins Mare Frigoris to Oceanus Procellarum

Figure 2 Craters on the 'far side' of the Moon. The hemisphere always turned away from Earth is as crater scarred as the familiar hemisphere, though it lacks 'seas' similar in type to Mare Imbrium.

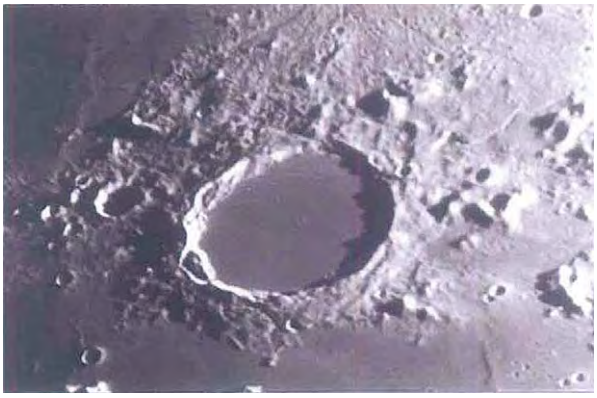


Figure 3 Plato, one of the most distinctive craters on the Moon. It is 109 km in diameter, and very regular. Its very dark grey floor makes it easy to locate whenever it is sunlit.

common, and so are crater-pairs, sometimes separated and sometimes joined. The brightest crater is the 40 km Aristarchus, which has even been mistaken for a volcano in eruption; the huge walled plain, Grimaldi, near the western limb, has the darkest floor.

The system of naming craters after people (usually, though not always, astronomers) was introduced by the Italian Jesuit Riccioli, who drew a lunar map in 1651. The system has been extended since, and is universally accepted. A selected list of some prominent craters is given in Table 4. One crater, the 84 km Wargentin, is filled with lava, so that it is in effect a large plateau.

Other features of the surface include wrinkle-ridges, crossing the maria; valleys, such as the magnificent valley which cuts through the Alps near Plato; low swellings or domes, with gentle slopes and often with summit pits, together with isolated peaks and the crack-like features known as rills, rilles, or clefts. Of special interest is the Straight Wall, in the Mare Nubium. In fact, it is not a wall, but a fault, appearing dark before full moon, because it casts a shadow, and bright after full moon, when its inclined face is illuminated.

Lunar Rocks

All the rocks brought back for analysis are breccias of igneous rocks; the Apollo astronauts brought back 2196 samples (Figure 4) (*see* **Analytical Methods:** Geochronological Techniques). Sedimentary and metamorphic rocks were absent. In the lavas, basalts are dominant; the youngest has been given a radiometric age of 3.08 thousand million years, while the oldest dates back 3.85 thousand million years. The basalts contain more titanium than terrestrial lavas – over 10% in the Apollo 11 samples – and there are small amounts of metallic iron. Many lunar rocks are also comparatively low in sodium and potassium, but one particular type of basalt is rich in potassium (chemical symbol K), the Rare Earth elements, and phosphorus (P), so that it is known as KREEP. Anorthosite – rock containing the minerals plagioclase, pyroxene, and/or olivine in various proportions – is plentiful; one specimen collected by the Apollo 15 astronauts is radiometrically dated as 4 thousand million years old, and is white. It is known as the Genesis Rock! A sample collected from Apollo 12 is about the size of a lemon; it consists largely of SiO_2, and is rich in uranium, potassium, and thorium, making it exceptionally radioactive. It is composed of a dark grey breccia, a light grey breccia, and a vein of solidified lava (Figure 5).

It is now known that some meteorites found on Earth have come from the Moon. Most are breccias, of the same type as mare basalts the anarthosites of the highlands or the regolith. Such lunar-sourced meteorites have been found in Antartica, Australia, North Africa and Oman (*see* **Solar System:** Meteorites). The subsurface rocks of the Moon have so far proved elusive to manned exploration, but lunar meteorite Dhofar 310 recovered from the Oman desert is reported by S.I. Demidova and others in 2003, to be a polymict breccia with deep-seated lunar crustal

Table 4 Some important craters

Name	*Latitude*	*Longitude*	*Diameter, km*	
Albategnius	11.7 S	4.3 E	114	Adjoins Hipparchus
Alphonsus	13.7 S	3.2 W	108	Ptolemæus chain
Anaxagoras	73.4 N	10.1 W	50	Ray centre
Archimedes	29.7 N	4.0 W	82	On Imbrium; trio with Aristillus, Autolycus
Aristarohus	23.7 N	47.4 W	40	Brilliant walls and central peak
Aristillus	33.9 N	1.2 E	55	Archimedes group
Aristotle	50.2 N	17.4 E	87	Pair with Eudoxus
Arzachel	18.7 S	1.9 W	96	Archimedes group
Autolycus	30.7 N	1.5 E	39	Archimedes group
Bailly	66.5 S	69.1 W	287	Field of ruins, near S. limb
Bessel	21.8 N	17.9 E	15	Bright; on Serenitatis
Bullialdus	20.7 S	22.2 W	60	On Nubium; massive walls, central peak
Catharina	18.1 S	23.4 E	104	Theophius group
Clavius	58.8 S	14.1 W	245	Southern highlands
Copernicus	9.7 N	20.1 W	107	Great ray crater
Cyrillus	13.2 S	24.0 E	98	Theophius group
Democritus	62.3 N	35.0 E	39	Highlands N of Mare Frigoris
Dionysius	2.8 N	17.3 E	18	Brilliant crater on edge of Tranquillitatis
Doppelmeyer	28.5 S	41.4 W	63	Bay leading off Humorum
Encke	4.7 N	36.6 W	28	On Procellarum; Kepler area
Endymion	53.9 N	57.0 E	123	Near Humboldtianum; darkish floor
Eratosthenes	14.5 N	11.3 W	58	End of Apennines
Fra Mauro	6.1 S	17.0 W	101	On Nubium, group with Bonpland, Parry
Fracastorius	21.5 S	33.2 E	112	Great bay off Nectaris
Gassendi	17.6 S	40.1 W	101	Edge of Humorum
Grimaldi	5.5 S	68.1 W	172	W of Procellarum, very dark floor
Hercules	46.7 N	39.1 E	69	Pair with Atlas
Hevel	2.2 N	67.6 W	115	Grimaldi chain
Hipparchus	5.1 S	5.2 E	138	Pair with Albategnius
Hyginus	7.8 N	6.3 E	9	Depression in Vaporum; great crater rill
Janssen	45.4 S	40.3 E	199	Southern Uplands, rim broken by Fabricius
Kepler	8.1 N	38.0 W	31	In Procellarum; ray centre
Langrenus	8.9 S	61.1 E	127	Patavius chain
Longomontanus	49.6 S	21.8 W	157	Clavius area
Macrobius	21.3 N	46.0 E	64	Crisium area
Maginus	50.5 S	6.3 W	194	Clavius area
Maurolycus	42.0 S	14.0 E	114	Stofler group
Menelaus	16.3 N	16.0 E	26	Edge of Serenitatis; brilliant
Moretus	70.6 S	5.8 W	111	Southern uplands
Newton	76.7 S	16.9 W	78	Moretus area
Olbers	7.4 N	75.9 W	74	Grimaldi area, ray centre
Petavius	25.1 S	60.4 E	188	Langrenus chain
Phocylides	52.7 S	57.0 W	121	Schickard area
Piccolomini	29.7 S	12.7 E	47	End of Altai Scarp
Pitatus	29.9 S	13.5 W	108	Sedge of Nubium; passes connected with Hesiodus
Plinius	15.4 N	23.7 E	43	Between Serenitatis and Tranquillitatia
Posidonius	31.8 N	29.9 E	95	Edge of Senenitatis
Ptolemæus	9.3 S	1.9 W	464	Trio with Alphonsus and Arzachel
Purbach	25.5 S	2.3 W	115	Walter group
Pythagoras	63.5 N	63.0 W	142	NW of Iridum
Riccioli	3.3 S	74.6 W	139	Adjoins Grimaldi; dark patches on floor
Rømer	25.4 N	36.4 E	39	Taurus area
Schickard	44.3 S	55.3 W	206	Great walled plain
Schiller	51.9 S	39.0 W	180 × 97	Schickard area; fusion of two rings
Stadius	10.5 N	13.7 W	60	'Ghost ring' near Copernicus
Stevinus	32.5 S	54.2 E	74	Petavius area; pair with Snellius
Taruntius	5.6 N	46.5 E	56	On Fœcunditatis
Thales	61.8 N	50.3 E	31	Near Strabo; ray centre
Theophilus	11.4 S	26.4 E	110	Trio with Gyrillus and Catharina
Triesnecker	4.2 N	3.6 E	26	Vaporum area; great rill system

Continued

Table 4 Continued

Name	*Latitude*	*Longitude*	*Diameter, km*	
Tycho	43.4 S	11.1 W	102	Southern highlands; brightest ray centre
Vendelinus	16.4 S	61.6 E	131	Petavius chain
Walter	33.1 S	1.0 E	128	Trio with Regiomontanus and Purbach
Wargentin	49.6 S	60.2 W	84	Schickard area; the famous plateau
Zucchius	61.4 S	50.3 W	64	Schiller area; pair with Segner

Figure 4 Astronaut Schmitt, of Apollo 17, standing by a huge boulder. Schmitt was a professional geologist who had been trained as an astronaut specially for the mission; December 1972, He and his companion, Eugene Cernan, are (so far) the last men to have been to the Moon. Courtesy of NASA.

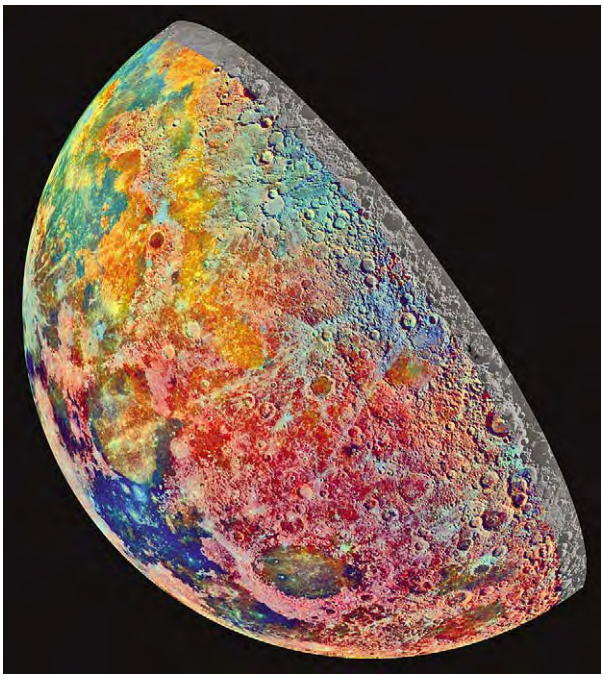


Figure 5 The minerals of the Moon. This is a mosaic of 53 images, obtained in 1992 by the Galileo space craft. The exaggerated false colour shows the differences in surface structure. Blue to orange indicate volcanic lava showing the dark blue Mare Tranquillitatis (lower left) is rich in titanium. Near the bottom of the image, right at Mare Tranquillitatis, is Mare Crisium, pink colours indicating material of the lunar highlands.

material within it as clasts – granulites and igneous rocks of anorthosite, gabbronorite and troctolite composition, also minor dunite and pyroxenite. A unique Al-rich orthopyroxenite/Al-spinel clast is compatible with pressure at a depth of ~20 km within the lunar crust at its source.

Origin of the Craters

It was long believed that the craters were of volcanic origin, similar in type to terrestrial calderæ, but it is now widely accepted that the vast majority are of impact origin, and we have at least a reasonable idea of the sequence of events.

When the Moon came into existence as a separate body, the heat generated melted the outer layers, and for a time there must have been a magma ocean many kilometres deep. Eventually, a crust was formed, thicker on the far side of the Moon than on the Earth-facing side; by cosmic standards it did not take long for the axial rotation to become synchronous. At that stage there was a vast amount of debris moving round the Sun, and the newly-formed planets and satellites swept it up. Between 4400 and 3900 million years ago came the Great Bombardment, when meteorites rained down on the Moon to produce the first major basins such as the Mare Tranquillitatis. Then, between 3900 and 3800 million years ago, came the tremendous impact which resulted in the Imbrian basin and affected the whole of the Moon. As the Great Bombardment ceased there was widespread vulcanism, with magma pouring out from below the crust and flooding the basins to produce structures such as the Mare Orientale. The lava flows ended rather suddenly; as the outpouring slackened, many craters were left undamaged, so that the youngest, such as Tycho and Copernicus, are unflooded. On the far hemisphere, with its thicker crust, there was less flooding, which explains the absence of major Maria and the presence of large,

light-floored walled plains known as palimpsests. Since then, the Moon has seen little activity, though it has been claimed that Copernicus is no more than a thousand million years old and Tycho even younger. One thing is certain: the dinosaurs would have seen the Moon looking very much as it does today!

The lunar cratered surface is remarkably like that of the planet Mercury (*see* **Solar System:** Mercury).

Transient Lunar Phenomena (TLP)

Many observations have been made of elusive glows and obscurations in parts of the Moon, notably in and near the brilliant crater Aristarchus. The reality of these events was demonstrated in 1992, when A. Dollfus, using the 83 cm refractor at the Meddon Observatory (Paris) saw and photographed moving glows inside the large walled plain Langrenus (Figure 6). TLP are almost certainly due to dust lifted above the surface under the effect of gas escaping from below the crust.

Ice on the Moon?

Some of the polar craters are deep by lunar standards, and their floors are always in shadow, so that they remain very cold indeed. The temperatures may be as low as −230°C. In 1966, NASA made the unexpected announcement that ice had been found in the bottoms of these deep polar formations.

The results came from an unmanned probe, Clementine, which had been orbiting the Moon since 1994. It carried a neutron spectrometer, reported effects which indicated the presence of hydrogen, which could combine with oxygen to produce water. Extravagant claims were made, and one NASA scientist went so far as to comment that there are a bunch of craters filled with water ice could a significant resource that would allow a modest amount of colonization for many years. Water would now be mined directly on the Moon instead of having to be shipped from Earth. Yet how could the ice have got there? Rock samples had shown no sign of hydrated material, and the impact of an icy comet would have generated a great deal of heat. Similar results were announced from the next probe, Prospector, launched into lunar orbit in 1998, but many authorities were sceptical. Finally, on 31 July 1999, a test was carried out. Prospector had come to the end of its career, and was deliberately crashed into a polar crater to see whether any signs of water could be found. Predictably, the results were negative. Further tests of the same kind are being planned, but it must be said that the idea of lunar ice seems decidedly far-fetched.

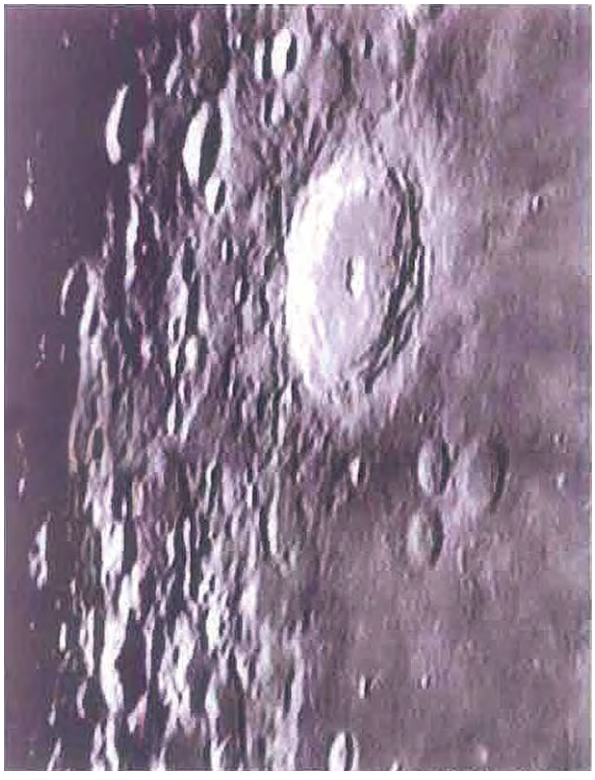


Figure 6 The lunar crater in Langrenus, diameter 127 km, with high walls and a central peak. It is one member of a large chain of craters, including Petavius and Vendelinus.

Life on the Moon?

There seems no chance that life has ever appeared on the Moon; it has been sterile throughout its long history. The crews of the first two Apollos, 11 and 12, were quarantined on return from the Moon to make sure that they had brought back nothing harmful, but quarantining was then abandoned as being unnecessary.

Occultations

As the Moon moves, it may pass in front of a star and hide or occult it. The star shines steadily until it is covered. When it snaps out abruptly; this was one of the early proofs that the Moon has practically no atmosphere.

Eclipses

Eclipses of the Moon are caused by the Moon's entry into the cone of shadow cast by the Earth. The supply of direct sunlight is cut off, but in general the Moon does not disappear, because some sunlight is refracted on to it by the layer of atmosphere surrounding the Earth; the Moon becomes dim and often coppery in colour.

Table 5 Lunar eclipses, 2004 2020

Date	Type	Time of mid eclipse, GMT	Duration of totality minutes	Percentage of Moon eclipsed
2004 May 4	Total	20.32	38	100
2004 Oct 28	Total	03.05	40	100
2005 Oct 17	Partial	12.04		6
2006 Sept 7	Partial	18.52		18
2007 Mar 3	Total	23.22	37	100
2008 Feb 21	Total	03.27	24	100
2008 Aug 16	Partial	21.11		81

Table 6 The Danjon scale

0	Very dark; Moon almost invisible at totality
1	Dark grey or brownish, details barely identifiable
2	Dark or rusty red, with a dark patch in the middle of the shadow: brighter edges
3	Brick red, sometimes a bright or yellowish border to the shadow
4	Coppery or orange red; very bright, with a bluish cast and varied hues

Lunar eclipses may be either total or partial. Because the Sun is a disk, not a point source of light, there is an area to either side of the main shadow cone (the umbra), through which the Moon has to pass; this region is termed the penumbra, and produces only a slight dimming of the surface. Some eclipses are penumbral only. The last eclipses were those of 16 May and 9 November 2003; a list of umbral eclipses up to 2008 is given in Table 5.

No two eclipses are alike; everything depends upon conditions in the Earth's upper air through which the sunlight has to pass. If there is an unusual amount of dust, following an event such as a volcanic eruption, the eclipses is liable to be dark. Observers use a scale given by A. Danjon (Table 6).

Lunar eclipses do not happen at every new moon, because of the inclination of the lunar orbit. Astronomically, they are not important, but they are certainly beautiful to watch, some displaying a wonderful range of hues.

See Also

Analytical Methods: Geochronological Techniques. **Earth Structure and Origins**. **Solar System:** Meteorites; Mercury; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus. **Tectonics:** Earthquakes. **Tektites**.

Further Reading

Harland B (1977) *Exploring the Moon*. Springer/Praxis.
Lindsay H (2001) *Tracking Apollo to the Moon*. New York, Berlin: Heidelberg, Springer Verlag.
Massey H, *et al.* (eds.) (1997) *The Moon: A New Appraisal*. London: The Royal Society.
Moore P (2002) *Patrick Moore on the Moon*. London: Cassell.
Schultz P (1976) *Moon Morphology*. University of Texas Press.
Shepard A and Slayton D (1994) *Moon Shot*. Turner Pub lishing Inc.
Taylor S (1975) *Lunar Science*. Oxford: Pergaman Press.
Wilhelus E (1999) *To a Rocky Moon*. Arizona: University of Arizona Press.
Wood CA (2003) *The Modern Moon*. Sky Publishing Corp.
Westfall JE (2000) *Atlas of the Lunar Terminator*. Cambridge: Cambridge University Press.
Rukl A (2001) *Atlas of the Moon*. London: Hamlyn.

Mars

M R Walter, A J Brown, and S A Chamberlain, Macquarie University, Sydney, NSW, Australia

Introduction

Mars, the fourth planet from the Sun, is Earth's second closest planetary neighbour. As a nearby terrestrial planet with an atmosphere, it shares some geological processes with Earth. However, our impressions of Mars are most clearly drawn into perspective by considering the differences between the geology of Earth and the geology of Mars.

Mars has an elliptical orbit, ranging from 207 to 249 million km from the Sun. Earth is much closer to the Sun (147 to 152 million km) and therefore is warmer. It takes 687 Earth days for Mars to orbit the Sun (670 Martian days, or 'sols'). A day on Mars lasts 24 h and 37 min. The diameter of the planet is 6780 km, about half that of Earth. Its surface area is about the same as that of the land area on Earth. At present, its axis is inclined at 25° to the ecliptic (the plane of rotation around the Sun), much like Earth's, and so it has similar seasons. Compared to the northern hemisphere, southern-hemisphere winters on Mars are more intense; springs and

summers are shorter but, because they occur when the planet is closer to the Sun, peak temperatures are as much as 30°C higher than northern-hemisphere spring and summer temperatures are. This is demonstrated by the fact that the southern polar cap has been known to disappear entirely during southern Summer, though the northern polar cap has never done so. Southern hemisphere winters on Mars are longer and colder than those of the northern hemisphere since Mars is further from the Sun during this period.

Interior of Mars

Some of the physical characteristics of Mars are listed in Table 1. An interpretation of the interior of Mars is presented at Figure 1. The interior structure of Earth is well constrained due to our ability to measure seismic reflections throughout the mantle and into Earth's core. At present, without similar seismic information on Mars, it is possible only to postulate what the Martian interior may be like – for example, does it have a solid core, or a partly liquid core like Earth has? The current view is that Mars has a liquid outer and a solid inner core with a radius of approximately 1700 km. The resultant reduced mantle depth compared to Earth implies that the Martian interior is likely to have cooled substantially faster than Earth did, and convection of the mantle, if it ever occured, may now have slowed or ceased. This implies that volcanism on the Martian surface is less likely today, and may provide an explanation for the current lack of a strong magnetic field on Mars.

The Mars Global Surveyor orbiter did find enigmatic strong local magnetism in rocks of the southern highlands, but an extremely weak overall global magnetic field on Mars. The absence of crustal magnetism near large impact basins such as Hellas and Argyre implies cessation of internal dynamo action during the Early Noachian epoch (similar to 4 billion years ago). Although massive tectonic events most probably caused the formation of Valles Marineris, plate tectonics as is understood on Earth is unlikely to have played a major role in Martian geology. This implies that granites, which are dependent on recycling of the crust in the presence of water, are unlikely to form on Mars. To date, this has been borne out by spectroscopic investigations of the planet.

Table 1 Physical characteristics of Earth and Mars

Characteristic	*Mars*	*Earth*
Orbit (million km)	207 249	147 152
Year (in Earth days)	687	356.25
Day (hours)	24.6	23.9
Mean radius (km)	3390	6371
Core radius (km)	~1700	3485
Average mantle depth (km)	1690	2886
Present obliquity to orbit (degrees)	25.19	23.45
Surface temperature variations (degrees C)	100 to +17	82 to +54

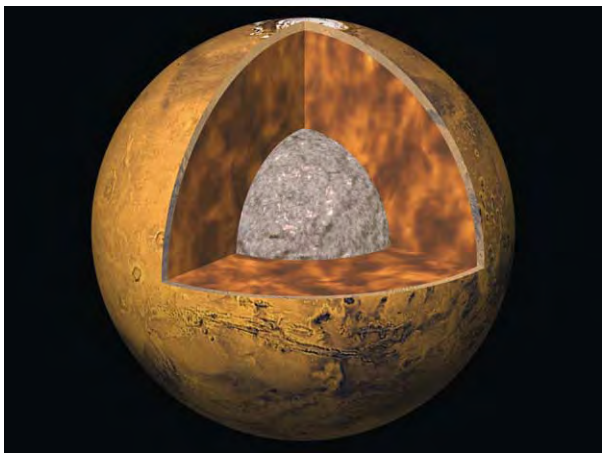


Figure 1 An interpretation of the interior of Mars, showing the core, mantle, and crust. Image courtesy of Calvin J. Hamilton.

Martian Atmosphere and Aeolian Processes

The atmosphere is 95% carbon dioxide, 2.7% nitrogen, 1.6% argon, and 0.13% oxygen, plus minute traces of other gases. In contrast, Earth's atmosphere is 78.1% nitrogen, 20.9% oxygen, 0.93% argon, and 0.03% carbon dioxide. Mars' atmosphere is thin, with the pressure at the surface of the planet (5.6 mbar) being less than one-hundredth of that on Earth. Despite the thin atmosphere of Mars, aeolian (wind-driven) processes have played a large role in shaping many surface features on Mars. These include dunes, yardangs, and etched and eroded terrains. Active dust devils have even been observed by the Mars Orbiter Camera on Mars Global Surveyor (Figure 2). Due to the low gravity of Mars, these features have greater height than their counterparts on Earth have. Dust, blown by the wind over the entire planet, tends to blanket rock features with a homogeneous layer that hampers identification of rock outcrops by orbital and telescopic spectroscopic methods. Dust storms, which periodically turn the Martian atmosphere into an opaque red layer, appear to be coupled with the heating of the atmosphere

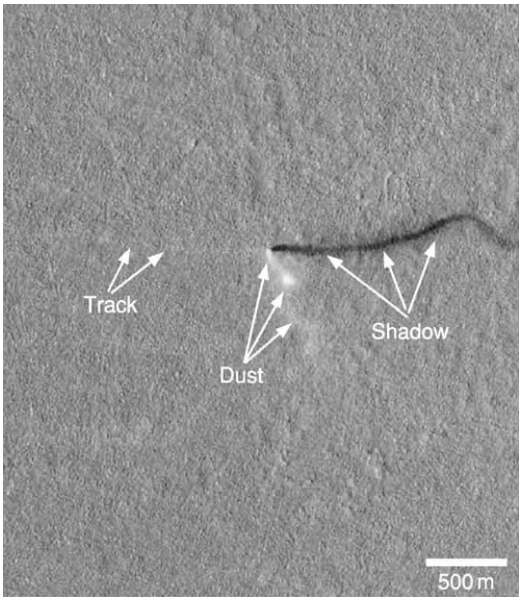


Figure 2 An image of a dust devil taken by the Mars Orbiter Camera. NASA/JPL/Malin Space Science Systems.

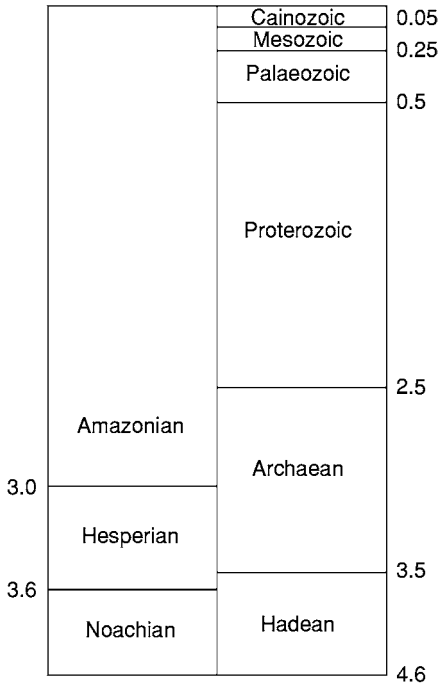


Figure 3 Geological time scales on Mars (left) and Earth (in billions of years).

by the Sun as Mars approaches perihelion (i.e., the closest part of its elliptical orbit).

Obliquity and Climate Variations

It has been calculated that over the past 10 million years, the angle of the spin axis of Mars (the 'obliquity') to the ecliptic, the plane of rotation around the Sun, has ranged from 13° to 4°. The obliquity varies chaotically, on a time-scale of hundreds of thousands to millions of years. In contrast, the obliquity of Earth is stabilized by the presence of the Moon. When the obliquity of Mars is at a minimum, the poles would have permanent caps of frozen carbon dioxide, because, as on Earth, little of the Sun's warmth would reach the poles; when the obliquity of Mars is at a maximum, the polar caps would melt in summer. At times of high obliquity, the water and carbon dioxide stored at the poles would vaporize and be released into the atmosphere, possibly raising the pressure to high enough levels to make liquid water stable for short times.

Cratering Record

The lack of a thick atmosphere and active fluvial processes relative to Earth gives Martian impact craters great preservation potential. Assuming no preference for the location of impact sites, observations of the distributions of craters on the Martian surface can be used to date the surface units. The great Martian impact basins of Hellas and Argyre are thought to have formed early in the history of the Solar System. By correlating the crater distributions with those seen on the Moon (assuming Mars and the Moon have similar impact histories), it is possible to assign ages to regions of the Martian crust, providing a rough guide to when areas of Mars were last resurfaced. Figure 3 relates the Martian geological time-scale to the major geological time periods of Earth. All values are approximate and are given in billions of years. It should be noted that since the current dating of the Martian surface depends on interpretation of impact craters, and these may be incorrect for several reasons, chief amongst them the possibility of one impact causing multiple craters, and also that volcano calderas may be mistaken for craters.

Global Hemispheric Dichotomy and Crustal Thickness

The analysis of images returned by the Viking orbiters of the 1970s revealed a hemispheric dichotomy on Mars. When these results were coupled with those of the Mars Global Surveyor laser altimeter instrument in the late 1990s, a picture of two halves of Mars emerged. The young and smooth northern lowlands have developed separately from the old, cratered southern highlands. The origin of this

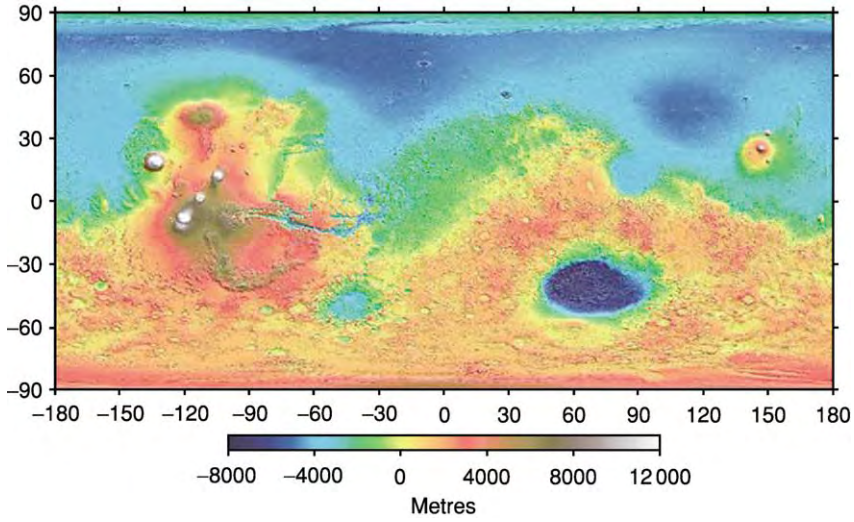


Figure 4 The elevation map of the topography of the Martian surface as determined by the Mars Orbital Laser Altimeter. Image courtesy of NASA.

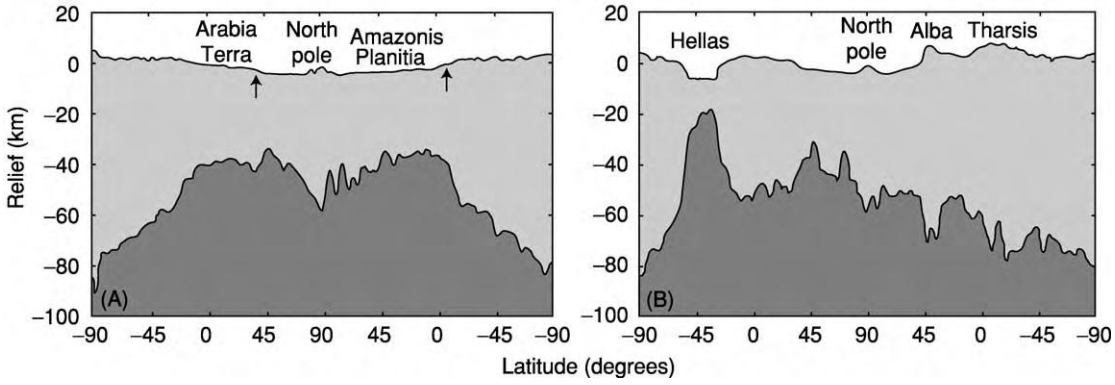


Figure 5 The estimated crustal depth on Mars. (A) Longitude line from 0° to 180°; areas indicate the regions of hemispheric dichotomy: note no major crustal depth variations that might indicate a collision origin of the dichotomy. (B) Longitude line from 70° to 250°. Reprinted with permission from Zuber MT, *et al.* (2000) *Science* 287: 1788 1793. Copyright 2004.

dichotomy is still the subject of debate. Suggestions have included the former presence of a shallow ocean in the northern hemisphere, or low-viscosity lava flows recently covering the northern plains.

An impact origin for the global dichotomy appears to have been ruled out by topography and gravity data obtained by Mars Global Surveyor. The data show that the crust thickness variations are fairly smooth across the dichotomy boundary, as shown in Figure 4. The crustal thickness shows a minimum crustal average depth of 50 km (Figure 5). This compares with approximate Earth values of 30 km for oceanic crust and 80 km for continental crust. The evidence from the Shergottite meteorites, as well as remote sensing from spacecraft and Earth, suggest that the upper crust of Mars is of volcanic origin and mainly of basaltic or andesitic origin, but there is very little in-depth information about the composition.

Mineralogy and Petrology

Thick, homogeneous layers of dust on the Martian surface have hampered efforts to map the surface mineralogy of Mars. The majority of orbital mapping has been conducted by the National Aeronautics and Space Administration (NASA) Viking mission, Mars Global Surveyor, and Mars Odyssey and European Space Agency (ESA) Mars Express missions. This has been complemented by results of landed missions, including Viking 1 and 2, Pathfinder (including the rover Sojourner), and the Mars Exploration rovers

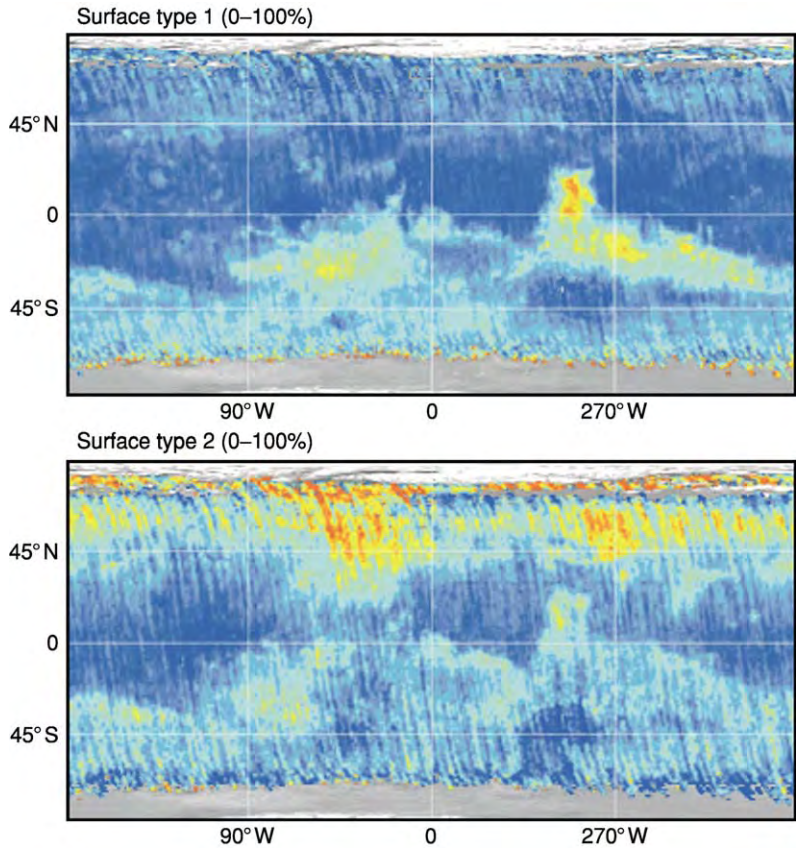


Figure 6 Map of surface types 1 and 2 as determined by thermal emission spectrometry. Red and yellow colours indicate high abundances; blue and light blue indicate low abundances. Type 1 is a basaltic type rock. Type 2 is a more silicic rock that matches the composition of andesite on Earth, though it is probably formed by different processes. Image courtesy of NASA.

(Opportunity and Spirit). Orbital thermal infrared spectroscopy has revealed two different predominant rock types which have been interpreted as the volcanic rocks basalt and andesite. Spatially limited areas of hematite and olivine have been recognized by remote sensing from orbit (Figure 6).

The detection of olivine, which weathers easily in the presence of water, supports the generally held view that water has not existed in its liquid form on the Martian surface for large periods of time. Five of the six possible modes of formation for the detected hematite deposits, however, involve water. A possible resolution to this conundrum is the proposal that frozen glacial deposits or permafrost existed early in the history of Mars, then melted slowly over time to produce water that seeped into the ground and was later heated by a shallow magma, thereby reacting in the subsurface to form the oxidized iron mineral, hematite. Although these analyses effectively represent point measurements on widely dispersed locations on the Martian surface, they give a landed context for the orbital observations. Generally, the Mars Pathfinder analyses are thought to represent the more andesitic basalt member discovered by Mars Global Surveyor, whereas Viking analyses match the more mafic basalt end member.

The results of soil and dust geochemical analyses on Mars by orbital, landed, and telescopic missions have revealed soils relatively high in Fe and S and low in Si and Al. The mineralogy of the soil is not well defined, but is inferred to consist of poorly crystalline weathering products of basalts, analogous to palagonite, which is a hydrous weathering product of basaltic glass. The presence of large amounts of S and Cl in the dust suggests the presence of soluble salts such as sulphates, possibly deposited in the soil as a result of deposition of volcanic aerosols from the atmosphere. To the current time, only small amounts of carbonate have been detected in the homogeneous dust covering Mars. No outcrops of carbonate have been discovered. The lack of carbonates is a strong argument against a long-lived ocean or standing water, which might be expected to deposit carbonate in the presence of the CO_2-rich Martian atmosphere, just as limestone forms in the shallow coastlines on Earth.

Water on Mars

At low latitudes, daily temperatures on Mars range from about −100° to +17 °C, and the average is −60°C. Because of the low pressure exerted by the very thin atmosphere, at these temperatures liquid water is everywhere unstable. Consequently, the water ice at the poles sublimes (goes straight from ice to vapour) into the atmosphere. Down to latitudes of about 40°, ice can exist in the ground as 'permafrost' as shallow as 1 m. Water ice has been detected at the north pole when it is exposed as the overlying carbon dioxide ice ('dry ice') sublimes in summer.

Since the 1970s, when Viking orbiters sent back high-resolution images of the surface of Mars, scientists have puzzled over features that resemble the gullies and water-eroded valleys seen on Earth. The lack of liquid water on the Martian surface today begs the question as to how the gully and valley features formed. Some scientists have postulated that a thicker atmosphere in the past may have led to liquid water being stable on the Martian surface. Probable water-influenced features on Mars have been recognized on three different scales and are commonly grouped into 'outflow channels', 'valley systems', and 'gullies', in order of descending scale.

Giant outflow channels on Mars are several tens to hundreds of kilometres across and many thousands of kilometres in length. They are mostly Hesperian in age and commonly start in chaotic terrain, as seen in the image of Hydaspis Chaos (Figure 7). They are usually associated with streamlined islands and terraces, indicating massive fluid flow, with the probable source being water, although some researchers have suggested CO_2 as a possible fluid.

Valley networks, such as those seen in the Thaumasia region in Figure 8, superficially resemble branching valley fluvial systems here on Earth; however, the lack of fine-scale structure, low drainage densities, and differing morphology argues against a rainfall and subsequent surface runoff origin. Instead, groundwater flow from seeps or hydrothermal systems is a plausible mode of formation. Most of the valley networks are Noachian in age, but there are younger systems, including those developed on Amazonian-age terrain around volcanic centres in the Tharsis region.

Martian gully systems often develop in the rims of impact craters, and have fine-scale features that suggest they are very young. They do not occur near the equator, appearing only at latitudes poleward of 30° in both hemispheres. A possible mode of formation is by melting of water beneath snow or ice packs. The lack of gullies around the equator may indicate that

Figure 7 (A) Hydaspis Chaos, the source of Tius Valles, as imaged by Viking, showing a typical giant outflow channel. Note streamlined island features. The image is approximately 200 km across (north is up). (B) Evidence of erosion by a liquid (water?) in the Athabasca part of the Marte Vallis channel and a stream lined island system, as imaged by the Mars Global Surveyor M21 01914. (B) Reproduced with permission from Hartmann WK (2003) *A Travellers Guide to Mars*. New York: Workman Publishing. NASA/JPL/Malin Space Science Systems.

formation of ice in impact craters was insufficient to produce these deposits.

The gamma ray spectrometer (GRS) on the orbiting spacecraft Mars Odyssey has been able to map hydrogen abundance in the top metre of the surface of Mars. The most likely source of the hydrogen in the Martian crust is molecular water or hydroxyl (OH^-) ions in weathered rocks. The distribution of hydrogen in the Martian surface is shown in Figure 9.

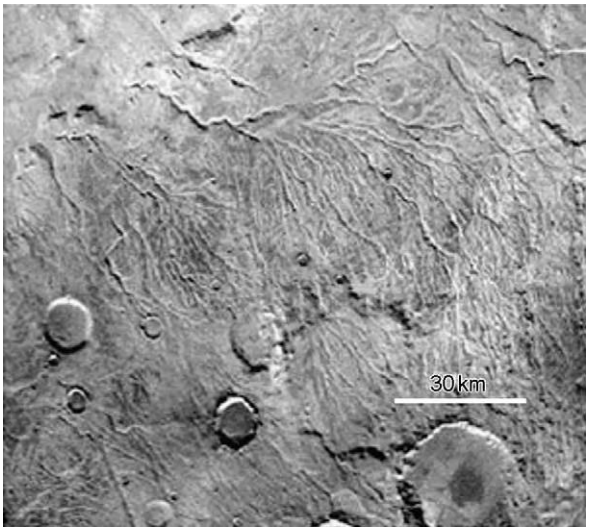


Figure 8 Martian valley networks in the Thaumasia region, as observed by Viking. Image is 100 km across. Image courtesy of NASA.

Large-Scale Features

Due to the smaller size of Mars in comparison to Earth, and the apparent lack of plate tectonics in the past 4 billion years, some topographic features have taken on exaggerated forms. The Tharsis bulge is a large region of volcanism and deformation crossing the equator. The formation process of the Tharsis uplift is still under discussion; it may be due to convection in the interior of the planet or to there being a region of intense volcanism. The Tharsis complex consists of two broad rises, the largest southern rise containing three of the largest volcanoes on Mars, Ascraeus Mons, Pavonis Mons, and Arsia Mons. The smaller northern rise is dominated by the volcanic construct Alba Patera.

Olympus Mons (Figure 10) is situated to the west of the Tharsis region and though not surficially connected to the region, it is thought to be connected in origin. Olympus Mons is the largest known volcano in the Solar System. Like some of Earth's volcanoes, it

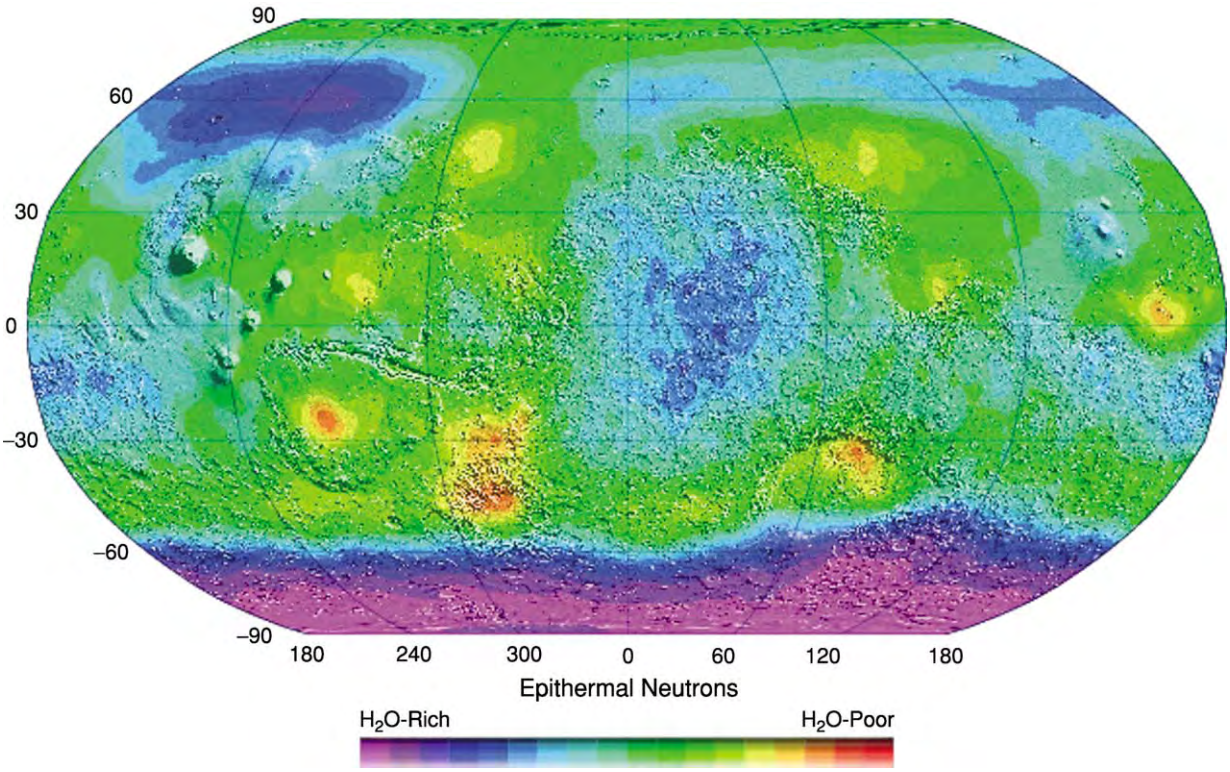


Figure 9 Map of hydrogen (or epithermal neutrons) in the shallow Martian crust (late southern summer), as detected by the gamma ray spectroscope on Mars Odyssey. Image courtesy of Dr William V. Boynton, University of Arizona.

Figure 10 Image of Olympus Mons, as observed by the NASA Viking Orbiters. Image courtesy of NASA.

is thought to have been an area of the crust situated above a region of advecting hot mantle material, known as a hotspot. Unlike Earth analogues, Olympus Mons continued to build, because at the time of formation Mars lacked the plate tectonics to move the volcano away from the source. At 21 183 m above the global reference datum, Olympus Mons is 2.5 times higher than Earth's largest shield volcano, Mauna Loa, and 100 times its volume.

Valles Marineris, named for the spacecraft that first discovered it (Mariner 9), is a giant canyon system that extends eastward some 4500 km from the central regions of the Tharsis complex. The troughs are generally about 50 to 100 km wide and the depth of the floor drops down to 5310 m below the global reference datum (six to seven times as deep as the Grand Canyon). This feature is thought to have formed as the uplift of Tharsis created tension in the cooled crust.

Martian Polar Regions

The polar regions are broadly composed of four geological units: the basal plains, the polar layered deposits, the residual ice, and the seasonal frost. The basal plains differ north from south, In the south, they are believed to date from the Noachian Period. There is evidence of intense cratering, contractional deformation, resurfacing of low areas, and local dissection producing valley networks. There is no evidence for polar deposits, ice sheets, or glaciation at these times. During the Hesperian, these deposits were altered by waning impacts and volcanism. In the north, the basal plains are much younger, being buried by a water or debris ocean during the Hesperian Period.

The polar layered deposits began building up during the late Hesperian or later and appear to be similar at both poles. Polar layered deposits extend further than do the icecaps that overlie them (to a greater and more asymmetric extent in the south). The deposits are composed of dust, water ice, and other volatiles such as CO_2, in varying ratios for each layer. Each of these layers, shown in Figure 11, was deposited under differing environmental conditions.

Due to the small number of superimposed craters at each polar icecap, the caps have been interpreted to be of late Amazonian age, with the southern cap being slightly older than the northern. This means that the Martian environment has only recently allowed the current polar icecaps to form. This is likely to be due to the changing obliquity (0°–60°) of the planet, causing the icecaps to reform continuously. There is evidence of advance and retreat at both poles, but exact timing remains unknown. The residual ice in the north appears to be mainly H_2O, whereas the south polar surface residual ice appears to be CO_2. Little is known of the bulk composition of residual ices at either pole, because thin overlying layers of CO_2 ice can mask signatures of the composition beneath. The seasonal frost is composed of CO_2

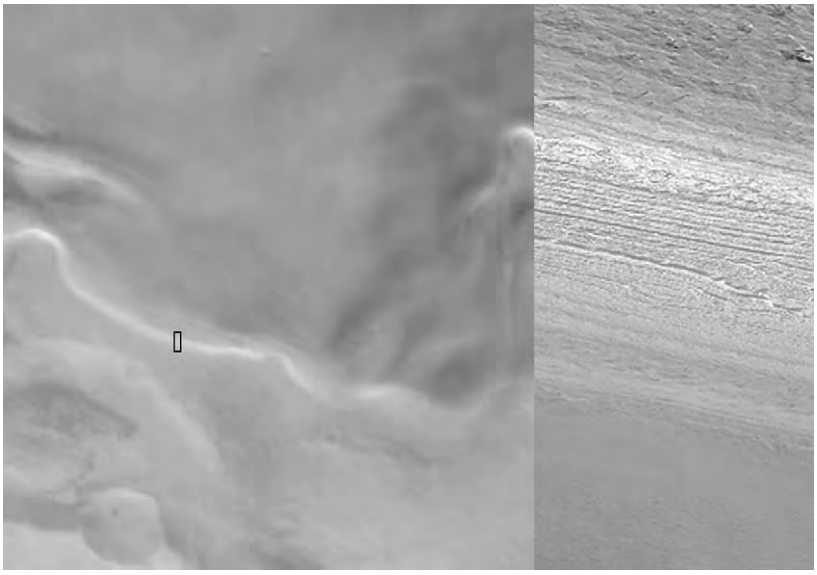


Figure 11 Martian layered polar terrain, as imaged by the Mars Orbiter Camera. Image on the right is 750 m across and sits within the small area outlined in the context image on the left. NASA/JPL/Malin Space Science Systems.

ice that condenses from the atmosphere during winter and sublimes again in the Martian summer.

Satellites of Mars

Mars has two small satellites, Phobos (27 km maximum diameter) and Deimos (25 km maximum diameter), both irregular potato-shaped bodies. Both have cratered surfaces and Deimos is also strongly grooved. Their origin is obscure – they may be captured asteroids. The orbit of Phobos is unstable and it will eventually crash into Mars.

Shergottite–Nakhlite–Chassigny Meteorites

Meteorites, widely accepted as being from Mars have been collected here on Earth. They are collectively called SNC meteorites after the discovery locations of the first three meteorites – Shergotty, Nakhla and Chassigny. They are believed to come from Mars because: 1.) their young age of formation relative to the Solar System, suggesting they were formed on a rocky planet, and 2.) trapped gases inside solidified molten rock produced by high shock pressures in some of the meteorites display chemical and isotopic characteristics matching those of the Martian *atmosphere* (as reported by the Viking landers). The trapped gases are interpreted to represent samples of the atmosphere when these rocks were ejected from Mars by impacts. The Martian meteorites have contributed greatly to knowledge about Mars, but naturally have also raised many scientific questions. The inability to pinpoint a source location on Mars is the greatest limitation on the utility of the SNC meteorites, but the ability to analyse the rocks with the most modern of terrestrial scientific equipment is a great advantage lacking as we do any rock samples directly recovered from the planet.

At the current time, there are 30 known SNC meteorites. The reported crystallization ages of SNC meteorites range from 154 Ma to 1.3 Ga, with the exception of ALH84001, which has a crystallization age of about 4 Ga. ALH84001 was discovered in Antarctica in 1984 and was held in storage by NASA for some years. In 1996, it was reported that a team of NASA scientists had found possible signs of life in small amounts of carbonate within ALH84001. Six lines of evidence were cited that would be explained best by invoking a biological origin. These lines of evidence have been debated within the scientific community in the intervening years and thus the question of life on Mars still remains open, lacking rock samples directly recovered from the planet.

Mars Exploration Rover Missions

The Mars Exploration Rovers, Spirit and Opportunity, have added to the data from the planet's surface. Designed as mobile remote geologists, they were equipped with various instruments to study the rocks they found, including panoramic cameras, infrared spectrometers, abrasion tools and a microscopic camera. '*Opportunity*' has been successful in

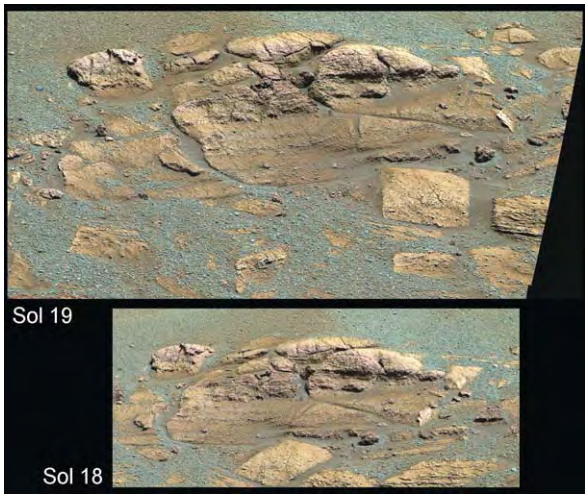


Figure 12 Images of the rock 'El Capitan', the first bedrock to be investigated by a Mars landed rover, on two days (sols). The outcrop is approximately 10 cm high. Image courtesy of NASA.

studying a number of fascinating outcrops which are only now starting to be interpreted. An image of the first competent rock outcrop found by a rover on Mars is shown in false colours in Figure 12. It reveals layered successions which could be volcanic or sedimentary in nature. Interplanetary investigations like these will continue to reveal the true extent of apparent differences in the geology of Earth and Mars.

See Also

Sedimentary Processes: Glaciers. **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Venus. **Volcanoes**.

Further Reading

Acuna MH, Connerney JEP, Ness NF, *et al.* (1999) Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* 284(5415): 790 793.

Barlow NG (1988) Crater size/frequency distributions and a revised Martian relative chronology. *Icarus* 75: 285 305.

Boyce JM (2002) *The Smithsonian Book of Mars*. Washington, DC and London: Smithsonian Institution Press.

Boynton WV, Feldman WC, Squyres SW, *et al.* (2002) Distribution of hydrogen in the near surface of Mars: Evidence for subsurface ice deposits. *Science* 297(5578): 81 85.

Christensen PR (2003) Formation of recent martian gullies through melting of extensive water rich snow deposits. *Nature* 422(6927): 45 48.

Clifford SM and Parker TJ (2001) The evolution of the Martian hydrosphere: Implications for the fate of a primordial ocean and the current state of the northern plains. *Icarus* 154(1): 40 79.

Gibson EK, McKay DS, Thomas Keprta KL, *et al.* (2001) Life on Mars: evaluation of the evidence within Martian meteorites ALH84001, Nakhla, and Shergotty. *Precambrian Research* 106(1 2): 15 34.

Gulick VC (1998) Magmatic intrusions and a hydrothermal origin for fluvial valleys on Mars. *Journal of Geophysical Research Planets* 103(E8): 19 365 19 387.

Hartmann WK (2003) *A Travellers Guide to Mars*. New York: Workman Publishing.

Kieffer HH, Jakosky BM, Snyder CW, and Matthews MS (1992) *Mars*. Tucson: The University of Arizona Press.

Kolb EJ and Tanaka KL (2001) Geologic history of the polar regions of Mars based on Mars Global Surveyor data: II. Amazonian period. *Icarus* 154: 22 39.

McKay DS, Gibson EK, ThomasKeprta KL, *et al.* (1996) Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science* 273(5277): 924 930.

Nimmo F and Stevenson DJ (2000) Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *Journal of Geophysical Research Planets* 105(E5): 11969 11979.

Sleep NH (1994) Martian plate tectonics. *Journal of Geophysical Research Planets* 99(E3): 5639 5655.

Squyres SW and Kasting JF (1994) Early Mars how warm and how wet. *Science* 265(5173): 744 749.

Tanaka KL and Kolb EJ (2001) Geologic history of the polar regions of Mars based on Mars Global Surveyor Data: I. Noachian and Hesperian periods. *Icarus* 154: 3 21.

Treiman AH, Gleason JD, and Bogard DD (2000) The SNC meteorites are from Mars. *Planetary and Space Science* 48(12 14): 1213 1230.

Zuber MT, Solomon SC, Phillips RJ, *et al.* (2000) Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* 287(5459): 1788 1793.

Jupiter, Saturn and Their Moons

P Moore, Selsey, UK

Introduction

Jupiter and Saturn, the largest and most massive planets in the Solar System, have no visible solid surfaces and are therefore not of real concern to the geologist. However, their satellite systems are of immense interest. Jupiter has four major satellites (the Galileans) which are of planetary size, and which differ markedly from each other geologically; there are also over four dozen small satellites, almost certainly captured asteroids, which seem to be icy. Saturn has one very large satellite (Titan), which has a dense atmosphere obscuring the surface; there are also eight medium-sized icy satellites as well as numerous ex-asteroids. These various bodies are discussed in this article.

Jupiter

Jupiter is the giant of the Sun's system. It is more massive than all the other planets combined, and is generally the brightest object in the sky apart from the Sun, the Moon, and Venus. Physical and orbital data are given in Table 1.

Telescopically, Jupiter shows a yellowish, obviously flattened disk, crossed by dark belts and bright zones (Figure 1). It has always been assumed that the belts are regions of descending gases while the bright zones are regions where gas is rising from the interior, though some recent (2003) observations may indicate that the reverse is true. There are generally two very prominent belts, the North Equatorial and the South Equatorial, with others in higher latitudes. Jupiter does not rotate in the way that a solid body would do. The region between the north edge of the South Equatorial Belt and the south edge of the North Equatorial Belt (System I) has a mean period of 9 h 50 m 30 s, while the period of the rest of the planet (System II) is 9 h 55 m 41 s, but individual features have periods of their own. Radio methods indicate that the interior (System III) has a period of 9 h 55 m 29 s, though this is subject to some uncertainty.

Much of our knowledge of Jupiter has been derived from five space-craft: two Pioneers, two Voyagers, and Galileo. Details are given in Table 2. Useful data were also obtained from two fly-by probes, Ulysses in February 1992 (the solar polar probe) and Cassini in December 2000 (en route for an encounter with Saturn).

Jupiter is assumed to have a hot silicate core at a temperature of at least 20 000°C, probably rather more. The core is surrounded by a thick shell of liquid metallic hydrogen, which is itself surrounded by a shell of liquid molecular hydrogen; above lies the atmosphere, made up chiefly of hydrogen and helium, with hydrogen compounds such as methane and ammonia. Windspeeds in the visible clouds are high, and the surface details are always changing. Of special note is the Great Red Spot, which has been visible for most of the time since regular observations began in the seventeenth century. Once thought to be a glowing volcano, it is now known to be a phenomenon of Jovian meteorology – a high-level anticyclonic vortex, elevated by 8 km above the adjacent cloud deck.

Jupiter's magnetic field is much the strongest in the Solar System. The magnetic axis is inclined to the rotational axis at an angle of 9.6°; the polarity is opposite to that of the Earth's field. The planet is a powerful radio source, and is surrounded by zones of radiation which would quickly be fatal to an astronaut unfortunate enough to enter them. There is a system of dark, obscure rings, probably formed from material released from the small inner satellites by meteoritic impact. They are quite unlike the glorious icy rings of Saturn.

Table 1 Data for Jupiter

Distance from the Sun	Max 815 700 000 km (5.455 astronomical units)
	Mean 778 350 000 km (5.203 a.u.)
	Min 740 900 000 km (4.951 a.u.)
Orbital period	11.86 years (4332.59 days)
Rotation period	System I, 9 h 50 m 30 s
	System II, 9 h 55 m 41 s
	System III, 9 h 55 m 29 s
Axial inclination	3° 4
Orbital inclination	1°18′15″8
Orbital eccentricity	0.048
Diameter	Equatorial 142 884 km
	Polar 133 708 km
Oblateness	0.065
Density, water 1	1.33
Mass, Earth 1	317.89
Volume, Earth 1	1318.7
Escape velocity	60.22 km/sec
Surface gravity Earth 1	2.64
Mean surface temperature	150°C
Albedo	0.43

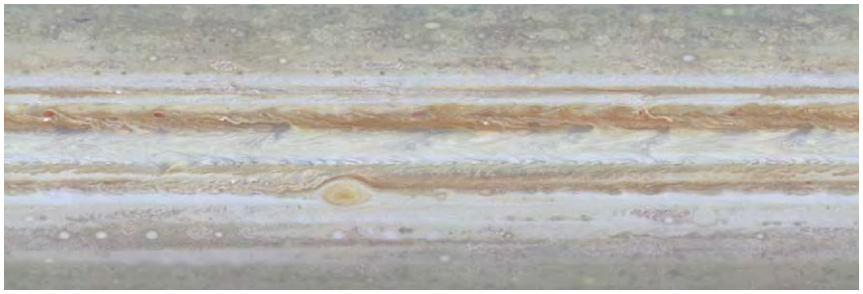


Figure 1 Linear zones of light and dark, rising and descending gases on Jupiter's surface. Reproduced from NASA.

Table 2 Space craft to Jupiter

Name	*Launch date*	*Encounter date*	*Nearest approach, km*	
Pioneer 10	2 Mar. 1972	3 Dec. 1973	131 400	Fly by
Pioneer 11	5 Apr. 1973	2 Dec. 1974	46 400	Fly by. Went on to Saturn
Voyager 1	5 Sept. 1977	5 Mar. 1979	150 000	Images of Jupiter and the Galileans. Went on to Saturn
Voyager 2	20 Aug. 1977	9 July 1979	714 000	Complemented Voyager 1. Went on to Saturn, Uranus, Neptune
Galileo	18 Oct. 1989	7 Dec. 1995	Entry	Orbiter and entry probe

Data were also obtained from Ulysses (1992) and Cassini (2001).

Satellites

The four Galilean satellites – Io, Europa, Ganymede, and Callisto – were observed by Galileo as long ago as 1610 (hence the name of the space-craft). Four small inner satellites were discovered between 1892 and 1979, and there are many small bodies moving round Jupiter beyond the path of the outermost Galilean, Callisto. The total number of known satellites by April 2004 was 62. Data for all the satellites over 8 km in diameter are given in Table 3.

Io Io is violently volcanic. In March 1979, S Peale and his colleagues in America suggested that since Io's orbit is not perfectly circular, the interior might be 'flexed' by the gravitational pulls of Jupiter, and also the other Galileans, sufficiently to produce active surface volcanoes. Only a week later the first volcanic plume was detected on an image from the Voyager 1 space-craft, and many dozens have since been identified, both from the space probes and with the Hubble Space Telescope. Lava-flows and lava lakes are plentiful; the average temperature of the lavas is about 1600°C. Many explosive eruptions are driven by sulphur dioxide gas emission; the surface is remarkably colourful, with yellow, orange, red, and black areas (Figure 2). The surface is 'young' and virtually without surviving impact craters. One volcano, Loki, is the most powerful in the Solar System, emitting more heat than all the Earth's active volcanoes combined. Io is connected to Jupiter by a strong flux tube, and has a marked effect upon the Jovian radio emission; material sent out by the volcanoes is spread along the orbit, producing a torus.

Europa Europa is only slightly smaller than Io, and rather further from Jupiter, but the two satellites are very different. Europa has a smooth, icy surface with very limited vertical relief and few impact craters, though one of these, Pwyll, shows bright rays extending outward and crossing all other features. There are plains, chaotic areas, and low ridges, together with shallow pits. Detailed views from space-craft (particularly Galileo) show what look remarkably like icebergs, and it is widely believed that an ocean of salty water lies below the visible surface, with the icebergs floating around (Figure 3). Fragmented blocks of ice seem to look very like the blocks in the Earth's polar seas during a springtime thaw.

Europa does not have a strong internal magnetic field, but it orbits within Jupiter's magnetosphere, and the instruments on Galileo have detected an induced magnetic field which produces significant effects linked with the rotational period of the planet. Jupiter's magnetic field at Europa changes direction every $5\frac{1}{2}$ h, and this indicates the existence of a layer of electrically conducting material, such as salty water, not far below the icy surface of Europa. It seems that the ocean may lie at a depth of less than 100 km. If it really does exist (and as yet there is no

Table 3 Satellites of Jupiter

Satellite	*Mean distance from Jupiter, km*	*Orbital period, days*	*Diameter, km (equator)*	*Density water 1*	*Orbital eccentricity*	*Orbital inclination*	*Escape velocity km s* ^{1}
Small inner satellites							
Metis	128 100	0.294	60	2.8	0.001	0.021	0.025
Adrastea	128 900	0.300	26	2?	0.002	0.027	0.014
Amaithea	181 100	0.498	262	1.8	0.003	0.389	0.084
Thebe	221 900	0.674	110	1.5	0.018	0.070	0.043
Galileans							
Io	421 800	1.769	3660	3.6	0.004	0.036	2.56
Europa	671 100	3.551	3130	3.0	0.009	0.470	2.02
Ganymede	1 070 400	7.154	5268	1.9	0.002	0.195	2.74
Callisto	1 882 700	16.689	4821	1.1	0.007	0.281	2.45
Outer prograde satellites							
Themisto	7 507 000	130.0	9	2?	0.242	43.08	Low
Leda	11 165 000	240.0	16	2.7	0.164	27.46	0.01
Himalia	11 461 000	250.6	186	2.8	0.162	27.50	0.12
Lysithea	11 717 000	259.2	38	3.1	0.212	28.30	0.02
Elara	11 741 000	259.6	78	3.3	0.217	26.63	0.05
Outer retrograde satellites							
Ananke	21 276 000	610.5	28	2.7	0.244	148.9	0.02
Carme	23 404 000	702.3	48	2.8	0.253	164.9	0.03
Pasiphae	23 624 000	708.0	58	2.9	0.109	151.4	0.03
Sinope	23 939 000	724.5	38	3.4	0.250	158.1	0.24

Ten of the other small retrograde satellites have been named; all are below 8 km in diameter. Harpalyke, Praxidike, Iocaste, Chaldene, Isonoe, Erinome, Tayrete, Kalyke, Megaclite, and Callirrhoe.
? doubtful value.

Figure 2 Io, imaged by the Galileo probe, the colour is very accurate. The volcanoes, such as Pele, are violently active. Reproduced from NASA.

final proof) there will be tidal effects. There have been the inevitable speculations about possible life-forms, but conditions in such a strange, sunless sea would not appear to be inviting!

Figure 3 Europa, imaged from the Galileo space craft in orbit round Jupiter. The icy surface, unlike any other in the Solar System, may cover a salty ocean. Hubble Space Telescope Image, NASA.

Ganymede Ganymede is the largest satellite in the Solar System, and is actually larger than the planet Mercury, though not so massive; the mean density of the globe is less than twice that of water. There is an excessively tenuous atmosphere, and a marked magnetic field. Presumably this indicates the presence of a metallic core surrounded by a mantle composed of a mixture of rock and ice; the surface is icy. There are two types of surface terrain, about equal in area, very ancient, thickly cratered dark regions and somewhat younger light regions, marked with an extensive array of ridges and grooves. Evidently there has been marked tectonic activity in the past. In general, the craters have little vertical relief. The largest indiviual

feature is a dark plain which has been given the appropriate name of Galileo.

Callisto Callisto is rather smaller than Ganymede, and is less dense, so that ice is a major constituent of the globe. The surface is heavily cratered, and there are two huge ringed plains, Valhalla and Asgard. Unexpectedly, it has been found that the local magnetic field fluctuates in the same way as that of Europa, and there may be a similar salty ocean deep inside the globe. Space-craft data indicate that the interior of Callisto is made up of compressed rock and ice, with the percentage of rock increasing with depth.

Amalthea Amalthea the only other Jovian satellite over 200 km in diameter, was discovered in 1892 by EE Barnard, using the 91 cm refractor at the Lick Observatory (this was the last visual discovery of a planetary satellite). It is irregular in form; the surface is red, due possibly to contamination from the volcanoes of Io. In November 2002, the Galileo space-craft flew past Amalthea and found that the density is very low; indeed the satellite has been described as an 'ice rubble pile'. The rotation is synchronous, with the longest axis pointing to Jupiter. Images from Galileo show that there are two craters, Gaea and Pan, which are very large relative to the overall diameter of Amalthea; there are ridges, troughs, and two bright patches which seem to be hills. Amalthea is one of four satellites known to move within the orbit of Io; the others are Metis, Adrastea, and Thebe. They are icy in nature, and impact craters have been imaged.

Outer Icy Satellites By April 2004, the total number of known Jovian satellites had risen to 62, but few of these were more than a few kilometres in diameter. Several satellites, including Himalia (diameter 184 km) move at between 11 and 12 million kilometres from Jupiter; these have direct motion, and those that have been imaged show the usual impact craters. Much further out move small satellites, many of which have retrograde motion; the largest, known before the space-probe era, are Ananke, Carme, Pasiphaë, and Sinope. They are so far from Jupiter that their orbits are not even approximately circular, and no two cycles are alike. There seems no doubt that all the outer icy satellites are captured asteroids; their small size means that they are very difficult to record.

Saturn

Saturn the largest and most massive planet in the Solar System, apart from Jupiter, is distinguished by its magnificent icy ring system, making it probably the most beautiful object in the entire sky. Physical and orbital data are given in Table 4.

Saturn's globe shows belts not unlike those of Jupiter, but much less prominent, and sensibly curved. Predictably, the globe is flattened; this is because of the rapid rotation. It is thought that there is a silicate core at a temperature of perhaps 15 000°C, overlaid by layers of metallic hydrogen, molecular hydrogen, and then the atmosphere, which is not unlike that of Jupiter.

There is no surface feature comparable with the Great Red Spot on Jupiter, but prominent, temporary white spots are seen occasionally, as in 1933, 1960, and 1990 (Figure 4). The 1933 spot was discovered on 3 August by WT Hay, using a 15 cm refractor; it remained identifiable until 13 September, and was considered to be of an eruptive nature. The 1990 spot, discovered by S Wilber, was in the same latitude.

Table 4 Data for Saturn

Distance from the Sun, km	Max 1 506 400 (10.069 astronomical unite)
	Mean 1 426 800 (9.359 a.u.)
	Min 1 347 600 (9.008 a.u.)
Orbital period	29.4235 years (10 746.94 days)
Equatorial rotation, period	10 h 13 m 59 s
Axial inclination	26.73°
Orbital eccentricity	0.05555.
Orbital inclination	2°29′21″
Diameter; km	Equatorial 120, 536
	polar 108, 728
Oblateness	0.098
Mass, Earth 1	95.17
Volume, Earth 1	752
Escape velocity	35.26 km/s
Surface gravity, Earth 1	1.19
Mean surface temperature	180°C

Figure 4 Saturn, image from the Hubble Space Telescope, 1 December 1994; showing the bright white spot, it consists of condensed ammonia ice crystals, and had changed little since its discovery in September 1994.

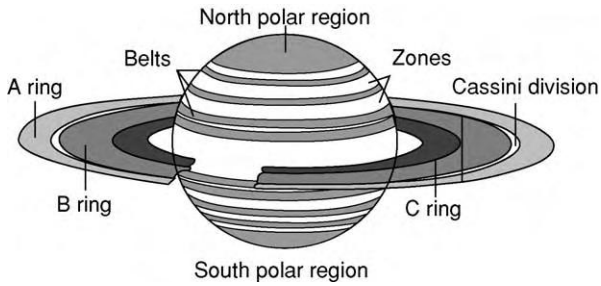


Figure 5 Diagram of Saturn and its ring system. There are some less prominent rings behind the main system.

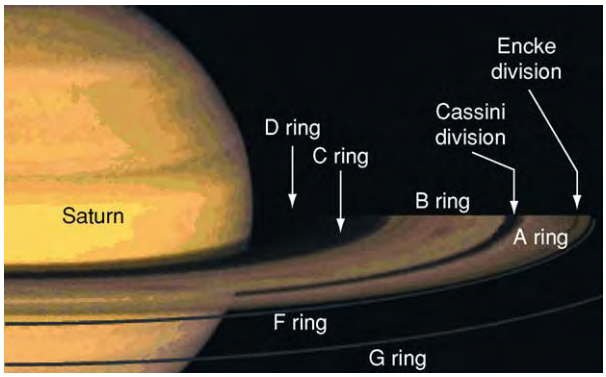


Figure 6 Diagram showing the main ring system. The bright rings are A and B, C is the Crepe or Dusky Ring. D is not well defined. Reproduced from Kennod R (1990) *The Journeys of Voyager NASA reaches for the planets, BDD*, New York.

Table 5 The ring system of Saturn

Feature	*Distance from centre of Saturn, km*
Inner edge of ring D	66 900
Outer edge of ring D	73 150
Inner edge of ring C	74 510
Outer edge of ring C	92 000
Inner edge of ring B	92 000
Outer edge of ring B	117 500
Centre of Cassini division	119 000
Inner edge of ring A	122 200
Centre of Encke division	135 700
Outer edge of ring A	136 800
Centre of ring F	140 210
Centre of ring G	168 000
Inner edge of ring E	180 000
Brightest part of ring E	230 000
Outer edge of ring E	480 000

Within a few days the spot had been spread out by Saturn's strong equatorial winds, and by October had been transformed into a bright zone all round the equator. Extra outbreaks were seen in it, clearly indicating an uprush of material from below. Saturn has a strong magnetic field, though less powerful than that of Jupiter; the rotational axis and the magnetic axis are almost coincident.

There are three main rings (Figure 5); A and B, separated by a gap known as the Cassini Division is honour of its discoverer, and an inner semi-transparent ring, C (the Crepe Ring) (Figure 6). Details of the ring system are given in Table 5. The rings are made up of ice particles, from grains up to several metres across; the brightest ring, B, shows curious 'spokes', presumably particles elevated away from the main ring plane by magnetic or electrostatic forces. The irregular F ring, outside the main system, is stabilized by two small 'shepherd' satellites, Prometheus and Pandora. Though the system is very extensive, with a total diameter of 270 000 km, it cannot be more than 200 m thick, and if all the particles could be combined they would make up an icy satellite less than 300 km across.

Three space-craft have now encountered Saturn; Pioneer 11, and the two Voyagers. Data are given in Table 6. The Cassini/Huygens probe, launched in 1997, was scheduled to reach its target in late 2004.

Satellites

The satellite system of Saturn is very different from that of Jupiter. There is one very large satellite, Titan, and seven icy satellites with diameters between 200 and 1600 km; the rest are much smaller. By May 2003, the total number of known satellites had risen to 31. Data are given in Table 7.

Titan Apart from Ganymede, Titan is the largest, satellite in the Solar System. Its surface is permanently hidden by its dense atmosphere, made up chiefly of nitrogen with appreciable amounts of methane and ethane; organic compounds are plentiful. The atmospheric pressure on the surface is 1.6 times that of the Earth's air at sea level; the surface temperature is −178°C.

Table 6 Space craft to Saturn

Name	*Launch date*	*Encounter date*	*Closest approach, km*	
Pioneer 11	5 Apr. 1973	11 Sept. 1979	20 880	Preliminary results
Voyager 1	5 Sept. 1977	12 Nov. 1980	124 200	Images included Titan
Voyager 2	20 Aug. 1977	25 Aug. 1981	101 300	Went on to Uranus and Neptune
Cassini/Huygens	15 Oct. 1997	2004		Scheduled orbiter and landing on Titan (Huygens)

Table 7 Satellites of Saturn

Name	*Mean distance from Saturn, km*	*Orbital period, d*	*Longest diameter, km*	*Density, water 1*	*Escape velocity, km/s*
Pan	133 583	0.525	19	?	Low
Atlas	137 640	0.602	37	?	Low
Prometheus	139 350	0.613	145	0.27	0.02
Pandora	141 700	0.625	114	0.70	0.28
Epimetheus	151 422	0.694	144	0.63	0.32
Janus	151 172	0.694	196	0.67	0.35
Mimas	185 520	0.942	421	1.17	0.16
Enceladus	238 020	1.370	512	1.24	0.21
Tethys	294 600	1.888	1038	0.98	0.44
Telesto	294 600	1.888	34	?	Low
Calypso	294 600	1.888	30	?	Low
Dione	377 400	2.737	1120	1.49	0.50
Helene	377 400	2.737	36	?	Low
Rhea	527 040	4.517	1528	1.33	0.66
Titan	1 221 850	15.945	5150	1.88	2.65
Hyperion	1 481 100	21.278	410	1.47	0.11
Iapetus	3 561 300	79.330	1460	1.21	0.59
Phoebe	12 952 000	550.48 (ret)	220	0.77	0.07

The remaining outer satellites are below 50 km in diameter.

Nothing definite is known about the nature of the surface. The Voyagers could do no more than send back images of the top of a layer of orange haze; infrared images taken with the Hubble Space Telescope and, the Keck II telescope in Hawaii have shown bright patches and darker regions. It has been suggested that there could be frozen landmasses and frigid hydrocarbon seas or lakes, but we must await the arrival of the Huygens probe, scheduled to land on Titan in early 2005. Certainly Titan is unlike any other body in the Solar System. It does not seem to have an internal magnetic field, but orbits near the outer edge of Saturn's vast magnetosphere. As with most of the other satellites, its rotation is synchronous, so that the same hemisphere always faces Saturn.

Medium-sized icy satellites These seem to form 'pairs'; Rhea/Iapetus (Figure 7), Tethys/Dione, and Mimas/Enceladus. Two very small satellites, Telesto and Calypso, move in the same orbit as Tethys, while Helene shares the orbit of Dione. Hyperion, moving between the orbits of Titan and Iapetus, is irregular in shape, and has a darkish surface often regarded as 'dirty ice'; the rotation is not synchronous, and on average the rotation period is around 13 days, There are several craters, and one long ridge or scarp. It is possible that Hyperion is part of a larger body which broke up. Phœoebe, outermost of the named satellites, is much further from Saturn, and has retrograde motion, so that it must be a captured asteroid. Little is known about its surface, but it was imaged in 2004 by the Cassini probe and found to be crater scarred. It

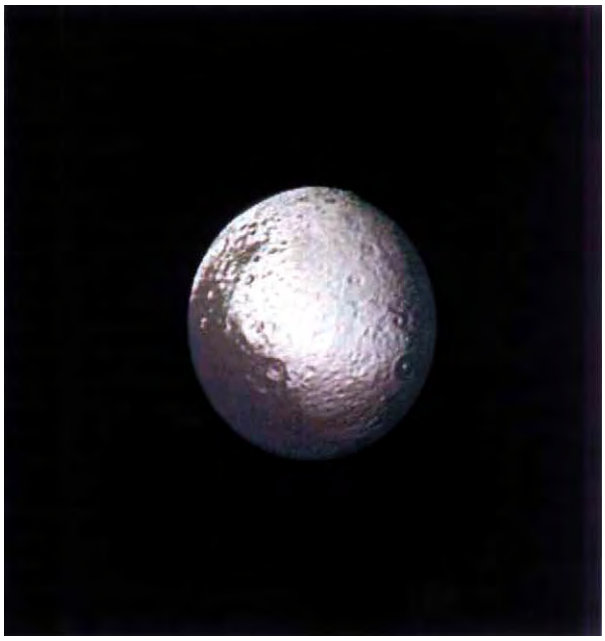


Figure 7 Iapetus, imaged by Voyager 2, 1981. The dark area is well defined; like the bright areas, it is cratered. Copyright Cawin J. Hamilton 1999.

was the first satellite to be discovered photographically (by WH Pickering in 1898).

Mimas Mimas is only slightly denser than water, and consists largely of ice, though there may well be a small rocky core. The surface is dominated by a huge, deep crater named Herschel in honour of the discoverer of Mimas. The 130 km crater is one-third

the diameter of Mimas itself, so that the impact which formed it must have come close to disrupting the entire satellite. Parallel grooves indicate that the surface must have been subjects to considerable strain.

Enceladus Enceladus is the most interesting of all the satellites from a geological point of view, because there are at least five different types of terrain. Craters exist in many areas, and give the impression of being young and sharp; there is also an extensive grooved plain which is crater-free. Surprisingly, Enceladus may be active, with a liquid interior; if so, we are seeing what is termed cryovulcanism, the icy equivalent of volcanic action. The interior of Enceladus is presumably being tidally flexed by the gravitational pulls of Saturn and the more massive satellite Dione, the orbital period of which is twice that of Enceladus. Re-surfacing has led to the obliteration of old craters.

Tethys Tethys has a very low density, and is probably composed almost entirely of ice. One crater, Odysseus, has a diameter of 400 km larger than Mimas. The main surface feature is a huge trench, running from near the north pole across the equator and along to the south polar region.

Dione Dione is only slightly larger than Tethys, but much denser and more massive. The surface is not uniform. The trailing hemisphere is relatively dark and heavily cratered, the leading hemisphere is much lighter. One prominent feature, named Amata, is associated with a system of bright wispy features which extend over the trailing hemisphere and are accompanied by narrow linear troughs and ridges. Geologically, Dione seems to have been much more active than Tethys.

Rhea Rhea is heavily cratered; as with Dione the trailing hemisphere is the darker of the two, and there are not many really large formations. There are two distinct types of terrain, the first contains craters over 40 km across, while the second, covering parts of the polar and equatorial regions, is characterized by craters of much smaller size. Rhea seems to have a rocky core, around which most of the material is ice.

Iapetus Iapetus is unusual in many ways. The trailing hemisphere is bright and icy, but the leading hemisphere is as dark as a blackboard. The demarcation line is not abrupt; there is a 200–300 km transition zone. The low density of Iapetus shows that the dark areas are due to surface materials, which have welled up from below, but their thickness is not known. It is also notable that many of the craters in the bright areas have dark floors. There have been suggestions that the dark material has been wafted on to Iapetus from the outer satellite Phœoebe, but this seems unlikely, partly because Phœoebe is so small and remote but mainly because the colours do not match. We know little about the dark area, but the bright regions contain craters of the usual type. As seen from Earth, Iapetus is very variable; it is brightest when west of Saturn, with the bright hemisphere facing us.

Minor satellites Pan, discovered on a photograph taken by the Voyager 2 space-craft, moves within the outer division of Saturn's ring system (the Encke Division). Atlas, moves near the edge of Ring A: nothing is known about its surface details. Prometheus and Pandora, the F-ring 'shepherds' have been imaged; both consist mainly of ice, and both are cratered, Prometheus shows ridges and valleys, while Pandora has two 30 km craters. Janus and Epimetheus have the same mean distance from Saturn; every four years they approach each other, and actually exchange orbits. They are irregular in shape, and may well be the remnants of a larger body which has broken up. Predictably, both are cratered.

Outer minor satellites All these are small, though one, known at present as S/2000 S3, may be almost 50 km in diameter. By May 2003 the total number of satellites had risen to 31. The outer minor satellites tend to form clusters, some with direct motion and others retrograde; most of them seem to be fragments of larger satellites which have broken up. Nothing is known about their surface features. No doubt many more tiny satellites await discovery.

Future space probes, and more powerful Earth-based telescopes, will add greatly to our knowledge. Certainly there is no doubt that to the geologist, the giant planets and their satellites are of surpassing interest.

See Also

Solar System: Asteroids, Comets and Space Dust; Neptune, Pluto and Uranus.

Further Reading

Alexander AFOB (1952) *The Planet Jupiter*. London: Faber and Faber.

Alexander AFOB (1958) *The Planet Saturn*. London: Faber and Faber.

Asimov I (1995) *The Ringed Planet Saturn*. Milwaukee: Gareth Stevens.

Beebe H (1991) *Jupiter*. Washington: Smithsonian Institution Press.
Burns M (ed.) (1986) *Satellites*. Arizona: University of Arizona Press.
Clustenis A and Taylor F (1999) *Titan: the Earthlike Satellite*. Singapore: World Scientific.
Greenberg R and Ehahic A (1984) *Planetary Rings*. University of Arizona Press.
Gehres T (ed.) (1986) *Jupiter*. Arizona: University of Arizona Press.
Gehres T (ed.) (1984) *Saturn*. Arizona: University of Arizona Press.
Hunt G and Moore P (1980) *Atlas of Jupiter*. London: Mitchell Beazley.
Hunt G and Moore P (1982) *Atlas of Saturn*. London: Mitchell Beazley.
Morrison B and Samz J (1982) *Voyager to Jupiter*. NASA.
Morrison B (1982) *Voyager to Saturn*. NASA.
Moore P (2001) *Astronomy Data Book*. London: Institute of Physics Publishing.
Peek BM (1981) *The Planet Jupiter*. London: Faber and Faber.
Rugers J (1990) *The Giant Planet Jupiter*. Cambridge: Cambridge University Press.
Rothery D (1992) *Satellites of the Outer Planets*. Oxford: Clarendon Press.
Lang K (2003) *Cambridge Guide to the Solar System*. Cambridge: Cambridge University Press.

Neptune, Pluto and Uranus

P Moore, Selsey, UK

Introduction

Five planets have been known since ancient times: Mercury, Venus, Mars, Jupiter, and Saturn, all of which are prominent naked-eye objects. Since the invention of the telescope, three more planets have been discovered beyond the orbit of Saturn: Uranus in 1781, Neptune in 1846, and Pluto in 1930, though the planetary status of Pluto is a matter for debate. This article describes the outer members of the Solar System, together with their numerous satellites.

Uranus

Uranus was discovered in March 1781 by William Herschel, one of the greatest of all observers. He was Hanoverian by birth, but spent most of his life in England; by profession he was a musician, but his main interest was in astronomy, and he built excellent reflecting telescopes. With one of these he began systematic 'reviews of the heavens', and came across an object which was certainly not a star. It showed a perceptible disk, and it moved from night to night. Herschel believed it to be a comet, but before long its planetary nature became evident.

Uranus is a giant world, but it is not of the same type as Jupiter and Saturn. Rather than being described as a gas-giant, it is better referred to as an 'ice-giant' (Figure 1). The outer atmosphere is made up chiefly of hydrogen (probably about 83% by number of molecules) and helium (15%); methane accounts for 2%, so that there are only traces of other substances. Methane freezes out at a very low temperature, and forms a thick cloud layer above which is the predominantly hydrogen atmosphere. Methane absorbs red light, which is why Uranus appears bluish-green. Minor constituents such as ethane C_2H_6 and acetylene (C_2H_2) play a role in forming 'hazes'.

Below the atmosphere come the 'ices', a mixture of water, methane, and ammonia with traces of other substances. These materials behave as liquids under the temperature and pressure conditions inside the globe. However, Uranus differs from the other giant

Figure 1 Uranus, imaged from Voyager 2 in January 1986. Compared with the other giants, Uranus is bland in appearance, and there are no well marked features.

Table 1 Data for Uranus and Neptune

	Uranus	*Neptune*
Distance from Sun	Max 3 005 200 000 km (20.088 a.u.)	4 347 000 000 km (30.316 a.u.)
	Mean 2 869 600 000 km (19.181 a.u.)	4 496 700 000 km (30.058 a.u.)
	Min 2 734 000 000 km (18.275 a.u.)	4 456 000 000 km (29.800 a.u.)
Orbital period	84.01 years.	164.8 years
Synodic period	369.66 days.	367.5 days
Rotation period	17.24 h (17 h 14.4 m)	16.4 h (16 h 7 m)
Mean orbital velocity	6.82 km s ^{1}	5.43 km s ^{1}
Axial inclination, degrees	97.86	28.48
Orbital inclination, degrees	0.773	1.769
Orbital eccentricity	0.0462	0.009
Diameter, km	Equatorial 51 118	50 538
	Polar 49 946	49 600
Mass, Earth 1	14.6	17.2
Volume, Earth 1	64	57
Escape velocity, km s ^{1}	21.1	23.9
Surface gravity, Earth 1	1.17	1.2
Density, water 1	1.27	1.77
Oblateness	0.023	0.02
Albedo	0.51	0.35
Mean surface temperature, °C	214	220

planets in that it seems to have at best a very weak internal heat-source. This means that the temperatures of the surfaces of Uranus and Neptune are almost equal, even though Neptune is so much farther from the Sun.

Uranus is unusual in one respect; its axial inclination makes more than a right angle to its orbit, so that the rotation is technically retrograde, though not usually classed as such (Table 1). From Earth the equator of Uranus is regularly presented, as was the case in 1923 and 1966; at other times a pole may lie in the centre of the disk – the south pole in 1985, for instance, and the north pole in 2030. The rotation period is only 17.2 h, so that the Uranian calendar is strange. During one orbit each pole has a 'night', lasting for 21 Earth years, with corresponding daylight at the opposite pole. The reason for the extreme axial inclination is not known.

Uranus has a magnetic field, the polarity of which is opposite to that of the Earth, and the magnetic axis is displaced from the axis of rotation by 57.6°, again for reasons which are unknown; neither does the magnetic axis pass through the centre of the globe – it is offset by 8000 km.

Our main knowledge of Uranus has been drawn from Voyager 2, the only probe to have encountered the planet. Launched in 1977, Voyager 2 encountered Jupiter in 1979 and Saturn in 1981, and on 24 January 1986, it passed over the north pole of Uranus, only about 80 000 km above the cloud-tops, before going on to rendezvous with Neptune in 1989. Clouds were recorded on the globe, and images were obtained of all the major satellites, as well as the ring system.

Table 2 Rings of Uranus

	Distance from Uranus, km	*Width, km*	*Eccentricity*	*Period, h*
6	41 837	1.5	0.0010	6.1988
5	42 235	2	0.0019	6.2875
4	42 571	2.5	0.0010	6.3628
Alpha	44 718	4 10	0.0008	6.5808
Beta	45 661	5 11	0.0004	7.0688
Eta	47 176	1.6	0.004	7.4239
Gamma	47 626	1 4	0.0001	7.5307
Delta	48 303	3 7	0.0	7.6911
Lambda	50 024	2	0.0	8.1069
Epsilon	51 149	20 96	0.0079	8.3823

The rings of Uranus are quite unlike those of Saturn; they are thin and dark, so that they are not easy to study with Earth-based telescopes. Details of the system are given in Table 2. Ten rings are known; their thickness is between 0.1 and 1 km, and only one, the outermost (the Epsilon ring), is of considerable width. This ring is appreciably eccentric; the tiny satellites Cordelia and Ophelia, discovered by Voyager 2, act as 'shepherds' to it. The rings seem to be made up mainly of particles a few metres in diameter.

Uranus has an extensive satellite system; data are given in Table 3. Only Miranda, Ariel, Umbriel, Titania, and Oberon were known before the Voyager mission (the names come from Shakespeare and Pope's poem *Rape of the Lock* – a strange departure from the usual tradition of mythology). All these were imaged by Voyager 2. The largest satellites, Titania and Oberon, are heavily cratered, but are not alike. Titania has clearly been the site of much past tectonic

Table 3 Satellites of Uranus

	Mean distance from Uranus, km	*Orbital period, d*	*Diameter, km*	*Magnitude*	*Escape velocity, km s* ^{1}	*Density, water 1*
Cordelia	49 471	0.330	26	24.2	V. low	?
Ophelia	53 796	0.372	32	23.9	V. low	?
Bianca	59 173	0.433	42	23.3	V. low	?
Cressida	54 777	0.463	62	22.3	V. low	?
Desdemona	62 676	0.473	54	22.5	V. low	?
Julia	64 352	0.493	84	21.7	V. low	?
Portia	66 085	0.513	105	21.1	V. low	?
Rosalind	69 941	0.558	54	22.1	V. low	?
Belinda	75 258	0.663	65	22.3	V. low	?
Puck	86 000	0.762	164	20.4	V. low	?
Miranda	129 400	1.414	481	16.3	0.5	1.3
Ariel	191 000	2.520	1158	14.2	1.2	1.6
Umbriel	256 300	4.144	1169	14.8	1.2	1.4
Titania	435 000	8.206	1578	13.7	1.6	1.6
Oberon	583 500	13.463	1523	13.9	1.5	1.5
Caliban	7 230 000	579	60	22.3	V. low	?
Stephano	8 002 000	676	39	24.0	V. low	?
S/2001 UI	8 571 000	769	20	25	V. low	?
Sycorax	12 179 000	1 283	120	20.7	V. low	?
Prospero	16 418 000	1 992	40	23	V. low	?
Setebos	17 459 000	2 202	40	23	V. low	?

activity; there are ice cliffs, fault valleys, and trench-like features, one of which extends for over 1450 km. Oberon has a brownish surface; some of the craters are dark-floored, and there are several systems of bright rays. Ariel is also cratered, but the dominant features are broad, branching, smooth-floored valleys which look as though they have been cut by fluid, though water is not a likely candidate because of Ariel's small size and low temperature. Umbriel has a darker and more subdued surface, with one prominent feature of uncertain nature near the edge of the best image (it must be remembered that Voyager could cover only half of the total surface). Miranda, passed by Voyager at a mere 3000 km, has an amazingly varied landscape; there are several distinct types of terrain – old, cratered plains, brighter areas with cliffs and scarps, and large, trapezoidal regions known as corona. Large craters are lacking, but there are fault valleys, parallel ridges, and graben up to 15 km across. Miranda presents real problems of interpretation, particularly because the various types of terrain seem to have been formed at different periods. It has been suggested that the satellite may have been shattered and re-formed several times, but this would involve considerable heating, which, in view of Miranda's small size, and icy nature, does not sound probable.

All the remaining satellites are small, only Puck is as much as 100 km in diameter. The five outer satellites have retrograde motion, and are presumably captured asteroids.

Neptune

Neptune was discovered in 1846 by J Galle and H D'Arrest, from the Berlin observatory. Irregularities in the movements of Uranus had enabled the French mathematician UJJ Le Verrier to calculate the position of the body responsible for them (similar calculations by JC Adams in England had given much the same result). Neptune is a twin of Uranus, but the two worlds are by no means identical. Neptune is very slightly the smaller of the two, but it is appreciably denser and more massive, and is a much more dynamic world (Table 1). Unlike Uranus, it has a strong internal heat source, and sends out 2.6% more energy than it would do if it depended entirely upon what it receives from the Sun. It does not have an exceptional axial inclination; at the time of the Voyager 2 pass it was the south pole which was in sunlight. There is a magnetic field, but the magnetic axis is displaced by 47° from the axis of rotation and does not pass through the centre of the globe, so that in this respect Neptune really does resemble Uranus. The interior of Neptune is presumably not unlike that of Uranus, apart from the greater internal heat source.

The atmosphere consists mainly of hydrogen, with an appreciable amount of helium and some methane together with traces of other substances such as hydrogen cyanide, acetylene, and ethane. Voyager identified various cloud layers, at a level where the pressure is 3.3 bars there is a layer which seems to be of

hydrogen sulphide, above which are layers of hydrocarbons, with a methane layer and an upper methane haze. Above the hydrogen sulphide layer there are discrete clouds with diameters up to 100 km, casting shadows on the cloud deck 50 to 75 km below. These clouds may be described as 'methane cirrus'. Temperature measurements from Voyager show that there is a cold mid-latitude region, with a warmer equator and pole (we know little about the north pole, which was in darkness during the Voyager encounter). There are strong winds; most of them blow in a westerly direction (that is to say, opposite to the planet's rotation), and are distinctively zonal. At the equator they blow westward at up to 450 m s ^{1}. Further south they slacken, and beyond latitude $-50°$ they become eastward (prograde) up to 300 m s ^{1}, decreasing once more near the south pole. There is, in fact, a broad equatorial retrograde jet between latitudes $+45°$ and $-50°$, with a relatively narrow prograde jet at around latitude $-70°$.

At the time of the Voyager encounter the most conspicuous feature on the disk was the Great Dark Spot, a huge oval with a longer axis of 10 000 km, drifting westward relative to the adjacent clouds, it was a high-pressure area, rotating counter-clockwise and showing all the signs of an atmospheric vortex. Hanging above it were methane cirrus clouds, and between these and the main cloud deck there was a clear region 50 km deep. Other, smaller spots were seen at different latitudes, and the whole disk was extremely active. Later images obtained with the Hubble Space Telescope show that the Great Dark Spot has disappeared, so that the surface shows marked changes over relatively short periods.

Neptune has an obscure ring system (Table 4). The outer ring, named after Adams, is 'clumpy', with three brighter arcs, while the Lassell ring is a diffuse band of material containing a high percentage of very small particles. There may be 'dust' extending from the inner Galle ring almost down to the cloud-tops.

Eleven satellites are known (Table 5), but of these only two, Triton and Nereid, were discovered before the Voyager 2 fly-by. Triton is one of the most remarkable bodies in the entire Solar System. It was found by the English astronomer W Lassell a few weeks after the discovery of Neptune itself, and is brighter than any of the satellites of Uranus; it is also more reflective, with an albedo in places of 0.8, and it is the coldest world ever encountered by a space-craft – the temperature is $-235°$C, a mere $18°$ above absolute zero. The globe seems to be made up of a mixture of rock and ice. There is an extensive though very tenuous atmosphere, made up almost entirely of nitrogen with a trace of methane. Triton has retrograde motion, and there seems no doubt that it is a captured body rather than a bona fide satellite.

The surface is very varied. There is a general coating of ice, presumably water ice overlain by nitrogen ice; there is little surface relief, and there are few craters. The area surveyed by Voyager 2 was divided into three parts: polar (Uhlanga Regio), eastern equatorial (Monad) and western equatorial (Bubembe Regio). The polar area is covered with a pink cap of nitrogen snow and ice, and there are geysers, completely unexpected before the Voyager mission. Apparently there is a sub-surface layer of

Table 4 Rings of Neptune

	Distance from centre of Neptune, km	*Width, km*
Galle	41 900	2000
Le Verrier	53 200	110
Lassell	53 200	4000
Arago	57 200	100
	61 950	(indistinct)
Adams	62 933	50

Table 5 Satellites of Neptune

	Mean distance from Neptune, km	*Orbital period, d*	*Orbital eccentricity*	*Orbital inclination*	*Diameter, km*	*Magnitude*
Naiad	48 227	0.29	0.0003	4.74	58	25
Thalassa	50 075	0.31	0.0002	0.21	80	24
Despina	52 526	0.33	0.0001	0.07	148	23
Galatea	61 953	0.43	0.0001	0.05	158	23
Larissa	71 548	0.55	0.0014	0.20	208	21
Proteus	1 17 647	4.12	0.0004	0.04	436	20
Triton	33 4 76	5.87	0.0000016	157.34	2705	13.6
Nereid	5513	360.14	0.7512	7.23	340	18.7
S/2002 N2	20 200	2525	0.17	57	40	25
S/2002 N3	21 390	2751	0.47	43	40	25
S/2002 N1	21 990	2868	0.43	121	40	25

liquid nitrogen. If any of this migrates upward, the pressure is relaxed and the nitrogen explodes in a shower of ice and vapour, travelling quickly up the nozzle of the geyser-like vent – fast enough to make it rise to several kilometres before falling back; the outrush sweeps dark débris along it, blown by winds in the tenuous atmosphere. The edge of the cap is well-defined, and north of it there is a darker, redder region. Monad Regio is part smooth, part hummocky, with rimless pits (paterae), mushroom-like features (guttae) and low-walled plains; Bubembe Regio is characterized by the so-called cantaloupe terrain – a nickname given to it because of its superficial resemblance to a melon-skin! Fissures cross it, meeting at elevated X or Y junctions; this is probably the oldest part of Triton's surface. Of course, we have no information about the hemisphere which was in darkness at the time of the Voyager encounter.

Nereid, the other satellite known before the Voyager mission, was discovered by G Kuiper in 1949. It is small, only 340 km in diameter, and has a very eccentric orbit, so that its distance from Neptune ranges between 1.35 million km and 9.62 million km. The orbital period is just over 360 days, but it is unlikely that its rotation period is synchronous; as with Hyperion in Saturn's system, the rotation period may be chaotic. It has direct motion, and although not well imaged from Voyager 2, it seems to be fairly regular in shape; there were vague indications of a few large craters. Of the other satellites, Proteus is actually larger than Nereid, but its closeness to Neptune means that it is not observable with Earth-based telescopes. The rotation is synchronous, and the albedo is low; it has been said that Proteus is 'as dark as soot'. The area imaged by Voyager is dominated by a circular depression, Pharos, 225 km in diameter and up to 15 km deep. The remaining inner satellites are small, icy, and presumably cratered; Voyager sent back one image of Larissa. The orbit of the 158 km satellite Galatea is very close to the Adams Ring. Three outer asteroidal satellites were discovered in 2001.

Pluto

Pluto, discovered by Clyde Tombaugh in 1930, is an enigma; it is smaller than the Moon or even Triton, and has an orbit which is both eccentric and inclined. It has one satellite, Charon. Data for Pluto and Charon are given in Table 6.

The calculations leading to the discovery of Pluto were made from 1905 by Percival Lowell, and were based upon perturbations of Neptune and (particularly) Uranus. Searches carried out with the Lowell refractor at Flagstaff in Arizona were unsuccessful, but the planet was finally identified at Flagstaff by Tombaugh, not too far from the position predicted by Lowell. After some discussion it was named Pluto, after the God of the Underworld. When near perihelion, it is closer-in than Neptune, as was the case between 1979 and 1999, but there is no danger of collision, because Pluto's orbit is inclined at an angle of 17°. After the discovery it was found that Pluto had been photographed at Flagstaff in 1915 and Mount Wilson in 1919, but had been missed because it was so much fainter than had been expected. A telescope of fair size is needed to show it at all; Voyager went nowhere near it, but some surface features have been recorded by the Hubble Space Telescope (**Figure 2**). There are indications of a dark equatorial band and brighter polar areas. The axial inclination is 122°, so that, as with Uranus, the rotation is technically retrograde; the rotation period is 6.34 days.

Pluto's density, twice that of water, indicates that the globe contains more rock than in the icy satellites of the giant planets, but our knowledge of the internal structure is decidedly meagre. Pluto may or may not be differentiated, but the gravitational pressure may not be adequate to increase the rock density deep inside the globe to a marked degree. The surface is coated with methane ice at least in some areas. There is a tenuous but surprisingly extensive atmosphere, made up chiefly of nitrogen together with methane. As Pluto moves out toward aphelion, due in the year 2114, the temperature will drop and the atmosphere may freeze out, so that for part of its long 'year' there may be no atmosphere at all.

Table 6 Pluto and Charon

	Pluto	*Charon*
Distance from Sun, km	Max 7 381 200 000 Mean 5 906 400 000 Min 4 445 800 000	
Orbital period, days	90 465 (247.7 years)	6 d 9 h 17 m (round Pluto)
Rotation periods	6 d 9 h 17 m	6 d 9 h 17 m
Mean orbital velocity	4.75 km s ^{1}	0.23
Axial inclination, degrees	122.46	
Orbital inclination, degrees	17.14	(to Pluto) c 9
Orbital eccentricity	0.2488	0.0076
Diameter, km	2 324	1270
Density, water 1	2.05	2
Volume, Earth 1	0.006	
Mass, Earth 1	0.0022	
Max. surface temperature, °C	about 233	
Escape velocity, km s ^{1}	1.18	0.58
Surface gravity, Earth 1	0.06	0.21
Albedo	0.55	0.36

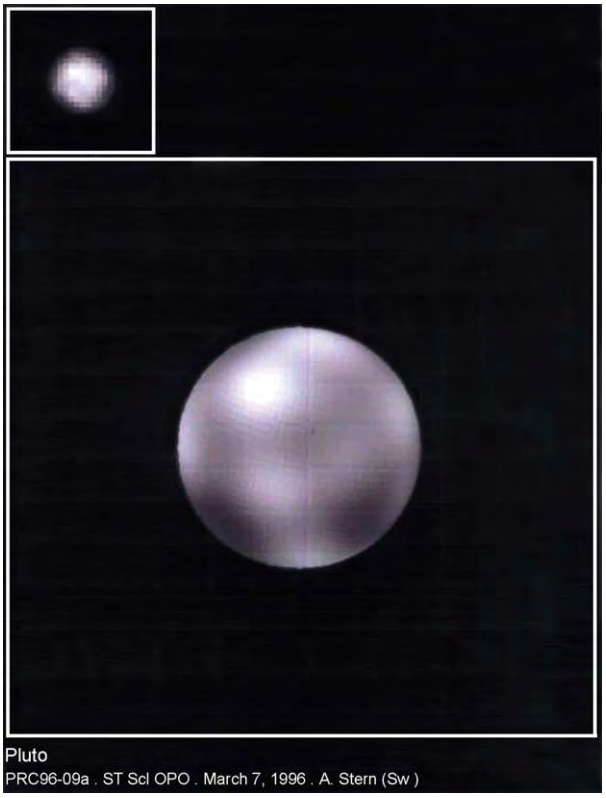


Figure 2 Pluto, imaged with the Hubble Space Telescope. Regions of different brightness are shown, but of course no definite surface features.

Pluto has a companion, Charon, discovered in 1978. Its diameter is more than half that of Pluto, so that it cannot be regarded as a conventional satellite. Its orbital period is the same as Pluto's rotation period, and the surface-to-surface distance between the two bodies is no more than 18 000 km. To an observer on Pluto, Charon would remain 'fixed' in the sky. The density of Charon is lower than that of Pluto; the surface is probably coated with water ice, and there is no trace of atmosphere. Mutual transit and occultation events for Pluto and Charon were observed between 1985 and 1990, and were very informative. When Charon passed behind Pluto it was completely hidden, and Pluto's spectrum could be seen alone, when Charon passed in front of Pluto the two spectra were seen together, so that of Pluto could be subtracted.

From 1992 many small bodies have been found orbiting the Sun beyond Neptune. These make up the Kuiper Belt, named after the Dutch astronomer GP Kuiper, who suggested its existence (*see* **Solar System:** Asteroids, Comets and Space Dust. One of these, Quaoar, may be as large as Charon, and others are around 1000 km in diameter. There are grounds for proposing that Pluto and Charon should be regarded simply as the largest known Kuiper Belt objects, though some astrnonomers are reluctant to deprive Pluto of its planetary status!

Is there another large planet moving far beyond Neptunes and Pluto? Slight irregularities in the movements of Uranus and Neptune indicated that this might be the case, and periodical searches for 'Planet X' have been carried out, but with no success. Recently it has been claimed that improved values for the masses of Uranus and Neptune show that no unexplained perturbations occur, and there is no need for Planet X, but one thing is certain; Pluto's mass is too small to cause the effects which led Lowell to calculate a position for it. Therefore, either Lowell's reasonably correct result was purely fortuitous, or else the planet for which he was searching has yet to be discovered. If Planet X really exists, it will no doubt eventually be found.

Sedna, discovered in 2004, may be 1300 km in diameter, its orbital period is over 12 000 years and it may recede almost as far as the Oort Cloud.

See Also

Solar System: Asteroids, Comets and Space Dust; Meteorites; Jupiter, Saturn and Their Moons.

Further Reading

Chuikshank D (ed.) (1995) *Neptune and Triton*. Arizona: University of Arizona Press.

Elliott and Kehh (1984) *Rings*. Cambridge Massachusetts: MIT Press.

Miner E (1998) *Uranus*. New York: Wiley/Praxis.

Hunt G and Moore P (1990) *Atlas of Uranus*. Cambridge: Cambridge University Press.

Hunt G and Moore P (1995) *Atlas of Neptune*. Cambridge: Cambridge University Press.

Miner ED and Wessen RR (2002) *Neptune*. Berlin, Heidel berg, New York: Springer Verlag.

Stern A and Mitton J (1999) *Pluto and Charon*. New York: Wiley.

Tombaugh C and Moore P (1980) *Out of the Darkness: the Planet Pluto*. London: Stackpole and Lutterworth.

Bergstralh, *et al.* (eds.) (1991) *Uranus*. Arizona: University of Arizona Press.

Moore P (1996) *The Planet Neptune*. New York: Wiley.

SPACE DUST

See **SOLAR SYSTEM: Asteroids, Comets and Space Dust**

STRATIGRAPHICAL PRINCIPLES

N MacLeod, The Natural History Museum, London, UK

Introduction

Stratigraphy is the branch of geology that deals with the formation, composition, sequence, and correlation of stratified rocks. Since the whole Earth is stratified, at least in a broad sense, bodies of all the different types of rock – igneous, sedimentary, and metamorphic – are subject to stratigraphic study and analysis. In most cases, however, stratigraphy focuses on the evaluation of sedimentary rock strata. Modern principles of stratigraphic analysis were developed in the eighteenth and nineteenth centuries by geologists such as Niels Stensen, James Hutton, Georges Cuvier, William Smith, and Charles Lyell. By 1900 all the intellectual tools needed for the description, sequence, and correlation of strata were in place. Shortly after 1900, the tools needed to establish the absolute ages of minerals containing unstable radioisotopes also became available, giving stratigraphers a physical basis for making chronostratigraphic correlations, at least in certain favourable stratigraphic situations. Since the 1950s efforts have been made to establish international standards for stratigraphic nomenclature and the usage of stratigraphic terms and the internationally agreed designation of 'type-sections' or stratotypes for various sorts of stratigraphic unit, especially those relating to chronostratigraphy.

First Principles

The study of stratigraphy began with attempts to understand common observations, such as what the rocks we call fossils are and how the rocks that comprise mountains came to be elevated above the land surface. Of course, both fossils and mountains were well known to ancient Greek natural historians, such as Plato, Hereodotus, Aristotle, Xenophanes, and Pliny. Although a variety of explanations were offered for these phenomena, no systematic investigations of modern aspect were carried out by these classical scholars, according to the intellectual style of their time. The organic nature of fossils was recognized by a number of Renaissance scholars, including Leonardo da Vinci (1452–1519) and Conrad Gesner (1516–1565). Da Vinci's writings were particularly prescient in that he recognized that fossil mollusc shells from the tops of mountains were similar to the shells of modern molluscs and that this similarity implied that sediments occupying the mountain tops must originally have been deposited beneath marine waters. These were isolated musings, however.

The first modern treatment of a stratigraphic problem was published by Niels Stensen (1638–1686, also known by his anglicized literary name, Nicholas Steno) in 1669 (*see* **Famous Geologists:** Steno). Most scholars mark Steno's *De solido intra solidum naturaliter contento disseratiinis prodomus* as the first stratigraphic treatise. In this short work–which was presented to Steno's patron, the Grand Duke Ferdinand II of Tuscany – Steno establishes three cardinal principles of stratigraphic analysis and then uses these to reconstruct the geological history of Tuscany. Steno's principles are as follows.

1. Original horizontality: unconsolidated sediments deposited on a solid base must have originally formed horizontal layers, since the sediment particles would have 'slithered' to the lowest point. Thus, consolidated strata inclined at an angle must have become tilted after consolidation.
2. Original continuity: layers of unconsolidated sediments deposited on a solid base would have formed continuous sheets of material. Thus, bands of consolidated sediments whose ends have been broken must have experienced this breakage and erosion after consolidation.
3. Superposition: since each layer of unconsolidated sediment deposited on a solid base must have formed after the basal layer had been deposited, overlying layers of sediment are younger than underlying layers.

Using these principles, Steno argued that Tuscan geology, and especially the stratified sediment layers

forming its mountains, represented the remains of a series of subterranean-erosion and land-surface-collapse events (Figure 1). Not only did this model reconcile the cyclic and directional aspects of the Tuscan stratigraphic record, it also established the principal of stratigraphic correlation as the matching of stratigraphic observations from distant outcrops in order to obtain a sense of a rock body's geometric structure (Figure 2).

The next significant contribution to stratigraphic principles was made in 1785 by the Scottish lawyer–gentlemen farmer James Hutton (1726–1797) (*see* **Famous Geologists:** Hutton), who stressed the cyclic aspects of the stratigraphic record in his doctrine of

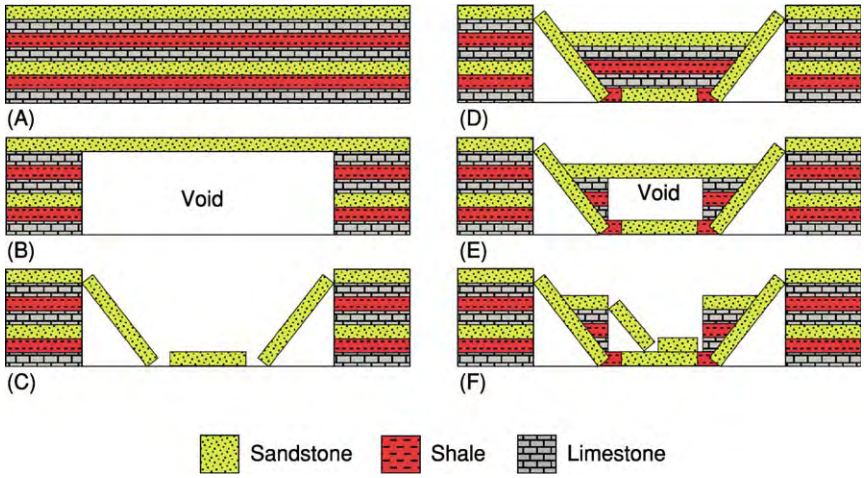


Figure 1 Steno's conceptual interpretation of the stratigraphic history of Tuscany. (A) Flat lying continuous sediments were deposited beneath marine waters. (B) Lithified sediments were uplifted, and subterranean voids or caverns developed through the erosive action of subsurface waters. (C) When the subterranean voids grew sufficiently large, the roofing layers collapsed, elevating the cavern walls, down dropping flat laying layers that remained intact, and tilting blocks adjacent to the elevated areas. (D) Submergence of the entire land surface once again caused flat lying continuous sediments to be deposited. (E) These new sediments were lithified and uplifted, after which new cavernous voids developed. (F) A new round of erosional collapse further modified the landscape. Note how Steno's model encompasses both the apparently directional nature and the cyclic nature of stratigraphic deposits and landscape formation. (Redrawn from Steno's diagram in *De solido intra solidum naturaliter contento disseratiinis prodomus*.)

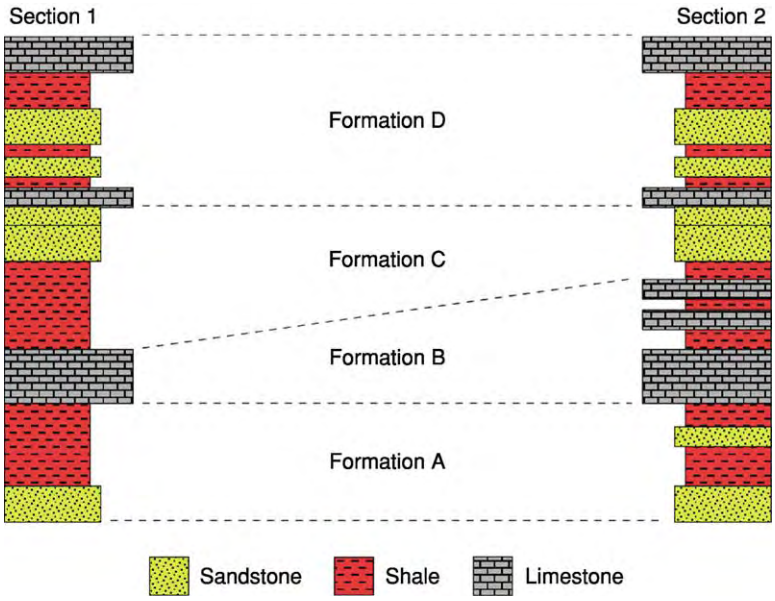


Figure 2 In addition to developing his theory of landscape formation, Steno stressed the importance of stratigraphic correlation the matching of stratigraphic sequences between outcrops. In this illustration two hypothetical outcrop sections have been correlated based on rock type and subdivided into lithologically unified packages of strata.

uniformitarianism. Citing evidence from the angular unconformities exposed at such Scottish localities as Jedburgh and Siccar Point, Hutton reasoned that the originally horizontal marine sediments of the lower succession must have been consolidated, then tilted as they were raised up above the water's surface, planed off by erosion, submerged, and buried by additional horizontally deposited sediments, which were then consolidated, before the entire sequence was lifted again to become the rock bodies we see before us at these, and other, localities. Hutton believed that these erosion–deposition–uplift cycles had been repeated endlessly in Earth history, implying that: first, the Earth itself is very old; second, the processes we see working today (e.g. erosion, deposition, gradual uplift) operated in the past; third, the power for uplift came from the heat generated by compaction, supplemented by heat at depth left over from the Earth's initial formation; and, fourth, the ultimate purpose of this system was to produce a self-renewing Earth that was 'adapted to the purposes of man'. In particular, Hutton denied that fossils provided any evidence for the directional passage of time because each uniformitarian cycle's biota was 'perfect'.

Slightly later (in 1812), the French Baron Georges Cuvier (1769–1832) (*see* **Famous Geologists:** Cuvier) published a summary of his palaeontological studies in the Paris Basin in his book *Recherches sur les Ossemens Fossils*, the first chapter of which took issue with Hutton's uniformitarian approach to stratigraphic analysis. Cuvier argued that the abrupt disappearances of entire fossil marine faunas that characterize several horizons within this basin, and the equally abrupt appearances of new terrestrial faunas in strata lying just above these marine beds, were evidence for the repeated, sudden, and, in ecological terms, catastrophic elevation of the land. In contrast to Hutton, who believed in endless cycles, Cuvier and his colleagues – who came to be known as 'catastrophists' – envisioned an Earth whose internal core was undergoing constant thermal contraction. As this core pulled away from the hard crust gaps opened up. It was these gaps that were responsible for the catastrophes. In a manner analogous to that proposed in Steno's model, crustal failure occurred when the subterranean gaps become too large to support the burden of the overlying crust. It was supposed that these failures happened suddenly, down-dropping entire regions, the surrounding parts of which would appear to be thrust up (in relative terms) as mountains. Unlike Hutton's endless uniformitarian cycles, Cuvier's hypothesis of Earth history was resolutely directional and finite. The Earth would eventually cool to the point where no more contraction would take place, thus bringing the catastrophes to an end. Also, unlike Hutton, the catastrophists saw extinction as a real phenomenon, with new biotas responding to the changed environment in unique ways.

The next major contribution to stratigraphy was made by the English canal surveyor and geologist William Smith (1769–1839) (*see* **Famous Geologists:** Smith). Smith was the first to recognize the difference between lithostratigraphy (the characterization of rock strata by the kind and/or arrangement of their mineralogical constituents) and biostratigraphy (the characterization of rock strata by their biological constituents). Before Smith, the remains of once-living creatures and the mineral particles of which sedimentary rocks are made were considered to be of equal value in recognizing strata. Smith made a conceptual distinction between lithological and palaeontological sources of stratigraphic information and, by careful analysis of the fossils contained in stratigraphic bodies, demonstrated that strata with very similar lithological constituents could be distinguished on the basis of their fossil content. Even more importantly, Smith showed that the successive biotas preserved in the sedimentary strata of the British Midlands always occurred in the same sequence, regardless of the character of the local lithological sequences. This key stratigraphic principle later became known as the 'principle of faunal succession' (Figure 3). By applying the principle of faunal succession to his biostratigraphical observations, Smith was not only able to predict more accurately the types of rock that would be encountered during canal construction, but also able in 1815 to produce the first modern geological map. While William Smith was not given to abstract theorizing, his commitment to field observations, his willingness to accept those observations at face value, and his use of fossil extinction events as a basis on which to recognize the directional passage of time were far more in line with the philosophical tenets of catastrophism than with those of uniformitarianism.

Uniformitarianism's champion was Charles Lyell (1797–1875) (*see* **Famous Geologists:** Lyell). Lyell accepted the cyclic nature of Huttonian uniformitarianism to the extent that he denied the possibility of both extinction and evolution (though, to be fair, it must be said that the latter was denied by Cuvier as well, albeit on different grounds). Lyell also emphasized and greatly developed Hutton's idea of a mechanistic uniformitarianism in which known natural laws and processes operated at rates comparable to those observed today. Lyell believed that these mechanisms were responsible for all features of the geological record. It is interesting to note that Cuvier, Agassiz, (*see* **Famous Geologists:** Agassiz) and the other

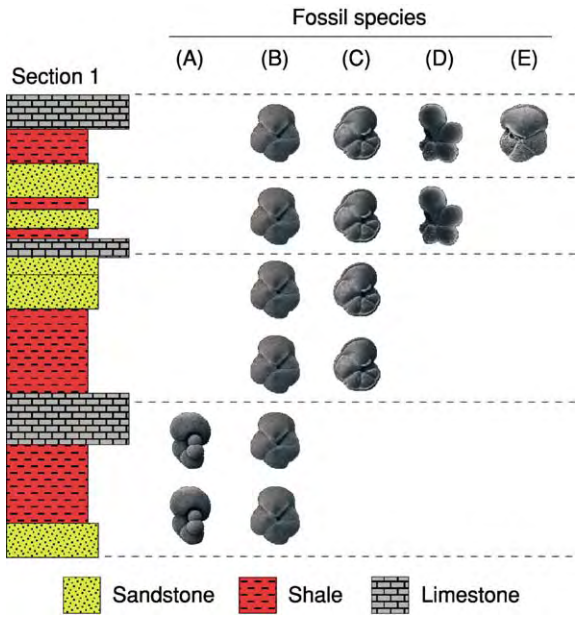


Figure 3 Organization of stratigraphic sequences into units based on their fossil content using the principle of faunal succes sion. In this illustration the distributions of five fossil planktonic foraminiferal species have been used to recognize stratigraphic units on the basis of unique associations of species. The principle of faunal succession can be used to recognize stratigraphic units because fossil species are individualized in the sense that they have definite and unique starting (speciation) and ending (global extinction) points. Thus, the stratigraphic range of a fossil species encompasses a distinct time interval. Note also that palaeontolo gically defined stratigraphic units differ in both number and kind from lithologically defined units (compare with **Figure 2**).

scientific catastrophists were aligned with Lyell in accepting the principle of mechanistic uniformitarianism. Lyell summarized his arguments, and supported them with examples drawn from his geological travels throughout Europe, in a massive three-volume work *Principles of Geology* published in 1830–1833.

While the uniformitarian–catastrophist debate has often been portrayed as a triumph of dispassionate scientific reason over theologically driven special pleading, with the Lyell uniformitarians founding the sciences of stratigraphy and sedimentary geology as we know them today, a more faithful description of the historical record reveals a far more interesting story. Lyellian uniformitarianism did indeed triumph, but not so much over the scientific catastrophism of Cuvier, Brongniart, d'Orbigny, and Agassiz as over the theological catastrophism embraced by the school of Natural Theology (especially in England) and sheer scientific fantasy. Lyell's reasoned approach, which emphasized modern processes working over long periods of time, appealed to many, not least Charles Darwin (*see* **Famous Geologists:** Darwin) who read Lyell's treatise during the *Beagle* voyage and used Lyellian principles as a basis for his geological explorations. Lyell's commitment to the basic uniformitarian doctrine of endless and ahistorical cyclicity, however, was not accepted even among Lyell's contemporaries. Lyell was caricatured for his position by Henry de la Beche in a famous cartoon (**Figure 4**) and was forced to retract from it by stages in subsequent editions of his *Principles* volumes. Neither was Lyell's view of the value of fossils in stratigraphic correlation – at least for higher taxonomic groups – accepted by his contemporaries, much less by contemporary stratigraphers. Modern uniformitarianism is a combination of the Huttonian–Lyellian emphasis on modern observable processes operating over long periods of time, but nevertheless allowing for the incorporation of processes that have no modern counterpart (e.g. Louis Agassiz's continental glaciations, enormous flood-basalt volcanic eruptions, asteroid impacts), and a thoroughly catastrophist emphasis on extinction, the existence of intervals of (geologically) rapid and widespread global change, and the directional nature of geological time.

Following Smith's demonstration of the power of biostratigraphy, the forefront of stratigraphic research turned to the identification of biostratigraphic zones that could be used to facilitate long-range stratigraphic correlation (e.g. intrabasinal, interbasinal, and intercontinental). This immediately raised a further conceptual problem. Did the identification of the same biozone in different localities mean that the resulting correlation located the two sections in terms of their position in the sequence of biotas preserved over geological time (homotaxis) or in terms of geological time itself (homochrony)? These concepts are distinct because the same sequence of events could be preserved at different localities without the individual events having taken place at the same time.

Until 1900 stratigraphers had been forced to couch their observations in terms of relative time (e.g. event A took place before or after event B) because there was no way to measure absolute time in stratigraphic successions. Of course, attempts to estimate absolute time were made, usually based on modern rates of sediment accumulation and estimates of compaction ratios for different types of sedimentary rock. Nevertheless, since these rates and ratios vary widely, and since there was no way of confirming that any given estimate was correct, such calculations were approximate at best.

This situation changed in the early 1900s, however, with the discovery of natural radioactivity and unstable radioisotopes of naturally occurring elements. Radioisotopes have unstable nuclei that spontaneously decay

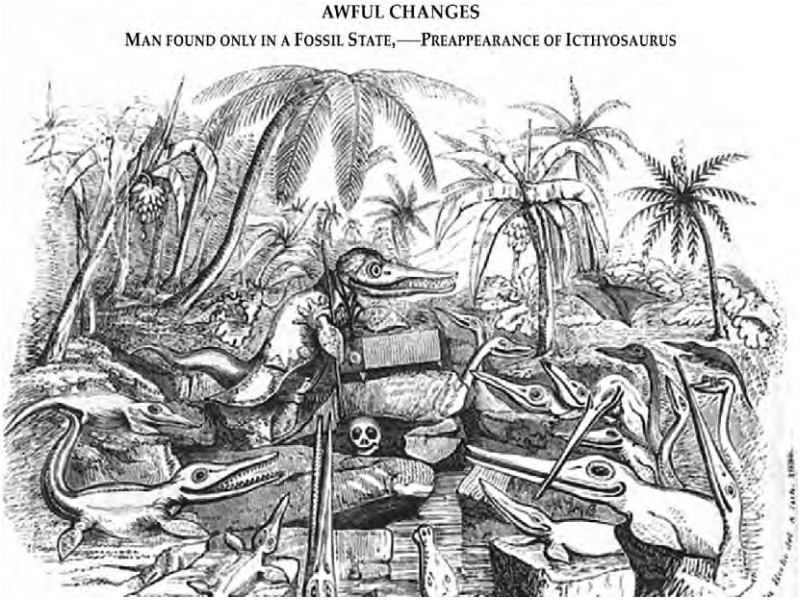


Figure 4 Henry de la Beche's caricature of Charles Lyell as 'Professor Icthyosaurus' lecturing to an eager audience of saurian students at some time in the future on the topic of an 'insignificant' and 'lower order' fossil animal from Earth's distant past (the representation of a human skull in the drawing's centre). This image, which was originally published as the frontispiece to Frank Buckland's *Curiosities of Natural History*, poked fun at Lyell's belief in the cyclic, or uniformitarian, nature of geological processes, which predicted the re emergence of extinct fossil forms when future environmental conditions matched those of the past. Modern uniformitarianism no longer embraces this aspect of Hutton and Lyell's original formulation but represents a dynamic amalgam of nineteenth century uniformitarian and catastrophist theory. See text for discussion.

through the emission of subatomic particles from the isotope's nucleus at a fixed and measurable rate. Daughter isotopes are produced as the products of this decay process, along with various types of radiation. If the amounts of original radioisotopic material of a specific type in a particular mineral species and the daughter-product isotope are known, the absolute age of the mineral can be calculated, subject, of course, to several assumptions (e.g. a correct value of the decay constant, accurate measurements, no loss of daughter-product isotope) (*see* **Time Scale**).

Unfortunately, accurate isotopic dating cannot usually be carried out on sediments directly. Most sedimentary rocks are composed of mineral grains whose origin predates that of the sedimentary rock body by a substantial time interval. In some instances, though, a layer of volcanic material (e.g. an ash-fall tuff) with newly formed mineral crystals can become interbedded in a suite of sedimentary rock. In such cases, the age obtained from the volcanic deposit can be used to constrain the ages of the immediately overlying and underlying sediments, subject, once again, to assumptions. By using isotopically datable materials located stratigraphically near major biostratigraphically defined boundaries in the stratigraphic record (*see* below and **Time Scale**), it is possible to estimate absolute ages for these boundaries.

Stratigraphic Classification

As stratigraphers combined the principles of stratigraphic analysis set down by Steno, Hutton, Cuvier, Smith, Lyell, and others with lithostratigraphic, biostratigraphic, and geochronological observations during the first half of the twentieth century, the true geometric relations between observed lithostratigraphic and biostratigraphic units emerged, along with their mutual relations to an entirely conceptual 'chronostratigraphy' (the characterization of rock strata by their temporal relations). These concepts are illustrated in Figure 5, and are usually discussed in terms of the distinction between rock stratigraphic units (that are distinguished by physical or biotic criteria that can be observed at the outcrop, core, well-log, etc.) and time stratigraphic units (that are in all cases inferences based on stratigraphic observations, but have the advantage of being referable to a common geological time-scale). There has been, and continues to be, much confusion over the use of these terms, primarily because of the genuine subtlety of the distinction, but also because of problems stemming from the definition of certain sorts of stratigraphic unit (e.g. biostratigraphic Oppel zones, which are defined on rock-stratigraphic criteria chosen for their supposed ability to achieve time-stratigraphic

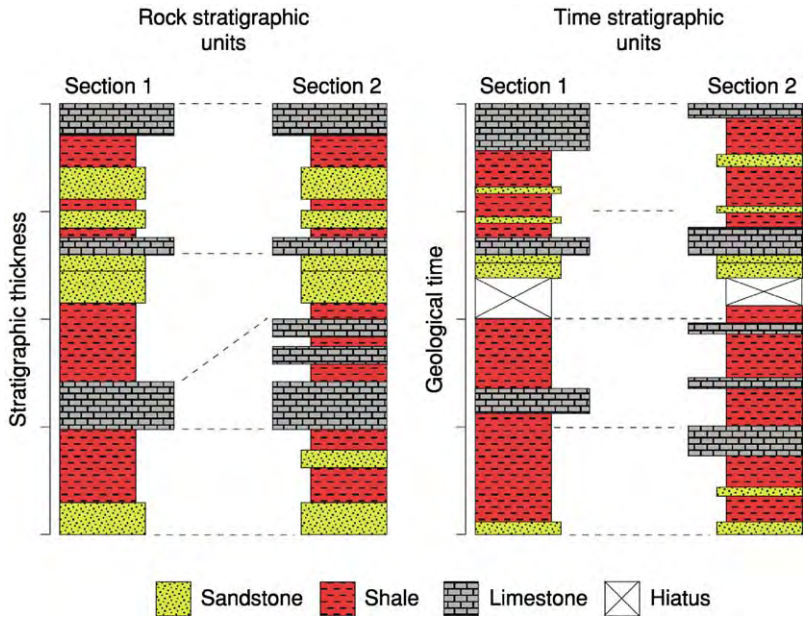


Figure 5 The difference between rock stratigraphic units (left) and time stratigraphic units (right). In this illustration the rock stratigraphic units, along with their lithostratigraphic correlation, are scaled to stratigraphic thickness, as they would be observed in a field study. When these same sections are portrayed as time stratigraphic units (and organized according to the time intervals during which they were deposited), however, the character of their comparative relations (both inter sectional and intra sectional, as well as their inter section correlations) changes.

correlations) and the fact that many stratigraphers prefer to report their rock-stratigraphic observations (e.g. position in a measured section or core) in terms of time-stratigraphic inferences.

In order to stabilize stratigraphic classification and nomenclature, the International Subcommisson on Stratigraphic Classification (ISSC) (originally the International Subcommission on Stratigraphic Terminology (ISST)) was created in 1952 at the 19th International Geological Congress (Algiers). From 1952 to 1965 the ISSC operated as a standing committee under successive international geological congresses. In 1965 responsibility for the ISSC was transferred to the International Union of Geological Sciences (IUGS), where it remains. The ISSC maintains a web site at http://www.geocities.com/issc arg/. The ISSC has several purposes. Among these is to publish and maintain the International Stratigraphic Guide, whose purpose is to promote international agreement on principles of stratigraphic classification and to develop a common internationally acceptable stratigraphic terminology and rules of stratigraphic procedure. The various stratigraphic-unit concepts and definitions currently recognized by the ISSC are summarized briefly below.

Lithostratigraphic Units

The basic unit of lithostratigraphy is the formation, which is the smallest mappable rock unit possessing a suite of lithologic characteristics that allow it to be distinguished from other such units. Formations need not be lithologically homogeneous, but the entire interval of strata should be diagnosable. Moving up the lithostratigraphic hierarchy to more inclusive units, a set of contiguous formations may be combined to form a group (e.g. the Lias Group), membership of which is usually identified based on common lithological characteristics (e.g. dominantly argillaceous facies) or genetic characteristics (e.g. a suite of formations bounded by two basin-wide unconformities). Occasionally, contiguous groups will themselves be placed into subgroups or supergroups (e.g. the Newark Supergroup, the Wealden Supergroup) based on genetic characteristics. Subgroups and supergroups may also include formations not previously assigned to a group. The most inclusive lithostratigraphic unit is a complex, which is distinguished by its diverse lithological composition – including sedimentary, metamorphic, and/or igneous rocks – and its intricate structure.

Moving down the lithostratigraphic hierarchy to more exclusive units, a member is a subdivision of a formation, recognized on lithologic criteria (e.g. the sandy member of a formation representing a suite of deltaic strata). Typically, members consist of more than a single bed, although some massive bodies with no internal stratification are recognized as members. The smallest formal lithostratigraphic unit is a bed,

which is a thin lithostratigraphically monotonous sequence with some locally unique lithological character (e.g. the *Hypsilophodon* Bed). A hypothetical example of this lithostratigraphic hierarchy is presented in Figure 6.

Igneous and/or metamorphic rock bodies of tabular form and stratified nature may be admitted within this lithostratigraphic classification, either in themselves or in combination with adjacent sedimentary units. Igneous rock bodies that cut across stratified rocks of any type can be handled within this scheme under the informal designation of being associated with (in the sense of 'bounded by' or 'included within') a larger, formal, lithostratigraphic unit.

Biostratigraphic Units

The basic unit of biostratigraphy is the biozone, which is any unit of rock distinguished from other such units on the basis of its fossil content. Unlike formations, biozones do not need to be mappable units and so can vary greatly in thickness and geographical extent. Biozones may be defined on the basis of a wide variety of criteria (*see* **Biozones**). Intervals of strata between biozones that lack fossils are referred to as barren interzones, while barren intervals within biozones may be termed barren intrazones. Moving up the biostratigraphic hierarchy, a set of contiguous biozones may be grouped into superbiozones. Superbiozones do not need to be genetically linked in the same way as do the higher-level lithostratigraphic units, but some justification for the designation should be made at the time of the superbiozone's proposal. Biozones may also be subdivided into sub-biozones in order to express finer levels of biostratigraphic detail or identify a biotically distinctive regional grouping of strata. The term zonule is used to refer to a biostratigraphically diagnosable unit that is subordinate to a subbiozone. Finally, individual stratigraphic surfaces characterized by a distinctive biotic component are referred to as biohorizons. A hypothetical example of this biostratigraphic hierarchy is presented in Figure 7.

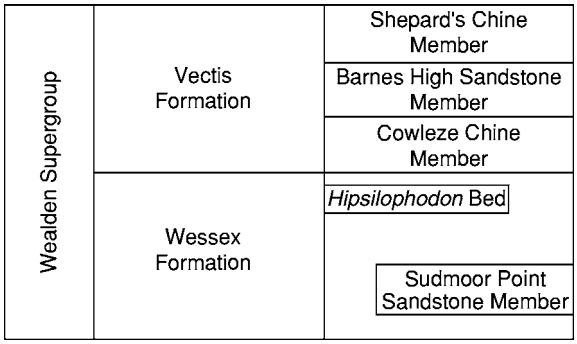


Figure 6 The use of lithostratigraphic units to subdivide a classic Lower Cretaceous suite of non marine sediments in the Wessex Basin of Great Britain. See text for discussion.

Chronostratigraphic and Geochronological Units

Chronostratigraphic units comprise groups of strata that are recognized as having formed during a specific interval of geological time. While chronostratigraphic terms are conceptual rock stratigraphic units, their classification is mirrored by the geochronological or time-stratigraphic classification scheme. To understand the difference between these two scales, consider an hourglass. Sand falling through the neck of the hourglass is deposited in the lower reservoir over a certain time interval (say 1 h). A chronostratigraphic unit is equivalent to the sand deposit, while the associated geochronological unit is equivalent to the amount of time over which the sand deposit accumulated (1 h). The chronostratigraphic unit accumulated over the time interval and can be said to represent that interval in terms of the deposit's thickness and

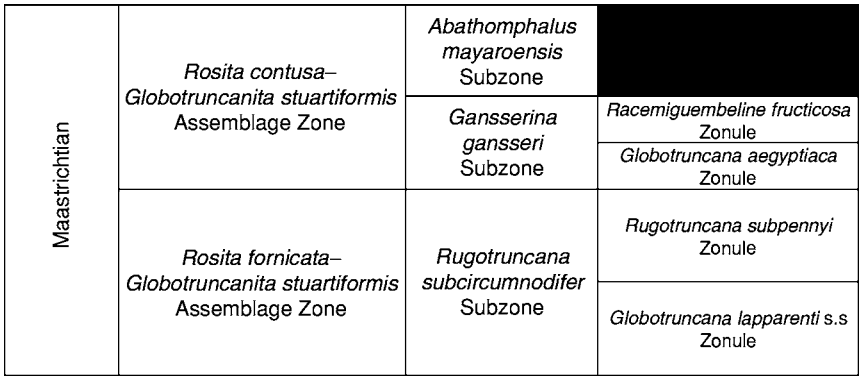


Figure 7 The use of biostratigraphic units to zone a classic Upper Cretaceous suite of deep marine sediments in north central Texas on the basis of their planktonic foraminiferal content. Note the chronostratigraphic series unit (Maastrichtian) and that not all sub biozones are divided into zonules. See text for discussion.

extent. But the sand deposit itself cannot be said to be time. Table 1 lists the chronostratigraphic and geochronometric unit equivalents.

The application of chronostratigraphic unit classification may be illustrated by the chronozone, which is equivalent to a geochronological chron. All stratigraphic intervals represent potential chronozones and chrons, as do all lithostratigraphic and biostratigraphic units. For example, the (hypothetical) *Exus alphus* biozone represents a chronozone that begins with the stratigraphic horizon deposited at the time of the speciation of this (hypothetical) species and ends with the stratigraphic horizon deposited at the time of its global extinction (Figure 8). This chronozone corresponds to the chron, which is defined as the time interval between this species' global speciation and extinction events. Both the chronozones and the chrons are worldwide in extent, though it may not be possible to recognize either in localities remote from the geographical range of the species. The chronozones and chrons will also be estimates (at least for biostratigraphic zones) and are subject to revision.

Stages (equivalent to geochronological 'ages') are the most common chronostratigraphic unit and are usually defined on the basis of the chronozones of a series of biozones (e.g. the Maastrichtian Stage/Age). Note that biozone boundaries themselves cannot be used to achieve a true chronostratigraphic system because they are inherently diachronous (see Figure 8). Stages

Table 1 Nomenclatural equivalents with examples

Chronostratigraphic units	*Geochronological units*	*Example*
Eonathem	Eon	Phanerozoic
Erathem	Era	Mesozoic
System	Period	Cretaceous
Series	Epoch	Upper Cretaceous
Stage	Age	Maastrichtian
Chronozone	Chron	*Belemnella occidentalis* Zone

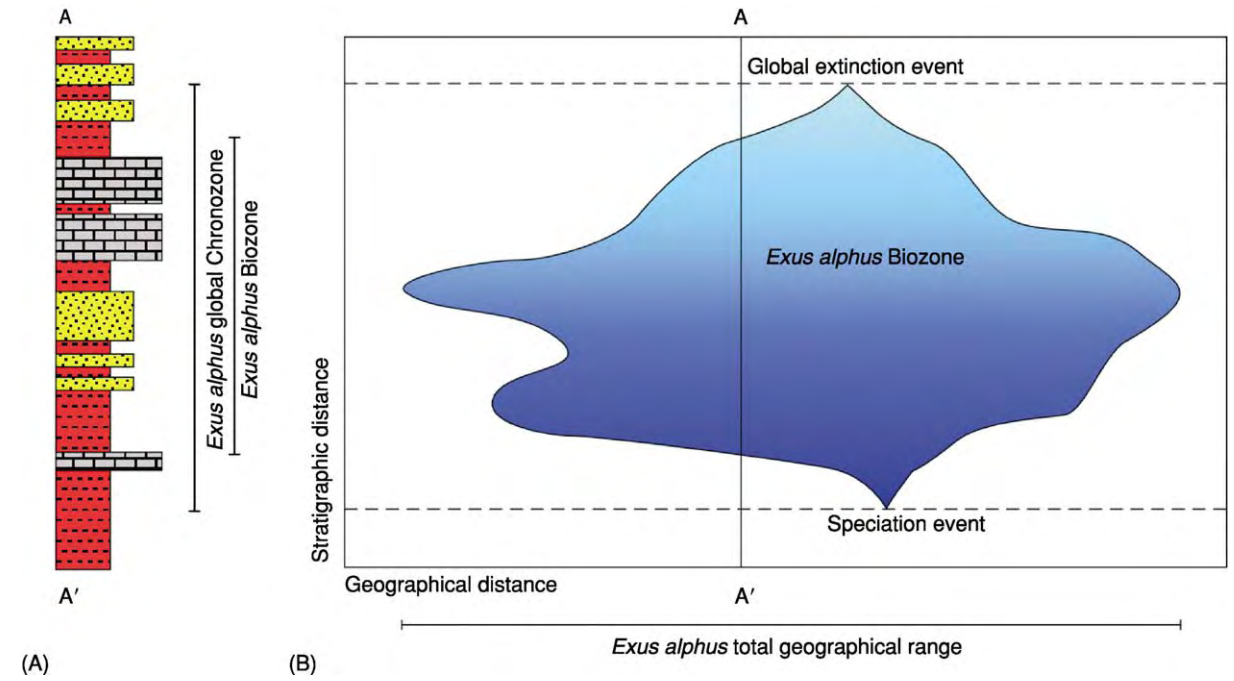


Figure 8 Relation between the (hypothetical) *Exus alphus* Biozone and its corresponding chronozone. (A) Zone expression at the level of a local stratigraphic sequence A A′. (B) Two dimensional slice through the *Exus alphus* Biozone containing the A A′ sequence. Note that the rock stratigraphic expression of the local biozone (which also represents a local chronozone) underestimates the extent of the global chronozones based on the interval between the species' speciation and global extinction horizons. See text for discussion.

may be subdivided into substages. Systems (equivalent to geochronological 'periods') are composed of a sequence of stages. For example, the Induan, Olenekian, Anisian, Laningian, Carnian, Norian, and Rhaetian stages/ages, all of which are defined on the basis of biochronozones, combine to form the Triassic System/Period. Similarly, erathems (equivalent to geochronological 'eras') are composed of a sequence of systems. Three erathems/eras are currently recognized, the Palaeozoic, Mesozoic, and Cenozoic. Finally, eonathems (equivalent to geochronological 'eons') are composed of a sequence of erathems. Thus, the Palaeozoic, Mesozoic, and Cenozoic combine to form the Phanerozoic Eonathem/Eon. This was preceded in geological time successively by the Proterozoic and Archaean eonathems/eons.

There have been several recent proposals (*see* **Time Scale**) to dispense with the dual chronostratigraphic/geochronological classifications of rock units and time units in favour of a single scheme based on the current geochronological classification. Under this scheme, little-used terms such as 'eonathem', 'erathem', 'age' and 'chron' would be considered redundant in favour of their more frequently seen geochronological equivalents. The reasoning behind this proposal does not challenge the logical distinction between the rock-stratigraphic and time-stratigraphic concepts of the dual classification, but instead appeals to the advantages of simplicity, greater ease of imparting the relevant concepts to students, editorial efficiency, and the fact that acceptance of the concept of global stratotype sections and points (GSSP) (see below) by most stratigraphers has rendered – at least to some – the dual rock–time system unnecessary. On the other side of this argument is the simple fact that, by blurring the distinction between rock-stratigraphic and time-stratigraphic units, one is, to some extent, blurring the distinction between observation (rock) and interpretation (time) in chronologically orientated stratigraphic investigations and, in doing so, losing the ability to distinguish between the two. What will become of these proposals remains to be seen.

Other Types of Stratigraphic Unit

With the advent of geophysical methods of analysis, several special types of lithostratigraphic classification have been developed to take advantage of the chronostratigraphic implications of such methods. Perhaps the best example is the study of rock magnetism, which can be used in some lithologies to determine the ancient polarity of the Earth's magnetic field. Based on such observations, a magnetozone can be defined as an interval of strata possessing a characteristic magnetic polarity, either normal or reversed (*see* **Magnetostratigraphy**). These can then be related to time through the use of the chronostratigraphic equivalent of the magnetozone, the magnetochron.

Magnetozones are particularly useful for chronostratigraphic analysis because the time interval over which the Earth's magnetic field changes polarity is short compared with the duration of the magnetozones, biozones, and formations. However, magnetozones can rarely be recognized on the basis of their magnetic properties alone, necessitating the use of other types of stratigraphic analysis – usually biostratigraphy – to achieve the identifications. This increases the complexity of the analysis (and the corresponding chance of error) significantly. Nevertheless, combined magneto-bio-chronostratigraphic analysis has resulted in marked improvements in our understanding of the stratigraphic record. Other types of lithostratigraphic observation that have proven useful in this context include chemical stratigraphy, isotope stratigraphy, seismic stratigraphy, climate stratigraphy, cycle stratigraphy, and orbital stratigraphy.

Stratotypes

With recognition of the distinction between rock stratigraphic units and time stratigraphic units, the ISSP recognized the need to designate 'type-sections' or stratotypes that would constitute standards of reference for various sorts of stratigraphic unit. There are two primary kinds of stratotype: unit stratotypes, which serve as the standard of definition for a stratigraphic unit, and boundary stratotypes, which serve as the standard of definition for a stratigraphic boundary. Unit stratotypes can be either single sections or suites of sections that, when taken together, form a composite unit stratotype. The primary requirement for a stratotype is that it adequately represents the concept of the stratigraphic unit or boundary in all essential particulars. This ideal, however, is rarely met in practice. All real stratigraphic sections exhibit a collection of generalized and idiosyncratic characteristics, and no stratigraphic section can be regarded as truly representative of all other sections and cores worldwide. In addition, disagreements over which section to select as an official ISSP-recognized stratotype have tended to incorporate appeals to historical precedent, priority, and even nationalism, as well as more objective scientific criteria. There is also the ever-present danger that new discoveries might render a designated stratotype incorrect. For example, the boundary at the base of the Cambrian System is defined as the level of the first occurrence of the trace fossil *Treptichnus pedum*, which was thought to occur

2.4 m above the base of Member 2 of the Chapel Island Formation at Fortune Head, Newfoundland, but which subsequent investigations have shown occurs at least 4 m below that horizon in the same section. (Note that, in recognition of the inherently provisional nature of stratigraphic boundary definitions, the ISSP now provides a procedural means for updating boundary stratotype definitions.) Despite these practical deficiencies, the stratotype concept has proven to be popular and has undoubtedly contributed to stabilizing the definitions of stratigraphic units.

One recent modification of the boundary-stratotype concept that has proven to be particularly useful is the 'topless' mode of boundary-stratotype designation. Under this convention, a boundary stratotype designated to serve as the reference for the base of one stage is automatically regarded as defining the top of the underlying stage. This convention elegantly solves the problem of designating unit stratotypes for two successive stages and then finding that the upper boundary of the lower unit and the lower boundary of the upper unit have been placed at different horizons, leading to the artificial production of a stratigraphic gap or overlap.

Conclusion

The principles of stratigraphic analysis were worked out during the nineteenth century. During the twentieth century they were applied at an intercontinental scale and modified to accommodate technological developments that allowed more and different types of geological observations to be employed in stratigraphic correlation. No doubt the former trend will be further refined, and the latter extended, during the twenty-first century. New developments will involve the creation of databases that summarize stratigraphic observations over the Earth's surface (and extending into its subsurface), the development of automated algorithms for comparing the data included in such databases and resolving conflicts between alternative sources of information, and the training of stratigraphers to better appreciate the proper use, strengths, and weaknesses of each source of stratigraphic information so that they may apply the age-old principles of stratigraphy to optimal effect.

Glossary

Angular unconformity A surface of erosion separating lower strata that dip at a different angle from the overlying younger strata.

Biochronozone The associated chronozone of a biozone.

Biostratigraphy The characterization of rock strata by their biological constituents.

Chronostratigraphy The characterization of rock strata by their temporal relations.

Depositional hiatus A horizon within a body of sedimentary rock that represents a gap in time due to the non-deposition of sediment, active erosion, or structural complications.

Diachrony The condition of taking place at different times.

Facies A stratigraphic body distinguished from other such bodies by a difference in appearance or composition.

Geochronology The geological study of absolute time.

Homochrony The condition of taking place at the same time.

Homotaxis The condition of occupying the same position in a sequence.

Isochrony The condition of being created at the same time.

Lithostratigraphy The characterization of rock strata by the kind and/or arrangement of their mineralogical constituents.

Radioisotope An isotope of an element that is capable of changing spontaneously into an isotope of another element by emitting a charged particle from its nucleus.

Stratotype The original or subsequently designated type of a named stratigraphic unit (unit stratotype) or stratigraphic boundary (boundary stratotype).

Stratum (plural strata) A tabular section of a rock body that consists throughout of the same type of rock material.

See Also

Biozones. **Famous Geologists:** Agassiz; Cuvier; Darwin; Hutton; Lyell; Smith; Steno. **Magnetostratigraphy**. **Sequence Stratigraphy**. **Time Scale**. **Unconformities**.

Further Reading

Adams FD (1938) *The Birth and Development of the Geological Sciences*. London: Williams & Wilkins.

Ager DV (1993) *The Nature of the Stratigraphical Record*, 3rd edn. New York: John Wiley & Sons.

Gould SJ (1987) *Time's Arrow, Time's Cycle: Myth and Metaphor in the Discovery of Geological Time*. Cambridge: Harvard University Press.

Hedberg HD (1976) *International Stratigraphic Guide: A Guide to Stratigraphic Classification, Terminology, and Procedure*. New York: John Wiley & Sons.

Rawson PF, Allen PM, Brenchley PJ, *et al.* (2002) *Stratigraphical Procedure*. London: The Geological Society.

Rudwick MJS (1972) *The Meaning of Fossils: Episodes in the History of Palaeontology*. London: MacDonald.
Salvador A (1994) *International Stratigraphic Guide*. Trondheim: International Union of Geological Sciences.
Shaw A (1964) *Time in Stratigraphy*. New York: McGraw Hill.
Zalasiewicz JA, Smith A, Brenchley PJ, *et al.* (2004) Simplifying the stratigraphy of time. *Geology* 32: 1 4.

STROMATOLITES

See **BIOSEDIMENTS AND BIOFILMS**

SUN

See **SOLAR SYSTEM: The Sun**

TECTONICS

Contents

Convergent Plate Boundaries and Accretionary Wedges

G K Westbrook, University of Birmingham, Birmingham, UK

Introduction

At subduction zones, some of the material on the subducting plate is scraped off and added to the leading edge of the overriding plate to form an accretionary wedge (prism, complex), so-called because it is predominantly wedge-shaped in cross-section. The material removed from the subducting plate is primarily sedimentary. Occasionally, igneous crust is transferred to the overriding plate, and when this process occurs at a large scale, it is usually called obduction. Accretionary wedges are analogous to the foreland fold-and-thrust belts of mountain ranges, but they are predominantly submarine and the rates of convergence are typically hundreds of times greater. Where wedges grow large ($\geq$200 km wide and $>$20 km thick), parts of them emerge as islands; examples include Barbados, which is off the Lesser Antilles island arc, and Nias, off Sumatra. Large areas of south-eastern Iran and south-western Pakistan form the Makran accretionary complex, which occupies a belt 240 km across and extends a further 160 km offshore. In the deeper parts of accretionary wedges, the accreted rocks are metamorphosed in a low-temperature/high-pressure environment to produce blueschists, which may be brought to the surface by exhumation following continental collision or other events that halt subduction. Accretionary wedges are not ubiquitous at subduction zones. Their presence is favoured by thick sequences of sediment on the subducting plate and by low rates of subduction. At most subduction zones, low sediment supply, sediment subduction, and tectonic erosion conspire to suppress the formation of accretionary wedges, which are absent or small and transitory.

Wedge Geometry and Fluid Pressure

The geometry of an accretionary wedge is controlled by the shape of the bounding basement surfaces and by the shear stress on the slip surface between the wedge and the subducting plate. If the surface of the crystalline crust of the overriding plate dips seaward, then the landward part of the accretionary wedge overlies it and forms the seaward margin of the fore-arc basin (Figure 1). If the surface of the crystalline crust of the overriding plate dips landward, then the landward part of the accretionary wedge lies beneath it, and a leaf of crystalline basement separates the accretionary wedge from the fore-arc basin.

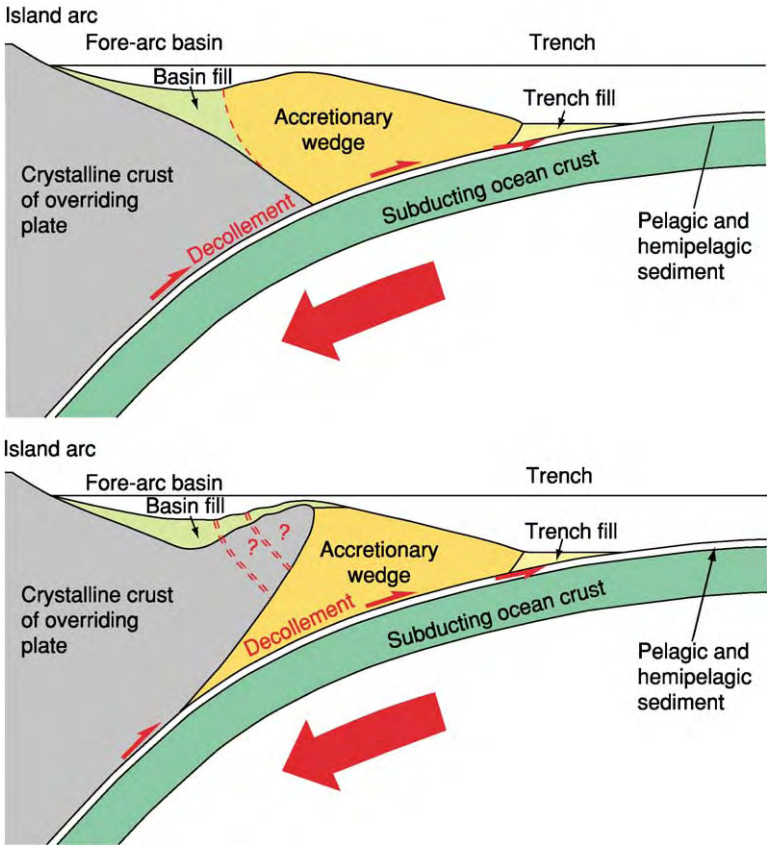


Figure 1 Accretionary wedges are formed at the leading edge of the overriding plate at a subduction zone. They occupy the region between the trench and the fore arc basin, filling the tectonic depression created by subduction, of which the trench is its bathymetric expression. The boundary with the fore arc basin may be provided by a leaf of crystalline crust of the overriding plate (lower), or may be a dynamic boundary (upper) at which fore arc basin sediments lap onto the deformed sediments of the accretionary wedge and may be progressively incorporated into the wedge by deformation as the wedge grows. Copyright Graham Westbrook.

The angle of taper of the wedge (the angle between the dip of the surface and the dip of the basement) depends on the shear stress along the base of the wedge and the strength of the wedge (Figure 2). There is a critical taper at which the force imparted by the basal shear stress is matched by the gravitational spreading force produced by the weight of the wedge. If the basal shear stress is reduced, then the critical taper is reduced, and vice versa. The major factor controlling the strength of the wedge and the shear stress along its base is the frictional resistance, which is the product of the coefficient of friction and the stress normal to any plane of potential movement. The accretion of material to the toe of the wedge lengthens the wedge and changes its taper. In response to this, the wedge deforms internally, thus maintaining its critical taper. So, during accretion, the wedge is continually deforming.

The shear strength of rocks is dependent on the pressure of fluid present in them. The ratio of fluid pressure to the lithostatic pressure (the pressure exerted by the weight of rock) is usually called λ (lambda). Shear failure is governed by the effective stress, which is the difference between the normal stress (across a potential plane of failure) and the fluid pressure. The shear stress, τ, on a plane of motion between two rock masses, such as the decollement at the base of the wedge, is $\tau = \mu\sigma(1 - \lambda)$, where μ is the coefficient of friction, σ is the stress normal to the decollement, and λ is the fluid pressure ratio. The normal stress, σ, is approximately equal to the weight of the sediment in the wedge above the decollement at any particular point. If the load acting on the rock increases more rapidly than the rock can respond by compacting and expelling water, the water bears some of the increased load and becomes overpressured (i.e., its pressure is greater than hydrostatic). If λ exceeds a value of 1, fractures in rocks can be opened by the pressure of water alone (hydrofracturing). Differences in the nature of the rocks in which the decollement is situated affect the stability of the wedge. Clays have low coefficients of friction and low

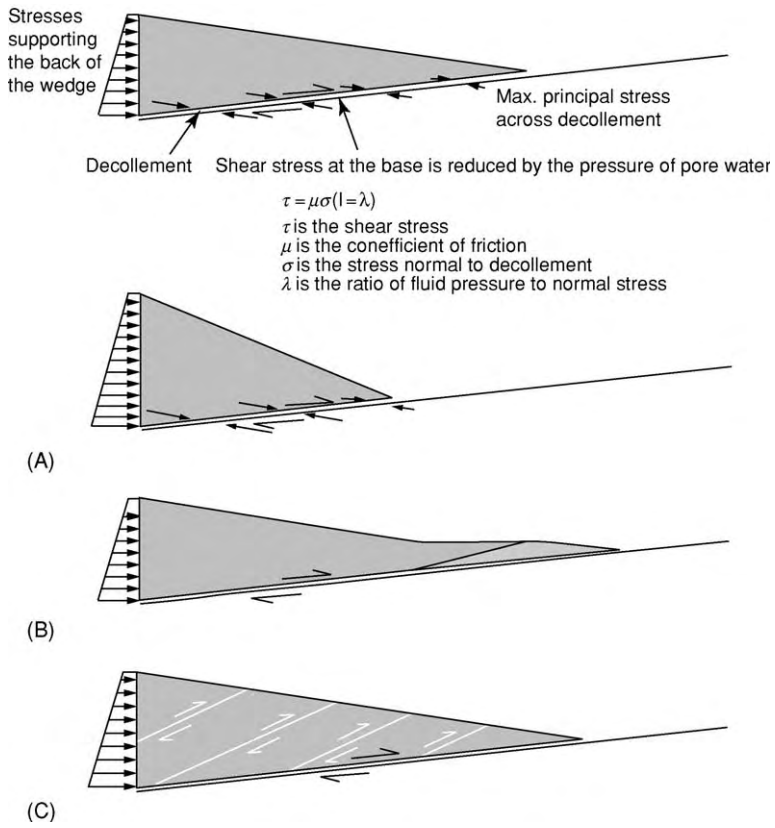


Figure 2 The stability of an accretionary wedge is dependent on the balance between the shear stress along its base, by the motion of the subducting plate, and the stresses generated by gravitational body forces within the wedge. For a wedge with Mohr Coulomb rheology and constant properties, a single of angle of taper between the top and bottom surfaces, termed the 'critical taper', is established. (A) If shear stress along the base of the wedge is increased, the critical taper is increased, and vice versa. (B) Accretion of material to the toe of the wedge increases wedge width, tending to decrease the critical taper, in response to which (C) the wedge thickens by internal deformation, in order to maintain the critical taper. Sedimentation onto the wedge, erosion, and subcretion (accretion of material to the base of the wedge) similarly induce internal deformation to maintain the critical taper. Copyright Graham Westbrook.

permeability; this favours the build-up of high fluid pressure. Both of these factors reduce the shear stress, leading to low angles of taper. Sands, which are more permeable and have a higher coefficient of friction, would produce a relatively high angle of taper. Several accretionary wedges have angles of taper that are less than 5° and values of the fluid pressure ratio, λ, that are greater than 0.9. The stratigraphy of the trench and minimization of the work required to move the wedge over the subducting plate favour the formation of decollements in clay-rich formations.

At subduction zones, sediments can be become overpressured in three ways:

1. In the trench, on the subducting oceanic plate, fine-grained pelagic and hemipelagic sediments of low permeability become overpressured by the rapid overlying deposition of trench-fill sediment (Figure 3). Deposition rates in trenches, which are as much as a few kilometres per million years, are among the fastest in the world. The greatest degree of overpressure is developed near, but not at, the top of the low-permeability sediment (Figure 4).
2. Sediment accreted to the wedge becomes overpressured by the tectonic thickening of the wedge. The sediment at the base of the wedge is most prone to becoming overpressured, because it has the longest drainage paths and also because low-permeability mud predominates in the lower part of the accreted section.
3. Sediment carried beneath the wedge on the subducting plate becomes overpressured by the weight of the increasing thickness of the wedge above as it passes below. This produces the most rapid increase in load and the highest overpressure.

Wedge Growth

The accretionary wedge grows in a number of ways: by frontal accretion, subcretion, migration into a fore-arc basin, and deposition and deformation of

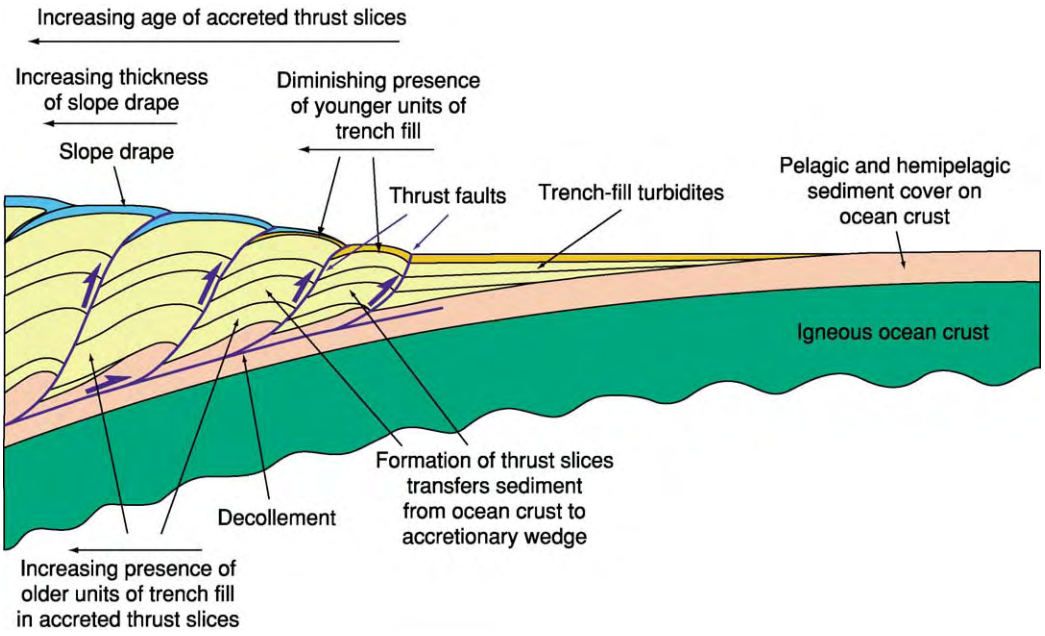


Figure 3 The process of accretion at the toe of an accretionary wedge. Commonly, but not invariably, the turbidites deposited in the trench form most of the accreted material. The accreted materials separate from the overlain pelagic and hemipelagic muddy sediments along a detachment surface (decollement) in the upper part of the sediment cover; the decollement is where loading from turbidite deposition has increased pore fluid pressure to a value nearing that of the lithostatic pressure imposed by the weight of the overlying sediments, weakening them and making them prone to failure by shearing. The thrust faults that detach the individual thrust slices that form the accreted section originate in the decollement, which propagates ahead of the toe of wedge. The accreted section always includes the youngest turbidites; these are deposited in the trench but not on older accreted thrust slices, which have been uplifted out of the zone of deposition. Consequently, the stratigraphy of each successively accreted thrust slice contains younger turbidites. Within each thrust slice, the sediments upward and landward, but the overall stratigraphy of the wedge are younger becomes younger seaward. Superimposed on the accreted sediments of the wedge is a drape of hemipelagic sediment, undiluted by turbidites, and the age of the base of this drape is youngest seaward. The drape is also deformed by the deformation of the wedge as it thickens, with the oldest part of the drape sequence being more deformed than the youngest. Copyright Graham Westbrook.

slope-drape and slope-basin sediments. In frontal accretion, thrusts propagating from a decollement in a weak, overpressured horizon at the toe of the wedge divide the overlying section into thrust slices, which become added to the toe of the wedge (Figure 3). The level of the decollement is commonly in the upper part of the pelagic–hemipelagic sequence on the subducting plate, which has been overpressured by the deposition of turbidites above it in the trench. The age of the accreted sediment changes with time, giving the wedge a characteristic tectonostratigraphy. Each thrust slice is youngest upward and landward, but the sequence of successively accreted thrust slices has the youngest thrust being seaward and downward. It is this characteristic stratigraphy that can be used to identify ancient accretionary wedges, such as the Ordovician–Silurian wedge of the Southern Uplands of Scotland.

In wedge growth by subcretion, sediment is added to the base of the accretionary wedge by the formation of duplexes at a ramp where the decollement changes level. These propagate successively forward because the work required to continue to move the wedge up a ramp becomes greater than that needed to propagate displacement along the lower decollement and generate a new ramp. In the process, the energy in the sediment between the ramps is transferred from the subducting plate to the accretionary wedge (Figure 5). It has also been suggested that the formation of a zone of tectonic melange along the decollement enables material from the subducting plate to be added to the accretionary wedge, but this can also operate in the opposite sense.

Accretionary wedge growth can occur when landward force imparted by the subducting lithosphere increases with increases in wedge width, pushing the wedge backward into the fore-arc basin and forming thrusts that incorporate fore-arc basin sediment into the wedge (Figure 5). In the mechanism involving slope-drape and slope-basin sediments, deposits directly onto the wedge are deformed by the continual deformation of the wedge beneath as it strives to maintain its critical taper. The sediment forming the slope drape is usually hemipelagic, but in some cases,

turbidites derived from erosion of the fore-arc basin or upper part of the wedge are deposited in synclinal troughs to form slope basins, so-called because they occur on the inner trench slope. In foreland fold-and-thrust belts, basins such as these are termed 'piggyback' basins. In the southern part of the Barbados accretionary complex, distributary channels of the Orinoco submarine fan run along synclinal valleys in the accretionary complex, as well as on the ocean floor before it (Figure 5).

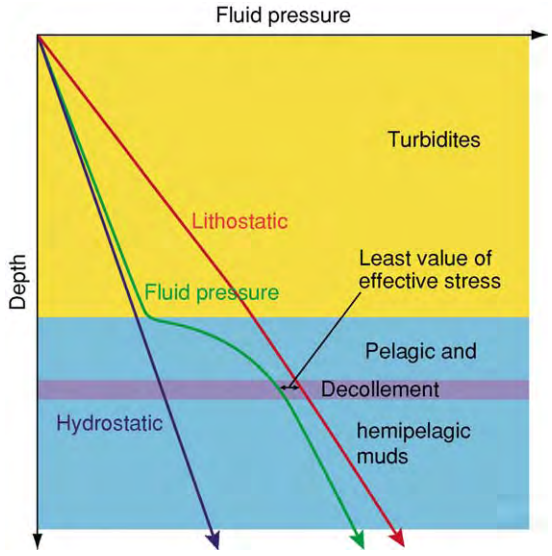


Figure 4 Fluid pressure varies with depth within the sediment in a trench, prior to accretion or subduction. Within the relatively permeable, rapidly deposited, turbidite fill of the trench, fluid pressure is only a little above hydrostatic. Within the low permeability pelagic and hemipelagic sediments, fluid pressure increases towards lithostatic pressure in response to the loading produced by the rapid deposition of the turbidites above the sediments. The depth at which fluid pressure is nearest to lithostatic pressure is favoured for development of the decollement, although local lithologically mediated variations in the coefficient of friction and cohesion may control its actual position. Copyright Graham Westbrook.

Thickening of an accretionary wedge in its frontal part is brought about by continued displacement on the thrusts by which sediment was accreted and by general horizontal shortening, which produces folding of the thrust slices and landward rotation and steepening of structures. The amount of thickening that this can produce is limited by the rotation of the thrust faults away from the optimum angle for thrusting. Consequently, thickening in the more landward parts of the wedge is produced by motion on new, out-of-sequence, thrusts (termed 'out of sequence' because they do not follow the normal sequence of the youngest thrust being the most seaward). Thickening is also produced by subcretion, adding material from below.

The size of an accretionary wedge might be expected to be simply a product of the thickness of sediment on the subducting oceanic crust, the rate of subduction, and the period over which subduction has occurred. It is certainly the case that the really large wedges occur where major submarine fan systems of considerable thickness are being subducted (Figure 6), but the correlation between wedge size and subduction rate is negative. The very large accretionary wedges occur at subduction zones with low rates of convergence. The frictional resistance to motion along the bases of these wide wedges (which, even with a fluid pressure ratio of $\lambda = 0.9$, is a few teranewtons per square metre, per unit length of the

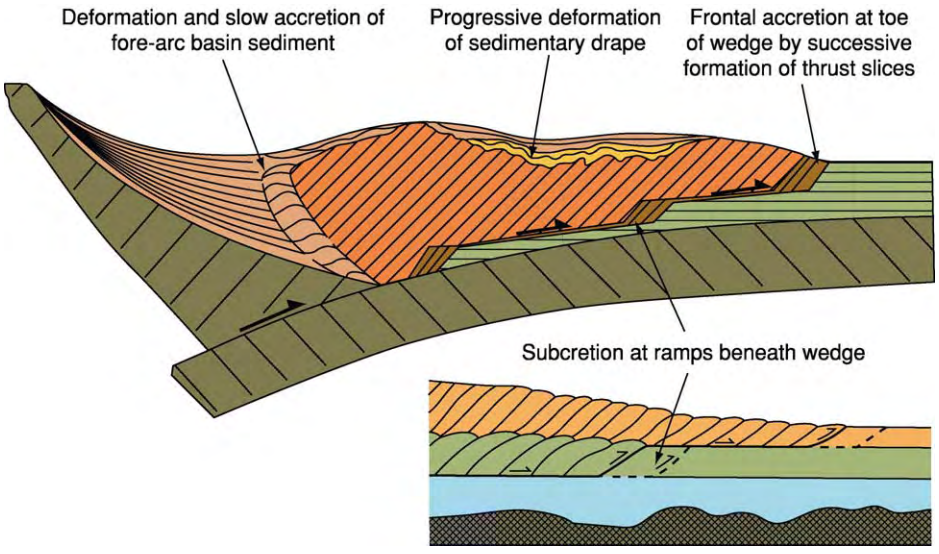


Figure 5 Different modes of accretion of sediment to an accretionary wedge. Copyright Graham Westbrook.

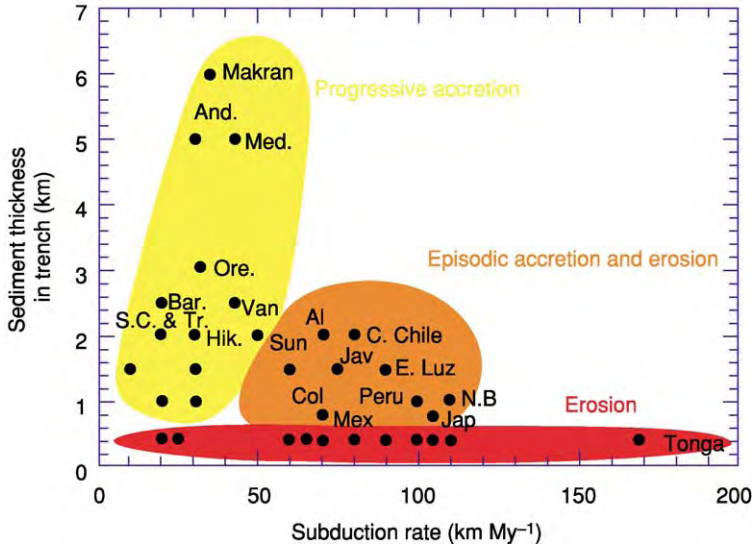


Figure 6 Thickness of sediment on a subducting plate in the trench at a subduction zone, plotted against the rate of subduction. Three discrete groups of subduction zones are identified worldwide: those that have large accretionary wedges with a history of near continuous growth (yellow), those that have accretionary wedges of small or moderate size in relation to their budget of sediment input and history and that have undergone episodes of tectonic erosion (orange shading), and those that have no, or insignificant, accretionary wedges and are dominated by tectonic erosion (red shading). The last category is typified by very thin sediment fill in the trench, irrespective of subduction rate, and tectonic erosion by basement relief is predominant. Copyright Graham Westbrook.

plate boundary, comparable to the net plate driving forces from ridge push and subduction pull) may be a contributory cause of the low convergence rates. A crucial aspect of the accretion process is the separation of accreted from subducted sediment. Off the coast of Costa Rica, all of the sediment in the trench, which has no cover of turbidites, is subducted. Lower permeability and a higher rate of loading, which correlates with faster subduction, develop higher fluid pressure in the sediments, which favours their subduction because the decollement surface can form at a shallow level. Studies of the global budget of material accreted at convergent plate boundaries, in comparison with material supplied by subducting plate and sedimentation in the trench, have shown that there is a net loss of material. Globally, sediment subduction and tectonic erosion (see below) predominate over accretion.

Fluid Flow, Seeps, and Methane Hydrate

As the sediment incorporated into an accretionary wedge and subducted beneath the wedge compacts, it expels water. Also, the dehydration of minerals in the sediment with increasing temperature and pressure, such the transformation of smectite to illite, releases water. This water flows through the wedge and subducting crust as it escapes to the ocean, carrying with it heat, solutes, and gases (Figure 7). The permeability of the mud-rich sediments that are subducted is generally too low for the water to escape along the subduction zone. The rate of flow of the water through the sediment is less than the rate of subduction. So, the water migrates into and then through high-permeability pathways, such as faults and permeable sediments such as sands, or even the igneous ocean crust, which is about a thousand times more permeable than compacted clay-rich mud deposited on the ocean floor. The warm water expelled beneath the trench increases the heat flow from the trench. Water driven into trench sediments or expelled along faults through the wedge can contain methane that was generated by methanogenic bacteria and other hydrocarbon gases that were created in conditions of higher temperature beneath the wedge. The methane and hydrogen sulphide expelled at seeps and vents sustain chemosynthetic communities of biota. (In this setting, the hydrogen sulphide is a product of anaerobic oxidation of methane by symbiotic communities of aquatic bacteria and archaea.) The seeps and vents are usually located along faults or in mud volcanoes, which occur on the accretionary wedge and on the ocean floor in front of the wedge. Mud volcanoes are created by mud diapirism, in which the tensional stresses at the top of a body of low-density mud are sufficient to create a pathway for the mud to rise buoyantly to the seabed, where it erupts. The diapiric bodies are initially created by deformation of underconsolidated

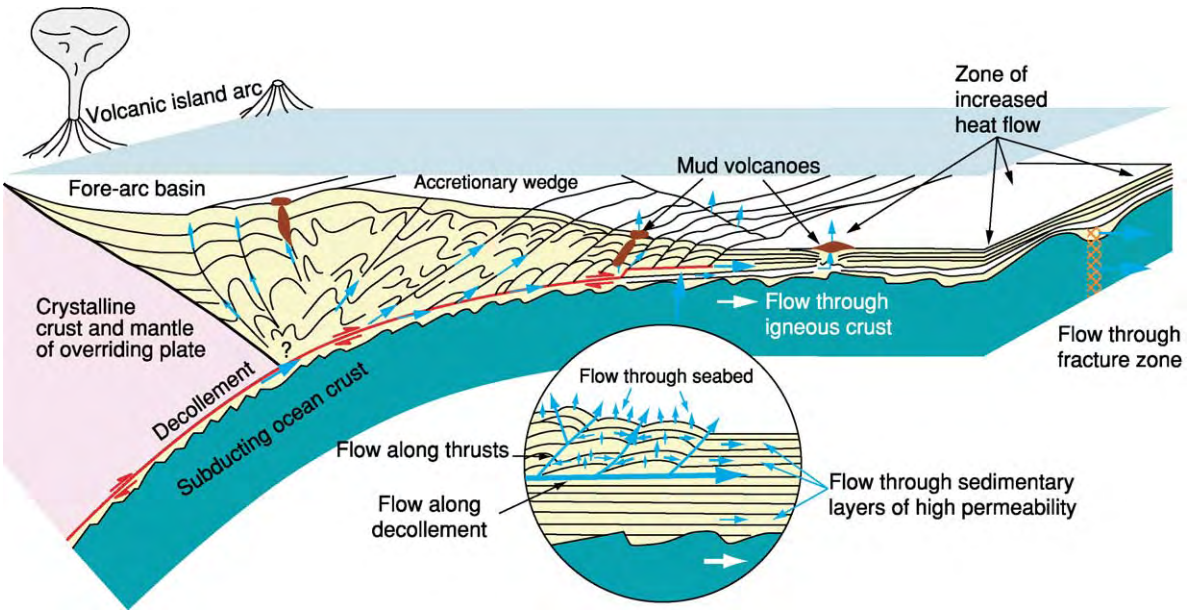


Figure 7 Fluid flow and modes of expulsion of fluid from an accretionary wedge. The sources of fluid are pore water expelled by compaction from sediment subducted beneath the accretionary wedge and sediment accreted to the wedge, and dehydration of hydrous minerals such as smectite, as temperature increases with increasing depth of subduction. Copyright Graham Westbrook.

mud-rich layers that are accreted into the wedge or subcreted beneath it. Mud volcanoes also appear to be created by the fluidization and entrainment of mud by water driven to the surface by tectonic expulsion, and this mode of formation is characteristic of mud volcanoes created in front of accretionary wedges on the ocean floor, or behind them in fore-arc basins.

The migration of methane-containing pore water through the sediments of an accretionary wedge as it compacts creates methane hydrate in the sediments occupying the first few hundred metres depth range beneath the seabed (*see* **Petroleum Geology:** Gas Hydrates). This occurs because this region lies within the stability field for methane hydrate (which is a solid clathrate formed from water and methane, in which the methane molecules are held within a cage of water molecules in an approximately 1:6 ratio). The hydrate stability field generally exists in Earth's major oceans in water depths greater than about 300 m, and is favoured by increasing pressure and decreasing temperature. Consequently, most of the sediments beneath continental margins are in the hydrate stability field down to the depth at which, because of the increase of temperature with depth, the geotherm crosses the stability boundary for hydrate. Beneath this boundary, methane can be present as free gas, in which case the boundary creates a seismic reflection because the presence of only a very small amount of free gas (less than 1% of the pore space is enough) reduces the seismic velocity of P waves significantly. The polarity of the reflection is negative, opposite to that of the seabed, because of the decrease in velocity beneath it, and is most clearly visible on seismic reflection sections where it cuts across the reflections produced by sedimentary bedding (Figure 8). This reflection is widespread in accretionary wedges, and is usually termed a bottom-simulating reflection (BSR) because its shape, to the first order, mimics that of the seabed, which, because of its nearly uniform temperature, controls the shape of the isotherms beneath it.

Uplift of the seabed produced by the thickening of the wedge continually moves the base of the zone containing hydrate upward out of hydrate stability field, causing hydrate to dissociate and release free gas that produces the BSR. Because the depth of the BSR below the seabed is controlled by the geothermal gradient, mapping the depth of the BSR has been used to map variations in heat flow from accretionary wedges, which is influenced by tectonic thickening and fluid flow. The tectonic expulsion of methane-rich pore water and the dissociation of hydrate to free gas caused by uplift results in methane hydrate and BSRs being widespread in accretionary wedges, whereas they occur only rarely in the sediments of passive continental margins.

Tectonic Erosion at Subduction Zones

The inner walls of trenches of arcs (e.g., the Mariana arc, Tonga arc, and South Sandwich arc) do not have significant accretionary wedges. Those that do occur

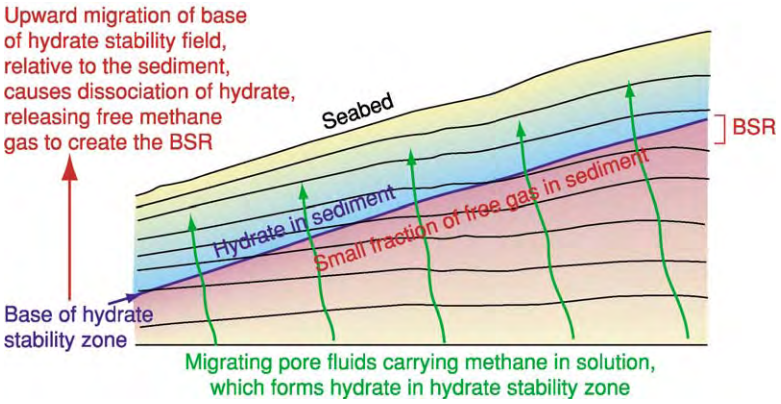


Figure 8 Model for the growth of methane hydrate within sediments in the uppermost part of an accretionary wedge, and for the formation of a bottom simulating seismic reflection (BSR). Methane, dissolved in pore water expelled from sediment by compaction within and beneath the wedge, is carried upward into the hydrate stability field, where it forms hydrate. Uplift of the seabed, caused by the thickening of the wedge with continuing growth, destabilizes the base of the hydrate stability zone, releasing methane gas and water from dissociated hydrate. The presence of free gas is the principal cause of the BSR. Copyright Graham Westbrook.

are very small and transitory. Most of the evidence suggests that material from the inner trench wall is being removed and subducted. The budgets of sediment available to be added to accretionary wedges in comparison with the mass of sediment in wedges indicate that many have lost sediment or have undergone periods when none was accreted. For example, the accretionary wedge off the coast of Honshu, Japan, is composed mainly of Cretaceous sediment, with only a very little Neogene sediment and no Paleogene sediment. This suggests that there is a process that removes material from the overriding plate at subduction zones. This process, or group of processes, is referred to collectively as tectonic erosion or subduction erosion.

What is the evidence for tectonic erosion? The absence of accretionary wedges from arcs such as Tonga might be explained if all of the sediment in the trench was subducted. Fore-arc subsidence such as that off the coast of Peru, where there has been 4 km of subsidence over a 100-km width of fore arc since the late Miocene, is difficult to explain other than by the removal of material from the base of the fore arc. The landward migration of the volcanic arc across the overriding plate, as exemplified by the Andes and by island arcs such as the South Sandwich or the northern end of the Lesser Antilles, although explicable in specific instances by a reduction in the angle of dip of the subducted lithosphere, can only be generally explained by removal of crust from the leading edge of the overriding plate. The mechanisms proposed for tectonic erosion are of two principal types: erosion by basement topography and erosion from the effects of high fluid pressure.

Erosion by Basement Topography

Tectonic erosion may occur when the basement topography physically breaks off and displaces parts of the fore arc, carrying it deeper into the subduction zone; this steepens the trench slope locally, causing submarine slides into the trench. The material from the slides may then also be subducted (see **Figure 9**). The mechanism depends on the basement of the subducting plate being stronger than the material in the tip of the overriding plate. Where an accretionary wedge, composed of sediment or metamorphosed sediment, forms the leading edge of the overriding plate, this is normally the case, but where the edge of the overriding plate is composed of igneous or high-grade metamorphic rocks, the situation can be reversed, resulting in the accretion/obduction of the basement feature on the subducting plate. There are two broad categories of basement topography. The first is general basement relief inherited from a mid-ocean ridge and accentuated by normal faulting in the outer trench slope. This is ineffective if sediment cover is more than several hundred metres thick, because the decollement forms well above the basement and the accretionary wedge rides over the relief, undisturbed. The second category comprises discrete features of high relief, such as seamounts, transform ridges and troughs, and hotspot ridges. These are more severe in their effects, but are not present everywhere, and so they produce spatially and temporally limited episodes of tectonic erosion, of which a good example is provided by the subduction of seamounts on the Cocos plate beneath the convergent margin of Costa Rica.

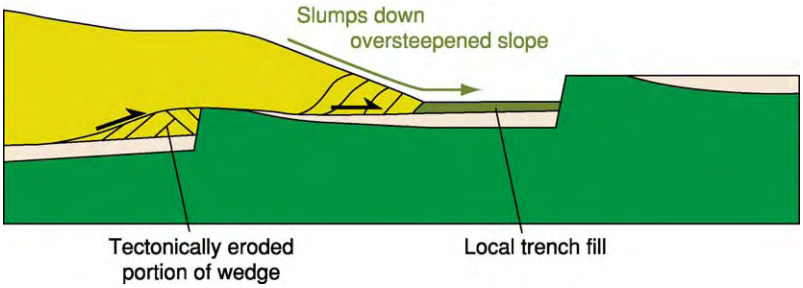


Figure 9 Tectonic erosion by the relief of the oceanic basement is a common process where sediment cover on the subducting plate is thin. Scarps originally formed by normal faulting at the mid ocean ridge are enhanced, or a new one is created by normal faulting in the outer part of the trench as the oceanic lithosphere flexes downward into the subduction zone. The tops of these scarps collide with the toe of the accretionary wedge or with the leading edge of the crystalline fore arc of the overriding plate, forcing the decollement to jump upward to the level of the top of the scarp, thereby transferring some of the material that was in the wedge to the subducting plate. The surface of the wedge is steepened by the removal of material from its toe, inducing slumps into the trench of material that is subsequently accreted and swept into the subduction zone by the basement scarps. The subduction of seamounts produces a similar, although more severe, process of tectonic erosion that is effective in thicker sediment cover, but is geographically more localized. Copyright Graham Westbrook.

Erosion from the Effects of High Fluid Pressure

The need to invoke this second type of mechanism is presented by those subduction zones where it is clear that tectonic erosion is or has been active, yet no features of basement topography appear to be responsible. The fluid pressure ratio, λ, is greatest at the top of any body of rock in which pore water is connected. This makes the uppermost rocks weakest and most liable to failure and displacement. As zones of high pressure are driven to migrate upward, progressive failure and displacement remove material from (tectonically erode) the section through which the high-pressure pore fluid migrates, until it dissipates or escapes to the surface. As subduction is continually feeding sediments with a high water content beneath wedges of accreted sediment or the basement of the overriding plate, the potential for this type of mechanism to operate is always present, if pore water expelled from the sediment cannot escape easily through the overlying wedge. There is evidence for two possible variants of this process, the first more general, the second more specific:

1. In 'stoping', high-pressure fluid causes disaggregation of the base of the wedge. The disaggregated rock is incorporated into a shear zone melange and is subducted (Figure 10). Shear zone melanges are exhibited by the exhumed deeper parts of old accretionary wedges that can be observed on land, such as the Franciscan wedge in California.
2. In the process of reactivation and upward migration of detachment surfaces, high-pressure fluid weakens the upper surfaces more than the lower ones. Rock beneath the reactivated surfaces is subducted (Figure 10). Seismic images of duplexes within the subducting section have provided the indication that this process occurs.

Obduction

Sometimes, features of oceanic basement are not subducted, but are sheared off against the crystalline crust of the overriding plate, leading to obduction of part of the oceanic crust. This appears to occur most commonly where asperities exist in the oceanic crust that is being subducted, such as seamounts or the ridges flanking transform faults. The Tres Montes Peninsula on the coast of southern Chile, close to the Chile triple junction, was probably brought about by this process. Very large-scale obduction of ocean crust to create ophiolite complexes is associated with continent/continent collision or continent/island-arc collision and may also be a consequence of the closing of back-arc basins following a change in plate motions, with young buoyant ocean crust thrust onto the adjacent plate, as exemplified by the Rocas Verdes complex of southern Chile.

Oblique Subduction

There are several convergent margins, including the eastern end of the Aleutian island arc, Sumatra, the north-west United States (Cascadia), and the northern Pacific margin of Colombia, where the direction of convergence is oblique; in these cases, as well as convergence between the two plates in the direction orthogonal to the plate boundary, there is a component of motion parallel to the plate boundary. At these margins, the directions of thrust-fault outcrops and of fold axes, which form at right angles to the direction of compression, run subparallel to the margin, not at right angles to the direction of convergence of the two plates. The reason for this is that the displacement between the accretionary wedge and the subducting plate is partitioned into a component orthogonal to

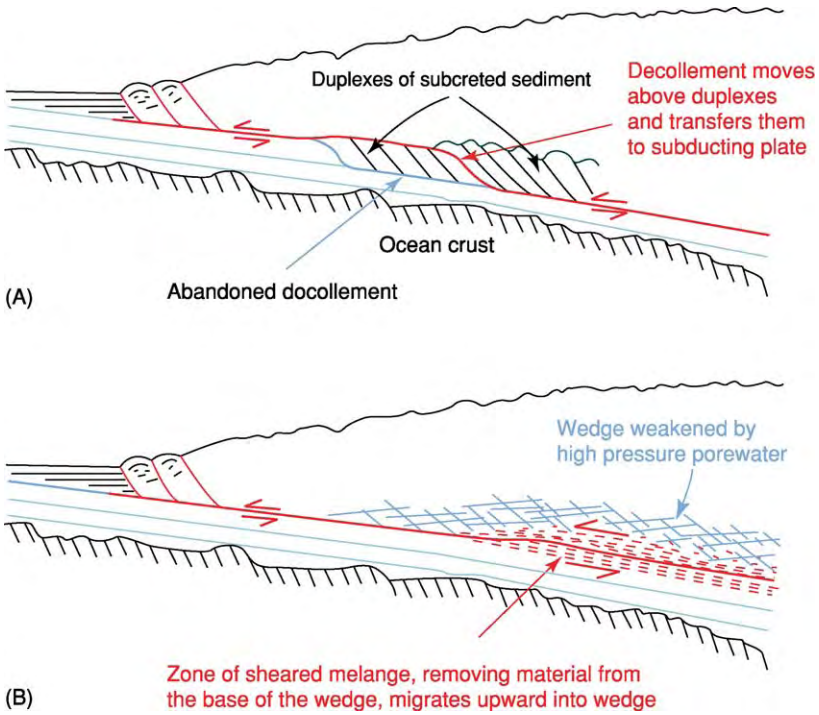


Figure 10 The hypothesis of tectonic erosion from the effect of high pore fluid pressure. (A) Loss of subcreted material from the accretionary wedge by relocation of the active decollement to the upper of the two levels separating the subcreted duplexes, because of an increase in drainage from the lower decollement and/or a restriction in the drainage from the upper level. (B) Material above the main decollement has been weakened by high fluid pressure; a melange of this disaggregated material is incorporated into a shear zone, which transports the material deeper into the subduction zone. Removal of this material makes it easier for high pressure fluid in the shear zone to infiltrate the zone and propagate the process of tectonic erosion upward into the overlying wedge. Copyright Graham Westbrook.

the margin and a component parallel to it. A strike–slip fault, or series of faults, separates the wedge, and often the fore-arc basin as well, from the volcanic arc and the rest of the plate. The horizontal motion between the fore-arc region and the remainder of the plate along this fault compensates the obliquity of plate convergence, so that the relative motion between the subducting plate and the accretionary wedge is nearly orthogonal to the margin (Figure 11). It has been demonstrated theoretically and by experiments with sand-box models that, when the direction of convergence is greater than about 15° from orthogonal to the margin, displacement can be partitioned in this way. This is because the work required to move the subducting plate the shortest distance, orthogonal to the margin, against the frictional resistance of the decollement surface, plus the work to move the fore-arc along the strike–slip fault, is less than the work required to move the plate obliquely a greater distance along the decollement in the direction of convergence between the two plates. The angle of obliquity at which partitioning occurs depends on the shear stress at the base of the wedge, and if this is very low, because of fluid overpressure,

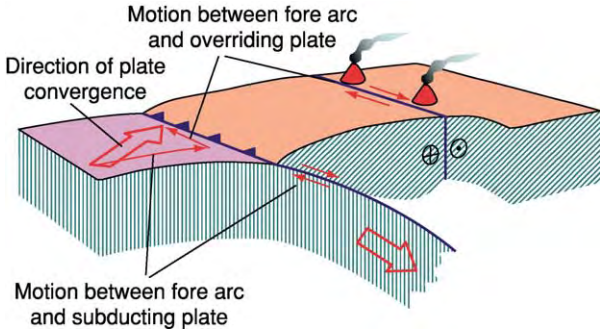


Figure 11 Where subduction is oblique to the margin, motion is partitioned into orthogonal underthrusting of the subducting plate beneath the fore arc (accretionary wedge plus or minus the fore arc basin) and strike slip movement between the fore arc and the overriding plate. The direction of plate motion is shown by the large arrows; the small red arrows on the subducting plate indi cate components of motion normal and parallel to the margin. Copyright Graham Westbrook.

the deviation from orthogonal convergence required to produce partitioning can approach 40°. Where partitioning occurs, the fore-arc can be translated large distances along the edge of the overriding plate

Table 1 Summary of controls on accretion, subduction, and tectonic erosion of sediment

Separation of accreted and subducted sediment	The level of detachment between accreted and subducted sediment is governed by Ratio of shear stress to effective normal stress Presence of a weak horizon, produced by Low intrinsic strength (low coefficient of friction and low cohesion) High fluid pressure from Low permeability High rate of loading, produced by Sufficient sediment supply to trench High subduction rate High angle of dip of subducting plate
Tectonic erosion	By basement topography General relief inherited from mid ocean ridge, and accentuated by normal faulting in outer trench slope; ineffective if sediment cover is thicker than several hundred metres Discrete features of high relief Seamounts Transform ridges and troughs Hotspot ridges (swells) From the effects of high fluid pressure 'Stoping' (disaggregation of base of wedge and incorporation into a shear zone melange) Reactivation and upward migration of detachment surfaces
Accretion vs. tectonic erosion	Progressive accretion is associated with high sediment thickness and low subduction rate Tectonic erosion is associated with low sediment thickness Episodic accretion and tectonic erosion is associated with high sediment thickness and high subduction rate

to become a displaced, or exotic, terrain. A summary of the controls on accretion, subduction, and tectonic erosion of sediment is given in Table 1.

See Also

Andes. **Europe:** Caledonides of Britain and Ireland; Mediterranean Tectonics. **Japan**. **Metamorphic Rocks:** Facies and Zones. **New Zealand**. **North America:** Northern Cordillera. **Oceania (Including Fiji, PNG and Solomons)**. **Petroleum Geology:** Gas Hydrates. **Seamounts**. **Sedimentary Processes:** Deep Water Processes and Deposits. **Seismic Surveys**. **Tectonics:** Mountain Building and Orogeny; Ocean Trenches. **Ultra High Pressure Metamorphism**.

Further Reading

Bebout GE, Scholl DW, Kirby SH, and Platt JP (eds.) (1996) *Subduction Top to Bottom, Geophysical Monograph 96*. Washington, DC: American Geophysical Union.

Carson B and Screaton EJ (1998) Fluid flow in accretionary prisms: evidence for focused, time variable discharge. *Reviews of Geophysics* 36: 329 351.

Davis DJ, Suppe J, and Dahlen FA (1983) Mechanics of fold and thrust belts and accretionary wedges. *Journal of Geophysical Research* 88: 1153 1172.

Fryer P (1996) Evolution of the Mariana convergent plate margin system. *Reviews of Geophysics* 34: 89 125.

Hyndman RD and Davis EE (1992) A mechanism for the formation of methane hydrate and sea floor bottom simulating reflectors by vertical fluid expulsion. *Journal of Geophysical Research* 97: 7025 7041.

Hyndman RD, Spence GD, Chapman R, Riedel M, and Edwards RN (2001) Geophysical studies of marine gas hydrate in northern Cascadia. In: *Geophysical Monograph 124, Natural Gas Hydrates: Occurrence, Distribution, and Detection*, pp. 273 295. Washington, DC: American Geophysical Union.

Lallemand SE, Schnurle P, and Malavieille J (1994) Coulomb theory applied to accretionary and nonaccretionary wedges possible causes for tectonic erosion and or frontal accretion. *Journal of Geophysical Research* 99: 12033 12055.

Moore JC and Vrolijk P (1992) Fluids in accretionary prisms. *Reviews of Geophysics* 30: 113 135.

Ranero CR and von Huene R (2000) Subduction erosion along the Middle America convergent margin. *Nature* 404: 748 752.

Stern RJ (2002) Subduction Zones. *Reviews of Geophysics* 40(4): 3.1 3.38.

Tarney J, Pickering KT, Knipe RJ, and Dewey JF (eds.) (1991) *The Behaviour and Influence of Fluids in Subduction Zones*. London: The Royal Society (From *Philosophical Transactions of the Royal Society, Series A* 335: 225 418.)

von Huene R and Scholl DW (1991) Observations at convergent margins concerning sediment subduction, subduction erosion, and the growth of continental crust. *Reviews of Geophysics* 29: 279 316.

Westbrook GK, Ladd JW, Buhl P, Tiley GJ, and Bangs N (1988) Cross section of an accretionary wedge: Barbados Ridge Complex. *Geology* 16: 631 635.

Earthquakes

G J H McCall, Cirencester, Gloucester, UK

Introduction

Earthquakes have many and diverse relationships with other Earth processes, and their study has a wide range of possible applications. Earthquakes will be considered here under the following headings:

1. The nature of earthquakes,
2. The importance of seismological records,
3. The global distribution of earthquakes, and
4. Earthquakes as a hazard – tectonic, volcanic, and man-made earthquakes.

The Nature of Earthquakes

An earthquake is a sudden movement of the Earth's surface, caused by a release of strain built up over long periods on faults. The rocks are elastic and can store energy in the same way as a compressed spring. Earthquakes are focused on faults in the rock mass. Most have foci within the crust but a few, in plate boundary zones and beneath stable cratonic areas (where they are related to events in nearby subduction zones), have foci at great depths, down to about 700 km, in the mantle; beyond this depth the rock mass is insufficiently rigid to rupture. The very deep earthquakes are not well understood. Most earthquakes have foci less than 30 km deep. Not all the built-up strain is relieved by earthquakes; much of it is relieved by continual small adjustments, a process of creep. However, where friction prevents such accommodation, the strain builds up until something has to give, and there is a sudden rupture of the weakest part of the solid rock, the forces being accommodated by sudden dislocation of the rocks on either side of the fault plane (Figure 1). This process can happen on all three types of fault: normal, reverse, and transcurrent. The point directly above the focus is the called the epicentre; here, the effects of the earthquake will be greatest. If the focus is shallow, the effects will be greater than if it is deep – the 1960 Agadir earthquake, Morocco, was not of great magnitude, but it was very shallow and the epicentre was right under the city.

The magnitude is a measure of the amount of energy released by the earthquake. It is calculated from the size or amplitude of the waves traced by the pen of a seismograph, an instrument that picks up the waves at some distance from the epicentre and records them in the form of a wavy trace on a rotating drum (Figure 2) coupled to a clock. The principle of the seismograph is illustrated by a chandelier that swung in the great Lisbon earthquake of 1755 – a freely pivoting horizontal strut is attached to an upright support (Figure 3). A heavy mass at the strut end is attached to the pen, which traces a continuous line on the paper. Most of the time there are no ground movements, so the trace is horizontal. All seismographs have to be standardized so that valid comparisons can be made between their traces. Waves weaken as they travel outwards from the earthquake focus, so allowance has to be made for distance between the focus and the instrument (this can be calculated by measuring the difference between the arrival times of P and S waves). Three seismographs are required to measure north–south and east–west horizontal movements and vertical movements. The Worldwide Standardized Seismograph Network was established in 1962. All the instruments are standardized as if they were situated 100 km away from the focus.

Waves produced by earthquakes spread through the Earth (Figure 4). They comprise body waves and surface waves. Body waves travel through the Earth and are of two types, primary (P) and secondary (S)

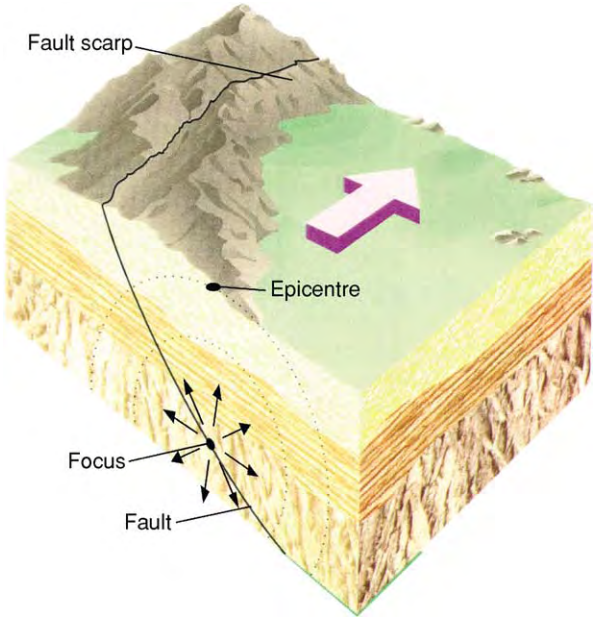


Figure 1 Block diagram showing the relationship between an earthquake focus, epicentre, and fault.

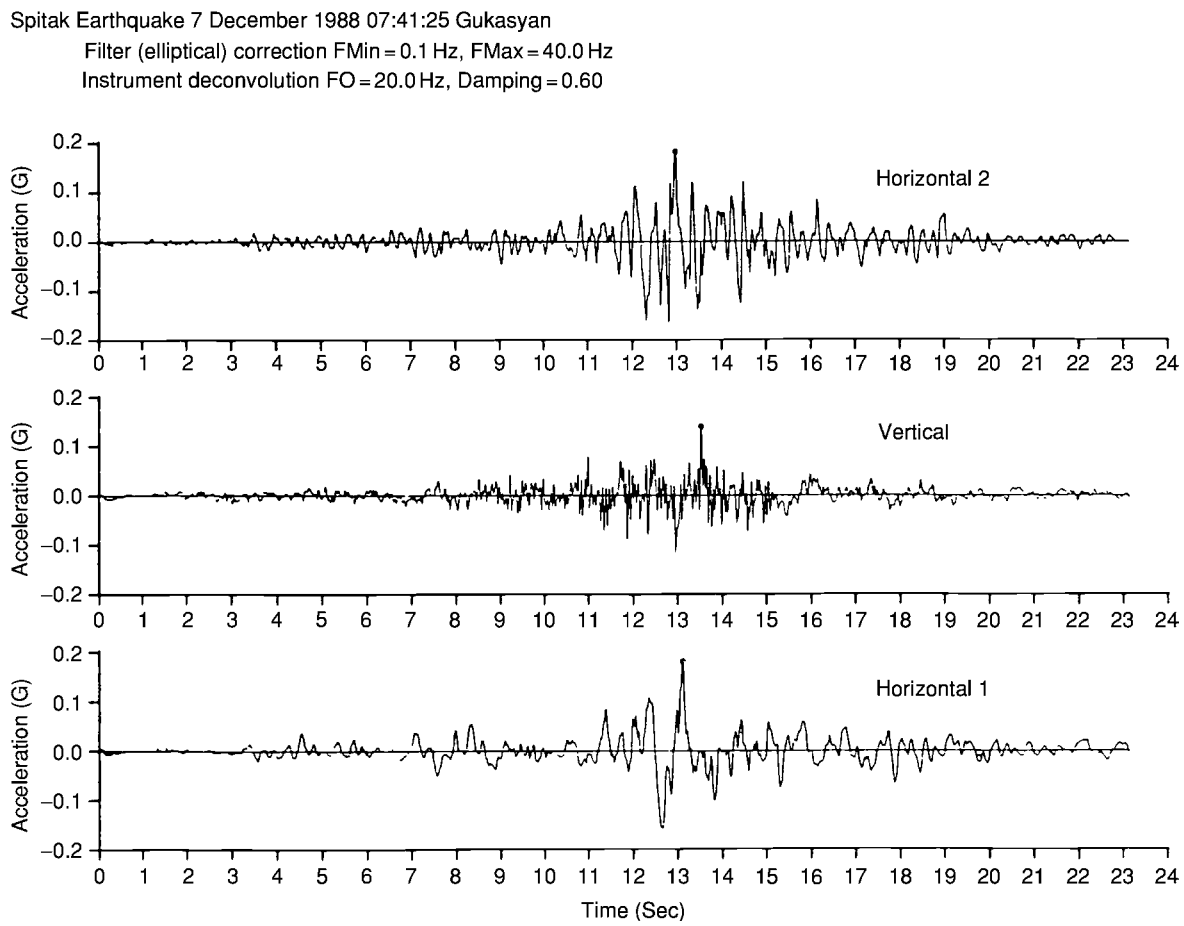


Figure 2 Record on three seismographs of the Spitak 1988 main shock (magnitude 6.9). Reproduced with permission from Rommer and Ambraseys (1989) *Earthquake Engineering and Dynamics*. Chichester: John Wiley & Sons. © John Wiley & Sons Limited.

waves. P waves can pass through both solid and molten material within the Earth's interior; they travel fastest and are the first to arrive at a given location, and they are also the first to be felt by the man in the street. They are longitudinal or compressional waves, vibrating forwards and backwards in the direction of travel. They travel at about 6 km s^{-1} through continental crust and 8 km s^{-1} through oceanic upper mantle. They may produce booming sound waves in the atmosphere. S waves travel about half as fast as P waves (3.6 km s^{-1} and 4.7 km s^{-1} in continental crust and oceanic upper mantle, respectively); they cannot pass through fluids and thus do not penetrate the liquid outer core. They are shear or transverse waves: as they pass through the rock they move particles both from side to side and up and down, at right angles to the direction of travel. Two kinds of surface waves, which travel just below the surface, are called Love and Rayleigh waves. They arrive shortly after the body waves. Love waves travel faster than Rayleigh waves and push the rock particles sideways, at right angles to the direction of travel. Like S waves, they shear buildings and constructions sideways, causing immense damage, but have no vertical motion. The slowest waves, Rayleigh waves, push particles upwards and backwards; the particles move in the vertical plane, following an elliptical path as the wave passes by.

Charles Richter in 1935, working in California, devised the Richter Scale of magnitude, in which the absolute strength at the focus can be calculated on a logarithmic scale: a rise of one unit of magnitude represents a 10-fold increase in absolute strength (i.e. a magnitude 5 earthquake is 10 times a strong as a magnitude 4 earthquake). The difference in energy release is even greater – an increase of one unit of magnitude represents 30–32 times as much energy being released. Theoretically, earthquakes with magnitudes of more than 10 could occur, but the greatest magnitude so far measured for any earthquake is about 9.5. The Richter scale is given in Table 1. It has been superseded as a scale for measuring the

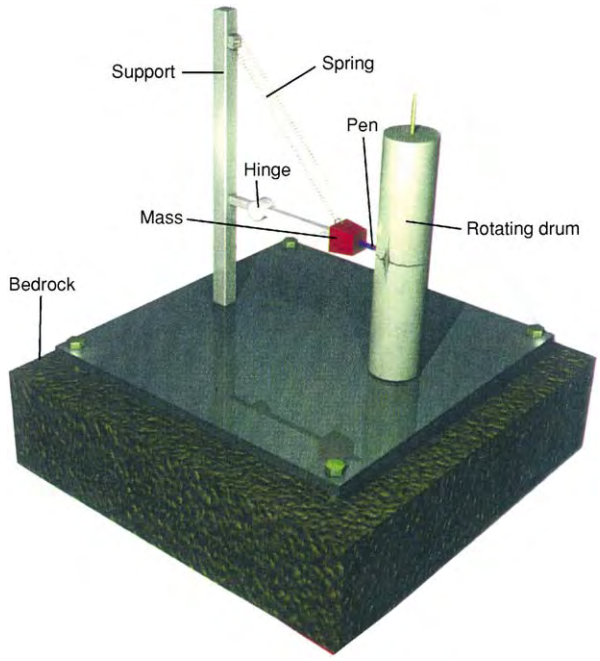


Figure 3 The components of a seismograph designed to record vertical ground movement.

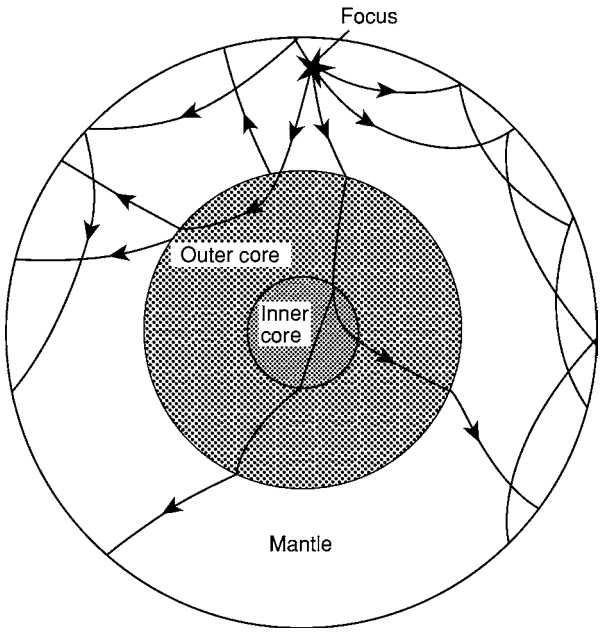


Figure 4 The way in which earthquake waves spread through the globe and are reflected at boundaries, returning to the surface. Measurement of the speed of such return is used to delineate the materials of the inner Earth according to density and physical state. Reproduced from Van Andel TJ (1994) *New Views on an Old Planet.* Cambridge: Cambridge University Press.

comparative intensity at the focus by the moment magnitude scale, but the principles are the same – the moment magnitude scale allows more refined methods of comparison.

Table 1 The Richter scale

Magnitude	*Qualitative description*	*Average number per year*	*Average intensity equivalent close to epicentre*
0 1.9		700 000	I V; recorded but not felt
2 2.9		300 000	I V; recorded but not felt
3 3.9	Minor	40 000	I V; felt by some
4 4.9	Light	6200	I V; felt by many
5 5.9	Moderate	800	V VII; slight damage
6 6.9	Strong	120	VII; damaging
7 7.9	Major	18	IX XI; destructive
8 8.9	Great	1 every 10 20 years	XII; widely devastating

The Importance of Seismological Records

We cannot directly study the rocks of the crust below the limits of borehole drilling (a few kilometres), though ancient rock systems do expose sections of the ancient deep crust (as in the Kapuskasing Belt, Ontario, Canada) and perhaps even the crust–mantle contact (as in Oman and Western Newfoundland). The behaviour of earthquake waves, however, provides us with invaluable evidence about the nature of the lower crust, mantle, and core because the velocities of P and S waves are functions of the density of the material through which they pass. Knowledge of rock density can tell us much about the physical state of the materials deep within the Earth, and the behaviour of S waves tells us that the outer core is molten. In Figure 5, the different densities of common rock materials are plotted against the P-wave velocity.

It is fair to say that earthquake waves form the basis of our knowledge of the mantle and core. Artificially produced seisms can also be picked up by seismographs, and, thus, nuclear explosions can be globally monitored. The explosion in the submarine Kursk in 2000 was picked up by distant seismograph stations in Africa, and this provided valuable evidence of what happened. Tomographic methods have recently been developed, producing three-dimensional images of the deep Earth, including subducted slabs of crust, using a technique akin to the use of tomography in medicine.

Earthquakes occur in sequences: slight foreshocks may give warning of a major earthquake, and aftershocks occur for long after the main shock. Foreshocks and aftershocks are generally of lower intensity than the main shock, but sometimes very

large shocks occur, as in the 1999 Izmit earthquake in Turkey. In the 1988 Spitak earthquake (magnitude 6.9) an aftershock of magnitude 6.2 occurred 4 min after the main shock.

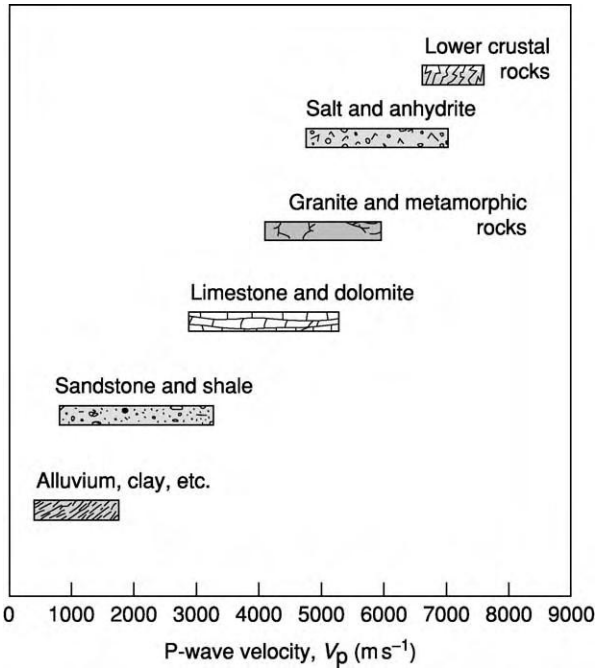


Figure 5 The different densities of common rock types plotted against P wave velocity.

Earthquakes may produce a trace of the rupture on the land surface, dislocating the land for many kilometres. In Figure 6 the trace produced across wheat fields by the Meckering 1968 (magnitude 6.9) earthquake is shown. Such traces are invaluable in studying the sense of the movement and displacement.

The Global Distribution of Earthquakes

Earthquakes do not occur to the same extent all over the globe. The major events are largely concentrated at the boundaries of tectonic plates, and the concentration and magnitude are greater in zones of plate convergence (subduction and collision) than in zones of plate divergence (mid-ocean ridges and rift valleys). This distribution is clearly shown in Figure 7.

Not all earthquakes occur on plate boundaries, however: the destructive Killari earthquake in India in 1993 occurred within a stable cratonic area. The immensely destructive Lisbon earthquake of 1755 was also nowhere near a plate boundary.

Earthquakes as a Major Hazard: Tectonic, Volcanic, and Man-Made Earthquakes

The most damaging earthquakes are not necessarily of high magnitude. The 1994 Kobe earthquake, one

Figure 6 Surface trace of the Meckering 1968 earthquake (magnitude 6.9) in Western Australia. Reproduced from Everingham I (1968) *Preliminary report on the 14th October 1968 earthquake at Meckering, Western Australia*. Record 1968/142. Canberra: Bureau of Mineral Resources, Geology and Geophysics.

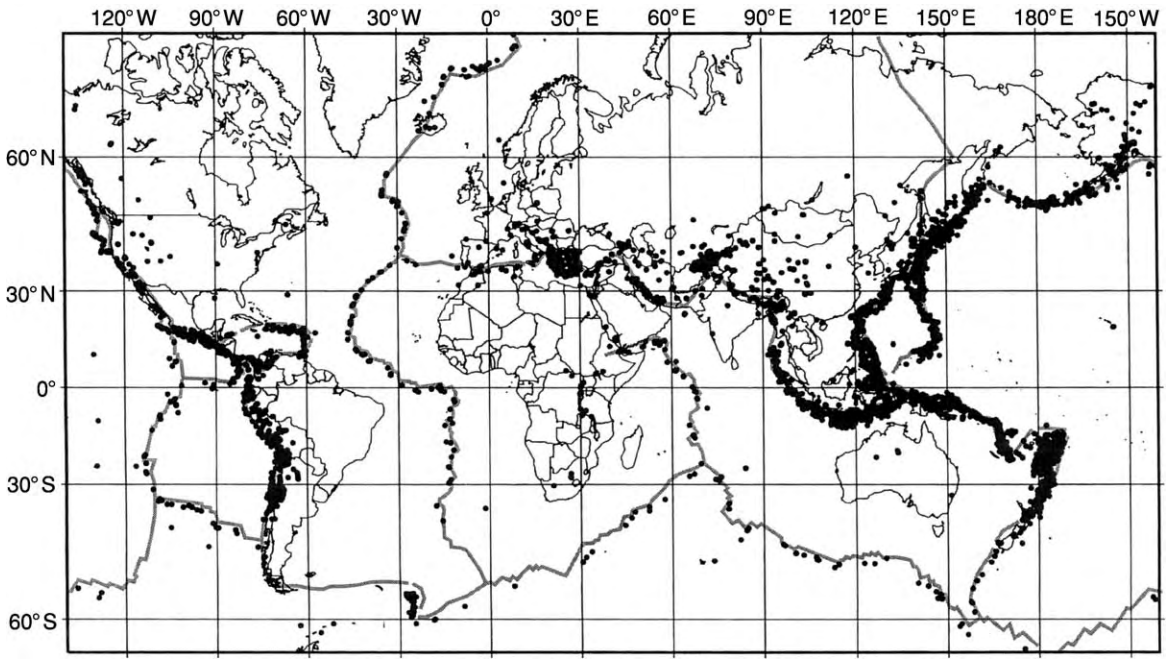


Figure 7 The global distribution of earthquakes that occurred in 1994. Reproduced from US National Earthquake Information Center.

of the most destructive and costly in living memory (55 000 houses destroyed as well as freeway, rail, and port installations), had its epicentre 20 km from the city and had a magnitude of only 6.8. The 1960 Agadir earthquake had a magnitude of only 5.8, but the focus was shallow and right beneath the city.

Thus, magnitude, though a valuable absolute measurement, tells us little about the degree of damage and the loss of property and life, even at the epicentre. The nature of the subsurface rocks can have a significant effect, especially if shock-induced liquefaction occurs, and thus we need another measurement scale. The Mercalli intensity scale measures the relative intensity of the effects felt at a specific site (the intensity will commonly decrease away from the epicentre, but secondary effects such as subsurface variation and shock liquefaction complicate this relationship). In Europe, a modification of the Mercalli scale, the MSK scale (named after Medvedev, Sponheuer, and Karnik), is now used. This scale is given in Table 2.

Tectonic Earthquakes

Earthquakes may be divided into tectonic earthquakes, volcanic earthquakes, and man-made earthquakes. In considering the natural-hazard aspect, it is the tectonic earthquakes that are by far the most destructive natural hazards. This hazard largely affects urban populations, and human design and construction has a unique role in mitigating this hazard.

The actual physical process of ground motion presents little threat to humans in the open: most casualties (other than the casualties of secondary tsunamis) occur inside buildings that partially or totally collapse. The correct design of buildings and constructions such as bridges and viaducts can thus greatly mitigate the damage and casualties resulting from an earthquake.

The vulnerability of a building to earthquake damage varies according to many factors. Vertical ground motion is the principal damaging component causing collapse, burial of people, and death. Lateral ground motion breaks or deforms power lines, pipelines, water pipes and sewers, roadways, railways (Figure 8), and bridges. Quite small lateral offsets can be very damaging. In the Mexico City earthquake of 1985, much damage was caused by adjacent high-rise buildings swaying with different wave motions and knocking each other down. It is noticeable that in Beijing, an earthquake-prone city, the high-rise buildings are widely spaced, with intervening areas of low-rise buildings, so that they cannot interact in this way. In the case of Kobe, the sixth floor of one high-rise building pancaked (Figure 9). The lower floors were built of steel-encased reinforced concrete and the upper floors of pure reinforced concrete; the junction on the sixth floor acted as an element of weakness.

Though earthquakes are mainly an urban hazard, catastrophic earthquakes may strike village populations where low-rise housing is substandard – as in

Table 2 The modified Mercalli (MSK) intensity scale

Intensity	*Effects*
I	Felt rarely. Sometimes dizzyness and nausea. Birds and animals uneasy. Trees, structures, liquids sway.
II	Felt indoors by a few, especially on upper floors. Delicately suspended objects swing.
III	Felt indoors, especially on upper floors by several people. Usually rapid vibration as if a lightly loaded lorry passing. Hanging objects and standing motor cars rock slightly.
IV	Felt indoors by many and outside by a few. Some awakened. No one usually frightened. Sensation of heavy object striking building. Vibration as of heavy lorries passing. Crockery, windows, doors rattle. Walls and frames creak. Hanging objects and standing motor cars sway.
V	Felt indoors by almost everyone, outdoors by most people. Many awakened, a few frightened and run outdoors. Buildings tremble. Crockery and windows sometimes break. Pictures and doors clatter. Small objects move. Some liquids spilt. Clocks stopped. Trees shaken. Animals anxious.
VI	Felt by all indoors and outdoors. Many frightened, some alarmed. All awakened. People, trees, bushes shaken. Liquids set in motion. Small bells ringing. Crockery broken. Plaster cracks and falls. Books and vases fall over. Some furniture moved. Domestic animals try to escape. Minor landslides on steep slopes.
VII	All frightened, run out of doors, general alarm. Some people thrown to ground. Trees shaken quite strongly. Waves and mud stirred up in lakes. Sandbanks collapse. Large bells ring. Suspended objects quiver. Much damage to badly constructed buildings and old walls. Slight damage to well built buildings. Chimneys crack. Much plaster, tiles, loose bricks fall. Heavy furniture overturned. Concrete ditches damaged.
VIII	Alarm approaches panic. Vehicle drivers disturbed. Trees broken and shaken. Sand and mud spurt from the ground. Marked changes to springs and wells. Much damage to ordinary and older buildings. Walls, pillars, chimneys, towers, statues, gravestones crack and fall. Very heavy furniture overturned.
IX	General panic. Ground cracked open (10 cm). Much damage to structures built to withstand earthquakes. Frequent partial or total collapse of other buildings. In reservoirs, underground pipes broken. Buildings dislodged from foundations, rock falls.
X	Widely cracked ground, fissures up to 1 m wide. Frequent river bank and coastal landslides and shifted sands. Water levels change. Water thrown onto riversides. Serious damage to dams, embankments, bridges. Severe damage to well built wooden structures. Masonry structures destroyed along with their foundations. Rails bent. Open cracks or waves on roads. Pipes torn apart.
XI	Widespread serious ground disturbance, broad fissures, landslips, landslides. Muddy water spurts upward. Tsunamis develop. Severe damage to all wooden structures. Great damage to dams. Few masonry structures remain upright. Pillars of bridges and viaducts wrecked. All pipelines wrecked. Rails badly bent.
XII	Total damage to all constructions. Great disturbance to ground with many shearing cracks. Many landslides on slopes, rockfalls common, rock masses dislocated, water channels altered and dammed. Ground surface waves like water and ground remains undulating. Objects thrown into the air.

the case of the Cairo earthquake in 1992 and the Killari, central India, earthquake in 1993. In the case of the Cairo earthquake, poorly constructed extra storeys had been added to moderate-rise housing. In Killari, stone-built low-rise houses were poorly constructed (**Figure 10**). In the catastrophic Bam earthquake of December 2003, the mud bricks of the low-rise dwellings crumbled and collapsed, leaving few air spaces to allow buried victims to breathe; another factor responsible for the scale of the fatalities was the fact that all the dwellings in southern Iran have basements cooled by wind towers designed for the sweltering summer heat, and many victims would have been asleep in them.

Some of the most devastating historic and recent earthquakes are listed in **Table 3**.

Where cities are situated in plate-boundary zones the effects are most disastrous. The San Francisco earthquake of 1906 (**Figure 11**) provides an example of this. The Kobe earthquake of 1995, in which the financial loss was US $200 billion, occurred in a city that experiences a tremor every few days. An analysis of the locations of 100 of the largest cities in the world, which accommodate 10% of the global population, shows that they can expect to experience an earthquake of intensity VI or more on the MSK scale within 50 years.

The earthquake hazard extends beyond high-risk cities such as those sited on plate boundaries. Entire countries may be at quite low risk, yet have some vulnerability. The UK is a low-risk country, and earthquakes of more than magnitude 5.5 are extremely unlikely (**Figure 12**). Charles Davidson published a list of 1191 recorded shocks in Britain between AD 974 and AD 1924. In Lincoln in 1185, "great stones were rent; houses of stone fell; the metropolitan church of Lincoln was rent from top to bottom" and there is a similar report from the cathedral city of Wells in 1248. Two apprentices were killed in London in 1580 as a result of an earthquake in the Dover Straits. The Colchester earthquake in 1884 (magnitude 4.7) peaked in intensity near the epicentre (between Pelden and Langenhoe) (**Figure 13**) at MSK VIII and caused widespread damage, which was

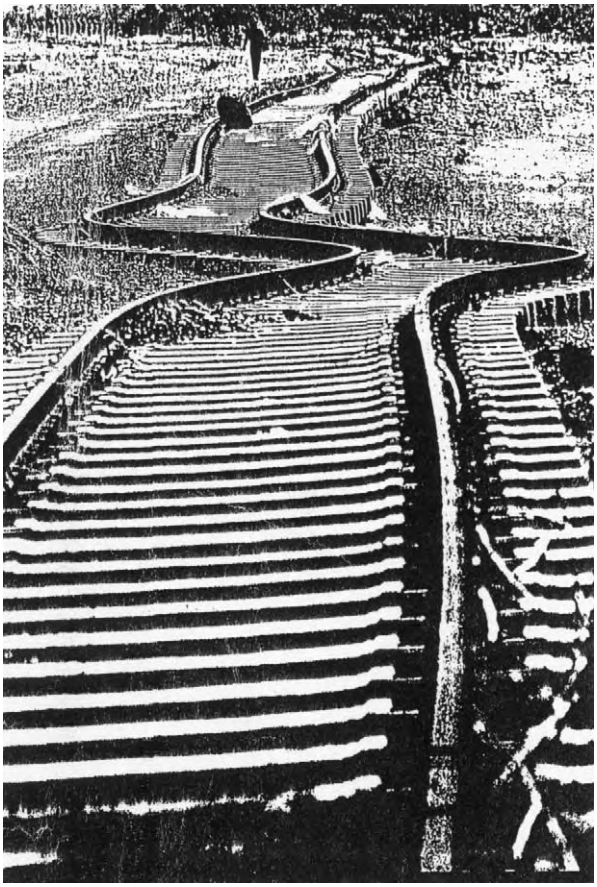


Figure 8 A railway track in the western USA twisted and shortened by lateral motion during an earthquake.

compensated by a Mansion House Fund that paid out £9000 (equivalent to £500 000 today). The Roermond earthquake in the Netherlands in 1998 (magnitude 5.8) caused £30–40 million of damage, despite the single fatality. It is predicted that a magnitude 5.7 earthquake focused at a depth of 5 km directly under Manchester would cause havoc. The increasing size of conurbations and cities increases their vulnerability: the Colchester area would suffer much more nowadays from a comparable earthquake to the 1884 event, because the population and industry are now much denser than at that time.

Earthquakes can occur in areas that are not considered to be at risk. The Spitak earthquake in Armenia in 1988 is such a case. The region was not considered to be at high risk, and a nuclear power station was planned for the Spitak area. This earthquake caused the whole process of earthquake risk assessment in the Soviet Union to be revised.

The New Madrid, Missouri, earthquakes of 1811–1812 are even more surprising. There is a reliable historical record of three earthquakes spaced over two months with magnitudes 8.2, 8.1, and 8.3. They rang the bells of Boston and rattled Quebec and provide a remarkable example of major interplate seismicity.

Secondary effects The secondary effects of earthquakes can be as destructive and lethal as the primary effects, or more so.

Figure 9 High rise buildings in Kobe after the 1995 earthquake, showing the sixth floor pancaked by vertical motion. Reproduced from Esper P and Tachibana E (1998) The Kobe earthquake. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology*, pp. 105 116. Engineering Geology Special Publication 15. London: Geological Society.

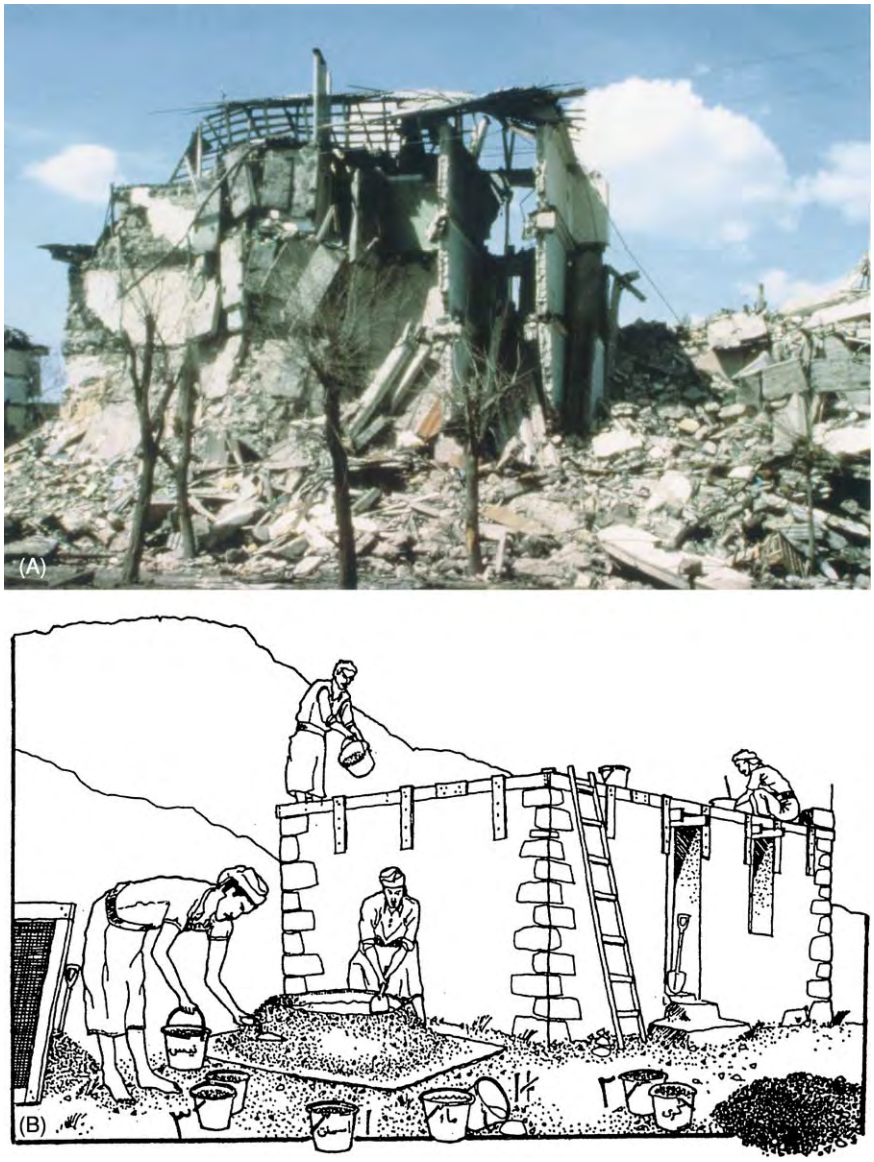


Figure 10 (A) Damage to a low rise poorly constructed stone building of the type affected by the Killari, India, 1991 earthquake (photograph National Geophysical Data Center USA). (B) Improved training in building similar low rise buildings in the Yemen. Reproduced from Degg MR (1995) Earthquakes, volcanoes and tsunamis: tectonic hazards in the built environment of southern Europe. *Built Environment* 21: 94 113. Courtesy of the Geological Society, London.

Tsunami The so-called tidal wave generated by earthquakes is probably the most lethal secondary effect: during the twentieth century earthquakes in Chile caused fatalities in Hawaii and Japan. The effects of tsunamis may be felt thousands of kilometres from the earthquake epicentre, but can be mitigated by systematic warning systems.

Fire Tokyo in 1923, San Francisco in 1906, and Kobe in 1995 all suffered from the secondary effects of fire. This was exacerbated by the fact that water supplies were cut off. In Tokyo there was a firestorm. In San Francisco, 70% of the damage was due to fire.

Liquefaction of sands, silts, and clays Another important secondary effect is that thixotropic sands and silts, which liquefy on shock, greatly increase the damage: examples of this are the waterfront area in the Messina earthquake, Sicily, of 1908, in which 98% of the houses were ruined and 160 000 died; Mexico City in 1985, where the worst damage was in building developments founded on old lake deposits (the wave amplitude was magnified 8–50 times here); and Anchorage, Alaska, where a magnitude 8.4 earthquake with an epicentre 130 km away caused devastation in a housing development founded on the thixotropic Bootlegger Clay

Table 3 Some important earthquakes in the last 2000 years (various sources): note that magnitudes are on various scales. An earthquake at Gujarat, India, in 2001, which killed more than 50 000 people has been omitted from the table

Year AD	*Place*	*Casualties*	*Estimated loss*
342	Antioch	40 000	
454	Sparta	20 000	
565	Antioch	30 000	
856	Corinth	45 000	
1170	Sicily	15 000	
1290	Chihli, China	100 000	
1456	Naples	60 000	
1556	Shensi, China (M 8.3?)	830 000	
1716	Algiers	20 000	
1737	Calcutta	300 000	
1755	North Persia	40 000	
1755	Lisbon (M 6.9?)	60 000	
1759	Baalbek	20 000	
1783 1786	Calabria	50 000	
1797	Quito	41 000	
1822	Aleppo	22 000	
1828	Honshu	34 000	
1896	Sanriku, Japan	28 000	
1897	Assam (M 8.7)	1542	6 major towns and all villages in 30 000 sq. miles leveled
1906	San Francisco (M 7.9?)	700	US$ 400 million
1908	Messina, Reggio (M 7.5)	160 000	98% of houses ruined
1915	Avezzano, Italy (M 7.0)	30 000	
1920	Kansu, China (M 8.5)	180 000	Vast landslides
1923	Tokyo (M 8.2)	143 000	Firestorm killed 38 000, 25 000 houses destroyed
1932	Kansu, China (M 8.5)	70 000	
1933	Long Beach, CA	120	US$ 50 million
1933	Sanriku, Japan	3000	8800 houses destroyed by tsunami
1935	Quetta (M 7.5)	60 000	
1939	Concepcion, Chile (M 8.5)	30 000	
1939	Erzincian, Turkey (M 8.0?)	40 000	30 000 dwellings destroyed
1948	Soviet Iran border	19 000	
1960	Agadir (M 9.5?)	60 000	
1960	Chile (M 9.5?)	10 000	58 600 houses destroyed
1964	Anchorage (M 8.4)	114	
1970	Peru (M 7.9)	80 000	Devastating rock and ice falls
1971	San Fernando, CA	64	US$ 1 billion
1975	Haicheng, China (M 7.5)	1328	
1976	Guatemala (M 7.5)	22 000	
1976	Tangshan, China (M 7.7)	240 000 (some estimates are as high as 850 000)	Vast damage
1985	Mexico City (M 8.1)	10 000	US$ 4 billion
1988	Spitak (M 6.9)	30 000	US$ 14 billion; accompanied by landslides and rockfalls
1989	Loma Prieta, CA (M 7.1)	63	US$ 7 billion
1989	Newcastle, Australia	10	US$ 7 billion
1990	Northwest Iran (7.7)	40 000	US$ 8 billion
1991	Killari, India	10 000	Immense destruction of village housing
1992	Cairo (M 5.5 5.9)	$<$500	40 000 homeless; may have been due to construction of Aswan Dam
1994	Northridge, CA (M 6.7)	60	US$ 20 billion
1995	Kobe, Japan (M 7.2)	5429	US$ 200 billion
1999	Izmit, Turkey (M 7.6)	$>$17 000	Immense destruction
2003	Bam, Iran	$\sim$45 000	Immense destruction of modern city and ancient citadel destroyed

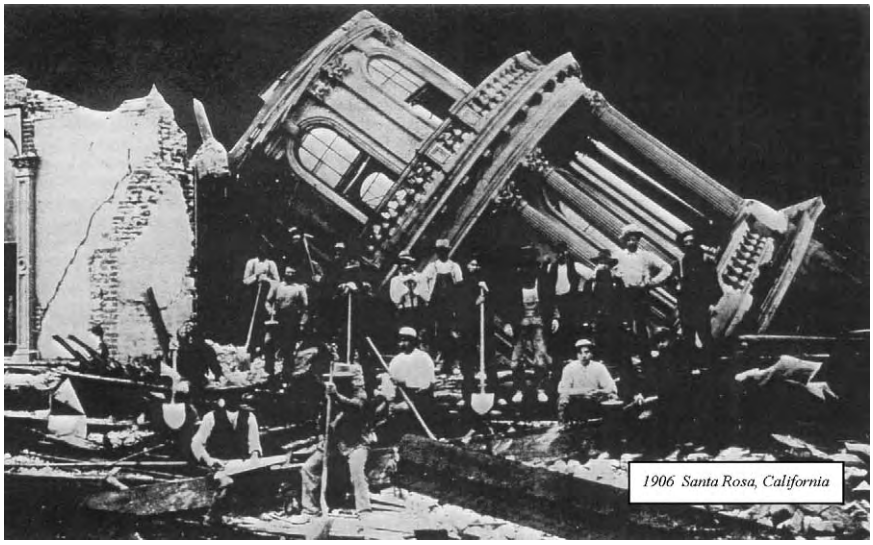


Figure 11 Gross displacement of a large building in the San Francisco 1906 earthquake.

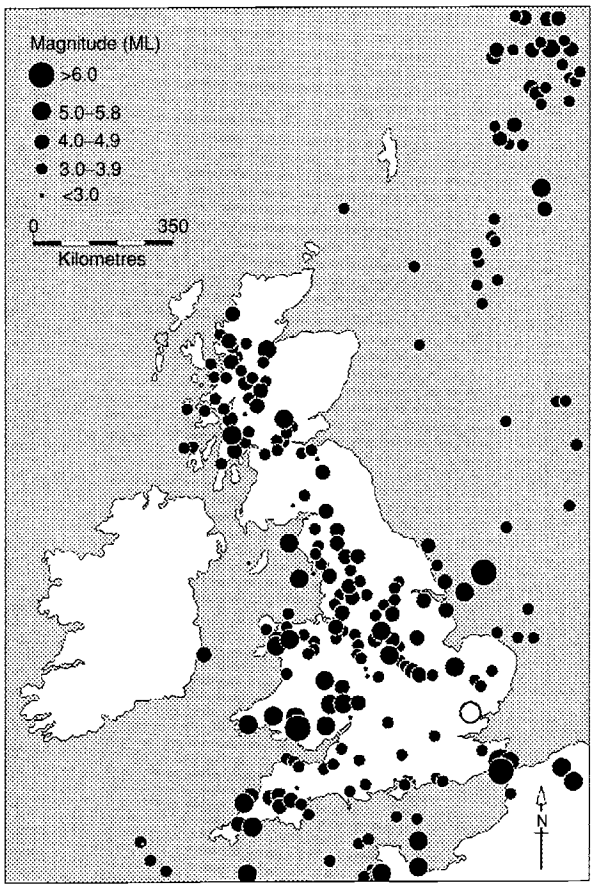


Figure 12 The magnitudes and locations of earthquakes in Great Britain greater than magnitude 3 after 1700 and greater than magnitude 4 before 1700. Reproduced from Musson R (1996) British earthquakes and the seismicity of the UK. *Geoscientist* 16: 24 25.

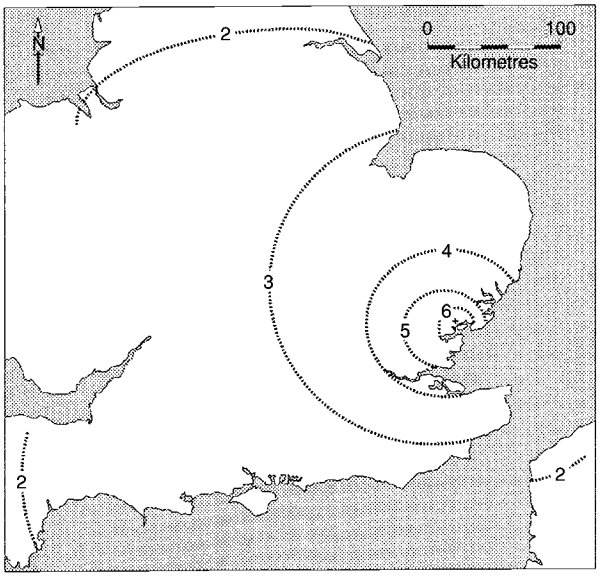


Figure 13 Intensity plot of the Great Colchester Earthquake of 1884 using the MSK scale. Reproduced from Musson R, Neislon G, and Burton PW (1990) *Microseismic Reports on Historic British Earthquakes XIV: 22 April 1984 Colchester*. BGS Seismology Report W1/90/33. Edinburgh: British Geological Survey.

Formation (Figure 14). Here, the risk was well known but there was a lack of communication between the geologists and the planners.

Landslides and rock falls Very damaging landslides or rock falls can be triggered by earthquakes and may occur some time after the main shock or aftershocks. In Montana in 1985, 30 million tonnes of rock were

Figure 14 Destruction of the housing development at Turnagain Point above the Bootlegger Clay, Anchorage, 1994. The bluff moved 606 m towards the bay, and 75 homes were destroyed. Sand lenses in the clay lost strength (photograph National Geophysical Data Center, USA).

set in motion, fatally burying 36 people at a camp site. In Peru in 1970, rock and ice slides triggered by an earthquake killed 20 000 people.

Disease The risk of disease is a major concern after earthquakes, particularly in hot climates. Water supplies may be cut off and the populace may resort to using polluted supplies of water mixed with sewage and drainage effluents, which may be contaminated by corpses.

Starvation Normal food supplies may be cut off and transport arteries may be blocked, so it is important to bring in food supplies immediately from the world outside.

Exposure The 1988 Spitak earthquake in Armenia illustrates the problem of exposure. Many people lost their dwellings and were living in the open in very cold December climatic conditions. The need for tents, warm clothing, and blankets was urgent.

Looting Looting is prevalent after earthquakes.

Mitigation The potential for mitigation of the earthquake hazard is limited. The main ways of mitigating the hazard are through good building and constructional design, planning development away from at-risk areas, and warning. However, warning is a very difficult matter. Research into earthquakes is at an interim stage, and the scientific community is at present by no means in consensus about the physical processes involved. There is the problem of how threatened populations react to warnings: if the event fails to occur, especially more than once, the population may not heed future warnings; alternatively, giving a warning to evacuate may engender panic. In countries with controlled political systems, such as China, warning and evacuation may be easier than in a democratic Western country. The Chinese did evacuate the city of Haicheng twice in 1974 and 1975, on 4 December and again on 4 February, based on seismology, community monitoring levels, radon gas in water, water temperature, tiltmeters, magnetometer readings, and patterns of animal behaviour. An earthquake of magnitude 7.3 struck at 7.36 AM on 4 February. However, the great Tangshan earthquake of 1976 struck without any prediction or warning and killed at least 240 000 people (possibly many more).

Millions of US dollars have been spent on research into earthquakes in California, and some improvement in prediction has been achieved, yet the existence of the Northridge Fault, the site of the 1994 earthquake, which killed 60 people and caused 20 billion dollars worth of damage, was not even known before the event.

The best mitigation procedure would be to have international teams ready with emergency supplies and equipment, trained personnel, and sniffer dogs, at a distance from earthquake-prone regions, ready to been flown in by plane and helicopter.

Research into earthquakes The most important research into earthquakes has involved statistical,

geographical, geological, theoretical, and mathematical studies of seismicity. An example is a study by Lya Tuliani in Russia, which addressed problems of geodynamics and seismology, tectonosphere layering, and lithostructure in seismically active regions in order to develop risk estimates. The procedure involved mathematical data processing. It was claimed that this study allowed highly accurate prediction of the coordinates of high-risk sites. This statistical, mathematical, and office-based approach contrasts with research in the USA (which involves actually drilling down to earthquake foci on faults), research into the Boothiel lineament, the site of the New Madrid earthquakes (which has revealed sand boils caused by the earthquake), and excavations in the alluvium of the rice paddies west of Beijing, China (where the actual earthquake fault of a seventeenth century event has been exposed cutting the clayey alluvium in open pits). Research has been carried out into much older earthquakes in Iran, based on the dislocation of qanats (horizontal wells). All these approaches and many more are invaluable, but the problem of predicting earthquakes is extremely complex and may never be completely solved.

Volcanic Earthquakes

There can be a connection between major tectonic earthquakes and volcanic eruptions. In Chile in 1960, a major earthquake triggered the eruptions of several volcanoes, and in Sicily there is a record of Etna erupting a day or two before a major tectonic earthquake. However, the swarms of small seisms that usually precede volcanic eruptions (though there may be no such prelude) pose little threat to life and property. They do, however, provide valuable warnings of forthcoming eruption, and arrays of instruments are mounted on dangerous volcanoes for this purpose.

Man-Made Earthquakes

Small seismic disturbances can be triggered by human activity. In the USA, the Boulder Dam and Lake Mead are constructed in a region that is highly strained; many small shocks have been correlated with changes in water depth. In Colorado, the injection of liquid wastes down boreholes has also been shown to trigger small seisms.

Moonquakes and Seisms on Other Planets

Very small earthquakes do occur on the largely quiescent Moon when it undergoes maximum tidal stresses resulting from the attraction of the Earth and Sun. Similar stresses must operate on the Earth and cause minor seisms, but the effect is of no importance in such a dynamic body. Moonquakes and artificial seisms produced on the surfaces of other extraterrestrial bodies – Mars, Venus, and Mercury – can provide a valuable insight into their internal make up. A fascinating project would be to site an instrument from an unmanned spacecraft on Io, Jupiter's volcanically active satellite, to obtain detail of its interior configuration – Io must be seismically active.

See Also

Earth: Mantle; Crust. **Earth Structure and Origins**. **Engineering Geology:** Aspects of Earthquakes; Natural and Anthropogenic Geohazards; Liquefaction. **Plate Tectonics**. **Tectonics:** Faults.

Further Reading

Bolt BA (1999) *Earthquakes*. New York: Freeman.

Bommer JJ and Ambraseys NN (1989) The Spitak, Armenia, USSR earthquake of 7 December 1988: a summary engin eering geology report. *Earthquake Engineering and Struc tural Dynamics* 18: 921 925.

Chen Y, Tsoi KL, Chen F, *et al*. (1988) *The Great Tangshan Earthquake of 1976*. Oxford: Pergamon.

Degg MR (1992) Some implications of the 1985 Maxican earthquake for hazard assessment. In: McCall GJH, Laming DJC, and Scott SC (eds.) *Geohazards Natural and Man Made*, pp. 93 114. London: Chapman and Hall.

Degg MR (1995) Earthquakes, volcanoes and tsunamis: tectonic hazards in the built environment of southern Europe. *Built Environment* 21: 94 113.

Degg MR (1998) Hazard mitigation in the urban environ ment. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology*, pp. 329 337. Engineering Geology Special Publication 15. London: Geological Society.

Esper P and Tachibana E (1998) The Kobe earthquake. In: Maund JG and Eddleston M (eds.) *Geohazards in Engin eering Geology*, pp. 105 116. Engineering Geology Special Publication 15. London: Geological Society.

Everingham I (1968) *Preliminary report on the 14th Octo ber 1968 earthquake at Meckering, Western Australia*. Record 1968/142. Canberra: Bureau of Mineral Re sources, Geology and Geophysics.

Keller GR (2000) Seismic properties of rocks. In: Hancock PL and Skinner BJ (eds.) *The Oxford Companion to the Earth*. Oxford: Oxford University Press.

McCall GJH (1996) Natural hazards. In: McCall GJH, de Mulder EFJ, and Marker BR (eds.) *Urban Geoscience*, pp. 81 125. Rotterdam: Balkema.

McCall GJH (2000) The great Colchester earthquake of 1884 revisited. *Geoscientist* 10: 4 6.

McCall GJH (2004) Remembering Bam. *Geoscientist* 14: 8 9.

Menard HW (1974) *Geology, Resources and Society*. San Francisco: WH Freeman and Co.
Musson R (1996) British earthquakes and the seismicity of the UK. *Geoscientist* 16: 24 25.
Musson R, Neislon G, and Burton PW (1990) *Microseismic Reports on Historic British Earthquakes XIV: 22 April 1984 Colchester.* BGS Seismology Report W1/90/33. Edinburgh: British Geological Survey.
Scarth A (1997) *Savage Earth*. London: HarperCollins.
Tuliani LI (1999) *Seismicity and Earthquake Risk: On the Basis of Thermodynamic and Rheological Parameters of the Tectonosphere*. Moscow: Scientific World.
Van Andel TJ (1994) *New Views on an Old Planet*. Cam bridge: Cambridge University Press.
Wong IG (2000) Earthquake mechanisms and plate tecton ics. In: Hancock PL and Skinner BJ (eds.) *The Oxford Companion to the Earth*, pp. 287 289. Oxford: Oxford University Press.

Faults

S Stein, Northwestern University, Evanston, IL, USA

Introduction

Faults are surfaces in the Earth along which one side moves or has moved with respect to the other. They are identified either when an earthquake occurs or by geological mapping showing that motion across the fault has occurred in the past. Many faults are inactive, in the sense that there has been no motion across them within some defined time interval, typically the past million years or less. Other faults are active, in the sense that recent motion has occurred and hence motion might be expected in the future. Faults, and the earthquakes on them, are studied to understand both the regional tectonics and the mechanics of faulting.

Typically, earthquakes occur on previously identified faults, demonstrate that the fault is active, and provide information on the fault's geometry and the motion on it. For example, in the famous 1906 San Francisco earthquake, one of the first earthquakes to be carefully studied, several metres of relative motion occurred along several hundred kilometres of the San Andreas Fault. Hence, H Reid proposed the elastic-rebound theory of earthquakes, in which materials on opposite sides of the fault move relative to each other, but friction 'locks' the fault and prevents it from slipping (**Figure 1**). Eventually more strain accumulates than the fault rocks can withstand, and the fault slips in an earthquake. The motion is sometimes revealed after earthquakes by linear features, including roads and rows of trees (**Figure 2**). Those who study earthquakes seek to understand both the geological processes causing earthquakes and the physics of faulting. These issues are important for society because knowing where and when earthquakes are likely and the expected ground motion during them can help to mitigate the risk that they pose.

The largest earthquakes occur at plate boundaries. We view them as the most dramatic part of the seismic cycle, which takes place on segments of the plate boundary over hundreds or thousands of years. During the interseismic stage, which makes up most of the cycle, steady motion occurs at a distance from the locked fault. Immediately prior to rupture there is the preseismic stage, during which small earthquakes (foreshocks) and other possible precursory effects may occur. The earthquake is the coseismic phase, during which rapid fault slip generates seismic waves. During these few seconds, metres of slip on the fault 'catch up' with the few millimetres per year of motion that have occurred over hundreds of years at a distance from the fault. Finally, a postseismic phase occurs after the earthquake, during which aftershocks and transient afterslip occur for a period of years before the fault resumes steady interseismic behaviour.

Because this cycle extends over hundreds of years, we do not have observations of it in any one place. Instead, our view of the seismic cycle is based on a combination of observations from different places. It is far from clear how good this view is and how well our models represent the complexity of the seismic cycle. As a result, earthquake and fault studies remain active research areas that integrate a variety of techniques. Seismology is used to study the motion during earthquakes. Historical records often provide data on the earthquake cycle for a given fault segment. Field studies provide information about the location, geometry, and history of faults. Geodetic measurements are used to study ground deformation before, during, and after earthquakes and thus provide information about the processes associated with fault locking and afterslip. Results for individual earthquakes are combined with those from other analyses, including laboratory studies of rock deformation, to understand how the earthquakes in a region reflect the large-scale tectonic processes causing them and to study the physics of faulting.

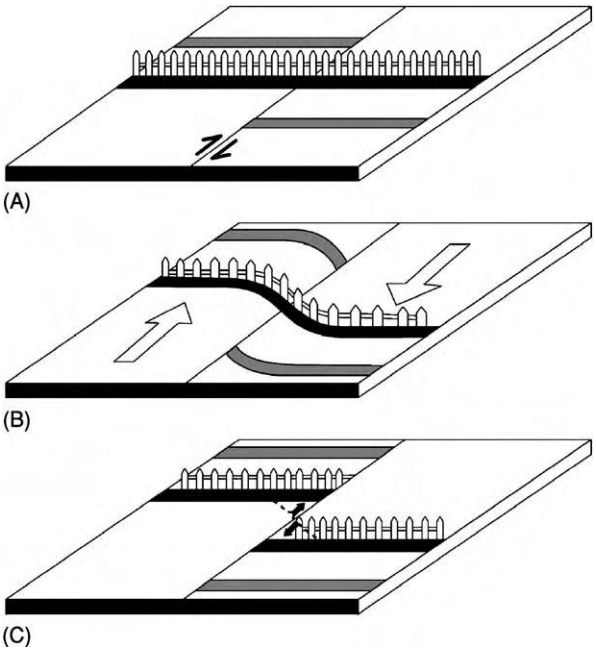


Figure 1 The elastic rebound model of earthquakes assumes that between earthquakes material at a distance from the fault undergoes relative motion. Because the fault is locked, features across it that were originally linear (A), such as a fence, are slowly deformed with time (B). Finally, the strain becomes so great that the fault breaks in an earthquake, offsetting the features (C). (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

Figure 2 Displacement of crop rows resulting from an earthquake on the Imperial fault, El Centro, California on 15 October 1979. (Courtesy of the National Geophysical Data Center.)

Fault Geometry

We treat faults as planar surfaces across which relative motion occurs during earthquakes. Geological observations of faults that reach the surface show that this is often approximately the case, although complexities are common. This assumption is usually also consistent with seismic data.

As shown in Figure 3, the fault plane is characterized by **n**, its normal vector. The direction of motion is given by **d**, the slip vector in the fault plane. The slip vector indicates the direction in which the upper side of the fault, known as the hanging-wall block, moves with respect to the lower side of the fault, known as the foot-wall block. Because the slip vector is in the fault plane, it is perpendicular to the normal vector.

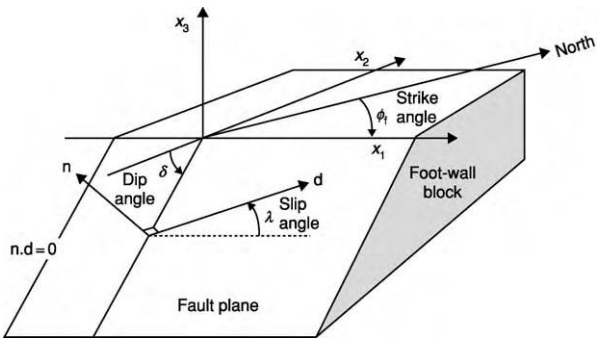


Figure 3 Fault geometry used in earthquake studies. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

A coordinate system for studying faults has the x_1 axis in the fault strike direction, the intersection of the fault plane with the Earth's surface. The x_3 axis points upwards, and x_2 is perpendicular to the other two axes. The dip angle, δ, gives the orientation of the fault plane with respect to the surface. The slip angle, λ, gives the motion of the hanging wall with respect to the foot wall. Fault geometries are described by the slip angle (Figure 4). When the sides slide by each other, pure strike-slip motion occurs. When $\lambda = 0°$, the hanging wall moves to the right, and the motion is called left-lateral. Similarly, for $\lambda = 180°$, right-lateral motion occurs. To tell which is which, look across the fault and see which way the other side moves. The other basic fault geometries describe dip-slip motion. When $\lambda = 270°$, the hanging wall slides downwards, causing normal faulting. In the opposite case, $\lambda = 90°$ and the hanging wall goes upwards, yielding reverse or thrust faulting. Seismologists often use the terms reverse fault and thrust fault interchangeably, whereas structural geologists reserve the term thrust fault for a shallow-dipping reverse fault. Most earthquakes consist of some combination of these motions and have slip angles between these values.

If we treat a fault as rectangular, the dimension along strike is called the fault length and the dimension in the dip direction is known as the fault width. Actual fault geometries can be much more complicated than a rectangle. The fault may curve, requiring a three-dimensional description. Rupture may occur over a long period and consist of several subevents on different parts of the fault with different

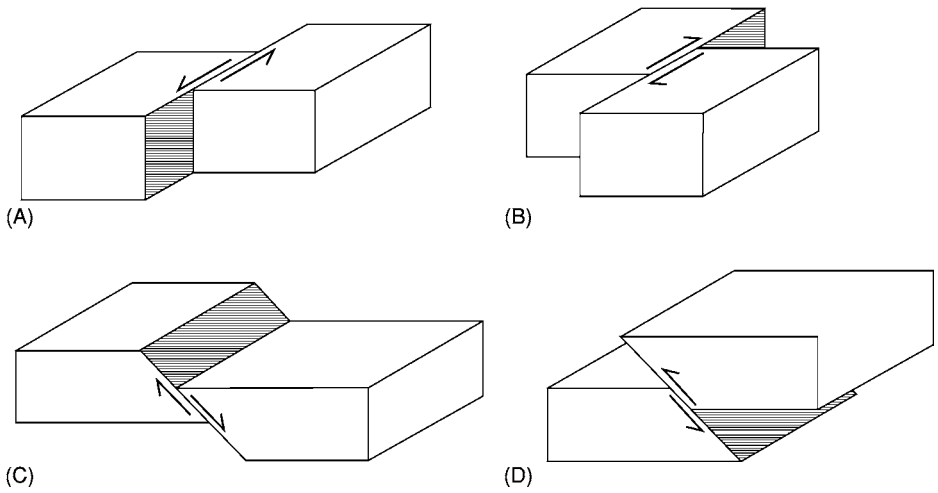


Figure 4 Basic types of faulting. (A) Left lateral strike slip fault, λ 0°. (B) Right lateral strike slip fault, λ 180°. (C) Normal dip slip fault, λ 90°. (D) Reverse dip slip fault, λ 90°. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

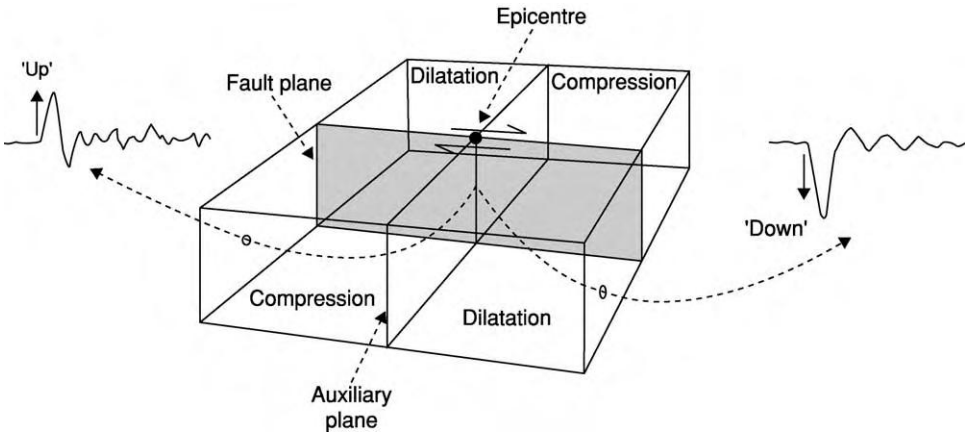


Figure 5 First motions of P waves observed at seismometers located in various directions about the earthquake provide information about the fault orientation. The nodal planes separate compressional ('up') and dilatational ('down') first arrivals. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

orientations. Such complicated seismic events, however, can be treated as a combination of simple events.

Seismological and Geodetic Studies

Seismological studies provide much of our information about earthquakes and the faults on which they occur. The arrival times of seismic waves at seismometers at different sites are first used to find the location of an earthquake, known as the focus or hypocentre. This location is often shown by the epicentre, the point on the Earth's surface directly above the earthquake. Next, the amplitudes and shapes of the radiated seismic waves are used to study the size of the earthquake, the geometry of the fault on which it occurred, and the direction and amount of slip.

The geometry of faulting during an earthquake, known as the focal mechanism, is found by using the fact that the pattern of radiated seismic waves depends on the fault geometry. Seismic waves are divided into P or compressional waves, in which material moves back-and-forth in the direction of wave propagation, and S or shear waves, where material moves at right angles to the propagation direction. P waves travel faster than S waves, so the first pulse to arrive is a P wave. The simplest method uses the first motion, or polarity, of the first-arriving P wave, which varies between seismic stations at different directions from an earthquake. As illustrated (Figure 5) for a strike-slip earthquake on a vertical fault, the first motion is either compression, for stations located such that material near the fault moves 'towards' the station,

or dilatation, where the motion is 'away from' the station. The first motions define compressional and dilatational quadrants, divided by two perpendicular nodal planes – the fault plane and the auxiliary plane, which is perpendicular to the fault plane. If these planes can be found, the fault geometry is known, and can be plotted using the familiar stereographic or 'beachball' representation.

Because the first motions from slip on the actual fault plane and the auxiliary plane would be the same, first motions alone cannot resolve which is the actual fault plane. However, additional information can often settle the question. Sometimes geological or geodetic information, such as the trend of a known fault or observations of ground motion, indicates the fault plane. Often, smaller aftershocks occur on and thus delineate the fault plane. If the earthquake is large enough, the finite time required for slip to progress along the fault causes variations in the waveforms observed at different directions from the fault; such directivity effects can be used to infer the fault plane.

More sophisticated techniques use the amplitudes and shapes of the seismic waves. These waves can be body waves, which travel through the Earth's interior, or surface waves, which propagate along paths close to the Earth's surface. The approach is to compare the observed body and surface waves with theoretical, or synthetic, waveforms computed for various source parameters, and find a model that best fits the data. Such analysis also gives information about the earthquake depths and rupture processes, which cannot be extracted from the first motions.

Figure 6 shows how body waves can be used to check the mechanism and estimate the depth. Synthetic seismograms were computed for various focal depths. The left-hand panel shows the expected timing and amplitudes of various arriving phases, and the right-hand panel shows the synthetic seismogram resulting from including the effect of the earthquake source and seismometer. The data are fitted well by a source at a depth of about 30 km.

Modelling surface waves can also help to resolve earthquake focal mechanisms and depths. Depending on the fault geometry, more energy is radiated in some directions than others. Figure 7 shows theoretical radiation patterns for the two kinds of surface waves, Love and Rayleigh, corresponding to several focal mechanisms with a fault plane striking north. The patterns are distinctive: a vertical strike-slip fault has two four-lobed patterns, such that Love-wave amplitude has maxima in the north, east, south, and west directions, whereas the Rayleigh-wave amplitude has maxima in the north-east, south-east, south-west, and north-west directions. In contrast, a dip-slip fault dipping at 45° has a four-lobed Love-wave pattern and a two-lobed Rayleigh-wave pattern. Such patterns can be generated for any fault geometry and compared with observations to find the best-fitting source geometry.

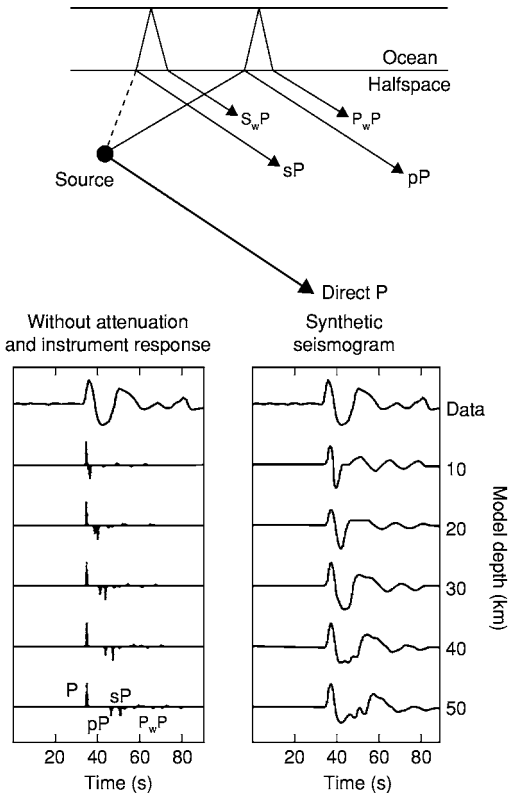


Figure 6 Body wave modelling for depth determination. Synthetic seismograms for an assumed fault geometry, including the effects of the seismometer and attenuation, are calculated for various depths. The data are best fitted by a depth of approximately 30 km. (Reproduced from Stein S and Wiens D (1986) Depth determination for shallow teleseismic earthquakes: methods and results. *Reviews of Geophysics and Space Physics* 24: 806 832.)

For earthquakes on land, additional information is derived by measuring the deformation of the Earth resulting from the earthquake using geodesy, the science of the Earth's shape. Most such techniques rely on detecting the motion of geological or manmade features or of geodetic monuments, which are markers in the ground. Until recently, these measurements were typically made by triangulation, which measures the angles between monuments using a theodolite, or trilateration, which measures distances with a laser. Vertical motion was measured using a precise level to sight on a distant measuring rod. However, the advent of geodetic methods using signals from space permits all three components of position to be measured with sub-centimetre precision. As a result, geodetic measurements before and after earthquakes can now determine

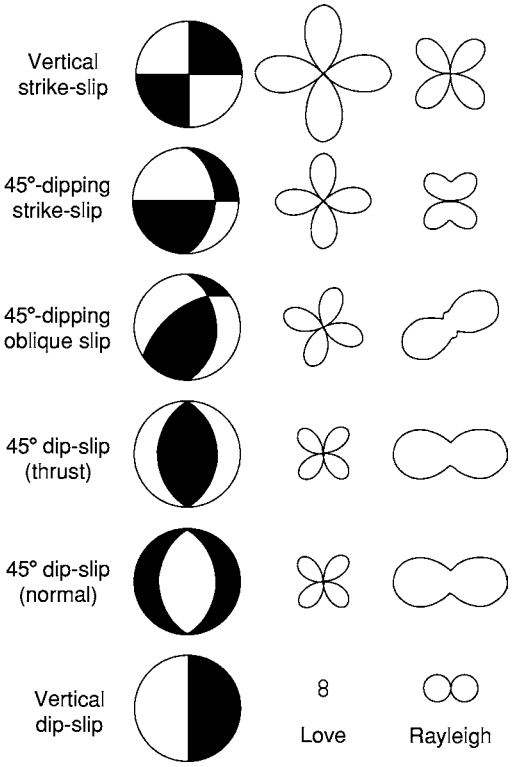


Figure 7 Focal mechanisms and surface wave amplitude radiation patterns for six fault geometries. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

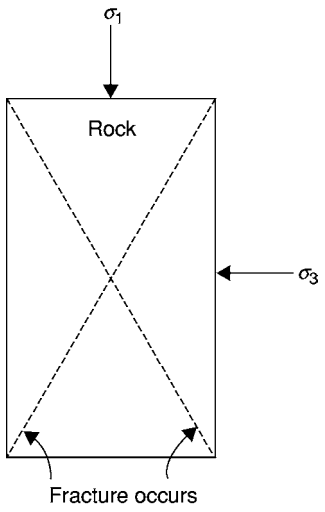


Figure 8 Schematic illustration of an experiment in which a rock sample is compressed along the direction of the maximum principal stress until fracture occurs. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

coseismic motion with high precision much more easily than was previously possible.

Faults and Stresses

The simplest theory for fracture predicts that faulting will occur on the plane on which the shear stress is highest. Although this is not exactly true, it provides an insight into the relation between fault orientations and regional tectonics. Consider an experiment in which a rock is compressed (Figure 8) with principal stresses $|\sigma_1| \geq |\sigma_2| \geq |\sigma_3|$. We expect fracture on the planes of maximum shear stress, which are 45° from the maximum and minimum principal stress axes and include the intermediate principal stress axis. Alternatively, if the experiment involves a situation known as uniaxial compression, where $|\sigma_1| \geq |\sigma_2| = |\sigma_3|$, failure should occur on any plane at 45° to the maximum principal stress (σ_1) axis. Experiments support the idea that fracture is controlled by shear stress, but in a more complicated way, so the fracture plane is often about 25° rather than 45° from the maximum principal stress direction.

For simplicity, however, we often assume that faults in the Earth form on the planes of maximum shear stress. The three basic fault geometries – strike-slip, normal, and thrust – are related to the stress axes (Figure 9). If the vertical principal stress is the most compressive, the fault dips at 45°, and normal faulting occurs. If, instead, the vertical principal stress is the least compressive, the fault geometry is the same but reverse or thrust faulting occurs. When the vertical principal stress is the intermediate principal stress, strike-slip motion occurs on a fault plane 45° from the maximum principal stress. Thus, the geometry of faults, which can be mapped geologically or inferred from seismograms of earthquakes, can be used to study stress orientations.

Fault Strength

Using seismic waves alone limits what we can learn about the earthquake process. The seismic waves radiated from an earthquake reflect the geometry of the fault and the motion on it, and so can give an excellent picture of the kinematics of faulting. However, they contain much less information about the actual physics, or dynamics, of faulting. Hence seismological results are combined with experimental and theoretical studies of rock friction and fracture to explore the physics of faulting.

Consider the strain that results from compressing a rock specimen. For small stresses the resulting strain is proportional to the applied stress, so the material is purely elastic (Figure 10A). Once the stress reaches the rock's fracture strength, σ_f, the rock breaks. Such brittle fracture is the simplest model

for an earthquake on a fault. Other materials show a change in the stress–strain curve for increasing stresses (Figure 10B). For stresses less than the yield stress, σ_o, the material acts elastically. If the stress is released, the strain returns to zero. However, for stresses greater than the yield stress, releasing the stress relieves the elastic portion of the strain but leaves a permanent deformation (Figure 10C). If the material is restressed, the stress–strain curve now includes the point of the permanent strain. The portion of the curve corresponding to stress above the yield stress is called plastic deformation, in contrast to the elastic region where no permanent deformation occurs. Materials showing significant plasticity are called ductile.

At low pressures rocks are brittle, but at high pressures they behave ductilely, or flow. Figure 11 shows experiments where a rock is subjected to a compressive stress that exceeds a confining pressure. For confining pressures less than about 4 Kb the material behaves brittlely – it reaches the yield strength and then fails. For higher confining pressures the material flows ductilely. These pressures occur not far below the Earth's surface – each 3 km increase in depth corresponds to a 1 Kb increase in pressure, so 8 Kb is reached at about 24 km.

How fault behaviour varies with depth is often discussed in terms of the strength – the maximum difference between the horizontal and vertical stresses that the rock can support. At shallow depths rocks fail either by brittle fracture or by frictional sliding on pre-existing faults, and strength increases with depth. However, at greater depths, rocks deform ductilely, as described by flow laws, which show that strength decreases as the temperature increases. This temperature-dependent behaviour is the reason that the cold lithosphere forms the planet's strong outer layer and that earthquakes occur only to a given depth. These variations are described by strength plots known as strength envelopes.

Figure 12 shows strength envelopes appropriate for old oceanic lithosphere and a stable continental interior. In the frictional region, curves are shown for various values of the ratio of pore pressure to lithostatic pressure, because higher pore pressures result in lower strengths. Ductile flow laws are shown for quartz and olivine, minerals often used as models

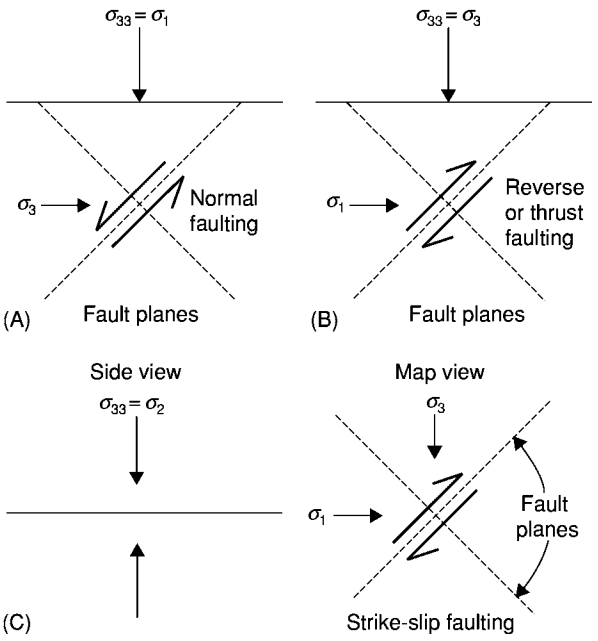


Figure 9 Stress fields associated with the three types of faulting, assuming that the earthquake occurred on a plane of maximum shear stress. (A) Normal, (B) reverse, (C) and strike slip faulting involve different orientations of the principal stresses. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

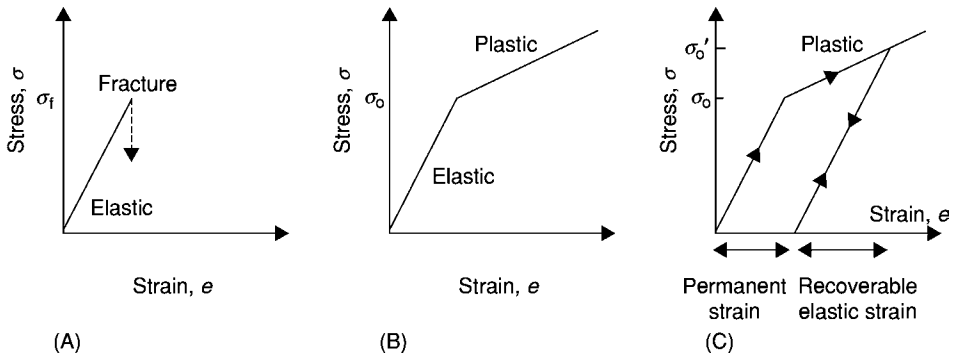


Figure 10 (A) Stress strain curve for a material that is perfectly elastic until it fractures when the applied stress reaches σ_f. (B) Stress strain curve for a material that undergoes plastic deformation when the stress exceeds a yield stress, σ_o. (C) Permanent strain results from plastic deformation when the stress is raised to σ_o' and released. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

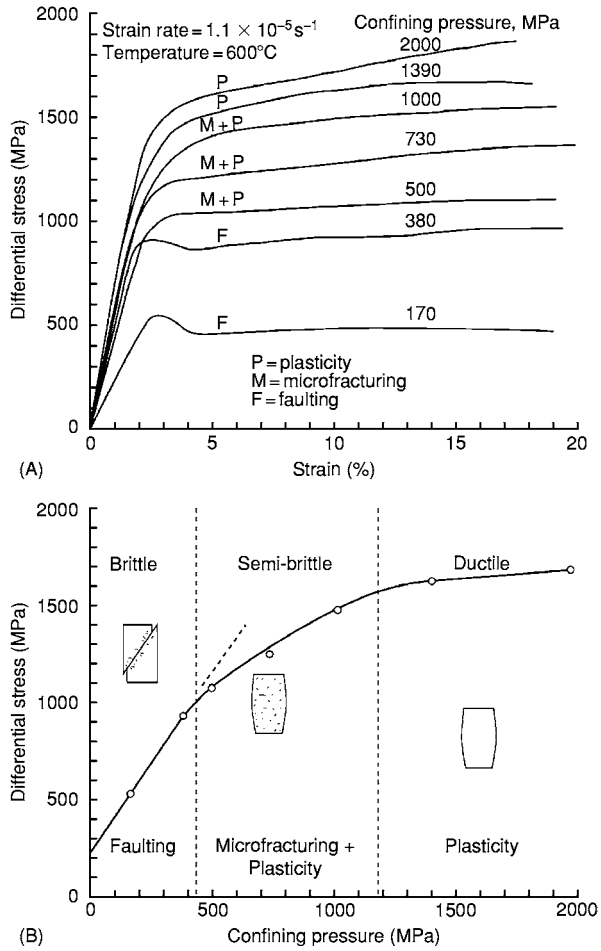


Figure 11 Results of an experiment in which rocks were subjected to a compressive stress greater than the confining pressure. (A) Differential stress strain curves (c.f. **Figure 10**) for various confining pressures. (B) At low (less than 4 Kb) confining pressures the material fractures and its strength increases with pressure. At higher pressures the material is ductile and its strength increases only slowly with pressure. A semi brittle transition regime, in which both microfractures and plasticity occur, separates the brittle and ductile regimes. (Reproduced from Kirby SH (1980) Tectonic stresses in the lithosphere: constraints provided by the experimental deformation of rocks. *Journal of Geophysical Research* 85: 6353–6363.)

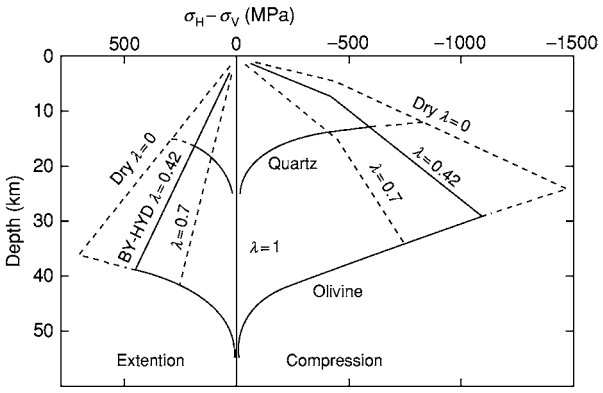


Figure 12 Strength envelopes as a function of depth for various values of λ, the ratio of pore pressure to lithostatic pressure. At shallow depths, strength is controlled by brittle fracture; at greater depths ductile flow laws predict rapid weakening. In the ductile flow regime, quartz is weaker than olivine. In the brittle regime, the lithosphere is stronger in compression (right side) than in extension (left side). (Reproduced from Brace WF and Kohlstedt DL (1980) Limits on lithospheric stress imposed by laboratory experiments. *Journal of Geophysical Research* 85: 6248–6252.)

for continental and oceanic rheologies, respectively. Strength increases with depth in the brittle region owing to the increasing normal stress, and then decreases with depth in the ductile region owing to increasing temperature. Hence strength is highest at the brittle–ductile transition. Strength decreases rapidly below this transition, so the lithosphere should have little strength at depths greater than about 25 km in the continents and 50 km in the oceans. As a result, the limiting temperature for continental seismicity is lower than for oceanic seismicity.

Faulting and Rock Friction

It is natural to assume that earthquakes occur when tectonic stress exceeds the strength, so steady motion across a plate boundary would give a series of successive earthquakes at regular intervals (**Figure 13**). However, the time between earthquakes on plate boundaries varies even though the plate motion causing the earthquakes is steady. Some of the variation may be due to the intrinsic randomness of the failure process, such that some small ruptures cascade into large earthquakes whereas others do not. Another cause of the variation may be features of rock friction.

Interesting insight emerges from considering an experiment in which stress is applied to a faulted rock, where motion occurs once the stress reaches a certain level. As stress is reapplied, this pattern of jerky sliding and stress release continues. This stick–slip pattern looks like a laboratory version of earthquakes on a fault: as the fault is loaded by tectonic stress, occasional earthquakes occur. The analogy is strengthened by the fact that at higher temperatures (about 300°C for granite) stick–slip does not occur. Instead, stable sliding occurs, in much the same way as earthquakes do not occur at depths where the temperature exceeds a certain value. Stick–slip results from a familiar phenomenon: it is harder to start an object sliding against friction than to keep it sliding. This is because the static friction stopping sliding exceeds the dynamic friction that opposes motion once sliding starts.

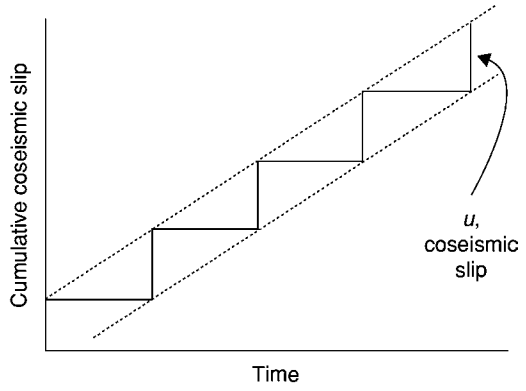


Figure 13 Slip history for an idealized earthquake cycle on a plate boundary, in which all earthquakes have the same coseis mic slip. (Reproduced from Shimazaki K and Nakata T (1980) Time predictable recurrence model for large earthquakes. *Geophysical Research Letters* 7: 279 282.)

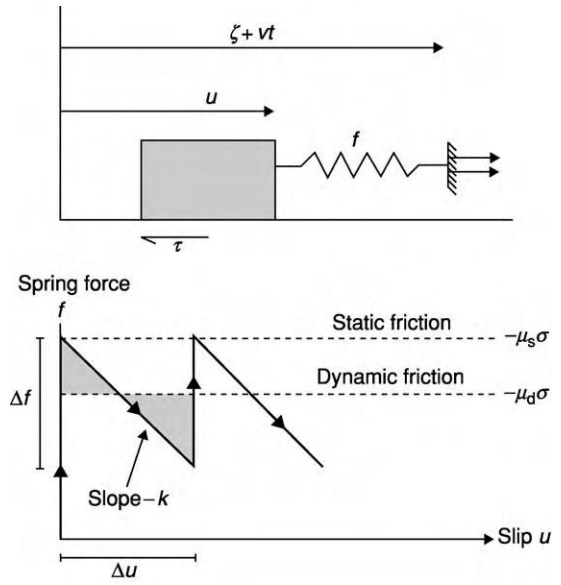


Figure 14 A simple spring and block analogue illustrating stick slip as a model for earthquakes. (Reproduced from Stein S and Wysession M (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

To gain an insight into stick–slip as a model for earthquakes, consider the experiment illustrated in Figure 14. If an object is pulled across a table with a rubber band, jerky stick–slip motion occurs. This situation can be modelled by assuming that a block is loaded by a spring that applies a force, f, that is proportional to its stiffness, k, and extension. If loading results from the spring's far end moving at velocity v, the spring force increases with time. The block starts sliding once the spring force exceeds the frictional force $\mu_s\sigma$ where μ_s is the static friction coefficient and σ is the normal stress due to the block's weight. Once sliding starts, the friction drops to its dynamic value μ_d, and the driving force decreases as the spring shortens, until it becomes less than the friction force. The block slows and eventually stops once the shaded area above the spring-force line equals that below the line, or when the work done accelerating the block equals that which decelerated it. If the spring end continues to move, loading continues until the spring force again equals the static friction force and another slip event occurs.

Laboratory experiments show that the difference between static and dynamic friction is more complicated than is assumed in this simple model. We can think of the lower dynamic friction as showing either velocity weakening, decreasing as the object moves faster, or slip weakening, decreasing as the object moves further. Frictional models, called rate-dependent friction and state-dependent friction, with a variable coefficient of sliding friction are used to describe these effects. Velocity weakening permits earthquakes to occur by stick–slip, whereas for velocity strengthening stable sliding is expected. Laboratory results show that for granite the transition occurs at about 300°C, which should be the limiting temperature for earthquakes. Thus, the frictional model predicts a maximum depth for continental earthquakes that is similar to that predicted by the rock-strength arguments.

These results can be used to simulate the earthquake cycle. Figure 15 shows the slip history as a function of depth and time for a model in which a strike-slip fault is loaded by plate motion. The fault has rate- and state-dependent frictional properties such that stick–slip occurs above 11 km. From time A to time B, stable sliding occurs at depth and a little precursory slip occurs near the surface. The earthquake causes 2.5 m of sudden slip at shallow depths, as shown by the curves for times B and B′. As a result, the faulted shallow depths 'get ahead' of the material below, loading that material and causing postseismic slip from time B to time F. Once this is finished, the 93-year cycle starts again with steady stable sliding at depth.

Such models replicate many aspects of the earthquake cycle. An interesting difference, however, is that the models predict earthquakes at regular intervals, whereas earthquake histories are quite variable. Some of the variability may be due to the effects of earthquakes on other faults or other segments of the same fault. Figure 16 shows this idea schematically for the block model. Assume that after an earthquake cycle the compressive normal stress is reduced. This 'unclamping' reduces the frictional force resisting sliding, so it takes less time for the spring force to rise to the level needed for the next slip event. Conversely,

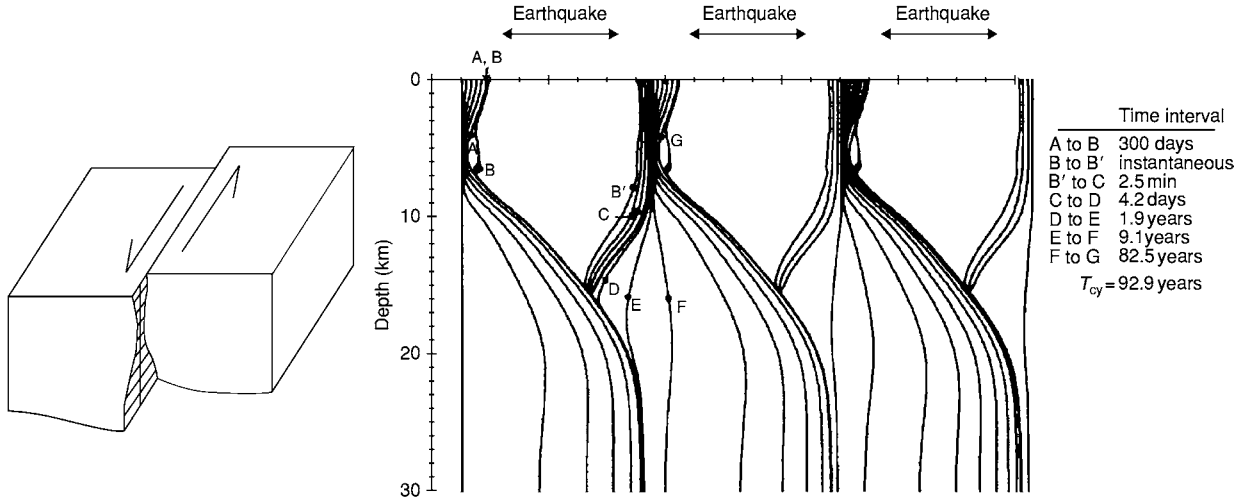


Figure 15 Earthquake cycle for a model in which a strike slip fault with rate and state dependent frictional properties is loaded by plate motion. The slip history for three cycles as a function of depth and time is shown by lines representing specific times. Steady motion occurs at depth, and stick slip occurs above 11 km. (Reproduced from Tse ST and Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *Journal of Geophysical Research* 91: 9452 9472.)

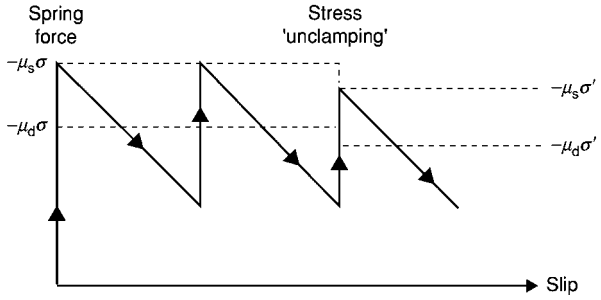


Figure 16 Modification of the block model shown in **Figure 14** to include the effects of changes in normal stress. Reduced normal stress reduces the frictional force, 'unclamping' the fault and decreasing the time to the next slip event. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure.* Blackwell Publishing.)

increased compression 'clamps' the block more, increasing the time to the next slip event. This analogy implies that earthquake occurrence on a segment of a fault may reflect changes in the stress on the fault resulting from earthquakes elsewhere. Some earthquake observations provide support for this idea.

See Also

Earth: Crust. **Engineering Geology:** Seismology; Natural and Anthropogenic Geohazards. **Plate Tectonics**. **Tectonics:** Earthquakes; Neotectonics.

Further Reading

Aki K and Richards PG (2002) *Quantitative Seismology.* Sausalito: University Science.

Brace WF and Kohlstedt DL (1980) Limits on lithospheric stress imposed by laboratory experiments. *Journal of Geophysical Research* 85: 6248 6252.

Kirby SH (1980) Tectonic stresses in the lithosphere: con straints provided by the experimental deformation of rocks. *Journal of Geophysical Research* 85: 6353 6363.

Kirby SH and Kronenberg AK (1987) Rheology of the lithosphere: selected topics. *Reviews of Geophysics* 25: 1219 1244.

Lay T and Wallace TC (1995) *Modern Global Seismology.* New York: Academic Press.

Moores EM and Twiss RJ (1995) *Tectonics*. New York: W H Freeman.

Scholz CH (1990) *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge University Press.

Shearer PM (1999) *Introduction to Seismology.* Cambridge: Cambridge University Press.

Shimazaki K and Nakata T (1980) Time predictable recur rence model for large earthquakes. *Geophysical Research Letters* 7: 279 282.

Stein S and Wiens D (1986) Depth determination for shallow teleseismic earthquakes: methods and results. *Reviews of Geophysics and Space Physics* 24: 806 832.

Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure.* Malden, MA: Blackwell Publishing.

Tse ST and Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *Journal of Geophysical Research* 91: 9452 9472.

Folding

J W Cosgrove, Imperial College London, London, UK

Introduction

A fold is defined as "...a curved arrangement of originally parallel surfaces ..." and a large range of terms have been used to describe geological folds. These including folds, flexures, inflexions, bendings, plications, undulations, and crenulations. Although today most geologists use the term fold for buckle folds and many reserve flexure for layer deflections caused by bending, the terms are also used as synonyms.

Folds form on all scales, from those with a wavelength of a few mm that can only be seen in rock thin sections under the microscope, to folds with wavelengths in excess of 10 km. The mechanisms of formation are independent of size; however, for large-scale folds it is important to take into account the effect of gravity when analysing their folding behaviour.

Folds in rocks form under a wide variety of conditions. For example, folds can occur at all depths in the crust from early, near-surface folding linked to the slumping of water saturated, uncemented sediments down continental slopes, to folding of rocks under the high pressures and temperatures encountered in the lower crust. Folding of rocks occurs by various processes, the three most important being buckling, bending, and flowing. Buckling, which is the most common type of fold, requires the rock to possess a mechanical anisotropy, usually layering, and for the maximum compressive stress to act parallel or subparallel to the layering. In contrast, bending is defined as a transverse deflection of a layer or beam by a transverse couple. Flow folding occurs during the flow of a material such as lava, salt, or ice. No mechanical anisotropy is necessary for this type of folding, only passive marker bands that reveal the flow patterns within the material.

Although folds are structures that are characteristic of ductile deformation, they are often found in association with fractures (*see* **Tectonics:** Fractures (Including Joints)). The fractures result from the same stress field responsible for folding and the resulting fracture patterns, which are controlled by the mechanisms by which the folds form, can play an important role in hosting mineralization and in the storage of fluids such as water and hydrocarbons.

Fold Geometry

The main geometric parameters relating to the three-dimensional geometry of a fold are shown in Figure 1A. However, although folds are three-dimensional structures they are most commonly exposed as two-dimensional sections on joint surfaces. Consequently, their detailed classification is based on the geometry of a section. Because of the dependence of fold geometry on the orientation of the section (Figure 1B), a particular section is used to determine the geometry, namely the profile section, i.e., the section at right angles to the fold hinge. Historically, the problem of describing layer shape developed around two geometrical models, the parallel fold and the similar fold (Figure 2A and B). As the name implies, in parallel folds the orthogonal thickness of the layer remains constant. In similar folds, the layer thickness, parallel to the hinge surface remains constant. Although similar folds show considerable variations in layer thickness, this type of folding can produce folds which can extend indefinitely in the profile section, whereas with parallel folds this is not possible (Figure 2A and B).

As can be seen from Figure 2C, these two fold types are examples from a complete spectrum of possible geometries. The spectrum has been divided into five classes based on the pattern of dip isogons, i.e., lines

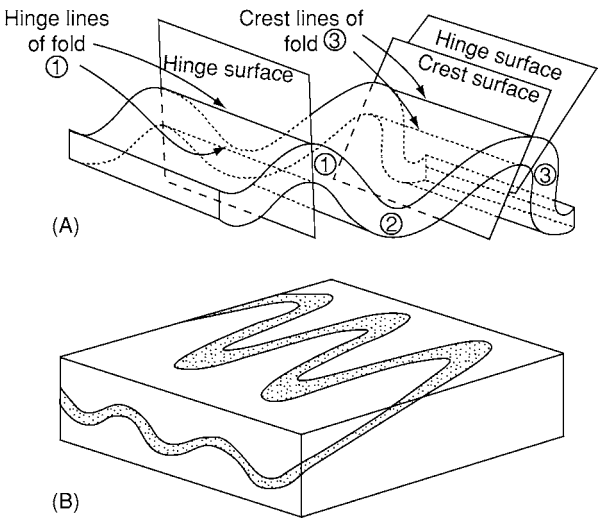


Figure 1 (A) Geometric features of a fold. Fold 1 is symmetric and folds 2 and 3 asymmetric. (B) Diagram showing the dependence of fold outcrop pattern on the orientation of the plane of exposure.

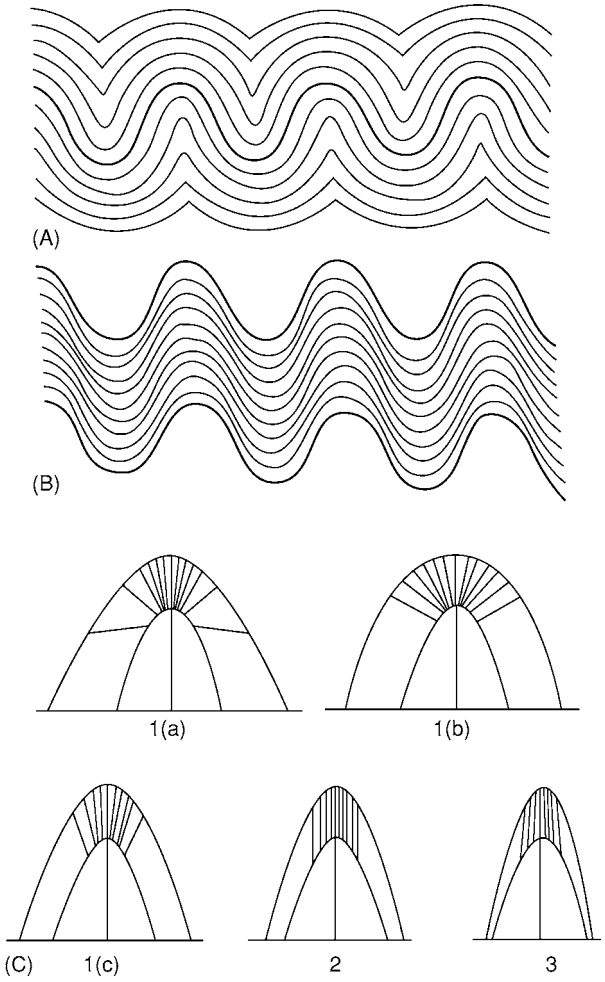


Figure 2 (A) parallel and (B) similar folds. (After Van Hise (1894).) (C) Classification of fold profiles using dip isogon patterns. 1(a) strongly convergent, 1(b) parallel, 1(c) weakly convergent, (2) similar, and (3) divergent (Ramsay (1967)).

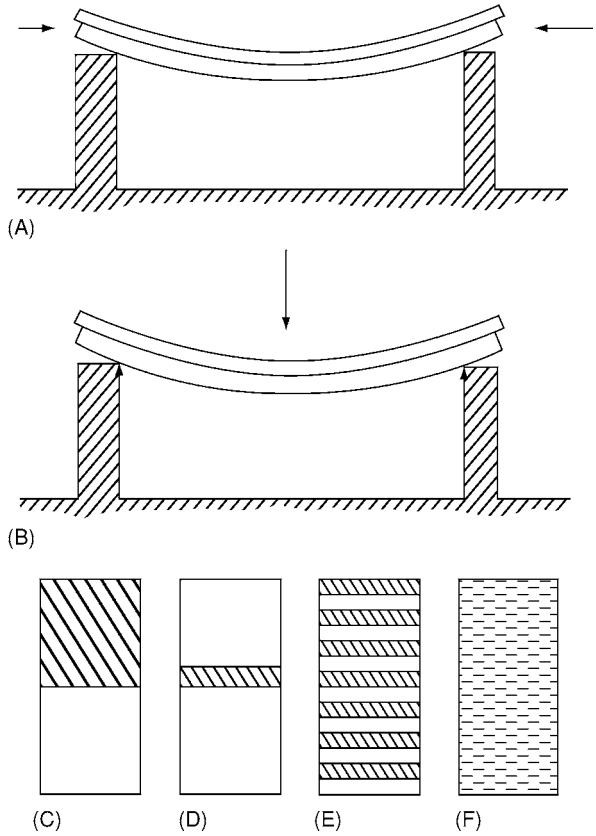


Figure 3 The orientation of the principal compression for (A) buckling and (B) bending of the layers. (C) an interface between two unlike materials; (D) a single layer; (E) a multilayer; and (F) a mechanically anisotropic material. The buckling behaviour of these systems is discussed in the text.

joining points of equal dip on the two surfaces defining the folded layer. The parallel folds and similar folds are class 1b and 2, respectively.

Mechanisms of Folding

The two most important mechanisms by which folds form in rocks are buckling, the result of compression parallel or sub-parallel to the rock layering and bending, the result of compression at a high angle to the layering, (Figure 3A and B, respectively).

Buckle Folds

The most commonly formed folds in the Earth's crust are buckle folds. In order for these folds to form, the rock must possess a mechanical anisotropy. This is generally a planar mechanical anisotropy because a variety of geological processes give rise to such materials. Rocks can be intrinsically anisotropic because of the process of their formation such as, for example, a bedded sedimentary succession, or the anisotropy can be induced as a result of subsequent deformation and metamorphism during which time they can develop planar and linear mineral fabrics as they are converted to slates, phyllites, schists, and gneisses. Buckling systems can be sub-divided into four groups, namely folds formed by: (i) the buckling of a single interface, (Figure 3C); (ii) the buckling of two interfaces, which define a single layer in a matrix (Figure 3D); (iii) the buckling of several layers (Figure 3E); and (iv) the buckling of a mineral fabric such as a slate or schist (Figure 3F).

Interface Buckling

Interface buckling occurs on many scales, from large buckles that form at the boundary between the Mesozoic cover rocks and older basement in the French Alps as a result of the collision of the African and

European plates (Figure 4A), to small-scale examples directly observable in the field (Figure 4B) and formed in analogue models in the laboratory (Figure 4C). The folds start as symmetric sinusoidal deflections but as they amplify change their geometry into the marked cusp geometry seen in Figure 4. The cusps always point into the stronger of the two materials.

Single Layer Buckling

Single layers occur commonly in nature, for example, an isolated sandstone or limestone bed in a thick shale or marl sequence or a sheet of igneous rock intruded into an unlayered matrix (Figure 5A), and they are commonly observed to buckle. Theoretical analyses

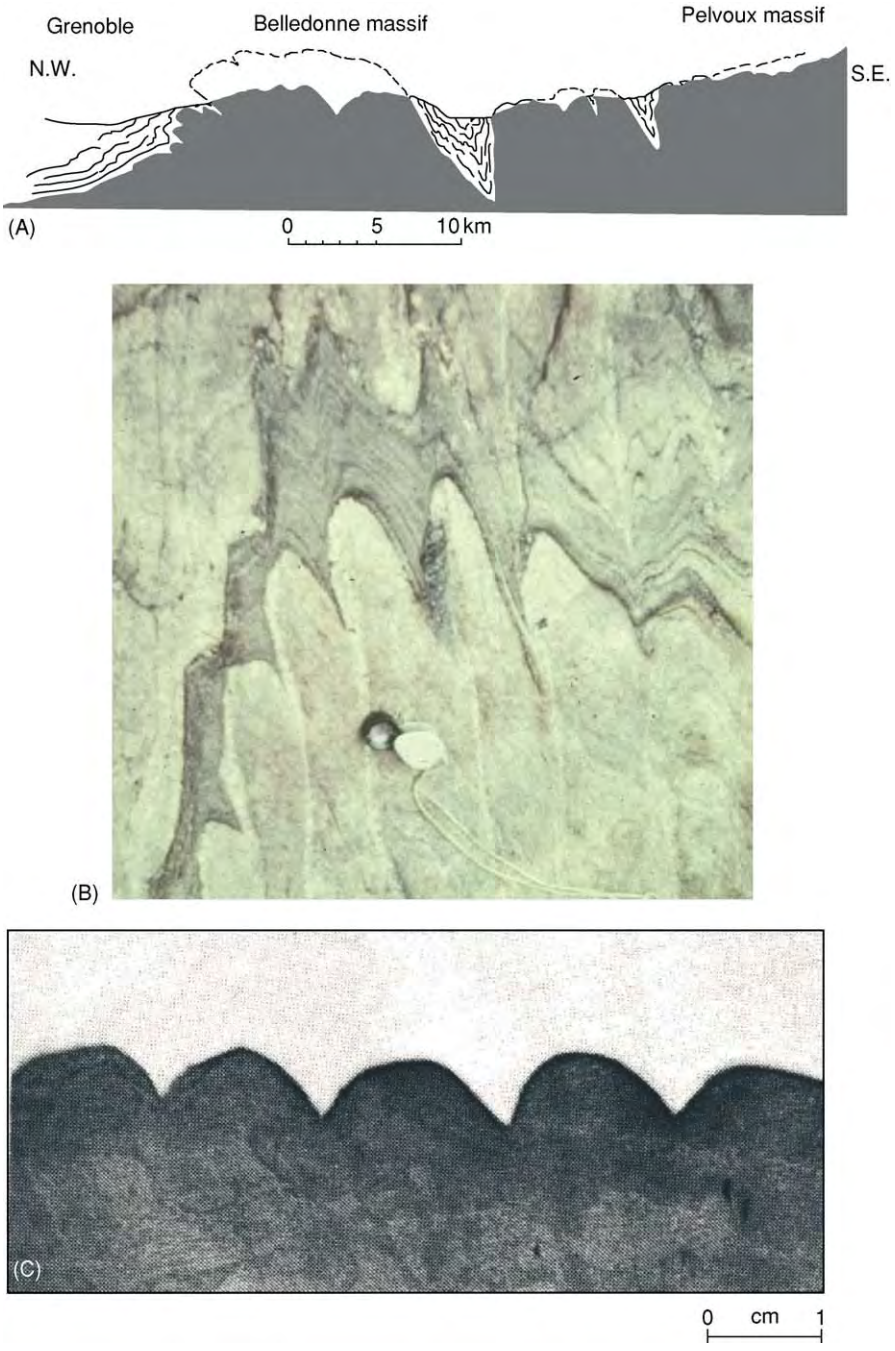


Figure 4 (A) The cuspate interface between the strong Hercynian basement rocks (black) and the weaker Mesozoic cover rocks, caused by the horizontal compression linked to the collision of the African and European plates. (B) Cusp structures formed at the interface between quartz rich (light) and mica rich (dark) bands, Loch Monar, Scotland. (C) Cusps generated by compression parallel to the interface between strong (black) and weak (light) plasticine.

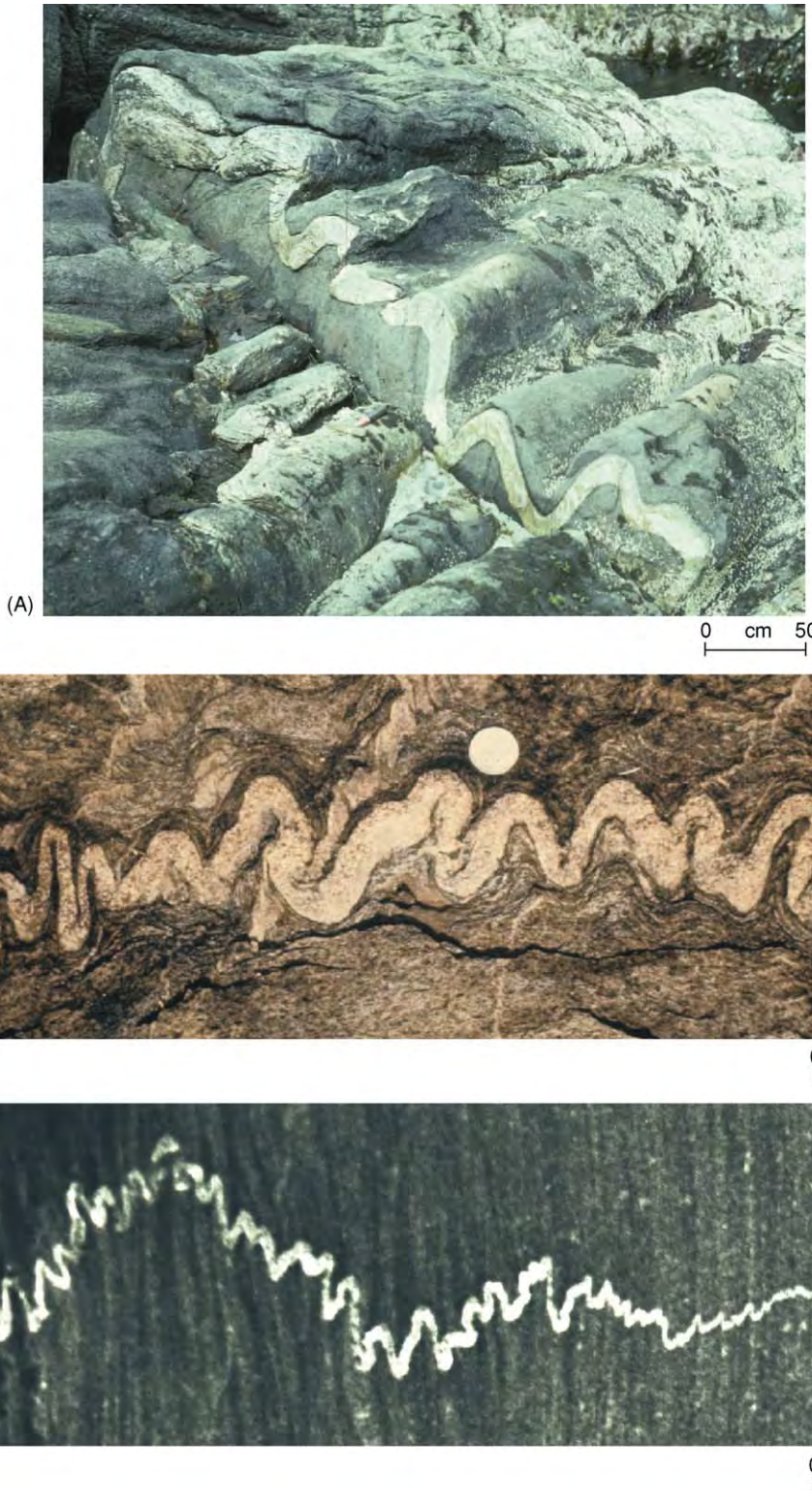


Figure 5 (A) A folded aplite dyke cutting a dolerite body, Outer Hebrides, Scotland. (B) A folded single layer, in which the folds are isoclinal. No further shortening of the layer can occur by limb rotation. Further compression may result in homogeneous flattening of the folds or in the folding of the layer into larger wavelength folds (see C). (C) A tapered quartz vein showing buckling on two scales. On both scales, the wavelength is seen to decrease with decrease in layer thickness.

predict and field observations confirm that there is a direct relationship between the wavelength (W) of folds that develop and the thickness (t) of the buckling layer (Figure 5C).

$$W = 2\pi t(\mu_1/6\mu_2)^{1/3} \quad [1]$$

The single layer buckling equation shows that in addition, the wavelength is also controlled by the ratio of the strength of the layer and matrix (μ_1/μ_2). The impact of the strength ratio (competence contrast) on fold style can be seen by rearranging eqn 1.

$$W/t = 2\pi(\mu_1/6\mu_2)^{1/3} \quad [2]$$

The wavelength/thickness ratio is determined by the relative strength of the layer and matrix. Folds with a range of W/t ratios are shown in Figure 6, which shows the control of the competence contrast between the layer and the matrix on the geometry of the buckles that form.

Once buckling has been initiated, layer shortening can continue by rotation of the fold limbs until the limbs become parallel to each other (i.e., the folds are isoclinal, Figure 5C), and no further shortening by limb rotation can occur. Shortening may continue by homogeneous flattening of the buckled layer and matrix, or alternatively continued compression may cause the buckled layer to buckle again. It has an effective thickness that is considerably greater than the original thickness of the layer and so, in accordance with the buckling equation eqn 1, will buckle with a larger wavelength (Figure 5C).

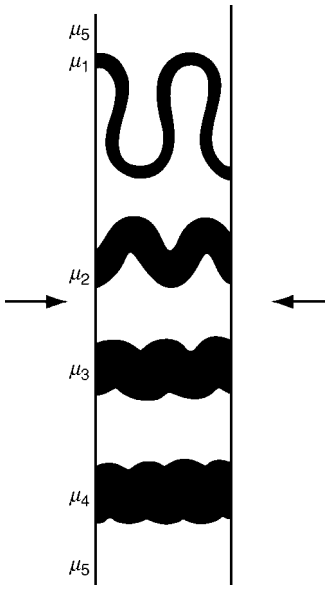


Figure 6 The effect of the ratio of the strength of the layer and the matrix on fold style. As the contrast decreases the wave length/thickness ratio also decreases. (After Ramsay (1982).)

Multilayer Buckling

The most common type of geological layering is that of a multilayer, i.e., a succession of different layers. Often the multilayer is made up of the regular alternation between two or three rock types. The regularity of these multilayers reflects the processes by which they were formed. For example, turbidites are formed as a result of the periodic, fluid-induced collapse of a sedimentary accumulation on the edge of the continental shelf. The fluidised sediments flow as dense turbidity currents down the continental slope, depositing first the heavy sands followed by the slower settling out of the fine shale particles. In this way a sequence of alternating sandstones and shales can be built up to form a turbidite.

The buckling behaviour of multilayers is controlled by the spacing of the strong units it contains. Figure 7A and B are both physical multilayers. However, their buckling behaviours are very different. The multilayer shown in Figure 7A behaves mechanically as a series of single layers, each layers folding

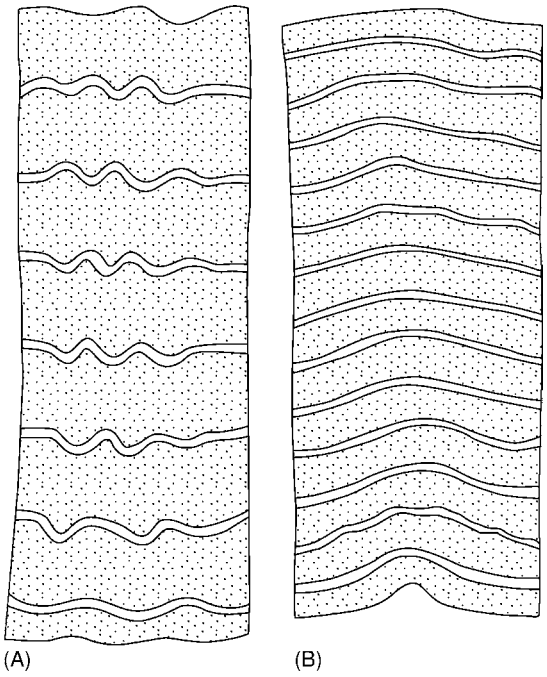


Figure 7 (A) and (B) show two compressed multilayers of alternating competent (white) and incompetent (stippled) rubber layers. The widely spaced layers (A) behave mechanically as single layers and form disharmonic folds; the more closely spaced layers behave mechanically as a multilayer and all layers form the same wavelength. (After Ramberg (1963).)

according to the single layer buckling equation, (eqn 1). This results in 'disharmonic folding'. In contrast, the multilayer shown in Figure 7B behaves mechanically as a multilayer, i.e., all the layers develop the same wavelength and amplitude. Such folding is called 'harmonic folding'.

These two types of buckling behaviour can be readily explained by considering the strain that develops in the matrix around a single, competent (i.e., strong with respect to the surrounding matrix) layer as it buckles. The zone of disturbance on each side of the buckling layer is known as the zone of contact strain. If the competent layers of a multilayer are sufficiently far apart for there to be no significant overlap of their zones of contact strain, then each layer will buckle as a single layer. If, however, the zones of contact strain of adjacent competent layers do significantly overlap, the layers can no longer buckle independently of each other. The zones of contact strain and associated zones of contact stress of adjacent layers must be compatible and, as a result, all the layers are subjected to the same stress field and develop the same wavelength.

In order to determine how close the competent layers of a multilayer must be for multilayer as opposed to single layer buckling to occur, it is necessary to know how far the zone of contact strain extends away from the layer into the matrix. For a viscous matrix it is found that the disturbance has died down to approximately 1% of its maximum value at a distance one wavelength from the layer (Figure 8).

The Buckling of Anisotropic Materials

The most complex buckling occurs in materials possessing a pervasive mechanical anisotropy. The anisotropy may be an intrinsic property of the rock resulting, for example, from the bedding parallel alignment of clay particles in a shale or induced during metamorphism when, for example, a slate or schist is formed from a mudstone. Theoretical studies and experimental work on such materials shows that there is a range of structures that can form when they are compressed parallel to the mineral fabric or layering. The two end members of this range are upright folds and box folds (Figure 9). The type of structure that forms is determined by the mechanical anisotropy of the material. As the anisotropy increases so the upright folds (Figure 9A), give way to folds with gently diverging axial planes (Figure 9B) and finally to box folds (Figure 9C).

It should be noted that if a geological multilayer has a sufficiently high mechanical anisotropy, it will buckle to form a box fold. An example of this is shown in Figure 10A, where a box fold with a wavelength of several 100 m has developed in Carboniferous turbidites from south-west England.

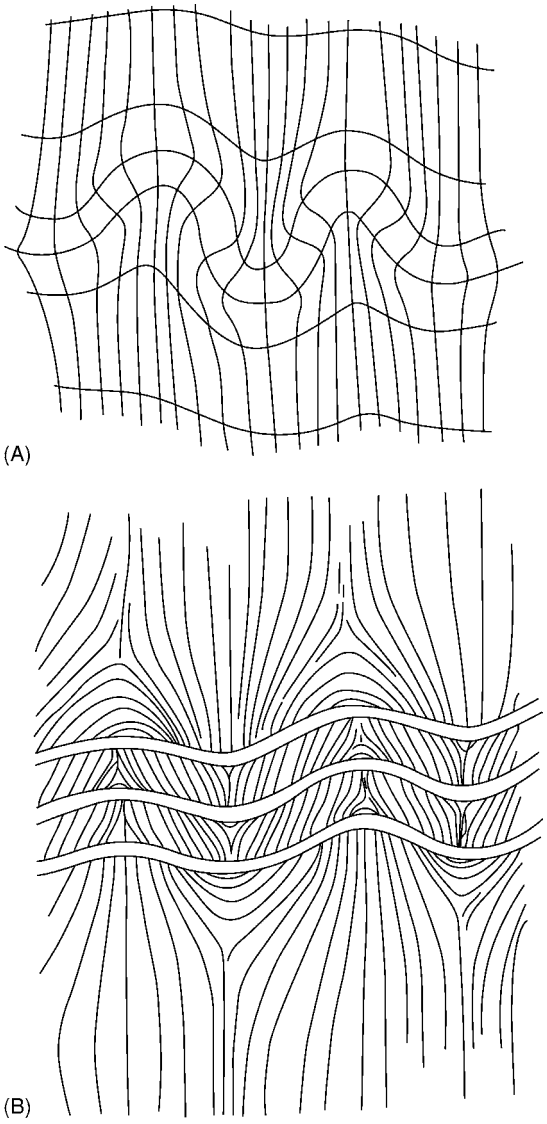


Figure 8 (A) Experimentally produced buckles in viscous materials showing the disturbance of the matrix (the zone of contact strain) by the buckling single layer (B) Folded multilayer where the proximity of the layers caused an overlap of the zones of contact stain and therefore the formation of multilayer buckles.

Experimental Work on Folds

One of the disadvantages of theoretical studies of folding is that they are often only valid for the first increment of buckling. Once buckling has been initiated, the assumptions of the analysis are violated and the theory cannot be used to predict the way in which the fold amplifies into a finite fold. In contrast, experimental work on models made from rock analogue materials such as gelatine, are ideal for the study of the amplification of folds. In addition, unlike

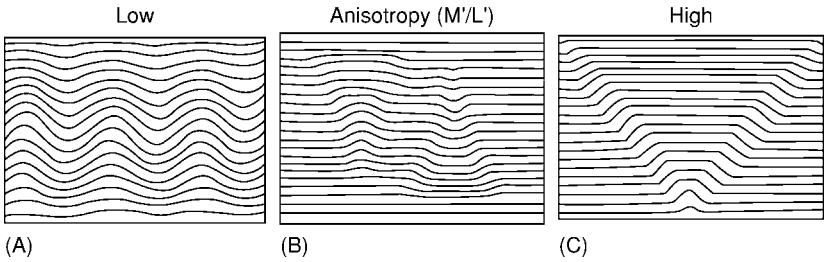


Figure 9 The structures that can form when a mechanically anisotropic material such as a mineral fabric of sedimentary layering is compressed parallel to the layering. Depending upon the anisotropy they range from upright folds with axial planes normal to the principal compression to box folds where the axial planes are inclined to the compression.

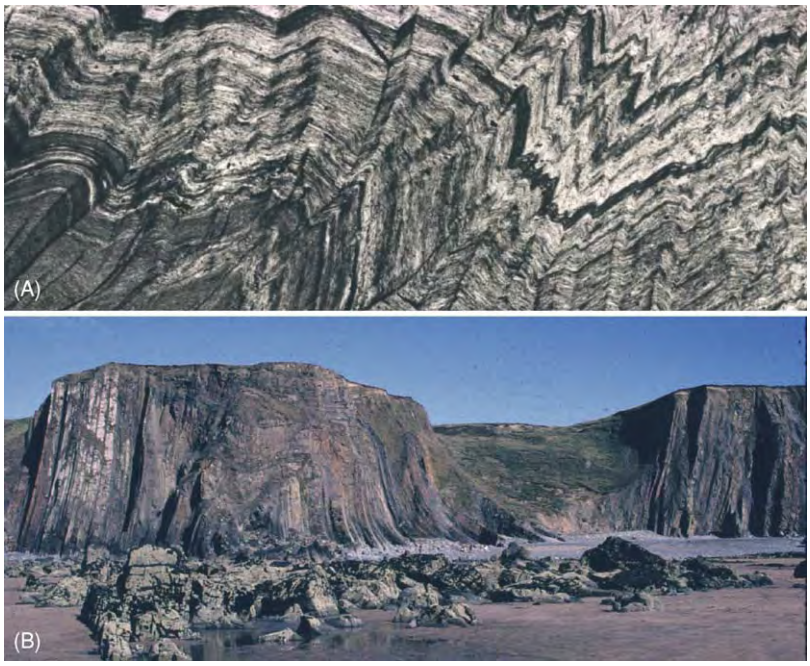


Figure 10 (A) Small scale box fold formed in a micaceous fabric and (B) large scale box fold formed by the buckling of a turbidite.

the theoretical treatments of folding which assume that the layered system is stressed instantaneously (i.e., that the process of stressing the material plays no part in the buckling), the experiments show that the process of stressing the layered system can influence the process of folding. Such experiments show that folds that form during the compression of the models do not generally develop synchronously. They form in a serial manner, either one after the other where the amplification of one fold stimulates the initiation and amplification of another next to it (Figure 11A), or one after the other at random positions within the model. Figure 11B shows such randomly positioned folds in the Jura mountains of Switzerland.

In the complex multilayers that occur in nature, it is often found that all types of buckling occur in the same multilayer. For example, in the multilayer shown in Figure 12, which is a finely laminated evaporite, examples of single layer buckling can be seen (layers 1 and 2) where the wavelength is determined by the thickness of the layers. The two layers are far enough apart for them to fold independently and they have done so producing an example of disharmonic folding. In area 3 of the multilayer, examples of multilayer folding where the folds have axial planes oriented at right angles to the layering can be seen (Figure 7B and 9A) and in area 4, box folds have formed with axial planes inclined to the layering (Figure 9C, 10A and B).

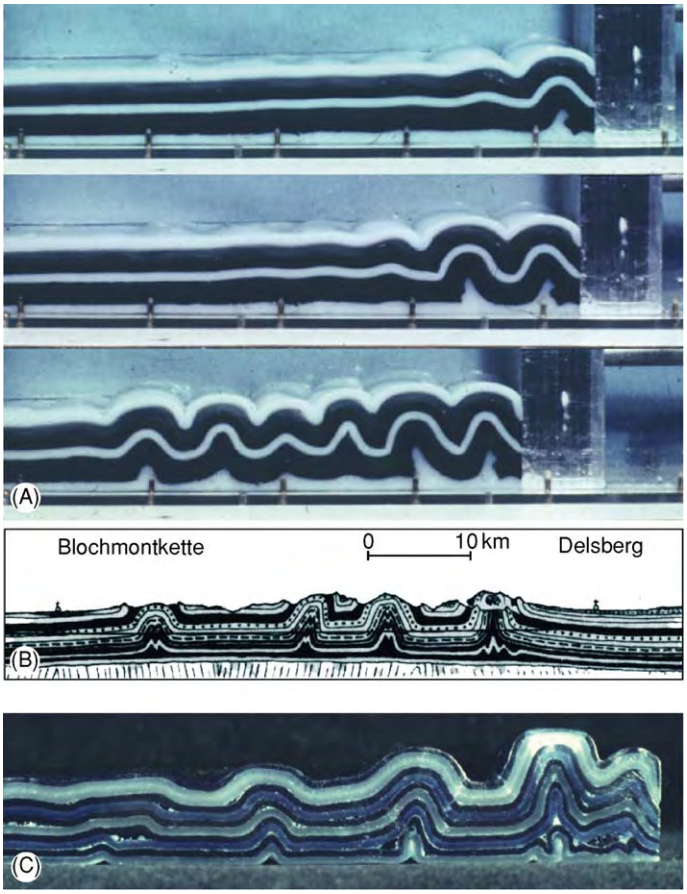


Figure 11 (A) The serial development of folds in a lubricated gelatine multilayer caused by a piston moving from right to left. (B) Isolated folds in the Jura mountains Switzerland. (C) Isolated folds in a lubricated wax multilayer.

Figure 12 Harmonic and disharmonic folding in a specimen of the Castile and Todilto evaporites, New Mexico, USA (c.f. layers 1 and 2). Some of the folding (layers 1 and 2), can be described by the theory of single layer buckling, some by the theory of multilayer buckling, (3) and some (4) by the theory of buckling of an anisotropic material.

Three-Dimensional Geometry of Buckle Folds

The three-dimensional geometry and spatial organization of buckle folds has been studied primarily by field observations and analogue modelling. These studies show that buckle folds have a periclinal geometry, i.e., have the form of an elongate dome, basin, or saddle (Figure 20A). The geometry of a pericline is often given in terms of the ratio of its half wavelength and its hinge length. This is termed the aspect ratio and although it will increase as the fold amplifies, it

is found that the majority of buckle folds in the upper crust have ratios of between 1:5 and 1:10. Typical geometries of geological folds are shown in Figure 13.

Bending

Bending is the term used to describe the flexuring of a layer induced by a compression acting at a high angle to the layering (Figure 1B). Geological flexures that are the result of bending are known as drape folds or forced folds and are frequently formed when sediments, which cover a more rigid basement, flex in response to components of vertical movement along basement faults (Figure 14). This may be normal movement, in which case the flexing of the layering involves layer parallel extension (Figure 14A), or reverse movement, in which case the folds that result will involve an element of buckling (Figure 14D).

A forced fold is defined as a fold in which the final overall shape and trend are dominated by the shape of some forcing member below. These are frequently fault blocks, movements of which produce linear fault scarps which in turn produce linear forced folds with an aspect ratio much higher that that associated with buckle folds and with different spatial organizations.

The two types of basement faults linked to the formation of forced folds are dip-slip faults (either normal or reverse Figure 14A and D respectively). These faults produce fault scarps at the basement–cover interface over which the forced folds form. When the third type of faulting, i.e., strike-slip or wrench faulting occurs in the basement, no fault scarp is produced and consequently no forced folds are produced in the overlying strata. However, movement on basement strike-slip faults can give rise to

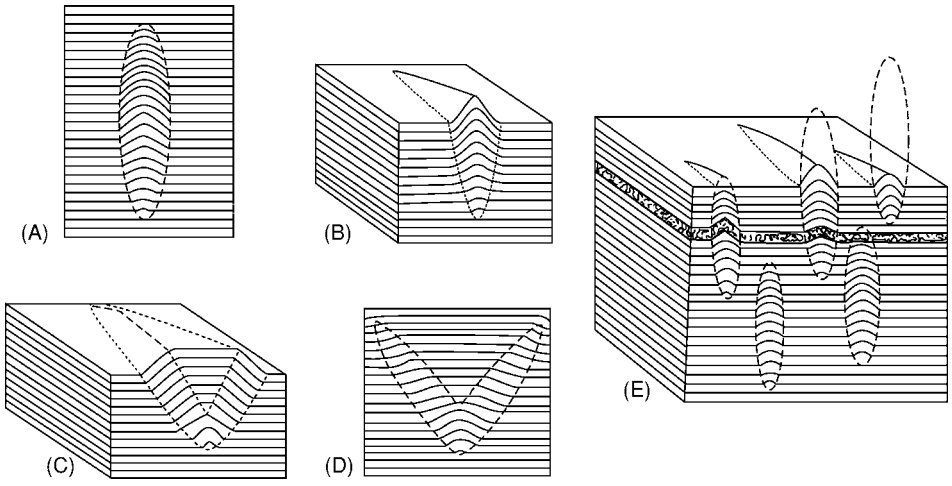


Figure 13 (A) Typical profile geometry of a fold in a multilayer. (B) Block diagram showing a fold dying out in both profile and plan view. (C) Block diagram and (D) profile geometry of a box fold. (E) The spatial organization of folds within a multilayer.

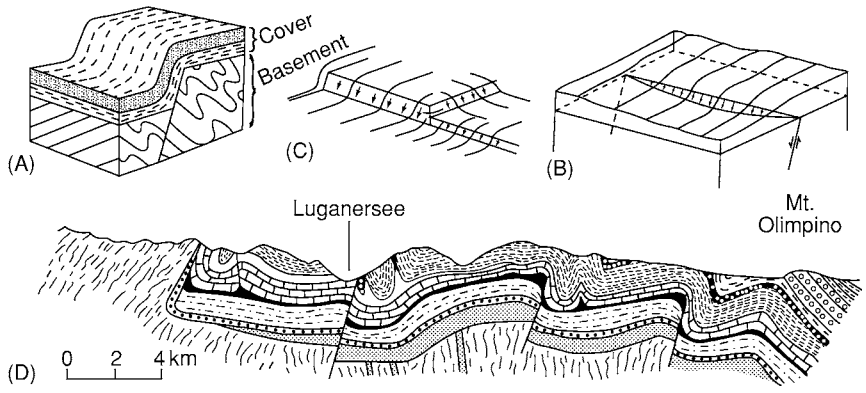


Figure 14 (A) and (B) block diagrams of drape or forced folds, the result of normal faulting in the basement. (C) shows the type of fold geometry which may be associated with block faulting in the basement. (D) Forced folds probably formed over reverse faults in the basement.

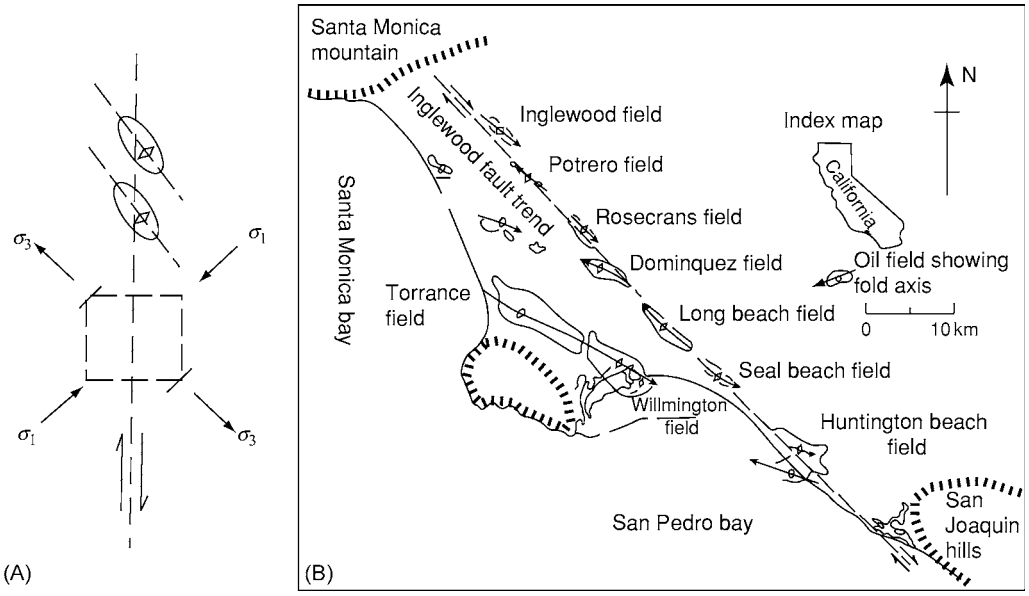


Figure 15 (A) Movement along a wrench fault in the basement produces horizontal compression in the cover rocks and the formation of a linear train of offset folds above the fault. (B) Such en echelon folds above the Inglewood fault, California.

folding in the overlying cover rocks. The stress generated in the cover rocks above a basement strikeslip fault is shown in Figure 15A. A local horizontal compression, inclined at 45° to the basement fault, is developed in the cover and this can give rise to a variety of structures, depending on the rheological condition of the cover rocks. If they behave in a ductile manner, then a series of folds may develop. They will form with their axial planes at right angles to the local maximum compression (σ_1) and will be arranged in an offset manner along the trace of the basement fault (Figure 15A). Natural examples of these linear arrays of buckle folds (see, for example, Figure 15B, which shows folding induced in the cover rocks above the San Andreas wrench fault) are excellent markers for locating major hidden faults.

Fault-Bend Folds

The forced folds shown in Figure 14 occur in cover rocks which respond in a ductile manner in response to movement on faults in a relatively rigid basement. Other well-documented relationships between faults and folds are known, one of the most familiar being fault-bend folding (Figure 16). In this type of forced folding, the folding is not the result of the movement of rigid fault blocks in the basement but rather the result of fault movement within the cover rocks. As can be seen from Figure 16, faults are not perfectly planar surfaces of slip. They generally have gentle undulations and may display substantial curvature or sharp bends. For example, the thrust fault shown in Figure 16 is made up of two horizontal portions (flats) linked by an inclined portion (ramp). As the two fault blocks slip past one another there must be deformation in at least one of them because rocks are not strong enough to support large voids. For this reason, many major folds within layered rock exist within the hanging-wall fault blocks, formed by bending the fault blocks as they slip over non-planar fault surfaces. This mechanism of folding is termed Fault-bend folding.

Flow Folding

In addition to the mechanisms of buckling and bending, folds can also be produced by flow. Impressive examples occur during the outpouring of lavas and during the slower flow of ice and salt. Salt is less dense than most rocks and therefore tends to rise diapirically though overlying strata. The resulting salt domes often emerge at the Earth's surface (Figure 17) where they can form salt glaciers (Namakiers). Figure 18 shows large flat lying flow folds formed as a result of flow within a salt glacier.

Implications of Folds Regarding the Properties of the Rock

The Earth's crust is characterized by an upper seismically active zone where brittle failure dominates, resulting in the formation of fractures and loss of continuity of the rocks, and a lower aseismic zone associated with ductile deformation, i.e., deformation that

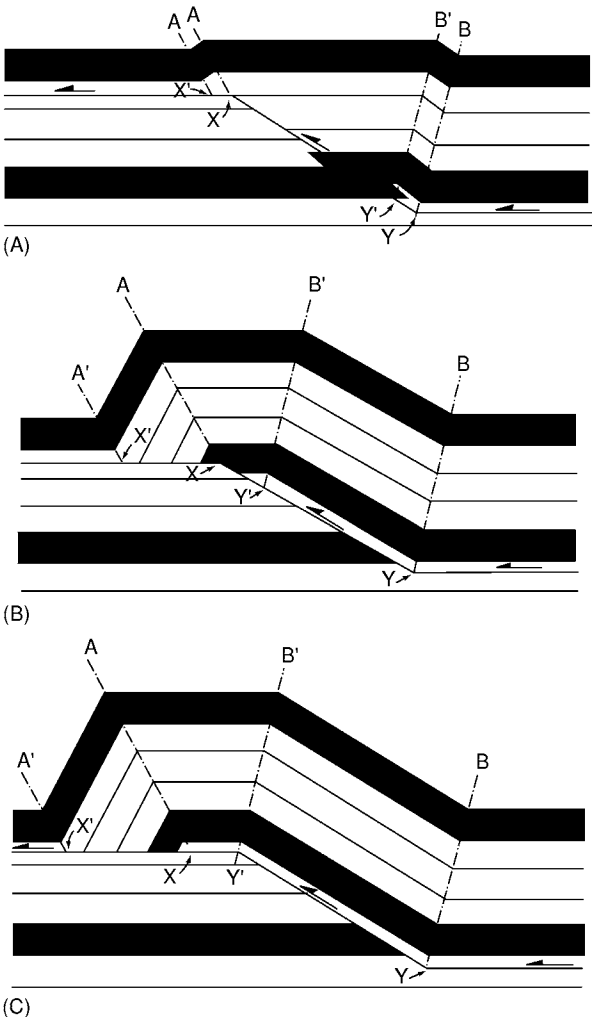


Figure 16 The development of a fault bend fold as a thrust sheet rides over a ramp in the detachment horizon. (After Suppe (1983).)

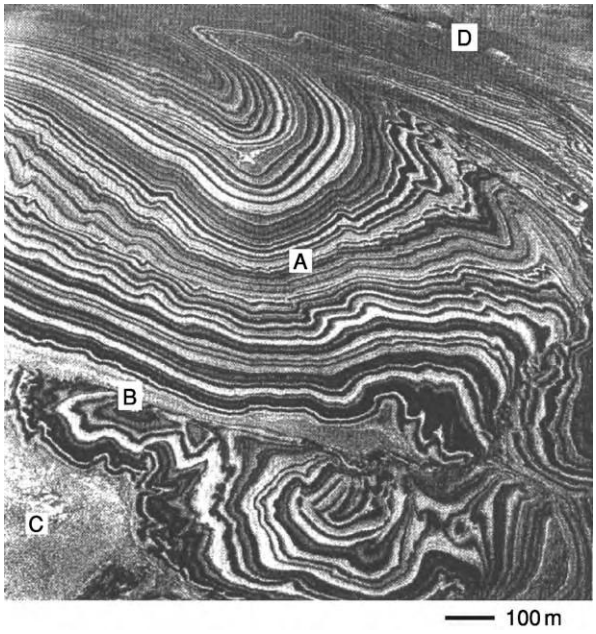


Figure 17 Oblique aerial view of a salt dome from the Great Kavir, Iran, showing the agate like regularity of bedding in the younger salt (A) and a fault contact (B) with the more massive older salt (C).

occurs without the loss of continuity of the material. Folding tends to occur within this ductile zone although, as noted earlier, folds can form at all depths in the crust. In addition, fractures can form in association with folding and the orientation of the fractures indicates clearly that they are formed by the same stress field that operated during folding (Figure 19).

Strain Within a Folded Layer and Associated Fracturing

The strain distribution (and therefore the fracture pattern) within a folded layer is dependant on the layer properties. In a homogenous isotropic layer, such as a uniform sandstone or limestone bed, the strain distribution is likely to be similar to that shown in Figure 20B, in which a layer parallel extensional field associated with the outer arc is separated from a layer parallel compression field associated with the inner arc by a neutral surface along which there is no strain. This model of strain distribution is termed Tangential Longitudinal strain folding. In contrast, a homogeneous anisotropic layer, such as a well-bedded shale, may fold by bedding parallel slip which results in the strain distribution shown in Figure 20C. This is known as flexural flow folding if the shear strain parallel to the layer boundary is uniformly distributed across the layer, and flexural slip folding if it is concentrated along distinct bedding planes. It is interesting to note that both models of folding (Figure 20B and C) produce parallel folds, i.e., folds with a constant orthogonal thickness. This illustrates the fact that the strain state within a fold cannot be deduced from the geometry of its profile section. However, the different strain patterns within the two models reflect the fact that they have very different stress fields within them which may lead to the formation of characteristic fracture patterns which enable the two fold types to be recognized in the field. For example, the extensional fractures in the outer arc of the pericline shown in Figure 20A and the shear fractures in the inner arc, indicate outer arc extension and inner arc contraction, respectively, i.e., a pattern compatible with the deformation associated with the Tangential Longitudinal Strain fold (Figure 20B).

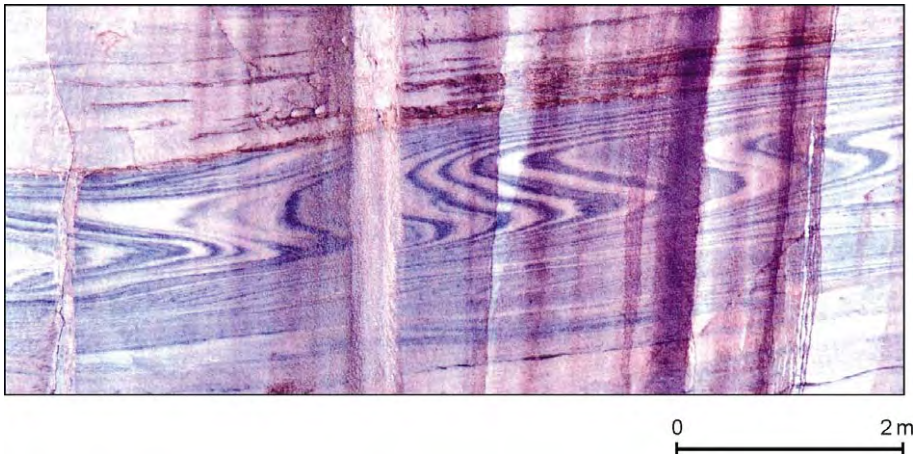


Figure 18 Large flat lying flow folds formed as a result of flow within a salt glacier, Iran.

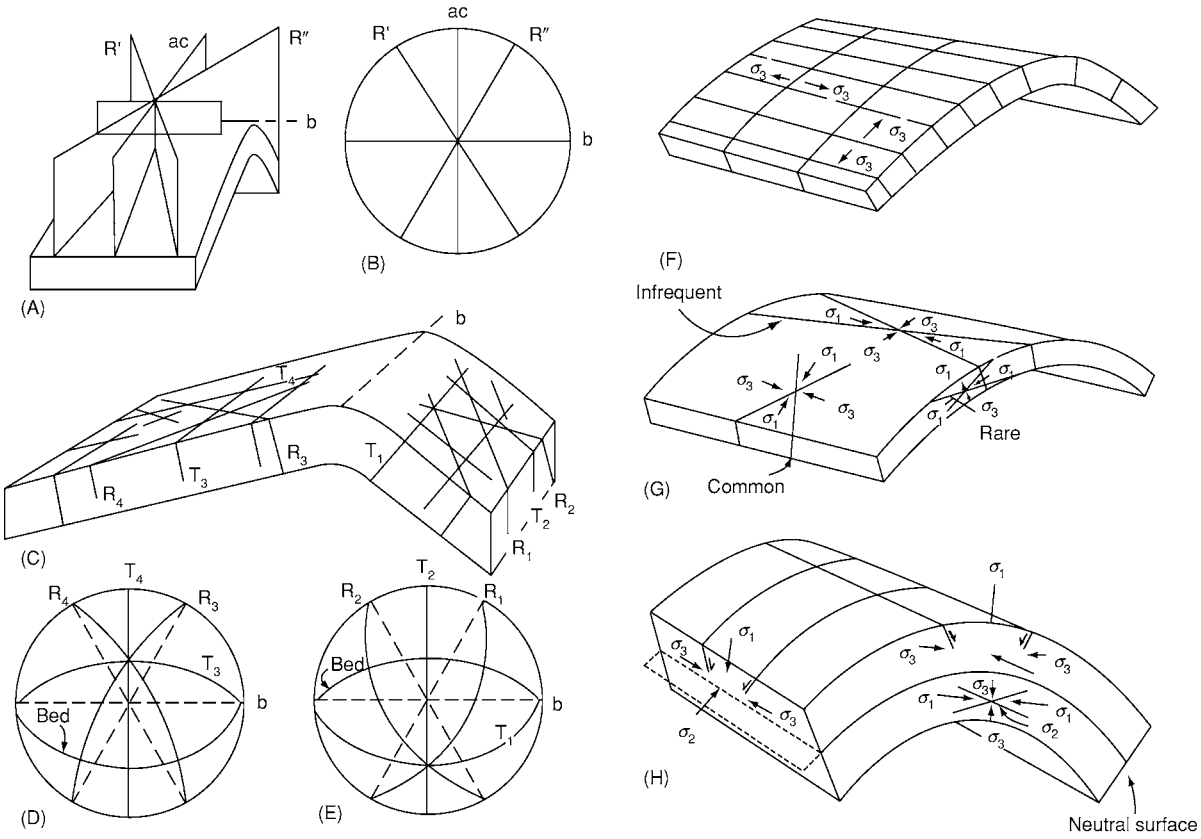


Figure 19 (A) Ideal relationship of master joints to a relatively small fold. (B) Stereographic plot of fractures shown in (A). (C) Trend of minor fractures in a folded competent unit. (D) and (E) Stereographic plots of fractures in the two limbs. R and T are shear and extension fractures, respectively. (F) Typical relationship of extensional fractures to a fold. The orientation of the least principal stress associated with each set (which are of different ages) is shown. (G) Typical orientation of shear fractures in a thin bedded layer, with associated stress systems. (H) Typical orientation of normal faults and thrusts which may develop in a thick flexed unit.

Thus, fractures formed in association with buckle folds may be the result of the local stresses generated as a result of buckling. Alternatively, they may be caused by the regional stress field. The existence of a buckle is likely to disturb the regional stress field. For example, folds are often accompanied by bedding plane slip, implying that the bedding planes cannot sustain a large shear stress. It follows, therefore, that

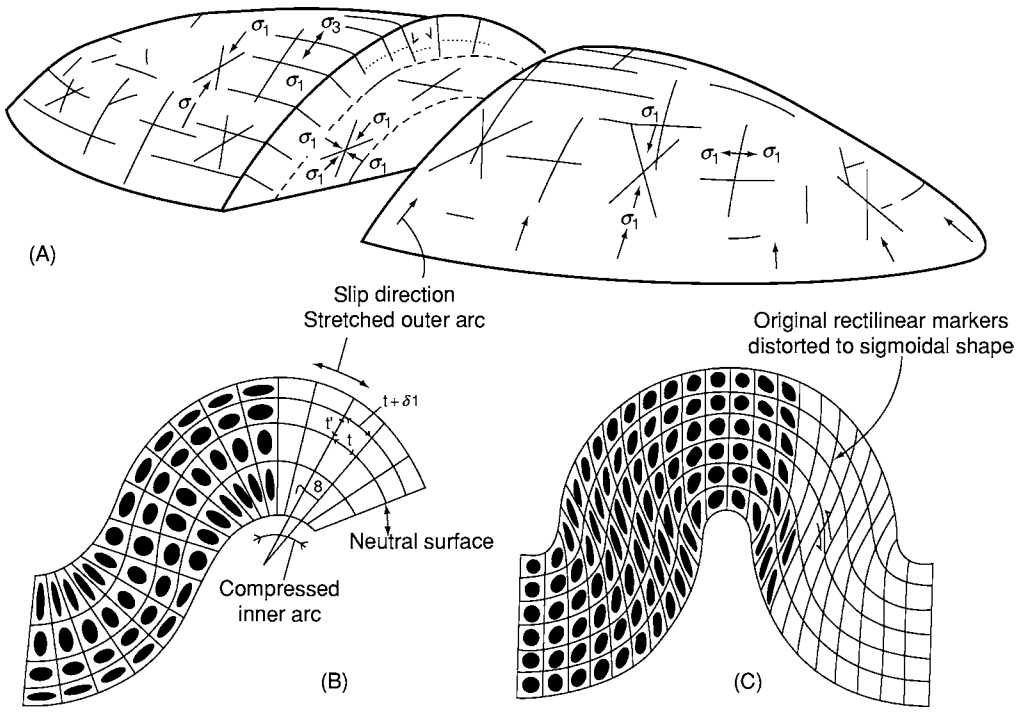


Figure 20 (A) Various fractures associated with a pericline. (After Stearns (1978).) (B) Strain distribution in a tangential longitudinal strain fold and (C) a flexural flow fold. (After Ramsay (1967).)

the principal stresses must be constrained to being either subparallel or subnormal to the folding layers. As a result of this stress deflection, the fractures also form normal to bedding. This is illustrated in Figure 19, which shows the predicted orientation of the shear and extensional fractures that would form in response to the regional compression generating the fold (Figure 19A), together with their projection on a stereographic plot (Figure 19B), and the frequently observed orientation of these fractures on the limbs of the fold which occurs as a result of the principal compressive stress following the layering (Figure 19C–E). The types and orientations of fracture found in association with buckle folds, which form as a result of both the regional and local stresses, are summarized in Figure 19F–H.

See Also

Tectonics: Fractures (Including Joints).

Further Reading

Cosgrove JW and Ameen MS (1999) A comparison of the geometry, spatial organization and fracture patterns associated with forced folds and buckle folds. In: Cosgrove JW and Ameen MS (eds.) *Forced Folds and Fractures*, pp. 7 21. Bath: Geological Society of London, Special Publication No. 169.

Jackson MPA, Cornelius RR, Graig CH, Stocklin J, and Talbot CJ (1990) Salt diapirs of the Great Kavir, Central Iran. *Geological Society of America Memoir 177*. pp. 139.

Price NJ and Cosgrove JW (1990) *Analysis of Geological Structures*. Cambridge: Cambridge University Press.

Ramberg H (1963) Fluid dynamics of viscous buck ling applicable to folding of layered rocks. *American Association of Petrology & Geology Bulletin* 47: 485 505.

Ramsay JG (1967) *Folding and Fracturing of Rocks*. New York: McGraw Hill.

Ramsay JG and Huber MI (1987) *The Techniques of Modern Structural Geology: Volume 2: Folds and frac ture*. London: Academic Press.

Stearns DW (1978) Faulting and Forced folding in the Rocky Mountain Foreland. *Geological Society of Amer ica*. Memoir 151: 1 38.

Suppe J (1983) Geometry and Kinematics of fault bend folding. *American Journal of Science* 283: 684 721.

Van Hise CR (1894) Principals of North American pre Cambrian geology. *US Geological Survey*, 16th annual report. pp. 571 843.

Weiss LE (1959) Structural analysis of the basement sys tem at Turoka, Kenya. London: *Overseas Geology and Mineral Resources* 7: 3 35, 123 153.

Fractures (Including Joints)

J W Cosgrove, Imperial College London, London, UK

Introduction

Fractures are the result of brittle failure which is the general term given to failure during which continuity of the material is lost. Deformation that does not involve loss of continuity is termed ductile. Two modes of brittle failure have been recognized, namely Shear failure and Tensile failure and these can be distinguished from each other on the basis of: (i) the orientation of the fractures with respect to the principal stresses that caused them; and (ii) the relative motion of the rock on each side of the fractures (Figure 1).

If the rock moves parallel to the fracture (Figure 1A) the fracture is a shear fracture and if it only moves normal to the fracture (Figure 1B) it is a tensile fracture.

Shear fractures in rocks are called faults and tensile fractures joints.

Mechanism of Formation of Fractures

The current understanding of brittle failure in rocks is the result of a combination of field observation, experimental work, and theoretical study. Field observations show the two types of fractures, and experimental work reveals that shear fracture occurs when the principal stresses are all compressive, and tensile failure when the least principal stress is tensile and equal to or greater than the tensile strength of the rock.

Shear Failure

Experiments have been performed in which cylindrical rock samples, surrounded by a hollow jacket into which a fluid can be injected to provide a confining stress (σ_3), are subjected to an axial loading (σ_1) until they fail. This results in a body of data recording the principal stress necessary for the rock to fail under a wide range of confining stress. These data can be represented graphically in two ways. The first is to plot the axial load against the confining stress (Figure 2A). For many rocks this plot produces a straight line whose intersection with the axial stress axis gives the uniaxial strength, i.e., the stress a rock can sustain with no confining stress. The graph

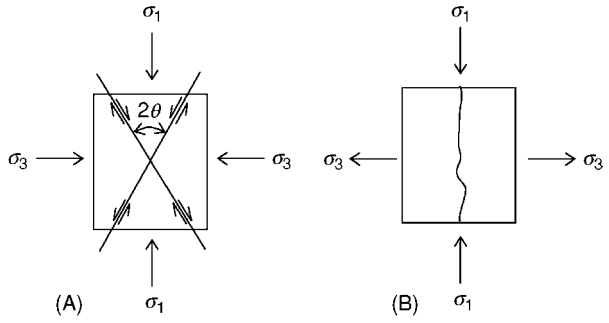


Figure 1 The two modes of brittle failure (A) Shear failure and (B) Tensile failure. They can be distinguished from each other on the basis of: (i) the orientation of the fractures with respect to the principal stresses that caused them; and (ii) the relative motion of the rock on each side of the fracture.

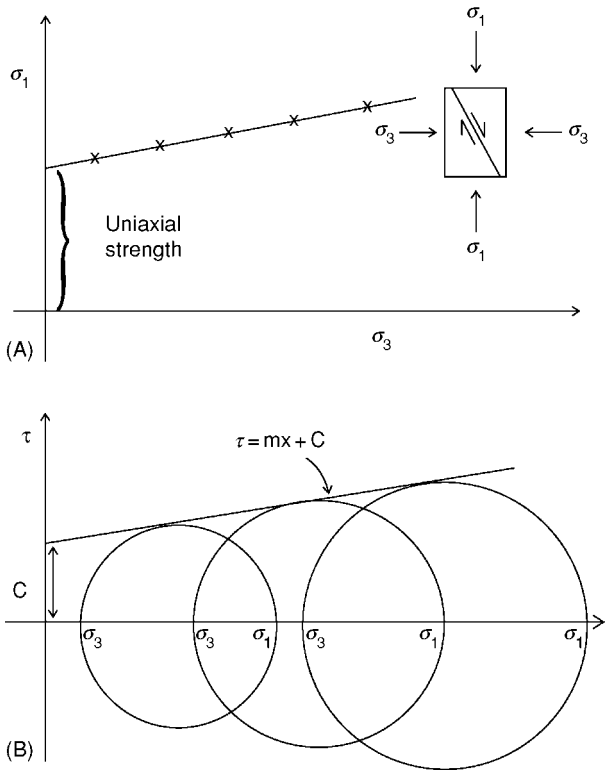


Figure 2 The graphical expression of the experimental data on shear failure. These data can be represented graphically in two ways. (A) is a plot the axial load against the confining stress at failure, (B) shows the same data plotted as a series of Mohr stress circles. The tangent to these circles represents the failure envelope for shear failure.

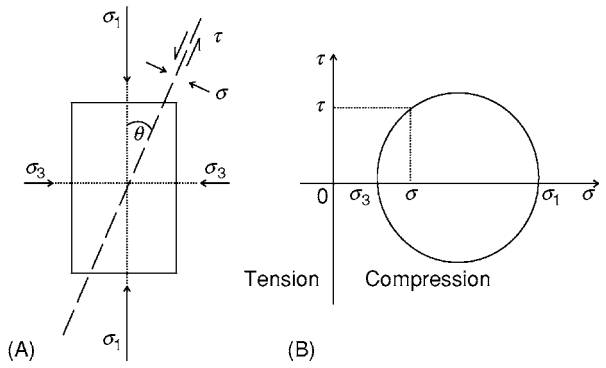


Figure 3 (A) The state of stress (σ and τ) on a plane inclined at θ° to the maximum principal compression, is given by the biaxial stress equations. (eqns [1] and [2]). The graph of σ against τ for values of θ between 0° and 180°; (B) defines a circle, the Mohr stress circle, which defines the state of stress.

indicates clearly that the strength of a rock (i.e., its ability to sustain a load without permanent deformation) is not a fixed value but depends (amongst other things) on the confining stress. The higher the confining stress, σ_3, the greater the axial load needed to cause failure. A rock's strength will, therefore, increase with increasing depth in the crust.

The second method of plotting the experimental data is to plot the stress state for each experiment (i.e., a stress state that caused the rock to fail) as a Mohr circle. The state of stress on any plane inclined at θ° to the maximum principal compression (Figure 3) is given by the biaxial stress equations:

$$\sigma = \sigma_1 \cos^2\theta + \sigma_3 \sin^2\theta \qquad [1]$$

$$\tau = (\sigma_1 - \sigma_3)\cos^2\theta \sin^2\theta \qquad [2]$$

These equations can be represented graphically by calculating σ and τ for values of θ between 0° and 180° and plotting the results on a graph of σ against τ. The resulting points define a circle, the Mohr stress circle (Figure 3B), which defines the state of stress. The diameter of the circle, which is a measure of the differential stress, $(\sigma_1 - \sigma_3)$, is determined by the values of the principal stresses, σ_1 and σ_3. As the experimental data represented in Figure 2A consists of values of the principal stresses that caused the rock to fail, these data can be plotted as a series of Mohr circles (Figure 2B). The tangent to these circles represents the failure criterion for shear failure.

For many rocks, this tangent is a straight line whose equation is:

$$\tau = m\sigma + C \qquad [3]$$

where m is the slope of the line and C its intersection with the shear stress axis.

Having established the shear failure criterion experimentally, it is important to compare it with the theoretically derived criterion. This was developed independently by Navier and Colomb, who argued that in order for a shear fracture to develop, the shear stress τ acting along the potential fracture plane (Figure 3A) must be sufficiently large to overcome the cohesion along that plane, C_0, plus the resistance to shear along the plane once it had formed. The resistance to slip is given by Amonton's law of frictional sliding which states:

$$\tau = \mu\sigma \qquad [4]$$

where τ and σ are the shear and normal stresses, respectively acting on the fracture plane and μ the coefficient of sliding friction. μ is defined as the tangent of the angle of sliding friction φ. Hence the complete criterion can be expressed in the form:

$$\tau = \sigma \tan\varphi + C_0 \qquad [5]$$

This is known as the Navier–Colomb criterion of shear failure and is identical to the criterion established experimentally (eqn [3]). The orientation of the planes where this condition is first met can be determined by substituting the biaxial stress equations (eqns [1] and [2]) into the shear failure criterion (eqn [5]) and solving for the minimum. The optimum orientations for the shear fractures are:

$$\theta = + \text{ or } - [45^\circ - \varphi/2] \qquad [6]$$

where θ is the angle between the maximum compressive stress (σ_1) and the shear fracture (Figure 1A). Note that two fracture orientations are predicted, inclined at $45^\circ - \varphi/2$ each side of σ_1. They are termed conjugate shear planes and although the magnitude of the shear stress along them is the same, the sense of shear is different. Faults are the geological expression of shear failure, and conjugate small-scale faults in a sequence of alternating sandstones and shales are shown in Figure 4.

Classification of Faults

As can be seen from Figure 1A, the orientation of a fault is controlled by the orientation of the principal stresses that generate them. Field observations reveal that faults fall into three classes, normal faults, wrench (or strike slip) faults, and thrust (or reverse) faults which, as can be seen from Figure 5, correspond to stress states where σ_1, σ_2, and σ_3 are vertical.

Figure 4 Conjugate normal faults in Carboniferous turbidites from Bude, Cornwall, England.

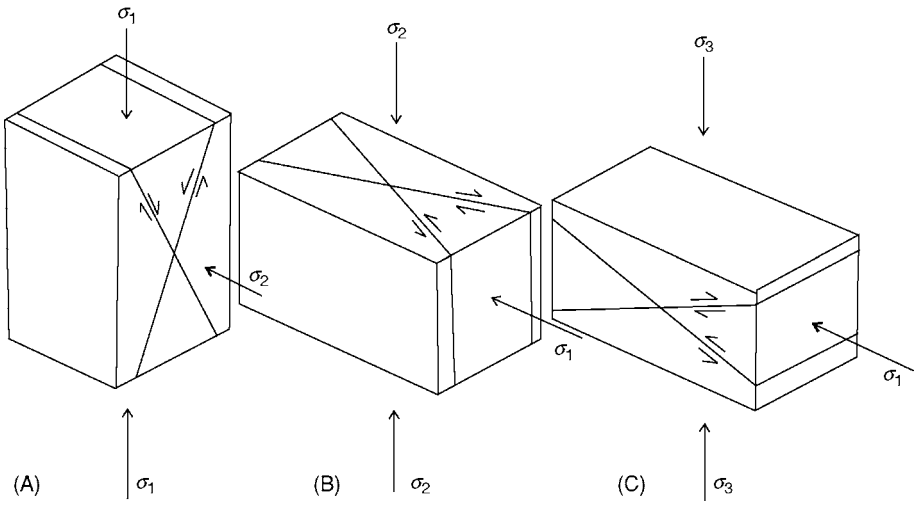


Figure 5 The orientation of shear fractures that form when (A) σ_1, (B) σ_2 and (C) σ_3 are vertical. The resulting faults are termed, normal faults, wrench (or strike slip) faults, and thrust (or reverse) faults, respectively.

This observation implies that the principal stresses tend to be oriented in one of these three orientations. In 1951, Anderson argued that this was because the Earth's surface is a free surface which cannot sustain a shear stress. Thus, in order not to generate a shear stress parallel to this surface, the principal stresses are constrained to remain either parallel or normal to it. As is discussed later, the three classes of faults characterize three different tectonic regimes.

Tensile Failure

Joints and veins are the most common geological expression of tensile failure. Experiments show that this type of failure generates fractures normal to σ_3 (Figure 1B). The theory of tensile failure was developed by Griffith (1925) who argued that in an ideal material the tensile strength of a material would be determined by the strength of the

inter-atomic bonds. However, experiments revealed that the measured tensile strength of a material is usually several orders of magnitude lower than that calculated on the basis of their inter-atomic bond strength. Griffith argued that this was because the materials contain small flaws or microfractures and that these resulted in a local stress magnification at the crack tips. This is illustrated diagrammatically in Figure 6, which shows the stress configuration of a plate subjected to a horizontal extension. In the first plate (Figure 6A), the stress is evenly distributed through the plate. In the second plate, which contains a flaw represented by an elliptical crack (Figure 6B), the tensile stress is locally magnified at the fracture tips. The amount of magnification depends primarily on the orientation and eccentricity of the crack. The greater the eccentricity, the greater the magnification. Griffith argued that by this means, a relatively low applied stress could be locally amplified at fracture tips within the material to the point where it reached the stress required to break the atomic bonds within the material causing the material to fail. He developed the following failure criterion (the Griffith criterion of tensile failure) based on this model:

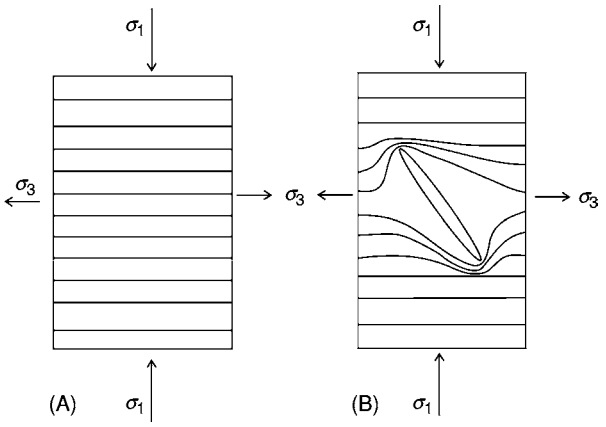


Figure 6 (A) A uniform stress field represented by uniformly spaced stress trajectories in a stretched layer. (B) The concentration of tensile stress at crack tips in a uniformly extended layer.

$$\tau^2 + 4T\sigma - 4T^2 = 0 \qquad [7]$$

where τ is the shear stress, σ the normal stress, and T the tensile strength of the material. The graphical expression of this failure criterion is shown in Figure 7A. It has the form of a parabola whose intersection with the normal stress axis gives the tensile strength and with the shear stress axis the cohesion.

The complete criteria for brittle failure, (the Griffith, Navier–Colomb criteria), is obtained by linking the two criterion (eqns [5] and [7]) at the point where their slopes are identical (Figure 7B). Any stress state can be represented on this graph as

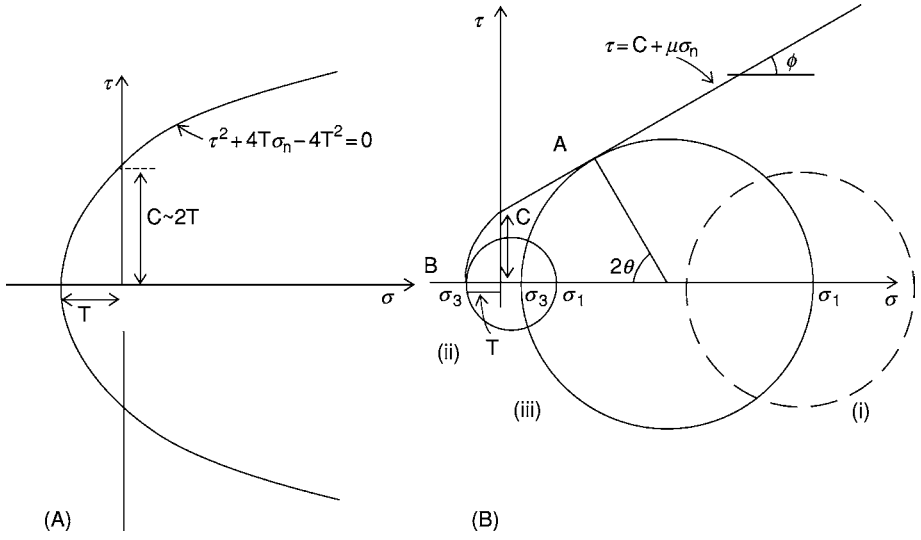


Figure 7 (A) The graphical expression of the Griffith criterion of tensile failure (eqn [7]). It has the form of a parabola whose intersection with the normal stress axis gives the tensile strength of the material and with the shear stress axis the cohesion. (B) The complete criteria for brittle failure, (the Griffith, Navier Colomb criteria), is obtained by linking the two criteria (eqns [5] and [7]) at the point where their slopes are identical.

a Mohr circle whose position is determined by the values of the principal stresses (σ_1 and σ_3). If a stress state, when plotted on the graph, does not touch or intersect the failure envelope, the stress state is stable, i.e., will not cause the rock to fail, e.g., stress field (i) Figure 7B. If, however, it does touch the envelope, failure will occur, either by tensile fracturing if the contact is with the tensile part of the envelope (Figure 7B (ii)), or shear fracturing if it is with the shear part (Figure 7B (iii)).

What Determines Whether Tensile or Shear Fractures Form?

It can be seen from Figure 7B that shear failure is associated with a large differential stress ($\sigma_1 - \sigma_3$) i.e., the Mohr's stress circle must be large in order to intersect the shear failure envelope, and that tensile failure is associated with a low differential stress, i.e., the Mohr's stress circle must be small in order to intersect the tensile failure envelope. The precise conditions necessary for the formation of the two types of fractures are:

$$\text{For tensile failure to occur} \quad (\sigma_1 - \sigma_3) < 4T \qquad [8a]$$

$$\text{For shear failure to occur} \quad (\sigma_1 - \sigma_3) > 4T \qquad [8b]$$

where T is the tensile strength of the material.

The geometrical relationships between the principal stresses and the fractures they produce (i.e., a conjugate set of shear fractures symmetrically about σ_1 and a single set of tensile fractures at right angles to σ_3) is shown in Figure 1 and, as noted below, the understanding of these relationships provides a powerful tool in fracture analysis.

It follows therefore that the orientation of the fractures that form in response to a stress field is determined by the orientation of the principal stresses (Figure 1), and the type of fracture (shear or tensile) by the magnitude of the differential stress.

The Effect of a Fluid Pressure on Fracturing

Fluid-Induced Failure

The state of stress in the crust is dominantly compressional. For example, in a nontectonic environment, the stress at any depth is generated by the overburden which produces a compressive vertical stress which induces a compressive horizontal stress. Thus, at any depth the Mohr stress circle will plot in the compressive regime in Figure 7B and there will be no possibility of tensile failure. Geologists were, therefore, perplexed to find that large numbers of tensional fractures occur in the crust. This paradox was resolved when the importance of fluid pressures within a rock was understood. Pore fluid pressure within a rock increases as the rock is buried. (*see* **Tectonics:** Hydrothermal Activity). The stress state within the pores is hydrostatic and the pressure acts so as to appose the lithostatic stress caused by the overburden. This effect can be shown diagrammatically by representing the lithostatic stress as an ellipse with the stress acting compressively and the fluid pressure as a circle with the pressure acting outwards (Figure 8A). The fluid pressure reduces all the lithostatic stresses by an amount P_{fluid} to give an effective stress. Thus, the principal stresses σ_1 and σ_3 become $(\sigma_1 - P_{fluid})$ and $(\sigma_3 - P_{fluid})$. This new stress field can be plotted as a Mohr stress circle (Figure 8B). It can be seen that the original lithostatic stress circle is moved towards the tensile regime but that the diameter of the circle, i.e., the differential stress, remains unchanged.

The amount of migration of the stress circle is determined by the magnitude of the fluid pressure. Thus, as the fluid pressure gradually increases during burial, the stress circle is pushed inexorably towards the failure envelope. When it hits the envelope, failure occurs. Such failure is termed fluid induced or hydraulic fracturing. In this way an originally compressional stress regime can be changed so that one or more of the principal stresses becomes effectively tensile and the conditions for tensile failure can be satisfied.

The Expression of Fluid-Induced Failure

In the example shown in Figure 8B the lithostatic stress had a small differential stress (i.e., less that 4T (see eqn [8]) and as a result the induced hydraulic fractures were tensile fractures. If it had been greater than 4T, shear fractures would have formed.

The Organization of Tensile Fractures

The Mohr circles shown on Figure 9 all intersect the failure envelope in the tensile regime, i.e., the differential stresses are all less than 4T and will all therefore result in tensile failure. Their differential stresses vary from just less than 4T (circle (i) Figure 9), to zero (circle (iv), Figure 9). Note that when the stress state is hydrostatic, the Mohr circle is reduced to a point.

As noted above, tensile fractures form normal to the minimum principal compressive stress σ_3 (Figure 1B), i.e., they open against the minimum compressive stress. The stress state represented by Mohr circle (i) in Figure 9, has a relatively large differential stress and there is, therefore, a definite direction of easy opening for the fractures. The fractures would

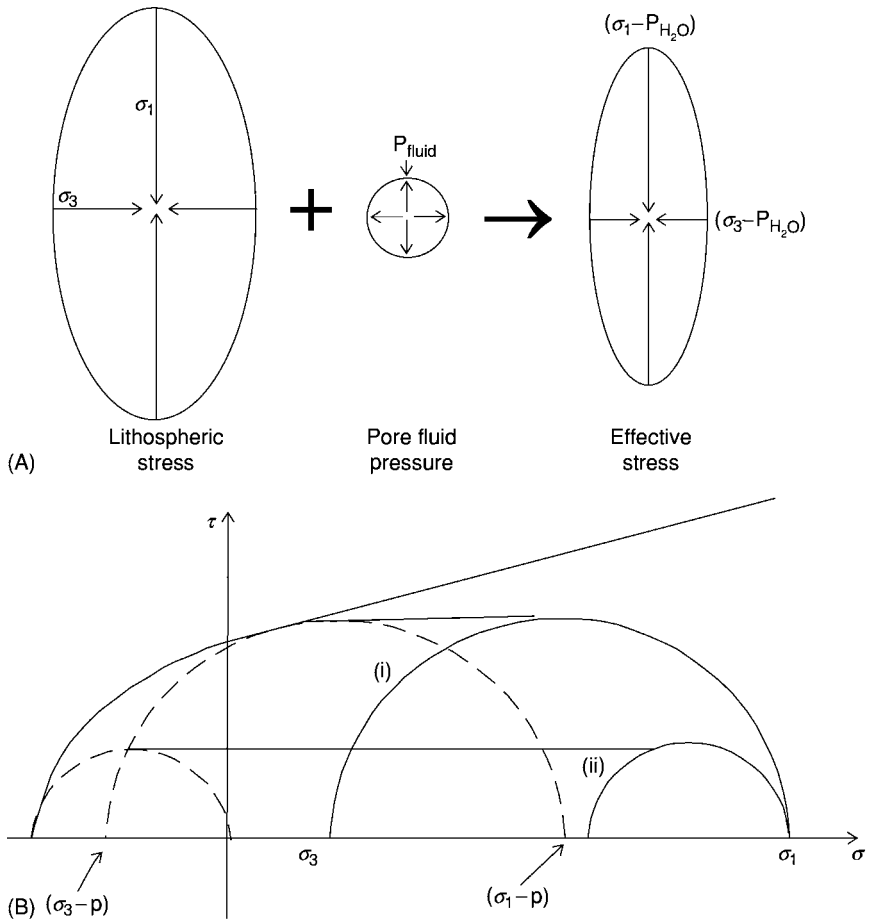


Figure 8 (A) Diagramatic representation of the effect of a fluid pressure (the circle with the outwardly acting stress) on the stress state in a rock (the ellipse with the inwardly acting stresses). All normal stresses are reduced to an effective stress (σ p_{fluid}) but the differential stress (σ_1 σ_3) remains unchanged. The effect is to cause the Mohr stress circle to move to the left by an amount equal to the fluid pressure; (B). Thus depending on the magnitude of the differential stress the induced fractures will be either shear (stress state (i)) or tensile (stress state (B)) (*see* **Tectonics:** Folding).

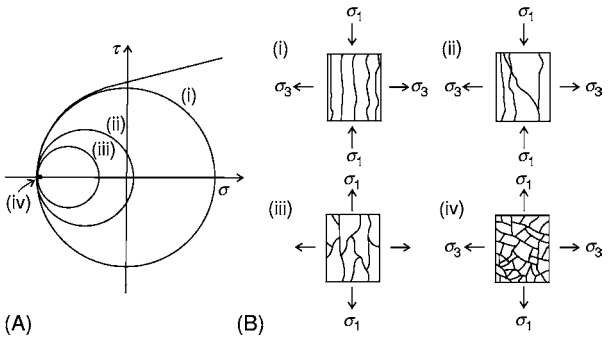


Figure 9 (A) Mohr stress circles (i) (iv) representing a range of stress states, all of which will lead to tensile failure. NB the Mohr circle (iv), that represents hydrostatic stress is a point. (B) Pat terns of tensile failure generated by the corresponding stress states shown in (A).

therefore exhibit a marked alignment normal to this direction (Figure 9B (i)). However, for the stress states represented by the Mohr circles (ii–iv), the differential stress becomes progressively smaller until, for the hydrostatic stress represented by circle (iv), the differential stress is zero. In a hydrostatic stress field the normal stress across all planes is the same and there is, therefore, no direction of relatively easy opening for the fractures. Thus, they will show no preferred orientation and, if they are sufficiently closely spaced and well developed, will produce a brecciation of the rock (Figure 9B (iv)). It can be seen that as the differential stress becomes progressively lower so the tendency for the resulting tensile fractures to form a regular array normal to σ_3 decreases. Tensile fracture systems ranging from well-aligned fractures to randomly

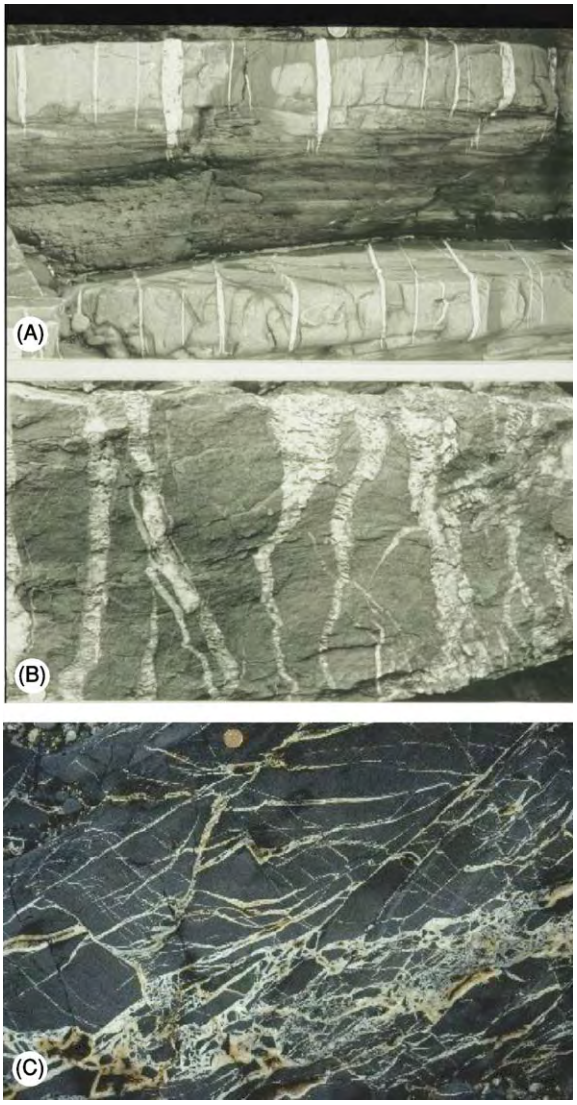


Figure 10 (A) A regular array of tensile fractures exposed on a bedding plane in Carboniferous sandstone, Millook, Cornwall, England. (B) Less well organized tensile fractures cutting Devonian sandstones, St. Anne's Head, South Wales. (C) Carboniferous sandstone cut by randomly oriented tension fractures.

Figure 11 Polygonal arrays of tensile fractures cause by the desiccation of a silt layer.

Figure 12 Polygonal arrays of tensile fractures cause by the cooling of a lava flow. (Giant's Causeway Northern Ireland).

oriented fractures are to be expected in rocks, and field observations support this idea, Figure 10.

As noted above, the problem of forming tensile fractures in the compressive stress field that generally characterizes the Earth's crust can be solved by appealing to high fluid pressures. However, tensile failure can occur in rocks without the aid of a high internal fluid pressure, for example, during the contraction of a layer as a result of desiccation of a sediment (Figure 11) or the cooling of an igneous body (Figure 12).

In both these examples, the fractures are organized into polygonal arrays showing that the tensile stresses generated were the same in all directions.

Fracture Sets

Generally, the state of stress in the Earth's crust is not hydrostatic. Consequently a single episode of deformation is likely to generate a set of fractures with the same orientation. However, most rocks experience several different stress regimes during their history with the result that several fracture sets are frequently found superimposed on each other to produce a fracture network (Figure 13). The interaction of late fractures with early fractures is illustrated in Figure 14.

The effect of early fractures on later ones is to arrest their propagation and to modify their orientation. It can be seen from Figures 14A and B that one

fracture often ends abruptly against another. This abutting relationship gives the relative age of the fractures, i.e., the later fracture abuts against the earlier fracture. If an early fracture is an open fracture then it will represent a free surface within the rock and, as discussed in the above section on classification of faults, will be unable to support a shear stress. Consequently, the principal stresses will reorient as they approach it into a position either normal or parallel to the fracture. This effect can be clearly seen in Figure 14, where later fractures curve into an orientation at right angles to the earlier fracture as they approach it.

Figure 13 A fracture network in a Liassic limestone bed from Lilstock, North Somerset, England. It was produced by the super position of individual fracture sets.

Fracture Networks

Structural geologists study the cross-cutting relationships of different fracture sets in order to determine their relative age. A variety of rules have been established to help in this task. It is found that early fractures tend to be long and relatively continuous and, as noted above, later fractures abut against these and are consequently shorter. Some of these features can be seen in Figure 15, which shows a fractured limestone pavement containing several fracture sets. The

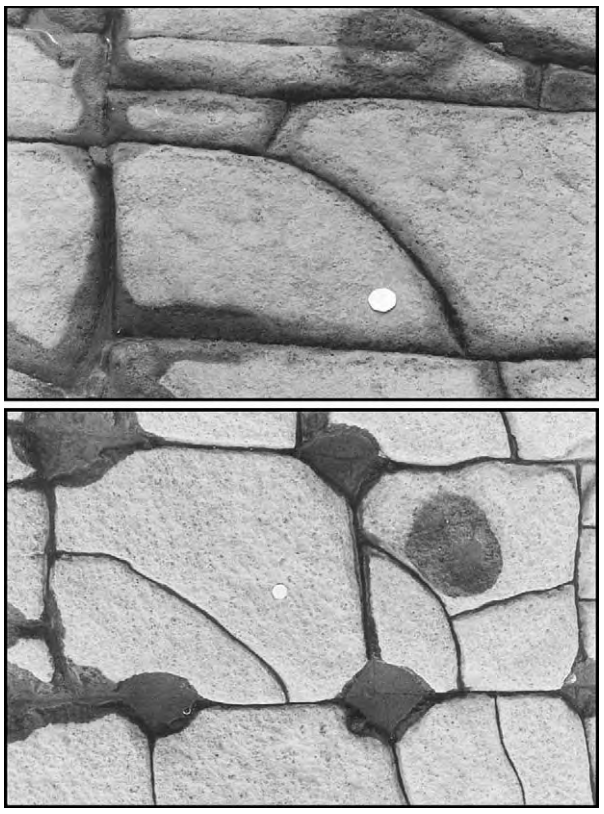


Figure 14 Details of the limestone pavement shown in Figure 13 illustrating the interaction of late fractures with early fractures. The effect of early fractures on later ones is to arrest their propagation and to modify their orientation. It can be seen that the later fractures are deflected by and abut against the earlier fractures.

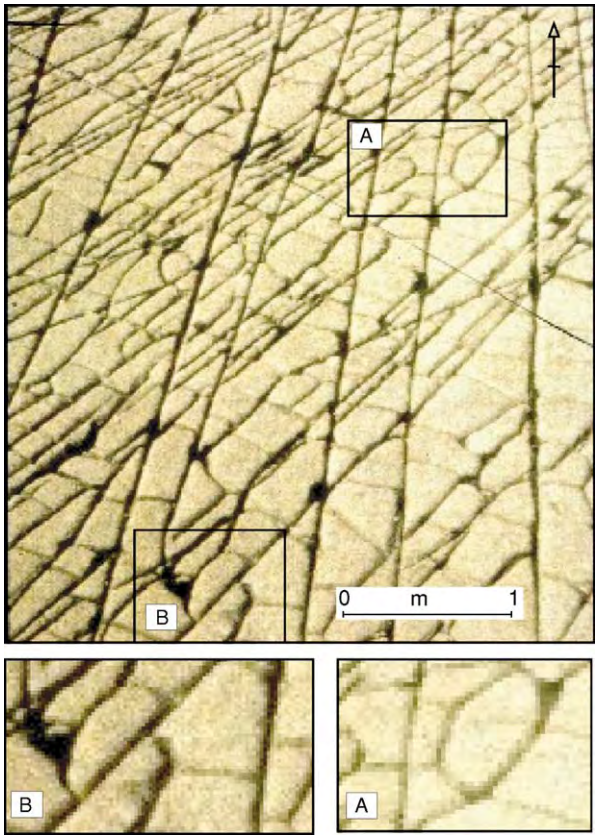


Figure 15 Fracture patterns in a limestone pavement at Lilstock, North Somerset, SW England. The older fracture sets are the most continuous and, as the sets become progressively younger, they become less continuous and less well oriented.

longest and most continuous runs approximately N–S. These are the oldest fractures and are crosscut by several younger sets which become progressively less continuous and less aligned as the regional stress fields responsible for their formation becomes progressively modified by the pre-existing fractures. The fracture set trending approximately NW–SE, the second set to form, shows a remarkable degree of continuity, being only affected by the N–S fractures; its orientation is related directly to the regional stress field.

However, as more fracture sets develop in the rock mass, modification of the stress orientation by the pre-existing fractures may result in there being a poor correlation between the fracture orientation and the regional stress field responsible for its formation. This is well illustrated in subarea A in Figure 15 which has been enlarged in the bottom left-hand corner of the figure. The influence of the pre-existing fractures on the orientation of the late fracturing is so marked that the later fractures display a polygonal organization and cannot be linked directly to the regional stress field responsible for their formation.

Fracture Analysis

A fracture analysis is the study of a fractured rock mass in order to: (i) establish the detailed geometry of the fracture network; (ii) determine the sequence of superposition of the different fracture sets that make up the fracture network; and (iii) deduce the stress regime associated with the formation of each fracture set. The reason why a detailed knowledge of the geometry of the fracture network is so important is that the bulk properties (e.g., strength, permeability) of a fractured rock mass (and most natural rocks are fractured) are generally determined by the fractures they contain rather than by the intrinsic rock properties.

Stages (ii) and (iii) of a fracture analysis are carried out using the principals outlined above relating to the interaction of fractures and the relationship between the stress field and fracture orientation (Figure 1).

Types of Faults a Plate Margins

The 'type' of plate margin is controlled by the relative motion of the two adjacent plates. They can be subdivided into three classes, convergent, divergent, and strike-slip. Convergent margins lead to compressional regimes at the plate margins which results in the formation of mountain belts. The stress regime is that appropriate for thrusts to form, namely a horizontal maximum principal compressive stress and a vertical minimum stress (Figure 5C). Divergent plate margins result in the formation of oceans and the separation of plates. The initial stage of this process is the fracturing of the lithosphere and the formation in the upper crust of major rift systems such as the East African Rift (*see* **Tectonics:** Rift Valleys). The stress regime of a horizontal minimum principal compressive stress and a vertical maximum stress is appropriate for the formation of normal faults (Figure 5A). When plates move parallel to each other at different velocities, conditions are appropriate for the formation of major wrench (strike-slip) faults (Figure 5B) such as the San Andreas Fault zone of California which separates the Pacific and North American plates.

Thus it can be seen that each of the three types of plate margins is characterized by a different types of fault.

Scale of Fracturing

Fractures occur on all scales within the Earth's crust, ranging from major faults that define plate margins, through faults that can be seen on seismic sections (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges), down to faults that can be observed directly in the field, e.g, Figure 4, to microscopic fractures only visible under the microscope. Detailed studies of the microfractures in rocks at different stages of the evolution of tensile fractures show, as predicted by Griffith's theory of stress magnification (1925) outlined above, that the microfractures grow by tensile failure at the crack tips and that suitably located microfractures link to form larger fractures oriented normal to σ_3, the minimum compressive stress (*see* **Tectonics:** Faults).

More remarkably, when the growth of shear fractures are studied in the same way, it is found that

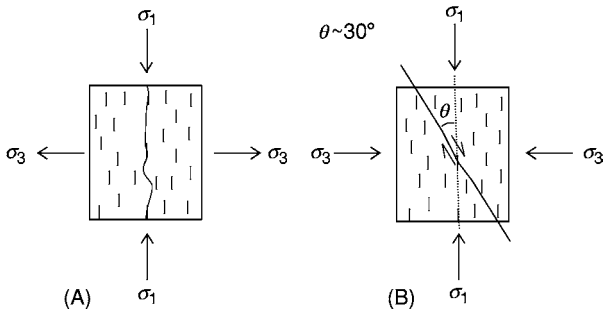


Figure 16 Randomly oriented micro fractures within a material and their growth by tensile failure and subsequent linkage to form (A) macroscopic tensile fractures and (B) macroscopic shear fractures.

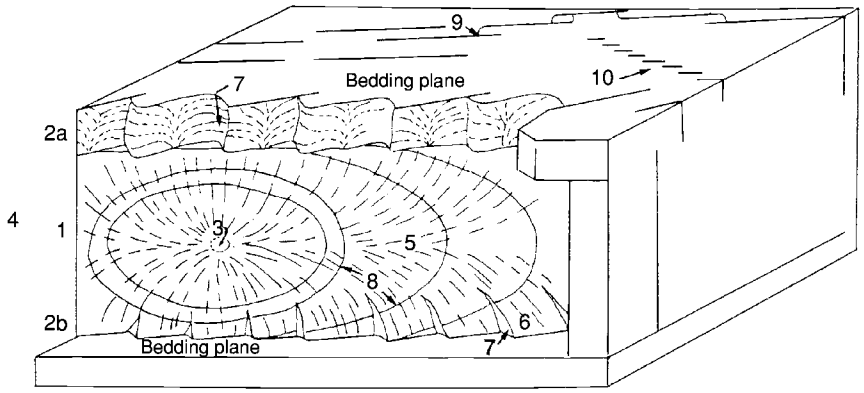


Figure 17 A block diagram illustrating the different types of surface structures (patterns) and fracture trace architecture. (Based on Kullander *et al.*, 1990.) (1) Main joint face, (2a) abrupt twist hackle fringe, (2b) Gradual twist hackle fringe, (3) Origin of fracture, (4) Hackle plume, (5) Plume axis, (6) Twist hackle face, (7) Twist hackle step, (8) Rig marks (front line of the fracture), (9) hooking, (10) En echelon fractures.

initially fracturing occurs by the growth of microfractures at their tips and in an orientation normal to σ_3. However, the macroscopic shear fractures are formed by the linking of offset microfractures, as shown in Figure 16. Thus, it can be seen that despite the two types of fractures having independent failure criteria and different orientations with respect to the principal stresses, they are nevertheless fundamentally linked on a microscopic scale. They are both the result of the growth of microfractures by tensile failure and differ only in the way in which these fractures are linked to a macroscopic fracture.

Surface Features of Fractures

Fractures display a variety of surface features and Figure 17 is a summary diagram showing some of these. Fractography is the science which deals with the description, analysis, and interpretation of fracture surface morphologies and links them to the causative stresses, mechanisms, and subsequent evolution of the fractures. It has been demonstrated that the diverging rays of the plumose structures (Figures 17 and 5) always remain parallel to the direction of propagation of the fracture. Thus, by constructing lines at right angles to these rays, the position and shape of the fracture front at different times of its evolution can be determined (Figures 17 and 8). When the exposures are sufficiently good, it is found that the fracture fronts form a series of concentric 'ellipses', the centre of which marks the site of fracture initiation.

See Also

Tectonics: Earthquakes; Faults; Folding; Hydrothermal Activity; Seismic Structure At Mid-Ocean Ridges; Rift Valleys.

Further Reading

Ameen MS (1995) *Fractography: fracture topography as a tool in fracture mechanics and stress analysis.* Special Publication Geological Soc. Of London, No. 92; p 240.

Anderson EM (1951) *The dynamics of faulting and dyke formation with application to Britain,* 2nd edn. Oliver & Boyd.

Griffith AA (1925) *The theory of rupture.* 1st International Conference of Applied Mechanics Proceeding Delft. P55.

Kullander BR, Dean SL, and Ward BJ (1990) *Fracture Core Analysis: Interpretation, Logging, and Use of Natural and Inducted Fractures in Core: Methods in Exploration Series, No. 8.* Tulsa, Oklahoma, USA: American Association of Petroleum Geologists.

Mandl G (1999) *Faulting in Brittle Rocks.* Springer.

Price NJ (1966) *Fault and joint development in brittle and semi brittle rock.* Pergamon press.

Hydrothermal Activity

R P Lowell, Georgia Institute of Technology, Atlanta, GA, USA
P A Rona, Rutgers University, New Brunswick, NJ, USA

Introduction

Hydrothermal activity results from the complex interplay of heat transfer, fluid–rock chemical reactions, and fluid circulation within Earth's continental and oceanic crust. Hydrothermal circulation redistributes heat energy in the crust, often giving rise to regions of concentrated thermal output that lead to the emplacement of economically important mineral deposits and that serve as geothermal energy resources. Hydrothermal activity is thus an important component of Earth's global heat engine whereby heat transferred to the lithosphere by mantle convection is transferred to Earth's surface by thermal conduction, volcanic extrusion, and hydrothermal venting. Lithospheric plate motions, volcanic and tectonic activity, and earthquakes are manifestations of Earth's global heat engine. There is a close connection between tectonic plate boundaries and sites of hydrothermal activity (Figure 1). In the following sections of this article we compare some aspects of terrestrial and submarine hydrothermal activity, describe the basic physics and chemistry of hydrothermal circulation, briefly discuss the importance of two-phase flow, and suggest some directions for future study.

Comparison between Terrestrial and Submarine Activity

Although hydrothermal activity in terrestrial and submarine settings has many similarities there are significant distinctions. Part of the reason for this distinction stems from the manner in which heat is transported in continental and oceanic lithosphere, respectively (Table 1). In terrestrial settings, nearly 40% of the heat flux stems from radiogenic heat production in the crust, and 60% is conducted from the underlying mantle. Terrestrial hydrothermal activity accounts for less than 1% of Earth's thermal budget. On the other hand, the process of plate creation and seafloor spreading along the roughly 60 000-km ocean ridge system dominates the thermal regime of the oceanic lithosphere. Conductive heat flux from the spreading lithosphere decreases as $\tau^{-1/2}$, where τ is the lithospheric age. Hydrothermal circulation transports a significant fraction of the lithospheric heat advectively, leading to lower than expected conductive heat flow in young lithosphere (Figure 2). Nearly 25% of Earth's global heat loss and 33% of the heat loss from oceanic lithosphere result from hydrothermal activity. Most seafloor hydrothermal heat loss occurs at low temperature. High-temperature hydrothermal activity, which accounts for less than 10% of the total seafloor hydrothermal heat loss (Table 1), appears to occur only in lithosphere less than 1 My old.

Another interesting distinction between terrestrial and submarine hydrothermal activity has been their role in human endeavours. Warm and hot springs on the continents have been used for bathing and medicinal purposes since antiquity. Thermal springs were utilized throughout the Roman empire, and early descriptions of springs in Europe appear in seventeenth-century writings. Terrestrial hydrothermal systems have also long been used as an energy resource. Geothermal waters in Iceland have been used for heating for centuries, and by the 1930s a centralized heating system was established for Reykjavik. Geothermal steam has been produced at Lardarello, Italy, since the latter half of the nineteenth century, and the Geysers geothermal field in California was first exploited in the 1920s. It is remarkable that commercial development of geothermal resources occurred long before measurements of geothermal heat flux and without detailed geophysical exploration.

On the other hand, direct detection of submarine hydrothermal activity did not occur until the 1960s (Red Sea) and the 1970s at mid-ocean ridges of the Atlantic and Pacific. As a result of these discoveries the understanding of biogeochemical processes on Earth was revolutionized. It became clear that submarine hydrothermal circulation significantly impacts global geochemical cycles of both the lithosphere and the ocean. Chemical transport of certain major and trace elements to the ocean by hydrothermal discharge equals or exceeds river inputs. Moreover, hydrothermal fluids also serve as an energy resource for complex chemosynthetic biological ecosystems. The discovery of chemosynthetic ecosystems at seafloor hydrothermal vents has led to a new awareness of life in extreme environments and has stimulated the discussion of the origin of life on Earth and other planetary bodies in the solar system.

Physics and Chemistry

The fundamental components of hydrothermal activity are a heat source and a fluid circulation system

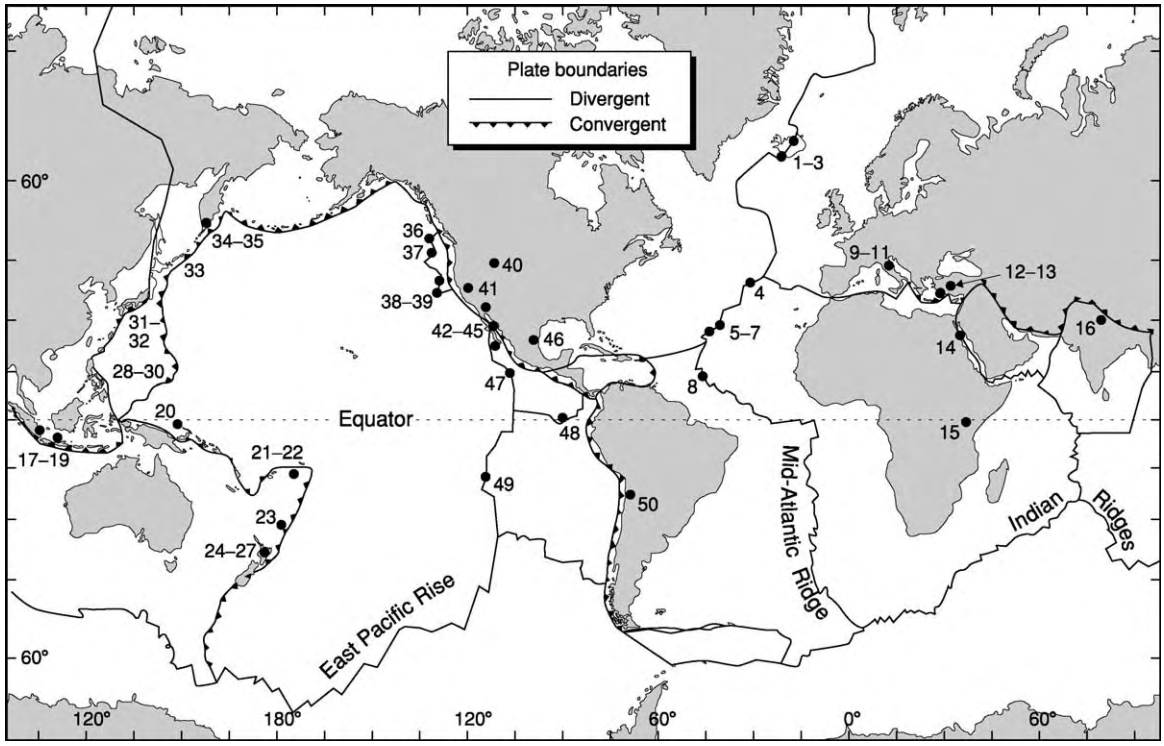


Figure 1 Plate tectonic map of the world showing locations (1 50) of selected submarine and terrestrial high temperature hydro thermal sites as follows: (1) Krafla; (2) Namafjall; (3) Svartsengi; (4) Rainbow; (5) Lost City; (6) TAG; (7) Snake Pit; (8) Logatchev; (9) Larderello; (10) Mt. Amiata; (11) Travale; (12) Kizildere; (13) Afyon; (14) Atlantis II Deep; (15) Olkaria; (16) Puga; (17) Kawah; (18) Kamodjang; (19) Dieng; (20) PACMANUS; (21) North Fiji Basin; (22) Lau Basin; (23) Brothers; (24) Kawerau; (25) Rotorua; (26) Broadlands; (27) Wairakei; (28) Tiwi; (29) Mariana Trough; (30) Okinawa Trough; (31) Otake; (32) Sunrise; (33) Matsukawa; (34) Paratunka; (35) Pauznetska; (36) Magic Mountain; (37) Main Endeavour; (38) Sea Cliff; (39) Escanaba; (40) Yellowstone; (41) Geysers; (42) Imperial Valley; (43) Cerro Prieto; (44) Guaymas Basin; (45) East Pacific Rise 21° N;(46) Pathe; (47) East Pacific Rise 9° N; (48) Galapagos; (49) Rapa Nui; (50) El Tatio.

Table 1 Heat flux from the Earth ($\times 10^{12}$ W)[a]

Type	*Value*
Continental Crust	
(a) Crustal radiogenic heat production	4.6
(b) Conductive heat flux from the mantle	6.8
(c) Extrusion of lavas	0.03
(d) Hydrothermal flux	0.1
Total	11.5
Oceanic Crust	
(a) Conduction	20.3
(b) Extrusion of lavas	0.3
(c) Axial high temperature hydrothermal flux	0.3
(d) Axial low temperature hydrothermal flux	2.7
(e) Off axis low temperature hydrothermal flux	7
Total hydrothermal flux	10
Total	30.6
Global Heat flux	42.2

[a]Compiled from Sclater JG, Jaupart C, and Galson D (1980) and Elderfield and Schultz (1996); modified after Lowell (1991).

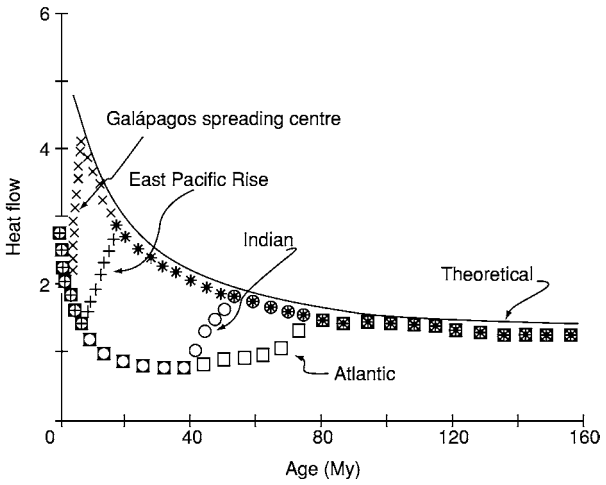


Figure 2 Observed mean heat flow for oceanic spreading centres compared with theoretical curve for conductive cooling of lithosphere. Reproduced from Anderson RN, Langseth MG Jr, and Slater JG (1977) The mechanisms of heat transfer through the floor of the Indian Ocean. *Journal of Geophysical Research* 82: 3391 3409.

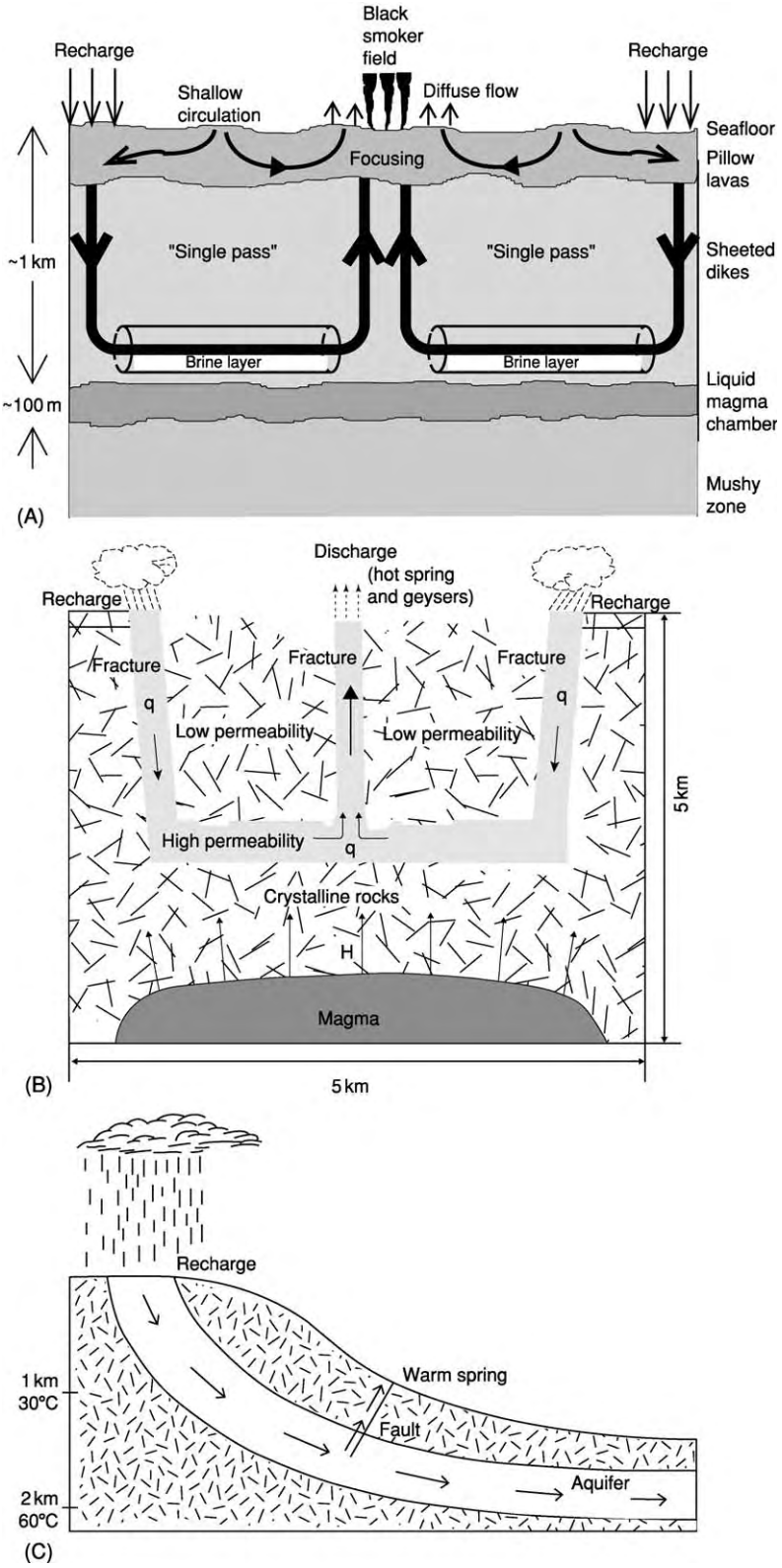


Figure 3 (A) Cartoon of single pass hydrothermal circulation model at an ocean ridge crest. The major single pass segment refers to a deep circulation cell in which seawater recharge penetrates through sheeted dykes to near the top of a magma body, takes up heat and undergoes water rock chemical reactions while flowing quasihorizontally, and ascends through faults or fractures to the seafloor as high temperature focussed black smoker flow. Mixing of the deep circulation with shallow cooler circulation in the basaltic pillow lavas may result in diffuse discharge. Reproduced from Germanovich LN, Lowell RP, and Astakhou DK (2000) Stress dependent

(*see* **Geysers and Hot Springs**). It is the nature of the heat source that generally determines whether hydrothermal activity occurs at high (>150°C) or low temperature. The circulation system consists of a recharge zone through which fluids enter the crust, a region in which the fluid takes up heat from its surroundings, and a discharge zone, through which the heated hydrothermal fluid emerges at the surface as a hot spring or hydrothermal vent. Although fluid may sometimes recirculate several times before exiting the system, it is often convenient to describe circulation in terms of a simple single-pass circulation model. Figure 3 shows cartoons of single-pass models envisioned for high-temperature terrestrial and submarine systems and a low-temperature warm spring system. In addition, all hydrothermal activity exhibits temporal variability, and chemical reactions between the circulating fluid and rock are often important.

Heat

Geothermal gradient Conductive heat flux, H, is related to the geothermal gradient by $H = \lambda\ dT/dz$, where λ is the thermal conductivity. For rocks, λ ranges from approximately 1.8 to 5 W $(m\,^{\circ}C)^{-1}$, with most igneous and metamorphic rocks falling into a narrower range between 2.0 and 2.5 W $(m\,^{\circ}C)^{-1}$. In older, stable continental cratons, the geothermal gradient may be as low as $10^{\circ}C\,km^{-1}$, whereas in active volcanic regions it may be more than $100^{\circ}C\,km^{-1}$. A typical geothermal gradient of $\approx 25^{\circ}C\,km^{-1}$ gives a conductive heat flux of $\approx 60\,mW\,m^{-2}$.

In terrestrial low-temperature hydrothermal activity, fluids driven by a topographic head circulate to a depth of ~1–3 km in the crust where they are heated by the geothermal gradient. The fluids emerge through faults at the surface as warm or hot springs with temperatures ranging from a few tens of degrees above ambient to the local surface boiling temperature (Figure 3C). Such springs are found worldwide in areas of both normal and elevated heat flow.

Low-temperature hydrothermal circulation in oceanic crust occurs from ridge axes to a lithospheric age of ~60 My. This circulation is partially controlled seafloor topography in combination with the geothermal gradient, with discharge occurring at highs and recharge occurring at topographic lows. Type and thickness of sediment cover also influences this circulation. More than 90% of all hydrothermal heat loss from the seafloor occurs at low temperature. This circulation impacts geochemical cycles as the equivalent of an ocean volume approximately evens 10^6 years.

Magmatic heat High-temperature hydrothermal activity (typically classified as > 150°C) is associated with active volcanism. In these settings, shallow magmatic intrusions provide the heat source. Part of this heat comes from the latent heat of crystallization and part of the heat is derived from the cooling pluton. Thermal buoyancy differences between the colder and hotter parts of the system drive convective fluid motions. As volcanism is associated with ocean ridges, hot spots, and island arc systems (fore-arc, arc, and back-arc settings) at subduction zones, it is not surprising that essentially all high-temperature hydrothermal activity occurs in these regions (Figure 1).

In terrestrial settings, boiling hot springs and geysers provide the surface expressions high-temperature hydrothermal activity. Reservoir temperatures of these systems typically lie between 200 and 350°C. In oceanic settings vigorous high-temperature hydrothermal activity is exhibited as "black smoker" venting at temperatures between 300 and 400°C (Figure 4). Lower temperature "white smokers" with temperatures ~150–200°C are also common. Because of the high pressure (~250 bars) at the seafloor, these high-temperature vents lie below the boiling temperature. As discussed later, however, boiling and phase separation appear to occur in the subsurface of both terrestrial and submarine high-temperature hydrothermal systems.

Chemical heat It has long been recognized that hydration of peridotite is an exothermic reaction that produces heat, that alters the chemistry of the rocks and hydrating solutions involved, and that expands the volume of the rocks (~40%). It is only now emerging how widespread this process called the "serpentinization reaction" may be beneath ocean basins and possibly continents. The reaction involves peridotite, the characteristic ultramafic rock type of the Earth's upper mantle, and either seawater or meteoric water. Serpentinization is commonly observed in ultramafic rocks recovered from the seafloor and in slices of ancient oceanic mantle exposed on land as ophiolites. This reaction yields distinctive chemical solutions characterized by high alkalinity, high ratios of Ca to Mn and other metals, and abiogenic

permeability and the formation of seafloor event plumes. *Journal of Geophysical Research* 105: 8341 8354. (B) Analogous cartoon representing single pass flow in a high temperature terrestrial system. Adapted from White (1973) Characteristics of geothermal resources. In: Kruger P and Otte C (eds.) *Geothermal Energy*. Stanford, CA: Stanford University Press. (C) Cartoon of a terrestrial low temperature warm spring system. Reproduced from Lowell (1992) Hydrothermal systems. In: *Encyclopedia of Earth System Science*, vol. 2, pp. 547 557. San Diego, CA: Academic Press.

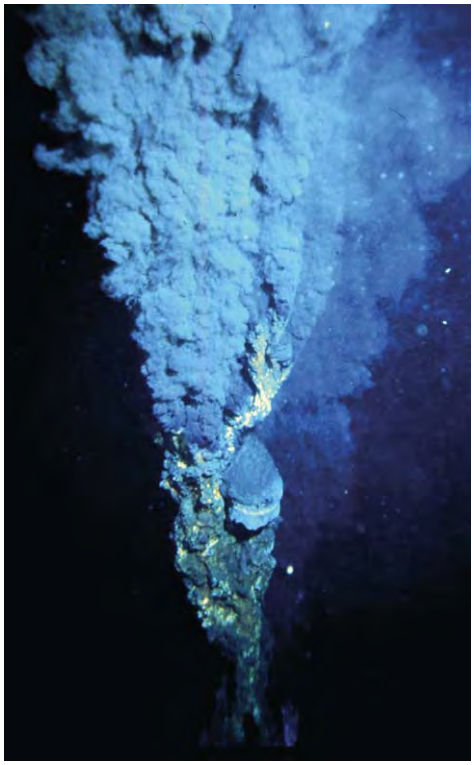


Figure 4 Black smoker vent at the East Pacific Rise 21° N hydrothermal field. (© Woods Hole Oceanographic Institution, Woods Hole, Massachusetts.)

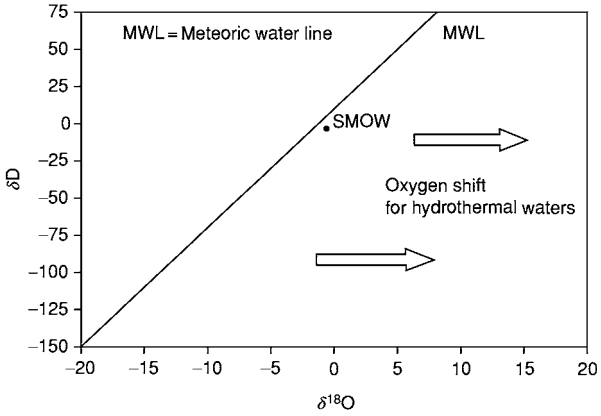


Figure 5 Schematic of δD δ^{18}O relationship in meteoric waters. Reproduced from Craig (1961). Horizontal arrows indicate the δ^{18}O shift generally found in hydrothermal fluids that results from isotopic exchange with δ^{18}O enriched igneous and metamorphic rocks.

generation of methane (CH_4) and hydrogen (H_2) gas. The heat released, depending on the volume and rate of serpentinization, may be sufficient to drive hydrothermal circulation over a range of fluid temperatures, typically low to intermediate (degrees to tens of degrees Celsius), and possibly up to several hundred degrees Celsius.

Serpentinization is favoured by conditions that facilitate access of water to large volumes of the upper mantle. In ocean basins the conditions include a low magma budget, which produces thin ocean crust, and tectonic extension and volume expansion that creates permeability through fractures and faults and that exposes rocks of the upper mantle on the seafloor. Such conditions generally occur at sections of slow-spreading ocean ridges in the Atlantic, Indian, and Arctic oceans. For example, fluids with the chemical signatures of serpentinization reactions are common along the mid-Atlantic ridge where several high-temperature (to 360°C) seafloor hydrothermal fields (Logatchev at 14°45′ N, 44°58′ W and Rainbow, 36°14′ N, 33°54′ W) at least partially situated in serpentinized ultramafic rocks of the upper mantle have been found. Only one of these fields appears to be an end member of a hydrothermal system entirely driven by serpentinization reactions (Lost City field, near 30° N, 42° W) located about 15 km west of the eastern intersection of the rift valley with the Atlantis Fracture Zone, where the field is apparently isolated from magmatic heat sources. There serpentinization-derived fluids are discharging at temperatures up to 75°C and precipitating calcium carbonate and magnesium hydroxide chimneys, which have grown up to 60 m high. Thermal and chemical fluxes from such serpentinization-driven seafloor hydrothermal systems have yet to be determined, but may be a significant fraction of global hydrothermal mass and heat budgets. Seawater and upper mantle rocks are ubiquitous in ocean basins, although the sites of serpentinization may be localized.

Fluid Sources

The aqueous fluid involved in hydrothermal activity can, in principle, have several different origins. Meteoric waters are the predominant fluid, but metamorphic or magmatic fluids may also contribute. The origin of the fluid is generally determined by examining their oxygen and hydrogen isotopic ratios (Figure 5). Meteoric waters are defined by a characteristic linear relationship between δD and δ^{18}O (MWL), whereas metamorphic and magmatic rocks and waters tend to be enriched in δ^{18}O relative to the MWL, and hence lie to the right of the curve. Ocean waters occupy a small range of δD–δ^{18}O space near the MWL (denoted by SMOW). Hydrothermal source waters are typically meteoric (or seawater) and hence lie somewhere along the MWL. At temperatures greater than 200°C, hydrothermal waters may be enriched in δ^{18}O as a result of isotopic exchange during water–rock reactions. The presence of a magmatic or metamorphic component may also move the

isotopic signature of the fluid to the right of the MWL (Figure 5). Isotopic evidence of a magmatic component in active hydrothermal systems is generally inconclusive, but the presence of CO_2 in some hydrothermal fluids points to the presence of magmatic volatiles. Seawater is by far the predominant fluid in submarine hydrothermal systems.

The Circulation System

Heat transport by fluid flow through the rock requires interconnected fluid pathways and a driving force for fluid flow. The relationship between the driving force, the gradient of hydraulic head $\partial\hat{H}/\partial x_j$, and volumetric flow rate per unit area per unit time, or specific discharge q_i, is generally given by the empirical relationship called Darcy's Law.

$$q_i = (gk_{ij}/v)(\partial\hat{H}/\partial x_j) \qquad [1]$$

where g is the acceleration due to gravity, v is the kinematic viscosity of the fluid, and k_{ij} is the intrinsic rock permeability tensor, respectively. Subscripts, i, j refer to the Cartesian coordinate directions. In many applications k_{ij} is treated as a scalar k; the units of k_{ij} are m^2.

Rock permeability is the single most important physical parameter that affects hydrothermal circulation. This parameter is a measure of the interconnectivity of pore spaces and fractures; however, these features and their interconnectivity may depend upon physical and chemical processes related to the flow itself. Consequently, in any given hydrothermal environment, rock permeability may be a complex function of time and space that is difficult to determine *in situ* at the field scale. Moreover, permeability is often heterogeneous and anisotropic; it is a scale-dependent parameter that may vary over several orders of magnitude on relatively small spatial scales. Considerable research effort has been devoted to the determination of permeability and its temporal evolution during hydrothermal activity. In the following subsections we discuss some approaches to describing permeability in hydrothermal systems and its temporal evolution.

Porous medium permeability The percentage of rock volume that may be occupied by fluid is termed the porosity. To the extent that this porosity is interconnected it may give rise to permeable pathways for fluid flow. Such porosity-related permeability is termed primary permeability. There have been several mathematical models attempting to relate effective, or interconnected, porosity, ϕ, with permeability, but these have had limited success. A mathematical relationship between porosity and permeability is highly desirable for two reasons. First, porosity is scale independent and therefore laboratory-based measurements of porosity are meaningful; secondly, *in situ* porosity can be estimated from both electrical and seismic data.

Mathematical models relating effective porosity ϕ to a scalar bulk permeability k are generally of the form

$$k = Cb^2\phi^n \qquad [2]$$

where b is the average grain size of the medium and C is a numerical constant, respectively. The exponent n ranges between 2 and 3 in most formulations. The well-known Carmen-Kozeny relation is similar to eqn [2]. Equations of the form [2] often fail in practice because in most systems, even those that have significant primary permeability, field-scale permeability is controlled by fractures.

Fracture- and fault-related permeability Permeability in essentially all hydrothermally active regions is controlled by fractures and faults. Such permeability is termed secondary. In igneous and metamorphic rocks, which host most high-temperature hydrothermal activity, cracks must provide the main permeability because porosity and, hence, primary permeability is low. When permeability is controlled by fractures, large permeability can exist in the presence of very low interconnected porosity. In fracture-controlled systems, permeability is related to crack density, the abundance of crack intersections and the cube of the crack aperture.

A generalized formulation can be written as

$$k = C'\frac{l^3}{h}Na^2 \qquad [3]$$

where C' is a dimensionless coefficient describing the degree of crack interconnectivity, l is the mean crack aperture, h is the crack spacing, a is the crack length, and N is the number of cracks per unit area. As special case of [3], one may consider a set of planar parallel cracks of aperture l and spacing h. In this case the permeability is

$$k = \frac{l^3}{12h} = \frac{h^2}{12}\phi^3 \qquad [4]$$

where the porosity $\phi = l/h$. Figure 6 depicts k vs ϕ for selected values of h; the results show that large values of crack permeability can exist for $\phi \leq 1\%$. These values are several orders of magnitude greater than 10^{-18}–$10^{-20}\,m^2$, which are typical laboratory values for unfractured granite.

Estimates of permeability for hydrothermally active regions have been determined from borehole

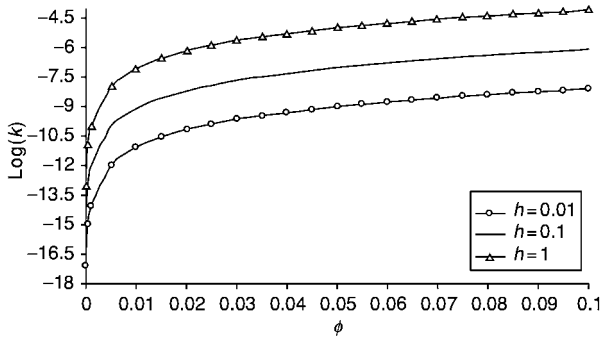


Figure 6 Graph of permeability versus porosity at a number of given fracture spacings for rock permeability resulting from planar parallel fractures. The curves show that high fracture permeability can occur in low porosity rocks.

measurements and from field measurements of crack distributions in fossil systems. Field-scale permeability has also been estimated from mathematical modelling of hydrothermal heat output. Measurements made in Deep Sea Drilling Project and Ocean Drillng Project boreholes give permeability values ranging from 10^{-18} m^2 in sheeted dykes to as high as 10^{-10} m^2 in pillows. Values in continental hydrothermal systems often fall between 10^{-12} and 10^{-15} m^2. Crack spacing and apertures in ophiolites yield permeability values ranging between 10^{-13} and 10^{-8} m^2. Mathematical modelling studies of high-temperature venting at ocean ridge crests give a similar range. The high values of permeability and the broad range of values estimated from field and modelling studies further indicate that permeability in hydrothermal regions is fracture-controlled.

Fracture-controlled permeability is typically heterogeneous and possibly anisotropic. Fracture concentration and orientation may result from tectonic stresses as well as processes related to magma emplacement and volcanic eruptions and thermal stresses. Zones of high fracture-controlled permeability are associated with dikes. In both continental and submarine hydrothermal systems discharge zones are often focussed along tectonic faults. Recharge zones are more problematical, but faulting could be important there as well.

Although fractures and faults control hydrothermal circulation patterns, one does not often consider flow in discrete fractures. Mathematical models usually treat flow in fractured rock as an equivalent porous medium and apply Darcy's Law. It is important in this regard, however, to recognize that fracture permeability can exist on several spatial scales and that the permeability may depend on time.

Temporal variations in permeability Temporal changes in permeability can result from several mechanisms. In addition to changes in response to tectonic and magmatic processes, both thermal and chemical processes can be significant. Because the processes have not been quantified in great detail, their relative importance is uncertain.

As circulating aqueous fluids encounter different pressure and temperature environments dissolution and precipitation of chemical constituents may occur. During water–rock reactions, hydrothermal fluids reach thermodynamic equilibrium with quartz. At pressures of a few hundred bars, the solubility of quartz reaches a maximum between 350 and 400°C. Thus if the hydrothermal solution is heated above 400°C quartz will precipitate and clog fractures and pore spaces. Similarly as the hydrothermal solution ascends towards the surface, both lower pressures and temperatures in the environment will foster quartz precipitation. Both quartz and amorphous silica are common vein minerals in hydrothermal systems, and precipitation of these phases may exert a strong influence on hydrothermal circulation over time. The development of a low permeability barrier as a result of quartz precipitation is likely an important factor in the evolution of vapour-dominated hydrothermal systems.

In submarine hydrothermal systems precipitation of anhydrite may be important in both recharge and discharge zones. Because the solubility of anhydrite decreases with increasing temperature, heating of seawater during hydrothermal recharge to $T \geq 150$°C results in precipitation of anhydrite in recharge zones (**Figure 3A**). Sulphate is removed from seawater by precipitation of anhydrite and reaction with crustal rocks; however, upon ascent, mixing of sulphate-poor, hot hydrothermal fluid with cold sulphate-rich seawater may again result in the precipitation of anhydrite. Mixing in the subsurface may contribute to focussing seafloor venting into black smokers; whereas mixing above the seafloor contributes to the formation of chimney structures.

Thermoelastic stresses result from the passage of either cold fluids through initially hotter rock or hot fluid through cooler rocks. In the former case, cooling of rock surfaces leads to thermal contraction and the enhancement of permeability. In the latter case, heating of the rock leads to thermal expansion and reduction of permeability. The dependence of permeability on temperature can be expressed as

$$k = k_0[1 - \gamma(T - T_0)]^3 + k_{\text{res}} \qquad [5]$$

where k_0 is the permeability of the main crack network, k_{res} is a finer-scale residual permeability, and γ is factor expressing the strength of the thermoelastic effect.

At temperatures exceeding 350–400°C, rocks begin to exhibit ductile behaviour. Although this behaviour depends upon the rate which stresses are applied, ductile behaviour will tend to seal cracks. Thus permeable pathways that may initially be opened by stresses resulting in brittle failure may gradually close. This process may limit the depth to which cracks remain open in the crust and the extent to which hydrothermal circulation may approach magma bodies.

Water–Rock Chemical Reactions

As aqueous fluids pass through hot subsurface rocks, chemical reactions occur. Some chemical constituents may be removed from the fluid, others may be extracted from the rock. The reactions may also involve isotopic exchange between the fluid and rock. These reactions are complex functions of temperature, pressure, lithology, permeability structure, duration of activity, and other factors. A detailed discussion of this topic is beyond the scope of this article; however, the use of geochemical thermometers and the formation of hydrothermal ore deposits are discussed briefly.

Geochemical thermometers The strong temperature dependence of solubility of certain chemical constituents in hydrothermal fluids, the temperature dependence of elemental partitioning between rock and solution, and the temperature dependence of isotopic partitioning between mineral and fluid phases have led to the development of a variety of geochemical thermometers to deduce subsurface conditions from surface samples. The quartz geothermometer utilizes the strong temperature dependence of quartz solubility and the slow kinetics of quartz precipitation at low temperature. As hydrothermal solutions in equilibrium with quartz at high temperature rise to the surface and cool, the high degree of disequilibrium in the measured quartz concentration permits a calculation of the equilibrium temperature at depth Other common geothermometers include:

1. Na/K, which makes use of the temperature dependence of partitioning of these elements between aluminosilicate rocks and hydrothermal fluid;
2. Na–K–Ca, which includes the effect of Ca in the partitioning; and
3. ratios of stable isotopes such as $\delta^{13}C$, $\delta^{18}O$, δD, and $\delta^{34}S$.

Various factors affect the resolution and reliability of each of these geothermometers, so often many independent ones are used.

Ore deposits–fossil hydrothermal systems As a result of water–rock chemical reactions at intermediate to high temperature, trace metallic ore-forming metals such as Fe, Cu, Zn, Sb, Au, Ag, and Pb are transferred from the rock to the hydrothermal solution. Because most of these metals form metallic sulphides that are highly insoluble in water, solubility is achieved by the formation of bisulphide or chloride ion complexes. Various mechanisms may cause local precipitation of these metal–ion complexes, resulting in a concentrated accumulation of metallic ore. A rapid drop in the solution temperature because of thermal conduction or mixing with cooler fluids, a change in solution pH, and boiling can all lead to ore deposition.

Many types of ore deposits in the geological record, such as porphyry ore deposits associated with silicic volcanism, Mississippi Valley-type lead-zinc deposits, and volcanically hosted massive sulphide deposits are linked to hydrothermal activity. Such ore deposits thus present an integrated fossil record of hydrothermal activity, and provide a window into subsurface heat transfer and fluid flow processes. By coupling this integrated fossil record with studies of active ore-forming processes on the seafloor and in other active hydrothermal environments, one can obtain a more complete picture of hydrothermal activity (*see* **Mining Geology:** Hydrothermal Ores).

Temporal Variability in Hydrothermal Activity

Temporal variability on a range of time-scales is a fundamental characteristic of hydrothermal activity (Table 2). Some of this variability is linked to episodicity in magmatic and tectonic activity, or climate changes. The occurrence of these processes ranges from scales of plate reorganization of $\sim 10^6$–10^7 years to magma replacement times of $\sim 10^1$–10^4 years at fast and slow spreading ridges, respectively. Temporal variability related to the fluid circulation system, mainly resulting from changes in crustal permeability, occurs on time-scales ~ 1–10^2 years. Seafloor hydrothermal activity is known to change on time-scales of hours to months following earthquakes, igneous intrusions (e.g., dykes), or volcanic eruptions. Climate changes may alter precipitation patterns, and hence fluid recharge, on time-scales of 10–10^3 years; ice ages and glaciation may affect high-altitude systems on similar time-scale.

Two-Phase Flow

Boiling and phase separation commonly occur in high-temperature hydrothermal systems. For pure water, boiling is defined by the boiling point curve as a function of pressure. Liquid phase occurs below

Table 2 Time scale of events and processes related to hydrothermal activity[a]

Time scale	*Activity or process*
10^6 10^7 years	Plate reorganization
10^6 10^7 years	Episodes of seafloor spreading
10^5 years	Magnetic polarity interval
10^1 10^6 years	Duration of ore formation processes
10^3 10^4 years	Eruption cycle on slow spreading ridges
10^1 10^3 years	Eruption cycle on fast spreading ridges
10^3 10^4 years	Glacial episodes
10^3 10^6 years	Duration of hydrothermal activity
10^1 10^3 years	Episodes of climate change
10^0 10^2 years	Duration of individual seafloor hydrothermal vent
10^0 10^1 years (hours to decade)	Duration of volcanic eruption
10^0 10^1 years	Residence time of hydrothermal fluid in oceanic crust
10^4 10^7 s	Transit time of upwelling hydrothermal fluid
10^5 10^7 s	Duration of earthquake swarms
10^5 10^7 s	Duration of dyke emplacement event
10^5 10^6 s	Duration of seafloor event plume
10^3 10^6 s	Period of tidal signals
0.1 3 s	Precipitation of sulphide particles during mixing of high *T* hydrothermal fluid with ambient seawater

[a]Modified from Lowell RP, Rona PA, and Von Herzen RP (1995) and Rona (1988).

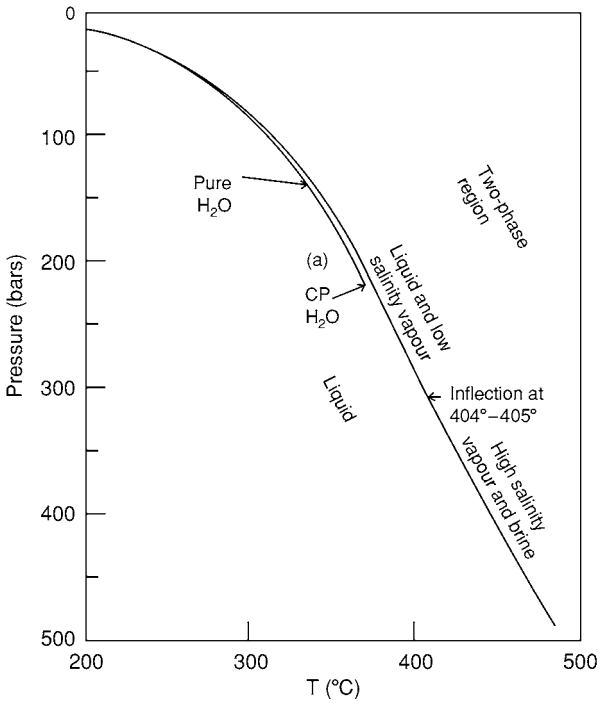


Figure 7 Boiling point curve for pure water compared with that for seawater. Note that the pure water curve ends at the critical point. The region above the pure water curve is pure vapour and the two phases only exist along the boiling curve. Reproduced from Bischoff and Rosenbauer (1984) The critical point and two phase boundary of seawater, 200 500°C. *Earth and Planetary Sci ence Letters* 68: 172 180.

the boiling point and the vapour phase occurs above it. The two phases are in equilibrium along the curve, and the volume fractions of each phase along the curve depend upon the enthalpy of the system. The phase diagram for water (Figure 7) most readily shows these relations. A key feature of the pure water phase diagram is the critical point defined by $P_c = 218$ bars, $T_c = 374°C$. Above the critical point only a single-phase "water substance" exists.

Most hydrothermal aqueous fluids contain some amount of dissolved salts; however, the presence of these dramatically alters the two-phase behaviour of water. First of all, both P_c and T_c increase as the salt content increases. For seawater, which can be represented by salinity $x \approx 3.2\%$ NaCl solution, $P_c \approx 300$ bars, $T_c \approx 405°C$. Secondly, two-phase flow is defined by a region of *P-T-x* space rather than a curve. Thirdly, because fluid density is a function of salinity, the fractionation of salt between the liquid and vapour phases affects the dynamics of the phases. Figure 8 depicts part of the phase diagram for NaCl–water solution.

In terrestrial systems subsurface boiling results in either liquid- or vapour-dominated systems. In liquid-dominated systems, fluid pressures are near hydrostatic. Hot spring fluids are generally neutral to alkaline pH and chloride rich. Liquid and vapour phases are intermingled within the two-phase zone, and the two-phase zone overlies a single-phase fluid. By contrast, in vapour-dominated systems low, nearly uniform, vapour-static fluid pressure occurs over a considerable thickness. Fluid discharge occurs at low pH and low chloride. The presence of a region of underpressure implies a permeable barrier between the vapour-dominated zone and the surrounding cold recharge. Vapour-dominated systems act as a heat pipe with near zero net mass flux; heat is carried upwards by high enthalpy vapour while small amounts of low-enthalpy liquid flow downwards. Most systems are liquid-dominated, including the geyser basins of Yellowstone National Park, USA, Wairakei and Broadlands, NZ, and Ahuachapan, MX; vapour-dominated systems include Geysers, USA, Lardarello, IT, Kamodjang, IND, and Matsukawa, JP (Figure 1).

In submarine systems venting black smoker fluids are mostly in the liquid phase; however, the chlorinity of vent fluids seldom corresponds to seawater (≈540 mmol kg). Rather it ranges from ≈30 to ≈1200 mmol kg (Table 3). The departure of vent salinity from that of seawater is attributed to phase separation. The lowest chlorinity values often occur shortly after magmatic eruptions or diking events and thus are indicative of active phase separation.

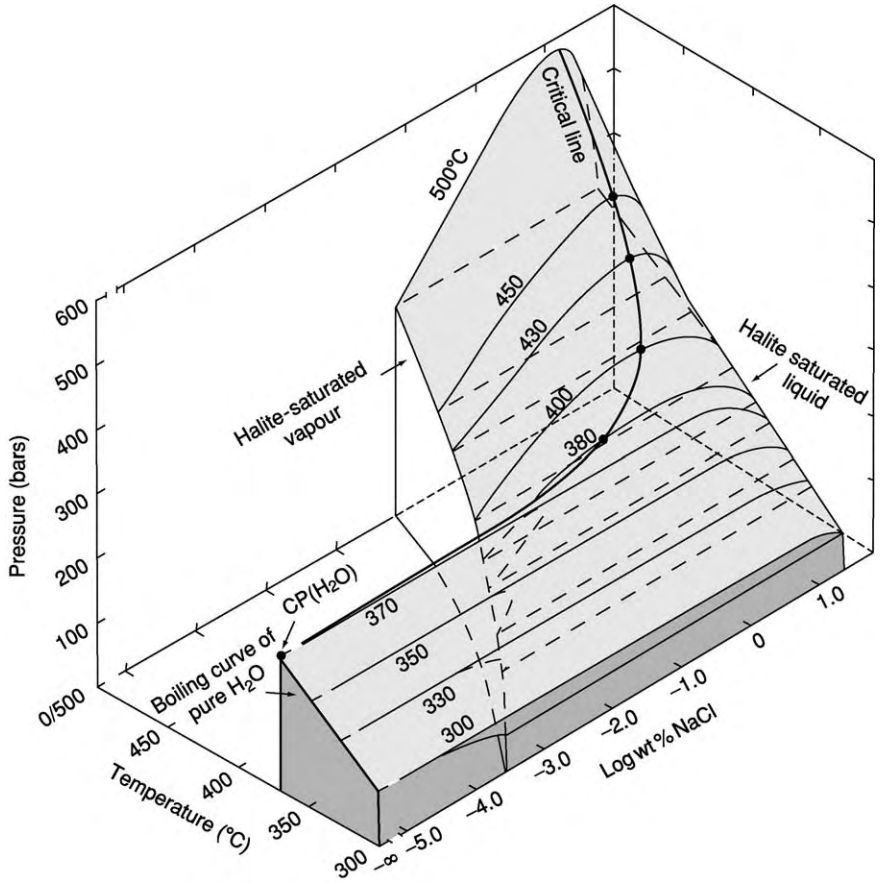


Figure 8 Three dimensional perspective of the NaCl H_2O phase diagram between 300 and 500°C. Reproduced from Bischoff and Pitzer (1989) Liquid vapor relations for the system NACL H_2O: Summary of the P T x surface from 300° to 500°C. *American Journal of Science* 289: 217 248.

Table 3 Chlorinity of selected high temperature seafloor hydrothermal Vents[a]

Vent site	*Year(s) sampled*	*Value (mmol kg ^{1})*[b]
East Pacific rise		
9 10° N	1991	32 860
21° N	1979, 81, 85	489 579
Juan de Fuca ridge		
North cleft	1990 92	730 1245
South cleft	1984	896 1087
Endeavour	1984 88	253 505
Axial volcano	1986 88	176 624
Mid Atlantic ridge		
TAG	1986	659
MARK	1986	559
Lau basin	1989	650 800

[a]Modified from Von Damm (1985).
[b]Normal seawater chlorinity 540 mmol kg ^{1}.

Phase separation also results in the formation of saline brines that, because of their high density, sink towards the base of the system. Later mixing of these saline brines with seawater may result in vent chlorinity greater than seawater. The formation of a brine layer at the base of a hydrothermal system may act as a thermal conductive barrier between the overlying hydrothermal circulation and the magma body (Figure 3A) and be a salinity source for saline vent fluids.

Future Directions

Hydrothermal activity represents an exciting dynamic area for future research. This is particularly true for submarine systems because of their links to studies of the origin of life, life in extreme environments, and the continued discovery of novel types of hydrothermal activity. The detailed sampling and data analysis and continued exploration for serpentinization-driven hydrothermal activity will likely grow during the next decade. At ridge crest and volcanic island arc systems, advances in ocean drilling technology, remote and autonomous sensing devices, long-term monitoring, integrated interdisciplinary experiments at various well-characterized seafloor

sites, and improvements in mathematical modelling techniques will stimulate the science over the next decade. Finally, we believe that analysis of seafloor hydrothermal activity over geological time, and attempts to discern the importance of hydrothermal activity elsewhere in the solar system will emerge as an important endeavour. Such studies are necessary to understand the links between hydrothermal activity and life.

In terrestrial systems continued exploitation as an energy resource will be important. Moreover, climatically induced precipitation changes, because of the link to hydrothermal recharge, may alter warm spring and geyser behaviour. Utilization of hydrothermal activity as a climate monitor has yet to receive attention.

See Also

Geysers and Hot Springs. **Igneous Processes**. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Origin of Life**. **Plate Tectonics**. **Tectonics:** Faults.

Further Reading

Anderson RN, Langseth MG Jr, and Sclater JG (1977) The mechanisms of heat transfer through the floor of the Indian Ocean. *Journal of Geophysical Research* 82: 3391 3409.

Bischoff JL and Pitzer KS (1989) Liquid vapor relations for the system NACL H_2O: Summary of the P T x surface from 300° to 500°C. *American Journal of Science* 289: 217 248.

Bischoff JL and Rosenbauer RJ (1984) The critical point and two phase boundary of seawater, 200 500°C. *Earth and Planetary Science Letters* 68: 172 180.

Craig H (1961) Standard for reporting concentrations of deuterium and oxygen 18 in natural waters. *Science* 133: 1833 1834.

Elder J (1981) *Geothermal Systems*. San Diego, CA: Aca demic Press.

Elderfield H and Schultz A (1996) Mid ocean ridge hydro thermal fluxes and the chemical composition of the ocean. *Annual Reviews of Earth and Planetary Science* 24: 191 224.

Germanovich LN, Lowell RP, and Astakhov DK (2000) Stress dependent permeability and the formation of sea floor event plumes. *Journal of Geophysical Research* 105: 8341 8354.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (eds.) (1995) *Seafloor Hydrothermal Systems: Physical, Chemical, and Biological Interactions*, AGU Geophysical Monograph 91. Washington, DC: American Geophysical Union.

Kelley DS, Baross JA, and Delaney JR (2002) Volcanoes, fluids and life at mid ocean ridge spreading centers. *Annual Review of Earth and Planetary Science* 30: 385 491.

Kruger P and Otte C (eds.) (1973) *Geothermal Energy*. Stanford, CA: Stanford University Press.

Lowell RP (1991) Modeling continental and submarine hydrothermal systems. *Reviews of Geophysics* 29: 457 476.

Lowell RP (1992) Hydrothermal systems. In: Nierenburg WA (ed.) *Encyclopedia of Earth System Science*, vol. 2, pp. 547 557. San Diego, CA: Academic Press.

Lowell RP, Rona PA, and Von Herzen RP (1995) Seafloor hydrothermal systems. *Journal of Geophysical Research* 100: 327 352.

Rona PA (1988) Hydrothermal mineralization at oceanic ridges. *Canadian Mineralogist* 26: 431 465.

Sclater JG, Jaupart C, and Galson D (1980) The heat flow through oceanic and continental crust and the heat loss of the Earth. *Review of Geophysics* 18: 269 311.

Von Damm KL (1995) Controls on the chemistry and tem poral variability of seafloor hydrothermal fluids. In: Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (eds.) *Seafloor Hydrothermal Systems: Physical, Chemical, and Biological Interactions*, AGU Geophysical Monograph 91, pp. 222 247. Washington, DC: American Geophysical Union.

White DE (1973) Characteristics of geothermal resources. In: Kruger P and Otte C (eds.) *Geothermal Energy*. Stan ford, CA: Stanford University Press.

Mid-Ocean Ridges

K C Macdonald, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

The mid-ocean ridge system is the largest mountain chain and the most active system of volcanoes in the solar system. In plate-tectonic theory, the ridge is located between plates of the Earth's rigid outer shell that are separating at speeds of approximately 10–170 mm year ^{1} (up to 220 mm year ^{1} in the past). The ascent of molten rock from deep within the Earth (*ca*. 30–60 km) to fill the void between the plates creates new seafloor and a volcanically active ridge. This ridge system wraps around the globe like the seam of a baseball and is approximately 70 000 km long (including the lengths of ridge offsets, such

as transform faults). Yet the ridge itself is only about 5–30 km wide, very small compared with the plates, which can be thousands of kilometres across (Figure 1).

Early exploration showed that the gross morphology of spreading centres varies with the rate of plate separation. At slow spreading rates (10–40 mm year ^{1}) a rift valley 1–3 km deep marks the axis, while for fast spreading rates (more than 90 mm year ^{1}) the axis is characterized by an elevation of the seafloor of several hundred metres called an axial high (Figure 2). The rate of magma supply is a second factor that may influence the morphology of mid-ocean ridges. For example, a very high rate of magma supply can produce an axial high even where the spreading rate is slow; the Reykjanes Ridge south of Iceland is a good example. Also, for intermediate spreading rates (40–90 mm year ^{1}) the ridge crest may have either an axial high or a rift valley depending on the rate of magma supply. The depth to the seafloor increases from a global average of approximately 2600 m at the spreading centre to more than 5000 m beyond the ridge flanks. The rate of deepening is proportional to the square root of the age of the seafloor because it is caused by the thermal contraction of the lithosphere. Early mapping efforts also showed that the mid-ocean ridge is a discontinuous structure, which is offset at right angles to its length by numerous transform faults that are tens to hundreds of kilometres long.

Maps are powerful: they inform, excite, and stimulate. Just as the earliest maps of the world in the sixteenth century ushered in a vigorous age of exploration, so the first high-resolution continuous-coverage maps of the mid-ocean ridge system stimulated investigators from a wide range of fields, including petrologists, geochemists, volcanologists, seismologists, tectonicists, and practitioners of marine magnetics and gravity, as well as researchers outside the Earth sciences, including marine ecologists, chemists, and biochemists. Marine geologists have found that many of the most revealing variations are observed by exploring along the axis of the active ridge. This along-strike perspective has revealed the architecture of the global rift system. The ridge axis undulates up and down in a systematic way, defining a fundamental partitioning of the ridge into segments bounded by a variety of discontinuities. These segments behave like giant cracks in the seafloor, which can lengthen or shorten and have episodes of increased volcanic and tectonic activity. In fact, elementary fracture mechanics can be used to explain the interaction between neighbouring ridge segments.

Another important change in perspective came from the discovery of hydrothermal vents by marine geologists and geophysicists. It became clear that,

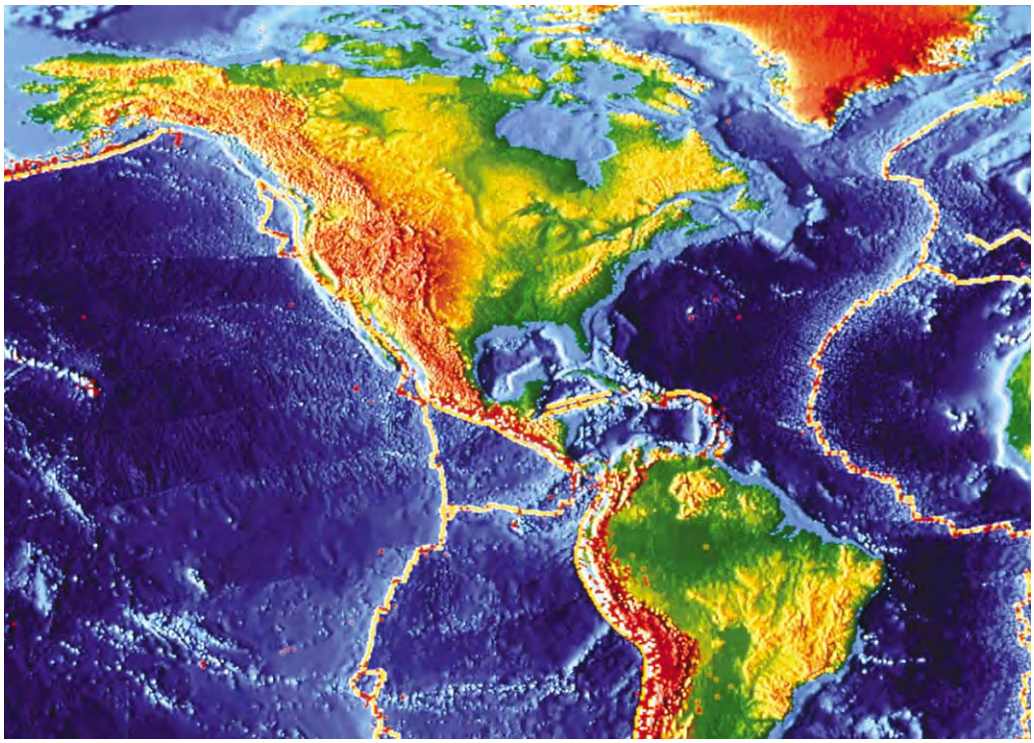


Figure 1 Shaded relief map of the seafloor showing parts of the East Pacific Rise, a fast spreading centre, and the Mid Atlantic Ridge, a slow spreading centre (courtesy of the National Geophysical Data Centre).

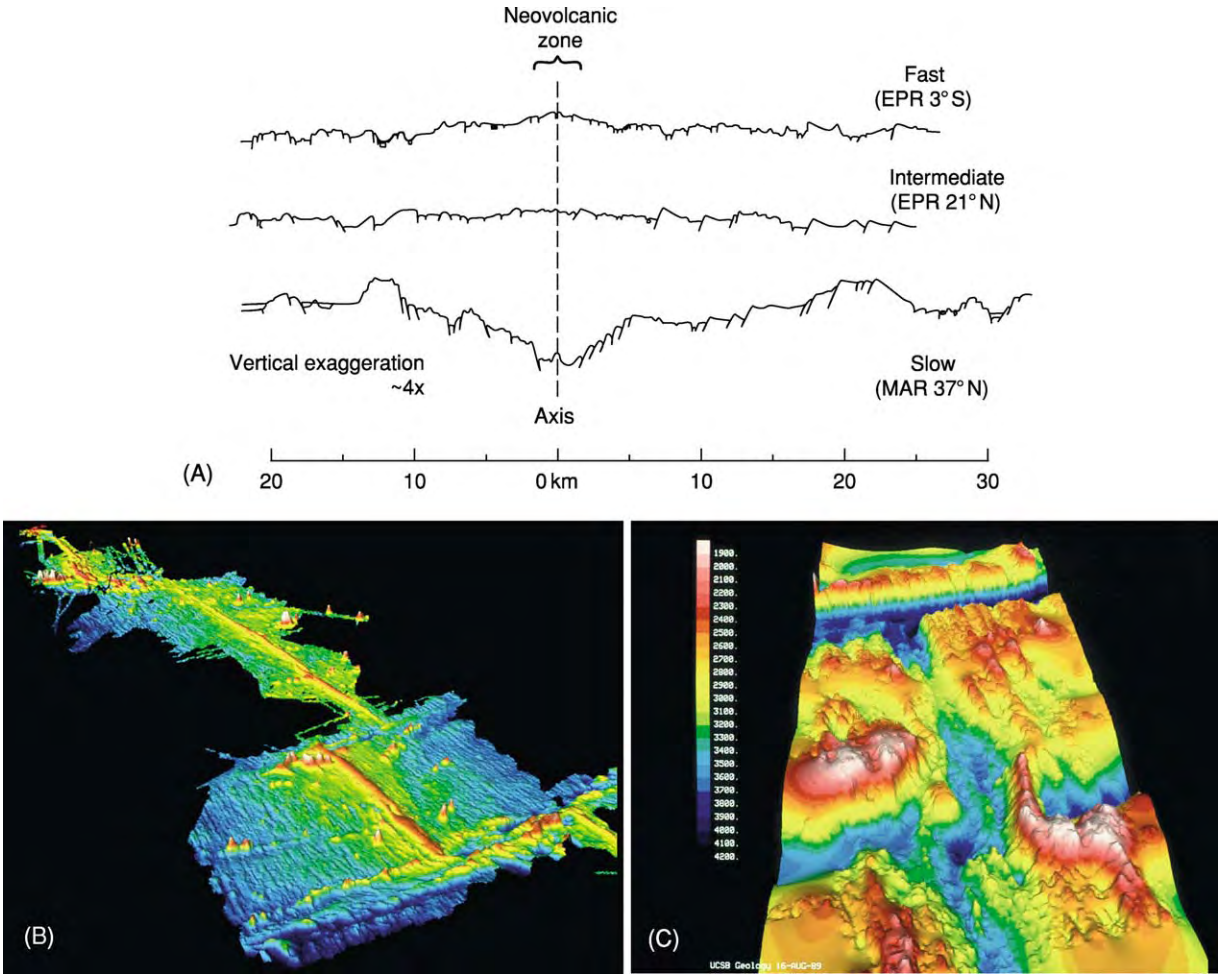


Figure 2 Topography of spreading centres. (A) Cross sections of typical fast intermediate and slow spreading ridges based on high resolution deep tow profiles. The neovolcanic zone (the zone of active volcanism) is indicated and is several kilometres wide; the zone of active faulting extends to the edge of the profiles and is several tens of kilometres wide. (Reproduced from Macdonald KC (1982) Mid ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annual Review of Earth and Planetary Sciences* 10: 155 190.) EPR, East Pacific Rise; MAR, Mid Atlantic Ridge. (B) Shaded relief map of a 1000 km stretch of the East Pacific Rise extending from 8° N to 17° N. Here, the East Pacific Rise is the boundary between the Pacific and Cocos plates, which are separating at a 'fast' rate of 120 mm year ^{1}. The map reveals two kinds of discontinuity: large offsets, about 100 km long, known as transform faults, and smaller offsets, about 10 km long, called overlapping spreading centres. Colours indicate depths of 2400 m (pink) to 3500 m (dark blue). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.) (C) Shaded relief map of the Mid Atlantic Ridge. Here, the ridge is the plate boundary between the South American and African plates, which are spreading apart at the slow rate of approximately 35 mm year ^{1}. The axis of the ridge is marked by a 2 km deep rift valley, which is typical of most slow spreading ridges. The map reveals a 12 km jog of the rift valley, a second order discontinuity, and also shows a first order discontinuity called the Cox transform fault. Colours indicate depths of 1900 m (pink) to 4200 m (dark blue). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

in studies of mid-ocean ridge tectonics, volcanism, and hydrothermal activity, the greatest excitement is in the linkages between these different fields. For example, geophysicists searched for hydrothermal activity at mid-ocean ridges for many years by towing arrays of thermistors near the seafloor. However, hydrothermal activity was eventually documented more effectively by photographing the distribution of exotic vent animals. Even now, the best indicators of the recency of volcanic eruptions and the duration of hydrothermal activity are found by studying the characteristics of benthic faunal communities. For example, during the first deep-sea mid-ocean-ridge eruption witnessed from a submersible, divers did not see a slow lumbering cascade of pillow lavas, as observed by divers off the coast of Hawaii. What they

saw was completely unexpected: white bacterial matting billowing out of the seafloor, creating a scene much like a mid-winter blizzard in Iceland, covering the freshly erupted glassy black lava with a thick blanket of white bacterial 'snow'.

Ridge Segmentation

The most recognizable segmentation of mid-ocean ridges is that defined by transform faults. These plate boundaries are usually perpendicular to the ridge segments they offset and are tens to hundreds of kilometres long, although some exceed 1000 km in length (e.g. the Romanche and San Andreas faults). In plate tectonics, a transform fault traces a small circle about the Euler pole of opening between any pair of plates. Thus the transform fault and its off-axis fracture zone traces may be used to determine the pole of opening as well as changes in the pole of opening. At a ridge–transform intersection, normal spreading processes are truncated. Normal faulting predominates on mid-ocean ridges, while strike-slip faulting dominates along transform faults. The transition can be very complex, with normal faults and strike-slip faults occurring along trends that are affected by shear stresses on the transform fault. Crustal accretionary processes are also affected by the juxtaposition of thick cold lithosphere against the end of a spreading segment. This effect increases with the age and thickness of the lithosphere that is sliding past the ridge–transform intersection. Transverse ridges occur along the length of some of the largest transform faults; some of these ridges have been elevated above sea-level for part of their history.

Between major transform faults, the axial depth profile of mid-ocean ridges undulates up and down with a wavelength of tens of kilometres and an amplitude of tens to hundreds of metres at fast-spreading and intermediate-spreading ridges. This pattern is also observed for slow-spreading ridges, but the wavelength of undulation is shorter and the amplitude is larger (Figure 3). In most cases, ridge-axis discontinuities occur at local maxima of the axial depth profile. These discontinuities include transform faults, as discussed above (first order), overlapping spreading centres (second order), and higher order (third and fourth order) discontinuities, which are increasingly short-lived, mobile, and associated with smaller offsets of the ridge (see Table 1 and Figure 4).

A much-debated hypothesis is that the axial depth profile (Figures 3 and 5) reflects the magma supply along a ridge segment. According to this idea, the magma supply is enhanced along shallow portions of ridge segments and is relatively starved at segment ends (discontinuities). In support of this hypothesis

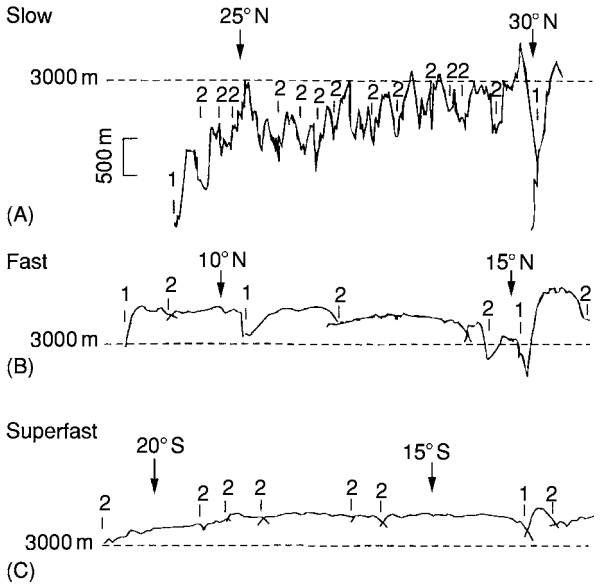


Figure 3 Axial depth profiles for (A) slow spreading, (B) fast spreading, and (C) superfast spreading ridges. Discontinuities of orders 1 and 2 typically occur at local depth maxima (discontinuities of orders 3 and 4 are not labelled here). The segments at faster spreading ridges are longer and have smoother lower amplitude axial depth profiles. These depth variations may reflect the pattern of mantle upwelling. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

is the observation at ridges with an axial high (fast-spreading ridges) that the cross-sectional area or axial volume varies directly with depth (Figure 6). Maxima in the cross-sectional area (more than 2.5 km^2) occur at minima along the axial depth profile (generally not near ridge-axis discontinuities) and are thought to correlate with regions where magma supply is robust. Conversely, small cross-sectional areas (less than 1.5 km^2) occur at local depth maxima and are interpreted to reflect minima in the magma-supply rate along a given ridge segment. On slow-spreading ridges characterized by an axial rift valley, the cross-sectional area of the valley is at a minimum in the mid-segment regions, where the depth is at a minimum. In addition, there are more volcanoes in the shallow mid-segment area, and fewer volcanoes near the segment ends. Studies of crustal magnetization show that very highly magnetized zones occur near segment ends; these are most easily explained by a local restriction of magma supply resulting in the eruption of highly fractionated lavas that are rich in iron.

Multichannel seismic and gravity data support the axial volume–magma supply–segmentation hypothesis (Figure 6). A bright reflector, which is phase-reversed in many places, occurs commonly (>60%

Table 1 Characteristics of segmentation, updated from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986 994 (see references therein). This four tiered hierarchy of segmentation probably represents a continuum in segmentation

	Order 1	*Order 2*	*Order 3*	*Order 4*
Segments				
Segment length (km)	600 ± 300[a] (400 ± 200)[b]	140 ± 90 (50 ± 30)	20 ± 10 $(15 \pm 10?)$	7 ± 5 $(7 \pm 5?)$
Segment longevity (years)	$>5 \times 10^6$	0.5 5×10^6 (0.5 30×10^6)	$\sim 10^4$ 10^5 (?)	$<10^3$ (?)
Rate of segment lengthening (long term migration) (mm y ^{1})	0 50 (0 30)	0 1000 (0 30)	Indeterminate: no off axis trace	Indeterminate: no off axis trace
Rate of segment lengthening (short term propagation) (mm y ^{1})	0 100 (?)	0 1000 (0 50)	Indeterminate: no off axis trace	Indeterminate: no off axis trace
Discontinuities				
Type	Transform, large propagating rifts	Overlapping spreading centres (oblique shear zones, rift valley jogs)	Overlapping spreading centers (intervolcano gaps), devals	Devals, offsets of axial summit caldera (intravolcano gaps)
Offset (km)	>30	2 30	0.5 2.0	<1
Offset age (years)[c]	$>0.5 \times 10^6$ $(>2 \times 10^6)$	0.5×10^6 (2×10^6)	~0	~0
Depth anomaly (m)	300 600 (500 2000)	100 300 (300 1000)	30 100 (50 300)	0 50 (0 100?)
Off axis trace	Fracture zone	V shaped discordant zone	Faint or none	None
High amplitude magnetization?	Yes	Yes	Rarely (?)	No? (?)
Breaks in axial magma chamber?	Always	Yes, except during OSC linkage? (NA)	Yes, except during OSC linkage? (NA)	Rarely
Breaks in axial low velocity zone?	Yes (NA)	No, but reduction in volume (NA)	Small reduction in volume (NA)	Small reduction in volume? (NA)
Geochemical anomaly?	Yes	Yes	Usually	~50%
Break in high temperature venting?	Yes	Yes	Yes (NA)	Often (NA)

[a]Values are ± 1 standard deviation.
[b]Where information differs for slow and fast spreading ridges (<60 mm y ^{1}), it is placed in parentheses.
[c]Offset age refers to the age of the seafloor that is juxtaposed to the spreading axis at a discontinuity.
NA, not applicable; ?, not presently known as poorly constrained; OSC, overlapping spreading centre.

of ridge length) beneath the axial regions of both the northern and southern portions of the fast-spreading and ultra-fast-spreading East Pacific Rise. This reflector has been interpreted as a thin lens of magma residing at the top of a broader axial magma reservoir. The amount of melt is highly variable along strike, varying from a lens that is primarily crystal mush to one that is close to 100% melt. This 'axial magma chamber' reflector is observed where the ridge is shallow and where the axial high has a broad cross-sectional area. Conversely, it is rare where the ridge is deep and narrow, especially near ridge-axis discontinuities. A reflector may occur beneath ridge-axis discontinuities during propagation and ridge-axis realignment, as may be occurring now on the East Pacific Rise near 9° N (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges; Propagating Rifts and Microplates At Mid-Ocean Ridges).

There is evidence that major-element geochemistry correlates with axial cross-sectional area (**Figure 7**). On the East Pacific Rise between 13° S and 21° S there is a good correlation between MgO wt% and cross-sectional area (higher MgO indicates a higher eruption temperature and perhaps a greater local magmatic budget). The abundance of hydrothermal venting (as measured by light transmission and backscatter in the water column and geochemical tracers) also varies directly with the cross-sectional area of the East Pacific Rise. It is not often that one sees a correlation between two such different kinds of measurement. It is all the more remarkable considering that the measurements of hydrothermal activity are sensitive to changes on a time-scale of days to months, while the cross-sectional area probably reflects a time-scale of change measured in tens of thousands of years.

On slow-spreading centres, such as the Mid-Atlantic Ridge, the picture is less clear. Seismic and gravity data indicate that the oceanic crust thins significantly near many of the transform faults, even

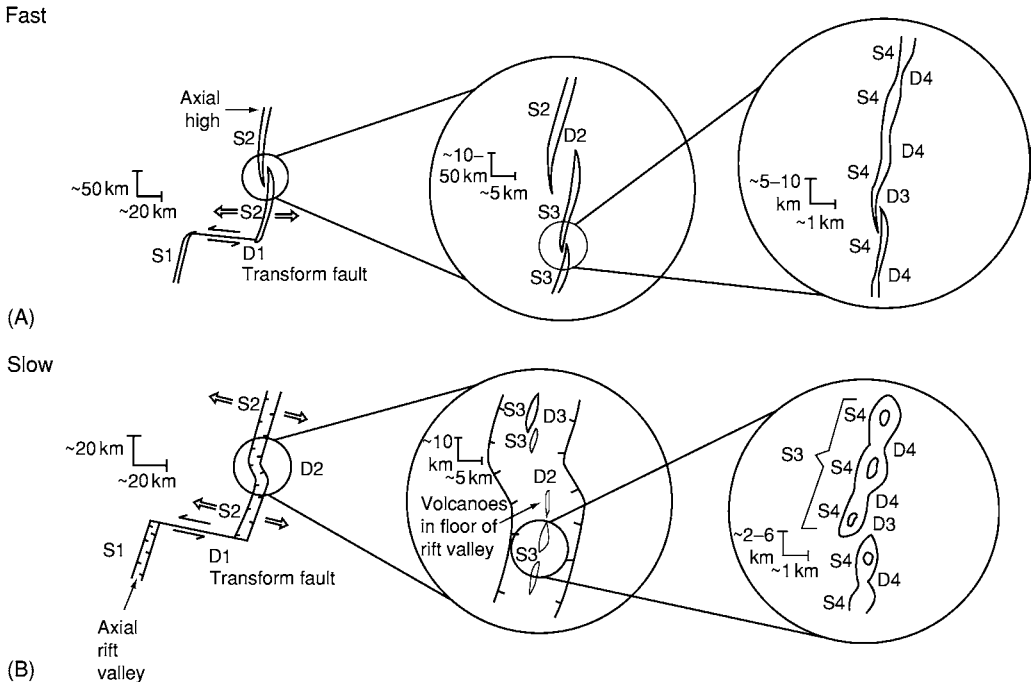


Figure 4 A possible hierarchy of ridge segmentation for (A) fast spreading and (B) slow spreading ridges. S1 S4 are ridge segments of orders 1 4, and D1 D4 are ridge axis discontinuities of orders 1 4. At both fast spreading and slow spreading centres, first order discontinuities are transform faults. Examples of second order discontinuities are overlapping spreading centres on fast spreading ridges and oblique shear zones on slow spreading ridges. Third order discontinuities are small overlapping spreading centres on fast spreading ridges. Fourth order discontinuities are slight bends or lateral offsets of the axis of less than 1 km on fast spreading ridges. This four tiered hierarchy of segmentation probably represents a continuum; it has been established, for example, that fourth order segments and discontinuities can grow to become third , second , and even first order features and vice versa at both slow spreading and fast spreading centres. (Reproduced from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986 994.)

those with small offsets. This is thought to be the result of highly focused mantle upwelling near the mid-segment regions, with very little along-axis flow of magma away from the upwelling region. Focused upwelling is inferred from 'bulls-eye'-shaped residual gravity anomalies and variations in crustal thickness that have been documented by seismic refraction and studies of microearthquakes. At slow-spreading centres, melt probably resides in small, isolated, and very short-lived pockets beneath the median valley floor (Figure 5) and beneath elongate axial volcanic ridges. An alternative view is that the observed along-strike variations in topography and crustal thickness can be accounted for by along-strike variations in mechanical thinning of the crust by faulting. There is no conflict between these models, so both focused upwelling and mechanical thinning may occur along each segment.

One might expect the same to hold at fast-spreading centres, i.e. crustal thinning adjacent to overlapping spreading centres. This does not appear to be the case at 9° N on the East Pacific Rise, where seismic data suggest a thickening of the crust towards the overlapping spreading centres and a widening of the axial magma chamber reflector. There is no indication of crustal thinning near the Clipperton transform fault either. And yet, as one approaches the 9° N overlapping spreading centres from the north, the axial depth plunges, the axial cross-sectional area decreases, the axial magma chamber reflector deepens, the average lava age increases, the MgO content of dredged basalts decreases, hydrothermal activity decreases dramatically, crustal magnetization increases significantly (suggesting eruption of more fractionated basalts in a region of decreased magma supply), crustal fracturing and inferred depth of fracturing increase (indicating a greater ratio of extensional strain to magma supply), and the throw of off-axis normal faults increases (suggesting thicker lithosphere and greater strain) (Figure 8A). How can these parameters all correlate so well, indicating a decrease in the magmatic budget and an increase in amagmatic extension, yet the seismic data suggest crustal thickening off-axis from and a wider magma lens near the overlapping spreading centres?

One possibility is that mantle upwelling and the axial magmatic budget are enhanced away from ridge-axis discontinuities even at fast-spreading

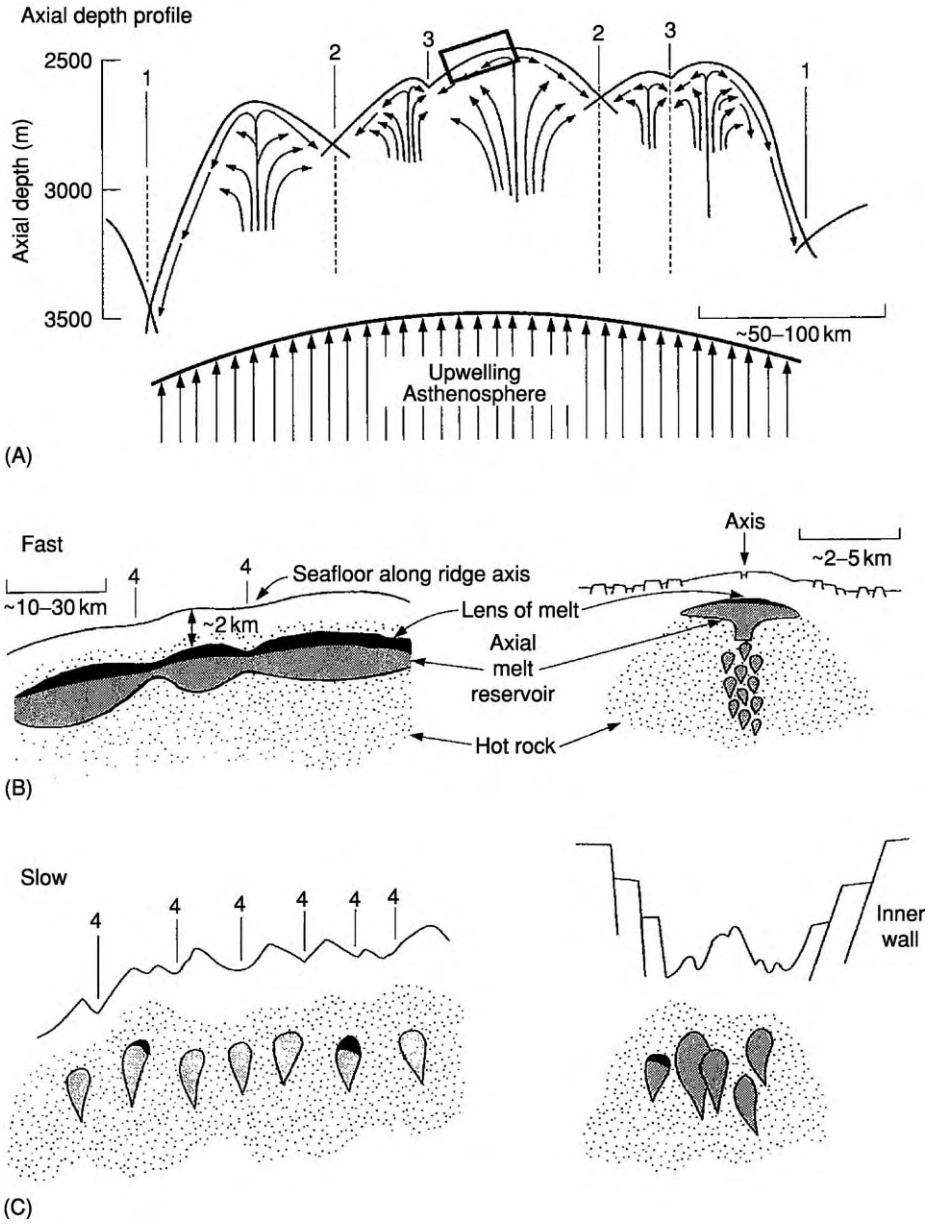


Figure 5 (A) How ridge segmentation may be related to mantle upwelling, and (B and C) the distribution of magma supply. In (A), the depth scale applies only to the axial depth profile; numbers denote discontinuities and segments of orders 1 3. Decompression partial melting in the upwelling asthenosphere occurs at depths of 30 60 km beneath the ridge. As the melt ascends through a more slowly rising solid residuum, it is partitioned at different levels to feed segments of orders 1 3. Mantle upwelling is hypothesized to be 'sheetlike' in the sense that melt is upwelling along the entire length of the ridge, but the supply of melt is thought to be enhanced beneath shallow parts of the ridge away from major discontinuities. The rectangle indicates the area enlarged to show fine scale segmentation for (B) a fast spreading example and (C) a slow spreading example. In (B) and (C) along strike cross sections showing the hypothesized partitioning of the magma supply relative to fourth order discontinuities (4s) and segments are shown on the left. Across strike cross sections for fast spreading and slow spreading ridges are shown on the right. (Reproduced from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986 994.)

centres, but subaxial flow of magma 'downhill' away from the injection region redistributes the magma episodically (Figure 5). This along-strike flow and redistribution of magma may be unique to spreading centres with an axial high, such as the East Pacific Rise or Reykjanes, where the axial region is sufficiently hot at shallow depths to facilitate subaxial flow. It is well-documented in Iceland and other volcanic areas analogous to mid-ocean ridges that magma can flow in subsurface chambers and dykes

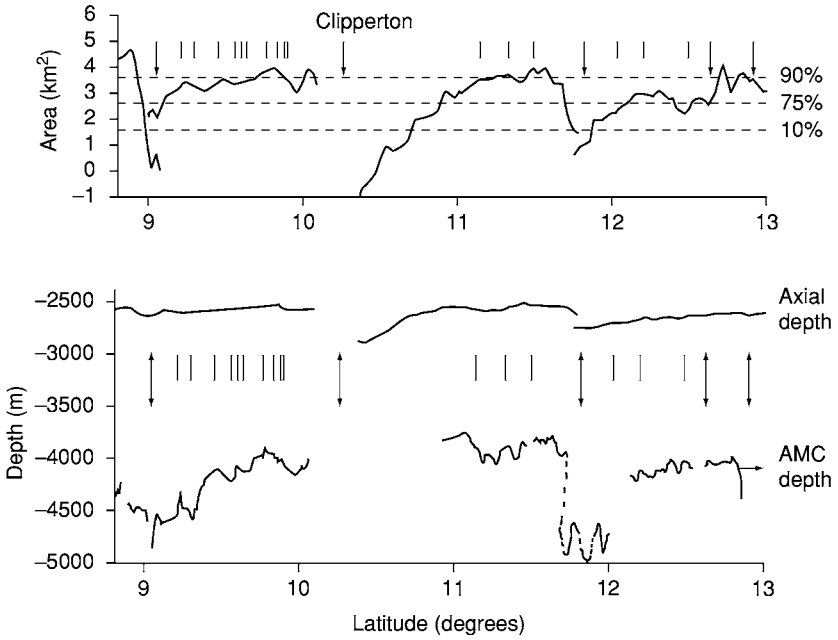


Figure 6 Profiles of the along axis cross sectional area, axial depth, and axial magma chamber (AMC) seismic reflector depth for the East Pacific Rise 9° 13° N. The locations of first and second order discontinuities are denoted by vertical arrows (first order discontinuities are named); each occurs at a local minimum of the ridge area profile and a local maximum of the ridge axis depth smaller discontinuities are denoted by vertical bars. There is an excellent correlation between ridge axis depth and cross sectional area; there is a good correlation between cross sectional area and the existence of an axial magma chamber, but detailed characteris tics of the axial magma chamber (depth, width) do not correlate. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

for many tens of kilometres away from the source region before erupting. In this way, thicker crust may occur away from the mid-segment injection points, proximal to discontinuities such as overlapping spreading centres.

Based on studies of the fast-spreading East Pacific Rise, a 'magma supply' model has been proposed, which explains the intriguing correlation between over a dozen structural, geochemical, and geophysical variables within a first-, second-, or third-order segment (Figure 9). It also addresses the initially puzzling observation that the crust is sometimes thinner in the mid-segment region, where upwelling is supposedly enhanced. Intuitively, one might expect the crust to be thickest over the region where upwelling is enhanced, as is observed on the Mid-Atlantic Ridge. However, along-axis redistribution of melt may be the controlling factor on fast-spreading ridges, where the subaxial melt region may be well-connected for tens of kilometres. In this model, temporal variations in along-axis melt connectivity may result in thicker crust near the mid-segment when connectivity is low (most often at slow-spreading ridges) and thicker crust closer to the segment ends when connectivity is high (most often, but not always, at fast-spreading ridges).

The basic concepts of this magma-supply model also apply to slow-spreading ridges that are characterized by an axial rift valley. Mantle melting is enhanced beneath the mid-segment regions. However, the axial region is colder (averaged over time), and along-strike redistribution of melt is impeded. Thus, the crust tends to be thickest near the mid-segment regions and thinnest near ridge-axis discontinuities (Figures 8B and 9).

It is possible that the segmentation of mid-ocean ridges and the observations that correlate with segmentation (e.g. axial depth, geochemistry of lavas, lava morphology, etc.; Figure 8) are not related to the supply of magma to the ridge. This is still an area of active research and debate. For example, it has been suggested that the supply of melt is uniform along fast-spreading ridges and that along-strike variations are caused by differences in hydrothermal heat loss. If heat loss were enhanced near segment ends, this could cause many of the along-strike variations noted in Figure 8. So far, however, there is no indication that hydrothermal heat loss is greater near segment ends. On the contrary, hydrothermal heat loss is least evident at segment ends and is often enhanced near the shallow mid-sections of first-, second-, and third-order segments.

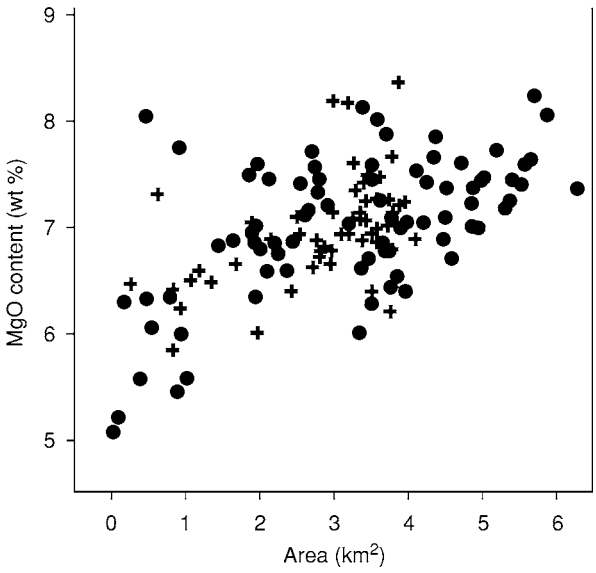


Figure 7 Cross sectional area of the East Pacific Rise plotted against the MgO content of basalt glass (crosses from 5 14° N; circles from 13 23° S). There is a tendency for high MgO contents (interpreted as higher eruption temperatures and perhaps a higher magmatic budget) to correlate with larger cross sectional areas. Smaller cross sectional areas are correlated with lower levels of MgO and a greater scatter in MgO content, suggesting magma chambers that are transient and changing. Thus shallow inflated areas of the ridge tend to erupt hotter lavas. Updated from Scheirer and Macdonald (1993) and references therein. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid ocean ridge tectonics, volcanism and geomorphology*, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

Fine-Scale Variations in Ridge Morphology Within the Axial Neovolcanic Zone

The axial neovolcanic zone occurs on or near the axis of the axial high of fast-spreading centres and within the floor of the rift valley of slow-spreading centres (**Figures 2A**). Studies of the widths of the polarity transitions of magnetic anomalies, including *in situ* measurements from ALVIN, document that approximately 90% of the volcanism that creates the extrusive layer of oceanic crust occurs in a region 1–5 km wide at most spreading centres. Direct qualitative estimates of lava age at spreading centres using submersibles and remotely operated vehicles tend to confirm this, as do recent high-resolution seismic measurements that show that layer 2A (interpreted to be the volcanic layer) achieves its full thickness within 1–3 km of the rise axis (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges). However, there are significant exceptions, including small-volume off-axis volcanic constructions and voluminous off-axis floods of basaltic sheet flows.

The axial high on fast-spreading and intermediate-spreading centres is usually bisected by an axial summit trough approximately 10–200 m deep, which is found along approximately 60–70% of the axis. Along the axial high of fast-spreading ridges, side-scan sonar records show that there is an excellent correlation between the presence of an axial summit trough and an axial magma chamber reflector as seen on multichannel seismic records (over more than 90% of the ridge length). Neither axial summit troughs nor axial magma chambers occur where the ridge has a very small cross-sectional area.

In rare cases, an axial summit trough is not observed where the cross-sectional area is large. In these locations, volcanic activity is presently occurring or has occurred within the last decade. For example, on the East Pacific Rise, near 9°45′–9°52′ N, a volcanic eruption documented from the submersible ALVIN was associated with a single major dyke intrusion, similar to the 1993 eruption on the Juan de Fuca Ridge. Side-scan sonar records showed that an axial trough was absent from 9°52′ N to 10°02′ N and, in subsequent dives, it was found that dyke intrusion had propagated into this area producing very recent lava flows and hydrothermal activity complete with bacterial 'snowstorms'. A similar situation has been thoroughly documented at 17°25′–17°30′ S on the East Pacific Rise, where the axial cross-sectional area is large but the axial summit trough is partly filled. Perhaps the axial summit trough has been flooded with lava so recently that magma withdrawal and summit collapse are still occurring. Thus, the presence of an axial summit trough along the axial high of a fast-spreading ridge is a good indicator of the presence of a subaxial lens of partial melt (axial magma chamber); where an axial summit trough is not present but the cross-sectional area is large, this is a good indicator of very recent or current volcanic eruptions; where an axial summit trough is not present and the cross-sectional area is small, this is a good indicator of the absence of a magma lens (axial magma chamber).

In contrast to the along-axis continuity of the axial neovolcanic zone of fast-spreading ridges, the neovolcanic zone of slower-spreading ridges is considerably less continuous and there is a great deal of variation from segment to segment. Volcanic constructions, called axial volcanic ridges, are most common along the shallow mid-segment regions of the axial rift valley. Near the ends of segments, where the rift valley deepens, widens, and is truncated by transform faults or oblique shear zones, the gaps between axial volcanic ridges become longer. The gaps between axial volcanic ridges are regions of older crust, characterized by faulting and a lack of recent volcanism.

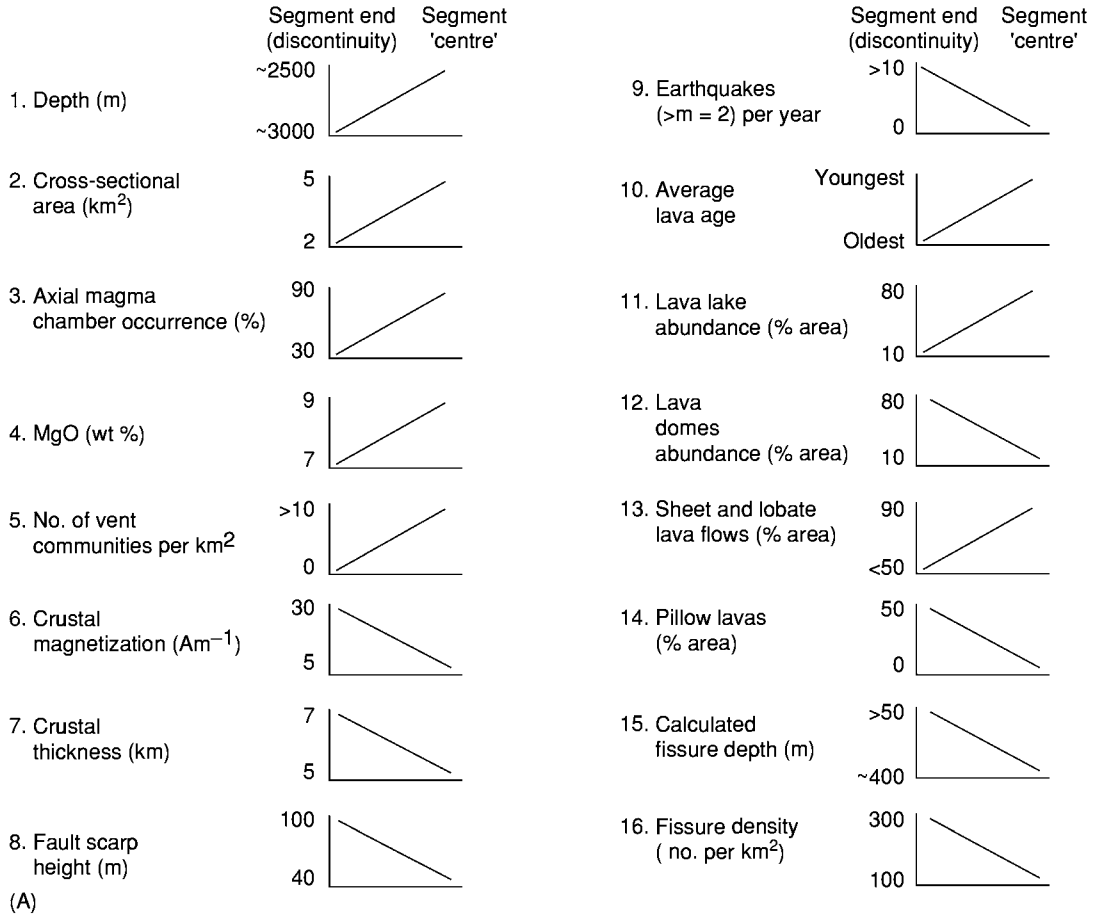


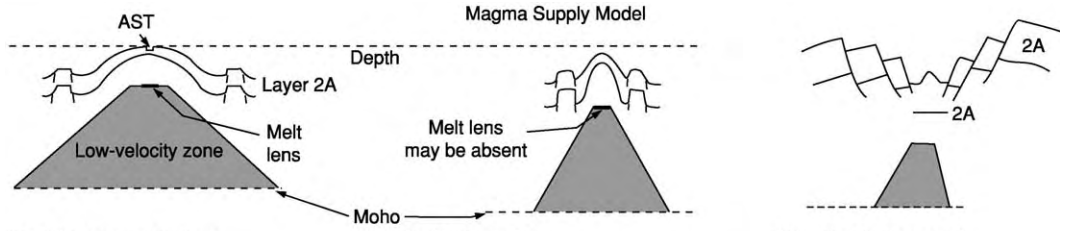
Figure 8 Summary of along axis variations in spreading centre properties from segment end (discontinuity of order 1, 2, or 3) to segment mid section areas for (A) fast spreading ridges with axial highs and (B) slow spreading ridges with axial rift valleys. A large number of parameters correlate well with location within a given segment, indicating that segments are distinct independent units of crustal accretion and deformation. These variations may reflect a fundamental segmentation of the supply of melt beneath the ridge.

Continued

These gaps may correspond to fine-scale (third- and fourth-order) discontinuities of the ridge.

There is another important difference between volcanism on fast-spreading and slow-spreading ridges. Axial volcanic ridges represent a thickening of the volcanic layer atop a lithosphere that may be 5–10 km thick, even on the axis. In contrast, the volcanic layer is usually thinnest along the axis of the fast-spreading East Pacific Rise (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges). Thus, the axial high of fast-spreading ridges is not a thickened accumulation of lava, while the discontinuous axial volcanic ridges of slow-spreading ridges are.

On both slow-spreading and fast-spreading ridges, pillow and lobate lavas are the most common lava morphologies. Based on laboratory studies and observations of terrestrial basaltic eruptions, this indicates that the lava effusion rates are slow to moderate on most mid-ocean ridges. High volcanic effusion rates, indicated by fossil lava lakes and extensive outcrops



(A) High magma budget	(B) Moderate to low magma budget	(C) Very low, sporadic magma supply
Shallow water depth	Deeper water depth	Deep rift valley floor
Broad cross-sectional area	Small cross-sectional area	Wide valley floor
Wide low-velocity zone High MgO content High temperature Low density	Narrow (except where 2 LVZ coalesce at an OSC)* Lower MgO content Lower temperature Higher density	Narrow low-velocity zone
Melt lens ubiquitous, depth and width variable*	Melt lens less common, highly variable width, depth where present, disrupted at RADs*	Melt lens very rare
Thinner crust (melt flows 'downhill' away from mantle upwelling zone)*	Thicker crust, but highly variable*	Highly variable crustal thickness
Axial summit trough present unless in eruptive phase*	Axial summit trough rare	
Hydrothermal venting abundant*	Hydrothermal venting rare*	Hydrothermal vents rare, fault controlled
Less crustal fissuring*	More crustal fissuring*	Extensive fissuring
Younger average lava age*	Older average lava age*	Highly variable but generally older lava age, mostly pillows
Lower crustal magnetization*	Higher crustal magnetization*	
Smaller throw on flanking fault scarps	Larger throw on flanking fault scarps	Large fault scarps, especially at inside corner highs
Layer 2A thinnest along axis	Layer 2A thinnest along axis	Layer 2A thick beneath axial volcanic ridges

Figure 9 Magma supply model for mid ocean ridges. (A) A segment with a robust magmatic budget, generally a fast spreading ridge away from discontinuities or a hotspot dominated ridge with an axial high. (B) A segment with a moderate magma budget, generally a fast spreading ridge near a discontinuity or a non rifted intermediate rate ridge. (C) A ridge with a sporadic and diminished magma supply, generally a rifted intermediate to slow spreading centre (for along strike variations at a slow ridge, see **Figure 8B**). AST, axial summit trough; LVZ, low velocity zone; OSC, overlapping spreading centres; RAD, ridge axis discontinuity. See references in Buck *et al.* (1998) (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

(Less than 1% of the ridge has been studied in sufficient detail to create this summary.) (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

of sheet-flow lavas, are very rare on slow-spreading ridges. High-effusion-rate eruptions are more common on fast-spreading ridges and are more likely to occur along the shallow inflated mid-segment regions of the rise, in keeping with the magma-supply model for ridges discussed earlier (Figure 9A).

Very little is known about eruption frequency. It has been estimated based on some indirect observations that, at any given place on a fast-spreading ridge, eruptions occur approximately every 50–100 years, and that on slow-spreading ridges eruptions occur approximately every 5000–10 000 years. If this is true, then the eruption frequency varies inversely with the square of the spreading rate. On intermediate- to fast-spreading centres, if one assumes a typical dyke width of around 50 cm and a spreading rate of 5–10 cm year ^{1}, then an eruption could occur approximately every 5–10 years. This estimate is in reasonable agreement with the occurrence of megaplumes and eruptions on the well-monitored Juan de Fuca Ridge. However, observations of sheeted-dyke sequences in Iceland and of ophiolites indicate that only a small percentage of the dykes reach the surface to produce eruptions.

On fast-spreading centres, the axial summit trough is so narrow (30–1000 m) and well-defined in most places that tiny offsets and discontinuities of the rise axis can be detected (Table 1 and Figure 2). This finest scale of segmentation (fourth-order segments and discontinuities) probably corresponds to individual fissure eruption events similar to the Krafla eruptions in Iceland and the Kilauea east rift zone eruptions in Hawaii. Given a magma chamber depth of 1–2 km, an average dyke ascent rate of approximately 0.1 km h ^{1}, and an average lengthening rate of approximately 1 km h ^{1}, typical diking events would give rise to segments 10–20 km long. This agrees with observations of fourth-order segmentation and the scale of recent diking events on the Juan de Fuca Ridge and in other volcanic rift zones. The duration of such segments is thought to be very short, of the order of 100–1000 years (too brief in any case to leave even the smallest detectable trace off-axis; Table 1). Yet, even at this very fine scale, excellent correlations can be seen between average lava age, density of fissuring, the average widths of fissures, and the abundance of hydrothermal vents within individual segments. In fact, there is even an excellent correlation between ridge cross-sectional area and the abundance of benthic hydrothermal communities (Figure 8).

A curious observation on the East Pacific Rise is that the widest fissures occur in the youngest lava fields. If fissures widen over time with increasing extension, one would expect the opposite: the widest fissures should be in the oldest areas. The widest fissures are approximately 5 m wide. Using simple fracture mechanics, it can be concluded that these fissures probably extend all the way through layer 2A and into the sheeted-dyke sequence. These have been interpreted as eruptive fissures, and this is where high-temperature vents (more than 300 °C) are concentrated. In contrast to the magma-rich dyke-controlled hydrothermal systems that are common on fast-spreading centres, magma-starved hydrothermal systems on slow-spreading ridges tend to be controlled more by the penetration of seawater along faults near the ridge axis.

Faulting

Extension at mid-ocean ridges causes fissuring and normal faulting. The lithosphere is sufficiently thick and strong on slow-spreading centres to support shear failure on the axis, so normal faulting along dipping fault planes can occur on or very close to the axis. These faults produce grabens that are 1–3 km deep. In contrast, normal faulting is not common on fast-spreading centres within 2 km of the axis, probably because the lithosphere is too thin and weak to support normal faulting. Instead, the new thin crust fails by simple tensional cracking.

Fault strikes tend to be perpendicular to the direction of least compressive stress; thus, they also tend to be perpendicular to the spreading direction. While there is some 'noise' in the fault trends, most of this noise can be accounted for by perturbations in the direction of least compressive stress due to shearing in the vicinity of active or fossil ridge-axis discontinuities. Once this is accounted for, fault trends faithfully record changes in the direction of opening to within $\pm 3^{\circ}$ and can be used to study plate-motion changes on a finer scale than that provided by the study of seafloor magnetic anomalies. Studies of the cumulative throw of normal faults, seismicity, and fault spacing suggest that most faulting occurs within 20–40 km of the axis irrespective of spreading rate.

The occurrence of inward- and outward-dipping faults depends on the spreading rate. Most faults (*ca*. 80%) dip towards the axis on slow-spreading centres, but there is a monotonic increase in the frequency of outward-dipping faults with spreading rate (Figure 10). Inward- and outward-facing faults are approximately equally abundant at very fast spreading rates. This can be explained by the smaller mean normal stress across a fault plane that dips towards the axis, cutting through thin lithosphere, than across a fault plane that dips away from the axis, cutting through a much thicker section of lithosphere. Given reasonable thermal models, the difference in the

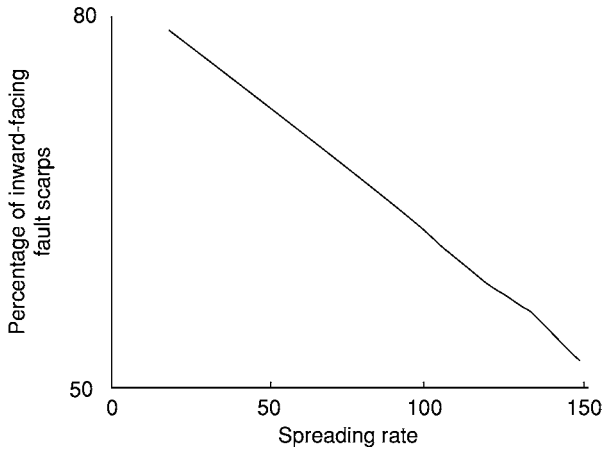


Figure 10 The effect of spreading rate on the percentage of fault scarps that are inward facing (facing towards the spreading axis rather than away from the spreading axis). A significant increase in the percentage of inward facing scarps occurs at slower spreading rates (mm yr). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid ocean ridge tectonics, volcanism and geomorphology, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

thickness of the lithosphere cut by planes dipping towards and away from the axis (and the mean normal stresses across those planes) decreases significantly with spreading rate, making outward-dipping faults more likely at fast spreading rates.

At all spreading rates, important along-strike variations in faulting occur within major (first-order and second-order) spreading segments. Fault throws (inferred from scarp heights) decrease in the mid-segment regions away from discontinuities (Figures 8 and 11). This may be caused by a combination of thicker crust, thinner lithosphere, greater magma supply, and less amagmatic extension away from ridge-axis discontinuities in the mid-segment region (Figure 11). Another possible explanation for along-strike variations in fault throw is along-strike variation in the degree of coupling between the mantle and crust. A ductile lower crust will tend to decouple the upper crust from extensional stresses in the mantle, and the existence of a ductile lower crust will depend on spreading rate, the supply of magma to the ridge, and proximity to major discontinuities.

Estimates of crustal strain due to normal faulting vary from 10–20% on the slow-spreading Mid-Atlantic Ridge to *ca*. 3–5% on the fast-spreading East Pacific Rise. This difference may be explained as follows. The rate of magma supply to slow-spreading ridges is relatively low compared with the rate of crustal extension and faulting, while extension and magma-supply rates are in closer balance on fast-spreading ridges. The resulting seismicity is different too. In contrast to faulting on slow-spreading ridges, where teleseismically detected earthquakes are common, faulting on fast-spreading ridges rarely produces earthquakes with magnitudes of more than 5. Nearly all of these events are associated with ridge-axis discontinuities. The level of seismicity measured at fast-spreading ridges accounts for only a very small percentage of the observed strain due to faulting, whereas fault strain at slow-spreading ridges is comparable to the observed seismic moment release. It has been suggested that faults in fast-spreading environments accumulate slip largely by stable sliding (aseismically) owing to the warm temperatures and associated thin brittle layer. At slower spreading rates, faults will extend beyond a frictional stability transition into a field where fault slip occurs unstably (seismically) because of a thicker brittle layer.

Disruption of oceanic crust by faulting may be particularly extreme on slow-spreading ridges near transform faults (Figure 12). Unusually shallow topography occurs on the active transform slip side of ridge–transform intersections; this is called the inside corner high. These highs are not volcanoes. Instead they are caused by normal faults that cut deeply into and perhaps all the way through the oceanic crust. It is thought that crustal extension may occur for 1–2 Ma on detachment faults with little magmatic activity. This results in extraordinary extension of the crust and exposure of large sections of the deep crust and upper mantle on the seafloor. Corrugated slip surfaces indicating the direction of fault slip are also evident and are called 'megamullions' by some investigators.

Disruption of oceanic crust by faulting may also reach extremes at the slowest known spreading rates of 0.5–1.0 cm year ^{1}, for example on the Gakkel Ridge in the Arctic Ocean and along parts of the Southwest Indian Ocean Ridge. There is evidence that crustal extension due to faulting exceeds 20% and that there are prolonged periods with no basaltic volcanism along the rift valley. Instead, upper-mantle peridotites upwell to fill the gap between the separating plates during the intervals between infrequent, but sometimes very large, basaltic eruptions.

At distances of several tens of kilometres from the axis, topography generated near the spreading centre is preserved on the seafloor with little subsequent change, except for the gradual accumulation of pelagic sediments at rates of approximately 0.5–20 cm per thousand years, until it is subducted. The preserved topographic highs and lows are called abyssal hills. At slow-spreading centres characterized by an axial rift valley, back-tilted fault blocks and half-grabens may be the dominant origins of abyssal hills (Figure 13),

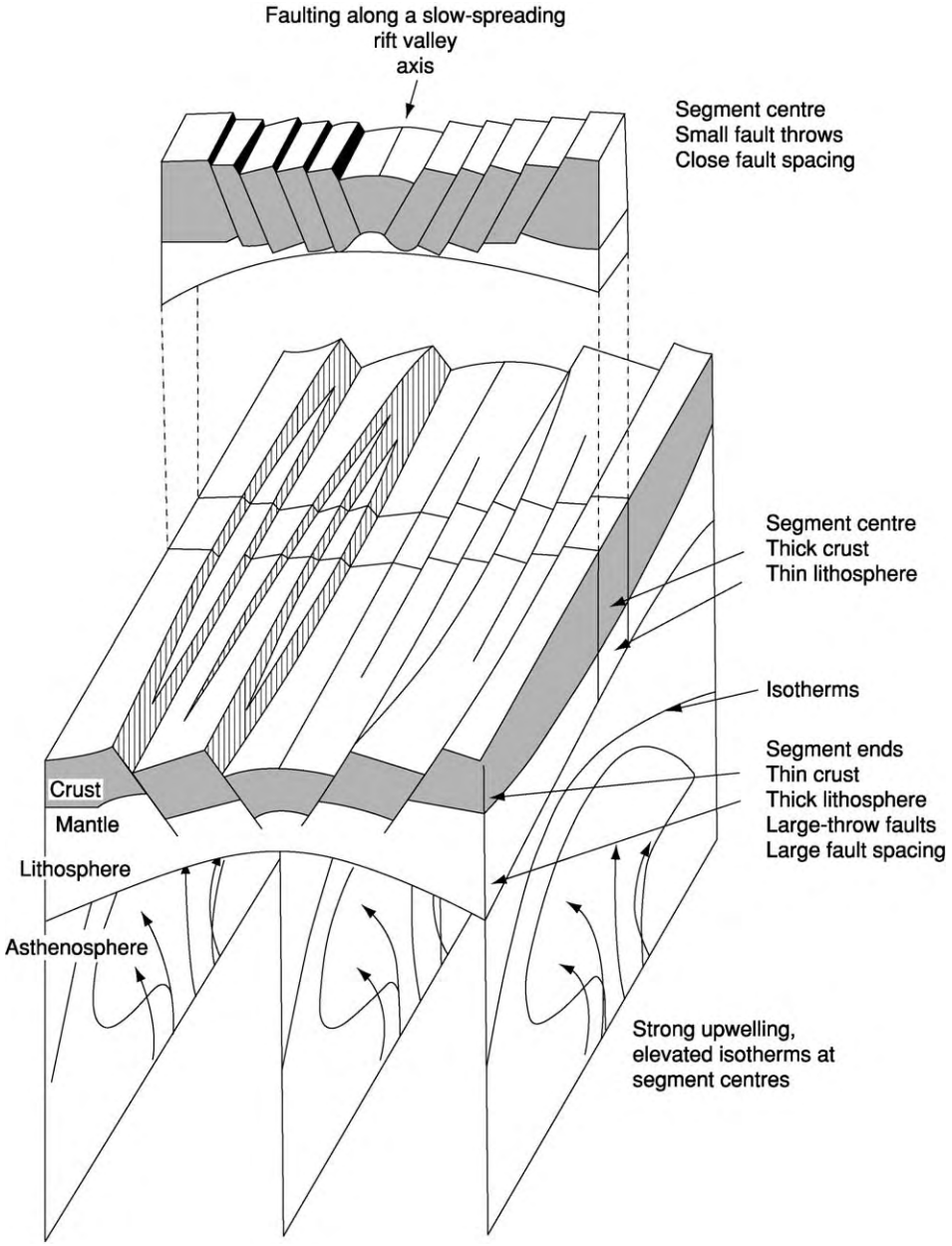


Figure 11 A geological interpretation of along axis variations in scarp height and the occurrence of more closely spaced scarps near mid segments on a slow spreading ridge. The cross section through the segment centre (top) shows more closely spaced smaller throw faults than at the segment ends (bottom). Focused mantle upwelling near the segment centre causes this region to be hotter: the lithosphere will be thinner, while increased melt supply will create a thicker crust. In contrast to observations at fast spreading centres, there may be very little melt redistribution along strike. Near the segment ends, the lithosphere will be thicker and magma supply will be less, creating thinner crust. Along axis variations in scarp height and spacing reflect these along axis variations in lithospheric thickness. Amagmatic extension across the larger faults near segment ends may also thin the crust, especially at inside corner highs. (Reproduced from Shaw PR (1992) Ridge segmentation, faulting and crustal thickness in the Atlantic Ocean. *Nature* 358: 490 493.)

although there is continued controversy over the roles of high-versus low-angle faults and listric faulting versus planar faulting and the possible role of punctuated episodes of volcanism versus amagmatic extension. At intermediate-rate spreading centres, abyssal-hill structure may vary with the local magmatic budget. Where the budget is starved and the axis is characterized by a rift valley, abyssal hills are

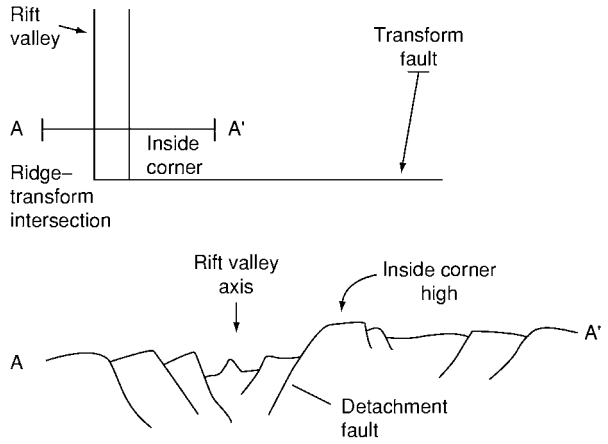


Figure 12 Inside corner high at a slow spreading ridge transform intersection. Extension is concentrated along a detach ment fault for up to 1 2 Ma, exposing deep sections of oceanic crust and mantle. The oceanic crust is thinned by this extreme extension; crustal accretion and magmatic activity may also be diminished. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid ocean ridge tectonics, volcanism and geomorphology*, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

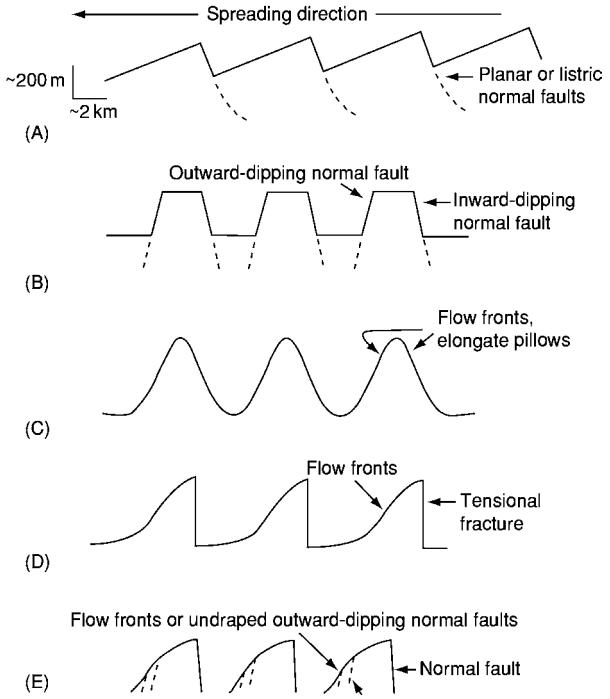


Figure 13 Five models for the development of abyssal hills on the flanks of mid ocean ridges. (A) Back tilted fault blocks (episodic inward dipping normal faulting off axis). (B) Horst and graben (epi sodic inward and outward dipping faulting off axis). (C) Whole volcanoes (episodic volcanism on axis). (D) Split volcanoes (epi sodic volcanism and splitt ng on axis). (E) Horsts bounded by inward dipping normal faults and outward dipping volcanic growth faults (episodic faulting off axis and episodic volcanism on axis or near axis). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid ocean ridge tectonics, volcanism and geomorphology*, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

generally back-tilted fault blocks. Where the magmatic budget is robust and an axial high is present, the axial lithosphere is episodically thick enough to support a volcanic construction, which may then be rafted away intact or split in two by the spreading axis, resulting in whole-volcano and split-volcano abyssal hills, respectively.

Based on observations made from the submersible ALVIN on the flanks of the East Pacific Rise, it would appear that the outward-facing slopes of the hills are neither simple outward-dipping normal faults, as would be predicted by the horst-and-graben model, nor entirely of volcanic construction, as would be predicted by the split-volcano model. Instead, the outward-facing slopes are 'volcanic growth faults' (Figure 14). Outward-facing scarps produced by episodes of normal faulting are buried near the axis by syntectonic lava flows originating along the axial high. Repeated episodes of dip-slip faulting and volcanic burial result in structures resembling growth faults, except that the faults are episodically buried by lava flows rather than being continuously buried by sediment deposition. In contrast, the inward-dipping faults act as tectonic dams to lava flows. Thus, the abyssal hills are horsts and the intervening troughs are grabens, with the important modification to the horst-and-graben model that the outward-facing slopes are created by volcanic growth faulting rather than traditional normal faulting. Thus, on fast-spreading centres, abyssal hills are asymmetric, being bounded by steeply dipping normal faults

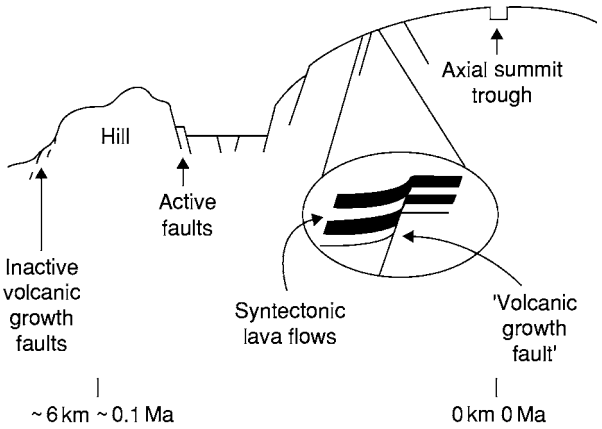


Figure 14 Cross sectional depiction of the development of volcanic growth faults. Volcanic growth faults are common on fast spreading centres and explain some of the differences between inward and outward facing scarps as well as the mor phology and origin of most abyssal hills on fast spreading cen tres. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid ocean ridge tectonics, volcanism and geomorphology*, pp. 1798 1813, Copy right (2001), with permission from Elsevier.)

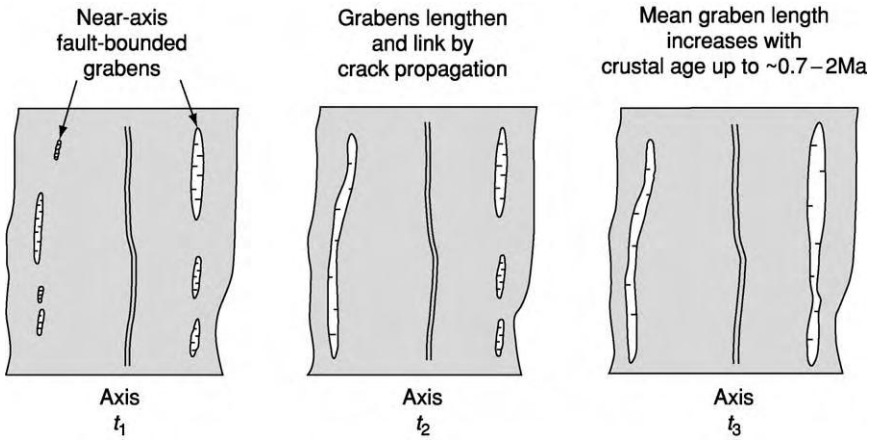


Figure 15 Proposed time sequence of along strike propagation and linkage of near axis faults and grabens, which define the edges of abyssal hills; time averaged propagation rates are approximately 20 60 km Ma ^{1}. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid ocean ridge tectonics, volcanism and geomorphology*, pp. 1798 1813, Copyright (2001), with permission from Elsevier.)

facing the spreading axis and volcanic growth faults on the opposing side. These faults lengthen at a rate of 20–60 Km/Ma reaching total lengths of 10–70 km to form abyssal hills and intervening grabens (Figure 15).

See Also

Analytical Methods: Gravity. **Lava**. **Magnetostratigraphy**. **Plate Tectonics**. **Seamounts**. **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges; Seismic Structure At Mid-Ocean Ridges.

Further Reading

Buck WR, Delaney PT, Karson JA, and Lagabrielle Y (1998) *Faulting and Magmatism at Mid Ocean Ridges*. Geophysical Monograph 106. Washington, DC: American Geophysical Union.

Fox PJ and Gallo DG (1986) The geology of North Atlantic transform plate boundaries and their aseismic extensions. In: Vogt PR and Tucholke BE (eds.) *The Western North Atlantic Region: The Geology of North America*, pp. 157 172. Boulder: Geological Society of America.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (1995) *Seafloor hydrothermal systems: physical, chemical, biological and geological interactions*. AGU Geophysical Monograph 91. Washington, DC: American Geophysical Union.

Langmuir CH, Bender JF, and Batiza R (1986) Petrological and tectonic segmentation of the East Pacific Rise, 5°30′ N 14°30′ N. *Nature* 322: 422 429.

Macdonald KC (1982) Mid ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annual Review of Earth and Planetary Sciences* 10: 155 190.

Macdonald KC and Fox PJ (1990) The mid ocean ridge. *Scientific American* 262: 72 79.

Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986 994.

Menard H (1986) *Ocean of Truth*. Princeton: Princeton University Press.

Phipps Morgan J, Blackman DK, and Sinton J (1992) *Mantle Flow and Melt Generation at Mid Ocean Ridges*. Geophysical Monograph 71. Washington, DC: American Geophysical Union.

Shaw PR (1992) Ridge segmentation, faulting and crustal thickness in the Atlantic Ocean. *Nature* 358: 490 493.

Hydrothermal Vents At Mid-Ocean Ridges

R M Haymon, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

On the Galapagos Rift in 1977, ~2600 m beneath the sunlit surface of the sea, scientists exploring the lightless, frigid seafloor were astonished to discover springs of warm water teeming with life (Figure 1A). It was quickly realized these remarkable ecosystems in the deep sea are supported by microbes capable of metabolizing chemicals dissolved in hydrothermal spring waters. In 1979, submersible divers exploring the crest of the East Pacific Rise came on superheated ($380 \pm 30°C$) hot springs, where plumes of scalding fluid, blackened by minuscule mineral particles, billowed into the ocean through tall mineral conduits (Figure 1B). The minerals formed at these 'black smoker' hydrothermal vents proved to be rich in copper, iron, zinc, and other useful metals.

Hundreds of hydrothermal vents have now been located on the crest and flanks of the mid-ocean ridge, and more are found every year. The mid-ocean ridge is a globe-encircling rift where plates of ocean lithosphere are repeatedly ripped apart (*see* **Tectonics:** Mid-Ocean Ridges). New, hot seafloor freezes in the cracks between the plates and is cooled by hydrothermal circulation that entrains enormous volumes of seawater. The great magnitude of hydrothermal fluid flow through the mid-ocean ridge influences the chemistry, biology, and physical oceanography of the world ocean; the thermal structure, physical properties, and chemical composition of ocean crust; and the nature and diversity of subseafloor microbial environments. Many now believe that submarine hot springs may have been the biochemical crucibles for the origin of life in the solar system.

For thousands of years before mid-ocean ridge hot springs were discovered in the oceans, people mined copper from mineral deposits that were originally formed on oceanic spreading ridges. These fossil deposits are embedded in old fragments of seafloor, called 'ophiolites', that have been uplifted and emplaced onto land by fault movements. The copper-rich mineral deposits in the Troodos ophiolite of Cyprus are well-known examples of fossil ocean ridge deposits that have been mined for at least 2500 years; in fact, the word 'copper' is derived from the Latin word *cyprium*, which means 'from Cyprus'.

The mineral deposits accumulating today at hot springs along the mid-ocean ridge are habitats for a variety of remarkable organisms, ranging in size from tiny microbes to 2-m-long tubeworms. The properties of the mineral deposits are inextricably linked to the organisms that inhabit them. The mineral deposits contain important clues about the physical–chemical environments in which some of these organisms live, and also preserve fossils of some organisms, creating

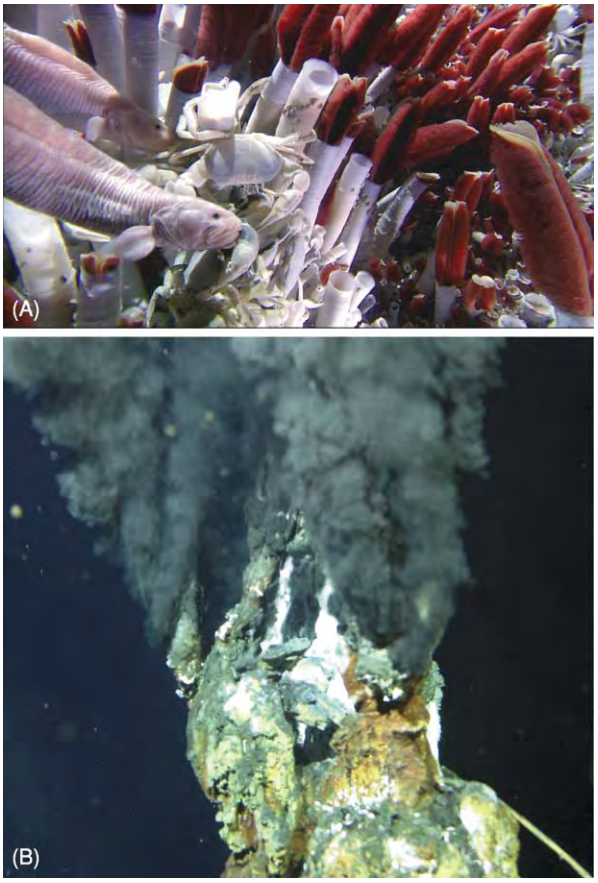


Figure 1 (A) A dense thicket of *Riftia* vestimentiferan tube worms on the crest of the East Pacific Rise near lat. 9° 50′ N; brachyuran crabs and pink zooarcid fish are visible in the fore ground. Chemosynthetic microbes, nourished by dissolved minerals in hydrothermal fluids, support many flourishing com munities such as this along the mid ocean ridge crest. Photo credit: Richard Lutz, Rutgers University, Stephen Low Produc tions, and Woods Hole Oceanographic Institution. (B) Vigorous plumes of mineral blackened fluid issue from metal enriched mineral chimneys on the northern East Pacific Rise crest. Photos were taken from the Alvin submersible (operated by the National Deep Submergence Facility at Woods Hole Oceano graphic Institution), during dives funded by the National Science Foundation.

a geological record of their existence. Hydrothermal vents thus help to balance the chemical composition of the oceans, and provide; sources of energy for deep-sea ecosystems; a renewable source of metals; and a depositional record of the physical, chemical, biological, and geological processes at modern and ancient submarine vents.

Where Deep-Sea Hydrothermal Vents and Deposits Form: Geological Controls

Less than 2% of the total area of the mid-ocean ridge crest has been studied at a resolution sufficient to reveal the spatial distribution of individual seafloor hydrothermal vents, mineral deposits, and other significant small-scale geological features. Nevertheless, because study areas have been carefully selected and strategically surveyed, and because methods for remotely mapping hydrothermal plumes have been developed, much has been learned about where vents and deposits form, and about the geological controls on their distribution. The basic requirements for hydrothermal systems include heat, to drive fluid circulation, and high-permeability pathways, to facilitate fluid flow through the seafloor. These requirements frequently are met at sites of seafloor volcanism and faulting on or near plate boundaries, or at 'mid-plate' seafloor volcanoes (for example, on Loihi Seamount near Hawaii). Along the mid-ocean ridge, vents and deposits are forming at sites where ascending magma intrusions introduce heat into the permeable shallow crust, and at sites where deep cracks provide permeability and fluid access to heat sources at depth.

Fast-Spreading Ridges

Near- and on-bottom studies along the fast-spreading East Pacific Rise suggest that most hydrothermal mineral deposits form at high-temperature vents ($\sim$100–400°C) concentrated along the summit of the ridge crest within a narrow 'axial zone' less than 500 m wide. Only a few active sites of mineral deposition have been located outside this zone. However, heat flow data indicate that $\sim$70% of the total hydrothermal heat loss from mid-ocean ridges occurs on ridge flanks, from lithosphere >1 million years old. Clearly, more exploration of the vast 'off-axis' region beyond the axial zone is needed to identify the nature and distribution of hydrothermal vents on ridge flanks.

Within the axial zone on fast-spreading ridges, the overall spatial distribution of hydrothermal vents and mineral deposits traces the segmented configuration of cracks and magma sources along the ridge axis. Active vents and mineral deposits are concentrated along the floors and walls of axial troughs created by volcanic collapse and/or faulting along the summit of the ridge crest. A majority of the deposits are located along fissures that have opened above magmatic dyke intrusions, and along collapsed lava ponds formed above these fissures by pooling and drainage of erupted lava. Where fault-bounded troughs have formed along the summit of the ridge crest, mineral deposition is focused along the bounding faults and also along fissures and collapsed lava ponds in the trough floor. Hydrothermal vents appear to be most abundant along magmatically inflated segments of fast-spreading ridges; however, the mineral deposits precipitated on the seafloor on magmatically active segments are often buried beneath frequent eruptions of new lava flows. The greatest number of deposits, therefore, can be observed on inflated ridge segments that are surfaced by somewhat older flows, i.e., along segments where much heat is available to power hydrothermal vents and mineral deposits have had time to develop but have not yet been buried by renewed eruptions.

Intermediate- and Slow-Spreading Ridges

Most hydrothermal deposits that have been found on intermediate- and slow-spreading ridge crests are focused along faults, fissures, and volcanic structures within large rift valleys that are several kilometres wide. The fault scarps along the margins of rift valleys are common sites for hydrothermal venting and mineral deposition. Some hydrothermal systems along rift valley bounding faults are known to have been episodically active over periods of thousands of years, and have produced large mineral deposits, several hundreds of metres in length and tens of metres thick. Fault intersections are thought to be particularly favorable sites for hydrothermal mineral deposition because they are zones of high permeability that can focus fluid flow. Vents and mineral deposits on rift valley floors also are observed above dike intrusions, along eruptive fissures and volcanic collapse troughs, or on top of volcanic mounds, cones, and other constructions. On the intermediate-rate Juan de Fuca Ridge, hydrothermal vents are located above seismically detectable magma lenses. Thus, both magmatic and tectonic controls on hydrothermal processes are observed at slow and intermediate spreading rates. In general, however, faults appear to play a greater role in controlling the distribution of hydrothermal vents and mineral deposits than they do at fast-spreading ridges, where magmatic fissures are clearly a dominant geological control on where vents and deposits are forming.

Spectacular fault-controlled hydrothermal mineral deposits have formed on older seafloor (approximately 1 million years old) on the west flank of the slow-spreading Mid-Atlantic Ridge, at the intersection of the Mid-Atlantic Ridge rift valley and the Atlantic Transform Fault near 30° N. This site, known as 'Lost City', is discharging warm hydrothermal fluids (<100°C) through more than 30 calcium carbonate mineral structures, the tallest of which is 60 m in height!

Structure, Morphology, and Size of Deposits

A typical hydrothermal mineral deposit on an unsedimented mid-ocean ridge accumulates directly on top of the volcanic flows covering the ridge crest. On sedimented ridges, minerals are deposited within and on top of the sediments. Beneath the seafloor are networks of feeder cracks through which fluids travel to the seafloor. Precipitation of hydrothermal minerals in these cracks and in the surrounding rocks or sediments creates a subseafloor zone of mineralization called a 'stockwork'. In hydrothermal systems in which fluid flow is weak or unfocused, or in which the fluids mix extensively with seawater beneath the seafloor, most of the minerals will precipitate in the stockwork rather than on the seafloor.

Hydrothermal deposits on mid-ocean ridges are composed of (1) vertical structures, including individual conduits known as 'chimneys' (Figure 2) and larger structures of coalesced conduits that are often called 'edifices', (2) horizontal 'flange' structures that extend outwards from chimneys and edifices, (3) mounds of accumulated mineral precipitates (Figure 2), and (4) horizontal layers of hydrothermal sediments, debris, and encrustations. Chimneys initially are built directly on top of the seabed around focused jets of high-temperature effluents. Chimneys and edifices are physically unstable and often break or collapse into pieces that accumulate into piles of debris. The debris piles are cemented into consolidated mounds by precipitation of minerals from solutions percolating through the piles. New chimneys are constructed on top of the mounds as the mounds grow in size. Hydrothermal plume particles and particulate debris from chimneys settle around the periphery of the mounds to form layers of hydrothermal sediment. Diffuse seepage of fluids also precipitates mineral encrustations on mound surfaces, on volcanic flows and sediments, and on biological substrates such as microbial mats or the shells and tubes of sessile macrofauna.

The morphology of chimneys is highly variable and evolves as the chimneys grow, becoming more complex with time. Black smoker chimneys often are colonized by organisms and evolve into 'white smokers' that emit diffuse, diluted vent fluids through a porous carapace of wormtubes (Figure 2). Fluid compositions and temperatures, flow dynamics, and biota are all factors that influence the development of chimney morphology. Both the complexity of the interactions between these factors and the high degree of spatial–temporal heterogeneity in the physical, chemical, biological, and geological conditions influencing chimney growth account for the diversity in

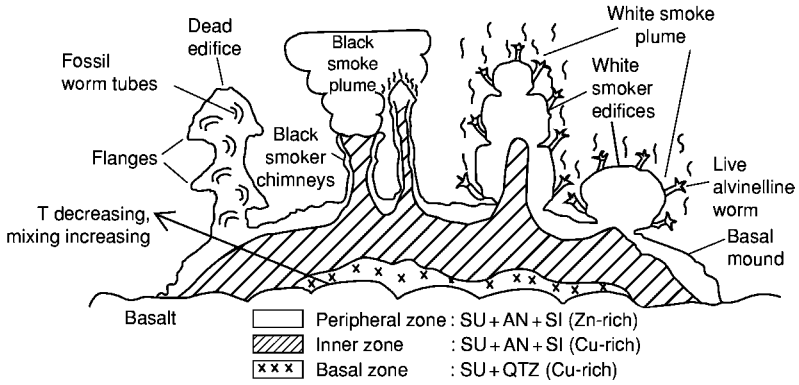


Figure 2 Composite sketch of the mineral structures and zones in hydrothermal mineral deposits on unsedimented ridge crests. Although mound interiors are seldom observed on the seafloor, the simplified sketch of m neral zoning within the mound is predicted by analogy with chimneys and massive sulphide deposits exposed in ophiolites. An outer peripheral zone of Zn rich sulphide (SU; dominantly $ZnS + FeS_2$), anhydrite (AN), and amorphous silica (SI) is replaced in the interior by an inner zone of Cu rich sulphide ($CuFeS_2 + FeS_2$), minor anhydrite, and amorphous silica. The inner zone may be replaced by a basal zone of Cu rich sulphide ($CuFeS_2 + FeS_2$) and quartz (QTZ). Zones migrate as thermochemical conditions within the mound evolve. Although not shown here, it is expected that zoning around individual fractures cutting through the mound will be superimposed on the simplified zone structure depicted in this sketch. T temperature. Modified with permission from Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America: The Eastern Pacific Ocean and Hawaii*, vol. N, pp. 125 144. Boulder: Geological Society of America.

morphology exhibited by chimneys, and presents a challenge to researchers attempting to unravel the processes producing specific morphologic features.

The sizes of hydrothermal mineral deposits on ridges also vary widely. It has been suggested that the largest deposits are accumulating on sedimented ridges, where almost all of the metals in the fluids are deposited within the sediments rather than being dispersed into the oceans by hydrothermal plumes. On unsedimented ridges, the structures deposited on the seabed at fast spreading rates typically are relatively small in dimension (mounds are typically less than a few metres in thickness and less than tens of metres in length, and vertical structures are <15 m high). On intermediate- and slow-spreading ridges, mounds are sometimes much larger (up to tens of metres in thickness, and up to 300 m in length). On the Endeavour Segment of the Juan de Fuca Ridge, vertical structures reach heights of 45 m, and the edifices are even taller at the Lost City site on the Mid-Atlantic Ridge. The size of a deposit depends on many factors, including the magnitude of the heat source, which influences the duration of venting and mineral deposition; whether deposition recurs episodically at a particular site, which depends on the nature of the heat source and plumbing system and on the rate of seafloor spreading; the frequency with which deposits are buried beneath lava flows; and the chemical compositions of vent fluids and chimney minerals, which are affected by site-specific subseafloor rock composition and water–rock reactions. The large deposits found on slower spreading ridge crests are located on faults that have moved slowly away from the ridge axis and have experienced repeated episodes of venting and accumulated mineral deposition over thousands of years, without being buried by lava flows. The tall Endeavour Segment edifices are formed because ammonia-enriched fluid compositions favour precipitation of silica in the edifice walls. The silica is strong enough to stabilize these structures so that they do not collapse as they grow taller. At Lost City, precipitation of sturdy calcium carbonate permits very tall edifices to form.

How Do Chimneys Grow?

A relatively simple two-stage inorganic growth model has been advanced to explain the basic characteristics of black smoker chimneys (Figure 3). In this model, a chimney wall composed largely of anhydrite (calcium sulphate) precipitates initially from seawater that is heated around discharging jets of hydrothermal fluid. The anhydrite-rich chimney wall precipitated during Stage I contains only a small component of metal sulphide mineral particles that crystallize from rapid chilling of the hydrothermal fluids. In Stage II, the anhydrite-rich wall continues to grow upwards and to thicken radially, protecting the fluid flowing through the chimney from very rapid chilling and dilution by seawater. This allows metal sulphide minerals to precipitate into the central conduit of the chimney from the hydrothermal fluid. The hydrothermal fluid percolates outward through the chimney wall, gradually replacing anhydrite and filling voids with metal sulphide minerals. During Stage II, the chimney increases in height, girth, and wall thickness, and both the calcium sulphate/metal sulphide ratio and permeability of the walls decrease. Equilibration of minerals with pore fluid in the walls occurs continuously along steep, time-variant temperature and chemical gradients between fluids in the central conduit and seawater surrounding the chimney. This equilibration produces sequences of concentric mineral zones across chimney walls that evolve with changes in thermal and chemical gradients and wall permeability (Figure 4).

The model of chimney growth is accurate but incomplete, because it does not include the effects on chimney development of fluid-phase separation, biological activity, or variations in fluid composition. Augmented models that address these complexities are needed to characterize fully the processes governing chimney growth.

Elemental and Mineral Compositions of Deposits

Typical ridge crest hydrothermal deposits are dominantly composed of iron, copper, and zinc sulphide minerals (*see* **Minerals:** Sulphides); calcium and barium sulphate minerals (*see* **Minerals:** Sulphates); iron oxide and iron oxyhydroxide minerals; and silicate minerals (Table 1). These minerals precipitate from diverse processes, including heating of seawater; cooling of hydrothermal fluid; mixing between seawater and hydrothermal fluid; reaction of hydrothermal minerals with fluid, seawater, or fluid–seawater mixtures; reaction between hydrothermal fluid and seafloor rocks and sediments; and reactions that are biologically mediated or catalysed. This diversity in the processes and environments of mineral precipitation results in deposition of many different minerals and elements (Tables 1 and 2). High concentrations of strategic and precious metals are found in some deposits (Table 2). The deposits are potentially valuable, if economic and environmentally safe methods of mining them can be developed.

Chimneys can be classified broadly by composition into five groups: sulphate-rich, copper-rich, zinc-rich,

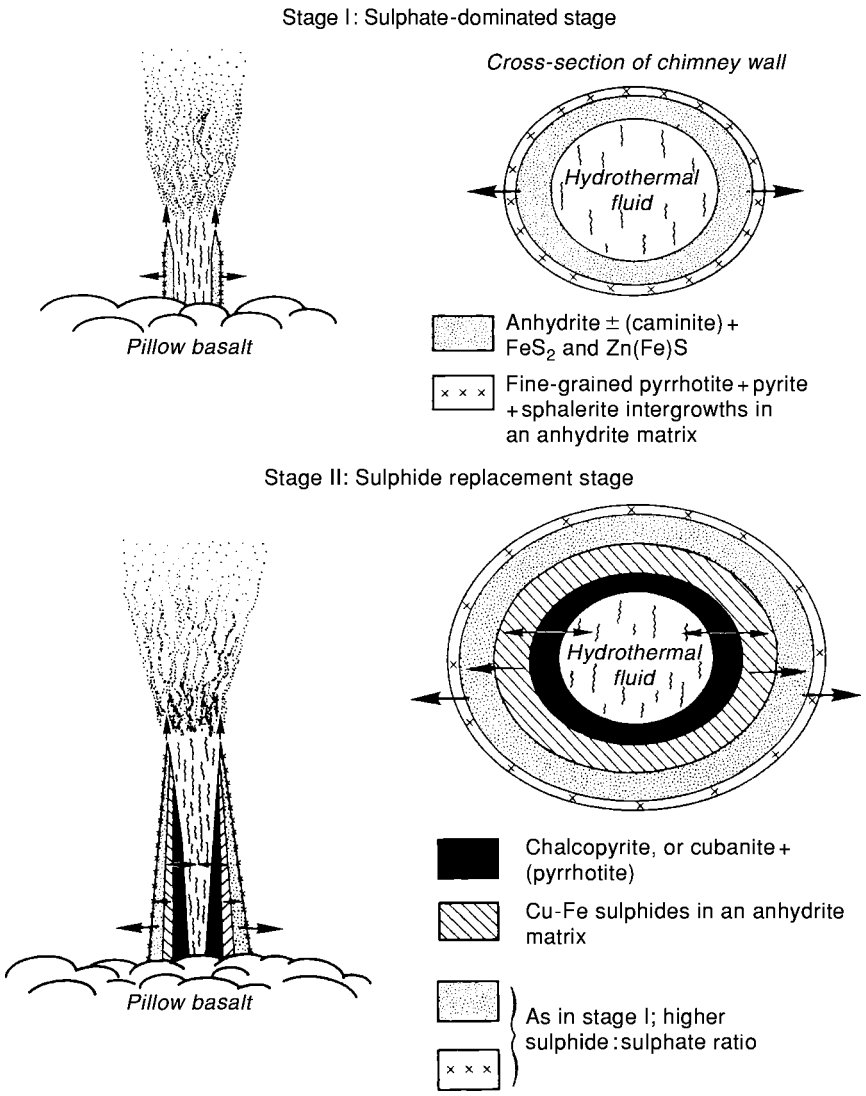


Figure 3 Two stage model of black smoker chimney growth. During Stage II, several different sulphide mineral zonation sequences develop, depending on permeability and thickness of chimney walls, hydrodynamic variables, and hydrothermal fluid composition. Arrows indicate directions of growth.

silica-rich, and carbonate-rich structures. Copper-rich chimney compositions are indicative of formation at temperatures above ~300°C. Sulphate-rich compositions are characteristic of active and immature chimneys formed at temperatures above ~150°C. Carbonate-rich chimneys are found in areas where discharging hydrothermal fluids have reacted with carbonate sediments or ultramafic igneous rocks, and may form at relatively low temperatures. Many chimneys are mineralogically zoned, with hot interior regions enriched in copper and cooler exterior zones enriched in iron, zinc, and sulphate (Figures 2–4). Mounds exhibit a similar gross mineral zoning, and those that are exposed by erosion in ophiolites often have silicified (quartz-rich) interiors (Figure 2). Seafloor weathering of deposits after active venting ceases results in dissolution of anhydrite and in oxidation and dissolution of metal sulphide minerals. Small deposits that are not sealed by silicification or buried by lava flows will not be well preserved in the geological record.

Chimneys as Habitats

Chimney and mound surfaces are substrates populated by microbial colonies and sessile organisms such as vestimentiferan worms (Figure 1), polychaete worms, limpets, mussels, and clams. It is likely that pore spaces in exterior regions of chimney walls and mounds also are inhabited by microbes. All of these chemosynthesis-dependent organisms benefit from the seepage of hydrothermal fluid through active

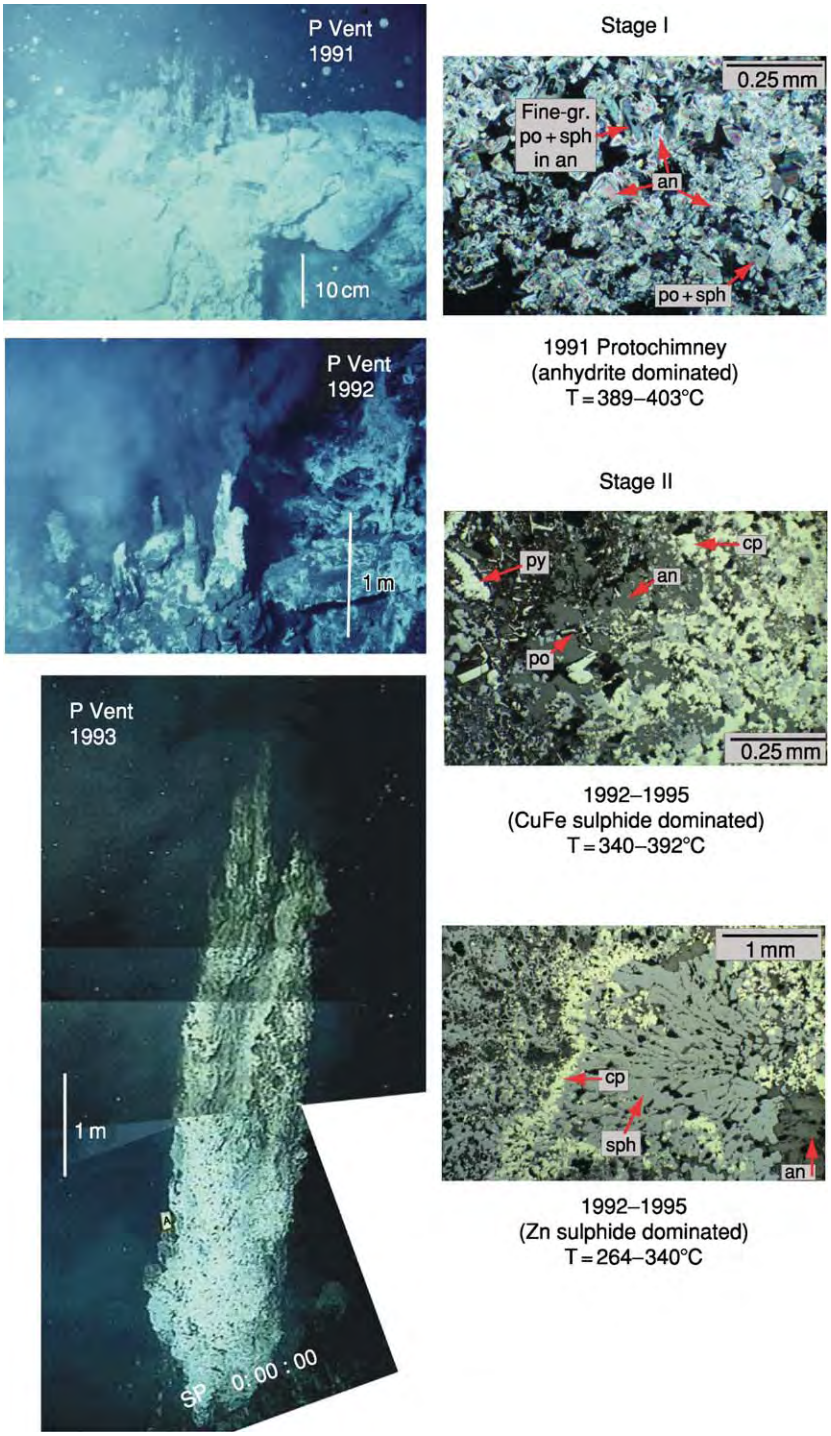


Figure 4 Morphological and mineralogical evolution of chimneys. (Left): A time series of seafloor photographs showing the morpho logical development of a chimney that grew on top of lava flows that erupted in 1991 on the crest of the East Pacific Rise near 9° 50.3′ N. Within a few days to weeks after the eruption, anhydrite rich Stage I 'protochimneys' a few centimetres high had formed where hot fluids emerged from volcanic outcrops that were covered with white microbial mats (top left). Eleven months later, the chimney consisted of cylindrical Stage II anhydrite sulphide mineral spires approximately 1 m in height, and as yet unpopulated by macrofauna (middle left). Three and a half years after the eruption, the cylindrical conduits had coalesced into a 7 m high chimney structure that was covered with inhabited Alvinelline wormtubes (bottom left). (Right): Photomicrographs of chimney samples from the eruption area that show how the chimneys evolved from Stage I (anhydrite dominated; top right) to Stage II (metal sulphide dominated) mineral compositions (see text). As the fluids passing through the chimneys cooled below ~330°C during Stage II, the CuFe sulphide minerals in the chimney walls (middle right) were replaced by Zn sulphide and Fe sulphide minerals (bottom right). Abbreviations: gr, grained; po, pyrrhotite; an, anhydrite; py, pyrite; cp, chalcopyrite. Photographs reproduced from Haymon R, Fornari D, Von Damm K, *et al.* (1993) Volcanic eruption of the mid ocean ridge along the East Pacific Rise at 9° 45 52′ N: direct submersible observation of seafloor phenomena associated with an eruption event in April, 1991. *Earth and Planetary Science Letters* 119: 85 101.

Table 1 Minerals occurring in ocean ridge hydrothermal mineral deposits

Mineral group/name	*Chemical formula*
Sulphides/sulphosalts	
Most abundant	
Sphalerite	Zn(Fe)S
Wurtzite	Zn(Fe)S
Pyrite	FeS_2
Chalcopyrite	$CuFeS_2$
Less abundant	
Iss (isocubanite)[a]	Variable ($CuFe_2S_3$)
Marcasite	FeS_2
Melnikovite	FeS_{2-x}
Pyrrhotite	$Fe_{1-x}S$
Bornite chalcocite	Cu_5FeS_4 Cu_2S
Covellite	CuS
Digenite	Cu_9S_5
Idaite	$Cu_{5\ 5}FeS_{6\ 5}$
Galena	PbS
Jordanite	$Pb_9As_4S_{15}$
Tennantite	$(Cu,Ag)_{10}(Fe,Zn,Cu)_2As_4S_{23}$
Valeriite	$2(Cu,Fe)_2S_23(Mg,Al)(OH)_2$
Sulphates	
Anhydrite	$CaSO_4$
Gypsum	$CaSO_4$ H_2O
Barite	$BaSO_4$
Caminite	$MgSO_4$ $x Mg(OH)_2$ (1 2x)H_2O
Jarosite natrojarosite	$(K,Na)Fe_3(SO_4)_2(OH)_6$
Chalcanthite	$CuSO_4$ $5H_2O$
Carbonates	
Magnesite	$MgCO_3$
Calcite	$CaCO_3$
Aragonite	$CaCO_3$
Elements	
Sulphur	S
Oxides/hydroxides	
Goethite	FeO(OH)
Lepidocrocite	FeO(OH)
Hematite	Fe_2O_3
Magnetite	Fe_3O_4
Brucite	$Mg(OH)_2$
Amorphous Fe compounds	
Amorphous Mn compounds	
Psilomelane	$(Ba,H_2O)_2Mn_5O_{10}$
Silicates	
Opaline silica	SiO_2 nH_2O
Quartz	SiO_2
Talc	$Mg_3Si_4O_{10}(OH)_2$
Nontronite	$(Fe,Al,Mg)_2(Si_{3\ 66}Al_{0\ 34})O_{10}(OH)_2$
Illite smectite	
Aluminosilicate colloid	
Hydroxychlorides	
Atacamite	$Cu_2Cl(OH)_3$

[a]Iss, Intermediate solid solution.

Table 2 Ranges of elemental compositions in bulk mid ocean ridge hydrothermal mineral deposits

Element	*Range*[a]
Cu	0.1 15.0 wt.%
Fe	2.0 44.0 wt.%
Zn	<0.1 48.7 wt.%
Pb	0.003 0.6 wt.%
S	13.0 52.2 wt.%
SiO_2	<0.1 28.0 wt.%
Ba	<0.01 32.5 wt.%
Ca	<0.1 16.5 wt.%
Au	<0.1 4.6 ppm
Ag	3.0 303.0 ppm
As	7.0 918.0 ppm
Sb	2.0 375.0 ppm
Co	<2.0 3500.0 ppm
Se	<2.0 224.0 ppm
Ni	<1.5 226.0 ppm
Cd	<5 1448 ppm
Mo	1.0 290.0 ppm
Mn	36.0 1847.0 ppm
Sr	2.0 4300.0 ppm

[a]Data from Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America*: *The Eastern Pacific Ocean and Hawaii*, vol. N, pp. 125 144. Boulder: Geological Society of America. Hannington MD, Jonasson IR, Herzig PM, and Petersen S (1995) In: Humphris SE, Zeirenberg RA, Mullineaux LS, and Thomason RE (eds), *Physical and chemical processes of seafloor mineralization at mid ocean ridges*, p. 115 157.

mineral structures and from the thermal and chemical gradients across mineral structures. The structures provide an interface between seawater and hydrothermal fluid that maintains tolerable temperatures for biota, and allows organisms simultaneous access to the chemical constituents in both seawater and hydrothermal fluid. However, organisms attached to active mineral structures must cope with changes in fluid flow across chimney walls (which sometimes occur rapidly) and with ongoing engulfment by mineral precipitation.

Some organisms actively participate in the precipitation and breakdown of minerals; for example, sulphide-oxidizing microbes mediate the crystallization of native sulphur crystals, and microbes are also thought to participate in the precipitation of marcasite and iron oxide minerals. Additionally, the surfaces of organisms provide favorable sites for nucleation and growth of amorphous silica, metal sulphide, and metal oxide crystals, and this facilitates mineral precipitation and fossilization of vent fauna.

Fossil Record of Hydrothermal Vent Organisms

Fossil moulds and casts of wormtubes, mollusc shells, and microbial filaments have been identified in both modern ridge hydrothermal deposits and in ancient deposits of Cretaceous, Jurassic, Devonian, and Silurian ages. This fossil record establishes the antiquity of vent communities and the long evolutionary history of specific faunal groups. The singular Jurassic fossil

assemblage preserved in a small ophiolite-hosted deposit in central California is particularly interesting because it contains fossils of vestimentiferan worms, gastropods, and brachiopods, but no clam or mussel fossils. In contrast, modern and Palaeozoic faunal assemblages described thus far include clams, mussels, and gastropods, but no brachiopods. Does this mean that brachiopods have competed with molluscs for ecological niches at vents, and have moved in and out of the hydrothermal vent environment over time? Fossilization of organisms is a selective process that does not preserve all of the fauna that are present at vents. Identification of fossils at the species level is often difficult, especially where microbes are concerned. Notwithstanding, it is important to search for more examples of ancient fossil assemblages and to trace the fossil record of life at hydrothermal vents back as far as possible, to shed light on questions about how vent communities have evolved and about whether life on Earth might have originated at submarine hydrothermal vents.

Summary

Hydrothermal activity is an integral aspect of seafloor accretion at mid-ocean ridges. The circulating fluids and plumes facilitate thermal and chemical exchange between the oceans and the lithosphere, support life above and below the seafloor, and affect current flow and biological activity at mid-water depths. The mineral deposits are valuable for their metals, for the role that they play in fostering hydrothermal vent ecosystems, for the clues that they hold to understanding spatial–temporal variability in hydrothermal vent systems, and as geological records of how life at hydrothermal vents has evolved. Vent organisms exhibit novel biochemistry, genetics, taxonomy, physiology, symbiosis, and community dynamics. From submarine hydrothermal systems, the insights gained about biogeochemical processes at high temperatures and pressures can be applied to biotechnology and to understanding life in inaccessible realms within Earth's crust, or the potential for life on other planetary bodies. As the complexities of hydrothermal systems on the mid-ocean ridge are unravelled through ongoing exploration and interdisciplinary studies, exciting applications of this knowledge are being discovered.

See Also

Minerals: Sulphates; Sulphides. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Origin of Life**. **Plate Tectonics**. **Seamounts**. **Tectonics:** Mid-Ocean Ridges.

Further Reading

Baker ET (1996) Geological indexes of hydrothermal venting. *Journal of Geophysical Research* 101(B6): 13 741 13 753.

Dilek Y, Moores E, Elthon D, and Nicolas A (eds.) (2000) *Ophiolites and Oceanic Crust: New Insights from Field Studies and the Ocean Drilling Program. Geological Society of America Memoir.* Boulder: Geological Society of America.

Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America: The Eastern Pacific Ocean and Hawaii,* vol. N, pp. 125 144. Boulder: Geological Society of America.

Haymon RM (1996) The response of ridge crest hydrothermal systems to segmented, episodic magma supply. In: MacLeod CJ, Tyler P, and Walker CL (eds.) *Tectonic, Magmatic, Hydrothermal, and Biological Segmentation of Mid Ocean Ridges, Special Publication,* vol. 118, pp. 157 168. London: Geological Society.

Haymon R, Fornari D, Von Damm K, *et al.* (1993) Volcanic eruption of the mid ocean ridge along the East Pacific Rise at 9° 45 52′ N: direct submersible observation of seafloor phenomena associated with an eruption event in April, 1991. *Earth and Planetary Science Letters* 119: 85 101.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thomson RE (eds.) (1995) *Seafloor Hydrothermal Systems: Physical, Chemical, Biological, and Geological Interactions, Geophysical Monograph*. vol. 91. Washington, DC: American Geophysical Union.

Kelley DS, Karson JA, Blackman DK, *et al.* (2001) An off axis hydrothermal field near the Mid Atlantic Ridge at 30° N. *Nature* 412: 145 149.

Little CTS, Herrington RJ, Haymon RM, and Danelian T (1999) Early Jurassic hydrothermal vent community from the Franciscan Complex, San Rafael Mountains, California. *Geology* 27: 167 170.

Schrenk MO, Kelley DS, Delaney JR, and Baross JA (2003) Incidence and diversity of microorganisms within the walls of an active deep sea sulfide chimney. *Applied and Environmental Microbiology* 69: 3580 3592.

Shank TM, Fornari DJ, Von Damm KL, *et al.* (1998) Temporal and spatial patterns of biological community development at nascent deep sea hydrothermal vents (9° 50′ N, East Pacific Rise). *Deep Sea Research II* 45: 465 515.

Tivey MK, Stakes DS, Cook TL, Hannington MD, and Petersen S (1999) A model for growth of steep sided vent structures on the Endeavour Segment of the Juan de Fuca Ridge: results of a petrological and geochemical study. *Journal of Geophysical Research* 104: 22 859 22 883.

Von Damm KL (2000) Chemistry of hydrothermal vent fluids from 9 10° N, East Pacific Rise: "Time Zero" the immediate post eruptive period. *Journal of Geophysical Research* 105: 11 203 11 222.

Propagating Rifts and Microplates At Mid-Ocean Ridges

R N Hey, University of Hawaii at Manoa, Honolulu, HI, USA

Introduction

Propagating rifts are extensional plate boundaries that gradually break through lithospheric plates, forming new plate boundaries and rearranging the geometries of old ones. If the rifting advances to the seafloor-spreading stage, propagating seafloor-spreading centres gradually extend through the rifted lithosphere. This evolution occurs rapidly ($\sim 10^5$ years) for oceanic propagators and much more slowly ($\sim 10^7$ years) for continental ones. The orthogonal combination of seafloor spreading and propagation produces a characteristic V-shaped wedge of lithosphere, with progressively younger and longer isochrons abutting the 'pseudofaults' that bound this wedge. Oceanic propagators generally replace pre-existing spreading centres, causing lithospheric transfer from one plate to another, and sequences of spreading-centre jumps, leaving failed rift systems in their wakes. This changes the classic plate tectonic geometry and results in asymmetric accretion of lithosphere to the plates. There is pervasive shear deformation in the overlap zone between the propagating and failing rifts, much of it accommodated by bookshelf faulting, in which, e.g., right-lateral plate motion shear produces high-angle left-lateral fault slip. When the scale or strength of the overlap zone becomes large enough, it stops deforming, and instead begins to rotate as a separate microplate between dual active spreading centres. Continental propagators break apart continents and can leave failed rifts (aulocogenes) along unsuccessful propagation paths, as well as predictable patterns of deformation ahead of the propagator tips. For continental rifting, the pseudofaults are the continental margins. Rift propagation appears to be the primary mechanism by which Earth's accretional plate boundary geometry is reorganized.

Oceanic Propagators

Figure 1 shows several variations of typical oceanic ridge propagation geometry, in which a pre-existing 'doomed rift' is replaced by the propagator. This always seems to result in at least slight spreading-centre reorientation. Whether this is because rifts propagate in response to changes in direction of seafloor spreading, or because the spreading direction changes while rifts propagate, rift propagation is the primary mechanism by which many seafloor-spreading systems have adjusted to changes in spreading direction. Propagation rates and local spreading rates are often similar in magnitude, although propagation rates as high as 1000 km My^{-1} have been discovered.

Figure 1A shows the discontinuous propagation model, in which periods of seafloor spreading alternate with periods of instantaneous propagation,

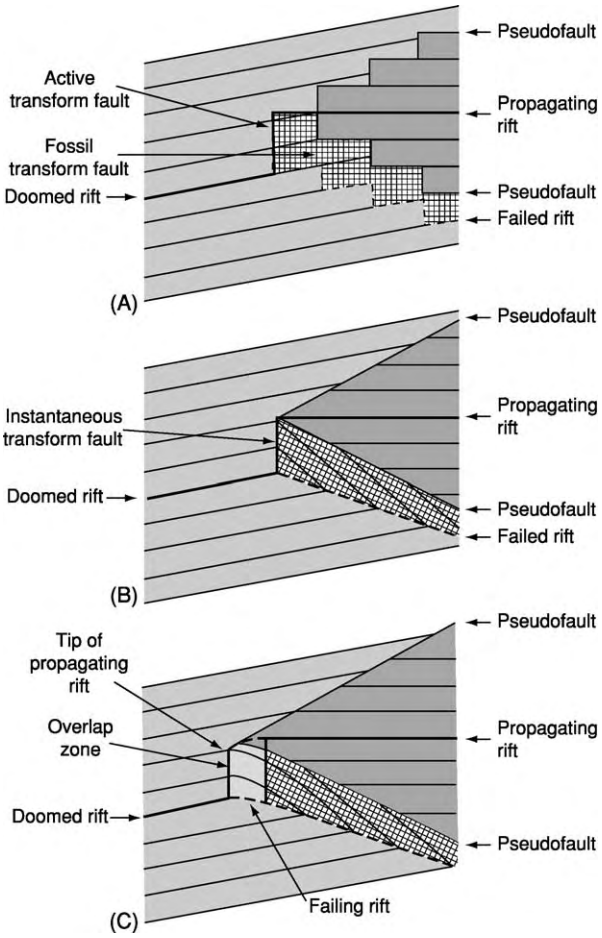


Figure 1 (A) Discontinuous, (B) continuous, and (C) non trans form zone oceanic propagating/failing rift models. Propagating rift lithosphere is marked by dark stipple, normal lithosphere created at the doomed rift is indicated by light stipple, and trans ferred lithosphere is cross hatched. Heavy lines show active plate boundaries. In the non transform zone model(C), active axes with full spreading rate are shown as heavy lines; active axes with transitional rates are shown as dashed lines. The overlap zone joins these transitional spreading axes. Reproduced with permission from Hey RN, Sinton JM, and Duennebier FK (1989) Propagating rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii*, pp. 161 176. Boulder, CO: Geo logical Society of America.

producing *en echelon* failed rift segments, fossil transform faults, and fracture zones, and blocks of progressively younger transferred lithosphere. Figure 1B shows the pattern produced if propagation, rift failure, and lithospheric transferal are all continuous. In this idealized model, a transform fault migrates continuously with the propagator tip, never existing in one place long enough to form a fracture zone, and thus V-shaped pseudofaults are formed instead of fracture zones. Figure 1C shows a more geologically plausible model, in which the new spreading centre requires some finite time to accelerate from zero to the full rate, with concomitant decreases on the failing spreading centre, so that lithospheric transferal is not instantaneous. Instead of a transform fault, a migrating broad 'non-transform' zone of distributed shear deformation connects the overlapping propagating and failing ridges during the period of transitional spreading. Deformation occurring in this overlap zone is preserved in the zone of transferred lithosphere. This zone, left behind as the overlap zone migrates, is bounded by the failed rifts and inner (proximal) pseudofault. Even more complicated geometries occur in some places where the doomed rift, instead of failing monotonically as the propagator steadily advances, occasionally propagates in the opposite direction.

Standard plate tectonic geometry holds for the area outside the pseudofaults and zone of transferred lithosphere, but rigid plate tectonics breaks down in the overlap zone where some of the lithosphere formed on the doomed rift is progressively transferred to the other plate by the rift propagation and resulting migration of the overlap zone. Shear between the overlapping propagating and failing rifts appears to be accommodated by bookshelf faulting, probably along the pre-existing abyssal hill faults. This produces oblique seafloor fabric, with trends quite different from the ridge-parallel and-perpendicular structures expected from classic plate tectonic theory.

Figure 2 is a relief map of the type-example propagating rift, at 95.5° W along the Cocos–Nazca spreading centre. This area closely resembles the geometry in Figure 1C, except that rift failure occurs more discontinuously. This propagator is breaking westward away from the Galapagos hotspot through 1-My-old Cocos lithosphere at a velocity of about 50 km My $^{-1}$. Well-organized seafloor spreading begins about 10 km behind the faulting, fissuring, and extension at the propagating rift tip. This 200 000-year time lag between initial rifting and the rise of asthenosphere through the lithospheric cracks to form a steady-state spreading centre suggests an asthenospheric viscosity of about 10^{18} Pa s. The combination of seafloor spreading (about 60 km My $^{-1}$) and propagation produces a V-shaped wedge of young lithosphere bounded by pseudofaults and pre-existing lithosphere. The propagating rift lithosphere is characterized by high-amplitude magnetic anomalies and by unusual petrological diversity, including highly fractionated ferrobasalts. This propagator is replacing a pre-existing spreading system about 25 km to the south, and thus spreading-centre jumps and failed rifts result. Although propagation is continuous, segmented failing rift grabens seem to die episodically on a time-scale of about 200 000 years. This has produced a very systematic pattern of spreading-centre jumps identified from magnetic anomalies, in which each jump was younger and slightly longer than the preceding jump. The spreading-centre orientation is being changed clockwise by about 13°, and more than 10^{4} km^{3} My $^{-1}$ of Cocos lithosphere is being transferred to the Nazca plate.

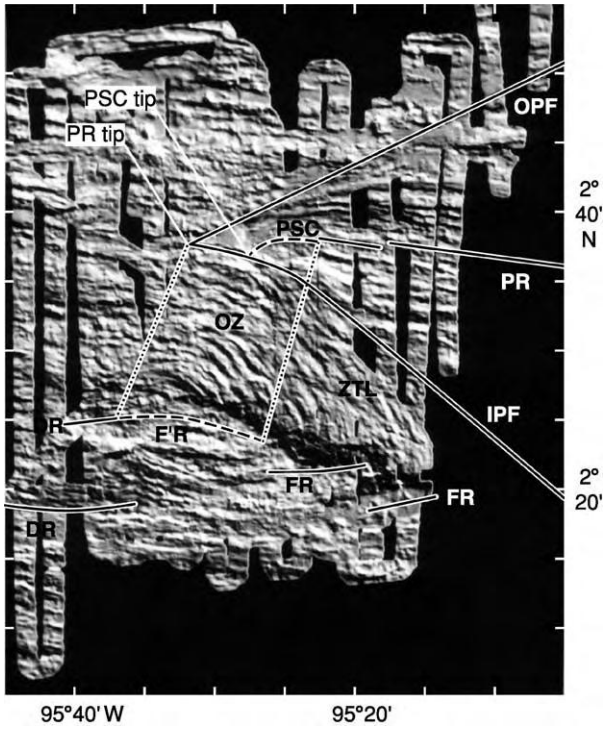


Figure 2 Shaded relief map of digital Sea Beam swath bathymetry at the Galapagos propagating rift system (95.5° W). The relative plate motion is nearly north south; propagation is to the west. The oblique structures in the overlap zone and its wake, the zone of transferred lithosphere, are clearly evident. PR, Propagating rift; PSC, propagating spreading centre; OPF, IPF, outer and inner pseudofaults; OZ, overlap zone; ZTL, zone of transferred lithosphere; DR, doomed rift; F'R, failing rift; FR, failed rift grabens. Adapted by permission from Hey RN, Sinton JM, and Duennebier FK (1989) Propagating rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii*, pp. 161 176. Boulder, CO: Geological Society of America.

The active propagating and failing rift axes overlap by about 20 km and are connected by a broad and anomalously deep zone of distributed shear deformation rather than by a classic transform fault. Most of the seismic activity occurs within this 'non-transform' zone, where the pre-existing abyssal hill fabric originally created on the doomed rift is sheared and tectonically rotated into new oblique trends. Simple equations accurately describe this geometry in terms of ratios of propagation and spreading rates, together with the observed propagating and doomed rift azimuths. For example, for the simplest continuous propagation geometry, if u is the spreading half-rate and v is the propagation velocity, the pseudofaults form angles $\tan^{-1}(u/v)$ with the propagator axis, and the isochrons and abyssal hill fabric in the zone of transferred lithosphere have been rotated by an angle $\tan^{-1}(2u/v)$.

The boundaries of the Galapagos high-amplitude magnetic anomaly zone, the ferrobasalt province, and the spreading-centre jumps are all coincident with the pseudofaults bounding the propagating rift lithosphere. All of these observations can be explained as mechanical and/or thermal consequences of a new rift and spreading centre breaking through cold lithosphere, with increased viscous head loss and diminished magma supply on the propagating spreading centre close to the propagator tip. This leads to an unusually deep axial graben and unusually extensive fractional crystallization and differentiation. The 95.5° W propagator tip is also a mantle geochemistry boundary, implying that this rift propagation is associated with plume-related subaxial asthenospheric flow away from the Galapagos hotspot.

Causes of Rift Propagation

One important observation is that many rifts and spreading centres propagate down topographic gradients away from hotspots or shallow ridge axis topography. For example, all six known active Galapagos propagators are propagating away from the hotspot. Plume-related asthenospheric flow generates gravitational stresses on the shallow spreading-centre segments near the hotspot that promote crack propagation away from the hotspot. Flow of asthenosphere into these cracks produces new lithosphere at propagating seafloor-spreading centres. Regionally high deviatoric tensile stresses associated with regional uplift provide a quantitatively plausible driving mechanism. Crack growth occurs when the stress concentration at the tip, characterized in elastic fracture mechanics by a stress intensity factor, exceeds the resisting stress intensity contribution. The spreading-centre propagation rate could be limited by the viscosity of the asthenosphere flowing into the rift, producing viscous suction forces at a local tip depression, or by process zone deformation at the rift tip. The overlap/offset ratio of propagating and failing rifts tends to be ~1, close to the ratio at which the stress intensity factor is maximized (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges).

Although many rifts appear to propagate in response to hotspot-related stresses, others appear to propagate because of stresses producing changes in plate motion. Subduction-related stresses appear to be a common mechanism for producing propagation in the North-east Pacific. This probably explains most of the massive reorganizations of the spreading geometry as the Pacific–Farallon ridge neared the Farallon–North America trench, as clearly evident even in the classic Raff–Mason magnetic anomaly data (Figure 3), although some propagation away from the Axial Seamount hotspot has also occurred in the Juan de Fuca area. Propagation may be produced in many ways over a wide range of scales, including small-scale propagation of overlapping spreading centres away from local magmatic centres. The larger reorganizations sometimes involve transient microplate formation, geometrically similar to the broad overlap zone model of Figure 1C, but on a much larger scale.

Microplates

Microplates are small, mostly rigid areas of lithosphere, located at major plate boundaries but rotating as more or less independent plates. They can form in many tectonic settings. The two main types along mid-ocean ridges, those formed at triple junctions and those formed away from triple junctions, share many similarities. Although it was once thought that stable, growing microplates could eventually grow into major oceanic plates, it now appears that these are transient phenomena resulting from large-scale rift propagation. When the overlap zone becomes too big and strong to deform by pervasive bookshelf faulting, it changes mechanical behaviour and accommodates the boundary plate motion shear stresses by beginning to rotate as a separate microplate. The most well-studied oceanic microplates are the Easter microplate along the Pacific–Nazca ridge and the Juan Fernandez microplate at the Pacific–Nazca–Antarctica triple junction (Figure 4). Despite their different tectonic settings, they show many striking similarities.

The scales of the Easter (~500 km diameter) and the Juan Fernandez (~400 km diameter) microplates are similar. The eastern and western boundaries of both microplates are active spreading centres, propagating north and south, respectively. Both microplates began forming about 5 Ma, and both

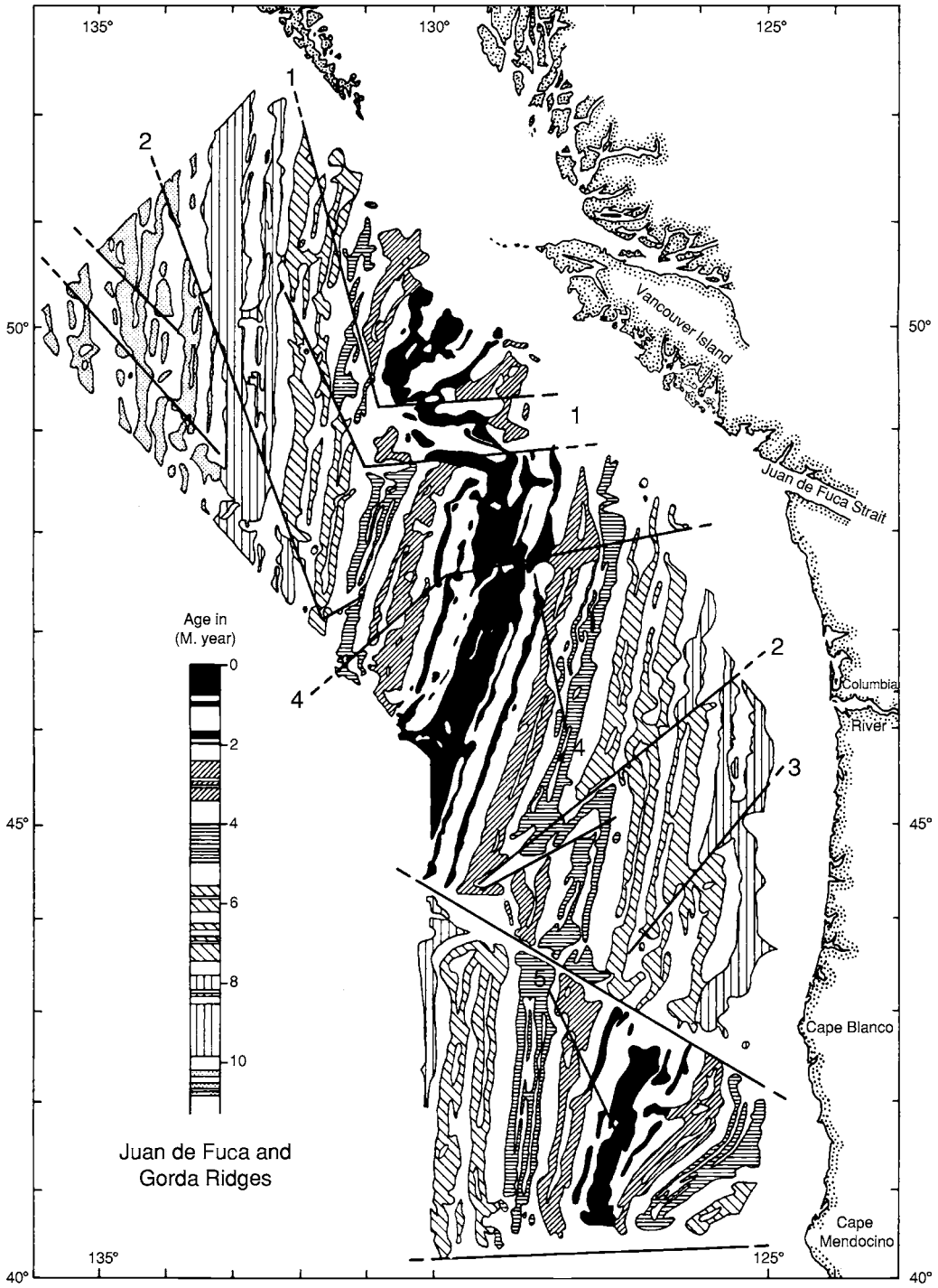


Figure 3 Raff Mason magnetic anomalies in the Juan de Fuca area. Positive anomalies are shaded. Numbers denote major propagation sequences. Reproduced with permission from Hey RN (1977) A new class of pseudofaults and their bearing on plate tectonics: A propagating rift model. *Earth and Planetary Science Letters* 37: 321 325.

East rifts have been propagating into roughly 3-My-old Nazca lithosphere. Extremely deep axial valleys occur at their tips, ~6000 m at Pito Deep at the north-eastern Easter microplate boundary, and ~5000 m at Endeavour Deep at the north-eastern Juan Fernandez microplate boundary. The northern and southern boundaries are complicated deformation zones, with zones of shear, extension, and significant areas of compression. The dominant (East) rift of the Easter microplate, the dominant (West)

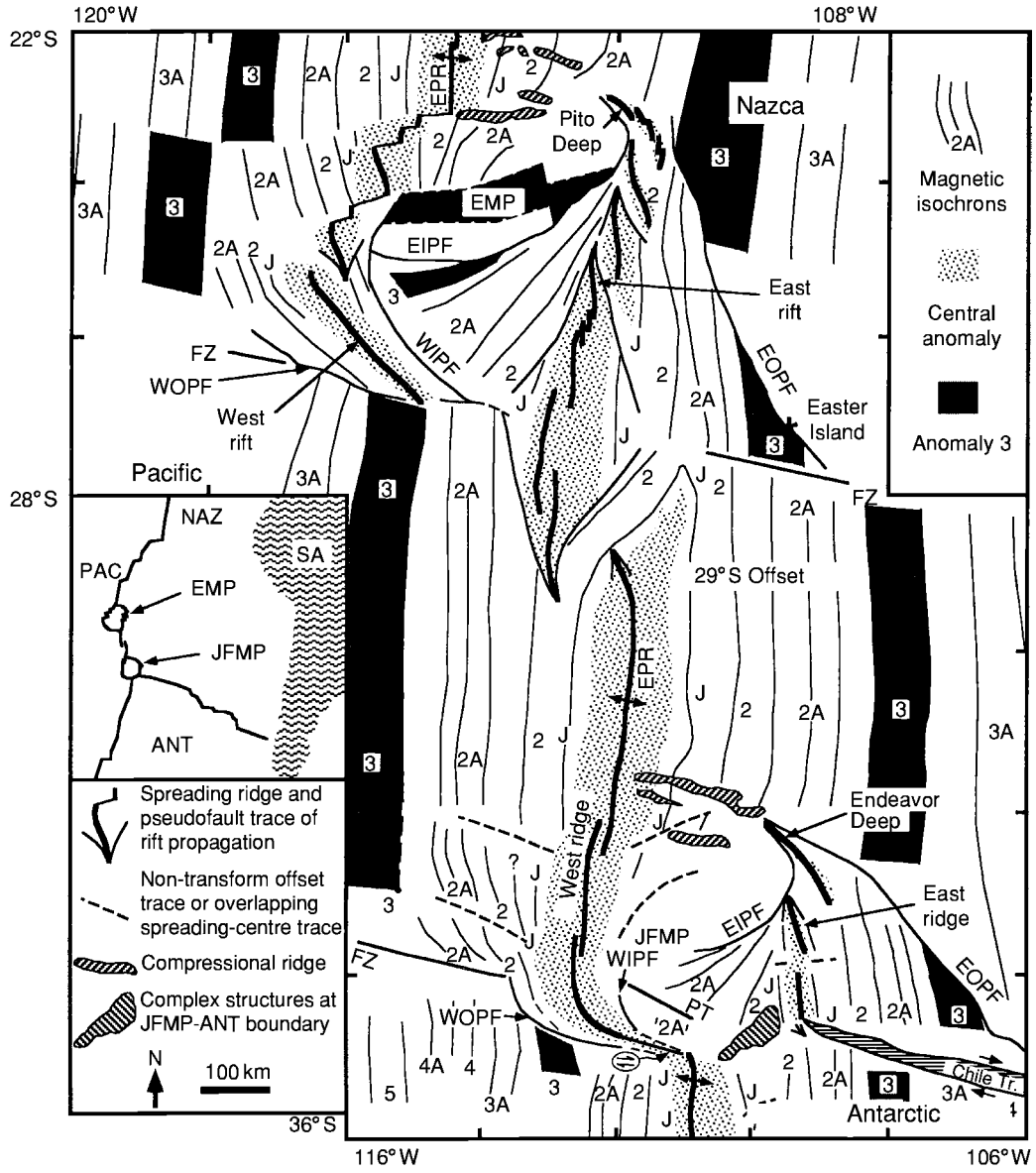


Figure 4 Tectonic boundaries, magnetic isochrons, and structures of the Easter microplate (EMP) and Juan Fernandez microplate (JFMP). EPR, East Pacific Rise; FZ, fracture zone; NAZ, Nazca; PAC, Pacific; SA, South America; ANT, Antarctic; WOPF, WIPF, EOPF, and EIPF are western outer and inner and eastern outer and inner pseudofaults, respectively. Numbers (e.g. 2, 2A) identify magnetic chrons from magnetic reversal time scale, J is Jaramillo reversal ~1 Ma, PT is paleotransform. Reproduced with permission from Bird RT and Naar DF (1994) Intratransform origins of mid ocean ridge microplates. *Geology* 22: 987 990.

rift of the Juan Fernandez microplate, and the dominant (West) rift of the duelling propagators between the microplates, are all propagating away from the Easter mantle plume (or the intersection of this plume with the ridge axis), suggesting that microplate formation as well as rift propagation can be driven by plume-related forces.

Both microplates have large ($\sim 100 \times 200$ km), complex, pervasively deformed cores, which may have formed by bookshelf faulting in overlap zones during an initial large-scale propagating rift stage of evolution. Both show more recent stages of growth as independent microplates, with deformation concentrated along the plate boundaries. At present, both microplates are rotating clockwise very rapidly about poles near the microplate centres, spinning like roller-bearings caught between the major bounding plates. The Easter microplate rotation velocity is about 15° My^{-1} and the Juan Fernandez velocity is about 9° My^{-1}. The roller-bearing analogy has been quantified in an idealized edge-driven model of microplate kinematics. If microplate rotation is

indeed driven by shear on the boundaries between the microplate and surrounding major plates, the rotation velocity (in radians) is $2u/d$, where $2u$ is the major plate relative velocity and d is the microplate diameter. This follows because the total spreading on the microplate boundaries must equal the major plate motion if the microplate did not exist. The rotation (Euler) poles describing the motion of the microplate relative to the major plates will lie on the microplate boundaries, at the farthest extensions of the rifts, which must continually lengthen by a different kind of rift propagation as the microplate rotates.

This idealized geometry (Figure 5) requires a circular microplate shape, yet also requires seafloor spreading on the dual active ridges, which must constantly change this shape. The more the microplate grows, the more deformation must occur as it rotates, and the less successful the rigid plate model will be. Although it would appear that this inevitable plate growth would soon invalidate the model, numerous episodes of rift propagation helping to maintain the necessary geometry are observed to have occurred at the Easter and Juan Fernandez microplates. All propagation was on the microplate interior side of the failing rifts, thus transferring microplate lithosphere to the surrounding Pacific and Nazca plates, shaving the new microplate growth at the edges and maintaining a shape circular enough for the edge-driven model to be very successful.

According to the edge-driven model, a microplate may stop rotating if one of the bounding ridge axes

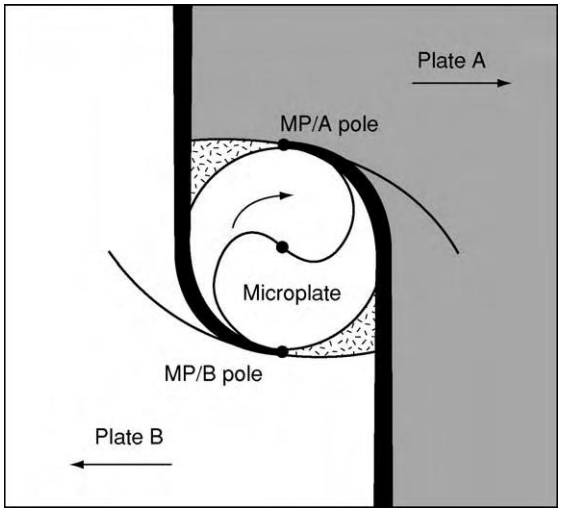


Figure 5 Roller bearing model of microplates based on a simple, concentrically rotating bearing. The microplate is approximated by a circular plate that is caught between two major plates (MP/A and MP/B). The main contacts between microplate and major plates are also the positions of the relative rotation poles (dots). Dark shading shows major spreading centres, overlapping about the microplate. Cross hatched corners are areas of compression. Medium curved lines are predicted pseudofaults; plate arrows show relative motions. This schematic model assumes growth from an infinitesimal point to a present circular shape; the model can be extended to take account of growth from a finite width, eccentric motions, and growth of the microplate. Reproduced with permission from Searle RC, Bird RT, Rusby RI, and Naar DF (1993) The development of two oceanic microplates: Easter and Juan Fernandez microplates, East Pacific Rise. *Journal of the Geological Society* 150: 965–976.

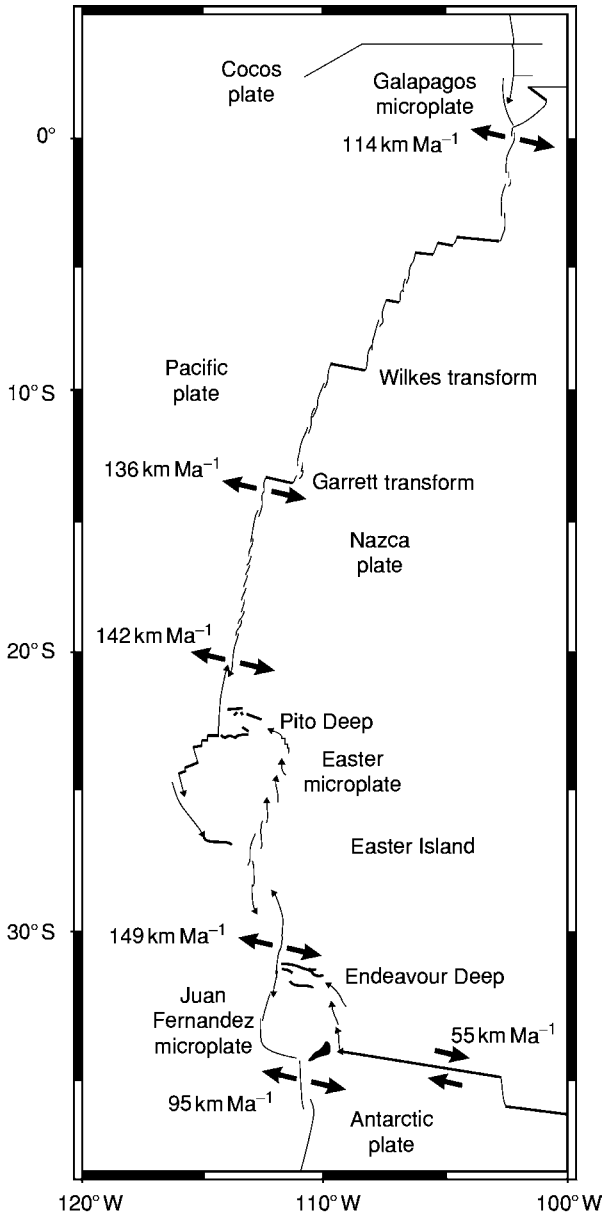


Figure 6 Plate tectonic geometry and relative plate motions along the southern East Pacific Rise. Light lines are ridges, and those with arrows are propagating. Heavy straight lines are transform faults. Reproduced with permission from Hey RN, Johnson PD, Martinez F, *et al.* (1995) Plate boundary reorganization at a large offset, rapidly propagating rift. *Nature* 378: 167–170.

propagates through to the opposite spreading boundary, eliminating coupling to one of the bounding plates. Dual spreading would no longer occur, spreading would continue on only one bounding ridge, and 10^6–10^7 km^3 of microplate lithosphere would accrete to one of the neighbouring major plates. Active microplates are thus modern analogues for how large-scale (hundreds of kilometres) spreading centre jumps occur. There is evidence in the older seafloor record that this happened many times along the ancestral East Pacific Rise.

All large right-stepping offsets along the Pacific–Nazca spreading centre are transform faults, whereas all large left-stepping offsets are microplates or the giant duelling propagators (possible protomicroplate) between the Easter and Juan Fernandez microplates (Figure 6). The Galapagos microplate at the Pacific–Cocos–Nazca triple junction also fits this pattern. This suggests that a recent clockwise change in Pacific–Nazca plate motion could have been an important factor triggering the formation of these microplates. Earth's fastest active seafloor spreading occurs in this area, and all parts of the plate boundary presently spreading faster than 142 km My^{-1} are reorganizing by duelling propagators or microplates (Figure 6). The combination of thin lithosphere produced at these 'superfast' seafloor spreading rates, and the unusually hot asthenosphere produced by an Easter mantle plume, would reduce the forces resisting propagation and thus make these plate boundary reorganizations easier, perhaps explaining their common occurrence in this area.

Microplates also occur in convergent settings, where small pieces of lithosphere are caught between large plates. A well-studied continental convergence example is the Mediterranean, where small plates the size of Turkey and the Aegean adapt to Africa–Eurasia convergence. Oceanic microplates can also form along convergent margins, e.g., in the West Pacific, to accommodate Pacific, Australia, and Asia convergence.

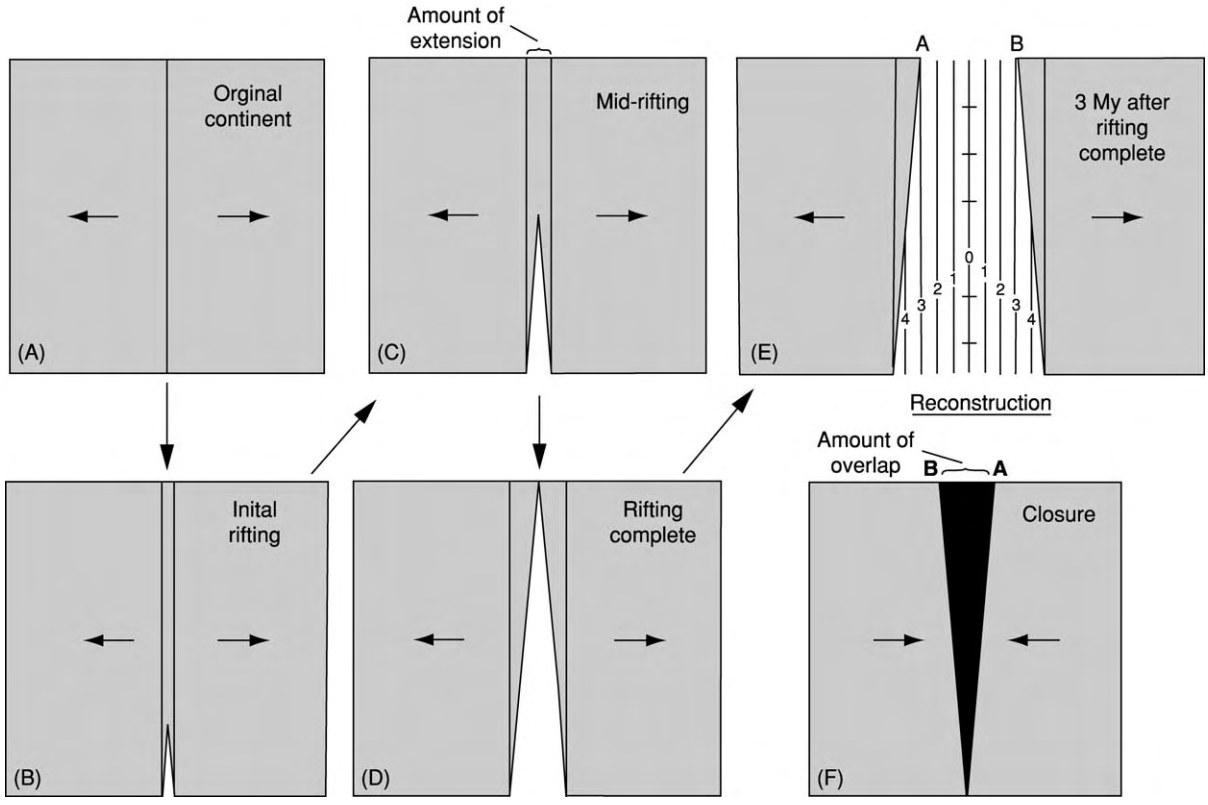


Figure 7 Propagating rift model for continental breakup (map view). (A) Original continent under tension; (B) initial rifting, with the amount of extension represented between parallel lines; (C) mid rifting (seafloor spreading is occurring in the lower half of the continent while crustal thinning and extension occur in the upper half); (D) rifting complete (continental edges have undergone extension that increases in the direction of rifting); (E) 3 My after rifting is complete (oldest seafloor is found in the part of the ocean where rifting began; oldest isochrons converge with the ocean continent boundary; continental edge is not an isochron); (F) reconstruction of the pre rift configuration (extension due to rifting results in apparent overlap when the continents are returned to their pre rift geometry). Reproduced with permission from Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677 10 688.

Continental Propagators

Continents break apart progressively as well, and thus continental margins are not isochrons but instead are a type of pseudofault boundary. During breakup, the margins bound V-shaped wedges of propagating rift lithosphere pointing in the direction of propagation. The relative azimuths of the margin and the first magnetic isochrons formed on the new seafloor-spreading centre give the propagation velocity. For example, the South Atlantic began forming at the southern tip of Africa–South America ~130 Ma and reached the area that became the great equatorial mid-Atlantic ridge (MAR) offset 30–40 My later, for an average northward propagation velocity of 10–15 km My ^{1}, similar to the spreading rate.

Predictable extensional rift deformation occurs ahead of the propagating spreading-centre tip (Figure 7). Plate reconstructions that take this new understanding of breakup deformation into account differ from classic reconstructions, such as the Bullard fits, which have equal amounts of continental overlap and underlap. Instead, increasing deformation ahead of the growing spreading centre means continental reconstruction fits should be exact where propagation begins, and should slow increasing overlap in the direction of propagation, with no areas of continental margin underlap (Figure 8).

Occasional rifts may propagate locally along globally unfavorable paths and eventually fail, resulting in failed rifts (aulocogenes) along margins. Systematic patterns of small failed rifts, or one huge one such as the Greenland–Canada failed rift in the Labrador Basin, would produce asymmetries in the conjugate margins. Occasional episodes of duelling propagation may also occur, forming ephemeral microplates eventually welded to one of the major plates, with corresponding margin asymmetries. One well-studied area of continental propagation is the Afar depression, where the Gulf of Aden propagator is presently breaking west into the African continent in Djibouti; a Red Sea propagator is simultaneously breaking south into the continent in Ethiopia (Figure 9). The interaction of these duelling propagators has produced changes in their propagation directions and the same kind of pervasive bookshelf faulting between the ridges as seen in oceanic rift propagation.

Implications

Rift propagation, which occurs on scales ranging from overlapping spreading centres with offsets of

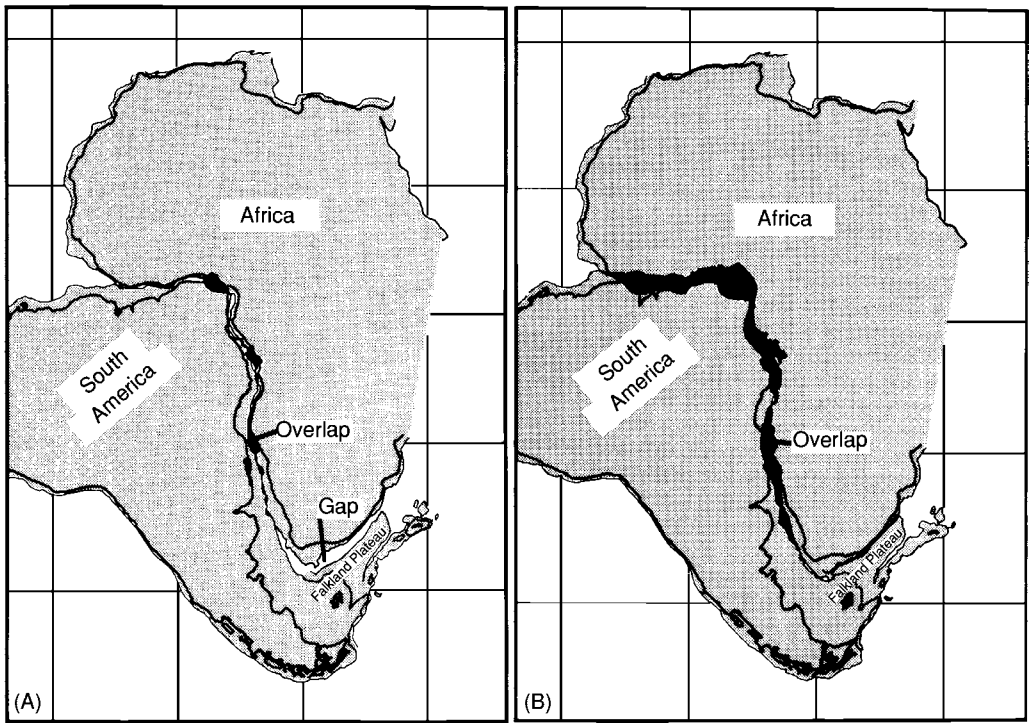


Figure 8 (A) The Bullard South America Africa reconstruction, using the 1000 m isobath to represent the continental edge and the 3000 m isobath for the Falkland Plateau. (B) The Vink reconstruction, using the propagating rift model and the same isobaths, obtained by rotating South America 58.00° counterclockwise around a pole at 47.00° N, 33.80°. Dark regions represent areas of overlap. Reproduced with permission from Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677 10 688.

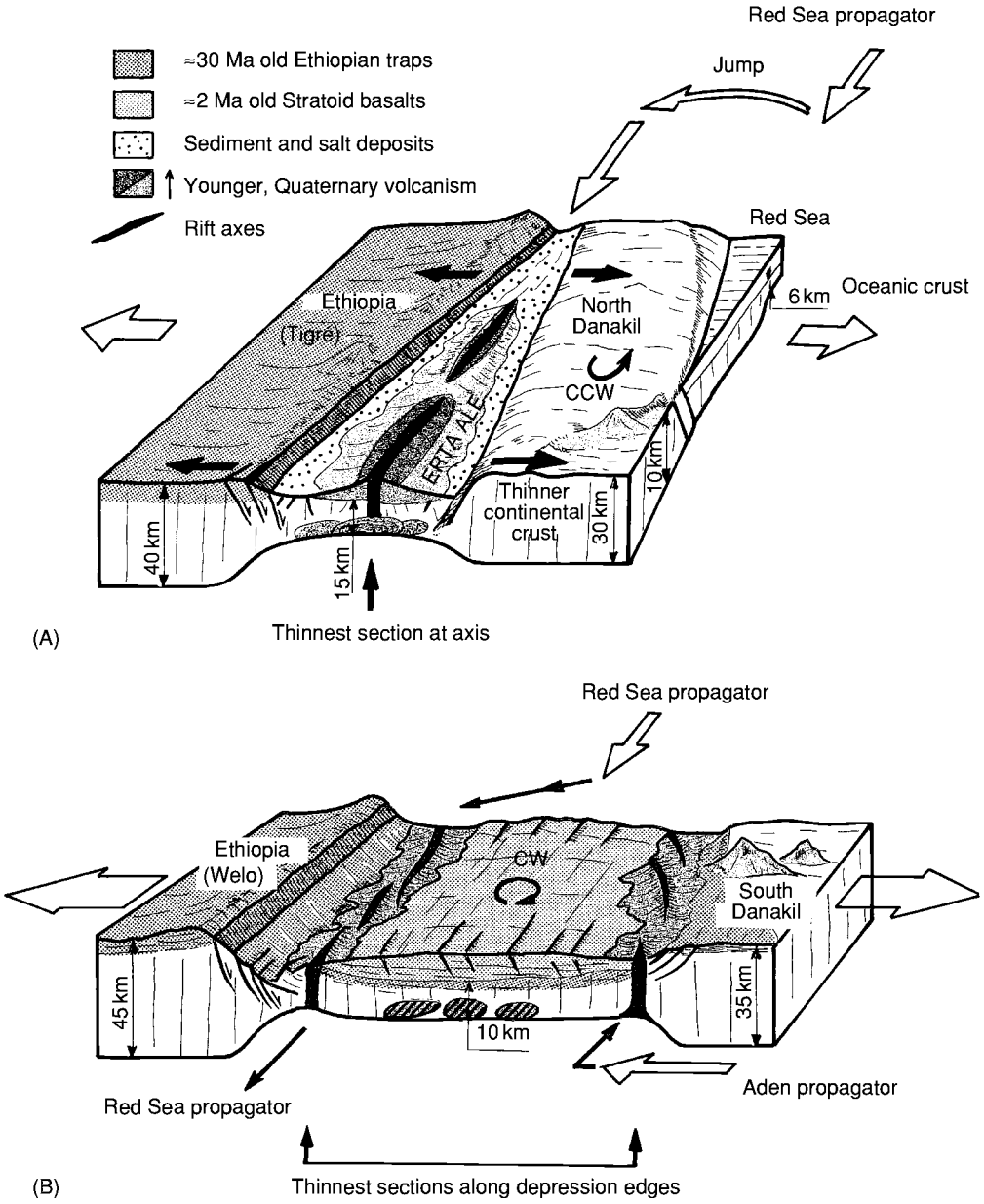


Figure 9 Schematic block diagrams of continental propagation in the (A) northern and (B) central Afar depression, Africa, viewed from the south east. The zone between the duelling Aden and Red Sea propagators is rotating clockwise (CW) and deforming by bookshelf faulting. CCW, Counter clockwise. Reproduced with permission from Manighetti I, Tapponnier P, Gillot PY, *et al.* (1998) Propagation of rifting along the Arabia Somalia plate boundary: into Afar. *Journal of Geophysical Research* 103: 4947 4974.

only a few kilometres, through oceanic propagators with offsets on the order of 10–100 km, up to offsets of several hundred kilometres at microplate tectonic scales, and several thousand kilometres at continental rifting scales, appears to be the primary mechanism by which Earth's accretional plate boundary geometry is reorganized. Although conceptually simple, the propagating-rift hypothesis has important implications for plate tectonic evolution. It explains quantitatively the existence of several classes of structures, including pseudofaults, failed rifts, and zones of transferred lithosphere, that are oblique to ridges and transform faults and thus previously seemed incompatible with plate tectonic theory. It explains why passive continental margins are not parallel to the oldest seafloor isochrons, but are instead pseudofaults, bounding lithosphere created on propagating spreading centres and indicating the direction of the continental breakup propagators. It also explains the large-scale reorganization of many seafloor-spreading systems, including both the origination and the termination of many fracture zones, as well as the

formation of some transient microplates that appear to be the modern analogues of large-scale spreading-centre jumps. This hypothesis provides a mechanistic explanation for the way in which many (if not all) spreading-centre jumps occur, why they occur in systematic patterns, and how spreading centres reorient when the direction of seafloor spreading changes. It also explains the origin of large areas of petrologically diverse seafloor, including the major abyssal ferrobasalt provinces. The common occurrence of rift propagation over a wide range of tectonic environments and spreading rates indicates that it represents an efficient mechanism of adjustment of extensional plate boundaries to the forces driving plate motions.

See Also

Geomorphology. **Plate Tectonics**. **Tectonics:** Seismic Structure At Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Bird RT and Naar DF (1994) Intratransform origins of mid ocean ridge microplates. *Geology* 22: 987 990.

Hey RN (1977) A new class of pseudofaults and their bearing on plate tectonics: A propagating rift model. *Earth and Planetary Science Letters* 37: 321 325.

Hey RN, Duennebier FK, and Morgan WJ (1980) Propa gating rifts on mid ocean ridges. *Journal of Geophysical Research* 85: 3647 3658.

Hey RN, Sinton JM, and Duennebier FK (1989) Propagat ing rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii,* pp. 161 176. Boulder, CO: Geological Society of America.

Hey RN, Johnson PD, Martinez F, *et al.* (1995) Plate bound ary reorganization at a large offset, rapidly propagating rift. *Nature* 378: 167 170.

Kleinrock MC and Hey RN (1989) Migrating transform zone and lithospheric transfer at the Galapagos 95.5° W propagator. *Journal of Geophysical Research* 94: 13 859 13 878.

Manighetti I, Tapponnier P, Gillot PY, *et al.* (1998) Propa gation of rifting along the Arabia Somalia plate bound ary: into Afar. *Journal of Geophysical Research* 103: 4947 4974.

McKenzie D and Jackson J (1986) A block model of distrib uted deformation by faulting. *Journal of the Geological Society, London* 143: 349 353.

Naar DF and Hey RN (1991) Tectonic evolution of the Easter microplate. *Journal of Geophysical Research* 96: 7961 7993.

Schouten H, Klitgord KD, and Gallo DG (1993) Edge driven microplate kinematics. *Journal of Geophysical Research* 98: 6689 6701.

Searle RC, Bird RT, Rusby RI, and Naar DF (1993) The development of two oceanic microplates: Easter and Juan Fernandez microplates, East Pacific Rise. *Journal of the Geological Society* 150: 965 976.

Tapponnier P, Armijo R, Manighetti I, and Courtillot V (1990) Bookshelf faulting and horizontal block rotations between overlapping rifts in southern Afar. *Geophysical Research Letters* 17: 1 4.

Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677 10 688.

Seismic Structure At Mid-Ocean Ridges

S M Carbotte, Columbia University, New York, NY, USA

Introduction

Oceanic crust is created at mid-ocean ridges (*see* **Tectonics:** Mid-Ocean Ridges) as mantle material upwells and undergoes pressure-release melting in response to ongoing seafloor spreading. As mantle melts rise to the surface and freeze, they form an internally stratified crust of extrusive basalts and sheeted dykes underlain by layered and massive gabbros. Spreading rate has long been recognized as a fundamental variable governing crustal accretion at ridges, with first-order differences observed in a wide range of ridge properties. However, significant changes in ridge properties are also observed along the ridge axis at any given spreading rate, which suggests that factors other than the rate of plate separation contribute to the local supply and distribution of magma from the mantle. Seismic methods permit imaging of structures within the crust that result from magmatic processes at mid-ocean ridges and provide important insights into the role of spreading rate and magma supply in crustal creation.

Since the early days of seafloor exploration, seismic studies, which rely on the propagation of sound waves through rocks, have been the primary tool used to investigate the internal structure of the oceanic crust (*see* **Seismic Surveys**). These studies reveal two primary seismic layers, which are generally believed to correspond to lithological structures in the crust: seismic layer 2 corresponds to the dykes and basaltic lava

flows that form the shallow crust, and layer 3 is associated with the massive and sheeted gabbros that form the lower crust. Seismic methods fall into two categories: reflection studies, which are based on the reflection of near-vertical seismic waves from interfaces where large contrasts in density and/or elastic properties are present, and refraction studies, which exploit the characteristics of seismic energy that travels horizontally as head waves through rock layers. Reflection methods provide continuous images of crustal horizons and permit efficient mapping of small-scale variations over large regions. Locating these horizons at their correct depths within the crust requires knowledge of the seismic velocity of crustal rocks, which is poorly constrained from reflection data. Refraction techniques provide detailed information on crustal velocity structure but typically result in relatively sparse measurements that represent large spatial averages. Hence, the types of information obtained from reflection and refraction methods are highly complementary, and these data are often collected and interpreted together.

At mid-ocean ridges, three crustal horizons are found where contrasts in elastic properties are sufficiently large that the horizons can be mapped with reflection techniques. These include seismic layer 2A (which is commonly assumed to correspond to the layer of lava flows (extrusives) that caps the oceanic crust), the shallow magma chamber from which the crust is formed, and the Moho (which marks the crust–mantle boundary). Each of these three structures and their main characteristics at mid-ocean ridges will be described here, and the implications of these observations for understanding how oceanic crust is created will be summarized. In the final section, changes in crustal structure at ridges spreading at different rates and the prevailing models to account for these variations will be described.

Seismic Layer 2A

Early Studies

Seismic layer 2A was first identified in the early 1970s from analysis of refraction data at the Reykjanes Ridge south of Iceland. This layer of low P-wave velocities (less than 3.5 km s $^{-1}$), which comprises the shallowest portion of the oceanic crust (Figure 1), was attributed to extrusive rocks with high porosities due to volcanically generated voids and extensive crustal fracturing. In the late 1980s a bright event corresponding to the base of seismic layer 2A was imaged for the first time using multichannel seismic-reflection data. This event is not a true reflection but rather a refracted arrival resulting from turning waves within a steep-velocity-gradient zone that marks the base of seismic layer 2A. Within this gradient zone, P-wave velocity rapidly increases to values typical of seismic layer 2B (more than 5.0 km s $^{-1}$) over a depth interval of about 100–300 m (Figure 1A). The 2A event is seen in the far offset traces of reflection data collected with long receiver arrays (more than

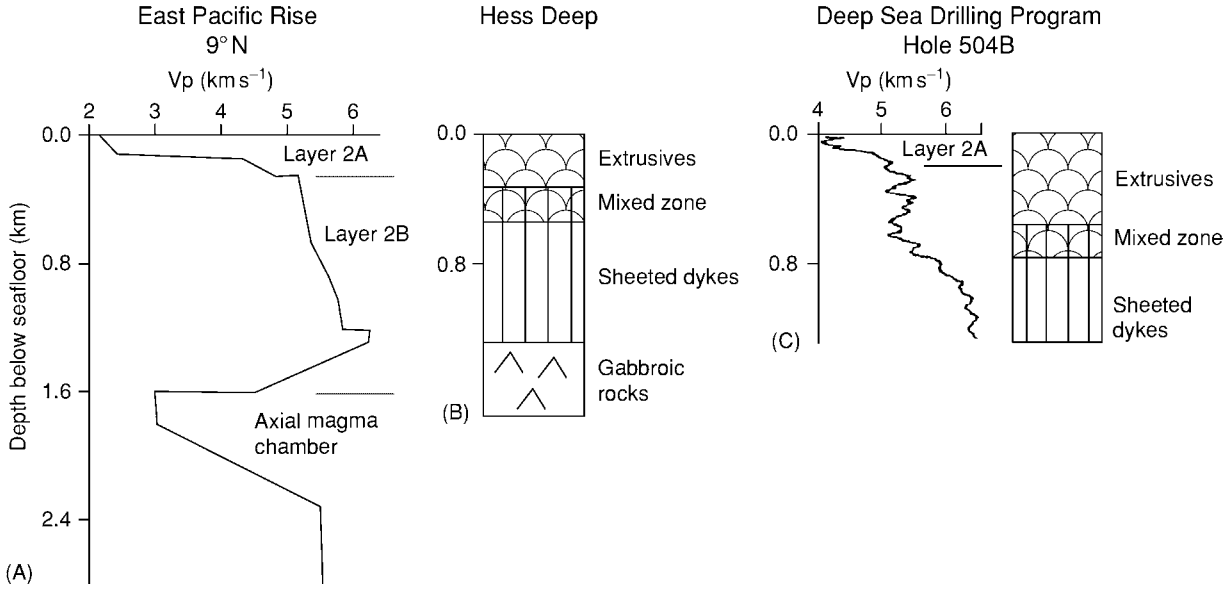


Figure 1 (A) Variations in seismic velocity with depth for newly formed crust at the East Pacific Rise. Layers 2A and 2B and the low velocities associated with the axial magma chamber are identified. (B) Lithological cross section through the upper crust at Hess Deep, derived from submersible observations. (C) Comparison of P wave velocities from *in situ* sonic logging within Deep Sea Drilling Program Hole 504B with the lithological units observed within the hole.

2 km) and can be successfully stacked, providing essentially continuous images of the base of layer 2A at mid-ocean ridges.

The Geological Significance of the Layer 2A–2B Transition: A Lithological Transition from Extrusives to Dykes or a Porosity Boundary within the Extrusives?

In most recent studies layer 2A near the ridge axis is assumed to correspond to extrusive rocks, and the base of layer 2A is assumed to correspond to a lithological transition to the sheeted-dyke section of the oceanic crust. The primary evidence for this lithological interpretation comes from studies at Hess Deep in the equatorial eastern Pacific. In this area, observations of fault exposures made from manned submersibles show that the extrusive rocks are on average 300–400 m thick, similar to the thickness of layer 2A measured near the crest of the East Pacific Rise (compare Figure 1A and B). However, the Hess Deep studies also revealed significant variability in the thickness of the extrusive layer (total range of 200–800 m) over horizontal distances of only a few hundred metres. Variability on this scale is well below the spatial resolution of seismic studies, which can provide only a smoothed view of the lithological heterogeneity that may be present.

Other researchers have suggested that the base of layer 2A may correspond to a porosity boundary within the extrusive section associated, perhaps, with a fracture front or hydrothermal alteration. This interpretation is based primarily on observations from a deep crustal hole located off the coast of Costa Rica, which was drilled as part of the Deep Sea Drilling Program (DSDP). Within this hole (504B) there is a velocity transition zone that is located entirely within the extrusive section (Figure 1C). Here, a thin high-porosity section of rubbly basalts and breccia with P-wave velocities of approximately 4.2 km s^{-1} overlies a thick lower-porosity section of extrusives with higher P-wave velocities (5.2 km s^{-1}). However, the relevance of these observations to the geological interpretation of the velocity structure of the ridge crest is questionable. Crust at DSDP hole 504B is 5.9 Ma old, and it is well established that the seismic velocity of the shallow crust increases with age owing to crustal alteration (see below). Indeed the velocities within the shallowest extrusives at DSDP 504B (*ca.* 4 km s^{-1}) are much higher than those observed at the ridge crest (2.5–3 km s^{-1}), indicating that significant crustal alteration has occurred (compare Figure 1A and C).

Conclusive evidence of the geological nature of seismic layer 2A will probably require drilling or observations of faulted exposures of the crust made where seismic observations are also available. At present, the bulk of the existing sparse information favours the lithological interpretation, and layer 2A is commonly used as a proxy for the extrusive crust. If this interpretation is correct, mapping the layer 2A–2B boundary provides direct constraints on the eruption and dyke-injection processes that form the uppermost part of the oceanic crust.

Characteristics of Layer 2A at Mid-Ocean Ridges

At the East Pacific Rise, layer 2A is typically 150–250 m thick within the region where crust is currently being formed (Figures 2 and 3). Only minor variations in the thickness of this layer are observed along the ridge crest, except near transform faults and other ridge offsets where the layer thickens.

Across the ridge axis, layer 2A approximately doubles or triples in thickness over a zone approximately 2–6 km wide, indicating extensive accumulation of extrusives within this wide region (Figures 4 and 5). This accumulation may occur through lava flows that travel up to several kilometres from their eruption sites at the axis, either over the seafloor or perhaps through subsurface lava tubes. Volcanic eruptions that occur off-axis may also contribute to building the extrusive pile.

Along the axis of the intermediate-spreading Galapagos spreading centre, located in the equatorial Pacific, layer 2A displays a wide range of thicknesses, which are closely linked with other characteristics of the ridge axis (Figure 6). Closest to the Galapagos Hotspot, a thin layer 2A is observed (150–350 m) beneath a shallow axial high, similar to that seen at fast-spreading ridges. Away from the hotspot-dominated portion of the ridge, the axial high disappears and layer 2A is thicker (*ca.* 300–600 m) and more variable along the axis. At the slow-spreading Mid-Atlantic Ridge the sparse existing data indicate that layer 2A does not systematically thicken away from the ridge axis, and it appears that the extrusive section is built largely within the floor of the median valley.

Axial Magma Chamber

Early Studies

Drawing on observations of the crustal structure of ophiolites (sections of oceanic or oceanic-like crust exposed on land) and the geochemistry of seafloor basalts, it was initially thought that mid-ocean ridges were underlain by large, essentially molten, magma reservoirs. However, prior to the 1980s, few actual constraints on the dimensions of magma chambers at ridges were available. Early seismic studies on the East Pacific Rise detected a zone of lower seismic

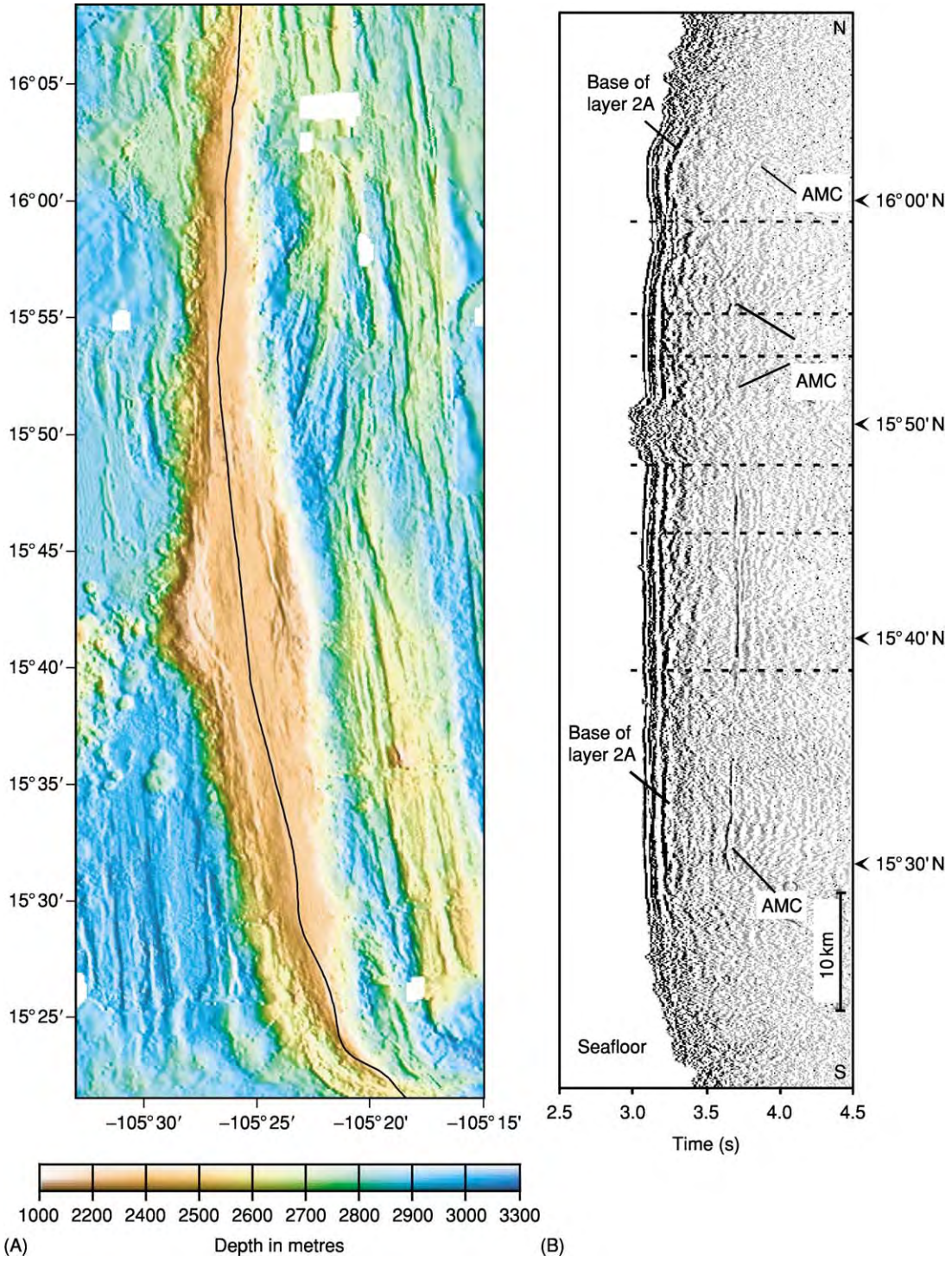


Figure 2 Example of a multichannel seismic line collected along the axis of the East Pacific Rise, showing the base of the extrusive crust (layer 2A) and the reflection from the top of the axial magma chamber (AMC). (A) The bathymetry of the ridge axis, with the location of the seismic profile indicated by the black line. (B) The dashed lines on the seismic section mark the locations of very small offsets that are observed in the narrow depression along the axis where most active volcanism is concentrated. From Carbotte SM, Ponce Correa G, and Solomon A (2000) *Journal of Geophysical Research* 105: 2737 2759.

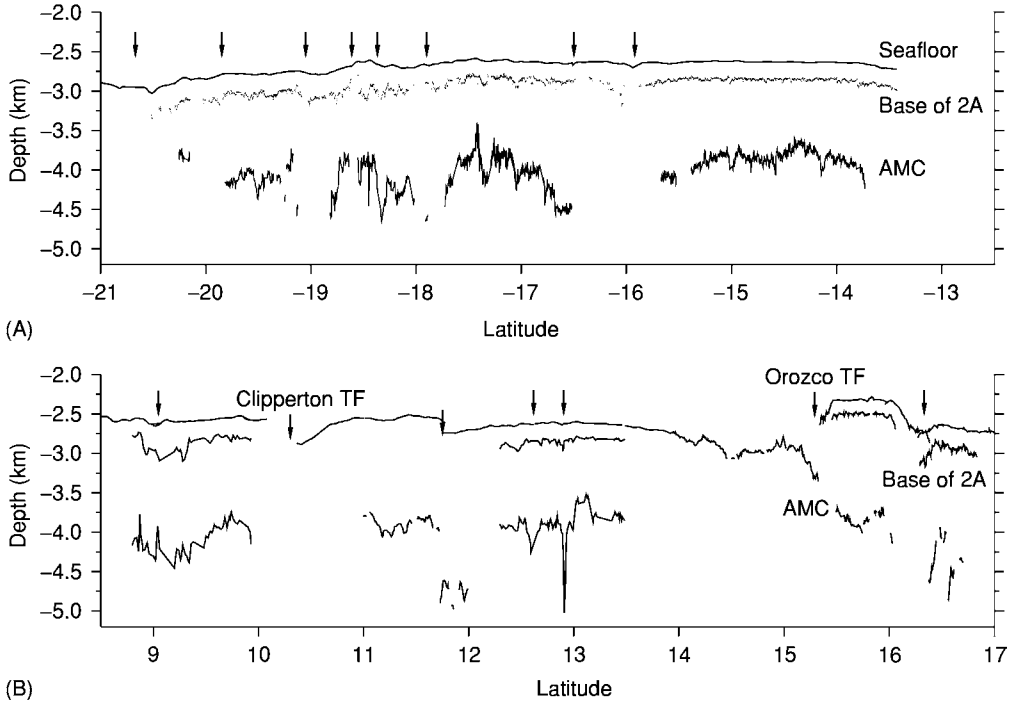


Figure 3 Cross sections along the axis of (A) the southern and (B) the northern East Pacific Rise, showing depths to the seafloor, the base of the extrusive crust, and the axial magma chamber (AMC) reflection. This compilation includes all multichannel reflection data available along this ridge. Labelled arrows show the locations of transform faults (TF). Other arrows mark the locations of smaller discontinuities of the ridge axis known as overlapping spreading centres. (Top panel) Hooft EE, Detrick RS, and Kent GM (1997) *Journal of Geophysical Research* 102: 27319 27340; (bottom panel) Kent GM., Harding AJ, and Orcutt JA (1993) *Journal of Geophysical Research* 98: 13945 13696; Detrick RS, Buhl P, Vera E, Mutter JC, Orcutt JA, Madsen J, and Brocher T (1987) *Nature* 326: 35 41; Babcock JM, Harding AJ, Kent GM, and Orcutt JA (1998) *Journal of Geophysical Research* 103: 30451 30467; Carbotte SM, Ponce Correa G, and Solomon A (2000) *Journal of Geophysical Research* 105: 2737 2759.

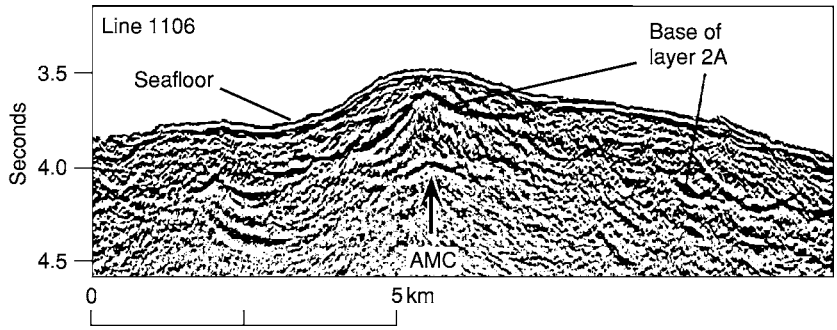


Figure 4 Example of a multichannel seismic profile shot across the ridge axis of the southern East Pacific Rise at 17°28′ S. The axial magma chamber (AMC) reflection and the event from the base of layer 2A can both be seen. From Carbotte SM, Mutter JC, and Wu L (1997) *Journal of Geophysical Research* 102: 10165 10184.

velocity beneath the ridge axis, as expected for a region containing melt. A bright reflector was also found, indicating the presence of a sharp interface with high acoustic impedance contrast within the upper crust. In the mid-1980s an extensive seismic reflection and refraction experiment was carried out on the northern East Pacific Rise by researchers from the University of Rhode Island, Lamont–Doherty Earth Observatory, and Scripps Institution of Oceanography. This study imaged a bright sub-horizontal reflector located 1–2 km below the seafloor along much of the ridge. In several locations this reflector was found to be phase reversed relative to the seafloor reflection, indicating that it resulted from an interface

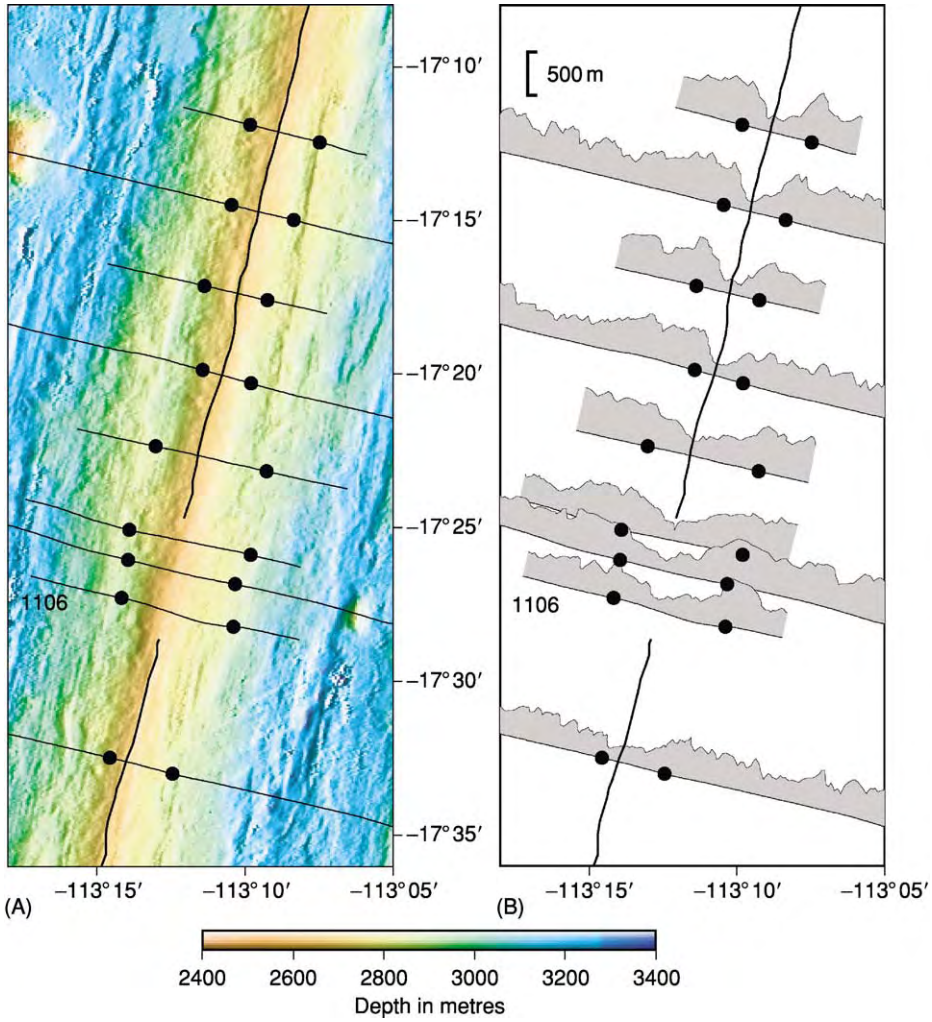


Figure 5 The thickening of the seismically inferred extrusive crust (layer 2A) across the axis of the southern East Pacific Rise. (A) Bathymetry map of the region with the locations of cross axis seismic lines shown as light lines. The bold black line shows the location of the narrow depression along the ridge axis where most volcanic activity occurs. The black dots indicate the width of the region over which the seismically inferred extrusives accumulate, as interpreted from the data shown in **Figure 5B**. (B) Thickness of the extrusive crust inferred from the seismic data along each cross axis line. Black dots mark where layer 2A reaches its maximum thickness away from the axis. Seismic line 1106 (shown in **Figure 4**) is labelled.

with an abrupt drop in seismic velocity. Based on its reversed phase and high amplitude, this event is now recognized as a reflection from a lens of magma located at the top of what is commonly referred to as an axial magma chamber.

Seismic refraction and tomography experiments show that this reflector overlies a broader region within which seismic velocities are reduced relative to normal crust (**Figure 7**). This low-velocity zone is approximately 5 km wide at shallow depths, possibly widening slightly at the base of the crust. Because of the relatively small velocity anomaly associated with much of this low-velocity zone (less than 1 km s $^{-1}$), this region is interpreted as hot largely solidified rock and crystal mush containing only a few percent partial melt.

Characteristics of the Axial Magma Chamber at Mid-Ocean Ridges

Several seismic-reflection studies have now been carried out along the fast-spreading East Pacific Rise, imaging over 1400 km of ridge crest (**Figure 3**). A reflection from the roof of the magma chamber is detected beneath about 60% of the surveyed region and can be traced continuously in places for tens of kilometres. This reflector is found at a depth of 1–2 km below the seafloor and deepens and disappears towards major offsets of the ridge axis, including transform faults and overlapping spreading centres (**Figure 3**). Most volcanic activity along the East Pacific Rise is concentrated within a narrow depression, less than 1 km wide, which is interrupted by

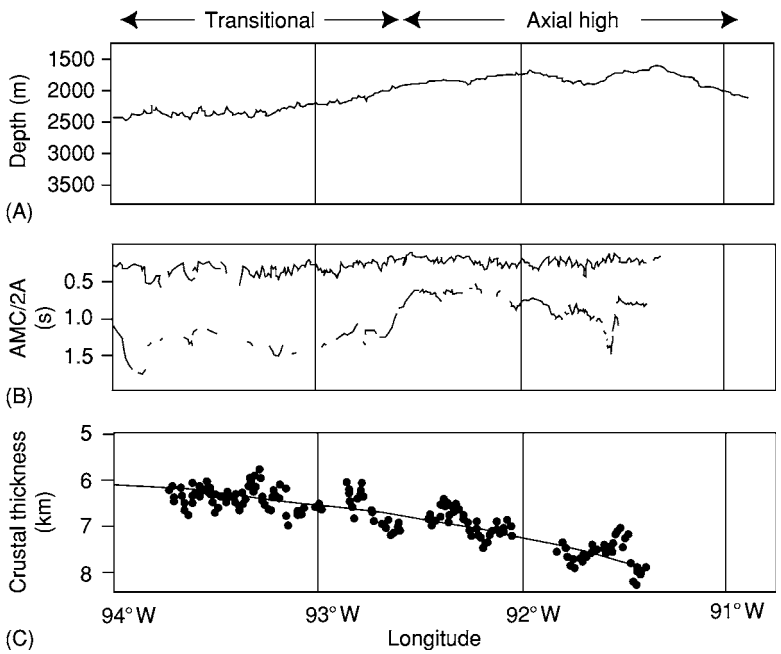


Figure 6 Crustal structure along the intermediate spreading Galapagos Spreading Centre. (A) Seafloor depth along the ridge axis. (B) Two way travel times to the base of layer 2A (top line) and the axial magma chamber (AMC) reflection (bottom line). (C) Crustal thickness derived from two way travel time to Moho on the basis of velocities derived from refraction data. Black line shows best fit polynomial to crustal thickness data. (Reproduced from Detrick RS, Sinton JM, Ito G, *et al.* (2002) Correlated geophysical, geochemical, and volcanological manifestations of plume ridge interaction along the Galápagos Spreading Centre. *Geochemistry Geophysics Geosystems* 3: 8501; DOI 10.1029/2002GC000350.)

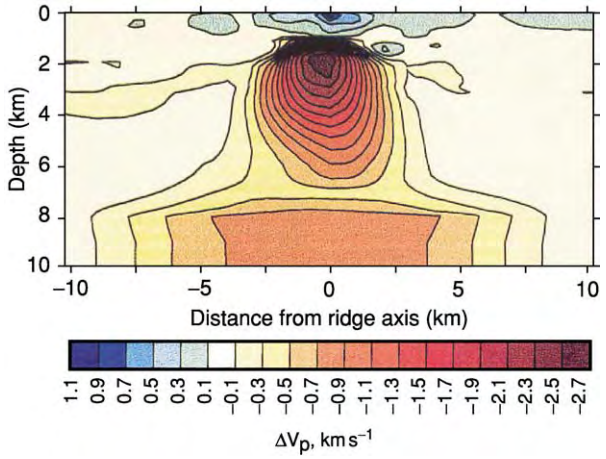


Figure 7 Seismic velocity structure of a fast spreading ridge, showing the region of low velocities associated with melt and hot rock at the ridge axis. The velocity model is derived from a tomography experiment at 9°30′ N on the East Pacific Rise. (Reproduced from Dunn RA, Toomey DR, and Solomon SC (2000) Three dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise at 9°30′ N. *Journal of Geophysical Research* 105: 23 537 23 555.)

small steps or offsets, which may be the boundaries between individual dyke swarms. In many places, the magma-chamber reflector does not disappear beneath these offsets (Figure 3). However, changes in the depth and width of the reflector are often seen. Seismic-tomography studies show that the broader region of low velocities associated with the crustal magmatic system pinches and narrows beneath these small offsets. These results suggest that segmentation of the axial magma chamber is associated with the full range of offsets observed along the ridge axis.

Migration of seismic profiles shot perpendicular to the ridge axis reveals that the magma-chamber reflection arises from a narrow feature that is typically less than 1 km wide (total range from 200 m to 4 km; Figure 4). Refraction data and waveform studies of the magma-chamber reflection suggest that it arises from a thin body of magma a few hundred to perhaps a few tens of metres thick, leading to the notion of a magma lens or sill. Initial studies assumed that this lens contained pure melt. However, recent research suggests that much of the magma lens may have a significant crystal content (more than 25%), with regions of pure melt limited to pockets only a few kilometres or less in length along the axis.

Magma-lens reflections similar to those imaged beneath the East Pacific Rise have been imaged along intermediate-spreading ridges, including the Galapagos Spreading Centre (Figure 6), south-east Indian Ridge, and Juan de Fuca Ridge, and at the back-arc spreading centre in the Lau Basin. In these areas,

magma-lens reflectors are typically found deeper in the crust, at depths of 2.5–3 km, although shallower reflections are observed in a few places.

Along the slow-spreading Mid-Atlantic Ridge, evidence for magma lenses has been found along the Reykjanes Ridge. Here, an intracrustal reflection at a depth of approximately 2.5 km is observed, which is similar to the depths of magma-lens events observed beneath the intermediate-spreading ridges. Seismic data have been collected elsewhere along the Mid-Atlantic Ridge with little evidence of magma-lens reflections, possibly owing to imaging problems associated with the rough seafloor topography typical of this ridge. However, there is evidence from refraction data and seismicity studies that large crustal magma bodies are not common beneath this ridge. Microearthquake data show that earthquakes can occur to depths of 8 km beneath the Mid-Atlantic Ridge, indicating that the entire crustal section is sufficiently cool for brittle failure. In other areas, slightly reduced velocities within the crust have been identified, indicating warmer temperatures and possibly the presence of small pockets of melt within the crust.

The prevailing model for magma chambers beneath ridges (Figure 8) incorporates both the geophysical constraints on chamber dimensions described above and geochemical constraints on magma-chamber processes. At fast-spreading ridges (Figure 8A), the magma chamber consists of a narrow and thin melt-rich magma lens, which overlies a broader crystal-mush zone and a surrounding region of hot but solidified rock. The dyke-injection events and volcanic eruptions that build the upper crust are assumed to tap the magma lens. The lower crust is formed from the crystal residuum within the magma lens and from the broader crystal-mush zone. Observations of ophiolites suggest that the injection of sills that tap magma directly from the upper mantle also contributes to the lower crustal section. Both seismic and seafloor compliance studies from the 9°–10° N region of the East Pacific Rise indicate melt accumulation at the base of the crust, within either melt-rich sills or a broader partially molten zone.

At slow-spreading ridges (Figure 8B) a short-lived dyke-like crystal-mush zone without a steady-state magma lens is envisioned. At these ridges volcanic eruptions occur and the crystal-mush zone is replenished during periodic magma-injection events from the mantle. Observations of seafloor fault exposures of crust created at slow-spreading ridges reveal a heterogeneous crustal section, where altered and deformed lower-crustal and upper-mantle rocks are unconformably overlain by lavas in some locations. Crustal accretion at these ridges is inferred to be a highly episodic process, with the internal structure of the crust being strongly disrupted by faulting.

Moho

The base of the crust is marked by the Mohorovicic Discontinuity, or 'Moho', where P-wave velocities increase from values typical of lower-crustal rocks (7–7.5 km s^{-1}) to mantle velocities (more than 8.0 km s^{-1}). The change in P-wave velocity is often sufficiently abrupt that a sub-horizontal Moho reflection is observed in reflection data, from which the base of the crust can be mapped. Depths to the Moho derived from seismic-refraction studies provide our best estimates of crustal thickness and are used to study how total crustal production varies in different ridge settings.

Characteristics of the Moho at Mid-Ocean Ridges

Reflection Moho is imaged in much of the data collected at the East Pacific Rise (Figure 9). It can often be traced beneath the region of lower-crustal velocities found at the ridge and occasionally beneath the magma-lens reflection itself. Where information on crustal velocities is available, average crustal thicknesses of 6–7 km are measured. There is no evidence for thickening away from the ridge crest, indicating that the crust acquires its full thickness within a narrow zone at the axis.

Unlike at the fast- and intermediate-spreading ridges, at the slow-spreading Mid-Atlantic Ridge the Moho is difficult to identify in seismic-reflection data, possibly owing to poor imaging conditions or a difference in the geological nature of this boundary. At this ridge, information on crustal thickness and variations in crustal structure is derived primarily from seismic-refraction studies. Average crustal thicknesses are similar to those observed at fast-spreading ridges (6–7 km). However, significant crustal thinning is observed (by 1–4 km) towards transform faults and smaller ridge offsets (Figure 10B). These results are interpreted as reflecting a three-dimensional pattern of mantle upwelling or melt migration to the ridge, resulting in greater crustal production within the central regions of ridge segments away from ridge offsets.

Variations in crustal thickness within ridge segments are also observed along the East Pacific Rise (Figure 10A). However, in the region with the best data constraints (9°–10° N), the spatial relationships are opposite to those observed at the Mid-Atlantic Ridge. Here, the crust is approximately 1 km thinner, not thicker, in the portion of the segment where a range of ridge-crest observations indicate that active

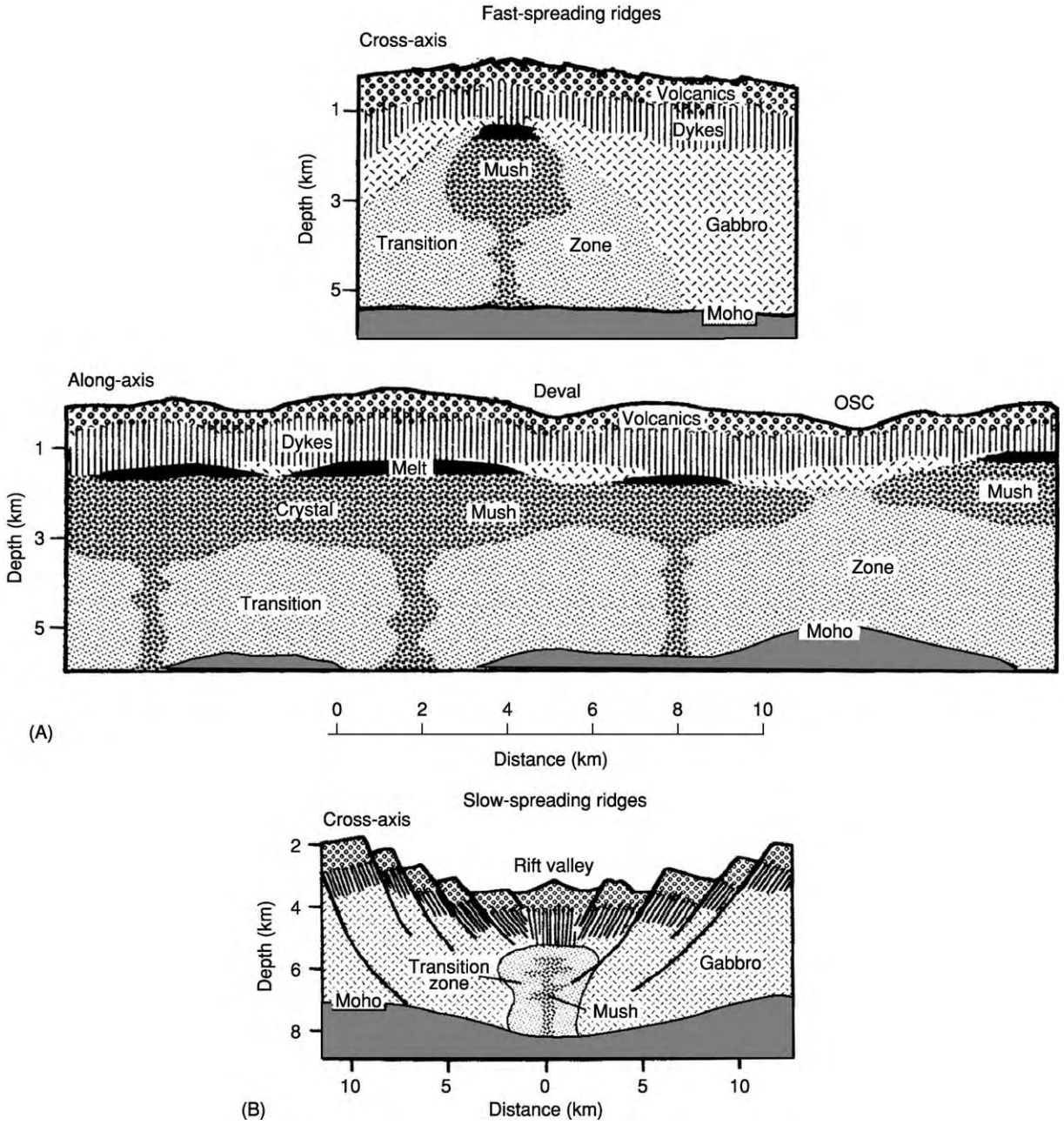


Figure 8 Schematic representation of the axial magma chamber beneath (A) fast and (B) slow spreading ridges. (A) At a fast spreading ridge a thin zone of predominantly melt (black region) is located 1–2 km below the seafloor, grading downwards into a partially solidified crystal mush zone. This region is in turn surrounded by a transition zone of solidified but hot rock. Along the ridge axis, the 'melt' sill and crystal mush zone narrows and may disappear at the locations of ridge discontinuities (labeled Deval and OSC in the along axis profile). (B) At a slow spreading ridge a steady state melt region is not present. Magma is periodically injected into the crust from the mantle with volcanic eruptions and the emplacement of small intrusive bodies, which crystallize to form the oceanic crust. (Reproduced from Sinton JA and Detrick RS (1992) Mid ocean ridge magma chambers. *Journal of Geophysical Research* 97: 197–216.)

crustal accretion is focused. At this fast-spreading ridge the presence of a steady-state magma chamber and a broad region of hot rock (Figure 8) may permit efficient redistribution of magma away from regions of focused delivery from the mantle. The absence of a steady-state magma chamber beneath the slow-spreading Mid-Atlantic Ridge may prohibit significant along-axis transport of magma, such that, at this ridge, thicker crust accumulates at the site of focused melt delivery.

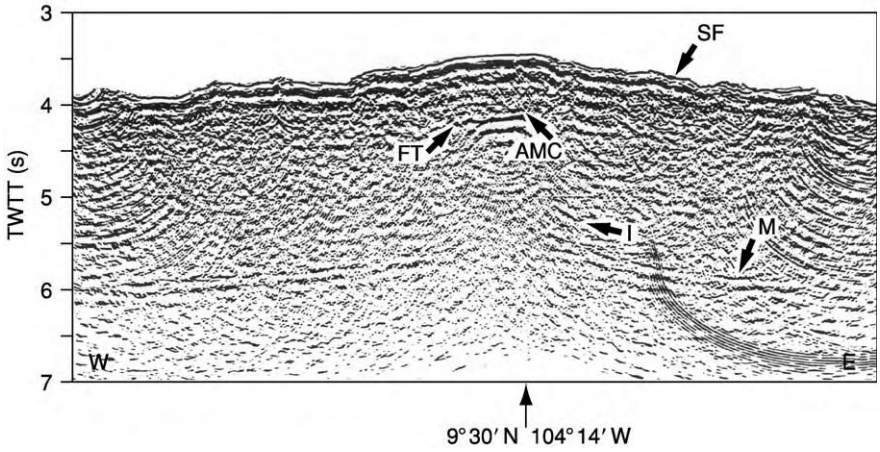


Figure 9 Multichannel seismic line crossing the East Pacific Rise at 9°30′ N, showing the Moho reflection (M), the seafloor (SF), the axial magma chamber (AMC) reflection, and other intracrustal reflections (FT, I). (Reproduced from Barth GA and Mutter JC (1996) Variability in oceanic crustal thickness and structure: multichannel seismic reflection results from the northern East Pacific Rise. *Journal of Geophysical Research* 101: 17 951 17 975.)

Variations in Mid-Ocean Ridge Structure

The seismic observations described in the previous sections reveal significant differences in the internal structure of the crust at fast- and slow-spreading ridges as well as within individual spreading segments, with important implications for crustal creation at mid-ocean ridges. Large gradients in crustal structure are observed at slow-spreading ridges, with crustal thickness often varying by a factor of two within individual ridge segments (Figure 10B). In comparison, at fast-spreading ridges only minor variations in crustal structure are observed within ridge segments, and crustal accretion appears to be a more uniform process at these rates. Indeed, throughout the fast-spreading-rate range (85–150 mm year ^{1}), average crustal structure is remarkably constant, with magma lenses imaged beneath much of the axis at similar widths and depths (Figure 11). In contrast, at slow-spreading ridges, steady-state magma bodies within the crust are rarely detected. Intermediate-spreading ridges (less than 85 mm year ^{1}) are of particular interest because they display characteristics from across the spreading-rate spectrum, the distribution of which appears to be closely linked with spatial variations in the supply of magma to the ridge. Where the ridge forms a shallow axial high, magma bodies have been observed at the shallow depths (less than 2 km) characteristic of fast-spreading ridges. Beneath the shallowly rifted sections that are more typical of intermediate-spreading ridges, magma bodies lie at a deeper level within the crust of 2.5–3 km (Figure 11).

At fast-spreading ridges, the shallowest crust, defined by seismic layer 2A, is uniformly thin along the ridge axis (*ca.* 200 m; Figure 12) and commonly thickens over a region several kilometres wide about the axis. In comparison, at slow-spreading ridges, the sparse available data suggest that a thicker layer 2A is developed along the axis and that full accretion of this layer occurs within a narrow region confined to the axial valley. Along some sections of intermediate-spreading ridges, layer 2A thickens away from the axis, as observed at the fast-spreading ridges, whereas in other regions, this layer appears to acquire its full thickness within the innermost axial zone. Assuming that layer 2A corresponds to the extrusive section, these differences in the accumulation of this layer could reflect differences in eruption parameters such as eruptive volumes, lava-flow viscosity and morphology, and the dominance of fissure versus point-source eruptions. Where a wide zone of extrusive-layer thickening is observed along fast- and portions of intermediate-spreading ridges, low-viscosity lobate and sheet flows may predominate, forming thin flows that travel for significant distances from eruptive fissures at the axis. Large-volume pillow-flow eruptions and eruptions that quickly localize at point sources forming local volcanic constructions may be more common at the intermediate- and slow-spreading ridges, where little thickening of the extrusive layer away from the axis is inferred from the seismic data. The bounding faults of the axial valleys typically present at these ridges may serve to dam any far-travelling lobate and sheet flows, giving rise to full accumulation of the extrusives at the axis.

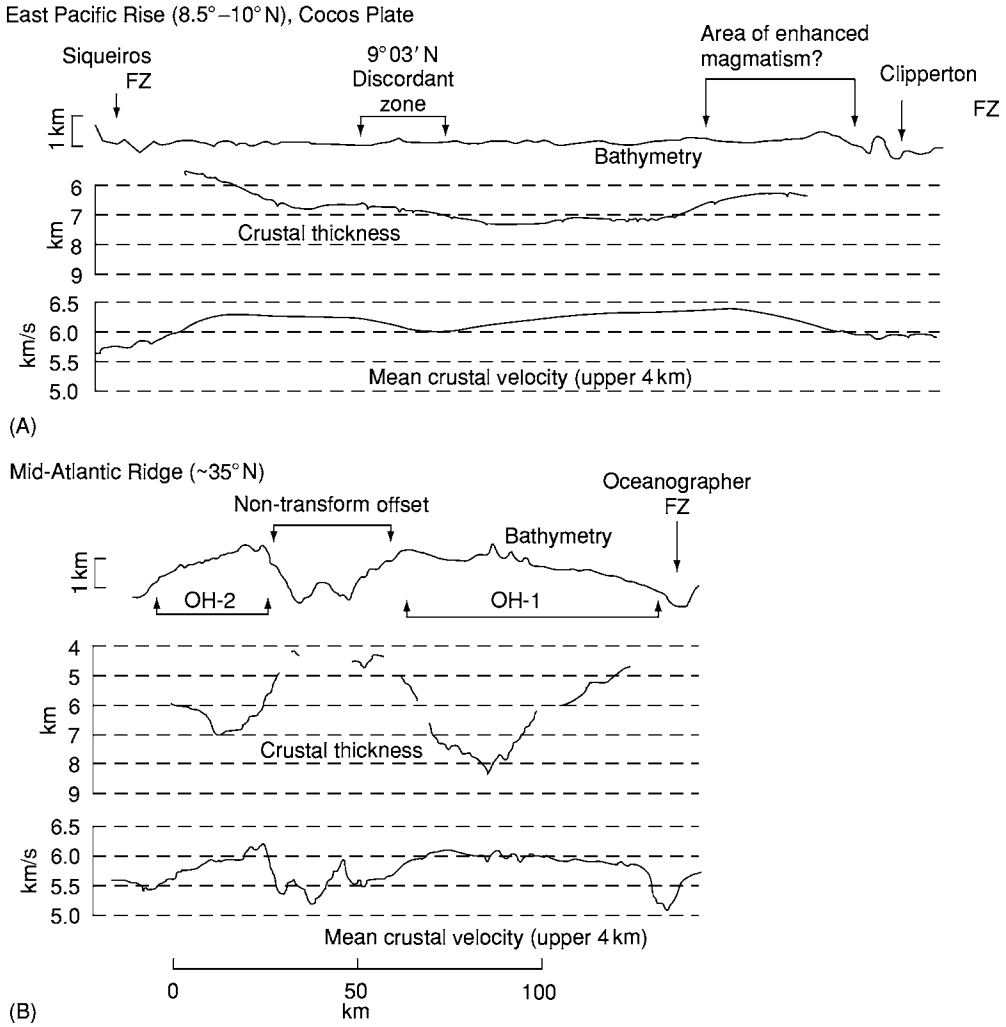


Figure 10 Comparison of crustal structures at (A) fast and (B) slow spreading ridges, showing along axis variations in crustal thickness and the mean seismic velocity of the upper crust. (B) At slow spreading ridges, the crust thins towards fracture zones and non transform offsets. (A) Changes in crustal thickness towards discontinuities are more modest at fast spreading rates. (Reproduced from Canales JP, Detrick RS, Toomey DR, and Wilcock WSD (2002) Segment scale variations in crustal structure of 150 to 300 ky old fast spreading oceanic crust (East Pacific Rise, 8°15′ N 10°15′ N) from wide angle seismic refraction profiles. *Geophysical Journal International* 152: 766 794.)

Although first-order differences are observed in a wide range of ridge properties with differences in spreading rate, several aspects of the seismic structure of ridges are surprisingly similar at all rates. The average thickness of the extrusive layer away from the ridge axis is comparable (*ca.* 350–650 m), and the total volume of extrusives produced by seafloor spreading may be largely independent of spreading rate. Average crustal thickness is also similar (6–7 km) across almost the entire spreading-rate range, and total crustal production does not depend on spreading rate except at the slowest rates (less than 15 mm year ^{1}; Figure 13). Below rates of 15 mm year ^{1}, the crust is thinner (2–4 km) and more variable in thickness, possibly because enhanced conductive heat loss in the uppermost mantle results in reduced melting.

What Controls the Depth at Which Magma Chambers Reside at Ridges?

Two main hypotheses have been put forward to explain the depths at which magma chambers are found at ridges. One hypothesis is based on the concept of a level of neutral buoyancy for magma within the oceanic crust. This model predicts that magma will rise until it reaches a level where the density of the surrounding country rock equals that of the magma. However, magma lenses at ridges lie at considerably greater depths than the neutral-buoyancy level

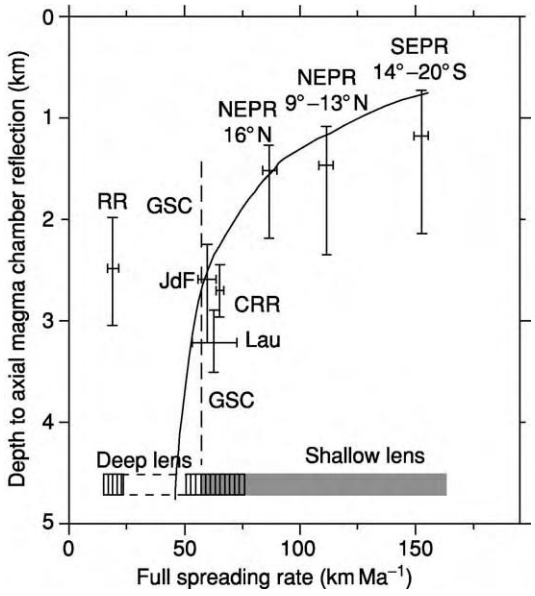


Figure 11 Average depth of magma lens reflections beneath ridges versus spreading rate. Magma lenses lie within two distinct depth ranges of 1–2 km for fast spreading ridges and 2.5–4 km for intermediate and slow spreading ridges. Both shallow and deep lenses are observed at some intermediate spreading ridges. The curved line shows the depth to the 1200°C isotherm calculated from the ridge thermal model of Phipps Morgan J and Chen YJ. Data from different ridges are labelled: RR, Reykjanes Ridge; JdF, Juan de Fuca Ridge; GSC, Galápagos Spreading Centre; CRR, Costa Rica Rift; Lau, Lau Basin; NEPR; northern East Pacific Rise; SEPR, southern East Pacific Rise.

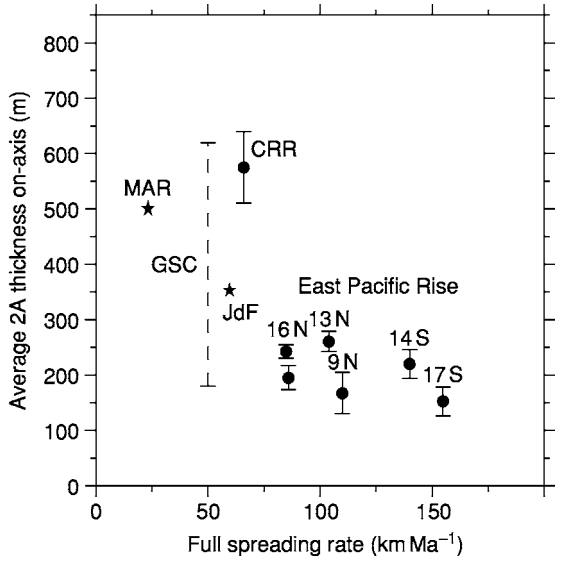


Figure 12 Thickness of the extrusive crust at the ridge axis versus spreading rate. For data obtained from detailed reflection surveys, average thicknesses are shown by black dots with standard deviations where available (solid lines) or thickness ranges (dotted lines). Data derived from other seismic methods are shown by stars. Data for the East Pacific Rise are labelled by survey location. CRR, Costa Rica Rift; MAR, Mid Atlantic Ridge; JdF, Juan de Fuca Ridge; GSC, Galápagos Spreading Centre.

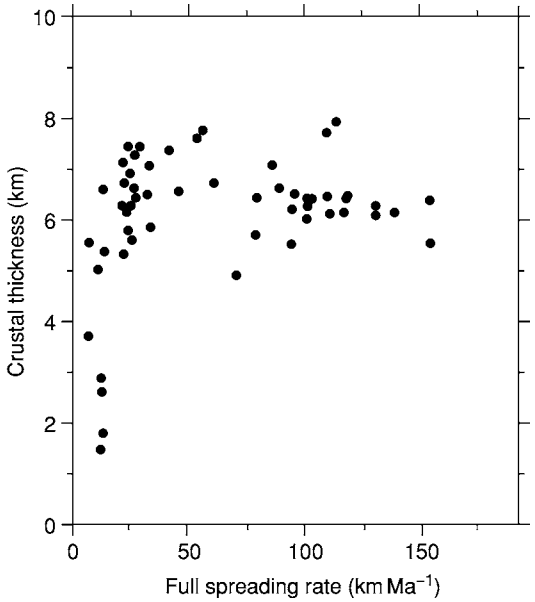


Figure 13 Crustal thickness versus spreading rate. Crustal thicknesses are determined from seismic data obtained away from fracture zones. (Reproduced from Bown JW and White RS (1994) Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth and Planetary Science Letters* 121: 435–449.)

predicted if the density of the magma is equivalent to that of lavas erupted onto the seafloor (2700 kg m^{-3}). Either the average density of magma is greater or mechanisms other than neutral buoyancy control magma-lens depth.

The alternative model hypothesizes that magma-chamber depth is controlled by the thermal structure of the ridge axis. In this model, a mechanical boundary, such as a freezing horizon or the brittle–ductile transition, prevents magma from rising to its level of neutral buoyancy. The depth of this boundary within the crust will be primarily controlled by the thermal structure of the ridge axis, which is expected to vary with spreading rate. The inverse relation between spreading rate and depth to low-velocity zones at ridges apparent in early seismic datasets provided compelling support for this hypothesis. Numerical models of ridge thermal structure have been developed that predict systematic changes in the depth to the 1200° C isotherm (a proxy for basaltic melts) with spreading rate that match the first-order depth trends for magma lenses (Figure 11). This model predicts a minor increase in lens depth within the fast-spreading-rate range and an abrupt transition to deeper lenses at intermediate spreading rates, consistent with the present dataset. The numerical models also predict that, at intermediate spreading rates, small variations in magma supply to the ridge can give rise to large changes in axial thermal structure. These models are supported by recent

observations of the Galapagos spreading centre and the Southeast Indian Ridge. At these ridges, abrupt steps in the depths of crustal magma bodies occur where the ridge axis changes from an axial high to a shallowly rifted valley, although differences in crustal thickness (a proxy for magma supply) are modest (e.g. Figure 6). These recent investigations of intermediate-spreading ridges highlight the important role of spatial variations in magma supply independent of spreading rate in the process of crustal accretion at mid-ocean ridges.

See Also

Earth: Mantle; Crust. **Igneous Processes**. **Plate Tectonics**. **Seismic Surveys**. **Tectonics:** Mid-Ocean Ridges; Propagating Rifts and Microplates At Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Bown JW and White RS (1994) Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth and Planetary Science Letters* 121: 435 449.

Buck WR, Delaney PT, Karson JA, and Lagabrielle Y (eds.) (1998) *Faulting and Magmatism at Mid Ocean Ridges*. Geophysical Monograph 106. Washington, DC: American Geophysical Union.

Hooft EE and Detrick RS (1993) The role of density in the accumulation of basaltic melts at mid ocean ridges. *Geophysical Research Letters* 20: 423 426.

Jacobson RS (1992) Impact of crustal evolution on changes of the seismic properties of the uppermost oceanic crust. *Reviews of Geophysics* 30: 23 42.

Karson JA and Christeson G (2003) Comparison of geological and seismic structure of uppermost fast spread oceanic crust: insights from a crustal cross section at the Hess Deep Rift. In: Goff J and Holliger K (eds.) *Heterogeneity in the Crust and Upper Mantle: Nature, Scaling and Seismic Properties*, pp. 99 129. New York: Kluwer Academic.

Kent GM, Singh SC, Harding AJ, *et al.* (2000) Evidence from three dimensional seismic reflectivity images for enhanced melt supply beneath mid ocean ridge discontinuities. *Nature* 406: 614 618.

Phipps Morgan J and Chen YJ (1993) The genesis of oceanic crust: magma injection, hydrothermal circulation, and crustal flow. *Journal of Geophysical Research* 98: 6283 6297.

Purdy GM, Kong LSL, Christeson GL, and Solomon SC (1992) Relationship between spreading rate and the seismic structure of mid ocean ridges. *Nature* 355: 815 817.

Sinton JA and Detrick RS (1992) Mid ocean ridge magma chambers. *Journal of Geophysical Research* 97: 197 216.

Solomon SC and Toomey DR (1992) The structure of mid ocean ridges. *Annual Review of Earth and Planetary Science* 20: 329 364.

Mountain Building and Orogeny

M Searle, Oxford University, Oxford, UK

Introduction

The term orogeny, derived from the Greek word 'oros', meaning mountain, and 'genesis', meaning birth or origin, encompasses all the processes of mountain building. Orogenic belts generally occur along plate margins and are characterized by thickened crust, metamorphism, magmatism, flexure of the lithosphere, and large-scale crustal deformation. The average thickness of the oceanic crust is about 5 km, and that of continental crust around 35 km. In the Himalaya–Tibet region the crust has reached 75–80 km thick, double the normal thickness. High topography along mountain belts usually accords with crustal and lithospheric thickening as a result of plate collision.

Most mountain ranges are the result of plate collision processes at convergent plate margins. Oceanic crust is dominantly composed of basaltic or gabbroic rocks, and the oceanic lithosphere is relatively strong and dense. Oceanic crust, composed mainly of olivine and pyroxene, can be subducted along Benioff zones and recycled back into the mantle. Continental crust is composed dominantly of granites, gneisses, and upper crustal sedimentary rocks. The quartz and feldspar-rich continental crust is weaker and more buoyant, and cannot easily subduct into the denser mantle.

Following the advent of plate tectonic theory in the 1960s it was proposed by J T Wilson that the process of orogeny was a 'cycle' beginning with rifting of continents and development of passive 'Atlantic-type' continental margins, followed by seafloor spreading and ocean basin formation, and ending with subduction, ocean closure, and finally, continental collision. This process became known as the Wilson cycle.

Mountain belts can be broadly categorized into three types of plate collision zones. These occur as the result of the collision between two oceanic plates

(e.g., Mariana–Philippine arc or the Caribbean island arc), secondly between a continental plate and an oceanic plate (e.g., Andes, or North American Cordillera), and thirdly between two continental plates (e.g., Alpine–Himalayan belt). Mountain belts can also form under the oceans, for example along the Mid-Atlantic ridge, or above hot spots, such as Hawaii or the Canary Islands, or in oceanic plateaux. Other mountain belts can form in the middle of continents, for example along the East African rift, where the mantle has domed up beneath linear rift valleys, as the continental plate begins to split apart.

Oceanic Island Arc Belts

Oceanic crust is generated along mid-ocean ridges, and becomes progressively older, colder, and denser with increasing distance from the ridge axis. When two oceanic plates converge, one plate flexes and bends beneath the other and begins to subduct. The subduction zone is marked by a deep ocean trench, an inclined zone of deep earthquakes and low heat flow (Wadati-Benioff zone), along which isotherms are buckled down. Subduction of slivers of sedimentary and basaltic rocks along the subduction zone can result in high-pressure metamorphism typical of the blueschist and eclogite facies. Typical examples of subduction zone plate boundaries are the Mariana trench between the Pacific and Philippine plates, or the Tonga–Kermadec trench between the Pacific and Indo-Australian plates. In both cases the mountain belt, comprising the fore arc and island arc complex is largely beneath sea-level. Andesitic island arcs are generated above subduction zones and are composed of explosive calc-alkaline volcanoes. Island arcs can produce small mountain ranges such as those along several West Pacific plate margins, or larger mountain ranges such as the Java–Banda arc in Indonesia. Marginal oceanic basins may open up behind the arc, forming back-arc spreading centres. Along the fore-arc region accretionary prisms of stacked thrust sheets can form as a result of scraping off sediments from the down-going plate. Subduction-related mélanges, including sepentinite mélanges with high-pressure rock clasts, are also typical of fore-arc regions. Ancient examples include the Franciscan complex in California.

Mature island arcs can produce significant mountain ranges, such as in Japan, where andesitic volcanoes are formed above a deep subduction zone. The Japan crust also consists of continental rocks, paired metamorphic belts, typical of fore-arc regions, and intrusive granitic magmas derived from partial melting of the down-going slab. The Japan Sea is a narrow ocean basin between the active margin in Japan and the passive continental margin along the east coast of mainland Asia.

Ophiolites and Mountain Building

Ophiolites are allochthonous sequences of oceanic crust and upper mantle emplaced onto continental margins. Some ophiolites are preserved as highly deformed slivers of serpentinized peridotites along ancient suture zones, or sites of plate collision. Some of the best preserved ophiolites occur as obducted thrust sheets emplaced over passive continental margins, typified by the Semail ophiolite in Oman, the Troodos ophiolite in Cyprus, and the Bay of Islands ophiolite in Newfoundland. In each of these cases a passive continental margin converged with an oceanic plate, which, instead of subducting downwards into the mantle, was detached and obducted on top of the continent. The subduction zone dipped away from the continental margin and the ophiolite complex was one of a series of thin-skinned thrust sheets emplaced onto the depressed continental margin.

In the Oman mountains of eastern Arabia, the ophiolite forms a thrust sheet over 15 km thick, approximately 700 km long, and over 100 km wide . It was formed at a spreading centre above an active subduction zone, along which old, cold oceanic basalts were subducted to depths of about 40 km, metamorphosed to garnet + clinopyroxene amphibolites, and accreted to the base of the mantle sequence peridotite. During the latest stages of ophiolite emplacement and mountain building the thinned leading edge of the passive continental margin was dragged down the subduction zone to depths of around 80 km or more and metamorphosed to eclogite facies. When the continental margin finally choked the subduction zone due to the buoyancy of the continental rocks, the deeply buried eclogites were exhumed rapidly back to the surface along the same subduction zone. The mountain building phase in Oman lasted approximately 25 My during the Late Cretaceous, when approximately 300–400 km width of the Tethyan ocean was closed.

Ophiolite obduction was the earliest phase of mountain building in the Alps and the Himalayas. Later continent–continent collision frequently overprints the earlier phases of orogeny, so mountain belts such as Oman are extremely important for deciphering subduction–obduction processes in the mountain-building cycle. A modern day example of a similar tectonic setting can be seen along the North Australian continental margin, which is currently subducting northwards beneath the accretionary prism of Timor island and the Banda volcanic arc of eastern Indonesia.

Andean-Type Mountain Building

The Andes mountains of South America extend for over 5000 km from Venezuela to southern Chile, and have formed as a result of the collision between the South American continental plate and the oceanic lithosphere of the Nazca and Pacific plates (Figure 1). The oceanic trench, the surficial trace of the subduction zone, lies about 70 km offshore the South American continental margin and reaches a depth of about 7 km off northern Chile. Earthquakes define a slab-like subduction zone inclined at 30°–40° eastwards beneath the Bolivian Andes, although to the south, the subducted lithosphere is almost horizontal beneath southern Peru and northern Chile before descending steeply into the mantle.

The geology of the Andes is dominated by linear belts of subduction-related granite–granodiorite batholiths and calc-alkaline volcanoes with shallow basins filled with continental red-beds. It has been estimated that around 300 km of crustal shortening has occurred in the central Andes as a result of east–west compression. Crustal shortening has resulted in increased crustal

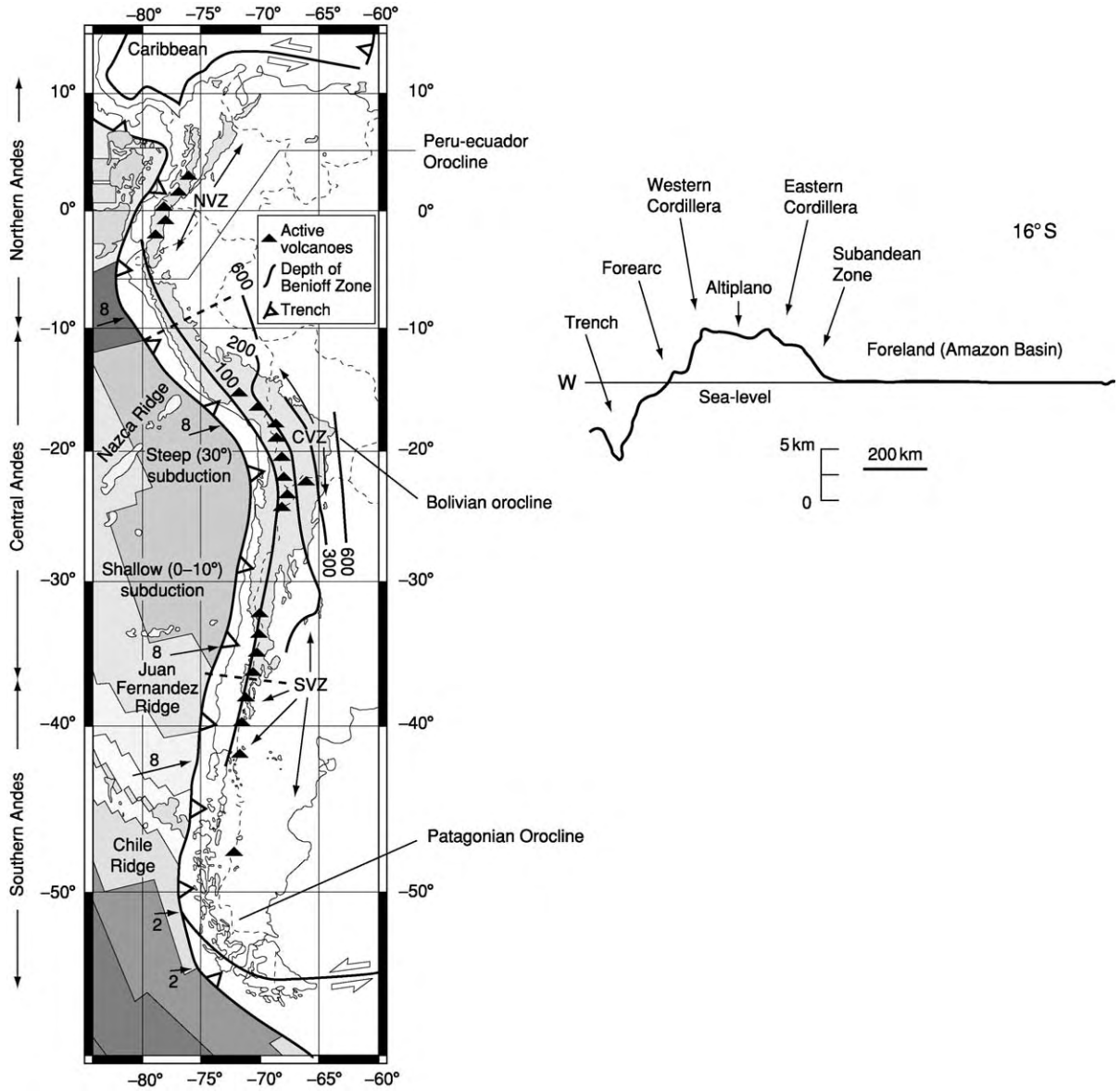


Figure 1 Map of the Andes in South America and topographic profile across the Central Andes showing the main tectonic features. Oceanic lithosphere is being subducted beneath the trench, and the oceanic crust is shaded according to age. Location of active volcanoes and the depth of the Benioff zone beneath the Andes is also shown. After Hancock PL and Skinner BJ (2000) *Oxford Companion to the Earth*, p. 21. © Oxford University Press.

thickening, ranging from ~35 km thickness east of the Andes, up to 80 km beneath the Bolivian Altiplano. The Altiplano has an average elevation of 4000 m above sea-level, and both the elevation and width of the Andes decrease both to the south and north. Along the eastern margin of the Bolivian Altiplano, a fold and thrust belt is present where relatively thin-skinned thrust sheets verge eastwards onto the Amazon foreland.

The North American Cordillera and Rocky Mountains have many geological features similar to those of the Andes, but also show important strike-slip, or transcurrent, faults like the San Andreas fault, movement along which has resulted in large-scale horizontal motion of crustal plates. Oceanic subduction beneath the continental margin has resulted in accretion of subduction-related deformed ophiolites, high-pressure metamorphic rocks, and mélanges. Granite batholiths, similar to the Andean batholiths, are aligned along mountain ranges like the Sierra Nevada, parallel to the continental margin. Andesitic volcanoes like the recently active Mount St Helens are regularly spaced along the range, along a linear belt roughly 200 km east of the trench.

Alpine-Type Mountain Building

The Alpine–Himalayan mountain belt formed as a result of the collision of the Africa–Arabia and Indian plates in the south with the Eurasian plate in the north. The Tethyan ocean, which separated these Gondwana continental fragments to the south from the Laurasian continent to the north, was mainly an east–west-aligned Permian and Mesozoic ocean that closed during the Late Cretaceous. The zone of collision stretches from the European Alps east along the Zagros mountains of Iran to the Himalaya. There are significant differences between the Alps, the Zagros, and the Himalayan mountain ranges, indicating that continental collision processes vary significantly along strike. The Oman mountains in Arabia show the early ophiolite obduction stage of the collision process, prior to continental collision. The Zagros mountains of Iran represent the initial continent–continent collision stage, and the Himalaya–Tibet orogeny represents the final stage of the continental collision process.

Seismic reflection and refraction profiles have revealed the deep crustal structure of the Alps. Together with surface geology, it has become possible to construct a geological section across the Alps (Figure 2). The Alpine orogeny includes a Cretaceous event and a Tertiary event, both of which involved subduction of crustal rocks down to eclogite facies depths and subsequent exhumation. Remnant ophiolites, marking the Tethyan suture zone, are preserved in the Valais region, and along the Ivrea zone. The Zermatt-Saas ophiolites have been subducted to over 80 km depth, metamorphosed to eclogite facies, and then exhumed rapidly during the early stage of Alpine orogeny. They were subsequently deformed by south-vergent backfolding and backthrusting along the Insubric line, a large-scale strike-slip and thrust fault that separates the European crust from the Adriatic crust.

Approximately 500 km of north–south convergence between Europe and Africa has been estimated from the restoration of cross-sections across the Alpine orogeny. Part of this shortening has been taken up by folding and thrusting of upper crustal, mainly sedimentary rocks of the Austro-Alpine, Penninic, and Helvetic nappes (large-scale recumbent folds overlying a thrust plane). The lower crust shows wedging of basement massifs including the Aar, Gotthard, and Mont Blanc thrust sheets in the north-west and the Monte Rosa, Gran Paradiso, and Dora Maira massifs in the south-east. The latter shows coesite-bearing eclogites, indicative of ultra-high-pressure metamorphism and deep subduction (around 100 km depth) of continental crustal material.

The later stage of the Alpine orogeny was characterized along the north and northwest by foreland-propagating thrusts extending from the Helvetic nappes into the Swiss molasse basin. The molasse basin developed in front of the rising Alps by flexural buckling of the European lithosphere caused by the excess load of the Alps. Along the south and south-east parts of the orogeny, the South Alpine nappes and Insubric Line structures show south-vergent backfolds and backthrusts, producing a retroshear, or bivergent orogeny with folds and thrusts verging in both directions.

Himalayan-Type Mountain Building

The Himalayan range stretches for over 2500 km from north-west Pakistan, eastwards across northern India, Nepal, Bhutan, and southern Tibet to south-west China, and is the type-example of a mountain range formed as a result of the collision of two continents. The Indian plate was a part of the southern supercontinent Gondwana until it rifted away from southern Africa, Madagascar, and Antarctica approximately 150 My ago. As the Indian Ocean spreading ridges formed in between the continents, India was pushed northwards at rapid plate tectonic rates of around 20 cm year ^{1}, until it collided with the southern margin of Asia approximately 50 My ago at equatorial latitudes. Palaeomagnetic data suggest that India underwent a marked reduction in the rate of northward motion since the time of collision to around 4–5 cm year ^{1}. The Indian plate has also undergone

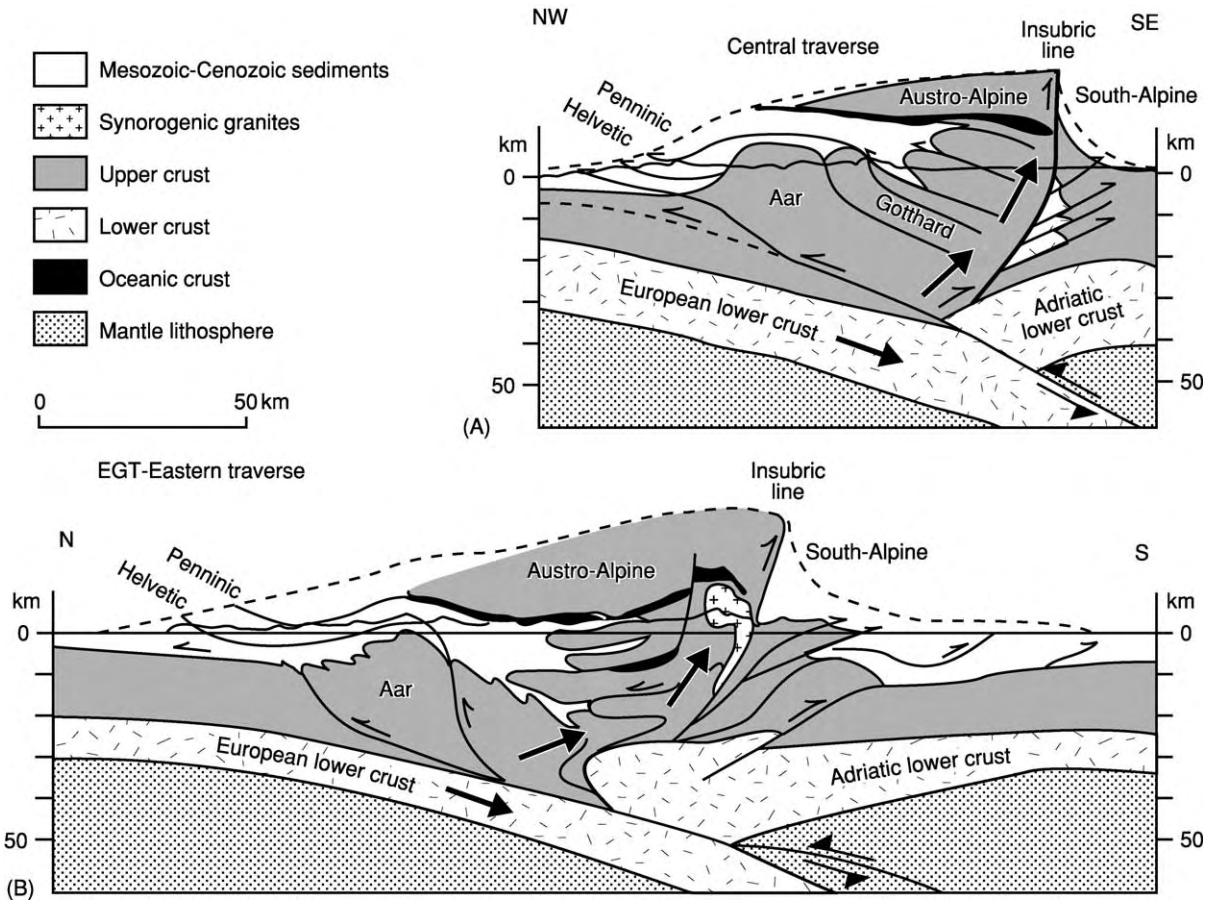


Figure 2 Two sections across the European Alps, showing the overall structure with the stacking of upper crustal sheets above the subducting lower crust. After Schmid *et al.* (1996) Geophysical geological transect and tectonic evolution of the Swiss Italian Alps. *Tectonics* 15: 1036 1064.

20°–30° of counterclockwise rotation. GPS measurements suggest that the present rate of contraction across the Himalaya is about $17.5 \pm 2\,\text{mm year}^{-1}$. Earthquake focal mechanisms show that the majority of earthquakes are a result of north–south compression and occur along the main Himalayan fault, the active plate boundary along which the Indian Shield underthrusts the Lesser Himalaya.

The collision of India with Asia was the most recent of a series of continental plate collisions, which successively accreted smaller continental blocks onto the stable Siberian shield. Since the initial continental collision, India has moved northwards, indenting into Asia by over 2500 km and creating renewed uplift of all the mountain ranges along the southern margin of Asia, including the Pamir, Hindu Kush, the Karakoram, and the Tibetan plateau (Figure 3). The effects of the Indian plate collision extend northwards as far as the Tien Shan and Altay ranges along the border of inner Mongolia.

The earliest effects of the India–Asia collision occurred along the Indus suture zone where the two plates first met. This zone is marked by oceanic rocks including ophiolites, slabs of oceanic crust, and upper mantle emplaced onto continental margins, deep-sea sediments, and occasionally high-pressure metamorphic rocks, indicative of subduction zones. The youngest marine sediments along the suture zone are Early Eocene *Nummulites*-bearing limestones deposited between 52 and 49 My ago. These rocks are frequently used as a proxy for dating the collision. As the collision progressed, the sedimentary rocks along the northern continental margin began to shorten by folding and thrusting processes. The northern mountain ranges of the Himalaya in Zanskar, Spiti, northern Nepal, and southern Tibet show spectacular folding and thrusting of these Tethyan sedimentary rocks. Approximately 150 km of shortening of upper crustal rocks has been estimated across the Tethyan zone. The lower crust that originally underlay these upper crust sediments must have been detached and underthrust northwards beneath the southern margin of the Asian plate (Lhasa block).

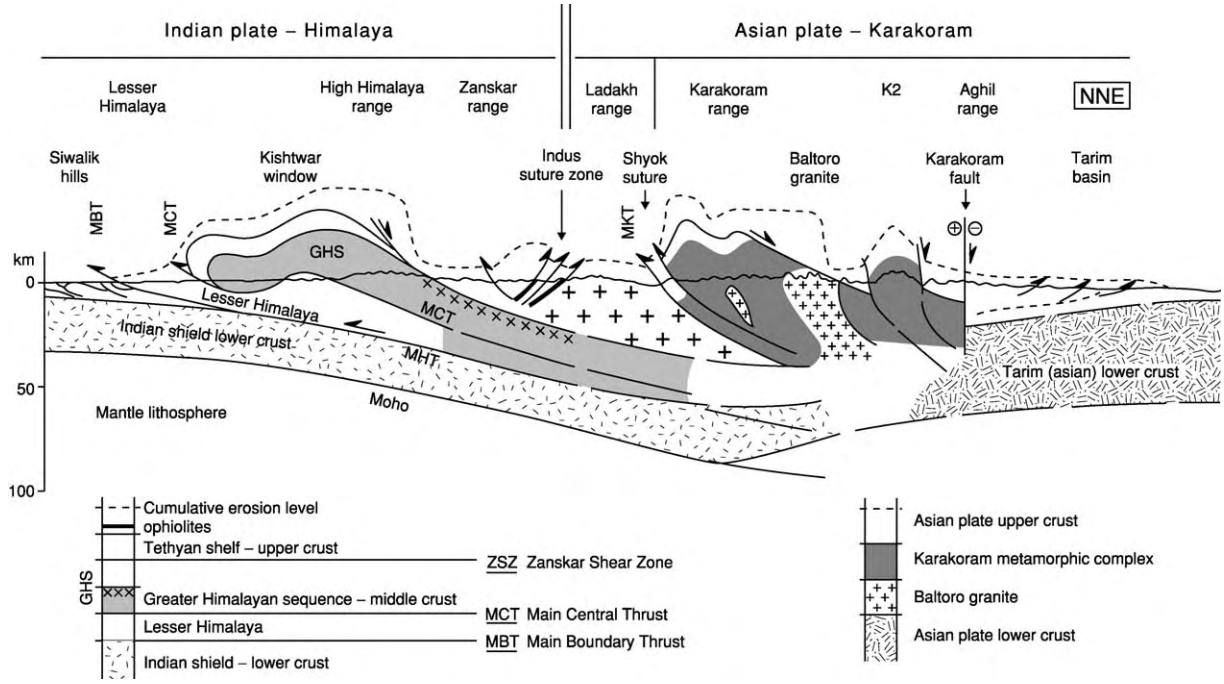


Figure 3 Geological profile across the western Himalaya and central Karakoram mountains. The middle crust metamorphic rocks in the Greater Himalaya and southern Karakoram are shaded and also contain crustal melt granites. The dashed line above shows the approximate cumulative erosion level, the amount of rock material eroded off the mountain range during the Tertiary orogeny. After Searle (1991) *Geology and Tectonics of the Karakoram Mountains*. Chichester, England: Wiley.

Crustal shortening and thickening resulted in increased temperatures and pressures in the deeper crust, causing regional metamorphism along the Greater Himalaya. Many of the highest peaks along the Himalaya are composed of the exhumed deep crustal metamorphic rocks. Dating of monazites by U–Th–Pb methods, and garnets using Sm–Nd isotopes from kyanite- and sillimanite-bearing gneisses in the Greater Himalaya shows that peak *P–T* conditions were initially reached between 35 and 30 My ago, and that temperatures remained high (above 600°C, in the kyanite and sillimanite stability fields) for at least 15 My after that. Around 20 My ago during the Early Miocene, temperatures peaked during a widespread sillimanite-grade metamorphic event that culminated in partial melting of the crust. This melting event resulted in migmatisation of the gneisses and generation of crustal melt leucogranites. The Himalayan leucogranites contain tourmaline, garnet, muscovite, and biotite, and they have very distinctive isotope chemistry (extremely high $^{87}Sr/^{86}Sr$ ratios) characteristic of granites produced from the melting of continental crust. The leucogranites were formed at relatively shallow depths (15–25 km) and were extruded southwards as giant sill complexes.

During the southward extrusion of the Greater Himalayan slab of metamorphic and granitic rocks, the metamorphic isograds were folded and inverted along a giant thrust fault shear zone termed the 'Main Central Thrust'. This thrust fault was active during the Early Miocene (around 20–15 My ago) and carried the entire Greater Himalaya southwards over the relatively unmetamorphosed rocks of the Lesser Himalaya. During this event the metamorphic isograds were recumbently folded and sheared by crustal scale thrusting. As a consequence, an inverted metamorphic isograd sequence characterizes the main central thrust high strain shear zone, with sillimanite- and kyanite-grade rocks structurally above low-grade biotite and garnet-grade rocks. The extrusion of this ductile mid-crustal layer, or channel, of partially molten high-grade gneisses, migmatites, and leucogranites was active during the Miocene (20–17 My ago) with major thrust-related shear zones along the base (Main Central Thrust) and a major normal fault shear zone along the top (South Tibetan detachment).

Around 10 My ago thrusting propagated southwards from the Main Central Thrust into the Lesser Himalaya and eventually to the 'Main Boundary Thrust', the southern boundary of the Himalaya. Active thrust faults have developed during Quaternary and recent times, extending south into the Siwalik hills where the active Himalayan front occurs today. The loading of the Himalaya caused the Indian

plate to flex down and create the Siwalik molasse basin. This basin accumulated all the debris eroded from the rising Himalaya, transported south by rivers. The rivers converge into the Indus River in the west and the Ganges in the east. Sediments eroded from the Himalaya have been transported by these rivers to the Indus Fan in the Arabian Sea and the Bengal Fan in the Bay of Bengal.

Tibetan Plateau

The Tibetan plateau is the largest area of uplifted crust on Earth. The plateau extends for over 3000 km east–west and 1500 km north–south. It lies at an average elevation of just over 5 km above sea-level. The interior of the plateau is very flat with internal drainage, low precipitation, and low erosion rates. The margins of the plateau are ringed by mountain ranges including the Himalaya along the south, the Karakoram and Pamirs to the south-west, the Tien Shan and Kun Lun to the north, and the Long Men Shan along the east. Earthquakes reveal that the high plateau is currently undergoing east–west extension, whilst the margins of the plateau show compression or strike-slip faulting.

The geology of Tibet shows that the plateau region includes several different continental plates that were progressively accreted to the southern continental margin of Asia throughout the Phanerozoic (Figure 4). The most recent and probably the largest of these was the final plate collision, that of India with Asia. The crust beneath the Tibetan plateau is between 65 and 70 km thick, double that of normal continental crust. Several different models have been proposed to account for the thick crust and approximately 1000 km of crustal shortening required. The extreme end-member models include underthrusting of India beneath the entire Tibetan plateau, a model proposed initially by Emile Argand in 1924, and homogeneous thickening of the plateau with very little underthrusting of Indian material at depth.

Recent deep crustal seismic reflection and refraction profiling of Tibet has revealed that the Indian plate lower crust probably underthrusts southern Tibet only as far as the Bangong suture zone approximately 450 km north of the Indus–Yarlung Tsangpo suture zone. The equivalent amount of shortening in the Indian plate upper crust has been taken up by intense folding, thrusting, and crustal thickening in the Tethyan zone and Greater Himalaya. The seismic profiling has also managed to trace the prolongation of the surface faults in the Himalaya, northwards beneath the southern part of the plateau. Seismic and structural data has revealed that the Main Central Thrust and the South Tibetan detachment normal fault bound a mid-crustal layer of hot, partially molten rock that extends southwards to the high Himalaya. Magnetotelluric studies have revealed the presence of 'bright spots' indicative of pockets of fluid or magma at relatively shallow depths beneath southern Tibet today. These have been interpreted as pockets of granitic magma forming today in a structural position similar to those of the 20- to 17-My-old leucogranites cropping out in the high Himalaya. This layer of partial-melt migmatites, high-grade gneisses, and leucogranites was extruded out from beneath the southern part of the Tibetan plateau as a channel of ductile-deforming rock bounded by a rheologically stronger upper crust (Tethyan zone) and lower crust (underplated Indian shield Precambrian and Early Palaeozoic rocks).

Whereas the Tibetan plateau shows little relief despite being 5 km above mean sea-level, the neighbouring Karakoram range of northern Pakistan and Ladakh shows the highest relief of all, with many 7000- to 8000-m-high mountains and deeply incised river valleys. The Karakoram crust has also been tilted, revealing the deep crustal geology not exposed in Tibet (Figure 3). The Karakoram shows multiple episodes of crustal thickening and regional metamorphism spanning the past 65 My and multiple episodes of crustal melting, resulting in granite magmatism. A series of pre-50-My-old granite–granodiorite intrusions indicate that the southern margin of Asia was probably a subduction-related Andean-type continental margin prior to the Indian plate collision 50 My ago. The climax of mountain building in the Karakoram occurred between 24 an – 15 My ago with the emplacement of the huge Baltoro granite batholith, a series of intrusions of biotite monzogranite to biotite–muscovite–garnet leucogranite. A suite of 24- to 22-My-old lamprophyre dykes intruding around the Baltoro granites indicates that parts of the upper mantle were melting at the same time as the lower crust. The ages of the Baltoro granites are similar to the age of the Greater Himalayan leucogranites to the south, which span 24–12 My ago, with the majority between 21 and 17 My ago. This suggests that following the India–Asia collision, crustal thickening, metamorphism, and magmatism spread both across the south Asian margin in the Karakoram and across the north Indian plate margin along the Himalaya.

Earthquake distribution across Tibet shows that the entire plateau region is deforming internally today, and not acting as a rigid plate. Earthquakes in the high plateau show that the crust is undergoing east–west extension, whereas earthquakes in the mountain ranges bordering the high plateau are

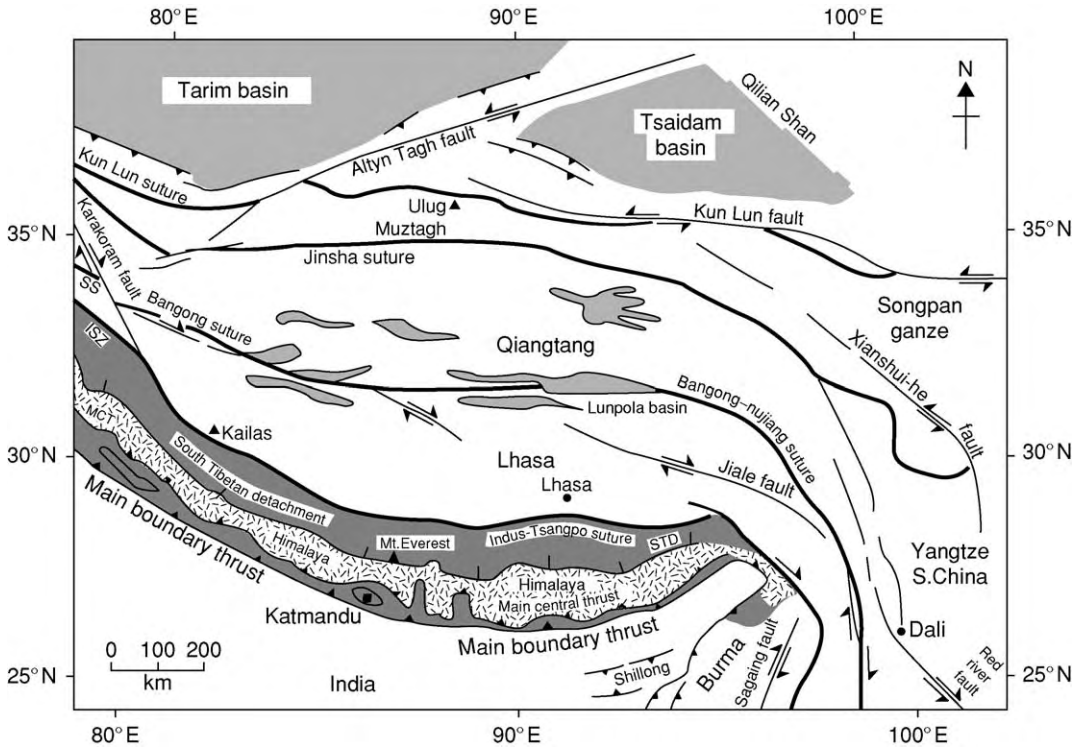


Figure 4 Map of the Tibet region showing the major plate boundaries. The Himalayas are shaded along the southern margin of the plateau. The stable continental blocks of the Tarim basin and Tsaidam basin in the north are shown, together with the major strike slip faults bounding the plateau. Suture zones are progressively younger towards the south from the Kun Lun to the Jinsha, Bangong Nujiang, and Indus Tsangpo suture zones.

mostly a result of either compression or strike-slip deformation. Southern Tibet shows about eight large graben systems, bounded by north–south-aligned steep normal faults. These extensional rifts cut across the northernmost Himalayan range, the Tethyan zone, but die out to the south and do not extend into the Greater Himalayan zone. The rifts are associated with hot springs and active geothermal systems.

The high plateau is bounded by a series of large-scale strike-slip faults, notably the sinistal Altyn Tagh and Kun Lun faults along the north, the dextral Karakoram fault along the south-west, and a series of arcuate, dextral strike-slip faults swinging around the eastern syntaxis (Jiale, Xianshui-he faults). The distribution and slip motion on these faults suggested that the thickened crust of the Tibetan plateau was being extruded horizontally, eastwards out of the way of the Indian plate indentor, a process known as continental extrusion. Despite being impressive faults, many of which show active offsets of Quaternary glacial features, there is relatively limited total geological offset along them, and their timing is generally not concomitant with collision.

The Tibetan plateau region is also remarkable for having a series of shoshonitic volcanic rocks that erupted intermittently over the past 45 My. These volcanic rocks were derived from partial melting of hot asthenospheric mantle. The ages of these volcanics reveal that by 13 My ago, southern Tibet no longer had a hot upper mantle, as relatively cold the Indian plate lithosphere was underplating from the south; however, central and northern Tibet did keep a hot mantle, indicating that any cold lithospheric root to the thickened plateau must have dropped off intermittently and diachronously, rather than as a single catastrophic event.

Conclusions

Although mountain belts can occur along constructive plate margins or even in some intraplate regions, most orogenic belts are the result of plate collision processes along destructive plate margins. Ocean–ocean plate collisions result in island arcs, accretionary prisms, and deep subduction along oceanic trenches. Ocean–continent plate collisions result in Andean-type mountain belts, characterized by calc-alkaline volcanic arcs and granite–granodiorite batholiths. Continent–continent collision can result in a wide array of mountain belts characterized by the

relatively simple folding and thrusting of the Zagros mountains, to the more complicated structure and metamorphism seen in the Alps and Himalaya. The largest and most extensive such collision, that between India and Asia, resulted in thousands of kilometres of crustal shortening, crustal thickening, metamorphism, melting, and exhumation as seen in the Himalaya, Karakoram, and Tibetan plateau.

See Also

Europe: The Alps. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges.

Further Reading

Harrison TM, Copeland P, Kidd WSF, and Yin An (1992) Raising Tibet. *Science* 255: 1663 1670.

Hodges KV (2000) Tectonics of the Himalaya and southern Tibet from two perspectives. *Geological Society of America Bulletin* 112: 324 350.

Keary P and Vine FJ (1996) *Global Tectonics*. Oxford: Blackwell Science.

Schmid SM, Pfiffner OA, Froitzheim N, Schönborn G, and Kissling E (1996) Geophysical geological transect and tectonic evolution of the Swiss Italian Alps. *Tectonics* 15: 1036 1064.

Searle MP (1991) *Geology and Tectonics of the Karakoram Mountains*. Chichester, England: Wiley.

Neotectonics

I Stewart, University of Plymouth, Plymouth, UK

Introduction

Neotectonics concerns the study of horizontal and vertical crustal movements that have occurred in the geologically recent past and which may be ongoing today. Though most crustal movements arise directly or indirectly from global plate motions (i.e., tectonic deformation), neotectonic studies make no presumption about the mechanisms driving deformation. Consequently, 'movements' is a vague catch-all term that encompasses a myriad of competing deformation processes, such as the gradual pervasive creep of tectonic plates, discrete (seismic) displacements on individual faults and folds, and distributed tilting and warping through isostatic readjustment or volcanic upheaval. The phrase 'geologically recent past' is also appropriately vague. Early attempts to define the discipline by arbitrary time windows (e.g., Late Cenozoic, Neogene, or Quaternary) have given ground to a more flexible notion that envisages neotectonism starting at different times in different regions. The onset of the neotectonic period, or the 'current tectonic regime', depends on when the contemporary stress field of a region was first imposed. For instance, the current tectonic regime began in the Middle Quaternary ($\sim$700 000 years ago) in the Apennines of central Italy, and even more recently ($<$500 000 years ago) in California; in contrast, in eastern North America, the present-day stress regime has been in existence for at least the past 15 million years.

Typically then, neotectonic movements have been in operation in most regions for the past few million years or so. Over such prolonged intervals, neotectonic actions are revealed by the stratigraphic build-up of sediments in inland and marine basins, the burial or exhumation histories of rocks, and the geomorphological development of landscapes. Geological studies of palaeobotany and palaeoclimate, numerical models of landscape evolution, and techniques such as fission-track analysis and cosmogenic dating are among the disparate tools unravelling this long-term tectonic activity. Over periods of many tens of to several hundreds of thousands of years, the actions of individual tectonic structures (faults and folds) can be determined, unmasked by their deformation of geomorphic markers, such as marine and fluvial terraces, and tracked with reference to the Late Pleistocene glacial–eustatic time-frame. The apparently smooth deformation rates discerned over intermediate time-scales are revealed to be episodic and irregular when faults and folds are examined over Holocene (10 000 years) time-scales. Over millennial time-scales, secular variations in the activity of tectonic structures can be gleaned from a diverse set of palaeoseismological approaches, from interpreting the stratigraphy of beds that have been affected by faulting, to detecting disturbances in the growth record of trees or coral atolls.

Active Tectonics

Although neotectonic movements continue up to the present day, the term 'active tectonics' is typically used to describe those movements that have occurred over the time-span of human history. Active tectonics deals with the societal implications of neotectonic deformation (such as seismic-hazard assessment, future

sea-level rise, etc.), because it focuses on crustal movements that can be expected to recur within a future interval of concern to society. Contemporary crustal movements may be discerned in Earth surface processes and landforms, such as in the sensitivity of alluvial rivers to crustal tilting. In addition, geomorphological and geological studies are important in recording the surface expression of Earth movements such as earthquake ground ruptures, which, due to their subtle, ephemeral, or reversible nature, are unlikely to have been preserved in the geological record. However, active tectonics also employs an array of high-technology investigative practices; prominent among these are the monitoring of ongoing Earth surface deformation using space-based or terrestrial geodetic methods (tectonic geodesy), radar imaging (interferometry) of ground deformation patterns produced by individual earthquakes and volcanic unrest, and the seismological detection and measurement of earthquakes (seismotectonics). These techniques are applied globally via the World-Wide Standardized Seismograph Network and regionally via local seismographic coverage.

The modern snapshots of tectonism can be pushed back beyond the twentieth century through the analysis of historical accounts and maps to infer past land surface changes or to deduce the parameters of past seismic events (historical seismology). In addition, earthquakes can leave their mark in the mythical practices and literary accounts of ancient peoples, recorded in the stratigraphy of their site histories and in the damage to their buildings (archaeoseismology). The time covered by such human records varies markedly, ranging from many thousands of years in the Mediterranean, Near East, and Asia to a few centuries across much of North America. Generally records confirm that regions that are active today have been consistently active for millennia, thereby demonstrating the long-term nature of crustal deformation, but occasionally records reveal that some regions that appear remarkably quiet from the viewpoint of modern seismicity (such as the Jordan rift valley) are capable of generating large earthquakes. In reality, the distinction between neotectonics and active tectonics is artificial; the terms simply describe different time slices of a continuum of crustal movement. This continuum is maintained by the persistence of the contemporary stress field, which means that inferences of past rates and directions of crustal movement from geological observations can be compared directly with those measured by modern geodetic and geophysical methods.

Although the terms 'neotectonic' and 'active' are somewhat blurred and are often used interchangeably, societal demands (for instance, regulatory authorities for seismic hazard, nuclear safety, etc.) often require the incidence of tectonic movements to be defined strictly. For instance, in the United States, under California law, an 'active fault' is presently defined as one that has generated surface-rupturing earthquakes in the past 11 000 years (i.e., the time period was established to relate to the time when the Holocene was considered to have begun). Other regulatory bodies recognize a sliding scale of fault activity: Holocene (activity in the past 10 000 years), Late Quaternary (activity in the past 130 000 years), and Quaternary (activity in the past 1.6 million years). Neotectonic faults, by comparison, are simply those that formed during the imposition of the current tectonic regime. 'Real' structures, of course, are unconstrained by such legislative concerns. Many modern earthquakes rupture along older (i.e., palaeotectonic) basement faults. Indeed, it is important to recognize that any fault that is favourably oriented with respect to the stress currently being imposed on it has the potential to be activated in the future, regardless of whether it has moved in the geologically recent past.

Global Tectonics

A useful way to differentiate styles and degrees of neotectonic activity is in terms of tectonic strain rate, which is a measure of the velocity of regional crustal motions and, in turn, of the consequent tectonic strain build-up. Crustal movements are most vigorous, and therefore most readily discernible, where plate boundaries are narrow and discrete. In these domains of high tectonic strain, frequent earthquakes on fast-moving ($>10\,\text{mm year}^{-1}$) faults ensure that a century or two of historical earthquakes and a few years of precise geodetic measurements are sufficient to capture a consistent picture of the active tectonic behaviour. Intermediate tectonic strain rates characterize those regions where plate–boundary motion is distributed across a network of slower moving ($0.1–10\,\text{mm year}^{-1}$). Examples of such broad deforming belts are the Basin and Range Province of the western United States or the Himalayan collision zone, where earthquake faults rupture every few hundred or thousand years, ensuring that the Holocene period is a reasonable time window over which to witness the typical crustal deformation cycle. In contrast, areas with low strain rates ensure that intraplate regions, often referred to as 'stable continental interiors', are low-seismicity areas with slow-moving ($<0.1\,\text{mm year}^{-1}$) faults that rupture every few tens (or even hundreds) of thousands of years, making the snapshot of human history an unreliable guide to the future incidence of tectonic activity.

The global pattern of present-day crustal motions can be accounted for by 'plate tectonic' theory, that elegant kinematic framework in which rigid plates variously collide, split apart, and slide along their actively deforming boundaries (*see* **Plate Tectonics**). Closer inspection, however, reveals that the basic rules that govern global plate motions (i.e., rigid blocks separated by narrow deforming boundaries) break down at the regional and local scales. This is particularly so on the continents, where a patchwork of preexisting geology and structure ensures that tectonic stresses are not applied in a uniform, straightforward fashion. Studies of how the contemporary stress field varies across Earth's surface distinguish between first- and second-order stress provinces. First-order provinces have stresses generally uniformly oriented across several thousands of kilometres. The largest of these are the midplate regions of North America and western Europe, where the stress fields are largely the far-field product of ridge push and continental collision. In contrast, first-order stress provinces in tectonically active areas are dominated by the downgoing pull of subducting slabs and the resistance to subduction. Second-order stress provinces are smaller, typically less than 1000 km across, and are related to crustal flexure induced by thick sequences of sediments and postglacial rebound, and to deep-seated rheological contrasts. Although the bulk of Earth's crust is in compression, significant regions of extension occur. In both the continents and the oceans, these extensional domains are long and narrow and correspond to topographically high areas, though notable exceptions are the Basin and Range provinces and the Aegean region of the eastern Mediterranean. Most first-order stress provinces, and many second-order stress provinces, coincide with distinct physiographic provinces.

Glacial Isostatic Adjustment

Plate-driving forces may exert the dominant control on the contemporary stress field, but another process contributes to crustal deformation at a global scale. That process is glacial isostatic adjustment (GIA), the physical response of Earth's viscoelastic mantle to surface loads imposed and removed by the cycles of glaciation and deglaciation, to which the planet has been subjected for the past 900 000 years (*see* **Tertiary To Present:** Pleistocene and The Ice Age). Because large ice-mass fluctuations induce the subcrustal flow of material, measurable crustal deformation extends for thousands of kilometres beyond the limits of the former ice margins; consequently, the effects of GIA are felt globally. In addition, though the crust's elastic response to ice-sheet decay is geologically immediate, the delayed viscoelastic response of the mantle ensures that GIA can persist long after the ice has gone. Although the effects of GIA can now be detected from space geodesy, its legacy is most clearly visible in the worldwide pattern of postglacial sea-level changes. Regions that were ice covered at the Last Glacial Maximum are uplifting (i.e., relative sea-level is currently falling) as a consequence of postglacial rebound of the crust. Likewise, the regions peripheral to the former ice sheets are subsiding (i.e., relative sea-levels are rising) due to collapse of the 'glacial forebulge'. The effect of this subsidence outside the area of forebulge collapse is to draw in water from the central ocean basins, which is compensated by uplift in the ocean basin interiors in the far-field of the ice sheets. The final GIA component is the hydro-isostatic tilting of continental coastlines due to the weight applied to Earth's surface by the returning meltwater load, which produces a 'halo' of weak crustal subsidence around the world's major landmasses. For the most part, geological studies of Holocene relative sea-level changes are consistent with the uplift/subsidence pattern predicted by global viscoelastic theory. The key areas of misfit are along plate boundary seaboards (especially subduction zones), where tectonic deformation dominates, and those areas 'contaminated' by local anthropogenic effects (groundwater extraction, etc.).

The neotectonic implications of GIA are not confined to the coastline. Glacial rebound is now widely considered as an effective mechanism for exerting both vertical and horizontal stresses, not only within the limits of the former ice sheets, but for several hundred kilometres outside. Within the former glaciated parts of eastern North America and northern Europe, both tectonic and rebound stresses are required to explain the distribution and style of both postglacial and contemporary seismotectonics. Outside in the ice-free forelands, predicted glacial strain rates are still likely to be one to three orders of magnitude higher compared to tectonic strain rates typical of continental interiors. Consequently, some workers argue that an apparent 'switching on' of Holocene earthquake activity in the eastern United States and the occurrence of atypically large seismic events, such as the great (magnitude >8) earthquakes that struck the Mississippi valley area of New Madrid in 1811–1812, may be associated with areas where glacial strains are particularly high. Glacial loading and unloading may also disturb the build-up of tectonic strain at glaciated plate boundaries, such as today in Alaska or previously when the Cordilleran ice sheet capped part of the Cascadia subduction zone. More recently, the isostatic component of glacier erosion in the mountain-building process is becoming appreciated.

Global Perspective

The worldwide patterns of vertical and horizontal crustal movements arise from the global effects of plate motions and glacial isostatic adjustment. Regionally and locally, this is augmented by flexure from eustatic or sediment loading, volcanic deformation, or anthropogenic change (e.g., dam impoundment). Though many neotectonic investigations seek to disentangle movements arising from the imposition of tectonic strains from those augmented by non-tectonic processes, this is often a fruitless holy grail; because deformation of Earth's crust typically induces compensatory flow underlying the mantle, neotectonic movements are applied globally. Nevertheless, these disparate contributory mechanisms, coupled with the varying time-scales over which their actions can de discerned, ensure that neotectonics encompasses a remarkable breadth of research disciplines. Few other fields easily blend topics as disparate as space science, seismology, Quaternary science, geochronology, structural geology, geomorphology, geodesy, archaeology, and history. It is this interdisciplinary marriage that makes neotectonics particularly exciting and especially challenging.

See Also

Plate Tectonics. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Burbank DW and Anderson RS (2001) *Tectonic Geomorphology*. Oxford: Blackwell.

Peltier WR (1999) Global sea level rise and glacial isostatic adjustment. *Global and Planetary Change* 20: 93 123.

Stewart IS and Hancock PL (1994) Neotectonics. In: Hancock PL (ed.) *Continental Deformation*, pp. 370 409. London: Pergamon Press.

Stewart IS, Sauber J, and Rose J (2000) Glacio seismotec tonics: ice sheets, crustal deformation and seismicity. *Quaternary Science Reviews* 14/15: 1367 1390.

Vita Finzi C (2002) *Monitoring the Earth*. Harpenden: Terra Publishing.

Ocean Trenches

R J Stern, The University of Texas at Dallas, Richardson, TX, USA

Introduction

An oceanic trench is a long, narrow, and generally very deep depression of the seafloor. Oceanic trenches are the deepest places on the Earth's solid surface and range down to 11 km below sea-level. These tremendous depths mark fundamental breaks in the Earth's lithosphere, the great plates that we all ride on (*see* **Plate Tectonics**). If mid-ocean ridges are where the Earth turns itself inside out, trenches are where the Earth swallows its skin. The asymmetry of trenches reflects a deeper phenomenon: as one plate bends down to return to the mantle, the other plate strains to fill the growing void. The depths of trenches are governed by many things, most importantly sediment flux but also the age of the downgoing lithosphere, the convergence rate, intermediate slab dip, and even the width of the sinking plate. Trenches are sites where fluids are 'squeezed' out of the subducted sediments and a newly recognized biosphere thrives.

Early Years of Study

Trenches are the most spectacular morphological features on the Earth's solid surface, but they were not clearly defined until the late 1940s and 1950s. The depths of the oceans were scarcely imagined until we began to lay telegraph cables between the continents in the late nineteenth and early twentieth centuries. The elongated bathymetric expression of trenches was not recognized early, and the term 'trench' does not appear in Murray and Hjort's classic oceanography overview. Instead they used the term 'deep' to describe the deepest parts of the ocean, such as the Challenger Deep, which is now recognized as the greatest gash on the solid surface of the Earth. Experiences in the World War I battlefields emblazoned the concept of a trench as an elongate depression defining an important boundary, so it is no surprise that the term 'trench' was used to describe natural features in the early 1920s. The term was first used in a geological context by SJ Scofield two years after the war ended to describe a structurally controlled depression in the Rocky Mountains. James Johnstone, in his 1923 textbook *An Introduction to Oceanography*, first used the term in its modern sense to describe a marked elongate depression of the seafloor.

During the 1920s and 1930s, Vening Meinesz developed a unique gravimeter that could measure gravity in the stable environment of a submarine and used it to measure gravity over ocean trenches. His gravity measurements revealed that trenches are sites of downwelling in the solid Earth. The concept of downwelling at trenches was characterized by DT Griggs in 1939 as the tectogene hypothesis, for which he developed an analogous model using a pair of rotating drums. The war in the Pacific led to great improvements in bathymetry, especially in the western and northern Pacific, and the linear nature of the trenches became clear. The rapid growth of deep-sea research efforts, especially the widespread use of echo-sounders in the 1950s and 1960s, confirmed the morphological utility of the term. The important trenches were identified and sampled, and their greatest depths were sonically plumbed. The heroic phase of trench exploration culminated in the 1960 descent of the bathyscaphe *Trieste*, which set an unbeatable world record by diving to the bottom of the Challenger Deep. Following Dietz' and Hess' articulation of the seafloor-spreading hypothesis in the early 1960s and the plate-tectonic revolution in the late 1960s, the term 'trench' has been redefined so that it now has tectonic as well as morphological connotations.

Trenches mark one of the most important types of natural boundary on the Earth's solid surface, that between two lithospheric plates. There are three types of lithospheric-plate boundary: divergent (where lithosphere and oceanic crust are created at mid-ocean ridges (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mid-Ocean Ridges), convergent (where one lithospheric plate sinks beneath another and returns to the mantle (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges), and transform (where two lithospheric plates slide past each other). Trenches are the spectacular and distinctive morphological features of convergent plate boundaries (Figure 1). Plates move together along convergent plate boundaries at rates that vary from a few millimetres to ten or more centimetres per year. Trenches form where oceanic lithosphere is subducted at a convergent plate margin,

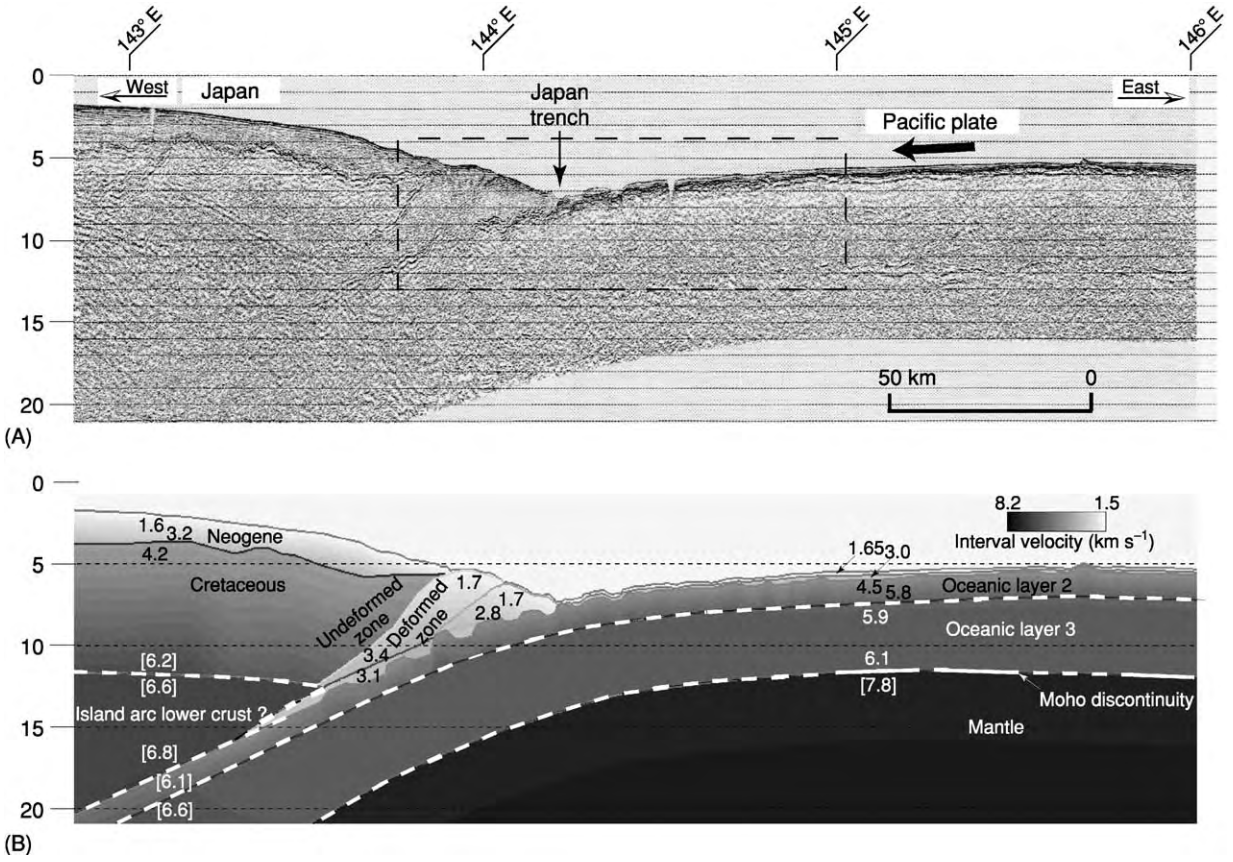


Figure 1 Profile across a typical trench (the Japan Trench near 39° N). (A) Seismic reflection image (pre stack depth migration) and (B) interpretation, including crustal units and seismic velocities (in km s ^{1}). Dashed box in (A) indicates the region shown in detail in Figure 3A. Vertical exaggeration in (B) is four times. The model is shaded according to the seismic velocities, and selected digital values are also shown. Modified from Tsuru T, Park J O, Takahashi N, *et al.* Tectonic features of the Japan Trench convergent margin off Sanriku, northeastern Japan, revealed by multichannel seismic reflection data. *Journal of Geophysical Research* 105: 16 403 16 413.

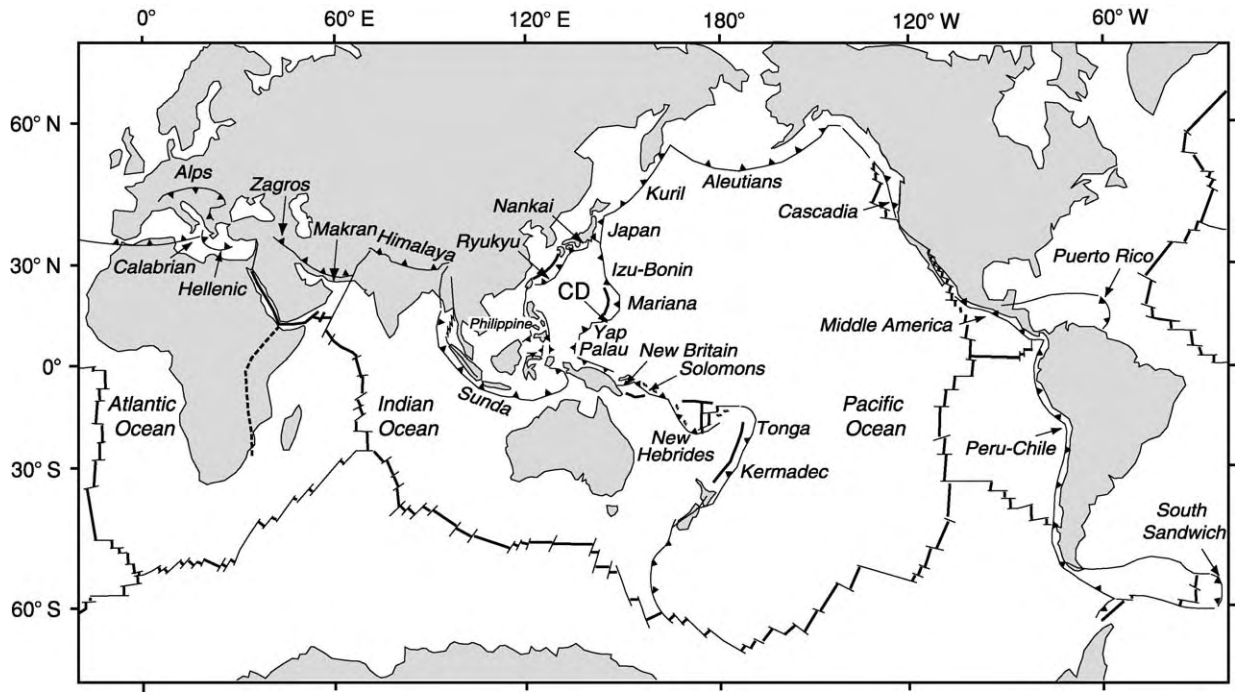


Figure 2 Locations of the world's major trenches and collision zones. CD is the Challenger Deep.

presently at a global rate of about a tenth of a square metre per second.

Geographical Distribution

There are about 50 000 km of convergent plate margins in the world, mostly around the Pacific Ocean – the reason that they are sometimes called 'Pacific-type' margins – but they also occur in the eastern Indian Ocean, and there are relatively short convergent-margin segments in the Atlantic and Indian Oceans and in the Mediterranean Sea (Figure 2). Trenches are sometimes buried and lack bathymetric expression, but the fundamental structures that they represent mean that the name should still be applied in these cases. This applies to the Cascadia, Makran, southern Lesser Antilles, and Calabrian trenches (Table 1). Trenches, magmatic (island) arcs, and zones of earthquakes that dip under the magmatic arc as deeply as 700 km are diagnostic of convergent plate boundaries and their deeper manifestations, subduction zones. Trenches are related to but distinguished from zones of continental collision, where continental lithosphere enters the subduction zone. When buoyant continental crust enters a trench, subduction eventually stops and the convergent plate margin becomes a collision zone. Features analogous to trenches are associated with collision zones; these include sediment-filled

Table 1 Maximum trench depth

Trench	*Maximum depth (m)*	*Accretionary prism?*
Challenger, Mariana	10 920	No
Tonga	10 800	No
Philippine (East Mindanao)	10 057	No
Kermadec	10 047	No
Izu Bonin	9 780	No
Kuril	9 550	No
North New Hebrides	9 175	No
New Britain	8 940	No
Yap	8 650	No
Puerto Rico	8 605	No
South Sandwich	8 325	No
South Solomons	8 322	No
Peru Chile	8 170	No
Japan	8 130	No
Palau	8 055	No
Aleutians	7 680	Yes
Ryukyu	7 460	Yes
Sunda	7 125	Yes
Middle America	6 660	No
Hellenic	5 092	Yes
Nankai	4 900	Yes
Calabrian	4 200	Yes, no morphological trench
Makran	3 200	Yes, no morphological trench
Cascadia	3 136	Yes, no morphological trench

foredeeps, referred to as peripheral foreland basins, such as that which the Ganges and Tigris–Euphrates rivers flow along.

Morphological Expression

Trenches are the centrepieces of the distinctive physiography of a convergent plate margin. Transects across trenches yield asymmetric profiles, with relatively gentle (*ca.* 5°) outer (seaward) slopes and steeper (*ca.* 10–16°) inner (landward) slopes. This asymmetry is due to the fact that the outer slope is defined by the top of the downgoing plate, which must bend as it begins its descent. The great thickness of the lithosphere requires that this bending be gentle. As the subducting plate approaches the plate boundary, it is first bent upwards to form the outer swell and then descends to form the outer trench slope. The outer trench slope is disrupted by a set of subparallel normal faults, which staircase the seafloor down into the trench (Figure 3). The plate boundary is defined by the trench axis itself. Beneath the inner trench wall, the two plates slide past each other along the subduction décollement, which intersects the seafloor along the base of the trench. The overriding plate generally contains a magmatic arc and a fore arc. The magmatic arc is created as a result of physical and chemical interactions between the subducted plate at depth and the asthenospheric mantle associated with the overriding plate. The fore arc lies between the trench and the magmatic arc. Fore arcs have the lowest heat flow of any place on Earth because there is no asthenosphere (convecting mantle) between the forearc lithosphere and the cold subducting plate.

The inner trench wall marks the edge of the overriding plate and the outermost fore arc. The fore arc consists of igneous and metamorphic basement, and this basement may act as a buttress to a growing accretionary prism, depending on how much sediment is supplied to the trench. If the sediment flux is high, material will be transferred from the subducting plate to the overriding plate. In this case an accretionary prism grows, and the location of the trench migrates away from the magmatic arc over the life of the convergent margin. Convergent margins with growing accretionary prisms are called accretionary convergent margins and make up nearly half of all convergent margins. If the sediment flux is low, material will be transferred from the overriding plate to the subducting plate by a process of tectonic ablation known as subduction erosion and carried down the subduction zone. Fore arcs undergoing subduction erosion typically expose igneous rocks. In this case,

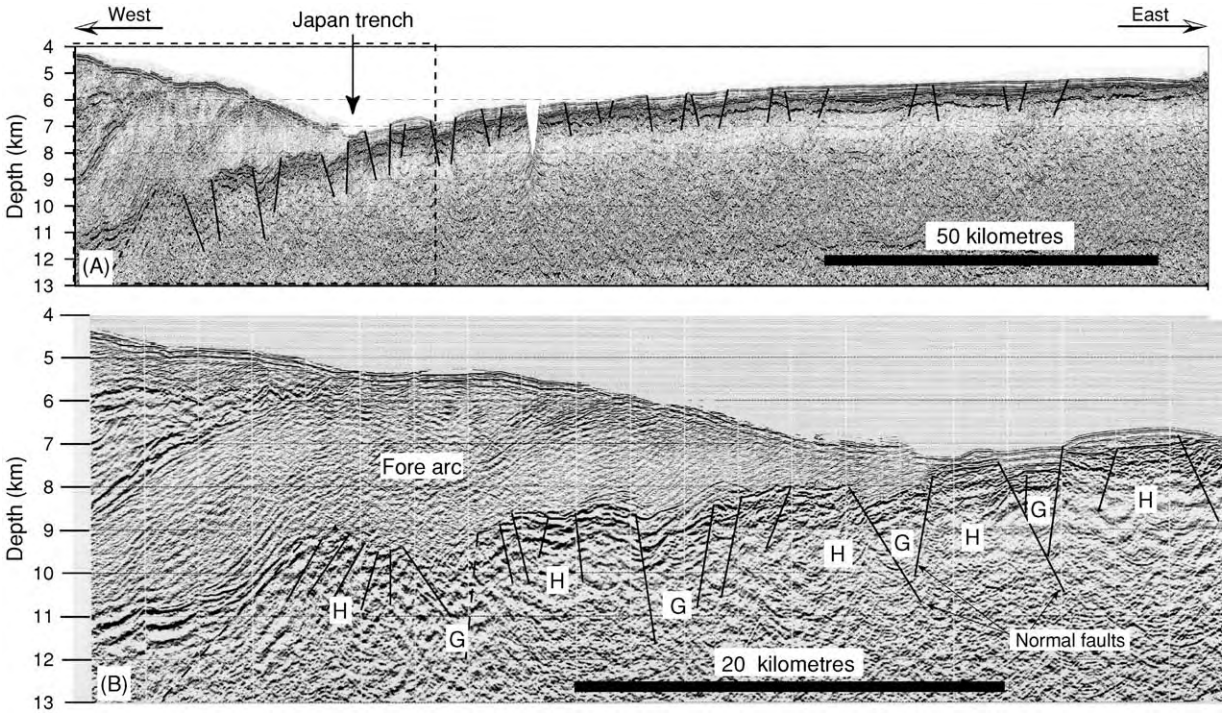


Figure 3 Horst (H) and graben (G) structures with normal faults associated with the subduction of the Pacific plate in the Japan Trench. Vertical exaggeration is four times in (A) and twice in (B). Dashed box in (A) indicates the region shown in detail in (B). Modified from Tsuru T, Park J O, Takahashi N, *et al.* Tectonic features of the Japan Trench convergent margin off Sanriku, northeastern Japan, revealed by multichannel seismic reflection data. *Journal of Geophysical Research* 105: 16 403 16 413.

the location of the trench will migrate towards the magmatic arc over the life of the convergent margin. Convergent margins experiencing subduction erosion are called nonaccretionary convergent margins and comprise more than half of convergent plate boundaries. This is an oversimplification, because a convergent margin can simultaneously experience sediment accretion and subduction erosion.

The asymmetric profile across a trench reflects fundamental differences in materials and tectonic evolution. The outer trench wall and outer swell are composed of seafloor, which takes a few million years to move from where subduction-related deformation begins near the outer trench swell to where the plate sinks beneath the trench. In contrast, the inner trench wall is deformed by plate interactions throughout the life of the convergent margin. The fore arc is continuously subjected to subduction-related earthquakes. This protracted deformation and shaking ensures that the slope of the inner trench wall is controlled by the angle of repose of whatever material it is composed of. Because they are composed of igneous rocks instead of deformed sediments, nonaccretionary trenches have steeper inner walls than accretionary trenches.

Filled Trenches

The composition of the inner trench slope is determined by sediment supply, which also exerts a first-order control on trench morphology. Active accretionary prisms are especially common in trenches that are located near continents where large rivers or glaciers reach the sea. These filled trenches cause some confusion because they are tectonically indistinguishable from other convergent margins but lack the bathymetric expression of a trench. The Cascadia margin off the north-west coast of the USA is a filled trench, the result of sediments delivered by the rivers of the north-western USA and south-west Canada. The Lesser Antilles convergent margin shows the importance of proximity to sediment sources for trench morphology. In the south, near the mouth of the Orinoco River, there is no morphological trench, and the fore arc and the accretionary prism have a total width of almost 500 km. The accretionary prism is so large that it forms the islands of Barbados and Trinidad. Northwards the fore arc narrows, the accretionary prism disappears, and north of 17 °N only the morphology of a trench is seen. In the extreme north, far away from sediment sources, the Puerto Rico Trench is over 8600 m deep, and there is no active accretionary prism. A similar relationship between proximity to rivers, fore arc width, and trench morphology can be observed from east to west along the Alaskan–Aleutian convergent margin. The convergent plate boundary off the coast of Alaska changes along its strike from a filled trench with a broad fore arc in the east (near the coastal rivers of Alaska) to a deep trench with a narrow fore arc in the west (near the Aleutian islands). Another example is the Makran convergent margin off the coasts of Pakistan and Iran, where the trench is filled by sediments from the Tigris–Euphrates and Indus rivers. Thick accumulations of turbidite deposits along a trench can be supplied by down-axis transport of sediments that enter the trench 1000–2000 km away, as is found in the Peru–Chile Trench south of Valparaiso and in the Aleutian Trench. Convergence rate can also be important in controlling trench depth, especially for trenches near continents, because slow convergence may mean that the capacity of the convergent margin to dispose of sediment is exceeded.

Thus, trench morphology can be expected to evolve continuously as oceans close and continents converge. While the ocean is wide, the trench may be far away from continental sources of sediment and so may be deep. As the continents converge, the trench may increasingly be filled with continental sediments and shoals. A simple definition of the transition from subduction to collision is when a plate boundary previously marked by a trench has filled sufficiently to rise above sea-level.

Accretionary Prisms and Sediment Transport

Accretionary prisms grow by frontal accretion – where sediments are scraped off, bulldozer-fashion, near the trench – or by underplating of subducted sediments and perhaps oceanic crust along the shallow parts of the subduction décollement. Frontal accretion over the life of a convergent margin results in the youngest sediments being found in the outermost part of the accretionary prism and the oldest sediments being found in the innermost portion. Older (inner) parts of the accretionary prism are much more lithified and have steeper structures than the younger (outer) parts. Underplating is difficult to detect in modern subduction zones but may be recorded in ancient accretionary prisms, such as the Franciscan Group of California, in the form of tectonic mélanges and duplex structures. Different modes of accretion are reflected in the morphology of the inner slope of the trench, which generally shows three morphological provinces. The lower slope comprises imbricate thrust slices, which form ridges. The middle part of the slope may comprise a bench or terraces. The upper slope is smoother but may be cut by submarine canyons.

Because accretionary convergent margins have high relief, are continuously deformed, and accommodate

a large flux of sediments, they are sites of vigorous sediment dispersal and accumulation. Sediment transport is controlled by submarine landslides, debris flows, turbidity currents, and contourites. Submarine canyons transport sediment from beaches and rivers down the upper slope. These canyons are formed by channelized turbidites and generally lose definition with depth because continuous tectonic readjustments disrupt the channels. Sediments move down the inner trench wall via channels and a series of fault-controlled basins. The trench itself serves as an axis of sediment transport. If enough sediment moves into the trench, it may be completely filled, and turbidity currents will then be able to carry sediment well beyond the trench and may even surmount the outer swell. Sediments from the rivers of southwest Canada and the north-western USA spill over where the Cascadia trench would be and reach the Juan de Fuca spreading ridge several hundred kilometres to the west.

The slope of the inner trench wall of an accretionary convergent margin continuously adjusts to the thickness and width of the accretionary prism. The prism maintains a 'critical taper', established by the Mohr–Coulomb failure criterion for the pertinent materials. A package of sediments scraped off the downgoing lithospheric plate will deform until it and the accretionary prism that it has been added to attain a critical-taper (constant slope) geometry. Once critical taper is attained, the wedge slides stably along its basal décollement. Strain rate and hydrological properties strongly influence the strength of the accretionary prism and thus the angle of critical taper. Fluid pore pressure can modify rock strength and is an important determinant of critical taper angle. Low permeability and rapid convergence may lead to pore pressures that exceed lithostatic pressure and result in a relatively weak accretionary prism with a shallowly tapered geometry, whereas high permeability and slow convergence lead to lower pore pressures, stronger prisms, and steeper geometry.

The Hellenic Trench system is unusual because its convergent margin subducts evaporites. The slope of the southern flank of the Mediterranean Ridge (its accretionary prism) is low, about 1°, which indicates very low shear stress on the décollement at the base of the wedge. Evaporites influence the critical taper of the accretionary complex, because their mechanical properties differ from those of siliciclastic sediments and because of their effect upon fluid flow and fluid pressure, which control effective stress. In the 1970s, the linear deeps of the Hellenic Trench south of Crete were interpreted as being similar to trenches in other subduction zones, but, with the realization that the Mediterranean Ridge is an accretionary complex, it became apparent that the Hellenic Trench is actually a starved fore arc basin, and that the plate boundary lies south of the Mediterranean Ridge.

Water and Biosphere

The volume of water escaping from within and beneath the fore arc results in some of the Earth's most dynamic and complex interactions between aqueous fluids and rocks. Most of this water is trapped in pores and fractures in the upper lithosphere and the sediments of the subducting plate. The average fore arc is underlain by a solid volume of oceanic sediment that is 400 m thick. This sediment enters the trench with 50–60% porosity. The sediment is progressively squeezed as it is subducted, reducing void space and forcing fluids out along the décollement and up into the overlying fore arc, which may or may not have an accretionary prism. Sediments accreted to the fore arc are another source of fluids. Water is also bound in hydrous minerals, especially clays and opal. The increasing pressure and temperature experienced by the subducted materials convert the hydrous minerals to denser phases that contain progressively less structurally bound water. Water released by dehydration accompanying phase transitions is another source of fluid introduced to the base of the overriding plate. These fluids may travel diffusely through the accretionary prism, via interconnected pore spaces in sediments, or may follow discrete channels along faults. Sites of venting may take the form of mud volcanoes or seeps and are often associated with chemosynthetic communities. Fluids liberated in the shallowest parts of the subduction zone may also escape along the plate boundary but have rarely been observed to drain along the trench axis. All of these fluids are dominated by water but also contain dissolved ions and organic molecules, especially methane. Methane is often sequestered in an ice-like form (clathrate) in the fore arc. Gas hydrates are a potential energy source and can rapidly break down. The destabilization of gas hydrates has contributed to global warming in the past and will probably do so in the future (*see* **Petroleum Geology:** Gas Hydrates).

Chemosynthetic communities thrive where fluids seep out of the fore arc. Cold seep communities have been discovered on inner trench slopes in the western Pacific, especially around Japan, in the Eastern Pacific, along the North, Central, and South American coasts from the Aleutian to the Peru–Chile Trenches, on the Barbados prism, in the Mediterranean, and in the Indian Ocean, along the Makran and Sunda convergent margins. These communities have been found down to depths of 6000 m. They have received much less attention than the chemosynthetic communities associated with hydrothermal vents. Chemosynthetic

communities are located in a variety of geological settings: above over-pressured sediments in accretionary prisms, where fluids are expelled through mud volcanoes or ridges (Barbados, Nankai, and Cascadia); along active erosive margins with faults; and along escarpments caused by debris slides (Japan Trench, Peruvian margin). Surface seeps may be linked to massive hydrate deposits and destabilization (e.g. Cascadia margin). High concentrations of methane and sulphide in the fluids escaping from the seafloor are the principal energy sources for chemosynthesis (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges).

Empty Trenches and Subduction Erosion

Trenches distant from an influx of continental sediments lack an accretionary prism, and the inner slope of such trenches is commonly composed of igneous or metamorphic rocks. Nonaccretionary convergent margins are characteristic of (but not limited to) primitive arc systems. Primitive arc systems are those that are built on oceanic lithosphere, such as the Izu–Bonin–Mariana, Tonga–Kermadec, and Scotia (South Sandwich) arc systems. The inner trench slopes of these convergent margins expose the crust of the fore arcs, including basalt, gabbro, and serpentinized mantle peridotite. These exposures allow easy access to materials from the lower oceanic crust and upper mantle, and provide a unique opportunity to study the magmatic products associated with the initiation of subduction zones. Most ophiolites are probably formed in a fore-arc environment during the initiation of subduction, and this setting favours ophiolite emplacement during collision with blocks of thickened crust. Not all nonaccretionary convergent margins are associated with primitive arcs. Trenches adjacent to continents where there is a low influx of sediments from rivers, such as the central part of the Peru–Chile Trench, may also lack an accretionary prism.

The igneous basement of a nonaccretionary fore arc may be continuously exposed by subduction erosion. This transfers material from the fore arc to the subducting plate and can take the form of frontal erosion or basal erosion. Frontal erosion is most active in the wake of seamounts being subducted beneath the fore arc. Subduction of large edifices (seamount tunnelling) oversteepens the fore arc, causing mass failures that carry debris towards and ultimately into the trench (Figure 4). This debris may be deposited in graben of the downgoing plate and subducted with it. In contrast, structures resulting from basal erosion of the fore arc are difficult to recognize on seismic-reflection profiles, so the occurrence of basal erosion is difficult to confirm. Subduction erosion may also diminish a once-robust accretionary prism if the flux of sediments into the trench diminishes.

Nonaccretionary fore arcs may also be sites of serpentinite mud volcanism. Serpentinite mud volcanoes form where fluids released from the downgoing plate percolate upwards and interact with the cold mantle lithosphere of the fore arc. Peridotite is hydrated into serpentinite, which is much less dense than peridotite and so will rise diapirically when there is an opportunity to do so. Some nonaccretionary fore arcs, for example the Marianas, are subjected to strong extensional stresses, which allows buoyant serpentinite to rise to the seafloor and form serpentinite mud volcanoes. Chemosynthetic communities are also found on non-accretionary margins such as the Marianas, where they thrive on vents associated with serpentinite mud volcanoes.

Outer Trench Swell

The outer rise is where the descending plate begins to flex and fault as it approaches the subduction zone. Here, the lithosphere is bent upwards by plate stresses, just as the plate is bent downwards in the trench – in neither case is the plate in isostatic equilibrium. Typically, the gravity over the outer swell is about 50 mGals higher than expected from isostasy, while gravity over the trench is about 200 mGals less than that expected from isostatic considerations. The bending of the plate is associated with tension in the upper 20 km, and shallow earthquakes, caused by tensional failure induced by the downward bending of the oceanic plate, are common: about 20 extensional outer-rise earthquakes of magnitude 5 or greater occur annually. Most axes of tension are perpendicular to the trench, regardless of the direction of relative motion between the two plates, indicating that failure is controlled by bending stresses in the plate. Plate bending also causes deeper (down to 50 km) earthquakes due to compression. The width of the outer rise is directly related to the flexural rigidity of the lithosphere. The thickness of the elastic lithosphere varies between 20 km and 30 km for most trench profiles. Faulting related to plate bending and stair-stepping of the descending slab into the trench may allow seawater to infiltrate deep into the crust and perhaps into the upper mantle. Faulting of the downgoing plate results in a horst-and-graben structure, which allows sediment that reaches the trench to be deposited in graben and carried downward. This faulting also breaks up seamounts as they approach the trench (Figure 5). The mechanism of frontal erosion may operate through the combined effects of seamount tunnelling, mass wasting and

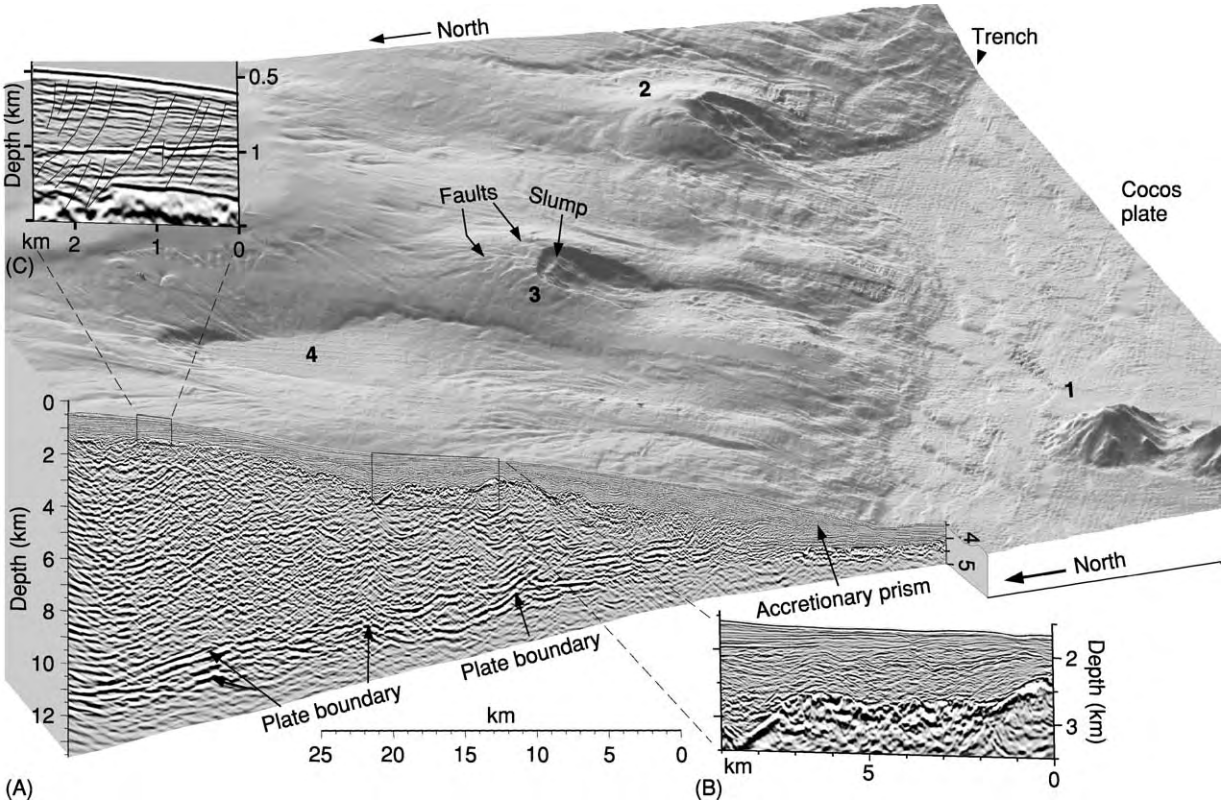


Figure 4 Downgoing seamounts and subduction erosion in the Middle America Trench off the coast of Costa Rica, where the Cocos Plate is being rapidly subducted (80 mm yr ^{1}). (A) Four seamounts in various stages of subduction (1 4) are particularly well manifested in the bathymetry of the inner trench wall. Seamount 1 (about 1 km tall) is approaching the trench and will enter it in about 200 000 years. Seamount 2 entered the trench about 200 000 years ago and is destroying the inner trench wall. The collision has caused oversteepening, with relief locally exceeding 0.5 km, leading to collapse at the sides and especially in the wake of the seamount. Note the slump and fractures. Oversteepening causes submarine landslides and flows of debris towards the trench, rebuilding the angle of repose and flooding the trench floor. Seamount 3 entered the trench about 400 000 years ago and has been swallowed beneath the accretionary prism. Debris continues to be shed from the impact zone and flows into the trench. Seamount 4 entered the trench about 600 000 years ago, and the region above it has almost completed its collapse above the sunken seamount. Subduction erosion usually occurs when debris flows fill graben (**Figure 3B**) in the downgoing plate and are carried down. Note also the seismic reflection profile across the Central American fore arc. (B) Detail of part of the seismic reflection profile. The frontal *ca.* 40 km of the margin has a rough margin wedge top produced by seamount subduction. (C) Where the margin wedge is more than 6 8 km thick its top is smooth, cut only by normal faulting. Modified from Ranero C and von Huene R (2000) Subduction erosion along the Middle America convergent margin. *Nature* 404: 748 752.

transport to the trench, deposition in a graben on the downgoing plate, and descent into the mantle.

Controls on Trench Depth

There are several factors that control the depths of trenches. The most important is the supply of sediment, which may fill the trench so that there is no bathymetric expression. It is therefore not surprising that the deepest trenches are all nonaccretionary. Table 1 shows that all trenches deeper than 8000 m are nonaccretionary. In contrast, all trenches with growing accretionary prisms are shallower than 8000 m. A second factor controlling trench depth is the age of the lithosphere at the time of subduction. Because oceanic lithosphere cools and thickens as it ages, it subsides. The older the seafloor, the deeper it lies, and this controls the minimum depth from which the seafloor begins its descent. This obvious correlation can be removed by looking at the relative depth (Δd), which is the difference between the regional seafloor depth and the maximum trench depth. The relative depth is affected by the age of the lithosphere at the trench, the convergence rate, and the dip of the subducted slab at intermediate depths. Finally, narrow slabs can sink and roll back more rapidly than broad plates, because it is easier for the underlying asthenosphere to flow around the edges of the sinking plate. Such slabs may have steep dips at relatively shallow depths and so may be

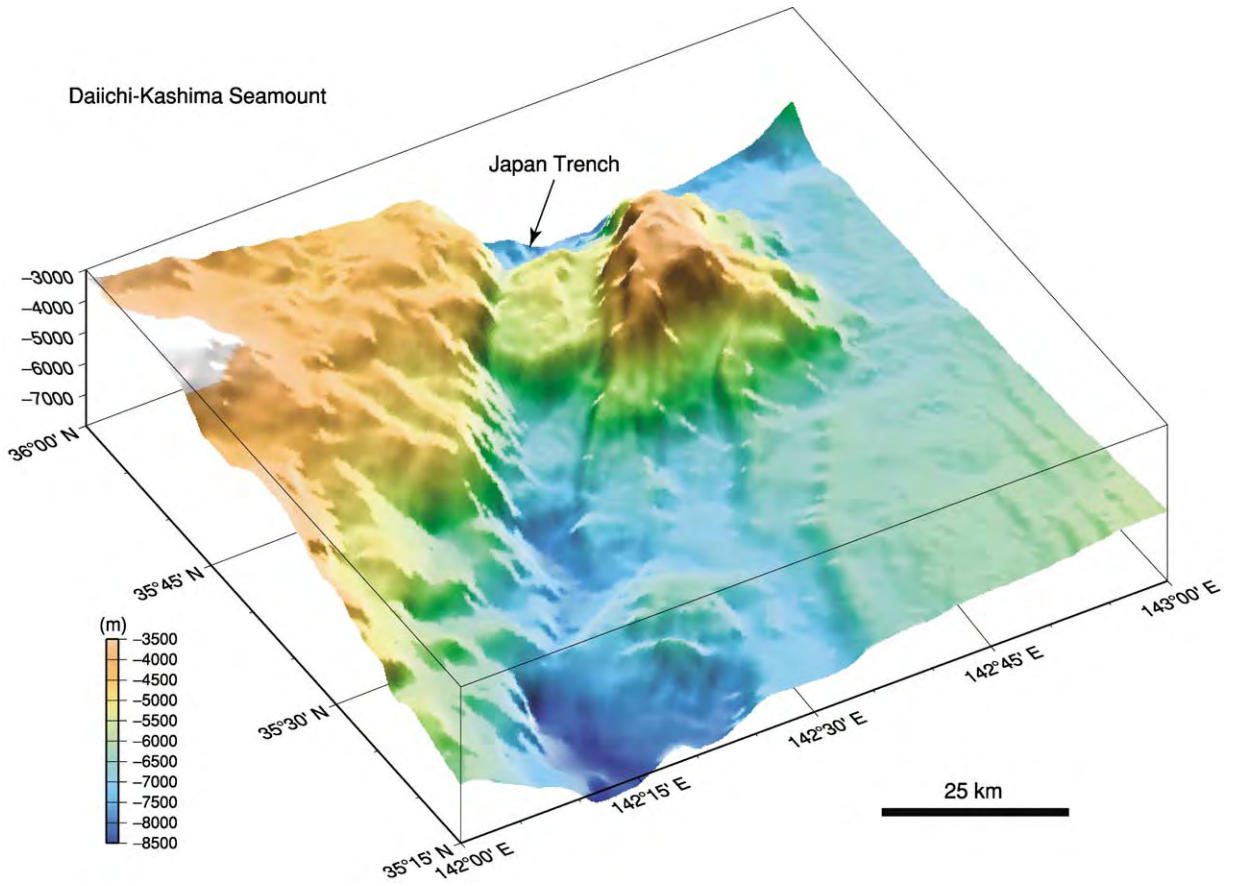


Figure 5 Bathymetric profile showing normal faulting affecting the Daiichi Kahima Seamount as it enters the Japan Trench, east of Tokyo. The seamount, which was originally a guyot (conical sides and a flat top), has been cut by a normal fault that drops its western third by about 1 km. Smaller normal faults related to bending of the plate as it approaches the trench affect the eastern part of the seamount and the seafloor around it. Data from JODC Expert Grid Data for Geography 500 m (J EGG500) Japan Oceanographic Data Center. View is from 225° (azimuth) and 30° (elevation), illumination is from the east. Vertical exaggeration is 5.5 times. Figure generated by Tomoyuki Sasaki of the Ocean Research Institute, University of Tokyo.

associated with unusually deep trenches, such as the Challenger Deep.

See Also

Analytical Methods: Gravity. **Earth:** Mantle; Crust. **Plate Tectonics**. **Seismic Surveys**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mid-Ocean Ridges.

Further Reading

Clift P and Vanucchi P (2004) Controls on Tectonic accre tion versus erosion in Subduction Zones: Implications for the Origin and Recycling of the Continental Crust. *Reviews of Geophysics* v. 42.

Fisher RL (1997) Deep sea trench. In: *Encyclopedia of Science and Technology*, 8th edn. McGraw Hill.

Fisher RL and Hess HH (1963) Trenches. In: Hill MN (ed.) *The Sea: volume 3. The Earth Beneath the Sea*, pp. 411 436. New York: Wiley Interscience.

Gvirtzman Z and Stern RJ (2004) Bathymetry of Mariana trench arc system and formation of the Challenger Deep as a consequence of weak plate coupling. *Tectonics* v. 23.

Hamilton WB (1988) Plate tectonics and island arcs. *Geological Society of America Bulletin* 100: 1503 1527.

Hawkins JW, Bloomer SH, Evans CA, and Melchior JT (1984) Evolution of intra oceanic arc trench systems. *Tectonophysics* 102: 175 205.

Jarrard RD (1986) Relations among subduction param eters. *Reviews of Geophysics* 24: 217 284.

Ladd JW, Holcombe TL, Westbrook GK, and Edgar NT (1990) Caribbean marine geology: active margins of the plate boundary. In: Dengo G and Case J (eds.) *The Geology of North America: volume H. The Caribbean Region*, pp. 261 290. Boulder: Geological Society of America.

Murray J, Sir, Hjort J, Johan J, Appellöf A, Gran HH, and Helland Hansen B (1912) *The depths of the ocean, a general account of the modern science of oceano graphy based largely on the scientific researches of the*

Norwegian steamer Michael Sars in the North Atlantic. xx, p. 821. London: Macmillan.
Sibuet M and Olu K (1998) Biogeography, biodiversity and fluid dependence of deep sea cold seep communities at active and passive margins. *Deep Sea Research Part II: Topical Studies in Oceanography* 45: 517 567.
Smith WHF and Sandwell DT (1997) Global sea floor topography from satellite altimetry and ship depth soundings. *Science* 277: 1956 1962.
Stern RJ (2002) Subduction zones. *Reviews of Geophysics* 10.1029/2001RG0001.
von Huene R and Scholl DW (1993) The return of sialic material to the mantle indicated by terrigeneous material subducted at convergent margins. *Tectonophysics* 219: 163 175.
Watts AB (2001) *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press.
Wright DJ, Bloomer SH, MacLeod CJ, Taylor B, and Goodlife AM (2000) Bathymetry of the Tonga Trench and forearc: a map series. *Marine Geophysical Researches* 21: 489 511.

Rift Valleys

L Frostick, University of Hull, Hull, UK

Introduction

Rift valleys are generally long narrow depressions in the Earths' crust that often contain major lakes and rivers (e.g. Lake Baikal and the Rhine and Rio Grande rivers; see Figure 1). They form as a result of the crust being stretched by plate-tectonic movements, for example during the extension that leads to the opening of new oceans. Rifts have been a feature of geological processes since the solid crust first formed, and early examples of rift valleys have been recognized in rocks as old as the Precambrian, more than 700 Ma ago. Those that have developed during more recent geological time form impressive features of the landscape, with the sides of the rift rising several kilometres above the rift floor. One good example is the East African Rift, which dominates the landscape from Ethiopia in the north to Malawi in the south (*see* **Africa:** Rift Valley).

Rift valleys have been the subject of considerable research and interest for more than a century, partly because of the inherent beauty of deep valleys but also because they are sites where both fossiliferous and economically important deposits are found (e.g. fossil hominids in East Africa, and the oil and gas found in ancient rift deposits of the North Sea; Figure 2). The reason rifts are so productive is linked directly to their mechanism of formation and to the geological structure that sits beneath the very visible surface valley.

Morphology and Structure

The surface expressions of all rifts are similar; there is a central depression or valley flanked by two uplifted shoulders, each of which is cut by a series of faults that step down towards the central lowland. The original interpretation of the geological structure that underlies this morphology was that it is like the keystone of an arch: a central piece of rock drops down between two faults (Figure 3A) to form a structure known by the German word 'graben'. This interpretation was schooled by the surface appearance of rifts and generally ignored the fact that the flat bottoms of rift valleys are almost always clothed in sediments, which mask the underlying geology. With the development of geophysical techniques, especially seismology, that have allowed geologists to glimpse the subsurface, it has become evident that the structure of many rifts is asymmetric, with one margin being higher than the other and the floor being tilted (Figure 3B). The higher margin is cut by the largest faults, which are referred to as the main border faults. The faults are generally normal ones, with one side of a slightly inclined fracture dropping downwards relative to the other and resulting in a near-vertical step in the geology. The lower of the two margins is generally associated with a series of smaller faults inclined both towards and away from the rift axis, which fragment the geology into a series of small horst blocks with intervening valleys. These are termed synthetic faults when the fault plane is inclined in the same direction as that of the border fault and antithetic faults when the inclinations are opposed.

The precise location and magnitude of the faulting is controlled by the thickness and character of the crust undergoing rifting and by the presence of pre-existing lines of weakness, which can be exploited by new faults if they have the right orientation. In some parts of rifts there are a few very large faults, which can have total vertical displacements in excess of 2 km; in other parts there are much smaller faults and the overall displacement is much less. Often the rocks between subparallel marginal faults are

Figure 1 Satellite remote sensing image of the Rio Grande Rift and River, USA. Image taken by the TERRA satellite using the MODIS instrument (moderate resolution imaging spectroradiometer) and enhanced with SRTM30 shaded relief.

tilted away from the rift axis, forming small valleys between the fault blocks.

The margins of rift valleys cut by the main border fault are almost always subject to uplift, which accentuates the topographical step down to the central floor. The distance from margin to margin varies considerably, but continental examples are normally between 30 km and 200 km wide. However, the largest rift system in the world is invisible beneath the surface of our major oceans. The mid-ocean ridges are the centres at which ocean floor is created, and they dominate the seafloor topography (*see* **Tectonics:** Mid-Ocean Ridges). Typically they are between 1000 km and 2000 km wide and 2–3 km high, and they are the foci of considerable volcanic and earthquake activity.

The length of continental rifts varies considerably. The East African Rift is perhaps the best-documented example, and it runs from Afar in Ethiopia in the north to the Zambesi river in the south, a total distance of over 3000 km. However, the rift is not equally deep all along its length; there are deeper areas, often occupied by lakes, and shallower areas, often flanking lake basins. The alternation of shallow and deep areas along the rift reflects segmentation in the underlying structure. Rifts are cut into segments by cross-rift structures that delimit sections dominated by faulting on a particular margin and in a particular orientation. In adjacent segments the main border fault can be on opposite margins, have different orientations, or move laterally with regard to the rift axis. Where such changes occur, the basement rocks come closer to the surface and effectively isolate parts of the rift, ponding water to form lakes. The cross-rift structures in recent rifts may be sufficiently prominent to be seen from space (**Figure 4**), and the string of lakes that results is clearly seen in the African rift (**Figure 5**).

Origin

Although it is generally accepted that rifts form where the crust is stretched as a result of tension, the plate-tectonic setting of rifts is very variable. Rifts can be associated with all three plate-margin types (constructive, destructive, and conservative) and are also found within otherwise-stable plates. There is therefore no single underlying mechanism of formation that can explain all rifts, and hence, the ways in which a rift develops and evolves are very variable. Whatever the mechanism, all rifts are situated in areas where the crust has been stretched and has thinned, much like the thinning that occurs in warm wax when it is pulled apart. As the crust thins, hot low-density mantle material is pulled upwards towards the surface, resulting in relatively high heat flows in the rocks around most rifts. In some rifts the heating of the crust is linked with the development of major domes, which cover vast areas, and with major volcanic outpourings. A good example of areas of doming linked with volcanism can be seen in the East African Rift, where there are two domes, each over 1000 km in diameter.

Volcanic activity can commence at an early stage of rift development and can be extensive; for example, in East Africa an area of over 500 000 km^2 is covered with rift-based volcanic rocks, and many of the most notable peaks of the area are, in fact, active or dormant volcanoes, e.g. Kilimanjaro. The rock types that spill out of the volcanoes are very varied and range from basic to acidic in nature. They also include some rocks that are rarely found outside rifts, such as carbonatites (*see* **Igneous Rocks:** Carbonatites), which

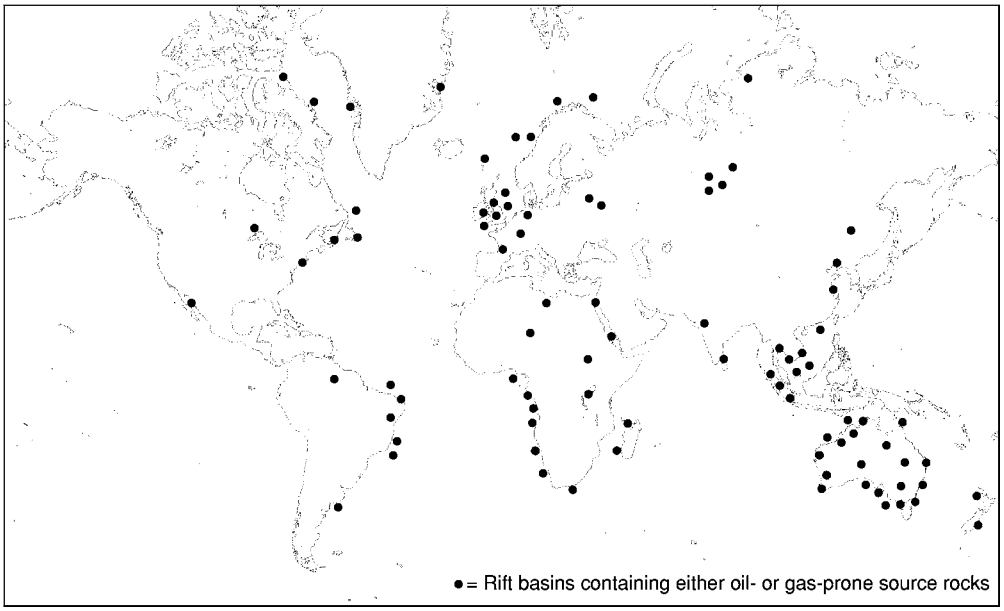


Figure 2 The locations of rift valley deposits that contain oil and gas (black spots). Adapted from Katz BJ (1995) A survey of rift basin source rocks. In: Lambiase JJ (ed.) *Hydrocarbon Habitat in Rift Basins*. Geol. Soc. London Spec. Publ. 80: 213 240.

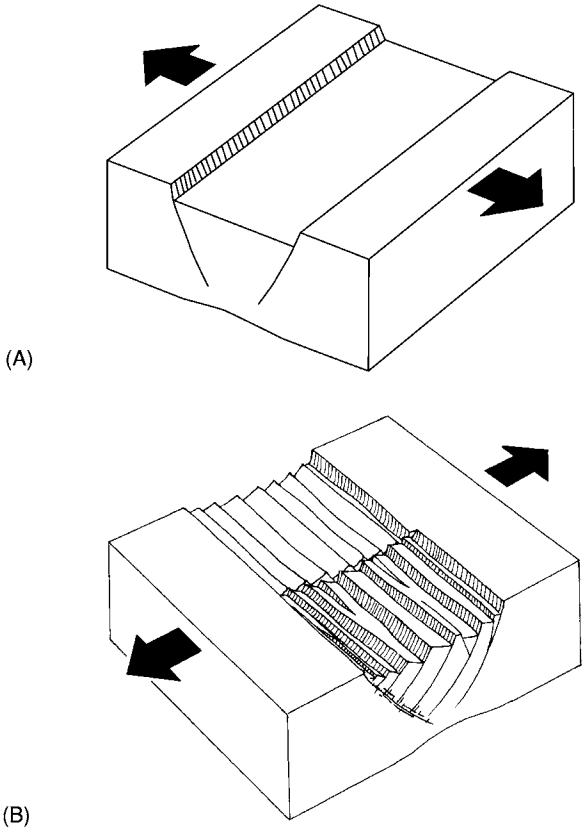


Figure 3 (A) Traditional interpretation of the structure of rifts a full graben with symmetrical fault scarps. (B) The half graben model of rift structure, with strongly asymmetrical fault scarps. Adapted from Perrodon A (1982) Rifts and Fossil Energy Rources, *Ancient Rifts and Troughs Symposium of the French National Centre of Scientific Research (CNRS) Marseilles, Nov Dec 1982.*

contain high levels of various salts that can be dissolved and swept into lakes, making them saline.

The Impact of Rifting on Hydrology, Climate, and Ecology

The development of a rift has a major impact on the overall environment. Continental rifts are often formed in major landmasses that are crossed by large rivers that have spent millions of years wearing down the surface. Rifting produces new topography to which the surface processes adapt, and the consequences of rifting can include the disruption of weather patterns, the diversion of rivers, and the formation of new lakes.

The uplift that occurs along the rift margins will decrease the ambient temperature and can increase rainfall. In contrast, the rift-valley floor remains warmer and can experience a rain-shadow effect since rain forms where clouds are forced to rise, and this occurs preferentially on the higher topographical areas of the flanks. As a result there are often different types of vegetation on the flanks and in the valley bottom. This can include the development of rainforest if the rift is in the tropics. Both the topography and the contrast in habitats across the rift valley can act as barriers to the migration of animals and, to a lesser extent, the spread of plant species. Where a rift forms in a previously undivided continent it produces new diversity of habitat and isolates individuals, both of which can boost evolution. This is one of the reasons

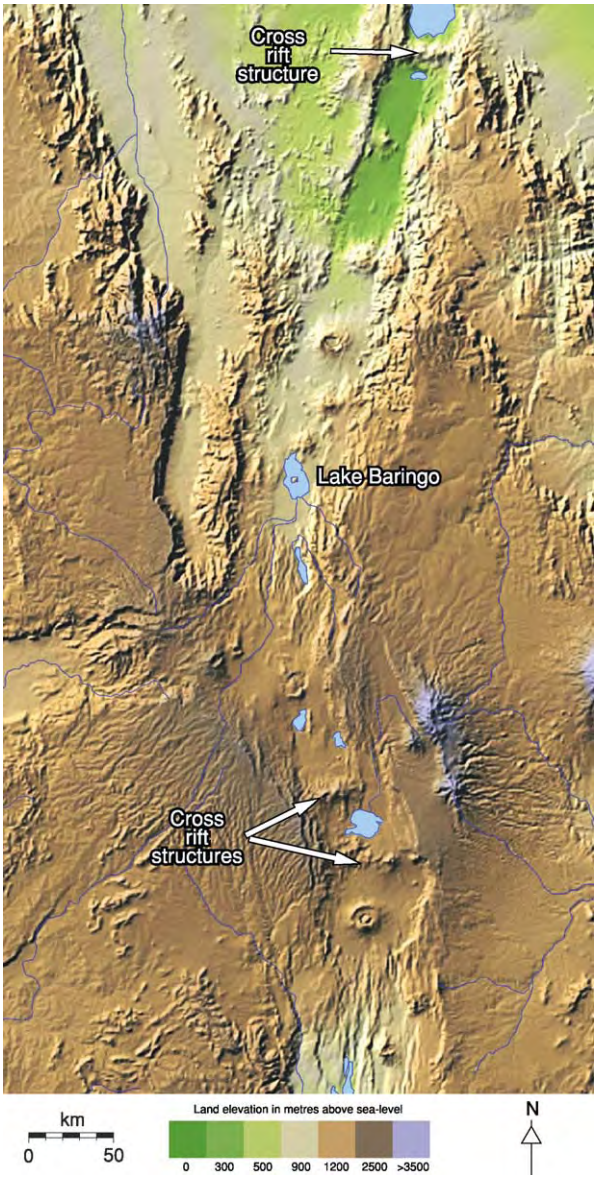


Figure 4 Cross rift structures seen clearly in this satellite image of the East African Rift. This is a shaded relief map produced from SRTM30 data with colour added to indicate land elevations.

Figure 5 Satellite remote sensing image of the northern Kenyan Ethiopian section of the East African Rift showing a string of separate lakes. Image taken by the TERRA satellite using the MODIS instrument (moderate resolution imaging spectroradiometer) and enhanced with SRTM30 shaded relief.

why many early hominid remains have been located in rift sediments. Of particular importance in this regard are the deposits in the East African Rift, which have been excavated and studied by many palaeoanthropologists, including successive members of the Kenyan Leakey family.

The development of the new topographical features of a rift will disrupt pre-existing continental drainage patterns and result in a very different landscape. Before rifting the topography is often subdued, with a very small number of large and ancient rivers draining towards the continental margin. The impact of rifting on the rivers will depend on their orientation. If the new rift parallels existing rivers, it can capture most or all of them, but if it cuts across the general trend of the drainage, streams can be beheaded, diverted, or even reversed. Domed sections of rifts are particularly effective at diverting rivers. The effect is similar to that of piling up soil into a heap so that water falling onto it is shed in all directions, giving rise to a radial pattern of channels.

Active faults form steps in the landscape, which can also affect rivers. Where there is more than one parallel fault, rivers can be caught between the two steps and flow parallel to the rift axis for many hundreds of miles. Margin uplift and tilting of fault blocks both tend to divert rivers away from rift basins. Drainage that has been affected in this way is often captured by continental river systems; for example, the Nile was augmented by the waters of rivers that were diverted away from the Red Sea and Gulf of Aden rifts. In some rifts cross-rift structures form topographical barriers that pond up drainage systems to form lakes. Good examples of this aspect of rifts are Lakes Baikal and Magadi in the Baikal and East African Rifts, respectively (Figure 5). In some rifts

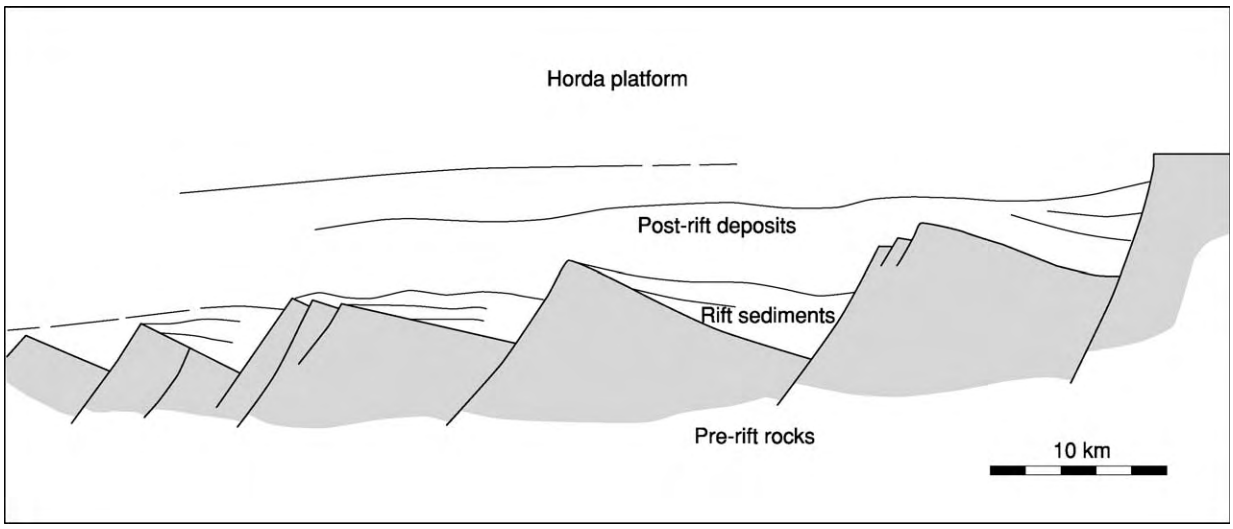


Figure 6 Diagrammatic cross section of the rift and post rift sedimentary fills of the Horda platform in the Norwegian North Sea. Note the tilting of the lower deposits, labelled 'rift sediments', which were emplaced during the period of active rift faulting. Adapted from Steel R and Ryseth A (1990) The Triassic Early Jurassic succession in the northern North Sea: megasequence stratigraphy and intra Triassic tectonics. In: Hardman RFP and Brooks J (eds.) *Tectonic Events Responsible for Britain's Oil and Gas Resources*. Geol. Soc. London Spec. Publ. 55: 139 168.

there are no large lakes, and in these there is often a large river that drains the length of the valley, for example in the Rhine and Rio Grande Rifts. Fault scarps are rarely crossed by large rivers, but small streams flow down the step created by the fault, cutting small steep valleys and depositing sediment in alluvial fans at the base of the slope.

Rifts as Sedimentary Basins

In an active rift the floor of the basin can continue to subside for millions of years. Throughout that time, sediments will be swept into the basin to fill the hole being created and will build up into a sequence of deposits with common characteristics that make it possible to identify a rift basin long after the surface expression has disappeared. The main characteristics are wedge-shaped deposits thickening towards the main border fault and river or lake deposits in the deepest part of the basin flanked by alluvial-fan and river deposits on either side. As subsidence is often asymmetrical, lower layers become progressively tilted relative to the surface (Figure 6) as whole sequences of sediment wedges are superimposed on each other. If subsidence ceases, the structure will eventually lose its topographical expression and the basin will stop accumulating sediments. Rifts can remain active for very long periods of time and will therefore accumulate many

Figure 7 Satellite image of the salt pans at the southern end of the Dead Sea. The very saline waters of this lake are directed into shallow pans and evaporated to extract salts rich in bromine. This is a LANDSAT 7 ETM image using bands 3, 2, and 1 for red, green, and blue colour channels.

kilometres of sediment. Good examples of ancient buried rifts are the Reconcavo basin in Brazil, the Jeanne d'Arc basin off the coast of Newfoundland, and the Triassic Richmond and Taylorsville basins of eastern Virginia, USA.

Economic Deposits in Rifts

The sedimentary sequences that fill rifts can host a range of economic deposits, which are exploited around the world. Of these, the most significant are oil and gas; for example the Jurassic basins of the North Sea between the UK and Norway and the Tertiary basins of the Red Sea are rifts that contain appreciable reserves of oil and gas (Figure 2). The oil is derived from organic-rich fine sediments that accumulated in the deeper quieter part of the basin and which, when buried and heated, migrated into adjacent deposits and filled the pore spaces. Salts that accumulate in both marine and nonmarine rifts have a long history of exploitation; for example, the bromine-rich salts of the Dead Sea Rift (Figure 7) have been extracted since the Bronze Age – a period of more than 4000 years. The processes of rock erosion and sediment deposition can lead to the enrichment of parts of the new rift deposit with heavy minerals such as metals and gemstones, and the sediments themselves can be an extractable resource, particularly if the rift is close to areas of human population where there is a demand for sands and gravels as building materials.

Perhaps the most significant property of rifts is the spectacular scenery associated with them, which attracts tourists and is economically significant. One good example of this is Death Valley in the USA (Figure 8), which is in an area of low rainfall and as such has sparse vegetation, allowing the structure to be seen very clearly even by non-geologists.

Figure 8 Satellite image of Death Valley, USA, showing clearly the fault scarps and surface deposits. This is a LANDSAT 5 TM image using bands 3, 2, and 1 for red, green, and blue colour channels.

See Also

Africa: Rift Valley. **Geomorphology**. **Igneous Rocks:** Carbonatites. **Petroleum Geology:** Reserves. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings; Lake Processes and Deposits. **Tectonics:** Faults; Mid-Ocean Ridges.

Further Reading

Allen PA and Allen JR (1990) *Basin Analysis Principles and Applications*. Oxford: Blackwells.

Frostick LE and Reid I (1989) Is structure the main control on river drainage and sedimentation in rifts? *Journal of African Earth Sciences* 8: 165 182.

Frostick LE and Steel RJ (eds.) (1993) *Tectonic Controls and Signatures in Geological Successions* Special Publica tion 20. International Association of Sedimentologist: Blackwell Scientific Publications.

Gawthorpe RL and Hurst JM (1993) Transfer zones in extensional basins their structural style and influence on drainage development and stratigraphy. *Journal of the Geological Society* 150: 1137 1152.

Gupta S and Cowie P (2000) Processes and controls on the stratigraphic development of extensional basins. *Basin Research* 12: 185 194.

Lambiase JJ (ed.) (1995) Hydrocarbon habitat in rift basins. Special Publication 90. London: Geological Society.

Leeder MR, Mack GH, and Salyards SL (1996) Axial trans verse fluvial interactions in half graben: Plio Pleistocene Rio Grande rift, New Mexico, USA. *Basin Research* 8: 225 242.

Palmason G (ed.) (1982) *Continental and Oceanic Rifts*. Geodynamics Series A, volume 8. Washington: American Geophysical Union.

Selley RC (ed.) (1997) *African Basins. Sedimentary Basins of the World 3*. Amsterdam: Elsevier.

Summerfield MA (1991) *Global Geomorphology: An Intro duction to the Study of Landforms*. Harlow: Longman.

TEKTITES

G J H McCall, Cirencester, Gloucester, UK

Introduction

Tektites, natural glass objects of unknown origin, were described in the tenth century in China and in the eighteenth century in Europe; Charles Darwin described one of these objects from Western Australia while on the Beagle, and four strewn fields were recognized by the twentieth century. An immense volume of work in the past 50 years, involving many state-of-the-art laboratory techniques, has conclusively shown that tektites are the product of a handful of the ~170 terrestrial large-scale events in the geological record. A number of similar events that produced analogous, but somewhat different, glass bodies are also recognized in the geological record, ranging back through the Pleistocene, Pliocene, Oligocene, Eocene, Cretaceous–Tertiary (K–T) boundary, Late Devonian, and possibly even the Archaean. The primary processes of formation of the splash forms of tektites (projection from the target site) and the secondary ablation process that produces flanged, glassy objects (on descent to the strewn field) are now well understood, but much is still unresolved concerning the processes at the target site and the manner of dispersion; the largest of the strewn fields represents a dispersion that covered one-tenth of Earth's surface with glassy objects.

Historical

For centuries, tektites were one of the mysteries of geology. Found in certain sediments of Cenozoic or Quaternary age or strewn on the surface of present-day salt lakes or sand dunes, restricted to four well-defined strewn fields of different geological ages, tektites were once the subject innumerable hypotheses, invoking such visitations as atmosphere-grazing or-skipping comets and lunar ejections. The late twentieth century saw the advent of the Space Age and intensified research showed conclusively that tektites are terrestrially sourced and are the product of immense-scale impact events. Further occurrences of related objects in rocks of Pleistocene, Pliocene, and Eocene ages, at the Cretaceous–Tertiary (K–T) time boundary, in the Late Devonian and possibly in the Precambrian, were recognized at the same time, and research on seafloor sediments in drill cores revealed microscopic tektites (microtektites) associated with three of the four originally defined strewn fields.

In Europe, Cro-Magnon man valued tektites 30 000 years ago, using the glass for ornaments. The Chinese recorded the occurrence of these objects in the *Record of Heterodoxy outside Nanling*, compiled in the tenth century. The objects found strewing fields in the Leizhou Peninsula were attributed to thunderstorms, being termed *Lei-gong-mo* or *Lei-gong-shih* (the 'ink stocks' or 'stools' of the 'thunder god') (**Figure 1A**). Tektites from the Austrian empire (the present-day Czech Republic) were described by Josef Mayer in 1788. Because of the glass industries there, a connection with prehistoric glass making was suspected. Charles Darwin was shown a tektite at Albany, Western Australia, in the 1830s and deemed it to be a volcanic 'bomb' (**Figure 1B**). In the 1930s, Alfred Lacroix described occurrences of these objects in Indochina and the Ivory Coast, and Henryk Stenzel described tektites from Texas.

Explanations advanced for the formation of tektites have included lightning strikes, volcanic eruption, prehistoric and historic glass manufacture, burning coal seams, and desiccation of silica gel masses. Extraterrestrial origin related to meteorites was advanced first by Charles Walcott in 1898 and Eduard Suess in 1900, and this was followed by a rash of extraterrestrial hypotheses (for example, oxidation of the tail of comets; plastic seepings from meteorites passing through the atmosphere; debris from an Earth-like celestial planet blanketed by sedimentary rocks; light-metal meteorites plunging into the atmosphere and producing glass; cometary flight skipping through Earth's atmosphere). Ejected material following lunar impacts became popular as a proposed source of tektites from 1940 up to the time of the Apollo XI landing and the Luna unmanned sample recovery mission, which together dealt a death blow to such provenance because of the data on the silica-poor nature of the lunar surface rocks. Spencer, in 1933, appears to have been the first scientist to come to the right answer – "some form of spallation of melts from impacts on the Earth". In 1962, Ross Taylor, in Australia, concluded from the isotopic evidence that the material of tektites came from Earth, and the immense volume of research carried out on tektites since the Apollo XI landing has confirmed this relationship; besides geochemical, mineralogical, and isotopic evidence, the Ivory Coast strewn field has been conclusively related to the Bosumtwi, Ghana, impact crater, and the Central European strewn field has been related to the Ries

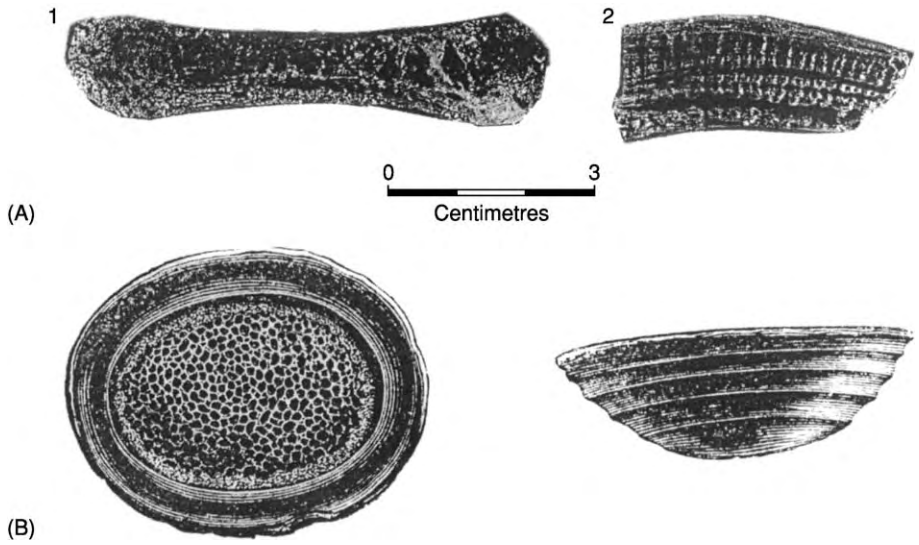


Figure 1 (A) Tektites from China, showing the (1) dumbbell shape (2) layered, Muong Nong type. (B) Charles Darwin's drawing of a flanged button australite: note the ring waves on the anterior surface. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

impact structure in south-east Germany (though there remain some aspects of this relationship still difficult to explain).

Strewn Fields

There are four strewn fields, all of which are of Late Eocene age or younger. However, there are other occurrences of related glassy objects and two of anomalous natural glasses, both of which are not tektites by definition, but are almost certainly closely related to them. The four strewn fields are summarized in Table 1; their distribution is shown together with other related occurrences in Figure 2.

Tektites commonly display primary regularly shaped 'splash' forms characteristic of spinning masses of melt solidifying while travelling through the atmosphere following ejection (though some are irregularly shaped); they may also have superimposed on these forms secondary shapes that were the product of ablation during their descent through the atmosphere hundreds or thousands of kilometres away from the target area. They may then be further modified by terrestrial agencies after falling to Earth.

North American Strewn Field

In North America, tektites were first discovered in Grimes County, Texas, USA, where they occur in the Eocene Jackson Formation and overlying Pleistocene gravels: they were called 'bediasites', a name derived from the Bidai Indians of south-eastern Texas. Bediasites consist of black glass, commonly deeply etched (Figure 3A), and shapes include splash-form spheres, discs, teardrops, and peardrops. Secondary ablation shapes are extremely rare, but at least one example has been reported (Figure 3B). Tektites subsequently recovered from Dodge County, Georgia, are similar, and the single recovery from Martha's Vineyard is like these, rather than those from Texas. All specimens yield the same radiometric age (K/Ar, Ar/Ar, or fission track), dating to 34.9 Ma, consistent with their presence in Eocene sediments. Microtektites of the same radiometric and microfossil controlled age are known from a number of drill cores in the Caribbean, and both microtektites and tektite fragments have been found in Uppermost Eocene sediments on land in Barbados; microtektites of this age have also been found in a drill core from as distant a site as the Weddell Sea, Antarctica. The source of this strewn field is believed to be the 85-km-diameter Chesapeake Bay impact structure in Delaware, which is of the right age.

Central European Strewn Field

Tektites have long been known to occur in the Czech Republic, Austria, and near Dresden and Kottbus in Germany, where they are called 'moldavites'. Many are of a greenish, translucent tint, and these have been used as gemstones. They occur in Miocene sediments, consistent with their radiometric age dating to 15.1 Ma, but also are found in reworked Pleistocene gravels. They exist in a variety of splash forms (ovoids, discs, teardrops, and rods) and their surface is commonly rough (Figure 4). Again, secondary shapes due to ablation are extremely rare.

Table 1 Tektite distribution

Strewn field	*Area covered*	*Age (isotopically determined)*
North American (~4000 tektites recovered)	Texas, Georgia, Martha's Vineyard, Barbados (microtektites in the Caribbean and Weddell Sea)	34.9 ± 2.5 Ma (Eocene)
Central European (55 000 tektites recovered)	Czech Republic, Austria, Germany	15.1 ± 0.1 Ma (Miocene)
Ivory Coast (~200 tektites recovered)	Ivory Coast (microtektites offshore West Africa)	~1 Ma (Pleistocene)
Australasian (South east Asia, >600 000 tektites recovered; Australia, ~100 000 tektites recovered)	China, Indochina, Thailand, Malaysia, Philippines, Indonesia, Australia, Central Indian Ocean (microtektites over wide area of the Indian Ocean, China Seas, and around Australia)	0.77 0.78 Ma (Pleistocene)

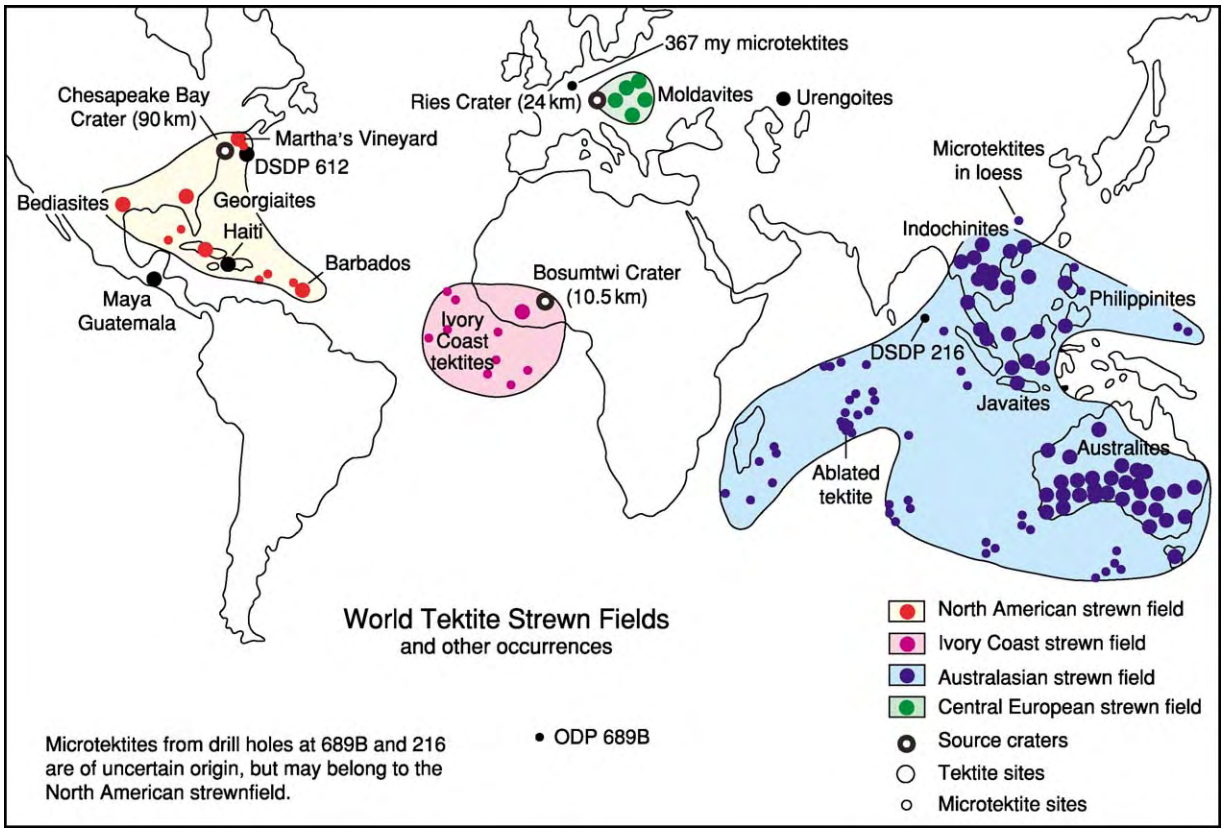


Figure 2 Strewn field distribution of tektites and sites of related occurrences. Deep Sea Drill ng Project (DSDP) and Ocean Drilling Program (ODP) sites are also indicated. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publish ng House, and McNamara K and Bevan A (2001) *Tektites*. Perth: Western Australian Museum.

Microtektites are unknown in this strewn field. The source is believed to be the 24-km-diameter Ries impact structure in south-east Germany, which is of the correct radiometric and stratigraphic age, but there are some unsolved questions relating to how exactly they were expelled and what rock formation supplied the material to form the melt and subsequent glass.

Ivory Coast Strewn Field

A very small number of tektites have been recovered from an area to the west of the Camoe River, near Ouelle, in the Ivory Coast, West Africa. They occur in surficial alluvial deposits over Precambrian (Birrimian) rocks in a gold-mining area. They are of black pitted glass and several splash forms are

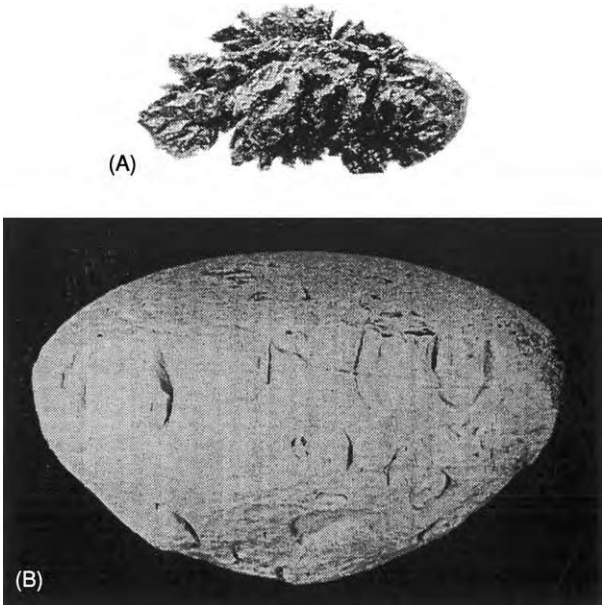


Figure 3 Two bediasites from Texas, USA. (A) Specimen deeply etched by terrestrial agencies. (B) Specimen showing secondary ablation (maximum dimension, 25 mm). Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

recorded (discs, spheroids, teardrops, and peardrops) (Figure 5). Radiometric ages are ~1 million years. An extensive area of the Atlantic Ocean off the coast of West Africa has yielded numerous microtektites in drill cores, and these occur close to the Jaramillo Magnetic Reversal dated at 0.97 Ma, consistent with the radiometric age of the tektites. This strewn field is attributed to the Bosumtwi Crater to the east, in Ghana, as the source impact structure, this being confirmed by radiometric dating, geochemistry, and isotopic methods. Very sophisticated methods related to Os/Re isotopes have shown that the bulk of the osmium in the tektites is extraterrestrial.

Australasian Strewn Field

The Australasian strewn field covers one-tenth of Earth's surface. The northern part of the field includes China, Indochina, Thailand, Malaysia, the Philippines, Borneo, and Indonesia (especially Java, Belitung, and Flores), and the microtektites from marine cores adjacent to these territories are also included. Though several names exist for these tektites (e.g., indochinites, thailandites, javanites, billitonites, rizalites, and philippinites), radiometric dating yields the same age of formation (0.77–0.78 Ma) for all specimens; thus they all represent the same event and are best referred to as 'South-east Asian tektites'. They differ greatly in physical form over the geographical

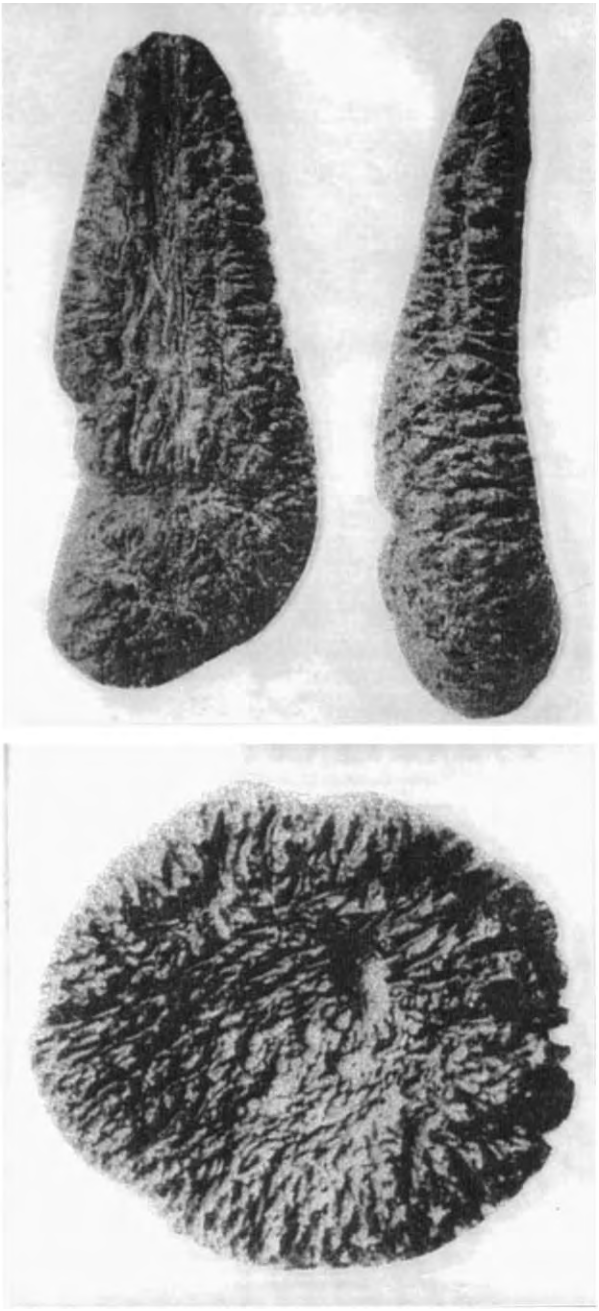


Figure 4 Moldavites from the Czech Republic, showing two teardrops and a discoidal form. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

range in which they are found; these differences are explained by climatic influences and the projected swathe of their descent to Earth. For example, the moist climate in the north part of the strewn field can be correlated with much more pitted, etched, and grooved surfaces, and the comparative lack of secondary ablation forms of northern specimens can be contrasted with specimens from the (distal) south.

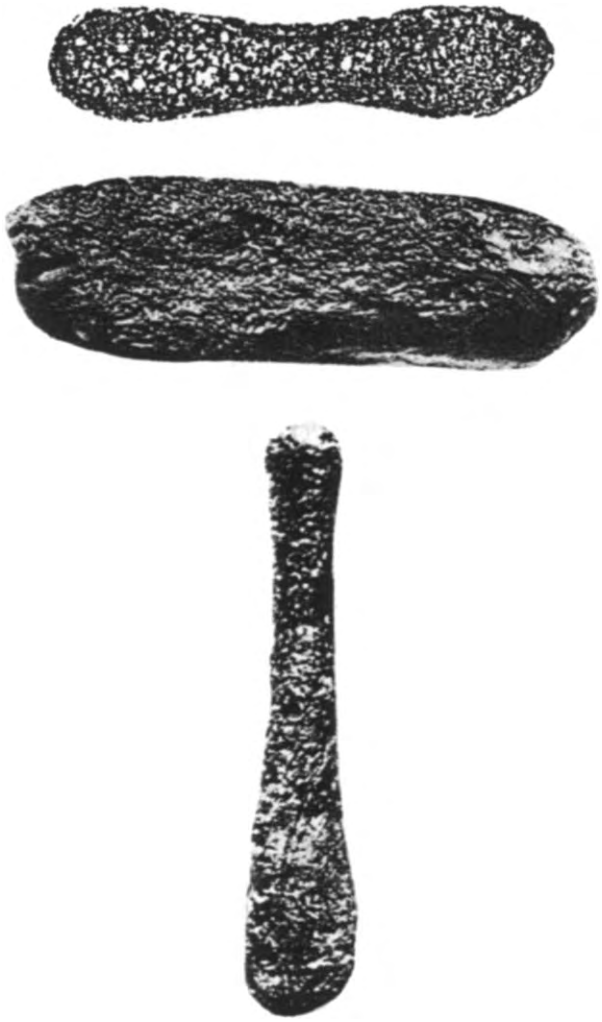


Figure 5 (Top) Dumbbell shaped tektite, 9 cm long, from the Ivory Coast; (middle and bottom) tektites from Indochina, showing the boat shape and teardrop form. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

Most specimens exhibit only splash-form shapes (spheres, ovoids. discs, dumbbells, boat shapes, teardrops, and peardrops). They typically occur in superficial deposit profiles and may be above or below laterite horizons.

Australasian strewn field tektites include the layered and irregularly shaped Muong Nong-type tektites, which were first identified in Laos. These may weigh up to 24 kg and they occur over hundreds, if not thousands, of kilometres, across the strewn field in Indonesia and Thailand (and possibly the Philippines), and the manner of their transport such distances from their source remains unexplained. These tektites also include suites of relict heavy and refractory minerals familiar in terrestrial sediments (quartz, corundum, monazite, rutile, zircon, chromite, andalusite, sillimanite, and kyanite). Lechatelierite and coesite are reported; the former is never found in volcanic glasses and the latter is a polymorph of quartz associated with high pressures in impact explosion processes, though it can occur in tectonic extreme pressure situations (*see* **Ultra High Pressure Metamorphism**). There are intermediate layered tektites in the Indochina collections, but these are splash-form shaped and are not irregular. Muong Nong-type tektites are of rare occurrence in the North American and Central European strewn fields; are not represented at all in the Ivory Coast strewn field or in the southern Australian part of the Australasian strewn field. Immense numbers of tektites may be recovered from a single site in Indochina, and several hundred thousand are reported from Da Lat alone.

The southern part of the Australasian strewn field covers the Australian continent. The tektites here show all the well-known splash forms, but ablated forms of these, particularly perfectly flanged buttons of relatively smooth black glass, are not uncommon (**Figure 6A**), though the flanges break off easily, leaving chatter-marked collars where the flanges have separated (**Figures 6B** and **12**), thus producing the most common form ('cores'). Dumbbells, boat shapes, and teardrops may also show ablation flanges (**Figure 7**). Some australites have ablated away so much of the body of the splash form that they are preserved as flat discs (**Figure 8**). Radiometric dating yields the same age as for South-east Asian tektites (0.77–0.78 Ma), and the specimens represent the same event. It has been claimed, based on stratigraphic evidence, that many Australian tektites fell to Earth very much later than this, and that an 'age paradox' is at work, but this is now refuted (though radiometric dating does indicate that there is a small cluster of older ~10-million-year-old Na-rich tektites near the Western Australia/South Australia border). A feature of the Australasian subfield is that large splash forms cluster around certain localities, an unexplained development. Australites are mainly found on the surface of salt-pan lakes in Western Australia, where they are washed in; they are also found in diamondiferous deposits in the north of Western Australia, on and between sand dunes in South Australia, and in Quaternary sediments in Victoria.

Microtektites were first discovered in the seas off the coasts of South-east Asia and Australia. Their stratigraphic and radiometric ages are consistent with being part of the strewn field. Some contain coesite and stishovite, both of which are high-pressure polymorphs of quartz associated with impact explosion processes; stishovite is found only associated with these types of processes. A single tektite has been

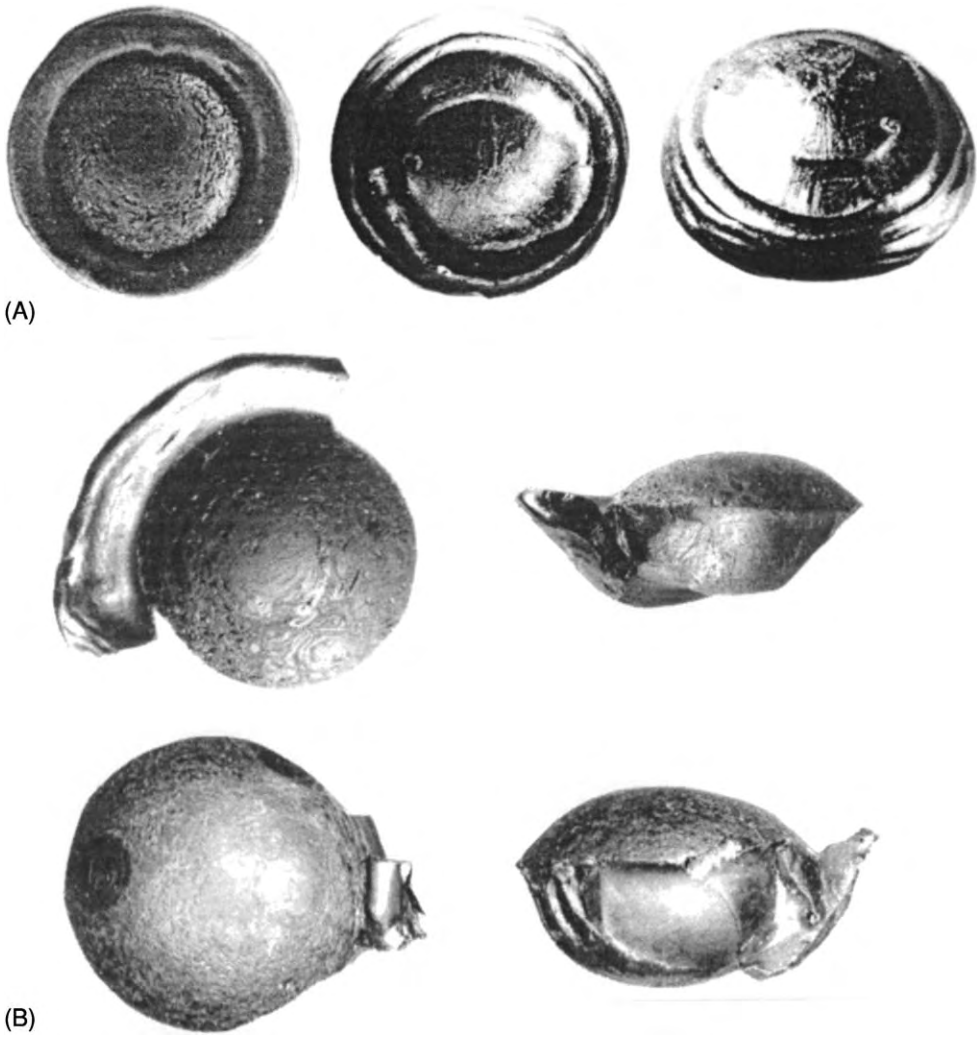


Figure 6 Australites from the Finke River, Central Australia. (A) Three views of a perfect flanged button, showing a posterior surface in flight and two anterior surfaces with ring waves produced by ablation (maximum dimension, 20 mm). (B) Four views of australites with flanges partly preserved as they separated from the remnant 'core' (maximum dimension, 17.5 mm). Photographs by WH Cleverly; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

recovered from a grab sample in the middle of the Indian Ocean (Figure 9) and is attributed to the same event leading to the other Australasian strewn field tektites. Microtektites have been recovered from cores all the way across the Indian Ocean, to sites not distant from Madagascar.

The source of the Australasian tektites remains a complete mystery. Tonle Sap Lake in Cambodia was investigated as a possible impact site, but no evidence was found of impact there. Studies of progressive changes in populations of microtektites and delineation of the restricted area of occurrence of coesite and stishovite content suggest a source in Indochina, in Cambodia, not far from Tonle Sap. The whole character of the strewn field suggests that China and Indochina are proximal to the source, and Australia is distal. It is inexplicable that the source structure, of such a geologically young age, and presumably larger than the 85-km-diameter Chesapeake Bay structure, in view of the strewn field dimensions, is not preserved to some extent. This has led to suggestions that the Australasian tektites derived from an explosive event in the atmosphere, as has been widely accepted for the Tunguska event in Siberia in 1908.

Microtektites

These microscopic glass bodies are seldom larger than 1 mm in diameter. They display the familiar splash forms in miniature (mostly spheres, but also

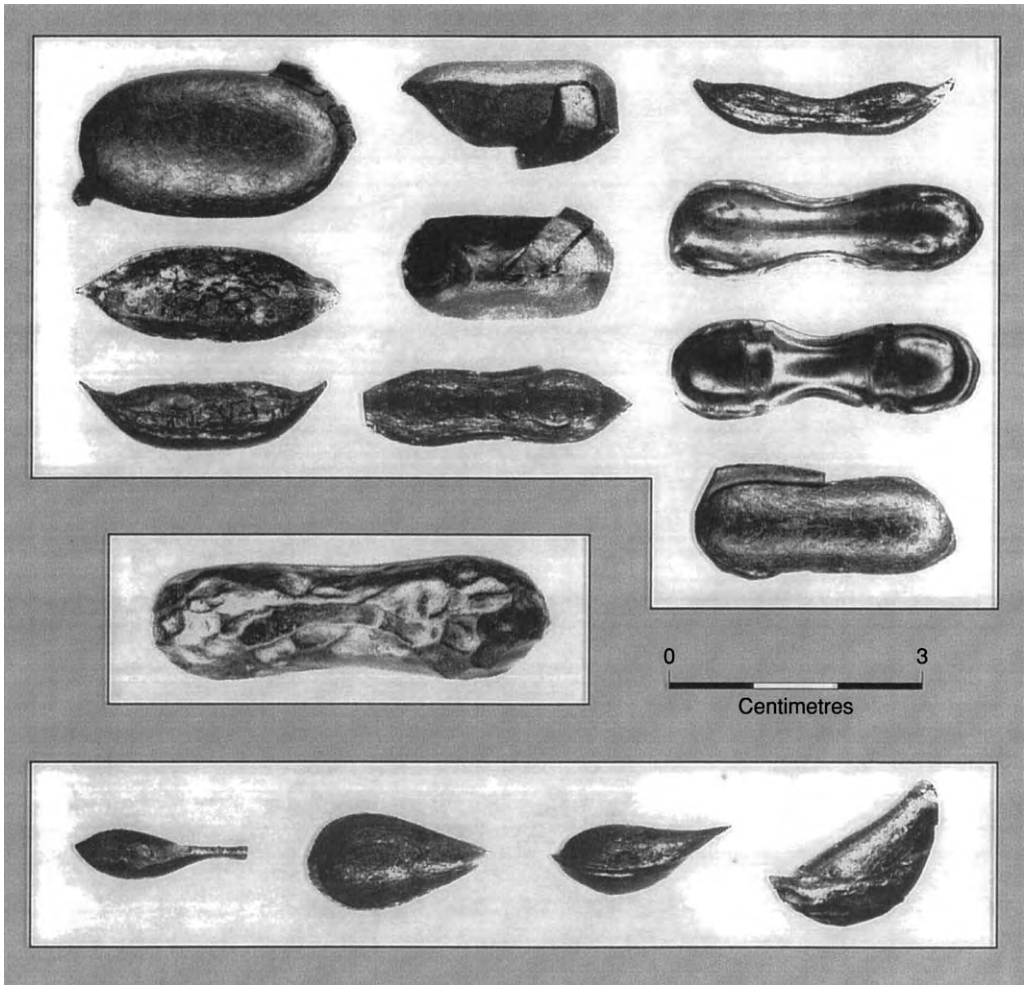


Figure 7 Boat shaped, dumbbell, and teardrop australites from Menangina and Gindalbie Sheep Stations, Western Australia, showing flanges in various stages of detachment from the 'cores'. Photographs by WH Cleverly; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

dumbbells, teardrops, peardrops, and irregular forms) (Figure 10A–C). They occur in deep-sea cores, but the horizons are commonly not entirely sharply defined because of the action of seafloor scavengers. Microtektites do not normally occur in soils on land, because, even in the time-span in which the australites have littered Australia (from 0.77 to 0.78 Ma) on land, the action of groundwater would have dissolved them. The one known occurrence on land is in Eocene sediments in Barbados.

Tektite Composition, Experimental Data, and Theoretical Considerations

Tektites are silica-rich glasses, ranging from an average of 68% (Ivory Coast) to 80% (North American, georgiaites) silica. Refractive indices range from 1.48 to 1.51 and specific gravities from 2.3 to 2.5. Each strewn field has its peculiar range of values. The whole-rock analyses have a different character, as compared to volcanic glasses, and the glass is almost anhydrous, in contrast to the values for volcanic glasses. Microtektites show greater compositional variation than do the tektites found on land because they have a much smaller volume and can be formed of anomalous fractions of the glass, whereas the larger tektites average out. Lechatelierite, coesite, and stishovite are commonly present in tektites and microtektites, whereas they are unknown in volcanic glasses.

Wind-tunnel experiments in 1963 on tektite glass and gelatine by Chapman and Larsen in the United States reproduced exactly the flanged button form of ablated australite tektite found in Australia, and left no doubt concerning the mode of origin of the australites (Figures 11 and 12). Theorizing on the origin of tektites in 1998, H Jay Melosh concluded that

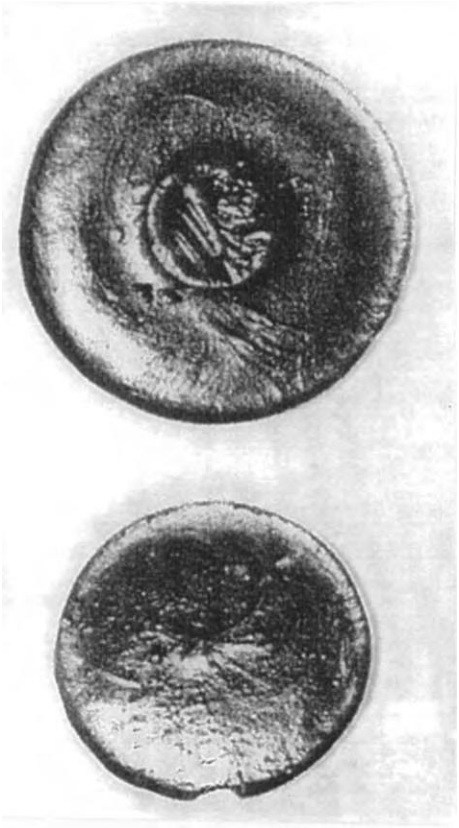


Figure 8 Ablated spherical australites that have lost all but a small relic of the splash form sphere, due to flowage of melt on ablation to the enlarged flange. This type of tektite is found only at Port Campbell, Victoria at the extreme distal end of the strewn field. Photograph by G Baker; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

immense shock pressures of 100 GPa occurred during impact explosion processes accompanied by temperatures of up to 50 000 K. Thus arguments related to requirements for high temperatures in glass technology, which have been advanced to refute terrestrial impact generation of tektites, are invalid, because the temperatures involved in impact processes are many orders greater than had been thought.

Related Occurrences

Libyan Desert Glass and Mount Darwin Glass

Anomalous natural glass objects have been found in an area in the extreme south-west of Egypt. Irregular masses weighing up to 800 kg strew the desert in interdune corridors. These objects are highly siliceous, lechatelierite-bearing glass, quite unlike tektites in that they contain 98.2% silica. They seem to be ejected impactites formed of Nubian Sandstone, thrown out

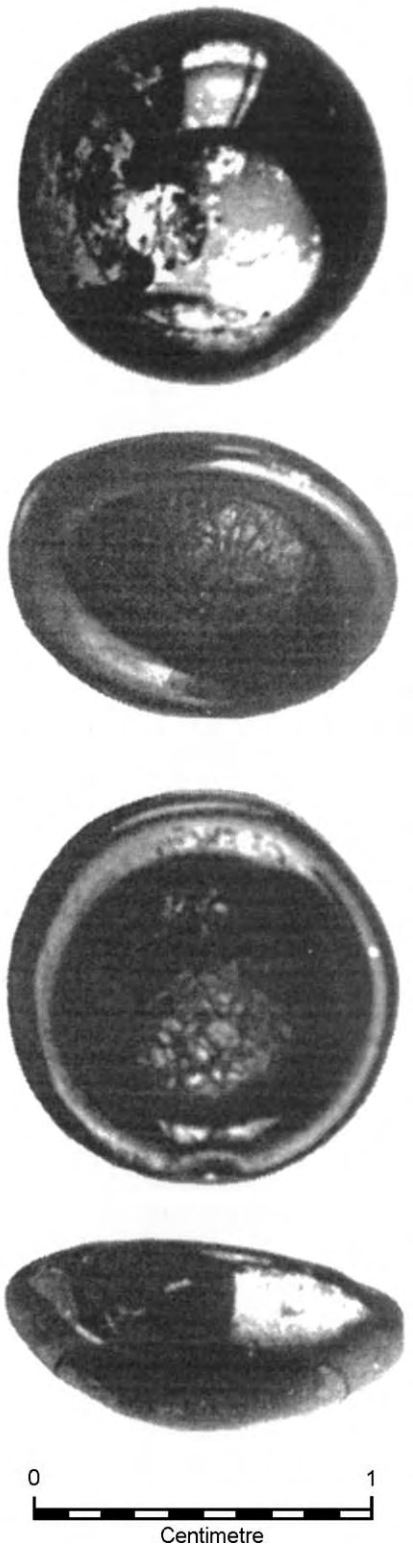


Figure 9 An ablated discoidal tektite with a weakly developed flange from a grab sample in the Central Indian Ocean. Photograph by BP Glass; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

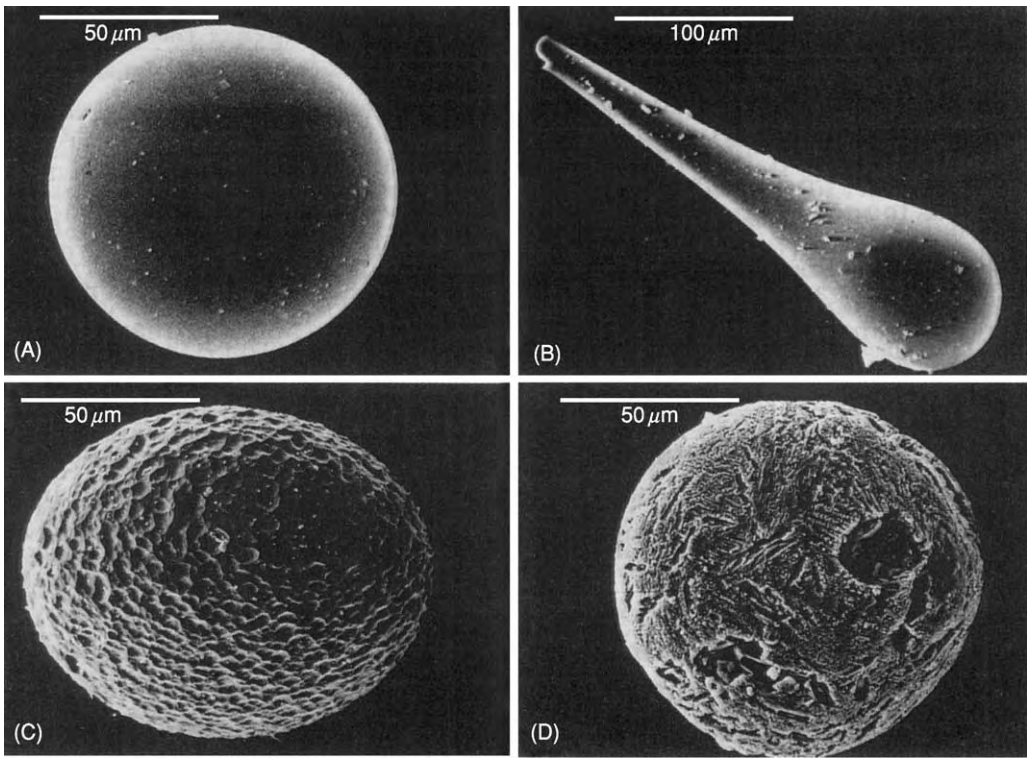


Figure 10 (A C) Microtektites and (D) microcrystite (clinopyroxene microspherule), all from Deep Sea Drilling Project 689, Maud Rise, Weddell Sea. The microtektites are equated with the North American strewn field and the microcrystite is equated with the Popigai impact structure, northern Siberia. Reproduced with permission from Glass BP and Koeberl C (1999) Ocean Drilling Project Hole 689B spherules and Upper Eocene microtektite and clinopyroxene bearing spherule strewn fields. *Science* 34: 197 208.

during the Oligocene 29 million years ago from two small craters (named BP and Oasis) that are situated about 100 km to the west. A minute fragment of an iron meteorite has been found with these glass objects. The glass was used to make Acheulian scrapers.

Mount Darwin Glass

Irregular masses of anomalous glass have found on the surface in western Tasmania and near Mt Macedon in Victoria. The glass mass is layered, not unlike Muong Nong-type tektites, but the silica content is much higher (88%, compared to 73%). Radiometric dating indicates formation 0.73 Ma, but the occurrence of these masses clearly has nothing to do with australites. A 1000-m-diameter crater has been recognized close to the occurrences of these glass objects in Tasmania, but there is no evidence of impact on excavation, though this site may be the source of the glass. The similar glass found at Mt Macedon, across the Bass Strait, 560 km to the north, is unexplained.

Zhamanshinites and Irghizites

Slags and glasses have been reported from the 13.5-km Zhamanshin impact structure in Kazakhstan, north of the Aral Sea (see Figure 2), overlying Palaeogene country rock. The best radiometric age derived dates these objects to 1.09 Ma. Zhamanshinites contain rock fragments and are impactite glasses. Irghizites occur within the bounds of the crater structure, not in an external strewn field, and are composed of small 'micro-irghizite' particles, welded together. The silica content of the irghizites is 72–79%, not unlike that of tektites, but the water content is slightly higher. The interest in these objects is that they may represent the separation of microtektites at source.

Urengoites

Three fragments of tektite-like glass have been found buried in Siberia at two sites 40 km apart (see Figure 2); based on radiometric dating, these urengoites were formed at 22–24 Ma.

The Eltanin Glasses

The 25-km-wide Pliocene Eltanin structure on the floor of the southern Pacific Ocean (see Figure 2) has both associated minute fragments of a mesosiderite (stony iron) meteorite and microscopic glass spherules; these objects have been recovered in piston

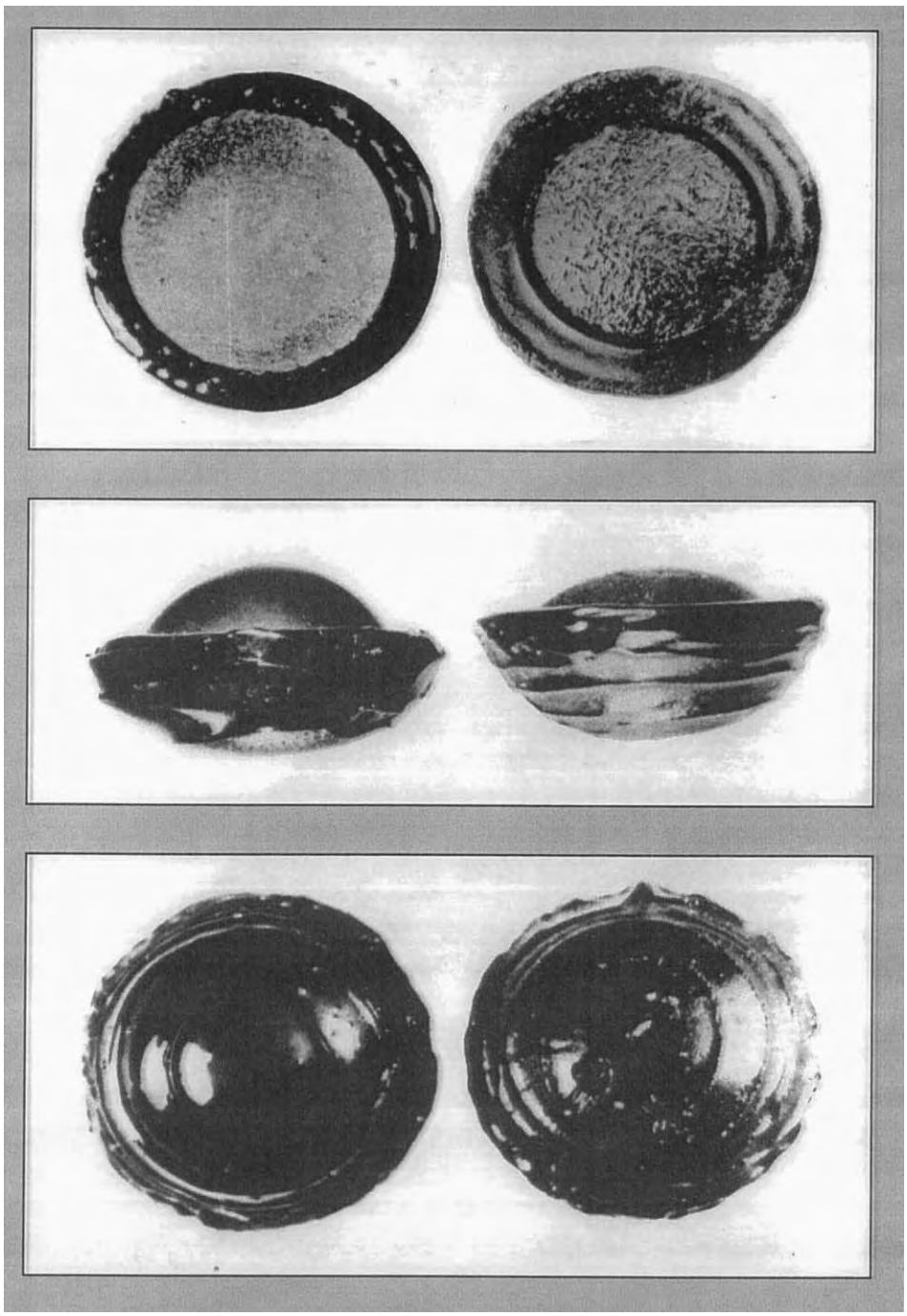


Figure 11 Comparison between a flanged australite from Port Campbell, Victoria (right) and an artificial product of a wind tunnel experiment (left) by DR Chapman and HK Larsen on tektite glass (NASA photograph, 1963). Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

cores close to the impact melt layer associated with a chaotic formation within the structure. The forms resemble microtektites in size and shape (spheres, teardrops), but the chemistry is quite different (silica average, 45%).

Late Eocene Microspherules

A microspherule layer has been recognized just below the North American microtektite layer in Caribbean cores and in the Weddell Sea core. These

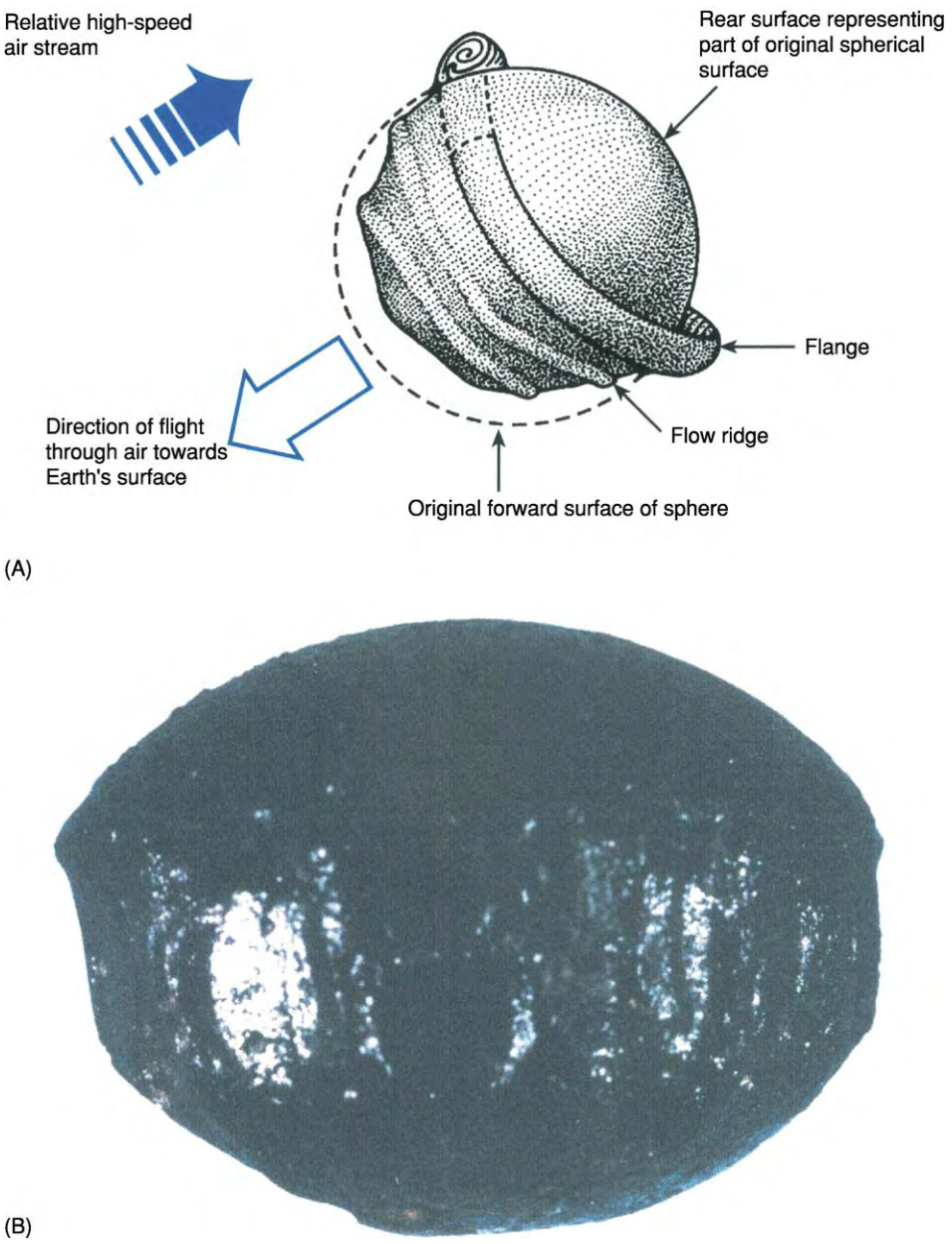


Figure 12 (A) Development of a spherical splash form australite during atmospheric ablation on descent. (B) The common 'core' form produced on separation of the flange, showing equatorial chatter marks where it has separated. Reproduced with permission from McNamara K and Bevan A (2001) *Tektites.* Perth: Western Australian Museum.

spherules differ from the Eltanin glasses in that they show a palimpsest of microcrysts, as well as glass (Figure 10D), and the chemical composition is much more varied. The mineral in the microcrysts is sometimes preserved and X-ray diffraction has shown it to be pyroxene. This layer has been equated with the Popigai impact structure in North Russia and the impact layer reported from Massignano, Italy, and other European sites. The average silica content of 64% is lower than in any tektites and matches closely the Popigai impactite glasses.

K–T Boundary Glass Bodies

Carbonate sediments at the K–T boundary in Haiti (see Figure 2) contain tektite-like bodies of glass enveloped in smectite; the glass objects have the usual splash forms (spheres, ellipsoids, teardrops, and elongate and dumbbell shapes). The glass is vesicular and mostly crystal free. Similar bodies are reported from clastic sediments at Beloc, Mexico (Figure 2), where foraminifera indicate the K–T boundary (*see* **Mesozoic:** End Cretaceous Extinctions). Spheres and

dumbbells are represented. In both cases, removal of smectite coating reveals the sculpturing seen on the surface of microtektites. Compositions are very different compared to those of tektites or microtektites: silica contents are about 63%, similar to melt rocks of andesitic composition revealed at the favoured source impact site, the 170-km-diameter Chicxulub Crater impact structure on land and offshore of Yucatan, Mexico. The contrast to the tektite composition is consistent with the country rock makeup at Chicxulub, and these glass bodies, though not fitting strictly the common definition of tektites, are clearly a form of tektite *sensu lato*. Microspherules are reported from the K–T boundary at Petruccio, Italy, and there are many examples of shocked quartz with planar structures at the same boundary in the United States (*see* **Impact Structures**).

Late Devonian Glass Bodies

Microspherules have been reported from the Senzeilles Shale, Belgium, a quiet-water deposit of Late Devonian age and very close above the Frasnian–Famennian stage boundary, the site of another, smaller scale extinction. The horizon is at the top of the *Palmatolepis triangularis* conodont zone, at the bottom of which is the extinction boundary. The spherules, up to 1 mm in diameter, are wholly of glass; they have been preserved from devitrification perhaps because of the anhydrous nature of the glass. Most of the bodies are spheres, but elongate, teardrop, and dumbbell shapes are present. The silica content varies from 38 to 80% and the chemical composition is variable. There are several possible source impact structures, though the Siljan, Sweden, 55-km structure is favoured.

Similar spherules occur at Qidong in Hunan Province, South China, in the *Palmatolepsis crepida* conodont zone of the Famennian. They reach up to 0.160 mm in diameter and are mostly spheres, though teardrop and peardrop forms are recorded. Lechatelierite is present in them. The chemistry is variable and the silica content ranges from 62 to 99%. They have been tentatively related to a possible impact structure at Taihu, south-west of Shanghai, but little is known of this structure. As for the Senzeilles Shale microspherules, these spherules do not correspond exactly to the extinction horizon.

Archaean Spherules

Microspherules in which no glass is preserved are known from the Wittenoom Formation in Western Australia (dating to 2500–2600 Ma) and also from rocks of the Onverwacht and Fig Tree Groups (3100–3500 Ma) near Barberton, South Africa. These microspherules have been attributed to impact processes and are related to microtektites.

Research Directions

Many state-of-the art techniques have been applied to studies of tektites, resolving many of the early questions about these objects. Geochemical, geophysical, isotopic, and statistical analyses that have been applied to tektites in the past 50 years have resolved questions concerning their source, the manner of formation of their primary and secondary flanged shapes, and the age of the four events forming the strewn fields in which they are found. The connection between tektites and microtektites has also been established. However, there are still questions that have not been fully resolved. Further research may reveal the process of melting at the target, and ejection from the target; the reason for the restricted nature of the target rocks involved in generating tektites; the reason for the restriction of tektite associations to a handful to the ~170 terrestrial megaimpact sites known; the exact relationship of tektites to microtektites in the processes occurring at the target, in flight, and during transport and dispersion to the strewn field; and the source of the Australasian strewn field.

See Also

Analytical Methods: Fission Track Analysis; Geochronological Techniques. **Gemstones**. **Igneous Rocks:** Obsidian. **Impact Structures**. **Mesozoic:** End Cretaceous Extinctions. **Shock Metamorphism**. **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Moon. **Ultra High Pressure Metamorphism**.

Further Reading

Barnes VE and Barnes MA (eds.) (1973) *Tektites*. Stroudsburg, PA: Dowden, Hutchinson and Ross.

Chapman DR and Larsen HK (1963) On the lunar origin of tektites. *Journal of Geophysical Research* 64: 4305 4368.

Glass BP (1968) Glassy objects (microtektites) from deep sea sediments off the Ivory Coast. *Science* 161: 861 862.

Glass BP (1990) Tektites and microtektites: key facts and inferences. *Tectonophysics* 171: 393 404.

Glass BP and Koeberl C (1999) Ocean Drilling Project Hole 689B spherules and Upper Eocene microtektite and clin opyroxene bearing spherule strewn fields. *Science* 34: 197 208.

Glass BP, Chapman DR, and Prasad S (1996) Ablated tek tite from the central Indian Ocean. *Meteoritics and Planetary Science* 31: 365 369.

Koeberl C (1986) Geochemistry of tektites and impact glasses: an overview. *Annual Reviews of Earth and Planetary Sciences* 14: 325 350.

Koeberl C (1987) *Geochemistry of Muong Nong Type Tektites: A Review,* Proceedings of the 2nd International Conference on Natural Glass, Prague, pp. 371 377.

Koeberl C and Shirey MB (1993) Detection of a meteorite component in Ivory Coast tektites with rhenium osmium isotopes. *Science* 261: 595 598.
McCall GJH (2000) The age paradox revisited. *Journal of the Royal Society of Western Australia* 83: 83 92.
McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.
McNamara K and Bevan A (2001) *Tektites*. Perth: Western Australian Museum.
Melosh HJ (1998) Impact physics constraints on the origin of tektites. *Meteoritics and Planetary Science* 33(Supplement): A104.
O'Keefe J (1963) *Tektites*. Chicago: University of Chicago Press.
Taylor SR (1962) The geochemical composition of australites. *Geochimica et Cosmochimica Acta* 26: 685 722.
Taylor SR (1969) Criteria for the source of australites. *Chemical Geology* 4: 451 459.

TERRANES OVERVIEW

L R M Cocks, The Natural History Museum, London, UK

Introduction

The word 'terrane' is used in a specialised sense by geologists, and should not be confused with the same-sounding 'terrain', which is used by many people, particularly the military, to denote characteristics of the countryside in a particular area. To a geologist, terrane is used for a discrete block of continental crust that is moving or has moved in relation to those blocks that surround it.

Definition

The Earth is today, and through geological time, made up of a number of moving plates (*see* **Plate Tectonics**). Each plate consists of heavier oceanic crust underlying lighter continental crust. Plates are constantly being enlarged through ocean-floor spreading, reduced by subduction or obduction, or displaced laterally by transform faulting, all of which processes affect both oceanic and continental crusts. However, because of its lighter density, continental crust tends to remain at the Earth's surface for far longer periods than does oceanic crust; consequently, very often the continental crust of an old plate remains at the surface today long after the oceanic crust on which it once rested has disappeared within Earth's interior, perhaps to be later remobilized into fresh crust. The oldest ocean crust known today in its original position is only about 160 million years old (Jurassic), whereas the continental crust includes rocks from modern times to more than 3 billion years ago, the oldest known. Terranes can be of varied size, ranging today from the vast Eurasian–African block down to the relatively small microplates found in the south-west Pacific within the East Indies. The difference between a 'continent' (as strictly defined) and a 'terrane' is that the former is invariably bounded by one or more oceans, whereas the latter is defined by its surrounding structural discontinuities. An accreted terrane is one that has been added to the margin of a larger one. Many areas may or may not have been real (i.e., separate) terranes in the past, and geological opinions can often differ widely as to their reality and status. When this uncertainty exists, the area is referred to as a 'suspect terrane'.

Boundaries of Terranes

The marginal boundaries of old terranes are termed 'sutures': when exposed, they are usually faults or fault systems, with the obvious characteristic that the rocks and stratigraphy are completely different on the opposite sides of the faults. Because movement of the crust is principally dominated by horizontal components, the suture faults are usually strike-slip or transform. For example, a major suture is the Tornquist–Teisseyre Line, or the Trans-European Suture Zone (TESZ), which stretches from the North Sea to the east of Aberdeen, through southern Denmark, north-eastern Germany, south-central Poland, Slovakia, Hungary, and Romania to the Black Sea. That suture represents what remains of the south-eastern margin of the old terrane of Baltica and separates that terrane from Avalonia (see later), Perunica (often termed Bohemia), and others to its south. The suture was originally formed during the Variscan Orogeny in Late Palaeozoic time, but movements along the TESZ area of crustal weakness have been reactivated during several subsequent geological periods and continue sporadically to the present day.

Principal Terranes

At two recognizable times in Earth history, at about 1000 Ma and 250 Ma, most of the continental

crust was together, forming vast supercontinents named Rodinia (*see* **Precambrian:** Overview) and Pangaea (*see* **Pangaea**), respectively. Prior to the aggregation of Rodinia, little is known of the preceding terranes, and so they are characterised and named only by the major earlier Precambrian shield areas, such as the Canadian Shield, and their positions relative to each other are currently poorly constrained and open to much scientific debate. However, after the breakup of Rodinia, which was well under away by 850 Ma, larger terranes have separate names, and the principal terranes were identified as follows:

- Gondwana. Easily the largest terrane (*see* **Gondwanaland and Gondwana**), comprising South America, Africa, India, Antarctica, and Australia, as well as a number of peripheral areas that formed parts of this huge terrane at different times.
- Siberia (otherwise known as Angara). An area that included only part of the modern political area of Siberia, but that was nevertheless very substantial.
- Laurentia. Most of North America and Greenland, and then adjacent areas, including Spitzbergen, northern Ireland, and Scotland. It was separated from Baltica and Gondwana by the Iapetus Ocean in the Lower Palaeozoic.
- Baltica. The northern part of mainland Europe eastward to the Ural Mountains and northward to include Novaya Zemlya and Franz Joseph Land in the Arctic. It was separated from Avalonia by the Tornquist Ocean and from Siberia by the Aegir Sea during the Lower Palaeozoic.
- Avalonia. An area including the western coast of the United States, the Maritime Provinces of Canada, Newfoundland, southern Ireland, Wales, England, and Belgium, which formed part of Gondwana until the Early Ordovician (about 490 Ma). It was a separate terrane only in the Ordovician. As Avalonia left Gondwana, the widening ocean between it and Gondwana is termed the Rheic Ocean.
- Laurussia. The terrane formed by the amalgamation of Laurentia, Baltica, and Avalonia during the Silurian, and which continued until the formation of Pangaea in the Late Palaeozoic. During the Upper Palaeozoic, Armorica, Perunica, the Rheno-Hercynian Terrane, and others drifted across the Rheic Ocean to become accreted to Laurussia, leaving a widening Neotethys Ocean behind them.
- Armorica-Iberia. This is sometimes termed the Armorican Terrane Assemblage (ATA); it consists of most of the western part of southern Europe, including Spain, Portugal, France, Sardinia, and parts of Germany. Some regard Perunica (Bohemia) as part of the ATA, but it moved independently of the ATA after its separation from Gondwana in the Ordovician.
- North China. The southern part of Siberia and the Korean Peninsula as well as northern China.
- South China. Most of southern China.
- Annamia. The Indochina Peninsula and adjacent areas.
- Sibumasu. The area running from eastern Burma (Myanmar) through Thailand, south-western China, and western Malaysia to Sumatra.

In addition to these named terranes, more than 50 additional terranes of variable size have been identified and named as existing during the Palaeozoic, as well as numerous discrete geological entities such as island arcs, which were independent units for differing geological times. Some of the many terranes that make up Eurasia today are shown in **Figure 1**. After the mutual accretion of the various terranes during the Palaeozoic, progressively forming Pangaea during the Upper Palaeozoic, the process of splitting and disintegration of Pangaea began in the Early Mesozoic to form the larger terranes that are known by their modern continental names today. The large area eastward of Pangaea was occupied first by the Neotethys Ocean and subsequently by the Tethys Ocean. The Mediterranean Sea can be considered as a remnant of the Tethys Ocean today. However, there are in addition many much smaller terranes; for example, it has been suggested that there are as many as 100 Early Mesozoic terranes in the collage that makes up the present-day Cordillera of North America.

Identifying the Positions of Old Terranes

To reconstruct the geography of Earth at different times in the geological past, it is necessary to locate the former position of the different terranes, and to understand how each moved with time. The ways of doing this using current knowledge are through evidence of ocean-floor magnetic stripes, movement over hotspots, palaeomagnetism, faunal provinces, distribution of sediments, and positioning of tectonic belts.

Ocean-floor magnetic stripes are studied by mapping out the modern ocean floor and its magnetic anomaly stripes, dating the stripes, and then progressively removing them so that it can be seen how the oceans have widened with time (*see* **Palaeomagnetism**). Magnetic field reversals and related issues such as the age of onset, the duration, and the frequency of superchrons (long periods of constant magnetic polarity) are now reasonably documented to the beginning of the Cretaceous. This is the only objective

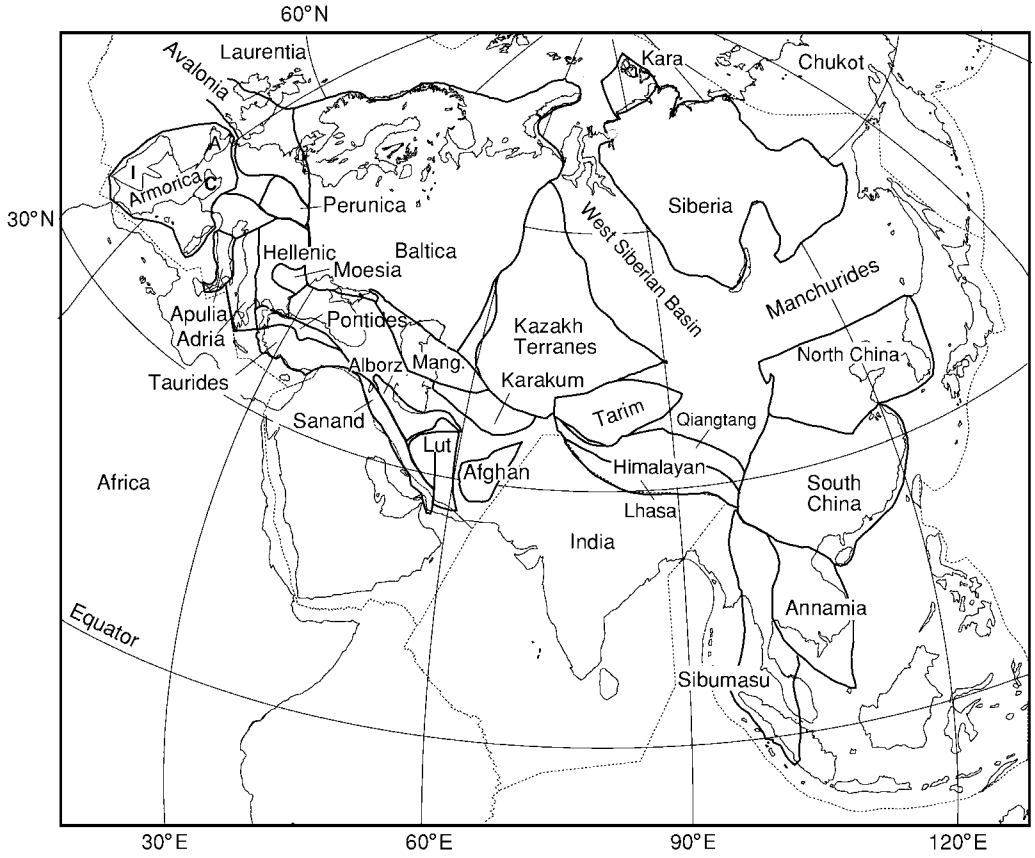


Figure 1 The boundaries of the major Palaeozoic terranes that have united to make up Eurasia today. C, Central France; I, Iberia; Mang., Mangyshlak Terrane. The large areas labelled West Siberian Basin and Manchurides are those occupied largely by continental crust that is post Palaeozoic in origin, as is the large area of south east Asia south eastwards of the Sibumasu and Annamia terranes. Thin dotted lines are modern plate boundaries. Modified by Trond Torsvik, Trondheim, from Torsvik and Cocks (2004).

method of discovering the previous position of terranes through time. Unfortunately, however, because the oldest ocean-floor crust known is of Jurassic age (and most of Earth's oceans floors are very much younger than that), this method is available only for determining terrane positioning in Tertiary and Late Mesozoic times.

Some terranes can be seen to have moved over hotspots with time, which thus gives progressive and definitive positioning of the terranes. This is unique in providing objective (as opposed to subjective) longitudinal control of terrane movement. Unfortunately, as with the plotting of the magnetic stripes, this method applies only to those few terranes that are drifting over active hotspots, and no consensus on results has been obtained for positions much older than the Tertiary. The only Mesozoic data are for the Tristan da Cunha and Great Meteor hotspots in the South Atlantic, which are traceable back to 130 Ma.

Palaeomagnetism occurs when an igneous rock is emplaced, then the magnetic (largely iron) constituents within it cool with their magnetic direction pointing towards the pole. This remnant magnetism is set after cooling, and so the study of an ancient igneous rock indicates the pole position at the time of deposition of the rock. Thus two things can be calculated: the palaeolatitude of the rock at the time of cooling and the subsequent rotation of the terrane. This is the best method of calculating the position of old terranes. There are, however, two drawbacks: first, there is no determination of the palaeolongitude of the terranes, and second, a great many igneous rocks have their original palaeomagnetism completely reset by subsequent tectonic events that involved enough heating to reset the magnetisation of the rock.

If the palaeoecology and the age of the fossils contained within an individual terrane are known, then some marine benthic or terrestrial fossils are found to be specific to one or more terranes, and quite different from those in other terranes that may be close to it today. These differences are often recognised as faunal provinces. The data from these fossil distributions provide terrane affiliations and positions that are reached completely independently of palaeomagnetic methods, and the two methods have

been used effectively together to discover where most of the terranes lay in Cambrian to Jurassic times. In addition, bioherms such as coral reefs are usually restricted to within 30° north and south of the equator, ancient or modern. Coals are most commonly found in two belts occurring in low temperate latitudes.

Distribution of sediments may be examined to locate the former position of the different terranes. Most clastic rocks give little indication of the climates within which they were deposited, but carbonates increase in abundance from high to low latitudes as the average surface temperature increases. Evaporites are seldom equatorial but are most common in two bands centring at about 20° north and south of the equator or palaeoequator. Glaciogenic sediments, such as tillites, are almost invariably found at high latitudes.

Positioning of tectonic belts may provide evidence of the position of old terranes. Some substantial mountain belts can be traced from one terrane to the adjoining one, but, in the past, several such correlations made by geologists have been shown to be incorrect. However, in the Precambrian, when there were no terrane-diagnostic fossils, mountain belts have proved to be the only indicators apart from palaeomagnetism.

In positioning old terranes, a key underlying precept must be kinematic continuity. This means that it must be remembered that terranes never leapt around the globe like sodium on water. Thus, if a certain terrane (for example, Baltica) appears to have been in one place, then 2000 km away 10 million years later, and then close to its original position 10 million years after that, it is probable that one or more of those postulated positions are not correct!

Ever since the acceptance of modern plate tectonic theory in the mid-1960s, geologists have realised that plates and their terranes have not been in a single position during geological time. Tuzo Wilson, in 1966, suggested that there was a substantially different terrane pattern and therefore geography before the supercontinent of Pangaea came together in the Late Palaeozoic. Since Wilson's observations,

Figure 2 The major terranes and Earth geography 400 Ma (the Early Devonian), assuming that Earth's magnetic field was a simple geocentric axial dipole. The dot dash line represents the margin of Gondwana. RH, Rheno Hercynian Terrane. Modified by Trond Torsvik, Trondheim, from Torsvik and Cocks (2004).

there have been many published models of where the various terranes lay through geological time. There is now considerable agreement on the identity, positions, and progression, through the Palaeozoic, of terranes surrounding what is today the North Atlantic area, around which the majority of academic geologists work. However, the many terranes that make up Central and South America, Africa, Asia, and Australasia are, in many cases, rather poorly defined and recognised, and their relative positioning through the Phanerozoic (let alone the Precambrian) is a matter for unresolved debate and geological argument. Figure 2 shows a possible terrane reconstruction for half the globe at about 400 Ma (the Early Devonian), at a time when Laurentia, Baltica, and Avalonia had fused to form Laurussia and when various terranes such as Armorica, Adria, the Pontides of Turkey, and the Hellenic Terrane (including Moesia) had all left the Gondwana superterrane following the opening of the Palaeotethys Ocean to their south. The other half of the globe, not shown in Figure 2, was largely occupied by the vast Panthalassic Ocean.

See Also

Gondwanaland and Gondwana. **Palaeomagnetism**. **Pangaea**. **Precambrian:** Overview. **Volcanoes**.

Further Reading

Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomag netic review. *Journal of the Geological Society, London* 159: 631 644.

Leitch EC and Scheibner G (eds.) (1987) *Terrane Accretion and Orogenic Belts*. Geodynamics Series 19. Washington DC: American Geophysical Union.

Stampfli GM and Borel GD (2002) A plate tectonic model for the Paleozoic and Mesozoic constrained by dynamic plate boundaries and restored synthetic oceanic isochrons. *Earth and Planetary Science Letters* 196: 17 33.

Torsvik TH and Cocks LRM (2004) Earth geography from 400 to 250 million years ago: a palaeomagnetic, faunal and sedimentological review. *Journal of the Geological Society, London* 161: 348 361.

Windley BF (1995) *The Evolving Continents*, 3rd edn. New York: John Wiley.

TERTIARY TO PRESENT

Contents

Paleocene

J J Hooker, The Natural History Museum, London, UK

Introduction

The Paleocene Epoch/Series is the first of the Cenozoic Era/Erathem. It is the first of five epochs in the Tertiary Period and the first of three in the Paleogene, which is treated either as a period in its own right or as a subdivision of the Tertiary. The Paleocene succeeds the Cretaceous Period/System and precedes the Eocene Epoch. The Paleocene lasted nearly 10 million years, from 65.5 till 55.8 Ma, and is divided approximately equally into three ages/stages (in order of decreasing age): the Danian, the Selandian, and the Thanetian (**Figure 1**). The naming of the Paleocene follows the earlier procedure of adding a prefix denoting degree of antiquity or modernity, in this case 'paleo', from the Greek *palaios*, meaning 'ancient', and 'cene', from the Greek *kainos*, meaning 'recent'. The Paleocene was the last of the Cenozoic epochs to be named, originally being proposed by Schimper in 1874. Schimper was a palaeobotanist and, in contrast to the definitions of the other Cenozoic epochs,

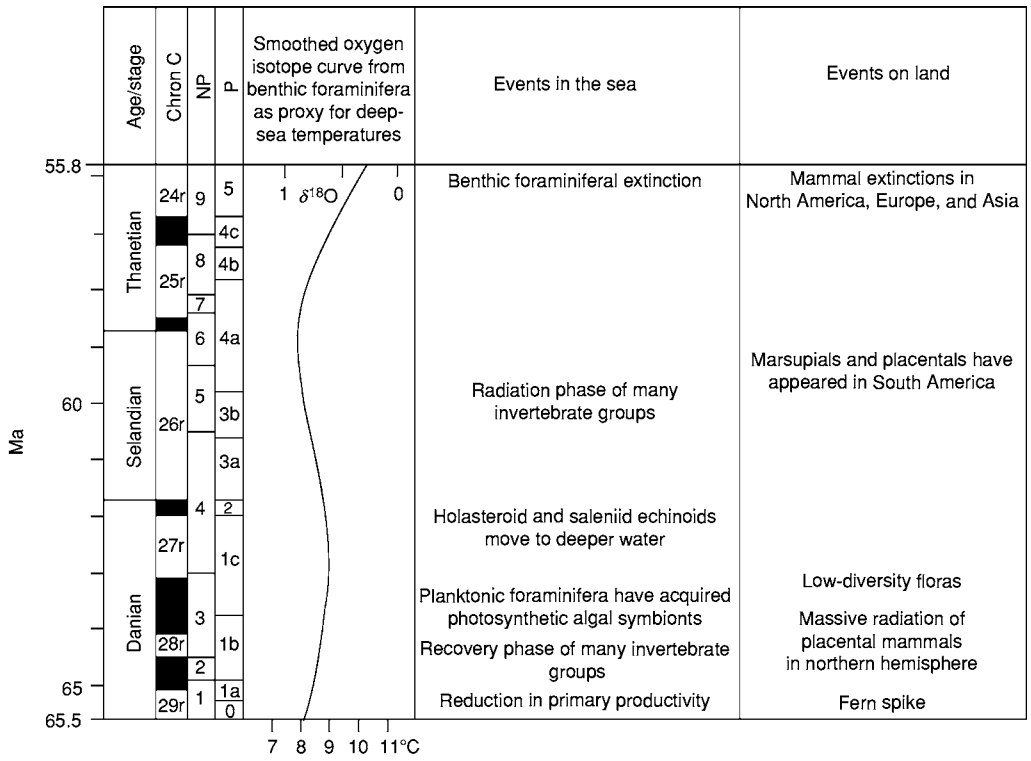


Figure 1 Time chart of the Paleocene, showing how it is divided up by ages/stages, magnetochrons (Chron C), global calcareous nannoplankton (NP), and planktonic foraminiferal (P) biozones. Also shown are an isotope proxied temperature curve and the main biotic and physical events in the sea and on land. Magnetochrons are divided into normal (black) and reversed (white) intervals; the normals, often composite, are younger than the reversals (r) that bear the same number. Data for the isotope curve from Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686 693.

he based his concept of the Paleocene on floras. Thus, he noted the presence in the Paris Basin of distinctive assemblages of plants that occur in strata overlying the Cretaceous and underlying the Eocene.

Unfortunately, some of the strata Schimper regarded as Paleocene are now known to be contemporaneous with those he regarded as Eocene. Partly as a consequence of this confusion, the Paleocene nomenclature did not gain worldwide acceptance until nearly a century after Schimper's work. The Paleocene was largely ignored by marine workers, but since the early years of the twentieth century, it has been championed by vertebrate palaeontologists, especially in North America, where it was recognized to be characterized by highly distinctive faunas of land mammals (Figure 2). Over the decades, the concept of its time-span has changed considerably and has also meant different things to palaeontologists in different fields. In Schimper's time, the first Paleocene age/stage, the Danian, was regarded as latest Cretaceous, mainly because of the continuation in northwest Europe of the typically Cretaceous chalk facies (Figure 3). The Danian was moved into the Paleocene when it was realized that it postdated a major extinction event in the sea and on land, which is now taken to mark the end of the Cretaceous Period and indeed of the Mesozoic Era. The official Global Stratotype Section and Point (GSSP) for the beginning of the Paleocene is at El Kef, Tunisia. The end of the Paleocene was also the subject of much discussion until a decade ago, when a sharp anomaly in the carbon isotope curve, known as the carbon isotope excursion (CIE), was recognized in a borehole core from the Weddell Sea, Antarctica. This has been interpreted as a climatic warming event, known as the Paleocene–Eocene thermal maximum (PETM). The PETM had profound and widespread effects on marine benthos and land mammal faunas, and is now accepted to mark the boundary between the Paleocene and Eocene epochs.

Plate Tectonics and Other Physical Phenomena

In many ways, the Paleocene represents a continuation of processes begun in the Cretaceous. Thus, Pangaea continued to fragment. Falling sea-levels since the beginning of the Late Cretaceous appear to

Figure 2 Kutz Canyon area, San Juan Basin, New Mexico, USA., showing the thick sequence of non marine fluvial mudstones, spanning much of the Paleocene. This is the basin that originally yielded distinctive Paleocene mammal faunas.

Figure 3 Stevns Klint, Denmark, showing marine Danian chalk facies, with basal fish clay unit, resting on Maastrichtian (latest Cretaceous) chalk.

have reached a threshold in the Paleocene, draining the very extensive Late Cretaceous marine carbonate platforms and dramatically increasing the non-marine area of the globe. A phase of the North American Laramide Orogeny also took place during the epoch. This involved continued sliding of the Farallon Plate under the western margin of the North American continent to produce early phases of the Rocky Mountains uplift. As a result, the North American Western Interior Seaway, which in the Late Cretaceous had split the continent from north to south, almost completely filled with sediment shed from the rising Rocky Mountains. The remnant Paleocene elongate inlet that had only a southerly opening is known as the Cannonball Sea; and the series of basins that formed in the west of the area were non-marine. The marine area of the Gulf Coast and Florida continued to subside.

The Farallon Plate also continued to be subducted beneath the western margin of South America, resulting in early phases of the Andes uplift. North America and South America were slowly moving away from each other at this time, the intervening ocean allowing deep-water circulation between the Pacific and Atlantic. Subduction in the Pacific continued on the eastern side (Kula Plate) beneath Kamchatka and Sakhalin. A land bridge across the Bering Straits was also intermittently developed between North America and Asia in response to fluctuating sea-levels.

On the other side of the world, the Atlantic Basin was continuing to spread. This particularly involved an extension of rifting northwards between Labrador

and West Greenland and between East Greenland and Europe. A hotspot formed beneath Greenland, producing outpourings of lava that intensified at the end of the epoch as the East Greenland–Europe area drifted over the hotspot. Between Europe and Asia, the epicontinental West Siberian Sea, although more restricted than in either earlier or later times, extended southwards from the Arctic Ocean, reaching the northeastern part of the Tethys Ocean (known as the Peritethys) separating Asia from Europe near the end of the epoch.

Africa moved and rotated north, pushing Apulia (comprising Italy, the former Yugoslavia, and western Greece) towards the main European craton, and producing the initial phases of the Alpine Orogeny. Eastward extension of this structural belt through Asia Minor and southern Iran partially isolated the Peritethys. India was an island continent still moving north towards Asia. The massive outpourings of hotspot-related basaltic lava that occurred in the Deccan region of West India during the latest Cretaceous continued for a brief interval in the earliest Paleocene. The Aluk Plate in the South Pacific continued to be subducted beneath the western edge of the Antarctic Peninsula. Rifting occurred between Antarctica and Australasia, but the two continents did not separate. In fact, during the Paleocene, the three major elements of Gondwana (South America, Antarctica, and Australasia) remained in contact. At the opposite pole, the Arctic Ocean was, for most of the epoch, an enclosed water body, separated by land from the rest of the world's oceans.

Biota

After the End-Cretaceous extinctions (*see* **Mesozoic:** End Cretaceous Extinctions), the earliest Paleocene biota was notable for the absence of such major and formerly diverse groups as ammonites, belemnites, rudists, plesiosaurs, mosasaurs, and non-avian dinosaurs, as well as for the low abundance and diversity of brachiopods, bivalve and gastropod molluscs, and marine reptiles. There is generally low abundance and diversity of marine life at the beginning of the Paleocene. In fact, some groups (both in the sea and on land) show a low-diversity recovery phase followed by radiation. However, the pattern is different for other groups of organisms. Key biotic events are discussed below.

Marine Realm

Calcareous nannoplankton Few members of the calcareous nannoplankton, a group of microscopic calcifying algae (coccolithophores and their possible relatives), survived the end of the Cretaceous. Paleocene nannofloras comprise these relict survivors plus an array of rapidly radiating new taxa (24 genera in the course of the epoch). By the end of the Paleocene, another major turnover resulted in over half of the Cretaceous relict species, and nearly a third of the newly evolved genera, becoming extinct. The rapidity of this evolutionary turnover and the widespread occurrence of these fossils in marine strata have resulted in the establishment of nine globally recognized Paleocene biozones (**Figure 1**).

Dinoflagellates The dinoflagellates, a group of cyst-forming unicellular algae, exhibit a stepwise origination pattern in the Early Paleocene. Nevertheless, rapid evolution and widespread occurrence make dinoflagellates important zone fossils. In particular, the biostratigraphically important genus *Apectodinium* originates during this epoch. A nearly worldwide acme of the genus occurs at the very beginning of the succeeding Eocene and this is one of the primary markers used for recognizing the boundary between the two epochs.

Foraminifera Rapid radiation of planktonic foraminifera typical of the Cenozoic continued in the earliest Paleocene from its beginnings in the last few hundred thousand years of the Cretaceous. The result was almost complete replacement of latest Cretaceous species by Cenozoic ones in an interval of less than 1.5 million years. It is thought that some planktonic foraminiferal species acquired photosynthetic algal symbionts during the Paleocene, which may have allowed them to spread into oligotrophic environments. Benthic foraminifera fared better than did the planktonics, and low-oxygen-tolerant species increased to dominate early in the epoch. Notable newcomers were the textulariids among the agglutinated-shelled forms and the nummulitids among the calcareous-shelled forms, these latter representing one of the best known groups of larger benthic foraminifera of the Cenozoic. A major extinction, the benthic foraminiferal extinction (BFE), affected benthic foraminifera at the end of the Paleocene. The evolutionary and cosmopolitan attributes of planktonic foraminifera, like those of the calcareous nannoplankton, have resulted in planktonic foraminifera being used to divide the Paleocene into eight global biozones (**Figure 1**).

Coelenterata and bryozoans The millepore hydrozoans made their appearance during the Paleocene, as did two families of octocorals. There were no innovations at family level within the scleractinian corals. However, this group does show a marked low in terms of diversity. Based on gross morphology, there is no

evidence that algal symbiosis collapsed at the end of the Cretaceous. Although the Paleocene reef recovery does not involve a return to Mesozoic levels of diversity, it does mark the first appearance of microbially cemented reefs since the Jurassic Period. This suggests that the Paleocene marks the emergence of modern coral reef communities, rather than a recovery from a eutrophically driven collapse. The Paleocene radiation of cheilostome and ascophoran bryozoans increased its pace; especially the cheilostomes, the family diversity of which increased exponentially.

Molluscs Cephalopods scarcely recovered from the end-Cretaceous extinctions, which eliminated ammonites and belemnites, although a new family of nautiloids, the Aturiidae, arose at the beginning of the Paleocene. Bivalve and gastropod recovery was more dramatic. Typically, the initial recovery phase in which diversity remained low varied in length and was followed by several pulses of increasing diversity separated by lows (initial radiation phase). Several families (Ostreidae, Carditidae, and Turritellidae), however, show a different pattern. In these cases, diversity increased or remained high at the beginning of the epoch, then suffered a decline. This means that these families formed a much higher percentage of the mollusc faunas early on than later in the epoch. The speciose nature of these families and their mixture of planktotrophic and brooding larval development mechanisms may have enhanced the Paleocene survival and early success of these opportunists. Once new forms arose later in the Paleocene, however, their competitiveness was low. Survival of Mesozoic taxa was more marked in high northern and southern latitudes than elsewhere.

Echinoids The end-Cretaceous extinctions resulted in a drop in diversity in the Early Paleocene and a change from roughly equal representation of regular and irregular groups to dominance by irregular echinoids. Irregular forms were also affected at the end of the Cretaceous, with holasteroids being decimated and survivors moving from shelf to deep-water environments as the Danian chalk facies disappeared. The regular family Saleniidae also shifted into deeper waters at the same time.

Vertebrates Marine fishes appear to have suffered little at the Cretaceous–Tertiary boundary. Among the cartilaginous elasmobranchs (sharks and rays), one family (the Torpedinidae) originated at the beginning of the Paleocene and four more (the Lamnidae, Otodontidae, Carcharinidae, and Mobulidae) first appear later in the epoch. Six teleost families appeared at the beginning of the Paleocene and a further 10 families appeared later in the epoch. Despite a patchy record, the Paleocene appears to mark the meagre beginning of a major Cenozoic radiation. In contrast, the only marine reptiles to survive into the Paleocene were turtles and dyrosaurid crocodilians.

Continental Realm

Land plants Schimper, when basing his Paleocene on distinctive floras, was aware that what he was observing might only be a local phenomenon of north-western Europe. In fact, Paleocene seed plant taxa and floral composition seem to represent a segment of a modernization trend, the origins of which lay in the latest Cretaceous, when angiosperms (flowering plants) became dominant over gymnosperms (broadly, conifers and cycads). There was turnover of seed plant taxa across the Cretaceous–Tertiary boundary and low-diversity opportunistic floras (particularly ferns) suggestive of abrupt ecological disruption in the very earliest Paleocene in western North America and probably elsewhere. Diversity, however, increased later in the epoch. Angiosperm fruits in the Paleocene were mainly small and dry.

Invertebrates Non-marine (mainly pulmonate) gastropods show no particular effect from end-Cretaceous events and the Paleocene saw essentially the beginning of a Cenozoic radiation of terrestrial families. For insects, there is little evidence of extinction, at least at family level, at the end of the Cretaceous, although there are few Paleocene sites yielding members of this group. Nevertheless, diversification that began in the Cretaceous seems to have increased in the Paleocene. Evidence from leaf damage by herbivorous insects suggests that early Paleocene insect herbivores were generalists. A recovery phase is lacking for at least the first million years of the Paleocene.

Vertebrates The pattern of events from freshwater teleost fishes is similar to that of the marine pattern. Five families have first records at the beginning of the Paleocene and a further 10 occur later in the epoch. Similarly, there appears to have been little effect of end-Cretaceous events on Paleocene amphibians, lizards, snakes, crocodilians, or turtles. The freshwater champsosaurs also survived into the Paleocene. The record of birds is sparse, but there is as yet no evidence of a major turnover at the end of the Cretaceous and no undoubted appearances of modern bird families in the Paleocene. There was, in contrast, a major change in mammals between the Cretaceous and the Paleocene. Although a boundary sequence exists only in western North America, Paleocene mammals differ radically from latest Cretaceous

ones in every continent where they are known. This implies a major turnover at or near the boundary. In North America, the Cretaceous marsupial versus placental-dominated fauna was replaced in the Paleocene by an almost exclusively placental fauna (*see* **Fossil Vertebrates:** Placental Mammals). In Europe (poorly known in the Late Cretaceous) and Asia, the Paleocene faunas are also dominated by placentals, but the closeness of their phylogenetic relationships with Late Cretaceous placental groups is disputed. The greatest difference is in South America, where primitive non-therian mammals are replaced by a diversity of marsupials and placentals. Prominent features of Paleocene mammal faunas are (1) their dominance by archaic types not closely related to modern orders and (2) their strong, continent-specific endemism. Nevertheless, the following modern orders do have their first fossil appearances during the Paleocene: carnivorans, edentates, rodents, and probably also lipotyphlans (typical insectivorans) and macroscelideans (elephant shrews). Mammals underwent rapid recovery and massive radiation following the end-Cretaceous extinction. Extinction of many of these archaic types took place at the end of the Paleocene in the northern hemisphere continents when major dispersal of more modern types displaced them. Various endemic groups, however, evolved in South America, which became isolated, first from North America and later from Antarctica. These groups survived long after the Paleocene.

Climate and Environments

After the end of the Cretaceous, oxygen isotopic records from benthic foraminiferal tests document a gradual warming of the world's oceans during the first third of the Paleocene. This warming reversed the overall but fluctuating cooling trend of the latest Cretaceous (Maastrichtian), but was minor compared to the brief warming event near the end of the Maastrichtian. A cooling in the Late Paleocene was followed by a major warming, punctuated at the very end of the epoch by the beginning of a brief intense warming episode that marks the onset of the Eocene (Figure 1). During the Paleocene, there was likely to have been little or no polar ice. The impact of climate and other factors on Paleocene environments are examined in the following sections.

Marine Environments

Following the end of the Cretaceous, it has been suggested that primary productivity in the oceans was strongly reduced. This is thought to explain the general change from infaunal to epifaunal dominance among benthic foraminiferal communities at the Cretaceous–Tertiary boundary. In some deep-water sites, the change was briefly delayed at the very beginning of the Paleocene by an opportunistic low-diversity infaunal assemblage that may have been responding to a large but short-term flux of organic matter to the seafloor, from the mass mortality of microplankton. Moreover, radiolarians and associated biosiliceous oozes, which infer high oceanic productivity, are generally rare across the Cretaceous–Tertiary transition. However, evidence of rich radiolarian assemblages across the boundary in New Zealand suggest that enhanced upwelling caused by climatic cooling characterized this southern high-latitude area at the beginning of the Paleocene. Carbon isotope studies of foraminifera also indicate stability of surface productivity at high, in contrast to low, latitudes, with resultant lower extinction levels. Another phenomenon is the apparent enhanced survival of both mollusc and ostracod taxa from the Cretaceous at high, compared to low, latitudes, with subsequent spread of the ostracods to lower latitudes later in the Paleocene.

Later Paleocene oceans had more normal higher primary productivity according to their greater diversity of planktonic microbiota. They are also likely to have had modern rates of thermohaline circulation and thus of nutrient flux from subsurface to surface waters. Major warming at the end of the epoch caused a slowing of the thermohaline circulation, with resultant reduction in rate of nutrient flux, which in turn expanded the geographic range of oligotrophic habitats. This trend eventually involved a selective warming of the deep ocean by 4°–6°C due to changed circulation and a much reduced pole-to-equator temperature gradient, resulting in reduced wind intensity. The sudden warming at the Paleocene–Eocene boundary is attributed to a massive injection of CO_2 into the oceans and atmosphere by thermal dissociation of methane hydrates and their release from marine sediments. This event is judged to be the cause of the contemporaneous major extinction of benthic organisms, mainly foraminifera (BFE) and ostracods.

Continental Realm

Leaf physiognomy climate proxies produce a land-based climate curve resembling that for the marine realm. Vegetation shows a shift from open-canopy broad-leaved evergreen woodland in the Late Cretaceous to rainforest in the Early Paleocene. This marks the first appearance of such a vegetation type, which extended to higher latitudes than the present day because of the absence or near absence of polar ice. The highest (polar) latitudes were occupied by

broad-leaved deciduous forests. Despite the enhanced equability, these floras had to survive extended winter darkness. Accordingly, their structure was open and their diversity low. The very beginning of the Paleocene is also locally marked by a dominance of the spores of ferns. This phenomenon, known as the 'fern spike', is recorded from regions as far apart as western North America and New Zealand and suggests that these plants were the first colonizers of a denuded landscape.

The mammals that lived in these Paleocene habitats were mainly small. Their Cretaceous ancestors had been mainly insectivorous and to a certain extent carnivorous, and they were only now expanding their dietary spectrum to include fruit. Larger size, which went hand in hand with leaf eating, was rare and began to be evolved late in the epoch. The ecological composition of most well-known faunas supports the presence of widespread forested environments. Land connections between continents, even if fleeting during times of low sea-level, allowed marsupials and placentals to disperse from North America to South America around the end of the Cretaceous or the beginning of the Paleocene. The marsupials went on to colonize Australasia via Antarctica before Australasia finally broke free.

See Also

Fossil Vertebrates: Fish; Placental Mammals. **Mesozoic:** Cretaceous; End Cretaceous Extinctions. **Microfossils:** Foraminifera. **Palaeoclimates**. **Sedimentary Environments:** Reefs ('Build-Ups'). **Tertiary To Present:** Eocene.

Further Reading

Aubry M P, Lucas SG, and Berggren WA (eds.) (1998) *Late Paleocene Early Eocene Climatic and Biotic Events in the Marine and Terrestrial Records*. New York: Columbia University Press.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Chaloner WG, Harper JL, and Lawton JH (eds.) (1991) The evolutionary interaction of animals and plants. *Philo sophical Transactions of the Royal Society of London, Series B* 333: 177 288.

Collinson ME (1990) Plant evolution and ecology during the early Cainozoic diversification. *Advances in Botanical Research* 17: 1 98.

Culver SJ and Rawson PF (eds.) (2000) *Biotic Response to Global Change. The last 145 million years*. Cambridge: Cambridge University Press.

Friis EM, Chaloner WG, and Crane PR (eds.) (1987) *The Origins of Angiosperms and their Biological Conse quences*. Cambridge: Cambridge University Press.

Hansen TA (1988) Early Tertiary radiation of marine mol luscs and the long term effects of the Cretaceous Tertiary extinction. *Paleobiology* 14: 37 51.

Hartman JH, Johnson KR, and Nichols DJ (eds.) (2002) The Hell Creek Formation and the Cretaceous Tertiary boundary in the northern Great Plains: an integrated con tinental record of the end of the Cretaceous. *Geological Society of America Special Papers* 361: 1 520.

Khain VE and Balukhovsky AN (1997) *Historical Geotectonics, Mesozoic and Cenozoic*. Russian Transla tion Series 117. Rotterdam: AA Balkema.

MacLeod N, Rawson PF, Forey PL, *et al*. (1997) The Cret aceous Tertiary biotic transition. *Journal of the Geological Society, London* 154: 265 292.

Scotese CR and Golonka J (1992) *Paleogeographic Atlas, PALEOMAP Progress Report 20 0692*. Dept. of Geology, University of Texas at Arlington.

Smith AG, Smith DG, and Funnell BM (1994) *Atlas of Mesozoic and Cenozoic Coastlines*. Cambridge: Cambridge University Press.

Thomas DJ, Zachos J, Bralower TJ, Thomas E, and Bohaty S (2002) Warming the fuel for the fire: evidence for the thermal dissociation of methane hydrate during the Paleocene Eocene thermal maximum. *Geology* 30: 1067 1070.

Wing SL, Gingerich PD, Schmitz B, and Thomas E (eds.) (2003) Causes and consequences of globally warm climates in the early Paleogene. *Geological Society of America Special Papers* 369: 1 614.

Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686 693.

Eocene

J J Hooker, The Natural History Museum, London, UK

Introduction

The Eocene epoch/series is the oldest of the four original subdivisions of the Tertiary period/system proposed in 1833 by Sir Charles Lyell in his *Principles of Geology*. The name derives from the Greek '*eos*', meaning dawn, and '*kainos*', meaning recent. This is "... because the very small proportion of living species contained in those strata indicates what may be considered the first commencement, or *dawn*, of the existing state of the animate creation" (vol. 3, p. 55). The time covered by Lyell's 'existing state' is what we now understand to be the Cenozoic era/erathem, which itself is broadly divided into either Tertiary and Quaternary or Paleogene and Neogene periods. Lyell subdivided the Tertiary into four epochs based on the proportions of living and extinct species of shelled organisms (molluscs and foraminifera) encountered as fossils in different strata. He recognized 1238 Eocene species, of which he considered only 42 (or 3.5%) remain alive today. Lyell's species are about equivalent to what modern marine biologists would rank as genera or even subfamilies. For this reason, no modern species of mollusc or foraminifer is currently recognized as occurring as far back as the Eocene.

The Eocene as recognized today has changed considerably in its definition and time-span since 1833. Its earliest parts have become the Paleocene and its later parts the Oligocene, both epochs that were described after 1833. The Eocene thus succeeds the Paleocene and precedes the Oligocene. It is currently estimated to have lasted nearly 22 million years, from 55.8 to 33.9 Ma. The Eocene itself is divided into four ages/stages (in order of decreasing age): the Ypresian, the Lutetian, the Bartonian, and the Priabonian (Figures 1 and 2). The beginning and end of the Eocene have only recently been stabilized by the Paleogene Subcommission of the International Union of Geological Sciences (IUGS). Its beginning is marked by a sharp dip in the carbon isotope curve, named the Carbon Isotope Excursion (CIE), interpreted as a global warming event, the Paleocene–Eocene Thermal Maximum (PETM). This climate event sparked major changes in both marine and continental biotas. The Global Stratotype Section and Point (GSSP) for the beginning of the Eocene is placed at Dababiya, Egypt. This geological section is the best available for demonstrating the boundary criteria and acts as a global reference. The end of the Eocene is marked by extinction of the planktonic foraminiferal family Hantkeninidae, representing the last in a cumulative series of extinctions caused by long-term global cooling. The GSSP for the Eocene–Oligocene boundary is at Massignano, Italy.

Plate Tectonics and Other Physical Phenomena

The processes of subduction around the Pacific rim continued from the Paleocene, with the Kula Plate disappearing beneath the Aleutian Arc of Alaska. In North America, the remnants of the Late Cretaceous Western Interior Seaway, still present in the Paleocene, disappeared completely. Caribbean, deep-water circulation between the Pacific and Atlantic began to be cut off late in the Eocene as the Central American Isthmus formed a structural unit, but remained submerged.

Seafloor spreading in the Atlantic continued to migrate northwards on either side of Greenland. Thus, a seaway formed early in the Eocene between Greenland and Europe, linking the earlier landlocked Arctic Ocean once again with the Atlantic. At the same time, the maximum outpourings of lava associated with the Iceland hotspot occurred in East Greenland and the then adjacent Hebridean Province. The Labrador Seaway between Greenland and North America also widened and the land connection between Europe and North America via Greenland finally severed. A land bridge developed intermittently between North America and Asia across the Bering Straits as a result of sea-level changes.

For the entire epoch, the epicontinental West Siberian Seaway linked the Arctic Ocean with the Peritethys, maintaining the isolation of Europe from Asia. This seaway narrowed at its southern end to form the Turgai Straits, which may occasionally have dried out at times of low sea-level. In fact, except for the first million or so years of the epoch, when Europe was connected to North America via Greenland, Eocene Europe comprised several island masses isolated from the rest of the world's continents. The Peritethys continued to be partially separated from the main mass of the Tethys Ocean by the narrow orogenic belt stretching intermittently from Italy through Asia Minor to southern Iran. In this complex area of the ancient Mediterranean, the islands of Corsica, Sardinia, and the Balearics were still part of the Iberian Peninsula. During the course of the Eocene,

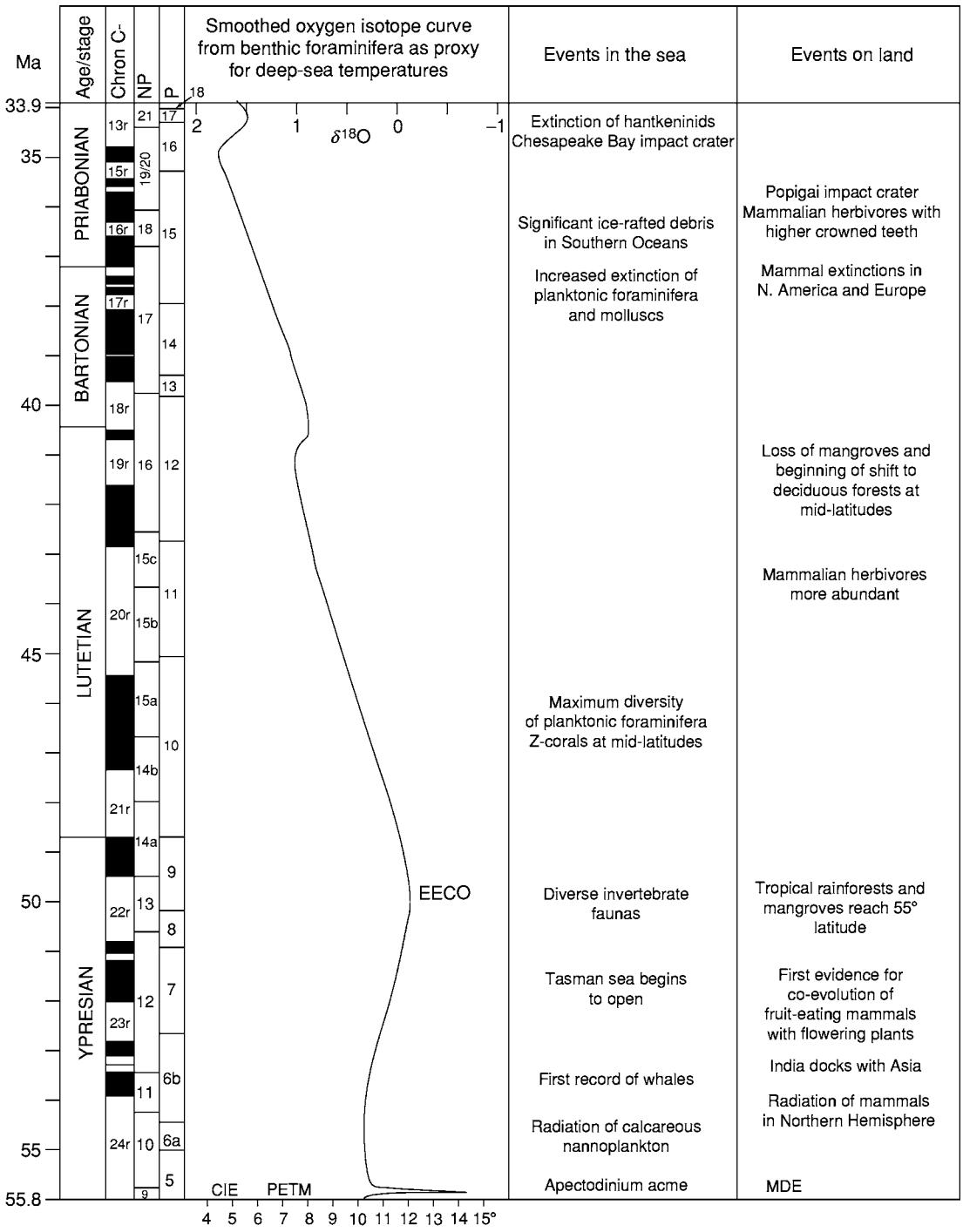


Figure 1 Time chart of the Eocene, showing how it is divided up by ages/stages, magnetochrons (chron C), and global calcareous nannoplankton (NP) and planktonic foraminiferal (P) biozones. Also shown are an isotope proxy temperature curve and the main biotic events in the sea and on land. Magnetochrons are divided into normal (black) and reversed (white) intervals; the normals (n), often composite, are younger than the reversals (r) that bear the same number. CIE, Carbon Isotope Excursion; EECO, Early Eocene Climatic Optimum; MDE, Mammalian Dispersal Event; PETM, Paleocene Eocene Thermal Maximum. Data for the isotope curve from Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686 693. Absolute dates from Gradstein F, Ogg J, and Smith A (in press) *A Geological Timescale 2004*. Cambridge University Press, Cambridge.

Figure 2 Coastal section in Alum Bay, Isle of Wight, UK, showing strata spanning almost the entire Eocene. Oldest is to the right, youngest to the left. The red and brown are Ypresian, the yellow with dark intercalations is Lutetian, the grey below the house is Bartonian (all vertical), and the white (mainly horizontal) is Priabonian.

anticlockwise rotation of Iberia resulted in initial uplift of the Pyrenees.

Further south, India completed its northward drift and docked with Asia in the Early Eocene, laying the foundations for the Himalayan uplift. Rifting between Australasia and Antarctica widened and developed into an (at first) narrow Tasman Sea. Australasia began its 50-million-year-long trek to lower latitudes at this time. Subduction beneath the west side of the Antarctic Peninsula led to its uplift and shallowing of the back-arc basin to the east. South America remained connected to the Antarctic Peninsula for most of the epoch. However, development of the Scotia Arc and formation of the Scotia Sea (Drake Passage) meant that, by the end of the Eocene, all major elements of Gondwana had separated and the modern arrangement of the continents was essentially in place.

Two important impact-ejecta strewn fields have also been recognized within Late Eocene sediments. It has been suggested that two large impact craters, Popigai in Siberia and Chesapeake Bay in the USA, were the sources for these (Figure 1).

Biota

The Eocene biota is characteristically diverse and abundant. Most surviving groups had recovered from the Early Paleocene diversity low by this time and had undergone or were undergoing radiation. Marine groups, such as molluscs, crustaceans, and echinoids, had a familiar modern appearance. On land, the same was true of many reptile and amphibian groups, but Eocene mammals differed radically from their living relatives. Key biotic events are discussed below.

Marine Realm

Calcareous nannoplankton An important turnover in this group of microscopic algae occurred during the first million years of the Eocene with the origination of many new taxa. The genera that had radiated and dominated in the Paleocene became extinct. Survivors were the previously low-diversity genera of modern aspect. This turnover is attributed to the PETM (see below). Later in the Eocene, diversity reduced regionally and deeper water habitats were vacated because of progressive cooling. New taxa evolved in response to the temperature changes and increased eutrophication. The rapid evolution and widespread occurrence of nannoplankton at this time have enabled the Eocene to be divided into 12 global biozones (Figure 1).

Dinoflagellates The PETM resulted in a near-global spread of dinoflagellate species belonging to the genus *Apectodinium* through middle and high latitudes. This event, known as the *Apectodinium* acme, is an important biostratigraphical marker for the Paleocene–Eocene boundary. Continued radiation produced increasingly diverse cyst assemblages during the Early and Middle Eocene, especially of the genus *Wetzeliella* and its relatives. Diversity remained high for most of the Eocene, despite the later cooling of sea surface temperatures. The very end of the epoch, though, bears witness to a reduction in abundance of low-latitude taxa, as well as the invasion of these areas by formerly higher latitude groups.

Foraminifera The initial Eocene turnover in planktonic foraminifera at the PETM was minor. Radiation continued from the Paleocene to reach a diversity maximum in the Middle Eocene. Then, deteriorating temperatures caused long-term stepwise extinctions, of which the most intense was at the end of the Middle Eocene (Bartonian). Surface-dwelling species were gradually replaced by cold-tolerant, subsurface species. Benthic foraminifera fared much worse at the PETM, suffering 30–50% extinctions for middle bathyal through abyssal forms. This is known as the Benthic Foraminiferal Extinction (BFE). Shallower water taxa fared better. The latter include the best-known Eocene calcareous foraminifera, some of which (e.g., the nummulites) became important rock formers and reached maximum diameters of 10 cm. Benthic foraminifera, like the planktonics, underwent stepwise extinctions throughout the Middle and Late Eocene. The evolutionary and cosmopolitan attributes of planktonic foraminifera, like those of the calcareous nannoplankton, have resulted in their use for dividing the Eocene into 13 global biozones (Figure 1).

Coelenterates and bryozoans Two families of octocorals originated at the beginning of the Eocene. Radiation of scleractinian corals occurred, with increasing species numbers throughout the epoch. Interestingly, there was no reduction in diversity when climates cooled in the Middle and Late Eocene. Amongst the bryozoans, the rapid diversification of cheilostome families in the Paleocene continued into the Eocene, but then began to slow, although there was a peak diversity of species in the Priabonian.

Molluscs After the terminal Cretaceous extinctions of ammonites and belemnites, the surviving coleoids (squids and cuttle fishes) were slow to recover and have left little fossil evidence in the Paleocene. Their appearance in the Eocene is therefore virtually as 'Lazarus' taxa. Most of the spirulid, and all of the sepiid, radiation took place in this epoch, with several new families appearing. Diversification in bivalves and gastropods continued from the Paleocene into the Eocene. A study in the Gulf Coast, USA, has shown that species numbers reached a peak in the Bartonian, when climates were already beginning to deteriorate, and plummeted during the Priabonian. The origination of many new families in the epoch gave these faunas a more modern aspect.

Echinoids The main Eocene evolutionary events surround the irregular group. This epoch saw a decline in cassiduloids and a corresponding rise in diversity of clypeasteroids, with five new families appearing. Clypeasteroids were the last group to evolve non-planktotrophic lecithotrophic or brooding development for their larvae, and this development occurred in the Eocene. Although this adaptation resulted in reduced dispersal ability, it is thought to have enhanced resistance to the rigours of increased seasonality associated with the deteriorating Middle and Late Eocene climate.

Vertebrates Amongst cartilaginous fishes (Elasmobranchii), there was a significant increase in rays (Rajiformes), with three new families making their first appearance at the beginning of the epoch. Eocene marine teleosts are characterized by a continuation of their Cenozoic radiation, particularly in tarpon and eels (Elopomorpha) and massively in spiny-ray fishes (Acanthomorpha). A caveat, however, is provided by the breadth of diversity of the many newly appearing acanthomorph families, implying an earlier record of these, which we have yet to discover.

Marine mammals are first recorded from the Eocene in the form of primitive whales (Cetacea) and sea cows (Sirenia). These represent the first marine tetrapod evolutionary innovation since the extinction of plesiosaurs and mosasaurs at the end of the Cretaceous. The earliest known cetaceans are from the Ypresian of Pakistan. These early members retained well-developed walking limbs, both fore and hind, and are interpreted to have been amphibious inhabitants of both freshwater and marine habitats. Limb structure demonstrates a close relationship with the entirely land-based artiodactyls (cloven-hoofed mammals) which appeared at about the same time.

Continental Realm

Land plants In the Eocene, angiosperms (flowering plants) diversified and became more modern in appearance. Many modern genera (although no modern species) can be recognized within families that are well represented today. Examples are magnolias (Magnoliaceae), grape vines (Vitaceae), citrus and allies (Rutaceae), spurges (Euphorbiaceae), dogwoods (Cornaceae), custard apples (Anonaceae), laurels (Lauraceae), moonseeds (Menispermaceae), icacinas (Icacinaceae), and palms (Arecaceae). Grasses, on the other hand, were notably rare.

Invertebrates Land gastropods underwent a major Eocene radiation with two new prosobranch and 16 new pulmonate families appearing. This gave a more modern aspect to land snail and slug faunas. The poor Paleocene record of insects makes it difficult to judge specific Eocene innovations. By the end of the Eocene, though, 223 families had appeared since the Cretaceous. In particular, there is a surge in origination of dragonflies (Odonata), flies (Diptera), butterflies and moths (Lepidoptera), and beetles (Coleoptera). Compared with the Paleocene, leaf damage indicative of insect herbivory shows an overall increase in abundance and diversity, which is expressed particularly by more specialist damage types, such as leaf mines and leaf galls.

Vertebrates Eight families of freshwater teleost fishes have their first records in the Eocene. These include such well-known families as carp (Cyprinidae), salmon (Salmonidae), and perch (Percidae). Five modern families of freshwater and terrestrial turtles appeared in the epoch. These included the well-known terrapins (Emydidae) and tortoises (Testudinidae). Nine modern families of birds have their first records in the Eocene, with doubtful records of a further 14. The undoubted records include parrots (Psittacidae), owls (Strigidae), nightjars (Caprimulgidae), swifts (Apodidae), colies (Coliidae), and rollers (Coraciidae). None, however, belong to the passerines, which were either very rare at this time or had not yet evolved.

Land mammals underwent their greatest innovation of the Cenozoic in the Eocene. Most modern

placental orders appeared at the beginning of the epoch in the northern hemisphere in what is known as the Mammalian Dispersal Event (MDE). The sudden appearance of so many specialized morphotypes implies an earlier evolution extending back into the Paleocene, of which we have no record. These newcomers were bats (Chiroptera), primates (Primates), cloven-hoofed mammals (Artiodactyla), odd-toed ungulates (Perissodactyla), and elephants (Proboscidea). Other modern orders appeared later in the Eocene, including hyraxes (Hyracoidea), pangolins (Pholidota), and tree-shrews (Scandentia). Many early members of these groups were small. Moreover, few hoofed plant-eaters developed a large size or an exclusively leaf-eating diet until relatively late in the epoch.

Climate and Environments

Eocene climates include the warmest of the entire Cenozoic era (early in the epoch), deteriorating later to lead eventually to the first Cenozoic ice build-up in Antarctica (Figure 1). Carbon and oxygen isotope records based on the analysis of foraminiferal tests, mammalian dental enamel, soil carbonates (Figure 3), and lignites all show a short, sharp, 200 000-year-long, negative anomaly (CIE) at the very beginning of the Eocene. This indicates an extreme warming perturbation (PETM), when deep-sea temperatures rose by 4–6°C. This spike was superimposed on a long-term warming trend that began in the Late Paleocene and culminated late in the Ypresian with a comparable, but longer duration, warming peak, known as the Early Eocene Climatic Optimum (EECO). During the PETM and EECO, there was probably no polar ice, pole to equator temperature gradients were much reduced, and overall wind circulation slowed. Subsequently, temperatures declined, with small warming interruptions near the end of the Lutetian and in the Priabonian (Late Eocene). Significant amounts of ice-rafted debris, indicative of glacial activity, first occur in southern high-latitude sediments at the end of the Bartonian.

Figure 3 Polecat Bench, Wyoming, USA, showing a thick sequence of latest Paleocene and earliest Eocene fluvial mudstones and palaeosols. The base of the Carbon Isotope Excursion (CIE) is recorded from mammalian dental enamel and soil nodules at a point intermediate between the two distinct red bands. The upper red band records the Mammalian Dispersal Event (MDE).

Marine Environments

The PETM perturbation is thought to have been caused by a massive injection of CO_2 into the oceans and atmosphere by the thermal dissociation of methane hydrates and their release from marine sediments. This event, in turn, caused a slowing of the thermohaline circulation with a resultant reduction in the rate of vertical marine nutrient flux and an expansion in the geographical range of oligotrophic habitats. The major rapid extinction of benthic foraminifera (BFE) and ostracods that resulted (*see* **Tertiary To Present:** Paleocene) meant low diversity for such organisms in bathyal and abyssal regions early in the Eocene. Post-extinction benthic foraminiferal faunas vary widely with geography compared with pre-extinction faunas. This is thought to be a phenomenon of highly perturbed communities. Early Eocene planktonic foraminiferal communities also show a greater diversity of oligotrophic forms than do those of the Late Paleocene. This is consistent with the proposed model of a uniformly warm Early Eocene ocean with reduced rates of circulation. Widespread deep-water anoxia, with associated calcium carbonate dissolution at the PETM, appears to have favoured the latitudinal spread of dinoflagellates of the genus *Apectodinium* (the *Apectodinium* acme) as far north as the Barents Sea and as far south as New Zealand. *Apectodinium* is thought to have been partially or fully holozoic and replaced earlier dominant photosynthetic dinoflagellate communities.

This period of warming and equability promoted high taxonomic diversity for many groups of organisms at much higher latitudes than today. Thus, tropical-type molluscs, zooxanthellate-like corals, and giant nummulites are well represented in northern Europe, a fact that impressed Lyell as long ago as 1833. The subsequent cooling reversed processes that took place early in the epoch. Thermohaline circulation rates accelerated, ocean mixing increased, low-nutrient, surface water habitats were reduced, and instability was created in high-nutrient, surface water areas through seasonal production. This

increase in seasonality, documented from oxygen isotopes in mollusc shells and fish otoliths (ear stones), is considered to be an important feature of the Late Eocene cooling. Extensive investigation of the two Late Eocene impacts has found no important, associated biotic extinctions. However, carbon and oxygen isotope analyses and microfossil assemblage composition from strata overlying the ejecta layers suggest that a minor, but prolonged, cooling followed at least one of these impact events. This prolongation of the environmental effect may have been caused by an increase in the Earth's albedo from ice-sheets that were extended in response to the cooling.

Continental Environments

Leaf physiognomy climate proxies give a land-based climate curve that, like the marine curve, shows cooling from an Early Eocene maximum, but declines in more widely fluctuating steps. Moreover, the study of leaf floras across the Paleocene–Eocene transition in the Bighorn Basin, Wyoming, USA, indicates an initial warming (PETM), followed by a cooling, and then renewed warming up to the EECO. The long-term warming phase that began in the latest Paleocene and culminated in the EECO had important effects on vegetation. During its later stages, highly diverse, multistratal rainforest with many lianas extended to latitudes of 55° or 60° north and south. This was accompanied by coastal mangrove vegetation, often dominated by the mangrove palm *Nypa*. These widespread rainforests were similar to those of south-east Asia today. They appear to have differed, however, in lacking epiphytic flowering plants and in having only very rare representatives of some taxa (e.g., dipterocarps) that are highly significant in modern south-east Asian forests. In addition to the multistratal nature of these Early Eocene forests, the plants differed from those of the Paleocene in their fruiting strategies. In the Paleocene, fruits were mainly small and dry with some drupes and nutlets. Fleshy fruits were rare. In contrast, in Eocene forests, fleshy fruits were abundant and larger nuts were also present. This change represents the beginning of the co-evolution of fleshy fruits with fruit-eating primates and of nuts with scatterhoarding rodents, that facilitates plant dispersal. At higher, polar latitudes, these rainforests were replaced by broad-leaved deciduous forests. There is no evidence in the Eocene for the coniferous forests that today clothe Earth's boreal regions. The later Eocene cooling is tracked by a shift to less diverse, more open, deciduous broad-leaved forests in middle and high latitudes, with the loss of many of the tropical taxa that dominated earlier in the epoch. Despite the development of more open vegetation, this did not include grassland, grasses being virtually absent from all Eocene floras.

The MDE that introduced many new mammals into the northern hemisphere continents at the time of the CIE (Figure 3) is partly attributed to the warm temperatures. These may have facilitated the well-documented dispersals through northern high latitudes (across the Greenland and Bering land bridges) through the extension of vegetation zones. However, minor dispersal also appears to have taken place across the Turgai Straits at lower latitudes. Low sea-levels in the vicinity of the Paleocene–Eocene boundary are also likely to have had a strong influence on dispersal. Mammal communities in the Early Eocene were dominated by small animals, many of which fed on insects and fruit and were adapted for life in the trees. Their patterns of ecological diversity resemble those of south-east Asian evergreen forests and thus support the plant evidence for the nature of the Eocene vegetation. As climates cooled later in the epoch, fruit-eating and climbing mammals diminished and larger ground-dwelling herbivores evolved. The patterns of ecological diversity that characterize these later faunas fit closely with more open wooded habitats as the plant fossils indicate. The structure and wear patterns of the teeth of these later Eocene herbivorous mammals indicate that they were browsers not grazers.

See Also

Biozones. **Fossil Invertebrates:** Insects. **Fossil Plants:** Angiosperms. **Fossil Vertebrates:** Placental Mammals. **Magnetostratigraphy**. **Microfossils:** Foraminifera; Palynology. **Palaeoclimates**. **Plate Tectonics**. **Tertiary To Present:** Paleocene; Oligocene. **Time Scale**.

Further Reading

Aubry M P, Lucas SG, and Berggren WA (eds.) (1998) *Late Paleocene Early Eocene Climatic and Biotic Events in the Marine and Terrestrial Records*. New York: Columbia University Press.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Chaloner WG, Harper JL, and Lawton JH (eds.) (1991) The evolutionary interaction of animals and plants. *Philo sophical Transactions of the Royal Society of London, Series B*, 333, pp. 177 288.

Collinson ME (1990) Plant evolution and ecology during the early Cainozoic diversification. *Advances in Botanical Research* 17: 1 98.

Culver SJ and Rawson PF (eds.) (2000) *Biotic Response to Global Change. The Last 145 Million Years*. Cambridge: Cambridge University Press.

Friis EM, Chaloner WG, and Crane PR (eds.) (1987) *The Origins of Angiosperms and Their Biological Consequences*. Cambridge: Cambridge University Press.

Khain VE and Balukhovsky AN (1997) *Historical Geotectonics, Mesozoic and Cenozoic*. Russian Translation Series 117. Rotterdam: A.A. Balkema.

Kobashi T, Grossman EL, Yancey TE, and Dockery DT III (2001) Reevaluation of conflicting Eocene tropical temperature estimates: molluskan oxygen isotope evidence for warm low latitudes. *Geology* 29: 983 986.

Prothero DR and Berggren WA (eds.) (1992) *Eocene Oligocene Climatic and Biotic Evolution*. Princeton: Princeton University Press.

Scotese CR and Golonka J (1992) *Paleogeographic Atlas*. Arlington, TX: PALEOMAP Project, Department of Geology, University of Texas.

Smith AG, Smith DG, and Funnell BM (1994) *Atlas of Mesozoic and Cenozoic Coastlines*. Cambridge: Cambridge University Press.

Thomas DJ, Zachos J, Bralower TJ, Thomas E, and Bohaty S (2002) Warming the fuel for the fire: evidence for the thermal dissociation of methane hydrate during the Paleocene Eocene thermal maximum. *Geology* 30: 1067 1070.

Vonhof HB, Smit J, Brinkhuis H, Montanari A, and Nederbragt AJ (2000) Global cooling accelerated by early late Eocene impacts? *Geology* 28: 687 690.

Wing SL, Gingerich PD, Schmitz B, and Thomas E (eds.) (2003) Causes and consequences of globally warm climates in the early Paleogene. *Geological Society of America Special Papers 369*, pp. 1 614. Boulder: Geological Society of America.

Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686 693.

Oligocene

D R Prothero, Occidental College, Los Angeles, USA

Introduction

The Oligocene Epoch was defined by Heinrich Ernst von Beyrich in 1854. This interval of geological time was based on marine strata in Belgium and Germany, thought to be younger than the Lyell's classic upper Eocene (*see* **Tertiary To Present:** Eocene) strata of the Paris Basin, but older than Lyell's (*see* **Famous Geologists:** Lyell) concept of Miocene rocks. Von Beyrich's original list of 'Oligocene' rocks contained a wide spectrum of units of varying ages, including those that are now clearly referable to the Eocene or Miocene. For example, one unit (the bone sand of Eppelsheim) produced a Late Miocene *Hipparion* fauna. In addition, the type strata of von Beyrich's Oligocene in Belgium and Germany do not overlie the type strata of the Paris Basin or Italian Eocene, so the Eocene–Oligocene boundary cannot be recognised in either area. As is true of the rest of the European Cenozoic, the type sections of the stages within the Oligocene are shallow-water deposits bounded by unconformities, and represent only a small portion of its duration.

For 130 years after von Beyrich's establishment of the Oligocene, there was considerable confusion over what was Eocene and what was Oligocene, not only in the western European type areas, but especially in other regions which could only be correlated indirectly to the stratotypes. For example, in North America, the Duchesnean land mammal age was thought to be late Eocene or Oligocene (it is now considered middle Eocene), the Chadronian land mammal age was correlated with the Early Oligocene (it is now known to be Late Eocene in age), and the Orellan and Whitneyan land mammal ages were thought to be Middle and Late Oligocene (they are now both regarded as Early Oligocene in age). Although these problems made the type Oligocene stages hard to correlate to other regions, the use of planktonic microfossils and magnetic stratigraphy has allowed geologists to correlate the classic shallow-marine European stratotypes and terrestrial sections to the global deep-marine standard (**Figure 1**). As a result, the Oligocene is now securely correlated around the world.

In 1989, the Eocene–Oligocene boundary was formally established at the last appearance of the spiny planktonic foraminiferal genus *Hantkenina* in a quarry section near Massignano, Italy. However, later work has since shown that part of the type upper Eocene Priabonian Stage is Early Oligocene by this definition, so there are still problems with this criterion. Most of the important climatic events that many scientists believe should mark the beginning of the Oligocene (e.g., the global oxygen isotope shift indicating the expansion of Antarctic glaciers, and related events such as the cooling on North America) are earliest Oligocene (magnetic Chron C13N, about 33 million years ago) using the hantkeninid criterion. Thus, there are grounds for revising

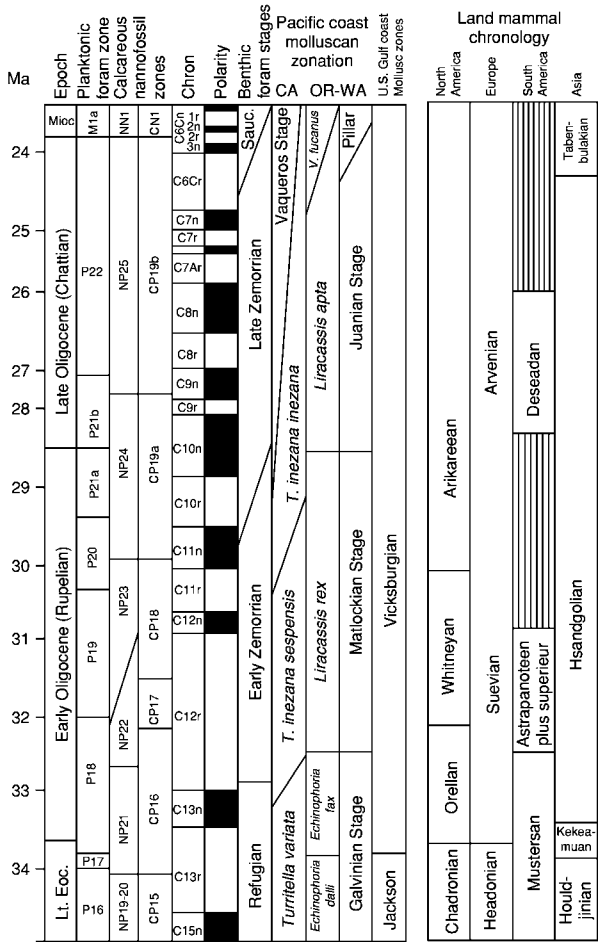


Figure 1 Correlation of various Oligocene biostratigraphic units to the global time scale (left) and magnetic polarity time scale (middle). Global time scale and planktonic zonation (after Berggren *et al.* (1995)). Pacific Coast marine zonation after Prothero (2001). US Gulf Coast molluscan zonation (after Prothero *et al.* (2003)). North American land mammal chronology (after Prothero and Emry (1996)). Asian land mammal chronology (after Meng and McKenna (1998)). European land mammal chronology (after Barbera *et al.* (2001)).

the 1989 definition to a more 'natural' boundary. This would also place the beginning of the Oligocene after the end of the type Late Eocene (Priabonian Stage). However, no such revision has been formally proposed to date.

Only two stages are recognized in the 11-million-year (34–23 Ma) span of the Oligocene. The Early Oligocene Rupelian Stage includes the interval from 34–28.5 Ma. The Late Oligocene Chattian Stage is dated between 28.5–23.8 Ma. There is no formally recognized Middle Oligocene.

Oligocene Climate

The 11 million years of the Oligocene marked an important climatic transition in Earth history. The latest Paleocene and Early Eocene (55–50 million years ago) was the peak of global warming, a 'greenhouse' climate that exhibited the warmest global conditions since the Late Cretaceous. Climates were so warm and mild that crocodilians and temperate plants lived above the Arctic Circle in regions that experienced six months of darkness. Beginning in the Middle Eocene, though, this balmy climate began to transform, with the greenhouse climate gradually changing to a colder, more extreme climate. The end of the Middle Eocene (37 Ma) was marked by a major cooling event, which caused the extinction of many marine organisms adapted to warm, tropical waters.

During the three million years of the Late Eocene (37–33 Ma), there was a slight warming and recovery from the long-term cooling trend. At least three major comet or asteroid impacts struck the Earth in the Middle of the Late Eocene (35.5–36.0 Ma), but these caused no significant changes in climate, nor extinction of any importance. As noted above, the Eocene–Oligocene boundary is now formally recognized by the extinction of hantkeninid foraminifera, although no other major climatic or extinction events occurred at this time. (Note that this invalidates an old idea from the 1970s and 1980s that a 'Terminal Eocene Event' – comparable to the event that ended the Cretaceous – also marked the end of the Eocene).

The most significant climatic event of this interval occurred in the earliest Oligocene (as currently defined, using the hantkeninid criterion), at about 33 Ma. This is now known as the Oi1 event. In the marine record, both benthic and planktonic foraminiferal oxygen isotopic ratios show about a 1.3 per mil increase (Figure 2). It was calculated that about 0.3–0.4 per mil of the change was due to a major expansion of Antarctic ice-sheets that lowered global sea-level by at least 30 m. The remaining 0.9–1.0 per mil is explained by about 5–6°C global cooling, which lowered global mean temperature from as high as 13°C in the Early Eocene and 7°C in the latest Eocene to values just a few degrees above freezing (as a global average – the poles were well below freezing for the first time, while the tropics remains relatively unchanged).

Abundant data suggests that this global cooling event was due largely to the growth of the first major Antarctic ice-sheet since the Permian (over 250 Ma). Drilling on the margin of the Antarctic continent and in oceanic plateaus in the Southern Ocean (e.g., Maud Rise and the Kerguelen Plateau) have produced unmistakable evidence of ice-sheet growth. Not only do the isotopic records show its effect, but many of the sediments drilled from the

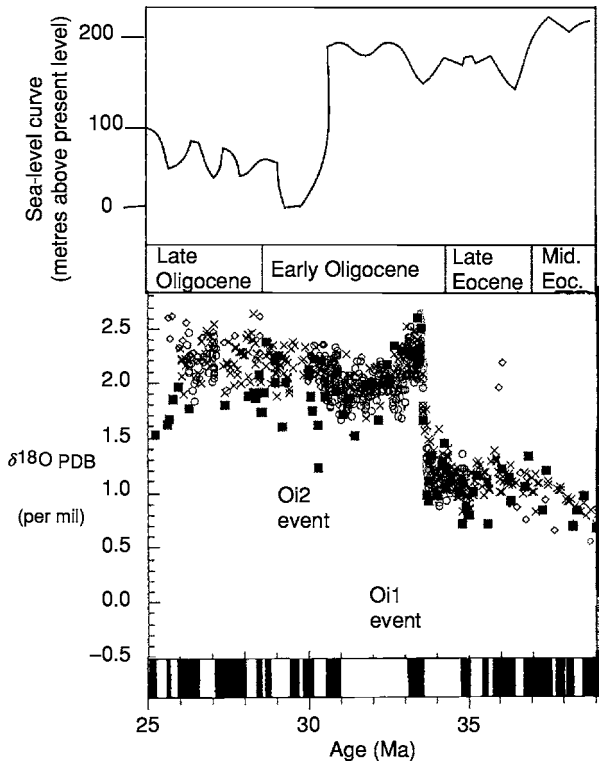


Figure 2 Details of oxygen isotope and sea level record in the Oligocene (modified from Prothero and Dott, 2003, and Zachos *et al.*, 1999).

Antarctic margin are glacial in origin. In addition, there are even observations of sediments dropped by melting icebergs well out into the Southern Ocean.

What caused this global cooling and the extinctions in the Early Oligocene? A few geologists have suggested that the Late Eocene impact events, or major volcanic eruptions in the Ethiopian Plateau, might have been responsible, but these ideas are challenged by the stratigraphical sequence of events. As noted above, the impacts occurred in the Middle of the Late Eocene, about two million years before the Early Oligocene cooling and two million years after the End–Middle Eocene cooling. Likewise, the volcanic eruptions that formed the Ethiopian traps are now dated in the Late Oligocene, when no significant extinctions are recorded. For more than 30 years, the primary mechanism responsible for the Early Oligocene cooling has been identified as the development of the Circum-Antarctic current. Today, this current circulates in a clockwise direction around the Antarctic continent, forming a 'refrigerator door' that locks the cold temperatures formed on the poles. This current is one of the largest in the ocean, moving as fast as 25 cm sec. The volume of water that passes between Antarctica and Australia is about 233 million cubic metres per second, or more than 1000 times the flow of the largest river on Earth, the Amazon. The 'refrigerator door' also separates the polar currents from subpolar and temperate currents, so that each is isolated from the other. By contrast, in the Early Eocene tropical waters in the Atlantic and Pacific mixed all the way to the poles, decreasing the difference in temperature between the poles and the equator.

As these cold waters circulate around the poles, they sink, generating the Antarctic bottom waters. These cold, oxygenated waters then flow along the bottoms of the world's oceans, all the way to the northern hemisphere. The effect of this on global oceanic circulation and climate is enormous. Antarctic bottom waters contain up to 59% of the world's marine water, and transport this cold water all over the bottom of the ocean. This, in turn, increases the stratification of shallow- and mid-level currents in the ocean, further accentuating the climatic differences between pole and equator.

So what triggered the development of the Circum-Antarctic current? The most obvious factor is plate Tectonics (*see* **Plate Tectonics**). In the Late Cretaceous, Australia and South America were still attached to Antarctica as remnants of the ancient Gondwana supercontinent. As noted above, this caused the tropical currents to mix with polar currents along a much broader front since there were barriers between the Atlantic, Indian, and Pacific oceans, and no southern ocean. The climatic effect was expressed by moderate global temperatures. Geophysical evidence shows that these three southern continents began their separation in the latest Cretaceous, but they were not far apart enough to allow deep-ocean currents to pass through until the latest Eocene or Early Oligocene. By the Early Oligocene, deep-sea cores south of New Zealand reveal a blast of cold water was passing between Australia and East Antarctica. As the Oligocene progressed, the between-continent separation grew wider, resulting in the development of larger and more powerful cold-water currents. Originally, geologists thought that the separation between the tip of South America and the Antarctic Peninsula did not occur until the end of the Oligocene, but recent evidence has suggested that this passage also opened in the Early Oligocene. This implies that the entire Circum-Antarctic current developed in a relatively short period of geological time.

In addition to these important currents, it is also thought that another body of water, the North Atlantic Deep Water, which flows out of the Arctic Ocean past Greenland into the bottom of the North Atlantic, originated some time in the Oligocene. Thus, the

global 'icehouse' conditions of the Oligocene can be largely attributed to the development of modern oceanic stratification and circulation patterns.

It has been suggested that a decline in greenhouse gases – especially CO_2 – might be a more important factor than the effects of oceanic circulation changes in steepening Oligocene climatic gradients. This interpretation is based on global circulation models of the atmosphere and oceans, coupled with simulation experiments in which the amount of atmospheric CO_2 was varied in the computer models. Certainly, there must have been a decline in CO_2 from the greenhouse world of the Cretaceous through Early Eocene to the icehouse of the Oligocene. But important questions have not yet been answered. Where is the global reservoir for all this carbon? There were no great bodies of unoxidized carbon locked up in coal deposits such as the one that terminated the Mid-Paleozoic greenhouse. Nor are there extensive Oligocene carbonates that might have locked up atmospheric CO_2. Those who favour this hypothesis suggest that the carbon was locked up in frozen methane hydrates on the sea-floor. Since the evidence of this frozen hydrate might not be preserved in the stratigraphical record, it is difficult to test this idea directly. Several other empirical studies also contradict this computer model. Chemical analyses of foraminifera suggests that Eocene CO_2 concentrations were not much higher than they are today. The number of stomata on the bottom of a leaf is strongly correlated with atmospheric CO_2 concentration, and it is observed that there is no evidence of higher CO_2 levels in Eocene leaf data. It has been argued that methane (CH_4), rather than CO_2, might have been the greenhouse gas responsible for Eocene warming. Given the clear evidence of oceanic circulation changes directly tied to sedimentological changes on Antarctica, and isotopic changes in the oceanic waters, there must have been a significant effect from the opening oceanic gateways and circulation changes.

The effects of these oceanic temperature changes are critical, not only to marine climates and organisms, but also to terrestrial climates as well. The most complete land-based record comes from North America, where palaeobotanical records show that mean annual terrestrial temperatures dropped 7–11°C in the earliest Oligocene. This is true of palaeofloral records all the way from Alaska and the Pacific North-west to the Gulf Coast. In addition to this rapid cooling, the record of ancient plants and soils also suggests that the continent underwent significant drying, with the establishment of much more seasonal, drought-prone climates. In the Big Badlands of South Dakota, Upper Eocene palaeosols (*see* **Soils:** Palaeosols) suggest over a metre of annual rainfall, supporting a dense forest. By contrast, in the Early Oligocene, mean annual rainfall was less than half a metre, and the vegetation was patchy scrubland with limited riparian forests. The land snails from the Badlands also change, from Late Eocene forms like those found today in tropical Central America, to Early Oligocene forms that are smaller and more drought-tolerant, and found today in Baja, California. In addition, the Late Eocene reptilian fauna that was dominated by crocodilians and pond turtles was replaced by dry land tortoises in the Early Oligocene.

Once the Early Oligocene climatic deterioration was completed, the Earth remained in this icehouse mode through the remainder of the Oligocene. The other significant Oligocene climatic event was several pulses of glaciation that occurred during the middle part of the Oligocene, about 30 Ma (the Oi2 event). Thick, extensive Mid-Oligocene glacial deposits are found throughout the Antarctic region, and benthic foraminiferal oxygen isotopes shifted by 1.6 per mil, suggesting another increase in ice volume and drop in global temperatures (**Figure 2**). As these ice-sheets grew, they pulled water out of the oceans, resulting in the largest drop in sea-level in the past 100 million years. Originally, it was suggested that sea-level dropped by almost 150 m, although more recent estimates suggest it was only half that amount (**Figure 2**). Whatever its magnitude, the Mid-Oligocene regression had a major effect on the shallow-marine realm, causing the continental shelves to become deeply incised once they were exposed to the subaerial erosion, and producing huge Mid-Oligocene unconformities in most marine rocks around the world.

The effect of the cooling and regression in the Middle Oligocene on land climates was less obvious. Sensitive tropical floral elements were already gone by the Mid-Oligocene, so the land plant record shows only minor cooling effects. The record of ancient soils from the Big Badlands of South Dakota shows that the climate became cooler and much drier, so that sand dune deposits became common in the Mid-west in the late Oligocene. These same soils suggest that vegetation was a mixture of scrublands and grasses, with few trees, by the Late Oligocene.

Oligocene Life

As the Eocene–Oligocene climatic deterioration began, the total diversity of land animals and both marine organisms decreased significantly from Eocene levels, reaching a Phanerozoic low in the Late Oligocene. The forests and jungles of the Early

Eocene were rapidly disappearing by the Late Eocene, so that by the Oligocene most of the temperate latitudes were covered by a mixture of forest and scrubland vegetation. This change in vegetation triggered by this cooling and drying was a major change in many of the land organisms. The Oligocene land-mammal fauna was dominated by primitive members of most living families. These included three-toed horses (which began to radiate into multiple lineages by the Late Oligocene), three different lineages of rhinoceroses, early camels, deer, and peccaries, as well as a handful of archaic land-mammal groups left over from the Eocene. Numerous modern carnivoran groups (especially early dogs, and the cat-like nimravids, as well as primitive members of the bear, weasel, and raccoon families) became the dominant predators as the last of the archaic carnivorous mammals (*see* **Fossil Vertebrates:** Placental Mammals), the creodonts, straggled on. On all the northern continents and Africa, rodents and rabbits both underwent a huge diversification as the niches for ground-dwelling seed-eaters increased, and the habitat for squirrel-like nut and fruit eaters diminished.

In Eurasia, many of the same trends were apparent. In the Early Oligocene, the Turgai Strait across the Obik Sea between Europe and Asia opened up, allowing Asian mammals (such as rhinoceroses and ruminants) to immigrate to Europe and drive many of the endemic natives to extinction. This Early Oligocene event is known as the Grande Coupure. However, there was only limited migration between Asia and North America via the Bering Strait. In Eurasia, the Oligocene saw a similar diversification of rhinoceroses (including one group, the giant indricotheres found in Mongolia and Pakistan, which reached 6 m at the shoulder and weighed 20 tonnes), plus some of the earliest members of the deer, giraffe, pig, and cattle families. Tree-dwelling mammals became much less common and vanished from many continents. For example, primates once flourished on all the northern continents during the Early and Middle Eocene. By the Oligocene, though, they became restricted to Africa and South America, where they evolved into Old World and New World monkeys, respectively. The remainder of the African fauna was also endemic to this island continent, which was not connected to Eurasia at the Arabian Peninsula until the Early Miocene. Instead, the African fauna was populated by archaic mastodonts, a wide diversity of hyraxes, and other peculiar endemic forms, such as the horned arsinoitheres. South America and Australia were also island continents, unconnected to the rest of the world, and each developed their own endemic faunas.

In the marine realm, the Early Oligocene extinctions triggered by global cooling were severe, causing major extinction in the planktonic and benthic foraminifera, and even in the planktonic algae (*see* **Fossil Plants:** Calcareous Algae) such as diatoms and coccolithophores. In the Gulf Coast of the US, 97% of the marine snail species and 89% of the clam species found in the Late Eocene did not survive into the late Early Oligocene, and over 50% of the sea urchins and sand dollars also became extinct. However, the overall taxonomic composition of the marine fauna remained essentially the same, with new species of clams, snails, and sea urchins replacing the extinct species (but at lower diversity), and making up the bulk of the fossilisable organisms in the Oligocene. By the end of the Early Oligocene, diversity was at an all-time low. Marine faunas were composed of groups tolerant of the cooler waters that began in the Oligocene. This is true especially in the molluscs (*see* **Fossil Invertebrates:** Molluscs Overview) of the Pacific Rim, which are mostly cold-water tolerant forms that migrated south to California from Alaska and Siberia during the Oligocene. Planktonic organisms were not only low in diversity, but occupied relatively few, simple biogeographic realms (since the area of the tropics had decreased), and evolved relatively slowly during the Oligocene.

Palaeogeography

With the Eocene separation of Australia from Antarctica and the collision of India with Asia, by the Oligocene most of the continents were approaching their present configuration. South America, however, would not finally separate from Antarctica until the beginning of the Oligocene, completing the breakup of Gondwana and opening the gateway for the full development of the Circum-Antarctic current. As discussed above, these continental movements brought about major changes in oceanic circulation, with the Circum-Antarctic current locking in cold conditions over Antarctica, initiating the first Antarctic ice-sheets, and also stimulating the flow of cold Antarctic bottom waters, which today control much of the world's oceanic circulation.

The growth of Antarctic glaciers meant that the high sea-levels of the Eocene greenhouse world were gone, and much of the seawater locked up in ice. The Late Oligocene regression turned the drowned coastal plains of the Atlantic and Gulf coast of North America into emergent floodplains, and the European archipelago largely dried up. This regression also dried up the Obik Sea and ended the separation of Europe and Asia. The Tethys Seaway was already partially disrupted by the collision of India

with Asia, but global regression destroyed the remaining vestiges of this seaway and its unique tropical biota.

On land, the Himalayas continued to develop, and the Alps began to rise rapidly as Africa began to collide with Europe and close the Mediterranean. The Andes began to erupt huge volumes of volcanic rocks, forming a mountain chain for the first time. In North America, the Rocky Mountains were no longer rising, but they continued to soar high above the western part of North America. The basins between the ranges began to fill up with sediments and volcanic debris erupted from the arc volcanoes to the west. Volcanic activity on the western edge of the continent, which had ceased when the Laramide Orogeny shut off the Sierran-Sevier arc, resumed in the Oligocene. The new arc was much further east than the previous arc, running in an irregular belt from central Mexico to New Mexico (the Mogollon-Datil volcanics) to south-west Colorado (the San Juan volcanics), Utah, and Nevada, and then up through Oregon and Washington (the ancestral Cascades), and British Columbia. These explosive volcanic centres erupted huge amounts of ash, much of which blew east and helped bury the Laramide uplifts of the Rocky Mountains. Much of this ash and sediment also spilled over onto the High Plains, forming the thick Oligocene deposits of the White River and Arikaree Groups (entombing their excellent record of fossils (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids)) and climates (*see* **Palaeoclimates**) in the Big Badlands of South Dakota and adjacent states. At about 30 Ma, the corner of the Pacific Plate was subducted under California, so the San Andreas transform fault began its activity. This, in turn, set off a wide variety of geologic events, including the beginning of the spreading of the Basin and Range Province in Arizona, Utah, and Nevada; the end of the eruptions in Nevada and California; the clockwise rotation of the Sierra-Cascade arc to the south-west by over 400 km; and the clockwise rotation of many tectonic blocks, including the Transverse Ranges of California. By the Miocene, all of these regional events (Basin and Range extension, Sierran rotation; cessation of southern volcanism; and Transverse Ranges rotation) were well developed and approaching their modern condition.

See Also

Famous Geologists: Lyell. **Fossil Invertebrates:** Echinoderms (Other Than Echinoids); Molluscs Overview. **Fossil Plants:** Calcareous Algae. **Fossil Vertebrates:** Jawless Fish-Like Vertebrates; Placental Mammals. **Palaeoclimates**. **Plate Tectonics**. **Soils:** Palaeosols. **Tertiary To Present:** Eocene.

Further Reading

Barbera X, Carbrera L, Marzo M, Pares JM, and Agusti J (2001) A complete terrestrial Oligocene magnetobiostra tigraphy from the Ebro Basin, Spain. *Earth and Planetary Sciences Letters* 187: 1 16.

Berggren WA, Kent DV, Swisher CC III, and Aubry M P (1995) A revised Cenozoic geochronology and chronostratigraphy: *SEPM Special Publication* 54: 129 212.

Davies R, Cartwright J, Pike J, and Line C (2001) Early Oligocene initiation of North Atlantic deep water forma tion. *Nature* 410: 917 920.

Diester Haass L and Zahn R (1996) The Eocene Oligocene transition in the Southern Ocean: history of water masses, circulation, and biological productivity inferred from high resolution records of stable isotopes and ben thic foraminiferal abundances (ODP Site 689). *Geology* 24(2): 16 20.

DeConto RM and Pollard D (2003) Rapid Cenozoic glaci ation of Antarctica induced by declining atmospheric CO_2. *Nature* 421: 245 249.

Exon N, *et al.* (2002) Drilling reveals climatic consequences of Tasmanian gateway opening. *EOS* 83(23): 253 259.

Meng J and McKenna M (1998) Faunal turnovers of Palaeogene mammals from the Mongolian plateau. *Nature* 394: 364 367.

Miller KG (1992) Middle Eocene to Oligocene stable iso topes, climate and deep water history: the Terminal Eocene Event? In: Prothero DR and Berggren WA (eds.) *Eocene Oligocene Climatic and Biotic Evolution,* pp. 160 177. Princeton, NJ: Princeton University Press.

Pearson PN and Palmer MR (1999) Middle Eocene sea water pH and atmospheric carbon dioxide concentra tions. *Science* 284: 1824 1826.

Prothero DR (1994) *The Eocene Oligocene Transition: Paradise Lost*. New York: Columbia University Press.

Prothero DR (ed.) (2001) *Magnetic Stratigraphy of the Pacific Coast Cenozoic*. Pacific Section SEPM Special Publication 91.

Prothero DR and Berggren WA (eds.) (1992) *Eocene Oligo cene Climatic and Biotic Evolution.* Princeton, NJ: Princeton University Press.

Prothero DR and Dott RH Jr (2003) *Evolution of the Earth,* (7th edn.) New York: McGraw Hill.

Prothero DR and Emry RJ (eds.) (1996) *The Terrestrial Eocene Oligocene Transition in North America*. New York: Cambridge University Press.

Prothero DR, Ivany LC, and Nesbitt ER (eds.) (2003) *From Greenhouse to Icehouse: The Marine Eocene Oligocene Transition.* New York: Columbia University Press.

Retallack GJ (1983) Late Eocene and Oligocene palesols from Badlands National Park, South Dakota. *Geological Society of America Special Paper* 193.

Royer DL, Wing SL, Beerling DJ, Jolley DW, Koch PIL, Hickey LJ, and Berner RA (2001) Paleobotanical evidence for near present day levels of atmospheric CO_2 during part of the Tertiary. *Science* 292: 2310 2313.

Sloan LC, Walker JCG, Moore TC Jr, Rea DK, and Zachos JC (1992) Possible methane induced polar warming in the early Eocene. *Nature* 357: 320 322.

Wolfe JA (1978) A paleobotanical interpretation of Tertiary climates in the Northern Hemisphere: *American Scientist* 66: 694 703.

Wolfe JA (1994) Tertiary climatic changes at middle latitudes of western North America: *Palaeogeography, Palaeoclimatology, Palaeoecology* 108: 195 205.

Zachos JC, Opdyke BN, Quinn TM, Jones CE, and Halliday AN (1999) Eocene Oligocene climate and seawater $^{87}Sr/^{86}Sr$: is there a link? *Chemical Geology* 161: 165 180.

Miocene

J M Theodor, Illinois State Museum, Springfield, IL, USA

Introduction

The Miocene (23.8–5.3 Ma) is the interval during which the world began to assume much of the configuration and topography we know today. Major tectonic changes in North America uplifted the Coast Ranges, formed the Basin and Range, and saw the evolution of the San Andreas Fault Zone. In Eurasia, the ongoing collision of India and Asia elevated the Tibetan Plateau, drastically altering global atmospheric circulation patterns. In South America, global atmospheric and oceanic circulation were greatly affected by the uplift of the Andes and the closure of the Panamanian seaway, in ways that persist to this day. The Miocene separation of Australia and Antarctica altered oceanic circulation patterns, changing the prevailing currents off Europe. In Africa, the development of the Great Rift Valley formed the environments that were home to the early hominids (*see* **Fossil Vertebrates:** Hominids). Much of the flora and fauna of the Miocene belong to groups that are still living today, but several groups experienced large radiations during this time interval, such as the whales and the snakes. Two major ecosystems – grasslands and kelp forests – evolved during the Miocene. These changes form the foundation for modern conditions.

The Miocene Epoch is the fourth subdivision of the Tertiary Period. In 1828, Sir Charles Lyell (*see* **Famous Geologists:** Lyell) noticed that different layers of rock in Europe contained different proportions of living and extinct species of marine molluscs. Concluding that the variation in this proportion indicated that the rock layers were formed at differing times, and assuming that the higher the proportion of living to extinct species, the more recently the layer was laid down, he defined three epochs – the Eocene (Dawn Recent), Miocene (Middle Recent), and Pliocene (More Recent), where the transitions between epochs were specified as a given percentage of living to extinct species.

Because the species composition of a rock layer can vary depending on the type of depositional environment, geologists no longer use this definition, and rely on other methods. Two specific, internationally recognized localities (also known as stratotypes) are currently designated as being the boundaries of the Miocene for international correlation. The base, or start, of the Miocene is defined at the Lemme Carrosio section in the upper part of the Rigoroso Formation in northern Italy. The base of the Pliocene – which is also, by definition, the end of the Miocene – is defined at the base of the Trubi Formation in the Evaclea Minoa section, on the southern coast of Sicily (*see* **Tertiary To Present:** Pliocene).

Geochronology

The European Miocene is generally divided into six marine stages (oldest to youngest): the Aquitanian, Burdigalian, Langhian, Serravallian, Tortonian, and Messinian (Figure 1). Terrestrial fossils in Europe have generally been associated with these marine stages as European marine rocks interfinger with terrestrial deposits, thus allowing a clear correlation between terrestrial and marine units. However, a number of more isolated fissure-fill localities are more difficult to correlate. As a result, two other systems based exclusively on terrestrial mammals have been developed, the European Land Mammal Ages (ELMAs) and the Mammalian Neogene Reference Level system (MN Zones).

In North America, there are fewer localities with interfingered marine and non-marine beds, but many units are associated with strata that can be radioisotopically dated. The majority of these deposits are correlated using mammalian biostratigraphic units. The North American Land Mammal Ages in use for the Miocene include the Arikareean, Hemingfordian, Barstovian, Clarendonian, and the Hemphillian. Most of these are well associated with radioisotopic dates and/or paleomagnetic assessments.

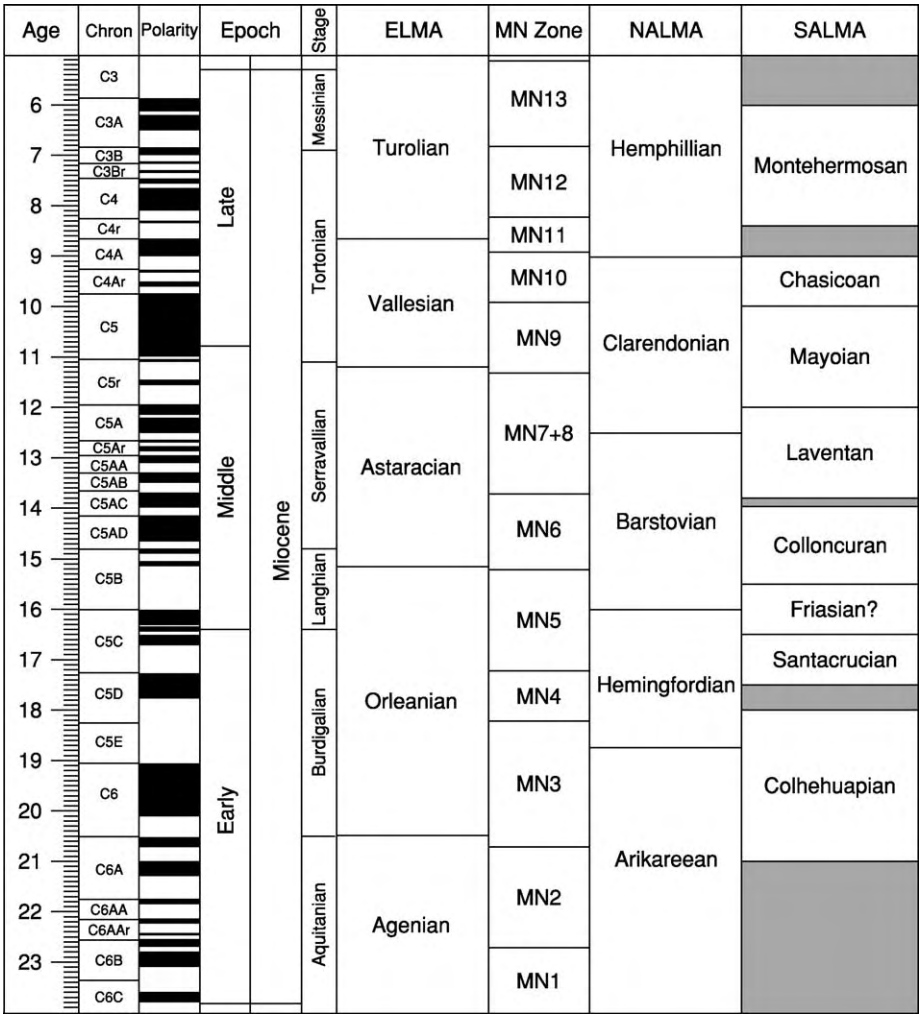


Figure 1 Chronostratigraphical correlation chart for the Miocene. ELMA European Land Mammal Age, MN Mammalian Neogene Reference Level Age, NALMA North American Land Mammal Age, SALMA South American Land Mammal Age.

Many regions are not as well described as Europe and North America. The South American Miocene Land Mammal Ages include the Colhehuapian, Santacrucian, Colloncuran, Friasian, Laventan, Mayoan, Chasicoan, Huayquerian, and Montehermosan. Some of these intervals are poorly constrained in terms of absolute ages. Asian and African faunas do not yet have a well-established system of named ages in use, but much work is being done to establish such a framework. The Australian Miocene can be correlated to the other continents using pollen and marine fossils, but the mammals are so different from those on other continents that they are not used in correlation outside of that continent.

Tectonics

Continental position During the Miocene, the continents moved into positions very close to their modern configuration (Figure 2). By the mid-Miocene, as Australia moved away from Antarctica, the modern oceanic circulation pattern was established. Miocene tectonic changes resulted in the creation of numerous important features of modern continental topography.

North America

North America underwent a complex sequence of tectonic events throughout the Cenozoic. A number of events occurred during the Miocene which combined to shape much of the landscape familiar to us today.

East Pacific Rise Spreading at the East Pacific Rise continued during the Oligocene, causing the complete subduction of the Farallon Plate near Los Angeles by 30.0 Ma, leaving two remnants to the north and south still undergoing subduction. By the Early

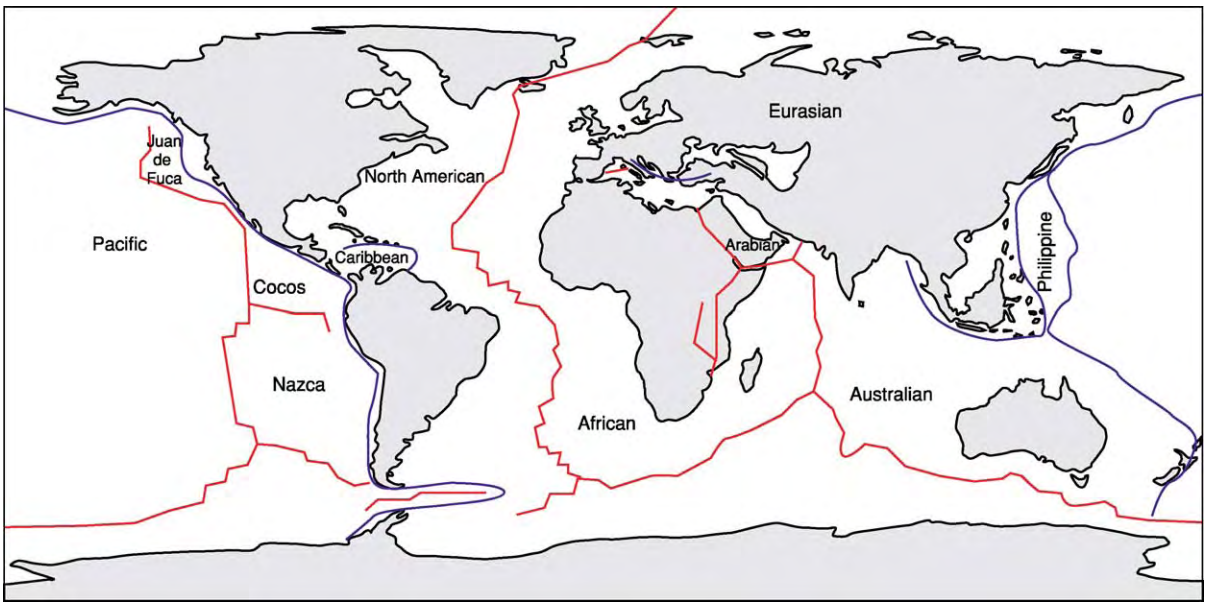


Figure 2 Palaeogeographical reconstruction of Miocene plate configurations and continent dispersions.

Miocene, most of the remaining Farallon Plate had been subducted under the North American Plate, bringing the Pacific Plate into contact with the North American Plate in southern California. The remaining two sections could no longer be considered a single plate by this time, and are henceforth regarded as independent plates with unique motions. The Juan de Fuca Plate is found to the north off the coast of Oregon and Washington and the Cocos Plate to the south off the coast of Mexico. Both continue to be subducted under the North American Plate. The contact of the Pacific, North American, and Juan de Fuca plates formed a triple junction at Mendocino, California by 20.0 Ma. During this transition, the subducting plate margin disappeared and the contact between the Pacific and North American plates developed into a transform fault system because the Pacific Plate is moving to the north, while the subducting Farallon Plate was moving more directly eastward. This transform fault system, which separates Baja California from the rest of California, runs under the continent between San Francisco and Los Angeles. The transform faulting formed along the Central Valley of California is known today as the San Andreas Fault Zone, and was in place by about 10.0 Ma.

Rocky Mountains The subduction of both the Farallon Plate and the more northerly Kula Plate are partially responsible for the character of the Rocky Mountains, which began their final uplift phase during the Miocene. The Farallon Plate was subducted under the North American Plate at an unusually shallow angle. This caused uplift of the Rockies in the western US and Mexico by compression, and probably caused some extension in the crust as the slab sank. By contrast, the Juan de Fuca and Kula plates were subducted at a much steeper angle (closer to 45°), resulting in more typical thrusted sedimentary sheets through the Canadian Rockies.

Basin and Range Changes in the continental crust to the west of the developing Rocky Mountains are poorly understood, but the thin crust in this region is evidence that this area became extended between 15.0 and 8.0 Ma. Crustal extension resulted in a series of faults that trend north–south, with dropped fault blocks (grabens) forming flat valley floors in between raised blocks (horsts). Erosion of the nearly 400 alternating mountain blocks formed the characteristic basin and range topography of the modern western United States.

Columbia River Flood Basalts From 17.0 to 15.5 Ma, a tremendous volume of volcanic rock was deposited in north-eastern Oregon, south-eastern Washington, and western Idaho, from at least 300 lava flows, some of which reached as far as the Pacific Ocean. The Columbia River Flood Basalts cover 164 000 km^2 and are situated between the Rocky Mountain and Cascade Ranges. The cause of these volcanic eruptions is complex and is tied to: (i) extension and thinning of the crust; (ii) formation of the Cascade Range; and (iii) the position of the Yellowstone hotspot (a mantle plume). Most flood basalt provinces are directly tied to a mantle hotspot, but the Yellowstone

hotspot was 300–400 km south of the flood basalt vents, and the basalt composition differs from classic hotspot flood basalts, indicating a more complex tectonic origin. The numerous basalt beds allow for precise radioisotopic age assessments for this area during the Middle Miocene.

Cascade and Sierra Nevada Ranges There is some evidence that these volcanic mountain ranges experienced a pulse of uplift beginning in the Late Miocene. Because there is also some evidence of uplift in the Coast Ranges at this time, it is possible that there was uplift of the entire north-western Cordilleran coast range. This uplift might have been a result of changes in plate motion, or of plate delamination and subsequent crustal rebound.

Central and South America

Two major events affected the topography of Central and South America during the Miocene, both of which had profound climatic and biogeographic effects: the beginning of closure of Panamanian seaway and the uplift of the Andes.

Isthmus of Panama The closure of the Panamanian seaway began at almost 13.0 Ma, with the gradual emergence of the Isthmus of Panama as a result of ongoing changes in tectonics and sea-level. The emergence of the isthmus resulted in separation of North and South America, and resulted in profound changes in Caribbean oceanic circulation in the Pliocene, cutting off circulation between the Atlantic and Pacific oceans. These drastic changes in oceanic circulation resulted in massive changes in oceanic heat transfer. The isolation of the Atlantic Ocean also helped produce the warm Gulf Stream current that today maintains warmer temperatures in northern Europe relative to other locations at the same latitude. It is probable that this change also affected northern-hemisphere glaciation.

Andes The major topographic feature of South America is the Andes Range, which provides a barrier to atmospheric circulation and creates a large rain shadow desert on its eastern margin. The Andes are unique because they formed at a non-collisional plate boundary. Crustal compression of the Andes region began in the Triassic, but by in the Early Miocene (~20.0 Ma) the majority of the range had reached only 25% of current elevation. By the later Miocene, the range had reached ~50% of the modern elevation, and uplift continued. The tectonic history of the Andes is complex and not yet well understood (*see* **Andes**).

Eurasia

The tectonic history of Eurasia is somewhat less complex than that of North America. Nevertheless there were several important Eurasian tectonic events that occurred during the Miocene.

Himalayas and the Tibetan Plateau Earlier in the Cenozoic, the northward plate motion of India brought that continental plate into contact with Eurasia (*see* **Indian Subcontinent**). The resulting continent-continent collision continued through the Miocene, causing uplift of the Himalayas, and, beginning in the Early–Middle Miocene, resulted in uplift of the Tibetan Plateau. This uplift exposed a considerable volume of carbonate rock to weathering, which is thought have affected global atmospheric CO_2 levels.

Messinian salinity crisis Considerable evidence from thick deposits of halite and anhydrite beneath the Mediterranean Sea indicate that, beginning at 5.96 Ma, the Mediterranean basin was repeatedly cut off from oceanic input and became a drying salt lake (Lago Mare) system. This dry-Mediterranean interval is termed the Messinian Salinity Crisis. It continued until 5.33 Ma, when the modern opening at the Strait of Gibraltar restored marine conditions. The cause of this Messinian Salinity Crisis is unclear. The likely major causes include several tectonic changes and a global decline in sea level of ~60 m as a result of glaciation. The most recent data favour a tectonic explanation, in which changes in the subduction of oceanic lithosphere of the Tethyan Ocean caused upwelling in the asthenosphere, resulting in uplift and consequent isolation of the Mediterranean Basin (*see* **Europe:** Mediterranean Tectonics).

Africa

During the Miocene, Africa, which had previously been relatively tectonically stable, underwent a series of tectonic changes as a result of the formation of a new spreading ridge running through the eastern edge of that continent. In the early part of the Miocene, the Afro-Arabian Plate moved north, contacting the Eurasian Plate. This contact reduced the size of the Mediterranean and may have contributed to the Messinian Salinity Crisis by reducing the amount of water available for rainfall and thus increasing aridity.

By the Early Miocene, uplift had begun in southern and eastern Africa, creating a rain shadow over eastern Africa. Around 18.0 Ma crustal extension from the spreading ridge began to take place along the East African Rift Valley, which forms one arm of the Afar Triple Junction – the other arms being the Gulf of

Aden and the Red Sea. Extension caused faulting and the formation of a horst-graben topography in the rift valley. This process fragmented the habitats of the rift valley, breaking up forest, increasing the diversity of the terrain and vegetation, and leading to the formation of numerous lakes along the valley floor. These lakes, such as Lake Turkana, provided drinking water and habitat for the early ancestors of humans. The lakes also provided better opportunities for preservation in the fossil record, which is much less likely in densely vegetated forest.

Climate

Sea-level Miocene sea-levels were generally higher than those of the modern day, and are tied to glaciations in Antarctica. During the Miocene, oceanic circulation patterns were strongly affected by the evolution of the circum-Antarctic current as Antarctica separated from Australia. This separation thermally isolated Antarctica and drastically altered global marine and atmospheric circulation.

In North America, much of the south-eastern United States was covered by shallow seas. The Early Miocene saw three major episodes of deposition of oceanic phosphorus, especially in eastern North America, when much of the eastern coast was underwater. These periods (21.0 to 20.0, 19.0 to 18.0, and 17.0 to 16.0 Ma) are likely to have been responses to global changes in sea-level, oceanic upwelling, and deep water currents, and may be related to the ongoing uplift and erosion of the Tibetan Plateau. Later in the Miocene (~6.0 Ma) sea levels dropped as much as 60 m because of glaciation, a drop which probably contributed to the Messinian Salinity Crisis.

Atmospheric CO_2 Early models of atmospheric CO_2 indicated that levels in the earlier Miocene were roughly double those of modern, pre-industrial times. However, more recent measurements of geochemical signatures (e.g., boron isotopes, plant alkenones) from marine sediments indicate that Early Miocene CO_2 levels were much lower. This is important because the hypothesised decline in the middle Miocene CO_2 levels have been thought responsible for the rapid spread of grasses that use the C4 photosynthetic pathway at that time. Plants that use only one enzyme to bind CO_2 in the chemical reactions of photosynthesis are termed 'C3' plants. However, this enzyme will also bind O_2, forming a compound that is harmful to the plant. At low concentrations of atmospheric oxygen this is not a significant disadvantage to C3 plants. The C4 pathway prevents these compounds from forming, but produces slightly less energy for the plant. At lower CO_2 levels, the C4 photosynthetic pathway is favoured, and it is generally assumed that the C4 pathway would only have evolved in intervals with lowered CO_2. The evolution of the C4 pathway allowed grasses to spread greatly during the Miocene, forming extensive grassland ecosystems for the first time.

Atmospheric CO_2 is a product of a number of different factors. One factor that is hypothesised to have affected global Miocene CO_2 levels is the uplift and subsequent erosion of the Tibetan Plateau, which is thought to have consumed massive amounts of atmospheric CO_2 as a result of limestone weathering.

Temperature and seasonality Global temperatures and seasonality are inferred from three sources: the ratio of oxygen and carbon isotopes measured in the skeletal fossils of benthic and microscopic planktonic organisms living in the ocean, and from the types of land plant fossils and the shapes of their leaves. Ice-rafted debris in the South Pacific indicates that Antarctic glaciation intensified at ~24.0 Ma. This was followed by a long warming interval, reaching a climax (Mid-Miocene Climatic Optimum) at ~15.0 Ma, with some brief episodes of glaciation in Antarctica. The warm interval was generally a continuation of Oligocene warm, temperate to sub-tropical conditions. Evidence from terrestrial floras indicates a mean annual temperature range from 11°C to 17°C in the northern United States during the Climatic Optimum, with the first grass macrofossils being found in the Great Plains. Following the Climatic Optimum, global temperatures decreased as the Antarctic ice sheets returned, and aridity increased in terrestrial ecosystems. In North America, temperatures warmed again ~8.0 Ma. A major glacial advance marked the end of the Miocene, resulting in lowered sea-levels.

Marine Life

Although the species found in the Miocene are largely extinct, most of the residents of Miocene oceans would be familiar to us, as they belong to modern groups of organisms (e.g., corals, bivalves, sharks, whales, sea cows). One major change in the oceans during the Miocene was the evolution of the kelp forest ecosystem. Fossil kelps and associated animals are known from the mid-Miocene Monterey Formation in southern California, their first appearance in the fossil record. Another important ecological development was the evolution of algal ridge formation on coral reefs by coralline algae. Earlier, coralline algae had formed parts of coral reefs, but had not formed these ridges, which protect living reef systems from heavy surf conditions. The appearance of the algal ridge systems allowed the expansion of coral reef

systems into more coastal environments. During the Miocene, there was also a large radiation of species of whales and dolphins, including the first appearance of the sperm and baleen whales.

Terrestrial Life

Most of the animals present in the Miocene in North America and Eurasia, and to some extent, Africa, would be generally familiar to the modern eye, as they also belong to extant groups, even though the particular species are extinct. However, many groups were living in regions they no longer occupy today. For example, although North America lacked many groups of mammals currently found there (e.g., pigs), various species of camels and proboscideans were common components of North American Miocene faunas. These three continents had sufficient connections so that animals and plants could migrate between them at various times during the Miocene. By contrast, Australia and South America were island continents during this time interval, and their vertebrate faunas evolved in virtual isolation. Large groups of species that evolved on these continents are extinct, with no extant relatives.

Plants

In the earlier part of the Miocene, up to the time of the mid-Miocene Climactic Optimum, broad-leafed evergreen vegetation and coniferous forest were broadly distributed, and broad-leafed deciduous forest was reduced in area. Cooling and drying in global climates resulted in the establishment and spread of savannahs and grasslands at the expense of forested ecosystems.

By 20.0 Ma, all the families of plants present in the fossil record belong to families still living today. By 10.0 Ma all the modern genera were present. Thus, the Miocene vegetation was composed of familiar elements, although the vegetational structure and distribution were different from their modern states.

Early in the Miocene, forests were widespread. As climate warmed, up until the time of the mid-Miocene Climactic Optimum, broad-leafed evergreen forests (similar to the type found in modern tropical forests) spread to higher latitudes. During the subsequent cooling, broad-leafed evergreen forests contracted their range, giving way to the spread of broad-leafed deciduous forests, and in higher latitudes, evergreen coniferous forests.

Although grass pollen is present in the fossil record much earlier than the Miocene, fossil remains of grass plants are rare elements of the flora until the Middle Miocene. During the Middle and Late Miocene, species diversity in the grasses increased dramatically, and evidence of more open savannah ecosystems appeared in arid regions. The greatest increases in grass diversity were among the groups of species utilising a C4 photosynthetic pathway. This has been interpreted as response to increasing aridity during the later Miocene and possibly as a response to decreasing levels of atmospheric CO_2. The spread of grasslands was facilitated by climactic cooling and drying, particularly as seen in rain shadows created by uplift in the Cordilleran regions of the Americas.

Animals

The fossil record of terrestrial animals in the Miocene is variable. The limited fossil record of insects shows that most of the modern genera were present by the Miocene.

The reptile fauna worldwide was generally similar to that of today, although the Miocene is characterized by a major radiation in the snakes. This snake radiation seems to be a response to the diversification of small mammals, especially murid rodents (rats and mice) during this time.

The Miocene also saw diversification of many groups of birds, especially the dominant modern group of songbirds, the passerines. The modern genera of owls also appeared during the Miocene, along with the appearance of daytime predators such as falcons, hawks, and eagles. Several species of very large vulture-like predators (vultures and teratorns) appeared in the Miocene in the New World, probably evolving in response to the evolution of large herbivores living in open grasslands.

In the island continents of South America and Australia, a number of flightless forms evolved, and in South America one family, the phorusrhachids, was an important terrestrial predator. By the Middle Miocene, the modern ostrich *Struthio* had appeared in Africa.

North America The Miocene mammalian history of North America is documented by a rich and well-dated fossil record, including 31 families of mammals. In the Early Miocene, faunas are dominated by ungulates: the even-toed artiodactyls, including oreodonts, camels, and peccaries, and a new family, the cervids (deer and their allies); and the odd-toed perissodactyls, with numerous species of horses, rhinos, tapirs, and a now-extinct family, the large, clawed chalicotheres. Carnivores increased in diversity in the Early Miocene, with increases in the canid (dog), ursid (bear), and amphicyonid (extinct bear-dog) families. Rodent diversity increased in the geomyids (gophers), sciurids (squirrels), and castorids (beavers). At the time of the Middle Miocene Climatic Optimum, ungulate diversity was much higher than it is today, with a much higher proportion of browsing (leaf-eating) species

than known in any ecosystem today, where ungulate faunas are dominated by grazing (grass-eating) mammals. By the Late Miocene, many browsing ungulate species had become extinct.

The Miocene was also punctuated by interchanges with Eurasia. In the Early Miocene, the first of two pulses of immigration across the Bering Land Bridge brought antilocaprids (pronghorn) and mustelids (weasels and skunks) to North America. The first proboscideans (elephant relatives) appeared ~16.0 Ma, and 11.0 Ma the horse *Cormohipparion* arrived in Europe from North America.

Eurasia The Early Miocene mammalian faunas of Europe are not well differentiated from earlier Oligocene faunas. Proboscideans immigrated to Europe from Africa by 18.0 Ma, and by 17.5–15 Ma dispersal routes connected Europe with Asia, Africa, and North America, bringing the horse *Anchitherium* and the chalicotheres to Europe in the Early Miocene. A major immigration (~15.0 Ma) brought the bovids (cattle) and suids (pigs) to western Europe from western Asia and Africa. The later Miocene of Eurasia is characterized by a diverse open country woodland fauna, which included horses, rhinos, bovids, felids (cats), hyaenids (hyaenas), and proboscideans.

The fauna from southern Asia is known primarily from the Siwalik sequence of Pakistan, and differs from European and western Asian faunas. The Pakistani faunas contain more primates and omnivorous artiodactyls, and fewer browsing ungulates than in western Asia, and a number of archaic groups persisted there. About 9.5 Ma, large giraffids (giraffes), suids (pigs), and new horses emigrated from Eurasia. Around 7.4 Ma, the fauna experienced turnover, and at the same time, soil carbonate measurements indicate a shift in the ecosystem away from plants using the C3 metabolic pathway to grasses using C4 pathways, indicating a shift in environment to a drier climate.

Africa The Early Miocene of Africa was a time of faunal transition, as 29 new families appeared during this time interval. The fauna at this time consisted of endemic proboscideans, creodont carnivores, hyraxes, giraffids, bovids, and aardvarks evolved early in the Miocene. Perissodactyls arrived from Eurasia in the form of rhinos and chalicotheres; archaic suid artiodactyls (pigs) represented the artiodactyls, with viverrids (civets) and felids representing the carnivores.

During the Middle Miocene, the bovids radiated and came to dominate the African ungulate fauna, and the hippos evolved. The first horse, *Anchitherium*, emigrated from Eurasia, along with mustelid and hyaenid carnivores. The Late Miocene saw extensive interchanges with Eurasia.

South America The faunas from South America are comparatively alien, consisting of diverse marsupials, including large marsupial carnivores (borhyaenids), xenarthrans (armadillos, glyptodonts, sloths) and a large number of extinct endemic animals, mostly ungulates, including the large toxodonts, which resemble hippos, rabbit-like hegetotheres, and the camel and horse-like litopterns. In the Early Miocene, two groups entered South America, probably by waif dispersal from Africa. These were the New World monkeys and the caviomorph rodents (guinea pigs, porcupines). Rodents were the only placental mammals in South America until the Pliocene, and they radiated into 16 families. Six to eight million years ago the first true carnivore, a procyonid (raccoon) entered South America, and around the same time, a sabre-toothed marsupial carnivore, *Thylacosmilus*, evolved. This fauna was to change tremendously in the Pliocene, when full-scale interchange began with North American faunas.

Australia As in South America, the Australian fauna was entirely isolated from other continents, and a highly endemic fauna evolved from the original mammalian stock, consisting of marsupials and monotremes. During the Miocene, the climate in Australia was much wetter than it is today, as evidenced by the Riversleigh fauna. Riversleigh dates to ~15.0 Ma, and the fauna includes the marsupial wolf (thylacine) and marsupial lion *Wakaleo*, quolls, possum-like marsupials, early kangaroos and diprotodonts, a group of large herbivorous marsupials. Wombats and koalas are first known from the Miocene. Around 6.0 Ma, Australian environments became much more arid, and many animals like the koalas adapted to eating tougher vegetation.

Glossary

C4 photosynthetic pathway This photosynthetic pathway is used by many grass species. In most plants, called C3 plants, CO_2 is fixed in a chemical reaction using an enzyme called Rubisco. This enzyme binds CO_2 into a compound used in forming sugars during photosynthesis. Rubisco will also bind O_2, forming a compound that is toxic to the plant, a process known as photorespiration. When atmospheric oxygen levels were relatively low, the poor selectivity of Rubisco did not pose a problem for C3 plants, but as carbon dioxide levels declined relative to oxygen levels, photorespiration would have posed a challenge. C4 plants have evolved a mechanism where an additional reaction using a more selective enzyme binds CO_2 into another compound, which is

sequestered in the plant tissues. This compound can then be broken down and the CO_2 passed to Rubisco away from the presence of oxygen, so that little photorespiration can occur.

Chalicotheres An extinct group of perissodactyls, in the family Chalicotheriidae. These animals were very large, with clawed feet. Some had very long forelimbs, and were probably bipedal, feeding on leaves from tall trees. They are known from the latest Eocene up to the Pleistocene.

Creodont carnivore The creodonts are an extinct group of carnivorous mammals. It was long thought that they were ancestral to the modern order Carnivora, but more recent work indicates they may not be close relatives. They looked somewhat dog-like, with slicing carnassial teeth similar to those of Carnivora. However, in Carnivora the carnassial teeth are the last upper premolar and first lower molar, and in creodonts the carnassial teeth are the upper first and/or second molars and the lower second or third molar.

Rain shadow Dry region on the leeward side of a mountain or mountain range. The rain shadow effect is caused when moist air rises against the windward slope and is lost to precipitation on the windward side, leaving little moisture on the leeward side.

Ungulate Hoofed mammals, such as the ariodactyls (cows, pigs, sheep, deer, etc.) or perissodactyls (horses, rhinos). The last phalanx, or ungual phalanx, of each toe in ungulates is modified with a thick hoof instead of a claw.

Viverrids The members of the extant carnivoran family Viverridae, the civets and mongoose, known today in Africa and Asia. Viverrids are small, arboreal and generally fruit-eaters.

Waif dispersal An unusual form of inter-continental dispersal across large bodies of water by terrestrial animals. In an episode of waif dispersal, one or several animals inadvertently travels between two continents on vegetation rafts, and populates the new region on arrival. Waif dispersal may be responsible for the arrival of monkeys into South America.

See Also

Analytical Methods: Geochronological Techniques. **Andes**. **Atmosphere Evolution**. **Europe:** Mediterranean Tectonics. **Famous Geologists:** Suess. **Fossil Vertebrates:** Mesozoic Mammals; Placental Mammals; Hominids. **Indian Subcontinent**. **Stratigraphical Principles**. **Tertiary To Present:** Oligocene; Pliocene.

Further Reading

Berggren WA, Kent DV, Swisher CC III, and Aubry M P (1995) A revised Cenozoic geochronology and chron ostratigraphy. In: Berggren WA, Kent DV, Aubry M P, and Hardenbol J (eds.) *Geochronology, Time Scales, and Global Stratigraphic Correlation: Unified Temporal FrameWork for an Historical Geology*, pp. 129 212. Tulsa: Society for Sedimentary Geology. SEPM Special Publication 54.

Flower BJ and Kennett JP (1993) Middle Miocene ocean climate transition: high resolution oxygen and carbon isotopic records from Deep Sea Drilling Project site 588A, southwest Pacific. *Paleoceanography* 8: 811 843.

Jackson JBC, Budd AF, and Coates AG (eds.) (1996) *Evolution and Environment in Tropical America*. Chicago: University of Chicago Press.

Jacobs BF, Kingston JD, and Jacobs LL (1999) The Origin of Grass Dominated Ecosystems. *Annals of the Missouri Botanical Garden* 86: 590 643.

Janis CM, Scott KM, and Jacobs LL (eds.) (1998) *Evolution of Tertiary Mammals of North America: Volume 1 Terrestrial Carnivores, Ungulates, and Ungulatelike Mammals*. Cambridge: Cambridge University Press.

Miller KG, Wright JD, and Fairbanks RG (1991) Unlocking the icehouse: Oligocene Miocene oxygen isotopes, eu stasy, and margin erosion. *Journal of Geophysical Research* 96(B4): 6829 6848.

Pazzaglia FJ and Kelley SA (1998) Large scale geo morphology and fission track thermochronology in topographic exhumation reconstructions of the Southern Rocky Mountains. *Rocky Mountain Geology* 33(2): 229 257.

Raymo ME (1994) The Himalayas, organic carbon burial, and climate in the Miocene. *Paleogeography* 9: 352 404.

Rössner GE and Heissig K (eds.) (1999) *The Miocene Land Mammals of Europe*. Munich: Verlag Dr. Friedrich Pfeil.

Scotese CR, Gahagan LM, and Larson RL (1989) Plate tectonic reconstructions of the Cretaceous and Cenozoic ocean basins. *Tectonophys*. 155: 27 48.

Vickers Rich P, Monaghan JM, Baird RF, and Rich TH (eds.) (1991) *Vertebrate Paleontology of Australasia*. Melbourne: Pioneer Design Studio.

Woodburne MO (ed.) (1987) *Cenozoic Mammals of North America: Geochronology and Biostratigraphy*. Berkeley: University of California Press.

Zachos JC, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global cli mate 65 Ma to present. *Science* 292: 686 693.

Pliocene

C Soligo, The Natural History Museum, London, UK

Introduction

The Pliocene is the second and terminal epoch of the Neogene period. It marks the end of the Tertiary and precedes the Pleistocene, the first epoch of the Quaternary period. The term Pliocene (from the Greek *pleiõn* 'more' + *kainos* 'new') was first proposed in 1833 by Sir Charles Lyell (*see* **Famous Geologists:** Lyell) in his *Principles of Geology*. His *Newer Pliocene* was characterized by a molluscan fauna, of which 90% of species were then extant, contrasting with the *Older Pliocene,* from which only 33–50% of molluscan species had survived. Lyell subsequently renamed his *Newer Pliocene* the Pleistocene.

The Pliocene was a time of substantial climatic change and, most significantly, saw the onset of northern hemisphere glaciation, marking the start of the modern-day cycles of glacials and interglacials. In contrast, the Pliocene was also the last period that sustained global temperatures that were (over a prolonged period of time) higher on average than at present. Fauna, flora and geography were also similar to the modern day. As such the Pliocene epoch holds particular promise for the modelling of potential effects of global warming on the modern world.

The tectonic processes and climatic variability that characterized the Pliocene had a lasting impact on the distribution and evolution of plants and animals. The emergence of a land bridge between the North- and South-American continents resulted in extensive biotic movements and profoundly changed the floral and faunal characteristics of South America. In Africa, meanwhile, the Pliocene saw the diversification of our evolutionary lineage, the Homininae, and the emergence of our own genus, *Homo* (*see* **Fossil Vertebrates:** Hominids).

Definition

The Mediterranean region has traditionally been the focus of stratigraphic research aimed at defining the boundaries of the Pliocene Series and its stages. Sometime towards the end of the Miocene, the Mediterranean Sea became isolated form the Atlantic Ocean. As a result, most of the Mediterranean basin dried out, an event that has been termed the Messinian salinity crisis. The start of the Pliocene is defined to coincide with the time when oceanic waters burst through the area of today's straits of Gibraltar, refilling the Mediterranean basin in a sudden catastrophic flooding event. The abrupt transition from evaporites, precipitated from the hypersaline waters of the drying Mediterranean, back to marls, representative of deposition under normal open-marine conditions, marks the stratigraphic start of the Pliocene series. The series is thus unusual in that its base is not located within a continuous marine sedimentation record, but rather marks the sudden return of open-marine sedimentation following a period of non-marine sedimentation. Since the formal acceptance and ratification of the *Gelasian* in 1996, the Pliocene series now consists of three stages (Table 1).

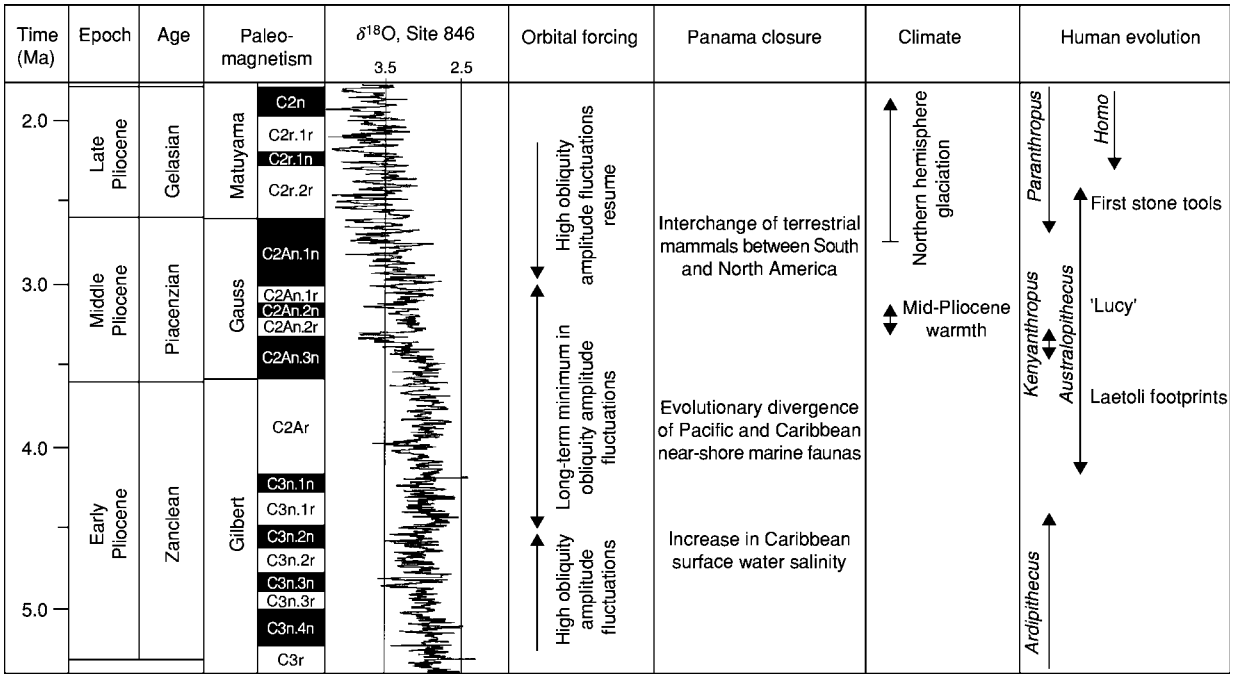
Zanclean (Lower Pliocene)

The Zanclean (from Zanclea, the classical name of Messina, Sicily) is the lowest stage of the Pliocene series. The Global Standard Stratotype section and Point (GSSP) for the Zanclean is located in the Eraclea Minoa section on the southern coast of Sicily (37°23′30″N; 13°16′50″E). The basal contact of this stage is where white Trubi Marl rests on dark brown Arenazzolo sands and marls. The base of the carbonate bed marking the small-scale Cycle 1 at the base of the stage corresponds to insolation Cycle 510 counted from the present, with an astrochronologic age of 5.33 Ma. Other tools for the global correlation of the base of the Zanclean stage include the base of the Thvera magnetic event (C3n.4n), dated to 5.236 Ma, coccoliths (the first occurrence of *Ceratolithus acutus* 5.37 Ma, the last occurrence of *Triquetrorhabdulus rugosus* 5.23 Ma and the last occurrence of *Discoaster quiqueramus* 5.537 Ma), and foraminifers (the first occurrences of *Globorotalia tumida* and *G. sphericomiozea* 5.6 Ma).

Piacenzian (Middle Pliocene)

The GSSP for the Piacenzian stage (after the town of Piacenza, Northern Italy) is located in the Punta Piccola section on the southern coast of Sicily (37°17′20″N; 13°29′36″E). The base of this stage is the base of the beige marl bed of the small-scale carbonate Cycle 77 and corresponds to the precessional excursion 347 counted from the present. Its astrochronologic age is estimated at 3.60 Ma. Further tools for the global correlation of the Piacenzian base include the Gilbert–Gauss magnetic reversal dated to 3.596 Ma and the location of the base within the obliquity-related $\delta^{18}O$ stage MG8 (O–176). Other correlation tools are geographically more restricted: within the Mediterranean region

Table 1 Composite table of Pliocene trends and events. $\delta^{18}O$, Site 846, from Shackleton *et al.*, 1995. *Proc ODP, Sci Results* 138: 337 355



the temporary disappearance of *Globorotalia puncticulata* (3.57 Ma), the first influx of *G. crassaformis*, the end of the paracme interval of *Discoaster pentaradiatus* (3.56 or 3.61 Ma) and the last occurrence of *Sphenolitus* spp. (3.73 or 3.70 Ma); in low- and mid-latitudes outside the Mediterranean, the last occurrences of *Globorotalia margaritae* (3.58 Ma) and of *Pulleniatina primalis* (3.65 Ma).

Gelasian (Upper Pliocene)

The Gelasian (after the town of Gela, Sicily) has only recently been ratified as the uppermost stage of the Pliocene series thereby establishing a threefold division of the Pliocene. The GSSP for the Gelasian stage is located in the Monte San Nicola section around 10 km N-NW of Gela in southern Sicily (Italy). The defined base of the Gelasian is the base of the marly layer overlying the Mediterranean Precession Related Sapropel (MPRS) 250 at 62 m in the Monte San Nicola section. The astrochronologic age of MPRS 250 (corresponding to precessional cycle 250 from the present) is 2.588 Ma. Other correlation tools include the Gauss–Matuyama boundary 20 Ky below the Gelasian base, nannofossils (last occurrences of *Discoaster pentaradiatus* and *D. surculus* 80 Ka above the base in most low- and mid-latitude areas) and foraminifers (last occurrence of *Globorotalia bononiensis* [*G. puncticulata* of some authors] 140 Ka above the boundary in the Mediterranean and North Atlantic). The last occurrence of *Stichocorys peregrina* (Radiolaria), the first occurrence of *Nitzschia joussaea* (Bacillariophycaceae/Diatomeae) in low-latitudes and the last occurrence of *Denticulopsis kamtschatica* (Bacillariophycaceae/Diatomeae) in the North-Pacific mid- and high-latitudes all approximate the Gauss–Matuyama boundary. The base of the Gelasian predates by 60 Ky the isotopic cold stage 100, which is marked by an increase in ice-rafted detritus in northern latitude oceanic sediments and the beginning of loess sedimentation in China. This base is also marked by changes in vegetation distribution patterns and continental vertebrate migrations (see below).

The end of the Pliocene series, of the Neogene and of the Tertiary period is determined by the Pliocene–Pleistocene boundary. The formal definition of that boundary has been the source of intense debate amongst stratigraphers. The currently validated Global Standard Stratotype Sections and Points (GSSP) for the base of the Pleistocene is located in the Vrica Section 4 km south of Crotone in southern Italy, just above the Olduvai subchron at an age of ~1.8 Ma.

Tectonics

The gradual formation of the isthmus of Panama(the land bridge connecting South and North America) cut off the Pacific Ocean from the Atlantic during the course of the Pliocene. The resulting rearrangement

of ocean currents profoundly influenced global climates. The isthmus formed as a result of a complex combination of lithospheric plate movements, which involved a north-western drift of the South American Plate and eastern drift of the Caribbean Plate. The gradual closure of the isthmus started during the Middle Miocene. By the Early Pliocene (4.6 Ma), it was advanced enough to affect deep-ocean circulation significantly. By 3.5 Ma the closure had progressed to the extent that evolutionary divergence had taken place between the Pacific and Caribbean shallow-water mollusc faunas. Evidence from nannofossils indicates that, between 3.65 and 2.76 Ma, the westward current between Caribbean and eastern Pacific ceased to exist while a northward intra-Caribbean current was established (Figure 1). However, the interchange of terrestrial mammals between the North and South American continents may have been delayed to 2.7 Ma, coinciding with the intensification of northern hemisphere glaciation and the associated sea-level drop. The substantial impact that the gradual shoaling of the Panama isthmus had on marine biota is exemplified by the observation that extinction rates of Pliocene molluscs in the western Atlantic and Gulf of Mexico were, on average, twice those in the eastern Pacific.

Other tectonic activities continued previous trends. In Africa-Arabia relative movements of the Nubian, Somalian and Arabian plates caused continued rifting along East Africa and initiated sea-floor spreading in the Red Sea. Many areas worldwide are thought to have undergone significant uplift throughout the Pliocene. Some of the most prominent areas being the South American Andes, the North American Sierra Nevada and Cascade Range, the African Atlas, the European Pyrenees and Alps, and the Himalayas and Tibetan Plateau in Asia.

The increased height of a large number of mountain ranges across the globe had a significant impact on local climates and may have influenced global climate. For example, the uplift of the Himalayas and Tibetan plateau, caused by the northward drift of the Indian plate and its crashing into the Asian plate, has been instrumental in shaping Asian climates at least since the Miocene. Continued uplift of northern and eastern parts of the Tibetan plateau have been linked to an intensification of both the summer and winter East Asian monsoons during the Middle Pliocene, 3.6–2.6 Ma. Increased volcanism is also likely to result from an increase in mountain building activity and a distinct increase in the frequency and thickness of tephra layers in ocean sediments from the Pliocene and the Quaternary support the notion of widespread orogenic activity during those periods. However, the absence of an obvious mechanism to underlie a global uplift of distinct

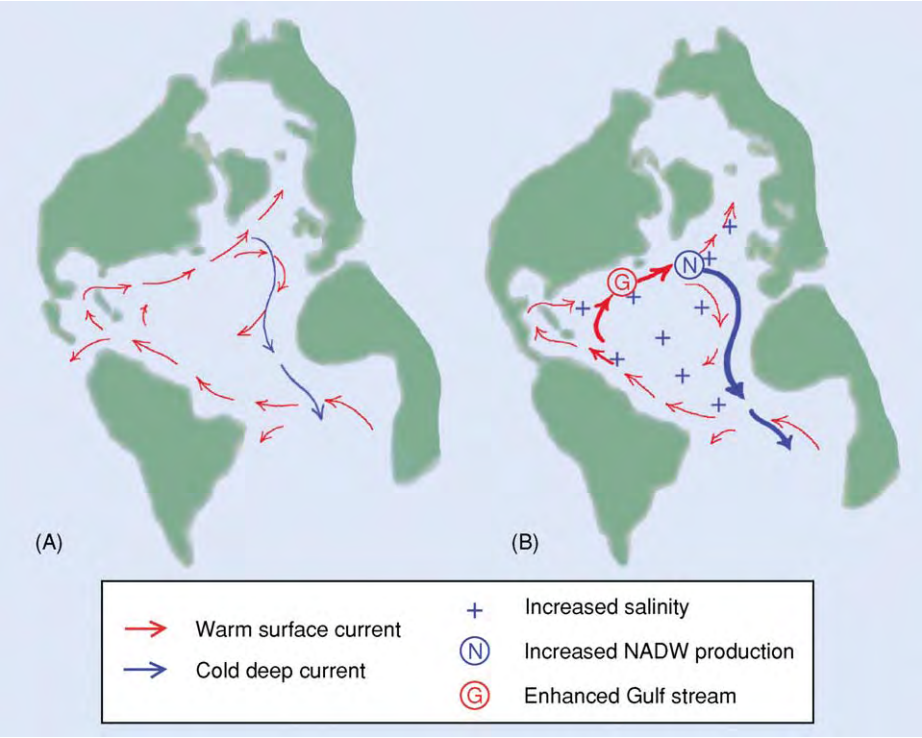


Figure 1 Atlantic marine currents before (A) and after (B) the emergence of the isthmus of Panama, showing intensified thermoha line circulation after the closure of the central American seaway. NADW: North Atlantic deep water.

mountain ranges of very varied ages has led some experts to question the existence of wide-ranging Plio-Quaternary uplift and to evoke the possibility that changes in climate have instead created the illusion of widespread uplift. Increased rates of erosion are widely cited as evidence for increased elevation gradients, but increased rates of erosion have also been associated with changes in global climate patterns, in particular with the enhancement of rainfall seasonality or the onset of rapid climatic oscillations in the later Pliocene.

Climate

The Pliocene was a time of profound climatic change. Starting with generally warm conditions up to the end of the Zanclean when global temperatures declined, resuming the trend of overall Cenozoic cooling. A subsequent phase of global warming, the 'mid-Pliocene warmth' (3.3–3.15 Ma) has been well documented and is a continued focus of intensive research due to its relevance to the prediction of future global warming events. The cooling that followed led to the northern hemisphere glacial-interglacial cycles, which continue to date. The onset of northern hemisphere glaciation was not synchronous. The first ice-sheets covered the Eurasian Arctic and north-east Asia by 2.75 Ma. Alaska became glaciated by 2.65 Ma, followed by the northeast American continent by 2.54 Ma. In terms of their potential to contribute to our understanding of the forces that drive and the consequences of global climatic change, the mid-Pliocene warm period and the onset of the modern ice age are of particular interest.

Mid-Pliocene Warming

With the exception of the last interglacial during which temperatures may have been briefly higher than today's, the period of mid-Pliocene warming (3.3–3.15 Ma) represents the most recent geological time span during which average global temperatures were higher than today's. Geography, flora and fauna of the mid-Pliocene were more similar to today's than during any other period of prolonged global warming. The mid-Pliocene, therefore, has the potential of providing the best indication of how the modern world may respond to future increases in global temperatures. Observational studies suggest that during the mid-Pliocene warmth sea levels were 20–40 m higher than today, the result of a reduced Antarctic ice-sheet area and the near complete absence of ice in the Arctic. Sea-surface temperatures were also substantially higher than today in the middle and higher latitudes. Warming was most pronounced in the north-eastern North Atlantic. Tropical sea-surface temperatures, however, were similar to those seen today.

Two basic factors have been suggested to explain the nature of Pliocene warmth: increased levels of atmospheric CO_2 (the greenhouse effect) and increased ocean heat transport. As increased atmospheric CO_2 is expected to increase temperatures globally and increased heat transport should result in a cooling of tropical waters, the reconstructed distribution of mid-Pliocene sea surface temperatures suggests that both mechanisms contributed to mid-Pliocene warming.

On land, there was a general shift of modern vegetation zones towards higher latitudes. Boreal forests occurred as far north as the Arctic coast, grading into temperate mixed-conifer and conifer-hardwood forest in the northern middle latitudes, and indicating the presence of temperatures that were much warmer than today in the Arctic regions of Iceland, Russia and North America as well as in western and central Europe. In contrast, some lower latitudinal areas may have seen temperatures drop below those of the modern day. The continental interiors of Africa, Asia and North America were more humid than they are today. Remains of pollen, leaves and wood of southern beech (*Nothofagus*) on Antarctica suggest a climate substantially warmer than at present. In the tropics, vegetation data point at temperatures lower than today in central America and East Africa, but warmer than at present in northern South America.

The Ice Age

The overwhelming climatic signature of the Pliocene is the increase in the levels of glaciation, and in particular the onset of large-scale glaciation of the northern hemisphere during the Middle Pliocene. At least three factors combined to initiate and sustain this glaciation. First, the overall trend of global cooling that prevailed through the Late Cenozoic ensured that precipitation over northern areas fell as snow rather than rain. Second, increased thermohaline circulation and Gulf Stream flow (a consequence of the closure of the isthmus of Panama) introduced the necessary moisture into far northern latitudes. Finally, a favourable Milankovitch orbital configuration between 3.1 and 2.5 Ma resulted in summers that were cool enough to prevent the snow from melting.

The onset of large-scale glaciation of the northern hemisphere is marked by an increase of $\delta^{18}O$ in benthic foraminifers between 3.0 Ma and 2.5 Ma and the appearance of substantial amounts of ice-rafted debris in the northern oceans from 2.7 Ma. The closure of the isthmus of Panama is likely to have

played a major role in the onset of northern hemisphere glaciation by redirecting surface currents within the Atlantic Ocean since 4.6 Ma and by influencing salinity levels. The resulting intensification of the Gulf Stream increased the transport of warm saline surface water to northern high latitudes and, consequently, the formation of North Atlantic deep water, which in turn, ensured increased import of warmer South Atlantic surface water into the northern hemisphere (**Figure 1**). Intensification of this marine 'conveyor belt' increased the potential for evaporation at high northern latitudes, thus supplying the additional moisture that would facilitate ice-sheet growth.

The gradual shoaling of the Panama isthmus and the associated changes in ocean currents and salinity gradients did not in themselves cause the onset of large-scale northern hemisphere glaciation. Initially, the enhanced Gulf Stream and increased transport of warm surface water into the high northern latitudes pushed global climates towards warmer conditions, culminating in the mid-Pliocene warm interval. The emergence of the Panama isthmus contributed significantly to creating the necessary preconditions for the initiation of large-scale northern hemisphere glaciation. However, other factors may have been involved (e.g. deepening of the Bering Straits, increased river discharge of fresh water into the Arctic Ocean from the Palaearctic). Continued Pliocene uplift or rejuvenation of various mountain ranges and plateaus across the world has also been suggested as a cause for the deterioration of Late Neogene climates. Invoked mechanisms include the influence of high elevation areas on atmospheric circulation as well as the reduction of atmospheric CO_2 concentrations through the weathering of newly exposed silicates and the resulting formation of carbonates, the increased burial of organic carbon due to increased rates of erosion, or increased ocean productivity through an increased delivery of phosphorus to the ocean, again as the result of increased erosion.

It is likely that several of these factors combined to create the preconditions necessary for the onset of northern hemisphere glaciation. The ultimate trigger, however, lay with the nature of fluctuations in the Earth's obliquity (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). Low tilt angles of the Earth's rotational axis result in cold northern hemisphere summers. Just prior to the onset of northern hemisphere glaciation, 4.5–3.1 Ma, there was a prolonged period during which the amplitude of Earth's obliquity fluctuated only minimally. Records of $\delta^{18}O$ indicate that during this time, limited glaciations of the Arctic were initiated several times, but failed to substantiate. From around 3.0 Ma the gradual increase of the amplitude of obliquity fluctuations set the final stage for the current pattern of pronounced glacial and inter-glacial periods, which have prevailed since the end of the Middle Pliocene.

Biotic Events

The Great American Interchange

The emergence of the isthmus of Panama in the Middle Pliocene resulted in the first contact of South America with another continent since the opening of the Drake Passage finalized its separation from West Antarctica toward the end of the Eocene. Having been rifted from Gondwanaland ~100.0 Ma and spent most of the Cenozoic in total isolation, South America was inhabited by a distinct biota with a large proportion of endemic taxa. In contrast, the North American continent had previously had a long history of intermittent connections and resulting floral and faunal interchange with the Old World.

The emergence of a permanent land connection between these two sub-continents resulted in a substantial level of terrestrial biotic interchange. However, not all taxa proved equally successful at dispersing and the biotas of the two sub-continents were not equally affected. With respect to long-term effects, the influence that southern taxa have had on North American biotas is near negligible. In contrast, the long-term effect that the great American interchange has had on the South American mammal fauna is substantial. Today nearly half of the families and genera of South American mammals are members of groups that emigrated from North America since the emergence of the Panama land bridge. This does not appear to be due to a substantially larger number of species initially dispersing from north to south as opposed to from south to north. In fact, when differences in source area are taken into account, the extent of north to south dispersal was similar to that from south to north. The difference in the proportions of modern representatives arose through the fact that North American immigrants to the south underwent substantial diversification in their new environments, whereas South American immigrants in the north did not. The reasons for the substantial differences in success rates between North American and South American immigrants are not clear.

Movements across the Panamanian isthmus were not restricted to mammals, but data for other groups are scarcer. For birds the available data suggest a net movement of taxa from north to south, whereas the majority of dispersals of reptiles and amphibians appear to have been from south to

north with significant levels of dispersal already occurring prior to the final emergence of the isthmus. In addition, while more than 90% of the angiosperm species of lowland rainforests in central America are estimated to be of South American origins, most montane plant species in South America have North American roots.

The Trans-Arctic Interchange

The Bering Strait between Siberia and Alaska initially opened during or just before the Early Pliocene. This new connection between the Arctic-Atlantic and the North-Pacific basins resulted in biotic movements referred to as the 'marine trans-Arctic interchange'. Data from invertebrates and algae suggest that, overall, and in line with the present prevailing currents, the trans-Atlantic interchange heavily favoured invasion from the Pacific to the Atlantic. However, data from before 4.8 Ma reveal the presence of Atlantic bivalves in the Pacific, whereas after 3.6 Ma North-Pacific molluscs suddenly became widespread in the North Atlantic. This confirms that, in accordance with existing models, currents through the Bering Straits flowed from north to south during the Early Pliocene, but were reversed some time between 4.8 and 3.6 Ma, probably as a consequence of the gradual closure of the central American seaway.

Other Biotic Movements

The emergence of the Panama land bridge and the opening of the Bering Strait resulted in a general intensification of marine currents. From a biotic perspective these also resulted in interchange between northern and southern temperate biota in the eastern Pacific and between temperate biota of the eastern and western Atlantic.

Plants

The Pliocene's climatic trends were reflected in the nature and distribution of plants worldwide. During periods of cooling, many areas saw a transition from broadleaved to coniferous woods or from closed to more open, grassy vegetations. Climate change evidently influenced the patterns of distribution of plants, but changes in the nature of the vegetation may in turn have influenced Pliocene climates. Around the Miocene to Pliocene transition, the biomass of plants using C_4 as opposed to C_3 photosynthetic pathways increased in Africa, Asia, North and South America, probably as a result of decreasing concentrations of atmospheric CO_2. The increase took place during the Late Miocene in the lower latitudes and during the Early Pliocene in the higher latitudes. However, some local environments remained C_3-dominated and no increase in the biomass of C_4-plants appears to have taken place in western Europe. The continued spread of C_4-plants during the Pliocene may have had a direct effect on global climates. Due to their shallower roots and their increased efficiency at CO_2 fixation, C_4-plants return less water from the soil to the atmosphere than C_3-plants. They thereby affect the hydrologic cycle and generally promote drier conditions downwind.

Hominin Diversification

In Africa, the Pliocene epoch saw the evolutionary diversification of the human lineage, the Hominidae (*see* **Fossil Vertebrates:** Hominids). Only three species of hominin are known from the Late Miocene. In contrast, up to 12 additional species (almost all from the eastern rift of the African rift valley and from South African cave sites) have been

Table 2 Hominin diversity during the Pliocene

Species	*Age*	*Localities*
Ardipithecus ramidus White *et al.*, 1994	5.8 4.4 Ma	Middle Awash (Ethiopia)
Australopithecus anamensis Leakey *et al.*, 1995	4.2 3.9 Ma	Kanapoi, Allia Bay (Kenya)
Australopithecus afarensis Johanson *et al.*, 1978	3.6 2.9 Ma	Laetoli (Tanzania); Koobi Fora, West Turkana (Kenya); Omo, Middle Awash, Hadar (Ethiopia)
Australopithecus bahrelghazali Brunet *et al.*, 1996	3.5 3.0 Ma	Bahr El Ghazal (Chad)
Kenyanthropus platyops Leakey *et al.*, 2001	3.5 3.3 Ma	Lomekwi (Kenya)
Australopithecus africanus Dart, 1925	3.0 2.4 Ma	Taung, Makapanskat, Sterkfontein (South Africa)
Australopithecus garhi Asfaw *et al.*, 1999	2.5 Ma	Bouri (Ethiopia)
Paranthropus aethiopicus Arambourg & Coppens, 1968	2.7 2.3 Ma	West Turkana (Kenya); Omo (Ethiopia)
Paranthropus boisei Leakey, 1995	2.3 1.4 Ma	Olduvai, Peninj (Tanzania); Chesowanja, Koobi Fora, West Turkana (Kenya); Omo, Konso Gardula (Ethiopia)
Paranthropus robustus Broom, 1938	1.9 1.4 Ma	Kromdraai, Swartkrans, Drimolen, Gondolin (South Africa)
Homo habilis Leakey *et al.*, 1964	2.3 1.6 Ma	Omo, Hadar (Ethiopia); Olduvai (Tanzania); East Lake Turkana (Kenya); Sterkfontein (South Africa)
Homo rudolfensis Alexeev, 1986	1.9 Ma	East Lake Turkana (Kenya)

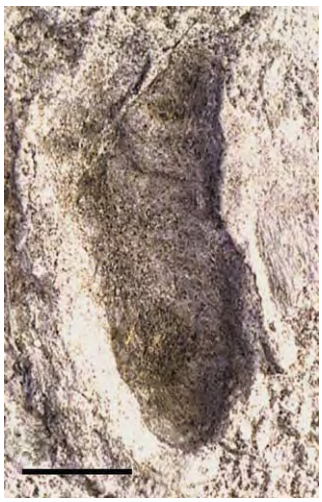


Figure 2 One of a series of footprints left behind in volcanic ash at the Early hominin site of Laetoli (Tanzania) some 3.6 Ma. Scale is 5 cm. Photograph by Peter Schmid.

described from the Pliocene to date (Table 2). Around 4.0 to 3.5 Ma a shin bone, probably of *Australopithecus anamensis*, and a series of fossilized footprints (Figure 2) provide the first clear evidence of bipedal walking. The most famous hominin fossil ('Lucy', an *Australopithecus afarensis)* lived just over 3.0 Ma and the oldest stone artifacts are known from the Middle to Late Pliocene transition, ~2.5 Ma. The Pliocene diversification of early hominins coincided with climatic fluctuations and a general trend towards more arid and less wooded habitats in eastern and southern Africa. The traditional view, however, that typical hominin features such as bipedality evolved as a direct response to the increased aridity and opening up of the East-African landscape is no longer considered valid. Recent palaeoecological reconstructions of early hominin environments generally point at wooded and well-watered habitats. Only with the advent of the genus *Homo*, around 2.5 Ma, can hominins be considered to have been fully adapted for life in open and arid environments. This is confirmed by data from marine sediments. Marine records show that the African climate saw an increase in aridity after 2.8 Ma, coinciding with the onset of large-scale glaciation in the northern hemisphere.

Before 2.8 Ma, variations in the intensity of the African monsoon were mainly the result of variations in low-latitude insolation due to Earth orbital precession and caused 19 Ka–23 Ka cyclical alternations between dry and wet conditions. After 2.8 Ma the African climate became sensitive to the increased amplitude of high northern latitude climate variations and the cycles changed to 41 Ka, paralleling the periodicity of the Earth's orbital obliquity variation and the Late Pliocene cycles of northern hemisphere glacials and interglacials. From 2.8 Ma, periodically cooler and drier conditions and the transition towards overall lower and more seasonal precipitation strongly favoured open savannah vegetation in east Africa. It is clear that the appearance and spread of open savannah cannot be considered to have contributed to shape the earliest evolution and diversification of the hominin lineage. Instead, it is more likely, that the gradual reduction of areas of humid and wooded habitat towards the Middle to Late Pliocene transition contributed to the disappearance of many taxa. In contrast, the emergence of the genus *Homo* coincided with substantial aridification of much of the African continent.

See Also

Atmosphere Evolution. **Carbon Cycle**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Famous Geologists:** Lyell. **Fossil Vertebrates:** Hominids. **Tectonics:** Mountain Building and Orogeny; Rift Valleys. **Tertiary To Present:** Miocene; Pleistocene and The Ice Age.

Further Reading

Castradori D, Rio D, Hilgen FJ, and Lourens LJ (1998) The Global Standard Stratotype section and Point (GSSP) of the Piacenzian Stage (Middle Pliocene). *Episodes* 21: 88 93.

Dowsett HJ, Barron JA, Poore RZ, Thompson RS, Cronin TM, Ishman SE, and Willard DA (1999) *Middle Pliocene Paleoenvironmental Reconstruction: PRISM2*. U.S. Geo logical Survey Open File Report, pp. 99 535, http//pubs.usgs.gov.

Driscoll NW and Haug GH (1998) A short circuit in ther mohaline circulation: a cause for northern hemisphere glaciation? *Science* 282: 436 438.

Haug GH and Tiedemann R (1998) Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermoha line circulation. *Nature* 393: 673 676.

Hay WH, Soeding E, DeConto RM, and Wold CN (2002) The Late Cenozoic uplift climate change paradox. *International Journal of Earth Sciences* 91: 746 774.

Jackson JBC, Budd AF, and Coates AG (eds.) (1996) *Evolu tion and Environment in Tropical America*. Chicago: University of Chicago Press.

Poore RZ and Sloan LC (eds.) (1996) *Climates and Climate Variability of the Pliocene*. Marine Micropa leontology 27.

Reed KE (1997) Early hominid evolution and ecological change through the African Plio Pleistocene. *Journal of Human Evolution* 32: 289 322.

Rio D, Sprovieri , Castradori D, and Di Stefano E (1998) The Gelasian Stage (Upper Pliocene): a new unit of the global standard chronostratigraphic scale. *Episodes* 21: 82 87.
Van Couvering JA, Castradori D, Cita MB, Hilgen FJ, and Rio D (2000) The base of the Zanclean Stage and of the Pliocene Series. *Episodes* 23: 179 187.
Vrba ES, Denton GH, Partridge TC, and Burckle LH (eds.) (1995) *Paleoclimate and Evolution with Emphasis on Human Origins*. New Haven: Yale University Press.
Wrenn JH, Suc J P, and Leroy SAG (eds.) (1999) *The Plio cene: Time of Change*. Dallas: American Association of Stratigraphic Palynologists Foundation.

Pleistocene and The Ice Age

A Currant, The Natural History Museum, London, UK

Introduction

Pleistocene is the name given to the geological epoch succeeding the Pliocene and preceding the Holocene. It is usually taken as beginning about 1.8 million years ago and ending with the termination of the last major cold stage and the beginning of the Holocene warm stage, about 10 000 years ago, but the start of the Pleistocene has proved difficult to define. Together, the Pleistocene and the Holocene form the Quaternary Period. The Pleistocene is characterized by great global climatic instability, particularly from the onset of major, cyclic, northern hemisphere, terrestrial ice accumulations about 700 000 years ago. This later period is popularly known as the Ice Age.

The term 'Pleistocene' was first used by Charles Lyell in 1839 to denote the age of sediments in and around the London and Paris basins, in which 70% of the molluscan fossils were considered to represent species that are still alive today. The deposits in question were clearly related to the ancient courses of the Thames and the Somme. Lyell's original definition was of limited value. It was very much at the mercy of changes in opinion on molluscan identification and taxonomy and had extremely restricted geographical utility, but the name remained and has been the subject of almost continual redefinition since Lyell's time. The current global stratigraphic section and point (GSSP) for the start of the Pleistocene at Virca in Italy still has unresolved dating problems. The start of the epoch is widely believed to be between 1.65 and 1.8 million years ago, although some authors push this back as far as 2.5 million years. The environmental distinction between the Late Pliocene and the Early Pleistocene is actually quite difficult to pin down, and there was considerable faunal and floral continuity between the two.

Historical Studies

In global terms, the Pleistocene was dominated by the changing extent of ice volume. Although sea-floor spreading and continental drift continued throughout the period, the effect of these longer term processes has been dwarfed by ice-related phenomena. The history of our changing and growing understanding of the Pleistocene is important to anyone examining the older literature related to the period, particularly that which deals with the 'Ice Age'. The pioneer geologist and geomorphologist Louis Agassiz published works on his studies of Alpine glaciers and glacial processes, *Étude sur les Glaciers* (1840) and *Système Glaciare* (1847), in which he put forward the idea of there having been a great Ice Age in which large parts of Europe had been covered by ice and the surface of the land was greatly modified under its influence. Agassiz then went on to make similar observations based on his travels in North America. The subsequent history of Ice Age studies has been one of very gradual recognition that there were probably multiple Ice Age events, widely referred to as 'glacials', punctuated by warmer phases, with temperatures as warm as or warmer than those now prevailing, known as 'interglacials'. Numerous models of Pleistocene climatic events have been put forward, but the one that held sway for longest and was most widely adopted was that proposed in 1909 by Penck and Bruckner in *Die Alpen im Eiszeitalter*, in which the authors suggested there had been four major Ice Age stages, the Gunz, Mindel, Riss, and Wurm. This system, based on the terrace deposits of Alpine rivers, became adopted all over the world, and although local names often replaced the Alpine originals, the fourfold division of Pleistocene ice ages was accepted almost everywhere. The intervening interglacial stages became named after the cold stages that delimited them, i.e., Gunz-Mindel, Mindel-Riss, and Riss-Wurm.

It is important to recognize that the early understanding of Pleistocene events was based on hybrid data. Evidence for the cold glacial stages was largely

derived from characteristic sediments and landforms whereas evidence for the warmer interglacial stages came mainly from organic sediments and the fossils they contained. The stratigraphic relationship between glacial and interglacial sequences was often obscure and the means of correlating between isolated terrestrial sequences was poorly defined. The Alpine model of Ice Age events became to a very large extent a self-perpetuating system. Although a number of lines of evidence began to emerge suggesting much greater overall complexity, the lack of clarity in the original stage definitions and the virtual impossibility to establish secure interregional correlations led to such evidence being subsumed into the model. Against the background of purely geological interpretations, there was also a growing body of archaeological evidence charting human prehistory through the Pleistocene in various parts of the world. It was not until the 1860s that it became widely accepted that the human race had any great antiquity, with stone tools that were clearly made by early people being found in undisturbed contexts along with the remains of extinct animals. Such a likelihood was first noted by the French antiquarian Jacques Boucher de Perthes, who from 1846 to 1857 published his findings of stone tools from the high-level (and therefore very ancient) terrace deposits of the River Somme near Abbeville. There was considerable opposition to the idea that there were human populations of great age represented in the fossil record, primarily from religious groups and individuals, and to some extent this opposition still survives today. For most people, the clinching discovery was that of prehistoric cave art in France and Spain, where people living in the last glaciation had painted pictures of contemporary woolly mammoths, woolly rhinos, reindeer, bison, and horses in the caves they also used for shelter – direct evidence with a ancient human signature. It quickly became clear that the record of human antiquity was longest in Africa, Europe, and southern Asia and relatively short in northern Asia, Australia, and the Americas.

There is probably no other period of Earth history that has been so intensively studied yet so badly misinterpreted than the Pleistocene, and all because the construction of theoretical models has tended to run ahead of the collection and interpretation of hard evidence. Far too little attention has been paid to testing models and far too much to reinforcing them. The Alpine Model of Pleistocene climatic events eventually fell victim to new evidence from the oceans. Cesare Emiliani and colleagues, working in the late 1950s and early 1960s on cores taken from deep-ocean sediments, found evidence from the shells of buried foraminifera that their oxygen isotope content varied through time. Emiliani had already worked on the palaeoclimate signals recovered from microfossils in older deposits, and he recognized his new data as being a direct record of changing global ice volume. The two significant isotopes of oxygen, ^{16}O and ^{18}O, behave differently during evaporation, leaving the oceans enriched with the heavier isotope. When water evaporates from the oceans and falls as rain or snow on the land, it is isotopically 'light', and if sufficient frozen water remains on land in large, stable ice-sheets, then the isotopic composition of the oceans will change to a measurable degree. When the ice melts, eventually the meltwater will ultimately return to the oceans, thus recentring the isotopic signal. Foraminifera living in the sea will deposit shells that reflect the chemistry of the seawater during their lifetime, and when they die, their shells fall to the seafloor and become incorporated in the sediments being deposited there, creating a retrievable record of changing global ice volume through time. Emiliani found evidence for many more episodes of global ice accumulation than were allowed for in the Alpine Model.

When plotted onto a graph, the marine isotope record bore a very strong resemblance to the theoretical curve of variations in the predicted amount of solar energy reaching a particular point on the planetary surface; this is determined by known variations in Earth's orbit. This curve had been calculated by the Serbian astrophysicist Milutin Milankovitch in the early twentieth century to explain long-term climatic variability, but, in the absence of any evidence to back his theory, Milankovitch's work remained largely ignored. Emiliani's recognition of the resemblance between the Milankovitch solar radiation curve and the oceanic oxygen isotope record has led to a completely new understanding of the driving forces behind global climatic change, now extending far back beyond the Pleistocene. Alongside plate tectonics, this must number as one of the greatest breakthroughs in geological science.

Milankovitch theory examines the combined effects of three known variables in Earth's orbit: the eccentricity of Earth's orbit around the Sun, the obliquity of Earth's axis relative to the Sun, and the precession of the equinoxes that changes the season at which Earth's axis is most tilted towards the Sun. Each of these has respective cycles of approximately 100 000, 41 000 and 23 000 years. In the past 700 000 years, the 100 000-year cycle has become dominant, giving major global cold stages at about this interval. Prior to this, there was considerable climatic instability during the earlier part of the Pleistocene, but not on anything like this dramatic scale. Amazingly detailed climatic signals for the last major climatic cycle

recovered by drilling through the Greenland ice-cap confirm and amplify the resolution of the marine isotope record for this later period.

It is now recognized that there have been about seven major global Ice Age events in the past 700 000 years, though these vary to some degree in intensity and progress. Armed with this information, it is now possible to re-examine the geological record and reinterpret the evidence for the relative order and extent of ice-sheet advance and for the intervening warmer interglacial stages. The modern terminology of Pleistocene subdivision is based on the marine isotopic signal. The isotope stages are numbered backwards from 1, with Marine Isotope Stage (MIS) 1 representing the Holocene. Warm stages have odd numbers and cold stages have even numbers. A historical anomaly in the last cold cycle has been divided up into three isotope stages (2, 3, and 4), so the last interglacial period is MIS 5, but otherwise the stages appear to represent single major climatic events (see **Figure 1**).

Life on Earth in the Pleistocene

The distribution of plants and animals on Earth's surface has been hugely influenced by large-scale high- to mid-latitude ice accumulation. Not only have ice-sheets periodically excluded life from large areas of land, but the effects of climate change have radically changed the normal distribution of many species, while significant areas of new land have become available for colonization at times of lowered sea-level. Interglacial sea-levels have generally been as high as those of today, or sometimes a few metres higher, but during cold stages, the amount of water locked up on land as ice has reduced global sea-levels by up to 120 m, exposing vast areas that are now continental shelf. Such drops in sea-level would unite the islands of south-east Asia into a single landmass, greatly extend the eastern seaboard of southern South America, make the British Isles a peninsula of north-west Europe, and connect Asia to North America by a broad northern land bridge (Beringia). In the southern hemisphere, the scope for significant increase in terrestrial ice cover is limited by the disposition of the continents. Antarctic ice masses grew and waned, but it was only in the Andean region of South America that very large, new, non-polar ice-sheets developed. The primary effect of Pleistocene ice accumulation is seen in the northern hemisphere, where the Hudson Bay area and Scandinavia became the central foci for the growth of major new ice-caps and large mountain glaciers developed in the Rockies and in the Alpine mountain belt.

The influence of Pleistocene cold-stage events is still very much with us today. Large regions of permanently frozen ground or permafrost in the arctic regions are attributable to the deep cold experienced by both glaciated and unglaciated higher latitudes, and the crustal distortion caused by the massive weight of former ice-sheets can be witnessed in a recovery phase around the Baltic Sea and Hudson Bay regions, where isostatic uplift is still very active. The Baltic Sea and Hudson Bay are no more than the slowly draining crustal depressions caused by former ice-sheets.

Rich sources of evidence for the faunal and floral history of the Pleistocene are to be found in the deposits laid down by major river systems. It is probably fortunate that the study of such deposits began in north-western Europe, where long-term regional uplift has created staircase-like terrace systems representing the former levels at which rivers ran as they progressively cut down to new base levels. The Pleistocene deposits of south-east England are an extremely good illustration of this kind of sequence, particularly those related to the Thames drainage. A fossil landscape is preserved beneath eastern England, buried beneath the debris of ice-sheets that invaded the area about 400 000 years ago. This glacial event (the Anglian glaciation, believed to represent Marine Isotope Stage 12) obliterated some of the major river systems of central England and diverted the Thames to something like its present course. Beneath the glacial tills, fossiliferous deposits representing these extinct river systems provide information on life in the area in the early Pleistocene and early middle Pleistocene.

From the Anglian glaciation onwards, the Lower Thames Valley preserves fragments of river deposits representing all the major climatic stages revealed by the oceanic isotope record. Because the terrace fragments can be related to the existing topography, it has been possible to build up a clear picture of the post-Anglian evolution of this particular landscape. The fact that the modern city of London is sited on the Thames plays no small part in unravelling the story. Almost every pocket of sand, gravel, and brickearth has at some stage in the recent past been excavated as an economic resource, and many of these sections have been carefully recorded and their fossils collected. It is a sad fact that the economic development that created the opportunity to understand the later Pleistocene history of the Thames has now destroyed much of the evidence on which that understanding was based, and it is one of the responsibilities of present-day geographers and Pleistocene specialists to make certain that what little remains is well documented and, where possible, conserved. This particular case history underlines the point that much of the evidence for terrestrial

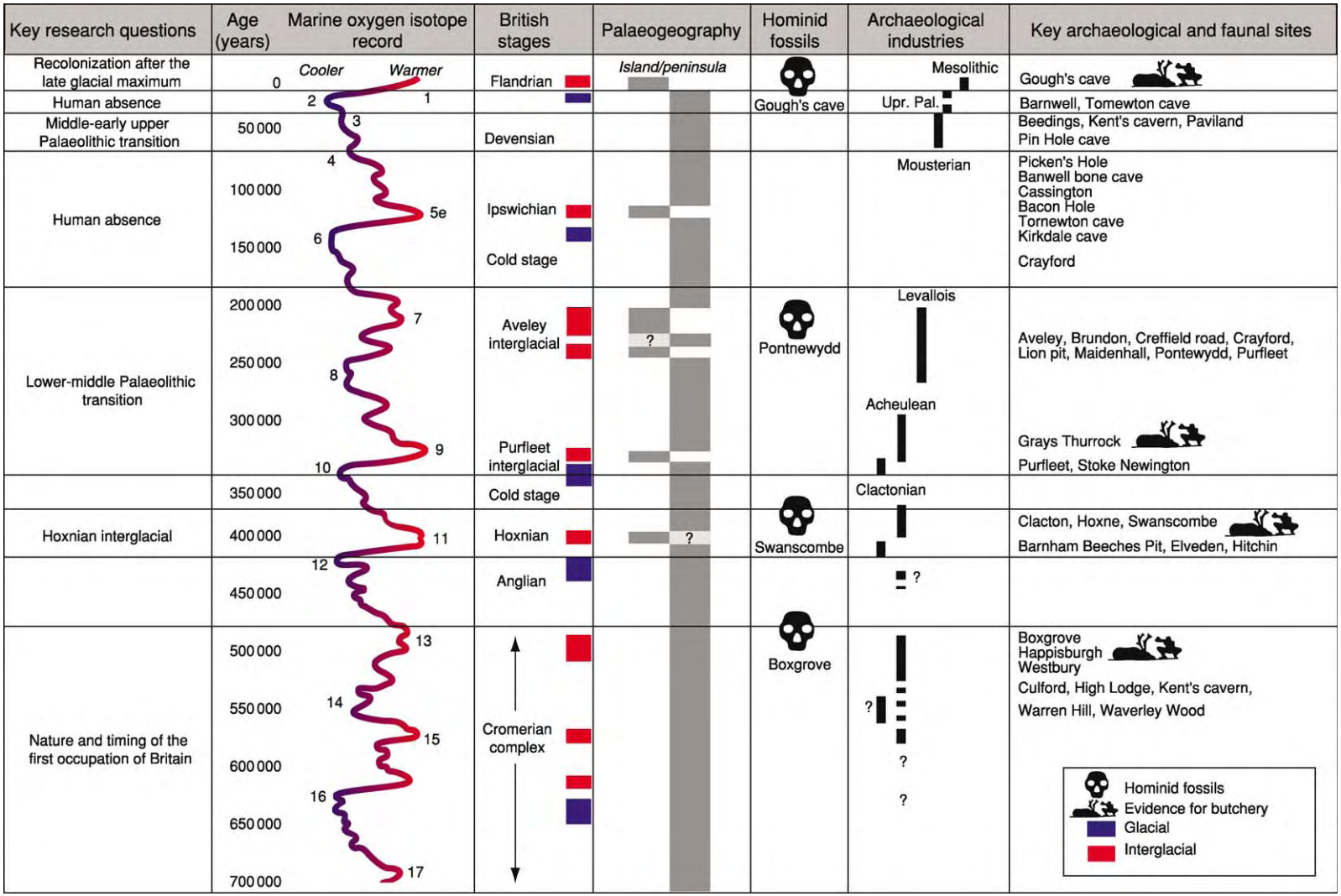


Figure 1 The relationship between apparent human occupation of the British Isles during the Pleistocene and climate change, based on the work of the Ancient Human Occupation of Britain research team directed by Professor C B Stringer FRS.

Pleistocene events lies buried within existing landscapes and can only usually be seen when it is artificially excavated.

In the Thames Valley, cold-stage deposits are usually poor in fossils, but the interglacial stages are represented by exceptionally rich organic levels containing bones, molluscs, beetles, and plants. Five distinct post-Anglian interglacials can be recognized, representing Marine Isotope Stages 11, 9, 7, 5, and 1 (stage 3 is not an interglacial, but is nonetheless well represented by cold-stage deposits laid down in the middle of the last glacial stage). Most of the known river terrace deposits have been found to have a cold–warm–cold sandwichlike structure, with an organic interglacial filling of varied lithology underlain and overlain by cold-stage gravels. The relationship and order of these terrace deposits can be worked out from their topographic position, but this interpretation has been greatly enhanced by study of their fossil content. Biostratigraphic data have been recovered from all of the major groups examined, but it has been found that the vertebrate faunas, particularly the mammals, give the most distinctive signature to each interglacial event. It has proved possible to use the distinctive mammalian faunas recovered from the Thames terrace sequence to correlate with other river and lake sequences around the country and even with fossiliferous cave deposits. The importance of cave sequences cannot be overstated, because they currently offer the only reliable means of direct dating (uranium series age determinations on clean stalagmite) that is available for Pleistocene sequences beyond the range of radiocarbon.

Many other river systems around the world are now being studied systematically to understand their distinctive depositional cycles and the ways in which these reflect Pleistocene climatic variation. Just as ice-sheets were restricted to particular areas of Earth's surface, so drainage systems have their own dynamics, according to the regions they drain. It is essential to understand those dynamics before trying to interpret the palaeoclimatic evidence that they may contain, in particular their fossil content. Returning briefly to the Thames, the great paucity of fossils relating to cold periods before the last one is a rather odd feature, particularly given the richness of interglacial sequences within the same region, but this is the signature of this particular river system. In other parts of the world, different parts of the climatic cycle may be better represented, and it is certainly the case that other European drainage systems have quite different depositional histories.

Historically, caves have been a great focus of interest. Caves are natural dustbins, anomalous sites for the temporary deposition of detritus that would otherwise be destroyed by natural processes (Figure 2). They are a natural focus for biological activity and wide range of animals use them for shelter and to raise their young. This includes amphibians, bats, a variety of mammalian and avian carnivores, and humans. Bone tends to be preserved preferentially in limestone environments, and most caves are to be found in limestones. Many caves have been found in which animal bone is one of the primary constituents of the cavern infilling. One of the most significant features of caves is that they are one of the few places in the terrestrial environment where there is at least the possibility of preserving quite long sequences of deposits representing multiple climatic cycles. Long stratigraphic sequences are comparatively rare on land in the Pleistocene.

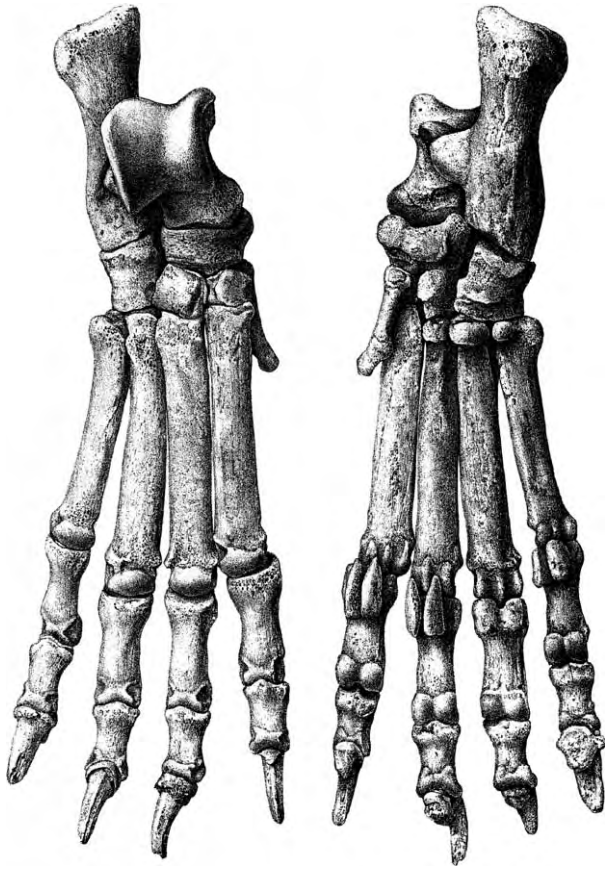


Figure 2 Engraving of the hind foot bones of a Pleistocene spotted hyena, *Crocuta crocuta*, from Tornewton Cave, Devon. Caves often contain exceptionally well preserved fossilized mammalian bones, particularly those of the carnivores that have used them as lairs or dens.

One of the great biological phenomena of the Pleistocene is the terminal Pleistocene extinction event, and a huge body of literature has been devoted

to its likely causes. Somewhere around the very end of the Pleistocene and extending in numerous cases into the early Holocene, a large number of large to medium-sized mammals died out in a relatively short time. It is informative to look at the anatomy of this extinction event. The areas in which the extinctions took place were Europe, northern Asia, Australia, and the Americas. Generally speaking, the faunas of southern Asia and Africa survived intact. There had previously been numerous rapid transitions from late glacial to early temperate environments, but no other event during the Pleistocene had seen global-scale extinctions. Some have suggested that environmental change alone caused mass extinction, but the evidence on the ground does not support this case very well.

Southern Asia and Africa both seem to have a long and continuous record of human activity, extending well back into the Pleistocene, and in Africa earlier still. The mammals of these areas are likely to have had long-standing adaptive responses to the presence of dangerous and effective bipedal predators. In Europe, the co-existence of people and native mammals was markedly sporadic, particularly during the later Pleistocene, whereas in much of northern Asia and in Australia and the Americas co-existence was either very limited or perhaps even completely absent. When anatomically modern humans began to spread across Europe, Asia, and beyond, the animals living in these regions would probably never have encountered people before and had little or no time to adapt to their presence. If this natural naivety of the animal populations was coupled with the great environmental stress of rapid environmental change, undoubtedly causing a natural crash in population numbers, then this is more likely to account for the widespread but highly regionalized pattern of extinction.

One particular case history is illuminating. By the time modern human populations had spread along the newly formed Arctic Ocean shoreline at the end of the Pleistocene, sea-levels were already beginning to rise, and a population of mammoths had become isolated beyond human reach on what is now Wrangel Island. Though mammoth populations on the European and Asian mainland and throughout North America dwindled to extinction by the end of the Pleistocene, the isolated animals on Wrangel Island survived to as recently as 4500 years ago, in one of the most inhospitable environments on Earth. It was only when humans eventually reached the island that the mammoths finally vanished. So when the great megalithic monument at Stonehenge was under construction on the grasslands of southern England, there were still woolly mammoths alive on an island in the Arctic Ocean.

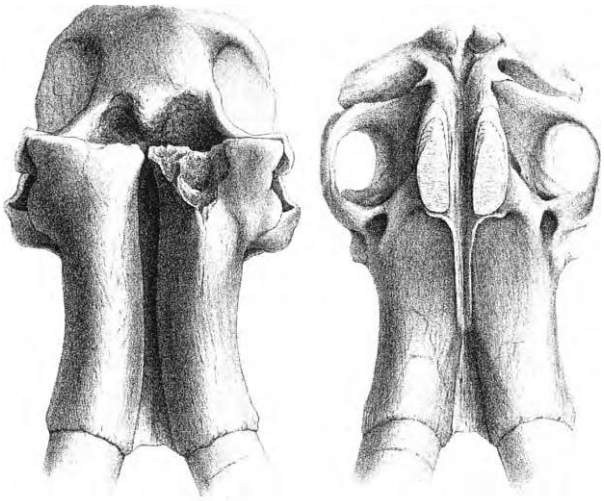


Figure 3 Engraving of the skull of a Pleistocene mammoth, *Mammuthus* cf. *primigenius*, from Ilford, Essex. Megaherbivores such as the mammoth were significant creators and modifiers throughout the Pleistocene and their activities helped maintain the preferred habitats of many other plants and animals.

The effects of mammalian extinctions in the Americas and in Australia were particularly severe, with the loss of all of the megaherbivores (animals weighting a tonne or more; Figure 3) as well as a large number of medium-sized species. Mammoths, mastodonts, giant sloths, glyptodons, native horses, notoungulates, litopterns, and many other forms were completely wiped out in the Americas, and Australia lost all of its larger marsupials. Large animals with long regeneration times and slow-growing, vulnerable young seem to have been pushed to extinction by the rapid spread of modern humans to new parts of the globe. The timing of both events coincides perfectly and it seems strange not to accept a direct link between the two phenomena. The significance of the loss of the Pleistocene megaherbivores cannot be overstated, for these animals were the environment creators of the higher latitudes. The vast cold, dry grasslands of Eurasia and North America were created and maintained by creatures such as mammoths throughout the Pleistocene. Once these animals became extinct, the environment that they had helped to create disappeared with them, destroying the habitat that had supported many smaller species, and the northern coniferous woodlands took over. In the longer term, human populations have effectively replaced the larger Pleistocene herbivorous mammals in living primarily on cultivated grasses such as wheat, maize, rice, and millet. In

summary, the Pleistocene was a period of large-scale, cyclic environmental change. Plants and animals showed considerable adaptive responses to these changes, but in the case of the environmentally important megaherbivores, their progress was cut short, probably by the rapid spread of modern humans.

Further Reading

Delcouty HR and Delcourt HA (1991) *Quaternary Ecology*. London: Chapman & Hall.

Lowe JJ and Walker MJC (1984) *Reconstructing Quaternary Environments*. England: Longman Scientific and Technical.

THERMAL METAMORPHISM

R Abart and R Milke, University of Basel, Basel, Switzerland

Introduction

In the course of geological processes, rocks may be subject to changing pressure-temperature and chemical conditions. This may induce mineral transformations and re-crystallization. If the mineralogical and (micro)structural changes take place, when the rock is in the solid state, this is referred to as metamorphism. Metamorphic transformations, which are primarily driven by elevated temperature at constant (low) pressure, may be referred to as thermal metamorphism. In this chapter, we summarize the settings in which thermal metamorphism may occur. We then describe characteristic mineralogical patterns of thermal metamorphism. We address the timing of thermal metamorphism and its implications on the extent of equilibration and the development of microstructures and reaction textures. Finally, we address fluid flow and associated chemical transport during thermal metamorphism.

Geological Settings of Thermal Metamorphism

Both pressure and temperature increase with depth. Whereas isobars, i.e., loci of equal pressure, constitute approximately horizontal surfaces, the thermal structure of the Earth's crust may be rather complex (Figure 1). Heat is constantly liberated by radioactive decay and by slow cooling and solidification of the Earth's interior. Heat is transported to the relatively cool surface primarily through thermal conduction. In the absence of mass flow associated with tectonic and magmatic activity and with fluid circulation, a time invariant relation between depth and temperature would be established in the subsurface, which may be referred to as a stable geothermal gradient. Conduction of heat is a relatively slow process, and the geothermal gradient may be modified, if heat is transported passively with moving matter. There are several processes which may cause deviations from a stable geothermal gradient towards high temperatures.

Processes, which may cause thermal perturbations on a regional scale

Large fragments of the lithosphere may be displaced by tectonic or gravity mechanisms. If the movement includes a vertical component, the uplifted lithosphere transports heat to shallower depths, thereby raising and condensing the isotherms in the crust. An efficient mechanism to transport thermal energy is magmatic underplating. This is the formation of extended mafic intrusions at deep-crustal or subcrustal level. Seismic signals indicate that underplating melt bodies may have thicknesses of several kilometres. Such large magma bodies increase the heat flow to the Earth's surface on a regional scale. Due to the lower density of the underplating magma compared to the lithospheric mantle, it may also give rise to surface uplift. The elevation of the highest cenozoic mountain ranges cannot be explained by isostatic movement of the post-collisional thickened lithosphere alone. They require a combination of isostatic uplift, magmatic underplating, and thinning of the lithospheric mantle from its root. Areas of increased heat flow are also found in zones of extensional geotectonic regime. Thinning of the lithosphere may occur by horizontal stretching or movement on normal faults. At the base of the lithosphere, mantle material rises to compensate for the lost volume. Since the rate of tectonic transport is generally faster than the rate of heat transfer, each point of the pre-extensional geotherm moves vertically to establish the new geotherm.

All regional scale processes that lead to thermal perturbations, increased heat flow, and raised geotherms involve vertical displacement of the crust

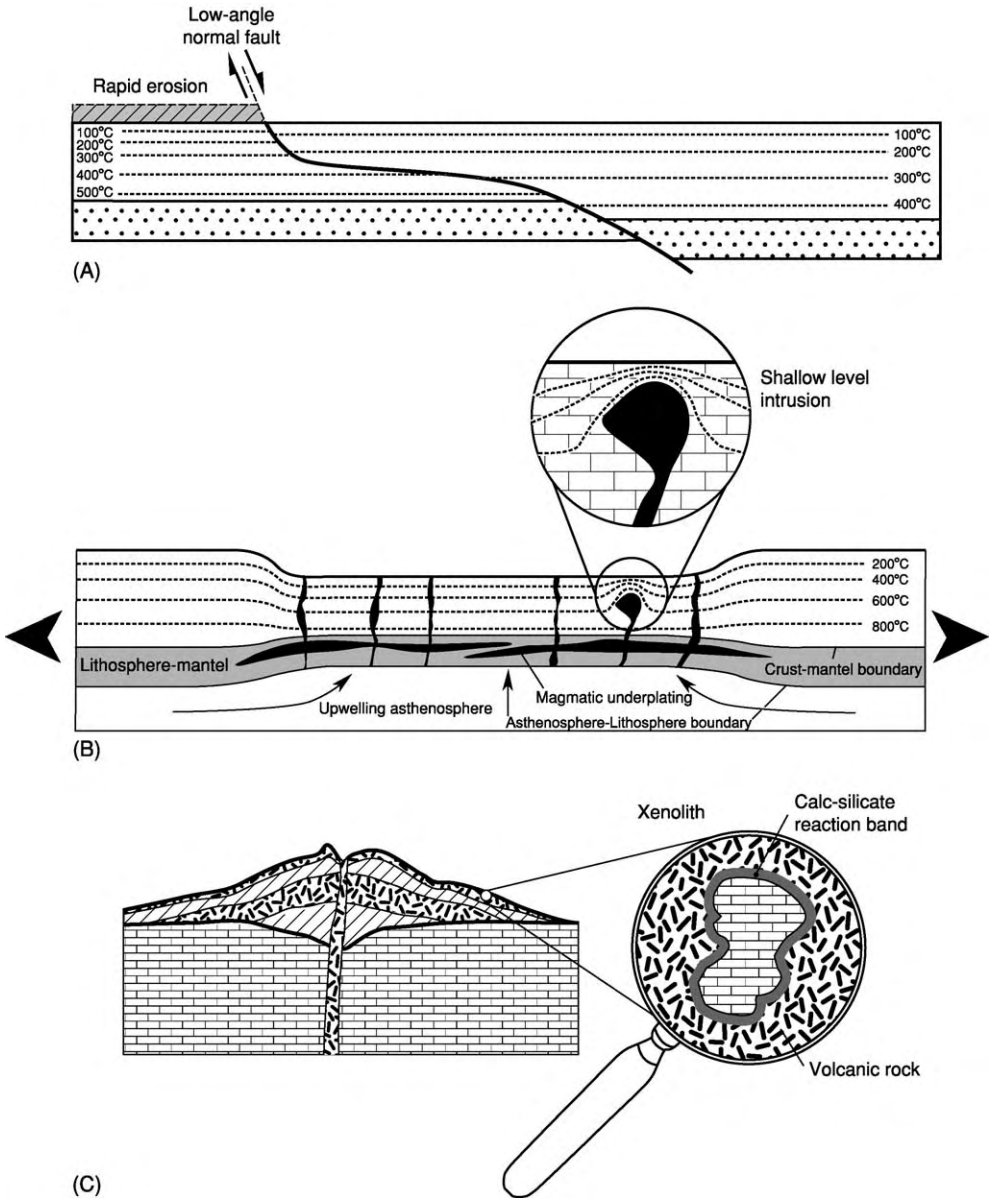


Figure 1 Settings of thermal metamorphism: (A) condensation of isotherms during tectonic exhumation of crustal fragments along low angle normal fault; (B) condensation of isotherms in the course of lithospheric thinning in an extensional regime; also shown is the effect of magmatic underplating, which is often associated with lithospheric thinning; insert shows the effect on the thermal structure of a shallow level intrusion; (C) thermal overprint of xenoliths in a volcanic setting.

and thus induce both changes in temperature and pressure conditions. Thermal metamorphism in a narrow sense, that is at constant pressure, is restricted to processes that operate on a local scale.

Processes, which may cause thermal perturbations on a local scale

Magmatic intrusions Large amounts of heat may be transported into shallow levels of the crust via the intrusion of magmas. Heat is liberated during solidification (latent heat of crystallization) and during cooling of the magma intrusion and it is transferred into the country rocks via thermal conduction and possibly via fluid circulation. The domain of thermal perturbation surrounding a shallow intrusion is referred to as the thermal aureole or the contact aureole and the associated metamorphic transformations are referred to as contact metamorphism. Within a thermal aureole peak metamorphic temperatures increase exponentially with decreasing distance from the intrusive contact, from background temperatures at the outer limit to maximum temperatures at the intrusive

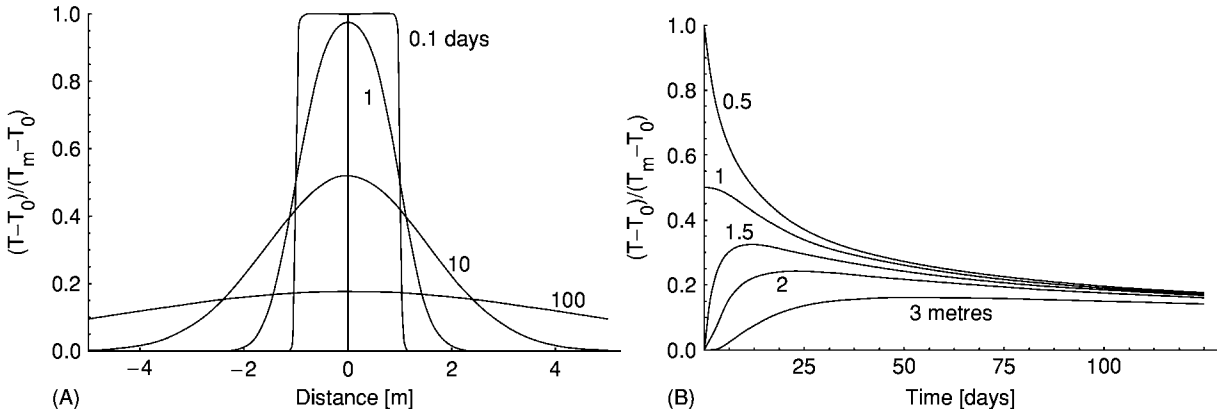


Figure 2 Temperature distribution around a dyke intrusion; for construction of this figure it was assumed that the magma intruded at a temperature T_m and the country rock was at a uniform temperature T_0 prior to magma intrusion; the magma and the country rock were assigned similar thermal diffusivities of 10^{-6} m^2/s; (A) shows the temperature distribution across the dike for different times after intrusion; (B) shows the evolution of temperature with time at different distances from the dyke centre.

contact. This gives rise to characteristic sequences of metamorphic parageneses (see below). Peak metamorphic temperatures at the intrusive contact, among others, depend on the temperature of the intruding magma, the temperature of the country rocks prior to magma intrusion, and on the rate and extent of heat production during crystallisation. The temperature pattern produced by the intrusion of a magmatic dike and its evolution with time are illustrated in Figure 2. It is important to note that the thermal evolution of a rock depends on its distance from the dike. Whereas the dyke material is subject to strictly monotonic cooling after intrusion the country rocks go through an initial heating stage before they start cooling. Peak metamorphic temperatures and heating rates are high close to the intrusive contact and relatively low further out in the country rock. Penetrative deformation associated with magma intrusion is usually confined to the immediate contact region and metamorphic crystallization is usually static in large portions of the contact aureole.

Volcanic activity Effusion of lavas in the course of volcanic activity is an obvious setting of thermal metamorphism. On the one hand, lavas may substantially heat the material over which they flow and cause sharp thermal aureoles of limited extent. On the other hand, on their way to the surface, magmas may incorporate fragments of country rocks, which then are quickly heated to magmatic temperatures. Many of the calc-silicate minerals have first been described from limestone derived xenoliths, which were metamorphosed in lavas of the Monte Somma volcano in southern Italy.

Thermal metamorphism on a regional scale is associated with the production of basaltic lavas at mid-ocean ridges. Thermal metamorphism in the mid-ocean-ridge environment is usually accompanied by intense hydrothermal activity, which may lead to pronounced chemical alteration of the original mid-ocean-ridge basalts.

Natural coal seam and hydrocarbon seepage burns, the production of fulgurites from lightning strikes, the formation of pseudotachylites from frictional heating, and the transformation of kinetic into thermal energy associated with meteoritic impacts may be considered as short-term and very localised special cases of thermal metamorphism.

Mineral Zones in Thermal Metamorphism

Mineral zones are defined by the systematic appearence of new minerals in a series of rocks with identical whole rock composition. These minerals are referred to as index minerals. Mineral zones can be mapped in the field.

Isograds are lines on a map connecting points of equal metamorphic grade. The practical application of this definition to rocks is problematic since there are sets of P-T-X conditions that satisfy reaction equilibrium for a given mineral assemblage. In practice, it is therefore difficult to exactly pin down the metamorphic grade and it is more practical to use the concept of reaction isograds to describe metamorphic rock series. Reaction isograds are lines joining points that are characterized by the equilibrium assemblage of a given reaction.

In thermal metamorphism, temperature is the leading variable controlling metamorphic grade. Contact aureoles are characterised by spatial patterns of reaction isograds that parallel the intrusive contact

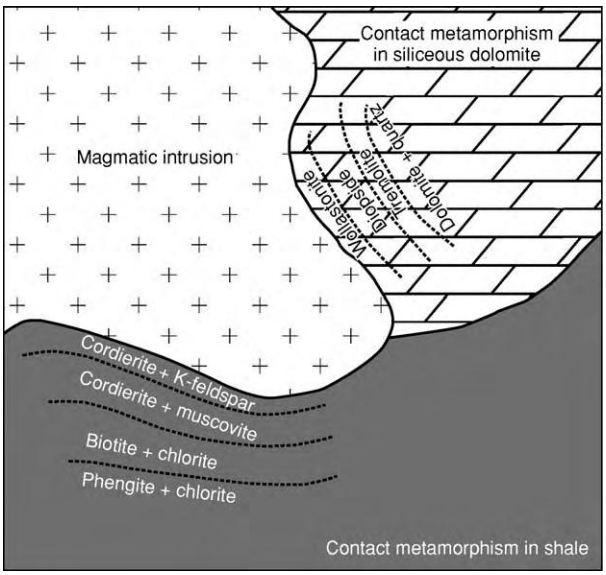


Figure 3 Schematic map of a shallow magmatic intrusion into country rocks of sedimentary origin. For a given lithology, different mineral zones characterized by index minerals evolve depending on the distance to the heat source.

and result in mineral zones (Figure 3). Contact metamorphism is most pronounced if the intruded rocks were previously unmetamorphosed or had only been subjected to low-grade regional metamorphism. The respective succession of mineral zones depends on the lithology of the metamorphosed rocks and comprises different index minerals for contact metamorphism of shales, impure carbonates, ultramafic rocks, or other country rocks. For the reconstruction of the thermal evolution in a contact aureole, simple rock compositions with clearly defined reactions are most convenient, e.g., siliceous dolomites (Figure 3).

H_2O and CO_2 are the prime constituents of metamorphic fluids. These species may be liberated or sequestred in the course of mixed-volatile mineral reactions. As a consequence, the composition of the pore fluid is influenced by mineral reactions and has an important influence on the stability of mineral parageneses.

The kinetics of mineral reactions may be slow compared to the time-scales of thermal perturbations. High heating rates in contact aureoles may lead to significant overstepping of mineral equilibria, such that the formation of mineral zones lags behind the thermal evolution.

Fluid-Rock Interaction

Thermal metamorphism, in particular contact metamorphism, is often accompanied by fluid migration. On the one hand, magmas may dissolve significant amounts of volatiles and fluid may be liberated during crystallization. On the other hand, thermal perturbations may cause bouyancy driven fluid circulation. Fluid flow tends to be pervasive during early stages of contact metamorphism, and it may become focused along joints and fractures at later stages. The fluid in the pore space of rocks is the most efficient transport medium in the solid crust. Fluid may transport dissolved species and isotopes. The chemical and isotopic signature of fluids is imprinted on the solid phases of the rocks via fluid rock interaction. If metamorphic transformations are accompanied by changes of the bulk rock chemical composition, this process may be referred to as allochemical metamorphism or metasomatism. Mineralogical and (stable)isotope alteration patterns are robust and enduring manifestations of palaeo fluid flow.

See Also

Igneous Processes. **Igneous Rocks:** Granite. **Metamorphic Rocks:** Classification, Nomenclature and Formation; Facies and Zones. **Mining Geology:** Magmatic Ores.

Further Reading

Bucher K and Frey M (1994) *Petrogenesis of metamorphic rocks*. Berlin: Springer.

Jamtveit B and Yardley BWD (1997) *Fluid Flow and Trans port in Rocks*. London: Chapman and Hall.

Kerrick DM (1991) *Contact Metamorphism. Reviews in Mineralogy*. Vol. 26. Mineralogical Society of America.

Kornprobst J (2002) *Metamorphic rocks and their geody namic significance*. Kluwer Academic Publishers.

Kretz R (1994) *Metamorphic Crystallization*. John Wiley & Sons.

Turcotte DL and Schubert G (1992) *Geodynamics*. John Wiley & Sons Inc.

Winter JD (2001) *An introduction to igneous and meta morphic petrology*. Prentice Hall.

Yardley B (1989) *An introduction to metamorphic pet rology*. Longman Earth Sciences Series.

TIME SCALE

F M Gradstein, University of Oslo, Oslo, Norway
J G Ogg, Purdue University, West Lafayette, IN, USA

Introduction

Boundary stratotypes of stages, high-resolution radioisotopic dating, earth-orbit tuning of cyclic sequences, advances in biostratigraphic scaling of stages, and detailed error analysis are keys to the standard geological time-scale. Construction and assembly of the geological time-scale follow several well-defined steps, including (1) construction of an updated global chronostratigraphic scale for Earth's rock record, (2) identification of key linear-age calibration levels for the chronostratigraphic scale using high-resolution radioisotopic (or other source of) absolute age dates, (3) application of earth-orbit tuning to intervals with cyclic sediments or stable isotope sequences that have sufficient biostratigraphic or magnetostratigraphic ties, (4) interpolation of the combined chronostratigraphic and chronometric scale when direct information is insufficient, and (5) calculation or estimation of error statistics on the combined chronostratigraphic and chronometric information to obtain a geological time-scale with estimates of uncertainty on boundaries and unit durations. The International Commission on Stratigraphy (ICS) cosponsors the standard globally applicable geological time-scale.

Human Time

Time is an indispensable tool for all of us. The time kept by innumerable watches and clocks regulates our everyday life, and the familiar calendar governs our weekly, monthly, and yearly doings. These sequences eventually condense into the historical record of events over centuries. The standard unit of modern timekeeping is the second, defined as the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the caesium-133 atom. This value was established to agree as closely as possible with the ephemeris second based on Earth's motion. The advantage of having the atomic second as the unit of time in the International System of Units is the relative ease, in theory, for anyone to build and calibrate an atomic clock with a precision of 1 part per 10^{11} (or better). In practise, clocks are calibrated against broadcast time signals, with frequency oscillations in hertz being the 'pendulum' of the atomic timekeeping device.

The tick of the second paces the quick heartbeat, and traditionally was the 60th part of the 60th part of the 24th part of the 24-h day, with the minute and the hour being convenient multiples to organize daily life. The day carries the record of light and dark, the month is marked by the regularly returning shapes of the moon, and the year is represented by the cycle of the seasons and the apparent path of the sun. All of these time passages are clearly understood, and humans have long recognized the notion that time is a vector, pointing from the present to the future. Events along its path mark the 'arrow of time', and the arrow is graded either in relative 'natural' units, or in units of duration – the standard second and its multiples, such as hours and years.

Geological Time and the Rock Record

What is often less clear is the concept of geological time, the bases of its units, and how to use these units properly. A good understanding of geological time is vital for every earth scientist, especially those who strive to understand geological processes and determine rates of change. This understanding takes place in a framework called 'Earth geological history', a kind of supercalendar of local and global events. The challenge to this understanding is reading, organizing, and sorting Earth's stone calendar pages, and, as best as we can, reconstructing the content of any missing pages. Stratigraphic correlation is a vital part of this event-reconstruction process.

One of the earliest reconstructions was made by Nicolas Steno (*see* **Famous Geologists:** Steno) (1631–1687), who made careful and original stratigraphic observations. Based on these observations, Steno concluded that Earth's strata contain the superimposed records of a chronological sequence of events that can be correlated worldwide. Geological correlation formally is expressed in terms of five consecutive operations and units:

- Rock units, such as formations or well log intervals (lithostratigraphic correlation; e.g., the Kimmeridge Clay Formation of England).
- Fossil units, such as zones (biostratigraphic correlation; e.g., the *Turrilina alsatica* benthic foraminifer range zone).
- Relative time units (geochronological ('Earth time') correlations; e.g., Jurassic Period, Eocene Epoch, Oxfordian Age, Magnetic Polarity Chron C29R).

- Rocks deposited during these time units (chronostratigraphic (time–rock) correlation; e.g., Jurassic System, Eocene Series, Oxfordian Stage, Magnetic Polarity Zone C29R).
- Linear time units or ages (geochronological correlation; e.g., 150 Ma, 10 ka).

Without correlation to a global reference scale, successions of strata (i.e., events in time derived in one area) are unique and contribute nothing to an understanding of Earth history elsewhere. The rules of hierarchy in geological correlation—from rocks and fossils to relative and linear time—are carefully laid down in the *International Stratigraphic Guide*. An abbreviated copy of this 'rule book' with further references may be found on the website of the International Commission on Stratigraphy (http://www.stratigraphy.org).

Before dealing further with linear geological time, consider the common geological calendar built from relative age units. This chronostratigraphic scheme is not unlike a historical calendar in which societal periods (e.g., the Minoan Period, the reign of Louis XIV, or the American Civil War) are used as building blocks, devoid of a linear scale. Archaeological relicts deposited during these intervals (the Palace of Minos on Crete, Versailles, or spent cannon balls at Gettysburg, respectively) comprise the associated physical chronostratigraphic record. A chronostratigraphic scale is assembled from rock sequences stacked and segmented in relative units based on their unique fossil and physical content. When unique local fossil and physical records are matched with those of other rock sequences across the globe – in a process known as stratigraphic correlation – a relative scale can be assembled; when calibrated to stage-type sections, this scale becomes a chronostratigraphic scale.

The standard chronostratigraphic scale, in downloadable graphics format, is available from the ICS website. This time-scale is made up of up of successive stages in the rock record (e.g., Cenomanian, Turonian, and Coniacian within the Cretaceous System). Originally, each stage unit was a well-defined body of rocks at a specific location of an assigned and agreed upon relative age span, younger than typical rocks of the underlying stage (Jurassic) and older than the typical rocks of the next higher stage (Paleogene). This is the concept of defining stage units with stage-type sections, commonly referred to as stratotype sections. The principles and building blocks of this chronostratigraphy were slowly established during centuries of study in many discontinuous and incomplete outcrop sections. Inevitably, lateral changes in lithology between regions and lack of agreement on criteria (particularly, which fossils were characteristic of which relative unit of rock), have always resulted in considerable confusion and disagreement with respect to stage nomenclature and use. Almost invariably, classical stage stratotypes turned out to represent only parts of stages. Hence, a suite of global subdivisions with precise correlation horizons was required.

Global Stratotype Section and Point

Now, relatively rapid progress is being made with definition of Global Stratotype Sections and Points (GSSPs) to fix the lower boundary of all geological stages, using discrete fossil and physical events that correlate well in the rock record. For the ladder of chronostratigraphy, this GSSP concept switches the emphasis from marking the spaces between steps (stage stratotypes) to fixing the rungs (stage boundaries).

Each progressive pair of GSSPs in the rock record also precisely defines the associated subdivision of geological time. Hence, philosophical arguments are being heard to cut out time–rock units, and deal only with rock, fossil, and time units *sensu strictu*. Essentially, the argument is that dual systems of precisely defined subdivisions of geological time and of parallel similarly defined subdivisions of the time–rock record are redundant. Or, even more radical, why not replace the hundreds of melodious, but confusing, 'ian' subdivisions of stages (Gelasian, Sinemurian, Spathian, Emsian, etc.) with simple 'real' ages – who needs the 'Victorian Age' when there is a 'nineteenth century'?

It is now 25 years ago that a 'golden spike' struck the first GSSP. This event (of historic proportions for development of the geological time-scale) involved the boundary between the Silurian and Devonian periods, or rather the lower limit of the Devonian, at a locality called Plonk in Czechoslovakia. The problem of the Silurian–Devonian boundary and its consensus settlement in the Klonk section hinged on a century-old debate, known as the 'Hercynian Question', that touched many outstanding geoscientists of the previous century. The issue came to the foreground after 1877, when Kaiser stated that the youngest stages (étages) of Barrande's 'Silurian System' in Bohemia correspond to the Devonian System in the Harz Mountains of Germany and other regions. Kaiser's findings contrasted with the conventional nineteenth-century wisdom that graptolites became extinct at the end of the Silurian. Eventually, it became clear that so-called Silurian graptolites in some sections occur together with so-called Devonian fossils in other sections, leading to the modern consensus that graptolites are not limited to Silurian strata.

A bronze plaque in the Plonk outcrop shows the exact position of the modern Silurian–Devonian Boundary, which is taken at the base of the Lochkovian Stage, the lowest stage of the Devonian. The base of the Lochkovian Stage is defined by the first occurrence of the Devonian graptolite *Monograptus uniformis* in Bed #20 of the Klonk Section, northeast of the village of Suchomasty. The Lower Lochkovian index trilobites with representatives of the *Warburgella rugulosa* group occur in the next younger limestone Bed #21 of that section.

The concept of the GSSP has gained acceptance among those stratigraphers who consider it a pragmatic and practical solution to the common problem that conventional stage-type sections inevitably leave gaps, or lead to overlap between successive stages. The boundary stratotype very much relies on the notion that it is possible to arrive at accuracy in correlation through the use of events (e.g., a geomagnetic reversal, a global change in a stable isotope value, or the evolutionary appearance of one or more prominent and widespread fossil taxa). Thus, the limits of a stage can now be defined with multiple event criteria that, using the best of current knowledge, are synchronous over the world. Delimiting successive stages in a clear and practical manner enhances their value as standard units in chronostratigraphy and ultimately in geochronology. Without standardised units, neither the (relative) stratigraphic scale nor the (absolute) time-scale can exist.

This is not to say that the classical concept of the stage stratotype has suddenly become obsolete, and should be abolished. Although the GSSP concept is not ruled by priority regulation, going back through the historical notions of how and where a stage was originally conceived, defined, and correlated sheds valuable light on the geological meaning and correlative content of stages and the historical notion of their boundaries. Nevertheless, it is clear that, at a time when scientific viewpoints are becoming increasingly more global, stage stratotypes have more regional rather than global significance. At present, nearly 50 GSSPs have been defined (**Figures 1** and **2**). Over 50 more Phanerozoic stages are in need of base definition, and ICS has set the year 2008 as the completion date for all remaining GSSPs.

Due to the fact that most of the Proterozoic lacks adequate fossils for correlation, a different type of boundary stratotype, the Global Standard Stratigraphic Age (GSSA), is in use for this interval of Earth history. The definition of a boundary by its linear age is the consequence of the fact that the Proterozoic now recognizes units of global stratigraphic subdivision, with the boundaries being defined in terms of the abstract age in millions of years. Summaries of ratified GSSAs may be found in **Figure 2**. Although there appears to be consensus that the subdivision of the Proterozoic in three eras (Palaeoproterozoic, Mesoproterozoic, and Neoproterozoic) is excellent, the finer subdivisions often contain no dated rocks, which makes their use haphazard. An intensive search is going on for physical events in the Proterozoic rock record suitable to qualify as 'golden spikes'.

It is desirable to add some documentation to a GSSP proposal about the historical significance of a new-stage GSSP in the hierarchy of stratigraphic units of higher rank. Careful consideration needs to be given to the fact that, for example, the GSSP for the base of the Triassic Period in Meishan, China, automatically defines the boundary of the Permian–Triassic, and the base of the Induan Stage. Because the Triassic is considered the lowest period in the Mesozoic, the golden spike in the Meishan section also separates the Palaeozoic from Mesozoic. In the same vein, the base of the Cambrian Period as defined in a GSSP in the Fortune Head section in south-west Newfoundland also defines the base of the Nemakitian-Daldynian Stage, often considered to be the lowest stage in the Cambrian. Because the Cambrian, by consensus, is considered to be the lowest period in the Palaeozoic Era and the lowest major unit in the Phanerozoic Aeon, the Fortune Head GSSP also defines the lower boundary of these major chronostratigraphic units. This leads to correlation in linear time units, called geochronological correlation, with reference to the geochronological calendar of Earth events. Whereas the chronostratigraphic scale is a convention to be agreed upon rather than discovered, calibration of the scale in seconds and (mega-) years is a matter for discovery and estimation rather than agreement. Like human time, linear geological time is expressed in units of standard duration – the second, and hence (thousands or millions of) years.

Building a Geological Time-Scale

The ideal time-scale is built from accurate radioisotopic ages, taken precisely at stage boundaries throughout the stratigraphic column in the Phanerozoic. There are two basic measuring (dating) tools to build the linear geological time-scale: (1) stratigraphically meaningfull readiometric dates in millions of years and (2) Earth-orbit-tuned sedimentary cycles in thousands of years. Both types of measurements are achieved in rocks that are stratigraphically discontinuous. Also required is a certain amount of luck in finding core or outcrop sections suitable for dating with index fossils, geomagnetic reversals, stable isotope anomalies, etc.

Global Boundary Stratotype Sections and Points (GSSPs)

Status in Oct 2003; see ICS website (http://www.stratigraphy.org) for updates.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
PHANEROZOIC							
Cenozoic Era							
Neogene System				*"Quaternary"* is traditionally considered to be the interval of oscillating climatic extremes (glacia and interglacial episodes) that was initiated at about 2.5 Ma, therefore encompasses the Holocene, Pleistocene and uppermost Pliocene. It is not a formal chronostratigraphic unit.			
Holocene Series							
Base Holocene	11.5 ka	0.00	Carbon-14 dating calibration	exactly 10,000 Carbon-14 years (= 11.5 ka calendar years BP) at the end of the Younger Dryas cold spell		Informal working definition	
Pleistocene Series							
Base Upper Pleistocene subseries	0.126	0.00	Astronomical cycles in sediments	base of the Eemian interglacial stage (= base of marine isotope stage 5e) before final glacial episode of Pleistocene	Potentially, within sediment core under the Netherlands (Eemian type area)	Informal working definition	
Base Middle Pleistocene subseries	0.781	0.00	Astronomical cycles in sediments	Brunhes-Matuyama magnetic reversal		Informal working definition	
Base Pleistocene Series	1.806	0.00	Astronomical cycles in sediments	Just above top of magnetic polarity chronozone C2n (Olduvai) and the extinction level of calcareous nannofossil *Discoaster brouweri* (base Zone CN13). Above are lowest occurrence of calcareous nannofossil medium *Gephyrocapsa* spp. and extinction level of planktonic foraminifer *Globigerinoides extremus.*	Top of sapropel layer 'e', Vrica section, Calabria, Italy	Ratified 1985	Episodes 8 (2), p. 116–120, 1985
Pliocene Series							
Base Gelasian Stage	2.588	0.00	Astronomical cycles in sediments	Isotopic stage 103, base of magnetic polarity chronozone C2r (Matuyama). Above are extinction levels of calcareous nannofossil *Discoaster pentaradiatus* and *D. surculus* (base Zone CN12c).	Midpoint of sapropelic Nicola Bed ("A5"), Monte San Nicola, Gela, Sicily, Italy	Ratified 1996	Episodes 21 (2), p. 82–87, 1998
Base Piacenzian Stage	3.60	0.00	Astronomical cycles in sediments	Base of magnetic polarity chronozone C2An (Gauss); extinction levels of planktonic foraminifers *Globorotalia margaritae* (base Zone PL3) and *Pulleniatina primalis.*	Base of beige layer of carbonate cycle 77, Punta Piccola, Sicily, Italy	Ratified 1997	Episodes 21 (2), p. 88–93, 1998
Base Zanclean Stage, base Pliocene Series	5.333	0.00	Astronomical cycles in sediments	Top of magnetic polarity chronozone C3r, ~100 kyr before Thvera normal-polarity subchronozone (C3n.4n). Calcareous nannofossils – near extinction level of *Triquetrorhabdulus rugosus* (base Zone CN10b) and the lowest occurrence of *Ceratolithus acutus.*	Base of Trubi Fm (base of carbonate cycle 1), Eraclea Minoa, Sicily, Italy	Ratified 2000	Episodes 23 (3), p. 179–187, 2000
Miocene Series							
Base Messinian Stage	7.248	0.00	Astronomical cycles in sediments	Astrochronology age of 7.251 Ma; middle of magnetic polarity chronozone C3Br.1r; lowest regular occurrence of the *Globorotalia conomiozea* planktonic foraminifer group.	Base of red layer of carbonate cycle 15, Oued Akrech, Rabat, Morocco	Ratified 2000	Episodes 23 (3), p. 172–178, 2000
Base Tortonian Stage	11.608	0.00	Astronomical cycles in sediments	Last Common Occurrences of the calcareous nannofossil *Discoaster kugleri* and the planktonic foraminifer *Globigerinoides subquadratus.* Associated with the short normal-polarity subchron C5r.2n.	Midpoint of sapropel 76, Monte dei Corvi beach section, Ancona, Italy	Ratified 2003	Episodes article in preparation

Figure 1 Overview of Global Stratotype Sections and Points in the Mesozoic and Cenozoic. The Internet version (see http://www.stratigraphy.org) is being updated regularly to reflect the growing number of ratified stratigraphic boundaries. Reprinted from the International Commission on Stratigraphy (2003), with permission.

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Serravallian Stage	13.64	0.00	Astronomical cycles in sediments	Near lowest occurrence of nannofossil *Sphenolithus heteromorphus,* and within magnetic polarity chronozone C5ABr		GSSP anticipated in 2004	
Base Langhian Stage	15.97	0.0	Calibrated magnetic anomaly scale	Near first occurrence of planktonic foraminifer *Praeorbulina glomerosa* and top of magnetic polarity chronozone C5Cn.1		GSSP anticipated in 2004	
Base Burdigalian Stage	20.43	0.0	Calibrated magnetic anomaly scale	Near lowest occurrence of planktonic foraminifer *Globigerinoides altiaperturus* or near top of magnetic polarity chronozone C6An		Guide event is undecided	
Base Aquitanian Stage, base Miocene Series, base Neogene System	23.03	0.0	Astronomical cycles in sediments	Base of magnetic polarity chronozone C6Cn.2n; lowest occurrence of planktonic foraminifer *Paragloborotalia kugleri*; near extinction of calcareous nannofossil *Reticulofenestra bisecta* (base Zone NN1).	35 m from top of Lemme-Carrosio section, Carrosio village, north of Genoa, Italy	Ratified 1996	Episodes 20 (1), p. 23–28, 1997
Paleogene System							
Oligocene Series							
Base Chattian Stage	28.4	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n. Arbitrary 100 kyr uncertainty assigned.	Planktonic foraminifer, extinction of *Chiloguembelina* (base Zone P21b)	Probably in Umbria-Marche region of Italy	GSSP anticipated in 2004	
Base Rupelian Stage, base Oligocene Series	33.9	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Planktonic foraminifer, extinction of *Hantkenina*	Base of marl bed at 19m above base of Massignano quarry, Ancona, Italy	Ratified 1992	Episodes 16 (3), p. 379–382, 1993
Eocene Series							
Base Priabonian Stage	37.2	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Near lowest occurrence of calcareous nannofossil *Chiasmolithus oamaruensis* (base Zone NP18)	Probably in Umbria-Marche region of Italy		
Base Bartonian Stage	40.4	0.2	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Near extinction of calcareous nannofossil *Reticulofenestra reticulata*			
Base Lutetian Stage	48.6	0.2	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Planktonic foraminifer, lowest occurrence of *Hantkenina*	Leading candidate is Fortuna section, Murcia province, Betic Cordilleras, Spain	GSSP anticipated in 2004	
Base Ypresian Stage, base Eocene Series	55.8	0.2	Astronomical cycles in sediments scaled from base-Paleocene	Base of negative carbon-isotope excursion	Dababiya section near Luxor, Egypt	Ratified 2003	Episodes article in preparation
Paleocene Series							
Base Thanetian Stage	58.7	0.2	Astronomical cycles in sediments scaled from base Paleocene, using base of magnetic polarity chronozone C26n. Arbitrary 0.1 (2 precession cycles, plus the base-Paleogene radiometric) uncertainty assigned to all estimates.	Magnetic polarity chronozone, base of C26n, is a temporary assignment	Leading candidate is Zumaya section, northern Spain	Guide event is undecided	
Base Selandian Stage	61.7	0.2	Astronomical cycles in sediments scaled from base Paleocene, using magnetic polarity chronozone placement of C27n.9	Boundary task group is considering a higher level - base of calcareous nannofossil zone NP5 – which would be ~1 myr younger.	Leading candidate is Zumaya section, northern Spain	Guide event is undecided	

Continued

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Danian Stage, base Paleogene System, base Cenozoic	65.5	0.3	Ar-Ar and U-Pb age agreement	Iridium geochemical anomaly. Associated with a major extinction horizon (foraminifers, calcareous nannofossils, dinosaurs, etc.);	Base of boundary clay, El Kef, Tunisia *(but deterioration may require assigning a replacement section)*	Ratified 1991	
Mesozoic Era							
Cretaceous System				*Most substages of Cretaceous also have recommended GSSP criteria*			
Upper							
Base Maastrichtian Stage	70.6	0.6	Estimated placement relative to Ar-Ar calibrated Sr-curve	Mean of 12 biostratigraphic criteria of equal importance. Closely above is lowest occurrence of ammonite *Pachydiscus neubergicus.* Boreal proxy is lowest occurrence of belemnite *Belemnella lanceolata.*	115.2 m level in Grande Carrière quarry, Tercis-les-Bains, Landes province, SW France	Ratified 2001	Episodes 24 (4), p. 229–238, 2001; Odin (ed.) IUGS Spec. Publ. Series, v.36, Elsevier, 910pp.
Base Campanian Stage	83.5	0.7	Spline fit of Ar-Ar ages and ammonite zones.	Crinoid, extinction of *Marsupites testudinarius*	Leading candidates are in southern England and in Texas		
Base Santonian Stage	85.9	0.7	Spline fit of Ar-Ar ages and ammonite zones.	Inoceramid bivalve, lowest occurrence of *Cladoceramus undulatoplicatus*	Leading candidates are in Spain, England and Texas		
Base Coniacian Stage	89.3	1.0	Spline fit of Ar-Ar ages and ammonite zones.	Inoceramid bivalve, lowest occurrence of *Cremnoceramus rotundatus* (*sensu* Tröger *non* Fiege)	Base of Bed MK47, Salzgitter Salder Quarry, SW of Hannover, Lower Saxony, northern Germany	GSSP anticipated in 2004	
Base Turonian Stage	93.6	0.8	Spline fit of Ar-Ar ages and ammonite zones.	Ammonite, lowest occurrence of *Watinoceras devonense*	Base of Bed 120, Rock Canyon Anticline, east of Pueblo, Colorado, west-central USA	Ratified 2003	Episodes article in preparation
Base Cenomanian Stage	99.6	0.9	Spline fit of Ar-Ar ages and ammonite zones, plus monitor standard correction. Then cycle stratigraphy to place foraminifer datum relative to ammonite zonation.	Planktonic foraminifer, lowest occurrence of *Rotalipora globotruncanoides*	36 m below top of Marnes Bleues Formation, Mont Risou, Rosans, Haute-Alpes, SE France	Ratified 2002	Episodes article in preparation
Lower							
Base Albian Stage	112.0	1.0	Estimated placement relative to bases of Cenomanian and Aptian, with large uncertainty due to lack of GSSP criteria. Ar-Ar age of 114.6 +/– 0.7 Ma from *Parahoplites nutfieldensis* below.	Calcareous nannofossil, lowest occurrence of *Praediscosphaera columnata* (= *P. cretacea* of some earlier studies), is one potential marker.		Guide event is undecided	
Base Aptian Stage	125.0	1.0	Base of M0r, as recomputed from Ar-Ar age from MIT guyot	Magnetic polarity chronozone, base of M0r	Leading candidate is Gorgo a Cerbara, Piobbico, Umbria-Marche, central Italy		
Base Barremian Stage	130.0	1.5	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at M5n.8.	Ammonite, lowest occurrence of *Spitidiscus hugii – Spitidiscus vandeckii* group	Leading candidate is Río Argos near Caravaca, Murcia province, Spain		
Base Hauterivian Stage	136.4	2.0	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at base M11n.	Ammonite, lowest occurrence of genus *Acanthodiscus* (especially *A. radiatus*)	Leading candidate is La Charce village, Drôme province, southeast France		
Base Valanginian Stage	140.2	3.0	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at M14r.3 (base *T. pertransiens*).	Calpionellid, lowest occurrence of *Calpionellites darderi* (base of – Calpionellid Zone E); followed by the lowest occurrence of ammonite *"Thurmanniceras" pertransiens*	Leading candidate is near Montbrun-les-Bains, Drôme province, southeast France		

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Berriasian Stage, base Cretaceous System	145.5	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base of *Berriasella jacobi* zone (M19n.2n.55)	Maybe near lowest occurrence of ammonite *Berriasella jacobi*		Guide event is undecided	
Jurassic System							
Upper							
Base Tithonian Stage	150.8	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M22An	Near base of *Hybonoticeras hybonotum* ammonite zone and lowest occurrence of *Gravesia* genus, and the base of magnetic polarity chronozone M22Ar		Guide event is undecided	
Base Kimmeridgian Stage	155.7	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M26r.2 (Boreal ammonite definition)	Ammonite, near base of *Pictonia baylei* ammonite zone of Boreal realm	Leading candidates are in Scotland, SE France and Poland	GSSP anticipated in 2004	
Base Oxfordian Stage	161.2	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M36An	Ammonite, *Brightia thuouxensis* Horizon at base of the *Cardioceras scarburgense* Subzone (*Quenstedtoceras mariae* Zone)	Leading candidates are in SE France and southernEngland	GSSP anticipated in 2004	
Middle							
Base Callovian Stage	164.7	4.0	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of the genus *Kepplerites (Kosmoceratidae)* (defines base of *Macrocephalites herveyi* Zone in sub-Boreal province of Great Britain to southwest Germany)	Leading candidate is Pfeffingen, Swabian Alb, SW Germany	GSSP anticipated in 2004	
Base Bathonian Stage	167.7	3.5	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of *Parkinsonia (G.) convergens* (defines base of *Zigzagiceras zigzag* Zone)			
Base Bajocian Stage	171.6	3.0	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of the genus *Hyperlioceras* (defines base of the *Hyperlioceras discites* Zone)	Base of Bed AB11, 77.8 m above base of Murtinheira section, Cabo Mondego, western Portugal	Ratified 1996	Episodes 20 (1), p. 16–22, 1997
Base Aalenian Stage	175.6	2.0	Duration of Aalenian-Toarcian from cycle stratigraphy	Ammonite, lowest occurrence of *Leioceras* genus	base of Bed FZ107, Fuentelsalz, central Spain	Ratified 2000	Episodes 24 (3), p. 166–175, 2001
Lower							
Base Toarcian Stage	183.0	1.5	Duration of Aalenian-Toarcian from cycle stratigraphy	Ammonite, near lowest occurrence of a diversified *Eodactylites* ammonite fauna; correlates with the NW European *Paltus* horizon.		Guide event is undecided	
Base Pliensbachian Stage	189.6	1.5	Cycle-scaled linear Sr trend	Ammonite, lowest occurrences of *Bifericeras donovani* and of genera *Apoderoceras* and *Gleviceras.*	Wine Haven section, Robin Hood's Bay, Yorkshire, England, UK	GSSP anticipated in 2003	
Base Sinemurian Stage	196.5	1.0	Cycle-scaled linear Sr trend	Ammonite, lowest occurrence of arietitid genera *Vermiceras* and *Metophioceras*	0.9 m above base of Bed 145, East Quantoxhead, Watchet, West Somerset, SW England, UK	Ratified 2000	Episodes 25 (1), p. 22–26, 2002
Base Hettangian Stage, base Jurassic System	199.6	0.6	U-Pb age just below proposed GSSP for base-Jurassic	Near lowest occurrence of smooth *Psiloceras planorbis* ammonite group		Guide event is undecided	
Triassic System							
Upper							
Base Rhaetian Stage	203.3	1.5	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Near lowest occurrence of ammonite *Cochlocera*,conodonts *Misikella* spp. and *Epigondolella mosheri*, and radiolarian *Proparvicingula moniliformis.*	Key sections in Austria, British Columbia (Canada), and Turkey	Guide event is undecided	

Continued

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Norian Stage	216.5	2.0	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Base of *Klamathites macrolobatus* or *Stikinoceras kerri* ammonoid zones and the *Metapolygnathus communisti* or *M. primitius* conodont zones.	Leading candidates are in British Columbia (Canada), Sicily (Italy), and possibly Slovakia, Turkey (Antalya Taurus) and Oman.	Guide event is undecided	
Base Carnian Stage	228.0	2.0	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Near first occurrence of the ammonoids *Daxatina* or *Trachyceras*, and of the conodont *Metapolygnathus polygnathiformis*	Candidate section at Prati di Stuores, Dolomites, northern Italy. Important reference sections in Spiti (India) and New Pass, Nevada (USA).	Guide event is undecided	
Middle							
Base Ladinian Stage	237.0	2.0	U-Pb array by Mundil *et al.* on levels near *Nevadites* (= *Secedensis*) ammonite zone in Dolomites, plus placement relative to magnetostratigraphy corrlations to cycle-scaled Newark magnetic polarity pattern	Alternate levels are near base of *Reitzi, Secedensis,* or *Curionii* ammonite zone; near first occurrence of the conodont genus *Budurovignathus.*	Leading candidates are Bagolino (Italy) and Felsoons (Hungary). Important reference sections in the Humboldt Range, Nevada (USA).	Guide event is undecided	
Base Anisian Stage	245.0	1.5	Proportional subzonal scaling	Ammonite, near lowest occurrences of genera *Japonites, Paradanubites, and Paracrochordiceras;* and of the conodont *Chiosella timorensis*	Candidate section probable at Desli Caira, Dobrogea, Romania; significant sections in Guizhou Province (China).	GSSP anticipated in 2004	
Lower							
Base Olenekian Stage	249.7	0.7	Composite standard from conodonts scaled to base-Anisian and base-Triassic	Near lowest occurrence of *Hedenstroemia* or *Meekoceras gracilitatis* ammonites, and of the conodont *Neospathodus waageni.*	Candidate sections in Siberia (Russia) and probably Chaohu, Anhui Province, China. Important sections also in Spiti.	Guide event is undecided	
Base Induan Stage, base Triassic System, base Mesozoic	251.0	0.4	U-Pb ages bracket GSSP (Bowring *et al.*, 1998)	Conodont, lowest occurrence of *Hindeodus parvus*; termination of major negative carbon-isotope excursion. About 1 myr after peak of Late Permian extinctions.	Base of Bed 27c, Meishan, Zhejiang, China	Ratified 2001	Episodes 24 (2), p. 102–114, 2001

The steps involved in modern time-scale construction can be summarized as follows:

Step 1. Construct an updated global chronostratigraphic scale for Earth's rock record.

Step 2. Identify key linear-age calibration levels for the chronostratigraphic scale using radioisotopic age dates.

Step 3. Apply Earth-orbit tuning to intervals with cyclic sediments or stable isotope sequences that have sufficient biostratigraphic or magnetostratigraphic ties.

Step 4. Interpolate the combined chronostratigraphic and chronometric scale when direct information is insufficient.

Step 5. Calculate or estimate error bars on the combined chronostratigraphic and chronometric information to obtain a geological time-scale with estimates of uncertainty on boundaries and on unit durations.

Step 6. Peer-review the resultant geological time-scale.

The first step, integrating multiple types of stratigraphic information in order to construct the chronostratigraphic scale, is the most time-consuming; it summarizes and synthesizes centuries of detailed geological research and tries to understand all relative correlations and calibration to the standard chronostratigraphic scale.

The second step, identifying which radiometric and cycle-stratigraphic studies to use as the primary constraints for assigning linear ages, is the one that has much evolved. Historically, Phanerozoic time-scale building went from an exercise with very few and relatively inaccurate radioisotopic dates, as available to the pioneer of the geological time-scale Arthur Holmes, to one with many dates with greatly varying analytical precision, as in the mid-1980s of the past century. Next, time-scale studies started

Global Boundary Stratotype Sections and Points (GSSPs)

Status in Oct 2003; see ICS website (www.stratigraphy.org) for updates.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
PHANEROZOIC							
Base Induan Stage, base Triassic System, base Mesozoic	251.0	0.4	Average of U-Pb constraints from Bowring *et al.* (1998)	Conodont, lowest occurrence of *Hindeodus parvus*; termination of major negative carbon-isotope excursion. About 1 myr after peak of Late Permian extinctions.	Base of Bed 27c, Meishan, Zhejiang, China	Ratified 2001	Episodes 24 (2), p. 102–114, 2001
Paleozoic Era							
Permian System			Permian-Carboniferous time scale is derived from calibrating a master composite section to selected radiometric ages				
Lopingian							
Base Changhsingian Stage	253.8	0.7	"	Conodont, near lowest occurrence of conodont *Clarkina wangi*	Leading candidates are in China		
Base Wuchiapingian Stage	260.4	0.7	"	Conodont, near lowest occurrence of conodont *Clarkina postbitteri*	Candidate section is Tieqiao rail-bridge section, Laibin Syncline, Guangxi Province, China	Ratification pending (Jan'04)	
Guadalupian							
Base Capitanian Stage	265.8	0.7	"	Conodont, lowest occurrence of *Jinogondolella postserrata*	4.5 m above base of Pinery Limestone Member, Nipple Hill, SE Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Base Wordian Stage	268.0	0.7	"	Conodont, lowest occurrence of *Jinogondolella aserrata*	7.6 m above base of Getaway Ledge outcrop, Guadalupe Pass, SE Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Base Roadian Stage, base Guadalupian Series	270.6	0.7		Conodont, lowest occurrence of *Jinogondolella nanginkensis*	42.7 m above base of Cutoff Formation, Stratotype Canyon, southern Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Cisuralian Series							
Base Kungurian Stage	275.6	0.7		Conodont, near lowest occurrence of conodont *Neostreptognathus pnevi-N. exculptu*	Leading candidates are in southern Ural Mtns.		
Base Artinskian Stage	284.4	0.7	"	Conodont, lowest occurrence of conodont *Sweetognathus whitei-Mesogondolella bisselli*	Leading candidates are in southern Ural Mtns.		
Base Sakmarian Stage	294.6	0.8	"	Conodont, near lowest occurrence of conodont *Streptognathodus postfusus*	Leading candidate is at Kondurovsky, Orenburg Province, Russia.		
Base Asselian Stage, base Cisuralian Series, base Permian System	299.0	0.8	"	Conodont, lowest occurrence of *Streptognathodus isolatus* within the *S. "wabaunsensis"* conodont chronocline. 6 m higher is lowest fusilinid foraminifer *Sphaeroschwagerina.*	27 m above base of Bed 19, Aidaralash Creek, Aktöbe, southern Ural Mountains, northern Kazakhstan	Ratified 1996	Episodes 21 (1), p. 11–18, 1998
Carboniferous System							
Pennsylvanian Subsystem				*Series classification and nomenclature are currently under discussion*			
Base Gzhelian Stage	303.9	0.9	"	Near lowest occurrences of the fusulinids *Daixina, Jigulites* and *Rugosofusulina, or lowest occurrence of* Streptognathodus zethu.		Guide event is undecided	

Figure 2 Overview of Global Stratotype Sections and Points and Global Standard Stratigraphic Ages in the Palaeozoic and Precambrian, respectively. The Internet version (see http://www.stratigraphy.org) is being updated regularly to reflect the growing number of ratified stratigraphic boundaries. Reprinted from the International Commission on Stratigraphy (2003), with permission.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Kasimovian Stage	306.5	1.0	"	Near base of *Obsoletes obsoletes* and *Protriticites pseudomontiparus* fusulinid zone, or lowest occurrence of *Parashumardites* ammonoid.		Guide event is undecided	
Base Moscovian Stage	311.7	1.1	"	Near lowest occurrences of *Declinognathodus donetzianus* and/or *Idiognathoides postsulcatus* conodont species, and fusulinid species *Aljutovella aljutovica*.		Guide event is undecided	
Base Bashkirian Stage, base Pennsylvanian Subsystem	318.1	1.3	"	Conodont, lowest occurrence of *Declinognathodus nodiliferus* s.l.	82.9 m above top of Battleship Wash Fm., Arrow Canyon, southern Nevada, USA	GSSP ratified 1996. Subsystem rank of Mississippian and Pennsylvanian names ratified 2000.	Episodes 22 (4), p. 272–283, 1999
Mississippian Subsystem				*Series classification and nomenclature are currently under discussion*			
Base Serpukhovian	326.4	1.6	"	Near lowest occurrence of conodont, *Lochriea crusiformis*.		Guide event is undecided	
Base Visean	345.3	2.1	"	Foraminifer, lineage *Eoparastaffella simplex* morphotype 1/morphotype 2	Leading candidate is Pengchong, south China		
Base Tournaisian, base Mississippian Subsystem, base Carboniferous System	359.2	2.5	"	Conodont, above lowest occurrence of *Siphonodella sulcata*	Base of Bed 89, La Serre, Montagne Noir, Cabrières, southern France	Ratified 1990	Episodes 14 (4), p. 331–336, 1991
Devonian System			Devonian time scale is a statistical fit of a composite biostratigraphic zonation (based on Figure 8 of Williams *et al.*, 2000) to selected radiometric ages				
Upper							
Base Famennian Stage	374.5	2.6	"	Just above major extinction horizon (Upper Kellwasser Event), including conodonts *Ancyrodella* and *Ozarkodina and goniatites of Gephuroceratidae and Beloceratidae*	base of Bed 32a, upper Coumiac quarry, Cessenon, Montagne Noir, southern France	Ratified 1993	Episodes 16 (4), p. 433–441, 1993
Base Frasnian Stage	385.3	2.6	"	Conodont, lowest occurrence of *Ancyrodella rotundiloba* (defines base of Lower *Polygnathus asymmetricus* conodont Zone)	Base of Bed 42a', Col du Puech de la Suque section, St. Nazaire-de Ladarez, SE Montagne Noir, southern France	Ratified 1986	Episodes 10 (2), p. 97–101, 1987
Middle							
Base Givetian Stage	391.8	2.7	"	Conodont, lowest occurrence of Polygnathus hemiansatus, near base of goniatite Maenioceras Stufe	Base of Bed 123, Jebel Mech Irdane ridge, Tafilalt, Morocco	Ratified 1994	Episodes 18 (3), p. 107–115, 1995
Base Eifelian Stage	397.5	2.7	"	Conodont, lowest occurrence of *Polygnathus costatus partitus*; major faunal turnover	Base unit WP30, trench at Wetteldorf Richtschnitt, Schönecken-Wetteldorf, Eifel Hills, western Germany	Ratified 1985	Episodes 8 (2), p. 104–109, 1985
Lower							
Base Emsian Stage	407.0	2.8	"	Conodont, lowest occurrence of *Polygnathus kitabicus* (= *Po. dehiscens*)	Base of Bed 9/5, Zinzil' ban Gorge, SE of Samarkand, Uzbekistan	Ratified 1995	Episodes 20 (4), p. 235–240, 1997
Base Pragian Stage	411.2	2.8	"	Conodont, lowest occurrence of *Eognathodus sulcatus*	Base of Bed 12, Velká Chuchle quarry, southwest part of Prague city, Czech Republic	Ratified 1989	Episodes 12 (2), p. 109–113, 1989

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Lochkovian Stage, base Devonian System	416.0	2.8	base-Devonian from scale in Cooper (this volume), which is 1 myr younger than Tucker et al (1998) estimate.	Graptolite, lowest occurrence of *Monograptus uniformis*	Within Bed 20, Klonk, Barrandian area, southwest of Prague, Czech Republic	Ratified 1972	Martinsson (ed.), The Silurian-Devonian Boundary, IUGS Series A, no.5, 349 pp., 1977
Silurian System			Silurian and Ordovician time scales are from calibrating a CONOP composite graptolite zonation to selecte radiometric ages				Holland and Bassett (eds), *A Global Standard for the Silurian System*, Nat. Mus. Wales, Geol. Series No.10, Cardiff, 325 pp., 1989
Pridoli Series							
Base Prídolí Series *(not subdivided in stages)*	418.7	2.7	"	Graptolite, lowest occurrence of *Monograptus parultimus*	Within Bed 96, Pozáry section near Reporje, Barrandian area, Prague, Czech Republic	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Ludlow Series		2.6					
Base Ludfordian Stage	421.3	2.6	"	*Imprecise.* May be near base of *Saetograptus leintwardinensis* graptolite zone.	Base of lithological unit C, Sunnyhill Quarry, Ludlow, Shropshire, southwest England, UK.	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Base Gorstian Stage	422.9	2.5	"	*Imprecise.*Just below base of local acritarch *Leptobrachion longhopense* range zone. May be near base of *Neodiversograptus nilssoni* graptolite zone.	Base of lithological unit F, Pitch Coppice quarry, Ludlow, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Wenlock Series							
Base Homerian Stage	426.2	2.4	"	Graptolite, lowest occurrence of *Cyrtograptus lundgreni* (defines base of *C. lundgreni* graptolite zone)	Graptolite biozone intersection in stream section in Whitwell Coppice, Homer, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Base Sheinwoodian Stage	428.2	2.3	"	*Imprecise.* Between the base of acritarch biozone 5 and extinction of conodont *Pterospathodus amorphognathoides.* May be near base of *Cyrtograptus centrifugus* graptolite zone.	Base of lithological unit G, Hughley Brook, Apedale, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Llandovery Series							
Base Telychian Stage	436.1	1.9	"	Brachiopods, just above extinction of *Eocoelia intermedia* and below lowest succeeding species *Eocoelia curtisi.* Near base of *Monograptus turriculatus* graptolite zone.	Locality 162 in transect d, Cefn Cerig road, Llandovery area, south-central Wales, UK	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Base Aeronian Stage	439.0	1.8	"	Graptolite, lowest occurrence of *Monograptus austerus sequens* (defines base of *Monograptus triangulatus* graptolite zone)	Base of locality 72 in transect h, Trefawr forestry road, north of Cwm-coed-Aeron Farm, Llandovery area, south-central Wales, UK	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Base Rhuddanian Stage, base Silurian System	443.7	1.5	"	Graptolites, lowest occurrences of *Parakidograptus acuminatus* and *Akidograptus ascensus*	1.6 m above base of Birkhill Shale Fm., Dob's Linn, Moffat, Scotland, UK	Ratified 1984	Episodes 8 (2), p. 98–100, 1985
Ordovician System							
Upper							
Base of sixth stage *(not yet named)*	450.2	1.6	"	Potentially near first appearance of the graptolite *Dicellograptus complanatus* and of the conodont *Amorphognathus ordovicicus.*		*Guide event is undecided*	

Continued

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base of fifth stage *(not yet named)*	460.9	1.6	"	Graptolite, lowest occurrence of *Nemagraptus gracilis*	1.4 m below phosphorite in E14a outcrop, Fågelsång, Scane, southern Sweden	Ratified 2002	Episodes 23 (2), p. 102–109, 2000 *(proposal; formal GSSP publication in preparation).*
Middle							
Base Darriwilian Stage	468.1	1.6	"	Graptolite, lowest occurrence of *Undulograptus austrodentatus*	Base of Bed AEP184, 22 m below top of Ningkuo Fm., Huangnitang, Changshan, Zhejiang province, southeast China	Ratified 1997	Episodes 20 (3), p. 158–166, 1997
Base of third stage *(not yet named)*	471.8	1.6	"	Conodont, lowest occurrence of *Tripodus laevis*			
Lower							
Base of second stage *(not yet named)*	478.6	1.7	"	Graptolite, lowest occurrence of *Tetragraptus approximatus*	Just above E bed, Diabasbrottet quarry, Västergötland, southern Sweden	Ratified 2002	Episodes article in preparation
Base of Tremadocian Stage, base Ordovician System	488.3	1.7	"	Conodont, lowest occurrence of *Iapetognathus fluctivagus*;just above base of *Cordylodus lindstromi* conodont Zone. Just below lowest occurrence of planktonic graptolites. Currently dated around 489 Ma.	Within Bed 23 at the 101.8 m level, Green Point, western Newfoundland, Canada	Ratified 2000	Episodes 24 (1), p. 19–28,
Cambrian System				Potential GSSP correlation levels include *Cordylodus proavus, Glyptagnostus reticulatus, Ptychagnostus punctuosus, Acidusus atavus,* and *Oryctocephalus indicus.*		Overview of potential subdivisions in Episodes 23 (3), p. 188–195, 2000.	
Upper ("Furongian") Series							
Upper stage(s) in Furongian				*Potential GSSP levels in upper Cambrian are based on trilobites and condonts*			
Base Paibian Stage, base Furongian Series	501.0	2.0	Radiometric ages near primary marker level. Estimated age and uncertainty only.	Trilobite, lowest occurrence of agnostoid *Glyptagnostus reticulatus.* Coincides with base of large positive carbon-isotope excursion.	369.06 m above base of Huaqiao Fm, Paibi section, NW Hunan province, south China	Ratified 2003	Episodes article in preparation
Middle	513.0	2.0	Radiometric ages near primary marker level. Estimated age and uncertainty only.	*Potential GSSP levels in Middle Cambrian are based mainly on trilobites*			
Lower				*Potential GSSP levels in Lower Cambrian are based on archaeocyatha, small shelly fossils, and to a lesser extent, trilobites*			
Base Cambrian System, base Paleozoic, base PHANEROZOIC	542.0	1.0	U-Pb age from Oman coinciding with the negative carbon excursion.	Trace fossil, lowest occurrence of *Treptichnus (Phycodes) pedum.* Near base of negative carbon-isotope excursion.	2.4 m above base of Member 2 of Chapel Island Fm., Fortune Head, Burin Peninsula, southeast Newfoundland, Canada	Ratified 1992	Episodes 17 (1&2), p. 3–8, 1994.
PROTEROZOIC				*Pre-Cambrian eras and systems below Ediacaran are defined by absolute ages, rather than stratigraphic points.*			
Neoproterozoic Era							
Base Ediacaran System	600		Vague estimation from bracketing radiometric ages	Termination of Varanger (or Marinoan) glaciation.	Base of the Nuccaleena Formation cap carbonate, immediately above the Elatina diamictite in the Enorama Creek section, Flinders Ranges, South Australia.	Age-definition (650 Ma) ratified in 1990, but ICS now voting on Australian GSSP	Episodes 14 (2), p. 139–140, 1991

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Cryogenian System	850		Defined chronometrically	Base = 850 Ma			
Tonian System	1000		Defined chronometrically	Base = 1000 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Mesoproterozoic Era							
Stenian System	1200		Defined chronometrically	Base = 1200 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Ectasian System	1400		Defined chronometrically	Base = 1400 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Calymmian System	1600		Defined chronometrically	Base = 1600 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Paleoproterozoic Era							
Statherian System	1800		Defined chronometrically	Base = 1800 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Orosirian System	2050		Defined chronometrically	Base = 2050 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Rhyacian System	2300		Defined chronometrically	Base = 2300 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Siderian System	2500		Defined chronometrically	Base = 2500 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
ARCHEAN							
Neoarchean Era	2800		Defined chronometrically	Base = 2800 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Mesoarchean Era	3200		Defined chronometrically	Base = 3200 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Paleoarchean Era	3600		Defined chronometrically	Base = 3600 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Eoarchean Era			Base is not defined				

to appear for selected intervals, such as Paleogene, Late Cretaceous, or Ordovician; these studies selected a small suite of radioisotopic dates with high internal analytical precision and relatively precise stratigraphic position. At the same time, a high-resolution Neogene time-scale started to take shape, using orbital tuning of long sequences of sedimentary and/or oxygen isotope cycles in the Mediterranean region and in Atlantic and Pacific pelagic sediments. The present trend for the pre-Neogene is to incorporate radioisotopic dates that have very small analytical and stratigraphic uncertainties, and pass the most stringent tests.

The third step, interpolating the stratigraphic and radiometric information, has also evolved. An early method had already constructed the basic two-way graph and was being used. This graph plotted the cumulative sum of maximum thicknesses of strata in thousands of feet per stratigraphic unit along the vertical axis and selected dates from volcanic tuffs, glauconites, and magmatic intrusives along the horizontal linear axis. This 'best-fit' line method, shown in Figure 3 as it was used in 1960, incorporated an uncertainty envelope from the errors on the radioisotopic age constraints. Despite its crudeness, the method was remarkably effective, but is a far cry from methods used today.

In the mid-1990s, Frits Agterberg and Felix Gradstein started to apply mathematical-statistical error analysis to the time-scale ages, which, for the first time, allowed them to assign fairly realistic error bars to ages of Mesozoic stage boundaries, a trend that persists today. A simplified introduction to the modern building tools is presented the following discussions, but first consider a frequently asked question: 'Which time-scale should be used?'.

Which Time Scale Should Be Used?

Of the several geological time-scales published in the past decade, a simple answer is to use the most recent one. A better answer is to investigate the pros and cons of these time-scales, the improvements that a new one might bring to a study, and to remember that it is not desirable to change units of measurement during execution of an experiment or survey or basin modelling exercise.

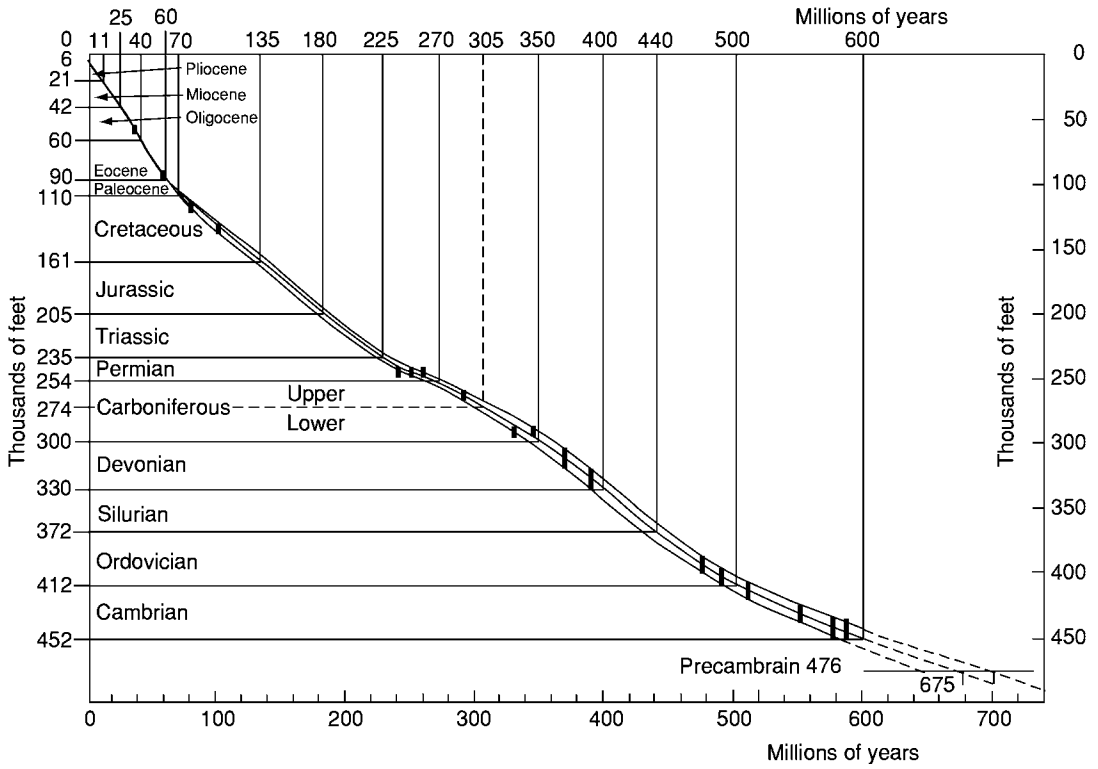


Figure 3 Scaling concept employed by Arthur Holmes in the first half of the twentieth century to construct the geological time scale. The cumulative sum of maximum thicknesses of strata, in thousands of feet per stratigraphic unit, is plotted along the vertical axis; selected radiometric dates from volcanic tuffs, glauconites, and magmatic intrusives are plotted along the horizontal linear axis. This version incorporated an uncertainty envelope from the errors on the radiometric age constraints. Modified from Holmes A (1960) A revised geological time scale. *Transactions of the Edinburgh Geological Society* 17: 183 216, with permission.

Certain qualities of a geological time-scale should be considered:

1. Does it use updated chronostratigraphic nomenclature of international standard?
2. Does it use all stratigraphic tools relevant to the interval for which the time-scale is intended, including magnetostratigraphy and its chronology, standard biozonations, stable isotope stratigraphy, cyclic stratigraphy interpolations, and/or orbital time-scale calibrations?
3. Are the stratigraphically meaningful radiometric age dates based on rigorous methodology, with the most up-to-date and accepted interlaboratory standards?
4. Does the scale bring out uncertainty in the age of individual stratigraphic boundaries?
5. Is the scale used in major geological projects and studies?

The ICS is actively engaged in time-scale construction and publication, which assists in bringing consensus and stability to the use of this international geological standard. Figure 4 shows the most up-to-date edition of the ICS scale.

Music of the Spheres

The sedimentary cycles approach to time-scale building, as is now standard for the time-scale of the past 23 My (Neogene), provides superior resolution and precision. Gravitational interactions of Earth with the sun, moon, and other planets cause systematic changes in Earth's orbital and rotational systems. These interactions give rise to cyclic oscillations in the eccentricity of Earth's orbit, and in the tilt and precession of Earth's rotational axis, with mean dominant periods of 100 000, 41 000, and 21 000 years, respectively (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). The associated cyclic variations in annual and seasonal solar radiation falling onto different latitudes alter long-term climate in colder versus warmer, and wetter versus dryer, periods. These, in turn, create easily recognizable sedimentary cycles, such as regular interbeds of limy and shaly facies. Massive outcrops of hundreds or thousands of such cycles are observed in numerous geological basins (e.g., around the Mediterranean) and in sediment cores from ocean-based drilling sites.

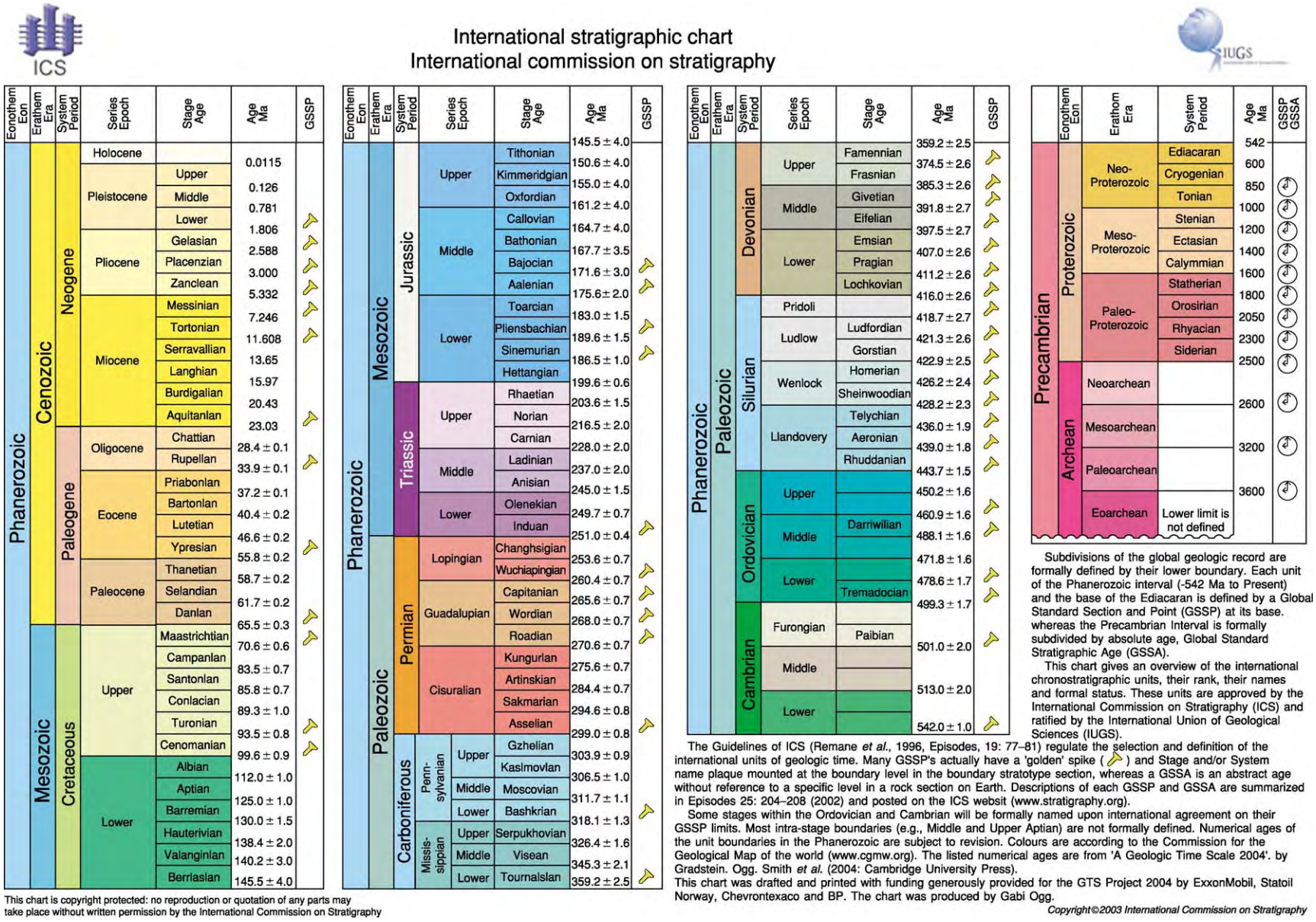


Figure 4 The international stratigraphic chart and time scale for the Phanerozoic, issued by the International Commission on Stratigraphy (ICS) (available on the Internet at http://www.stratigraphy.org). Reprinted from the International Commission on Stratigraphy (2003), with permission.

Greatly detailed counts these centimetre- to metre-thick cycles over land outcrops and in ocean drilling wells, combined with the additional correlation aids provided by magnetostratigraphy, oxygen isotope stratigraphy, and biostratigraphy, have been used to produce a very detailed Neogene sediment/orbital cycle pattern. The critical step is the direct linkage of each cycle to the theoretical computed astronomical scale of the 21 000-, 41 000-, and 100 000-year palaeoclimatic cycles. This astronomical tuning of the geological cycle record from the Mediterranean and Atlantic by earth scientists at Utrecht and Cambridge universities (e.g., Luc Lourens, Frits Hilgen, and Nick Shackleton) has led to unprecedented accuracy and resolution for the past 23 My. In New Zealand, Tim Naish and colleagues have calibrated the Upper Neogene record to the standard Neogene time-scale. Using the high-resolution land-based cycle, the isotope and magnetic record in the Wanganui Basin, these authors thereby transferred precise absolute ages to local shallow marine sediments and demonstrated the link between sequence and cycle stratigraphy.

Efforts are under way to extend the continuous astrochronological scale back into the Oligocene and Eocene by applying a combination of cycle stratigraphy, improved astronomical projections, oxygen isotope stratigraphy, and magnetostratigraphy to the deep-sea stratigraphic record. A special application of orbitally tuned cyclic sediment sequences is to 'rubber band' stratigraphically floating units, such as parts of the Paleocene, the Albian, and parts of the Lower Jurassic. A quantitative estimation of the duration of all cycles within a stratigraphic unit allows estimating their duration.

Decay of Atoms

For rocks older than the Neogene, the derivation of a numerical time-scale depends on the availability of suitable radioisotopic ages. Radioisotopic dating generally involves measuring the ratio of the original element in a mineral, such as sanidine feldspar or zircon, to its isotopic daughter products. The age of a mineral may then be calculated by means of the isotopic decay constant. Depending on the half-life of the element, several radioisometric clocks are available; $^{40}Ar/^{39}Ar$ and the family of U/Pb isotopes are the most common suites currently applied to the Phanerozoic, because of analytical precision and utility with tuffaceous beds in marine or non-marine sequences.

Radioisometric dating of sedimentary rocks follows several geological strategies:

1. Dating of igneous intrusions within sediments records the time of primary cooling, when the igneous rocks were emplaced and had cooled sufficiently (to a few hundreds of degrees centigrade) to set the radiometric decay clock in action. (Note: Because of uncertainty in the relation of the intrusion to the host sediment, such dates may be of limited stratigraphic use.)
2. Dating of volcanic flows and tuffs as part of the stratified sedimentary succession.
3. Dating of authigenic sedimentary minerals (mainly involving glauconite) found commonly in many marine sediments.

Note that mild heating or overburden pressure after burial may lead to loss of argon, the daughter product measured in the $^{40}K/^{40}Ar$ clock in glauconite. Another problem is that glauconite also contains an abundance of tiny flakes that allow diffusion of Ar at low temperatures. The result is that glauconite dates may be too young. Because of such problems, which may be difficult to detect, modern geological time-scales avoid dates based on glauconite.

Calibration of the decay constants or measurement standards can be enhanced by intercalibration to other radioisotopic methods, or by dating rocks of a known age (for example, a volcanic ash within an astronomically tuned succession). Astrochronological and interlaboratory recalibration of the $^{40}Ar/^{39}Ar$ monitor standard indicates that many of the $^{40}Ar/^{39}Ar$ ages used in previous Phanerozoic time-scales are too young by about 0.5–1.0%. For example, the age dating to 65.0 Ma that was assigned 10 years ago to the top Cretaceous is now 65.5 Ma.

Radioisotopic dating techniques with less than 1.0% analytical error are providing suites of high-precision U/Pb and Ar/Ar dates for the Palaeozoic and Mesozoic. The integration of this level of chronometric precision with high-resolution biostratigraphy, magnetostratigraphy, or cyclic scales is a major challenge to time-scale studies. Even the most detailed biostratigraphic scheme probably has no biozonal units of less than half a million years in duration, not to speak of the actual precision in dating a particular 'stratigraphic piercing' point, for which a U/Pb age estimate would be available with an analytical uncertainty of 100 000 years or more. Similarly, combining analytically less precise K/Ar dates with much more precise Ar/Ar or U/Pb dates in statistical interpolations creates a strong bias towards the latter, despite the fact that both may have equal litho-, bio-, and chronostratigraphic precision. Nevertheless, the combination of precise stratigraphic definitions through GSSPs and accurate radiometric dates near these levels is paving the way for a substantial increase in the precision and accuracy of the geological time-scale. The bases of Palaeozoic, Mesozoic, and

Cenozoic are bracketed by analytically precise ages at their GSSP or primary correlation markers – 542 ± 1.0, 251.0 ± 0.4, and 65.5 ± 0.3 Ma, respectively – and there are direct age dates for basal Carboniferous, basal Permian, basal Jurassic, basal Aptian, basal Cenomanian, and basal Oligocene. Most other period or stage boundaries lack direct age control. Therefore, the third step, linear interpolation, also plays a key role for development of the time-scale.

Interpolation and Statistics

Despite progress in standardization and dating, parts of the Mesozoic and Palaeozoic have sparse radioisotopic records. Ideally, each of the 90 or more stage boundaries that comprise the Palaeozoic, Mesozoic, and Cenozoic eras of the Phanerozoic should coincide with an accurate radioisotopic date from volcanic ashes that coincide with each of the stage boundaries. However, this coincidence is rare in the geological record. The combined number of fossil events and magnetic reversals far exceeds the total number of radioisotopically datable horizons in the Phanerozoic. Therefore, a framework of bio-, magneto-, and chronostratigraphy provides the principal fabric for stretching of the relative time-scale between dated tiepoints on the loom of linear time. For such stretching, interpolation methods are employed that are both geological and statistical in nature.

The outdated method of plotting the cumulative global thickness of periods against selected linear age dates was previously mentioned. Among the modern geological scaling methods, an assumption of relative constancy of seafloor spreading over limited periods of time is a common tool for interpolating the latest Cretaceous through Palaeogene relative scale. Magnetic polarity chrons – the units of magnetochronology – can be recognized both on the ocean floor, as magnetic anomalies measured in kilometres from the midocean spreading center, and in marine sediments, as polarity zones that contain biostratigraphic events and can be linked to linear time (*see* **Magnetostratigraphy**). Knowing the linear age of a few ocean crust magnetic anomalies (Earth magnetic reversals, or magnetochrons) allows interpolation of the ages of the intervening magnetic pattern, which, in turn, can be correlated to the fossil record and geological stage boundaries. The subduction of pre-Late Jurassic oceanic crust precludes such an interpolation approach for older Mesozoic and Palaeozoic strata.

A second geological method involves building a zonal composite to scale stages. Several outstanding examples are documented in the geological time-scale built by a large team of International Geological Congress (IGC) scientists in 2004. For this scale, Roger Cooper and colleagues have built a very detailed composite standard of graptolite zones from 200 or more sections in oceanic and slope environment basins for the uppermost Cambrian, Ordovician, and Silurian intervals. With zone thickness taken as directly proportional to zone duration, the detailed composite sequence was scaled using selected, high-precision age dates. For the Carboniferous through Permian, a composite standard of conodont, fusulinid, and ammonoid events from many classical sections can now be calibrated to a combination of U/Pb and $^{40}Ar/^{39}Ar$ dates. A composite standard of conodont zones was used for the Early Triassic. This procedure directly scales all stage boundaries and biostratigraphic horizons.

The two-way graphs of linear age versus scaled stages require a best-fit method, and that is where statistics comes into play, with cubic spline fitting and maximum-likelihood interpolation most suitable. On the time-scale of Figure 4, a majority of Phanerozoic stage boundaries for the first time show error bars; an exception is the Neogene Period, wherein analytical errors are negligible. These error bars reflect both radioisotopic and stratigraphic uncertainty. In addition, error bars were calculated on stage duration. Uncertainty in the duration of the age units is less than the error in age of their boundaries.

See Also

Analytical Methods: Geochronological Techniques. **Biozones**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Famous Geologists:** Steno. **Magnetostratigraphy**. **Stratigraphical Principles**.

Further Reading

Agterberg FP (2004) Geomathematics. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Bleeker W (2004) Towards a natural Precambrian time scale. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Bowring SA, Erwin DH, Jin YG, Martin MW, Davidek K, and Wang W (1998) U/Pb zircon geochronology and tempo of the end Permian mass extinction. *Science* 280: 1039 1045.

Cande SC and Kent DV (1995) Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *Journal of Geophysical Research* 100: 6093 6095.

Carter RM and Naish TR (eds.) (1999) *The High Resolution, Chronostratigraphic and Sequence Stratigraphic*

Record of the Plio Pleistocene, Wanganui Basin, New Zealand. Folio Series 2. Lower Hutt, NZ: Institute of Geological and Nuclear Sciences.

Gradstein FM, Agterberg FP, Ogg JG, Hardenbol J, van Veen P, Thierry J, and Huang Z (1995) A Triassic, Jurassic, and Cretaceous time scale. *SEPM Special Publication* 54: 95–126.

Gradstein FM, Ogg JG, Smith AG, *et al.* (2004) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Harland WB, Armstrong RL, Cox AV, Craig LE, Smith AG, and Smith DG (1990) *A Geologic Time Scale 1989*. New York: Cambridge University Press.

Herbert TD, D'Hondt SL, Premoli Silva I, Erba E, and Fischer AG (1995) Orbital chronology of Cretaceous–Early Palaeocene marine sediments. In: Berggren WA, Kent DV, and Hardenbol J (eds.) *Geochronology, Time Scales and Global Stratigraphic Correlations: A Unified Temporal Framework for a Historical Geology, SEPM Special Volume, No. 54*, pp. 81–94. Tulsa, OK: SEPM.

Hilgen FJ, Krijgsman W, Langereis CG, and Lourens LJ (1997) Breakthrough made in dating of the geological record. *Eos (Transactions of the American Geophysical Union)* 78(28): 285, 288–289.

Holmes A (1947) The construction of a geological time-scale. *Transactions Geological Society of Glasgow* 21: 117–152.

Holmes A (1960) A revised geological time scale. *Transactions of the Edinburgh Geological Society* 17: 183–216.

Kamo SL, Czamanske GK, Amelin Y, Fedorenko VA, Davis DW, and Trofimov VR (2003) Rapid eruption of Siberian flood volcanic rocks and evidence for coincidence with the Permian–Triassic boundary and mass extinction at 251 Ma. *Earth and Planetary Science Letters* 214: 75–91.

Lourens L, Hilgen F, Shackleton NJ, Laskar L, and Wilson D (2004) The Neogene Period. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Martinsson A (ed.) (1977) *The Siluro–Devonian boundary. International Union of Geological Sciences, Series A*, vol. 5. Vienna: International Union of Geological Sciences.

Obradovich JD (1993) A Cretaceous time scale. In: Caldwell WGE (ed.) *Evolution of the Western Interior Basin, Geological Association of Canada Special Paper 39*, pp. 379–396. St. John's, NL: Geological Association of Canada.

Ogg JG (2004) The Triassic, Jurassic and Cretaceous. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Renne PR, Deino AL, Walther RC, Thurrin BD, Swisher CC, Becker TA, Curtis GH, Sharp WD, and Jaouni AR (1994) Intercalibration of astronomical and radioisotopic time. *Geology* 22: 783.

Röhl U, Ogg JG, Geib TL, and Wefer G (2001) Astronomical calibration of the Danian time scale. *Geological Society, Special Publication* 183: 163–184.

Shackleton NJ, Crowhurst SJ, Weedon GP, and Laskar J (1999) Astronomical calibration of Oligocene–Miocene time. *Philosophical Transactions of the Royal Society of London, Series A* 357: 1907–1929.

Villeneuve M (2004) Radiogenic isotope chronology. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge UK: Cambridge University Press.

Weedon GP, Jenkyns HC, Coe AL, and Hesselberg SP (1999) Astronomical calibration of the Jurassic time scale from cyclostratigraphy in British mudrock formations. *Philosophical Transactions of the Royal Society of London, Series A* 357: 1787–1813.

TRACE FOSSILS

P J Orr, University College Dublin, Dublin, Ireland

Introduction

Trace fossils and ichnofabrics offer an alternative source of data on the ecology of any palaeoecosystem to that provided by the body fossil record. Different preservational biases apply to the trace and body fossil records. 'Soft-bodied' organisms, i.e., those lacking biomineralized tissues, can produce trace fossils, yet their fossilization potential is minimal. Most trace fossils are emplaced into unconsolidated sediment, cannot survive reworking, and are thus autochthonous (in marked contrast to the vast majority of body fossils). The study of trace fossils and ichnofabrics has made a fundamental and growing contribution to our understanding of the evolutionary palaeoecology of the Earth's biosphere. Furthermore, bioturbation impacts on the physical and chemical properties of sediments, including their diagenesis. Key terms employed in the taxonomic, preservational, and ethological classification of trace fossils are defined; how trace fossils and ichnofabrics are used in palaeoenvironmental and community reconstructions is discussed.

'Trace fossil' is used herein to describe any discrete structure produced by the reworking of sediment or the bioerosion of lithic (rock) or xylic (wood) substrates by infauna or epifauna. Eggs and sedimentary structures produced by physical processes do not represent trace fossils. Some authorities include stromatolites within this term, on which basis

Figure 1 Wrinkle structures in siliciclastic lithologies attributed to the former presence of a microbial mat. Image courtesy of Séan Burke.

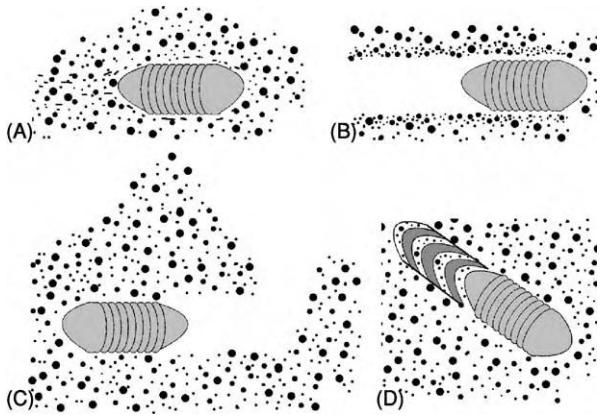


Figure 2 Style of burrowing reflects substrate consistency. (A) Intrusive burrowing typical of near surface soupgrounds. (B) Compression burrowing generates an open burrow and no spoils. (C) Excavation, yielding an open burrow and advected spoil. (D) Excavation, producing a backfilled burrow. In the latter case, the active infill, i.e., by the producer, may be structureless, but is often sorted; examples include separation into a distinct core and marginal or (as here) meniscate structures. The geom etry of individual menisci can vary from shallow and saucer like to sharply pointed cones; successive menisci are often defined by alternations in their colour and/or composition.

'suspect microbial structures' (e.g., wrinkle structures (Figure 1), roll-up structures, and elephant skin texture) would be included; these are attributed to the presence, including plastic deformation, of microbially bound crusts or veneers on the top of the sediment column. An ichnofabric, all aspects of the sedimentary texture that result from biogenic reworking, includes any remnant sedimentary structures, trace fossils, or indistinct structures produced by macrofauna ('sediment churning', 'burrow mottling', and biodeformational structures) that cannot be accorded formal taxonomic status. The poor definition of the latter usually reflects their emplacement in sediments with a high pore water content and low sediment shear strength (soupgrounds); it may be compounded by the employment of a different burrowing strategy (Figure 2). Intrusive burrowing in uncompacted sediment involves simple deflection of the sediment around the body; the sediment collapses behind the body to occlude any void immediately (sediment swimming). In more compacted sediments (softgrounds and firmgrounds), compression burrowing and, especially, excavation (accompanied by either advection or backfilling) produce more complex, open, and actively infilled structures. Cryptobioturbation, resulting from the activities of meiofauna and microfauna, may obliterate any primary sedimentary depositional fabric and homogenize sediment without producing either discrete trace fossils or biodeformational structures.

Taxonomy, Preservation, and Ethology

Taxonomy

Trace fossils are classified using ichnogenus and ichnospecies; the prefix 'ichno' distinguishes these ichnotaxa from either living organisms or body fossils. Some higher rank categories (ichnofamilies), or informal groupings, have been established, but are not used widely. As a sedimentary structure, the detailed morphology of a trace fossil reflects both the producer's behaviour and the properties of the host sediment; characteristics of the latter (such as water content or grain size) can vary between beds, or even the laminae of a bed (Figures 3A and 3B). The abundance, probably surfeit, of ichnotaxa, particularly ichnospecies, in the geological literature has its origin in the middle to late nineteenth century, when many trace fossils were interpreted as body fossils; every subtle variation in morphology was therefore considered to be significant, and the structure worthy of separate classification. Further, and more easily resolved by comprehensive study, the morphology of a trace fossil *in partim*, and thus the ichnotaxobases (the features used to classify it), may vary depending on the orientation in which it is observed (Figures 3B and 3C).

Identifying the producer Trace fossils are very rarely preserved in association with their producer; if they are, each should be named separately. The identification of the causative organism is not a prerequisite for the naming of a trace fossil. One organism may produce more than one trace fossil depending on its behaviour (Figure 4A). A structure may also be produced by different coexisting organisms (Figure 4B), or modified later by a different organism. Potential producers may be constrained (usually at the

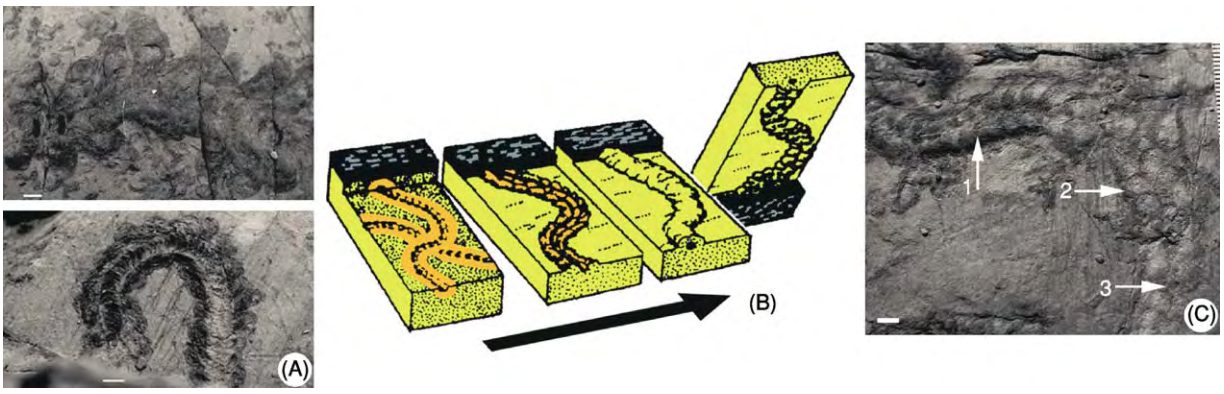


Figure 3 Key concepts in the study of trace fossils 1. (A) Preservational variants of the same ichnotaxon, reflecting emplacement in sediments of different consistency. (B) Different views of the same ichnotaxon reflect slight differences in the level of burrowing relative to the clay sand interface. From Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15.* Tulsa, OK: Society of Economic Paleontologists and Mineralogists. (C) A slight angular difference between the plane of bedding/weathering and that containing the trace fossil results in different two dimensional views of the trace fossil (1 3). Scale bars, 5 mm.

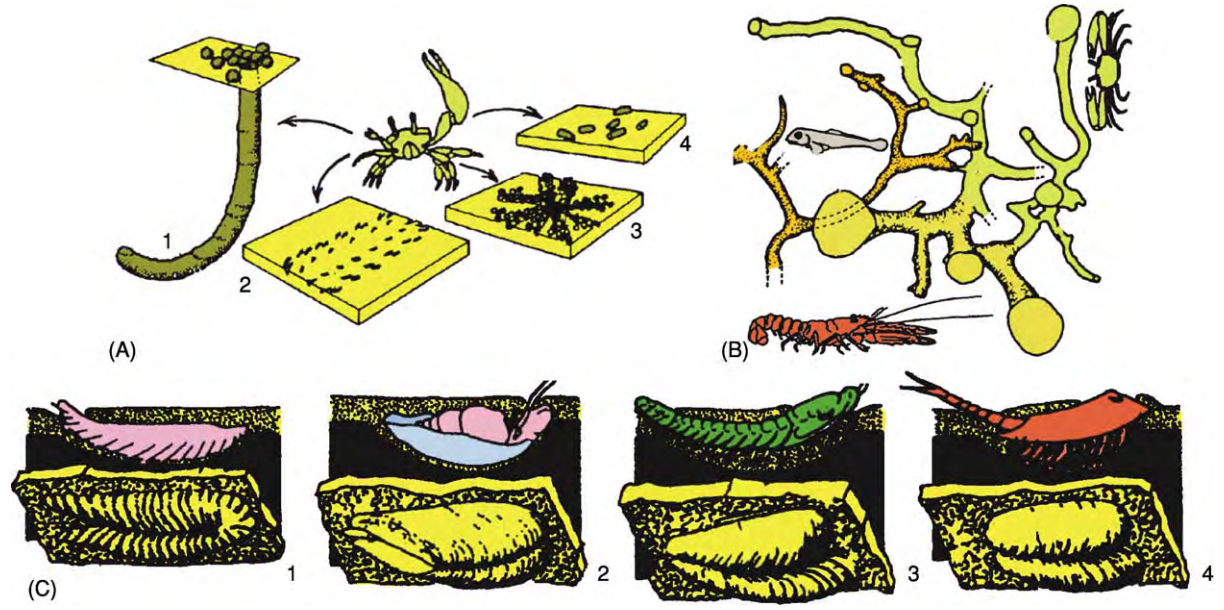


Figure 4 Key concepts in the study of trace fossils 2. (A) Open J shaped dwelling burrow and excavated pelleted sand (1), trackway (2), feeding structure (3), and faecal pellets (4) produced by the fiddler crab, *Uca*. (B) Plan view of open burrow network produced by a crab, lobster, and fish. (C) Examples of *Rusophycus* produced by a polychaete, snail, trilobite, and notostracan. From Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15.* Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

resolution of phylum or class) by the morphology and inferred function of the trace fossil (e.g., 'arthropod-produced trackway'). The body fossil content of the same, or surrounding, lithologies may be suggestive; for example, the breadth of a trackway may correspond to that of one, but not other, potential candidates. The potential producers of trace fossils, however, include organisms with minimal preservation potential. Furthermore, the same trace fossil may be produced by a variety of animals, precluding extrapolation of the producer's identity between case studies. The resting trace *Rusophycus*, examples of which in Palaeozoic marine sediments are often attributed to trilobites (*see* **Fossil Invertebrates:** Trilobites), also occurs in Mesozoic strata (after trilobites became extinct) and in sediments from (non-marine) environments that were never colonized by trilobites (Figure 4C).

Preservation

Exogenic trace fossils are emplaced at the sediment–water or sediment–air interface, and endogenic traces within the substrate; intergenic trace fossils are those emplaced endogenously at the interface between two beds (Figure 5A). Exogenic trace fossils have limited fossilization potential; they must be covered and thus cast, not eroded, during later deposition. Splitting at the original interface will yield both epirelief and hyporelief views of the structure. Loading of the substrate during emplacement of exogenic traces may depress underlying sediment; in finely laminated lithologies, splitting a short vertical distance below the original interface may reveal such undertracks, the resolution of which is often poorer than that of the corresponding exogenic expression; furthermore, only those parts of the trace fossil where loading was greater may be expressed as undertracks. Endogenic trace fossils observed in three dimensions are in full relief; those exposed on bedding-parallel surfaces

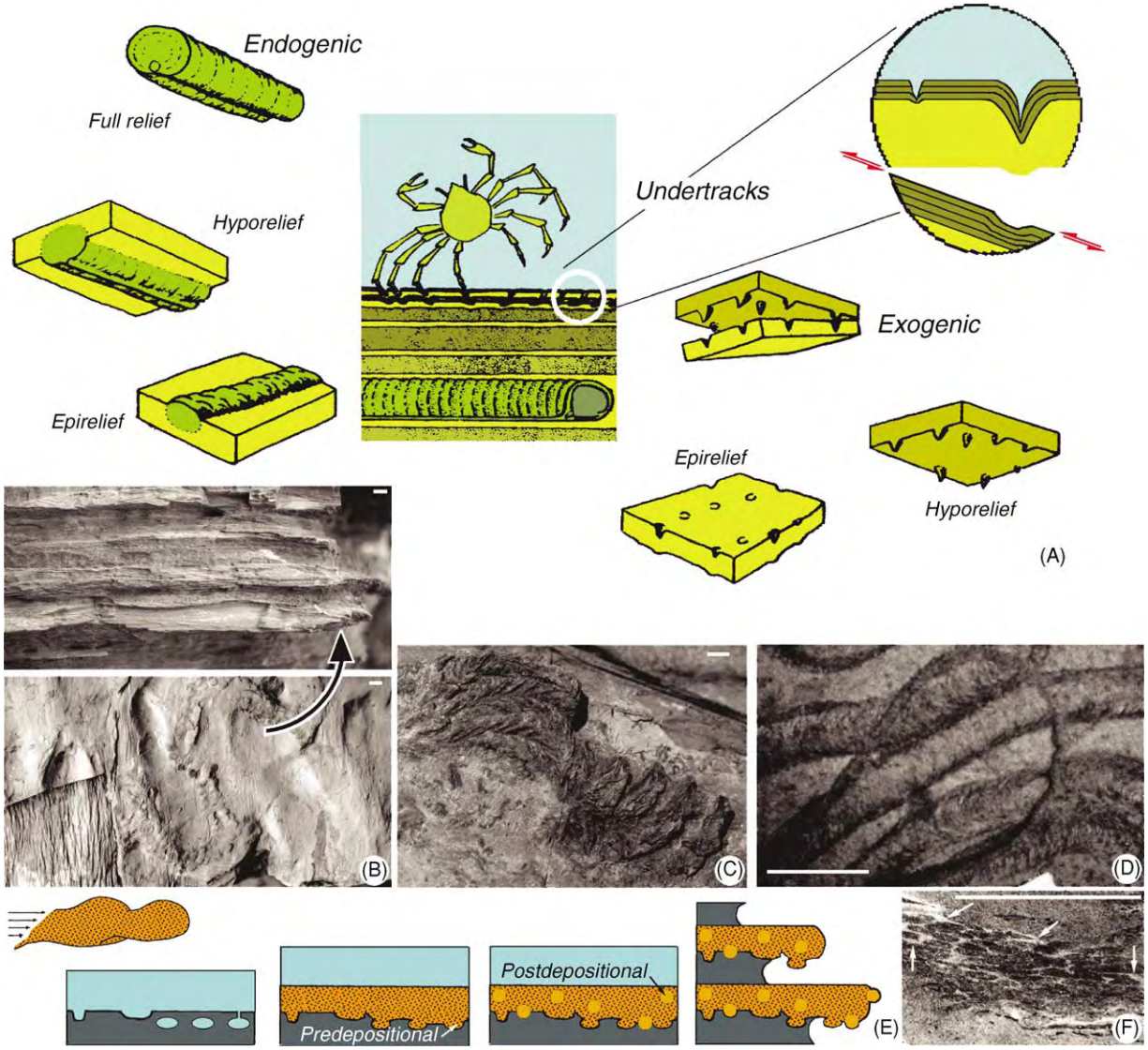


Figure 5 Key concepts in the study of trace fossils 3. (A) Preservational classification of trace fossils. Adapted from Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15.* Tulsa, OK: Society of Economic Paleontologists and Mineralogists. (B) The intergenic trace fossil is exposed in semirelief (positive hyporelief) on the sole of a bed of sandstone following preferential weathering of the underlying finer grained bed. (C) Positive epirelief view of *Lophoctenium* showing differential weathering of alternate menisci. (D) Polished horizontal cross section through an actively infilled burrow comprising a thick marginal wall structure and meniscate core. (E) Formation of predepositional and postdepositional trace fossils on the sole of an event bed. (F) Although the faecal pellets are the same colour and sourced from the host lithology, the light coloured sediment in the interstices between them indicates that the trace fossil is secondary. Scale bars, 5 mm.

are in semirelief; hyporelief and epirelief are used when the lower or upper surface, respectively, is that exposed. The orientation of specimens in the field should be determined at the time of their collection using independent criteria provided by sedimentary structures. The prefixes positive and negative describe the relief of the trace fossil relative to the host lithology. Full relief, and particularly semirelief, views can result from planes of splitting deflecting to follow the external surface of, rather than continuing across, a trace fossil. Semirelief views of intergenic structures are most common when successive lithologies have a different resistance to weathering (Figure 5B). Parts of an individual structure may weather differently; in the example in Figure 5C, alternate chevron-shaped menisci are either more sand-rich or mud-rich than, and thus have weathered positive and negative with respect to, the host lithology. Localized differences in the sedimentary fabric that result from bioturbation can be exacerbated during diagenesis. Trace fossils are often sites for the precipitation of early diagenetic minerals; this may be encouraged by the presence of mucus secreted during burrowing as an aid to locomotion, excretion, or to maintain an open structure. This can impact on economically important variables such as sediment texture, organic content, porosity, and permeability.

An open structure connected to either the air or water column may be later infilled gravitationally (passive infill) by sediment of a different composition and/or grain size. Active infill of a structure is by its progenitor. Modifications resulting from the reworking of sediment, for example during its ingestion and passage through the gut of an animal, include (1) excretion as pellets or a faecal string; (2) the exclusion or selection of grains by shape, size, or type; and (3) the reorientation of grains to produce a localized sedimentary fabric (Figure 5D).

Predepositional and postdepositional trace fossils Postdepositional trace fossils are emplaced after the deposition of a bed; the term 'predepositional' is used for trace fossils, now preserved on the sole of a bed, that were emplaced on the surface of, or within, the underlying bed. Thus, while an exogenic trackway in epirelief view is postdepositional, the secondary cast produced during deposition of the next bed, and exposed subsequently in hyporelief on the sole thereof, is predepositional. For exogenic trace fossils, the process requires there to be no erosion of the existing sediment. More extensive erosion, however, may expose endogenic open burrow systems, or wash away the active infill of burrows whilst preserving their outlines; infill of the voids created generates predepositional secondary casts (Figure 5E). The critical conditions, erosion closely followed by deposition, are often met during deposition of event beds, such as turbidites or storm deposits.

Postdepositional intergenic burrows may be emplaced later onto a surface containing predepositional secondary casts (Figure 5E). As well as the difference in their relative age, each suite may reflect different environmental conditions. Only the lower surface of a predepositional endogenic burrow system will be preserved; its infill will comprise the host lithology. Active infill of a postdepositional burrow may include modification of the host lithology (see above); in vertical cross-section, the outline of the burrow (whether infilled actively or, later, passively) will be complete.

Primary and secondary trace fossils Primary trace fossils represent either reworking of the host lithology *in situ* (but not necessarily at the time of its deposition), or the piping downwards of sediment of identical composition from one layer into another. Their identification relies on the properties of the host lithology being altered in the process (see above). Secondary trace fossils originate when sediment with different properties is 'piped' (usually downwards) into the host lithology; the evidence can be subtle (see example in Figure 5F).

Ethology

A behavioural, or ethological, classification of trace fossils is presented in Figure 6. Few trace fossils represent just one activity, and classification is thus on the basis of what is considered to be the dominant, most significant, behaviour. Cubichnia, temporary resting traces, are usually shallow, exogenic excavations (Figure 7A). Repichnia, locomotion traces, include both endogenic continuous burrows and exogenic structures. The latter may be continuous (trails), or comprise a series of discrete sequential footfalls (tracks or imprints) made by an appendage or limb (trackways) (Figure 7B). Pascichnia combine continuous locomotion parallel to bedding with feeding; emphasis is on the systematic extraction of food resources within an area (cf., strip mining). This is achieved by phobotactic (the avoidance of crossing previously formed parts of the structure), combined with strophotactic (episodic or periodic 180° turns) and/or thigmotactic (tendency to keep close to a previously formed part of the structure), behaviour patterns (Figures 7C and 7D). Trails in negative epirelief with similar geometries on the surfaces of modern ocean floors are exogenic structures with limited fossilization potential; at least, the vast majority of fossil pascichnia were actively filled endogenic burrow systems. Agrichnia are endogenic burrow

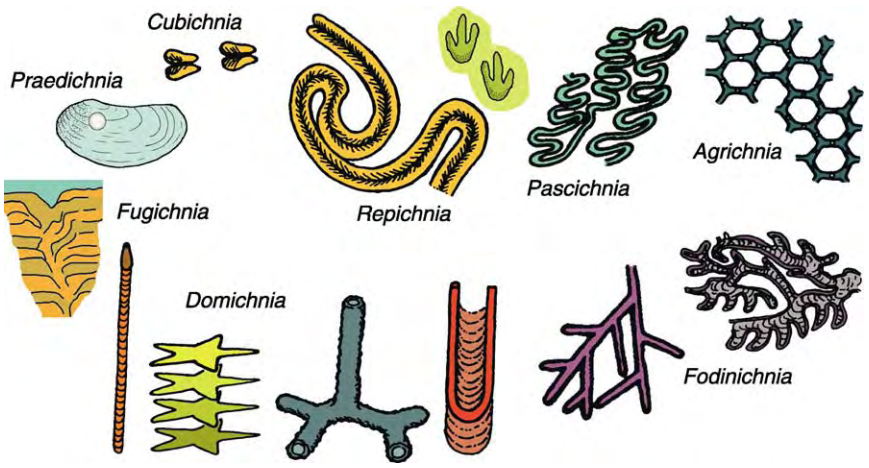


Figure 6 An ethological classification of trace fossils. Adapted from Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15.* Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

systems that were maintained as open structures, and are therefore almost always preserved in positive hyporelief as predepositional secondary casts (Figures 5E and 7E). Their characteristic, complex geometries in plan view include fish scale-like, polygonal, and planispiral patterns. Spiralling patterns maintain a constant spacing between successive whorls, i.e., they do not exhibit thigmotactic behaviour. Many agrichnia are essentially two-dimensional, contained within a single plane parallel to the sediment–water interface; three-dimensional structures comprise several such levels, separated from each other vertically and connected by a more steeply inclined continuous open burrow. Agrichnia possessed single or multiple vertical connections to the overlying water column. As with domichnia (see below), the introduction of oxygenated seawater to depth within the sediment could have reduced the idealized vertical stratification of electron acceptors to a patchwork of biogeochemical microenvironments (Figure 8). Their exact function is unknown, but agrichnia may represent traps or 'microbial gardens'; microbes may be cultured on, and harvested from, the walls. Fodinichnia combine semistationary behaviour with deposit feeding. Geometries are highly variable and include two-dimensional, bedding-parallel, and three-dimensional forms. The key morphological element is the derivation of a series of branching, non-interpenetrating, straight or curved shafts, or spreiten-infilled lobes, from a common source (Figures 7F–7H). Domichnia are open burrow systems, or borings into xylic and lithic substrates (Figures 7I–7L), utilized on a semipermanent basis; most are for habitation, but structures specifically constructed for other functions (e.g., brood chambers) have been identified; some examples of the fossilized cases of caddis larvae contain pupae (Figure 7K). Domichnia in soft sediments often exhibit passive infill from the overlying layer and a thick marginal lining (to prevent collapse of the open burrow system); examples include the pelleted walls typical of *Ophiomorpha*. A domichnion in a firmground may have a bioglyph on its internal surface (Figure 7L). A 'spreite' (plural spreiten) is produced by shifting the position of a U-shaped vertical burrow in order to maintain its base a constant distance below the sediment–water interface. The spreite, which marks the former position of the basal part of the burrow, may be protrusive (occurring inside the limbs of the open burrow system in response to sediment being removed; Figure 7I), and/or retrusive (occurring below the active burrow in response to sediment aggradation). Equilibrichnia, formed by the regular, incremental shift upwards of burrows in response to the semicontinuous accretion of sediment, are considered here as a variety of domichnion. Rapid upward (attempted) escape, for example following burial by sediment, results in poorly structured escape traces, fugichnia (fugichnion). Unequivocal examples of praedichnia, trace fossils indicating predation, include the holes drilled in the shells of other molluscs by some gastropods by mechanical and/or chemical means (Figure 7M); the acid secreted by muricid gastropods can have a pH as low as 3.8. More equivocal examples in soft substrates include the intersection of two trace fossils, following which only one continues.

This classification scheme works well for the majority of trace fossils, although it has its limitations. Categories can grade into each other; for example,

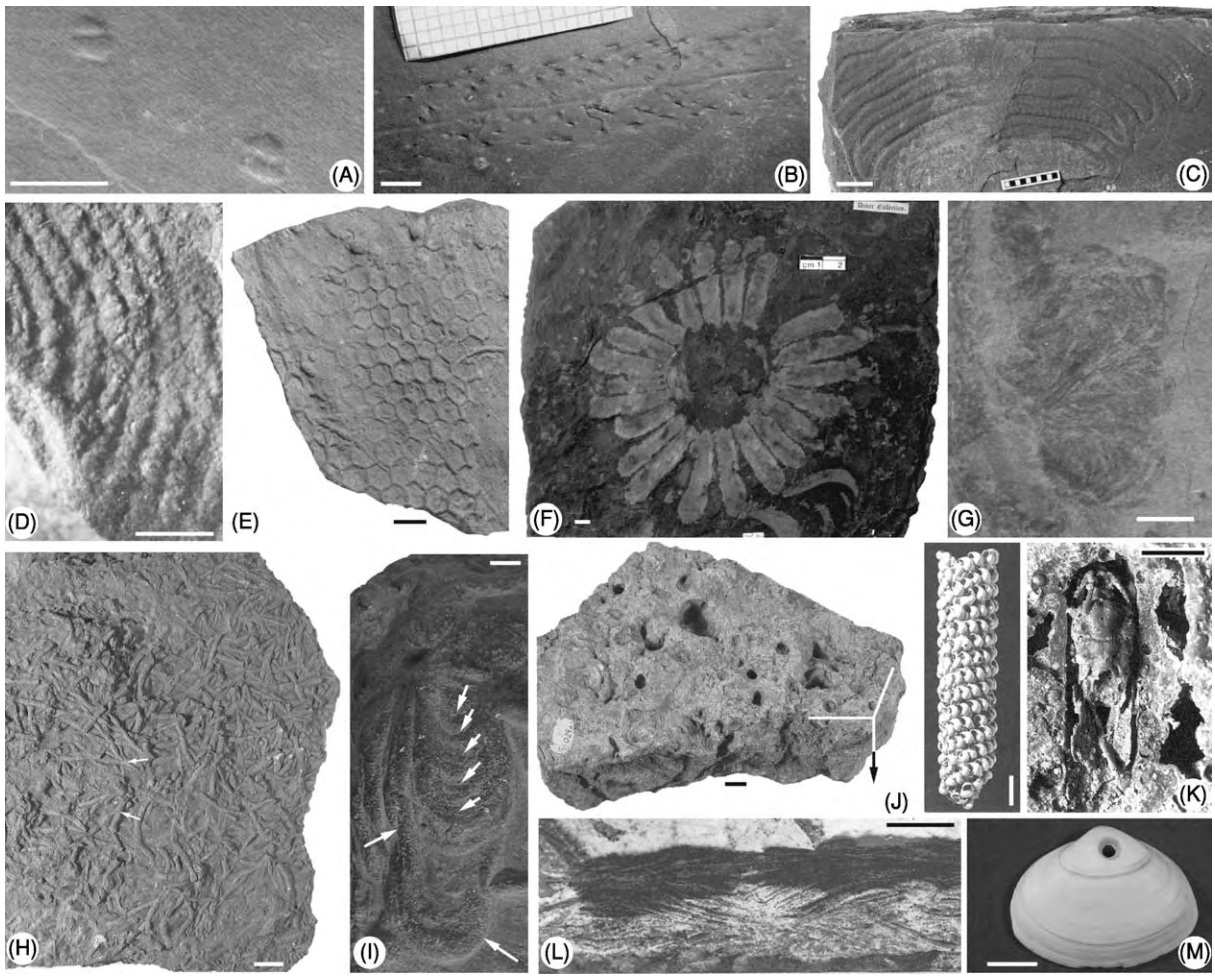


Figure 7 Variation in behaviour represented by trace fossils. (A) The cubichnion, *Rusophycus*. (B) An arthropod produced repich nion. (C) The pascichnion, *Nereites*, exhibiting strophotactic and thigmotactic behaviour. (D) The consistent sense of braiding between successive burrows indicates thigmotactic behaviour achieved by spiralling alone. (E) The agrichnion, *Paleodictyon*. (F) Fodinichnion comprising unbranched shafts radiating parallel to bedding from a central area. (G) Fodinichnion comprising a series of lobate spreiten. (H) High density, monospecific occurrence of the fodinichnion, *Chondrites*; each burrow system comprises a series of branching shafts (at arrows) fanning outwards and downwards from a central shaft. (I) Protrusive spreite (short arrows) contained inside a U shaped burrow (long arrows). (J) Oblique view of block of shallow marine oolitic limestone that was later bored and encrusted during emplacement of a hardground community. (K) External view and rare example of a pupa preserved inside the case constructed by a caddis fly larva. From Hugueney M, Tachet H, and Escuillié F *Caddisfly Pupae* from the Miocene indusial Limestone of saint Gerand le Puy, France (1990) *Palaeontology* 33: 495 502, Plate 1, Figures 3 and 4. (L) Bioglyph on the internal surface of a burrow excavated in semilithified sediments. Image courtesy of Richard Bromley, Copenhagen. (M) Praedichnion: a hole drilled by a gastropod through the shell of a bivalve. Scale bars, 5 mm.

at what point is a meandering pattern sufficiently regular to warrant the descriptor pascichnion, not repichnion? Opinions may vary between authors as to what the dominant behaviour is; the burrow *Planolites* has been considered a fodinichnion, pascichnion, and repichnion, as it involves reworking of the sediment whilst on the move in a straight to sinuous curve. Finally, each ichnospecies within an ichnogenus need not have the same ethology; more rarely, the ethology may vary (e.g., between a repichnion and a pascichnion) amongst a set of specimens of an ichnospecies.

Use as Environmental Indicators

Information from trace fossils and ichnofabrics can be incorporated into palaeoenvironmental reconstructions at scales ranging from the individual bed (reconstruction of a single endobenthic community) to depositional sequences tens to hundreds of metres thick (e.g., ichnofacies).

Ichnofacies

Ichnofacies are recurrent combinations of sedimentary facies and trace fossils; the Skolithos, Cruziana,

Zoophycos, and Nereites ichnofacies, emplaced within marine softgrounds, characterize successively greater water depths (Figures 9 and 10). The use of trace fossils as a palaeobathymetric indicator exploits the fact that many, particularly ichnogenera, have a long time range, but are restricted to, or most common in, specific environmental conditions. These conditions include wave or current energy, temperature, chemistry (including salinity and quantity of dissolved gases such as oxygen), light penetration, nutrient supply, competition for ecospace and resources, sedimentation rate, and substrate character (including the grain size and geotechnical properties of soft sediments). Changes in these conditions tend to correlate with changes in absolute water depth, and thus the palaeobathymetry is depth related rather than depth controlled. Environmental conditions are, however, far from uniform at any given water depth. Local factors may control the distribution of ichnofacies; for example, depositional conditions in the proximal or channellized parts of a submarine fan may resemble the high-energy, mobile substrates typical of nearshore environments, and the Skolithos ichnofacies occur in each. A change in environmental conditions can produce a succession of ichnofacies that mimics, but is not the product of, significant changes in water depth; progradation of the submarine fan in Figure 9A would result in lithologies with a Skolithos ichnofacies succeeding those with a Nereites ichnofacies, unaccompanied by any significant decrease in water depth. The Glossifungites, Trypanites, and Teredolites ichnofacies are emplaced into firmgrounds, lithic substrates, and xylic substrates, respectively (Figure 10); water depth is not a controlling factor (Figure 9B). Some authors recognize the Psilonichnus (between the foreshore zone and the terrestrial realm) and Arenicolites (opportunistic colonization of newly deposited event beds) ichnofacies. Originally identified as the Scoyenia ichnofacies, the heterogeneity of non-marine environments does not lend itself to classification; several alternative detailed subdivisions have been proposed, but no consensus has emerged.

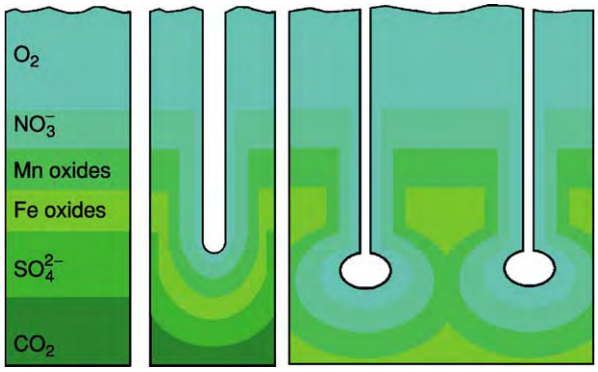


Figure 8 Modification of the idealized vertical stratification of electron acceptors via the emplacement of an open burrow structure at depth.

The presence or absence of an individual ichnogenus, even that after which the ichnofacies is named, is not strong evidence for a particular water depth. Furthermore, the bathymetric ranges of some ichnotaxa are known to have changed over

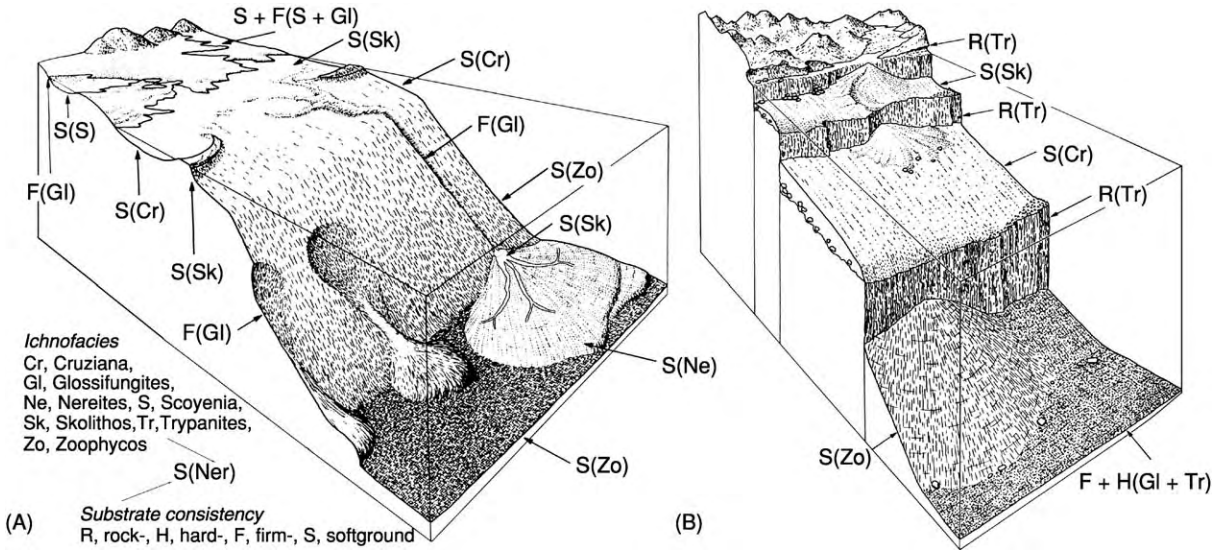
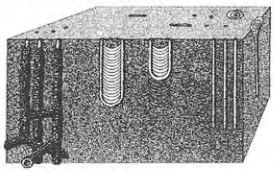


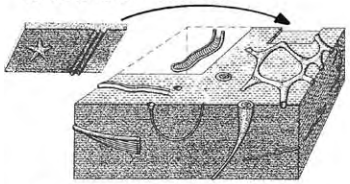
Figure 9 Schematic representation of the distribution of ichnofacies in marine environments. (A) Passive continental margin. (B) Sediment starved active continental margin. Reprinted from Bromley RG and Asgaard U (1991) Ichnofacies: a mixture of taphofacies and biofacies. *Lethaia* 24: 153 163 (www.tandf.no/leth), by permission of Taylor and Francis AS.

Skolithos ichnofacies



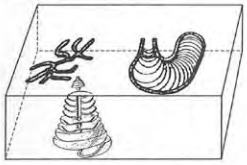
- *Typical environments:* moderate to relatively high energy conditions; muddy to clean, well-sorted, shifting substrates, subject to abrupt deposition and erosion.
- *Trace fossil content:* vertical, cylindrical, or U-shaped domichnia; latter may have protrusive and/or retrusive spreiten; also fugichnia.

Cruziana ichnofacies



- *Typical environments:* infralittoral to shallow circalittoral substrates; below daily wave base, but not storm wave base to offshore shelf; moderate to low energy; well-sorted to muddy silts and sands.
- *Trace fossil content:* varied; bedding-parallel epigenic and endogenic repichnia. Domichnia often with protrusive spreiten. Fodinichnia varied. Cubichnia common.

Zoophycos ichnofacies

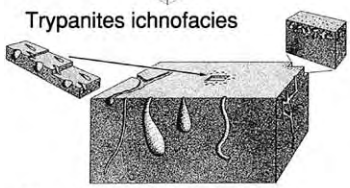


- *Typical environments:* circumlittoral to littoral; quiet-water, ?reduced oxygen; firm organic-rich sands and muds; sediment accretion continuous.
- *Trace fossil content:* simple to complex fodinichnia often with planar, inclined, or helicoidal spreiten; deposit feeding predominates; reduced diversity may reflect poorly oxygenated waters.

Nereites ichnofacies

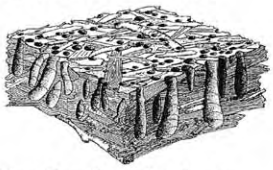
- *Typical environments:* bathyal to abyssal, quiet, oxygenated waters; episodic turbiditic deposition or continually accreting pelagic surfaces.
- *Trace fossil content:* bedding-parallel pascichnia and agrichnia dominate in distal parts of turbiditic fans; their distribution can be mutually exclusive reflecting partitioning into specific environmental niches. Proximal parts of turbidite fans often with vertical, cylindrical, or U-shaped domichnia; pelagic and hemipelagic sequences including inter-trubidite deposits usually dominated by fodinichnia and actively infilled repichnia.

Trypanites ichnofacies



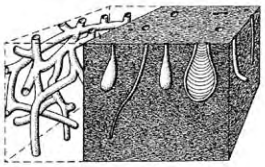
- *Typical environments:* lithic substrates, notably littoral omission substrates; also organic (shell, bone) substrates.
- *Trace fossil content:* cylindrical or vase-shaped endolithic domichnia; shallow cubichnia produced by epifauna; praedichnia in biomineralized tissues, e.g., shells.

Teredolites ichnofacies



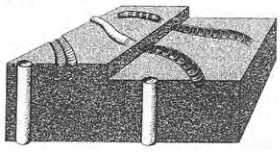
- *Typical environments:* xylic substrates, including forested marginal marine settings (mangrove swamps); high-density occurrence of allochthonous bored wood can be associated with flooding surfaces.
- *Trace fossil content:* lined or unlined domichnia, diversity often low or monospecific.

Glossifungites ichnofacies



- *Typical environments:* firmgrounds, often marine littoral and sublittoral omission surface; suitable substrates can also occur following incision into partially consolidated sediments.
- *Trace fossil content:* varied domichnia, including vertical, cylindrical, tear-shaped, or U-shaped burrows, and branching and gallery-like open networks.

Scoyenia ichnofacies



- *Typical environments:* terrestrial environments are extremely varied, but preservation potential highest in softgrounds and firmgrounds at margins of aquatic environments. Subdivisions or alternatives are available for specific environmental settings.
- *Trace fossil content:* bedding-parallel, meniscate burrows; vertical or gallery-like domichnia; invertebrate, particularly arthropod-produced, cubichnia and repichnia; vertebrate-produced repichnia.

Figure 10 Summary of the ethologies, lithologies, and sedimentary processes characteristic of each of the main ichnofacies. After Frey RW and Pemberton SG (1984) Trace fossil facies models. In: Walker RG (ed.) *Facies Models*, 2nd edn., pp. 189 207. Ontario: Geological Association of Canada.

time. During most of the Palaeozoic, the ichnogenus *Zoophycos* occupied a broad range of marine water depths; since the Early Permian, it has become progressively restricted to greater water depths and is only found in continental slope and deep basin settings today.

Use of Infaunal Ecospace: Endobenthic Tiering

Bioturbation in modern, fine-grained substrates undergoing accretion that is the work of a single community can be divided into three general levels: the surficial mixed and underlying transition layers in which bioturbation occurs, and the lowest historical layer which contains the ichnofabric preserved after diagenesis and lithification (Figure 11A). As sediment properties change progressively with depth, the boundary between the mixed and transition layers is gradational, not abrupt; it is convenient, however, to model the two as distinct layers. Mixed layer sediments are often soupgrounds and thus, although bioturbated completely, the trace fossils are often poorly preserved. Bioturbation in the transition layer is heterogeneous and occurs as discrete burrows. These are usually well defined because of the high shear strength of these more dewatered sediments (softgrounds or even firmgrounds); other progressive changes with depth include reductions in both the volume and degree of oxygenation of the interstitial pore waters. Vertical partitioning, tiering, of the transition layer infauna (and thus the trace fossils they emplace) occurs in response to such changes in the physical and chemical properties of the sediments (Figure 11A). Reconstruction of the tiering profile therefore provides a measure of the community complexity; this can include the depth to which sediments were bioturbated and thus the volume of ecospace exploited. As sediment accretes, organisms will move upwards to maintain the same level or ecological niche; trace fossils produced at shallower depths will thus be cross-cut by those emplaced at successively greater depths. Dominant to unilateral cross-cutting of one trace fossil by another implies that the latter belonged to a deeper tier; the mutual intersection of two or more trace fossils implies that they are

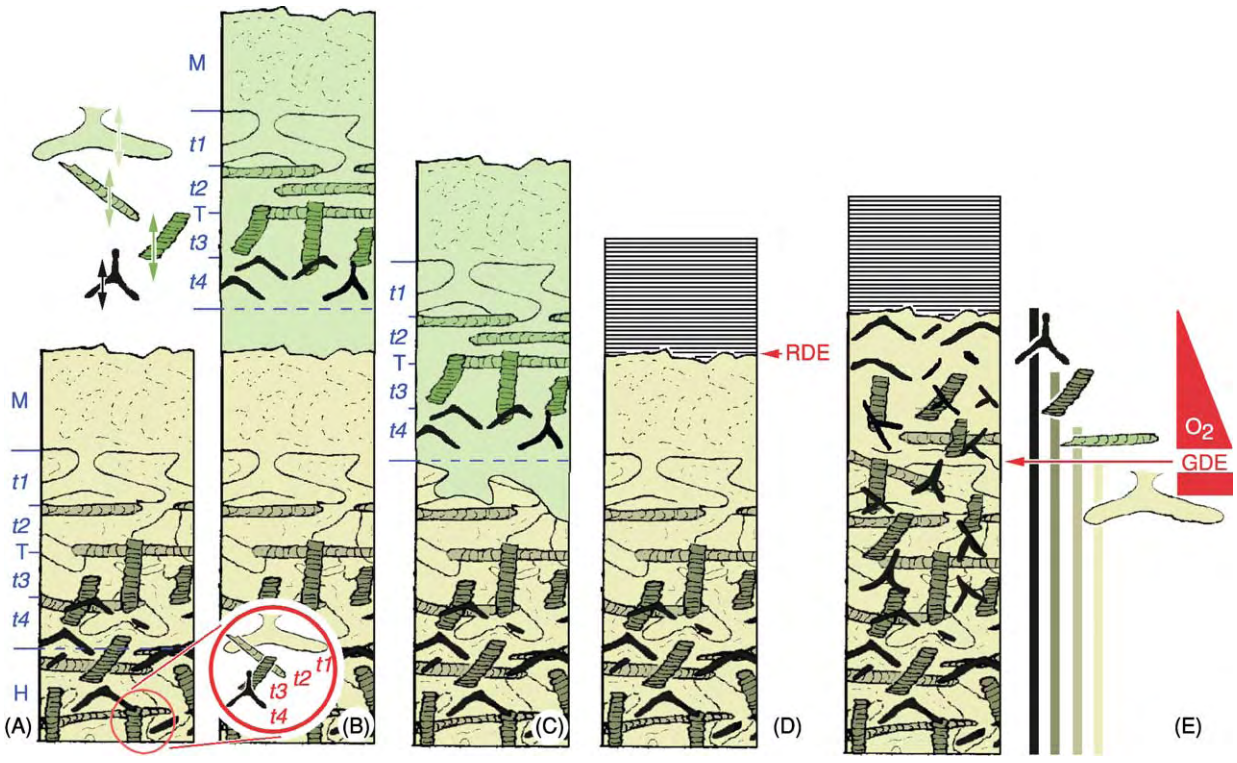


Figure 11 Endobenthic tiering revealed by the ichnofabrics produced 1. (A) Subdivision of the sediment column into the mixed (M), transition (T), and historical (H) layers; inset shows dominant to unilateral cross cutting relationships amongst the transition layer trace fossils allowing the identification of four tiers (*t1 4*). (B, C) Rapid relocation of the infauna following event bed deposition leaves a 'frozen profile' in the older sediments, all of which (B) or, if the event bed is erosively based, only the lower part of which (C), may be preserved. (D) Production of a 'frozen profile' resulting from the evacuation of the sediment column by infauna following a rapid deoxygenation event (RDE). (E) A gradual deoxygenation event (GDE), in which the oxygen content of the interstitial pore waters declines gradually, results in exclusion of the infauna of successively deeper tiers. Adapted from Savrda CE and Bottjer DJ (1986) Trace fossil model for reconstruction of paleo oxygenation in bottom waters. *Geology* 14: 3 6.

components of the same tier. In practice, deviations from the random cross-cutting of older structures by younger can occur; these include preferential re-exploitation as well as avoidance of earlier formed structures. Complete reworking of sediment at depth will obliterate the record of earlier activity in shallower tiers, and thus reduce the diversity of the trace fossil community (the palaeoichnocoenosis). The preservation of a complete tiering profile requires the intensity of bioturbation to decline with depth; earlier formed parts of the ichnofabric, including occasionally the mixed layer ichnofabrics, occur as relict patches between later structures.

The deposition of an event bed will result in a rapid upward relocation of the infauna; if this event bed is sufficiently thick, the base of the transition layer will be moved above the older sediments, leaving a frozen tiering profile preserved within them; the distance from the sediment surface to the base of the transition layer indicates the thickness of sediment occupied (**Figure 11B**). The upper part of the sediment column is often a soupground, and thus remobilized relatively easily; the upper parts of the frozen tiering profile, notably the mixed layer, may be eroded during the deposition of the next bed (**Figure 11C**). Evacuation of the sediment column following a rapid deoxygenation event, in which the redox threshold boundary is moved above the sediment–water interface, will leave a frozen profile preserved below the succeeding sediments (**Figure 11D**). Mixed layer ichnofabrics, rare in the geological record, are more likely to be preserved in this manner than by burial below an event bed.

Other means of reconstructing the tiering profile include the identification of the depth to which an intergenic postdepositional trace fossil penetrated the substrate. In a sequence of event beds of different thickness, the components of shallower tiers occur only on the soles of thinner beds (**Figure 12A**). Secondary trace fossils (see above) occur within a 'piped zone' (**Figures 12B and 12C**). The thickness of the piped zone indicates the volume of ecospace used by the endobenthic community; components of progressively deeper tiers occur closer to its base. In the simple model in **Figure 12B**, the secondary trace fossils were sourced from the immediately overlying layer whilst it was being deposited, and the piped zone is contained entirely within one layer. The example in **Figure 12C** shows a more complex situation; the lowest bed was rebioturbated during deposition of the second, not the first, overlying layer, but the latter was too thin to contain the piped zone entirely. In this idealized example, the colour of each layer is different, and the source of sediment in the piped zone is obvious. In practice, however, many hemipelagic

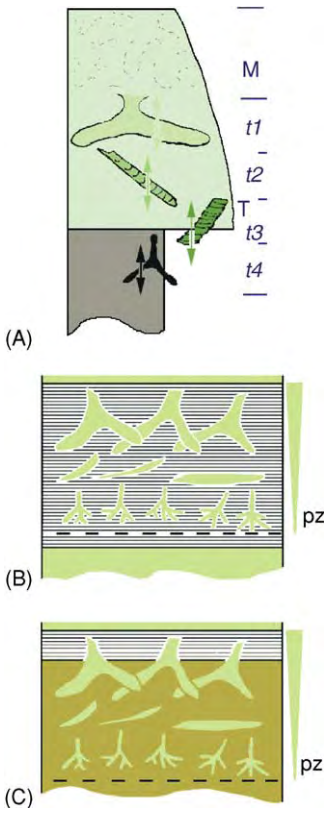


Figure 12 Endobenthic tiering revealed by the ichnofabrics produced 2. (A) The base of the event bed cannot be reached by the infauna of the shallower tiers; the trace fossils at its sole will comprise those of tier 3 and, if the accretion of sedi ment subsequently is gradual, tier 4. (B, C) Emplacement of secondary trace fossils defines a piped zone (pz), the thick ness of which is a measure of the volume of ecospace utilized. For clarity, earlier formed parts of the ichnofabrics have been omitted.

and pelagic sequences are characterized by alternating lithologies, the differences in colour and composition of which can be the result of orbital forcing. In such cases, trace fossils within the piped zone may be actively infilled by the same sediment as the host lithology.

Natural systems are obviously more complex, but more dynamic models can be produced by allowing variables, such as the rate of sediment accretion, including negative values (erosion), or the levels of oxygenation, to fluctuate over time. Not all circumstances will satisfy the assumptions within the model. Thin-bedded siliciclastic turbidites often exhibit a bipartite division into a sand-rich lower part and a mud-rich upper part, and infauna can position themselves with respect to the interface between the two, independent of its depth below the sediment–water interface. Cross-cutting relationships can also be generated by the superimposition of two successive

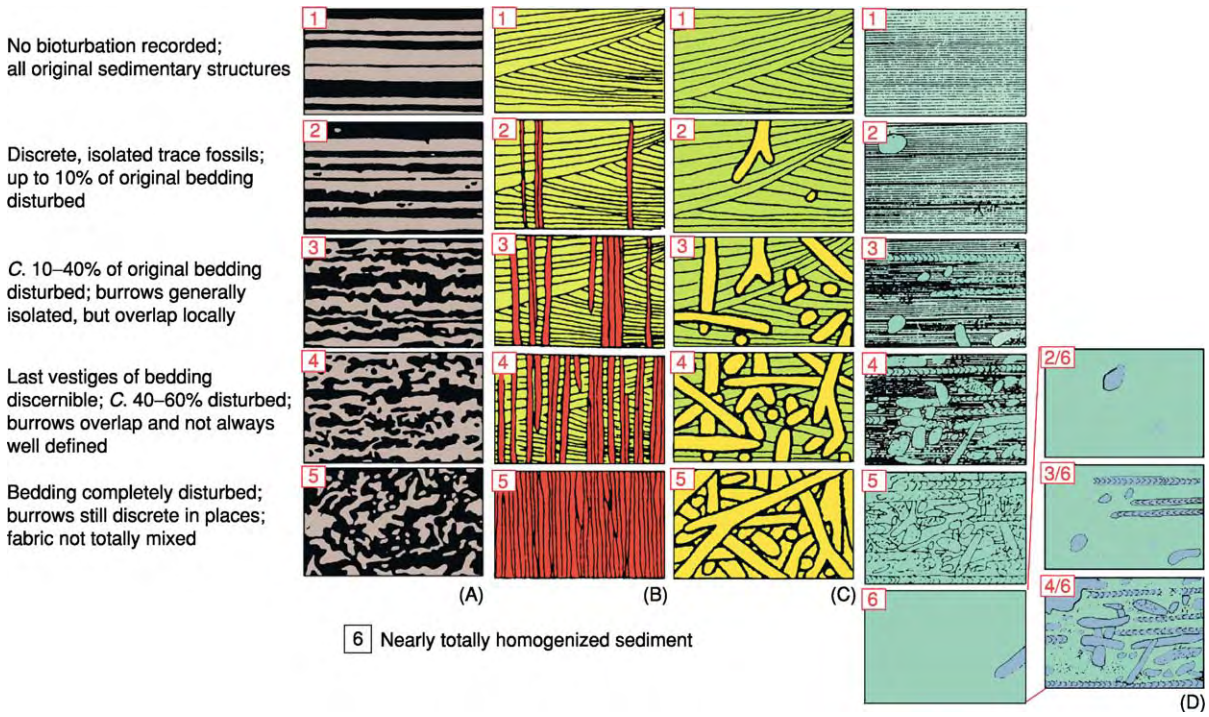


Figure 13 Definitions and schematic illustrations of ichnofabric indices for strata deposited in: (A) shelf environments; (B) high energy nearshore sandy environments dominated by *Skolithos*; (C) high energy nearshore sandy environments dominated by *Ophiomorpha*; (D) deep sea deposits, including examples of homogenized sediment that has been reworked subsequently. Reprinted with permission from Droser ML and Bottjer DJ (1993) Trends and patterns of Phanerozoic ichnofacies. *Annual Review of Earth and Planetary Sciences* 21: 205 225. © 1993 by Annual Reviews www.annualreviews.org

communities; following event bed deposition, it is not unusual for a short-lived opportunistic community to exploit the new ecospace and resources, before the longer term equilibrium community is re-established.

Palaeo-oxygenation

As bottom water oxygenation declines, the general tendency is for the thickness of the mixed and transition layers, and the size (reflected in the burrow diameter) and diversity of actively infilled transition layer trace fossils, to decrease (Figure 11E). It has been suggested that ichnofaunal assemblages dominated by domichnia, pascichnia, and fodinichnia indicate a progressive decline in pore water oxygenation. Related to this, certain ichnogenera, notably the fodinichnion *Chondrites*, especially if abundant in a low-diversity or monospecific assemblage, have been considered to be diagnostic of low oxygen conditions; however, this should not be assumed without supporting ichnological and sedimentological evidence. The shaft diameter is relatively large in the monospecific assemblage of *Chondrites* emplaced in a storm bed in Figure 7H, and an interpretation as an opportunistic colonization of newly deposited sediment is favoured.

Intensity of Bioturbation

The ichnofabric index (ii) is a semiquantitative measure of the intensity of bioturbation, based on the extent to which the original stratification is disrupted; as both the nature of bioturbation and host lithology will vary, the schematic illustrations are specific to certain environmental settings (Figure 13). The ichnofabric is observed in vertical sections; a standard horizontal field of view should be used and reported (500 mm is often used in outcrop studies).

See Also

Biosediments and Biofilms. **Diagenesis, Overview**. **Fossil Invertebrates:** Trilobites. **Palaeoecology**. **Palaeontology**. **Sedimentary Environments:** Depositional Systems and Facies; Storms and Storm Deposits.

Further Reading

Bottjer DJ, Hagadorn JW, and Dornbos SQ (2000) The Cambrian Substrate Revolution. *GSA Today* 10: 1 7.

Bromley RG (1996) *Trace Fossils. Biology, Taphonomy and Applications*. London: Chapman and Hall.

Bromley RG and Asgaard U (1991) Ichnofacies: a mixture of taphofacies and biofacies. *Lethaia* 24: 153 163.

Donovan SK (ed.) (1994) *The Palaeobiology of Trace Fossils*. Chichester: Wiley.

Droser ML and Bottjer DJ (1993) Trends and patterns of Phanerozoic ichnofacies. *Annual Review of Earth and Planetary Sciences* 21: 205 225.

Ekdale AA, Bromley RG, Pemberton SG (1984) *Ichnol ogy. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineral ogists.

Frey RW and Pemberton SG (1984) Trace fossil facies models. In: Walker RG (ed.) *Facies Models,* 2nd edn., pp. 189 207. Ontario: Geological Association of Canada.

Frey RW, Pemberton SG, and Saunders TDA (1990) Ichno facies and bathymetry: a passive relationship. *Journal of Paleontology* 64: 155 158.

Lockley M (1991) *Tracking Dinosaurs*. Cambridge: Cambridge University Press.

Maples CG and West RR (eds.) (1992) *Trace Fossils. Short Courses in Paleontology No. 5*. Tulsa, OK: Paleonto logical Society.

Pemberton SG (1992) (ed.) *Applications of Ichnology to Petroleum Exploration A Core Workshop. SEPM Core Workshops No. 17*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Savrda CE (1995) Ichnologic applications in paleocea nographic, paleoclimatic and sea level studies. *Palaios* 10: 565 577.

Savrda CE and Bottjer DJ (1986) Trace fossil model for reconstruction of paleo oxygenation in bottom waters. *Geology* 14: 3 6.

Taylor A, Goldring R, and Gowland S (2003) Analysis and application of ichnofabrics. *Earth Science Reviews* 60: 227 259.

Wetzel A and Aigner T (1986) Stratigraphic completeness; tiered trace fossils provide a measuring stick. *Geology* 14: 234 237.

ULTRA HIGH PRESSURE METAMORPHISM

H-J Massonne, Universität Stuttgart, Stuttgart, Germany

Introduction

Ultra high pressure (UHP) metamorphic rocks of common basic to felsic nature are defined by the occurrence of coesite, a silica polymorph that is denser than quartz. According to several experimental studies, the transition from quartz to coesite at 600°C requires a pressure (P) of around 27 kbar, a temperature (T) of conditions that occur on Earth at depths close to 100 km. Coesite in nature was detected first in rocks affected by impact metamorphism, but a coesite-bearing rock that had been subjected to regional metamorphism was described in 1983 by Chopin. Based on the coesite–quartz transition pressure and temperature curve, this rock, which was from the Dora Maira Massif of the Western Alps, must have been buried at depths of about 100 km or more. Transition curves of SiO_2 polymorphs show a moderate pressure (P) and temperature (T) slope, dP/dT, of only 10 bar °C^{-1} (Figure 1).

Prior to discovery of the Dora Maira Massif rock, it was thought that metamorphic rocks (excepting ultrabasites) that were formed during orogenic events, and now exposed at Earth's surface, represent, in general, a fossil record of pressure and temperature conditions only of Earth's crust, equivalent to maximum depths of 70 km and pressures up to ~18 kbar. Eclogites (basalts that have been metamorphosed under high pressure) were believed to have formed, depending on temperature, at pressures between 12 and 16 kbar, corresponding to the jadeite content of omphacite, the mineral that characterizes eclogites. However, these estimates are only justified when omphacite co-exists with plagioclase and quartz (but plagioclase is often only a retrograde product in eclogites, due to the breakdown of omphacite). Rare jadeite occurrences in felsic rocks with plagioclase supported the view of metamorphic pressures not exceeding 18 kbar for crustal rocks. In geodynamic models of subduction of oceanic crust under oceanic or continental plates, it was assumed that the return of subducted material was possible only for shallower regimes of the collision wedge. Such subducted material then became, for instance, part of an accretionary wedge complex. Within the framework of such a scenario, the discovery of coesite in regional metamorphic rocks was sensational and a real turning point in scientific thinking about deep burial and subsequent exhumation of crustal rocks. Soon after Chopin's report in 1983, it became evident that UHP rocks are more widespread than had been thought. Coesite was recognized in rocks of the Norwegian Caledonides, the Dabie Shan in China, and orogenic regions elsewhere. Moreover, in 1990, microdiamonds, another indicator mineral for UHP metamorphism, were reported by Sobolev and Shatsky in marbles and gneisses from the Kokchetav Massif, Kazakhstan. These diamonds provided evidence for burial of crustal rocks to depths of at least 130 km.

Identification of UHP Rocks

The discovery of UHP rocks at Earth's surface at a relatively late date in geological science history can be attributed to the nature of the processes involved. Retrogression of rocks during exhumation can result

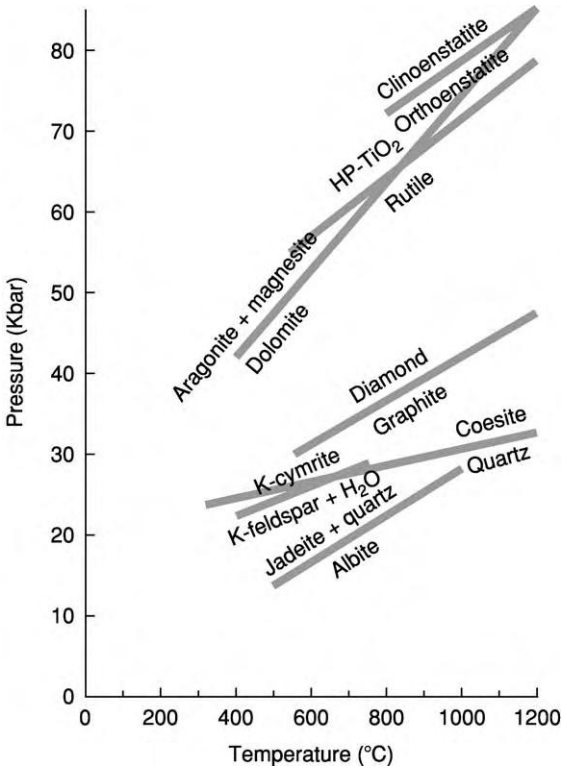


Figure 1 Pressure and temperature stability of various mineral phases that are of relevance to UHP metamorphism. Except for the transition curves below 30 kbar, the experimental error is even higher than is expressed by the thickness of the lines. The clinoenstatite orthoenstatite transition refers to a composition with 10 mol% of ferrosilite component.

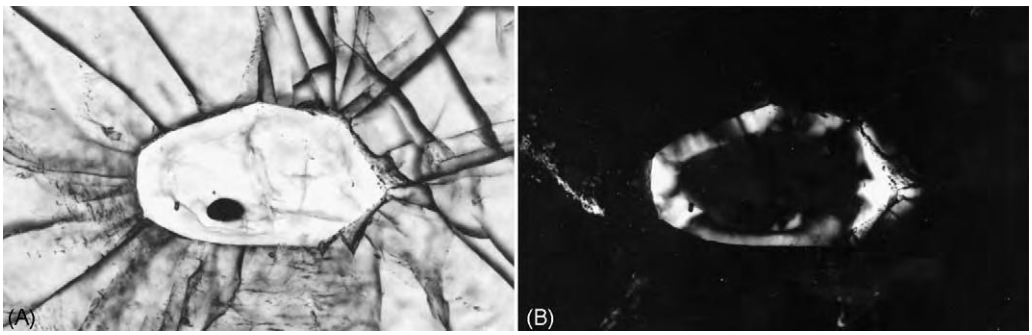


Figure 2 Photomicrograph of a coesite relic enclosed in garnet. Left hand side shows the appearance of palisade quartz using plane polarized light; the specimen is best seen under crossed nicols (i.e., a polarizing filter; right hand side). Typically, coesite has been partially replaced by the palisade quartz, causing radial cracks in the host mineral. Image width is 0.6 mm. The example is from a thin section of an eclogite from the central Saxonian Erzgebirge.

in a complete overprint of the UHP mineral assemblage, thus erasing the 'memory' (record) of the UHP event in the metamorphic rock. Therefore, it is important for the geoscientist working with such rocks to look for mineralogical hints of UHP metamorphism (e.g., tiny inclusions in resistant garnet porphyroblasts and unusual microfabric textures and appearances). Coesite, the 'indicator' mineral that points to UHP, has never been found to be a major constituent of UHP rocks. On the contrary, quartz dominates among the SiO_2 polymorphs, even in the best preserved UHP rocks, and coesite commonly occurs only as inclusions in minerals such as garnet and zircon (Figure 2). Even in such inclusions, coesite is normally partly decomposed to quartz. A typical decomposition fabric is formed of palisade quartz, with lamellae perpendicular to the replaced coesite. Cracks in the host mineral around the coesite inclusion form due to the volume increase during the decomposition of coesite to quartz. Thus, even completely replaced coesite can be recognized by such features. Tiny coesite relics can be identified by confocal micro-Raman spectroscopy even when the relics are not at the surface of a rock thin section.

Identifying UHP rocks on the basis of coesites is problematic when cracks around quartz inclusions in the host mineral are discernible but no coesite relic or palisade quartz is detectable. Palisade quartz formations may have recrystallized, forming polycrystalline quartz, but this cannot be considered a clear indication for UHP metamorphism. No doubt there have been a few reports in the literature of UHP rocks as a result of mistaken identification of unusual quartz inclusions as former coesite. Symplectites consisting of K-feldspar and quartz resemble quartz pseudomorphs of replaced coesite, and there are usually cracks in the host mineral, typically eclogitic omphacite, around such inclusions. Massonne, Dobrzhinetskaya, and Green have described such

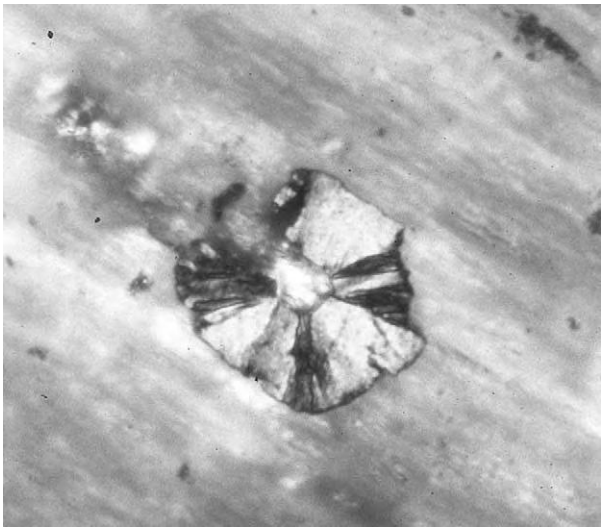


Figure 3 Inclusion of a pseudomorph of radially oriented graphite after diamond in clinopyroxene from a siliceous calcite marble of the Kokchetav Massif. A relic microdiamond is present in the centre of the pseudomorph. Image width is 0.11 mm.

symplectites as being decomposition products of K-cymrite ($KAlSi_3O_8 \cdot H_2O$). K-Cymrite has not yet been found as a relic, but it has potential for identifying UHP rocks due to its lower pressure stability limit (Figure 1), which is similar to that of coesite.

Another mineral that is diagnostic for UHP is diamond. Like coesites in UHP rocks, diamond can be easily transformed. During exhumation, unless it is enclosed as tiny grains in resistant porphyroblasts, diamond is transformed to graphite. Even when enclosed in porphyroblasts, diamond rarely survives, but it can react to form pseudomorphs of radially oriented graphite (Figure 3). Although microdiamonds are not as widespread in UHP rocks as coesite is, they are found in all kinds of UHP rocks, from ultrabasic to felsic or silicate to carbonate-rich rocks,

whereas coesite can occur only in felsic to basic rocks when SiO_2 is in excess. Phengite, a silicon-rich potassic white mica common in basic to felsic HP to UHP rocks, also has potential to be diagnostic of UHP rocks. In contrast to coesite and diamond, which are pure phases, potassic white mica is chemically complex. Mica compositions of the corresponding solid solution series can be stable at both high and low pressures, which is probably the reason that the potassic white mica that formed under HP to UHP conditions often remains in rocks during exhumation. However, phengite is diagnostic for UHP only if it is possible to detect its precise composition at UHP, and the compositions of its coexisting partners (e.g., garnet and omphacite). Although numerous experiments at high-pressure conditions have contributed to an understanding of the relationship between mica compositions and coexisting phases as a function of pressure and temperature, the degree of uncertainty in determining pressure conditions for phengite-bearing rocks is at less than 2 kbar, even when advanced thermodynamic calculation methods and suitable mineral paragenesis are considered. In any case, metamorphic temperature determinations are required using a well-calibrated geothermometer, such as the Fe^{2+}–Mg exchange reaction for different mineral pairs (e.g., garnet and omphacite). A problem arises when phengite inclusions or core compositions are used: for example, subduction of a basalt of the oceanic crust first reaches pressure and temperature conditions of the blueschist facies. At that stage, Si-rich phengites ($Si > 3.5$ per formula unit) can form and coexist with a typical blueschist facies assemblage. Further subduction at elevated pressure and temperature causes garnet growth. During this process, the blueschist-facies phengite is enclosed in garnet and thus survives the subsequent metamorphism, leading to a kyanite-bearing eclogite. At peak temperature conditions, Si contents of re-equilibrated phengites reach 3.3 per formula unit in the kyanite-bearing assemblage; this is typical for near-ultrahigh pressure conditions. If the blueschist-facies phengite is related to the peak temperature assemblage, pressures of the UHP regime would be indicated. Thus it is likely that in a few cases, rocks lacking coesite and diamond have been erroneously assigned to UHP metamorphism because a silicon-rich phengite was found in an ordinary high-pressure rock. This may have occurred with eclogite blocks in the low-grade metasediments of the Franciscan Formation in California, where phengites with high Si contents ($Si > 3.55$ per formula unit) were reported. Nevertheless, phengite geobarometry combined with geothermometers is a powerful tool and allows determination of pressure and temperature conditions in the UHP regime, whereas coesite and diamond can yield only minimum pressures. This is also true for tetrahedrally coordinated Al in clinopyroxene and orthopyroxene, which act as suitable geobarometers in the presence of garnet, especially for ultrabasic rocks lacking phengites.

In the known UHP regions of the world, minerals with specific exsolution fabrics have been observed and assigned to UHP conditions. Clinopyroxenes can contain rods of SiO_2. Garnets show clinopyroxene and orthopyroxene exsolution lamellae and precipitates (Figure 4). Both phenomena are explained by introduction of Si into the octahedral site of the corresponding mineral structure at UHP. Subsequent pressure release results in dissolution of the octahedrally coordinated Si and formation of specific minerals. Experimental constraints are related only to garnet, whereby small amounts of majorite component, $(Mg,Ca)_4Si_4O_{12}$, can be dissolved in the garnet structure at pressures exceeding 50 kbar. However, the dissolution products, clinopyroxene and orthopyroxene, can form from ordinary garnet by pressure release as well, but in that case, the pyroxenes should contain significant amounts of Tschermak's component, $CaAl_2SiO_6$. Titanite in marble from the Kokchetav Massif is yet another example for the likely introduction of Si into the octahedral site. This mineral contains coesite precipitates as dissolution product. Clinopyroxene from UHP areas can also show K-feldspar lamellae, which are interpreted as exsolution from K-bearing clinopyroxene. Such clinopyroxenes, with significant

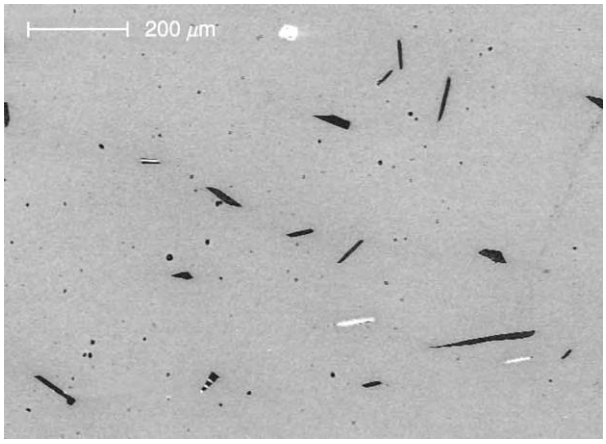


Figure 4 Exsolution of orthopyroxene and clinopyroxene from majoritic garnet, shown by a back scattered electron image (pyroxenes are darker than garnet). This sample is from the Western Gneiss Region of the Norwegian Caledonides. Reproduced with permission from Van Roermund HLM, Drury MR, Barnhoorn A, and De Ronde AA (2000) Super silicic garnet microstructures from an orogenic garnet peridotite, evidence for an ultra deep (>6 GPa) origin. *Journal of Metamorphic Geology* 18: 135 147.

amounts of K_2O (>0.5 wt.%), have been reported from diamondiferous siliceous marbles of the Kokchetav Massif. Experiments have proved that K is introduced into the clinopyroxene lattice at high pressures, but conclusions on the metamorphic pressures of the Kokchetav rocks have so far been only semiquantitative. In high-pressure experiments, TiO_2 with the α-PbO_2 structure is stable at high temperatures above 60 kbar instead of rutile (Figure 1). A nanocrystal of HP-TiO_2 was observed in a diamondiferous quartzofeldspathic rock from the Saxonian Erzgebirge, Bohemian Massif. Clinoenstatite lamellae in pyroxenes were reported from the Dabie Shan–Sulu terrane, China. This would point to pressures of 80 kbar and more (see Figure 1). Magnesite and calcite, probably former aragonite, in direct contact or separated by dolomite, were observed in the Dabie Shan. This allows the conclusion that the corresponding rocks experienced pressures of at least 60 kbar (Figure 1). Ilmenite rods in olivine were reported from several ultrabasic rocks of different UHP terranes. Relatively high concentrations of these rods were found in olivine from Alpe Arami, in the central Alps. Dobrzhinetskaya, Green, and Wang argued that this feature points to depths of more than 300 km as the origin of the corresponding rock. Certainly a number of other potential candidates, either minerals or specific mineral assemblages, could serve to diagnose UHP conditions; experimental studies and calculations of mineral equilibria using thermodynamic data have indicated this possibility, but corresponding observations have not yet been made.

Global Distribution

Following Chopin's 1983 report of finding the first UHP metamorphic rock of crustal origin in the Western Alps, several other finds were reported from around the globe, relating to more than 10 different orogens or far-distant sections of an orogenic chain. The locations of these rock finds are shown in Figure 5, together with some potential UHP terranes where, so far, no coesite or diamond relic has been found. There are certainly even more UHP potential terranes than are indicated in Figure 5 (note also that neither the Franciscan Formation in California nor Neoproterozoic areas with eclogites, for which geothermobarometrically derived pressure and temperature estimates lie in the UHP field, are indicated in Figure 5). Typically, the confirmed UHP areas worldwide are parts of young orogens resulting from

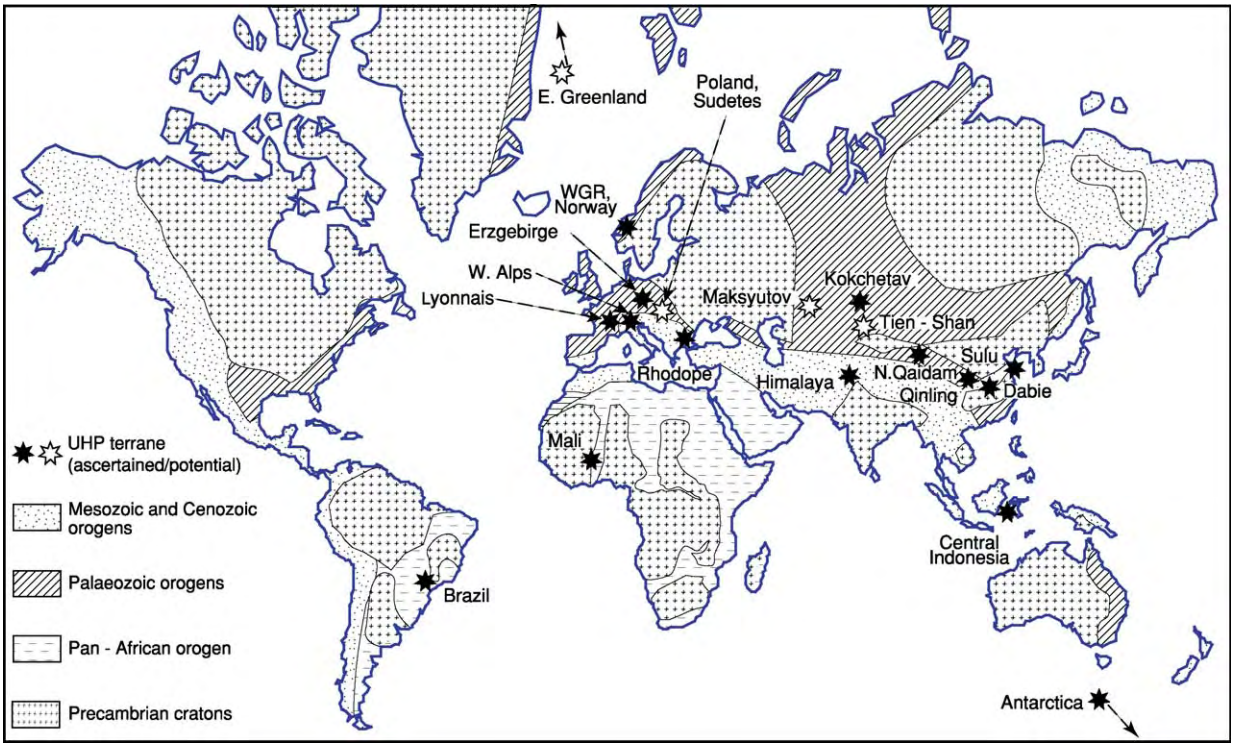


Figure 5 Occurrences of ascertained and potential UHP rocks worldwide. The group of potential UHP rocks includes pseudomorphs replacing coesite and/or diamond. WGR, Western Gneiss Region. Reproduced with permission from Chopin C (2003) Ultrahigh pressure metamorphism: tracing continental crust into the mantle. *Earth and Planetary Science Letters* 212: 1 14.

continent–continent collision, but relatively few of these areas have been well explored for UHP rocks. Those areas that have been examined closely are in the European Alps, the Norwegian Caledonides, the Chinese Qinling Shan–Dabie Shan–Sulu Terrane belt, the Kokchetav Massif in northern Kazakhstan, the Mid- and West-European Variscides, and, to some extent, the Himalaya.

European Alps

UHP rocks in the Western and Central Alps occur exclusively in the Penninic unit. Potential UHP rocks also exist in the Eastern Alps; these rocks are mainly eclogites, whereas in the Western and Central Alps, different metapelites, metagranitoids, and ultrabasic rocks are also involved in UHP metamorphism. The lateral extent of UHP units in the Alps is in the range of several to a few tens of kilometres. The thickness of these units is commonly less than 1 km, because they are part of the nappe stack of the Alpine Orogen. Age dating of the UHP event yielded values close to 35 Ma. Potential UHP eclogites in the Eastern Alps were formed in the Cretaceous. The protoliths of felsic UHP rocks of the Western Alps can be at least partially assigned to former Variscan metamorphic and plutonic rocks. This implies that Variscan continental crust was subducted to great depths during a late Alpine continent–continent collisional event.

Geothermobarometric evolution has resulted in a pressure and temperature path for the majority of the UHP rocks that is characterized by a prograde branch along geothermal gradients, decreasing from 10 to 6°C km^{-1} towards the pressure and temperature climax, which is settled between 30 and 40 kbar and around 750°C. The retrograde path typically shows decreasing temperatures during pressure release, but the temperatures of the retrograde path are always higher than those of the prograde path at a given pressure. Age dating, for instance, on zoned zircon grains has demonstrated that significant exhumation from great depths happened within a few million years. Thus the uplift rates must be in the range of several centimetres per year. This rate and dating to 35 Ma for UHP metamorphism were also found for ultrabasic rocks, for instance, outcropping at Alpe Arami, Central Alps.

Norwegian Caledonides

Felsic, basic, and ultrabasic UHP rocks are widespread in the Western Gneiss Region (WGR) of the Norwegian Caledonides. Similar rocks also occur in the Caledonian Orogen further north (for instance, in the Lofoten range) and at the eastern coast of Greenland. The widely scattered occurrences of UHP rocks within the WGR may indicate the existence of a coherent UHP terrane of up to $350 \times 150\,km^2$. Prograde metamorphism is characterized by geotherms around 6°C km^{-1}, reaching peak temperatures of 550°C in the south-east inland area and more than 800°C in the north-west coastal area of the WGR. In the latter area, microdiamonds were found on the island of Fjørtoft. Age data obtained from UHP metamorphic rocks scatter between about 440 and 390 Ma, but it has been concluded that UHP metamorphism happened close to 400 million years after closure of the Iapetus Ocean. Early exhumation rates of 1 cm or more per year are assumed. The coeval garnet peridotites originated at mantle depths of about 200 km (60–65 kbar) and temperatures significantly above 1000°C. However, it is possibly that they had resided there since mid-Proterozoic times, until they were uplifted by a mantle diapir in the Lower Devonian.

Chinese Qinling Shan - Dabie Shan - Sulu Terrane Belt

In central and eastern China, the Triassic orogenic belt resulting from collision of the Yangtze Craton and the Sino-Korean Craton contains sections with UHP metamorphic rocks. Although intruded by enormous volumes of Cretaceous granitoids, these areas can be traced over 100 km and more, leading to the conclusion of the existence of wide, coherent UHP terranes. Under this scenario, orthogneisses with Proterozoic protolith ages would be the dominant UHP rock type there. The pressure and temperature conditions of the Chinese UHP rocks resemble those of the WGR in regard to spread of peak pressure and temperature ($\leq$600–$\geq$800°C) conditions. However, a significant difference between the WGR and the Triassic orogenic belt is related to the peak temperature conditions of ultrabasic rocks. Garnet peridotites of the Sulu Terrane were metamorphosed at only 800°C, and those of the Dabie Shan were metamorphosed below 800°C. In spite of the relatively low temperatures, pressure conditions were estimated to be as high as 55 kbar or even more, thus yielding geothermal gradients somewhat below 5°C km^{-1}. It is assumed that rocks as well as other UHP rocks of the Qinling Shan–Dabie Shan–Sulu Terrane belt dates close to 220 Ma.

Kokchetav Massif in Northern Kazakhstan

An enormous variety of identified UHP rocks is known to occur at local sites (e.g., close to Lake Kumdy Kol) of the Kokchetav Massif. Although the development of such rocks can be traced over several tens of kilometres, the recent view of a

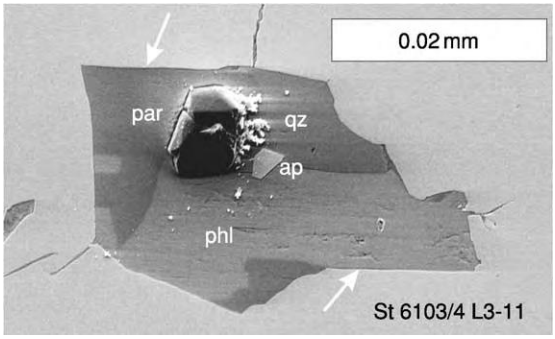


Figure 6 Scanning electron microscope image of a polyphase diamond bearing inclusion in garnet, interpreted as trapped sili ceous fluid or melt before crystallization. The example is a quart zofeldspathic rock from the Saidenbach reservoir, central Saxonian Erzgebirge. Arrows point to rational mica garnet inter faces. qz, quartz; par, paragonite; phl, phlogopite; ap, apatite.

'megamelange', including medium-pressure rocks, may best describe the geological situation. Nevertheless, two types of UHP rocks can be distinguished in terms of maximum pressure and temperature conditions. One type has experienced pressure and temperature conditions similar to those of UHP rocks of the Western Alps ($T_{max} \sim 750°C$). The other type, characterized by abundant microdiamond inclusions in various host minerals, was metamorphosed at pressures as high as 70 kbar. At these high pressures, temperatures exceeded 1000°C, probably resulting in partial melting. Polyphase aggregates (Figure 6), commonly containing microdiamonds, are enclosed in garnet and other phases. These aggregates serve as evidence for siliceous fluids or melts. The silica-rich material of these inclusions was probably trapped in growing minerals during exhumation of the (partially molten) rocks at decreasing temperatures. This happened at 530 Ma, as deduced from zircon dating, also indicating a fast uplift of the rock in the range of centimetres per year.

Mid- and West-European Variscides

An abundance of ascertained and potential UHP rocks have been detected in several crystalline complexes of the Variscides, including those later involved in the Alpine Orogeny. However, due to a significant fragmentation of specific major units by a late orogenic event, a former, possibly extended, coherency of UHP terranes has been broadly lost. Nevertheless, HP to UHP metamorphism cannot be related to a single event. For instance, in the Bohemian Massif, representing the north-eastern section of the Variscides, two major HP–UHP events can be discriminated from abundant age data. The earlier event, around 395 Ma, led to eclogites, probably former oceanic crust, which experienced maximum pressure and temperature conditions close to 700°C and 23–30 kbar. A common feature of these eclogites is the replacement of omphacite by amphibole porphyroblasts due to the influx of hydrous fluids, probably close to the pressure and temperature climax. In the younger UHP event, occurring about 340 Ma, felsic, basic, and ultrabasic rocks were involved. Peak temperature conditions can exceed 1000°C–1200°C and 80 kbar are indicated by a diamondiferous quartzofeldspathic rock from the Saxonian Erzgebirge; this rock has polyphase inclusions in garnet (Figure 6), similar to rocks from the Kokchetav Massif. The anatectic evolution of the Erzgebirge rock is inferred to be similar to that of diamondiferous rocks from the Kokchetav Massif, in addition to the fast exhumation. The diamondiferous rock from the Erzgebirge forms lenses up to 1 km long, like other UHP rocks in the Variscides. The surrounding gneisses display a medium to high pressure signature, but also a clear peak pressure contrast to the UHP rocks.

Himalaya

UHP rocks in the Himalaya are very rare, but their geodynamic context is much clearer compared to other UHP regions. The protoliths of coesite-bearing eclogites and surrounding metasediments in the north-western Himalaya were part of the continental margin of India that was subducted beneath Asia. The prograde metamorphism during this event followed a geothermal gradient of $6°C\,km^{-1}$ reaching peak pressure and temperature conditions at depths between 90 and 120 km about 50 Ma ago. Amphibole blastesis at the expense of clinopyroxene is common. The subsequent exhumation of the UHP rocks is characterized by moderate cooling at uplift rates in the range of centimetres per year, slowing down to millimetres per year for the past 40 million years.

Mechanisms

The known UHP rocks at Earth's surface can be subdivided into two groups. One group suffered from metamorphism along geothermal gradients of about $7°C\,km^{-1}$, reaching maximum pressures generally below 40 kbar (Figure 7). Frequently observed peak pressure and temperature conditions are 30 kbar at 650–700°C. Moderate cooling and uplift rates in the range of centimetre per year characterize the exhumation of this group of UHP rocks, commonly starting with the influx of hydrous fluids (for instance, into eclogites). The subduction of oceanic crust under continental crust, including the exhumation within a subduction channel, best explains the characteristics of these UHP rocks. However,

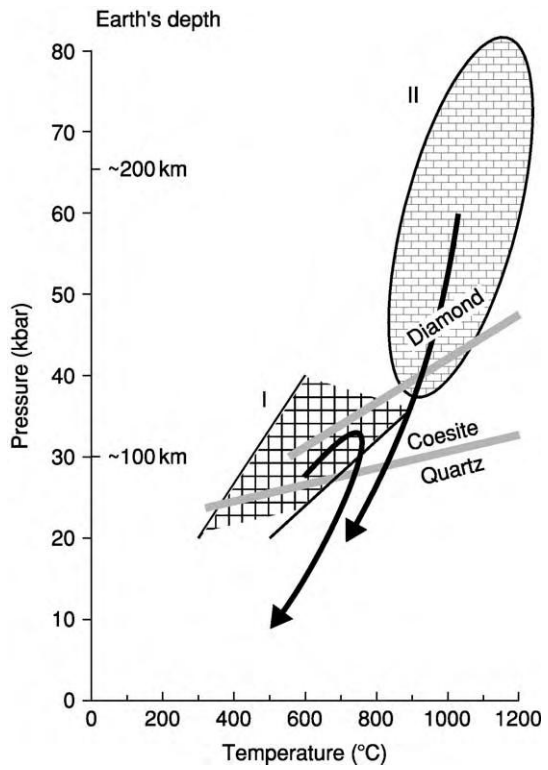


Figure 7 Range of peak pressure and temperature conditions of UHP rocks (patterned areas). The typical shapes of common pressure and temperature paths for UHP/ near UHP (I) and UHP (II) are depicted.

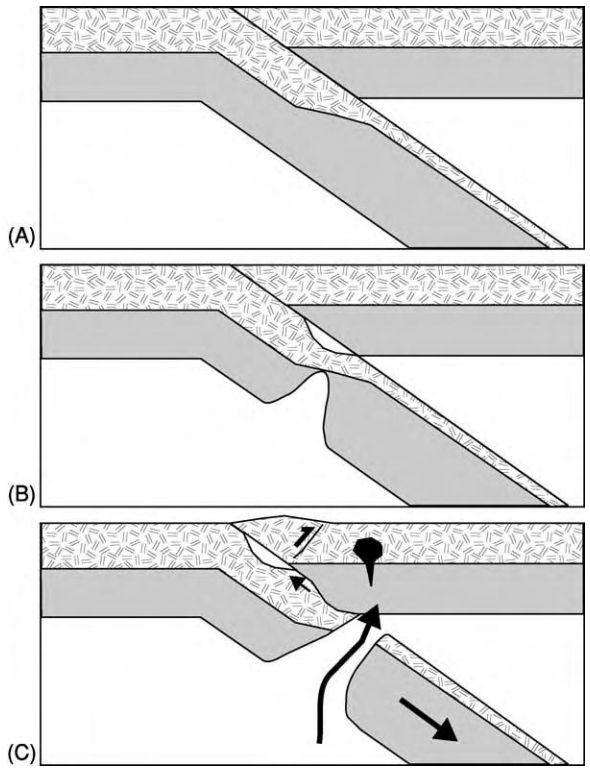


Figure 8 The slab breakoff model, evolution (from top to bottom). (A) Subduction; (B) slab weakening and narrow rifting; (C) slab breakoff, magmatism, and uplift of UHP sheets. Reproduced with permission from Hacker BR and Liou JG (1998) *When Continents Collide: Geodynamics and Geochemistry of Ultrahigh Pressure Rocks*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

formation of larger coherent UHP terranes consisting broadly of continental crust, as inferred from the WGR in Norway and the Triassic UHP belt in China, cannot be explained by this mechanism alone. If continental crust adherent to a slab of subducted oceanic crust is drawn into depth at the beginning of a continent–continent collision, as suggested by the slab-breakoff model (Figure 8), extended regions of continental crust can be affected by UHP metamorphism. Fast exhumation is caused by buoyancy forces exerted by continental material that is less dense than eclogites and garnet-bearing ultrabasic rocks, after the oceanic slab has been broken off to be subducted further down.

A second group of UHP rocks experienced significantly higher temperatures, or at least higher pressures (Figure 7). Often both groups of UHP rocks occur together in one crystalline complex. Moreover, both crustal material and mantle material were metamorphosed at peak pressure conditions, between 60 and 80 kbar, as has been proved at least for the UHP regions of the Kokchetav Massif and the Bohemian Massif. The slab-breakoff model may explain this situation as well, because it is believed that continental crust can be dragged to depths of about 200 km by the adherent oceanic slab, despite the buoyancy forces of the continental material. An alternative explanation, however, is delamination of continental lithosphere after continent–continent collision and significant thickening of continental crust, a process that is currently observable in the range of the Himalaya and in the Tibetan Plateau. Modelling experiments suggest that material from the continental crust can be involved in the delamination process and deeply submerged into Earth's mantle. There, anatectic processes in the continental material are caused by the hot environment before fast uplift starts. Because there is limited coherency among the hot-temperature UHP rocks of the Bohemian Massif or the Kokchetav Massif, the debate continues as to whether the slab-breakoff model, lithospheric delamination, or any other process can sufficiently explain the UHP rocks there.

Conclusions

Continent–(micro)continent collision leads to the formation of UHP rocks. Such rocks, although a

minority among metamorphic rocks in crystalline basement areas, are common in denuded Phanerozoic orogens. UHP rocks that originate from crustal protoliths are difficult to detect and it seems that they did not appear at Earth's surface until about 600 Ma ago. What could be responsible for that? Crustal thickening during continent–continent collision over a wide lateral area, resulting in a crust 60–70 km thick, may have been uncharacteristic in Proterozoic and Archaean times, thus limiting lithospheric delamination with continental crust involved. The magnitude of crustal thickening might have been influenced by the thermal structure of the (lower) crust and lithospheric mantle, where geothermal gradients are presently lower than they were a long time ago. The present thermal structure could also be responsible for lower geothermal gradients in subducted and metamorphosed oceanic crust, compared to those in pre-Phanerozoic times. Thus, dipping into the UHP field before melting would have been possible for oceanic crust only since the beginning of the Phanerozoic. This would also apply to continental crust adherent to the subducted oceanic crust when the slab-breakoff model would mirror the true process in overprinting continental material at UHP conditions.

See Also

Igneous Rocks: Kimberlite. **Impact Structures**. **Minerals:** Definition and Classification; Quartz. **Regional Metamorphism**. **Shock Metamorphism**. **Solar System:** Meteorites.

Further Reading

Carswell DA (2000) Special issue: Ultra high pressure metamorphic rocks. International Lithosphere Programme contribution 345. *Lithos* 52.

Carswell DA and Compagnoni R (2003) Ultrahigh pressure metamorphism. *EMU Notes in Mineralogy* 5.

Chopin C (1984) Coesite and pure pyrope in high grade blueschists of the Western Alps: a first record and some consequences. *Contributions to Mineralogy and Petrology* 86: 107 118.

Chopin C (2003) Ultrahigh pressure metamorphism: tracing continental crust into the mantle. *Earth and Planetary Science Letters* 212: 1 14.

Coleman RG and Wang X (1995) *Ultrahigh Pressure Metamorphism*. Cambridge, UK: Cambridge University Press.

Dobrzhinetskaya L, Green HWII, and Wang S (1996) Alpe Arami: a peridotite massif from depths of more than 300 kilometers. *Science* 271: 1841 1846.

Ernst WG and Liou JG (2000) *Ultrahigh Pressure Metamorphism and Geodynamics in Collision type Orogenic Belts*. Columbia, MD: Bellwether Publishing.

Hacker BR and Liou JG (1998) *When Continents Collide: Geodynamics and Geochemistry of Ultrahigh Pressure Rocks*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Liou JG, Zhang RY, Ernst WG, Rumble D III, and Maruyama S (1998) High pressure minerals from deeply subducted metamorphic rocks. In: Hemley RJ (ed.) *Ultrahigh Pressure Mineralogy: Physics and Chemistry of the Earth's Deep Interior. Reviews in Mineralogy,* vol. 37, pp. 33 96.

Massonne H J (2003) A comparison of the evolution of diamondiferous quartz rich rocks from the Saxonian Erzgebirge and the Kokchetav Massif: are so called diamondiferous gneisses magmatic rocks? *Earth and Planetary Science Letters* 217: 1 19.

Massonne H J, Dobrzhinetskaya L, and Green HW II (2000) Quartz K feldspar intergrowths enclosed in eclogitic garnet and omphacite. Are they pseudomorphs after coesite? *Extended Abstracts of the 31st International Geological Congress at Rio de Janeiro, Brazil, 6 17 August 2000* (on CD; search for Massone).

Parkinson CD, Katayama I, Liou JG, and Maruyama S (2002) *The Diamond Bearing Kokchetav Massif, Kazakhstan*. Tokyo, Japan: Universal Academic Press.

Schreyer W and Stöckhert B (1997) Special issue: High pressure metamorphism in nature and experiment. International Lithosphere Programme contribution 327. *Lithos* 41.

Sobolev NV and Shatsky VS (1990) Diamond inclusions in garnets from metamorphic rocks: a new environment for diamond formation. *Nature* 343: 742 745.

Van Roermund HLM, Drury MR, Barnhoorn A, and De Ronde AA (2000) Super silicic garnet microstructures from an orogenic garnet peridotite, evidence for an ultra deep (>6 GPa) origin. *Journal of Metamorphic Geology* 18: 135 147.

UNCONFORMITIES

A R Wyatt, Sidmouth, UK

Introduction

An unconformity is a surface that separates rocks of significantly different ages. This was at one time an exposed part of the Earth's land surface or the rock surface below a body of water (for example, a lake or the sea), and the younger rocks were deposited on this surface. Juxtaposition of rocks of different ages caused by faulting does not give rise to an unconformity. An unconformity represents a substantial break or gap in the local or regional depositional record. In modern usage this break or gap may have been caused by the erosion of previously deposited rocks or by a long period of non-deposition of sediments (that is, a long enough period that the absence of sediments of the relevant age can be recognized).

Early workers confined the use of the term unconformity to places where the older rocks had been deformed and eroded, so that the unconformity cut across the truncated beds of the lower deposits. The idea that structural discordance is an essential feature of an unconformity continued for much longer in the UK than in many other parts of the world, such as, for example, the USA. Other terms were introduced for breaks where there was no structural discordance. In the Phanerozoic these would normally be identified by gaps in the expected fossil sequence. For the simple case of non-deposition, the terms diastem and non-sequence were used. Although these terms have often been considered to be synonymous, some workers have suggested that a diastem is a break of shorter duration than a non-sequence. Where the break can be shown to be associated with erosion but the upper beds are still parallel to the lower beds, the term disconformity was used. The disconformity surface is often parallel to the bedding surfaces, but it may also show major relief.

Terms such as unconformity and disconformity refer to the surface (and, by implication, the time) that separates the older from the younger rocks. Terms have also been introduced to refer to the relationship between the bedding of the upper (younger) rocks and that of the lower (older) rocks. Where there is no structural discordance, so that the attitude of the upper beds is the same as that of the lower beds, the upper beds are said to be conformable. When there is structural discordance, the upper beds are unconformable.

Where there is structural discordance, as we follow the base of the overlying bed we find that it moves from one to another member of the lower, truncated, series. This is known as overstep (see **Figure 1A**, side face of block). The term is chiefly used when the angular nature of the unconformity is not obvious but is made evident by detailed mapping. One of the earliest recorded examples was the overstep of the base of the Cretaceous across the underlying Jurassic

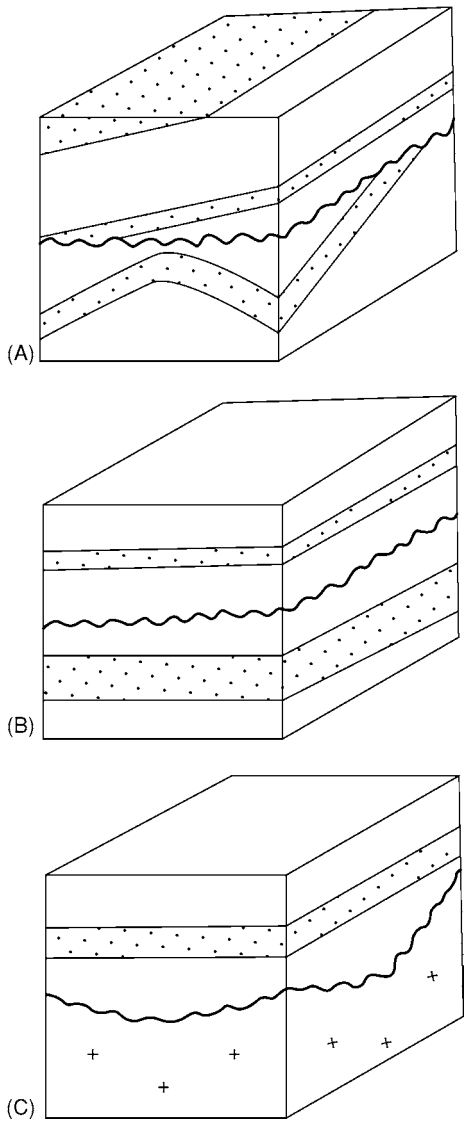


Figure 1 Types of unconformity. (A) Angular unconformity. The front face of the cube shows overlap, caused by onlap from the right. The side face shows overstep of the base of the overlying beds over the dipping lower beds. (B) Disconformity. (C) Heterolithic unconformity.

formations in Yorkshire, which was first noticed in the late eighteenth century.

The rocks overlying an unconformity often show features that record changes in the areal extent of deposition through time, particularly when looked at on a regional scale. For example, where the sea transgresses over the land each bed will cover a slightly greater area than the bed below. This process is known as onlap, and the result, as seen in the rocks, is referred to as overlap (see **Figure 1A**, front face of block). In recent years many authors have failed to distinguish between the process and the product, using the term onlap to describe both. In traditional usage the opposite of onlap – that is, the successive contraction in the lateral extension of beds in an upward succession – is known as offlap. It should be noted that offlap has been used in a very different sense by seismic stratigraphers (for reflection patterns generated by strata prograding into deep water).

History of the Concept

Some early workers published sketches of what would later be called unconformities, although they did not discuss their significance. Nicolaus Steno (1638–1687) (*see* **Famous Geologists:** Steno) produced a series of diagrams suggesting how unconformable beds in Tuscany could have been produced by cave formation and the subsequent collapse of the roof of the cave. John Strachey (1671–1743) published a diagram showing a sequence of horizontal Triassic and Jurassic rocks overlying inclined Carboniferous Coal Measures in Somerset. Some of his other diagrams suggest that he had a very vague understanding of what he was showing. Jean Etienne Guettard (1715–1786), working in northern France, produced some of the first geological maps, and two of the published maps included sections that clearly show unconformities. Unfortunately, no explanation of the observations was published.

James Hutton (1726–1797) (*see* **Famous Geologists:** Hutton) was the first author whose writings show that he understood the significance of unconformities (though he did not name them). In 1785 he presented his theory of the Earth at a meeting of the Royal Society of Edinburgh. Included in this theory was the concept of geostrophic cycles: the idea that the denudation of the landmasses produces sediments that are deposited on the seafloor and that these sediments are consolidated into rocks, elevated, folded, and denuded. Hutton theorized that there should be places where rocks from one cycle are overlain by rocks from a younger cycle, but there is no evidence that, at the time of giving his paper, Hutton had either seen or read about actual examples. If he had seen the work of Steno, Strachey, or Guettard, he would have been able to point to their diagrams as evidence for his theory.

Over the next few years Hutton searched for, and eventually found, field examples. In 1787 he discovered the unconformity at North Newton, near Lochranza, Arran (**Figure 2**). Here, reddish and yellowish sandstones, associated with some bands of caliche palaeosol, probably of Late Devonian age, rest with a marked discordance on Dalradian Schists of Late Cambrian age.

Almost immediately after finding the Arran locality, Hutton discovered horizontal beds of the Upper Old Red Sandstone lying on highly inclined (almost vertical) Silurian greywackes near Jedburgh. In the spring of the following year (1788) he found another example of almost horizontal Upper Old Red Sandstone unconformably overlying highly inclined Silurian greywackes at Siccar Point, north of Berwick.

Other examples were found by Hutton's friend John Playfair (1748–1819), whose book *Illustrations of the Huttonian Theory*, published in 1802, did much to draw attention to Hutton's work. Both Hutton and Playfair lacked a simple term to name what they were describing. The term unconformable was introduced in 1805 by one of their geological opponents, Robert Jameson (1774–1854), as an English translation of the German expression *abweichende Lagerung* ('deviating bedding or stratification') used by followers of Abraham Gottlob Werner (1749–1817). For some decades after 1805 geologists described examples of unconformable rocks, without, it appears, paying much attention to the cause of the relationship. It was only in the late 1830s, after Charles Lyell (1795–1875) (*see* **Famous Geologists:** Lyell) started publishing his extremely popular books, that the concept of, and name, unconformity really became incorporated in the thinking and language of geology.

For the rest of the nineteenth century the term unconformity was used to describe an angular discordance between two sets of strata. It was in 1905 that Amadeus William Grabau (1870–1946) extended the use of the term to include cases where there was an obvious erosional break in otherwise parallel strata. Grabau called this a case of disconformable strata, which soon became known as a disconformity.

In 1909 Eliot Blackwelder (1880–1969) suggested that the contact between sedimentary rocks and underlying igneous or metamorphic rocks should also be called an unconformity. Some authors began to apply the term nonconformity to this type of unconformity. Unfortunately, the term nonconformity was already in use as an alternative name for an angular unconformity. To try to avoid this confusion later authors used the term heterolithic unconformity. (Although at first

Figure 2 Hutton's unconformity at North Newton, near Lochranza, Isle of Arran, Scotland. The arrows mark the obvious change from steeply dipping Dalradian Schists in the lower part of the photograph to sub horizontal Devonian sandstones in the upper part.

sight 'heterolithic' (Greek for unlike rocks) may appear to be an appropriate term, it could also apply to clastic rocks overlying carbonates or to marine rocks overlying non-marine rocks. In some parts of the world heterolith is used to refer to interbedded sandstones, siltstones, and mudstones).

In all of the cases discussed above there was an assumption that the surface of unconformity represented a subaerial erosion surface. If we look at many parts of the world today, we see that the land surface is not a smooth plane. It is therefore not surprising that many unconformities that originated as subaerial surfaces are also irregular. This is often referred to as a buried-landscape type of unconformity. A good example is the Torridonian unconformably overlying the Lewisian in north-western Scotland.

In 1910 Bailey Willis (1857–1949) included surfaces of non-deposition in marine sediments as a variety of unconformity. Over the next few decades several authors stressed the importance of subaqueously formed breaks, but it was not until 1957 that John Essington Sanders (1926–1999) proposed a complex Greek-based nomenclature that attempted to distinguish clearly between subaerial and subaqueous breaks. It is, perhaps, not surprising that most people were put off by the nomenclature or thought that it was all an elaborate joke. But in ignoring the terms many people also ignored the attempt to refine geological thinking.

It was inevitable that the expansion of the use of the term unconformity would give rise to some nomenclatorial confusion (Table 1). The same word was used by different authors to describe different concepts, and the same concept was given different names. In recent years there has been some convergence of views on the nomenclature. Authors have tended to use the descriptive terms angular unconformity, disconformity, and heterolithic unconformity (Figure 1). This consensus has, however, been challenged by the specific definition of unconformity that has been used by the proponents of seismic and sequence stratigraphy (see below).

Lateral Variation

One descriptive term is usually adequate to describe an unconformity at a single exposure, but when the surface is traced over wide areas it is common to see the nature of the unconformity vary. A good example is the North Sea Unconformity Complex, often called the 'base-Cretaceous unconformity' or the 'Late-Cimmerian unconformity'. This is perhaps the most easily identifiable surface of the Phanerozoic succession of the Norwegian continental shelf. It displays great local complexity and great variability on a regional scale, such that in different places it has been classified as a nonconformity (in the sense of heterolithic unconformity), a disconformity, and an angular

Table 1 Names that have been given to different types of unconformity

Angular	*Parallel*	*Non depositional*	*Overlying igneous or metamorphic rock*
Angular discordance	Accordance	Concordant leuroatmodialeima	Heterolithic unconformity
Angular unconformity	Concordant trachyatmodialeima	Concordant leurodiscontinuity	Nonconformity
Clinounconformity	Concordant trachydiscontinuity	Concordant leurohydrodialeima	
Discordance	Concordant trachyhydrodialeima	Diastem	
Discordant atmodialeima	Disconformity	Marine unconformity	
Discordant discontinuity	Eroded surface	Nonevident disconformity	
Discordant hydrodialeima	Evident disconformity	Non sequence	
Nonconformity	Parallel unconformity	Paraunconformity	
Unconformity	Paraunconformity	Surface of non deposition	

unconformity. This variation in the nature of the surface reflects local differences in the processes of formation.

The unconformity complex developed during the transition from the synrift stage (active stretching) to the post-rift stage (thermal subsidence and sediment loading) in the development of the northern North Sea basin. A transgression coincided with the transition. This combination of differential subsidence, block rotation, changing patterns of sediment input, and sea-level rise caused local differences in patterns of erosion and sedimentation, which are reflected in the spatial variation of the type of unconformity.

Although varied in detail, there is a general distribution pattern of the different types of unconformity. At the rift margins the rising sea covered the previously exposed basement, producing nonconformities. On the rift flanks, where faulted blocks subsided and rotated, angular unconformities were normally developed. In the centre of the rift, subsidence dominated, generally giving rise to disconformities.

This example, produced during the development of a passive margin, helps to demonstrate that unconformities can originate in a variety of tectonic and sedimentary settings and are not just products of erosion at the end of a geostrophic cycle.

Unconformities and the Stratigraphic Record

Once the Huttonian theory of geostrophic cycles became commonly known, geologists started to apply the reasoning in their efforts to understand and classify the rock record. It became clear that there were major periods of deformation, uplift, and erosion, known as orogenies, which could be recognized over large areas, and the consequent unconformities were used to subdivide the geological column. It soon became apparent that much of Britain and Scandinavia had been affected by the Caledonian orogeny, which was originally thought to have culminated in the Late Silurian. The three classic unconformities discovered by Hutton were all produced by deposition after Caledonian deformation. South-western Britain, and much of the adjacent continent, had been affected by an orogeny that culminated in the Late Carboniferous, which was variously termed the Armorican, Hercynian, or Variscan (*see* **Europe:** Variscan Orogeny). Geologists in North America recognized a similar pattern of orogeny. It was also apparent that there is an ongoing Alpine–Himalayan orogeny.

Although the causes of these orogenic episodes were unclear and were to remain so until the development of plate-tectonic theory in the 1960s, the practical result was the rapid development of the broad outlines of the stratigraphic column. In addition to the major unconformities associated with the final phases of uplift and erosion, other unconformities were discovered that helped in the processes of subdivision and classification. Stratigraphers and palaeontologists could then look in more detail at the rocks bounded by these unconformities.

In many cases the first rocks deposited on an unconformity surface are conglomerates, often containing pebbles eroded from locally weathered rocks. These pebbles can give us information about the rocks that were exposed at that time. For example, the basal Carboniferous conglomerates that lie unconformably on Silurian shales in the east of the English Lake District contain distinctive fragments of the Shap granite. We know that the granite is intrusive into the Silurian (up to and including the Upper Ludlow), so we have some constraints on the timing of cooling, crystallization, uplift, and erosion of the granite.

At a higher stratigraphic level, although in the same area of England, we find the Lower Brockram of Permian age unconformably overlying the Carboniferous Limestone. The Lower Brockram is formed of pebbles that are mostly Carboniferous Limestone. The Upper Brockram, found slightly higher in the

succession, contains a large proportion of fragments of Ordovician and Silurian sedimentary and volcanic rocks, demonstrating that by the time the Upper Brockram was deposited an area of these older rocks was exposed and being eroded.

With the development of stratigraphic thinking and knowledge it became clear, however, that unconformities were a poor choice for defining widely correlatable boundaries. An unconformity necessarily implies that there is a gap in the record. This gap may represent different durations of time in different areas. For good correlation, detailed information about fossil occurrences and other temporal markers, such as ash bands, is needed. A gap cannot provide this kind of detail. So a lot of effort was put into searching for areas where a continuous sedimentary record, deposited during the time-span represented by the gap, could be demonstrated. In many cases this involved looking outside the European area, where most of the stratigraphic units had first been defined.

Unconformities and Sequences

For many practical problems, the refinements of stratigraphy are less important than the local rock distribution, so not all geologists abandoned the use of unconformities. From the 1930s onwards Arville Irving Levorsen (1894–1965) interpreted the geology of the mid-continent region of North America in terms of large-scale unconformity-bounded tectonostratigraphic units. Levorsen did not propose names for these units, referring to them as layers of geology, but he demonstrated their importance in petroleum exploration. It was in the late 1940s that Laurence Louis Sloss (1913–96) began to use the term sequence for such major unconformity-bounded units, eventually proposing a formal name for each sequence. In order to distinguish these sequences from standard chronostratigraphic units they were given the names of Indian tribes (Table 2).

Sloss, and many other North American stratigraphers, had encountered problems when attempting to apply the mostly European-based stratigraphic divisions to North American rocks. However, a small number of major craton-wide unconformities could be recognized, based on an integrated study of outcrop and subsurface data. Sloss stressed that at the scale of an individual exposure there was no obvious characteristic by which these inter-regional unconformities, which were used to separate sequences, could be distinguished from the many local unconformities. This was graphically illustrated when the unconformity at the base of the Absaroka Sequence was redefined.

The inter-regional unconformities represented major breaks in the depositional record and were associated with a great degree of overstep and overlap. For example, the rocks at the base of the Sauk Sequence range in age from latest Precambrian to Late Cambrian. They were deposited on an unconformity that cut across rocks of a great range of Precambrian ages. The boundary at the top of a sequence was interpreted as representing a time of major regression of the sea from the continental craton, with associated subaerial erosion. The base of the next sequence represented the re-flooding of the craton. Local unconformities were thought to be produced by minor fluctuations in the rate of sea-level rise or fall, by local tectonics, or by local changes in sediment supply.

Sequences, as promoted by Sloss, are major units, with demonstrably diachronous boundaries. Although they are useful as major subdivisions of North American strata, they were not intended to replace the standard chronostratigraphic units. They were erected because of the perceived differences between the rock record in North America and that in Europe. By no means all stratigraphers agreed with this approach, arguing that better-defined chronostratigraphic boundary sections and improved techniques of correlation would eventually help to solve the difficulties.

Unconformities, Seismic Stratigraphy, and Sequence Stratigraphy

The development of high-quality seismic-reflection profiles, primarily as a result of the intensive search for hydrocarbons, and the calibration of these profiles

Table 2 North American sequences

Sequence name	*Age of rocks included in the sequence*
Tejas	Late Paleocene to Holocene
Zuni	mid Jurassic to mid Paleocene
Absaroka	latest Mississippian (post Chesterian) to Early Jurassic[a]
Kaskaskia	late Early Devonian to Late Mississippian
Tippecano	mid Ordovician to Early Devonian
Sauk	latest Precambrian to Early or possibly early mid Ordovician

[a]The Absaroka was originally defined as from the Chesterian.

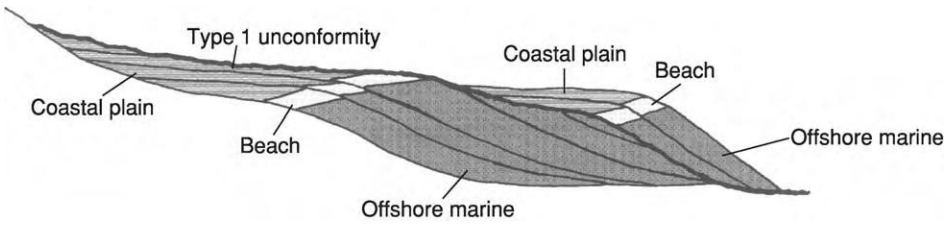


Figure 3 A type 1 unconformity, produced during a rapid relative fall in sea level, when the rate of eustatic fall exceeded the rate of basin subsidence. The diagram shows the erosion associated with the sea level fall, the ongoing subaerial erosion, the basinwards shift in facies, the downward shift in coastal onlap, and the abrupt change in facies. Only parts of the sequences above and below the unconformity are represented, but minor marine flooding surfaces show how the sequences can be subdivided into parasequences.

by geophysical logging techniques applied to relevant boreholes, has stimulated a major development of interest in unconformities. Seismic stratigraphy was originally developed by members of an Exxon research team and was presented and promoted to the worldwide geological community from the mid-1970s onwards. Seismic stratigraphy is simply the geological interpretation of seismic data. Its basic premise is that primary seismic reflectors represent either major bedding surfaces (with the assumption that the reflections are following isochronous horizons) or unconformities. By the 1980s the originators had broadened their concepts and were talking about sequence stratigraphy rather than seismic stratigraphy (*see* **Sequence Stratigraphy**).

Data analysis is based on the identification of stratigraphic units composed of genetically related strata, known as depositional sequences. The lower and upper boundaries of these depositional sequences are unconformities or their correlative conformities. When sequence stratigraphy was first promoted, an unconformity was defined as a surface of erosion or non-deposition that separates younger rocks from older rocks and represents a significant hiatus. A conformity was defined as a surface along which there is no evidence of erosion or non-deposition and along which no significant hiatus is indicated. (Purists might point out the difficulty of correlating a gap with a surface or argue whether theoretically there needs to be a correlative conformity, but most workers seem to accept the overall concept). The sequences of seismic and sequence stratigraphy are much smaller units than the North American sequences named by Sloss, and the time-transgressive nature of their bounding unconformities is not considered to be significant.

By the late 1980s the originally proposed concept of sequences was being refined and extended. Two types of sequence were now recognized, which are known as type 1 and type 2. They are differentiated on the basis of their lower boundaries, which have come to be called type 1 and type 2 unconformities. A type 1 unconformity is characterized by subaerial exposure and erosion associated with stream rejuvenation, a basinwards shift in facies, a downward shift in coastal onlap, and an abrupt change in facies, for example non-marine or very shallow-water marine rocks overlying deeper-water marine rocks (Figure 3). It is interpreted to form when there is a relative fall in sea-level at the depositional shoreline break, i.e. when the rate of eustatic fall exceeds the rate of basin subsidence. A type 2 unconformity lacks both subaerial erosion associated with stream rejuvenation and a basinwards shift in facies (Figure 4). It is interpreted to form when no relative fall in sea-level occurs at the depositional-shoreline position, i.e. when the rate of eustatic fall is less than the rate of basin subsidence.

Sequences were now subdivided into parasequence sets and parasequences. A parasequence is a relatively conformable succession of genetically related beds bounded by marine-flooding surfaces and their correlative surfaces. A parasequence set is a succession of genetically related parasequences that form a distinctive stacking pattern bounded, in many cases, by major marine-flooding surfaces and their correlative surfaces.

These definitions require a clear distinction to be made between a marine-flooding surface and an unconformity. A marine-flooding surface is defined as a surface that separates younger strata from older strata across which there is evidence of an abrupt increase in water depth. The deepening is commonly associated with minor submarine erosion (but no subaerial erosion or basinwards shift of facies) and non-deposition, and a minor hiatus may be indicated. An abrupt increase in water depth implies a transgression at the basin margins, with marine sediments overlying an exposure surface.

An unconformity is now defined as a surface separating younger strata from older strata, along which there is evidence of subaerial erosional truncation (and, in some cases, correlative submarine erosion) or subaerial exposure, with a significant hiatus. This use of the term unconformity is obviously more restrictive than that used when seismic stratigraphy was

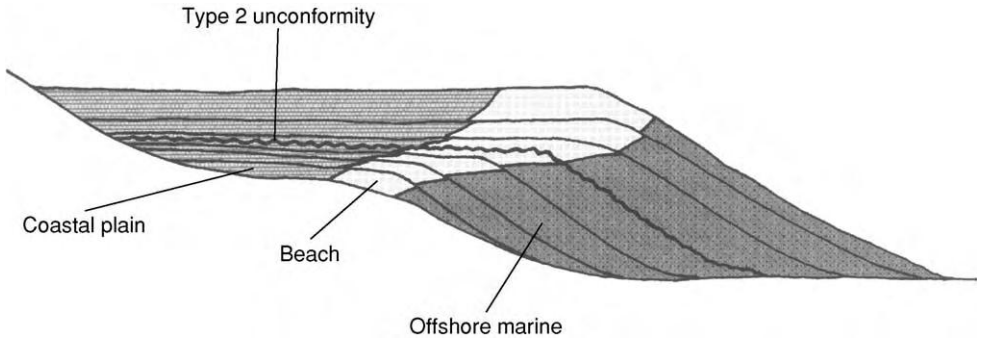


Figure 4 A type 2 unconformity, produced when the rate of eustatic fall is less than the rate of basin subsidence such that there is no relative fall in sea level. There is no major erosion of the underlying sediments. A major feature associated with a type 2 unconformity is a change from prominent progradation below the unconformity, caused by an increasing rate of regression, to prominent aggradation above the unconformity as the rate of regression slows.

first proposed and is also more restrictive than traditional use. This definition is required to differentiate between sequence and parasequence boundaries. Although the promoters of sequence stratigraphy proposed this restricted definition only in the context of their work, their definition has been applied more widely by some workers.

Although Hutton's original concept of an unconformity was based on his geostrophic theory, subsequent workers have tended to use the term to describe what has been observed, with explanations and hypotheses being separated from the observation. To restrict the use of the term to fit a particular hypothesis, as has been done by the promoters of sequence stratigraphy, is, to some people, a retrograde step. With such a well-known term as unconformity it can also lead to confusion, as many workers continue to follow traditional usage.

One early product of the sequence-stratigraphic model was a series of charts of global cycles of relative change in sea-level through time, based on the interpretation of unconformities and marine-flooding surfaces as products of eustatic sea-level change. Much of the data from which the charts were developed came from North America, with a small number of regional studies from elsewhere. The fact that many of the unconformities on these global cycle charts do not match traditional stratigraphic-unit boundaries, many of which were originally erected based on unconformities, does suggest a rather more complex interplay between local, regional, and global events than that proposed in the model.

Despite these caveats, work on seismic and sequence stratigraphy has promoted a huge increase in our knowledge of unconformities. Although much of this data remains in confidential commercial files, there is sufficient in the public sector to keep geologists arguing and theorizing for years to come.

See Also

Europe: Variscan Orogeny. **Famous Geologists:** Hutton; Lyell; Steno. **History of Geology From 1780 To 1835**. **Seismic Surveys**. **Sequence Stratigraphy**. **Stratigraphical Principles**. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Sanders JE (1957) Discontinuities in the stratigraphic record. *New York Academy of Science Transactions, Series 2,* 19: 287 297.

Sloss LL (1984) The greening of stratigraphy, 1933 1983. *Annual Reviews of Earth and Planetary Sciences* 12: 1 10.

Tomkeieff SI (1962) Unconformity an historical study. *Proceedings of the Geologists' Association* 73: 383 416.

Vail PR, Mitchum RM, Todd RG, *et al.* (1977) Seismic stratigraphy and global changes of sea level. In: Payton CE (ed.) *Seismic Stratigraphy Applications to Hydro carbon Exploration,* pp. 49 212. Memoir 26. Tulsa: American Association of Petroleum Geologists.

Van Wagoner JC, Posamentier HW, Mitchum RM, *et al.* (1988) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hastings BS, Posamentier HW, *et al.* (eds.) *Sea Level Changes An Integrated Approach,* pp. 39 45. Special Publication 42. Tulsa: Society for Sedimentary Geology.

UNIDIRECTIONAL AQUEOUS FLOW

J Best, University of Leeds, Leeds, UK

Introduction

Unidirectional water flows are vital agents of erosion, transportation, and deposition in many Earth surface environments and can occur in a wide variety of depositional settings from continental rivers to flows in the deep sea. Unidirectional flows move in one principal direction, with no time-averaged reverse flows within the depth-averaged fluid and, apart from any local deviations caused by bed topography, experience no reverse or oscillatory motion, such as may be produced by waves and tides. Unidirectional flows can be either uniform, where the flow does not vary in velocity or cross-sectional area along its path, or non-uniform, where the fluid velocity and cross-sectional area do change spatially. Non-uniform flows show convective acceleration, where the cross-sectional area decreases and velocity increases, or convective deceleration, where the cross-sectional area expands and the flow slows. In addition to this spatial change in flow properties, unidirectional flows may vary temporally in their behaviour. Flows that show no temporal change in their behaviour are termed steady, whereas those whose velocity changes over time are termed unsteady. Unsteady flows show temporal increases and decreases in velocity, which are often related to the passage of a discrete event such as a flood.

Water flowing over a boundary, whether solid or mobile, develops a flow structure that depends on the velocity and depth of the fluid together with its density and viscosity. The surface over which the fluid moves exerts a frictional drag on the flow, and the region of flow near the bed that is retarded by this friction is termed the boundary layer. Minor friction at the upper atmospheric interface can also cause a small decrease in velocity at the top of the flow in open channels, whereas in unidirectional flows that have solid boundaries all around the flow (such as flow in ice-covered channels or flows through conduits and pipes) significant boundary layers develop from all surfaces. Additionally, unidirectional flows that propagate within another fluid, such as unidirectional density currents, experience significant mixing at their upper boundary owing to shear at this surface. The boundary-layer structure near the surface (or wall) generates a stress on the bed, which initiates and causes sediment transport and ultimately the development of bed morphology. However, the sediment in transport and exact shape and nature of the topography also exert significant feedbacks upon the flow.

Flow Types

Unidirectional flows may be either laminar or turbulent. Laminar flows are dominated by viscous forces rather than the inertial forces acting on the fluid, whereas turbulent flows are dominated by inertial forces. The laminar or turbulent state of flow is expressed by the Reynolds number, Re, where

$$\mathrm{Re} = \frac{\rho \overline{u} Y}{\mu} = \frac{\textit{inertial forces}}{\textit{viscous forces}}$$

and ρ is fluid density, $\overline{u}$ is a characteristic velocity of the flow (such as the depth-averaged mean downstream velocity), Y is a characteristic length scale (such as the flow depth), and μ is the molecular viscosity of the fluid. Flows are termed laminar when $\mathrm{Re} < 500$ and any mixing that occurs is on a molecular scale, turbulent when $\mathrm{Re} > \mathrm{O}2000$ and mixing occurs through the action of turbulent eddies or coherent flow structures at various scales, and transitional when $500 < \mathrm{Re} < 2000$. Many aqueous flows are fully turbulent in their behaviour, although it should be remembered that, owing to their small size, many organisms living in unidirectional flows may experience the overwhelming effects of viscosity and thus live in laminar worlds. Additionally, changes to the turbulent nature of flow can be caused by increasing concentrations of fine suspended sediment, which may modify the nature of the velocity profile and mechanisms of turbulence generation.

Additionally, unidirectional flows are significantly affected by the relative influence of gravitational forces compared with inertial forces, as expressed through the Froude number, Fr.

$$\mathrm{Fr} = \frac{\overline{u}}{\sqrt{gY}} = \frac{\textit{inertial forces}}{\textit{gravitational forces}},$$

where g is acceleration due to gravity. Flows are termed subcritical when $\mathrm{Fr} < 1$, supercritical when $\mathrm{Fr} > 1$, and critical when $\mathrm{Fr} \cong 1$. This dimensionless number expresses the relative celerity of a gravity wave ($\sqrt{gY}$) on the flow: subcritical flows are able to experience the upstream effects of the wave, which has a velocity greater than the flow velocity, whereas the effects of the wave are felt only downstream for

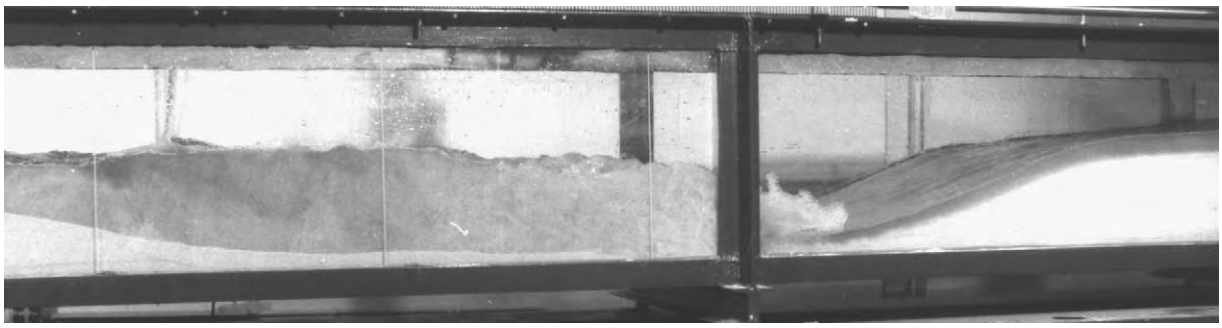


Figure 1 Flow in a laboratory channel, showing the transition from supercritical flow (Fr > 1, right) to subcritical flow (Fr < 1, left) through a hydraulic jump. Image courtesy of John Bridge.

supercritical flows. The form of the Froude number is slightly different for unidirectional density-current flows, where one fluid flows into another, as the effects of reduced gravity must be taken into account. However, the behaviour of the flow, and especially the interaction between the bed and the flow surface and hence the nature of energy loss within the flow, is strongly linked to the Froude number, as exemplified by the transition from supercritical to subcritical flow through a hydraulic jump (Figure 1).

In addition to these properties of flow, unidirectional aqueous flows can show a range of behaviours dependent on their rheology, or how their internal rate of strain responds to an applied external stress (Figure 2). Newtonian fluids, such as pure water, show a linear relation between applied shear stress and strain rate, and hence their viscosity is invariant with respect to the applied stress. Non-Newtonian fluids, however, do not behave in this manner, and either strain rate changes with applied stress (pseudoplastic and dilatant behaviours; Figure 2) or there is an initial yield stress with a subsequent linear stress–strain relationship (Bingham plastic; Figure 2). Although the majority of aqueous flows behave in a Newtonian manner, the addition of significant quantities of fine sediment can cause a change in behaviour and result in flows that have non-Newtonian characteristics, eventually leading to mud or debris flows in which the percentage of water may be very low and the rheology markedly non-Newtonian.

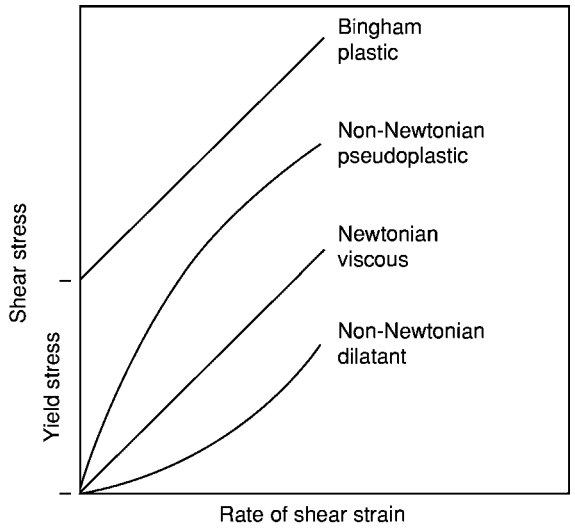


Figure 2 The relationship between the shear stress applied to a fluid and its strain rate, illustrating the various types of behaviour. The viscosity of the fluid is given by the slope of the line.

Velocity Profiles and Boundary Layer Structure

Unidirectional aqueous flows moving over a solid impermeable surface develop a distinct velocity profile away from the wall (Figure 3), and the boundary layer extends into the outer flow until the effects of wall friction become minimal (where the velocity at a point is approximately 95% of the maximum velocity

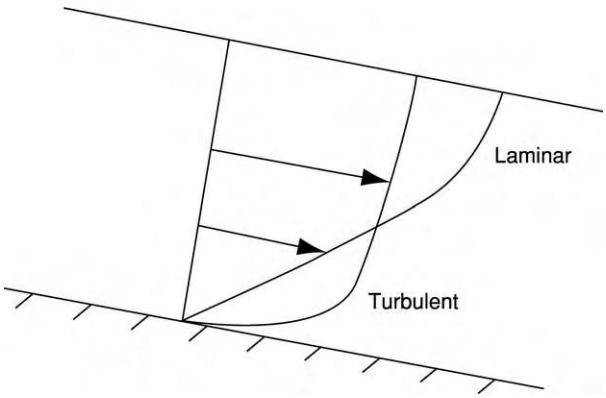


Figure 3 Comparison of a laminar flow and a turbulent flow over an impermeable flat surface. Both flows have the same mean discharge through a cross section, but they adopt different velocity profiles. (After Middleton GV and Southard JB (1984) *Mechanics of Sediment Movement*. SEPM Short Course Notes 3. Tulsa: Society of Economic Paleontologists and Mineralogists).

in the profile). Turbulent flows, since they possess appreciable mixing between adjacent fluid layers, have a steeper velocity gradient near the wall than do laminar flows (Figure 3). The stresses generated in a laminar flow, τ, are a product of the diffusion of momentum within the fluid on a molecular level and can be expressed as

$$\tau = \mu\left(\frac{\partial \overline{u}}{\partial y}\right)$$

where $\overline{u}$ is the time-averaged velocity at a point and y is the vertical height above the boundary.

However, for turbulent fluids the flow at each point can be broken down into a mean flow velocity, $\overline{u}$, and the deviation from that mean, u' (i.e. $u = \overline{u} + u'$ for the downstream component of the flow). This decomposition of the turbulent-flow signal can be applied to all three components of velocity (i.e. $u = \overline{u} + u'$, $v = \overline{v} + v'$, and $w = \overline{w} + w'$, with u, v, and w denoting the downstream, vertical and spanwise components of velocity, respectively, in the x, y, and z directions). The stresses within turbulent flows are thus a function of both transfer of momentum on a molecular level (the viscous shear stress) and mixing caused by the movement of turbulent eddies within the flow, such that shear stress within a turbulent flow, τ, is given by

$$\tau = (\mu + \eta)\frac{\partial \overline{u}}{\partial y}$$

where η is the so-called eddy viscosity.

The effect of turbulent mixing is that packets of low-momentum fluid from near the bed may be mixed upwards in the flow and relatively faster parcels of fluid from higher in the flow may be carried downwards towards the bed. This exchange of downstream momentum is expressed by $-\rho u'$, and, for a parcel of fluid moving upwards or downwards in the flow (considering the x–y plane with u and v as the components of velocity), the rate of change of downstream momentum through a given area is expressed by $-\rho u'v'$. Hence, the time-averaged shear stress in a turbulent flow can be given by

$$\tau = \mu\left(\frac{\partial \overline{u}}{\partial y}\right) - \rho\overline{u'v'}.$$

Similar expressions can be written for the shear stresses exerted by the combination of the velocity fluctuations between the three components of flow, and these are termed the Reynolds stresses.

The velocity profile of a turbulent boundary layer developed over a smooth surface can be divided into several distinct regions (Figure 4): the viscous sublayer, in which the influence of viscosity is dominant but in which some turbulent eddies are initiated (there is a linear increase in velocity with height within the viscous sublayer); the turbulence-generation layer (or buffer zone), which is the region of largest velocity gradient and where the majority of turbulence is generated, which is characterized by a logarithmic velocity distribution with height above the bed; and an outer zone, which is characterized by the dissipation of turbulent eddies generated near the boundary.

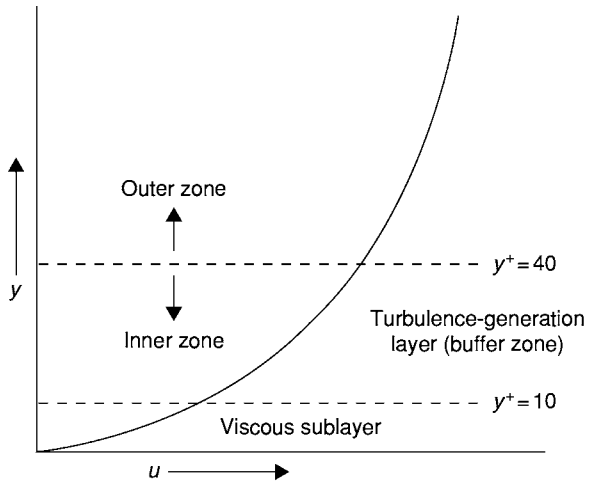


Figure 4 The different regions of a turbulent boundary layer.

The Estimation of Boundary Shear Stress

A key aim of understanding and quantifying boundary-layer structure is to enable prediction of the boundary shear stress exerted on the wall, in order that erosion thresholds and sediment transport can be estimated. Six methods are commonly used to estimate the mean value of the shear velocity, u_*, which is related to the boundary shear stress, τ_B, by $\tau_B = \rho u_*^2$.

1. The slope method uses the slope of the water surface, S, and flow depth, Y, such that $u_* = gYS^{0.5}$.
2. The best fit to the linear velocity profile within the viscous sublayer, $u^+ = y^+$, where $u^+ = U/u_*$ and $y^+ = yu_*/v$, where U is the mean velocity at a point and v is the kinematic viscosity of the fluid (μ/ρ).
3. The best fit to the logarithmic 'law of the wall', which describes the logarithmic shape of the downstream velocity profile in the lower part of a flow (up to the top of the turbulence-generation layer), such that $u^+ = (1/\kappa)\ln(y^+) + C$, where κ is the von Kármán constant (*ca.* 0.40 in clear-water flows) and C is a function of the roughness of the bed.

4. Using linear extrapolation of the Reynolds stress profile to the bed at a height of $y = 0$, such that $-(\overline{u'v'})/u_*^2 = 1 - (y/Y)$.
5. Adopting the spectral method, which uses estimated values of the turbulent dissipation rate, ε, in the inertial region of the spectral domain, such that $u_* = (\varepsilon\kappa y)^{1/3}$.
6. Using the normalized vertical flux of turbulent kinetic energy, which has been proposed to adopt a universal value irrespective of wall roughness, such that $0.5q^2\nu/u_*^3 \approx 0.30$, where q is the turbulent kinetic energy $(q = 0.5(\overline{u'}^2 + \overline{v'}^2 + \overline{w'}^2))$.

The Structure of Turbulent Boundary Layers

The nature of mixing within a turbulent boundary layer depends on the exact nature of the turbulent eddies, or coherent flow structures, that are present within the flow. These coherent flow structures are generated within the viscous sublayer and turbulence-generation region, and can be investigated through their velocity signatures and their temporal and spatial length scales. If a two-dimensional flow is considered, four quadrants of flow behaviour can be defined based on the deviations of the u and v components of flow (downstream and vertical velocities) from their respective mean values (Figure 5). This allows recognition of regions of relatively slow downstream-momentum fluid moving upwards within the flow (quadrant 2 events; Figure 5); relatively fast downstream-momentum fluid moving downwards within the flow (quadrant 4 events; Figure 5); and outward and inward interactions of flow (quadrants 1 and 3 events, respectively; Figure 5). This simple quadrant analysis has been widely used in turbulent-boundary-layer research and is appropriate if the flow is largely two dimensional; in fully three-dimensional flows, all three components of velocity should be considered, and octant analyses may be required to characterize truly the fluctuations in fluid flow.

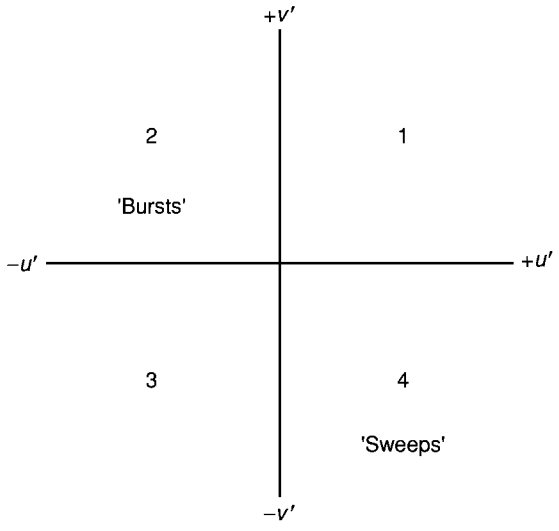


Figure 5 Quadrant classification of a turbulent flow, according to the deviations from the mean values of downstream (u') and vertical (v') velocity.

Much research over the past 40 years has been devoted to elucidating the form of coherent flow structures within turbulent boundary layers, linking these structures to their velocity signatures, and assessing their contributions to both the Reynolds stresses and turbulent kinetic energy budget. Studies have progressed from largely qualitative flow visualizations to quantitative measurements and more recent numerical simulations. Coherent flow structures within a flat-bed turbulent boundary layer are principally composed of:

- low-speed streak areas in the region $0 < y^+ \leq 10$ (where $y^+ = yu^*/\nu$), which are areas of relatively low downstream velocity that are aligned parallel to the flow and form a series of spanwise areas separated by regions of slightly higher flow velocity,
- ejections of low-speed fluid away from the wall (quadrant 2 events; Figure 5),
- sweeps of relatively high-momentum fluid towards the wall (quadrant 4 events; Figure 5), and
- vortical structures of several different kinds, including larger-scale structures, which may occupy a significant fraction of the flow depth and could be generated by amalgamations of smaller groups of vortices originating near the wall.

These structures have been visualized and modelled as a series of longitudinal vortices near the bed, which link through to the legs of 'horseshoe'-, 'hairpin'-, or 'arch'-shaped vortices higher in the flow, which constitute the ejections (quadrant 2 events) (Figure 6). The majority of the Reynolds stresses, approximately 70% in a smooth-wall boundary layer, may be linked to these quadrant 2 and quadrant 4 events, which may be critical in both the suspension of sediment (quadrant 2 events or 'bursts') and the entrainment of sediment as bedload (quadrant 4 events), although several studies have also highlighted the significance of quadrant 1 and quadrant 3 events. The transition from laminar to turbulent flow also appears to be linked to the generation of small 'packets' of vortices near the wall, which first appear as 'turbulent patches' near the bed (Figure 7). These 'patches' or 'turbulent spots' bear a striking resemblance to the groupings of hairpin vortices generated in fully turbulent boundary layers, which extend through a significant percentage of the flow depth.

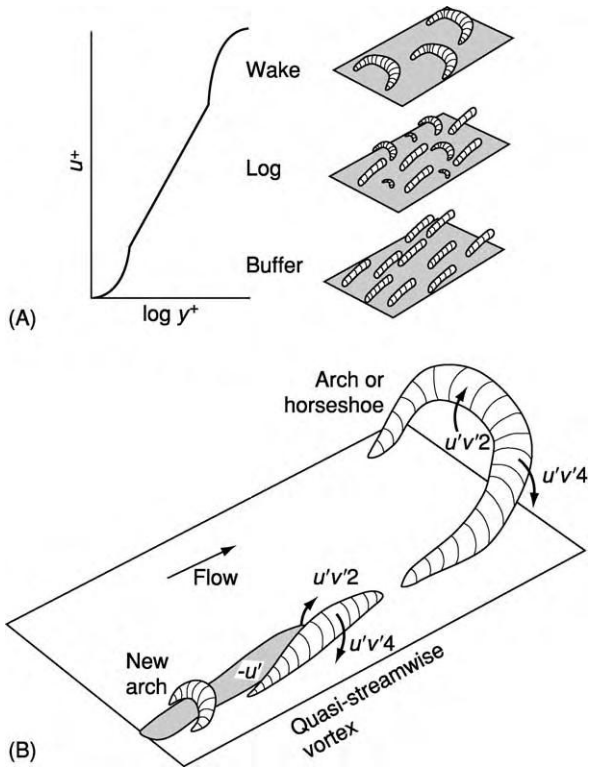


Figure 6 (A) Idealized model of populations of vortices in dif ferent regions of a turbulent boundary layer. (B) Schematic model of the links between ejection and sweep motions and streamwise vortices and 'arch' or 'horseshoe' shaped vortices in a turbu lent boundary layer 2 and 4 refer to quadrant 2 and 4 events (see Figure 5); Reproduced from Robinson SK (1991) Coherent motions in the turbulent boundary layer, *Annual Review of Fluid Mechanics* 23: 601 639.

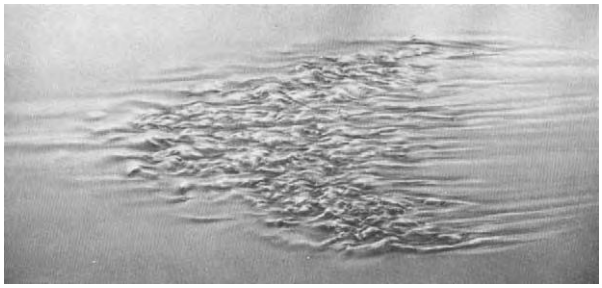


Figure 7 Photograph of a developing 'turbulent spot' at the laminar turbulent transition, as visualized by transition, as visu alized by reflective aluminum tracer particles at the base of the boundry layer and as viewed from above the flow. Flow is right to left. Reproduced from Cantwell B, Coles D, and Dimotakis P (1978) Structure and entrainment in the plane of symmetry of a turbulent spot, *Journal of Fluid Mechanics* 87: 642 672, with permis sion from Cambridge University Press.

Flow Separation

If a flow encounters a marked change in the gradient of the wall over which it is flowing or is subject to an adverse positive pressure gradient (for example, as a result of injection of fluid through a porous wall), then the fluid may be forced away from the wall and separate from the boundary, with subsequent re-attachment of the flow to the bed at some distance downstream, creating a zone of recirculating, or separated, flow near the bed (Figure 8). Flow separation is a key process under most unidirectional water flows and frequently occurs at both positive and negative steps or changes in bed topography. Such areas of flow separation are critical in many unidirectional flows, occurring at a range of scales from, for example, separation behind individual grains, bedforms (Figure 8), and bars to larger-scale features such as those associated with changes in channel

Figure 8 Flow separation visualized behind a dune bedform. Flow right to left. Flow is visualized by the path of neutrally buoyant particles within the water, which shows the recirculating flow within the dune leeside.

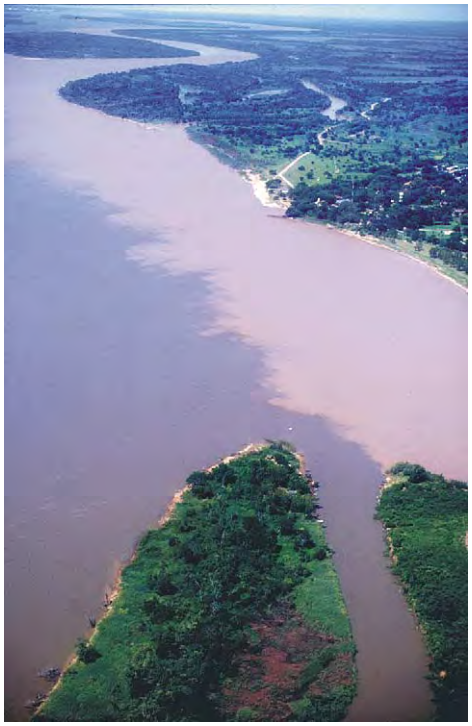


Figure 9 Large scale Kelvin Helmholtz instabilities generated along the mixing layer between the Rio Paraná (left, clearer water) and Rio Paraguay (right, higher sediment concentration), Argentina. Flow is away from the viewer, and the width of the image is approximately 1.5 km.

curvature, abrupt gradient changes at the edge of subaqueous slopes, and subaqueous topography. One key consequence of flow separation is that a steep velocity gradient and shear layer are generated between the separation zone and the faster free-stream fluid outside: large-scale coherent vortices, termed Kelvin–Helmholtz instabilities, are generated along this shear layer. Such large-scale vortices are highly turbulent and may be responsible for generating large instantaneous Reynolds stresses, which are often critical in erosion of the bed and sediment transport.

Free Shear Layers

In addition to shear layers associated with flow separation, zones of distinct differential velocity and rapid change in velocity may be present within the body of a unidirectional flow, owing to flow convergence around topography or in combining channels or to shear at the top of a subaqueous density current, for example. Turbulence and mixing across such 'free' shear layers depend on the velocity differential across the shear layer and the relative densities and viscosities of the two incoming flows, but these free shear layers often create large-scale Kelvin–Helmholtz instabilities (Figure 9), which dominate both fluid mixing and the instantaneous Reynolds stresses. For example, fluid mixing at channel confluences (Figure 9) has been shown to be greatly influenced by the shear-layer dynamics between the incoming flows, and interactions between free shear layer and bed topography are thought to control the downstream dispersal of suspended sediments and pollutants.

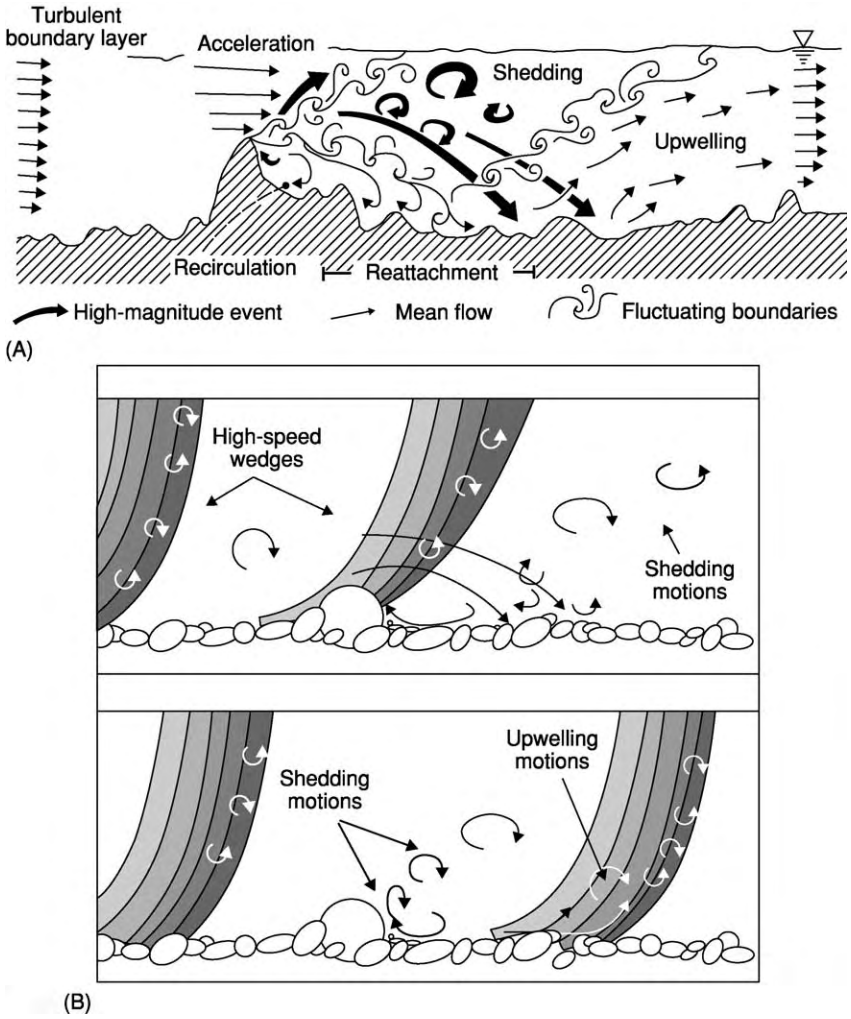


Figure 10 (A) Flow regions associated with the presence of a pebble cluster on the turbulent flow field. (B) Schematic diagram of the large scale flow structures proposed to develop over a rough gravel surface, showing (top) how the passage of a large scale high speed flow structure expands the flow separation zone in the lee of the clast and (bottom) how this generates vortex shedding from the separation zone and upwelling in the region of flow reattachment. Both A and B are reproduced from Buffin Belanger T and Roy AG (1998) Effects of a pebble cluster on the turbulent structure of a depth limited flow in a gravel bed river. *Geomorphology* 25: 249 267.

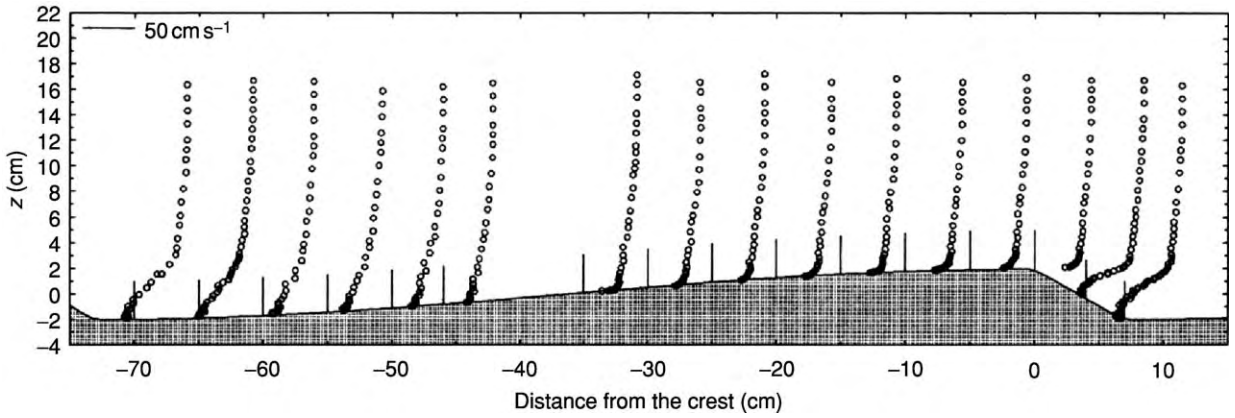


Figure 11 Velocity profiles measured over a dune bedform, showing flow separation in the dune leeside and boundary layer recovery over the stoss side of the next downstream dune. Flow left to right. Reproduced from Nelson JM, McLean SR, and Wolfe SR (1993) Mean flow and turbulence fields over two dimensional bedforms. *Water Resources Research* 29: 3935 3953, with permission from American Geophysical Union.

Other Factors Influencing Boundary Layer Structure

In many unidirectional aqueous flows, the precise nature of the mean and turbulent flow is influenced by a range of variables that can significantly alter the flow structure, bed shear stress, patterns of sediment transport, and, hence, development of bed morphology. Some of the most significant influences on the characteristics of unidirectional aqueous flows are described below.

The Nature of Bed Grain Roughness

Particle roughness significantly increases the potential for turbulent mixing near the bed and often results in an increase in the gradient of the near-bed velocity profile, with a concomitant increase in the bed shear stresses derived from the velocity gradient, Reynolds stress, or turbulent kinetic energy budget. Grain roughness may destroy the viscous sublayer and also increase the generation of turbulence near the bed, through either encouraging intensified bursting (quadrant 2 events) from between the grains (and thus larger-scale return quadrant 4 events) or generating regions of flow separation around individual grains or groups of particles, which may both create significant velocity gradients near the bed and generate large-scale coherent vortices, associated with flow separation both in front of and behind the particles, which can penetrate the entire flow depth (Figure 10).

The Presence and Type of Bedforms

Many bedforms such as ripples, dunes and larger-scale bar forms, create their own flow field through topographic, convective accelerations and decelerations of fluid that may significantly change the nature of a unidirectional flow. In addition, many bedforms are also associated with flow separation on their upstream stoss side or downstream lee side, which generates appreciable turbulence and a boundary layer that is recovering from flow separation downstream of the region in which the flow reattaches to the bed (Figure 11).

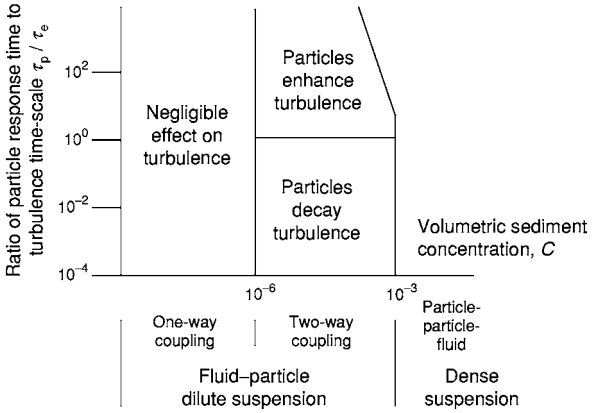


Figure 12 Schematic diagram of the attenuation or enhancement of turbulence due to the presence of sediment in a flow as a function of the volumetric sediment concentration, C, and the ratio of the particle response time to the turbulence time scale, τ_p/τ_e (after Elghobashi S (1994) On predicting particle laden turbulent flows. *Applied Scientific Research* 52: 309 329).

The Type and Quantity of Suspended Sediment

Many turbulent flows transport appreciable quantities of suspended sediment, with the suspended concentration in some flows reaching levels at which the flows become markedly non-Newtonian, such as in the Huanghe River in China, where concentrations of up to 1290 kg m ^{3} have been recorded. Suspension

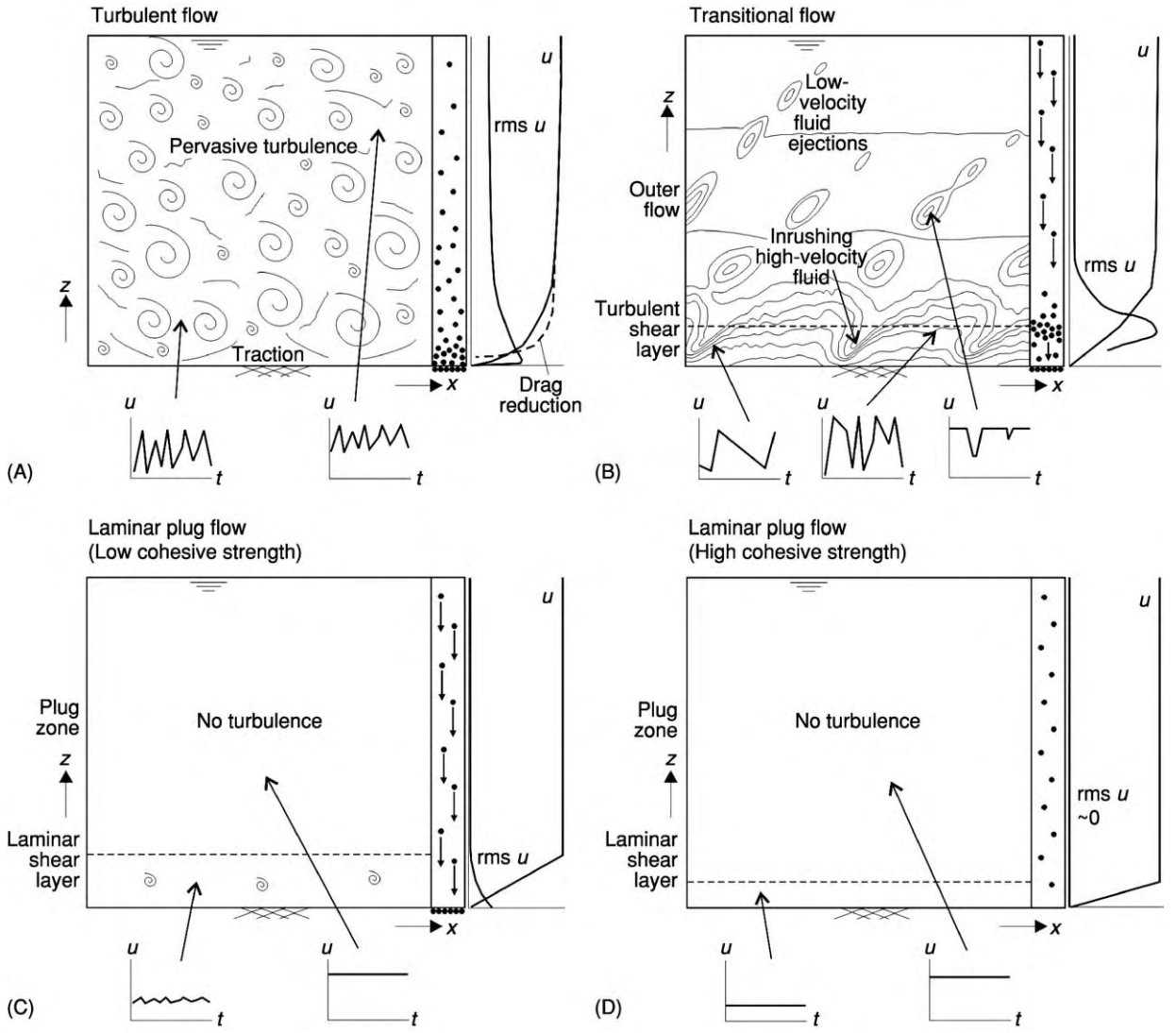


Figure 13 Schematic model for flows with increasing clay concentrations. Each model depicts the characteristic velocity profile and nature of flow, a vertical profile of the rms *u*, representative time series at several heights, and a representation of sediment settling. rms refers to the root mean square value of the downstream velocity component, or level of turbulence, within the flow, and indicates the principal regions of turbulence generation within each profile. (A) Turbulent flow with a logarithmic velocity profile and turbulence generation near the bed. At low clay concentrations, drag reduction may begin to occur. Coarser sediment is supported through turbulence, and sedimentary structures can develop. (B) Transitional flows with a developing shear layer, which separates a lower region of high velocity gradient from an upper layer of reduced shear. Sketch of flow depicts streamlines. Turbulence is strongest in the shear layer, along which Kelvin Helmholtz instabilities are developed with a distinctive velocity signature (see inset sketches). Sediment entrained into the basal region is trapped, and parallel lamination may be produced by the variable shear stresses induced by the shear layer instabilities. (C) Laminar plug flow without turbulence and with low cohesive strength. The cohesive strength of the flow is unable to support coarser sediment, which settles to the bed. (D) Laminar plug flow with high cohesive strength is able to support coarser sediment suspended within the flow. Reproduced from Baas JH and Best JL (2002) Turbulence modulation in clay rich sediment laden flows and some implications for sediment deposition. *Journal of Sedimentary Research* 72: 336 340. SEPM (Society for Sedimentary Geology).

of sediment requires turbulence within the flow, but a feedback is exerted where at some point the turbulence begins to be modified by the sediment in suspension. This complex feedback mechanism is poorly understood, with factors such as the concentration of sediment and ratios of grainsize: turbulent length and time scales thought to be important in causing either a decrease or increase in turbulence (turbulence attenuation and enhancement respectively) within the flow (Figure 12). Turbulence modulation is thus a key feedback mechanism within many unidirectional flows, acting to both enhance and suppress turbulence production, and is also known to change the downstream velocity profiles significantly with subsequent implications for sediment transport and sorting (Figure 13).

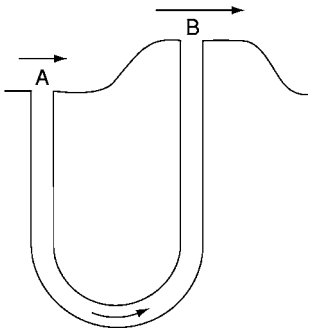


Figure 14 Flow acceleration above the raised limb of a burrow creates a lower pressure at B than at A, inducing flow within the tube (after Vogel S (1994) *Life in Moving Fluids*. Chichester: Princeton University Press).

The Porosity of the Bed Surface

Many studies of unidirectional flows have assumed that the bed is impermeable and that the subsurface flows exert little influence on the overlying boundary layer. However, this assumption is invalid, and the flow within porous beds, which comprise the surface of most sedimentary environments, can significantly alter the nature of the overlying boundary layer. This is especially true in the presence of bed morphology, which may generate differential velocities and fluid pressures around the topography. Such flow is seen, for instance, out of relict burrows that have a raised rim (Figure 14), where flow through the burrow and towards the area of raised topography results from lower fluid pressure associated with increased velocities at this raised opening. However, flow within porous beds may also lead to modification of the near-bed velocity profiles, and it has been suggested that the presence of a porous bed can decrease near-bed velocities and cause the velocity profile to deviate from a logarithmic form, with turbulence penetrating the top part of the porous bed.

See Also

Sedimentary Environments: Alluvial Fans, Alluvial Sediments and Settings; Deserts; Storms and Storm Deposits. **Sedimentary Processes:** Erosional Sedimentary Structures; Depositional Sedimentary Structures; Particle-Driven Subaqueous Gravity Processes; Deposition from Suspension; Fluxes and Budgets.

Further Reading

Ashworth PJ, Bennett SJ, Best JL, and McLelland SJ (1996) *Coherent Flow Structures in Open Channels*. Chichester: John Wiley and Sons.

Baas JH and Best JL (2002) Turbulence modulation in clay rich sediment laden flows and some implications for sediment deposition. *Journal of Sedimentary Research* 72: 336 340.

Bridge JS (2003) *Rivers and Floodplains*. Oxford: Blackwell Publishing.

Buffin Belanger T and Roy AG (1998) Effects of a pebble cluster on the turbulent structure of a depth limited flow in a gravel bed river. *Geomorphology* 25: 249 267.

Cantwell B, Coles D, and Dimotakis P (1978) Structure and entrainment in the plane of symmetry of a turbulent spot. *Journal of Fluid Mechanics* 87: 641 672.

Elghobashi S (1994) On predicting particle laden turbulent flows. *Applied Scientific Research* 52: 309 329.

Leeder MR (2000) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Publishing.

López F and García MH (1999) Wall similarity in turbulent open channel flow. *Journal of Hydraulic Engineering* 125: 789 796.

Middleton GV and Southard JB (1984) *Mechanics of Sediment Movement*. SEPM Short Course Notes 3. Tulsa: Society of Economic Palaeontologists and Mineralogists.

Nelson JM, McLean SR, and Wolfe SR (1993) Mean flow and turbulence fields over two dimensional bedforms. *Water Resources Research* 29: 3935 3953.

Nezu I and Nagkagawa H (1993) *Turbulence in Open Channel Flows*. Balkema: International Association for Hydraulic Research.

Pope SB (2000) *Turbulent Flows*. Cambridge: Cambridge University Press.

Robinson SK (1991) Coherent motions in the turbulent boundary layer. *Annual Review of Fluid Mechanics* 23: 601 639.

Van Rijn LC (1990) *Principles of Fluid Flow and Surface Waves in Rivers, Estuaries, Seas and Oceans*. Oldemarkt: Aqua Publications.

Vogel S (1994) *Life in Moving Fluids*. Chichester: Princeton University Press.

Williams J (1996) Turbulent flow in rivers. In: Carling PA and Dawson M (eds.) *Advances in Fluvial Dynamics and Stratigraphy*, pp. 67 125. Chichester: Wiley and Sons.

URALS

See **EUROPE: The Urals**

URBAN GEOLOGY

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

Most of the world's population lives in relatively crowded conditions in urban areas, affording them immediate contact with all forms of sustenance. At the same time, these teeming populations require huge imports of potable water, treatment of sanitary wastes, and export of solid, special, and hazardous wastes, along with a degree of infrastructure that impinges on and relies heavily on the constraints represented by the geological setting. Four major geological themes govern the application of geology to human life in cities and urban centres. First, although an abundance of regional geological information is applicable to urban development and life, the integration of most of such data is not readily discernible by the majority of the regional population. Second, urban life is concentrated such that there is 'loading' of the geological environment under various types of 'footprints' of engineered structures. Third, as the trend of importance of urban life for people continues to expand, 'megacities' become the centrepiece of new form of urban geology. Last, the concentration of urban populations in coastal regions interfaces with growing concerns over sea-level rise and global climate changes.

Cities historically have grown and developed around geological core areas where geological conditions were favorable to defense or security of construction. Although the form of these settings may seem geographical in nature, it is the underlying geology that has created such conditions. The following geological situations and the cities with which they are associated exemplify this:

- Natural, hard-rock sheltered seaports (Plymouth, England; Hong Kong; San Francisco, California; New York City).
- Confluences of major rivers (Pittsburgh, Pennsylvania).
- Mouths of major navigable rivers (Alexandria, Egypt; New Orleans, Louisiana; Para, Brazil).
- Heads of navigation of major rivers (Minneapolis–St. Paul, Minnesota).
- Confluences of rivers and pre-railroad trails (Paris; Rome).
- Defensive positions, underpinned by bedrock (Seoul; Rome).
- Sea-lane confluences for early trade (Singapore; Capetown, South Africa; Boston, Massachusetts).
- Confluences of pre-railroad trails (Kansas City, Kansas; Santa Fe, New Mexico; Edmonton, Canada).
- Mouths of mountain passes (Salt Lake City, Utah; Denver, Colorado; Reno, Nevada; Missoula, Montana).

Geoscience is a major potential contributor to maintenance of the health and welfare of cities and their populations. Successful implementation requires planning parameters for growth and redevelopment of the built environment and assessments of least-impact and least-cost alignments (e.g., the constantly needed improvements for rapid transit to move people around and in and out of cities). Geoscience plays an important role in the location, development, and delivery of potable water supplies and in the effective disposal of wastewater effluents. Previously used land must be characterized with respect to toxic contamination and its remediation, and risk from death, injury, and property losses stemming from geological hazards must be mitigated. Geological science is critical to understanding the potential for flooding, earthquakes, volcanic eruptions, ground collapse, mass movements, and seismic sea waves in densely populated areas. In the twenty-first century, yet another application of geoscience is in the consideration of how certain geological conditions might serve to enhance acts of terrorism.

Geological Influences on Urban Development

Modern engineered works in the urban environment represent relatively large and high-magnitude impacts on the substrate on which they are built. Geological site characterization of regional earth materials is thus an absolute necessity before laying the foundation for any large engineered structure, for economic reasons (i.e., to reduce construction costs) as well as for successful structure operation and maintenance and for basic environmental acceptability (Figure 1). Most of the high-impact concerns for site characterization are devoted to a few simple themes. The physical nature of foundation earth materials (soil, weak rock, or rock) must be identified as being able to support engineered works without unacceptable structural deformation or outright collapse. Geological anomalies that might compromise the integrity of building components or entire engineered work must be detected and delimited. Furthermore, 'bad ground' that would require premium foundations or difficult construction efforts and higher building costs must be recognized and presented to planners (Figure 1).

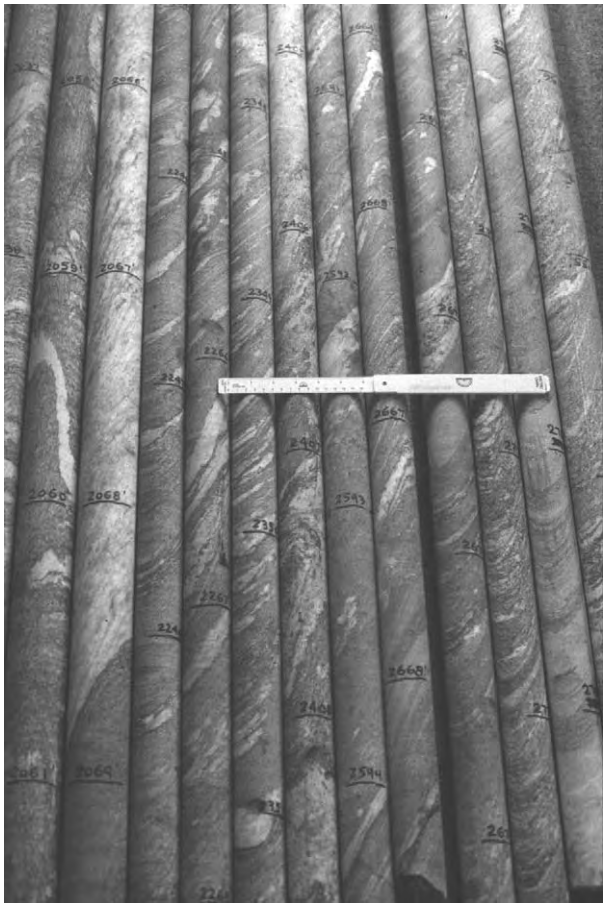


Figure 1 NX size (2.125 inches) rock core recovered by oriented, triple tube coring technique from more than 650 m below Kennedy International Airport, Long Island, New York City, at the site of a proposed liquefied natural gas storage cavern planned by the Brooklyn Union Gas Company (now Key Span Energy Co., Inc.). The rock is complexly mixed Fordham Gneiss and Manhattan Schist and these are select segments of 3 m lengths, all unbroken by natural discontinuities. The photograph shows the generally excellent foundation and deep cavern characteristics of New York City bedrock. Ruled portion of scale is 15 cm in length; depth marks on core are in metres. Photograph by the author.

Cities of The World Literature Series

For more than 20 years, the Association of Engineering Geologists (AEG) has fostered the incorporation of geoscience within urban development by publishing an international series of papers (*Cities of The World*) dedicated to the memory of the outstanding efforts of the late Canadian geological engineer, Robert F Legget (1904–94). The quarterly journal of the AEG (now a joint effort with The Geological Society of America), *Environmental and Engineering Geology*, seeks and publishes detailed accounts of urban geology, utilizing a standard format (Table 1; see also www.aegweb.org) that serves as a codification of the important elements of urban geology.

Problematic Conditions of Urban Construction

Construction in the built environment has to address a variety of geological 'constraints' that may be hidden from view during general observation of the existing ground surface. Without accurate geological information at the planning and design stages, new construction and urban renewal efforts will almost certainly encounter cost overruns, regulatory compliance infractions, and some type of construction failure. Three primary geological considerations (soil, groundwater, and geological dicontinuities) are integral to avoiding such difficulties. The nature and thickness of soil units that will bear the dead and live loads transferred downward from the intended construction must be established. Likewise, the presence, depth, and potential fluctuation of the groundwater surface below the intended engineered works must be evaluated and related to needs for dewatering without detriment to stability of existing nearby engineered works. The soil and groundwater profiles may pose potential excavation problems when deep basements are required to accommodate vehicle parking. The presence of and adverse geometrical orientation of geological discontinuities (bedding, joints, shear zones, and faults, to name a few) may

Table 1 Standard elements of urban geological considerations for *Cities of the World* journal series[a]

Chapter	*Section*
1. Background	1.1 Location
	1.2 History of founding
	1.3 Geological influences affecting founding
2. Geological setting	2.1 Brief on regional geology
	2.2 Geology of the city
	2.2.1 Basement rock
	2.2.2 Surficial units (soils)
	2.2.3 Stratigraphic chart
3. Geotechnical characteristics	3.1 General foundation related geological units
	3.2 Exploration methods
	3.3 Typical foundation types in use
	3.4 General laboratory test methods
4. Materials	4.1 Traditional types and uses
	4.2 Sources and extraction methods
	4.3 Regulations and zoning affecting extraction
	4.4 Environmental impact of extraction
5. Geological constraints	5.1 Classification
	5.2 Recurrence
	5.3 Mitigation
6. Historic resource extraction	6.1 History
	6.2 Classification of extracted ground; mines and fluids (water, oil, gas)
	6.3 Areal extent
	6.4 Constraints related to extracted ground
	6.5 Mitigation of extracted ground threats
7. Seismicity of the city	7.1 Historic record
	7.2 Notable events
	7.3 Generalized recurrence interval
	7.4 Ground motion amplification factors
	7.5 Seismic design provisions in force
8. Environmental concerns	8.1 Water supply
	8.2 Wastewater treatment
	8.3 Waste management (solid, special, and hazardous)
	8.4 Remediation of uncontrolled wastes
	8.5 Wetlands factor
9. Major engineered structures (tabulated)	Detail as appropriate
10. Use of underground space	
11. Summary	11.1 Conclusions
	11.2 Predictions for the future
References	As appropriate
Illustrations (key illustrations of the geological situation of the city)	Frontispiece (color oblique emphasizing major geological features)
	Index map
	General geological planimetric map
	Stratigraphic column
	Geotechnical cross section
	Seismicity plot
	Optional photographs

[a]Content and format recommendations of the Association of Engineering Geologists for papers submitted for publication in the series dealing with continuing development of the world's cities. Papers are published in the *Environmental and Engineering Geoscience Journal*.

affect the stability of basement excavations and the integrity of surrounding buildings and other structures. These aspects of engineering geology require the geological team to interface with urban historians, archaeologists, architects, urban planners, insurers, financiers, and others related to the design and construction processes (Tables 2 and 3).

Role of the Engineering Geologist

Geological information is critical to the siting, design, and construction of all engineered works. This is particularly important in the urban environment, where all physical aspects of construction are compounded in their effects by mandates in scheduling

Table 2 Geotechnical influences on urban construction

Geotechnical influence	*Emphasis*	*Key considerations*
Site geological knowledge reduces the risk of unknown ground conditions	Unknown, undetected, or undisclosed geological conditions can compromise scheduling, cost, and operational performance of engineered works	Capacity of foundation soils to support loads of the structure; stability of surrounding facilities; excessively 'weak' or 'strong' ground exacerbates the construction effort
Some earth materials have undesirable properties or characteristics	Avoidance of constraints or slowdowns to construction process	Detection in the site exploration process, notification of owner and design engineer; incorporation into the design and construction specifications
'Faces' across which geologic character of construction ground changes	When not anticipated by the contractor, can cause unwanted perturbations in schedule and in project cost; where spanned by one 'bay' (segment) of construction, may lead to unacceptable differential settlement of portions of the foundation	Avoidance requires adequate site characterization, funding by the owner, and judicious selection of geological and geotechnical support consultants
Pockets or zones of 'bad' ground	Three dimensional bodies of degraded earth material not able to support design, construction, or operational efforts or roles of project	Generally related to geomorphic or tectonic considerations of origin and may have characteristics detrimental to construction or of performance of the facility
Near surface groundwater	Always a problem	Generally interferes with construction, particularly in placing the foundation; may require dewatering, which may affect performance of surrounding existing foundations
Perched water or groundwater is generally detrimental to the construction process	Control and removal without impairment of construction or with performance of completed facility	When truly 'perched', drains into the construction excavation within hours and does not replenish
Nature of site preparation or construction 'spoil' (soil) or 'muck' (tunnel spoil)	Must be removed from the construction site and reused in some worthwhile manner acceptable to the community	Spoil and muck have geotechnical characteristics that must be heeded in considering their reuse

Table 3 Engineering geological contribution to urban geology

Contribution	*Purpose*	*Important considerations*
Stratigraphy	Define the nature and bounds of soil types and of geological formational units	Controls the suitability and relative cost of siting and dimensional design of virtually all projects
Engineering properties of the foundation soil or rock	Must be capable of bearing the combined live and dead loads of the engineered works to be constructed	Acceptability measured in terms of nil compressibility and sufficiently high shear strength to support the loads of the project
Geological structure of bedrock exposed in construction	Locate and define fault displaced geological units and discontinuities that are of sufficient length of exposure to cross any one dimension of site excavation	Avoidance of adverse geological structures, premium foundation conditions, and expensive or 'bad' ground in terms of underground construction
Occurrence of groundwater	Protection from damage by human activities; avoidance of premium costs for foundations	In no way can groundwater be of beneficial consequence to the construction process
Surface water hydrology	Using geological evidence to define the nature of flooding as an economic and human welfare concern	Mainly involves interpretation of present geomorphic features that control the path, depth, and velocity of low frequency/high impact flood events
Avoidance of surrounding unstable ground	Protect from what may slip onto the site or move down slope from the site	Mainly ground that is unstable under gravitational and slope water conditions, along with rock falls
Avoidance of existing subsurface voids	Avoid presence of abandoned tunnels, mine shafts, mines in general, and natural voids such as karst caverns	All subsurface voids suffer from decreasing 'stand up' time, which is the ability to span loads imposed at the ground surface, and from the surrounding overburden weight of the geological column of materials

Table 4 Problems of urban construction: naturally troublesome geological conditions

Condition	*Impact*	*Geological considerations*
Soft and/or otherwise compressible soils	When detected will lead to engineering selection of medium to deep foundations, in turn leading to premium foundation costs	Afflicted typically by clay mineral rich soils such as glacial, glaciomarine, and glaciolacustrine clay soils, as well as a variety of slack water fluvial clays, silts, and 'muds'; learning to expect the presence of such conditions is based on regional and local geological knowledge
Certain marine silts and clayey silts are geotechnically 'quick' (unstable) by virtue of collapse prone 'card house' structure brought about by flocculated structure under saline depositional conditions	Can become unstable and can collapse under dynamic shock of earthquakes and a variety of man made shocks such as pile driving, blasting, and dropping of large loads	Where bordering river and stream banks and other steep terrain; type formation is the Leda Clay of Quebec Province and the marine clay soils of Scandinavia where encountered on land
Clay ('mud') 'plugs' formed as still water deposits of truncated river meanders	Creates unacceptable differential settlement of foundations, roads, and bridge abutments and piers	Endemic geotechnical problem in the Lower Mississippi River Valley of the United States
Organic soil of all types (notably peat)	Causes foundation deformation of all manner of engineered structures with all but the lightest bearing loads	Tends to have irregular but often oval to elliptical bounding contacts in the lateral sense; two type occurrences for this problem are the Newport Beach Fountain Valley areas of the Los Angeles basin, especially at the Pacific Ocean mouth of the ancestral Santa Ana River and the southern metropolitan area of Edmonton, Alberta, Canada
'Highs' in bedrock surface	Escalates costs of site preparation for construction	Sites must be 'brought to grade' as a preparation for installation of foundation; rock can cost ten times or more to remove, compared to soils
Buried zones of degraded bedrock	Increased costs of tunnels for water supply, sewerage, and rapid transit	Often the most difficult to anticipate; some are due to fault movement and displacement, others are due to geochemical cation exchange by water
Buried valleys	Channels of unconsolidated, porous earth material ('soils'), with preferential movement of groundwater	May place earth materials of vastly different engineering properties in the line of linear projects such as roads and tunnels, representing unanticipated design/construction conditions
Near surface groundwater	Calls for dewatering of basement excavations	Avoidance of disturbance of adjacent foundations and alteration of static ground surface, the latter leading to rot in older driven timber piles
Dumps of organic debris	Excessive settlement of foundations of successor buildings	Early residents routinely filled all topographic declivities with refuse, most of which is vertically compressible with time
Dumps of industrial (hazardous) wastes	Threats to health, construction workers, nearby residents, and occupants of the new facility	Mainly from low level, chronic emissions of toxic vapors from semivolatile and volatile organic compounds
Floodplain development	Unwise development of land subject to poorly predictable low frequency flooding	Geomorphic evidence of past flooding often crucial in defining the need for zoning against development
Proximity to seismically active faults and fault zones	The major seismotectonic zones are now well known throughout most of the world	Avoidance of high density human occupancy along known, 'active' fault traces and in areas of 'bad ground' subject to loss of foundation stability in large magnitude earthquakes
Proximity to major volcanoes	Ashfalls and lahars (volcanically induced mudflows)	In mild and humid climates, geochemical weathering rapidly produces volcanic soil, which attracts people because of cultivation advantages; geological evidence can predict rough levels of risk by time and location, for future exposure

and in costs of construction. The process by which geological input is discovered and provided to permitting officials, planners, owners, designers, and construction contractors is termed 'site characterization', and this task must include the discovery and description of man-made wastes (perhaps of archaeological/historical importance) left from previous site occupation. Site characterization as a concept generally arose in about 1900, but has taken on the particular meaning described here only since about 1980. It is a well-accepted practice in the field of engineered construction and environmental permitting.

Engineering geologists are generally tasked with planning and conducting site and waste characterizations, and the most important aspects related to the effort are experience, training, and professional competence of the geologist and the provision of adequate exploration funding by the builder/owner. In this respect, there has been a truly unfortunate trend towards unwarranted price-driven selection ('bid-shopping', or commoditization) of site-characterization consultants, by owners and some engineers, worldwide, since about 1980. Suffice it to say that the geologist who will perform the site characterization should be selected on the basis of qualifications and experience, rather than on the cost (budget) accorded to the effort.

Engineering Geological Site Characterization

Each construction site needs to be characterized for its geological, hydrological, and waste conditions, so that architects and design engineers have the necessary parameters when considering feasibility and design. In some cases, when geological constraints (the preferred term for 'geological hazards') are identified, the proposed construction may be shifted to a geologically 'less expensive' site in terms of building costs, operation, and maintenance of the proposed works. More often, however, for a variety of reasons, the owner of the planned construction, public or private, is committed to the site and the geological knowledge becomes essential to the success of the project.

Engineering geologists generally begin site work with a walkover of the site, attempting to 'peer' below the ground surface to generate a working hypothesis and conceptual geological model of what is likely present and what may be hidden. The ultimate goal is to assess how geotechnical factors could impact the construction or performance of the engineered works. Tables 4 and 5 provide a summary of geological conditions that might be expected to impact project feasibility, design, and construction processes

Table 5 Problems of urban construction: societal pressures causing geological impacts

Condition	*Impact*	*Geological considerations*
Cities require off street vehicle parking for a significant percentage of building occupants and visitors	Above ground space is too valuable for parking vehicles, hence deep basements are required	Hardrock excavation costs are extreme; bedded sedimentary and jointed or foliated crystalline rock present slope stability problems for the basement excavation during construction; stability of adjacent buildings is a constant concern
Water must be supplied to the city	Need for conveyance projects and tunnels; importation of the water for the distribution system	Linear projects such as pipelines, aqueducts, highways, and railroads are particularly susceptible to earthquake ground motion
Transportation routes must not interfere with much of the existing ground surface infrastructure	Reliance on tunnelling to achieve minimal grades between stations and minimal depth below ground for people movement	Places geological information at a premium for design and construction of stations; keeps transit tunnel alignment in sound rock, free of major ground support considerations
Abandoned waste dumps of all types	Organic debris will naturally compress as much as 300% vertically (becoming as little as one third of its preconstruction thickness, hence causing considerable settlement)	Use archival records, old topographic maps, historic aerial photographs, and confirming geophysical traverses
Previous quarrying, mining, and other mineral extraction activity	Many cities once relied on extraction of coal, building stone, gypsum, clay mineral, or mineral ore, leaving unstable ground	Present day hazards include rotting support timber, collapse prone workings, and movement of contaminated groundwater in workings
Excessive abstraction of groundwater	Causes rotting of historic timber foundation piles; these may also rise on termination of water abstraction	Saline intrusion in coastal areas, activation of non seismic 'growth' faults in post Cretaceous coastal embayments

Figure 2 An example of a zone of 'bad ground' discovered and delimited in an engineering geological site characterization ex ploration. This zone represents poor engineering properties of foundation materials in terms of their ability to support structural loads that would be imposed by engineered works. Shown is a shear zone of ancient tectonic origin, as it perpendicularly enters the foundation excavation for the Lahey Medical Clinic, urban Boston, Massachusetts. Such features are important when they are planar to and cross one or more dimension of the foundation excavation, because they have the potential to induce instability of the surrounding ground and existing buildings. Timbers are 200 mm^2 in section; survey tape in midview. Photograph by the author.

and therefore must be incorporated into the site engineering geological characterization process.

Geological conditions decidedly influence geotechnical design measures to transfer dead and live loads of proposed structures into the foundation soil or rock (Figure 2). Transfer of load to geological horizons is integral to ensuring adequate structural support for engineered works. During the site characterization process, appropriate site exploration techniques must be incorporated to determine the engineering geological conditions likely to adversely impact the project.

Summary

The struggle to provide decent living conditions for humanity on earth is characterized to greatest degree by the interwoven impact of geology on the human habitat and the impact of humans on geologic conditions that serve to support that life. Cities are the focus of human activity and this activity takes place on and in the ground, which is itself only a complex of geologic conditions. Of all of the qualities of human life in the cities is its dynamism, all of which tends to obscure and often obliterate the delicate evidence of the nature of geologic conditions supporting the city. For this reason, we must turn to urban geologic observations and recording in order to compile a sequence of snap-shot views and vignettes of urban geology. When recorded, this information is invaluable in the continual struggle to accommodate the great demands.

See Also

Engineering Geology: Aspects of Earthquakes; Natural and Anthropogenic Geohazards; Made Ground; Site Classification; Ground Water Monitoring at Solid Waste Landfills. **Europe:** Holocene.

Further Reading

Association of Engineering Geologists (1982 2004) Cities of the world. *Environmental and Engineering Geoscience Journal* (annual series).

Baskerville CA (1992) *Bedrock and Engineering Geologic Maps of Bronx County and Parts of New York and Queens Counties, New York: USGS Miscellaneous Investigation Series MAPI 2003, two sheets, 1:24,000.* Washington, DC: US Geological Survey.

Culshaw MG (2005) Urban geoscience (the Glossop Lecture for 2004). *Quarterly Journal of Engineering Geology and Hydrogeology.*

Kaye CA Jr (1959) *Geology of the San Juan Metropolitan Area, Puerto Rico: US Geological Survey Professional Paper 417 A.* Washington, DC: US Geological Survey.

Kaye CA Jr (1976) *The Geology and Early History of the Boston Area of Massachusetts; A Bicentennial Approach: Geological Survey Bulletin 1476.* Washington, DC: US Government Printing Office.

Legget RF (1962) *Geology and Cities.* New York: McGraw Hill.

Legget RF and Hatheway AW (1988) *Geology and Engin eering.* New York: McGraw Hill.

McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience.* Rotterdam: Balkema.

Schlocker J (1974) *Geology of the San Francisco North Quadrangle, California: US Geological Survey Profes sional Paper 782.* Washington, DC: US Geological Survey.

Schuberth CJ (1968) *The Geology of New York City and Environs.* New York: Natural History Press.

United States Geological Survey (1894 1955) *Folio series.* Select titles pertaining to cities; maps at 1:125,000 plus text and photographs. Washington, DC: US Geological Survey.

VENUS

See **SOLAR SYSTEM: Venus**

VOLCANOES

G J H McCall, Cirencester, Gloucester, UK

Introduction

Volcanoes are a major component of the Earth's present surface geology, both active and extinct; volcanic processes can be recognized throughout geological history and are important in generating certain types of mineral deposits (e.g., metallic sulphides, sulphur). They are studied in their relationship to both mantle (*see* **Earth:** Mantle) and crustal (*see* **Earth:** Crust) processes of igneous rock generation and plate tectonics. Volcanic eruptions comprise one of the most important of natural hazards (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards) threatening populations living close to them. Volcanoes have been recognized on other bodies of the solar system – Venus (*see* **Solar System:** Venus), Mars (*see* **Solar System:** Mars) and Jupiter's satellite Io (*see* **Solar System:** Jupiter, Saturn and Their Moons) – the latter is the most volcanically active body in the solar system. Mars has the largest single volcano, Olympus Mons (440 km diameter; 24 km altitude). Venus, which we can only study from radar images on account of its dense volcanogenic CO_2-rich atmosphere, appears like Mars, to have no active volcanoes now, though there are many inactive ones.

Volcanoes and the Mantle

Liquid rock, poured out from volcanoes as lava (*see* **Lava**), makes up only a small portion of the planet, though a large part of the core is molten metal. The outer layer of the Earth, the lithosphere, is relatively cool, but the mantle below is so hot that rocks lose their cohesion. Indirect evidence obtained from seismology suggests that they move very slowly; and the theory of plate tectonics requires such movements to take the form of convection currents and localized upwellings (hot spots, mantle plumes) (*see* **Mantle Plumes and Hot Spots**). However, there is no agreement among geoscientists as to the depth of the lower boundary of these circulations. The hot, soft, solid material of the mantle, of peridotite and related compositions, only partially liquifies when the temperature exceeds the melting point of minerals in the rock. Because it is lighter than the solid rock above it, it will rise towards the surface, entering the lithosphere. It may pass through the crust directly and swiftly appear at the surface as a volcano, or it may be halted and form a large concentration of molten rock down in the crust, a magma chamber: from this it may later burst out to the surface as a residue, changed after some crystallization in the magma chamber. In narrow conduits to the surface, it may cool and crystallize, forming wall-like intrusions (dykes) or sheet-like intrusions (sills). Intrusive rocks formed by crystallization of the magma in the magma chamber, dykes, and sills, form the underworks of volcanoes, and may be all that is left after erosion: they frequently comprise ring complexes as on the Isle of Rum in Scotland. A quirk of erosion has left these underworks exposed, surrounded by Precambrian granite, in a hollow central to the remains of the lava and tuff pile in the 100-km wide Miocene Kisingiri volcano in western Kenya – Howel Williams called this the best preservation yet seen of a volcano complete with its upper- and under-works.

Distribution of Volcanoes

Volcanoes do not occur everywhere on the Earth's surface. The global distribution shown in **Figure 1** reflects the present distribution of more than a dozen tectonic plates, rigid plates of the lithosphere, which move laterally across the Earth's surface at minute velocities in the order of centimetres per year. Of course, they cannot do this indefinitely without colliding with each other and this pattern is controlled by upwelling of magma on the lines of separation (the mid-ocean ridges) and either the diving down of the spreading oceanic plate under the one it spreads against (subduction) or, where continental parts of plates meet, collisions such as formed the Himalayas. In subduction there is both

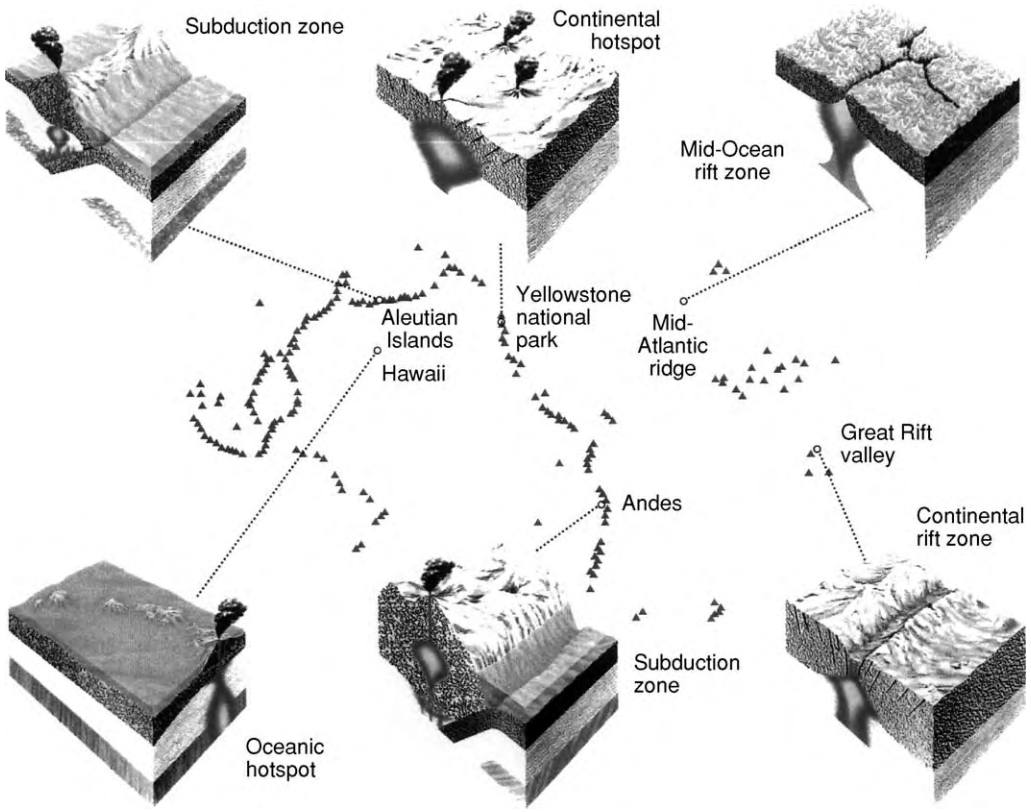


Figure 1 The global distribution of volcanoes and their geotectonic setting (after Scarth (1997)).

upwelling of magma from the mantle and melting in the lithosphere, after the descending oceanic plate has penetrated to some depth, and volcanic arcs are so formed, not at the surface boundary between the plates (the oceanic trench), but at some distance beyond it within the overriding plate.

As shown in Figure 1, most of the existing volcanoes of the Earth are concentrated in the arc zones; the ridge zones of upwelling produce much eruption deep under the oceans, but few subaerial volcanic piles occur on them: exceptions being the volcanoes on Iceland and Tristan da Cunha. Some volcanoes occur in the oceans unrelated to either type of plate tectonic zone, but are attributed to hot spots or plumes in the mantle: the Hawaiian Islands and the Galapagos are examples of these. Seamounts are stumps of such volcanoes. There are also volcanoes in rift zones within continents, for example Lengai, Tanzania, in the eastern rift and the volcanoes of the western rift (e.g., Nyiragongo), but such rift valleys are really aborted oceanic ridges, where the updoming and eruption has occurred, but the two flanks never spread and diverged away from the rift.

Classification of Volcanoes

There are two types of eruption:

Fissure Eruption

Concentrated in long, narrow fissures, the lava generally spreads to form extensive lava plateaus. If the eruption is explosive, plateaus of ash flows may form. Most fissure eruptions are, however, associated with fluid basalt lavas, which have given rise to the terms flood basalts and plateau basalts, but phonolites in Kenya also form extensive lava plateaus. Such eruptions have not been common in historical times, though the Laki fissure eruption in 1783 was of this type: however, extensive plateaus hundreds of kilometres wide of superimposed, quite thin fluid flows were formed in the geological past (e.g., the Deccan Plateau in India).

Central Eruption

Volcanic activity is more commonly concentrated in a central vent and gives rise to volcanic cones or, in the case of some viscous lavas, domes or necks with no summit crater. Single volcanoes, clusters, or chains are formed in this way. The chains may be associated with belts of extreme seismic disturbance (e.g., the Aleutian islands, Chile). Central volcanoes may have satellite cones superimposed on them: these are said to be nested (Figure 2) if they occur within the crater, parasitic or adventive if they occur on the outer slopes.

Central volcanoes of substantial dimensions may develop large circular cavities, calderas, focused on their summits: these may be circular, or of rather irregular annular shape; they may even be multiple, one inside or superimposed on the other. The Suswa caldera in Kenya is annular, a central plateau being preserved. Calderas may be formed either by subsidence, due to withdrawal of magma/lava support, or volcanic explosion. Calderas are evident on Mars and Io.

There are three types of central volcanic pile, all of which may develop calderas:

i. shield volcanoes: all lava flows
ii. ash cones: all pyroclastic
iii. composite or strato-volocanoes: alternating lava and pyroclastic (Figure 3).

The material that flows out as lava is termed effusive. Extrusive includes both lavas and pyroclastic rocks.

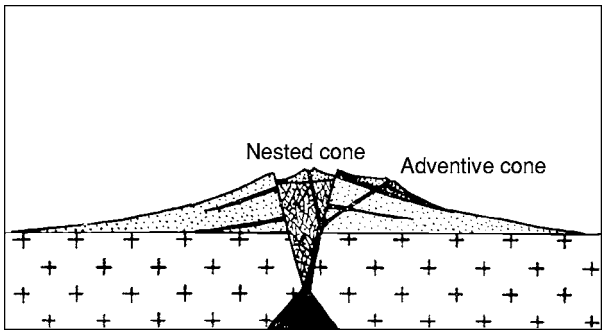


Figure 2 Diagram of a composite volcano: lava (black) and tephra (stippled) interbedded (from Green and Short (1971)).

There are seven types of volcano, classified according to the nature of the eruption. Four of the names stem from southern Italy, a field of classical study which can be termed the 'Birthplace of Vulcanology'; and which has lately been described by Guest *et al*. in modern terms. Iceland, Hawaii, and Martinique supply the remaining names. The list is given in Table 1.

This scheme is a very useful subdivision, but most volcanoes show some departure from these classical definitions derived from a handful of famous volcanoes. No other volcano, for example, matches Stromboli, where explosions occur every 10 minutes and activity has been continuous for 2500 years. Other types can be long dormant, Mt Pelée in Martinique, lay dormant for centuries, only, in 1902, to send off sudden blasts lasting a few minutes separated by weeks of relative inactivity while the plug in the vent built up into a spine (Figure 5).

A very useful table of explosivity of volcanoes was published in 1982 (Table 2):

Products of Volcanoes

Lavas

Magma erupted from volcanoes as lava consists of molten rock, crystals, and gas – carried as bubbles, mainly water, and carbon dioxide. The less siliceous lavas such as basalt flow freely and build up extensive shield volcanoes as in Hawaii. Other lava types may flow freely – a phonolite flow forming the Yatta Plateau in Kenya followed a valley for >300 km from its source. Less fluid flows such as trachyte

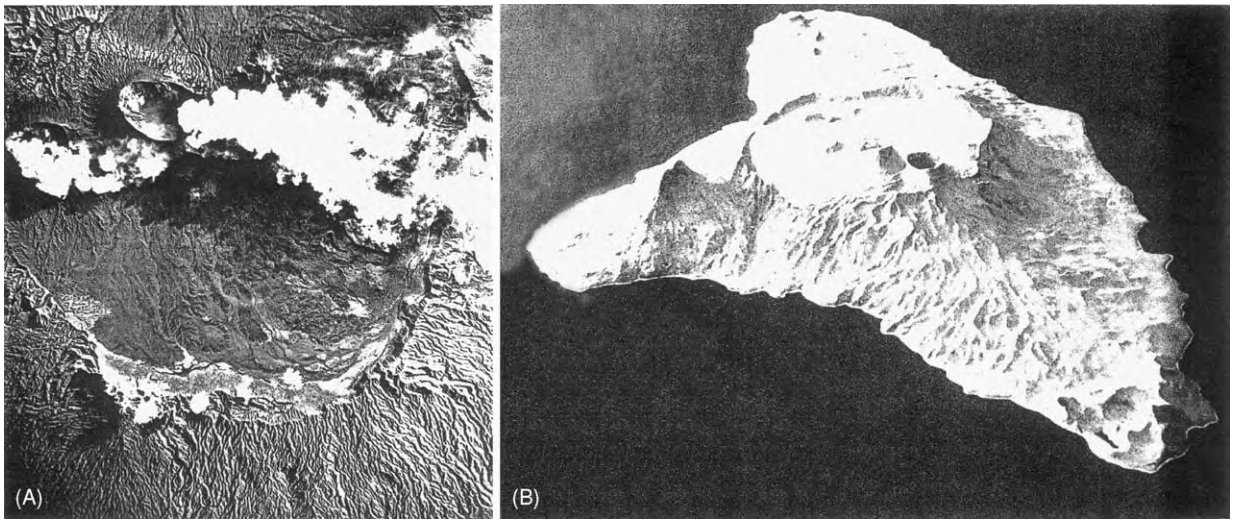


Figure 3 The 13 km diameter basaltic Ambrym volcano, Vanuatu. (A) Plaster model (by Jon Stephenson) (B) Air photo showing the caldera with two active craters nested within (Benbow and Marum).

Table 1 The seven types of volcano

Type	*Characteristics*
Icelandic	Fissure eruptions, releasing free flowing basaltic magma; quiet, gas poor; great volumes of lava flowing as sheets over large areas to build up plateaus
Hawaiian	Fissure, caldera, and pit crater eruptions: mobile lavas with some gas; quiet to moderately active eruptions: occasional rapid emissions of gas charged lava produce fire fountains; mainly basaltic; minor amounts of ash; builds up lava domes
Strombolian	Stratocones (composite) with summit craters: moderate, rhythmic to nearly continuous explosions, resulting from spasmodic gas escape; clots of lava ejected producing bombs and scoria; periodic more intense activity with lava outpourings; light coloured clouds, mostly steam reach only to moderate heights
Vulcanian	Stratocones: central vents; associated lavas more viscous; lavas crust over vent between eruptions, allowing gas buildup below surface; eruptions increase over long period of quiet until crust is broken up, clearing vent and ejecting bombs, pumice, and ash; lava flows from the top of flank after main explosive eruption; dark ash laden clouds, convoluted and cauliflower shaped, rise more or less vertically to moderate heights, depositing ash along the flanks of the volcano (note: other types, such as Hawaiin can produce similar effects when they suffer interference with groundwater, and phreatic eruption ensues, large steam clouds carrying fragmental material)
Vesuvian	More paroxysmal than the above two: extremely violent expulsion of gas charged magma from stratocone vent; eruption occurs after long interval of quiescence or mild activity: vent tends to be emptied to considerable depth: lava ejects in explosive spray (glow above vent), with repeated clouds (cauliflower) that reach great heights and deposit ash
Plinian	More violent form of Vesuvian eruption (**Figure 4**): last major phase is uprush of gas that carries cloud vertically upward in a column for kilometres; narrow at base but expands outwards at upper elevations; cloud generally low in ash
Peléan	Results from high viscosity lava and delayed explosiveness, conduit of stratovolcano being usually blocked by dome or plug (**Figure 5**); gas (+ some lava) escaped from lateral (flank) openings or by destruction/uplift of plug; gas, ash, and blocks move with high velocity downslope in one or more blasts as nuees ardentes or glowing avalanches, producing directed deposits

After Guest *et al.* (2003).

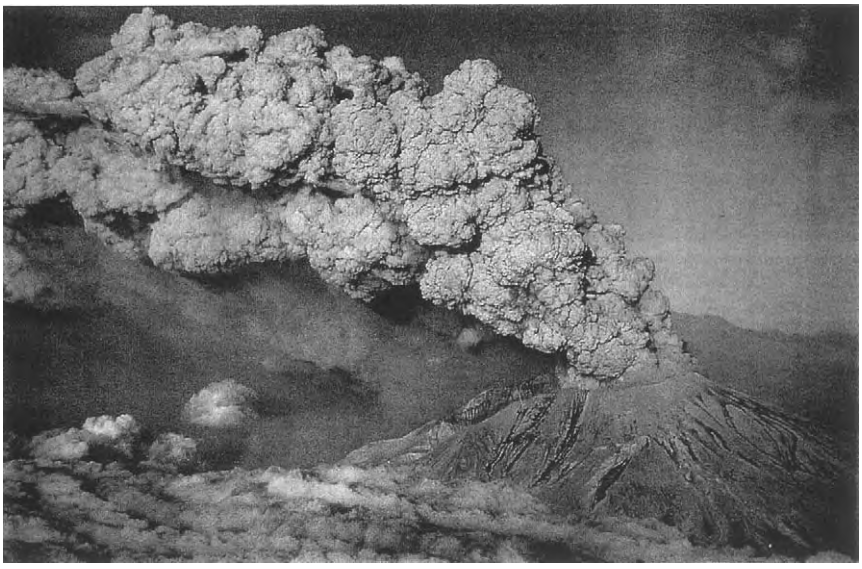


Figure 4 Plinian eruption of Mt St Helens, Oregon: ash, gas, and pulverized rock shooting out directionally (second eruption, July 1980) (from Pyle, 1998).

may terminate close to the vent (Figure 6). Even more siliceous lavas, such as rhyolite (equivalent in composition to granite), are sticky and form heaps close to the vent, or block the vent as plugs which may slowly extrude as spines. Lavas of this type are very dangerous as the plug finally breaks and there is a sudden explosive outburst which cannot be predicted. Both basalts and rhyolites tend to break up as they crust over, yielding slaggy flows. The solidified crust may be arrested and the molten lava flow underneath

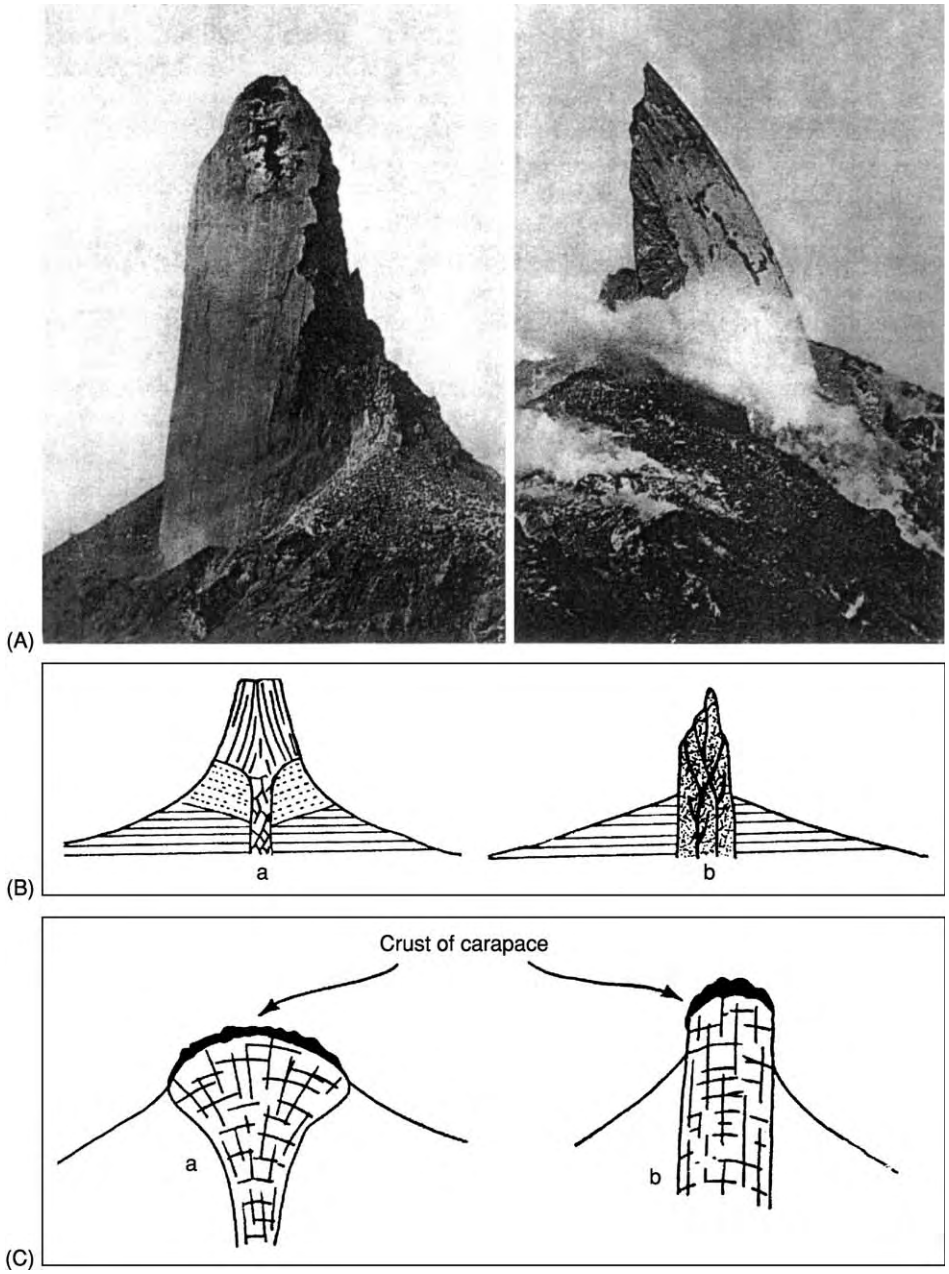


Figure 5 Mt Pelée, Martinique, (A) Two views of the rhyolite spine that plugged the vent prior to the catastrophe in 1902. (B) Types of lava domes: with spreading and no spreading. (C) Volcanic necks left behind by the above (from Green and Short (1971)).

it in a lava tube. There are four types of flow distinguished: pahoehoe (ropey), a'a' (slaggy); block lava; and pillow lavas (Figure 7). The latter form when the lava flows into water: sack-like glass covered bodies form, with concentric zones of vesicles or varioles (gas concentrations): they fissure radially and tend to break up easily, forming angular hyaloclastite breccias, containing pillow fragments. Fine fragmental glassy material formed in water or by deposition on wet surfaces is called peperite.

Some volcanoes erupt quite unusual peralkaline lavas; the Miocene 60 km diameter Kisingiri volcano in the Kavirondo Rift Valley, Kenya, erupted nephelinite and melilitite lavas; the magma chamber rocks below are nephelinitic ijolites and melilitic uncomprahgrites and turjaites, and the throat of the eroded volcano is a huge plug of carbonatite. Carbonatite has, in the past, been erupted as lavas from volcanoes elsewhere, but the only active emission known is of thin fluid flows of natrocarbonatite in

Table 2 Volcanic explosivity index

	Non explosive	*Small*	*Moderate*	*Moderate large*	*Large*	*Very large*			
General description	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Volume of tephra (m^3)	10^4	10^6	10^7	10^8	10^9	10^{10}	10^{11}	10^{12}	
Cloud column height (km)*	<0.1	0.1 1	1 5	3 15	10 25	25			
Qualitative description	Gentle, effusive ←		Explosive			Cataclysmic, paroxysmal, colossal →			
						Severe, violent, terrific →			
Classification	←	Strombolian	→			Plinian →			
	Hawaiian ←			Vulcanian	→	Ultra Plinian →			
Total historic eruptions	487	623	3176	733	119	19	5	2	0
1975 1985 Eruptions	70	124	125	49	7	1	0	0	0

*For VEI 0 2, data are km above crater; for VEI 3 8. Data are in km above sealevel.
From Newhall and Self (1992).

the crater of Lengai volcano, Tanzania. The volcanoes Nyamuragira and Nyiragongo in the Western Rift Valley near Lake Kivu erupt potassic lavas rich in leucite.

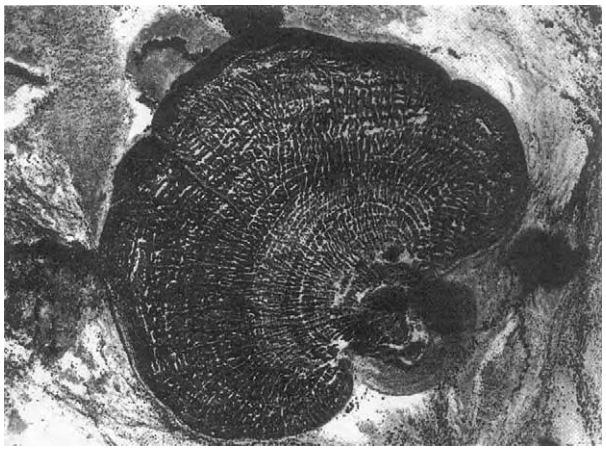


Figure 6 Trachyte flow extending only a few metres from the source vent in an ash cone (Silali volcano, Kenya) (from Green and Short (1971)).

Tephra

All clastic (fragmental) material issued from a volcano is covered by the term 'tephra'. It is classified by size, as in Table 3:

Phreatic Eruptions

Interaction of rising magma and groundwater produces explosive eruptions, commonly with little material emission. Shallow depressions called 'maars' are formed this way and the extreme case is the crater with no material eruptive association except fragmented country rock (e.g., Hole-in-the-Ground, Oregon, Figure 9).

Volcanic Clouds

These may be of steam or ash rising to many kilometres and moving with the wind (Figure 10): the cloud from Stromboli is often carried as far as Greece. Ash clouds may be incandescent. An acid-bearing ash cloud from Klyuchevskaya, Kamchatka, rose to 20 000 m in 1994 and was carried 1000 km to the east by 240 km per hour winds, interfering with air travel.

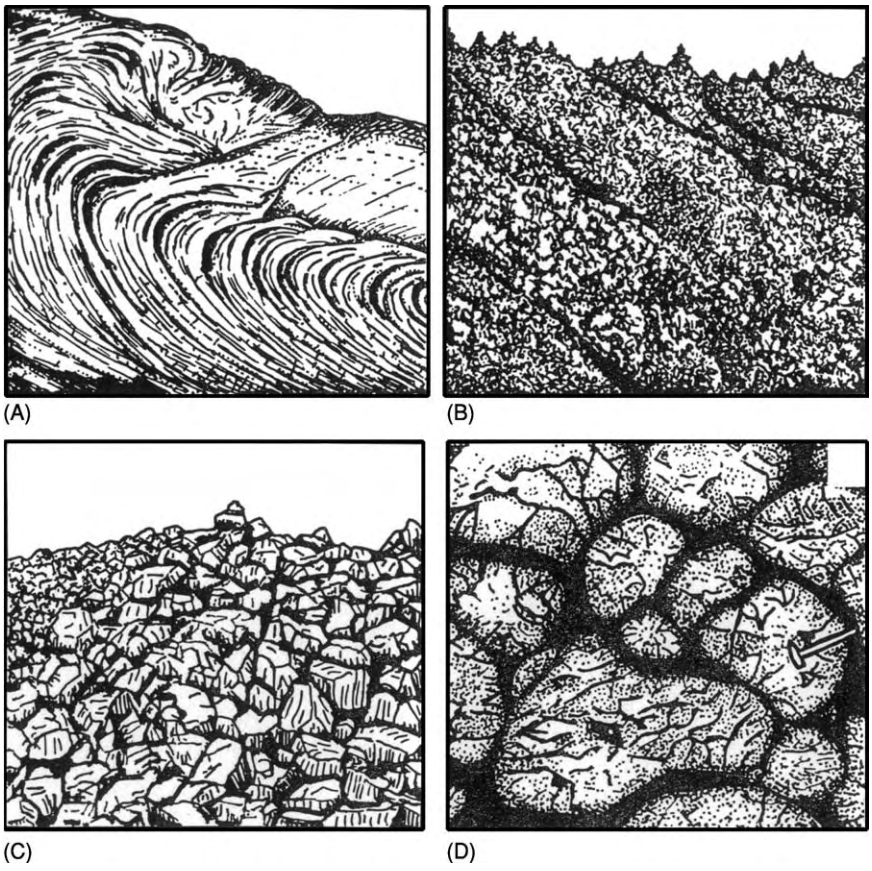


Figure 7 Four types of lava: (A) Pahoehoe; (B) A'a'; (C) Block lava; (D) Pillow lava (from Green and Short (1971)).

Table 3 Tephrà classified by size

Name	*Size range*	*Remarks*
Dust	Less than 0.25 mm diameter	
Ash or sand	0.25 4 mm diameter	When lithified, the finer material is referred to as 'tuff' (**Figure 8**). The terms 'crystal' or 'lithic' are added according to the dominant component. Also the word 'pumice' may be added if pumice is dominant
Lapilli	0 32 mm diameter	
Bombs or blocks	More than 32 mm diameter	When lithified, these coarse deposits are referred to as 'agglomerates' or 'volcanic breccias'. The word 'scoria' is used for clasts with numerous open gas cavities, appearing like a sponge.
Bentonites		Bentonites are montmorillonitic clayey rocks formed by devitrification of volcanic glasses: palagonite is a yellow isotropic material formed by alteration of basaltic glass shards or hydration of them when hot

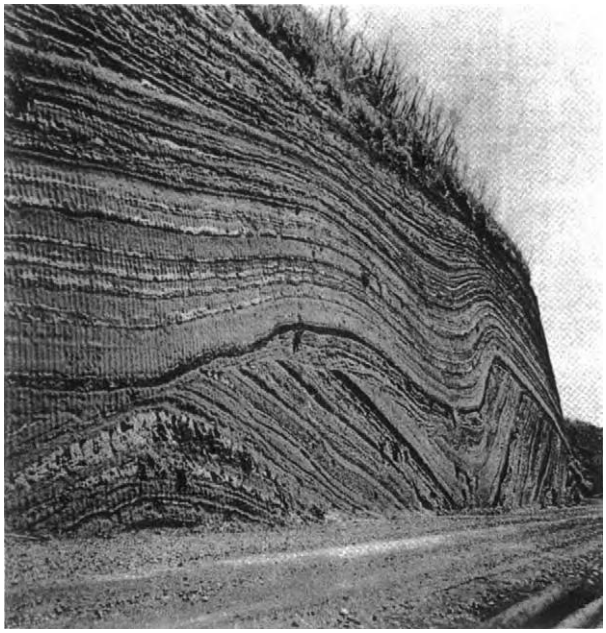


Figure 8 Bedded basaltic tuff of Pleistocene age, O Shima volcano, Japan (from Green and Short (1971)).

Nuées Ardentes

These are glowing avalanches of unsorted tephra or incandscent ash flows. They are characteristic of volcanoes erupting siliceous magma (rhyolite or andesite), and occur when plugs rupture or a sector of the volcanic cone collapses. Another cause is loss of support from below of a lava column. The eruption of Vesuvius in AD 79 killed many by asphyxiation (**Figure 11**).

Lahars

These are volcanic debris flows and may be hot or cold. They may arise from interaction of hot pyroclastic flows and surges with ice and snow covering the volcanic cone summit, as at Nevado del Ruiz, Colombia, in 1985 (**Figure 12**); in this case the duration of the horrific event was small, but in the case of Pinatubo, Philippines, in 1993, pyroclastic flows continued to be reworked by heavy rains, forming lahars which continued for years.

Gases

Volcanic gases are commonly dominate by CO_2, which asphyxiates: such gases are invisible and concentrate in depressions in the land, as in Iceland: these concentrations are called 'mofettes': The Lake Nyos disaster in Cameroon was due to a sudden outburst of these gases from under the lake.

Volcanoes as a Major Natural Hazard

Volcanoes are distributed along plate boundaries and along rifts and fractures within plates, and they are thus fixed features. The hazards they pose are easier to predict and restrict geographically than those of earthquakes, but they pose a great hazard–particularly because the volcanic rocks form rich soils for agriculture and thus populations concentrate around them. Not only active volcanoes pose hazards: in 1943, Paricutin, Mexico, originated as a fissure opened up in a ploughed field and built up to a 2400 m high cone, from which 'a'a' lavas flowed down and buried buildings. Mt Pelée in Martinique was supposed, in 1902, be extinct and a nuée ardentes killed all but one of its 29 000 inhabitants (a prisoner in a gaol).

The toll on human life is an order less than from earthquakes. Since 1980 there have, however, been more than 30 000 deaths and some 1 million people have been detrimentally affected by eruptions.

In **Table 4**, some selected volcanic disasters, since the eighteenth century, are listed. Fatalities are not a

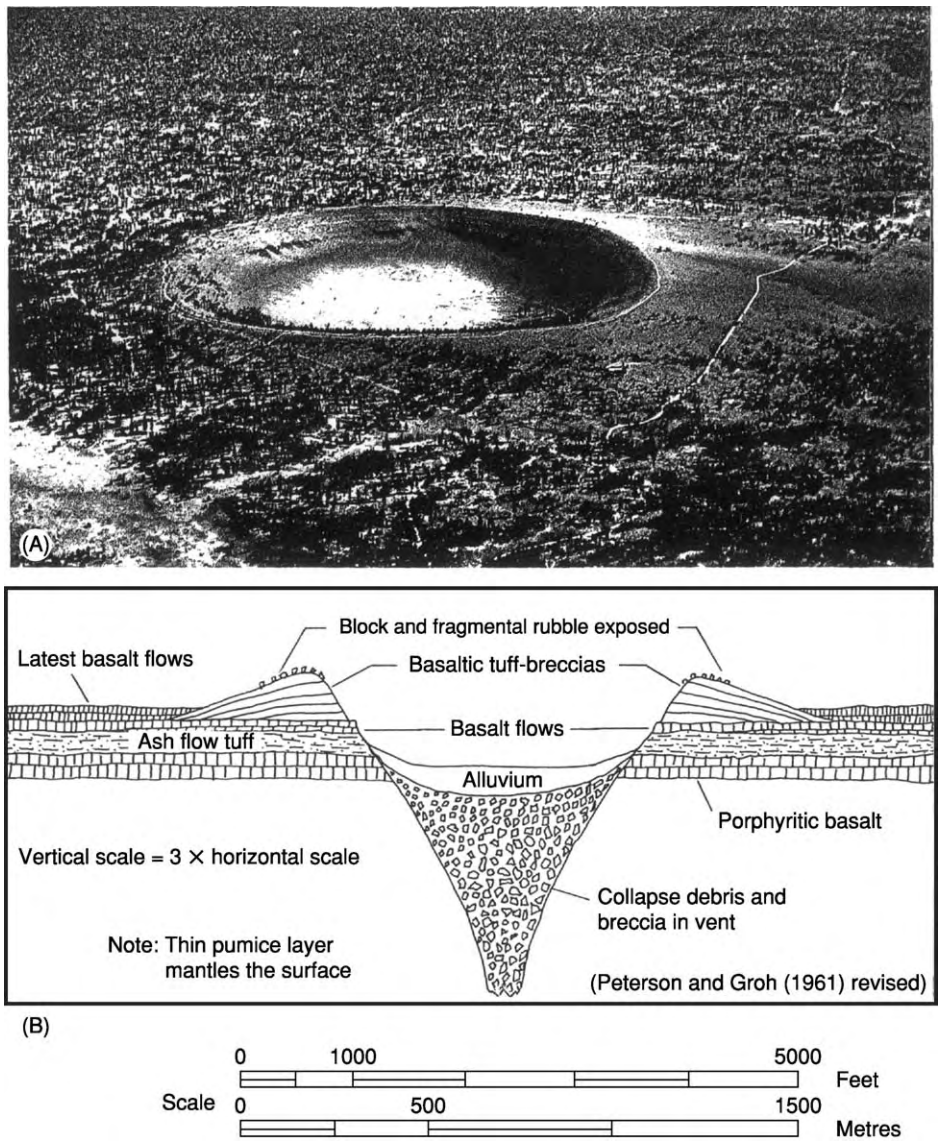


Figure 9 Hole in the Ground, Oregon, a 1.6 km diameter Maar (from Green and Short (1971)).

satisfactory measure of disaster – for example, the 1997 collapse of the old crater wall surrounding the dome complex of Soufriere, Montserrat, cause only 19 fatalities, but the ongoing eruptions over several years caused the capital of the small island and more than half of the island's population to be evacuated indefinitely.

The principal volcanic hazards and effects are listed below:

Hazards

- Lava flows
- Ash falls
- Pyroclastic flows and surges
- Directed blasts and atmospheric shock waves
- Lahars and floods
- Landslides
- Volcanic gases
- Tsunamis
- Climate modification
- Crater lake emptying suddenly
- Ice/snow lava interaction

Effects

- Loss of land and buildings
- Disruption of social and economic infrastructure
- Famine
- Water pollution

- Disease
- Drowning
- Asphyxiation

Lava flows represent the principal hazard of basaltic volcanoes. On Etna, they stream down the flanks and threaten villages and coastal towns. The recent eruptions of Nyiragongo volcano above Goma in the Congo drove a lava flow through the middle of the town, destroying much property. They destroy buildings and usable land. Mitigation by diversion has been tried on Etna but is of limited success.

Pyroclastic flows may attain velocities of tens to hundreds of metres per second and temperatures of 300–800°C. They may generate secondary debris flows. Block and ash flows are slightly less deadly than pumice-rich ignimbrites. The worst of these flows and blasts destroy and kill everything in their path.

Debris avalanches, as at Mt St Helens in 1980 (Figure 4), travel rapidly and are equally lethal and destructive.

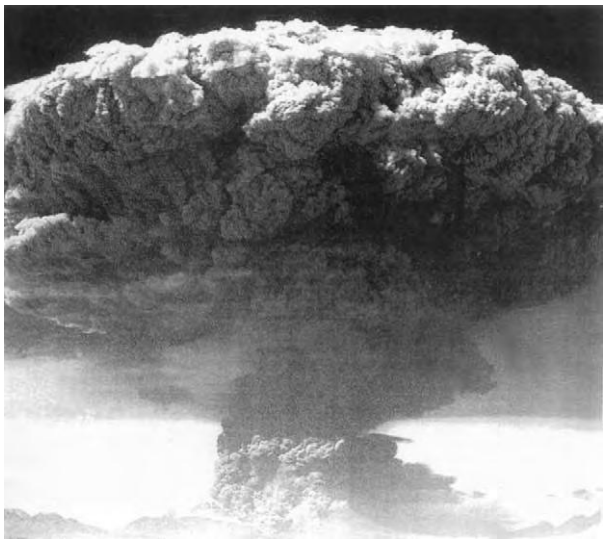


Figure 10 Ash cloud from a devastating eruption of Lascar volcano, Chile (from Pyle (1998)).

Figure 11 Cast of a child's body, asphyxiated by pyroclastic surge of Vesuvius, 79 AD (from Guest *et al.* (2003)).

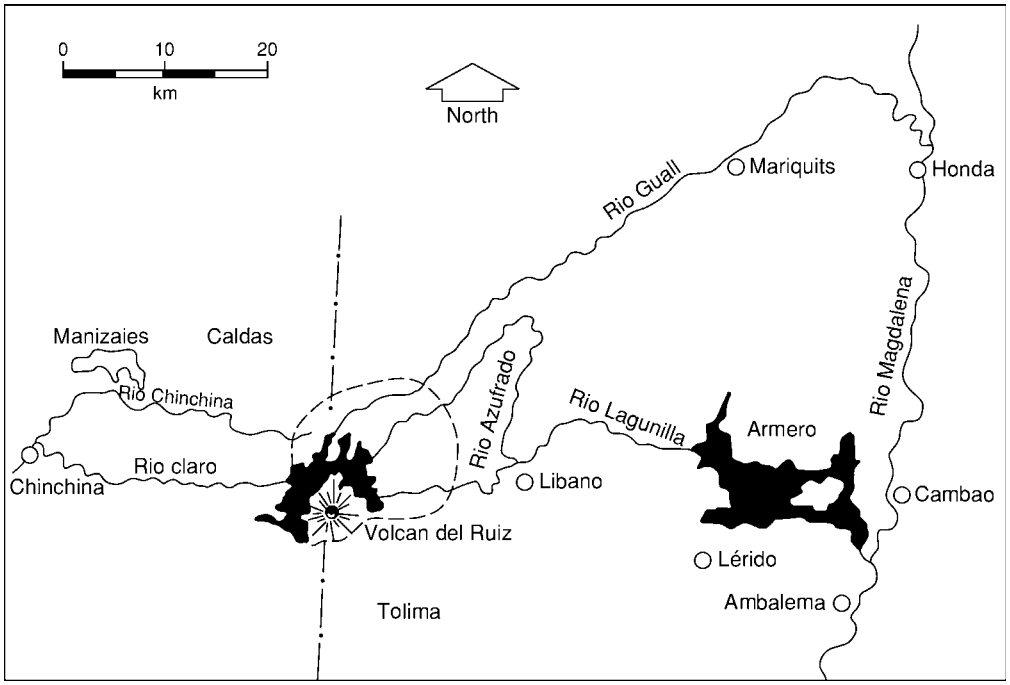


Figure 12 The pathway of the lahar from Nevado del Ruiz volcano, Colombia, to destroy Armero (from Hall, Geohazards Natural and Man Made, Chapman and Hall (1992)).

Table 4 Selected Volcanic disasters since the eighteenth century

Volcano (country)	*Year*	*Fatalities*	*Cause*
Laki, Iceland	1783	10 500	Famine
Unzen, Japan	1792	15 188	Tsunami
Tambora, Indonesia	1815	92 000	Mainly famine
Krakatau, Indonesia	1883	36 417	Tsunami
Mt Pelée, Martinique	1902	29 000	Pyroclastic flows
Kelut, Indonesia	1919	5110	Debris flows
Lamington, Papua New Guinea	1951	2940	Pyroclastic flows
El Chichon, Mexico	1982	1700	Pyroclastic flows
Nevado del Ruiz, Colombia	1985	25 000	Debris flows
Lake Nyos, Cameroon	1986	1746	Volcanic gases (CO_2) asphyxiation*
Pinatubo, Philippines	1991	500	Various
Soufriere, Montserrat	1997	19	Pyroclastic flows
Goma, Congo	2002	40	Lava flows

*This was not associated with a volcano: the gases came from a lake but were of volcanic origin.
After WJ Maguire.

Table 5 Volumes and non out distances for volcanic debris

Volcano	*Volume (km^3)*	*Run out (km)*
Nevado di Colima	22 33	120
Socompa	17	35
Volcan di Colima	6 12	43
Shasta	26	50
Popocatapetl	28	33
Chimborazo	8.1	35
Mawenzi	7.1	60
Galunggung	2.9	25
Mt St Helens	2.5	24
Fuji	1.8	24
Shiveluch (1964)	1.5	12
Bandai San (1888)	1.5	11
Egmont	0.35	31
Unzen (1792)	0.34	6.5
Asakusa	0.04	6.5

After WJ Maguire.

Lahars can be deadly, as at Nevado del Ruiz in 1985, where a flash flood was generated and spread out 40 km to destroy Armero completely (Figure 12).

A tsunami generated by eruption in the Kurile Is in 1737 swept up a fjord to a height of 65 m.

Tables 5 and 6 list volumes and run-out distances for volcanic debris avalanches and principal mitigating measures for these hazards. An interesting fact, as described by P Delos Reyes, is that animals give indications of forthcoming eruption by their unusual actions.

In Figure 13A and B, the distribution of hazards around a volcano is illustrated and the relative dispersal – the most dispersive fine ash carrying sulphuric acid droplets may extend over the entire circumference of the planet, high in the stratosphere and cause spectacular sunsets for years. In Figure 14, mitigation measures are illustrated.

Volcanoes listed for special study under UN IDNDR are named below (from WJ Maguire):

Decade volcanoes (UN sponsored)

- Colima (Mexico)
- Galeras (Colombia)
- Mauna Loa (USA)
- Merapi (Indonesia)
- Mount Rainier (USA)
- Nyirogongo (Zaire)
- Sakurajima (Japan)
- Santa Maria (Guatemala)
- Ta'al (Philippines)
- Ulawan (Papua New Guinea)
- Unzen (Japan)
- Vesuvius (Italy)

Laboratory volcanoes (EU sponsored)

- Etna (Sicily)
- Furnas (Azores)
- Krafla (Iceland)
- Piton de la Fournaise (Reunion Is)
- Teide (Tenerife)
- Santorini (Greece)

It is noteworthy that none are in Kamchatka, Siberia, often cited as the world's most active volcanic region: 20 volcanoes erupt there, 3–5 times per year.

Volcanoes and Earthquakes

Volcanic earthquakes provide a valuable warning of impending eruption, clusters of small seisms being detected on seismograph arrays. Major tectonic earthquakes are quite distinct from these seisms, but major earthquakes and eruptions can be interrelated:

Table 6 Principal mitigating measures for volcanic hazards

Hazard	*Principal mitigating measures*
Lava flows	Damming or diversion: flow front water cooling
Debris flows (lahars) and floods	Judicious siting of settlements; construction of elevated refuges, sediment dams, and baffles; dredging and levee construction; seismometer and trip wire warning systems
Pyroclastic flows and surges	Judicious siting of settlements; pre evacuation
Tephra	Evacuation of poorly constructed buildings; accumulation of accumulated ash, etc. from roofs; availability of face masks/protective headgear; appropriate medical care for respiratory problems and ingestion of glass microshards; contingency plans for power cut off, communications and transport disruption; availability of uncontaminated water supplies; measures to minimise crop and livestock damage; warnings to air traffic
Landslides and debris avalanches	Identification of collapse prone areas; slope stability monitoring; pre evacuation
Directed blasts and shock waves	Pre evacuation
Volcanic gases	Gas monitoring; resettlement if a permanent problem: pre evacuation if episodic and predictable: public safety guidelines and warning notices; construction of elevated refuges where appropriate
Tsunami ('tidal wave')	Identification of unstable slopes adjacent to water bodies; slope stability monitoring; pre evacuation; establishment of a tsunami warning network both regionally and internationally

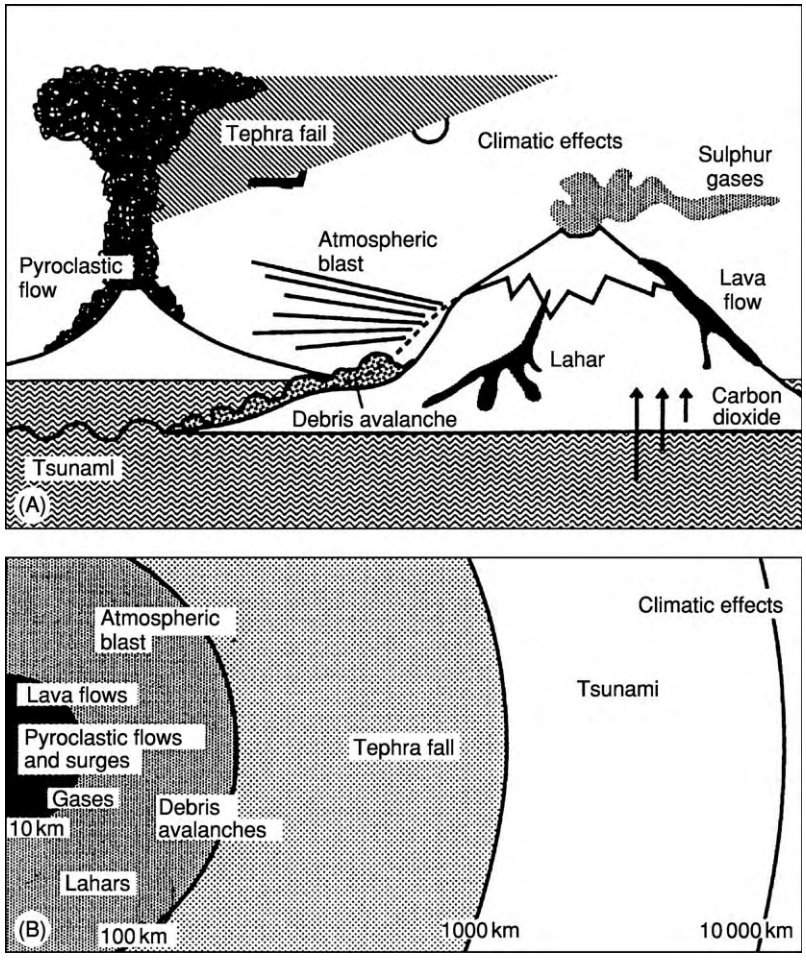


Figure 13 The destructive processes generated by volcanoes: (A) showing those confined to the immediate vicinity. (B) showing those dispersing further, even globally (from Maguire (1998)).

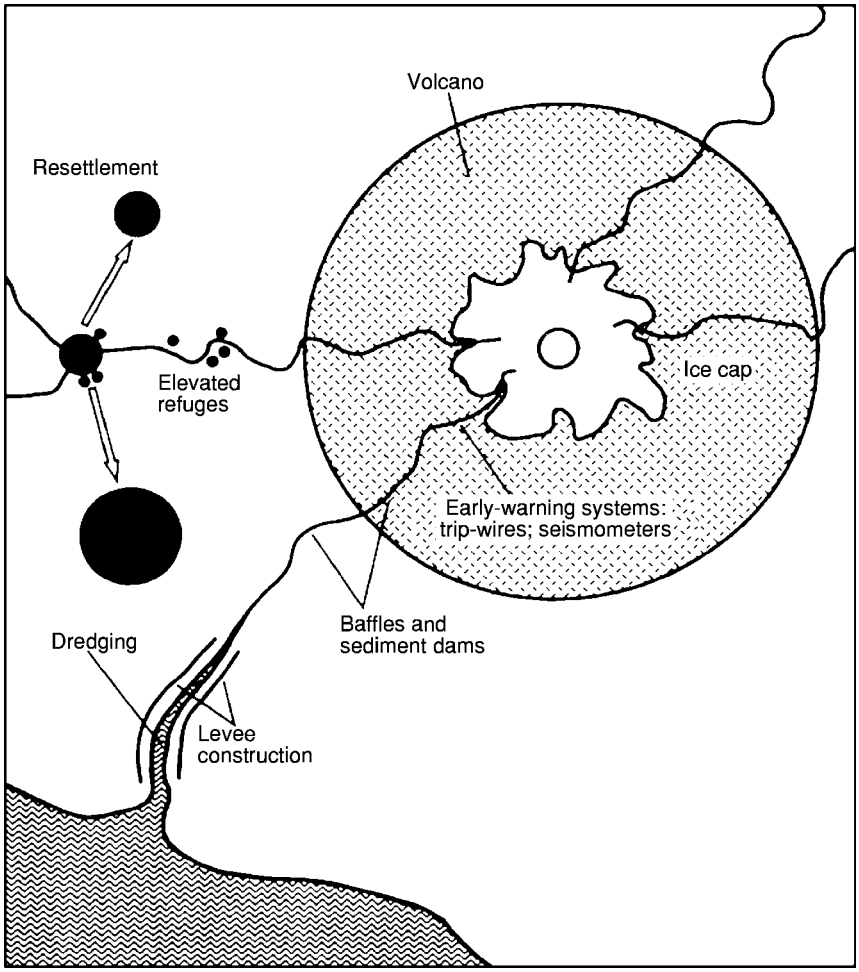


Figure 14 Mitigation measures possible in the case of lahars (from Maguire (1988)).

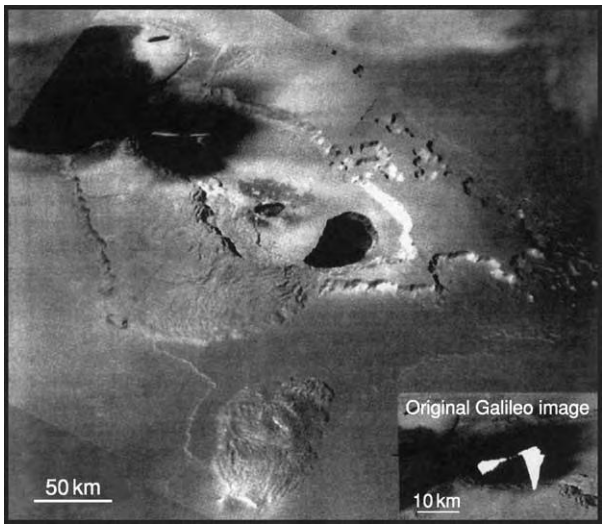


Figure 15 Tvashtar Catena, Io: a 160 km long caldera containing two nested smaller lava filled pits and an adventive volcano at the far end, associated with huge lava fountains and black lava effusion down the flanks of the host volcano. There is also a volcano of more viscous lava with a small summit crater in the foreground. Lighter areas are due to sulphur (NASA Galileo 1999 image).

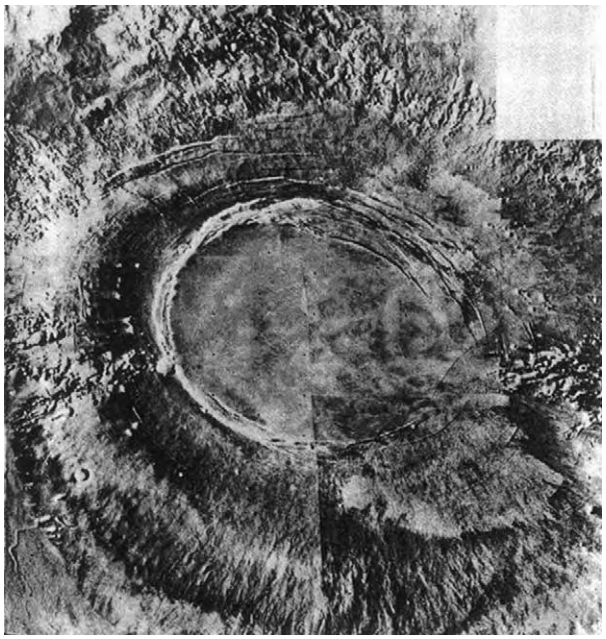


Figure 16 Arsia Mons volcano, Mars, with a 99 km wide caldera (NASA Viking Orbiter 1 image).

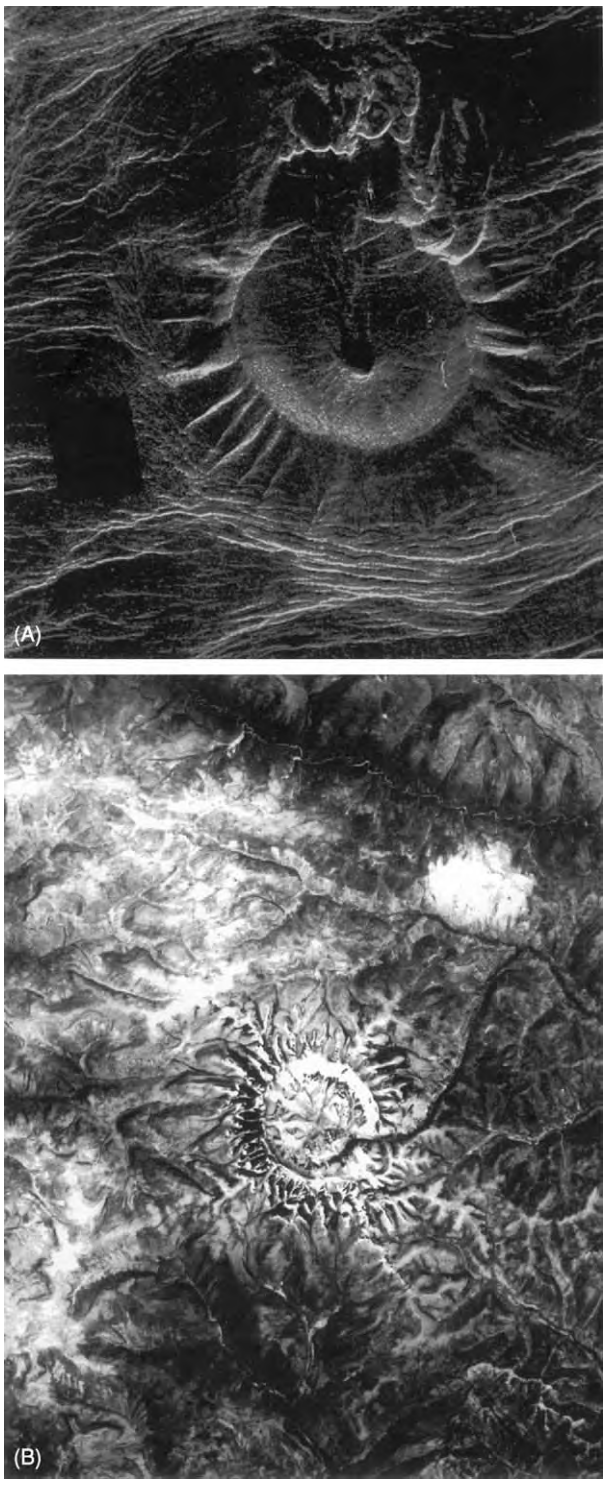


Figure 17 (A) False colour radar image of a volcanic crater on Venus (35 km diameter at summit). Note lava flow which has passed through a breach in the wall at top; also the eroded (?) lateral slope ridges and gullies which suggest Earth like processes operating (NASA image). (B) The Gora Konder crater, diameter 22 km, Siberia: in remote terrain, the origin, volcanic or impact is unknown, illustrating the difficulty of differentiating impact and volcanism on physiography alone: the similarity of the flank ridges and gullies to a) is striking (NASA, Shuttle Discovery image).

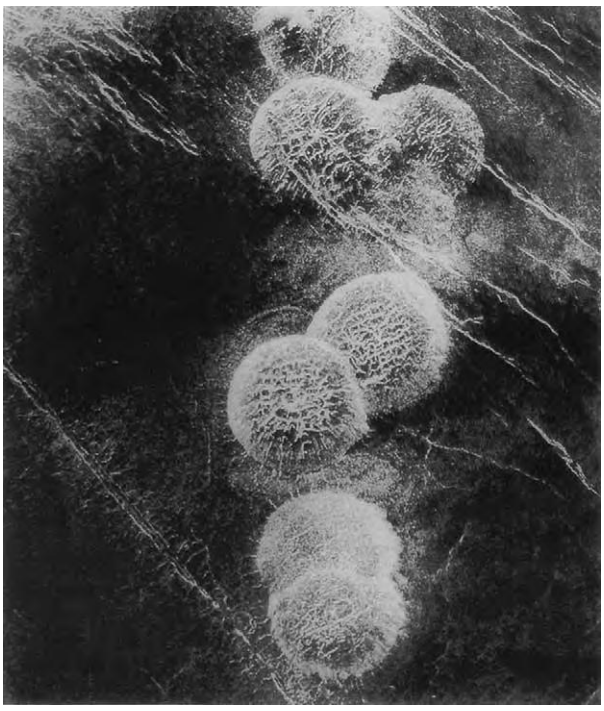


Figure 18 False colour radar image of seven eruptive domes, each 25 km in diameter, overlying one another, on Venus: possibly domes of viscous silicic lava? (NASA image).

a major earthquake in Chile in 1960 is recorded as activating several Andean volcanoes and a major earthquake in Sicily is recorded as having been preceded three days earlier by an eruption of Etna.

Extra-terrestrial volcanoes

Active volcanoes are known on Jupiter's satellite Io (Figure 15), and huge extinct volcanoes occur on Mars and Venus (Figures 16, 17 and 18). Venus's thick CO_2 rich atmosphere is believed to be volcanically sourced.

See Also

Earth: Mantle; Crust. **Engineering Geology:** Natural and Anthropogenic Geohazards. **Lava**. **Mantle Plumes and Hot Spots**. **Solar System:** Venus; Mars; Jupiter, Saturn and Their Moons.

Further Reading

Davies A and Bowler S (2001) Extra terrestrial active volcanism. *Geoscientist* 11(8): 4 7.

Green J and Short NM (1971) *Volcanic Landforms and Surface Features A Photographic Atlas and Glossary.* New York, Heidelberg, Berlin: Springer.

Guest JE, Cole PD, Duncan AM, and Chester DK (2003) *Volcanoes in Southern Italy*. London: Geological Society, Earth in View Series.

Maguire WJ (1998) Volcanic hazards and their mitigation. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology*, pp. 79 95. London: Geological Society, Engineering Geology Special Publication 15.

McCall GJH (1956) Geology of the Gwasi Area. *Report No. 45, Geological Survey of Kenya*.

McCall GJH and Bristow CM (1965) An introductory account of Suswa volcano. *Bulletin Volcanologique* 28: 1 35.

McCall GJH, Laming DJC, and Scott SC (1992) *Geohazards: Natural and Man made*. London: Chapman and Hall.

McCall GJH, Le Maitre RW, Malahoff A, Robinson GP, and Stephenson PJ (1970) The Geology and Geophysics of the Ambrym Caldera, New Hebrides. *Bulletin Volcanologique* XXXIV 3: 681 696.

Pyle D (1998) *Volcanoes*. London: Oceania.

Scarth A (1997) *Savage Earth*. London: HarperCollins.

Smith M, Dunkley PN, Deino A, Williams LAJ, and McCall GJH (1995) Geochronology, stratigraphy and structural geology of Silali volcano. *Journal of the Geological Society* 152: 293 310.

WEATHERING

W B Whalley and P A Warke, Queen's University Belfast, Belfast, UK

Introduction

Weathering can be defined as the irreversible structural and/or mineralogical breakdown of rock through the cumulative effects of chemical, physical, and biological processes operating at or near the Earth's surface (Table 1). However, this seemingly straightforward definition masks the complexity of rock weathering, in which interactions between the many different components of the weathering system give rise to an element of unpredictability that is characteristic of non-linear systems. The weathering behaviour of rock is a response to subaerial (Earth surface) conditions, where temperatures and pressures differ from those under which the minerals were formed. Consequently, adjustment to surface environments is manifest through rock breakdown, the rate of which is controlled by many factors: characteristics of the rock itself, the availability of weathering agents such as salt and moisture, biological agents such as lichens, and, especially, the microclimatic environment to which the rock is exposed.

Without weathering and, in particular, the breakdown of one mineral type into another, there would be no soils of any significance and little scope for widespread development of flora and fauna on land. Thus, long-term weathering is of paramount importance to the biosphere and is a crucial element of both long-term and short-term landscape development.

Nonequilibrium Conditions and the Lithological Cycle – General Significance

The lithological cycle provides a useful starting point when considering the role of weathering in landscape development. Erosion is generally preceded by a combination of weathering processes, which are usually crucial in the formation of silt, clays, and resultant solutes and in the release of residual components of crustal materials.

As rock approaches the Earth's surface, either through tectonic uplift or through erosion of the overburden, associated changes in pressure and/or temperature mean that it is no longer in a state of equilibrium. In this context, 'surface' should be taken to include the range of locations where the hydrosphere, atmosphere, and biosphere interact with the lithosphere; the maximum extension of these interactions is *ca*. 100 m (e.g. where there is tropical deep weathering), although it is normally much less than this. Differences in pressure and temperature at or near the Earth's surface give rise to important (intrinsic) aspects of rock breakdown that can create or reinforce positive-feedback conditions in weathering systems.

- Thermodynamically and chemically different conditions at the Earth's surface can destabilize minerals, thus increasing their susceptibility to subaerial weathering processes.
- Changing chemical and thermodynamic conditions may be accompanied by volume changes resulting from decreased overburden pressures, which lead to differential stresses that are realized as discontinuities at various scales, from joints and cracks to microcracks. Volume changes can also occur when one mineral is altered to another.

When the rock arrives at the Earth's surface, the interplay of hydrosphere, atmosphere, and biosphere provides more complexity, although, for the most part, we can reduce this to a small number of extrinsic factors, namely water (in all three phases), temperature (usually between -40°C and $+40^{\circ}$C at or near the Earth's surface), and biotic activity (which depends on water and temperature). The interplay of these factors has had important consequences for long-term global climate change as well as landscape development, as discussed below.

Joints, Cracks, and Microcracks

One major result of the uplift of rocks to the Earth's surface is that, as pressure decreases, there is a volumetric expansion. The most significant way in which this manifests is through the creation of crack systems, from joints to microcracks. The intersection of joints, which can be many metres in length and depth, can substantially weaken a mass of rock. This gives rise to the concept of 'rock-mass strength' (as opposed to the 'intact strength' of small blocks). Therefore the 'strength' of the rock on a face in a quarry (rock mass) will differ from the crushing strength of the aggregate (intact strength). At a smaller scale, microcracks are plentiful in many rocks; they are usually a few micrometres wide and perhaps a few centimetres long.

Table 1 Examples of weathering processes and mechanisms

Weathering process	*Weathering mechanisms* (*Main mechanisms outlined, not a full discussion of hypotheses*)
Sometimes referred to as physical or mechanical weathering	
Salt weathering (haloclastis)	**Salt crystallization** In pores and microfractures salt crystallization can result in the creation of expansive stresses in excess of the tensile strength of the rock. Repeated exposure to the stress effects of salt crystallization can result in the disruption of intergranular bonds and a reduction in structural coherence
	Crystalline phase change Changes from dehydrated to hydrated states through the absorption of atmospheric humidity result in volumetric expansion of salt crystals in pores and microfractures. Typically, take up of moisture by an anhydrous salt forms a crystallographically different mineral
	Thermal expansion and contraction Interstitial salt crystals exhibit coefficients of thermal expansion that are often greater than those of the rock minerals that surround them
	Mobilization of silica Under highly alkaline conditions disruption of aluminosilicate minerals can occur together with the dissolution of quartz and silica cement
Frost weathering (macrogelivation)	**Freeze thaw** Repeated freezing results in volumetric expansion of water in pore spaces and fractures. This can also enhance the disruptive hydration effects of swelling clay minerals
	Hydrofracture Moisture freezes in a microfracture sealing off the surface end, and any unfrozen water trapped in the substrate may be forced under pressure, through volumetric expansion of the ice, towards the tip of the microfracture and thus extend it
Thermal weathering (insolation weathering, thermoclastics)	**Thermal fatigue** The effects of insolation weathering arise from differential volumetric expansion of individual mineral grains and/or surface and near surface rock layers in response to repeated (diurnal) heating and cooling
	Thermal shock is a rapid increase in rock surface temperature (typically associated with bush fires). Thermal gradients develop quickly, giving rise to tensile stresses between expanding heated surface rock layers and cooler substrate material
Chemical weathering	**Solution of minerals** occurs as a result of exposure to water, its effectiveness is influenced by contact time, the pH of the water, and the solubility characteristics of the elements of the individual minerals
	Carbonation is the reaction of minerals with 'carbonic acid' (CO_2 dissolved in water) and is a particularly important reaction in limestone weathering
	Hydrolysis is the chemical reaction between hydrogen ions in water and the ions in any mineral structure
	Oxidation is when electrons are lost from an atom and usually describes the reaction with oxygen to form oxides, e.g. the conversion of ferrous to ferric iron (such as the weathering of olivine) with an associated volume increase
	Reduction is when an atom gains an electron, usually under anaerobic conditions (typically, gleys in soils)
	Hydration is an exothermic reaction involving the addition of water to a mineral. It is an important weathering characteristic of many clay minerals, which undergo considerable volumetric expansion when water is incorporated into their crystal lattices, e.g. swelling clay minerals such as bentonite from weathered volcanic ash
Biological weathering	**Chemical dissolution of minerals** through the action of organic and inorganic acids
	Chelation is the removal of metallic ions by chelating agents of organic origin
	Plucking is the dislodgement of rock or mineral fragments through the contraction of lichen thalli and fungal hyphae on drying
	Boring into rock by biota (e.g. snails, sea urchins)
	Fracturing by root penetration and exploitation of joints and cracks.

This table gives only the main mechanisms considered to be involved in rock weathering.

The surface area of rock components (say intact rock blocks) is the area upon which physical and chemical weathering processes act. This surface area is extended substantially by the presence of microcracks, joints, and cracks because, as long as water can reach these surfaces, weathering reactions can act upon them. Such physical and chemical action can further extend microcrack and joint systems by increasing the tensile stresses at crack tips. The interpenetration of discontinuities at any of these scales aids weathering in general.

Minerals and Rates of Weathering

Minerals in rocks can be divided into those that are relatively resistant to weathering (such as quartz), those that are rather vulnerable to weathering (such as olivine), and those that have been produced by weathering. Figure 1 shows the Goldich weathering series (which is the inverse of the Bowen reaction series, in that the minerals that crystallize at the lowest temperatures in the melt are also the most stable in the conditions found at the Earth's surface).

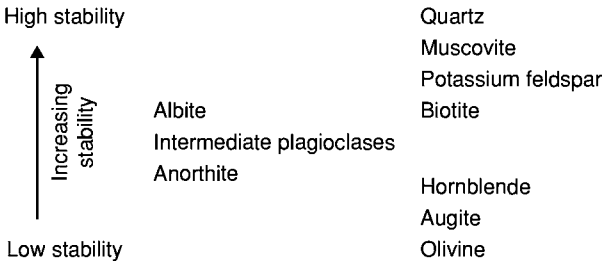


Figure 1 Weathering and the stability of some rock forming minerals. The continuous plagioclase series is given in a separate column.

With the exception of thermoclastis, all Earth-surface weathering (physical, chemical, and biological) involves water, which enables chemical reactions, facilitates the mobilization and ingress of salt into rock fabrics, and, in its solid phase (ice), contributes directly to rock breakdown. Chemical weathering is particularly complex in that the efficacy of reactions depends on moisture availability, temperature, and the nature of the mineral assemblage. Mineral reactions are therefore controlled by the mineralogical nature of the reactants and the acidity (pH) and reduction–oxidation potential (Eh or redox potential) of the water. The dissociation of water into H^+ (protons or hydrogen ions) and OH (hydroxyl radicals) is important in weathering reactions, especially in hydrolysis, where one or other of these ions replaces ions in a mineral structure.

Weathering products are of three main types: layer silicates in the clay minerals, silica in solution (usually given the formula H_4SiO_4 – 'silicic acid'), and alkali-metal ions in solution.

The example of orthoclase weathering to muscovite can be expressed as:

$$3KAlSi_3O_8 + 12H_2O + 2H^+ \rightarrow KAl_3Si_3O_{10}(OH)_2 + 6H_4SiO_4 + 2K^+.$$

Muscovite can then degrade into kaolinite:

$$2KAl_3Si_3O_{10}(OH)_2 + 3H_2O + 2H^+ \rightarrow 3Al_2Si_2O_5(OH)_4 + 2K^+$$

(It should be noted that there are several ways of expressing these reactions and that these two examples are rather generalized because of the complex nature of feldspars and clay minerals.)

The significance of water-soluble products, which may be removed by agents such as groundwater, is clearly seen as the solute loads of rivers. Additionally, the relict nature of weathering products not removed in solution can be seen in saprolites and duricrusts, which may be legacies of previous weathering stages.

In terms of rock weathering, one of the most dramatic examples of the role of chemical weathering in landscape development is provided by 'karst' landforms, where the reactions between rainwater, atmospheric gases, and calcareous rocks (typically limestones) can over time create distinctive and dramatic surface features (Figure 2). Rainwater dissolves atmospheric CO_2 to give bicarbonate ions, HCO_3, in solution as a weak acid, carbonic acid, as a result of a reversible reaction:

$$H_2O(l) + CO_2(g)\{ReversReact\}H^+(aq) + HCO_3(aq)$$

Calcium carbonate is an important constituent of chalks and limestones and is soluble in carbonic acid:

$$CaCO_3 + H_2CO_3\{ReversReact\}Ca^{++} + 2HCO_3$$

These reactions also create subsurface weathering structures, such as cave systems, and features such as stalagmites, which are formed when the reaction is reversed and calcite is deposited.

The rate laws of chemical reactions show how rates depend on the concentrations of the reactants. Although this is significant in reactions that take place over very short periods of time (usually less than a few hours), it can be ignored for our purposes. More significant is the application of Arrehenius' equation to chemical reactions. In particular, increases in temperature speed up reactions. A general rule is that an increase in temperature of 10 K doubles the 'weathering rate'. Clearly, there are direct links with the extrinsic factor of climatic temperature. Assuming a constant (mean annual) temperature at a location, we might reasonably expect, at least over relatively short periods of time such as a few thousand years, the rate of weathering to be constant. When a distinctive weathering product is produced, this can be used as a marker or even as a somewhat primitive chronometer. Rock varnish is a good example of the latter. Weathering horizons in soils buried by subsequent deposition (palaeosols) also provide distinctive marker horizons.

Weathering in Surface Landscape Interpretation

Weathering studies have tended to be somewhat overshadowed by other geomorphological disciplines, and yet weathering is extremely significant because it is the principal means of sediment production for erosion, transport, and deposition by aeolian and fluvial processes. There has been a tendency to investigate more visually prominent weathering features, such as tafoni (shallow rounded cavities in rocks produced by weathering), but it is now widely accepted that such features are exceptional and not truly

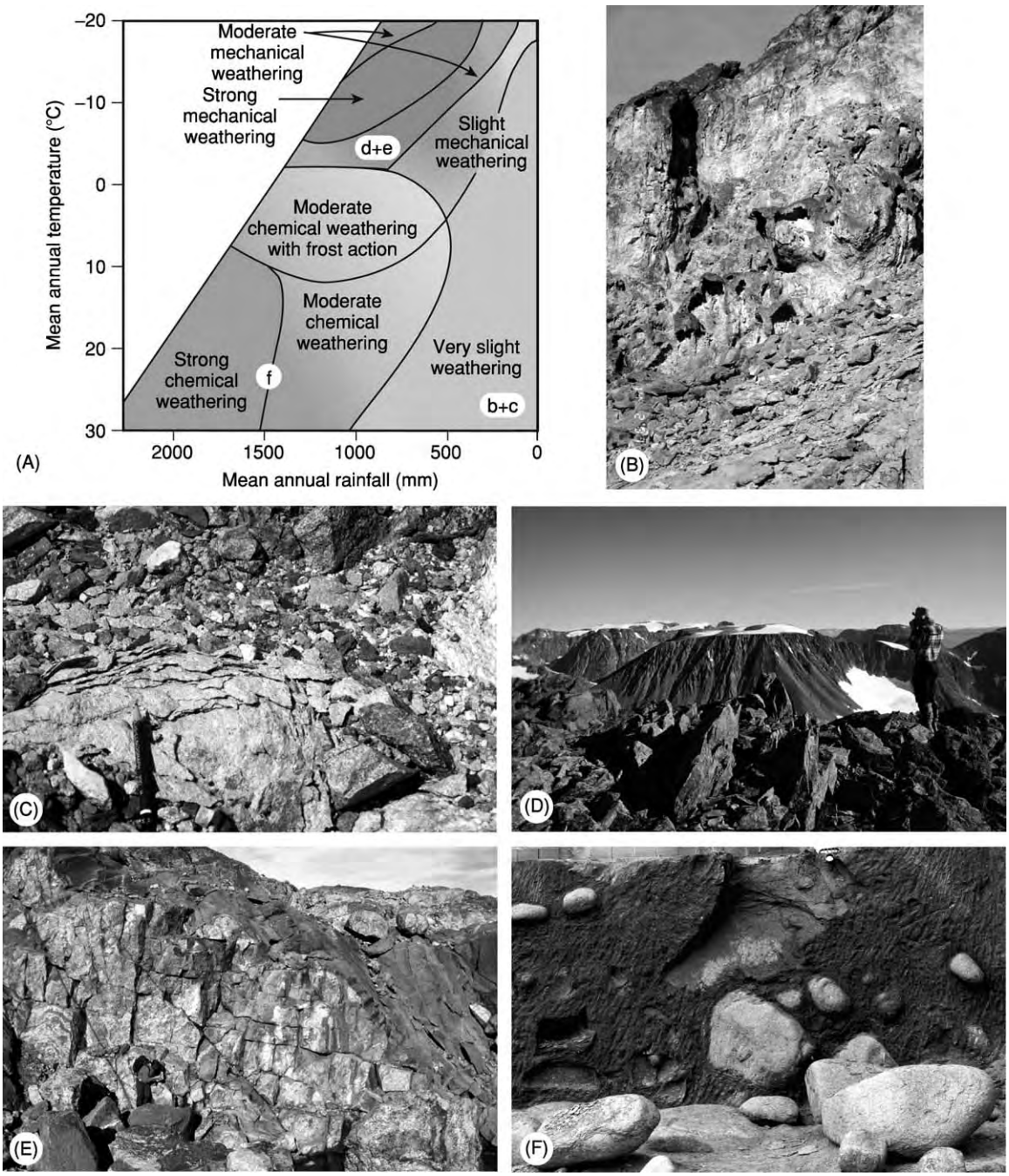


Figure 2 (A) Peltier's zonal classification of weathering, based on average annual temperature and moisture availability. Images shown in parts (B) (F) are marked on the corresponding weathering zones. (B) Present day weathering of a sandstone outcrop in Death Valley, California (a hyperarid hot desert); breaching of the casehardened outer layers of stone and subsequent development of basal tafoni is contributing to undercutting and periodic collapse and retreat of the rock face. Present day weathering processes comprise a complex assemblage of physical, chemical, and biological mechanisms, which are of both spatially and temporally variable effectiveness (cliff height is approximately 4 m). (C) Also in Death Valley, the active disintegration of clasts occurs under present day conditions primarily but not solely because of the effect of groundwater that is rich in a complex assemblage of salts, which facilitates salt weathering mechanisms and the preferential chemical weathering of silicates owing to the strongly alkaline conditions. Owing to the often prolonged absence of surface water, debris disintegrates *in situ*, leading to a local accumulation of sediment, which is periodically transported away from the site during flash flood events. (D) High latitude regions, such as this site in the Lyngen Alps, northern Norway (70°N), tend to be associated with physical weathering by frost action because of the present day climatic conditions, which are dominated by low average annual temperatures. However, this inference is not wholly accurate as the gabbroic rock of the region has been subject to long term chemical weathering (perhaps extending as far back as the Tertiary), leading to an accumulation of clay minerals in a weathering mantle on the plateaux that can be up to 2 m deep. Thus, the present day environment belies the long term activity and changing environmental conditions, especially temperature. This inheritance is also

representative of most landscapes, which are dominated by extensive debris mantles (saprolites) and where weathering phenomena tend to be more mundane. Increasing understanding of the factors controlling weathering phenomena in a wide variety of climatic regions, in the latter half of the twentieth century, has highlighted several major conceptual issues that underpin much of contemporary rock-weathering research – issues that have significant implications for the understanding of landscape development. These issues include: climate, rock weathering, and classification; magnitude and frequency of rock weathering; feedback mechanisms; equifinality (form convergence) – the problem of linking process and form; and inheritance effects.

Climate, Rock Weathering, and Classification

The traditional classification of weathering tends to be climatically zonal, reflecting temperature characteristics and precipitation, with the implication that maximum weathering occurs in the humid tropics and minimum weathering occurs in hot and cold desert regions. However, better understanding of the conditions under which specific weathering processes operate demonstrates that the efficacy of these processes cannot necessarily be clearly associated with specific climatic zones, primarily because of the temporal and spatial variability of conditions at the rock–air interface (Figure 3). For example, chemical weathering is relevant to rock breakdown in arid environments because the spatial variability of microenvironments can create atypical 'pockets' of activity. Many early schemes neglected the temporal aspects of weathering systems, whereby, if sufficient time is allowed, many weathering features hitherto ascribed to specific climatic zones can develop in regions with quite different climatic parameters. For example, the development of karstic (limestone) weathering phenomena associated with moisture availability proceeds slowly in many present-day hot arid environments but rapidly in colder conditions (because of the greater solubility of carbon dioxide in water at low temperatures).

Further complexity is introduced by the temporal variability of conditions at the rock–air interface, whereby long-term (tens to thousands of years) and/or short-term (diurnal and seasonal) environmental changes can alter the nature and extent of rock weathering at a specific location. This is particularly important in hot and cold deserts, where small changes in environmental parameters may have significant effects on weathering activity. Furthermore, viewing weathering as a consequence of the existing climate neglects the effects of inheritance (see below), which are important when considering the landscape as a whole and as a continuum. It is now accepted that, rather than trying to ascribe particular weathering processes to specific climatic zones, it is better to view weathering features as azonal phenomena *per se*.

Magnitude and Frequency in Rock Weathering

Sudden high-magnitude increases in the stress burden on rock through the operation of weathering processes may result in rapid or catastrophic breakdown of the rock fabric. Breakdown might not have occurred, or might have occurred less dramatically, if the stress had been applied more gradually. For example, the freezing rate and the number of oscillations across the freezing point of water in rock (which may be several degrees below 0°C) have been shown to be important in determining the efficacy of frost weathering. The rapid freezing of moisture within the rock fabric reduces losses through evaporation and 'cryosuction' (converting liquid water to solid), increasing the energy available for shattering. This in turn is related to the threshold concept, whereby, if the critical threshold of material strength is greater than the stress applied, no apparent change will occur. The absence of obvious visual damage in a rock can, however, be misleading, as it implies that the material is unaltered, even though microscopic external and internal changes may have occurred, the accumulation of which may eventually reduce cohesive strength and lead to 'fatigue' failure. Similarly, the effects of extreme heating and high rates of surface temperature change (thermal shock) are clearly demonstrated in the natural environment during bush fires, where widespread splitting and spalling of natural rock outcrops occurs as critical thresholds of strength are exceeded over very short periods of time (minutes). Depending on the nature of the

seen in the concordance of plateau summits, which are thought to be Mesozoic in age and probably the result of long term *in situ* chemical weathering producing an 'etch plain'. The effects of glacial activity and periglacial weathering in this environment are far less important than would be implied by a superficial view regarding only present day processes. (E) The gabbroic cliffs (Øksfordsjøkel plateau, northern Norway) show clear evidence of jointing and removal by small rockfalls as well as glacial activity, illustrating the importance of weathering and crack or joint frequency in the wider context of erosive agents. Unlike in (D), the deeply weathered plateau blockfield material has been removed, and the weathering seen is mainly 'cryogenic'; nevertheless, chemical weathering continues to act on mafic minerals, producing a weathering rind. (F) Corestones remain in a deeply weathered Brazilian saprolite as a consequence of long term chemical weathering processes. Without a good stratigraphical marker, the age of these deposits and time taken to form such materials is indeterminate (profile section depth is approximately 2 m).

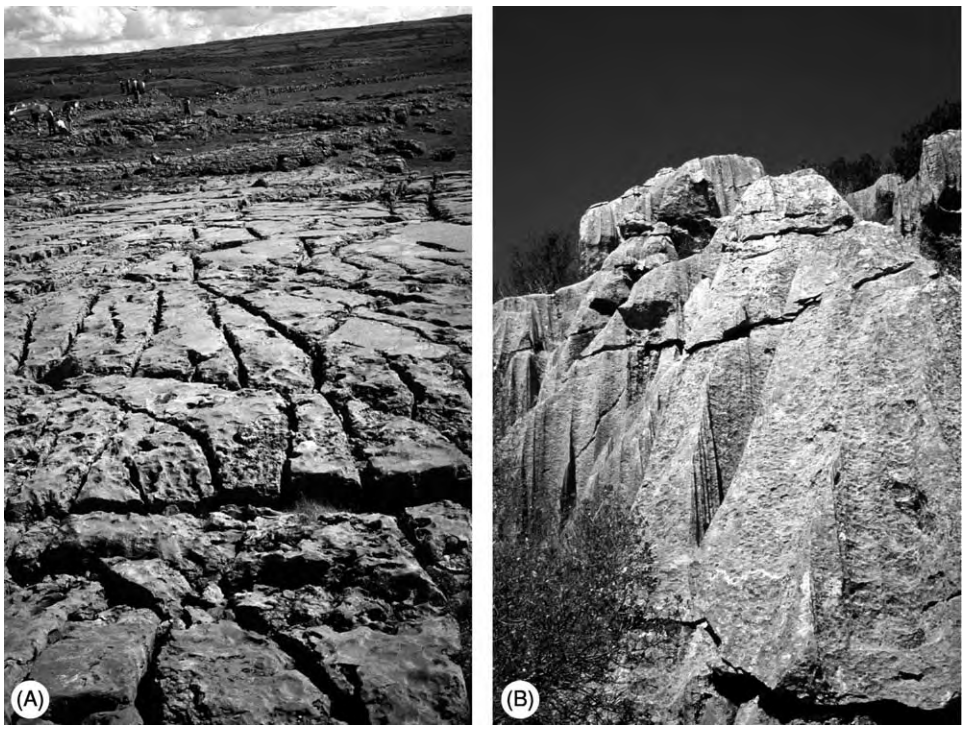


Figure 3 (A) Limestone pavement ('clints' and 'grykes') developed in the horizontally bedded Carboniferous limestone of the Burren, County Clare, Ireland. This limestone has a low porosity, resulting in the chemical dissolution effect of rainwater being concentrated along vertical joints, which form naturally occurring lines of weakness. (B) In contrast, the Jurassic limestone outcrops of the Sierra Norte mountains in Mallorca have been contorted and tilted by orogenesis, with the resultant flow of rainwater over surfaces preferentially weathering joints to create a different limestone landscape of pinnacles and towers on which a hierarchy of smaller solutional weathering features have formed.

environmental stress, prolonged and continuous stressing of rock may be less destructive than – or not so obviously destructive as – repeated high-magnitude short-term events.

Although the magnitudes and frequencies of stress events are important, material characteristics must also be considered. For example, the impermeability of some rock types may leave them relatively unaffected by high-magnitude weathering events, primarily because weathering agents such as moisture and salt cannot penetrate the rock fabric because of a lack of microcracks. Other rock types, exposed to the same environmental conditions, may be more susceptible because their structural and mineralogical characteristics facilitate the ingress of exploitative weathering agents. Consequently, in a landscape where different lithologies are present, it is probable that the nature and extent of the weathering response will vary.

Feedback Mechanisms

The weathering system may be best characterized as a nonlinear system primarily because of the variety of pathways and the unpredictability of outcomes. The unpredictable nature of rock-weathering pathways reflects the operation of positive- and negative-feedback conditions, with the former resulting in an overall change in system state through the acceleration of weathering activity while the latter maintains the system status quo or acts to retard weathering. The trigger for a change in feedback conditions can be seemingly insignificant but may result in considerable system destabilization. For example, a change in groundwater conditions can be sufficient to initiate widespread salt weathering of both natural rocks and manmade building material. Similarly, a change in microenvironmental parameters such as atmospheric humidity may activate salts that have hitherto been present but 'dormant' as regards weathering of rock.

Feedback conditions may help to explain some anomalous rock weathering responses. Part of a rock outcrop, for example, may exhibit well-developed weathering phenomena while an adjacent part of the same outcrop with the same exposure conditions shows no obvious evidence of deterioration. Examples of this are shown in Figure 4. Figure 4A shows a combination of basal tafoni with smaller hollows (alveoli) forming above; in Figure 4B there are no basal hollows but there are well-developed

Figure 4 (A) Sandstone outcrop at Capitol Reef, Utah, showing well developed basal tafoni and alveolar weathering features some 2 3 metres above ground level. (B) Sandstone monolith in Arches National Park, Utah, with well developed alveoli formed some 2 3 m above ground level.

alveoli perched several metres above ground level. In both examples alveoli have formed adjacent to seemingly intact rock that has been exposed to the same environmental conditions and has the same general lithological characteristics. Positive-feedback will have influenced decay pathways initially through the creation of conditions conducive to the initiation of these hollows and subsequently through their progressive development. Initiation of the alveoli may have occurred as a result of structural or compositional anomalies within the sandstone, which were preferentially weathered because they allowed moisture or salt to penetrate. Hollow development may have been slow until increasing size allowed the establishment of microenvironmental conditions that were conducive to the more intense action of various weathering processes not effective on the outcrop surface. The weathering significance of such features lies not only in their development but also in the reasons why the adjacent rock remains relatively unaffected.

Equifinality – The Problem of Linking Process and Form

Another complicating factor affecting rock weathering is the issue of equifinality. This arises when different weathering processes produce similar weathering forms, thus preventing the identification of simple correlations between process, climate, and form. For example, angular shattered debris is common in both hot desert and high-latitude or mountainous regions. In the former, rock typically weathers *in situ* with little or no fluvial abrasion in an environment characterized by large diurnal temperature fluctuations and limited moisture availability. In the latter, moisture is normally more abundant and air temperatures frequently fall below 0°C. In these different climatic environments different groups of weathering processes contrive to produce similar debris forms. The similarity of clasts from these two environments is a consequence of the elongation of crack tips, typically by salt crystals in hot deserts and by ice in cold deserts. However, it should be noted that salt weathering is not exclusive to hot deserts: the sculptured cavernous weathering features (tafoni) observed in the Dry Valleys of Antarctica are attributed to the action of salt weathering rather than freeze–thaw processes.

At the landform scale, a classic example is that of 'U'-shaped valleys, which were once (and still are by some) viewed as being characteristic of glaciated landscapes. However, these features can also be found in the subtropics, where intense chemical weathering at the water table contributes to 'basal sapping' (weathering plus erosion of products) and

the maintenance of steep-sided valleys. In glaciated regions vertical-sided 'U' shapes can be a consequence of rock sheeting ('exfoliation'; as seen in the granites of Yosemite) and massive block failure. Both can be considered to be large-scale weathering phenomena involving overburden removal by long-term denudation.

Inheritance Effects

Many weathering studies have sought to explain patterns and rates of breakdown in terms of the prevailing environmental conditions and have sometimes relied upon the inference of process from 'form' alone at various scales, from landscape through to individual mineral grains. Too often, weathering processes that are immediately associated with observed rock weathering forms are judged to be solely responsible for their formation. In fact, this 'guilt by association' obscures the role of previous events and conditions – the weathering history of the material.

Exposure to various environmental conditions through either long-term climatic change or the spatial relocation of clasts gives a rock a complex weathering history. Rock outcrops and debris in many present-day landscapes therefore carry within their fabric an inheritance of structural and mineralogical weaknesses incurred under former conditions, which influence their response to present-day processes.

In addition to depending on the prevailing environmental conditions, weathering rates depend on the physical and mineralogical characteristics of the rock, and any factor that alters these characteristics will therefore influence subsequent rates of decay. Such changes may result from present-day weathering processes. However, many changes may be cumulative, each reflecting a set of environmental conditions and/or processes that acted in the past. Indeed, the significance of inheritance effects lies in the fact that they are frequently unrelated to present-day environmental conditions but were incurred under former conditions. For example, in arid environments many weathering features owe their initiation and development to the greater availability of water in the past. In Death Valley, now a hyperarid desert, the influence of water on landscape development is clearly evident, with extensive alluvial fans and, at a smaller scale, corroded cobbles, which can be found along the ancient strandlines that mark lake high stands from the early Holocene. These cobbles exhibit well-developed hollows covered by a manganese- and iron-rich rock varnish, indicating that the processes responsible for their development are no longer active under the present-day hyperarid conditions. Such weathering features created under former conditions are preserved within the present-day landscape and while some may be 'inactive' others may be undergoing some limited modification.

Weathering, Near-Surface Diagenesis and the Geological Record

Weathering of a variety of rocks, and sediments in particular, provides useful evidence of past climatic conditions. The weathering products can be substantial or present just as relict traces. Soils in the geological record (palaeosols) are found in a wide variety of sediments (including loess and river gravels) and in extrusive igneous rocks (typically Tertiary basalts). Substantial deposits related to weathering are exemplified by 'duricrusts'. These are hard durable layers or concretions found exposed (and even actively forming) on the surface of the Earth, typically in semiarid areas. The best known are laterite (Fe and Al rich), calcrete ($CaCO_3$), ferricrete (Fe^{3+}), gypcrete ($CaSO_4$), and silcrete (SiO_2). The main formative process is the upward flow of water with evaporation at the surface. This tends to be a self-limiting process, as the thickening duricrust eventually limits its own development by restricting evaporation.

Silcretes are orthoquartzites formed from fluids containing high concentrations of silica in solution (silicic acid), and silcrete remnants form the triliths at Stonehenge. The deposit from which these were derived (by weathering and removal of the surrounding uncemented sand) has been denuded but *in situ* silcretes have been found in the Paris basin, which forms the southern side of the Tertiary basin of southern England.

Weathering Related to Engineering and Economic Geology

As weathering and weathering-derived products are near-surface phenomena, they have important implications for engineering and economic geology. Frequently these implications are problematical; for example, deep weathering in relation to dams, other containment structures, foundations etc. In tropical areas, deep piling techniques may have to be used for high-rise buildings or substantial thicknesses of deep weathering may have to be removed to reach rockhead. Weathering processes influence many aspects of economic geology, not only by making ores (or traces of ore deposits) visible but also by producing valuable deposits in their own right. Of the former, perhaps the weathering of kimberlite to leave diamonds in alluvial deposits is the most famous.

It is important to remember that the weathering processes that affect rock in the natural environment also operate in urban environments, where their effectiveness can be enhanced by atmospheric pollutants, particularly through the effects of 'acid deposition', which results from increased concentrations of atmospheric SO_2, SO_3, NO_2, and NO (sometimes referred to colloquially as SOx and NOx) dissolved in rainwater. Although carbonates suffer most from such attacks, especially but not uniquely in industrial areas, other rock types used for building facades also suffer. Feldspars in certain granites are susceptible, as are calcareous sandstones. The main reactions are the result of atmospheric CO_2 dissolved in water (carbonic acid) and acid deposition (most commonly as aqueous H_2SO_4) on calcium carbonate, which is a common constituent of many rock types.

$$CaCO_3(s) + H_2SO_4(aq) \rightarrow Ca^{2+}(aq) + SO_4^{2}\ (aq) + H_2O(l) + CO_2(g)$$

In this reaction (which is also believed to occur in some natural karstic systems in addition to the usual CO_2 reaction), rainwater removes the dissolved, although sparingly soluble, calcium sulphate. Where there is no running water, or where there is air containing water droplets of low pH (occult precipitation), hydrated calcium sulphate (gypsum; $CaSO_4 \bullet 2H_2O$) can form. Where 'soot' particles are also scavenged from the local atmosphere and precipitated with the gypsum, 'black crusts' may form. There is still debate as to whether such crusts protect the underlying stone or exacerbate decay.

Biological Influences on Weathering

The traditional view of biological factors in rock weathering relates to tree roots penetrating preexisting cracks in rocks, forcing mechanical breakdown. This is probably insignificant in comparison with the effects of micro-organisms on (and very near) rock surfaces and, in particular, bacterial action and root gas–water exchanges in soils. The metabolism of plants (and bacteria) produces organic acids (e.g. humic acid and fulvic acid), which can lower the pH of soil water and directly attack some minerals such as calcite. In addition, roots can increase gaseous CO_2 thus giving rise to high concentrations of HCO_3 and increased acidity. The earliest soils (which were poor in metabolically important elements) were probably altered by the action of prokaryotic and early eukaryotic organisms, which broke down minerals and enriched soils with iron and phosphorous by chelation mechanisms.

The very presence of root structures and decaying vegetation not only increases CO_2 concentrations but also helps to retain moisture, allowing chemical weathering reactions to take place around soil particles. It has been shown that in modern streams the rate of chemical weathering is about seven times higher in a forested area than in surrounding bare areas.

Long-Term Changes in Weathering – Some Complicating Factors

There are several ways in which changes in weathering rates have had a marked effect on the geological record. The weathering of calcium and magnesium silicate minerals removes CO_2 from the atmosphere. This may be represented by

$$CO_2 + CaSiO_3\{\text{Revers React}\}CaCO_3 + SiO_2$$

The concentration of CO_2, moderated by weathering, may influence climate. A computer simulation by Berner estimated levels of CO_2 in the Phanerozoic and showed a marked drop in the later Palaeozoic. This has been interpreted as the result of an increase in weathering during this period. The following aspects are significant.

In the Devonian land plants became more diverse and colonized previously barren land (especially upland). This extension of forested areas resulted in an increase in weathering. Atmospheric CO_2 continued to fall during the Carboniferous, even after the spread of forests, and this is attributed to burial of those forests in swampy areas (CO_2 sequestration). This total decline in atmospheric CO_2 is assumed to have weakened the effect of greenhouse warming and thus induced, or at least exacerbated, an extension of the southern hemisphere ice-sheets at about this time. An increase in levels of atmospheric CO_2 in the Mesozoic is attributed in part to minimal amounts of orogenesis (which increases weathering and removes CO_2 from the atmosphere) but mainly to increased emission during the metamorphism of calcareous oozes in subduction zones. In the Cenozoic a slow decline in the CO_2 concentration is attributed to orogenesis, in particular that of the Himalaya–Karakorum chains. This is an area of investigation that has been prompted by the work of Ruddiman and Raymo and is still an area of considerable research. A further factor is the position of land masses relative to the tropics: higher temperatures and precipitation increase weathering, and a lack of vegetation in more extreme conditions may reduce weathering (thus decreasing CO_2 sequestration). Clearly, the area of continental masses exposed to

prevailing climatic belts also affects the total amount of weathering and is part of the self-regulating system of global climate. For example, increases in precipitation as a result of maritime influences increase terrestrial vegetation, weathering, and CO_2 capture, reducing the greenhouse effect and leading to decreased temperatures (and evaporation rates). Overall, temperature and precipitation affect the amounts and the extent of weathering. Thus, in moderating atmospheric CO_2, weathering on the Earth's surface exerts substantial negative feedback on global climatic change.

Concluding Comments

Weathering processes rarely, if ever, operate in isolation. Consequently, the weathering forms that we see are probably products of the cumulative and sequential effects of a variety of physical, chemical, and biological processes. Recent investigations focusing on the mechanisms of weathering rather than on the resultant forms have led to a wider and deeper understanding of the importance of weathering in topics ranging from building-stone conservation to long-term controls of climate and evolutionary processes.

See Also

Building Stone. **Clay Minerals**. **Geomorphology**. **Geotechnical Engineering**. **Mining Geology:** Hydrothermal Ores. **Sedimentary Processes:** Karst and Palaeokarst. **Soils:** Modern; Palaeosols.

Further Reading

Berner RA (1991) A model for atmospheric CO_2 over Phanerozoic time. *American Journal of Science* 291: 339–375.

Bland W and Rolls D (1998) *Weathering*. London: Arnold.

Drever JI and Clow DW (1995) Weathering rates in catchments. In: White AF and Brantley SL (eds.) *Chemical Weathering Rates of Silicate Minerals*, pp. 463–483. Reviews in Mineralogy. Washington DC: Mineralogical Society of America.

Goudie AS and Viles HA (1997) *Salt Weathering Hazards*. Chichester: Wiley.

Lsaga AC, Soler JM, Ganor J, Burch TE, and Nagy KL (1994) Chemical weathering rate laws and global geochemical cycles. *Geochimica et Cosmochimica Acta* 58: 2361–2386.

Ollier CD (1984) *Weathering*. London: Longman.

Peltier L (1950) The geographic cycle in periglacial regions as it is related to climatic geomorphology. *Annals of the Association of American Geographers* 40: 214–236.

Phillips JD (1999) *Earth Surface Systems*. Oxford: Blackwell.

Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Ruddiman WF and Raymo ME (1988) Northern Hemisphere climate regimes during the past 3 Ma: possible tectonic connections. *Philosophical Transactions of the Royal Society, Series B* 318: 411–430.

Thomas MF (1994) *Geomorphology in the Tropics: A Study of Weathering and Denudation in Low Latitudes*. Chichester: Wiley.

Twidale CR (1982) *Granite Landforms*. Amsterdam: Elsevier.

White WB (1988) *Geomorphology and Hydrology of Karst Terrains*. New York: Oxford University Press.

Yatsu E (1988) *The Nature of Weathering*. Tokyo: Sozosha.

Index

NOTE

Bold page number locators refer to complete articles on the various topics covered by this encyclopedia. Illustrations and tables are indicated by *italic* page numbers.

Text is located by page numbers in normal print.

Cross references, prefixed by *see* and *see also*, are also listed at the end of each article.

C

D

E

F

H

K

M

N

O

P

T

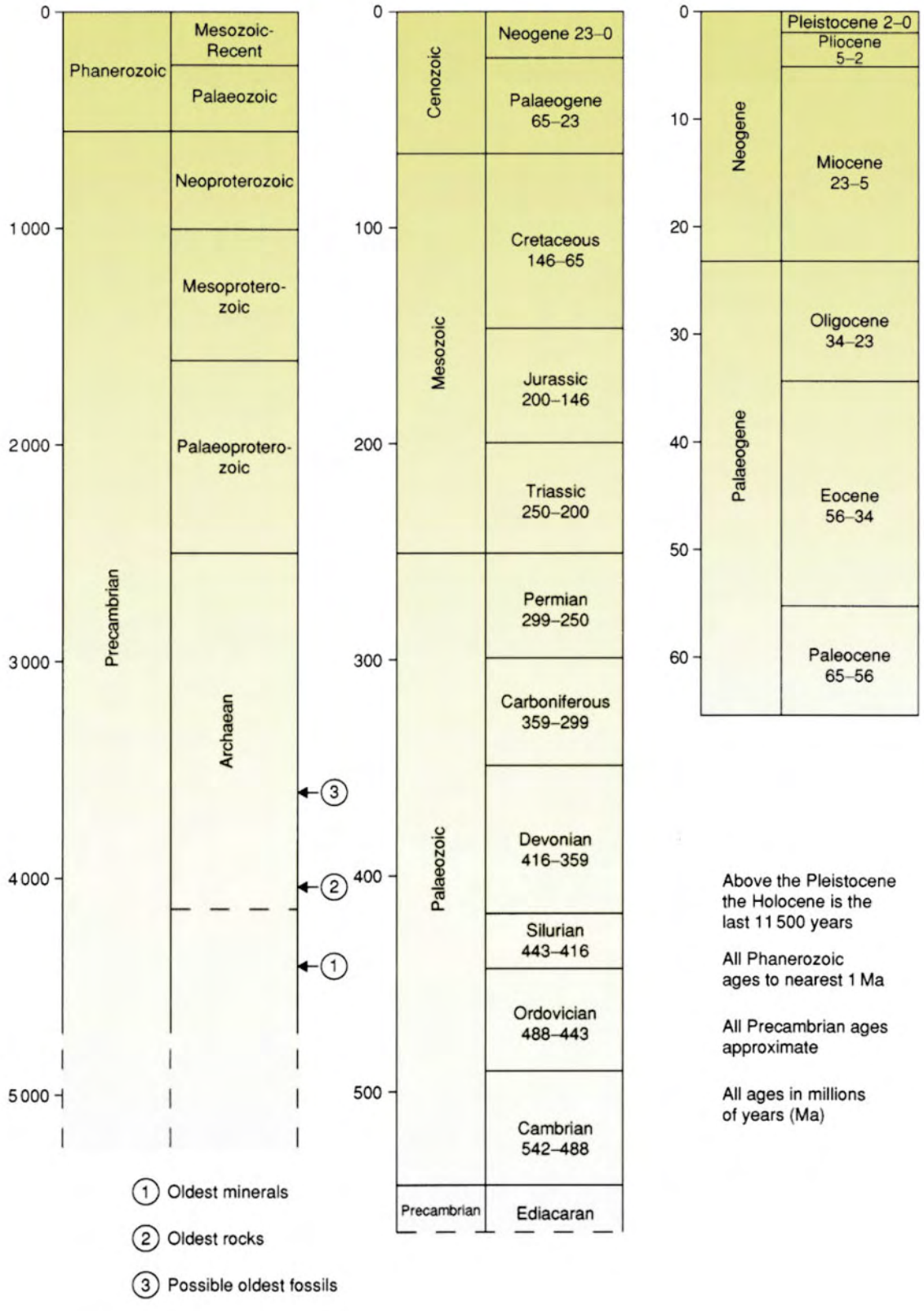
U

W

X

Y

Z


0
1 000
2 000
3 000
4 000
5 000
Phanerozoic
Precambrian
Mesozoic-Recent
Palaeozoic
Neoproterozoic
Mesoprotero-zoic
Palaeoprotero-zoic
Archaean
3
2
1
0
100
200
300
400
500
Cenozoic
Mesozoic
Palaeozoic
Precambrian
Neogene 23–0
Palaeogene 65–23
Cretaceous 146–65
Jurassic 200–146
Triassic 250–200
Permian 299–250
Carboniferous 359–299
Devonian 416–359
Silurian 443–416
Ordovician 488–443
Cambrian 542–488
Ediacaran
0
10
20
30
40
50
60
Neogene
Palaeogene
Pleistocene 2–0
Pliocene 5–2
Miocene 23–5
Oligocene 34–23
Eocene 56–34
Paleocene 65–56
1 Oldest minerals
2 Oldest rocks
3 Possible oldest fossils
Above the Pleistocene the Holocene is the last 11 500 years
All Phanerozoic ages to nearest 1 Ma
All Precambrian ages approximate
All ages in millions of years (Ma)